



第四届全国集成微系统建模 与仿真学术交流会

2025年8月22-24日 || 浙江·湖州

会议文集



目录

A: 异质异构集成/先进材料/多物理场建模仿真

报告/文章编号	题目	页码
特邀-鲍华	GiftBTE: 基于声子玻尔兹曼输运方程的高效跨尺度热输运仿真	1
特邀-陈文超	三维异质异构集成工艺-多物理场联合建模仿真方法与应用	2
特邀-宋志刚	铋化物超晶格红外探测器建模仿真	3
特邀-张欢欢	射频系统多物理场仿真: 从数值算法到智能驱动	4
特邀-陈士涛	新型 MEMS 集成器件多物理场电磁、微磁、力、热协同建模及仿真设计	5
特邀-蒲波	异质异构集成芯片系统信号和电源完整性及多物理场仿真技术研究	6
企业特邀-陈士涛	新型 MEMS 集成器件多物理场电磁、微磁、力、热协同建模及仿真设计	7
企业特邀-王明玉	一种高效的基于体积分方程的直流电阻电感参数提取仿真工具	8
企业-Dave Tay	XYALIS 自动化掩模数据准备 (MDP) 解决方案	9
M101	片上多层叉指电容的几何缩放模型	10
M102	InP 基异质结双极型晶体管质子辐照损伤机理与非理想效应研究	17
M105	宇航 2.5D/3D 微系统集成结构基于有限元仿真技术的结构、材料及工艺优化实践	25
M106	基于宏模型的热电制冷封装系统仿真设计	32
M107	基于异构集成技术的小型化宽带太赫兹集成电路-波导互联结构	40
M108	三维系统级封装的单粒子效应 Geant4 模拟方法研究	51
M109	面向芯粒尺度的多物理场耦合研究	67
M110	低温硅基器件热点邻域的多尺度建模与声子输运机制研究	73
M111	基于 RDL-first 工艺的封装材料仿真分析	80
M112	基于 TSV 硅桥的光电微系统三维异构集成设计与仿真研究	86
M113	多通道高速光电微系统三维异构集成工艺仿真及开发	93
M114	用于 2.5D/3D 射频微系统的硅转接板中互连结构传输特性建模仿真	103
M116	硅基异构集成射频微模组仿真设计	109
M117	RRAM 器件总电离剂量效应建模: 物理机理与紧凑模型方法	119

B: 射频/光电/太赫兹微系统建模仿真

报告/文章编号	题目	页码
特邀-李达	先进封装电磁兼容智能化设计研究	124
特邀-唐旻	射频电路和集成系统的多物理场分析	125
特邀-王大伟	集成微系统多物理耦合精准建模与快速仿真	126
特邀-王维波	微波异构集成芯片系统仿真及三维器件模型提取技术研究	127
特邀-虞绍良	高速互连中的光学设计与光电链路仿真	128

特邀-张傲	InP 高电子迁移率晶体管亚太赫兹建模与表征	129
特邀-周浩	射频微系统技术与先进相控阵天线	130
特邀-施永荣	先进三维集成相控阵前端技术研究探讨	131
企业特邀-赵佳劼	射频前端模组设计与仿真关键技术	132
企业-尹朝卿	基于板级 EDA 的精细化建模解决方案	133
M201	宽带射频微系统器件 S 参数模型时域建模方法研究	134
M202	陶瓷基毫米波变频收发微系统仿真技术研究	144
M204	一种面向射频电路的可连续调节变压器	155
M205	基于不确定性量化的微带滤波器逆向设计	161
M206	铟磷异质结双极晶体管大信号建模及参数提取方法研究	165
M207	110GHz 铟磷异质结双极晶体管小信号建模及参数提取研究	170
M208	基于螺吡喃光致变色的可重构光学物理不可克隆函数仿真研究	175
M209	70GHz 电吸收激光调制器射频模型研究	180
M210	芯粒轻量化热模型与集成微系统快速热仿真应用	184
M211	金属-介电混合 O-PUF 中散斑相关性的调控机制及其响应熵增强效应研究	196
M212	面向多端口互连系统宏建模的自动化极点确定方法	203
M213	四种去嵌入方法在氮化镓高电子迁移率晶体管器件建模中的应用与比较	209
M214	基于快速时域仿真和迁移学习的信号完整性建模方法	214
M215	具有超宽阻带性能的紧凑型双模 SIW 带通滤波器	236
M216	低寄生电感 SIC 模块双面冷却结构协同设计	240
M217	SIC 半桥功率模块电热力协同设计优化与可靠性仿真	243
M218	基于金属光栅-铱-金属镜结构的光电探测器仿真研究	248
M219	基于 Angelov 模型的 GaN HEMT 非线性电容模型的研究	256
M220	多通道高速光电微系统设计与仿真研究	265
M221	超宽带 3D- SiP 仿真技术研究	272
M222	基于高精度多场协同仿真的异构集成 X 波段前端设计	277
M223	基于相位调制的宽范围可调谐射频电感建模与仿真	282

C: MEMS/NEMS 微系统建模仿真

报告/文章编号	题目	页码
特邀-耿文平	集成铁电 MEMS 器件与系统	289
特邀-王坤锋	用于微弱肺音检测的微纳压电谐振式麦克风研究	290
特邀-易志然	压电 MEMS 微型散热风扇系统仿真与优化分析	291
特邀-周再发	MEMS 综合性设计技术发展思考	292

特邀-贾浩	MEMS 多物理场建模仿真与超灵敏传感器	293
企业-孙岩	基于模型的微系统研发数字化转型实践	294
M301	基于多层精密微组装的石英振梁加速度计模态抑制研究	295
M302	一种基于 POD 降阶和状态空间方程的 MEMS 热式流量传感器宏模型提取方法	304
M304	应变下单层 2H-MoTe ₂ 的振动响应	317
M305	MoSe ₂ 纳米机电谐振器的非线性动力学特性研究	320
M306	室温下具备高频率-激光功率响应率的少层 MoSe ₂ 纳米机械谐振器	322
M307	激光可调控的金纳米机电谐振器	325
M308	高精度测定超薄 ReSe ₂ 中复折射率以提高光机械转导效率	326
M309	纳米机电谐振器模式序列的调控研究	330
M310	基于双源混频的频率可重构纳机电频率梳	331
M311	高频-甚高频 CrOCl 纳米机电谐振器	333
M312	基于超微型 MEMS 可调谐超表面的动态太赫兹滤波器	336
M313	聚甲基丙烯酸甲酯辅助层热裂解法制备二维纳机电谐振器阵列	339

D: Chiplet/SOC/IC 建模仿真与 EDA

报告/文章编号	题目	页码
特邀-卜建辉	基于机器学习的先进 CMOS 电路单粒子软错误 SPICE 仿真	342
特邀-游海龙	数据驱动的半导体器件建模方法	343
特邀-张国和	基于强化学习与图神经网络的晶圆级芯片自动化设计框架	344
特邀-赵晨曦	毫米波负载调制式平衡放大器及其衍生结构	345
特邀-侯赛飞	基于二维半导体 DRAM 的工艺优化和存算一体应用探索	346
企业特邀-代文亮	多物理场仿真 EDA 赋能加速 Chiplet 设计	347
M401	集成微系统 AI-EDA 技术	348
M402	芯和三维全波瞬态产品关于 IC 建模的关键网格技术	354
M403	SZ0501 型 SiP 总剂量效应自动化仿真研究	357
M404	CMOS 反相器 X 射线与电磁脉冲协和效应研究	369
M405	一种基于 SPICE 的 LM4040 带隙基准源的电路模型建模方法	380
M406	高精度霍尔式磁编码器信号处理系统的建模与仿真	388
M407	基于 LLMs 生成的模拟芯片行为模型研究	392
M408	用于多芯片集成系统快速稳态热预测的多倍率多尺度图卷积神经网络算法	401
M409	基于 Verilog_A 的系统级封装建模与仿真	408

E: 神经形态芯片与微系统

报告/文章编号	题目	页码
特邀-李炎	领域专用处理器芯片技术及其产业化应用	418
特邀-曲益明	基于图论的栅介质击穿的渗流模型研究	419
M501	跨层误差建模与压缩感知目标检测在存算一体系统上的部署	420
M502	面向存内计算 SoC 的点对点特征处理负荷检测方案	428
M503	基于 STT-MRAM 模拟存内计算的阵列规模分析框架	432
M504	面向多芯片异构集成的 SiP 封装技术	437
M505	55nm 工艺存算一体芯片的架构创新与能效优化	441
M506	面向存算一体架构的 AI 编译器研究进展	445
M507	面向脉冲神经网络的存算一体芯片综述与展望	451
M508	通过噪声感知训练提升存算芯片的推理精度	456
M509	基于注意力的磁随机存储器缺陷分类检测	462

F: 功率器件建模仿真与 EDA

报告/文章编号	题目	页码
特邀-刘斯扬	GaN 功率 HEMT 器件 SPICE 模型研究进展	466
特邀-王曦	SiC 光控功率器件的建模与仿真	467
M601	基于神经网络的碳化硅 MOSFET PSpice 行为模型	468
M602	1200V SiC 凹状基区光控晶体管的特性仿真与分析	472
M603	一种考虑温度效应的 3.3 kV SiC MOSFET 行为模型	474
M604	具有 2.33mV/°C 温度敏感性的氧化镓 SBD 器件能够实现功率器件的实时结温监测	479
M605	基于 N-P-N 夹层外延片制造全 P 型包裹的 4H-SiC 沟槽 MOSFET 器件的全面研究	485
M606	20kV SiC n-IGBT 集电区参数优化设计	491
M607	温度升高对碳化硅快速离化器件特性影响的分析	496
M608	短路瞬态应力下不同平面栅结构对 SiC MOSFET 内多物理场行为的影响	502
M609	基于单粒子辐照损伤的碳化硅 MOSFET 栅极泄漏电流退化模型研究	507
M610	扩流型 SiC MOSFET 动态电阻的机理研究	511
M611	SiC 双沟槽 MOSFET 短路特性与失效机制分析	519
M612	基于电-热-力多物理场的 SiC 功率模块封装结构优化	524
M613	超结功率 MOSFET 反向恢复特性仿真研究	530
M614	超结 MOSFET 器件 EMI 优化仿真分析	537

GiftBTE：基于声子玻尔兹曼输运方程的高效跨尺度热输运仿真

鲍华

上海交通大学

摘要：在亚微米尺度下，基于傅里叶定律的传统热扩散方程不再适用。为了解决亚微米尺度热输运的高效仿真问题，课题组开发了一套开源通用软件包 GiftBTE，可以用于高效的求解非灰声子玻尔兹曼输运方程（BTE）。GiftBTE 具有极高的计算效率，能够实现三维体系的非灰声子 BTE 的全尺寸仿真。通过进一步与第一原理计算相结合，可以对任意结构的亚微米热传输过程进行无拟合参数的准确的计算。GiftBTE 的应用包括但不限于计算各种纳米结构的等效导热率、预测纳米器件中的温度和热流、以及模拟涉及小热点或超快过程的材料上的激光加热。基于该方法开展了系列微纳尺度导热的研究，揭示了分子动力学模拟中的声子输运机制，量化多孔石墨烯中热输运的多种波动效应，研究了纳米热点附近的热输运过程，开展了鳍式场效应晶体管中的自热效应仿真。

三维异质异构集成工艺-多物理场联合建模仿真方法 与应用

陈文超

浙江大学

摘要：异质异构集成工艺过程中退火、晶格失配引起残余应力，这些工艺过程残余应力与自热效应引起的电-热-力耦合过程叠加，对有源器件性能、无源器件电迁移演化产生重要影响，从而通过影响寄生参数、晶体管开关速度等性能恶化信号完整性和电源完整性。报告从工艺过程建模、多物理效应对先进有源器件性能影响机制与分析、多物理效应对键合性能及可靠性影响等层面展开，揭示工艺过程效应与多物理效应叠加对多种器件性能影响机理，探索器件工艺协同优化（DTCO）及系统工艺协同优化技术（STCO）。

铋化物超晶格红外探测器建模仿真

宋志刚

中国科学院半导体研究所

摘要：本报告主要介绍团队近年来在铋化物超晶格红外探测器建模仿真领域取得的部分进展，基于8带模型和平面波展开方法，给出了超晶格能带和光吸收特性。并结合漂移扩散输运方法，模拟超晶格载流子输运行为。进一步，从单元探测器到焦平面阵列，建立了串扰、噪声等效温差及调制传递函数的计算方法。

射频系统多物理场仿真：从数值算法到智能驱动

张欢欢

西安电子科技大学

摘要：在射频系统设计中，以电磁场为核心的多物理场耦合问题（如电磁场-温度场-应力场相互作用）已成为制约高性能系统研发的关键瓶颈，使得先进射频系统的设计高度依赖精准高效的多物理场仿真技术。早期多物理场仿真主要通过数值算法（如有限元法、有限差分法等）求解物理场控制方程实现，近年来，随着人工智能技术的快速发展，基于数据驱动与物理驱动的智能多物理场仿真方法逐渐成为该领域的研究前沿。本报告将系统介绍报告人团队近年来在射频系统多物理场仿真算法方面的工作，包括基于数值算法的多物理场仿真方法、基于人工智能的多物理场仿真方法及相关工程应用。

新型MEMS集成器件多物理场电磁、微磁、力、热 协同建模及仿真设计

陈士涛

合肥电热芯电子科技有限公司

摘要：万物互联和智能工业化发展为射频集成电路器件和光 MEMS 微纳结构器件的高密度异质集成带来了全新的发展机遇，构建逼近物理真实的建模和工程 EDA 难度很大，但对芯片高质量工艺和性能的发展至关重要。传统的射频系统内部的高密度异质集成的多物理场电磁—力—热往往是弱耦合效应，两两之间是单向耦合作用。但是随着集成系统不停的小型化需求，这就对器件微型化提出更高的要求，需要进行新材料、新器件和新机理的研究。本报告将以具体微纳尺度射频 MEMS 器件为例，讨论微波集成电路新型器件 BAW 滤波器、乃至磁电天线芯片内部的电磁、微磁、力、热强相互耦合效应，以及 BAW 滤波器芯片设计方法的研究。除此之外，针对光 MEMS 微镜芯片的力—电磁—热、声阻尼场的多场耦合问题提出新的解决方案和思路。针对高密度异质集成的多物理场电磁、微磁、力、热强相互耦合问题，报告人还提出了自主多物理场仿真技术，解决国外 ANSYS 软件多物理场仿真“卡脖子”技术难题，本报告人提出的该项技术目前全球范围只对标 COMSOL 软件，甚至计划未来 5-10 年超越 COMSOL 软件仿真能力！

异质异构集成芯片系统信号和电源完整性及多物理场仿真技术研究

蒲波

宁波德图科技有限公司

摘要：当前半导体逐渐接近现有工艺和材料体系下的摩尔定律极限，实现在特定面积下的大规模晶体管集成存在技术和成本上的重大挑战。为了克服这个挑战，异质异构集成芯片技术逐渐获得了学术界和工业界的重要关注。由于 Chiplet 系统具有高集成度、高带宽、高功耗和多尺度的结构和场景特点，带来在高密度信号完整性、高功率供电电源完整性以及散热和应力可靠性等一系列多物理场问题。本报告将集中围绕 Chiplet 信号完整性的设计和优化、实现可靠电源完整性及通过多物理场分析保障 Chiplet 可靠性进行阐述，以寻求一条在后摩尔时代有效的 Chiplet 设计和仿真优化之路。

新型MEMS集成器件多物理场电磁、微磁、力、热 协同建模及仿真设计

陈士涛

合肥电热芯电子科技有限公司

摘要：万物互联和智能工业化发展为射频集成电路器件和光 MEMS 微纳结构器件的高密度异质集成带来了全新的发展机遇，构建逼近物理真实的建模和工程 EDA 难度很大，但对芯片高质量工艺和性能的发展至关重要。传统的射频系统内部的高密度异质集成的多物理场电磁—力—热往往是弱耦合效应，两两之间是单向耦合作用。但是随着集成系统不停的小型化需求，这就对器件微型化提出更高的要求，需要进行新材料、新器件和新机理的研究。本报告将以具体微纳尺度射频 MEMS 器件为例，讨论微波集成电路新型器件 BAW 滤波器、乃至磁电天线芯片内部的电磁、微磁、力、热强相互耦合效应，以及 BAW 滤波器芯片设计方法的研究。除此之外，针对光 MEMS 微镜芯片的力—电磁—热、声阻尼场的多场耦合问题提出新的解决方案和思路。针对高密度异质集成的多物理场电磁、微磁、力、热强相互耦合问题，报告人还提出了自主多物理场仿真技术，解决国外 ANSYS 软件多物理场仿真“卡脖子”技术难题，本报告人提出的该项技术目前全球范围只对标 COMSOL 软件，甚至计划未来 5-10 年超越 COMSOL 软件仿真能力！

一种高效的基于体积分方程的直流电阻电感参数提取仿真工具

王明玉

芯和半导体科技（上海）股份有限公司

摘要：本研究提出了一种网格友好的基于体积分方程的仿真器，用于精确高效地提取直流电阻和电感（DCRL）。该仿真器在离散体积电流时采用一种新颖的质心基函数变换，有效地从系统矩阵中提取体积电流的方向性。这种变换不仅简化了快速多极子方法的实现，还增强了系统矩阵的压缩性，从而与传统方法相比显著降低了内存消耗和计算时间。此外，该模拟器采用了网格友好的方案，无需对细长键合线进行体网格划分，同时保持了高精度。该模拟器可以精确高效地计算任意形状和大规模结构的 DCRL 参数。

XYALIS 自动化掩模数据准备 (MDP) 解决方案

Dave Tay

Advinno Technologies Pte Ltd

摘要: XYALIS 自动化掩模数据准备 (MDP) 解决方案简介

XYALIS 提供从设计规格 (Specs) 到完整掩模布局 (Mask Layout) 的全自动化流程, 显著减少手动操作时间并提升生产效率。其工具组合包括 GOTframe、GOTmask、GOTfiller 等模块, 支持:

- 多芯片阵列生成与优化 (适用于 MPW 和 3D-IC)
- 掩模顺序表与版图文档自动生成 (支持 .pdf、.csv、.xls、.html 等格式)
- CMP dummy fill、芯片完整性检查、版图合并及调试工具 • Python、Tel/Tk 与 SQLAPI 接口, 实现定制化集成与高效脚本化执行

关键优势

- 自动化流程大幅缩短周期时间 (如 STMicro 从 3 周缩短至 1 小时)
- 提升硅片使用效率与良率 (如 Mosis 的 MPW 优化)
- 强大的 AI 求解器与图形化调试工具, 支持可视化与规范验证
- 适用于多层光掩模、曲线版图、光子芯片与先进封装设计

XYALIS 拥有超过 25 年行业经验, 服务于全球前十的半导体厂商, 致力于提供定制化的高性能解决方案。

片上多层叉指电容的几何缩放模型

刘依玟¹, 王欢鹏², 吴韵秋^{1*}, 刘杰^{1, 3}, 刘辉华¹, 余益明¹, 赵晨曦¹,
张青凤⁴, 康凯¹

¹ 电子科技大学电子科学与工程学院, 成都 611731

² 电子科技大学长三角研究院(湖州)集成电路与系统研究中心, 湖州 313098

³ 中国电子科技集团公司信息科学研究院, 北京 100043

⁴ 电子科技大学(深圳)高等研究院, 深圳 518110

通讯地址: yqw@uestc.edu.cn

联系方式: 028-61830775

摘要: 多层叉指电容凭借其高集成度特性, 在紧凑空间内展现出优异的性能表现。本文构建了一种适用于多层叉指电容的可缩放等效电路模型, 通过叉指结构尺寸准确推导电路参数, 并实现对其电学特性的有效预测。经验证, 该缩放模型在 0~40 GHz 频率范围内的误差小于 7.4%。这种基于几何尺寸缩放的建模方法, 不仅能够实现电路参数的精准预测, 也为片上无源器件的快速设计与优化提供了可靠的理论支撑。

关键词: 叉指电容; 等效电路模型; 多层结构; 集成电路

1 引言

叉指电容的结构特征在于其交叉的指状电极, 该结构使其具备优异的高频性能, 以及集成度高的特点。叉指电容可广泛应用于射频滤波器、阻抗匹配网络等电路设计中; 凭借其出色的匹配特性, 还可应用于数模混合电路中以提升线性度[1]。随着芯片集成密度的持续提升, 多层金属结构使叉指电容能够在更小的面积内实现更优越的性能。

针对叉指电容的等效电路建模, 传统方法主要依赖于从测试数据[2]中提取电路参数, 或针对单层叉指电容结构[3]进行计算。单层叉指结构中的主要等效电容可通过经验公式[4]获得。文献[5]计算了叉指结构中的特定耦合电容, 并在 12 GHz 频率下进行了验证。近年来, 一些研究创新性地利用神经网络构建了电容的可缩放模型[6], 该模型对单层叉指电容展现出优异的预测精度。然而, 针对多层金属结构叉指电容的等效电路参数计算方法仍较为缺乏, 难以准确预测其电学特性。

本文提出了一种适用于片上多层叉指电容的可缩放等效电路模型, 该模型建立了电容高频特性与其尺寸参数之间的显式依赖关系。模型精度在 40 GHz 以下频段经多个不同尺寸样品的测试数据验证。该模型能有效预测片上叉指电容特性, 为射频集成电路设计提供了有效工具, 在早期设计探索与电路优化过程中具有重要价值。

2 多层叉指电容的可缩放等效电路模型

为建立引言所述适用于片上多层叉指电容的可缩放等效电路模型，本研究首先建立其电路基本拓扑结构，继而结合物理参数与电磁场理论推导其尺寸缩放规律。

2.1 基础结构及等效电路模型

本文涉及的片上多层叉指电容的基础结构示意图及其等效电路模型如图 1 所示。

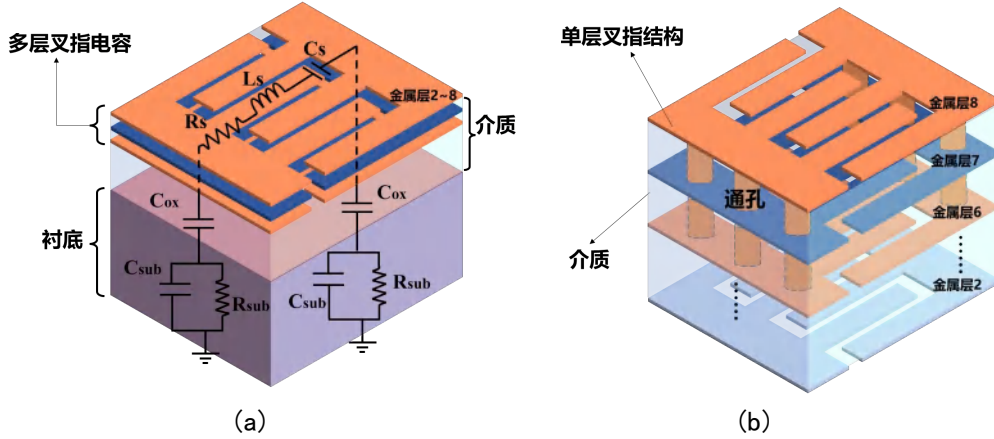


图 1 多层叉指结构示意图(a)等效电路模型；(b) 层间细节示意图

在图 1(a)中所示的元件 C_s 表征多层叉指结构的主要串联等效电容，分布于两端叉指电极的电容 C_{ox} 主要表示由金属结构下方氧化层中的垂直电场引起的容性效应。串联电阻 R_s 源于电容金属的阻性损耗，其在信号传输方向上的性质与矩形金属相似，其阻值与叉指长度成正比，与叉指宽度成反比。对于单指结构，可以基于传输线理论得到感性损耗 L_s 。多指结构的电感为多个单指电感串联构成，而多层结构的电感则按并联方式处理。由电容 C_{sub} 和电阻 R_{sub} 组成的并联支路用于表征衬底损耗效应。在相同工艺条件下，衬底相关参数可视为固定值。层间垂直互连的细节结构如图 1(b)所示，可见连接多层结构的互连通孔主要分布在两端电极区域。

2.2 多层等效电容 C_s

对于单层叉指电容，可以用[3]中提出的方法计算：

$$C_p = (\epsilon_r + 1)L[A_1 \times (2N - 3) + A_2] \quad (1)$$

上式中，由内侧叉指和外侧叉指引入的有效面积 A_1 和 A_2 主要取决于氧化物厚度和指宽。参数 L 和 N 为指长和指数， ϵ_r 为氧化物材料的相对介电常数。

然而，在多层结构叉指电容结构中，每一层叉指下方的“接地面”并非理想无限大平面，而是结构相同的下层叉指结构。因此，电场无法被约束在单层叉指区域内，而是会通过叉指缝隙向层间区域泄露，形成附加的边缘电容分量。该横向电容分量可由公式(2)来估计：

$$C_l = \frac{\lambda h_m (2N - 1) L \varepsilon_0 \varepsilon_r}{d_g} \quad (2)$$

层间介质的厚度 h_m 和指隙 d_g 对电容有较大影响，其中系数 λ 用于校正该缝隙区域电场泄露的实际程度。

此外，公式(1)的计算仅考虑了叉指主体结构本身的电容。而在多层结构中，两端电极抽头区域同样贡献电容分量。该分量可采用类似平行板电容的计算思路进行建模：

$$C_0 = \frac{\varepsilon_0 \varepsilon_r h W_t}{d_t} \quad (3)$$

式中， W_t 和 d_t 表示电极两端抽头部分的宽度和间距， W_t 与指宽正相关，且 d_t 与指长正相关。

综上，多层叉指电容模型中的主要串联电容 C_s 值可通过式(4)将上述三个构成部分叠加得到。

$$C_s = C_p + C_l + C_0 \quad (4)$$

2.3 垂直氧化层电容 C_{ox}

在传统模型[2]中，垂直氧化层电容通常被简化为每个电极单独贡献的平行板电容，其有效面积基于叉指平面投影计算。

然而，这种简化模型对于多层叉指电容结构而言并不准确，可能引入不可忽略的误差。多层金属结构在电路中所激发的电场分布更加复杂，电极下方氧化层中的电场并非均匀且严格垂直于两个电极表面分布。此外，由于叉指结构两端指状电极区域高度紧凑，两个端口电极会受到部分共同电场的影响。因此，本文模型中，氧化层电容仍采用平行板电容公式进行估算，并将其等效地作用于电容两端电极。需特别指出，在本研究所涉及的样品中，叉指电容两端电极具有类似传输线的延伸结构，该部分连接至两侧测试焊盘，其所占据面积引入的氧化层电容分量不可忽视。

除上述主体氧化层电容外，多层叉指电容的金属侧壁厚度由于密集通孔连接多层金属而不可忽视。侧壁金属通过周围介质产生额外的寄生电容，其值可由下式计算：

$$C_{fringe} = \varepsilon_0 \varepsilon_r W_{total} \left(\frac{4}{\pi} \right) \ln \left(1 + \frac{h_{top}}{h_{bottom}} \right) \alpha \quad (5)$$

其中，参数 W_{total} 表示多层叉指电容的总侧壁宽度，和叉指宽度与指隙强相关。参数 h_{top} 和 h_{bottom} 分别代表从衬底表面到顶层金属和底层金属的距离。此外，引入系数 α 对侧壁金属产生的边缘效应进行校正。根据本研究样品的尺寸特征，系数 α 的取值确定为 $(90/WN)$ ，表明其边缘效应的影响程度与叉指宽度和指数成负相关，总宽度越大，边缘效应的影响越弱。

3 模型验证与讨论

为验证本文提出的可缩放等效电路模型,选取多组具有不同尺寸特征的多层叉指电容样品进行测试,并在 0~40 GHz 频率范围内进行模型验证。验证工作聚焦于叉指宽度和长度变化对模型参数及电容特性的影响。

3.1 叉指宽度对模型参数及性能的影响

选取叉指长度为 100 μm 、叉指数为 3、金属层数为 7 层的三组样品,仅改变叉指宽度,保持其他工艺与尺寸参数一致。等效电路模型中的主要参数由本文所提出的缩放规律进行计算,具体数值与模型准确度在下表中列出:

表 1 不同叉指宽度的等效电路模型参数值及准确度

样品编号	叉指宽度/ μm	R_s/Ω	L_s/pH	C_s/fF	C_{ox}/fF	准确度
W-1	20	2.85	13.23	89.5	111.6	96.4%
W-2	30	2.10	12.00	91.3	138.4	95.2%
W-3	40	1.42	13.26	93	167.1	92.6%

由表 1 所展示的信息,可知在该叉指宽度变化范围内,等效串联电阻 R_s 所受影响较小,但依然呈现随着指宽增大而减小的趋势,这主要源于金属横截面积的增大。表征感性损耗的 L_s 与特性阻抗相关,应随着指宽的增大呈现先减小后增大的趋势,表 1 所选样品数据也能够体现这一规律。叉指宽度的变化对垂直氧化层电容 C_{ox} 的影响最为显著,因为指宽的增加直接增大了氧化层电容的正对面积以及侧壁金属的宽度,从而显著影响 C_{ox} 的两个主要构成分量。同时,多层等效电容 C_s 也随着叉指宽度增大而缓慢增加,这与缩放公式的预期一致,因为构成 C_s 的单层叉指主体电容 C_p 与基础结构电容 C_0 均与指宽正相关。

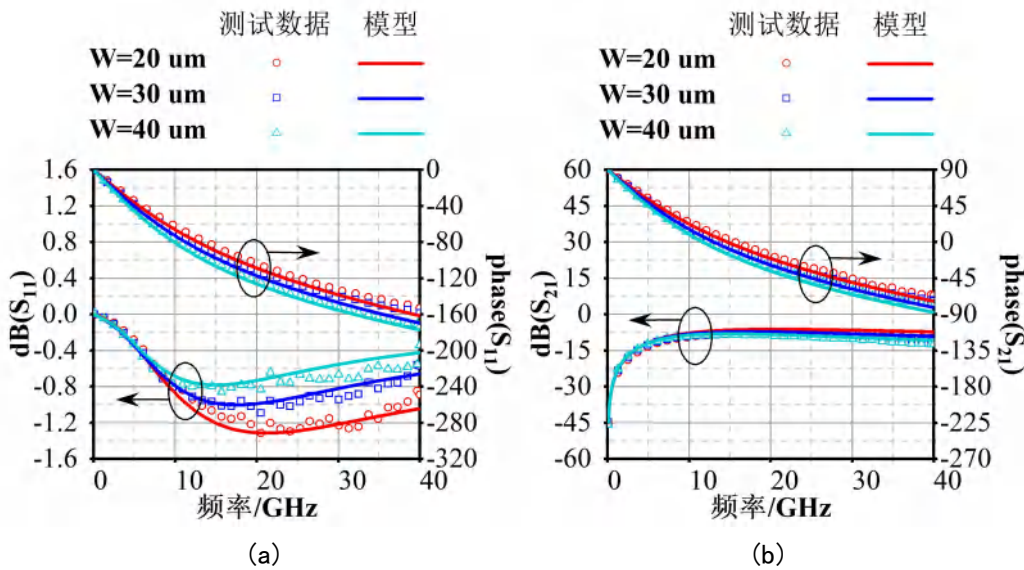


图 2 (a) S_{11} 参数在不同指宽下的比较; (b) S_{21} 参数在不同指宽下的比较

以电路的 S 参数为电学特性的参考指标,图 2 展示了测试数据及等效电路模

型的对比情况。如图 2 所示，本文所提出的可缩放等效电路模型，对不同叉指宽度的样品均展现出优异的电学特性预测能力。

3.2 叉指长度对模型参数及性能的影响

为探究叉指长度的影响，选取三个叉指宽度为 30 μm 、叉指数为 3、由 5 层金属构成的样品，其余尺寸及工艺参数保持一致。利用本文提出的可缩放建模方法进行模型参数计算并展示在下表：

表 2 不同叉指长度的等效电路模型参数值及准确度

样品编号	叉指长度/ μm	R_s/Ω	L_s/pH	C_s/fF	C_{ox}/fF	准确度
L-1	50	0.93	9.23	45.9	34.9	95.7%
L-2	100	1.86	18.00	67.4	43.5	94.2%
L-3	150	2.79	27.69	91.5	52.1	94.5%

根据表 2 中数据可知，较长的叉指长度显著增强了电感效应，并因损耗路径延长而增大了等效电阻 R_s ，符合此前的缩放规律分析。多层等效电容 C_s 主要由多层结构的叉指主体电容 C_p 叠加与边缘泄露电容 C_l 构成，两者均与指长正相关，该规律在表 2 数据中得到了明确体现。另外，垂直氧化层电容 C_{ox} 随着叉指长度的增大而增大，这是由于指长增加了氧化层电容的正对面积。在本文提出的缩放模型中，指长并不影响侧壁电容的大小，因此其对 C_{ox} 的整体影响程度弱于指宽。

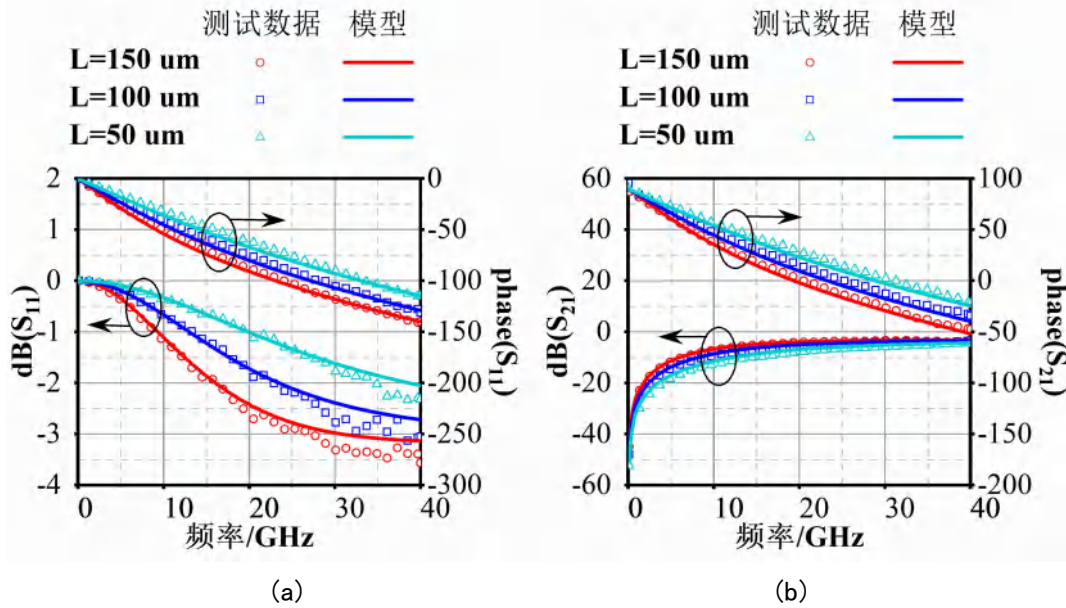


图 3 (a) S_{11} 参数在不同指长下的比较；(b) S_{21} 参数在不同指长下的比较

图 3 对比了该组样品的 S 参数测试数据与模型拟合结果，在 0 ~ 40 GHz 范围内，模型能够准确描述叉指长度变化下多层叉指电容的电学特性。

3.3 关键尺寸参数影响的综合讨论

综合上述针对叉指宽度和长度的验证结果，多层结构本身有效降低了阻性损耗，使得所有样品的 R_s 均维持在较低水平。同时，电感效应对叉指长度的变化

表现出更高的敏感性。在相同的变化比例下，叉指长度对多层等效电容 C_s 的影响显著大于叉指宽度。这一现象主要源于边缘场泄露引起的电容效应影响大于基础结构电容 C_0 。

金属层数对电容特性亦有显著影响。以 W-2 和 L-2 号样品为例，二者仅有金属层的差异，其余关键尺寸完全相同。五层结构的 L-2 样品中多层等效电容 C_s 明显小于七层结构的 W-2 样品，符合多层电容叠加的物理预期。此外，L-2 的底层金属距离衬底更远，导致其垂直氧化层电容 C_{ox} 显著减小。

以上全部模型在 0~40 GHz 范围内与实测数据对比的准确度均达到 92% ~ 97%，这充分表明所提出的可缩放模型能够基于叉指尺寸准确预测多层叉指电容的性能。

4 结论

本文构建了一种适用于片上多层叉指电容的可缩放等效电路模型，并通过实测数据，验证了叉指宽度与长度对其电学特性的影响规律。模型在 0-40 GHz 频段范围内预测准确度可达 92.6% 以上。该模型能够基于叉指几何尺寸参数实现精确的比例缩放，为电路设计及其前期优化提供了可靠的分析工具。

致谢

本文由国家自然科学基金（批准号：62025106）和四川省自然科学基金（批准号：2024NSFSC0466）资助。

参考文献

- [1] Hua Fan, Zhuorui Chen, Tongrui Xu, et al., "14-bit SAR ADC with on-chip digital bubble sorting calibration technology," Chinese Journal of Electronics, vol. 34, no. 1, pp. 125-136, 2025.
- [2] K. Subramaniam, A. V. Kordesch and M. Esa, "Nine-element lumped metal finger capacitor model extraction method for RF applications," in 2006 4th Student Conference on Research and Development, 2006, Shah Alam, Malaysia, pp. 25-28.
- [3] G. D. Alley, "Interdigital capacitors and their application to lumped-element microwave integrated circuits," IEEE Transactions on Microwave Theory and Techniques, vol. 18, no. 12, pp. 1028-1033, Dec. 1970.
- [4] S. Jeyaraman, V. N. R. Vanukuru, D. Nair and A. Chakravorty, "Modeling of high-Q conical inductors and MOM capacitors for millimeter-wave applications," IEEE Transactions on Electron Devices, vol. 67, no. 12, pp. 5646-5652, Dec. 2020.
- [5] C. Iversen, "A high density MIM capacitor in a standard CMOS process," Journal of Semiconductor Technology and Science, vol. 1, no. 3, pp. 189-192, 2001.

- [6] Y. Xiong, S. Guo, S. Hu, Y. Gu, Y. Fang and Y. Guo, "A scalable model for on-chip interdigital capacitor based on space-mapping neural network," in 2024 9th International Conference on Integrated Circuits and Microsystems (ICICM), 2024, Wuhan, China, pp. 93-97.

InP 基异质结双极型晶体管质子辐照损伤机理与非理想效应研究

柳润坤, 张宏伟, 李五星

中国电子科技集团第二十七研究所, 郑州 450047

摘要: InP 基双异质结双极晶体管 (DHBTs) 的最大直流增益的电压偏置点 (0.6 - 0.9V) 与最大频率工作电压 (0.9V 以上) 存在显著偏差, 高偏压下非理想效应及质子辐照会导致增益退化, 制约其空间应用。本文通过 2MeV 质子辐照实验 (剂量 $5 \times 10^{13} \sim 1 \times 10^{14} \text{ H}^+/\text{cm}^2$), 结合 TCAD 仿真, 系统分析高偏压下异质结势垒效应 (HBE) 及辐照损伤机理, 验证了质子辐照对 HBE 的强化作用。

关键词: InP/InGaAs 异质结双极型晶体管; 质子辐照; 异质结势垒效应

1 引言

InP/InGaAs 异质结双极晶体管 (HBT) 凭借 InP 低表面复合速率、高热导率及 InP 与 InGaAs 异质结优异的电子传输特性, 在高频特性与输出功率方面表现突出 [1-4], 为空间领域高频、高带宽信号收发系统提供核心支撑, 展现出广阔应用前景 [5-7]。然而, 太空环境中高能粒子与宇宙射线的持续轰击, 极易引发器件性能劣化乃至失效 [8-10]。其中, 质子因丰度高、能量辐射强且屏蔽难度大 [11-13], 其辐照损伤对器件可靠性的威胁尤为突出。

尽管质子辐照效应研究已受关注, 但高电流密度工况 (器件实现高输出功率的关键场景) 下辐照损伤机制仍不明确, 亟待深入研究。针对上述问题, 本文以 InP 基 HBT 为研究对象, 开展 2 MeV 质子辐照实验, 通过分析不同偏置条件下直流参数 (如电流增益 β 、基区电流) 的退化规律, 发现低注入与高注入偏置均导致器件直流性能显著退化, 高电流密度下异质结势垒效应 (HBE) 的耦合作用, 进一步加剧了损伤程度, 为解析 InP/InGaAs HBT 太空辐照失效机理提供关键依据。

2 器件结构和实验

本研究使用的 InP/InGaAs HBTs 的外延层是通过气源分子束外延 (GSMBE) 在 4 英寸 (100) 晶体取向的半绝缘 InP 衬底上生长的。所有的 InGaAs 和 InGaAsP 层都与 InP 层及衬底层晶格匹配, 器件结构参数如表 1 所示。

表 1 器件的外延 InP 基 HBTs 器参数

Layer	Material	Thickness (nm)	Doping (cm^{-3})
Emitter cap	InP	60	Si 2×10^{19}
Emitter	InP	130	Si 1.2×10^{19}

Emitter	InP	40	Si 2×10^{17}
Base	$\text{In}_{(1-x)}\text{Ga}_x\text{As}$	65	Graded C 3×10^{19}
Collector	$\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$	50	Si 1×10^{16}
Collector	$\text{In}_{0.73}\text{Ga}_{0.27}\text{As}_{0.58}\text{P}_{0.4}$ 2	20	Si 1×10^{17}
Collector	$\text{In}_{0.88}\text{Ga}_{0.12}\text{As}_{0.27}\text{P}_{0.7}$ 3	20	Si 1×10^{17}
Collector	InP	200	Si 1×10^{16}
Sub-Collector	InP	50	Si 1.2×10^{19}
Sub-Collector	$\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$	50	Si 2×10^{19}
Sub-Collector	InP	300	Si 1.2×10^{19}
Buffer	InP	50	undoped
Substrate	InP		Fe

图 1 展示了扫描电子显微镜(SEM)拍摄的 InP 基 HBTs 的俯视图和发射极基极接触的正视图。

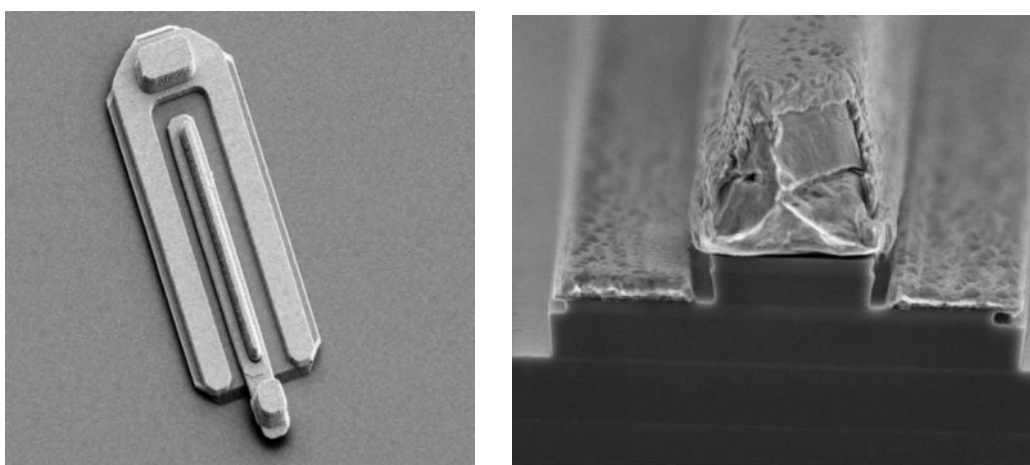


图 1 SEM 拍摄的 InP 基 HBTs 器件的俯视图和基极接触正视图

使用串列静电加速器 EN-18 对 InP 基 HBTs 进行了 2MeV 能量下不同剂量辐照损伤的实验, 辐照剂量分别为 $5 \times 10^{13} \text{ H}^+/\text{cm}^2$ 和 $1 \times 10^{14} \text{ H}^+/\text{cm}^2$, 预计的入射深度约为 $38.4 \mu\text{m}$, 因此 2 MeV 的质子可以很容易地通过晶体管的有源区域。为了防止离子通道效应, 将入射角设置为离器件法线方向 7 度左右。平均剂量率约 $4 \times 10^9 \text{ H}^+ \cdot \text{cm}^{-2} \cdot \text{s}^{-1}$, 辐照过程中器件处于未加偏置的条件下。

3 结果分析

图 2 显示了高 V_{BE} 偏压下 InP 基 DHBTs 器件 2MeV 质子辐照后不同剂量下的正向 Gummel 特性曲线, 质子辐照剂量为 $5 \times 10^{13} \text{ H}^+/\text{cm}^2$ 和 $1 \times 10^{14} \text{ H}^+/\text{cm}^2$, 器件发射极尺寸为 $0.8 \times 10 \mu\text{m}^2$ 。

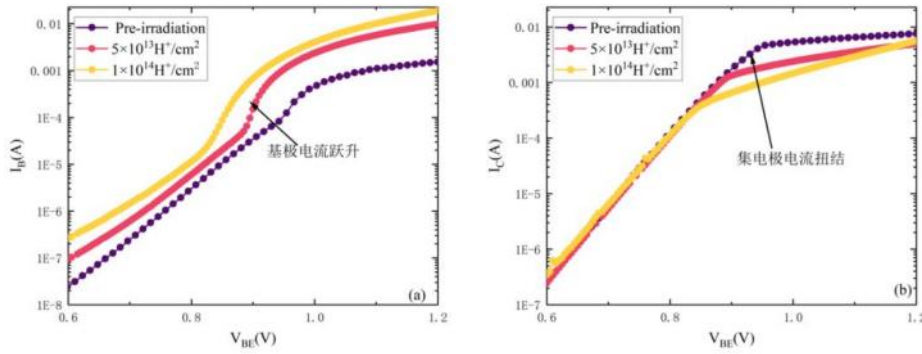


图 2 高偏压下质子辐照后 InP 基 DHBTs 器件正向 Gummel 特性曲线 (a) 基极电流 (b) 集电极电流

结果显示，质子辐照前后 InP 基 DHBTs 器件 Gummel 特性曲线中基极电流和集电极电流都在高 V_{BE} 偏压范围内发生了与以往研究不同异常现象，首先器件基极电流 I_B 在 0.8-0.9V 偏压范围内出现了异常的剧烈增长的跃升现象，而器件输出的集电极电流 I_C 在高偏压范围内的增长速率发生了异常下降的扭结现象。并且质子辐照使 InP 基 DHBTs 器件发生这些现象的 V_{BE} 偏压减小，在质子辐照剂量达到 $1 \times 10^{14} \text{ H}^+/\text{cm}^2$ 时发生这些异常现象的 V_{BE} 偏压减小了 0.12V。

对比图 2 器件正向 Gummel 特性曲线，发现高偏压下器件发生退化的主要原因应该归结于基极电流的剧烈增加，因此需要分析基极电流中何种成分的增加造成了这些非理想效应。通过正向 Gummel 特性曲线计算了 InP 基 DHBTs 器件的基极电流的理想因子，如图 3 所示。

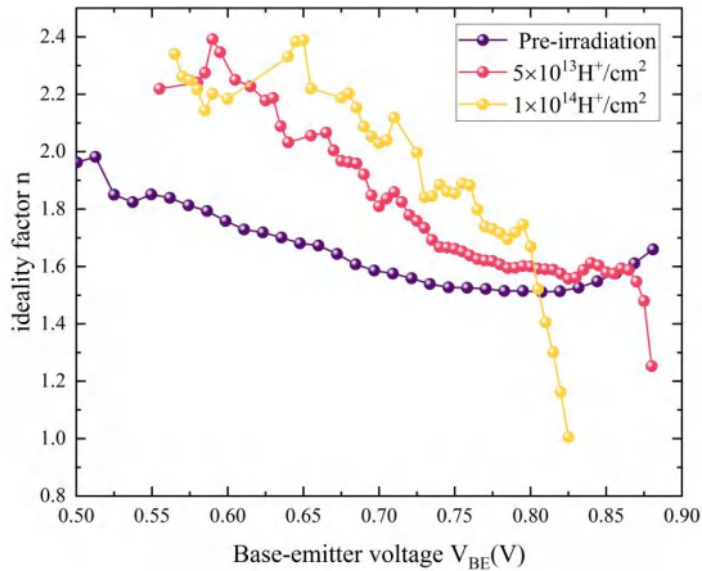


图 3 辐照前后基极电流的理想因子

结果显示质子辐照后当 V_{BE} 大于某个临界电压后基极电流理想因子迅速减小到 1，而随着质子辐照剂量的增加基极电流理想因子迅速减小时的临界电压减小。理想因子接近于 1 表明此时基极电流占据主导地位的为中性基区体复合电流。InP 基 DHBTs 器件工作在高偏压状态下发生的非理想效应会造成器件电学特性的退化，并且质子辐照会使这些非理想效应更早更严重的发生。这些非理想效应

的发生与大注入条件下的异质结势垒效应相关。

根据 InP 基 DHBTs 器件电学特性测试结果和现有器件模拟研究成果，选择了流体动力学模型来描述载流子输运过程，在基本物理模型中添加了辐射复合模型、俄歇（Auger）复合模型、肖克莱（SRH）复合模型、热电子发射模型、隧穿模型以及高迁移率模型。图 4 为 Sentaurus 仿真中 InP 基 DHBTs 器件的三维立体截面图，仿真中的器件尺寸与质子辐照实验使用的尺寸保持一致。

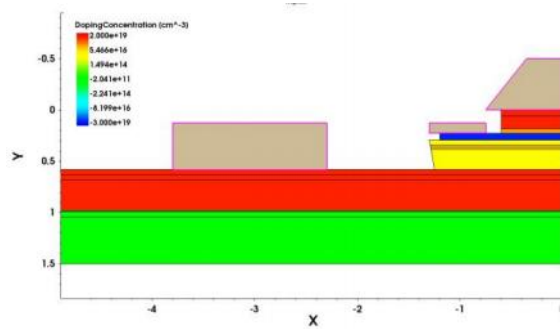


图 4 InP 基 DHBTs 器件 Sentaurus 仿真截面图

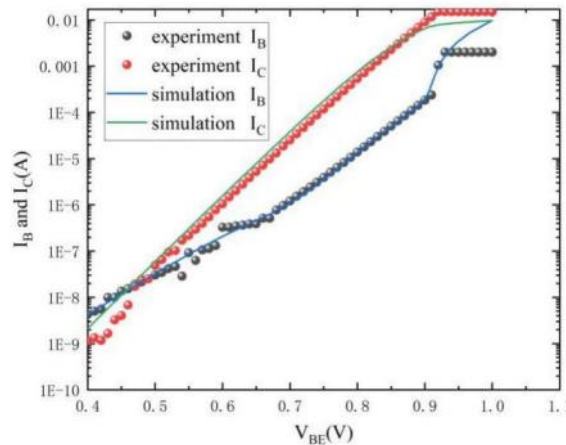


图 5 InP 基 DHBTs 器件 Sentaurus 仿真与测试正向 Gummel 特性曲线

图 5 为 InP 基 DHBTs 器件正向 Gummel 特性实验测试曲线和 Sentaurus 仿真曲线，测试数据和 Sentaurus 仿真数据在高低偏压范围内都能很好的拟合一致并且仿真结果中同样出现了和实验测试数据中的异质结势垒效应。因此可以使用 Sentaurus 仿真对器件异质结势垒效应进行仿真分析。

高能质子辐照不同辐照剂量在 InP 基 DHBTs 器件内产生不同浓度的缺陷，造成器件在不同质子辐照剂量下，产生不同的退化程度。通过 Sentaurus 仿真，模拟出基区和集电区不同缺陷浓度对 InP 基 DHBTs 器件异质结势垒效应的影响。缺陷性质如表 2 所示

表 2 Sentaurus 仿真中基区集电区加入缺陷性质

	缺陷性质	缺陷能级	缺陷浓度
Simulation 1	Donor	0.2 eV	无缺陷
Simulation 2	Donor	0.2 eV	$1 \times 10^{15} \text{cm}^{-3}$

Simulation 3	Donor	0.2 eV	$1.5 \times 10^{16} \text{cm}^{-3}$
Simulation 4	Donor	0.2 eV	$1.2 \times 10^{18} \text{cm}^{-3}$
Simulation 5	Donor	0.2 eV	$5 \times 10^{18} \text{cm}^{-3}$

图 6 给出了不同缺陷浓度下器件仿真的正向 Gummel 曲线图,可以看出当缺陷浓度达到 $5 \times 10^{18} \text{cm}^{-3}$ 时器件输出特性出现了较大退化,并且此缺陷浓度下异质结势垒效应增强,随着缺陷浓度升高,基区电流发生跃升时的偏压减小,并且电流发生跃升后扭结效应增强。当缺陷浓度低于 $1.5 \times 10^{16} \text{cm}^{-3}$ 时器件输出特性几乎没有变化,InP 基 DHBTs 器件 2MeV 质子辐照实验中当质子注入剂量低于 $5 \times 10^{13} \text{H}^+/\text{cm}^2$ 器件直流特性几乎不发生退化,仿真结果与实验结果表明 InP 基 DHBTs 器件具有较好的抗辐照性能。

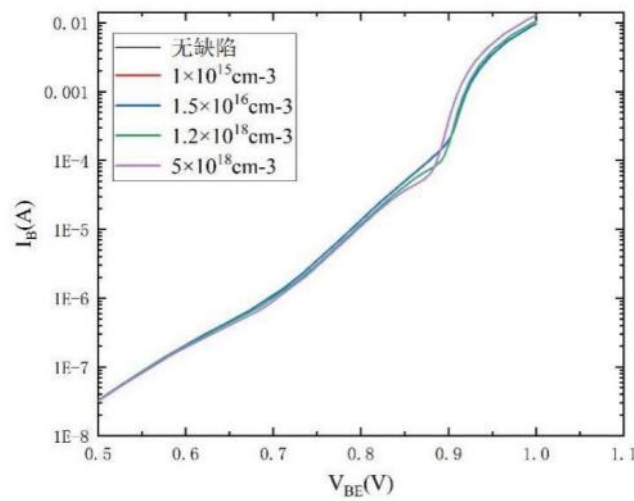


图 6 不同缺陷浓度的正向 Gummel 特性仿真曲线

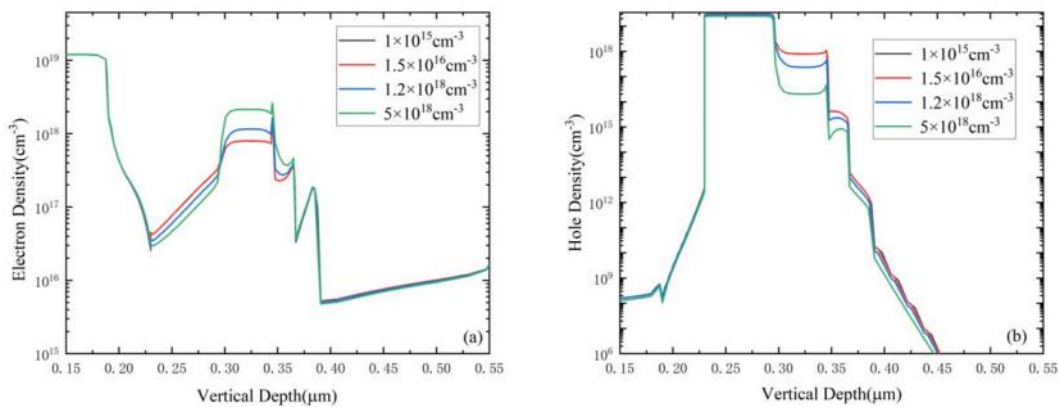


图 7 不同缺陷浓度下 InP 基 DHBTs 器件载流子浓度 (a) 电子浓度 (b) 空穴浓度

图 7 给出了不同缺陷浓度下 InP 基 DHBTs 器件 BC 结异质界面附近的电子与空穴浓度,基区和集电区的缺陷造成异质界面两侧电子浓度升高,空穴浓度下降,图 8 为不同缺陷浓度下 InP 基 DHBTs 器件空间电场图,缺陷的引入导致了集电区内异质界面附近的电场恶化,降低了载流子的运输速度,造成

电子浓度升高，并且缺陷的引入使空间电场的峰值向基区偏移，导致异质结势垒效应发生在更小的 V_{BE} 偏压下。

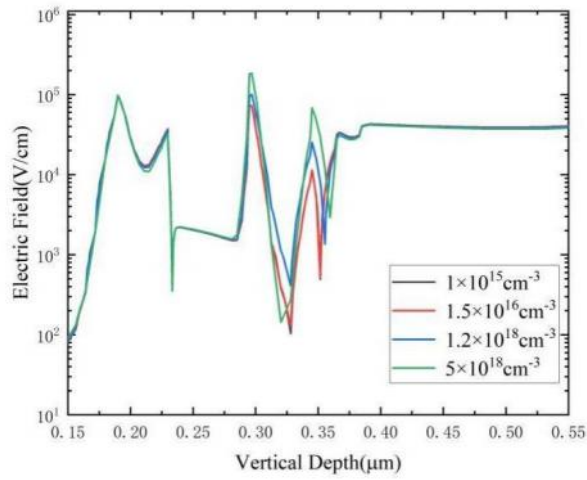


图 8 不同缺陷浓度下 InP 基 DHBTs 器件的空间电场分布

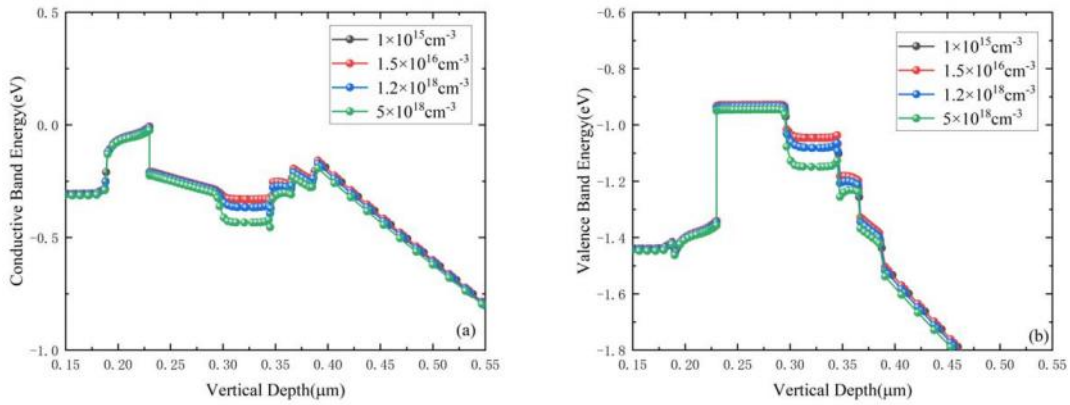


图 9 不同缺陷浓度下器件能带图 (a) 导带 (b) 价带

图 9 显示了 InP 基 DHBTs 器件不同浓度缺陷下的能带图，缺陷导致集电区内异质界面基区侧的导带和价带降低，使得异质界面的能带偏移增大，异质结的电子与空穴势垒升高，加剧了异质结势垒效应。

图 10 给出了器件不同缺陷浓度下的电流密度，基区引入的缺陷导致基区的电流密度增加，使得中性基区内的复合电流增加，造成基极电流在发生异质结势垒效应后剧烈增加，造成器件直流特性发生较大程度的退化

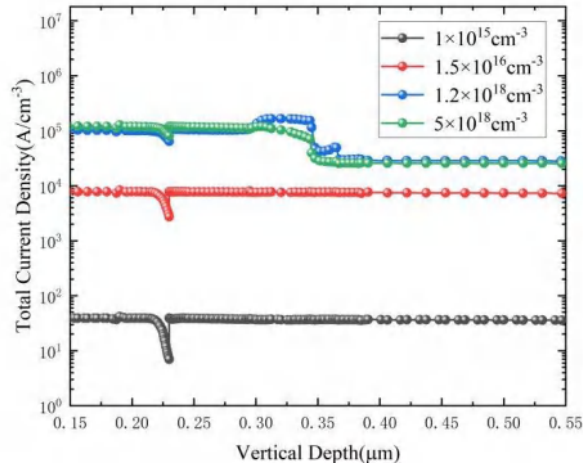


图 10 不同缺陷浓度下的电流密度

总的来说，质子辐照通过在器件有源区引入大量缺陷，中性基区和集电区内的缺陷造成异质界面的电子势垒升高、BC 结空间电荷区电场恶化，使得 InP 基 DHBTs 器件异质结势垒效应更加严重的发生，并且使异质结势垒效应发生在更小的偏压下，造成器件直流特性剧烈的退化。

4 结束语

本文系统研究了质子辐照对 InP/InGaAs HBTs 的损伤效应及机理。结果表明：低基极-发射极偏压下，辐照通过引入复合中心增强基区复合电流；高偏压时，异质结势垒效应（HBE）增强抑制集电极电流增长。随辐照剂量增加，电流增益衰减临界注入电流密度向低值偏移。机理上，辐照引入的缺陷使异质界面电子势垒升高、BC 结电场分布破坏，协同加剧 HBE 效应，导致器件直流特性劣化。本研究为优化异质结能带结构、提升器件抗辐照性能提供了理论依据，对空间及核环境电子器件设计具指导意义。

参考文献

- [1] Sengupta K, Nagatsuma T, Mittleman DM. Terahertz integrated electronic and hybrid electronic-photonic systems. *Nature Electronics*. 2018;1:622-35. doi:10.1038/s41928-018-0173-2
- [2] Urteaga M, Griffith Z, Seo M, Hacker J, Rodwell MJW. InP HBT Technologies for THz Integrated Circuits. *Proceedings of the IEEE*. 2017;105:1051-67. doi:10.1109/jproc.2017.2692178
- [3] Zhang J-J, Ding P, Jin Y-N, Meng S-H, Zhao X-Q, Hu Y-F, et al. A comparative study on radiation reliability of composite channel InP high electron mobility transistors*. *Chinese Physics B*. 2021;30. doi:10.1088/1674-1056/abe2fd
- [4] Yang B, Zhang JJ, Ding P, Sun SX, Jin YA, Zhao XQ, et al. PKA distributions in InAlAs and InGaAs materials irradiated by protons with different energies. *Nuclear Instruments & Methods in Physics Research Section B-Beam Interactions with Materials and Atoms*. 2020;484:42-7. doi:10.1016/j.nimb.2020.09.024
- [5] Kashio N, Kurishima K, Ida M, Matsuzaki H. Over 450-GHz $f(t)$ and $f(max)$ InP/InGaAs

- DHBTs With a Passivation Ledge Fabricated by Utilizing SiN/SiO₂ Sidewall Spacers. *Ieee Transactions on Electron Devices*. 2014;61:3423-8. doi:10.1109/ted.2014.2349872
- [6] Ida M, Kurishima K, Watanabe N. Over 300 GHz $f(T)$ and $f(max)$ InP/InGaAs double heterojunction bipolar transistors with a thin pseudomorphic base. *Ieee Electron Device Letters*. 2002;23:694-6. doi:10.1109/led.2002.806300
- [7] Kashio N, Kurishima K, Fukai YK, Ida M, Yamahata S. High-Speed and High-Reliability InP-Based HBTs With a Novel Emitter. *Ieee Transactions on Electron Devices*. 2010;57:373-9. doi:10.1109/ted.2009.2037461
- [8] Hoeffgen SK, Metzger S, Steffens M. Investigating the Effects of Cosmic Rays on Space Electronics. *Frontiers in Physics*. 2020;8.doi:10.3389/fphy.2020.00318
- [9] Tripathi RK. Radiation Effects In Space. 21st International Conference on Application of Accelerators in Research and Industry (CAARI). Ft Worth, TX2010. p. 649-54.doi:10.1063/1.3586182
- [10] Youssouf AS, Habaebi MH, Hasbullah NF. The Radiation Effect on Low Noise Amplifier Implemented in the Space-Aerial-Terrestrial Integrated 5G Networks. *IEEE Access*. 2021;9:46641-51.doi:10.1109/access.2021.3065497
- [11] Zhang J-L, Ding P, Mei B, Meng S-H, Zhang C, Ma L-H, et al. The effects and mechanisms of 2 MeV proton irradiation on InP-based high electron mobility transistors. *Applied Physics Letters*. 2022;120.doi:10.1063/5.0084257
- [12] Zhong Y-H, Yang B, Chang M-M, Ding P, Ma L-H, Li M-K, et al. Enhancement of radiation hardness of InP-based HEMT with double Si-doped plane*. *Chinese Physics B*. 2020;29.doi:10.1088/1674-1056/ab6962
- [13] Sun SX, Chang MM, Li MK, Ma LH, Zhong YH, Li YX, et al. Effect of defects properties on InP-based high electron mobility transistors. *Chinese Physics B*. 2019;28. doi:10.1088/1674-1056/28/7/078501

Research Article

Optimization Techniques for Structure-Material-Process in Aerospace 2.5D/3D Integrated Microsystems: A Finite Element Analysis-Based Approach

LI Zhihao¹, WANG Zhihui¹, GUO Hengtong², LIU Yiguang¹, FU Wanyue¹, LIU Pei¹

¹ China Academy of Aerospace Standardization and Product Assurance, Beijing 100071, China

² Beijing Microelectronic Technology institute, Beijing 100076, China

Corresponding author: LIU Pei; Email:57269529@qq.com .

Manuscript Received June XX, 2025; Accepted June XX, 2025; Published Online June XX, 2025.

Abstract — 2.5D/3D heterogeneous integration packaging stands as the dominant core technology for enabling large-scale integration, high-density interconnects, and flexible reconfiguration in next-generation aerospace microsystems. This approach incorporates complex structures such as through-silicon vias (TSVs), μ bumps (μ Bumps), large-sized chips, and multilayer substrates, posing significant challenges for optimization of integration architecture and fabrication processes. Addressing critical issues—including cross-scale modeling of heterogeneous integration structures and simulation of complex nonlinear manufacturing processes—this study proposes a process-structure co-optimization methodology based on finite element analysis (FEA) for a representative aerospace microsystems. We conducted four key implementations: multi-physics constitutive modeling of micro-interconnects, FEA-driven TSV design optimization in silicon interposers, structural simulation of multilayer substrates, and dynamic simulation of flip-chip reflow processes. Our findings yield optimized structural and process enhancements in microsystem longevity, thermomechanical responses, and pressure-assisted reflow performance, ultimately providing a high-precision, high-efficiency solution for advanced aerospace packaging design.

Keywords — Constitutive modeling; TSV design; Reflow simulation; Warpage control.

I. Introduction

2.5D/3D packaging technology stands as a pivotal pillar driving the rapid advancement in the field of integrated circuits. It represents a new paradigm for high-density chip integration by bridging the internal chip system with the external system^[1-3]. With the continuous escalation in the complexity of space missions, aerospace equipment urgently demands miniaturized, intelligence-enhanced, and integrated electronic systems. Consequently, the functional density of spacecraft has experienced exponential growth. Under these circumstances, 2.5D/3D heterogeneous integration technology has emerged as the mainstream approach for manufacturing avionics microsystem products^[4-5]. These stacked structures achieve multi-chip heterogeneous integration within a three-dimensional space by leveraging vertical interconnects via through-silicon vias (TSVs) in silicon interposers and layer-to-layer interconnects using microbumps. The technology is characterized by several key features: the implementation of interposers with high-density TSV arrays to provide vertical interconnection channels in the form of blind vias or through vias^[6]; the realization of fine planar interconnects with line widths/spacing ranging from several micrometers (μm) to tens of micrometers (μm) through redistribution layers (RDLs)^[7]; and the implementation of layer-to-layer vertical interconnects via microbumps,

which exhibit a solder joint diameter reduced by an order of magnitude compared to traditional solder ball interconnects^[8].

However, the extreme space environment imposes significantly more stringent reliability requirements on 2.5D/3D packaging. Optimizing structural and material design, as well as process design, through digital simulation methods thus demonstrates significant potential for application in aerospace equipment^[9-11].

II. Primary Reliability Challenges Aims and Scope

Aerospace 2.5D/3D microsystem integrated packaging structures are typically composed of through-silicon vias (TSVs), silicon interposers, microbumps and solder ball arrays, chiplets, redistribution layers (RDLs), package substrates, and molding compounds. Their characteristic failure modes include interfacial delamination, solder joint fatigue cracking, microvoids within the packaging materials, and fatigue cracks^[12-15], as illustrated in Figure 1.

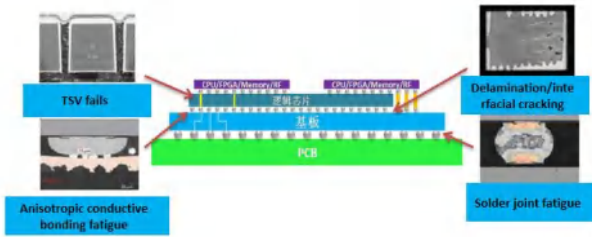


Figure 1 Typical Failure Mechanisms in Aerospace 2.5D/3D Microsystem Integration Structures

(1) TSV-Cu Protrusion Induced by Thermomechanical Stress: Under various process temperatures (e.g., annealing temperatures of 400-425 ° C) and operating temperatures (-25 to 125 ° C), significant thermomechanical stresses develop within both the TSV copper (TSV-Cu) and the surrounding silicon^[16]. This TSV-Cu protrusion adversely impacts device reliability. On one hand, the protruding TSV-Cu can cause cracking in the surrounding dielectric materials; on the other hand, it can pull on structures at both ends of the TSV (such as the BEOL layers), leading to fracture of the metal interconnect layers, as shown in Figure 2.



a) Dielectric Material Cracking b) Lead Layer Failure
Figure 2 Reliability Challenges Induced by Electroplated Cu Extrusion

(2) Solder Joint Thermal Failure: Under thermal cycling conditions, the creep rate of solder interconnect materials increases dramatically with rising temperature. Lead-free solders exhibit higher creep rates than traditional 63Sn37Pb solder. From the perspective of microstructural evolution, the coarsening of second-phase particles significantly reduces the capacity for dislocation motion and grain boundary sliding, consequently degrading the material's creep resistance and mechanical strength^[17-19].

Failure analysis reveals that crack initiation and propagation in solder joints occur predominantly within the bulk solder material. As the number of thermal cycles increases, crack lengths in SnPb solder joints become longer than those in lead-free joints. Plastic strain and creep strain dominate solder joint failure, with higher inelastic strain correlating directly with poorer resistance to thermal cycling, as depicted in Figure 3.

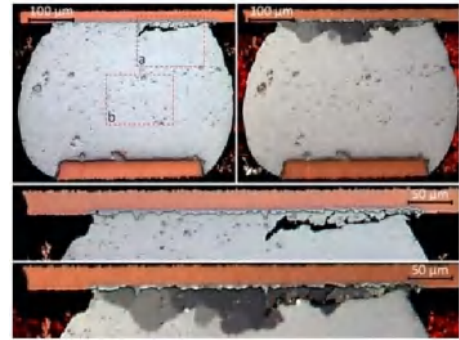


Figure 3 Solder Joint Failure Modes Under Thermal Cycling Conditions

(3) TSV Electromigration Failure: Two locations in the TSV interconnect structure exhibit high susceptibility to electromigration (EM) due to diffusion inhomogeneity: the transition between the TSV and the first metal layer of the back-end-of-line (BEOL), and the transition between the TSV and the redistribution layer (RDL). These structural inhomogeneities cause local current crowding^[20-21].

The resultant electromigration induces non-uniform directional atomic migration, leading to the formation of microvoids and ultimately open-circuit failure. The differing diameters at the top and bottom of the TSV cause void formation at distinct locations determined by current direction, as shown in Figure 4. When the electron flow is from the top TSV towards the bottom, voids form near the bottom Cu/SiN interface; conversely, reversing the current direction leads to void formation near the top interface within the metal layer. Furthermore, EM-induced failure phenomena in metal layers during TSV interconnection have also been observed in other studies^[22-23].

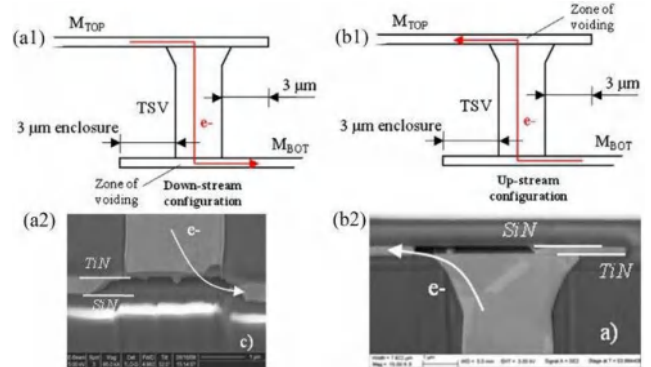


Figure 4 Electromigration Failure Mechanism Caused by Microvoids Near Cu/SiN Interface in TSV Interconnect

III. Cross-Scale Multi-Component Modeling and Simulation Methodology

1. Material Nonlinear Modeling of Key Structures

Within aerospace 2.5D/3D microsystem integrated structures, flip-chip interconnect bumps and copper interconnect structures are critically linked to system-level reliability. These micro/nano-scale interconnects exhibit distinct material nonlinear behaviors under external loading^[24].

The unified viscoplastic Anand constitutive model is adopted for flip-chip interconnect bumps to describe their

mechanical behavior. This model accounts for both time-dependent high-temperature creep and time-independent low-temperature plastic strain occurring simultaneously during deformation. The Anand constitutive equations are given as follows:

$$\dot{\epsilon}_p = A \exp\left(-\frac{Q}{RT}\right) \left[\sinh\left(\xi \frac{\sigma}{s}\right) \right]^{1/m} \quad (1)$$

where $\dot{\epsilon}_p$ is the strain rate, A is a constant, Q is the activation energy, m is the strain rate sensitivity exponent, ξ is the stress factor, T is the absolute temperature, s is the internal state variable, and R is the gas constant.

For copper interconnect structures, copper's melting point (1083°C) is significantly higher than solder, causing it to exhibit predominantly elastic-plastic characteristics under thermal cycling loads. A bilinear isotropic hardening model is applied:

$$E_T = \frac{\sigma_{lim} - \sigma_y}{\epsilon_{lim} - \zeta_y} \quad (2)$$

where E_T is the tangent modulus, σ_{lim} is the material ultimate strength stress, σ_y is the material yield stress, ϵ_{lim} is the material ultimate total strain, and ζ_y is the material total strain at yield.

2. Cross-Scale Submodeling of Package Interconnect Structures

Aerospace 2.5D/3D microsystem integration involves multiple chiplets connected to an organic interposer substrate via C4 solder joints, necessitating modeling approaches addressing significant scale disparities^[25]. TSV structures, serving as critical vertical interconnects, require detailed finite element models (Submodeling technique^[26-27]). The submodel is positioned to correspond precisely with the relevant region of the global (coarse) model. The calculated displacement values from the global model's cut boundaries are then applied as boundary conditions to the submodel. The degrees of freedom (DOF) values from the corresponding elements in the global model are determined, and these values are interpolated onto the submodel's cut boundary using element shape functions. The boundary conditions and detailed model of the TSV structure are illustrated in Figures 5 and 6.

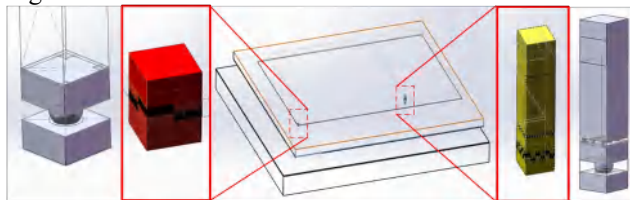


Figure 5 Submodel Boundary Conditions Extracted from Global Model

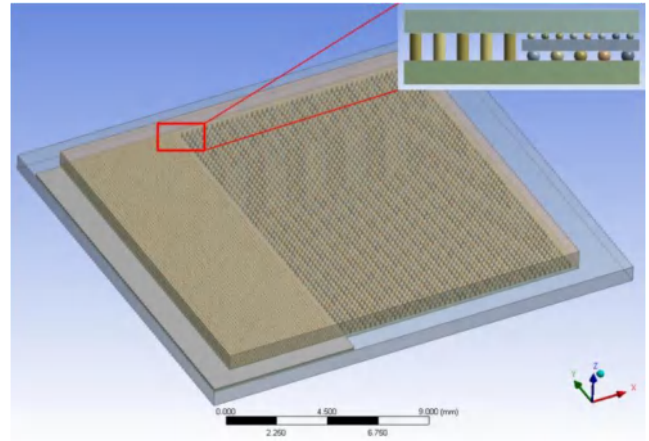


Figure 6 Schematic of TSV Structure Simulation Model

3. Cross-Scale Modeling of Multi-Layer Substrate Wiring Structures

The multi-layer copper wiring structure within high-density chiplet-integrated molded substrates directly influences physical performance. Neglecting this wiring structure in thermo-mechanical simulation analyses compromises result accuracy. However, the line width/spacing ($\approx 1 \mu\text{m}$) on chiplet-integrated organic substrates differs by four orders of magnitude from the substrate's overall dimensions ($\approx 50 \text{ mm}$). Explicit geometric modeling and meshing of the wiring structure would lead to prohibitively large mesh counts and computational burden. To maintain simulation accuracy while controlling computational cost, the physical properties of the wiring structure are equivalently mapped, bypassing explicit geometric modeling and meshing. Software is used to calculate the equivalent physical parameters of the substrate wiring layers, which are then incorporated into the simulation model, as shown in Figure 7.

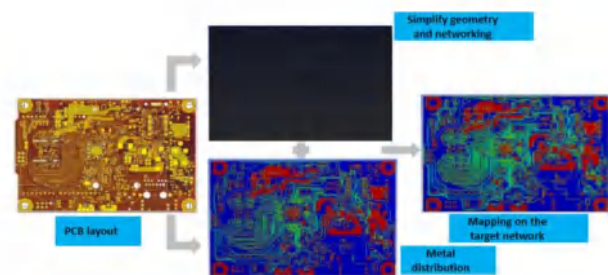


Figure 7 Equivalent Simulation Model Incorporating Multilayer Interconnect Materials

4. Equivalent Modeling of the Global Structure

In finite element analysis of the entire aerospace microsystem, structural details corresponding to second and third strain invariants are sacrificed for simulation efficiency and feasibility^[28]. To ensure simulation accuracy, however, detailed submodels are established for critical components including flip-chip interconnects, TSV interposers, internally stacked die, internal wire bonds, and solder joints. The equivalent model of the global structure is depicted in Figure 8.

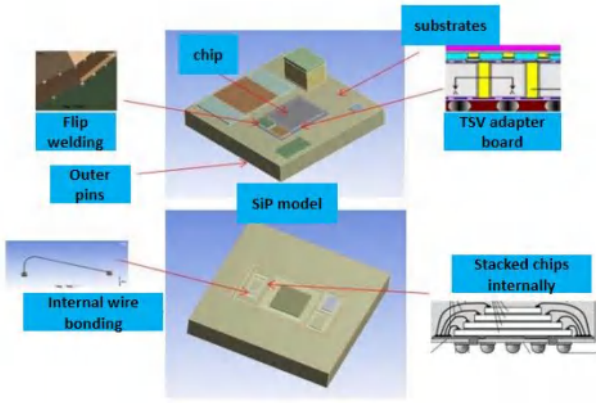


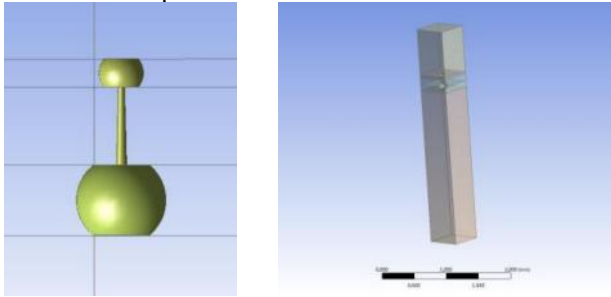
Figure 8 Equivalent Modeling of Aerospace 2.5D/3D Integrated Microsystem Architectures

IV. Simulation - Based Optimization Design of Structure and Materials for TSVs

As a novel interconnect process structure for 2.5D/3D microsystems, through-silicon vias (TSVs) exhibit diverse structural forms and material combinations, yet their reliability remains incompletely characterized^[29]. Simulation-based approaches enable the reliability optimization of TSV structures, identifying the optimal structural configuration and material combination to guide subsequent microsystem integration design.

1. Structural Optimization Design

Building upon existing manufacturing processes, typical TSV via shapes include cylindrical and frustum (tapered) configurations, which influence their lifetime. The cylindrical and frustum TSV structures are depicted in Figure 9. Four models (MODEL I - IV) with distinct via shapes (frustum or cylindrical), TSV heights, and aspect ratios, were designed using an orthogonal methodology; their structural parameters are summarized in Table 1.



(a) MODEL I-Frustum of a cone (b) MODEL I

I-MODEL IV-Cylindrical

Figure 9 2.5D Device Architectures Implemented in MODEL I-MODEL IV

Table 1 Structural Configuration of MODEL I-MODEL IV Simulation Models

Finite element model	TSV height	Through-hole shape	TSV diameter	AR(Aspect ratio)
MODEL I	50 μm	Frustum of a cone	10 μm at the top and 14 μm at the bottom	5:1
MODEL II	50 μm	Cylindrical	10 μm	5:1
MODEL III	100 μm	Cylindrical	10 μm	10:1

MODEL IV	150 μm	Cylindrical	10 μm	15:1
----------	-------------------	-------------	------------------	------

Finite element analysis results for maximum plastic strain (Figure 10) indicate values of $4.3712\text{e-}2$ m/m, $2.0251\text{e-}2$ m/m, $1.8658\text{e-}2$ m/m, and $1.8955\text{e-}2$ m/m for MODEL I to MODEL IV, respectively. For frustum vias, the maximum plastic strain occurs at the bumps on the lower side of the interposer, while for cylindrical vias, it appears at the μ bumps on the upper side. The maximum plastic strain decreases monotonically with increasing aspect ratio.

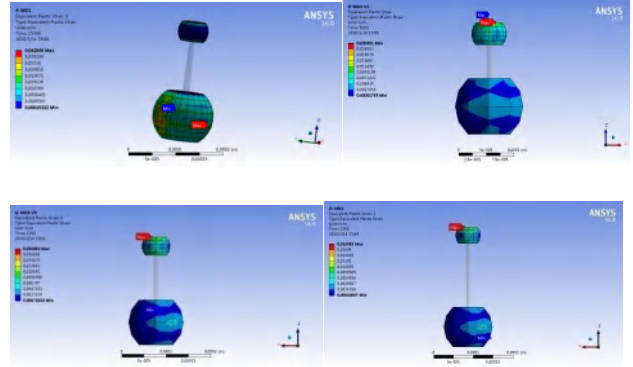


Figure 10 Extreme Plastic Strain Distribution in MODEL I-MODEL IV

The variation trend of maximum plastic strain (Figure 11) reveals that the bump in MODEL I experiences the highest peak plastic strain, which remains persistently elevated. The order of magnitude of the maximum plastic strain across MODEL I to MODEL IV is consistent, indicating relatively stable plastic strain evolution.

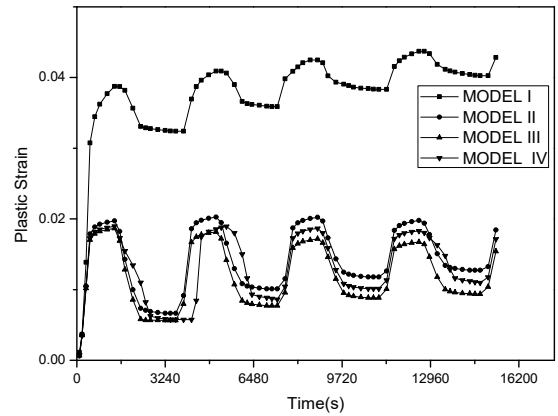


Figure 11 Evolution of Maximum Plastic Strain in MODEL I-MODEL IV Simulation Models

Applying the Engelmaier fatigue life prediction model (Equation 1), the predicted fatigue lives for MODEL I to MODEL IV are listed in Table 2.

$$N_f = \frac{1}{2} \left(\frac{\Delta\gamma}{2\epsilon_f} \right)^{1/c} \quad (3)$$

The predictions demonstrate that MODEL III, featuring a cylindrical TSV structure with a via diameter of 10 μm and an aspect ratio of 10:1, exhibits the optimal

thermo-mechanical performance under thermal cycling loads, achieving the highest thermal fatigue life of 10,234 cycles. This structure was consequently selected for subsequent TSV material simulation studies.

Table 2 Shear Strain Range and Fatigue Life Prediction in MODEL I-MODEL IV

Finite element model	Plastic strain maximum	Plastic strain shear range	Fatigue life N_f
MODEL I	4.3712e-002 m/m	9.85E-03	5802
MODEL II	2.0251e-002 m/m	8.00E-03	9911
MODEL III	1.8658e-002 m/m	7.90E-03	10234
MODEL IV	1.8955e-002 m/m	8.17E-03	9389

2. Material Optimization Design

TSVs comprise various materials for the via itself, the insulation layer, and the seed layer. Simulation enables the analysis of reliability for different material combinations. The cylindrical TSV structure identified in Section 4.1 (10 μm diameter, 10:1 aspect ratio) served as the initial model for TSV material optimization. Four models (MODEL V - MODEL VIII) were designed using an orthogonal distribution; their structural and material information is provided in the table below.

Table 3 Geometric Parameters of MODEL V-MODEL VIII Simulation Models

Table 3 Geometric Parameters of MODEL V-MODEL VIII Simulation Models

Finite element model		insulation	Seed layer	Through-hole material
size		5 μm	2 μm	The diameter is 10 μm, and the AR ratio is 10:1
TSV structural model	MODEL V	SiO ₂	Cu	Cu
	MODEL VI	SiO ₂	TiN	Cu
	MODEL VII	Si ₃ N ₄	Cu	Cu
	MODEL VIII	SiO ₂	Sn	W

Structural mechanics analyses were performed under thermal loading conditions specified by GJB548B-2005 Method 1010.1, Temperature Cycling (Condition C: -65° C to 150° C). The distribution of maximum equivalent stress on the seed layer (Figure 12) indicates that peak stresses are consistently concentrated at the interface between the seed layer material and the TSV via material. The maximum equivalent stresses for MODEL V to MODEL VIII are 1.417 GPa, 6.4172 GPa, 5.1813 GPa, and 4.2082 GPa, respectively.

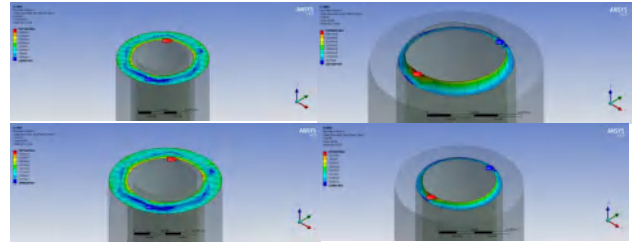


Figure 12 Maximum Equivalent Stress Distribution in Seed Layers of MODEL V-MODEL VIII

Comparison of these simulation results demonstrates that MODEL V, utilizing an SiO₂ insulation layer combined with a Cu seed layer and Cu via material, exhibits the most favorable thermo-mechanical behavior under thermal cycling loads.

V. Simulation-Based Process Optimization Design

Multi-physics coupled simulation serves as a critical tool for process optimization design^[30]. By establishing a simplified global structural model, it enables quantitative analysis of the thermo-mechanical-electrical coupling effects during the manufacturing process, thereby guiding the optimization of microsystem process parameters. The flip-chip bonding process involves contact and collapse between the chip bumps and the substrate pads. If the gap between the chip and the pad exceeds the bump height, effective contact wetting cannot occur, leading to solder joint non-wetting (opens)^[31-32]. Consequently, controlling the gap between bumps and pads is paramount for forming effective interconnects in molded assemblies during flip-chip bonding^[33-34].

As illustrated in Figure 13, defining the chip bump height as A and the solder joint height as B, the critical gap for forming a bump interconnect is defined as Gap_{critical} = A - B. The maximum gap between the molded assembly and the pad is defined as C, and the warpage of the molded assembly is defined as D, where C = B + D. The prerequisite for forming an effective interconnect is that the molded assembly warpage remains within the critical gap, i.e., D ≤ Gap_{critical} = A - B. Therefore, reducing the molded assembly warpage D enables effective solder joint interconnection.

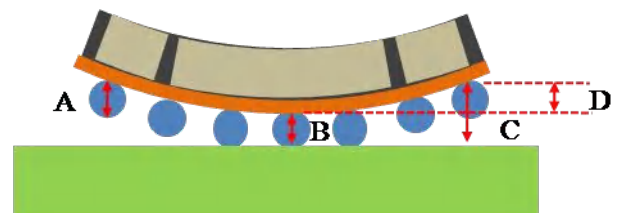
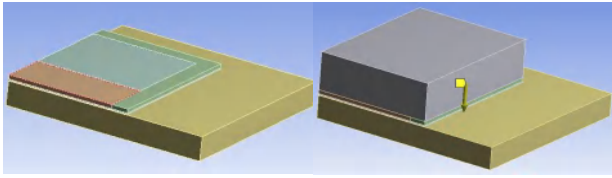


Figure 13 Flip-Chip Interconnect Characterization

Molded assembly warpage primarily occurs within the filled encapsulant regions^[35-36]. Finite element analysis can be employed to compare warpage behavior during pressureless reflow versus Pressure-assisted reflow processes; a model of the molded assembly reflow process is shown in Figure 14.



(a) Pressure-free model (b) Pressurized model

Figure 14 Reflow Process Model for Molded Module Assembly

Structural mechanics analyses were conducted under thermal loading conditions at a peak temperature of 265° C with a duration of 90-120 seconds. Simulation results (Figures 15 and 16) reveal that during pressureless reflow, the high-temperature warpage of the assembly reaches 42.5 μm. In contrast, employing a Pressure-assisted reflow method effectively constrains deformation at the assembly corners during high-temperature reflow through the applied weight of the pressure block, resulting in a high-temperature warpage value of only 12.3 μm. This thereby reduces deformation warpage by approximately 70%, ensuring effective interconnects for large-sized flip-chip assemblies.

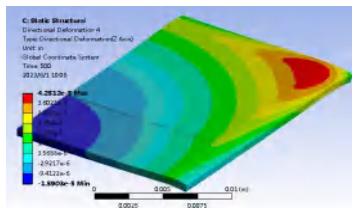


Figure 15 Simulation of Flip-Chip Reflow Process in Molded Module

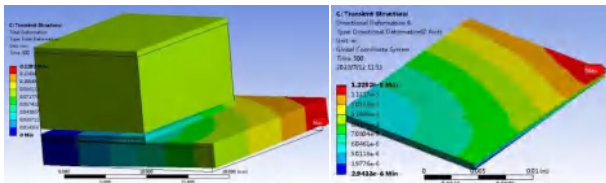


Figure 16 Counterweight-Pressurized Flip-Chip Reflow Process Simulation for Molded Module

Therefore, implementing counterweight pressure during reflow provides an effective approach for flip-chip bonding warped assemblies. Applying a pressure block atop the chip allows control over the relative warpage behavior between the multi-chip molded assembly and the package substrate.

VI. Conclusion

This study first investigated the characteristic failure modes — including TSV-Cu protrusion induced by thermomechanical stress, solder joint thermal failure, and TSV electromigration — observed in selected representative aerospace 2.5D/3D microsystem products. A submodeling approach was employed to investigate cross-scale modeling methodologies for TSV via structures. Furthermore, based on the Anand constitutive model and the bilinear isotropic hardening method,

material nonlinear modeling techniques were developed for flip-chip interconnect structures and copper interconnects. Detailed local models were established for critical structural components. To ensure computational accuracy while enhancing efficiency, a "global-to-local finite element mapping" methodology was implemented to simplify non-critical regions of the microsystem integrated structure, thereby constructing an equivalent model of the global assembly. Focusing on the TSV structure, structural mechanics analyses under thermal loading conditions were conducted to investigate the influence of via shape and aspect ratio on TSV fatigue life, leading to the identification of the optimal TSV geometry. Subsequently, using the TSV structure with the best structural parameters identified above, the maximum equivalent stress behavior for different combinations of interposer, insulation layer, seed layer, and via materials was studied, resulting in the determination of the optimal material combination. Finally, taking the molded assembly as the research object, simulations successfully replicated the warpage phenomenon occurring during the reflow process. A Pressure-assisted reflow process scheme was designed and simulated, demonstrating a significant 70% reduction in warpage during molded assembly reflow.

Acknowledgement

I sincerely thank all the leaders and colleagues who provided selfless help and support during the drafting of this paper. Without your careful guidance, valuable suggestions, and generous assistance, I could not have completed this work.

References

- [1] Yang J, Zhang J Z, Zhu J, et al. Development and typical applications of RF integrated microsystem technology[J]. Aerospace Shanghai, 2024, 41(4): 102-112.
- [2] Ma Q L, Ma Y H, Zhong C, et al. Research progress on key technologies of 2.5D packaging[J]. Electronics & Packaging, 2025, 25(5): 050107.
- [3] Cao L Q, Hou F Z, Wang Q D, et al. Development and opportunities in advanced packaging technology[J]. Science and Technology Foresight, 2022, 1(3): 101-114.
- [4] WANG Z Y, SUN J B, GOKSOY A, et al. Exploiting 2.5D/3D heterogeneous integration for AI computing[C]//Proceedings of the 29th Asia and South Pacific Design Automation Conference. New York: IEEE, 2024: 758-764.
- [5] MANDALAPU C S, BUCH C, et al. 3.5D advanced packaging enabling heterogeneous integration of HPC and AI accelerators[C]//Proceedings of the IEEE 74th Electronic Components and Technology Conference. New York: IEEE, 2024: 798-802.
- [6] Gao H, Zhang W Q, Gu Y, et al. Research on hole-forming process of high-aspect-ratio TSV interposer[J]. Electro-Mechanical Engineering, 2025, 41(2): 41 - 45, 51.
- [7] Zhu Z, Huang C H, Li L S, et al. Study on thin-film redistribution layer process and assembly reliability based on ceramic substrate[J]. Electronics & Packaging, 2023, 23(10): 100208.
- [8] Huang M L, Wang S B, You H C, et al. Electromigration in Sn-based microbumps for IC chips: From physical mechanisms to reliability enhancement[J]. Acta Metallurgica Sinica, 2025, 24(3).
- [9] Wang Y. Status and development of high-density high-performance electronic packaging technology[J]. Microelectronics Technology, 1998(S1): 1-22.
- [10] Hao R, Wu J, Wu K, et al. Influence of reflective layer on the performance of flip-chip LED[J]. Electronics Process Technology, 2019(3): 125-129.
- [11] Xiang Y Y, Zhou Z, Zhang J, et al. Research on thermal stress field of self-propagating interconnection for large-area chip/DBC

substrate[J]. *Electronics Process Technology*, 2019(4): 187-191.

- [12] PU H P, KUO H J, LIU C S, et al. A novel submicron polymer re-distribution layer technology for advanced InFO packaging[C]. 2018 IEEE 68th Electronic Components and Technology Conference(ECTC), San Diego, CA, USA, 2018: 45-51.
- [13] ZHOU J, CAO X, ZHANG N, et al. Micro-channel heat sink: a review[J]. *Journal of Thermal Science*, 2020, 29(6): 1431-1462.
- [14] CHEN S, EN Y F, LI G Y, et al. An ion beam layer removal method of determining the residual stress in the as-fabricated TSV-Cu/TiW/SiO₂/Si interface on a nanoscale[J]. *Microelectronics Reliability*, 2020(112): 113826.
- [15] XIAO H, WANG F, WANG Y, et al. Effect of ultrasound on copper filling of high aspect ratio through-silicon via (TSV)[J]. *Journal of The Electrochemical Society*, 2017, 164(4): D126-D129.
- [16] Chen S. Study on protrusion behavior of TSV-Cu key structure in 3D electronic packaging[D]. Beijing: Beijing University of Technology, 2017.
- [17] Ning Y X, Pan K L, Li N. Failure analysis of solder joints in electronic assembly[J]. *Electronic Industrial Equipment*, 2007(9).
- [18] Ren N, Tian Y, Long D F. Effect of temperature on reliability of PBGA solder joints under thermal cycling[J]. *Welding Technology*, 2016, 45(4): 8-11.
- [19] Wang K F. Creep fracture failure and prevention of welded structures[J]. *Petroleum Engineering Construction*, 1999, 25(4): 1-5.
- [20] ZHANG P, XUE S B, WANG J H. New challenges of miniaturization of electronic devices: Electromigration and thermomigration in lead-free solder joints[J]. *Materials & Design*, 2020, 192: 108726.
- [21] TAE-KYU L, HANRY Y, INDRANATH D. Damage mechanisms in through-silicon vias due to thermal exposure and electromigration[J]. *Journal of Electronic Materials*, 2024, 53(3): 1214-1222.
- [22] DONG C C, HUANG P ZH. Phase-field simulation of conductive inclusion evolution in highly symmetric oriented single crystal metal interconnects under anisotropic interface diffusion induced by electromigration[J]. *Journal of Mechanics of Materials and Structures*, 2024, 19(3): 373-395.
- [23] SHAN Y P, HAO H J, YIN Y ZH, et al. Effects of temperature and dc electric fields on perfluorooctanoic acid sorption kinetics to activated carbon[J]. *Environmental Science Technology*, 2024, 16(4): 123-147.
- [24] GAO X G, LI L, SONG Y D. A temperature-dependent constitutive model for fiber-reinforced ceramic matrix composites and structural stress analysis [J]. *International Journal of Damage Mechanics*, 2014, 23(4): 507 - 522.
- [25] Liu C, Ren P, Ding C Q, et al. Substructure-based cross-scale modeling method for BGA in SiP packaging[J]. *Electro-Mechanical Engineering*, 2024, 40(6): 60 - 64.
- [26] Wang X C. Finite element method[M]. Beijing: Tsinghua University Press, 2003: 186 - 190.
- [27] Han Y, Dai C L, Li K. Research on consistent multi-scale structural behavior modeling based on substructure method[J]. *Journal of Transportation Science and Engineering*, 2022, 38(3): 48 - 56.
- [28] S. Zhao, Y. Xu, C. Pan, et al. Microstructural Modeling and Strengthening Mechanism of TiB/Ti-6Al-4V Discontinuously Reinforced Titanium Matrix Composite[J]. *Materials*, 2019, 12.
- [29] Dong Z X, Cang D Q, Zhao J C, et al. Research progress on thermo-mechanical reliability of through-silicon via interconnect structures[J/OL]. *Semiconductor Technology*.
- [30] Haodong Yu, Pengsen Yang, Hongwang Yao, et al. Multi-physics coupling-based numerical simulation and optimization of heat pipe cooling systems for sealed switchgear[J]. *International Journal of Thermal Sciences*. 2025, 5.
- [31] Lyu X R, Lin P R, Liu J S, et al. Research on defects and detection methods in flip-chip packaging process[J]. *Special Topic on Integrated Circuit Technology*, 2021(07).
- [32] Zhu G L, Han X, Xu X M, et al. Failure analysis and improvement of flip-chip on organic substrate for integrated circuits[J]. *Electronics & Packaging*, 2024, 24(2): 020111.
- [33] Chen Z H, Zhang C, Xu P, et al. A failure mechanism and optimization method for micro-vias in flip-chip package[J]. *Microelectronics*, 2024, 54(1).
- [34] Zhu G L, Ji Z K, Ji P, et al. Analysis and improvement of flip-chip solder joint cracking in plastic package[J]. *Electronics & Packaging*, 2023, 23(2).
- [35] Gao N Y, Chen X X, Tong L Y, et al. Failure analysis of warpage in flip-chip plastic package[J]. *Electronic Product Reliability and Environmental Testing*, 2020, 38(2).
- [36] Zhou X F, Xu Z G, Zhang Z Y, et al. Study on reflow behavior of

flip-chip on high-density organic substrate[J]. *Electronic Product Reliability and Environmental Testing*, 2022, 40(2).



LI Zhi Hao received the master degree in mechanical engineering technology from Beijing Jiaotong University, Beijing, China, in 2024. His main research interest is component assurance. (Email: 1067577718@qq.com)



LIU Pei is currently a Researcher at the Electronic Components Research Institute of the China Academy of Aerospace Standardization and Product Assurance. He has long been engaged in strategic development research of complex electronic components, including microwave devices, optoelectronic and microelectronic chips, and microsystems. His work focuses on aerospace component standardization, testing, product evaluation, and quality assurance technologies. He has presided over and participated in multiple national-level research projects. His scientific achievements have been honored with several awards, including the First Prize of Shaanxi Provincial Science and Technology Award, the Second Prize of National Defense Science and Technology Award, and the First/Second Prize of Aerospace Science and Technology Progress Award. (Email: 57269529@qq.com)

基于宏模型的热电制冷封装系统仿真设计

李龙飞, 唐 旻

(上海交通大学 信息与电子工程学院, 上海 200240)

摘要: 针对电子封装系统热管理中热电制冷器的仿真与设计, 提出一种基于宏模型的高效方法。首先, 整个电子封装系统划分为两个区域: 热电器件区域 (TDR) 和非热电区域 (NTR)。在 TDR 中, 提出一种新颖的一维等效热阻模型 (ETRM) 来表征热电器件; 在 NTR 中, 采用宏建模技术将 NTR 对 TDR 的影响转化为在接口处的多端口热阻网络 (MTRN)。通过上述方式, TDR 与 NTR 仅在端口处通过温度连续性和热流连续性条件进行耦合, 从而快速求解所有端口的热流密度。最后, 通过端口热流重构芯片封装的温度分布。数值算例验证了本文方法的准确性, 基于宏模型技术, 计算成本得到了显著降低, 与商业软件相比计算效率提升了约 27 倍。

关键词: 区域分解; 等效热阻模型; 多端口热阻网络; 热电制冷器

1 引言

由于具有重量轻、可靠性高和噪声低等优点, 热电制冷器 (TEC) 已成为集成电路及其封装中热管理最重要的冷却技术之一[1]。在实际应用中, 热点区域通常表现出极高的局部热流密度, 这对集成电路的热管理提出了严峻挑战[2]。根据帕尔贴效应, TEC 在受限区域内提供直接且高效的局部冷却方面具有显著优势, 能够很好地应对上述问题。然而, TEC 系统的高效仿真与优化并非易事。

在基于 TEC 制冷的集成电路及封装系统的仿真中, 通常面临一个涉及多物理场耦合的三维非线性问题。为了评估 TEC 器件的性能, 人们已提出了多种数值方法用于求解电热耦合方程。例如, 文献[3]采用三维有限元法 (FEM) 对分段环形热电制冷器进行了数值仿真; 文献[4]通过有限差分法 (FDM) 分析了包含多个热电元件在瞬态与稳态下的行为表现; 文献[5]则利用有限体积法 (FVM) 评估了使用 TEC 进行光伏组件主动热管理的可行性。然而, 由于热电效应本质上是一个需要复杂迭代求解的非线性问题, 这些数值方法通常伴随着较大的计算开销, 难以满足 TEC 器件实时优化的需求。为解决该问题, 文献[6]提出了一种简单的一维能量平衡模型来等效表示热电器件。尽管该模型结构简单, 但在存在较大温度梯度的情况下可能会产生明显的误差。文献[7]提出了一种基于变量分

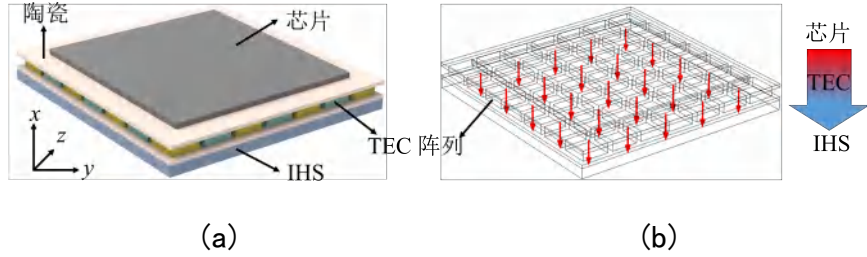


图1 (a)集成了热电制冷器的功率芯片示意图 (b)热流示意图

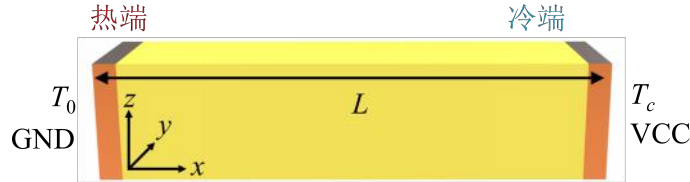


图2 P型热电臂示意图

离的一维解析方法来描述 TEC 的热行为。然而，该方法忽略了电场与热场之间的耦合效应。作为改进，文献[8]提出了一种新的边界条件，并推导出了最优电流密度的解析解，以实现最大冷却热流密度。但该方法未考虑 TEC 器件中的塞贝克效应，在某些应用场景下可能导致结果不准确。

本文提出了一种用于热电制冷封装系统仿真设计的高效方法，以提升其热管理性能。首先将整个系统划分为热电器件区域（TDR）和非热电区域（NTR）。为了实现 TEC 器件的实时优化，基于宏模型对系统中的 TDR 和 NTR 进行处理。首先，推导出一种新颖的一维等效热阻模型（ETRM）来表征热电器件。随后，基于宏模型技术，将 NTR 对 TDR 的热影响转化为在接口处的多端口热阻网络（MTRN）。TDR 与 NTR 仅在指定的端口处通过温度连续性和热流连续性条件进行耦合，从而可以快速求解端口处的温度与热流。最后，利用端口热流重构芯片封装的温度分布。

2 模型建立

2.1 电热耦合模型

在热管理中，TEC 器件通常用于将芯片产生的热量传递，并通过集成散热器(IHS)进行耗散，如图 1(a)所示。在电场中，考虑塞贝克效应的电流密度 J 的表达式为

$$\mathbf{J} = \sigma \mathbf{E} - \sigma S \nabla T \quad (1)$$

式中， $\mathbf{E} = -\nabla\phi$ 是电场， ϕ 是电势， σ 是电导率， S 是塞贝克系数， T 是温度。对于电学边界条件，热电臂的热端接地，冷端由一个理想的电压源作为激励

$$\phi_c = V \quad (2)$$

在热场中，由傅里叶传热 q_F 和帕尔贴效应 q_P 产生的热流可表示为

$$\mathbf{q} = \mathbf{q}_F + \mathbf{q}_P = -\kappa \nabla T + \Pi \mathbf{J} \quad (3)$$

式中， κ 是热导率， $\Pi = S \times T$ 是帕尔贴系数。

考虑焦耳热效应的热传导方程为

$$\nabla \cdot \mathbf{q} = \nabla \cdot (-\kappa \nabla T + \Pi \mathbf{J}) = \mathbf{J} \cdot \mathbf{E} \quad (4)$$

对于热学边界条件，热电臂的热端位于散热器下方，并将温度 T_h 设为恒定值

$$T_h = T_0 \quad (5)$$

冷端连接到芯片上来移除热量

$$\mathbf{n} \cdot \mathbf{q}_\Gamma = -\kappa (\mathbf{n} \cdot \nabla T) + ST (\mathbf{n} \cdot \mathbf{J}) \quad (6)$$

其中， \mathbf{n} 是单位法向量， \mathbf{q}_Γ 是界面上的热流。

2.2 一维等效热阻模型(1-D ETRM)

如图 1(b)所示，热量主要沿垂直方向流经 TEC 区域。因此，建立一个一维 TEC 模型来表示三维 TEC 热电臂是合理的。如图 2 所示，热流方向被指定为 x 轴。TEC 热电臂的长度为 L ，冷端热流为 q_c ，冷端温度为 T_c ，热端温度为 T_0 。

热电臂中的一维电热耦合方程可以表示为

$$\sigma \frac{d^2 \varphi}{dx^2} + \sigma S \frac{d^2 T}{dx^2} = 0 \quad (7)$$

(7)的通解为

$$\varphi(x) = -ST(x) + D_1 x + D_2 \quad (8)$$

其中 D_1 和 D_2 是未知系数。基于电学边界条件，我们可得

$$\varphi(0) = -ST(0) + D_2 = 0 \quad (9)$$

$$\varphi(L) = -ST(L) + D_1 L + D_2 = V \quad (10)$$

其中 $T(L)$ 等于冷端温度 T_c 。因此，未知系数可以计算得

$$D_1 = \frac{V}{L} + \frac{S}{L} T_c - \frac{D_2}{L} \quad (11)$$

$$D_2 = ST_0 \quad (12)$$

电流密度 J 可以写为

$$J = -\frac{\sigma}{L} (V + ST_c - ST_0) \quad (13)$$

其中， J 沿 x 轴的负方向 ($J < 0$) 且与位置无关。从(13)可知，电流密度 J 与电压 V 、冷端温度 T_c 成正比。

对于温度场分布，一维热传导方程可表示为

$$\kappa \frac{d^2 T}{dx^2} - S \left(\mathbf{J} \cdot \frac{dT}{dx} \right) + \mathbf{E} \cdot \mathbf{J} = 0 \quad (14)$$

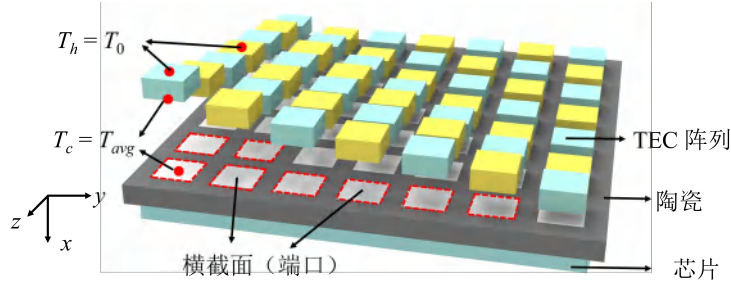


图 3 位于热电臂和陶瓷之间的横截面示意图

一维场强 E 表示为

$$E = \frac{J}{\sigma} + S \frac{dT}{dx} \quad (15)$$

将(15)带入(14)，我们可以得到

$$\kappa \frac{d^2T}{dx^2} + \frac{J^2}{\sigma} = 0 \quad (16)$$

此处，热导率 κ 被假设为常数。由于电流密度 J 与位置无关，因此可以推导出方程(16)的通解为

$$T(x) = -\frac{J^2}{2\kappa\sigma} x^2 + c_1 x + c_2 \quad (17)$$

其中 c_1 和 c_2 是未知系数。

施加图 2 所示的热学边界条件可得

$$c_1 = \frac{1}{(\kappa - SJL)} \left(\frac{J^2 L}{\sigma} - \frac{SJ^3 L^2}{2\kappa\sigma} + SJT_0 + q_c \right) \quad (18)$$

$$c_2 = T_0 \quad (19)$$

因此，通过(13)和(17)的表达式成功构建了一维 ETRM 模型，该模型建立了电压 V 、电流密度 J 、冷端热流 q_c 与温度 T 之间的关系。一旦给定 q_c ，就可以通过以下迭代方法计算热电臂的温度分布：对于一个初始的 T_c ，通过(13)可以得到一个临时的 J 值；然后，利用(17)更新冷端的温度 T_c 。在上述过程中收敛后即可获得最终的温度分布。

2.3 多端口热阻网络(MTRN)

除了 TEC 器件外，封装系统通常还包括基板、导热界面材料 (TIM)、有源芯片、散热器等。NTR 的三维复杂结构会在仿真和优化过程中带来较高的计算成本。在本工作中，提出利用宏建模将 NTR 对 TDR 的影响转化为界面上的 MTRN。

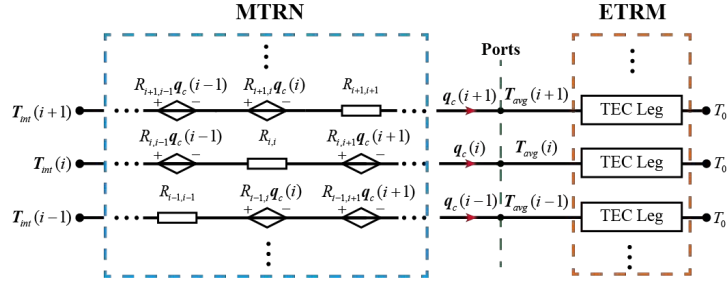


图 4 多端口热阻网络示意图

首先，基于三维有限元法（FEM），建立 NTR 的矩阵方程如下：

$$\mathbf{AT} = \mathbf{b} \quad (20)$$

其中， \mathbf{A} 是刚度矩阵， \mathbf{T} 是 NTR 中的温度， \mathbf{b} 是激励向量。

如图 3 所示，假设 NTR 与 M 个热电臂相连，每一个对应一个端口。在实际中，根据有限元对其进行离散化，每个端口上存在多个小单元，且每个单元的温度不同。由于端口连接着金属焊盘，所以端口处的温度梯度非常小。因此，将每个端口的平均温度 T_{avg} 作为冷端温度 T_c 是合理的。为了获得 NTR 中的传热特性，依次在各个端口施加热流密度激励。第 i 个 ($i = 1, 2, \dots, M$) 端口上的热流表示为 q_i ，其平均温度定义为 T_{avg}^i 。值得指出的是，当施加不同的热流时，矩阵 \mathbf{A} 保持不变。因此，在使用稀疏线性直接求解器时，该方法便于重复利用矩阵 \mathbf{A} 的 LU 分解信息，提高了计算效率。从第 j 个端口到第 i 个端口的等效热阻 R_{ij} 可通过下式计算：

$$R_{ij} = \frac{\mathbf{T}_{avg}^j(i) - \mathbf{T}_{int}(i)}{q_j} \quad (21)$$

其中， $\mathbf{T}_{avg}^j(i)$ 表示当在第 j 个端口施加热流 q_j 时，第 i 个端口上的平均温度， $\mathbf{T}_{int}(i)$ 表示所有端口均为绝热边界时，第 i 个端口上的初始温度。然后建立 MTRN，以矩阵形式构建温度差与热流之间的关系，表示为：

$$\mathbf{R} \cdot \mathbf{q}_c = \mathbf{T}_{int} - \mathbf{T}_{avg} \quad (22)$$

其中， \mathbf{R} 是一个 $M \times M$ 矩阵， \mathbf{q}_c 是一个 $M \times 1$ 的向量，表示端口上的冷却热流密度， \mathbf{T}_{avg} 是一个 $M \times 1$ 的向量，表示端口上的平均温度。MTRN 的示意图如图 4 所示。需要注意的是，MTRN 与 NTR 内部的材料属性和几何结构有关，但不依赖于激励值 q_j 。通过将一维 ETRM 与 MTRN 相结合，可以高效率地获得冷端热流 q_c 。最终，将得到的热流 q_c 作为 NTR 的激励，用于重构其温度分布。

3 数值结果

为验证所提出方法处理复杂结构的有效性和效率，本文对一个集成 TEC 冷却系统的 2.5-D 封装结构进行了仿真。多个芯片紧密地布置在一个中介层上。芯

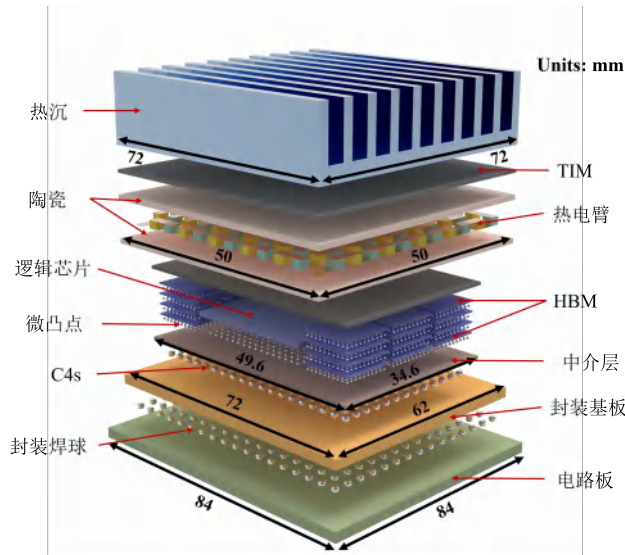


图 5 集成 TEC 冷却系统的 2.5-D 封装示意图

表 1 几何尺寸与材料参数

结构	尺寸(length × width × thickness) (mm)	热导率 (W/[m·K])
逻辑芯片	32.5×25.4×0.72	120
HBM	10×10×0.14	120
中介层	34.6×49.6×0.35	50
微凸点	10×10×0.04	// 43, ⊥ 60
衬底	72×62×1.9	3
TIM	34.6×49.6×0.1	6
EMC	34.6×49.6×0.72	0.7
C4s	34.6×49.6×0.07	// 43, ⊥ 60
陶瓷	50×50×0.5	20
热电臂	1×1×1.5	1.4

// 和 ⊥ 分别代表水平和垂直方向

片与中介层之间的最小距离使得局部热量更容易积聚,从而缩短集成电路的使用寿命。因此,温度分布分析对于维持整个系统的性能和可靠性至关重要。

如图 5 所示,该 2.5-D 封装结构包括逻辑芯片、高带宽存储器(HBM)、导热界面材料(TIM)、微凸块(micro bumps)、物理层(PHY)、受控塌陷芯片连接(C4s)、环氧模塑料(EMC)等组件。需要指出的是散热片并没有实际建模,而是利用恒温边界条件进行代替。EMC 填充在芯片周围及之间的间隙中,其顶面与芯片表面齐平。需要注意的是,微凸块和 C4s 的热特性通过等效热导率进行建模[9]。P 型热电臂的塞贝克系数为 $S = 240 \mu\text{V/K}$, N 型为 $S = -240 \mu\text{V/K}$,电导率为 $\sigma = 7 \times 10^4 \text{ S/m}$ 。外部环境温度 $T_{\text{ext}} = 293.15 \text{ K}$,封装区域的对流换热系数为 $10 \text{ W/(m}^2\cdot\text{K)}$ 。总共 287 对热电臂元件均匀分布在 TEC 的两个陶瓷基板之间。几何与材料参数详见表 1。

表 2 本文方法与 COMSOL 的对比

方法	COMSOL	本文	提升
自由度(DoFs)	420,470	262,988	1.6×
预仿真时间	None	2 m 58 s	—
总计算时间	36 h 4 m 16 s	1h 20 m 2 s	27×

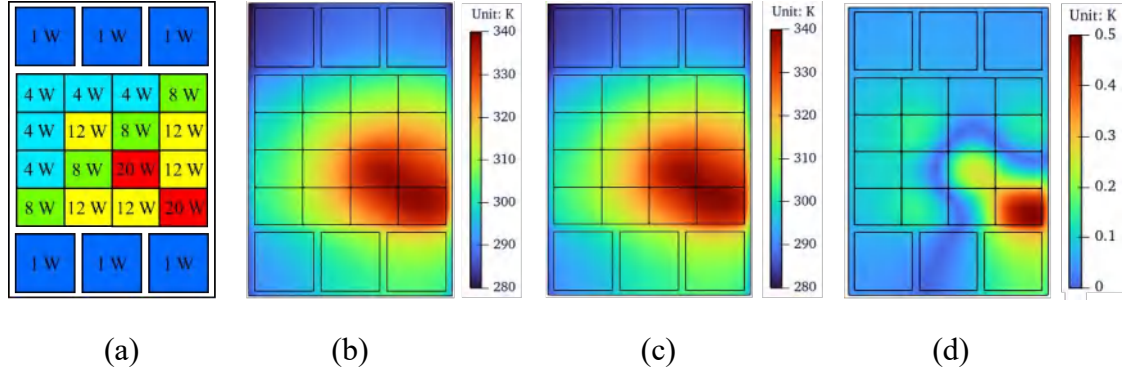


图 6 (a) 2.5-D 封装的功率分布图。(b) COMSOL 温度分布图。(c) 本文所提方法的温度分布图。(d) 绝对误差。

为了验证所提方法在处理复杂问题时的有效性，逻辑芯片和 HBM 的功率分布如图 6(a)所示。在电热仿真中，当施加的 TEC 总电压为 36 V 时，图 6(b)-(c) 分别展示了利用所提出方法和 COMSOL 得到的温度分布图。图 6(d)展示了两者之间的绝对误差分布，最大值小于 0.5 K。这表明在处理复杂结构的时候，本文所提方法可以保证准确性。在计算效率方面，本文分析了 370 种不同参数组合的情况：施加电压从 1 V 增加到 85 V，热电臂长度从 1.1 mm 变化到 2 mm。电热仿真的 CPU 时间对比见表 2。由于需要构建 MTRN，所以本文方法需要有预仿真时间。当调整热电臂长度时，COMSOL 需要重新构建几何结构；而本文提出的 1-D EHFm 无需进行网格划分，在优化热电支腿结构方面更具效率。COMSOL 仿真时间为 36 h 4 m 16 s，本文方法为 1 h 20 m 2 s（包含预仿真时间）。与 COMSOL 相比，所提出的方法实现了 27 倍的加速，展现出其高效的计算性能。

4 结论

本文提出了一种用于热电制冷封装系统设计与仿真的高效方法。为了解决 TDR 仿真效率低的问题，开发一种新颖的 ETRM，将原始的三维模型转化为紧凑的一维模型，从而显著降低计算复杂度。随后，引入宏建模将 NTR 对 TDR 的影响转化为界面上的 MTRN，使得 TDR 与 NTR 仅在端口处耦合，从而实现了界面热流的快速求解。接着，利用所得的端口热流重构 NTR 的温度分布。通过数值算例验证了所提方法的准确性，与商业软件相比，本文方法在计算效率上提

升了 27 倍。未来的工作将聚焦于将该方法扩展到更复杂的系统，并探索其在实时热管理应用中的潜力。

参考文献

- [1] L. Xiu. Time Moore: Exploiting Moore's Law from The Perspective of Time[J]. IEEE J. Solid-St. Circ., 2019, 11(1): 39-55.
- [2] D. M. Rowe. Thermoelectrics Handbook: Macro to Nano. Boca Raton[M]. FL, USA: CRC Press, 2005.
- [3] L. Mu, H. Zhang, J. Cui, J. Zhao. Numerical study of transient supercooling performance and thermal stress analysis of segmented annular thermoelectric cooler[J]. Appl. Therm. Eng., 2024, 245.
- [4] R. A. Rathnayake, H. Wijekoon, et al. Time-dependent finite-difference model for transient and steady-state analysis of thermoelectric bulk materials[J]. Phys. Scr., 2022, 97(12).
- [5] R. Hamzehzarghani, M. Eslami. Feasibility evaluation of a thermoelectrically cooled photovoltaic module at different arrangements and working conditions: A numerical study[J]. Therm. Sci. Eng. Prog., 2024.
- [6] L. Chen, R. Liu, and X. Shi. Thermoelectric Materials and Devices[M]. Cambridge, MA, USA: Elsevier, 2020.
- [7] Y. Sheikhejad, R. Bastos, Z. Vujicic, A. Shahpari, and A. Teixeira. Laser thermal tuning by transient analytical analysis of peltier device[J]. IEEE Photon. J., 2017, 9(3): 1–13.
- [8] L. Chen, S. Sadiqbatcha, H. Amrouch, and S. X.-D. Tan. Electrothermal simulation and optimal design of thermoelectric cooler using analytical approach[J]. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 2022, 41(9): 3066–3077.
- [9] X. Rao, J. Song, Q. Tian, et al. A universal and efficient equivalent modeling method for thermal analysis of 3D ICs containing tapered TSVs[J]. Int. Commun. Heat Mass Transf., 2022.

基于异构集成技术的小型化宽带太赫兹集成电路

-波导互联结构

杜凯源^{1,2}, 夏鹤翔^{1,*}, 袁斌¹, 喻忠军^{1,2,*}

¹ 中国科学院空天信息创新研究院

² 中国科学院大学电子电气与通信工程学院

通讯作者: 夏鹤翔 xiahx@aircas.ac.cn 13487076137

喻忠军 yuz.j@ucas.ac.cn 18612208244

摘要: 本文提出一种基于异构集成技术的高集成度、高性能太赫兹集成电路 (TMIC) -波导 (RWG) 互联结构, 利用硅基 MEMS 工艺以及 BCB 薄膜工艺将 TMIC、RWG、波导过渡结构一体化集成在同一硅基板上。仿真结果表明该互联结构在其在 262 - 297 GHz (35 GHz, 12.5%) 内回波损耗低于 15 dB, 插入损耗低于 0.6 dB。高集成度、大带宽与低损耗特性使其在下一代小型化高性能太赫兹系统中具有广阔的应用前景。

关键词: 太赫兹; 异构集成; MEMS; BCB; 宽带波导过渡;

1 前言

近年来,随着物联网、5G/6G 及自动驾驶等新一代无线通信技术的飞速演进,对超高速数据传输与高分辨率成像的需求不断攀升。太赫兹技术凭借其丰富的频谱资源,可在信号速率与成像分辨率上超越传统微波与光学方案。要实现系统的小型化与高性能化,关键在于构建高带宽、低损耗、高集成度的太赫兹封装技术^[1]。矩形波导 (RWG) 以结构简洁、传输损耗低的优势,成为太赫兹封装系统中信号传输的理想结构^[2]。因此,设计高性能、高集成度的太赫兹集成电路 (TMIC) -RWG 互联结构,对于提升系统性能、降低系统体积至关重要。

过去几十年间,为实现 TMIC-RWG 互联先后提出了多种方案,其中微带探针结构因其实现简便、结构紧凑而被广泛应用。然而,该结构通常会加载基于四分之一波长短路矩形波导的谐振腔以扩展带宽并提升转换效率,这显著增加了系统体积,并大幅提高了加工与装配难度。该结构通常会使用金属机加工加工出波

导腔体，并使用以石英、玻璃、陶瓷为代表的厚膜基板电路加工出微带探针用于激励波导；TMIC 放置在基板上并基板电路连接；这使得该结构集成度低，且工艺精度难以适配太赫兹频段下极小的结构尺度，造成性能下降。目前已有针对该方案结构与工艺上的改进，包括使用端壁（End-Wall）激励结构替代四分之一波长短路矩形波导^[3,4]，从而实现更加紧凑的电路结构；使用陶瓷、硅基板实现集成化波导代替金属加工波导，从而提升结构集成度等^[5,6]。但目前这些方案的带宽、插入损耗等性能仍不够优秀，且难以与 TMIC 进行高性能集成，需要引入额外的基板，并使用金丝键合、芯片倒装等精度较低微组装工艺。波导过渡结构与 TMIC 集成结构两者不同的工艺不仅增加了复杂度，还大幅提升了因装配误差所导致的性能恶化。

近年来，异构集成技术凭借其对不同功能、不同材料芯片以及无源电路进行高性能集成的能力，被认为是实现太赫兹系统小型化、高性能化的最有希望的途径。因此，针对上述挑战，本文提出一种基于硅基微机电系统（MEMS）工艺的 TMIC-RWG 互联结构，利用异构集成技术将 TMIC、RWG、波导过渡结构一体化集成在同一硅基板上，在 300 GHz 频段实现了高集成度、高带宽、低损耗的太赫兹传输通路，为下一代小型化、高性能太赫兹系统提供了新的可行思路。

2 异构集成太赫兹芯片-波导互联结构

本文提出的太赫兹芯片-波导互联结构主要包含四个部分，由下至上分别为硅基板、底层金属、介质层以及顶层金属，如图 1 所示。其中，硅基板通过 MEMS 工艺刻蚀出能够容纳 TMIC 的腔体，从而将 TMIC 嵌入基板内。这种 TMIC 集成结构拥有极高的集成度与工艺精度，同时能够极大的缩小互联距离，降低太赫兹信号的传输损耗。底层金属、介质层与顶层金属组成接地共面波导（GCPW）结构，与 TMIC 焊盘通过介质层通孔进行互联。GCPW 末端为圆形激励探针，与介质层通孔形成的腔体、U 型匹配膜片共同组成紧凑的 End-Wall 激励结构，用以激励下方的 RWG。下方的 WR-3 RWG 通过 MEMS 技术以及穿硅环形沟槽（Through Silicon Ring Trench, TSRT）技术在硅基板上实现，拥有极高的集成度

以及加工精度。

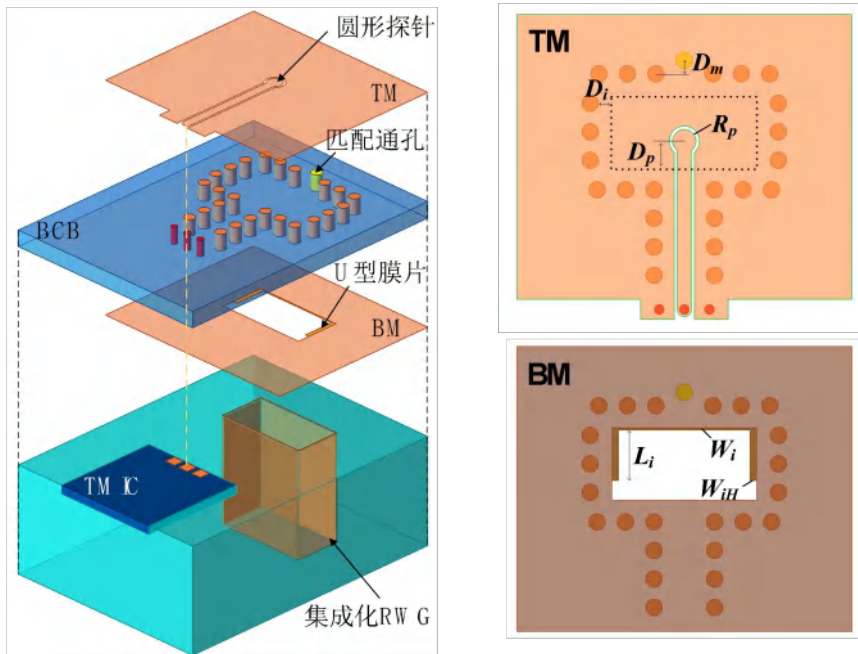


图 1 异构集成 TMIC-RWG 互联结构示意图

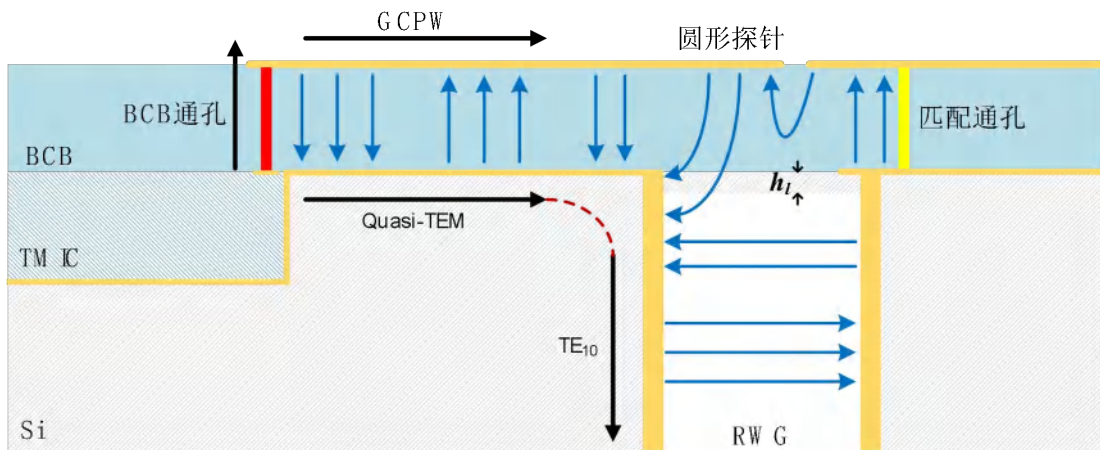


图 2 异构集成 TMIC-RWG 互联结构横截面与电磁场模式示意图

该异构集成互联结构剖面图以及信号通路示意图如图 2 所示（以信号流出 TMIC 为例）。太赫兹信号从 TMIC 焊盘输出，通过介质通孔与两侧地孔形成的类共面波导（CPW）结构馈入 50Ω GCPW 传输线。在 GCPW 上，电磁波以类横电磁波（Quasi-TEM）的形式传播。信号传输至末端圆形开路探针，在由一圈屏蔽地孔围成的类金属腔体中激励出谐振模式，该模式电场分布与 RWG 的主模 TE_{10} 模的电场分布近似，从而激励起 RWG 的主模传输，实现信号从平面 GCPW

到 RWG 的过渡。

本文所提出的太赫兹芯片-波导互联结构拥有下列优势：

(1) 互联结构基于硅基 MEMS 工艺以及集成电路薄膜工艺设计，拥有极高的制造精度以及较小的特征尺寸，有利于结构的小型化。

(2) TMIC 集成结构、RWG 以及波导过渡均使用同一类工艺在同一硅基板上集成，不仅拥有极高的集成度，还减少了工艺流程，避免了多次装配流程可能造成的性能恶化。

(3) 采用 End-Wall 激励结构，无需额外短路枝节，大幅减小了结构体积，降低了结构复杂度，并且解放了基板上方空间，可进一步实现三维堆叠等设计。

2.1 基于 MEMS 与 BCB 工艺的异构集成互联结构

通过硅基 MEMS 以及 TSRT 工艺制作的腔体、波导结构，相比于传统的金属机加工工艺，拥有更小的尺寸、更高的加工精度以及更低的金属表面粗糙度，从而获得更高的 TMIC 装配精度以及太赫兹信号传输效率。苯并环丁烯 (Benzocyclobutene, BCB) 作为一种新型的光敏性树脂材料，能够适配集成电路薄膜工艺，且拥有较低的相对介电常数 ($\epsilon_r \approx 2.65$) 以及极低的损耗角正切 ($\tan \delta \approx 0.0008$)，这使其极为适合在太赫兹频段用作信号传导介质。通过结合这两种工艺，实现高集成度、高性能的异构集成互联结构。

该互联结构所使用的异构集成工艺大致流程如图 3 所示。图 3 (a)：准备厚度 $1000\mu m$ ，双面涂覆 Si_3N_4 保护层的硅片作为衬底；图 3 (b)：正面旋涂光刻胶，经过曝光、显影后形成保护层刻蚀掩模；图 3 (c)：使用湿法刻蚀工艺，在硅片表面嵌入芯片位置刻蚀出初级沟槽；图 3 (d)：正面经过光刻流程，定义 TSRT 刻蚀区域，利用深硅刻蚀技术得到环形沟槽；图 3 (e)：在环形沟槽内溅射金属种子层，随后通过电镀工艺在沟槽中填充金属，实现 TSRT；图 3 (f)：通过化学机械抛光 (CMP) 研磨多余铜层并抛光硅片表面；图 3 (g)：通过光刻流程，在硅片上定义出芯片嵌入区域，随后通过干法刻蚀工艺在衬底上形成嵌入式沟槽；图 3 (h)：在衬底上溅射种子层，通过光刻定义底层金属图形，随

后电镀制作底层金属地，最后通过离子束刻蚀移除种子层；图 3 (i)：将芯片嵌入沟槽，使用银浆将芯片粘接于沟槽内，实现芯片与衬底的电气及机械连接；图 3 (j)：正面旋涂光敏 BCB 介质，通过曝光定义 BCB 通孔位置；图 3 (k)：溅射种子层，经过光刻与电镀工艺填充通孔以及形成顶层金属图案；图 3 (l)：晶圆背面减薄至 $500\mu\text{m}$ ，暴露出 TSRT 金属；图 3 (m)：背面旋涂光刻胶，经过光刻流程定义 RWG 位置；经过刻蚀工艺移除 TSRT 内部的硅介质，得到集成化 RWG；图 3 (n)：移除光刻胶，对晶圆进行清洗，得到异构集成 TMIC-RWG 互联结构。

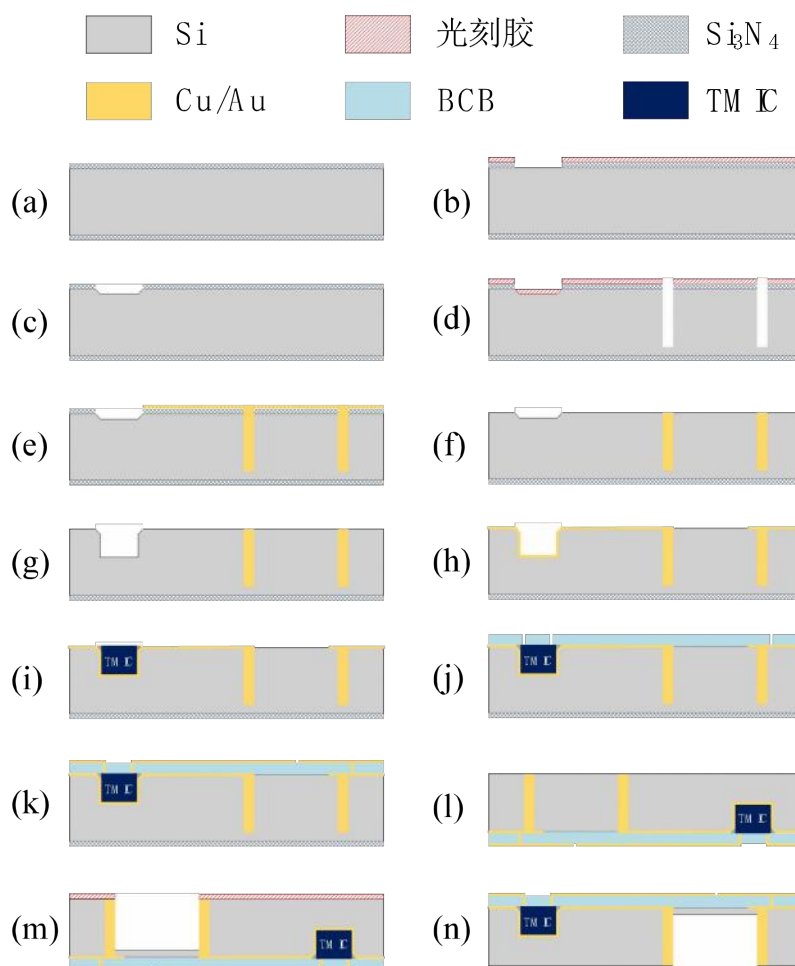


图 3 异构集成互联结构工艺流程

2.2 宽带 End-Wall 波导过渡结构设计

由 GCPW 到 RWG 的过渡结构设计如图 1 所示, RWG 位于 GCPW 金属地下

方，圆形开路探针直接插入 RWG 的端壁进行激励，形成 End-Wall 结构。分析图 2 中的电场分布可知，电磁场由 GCPW 中的 Quasi-TEM 模式转换为 RWG 中的 TE_{10} 模式；为匹配两种模式下电磁场分布，从而提高转换效率，圆形开路探针应放置在 RWG 短边中心附近^[7]。因此，优化圆形开路探针的位置 D_p 以及半径 R_p 对提升电路匹配效果、降低回波损耗与插入损耗是十分必要的，如图 4 (a)所示。值得注意的是，当 R_p 较大时，圆形探针在高频段将会向上方空间辐射信号，造成额外的辐射损耗，导致电路损耗增加，如图 4 (b)所示。因此，经过参数扫描优化，最终选择圆形开路探针半径为 $R_p = 70\mu\text{m}$ ，位置 $D_p = 90\mu\text{m}$ 。

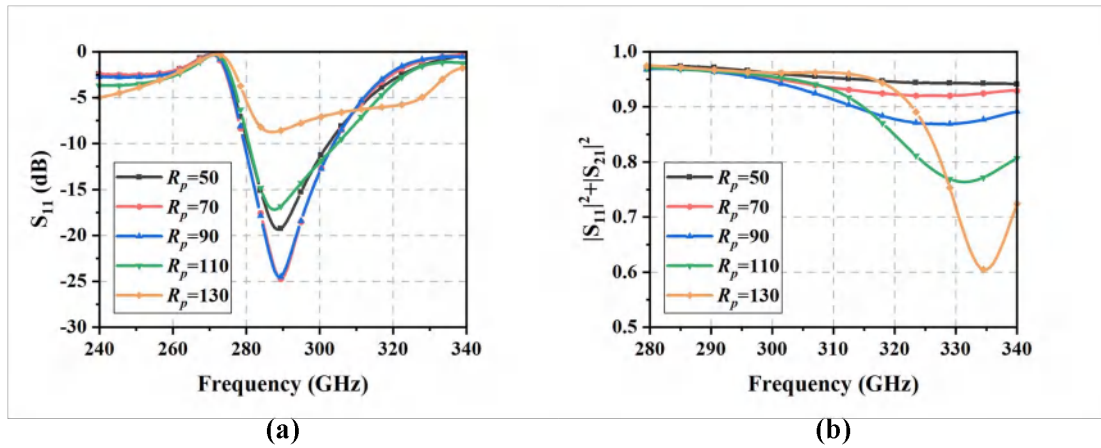


图 4 圆形探针直径 R_p 对互联性能的影响 (a) S_{11} . (b) $|S_{11}|^2 + |S_{21}|^2$

探针周围由一圈屏蔽孔以及上、下层金属地形成金属谐振腔体，其腔体大小主要由屏蔽孔到 RWG 开口的距离 D_i 决定。该腔体大小将极大影响过渡结构的中心频率：腔体增大，中心频率将向低频方向移动，如图 5 (a)所示。当腔体过小时，中心频率过度向高频偏移，使得较低频区域难以匹配；当腔体过大时，腔体内电磁场与 RWG 主模电磁场分布差异过大，激发高次模并造成信号大量反射。经过参数优化，最终设置 $D_i = 80\mu\text{m}$ 。同时，可移动上方中央的屏蔽通孔（匹配通孔），改变腔体内电磁场分布，从而进一步提高转换效率并调整中心频率，如图 5 (b)所示。结合对 D_i 与 D_m 的参数优化，最终该过渡结构中心频率被设置在 290 GHz 附近。

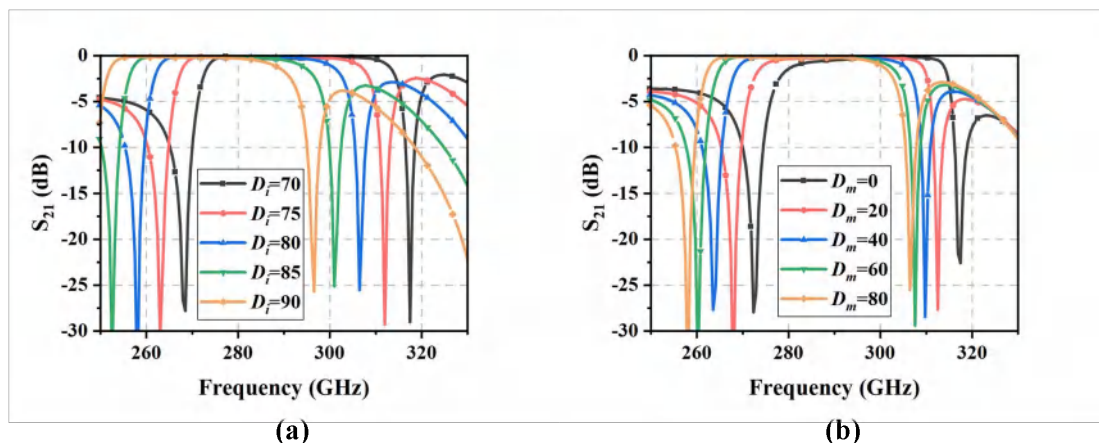


图 5 BCB 通孔对中心频率的影响 (a) 屏蔽通孔围成的腔体大小. (b) 匹配通孔位置

上述结构虽然在 290 GHz 附近实现了良好的匹配 ($S_{11} \approx -25$ dB)，但带宽较窄，难以满足太赫兹超宽带应用。为了进一步拓展带宽，在底层金属 RWG 端口处添加 U 型膜片，如图 1 所示。该膜片由宽边的 E 型膜片以及短边的 H 型膜片组成，可分别等效为波导传输线中并联的电容与电感。通过引入新的电抗器件，能够对过渡结构的频率响应进行进一步的调谐，降低结构在低频处的信号反射，从而提高整体带宽。图 6 展示了 E 型膜片宽度 W_i 以及 H 型膜片长度 L_i 对系统频率响应的影响。可以看出， W_i 主要影响低频响应，能够将系统通带的截止频率向低频扩展；而 L_i 主要影响高频响应，能够将系统通带的截止频率向高频扩展。

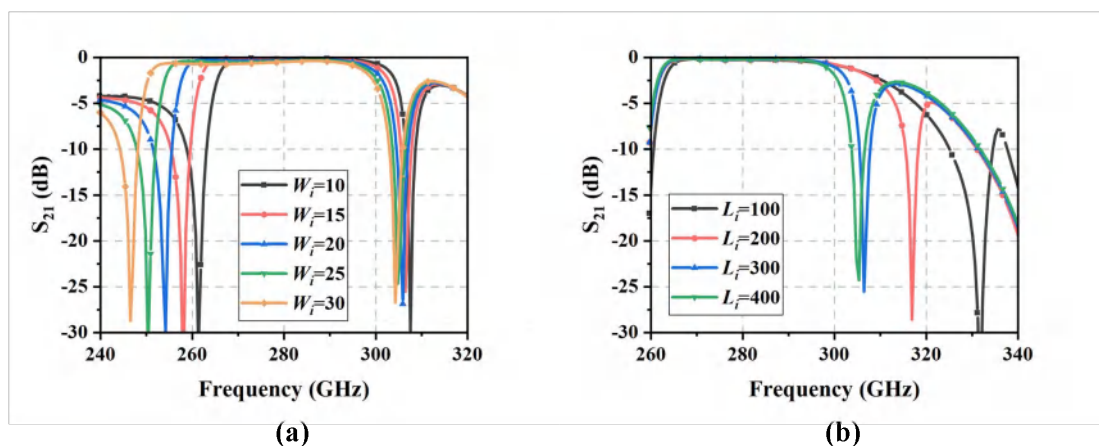


图 6 U 型膜片参数对拓宽互联结构带宽的效果 (a) E 型膜片宽度. (b) H 型膜片长度

集成化波导结构通过硅基 MEMS 以及 TSRT 工艺实现，如图 3 所示。波导尺寸选择标准的 WR-3 RWG，为 $0.8636 \times 0.4318 \text{ mm}^2$ ，方便与外部电路连接。在图 3 (m) 所示的波导刻蚀阶段，由于硅基板还需承载正面的金属电路以及 BCB

层，因此不能全部移除，剩余的硅厚度 h_f 将对系统性能产生影响。由于本文所使用的高阻硅基板介电常数为 11.9，远高于 BCB 以及空气的介电常数，因此最终谐振腔体中的介质等效相对介电常数 $\epsilon_e > \epsilon_r$ ，使得系统中心频率向低频移动，如图 7 (a)所示。

实际设计中，需要综合分析上述所有参数对互联频率响应的影响。通过参数调谐与优化，最终得到宽带、低损耗的波导过渡结构。

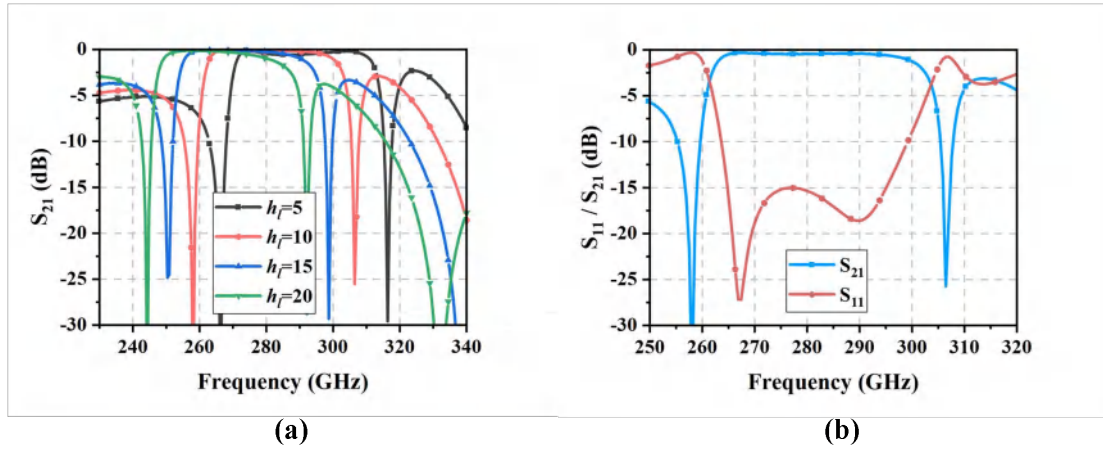


图 7 (a) 刻蚀后硅片剩余厚度对互联性能的影响. (b) 异构集成 TMIC-RWG 互联结构性能

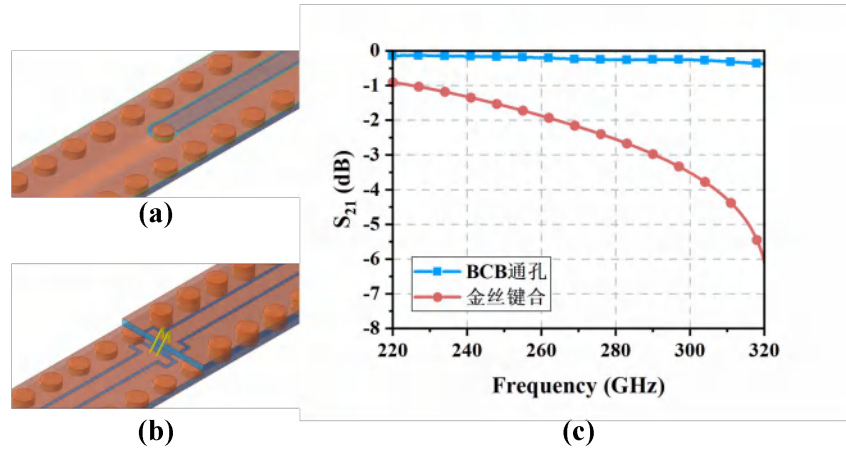


图 8 BCB 通孔互联性能对比 (a) 互联模型. (b) 金丝键合模型. (c) 互联损耗对比

3 仿真结果

本设计中 TMIC 嵌入在硅基板内，通过 BCB 通孔形成的类 CPW 结构与 GCPW 传输线互联。对该结构进行建模分析，如图 8 所示。在对 GCPW 传输损耗去嵌入后，可得 BCB 通孔互联插入损耗在 300 GHz 约为 0.23 dB。相比于使用金丝键合、芯片倒装工艺实现的 TMIC 集成，该方案拥有显著更低的损耗以及更

高的集成度。

对波导过渡结构进行优化迭代后，最终设计参数值如下： $R_p = 0.07 \text{ mm}$ ， $D_p = 0.16 \text{ mm}$ ， $D_i = 0.13 \text{ mm}$ ， $D_m = 0.08 \text{ mm}$ ， $W_i = 15 \mu\text{m}$ ， $W_{itl} = 40 \mu\text{m}$ ， $L_i = 0.3 \text{ mm}$ 。最终得到的异构集成 TMIC-RWG 互联结构仿真性能如图 7 (b)所示，其在 262 – 297 GHz (35 GHz, 12.5%) 内回波损耗 $RL > 15 \text{ dB}$ ，插入损耗 $IL < 0.6 \text{ dB}$ 。表将结果与目前前沿的工作进行了对比，表明本设计在实现了 TMIC、RWG、波导过渡异构集成的同时，拥有较为理想的互联性能。

	工艺类型	互联类型	频率 (GHz)	插入损耗	回波损耗
[8]	金属机加工 介质基板	SIW-RWG	75 – 110 (37.8%)	0.6 dB [#]	> 10 dB
[9]	金属机加工 介质基板	MS-RWG	73.5 – 79 (12.5%)	1.5 dB [#]	> 15 dB
[10]	金属机加工 石英探针	TMIC-RWG	92 – 96 (4.3%)	1.5 dB [#]	> 19 dB
[11]	异构集成 金丝键合	TMIC-RWG	93 – 95 (2.1%)	2 dB [*]	> 15 dB
本文	异构集成	TMIC-RWG	262 - 297 (12.5%)	0.4 dB[#] / 0.6 dB[*]	> 15 dB

[#] 不包含 TMIC 互联部分. ^{*} 包含 TMIC 互联部分

表 1 本设计与目前前沿工作的对比

4 结论

本文提出了一种新型的基于异构集成技术的宽带高性能太赫兹集成电路-波导互联结构。该互联结构使用硅基 MEMS 工艺以及 BCB 薄膜工艺，将太赫兹集成电路、矩形波导、波导过渡结构异构集成在同一硅基板上。过渡结构使用了端壁激励结构，并结合屏蔽孔腔体、匹配通孔以及 U 型膜片，实现了宽带、高效、

紧凑的 GCPW Quasi-TEM 模式到 RWG 中的 TE_{10} 模式转换电路。仿真结果显示了该互联结构在带宽、回波损耗与插入损耗上性能优异。结合其高集成度的优势，该结构能进一步推动太赫兹系统的小型化、高性能化。

致谢

本项目由中国科学院空天信息创新研究院科学与颠覆性技术研究基金项目（2024-AIRCAS-SDTP-02）资助。

参考文献

- [1] Yang X, Huang X-L, Zhou L, et al. A 3-D Heterogeneously Integrated Application of the RF-Module With Micro-Bump Filter and Embedded AiP at W-Band [J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2024, 14(8): 1422-33.
- [2] Yu B, Wang Z, Li O, et al. A 200-GHz Four-Element Phased-Array Receiver System-in-Package Using HTCC Technology for Sub-Terahertz Communications [J]. IEEE Transactions on Microwave Theory and Techniques, 2023: 1-15.
- [3] Yuan B, Wu P, Yu Z, et al. Wideband End-Wall Transition From Microstrip to Waveguide With via-Less Choke Structure for Terahertz Application [J]. IEEE Transactions on Terahertz Science and Technology, 2022, 12(3): 317-20.
- [4] Khani B, Makhoul S, Steffan A G, et al. Planar 0.05–1.1 THz Laminate-Based Transition Designs for Integrating High-Frequency Photodiodes With Rectangular Waveguides [J]. Journal of Lightwave Technology, 2019, 37(3): 1037-44.
- [5] Cheng H, Zhu X, Hou F, et al. Compact 31-W 96-GHz Amplifier Module in GaN-MEMS for Wireless Communications [J]. IEEE Transactions on Microwave Theory and Techniques, 2022, 70(2): 1233-41.
- [6] Yu B, Wang Z, Wu P, et al. High-Integration and Low-Cost Transmitter Packaging Solution for 0.2 THz SiP Application Using HTCC Technology [J]. IEEE Microwave and Wireless Components Letters, 2022, 32(6): 680-3.
- [7] Yuan B, Wu P, Yu Z, et al. A Novel End-Wall Waveguide Excitation With Wide Bandwidth and Simple Structure for Millimeter-Wave/Terahertz Application [J]. IEEE Microwave and Wireless Components Letters, 2022, 32(7): 831-4.
- [8] Hansen S, Pohl N. A W-Band Stepped Impedance Transformer Transition from SIW to RWG for Thin Single Layer Substrates with Thick Metal Cladding [J]. 2019 49th European Microwave

Conference (Eumc), 2019: 352-5.

[9] Thanh Tuan N, Sakakibara K, Kikuma N. Bandwidth Extension of Planar Microstrip-to-Waveguide Transition by Controlling Transmission Modes Through Via-Hole Positioning in Millimeter-Wave Band [J]. IEEE Access, 2019, 7: 161385-93.

[10] Cheng H, Hou F, Guo J, et al. MEMS Chip With Amplifier for 4-W Power Combining up to 100 GHz [J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2020, 10(5): 779-85.

[11] Yang X, Huang Y-S, Zhou L, et al. Low-Loss Heterogeneous Integrations With High Output Power Radar Applications at W-Band [J]. IEEE Journal of Solid-State Circuits, 2022, 57(6): 1563-77.

三维系统级封装的单粒子效应Geant4模拟方法研究

李洋^{1,2}, 李亚勇¹, 向爽¹, 张国和², 贺朝会^{1,*}

核科学与技术学院, 西安交通大学

微电子学院, 西安交通大学

(*Email: hechaohui@xjtu.edu.cn, 西安市碑林区咸宁西路28号)

摘要: 三维系统级封装 (3D SiP) 具有三维堆叠多芯片结构和高功能/存储密度等新特征。其在地面重离子单粒子效应评估中面临重离子能量和射程不足的难题, Geant4模拟仿真成为解决该问题的重要途径之一。本文开展三维系统级封装的单粒子效应Geant4模拟方法研究, 建立在三维系统级封装器件的全尺寸Geant4模型, 包含通孔、焊盘、器件、封装等结构; 建立了三维封装器件的虚拟灵敏体积阵列和粒子追踪计算方法; 开展了重离子入射模拟, 统计了粒子径迹和能量沉积规律, 以及单粒子效应在封装内的空间分布规律。本文研究结果将对3D SiP的重离子评估和加固设计提供技术指导。

关键词: 系统级封装; 单粒子效应; 重离子; Geant4

1 引言

随着半导体特征尺寸逼近物理极限, 摩尔定律难以延续, 微系统技术成为突破瓶颈、延续芯片性能增长的关键路径。国际半导体路线组织 (International Technology Roadmap for Semiconductors, ITRS) 明确将系统级封装 (System-in-Package, SiP) 和系统级芯片 (System-on-Chip, SoC) 共同列为后摩尔时代集成电路的高价值系统, 二者成为集成电路的主要发展方向^[1,2]。其中, 系统级封装是通过应用先进封装技术将多个具有不同功能的电子元器件裸芯片封装成一个整体, 集成一个微系统。通过封装集成实现系统功能密度的提升, 降低了对小特征尺寸、高性能芯片的技术依赖, 但系统同样具备小型化、低功耗、高性能的优点。

封装技术发展趋势如图1所示。在传统封装的基础上, 先进封装通过倒装芯片、晶圆级封装、2.5D/3D封装技术增加封装内的功能密度、缩短互连长度和进行系统功能重构, 实现系统综合性能的提升^[3]。2.5D/3D系统级封装是当前高性

能微系统的典型代表，在空间电子学系统中具有广泛的应用前景。对于单粒子效应，SiP系统的2.5D/3D堆叠架构将带来不同于以往平面器件和PCB板的特殊辐射效应问题，并成为制约其可靠应用的突出因素^[4,5]：三维堆叠使部分裸芯片的位置较深，实验评估中对重离子的能量和射程提出了挑战；3D封装内部多层封装、互连、芯片材料与粒子作用下的能量沉积与电荷收集规律复杂；高能粒子穿过多个芯片模块的几率提升，同一个粒子事件可能对多层芯片造成影响，造成单粒子多芯片错误；高密度片间互联下的高速信号耦合加强，辐射与电源、信号完整性的耦合加深，等等。目前针对二维系统级封装的辐射效应研究和模拟已有一定成果（单粒子效应^[6-8]、总剂量效应^[9,10]、瞬时剂量率效应^[11,12]、电磁脉冲效应^[13]），但针对三维堆叠立体结构的SiP辐射效应的研究仍在起步阶段。

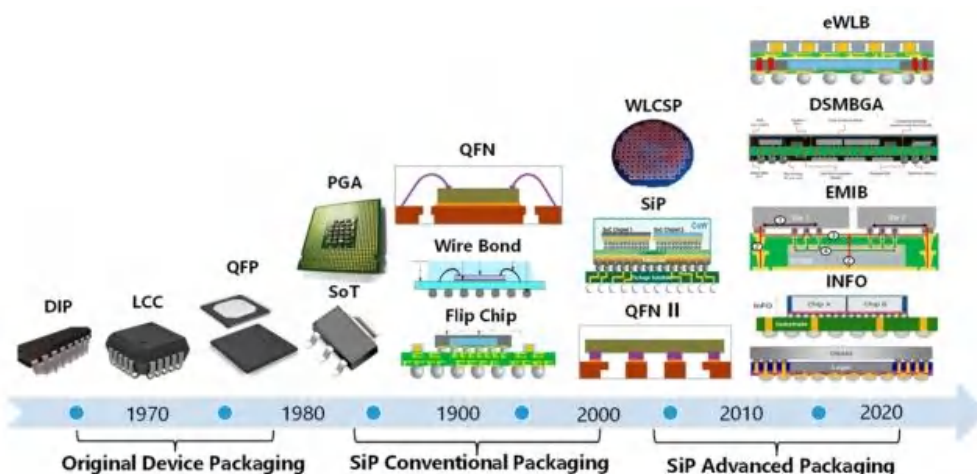


图 1 封装技术的发展趋势^[5]

空间环境中重离子种类多、能量分布范围极宽，器件重离子单粒子效应复杂，地面加速器评估过程中需同时满足重离子能量和射程的要求。然而，3D SiP在地面重离子单粒子效应实验评估中面临重离子能量和射程不足的难题。加速器产生的重离子因能量受限可能无法穿透3D SiP到达内部灵敏区域，造成器件空间单粒子效应敏感性评估误判，威胁航天器空间的安全。同样，加速器维护成本高昂，重离子实验评估中影响因素遍历会造成成本高昂的困境。

因此，有必要开展三维系统级封装的单粒子效应数值模拟计算研究，通过建立3D SiP的单粒子效应分析模型，模拟粒子在3D SiP模型中的能量沉积与单粒子效应分布规律，分析单粒子效应在封装内的影响机制，为3D SiP的辐射效应评估

和器件在空间辐射环境的应用提供支撑。

2 三维系统级封装Geant4建模

本文选取了典型结构的3D SiP开展了重离子单粒子效应模拟方法研究。首先，采用Geant4建立了全尺寸3D SiP单粒子效应分析模型，包括五层堆叠芯片与填充、芯片间互连通孔与焊盘等结构。其次，探索并建立了3D SiP的全局粒子追踪算法，以解决全尺寸、多芯片、超大规模存储阵列引发的计算需求暴增的问题。接着，开展了重离子入射模拟研究，统计了粒子径迹和能量沉积规律。最后，总结了单粒子效应在3D SiP内部的空间分布规律。

2.1 基于硅转接基板互连的三维系统级封装

三维系统级封装通常采用芯片垂直方向堆叠封装组成微系统，系统内部多芯片采用晶圆级互连或者硅转接基板互连。鉴于中国在发展先进制程、晶圆级互连技术等方面面临一定的外部限制，将较大尺寸裸芯片通过硅转接基板进行集成，成为了国产三维系统级封装的重要方案。

图2为基于硅转接基板互连的三维系统级封装典型结构。在封装时，从基板到顶部，芯片呈数层分布，各层间通过RDL布线、TSV通孔、焊球或混合键合等方式实现与其他芯片层的相互连接，封装中可采用的工艺包括凸块（Bumping）工艺、扇外型封装、FC焊接等。在整个芯片微模组焊接组装在一起后，使用填充材料进行焊球层的填充，使模组形成一个完整的封装整体。

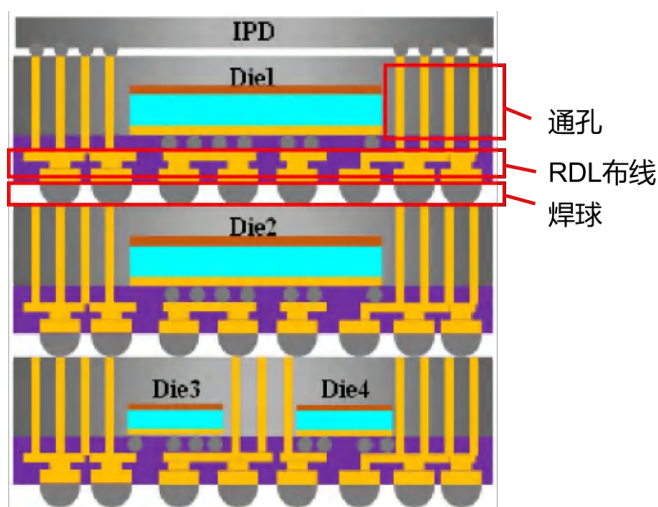
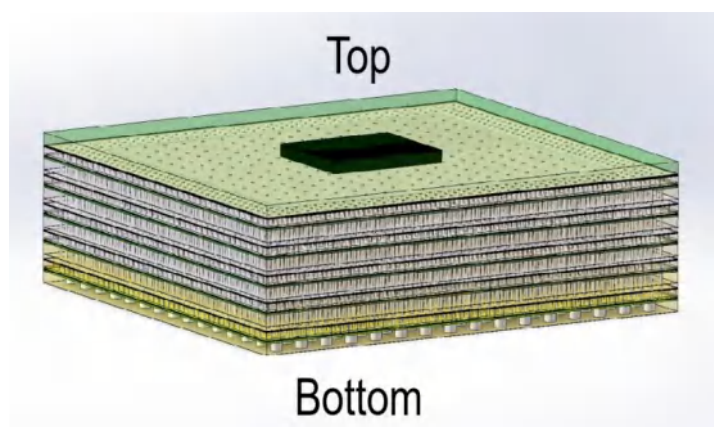


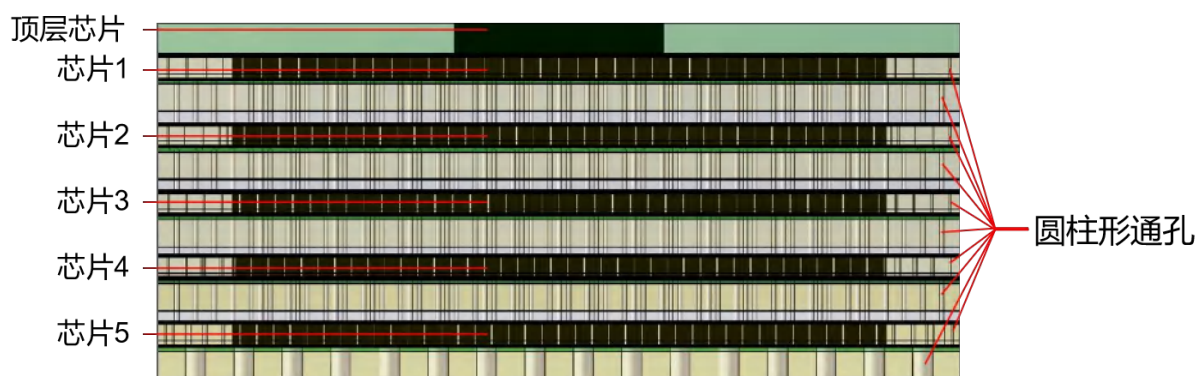
图 2 基于硅转接基板互连的三维系统级封装典型结构示意图

2.2 三维系统级封装Geant4建模

如图3所示，建模采用的三维系统级封装除顶层芯片外还包同型号的5层芯片（芯片1-芯片5），其整体外形尺寸为10.77 mm（长）×9.9 mm（宽）×2.6814 mm。每层硅转接基板设置腔体，芯片放置在腔体中。整个微系统采用了自上而下的立体化逐层建模方式，共分成了67个层结构，多层结构根据材料属性与功能定位可分为两大类型：以导电特性为主的金属层（Metal Layer）和以绝缘介质为主的介电层（Dielectric Layer）等。其中每个封装芯片的相关层包括：转接基板层、芯片层、电源、地、隔离层、焊盘与填充层等。



(a)



(b)

图 3 3D SiP模型示意图：(a)整体结构示意图，(b)正视图

在利用Geant4建模的过程中，本文首先通过由下至上堆叠67层结构搭建出全尺寸模型（10.77 mm×9.9 mm×2.6814 mm），材料属性根据设计文件中对应层的主要功能确定。顶层硅转接基板厚度为0.22 mm，其余芯片层硅转接基板厚度为

0.166 mm，焊盘层的厚度为0.08 mm，二次布线层厚度为0.016 mm。转接基板间焊球直径为90 μm ，底层对外引出焊球直径为400 μm 。应用布尔运算，在每层结构中替换部分结构（如芯片、通孔、焊盘、电源、地、信号等）及其对应的材料属性。3D SiP建模使用材料属性如表1所示。

表 1 3D SiP建模主要材料属性

名称	厚度/mm	材料	密度/ g/cm^3
顶层芯片	0.22	Si	2.33
芯片1-芯片5	0.166	Si	2.33
焊盘	0.08	Cu	8.96
通孔	0.51168	Cu	8.96
材料填充	0.43168	$\text{C}_{11}\text{H}_{12}\text{O}_3$ & SiO_2	1.75

图4为Geant4中所建的3D SiP模型。顶层芯片由于采用了倒装工艺，器件的灵敏体积（Sensitive Volume, SV）区域在下表面（SV0）；对于其余层芯片（芯片1-芯片5），芯片放置在硅转接基板空腔中，灵敏体积在上表面（SV1-SV5）。随着工艺尺寸的缩小，晶体管密度显著增加，因此通常采用平行六面体实体阵列（RPP）方法来评估器件敏感性或计算截面。然而，由于3D SiP内部封装了多个芯片，且全尺寸下单个芯片的晶体管/存储阵列密度极高，传统的晶体管灵敏体积阵列建模方法面临计算量急剧增加的问题，并不适用于三维全尺寸多芯片下的超大晶体管建模与计算。本文采用“片层整体灵敏体积”+“虚拟灵敏体积阵列”的方法来解决这一关键问题，即在每层芯片中采用一个片层状整体来代表灵敏体积阵列，通过建立粒子追踪计算方法，在片层整体灵敏体积中划分网格来设定“虚拟”灵敏体积阵列，追踪载能粒子径迹并统计入网格（即“虚拟”灵敏体积）的能量沉积中，进而分析器件的单粒子效应规律和机制。

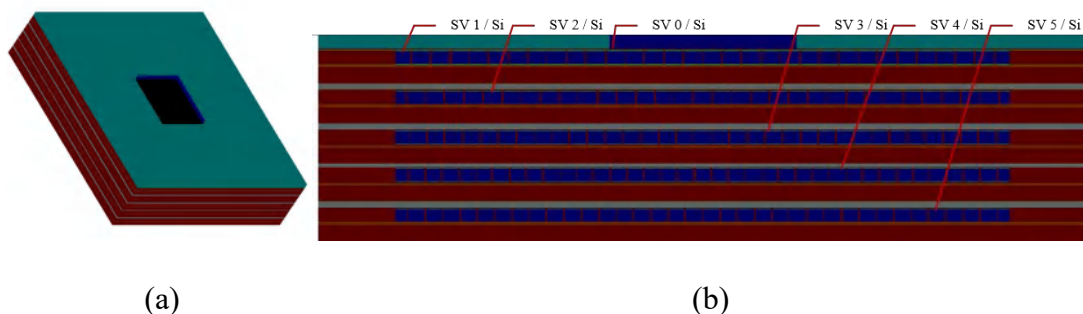


图 4 Geant4中的3D SiP模型示意图：(a)整体结构，(b)灵敏体积层位置

2.3 粒子追踪计算方法

为了处理超大存储规模（Gbits以上）灵敏体积（SV，通常设置为平行六面体）因数量过多导致计算内存占用过大的问题，本方法的核心思路是通过使用“虚拟”SV进行处理，避免创建大量实际的逻辑体或实体，从而降低内存资源需求，并采用程序并行运算提高效率。

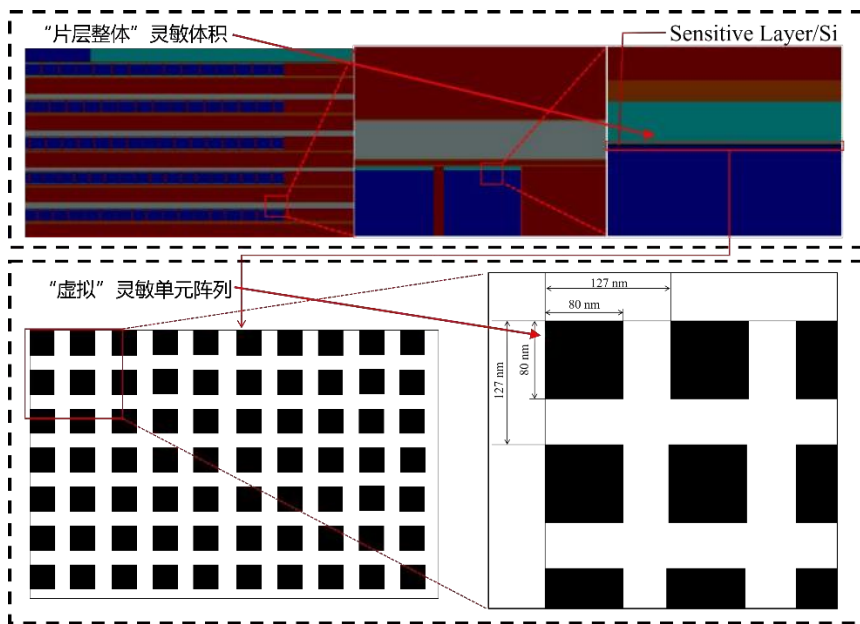


图5 3D SiP底部五层堆叠芯片中单层芯片的“片层整体”灵敏体积和“虚拟”灵敏单元阵列示意图

图5为3D SiP底部五层堆叠芯片中单层芯片的“片层整体”灵敏体积(Sensitive Layer)和“虚拟”灵敏单元阵列示意图。芯片与“片层整体”灵敏体积具有相同的横向尺寸，纵向尺寸由灵敏单元的纵向尺寸确定，根据SRAM芯片平行六面体（RPP）模拟评估方法中采用的纵向深度信息确定。其次，在“片层整体”中

划分虚拟网格，即不通过建立物理体（Physical Volume）形成灵敏体积阵列，而是通过划分的虚拟网络来建立“虚拟”灵敏体积网格阵列，每个网格代表一个灵敏体积单元。例如，对于给定粒子的位置坐标，通过坐标转换和计算，便可确定该粒子是否落在某个虚拟子SV内。需要注意，由于虚拟SV的计算结果严重依赖于用户设置的步长，需要将SV内部步长控制在“虚拟”灵敏单元横向尺寸至少1/5以内。因此，本文将粒子在灵敏层内部的Step最大步长设置为10 nm。表2统计了各层芯片的灵敏体积阵列参数。

表 2 3D SiP建模中芯片的灵敏体积阵列参数

芯片	“片层整体” 尺寸	“虚拟”灵敏单元 尺寸	“虚拟”灵敏单元 间隔	芯片灵敏单元容 量
顶层芯片	2.4 μm*2.8 μm *0.22 μm	0.08 μm*0.34 μm / 0.12 μm*0.26 μm	1.04 μm *0.52 μm	12 Mbits
芯片1-芯片 5	8.77 μm*7.9 μm *0.166 μm	0.08 μm*0.08 μm	0.127 μm*0.127 μm	4 Gbits

最后，追踪“片层整体”中的粒子径迹，并统计“虚拟”灵敏单元中的能量沉积：在粒子模拟的每一步（Step）中，获取粒子所在的逻辑体名称，判断是否为代表“片层整体”的逻辑体。如果是，则进一步获取粒子的位置坐标，通过位置信息判断确定粒子是否在某个“虚拟”灵敏单元内。若在，则统计该步的能量沉积。具体步骤包括：首先，开展粒子所在逻辑体名称判断。在G4UserSteppingAction类的UserSteppingAction方法中，通过GetLogicalVolume()获取粒子所在的逻辑体，然后比较其名称是否为代表“片层整体”的名称。其次，对粒子位置进行判断。当逻辑体名称匹配时，获取粒子的位置信息，通过坐标转换与位置映射规则，判断粒子是否在某个“虚拟”灵敏单元内。最后，进行虚拟子SV中的能量沉积统计。如果粒子在“虚拟”灵敏单元内，则获取该步的能量沉积值（ E_{dep} ），并将其累加到相应“虚拟”灵敏单元的统计变量中，可纳入EventAction类的数组里进行统计。若单个“虚拟”灵敏单元内沉积能量对应的电

荷产生大于预设临界电荷 ($Q_{critical}$)，则判断为单粒子事件。

2.4 仿真流程

本文开展Geant4粒子输运计算，采用重离子入射3D SiP模型，统计了各层芯片中的粒子径迹、能量沉积规律，进而计算单粒子效应截面，分析了单粒子效应在3D SiP内部的空间分布规律。Geant4模拟中参考了兰州近物所加速器的重离子能量范围，针对建立的3D SiP模型，模拟了1.4 GeV和16.8 GeV的Fe离子、31.35 GeV和1.9588 GeV的Bi离子从距离顶层和底层5 mm两处面源正入射的情况，面源的大小为10.77 mm×9.9 mm，设置每次入射1e5个粒子。分析了重离子种类、能量，以及入射方向的影响。

3 模拟结果与讨论

3.1 重离子粒子径迹和能量沉积规律分析

3.1.1 重离子入射模拟

图6展示了16.8 GeV Fe离子从顶层入射3D SiP模型的三维粒子轨迹侧视图，该能量的Fe离子足以穿透该款3D SiP模型，且64.57%的初级粒子进入了底部灵敏层(Sensitive Layer)。同样，16.8 GeV Fe离子底层入射时，进入顶层Sensitive Layer的比例为64.52%，二者相近。31.35 GeV的Bi离子和16.8 GeV的Fe离子入射3D SiP模型后芯片1-芯片5的灵敏层记录到的初级粒子数如图7所示，与Bi (31.35 GeV)相比，随入射深度的增加，Fe (16.8 GeV) 在各灵敏层中的初级粒子变化量更小，且线性特征更明显。

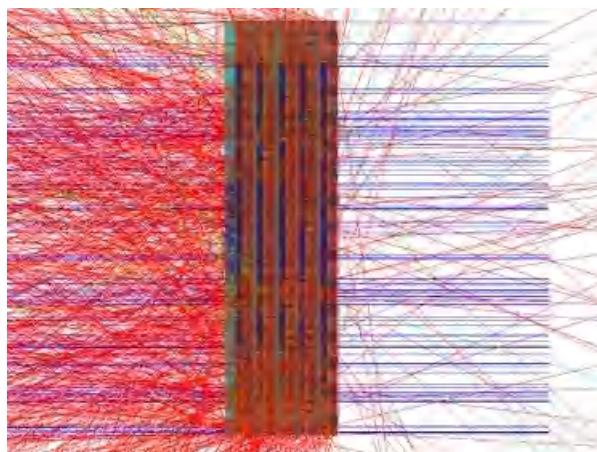


图 6 Fe (16.8 GeV) 离子顶层入射的粒子轨迹侧视图

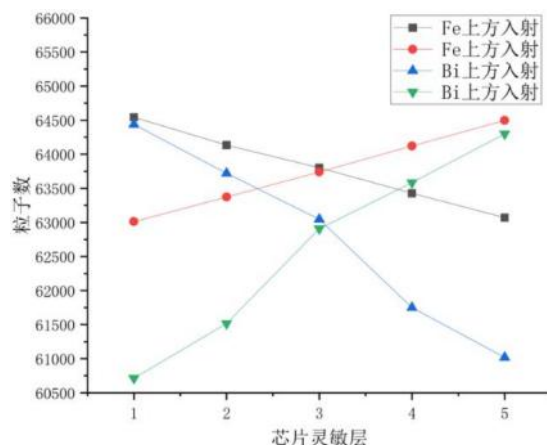


图 7 Fe (16.8 GeV) 与Bi (31.35 GeV) 入射时不同深度的芯片灵敏层 (芯片1-芯片5) 记录到的初级粒子数

在单次模拟中，16.8 GeV的Fe离子入射时芯片1-芯片5灵敏层的能量沉积总量如图8所示，随着入射深度的增加，其造成的能量沉积整体呈增长趋势，但变化幅度较小，在同一入射情况下，各灵敏层记录到的初级粒子数相比最浅层的变化量均未超过最浅层粒子数的3%。1.9588 GeV的Bi离子在3D SiP模型中的射程较短，两种入射情况下均未进入第一灵敏层，未记录到能量沉积。对于1.4 GeV的Fe离子，其初级粒子的穿透距离较短，上方入射时只有64.86%的初级粒子进入上方灵敏层，且仅在芯片0与芯片1中记录到初级粒子。与顶部较薄的Metal Layer相比，3D SiP模型底部的金属层相对更厚，且底部的灵敏体积相比于顶层灵敏单元放置的位置更深，低能Fe离子（1.4 GeV）从下方入射时，难以穿透更厚的金属层并进入放置位置更深的灵敏层，因此在底部灵敏层中未记录到有初级粒子进入。

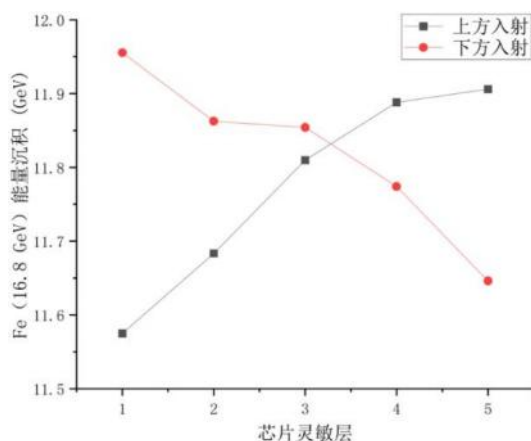


图 8 Fe (16.8 GeV) 入射时不同深度的芯片灵敏层 (芯片1-芯片5) 的能量沉积

3.1.2 重离子能量沉积规律总结

模拟结果表明重离子入射时,次级粒子的能量沉积远小于初级入射粒子的贡献,且对于初级粒子无法到达的芯片敏感区域,次级粒子能量沉积造成的单粒子效应概率极低。在初级粒子的射程范围内,各灵敏层中的能量沉积随灵敏层与粒子源距离的增加而提高。

随着重离子能量的提升,其入射深度也随之增加,粒子在各灵敏层内的能量沉积将呈现一种先急速上升后缓慢下降的趋势:重离子能量低,没有粒子进入灵敏层内,无能量沉积;重离子能量升高,灵敏层内有少量能量沉积;重离子射程达到灵敏层附近,在灵敏层内有大量能量沉积;多数重离子穿透灵敏层,粒子径迹造成能量沉积,可引单粒子效应。

低能Fe (1.4 GeV) 入射深度低,上方入射时多数初级粒子仅能达到芯片1的灵敏区中。在单次模拟中,即使多数16.8 GeV的Fe离子可直接穿透3D SiP模型,但其在整个3D SiP模型中造成的能量沉积总量仍低于低能Fe离子 (1.4 GeV) 在芯片1灵敏区中的沉积。

3.2 单粒子效应在器件中的分布规律分析

3.2.1 单粒子效应截面

粒子在灵敏层内的电离沉积能量全部用于产生电荷,并且这些电荷被完全收集。因此,粒子在单个SV内沉积的能量超过阈值 E_{th} 时,会发生单粒子翻转 (Single Event Upset, SEU), 该阈值 E_{th} 由临界电荷 Q_c 通过以下公式获得:

$$E_{th}[\text{MeV}] = \frac{E_{pair}Q_c}{e} = \frac{Q_c[\text{fC}]}{44.5} \quad (1)$$

式中, e 是单个电子电荷量, E_{pair} 是在材料内生成单个电子-空穴对所需的能量。

根据参考文献^[1,14], 顶层灵敏层SV0中的SVs的 $Q_c=1.1$ fC, $E_{th}=25$ keV (以SRAM型器件估算); 下方五层堆叠灵敏层SV1-SV5的 $Q_c=2.2$ fC, $E_{th}=50$ keV (以DRAM型器件估算)。模拟中粒子在模型正上方或正下方 10.77×9.90 mm范围内

发射100000个Event。因此，粒子注量F为

$$F = \frac{I}{S} = \frac{100000}{10.77 \text{ mm} \times 9.9 \text{ mm}} = 937.8839 \text{ mm}^{-2} = 93788.39 \text{ cm}^{-2} \quad (2)$$

式中，I为一次模拟的入射粒子总数，粒子束I = 100000；S为入射面积。

根据公式，翻转截面 σ 的计算方法为

$$\sigma = \frac{N}{F} \quad (3)$$

式中，N为各灵敏层内出现翻转的SV数目；F为计算所得的粒子注量。据公式计算得到不同重离子入射时的芯片1-芯片5的单粒子效应截面，如图9所示。

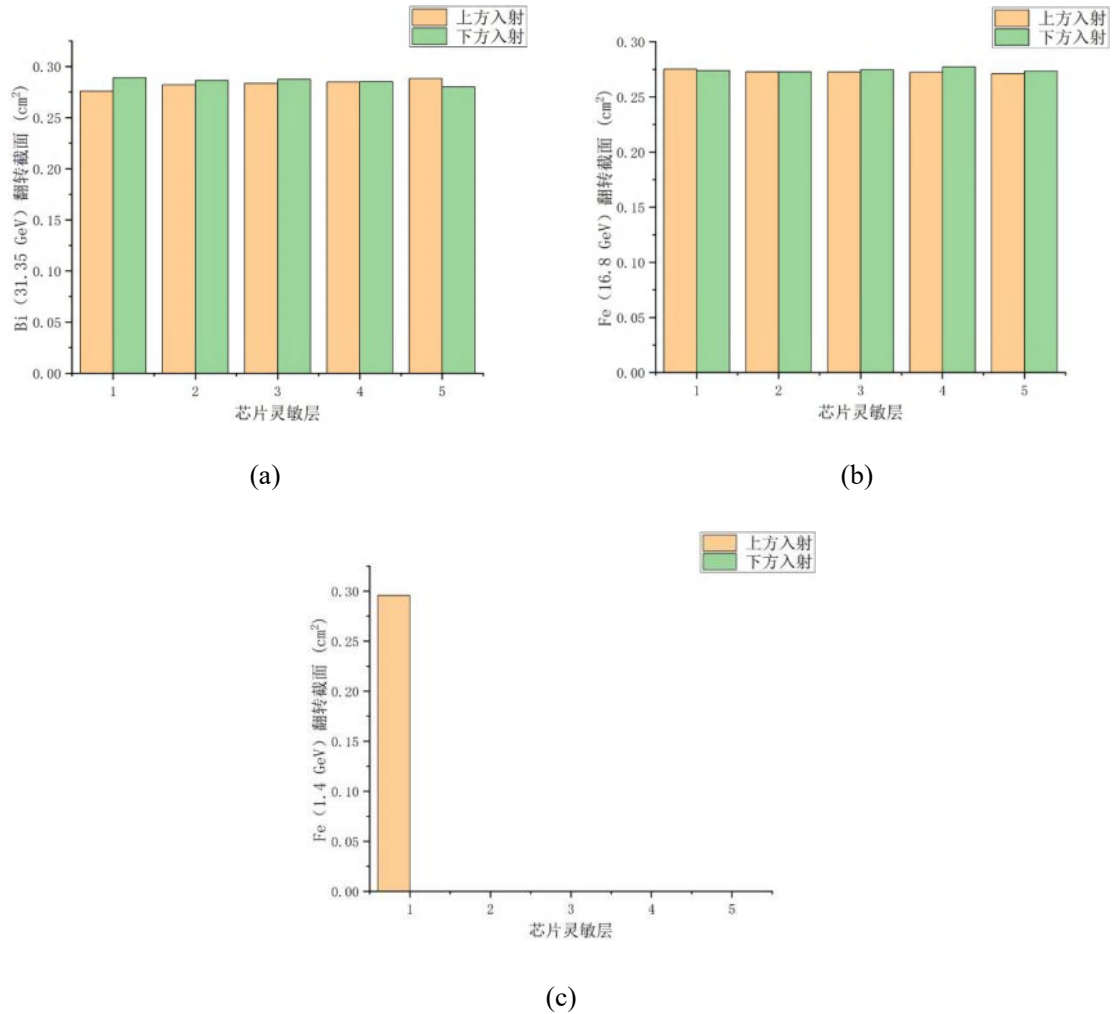


图 9 不同重离子入射时各芯片灵敏层（芯片1-芯片5）的翻转截面统计：(a)Bi(31.35 GeV)，(b)Fe(16.8 GeV)，(c)Fe(1.4 GeV)

3.2.2 单粒子效应分布规律分析

图 10、图 11 分别展示了 Fe (16.8 GeV) 入射时各灵敏层内翻转截面与记录到的初级粒子总数、初级粒子进入 SV 的几率趋势比较，观察到翻转截面在一个区间内随深度变化的涨跌趋势并不明显。对于进入各灵敏层的初级粒子数和各灵敏层内初级粒子进入 SV 的概率，二者的变化趋势均与重离子在不同芯片灵敏层的翻转截面变化趋势相似，因此推断翻转截面和粒子初始能量以及重离子在 3D SiP 中受影响偏转有关。

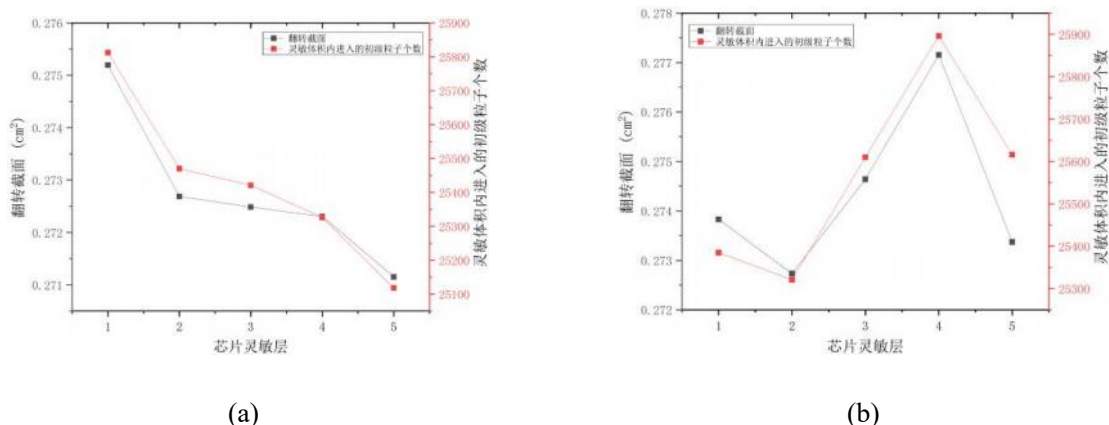


图 10 Fe (16.8 GeV) 粒子入射时各芯片灵敏层 (芯片1-芯片5) 的翻转截面和进入各灵敏层的初级粒子数量变化趋势比较: (a)上方入射, (b)下方入射

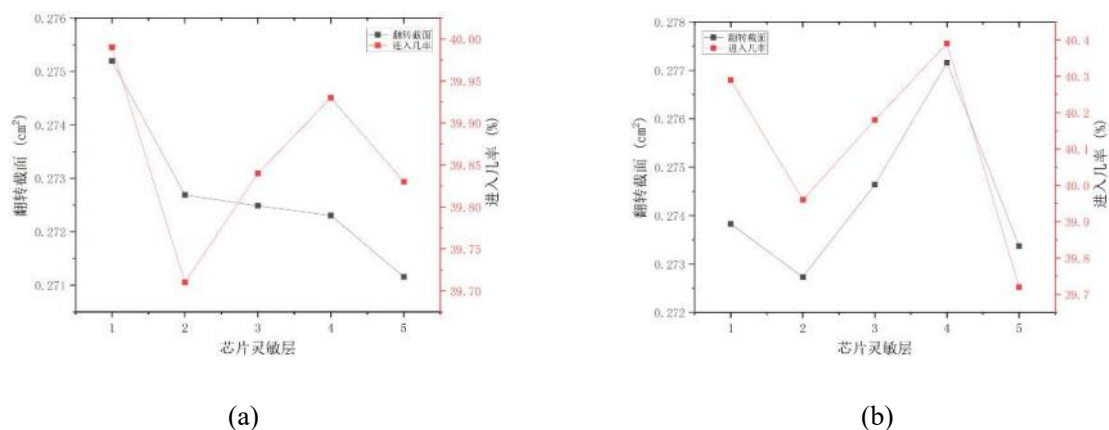


图 11 Fe (16.8 GeV) 粒子入射时各芯片灵敏层 (芯片1-芯片5) 的翻转截面和灵敏层内初级粒子进入SV的概率变化趋势比较: (a)上方入射, (b)下方入射

本文还对重离子入射时的单粒子效应翻转类型进行了预估。在Fe离子 (16.8 GeV和1.4 GeV) 与Bi离子 (31.35 GeV) 入射模拟中, 约60%的粒子Event会引发

单粒子翻转。单层芯片中，单粒子翻转类型基本均为单比特翻转（Single Bit Upset, SBU），存在少量单个粒子Event在同一层级造成多比特翻转（Multiple Bit Upset, MBU）的情况，但发生概率极低，可忽略不计。图12展示了两种重离子入射情况下造成单粒子翻转的Event中，引发单芯片翻转（Single Chip Upset, SCU）或两层、三层等多芯片翻转（Multiple Chip Upset, MCU）的Event所占百分比，部分入射情况下重离子无法穿透模型至第一灵敏层，因此不存在单粒子事件。对于16.8 GeV的Fe离子与31.35 GeV的Bi离子，粒子Event引发的单粒子翻转类型主要为SCU与跨两层的MCU，且二者所占比例之和超过60%。跨五层的MCU发生概率极低，仅占有所有Event的约1%。对于1.4 GeV的Fe离子，其上方入射时绝大部分能量沉积在SV1层级，因此造成的单粒子翻转多发生在芯片1灵敏层，且基本为SCU。

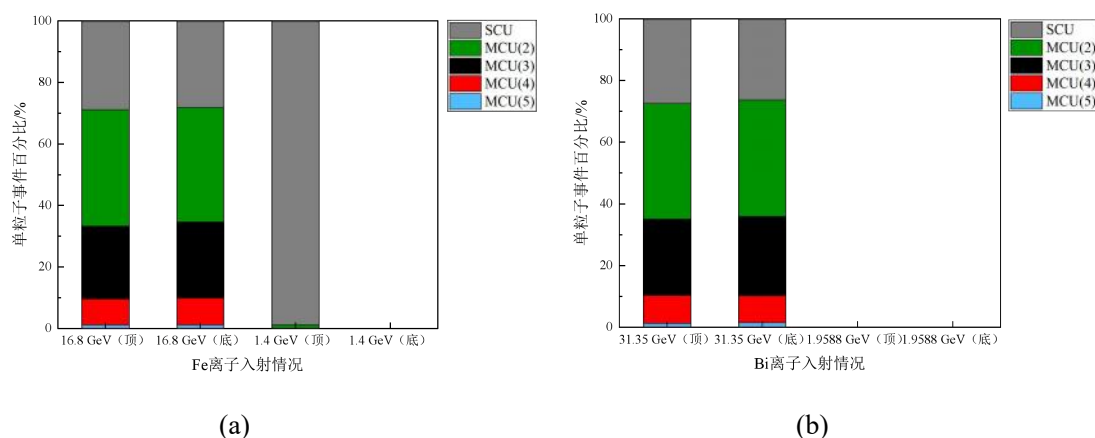


图 12 不同重离子入射情况下，不同类型的单粒子事件百分比：(a)Fe离子，(b)Bi离子

3.2.3 单粒子效应分布规律总结

重离子入射 3D SiP 时，若重离子在射程内灵敏层中沉积的能量超过该灵敏区的阈值 E_{th} ，便会发生单粒子翻转。在 Fe 离子（16.8 GeV 和 1.4 GeV）与 Bi 离子（31.35 GeV）入射模拟中，约 60%的 Event 会引发 SCU，且在单层芯片中，由单个 Event 造成的单粒子翻转类型基本均为单比特翻转；16.8 GeV 的 Fe 离子与 31.35 GeV 的 Bi 离子 Event 引发的单粒子翻转类型主要为 SCU 与跨两层的 MCU，跨五层的 MCU 仅占约 1%；1.9588 GeV 的 Bi 离子无法穿透模型至第一

灵敏层，因此无法造成单粒子翻转。

3D SiP 翻转截面强烈依赖于重离子粒子初始能量和在器件中的径迹。对于进入各灵敏层的初级粒子数和各灵敏层内初级粒子进入 SV 的概率，二者的变化趋势均与重离子在不同芯片灵敏层的翻转截面变化趋势相似。

4 总结

本文针对 3D SiP 在重离子辐射环境下面临的单粒子效应问题，开展了三维系统级封装的单粒子效应 Geant4 模拟方法研究，建立了 3D SiP 器件的全尺寸 Geant4 模型、虚拟灵敏体积阵列与粒子追踪联合计算方法，并进行了 3D SiP 模型的重离子入射模拟，统计了粒子径迹和能量沉积以及单粒子效应在封装内的空间分布规律。模拟结果表明，随着重离子能量的提升，粒子在各灵敏层内的能量沉积将呈现一种先急速上升后缓慢下降的趋势，初级粒子射程范围内灵敏层的能量沉积随入射深度增加而升高，而次级粒子整体造成的能量沉积随深度衰减。在 Fe 离子（16.8 GeV 和 1.4 GeV）与 Bi 离子（31.35 GeV）入射模拟中，约 60% 的重离子 Event 会造成单粒子翻转。在单层芯片中，由单个 Event 基本只会引发单比特翻转，多比特翻转发生的概率极低；高能重离子（如 16.8 GeV 的 Fe 离子与 31.35 GeV 的 Bi 离子）Event 引发的单粒子翻转类型主要为 SCU 与跨两层的 MCU，跨五层的 MCU 仅占约 1%；低能重离子（如 1.9588 GeV 的 Bi 离子）若无法穿透模型至第一灵敏层，则无法造成单粒子翻转。对于 3D SiP，单粒子多芯片错误将不可避免。

本文建立了 3D SiP 的全尺寸 Geant4 模型及单粒子效应分析方法，为揭示 3D SiP 的重离子辐射效应评估提供了一定的参考。未来工作可考虑进一步精细化模型、完善灵敏体积结构，并开展更多能量与种类的粒子入射模拟，以完善 3D SiP 器件的辐射可靠性评估体系。

参考文献

- [1] Li Y , Yoshida M , Gomi Y ,et al. Experimental Study of Proton-Induced Radiation Effects on DDR5 Modules[J]. IEEE Transactions on Nuclear Science, 72[2025-06-20].
- [2] 贺朝会, 陈伟, 韩建伟, 等.新型微系统的辐射效应与抗辐射加固技术[J].中国科学:物理学 力学 天文学, 2024, 54(03):5-22.
- [3] Wang H, Ma J, Yang Y, et al. A Review of System-in-Package Technologies: Application and Reliability of Advanced Packaging [J]. Micromachines, 2023, 14(6): 1149.
- [4] NASA. The NASA electronic parts and packaging (NEPP) program: roadmap for FY15 and beyond and recent radiation highlights[EB/OL]. [2015-09]. <https://ntrs.nasa.gov/api/citations/20150023279/downloads/20150023279.pdf>.
- [5] NASA. Radiation effects and analysis lessons: a scientist's field instruction to explain radiation testing[EB/OL].[2022-04]. https://ntrs.nasa.gov/api/citations/20220002512/downloads/Wyrwas-Rad-Effects-Testing-Primer_7f.pdf.
- [6] Zheng S , Zhang Z , Ye J ,et al.Experimental Study of the Impact of Temperature on Atmospheric Neutron-Induced Single Event Upsets in 28 nm Embedded SRAM of SiP[J].Electronics (2079-9292), 2024, 13(11): 2012.
- [7] Rajkowski T, Saigne F, Pouget V, et al. Analysis of SET propagation in a system in package point of load converter[J]. IEEE Transactions on Nuclear Science, 2020, 67: 1494–1502.
- [8] 田佳玉, 曹荣幸, 张涛, 等.典型SiP微系统器件质子单粒子效应研究[C]. 2024第六届全国辐射物理学术交流会. 中国核学会, 2024: 16.
- [9] Zhang G, Yang Z, Li X, et al. Gamma-ray irradiation induced dielectric loss of SiO₂/Si heterostructures in through-silicon vias (TSVs) by forming border traps[J]. ACS Applied Electronic Materials, 2024, 6(2): 1339-1346.
- [10] Li K, Zhang E X, Gorchichko M, et al. Impacts of through-silicon vias on total-ionizing-dose effects and low-frequency noise in FinFETs[J]. IEEE Transactions on Nuclear Science, 2021, 68(5): 740-747.
- [11] Li Y, Li J, Guo Y, et al. Experimental study of transient dose rate effect on system-in-package SZ0501[J]. IEEE Transactions on Nuclear Science, 2022, 69(8): 1840-1849.
- [12] Li Y, Guo Y, Li J, et al. Transient dose rate effect between system-in-package and printed circuit boards: a comparative experimental study[J]. IEEE Transactions on Nuclear Science, 2023, 70(8): 2106-2115.

- [13] Li N, Li Y, Guo Y, et al. Simulation analysis of electromagnetic pulse susceptibility and hardening design for system-in-package SZ0501[J]. *Microelectronics Reliability*, 2023, 141: 114892.
- [14] DENG, Y. Study of Single Event Upset Phenomena Induced by Muons and Protons in 65-nm Planar Bulk SRAMs[EB/OL]. Kyushu University Institutional Repository. [2024-02]. https://catalog.lib.kyushu-u.ac.jp/opac_download_md/7182513/tj1318.pdf.

面向芯粒尺度的多物理场耦合研究

刘娜, 吴秋月, 乔守宇, 蔡国雄

(厦门大学 电子科学与技术学院, 厦门 361000)

摘要: 随着 Chiplet 技术的发展, 高密度集成引发的热串扰、热应力和电磁问题日益突出。本研究针对 2.5D/3D 封装中 TSV、bump 等微小结构, 建立等效模型并结合区域分解技术, 开发了基于时域有限元法的瞬态电热多物理场仿真方法。相比 COMSOL, 计算效率提升 3 倍, 可精确分析芯片布局、材料及散热方案对热分布的影响。研究成果为高性能 Chiplet 设计提供了高效的多物理场仿真工具, 有助于优化散热方案并提升可靠性。

关键词: 多物理场; 区域分解法; 有限元法; Chiplet; 2.5D/3D 封装

1 引言

集成芯片通过将特定功能的芯粒集成制造而成。随着晶体管尺寸逼近物理极限, Chiplet 技术通过模块化设计有效突破传统单片集成电路的限制, 成为重要发展方向[1]。先进封装是 Chiplet 技术发展的基础, 但随着集成度提高, 散热和热应力问题日益突出, 高功率密度影响性能与寿命, 热应力可能导致结构失效, 高温还会损害信号完整性[2][3]。Chiplet 的电-热-应力多物理效应严重影响系统稳定性, 因此, 电-热一体化仿真成为关键技术, 对优化设计至关重要。

Chiplet 可靠性问题的电-热分析主要是为了得到模型中电场、温度场的分布以及它们之间的耦合关系。多物理场的仿真主要包括电场、热场分析及它们之间的耦合关系。目前主要的计算方法有: 时域有限差分法、部分等效电路法、时域有限元法、边界元法等[4]-[8]。其中, 有限元法在多物理场数值模拟领域中的研究和应用广泛, 是应用较为普遍的数值模拟方法。国内外学者和企业用电-热算法领域已开展大量的研究[9]。如国内上海交通大学毛军发院士课题组提出 LB-ADI 法, 进行了 Chiplet 芯片 2.5D 封装和 3D 封装的瞬态热仿真, 便于快速得到 Chiplet 芯片的热分布[10]。美国伊利诺伊大学香槟校区 Jin 教授课题组在有限元法的基础上组合运用基于比例-积分-微分(Proportional-Integral-Derivative Control, PID)控制算法的自适用时间步进和区域分解(Finite Element Tearing and Interconnecting, FETI)法完成了高功率射频/微波器件的电热联合仿真, 进行了集成全波电磁分析和瞬态热分析[11]。此外, Jin 教授课题组还开发了有限元区域分解方法仿真大规模 3D 集成电路键合线多物理场特性[12]。

尽管三维有限元法在处理复杂电路多物理场问题方面具有显著优势, 但在实

际应用中仍面临严峻挑战。高密度三维集成电路中多尺度存在，且电路边界条件复杂以及多物理场耦合求解未知量太多，计算速度太慢，满足不了实际快速仿真设计的需求。因此，发展快速收敛、高精度有限元算法无疑是十分重要。本工作开展电-热多物理场耦合中高精度区域分解有限元法算法研究，为Chiplet芯片的布局分析提供更高效率的解决方案。

2 区域分解有限元方法

本工作基于电流连续性方程和瞬态热传导方程，通过建立电-热耦合本构关系及考虑材料参数的温度依赖性，构建了集成电路多物理场仿真模型。采用有限元法与区域分解法相结合的方法，为集成电路电-热可靠性分析提供了新的理论和技术支撑。电-热方程和耦合本构关系如下：

$$\text{电流连续性方程： } \nabla \cdot (\boldsymbol{\sigma}^{(T)} \mathbf{E} + \epsilon \dot{\mathbf{E}}) = 0 \quad (1)$$

$$\text{热传导方程： } \dot{\mathbf{T}} = \frac{1}{\rho^{(T)} c_p^{(T)}} \nabla \cdot (\mathbf{K}^{(T)} \nabla \mathbf{T}) + \mathbf{q} \quad (2)$$

$$\text{电-热耦合关系： } \mathbf{q} = \boldsymbol{\sigma}^{(T)} \mathbf{E}^2 \quad (3)$$

电流脉冲在电导率 $\boldsymbol{\sigma}$ 和介电常数 ϵ 中产生电场 \mathbf{E} 和热源项 \mathbf{q} 。热源 \mathbf{q} 在比热容 c_p 和各向异性导热系数 \mathbf{K} 下引起温度场的扩展，温度 \mathbf{T} 的变化使 $\boldsymbol{\sigma}$ 、 \mathbf{K} 等参数跟随温度的变化。运用数值仿真算法求解前，需要对模型网格剖分，而正则的网格质量是获得一定数值精度的保证。区域分解算法可以将不同材料、不同模块之间灵活处理，有效减少整个大规模系统产生。在处理区域分解界面时，电场、热场我们将运用如惩罚因子等不同的数值通量保持界面场的连续性条件，进而保证算法的精度及稳定性。

3 数值结果与分析

3.1 基于 2.5D 封装的 Chiplet 电热耦合仿真

针对 Chiplet 结构复杂、材料特性差异大的特点，本研究选取典型复杂芯粒结构，分析不同介质材料特性，建立电路理论模型和边界条件。基于电-热耦合算法开展多物理场仿真，研究材料热导率、功率分布等因素对电热行为的影响。通过参数化仿真分析芯片布局。图 1 给出了基于 2.5D 封装的 4 芯粒 Chiplet 模型的两片芯片布局方案，对比研究了芯片布局与加电位置之间的影响。

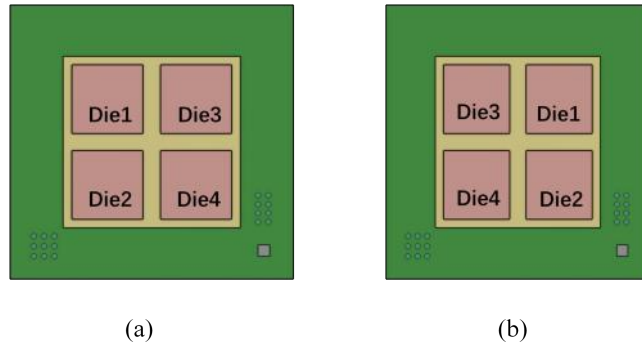


图 1 不同 Chiplet 芯片布局方案 (a) 布局 1 (b) 布局 2 (Die1, Die2 为恒定低功耗, Die1、Die4 为脉冲高功耗, 布局 1 为高功耗芯粒靠近加电位置, 布局 2 为高功耗芯粒远离加电位置)

图 2 为两种布局下加 1.75V 电压后电热耦合的芯片热分布, 由图可知, 布局 1 时, 峰值功率最大的 Die 4 离加电位置最近, 在 2.5 s 时最高温为 367.754 K。而在布局 2 中, 热点随芯粒 4 位置迁移, 最高温为 366.577 K, 下降了 1.177 K。由图 3 可观察到两种芯片布局下截点温度随时间的变化曲线, 在 3 s 之后, 布局 (a) 的最高温度明显低于布局 (b)。结果表明, 在 Chiplet 系统设计中, 高功率的芯粒应远离加电位置, 有助于降低整个系统的最高温度。

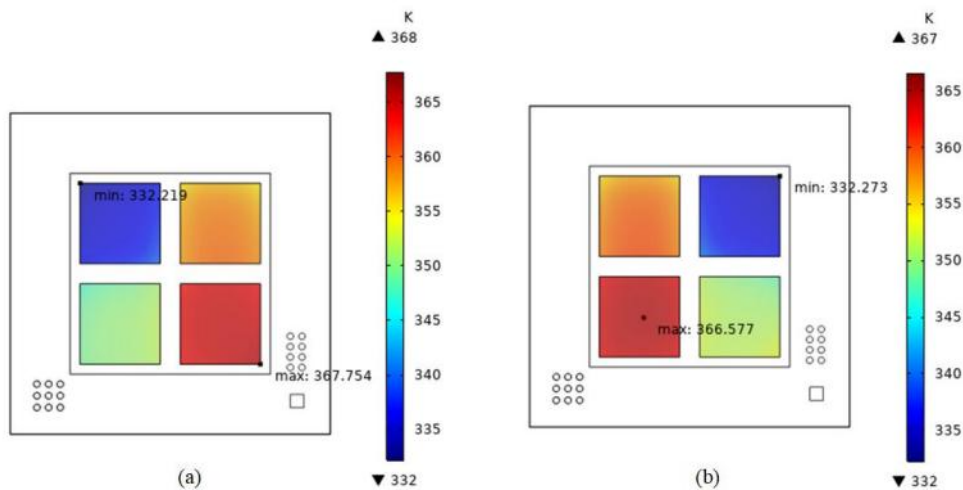


图 2 2.5D 封装 4 芯粒 Chiplet 模型的两种不同芯粒布局下温度分布: (a)高功率芯粒靠近加电位置的布局 (b)高功率芯粒远离加电位置的布局

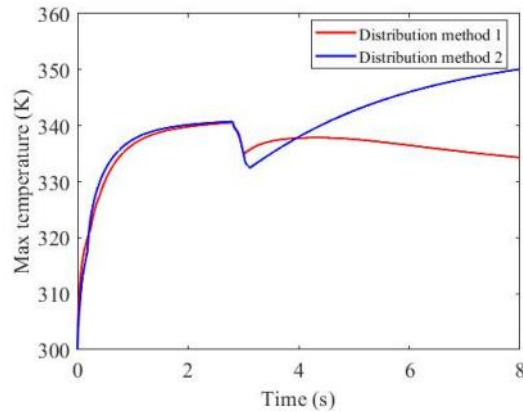


图 3 2.5D 封装 4 芯粒 Chiplet 模型的两两种芯粒布局下点(0,0,0.15) mm 温度随时间的变化

表 1 中给出电热耦合新算法和 COMSOL Multiphysics 所消耗的 CPU 时间和自由度对比,可以看出在自由度基本一致的情况下, CPU 的时间成本显著降低,采用基于区域分解算法的计算时间仅用了 192 s,对比于商业软件 COMSOL 计算效率提高了 3.3 倍。

表 1 2.5D 封装 4 芯粒 Chiplet 模型在提出算法和 COMSOL 中的自由度和 CPU 时间

方法	自由度	CPU 时间(s)
COMSOL	431,177(电 184,458, 热 246,719)	649
区域分解算法	415,787(电 176,552, 热 239,235)	192

3.2 不同芯粒间距 3D Chiplet 的电热耦合仿真

为进一步探究三维堆叠构型下的热传导特性,研究了建立了 3D 封装的 11 芯粒 Chiplet 模型多物理场耦合模型。芯粒模块包括芯粒、热界面材料和凸点层,将芯粒模块间距设置以 1 mm 的步长从 0 mm 增加到 4 mm,具体如图 4 所示,图中黄色部分为所设定的芯粒模块。图 5 示出了随所间距变化的芯粒的结温,可见随着芯粒间距增大,结温逐渐下降。这是由于芯粒间距会影响芯粒之间的热串扰,随着间距减小,热串扰效应会增强,导致芯粒温度升高。Chiplet 电路集成度高,若芯粒分布过于密集,会引起芯粒间强热串扰效应,进一步导致热点问题。短间距芯粒布局将为多个芯粒集成带来散热挑战,需要进行合理的芯粒间距布局设计规划。

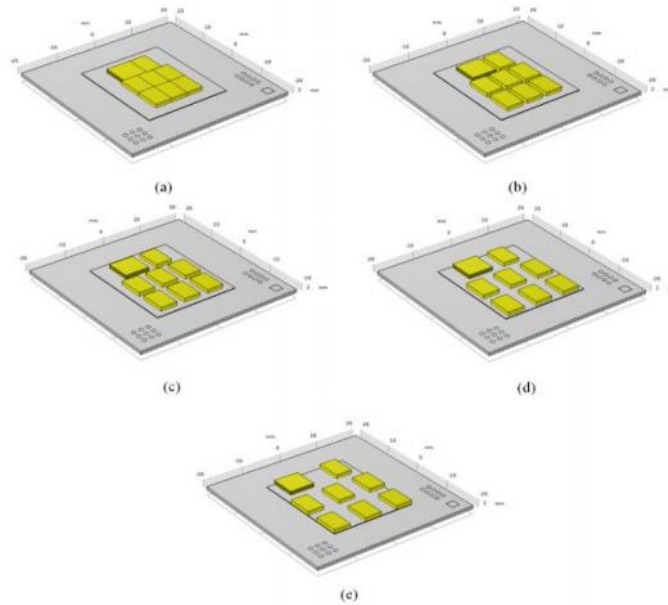


图 4 不同芯粒间距的 3D 封装 11 芯粒 Chiplet 模型:(a) 0 mm (b) 1 mm (c) 2 mm (d) 3 mm (e) 4 mm (黄色部分为所设定的芯粒模块, 包括芯粒、热界面材料和凸点层)

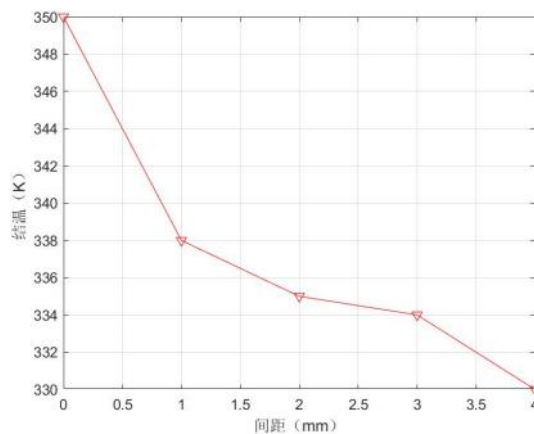


图 5 3D 封装 11 芯粒 Chiplet 模型的芯粒间距对结温的影响

4 总结

本工作针对 2.5D/3D 封装的 Chiplet 集成电路开发了基于区域分解的时域有限元多物理场仿真方法, 通过建立等效电路模型将电磁问题简化为泊松方程求解, 并结合瞬态热传导模型实现了高效电热耦合分析。研究系统考察了芯粒布局、热界面材料对热分布的影响, 优化了多芯粒系统散热方案。结果表明所提算法计算效率显著优于传统商业软件, 适用于复杂封装结构的快速仿真。研究成果为高密度 Chiplet 集成电路提供了高效的多物理场仿真框架, 可扩展应用于不同芯片结构, 随着 3D 集成技术的发展, 该方法将为优化散热设计和封装策略提供重要支

撑，为下一代高集成度、高性能计算芯片的开发奠定技术基础。

参考文献

- [1] A. Shanmugavel, Robust Design for Integrated Circuits, *ANSYS Advantage*, 2013.
- [2] H. H. Zhang, Z. L. Jia, P. F. Zhang, Y. Liu, L. J. Jiang, and D. Z. Ding, Electromagnetic-Circuit-Thermal-Mechanical Multiphysics Numerical Simulation Method for Microwave Circuits, *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, vol. 9, pp. 129-141, 2024.
- [3] C. J. Nie, Q. Z. Xu, C. H. Wang, H. Cao, J. Y. Liu, and Z. Q. Li, Efficient Transient Thermal Analysis of Chiplet Heterogeneous Integration, *Applied Thermal Engineering*, vol. 229, 2023.
- [4] E. L. Tan, D. Y. Heh, Stability Analyses of Nonuniform Time-Step LOD-FDTD Methods for Electromagnetic and Thermal Simulations, *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, 2017, 2: 183-193.
- [5] L. Jiang, Y. Liu, and M. C. Cheng, Fast-Accurate Full-Chip Dynamic Thermal Simulation with Fine Resolution Enabled by a Learning Method, *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, vol. 42, no 8, pp. 2675-2688, 2023.
- [6] C. Wang, D. G. Yang, J. M. Lyu, Y. Dai, C. Zhuo, and Q. Chen, On Model Order Reduction and Exponential Integrator for Transient Circuit Simulation, *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, vol. 43, no. 1, pp. 328-339, 2024.
- [7] Q. K. Liu, Z. Y. Mo, A. Q. Zhang and Z. Yang, JAUMIN: A Programming Framework for Large-scale Numerical Simulation on Unstructured Meshes, *CCF Transactions on High Performance Computing volume*, vol. no. 1, pp. 35-48, 2019.
- [8] H. H. Zhang, P. P. Wang, L. J. Jiang, W. E. I. Sha, M. S. Tong, Y. Liu, W. J. Wu, G. M. Shi, Parallel Higher Ord[12]S. Zuo, Z. Lin, D. García-Doñoro, Y. Zhang and X. W. Zhao, A Parallel Direct Domain Decomposition Solver Based on Schur Complement for Electromagnetic Finite Element Analysis, *IEEE Antennas and Wireless Propagation Letters*, vol. 20, no. 4, pp. 458-462, 2021.
- [9] 丁大志, 成爱强, 王林, 张天成, 陈如山, 高功率微波效应的多物理场建模方法研究, *电波科学学报*, 2020, 35(1):93-105.
- [10] J. Li, M. Tang, L. S. Wu, L. G. Jiang, W. L. Dai, and J. F. Mao, LB-ADI: An Efficient Method for Transient Thermal Simulation of Integrated Chiplets and Packages[J], *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, vol. 9, pp. 149-156, 2024
- [11] T. J. Lu, J. M. Jin, Electrical-Thermal Co-Simulation for Analysis of High-Power RF/Microwave Components, *IEEE Transactions on Electromagnetic Compatibility*, vol. 59, no. 1, pp. 93-102, 2017.
- [12] T. Lu, and J. M. Jin, Coupled Electrical-Thermal-Mechanical Simulation for the Reliability Analysis of Large-Scale 3-D Interconnects, *IEEE Transactions on Components Packaging and Manufacturing Technology*, 2017, 7(2): 229-237.

低温硅基器件热点邻域的多尺度建模与声子输运机制研究

贾如¹, 张新宇¹, 盛宇飞¹, 徐家璇¹, 夏涌林¹, 鲍华^{1,*}

¹上海交通大学溥渊未来技术学院, 上海市, 200240, 中国

摘要: 低温硅基器件中的热点温升较常温更严重。宏观传热方程难以描述低温复杂声子输运。本文结合第一性原理计算与声子玻尔兹曼输运方程开展多尺度建模, 系统研究低温硅器件热点处的声子输运机制。结果表明, 弹道声子、温度依赖的声子热物性及热点尺寸是关键因素。低温下, 弹道声子主导硅基区热输运, 热物性温度依赖导致非线性温升, 且热点尺寸对热阻影响显著增强。本文系统解析了低温热点声子输运机理, 为低温硅器件性能优化提供参考。

关键词: 低温硅基器件; 自热效应; 声子玻尔兹曼方程; 多尺度建模

0 前言

低温硅基 CMOS 器件在量子计算、天文探测与空间探索等领域具有广泛应用前景[1-3]。在低温环境下, 器件热点会引发比常温更显著的温升, 显著影响其电学性能[4, 5]。因此, 深入理解低温下热点区域的热输运机制, 对于提升器件性能具有重要意义[4]。

现有研究主要聚焦于低温热输运的实验测量, 通常通过金属电阻的温度敏感性来测量温升, 并提取热阻以表征热点行为[5]。实验表明, 温升与功率输入呈非线性关系, 且低温下的热阻远高于室温[6, 7]。这些现象通常采用温度依赖热导率或包含多个拟合参数的紧凑热阻模型加以解释[6-9]。然而, 此类有效热导率多来源于薄膜实验, 难以揭示具体器件结构中的微观声子散射机制。在这种情况下, 热输运更适通过声子玻尔兹曼输运方程(phonon Boltzmann Transport Equation, BTE)描述[10]。目前, 多数声子 BTE 求解器仅适用于室温附近, 其为简化计算而忽略了声子性质的温度依赖性[11-14]。因此, 为了能够准确刻画低温下非线性声子输运行为, 需要考虑非线性温度依赖的声子热物性。

本研究结合第一性原理计算与非线性温度依赖的声子玻尔兹曼输运方程, 系统分析了低温器件热点邻域的声子输运特性。首先揭示了低温下功率密度与温升之间非线性关系的来源, 并进一步探讨了热点尺寸对热阻的影响, 解释了已有实验中观测到的现象。

1 方法

在器件热仿真中，最常用的方法是基于傅里叶定律的热扩散方程建模，该方法在宏观尺度下具有良好适用性[15]。然而，随着器件尺寸不断缩小，特征尺度 L 逐渐接近甚至小于声子平均自由程 MFP[15, 16]，热输运行为逐渐从扩散（MFP $\ll L$ ）过渡至弹道输运（MFP $\gg L$ ）。特别是在低温条件下，由于声子平均自由程显著增加，弹道声子输运效应会更显著[17]，其导致的局域温升会直接影响器件的电学性能[10]。为准确捕捉微纳尺度下的声子输运行为，应采用声子玻尔兹曼输运方程进行建模，该方法已被广泛用于纳米尺度器件的自热效应[11, 12]。在小温差近似条件下，稳态声子 BTE 形式如下[18]：

$$\mathbf{v} \cdot \nabla e = -\frac{e - e^0}{\tau} + \dot{q}, \quad (1)$$

其中， \mathbf{v} 为声子群速度， τ 为声子弛豫时间， e 为声子能量分布函数， e^0 为满足玻色-爱因斯坦分布的平衡能量分布， \dot{q} 为体热源项。等式左侧为漂移项，右侧包含散射项与热源项[18]。在求解 BTE 时，为简化计算，通常采用小温差近似（ $\Delta T \ll T$ ），即将平衡态能量分布函数线性化为温度的函数（ $e^0 = C_p T$ ），并假设声子热容、群速度与弛豫时间（ C_p, τ, \mathbf{v} ）为常数[19]。在低温下求解声子 BTE 面临两项主要挑战。第一，缺乏高精度、模式分辨的声子参数数据，原因在于第一性原理计算代价较高。为此，我们在前期研究中开发了加速的第一性原理计算方法，可以获得热容、群速度和弛豫时间等声子输入参数[20]。第二，声子参数在低温下具有强烈的温度依赖性[13, 14]，此时需考虑平衡能量的非线性项与温度依赖的物性参数。对应的非线性 BTE 表达式为：

$$\mathbf{v}(T) \cdot \nabla e = -\frac{e - e^0}{\tau(T)} + \dot{q}, \quad (2)$$

其中，平衡能量项的非线性以 $e^0(T) = \int_0^T C_p(T) dT$ 表示。

2 结果分析

1.1 低温下非线性温升功率关系

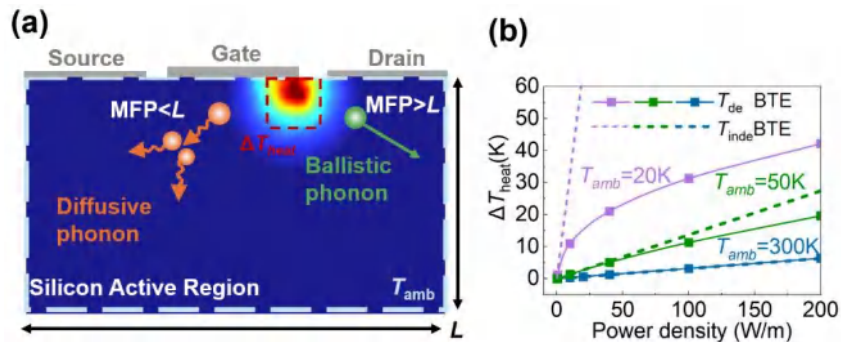


图 1 (a) 硅基MOSFET中有源区的结构示意图；(b) 不同功率密度下热点区域的平均温升，环境温度分别为20K、50K和300K。仿真区域中边长设为 $L=1\mu\text{m}$ 。

尽管已有实验观察到低温下的自热现象，其背后的声子输运机制仍缺乏深入理解。由于热点通常形成于漏极-沟道结附近，本文将分析重点放在图 1 (a) 所示的硅有源区内的声子输运[5,6]。仿真区域设置为边长为 L 的正方形硅结构。在低温条件下 ($L \ll \text{MFP}$)，我们已验证声子输运对 L 的变化 ($0.5\text{--}10\mu\text{m}$) 并不敏感。对于硅 MOSFET 的仿真结构，其顶部边界设置为漫反射绝热边界，其余三个边界保持在恒定环境温度，以考虑热量向衬底的耗散[5]。

本研究中自热效应的定量表征与热阻的定义参考已有实验研究[6, 14, 15]。自热引起的温升通过热点区域内的平均温度升高 $\Delta T_{\text{heat}} = T_{\text{heat}} - T_{\text{amb}}$ 来描述，该方法与实验中采用的栅电阻测温方式一致 [16]。同时，热阻定义为 $R^* = \partial \Delta T_{\text{heat}} / \partial P|_{T_{\text{amb}}}$ ，其中 P 为单位栅宽的功率密度。热点的边长 L_{heat} 设为 10nm，接近实际器件中焦耳加热区域的典型尺寸[17]。

为进一步探究自热背后的声子输运机制，我们分析了弹道声子输运与温度依赖的声子热物性的影响。具体而言，比较了使用温度依赖与非温度依赖声子参数所得的温升结果，分别对应式 (1) 与式 (2)。图 1 (b) 中用虚线与点划线分别表示两种结果。结果显示，温升非线性仅在考虑温度依赖声子参数的低温仿真中出现，而该非线性在环境温度升高后逐渐减弱，与实验观察一致[4]。这表明低温下的非线性温升主要来源于声子热物性的强温度依赖性。另一方面，图 1 (b) 中无论是否考虑声子参数的温度依赖性，仿真均显示环境温度降低时温升会显著增加。这说明低温下增强的自热效应主要是来自于弹道声子的贡献增强 ($\text{MFP} \gg L$)，而非声子物性的温度依赖性。

1.2 不同热点尺寸的影响

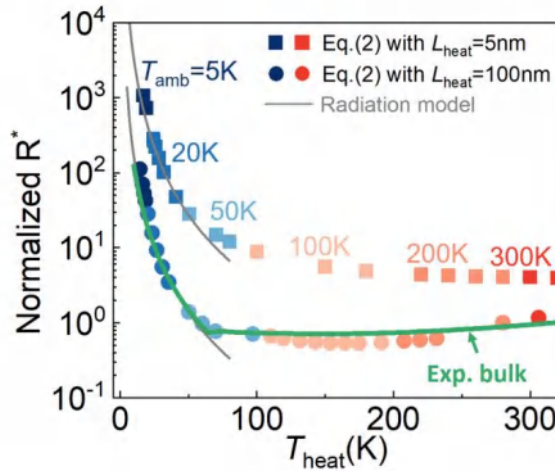


图 2 方形点与圆形点分别表示在环境温度为5K、20K、50K、100K、200K和300K下，两种热点尺寸对应的归一化热阻。灰色曲线为公式（6）的计算结果，绿色曲线为参考文献[6]中的实验数据结果。仿真区域中边长设为 $L=10\mu\text{m}$ 。

MOSFET 中的热点尺寸通常处于数纳米至数微米的范围内，其大小受到偏置条件的影响[12,15,21]。已有实验表明，在低温下，偏置条件对热阻的影响远大于常温，但其背后的声子输运机制尚不明确[6]。为了理解这一现象，本文进一步探讨了热点尺寸对热输运行行为的影响。已有研究指出，由热点温升定义的热阻 $R^* = \partial\Delta T_{\text{heat}} / \partial P|_{T_{\text{amb}}}$ 主要取决于热点温度，而非环境温度[6, 7]。因此，不同环境温度下的热阻可以统一为热点温度的函数。在图 3 中，我们基于两个典型热点尺寸 $L_{\text{heat}}=5\text{nm}$ 和 $L_{\text{heat}}=100\text{nm}$ ，计算了不同环境温度下，热阻随热点温度变化的结果。

由于本研究采用二维仿真，忽略了栅宽方向的声子输运，因此为确保仿真结果与实验数据具备可比性，需通过归一化来修正维度差异 $\text{Normalized } R^* = R^* / R^*_{T_{\text{heat}}=300\text{K}}$ 。对于实验数据，以其在 300K 下的热阻值为归一化基准[9]。为了体现热点尺寸的影响，BTE 的计算结果均用 $L_{\text{heat}}=100\text{nm}$ 在 300K 下的热阻值归一化。如图 2 所示， L_{heat} 为 100nm 的 BTE 计算结果与文献[6]中 MOSFET 的实验数据（绿色曲线）高度一致。

除声子 BTE 外，本文还引入声子黑体辐射模型（如图 3 中的灰色曲线所示），以捕捉低温下热流的 T^4 依赖关系，从而表征纯弹道声子输运行为。该模型在超低温区间与实验数据表现出良好一致性[22]：

$$R^*(T_{\text{heat}}) \cdot W = \frac{dT_{\text{heat}}}{dP/W} = \frac{1}{4\sigma_p(w/W)T_{\text{heat}}^3}. \quad (3)$$

$\sigma_p = \pi^2 k_B^4 / 40\hbar^3 v_{\text{ave}}^2 \approx 485\text{Wm}^{-2}\text{K}^{-4}$ 是硅中声子的等效斯特藩-玻尔兹曼常数， $P/W = \dot{q}L_{\text{heat}}^2$ 表示单位栅宽下的热点热源项，其中 W 为器件的深度， $w/W = 3L_{\text{heat}}$ 为单位栅宽下的声子辐射区域面积。

在 5K 至 60K 的低温区间，两种热点尺寸对应的热阻相差 1–2 个数量级，明显大于常温下的差异。此时，由式（2）计算所得热阻逐渐趋近于声子黑体辐射极。这是由于当声子平均自由程远大于热点尺寸（ $\text{MFP} \gg L$ ）时，热传导以弹道声子为主，这一结论也与文献[23]相符。式（3）右端表明，在 5K–60K 范围内，热阻与热点尺寸成反比。然而，当热点温度升高超过 60K 后，扩散声子输运逐渐出现，使得公式（3）不再适用。这说明，尽管黑体辐射模型可有效表征低温下的弹道声子特性，但要全面描述从弹道到扩散过渡的声子输运过程，仍需发展具备温度依赖性的 BTE 模型。当热点温度超过 100K 时，两种热点尺寸下的热

阻表现出相反的温度依赖趋势：对于 100nm 热点，热阻在 150K 以上开始上升；而对 5nm 热点，热阻仍持续下降。这一现象与实验观察一致[7, 9, 23]：热阻上升趋势对应 bulk 器件的表现，而下降趋势则符合 SOI 器件的行为。

上述现象可归因于两种相互竞争的机制：其一，硅体块材料的热导率随温度升高而降低，导致热阻上升；其二，声子平均自由程随温度升高而缩短，弹道输运占比减小，从而降低热阻。对于较小的热点（5nm），弹道声子在较宽温度范围内占主导地位。温度升高后 MFP 缩短，弹道效应减弱，热阻下降，此时机制二占主导。而对于较大的热点（100nm），弹道效应较弱，温度升高后机制一主导，导致热阻上升。

综上所述，本节系统分析了热点尺寸对 MOSFET 有源区热阻的影响。在 60K 以下，热阻与热点尺寸呈反比关系；而在 150K 以上，小热点尺寸对应的热阻开始上升，而大热点对应的热阻则持续下降。该现象首次从声子输运机制角度进行了理论解释，并与实验结果保持一致。

3 结论

综上所述，本文结合第一性原理方法与温度依赖的声子玻尔兹曼输运方程（BTE），实现了低温器件中热点邻域的多尺度建模并研究了声子输运行为。仿真结果与实验现象吻合，揭示了低温下的声子输运机制。研究表明，弹道声子、声子热物性的温度依赖性以及热点尺寸是影响低温热输运的关键因素。首先，弹道声子显著增强了热点处的温升；其次，强温度依赖的声子热物性导致了温升对功率的非线性响应。此外，硅有源区中的热点尺寸对热阻具有显著影响：在 5K 至 60K 范围内，热阻与热点尺寸呈反比关系；在温度高于 150K 时，热阻的温度依赖趋势与实验结果一致，并可由声子输运机制予以解释。本研究深入探讨了硅基 MOSFET 中有源区的声子输运特性，为理解低温器件热点热输运机制提供了重要理论支持。

致谢

本研究得到了国家自然科学基金（项目编号：52476080）的资助；本研究的计算结果得到了上海交通大学高性能计算中心的支持和帮助；

参考文献

- [1] E. Charbon, M. Babaie, A. Vladimirescu, F. Sebastiano, Cryogenic CMOS Circuits and Systems: Challenges and Opportunities in Designing the Electronic Interface for Quantum Processors, IEEE Microwave Magazine, 22 (2021) 60-78.
- [2] F. Jazaeri, A. Beckers, A. Tajalli, J.M. Sallese, A Review on Quantum Computing: From Qubits to Front-end Electronics and Cryogenic MOSFET Physics, 2019 MIXDES - 26th

- International Conference "Mixed Design of Integrated Circuits and Systems", 2019, pp. 15-25.
- [3] R. Cheng, B. Li, Z. Wang, J. Zhang, W. Shan, J. Zhang, Y. Cai, G. Han, Low-temperature CMOS technology for high-performance computing: development and challenges, *SCIENTIA SINICA Informationis*, 54 (2024).
- [4] C. Prasad, A Review of Self-Heating Effects in Advanced CMOS Technologies, *IEEE Transactions on Electron Devices*, 66 (2019) 4546-4555.
- [5] G. Pavlidis, B. Foley, S. Graham, 9 - Gate resistance thermometry: An electrical thermal characterization technique, (2022) 201-221.
- [6] A.A. Artanov, E.A. Gutiérrez-D, A.R. Cabrera-Galicia, A. Kruth, C. Degenhardt, D. Durini, J. Méndez-V, S.V. Waasen, Self-Heating Effect in a 65 nm MOSFET at Cryogenic Temperatures, *IEEE Transactions on Electron Devices*, 69 (2022) 900-904.
- [7] F. Bergamaschi, T. Mota Frutuoso, B. Paz, G. Billiot, A. Jansen, P. Galy, E. Vincent, F. Gaillard, B. Duriez, M. Cassé, Experimental Analysis and Modeling of Self-Heating and Thermal Coupling in 28 nm FD-SOI CMOS Transistors Down to Cryogenic Temperatures, *IEEE Transactions on Electron Devices*, pp (2024) 1-7.
- [8] M. Cassé, B.C. Paz, F. Bergamaschi, G. Ghibaudo, F. Serra, G. Billiot, A.G.M. Jansen, Q. Berlingard, S. Martinie, T. Bedecarrats, L. Contamin, A. Juge, E. Vincent, P. Galy, M.A. Pavanello, M. Vinet, T. Meunier, F. Gaillard, FDSOI for cryoCMOS electronics: device characterization towards compact model, (2022) 34.36.31-34.36.34.
- [9] P.A.T. Hart, M. Babaie, A. Vladimirescu, F. Sebastiano, Characterization and Modeling of Self-Heating in Nanometer Bulk-CMOS at Cryogenic Temperatures, *IEEE Journal of the Electron Devices Society*, 9 (2021) 891-901.
- [10] Y. Sheng, S. Wang, Y. Hu, J. Xu, Z. Ji, H. Bao, Integrating First-principles-based non-Fourier thermal analysis into nanoscale device simulation, *IEEE Transactions on Electron Devices*, 71 (2024) 1769-1775.
- [11] Y. Shen, X.-S. Chen, Y.-C. Hua, H.-L. Li, L. Wei, B.-Y. Cao, Bias dependence of non-Fourier heat spreading in GaN HEMTs, *IEEE Transactions on Electron Devices*, 70 (2022) 409-417.
- [12] Y. Shen, H.-A. Yang, B.-Y. Cao, Near-junction phonon thermal spreading in GaN HEMTs: A comparative study of simulation techniques by full-band phonon Monte Carlo method, *International Journal of Heat and Mass Transfer*, 211 (2023) 124284.
- [13] R. Li, J.-X. Wang, E. Lee, T. Luo, Physics-informed deep learning for solving phonon Boltzmann transport equation with large temperature non-equilibrium, *npj Computational*

Materials, 8 (2022) 29.

- [14] C. Zhang, S. Huberman, X. Song, J. Zhao, S. Chen, L. Wu, Acceleration strategy of source iteration method for the stationary phonon Boltzmann transport equation, *International Journal of Heat and Mass Transfer*, 217 (2023).
- [15] J. Sun, X. Li, Y. Sun, Y. Shi, Impact of Geometry, Doping, Temperature, and Boundary Conductivity on Thermal Characteristics of 14-nm Bulk and SOI FinFETs, *IEEE Transactions on Device and Materials Reliability*, 20 (2020) 119-127.
- [16] C.W. Chang, D. Okawa, H. Garcia, A. Majumdar, A. Zettl, Breakdown of Fourier's Law in Nanotube Thermal Conductors, *Physical Review Letters*, 101 (2008) 075903.
- [17] R. Li, E. Lee, T. Luo, Physics-informed neural networks for solving multiscale mode-resolved phonon Boltzmann transport equation, *Materials Today Physics*, 19 (2021) 100429.
- [18] C. Kittel, P. McEuen, *Introduction to solid state physics*, 2018.
- [19] Y. Hu, R. Jia, J. Xu, Y. Sheng, M. Wen, J. Lin, Y. Shen, H. Bao, GiftBTE: an efficient deterministic solver for non-gray phonon Boltzmann transport equation, *Journal of Physics: Condensed Matter*, 36 (2023) 025901.
- [20] X. Zhang, C. Shao, H. Bao, Cryogenic thermal transport properties from accelerated first-principles calculations: Role of boundary and isotope scattering, *Physical Review B*, 110 (2024) 224301.
- [21] X. Chen, S. Boumaiza, L. Wei, Modeling bias dependence of self-heating in GaN HEMTs using two heat sources, *IEEE Transactions on Electron Devices*, 67 (2020) 3082-3087.
- [22] K. Triantopoulos, M. Cassé, S. Barraud, S. Haendler, E. Vincent, M. Vinet, F. Gaillard, G. Ghibaudo, Self-Heating Effect in FDSOI Transistors Down to Cryogenic Operation at 4.2 K, *IEEE Transactions on Electron Devices*, (2019) 1-8.
- [23] A.Y. Choi, I. Esho, B. Gabritchidze, J. Kooi, A.J. Minnich, Characterization of self-heating in cryogenic high electron mobility transistors using Schottky thermometry, *Journal of Applied Physics*, 130 (2021).

基于 RDL-first 工艺的封装材料仿真分析

张需

中国电子科技集团公司第五十八研究所¹

摘要：RDL-first（再布线先行）工艺后封装特点避免了封装材料热冲击变形，对封装材料具有宽松的热性能要求，支持异质集成。相比于传统 EMC（环氧树脂料）材料，PDMS（聚二甲基硅氧烷）具有高导热性、柔韧性与生物相容性，应用于微流控硬件及生物模型制作等。文章基于 RDL-first 工艺对于 EMC、PDMS 材料进行有限元结构仿真分析，总结两种材料受相同大小压力的变形、应力响应，突出了 PDMS 材料超弹性、高可靠性特点，对不同场景下封装材料选择提供支持。

关键词：RDL-first；EMC；PDMS；有限元分析；

1 前言

RDL-first（Redistribution Layer First，再布线先行）是一种先进的半导体封装技术，与传统封装工艺顺序相反。它先在载体上制作再分布层（RDL），然后再进行芯片贴装和其他工艺步骤^[1]。RDL-first 工艺流程示意图如图 1 所示。

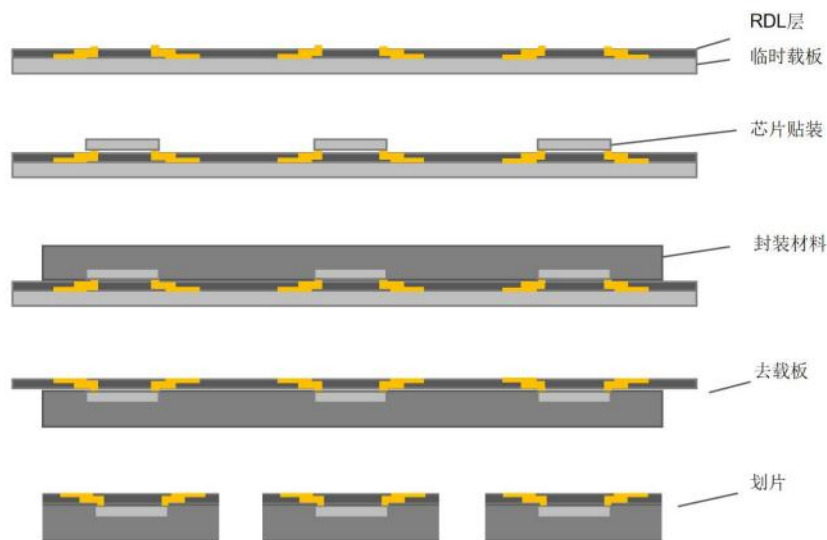


图 1 RDL-first 工艺流程示意图

RDL-first 工艺相比传统封装技术，能够实现更精细的线宽/线距，显著提升集成密度和互连性能^[2]；同时缩短了信号传输路径，降低了寄生效应，使封装整体电性能更优。这种工艺还突破了传统封装对芯片凸点间距的限制，并赋予设计

¹ 江苏省无锡市滨湖区建筑西路 777 号，214072

更大灵活性，便于实现多芯片异构集成和 2.5D/3D 先进封装，特别适合高性能计算、AI 芯片和植入式医疗芯片等对高密度互连有严苛要求的应用场景。

传统 EMC 材料以硅微粉等高脆性、高模量材料掺杂，以提供较好的支撑性，但其高脆性特点大大限制了封装体的应用场景^[3]。一些芯片应用如植入式医疗设备芯片、表皮电子设备、体内降解芯片等对封装材料生物相容性、柔性要求较高，EMC 材料已经无法满足封装需求。PDMS 材料是一种高分子有机硅化合物，具有优异的光学透明性、生物相容性和可调弹性，在上述应用领域被广泛关注^[4]。

本文利用 ANSYS 有限元仿真分别对于使用 EMC、PDMS 材料封装的 RDL-first 工艺晶圆进行结构仿真，以更直观的比较两种材料的结构特性。

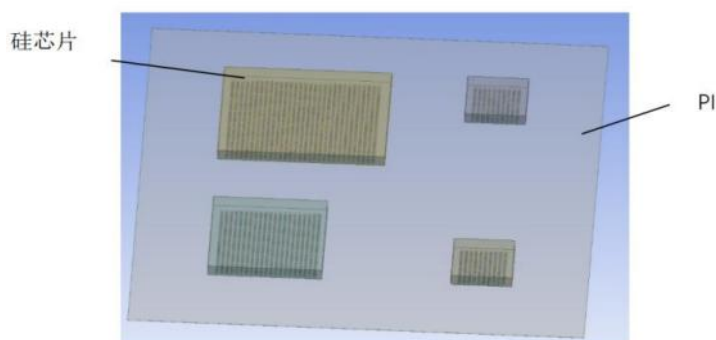
2 方法

2.1 模型建立

以一种典型的功能芯片设计为基础，在 ANSYS 中建立了包括 PI(聚酰亚胺)、Cu-pillar (铜柱)、硅芯片 (Si)、封装材料的封装体仿真模型，模型布局图如图 2 所示，各组分材料参数如表 1 所示。

表 1 仿真材料参数表

材料	热膨胀系数/ C^{-1}	杨氏模量/GPa	泊松比
EMC	7.2E-06	23.8	0.371
PDMS	9.6E-04	0.0015	0.5
Cu	1.7E-05	115	0.34
PI	5.0E-05	3.4	0.19
Si	2.6E-06	131	0.28



(1) 仿真模型中的单模块正面布局示意图



(2) 仿真模型中的单模块侧面示意图

图 2 单模块基础仿真模型示意图

涂覆封装材料后晶圆表面示意图如图 3 所示。

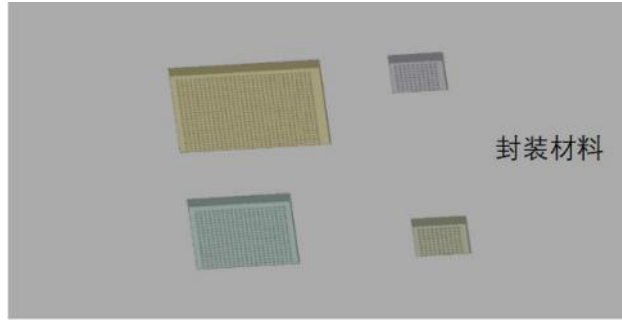


图 3 涂覆封装材料后的晶圆表面示意图

2.2 分析设置

(1) 网格设置

总体网格尺寸设置为 1mm，对 Cu-pillar 进行局部尺寸加密设置为 0.5mm，由于 PI 层厚度仅 $10\ \mu\text{m}$ ，远小于 PI 层 XY 方向的尺寸，因此将 PI 层简化为薄壳结构，不进行厚度层面的划分。

(2) 边界条件

在封装体中心点设置为固定支撑点，在 X 轴方向施加约束条件 $U_Y=0$ ，Y 轴方向施加约束条件 $U_X=0$ 。在进行 Pdms 材料封装仿真时，由于材料符合超弹性模型，在结构分析设置中开启大变形。

(3) 施加载荷

为模拟封装体在实际使用过程中的形变，对单模块 (Block) 四角施加压力，以观察封装体的变形量及表面应力，在仿真中为了计算方便将力值大小取 1N，如图 4 所示。

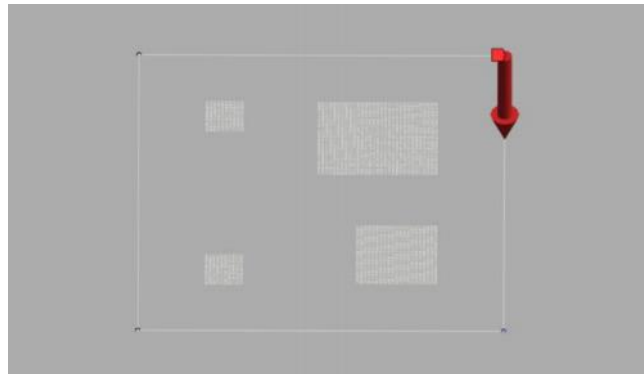


图 4 仿真过程施加载荷方向及施力点

3 结果分析

对封装结构分别赋 EMC、PDMS 两种材料属性，进行总变形、等效应力与最大主应力的仿真。仿真对比结果如表 2 所示。PDMS 材料的仿真变形结果如图 5 所示。

表 2 两种材料的仿真结果对比

材料	总变形/mm	等效应力/MPa	最大主应力/MPa
EMC	1.2	28.5	3.8
PDMS	27.4	520.6	546.6

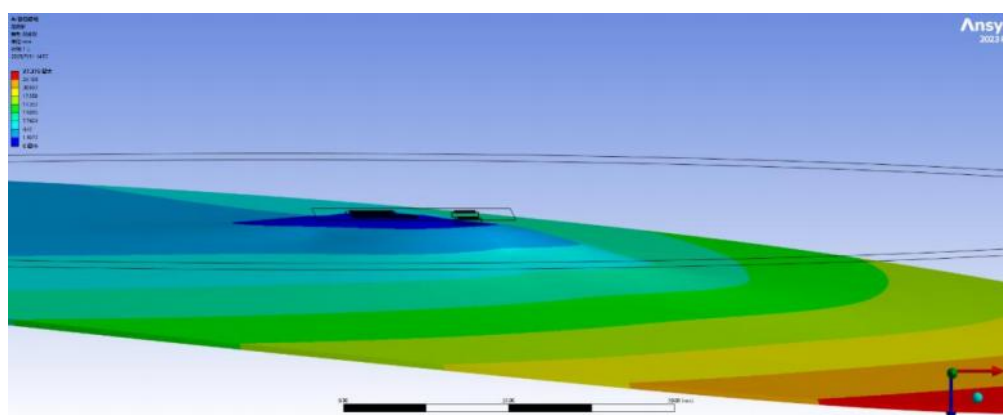


图 5 PDMS 材料仿真总变形示意图

根据两种材料的仿真结果对比表，可以直观地看出 EMC 材料具有良好的封装支撑性，在同样的载荷条件下，变形量与应力都较小。而 PDMS 材料仿真变形量、应力远大于 EMC 材料的仿真响应。主要原因是二者材料刚度（杨氏模量）的巨大差异，PDMS 的低模量直接导致其在相同的载荷下产生大得多的变形，大变形情况下，材料几何形状发生很大改变、泊松比效应放大，导致局部区域应力状态复杂且集中。PDMS 材料最大主应力仿真结果示意图如图 6 所示。

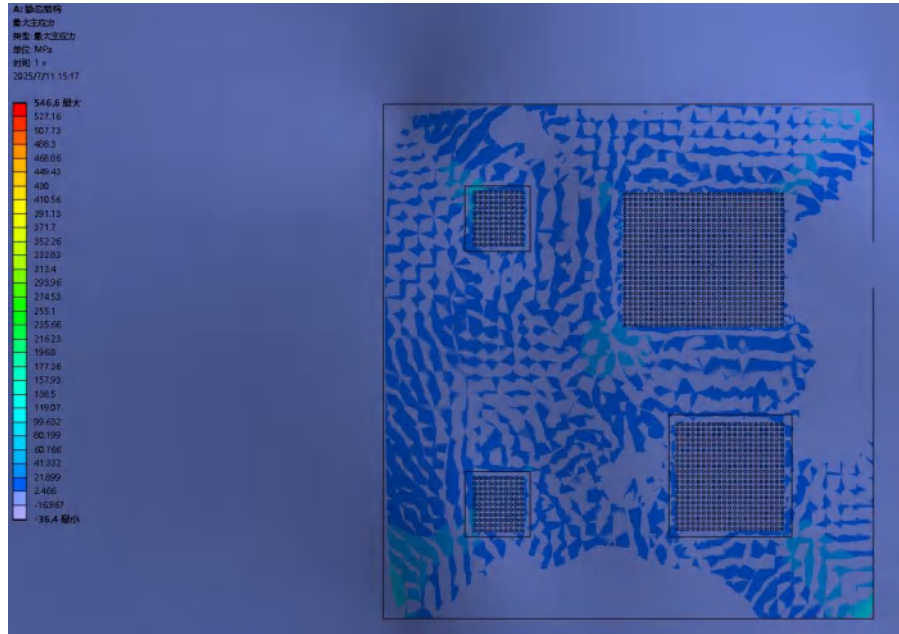


图 6 PDMS 材料仿真最大主应力分布示意图

较大的应力主要集中分布在受力点、固定支撑点及 Cu-pillar 材料连接位置。单一的 Pdms 材料处表面等效应力最小。虽然 EMC 材料相比，PDMS 封装支撑性差，其变形会产生的巨大应力，但该应力大小远小于硅芯片的断裂强度，对芯片功能性不会有破坏性的影响。

4 结论

文章基于 RDL-first 工艺对于 EMC 材料及 PDMS 材料进行同等条件下的变形量及应力仿真，所得结论总结如下：

- (1) EMC 材料具有更好的结构支撑性及抗压能力；
- (2) PDMS 材料具有超弹性、高柔韧性，其封装可靠性也在信任范围。

根据仿真结论，可以总结这两种材料所适应的不同应用场景。EMC 材料凭借高刚性、耐高、低成本优势，主要用于电子领域刚性封装，如芯片保护、功率模块，提供结构支撑和绝缘防护。而 PDMS 因其超弹性、生物相容性及高透明度，不可替代地应用于生物医学、软体机器人和仿生光学等柔性场景，有效弥补了传统封装材料的局限。

参考文献

- [1] 吉勇,王成迁,李杨. 扇外型封装发展、挑战和机遇[J]. 电子与封装,2020,20(8):1-6.
- [2] 张政楷,戴飞虎,王成迁. 先进封装 RDL-first 工艺研究进展[J]. 电子与封装,2023,23(10):26-35.

- [3] 宗小雪,苏梅英,周云燕,等. 基于扇外型封装塑封材料性能表征研究[J]. 电子元件与材料,2020,39(5):90-96.
- [4] 张英豪. 面向柔性电子封装的抗疲劳拉伸 PDMS 薄膜工艺及性能研究[D]. 湖北:华中科技大学,2022.

基于 TSV 硅桥的光电微系统三维异构集成设计与仿真研究

闵成彧^{1#}, 张文亚^{1,2*#}, 田斯劼¹, 陈鹏¹, 王奥培¹, 涂凌霞¹, 邵斯竹¹, 吴罚¹

*联系作者: wenya.zhang@cumec.cn

#共同一作: 闵成彧、张文亚

1 联合微电子中心有限责任公司 重庆 401332, 2 西安理工大学 西安 710048

摘要: 硅桥芯粒是实现多芯粒光电三维异构集成的重要部件, 针对多路高速光模块典型应用场景, 本文首先采用 CUMEC 硅通孔 (TSV)、大马士革 RDL 等工艺, 进行了高纵横比 TSV 硅桥芯粒设计及和流片; 并提出一种基于 TSV 硅桥的高密度光电集成三维封装结构及制造方法, 实现了 TIA 芯片、硅光芯片和 TSV 桥芯片的多芯粒典型光电微系统集成设计; 最后, 采用了电磁仿真软件 HFSS, 进行了光电微系统三维异构集成结构的信号完整性仿真与分析。硅桥芯粒切片及 WAT 测试结果表明, 芯粒内部 TSV 侧壁光滑, 纵横比达到 10: 1, TSV 串联阻值平均值为 $0.48\ \Omega$, 连通性能良好, 可实现 TIA 电信号垂直传输。链路仿真结果表明, TIA 的 16 条输入单端共面波导信号通道 $S_{21} \leq -1.1\text{dB/mm @}25\text{GHz}$; TIA 的 16 条输出差分共面波导信号通道信号通道 $S_{21} \leq -0.6\text{dB/mm @}25\text{GHz}$ 。本文为设计更高性能的多通道高速光电微系统异构集成提供了指导和支撑。

关键词: 光电微系统; 三维异构集成; TSV 硅桥; 信号完整性

1 引言

随着全球网络流量的不断增长, 高速率数据通信技术带宽不断提升, 数据中心对光模块在传输速率、带宽、高能效的要求越来越高。据统计, 数据中心处理的数据量每年以 30% 的速率快速增长, 2010 到 2022 年间, 服务器用的交换芯片容量从 0.646Tb/s 增加到 25.6Tb/s, 激增了 80 倍, 功耗提升 8 倍^[1-2]。传统光芯片和电芯片主要采用引线进行互连, 并且通过采用板上芯片封装技术 (Chips on Board, COB) 实现二维集成的组装。此种方式易于实现, 但是光芯片与电芯片的高速信号互连走线较长, 且封装体积较大, 严重限制了 I/O 数量和信号通道传输速率, 难以满足高传输速率、高带宽、低损耗等方面的要求^[3-5]。

光电合封 (CPO, Co-Packaged Optics) 是一种新型的光电子集成技术, 通过将光芯片与电芯片在芯片级集成在同一基板上, 使得光信号和电信号在芯片内直接转换, 可大幅减小封装尺寸、提高数据转换效率、降低功耗, 为实现高带宽、低延迟的光电互连提供了新的解决方案^[6-7]。TSV 硅桥可作为一种局部高密度互连中介层, 可嵌入封装基板内部, 通过微米级 TSV 和重布线层 (RDL), 在微米尺度上实现多芯粒间超短距离、超高带宽的互连。另一方面, 在 TSV 硅桥构建的多通道光电互连系统中, 高速信号传输面临着严峻的信号完整性挑战。

硅桥内高密度 TSV 阵列、微凸点及细间距 RDL 走线会引入显著的寄生参数，导致阻抗不连续、串扰、损耗等效应。因此，开展多通道全链路电磁协同仿真具有至关重要的意义。

本文聚焦于基于 TSV 硅桥的光电微系统三维异构集成架构，结合多物理场仿真手段，开展系统级设计与性能验证，以支撑高密度光电微系统的设计与应用。

2 TSV 硅桥芯粒与多芯粒封装版图设计

2.1 多芯粒扇出重布线层版图设计

如下图 1(a)所示，本封装结构基于晶圆重塑工艺、重布线等工艺将 4 个四通道 TIA 芯片、1 块硅光芯片和 3 块硅桥芯粒高密度互联集成，形成一个 16 收 16 发的光电微系统，其封装尺寸为 $9850 \times 4150 \mu\text{m}$ 。硅光芯片高频信号通过多层重布线连接到 TIA 芯片的输入端，然后通过 TIA 芯片输出，然后再通过布线层输入到硅桥芯粒底部，最后通过 TSV 硅桥芯粒将高频信号垂直输出到封装顶部与外界互联。其中左硅桥芯粒将 TIA1 和 TIA2 两个电芯片 8 个高频通道信号和部分地、电源和控制信号引出；右硅桥芯粒将 TIA3 和 TIA4 两个电芯片 8 个高频通道信号部分地、电源和控制信号引出；上硅桥芯粒将整个系统剩余地、电源和控制信号引出。此封装方案版图采用沿 Y 轴对称结构设计，左边 8 个通道高频信号版图和右边 8 个通道高频信号版图层叠结构完全一致。

如图 1(b)所示，光电微模组封装步骤如下：1)在塑封磨具正旋涂一层可去除的牺牲层；2)在牺牲层上按设计要求一次放置 TSV 桥芯片，TIA 跨阻放大器芯片，芯片 I/O 口朝下。其中 TIA 芯片在塑封前会在芯片背面用银增厚，与 TSV 桥芯片高度一致；3)TSV 桥芯片和 TIA 跨阻放大器芯片放置完成后，在磨具中注入树脂材料将芯片完全覆盖；4)待树脂完全固化后去除塑封磨具，将芯片 I/O 面朝上；5)在芯片 I/O 面做重布线层，使用 4P4M 将跨阻放大器的 I/O 重新分布，以方便信号与硅光芯片、外接基板相连。待重布线层完成后，对“七芯片微模组”背面进行减薄露出 TSV 芯片 PAD。然后对 TSV 芯片电镀，以便后续与基板打线连接；6)最后对“七芯片微模组”植球后，将完成植球的微模组贴到硅光芯片上，然后通过与基板相连，构成一个 16 收 16 发的光电微系统。

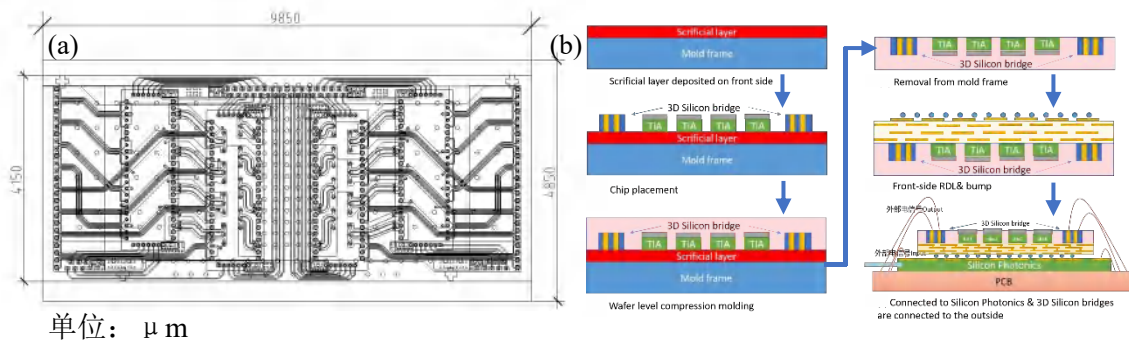


图 1 (a)光电微系统重布线层版图；(b)封装设计流程图

2.2 20 X 200 TSV 桥芯片设计

图 2 为 CUMEC 微系统中心自主设计、流片的三种 TSV 硅桥芯粒。其中上 TSV 桥芯片尺寸为 $3150 \times 235 \mu\text{m}$ 、左 TSV 桥芯片和右 TSV 桥芯片尺寸均为 $1030 \times 3750 \mu\text{m}$ 。

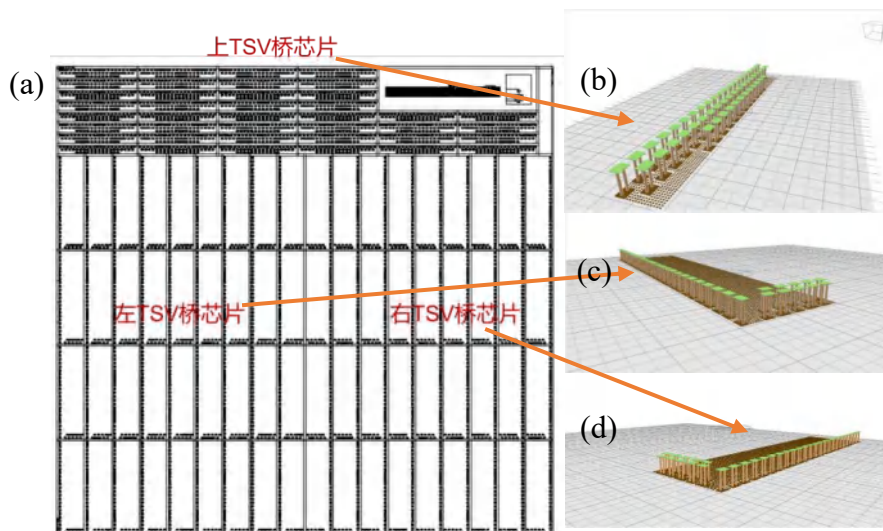


图 2 (a)硅桥芯粒设计图；(b)左 TSV 桥芯片；(c)上 TSV 桥芯片；(d)右 TSV 桥芯片

硅桥芯粒 I/O 焊盘为 $100 \times 70 \mu\text{m}$ ，相邻焊盘中心间距为 $150 \mu\text{m}$ ，每个 I/O 焊盘下均设计了两个 $20 \times 200 \mu\text{m}$ TSV，两个 TSV 中心间距 $50 \mu\text{m}$ 。硅桥芯粒设计完成后，采用 CUMEC 大马士革、Bosch 等工艺实现三种硅桥芯粒成功流片。

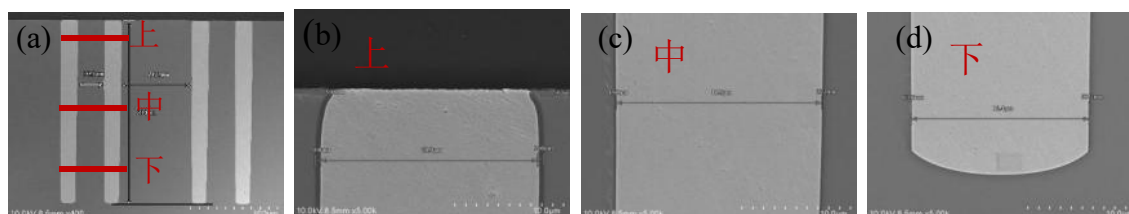


图 3 硅桥芯粒侧壁上、中、下 SEM 扫描电镜图

图 3 所示为硅桥芯粒 FA 切片横截面扫描电镜(SEM)图，结果显示硅桥芯粒内部 TSV 的上中下三段 CD 值分别为 $20 \mu\text{m}$ 、 $18.9 \mu\text{m}$ 、 $16.4 \mu\text{m}$ ，TSV 深度为 $210 \mu\text{m}$ 、垂直角度约为 89.51° ，且 TSV 侧壁光滑、无空洞。如图 4(b)所示，为了监控硅桥芯粒流片良率，特意设计了 6 个 TSV 通孔串联 WAT 测试结构，放置在每个光罩曝光窗口中，以监控 TSV 通孔连接性。如图 4(a) 所示，使用四探针法对晶圆上每个曝光窗口中 WAT 结构进行测试，测试结果表明 6 个 TSV 通孔串联阻值在 $0.47\text{-}0.495 \Omega$ 之间。如图 4(c)所示，整个晶圆中 WAT 测试结果平均值为 0.48Ω 。SEM 扫描电镜和 WAT 结果表明，CUMEC 加工的硅桥芯粒内部 TSV 侧壁光滑笔直，纵横比达到 10: 1 且连通性良好，可以实现光探测器经过 TIA 放大生成的

电信号向外部的垂直传输的功能。

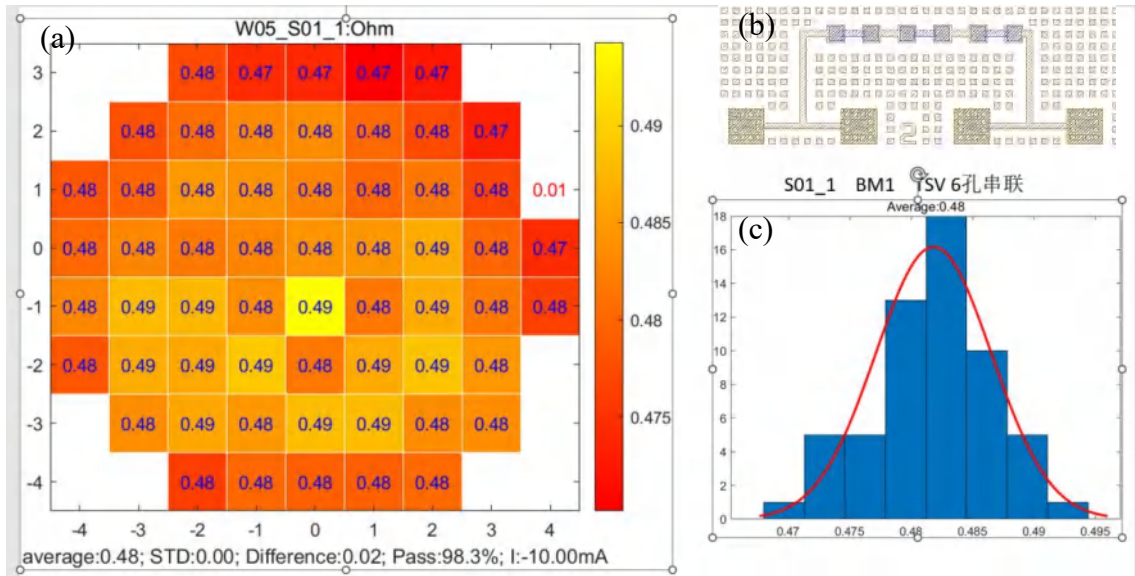


图4 纵横比 10:1 的 TSV 硅桥芯粒 WAT 测试结果,(a) 晶圆上每个曝光窗口 WAT 阻值分布; (b) TSV 通孔串联 WAT 测试结构; (c) WAT 阻值分布统计

3 光电互联通道信号完整性分析

3.1 信号完整性仿真设置

仿真设置流程如下：基于图 4(a)中光电微系统异构集成结构，将版图文件导入三维全电磁仿真软件 HFSS，进行仿真参数设置，完成频域仿真。HFSS 中异构集成结构的层叠结构如图 5(a)所示，仿真参数设置如图 5(b)所示。层叠结构中，最下方是厚度为 200 μm 的硅基衬底，衬底和金属间有 5 μm 厚度的绝缘层。金属层 M1-M3 厚度 3 μm , M4 层厚度 8 μm , Via 层厚度均为 5 μm , 传输线集中在 M1-M3 层。金属采用铜，电导率为 58000000 s/m，金属层介质采用 PI，介电常数为 3.1。仿真频率范围为 0-35GHz，采用步进 0.2GHz，收敛误差精度 0.02，最小收敛误差次数 2 次。



图 5 (a)层叠结构示意图; (b) 仿真参数设置

3.2 信号通路设计及仿真结果分析

图 6 (a) 是基于 TSV 硅桥的光电微系统三维示意图, 此微系统版图采用沿 Y 轴对称结构设计, 左边 TIA1、TIA2 高频信号版图和右边 TIA3、TIA4 高频信号版图的走线方式和层叠结构完全一致, 因此仿真时仅选取了左边 TIA1 和 TIA2 的光电探测器与 TIA 芯片互连的 8 条单端共面波导信号通道和 TIA 与硅桥芯粒互连的 8 条差分共面波导信号通道进行仿真, 如图 6 (b)。

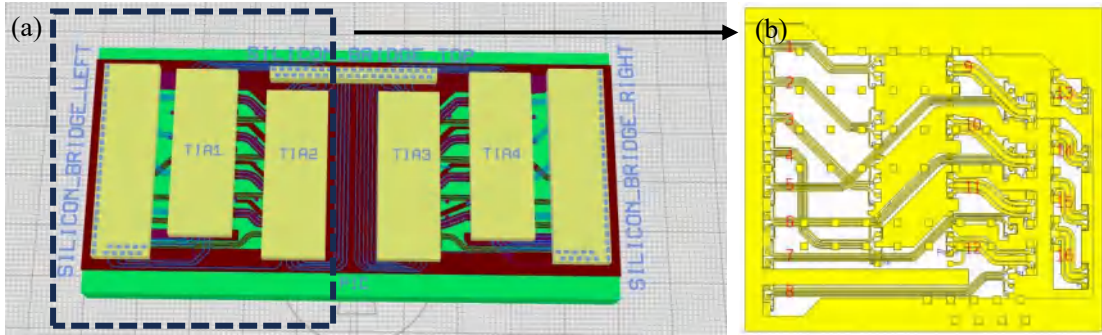


图 6 (a) 光电微系统三维示意图; (b) 仿真版图

光电探测器与 TIA 芯片互连的 8 条信号通道均采用单端共面波导走线, 信号线线宽均为 $70\mu\text{m}$, 地线线宽 $40\mu\text{m}$, 信号线与地线间距 $15\mu\text{m}$, 单端阻抗控制在 50Ω 。单端信号线从 M1 层通过 V1 到达 M2, 然后经过 V2 到达 M3, 再经过 M3 的长走线后走到 M4 层, 信号线主要在 M3 层, 每条信号线端口一端设置在 M1 层, 另一端设置在 M4 层, 信号通路模型如下图 7(a)所示。8 组单端共面波导信号走线仿真结果如图 7(b)所示。结果表明在信号频率 25GHz 时, 8 组单端共面波导信号通道均满足 $S_{21} \leq -1.1\text{dB/mm}$ 。连接 TIA1 的通道 9、10、11 和 12 传输信号线长度接近, 带宽接近; 连接 TIA2 的通道 13、14、15 和 16 传输信号, 带宽接近。因为通道 13 传输线最短, 约 $500\mu\text{m}$, 所以在 30GHz 以内损耗最小。

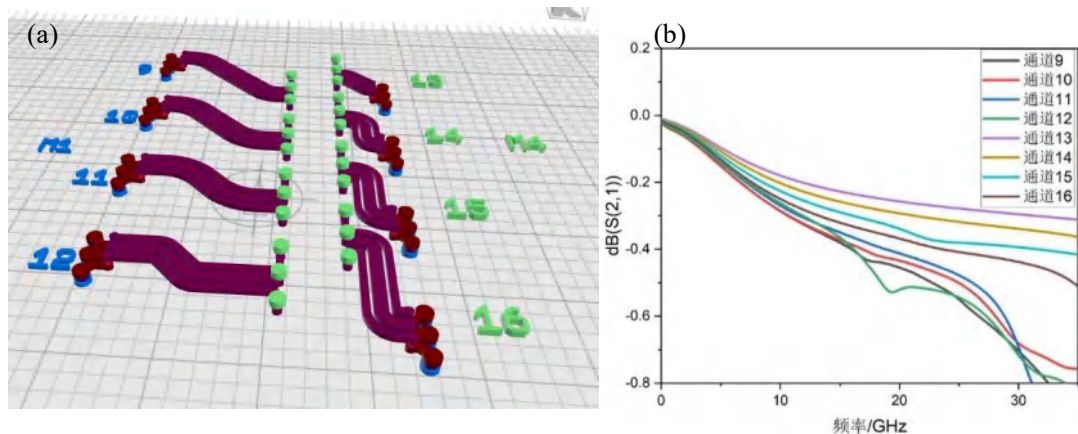


图 7 (a) 信号通路模型图; (b) 8 组单端共面波导信号走线仿真结果

TIA 与硅桥芯粒互连的 8 条信号通道均采用共面波差分导走线, 其中 TIA1 芯片的 4 对差分共面波导线在 M1 层, 差分共面波导信号线与地线线宽均为

15 μm ，差分信号线间线距为 18.5 μm ，信号线与地线间的距离为 18.5 μm ，差分对阻抗控制在 100 Ω ；TIA2 芯片的 4 对差分共面波导线从 M1 层经过 V1 到达 M2，然后 M2 层长走线后经过 V1 回到 M1，信号线主要在 M2 层，差分共面波导信号线与地线线宽均为 15 μm ，差分信号线间线距为 15 μm ，信号线与地线间的距离为 15 μm ，差分对阻抗控制在 100 Ω 。仿真时 8 组信号通路共计 32 个端口均设置在 M1 层，信号通路模型如下图 8 (a)所示。8 组差分共面波导信号走线仿真结果如图 8 (b) 所示，结果表明在信号频率 25GHz 时，8 组差分共面波导信号通道均满足 $S_{21} \leq -0.6 \text{ dB/mm}$ 。其中通道 4 带宽最窄，因为信号线与地线均经过两次 135°的拐角，且受到上方 M2 层的差分走线的影响。

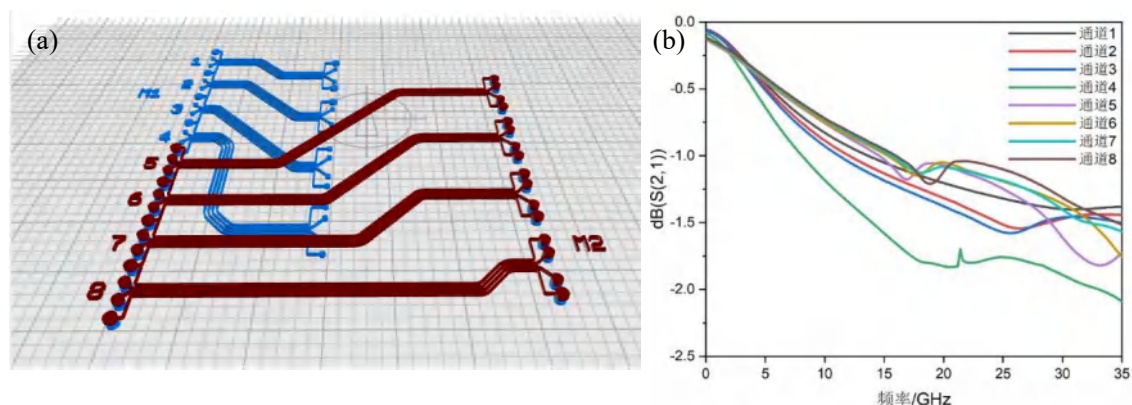


图 8 (a) 信号通路模型图；(b) 8 组差分共面波导信号走线仿真结果

基于电磁仿真软件 HFSS，对整个光电封装微模组 TIA1、TIA2 的输入端 8 条单端共面波导信号通道和输出端 8 条差分共面波导信号通道进行仿真，结果表明封装设计中的链路通道均满足在 25GHz 以内 $S_{21} \leq -3 \text{ dB}$ 信号传输要求。又因为封装结构中版图中右边 TIA3、TIA4 的链路通道与左边 TIA1、TIA2 链路通道的走线方式和链路结构完全一致，因此可推定链路通道也满足在 25GHz 以内 $S_{21} \leq -3 \text{ dB}$ 信号传输要求。

4 结论

本文基于 CUMEC 的硅通孔、大马士革 RDL 等工艺完成了高纵横比 TSV 芯粒设计并流片，提出一种基于高纵横比 TSV 硅桥的高密度光电集成的三维封装结构及其制造方法，实现 4 个四通道跨阻放大器(TIA)芯片、1 个硅光芯片和 3 个 TSV 桥芯片的典型光电微系统集成设计。硅桥芯粒切片及 WAT 测试结果表明，芯粒内部 TSV 侧壁光滑，垂直角度为 89.51°，纵横比达到 10: 1；TSV 串联阻值平均值为 0.48 Ω ，连通性能良好，可实现 TIA 电信号垂直传输。基于电磁仿真软件 HFSS，电磁仿真结果表明，光电微系统集成结构中，光电探测器与 TIA 芯片互连的 16 条单端共面波导信号通道均满足 $S_{21} \leq -1.1 \text{ dB/mm} @ 25 \text{ GHz}$ ，TIA 与硅桥芯粒互连的 16 条差分共面波导信号通道均满足 $S_{21} \leq -0.6 \text{ dB/mm} @ 25 \text{ GHz}$ 。

本文的多芯片封装方案和仿真实验为后续高密度光电微系统的设计与应用提供了一定的参考意义。

参考文献

- [1] 孙佳琪, 宋艳飞, 王睿哲. 片间光电融合技术分析[J]. 中国集成电路, 2024, 33(5):19-23.
- [2] Min Tan, Jiang Xu et al. Co-packaged optics (CPO): status, challenges, and solutions [J]. *Frontiers of Opto electronics*, 2023.16 (1)
- [3] 孙瑜, 刘丰满, 薛海韵. 高速高密度光电共封装技术[J]. 中兴通讯技术, 2018, 24(4): 27-32.
- [4] Margalit N, Xiang C, Bowers S M, et al. Perspective on the future of silicon photonics and electronics[J]. *Applied Physics Letters*, 2021, 118(22): 1-10.
- [5] 卞玲艳, 曾艳萍, 蔡莹, 等. 大数据时代光电共封技术的机遇与挑战[J]. *激光与光电子学进展*, 2024, 61(9):0900006.
- [6] Minkenber C, Krishnaswamy R, Zilkie A, et al. Co packaged datacenter optics: opportunities and challenges [J]. *IET Optoelectronics*, 2021, 15(2): 77-91.
- [7] 欧祥鹏, 杨在利, 唐波, 等. 2.5D/3D 硅基光电子集成 技术及应用[J]. *光通信研究*, 2023(1): 1-16.

多通道高速光电微系统三维异构集成工艺仿真及开发

谢寒^{a)}，吕宏瑞^{a)}，唐娟，曾淑文，王攀，罗雪，徐鹤沼，武吉龙，吴罚^{b)}

联合微电子中心有限责任公司，重庆，401332

a) 同等贡献；b) 通讯作者：fa.wu@cumec.cn

摘要：三维异构集成是实现光电共封的重要手段，针对多通道高速光电微系统应用场景，本文对大马士革 RDL（重布线层）及高密度低寄生 TSV（硅通孔）等三维异构集成关键工艺结构进行了仿真研究，在此基础上完成了大马士革 RDL 工艺、高密度低寄生 TSV 工艺以及集成流程的开发。基于设计的多通道高速光电微系统制备了工艺样片并完成了相应的工艺测试及仿真结果对比。其中，样片上实际观测到 RDL、TSV 及 bump 结构形貌正常；铜互连大马士革 RDL 工艺能够满足最小线宽 $\leq 2\ \mu\text{m}$ 、最小间距 $\leq 2\ \mu\text{m}$ 、布线层数 ≥ 4 层的设计应用要求；TSV 工艺能够满足直径 $\leq 5\ \mu\text{m}$ ，深度 $\geq 50\ \mu\text{m}$ ，深宽比 $\geq 10:1$ 的设计要求；实际测得 TSV 传输损耗在 67GHz 内较为稳定，测试和仿真结果基本吻合。结果表明，在工艺仿真指导下研发的 RDL 工艺及 TSV 工艺满足当前的主流设计应用要求，在应用仿真优化带宽密度指标方面仍有进一步提升的空间。

关键词：三维异构集成工艺；高速光电微系统；工艺仿真与开发；大马士革 RDL；硅通孔 TSV

1 引言

光电微系统的概念自提出以来得到了快速发展，其探测精度高、处理速度快、传输通量大，与传统光电架构相比优势明显^[1]。随着集成电路的纳米制程工艺向物理极限逼近，为了提高集成度、降低功耗、提升性能，在二维芯片技术的基础上发展出了三维集成及封装技术^[2]。近年来，随着 5G 网络、物联网、数据计算中心的提出，对光电微系统的性能和功耗提出了更高的要求，三维异构集成架构的实现成为了延续摩尔定律的重要突破点^[3]。该领域制备工艺较为有代表性的有大马士革 RDL 工艺^[4]以及 TSV 工艺^[5]等，其尺寸较为灵活、传输特性良好、整合相对容易、未来发展潜力较高。

本文围绕三维异构集成的关键工艺，介绍了大马士革 RDL 工艺流程、高密度低寄生 TSV 工艺流程以及相应的集成工艺流程仿真及开发。同时，基于开发的成套工艺和主流的应用需求，设计了多通道高速光电微系统进行工艺测试验证，并完成了相应的工艺测试及仿真结果对比。其整体工艺集成架构图如图 1 所示。

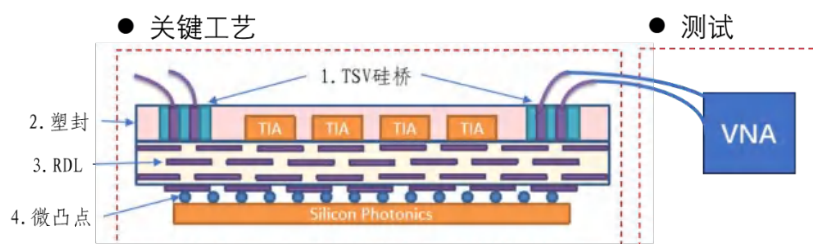


图 1 多通道光电微系统工艺集成架构图

2 关键工艺结构仿真及流程开发

三维异构集成关键工艺流程主要包括铜互连工艺中的大马士革 RDL 工艺流程以及高密度低寄生 TSV 工艺流程中的系列工艺，如 CMP 工艺、铜电镀工艺、深硅刻蚀工艺、介质层沉积工艺、背面露头工艺等。在三维异构集成关键工艺基础上，可以进一步进行集成工艺开发。对三维异构集成关键工艺结构进行仿真，对高效、准确地开发工艺流程具有重要的指导意义。本文选取部分核心工艺的仿真及流程开发进行介绍。

2.1 大马士革 RDL 工艺结构仿真及流程开发

光电微系统通常需要多层 RDL 层进行布线，因此在布线前，需要对叠构方式进行设计和优化，包括地平面的确定、阻抗匹配问题等，都需要采用 EDA 仿真工具进行优化设计。另外，面向高速光电微系统较高的工作带宽的需求（如 40GHz 以上），在现有工艺条件和设计规则下如何建立 RDL 布线方式，也需要通过建模进行仿真分析。为了满足翘曲度的要求，RDL 通常不能大面积的制作电源或地平面。因此，采用共面波导（CPW）设计以及较粗的地线代替回流平面的设计来构成微带线（Microstrip）结构，对相关方案进行建模仿真，以评估链路性能。

对硅衬底对 RDL 布局布线的影响展开仿真研究。如图 2 所示，RDL 模型与硅衬底间距离各不相同，受 RDL 和衬底间寄生电容影响的传输线特性阻抗各不相同，因此对于 CPW 等传输线结构，可以通过建模仿真来优化调整 G 和 S 的间距以获得合理的结构。此外，衬底采用高阻硅由于具有更小的损耗角正切而具有明显优势，也可以结合仿真来指导硅衬底的材料参数的选择。

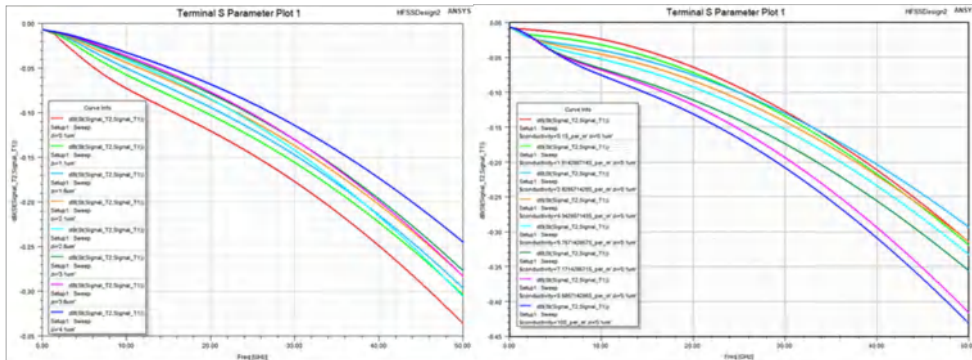


图 2 左图：CPW 传输线与衬底不同距离下 S21 参数随频率变化情况；

右图：不同衬底电导率下 S21 参数随频率变化情况

可以看到，距离衬底越远，由于寄生电容越小，因此 CPW 的传输损耗越小。右图展示了 CPW 传输线结构的 S21 参数随着衬底电导率变化曲线，传输损耗随着衬底由低阻变为高阻而逐渐减小。上述仿真结果为工艺参数优化和设计规则制定指明了方向。

依照仿真结果确定大马士革 RDL 工艺的部分工艺参数。首先对单层大马士革

RDL 工艺进行开发，其基本流程如图 3 (a) 所示。多层大马士革工艺可以理解为单层大马士革工艺的重复和堆叠，只是每层的线宽、线距和铜线的厚度不太相同。基于此项研究，设计合理的大马士革 RDL 工艺流程并完成多轮迭代优化。图 3 (b) 为三层铜大马士革工艺的具体流程。

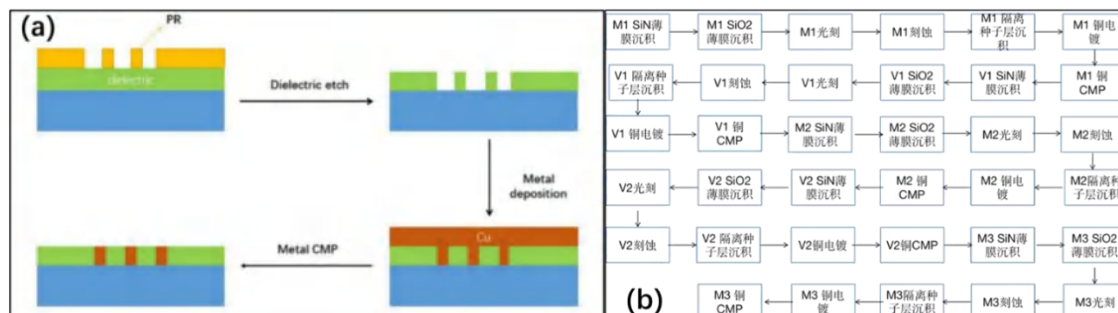


图 3 (a) 大马士革 RDL 工艺流程图；(b) 三层大马士革 RDL 工艺流程图

大马士革工艺的 Metal 刻蚀的 FA 结果如图 4 (a) 所示，可以看到刻蚀的边角 profile 平整度良好。Metal 电镀的 FA 结果如图 4 (b) 所示，可以看到经过多轮开发，铜填充整体良好，OM 显示表面无铜残留，表面 dishing 不明显。Via 的刻蚀和电镀等工艺开发结果如图 4 (c)、(d) 所示。

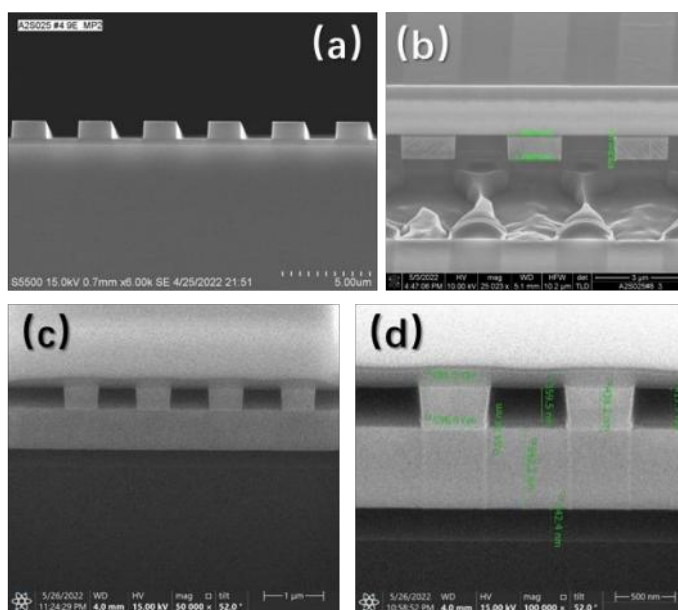


图 4 大马士革 RDL 工艺：(a) Metal ETCH, (b) Metal ECP, (c) Via ETCH, (d) Via ECP 短流程开发

大马士革的短流程工艺开发完成后，进行了铜互连工艺整合（5 层大马士革布线）。铜互连工艺整合的 flow 及 FA 切片如图 5 所示，实际的走线呈现顶面宽、底面窄的情况。从图中可以看到，已经满足大马士革 Cu 布线层数 ≥ 4 层的应用要求。

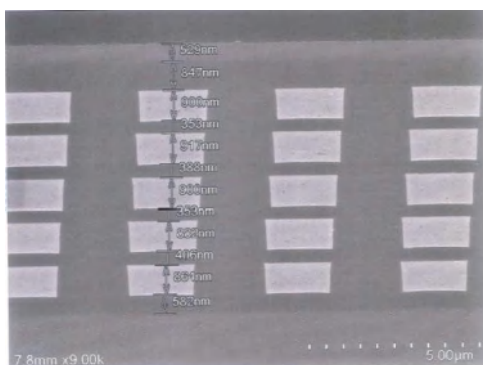


图 5 铜互连工艺整合的 flow 及 FA 切片

2.2 高密度低寄生 TSV 结构仿真及工艺开发

针对高密度低寄生 TSV 工艺的设计应用需求,开展 TSV 结构的建模仿真工作,其模型如图 6 所示^{[6][7]}。

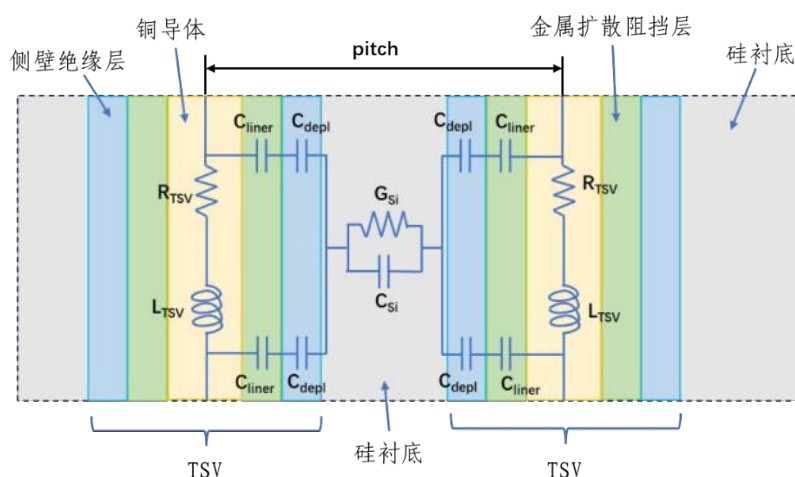


图 6 TSV 模型示意图

其中,寄生电容 C_{depl} 可以通过优化侧壁绝缘层介质工艺,增加绝缘层厚度的方式较小;而寄生电容 C_{Si} 可以使用高阻硅衬底代替常规的硅衬底得到优化。仿真数据表明(图 7),随着衬底电阻率由低阻($10 \Omega \cdot \text{cm}$)过渡到高阻($1000 \Omega \cdot \text{cm}$),40GHz 信号对应的 S21 参数从 -0.4dB 变化到 -0.12dB 左右。低阻衬底 TSV 的 -0.3dB 带宽为 10GHz,而高阻衬底 TSV 的 -0.3dB 带宽可以增大到 75GHz。随着 TSV 氧化层厚度由 $0.1 \mu\text{m}$ 增加到 $1.3 \mu\text{m}$,40GHz 信号对应的 S21 参数从 -0.42dB 变化到 -0.29dB 左右。 $0.1 \mu\text{m}$ 氧化层 TSV 的 -0.3dB 带宽仅为 2GHz,而 $1.3 \mu\text{m}$ TSV 的 -0.3dB 带宽可以增大到 40GHz。

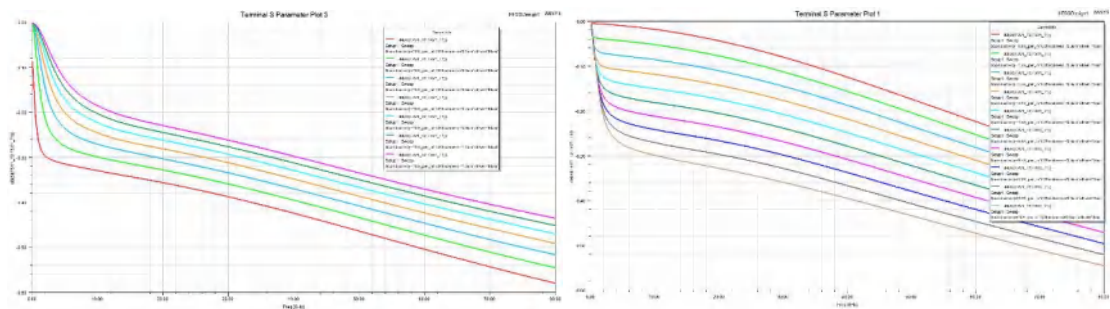


图 7 衬底电阻率和 TSV 侧壁绝缘层厚度对不同频率下 S21 参数的影响

以上述仿真结果作为依据,确定 TSV 的工艺参数并进行工艺流程的开发。TSV 刻蚀的 FA 结果如图 8 (a) - (d) 所示,可以看到 TSV 刻蚀的实际深度约为 $100.8 \mu\text{m}$ - $103.2 \mu\text{m}$, TSV 的顶部开口 CD 约为 $12.87 \mu\text{m}$, 底部开口 CD 约为 $11.8 \mu\text{m}$ 。

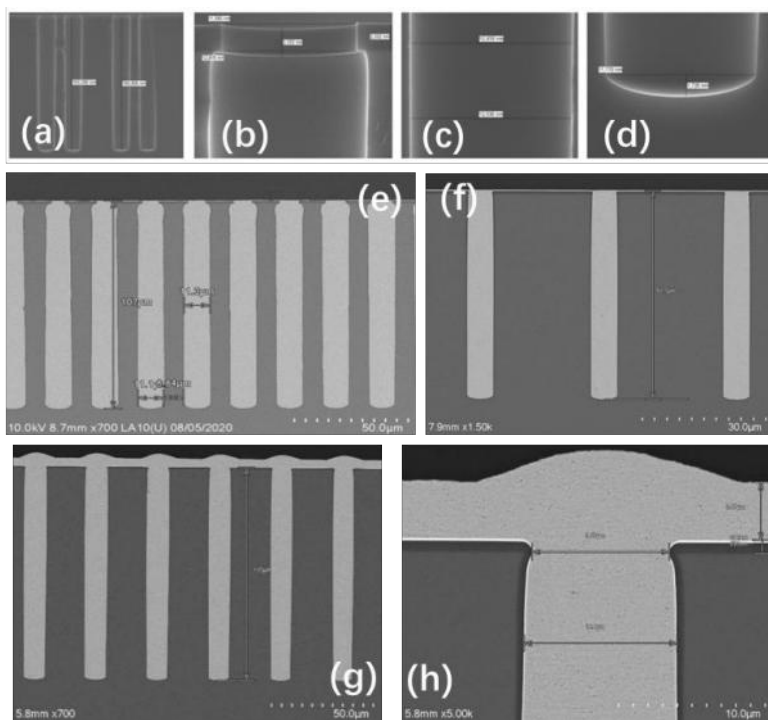


图 8 (a) - (d) :TSV 刻蚀单工艺开发; (e) (f): TSV 电镀单工艺开发;
(g) (h): 高密度低寄生 TSV FA 切片图

尺寸为 $10 \mu\text{m} \times 100 \mu\text{m}$ 和 $5 \mu\text{m} \times 50 \mu\text{m}$ 的 TSV 铜电镀的 FA 结果分别如图 8 (e)、(f) 所示,铜填充情况良好,没有明显的 Void。TSV 短流程工艺开发完成后,进行高密度低寄生 TSV 工艺开发。FA 切片测试如图 8 (g)、(h) 所示,可以看到 TSV 的深宽比 $\geq 10:1$, TSV 最小直径为 $5 \mu\text{m}$, TSV 中心间距为 $30 \mu\text{m}$,铜填充情况良好。

2.3 集成工艺开发

多通道光传输微系统采用多芯片扇出的方式实现 N 个电芯片和 N-1 个 TSV 桥芯片的集成,并最终通过倒装工艺实现与硅光芯片的三维集成。多芯片扇出集成实际工艺流程图如图 9 所示。

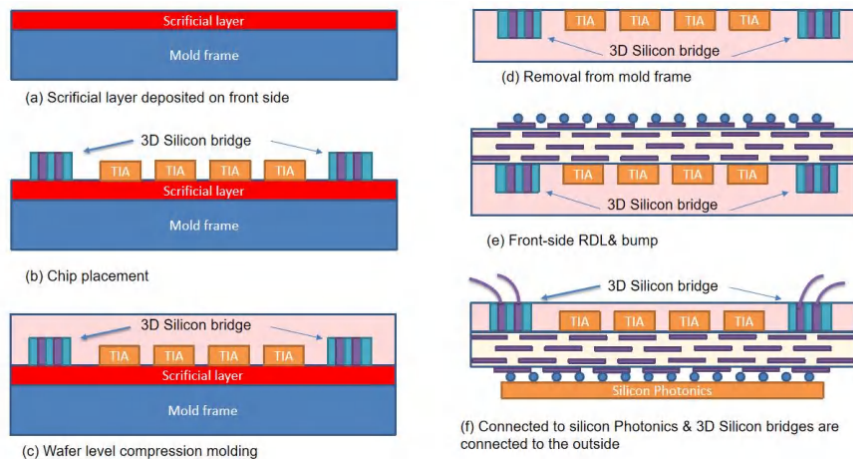


图 9 多芯片扇出工艺流程图

在 Mold frame 硅基板上进行牺牲层的沉积制备；然后将 N 个电芯片和 N-1 个 TSV 硅桥芯片在牺牲层上按照互连设计进行贴装，并进行晶圆级压塑，如图 10 所示，为真片塑封的光学照片。

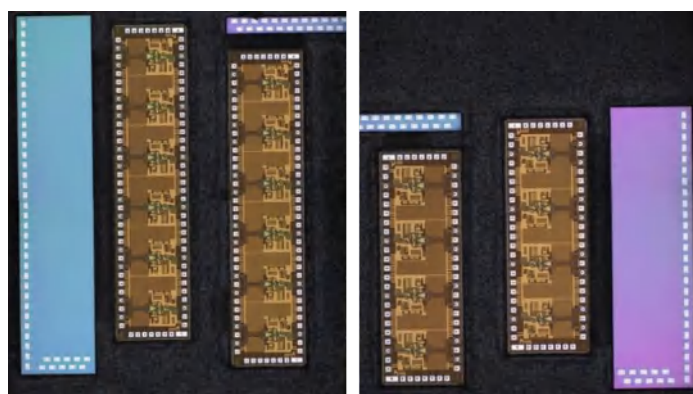


图 10 芯片装片塑封（真片）

在多芯片完成塑封埋入后，在芯片正面进行 PI 介质的沉积，固化后形成第一层介质层，并对接触焊盘进行光刻开窗工艺，如图 11 (a) 所示。对光刻漏出区域进行电镀并去除光刻胶，对种子层进行刻蚀，形成第一层 RDL，线宽 $15.26 \mu\text{m}$ ，如图 11 (b) 所示。四层 RDL 形成之后，在金属布线上需要植焊球的区域进行开窗，形成 bump，如图 11 (c) 所示。最后转移到减薄膜进行背面的减薄，并进行划片，如图 11 (d) 所示。

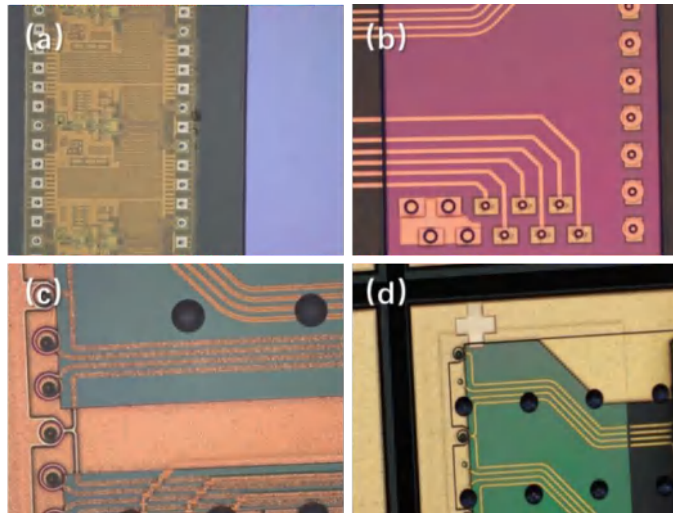


图 11 实际工艺样片：(a) 介质沉积，固化；(b) 刻蚀 RDL；
(c) 形成 bump；(d) 划片

3 工艺测试及仿真结果对比

在三维异构集成工艺开发过程中，对各个短流程的重复性、可靠性进行了验证，进行了多轮次投片。完成大马士革 RDL 工艺流程、高密度低寄生 TSV 工艺及集成工艺开发后，基于设计的多通道高速光电微系统进行了工艺样片的制备以及相应的工艺测试，对 RDL 及 TSV 的测试结果与仿真结果进行了对比验证。

3.1 ADI CD 测试

顶层 Metal 的 ADI CD 的监测结果如图 12 所示。从批量制备的 4 批样片中分别随机抽取 1 片、2 片、2 片、3 片，每片测量 5 个点；目标 Metal 的 CD 是 0.9 μm，设置 10% 的卡控标准 (0.81 μm-0.99 μm)。从图中可以看到，除了其中一片有异常，其余样片的所有测试点的 CD 均在卡控标准线以内，表明工艺稳定性良好。基于上述结果可以判定，当前的铜互连大马士革 RDL 工艺能够满足最小线宽 ≤2 μm，最小间距 ≤2 μm 的设计开发要求。

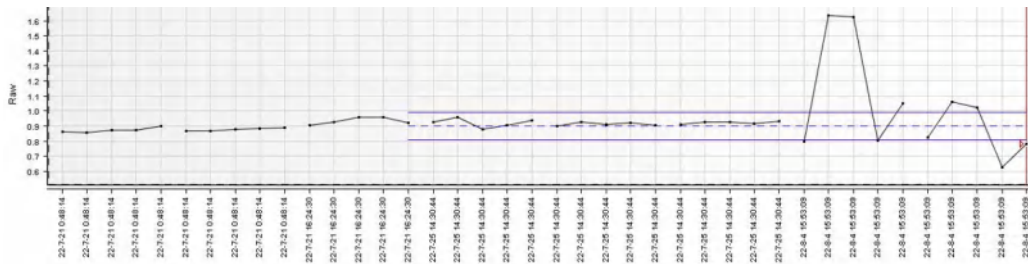


图 12 TM1 ADI CD

3.2 切片测试

对工艺样片进行 SEM 切片测试，检测 TSV 的结构完备性。切片示意图如图 13 所示。

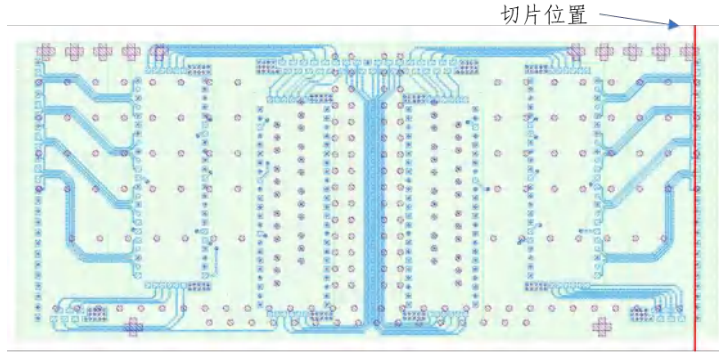


图 13 切片测试示意图：图中红线是切片的位置

SEM 切片如图 14 所示。正面再布线和 bump 植球 ($100\ \mu\text{m}$)、TSV 铜柱以及芯片背面增厚的位置和形貌均清晰可见，各结构对准情况良好；TSV 铜柱尺寸达到设计要求，一致性较好。

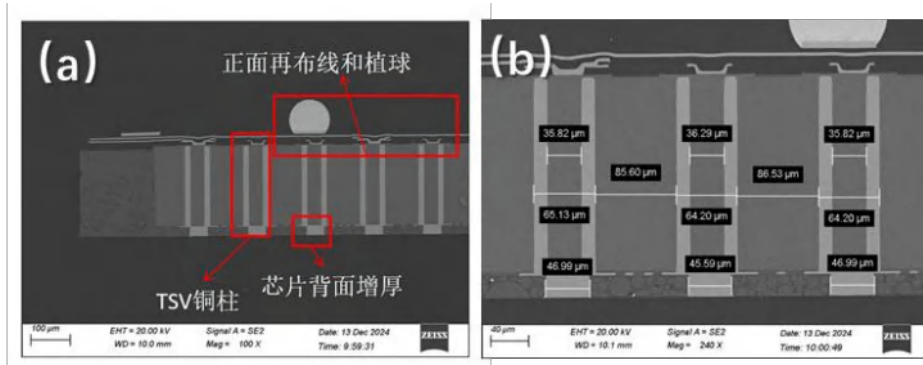


图 14 SEM 实测照片：(a) 各结构形态，(b) TSV 铜柱及尺寸

3.3 3D X-ray 测试

对工艺样片进行 3D X-ray 测试，检测 bump 的结构完备性。检测过程中 bump 的位置正确，未见明显形貌异常。典型切面成像如图 15 所示。

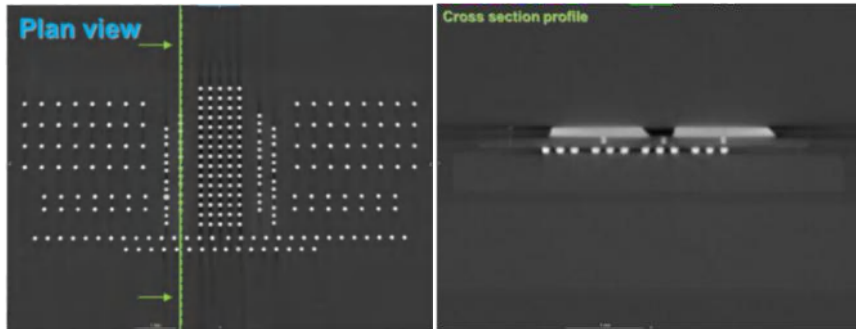


图 15 3D X-ray 的典型切面成像图

3.4 PCM TSV 射频测试

设计的高密度低寄生 TSV 特性测试结构及其特性测试结果的 S21 参数如图 16 所示，其中包括 100 个等间距 TSV ($30\ \mu\text{m}$ 间距)。该结果表明，被测的 TSV 阵列结构在 67GHz 以内的频率范围内均能保持较为稳定的传输损耗，满足设计要求；同时，随机选取的三个 die 的测试结果较为一致，说明工艺一致性较好。

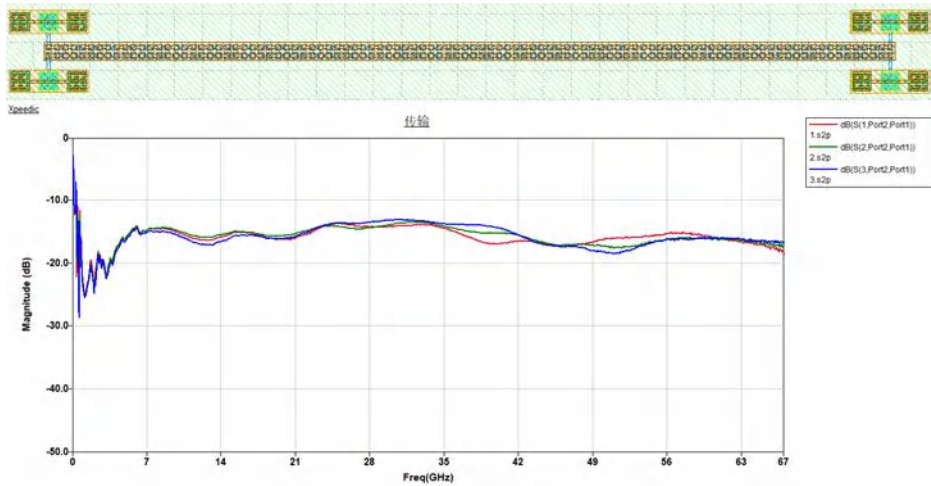


图 16 上图：低损耗 TSV 特性测试结构图；下图：低损耗 TSV 特性测试 S21 参数结果

针对单个 TSV 结构进一步进行射频测试，同时按照实际的工艺参数进行仿真。同时比较测试数据和仿真数据的典型去嵌结果，其误差为 5.7%，吻合情况较好，如图 17 所示。

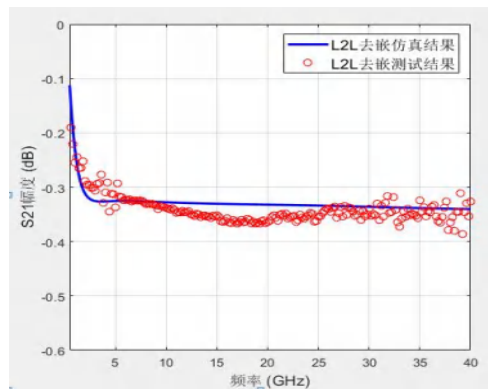


图 17 TSV 射频测试与仿真对比

3.5 RDL 射频测试

对多个线型、线长、线宽的 RDL 布线进行测试验证，将测试结果与仿真结果进行对比，其典型结果如图 18 所示。

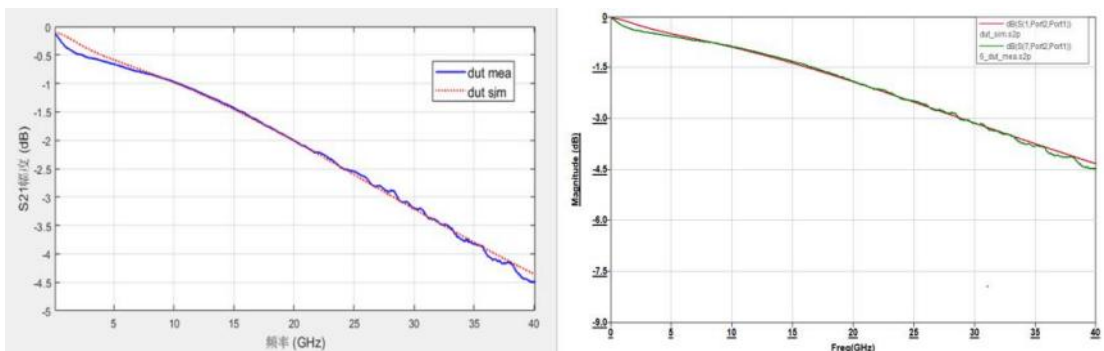


图 18 左图：长 200um 宽 9um 微带线仿真和测试的 S21 参数；

右图：长 100um 宽 9um 带状线仿真和测试的 S21 参数

长200 μm 、宽9 μm 的微带线, 仿真和测试结果平均误差为5.35%; 长100 μm 、宽9 μm 带状线, 仿真和测试结果平均误差为6.00%。仿真数据与测试数据基本吻合。

4 结论

本文针对多通道高速光电微系统应用场景下的大马士革 RDL (重布线层) 及高密度低寄生 TSV (硅通孔) 等三维异构集成关键结构进行了仿真研究, 在此基础上完成了大马士革 RDL 工艺、高密度低寄生 TSV 工艺以及集成流程的开发。基于设计的多通道高速光电微系统制备了工艺样片并完成了相应的工艺测试及仿真结果对比。其中, 铜互连大马士革 RDL 工艺达到最小线宽 $\leq 2 \mu\text{m}$ 、最小间距 $\leq 2 \mu\text{m}$ 、布线层数 ≥ 4 层的应用要求的设计指标; TSV 工艺能够满足直径 $\leq 5 \mu\text{m}$ 、深度 $\geq 50 \mu\text{m}$ 、深宽比 $\geq 10:1$ 的设计要求。样片上实际观测到 RDL、TSV 及 bump 结构形貌正常, 实际测得 TSV 传输损耗在 67GHz 内较为稳定; 典型结构的测试和仿真数据基本吻合, 结果平均误差约为 5%至 6%。结果表明, 在工艺仿真指导下研发的 RDL 及 TSV 工艺满足当前的主流设计应用要求, 在应用仿真优化带宽密度指标方面仍有进一步提升的空间。

参考文献

- [1] Nan Marie Jokerst, Martin A. Brooke, Sang-Yeon Cho, Dae-Ik Kim, Mikkel Thomas, and Suzanne Fike, "Optoelectronic Microsystems Integration," *Optics & Photonics News*, 2003, 14(2), 26-31, doi: 10.1364/OPN.14.2.000026.
- [2] D. K. Nayak, S. Banna, S. K. Samal and S. K. Lim, "Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs," 2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Rohnert Park, CA, USA, 2015, pp. 1-2, doi: 10.1109/S3S.2015.7333538.
- [3] G. Murali and S. K. Lim, "Heterogeneous 3D ICs: Current Status and Future Directions for Physical Design Technologies," 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 2021, pp. 146-151, doi: 10.23919/DATE51398.2021.9474057.
- [4] Lau, J., P. Tzeng, C. Lee, C. Zhan, M. Li, J. Cline, K. Saito, et al. 2013. "Redistribution Layers (RDLs) for 2.5D/3D IC Integration." *IMAPSource Proceedings 2013* (1): 434-41, doi: 10.4071/isom-2013-WA12.
- [5] M. Motoyoshi, "Through-Silicon Via (TSV)," in *Proceedings of the IEEE*, 2009, vol. 97, no. 1, pp. 43-48, doi: 10.1109/JPROC.2008.2007462.
- [6] Salah, K., Ismail, Y., El-Rouby, A. "TSV Modeling and Analysis," In: *Arbitrary Modeling of TSVs for 3D Integrated Circuits. Analog Circuits and Signal Processing*. Springer, Cham., 2015, doi: 10.1007/978-3-319-07611-9_3.
- [7] D. M. Jang et al., "Development and Evaluation of 3-D SiP with Vertically Interconnected Through Silicon Vias (TSV)," 2007 Proceedings 57th Electronic Components and Technology Conference, Sparks, NV, USA, 2007, pp. 847-852, doi: 10.1109/ECTC.2007.373897.

用于 2.5D/3D 射频微系统的硅转接板中互连结构传输

特性建模仿真

陈鹏, 田斯劼**, 王攀, 张文亚, 唐娟, 阮雨杉, 袁恺, 闵成彧

**sijie.tian@cumec.cn

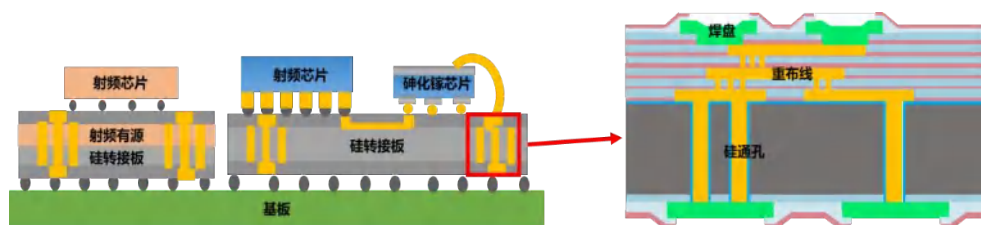
联合微电子中心有限责任公司, 重庆 401332

摘要: 基于硅转接板的 2.5D/3D 先进封装技术被视为一种应对射频微系统小型化、异质异构集成化发展需求的有效手段。用于 2.5D/3D 射频微系统的硅转接板需要提供高性能、小尺寸的互连结构。本文从工程实际出发, 对射频硅转接板中的重布线 (RDL) 和硅通孔 (TSV) 互连结构传输特性做建模, 通过两个测试结构仿测对比验证了建模方法的有效性, 仿真分析了大马士革布线金属密度控制、准同轴冗余结构等因素对硅转接板互连结构传输特性的影响, 以支撑 2.5D/3D 射频微系统的工程设计与应用。

关键词: 2.5D/3D 射频微系统; 硅转接板; 互连结构; 先进封装; 三维集成

1 引言

基于硅转接板的 2.5D/3D 先进封装技术被视为一种应对射频微系统小型化、异质异构集成化发展需求的有效手段[1-4]。用于 2.5D/3D 射频微系统的硅转接板需要提供高性能、小尺寸的互连结构, 且与数字硅转接板不同, 射频硅转接板一般重布线 (RDL)、硅通孔 (TSV) 密度较低, 对关键互连结构进行针对性建模仿真以优化传输特性具有设计时间与版图空间上的可行性[5-7]。本文从工程实际出发, 对射频硅转接板中的 RDL 和 TSV 互连结构传输特性做建模仿真, 以支撑 2.5D/3D 射频微系统的工程设计与应用。



(a) 2.5D/3D 射频微系统示意 (b) 射频硅转接板互连结构示意

图1 用于2.5D/3D射频微系统的硅转接板互连结构示意图

2 设计与工艺信息

对设计与工艺信息的了解是开展建模仿真工作的前提。射频硅转接板内部互连结构 (RDL、TSV) 设计的工艺规则由硅转接板晶圆厂工艺设计套件 (PDK)

给出。面向射频微系统应用，联合微电子中心有限责任公司（以下简称 CUMEC）开发了多种硅转接板工艺平台，以满足客户多样化的工程需求。对于低成本大尺寸高密度硅基板需求，CUMEC 可提供大马士革厚铜布线一倍光罩硅转接板（使用硅基无机介质，微米级布线，设计尺寸最大可达八英寸晶圆）。相对于聚酰亚胺（PI，有机介质）布线工艺，大马士革布线工艺通常支持的布线与布孔精度与密度更高，兼容现有硅基工艺，如与硅芯片热机械参数匹配、结合硅帽后气密性更好；但成本较高，且布线设计工艺约束较多，如和射频集成芯片设计一样需要通过填充金属（dummy metal）和开槽镂空（dummy slot）来控制金属密度。对于射频-数字一体化、微波光子等领域的高性能超高密度混合集成需求，CUMEC 可提供大马士革厚铜布线五倍光罩硅转接板（线宽/间距 $\geq 0.9\ \mu\text{m}$ ，铜厚 $2\ \mu\text{m}$ ），和大马士革-PI 混合布线硅转接板，后者结合了针对数字互连的超高布线密度的大马士革薄铜布线（线宽/间距 $\geq 0.4\ \mu\text{m}$ ，铜厚 $0.9\ \mu\text{m}$ ）与射频性能较优的大马士革/PI 厚铜布线。

CUMEC 硅转接板工艺支持的典型 TSV 规格为 $5*50\ \mu\text{m}$ 、 $10*100\ \mu\text{m}$ 和 $20*200\ \mu\text{m}$ ，且均支持使用高阻硅衬底。为满足射频硅转接板对 TSV 电学性能的高要求，上述 TSV 均为实心填充铜，不采用中空或间隙结构，不使用多晶硅或钨等其他材料填充[8]。由于射频集成微系统的先进封装工艺流程通常不使用塑封加固[1,5]，为满足加工过程对组件机械性能的要求，射频硅转接板会使用 $\geq 200\ \mu\text{m}$ 晶圆厚度，即 $20*200\ \mu\text{m}$ 规格 TSV，这种大尺寸实心深孔对 TSV 铜填充工艺是种挑战。图 2 为 CUMEC 加工的某产品 $20*200\ \mu\text{m}$ TSV 结构扫描电镜（SEM）结果[9]，可见 TSV 无空洞、侧壁光滑垂直。这意味着 TSV 可以近似为圆柱结构，简化了建模仿真的复杂度。未来 CUMEC 硅转接板工艺平台可组合硅基集成无源器件（IPD）和超高电容密度的深刻槽硅电容（DTC）工艺，实现互连结构与无源器件的深度集成；可融合硅基有源工艺，提供用于 3D 堆叠的有源 TSV 射频芯片。

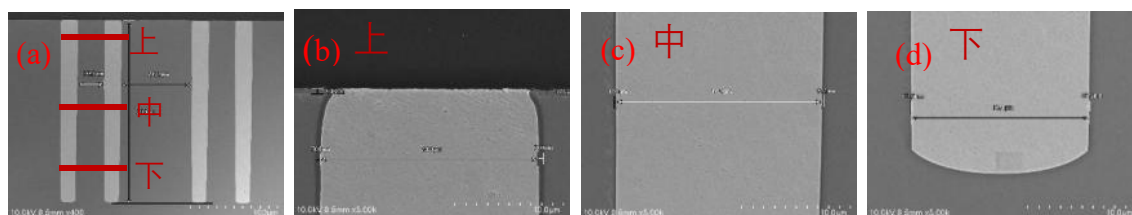


图2 CUMEC某产品 $20*200\ \mu\text{m}$ TSV结构SEM结果[9] 图片已获得授权

射频硅转接板的设计还需要综合考虑内部互连结构与外部互连结构，后者一般为凸点，相应地会有凸点下金属（UBM）等结构。以 2.5D 射频微系统为例，硅转接板正面微凸点和背面凸点，会分别通过倒装工艺与上层芯片和下层基板键合，如图 1（a）所示。除开焊料球和铜柱锡帽等数字领域常用的凸点类型，射频硅转接板还会使用金凸点，比如为了兼容某些原设计为使用引线键合的射频芯

片，这些芯片的焊盘分布可能不兼容焊料球倒装工艺。特别是，当被用作高密度硅基板时，射频硅转接板还会面临兼容引线键合、导电胶粘合或表贴（SMT）等工艺的复杂需求。为了提高先进封装工艺流程可行性和射频微系统可靠性，有时不得不考虑底填加固，因此设计时需要评估底填对硅转接板互连结构以及其他组件电学性能的影响。在 2.5D/3D 射频微系统设计方案自顶向下分解的过程中，功能结构与工艺流程会协同做概要设计；各组件自下向上做详细设计实现过程中，建模仿真需要考虑组件内部的同时兼顾组件间交界面的影响。

3 方法

3.1 建模仿真方法

对硅转接板中具体互连结构建立物理模型后，业界一般借助 3D 电磁有限元仿真工具如 Ansys HFSS，数值离散频域求解全波麦克斯韦方程组，得到表征传输特性的散射（S）参数。因此，以下结合仿真术语来介绍互连结构传输特性建模方法。

首先，需要确定互连结构材料参数与表面特性。硅转接板互连结构中，金属导体包括铜 RDL、TSV 与可能的铝焊盘，主要关注电导率；有损介质包括绝缘介质二氧化硅、氮化硅等，主要关注介电常数和损耗角正切；对于互连结构仿真，半导体硅衬底可以视为一种有损介质，通过介电常数和电导率表征。此外，通过 Huray 模型描述金属导体的表面粗糙度。值得注意的是，对于具体工艺平台，应该通过严谨的实验测量来校准上述数据，这个工作超出了本文的讨论范围。因此本文仿真使用的是各材料参数典型值，并忽略了互连结构的表面粗糙度，具体如表 1 所示。其中使用高阻硅衬底，忽略了二氧化硅的损耗正切。

表1 本文仿真使用的互连结构材料参数

	相对介电常数	相对磁导率	电导率 (S/m)
二氧化硅	4.1	1	0
硅	11.9	1	0.04
铜	1	1	58000000
铝	1	1	38000000

其次，需要确定边界条件，即仿真端口。对于实际结构，需要结合理论分析、迭代仿真甚至测试验证来确定端口设置，是建模仿真工作的关键环节。对于适用于典型传输线模型的 RDL 结构，可以理论分析其传输场型，使用波端口激励特定模式。对于两端垂直连接 RDL 的 TSV 结构，也可以先分析典型传输线模型的适用性。但对于工程实际使用的 TSV 设计，比如图 1（b）所示的两端水平连接 RDL 传输线的 TSV 结构，不适合将 TSV 部分孤立出来做仿真，一种常用做法是在 TSV 两端保留适当长的水平连接 RDL 传输线，在 RDL 一端施加波端口做仿真，综合去嵌得到 TSV 部分的传输特性[7,10]。实践中由于 TSV 上下的水平连

接结构往往不严格对称，只能近似去嵌；如果去嵌方法有效，在离 TSV 足够远处的 RDL 一端施加波端口抑或集总端口，去嵌后 TSV 部分传输特性应无显著差异。

考虑外部互连结构的存在，硅转接板混合互连结构（包含 TSV，RDL）会以表面焊盘作为两端，通常在焊盘处施加集总端口。根据焊盘附近的具体结构，考虑使用类同轴（coax，适用于凸点、引线等情况）抑或边缘（edge）集总端口，而端口阻抗则根据实际情况需要来确定。比如为了便于测试验证，一般会设计传输线结构连接地-信号-地（GSG）测试焊盘的测试结构。对于这种测试结构仿真，可以考虑使用适当尺寸的 $50\ \Omega$ coax 集总端口来近似 GSG 探针扎到测试焊盘上的作用。

最后，需要根据目标几何结构，选择合适的几何建模与仿真执行流程。通常使用 HFSS 3D Layout 导入实际硅转接板版图，建立互连结构的三维几何模型。但对于实际复杂互连结构，dummy metal、dummy slot、过孔阵列等大量精细图形会导致仿真计算量过大，此时需要考虑做相应简化，评估简化几何对仿真结果的影响。此外，对于层叠简单的互连结构，也可以使用 HFSS CAD 建立几何模型并执行参数化仿真；对于不适合用层叠描述的复杂结构，如非圆柱结合的 TSV 或带缺陷结构，则无法使用 HFSS 3D Layout，只能使用 HFSS CAD 建立几何模型。在确定上述建模方法后，结合自动化版图设计与仿真脚本，可以高效地对互连结构做批量仿真分析。

3.2 测试结构仿真

为了验证上述建模仿真方法，本文对 CUMEC 某射频产品两个测试结构做了仿测对比。测试结构 1 为接地共面波导（GCPW），测试结构 2 为跨硅微带线，二者均为包含 TSV、RDL 和测试焊盘的混合互连结构，并且考虑金属密度控制做了相应的 dummy metal 和 dummy slot 设计。使用表 1 的材料参数和表面特性，并使用适当尺寸的 $50\ \Omega$ coax 集总端口来近似 GSG 探针扎到测试焊盘上的作用，在 HFSS 3D Layout 中自动化导入加工版图、仿真设置与求解。

3.2 RDL 结构仿真

【进行中】 RDL 结构设计采用大马士革厚铜布线五倍光罩硅转接板工艺，采用 GCPW 线型，为实现 $50\ \Omega$ 特性阻抗，设计层叠与线宽线距。为研究大马士革布线金属密度控制对传输特性的影响，分别采用不同的 dummy metal 和 dummy slot 模式（dummy pattern）版图做仿真，如图 3 所示。

图3 RDL结构仿真模型示意图

3.3 TSV 结构仿真

【进行中】 TSV 结构设计采用 $10*100$ 规格。先对比 GSG 准同轴 TSV 与阵

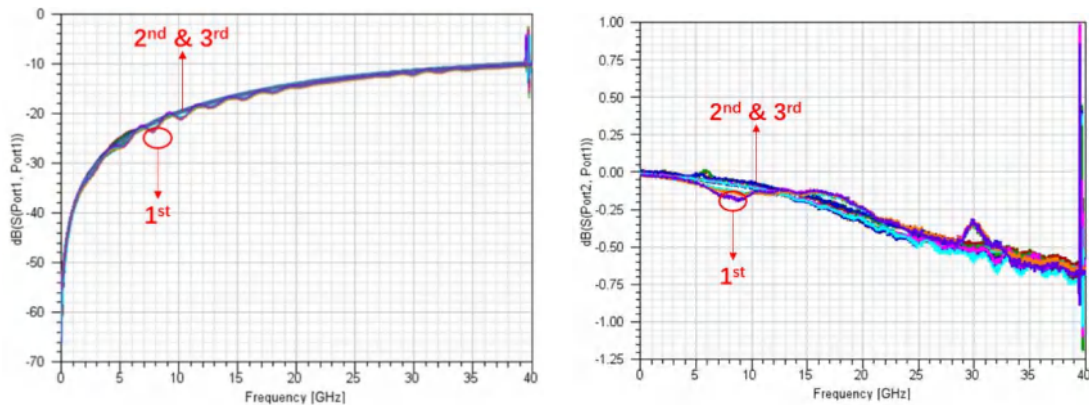
列 TSV，然后对于准同轴结构对比不同冗余 TSV 的影响，如图 4 所示。

图4 TSV结构仿真模型示意图

4 结果与讨论

4.1 测试结构仿测对比结果

为了验证建模仿真方法，本文对 CUMEC 某射频产品两个测试结构做了仿测对比。首先实测传输特性，使用矢量网络分析仪及射频探针得到测试结构的 S 参数表征传输特性，并结合简单可靠性试验处理进行重复测试，确保测试的有效性。以下以测试结构 1 的实测结果为例，做简单分析。如图 5 (a) 与 (b) 所示，对晶圆上 10 个点位测试结构 1 各做三次测试，实测回损与插损偏差在工程容差范围内，表明了高度的片内一致性。40 GHz 处毛刺为测试时环境射频干扰导致，后两次测试由于测量环境更一致，测试结果基本重合，这表明经如图 1 (b) 所示可靠性试验处理后，测试结构无失效且无传输特性劣化。因此，该 GCPW 传输线实测未去嵌传输特性代表值如图 1 (c) 所示。经过细节性设计，大马士革硅基传输线的传输特性将可以进一步优化。



(a) 三次测试的回损对比

(b) 三次测试的插损对比

测试序号	测试前状态
1 st	晶圆流片完成后，未处理
2 nd	5次-35°C至85°C的温度冲击试验处理后
3 rd	30次-55°C至125°C的温冲温度冲击试验处理后

(c) 三次测试前状态

	传输特性	
	@20GHz	@40GHz
IL	~0.7 dB/mm	~1.5 dB/mm
RL	~14.5 dB	~10.1 dB

(d) 实测未去嵌传输特性代表值

图5 CUMEC某射频产品测试结构1实测传输特性

使用本文方法对两个测试结构做建模仿真，仿测对比结果如图 6 所示，可见仿测结果基本一致，均在工程容差范围内，验证了本文建模仿真方法的有效性。相较于测试结构 1，测试结构 2 由于跨硅衬底、阻抗不连续点更多等缘故，传输特性较差，仿测结果偏差也更大。后续若通过实验测量校准材料参数和表面特性，综合去嵌算法，可以进一步降低仿真误差。

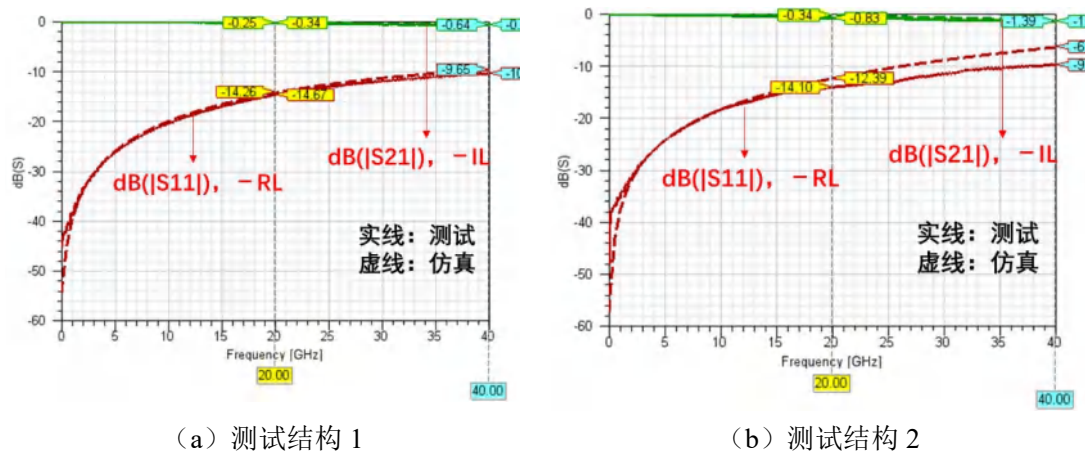


图6 两个测试结构传输特性的仿测对比结果

4.2 RDL 结构仿真结果分析【进行中】

4.3 TSV 结构仿真结果分析【进行中】

5 结论【进行中】

参考文献

- [1] 吴磊. 射频系统先进封装技术研究进展[J]. 电子元件与材料, 2024, 43(9): 1053-1062.
- [2] 王豪杰, 崔碧峰, 王启东, 等. 射频系统 2.5D/3D 封装结构的研究进展[J]. 电子与封装, 2021, 21(9): 090205.
- [3] 龙乐. 射频系统封装的进展现状和影响[J]. 电子与封装, 2011, 11(7): 000905.
- [4] 夏晨辉, 王刚, 王波, 明雪飞. 用于 5G 通信的射频微系统与天线一体化三维扇出型集成封装[J]. 电子学报, 2023, 51(6): 1572-1580.
- [5] 朱瀚翔. 硅基叠层射频封装多物理场协同分析技术研究[D]. 北京: 中国科学院微电子研究所, 2021.
- [6] 田斯劼, 陈鹏, 王攀, 袁恺, 闵成彧. 一种高频 TSV 硅转接板及其设计方法: 中国, 2024113144669[P]. 2024-09-20.
- [7] Wang, M.; Ma, S.; Jin, Y.; Wang, W.; Chen, J.; Hu, L.; He, S. A RF Redundant TSV Interconnection for High Resistance Si Interposer[J]. Micromachines, 2021.
- [8] 王根旺, 等. 多芯粒 2.5D/3D 集成技术研究与应用现状[J/OL]. 电子科技大学学报, 205:1-13.
- [9] 闵成彧, 等. 基于 TSV 硅桥的光电微系统三维异构集成设计与仿真研究. (已投稿至本次会议).
- [10] Yen, H., 等. TSV RF de-embedding method and modeling for 3DIC. 2012 SEMI Advanced Semiconductor Manufacturing Conference, Saratoga Springs, NY, USA, 2012, pp. 394-397.

硅基异构集成射频微模组仿真设计

孙志强¹, 高阳¹, 张文锋¹, 罗明¹, 孙岩¹, 卢茜¹, 葛菊祥¹,

刘江洪¹, 杨思怡¹, 汪昌思¹, 叶雨农²

宽带微波电路高密度集成四川省工程研究中心¹

中国电子科技集团公司信息科学研究院²

摘要: 面向未来信息系统对微型化射频组件的迫切需求, 本文提出了一种基于硅基异构集成的射频微模组。目前射频接收机的前端模块和变频通道主要采用微波混合集成电路 (Microwave Hybrid Integrated Circuit, MHIC) 的工艺形式, 体积和重量偏高, 在一些小型化、轻量化应用场景很受限。本设计采用先进三维堆叠工艺, 主要通过球栅阵列 (Ball Grid Array, BGA) 将扇出 (Fan-out) 模组垂直互联到硅通孔 (Through Silicon Via, TSV) 基板, 且其余裸芯片间及其与表层微带传输线间通过金丝键合实现互联, 硅基板底部植球焊接到测试母板上实现完整性能验证。在单模组内异构集成砷化镓 (Gallium Arsenide, GaAs) 工艺低噪声放大器 (Low Noise Amplifier, LNA)、混频器、开关预选滤波器多功能芯片及 Fanout 微模组, 实现 Ku 波段以内射频信号下变频到 C 波段以下中频的完整接收链路功能。传统的仿真设计方法主要聚焦在局部结构的场性能优化, 链路模型精度不够, 仿真效率低, 而本文通过从局部过渡结构的场仿真到整板电磁性能评估, 从芯片实测数据级联模型分析到场路联合谐波平衡仿真研究, 并综合考虑核心散热器件的影响, 构建了电热耦合多物理场模型, 提高了模型的准确度和仿真效率。本设计联合三维电磁全波仿真 (High Frequency Structure Simulator, HFSS)、电路系统分析 (Advanced Design System, ADS) 及热-电耦合仿真 (FloEFD), 优化了射频微模组的变频增益 ($\geq 22\text{dB}@18\text{GHz}$) 及输入 $P_{-1\text{dB}}$ ($\geq -10\text{dBm}@18\text{GHz}$), 简化了传统组件模块通过实物多次迭代设计的研制周期。实测结果表明, 本文提出的硅基异构集成射频微模组在全频段内变频增益达 22 dB, 实物尺寸仅 $20\times 20\times 1.68\text{ mm}^3$, 且平面尺寸为目前 200 um 厚度硅基板含双面焊盘加工的工艺极限状态。在性能指标相同的情况下, 较传统微组方案体积尺寸大幅缩小, 高度降低至 20%, 面积缩小至 15%。本文工作为变频通道的系统级芯片集成提供了高可靠性解决方案。

关键词: 硅基异构集成; 宽带变频; 多物理场仿真; 三维堆叠

1 前言

目前射频、微波、毫米波电路系统需要变频通道具备至少 L 波段到 Ku 波段连续带宽与高功率耐受性^[1-5]，而传统的变频组件设计从 MHIC 工艺架构逐步发展到基于低温共烧陶瓷（Low Temperature Co-fired Ceramic, LTCC）、高温共烧陶瓷（High Temperature Co-fired Ceramic, HTCC）、直接覆铜（Direct plated copper, DPC）基板搭建的射频微系统（System in Package, SiP），主要采用多芯片分段拼接的单层平面互联方式^[6-8]，导致体积膨胀和互连损耗大的问题^[9-10]。同时，随着印刷电路板（Printed Circuit Board, PCB）从简单的单面板、双面板发展到多层板，其中高密度互联（High Density Interconnect, HDI）技术在多层高频板的应用也逐步增加，基于国产复合基板的小型化变频组件也开始实现应用^[11-12]。总之，传统的变频模块主要依赖分立器件和板级集成，腔体内射频同轴电缆与低频控制线交叉缠绕，存在体积大、功耗高及高频性能受限（传输插损逐级增加、多频段兼容性差）等技术瓶颈^[3-11]。

硅基异构集成技术通过融合硅基互补金属氧化物半导体（Complementary Metal Oxide Semiconductor, CMOS）工艺、III-V 族化合物宽禁带半导体（如第二代砷化镓、磷化铟、第三代氮化镓等）器件优良的高频特性、重布线层（Re-Distributed Layer, RDL）技术、低损耗小型化 TSV 技术的优势，被视为突破上述限制的关键技术路径^[13-15]。

硅基异构集成技术当前面临的关键技术挑战是：电磁性能耦合分析、高密度集成空间的热管理和结构微应力评估^[16-20]。在电磁性能方面，异质材料界面阻抗失配导致信号反射和谐波失真；在热管理方面，多功能单元功率密度集中引发的热积累造成电性能频率漂移和性能波动；在结构方面，三维堆叠中 TSV 与键合界面的机械微应力降低器件焊接的可靠性等。

在高密度集成的发展趋势下，亟需高效率、高精度、多场耦合综合考虑的建模方法，但当前的研究方法多聚焦单一领域优化，具体代表性的局限有主要如下 3 点：

1) 在电路设计方面，如文献[5]，基于传统的链路指标计算和多芯片组件（Multi-Chip Module, MCM）工艺，提出一种超宽带、多通道、低噪声、低灵敏度、大动态范围接收前端的设计方案，样机在噪声系数和开关隔离度性能方面有所提升。并通过设计仿真和实物样机的研制测试，验证了该方案的可行性，实物尺寸为 53 mm×115 mm×9 mm。但总的体积还是偏大，老的工艺技术，忽略链路的电磁耦合效应、封装寄生效应。

2) 在工艺集成方面，如文献[21]，设计并制备了一款基于微电子机械系统（Micro Electro Mechanical Systems, MEMS）工艺的小型化 X 波段硅基射频微系

统,该系统实现了硅基片上有源及无源芯片的异构集成,通过在硅基片上直接集成 MEMS 双滤波器及紧凑型布线分布实现了小型化,通过系统内部的设计优化解决芯片间级联匹配带来的寄生问题。虽在尺寸 $8.5\text{ mm}\times 11\text{ mm}\times 1\text{ mm}$ 的空间内实现了带外抑制 $> 80\text{ dBc}@7.5\text{-}9\text{ GHz}$,中频输入端口驻波比 ≤ 1.7 ,射频输出端口驻波比 ≤ 1.4 的性能指标要求,但 $8\sim 8.5\text{ GHz}$ 的工作频段较窄,需在高密度集成的工艺条件下提升射频性能。

3) 在仿真方法方面,如文献[22],利用硅基三维异构集成工艺设计一款射频微系统,且根据工艺特征以及产品在多物理场中的耦合现象,建立了一种面向硅基三维异构集成工艺射频微系统的多物理场一体化仿真流程,但缺乏对电磁-热-应力多物理场协同作用的量化建模分析,无法满足宽带射频系统的全频段性能指标评估需求,导致设计与实测性能偏差大。

2 本文贡献

针对上述问题,本文采用三维堆叠架构的工艺技术,实现 LNA、开关预选滤波器、混频器、Fanout 微模组等下变频链路器件和 TSV 基板的异构集成,达到了在密度空间中实现高性能的射频通道需求,本文的具体贡献有如下 3 部分。

2.1 异构集成架构

Fanout 微模组通过植 $100\text{ }\mu\text{m}$ 直径小球垂直互联到硅基板表面的中频链路中,相较于传统射频裸片间的金丝键合,该方式提高了焊接可靠性,增强了中频模组于射频链路间的隔离度。其余下变频通道中的射频芯片间采用金丝热声焊接 (Wire Bonding, WB),支持 Ku 波段以内射频信号的传输。

2.2 多物理场仿真框架

建立 HFSS-ADS-FloEFD 联合仿真流程,如图 1 所示。通过 HFSS 提取输入输出过渡结构的 S 参数,并将其带入基于 ADS 构建的谐波平衡电路系统模型,通过 FloEFD 分析微模组中核心散热器件的热应力分布,并将评估结果反馈给三维场模型优化布局设计,将两者综合迭代反馈给 ADS 模型优化电路系统性能。

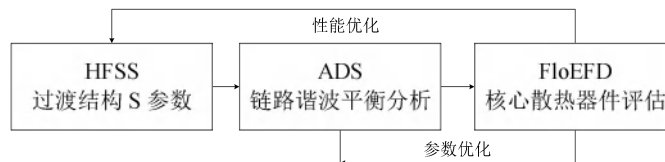


图 1 HFSS-ADS-FloEFD 联合仿真流程

2.3 实测验证

制备了尺寸为 $20\times 20\times 1.68\text{ mm}^3$ 的射频微模组,实现 Ku 波段以内射频信号下变频到 C 波段以下中频的完整接收链路功能,在全频段内的变频增益达 21 dB ,验证实测与仿真指标变频增益的误差 $< 4.89\%$ 。

3 异构集成架构设计

3.1 电路系统架构

本文采用一次下变频的电路系统架构，其中射频部分电路实现了对接收到的 Ku 波段以内射频信号的频率选择、放大功能，本振电路对需要混频的本振信号进行放大、滤波处理，两者通过混频器实现射频到中频信号的频率变换。最后的中频电路有滤波、衰减、放大等功能，实现数字信号处理之前的 C 波段以下中频信号处理，具体的电路系统架构如图 2 所示。

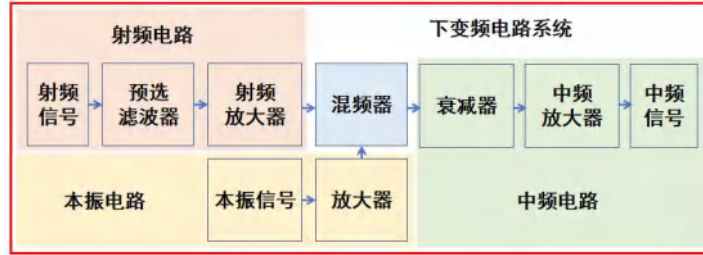


图 2 变频微模组的电路系统架构

3.2 关键工艺技术

本文提出的射频微模组三维异构集成的堆叠形式如下，首先将植了直径 100um SAC305 小球的 Fanout 微模组焊接到 TSV 基板的正面，然后用导电胶将 GaAs 裸片粘接到基板同一侧，且链路中芯片间及其与表层微带传输线间通过金丝键合实现互联。最后在 TSV 基板的背面植 450 um 锡铅焊球，便于在射频 PCB 母板上进行使用验证。这种先进的工艺集成方式极大的提高了空间资源利用率，极大的压缩了变频组件的体积，且保障射频与中频电路性能，具体的三维异构集成方式如图 3 所示。这种异构集成的方法集成度高、难度大、新颖性强，也给设计仿真、调试带来了巨大困难。

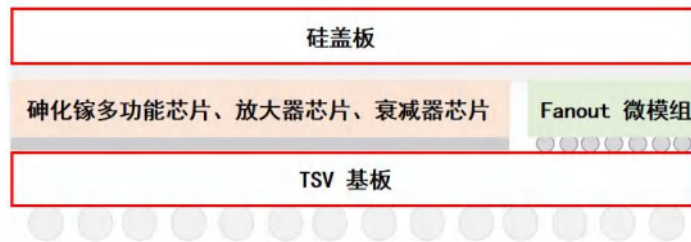


图 3 三维异构集成方式

4 多物理场协同仿真

4.1 三维电磁全波仿真

本文提出的三维电磁全波仿真模型主要分为以下 8 部分，接收机射频信号从母板通过微模组底部 BGA 球再经过 TSV 基板中的类同轴进入，接着到硅基板表面后，通过表层微带传输线金丝键合到 LNA 芯片，再到开关预选滤波器多功能

芯片，接着通过微带线与混频器芯片进行互联，本振信号的输入路径类似；然后在混频器的中频输出口，通过传输线与中频放大器 and 滤波器实现互联，最终通过 Fanout 微模组实现下变频处理后的中频信号输出，同样再经由类同轴和 BGA 球传输到母板中频输出口，整板三维电磁全波仿真模型与 TSV 基板实物对比如图 4 所示。

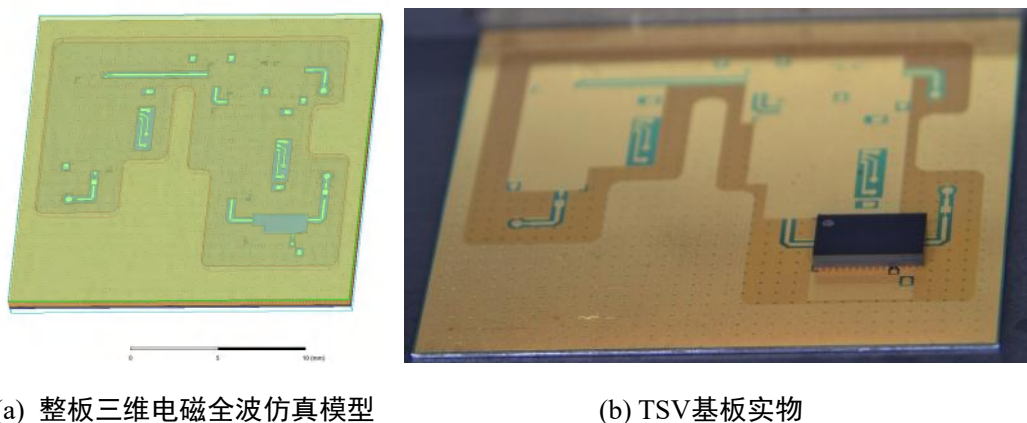


图 4 整板三维电磁全波仿真模型与TSV基板实物对比

针对一典型无源过渡结构进行局部详细电磁场仿真分析，接收机射频信号从测试母板的板内带状线进入，再通过 PCB 板层中的类同轴到复合基板表面，接着通过射频微模组 BGA 球经硅基板内类同轴到 TSV 表层微带线，完整过渡结构的三维电磁全波仿真模型如图 5 所示。

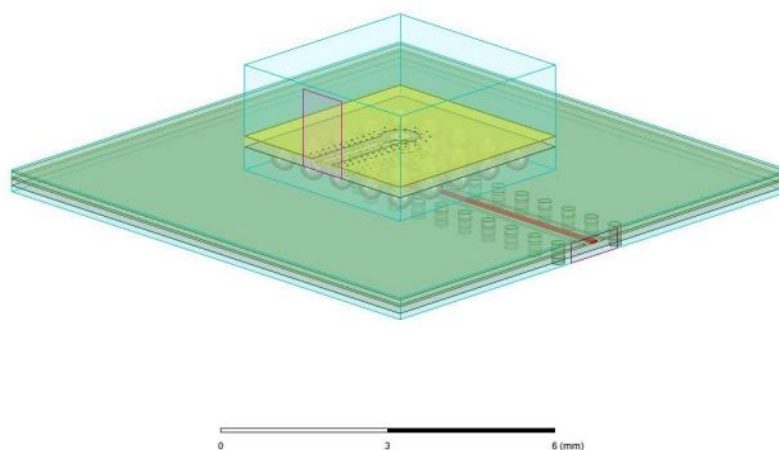


图 5 母板带线到TSV基板表层微带线的局部结构模型

本文提出的射频微模组测试母板内的带状线到 TSV 基板表层微带线的局部结构仿真结果如图 6 所示，其中端口 1 为 TSV 基板表层微带线的射频输出口，端口 2 为测试母板内带状线的射频输入口。最终该 2 端口过渡结构的反射系数 S_{11} 与 $S_{22} \leq -11.4 \text{ dB}@18\text{GHz}$ ，插入损耗 $S_{21} \leq -0.5 \text{ dB}@18\text{GHz}$ ，根据无源结构的互异性^[23]，该端口间隔离度 $S_{12} \leq -0.5 \text{ dB}@18\text{GHz}$ ，均满足设计指标要求。

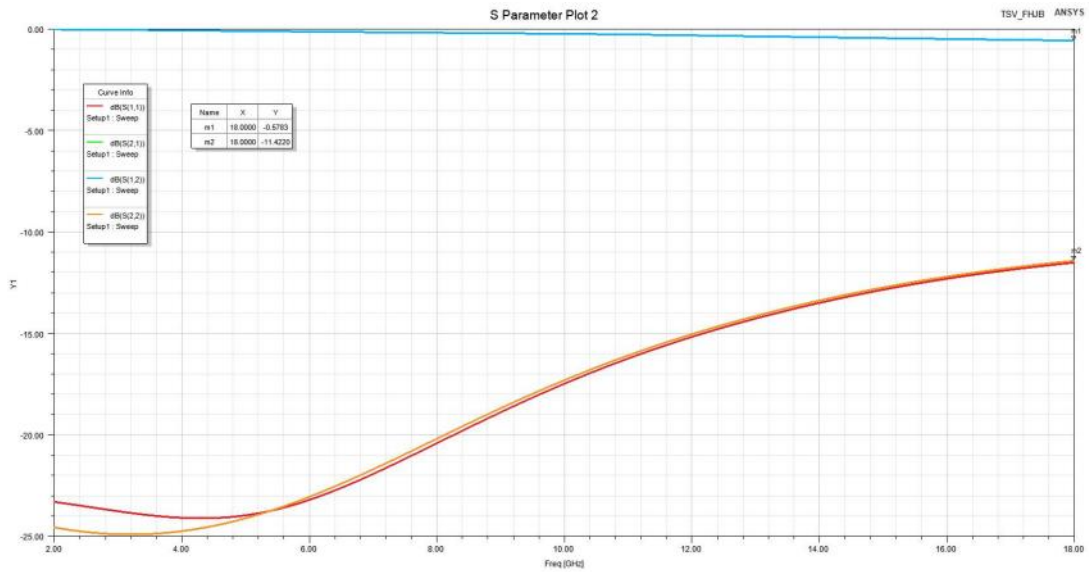


图 6 母线带线到TSV基板表层微带线的局部结构仿真结果

4.2 场路联合仿真

根据本文提出的射频微模组电路系统架构,首先对其中每款芯片建立对应的表格基行为模型^[24-25],然后将其无源输入输出过渡结构的仿真结果带入芯片实测数据链路中,实现如 7 图所示的场路联合仿真模型。

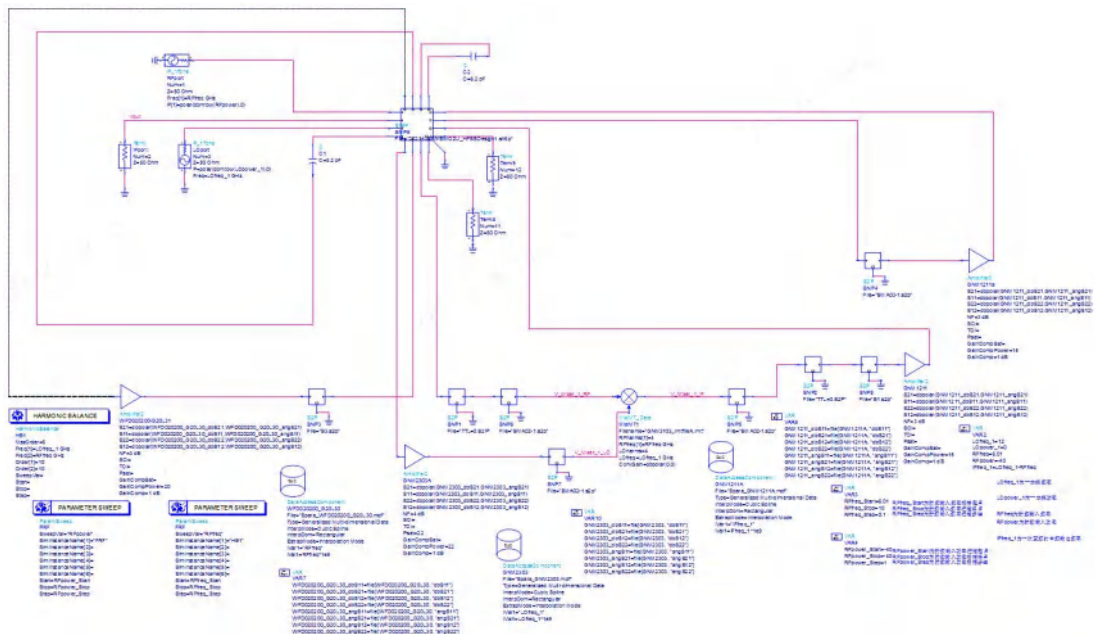
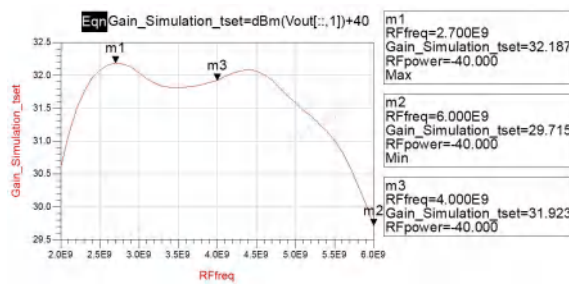
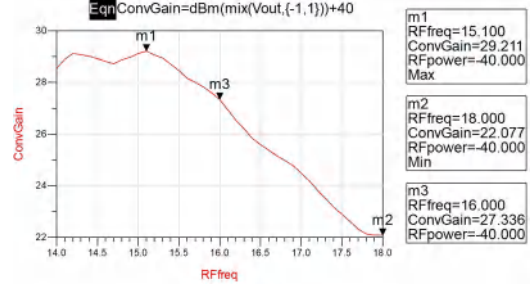


图 7 场路联合仿真模型

通过谐波平衡 (Harmonic Balance, HB) 仿真分析法,对整个射频微模组的链路进行研究。如图 8 所示, SC 波段的变频增益 $\geq 29.7 \text{ dB}@6\text{GHz}$, Ku 波段的变频增益 $\geq 22.1 \text{ dB}@18\text{GHz}$, 均满足设计指标要求。



(a) 2-6 GHz 工作频段



(b) 14-18 GHz 工作频段

图 8 场路联合仿真结果

4.3 热电耦合仿真

首先对 4 款核心发热芯片建立整版三维结构热模型，如图 9 所示。其中基板相关材料的仿真参数如表 1 所示，GaAs 芯片的热阻为 20 °C/W。然后根据单个芯片功耗在 BGA 球底部为 +70 °C 的情况，计算出芯片的结温如表 2 所示。最后根据热模型仿真结果，测试 4 款核心发热芯片在该温度下各自的数据，反馈到场路联合仿真模型，即可得到可调温度的热电耦合仿真模型，这在实际工程应用中具有可操作的实践意义。

表 1 基板相关材料仿真参数

名称	热导率
封装底部锡球	50 W/mK
芯片粘接材料	2W/Mk
围框、盖板、基板材料	150W/mK
Fanout 微模组 BGA 球	35W/mk
PI	0.15 W/mk
树脂	0.5 W/mk

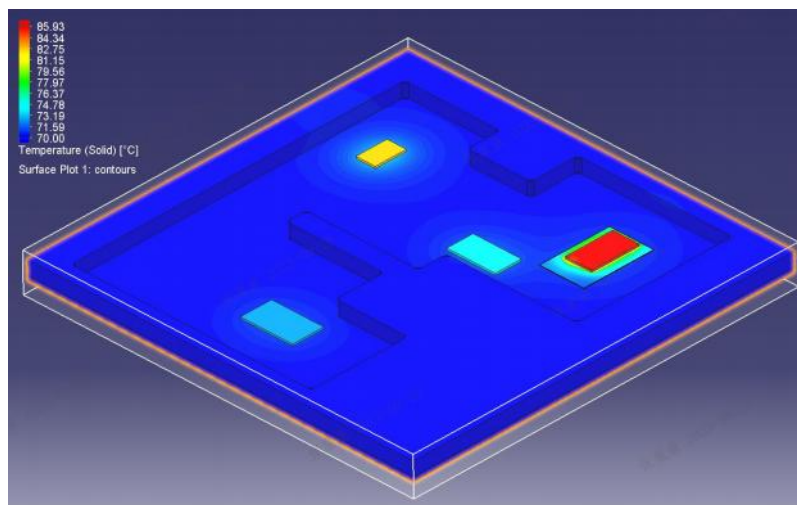


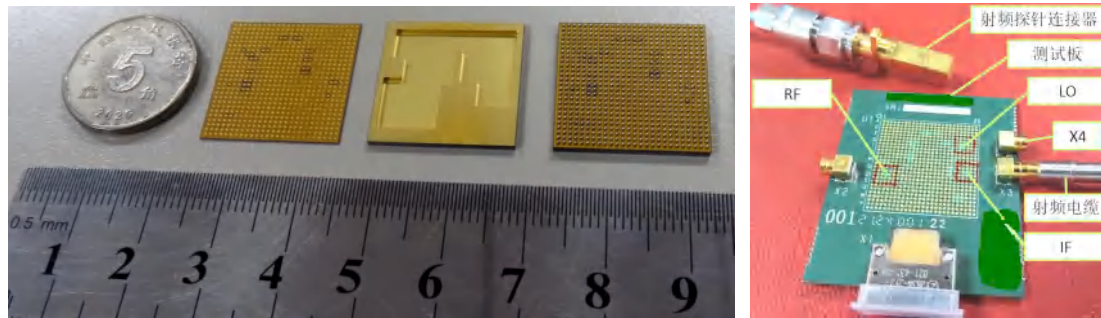
图 9 三维结构热模型

表 2 芯片的结温

芯片种类	芯片壳温 (°C)	芯片结温 (°C)
芯片 1	72.96	79.16
芯片 2	82.35	95.85
芯片 3	74.57	82.37
芯片 4	85.93	93.73

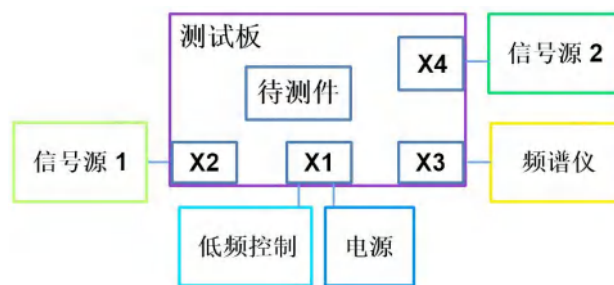
5 实测验证与仿真对比

基于电热耦合多物理场模型，实现了本文提出的射频微模组仿真设计，实物如图 10 所示，其主要由硅基板、盖板和其中裸片及 Fanout 中频衰减放大微模组构成。性能指标实现了 Ku 波段以内射频信号下变频到 C 波段以下中频信号的完整接收链路功能。实测结果表明，本文提出的射频微模组在全频段内变频增益达 22 dB，实物尺寸仅 20×20×1.68 mm³，且平面尺寸为目前 200 um 厚度硅基板加工的工艺极限状态。在性能指标相同的情况下，本文的研究较传统微组装方案体积尺寸大幅缩小，高度降低至 20%，面积缩小至 15%。



(a) 实物

(b) 测试夹具



(c) 测试框图

图 10 宽带变频微模组实物及测试系统

针对核心指标变频增益 (Gain)，将其实测 ($Gain_{test} = 28.5$) 与仿真 ($Gain_{simulation} = 29.8$) 数据进行对比分析，根据式 (1) 计算得到两者的误差 (Err) 为 4.8%，这个结论很大程度上验证了射频微模组的模型实际应用效果与可行性。

$$Err = | Gain_{simulation} - Gain_{test} | / Gain_{test} * 100 \% \quad (1)$$

6 结论

本文的射频微模组设计采用硅基三维堆叠工艺,很好的结合了 BGA、TSV、RDL 三种先进封装工艺,通过 BGA 将 Fanout 微模组垂直互联到 TSV 基板上,并将 GaAs 裸片粘接到硅基板表面的同一侧,芯片间及其与表层微带传输线间通过金丝键合实现互联,最后在硅基板底部植球,然后焊接到测试母板上实现完整的性能验证。在单模组内异构集成了 GaAs 工艺射频芯片及 Fanou 封装微模组,构建电热耦合多物理场模型,最终实现了 Ku 波段以内射频信号下变频到 C 波段以下中频信号的完整接收链路功能,为变频通道的系统级芯片集成提供了高可靠性解决方案。典型指标变频增益的实测与仿真数据对比误差精度较高,完善了国内射频微系统产品中芯片到微模组及组件级的高精度系统模型研究和实际应用探索链条。随着射频前端组件模块通道数的增加,信号总功率增加、信号的选择控制也会变得更加复杂,射频微模组及板级集成的电热耦合性能研究、微互联结构的长期可靠性分析也是未来的一个有效聚焦方向^[26-27]。

参考文献

- [1] Razavi. RF microelectronics[M]. New York: Pearson Education, Inc., 2012.
- [2] 夏晨辉, 王刚, 王波, 等. 用于 5G 通信的射频微系统与天线一体化三维扇出型集成封装[J]. 电子学报, 2023, 51(6): 1-6.
- [3] 徐锐敏, 樊元东, 薛良金, 等. 18-40 GHz 毫米波电子侦察接收机高频前端[C]. 1999 年全国微波毫米波会议论文集, 长沙, 1999: 436-438.
- [4] 王毅. 0.8~18 GHz 超宽带接收组件设计与实现[J]. 电子技术应用, 2022, 48(1): 146-151+156.
- [5] 李翀, 舒安刚, 胡俊丰, 等. 超宽带多通道接收前端设计与研制[J]. 压电与声光, 2024, 46(4): 458-462.
- [6] 谢小强, 徐跃杭, 夏雷. 微波集成电路[M]. 北京: 电子工业出版社, 2018.
- [7] 徐锐敏, 陈志凯, 赵伟. 微波集成电路的发展趋势[J]. 微波学报, 2013, 29(S1): 55-60.
- [8] 过方舟, 徐锐敏. 系统级封装关键技术研究进展[J]. 微波学报, 2014, 30(S1): 588-593.
- [9] 郝继山, 向伟玮. 微系统三维异质异构集成与应用[J]. 电子工艺技术, 2018, 38(6): 317-321.
- [10] 李扬. 基于 SiP 技术的微系统[M]. 北京: 电子工业出版社, 2021.
- [11] 段超, 周文胜, 骆明伟, 等. 基于复合基板的高集成超宽带变频组件设计[J/OL]. 微波学报, 2024: 1-8.
- [12] 麻泽龙, 吴景峰, 余小辉. 基于系统级封装技术的 X 频段变频模块研制[J]. 电子技术应用, 2024, 50(12): 105-111.

- [13] 毛军发. 毫米波电路系统异质集成技术[C]. 2018 年全国微波毫米波会议论文集, 成都, 2018: 10-10.
- [14] 汪志强, 杨凝, 戴扬, 等. 异构集成路线图对我国微系统发展的启示[J]. 导航与控制, 2022, 21(3/4): 40-45.
- [15] 汪志强, 杨凝, 张劲春, 等. 异质异构集成微系统可靠性技术发展的挑战和机遇[J]. 微电子学与计算机, 2023, 40(11): 61-71.
- [16] 徐锐敏, 王欢鹏, 徐跃杭. 射频微系统关键技术进展及展望[J]. 微波学报, 2023, 39(5): 70-78.
- [17] 段宝岩. 电子装备机电耦合研究的现状与发展[J]. 中国科学:信息科学, 2015, 45(3): 299-312.
- [18] 余怀强, 唐光庆, 桂进乐, 等. 微系统热管理技术的新发展[J]. 压电与声光, 2018, 40(6): 931-935.
- [19] Ming Luo, Qing Xiao, Qian Lu, and et al. A compact RF front-end SiP using silicon-based integrated passive devices[C]. 2018 International Conference on Microwave and Millimeter Wave Technology (ICMMT), 2018: 1-3.
- [20] 朱臣伟, 刘娟, 唐昊, 等. 基于协同仿真技术的超宽带射频微系统热电设计[J]. 固体电子学研究进展, 2022, 42(4): 269-274+286.
- [21] 吴小玲, 阮文州, 刘德志. X 波段高集成硅基射频微系统[J]. 微纳电子技术, 2024, 61(7): 111-115.
- [22] 张睿, 朱旻琦, 杨兵, 等. 硅基三维异构集成射频微系统的多物理场耦合仿真与设计[J]. 电子技术应用, 2024, 50(5): 1-6.
- [23] D. M. Pozar. Microwave engineering, third edition[M]. New York: John Wiley & Sons, Inc., 2005.
- [24] 吴洪学, 高学邦. 雷达收发组件芯片技术[M]. 北京: 国防工业出版社, 2017.
- [25] Zhiqiang Sun, Huanpeng Wang, Jiahui Miao, and et al. Design of a X-Band 2×2 Phased-Array Transceiver Module Using Chiplet Method[C]. IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), 2021: 24-26.
- [26] Huanpeng Wang, Jingrou Wang, Yuehang Xu. Reliability Design of a X-Band 2×2 Tile-Type T/R Module under Thermal Shock and Random Vibration[C]. IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), 2020: 100-101.
- [27] 王从思, 王璐, 王志海. 微波组件机电热耦合建模与影响机理分析[M]. 北京: 科学出版社, 2018.

Research Article

Total Ionizing Dose Effect Modeling of RRAM Devices: Mechanisms and Compact Model

GUO Huimeng^{1,2}, LI Yujia², LI Yiqing^{1,2}, REN Tingrui^{2,3}, WANG Liang², ZHAO Yuanfu²

¹ School of Electronic and Information, Hangzhou Dianzi University, Hangzhou 310018, China

² Beijing Microelectronics Technology Institute, Beijing 100076, China

³ School of Integrated Circuit Science and Engineering, Beihang University, Beijing 100191, China

Corresponding author: Guo Huimeng; Email: guohm163@163.com.

Manuscript Received March 22, 2022; Accepted March 22, 2022; Published Online March 22, 2022.

Abstract—This study investigates the total ionizing dose (TID) effects on RRAM devices by establishing a complete test system combining pulse operation and in-situ radiation characterization, where an optimized pulse programming scheme with adaptive voltage and timing adjustment was developed to address radiation-induced degradation. The results reveal that radiation exposure leads to significant modifications in device characteristics, including converged resistance state distributions between HRS and LRS, distorted I-V curves with altered switching thresholds, and increased operation variability, while the developed radiation-aware compact model successfully captures these effects by incorporating trap-modulated conduction, dose-dependent oxygen vacancy mobility, and modified filament dynamics, providing critical insights for radiation-hardened RRAM design.

Keywords—RRAM, Total ionizing dose, Reliability.

I. Introduction

Resistive random-access memory (RRAM) has emerged as a promising candidate for next-generation non-volatile memory owing to its advantages of simple structure, high scalability, fast switching speed, and low power consumption [1], [2]. With the increasing deployment of memory systems in aerospace, nuclear, and other radiation-prone environments, understanding and mitigating radiation-induced degradation in RRAM devices have become critical for ensuring system reliability [3], [4]. Among various radiation effects, the total ionizing dose (TID) effect, caused by the accumulation of radiation-generated charges in dielectrics and interfaces, can lead to significant parameter shifts, resistance state instability, and even functional failure in RRAM devices [5][6][7].

Radiation-induced effects pose significant reliability challenges for memory devices operating in aerospace, nuclear, and other harsh environments. As an emerging non-volatile memory, RRAM has attracted considerable attention due to its simple metal-insulator-metal structure, low power consumption, and scalability. In recent years, the radiation resistance of RRAM devices has become a focal point of research, particularly concerning their response to TID, single-event effects (SEE), and displacement damage [8]. Studies reveal that RRAM's switching mechanism, driven by conductive filament formation and rupture, exhibits inherent tolerance to certain radiation-induced charge trapping effects [9]. However, performance degradation, such as resistive state instability and increased variability under high-dose exposure, has also been observed. Researchers are actively exploring material optimization, structural design

improvements, and radiation-hardening techniques to enhance RRAM's operational reliability in radiation environments [10]. Overall, RRAM demonstrates promising potential for radiation-resilient memory applications, yet further systematic investigation is necessary to fully understand and mitigate its failure mechanisms under complex radiation conditions. While extensive research has been conducted on TID effects in conventional CMOS technologies, experimental studies and physical modeling of TID-induced degradation in RRAM devices remain insufficient. The lack of reliable, experimentally validated models severely limits circuit-level radiation-hardened design and system-level reliability prediction for RRAM-based memory.

In this work, we present a comprehensive experimental investigation and compact modeling of the TID effects in RRAM devices. A series of controlled radiation experiments were performed to characterize the electrical parameter shifts and resistance state evolution under different total dose levels. Based on the experimental observations and physical mechanisms analysis, a physics-based compact model is proposed to quantitatively describe the TID-induced degradation behaviors of RRAM. The developed model is implemented in a SPICE-compatible framework, enabling accurate circuit simulation and providing essential design guidance for RRAM-based memory systems operating in radiation-harsh environments.

II. Characteristics of RRAM

1. Physical Model of RRAM

RRAM devices operate based on resistive switching, where the resistance state is modulated by controlling the

formation and rupture of conductive filaments (CFs) within the oxide layer. The device consists of a trilayer structure: a top electrode (TE), a transition metal oxide (TMO) switching layer, and a bottom electrode (BE). By applying voltage biases of different polarities across the electrodes, the electric field drives the redistribution of oxygen vacancies (Vo) within the TMO layer, thereby regulating the formation and dissolution of CFs. This process dynamically alters the resistance of device between high-resistance (HRS) and low-resistance (LRS) states.

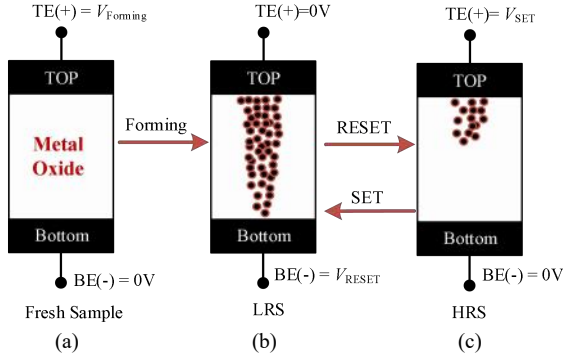


Figure 1 Oxygen vacancy migration process in RRAM. (a) Fresh sample: HRS before forming. (b) After forming: CF formation, switching to LRS. (c) After reset: CF rupture, switching back to HRS.

In the SET process, a positive voltage V_{SET} applied to the TE induces oxygen ion migration, leading to the aggregation of oxygen vacancies and the formation of a conductive filament, switching the device to LRS. Conversely, in the RESET process, a negative voltage V_{RESET} at the BE causes filament rupture due to oxygen vacancy recombination, reverting the device to HRS.

For modeling simplicity, the complex electrochemical processes—such as oxygen ion migration and vacancy diffusion—are often abstracted into a conductive filament growth/rupture mechanism. This simplified approach effectively captures the resistive switching behavior while enabling compact modeling for circuit simulations [11].

$$\begin{cases} I_{hop} = I_0 \left(\frac{\pi w^2}{4} \right) \times \exp\left(-\frac{x}{x_T}\right) \times \sinh\left(\frac{V_{gap}}{V_T}\right) \\ I_{CF} = \pi w^2 V_{CF} / 4 \rho (x_0 - x) \end{cases} \quad (1)$$

The SET transition involves the field-driven migration of oxygen vacancies and subsequent CF growth, a dynamic process described by:

$$\begin{cases} \frac{dx}{dt} = af \exp\left(-\frac{E_a - \alpha_a ZeE}{k_B T}\right) \\ \frac{dw}{dt} = (\Delta w + \Delta w^2 / 2w) f \exp\left(-\frac{E_a - \alpha_a ZeE}{k_B T}\right) \end{cases} \quad (2)$$

The RESET transition is triggered by Joule heating and oxygen ion back-diffusion, leading to CF rupture. The kinetics are modeled as:

$$\begin{cases} \frac{dx}{dt} = af \exp\left(-\frac{E_i - \gamma ZeV}{k_B T}\right) \\ \frac{dx}{dt} = af \exp\left(-E_h / k_B T\right) \sinh\left(\frac{\alpha_h ZeE}{k_B T}\right) \\ \frac{dx}{dt} = af \exp\left(-\frac{\Delta E_r}{k_B T}\right) \end{cases} \quad (3)$$

In summary, the resistive switching behavior of RRAM devices is fundamentally governed by the migration of oxygen vacancies and the corresponding formation and rupture of conductive filaments within the oxide layer. This filamentary switching mechanism not only enables scalable and energy-efficient memory operation but also lends itself to simplified physical models that facilitate device characterization and circuit-level simulations.

2. Electrical Characterization

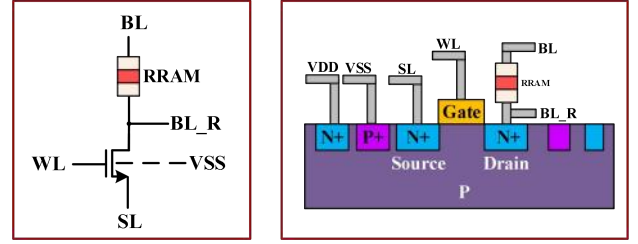


Figure 2 (a) Circuit schematic of the 1T1R test structure, showing the series connection between the selector transistor and RRAM device with critical measurement nodes. (b) Corresponding layout diagram illustrating the signal distribution and interconnect architecture.

The 1T1R test structure employs a series-connected NMOS transistor to provide intrinsic current compliance during RRAM operation, where the gate voltage (V_{WL}) sets the saturation current to prevent dielectric breakdown. For electrical characterization, SET operations are performed by applying 3V/500ns pulses to the bit line (BL) with $V_{WL}=2.5V$ (limiting current to $\sim 100\mu A$), while RESET uses $-2V/1\mu s$ pulses with $V_{WL}=3.3V$. A dedicated read circuit (BL_R path) incorporating a sense amplifier and $10k\Omega$ reference resistor enables accurate resistance measurement by applying 0.2V read voltage and detecting the current through the RRAM cell, allowing for LRS/HRS distinction without disturbing the memory state.

The test system, as illustrated in Fig. 3, integrates three key components: a host PC, test board, and DUT socket. The host PC transmits control commands via UART to the test board, which incorporates an FPGA and precision analog circuits to generate programmable voltage/pulse signals for RRAM stimulation through probe station connections (BL/WL/SL pins). The system executes four operational modes: READ, SET, RESET, and DC-IV sweep. Device responses are captured by a 16-bit ADC (1MS/s sampling rate) on the test board and transmitted via isolated UART to the host PC, where custom Python software performs real-time data processing.

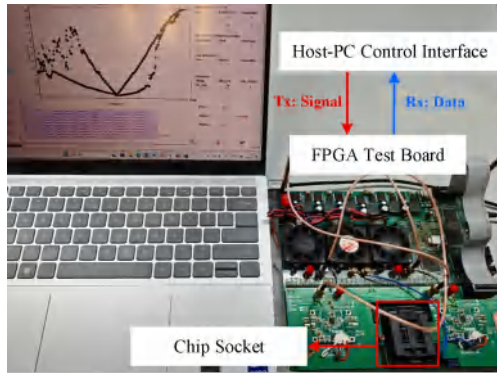


Figure 3 RRAM test system architecture.

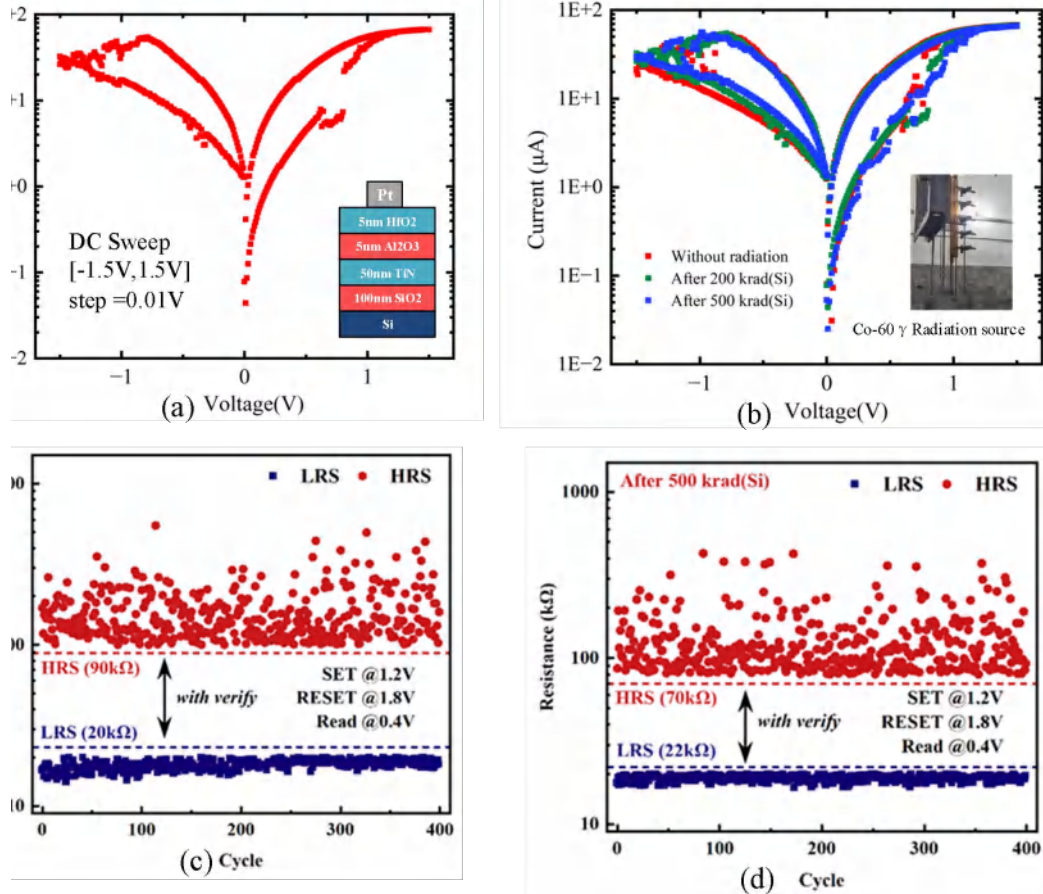


Figure 4 (a) Voltage sweep ($0V \rightarrow 1.5V \rightarrow -1.5V \rightarrow 0V$, $step=0.01V$) demonstrating SET (LRS) and RESET (HRS) transitions. (b) voltage sweep characteristics before and after radiation (c) Resistance state distributions in RRAM cycling test before radiation. (d) statistical distribution of resistance states across multiple cycles after 500 krad (Si) radiation

Figure 4 illustrate the impact of radiation exposure on RRAM switching behavior and resistance stability. The DC I-V characteristics were measured with voltage sweep from $-1.5V$ to $1.5V$ ($step=0.01V$). During the forward sweep ($0 \rightarrow 1.5V$), the device exhibits abrupt SET transition at $\sim 1.2V$, switching to LRS. In the reverse sweep ($1.5V \rightarrow -1.5V$), RESET occurs at $\sim -1.0V$, reverting to HRS. Figure 4(cs) presents the resistance state distributions of the RRAM device over 400 switching cycles. The high-resistance state (HRS) shows a broad distribution ($>90k\Omega$), while the low-resistance state (LRS) exhibits a tightly clustered distribution ($<20k\Omega$). The cycling tests were performed under SET ($1.2V/500ns$), RESET ($-1.8V/5\mu s$), and READ ($0.4V$) conditions. The results demonstrate stable switching behavior with a consistent HRS/LRS resistance ratio exceeding $5\times$, confirming excellent endurance characteristics, the

radiation test results will be discussed in detail in the following section.

III. Radiation Experiments and Analysis

1. Experimental results

Figure 4(b) shows the I-V curves of the resistive unit of the Pt/HfO₂/Al₂O₃/TiN structure device before and after irradiation. The device still has bipolar resistive characteristics after doses of 200 krad (Si), 500 krad (Si). Most devices can still perform write-erase operations at the set voltage, and overall exhibit good radiation resistance, indicating that irradiation has little impact on RRAM. Unrestricted current. With increasing radiation dose, the resistance window of the RRAM devices gradually narrows, primarily due to a significant reduction in the high-resistance state (HRS) and a slight

increase in the low-resistance state (LRS). To address this degradation, an adaptive pulse control strategy will be proposed in the subsequent sections to enhance the device endurance under radiation exposure.

2. RRAM Radiation Effects Modeling

To accurately capture the degradation mechanisms of RRAM devices under radiation environments, a comprehensive modeling framework is established by incorporating ionizing damage, displacement damage, and radiation-enhanced thermal effects. Ionizing radiation leads to increased oxygen vacancy concentration and reduced spacing between adjacent vacancies, resulting in the broadening of the initial vacancy distribution and weakening the stability of the conductive filaments. Simultaneously, displacement damage introduces lattice defects, which cause potential barrier fluctuations and local electric field distortions, further exacerbating filament instability and variability in the switching process. In addition, radiation-induced heating effects raise the local temperature of the device, accelerating defect generation and vacancy migration. These combined effects significantly alter the conductive filament evolution dynamics, leading to reduced resistance window, threshold voltage shifts, and increased resistance state variability. By integrating these physical mechanisms

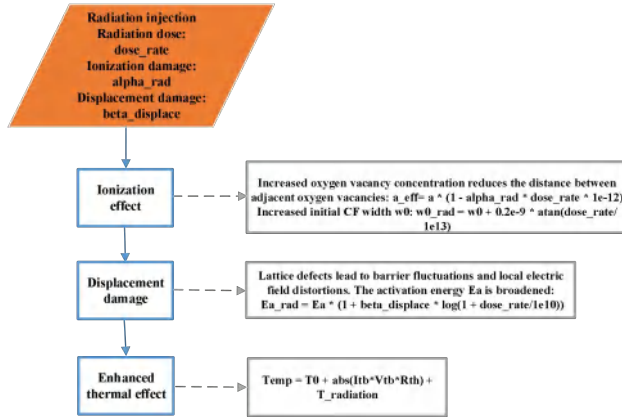


Figure 5 RRAM radiation effects modeling process.

into a compact model, the radiation-induced degradation of RRAM devices can be effectively predicted, providing theoretical guidance for device hardening and control strategy optimization in radiation-intensive applications.

IV. Conclusions

This study systematically investigated the total ionizing dose (TID) effects on RRAM devices by developing a dedicated test system and establishing a comprehensive radiation effects model. The experimental setup enabled precise characterization of TID-induced degradation in switching parameters, including SET/RESET voltages and resistance distributions. By correlating electrical measurements with physical analysis, we identified three key radiation damage mechanisms: trap-assisted conduction, oxygen vacancy migration enhancement, and interfacial layer degradation. The proposed compact model successfully predicts

post-irradiation device behavior by incorporating dose-dependent modifications to filament dynamics and charge transport, demonstrating good agreement with experimental data. These findings provide critical insights for designing radiation-hardened RRAM technology suitable for space and nuclear applications. The test methodology and modeling framework developed in this work offer valuable tools for evaluating and optimizing the radiation tolerance of emerging nonvolatile memory devices.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB4400401.

References

- [1] S. Sing *et al.*, "A New High Density 3D Stackable Via RRAM for Computing-in-Memory SOC Applications," *IEEE Trans. Electron Devices*, vol. 71, pp. 2399-2403, April 2024.
- [2] J. Chen, H. Wu, B. Gao, J. Tang, X. S. Hu and H. Qian, "A Parallel Multibit Programming Scheme With High Precision for RRAM-Based Neuromorphic Systems," in *IEEE Transactions on Electron Devices*, vol. 67, no. 5, pp. 2213-2217, May 2020.
- [3] H. Liu *et al.*, "Uniformity Improvement in 1T1R RRAM With Gate Voltage Ramp Programming," in *IEEE Electron Device Letters*, vol. 35, no. 12, pp. 1224-1226, Dec. 2014.
- [4] A. M. S. Tosson, S. Yu, M. H. Anis and L. Wei, "A Study of the Effect of RRAM Reliability Soft Errors on the Performance of RRAM-Based Neuromorphic Systems," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 11, pp. 3125-3137, Nov. 2017.
- [5] L. Zhang *et al.*, "Total Ionizing Dose (TID) Effects on TaOx -Based Resistance Change Memory," in *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2800-2804, Aug. 2011.
- [6] P. Apsangi, N. Chamele, H. J. Barnaby, Y. Gonzalez-Velo, K. E. Holbert and M. N. Kozicki, "Impact of TID on the Analog Conductance and Training Accuracy of CBRAM-Based Neural Accelerator," in *IEEE Transactions on Nuclear Science*, vol. 70, no. 12, pp. 2572-2577, Dec. 2023.
- [7] X. Han, A. Privat, K. E. Holbert, J. Seo, S. Yu and H. J. Barnaby, "Total Ionizing Dose Effects on Multistate HfO_x-Based RRAM Synaptic Array," in *IEEE Transactions on Nuclear Science*, vol. 68, no. 5, pp. 756-761, May 2021.
- [8] K. Takeuchi *et al.*, "Single-Event Effects Induced on Atom Switch-based Field-Programmable Gate Array," in *IEEE Transactions on Nuclear Science*, vol. 66, no. 7, pp. 1355-1360, July 2019.
- [9] B. Traoré *et al.*, "On the Origin of Low-Resistance State Retention Failure in HfO₂-Based RRAM and Impact of Doping/Alloying," in *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4029-4036, Dec. 2015.
- [10] A. Lodhi, A. Dwivedi, S. Saini, N. Jingar, A. Khandelwal and S. P. Tiwari, "Radiation-Hardened ReRAM Devices for Flexible Electronics," in *IEEE Journal on Flexible Electronics*, vol. 2, no. 5, pp. 402-407, Sept. 2023.
- [11] H. Li, P. Huang, B. Gao, B. Chen, X. Liu and J. Kang, "A SPICE Model of Resistive Random Access Memory for Large-Scale Memory Array Simulation," in *IEEE Electron Device Letters*, vol. 35, no. 2, pp. 211-213, Feb. 2014.



Huimeng Guo received the B.Eng. degree in electronic science and technology from the North China University of Water Resources and Electric Power, Henan, China, in 2021. She is currently pursuing the Ph.D. degree in electronic science and technology with Hangzhou Dianzi University, Hangzhou, Zhejiang, China. Additionally, she is also studying at the

Beijing Institute of Microelectronics Technology. Her current research interests include RRAM, memristive systems, and non-volatile memory.



Yujia Li received the B.E. degree from the North University of China, Taiyuan, Shanxi, China, in 2013. She then pursued the M.E. degree at the China Academy of Electronics and Information Technology (CAEIT), Beijing, China, graduating in 2016. In 2021, she earned her Ph.D. degree from the Beijing University of Technology, Beijing, China. After completing her doctoral

studies, she joined the postdoctoral research center at the Beijing Institute of Microelectronics Technology. Her current research interests include RRAM and device reliability.

and hardening of devices and integrated circuits.



Yiqing Li received the B.Eng. degree in electronic and information engineering from the North China Institute of Science and Technology, Hebei, China, in 2022. She is currently pursuing a Ph.D. degree in electronic science and technology at Hangzhou Dianzi University in Zhejiang, China. Additionally, she is also studying at the Beijing Institute of Microelectronics

Technology. Her current research interests include locally active memristors, nonlinear circuits, neural networks and radiation-hardened integrated circuit design.



Tingrui Ren received the B.E. degree from Changchun University of Science and Technology in China. He is currently pursuing the Ph.D. degree in Integrated Circuit Science and Engineering in Beihang University, Beijing, China. His research interests include neural architecture search for computing-in-memory (CIM) systems and hardware acceleration of deep learning.



Liang Wang received the B.S. degree in Science and Technology of Electronics from the Harbin Institute of Technology, Harbin, China, in 2004, the M.S. degree in Microelectronics from the Beijing Institute of Microelectronics Technology (BMTI), Beijing, China, in 2007, and the Ph.D. degree in Microelectronics and Solid-State Electronics from the BMTI, Beijing, China,

in 2022. He is currently a professor in BMTI. His research interests include reliability mechanisms, simulation and design optimization of CMOS devices and integrated circuits. He has authored more than 70 papers on academic journals and conferences.



Yuanfu Zhao (SM'07) received the B.S. degree in semiconductor physics and devices from Zhejiang University, Hangzhou, China, in 1983, the M.S. degree from Xidian University, Xi'an, China, in 1986, and the Ph.D. degree from Shaanxi Microelectronics Technology Institute, Xi'an, in 1989. He is currently a Ph.D.

Advisor with the Beijing University of Technology, Beijing, China, and he is also the Ph.D. Advisor with the Harbin Institute of Technology, Harbin, China, and the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include radiation effects

先进封装电磁兼容智能化设计研究

李达

浙江大学

摘要：针对先进封装高密度互连中日益严峻的信号/电源完整性与电磁干扰挑战，本报告将汇报课题组构建的面向封装引脚分配优化的智能化设计框架。开发了虚边界广义 T 矩阵与混合模态级联等快速电磁分析技术，实现了近远场辐射及封装互连 S 参数的高速、精确评估。此外，通过开发深度强化学习模型和物理辅助增强多目标遗传算法，实现了焊球引脚阵列电磁自屏蔽优化，以及多层过孔引脚阵列布局的多目标、多约束 SI/PI 协同优化。

射频电路和集成系统的多物理场分析

唐旻

上海交通大学

摘要：随着集成电路不断向高性能/多功能、高集成度快速发展，电路结构和工艺复杂度明显增加，电磁、温度、应力等多物理耦合效应愈加显著，对其进行准确高效的仿真分析与设计面临极大挑战，存在诸多科学技术问题亟需解决。本报告主要针对射频电路与集成系统的多物理场建模方法、高效仿真技术、以及多物理场 EDA 软件开发等方面进行阐述，具体内容包包括射频无源器件的多物理协同分析、高功率半导体器件的多物理效应研究、复杂封装系统的多物理场仿真技术等，最后简要介绍基于多物理场的射频集成系统多功能协同设计案例。

集成微系统多物理耦合精准建模与快速仿真

王大伟

杭州电子科技大学

摘要：针对集成微系统中多物理耦合行为精准建模与快速仿真需求，开展了器件-封装-系统多个层级建模仿真方法研究。首先，充分考虑各层级数理模型数值特征，制定了谱元、有限元、有限体积等算法为基础的多算法协同方案；同步，引入了区域分解、多线程并行计算实现了多物理耦合问题高效求解。随后，以数值计算方案为基础，开发了多款自研多场仿真器，可支持半导体器件、微波电路、集成封装系统等多类型器件/功能单元多物理特性仿真分析。最后，为实现微系统多物理仿真分析，探索了等效原理、模态分解、正交分解等模型降阶技术在射频/微波/毫米波电路协同仿真中的应用，形成了多种机制模型协同建模与快速仿真方法。

微波异构集成芯片系统仿真及三维器件模型提取技术研究

王维波

中国电子科技集团公司第五十五研究所

摘要：三维射频异构系统集成技术是当前电子信息技术领域的前沿研究方向，它通过将不同材料、工艺和功能的芯片与元器件在三维空间内高密度集成，实现了微波系统性能的飞跃提升。随着军事、通信、雷达等领域对微波组件性能要求的不断提高，传统的二维集成技术已难以满足多功能、高性能、小型化、低功耗等迫切需求，三维射频异构系统集成技术已经成为研究热点。

对于三维射频异构集成系统而言，其采用的 MMIC 芯片产生的空间辐射电磁场将会与三维结构产生相互作用，从而导致在近似二维平面尺度上设计的 MMIC 性能发生较大变化，这种变化随着工作频率的升高而愈加明显，往往会导致三维射频异构集成系统研制异常困难，本文将针对这个典型问题，面向三维射频异构集成系统研究无源、有源器件产生的这种变化，以及其对电路性能产生的影响，通过与传统二维平面器件的模型参数对比，总结其模型变化的一般性规律，并结合典型的三维射频异构集成芯片设计验证了一种器件模型的提取技术，从而为三维 MMIC 正向设计提供了一种方法和思路，最终对典型 3D 微模组设计中存在的问题和方法进行了总结展望，期望能为相关领域的科研同行提供一些启发和参考。

高速互连中的光学设计与光电链路仿真

虞绍良

之江实验室

摘要：人工智能技术对算力提出了前所未有的需求，基于光互连技术构建超大规模集群已经成为了发展 AI 算力的重要趋势。本报告将介绍团队围绕高速光互连开展的相关光电建模仿真工作。一方面，对于微尺度的光学结构，如何高效设计多自由度光学自由曲面，实现低损耗光耦合。另一方面，如何对信号的发射和接收过程进行建模，实现电-光-电的全链路仿真。

InP高电子迁移率晶体管亚太赫兹建模与表征

张傲

南通大学

摘要：本报告讨论了对太赫兹频段 InP 基高电子迁移率晶体管 (HEMT) 建模和表征技术。提出了用于太赫兹频段的 InP HEMT 小信号和噪声可缩放模型，推导了噪声参数的表达式，构建了基于器件栅宽的噪声和信号模型参数的缩放规则。在 1-325GHz 的频率范围内，噪声参数 40GHz 频率范围内，栅宽为 $2 \times 15 \mu\text{m}$ 、 $2 \times 20 \mu\text{m}$ 和 $2 \times 25 \mu\text{m}$ (栅指数 \times 栅宽) HEMT 器件，模拟和测试结果吻合很好，模型可以精准预测相同栅长不同栅宽 InP HEMT 器件的交流特性和噪声特性。

基于所提出的太赫兹频段 InP HEMT 器件噪声模型，设计了一款基于 InP HEMT 工艺的 230-250 GHz 低噪声太赫兹单片集成电路。该放大器采用五级共源极放大结构；基于噪声匹配技术设计了放大器的第一级和第二级，基于功率匹配技术设计了中间两级，最后一级重点完成输出匹配。实测结果表明，230-250 GHz 频率范围内，低噪声放大器的小信号增益大于 20 dB，采用 Y 因子法对封装后的低噪声放大器模块完成了噪声测试，频率为 243-248 GHz 时该 MMIC 放大器噪声系数为优于 7.5dB，与 HBT 和 CMOS 工艺相比，基于 HEMT 工艺的低噪声放大器具有 3 dB 以上的噪声系数优势。

射频微系统技术与先进相控阵天线

周浩

雷达探测感知全国重点实验室

摘要：微系统技术来源于在微观尺度实现高密度集成的概念，集合了先进器件、尖端工艺、智能算法和开放架构的优势，让系统能够以更小的体积、更低的成本、更低的能耗去承载更庞大、更复杂的功能。相控阵天线技术经过了从无源到有源、从模拟到数字、从自适应到智能化的发展历程，正与微系统等新兴技术互相融合，在系统引领下不断完成从芯片到复杂系统、从硬件到软件的完整集成，实现系统从单一功能向多功能的跨越。报告在分析微系统技术发展及应用的基础上，重点介绍了射频微系统技术应用于相控阵天线的典型案例——晶圆集成相控阵微系统的概念内涵、技术布局以及研究进展，以此说明射频微系统技术对于先进相控阵天线系统的巨大助益。报告还将从系统示范、开放硬件、高端芯片、集成工艺、仿真测试等方面分享对于微系统相控阵天线的发展思考，希望对该方向的探索研究和工程应用起到借鉴和推动作用。

先进三维集成相控阵前端技术与探讨

施永荣

南京航空航天大学

摘要：报告围绕雷达探测与电磁频谱感知相控阵前端系统三维集成存在的问题与挑战，探讨相控阵滤波天线阵列、阵面孔径信号互连、有源通道异质异构集成架构和新型功分合成网络拓扑方面的工作，并对未来相控阵前端系统面临的科学问题进行了展望。

射频前端模组设计与仿真关键技术

赵佳劼

上海华大九天科技股份有限公司

摘要：射频前端模组设计是化合物工艺半导体的重要应用方向。本报告将首先回顾无线通信中射频前端模组设计特点，对 EDA 软件的要求，然后介绍全流程中涉及的关键技术，包括前端和后端设计、电路仿真、多物理场仿真及物理验证等，最后介绍华大九天在该领域已开展的工作和取得的主要进展。

基于板级EDA的精细化建模解决方案

尹朝卿

三微电子科技（苏州）有限公司

摘要：摘要：针对板级电子系统（如 PCB、封装基板）三维建模中精度不足、参数化灵活性差及多源数据集成效率低的问题，本文提出精细化建模解决方案，聚焦“参数化建模-多源数据融合-自动化转换”技术路径。

参数化建模引擎开发：通过改进映射表结构（增加操作 ID 与几何属性关联），支持 TSV、微凸块等复杂互连结构的灵活参数化定义，解决传统固定映射关系无法表达多层级操作逻辑的缺陷。

混合网格划分技术：采用自适应网格密度算法，在关键区域（如焊盘、过孔）提升网格精度至微米级，非关键区域优化密度以降低数据量，实现几何与渲染属性的高效协同。

多源数据集成机制：支持 CAD 几何模型（STEP/STL）与 EDA 规则库的自动转换，通过特征识别技术提取版图中的封装尺寸、布线规则等参数，直接生成带物理属性的三维模型，减少人工建模工作量。实验表明，该方案在 16 层 PCB 设计中实现三维模型文件体积压缩 40%，关键结构细节精度提升 50%，显著提升数据可复用性与制造兼容性。

宽带射频微系统器件 S 参数模型时域建模方法研究

陈智源, 葛菊祥, 郝继山, 张文锋, 黄俊杰

中国电子科技集团公司第二十九研究所, 成都 610036

摘要: 宽带射频微系统链路仿真依赖时域信号流驱动, 传统方法构建的器件模型往往难以直接调用, 而基于实测 S 参数模型的精度较高, 但在时域仿真中调用需频繁进行时频变换, 导致效率低下。本文提出一种基于 S 参数的时域表征方法, 通过频率抽样法构建自适应阶数 FIR 数字滤波器, 将频域 S 参数模型等效为时域 FIR 滤波器, 实现输入信号的时域卷积和延迟校正。仿真与实测验证表明, 该方法有效提升仿真效率, 避免信号混叠, 具有工程应用价值。

关键词: 微系统链路仿真; S 参数模型; 自适应阶数 FIR 数字滤波器; 时域卷积; 延迟校正

1 引言

宽带射频微系统链路仿真分析涉及天线、微波、数字等多个部分的全线贯通, 通过复杂组合和参数控制各类器件模型, 实现对系统性能的评估。其中, 器件模型的构建是确保射频微系统分析高效、精准的关键^[1-2]。然而, 传统的单学科建模仿真方法难以满足宽带射频微系统全链路分析的需求, 因为该系统需要基于信号流驱动的时域仿真^[3], 而天线、微波等学科采用的三维电磁场有限元分析或频域电路分析方法所构建的器件模型无法直接调用。

在众多器件建模方法中, 多端口网络 S 参数模型因其参数易获取、计算效率高等优势, 被广泛应用于无源器件或小信号有源器件的建模^[4-5]。与基于公式的理想器件模型相比, 基于实测 S 参数的模型具有更高的仿真精度。S 参数表征了端口入射电压波和反射电压波的映射关系, 从信号流角度看, 它反映了器件对输入信号的频域响应, 可视为器件的频域电路模型。

在宽带射频微系统链路的时域信号流仿真中, 若直接调用 S 参数模型, 则需在每个时间步对输入信号进行时频变换^[6]: 首先将输入时域信号转换为频域信号, 再将经过器件模型计算后的频域信号转换为时域信号输出。然而, 宽带射频微系统通常包含大量器件, 频繁的时频变换不仅严重影响仿真效率, 还可能导致时域信号混叠问题。

针对这一难题, 本文提出了一种宽带射频微系统器件 S 参数的时域表征和建模方法。该方法基于 S 参数的频率响应, 采用频率抽样法建立自适应阶数的 FIR 数字滤波器, 将频域 S 参数模型等效为时域 FIR 数字滤波器模型, 从而实现输入信号的时域卷积和延迟量校正。通过与实测数据对比, 验证了该方法在模型精度和仿真效率方面的优势, 为构建宽带射频微系统时域信号流仿真模型提供了新

的解决方案。

2 基本原理

为解决直接调用频域 S 参数模型导致的仿真效率低的问题, 本文提出一种 S 参数模型的时域表征和建模方法。该方法将 S 参数表征的器件均看作滤波器, S 参数确定了滤波器的预期频率响应, 基于预期频率响应构建一个 FIR 数字滤波器, 进而实现对输入信号的时域卷积和延迟量校正。该方法主要包含四个步骤: 基于 S 参数的频率响应数据插值, 自适应阶数 FIR 数字滤波器设计, 基于 FIR 数字滤波器对输入信号时域卷积, 输出信号延迟量校正。

2.1 基于 S 参数的频率响应数据插值

由于 S 参数模型通常以模型频率表征, 而 FIR 数字滤波器通常以数字角频率表征, 因此为构建 FIR 数字滤波器, 首先需将原始 S 参数的模拟频率映射到数字角频率, 并完成频率响应数据插值。

对于宽带宽带射频微系统而言, 假设系统仿真的采样率为 f_s , 而器件 S 参数的最大频率为 f_{\max} , 则得到在该器件工作频段内使得信号不发生失真和混叠的器件采样率 $f_s^{\text{component}}$ 为^[7]:

$$f_s^{\text{component}} = \max \{ f_s, 2 \cdot f_{\max} \} \quad (1)$$

从而基于式(2)将 S 参数的模拟频率 f 映射到 FIR 滤波器的数字角频率 ω :

$$\omega = \frac{2\pi f}{f_s^{\text{component}}} \quad (2)$$

此时, 原始 S 参数可表示为式(3), 其中为器件的端口数。

$$S = \begin{bmatrix} \omega_{f_1} & S_{f_1}^{11} & S_{f_1}^{12} & S_{f_1}^{13} & \dots & S_{f_1}^{LL} \\ \omega_{f_2} & S_{f_2}^{11} & S_{f_2}^{12} & S_{f_2}^{13} & \dots & S_{f_2}^{LL} \\ \omega_{f_3} & S_{f_3}^{11} & S_{f_3}^{12} & S_{f_3}^{13} & \dots & S_{f_3}^{LL} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \omega_{f_i} & S_{f_i}^{11} & S_{f_i}^{12} & S_{f_i}^{13} & \dots & S_{f_i}^{LL} \end{bmatrix} \quad (3)$$

为适应各类器件的任意频率响应特性, 需构建一个阶数 N 为奇数的 FIR 数字滤波器来逼近原始 S 参数的频率响应。下面以 S^{21} 为例来说明, 式(3)中的 S 参数矩阵其他维度的 S 参数建模方法与 S^{21} 一致。

为基本频率抽样法构建该 N 阶 FIR 数字滤波器, 要求预期频率响应样本点数等于滤波器长度 N , 并在单位圆上等间隔分布, 且样本点对应的频率响应关于 $\frac{N-1}{2}$ 共轭对称^[8]。

首先根据原始 S^{21} 参数通过线性插值得到数字角频率在 $[0, \pi]$ 范围的频率响应, 然后根据共轭对称原则扩展得到数字角频率在 $(\pi, 2\pi]$ 范围的频率响应, 从

而得到 FIR 数字滤波器完整的频率响应。

将数字角频率 ω 在 $[0, \pi]$ 范围上进行 $\frac{N-1}{2}$ 等分，则待插值样本点

$\left[\omega_0, \omega_1, \dots, \omega_{\frac{N-1}{2}} \right]$ 可通过式(4)计算得到：

$$\omega_i = i \cdot \frac{2\pi}{N-1} \quad \left(i = 0, 1, 2, \dots, \frac{N-1}{2} \right) \quad (4)$$

插值样本点所对应的 FIR 数字滤波器的幅度响应 $\left[h_0, h_1, \dots, h_{\frac{N-1}{2}} \right]$ 可通过式(5)

计算得到：

$$h_i^{21} = \frac{\omega_i - \omega_{\text{Interp1}}}{\omega_{\text{Interp2}} - \omega_{\text{Interp1}}} \cdot |S_{\text{Interp1}}^{21}| + \frac{\omega_{\text{Interp2}} - \omega_i}{\omega_{\text{Interp2}} - \omega_{\text{Interp1}}} \cdot |S_{\text{Interp2}}^{21}| \quad \left(i = 0, 1, 2, \dots, \frac{N-1}{2} \right) \quad (5)$$

其中， ω_{Interp1} 、 ω_{Interp2} 、 S_{Interp1}^{21} 、 S_{Interp2}^{21} 可依据式(6)得到：

$$\begin{cases} \omega_{\text{Interp1}} = 0, \omega_{\text{Interp2}} = \omega_{f_i}, S_{\text{Interp1}}^{21} = S_{f_i}^{21}, S_{\text{Interp2}}^{21} = S_{f_i}^{21} & \text{if } \omega_i < \omega_{f_i} \\ \omega_{\text{Interp1}} = \omega_{f_k}, \omega_{\text{Interp2}} = \omega_{f_{k+1}}, S_{\text{Interp1}}^{21} = S_{f_k}^{21}, S_{\text{Interp2}}^{21} = S_{f_{k+1}}^{21} & \text{if } \omega_{f_k} \leq \omega_i < \omega_{f_{k+1}} \\ \omega_{\text{Interp1}} = \omega_{f_L}, \omega_{\text{Interp2}} = \omega_{f_L}, S_{\text{Interp1}}^{21} = S_{f_L}^{21}, S_{\text{Interp2}}^{21} = S_{f_L}^{21} & \text{if } \omega_i \geq \omega_{f_L} \end{cases} \quad (6)$$

插值样本点所对应的 FIR 数字滤波器的线性相位响应 $\left[\varphi_0, \varphi_1, \dots, \varphi_{\frac{N-1}{2}} \right]$ 可通过

式(7)计算得到：

$$\varphi_i = -\omega_i \cdot \frac{N-1}{2} \quad \left(i = 0, 1, 2, \dots, \frac{N-1}{2} \right) \quad (7)$$

基于式(4)、式(5)和式(7)可得到 FIR 数字滤波器在 $[0, \pi]$ 范围上的频率响应为：

$$H_i = h_i e^{-j\varphi_i} \quad \left(i = 0, 1, 2, \dots, \frac{N-1}{2} \right) \quad (8)$$

根据 FIR 数字滤波器频率响应关于 $\frac{N-1}{2}$ 共轭对称的特点，可基于式(9)得到

$(\pi, 2\pi]$ 范围上的频率响应为：

$$H_i = h_{N-i} e^{-j(-\varphi_{N-i})} \quad \left(i = \frac{N+1}{2}, \frac{N+3}{2}, \dots, N-1 \right) \quad (9)$$

将式(8)和式(9)合在一起就组成了 FIR 数字滤波器在 $[0, 2\pi]$ 范围上的完整频率响应，如式(10)所示：

$$\mathbf{H} = \left[h_0 e^{-j\varphi_0}, h_1 e^{-j\varphi_1}, \dots, h_{(N-1)/2} e^{-j\varphi_{(N-1)/2}}, h_{(N-1)/2} e^{-j(-\varphi_{(N-1)/2})}, \dots, h_2 e^{-j(-\varphi_2)}, h_1 e^{-j(-\varphi_1)} \right] \quad (10)$$

以上是以 S^{21} 为例进行插值方法的说明，通过类似的方法可以得到其他端口参数 S^{ij} 所对应的频率响应。

2.2 自适应阶数 FIR 数字滤波器设计

FIR 数字滤波器的阶数直接影响了其逼近预期频率响应的精度，提高滤波器的阶数虽然可以有效提升精度，但也会带来计算量的增加^[9-11]，因此需找出满足预期精度要求的最小滤波器阶数，这一优化问题可表示为式(11)所示。

$$\min \{N\} \quad s.t. \quad e^{MSE} < \varepsilon \quad (11)$$

其中， e^{MSE} 代表 FIR 数字滤波器的幅频响应曲线与原始 S 参数曲线的均方误差，均方误差 MSE 可通过式(12)得到。

$$MSE = \frac{1}{N} \sum_{i=0}^{N-1} (h_i - |S_i^{21}|)^2 \quad (12)$$

自适应阶数优化算法采用粒子群优化 (Particle Swarm Optimization, PSO) 算法^[12-14]，流程如下：

- ① 首先，初始化粒子群并为每个粒子随机分配滤波器阶数 N_i ($i = 1, 2, L, m$)，其中 m 为粒子总数；
- ② 其次，计算每个粒子的适应度值，并根据适应度值更新个体历史最优值与全局历史最优值；
- ③ 然后，验证是否达到迭代停止条件，若条件满足，则终止优化并输出当前最优解，否则继续执行后续步骤；
- ④ 最后，根据公式(13)更新每个粒子的运动速度，随后根据公式(14)调整每个粒子的位置 (即 FIR 数字滤波器的阶数)，并跳转回步骤 2 进行下一次迭代。

$$\delta_{N_i}^{\%} = \omega \delta_{N_i} + c_1 (N_i^{p_best} - N_i) + c_2 (N_i^{g_best} - N_i) \quad (i = 0, 1, L, m) \quad (13)$$

$$N_i^{\%} = N_i + \text{floor} [\alpha \delta_{N_i}^{\%}] \quad (14)$$

其中， N_i 和 δ_{N_i} 分别表示第 i 个粒子的当前速度和位置， $N_i^{\%}$ 和 $\delta_{N_i}^{\%}$ 分别表示更新后的第 i 个粒子的速度和位置， $N_i^{p_best}$ 和 $N_i^{g_best}$ 分别表示第 i 个粒子的个体历史最优值和种群的全局历史最优值， ω 是惯性权重， c_1 和 c_2 分别是认知和社会学习因子，分别控制粒子的自我学习和群体学习行为，而 α 是控制更新过程的权重。

根据上述描述的 PSO 优化过程，可以确定满足所需建模精度的 FIR 数字滤波器的最小阶数。

2.3 基于 FIR 数字滤波器对输入信号时域卷积

在宽带宽带射频微系统的每个仿真步中,链路上器件之间传递的具有一定长度的时域信号流数据,信号流的长度由系统采样率和仿真时间窗共同确定。

输入信号与器件 S 参数频域模型的计算过程,可等效为输入信号与器件 FIR 数字滤波器模型的时域卷积计算过程。假设每个仿真步的信号流长度为 M , 则其与 N 阶 FIR 数字滤波器进行时域卷积计算后,将得到长度为 $M + N - 1$ 的输出信号流,其中各元素可通过式(15)计算得到^[15]:

$$y(k) = \sum_{i=0}^{N-1} b_i x(k-i) \quad (k=1,2,\dots,M+N-1) \quad (15)$$

2.4 输出信号延迟量校正

长度为 M 的输入信号流与 N 阶 FIR 数字滤波器进行时域卷积,将得到长度为 $M + N - 1$ 的输出信号流,其中只有与输入信号等长的数据是输出信号的有效部分。此外, N 阶 FIR 数字滤波器的群延迟由式(16)计算得到^[16-17]:

$$\tau = -\frac{d\varphi(\omega)}{d\omega} = \frac{N-1}{2} \quad (16)$$

从式(16)可知, N 阶 FIR 数字滤波器将对输入信号产生 $\frac{N-1}{2}$ 个抽样间隔的延迟量,因此从 $M + N - 1$ 的输出信号流中的第 $\frac{N-1}{2} + 1$ 开始,读取为 M 个数据点,即可实现对延迟量的校正,提取到输出信号的有效部分。

3 实例验证

本文通过以下两个实施例对上述 S 参数模型时域表征和建模方法进行验证。

实例 1:某低通滤波器 S 参数频率范围 100MHz~18GHz,工作频段 0.1~2GHz,输入信号为 $y(t) = \sin(2\pi f_1 t) + \sin(2\pi f_2 t)$, $f_1 = 2\text{GHz}$, $f_2 = 12\text{GHz}$, 设定 FIR 数字滤波器的预期精度为 10^{-2} 。

基于 PSO 算法的自适应结果如图 1 所示,最优阶数为 273。输出信号延迟修正结果如图 2 所示。图 3、图 4 所示的功率谱表明,输出信号中的高频信号分量被有效过滤。

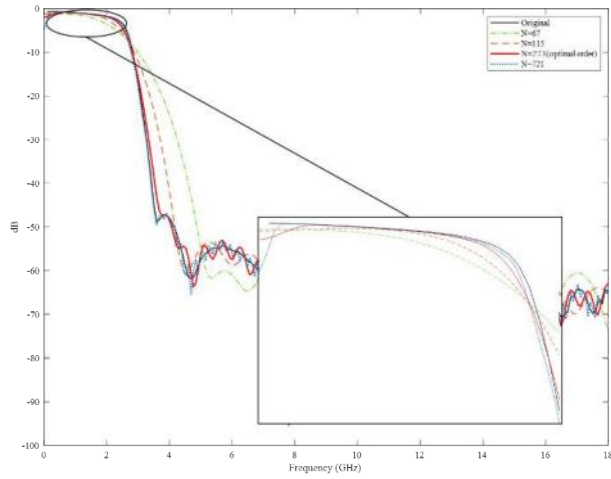


图 1 低通滤波器自适应阶数 FIR 滤波器设计结果

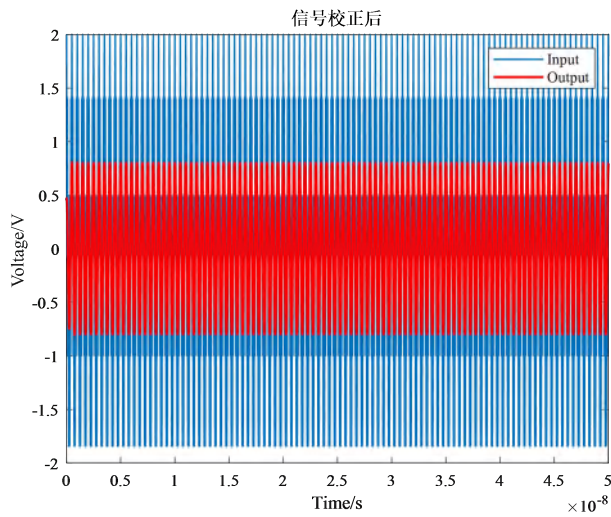
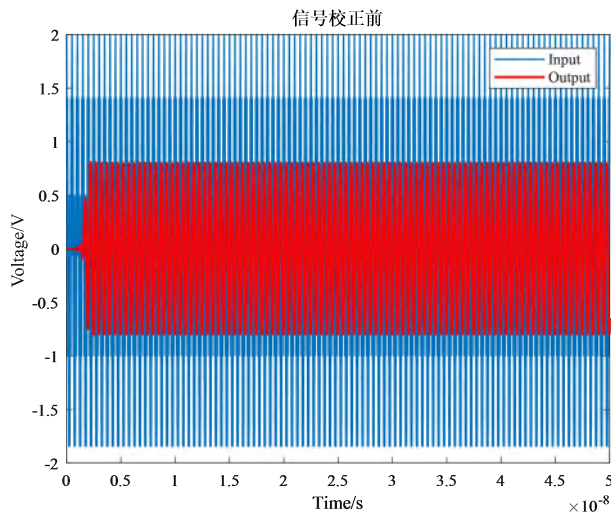


图 2 低通滤波器输出信号延迟量修正

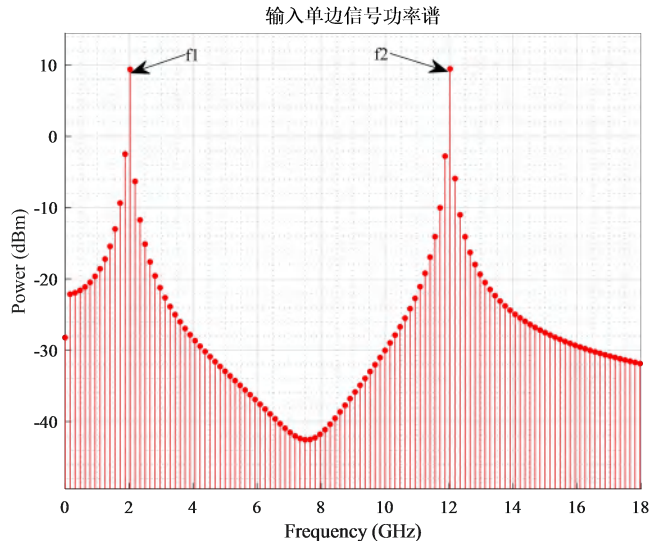


图 3 低通滤波器输入信号功率谱

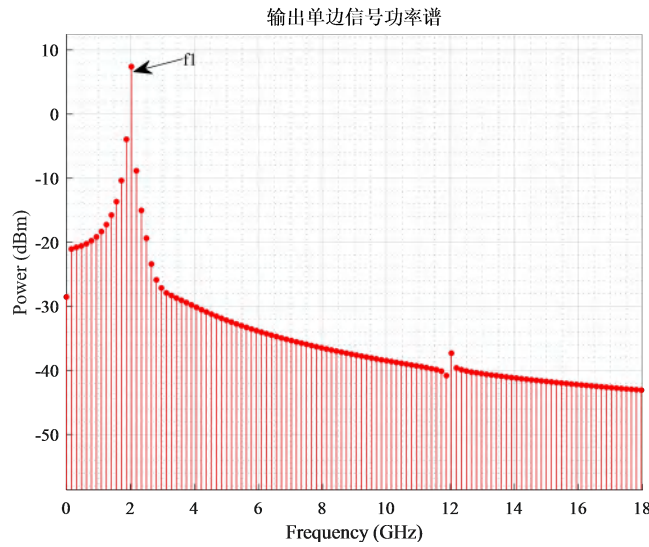


图 4 低通滤波器输出信号功率谱

实例 2：某低噪声放大器 S 参数频率范围 1000MHz~10GHz，工作频段 0.7~5GHz，输入信号为 $y(t) = \sin(2\pi f_1 t) + \sin(2\pi f_2 t)$ ， $f_1 = 3\text{GHz}$ ， $f_2 = 6.2\text{GHz}$ ，设定 FIR 数字滤波器的预期精度为 10^{-2} 。

基于 PSO 算法的自适应结果如图 5 所示，最优阶数为 373。输出信号延迟修正如图 6 所示。图 7、图 8 所示的功率谱表明，信号处于工作频段内的放大倍数和处于工作频段外的放大倍数有明显差异。

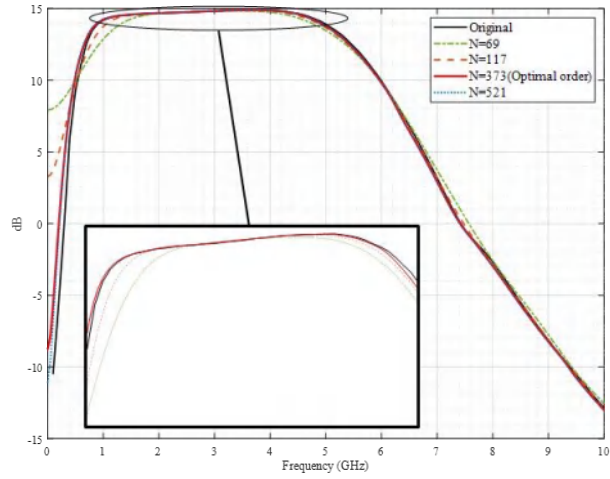


图 5 低噪声放大器自适应阶数 FIR 滤波器设计结果

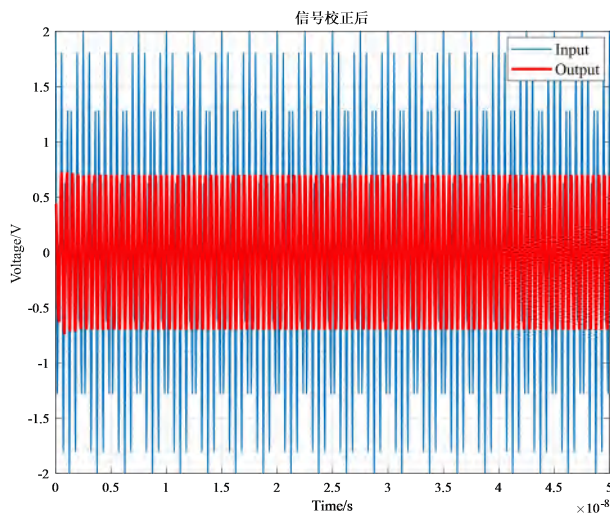
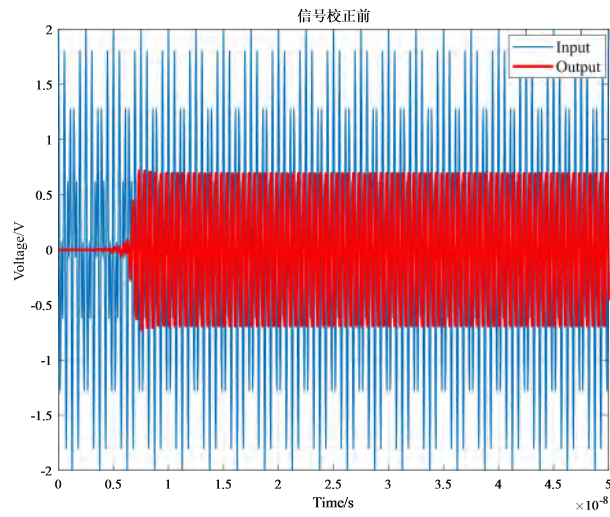


图 6 低噪声放大器输出信号延迟量修正

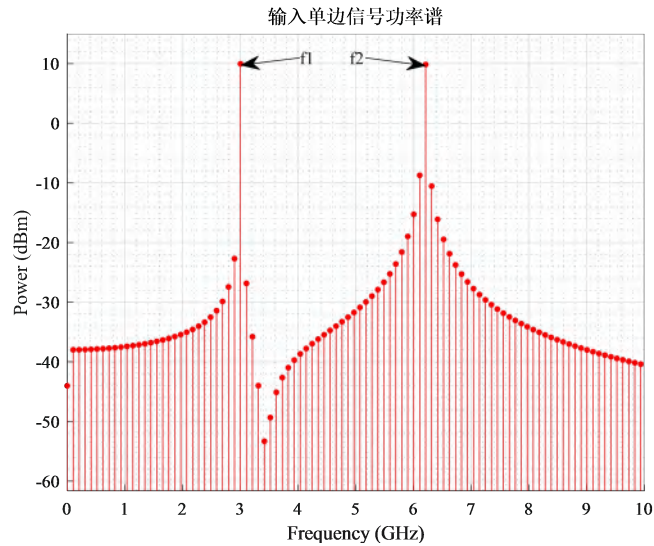


图 7 低噪声放大器输入信号功率谱

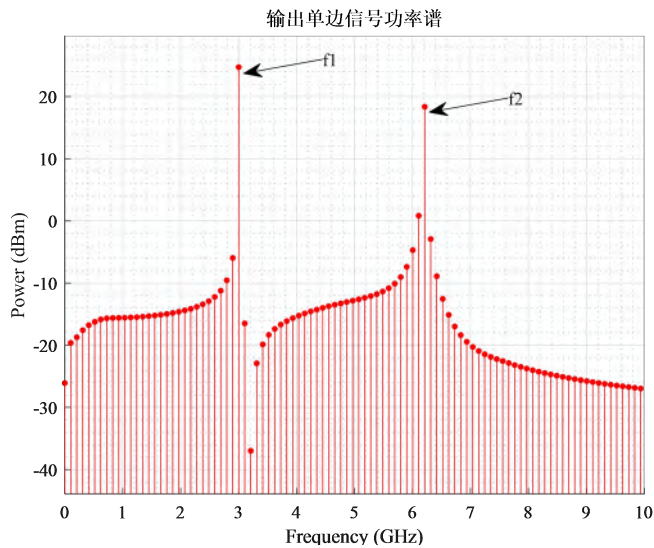


图 8 低噪声放大器输出信号功率谱

4 结论

针对宽带射频微系统链路在时域信号流仿真中存在的问题,本文提出了一种基于 FIR 数字滤波器的 S 参数模型时域表征方法。该方法通过构建符合器件 S 参数频率响应特性的 FIR 数字滤波器,实现了对输入信号的直接时域卷积运算。与传统的频域 S 参数模型相比,本文方法无需在每个仿真步进行时频变换,从而显著提高了仿真效率。同时, FIR 数字滤波器的采用不仅避免了频域离散化引起的信号混叠问题,还确保了信号传输过程中的相位完整性,相较于传统方法展现出更为显著的技术优势。

参考文献

- [1] Ronny Frevert. Modeling and Simulation for RF System Design[M]. New York: Springer US, 2005.
- [2] L. Pantoli, D. Spina, D. Romano, G. Antonini, G. Leuzzi & T. Dhaene. Advanced time-domain analysis of RF and microwave circuits for enhanced system-level

- simulations[J]. IEEE Transactions on Microwave Theory and Techniques, 2023, 71(5): 2105-2117.
- [3] Liu W, Zhu L, Feng F, et al. A Time Delay Neural Network Technique for Nonlinear Microwave Device Modeling[J]. Micromachines, 2020, 11(9): 831.
- [4] Hung, Chih De, et al. A Unified Behavior Model of Low Noise Amplifier for System-level Simulation[C]. European Conference on Wireless Technology IEEE, 2006.
- [5] Michel C. Jeruchim. Simulation of Communication Systems: Modeling, Methodology and Techniques[M]. New York: Springer US, 2000.
- [6] S. Koziel, A. Pietrenko-Dabrowska & U. Ullah. Reduced-cost microwave modeling using constrained domains and dimensionality reduction[J]. Scientific Reports, 2023, 13(1): 18509.
- [7] 王颖,刘娟,郝鑫,等.用于太赫兹高速通信的 FIR 滤波器 64 并行实现算法[J].太赫兹科学与电子信息学报,2020,18(4):560-564.
- [8] 张书玉,王婷.FIR 数字滤波器的智能优化算法设计综述 [J].电子技术应用,2021,047(5):29-34.
- [9] 孙广富,宋捷,鲁祖坤,等.滤波器阶数对导航接收机时域抗干扰的影响分析[J].国防科技大学学报,2023,45(5):60-71.
- [10] 张勇刚,李宁,郝燕玲.改进的变阶数 LMS 自适应滤波算法[J].哈尔滨工程大学学报,2010,31(3):350-354.
- [11] 隋天宇,张宝华,李宇,黄海宁,张春华.一种自适应均衡器最优阶数确定算法[J].仪器仪表学报,2012,33(5):1187-1194.
- [12] 林川,冯全源.应用于自适应格型 RLS 滤波器的变阶数算法 [J].信号处理,2010,26(2):298-302.
- [13] 李爱国,覃征,鲍复民,等.粒子群优化算法[J].计算机工程与应用,2002,38(21):1-3.
- [14] Lv Zhensu, Hou Zhirong. Particle Swarm Optimization with Adaptive Mutation[J]. Acta Electronica Sinica, 2004, 32(3): 416-420.
- [15] 郭勇,杨欢.基于 IP 核 FIR 滤波器的设计与 FPGA 实现[J].无线电工程,2017,47(01):79-82.
- [16] 伏燕军,程强强,于润桥,等.信号 FIR 数字滤波后相位延迟的消除[J].计算机工程与应用,2012,48(7):146-149.
- [17] 张红梅,韩万刚.FIR 滤波器时延分析及其在信号融合中的应用[J].武汉大学学报:工学版,2016,49(2):6.

陶瓷基毫米波变频收发微系统仿真技术研究

余希猛, 杨振涛, 于斐, 刘洋, 刘林杰

(中国电子科技集团公司 第十三研究所, 石家庄 050051)

摘要: 随着现代通信技术的飞速发展, 变频收发微系统向着多通道、高集成、高频率、小型化的方向发展。基于高温共烧陶瓷 (HTCC) 工艺, 提出了一种适用于毫米波频段的高可靠陶瓷基变频收发微系统封装方案。借助仿真软件对传输路径进行了优化, 仿真表明, 在 DC~20GHz 频段内, 变频收发垂直通道的回波损耗优于 12 dB, 插入损耗优于 0.9 dB。对加工的样品进行了测试, 测试与仿真具有较好的一致性, 本研究为高可靠陶瓷基集成微系统的封装设计与制造提供了参考。

关键词: 高温共烧陶瓷 (HTCC); 球栅阵列 (BGA); 变频收发微系统; 阻抗匹配; 电磁仿真

作者简介:

余希猛 (1993—), 男, 河南信阳人, 硕士, 高级工程师, 主要从事射频微系统封装研发工作。邮箱: 2686867405@qq.com, 联系电话: 15188578663

0 引言

随着现代通信技术的不断更新, 变频收发微系统正朝着多通道、高集成度、高频率及小型化方向迅速发展, 以满足未来在高速数据通信、信息探测及空间电子等领域的苛刻应用需求。毫米波频段的信道容量与传输速率的优势使得毫米波变频收发系统在新一代通信与感知系统中占据着越来越重要的地位^[1-2]。传统毫米波变频收发系统多采用金属箱体封装, 其封装结构虽然具有良好的电磁屏蔽性能, 但同时带来了体积大、重量重、组装复杂等问题, 严重限制了系统的小型化与模块化发展^[3-4]。为实现对高性能、高可靠毫米波变频收发微系统的有效封装, 陶瓷基毫米波变频收发微系统采用陶瓷基板作为载体, 融合了微细加工、厚膜印刷、多层布线等先进工艺技术, 成为提升射频微系统性能的关键方向^[5-7]。科研人员基于该技术针对毫米波变频收发微系统开展了很多研究。例如, 张鸣一等人^[8]制备了一款三维集成 Ku 波段四通道收发模块, 其利用球栅阵列高密度连接的优点, 结合高温共烧陶瓷基板, 实现了收发模块的小型化。

本文设计了适用于毫米波频段的高可靠陶瓷基变频收发微系统封装结构, 采用高温共烧陶瓷 (HTCC) 工艺制备封装基板, 通过理论分析和仿真优化对水平互连结构和垂直互连结构的传输性能进行了分析研究。通过优化键合金丝互连结构和焊球互连结构处的阻抗, 有效实现毫米波信号的高效传输。最后, 通过实验测试和对比分析, 验证了传输通道良好的传输性能。

1 陶瓷基毫米波变频收发微系统结构设计

毫米波变频收发微系统的主要任务是完成射频信号与本振信号之间的频率变换,同时实现收发信号链路的放大、滤波及路径控制。本文设计了一种适用于毫米波频段的高可靠陶瓷基变频收发微系统封装结构,整体结构设计如图1所示。其集成了四个接收模块(R1~R4)和四个发射模块(T1~T4),在发射链路中,本振信号经上变频器变换至目标毫米波频段,经功率放大器放大后再通过功分器将能量分配至多个天线,实现空间多路发射;在接收链路中,多个天线接收到的信号通过功分器合路后送入低噪声放大器,再经滤波与下变频处理,输出到后端处理系统。收发模块均通过直径0.76 mm、节距1.00 mm的焊球与印制电路板(PCB)连接,实现板级系统集成。从图中可以看出,本振信号LO1通过功分链路A1、A2实现信号的高效分配,LO2通过功分链路A3、A4实现信号的高效分配,B1~B4和C1~C2为滤波路径。

整个陶瓷基封装微系统的尺寸为48.00 mm×48.00 mm×4.65 mm,在有限体积内实现了多通道信号的高效收发,显著提升了系统的集成度与可靠性,满足毫米波频段高频率、低损耗、高隔离度的电性能要求。

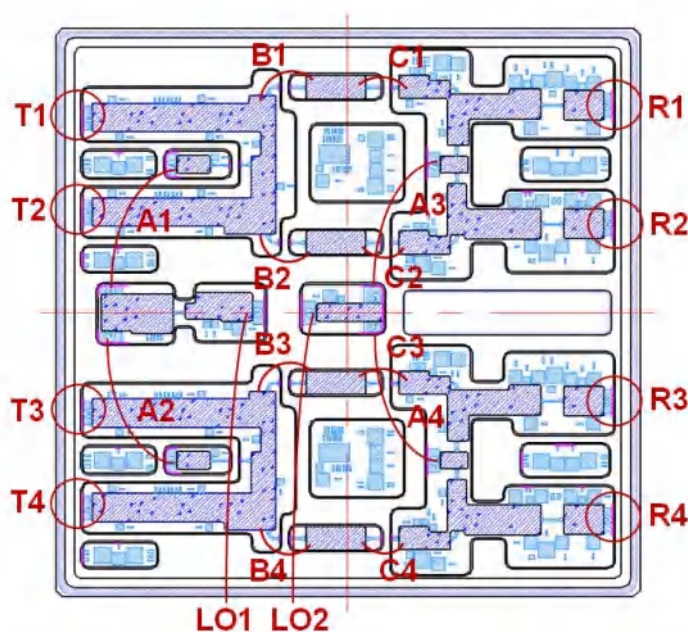


图1 陶瓷基毫米波变频收发微系统封装结构示意图

2 传输性能分析

陶瓷基毫米波变频收发微系统主要的传输路径包括了水平传输结构和垂直传输结构,此外还包括键丝互连结构和焊球互连结构。为了保证毫米波信号的传输性能,需要仔细设计传输线结构,主要从三个方面进行设计优化,一是实现传输损耗的有效控制,使信号在封装结构内传输时保持较高的幅度,从而保证信号强度和传输质量。二是保证传输线阻抗的一致性,避免由于阻抗不连续性导致的反射

和驻波效应。三是保证合理的屏蔽层和接地层结构，减小辐射损耗和串扰，避免信号串扰和电磁兼容性问题。本文首先研究了水平互连简化结构和垂直互连简化结构的基本结构参数，然后以基本结构参数为基础，对实际产品的传输路径进行了仿真优化。

2.1 水平互连简化结构

如图 2 所示，水平互连传输路径采用了带状线传输结构，其传输结构由上下接地层、内部信号线和介质层组成。图中 w 为信号线线宽、 t 为信号线厚度、 b 为上下接地层间距、 h 为信号线与接地层距离。由于信号线被上下接地层包围，电磁场被限制在介质层内，其抗干扰能力较强，可以有效降低信号线间的电磁耦合。

图 3 是水平互连简化结构的阻抗计算结果，选取的陶瓷材料为氧化铝陶瓷，其相对介电常数 (ϵ_r) 为 9.8，介电损耗因子 ($\tan \delta$) 为 0.003。基于高温共烧陶瓷技术，每层陶瓷片的厚度 ($H1$) 为 0.20 mm，金属化厚度 ($T1$) 为 0.01 mm，信号线宽度 ($W1$) 为 0.06 mm，信号线与地平面间距离 ($D1$) 为 0.20 mm，通过图 3 可以看出其阻抗值约为 48.18Ω 。

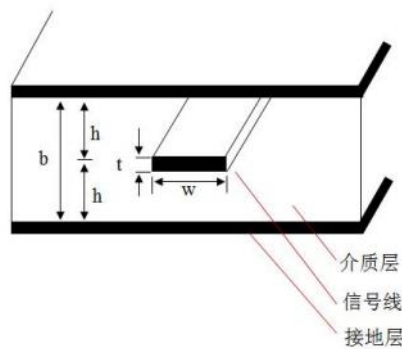


图 2 水平互连简化结构图

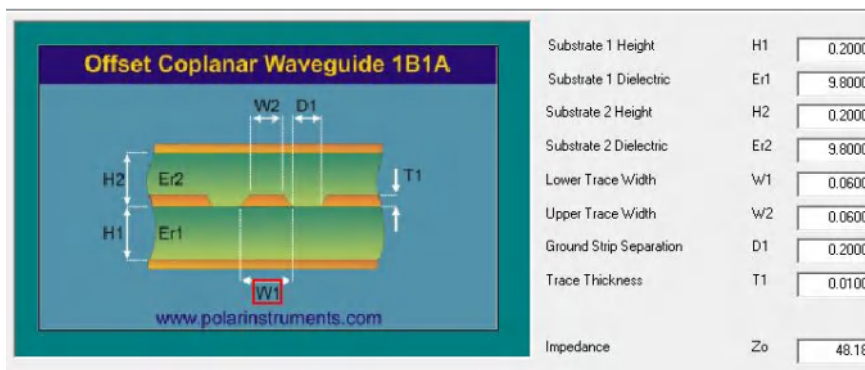


图 3 水平互连简化结构阻抗计算图

2.1 垂直互连简化结构

如图 4 所示，垂直互连传输路径采用了类同轴传输结构，其中主要影响垂直

传输结构阻抗的是信号孔半径 b ，信号孔与地孔之间的距离 a 。该类同轴结构可以采用同轴线相关理论公式进行设计。

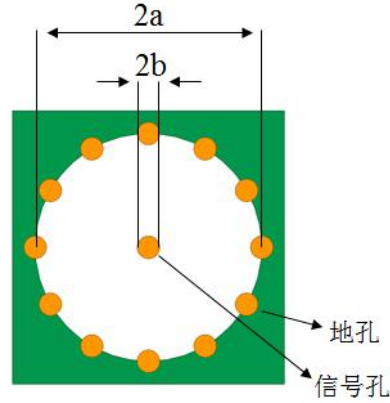


图 4 垂直互连简化结构图

同轴线特征阻抗 (Z_c) 计算公式为

$$Z_c = \frac{1}{2\pi} \sqrt{\frac{\mu_0}{\epsilon_r \epsilon_0}} \ln\left(\frac{a}{b}\right) \quad (1)$$

首先需要确定垂直过孔的直径,再根据同轴线特征阻抗计算公式推导回流地孔与信号孔的距离。设定信号孔半径为 $b = 50\mu\text{m}$ ，计算得到回流地孔与信号孔的距离为 $a = 620\mu\text{m}$ ，式中 ϵ_0 为真空介电常数； ϵ_r 为氧化铝陶瓷相对介电常数； μ_0 为真空磁导率。 μ_0 为 $1.26 \times 10^{-6} \text{ H/m}$ ， ϵ_0 为 $8.85 \times 10^{-12} \text{ F/m}$ ， ϵ_r 为 9.8，带入各项参数计算得到

$$Z_c = \frac{1}{2\pi} \sqrt{\frac{1.26 \times 10^{-6}}{9.8 \times 8.85 \times 10^{-12}}} \ln\left(\frac{0.620}{0.050}\right) \approx 48.1(\Omega) \quad (2)$$

从计算结果可以看出，其特征阻抗理论值约为 48.1Ω ，基本满足单端 50Ω 的设计要求。图 5 展示了垂直互连简化结构的电磁场分布图及传播模式，垂直传输结构的主模是 TEM 模，低次模是 TE₁₁ 模，其截至波长 (λ_c) 为

$$\lambda_c = \pi(a+b) \quad (3)$$

则上限工作频率 (f_c) 为

$$f_c = \frac{c_0}{\sqrt{\epsilon_r} \pi(a+b)} = \frac{3 \times 10^8}{\sqrt{9.8} \pi (0.620 + 0.050) \times 10^{-3}} \approx 45.5(\text{GHz}) \quad (4)$$

式中 c_0 为光速，值为 $3 \times 10^8 \text{ m/s}$ 。由此可知，TE₁₁ 模式只有在系统频率超过

46GHz 时, 才会被激发。从图 5 可以看出, TE₁₁ 模式被激发的频率约为 50GHz, 其值基本与理论计算值相符。

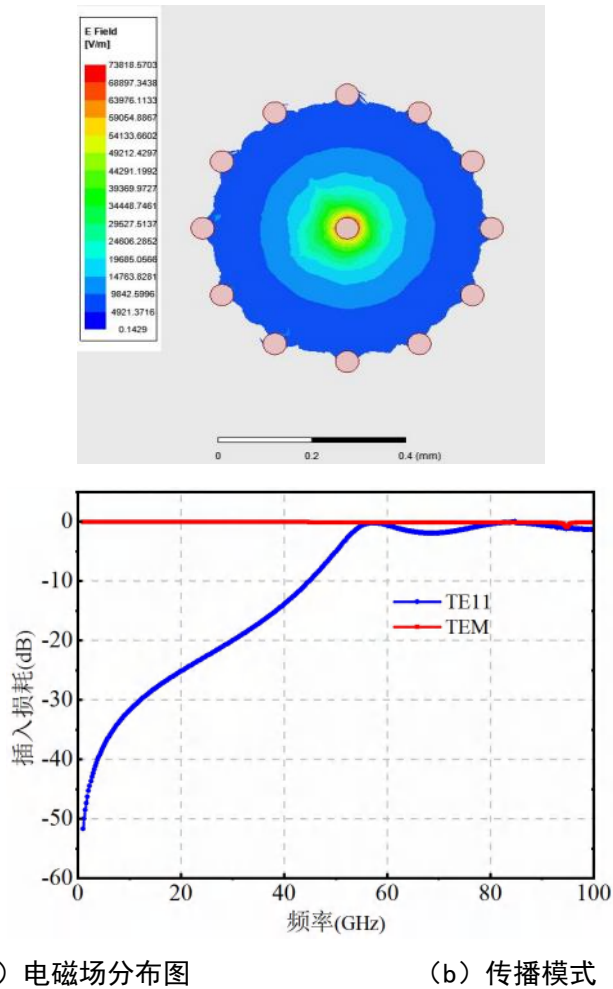


图 5 垂直互连简化结构的仿真结果

2.3 水平互连复杂结构

基于水平互连简化结构的基本结构参数, 进行了实际产品的水平传输路径设计。图 6 展示了实际产品的水平传输路径仿真模型, 模型包括了芯片、键合金丝和水平传输路径。其中芯片材质为砷化镓 (GaAs), 芯片厚度为 0.10mm, 芯片的相对介电常数为 12.9。键合金丝的材质为金 (Au), 直径为 0.03mm, 长度为 0.30mm, 拱高为 0.20mm。图 7 (a) 展示了键合金丝对传输性能的影响情况。键合金丝起到连接芯片和水平传输路径的作用, 键合金丝的高感性是导致整个传输路径失配的最大因素, 通过双丝并联, 降低感值, 极大的改善了路径传输性能。由图 7 (a) 可见, 双丝键合互连结构在 DC~40 GHz 频段内, 回波损耗 ≤ -13 dB, 插入损耗 ≥ -4.0 dB。图 7 (b) 展示了垂直过孔形式对传输性能的影响情况。采用阶梯过孔可以有效改善垂直转水平结构的阻抗失配, 从而改善 30GHz~40GHz 频段的传输参数, 使得传输性能提升到 40GHz 以上。由图 7 (b) 可见, 阶梯过

孔互连结构在 DC~40 GHz 频段内，回波损耗 ≤ -13 dB，插入损耗 ≥ -4.0 dB。

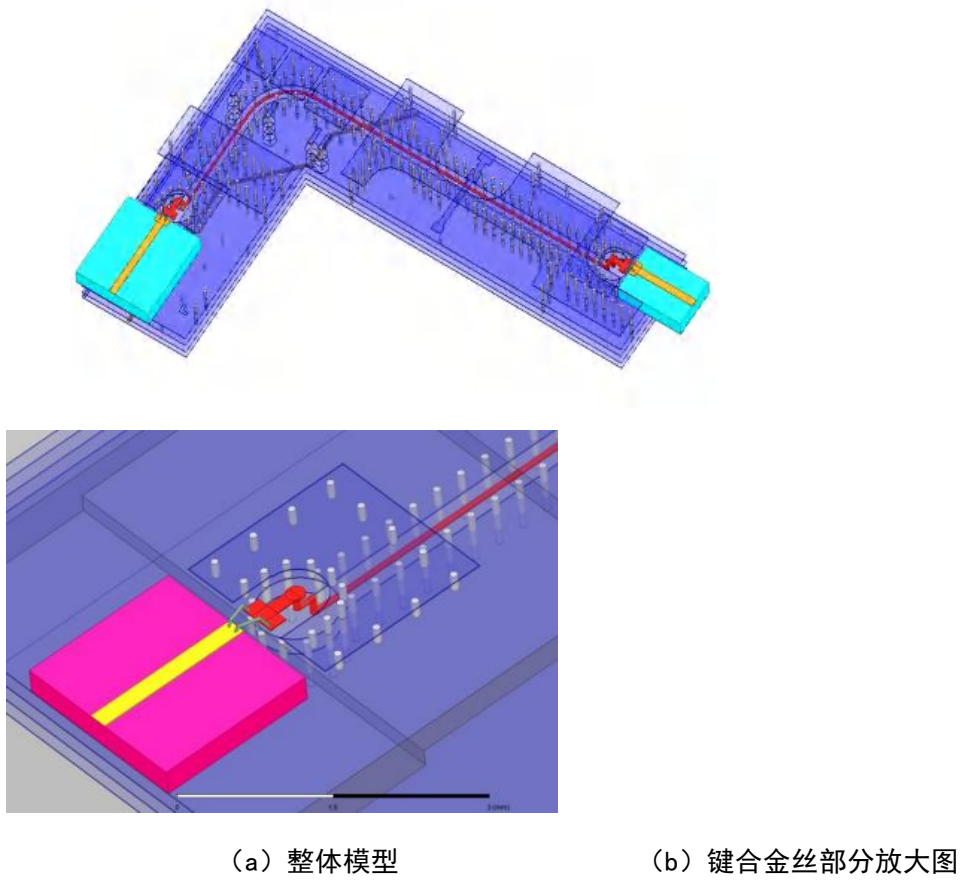
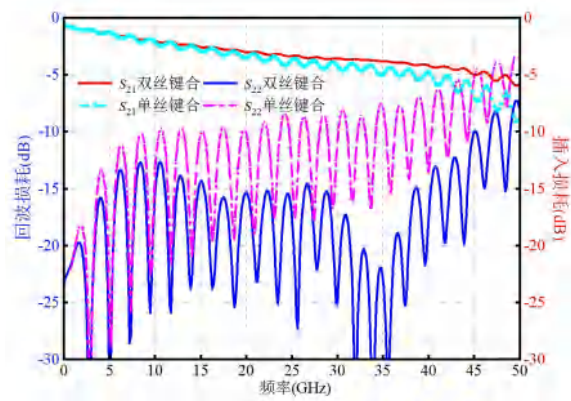
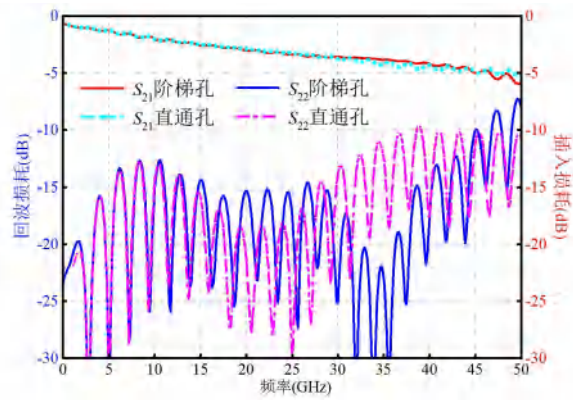


图 6 实际产品的水平传输路径仿真模型



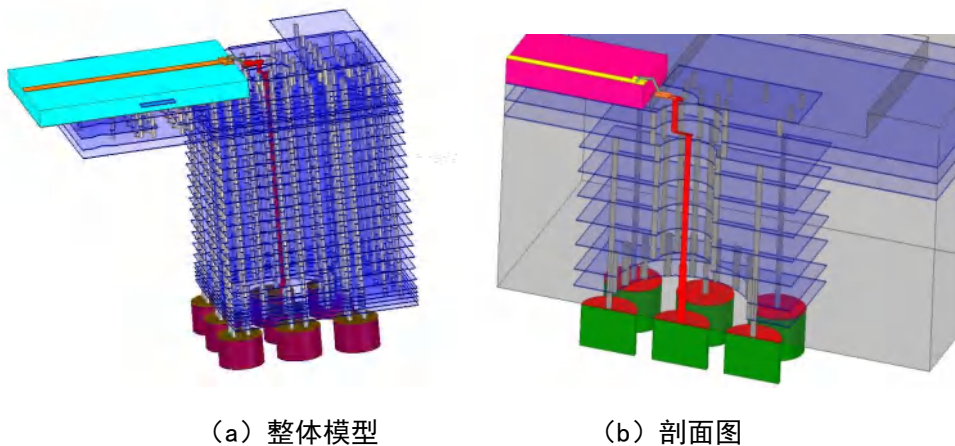


(a) 键合金丝对传输性能的影响 (b) 垂直过孔形式对传输性能的影响

图 7 水平互连复杂结构的仿真结果

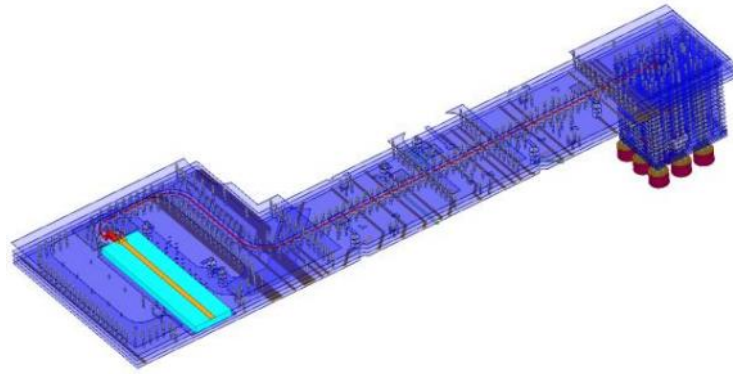
2.4 垂直互连复杂结构

基于垂直互连简化结构的基本结构参数,进行了实际产品的垂直传输路径设计。图 8 展示了实际产品的垂直传输路径仿真模型,模型包括了芯片、键合金丝、垂直传输路径以及 BGA 焊球。其中芯片参数和键合金丝的参数与水平传输路径仿真模型中的参数一致, BGA 焊球的材质为铅锡合金 (Sn63Pb37),焊球之间的节距为 1.00mm,焊球的直径为 0.76mm,焊球的高度为 0.50mm。图 9 展示了垂直互连结构的传输性能仿真结果。BGA 焊球是导致整个传输路径失配的最大因素,增大 BGA 焊球附件的反焊盘可以改善焊球结构的阻抗失配,但改善效果不明显,且回流地孔与信号孔距离过大可能导致电磁泄露,引入谐振。由图 9 可见,在 DC~20 GHz 频段内,无水平互连复杂结构时,回波损耗 ≤ -13 dB,插入损耗 ≥ -1.0 dB;增加水平互连复杂结构时,回波损耗 ≤ -13 dB,插入损耗 ≥ -4.0 dB。



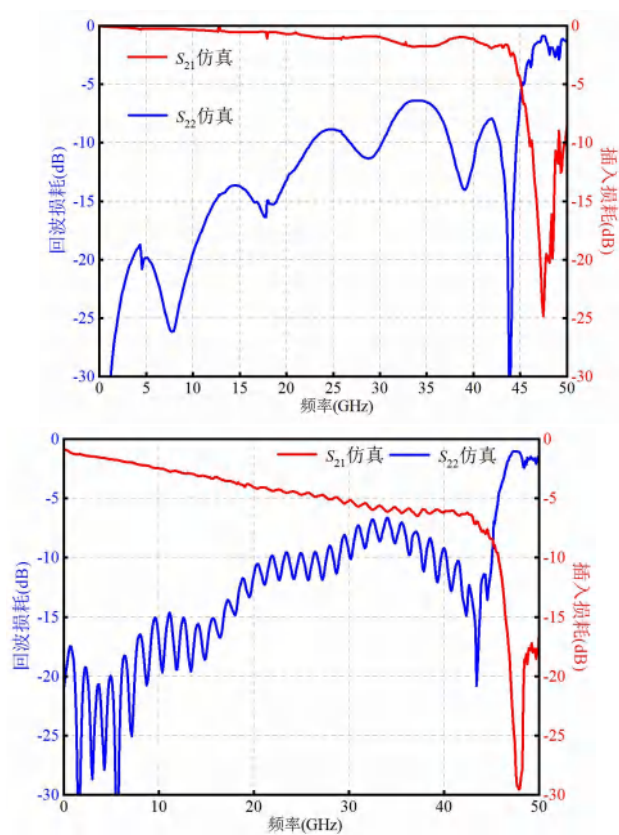
(a) 整体模型

(b) 剖面图



(c) 水平互连复杂结构+垂直互连复杂结构

图 8 实际产品的垂直传输路径仿真模型



(a) 垂直互连复杂结构

(b) 水平互连复杂结构+垂直互连复杂结构

图 9 垂直互连复杂结构的仿真结果

3 样品制备与测试

基于高温共烧陶瓷技术制备了样品，整体加工流程如图 10 所示。陶瓷生瓷片经过冲孔填孔、印刷、叠片、层压和热切形成生瓷件，然后生瓷件经过烧结形成瓷件，瓷件再经过镍、金镀覆形成陶瓷成品，最后对成品进行电性能测试，测试合格的成品入库保存。

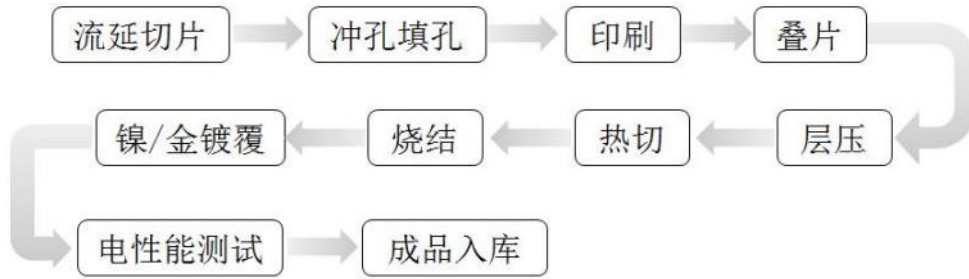
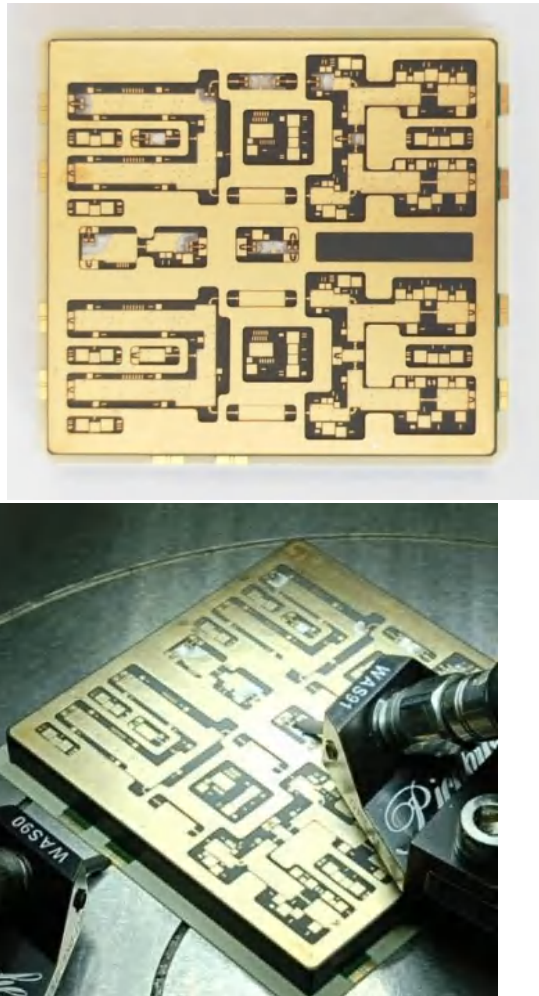


图 10 高温共烧陶瓷加工流程图

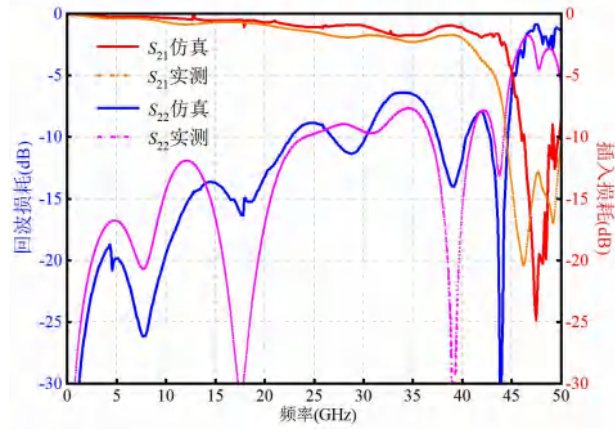
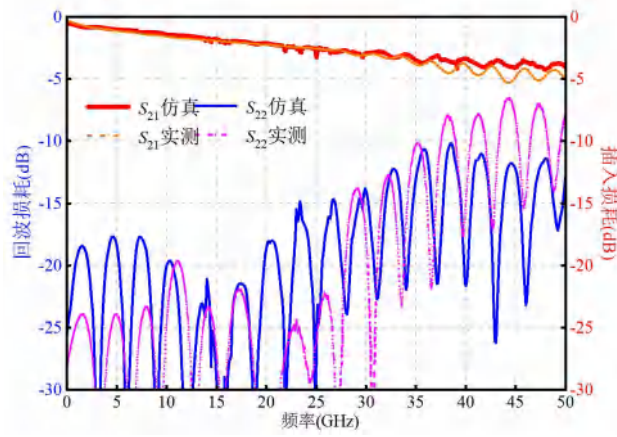
使用带宽67 GHz的矢量网络分析仪对S参数进行测量，利用地-信号-地（GSG）结构的高频探针与陶瓷基毫米波变频收发微系统搭接。测试示意图如图11所示。陶瓷基毫米波变频收发微系统传输性能测试与仿真对比结果如图12所示，从图12（a）中可以看出，在DC~40 GHz频段，水平互连复杂结构的实测回波损耗 ≤ -10 dB，插入损耗 ≤ 4.0 dB。从图12（b）中可以看出，在DC~20 GHz频段，垂直互连复杂结构的实测回波损耗 ≤ -12 dB，插入损耗 ≤ 0.9 dB。从图12（c）中可以看出，在DC~20 GHz频段，水平互连复杂结构+垂直互连复杂结构的实测回波损耗 ≤ -15 dB，插入损耗 ≤ 4.5 dB。测试结果与仿真结果基本相符。



(a) 样品

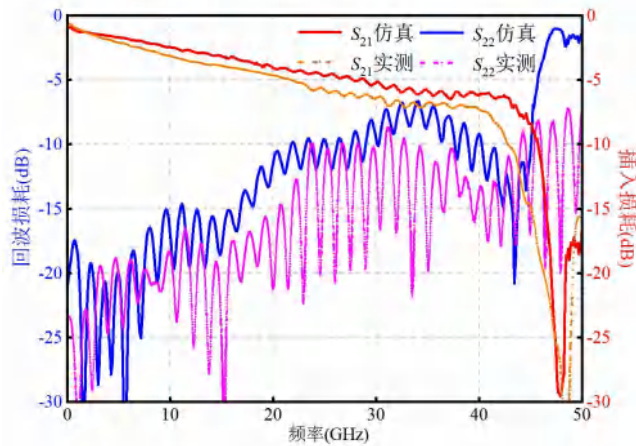
(b) 测试示意图

图 11 陶瓷基毫米波变频收发微系统封装样品和传输性能测试图



(a) 水平互连复杂结构

(b) 垂直互连复杂结构



(c) 水平互连复杂结构+垂直互连复杂结构

图 12 陶瓷基毫米波变频收发系统传输性能测试与仿真对比图

4 结论

本文研究了一种高可靠陶瓷基毫米波变频收发微系统封装基板。通过理论分

析和仿真优化对水平互连结构和垂直互连结构的传输性能进行了分析研究。通过优化键合金丝互连结构和焊球互连结构处的阻抗,有效实现毫米波信号的高效传输。实测结果表明在 DC~40 GHz 频段,水平互连复杂结构的实测回波损耗 ≤ -10 dB,插入损耗 ≤ 4.0 dB。在 DC~20 GHz 频段,垂直互连复杂结构的实测回波损耗 ≤ -12 dB,插入损耗 ≤ 0.9 dB。本文的封装结构与设计方法可为陶瓷封装设计领域的科研人员进行射频集成微系统设计提供参考。

参考文献

- [1] 邹雷. Ka 波段上下变频模块研究[D]. 成都: 电子科技大学, 2017.
- [2] 童凌云. 毫米波变频与放大芯片及接收电路技术研究[D]. 成都: 电子科技大学, 2022.
- [3] 夏令希,刘勇,马传武,等.Ka 波段宽带收发前端设计[J].微波学报, 2023, 39(S01):285-287.
- [4] 张生春,郭垚,王洁,等.一种高集成度 X 波段八通道变频收发组合的设计[J].火控雷达技术, 2023, 52(3):90-96.
- [5] 韩世宏. 基于 HTCC 的射频微系统关键技术研究[D]. 成都: 电子科技大学, 2022.
- [6] 过方舟,徐锐敏. 系统级封装关键技术研究进展 [J]. 微波学报, 2014, 30(S1): 588-593.
- [7] 赵燕,吴鹏,王志凌,等. 微波毫米波 SiP 三维集成陶瓷封装基板研究进展 [J]. 微波学报, 2024, 40(S1): 256-259.
- [8] 张鸣一,朱春雨,刘文豹,等. 基于三维集成技术的 Ku 波段四通道 T/R 模块 [J]. 半导体技术, 2021, 46(4):286-289.

作者简介:

余希猛 (1993—), 男, 河南信阳人, 硕士, 高级工程师, 主要从事射频微系统封装研发工作。邮箱: 2686867405@qq.com, 联系电话: 15188578663



一种面向射频电路的可连续调节变压器

汪芳倩, 单光宝, 张攀, 郑彦文

西安电子科技大学集成电路学部

摘要: 本文提出了一种面向射频电路的宽带可连续调谐变压器。可以通过调制施加的电压来改变磁通量。首先, 探明了变压器的感值随电压幅值和相位差的变化规律。其次, 建立几何模型验证理论分析, 仿真结果表明, 在 DC~25 GHz 范围内, 主、次电感的调谐范围分别为 75.0%和 71.4%。结果证实, 可调谐变压器提供了需要动态阻抗匹配和多频段操作的可重构电路应用。

关键词: 片上变压器; 可重构; 射频电路

1 前言

在电子设备小型化和多功能集成的实现过程中, 面临集成能力有限和功能固化的挑战。片上变压器是射频电路的基本组成部分, 由于其布局紧凑且能完成多种电路功能, 所以常用于集成电路中。变压器可以实现阻抗匹配, 抑制带外响应和产生传输零点[1],[2],[3]等功能。因此, 变压器调谐是实现射频电路重构的有效途径。目前, 变压器重构主要依赖于开关[4]、变容二极管[5]、电感[6]、MOS[7]和双路变压器(DPT), DPT通过激励信号[8]的相位差进行调谐。然而, 基于MOS的可调变压器的调谐范围受到耦合系数的限制。在DPT中, 控制方法复杂, 通用性差。而且, 上述方法难以连续调节变压器感值。同时, 在实现过程中引入了寄生参数, 增加了版图面积。基于磁耦合[9]的调谐方法具有控制方式统一、功耗低、连续调节、调节范围大等优点, 具有很大的发展潜力。由于没有额外的元素, 避免了引入外部寄生参数。本文提出了一种基于磁耦合的可调变压器, 为射频可重构电路的实现提供了新方法。

2 调谐方法

2.1 基本原理

本文提出的可调变压器原理如图1所示, 其中 R_1 、 R_2 和 R_3 分别是是电感的寄生电阻, L_1 、 L_2 和 L_3 是自感值, V_1 、 V_2 和 V_3 是电感两端的电压, k_{12} 、 k_{13} 和 k_{23} 表示两两电感之间的耦合系数。

图1(a)表示可调变压器的原理图, L_1 为控制电感, L_2 和 L_3 共同构成受控变压器。图1(b)的电流电压关系可以表示为

$$\begin{aligned}
V_1 &= j\omega L_1 i_1 + j\omega M_{12} i_2 + j\omega M_{13} i_3 \\
V_2 &= j\omega L_2 i_2 + j\omega M_{12} i_1 + j\omega M_{23} i_3 \\
V_3 &= j\omega L_3 i_3 + j\omega M_{13} i_1 + j\omega M_{23} i_2
\end{aligned} \tag{1}$$

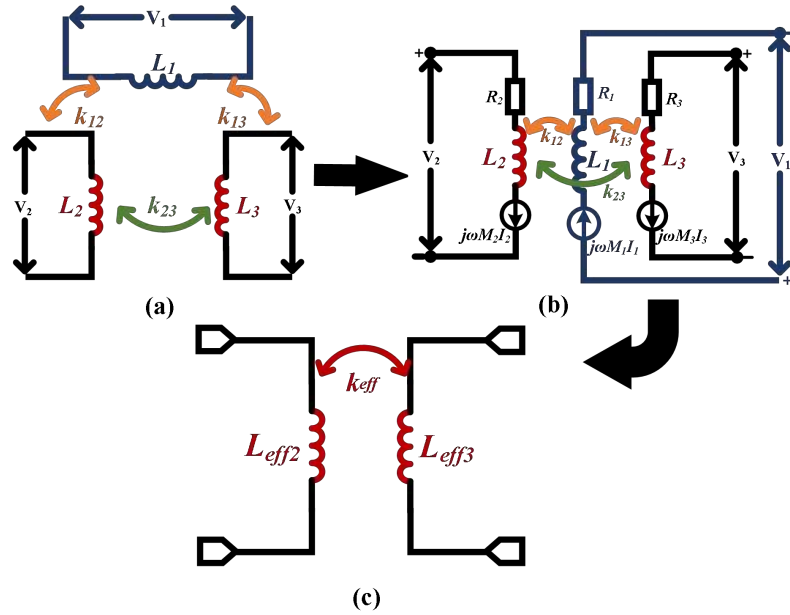


图 1 可调变压器 (a)原理图 (b) 等效电路图 (c) 常规等效图

假设 $\alpha = \frac{V_2}{V_1} = \left| \frac{V_2}{V_1} \right| e^{j\angle(\phi_2 - \phi_1)}$, $\beta = \frac{V_3}{V_1} = \left| \frac{V_3}{V_1} \right| e^{j\angle(\phi_3 - \phi_1)}$, $M_{12} = M_{13} = M$, $L_1 = L_2 = L_3 = L$, 则

$$\begin{aligned}
\frac{i_1}{i_2} &= \frac{(M_{23}^2 - L^2) + (\alpha + \beta)M(L - M_{23})}{M(L - M_{23}) + \alpha(M^2 - L^2) + \beta(LM_{23} - M^2)} \\
\frac{i_1}{i_3} &= \frac{(M_{23}^2 - L^2) + (\alpha + \beta)M(L - M_{23})}{M(L - M_{23}) + \alpha(LM_{23} - M^2) + \beta(M^2 - L^2)}
\end{aligned} \tag{2}$$

将图 1 (b) 降阶为图 1 (c) , 用三端口网络降阶的方法可以得到

$$V_1 = -(i_2 + i_3)Z_1 \tag{3}$$

$$Z_1 = -(j\omega L_1 \frac{i_1}{i_2 + i_3} + j\omega M) \tag{4}$$

因此

$$V_2 = j\omega(L + M \frac{i_1}{i_2 + i_3})i_2 + j\omega(M_{23} + M \frac{i_1}{i_2 + i_3})i_3 \tag{5}$$

根据变压器的电路规律可知, $L_{eff} = L + M \frac{i_1}{i_2 + i_3}$, $M_{eff} = M_{23} + M \frac{i_1}{i_2 + i_3}$ 。因

为变压器结构对称，所以 $L_{eff2} = L_{eff3} = L_{eff} = L + M \frac{i_1}{i_2 + i_3}$ 。由上述推导可知，变压

器的等效电感和互感均可以由电压比调节。

2.2 调谐范围

外加电压可以通过幅值和相位共同调节变压器，假设 $\alpha + \beta = r(\cos \theta + i \sin \theta)$ ，其中 $\left| \frac{V_2}{V_1} \right| = r$ ， $\theta = \angle(\phi_2 - \phi_1)$ 。带入 L_{eff} 、 M_{eff} 和 V_2 ，得到：

$$L_{eff} = \text{Re}(L_{eff}) + i \text{Im}(L_{eff})$$

$$\text{Re}(L_{eff}) = \frac{(2Mr \cos \theta - r^2 L)(M^2 - L^2) + (2M^2 - MLr \cos \theta)(L - M_{23})}{4M^2 - 4Mr \cos \theta L + (rL)^2} \quad (6)$$

$$\text{Im}(L_{eff}) = \frac{2Mr \sin \theta (M^2 - L^2) + rML(L - M_{23})}{4M^2 - 4Mr \cos \theta L + (rL)^2}$$

$$M_{eff} = \text{Re}(M_{eff}) + i \text{Im}(M_{eff})$$

$$\text{Re}(M_{eff}) = \frac{(2Mr \cos \theta - r^2 L)(M^2 - LM_{23}) - (2M^2 - MLr \cos \theta)(L - M_{23})}{4M^2 - 4Mr \cos \theta L + (rL)^2} \quad (7)$$

$$\text{Im}(M_{eff}) = \frac{2Mr \sin \theta (M^2 - LM_{23}) - rML(L - M_{23})}{4M^2 - 4Mr \cos \theta L + (rL)^2}$$

$$V_2 = j\omega[\text{Re}(L_{eff})i_2 + \text{Re}(M_{eff})i_3] - \omega[\text{Im}(L_{eff})i_2 + \text{Im}(M_{eff})i_3] \quad (8)$$

从上式中可以看出，等效电感虚部的出现导致变压器寄生电阻的变化，在特定状态下，通过电压比可以调节变压器的品质因数 Q。

接下来，逐一分析 $\text{Re}(L_{eff})$ 和 $\text{Im}(L_{eff})$ 的变化趋势， M_{eff} 与 L_{eff} 变化趋势相同，因此不多做赘述。

2.2.1 $\text{Re}(L_{eff})$ 的变化趋势

首先，分别以 r 和 θ 为变量对 $\text{Re}(L_{eff})$ 求导：

$$\text{Re} L_{eff}(r)' = \frac{M(2M^2 - L^2 - LM_{23})(L^2 \cos \theta r^2 - 4MLr + 4M^2 \cos \theta)}{[4M^2 - 4Mr \cos \theta L + (rL)^2]^2} \quad (9)$$

$$\text{Re} L_{eff}(\theta)' = \frac{M \sin \theta (L^2 + LM_{23} - 2M^2)r[4M^2 - (rL)^2]}{[4M^2 - 4Mr \cos \theta L + (rL)^2]^2} \quad (10)$$

以 r 为变量时，导数有两个零点：

$$\begin{aligned} r_1 &= \frac{2M(1 - \sin \theta)}{L \cos \theta} \\ r_2 &= \frac{2M(1 + \sin \theta)}{L \cos \theta} \end{aligned} \quad (11)$$

由于 $r > 0$ ，所以 $\theta > \frac{\pi}{2}$ 时， $\text{Re}(L_{eff})$ 没有极点， $\theta < \frac{\pi}{2}$ 时， $\text{Re}(L_{eff})$ 有两个极点。

$\text{Re} L_{eff}(r)$ 的正负由电感之间的耦合系数决定。在没有极点的情况下， $2k_{12}^2 < 1 + k_{23}$ 时， $\text{Re}(L_{eff})$ 随着 r 的增加单调减小； $2k_{12}^2 > 1 + k_{23}$ 时， $\text{Re}(L_{eff})$ 随着 r 的增加单调增加。在两个极点的情况下， $2k_{12}^2 < 1 + k_{23}$ 时， $\text{Re}(L_{eff})$ 随着 r 的增加先增加后减小再增加； $2k_{12}^2 > 1 + k_{23}$ 时， $\text{Re}(L_{eff})$ 随着 r 的增加先减小后增加再减小。

以 θ 为变量时，由于 $\theta \in (0, \pi)$ ，导数没有零点。 $\text{Re}(L_{eff})$ 在以下两种情况下随 θ 增加而增加，其余情况递减： $2k_{12}^2 < 1 + k_{23}$ 且 $2k_{12} > r$ 时，以及 $2k_{12}^2 > 1 + k_{23}$ 且 $2k_{12} < r$ 时单调递增。

2.2.2 $\text{Im}(L_{eff})$ 的变化趋势

以 r 和 θ 为变量对 $\text{Im}(L_{eff})$ 求导：

$$\text{Im} L_{eff}(r)' = \frac{M[2(M^2 - L^2) \sin \theta + L(L - M_{23})](4M^2 - r^2 L^2)}{[4M^2 - 4Mr \cos \theta L + (rL)^2]^2} \quad (12)$$

$$\text{Im} L_{eff}(\theta)' = \frac{2Mr\{(M^2 - L^2)[(4M^2 + r^2 L^2) \cos \theta - 4MLr] - 2MrL^2(L - M_2)}{[4M^2 - 4Mr \cos \theta L + (rL)^2]^2} \quad (13)$$

在 r 的值域内， $\text{Im}(L_{eff})$ 有一个极点： $r_1 = \frac{2M}{L}$ 。当 $\sin \theta < \frac{k_{23} - 1}{2(k_{12}^2 - 1)}$ 时， $\text{Im}(L_{eff})$

先增加后减小；当 $\sin \theta > \frac{k_{23} - 1}{2(k_{12}^2 - 1)}$ 时，先减小后增加。

以 θ 为变量时， $\text{Im}(L_{eff})$ 有一个极点： $\theta = \arccos\left(\frac{-AC - B\sqrt{A^2 + B^2 - C^2}}{A^2 + B^2}\right)$ ，其

$A = (M^2 - L^2)(4M^2 + r^2L^2)$, $B = -2ML^2r(L - M_{23})$, $C = 4MLr(L^2 - M^2)$, 经过判断 θ 存在且满足值域要求。根据三角函数的性质, $\text{Im}(L_{\text{eff}})$ 先减小后增加。

3 仿真验证

如图 2 所示, 可调变压器由主变压器和辅助电感两部分构成, 主变压器外接被控源 V_2 和 V_3 , 辅助电感外接调制源 V_1 。

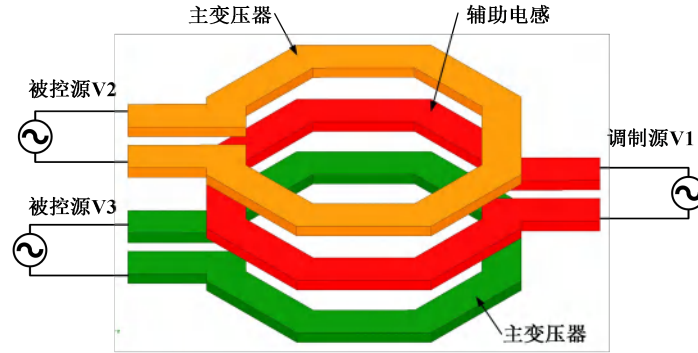


图 2 可调变压器结构图

改变电压幅值比, 变压器感值的变化如图 3 所示。

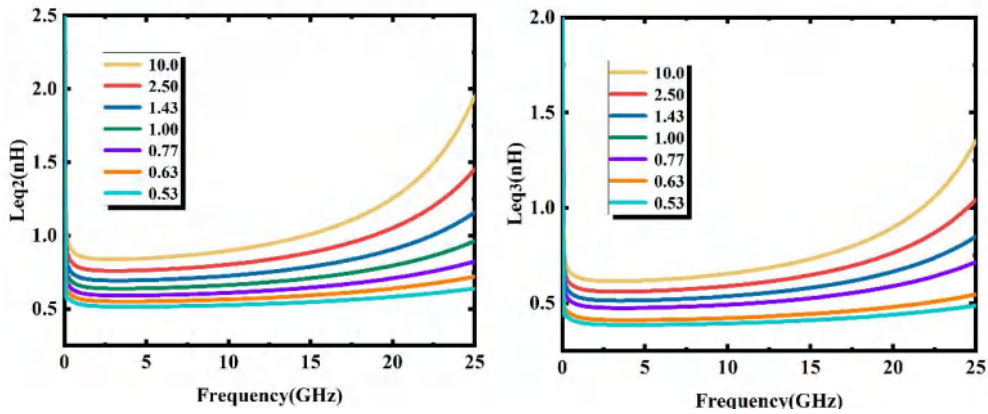


图 3 可调变压器电感值

从图 3 可知, 在 DC~25GHz 范围内, 变压器的感值可以由外加电压的幅值比调节, 调谐范围分别为 75.0%和 71.4%, 验证了磁耦合调节的可行性。

参考文献

[1] M. J. Zavarei, K. Kim, and H.-J. Song, "A 26–40 GHz wideband power amplifier with transformer-based high-order matching networks in 28-nm CMOS FD-SOI," IEEE

- Microwave and Wireless Components Letters, vol. 32, no. 9, pp. 1079–1082, 2022.
- [2] H. Jia, C. C. Prawoto, B. Chi, Z. Wang, and C. P. Yue, “A full K-band power amplifier with 32.9% PAE and 15.3-dBm power in 65-nm CMOS,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 9, pp. 2657–2668, 2018
- [3] B. Bae, E. Kim, S. Kim, and J. Han, “Dual-band CMOS low-noise amplifier employing transformer-based band-switchable load for 5G NR FR2 applications,” *IEEE Microwave and Wireless Technology Letters*, vol. 33, no. 3, pp. 319–322, 2022
- [4] J. Ko and S. Nam, “A Two-Stage S/X -Band CMOS Power Amplifier for High-Resolution Radar Transceivers,” *IEEE Microwave and Wireless Components Letters*, vol. 28, no. 7, pp. 606–608, 2018.
- [5] B. Liu, R. Zhou, and Z. Zhu, “Reconfigurable 2.4/5.0-GHz dual-band CMOS power amplifier for WLAN 802.11 ax,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024.
- [6] S. Oh, J. Kim, and J. Oh, “Sub-THz switch-less reconfigurable triple-/push-push dual-band VCO for 6G communication,” in *2022 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*. IEEE, 2022, pp. 219–222.
- [7] N.-Z. Sun, L. Gao, H.-Y. Li, J.-X. Xu, and X. Zhang, “A compact millimeter-wave reconfigurable dual-band LNA with image-rejection in 28-nm bulk CMOS for 5G applications,” *IEEE Journal of Solid-State Circuits*, 2024
- [8] M. A. Mokri, S. Miraslani, M. A. Hoque, and D. Heo, “A dualpath transformer-based multiband power amplifier for mm-wave 5G applications,” *IEEE Journal of Solid-State Circuits*, 2024.
- [9] Y. Zheng, G. Shan, X. Fan, Z. Yang, and G. Li, “An 0.04–10.32 nH ultracompact tunable inductor based on flexible magnetic coupling tuning technique,” *IEEE Transactions on Electron Devices*, 2024.

基于不确定性量化的微带滤波器逆向设计

胡家宁, 邓皓千, 徐魁文

杭州电子科技大学绍兴集成电路研究院

摘要: 尽管基于学习的代理模型可实现快速预测, 但数据与预测中的模型不确定性仍是重大挑战。本文提出一种基于学习的不确定性分析方法, 该方法利用集成对抗学习分析逆模型进行不确定性分析, 得到逆模型输出的置信区间。随后采用差分进化算法在置信区间内优化逆模型输出的几何参数。为验证该方法有效性, 对具有多个频带和带宽的发夹型滤波器进行了仿真、实测。相较于优化前的电路响应, 优化后响应与输入标签的拟合度上取得了显著提升。

关键词: 神经网络; 射频电路逆建模; 不确定性分析

1 引言

近年来, 深度学习在微波电路设计领域获得了显著关注, 为正建模与逆建模提供了变革性的解决方案 [1]。神经网络 (ANN) 擅长学习几何参数与电磁 (EM) 响应之间的非线性关系, 从而能够实现高效的代理建模。在正建模中, 几何参数作为输入用于预测电磁响应 (例如, S 参数、电场分布或功率分布)。传统的全波仿真计算密集且耗时, 而充分训练的正模型可以取代全波仿真器, 实现快速的设计迭代。反之, 逆建模则利用神经网络直接生成满足预定性能指标的几何参数。该方法为传统的全波优化提供了一种快速且系统化的替代方案。然而, 逆建模面临着诸多挑战, 例如解的非唯一性 (即多种几何构型可产生相同的电磁响应) 以及高维输入空间问题。

为应对这些局限, 本文提出一种用于微波滤波器逆设计的建模优化方法。该框架包含两个核心组件: 建模方法神经网络 (ANN) 和标签生成算法合成少数类过采样技术 (MLSMOTE) [2]。尽管神经网络能加速设计流程, 但其固有的不确定性 (源于有限的数据和模型近似) 亟需严谨量化。为此, 我们集成了集成对抗学习 (Ensemble Adversarial Learning) 来量化滤波器逆代理模型输出的不确定性, 为预测的几何参数提供置信区间。

2 方法

2.1 提出的方法概述

本文提出的方法由逆向建模、不确定性分析和优化 3 个步骤组成。在逆向模型中, 通过输入 s 参数, 逆模型可以很好地预测几何空间。利用逆模型, 将不确

定性分析方法嵌入到集成对抗样本中，生成一组多样化的专家子模型。然后合并这些模型以定量地量化模型的不确定性。然后，利用差分进化优化器对逆模型的输出进行优化，从而提高模型的整体精度。

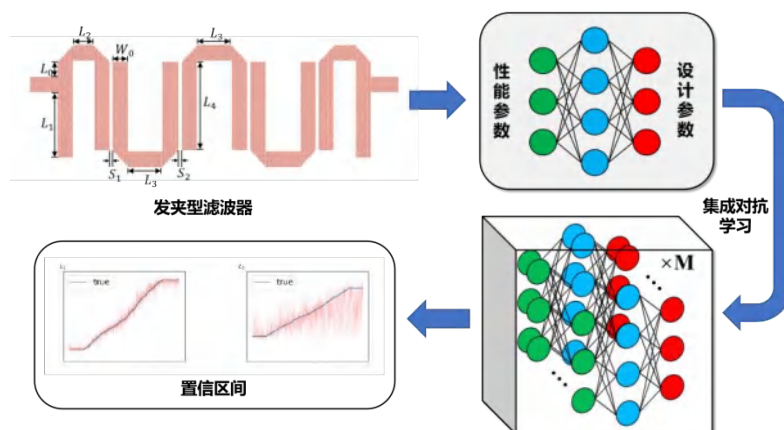


图 1 不确定性分析流程

2.2 基于人工神经网络的逆设计

利用全连接神经网络可以拟合高度非线性关系的能力，建立电路几何参数与散射参数之间的逆设计模型映射。经过训练的神经网络模型可以在毫秒内给出响应，从而实现快速设计。然而，由于精确的网络需要大量训练数据，随着电路结构复杂度提高，仿真时间延长，数据集往往无法覆盖整个需求的几何参数分布，导致直接映射方法在模型精度和可靠性方面存在一定的缺陷。因此，本文将利用不确定性量化分析来提高模型的准确性和稳定性。

2.3 基于学习的不确定性量化分析

在获得相对精确的神经网络模型后，利用集成对抗学习[3]对模型的不确定性进行量化，从而收缩几何空间范围，加速优化器迭代收敛。该方法在不改变原有网络结构的情况下易于实现部署。其基本思想是在原始模型的基础上，利用对抗学习生成多个相似的专家子模型，并将所有子模型的均值和方差分别作为输出的期望值和不确定性。

表 1 优化前后s参数对比

Algorithm 1 Algorithm Of Ensembles Adversarial Learning

- 1: Use a proper scoring rule as the training criterion $\nabla \ell(\theta, X, Y)$ for all neural network. Default values are $M = 20$ and $\alpha = 0.01$.
 - 2: Initialize $0, 1, \dots, M$ randomly.
 - 3: train first model 0 by randomly sampling data point.
 - 4: **for** each $i \in [1 : M]$ **do**
-

5: generate adversarial example using

$$X_{i+1} = X_i + \alpha \nabla_{X_i} \ell(\theta, X_i, Y)$$

6: Minimize $\nabla(\theta_{i+1}, X_{i+1}, Y)$

7: end for

3 实验验证

在集成对抗学习后，逆模型可以分析预测结果的不确定性，并生成预测几何参数的置信区间。为了进一步优化逆模型的输出结果，将置信区间作为差分进化优化器的搜索空间。图 2 左为使用差分进化优化后的电路响应。从标签拟合的角度来看，优化后的电路响应与标签响应的相似度比优化前提高了 37.22%。同时，在整个样本空间上进行相同设置的前向电路优化作为对比。图 2 右为无逆模型直接优化方案，电路性能明显差于不确定性分析后逆模型产生的电路性能。上述优化器被设置为 10 次迭代、10 个种群和 100 次对电路模拟器的调用。

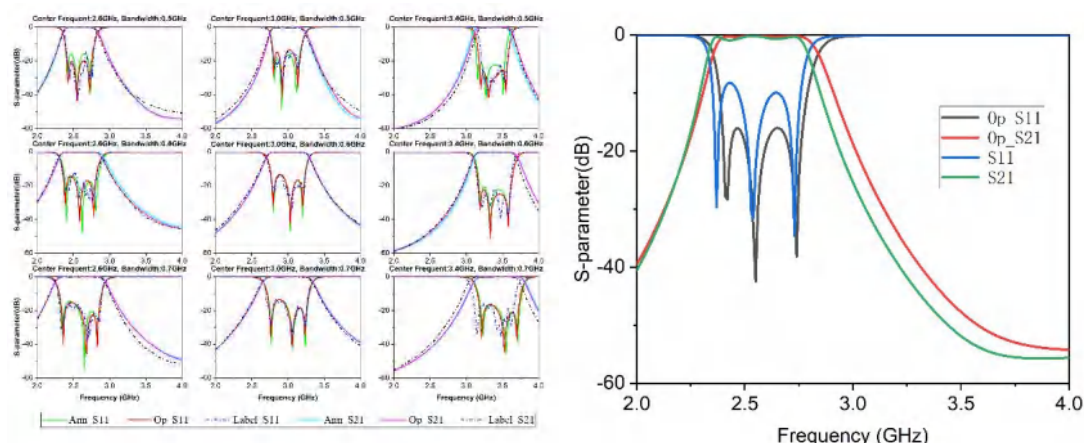


图 2 优化前后s参数对比

为了使仿真结果与实测结果较好地吻合，采用空间映射技术将 2.5 维仿真器 ADS 生成的样本映射到全波仿真器 CST 中。在得到全波仿真结果后，为了进一步验证方法的有效性，制作了所提滤波器的原型，并利用矢量网络分析仪进行了测试。测量结果如图 3 所示。

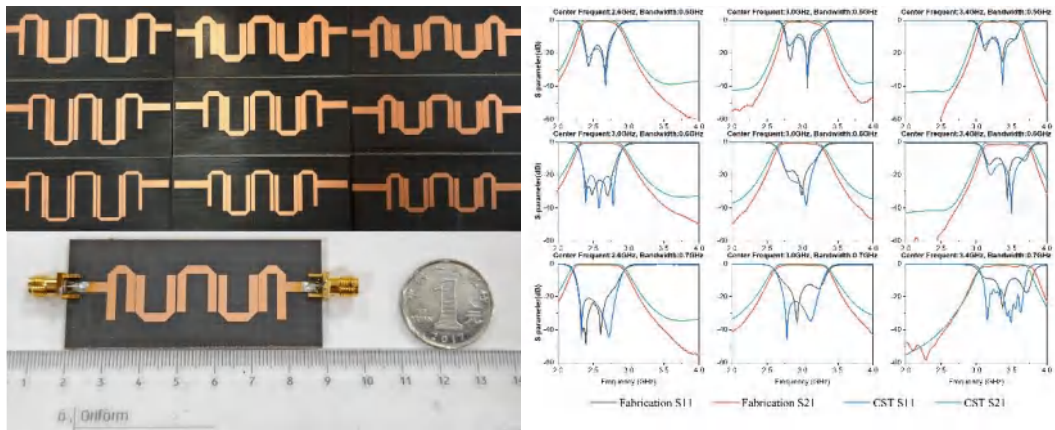


图 3 发夹型滤波器加工实测

4 结论

本文提出了一种基于不确定性分析的逆模型优化方法。该方法基于集成对抗学习和人工神经网络，采用集成对抗学习对神经网络模型的不确定性进行分析。该方法训练的模型输出与标签数据具有更好的拟合程度。

参考文献

- [1]. F. Feng, W. Na, J. Jin, J. Zhang, W. Zhang, and Q.-J. Zhang, "Artificial neural networks for microwave computer-aided design: The state of the art," *IEEE Transactions on Microwave Theory and Techniques*, 2022.
- [2]. F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385-397, 2015.
- [3]. BB. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

铟磷异质结双极晶体管大信号建模及参数提取方法研究

罗坤¹, 冯长乐¹, 张傲^{2,3*}

(1. 南通大学, 信息科学技术学院, 江苏, 南通, 226019;

2. 南通大学, 微电子学院, 江苏, 南通, 226019;

3. 新加坡南洋理工大学 电气与电子工程学院, 新加坡 639798)

摘要: 本文对铟磷异质结双极晶体管 (InP HBT) 大信号模型进行了研究, 使用基于 VBIC 模型的符号定义器件 (SDD) 对 HBT 进行了建模, 并通过 HBT 器件的直流 I-V 曲线和 S 参数测量数据对模型的直流和交流参数进行了精确提取。使用发射极面积为 $0.6 \times 10 \mu\text{m}^2$ 的 InP HBT 器件分别在 25°C 、 80°C 和 120°C 的直流特性以及在 200MHz-110GHz 频率范围内的交流特性对模型进行了验证。研究结果表明, 模拟数据与测量数据吻合良好, 验证了模型和参数提取方法的准确性。

关键词: 铟磷异质结双极晶体管; VBIC大信号模型; 参数提取

Study on Large-Signal Modeling and Parameter Extraction Methods for InP Heterojunction Bipolar Transistors

LUO Kun¹, FENG ChangLe¹, ZHANG Ao^{2,3*}

(1.School of Information Science and Technology, Nantong University, Nantong 226019, China; 2. School of Microelectronics, Nantong University, Nantong 226019, China; 3. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore)

Abstract: This paper investigates the large-signal modeling of InP heterojunction bipolar transistors (InP HBTs), using the Symbolic Definition Device (SDD) based on the VBIC model for device modeling. The DC and AC parameters of the model are accurately extracted through DC I-V characteristics and S-parameter measurement data of HBT devices. The model is validated by DC characteristics of InP HBT devices with an emitter area of $0.6 \times 10 \mu\text{m}^2$ at 25°C , 80°C , and 120°C , as well as AC characteristics in the frequency range of 200 MHz to 110 GHz. The results show good agreement between simulated and measured data, verifying the accuracy of the proposed model and parameter extraction method.

Keywords: InP HBT; VBIC Large-Signal Model; Parameter Extraction

1 引言

近年来，化合物半导体器件在高速、高频领域的应用不断拓展，InP 基器件凭借优异性能脱颖而出。InP 异质结双极晶体管（HBT）作为其中的重要分支，因其独特的材料优势备受关注。III-V 族化合物材料 InP 相比传统半导体，具备更高的电子迁移率与更宽的禁带宽度，展现出卓越的击穿特性。与同属 III-V 族的 GaAs 材料相比，InP 在电子饱和速度、电流密度以及功率密度等关键性能指标上更具优势，这使得 InP HBT 器件在通信、雷达等高频功率放大领域展现出巨大的应用潜力，逐步成为推动相关技术发展的核心器件之一^[1-6]。

本文研究了基于 VBIC 大信号模型的 SDD 模型及其参数提取方法，利用 $0.6 \times 10 \mu\text{m}^2$ 的 InP HBT 器件对模型的直流特性和交流特性进行了验证，实验结果表明在不同环境温度下模型能准确表征器件的直流特性曲线，在 200MHz-110GHz 频率范围内模拟和测试 S 参数吻合很好，验证了模型及参数提取方法的准确性。

2 VBIC 大信号等效电路模型

图 1 给出了 VBIC 大信号等效电路模型，由两个互补型的寄生晶体管和本征晶体管构成，有四个端口（基极、集电极、发射极、衬底），VBIC 模型采用一阶 RC 网络热模型。相比于 SGP 模型，VBIC 模型不仅通过修改的表达式使得对厄利效应更加精确的建模，还考虑了准饱和效应、雪崩效应、自热效应等。

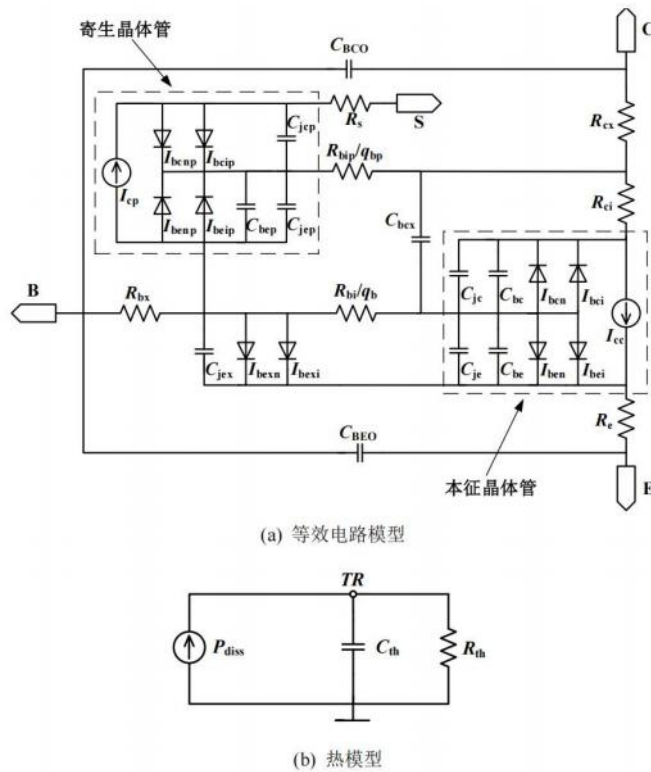


图 1 VBIC 大信号等效电路模型

由于 npn 型 HBT 器件不存在基底寄生 pnp 晶体管，相关基底寄生效应影响微弱等因素，在精度影响较小的情况下对模型进行了简化，图 2 给出了简化的 VBIC 大信号模型，与标准的 VBIC 模型相比，具有模型参数少、SDD 模型建立简单等优势。

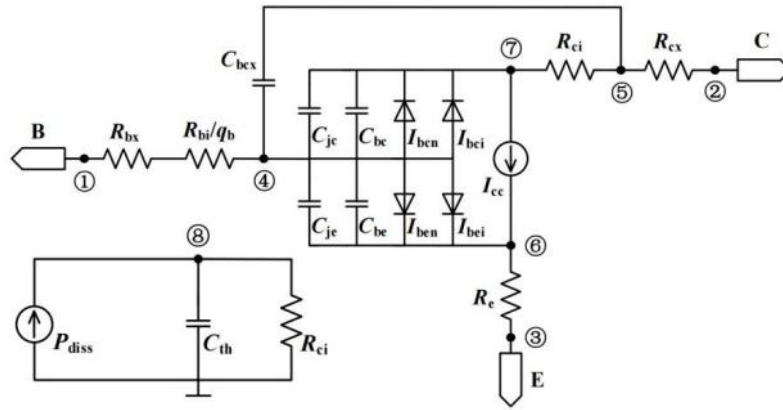


图 2 简化的 VBIC 大信号模型

模型参数提取主要包括直流参数提取和交流参数提取两部分；直流参数提取时使用器件测量的正向和反向 Gummel-Plot 数据直接提取，然后根据直流 I-V 曲线进行整体优化；交流参数提取时采用小信号等效电路模型提取结电容，对结电容进行非线性拟合来提取交流参数，参数提取结束后对模型进行 S 参数仿真，再对模型参数进行整体优化。

3 结果分析

图 3-5 给出了分别环境温度在 25°C、80°C 和 120°C 下，发射极面积为 $0.6 \times 10 \mu\text{m}^2$ 的 InP HBT 器件测量和模拟 I_c-V_{ce} ， $V_{be}-V_{ce}$ 输出特性曲线对比，模拟和测试数据吻合很好。

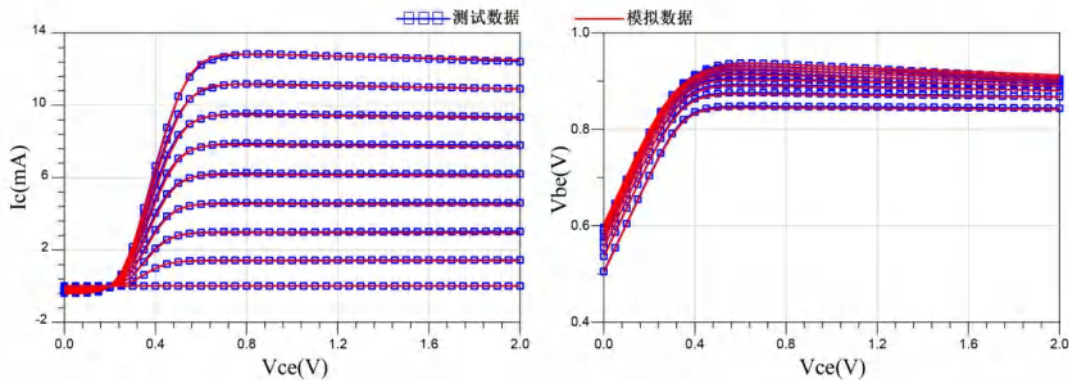


图 3 25°C 时 $6 \mu\text{m}^2$ HBT 器件 I_c-V_{ce} ， $V_{be}-V_{ce}$ 输出特性曲线对比

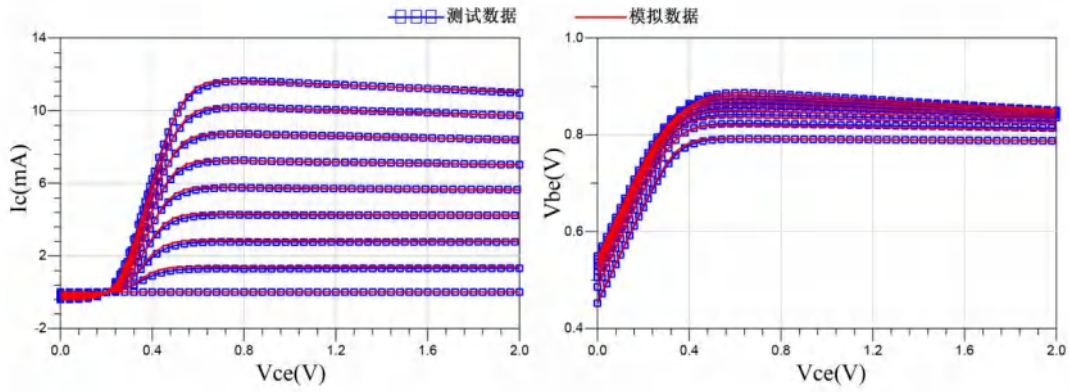


图 4 80°C时 6 μm^2 HBT 器件 I_c - V_{ce} , V_{be} - V_{ce} 输出特性曲线对比

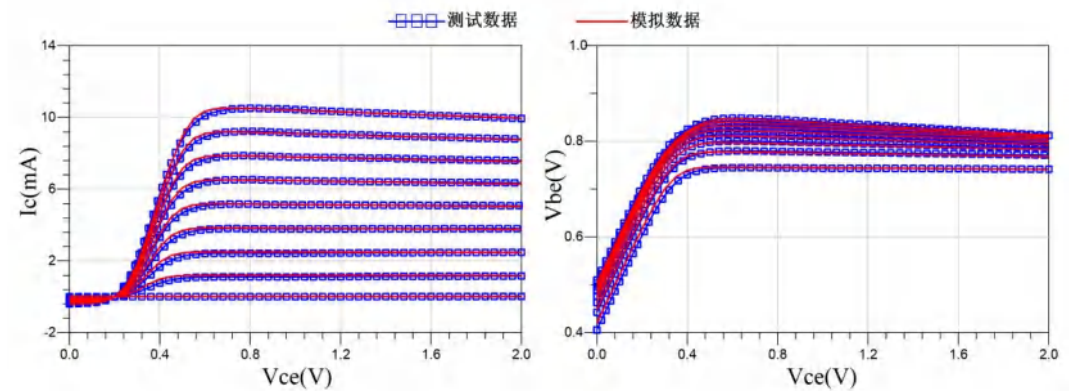


图 5 120°C时 6 μm^2 HBT 器件 I_c - V_{ce} , V_{be} - V_{ce} 输出特性曲线对比

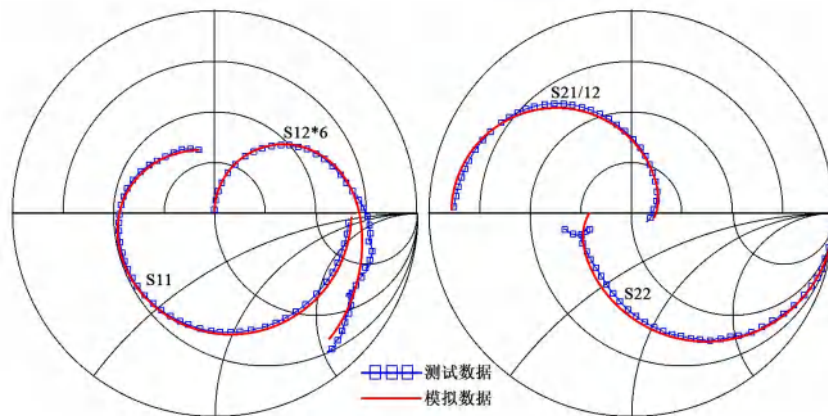


图 6 模拟和测试S 参数对比 (偏置: $V_{ce}=1\text{V}$, $I_{be}=250\mu\text{A}$)

图 6 给出了环境温度在 25°C 下 InP HBT 器件 S 参数的模拟结果和测试结果的对比曲线, 可以看到, 在 $V_{ce}=1\text{V}$, $I_{be}=250\mu\text{A}$ 偏置下, 200MHz-110GHz 频率范围内, 模拟与测试结果吻合很好, 模型及参数提取方法的准确性得以验证。

4 结论

本文对 InP HBT 器件大信号模型及其参数提取方法进行了研究。针对 InP HBT 器件, 对 VBIC 模型进行了简化, 并使用 SDD 器件构建模型, 使用器件 I-V 测量数据和 S 参数测量数据精确提取了模型的直流参数和交流参数。利用 $0.6 \times 10 \mu\text{m}^2$ 的 InP HBT 器件对模型的直流特性和交流特性进行了验证, 实验结果表明在不同环境温度下模型能准确表征器件的直流特性曲线, 在 200MHz-110GHz 频率范围内模拟和测试 S 参数吻合很好, 验证了模型及参数提取方法的准确性。

参考文献

- [1] 张傲, 高建军. 《III-V族异质结双极晶体管: 从器件工作原理到电路模型》[M]. 高等教育出版社, 2023.
- [2] Zhang A, Gao J. An improved nonlinear model for millimeter-wave InP HBT including DC/AC dispersion effects[J]. IEEE Microwave and Wireless Components Letters, 2021, 31(5): 465-468.
- [3] Mukherjee C, Raya C, Ardouin B, et al. Scalable compact modeling of III-V DHBTs: Prospective figures of merit toward terahertz operation[J]. IEEE Transactions on Electron Devices, 2018, 65(12): 5357-5364.
- [4] Urteaga M, Griffith Z, Seo M, et al. InP HBT technologies for THz integrated circuits[J]. Proceedings of the IEEE, 2017, 105(6): 1051-1067.
- [5] Rode J C, Chiang H W, Choudhary P, et al. Indium phosphide heterobipolar transistor technology beyond 1-THz bandwidth[J]. IEEE Transactions on Electron Devices, 2015, 62(9): 2779-2785.
- [6] Davy N, Nodjiadjim V, Riet M, et al. InP DHBT analytical modeling: Toward THz transistors[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 42(11): 4102-4111.

110GHz 铟磷异质结双极晶体管小信号建模及参数提取研究

冯长乐¹, 罗坤¹, 张傲^{2,3*}

(1. 南通大学, 信息科学技术学院, 江苏, 南通, 226019;

2. 南通大学, 微电子学院, 江苏, 南通, 226019;

3. 新加坡南洋理工大学 电气与电子工程学院, 新加坡 639798)

摘要: 异质结双极晶体管(HBT)因其优越的高频性能在微波与射频集成电路中具有重要应用, 而精确的小信号等效电路模型是提升电路设计效率的关键。本文首先构建了包含本征和非本征部分的小信号等效电路模型。在此基础上, 利用特殊结构方法对 PAD 电容和寄生电感进行提取, 运用截止法对寄生电阻进行提取。随后, 研究了去嵌方法和提取本征元件的直接提取方法, 并在三组偏置条件下对 $0.6 \times 10 \mu\text{m}^2$ 的 InP HBT 器件在 200MHz-110GHz 的频率范围内进行了验证。研究表明, 测试数据与模拟数据吻合良好, 验证了模型和提取方法的准确性。

关键词: 铟磷异质结双极晶体管; 小信号等效电路模型; 参数提取

Study on Small-Signal Modeling and Parameter Extraction for InP Heterojunction Bipolar Transistor up to 110 GHz

FENG Changle¹, LUO Kun¹, ZHANG Ao^{2,3*}

(1.School of Information Science and Technology, Nantong University, Nantong 226019, China; 2. School of Microelectronics, Nantong University, Nantong 226019, China; 3. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore)

Abstract: Heterojunction bipolar transistors (HBTs) have significant applications in microwave and radio frequency integrated circuits due to their superior high-frequency performance. An accurate small-signal equivalent circuit model is crucial for enhancing circuit design efficiency. In this paper, a small-signal equivalent circuit model including both intrinsic and extrinsic parts is first constructed. Based on this, the PAD capacitance and parasitic inductance are extracted using a special structure method, and the parasitic resistance is extracted using the cutoff method. Subsequently, the de-embedding method and the direct extraction method for intrinsic elements are studied. The $0.6 \times 10 \mu\text{m}^2$ InP HBT device is verified under three sets of bias conditions in the frequency range of 200 MHz to 110 GHz. The research results show that the test data and simulation data are in good agreement, verifying the accuracy of the model and extraction method.

Keywords: InP HBT; small-signal equivalent circuit model; parameter extraction

1 引言

III-V 族半导体具有深厚的基础物理学，例如高载流子迁移率和高电子漂移速度，这使它们在高速、高频电子设备中具有巨大潜力。InP HBT 相较于 GaAs HBT 有着更高的工作频率和更好的散热性能。小信号模型是构建非线性模型和噪声模型的基础，其与器件的物理结构直接相关。建立精确的 HBT 小信号等效电路模型对表征器件性能和提升单片集成电路设计的成功率至关重要^[1-7]。

本文研究了 InP HBT 的小信号等效电路模型和参数提取方法。利用特殊测试结构方法提取 PAD 电容和寄生电感，利用反向截止法提取寄生电阻。研究了去嵌方法，并进一步提取了本征元件。研究表明， $V_{CE}=1.4V$ ， I_b 分别为 $200\mu A$ 、 $300\mu A$ 和 $400\mu A$ 时， $0.6\times 10\mu m^2$ 的 InP HBT 在 200MHz-110GHz 频率范围内模拟和测试 S 参数吻合效果很好。

2 小信号等效电路模型

图 1 给出了 HBT 器件的 π 型小信号等效电路模型，外部围绕着并联寄生部分和串联寄生部分。其中， C_{pb} 、 C_{pc} 和 C_{pbc} 分别表示基极-发射极、集电极-发射极和基极-集电极 PAD 电容， L_b 、 L_c 和 L_e 分别表示基极、集电极和发射极寄生电感， R_{bx} 、 R_c 和 R_e 分别表示基极、集电极和发射极寄生电阻， C_{ex} 表示非本征基极-集电极电容， R_{bi} 表示本征基极电阻， R_{π} 表示本征基极-发射极电阻， C_{π} 表示本征基极-发射极电容， C_{bc} 表示本征基极-集电极电容， g_m 为跨导。

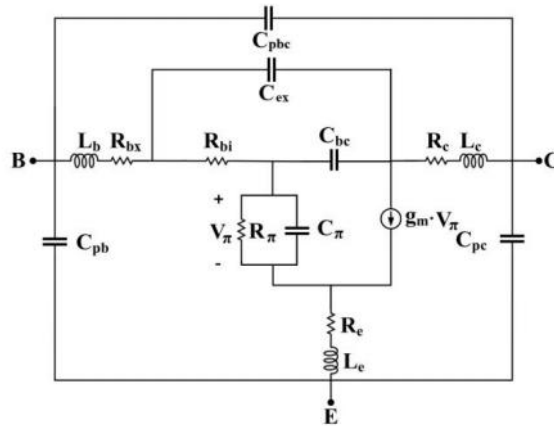


图 1 HBT 小信号等效电路模型

3 参数提取

3.1 寄生元件提取

图 2-4 分别为提取 PAD 电容、寄生电感的特殊测试结构等效电路模型和提取寄生电阻的截止状态下小信号等效电路模型。依据这三种情况下的测试数据和等效电路模型的网络参数可以提取出对应的寄生元件。

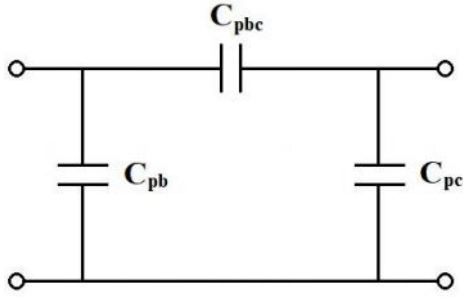


图 2 开路测试结构等效电路模型

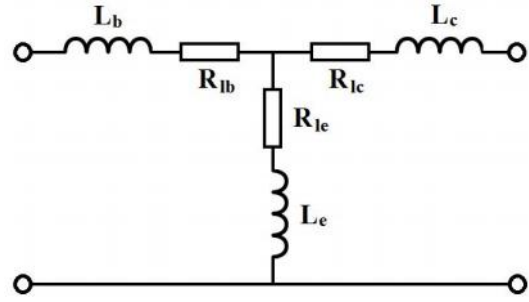


图 3 短路测试结构等效电路模型

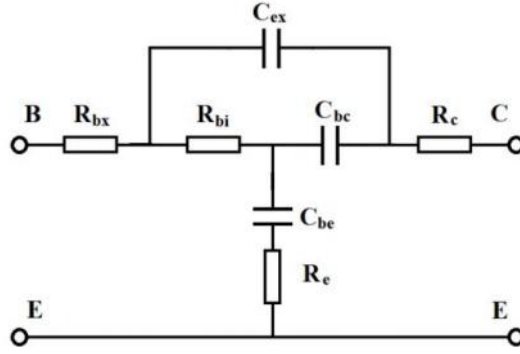


图 4 截止状态下小信号等效电路模型

3.2 本征元件提取

如式(1)所示， Y_{MEAS} 是测试 Y 参数， Y_{PAD} 是 PAD 电容的 Y 参数， Z_{RL} 是寄生电感和寄生电阻的 Z 参数。在提取完寄生元件之后，利用负元件去嵌技术对 PAD 电容、寄生电感和寄生电阻进行去嵌^[4]，得到本征 Y 参数。进一步得到各本征元件的表达式，直接提取得到本征元件。

$$Y_{INT} = \left[Y_{MEAS} - Y_{PAD} \right]^{-1} - Z_{RL} \quad (1)$$

提取本征元件的公式如式(2)-(7)所示：

$$C_{bc} + C_{ex} = \frac{1}{\omega} \text{Im} \left(\frac{Y_{11}Y_{22} + Y_{12}Y_{21}}{Y_{11} + Y_{21}} \right) \quad (2)$$

$$R_{bi}C_{bc} = \frac{1}{(C_{bc} + C_{ex}) \text{Re} \left(\frac{Y_{11}Y_{22} - Y_{12}Y_{21}}{Y_{22} + Y_{12}} \right)} \quad (3)$$

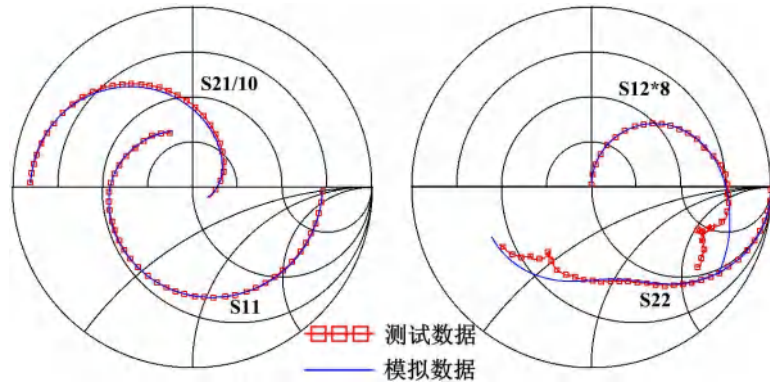
$$R_{bi} = \frac{1}{\text{Re} \left(\frac{1}{Y_{11}} \right)} \quad (4)$$

$$R_{\pi} = \text{Re} \left[\frac{1 + j\omega R_{bi}C_{bc}}{1/(Y_{11} + Y_{12}) - R_{bi}} \right] \quad (5)$$

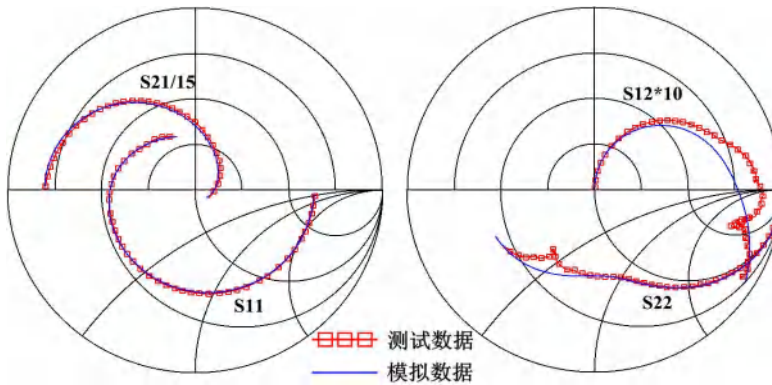
$$C_{\pi} = \frac{1}{\omega} \text{Im} \left[\frac{1 + j\omega R_{bi}C_{bc}}{1/(Y_{11} + Y_{12}) - R_{bi}} \right] \quad (6)$$

$$g_m = (Y_{21} - Y_{12}) \{1 + R_{bi} [1/R_{\pi} + j\omega(C_{bc} + C_{\pi})]\} \quad (7)$$

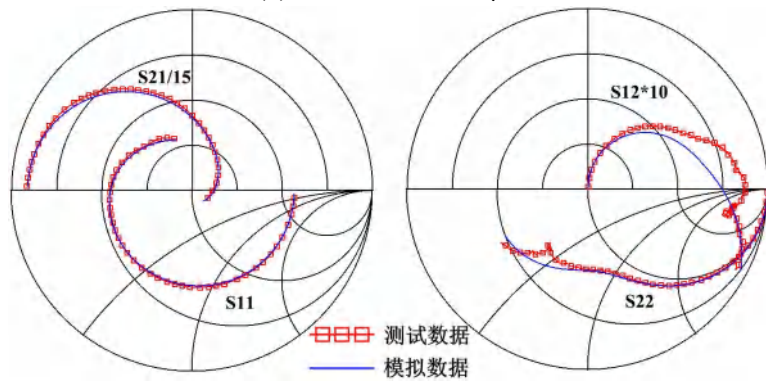
在偏置条件为 $V_{CE}=1.4V$, I_b 分别为 $200\mu A$ 、 $300\mu A$ 和 $400\mu A$ 时, 对发射极长度为 $10\mu m$, 宽度为 $0.6\mu m$ 的 InP HBT 在 $200MHz-110GHz$ 频率范围内进行参数提取, 将提取得到的元件参数代入小信号等效电路模型中, 得到测试和模拟 S 参数的对比曲线图, 如图 5 所示:



(a) $V_{CE}=1.4V, I_b=200\mu A$



(b) $V_{CE}=1.4V, I_b=300\mu A$



(c) $V_{CE}=1.4V, I_b=400\mu A$

图 5 三种不同偏置下 HBT 的 S 参数对比曲线

在 $V_{CE}=1.4V$, I_b 分别为 $200\mu A$ 、 $300\mu A$ 和 $400\mu A$ 时, 图 5 给出了发射极面积为 $0.6 \times 10 \mu m^2$ 的 InP HBT 在 $200MHz-110GHz$ 频率范围内测试结果和模拟结果的对比曲线, 测试结果和模拟结果吻合很好, 验证了模型和提取方法的准确性。

4 结论

本文对 110GHz InP HBT 小信号建模及参数提取方法进行了研究。依据开路测试结构、短路测试结构和反向截止方法对 PAD 电容、寄生电感和寄生电阻进行提取,然后利用负元件去嵌方法去除寄生元件的影响,接着研究了本征元件的直接提取方法。在 $V_{CE}=1.4V$, I_b 分别为 $200\mu A$ 、 $300\mu A$ 和 $400\mu A$ 的偏置下,发射极面积为 $0.6\times 10\mu m^2$ 的 InP HBT 在 200MHz-110GHz 频率范围内测试结果和模拟结果吻合很好,验证了建立的 InP HBT 小信号等效电路模型及参数提取方法的准确性。

参考文献

- [1] Chen Y, Liu J, Zeng M, et al. Universal growth of ultra-thin III–V semiconductor single crystals[J]. Nature Communications, 2020, 11(1): 3979.
- [2] Urteaga M, Griffith Z, Arias-Purdue A, et al. InP HBT technologies for sub-THz communications[C]//2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022: 122-123.
- [3] 高建军著. 异质结双极晶体管: 射频微波建模和参数提取方法[M]. 北京: 高等教育出版社, 2013.
- [4] 张傲, 高建军著. III-V族异质结双极晶体管: 从器件工作原理到电路模型[M]. 北京: 高等教育出版社, 2023.
- [5] Zhang A, Gao J. An improved small signal model of InP HBT for millimeter-wave applications[J]. Microwave and Optical Technology Letters, 2021, 63(8): 2160-2164.
- [6] Davy N, Nodjiadjim V, Riet M, et al. InP DHBT analytical modeling: Toward THz transistors[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 42(11): 4102-4111.
- [7] Mahadev Arabhavi A K. Terahertz InP/GaAsSb Double Heterojunction Bipolar Transistors[D]. ETH Zurich, 2022.

基于螺吡喃光致变色的可重构光学物理不可克隆函数 仿真研究

方浩年^{1, 2, 3}, 惠瑞杰^{1, 2, 3}, 陈飞良^{1, 2, 3}, 李沫^{1, 2, 3*}, 张健^{1, 2, 3}

1. 电子科技大学电子科学与工程学院, 成都 611731

2. 电子科技大学先进毫米波技术集成攻关研究院, 成都 611731

3. 太赫兹辐射与应用四川省重点实验室, 成都 611731

limo@uestc.edu.cn

摘要: 本文提出一种基于螺吡喃光致变色效应的动态可重构光学物理不可克隆函数。通过将螺吡喃分子掺杂于含随机分散二氧化钛纳米粒子的聚甲基丙烯酸甲酯基体中构建复合无序介质, 利用螺吡喃-部花菁异构化反应中折射率与吸光系数的协同变化并相互耦合实现散斑动态重构。三维时域有限差分仿真结果表明: 螺吡喃至部花菁转化可动态重构散射场, 汉明随吸光系数曲线呈现剧增-饱和特性, 这种强非线性响应特征从物理层面增强了 PUF 防伪过程的可重复性并具有光控快速可重构特性。

关键词: 光学物理不可克隆函数; 可重构; 螺吡喃; 光致变色

1 前言

全球假冒商品(如电子设备^[1]及奢侈品)每年造成约 1.7 万亿美元经济损失, 且严重威胁人类的生命健康和信息安全, 凸显防伪技术升级的紧迫性^[2]。传统防伪标签(如荧光图案、微指纹)依赖可复制的物理特征, 易被专业伪造技术破解。

物理不可克隆函数(Physical Unclonable Function, PUF)通过制造过程的固有随机性生成唯一“物理指纹”, 为高安全性防伪提供新路径^[3]。但现有技术仍面临核心瓶颈: 电学 PUF(如仲裁器 PUF)易受建模攻击, 攻击者可通过部分质询-响应对(CRP)预测全部响应, 丧失数学不可克隆性^[4]; 基于 3D 无序结构的传统激光散射光学 PUF 虽具备物理与数学双重不可克隆性, 但其无序结构固化导致 CRP 容量受限, 导致编码容量相对较低且假阳性阈值较大^[5]。可重构光学 PUF(R-PUF)通过动态调整内部材料光学参数(如折射率^[6])实现 CRP 动态刷新, 有望突破传统编码容量限制并实现动态防伪功能。但现有方案仍受限于材料循环寿命差、响应速度慢、需电^[7]/热^[8]等物理场调控等问题, 难以满足可重复、快速高效等应用需求。因此, 开发兼具可重复性、快速响应的低成本 R-PUF 对高安全动态防伪具有重要意义。

螺吡喃(Spiropyran, SP)作为经典有机光致变色分子, 其独特的 C-O 键可在紫外光照射下断裂, 由闭环 SP 态异构化为开环部花菁(MC)显色态, 并在可见光/热刺激下可逆恢复^[9]; 该异构化过程伴随显著的光学特性动态切换(吸收

系数/折射率变化)，且具备秒级响应速度与超千次循环寿命^[10]，可用于实现 R-PUF。基于上述原理，本研究提出了一种利用螺吡喃光致变色效应的可重构光学物理不可克隆函数（SPR-PUF），并通过仿真对其特性进行研究。具体而言，将 SP 分子掺杂于分散有二氧化钛(TiO₂)纳米粒子的聚甲基丙烯酸甲酯(PMMA)基体中，构建复合无序光学介质。螺吡喃-部花菁异构化反应通过调控介质折射率与不同波长处吸光系数，实现 SPR-PUF 的可重构。本文通过仿真研究了特定波段处（532 nm）激光下螺吡喃光致变色效应对 SPR-PUF 散射场分布的动态调控机制，证明了 SPR-PUF 具有快速重构响应能力与防伪过程的高可重复性。

2 模型建立和结果分析

2.1 SPR-PUF 的结构设计与仿真模型

采用三维时域有限差分法（FDTD）仿真螺吡喃光致变色导致的的散射场动态重构过程，通过改变材料光学参数模拟紫外光调控 SP 可逆化异构过程。仿真模型如图 1(a)所示。其中包括：掺杂 SP 分子的 $10 \times 10 \times 4 \mu\text{m}^3$ 的 PMMA 层（介质层）；10% 体积分数粒径 200 nm 的 TiO₂ 纳米粒子随机分布（散射体）；以 532 nm 平面波垂直入射介质层上表面，并在介质层下表面 $2 \mu\text{m}$ 处设置 $10 \times 10 \mu\text{m}^2$ CCD 监视器，收集散射场强分布信息（光源与监视器）。

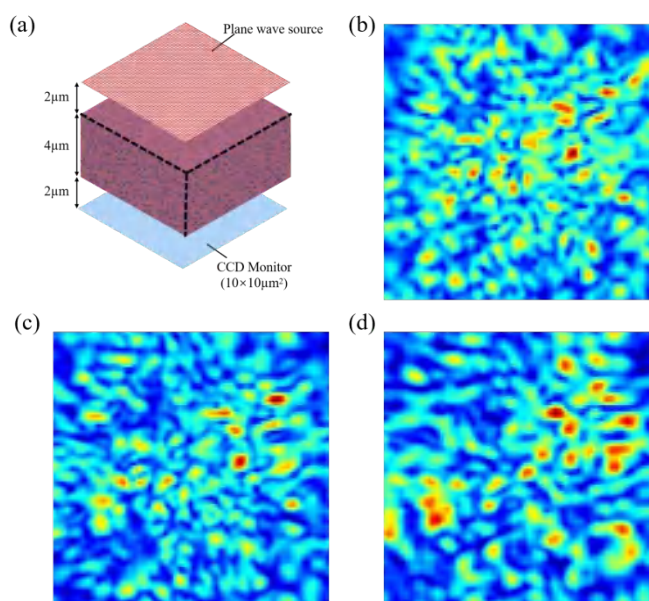


图 1 SPR-PUF 仿真模型图及光学散射场分布。(a) SPR-PUF 仿真模型图；(b)未加入 SP 分子的 PUF 散射场分布；(c)体积分数 0.1%的 SPR-PUF 散射场分布；(d)转化为 MC 后体积分数 0.1%的 SPR-PUF 散射场分布。

2.2 仿真结果与分析

SPR-PUF 动态可重构的核心机制在于：紫外光触发无色或浅色的 SP 分子 C-O 键断裂开环，形成部花菁（MC）异构体，导致介质折射率与不同波长处吸

光系数协同变化。为量化该过程对散射场的影响，采用三维时域有限差分（FDTD）方法，模拟特定波长（532 nm）平面波激光入射下 SP 态向 MC 态转化过程中的散射场强度分布变化，揭示光控可重构的散斑动态变化规律。

图 1(b)-(d)分别为无 SP、SP 和 MC 三种不同状态下散射场强的分布。SP 分子向 MC 分子异构化过程中，分子折射率与吸光系数的协同变化显著增大了散射截面和光程差，导致多重散射路径重构，从而导致散射场分布及散斑图案的动态重置。为了进一步探究 SPR-PUF 可逆异构化过程中折射率与吸光系数的变化对散斑图案重构的影响，我们引入汉明距离量化重构前后 SPR-PUF 生成的相同长度的二进制字符串之间的差异，用于度量散斑图案的重构程度。图 2(a)显示了 SP 分子和 MC 分子可逆异构化过程中 532 nm 波长处吸光系数变化^[11]，根据图中吸光系数的值计算了 SP 分子在可逆异构化过程中的消光系数并带入 FDTD 中仿真计算。

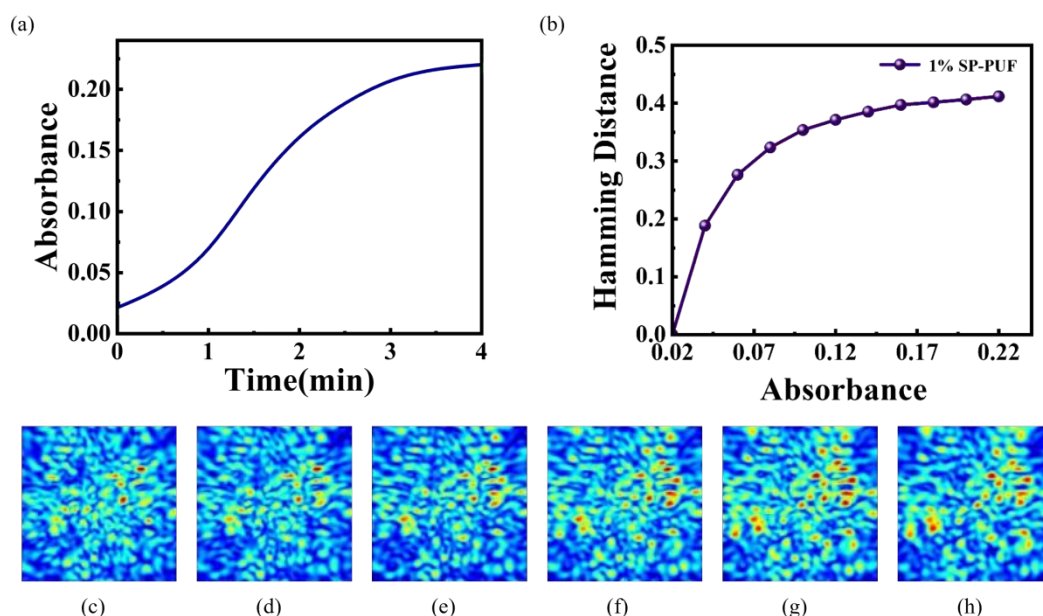


图 2 (a)SP 向 MC 异构化过程中 532 nm 波长处吸光系数变化图； (b)SP（吸光系数=0.02）散斑图案与其他吸光系数散斑图案的汉明距离对比结果；吸光系数分别为 0.02(c)、0.08(d)、0.12(e)、0.16(f)、0.2(g)、0.22(h)散射场分布。

相应的汉明距离结果如图 2 (b)所示，随着 C-O 键断裂开环，SP 向 MC 结构的变化引起 PUF 折射率的变化，汉明距离-吸光系数关系曲线先骤然上升后趋近平缓，呈现“剧增-饱和”双阶段的强非线性响应特征。从不同吸光系数散射场分布来看，在低吸光系数区间（0.02-0.12），少量的 SP 分子转化为 MC 分子，初始转化的 MC 分子（高折射率）和其周围 PMMA（较低折射率）以及 TiO₂（高折射率）之间的折射率对比度的改变强烈影响了光子的散射路径和局域场分布。同时由于 SPR-PUF 介质的高度无序和多重散射特性，微小的局部折射率变化会被极大地放大，少量 MC 分子生成显著改变散斑图案。汉明距离的剧增表明 CRP

实现了分钟级的快速动态刷新，显著提升了 SPR-PUF 的安全防伪性能与快速重构响应能力。在高吸光系数区间（0.12-0.22），随着光照持续，越来越多的 SP 分子转化为 MC 分子，大量区域已被 MC 分子占据，新增的 MC 分子带来的局部折射率提升相对于已改变的背景环境来说，其相对变化幅度减小。汉明距离增速趋缓，意味着散射场重构接近饱和状态，SPR-PUF 趋向一种新的，MC 分子主导的准稳态。饱和区汉明距离的低波动性显著提升响应可重复性，SPR-PUF 可设置更宽的容错阈值降低假阳性率，在保障防伪安全性的同时增加了系统防伪过程的可重复性。

3 结论

本研究提出一种基于螺吡喃光致变色的 PUF 设计方法，通过构建含 SP 分子与 TiO₂ 纳米粒子的 PMMA 复合无序介质，结合 FDTD 仿真，验证了 532 nm 激光下 SP 分子向 MC 分子异构化过程中折射率与吸光系数的协同变化过程对多重散射场的动态调制作用。仿真结果证实：SP 分子的引入显著改变了 TiO₂/PMMA 体系的散射场分布；汉明距离随吸光系数的变化呈现“剧增-饱和”双阶段特性，这种强非线性响应特征从物理层面协同优化了 PUF 快速可重构特性与防伪过程的可重复性。此外，SPR-PUF 的光控可重构机制无需依赖电^[7]/热^[8]等外场调控介质，简化了系统架构，在操作便捷性、制造成本方面具备显著优势。本研究提出的 SPR-PUF 具有光控快速重构响应能力与非线性响应特性为高安全性动态防伪标签的设计提供了可行的理论基础。

参考文献

- [1] Labrado C, Thapliyal H. Design of a piezoelectric-based physically unclonable function for IoT security [J]. *IEEE Internet of Things Journal*, 2019, 6(2): 2770–2777.
- [2] Kelesidis T, Falagas M E. Substandard/counterfeit antimicrobial drugs [J]. *Clinical Microbiology Reviews*, 2015, 28(2): 443–464.
- [3] Pappu R, Recht B, Taylor J, et al. Physical one-way functions [J]. *Science*, 2002, 297(5589): 2026–2030.
- [4] Zeitouni S, Oren Y, Wachsmann C, et al. Remanence decay side-channel: the PUF case [J]. *IEEE Transactions on Information Forensics and Security*, 2015, 11(6): 1106–1116.
- [5] Arppe R, Sørensen T J. Physical unclonable functions generated through chemical methods for anti-counterfeiting [J]. *Nature Reviews Chemistry*, 2017, 1(4): 1–13.
- [6] Gan Z, Chen F, Li Q, et al. Reconfigurable optical physical unclonable functions enabled by VO₂ nanocrystal films [J]. *ACS Applied Materials & Interfaces*, 2022, 14(4): 5785–5796.
- [7] Nocentini S, Rührmair U, Barni M, et al. All-optical multilevel physical unclonable functions [J]. *Nature Materials*, 2024, 23(3): 369–376.
- [8] Kim Y, Lim J, Lim J H, et al. Reconfigurable multilevel optical PUF by spatiotemporally programmed crystallization of supersaturated solution [J]. *Advanced Materials*, 2023, 35(22): 2212294.
- [9] Klajn R. Spiropyran-based dynamic materials [J]. *Chemical Society Reviews*, 2014, 43(1):

148–184.

- [10] Kortekaas L, Browne W R. The evolution of spiropyran: fundamentals and progress of an extraordinarily versatile photochrome [J]. *Chemical Society Reviews*, 2019, 48(12): 3406–3424.
- [11] Du L, Dai J, Xu Z-Y, et al. From shape and color memory PCL network to access high security anti-counterfeit material [J]. *Polymer*, 2019, 172: 52–57.

70GHz 电吸收激光调制器射频模型研究

赵小龙*, 陈发涛, 谢常叶

电子与信息学部微电子学院, 西安市微纳电子与系统集成重点实验室, 西安交通大学, 陕西 西安 710049

zhaoxiaolong@xjtu.edu.cn

摘要: 超高带宽电吸收激光调制器是高速光通信系统的关键元件。建立电吸收激光调制器的射频模型, 对于其超高带宽优化设计具有指导意义。本文结合物理先验知识和 S 参数测量结果特征, 建立电吸收调制激光器的射频等效电路模型。验证了该模型在 40GHz 以下与实验结果偏差 $<0.5\text{dB}$, 并且能够模拟器件在 70GHz 范围内响应的主要特征, 为更高带宽电吸收激光调制器芯片和模块优化设计提供了模型参考。

关键词: 电吸收激光调制器; 射频模型; S 参数; 宽带

1 前言

电吸收调制激光器 (Electro-absorption Modulated Laser, EML) 具有高带宽和低功耗特性, 是高速光通讯系统中的重要光源之一。为了提升通信速率需要设计尽量高带宽的 EML 模块。影响 EML 模块带宽的因素包括 EML 芯片结构、COC (Chip On Carrier) 基板结构、键合线、匹配电阻等。

韩国电子通信研究院 (ETRI) 的研究[1,2]表明通过设计合适的键合线长度和匹配电阻值, 可以显著改善 EML 模块带宽, 优化后带宽可达 55 GHz[2]。华中科技大学张敏明教授团队也对键合线设计优化 EML 模块带宽开展了研究, 优化后模块带宽可达 53.4 GHz [3]。

根据射频电路理论, 上述的基板结构、键合线、匹配电阻等参数优化设计过程, 可以看作是对 EML 芯片进行阻抗匹配, 以获得最大带宽。根据 Bode Fano 准则, 通过匹配所能获得的带宽和反射系数之间相互制约, 这主要由 EML 芯片本身结构决定, 即一旦芯片结构确定, 为了满足一定的反射系数约束, 则可获得的最大带宽由 EML 芯片决定。因此, 为了进一步提升 EML 模块带宽以及更容易实现宽带匹配, 需要 EML 芯片本身具有“宽带特性”。

基于此, 本文主要开展 EML 芯片的射频电路模型研究, 为进一步揭示 EML 芯片带宽限制因素以及优化设计奠定基础。

2 EML 射频模型建模方法

如图 1 所示为 EML 芯片的照片和结构示意图。EML 芯片主要由 DFB (Distributed FeedBack) 激光器和 EA (Electro-Absorption) 调制器构成。DFB 激光器和 EA 调制器的主要结构为量子阱, 可以视为复杂的 pn 结结构。在射频

电路中 pn 结通常可以采用电阻和电容的并联电路等效，本文所用的 EA 调制器施加反向直流偏置，这种情况下通常只需要考虑电容的影响。在 DFB 激光器和 EA 调制器之间的互联段将引入一定的串联电阻。

基于这些先验知识，再结合 S 参量测量结果的特征，即可给出 EML 芯片的关键电路结构，再通过优化调整，即可获得符合物理实际的等效电路模型。

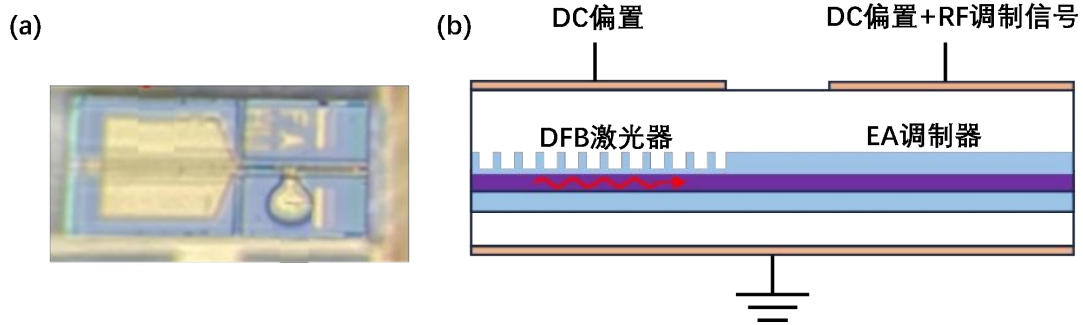


图 1 电吸收调制激光器 (EML) 芯片 (a) 照片和 (b) 结构示意图

3 结果分析

基于以上建模思路，提出图 2 所示的射频等效电路模型。这里考虑当激光器偏置为零时，无光输出，此时激光器的阻抗将较大，因此这部分对系统的影响可以忽略。当给激光器施加正向偏压，有光输出时，则需要考虑激光器的等效电路的影响。而其他等效元件参数无论在出光状态和不出光状态，其参数取值应当保持不变。

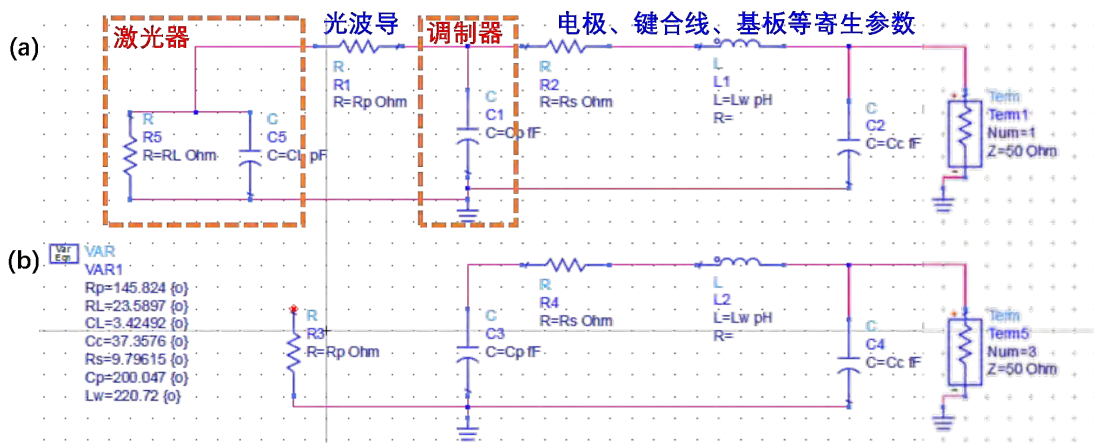


图 2 40GHz EML 射频等效电路模型，(a) 出光状态，(b) 不出光状态

仿真计算结果与测试结果的对比如图 3 所示。可见，在 40GHz 范围内，模型计算结果与实验值非常吻合，偏差在 0.5dB 以内，验证了模型的有效性。

当测试频段扩展到 70GHz 范围后，由于对于 70GHz 电信号其在芯片中的波长已经与芯片长度尺寸可比拟，因此，高频模型中需要进一步考虑分布寄生效应的影响。对图 2 电路结构进行优化后的在 70GHz 频段范围内的仿真计算结

果与测试结果的对比如图 4 所示。可见，等效模型依然能够反映实验结果主要特征。存在的偏差可能来自对 COC 以及测试探针的建模不完备。

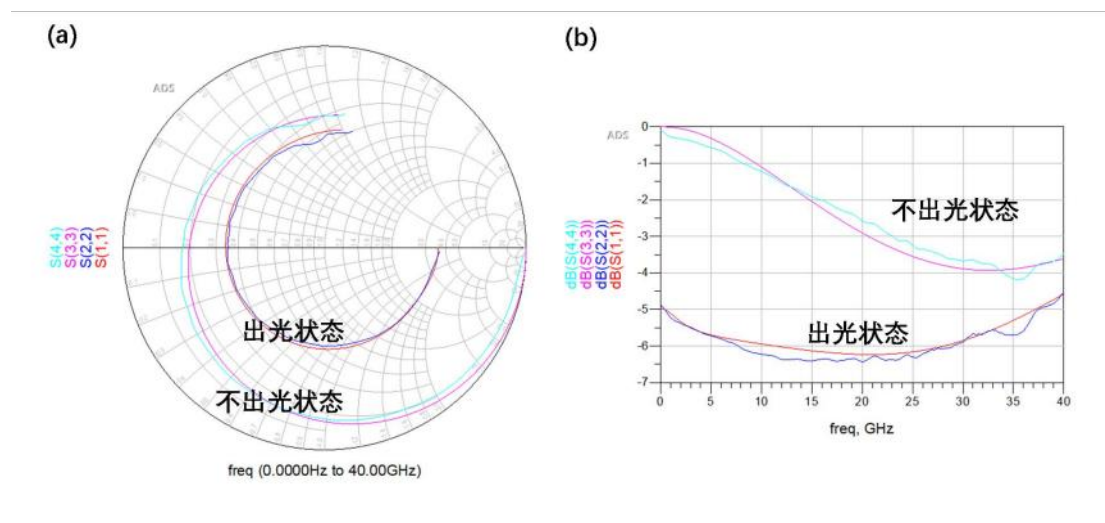


图 3 基于 40GHz EML 射频等效电路模型的 S 参数计算结果，(a) Smith 圆图，(b) S 参数幅度,其中蓝色和青色曲线为测量结果，红色曲线为等效模型计算结果

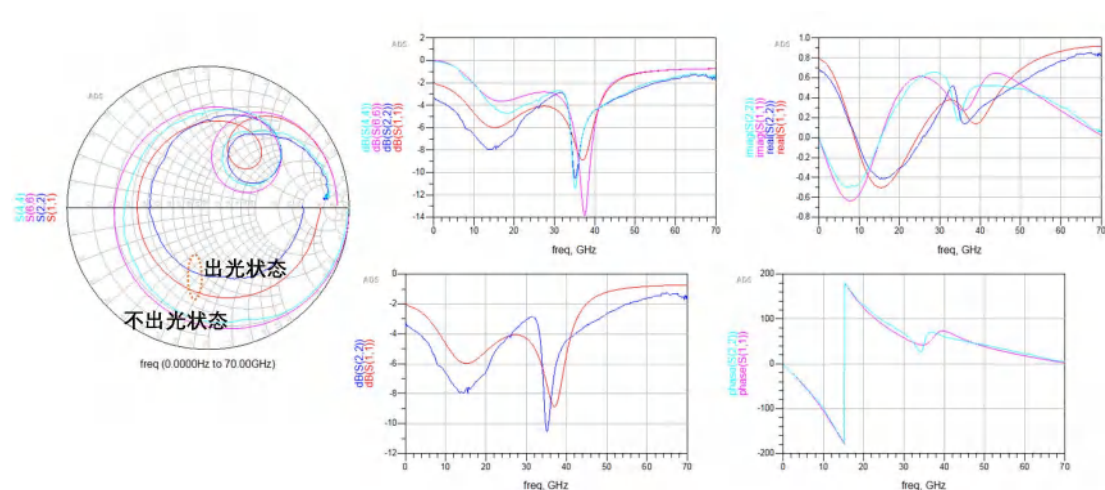


图 4 70GHz EML 射频等效电路模型理论计算与试验结果比较，其中蓝色和青色曲线为测量结果，红色曲线为等效模型计算结果

4 结论

本文结合物理先验知识和 S 参数测量结果特征，建立了电吸收调制激光器射频等效电路模型。在 40GHz 以下，该模型与实验结果吻合很好，偏差在 0.5dB 以内，验证了该模型的有效性。同时，当频率扩展到 70GHz，该模型能够揭示器件响应的主要特征。本文研究对更高带宽 EML 芯片和模块优化设计提供模型参考。

参考文献

- [1] S.-J. Yun et al., "112-Gbaud PAM4 Operation of Lumped-EML With 150-um EAM Length

- Using LC Resonance Based on Matching Resistance Optimization," *Journal of Lightwave Technology*, vol. 42, no. 1, pp. 229-235, 2024.
- [2] O.-K. Kwon, Y.-T. Han, Y.-S. Baek, and Y.-C. Chung, "Improvement of modulation bandwidth in electroabsorption-modulated laser by utilizing the resonance property in bonding wire," *Opt. Exp.*, vol. 20, no. 11, pp. 11806–11812, 2012.
- [3] Y. Han et al., "Bandwidth Extension of EML Chip-on-Carrier Submodule Using Precise Parameters Extraction," in *IEEE Photonics Technology Letters*, vol. 35, no. 23, pp. 1251-1254, 1 Dec.1, 2023.

芯粒轻量化热模型与集成微系统快速热仿真应用

王大伟, 王乐天, 刘军, 孙玲玲, 李文钧, 赵文生

杭州电子科技大学

摘要: 先进封装高集成度与高功率密度使其热问题尤为突出, 而其几何跨尺度与物理多尺度等特性导致传统数值仿真难以实现系统级高效仿真。本文提出一种基于轻量化芯粒热模型的堆叠式系统级高效热建模仿真方法。首先, 通过兼容降阶或神经网络模型, 无需暴露芯粒细节, 极大的提升仿真效率的同时支持芯粒-封装内温度场重构; 随后, 提出多模型空间协同建模方法, 通过构建信息交换接口模型, 实现了轻量化与传统热模型协同仿真; 最后, 通过芯粒与微系统建模仿真案例, 对提出的建模仿真方法进行了验证, 结果表明, 其在保证仿真精度的同时, 可实现复杂系统建模仿真速度提升超过两个数量级。

关键词: 集成封装微系统; 芯粒轻量化热模型; 多类型模型空间协同; 快速热仿真方法

1 前言

近年来, 先进封装技术因其可通过 2.5D 或 3D 堆叠实现系统集成度的持续提升, 逐渐成为业界关注的焦点^[1]。其中, 基于芯粒的架构因其相较传统单芯片系统 (SoC) 设计, 支持模块化设计与异构集成, 具备更高良率、更强灵活性与可扩展性, 并能有效降低成本, 成为一种极具前景的解决方案^[2]。然而, 极致紧凑的封装结构带来了显著提升的集成度和功率密度, 使热问题成为制约集成微系统性能的关键因素。因此, 在设计阶段通过建模与仿真技术开展热性能分析, 并指导热管理优化, 对实现高性能集成微系统设计至关重要^[3]。

然而, 现有大多数热建模与仿真技术难以实现仿真精度与效率平衡。热阻网络模型是一种适用于大规模电子器件高效热分析的方法^{[4]-[7]}, 但通常只能预测少量预设观测点的温度, 精度较低, 且无法提供详细的热分布信息。而基于有限元等数值方法的热建模虽然能提供高精度热分析^{[8]-[9]}, 但由于集成封装微系统的特征尺寸从纳米到厘米跨度巨大, 几何模型空间离散后网格数量庞大, 导致热模型规模庞大, 仿真效率极低^[10]。尽管可以通过模型简化提升仿真效率, 但往往会牺牲仿真精度^{[11]-[12]}。此外, 在基于芯粒的集成封装系统中进行数值仿真时, 往往需要获取芯粒内部的结构细节信息, 这可能会导致敏感设计信息的泄露。

近年来, 基于模型降阶与神经网络的轻量化建模方法被广泛应用于电子器件热仿真加速中^{[13]-[16]}。例如, Xiao Li 等提出了一种基于科学机器学习算法的建模方法, 用于实现绝缘栅双极晶体管 (IGBT) 的高效热评估^[15]; Hui Li 等, 构建了一种面向 Press-Pack IGBT 的热孪生模型, 可在稳定电流输入下实现器件温度

场的实时预测。这类轻量化方法能够显著提升热仿真的计算效率^[4]。

然而，现有轻量化建模方法普遍面临一个问题：由于模型降阶通常直接作用于完整器件的热模型，当器件设计发生变更时，需重新训练模型，导致其在集成封装系统热优化设计中的实用性较差。相比之下，重新训练轻量化模型所需的代价甚至可能高于直接进行数值仿真^{[17]-[20]}。为解决上述问题，本文提出了一种面向集成封装微系统的高效热建模与仿真方法，该方法基于轻量化模型与传统数值仿真模型的空间协同建模机制，能够支持系统级热设计优化。本文的主要贡献如下：

(1) 提出了一种基于接口热流连续性的热模型分解策略，实现了芯粒与中介层模型空间分离；

(2) 提出了一种基于模型降阶与神经网络结合的芯粒轻量化热建模方法，所构建的芯粒轻量化热模型（LCTM）在大幅减少内存占用的基础上，可根据边界分解模型提供的热流密度边界条件，实现毫秒级快速重建芯粒内部温度场，并保持较高精度与良好泛化能力；

(3) 提出了一种基于 LCTM 与传统封装结构热模型的空间协同建模与仿真方法。通过小尺寸接触面的热阻矩阵计算相邻模型间的热流矩阵，从而实现对整个基于芯粒的集成微系统的快速热仿真。

2 热建模方法

根据 2.5D 集成微系统的几何尺度特性及其设计优化需求，封装内部的对象可分为两类：一类是特征尺寸较小、设计保持不变的芯粒，另一类是特征尺寸较大、需要进行优化的封装结构。基于上述考虑，芯粒与封装结构将采用不同的建模策略。其中，对芯粒结构采用基于模型降阶与神经网络构建的可复用轻量化热模型，而与其相连的中介层结构则采用数值热模型进行建模以满足优化需求。在构建集成微系统模型时，LCTM 可通过边界接口与数值模型进行信息交互，最终实现整个集成微系统的热仿真。

2.1 基于热流连续条件的模型域分解

集成微系统中的热传导过程可由热传导方程描述^[21]：

$$\begin{cases} 0 = \nabla \cdot (\kappa \nabla T) + Q_s \\ \kappa \frac{\partial T}{\partial n} \Big|_{\Gamma_c} = -h(T - T_{amb})_{\Gamma_c} \end{cases} \quad (1)$$

其中， T 表示温度， Q_s 表示热源项， κ 表示材料的热导率， h 和 T_{amb} 分别表示接触界面上的对流换热系数与环境温度。如图 1 所示，基于热流守恒原理^[22]，热模型中相邻的两个区域可以通过在界面上施加纽曼边界条件实现分离。所施加的纽曼边界条件应保证界面温度的一致性和热流连续性，即满足如下条件：

$$\begin{cases} -\kappa_1 \frac{\partial T_1}{\partial n} = -\kappa_2 \frac{\partial T_2}{\partial n} \Leftrightarrow q_1 = q_2 \\ T_1 = T_2 \end{cases} \quad (2)$$

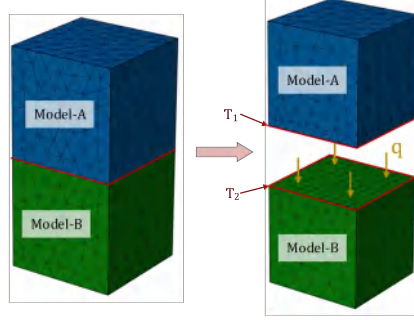


图 1 热模型区域划分过程的示意图

其中， T_n 与 $\frac{\partial T_n}{\partial n}$ 分别表示区域 n 在界面上的温度和法向温度梯度。这表示在理想接触的假设下，热模型中相邻区域在界面上的温度连续，为构建多区域耦合的热模型提供了理论基础。在此基础上考虑将式(1)所示的偏微分方程应用有限元法进行离散后，可得：

$$\mathbf{KT}_{model} = \mathbf{B} \quad (3)$$

其中， \mathbf{K} 表示刚度矩阵， \mathbf{B} 表示质量矩阵， \mathbf{T}_{model} 表示温度节点值向量。考虑 Neumann 边界条件与 Robin 边界条件两种常见边界条件，基于刚度矩阵与质量矩阵在有限元离散中的组装过程，式(3)可进一步分解为以下矩阵形式：

$$\mathbf{KT} = \mathbf{B}_P(P) + \mathbf{B}_h(h, T_{amb}) + \mathbf{B}_q(q) \quad (4)$$

其中， P 表示芯粒内部的功率密度， q 表示接触界面处的热流密度。通过矩阵变换，可以将温度场表示为多个温度模态的线性组合，式（4）可进一步转化为：

$$\mathbf{T}_{model} = \boldsymbol{\vartheta}_P(P) + \boldsymbol{\vartheta}_{h, T_{amb}}(h, T_{amb}) + \boldsymbol{\vartheta}_q(q) \quad (5a)$$

$$\boldsymbol{\vartheta}_P(P) = \mathbf{K}^{-1} \mathbf{B}_P(p) \quad (5b)$$

$$\boldsymbol{\vartheta}_{h, T_{amb}}(h, T_{amb}) = \mathbf{K}^{-1} \mathbf{B}_h(h, T_{amb}) \quad (5c)$$

$$\boldsymbol{\vartheta}_q = \mathbf{K}^{-1} \mathbf{B}_q(q) \quad (5d)$$

其中， $\boldsymbol{\vartheta}_P$ ， $\boldsymbol{\vartheta}_{h, T_{amb}}$ ， $\boldsymbol{\vartheta}_q$ 分别对应热源功率 P ，所施加 Robin 边界条件和 Neumann 边界条件对应的温度模态。对于纽曼边界条件所对应的温度模态 $\boldsymbol{\vartheta}_q$ ，由于 $\mathbf{B}_q(q)$ 的每一列只存在一个有效值（其余都为0），所以 $\boldsymbol{\vartheta}_q$ 又可以表示为：各节点热流密度与其对应温度模态 $\boldsymbol{\vartheta}_{q,n}$ 的加权叠加之和，即：

$$\boldsymbol{\vartheta}_q(q) = \sum_{n=1}^m q_n \times \boldsymbol{\vartheta}_{q,n} \quad (6)$$

其中， q_n 表示接口上第 n 个节点对应的热流密度， m 为接口上的节点总数，其取值由模型的复杂度决定。至此，我们已经获得了一个描述温度模态与接口节点热

流密度之间映射关系的模型。获得接口上热流分布后，即可计算子区域内部的温度分布。

2.2 轻量化芯粒热模型

综上，2.5D 集成微系统的热仿真模型可被划分为三个部分：芯粒热模型、界面信息交换模型以及与芯粒相连接的封装结构热模型。界面信息交换模型起到了信息传递通道的作用，使得芯粒与封装结构的热模型在满足热流连续性与守恒性原则的同时，可实现解耦计算。在本研究中，封装结构部分仅考虑了中介层。鉴于芯粒热模型在集成微系统设计过程中保持不变且可独立求解，本文在本小节提出了一种基于模型降阶与人工神经网络的轻量化芯粒热建模方法，用于构建可复用的轻量化芯粒热模型。该模型不仅能显著提升热仿真计算效率，还可大幅降低计算资源消耗。

依据公式（6），可以在拥有边界热流密度分布的情况下，对芯粒的温度分布进行精确求解。然而，随着模型复杂度的不断提升，接口处的节点数量 m 也随之增加，从而显著增加模型求解的复杂度与耗时。为应对此问题，本文提出一种基于 ANN 的轻量化芯粒热建模方法，图 2 展示了该方法的整体流程。

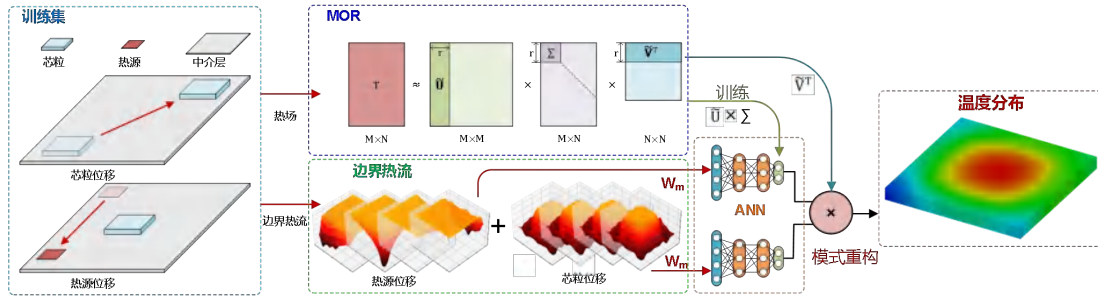


图 2 基于人工神经网络的芯粒轻量化热模型建模流程图

该方法首先构建一个包含中介层、芯粒和热源的基础模型，以获得神经网络的训练数据集。在数据生成过程中，通过持续改变芯粒与热源在中介层上的相对位置，模拟其间的耦合效应。对于特定的芯粒，其与中介层之间接口的热流分布通常具有空间连续性，即相邻节点之间的热流密度不会发生剧烈变化。这使得接口上的热流密度呈现出一定的规律性而非完全随机分布。

基于这一特性，可通过神经网络对上一届得到的模型进行降阶。利用一组维度远小于总节点数 m 的基函数及其对应系数，便可对整个温度场进行有效逼近与重建，从而实现对芯粒热特性的高效表征与快速求解。这些基函数可通过在构建训练集时经过仿真得到的一系列 ϑ_q 温度模态执行奇异值分解得到，其数学形式可表示为：

$$\vartheta_q = \sum_{k=1}^r \mathbf{U}_k \cdot \lambda_k \cdot \mathbf{V}_k^T = \sum_{k=1}^r \phi_k \mathbf{V}_k^T \quad (7)$$

其中， \mathbf{U}_k 表示左奇异矩阵 \mathbf{U} 的第 k 列， λ_k 表示奇异值矩阵 λ 中的第 k 个奇异值， \mathbf{V}_k^T

表示右奇异矩阵 V^T 的第 k 行。 ϕ_k 是一个确定性的空间基函数，其由 U_k 与 λ_k 的乘积得到。 λ_k 的大小反映了对应的 ϕ_k 所包含的信息量，且通常随着 k 的增加迅速衰减。保留的奇异值数量记为 r ，通常满足 $r \ll m \ll M$ ，其中 m 表示界面上的节点总数， M 为奇异值的总数。以 ϕ_k 为基函数，并引入神经网络形式替代式(7)中的热流项 $T_q(q)$ ，则该项可重新表示为：

$$\vartheta_q(q) = \sum_{k=1}^r \phi_k \cdot N(q_m) \quad (8)$$

其中， $N(q_m)$ 表示神经网络函数，其输入为界面上各节点的热流密度，输出为对应的降维系数。考虑到总热流 q_0 与 ϑ_q 之间的线性关系，可以对 ANN 模型进行训练，以预测热流分布权重 w_m ，从而减少训练数据的需求。据此，式(8)可以进一步改写为：

$$\vartheta_q(q) = q_0 \sum_{k=1}^r \phi_k \cdot N(w_m) \quad (9a)$$

$$w_m = \frac{q_m}{q_0} \quad (9b)$$

按照上述方法，模型的复杂度由原始的 m 降维至 r ，显著降低了计算开销，使得模型能够在保持较高精度的同时实现高效求解，满足实际集成封装系统中对快速热仿真的需求。

2.3 多机制模型协同建模仿真

本小节提出了一种将 LCTM 与有限元模型的空间协同建模与仿真方法。该方法通过构建接口信息交换模型，实现芯粒轻量化热模型与相邻中介层热模型之间的界面热流信息传递。需要强调的是，对于 LCTM，其对应的热阻抗矩阵仅需在建模阶段进行单次计算，在后续的热仿真过程中，即使需要更换不同材料参数的中介层与 LCTM 进行组装，该热阻抗矩阵亦可直接复用，从而大幅提升了仿真效率。下文以图 3 所示封装结构为例，详细介绍其提出的建模过程。

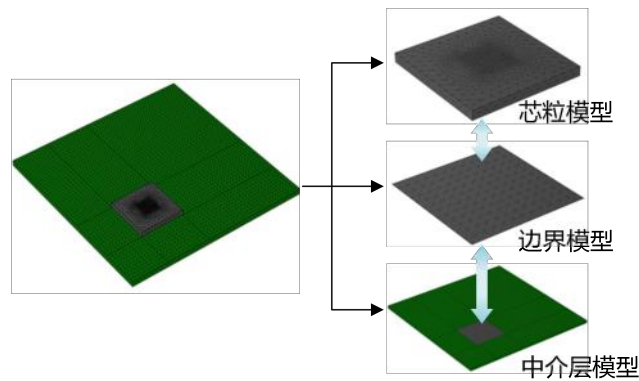


图 3 单芯粒封装示意图

芯粒与中介层的几何模型经过离散化后，可构建基于有限元方法的热模型。

根据式（5），可分别构建刚度矩阵与质量矩阵，进而可得：

$$\mathbf{T}_C = \mathbf{K}_C^{-1}(\mathbf{B}_{p,C}(P) + \mathbf{B}_{h,C}(h, T_{amb}) + \mathbf{B}_{q,C}(q)) \quad (10a)$$

$$\mathbf{T}_I = \mathbf{K}_I^{-1}(\mathbf{B}_{h,I}(h, T_{amb}) + \mathbf{B}_{q,I}(q)) \quad (10b)$$

进一步提取芯粒与中介层接口上的节点信息，并将式（10）重构为块矩阵形式，可得如下表达式：

$$\begin{bmatrix} \mathbf{T}_{C,O} \\ \mathbf{T}_{C,B} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{C,O} \\ \mathbf{R}_{C,B} \end{bmatrix} (\mathbf{B}_{p,C}(P) + \mathbf{B}_{h,C}(h, T_{amb}) + \mathbf{B}_{q,C}(q)) \quad (11a)$$

$$\begin{bmatrix} \mathbf{T}_{I,O} \\ \mathbf{T}_{I,B} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{I,O} \\ \mathbf{R}_{I,B} \end{bmatrix} (\mathbf{B}_{h,I}(h, T_{amb}) + \mathbf{B}_{q,I}(q)) \quad (11b)$$

$$\begin{bmatrix} \mathbf{R}_{X,O} \\ \mathbf{R}_{X,B} \end{bmatrix} = \mathbf{K}_X^{-1} \quad (11c)$$

当仅考虑接触边界的温度计算时，式（11）可进一步简化为：

$$\mathbf{T}_{C,B} = \boldsymbol{\vartheta}_{P,C,B}(P) + T_{amb} + \mathbf{R}_{C,B} \cdot \mathbf{B}_{q,C}(q_C) \quad (12a)$$

$$\mathbf{T}_{I,B} = T_{amb} + \mathbf{R}_{I,B} \cdot \mathbf{B}_{q,I}(q_I) \quad (12b)$$

由式（2）可知，式（12a）与式（12b）右端项相等，因此有：

$$\mathbf{T}_{P,C,B}(P) = \mathbf{R}_{C,B} \cdot \mathbf{B}_{q,C}(q_B) + \mathbf{R}_{I,B} \cdot \mathbf{B}_{q,I}(q_B) \quad (13a)$$

$$q_B = q_I = -q_C \quad (13b)$$

由于矩阵 $\mathbf{B}_q(q)$ 的非零元素仅分布在对应于接触界面节点的区域。基于上述特性，可对公式(13)中的相关矩阵进一步简化与合并，最终可得到一个用于求解热流密度矩阵 \mathbf{Q}_B 的紧凑形式表达式如下：

$$\mathbf{Q}_B = \left((\mathbf{R}_{C,B,B} + \mathbf{R}_{I,B,B}) \mathbf{B}_q(q_B) \right)^{-1} \mathbf{T}_{P,C,B}(P) \quad (14a)$$

$$\mathbf{R}_{C,B,B} = \mathbf{R}_{C,B} \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_B \end{bmatrix} \quad (14b)$$

$$\mathbf{R}_{I,B,B} = \mathbf{R}_{I,B} \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_B \end{bmatrix} \quad (14c)$$

其中， \mathbf{I}_B 是一个 m 阶的单位矩阵， $\mathbf{R}_{C,B,B}$ 和 $\mathbf{R}_{I,B,B}$ 是 m 阶的方阵， m 的大小由接触界面上的节点数决定。由于其在结构上与电流传导方程具有相似性，因此这些矩阵也被称为热阻网络矩阵。对于同一个轻量化 Chiplet 热模型，其热阻矩阵 $\mathbf{R}_{C,B,B}$ 只需要在训练阶段提取一次。基于上述方法，相邻热模型之间可以实现有效的信息交换。由于提取得到的热阻网络矩阵大小仅与接触界面上的节点数 m 有关，且 m 远小于模型的总节点数 n ，其计算复杂度远低于整体模型的求解过程。因此，

Chiplet 的温度分布可通过将热流密度矩阵 Q_B 输入至 LCTM 中高效计算获得。该方法在实现高效温度场仿真的同时，无需暴露 Chiplet 的内部结构与设计细节，从而兼顾了性能与数据安全性。

3 结果分析

本节通过一系列实验验证所提出建模方法的有效性、准确性与鲁棒性。为了保证对比的公平性，所有模型构建与仿真均在配备了一块 Intel i5-13400 CPU 的计算机上完成。

3.1 单芯粒模型

根据第 2.2 节描述的建模方法，首先构建芯粒的 LCTM 模型，随后通过与不同体积和导热系数的中介层进行协同仿真来验证 LCTM 的鲁棒性，如图 4 所示。在本案例研究中，为获取训练数据集，通过在尺寸为 $50\text{mm} \times 50\text{mm} \times 1\text{mm}$ 的中介层上移动芯粒的位置进行了多组热仿真模拟。每次仿真后，将提取芯粒与中介层界面间的热通量以及芯粒表面温度分布作为训练数据。仿真过程中，中介层和芯粒的导热系数均设置为 $148\text{W}/(\text{m}\cdot\text{K})$ ，模型边界条件设为对流边界，对流系数 $h = 4000\text{W}/(\text{m}^2\cdot\text{K})$ ，环境温度设为 293.15K 。获取数据集后，通过降阶和神经网络训练完成芯粒轻量化热模型构建。

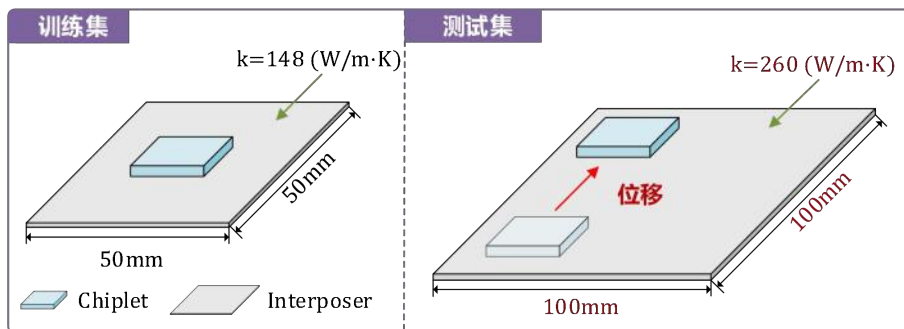


图 4 数据集生成过程中训练场景与测试场景的对比

为验证所提出芯粒轻量化热模型的鲁棒性，本文在超出模型训练范围的条件下进行协同仿真。通过调节中介层的热导率及几何尺寸，模拟了金属互连布线密度和封装布局尺度变化对热特性的影响，从而系统评估 LCTM 的泛化能力。在测试时将中介层的参数设置为导热系数为 $260\text{W}/(\text{m}\cdot\text{K})$ 、尺寸为 $100\text{mm} \times 100\text{mm} \times 1\text{mm}$ ，以检验模型在面对不同封装设计参数时的适应能力。仿真结果如图 5 所示，在图 5 (a) 中， M_x 和 M_y 分别表示芯粒在中介层表面沿 x 和 y 方向的平移距离；图 5 (b) 展示了 LCTM 与有限元模型计算结果之间的温度均方根误差随平移距离变化的趋势，最大均方根误差不超过 0.23K ，表明两种模型的结果高度一致。图 5 (c) 展示了得到最大均方根误差情况下的温度分布。在该场景中，LCTM 的平均绝对温度误差为 0.2K ，最大绝对误差为 1.2K ，仍能较好地保持预测精度。为了评估计算效率，将基于 LCTM 的组装系统热模型与有限元模型的仿真

耗时进行对比并在表 1 中展示了对比结果，相对于有限元模型，组装模型实现了 11.7x 的加速比。

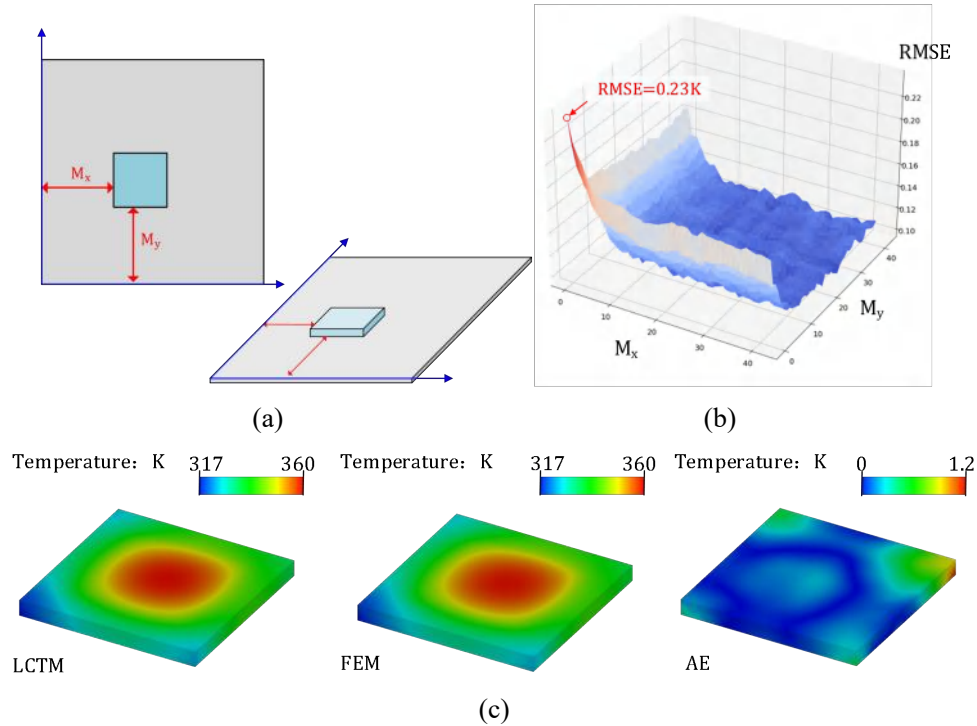


图 5 LTCM 的性能验证。(a) 模型化封装的示意图；(b) 作为小芯片位置的函数的温度的 RMSE；(c)

在 RMSE 最大处的温度分布的对比

表 1 单个芯粒组装模型计算时间对比

	DoF	有限元模型	组装模型	加速比
平均计算时间	7882	2.70s	0.23s	11.7x

综上所述，所提出的 MOR-ANN 建模方法所构建的热模型在多种封装结构参数变化下依然具备良好的有效性、准确性与鲁棒性，具备在实际系统热仿真中的广泛应用潜力。

3.2 多芯粒集成微系统模型

本节通过图 6 所示集成微系统进行热建模与仿真对所提出方法的精度与效率进行了验证。首先，基于提出的芯粒轻量化热建模方法，分别构建该系统中各芯粒（包括 HBM、GPU 和 CPU）的热模型。随后，将这些 LCTM 与中介层的有限元热模型进行空间组装，构建出整个集成微系统的热仿真模型。该系统的几何尺寸参数及功率配置如表 2 所示，通过上述方法构建的集成微系统热模型不仅显著降低了仿真计算的复杂度，同时在保持精度的基础上提升了建模效率，为复杂封装系统的热性能评估提供了有效手段。

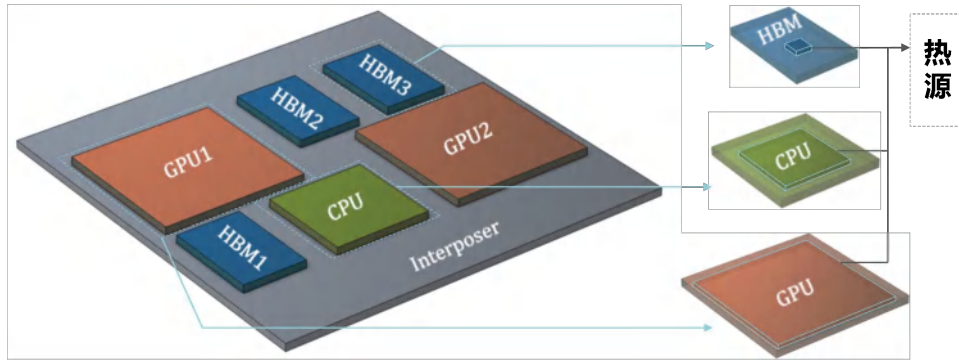


图 6 多GPU微系统的结构示意图^[23]

表 2 热模型中所用的尺寸与功率参数

模型	几何尺寸[长,宽,高] (毫米)	功耗(瓦)
GPU	[18,18,1]	295
CPU	[12,12,1]	105
HBM	[8,12,1]	20
中介层	[50,50,1]	-

在该系统中芯粒和中介层的表面边界条件统一设置为对流边界,其中对流换热系数设为 $h = 4000 W/(m^2 \cdot K)$,环境温度设为 $T_{amb} = 293.15K$ 。

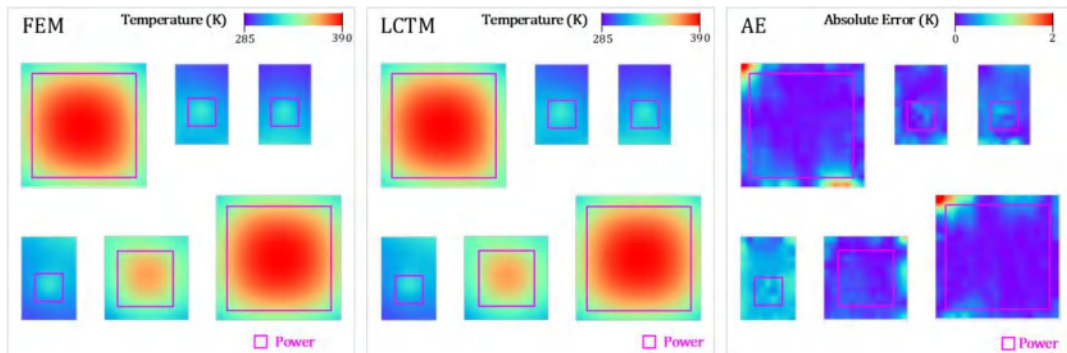


图 7 (a) 基于有限元模型获得的界面温度分布; (b) 基于LCTM模型获得的界面温度分布; (c) 两组数据之间的绝对误差分布图

表 3 多GPU集成微系统组装模型计算时间对比

	DoF	有限元模型	组装模型	加速比
平均计算时间	35,931	5.60s	0.47s	11.9x

图 7 对比展示了所提出的装配热模型与传统有限元方法在热仿真中的模拟结果。从图中可以看出,两种模型在界面区域所预测的温度分布高度一致,最大温度差异不超过 2K,表明所提出模型在保证精度方面具有良好表现。在表 3 中进一步对比了两种方法的计算时间,结果表明,在保持较高仿真精度的同时,所

提出方法实现了 11.9 倍的加速效果，充分验证了其在计算效率方面的显著优势。

最后，基于 RQS-TTC-SC01 热测试模拟芯片，设计了如图 8 (a) 所示热测试电路，进一步验证了所提出建模仿真方法。所使用模拟芯片通过加热元件精确模拟实际工作芯片的发热特性，并利用测温元件监测芯片在给定激励条件下指定位置下的实时温度。温度监测系统由待测模块、红外热像仪、数据采集器、三路直流电源和计算机组成，直流电源为待测模拟芯片供电，红外热像仪用于实时捕捉测试模块表面温度分布；数据采集器结合数据处理软件，实现对内部特定测试点温度的实时监测。对于测试点温度监测，利用数据采集器获取温度敏感二极管的实时电压，并将采集到的数据传输到数据处理软件，计算出相应的温度值，所构建模型预测温度与测试温度对比如图 8 (b) 和 (c) 所示，可以看出两组数据吻合较好。

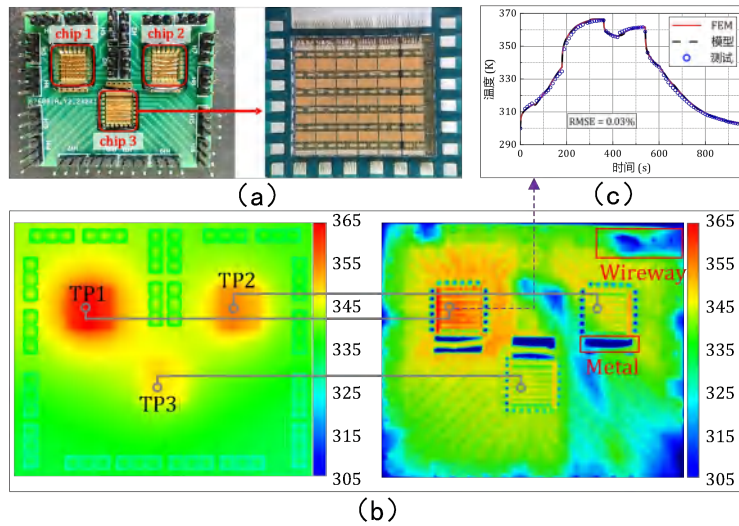


图 8 (a) 温度测试芯片与测试电路；(b) $t=360$ s 时测试电路表面温度；(c) 测试点 TP1 处温度随时间变化曲线

4 结论

本文提出了一种轻量化的芯粒热建模方法，用于加速集成封装系统的三维热场仿真。仿真结果表明，所构建的芯粒轻量化热模型能够在毫秒级的时间内完成芯粒的热仿真，同时保持较高的计算精度。与基于有限元方法的传统模型相比，该方法在保证精度的前提下，热仿真加速比可达一个数量级以上。

致谢

本研究得到国家自然科学基金项目（编号 62293493、92373117、U24A20296、62222401）以及国家重点研发计划项目（编号 2025YFE0102500）资助。

参考文献

- [1] J. H. Lau. Recent advances and trends in advanced packaging[J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2022, 12(2): 228-252.

- [2] M. Zhou, L. Li, F. Hou, G. He, J. Fan. Thermal modeling of a Chiplet-based packaging with a 2.5-D through-silicon Via interposer[J]. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2022, 12(6): 956-963.
- [3] X. Wang, Y. Yang, D. Chen, D. Li. Intelligent design method of thermal through silicon via for thermal management of Chiplet-based system[J]. *IEEE Transactions on Electron Devices*, 2023, 70(10): 5273-5280.
- [4] H. Li, et al. A thermal twin modeling method of press pack IGBT based on power loss[J]. *IEEE Transactions on Electron Devices*, 2022, 69(12): 6922-6928.
- [5] Y. Shi, J. Liu, Y. Ai, S. Chen, Y. Bai. Dynamic IGBT three-dimensional thermal network model considering base solder degradation and thermal coupling between IGBT chips[J]. *IEEE Transactions on Transportation Electrification*, 2023, 9(2): 2994-3011.
- [6] X. Sun, et al. Design and evaluation of a face-down embedded SiC power module with low parasitic inductance and low thermal resistance[J]. *IEEE Transactions on Power Electronics*, 2023, 38(3): 2799-2804.
- [7] F. Qin, Y. Zhang, T. An, R. Zhou. An improved thermal network model of press-pack IGBT modules considering contact surface damage[J]. *IEEE Transactions on Device and Materials Reliability*, 2023, 23(4): 444-452.
- [8] A. S. Bahman, K. Ma, F. Blaabjerg. A lumped thermal model including thermal coupling and thermal boundary conditions for high-power IGBT modules[J]. *IEEE Transactions on Power Electronics*, 2018, 33(3): 2518-2530.
- [9] K. Ma, A. S. Bahman, S. Beczkowski, F. Blaabjerg. Complete loss and thermal model of power semiconductors including device rating information[J]. *IEEE Transactions on Power Electronics*, 2015, 30(5): 2556-2569.
- [10] J. Li, M. Tang, L.-S. Wu, L. Jiang, W. Dai, J. Mao. LB-ADI: An efficient method for transient thermal simulation of integrated Chiplets and packages[J]. *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, 2024, 9: 149-156.
- [11] Q. Shi, Y.-M. Cheung, J. Lou. Robust tensor SVD and recovery with rank estimation[J]. *IEEE Transactions on Cybernetics*, 2022, 52(10): 10667-10682.
- [12] H. Zhao, Z. Zhang, Y. Yang, J. Xiao, J. Chen. A dynamic monitoring method of temperature distribution for cable joints based on thermal knowledge and conditional generative adversarial network[J]. *IEEE Transactions on Instrumentation and Measurement*, 2023, 72: 4507014.1-14.
- [13] L.-T. Wang, D.-W. Wang, B.-W. Zhang, X.-F. Yang, W.-S. Zhao. Proper orthogonal

- decomposition and long short-term memory neural network-based multiphysics digital twin model for electronic device online condition monitoring[J]. *IEEE Transactions on Instrumentation and Measurement*, 2025, 74: 1-13. Art No. 2523113.
- [14] J. Kuprat, K. Debbadi, J. Schaumburg, M. Liserre, M. Langwasser. Thermal digital twin of power electronics modules for online thermal parameter identification[J]. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2024, 12(1): 1020-1029.
- [15] X. Li, et al. Scientific machine learning enables multiphysics digital twins of large-scale electronic chips[J]. *IEEE Transactions on Microwave Theory and Techniques*, 2022, 70(12): 5305-5318.
- [16] Y. Yang, X. Zhao, Q. Cheng, R. Guo, M. Li, J. Zhou. POD-ANN as digital twins for surge line thermal stratification[J]. *Nuclear Engineering and Design*, 2024, 428: 113487.
- [17] L. Chen, W. Jin, J. Zhang, S. X. -D. Tan. Thermoelectric cooler modeling and optimization via surrogate modeling using implicit physics-constrained neural networks[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023, 42(11): 4090-4101.
- [18] C. Entzminger, W. Qiao, L. Qu, J. L. Hudgins. Automated extraction of low-order thermal model with controllable error bounds for SiC MOSFET power modules[J]. *IEEE Transactions on Power Electronics*, 2024, 39(1): 538-551.
- [19] L. Codecasa, V. d'Alessandro, A. Magnani, A. Irace. Circuit-based electrothermal simulation of power devices by an ultrafast nonlinear MOR approach[J]. *IEEE Transactions on Power Electronics*, 2016, 31(8): 5906-5916.
- [20] J. Zhang, et al. Hybrid data-driven and mechanistic modeling approach for power module rapid thermal analysis[J]. *Engineering Applications of Artificial Intelligence*, 2024, 39(11): 14617-14629.
- [21] X. Liu, W. Peng, Z. Gong, W. Zhou, W. Yao. Temperature field inversion of heat-source systems via physics-informed neural networks[J]. *Engineering Applications of Artificial Intelligence*, 2022, 113: 104902.
- [22] P. Xie, L. Xu, J.-H. Yin, H. Wang, Z.-H. Yang, B. Li. An interior penalty domain decomposition method for thermal analysis of 3-D integrated systems[J]. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2021, 11(3): 395-406.
- [23] Y. Ma, L. Delshadtehrani, C. Demirkiran, J. L. Abellan, A. Joshi. TAP-2.5D: A thermally-aware Chiplet placement methodology for 2.5D systems[C]. *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2021: 1246-1251.

金属-介电混合 O-PUF 中散斑相关性的调控机制及其 响应熵增强效应研究

惠瑞杰^{1, 2}, 陈飞良^{1, 2}, 方浩年^{1, 2}, 李沫^{1, 2, *}, 张健^{1, 2, *}

1. 电子科技大学电子科学与工程学院 成都 611731

2. 太赫兹辐射与应用四川省重点实验室 成都 611731

jianzhang@uestc.edu.cn; limo@uestc.edu.cn

摘要: 光学物理不可克隆函数 (O-PUF) 凭借其物理本质的高度随机性和不可预测性, 成为硬件安全领域备受关注的技术候选者。然而固有的散斑相关性物理机制限制了纯介电粒子体系 O-PUF 响应的随机性。本研究提出金属-介电粒子混合的 O-PUF, 仿真建立了二氧化钛和金纳米粒子在聚甲基丙烯酸甲酯随机分布的 O-PUF 三维模型。通过时域有限差分仿真与带吸收系数修正的散斑相关函数模型分析了 O-PUF 的散斑模式随位移、动量和金属体积分数变化的相关性, 发现混入更多金纳米粒子的 O-PUF 散斑短程相关性增强, 独立激励-响应对数量减少; 散斑长程相关性降低, 静态响应熵升高。这种增强的物理随机性使金属-介电粒子混合 O-PUF 成为安全认证与随机数生成的理想方案之一。

关键词: 光学物理不可克隆函数 (O-PUF); 纳米粒子; 时域有限差分法; 散斑相关函数; 静态响应熵

1 前言

防伪认证作为安全交互的核心机制, 通过对实体身份的核验保障硬件真实性或信息传输可信性。当前假冒身份引发的社会经济风险持续加剧, 尤其伪劣产品泛滥已危及公共健康安全^[1]。物理不可克隆函数 (Physical Unclonable Function, PUF) 可将身份信息固化于硬件微纳结构的固有随机性中, 通过动态激励-响应对 (Challenge-Response Pairs, CRPs) 实时生成高熵密钥, 其静态熵直接源自物理熵源的随机特性, 无需密钥存储即可构建本质可靠的硬件信任根。

近年来 PUF 的研究大多聚焦于电学 PUF (E-PUF) 与光学 PUF (O-PUF)。E-PUF 具有良好的集成兼容性, 但容易受建模攻击与侧通道攻击^[2,3]; O-PUF 例如物理机制依赖于多重散射, 激光束通过非均匀材料传播时产生的随机干涉图案 (散斑), 被证明具有高安全性, 适用于对认证安全性要求更高的场景^[4,5]。当前 O-PUF 主流方案采用高折射率介电粒子 (如 TiO₂、ZnO、GaP 等) 嵌埋于透明基底。虽然纳米粒子引入的短平均自由程有利于散射, 但其散射路径的重合却导致了较强的长程相关性, 降低了单次测量散斑图样的随机性, 最终限制了静态响应熵的提升。本研究提出一种金属-介电混合结构的 O-PUF, 利用金属的吸收效应抑制散射长程相关性, 从而突破纯介电粒子体系对静态响应熵的限制。针对该方案的实际应用安全性, O-PUF 需要保障充足的独立 CRPs 数量与静态响应熵,

其随金属组分混入的变化亟待探究。

针对上述问题，本研究揭示了纯介电粒子体系 O-PUF 中金属组分的引入对散斑短程/长程相关性的调控机制，建立了金属-介电混合 O-PUF 中独立 CRPs 数量与静态响应熵的变化规律。以聚甲基丙烯酸甲酯 (PMMA) 基底嵌埋金红石态的 TiO_2 粒子的 O-PUF 为研究对象，以额外加入金纳米粒子的 O-PUF 作为对照组，分别建立了对应的三维均匀随机分布模型，并进行了时域有限差分 (FDTD) 仿真，通过 PUF 位移对相同透射模式的影响研究了散斑短程相关性，通过不同透射模式的空间平均研究了散斑长程相关性。仿真结果证明金纳米粒子的引入增强了散斑短程相关性，独立 CRPs 数量减少，散斑长程相关性降低，静态响应熵升高。进一步地，我们比较仿真和理论模型，发现加入金属后散斑相关性的变化可以近似地被带吸收系数修正的散斑相关函数理论描述，分别利用带吸收系数修正的散斑短程相关函数与散斑长程相关函数量化了引入金属对 O-PUF 的独立 CRPs 数量与静态响应熵的影响。

2 模型构建

2.1 FDTD 仿真模型

为模拟 O-PUF 的激励-响应行为，采用波长为 633 nm 的氩氦激光光源。建模构建的 PMMA 基底尺寸为 $10\ \mu\text{m} \times 5\ \mu\text{m} \times 2\ \mu\text{m}$ (图 1a, b)，其在 633 nm 波长下的折射率为 1.489。通过数值计算 TiO_2 和金粒子不同粒径的归一化消光/散射/吸收截面如图 1e, f 所示，发现 TiO_2 粒径约为 300nm 时对应第一个米氏共振峰，金粒径约为 100nm 时对应第一个吸收共振峰，同时保留了与 TiO_2 近似的归一化散射截面，在这两个尺度下 TiO_2 和金均有良好的散射效率，保障了 O-PUF 的随机性。因此，我们在 PMMA 基底内构建了含 300 nm TiO_2 粒子 (5 vol%) 均匀随机分布的纯介电 O-PUF 三维仿真模型；作为对照，额外引入 100 nm 金粒子 (0.2 vol%)，形成金属-介电混合 O-PUF 三维仿真模型。

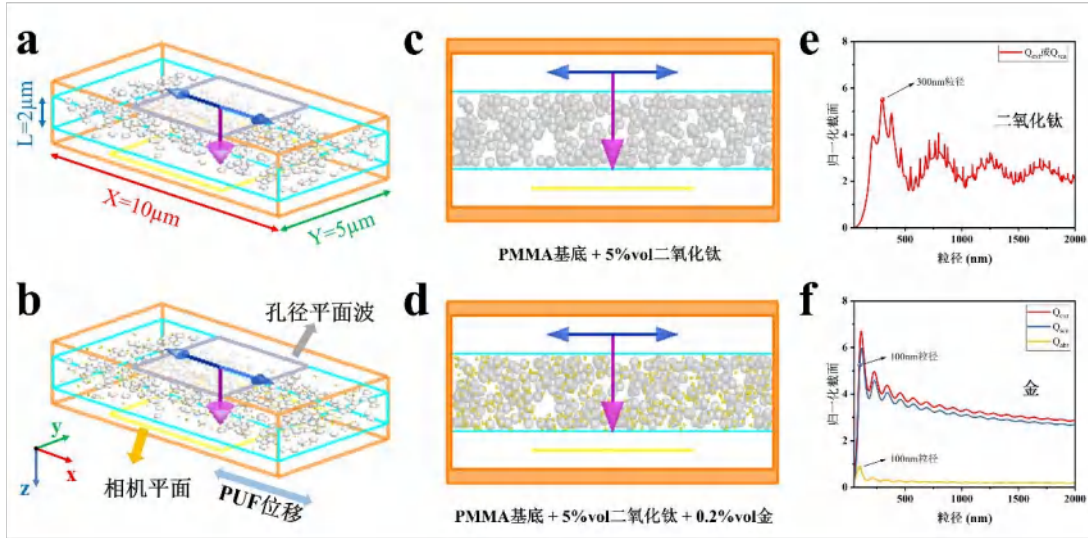


图 1 (a, b) 纯介电与金属-介电混合 O-PUF 三维仿真模型示意图；(c, d) 纯介电与金属-介电混合 O-PUF 三维仿真模型正视图；(e, f) TiO_2 与金粒子分别在 PMMA 基底中的归一化散射截面，其中，消光截面 Q_{ext} ，散射截面 Q_{sca} ，吸收截面 Q_{abs} 。

我们采用 FDTD 方法仿真平面波对纯介电与金属-介电混合 O-PUF 模型的透射电场。仿真设置完美匹配层边界条件，针对非周期结构使用 $4\mu\text{m} \times 4\mu\text{m}$ 孔径平面波（隔离边界效应），光源以 x 方向偏振、沿 z 轴传播照明 O-PUF，在模型 z 轴正方向 $0.5\mu\text{m}$ 处设置 $5\mu\text{m} \times 5\mu\text{m}$ 监测平面记录电场分布作为散斑图样。

为模拟 O-PUF 的散斑短程相关性，通过笛卡尔坐标系变换对 O-PUF 模型施加 $\pm x$ 方向位移，同步位移照明与监测区域以模拟位移后透射电场分布；按奈奎斯特准则以半波长间距采样电场数据，经快速离散傅里叶变换获得横向动量 \mathbf{k} 空间中输入模式 **a** 至输出模式 **b** 的透射复振幅 \mathbf{E}_{ab} 。输入振幅归一化后得透射系数 $T_{\text{ab}} = |\mathbf{E}_{\text{ab}}|^2$ ，计算位移前后全模式透射系数的皮尔逊相关系数，即归一化散斑相关函数。因照明/监测区域同步位移匹配，散斑相关函数主要表征短程相关性^[6]。

为模拟 O-PUF 的散斑长程相关性并抑制短程干扰，基于记录的透射电场数据，计算固定动量差 $\Delta\mathbf{q} = 25\delta\mathbf{q} \gg \delta\mathbf{q}$ 的模式间皮尔逊相关系数绝对值作为长程相关性估计。其中 $\delta\mathbf{q}$ 为相邻模式动量差，因大动量差下散斑短程相关性贡献呈指数级衰减，该方法可有效分离长程相关性分量。

2.2 理论近似模型

FDTD 仿真模型虽可初步表征金属-介电混合 O-PUF 中散斑短程与长程相关性的变化趋势，但受限于结构无序性导致的相关性概率波动，独立 CRPs 数量无法定量描述；同时短程相关性对同动量模式长程贡献的掩盖，致使静态响应熵变化亦难以量化。为此，本研究提出采用含吸收系数修正的散斑相关函数模型近似

量化金属-介电混合体系相关性，并通过比较 FDTD 模型仿真的结果进行定性验证。

鉴于引入金属粒子体积分数远低于介电粒子 (<0.1)，其对 O-PUF 传输自由程 l_t 的影响可忽略，但金属共振吸收显著增大体系吸收系数 K 。由辐射传递方程推导， K 与金属体积分数 $f_{\text{金属}}$ ，归一化吸收截面 $Q_{\text{abs-金属}}$ 和粒径 $d_{\text{金属}}$ 有以下近似关系^[7]：

$$K = \sqrt{\frac{9Q_{\text{abs-金属}}f_{\text{金属}}}{2d_{\text{金属}}l_t}} \quad (1)$$

含吸收系数修正的散斑短程和长程相关函数 C_1 和 C_2 表达如下^[8,9]：

$$C_1 = D_1 \delta_{\Delta\mathbf{q}_a, \Delta\mathbf{q}_b} \frac{|M|^2}{K^2} \left| \frac{(1 + K^2 z_0^2) \sinh(KL) + 2Kz_0 \cosh(KL)}{(1 + M^2 z_0^2) \sinh(ML) + 2Mz_0 \cosh(ML)} \right| \quad (2)$$

$$C_2 = \frac{D_2}{gN_{\text{mod}}} F_2(\Delta\mathbf{q}, K) \quad (3)$$

其中， $\Delta\mathbf{q}_a$ 和 $\Delta\mathbf{q}_b$ 分别是输入模式和输出模式的横动量差， $M = \sqrt{(\Delta\mathbf{q})^2 + K^2}$ 是沿传播方向指数衰减的质量， z_0 是外推长度， g 是平均透射系数， N_{mod} 是照明面积支持的模式数， D_1 和 D_2 是单位阶常数（约为 1）， F_2 是归一化长程相关函数的一般表达式。

为描述独立 CRPs 数量的变化，考虑 C_1 的半高全宽主导了散斑的去相关长度，与独立 CRPs 数量成反比关系，因此独立 CRPs 数量增益 G_I 可以表示为：

$$G_I = \frac{[C_1^{-1}(0.5)]_{\text{介电}}}{[C_1^{-1}(0.5)]_{\text{金属-介电}}} \quad (4)$$

为描述静态响应熵的变化，考虑 C_2 导致散射路径合并对响应的贡献^[10]，O-PUF 的静态响应熵 H 和熵增益 G_H 可以表示为：

$$H = \frac{D_1 N_{\text{mod}} g}{D_2 F_2(0, K)} \log\left(\frac{\pi e}{2} \frac{N_\phi}{N_{\text{mod}}}\right) \quad (5)$$

$$G_H = \frac{[F_2(0, 0)]_{\text{介电}}}{[F_2(0, K)]_{\text{金属-介电}}} \quad (6)$$

其中表示 N_ϕ 输出模式的总光子数。基于上述近似模型，可定量揭示不同金属体积分数下，金属-介电混合 O-PUF 相较于纯介电体系的独立 CRPs 数量与静态响应熵的演化规律。

3 结果分析

3.1 FDTD 仿真结果

基于扫描位移的模拟方法获得的短程散斑相关函数（图 2a, b）显示：x 偏振方向相关性显著强于 y 偏振，源于 x 偏振入射光的弹道光子主导效应；不同曲线随位移的去相关速率差异表明，金纳米粒子引入后形成新增散射和吸收位点，使金属-介电混合 O-PUF 散斑短程相关性增强，导致其支持的独立 CRPs 数量降低。

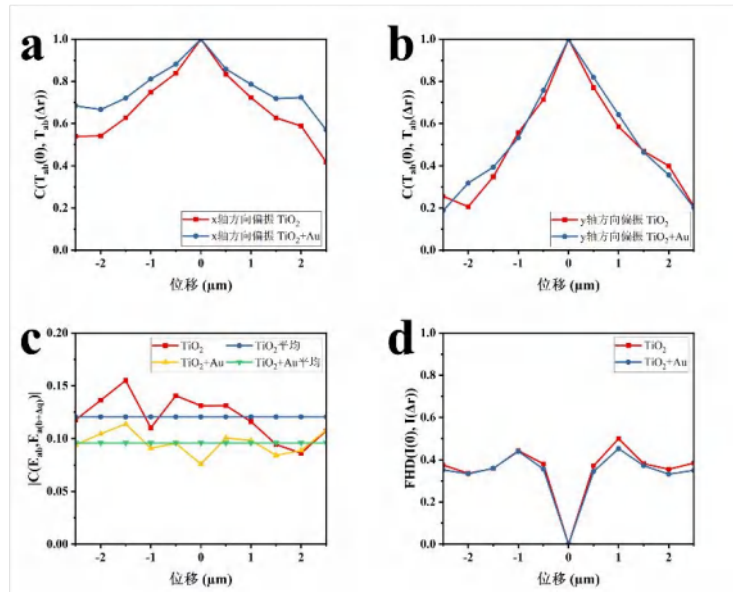


图 2 纯介电与金属-介电混合 O-PUF 三维模型 FDTD 仿真结果。（a, b）电场沿 x 轴偏振与沿 y 轴偏振的散斑短程相关性曲线；（c）模式动量差为 $\Delta q = 25\delta q$ 的散斑长程相关性曲线；

（d）位移后的散斑与初始位置的散斑之间的汉明距离变化

长程相关性模拟结果（图 2c）显示：与短程相关性趋势相反，金属-介电混合 O-PUF 的长程相关性强于纯介电体系。该现象表明金粒子引入降低了散斑长程相关性，致使静态响应熵升高。图 2d 进一步通过 Gabor 算法将散斑强度转化为密钥，其位移变化前后之间的分数汉明距离呈现关键特征：趋近 0 表征散斑高度相似，趋近 0.5 则标志完全不相关。混合 O-PUF 的去相关速率显著减缓（与图 2a, b 短程增强趋势一致），印证了散斑短程相关性增强与独立 CRPs 数量减少的强关联性。

3.2 理论近似结果

我们通过基于 Van Rossum M C W 等人含吸收系数修正公式计算的归一化

函数 C_1 和 F_2 (图 3a, b), 定性验证带吸收系数修正的散斑相关函数模型与 FDTD 仿真模型的吻合度。结果显示金属-介电混合 O-PUF 中金属组分等效为吸收系数的处理方式得到的结果与仿真结果一致, 其短程相关性增强而长程相关性降低的典型特征, 证实了该近似模型的有效性。

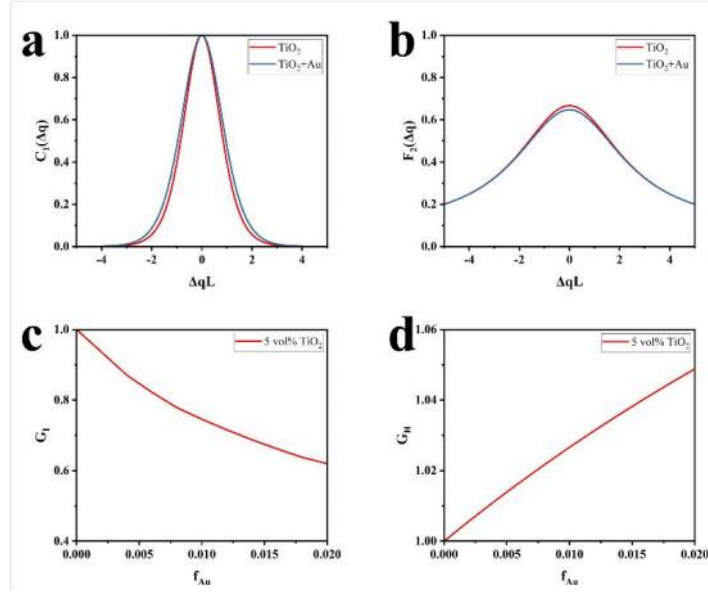


图 3 纯介电与金属-介电混合 O-PUF 散斑相关理论近似结果。(a, b) 归一化散斑短程相关函数和散斑长程相关函数; (c) 金属-介电混合 O-PUF 的独立 CRPs 数量随金属体积分数的增益曲线; (d) 金属-介电混合 O-PUF 的静态响应熵随金属体积分数的增益曲线。

进一步地, 基于散斑短程相关函数 C_1 半高全宽的变化量, 我们计算了金粒子体积分数梯度下 O-PUF 独立 CRPs 数量的演变规律 (图 3c)。结果表明金属体积分数增加时, 金属-介电混合 O-PUF 的独立 CRPs 数量持续下降。与此同时, 金粒子体积分数增大导致散斑长程相关性降低, 这反映了 O-PUF 中不同动量模式间关联通道的缩减。通过对带吸收系数修正的 F_2 进行化简, 最终获得静态响应熵增益的具体表达式如下:

$$G_H = \frac{2}{3F_2(0, K)} = \frac{8KL}{3} \frac{\cosh(2KL) - 1}{\sinh(2KL) + 2KL \cosh(2KL) - 4KL} \quad (7)$$

如图 3d 所示, 随金粒子体积分数的增加, 静态响应熵也呈现线性增大趋势, 响应的随机性更强。

最终研究表明, 金属-介电混合 O-PUF 中金属体积分数的增加会同时压缩独立 CRPs 数量并提升静态响应熵。我们建议针对不同应用场景对响应随机性或安全使用次数的侧重, 需在 O-PUF 设计中权衡优化金属体积分数的引入比例。

3 结论

针对现有散斑相关性物理机制限制了纯介电粒子体系 O-PUF 响应的随机性的问题, 本文提出了金属-介电混合 O-PUF 方案以突破纯介电粒子体系 O-PUF 静态响应熵的限制。基于 FDTD 仿真研究了在纯介电粒子体系中引入金属后散斑短程相关性和长程相关性的变化, 定性验证了金属-介电混合 O-PUF 可被含吸收系数修正的散斑相关函数模型有效近似。进一步通过理论近似模型研究表明, 在纯介电粒子体系 O-PUF 中引入金属粒子体积分数增加, 导致系统支持的独立激励响应数量逐渐减少, 同时静态响应熵逐渐增加, 随机性得以增强。这一特性使此类混合 O-PUF 在高安全性认证和高效随机数生成等场景中具有应用潜力。

参考文献

- [1] Klausen M, Zhang J, Stevens M M. Designing Physical Unclonable Functions From Optically Active Materials[J]. *Advanced Materials*, 2025: 2502059.
- [2] Rührmair U, Sehnke F, Sölter J, et al. Modeling attacks on physical unclonable functions[C]. *Proceedings of the 17th ACM conference on Computer and communications security*. 2010: 237-249.
- [3] Mahmoud A, Rührmair U, Majzoobi M, et al. Combined modeling and side channel attacks on strong PUFs[J]. *Cryptology ePrint Archive*, 2013.
- [4] Pappu R, Recht B, Taylor J, et al. Physical one-way functions[J]. *Science*, 2002, 297(5589): 2026-2030.
- [5] Pavanello F, O'Connor I, Rührmair U, et al. Recent advances in photonic physical unclonable functions[C]. *2021 IEEE European Test Symposium (ETS)*. IEEE, 2021: 1-10.
- [6] Osnabrugge G, Horstmeyer R, Papadopoulos I N, et al. Generalized optical memory effect[J]. *Optica*, 2017, 4(8): 886-892.
- [7] van Rossum M C W, Nieuwenhuizen T M. Multiple scattering of classical waves: microscopy, mesoscopy, and diffusion[J]. *Reviews of Modern Physics*, 1999, 71(1): 313.
- [8] Berkovits R, Feng S. Correlations in coherent multiple scattering[J]. *Physics Reports*, 1994, 238(3): 135-172.
- [9] Van Rossum M C W, Nieuwenhuizen T M. Influence of skin layers on speckle correlations of light transmitted through disordered media[J]. *Physics Letters A*, 1993, 177(6): 452-458.
- [10] Tuyls P, Škorić B, Stallinga S, et al. Information-theoretic security analysis of physical uncloneable functions[C]. *Financial Cryptography and Data Security: 9th International Conference, FC 2005, Roseau, The Commonwealth Of Dominica, February 28–March 3, 2005. Revised Papers 9*. Springer Berlin Heidelberg, 2005: 141-155.

Research Article

A Novel Automatic Pole Determination Method for Macromodeling Multi-Port Interconnect Systems

Zi-Xing Ye¹, Hong-Shun Yin¹, Wen-Yong Zhou², Da-Wei Wang¹, and Wen-Sheng Zhao¹

¹ School of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China

² Shanghai Huada Emphyrean Information Technology Company Ltd., Shanghai 201306, China

Corresponding author: Wen-Sheng Zhao; Email: wsh.zhao@gmail.com.

Abstract — With the increasing bandwidth and frequency, the geometric complexity of packaging, board-level, and interconnect systems are continuously increased. Direct full-wave electromagnetic simulation of these passive systems consumes enormous computational resources. To simplify the analysis, macromodeling the electromagnetic responses of passive systems has become common practice. Although vector fitting is an excellent macromodeling method, it typically requires users to manually specify the fitting order, which determines the accuracy. In this paper, a universal automatic pole determination method is proposed for vector fitting to achieve accurate macromodeling for single-input/multiple-output systems. This approach directly produces a rational function macromodel from the given EM response, and it performs robustly on both systems with noise and large-scale multipoint interconnect systems.

Keywords — Automatic pole determination, multi-port interconnect systems, reduced-order model, Macromodeling.

I. Introduction

In the analysis and design of high-speed, high-frequency electronic systems, accurate and efficient modeling and simulation of increasingly complex passive structures are essential for ensuring system-level design accuracy and reliability [1]. Behavioral-level macromodeling and characterization of large-scale passive circuits represents a highly practical and effective method [2], [3]. Macromodels are SPICE-compatible equivalent circuit models constructed from the frequency-domain responses of passive device and interconnect systems. Once built, they can be flexibly embedded into the overall electronic system for rapid transient analysis.

The key to behavioral macromodeling of passive devices and multi-port interconnect systems is to extract a state-space model from their electromagnetic (EM) frequency responses and convert it into a rational function form for system-level description. Subsequently, by mapping the parameters of the rational model to the corresponding circuit elements, the model can be transformed into an equivalent circuit for subsequent time-domain analysis. To solve this problem, vector fitting (VF) [4]-[6] is widely used in both industry and academia for its robustness and generality.

VF demonstrates excellent performance through its pole relocation module; however, its application typically requires the user to determine the number of fitting poles based on experience, preventing an end-to-end process that directly yields a rational function model from the given EM response of a passive system. Traditionally, there are two approaches for determining the number of fitting poles. One requires the user to manually specify the pole count based on prior experience or by adjusting it

to meet accuracy requirements. The other involves automatic estimation by the program, where a maximum number of poles is predefined, and the fitting is performed incrementally from a lower to a higher number of poles [7], [8]. When the fitting error meets the predefined accuracy threshold, the corresponding number of poles is identified as the final pole count. It is evident that this method becomes time-consuming when applied to large-scale multi-port systems.

There has also been a pole-collapsing column fitting (PC-VF) [9] proposed for multiple-input/multiple-output systems in the field of power electronics. Its pole determination method still relies on an empirically predefined order, and the absence of an iterative verification process makes the determination overly confident. Therefore, for the VF process of passive interconnect systems, a universal and fast pole identification method remains an open problem.

To overcome this issue, for the first time, a novel automatic pole determination strategy is proposed to further enhance the versatility of VF method. The VF-Based macromodeling overall workflow is shown in Fig.1. The main contributions of this work are summarized as follows:

- 1) A combined strategy of capacity space analysis and wavelet-based filtering is proposed to obtain a reasonable and reduced initial pole count.
- 2) A convolution-based dimensionality reduction method is proposed for iterative pole relocation, leveraging the structural similarity between the reduced and original multi-port networks. This approach significantly reduces computation time while ensuring accurate pole positioning.

It should be noted that the VF used in this paper is the

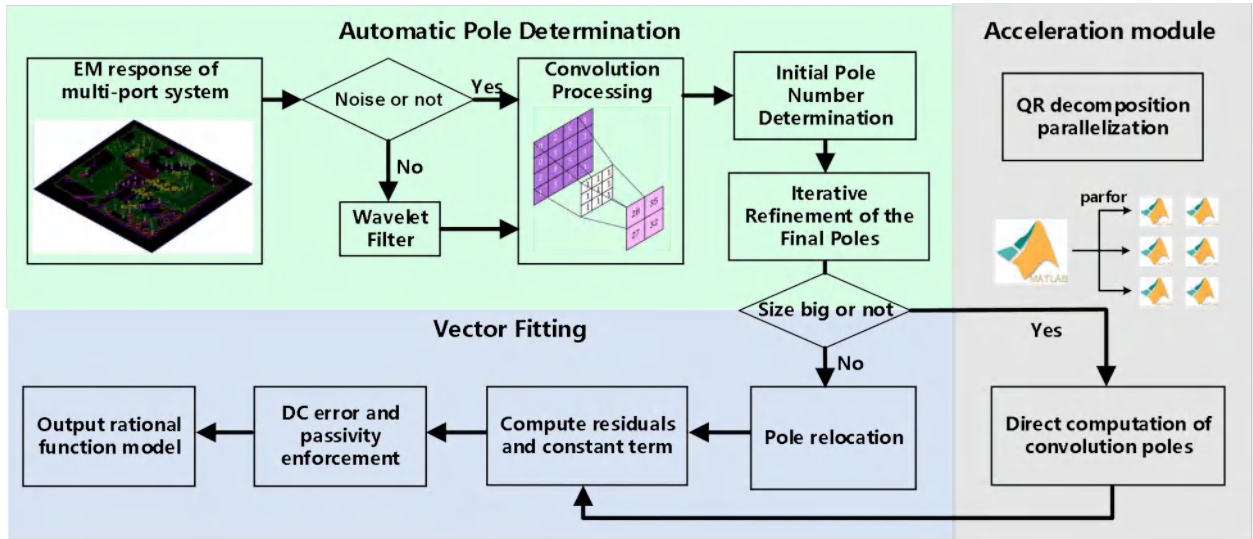


Figure 1 The VF-Based macromodeling overall workflow

standard version distributed in [10], which includes fast implementation [11] and adopts relaxed scaling function condition [12]. The DC error and passivity enforcement used here are detailed in [13].

II. Data Pre-Processing Procedure

To enhance the efficiency and accuracy of the automatic pole determination method, a new criterion for determining the poles and necessary data pre-processing procedure will be introduced as a preliminary step in this Section.

1. New Criterion for Determination of Poles

As this study aims to provide a general automatic pole determination scheme for VF, a comprehensive consideration of the multi-port system to be fitted is necessary. Here, a more practical criterion is proposed to guide the selection of the number of fitting poles. Specifically, the product of the number of the fitting poles and the fitting error is used as a metric to evaluate the reasonableness of the selected pole count.

The goal of VF is to approximate the original data using a rational function:

$$\hat{H}(s) \approx H(s) = \sum_{n=1}^N \frac{R_n}{s - P_n} + S_\infty \quad (1)$$

Here, P_n denotes the poles, R_n represents the residues, and n is the number of poles. Subsequently, an auxiliary function is constructed to enable a least-squares solution.

$$\sigma(s) = c_0 + \sum_{n=1}^N \frac{c_n}{s - q_n} \quad (2)$$

$$\sigma(s) \cdot \hat{H}(s) \approx r_0 + \sum_{n=1}^N \frac{r_n}{s - q_n} \quad (3)$$

Based on the obtained values of c_0 , c_n , and r_n , the auxiliary function can be represented in the following form to facilitate iterative pole relocation:

$$\sigma(s) = c_0 \frac{\prod_n (s - z_n)}{\prod_n (s - q_n)} \quad (4)$$

By continuously replacing q_n with z_n during the

iterations, the rational function gradually approximates the original data with increasing accuracy.

The core idea of VF lies in pole relocation.² Parameter sweeps show that increasing the number of poles yields diminishing returns in accuracy, suggesting an optimal pole count. This paper uses the product of pole count and fitting error (K) as a selection criterion—lower K balances accuracy and efficiency.

$$err = \frac{\sum_{p=1}^N \text{norm}(H_p - S_p)}{\sum_{p=1}^N \text{norm}(S_p)} \quad (5)$$

$$K = err \times order \quad (6)$$

where S_p represents the target frequency-domain response matrix at frequency f_p , while H_p denotes the S -parameter matrix computed using the rational function at frequency f_p . The operator $\text{norm}(\cdot)$ refers to the matrix 2-norm, and N represents the number of sampled frequency points.

During the validation process using EM simulation results from various systems, two complex scenarios pose significant challenges to rational function modeling:

- ✧ First, when modeling S -parameter curves obtained from real-world measurements, the presence of noise-induced spikes must be considered. These artifacts complicate the determination of the initial number of poles.
- ✧ Second, if the target system has a large number of ports and the simulated EM response curves contain a high density of sampling points, the computational resources required for iterative pole determination become extremely demanding.

2. Wavelet Filtering

To address the first challenge, this study proposes preprocessing both real and imaginary parts of the EM response separately with wavelet filtering. This enhances the algorithm's noise capacity space and improves the determination of the initial number of poles. Each S -parameter curve is decomposed into low-frequency approximations and multi-scale high-frequency details. Then noise thresholding is applied to the high-frequency detail coefficients—zeroing those below the noise

threshold to remove spikes. The denoised curve is then reconstructed via inverse wavelet transform. The specific implementation process of wavelet filtering follows the steps below:

Algorithm 1: Wavelet filtering method

Input: Number of ports: n ; S -parameter: $S_{ij}(i, j = 1, 2, 3, \dots, n)$

Output: Denoised S -parameters: \widehat{S}_{ij}

1. Extract real and imaginary Parts:

$$S_{ij} = \text{Re}(S_{ij}) + j * \text{Im}(S_{ij})$$

2. Perform wavelet decomposition

$$\text{Re}(S_{ij}) = A_{2re} + D_{2re} + D_{1re}$$

$$\text{Im}(S_{ij}) = A_{2re} + D_{2re} + D_{1re}$$

3. Filter for high-frequency components

$$\widehat{D}_i[k] = \begin{cases} \text{sign}(D_i[k]) \cdot (|D_i[k]| - T), & |D_i[k]| > T \\ 0, & |D_i[k]| \leq T \end{cases}$$

4. Reconstruct S -Parameters S -parameter:

$$\widehat{\text{Re}}(S_{ij}) = \widehat{A}_{2re} + \widehat{D}_{2re} + \widehat{D}_{1re}$$

$$\widehat{\text{Im}}(S_{ij}) = \widehat{A}_{2re} + \widehat{D}_{2re} + \widehat{D}_{1re}$$

$$\widehat{S}_{ij} = \widehat{\text{Re}}(S_{ij}) + j * \widehat{\text{Im}}(S_{ij})$$

3. Convolutional Dimensionality Reduction

For the second scenario, a convolutional compression approach is proposed to process the first two dimensions of multi-port EM response matrices. The conventional approach of performing iterative operations directly on large-scale matrices to determine the final poles proves time-consuming. To address this challenge, a convolution-based model order reduction technique is introduced to accelerate the computational process inspired by convolutional neural network [14], [15].

The subsequent analysis considers a multi-port system with N_c ports, N_s sampling points, and N predetermined poles, where the EM response matrix maintains dimensions $[N_c, N_c, N_s]$. Specifically, the square identity matrices of size $\sqrt{N_c}$ serves as convolution kernels, with stride lengths matching the kernel dimensions. This configuration enables effective compression of the matrix's first two dimensions through convolution operations. The final compressed EM response matrix exhibits reduced dimensions $[N'_c, N'_c, N_s]$, where N'_c denotes the compressed port dimension.

III. Automatic Pole Determination Method

The automatic pole determination method consists of two parts: initial pole number determination, and iterative refinement of the final poles, as shown in Fig. 1. In the first stage, the initial pole locations are determined in a rational and low-order manner, while in the second stage, iterative refinement is employed to further improve the accuracy of the pole determination. It is worth noting that the entire automatic pole determination process included the data pre-processing procedure serves as a plug-and-play preprocessing module for VF-based macromodeling as shown in Fig.1. It autonomously determines the number and initial values of poles without interfering with the vector fitting procedure itself, thereby ensuring the stability and robustness of the overall modeling framework.

1. Initial Pole Number Determination

In the VF process, the preliminary determination of pole quantities, which is equivalent to the fitting order, is primarily governed by the number of spectral peaks in the response curve to be fitted. This relationship is expressed as follows:

$$P_0 = 2n_{peak} \tag{7}$$

where n_{peak} denotes the number of spectral peaks in the target response curve, and P_0 represents the initially determined number of poles. For curves exhibiting resonant peaks, conjugate complex pole pairs are generated corresponding to individual peaks, resulting in $P_0 = 2n_{peak}$. To ensure the pole pairing for system stability, causality preservation, and numerical robustness, the total number of poles in VF is constrained to even integers.

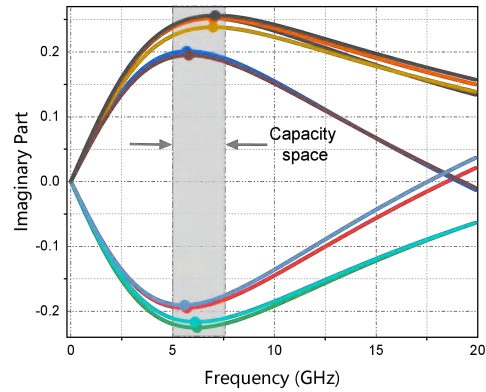


Figure 2 Capacity space strategy diagram

To minimize computational overhead during subsequent iterative refinement of pole positions, the initial pole determination prioritizes precision with minimal number. Notably, the fitting process operates on the EM response matrix $[N_c, N_c, N_s]$ characterizing the entire multi-port network, employing common poles across all matrix entries. Conventional approaches summing per-curve peak counts for matrix-level pole determination incur excessive pole quantities.

A capacity space strategy is proposed to suppress redundant identification of proximal poles across distinct curves. This method enforces single registration of geometrically adjacent poles within a threshold derived from the EM response matrix, even when multiple peaks exist on individual curves. For an EM response matrix of dimensions $[N_c, N_c, N_s]$, the capacity space threshold T_{cs} is defined as:

$$T_{cs} = \min((\log_{10} N_s + \log_{10} N_c), N_s/10) \tag{8}$$

The EM response matrix, being a complex-valued matrix, requires peak identification to be performed on its imaginary components. The pole determination process within the capacity space for a given dataset is illustrated in Fig. 2. In general, multiple S -parameter curves tend to exhibit several local extrema in close proximity within the capacity space. Selecting all these extrema as poles for modeling would result in a significant increase in unnecessary computational overhead. Therefore, a selection criterion is introduced in the capacity space strategy. This is, multiple poles appearing on distinct response curves within the designated capacity space are consolidated into a single representative pole, thereby reducing redundancy and improving efficiency.

When processing the experimental data, inherent

noise and glitches induce excessive spectral peaks, resulting in prohibitively large initial pole counts even with capacity space implementation. To address this, a noise assessment criterion is incorporated during preliminary pole determination. When detected pole number exceeds 50 times the port number N_c , the data is identified as noise-contaminated and subjected to the wavelet-based filtering algorithm detailed in Section II-B. The filtered data subsequently undergoes capacity space processing to derive optimized preliminary poles. Taking S_{14} -parameter with noisy S4P datasets as an example, Fig. 3(a) demonstrates the effect of wavelet filtering on the S -parameter curve. The K-value evolution versus pole number is also illustrated in Fig. 3(b) through a logarithmic plot for the S4P dataset. Comparative analysis reveals that wavelet filtering reduces initial pole counts from 642 (exhibiting elevated K-value deviations) to 64 poles, with corresponding K-values converging within the theoretical optimum range.

It is important to note that wavelet filtering is only applied during the Initial Pole Number Determination stage and is not used in subsequent processing steps. Therefore, it does not affect the fitting of the original data. Its sole purpose is to provide a more reasonable initial pole estimate, which improves the stability and efficiency of the iterative process.

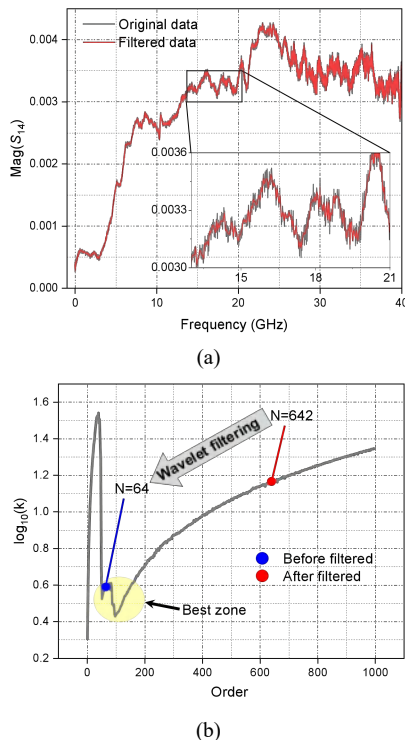


Figure 3 The noise suppression effect of wavelet filtering: (a) Magnitude of S_{14} -parameter and (b) Initial number of poles before and after filtering.

2. Iterative Refinement of the Final Poles

Following initial pole determination, a stepwise refinement protocol is implemented to approach optimal pole number. The iterative process initiates at 50% of the preliminary pole count, with adaptive step sizes governed by both initial pole distribution and frequency sampling density. Multiple VF iterations are executed, where the K-value convergence characteristics systematically guide the selection of both final pole number and initial

parameterization. The step of iterations is given as:

$$S_{iter} = 0.25 \cdot \min[2N_{initial}(iter - 1), (N_{max} - N_{min}) \cdot (iter - 1)] \quad (9)$$

where $iter$ represents the iterative algebra, $N_{initial}$ is the initial number of poles, and N_{max} and N_{min} denote the maximum and minimum values of the order. Here, $N_{initial}$, N_{max} , and N_{min} are set as 2, the number of frequency points, and half of the initial number of poles, respectively. A 2-pass strategy is used during iteration, that is, when the K-value does not decrease after two iterations, the pole at this time is taken as the final pole.

If the complete S -parameter matrix is directly fitted during iteration, fewer computing resources are required when the port scale is small. As the port scale increases, however, the time and computing resources required for fitting will increase exponentially due to the increase in the size of matrix to be solved during the fitting process. Therefore, it is a wise choice to use the reduced-dimensional data instead of the original complete matrix for iteration after reducing the data dimension and compressing the information.

Taking a 64-port interconnect system as an example, the convolution-based dimensionality reduction method proposed in Section II-C is applied to the target matrix. As shown in Fig. 4, it is evident that the trend and magnitude of the K-value, i.e., the product of fitting error and pole number, of the reduced matrix closely match those of the original full matrix. Therefore, the reduced matrix obtained through convolution can be effectively used for fast iterative determination of the optimal number of poles. The complete procedure of automatic determination of the pole number is also illustrated in Fig. 4. Initially, a preliminary pole number of 12 is estimated using the capacity-space strategy. Starting from half of this value, i.e., 6, a two-pass iteration strategy is used to search for the optimal pole number. Eventually, the determined pole number converges to 12, and the corresponding K-value is located near the global minimum, demonstrating the effectiveness and accuracy of the proposed automatic pole determination method.

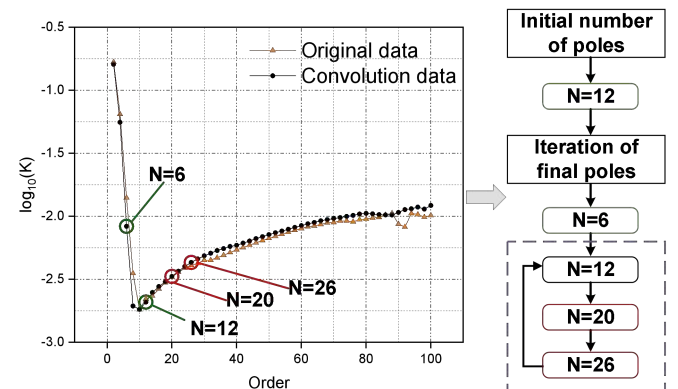


Figure 4 Automatic pole determination flowchart

After determining the poles using the automatic pole determination method, the VF algorithm is applied to perform pole relocation. By imposing DC point and passivity constraints during this process, a rational macromodel corresponding to the multi-port passive system can be accurately constructed.

IV. Validation and Discussion

In this section, the versatility and robustness of the proposed automatic pole determination method are demonstrated by macromodeling a variety of multi-port interconnect systems in different ranges. The cases used for verification, which are available as required, cover wide-band, multi-port, multi-sampling point, noisy simulated and measured data [16]. The code, written in MATLAB R2022b, can be embedded in Linux system and directly called through MATLAB runtime. Only the SNP file and the specified location of the EM response of the target multi-port passive system are required. The rational fraction model including poles, residuals, and constant terms is output to the specified location.

In the VF operation process, QR decomposition process takes the longest time [17], and it accounts for nearly half of the total time. In view of this, eight-core parallelization for QR decomposition operation is enabled during the algorithm operation to improve the efficiency. When the port is small, the operation is already quite fast, and the parallel operation starts. The extra time required to start the operation makes the operation slower. This algorithm uses the number of ports required for fitting as the starting condition for parallel operation.

When performing VF pole relocation iterations, unlike the original setting of iteration until the maximum number of times, 2-pass iterations are also used in this study. This is, the iteration is stopped when the error of two iterations no longer decreases. In the automatic pole determination module, the maximum number of pole relocations is set as 10. When the final rational fraction model is obtained by VF, the maximum number of pole relocations is set as 20.

1. Macromodel of Conventional System

The rational fraction macromodeling of the conventional system are shown in Table I. They are arranged from small to large according to the number of ports, and the number of poles required for fitting does not exceed 100. It is evident that the proposed method achieves high accuracy with few poles when modeling in such cases, and the longest time used is about 1.5 minutes.

Table 1 Macromodel of Conventional Systems

Ports	Poles	Time (s)	K	Error
7	16	8.873	1.6E-4	1.00266E-5
11	84	5.279	0.2301	0.002739
34	34	18.58	0.03193	0.000939
44	14	27.186	6.42E-4	4.58436E-5
64	24	91.529	0.002459	1.024E-4
229	34	90.03	0.059665	0.0018393

2. Macromodel of Noisy System

For a four-port noisy system with 4000 sampled points, the fitting results with and without wavelet filtering are compared in Table II. With the implementation of filtering module, the error slightly increases but remains within the acceptable threshold, and the K-value is significantly reduced by 50%. In addition, the computational time is dramatically decreased. The

S-parameters of the noisy system are shown in Fig. 5

Table 2 Macromodel of Noisy System

	Poles	Time (s)	K	Error
w/o wavelet filtering	320	50.914	5.3787	0.0168
w/ wavelet filtering	124	20.534	2.326	0.0187

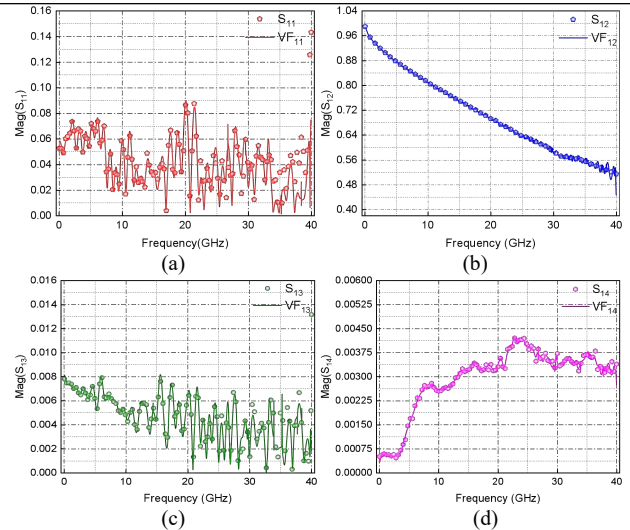


Figure 5 Magnitudes of (a) S11-, (b) S12-, (c) S13-, and (d) S14-parameters of the noisy system.

3. Macromodel of Large-Scale System

For large-scale system, where the product of the port number and the number of sampling points is greater than 200,000 (set according to experience), the pole relocation process needs to calculate $M = N_c \cdot (N_c + 1)/2$ times of QR decomposition operations of a matrix of size $[2N_s + 1, N + 1]$. If a pole relocation is performed on a multi-port system with 384 ports and 601 sampling points and the number of poles is 100, it is necessary to calculate 73728 times of QR decomposition of a matrix of size $[1203, 101]$, which undoubtedly leads to huge consumption of computing resources. Based on the consistency between the system after convolution dimensionality reduction and the original system in the pole-K value logarithmic graph which is shown in Fig.4, the poles determined by the pole relocation of the dimensionality reduction system are directly used to calculate the rational fraction model for ultra-large-scale port systems, thereby greatly improving the computational efficiency while ensuring accuracy. The results for two large-scale port systems are summarized in Table III. It is evident that compared with traditional method which needs two-time pole relocation operation on convolutional data and original data respectively, the convolutional method can achieve higher accuracy with a five to six times speed improvement.

Table 3 Macromodel of Large-Scale Systems

	Ports	Poles	Time (s)	K
Traditional	384	225	3830.11	2.1115
	218	200	4405.23	3.5E-2
Convolutional	384	300	606.31	0.8498
	218	492	909.31	1.5E-3

V. Conclusions

In this paper, a novel and general automatic pole determination method for passive macromodeling was proposed to directly transform the EM response of a passive system into rational function model. It combined capacity space evaluation, wavelet filtering, and convolution-based model order reduction. The proposed approach demonstrated robust and versatile performance across various types of systems, including noisy systems and large-scale multi-port systems, while significantly improving computational efficiency. Moreover, the proposed method was designed as a plug-and-play module based on VF, facilitating the integration of any more advanced algorithms to replace any part of the process, indicating substantial potential for further enhancement.

References

- [1] Z. Z. Peng and D. L. Su, "Analytical models of passive linear structures in printed circuit boards," *Chinese Journal of Electronics*, vol. 30, no. 2, pp. 275–281, 2021.
- [2] Y. Q. Xiao, S. Grivet-Talocia, P. Manfredi, and R. Khazaka, "A novel framework for parametric Loewner matrix interpolation," *IEEE Transactions on Component, Packaging and Manufacturing Technology*, vol. 9, no. 12, pp. 2404–2417, Dec. 2019.
- [3] Z. Du, N. Dong, and Y.-Z. Xie, "Behavioral modeling method of macromodels for interconnected systems with frequency characteristics and nonlinear termination networks," *IEEE Trans. Circuits Syst. II Exp. Briefs*, vol. 70, no. 9, pp. 3579–3583, Sep. 2023.
- [4] D. Deschrijver, B. Haegeman, and T. Dhaene, "Orthonormal vector fitting: A robust macromodeling tool for rational approximation of frequency domain responses," *IEEE Trans. Advanced. Packaging*, vol. 30, no. 2, pp. 216–225, May 2007.
- [5] A. Carlucci, T. Bradde, and S. Grivet-Talocia, "Addressing load sensitivity of rational macromodels," *IEEE Transactions on Component, Packaging and Manufacturing Technology*, vol. 13, no. 10, pp. 1591–1602, Jul. 2023.
- [6] S. B. Olivadese and S. Grivet-Talocia, "Compressed passive macromodeling," *IEEE Transactions on Component, Packaging and Manufacturing Technology*, vol. 2, no. 8, pp. 1378–1388, Aug. 2012.
- [7] T. M. Campello, S. L. Varricchio, G. N. Taranto, and A. Ramirez, "Enhancements in vector fitting implementation by using stopping criterion, frequency partitioning and model order reduction," *International Journal of Electrical Power & Energy Systems*, vol. 120, p. 105905, Sep. 2020.
- [8] S. Grivet-Talocia and M. Bandinu, "Improving the convergence of vector fitting for equivalent circuit extraction from noisy frequency responses," *IEEE Transactions on Electromagnetic Compatibility*, vol. 48, no. 1, pp. 104–120, Feb. 2006.
- [9] A. M. Smith, S. D'Arco, J. A. Suul, and B. Gustavsen, "Improved pole placement and compaction of MIMO vector fitting applied to system identification," *IEEE Transactions on Power Delivery*, vol. 39, no. 2, pp. 1259–1270, Apr. 2024.
- [10] Matrix fitting toolbox, [Available Online]: <https://www.sintef.no/projectweb/vectorfitting/downloads/matrix-fitting-toolbox> [Accessed 20 October, 2024].
- [11] D. Deschrijver, M. Mrozowski, T. Dhaene, and D. De Zutter, "Macro-modeling of multiport systems using a fast implementation of the vector fitting method," *IEEE Microwave and Wireless Components Letters*, vol. 18, no. 6, pp. 383–385, Jun. 2008.
- [12] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *IEEE Transactions on Power Delivery*, vol. 14, no. 3, pp. 1052–1061, Jul. 1999.
- [13] Z. Ye, "PMM: A Matlab toolbox for passive macromodeling in RF/mm-wave circuit design," in *Proceedings of IEEE 10th International Conference on ASIC*, Shenzhen, China, pp. 1–4, 2013.
- [14] J. Y. Gan, K. Y. Jiang, H. Y. Tan, G. H. HE, "Facial Beauty Prediction Based on Lighted Deep Convolution Neural Network with Feature Extraction Strengthened," *Chinese Journal of Electronics*, vol. 29, no. 2 pp. 312–321, 2020.
- [15] A. Krizhevsky et al. "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2012.
- [16] 2024 China Postgraduate IC Innovation Competition (EDA Elite Challenge Contest). [Available online]: <http://edachallenge.cn> [Accessed 10 October, 2024].
- [17] A. China and S. Grivet-Talocia, "On the parallelization of vector fitting algorithms," *IEEE Transactions on Component, Packaging and Manufacturing Technology*, vol. 1, no. 11, pp. 1761–1773, Nov. 2011.

四种去嵌入方法在氮化镓高电子迁移率晶体管器件建模中的应用与比较

白晶¹, 张傲^{2,3*}, 高建军^{1*}

(1.华东师范大学, 物理与电子科学学院, 上海, 200241;

2.南通大学, 微电子学院, 江苏, 南通, 226019;

3.新加坡南洋理工大学 电气与电子工程学院, 新加坡 639798)

摘要: 为了描述有源器件的射频行为, 精确的片上表征是至关重要的。本文比较了四种去嵌入方法, 并研究了它们对氮化镓高电子迁移率晶体管 (GaN HEMT) 建模的影响。通过三维电磁 (EM) 仿真器获取无源测试结构的 S 参数。在此基础上, 建立了小信号等效电路模型, 以比较这四种去嵌入方法在 1 到 50 GHz 频率范围内的影响。此外, 还研究了这些去嵌入方法对计算截止频率 f_c 和最大振荡频率 f_{max} 的影响。

关键词: 去嵌入; 电磁 (EM) 仿真; 氮化镓 (GaN); 高电子迁移率晶体管 (HEMT); 小信号模型

1 引言

氮化镓高电子迁移率晶体管 (GaN HEMT) 因其高功率密度、高击穿电压、优异的高频性能以及极小的开关损耗而受到广泛关注。精确的测量是器件建模的基础。在典型的微波测量中, 待测器件被嵌入测试夹具中, 以提供测量探针所需的接口。然而, 这些夹具不可避免地引入了不期望的寄生效应, 因此去嵌入是必不可少的步骤^[1]。

在过去几十年中, 众多去嵌入方法被相继提出。基于集总等效电路模型的去嵌入方法通过测量几个特定测试结构来提取探针焊盘、输入和输出互连线的寄生元件。然后, 通过在导纳和阻抗中减去寄生元件的影响来获得待测器件的本征参数^[2, 3]。基于级联矩阵计算的去嵌入方法考虑了分布效应^[4, 5]。在这种方法中, 过渡部分被视为与器件级联的双端口网络, 其影响通过矩阵计算从测量数据中移除。一般来说, 以上去嵌入方法都需要制造相应的测试结构, 这导致了晶圆面积的消耗。因此, 商业化三维电磁 (EM) 仿真器也被应用于去嵌入^[6, 7]。在本文中, 比较了四种去嵌入方法。通过三维电磁仿真获得无源测试结构的 S 参数。为了研究去嵌入方法的影响, 基于这四种去嵌入方法分别建立了 GaN HEM 器件的小信号模型, 并进行了对比。此外, 还研究了这些去嵌入方法对计算器件截止频率 f_c 和最大振荡频率 f_{max} 的影响。

2 去嵌入方法

2.1 Open-Short 方法 1

开路-短路（Open-Short）去嵌入方法是目前应用最广泛的方法之一^[2]。该方法基于假设模型中的寄生元件是并联-串联结构，需要额外的开路、短路测试结构。无源测试结构中的焊盘和金属互连的布局与被测件相同。晶体管的实际 Y 参数可由公式(1)获得：

$$Y_{\text{trans}} = ((Y_{\text{DUT}} - Y_{\text{open}})^{-1} - (Y_{\text{short}} - Y_{\text{open}})^{-1})^{-1} \quad (1)$$

其中的 Y_{DUT} 为晶体管的实测 Y 参数矩阵， Y_{open} 和 Y_{short} 分别为开路结构和短路结构的 Y 参数矩阵。

2.2 Open-Short 方法 2

Open-Short 方法 2 需要三个片上测试结构：两个开路结构和一个短路结构。与方法 1 相比，额外的开路结构中的微带线长度与过渡结构的长度相同。从短路结构的导纳中去除并联导纳时，采用只有焊盘的开路结构。晶体管的实际 Y 参数可从公式(2)中获得：

$$Y_{\text{trans}} = ((Y_{\text{DUT}} - Y_{\text{open2}})^{-1} - (Y_{\text{short}} - Y_{\text{open1}})^{-1})^{-1} \quad (2)$$

其中的 Y_{open1} 为只带焊盘的开路结构的 Y 参数矩阵， Y_{open2} 为带有长微带线结构的开路结构的 Y 参数矩阵。

2.3 TRL 方法

直通-反射-延迟线（TRL）去嵌入方法使用三种去嵌入结构，即直通、反射和延迟线^[4]。如图 3 所示，该方法将 DUT 结构表示为三个双端口网络的级联。在这种情况下，整体的传输矩阵可以表示为三个双端口网络传输矩阵的级联：

$$[T_{\text{M}}^{\text{DUT}}] = [T^{\text{A}}][T^{\text{DUT}}][T^{\text{B}}] \quad (3)$$

其中， $[T^{\text{A}}]$ 和 $[T^{\text{B}}]$ 为左右过渡结构的传输矩阵， $[T^{\text{DUT}}]$ 为晶体管的实际传输矩阵。

2.4 Half-Thru 方法

获得过渡结构 S 参数的一种简单方法是测量如图 4 所示的直通结构^[5]。假定左、右过渡结构是相同的，则相应过渡结构的 S 参数可表示为：

$$S_{11}^{\text{In,Out}} = S_{22}^{\text{In,Out}} = \frac{(S_{11}^{\text{Thru}} + S_{22}^{\text{Thru}})}{(2 + S_{21}^{\text{Thru}} + S_{12}^{\text{Thru}})} \quad (4)$$

$$S_{12}^{\text{In,Out}} = S_{21}^{\text{In,Out}} = \sqrt{\frac{1}{2}(S_{12}^{\text{Thru}} + S_{21}^{\text{Thru}})(1 - S_{11}^{\text{In,Out}^2})} \quad (5)$$

其中， S^{Thru} 为测得的直通结构的 S 参数。半直通（Half-Thru）去嵌入方法只使用一个与过渡结构长度相同的测试结构进行 EM 仿真。各种方法对应的被测器件电路模型及无源测试器件的三维结构如图 1 所示。

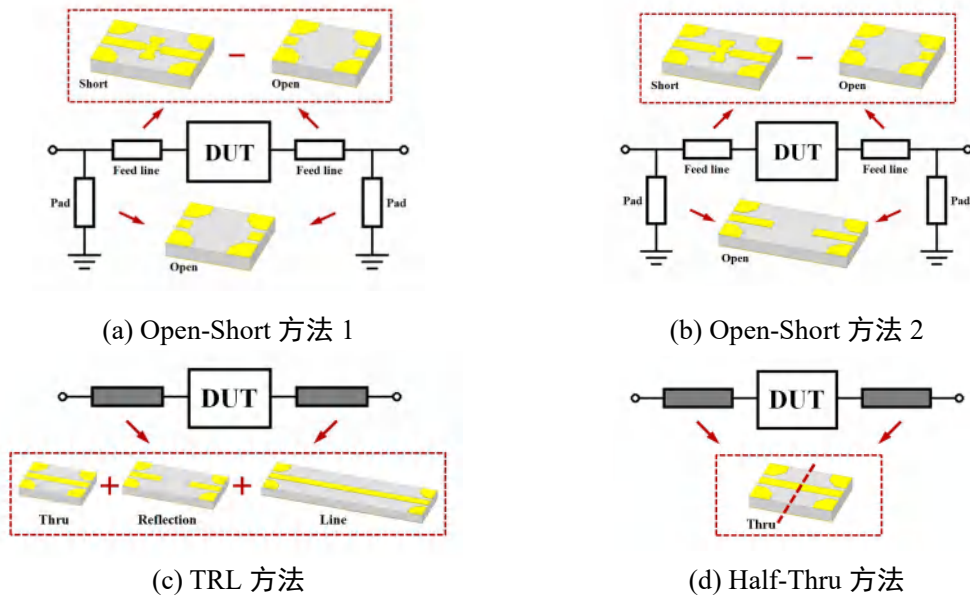


图 1 不同去嵌方法的被测器件电路模型及测试结构三维模型

3 小信号模型

使用 PNA 网络分析仪 (Keysight N5227B) 在 1 至 50 GHz 频段内完成了 S 参数的测量。通过三维电磁仿真得到了各无源测试结构的 S 参数, 有效地节省了晶圆面积。首先, 在片上测量之前使用陶瓷阻抗标准衬底 (ISS) 进行校准, 将校准参考平面移动至探针尖端。然后, 结合去嵌入方法, 将参考平面平移至待测器件根部。

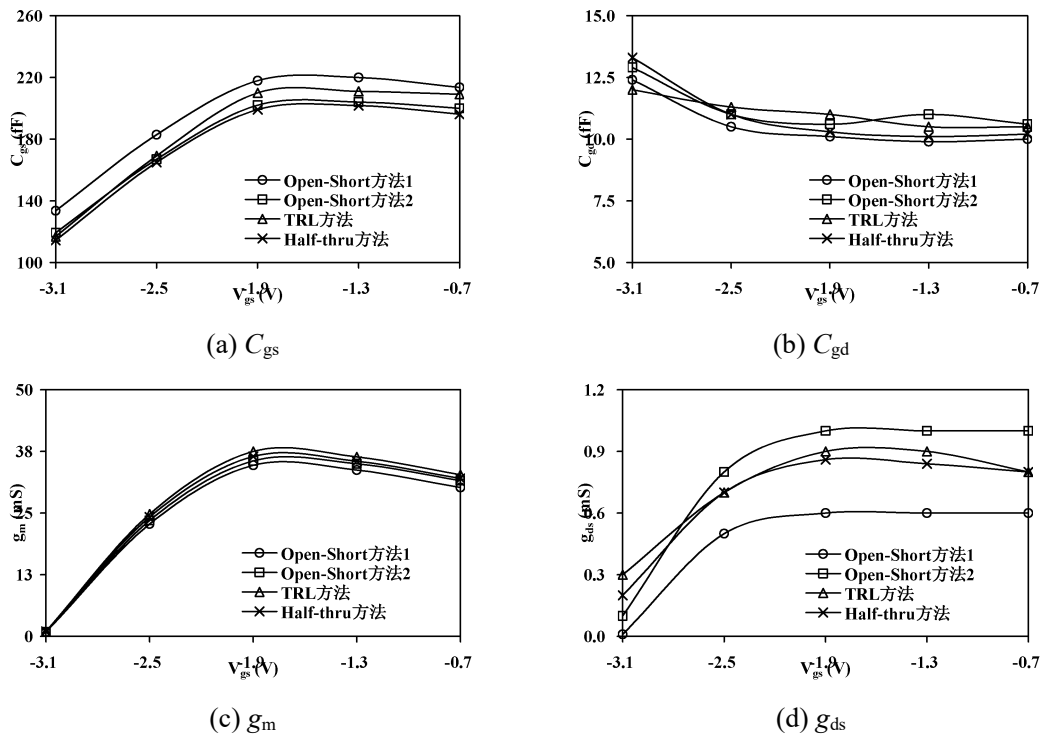


图 2 不同去嵌方法去嵌后的本征参数提取 ($V_{ds}=20V$)

使用网络分析仪测量了被测器件 $2 \times 50 \mu m$ GaN HEMT 的 S 参数, 同时, 通

过 EM 仿真得到了各测试结构的 S 参数。然后, 采用不同的去嵌入方法得到晶体管的实际 S 参数。去嵌入后, 在截止条件下提取被测件的寄生电阻, 得到的值分别为 $R_g = 2.9 \text{ Ohm}$ 、 $R_d = 4.1 \text{ Ohm}$ 和 $R_s = 0.9 \text{ Ohm}$ 。利用传统的提取公式^[8], 直接从本征 Y 参数中确定不同偏置条件下的本征元件值。为了研究不同的去嵌入方法对建模的影响, 在 $V_{ds} = 20 \text{ V}$ 的条件下, 不同 V_{gs} 电压下提取的本征参数如图 2 所示, TRL 和 Half-Thru 方法得到的本征参数值相对接近。

将模拟结果与去嵌入后的测量数据进行对比, 其小信号电路模型的精度对比如图 3 所示。对于 Open-Short 方法 1, S_{11} 和 S_{12} 的误差最小, 但 S_{22} 的误差相对较高。Open-Short 方法 2 有效降低了 S_{22} 的误差, 但牺牲了 S_{11} 的精度。TRL 方法和 Half-Thru 方法得到的结果较为接近。

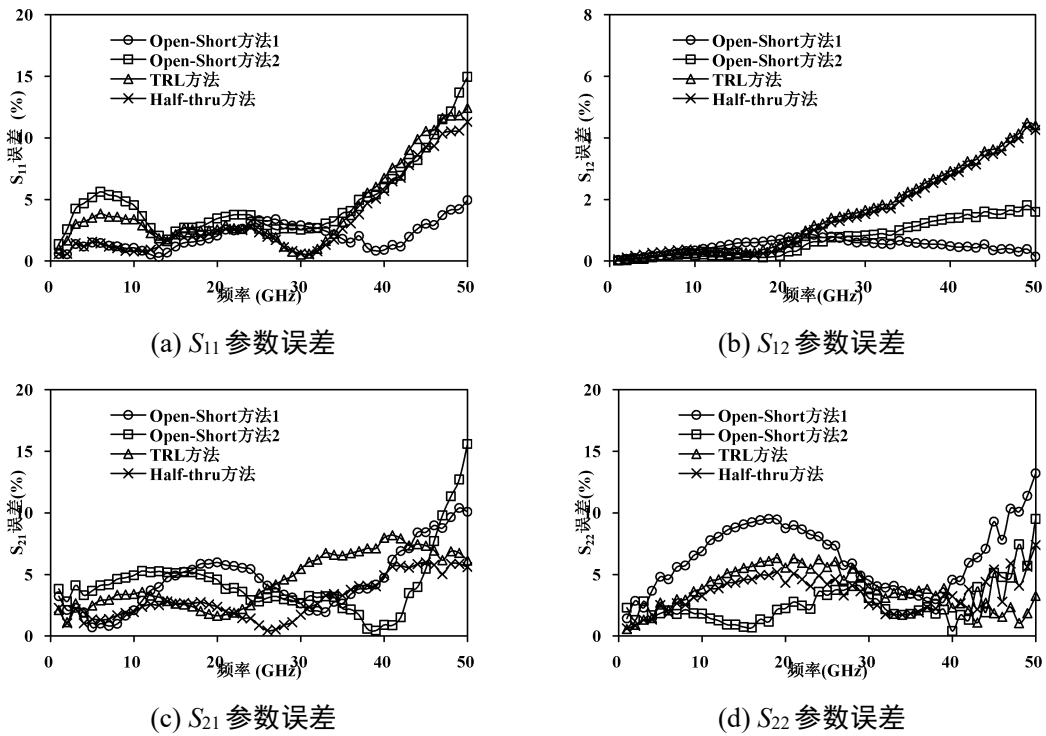


图 3 不同去嵌方法去嵌后的 S 参数误差

截止频率 f_T 和最大振荡频率 f_{max} 是毫米波和亚毫米波电路设计中有源器件的重要因素。其表达式为:

$$f_T \approx \frac{g_m}{2\pi(C_{gs} + C_{gd})} \quad (6)$$

$$f_{max} \approx \frac{f_T}{2\sqrt{(R_i + R_s + R_g)/R_{ds} + (2\pi f_T)R_g C_{gd}}} \quad (7)$$

为了清楚地显示不同的去嵌入方法对器件的影响, 根据不同的去嵌入结果计算器件的 f_T 和 f_{max} , 列于表 1 和表 2。通过比较, Open-Short 方法 1 的去嵌入结果计算出的 f_T 值相对较低, 而使用同一方法的去嵌入结果计算出的 f_{max} 值相对较高。相比之下, 从其他三种方法得到的结果表现出更高的一致性。Open-Short 方法 1

假设所有并联寄生电容均位于焊盘上。然而，实际上，这些电容也部分分布在焊盘与晶体管之间的互连线上。Open-Short 方法 2 考虑了这种分布效应，但代价是需要额外的开路结构。无论是 TRL 还是 Half-Thru 方法，均基于级联网络去嵌入技术，同样考虑了分布效应。

表 1 不同去嵌入方法提取的器件 f_T
(单位: GHz, $V_{ds} = 20$ V)

V_{gs}	方法 1	方法 2	方法 3	方法 4
-2.5V	18.8	21.0	21.9	21.9
-1.9 V	24.2	26.7	27.0	27.7
-1.3 V	23.3	25.9	26.2	26.7
-0.7 V	21.5	23.9	23.8	24.7

表 2 不同去嵌入方法提取的器件 f_{max}
(单位: GHz, $V_{ds} = 20$ V)

V_{gs}	方法 1	方法 2	方法 3	方法 4
-2.5V	123.0	112.3	114.4	118.1
-1.9 V	139.4	128.7	123.8	132.4
-1.3 V	133.9	119.9	120.2	127.9
-0.7 V	123.9	115.0	113.2	119.5

4 结论

本文研究了四种不同的去嵌入方法对 GaN HEMT 器件建模的影响。在 1 到 50 GHz 频率范围内，使用网络分析仪测量了栅宽为 $2 \times 50 \mu\text{m}$ 的 GaN HEMT 的 S 参数，同时，通过 EM 仿真得到了各测试结构的 S 参数。通过采用不同的去嵌入方法对测试数据进行处理，分别提取了模型参数，对比了 S 参数，并且进一步比较了不同的去嵌结果对截止频率 f_T 和最大振荡频率 f_{max} 计算结果的影响。

参考文献

- [1] Loo X S, Yeo K S, Chew K W J, et al. A new millimeter-wave fixture deembedding method based on generalized cascade network model[J]. IEEE electron device letters, 2013, 34(3): 447-449.
- [2] Koolen M, Geelen J A M, Versleijen M. An improved de-embedding technique for on-wafer high-frequency characterization[C]//Proc. Bipolar Circuits Technol. Meeting. 1991: 188-191..
- [3] Costa D, Liu W U, Harris J S. Direct extraction of the AlGaAs/GaAs heterojunction bipolar transistor small-signal equivalent circuit[J]. IEEE Transactions on electron Devices, 2002, 38(9): 2018-2024.
- [4] Engen G F, Hoer C A. Thru-reflect-line: An improved technique for calibrating the dual six-port automatic network analyzer[J]. IEEE transactions on microwave theory and techniques, 1979, 27(12): 987-993.
- [5] Zhang B, Xiong Y Z, Wang L, et al. On the de-embedding issue of millimeter-wave and sub-millimeter-wave measurement and circuit design[J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2012, 2(8): 1361-1369.
- [6] Kikuchi K, Raffo A, Vadalà V, et al. A New Calibration Technique of Electromagnetic Simulators for Accurate Analyses of Microwave Components on Epitaxial Wafers[J]. IEEE Access, 2024.
- [7] Bousnina S, Falt C, Mandeville P, et al. An accurate on-wafer deembedding technique with application to HBT devices characterization[J]. IEEE Transactions on Microwave Theory and Techniques, 2002, 50(2): 420-424.
- [8] 高建军. 场效应晶体管射频微波建模技术[M]. 电子工业出版社, 2007.

基于快速时域仿真和迁移学习的信号完整性建模方法

虞梦怡, 毛昌盛, 王大伟, 赵文生

杭州电子科技大学

摘要: 本文提出一种融合快速时域仿真与迁移学习的建模方法: 其中源域采用电路仿真与多边沿响应方法生成瞬态波形及最坏眼图, 目标域基于电磁仿真提取 S 参数并进行瞬态仿真, 分别构建几何参数映射眼高的深度神经网络模型与输入映射输出波形的时间卷积网络模型, 通过多种迁移策略重用知识, 显著降低数据需求与训练成本, 能够有效解决三维集成微系统中信号完整性建模方法面临的电磁与瞬态仿真开销大、数据依赖强的小样本问题。

关键词: 信号完整性; 迁移学习; 时域仿真; 神经网络

1 引言

随着集成电路工作频率和集成度的不断提升, 其特征尺寸已逼近材料物理极限。在半导体与微电子领域, 基于先进封装的三维集成微系统已经成为满足诸如更高带宽、功能密度和更小占位等芯片性能和尺寸要求的可行解决方案。面向三维集成微系统的高速链路在追求高带宽密度与低延迟传输的同时, 也面临电磁干扰、串扰、时序失配等更严峻的信号完整性(Signal Integrity, SI)挑战[1]。作为高速信号质量评估的核心工具, 眼图通过可视化呈现符号间干扰、幅度失真和抖动等问题, 帮助工程师进行诊断与优化。

在此背景下, 为了实现高速链路的 SI 性能预测、优化与设计反馈闭环, 迫切需要高效且精准的建模方法, 以辅助工程决策。典型的 SI 建模任务主要包括两类: 一是从几何结构参数预测关键 SI 指标(如眼高); 二是建立输入激励与输出波形的映射关系, 揭示信号在复杂结构中的传输行为。这些任务常常具有高维输入、强非线性和剧烈响应变化特征, 令传统建模方法难以胜任。

近年来, 机器学习技术(Machine Learning, ML)为 SI 分析提供了新的突破性思路[2]-[4]。传统的人工神经网络(Artificial Neural Network, ANN)因其优秀的非线性建模能力被广泛应用于代理建模[5]-[9]、设计优化[10][11]和不确定性量化[12]等多个环节。然而, ANN 模型高度依赖训练数据, 需要大量高质量数据来训练模型, 以确保其 SI 特性的可靠模拟。另一方面, ANN 方法存在明显的扩展性局限: 目标对象结构、环境发生细微修改, 模型就需重新开始采样新的大量数据并重新训练模型, 使其缺乏灵活性与效率。同时各个神经网络模型孤立运行, 针对不同任务对象独立训练, 不考虑任务间的相关性, 即使任务间具有高度相似性, 也难以复用已有共性知识。

SI 建模数据的生成主要依赖两类仿真：一是基于全波电磁(Electromagnetic, EM)仿真，通过细致网格求解偏微分方程，准确捕捉寄生参数和耦合效应，但计算量大、耗时长；二是基于时域瞬态仿真，能完整还原波形细节和高频效应，但仿真时间随比特序列长度线性增长，对数百万甚至数十亿比特的响应分析使效率大幅下降。实际流程中，EM 仿真用于提取高保真参数，瞬态仿真则基于这些参数评估波形与眼图。

为降低仿真开销，一些研究采用电路仿真或等效电路模型替代 EM 仿真，以传输线模型简化复杂结构，提升效率；为应对瞬态仿真在处理长比特序列时的计算负担，研究者提出了单比特响应(Single Bit Response, SBR)、双边沿响应(Double Edge Response, DER)和多边沿响应(Multiple Edge Response, MER)等快速时域仿真方法[13]-[17]，仅对脉冲或边沿建模，再通过时移叠加重构完整响应，实现最坏眼图和统计眼图的快速求解。然而，电路仿真难以充分模拟高频耦合效应，难以还原更复杂的器件结构；快速时域方法在高非线性场景下的精度仍逊于瞬态仿真。因此，如何在电路仿真和快速时域仿真方法所带来的高效性与电磁仿真和瞬态仿真所提供的高精度之间寻求有效平衡和协同建模策略，仍是 SI 分析中亟待解决的重要研究问题。

基于电路仿真的快速时域仿真虽在精度上有所局限，但其计算高效、实现简便，且具备物理一致性，能保留模型结构特征，可成为实际仿真的有力补充。现有研究中有将模型先验知识注入到神经网络建模任务的方法，能够有助于弥补数据不足对特征提取能力的影响，解决“小样本建模”的现实困境。例如基于知识的神经网络(Knowledge-Based Neural Network, KBNN)[18]和空间映射(Space Mapping)[19]方法，通过将物理模型或经验公式融入网络架构或构建精细到粗糙模型的映射函数。但这些方法往往依赖手工设计的融合策略或特定经验。快速时域和瞬态仿真间的非线性多维差异，使得上述策略难以直接应用。

基于此背景，迁移学习(Transfer Learning, TL)作为一种通用建模范式，提供了自动化知识重用的新思路[20]。TL 可将源任务中学到的参数和特征迁移到目标任务，降低对新数据的依赖，提升训练效率。已有研究显示，TL 在频率响应预测[21]、微波器件建模[22]和多精度仿真建模[23][24]等领域具有良好应用前景，但大多方法只限于特定结构或单一任务的局部迁移。

本文提出了一种结合快速时域仿真与迁移学习的通用 SI 建模方法：源域模型采用电路仿真与 MER 方法高效生成最坏眼图，目标域模型基于 EM 提取 S 参数并进行瞬态仿真以获得高精度响应；二者共享相同网络结构，通过多种迁移策略实现知识重用，显著降低数据需求和训练成本，并在多任务和不同样本量条件下验证了方法的鲁棒性与扩展性。本研究为构建满足不同仿真精度需求与建模任务的统一 SI 建模流程提供了可行思路。本文的创新性如下：

(1) 首次创新性地提出了从快速时域仿真方法到瞬态仿真的跨域迁移学习框架，

覆盖从电路仿真到电磁仿真的建模迁移路径。源域使用快速时域仿真技术并利用电路仿真中的传输线模型构建互连结构，高效提取 SI 特性，目标域仅需有限次高精度电磁仿真和瞬态仿真，实现数值计算效率和高精度仿真的有机结合。

(2) 所提出的方法不依赖于特定的电路拓扑结构，且不局限于固定回归任务，适用于任何现代非线性高速链路 SI 代理模型的建立，有效规避了传统神经网络建模方法和迁移学习方法的局限性。

(3) 针对不同的任务如波形映射任务或眼高预测任务分别设计专用神经网络结构，系统化比较不同迁移策略对训练损失、数据依赖和模型鲁棒性的影响；结果表明框架在多种复杂任务中均表现出优异的可扩展性与稳健性。

全文其余部分组织如下：第二章介绍了快速时域仿真技术，重点阐述如何利用 MER 技术构建瞬态波形以及得到最坏眼图。第三章具体介绍本文提出的迁移学习策略，并针对波形预测和眼高映射两类任务，分别建立神经网络以应对不同的输入输出需求；第四章从数据规模、训练时间等维度设置对比实验，证明所提出方法的性能，并从两个不同的迁移任务中探究迁移策略对模型性能的影响，验证方法的有效性和鲁棒性；第五章总结研究成果。

2 快速时域仿真技术

2.1 传统眼图与最坏眼图

眼图是一种将数字信号在单个比特周期内的多个符号电平及其边沿转换叠加而成的图形表示。眼图的开口量化了信号失真程度，直接反映了符号间串扰 (Inter Symbol Interference, ISI) 的严重性和噪声水平。由常规眼图演变而来的最坏情况眼图，则描绘了平均眼图的最内轮廓曲线，代表在最坏情况下电路模型所能容忍的最小时序和电压裕量。如果最坏情况眼图满足设计裕量，那么平均眼图的裕量也必然符合规格。

传统上，工程师使用长伪随机二进制序列 (Pseudo-Random Binary Sequence, PRBS) 作为链路的输入激励源，并在输出端捕获失真后的模拟波形。由于 PRBS 码的频域特征和白噪声频域特征十分相似，这种毫无规律的特点可以用来模拟真实环境中的输入信号。接收机把每个比特周期内的连续波形压缩到一个单位间隔内，叠加而形成平均眼图。但此方法需极长的 PRBS 序列才能覆盖足够的码型变化，需要耗费大量资源。

2.2 瞬态波形构建

SBR 假设系统线性时不变，但在高速链路中常因驱动器非线性或上下边沿不一致而失效；DER 通过分别建模上升沿和下降沿来捕捉它们的非对称性；MER 则在此基础上引入前序多比特的模式响应，以考虑非线性影响[14]。例如，考虑边沿跳变前 1 个比特时，输入序列“101”和“001”都包含一个“01”跳变，但由于前导位不同，其实际边沿响应也有所差异[25]。

利用比特响应构造考虑前 2 个前导位的边沿响应可以描述为:

$$f^{11}(t) = y^{110}(t) \quad (1)$$

$$f^{01}(t) = y^{010}(t) - y^{001}(t + T) + V_{\text{high}} \quad (2)$$

其中, $f^{b_{-2}1}(t)$ 表示对于当前位为 0, 考虑前面比特模式为 $b_{-2}1$ 的下降边, 去除了前面边沿跳变的影响。 V_{high} 为高电平的稳态响应值。

同理, 上升边可以由下式表示:

$$r^{00}(t) = y^{001}(t) \quad (3)$$

$$r^{10}(t) = y^{101}(t) - y^{110}(t + T) + V_{\text{low}} \quad (4)$$

将二阶多边沿响应推广到更高阶以达到更高的精度, 当考虑 m 个前导位时, 多边沿响应就可以表示为:

$$r^{b_{-m}\cdots b_{-2}0}(t) = y^{b_{-m}\cdots b_{-2}01}(t) - y^{b_{-m}\cdots b_{-2}00}(t) + V_{\text{low}} \quad (5)$$

$$f^{b_{-m}\cdots b_{-2}1}(t) = y^{b_{-m}\cdots b_{-2}10}(t) - y^{b_{-m}\cdots b_{-2}11}(t) + V_{\text{high}} \quad (6)$$

其中, b_{-m} 为第 m 个前导位码元, b_0 表示当前位码元, $y^{b_{-m}\cdots b_{-1}b_0}(t)$ 表示比特序列对应的输出响应, $r^{b_{-m}\cdots b_{-2}0}(t)$ 表示前导位模式为 $b_{-m}, \cdots, b_{-2}, 0$ 的上升边, $f^{b_{-m}\cdots b_{-2}1}(t)$ 表示前导位模式为 $b_{-m}, \cdots, b_{-2}, 1$ 的下降边。

系统响应是通过对上升沿响应和下降沿响应进行线性移位叠加操作构建而成的。系统响应 $v(t)$ 具体可以表示为:

$$v(t) = \sum_{i=2}^N (b(i) - b(i-1)) \cdot S(t + (i-2)T) + v_{-\infty} \quad (7)$$

$$S(t) = \begin{cases} r(t) - V_{\text{low}} & b(i) - b(i-1) = 1 \\ 0 & b(i) - b(i-1) = 0 \\ V_{\text{high}} - f(t) & b(i) - b(i-1) = -1 \end{cases} \quad (8)$$

其中, $v_{-\infty}$ 为初始直流电压。

MER 方法中, 上升沿响应 $r(t)$ 和下降沿响应 $f(t)$ 由考虑的前导位码型模式决定。图 1 为 MER 构建系统响应的示意图。

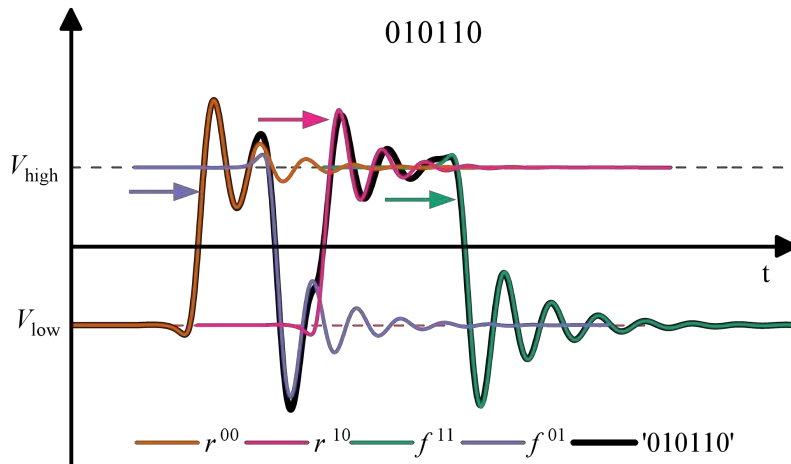


图 1 基于多边沿响应移位叠加的瞬态波形生成示意图

预先确定 MER 阶数并通过瞬态仿真获得基础系统响应后, 根据式(5)和式(6)

计算得到去除符号间干扰的纯净多边沿响应。对目标码型序列逐位扫描，识别所有发生边沿跳变的位置。对于每一个第*i*位的跃迁，提取其前面的*m*位比特(即第*i - m*位到第*i - 1*位)作为前导比特模式，用以选取对应的边沿响应。将选取的边沿响应整体右移 *m* 个比特周期，以对齐第 *i* 位边沿跳变的时间位置。遍历所有边沿跳变位置，叠加成最终的瞬态波形。

2.3 最坏眼图构建

基于单位脉冲响应的峰值失真方法(Peak Distortion Analysis, PDA)，可以在不做随机长序列仿真的情况下，直接算出最坏眼图[13]。PDA 方法的核心思想在于计算最坏符号间干扰。由反射、信道谐振和损耗引起的 ISI 在单比特响应中表现为拖尾效应，使得当前采样点受到前后比特的影响而发生失真。因此，最大失真可以通过叠加所有整数倍比特周期的 ISI 直接计算。对于最坏情况“1”，其含义是：当前比特为“1”时，所有负向的 ISI 将当前响应朝负方向推移，造成的干扰最大，可以表示为：

$$s_1(t) = y(t) + \sum_{\substack{k=-\infty, \\ k \neq 0}}^{\infty} y(t - kT)|_{y(t-kT) < 0} \quad (9)$$

同理，对于最坏情况“0”，其含义为当前比特为“0”时，所有正向 ISI 干扰将当前响应朝正方向推移，可以表示为：

$$s_0(t) = \sum_{\substack{k=-\infty, \\ k \neq 0}}^{\infty} y(t - kT)|_{y(t-kT) > 0} \quad (10)$$

最坏情况的眼图*e(t)*定义为：

$$s_1(t) > e(t) > s_0(t) \quad (11)$$

在 SBR 方法中，可以使用贪心策利用线性叠加方式最大化正向 ISI 并最小化负向 ISI 直接构建。然而，在 DER 或 MER 方法中，边沿跳变造成的 ISI 不再固定，取决于该边沿跳变的前导位比特模式，无法通过简单的符号序列选择来推导最坏眼图。借用 PDA 算法的思想，生成最坏情况眼图的核心思想在于找到在所有可能的比特序列中，对当前采样点造成最严重 ISI 干扰、使输出电压达到极值的最坏码型。随着码型长度的增加，穷举所有排列的复杂度呈指数级增长。为应对这一问题，已有研究提出了一种动态规划(Dynamic Programming, DP)算法，可以在线性时间内求解最坏情况码型[16][17]。

对于一条连续的时域波形多边沿响应*r(t)*和*f(t)*，采样点固定为第*N*个比特结束的时刻，即全序列的末尾，以捕捉所有已经输入的比特对当前采样点位的影响。每隔一个比特周期，在式(5)和式(6)计算得到的纯净边沿响应的同一相对位置进行采样，将连续的多边沿响应离散化位边沿响应数组*R[k]*和*F[k]*，定义为：

$$R[1] = r(\tau + T), \dots, R[N] = r(\tau + NT) \quad (12)$$

$$F[1] = f(\tau + T), \dots, F[N] = f(\tau + NT) \quad (13)$$

其中， $R[k]$ 和 $F[k]$ 分别代表边沿跳变发生后，经过 kT 时间的延迟，对当前的采样点状态的影响。对于阶数为 m 的MER方法，DP状态 $A_{b_{-m}\dots b_{-1}}(i)$ 表示输入到第 i 位时，已输入的序列模式中最后 m 位码型为 b_{-m}, \dots, b_{-1} 的最坏情况累积电压。对于第 $i+1$ 位比特， $A_{b_{-m}\dots b_{-1}}(i+1)$ 只能通过 $A_{1,b_{-m}\dots b_{-2}}(i)$ 和 $A_{0,b_{-m}\dots b_{-2}}(i)$ 叠加的“ $b_{-2} \rightarrow b_{-1}$ ”码元跳变的影响得到。以当前采样点位参考系，“ $b_{-2} \rightarrow b_{-1}$ ”码元跳变的影响会延迟 $(N-i)T$ 时间到达采样点处，这一影响表示为 $S[N-i]$ 。动态规划法的具体公式如下：

$$A_{b_{-m}\dots b_{-1}}(i+1) = \min \begin{cases} A_{1,b_{-m}\dots b_{-2}}(i) + (b_{-1} - b_{-2})S_{\Delta t}^{1,b_{-m}\dots b_{-2}}[N-i] \\ A_{0,b_{-m}\dots b_{-2}}(i) + (b_{-1} - b_{-2})S_{\Delta t}^{0,b_{-m}\dots b_{-2}}[N-i] \end{cases} \quad (14)$$

其中，

$$S_{\Delta t}^{1,b_{-m}\dots b_{-2}}[N-i] = \begin{cases} R_{\Delta t}^{1,b_{-m}\dots b_{-2}}[N-i] - V_{\text{low}}, & b_{-1} - b_{-2} = 1 \\ F_{\Delta t}^{1,b_{-m}\dots b_{-2}}[N-i] - V_{\text{high}}, & b_{-1} - b_{-2} = -1 \\ 0, & b_{-1} - b_{-2} = 0 \end{cases} \quad (15)$$

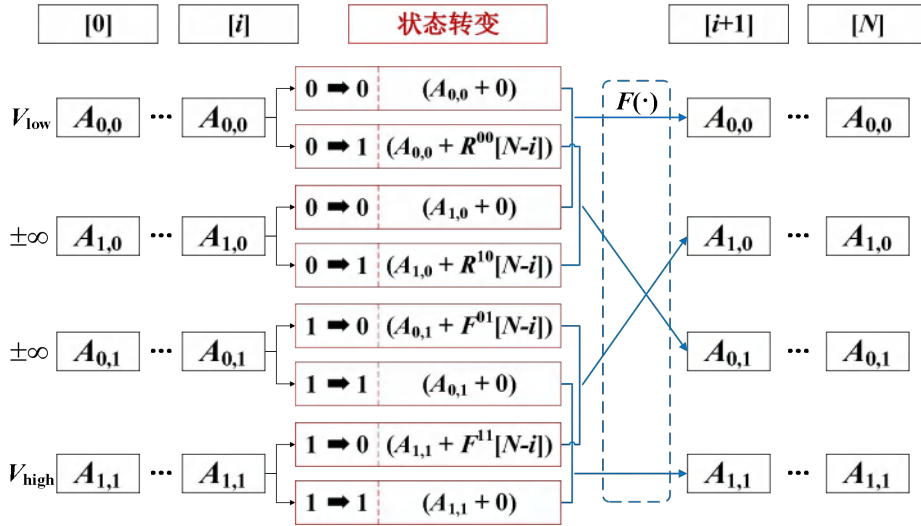
其中， $S_{\Delta t}^{1,b_{-m}\dots b_{-2}}[N-i]$ 代表 Δt 时，第 i 个位置发生的前导位为1， b_{-m}, \dots, b_{-2} 的边沿跳变在采样点位的造成的电压变化值。 $S_{\Delta t}^{0,b_{-m}\dots b_{-2}}[N-i]$ 同样定义。不考虑任何ISI影响时，初始化稳态状态：

$$A_{0,\dots,0}(0) = V_{\text{low}} \quad (16)$$

$$A_{1,\dots,1}(0) = V_{\text{high}} \quad (17)$$

$$A_{\text{others}}(0) = \infty \quad (18)$$

图2展示了该动态规划方法的详细流程图。



最坏情况“1”： $F(\cdot)=\max(\cdot)$ ；最坏情况“0”： $F(\cdot)=\min(\cdot)$

图2 基于动态规划的最坏情况眼图构建算法

通过填充大小为 $N \times M$ 的状态矩阵 A ，最终当前采样点位的最坏情况“1”为

矩阵最后一列的最小值。对整个单元间隔周期内的采样点进行遍历，即可绘制出最坏情况“1”的眼图。通过回溯矩阵中获得到最小值的路径，可提取对应导致最坏眼图的最坏码型。对于最坏情况“0”，和最坏情况“1”类似，只需将初始化稳态状态 $A_{\text{others}}(0) = -\infty$ ，同时将上述的最小化操作替换为最大化，寻找使得采样点电压最大的最坏码型序列。

本章所述的快速时域仿真方法相比传统瞬态仿真，能够以更低的资源消耗和更高的仿真效率生成瞬态波形以及最坏眼图，从而构建足够规模的数据集，为后续迁移学习源域模型的建立奠定基础。此外，该方法流程不依赖于特定电路拓扑结构，具有良好的通用性，适用于各种现代非线性高速链路模型。

3 迁移学习

迁移学习是一种系统化的机器学习方法。传统的机器学习方法通常从零开始学习每一个任务，而迁移学习则利用源域 D_S 及其相关学习任务 T_S 中获得的知识，来提升目标域 D_T 上目标任务 T_T 的预测函数 $f(\cdot)$ 的性能，其中 $D_S \neq D_T$ 或 $T_S \neq T_T$ [26]。通常，在目标域样本数量远少于源域样本数量的情况下，迁移学习方法尤为有效。本章将详细介绍迁移学习策略，并在小样本条件下，将其应用于构建用于几何参数到眼高预测任务下的深度神经网络(Deep Neural Network, DNN)模型和输入输出波形映射任务下的时间卷积网络(Temporal Convolutional Network, TCN)模型。

3.1 通用迁移学习策略

在迁移学习中，“迁移什么(what to transfer)”和“如何迁移(how to transfer)”是两个核心问题。“迁移什么”关注于识别哪些知识可以在不同领域或任务之间有效共享，而“如何迁移”则涉及设计相应的学习算法以实现这些知识的迁移[27]。当所迁移的知识能够反映两个高度相关模型之间的共性特征时，迁移学习可以显著提升目标域模型的性能；相反，如果强行迁移彼此关联性较弱的知识，反而可能导致性能下降，即负迁移现象。

在本文提出的方法中，源域模型与目标域模型模拟的是同一个电路结构，二者主要区别体现在互连线建模精度与时域信号生成策略上。源域模型使用快速时域仿真，并采用电路仿真中内置的传输线模型对互连结构进行建模，能够以极高的仿真效率生成所需的时域波形与最坏情况眼图，避免了瞬态仿真中对长伪随机序列的高资源消耗；而目标域模型则采用瞬态仿真，并基于电磁仿真提取的S参数对互连线进行高精度建模，从而获得更加真实和精确的信号响应。

尽管两者的整体仿真框架均为电路级时域仿真，但在互连建模精度与时域波形生成方式上的差异，使源域与目标域在数据质量和计算代价上形成互补。两个模型共享相同的神经网络结构，并服务于一致的任务目标，这种高度相关性为迁移学习的有效实施提供了坚实基础。

相较于目标域的高精度仿真，源域数据在保持物理一致性的同时具有更高的生成效率，适合作为预训练阶段的数据来源。源域模型 ANN_S 利用源域大规模数据集 $\{\mathbf{x}_S, \mathbf{y}_S\}$ 完成预训练，可构建出输入到输出的映射函数：

$$\mathbf{y}_{Pred} = f_S(\mathbf{x}_S) \quad (19)$$

其中， \mathbf{x}_S 为源域输入数据， \mathbf{y}_{Pred} 为对应预测目标。训练过程中，网络不断调整权重和偏置参数，以最小化损失函数。本文采用的模型均使用预测值与实际值的均方误差 (Mean Squared Error, MSE) 作为损失函数，训练过程可具体由下列优化问题表示：

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (20)$$

$$(\mathbf{w}_S, \mathbf{b}_S) = \operatorname{argmin} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) \quad (21)$$

其中， n 是样本数量， y_i 是样本的实际结果， \hat{y}_i 是模型输出的样本的预测值。

根据源域与目标域之间的关系，可以采用不同的迁移学习方法。本文主要探讨两种主要的迁移学习策略：全模型微调方法 (Fine-Tuning) 和部分冻结方法 (Partial-Freezing)。

3.1.1 全模型微调方法

源域与目标域模型采用相同的神经网络结构，输入层与输出层的节点数量一致。Fine-Tuning 方法将源域神经网络模型 ANN_S 中的所有参数（权重和偏置）直接迁移到目标模型 ANN_T 中作为初始权重：

$$\{\mathbf{w}_{l,T}, \mathbf{b}_{l,T}\}_{l=1}^m = \{\mathbf{w}_S, \mathbf{b}_S\} \quad (22)$$

其中， $\{\mathbf{w}_{l,T}, \mathbf{b}_{l,T}\}_{l=1}^m$ 表示从 $l = 1$ 到 $l = m$ 各隐藏层的权重和偏置。由此，源域模型的知识被完全转移到了目标域模型中。 ANN_T 在少量高精度瞬态仿真数据上进行再训练，微调权重和参数，使得损失函数 MSE 最小。优化过程中采用 Adam 优化器对各层参数 θ_t 进行微调，其更新公式如下所示：

$$\theta_t^{(k+1)} = \theta_t^{(k)} - \alpha \frac{\mathbf{m}_k / (1 - \beta_1^k)}{\sqrt{\mathbf{v}_k / (1 - \beta_2^k) + \varepsilon}} \quad (23)$$

其中， α 为学习率， \mathbf{m}_k 和 \mathbf{v}_k 分别表示第 k 次迭代中的一阶矩估计和二阶矩估计， β 为对应的衰减因子。

由于初始化的权重参数已较接近目标优化的最优值，采用“热启动”方式能够加速神经网络的收敛过程。然而，固定模型结构可能限制其实用灵活性。当目标域样本充足时，全模型微调能够充分适应目标域的特征；但在目标域样本稀缺的情况下，大量参数和较高的训练自由度容易导致模型过拟合，反而不利于知识迁移的泛化效果。

3.1.2 部分冻结方法

源域模型与目标域模型采用相同的网络模型,结构上划分为特征提取层与任务适应层。**Partial-Freezing** 仅迁移源域模型中前 k 层特征提取层的通用特征参数,其余层的参数随机初始化:

$$\{\mathbf{w}_{l,T}, \mathbf{b}_{l,T}\}_{l=1}^k = \{\mathbf{w}_S, \mathbf{b}_S\}_{l=1}^k, k < m \quad (24)$$

在获得瞬态仿真数据 $\{\mathbf{x}_T, \mathbf{y}_T\}$ 后,目标模型训练过程中冻结特征提取层的权重和偏置,仅对任务适应层的权重进行训练。在网络训练时,仅计算并更新任务适应层参数的梯度 θ_t ,而特征提取层参数 θ_f 保持不变,即:

$$\nabla_{\theta_f} \text{MSE} = 0 \Rightarrow \theta_f^{(k+1)} = \theta_f^{(k)} \quad (25)$$

该模型同样采用 Adam 优化器进行优化,其任务适应层参数 θ_t 更新遵循式(23)所示的更新规则。

Partial-Freezing 方法利用了深度神经网络中浅层负责提取通用特征、深层学习任务特定表示的结构特性。通过保持前 k 层与源域模型结构一致并冻结其参数,目标域模型可在第 k 层之后根据任务需求灵活调整,如增加隐藏神经元数量或扩展网络深度。不同任务及迁移学习场景对冻结层数的需求存在差异:冻结过多层会限制模型对目标域特征的适应能力,而冻结过少则可能导致知识迁移不足。后续研究将结合具体迁移任务与模型配置,探讨最优的层冻结策略。

3.2 几何参数映射眼高的迁移学习模型

构建电路模型几何参数到眼高的映射是机器学习方法的典型应用。在迁移学习框架下,该任务定义如下:由快速时域仿真配合电路仿真中传输线模型构建的最坏情况眼图,作为一种低成本、高效率的方式,生成大规模包含几何参数与对应眼高的源域数据。而目标域数据则基于瞬态仿真结合 EM 仿真提取的互连结构 S 参数进行建模,具备更高的仿真精度。本文采用 DNN 分别构建源域与目标域的代理模型,并通过迁移学习策略,在有限的高精度目标域数据上实现几何参数到眼高的有效映射,从而降低对高成本数据的依赖。

如图 3 所示,典型的 DNN 架构由多个层组成,包括输入层、若干隐藏层以及输出层。电路模型的几何参数由随机变量表示,记为 $\mathbf{x} = [x_1, x_2, \dots, x_n]$ 。每一隐藏层由权重矩阵、偏置向量与激活函数组成,用于从输入中逐层提取特征并传递至下一层,从而实现输入到输出之间的非线性映射建模:

$$\mathbf{y}^{(h)} = \sigma(\mathbf{W}^{(h)}\mathbf{y}^{(h-1)} + \mathbf{b}^{(h)}) \quad (26)$$

其中,权重矩阵 $\mathbf{W}^{(h)}$ 和偏置向量 $\mathbf{b}^{(h)}$ 为网络中的可训练参数, σ 表示激活函数。模型参数通过 Adam 优化器进行更新,该优化器根据梯度方向和幅度对参数进行自适应调整,使损失函数值逐步收敛至全局最小值。

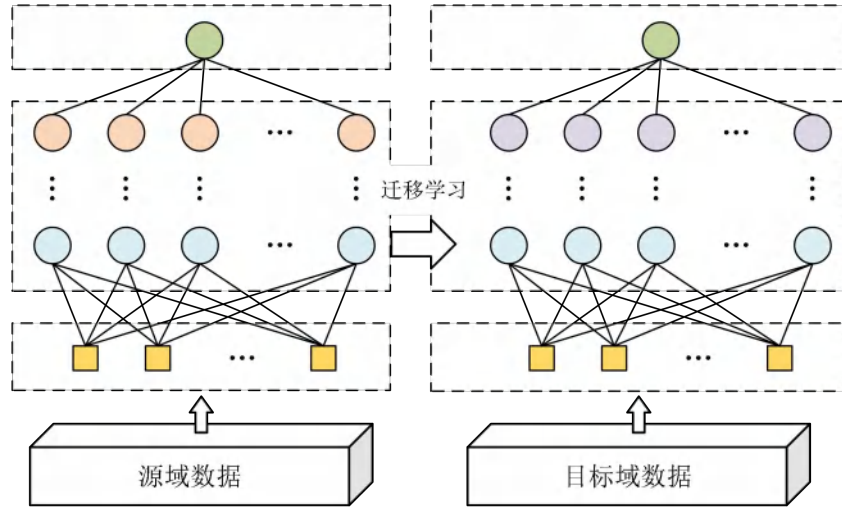


图 3 基于DNN模型的迁移学习方法

本文源域模型 DNN_S 与目标域模型 DNN_T 针对同一类似电路结构共享相同的参数空间，均采用具有相同隐藏层数量与神经元规模的 DNN 架构。基于前述两种迁移学习策略，第一种方法将 DNN_S 中的所有权重与偏置参数完整迁移至 DNN_T ，并在目标域数据上进行全模型微调；第二种方法则固定前 k 个隐藏层的参数，仅对剩余 $m - k$ 层进行 Xavier 初始化，并在目标域数据集上训练这些任务适应层。通过改变冻结层数，对不同参数迁移深度下的迁移学习效果进行对比分析，可以有效评估各策略在建模性能与泛化能力方面的差异。

3.3 输入输出波形映射的迁移学习模型

序列建模任务是引入机器学习方法解决高速链路仿真限制的另一种成功范式[28]。快速时域仿真通过移位叠加边沿响应，等效模拟了信号经过信道后的输出波形，在这种建模任务上迁移学习同样可以充分适用。

循环神经网络(Recurrent Neural Networks, RNN)通过维护随时间传播的隐藏状态向量，在捕捉序列数据中的长期依赖性方面表现出色。然而，近期研究表明，某些卷积神经网络结构在音频合成、机器翻译等序列处理任务中已能达到甚至超过 RNN 的性能[29]。TCN 是一种专为时间序列建模设计的卷积结构。通过引入膨胀卷积与残差连接，TCN 在保持训练稳定性的同时，能够显著扩大感受[30]。

TCN 架构主要包含三部分组成：因果卷积、膨胀卷积和残差连接。因果卷积确保模型遵循时间顺序，即输出仅依赖于当前及之前的输入，而不包含任何未来信息。为了保持输出与输入序列长度一致，通常对输入在前端进行零填充，长度为卷积核大小-1。对于一维输入信号 $\mathbf{x} \in \mathbf{R}^n$ 和卷积核 $f: \{0, \dots, k - 1\} \rightarrow \mathbf{R}$ ，常规的一维因果卷积层可以定义为：

$$F[\mathbf{x}(t)] = (\mathbf{x} * f)(t) = \sum_{i=0}^{k-1} f(i)\mathbf{x}(t - i) \quad (27)$$

其中， k 是卷积核大小， $F(\cdot)$ 为卷积操作。

为了在不过多增加模型参数的前提下扩大感受野，TCN 引入了膨胀卷积，这是一跳过某些输入单元的卷积形式，从而更好地理解输入数据中的上下文信息。扩张因子决定了卷积核中元素之间的间距。当与因果卷积相结合时，第 r 层的卷积过程可以表示为：

$$F[x(t)] = (x *_{d_r} f)(t) = \sum_{i=0}^{k-1} f(i)x(t - d_r i) \quad (28)$$

其中， d_r 是第 r 层的扩张因子，可以设置为 2^{r-1} 。膨胀卷积相当于在两个相邻的滤波器之间引入一个固定的步长，使得高层输出能够覆盖更宽范围的输入，扩展卷积网络的感受野。膨胀因果卷积示意图如图 4 所示，

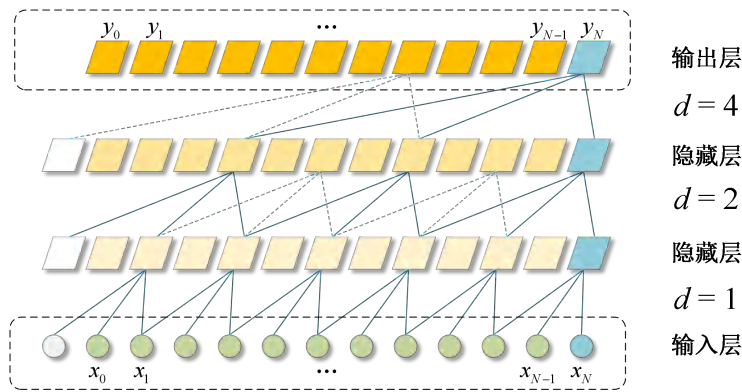


图 4 采用膨胀因子 $d = 1, 2, 4$ 且滤波器尺寸 $k = 3$ 的膨胀因果卷积

残差连接已被证明可有效缓解深度神经网络训练过程中的梯度消失与梯度爆炸问题，同时提升训练效率与最终性能。因此，标准的深层 TCN 架构通常将多个卷积层组成残差块，并引入残差连接形成“跳跃连接”结构，其输出形式为：

$$\text{Output} = \text{Activation}(x + \mathcal{F}(x)) \quad (29)$$

其中， \mathcal{F} 表示由卷积层与激活函数组成的组合映射。如图 5 所示，残差块结构由两个扩张因果卷积层组成，每层后接 ReLU 激活函数。

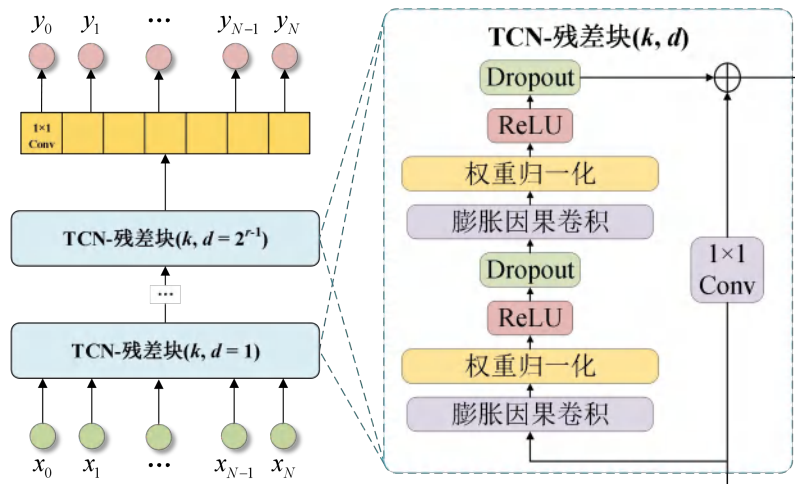


图 5 TCN残差块示意图

所有卷积核均进行权重归一化，并采用 dropout 技术实现正则化以防止过拟合。最终，TCN 的输出层通过一维卷积映射得到目标输出波形。在电路模型的输入输出波形映射任务中，调整输入信号的比特率来扩充数据多样性。源域模型 TCN_S 训练所用的数据集是通过传输线模型仿真得到的边沿信号进行时移与叠加后生成的输出响应波形。在此基础上，本文同样采用上述两种参数迁移策略进行建模，冻结 TCN_S 中部分残差块的权重与参数，仅对其余部分在目标域模型 TCN_T 中进行再训练，从而实现已学知识的高效迁移与适应。

3.4 总体方法阐述

图 6 展示了本文所提出的方法的总体流程和框架。

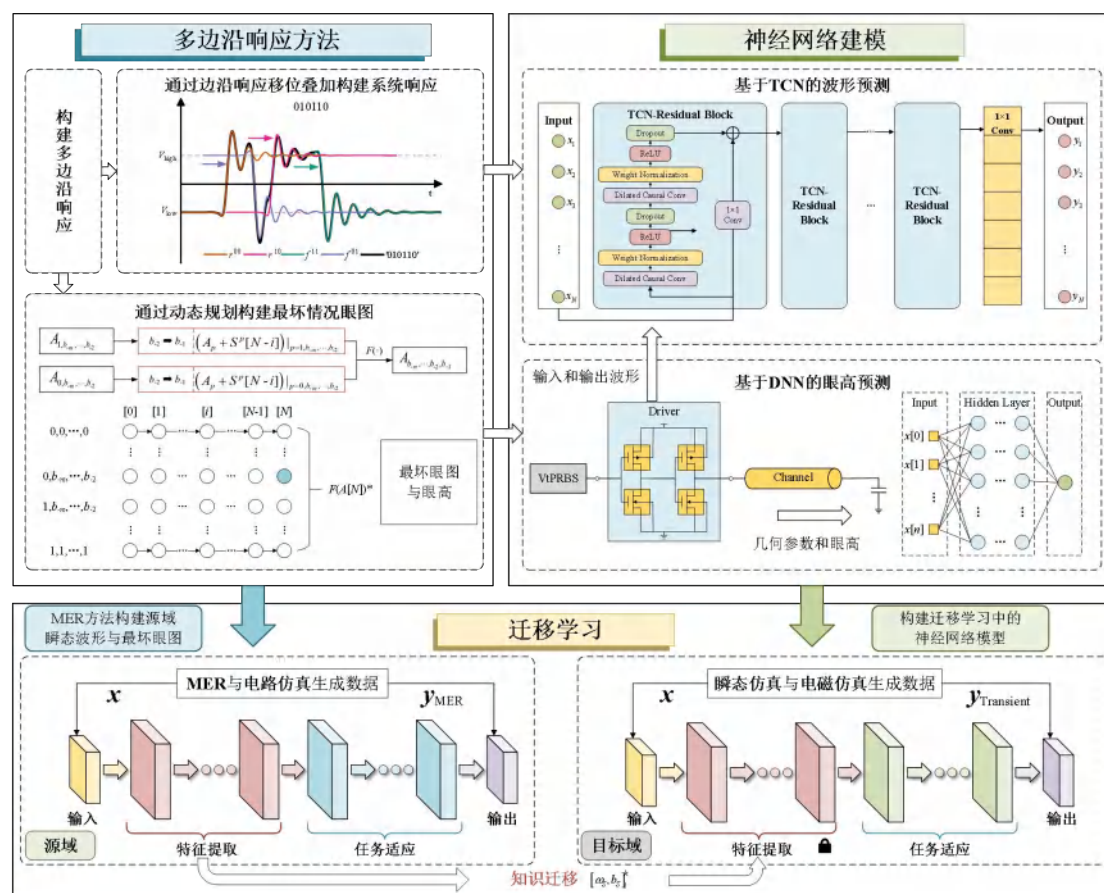


图 6 本文采用的总体建模方法框架

本文所提出的建模方法，首先在源域通过基于电路仿真模型的高效时域仿真生成两类训练数据：一方面利用第 2.2 节的内容，将任意比特序列的脉冲或边沿响应按时间偏移叠加，快速合成高速链路的瞬态输出波形，作为输入输出波形映射任务的源域数据生成方式；另一方面利用第 2.3 节的内容，通过动态规划扫描各采样点，对纯净边沿响应进行最坏情况叠加，得到最坏眼图及相应的眼高，形成几何参数到眼高预测任务所需的源域样本。

在神经网络建模与迁移学习阶段，本文针对两项建模任务分别设计了架构并验证了两种迁移策略。对于几何参数到眼高预测任务，采用 DNN 网络模型，对

于输入输出波形映射任务，选用 TCN 模型作为迁移学习的代理模型。使用 Fine-Tuning 和 Partial-Freezing 两种迁移学习策略，通过将源域预训练得来的网络参数直接迁移至目标模型，或在此基础上冻结若干前置残差块再微调后续层，来实现对有限高精度瞬态仿真样本的高效利用。

4 实验与结果分析

本章通过具体的高速链路设计示例，验证所提出的从源域基于传输线模型的快速时域仿真向目标域基于电磁仿真提取的 S 参数模型的瞬态仿真的通用迁移学习方法在不同应用场景下的有效性。

针对眼高预测与瞬态波形建模任务，分别构建基于 DNN 与 TCN 的神经网络架构，并结合迁移学习策略进行建模。通过将这些迁移学习模型与传统从零训练的基础模型进行对比，可全面评估源域先验信息在目标域建模中的迁移效果，以及其对模型性能与泛化能力的提升作用。

4.1 高速链路结构

图 7 展示了高速链路的基本结构，由发送端(Transmitter, TX)、互连通道和接收端(Receiver, RX)组成[31]。发送端采用伪随机二进制序列进行激励，信号电压范围为 0~1.1V。本文选用了由四个 CMOS 晶体管构成的两级驱动器作为 TX 的核心驱动电路(PMOS: PM_i , NMOS: NM_i)[32]。其中，晶体管的栅宽和栅长作为关键物理设计变量。

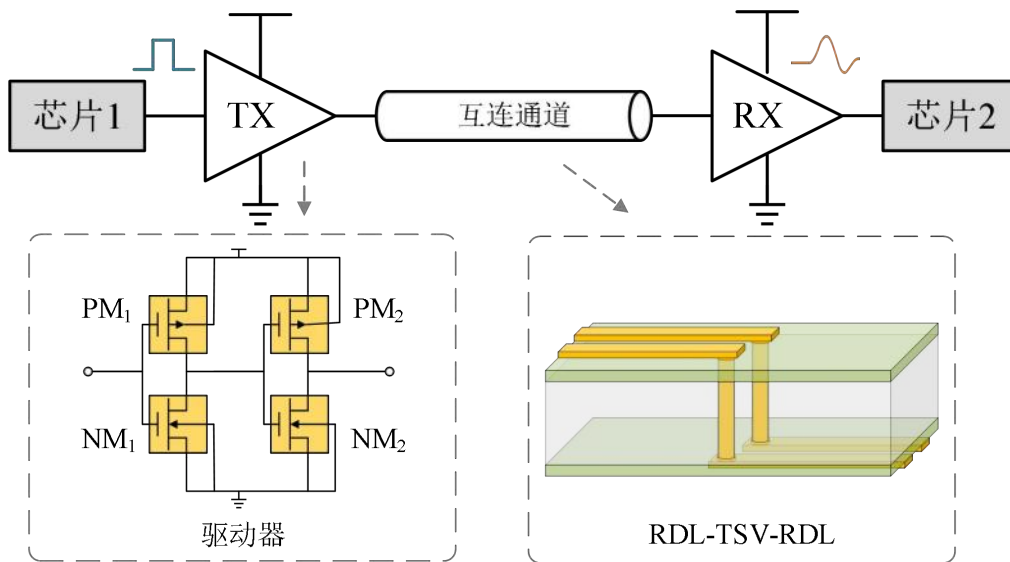


图 7 高速链路结构示意图

随着系统级芯片(SoC)技术节点的不断缩小，芯片良率下降且制造成本上升。三维集成通过硅通孔(Through-Silicon Via, TSV)实现垂直连接，并通过重分布层(Redistribution Layer, RDL)完成水平布线，成为提升系统性能和集成密度的关键手段。本文选取 RDL-TSV-RDL 互连结构作为 TX 与 RX 之间的互连信道。

图 8 为 RDL-TSV-RDL 结构示意图，包含 TSV 的垂直互连结构和 RDL 的水平互连结构。RDL 通过凸点(Bump)与 TSV 相连，形成三维封装的标准互连结构。TSV 由直径为 d_{TSV} 、高度为 h_{TSV} 的圆柱形铜柱构成，嵌入硅基板中。铜柱与硅基板之间填充二氧化硅绝缘层，实现电隔离。凸点直径为 d_{Bump} ，高度为 h_{Bump} ，用于连接 TSV 与 RDL。相邻 TSV 中心之间的平面距离定义为 $pitch$ ，代表互连密度的关键指标[33]。用以描述互连密度。其中 TSV 半径、高度、氧化层厚度和 RDL 金属层厚度由具体的工艺决定，本文将相邻 TSV 间的 $pitch$ 与 RDL 宽度 w_{RDL} 作为关键物理设计变量，利用神经网络学习其耦合效应等特性。具体参数选取如表 1 所示。

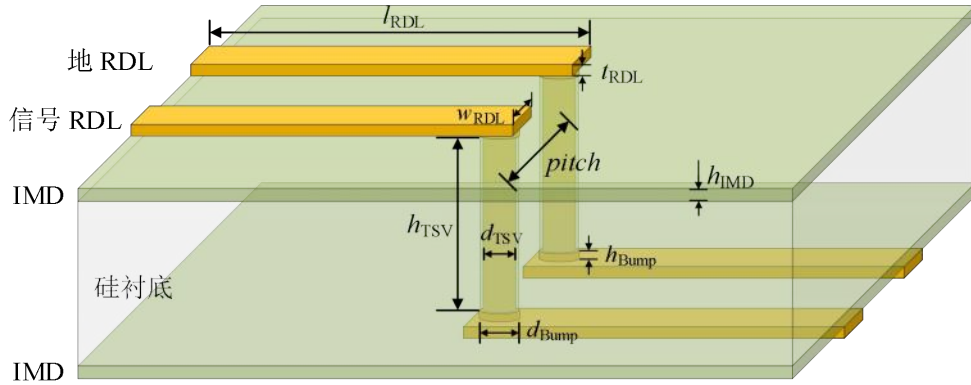


图 8 RDL-TSV-RDL 结构模型及结构参数示意图

表 1 高速链路结构参数选取范围表

	参数	范围（眼高任务）	范围（波形任务）
RDL-TSV-RDL	$pitch$ (μm)	40~60	40
	w_{RDL} (μm)	10~30	12
驱动器	L_{N} (μm)	0.1~0.2	0.15
	W_{N} (μm)	50~100	100
	L_{P} (μm)	0.1~0.2	0.2
	W_{P} (μm)	50~250	200
	比特率 (Gbps)	6.4	2~8

在源域中，互连通道采用内置传输线模型，通过 MER 方法计算瞬态波形和最坏眼图；在目标域中，对 RDL-TSV-RDL 结构首先在 EM 仿真平台中建立详细模型并对该模型的 S 参数、瞬时阻抗等电磁特性进行仿真分析。生成网络参数文件嵌入电路仿真中进行时域瞬态仿真，保证该结构既能保留器件的场耦合效应，也能兼顾电路特征。图 9(a)分别展示了瞬态仿真与 MER 方法计算得到的最坏眼图。结果显示 MER 方法生成的最坏眼图基本吻合瞬态仿真平均眼图的内轮廓。图 9(b)对比了 MER 方法重构的瞬态输出波形与瞬态仿真波形。两者在波形形态和关键时序特征上高度一致，由此验证了使用 MER 方法作为源域模型进行知识

迁移的合理性,从本质上规避了源域和目标域间差异过大而导致的负迁移问题。

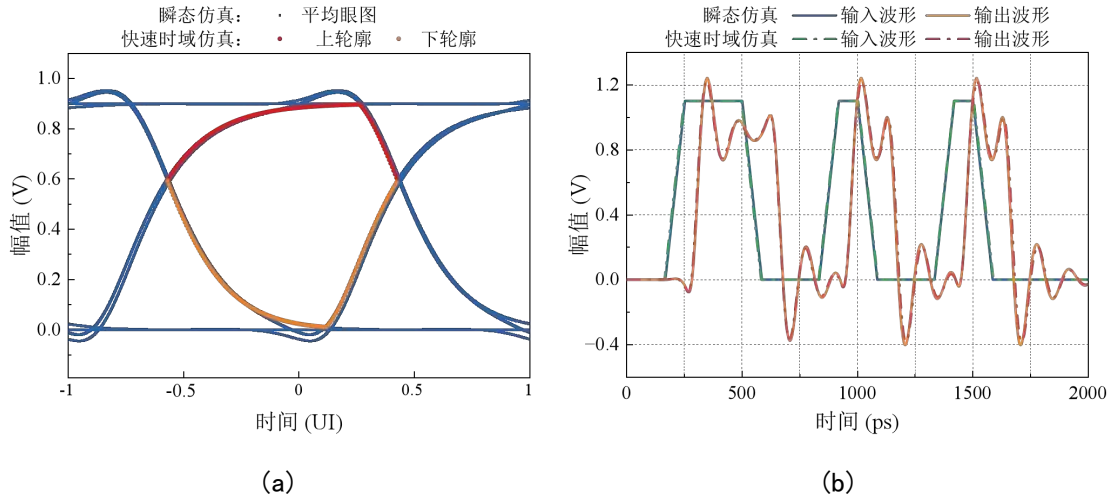


图 9 二阶MER方法与瞬态仿真的对比：(a) 最坏情况眼图；(b) 瞬态波形

4.2 眼高映射任务

驱动器的输入信号速率设定为 6.4 Gbps, 上升与下降时间均为 85 ps。几何设计参数作为模型输入, 瞬态仿真或快速时域仿真获得的眼高作为模型输出, 建立从几何结构到眼高的映射关系。训练数据采用六维参数空间中的拉丁超立方体采样 (Latin Hypercube Sampling) 方法生成。

所构建的 DNN 网络架构包含五层全连接层, 隐藏层神经元数量依次为 [128, 256, 256, 128, 128], 激活函数采用 Tanh, 使用 Xavier 进行初始化, 优化器为 Adam, 学习率设为 0.001。DNNs 在 300 组源域数据样本上进行预训练。在目标域中, 分别构建包含 50 组、100 组、200 组及 400 组样本的训练数据集, 基于源域模型进行迁移学习构建 DNN_T。验证集重新采样 100 组不同的数据集, 对比针对同一验证集的迁移学习后的模型与传统模型的预测能力, 系统评估迁移学习策略在不同样本规模下的效果。

在模型训练前, 对所有输入特征与输出标签均采用最大最小值归一化 (Min-Max Normalization) 方式, 将数据缩放至 [0,1] 区间, 以提升模型训练稳定性与收敛效率:

$$\mathbf{x}_{\text{normal}} = \frac{\mathbf{x} - \mathbf{x}_{\min}}{\mathbf{x}_{\max} - \mathbf{x}_{\min}} \quad (30)$$

图 10 展示了 DNN 与迁移学习后的 DNN 模型 (DNN-TL) 在验证集上的预测精度。横轴为真实眼高, 纵轴为模型预测结果。参考线 $y = x$ 表示理想预测情况。可以观察到, 两种模型的预测结果整体均较好地贴合参考线, 说明 DNN 模型具备良好的任务适应性。相较之下, 传统 DNN 模型在高低端眼高区间表现出明显偏差, 说明 DNN-TL 对整体趋势的捕捉更为精准, 对边缘情况也有较好的泛化能力。图 10(b) 展示了两种模型在训练过程中的损失收敛情况。DNN-TL 在源域模型的预训练基础上进行初始化, 具备先验知识, 仅需约 5 到 10 个轮次即可收

敛到近似最小损失；而传统 DNN 模型则需数十个轮次才能收敛。这种“热启动”优势显著提高了训练效率，同时减少了对超参数调优与模型结构调整的依赖。

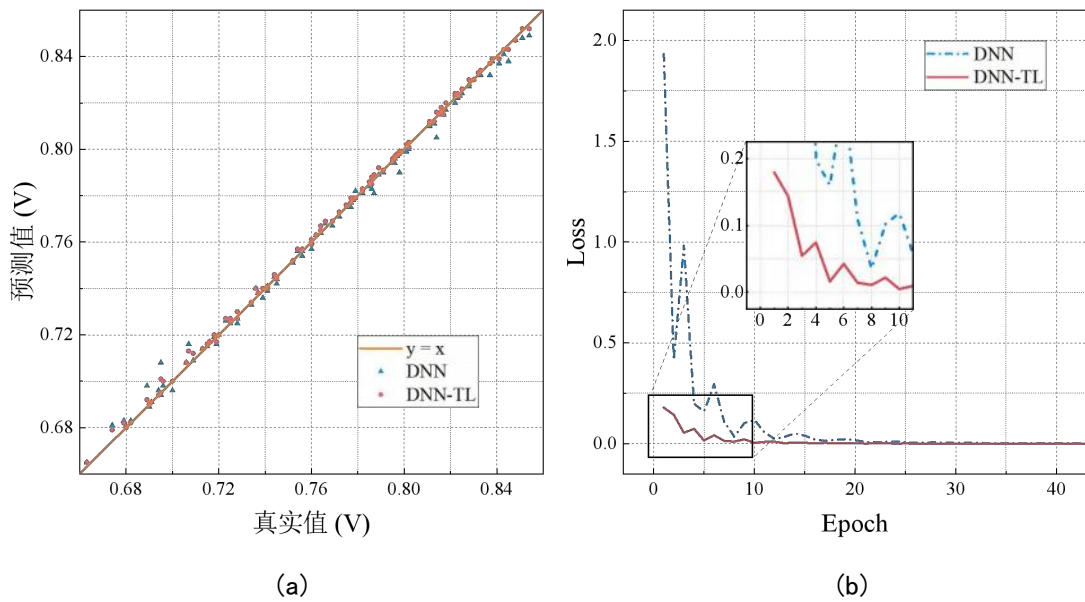


图 10 DNN与DNN-TL性能对比图，(a) 眼高预测准确率 (b) 训练损失收敛对比

为评估模型的稳定性，对每个模型进行五次独立训练，并对其均值与标准差进行统计比较。图 11 比较了在不同训练集规模下，两种模型的均方根误差(Root Mean Squared Error, RMSE)与平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)。

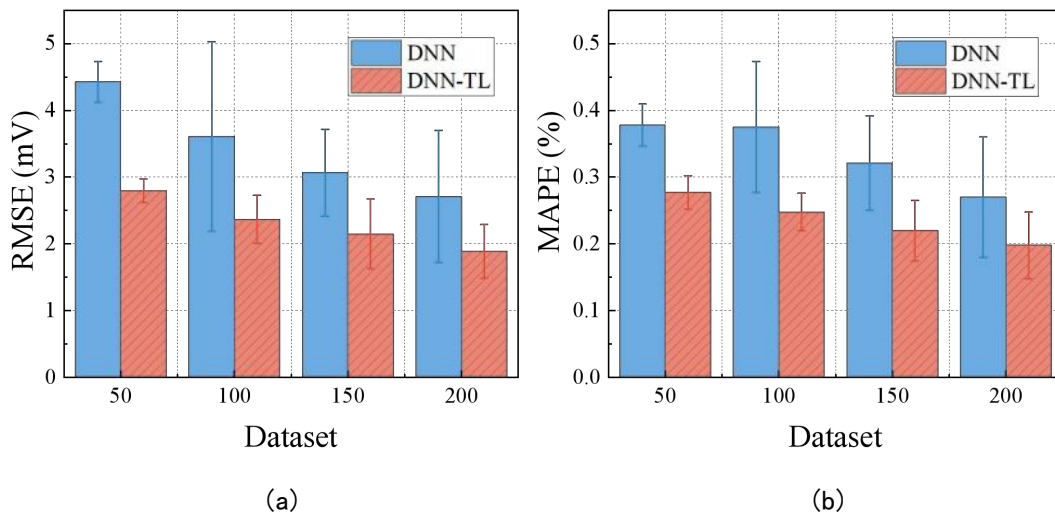


图 11 不同数据规模下DNN与DNN-TL模型的精度对比图，(a) RMSE, (b) MAPE

结果显示，随着样本数量的增加，两种模型的误差均逐渐减小，但所提出的 DNN-TL 模型在所有样本规模下始终优于传统 DNN 模型，表现出更低的误差标准差与更强的稳定性。表 2 对比了两种模型在四种数据集规模下的预测精度与训练时间。与传统 DNN 相比，DNN-TL 模型在 RMSE 与 MAPE 指标上实现了约 30%~40%的显著提升。

表 2 DNN与DNN-TL的准确率与计算效率对比

	DNN			DNN-TL		
	RMSE (mV)	MAPE (%)	Time (s)	RMSE (mV)	MAPE (%)	Time (s)
50	4.428	0.38	6.82	2.796	0.28	4.18
100	3.607	0.37	10.55	2.374	0.25	6.72
150	3.079	0.32	14.21	2.157	0.22	9.47
200	2.718	0.27	199.07	1.891	0.19	12.30

4.3 波形映射任务

在高速链路的输入输出波形映射任务中，为增强数据多样性，本文输入信号的比特率设置为 2 Gbps 至 8 Gbps，仿真时长固定为 2 ns，时间步长设为 1 ps，得到 2000 个采样点的波形序列。

本文源域采用二阶 MER 方法，仅需考虑两个前导位码元，因此仅需预处理 $2^3 = 8$ 个基本码型组合，通过在目标位后拼接重复符号以提取纯净边沿响应。通过对这些边沿响应进行时移与叠加，可在任意比特率下高效生成 1024 组波形样本，无需重复仿真。最终，从多速率、多符号组合的 MER 数据集中随机选取 500 条样本作为源域训练集。目标域样本使用相同的比特率范围和 10 个比特位的随机组合，共仿真生成 200 组样本按 1:1:2 比例划分为训练集、验证集和测试集。

本文采用了较深的 TCN 结构，由 6 个残差模块构成。每个残差模块内部包含 2 层膨胀因果卷积，卷积核大小统一设定为 5。各块的通道数依次设为 [64, 128, 128, 256, 256, 512]，每层设置丢弃率为 0.2 的 Dropout 策略以防止过拟合。通过指数增长的扩张因子，该结构可捕捉覆盖百余时间步的历史依赖信息。最终一维卷积层将高维特征映射回与输入等长的波形，实现端到端序列映射。本文将采用迁移学习的 TCN 模型(TCN-TL)与标准 TCN 进行比较，同时也引入传统序列建模网络 LSTM 和 GRU 作为对比模型。结果如图 12 所示。

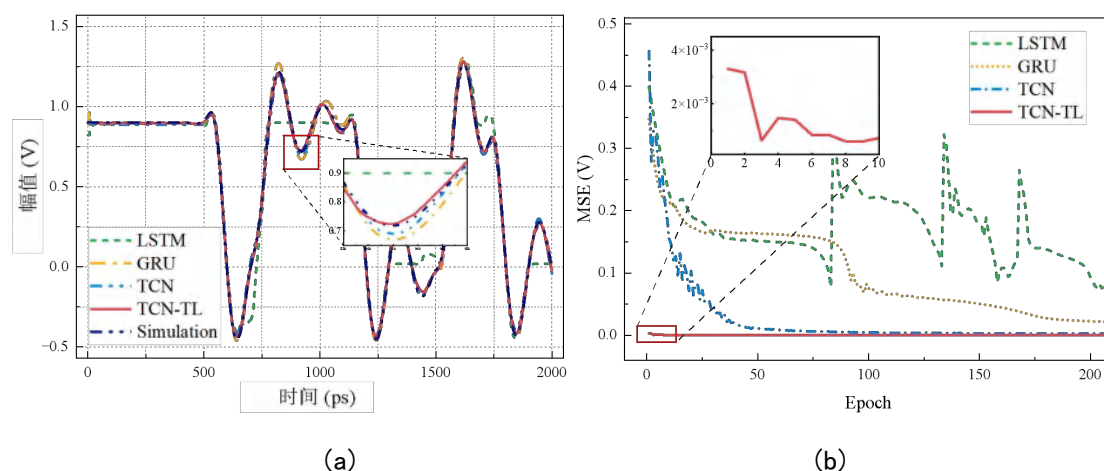


图 12 LSTM、GRU、TCN与TCN-TL对比图：(a) 预测结果，(b) 训练损失收敛

如图 12(a)所示, 在存在强振铃和严重符号间干扰的情况下, 只有 TCN-TL 模型能够紧密跟随真实仿真波形, 尤其是在快速跳变和细小尖峰区域。局部细节表明, 其预测的波峰与波谷几乎与参考波形重合, 而其他模型则存在滞后或振幅低估的现象。图 12(b)展示了各模型在训练过程中的均方误差变化趋势, TCN-TL 在初始阶段就迅速下降至极低的误差水平, 相比之下, LSTM 与 GRU 模型在训练初期存在更高的初始误差, 且收敛速度较慢, 需经过更多轮训练才能达到相似的精度水平。

表 3 详细列出了各类网络模型预测精度的对比, 采用均方误差和判定系数 (Coefficient of Determination, R^2) 作为评价指标。TCN-TL 的 MSE 为 $2.689 \times 10^{-4} V^2$, 分别比传统 TCN、GRU 和 LSTM 降低了 57.8%、76.8% 和 97.3%。其在训练集和验证集上的 R^2 值分别达到 0.9999 和 0.9998, 几乎与真实波形完全吻合。

表 3 TCN-TL 与 TCN、LSTM、GRU 的准确率对比

	训练集		验证集	
	MSE ($\times 10^{-4} V^2$)	R^2	MSE ($\times 10^{-4} V^2$)	R^2
LSTM	99.905	0.9534	108.587	0.9511
GRU	11.583	0.9944	11.511	0.99443
TCN	6.371	0.9970	6.552	0.9970
TCN-TL	2.689	0.9987	2.762	0.9987

4.3 迁移学习策略对比

本小节系统比较了不同迁移学习策略对神经网络性能的影响。在眼高预测任务中, 使用 100 个训练样本训练一个传统的 DNN 模型对比。随后对目标模型分别采用两种迁移学习策略: Fine-Tuning 和 Partial-Freezing。后者通过冻结预训练模型的前 1~4 层权重, 并将剩余层用 Xavier 方法初始化。在波形预测任务中, 基于 TCN 架构, 使用 50 个训练样本建立 TCN 模型, 同时采用全微调策略和部分冻结策略对每一个残差块进行迁移。在 Partial-Freezing 方法中, 从冻结前 1 至 5 个残差块展开实验, 对比不同冻结深度下模型性能的变化。所有模型均在统一的验证集上评估, 量化冻结层或残差块数量对迁移效果的影响。

不同迁移策略下的性能对比如图 13 所示, 可以看出: 不采用迁移的基础模型性能最差, 冻结更多层或模块虽然降低了可训练参数的比例, 但却带来了显著性能提升。从表 4 可以看出, 在 DNN 眼高映射任务中, 冻结前 3 层可获得最优的 RMSE 和 MAPE, 进一步冻结第 4 层后性能略有下降, 表明可训练层数过少将限制模型对目标任务的适应能力。类似地, 在 TCN 波形预测任务中, 表 4 显示冻结前 5 个残差块可获得最低的 MSE 和最高的 R^2 值。虽然全微调策略也优于非迁移模型, 但相比精细设计的部分冻结策略, 性能仍存在差距。

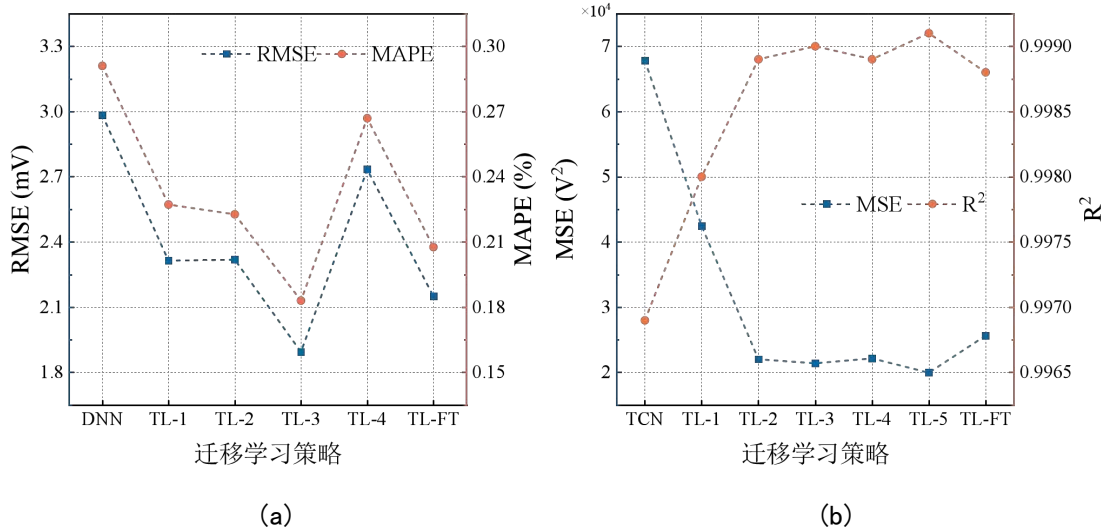


图 13 不同迁移学习策略下的模型性能对比图：(a) DNN，(b) TCN

表 4 不同迁移学习策略下DNN与TCN模型的性能对比

	Non-TL	Partial-Freezing					Fine-Tuning
		TL-1	TL-2	TL-3	TL-4	TL-5	
RMSE (mV)	2.984	2.315	2.319	1.895	2.734	/	2.151
MAPE (%)	0.291	0.227	0.223	0.183	0.267	/	0.208
参数 ($\times 10^5$)	1.49	1.48	1.15	0.49	0.16	/	1.49
MSE ($\times 10^{-4}V^2$)	4.126	3.239	2.693	2.571	2.290	2.266	2.479
R ²	0.9981	0.9985	0.9987	0.9988	0.9989	0.9990	0.9988
参数 ($\times 10^6$)	3.60	3.58	3.45	3.28	2.75	2.10	3.60

实验表明，不同架构下的最优冻结深度不同，但总体趋势一致，表明迁移策略在不同链路结构间具有良好的一致性。当目标域数据量有限时，建议优先冻结约 50% 的总层数或残差块，并根据验证结果进一步调整冻结深度，以实现性能与训练效率的平衡。

5 结论

本文提出了一种融合快速时域仿真与迁移学习的建模方法，以应对高速信号完整性分析中面临的两大挑战：瞬态仿真与电磁仿真计算开销大与机器学习建模数据依赖性强的问题。源域模型采用电路仿真与 MER 方法高效生成最坏眼图，目标域模型基于 EM 提取 S 参数并进行瞬态仿真以获得高精度响应。面向两种经典任务分别构建代理模型：用于几何参数到眼高预测的 DNN 模型，以及用于输入输出波形建模的 TCN 模型。通过多种迁移策略实现知识重用，显著降低数据需求和训练成本，并在多任务和不同样本量条件下验证了方法的鲁棒性与扩展性。在眼高预测任务中，DNN-TL 模型的 RMSE 降低了 30%~40%；在波形映射任务

中, TCN-TL 架构相较于 TCN、GRU 和 LSTM 基线模型 MSE 降低了约 70%, 且预测精度达到 $R^2 \approx 0.999$ 。本方法突破了传统数据密集型建模范式的限制, 在小样本场景下展现出卓越的预测精度和收敛速度, 同时在不同电路结构和任务类型中表现出良好的通用性与可扩展性。

参考文献

- [1] Fan J, Ye X, Kim J, Archambeault B, Orlandi A. Signal integrity design for high-speed digital circuits: progress and directions[J]. IEEE Transactions on Electromagnetic Compatibility, 2010, 52(2): 392-400.
- [2] Feng F, Na W, Jin J, Zhang J, Zhang W, Zhang Q-J. Artificial neural networks for microwave computer-aided design: The state of the art[J]. IEEE Transactions on Microwave Theory and Techniques, 2022, 70(11): 4597-4619.
- [3] Mao C-S, Wang D-W, Zhao W-S, Hu Y. Pseudo-labeling based semi-supervised learning for signal integrity analysis of high-bandwidth memory (HBM) interposer[J]. IEEE Transactions on Electromagnetic Compatibility, 2024, 66(6): 2056-2064.
- [4] Li X-R, Zhang P, Wang D-W, Liu J, Sun L, Zhao W-S. Equalizer optimization method based on local multi-constraint modeling-Bayesian optimization with region partitioning[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2025, early access.
- [5] Lu T, Sun J, Wu K, Yang Z. High-speed channel modeling with machine learning methods for signal integrity analysis[J]. IEEE Transactions on Electromagnetic Compatibility, 2018, 60(6): 1957-1964.
- [6] Lho D, et al. Channel characteristic-based deep neural network models for accurate eye diagram estimation in high-bandwidth memory (HBM) silicon interposer[J]. IEEE Transactions on Electromagnetic Compatibility, 2022, 64(1): 196-208.
- [7] Lu R, et al. A novel sequence-to-sequence-based deep learning model for multistep load forecasting[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 638-652.
- [8] Luo Y, Chu X, Yuan H, Wei T, Wang J, Wu F, Li Y. Fast response prediction method based on bidirectional long short-term memory for high-speed links[J]. IEEE Transactions on Microwave Theory and Techniques, 2023, 71(6): 2347-2359.
- [9] Dong X, et al. SPIRAL+: Efficient signal-power integrity co-analysis for inter-chiplet links validation[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2025: 1-1.
- [10] Zhang H H, Xue Z S, Liu X Y, Li P, Jiang L, Shi G M. Optimization of high-speed channel

- for signal integrity with deep genetic algorithm[J]. *IEEE Transactions on Electromagnetic Compatibility*, 2022, 64(4): 1270-1274.
- [11] Ma H, Li D, Li E-P, Cangellaris A C, Chen X. A fast optimization method for high-speed link inverse design with SVR-AS algorithm[J]. *IEEE Transactions on Signal and Power Integrity*, 2022, 1: 22-31.
- [12] Ma H, et al. Uncertainty quantification of signal integrity analysis for neuromorphic chips[J]. *IEEE Transactions on Signal and Power Integrity*, 2022, 1: 160-169.
- [13] Casper K, Haycock M, Mooney R. An accurate and efficient analysis method for multi-Gb/s chip-to-chip signaling schemes[C]. *Symposium on VLSI Circuits Digest of Technical Papers*, 2002: 54-57.
- [14] Yu W, Shi R, Cheng C-K. Accurate eye diagram prediction based on step response and its application to low-power equalizer design[J]. *IEICE Transactions on Electronics*, 2009, E92-C(4): 444-452.
- [15] Shi R, Yu W, Zhu Y, Cheng C-K, Kuh E S. Efficient and accurate eye diagram prediction for high speed signaling[C]. *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 2008: 655-661.
- [16] Ren J, Oh K S. Multiple edge responses for fast and accurate system simulations[J]. *IEEE Transactions on Advanced Packaging*, 2008, 31(4).
- [17] Chu X, Li N, Wang J, Luo Y, Wu F, Li Y. Calculation of the worst-case data patterns and eye diagram for nonlinear high-speed links[J]. *IEEE Transactions on Electromagnetic Compatibility*, 2021, 63(6): 2159-2163.
- [18] Wang F. Knowledge-based neural models for microwave design[J]. *IEEE Transactions on Microwave Theory and Techniques*, 1997, 45(12): 2378-2384.
- [19] Bandler J W, Ismail M A, Rayas-Sanchez J E, Zhang Q-J. Neuromodeling of microwave circuits exploiting space-mapping technology[J]. *IEEE Transactions on Microwave Theory and Techniques*, 1999, 47(12): 2417-2427.
- [20] Ma J, Dang S, Li P, Watkins G, Morris K, Beach M. Transfer learning for the behavior prediction of microwave structures[J]. *IEEE Microwave and Wireless Components Letters*, 2023, 33(2): 126-129.
- [21] Zhang T, Chen S, Wei S, Chen J. A data-efficient training model for signal integrity analysis based on transfer learning[C]. *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, 2019: 186-189.
- [22] Ma J, Dang S, Watkins G, Morris K, Beach M. A high-performance transfer learning-based

- model for microwave structure behavior prediction[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2023, 70(12): 4394-4398.
- [23] Qiu J, Ma H, Zhang F, Sun Z, Li E-P. Transient modeling of high-speed links using transfer learning-based neural network initialization[J]. IEEE Transactions on Electromagnetic Compatibility, 2024, 66(6): 2065-2073.
- [24] Guglani S, et al. Artificial neural networks with fast transfer learning for statistical signal integrity analysis of MWCNT and MLGMR interconnect networks[J]. IEEE Transactions on Electromagnetic Compatibility, 2024, 66(3): 939-948.
- [25] Chou C-C, Chuang H-H, Wu T-L, Weng S-H, Cheng C-K. Eye prediction of digital driver with power distribution network noise[C]. IEEE 21st Topical Meeting on Electrical Performance of Electronic Packaging, 2012: 131-134.
- [26] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.
- [27] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?[C]. Advances in Neural Information Processing Systems, 2014.
- [28] Ma H, Qiu J, Sheng G, Chen W, Li E-P. Deep simple recurrent unit-based transient modeling method for high-speed circuits[J]. IEEE Transactions on Microwave Theory and Techniques, 2025, 73(2): 736-744.
- [29] Yuan X, Qi S, Wang Y, Wang K, Yang C, Ye L. Quality variable prediction for nonlinear dynamic industrial processes based on temporal convolutional networks[J]. IEEE Sensors Journal, 2021, 21(18): 20493-20503.
- [30] X. Zhou, C. Pang, X. Zeng, L. Jiang, and Y. Chen, A short-term power prediction method based on temporal convolutional network in virtual power plant photovoltaic system[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.
- [31] Li Z, Li X-C, Wu Z-M, Zhu Y, Mao J-F. Surrogate modeling of high-speed links based on GNN and RNN for signal integrity applications[J]. IEEE Transactions on Microwave Theory and Techniques, 2023, 71(9): 3784-3796.
- [32] Cao Y, Zhang Q-J. A new training approach for robust recurrent neural-network modeling of nonlinear circuits[J]. IEEE Transactions on Microwave Theory and Techniques, 2009, 57(6): 1539-1550.
- [33] Li Y, Yu H, Li E. Signal integrity analysis of neuronal spike signal in 3-D packaging[J]. IEEE Transactions on Signal and Power Integrity, 2023, 2: 84-93.

具有超宽阻带性能的紧凑型双模 SIW 带通滤波器

王倩颖¹, 张傲^{2,3*}, 高建军^{1*}

(1.华东师范大学, 物理与电子科学学院, 上海, 200241;

2.南通大学, 微电子学院, 江苏, 南通, 226019;

3.新加坡南洋理工大学 电气与电子工程学院, 新加坡 639798)

摘要: 在现代通信系统中, 滤波器作为射频前端的重要组成部分, 不仅需具备优异的通带特性, 还应具有宽阻带性能, 从而抑制系统内外的干扰信号。本文提出了一种基于刻蚀矩形槽的基片集成波导带通滤波器, 结合 C 形分裂环槽、金属化通孔扰动以及带耦合槽的共面波导馈电结构, 有效提升了滤波器的阻带抑制范围。仿真与测试结果表明, 该滤波器实现了超过 30 dB 的上阻带抑制, 频率抑制比高达 $9.84 f_0$, 展现出优异的宽阻带特性和工程应用潜力。

关键词: 宽阻带; 带通滤波器; 基片集成波导; 矩形槽

1 引言

在现代通信系统中, 高性能微波滤波器是射频前端的关键部件, 不仅需具备优异的通带特性, 还应具有良好的宽阻带性能, 以有效抑制系统内外的干扰信号。然而, 高阶模的存在往往导致杂散通带, 影响滤波器的整体性能。针对这一问题有很多种高阶模抑制技术。例如, 串联低通滤波器可实现通带与阻带的独立设计与调节, 但通常会带来插入损耗增加与尺寸扩大的问题^[1]; 通过引入传输零点, 可在特定频率处形成零陷, 从而有效抑制杂散通带^[2]; 采用高阻抗比的阶梯阻抗谐振器 (SIR) 也能显著上移二次谐波频率, 改善滤波器的带外响应^[3]。

近年来, 基于槽结构的扰动技术逐渐受到关注。例如, 多层槽阵列可显著拓宽阻带范围^[4]; 而通过调节径向槽的位置与长度, 可实现多模谐振频率的独立控制, 从而提升滤波器的频率选择性^[5]。尽管上述方法在一定程度上改善了高阶模抑制效果, 但单一手段往往难以兼顾结构紧凑性与宽阻带性能, 仍存在一定的设计瓶颈。

本文提出一种新型紧凑型双模基片集成波导带通滤波器。该结构通过在腔体顶部刻蚀 C 形分裂环槽, 有效降低 TM_{10} 模的谐振频率, 增大其与高阶模之间的频率比; 引入中心金属化通孔以激发 TM_{00} 模, 实现双模工作并拓展阻带范围; 采用带耦合槽的共面波导馈电进一步提升阻带抑制性能。实测结果表明, 该滤波器实现了最高达 $9.84 f_0$ (>30 dB) 的超宽上阻带抑制比, 在现代无线通信系统中具有良好的干扰抑制应用潜力。

2 分析与设计

2.1 改进型 SIW 腔体结构与本征模分析

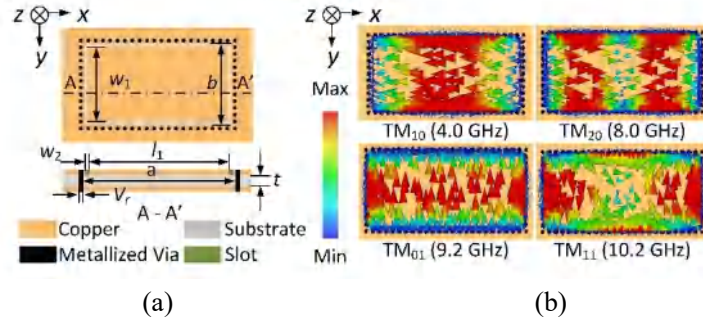


图 1 (a) 所提结构的俯视图与剖面图 (b) 前四阶谐振模的电流分布

图 1(a)展示了所提出的改进型 SIW 腔体结构示意图。腔体上表面刻蚀的矩形槽形成磁壁，槽边沿布置的金属化通孔用于连接上下金属层，从而防止电磁泄漏并提高腔体的品质因数。基于本征模分析，图 1(b)给出了前四阶模的电流分布情况。由于腔体厚度远小于腔体的长和宽，因此仅需考虑横向磁模，其模式可表示为 TM_{mn} ，各模式谐振频率 $f_{TM_{mn}}$ 可由下式确定：

$$f_{TM_{mn}} \approx \frac{c_0}{2\sqrt{\epsilon_r \mu_r}} \sqrt{\left(\frac{m}{a_{eff}}\right)^2 + \left(\frac{n}{b_{eff}}\right)^2} \quad (1)$$

其中， c_0 为真空中的光速， ϵ_r 和 μ_r 分别为介质的相对介电常数和磁导率， m 和 n 为腔体沿 x 轴与 y 轴方向的模态阶数， a_{eff} 和 b_{eff} 分别代表腔体的等效电长度与等效电宽度。

2.2 扰动加载与模式调控机制

不同于文献^{[4][5]}中仅侧重于抑制杂散通带的策略，本文通过引入 C 形分裂环槽，有效降低了主模 TM_{10} 的谐振频率，从而显著提升其与高阶模之间的频率比。由于槽结构的加载使电流路径发生偏转，腔体的等效电长度 a_{eff} 增加，导致 TM_{10} 和 TM_{20} 模的谐振频率明显下降，而其他高阶模则受槽影响较小。如图 2(a)所示，随着槽长 l_2 的增加，除 TM_{20} 模外，其余高阶模与基模之间的频率比均呈上升趋势，显著拓宽了低阶模与高阶模之间的阻带范围。

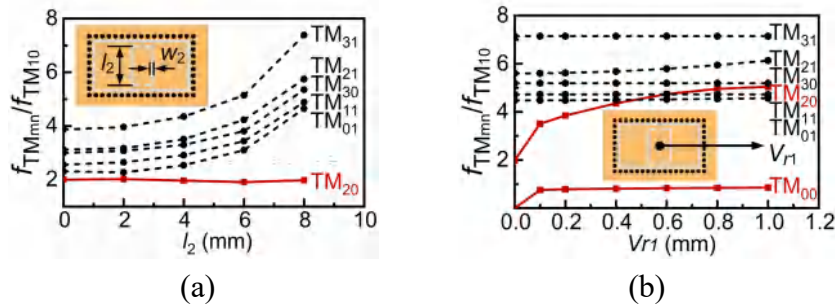


图 2 (a) 高阶模与基模之间的频率比随 C 形分裂环槽长度 l_2 变化的关系曲线 (b) 高阶模与基模之间的频率比随中心金属化通孔半径 V_{r1} 变化的关系曲线

为进一步抑制 TM_{20} 模并扩展阻带，本文在腔体中心引入金属化通孔作为短路点。该扰动对偶模影响显著，而对奇模影响较小。如图 2(b)所示，随着通孔半径 r_1 的增大，偶模的谐振频率明显上移，而奇模几乎保持不变。最终 TM_{01} 模成为新的“二次”谐波，其谐振频率约为基模频率的 4.42 倍。此外，金属化通孔还激发出一个频率接近 TM_{10} 模的 TM_{00} 模，使腔体实现双模共振，不仅拓展了通带，还提升了通带内的平坦度，进一步优化了滤波器的带内性能。

2.3 馈电结构与性能优化

依据镜像阻抗原理^[6]，采用如图 3(a)所示的带耦合槽共面波导 CPW 结构作为馈电方式。当耦合槽长度设计为高次模谐振频率处的半导波波长时，可实现对该模式的有效抑制。具体而言，将耦合槽长度设置为 $f_1 (= 10.8 \text{ GHz})$ 处的半个导波波长时， TM_{20} 与 TM_{30} 模可被显著抑制，从而有效拓展了滤波器的上阻带范围并提升了带外抑制性能。在滤波器优化设计中，耦合矩阵至关重要，图 3(b)展示了所采用的耦合拓扑结构，其耦合系数根据公式(2-4)^{[7][8]}进行计算。其中如图 3(c)所示，在 CPW 馈电结构中，外部品质因数 Q_e 可通过调节 CPW 缝隙宽度 w_4 或嵌入式耦合槽长度 l_3 来实现控制，且增加耦合槽长度和缝隙宽度可有效提高 Q_e ，进而减小输入/输出耦合强度。由于所设计的滤波器是对称结构，满足 $|M_{S1}| = |M_{1L}|$ 且 $|M_{S2}| = |M_{2L}|$ 。此外，奇模与偶模之间的耦合系数为零。同时，由于奇模 TM_{10} 与外部端口的耦合强度始终大于偶模 TM_{00} ，因此有 $|M_{S1}| > |M_{S2}|$ ，即 $Q_{e1} < Q_{e2}$ 。

$$Q_e = \frac{\omega_0 \times \tau_{11}(\omega_0)}{4} \quad (2)$$

$$M_{Sn} = \frac{1}{\sqrt{Q_{en} \times FBW}} \quad (3)$$

$$M_{mn} = (1 - \frac{f_n}{f_0}) \times \frac{2}{FBW} \quad (4)$$

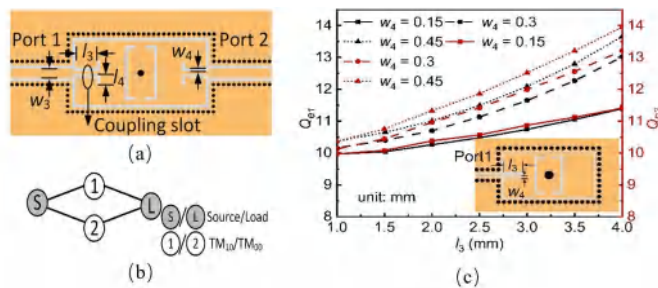


图 3 (a)滤波器的几何结构示意图 (b) 耦合矩阵(c) 外部品质因数随缝隙宽度和嵌入式耦合槽长度变化曲线

3 测试与结果

测试采用 Agilent E8363C 矢量网络分析仪，所测结果包含 2.4 mm 端接连接器的影响。腔体采用单层 Rogers 4003 板材加工，板厚为 0.2 mm，相对介电常数为 3.38，损耗正切为 0.0027。



图 4 电磁仿真、耦合矩阵合成、实测的 S 参数结果对比及实物照片。

图 4 展示了电磁仿真、耦合矩阵合成及实测的 S 参数结果，并附有原型实物图。从图中可见，三者吻合良好，仅在插入损耗方面存在约 0.3 dB 偏差，主要来源于连接器损耗及铜面粗糙度。滤波器的实测中心频率为 1.9 GHz，3 dB 带宽为 23.1%，插入损耗为 0.96 dB，阻带抑制范围达 18.7 GHz。

4 结论

本文提出了一类新型混合扰动型带通滤波器，在基片集成波导（SIW）腔体中引入矩形槽结构，通过在腔体顶部刻蚀 C 形分裂环槽，有效降低了基模 TM_{10} 的谐振频率，从而显著提升了其与高阶模之间的频率比。结合金属化通孔与带耦合槽的馈电结构，所设计的滤波器最终实现了高达 $9.84 f_0$ 的超宽阻带抑制性能，能适用于现代无线通信系统中的干扰抑制场景。

参考文献

- [1] Yun T S, Nam H, Kim J Y, et al. Harmonics suppressed substrate-integrated waveguide filter with integration of low-pass filter[J]. Microwave and Optical Technology Letters, 2008, 50(2): 447-450.
- [2] Tang C W, Chen M G. Wide stopband parallel-coupled stacked SIRs bandpass filters with open-stub lines[J]. IEEE microwave and wireless components letters, 2006, 16(12): 666-668.
- [3] Kuan H, Lin Y L, Yang R Y, et al. A multilayered parallel coupled microstrip bandpass filter with embedded SIR cells to have a broad upper rejection band[J]. IEEE microwave and wireless components letters, 2009, 20(1): 25-27.
- [4] Chu P, Feng J, Zhu P, et al. Substrate Integrated Waveguide Filter With Wide Stopband Up to $(2k + 3) f_0$ [J]. IEEE Transactions on Microwave Theory and Techniques, 2023, 71(12): 5358-5366.
- [5] Xie H Y, Wu B, Xia L, et al. Miniaturized half-mode fan-shaped SIW filter with extensible order and wide stopband[J]. IEEE microwave and wireless components letters, 2020, 30(8): 749-752.
- [6] Liu Y W, Lu Y J, Hsu P. Harmonic suppressed slot loop antenna fed by coplanar waveguide[J]. IEEE Antennas and Wireless Propagation Letters, 2014, 13: 1292-1295.
- [7] Fu S, Wu B, Chen J, et al. Novel second-order dual-mode dual-band filters using capacitance loaded square loop resonator[J]. IEEE Transactions on Microwave Theory and Techniques, 2012, 60(3): 477-483.
- [8] Hong J S G, Lancaster M J. Microstrip filters for RF/microwave applications[M]. John Wiley & Sons, 2004.

低寄生电感 SiC 模块双面冷却结构协同设计

李温昊, 赵文生, 王大伟, 于成浩

杭州电子科技大学

摘要: 双面冷却封装因其低寄生电感与优越散热特性, 已成为碳化硅 (SiC) 功率模块封装的重要研究方向。本文提出一种低电感、低热阻、高可靠性双面冷却封装设计。首先, 引入回路间负耦合效应, 显著抑制电流回路寄生电感效应; 其次, 提出了上下交错芯片布局设计, 有效地改善了芯片间热耦合效应均衡性、提升了封装等效热通量; 最后, 结合遗传算法与电-热多物理场仿真, 对关键几何参数进行协同优化, 获得寄生电感与热性能均衡的最优布局。

关键词: 双面冷却; 寄生电感; 负耦合; 热优化

1 前言

碳化硅 (SiC) 功率模块因其高击穿场强与优异热导率等本征物理特性, 在电力电子系统高性能化进程中展现出显著技术优势。作为宽禁带半导体材料, SiC 的禁带宽度、临界击穿电场及热导率等关键参数均超越传统硅 (Si) 材料, 使其功率器件在高温、高频及高压工况下具有代际性能优势。传统封装材料和结构已难以满足 SiC 器件性能需求。在传统单面冷却 (SSC) 功率模块中, 功率器件裸片底部键合于直接覆铜 (DBC) 基板实现底部电气互连与散热, 顶部则依赖键合线完成电气互连。然而, 键合线导致模块高寄生电感, 且连接处热膨胀系数 (CTE) 失配与长期高温波动使键合强度劣化, 同时阻碍器件顶面散热。这些缺陷导致 SiC 器件的物理性能优势无法被充分释放。

相较而言, 双面冷却 (DSC) 封装采用平面互连替代键合线, 可显著改善封装结构的电学及热力学性能, 已成为碳化硅功率模块封装研究的重要方向与热点。理论表明, DSC 模块可将器件至环境的热阻降低 50%——更低的热阻提升模块电流承载能力, 增强的热管理支持更高电流密度; 同时紧凑布局缩小模块体积, 实现系统级功率密度跃升。文献[1]提出了一种紧凑交错封装方法, 在单芯片功率模块低感化设计领域取得进展。实验验证其功率回路总寄生电感降至 2.92 nH。但该研究并未涉及多芯片并联场景。文献[2]提出了一种多芯片并联场景下的交错封装方法。实验验证该方法功率回路总寄生电感为 3.8 nH。

2 方法

2.1 低寄生电感设计

在半桥功率模块的拓扑构型中, 直流母线端子 (dc+至 dc-) 间形成的功率回路寄生电感本质上是制约高频开关性能的核心瓶颈。当功率模块采用多芯片并联

架构时，电流通路的几何分布将构成复杂的电磁耦合网络：该网络包含各并联支路内部的自感分量以及支路间的互感分量。通过电磁场优化设计策略，对并联桥臂实施特定的空间取向调控与相位对齐布局，可促使相邻电流回路产生的瞬变磁通呈现反向叠加特性。这种基于楞次定律的磁通抵消机制，能在不引入额外磁屏蔽的前提下，显著削弱回路总磁链，从而有效降低功率回路的等效寄生电感^[3]。

2.2 热性能均衡设计

采用以几何中心对称的芯片布局方案，同时将上下半桥臂芯片实施空间交错排列。此结构设计显著优化了多芯片间的热耦合均衡性：中心对称性消除了热传导路径差异引起的温度梯度畸变，而桥臂交错排列则有效阻断相邻芯片的热干涉区，降低局部热积聚风险。两者协同作用重构了封装内部热流分布，通过增强热场的空间均匀性与热通量传输效率，整体提升了模块的等效散热性能。

3 结果分析

3.1 电学性能分析

基于电磁场数值分析方法，利用 Ansys Q3D 软件对所设计功率模块的电学性能进行高精度建模仿真。图 1(a)与(b)分别为模块底板铜层和顶板下铜层的表面电流密度空间分布特征，10MHz 激励频率下提取的功率回路总寄生电感参数为 1.589 nH。

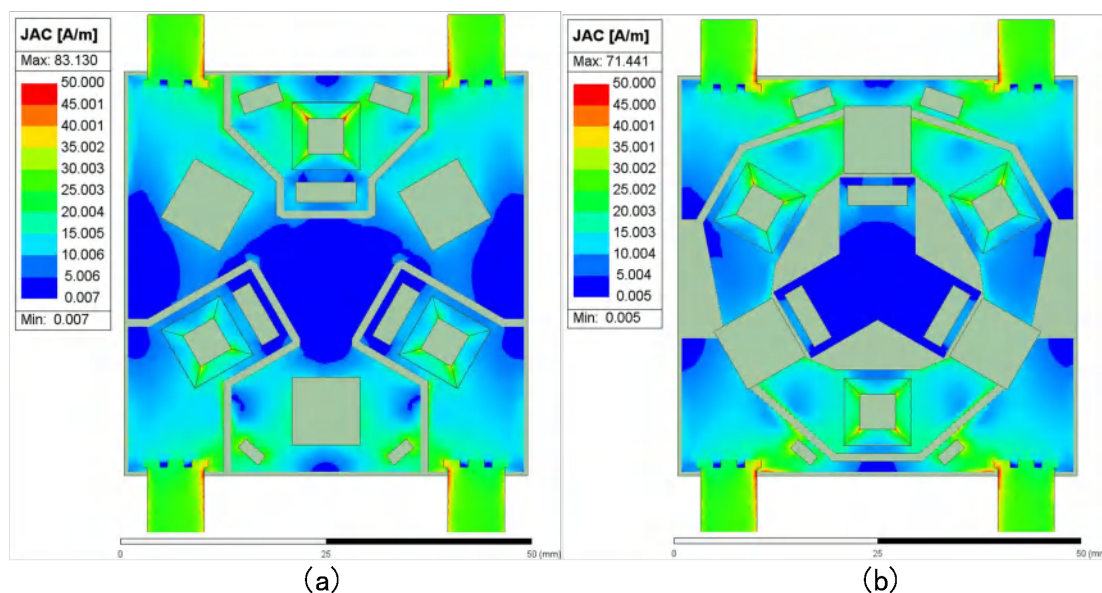


图 1 功率模块电学仿真结果

3.2 热学性能分析

图 2 展示了基于 Ansys 稳态热传导模型的仿真结果。其中图 2(a)与(b)分别量化表征了模块顶面和底面的三维温度场分布特性。仿真数据表明：中心对称的芯片布局结合上下桥臂交错排列的策略，显著优化了多芯片热耦合均衡性。

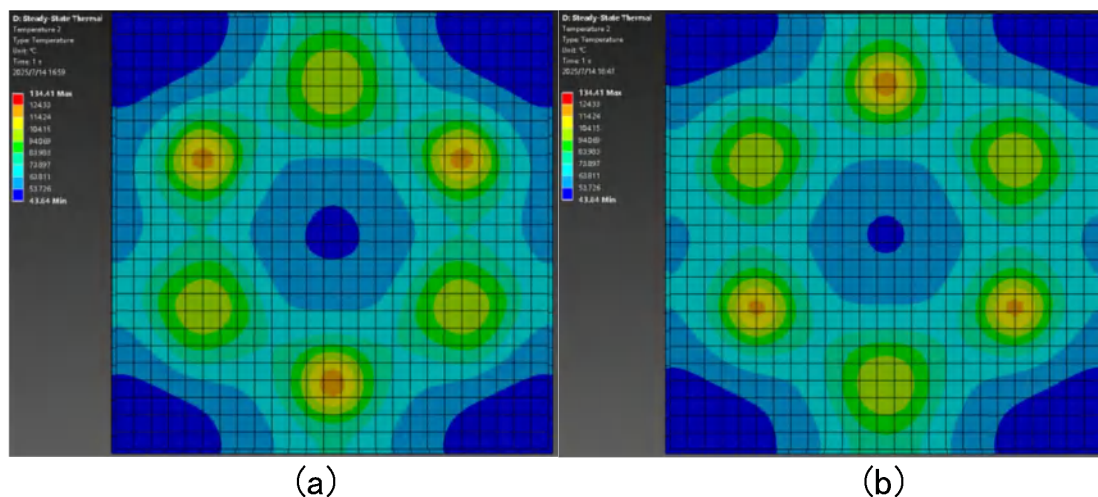


图 2 上下表面热仿真结果图

4 结论

本文通过引入回路间负耦合效应、上下交错芯片布局设计，显著降低了功率回路总寄生电感，有效地改善了芯片间热耦合效应均衡性，实现了低电感、低热阻、高可靠性双面冷却封装设计。然而，本文设计的布局结构仍存在缺陷：各功率支路寄生电感不同，导致静态、动态均流失衡问题。

参考文献

- [1] Yang F, Wang L, Kong H, et al. Compact-Interleaved Packaging Method of Power Module With Dynamic Characterization of 4H-SiC MOSFET and Development of Power Electronic Converter at Extremely High Junction Temperature[J]. IEEE Transactions on Power Electronics, 2023, 38(1): 417-434.
- [2] Yang F, Jia L, Wang L, et al. Interleaved Planar Packaging Method of Multichip SiC Power Module for Thermal and Electrical Performance Improvement[J]. IEEE Transactions on Power Electronics, 2022, 37(2): 1615-1629.
- [3] Wang J, Yu S, Zhou W. A Comprehensive Design Method for Multichip Double-Sided Cooling Power Module With Multidimensional Self-and Mutual Inductances[J]. IEEE Transactions on Power Electronics, 2024, 39(8): 9526-9539.

SiC 半桥功率模块电热力协同设计优化

与可靠性仿真

王晓宇, 赵文生, 王大伟, 于成浩

杭州电子科技大学

摘要: 双面冷却堆叠封装结构低寄生、低热阻、高功率密度等优点使其成为高电热性能和高可靠性碳化硅功率模块设计的基础架构。本文提出了一种低电感、低热阻、低应变双面冷却堆叠封装半桥单元设计, 通过回路互感耦合设计显著降低了寄生电感, 回路采用芯片上下交错布置有效优化散热路径。在此单元基础上构建全对称四半桥并联结构, 通过物理路径的对称有效缓解芯片的不均流问题。进一步建立二维封装截面模型, 获得粘塑性应变最小的最优结构, 实现封装热-力性能的协同优化。

关键词: 双面冷却; 寄生电感; 负耦合; 热优化

1 前言

双面冷却堆叠封装结构因其低寄生电感、低热阻与高功率密度等优势, 已成为高电热性能与高可靠性碳化硅 (SiC) 功率模块设计的关键架构。该技术通过取消传统线焊并引入平面互连结构, 实现芯片上下表面的同时散热, 显著增强了热管理能力, 降低了寄生电感, 电流承载能力提升可达 30%–60%, 有效突破了电动汽车逆变器在热管理与电气性能方面的瓶颈^[1]。

已有研究如文献^[2]在双面冷却结构基础上提出交错功率环路封装设计, 尽管在降低寄生电感方面取得进展, 并分析了源极寄生电感引起的均流问题, 但未能彻底解决模块的电流均衡性。文献^[3]提出多芯片交错式平面封装方法, 在改善上下层散热均匀性、降低耦合热阻方面具有一定优势, 然而其结构复杂, 导致较高的寄生电感, 限制了器件的开关性能, 且未能完全消除热耦合引起的芯片表面温度不均问题。文献^[4]则从材料角度出发, 研究了不同封装材料组合下的热场与热应力分布, 获得了适用于高功率密度条件下的材料最优组合, 但对其物理机理缺乏深入分析, 同时未涉及在工艺可行性约束下材料厚度对热-力性能的影响。

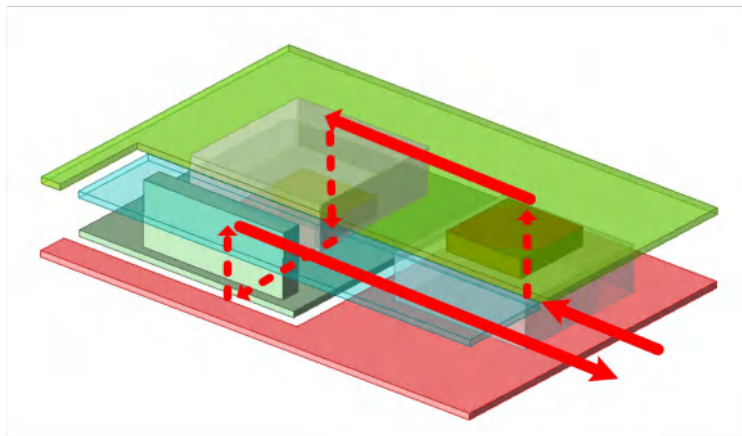
针对上述问题, 本文提出了一种兼具低寄生电感、低热阻与高可靠性的双面冷却封装结构, 并基于贝叶斯优化算法对模块各材料层厚度进行了全局寻优, 显著提升了模块的热-力耦合性能和整体可靠性。

2 方法

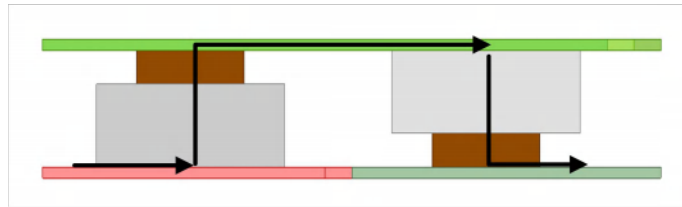
2.1 双面冷却全对称布局

为提升功率模块的使用寿命,避免因局部芯片温度过高导致的提前老化和性能退化,本文提出一种基于双面冷却的全对称布局结构。该结构通过物理路径的对称性,有效缓解封装寄生电感引起的芯片间不均流问题,同时削弱热耦合效应。通过优化芯片排布,实现上下半桥芯片间温度的均衡分布,从源头抑制局部过热,进一步提升电流均衡性与模块的热可靠性。

图 1 所示为所设计的功率模块单元半桥结构,完整的全对称四半桥结构由该半桥回路沿 X 轴和 Y 轴对称复制构建。下文将在此单元半桥结构的基础上开展寄生电感特性分析。



(a) 单元半桥结构



(b) 半桥截面图

图 1 单元半桥结构图

2.2 电感负耦合

为进一步降低双面冷却功率模块的寄生电感,本文引入电感负耦合设计思想。通过优化电流路径布局,使功率回路间形成多个方向相反的磁通通量,从而产生显著的负互感效应。这种磁通相互抵消的机制,有效削弱了总回路感应磁场,降低了等效寄生电感,进而提升器件的开关性能与电磁兼容性。

3 结果分析

3.1 寄生电感分析

利用 3D 电磁仿真软件对设计的模块进行寄生电感仿真,仿真结果见下表 1 中,其中, L_D 是指从 DC+端子到上半桥芯片的寄生电感, L_O 指 AC 端口两端的

寄生电感， L_s 指从下半桥芯片到 DC-端口的寄生电感。结果表明，功率模块回路中存在多重负耦合路径，如电感矩阵中明显的负互感项(如 D-S 和 O-S 之间)，体现出电流路径之间的磁通抵消效应。这种多点分布的负耦合显著降低单个半桥的等效寄生电感，有助于抑制电压尖峰、提升模块的开关性能与系统电磁兼容性。

表 1 单个半桥的寄生电感矩阵

电感 (nH)	100MHZ	L_D	L_O	L_S	总计
L_D		0.66	0.04	-0.36	0.34
L_O		0.04	1.58	-1.08	0.54
L_S		-0.36	-1.08	5.61	4.17

从表中可以得到回路总寄生电感为

$$L = L_D + L_O + L_S = 0.34 + 0.54 + 4.17 = 5.05nH \quad (1)$$

由此，四半桥模块总的寄生电感为：

$$L_{total} = \frac{L}{4} = 1.26nH \quad (2)$$

但是由于四个半桥模块之间的正耦合的存在，实际 3D 电磁仿真软件仿真出的四个半桥的总寄生电感为 1.47nH，略高于理论值。

3.2 热耦合分析

由于采用全对称结构，功率模块中上半桥芯片分布于模块四角，下半桥芯片则集中于模块中心，整体布局呈现中心对称特征。这种结构使上下半桥内部各芯片处于相似的热耦合环境中，有效实现了上半桥芯片之间以及下半桥芯片之间的温度均衡，有助于抑制局部过热现象，提升模块的热可靠性。图 2 展示了基于有限元仿真软件所获得的模块稳态热分布结果。仿真设定的热边界条件为：每个芯片功耗 50 W，环境温度为 25 °C，上下基板的传热系数均为 5000 W/(m²·K)。结果表明，全对称布局显著缓解了芯片间热耦合导致的温度不均，成功将上下半桥的芯片温度控制在同一水平，有效提升了模块整体的热稳定性与可靠性。

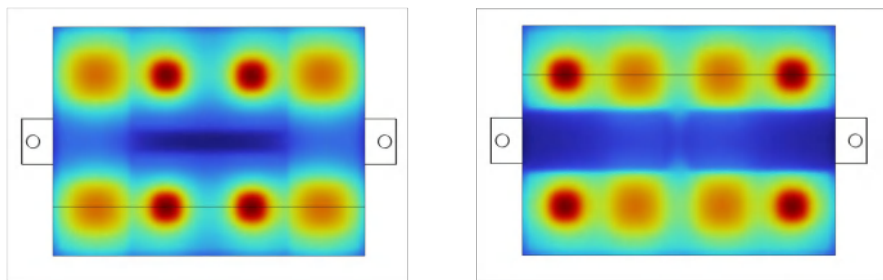


图 2 上下基板表面温度分布图

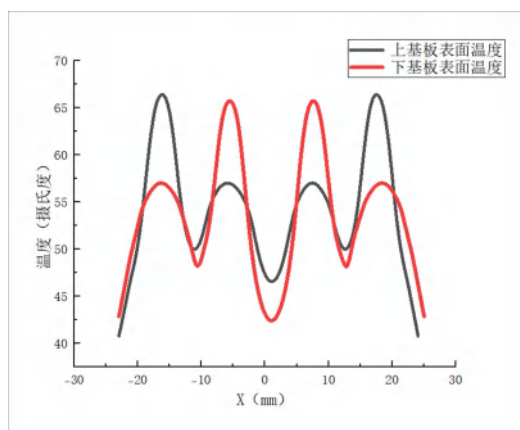


图 3 上下基板截线上的温度

3.3 贝叶斯优化

基于构建的双面冷却二维封装截面模型，采用贝叶斯优化算法对各封装材料的厚度在给定约束条件下进行全局寻优，获得热—力性能最优的结构参数组合。仿真结果表明，该优化方案有效降低了功率模块的最高结温与最大热应变，显著提升了封装的热—力协同性能。

4 结论

本文针对高功率密度应用中双面冷却 SiC 功率模块在电热性能与可靠性方面面临的挑战，提出了一种具备低寄生电感、低热阻与高热力协同性能的全新封装结构。通过引入全对称芯片布局设计，有效缓解了热耦合与不均流问题；基于电感负耦合原理优化电流路径，显著降低模块等效寄生电感，提升电气性能；构建二维封装模型并结合贝叶斯优化算法，在厚度约束条件下实现多层结构的全局优化，进一步减小最高结温与热应变，提升模块热—力可靠性。上述研究为高性能 SiC 功率模块的封装结构设计提供了理论依据与优化方法支持，对推动电动汽车功率器件的集成化、高可靠发展具有重要意义。

参考文献

- [1] Liu M, Coppola A, Alvi M, *et al.* Comprehensive Review and State of Development of Double-Sided Cooled Package Technology for Automotive Power Modules[J]. IEEE Open Journal of Power Electronics, 2022, 3: 271-289.
- [2] Yan Y, Liu B, Lv J, 等. A Novel Double-Sided Cooling Silicon Carbide Power Module With Ultralow Parasitic Inductance Based on an Interleaved Power Loop[J]. IEEE Transactions on Power Electronics, 2024, 39(10): 12570-12588.
- [3] Yang F, Jia L, Wang L, *et al.* Interleaved Planar Packaging Method of Multichip SiC Power Module for Thermal and Electrical Performance Improvement[J]. IEEE Transactions on Power Electronics, 2022, 37(2): 1615-1629.
- [4] Ren Z, Guo X, Fu J, *et al.* A Double-sided Cooling SiC Power Module Applied to Electric

Vehicles[C]//2023 26th International Conference on Electrical Machines and Systems (ICEMS). 2023: 5143-5148.

基于金属光栅-锗-金属镜结构的光电探测器仿真研究

孙立新^{1,2}, 陈飞良^{1,2*}, 张健^{1,2,3}, 李沫^{1,2,3*}

1. 电子科技大学电子科学与工程学院 成都 611731

2. 电子科技大学先进毫米波技术集成攻关研究院 成都 611731

3. 太赫兹辐射与应用四川省重点实验室 成都 611731

flchen@uestc.edu.cn, limo@uestc.edu.cn

摘要:面向太赫兹混频产生, 本文提出一种基于金属光栅-锗-金属镜 (MG-Ge-Mirror) 结构的光电探测器。顶部亚波长光栅通过局域表面等离子体共振 (LSPR) 与异常光学透射 (EOT) 效应增强光场局域化, 底部金属镜则与光栅构成法布里-珀罗微腔, 显著提升超薄锗的光吸收效率。数值仿真结果表明, 该结构在 1550 nm 处具有高量子效率 (EQE=73.8%) 与超快的响应速度 (37.6 fs)。光栅同时作为偏置电极, 结合共面波导 (CPW) 集成方案, 有望提高传统光电探测器的功率-带宽性能。

关键词: 光电探测器; 锗; 亚波长光栅; 金属镜

1 引言

太赫兹 (THz) 波因其独特的光谱特性, 在高速通信、无损检测及生物医学成像等领域展现出重大应用价值^[1]。光混频技术是产生连续可调谐太赫兹波的重要方案之一, 其转换效率直接受限于光电探测器的性能^[1,2]。其中, 金属-半导体-金属 (Metal-Semiconductor-Metal, MSM) 光电二极管结构因其易制造和低电容特性被广泛研究^[3-7]。锗 (Ge) 因其在近红外 (特别是 1550 nm) 的高吸收系数 ($\alpha \approx 10^5 \text{ cm}^{-1}$ @1550 nm) 及高载流子迁移率 ($\mu_e \approx 3900 \text{ cm}^2/\text{V}\cdot\text{s}$), 是实现硅基高速光电探测的理想材料^[8]。然而, 传统锗基 MSM 探测器难以兼顾超薄吸收层 (实现高速) 与高效率 (克服电极遮光和锗在 1550 nm 吸收相对较弱的问题)。

为解决以上挑战, 本文提出一种基于金属光栅-锗-金属镜 (MG-Ge-Mirror) 结构 (图 1) 的光电探测器。首先, 锗层表面的亚波长光栅不仅作为垂直偏置电极, 更通过异常光学透射 (EOT) 和局域表面等离子体共振 (LSPR) 效应, 有效增强光透射并将光子能量局域于光栅间隙下方的纳米级区域内, 实现超薄锗层内的宽谱近场光场增强。其次, 该结构在锗衬底底部集成高反射金属镜, 与表面金属光栅共同构成法布里-珀罗 (Fabry-Pérot) 光学微腔, 入射光在光栅与镜面间经历多次反射, 有效增加光程。该微腔结构可使超薄吸收层内的等效光程提升至物理厚度的 4 倍以上, 大幅提高弱吸收波段 (如锗的 1550 nm 窗口) 的光吸收^[9-10]。特别地, 当器件集成于共面波导 (CPW) 传输线时, 光栅指状结构直接作为垂直电场施加单元, 显著缩短载流子渡越路径。基于该设计, 本文通过电磁波-半导体多物理场耦合仿真 (COMSOL Multiphysics), 模拟研究了 MG-Ge-Mirror

结构在 500-1800 nm 宽谱范围内的光吸收特性，证明器件在 1550 nm 处实现 99% 超高吸收率与 0.92 A/W 响应度，响应速度可达 37.6 fs，为太赫兹光混频系统提供了一条可能得技术途径。

2 模型设计及方法

图 1a 为本文提出的 MG-Ge-Mirror 光电探测器的 3D 结构示意图，其截面图如图 1b 所示。该结构基底选用电阻率 $\geq 10^{10} \Omega \cdot \text{cm}$ 的高阻 SiC 衬底，其热导率 ($>400 \text{ W/m}\cdot\text{K}$) 可有效抑制高频工作时的热积累效应。在 SiC 表面沉积 300 nm 厚 Au 薄膜，其在 500-1800 nm 波段的光透射率趋近于零 ($T \approx 0$)，兼具背景光屏蔽与反射镜功能。在 Au 薄膜上键合厚度可调的本征 Ge 层作为光吸收区，最后在 Ge 层表面制备周期性 Au 光栅电极，其几何参数 (周期 p 、狭缝占比 a 、高度 t_{Au}) 经优化设计以增强光学性能。考虑毫米波太赫兹光混频的应用需求，器件电极集成了共面波导 (CPW) 耦合结构。

基于该器件构型，首先我们需要利用商业有限元软件 COMSOL Multiphysics 的“电磁波，频域 (ewfd)”模块来确定 MG-Ge-Mirror 结构在 1550 nm 波长处光吸收增强的结构参数，然后深入分析其光吸收增强的物理机制。在模型构建中，器件周期性光栅的周期沿 x 方向，假设光栅沿 y 轴方向为无限长，TM 偏振的光从 z 轴反方向垂直入射，因此可以将什么简化为二维模拟。如图 1b 中红色方框所示，我们选取有源部分的 x 方向采用周期性边界条件 (periodic boundary conditions, PBC)，z 轴方向采用完美吸收边界条件 (perfectly matched layer, PML)。最后，在通过参数化扫描确认了 MG-Ge-Mirror 结构在 1550nm 处的最优的结构参数后，使用 COMSOL Multiphysics 的“半导体-电磁波 (semc) 耦合”模块对光电探测器的静态和动态光电性能进行仿真。

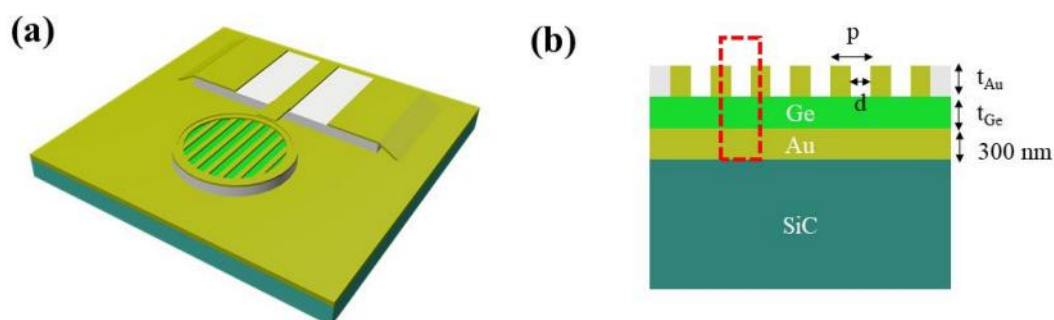


图 1. (a) MG-Ge-Mirror 光电探测器和 CPW 波导耦合的示意图。(b) 器件有源区横截面视

图。

3 结果与讨论

光电探测器的响应度与响应速度决定着光混频产生太赫兹信号的功率和带宽。MG-Ge-Mirror 光电探测器的响应度是由亚波长金属光栅的透射率和 Ge 谐振的量子效率两个方面决定的，而透射率和量子效率又取决于光栅的结构参数和 Ge 的厚度。本节首先分析了 Ge 的厚度和光栅的结构参数对吸收率的影响，然后讨论了光吸收增强的物理机制，最后根据优化后的光吸收增强结构，分析 MG-Ge-Mirror 光电探测器的光电响应。

首先，对 Ge 层厚度 t_{Ge} 进行优化仿真， t_{Ge} 厚度范围为 100-400 nm，调节步长为 10 nm。取光栅周期 p 为 200 nm，光栅狭缝占比 a 为 0.5，光栅高度 t_{Au} 为 350 nm，图 2a 显示了结构在波长为 1550 nm 处的光吸收率。在整个范围内，吸收率最大值超过 0.9 而最小值不到 0.5，这说明了 Ge 层厚度对结构吸收率的影响较大，吸收率最大值对应的 Ge 层厚度为 200 nm。其次，对光栅高度 t_{Au} 进行仿真， t_{Au} 高度范围为 200-400 nm，步长为 10 nm。取光栅周期 p 为 200 nm，光栅狭缝占比 a 为 0.5，Ge 层厚度 t_{Ge} 为 200 nm，图 2b 显示了结构在波长为 1550 nm 处的光吸收率。在整个范围内，光栅高度对结构吸收率的影响较小。因此，光栅高度主要影响电极电阻，对 LSPR 局域场分布影响较小，所以选择吸收率最大值对应的光栅高度 350 nm。然后，考虑光栅周期 p 对结构光吸收的影响， p 的变化范围为 100-300 nm，步长为 10 nm。取 Ge 层厚度为 200 nm，光栅狭缝占比 a 为 0.5，光栅高度 t_{Au} 为 350 nm，图 2c 显示了结构在波长为 1550 nm 处的光吸收率。在整个范围内，结构的吸收率随着光栅周期增出现了先增大在减小的情况，吸收率峰值对应的光栅周期为 200 nm。最后，对光栅狭缝占比 a 对光吸收的影响进行了仿真，范围为 0.2-0.8，步长为 0.1。取光栅周期 p 为 200 nm，光栅高度 t_{Au} 为 350 nm，Ge 层厚度 t_{Ge} 为 200 nm，图 2d 显示了结构在波长为 1550 nm 处的光吸收率。可以清晰的观察到，当光栅狭缝占比为 0.5 时，结构具有最高的吸收率，此时金属遮光区与透光区面积相等。

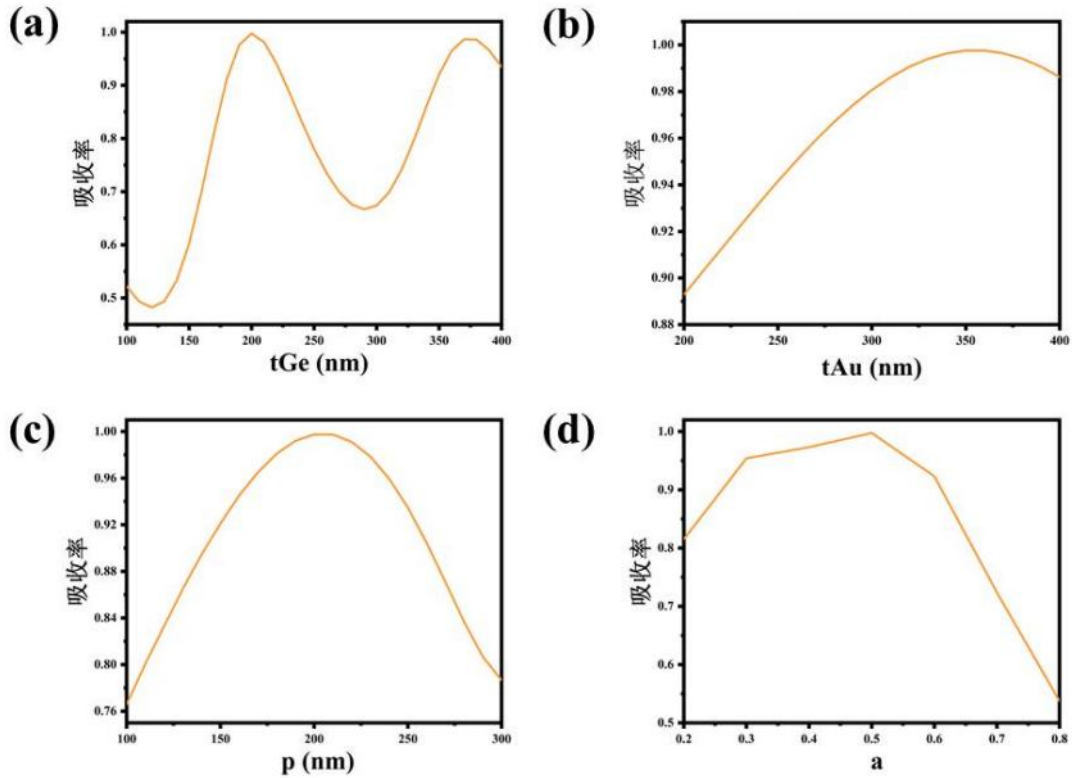


图 2. MG-Ge-Mirror 结构在波长为 1550 nm 处的光吸收率 (a) 随 Ge 层厚度 t_{Ge} 变化的曲线; (b) 随光栅高度 t_{Au} 变化的曲线; (c) 随光栅周期 p 变化的曲线; (d) 随光栅狭缝占比 a 变化的曲线。

基于上述仿真结果，我们在后续研究中，选取金属光栅结构参数为：周期 $p=200$ nm、光栅间距占比 $a=0.5$ 、光栅度 $t_{Au}=350$ nm, Ge 层厚度 ($t_{Ge}=200$ nm)。MG-Ge-Mirror 光电探测器在 500-1800 nm 波段展现出卓越的光吸收特性，如图 3a 所示, MG-Ge-Mirror 结构吸收光谱呈现双主峰结构: 630 nm 处峰值吸收率 87%，1550 nm 处达 99%，后者在 1535-1552 nm 带宽内波动小于 0.1%，满足宽带可调谐太赫兹光混频的需求。电磁场分布解析揭示了波长依赖的吸收机制演化，如图 3b 所示。短波长区（500 nm）磁场局域于金/空气界面并指数衰减，呈现传播型表面等离子极化激元（SPPs）特征，结构吸收率受限于金属欧姆损耗约为 82%；在可见光区（630 nm），光栅狭缝内形成法布里-珀罗（F-P）驻波，同时 Ge 表面激发近场吸收，二者协同实现近 87% 光吸收；在近红外区（1550 nm），超高吸收源于亚波长光栅激发异常光学透射（EOT）和局域表面等离子体共振（LSPR）效应，高效耦合入射光以及底部 Au 镜全反射 ($R>99\%$) 与光栅构成 F-P 腔，满足共振条件，腔内形成强局域化驻波场，使超薄有源层实现等效光程倍增，吸收率约为 99%，同时我们观察到光场主要局域在光栅以下几 nm 的 Ge 层内；在长波边缘（1800 nm）：EOT 效应减弱导致吸收降低（63%）。

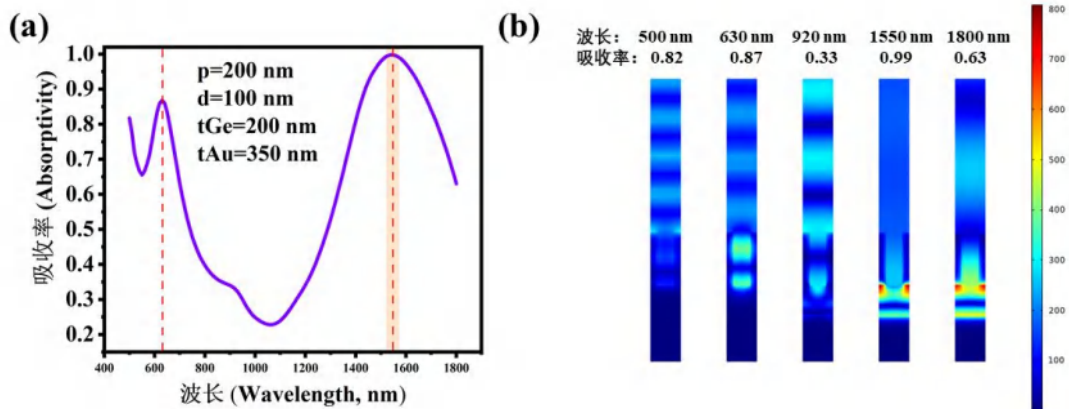


图 3. (a) MG-Ge-Mirror 结构的吸收光谱; (b) 500 nm, 630 nm, 920 nm, 1550 nm, 1800 nm 处的磁场分布图。

为了理解 MG-Ge-Mirror 结构光吸收增强的物理机制，本节进一步研究了什么的几何参数关键几何参数对其光谱响应行为的调控规律。如图 4a 所示，当 t_{Ge} 从 100 nm 增至 400 nm 时，光栅狭缝内 F-P 腔共振峰位置与强度保持稳定（630 nm），证实该模式对 Ge 厚度变化不敏感。但 1550 nm 处微腔驻波峰呈现周期性震荡，满足以下公式：

$$\Delta t_{Ge} = \frac{\lambda}{2n_{Ge}} \approx 195 \text{ nm} \quad (1)$$

其中， n_{Ge} 为锗材料在特定波长下的折射率实部（Real part of the complex refractive index），其值在 1550nm 的波长下约为 4。因此，峰值吸收率在 $t_{Ge} = 200 + m \times 195 \text{ nm}$ ($m = 0, 1, 2$) 处达 99% 以上，严格遵循法布里-珀罗共振条件，直接证实 Ge 层内多次反射干涉的物理图景。

如图 4b 所示，光栅狭缝内 F-P 腔共振峰随 t_{Au} 增加发生红移，这源于光栅高度引起的透射相位延迟，相位延迟公式为：

$$\Delta\phi = 2\pi\lambda \cdot \text{Re}(n_{Au}) \cdot t_{Au} \quad (2)$$

其中， $\Delta\phi$ 为光通过金光栅产生的相位延迟， λ 为入射光波长， $\text{Re}(n_{Au})$ 为金的复折射率实部， t_{Au} 为光栅高度。

如图 4c 所示，随着光栅周期由 100 nm 增加到 300 nm，光栅间隙内 F-P 腔共振峰蓝移，这是由于光栅周期增大使衍射角减小 ($\sin\theta = \frac{\lambda}{p}$)，有效光程缩短

($L_{eff} = \frac{2n_{Ge}t_{Ge}}{\cos\theta}$)，共振波长蓝移 ($\frac{\Delta\lambda}{\Delta p} \approx -0.6 \text{ nm/nm}$)。Ge 层内 F-P 腔共振峰红

移，这是由于光栅周期增大降低光栅等效折射率 ($n_{eff} \propto \frac{1}{p}$)，反射相位延迟减小 ($\phi_g \propto n_{eff}$)，为维持 $2n_{Ge}t_{Ge} + \phi_g = m\lambda$ 共振条件，波长需红移 ($\frac{\Delta\lambda}{\Delta p} \approx +$

0.7 nm/nm)。如图 4d 所示，光栅间隙占比 a 增大使双峰同步蓝移，这是由于间隙增大削弱狭缝等离子体耦合，降低光子注入效率 ($\eta_{EOT} \propto e^{-\frac{a}{a_c}}$ ，其中 $a_c \approx 80 \text{ nm}$ 为耦合衰减长度)，使共振能量向短波迁移；同时光栅间隙增大使衍射角增大，延长了有效光程，产生微弱红移趋势，但被 EOT 主导的蓝移覆盖。

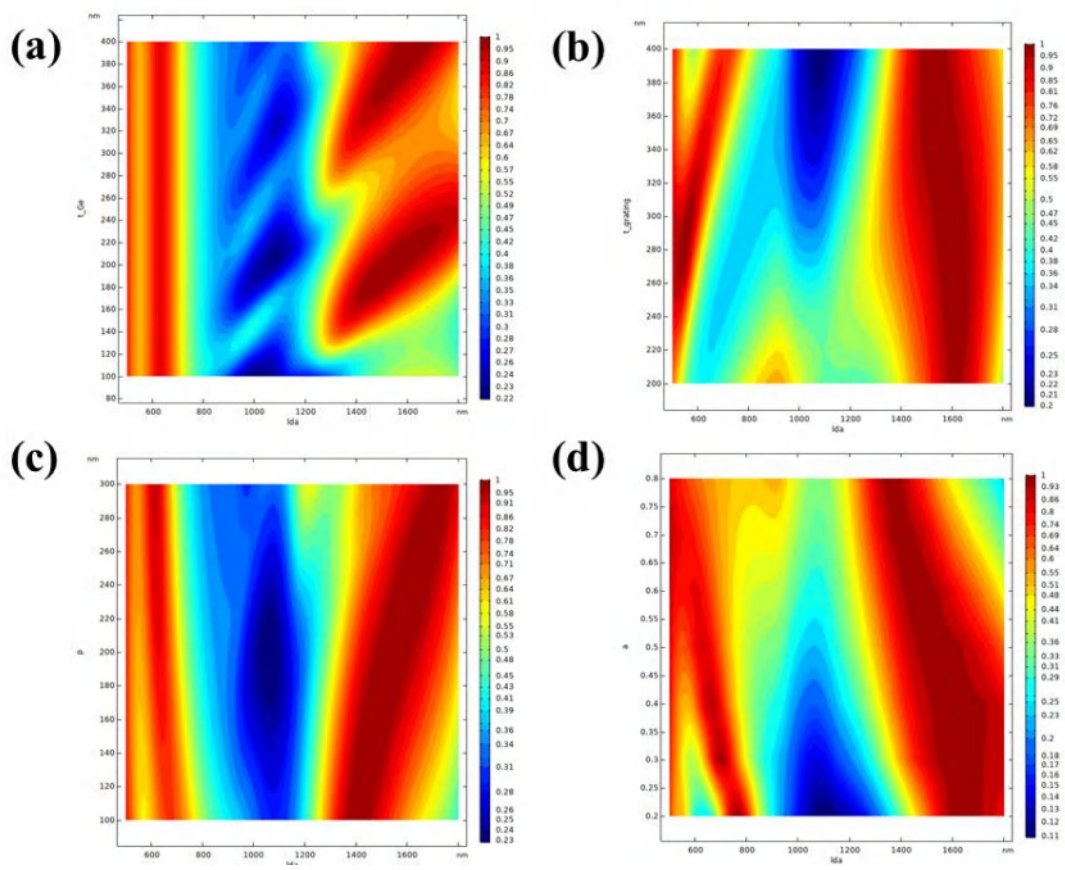


图 4. (a) MG-Ge-Mirror 结构吸收光谱随结构中的 Ge 层厚度 t_{Ge} 的变化情况；(b) MG-Ge-Mirror 结构吸收光谱随结构中的光栅高度 t_{Au} 的变化情况；(c) MG-Ge-Mirror 结构吸收光谱随结构中的光栅周期 p 的变化情况。(d) MG-Ge-Mirror 结构吸收光谱随结构中的占空比 a 的变化情况。

根据优化后的光吸收增强结构，分析 MG-Ge-Mirror 光电探测器的光电响应特性。当激光入射到 MG-Ge-Mirror 结构内部时，光子通过半导体光吸收作用产生电流。我们首先使用稳态求解器计算了 MG-Ge-Mirror 光电探测器的电流密度、量子效率 (EQE) 与响应度。如图 5a 所示，MG-Ge-Mirror 光电探测器的电流响应随着光强的增大而增大。同时，在 0.2 V 左右会出现电流饱和，这是因为光栅间隙处载流子堆积形成局部势垒。如图 5b 所示，MG-Ge-Mirror 结构的光电流密度峰值发生在 1610 nm 处，相较于吸收率的 1550 nm 处的峰值红移了 60 nm，此现象归因于热载流子的声子散射弛豫^[11]。最大电流密度在 1610 nm 处为 $3.41 \times 10^4 \text{ mA/cm}^2$ ，而 1550 nm 处为 $3.27 \times 10^4 \text{ mA/cm}^2$ 。图 5c 展示了 MG-Ge-Mirror 结构

工作在 1V 偏压下，在 1200-1800 nm 范围内，MG-Ge-Mirror 结构光电探测器的响应度和外量子效率（EQE），在 1550nm 处，响应度 0.92 A/W，外量子效率 73.8%。最后，我们使用了瞬态求解器计算了 MG-Ge-Mirror 光电探测器的响应速度。图 5d 展示了探测器工作在 1V 偏压，飞秒激光照射下，MG-Ge-Mirror 结构光电探测器的载流子响应速度。模拟的 fs 激光为功率 1 mW 的 10 fs 的高斯脉冲，MG-Ge-Mirror 结构光电探测器的半峰全宽（ t_{FWHM} ）为 37.6 fs。如图 3b 所示，超快的响应速度源于光子能量局域于光栅间隙下方的纳米级吸收区内，因此载流子具有超短的传输距离。

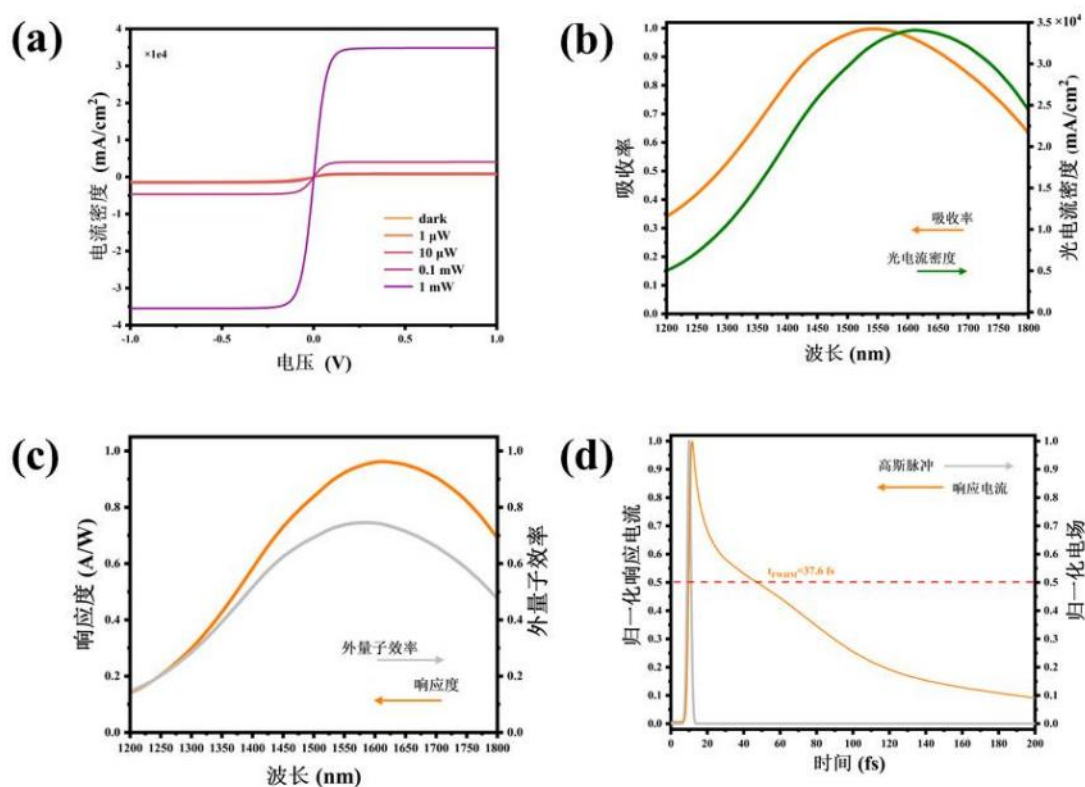


图 5. (a) MG-Ge-Mirror 结构在-2 到 2 V 偏压下，不同光强激光照射的电流-电压曲线；(b) 1 V 偏压下，1 mW 不同波长激光照射下电流密度曲线；(c) 响应度和外量子效率曲线；(d) 1V 偏压下，1 mW 飞秒（10 fs）高斯脉冲激光照射下电流响应曲线（归一化）。

总结

本文提出并仿真研究了一种金属光栅-锗-金属镜（MG-Ge-Mirror）结构的光电探测器，该器件通过亚波长光栅的异常光学透射（EOT）与局域表面等离子体共振（LSPR）效应，将 1550 nm 光子能量局域于光栅间隙下方几 nm 薄锗层内；同时，底部金属镜与光栅构成法布里-珀罗（F-P）谐振腔，使入射光在锗层内多次反射，进一步增强其光吸收能力。通过光学仿真，研究了光栅周期、狭缝占比、厚度等参数 MG-Ge-Mirror 结构对光吸收的影响，在 500–1800 nm 宽谱范围内实现了高效的光吸收。理论分析表明，LSPR 近场增强与 F-P 腔谐振模式的协同

效应是突破超薄吸收层量子效率限制的关键机制。光电联合仿真结果显示,该器件在 1550 nm 通信窗口实现 99%超高光吸收率与 0.92 A/W 响应度(对应外量子效率 73.8%);本征响应速度达 37.6 fs。本研究的工作为解决高速光电探测器功率-带宽性能限制提供了一种可能的解决思路,在光子学太赫兹领域具有良好的应用前景。

参考文献

- [1] Sizov F F. Detectors and Sources for THz and IR[C]. Materials Research Forum LLC, 2020.
- [2] Chen G, Yu Y, Shi Y, et al. High-Speed Photodetectors on Silicon Photonics Platform for Optical Interconnect[J]. Laser & Photonics Reviews, 2022, 16(12): 2200117.
- [3] Sze S M, Coleman Jr D J, Loya A. Current transport in metal-semiconductor-metal (MSM) structures[J]. Solid-State Electronics, 1971, 14(12): 1209-1218.
- [4] Buca D, Winnerl S, Lenk S, et al. Metal-germanium-metal ultrafast infrared detectors[J]. Journal of applied physics, 2002, 92(12): 7599-7605.
- [5] Rouvière M, Vivien L, Le Roux X, et al. Ultrahigh speed germanium-on-silicon-on-insulator photodetectors for 1.31 and 1.55 μm operation[J]. Applied Physics Letters, 2005, 87(23).
- [6] 蔡志猛.Si基 Ge MSM 光电探测器的研制[D].厦门大学,2008.
- [7] 杜鸣笛.基于金属交叉电极耦合谐振腔增强结构的高性能光电探测器研究[D].华中科技大学,2014.
- [8] Vorobyev L E. Germanium (Ge)[J]. Handbook series on semiconductor parameters, 1996, 1: 33-57.
- [9] Kats M A, Capasso F. Optical absorbers based on strong interference in ultra-thin films[J]. Laser & Photonics Reviews, 2016, 10(5): 735-749.
- [10] Billet M, Latzel P, Pavanello F, et al. Resonant cavities for efficient LT-GaAs photoconductors operating at $\lambda=1550\text{ nm}$ [J]. APL Photonics, 2016, 1(7).
- [11] Tea E, Hamzeh H, Aniel F. Hot carriers relaxation in highly excited polar semiconductors: Hot phonons versus phonon-plasmon coupling[J]. Journal of Applied Physics, 2011, 110(11).

基于 Angelov 模型的 GaN HEMT 非线性电容模型的研究

缪宇辰¹, 袁清宇², 王创业¹, 程加力¹

¹江苏海洋大学 电子工程学院 连云港 222005

²杭州电子科技大学 空间信息研究院 浙江省空间信息感知与传输重点实验室

杭州 310018

摘要: 本文研究了 Angelov 模型对于 GaN HEMT 器件特性的适用情况, 测试了基于 UMS GH15-10 工艺制造的 GaN HEMT 晶体管的 S 参数。剥离所有寄生元件后, 在多个偏置条件下提取出 GaN HEMT 的非线性栅源电容 C_{gs} 和栅漏电容 C_{gd} 。并基于 Angelov 模型建立了 GaN HEMT 的非线性电容模型, 最后分析了电容提取结果与仿真结果相对误差 R^2 和均方根误差 RMSE。

关键词: 本征电容; 参数提取; 非线性电容; 小信号等效电路; Angelov 模型

1 引言

近年来, GaN 高电子迁移率晶体管 (HEMT) 因其优异的大功率、高压和高频特性, 在电力电子与射频领域得到广泛应用。建立精确的 GaN HEMT 大信号模型对提升电路仿真精度与收敛性至关重要。现在的主流大信号模型有物理模型, 经验模型, 神经网络模型和查找表模型。物理模型是从半导体物理知识中物理推导而来; 经验模型主要利用数学公式拟合测试数据来表征器件的电流和电荷特性; 神经网络模型是通过神经网络描述模型的电流源和电荷源; 对于查找表模型, 电流源和电荷源存储在文本查找表中, 电路仿真时, 仿真器利用内插和外插算法产生新的包含高阶导数的数据。其中, 经验模型因结构简洁, 精度可控, 易于嵌入 EDA 工具等优势, 成为工业界主流选择^[1]。

当前 HEMT 研究主要依赖经验模型表征性能, 其表达式简单且参数可调, 已广泛应用于各类场景。在经验模型中, Angelov 模型因其精确拟合跨导和线性度、且保持导数连续从而确保仿真收敛性而被广泛应用。鉴于 GaN 与 GaAs 的材料相似性, 现在大部分 GaN HEMT 经验模型都参考的 MOSFET 经验模型^[2]。然而, HEMT 的核心是异质结结构, 即在两种不同带隙材料的界面处 (如 AlGaIn/GaN) 通过能带工程形成高迁移率的二维电子气 (2DEG)^[3]。2DEG 的主要特点在于其电子被量子力学效应强烈限制在垂直于界面的维度 (通常在几纳米范围内), 从而形成一种极薄 (约 10nm 量级)、高密度、高迁移率的电子气层。这种二维约束特性极大程度地减少了晶格散射和界面粗糙度散射效应, 赋予了电子极高的低场迁移率。这一工作机理与 MOSFET 中通过栅极电压在半导体表面感应形成的反型层电荷有着本质差异^[4]。因此, 传统 Angelov 电容模型, 并不一定能准确描述 GaN HEMT 的电容非线性行为。

2 模型介绍及参数提取

2.1 传统 Angelov 模型介绍

Angelov 模型是一种应用广泛的经验模型,该模型是由 Iltcho Angelov 在 1992 年提出。该模型基于 Statz 和 Curtice 模型开发,后来,他在 1996 年和 1999 年进行了两次重大改进。改模型的有效应已经在许多论文和实践应用种得到验证,可以很好的运用于高频应用,阻挡效应建模,电热效应建模。其在机械学习领域种也显示出良好的适应性。

Angelov 模型采用的非线性 C_{gs} 和 C_{gd} 电容模型,他们的表达式如下所示:

$$C_{gs} = C_{gsp} + C_{gs0} [1 + \tanh(\varphi_1)] [1 + \tanh(\varphi_2)] \quad (1)$$

$$C_{gd} = C_{gdp} + C_{gd0} [1 + \tanh(\varphi_3)] [1 - \tanh(\varphi_4)] \quad (2)$$

其中

$$\varphi_1 = P_{10} + P_{11}V_{gs} + P_{12}V_{gs}^2 + P_{13}V_{gs}^3 + \dots \quad (3)$$

$$\varphi_2 = P_{20} + P_{21}V_{ds} + P_{22}V_{ds}^2 + P_{23}V_{ds}^3 + \dots \quad (4)$$

$$\varphi_3 = P_{30} + P_{31}V_{gs} + P_{32}V_{gs}^2 + P_{33}V_{gs}^3 + \dots \quad (5)$$

$$\varphi_4 = P_{40} + P_{41}V_{ds} + P_{42}V_{ds}^2 + P_{43}V_{ds}^3 + \dots \quad (6)$$

$P_{10}, P_{11}, P_{12}, P_{13}$ 等参数均为拟合因子。

当电容精度可以保证再 5%~10% 的时候, 可以用一阶近似来表征电容:

$$C_{gs} = C_{gsp} + C_{gso} [1 + \tanh(P_{10} + P_{11}V_{gs})] [1 + \tanh(P_{20} + P_{21}V_{ds})] \quad (7)$$

$$C_{dg} = C_{gdo} + C_{gdo} [1 + \tanh(P_{30} + P_{31}V_{gs})] [1 - \tanh(P_{40} + P_{41}V_{gd})] \quad (8)$$

2.2 参数提取

本文采用直接提取法与数值优化相结合的方法提取模型参数。其核心思想遵循“提取寄生电容→去嵌寄生电容→提取寄生电感及寄生电阻→去嵌所有寄生参数→提取本征参数”的流程。随着器件工作频率的不断提升,尤其是进入毫米波段,衬底效应等寄生效应对器件特性的影响已不可忽略。在此背景下,主要适用于 40GHz 以下频率的传统小信号模型已难以满足当前电路设计需求。针对这一挑战,本文在传统小信号等效电路模型的基础上,创新性地引入了寄生元件 R_{pg} 和 R_{pd}^[7], 以有效表征衬底效应,小信号模型如下图 1 所示:

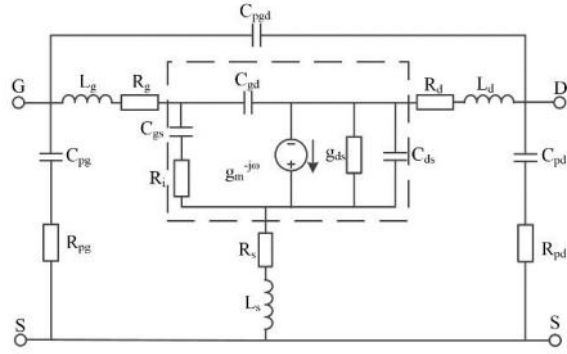


图1 改进的16元件小信号电路模型

首先，GaN HEMT 器件在低频 ($f < 5\text{GHz}$) 且“冷”夹断 ($V_{gs} < V_{th}, V_{ds} = 0\text{V}$) 工作状态下,寄生电感和寄生电阻对电路的影响可以忽略不计^[5], 主要是电容起作用, 此时的小信号等效电路拓扑结构如图红色所示。本文采用的是一种基于尺寸比例模型的寄生电容提取方法, 相比于传统的提取方法对于等效电路没有任何限制和假设, 利用四个符合尺寸比例关系的 HEMT 器件提取寄生电容, 可以同时提取所有的寄生电容 C_{pg} , C_{pd} 和 C_{pgd} , 无须复杂的计算和推导过程。本文基于 UMS GH15-10 工艺制作了五个不同尺寸的 GaN HEMT, 分别是 TZ0210, TZ0420, TZ0630, TZ0840, TZ0860。晶体管栅极长度均为 150nm。尺寸分别为 (栅宽×栅指): 10×2、20×4、30×6、40×8 和 60×8。上述器件均采用碳化硅衬底。

基于尺寸比例模型步骤的方法如下: 首先, 测量了不同栅宽器件在零偏压状态下的小信号 S 参数。然后, 将 S 参数转换为 Y 参数。最后由 Y 参数虚部得到寄生电容的值。

Y 参数的导纳矩阵表示为:

$$Y = \begin{bmatrix} \frac{j\omega C_{pg}}{1 + j\omega C_{pg} R_{pg}} + j\omega C_{pgd} & -j\omega C_{pgd} \\ -j\omega C_{pgd} & \frac{j\omega C_{pd}}{1 + j\omega C_{pd} R_{pd}} + j\omega C_{pgd} \end{bmatrix} \quad (9)$$

基于尺寸比例模型的方法如下:

$$C_{pg} = \frac{\text{Im}(Y_{11})}{w} \Big|_{w \rightarrow 0} - C_{pgd} \quad (10)$$

$$C_{pd} = \frac{\text{Im}(Y_{22})}{w} \Big|_{w \rightarrow 0} - C_{pgd} \quad (11)$$

$$C_{pgd} = \frac{\text{Im}(Y_{12})}{w} \Big|_{w \rightarrow 0} \quad (12)$$

式中的 w 为器件栅宽。

通过对五种不同器件在零偏压状态下的 S 参数和 Y 参数的测量，去低频（0.5~10GHz）范围的平均值作为最终值。寄生电容与栅宽之间的关系如图 2 所示。由于 C_{pgd} 的值非常小，几乎可以忽略不计，因此我们可以认为截距值是 C_{pd} 和 C_{pg} 的值。

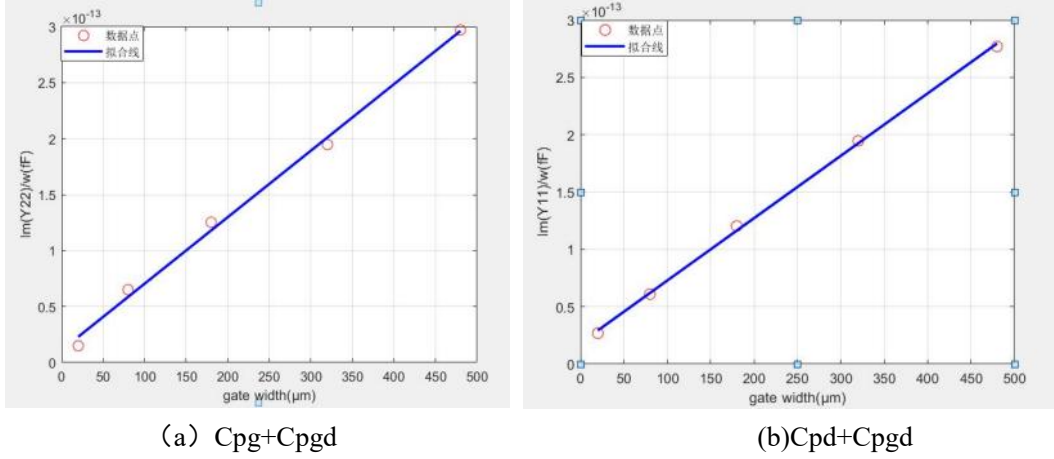


图 2 基于尺寸比例模型方法的提取结果：(a) $C_{pg}+C_{pgd}$ ；(b) $C_{pd}+C_{pgd}$

衬底的电阻的值可由以下的公式提取计算得到：

$$R_{pg} = \text{Re}\left(\frac{1}{Y_{11} + Y_{12}}\right) \quad (13)$$

$$R_{pd} = \text{Re}\left(\frac{1}{Y_{22} + Y_{12}}\right) \quad (14)$$

2.2 寄生电感和寄生电阻的提取方法

本文采用的是短路测试法^{[9][10]}，寄生电感和寄生电阻的值可由以下式求得：

1. 利用测得的开路结构的 S 参数矩阵和短路结构的 S 参数矩阵，将其转换成对应的 Y 参数矩阵 Y^{PAD} 和 Y^{short} ；

2. 利用电路分析理论消去寄生电容和衬底电阻的影响，得到只含有引线寄生电感和寄生电阻的 Y 参数矩阵 Y^{short1} 。

$$Y^{short1} = Y^{short} - Y^{PAD} \quad (15)$$

3. 将 Y 参数矩阵 Y^{short1} 转换为 Z 参数矩阵 Z^{short1} ：

$$Z^{short1} = \begin{bmatrix} Rg + jwLg + Rs + jwLs & Rs + jwLs \\ Rs + jwLs & Rd + jwLd + Rs + jwLs \end{bmatrix} \quad (16)$$

4. 通过下面几个公式计算出三个引线的寄生电感和寄生电阻值。

$$R_g = \operatorname{Re}(Z_{11}^{short1} - Z_{12}^{short1}) \quad (17)$$

$$R_d = \operatorname{Re}(Z_{22}^{short1} - Z_{12}^{short1}) \quad (18)$$

$$R_s = \operatorname{Re}(Z_{12}^{short1}) \quad (19)$$

$$L_g = \frac{\operatorname{Im}(Z_{11}^{short1} - Z_{12}^{short1})}{\omega} \quad (20)$$

$$L_d = \frac{\operatorname{Im}(Z_{22}^{short1} - Z_{12}^{short1})}{\omega} \quad (21)$$

$$L_s = \frac{\operatorname{Im}(Z_{12}^{short1})}{\omega} \quad (22)$$

由此，被测器件的寄生参数全部通过计算公式计算得到。将得到的数值作为初始值，利用优化算法进行迭代，最终得到的结果见下表 1。

表 1 寄生参数数值

参数名称	单位	值
Cpgd	fF	0.0035
Cpg	fF	18.4245
Cpd	fF	16.1755
Rpg	Ω	43.54
Rpd	Ω	21.83
Rd	Ω	5.1713
Rg	Ω	2.3050
Rs	Ω	0.315
Ld	pH	116.232
Lg	pH	117.655
Ls	pH	45.334

2.3 本征参数提取

寄生电阻和寄生电感在低频条件下可的影响可以忽略。但当工作频率增加时，尤其是当工作频率进入毫米波段时，器件上的互连寄生电感就不能忽略不记了，因此我们需要在高频条件下通过去嵌寄生元件的影响，对本征元件进行提取；具体的去嵌过程如下：

- 1.对待测器件进行在片测试，得到对应的 Y 参数矩阵 Y^{DUT}
- 2.剥离寄生电容和衬底电阻，得到器件的 Y 参数矩阵 Y^{HEMT}

$$Y^{HEMT} = Y^{DUT} - Y^{open} \quad (23)$$

3.将 Y 参数矩阵 Y^{HEMT} 转换为 Z 参数矩阵 Z^{HEMT} ,然后去除外部的寄生电感和寄生电阻,得到只含有本征元件的 Z 参数 Z^{int}

$$Z^{int} = Z^{HEMT} - Z^{short} \quad (24)$$

4.器件的本征参数提取公式:

$$C_{gd} = -\frac{Im(Y_{12}^{int})}{\omega} \quad (25)$$

$$C_{gs} = \frac{Im(Y_{11}^{int} - Y_{12}^{int})}{\omega} \quad (26)$$

$$g_{ds} = \frac{1}{Re(Y_{22}^{int})} \quad (27)$$

$$g_m = |Im(Y_{21}^{int}) - Im(Y_{12}^{int})| \quad (28)$$

$$\tau = -\frac{1}{\omega} \tan^{-1} \left[\frac{Im(Y_{11}^{int} + Y_{12}^{int})}{Re(Y_{11}^{int} + Y_{12}^{int})} \right] \quad (29)$$

我们提取了器件 TZ0210 和 TZ0420 的本征参数, V_{ds} 的测量范围从 0V 到 20V, V_{gs} 的测量范围从 -5V 到 0V, 提取出的 C_{gs} 结果如下图 3 和图 4 所示:

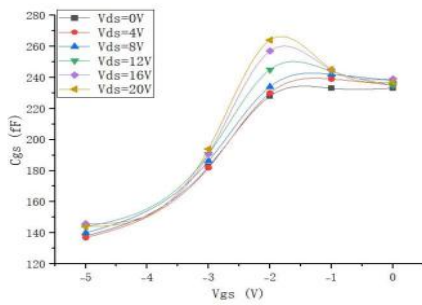


图 3 TZ0210 的 C_{gs} 提取结果

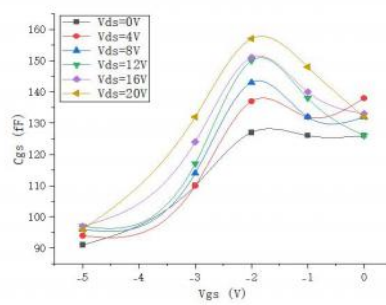


图 4 TZ0420 的 C_{gs} 提取结果

3 数据拟合与误差分析

我们将测量的 TZ0210 和 TZ0420 的 C_{gs} 电容值带入模型, 得到拟合参数结果如表 2 所示。

表 2 C_{gs} 拟合参数结果

	Cgsp	Cgso	P11	P12	P21	P22
TZ0210	94.6069	16.8038	10.0000	3.2842	-0.1135	0.0351
TZ0420	141.3553	344.5356	10.0000	3.3360	-1.3367	0.0055

分别选取了两个器件的 C_{gs} 的三维图和 C_{gs} 关于 V_{gs} 的横截面：

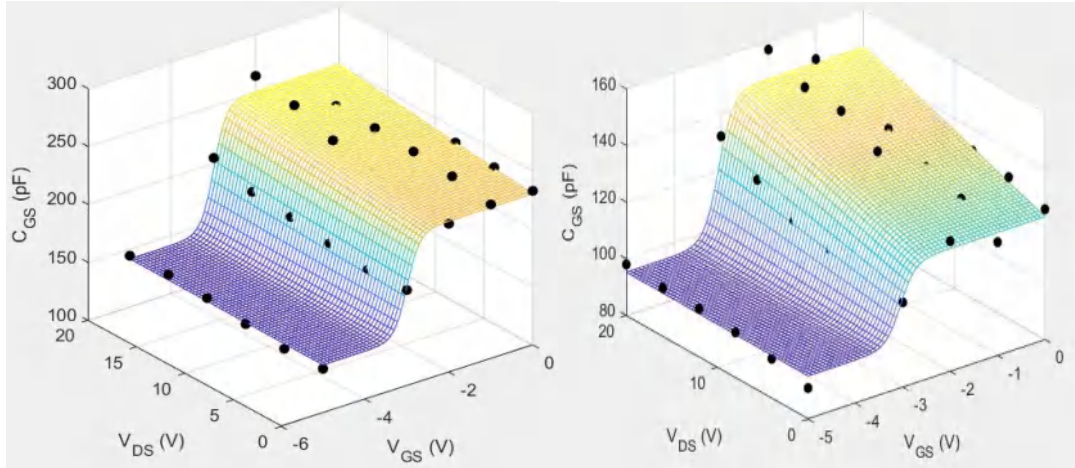


图 5 TZ0210 三维曲面图

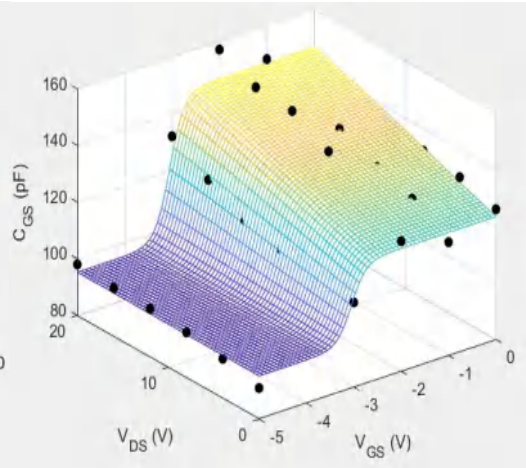


图 6 TZ0420 三维曲面图

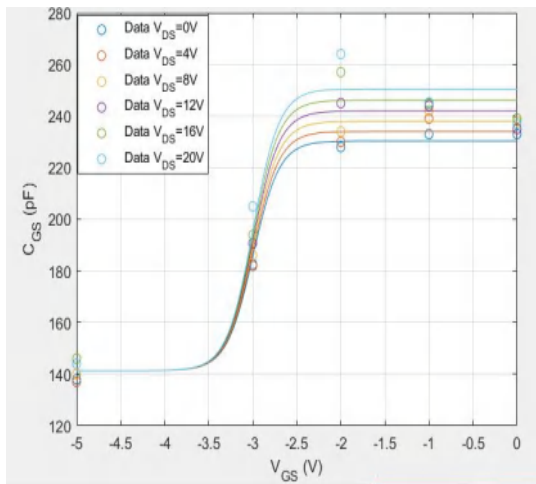


图 7 TZ0210 C_{gs} 关于 V_{gs} 的横截面

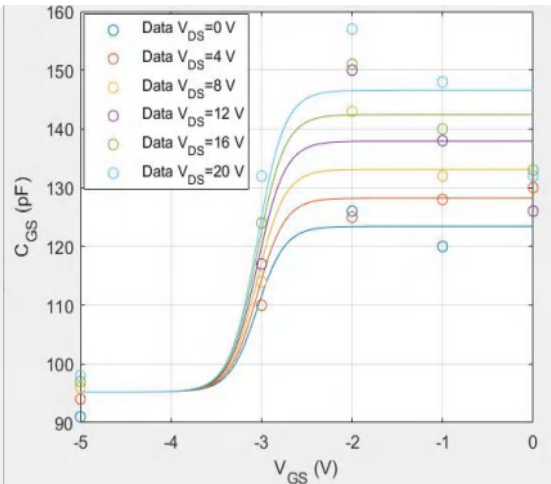


图 8 TZ0420 C_{gs} 关于 V_{gs} 的横截面

求得仿真结果 R^2 (相对误差) 和 RMSE(均方根误差)见表 3。

表 3

	RMSE	R^2
TZ0210	3.7160	0.9587
TZ0420	5.5651	0.9809

我们将两个器件测量的 C_{gd} 数据带入 Angelov 模型，得到拟合参数如下表 4 所示。

表 4 C_{gd} 拟合参数结果

	Cgdp	Cgdo	P31	P32	P41	P42
TZ0210	115.1728	-25.2927	0.3977	0.0793	0.3394	0.1868

TZ0420 84.1504 -18.4612 0.0274 0.1613 -0.0676 0.2256

分别选取两个器件的 C_{gd} 的三维图和 C_{gd} 关于 V_{gs} 的横截面:

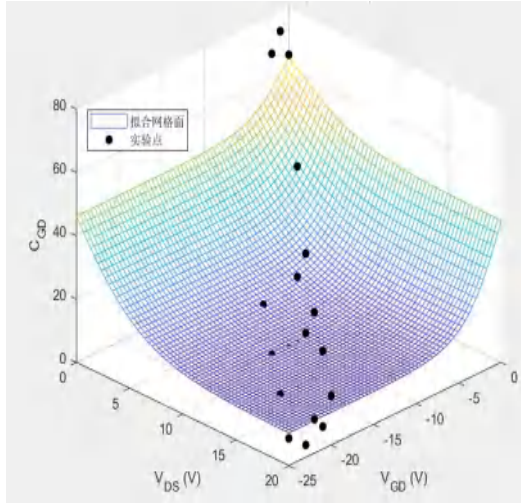


图 10 TZ0210 C_{gd} 三维曲面图

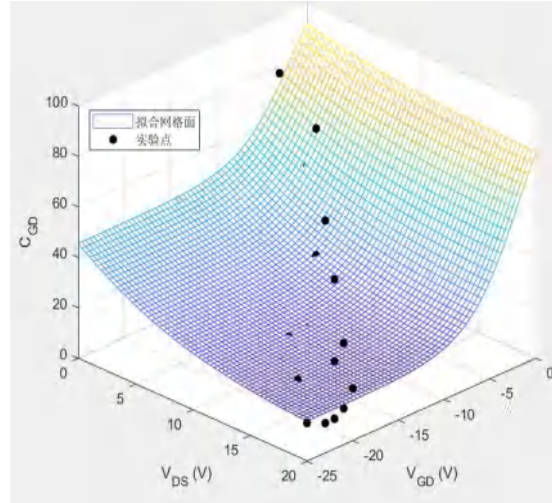


图 11 TZ0420 C_{gd} 三维曲面

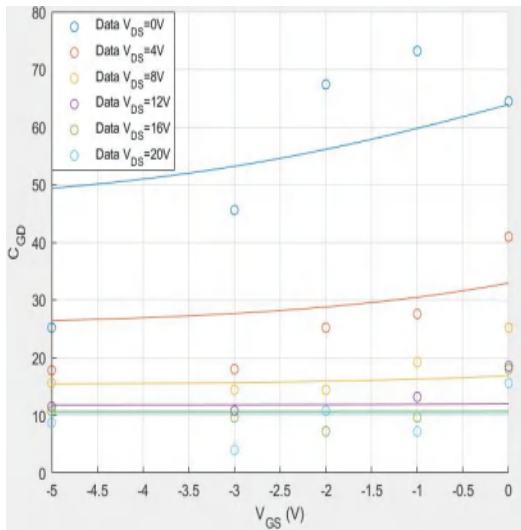


图 12 TZ0210 C_{gd} 关于 V_{gs} 的横截面

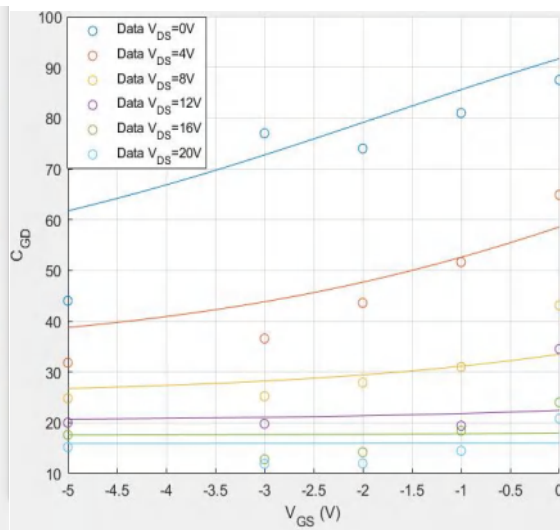


图 13 TZ0420 C_{gd} 关于 V_{gs} 的横截面

求得仿真结果 R^2 (相对误差) 和 RMSE(均方根误差)如表 5 所示。

表 5

	RMSE	R^2
TZ0210	7.0274	0.8468
TZ0420	5.5651	0.9809

3.2 文章总结

本文研究了 Angelov 模型对于 GaN HEMT 器件特性的适用情况, 测试了基于 UMS GH15-10 工艺制造的 GaN HEMT 晶体管的 S 参数。剥离所有寄生元件后, 在多个偏置条件下提取出 GaN HEMT 的非线性栅源电容 C_{gs} 和栅漏电容 C_{gd} 。并基于 Angelov 模型建立了 GaN HEMT 的非线性电容模型, 最后分析了电

容提取结果与仿真结果相对误差 R^2 和均方根误差 RMSE。

参考文献

- [1] 周影, 于盼盼, 高建军. 基于微波 S 参数测试的 MOSFET 小信号参数半分析法提取[J]. 南通大学学报(自然科学版), 2016, 15(03): 6-11.
- [2] 潘传奇, 王登贵, 周建军, 等. 高压硅基 p-GaN 栅结构 GaN HEMT 器件阈值电压稳定性研究[J]. 固体电子学研究与进展, 2025, 45(03): 15-19.
- [3] 毕磊. 射频 GaN HEMT 经验基大信号模型参数提取方法及建模技术研究[D]. 天津大学, 2021. DOI: 10.27356/d.cnki.gtjdu.2021.003575.
- [4] 黄磊, 毛书漫, 武庆智, 等. GaN HEMT 开关器件物理基大信号模型研究[C] //中国电子学会. 2023 年全国微波毫米波会议论文汇编(三); 2023: 39-41. DOI: 10.26914/c.cnkihy.2023.042099.
- [5] 程加力, 韩波, 李寿林, 等. 基于 GaAs STATZ 模型的 RF MOSFET DC 建模技术[J]. 半导体技术, 2011, 36(08): 591-594+603.
- [6] 王凯, 吴畅, 邢绍琨, 等. AlGaIn/GaN 双沟道 HEMT 下势垒层厚度对器件性能影响的研究[C]//中国电子学会.2024 年全国微波毫米波会议论文汇编(上册). 九峰山实验室, 2024: 768-770. DOI: 10.26914/c.cnkihy.2024.019822.
- [7] 张傲, 高建军. 毫米波寄生电容去嵌 S 参数影响性分析[C]//中国电子学会. 2021 年全国微波毫米波会议论文集(上册). 华东师范大学, 2021: 431-433. DOI: 10.26914/c.cnkihy.2021.030938.
- [8] 张佳琪, 徐跃杭. 一种微波 GaN HEMT 器件 IV 模型高效参数提取研究[C]//中国电子学会. 2018 年全国微波毫米波会议论文集(下册). 电子科技大学电子工程学院, 2018: 213-215.
- [9] Qingyu Yuan, Yixin Zhang, Xiaodong Luan, et al. An Improved Nonlinear I-V Model for GaN HEMTs [J]. International Journal of RF and Microwave Computer-Aided Engineering, 2024. 2024(1): Article ID 8834864.
- [10] Jing Bai, Ao Zhang, Jiali Cheng, et al., An improved small-signal model for GaN HEMTs devices [J]. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, 2024. 37(3): e3237.

多通道高速光电微系统设计与仿真研究

王皓岩^{1†}, 罗伟^{2†}, 曾虹铭², 刘静², 宋航宇², 廖海军², 刘杰¹, 袁恺^{2†}

1. 中国电子科技集团公司信息科学研究院, 重庆, 40030

2. 联合微电子中心有限责任公司, 重庆, 401332

‡ 同等贡献; †通讯作者: kai.yuan@cumec.cn

摘要: 多通道高速光电微系统在高带宽通信、数据中心、高性能计算等领域有着广阔的应用前景, 传统封装难以满足高速、低功耗的需求, 而光电合封技术通过缩短信号传输路径, 有望解决目前高速光电微系统面临的带宽及功耗瓶颈。本文基于先进封装工艺, 针对典型 16 通道高速光电微系统, 进行光电集成链路联合仿真设计, 提出了一种基于三维异构集成方案, 通过晶圆重塑、TSV 硅桥与重布线技术, 实现硅光芯片和电芯片的多芯粒三维互连。测试结果表明, 采用 NRZ 调制方式的 16 通道光模块速率可达 400Gbps, 误码率约 $1e-9$, 眼图情况良好。该封装结构突破传统平面集成的限制, 为今后的高密度高速光电微系统集成提供了技术支持。

关键词: 多通道高速光电微系统; 三维异构集成; 光电合封; 仿真设计

1 引言

人工智能、大数据、5G、边缘计算等领域的发展推动全球网络流量和算力的增长, 数据中心和 AI 领域对高速、低延迟、低功耗的数据传输和处理技术需求激增。基于传统封装技术的光电微系统如可插拔光模块, 已难以满足数据中心和 AI 等应用场景的高带宽低功耗需求。光电合封 (Co-packaged Optics, CPO) 技术通过将光引擎与 ASIC 芯片集成于同一封装体内, 大幅缩短光电互连距离至 mm 级, 减小光/电信号传输路径, 可有效降低功耗与延迟、提升带宽, 有望突破“算力-带宽-功耗”瓶颈[1-3]。区别于传统平面集成或简单堆叠方案, 基于硅通孔 (TSV)、硅桥等互连结构, 实现光电芯片、驱动芯片等多芯粒的三维异构集成[4,5], 能通过垂直布线缩短信号传输路径以降低损耗, 不同功能芯粒的空间重构有助于热分布优化, 从而降低模块整体功耗与体积, 为高速光电微系统模組的性能提升提供新的技术路径[6]。

本研究针对高速光电微系统应用, 提出基于硅桥 TSV 互连的三维异构集成方案, 设计了一种典型 16 通道高速光电微系统模組, 通过低损耗 TSV、晶圆级 RDL、凸点、晶圆重塑、Die-to-Wafer 键合等工艺, 完成了 16 通道高速光电微系统模組光电合封, 光电测试结果表明, 16 通道眼图情况良好, 达到了设计要求, 为设计更高性能光电微系统模組提供了技术基础。

2 系统设计

2.1 链路方案

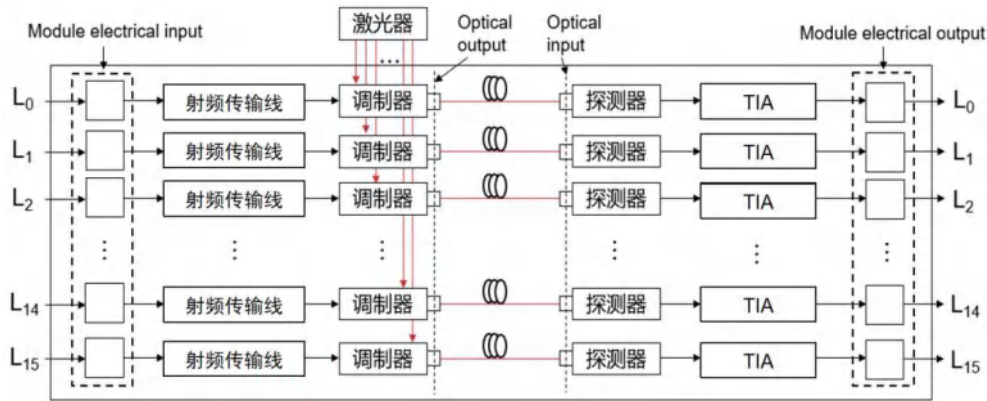


图 1 多通道高速光电微系统原理框图

如图 1 所示，高速光电微系统模组采用 16 路并行传输，包括 16 路发射机和 16 路接收机。在发射端，通过对调制器直接驱动的方案调制激光器的输入光，这样将输入的电信号转换为光调制信号输出，同时省去 Driver 芯片；在接收端，输入的光调制信号通过探测器转换为调制的电流信号，然后通过 TIA 将电流信号转换为放大电压信号。

通过仿真与测试，确定硅光芯片各项指标，单个端口带宽需达到 25Gbps，总传输速率达到 400Gbps ($16 \times 25\text{Gbps}$)，发射端消光比 $\geq 3.5\text{dB}$ ，平均输出功率 $\geq -10\text{dBm}$ ，接收灵敏度 $> -11\text{dBm}$ 。

2.2 三维集成方案

本项目研制的 16 通道高速光模块组采用第三代晶圆级硅光三维集成微模组技术途径。该微模组自下而上分别是 LTCC 基板、硅光芯片与激光器、三维异构集成结构 (eWLB 基板) 与 SiN 基板、4 个 TIA 芯片，其集成封装结构如图 2 所示。本产品设计发端采用直接驱动的方案，收端采用 PD+TIA 的方案，因此无需 Driver 电芯片，只需 TIA 芯片。每个 TIA 芯片有 4 个通道，因此需要 4 片 TIA 芯片。在集成架构上采用硅桥-eWLB 方案，将四颗 TIA 芯片和 TSV 桥芯片通过扇出工艺埋入塑封料中。本方案直接通过扇出工艺连接 TIA 芯片和硅光芯片，既保证了 TIA Pad 的重新扇出，又省去了硅转接板，只需要一层凸点工艺，可靠性更高。从图中可以看到，光口主要位于硅光芯片上，发端的电接口位于硅光芯片上，收端的电接口位于顶部的 eWLB RDL 上。

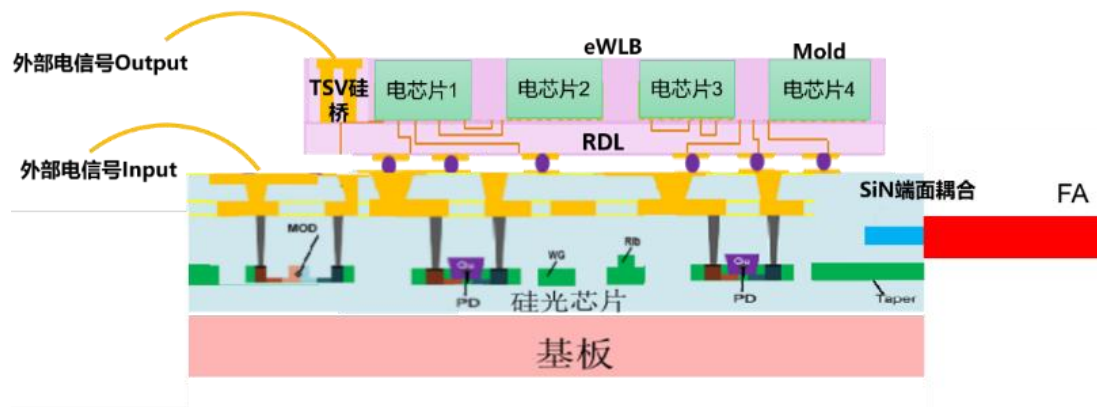


图 2 多通道高速光电微系统模组三维集成方案

图 3 对光模块三维集成工艺流程进行了说明。首先，在前层上沉积一层牺牲层，然后将 4 个 TIA 芯片和 3 个 TSV 硅桥芯片放置于牺牲层上面，然后进行晶圆级的压膜注塑工艺，随后去除牺牲层和注塑膜架，并进行正面的晶圆级 RDL 和凸点工艺。最后进行切片，并进行 Die-to-Wafer 的键合。

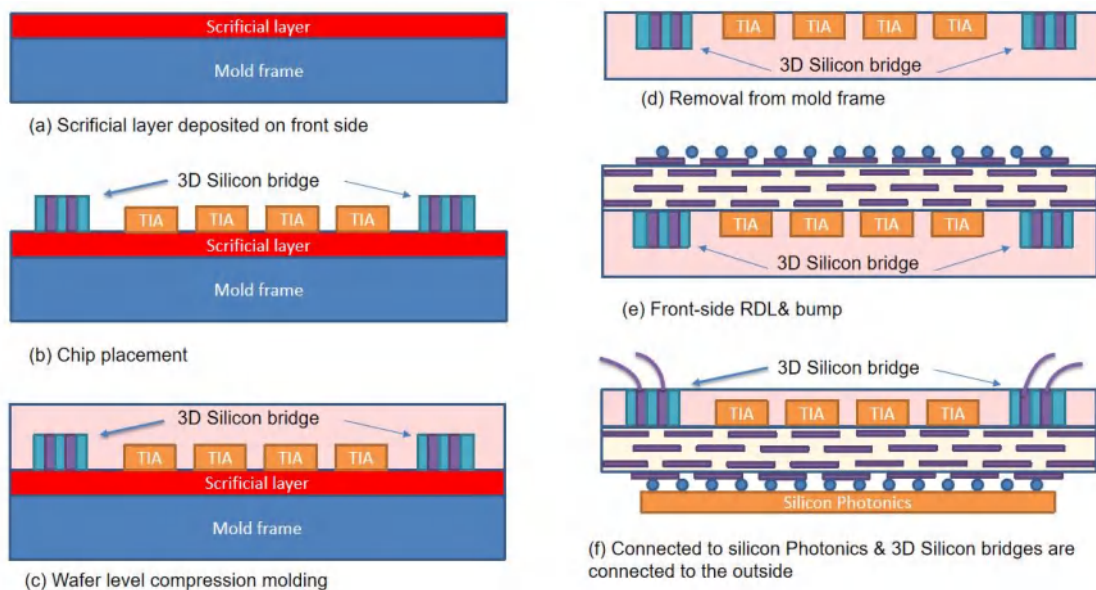


图 3 多通道光电微系统三维集成工艺流程

通过对 TIA 进行选型，并基于产品封装完成芯片布局及封装设计，如图 4 所示。硅光芯片的尺寸约为 $9850\mu\text{m} \times 4850\mu\text{m}$ ，符合合同限定的 $5\text{mm} \times 10\text{mm}$ 的平面尺寸。TIA 芯片的尺寸约为 $3.23\text{mm} \times 1.15\text{mm}$ ，四个 TIA 芯片左右对称放置于中央，左右以及上边各一个硅桥芯片。

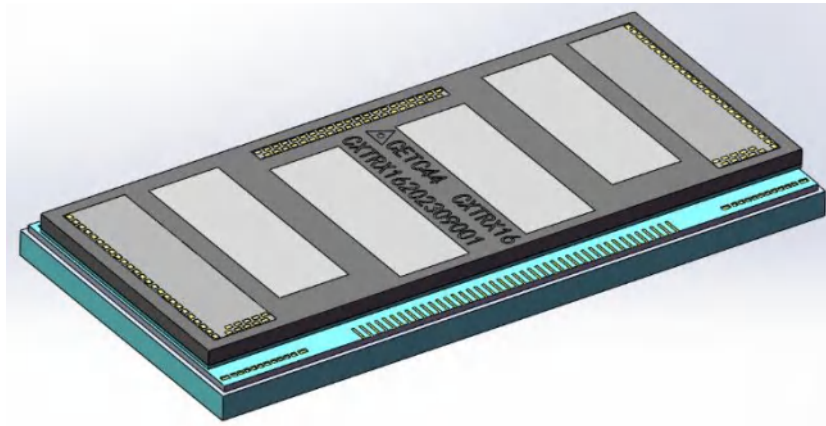


图 4 多通道光电微系统模组封装外形图

2.3 硅光芯片设计

图 5 所示为 16 通道收发硅光芯片，总尺寸为 $4.85\text{mm} \times 9.85\text{mm}$ 。16 通道的收发芯片采用 16 个调制器和 16 个探测器。调制器设计为 DSP 直驱，驱动电压 V_{pp} 大概在 1V。光接口采用周期为 127mm 的 42 通道阵列的端面耦合器，端面耦合器的模斑大概为 $3\mu\text{m}$ 。

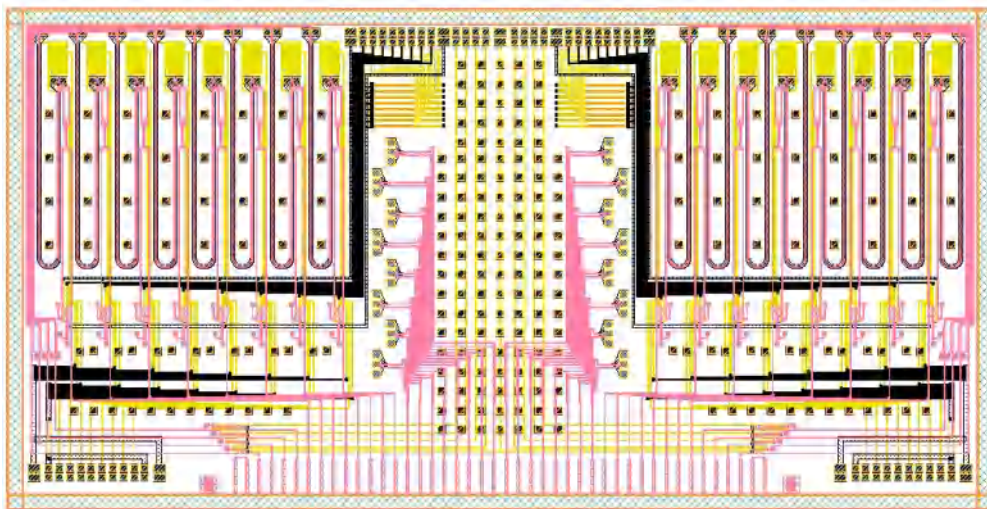


图 5 多通道收发硅光芯片总版图

硅光芯片一共 42 个端口，从左到右 1 和 42 端口相连，用于光纤耦合封装；光源输入后通过 MMI 分为两束，每一束进入到 2% 的功分器，98% 的分光进入到调制器的输入端，2% 的光进入到监测光电探测器，调制器的两端口的输出，每一端口通过 5% 功分器，95% 的光通过端面耦合器输出，两端口的 5% 光进入差分光电探测器。

2.4 三维集成 RDL 仿真设计

如下图所示为 16 通道高速光模块组硅光芯片的 RDL 接口方案,包括左右对称的 8 路 TIA 射频输出接口、探测器输入接口、TIA 直流接口。产品设计采用了 eWLB 方案进行 TIA 重塑晶圆布线的版图设计,包含了对 4 个 TIA 芯片进行重塑晶圆,与硅光芯片互连。而通过 TIA 放大的高速差分电信号输出以及外部电源、GND 以及 TIA 直流控制信号则通过 TSV 桥芯片从顶部引出。使用有限元/多物理场仿真等软件对高速差分信号、热分布/热应力完成了仿真设计,端-端的信号满足微模组产品要求。

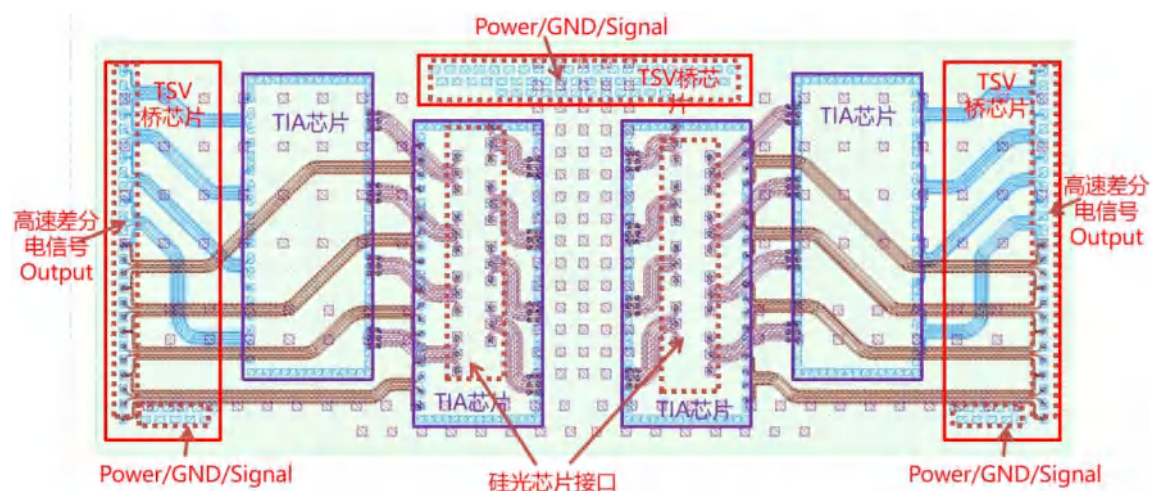


图 6 基于 TSV 硅桥的光电微系统模组三维集成 RDL 设计

3 多通道光电微系统测试

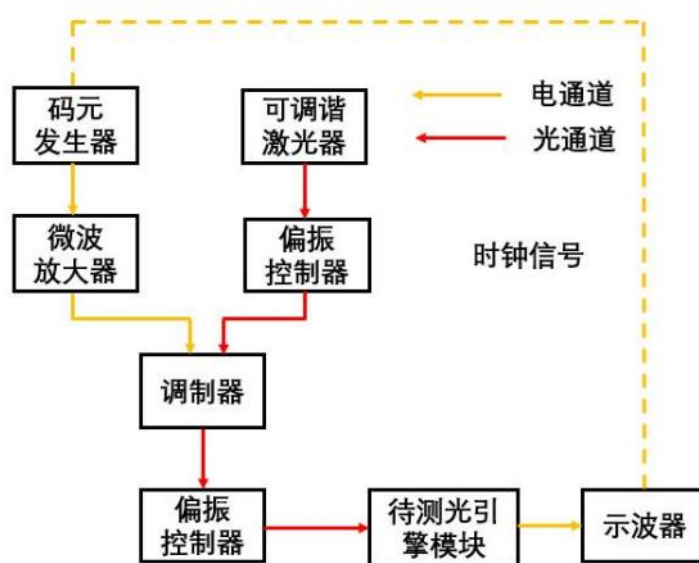


图 7 光模块集成样件收端工作速率测试方法框图

图 7 所示光模块集成样件收端工作速率测试方法框图, 码元发生器的输出信号经微波放大器放大后加载到调制器中, 通过调制器调制可调谐激光器输入的光信号, 然后输入到待测光模块集成样件中。输入的光信号通过光模块集成样件的发端转换为电信号, 并输入至示波器, 进行眼图速率测试。

对硅光芯片进行了大信号眼图测试, 如图 8 所示。可以发现, 3mm 调制臂长的调制器在 25Gbps 的传输眼图, 驱动电压 $< 3V_{pp}$, 消光比为 4.7dB。

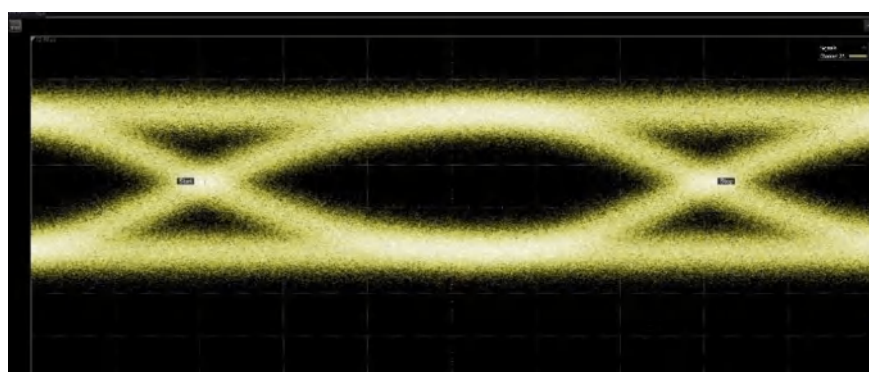


图 8 调制器眼图测试

多通道光电微系统模组研制完成后, 与 PCB 板进行贴装, 然后完成金丝键合, 最后进行封装测试。图 9 为多通道高速光模块组眼图速率测试结果, 其中 L1-L8 为左侧 8 个光电链路通道, 其中 R1-R8 为右侧 8 个光电链路通道。调出的示波器光口的眼图信号速率显示, 通道的速率均 $\geq 25Gbps$, 总速率为 $16 \times 25Gbps$, 误码率达 $1e-9$ 。通过改善封装工艺及倒装焊工艺, 完成了后续第 2 颗、3 颗多通道光电微系统模组的研制。

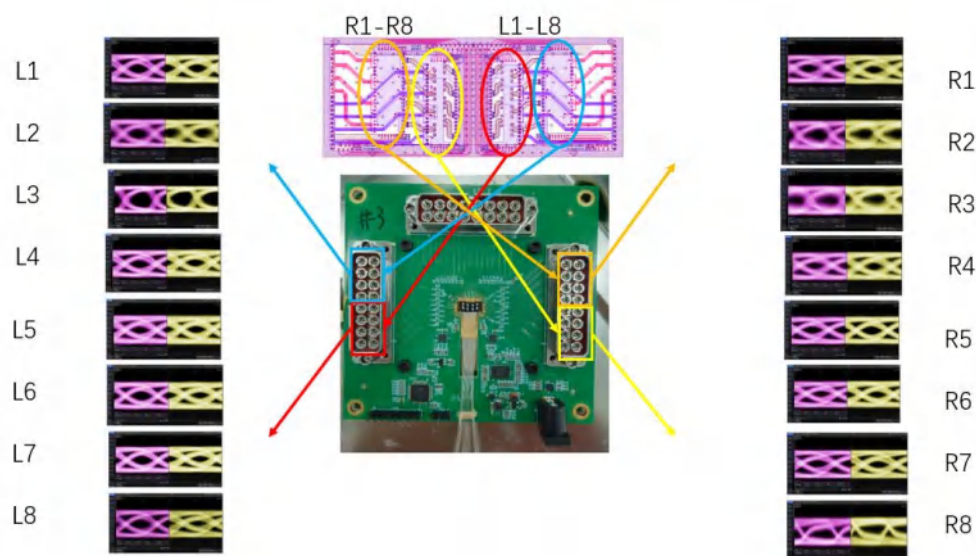


图 9 多通道高速光电微系统模组眼图速率测试

4 结论

本文基于先进封装工艺，针对典型 16 通道高速光电微系统，采用第三代晶圆级硅光三维集成微模组技术途径，首先进行了光电集成链路联合仿真设计，提出了一种基于三维异构集成方案，设计了 16 通道收发硅光芯片。通过晶圆重塑、TSV 硅桥与重布线技术，实现硅光芯片和电芯片的多芯粒三维异构集成互连。测试结果表明，采用 NRZ 调制方式的 16 通道光模块速率可达 400Gbps，误码率约 $1e-9$ ，眼图情况良好。该封装结构突破传统平面集成的限制，为今后的高密度高速光电微系统集成提供了技术支撑。

参考文献

- [1] 孙瑜, and 刘丰满. "高速高密度光电共封装技术." ZTE TECHNOLOGY JOURNAL (2018).
- [2] Tan, Min, et al. "Co-packaged optics (CPO): status, challenges, and solutions." *Frontiers of optoelectronics* 16.1 (2023): 1.
- [3] 蔡雪芳, et al. "基于封装集成的四通道微波光子电/光转换组件." *光通信技术* 44.6 (2020): 32-35.
- [4] Lau, John H. "Evolution, challenge, and outlook of TSV, 3D IC integration and 3D silicon integration." 2011 International symposium on advanced packaging materials (APM). IEEE, 2011.
- [5] 湛可馨, et al. "先进封装中硅通孔 (TSV) 铜互连电镀研究进展." *科技导报* 41.5 (2023): 15-26.
- [6] 卞玲艳, et al. "大数据时代光电共封技术的机遇与挑战." *Laser & Optoelectronics Progress* 61.9 (2024): 0900006.

超宽带 3D-SiP 仿真技术研究

闫玉凯¹, 彭亦童¹, 王曦茂¹, 侯健宏¹

(1 成都航天博目电子科技有限公司, 中国成都, 610036)

摘要:当前随着电子系统对小型化, 多功能, 高可靠性的要求越来越高, 电子技术正朝着三维集成的方向发展。系统级封装 (System in Package) 技术特别是 3D-SiP 可以实现立体封装, 较传统平面型封装密度更高, 更适用于系统的小型化需求。本文对超宽带 3D-SiP 的仿真技术进行了研究, 加工实物并进行测试对比, 验证了仿真方法的可行性准确性。

关键词:垂直互连; SiP; HTCC

1 前言

随着市场发展和技术进化, 系统级封装 (System in Package) 正成为电子设备发展的主流方向, 其具有集成度高, 成本低, 开发周期短, 生产效率高的特点。SiP 技术实现了从单片集成平面结构到多芯片集成立体结构的转变, 使高集成度、微小型化模块成为可能^[1-4]。文献[5]设计了一款基于低温共烧陶瓷 (Low temperature co-fired ceramic, LTCC) 技术的多功能收发模块, 该模块工作频率为 6~18GHz, 芯片安装在总层数为 13 层的 LTCC 基板两侧, 尺寸为 22mm*15mm*2mm。文献 [6] 基于晶圆级键合和硅通孔 (Through silicon via, TSV) 等三维集成技术研制了一款 X 波段四通道 TR 模块, 由两个子模块堆叠而成, 尺寸为 20mm*20mm*3mm。文献[7]基于陶瓷管壳对超宽带的互连结构进行了研究并对互连结构进行组合仿真得到整个模块集成设计的结果。

SiP 技术路线根据基材材质主要分为陶瓷基、有机基板和硅基。本研究聚焦陶瓷基板在超宽带收发模块中的应用, 通过建立三维全波电磁模型

对一款陶瓷基的超宽带收发模块进行了仿真技术研究, 并加工实物进行测试对比, 验证了仿真方法的可行性准确性。

2 模块叠层架构

图 1 示出了超宽带收发模块的叠层架构。模块共由三块基板组成, 分别为 DPC 工艺氮化铝基板、HTCC 工艺多层氧化铝基板和单层氧化铝基板, 各基板间通过可伐围框进行气密焊接, 同时基板间通过焊接铜柱实现射频, 电源和控制等信号的互连传输。

功放芯片置于氮化铝层以达到模块高效的散热, 多功能芯片位于多层的 HTCC 基板, 模块对外接口为 0.5mm 焊球, 置于单层的 HTCC 基板。

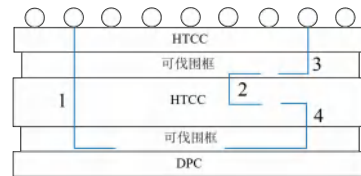


图 1 叠层架构

3 互连传输仿真设计

传统模块的射频链路通常在单一平面内进行传输, 而 3D-SiP 模块是多芯片立体集成的结构, 通常需要更加

复杂的传输转换。本文收发模块中的放大器芯片，多功能芯片位于不同层，同时模块输入输出口位于模块最顶层，由焊球作为对外接口，整个模块传输链路较为复杂，且需要实现超宽带的传输性能，因此超宽带的垂直互连传输结构成为实现模块性能的关键。

通过对模块划分得到了四种垂直互连传输结构：ALN层芯片-BGA、芯片-带状线、芯片-铜柱-焊球、芯片-铜柱-过孔-带状线，分别对应了图1中的路径1、2、3和4。

ALN层芯片-BGA传输结构用于功放芯片的信号发射输出。图2示出了ALN层芯片-BGA传输结构的仿真模型，信号由芯片输出至微带线，沿微带线传输至半包围铜柱，再依次经过多层AL₂O₃基板、全包围铜柱、单层AL₂O₃基板。该模型垂直传输段可视为类同轴结构，可根据公式(1)计算出50欧姆对应的初始尺寸信息。

$$Z_0 = \frac{60}{\sqrt{\epsilon_r}} \ln \frac{D}{d} \quad (1)$$

结合加工实际对类同轴结构做进一步调整最终得到传输结构的仿真结果如图2所示。可以看出2-18GHz范围内，反射系数不高于-20dB，插入损耗小于0.2dB。

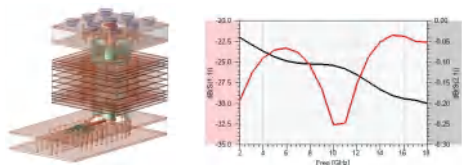


图2 ALN层芯片-BGA传输结构仿真模型及结果

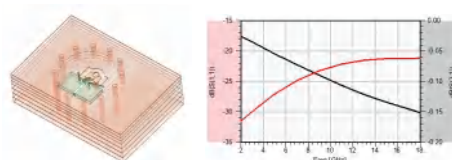


图3 芯片-过孔-带状线传输结构仿真模型及结果

芯片-过孔-带状线传输结构用于多功能芯片到内层带状线的垂直互连传输。该传输结构减少了对基板表层空间的占用，芯片输出后直接与表层匹配盘键合后垂直传输至内层带状线，提升布板密度及灵活性。

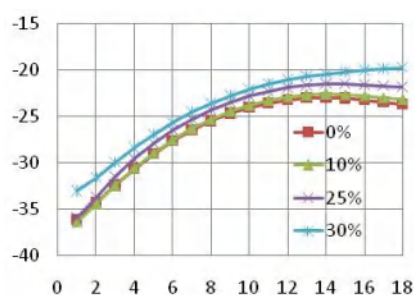


图4 反射系数随偏心度变化情况

由于多功能芯片收发焊盘间距过窄，此处通过调整中心过孔使其处于偏置状态，如图4所示，偏心度不超过35%情况下，传输性能变化有限，在偏心度不超多35%的情况下既能保证信号的良好传输，又解决了芯片收发焊盘间距过小无法直接键合传输的问题。图3示出了芯片-过孔-带状线传输结构的仿真结果。结果显示在2-18GHz范围内，反射系数不高于-20dB，插入损耗小于0.15dB。

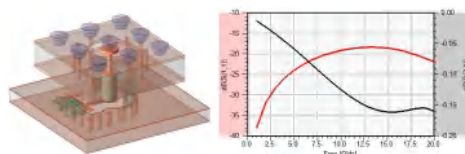


图5 芯片-铜柱-焊球传输结构仿真模型及结果

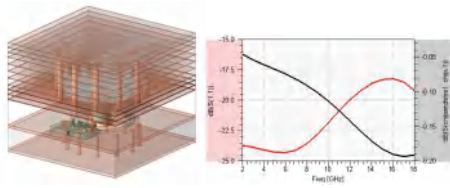


图 6 芯片-铜柱-过孔-带状线传输结构仿真模型及结果

芯片-铜柱-焊球传输结构（图 5）实现模块 BGA 至多功能芯片的信号传输；芯片-铜柱-过孔-带状线结构（图 7）实现 HTCC 多层基板内带状线至氮化铝(AIN)基板上芯片的信号传输。在芯片与铜柱的互连处，均采用了 T 型节 (T-junction) 匹配设计以优化阻抗连续性。仿真结果表明，在 2-18GHz 频段内，两种结构均表现出良好的传输特性：芯片-铜柱-焊球传输结构的 S11 优于 -18 dB，S21 低于 0.15 dB，芯片-铜柱-过孔-带状线结构 S11 优于 -18 dB，S21 低于 0.2dB。

从垂直互连结构的仿真结果来看，各个传输结构在 2-18GHz 超宽带范围内传输性能优良，适用于超宽带 3D-SiP 模块。

4 系统级封装仿真

根据垂直互连传输结构可以进行最终的布板设计，进而得到最终的版图文件。

HFSS 3D Layout 专门用于处理 2.5D 结构的仿真，尤其适合 PCB 这种叠层结构。本文采用 HFSS 3D Layout 对整个 SiP 模块封装进行建模，仿真模型如图 7 所示，图 8 为仿真模型的拆分示意图。

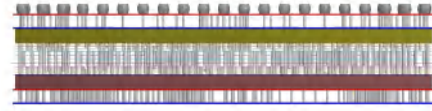


图 7 封装仿真模型

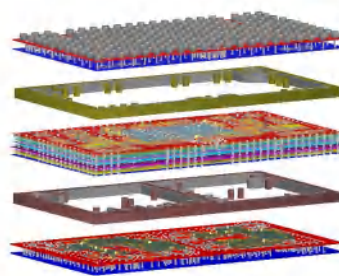


图 8 拆分示意图

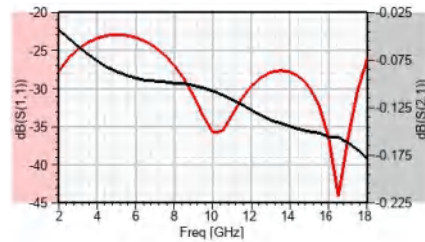


图 9 功放芯片输出口反射系数和插损

图 9 示出了功放芯片到模块焊球输出口的传输性能，从仿真结果看，反射系数不高于 -20dB，插入损耗小于 0.2dB，验证了垂直传输结构的正确性，同时反射系数和插入损耗曲线光滑无跳变反映出频带内无谐振频点且传输性能较好。

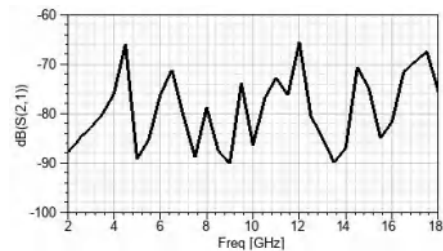


图 10 功放输出到驱放输入隔离

图 10 为功放芯片输出口和驱放芯片入口之间的隔离度仿真情况，从仿真结构看隔离度不低于 -65dB。

模块封装仿真得到整个管壳的无源射频传输性能，并联合芯片 S 参数

仿真得到整个 SiP 模块的射频传输性能。图 11 为场路联合的仿真链路图。

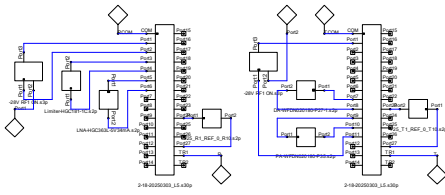


图 11 场路联合仿真链路

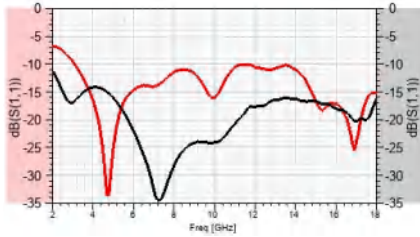


图 12 模块接收输入和输出反射系数

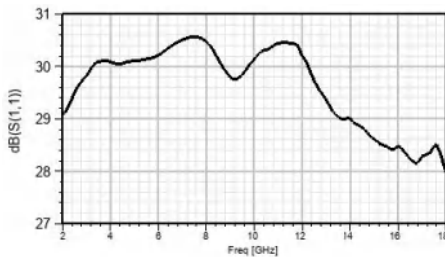


图 13 模块接收增益

图 12 示出了模块的接收输入和输出反射系数，从仿真结果看接收输入反射系统 2-3GHz 范围内不高于-7dB，3-18GHz 范围内优于-10dB；图 13 示出了模块接收增益仿真结果，全频带增益不低于 28dB。

5 测试与验证

模块尺寸为 9.6mm*18mm*3.7mm，对外接口为 0.5mm 焊球，加工实物如图 14 所示。

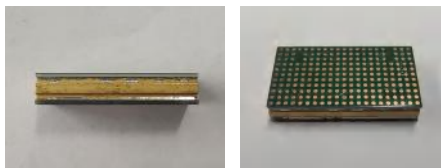


图 14 模块实物图

根据该模块的接口定制相应测试工装，通过测试工装对模块进行测试，最终测试结果如图 15-17 所示。

结果显示模块输入反射系数和输出反射系数的实测结果和仿真结果整体水平一致，接收增益的实测结果和仿真结果趋势一致，最大增益偏差为 0.68dB。

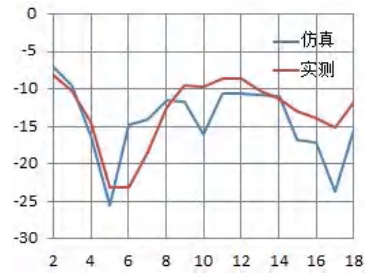


图 15 接收输入反射系数结果

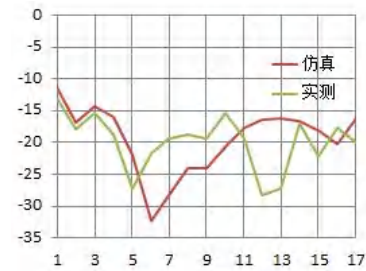


图 16 接收输出反射系数结果

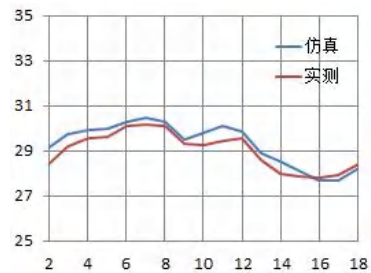


图 17 接收增益

从实测结果和仿真结果的对比可以看出模块的实测结果和仿真结果基本吻合，说明了仿真方法的正确性和可行性，能够为后续工程应用提供重要借鉴意义。

6 结论

本文通过建立三维全波电磁模型对一款陶瓷基的超宽带收发模块进行了仿真技术研究。首先,针对模块中涉及到的互连传输结构进行无源仿真,得到局部的具有宽带传输的互连传输模型,其次,结合芯片和封装进行了系统级封装的电磁仿真,最后对加工的实物进行相关测试。从测试结果和仿真结果的对比来看,仿真结果和实物测试结果对比一致性较高,满足了相关技术要求,同时验证了仿真方法可行性,具有一定借鉴意义。

参考文献

- [1]陆燕菲. 集成电路封装技术现状分析与研究[J]. 电子技术, 2020, 49(08):8-9.
- [2]赵燕, 吴鹏, 王志凌, 等. 微波毫米波 SiP 三维集成陶瓷封装基板研究进展[J]. 微波学报, 2024, 40(S1):256-259.
- [3]李庆东, 姚剑平, 苟明艺, 等. 基于多层有机复合基板的射频 SIP 研究[J]. 微波学报, 2024, 40(S1):244-247.
- [4]姚剑平, 李庆东, 管慧娟, 等. 基于 SiP 应用的多层有机复合基板的三维堆叠[J]. 电子与封装, 2025, 25(04):52-55. DOI:10.16257/j.cnki.1681-1070.2025.0044.
- [5]张鸣一, 朱春雨, 刘文豹, 等. 基于三维集成技术的 Ku 波段四通道 T/R 模块[J]. 半导体技术, 2021, 46(04):286-289+329. DOI:10.13290/j.cnki.bdtjs.2021.04.005.
- [6]王欢鹏. 射频微系统互连可靠性建模与仿真研究[D]. 四川:电子科技大学, 2021.
- [7]吕嘉然, 阚尧, 刘士杰, 等. 面向 3D-SiP 模

组的超宽带互联技术研究[J]. 固体电子学研究与进展, 2025, 45(02):51-57.

Research Article

Design of Heterogeneous Integrated X-band Front-end Module Based on High-Precision Multi-Field Co-simulation

Wang Huanpeng^{1,2}, Tang Yue^{1,2}, Song Weixuan^{1,2}, Tang Can¹, Xu Yuehang^{1,2}

¹ University of Electronic Science and Technology of China, No. 2006 Xiyuan Avenue, High-Tech Zone (West District), Chengdu, 611731, China

² Yangtze River Delta Research Institute (Huzhou) of University of Electronic Science and Technology of China, No. 819 Xisai Mountain Road, Huzhou, 313001, China

Corresponding author: Xu Yuehang; Email: yuehangxu@uestc.edu.cn.

Abstract — A model-driven heterogeneous integrated multi-field synergy simulation is proposed in this paper to address the issue of lacking high-precision simulation results for radio frequency (RF) front-end microsystems under heterogeneous integration processes. Firstly, the circuit simulation model of active chips, and the electromagnetic field models of passive chips and heterogeneous integrated interconnections are constructed. Secondly, the impact of heterogeneous integrated packaging on the chip circuit is analyzed, and an equivalent circuit including a compensation form is established to improve the link simulation of the X-band four-channel front-end module. Finally, by comparing the simulation results with the measured data, it is obtained that the four-channel transmit power error, four-channel transmit gain error, four-channel noise reception error, and four-channel receive gain error are all $\leq 11.5\%$.

Keywords — multi-field Co-simulation, heterogeneous integration, X-band, RF front-end.

I. Introduction

Miniaturization is an inevitable trend in the development of microwave and millimeter-wave integrated circuits and systems[1]. With the increasing complexity of integrated circuit manufacturing processes, the cost of tape-out has risen sharply, posing significant challenges to chip design. The transistor size of advanced process chips has approached the quantum physical limit, which has led to huge investment and difficulties in developing system-on-chip (SoC) under advanced processes [2-4]. Thus, people have begun to explore the possibility of multi-chip heterogeneous integration. From multi-chip modules (MCM) to 2.5D/3D packaging, and then to hybrid bonding, heterogeneous integration technology integrates chips and devices of different materials, processes, and functions into the same system through Through-Silicon-Via (TSV), Redistribution Layer (RDL), micro-bump/copper pillar and other processes, continuously improving functional density and reducing costs, thereby promoting the reduction of electronic product costs and the advancement of performance. Heterogeneous integration has become a research hotspot and development trend in the field of RF microsystems due to its significant advantages in improving system performance, optimizing functional density, and controlling costs, attracting extensive attention and in-depth exploration from academia and industry. In 2020, the Defense Advanced Research Projects Agency (DARPA) of the United States vertically stacked InP low-noise amplifiers, GaN power amplifiers, and

SiCMOS digital beamforming chips

using three-dimensional heterogeneous integration technologies such as copper pillar bumps and hot vias, developing a component-level digital millimeter-wave phased array [5]. The array size was reduced by 50% compared with traditional schemes, and the beamforming time was less than 1 μ s. In 2022, Shanghai Jiao Tong University integrated X-band CMOS-based phase-locked loop chips, W-band SiGe millimeter-wave chips, GaN power amplifier chips, capacitors, TaN resistors, filters, and other passive components to design a frequency-modulated continuous wave radar sensor [6], with an output power of 22 dBm and an overall volume of only 60×40×8 mm³. In 2024, the 13th Research Institute of China Electronics Technology Group Corporation (CETC) adopted 5-layer silicon wafer stacking and TSV technology to integrate functional modules such as 6-bit digitally controlled phase shifters, attenuators, and serial-to-parallel circuits in the transmit / receive (T/R) module [7], with an overall size of only 15×8×3.8 mm, a Ku-band transmitting channel saturated output power of >24 dBm, and a receiving channel noise figure of <3.0 dB.

The high-density integration characteristic of RF heterogeneous integrated microsystems leads to strong coupling effects among multiple physical fields (such as electromagnetic fields, thermal fields, mechanical stress fields, and even chemical fields) in their operating environment and internal interactions. With the development trend of RF heterogeneous integrated

microsystems towards higher density, frequency, and power, multi-physical field coupling effects have become a key factor restricting system performance and reliability, and multi-physical field collaborative simulation analysis has also become an effective method for evaluating heterogeneous integration designs. LI [8] et al. considered the influence of factors such as the internal and external operating environments of chips from a multi-level perspective and analyzed electromagnetic fields, thermal fields, and mechanical fields using collaborative simulation. Ma [9] et al. considered the Joule heating effect and the impact of temperature on IR drop, proposed an electro-thermal coupling model and an equivalent conductivity model for chiplet heterogeneous integration (CHI), and realized the collaborative simulation of voltage and temperature fields in multi-scale CHI systems. KIMMO [10] et al. proposed a multi-physical modeling and simulation workflow for heterogeneous integrated GaN-on-Si modules in millimeter-wave communications, which effectively predicts the performance of millimeter-wave integrated modules in different spatio-temporal scenarios and provides a quantitative analysis tool for solving thermal management, electromagnetic compatibility, and reliability issues caused by high-density integration. In radio frequency (RF) heterogeneous integrated microsystems, digital, analog, and digital-analog hybrid signal modes coexist simultaneously, along with the coexistence of small signals and high-power signals. This often leads to problems such as abnormal transceiver functions and failure to work normally due to electromagnetic interference and leakage in the cavity [11]. Traditional active circuit simulations fail to take into account the impact of electromagnetic energy on device performance, resulting in large deviations between the simulation results and the actual performance.

Therefore, for an X-band front-end module, 3D electromagnetic modeling is performed for passive chips and heterogeneous integrated interconnection structures in this paper. Through active chip circuit models combined with S-parameter files, a model-driven heterogeneous integrated multi-field collaborative simulation model is further established. Moreover, the influences of heterogeneous integrated packaging on internal circuits, such as packaging fan-out wiring effects, on-chip microstrip line proximity effects, and inductor proximity effects, are incorporated to achieve multi-physical field link simulations with higher precision.

Co-simulation modeling method

1. Chips and Interconnection Model Design

1.1 RF Chips Modeling

The X-band four-channel front-end module used in this paper integrates active chips such as single-pole double-throw switches (SPDT), low-noise amplifiers (LNA), power amplifiers (PA), and amplitude-phase multi-functional chips(MFC), as well as a one-to-four power divider (PD) passive chip and 3D heterogeneous interconnection structures connecting various chips.

Common modeling methods for equivalent circuit modeling of chips include the modeling method based on deriving electrical characteristic equations from physical structures, the modeling method based on test data combined with mathematical fitting and parameter extraction, and the semi-empirical modeling method combining partial physical principles and data correction. Since traditional S-parameters have certain limitations in dealing with networks with non-reciprocal characteristics or complex port relationships, to describe the characteristics of multi-port networks in RF and microwave circuits and effectively handle signal transmission problems under mismatched impedances of different ports, this paper adopts an improved K-parameter model to model the power amplifier chip. It has been verified that the simulation results of this model are highly consistent with those of the power amplifier chip IP core model.

To reflect the electromagnetic coupling and signal transmission loss characteristics of the chip structure in two-dimensional space, the S2D model is constructed by extracting the scattering parameters of other active chips through electromagnetic field simulation. It has been verified that the simulation results of the model are highly consistent with the actual test results. The circuit models of the four active chips are packaged as IP core circuit models, as shown in Figure 1.

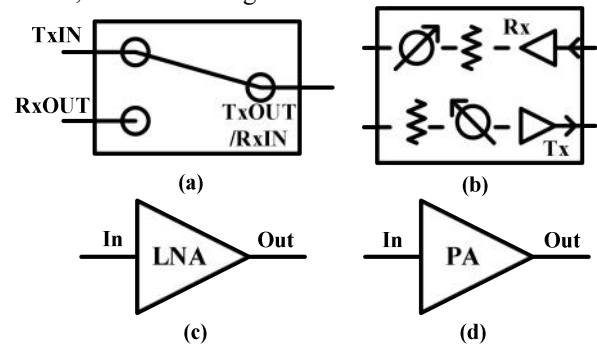
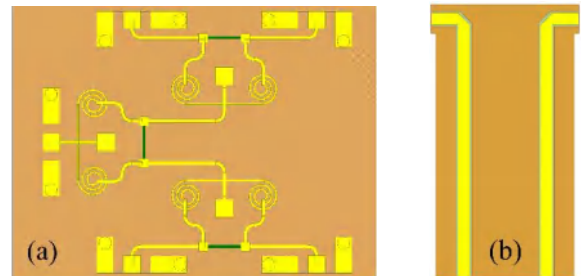


Figure 1 (a)SPDT chip (b) MFC chip (c) LNA chip (d) PA chip

1.2 Electromagnetic field model of PD and interconnected structures



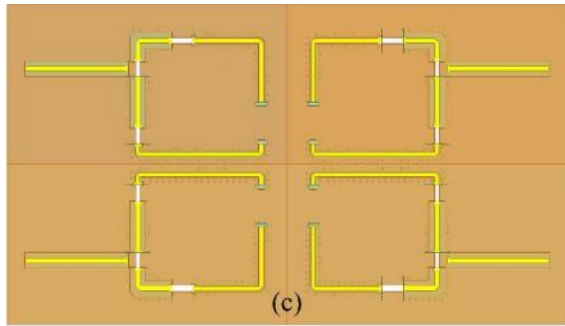


Figure 2 Electromagnetic field models (a) One-to-four PD chip (b) Interconnection structure connecting the PD and MFC chip (c) Module RF substrate

The full-wave electromagnetic field models of the one-to-four PD chip and module interconnection structure are constructed in electromagnetic simulation software, as shown in Figure 2. Its S-parameters are extracted and encapsulated in the custom component library of EDA simulation software in the form of an equivalent network model, forming a directly callable collaborative simulation module that provides high-precision characterization of the electromagnetic properties of the interconnection structure.

2 Design of Compensation Circuit for Packaging ImpactAnalysis of Packaging Parasitic Parameters

In heterogeneous integrated systems, packaging, as a key link connecting chips and external circuits, its parasitic parameters will have an undeniable impact on signal transmission and system performance. To offset the negative impact of packaging parasitic parameters and ensure that signals can be transmitted stably and accurately according to design requirements, a compensation circuit designed for the impact of heterogeneous integrated packaging is shown in Figure 3 in this paper. Firstly, the parasitic capacitor C_p is used to characterize the parasitic effect between the RDL and the ground plane of the underlying PCB. Due to the flip-chip packaging method adopted for the component, the chip surface is adjacent to the underlying PCB, and electromagnetic crosstalk will occur between them, which leads to an increase in the dielectric constant of the chip. This further results in transistor impedance mismatch, namely the detuning effect [9-10], and this phenomenon is particularly severe when there is metal present on the underlying PCB. Therefore, a parallel branch composed of C_{GND} and R_{GND} is added at this position. When the spiral inductor of the flip chip is adjacent to the underlying PCB ground plane, a mirror current with the opposite direction to the inductor current will be generated in the underlying ground plane [11], which causes a reduction in the actual inductance value of the on-chip spiral inductor. Hence, a parasitic inductor L_{chip} is added in this region to characterize the proximity effect of the on-chip spiral inductor.

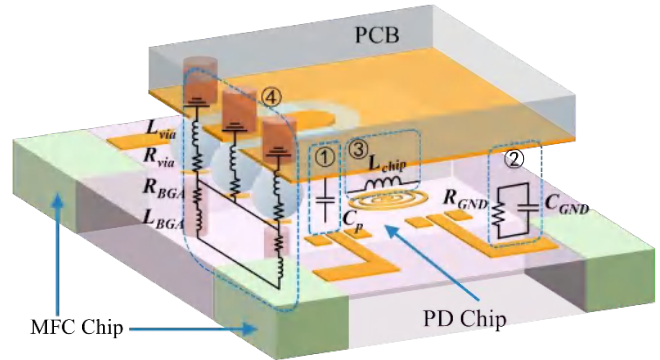


Figure 3 Parasitic parameters generated by packaging

The multi-functional chip was originally grounded through vias on the back of the chip, but after packaging, the chip is only connected to the underlying PCB ground plane through four solder balls, resulting in poor grounding. At the same time, the grounding path of other grounding points except the chip input/output ports is increased. During on-chip testing, the return current only needs to pass through the via to reach the ground plane. When a return current is generated by any grounding via on the chip close to the input/output port, the path of this current includes two chip vias, the chip ground plane, and the solder ball. From the equivalent circuit of the solder ball and via, it can be concluded that this situation will lead to an increase in the loop inductances L_{BGA} , L_{via} and resistances R_{BGA} , R_{via} [12].

2.2 Model-Driven Heterogeneous Integrated Multi-Field Co-Simulation

The RF substrate of the module integrates PA chips, switch chips, and LNA chips. The MFC chip and PD chip packaged as a whole are connected to the RF substrate through flip-chip bonding to form a complete RF chain. Therefore, the schematic diagram of the model-driven heterogeneous integrated multi-field co-simulation model is shown in Figure 4.

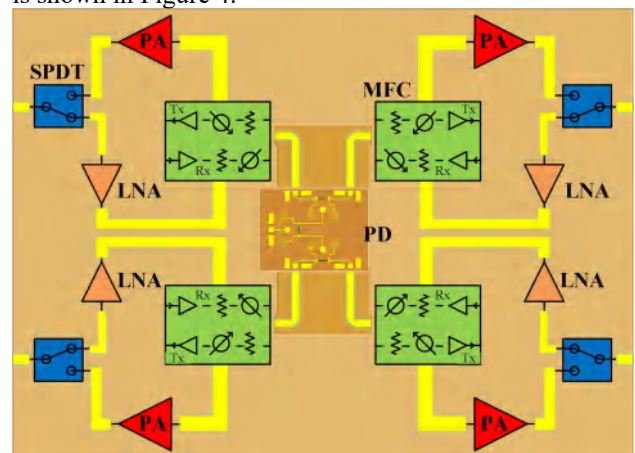


Figure 4 Schematic diagram of the model-driven heterogeneous integrated multi-field co-simulation

Figure 4 shows the simulation schematic of the RF chain. The circuit models corresponding to active chips such as SPDT switches, LNA, PA, and MFC chips, as well as the simulation module with packaging S-parameters corresponding to the one-to-four PD passive chip and module interconnection structure, are used for

co-simulation. The impact of electromagnetic coupling on device performance is considered in the model to improve the prediction accuracy of the simulation.

II. Result and Discussion

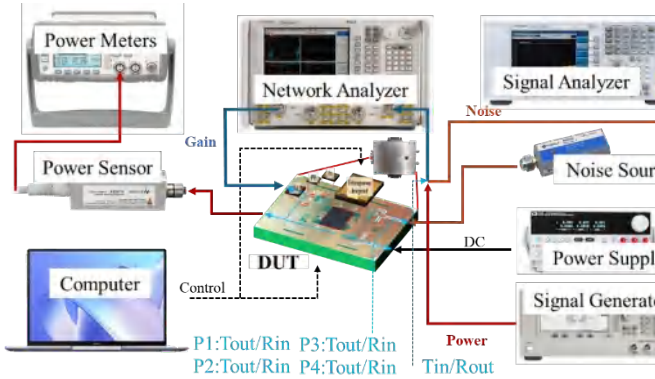


Figure 5 The block diagram of the module test

The physical diagram of the module and the block diagram of the measured setup is shown in Figure 5. The T/R mode as well as the attenuation and phase shift states are controlled by a computer. Based on the connection method shown in the diagram, model-driven multi-field collaborative simulations are performed on the receiving and transmitting links respectively. Index parameters such as the transmitting output power, transmitting branch gain, receiving noise figure, and receiving branch gain of the module is obtained. The comparison between these parameters and the actual test results of the module is shown in Figure 6.

The average error of the four-channel transmit power in co-simulation is 2.47%, the maximum error is 9.30%, and the minimum error is 0.04%; the average error of the four-channel transmit gain in co-simulation is 2.19%, the maximum error is 8.60%, and the minimum error is 0.01%; the average error of the four-channel receive noise in co-simulation is 4.25%, the maximum error is 11.49%, and the minimum error is 0.14%; the average error of the four-channel receive gain in co-simulation is 2.63%, the maximum error is 10.01%, and the minimum error is 0.04%. Through comparative analysis, the simulation results are in good agreement with the measured data, and the four-channel transmit power error, transmit gain error, receive noise error, and receive gain error are all within 11.5%, which proves that this model has strong predictive ability and simulation accuracy.

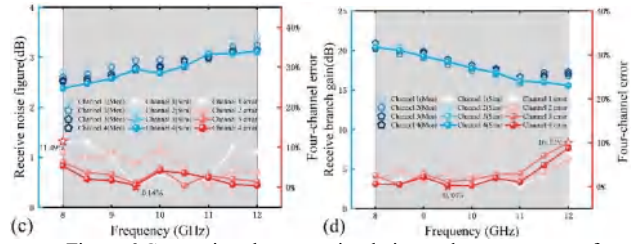
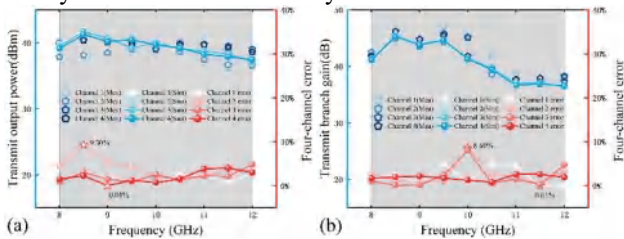


Figure 6 Comparison between simulation and measurement of X-band front-end module (a) Transmit output power (b) Transmit branch gain (c) Receive noise figure (d) Receive branch gain

III. Conclusion

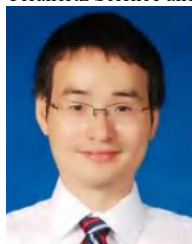
In this paper, a model-driven heterogeneous integrated multi-field synergy simulation model is proposed for the link simulation of the X-band four-channel front-end module. By establishing circuit models for active chips, constructing electromagnetic field models for passive chips and heterogeneous integrated interconnections and packaging them as simulation modules with S-parameter characteristics, and establishing a compensation equivalent circuit for the chip considering the fan-out wiring effect, on-chip microstrip line proximity effect, and inductor proximity effect caused by heterogeneous integrated packaging, a model-driven heterogeneous integrated multi-field synergy simulation model is constructed. Comparison and verification between link simulation and measured data show that the model exhibits excellent prediction accuracy within the X-band: the four-channel transmit power error is within 9.3%, the four-channel transmit gain error is within 8.6%, the four-channel noise reception error is within 11.49%, and the four-channel receive gain error is within 10.01%. The multi-field collaborative simulation model proposed in this paper not only provides a high-precision simulation method for the design verification of X-band heterogeneous integrated front-end modules, but also its modeling ideas and compensation methods can serve as a reference for the electromagnetic-circuit collaborative simulation of multi-chip heterogeneous integrated systems in the millimeter-wave band, which is beneficial to reducing the research and development cost and cycle of high-end RF microsystems.

References

- [1] XU Rui-min, WANG Huan-peng, XU Yue-hang. Progress and Prospects of Key Technologies in RF Microsystems[J]. *Journal of Microwaves*, 2023, 39(5): 70-78.
- [2] AGARWALR,CHENGP,SHAHP,etal. 3D packaging for heterogeneous integration[C]// 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), San Diego, 2022.
- [3] SHEIKH F, NAGISETTY R, KARNIK T, et al. 2.5D and 3D heterogeneous integration: Emerging applications [J]. *IEEE Solid-State Circuits Magazine*, 2021, 13(4): 77-87.
- [4] WOLF M J, RAMM P, KLUMPP A, et al. Technologies for 3D wafer level heterogeneous integration[C]// 2008 Symposium on Design, Test, Integration and Packaging of MEMS/MOEMS,Nice, 2008.
- [5] T. M. Hancock et al., "The DARPA Millimeter Wave Digital Arrays (MIDAS) Program," 2020 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS), Monterey, CA, USA, 2020, pp. 1-4, doi: 10.1109/BCICTS48439.2020.9392956.
- [6] X. Yang et al., "Low-Loss Heterogeneous Integrations With High Output Power Radar Applications at W-Band," in *IEEE Journal of Solid-State Circuits*, vol. 57, no. 6, pp. 1563-1577, June 2022, doi:

10.1109/JSSC.2021.3106444.

[7] Chen Xing, Zhang Chao, Li Xiaolin, et al. A T/R module based on silicon-based MEMS 3D integration technology[J]. Journal of Terahertz Science and Electronic Information Technology, 2024, 22(03): 331-336.



[8] LI Z R, XU G, HUANG X J, et al. Evaluation method of electronic performance margins for SiP based on field-circuit and multiple physical cooperative simulation[C]. 2019 22nd European Microelectronics and Packaging Conference & Exhibition (EMPC), 2019: 1-6.

[9] X. Ma et al., "An Electrical-Thermal Co-Simulation Model of Chiplet Heterogeneous Integration Systems," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 32, no. 10, pp. 1769-1781, Oct. 2024, doi: 10.1109/TVLSI.2024.3430498.

[10] KIMMO R, KOEN B, KRISTOFFER A, et al. Multi-physical simulations and modelling of an integrated GaN-on-Si module concept for millimetre-wave communications[C]. 2020 IEEE 70th Electronic Components and Technology Conference, 2020, 1369-1375.

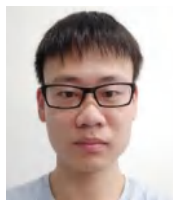
[11] Axelsson K. Generic packaging technologies for T/R-modules [C] //Microwave Conference, 2005 European. IEEE, 2005.

[12] Nagy, P. Heide, A. Springer and R. Weigel, "An Investigation of the Proximity Effect of Millimeter-Wave MMICs in Flip-Chip Configuration," 2001 31st European Microwave Conference, London, UK, 2001, pp. 1-4, doi: 10.1109/EUMA.2001.339156.

[13] W. Heinrich, "The flip-chip approach for millimeter wave packaging," in IEEE Microwave Magazine, vol. 6, no. 3, pp. 36-45, Sept. 2005, doi: 10.1109/MMW.2005.1511912.

[14] K. Song, J. Gao, G. T. Flowers, Z. Wang, W. Yi and Z. Cheng, "Modeling and Analysis of Signal Integrity of Ball Grid Array Packages With Failed Ground Solder Balls," in IEEE transactions on Components, Packaging and Manufacturing Technology, vol. 12, no. 2, pp. 306-315, Feb. 2022, doi: 10.1109/TCPMT.2022.3143110.

[15] K. . -C. Lu et al., "Packaging effects on a CMOS low-noise amplifier: Flip-chip versus wirebond," 2009 59th Electronic Components and Technology Conference, San Diego, CA, USA,



2009, pp. 20642069, doi: 10.1109/ECTC.2009.5074308.



Huanpeng Wang received the B.S. degree in electronic information science and technology from Fuzhou University, Fuzhou, China, in 2017, and the M.S. degree in electromagnetic field and electromagnetic waves from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2021. He is currently pursuing the Ph.D. degree. His research interests include RF circuits, RF microsystems, and advanced packaging technologies. (Email: uestcwhp@foxmail.com)



Yue Tang received the B.S. degree in communication engineering from the Shandong University, Qingdao, China, in 2024. She is currently pursuing the M.S. degree in electronic information with the University of Electronic Science and Technology of China (UESTC), Chengdu, China. Her research interests include radio frequency (RF) microsystems and advanced packaging technologies (Email: tyawct@163.com)



Can Tang received the B.S. degree from Sichuan University, Sichuan, China, in 2024. She is currently an M.S. candidate at the School of Electronic Engineering, University of Electronic Science and Technology of China, Sichuan, China.

Her research interest is RF microsystems. (Email: 202421021324@std.uestc.edu.cn)

Weixuan Song was born in Hulunbeier, Inner Mongolia, he is currently studying at the University of Electronic Science and Technology of China. His inerests include T/R module and chip packaging. (Email: swx1347@163.com)

Yuehang Xu (Senior Member, IEEE) received the B.S. and M.S. degrees in electromagnetic field and microwave techniques from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2004 and 2007, respectively, and the joint Ph.D. degree from UESTC and Columbia University, New York City, NY, USA, in 2010.

In 2010, he joined the Faculty of the Department of Electronic Engineering, UESTC, where he is currently a Professor. He was a Visiting Associate Professor with Case Western Reserve University, Cleveland, OH, USA, in 2016. He is the first author of Equivalent Circuit Modeling of Microwave GaN Power Transistors (China Science Press, 2017) and has authored or coauthored more than 150 scientific papers in international journals and conferences. His current research interests include the modeling of radio frequency (RF) electronic devices and microwave integrated circuits.

Dr. Xu has been a Technical Program Committee (TPC) Member of IEEE International Conference on Integrated Circuits Technologies and Applications (ICTA) since 2019. He has won three provincial and ministerial science and technology awards since 2016. He was awarded the National Natural Science Foundation of China (NSFC) Excellent Young Scholars in 2019. He was the General Chair of MOS-AK in 2019. He has served as an Associate Editor for International Journal of Numerical Modelling: Electronic Networks, Devices and Fields (IJNM), a Special Issue Guest Editor for Microwave and Optical Technology Letters (MOTL), and an Editorial Committee Member for Research and Progress of Solid-State Electronics and International Journal of High Speed Electronics and Systems.

(Email: yuehangxu@uestc.edu.cn)

基于相位调制的宽范围可调谐射频电感建模与仿真

郑彦文, 杨纯钰, 单光宝

西安电子科技大学

摘要: 可重构射频收发器在脑机接口、便携式穿戴、植入式医疗以及航天航空等领域发挥着关键作用。它的面积和成本很大程度上取决于电感的性能和尺寸。现有的可调谐电感的调谐范围有限, 体积庞大, 工作频率低, 限制了可重构射频收发器的应用。本文提出了一种磁耦合电感调谐技术, 建立机硅基电感模型, 仿真结果表明电感值调谐范围为 $0.2 \sim 3.2$ nH (自感 1.03 nH), 调谐范围为 93.75% , 相位差为 3° 时品质因子(Q)为 32.8 。这意味着在不增加面积的情况下, 可以获得 3.11 倍的电感值。

关键词: 磁耦合; 可调谐; 电感

1 引言

小型化、低成本的可重构射频收发器为脑机接口和植入式医疗的发展提供了关键技术支撑。应用该技术能够实现微创化植入, 不仅降低了对患者组织的损伤风险, 也提升了患者的舒适度^[1,2]。此外, 可重构射频收发器在便携式穿戴、航天航空等领域也发挥着关键作用。然而, 射频收发器中电感器在性能和尺寸上的瓶颈, 以及高制造成本限制了其在脑机接口和植入式医疗中的普及应用。为克服上述电感器瓶颈, 可调谐电感器被提出作为优化可重构射频收发器的关键途径。现有的主流可调谐电感器技术面临诸多挑战: 有源电感在高频段 Q 因子会显著降低^[4-6]; 微带线电感虽成本较低, 但面积大且集成度差^[7,8]; 开关电感存在面积过大的问题^[9,10]; 磁芯电感调制方式简单, 但其调谐范围较窄且面积较大^[11-13]。针对现有可调谐电感器工作频率受限、体积大、调谐范围窄等限制, 本文提出了一种宽频率覆盖、宽调谐范围的磁耦合电感调谐技术, 并基于硅基集成无源器件 (IPD) 工艺进行建模和仿真验证, 为低成本制造提供了技术基础。

2 调谐理论分析

2.1 可调谐电感模型

如图 1 所示, 建立了基于 IPD 技术的可调谐电感器模型。 C_{p1} 、 C_{p2} 为钝化层的寄生电容, C_{ox1} 、 C_{ox2} 为绝缘层的寄生电容, L_1 、 R_1 、 L_2 、 R_2 为上下金属线圈的等效电感和等效电阻, C_{s1} 、 C_{s2} 为同层金属线圈之间的等效电容, C_{si1} 、 C_{si2} 分别为衬底的寄生电容。这里将 L_1 所在回路称为辅助回路, L_2 所在回路称为主回路。可调谐电感使用时候, 主回路接入应用电路, 辅助回路接入控制电路。

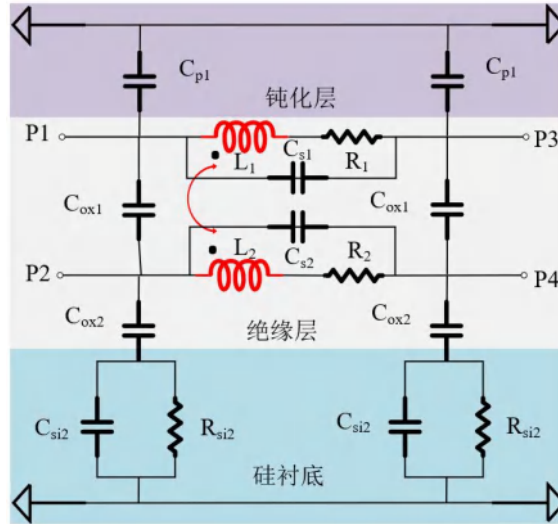


图 1 硅基可调谐电感模型

2.2 调谐原理分析

当主回路与辅助回路端口电压分别为 V_1 、 V_2 且存在相位差时，根据楞次定律， I_1 产生的磁场总是要阻碍主电路中 I_2 产生的磁场变化趋势，产生感应电流。此时在主电路和辅助电路中存在磁耦合现象，等效电路如图 2 所示。

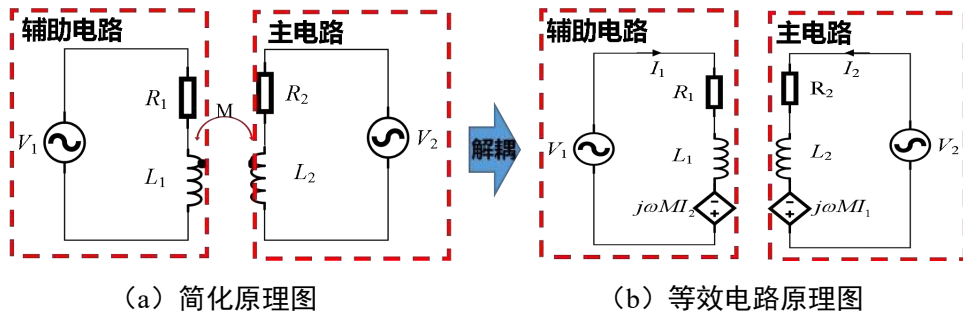


图 2 电感调谐原理图

主电路和辅助电路的电压方程分别如 (1) 和 (2) 所示。

$$V_1 = R_1 I_1 + j\omega L_1 - j\omega M I_2 \quad (1)$$

$$V_2 = R_2 I_2 + j\omega L_2 - j\omega M I_1 \quad (2)$$

(1) 和 (2) 可以用 (3) 所示的矩阵形式进行表示

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} R_1 + j\omega L_1 & -j\omega M \\ -j\omega M & R_2 + j\omega L_2 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \quad (3)$$

利用上述矩阵可以求解主电路和辅助电路中的电流 I_1 、 I_2 。 I_1 和 I_2 的表达式分别如式 (4) 和 (5) 所示

$$I_1 = \frac{R_2 V_1 + j\omega(L_2 V_1 + M V_2)}{(\omega M)^2 + R_1 R_2 - \omega^2 L_1 L_2 + j\omega(L_1 R_2 + L_2 R_1)} \quad (4)$$

$$I_2 = \frac{R_1 V_2 + j\omega(L_1 V_2 + M V_1)}{(\omega M)^2 + R_1 R_2 - \omega^2 L_1 L_2 + j\omega(L_1 R_2 + L_2 R_1)} \quad (5)$$

电流比例系数 α 形式如（6）所示

$$\alpha = -\frac{I_1}{I_2} = -\frac{R_2 V_1 + j\omega(L_2 V_1 + M V_2)}{R_1 V_2 + j\omega(L_1 V_2 + M V_1)} \quad (6)$$

如果 R_1 、 R_2 很小， α 可以进一步化简为（7）所示的形式

$$\alpha = -\frac{L_2 \frac{V_1}{V_2} + M}{L_1 + M \frac{V_1}{V_2}} = -\frac{L_2 \frac{|V_1|}{|V_2|} \angle(\varphi_1 - \varphi_2) + M}{L_1 + M \frac{|V_1|}{|V_2|} \angle(\varphi_1 - \varphi_2)} \quad (7)$$

电压比例系数 β 具有如（8）所示的表达式

$$\beta = \frac{|V_1|}{|V_2|} \angle(\varphi_1 - \varphi_2) \quad (8)$$

此时电流比例系数 α 形式如（9）所示

$$\alpha = -\frac{I_1}{I_2} = -\frac{L_2 \beta + M}{L_1 + M \beta} \quad (9)$$

则（2）可以写为（10）所示形式

$$V_2 = I_2 [R_2 + j\omega(L_2 - \alpha M)] \quad (10)$$

主电路等效电感 L_{eq} 可以用（11）式计算

$$L_{\text{eq}} = L_2 - \text{Re}(\alpha) \cdot M \quad (11)$$

L_{eq} 关于相位差 φ 的函数图像如图 3 所示，可以看出 L_{eq} 存在一个单调递减区和一个单调递增区间，且关于 θ_2 对称。

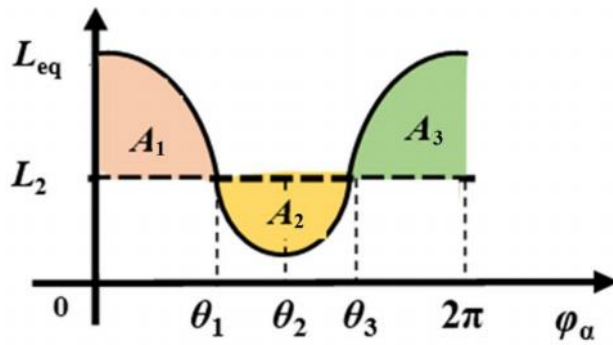


图 3 φ_α 关于 L_{eq} 的函数图像

3 仿真验证

如图 4 所示，在 10GHz 下可调电感器的自感和互感值分别为 1.03nH 和 0.48nH。

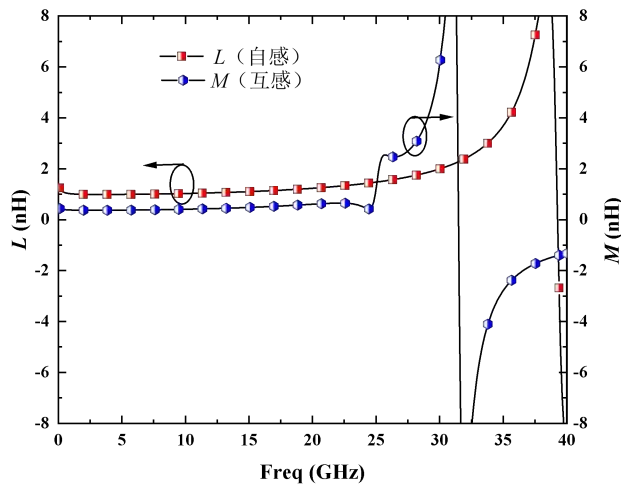
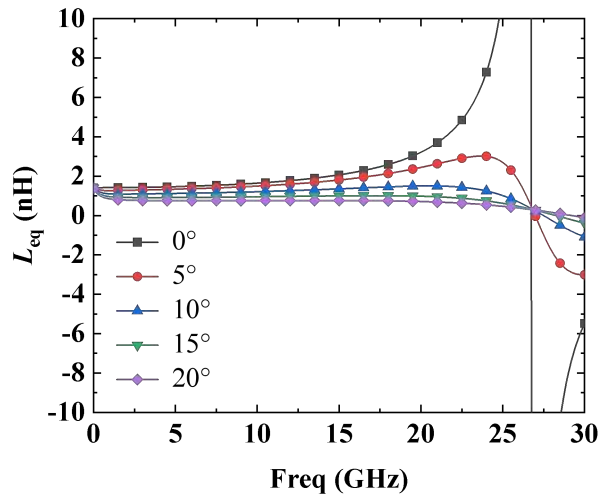


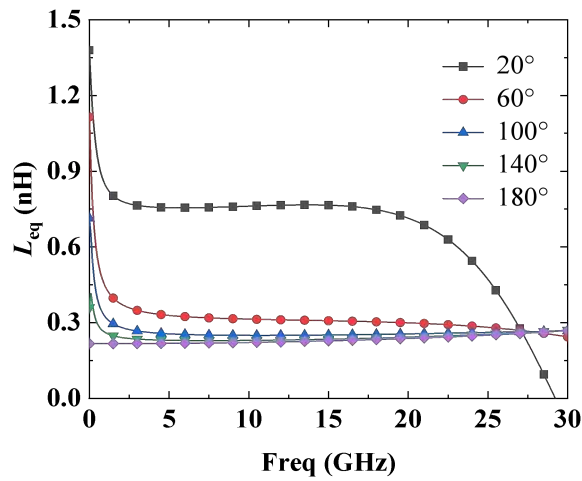
图 4 可调电感的自感和互感

当 V_1 和 V_2 等幅值并存在相位差的时候，如图 5 (a) 所示，在 $0\sim 20^\circ$ 范围内，随着相位差从 0° 开始以 5° 的步长增加， L_{eq} 随着相位差的增加逐渐减小， L_{eq} 的变化趋势与图 3 中 A_1 区域的变化趋势一致。当相位差为 0° ，20 GHz 的时的 L_{eq} 为 3.1 nH。如图 5 (b) 所示，相位差在 $20\sim 180^\circ$ 范围内，随着相位差从 20° 开始以 40° 的步长增加， L_{eq} 随着相位差的增加逐渐减小， L_{eq} 在 180° 的时候下降到 0.2 nH。如图 5 (c) 所示，相位差在 $180\sim 340^\circ$ 范围内， L_{eq} 随着相位差的增加从 0.2 nH 开始逐渐增加，当相位差增加到 340° 的时候 L_{eq} 为 1 nH。 L_{eq} 的变化趋势与图 3 中 A_2 区域的分析一致。如图 5 (d) 所示，相位差在 $340\sim 360^\circ$ 范围内， L_{eq} 随着相位差持续增加从 0.8 nH 开始逐渐增加，当相位差为 360° ，20 GHz 时的 L_{eq} 为 3.2 nH。 L_{eq} 的变化趋势与图 3 中 A_3 区域的分析一致。从仿真结果可得电感值调谐范围为 $0.2\sim 3.2$ nH，调谐范围为 93.75% 可根据公式 (12) 计算得出。如图 6 所示为不同相位差对应品质因子 (Q) 的仿真结果。

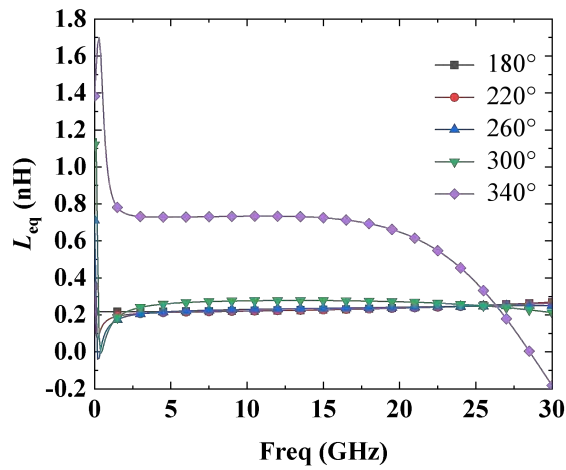
$$\text{调谐范围} = \frac{\Delta L}{L_{\max}} \quad (12)$$



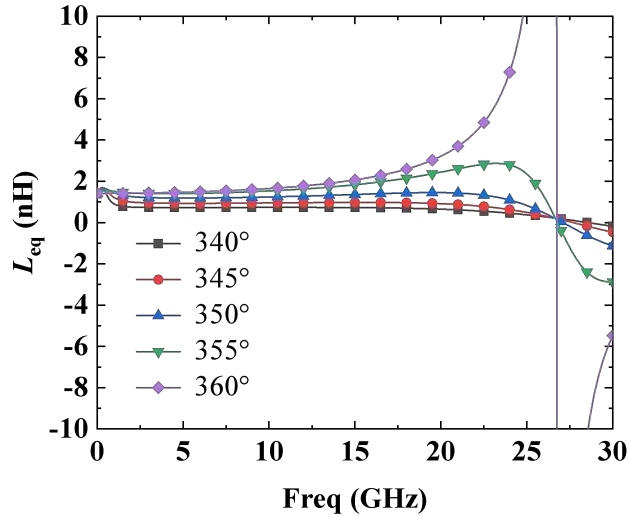
(a) $0\sim 20^\circ$



(b) $20\sim 180^\circ$



(c) $180\sim 340^\circ$



(d) 340~360°

图 5 不同相位差时的 L_{eq}

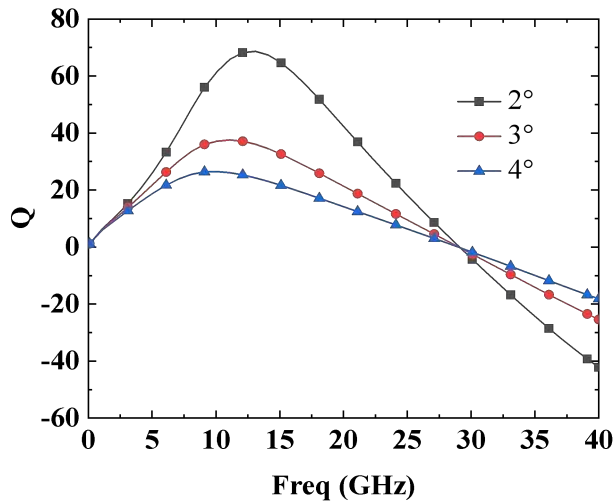


图 6 不同相位差时的品质因子 (Q)

4 结论

由上述分析可知，在相位差为 $0\sim 360^\circ$ 的范围内，仿真结果与理论分析中 L_{eq} 均存在单调递增与递减区域，各自单调区间内分别存在 α 大于 0 与小于 0 的情况，因此存在 L_{eq} 大于 L_2 与小于 L_2 的情况。在整个相位差的取值范围内，不同相位差对应的 L_{eq} 在 $0\sim 20$ GHz 范围内保持近似不变，即表现与频率近似无关的特性。在 $0\sim 20$ GHz 范围内采用相位调制的可调谐电感可以被用于设计匹配网络、滤波器等射频器件。

参考文献

- [1] A. Basir and H. Yoo, "Efficient Wireless Power Transfer System with a Miniaturized Quad-Band Implantable Antenna for Deep-Body Multitasking Implants," IEEE Trans. Microw. Theory Techn., vol. 68, no. 5, pp. 1943-1953, May. 2020.
- [2] J. S. Liu, Y. Wang, R. Guo, Q. Y. Wang, J. F. Zheng, K. Kurpad, W. Kainz and J. Chen, "A Cascaded Heterogeneous Equivalent Network for Evaluating RF-Induced Hazards on Active

- Implantable Medical Devices," *IEEE Trans. Electromagn. Compat.*, vol. 64, no. 2, pp.286-294, Apr. 2022.
- [3] S. Reich, M. Sporer, M. Haas, J. Becker, M. Schuttler and M. Ortmanns, "A High-Voltage Compliance, 32-Channel Digitally Interfaced Neuromodulation System on Chip," *IEEE J. Solid-State Circuits*, vol. 56, no. 8, pp. 2476-2487, Aug. 2021.
- [4] C. L. Ler, A. K. bin A'ain and A. V. Kordesch, "Reply to Comments on "Compact, High-Q, and Low-Current Dissipation CMOS Differential Active Inductor"," *IEEE Microw. Wireless Compon. Lett.*, vol. 18, no. 10, pp. 683-685, Oct. 2008.
- [5] T. B. Herbert, J. S. Hyland, S. Abdullah, J. Wight and R. E. Amaya, "An Active Bandpass Filter for LTE/WLAN Applications Using Robust Active Inductors in Gallium Nitride," *IEEE Trans. Circuits Syst. II Exp. Briefs*, vol. 68, no. 7, pp. 2252-2256, Jul. 2021.
- [6] D. M. Zaiden, J. E. Grandfield, T. M. Weller and G. Mumcu, "Compact and Wideband MMIC Phase Shifters Using Tunable Active Inductor-Loaded All-Pass Networks," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 2, pp. 1047-1057, Feb. 2018.
- [7] S. Rabbani, S. Narayana and Y. K. Singh, "A Novel Concurrent Dual Band Matching Network for Complex to Real Impedance Matching for RF Applications," *IEEE Trans. Circuits Syst. II Exp. Briefs*, vol. 70, no. 1, pp. 66-70, Jan. 2023.
- [8] L. J. Xu, J. P. Xu, Z. J. Chu, S. Liu and X. W. Zhu, "Circularly Polarized Implantable Antenna with Improved Impedance Matching," *IEEE Antennas Wireless Propag. Lett.*, vol. 19, no. 5, pp. 876-880, May 2020.
- [9] N. Wainstein and S. Kvatinsky, "TIME—Tunable Inductors Using MEMristors," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 65, no. 5, pp. 1505-1515, May. 2018.
- [10] K. Y. Son, T. Kim and K. Kwon, "A Dual-Band CMOS Tunable Duplexer Employing a Switchable Autotransformer for Highly Integrated RF Front Ends," *IEEE Microw. Wireless Compon. Lett.*, vol. 29, no. 7, pp. 495-497, Jul. 2019.
- [11] H. H. Chen, X. J. Wang, Y. Gao, X. L. Shi, Z. G. Wang, N. Sun, M. Zaeimbashi, X. F. Liang, Y. F. He, C. Z. Dong, Y. Y. Wei, J. G. Jones, M. R. Page, B. M. Howe, G. J. Brown and N. X. Sun, "Integrated Tunable Magnetolectric RF Inductors," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 3, pp. 951-963, Mar. 2020.
- [12] M. Vroubel, Yan Zhuang, B. Rejaei, J. N. Burghartz, "Integrated tunable magnetic RF inductor," *IEEE Electron Device Lett.*, vol. 25, no. 12, pp. 787-789, Dec. 2004.
- [13] S. S. Du, Q. H. Yang, X. N. Fan, M. Wang and H. W. Zhang, "A Compact and Low-Loss Tunable Bandpass Filter Using YIG/GGG Film Structures," *IEEE Microw. Wireless Compon. Lett.*, vol. 27, no. 5, pp. 431-433, May. 2017.

集成铁电MEMS器件与系统

耿文平

中北大学

摘要：报告人多年来致力于集成铁电 MEMS 器件与系统领域的基础科学问题和关键技术研究，在铁电畴制造方面：将铁电单晶引入微/纳机械电子系统（M/NEMS）技术领域，突破了硅基铁电单晶薄膜晶圆异质集成和铁电畴工程微纳制造共性关键技术；在铁电畴效应方面：揭示了铁电单晶薄膜电畴反转形成畴壁的导电机理、导电畴壁温度敏感机制及电场调控导电畴壁开关效应创新理论，验证了铁电畴抗辐照能力；在畴器件方面：设计实现了基于铁电单晶薄膜的力-电、热-电转换微纳结构，研发了抗辐照的铁电 MEMS 温度/振动/压力传感器件，为空间技术应用和极端环境力学参量高精度原位测量提供理论基础和技术支撑。

用于微弱肺音检测的微纳压电谐振式麦克风研究

王坤锋

中国科学院空天信息创新研究院

摘要：呼吸系统疾病是全球第三大死因，可穿戴肺音检测设备成为无创、经济且高效的临床辅助诊断工具。然而，现有用于肺音采集的麦克风在灵敏度和信噪比等关键性能指标上仍存在不足，难以满足复杂环境下对微弱肺音信号的准确提取与识别。为此，本研究结合微纳结构的多物理场仿真建模与优化设计，提出并实验验证了一种具备超高未放大灵敏度的微纳压电谐振式麦克风阵列。所设计并加工悬臂梁压电麦克风阵列，谐振频率覆盖 230 至 630 Hz，有效匹配典型微弱肺音信号的频率分布。器件实测结果显示，阵列未放大灵敏度最高可达 265 mV/Pa，信噪比达 98 dBA，性能显著优于传统平带麦克风。该阵列能够在低信号水平下准确识别喘息声特征，结合深度学习算法，实现了高准确率分类，展现出其在可穿戴肺音监测与呼吸疾病智能诊断中的广阔应用前景。

压电MEMS微型散热风扇系统仿真与优化分析

易志然

上海交通大学

摘要：针对微电子设备日益严峻的散热挑战，本研究聚焦于高性能压电 MEMS 微型散热风扇系统的开发。系统核心采用压电驱动器结合 MEMS 工艺制备的微扇叶结构，通过逆压电效应产生高频振动驱动气流。研究建立了涵盖压电-结构-流体多物理场的全耦合仿真模型，深入分析扇叶振动模态、流场分布及散热性能等关键特性。仿真揭示了风扇性能高度依赖于扇叶结构固有频率、振幅、倾斜角度及驱动电压频率/幅值。基于此，本研究系统开展了结构拓扑优化与驱动参数协同优化，优化扇叶形状与支撑结构以提升振幅与刚度匹配度；精细调节驱动频率至系统谐振点附近；优化扇叶倾角以增强气流定向性并抑制涡流效应。优化后风扇在低驱动电压下实现了显著提升的风量与散热效率。本工作验证了多物理场仿真结合参数优化在压电 MEMS 风扇设计中的有效性，为满足下一代微电子设备的高效、紧凑热管理需求提供了可行方案与理论依据。

MEMS综合性设计技术发展思考

周再发

东南大学

摘要：MEMS 设计过程涉及集成电路、机械、声学、光电、化学、生物等多个学科，设计过程难度大。研发 MEMS 智能化综合设计技术并开发相应软件工具，可以提升 MEMS 的设计水平和效率，有利于更多设计公司加入到 MEMS 研发领域中来，促进 MEMS 产业的发展。MEMS 的智能化综合性设计技术包括系统级、器件级和工艺级三个层次，报告从以上三个层次讨论分析了各自需要解决的一些关键问题，并介绍了针对这些问题取得的最新研究进展。最后，展望了 MEMS 智能化综合性设计的一些研究思路。

MEMS多物理场建模仿真与超灵敏传感器

贾浩

中国科学院上海微系统与信息技术研究所

摘要：MEMS 传感器在电子信息、物联网、环境监测等众多领域应用广泛。MEMS 传感器工作过程中往往涉及多物理场的耦合作用，其建模仿真在器件设计中具有不可替代的重要意义。通过对传感器结构的建模和参数优化，可实现其对微弱信号的超灵敏感知能力。本报告聚焦 MEMS 传感器的力、热、电多物理场耦合建模及其在超灵敏质量和温度传感方面的应用。

基于模型的微系统研发数字化转型实践

孙岩

中国电子科技集团公司第二十九研究所

摘要：随着微系统技术朝着超宽带、高集成度、多功能性及低成本的方向迅速发展，传统设计方法正面临迭代周期冗长、跨领域与层级协调效率低下以及制造验证成本高昂等瓶颈挑战。针对这些问题，本报告以构建系统-封装-芯片三级协同的跨尺度模型传递机制为突破口，深入探讨了在微系统研发数字化转型过程中，如何通过数字模型实现专家经验、数据资源、工具软件和企业信息系统的深度融合。报告分享了一系列方法与实践成果，展示了数字化转型对提升微系统研发效能的关键作用及其潜在价值。

基于多层精密微组装的石英振梁加速度计模态抑制研究

杨贵玉, 杨挺, 陈家林, 焦海龙
(北京遥测技术研究所, 北京, 100076)

摘要: 根据石英振梁加速度计的高精度测量特点, 建立了双端固支石英谐振梁的动态力学模型和多层微组装结构的压膜阻尼模型, 对石英振梁加速度计建立有限元模型并进行模态仿真分析, 表明结构存在较低频率的模态振型。基于多层微组装结构模型对不同装配间隙条件下动态响应进行了仿真分析, 并利用振动台进行了石英振梁加速度计频带特性测试。实验结果表明, 通过优化设计和精确的装配工艺控制, 可以实现石英振梁加速度计低阶模态耦合的抑制, 有效解决高精度加速度测量的宽带宽的难题。

关键词: 石英振梁加速度计; 微组装; 压膜阻尼; 模态抑制

1 引言

高精度石英振梁加速度计是采用全石英晶体材料, 基于石英晶体的压电效应和谐振梁力频特性的高精度固态型加速度传感器。由于采用谐振式检测原理, 输出信号为准数字的频率信号, 避免了信号转换引入的分度误差, 可直接与数字系统进行对接, 减少了因 A/D 转换所需的资源消耗和量化误差, 因此功耗低、精度高。同时敏感芯片结构采用石英基 MEMS 工艺制备, 力敏谐振器与质量块为一体式加工, 避免了因异质材料装配导致的热性能不匹配而产生的热应力, 保证了优良的长期稳定性能。

石英振梁加速度计具有精度高、体积小、频带宽、功耗低、抗干扰能力强、长期稳定性佳、适宜 MEMS 批量化生产等优势, 是当前惯性测量领域的研究热点, 已应用于航天、航空、航海以及相关的惯性测量、惯性制导等领域, 也一直被业界公认为是新一代中高精度导航系统首选的加速度测量仪表, 具有进入战略级惯性导航的潜在能力。目前在产品工程化应用方面处于领先地位的是美国 Honeywell 公司和法国 ONERA 公司。Honeywell 公司的 RBA500 系列加速度计被广泛用于美国空军各类战术导弹的中段制导^[1], 在多次局部战争中表现不俗, 年配套量超 1 万只; 法国 ONERA 研制的集成式全石英振梁加速度计 VIA 系列产品^[2-3], 相关专利技术已转让给 Thales、Sagem 和 iXBlue 等公司, 产品已在海军潜艇、飞机、战术导弹等导航系统得到大量应用^[4-5], 2019 年 iXBlue 公司推出首款批量生产的高性能导航级振梁加速度计 iXal A5, 其测量范围为 $\pm 80g$, 年偏值重复性优于 1mg。

对于高精度石英振梁加速度计, 在实际应用过程中除了关注零偏稳定性、标度因数稳定性、分辨率等静态性能指标, 还尤其要重视其动态性能指标, 如带宽

特性、振动整流特性。在航天应用领域，产品实际的工程力学环境往往存在较大量级的随机振动，通常随机振动的频率范围为 5Hz~2000Hz，振动的量级一般根据型号的不同而存在较大的差异。随机振动环境是包含低频、高频等复杂的频率成分，加速度计在进行结构设计时要尤其关注结构的低阶模态频率，避免在随机振动频率范围内引起结构局部的谐振放大，造成模态的交叉耦合，因此需要在加速度计高灵敏度设计目标与避免模态耦合间进行综合衡量。如果无法实现避免模态耦合的设计，则需要进行模态抑制的设计，通过结构设计、封装等措施实现对耦合模态的抑制，将模态耦合的影响尽可能降低。

本文针对项目团队研制的高精度石英振梁加速度计，首先建立双端固支石英谐振梁的动态力学模型，并基于多层微组装设计建立三层结构的压膜阻尼模型；其次对石英振梁加速度计核心敏感层建立有限元模型，基于 ANSYS 有限元软件进行模态仿真分析，对于影响加速度计动态性能的低阶耦合模态提出抑制方法；基于 COMSOL 多物理耦合分析多层微组装不同的控制层间间隙时，核心敏感层的动态响应特性，得出了微组装间隙与模态抑制的关系。最后利用振动台对实际研制的石英振梁加速度计进行了振动试验，验证了模态抑制方法的有效性，对后续多层精密微组装控制的研究提出了展望。

2 石英振梁加速度计动态模型

2.1 谐振梁动态模型

石英振梁加速度计核心敏感元件结构示意图如图 1，由质量摆、谐振梁和挠性摆组成^[6]。谐振梁为双端固支单梁结构形式，在 Z 向加速度输入时，质量摆会围绕挠性支撑摆的质心发生转动，从而对谐振梁产生轴向的拉应力或压应力作用，引起谐振梁等效刚度的变化，进而引起固有谐振频率的变化，实现加速度测量。

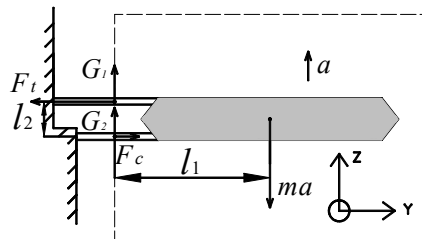


图 1 石英振梁加速度计结构示意图

在小位移模式下，整体系统可等效为质量-弹簧-阻尼系统。受到外界加速度为 $a(t)$ 时，谐振梁轴向力 F 与加速度的传递函数模型为^[7-8]：

$$\frac{F(s)}{a(s)} = \frac{mk_1 l_2 / l_1}{ms^2 + bs + k} = \frac{k_1 / \lambda}{s^2 + 2\xi\omega_n s + \omega_n^2} \quad (1)$$

式中： m 为质量摆质量， b 为阻尼系数， k 为系统刚度， k_1 为谐振梁轴向拉伸刚度， k_2 为挠性支撑的弯曲刚度， l_1 为质量摆质心的惯性力臂， l_2 为谐振梁轴向

力的力臂，定义 $\lambda = l_1/l_2$ 为放大系数。谐振梁自振角频率 $\omega_n = \sqrt{k/m}$ ，阻尼比 $\xi = b/(2\sqrt{km})$ ，系统刚度 $k = (2k_2 + k_1l_2^2)$ 。

根据双端固支谐振梁弯曲振动模式的自由振动方程，可得谐振梁固有谐振频率为^[8]：

$$f_0 = 1.0284 \frac{w}{l^2} \sqrt{\frac{E}{\rho}} \quad (2)$$

受轴向力作用状态下，其谐振频率为：

$$f = f_0 \sqrt{1 + 0.2949 \frac{Fl^2}{Ehw^3}} \quad (3)$$

其中， l 为谐振梁的长度， w 为谐振梁的宽度， h 为谐振梁的厚度， ρ 为石英材料密度， E 为弹性模量， F 为振梁受到的轴向力。

振梁加速度计采用差分工作模式，在理想情况下，处于差分的两个振梁加工参数完全一致，略去高阶项，可以得到频率变化量与输入加速度关系模型为^[9]：

$$\Delta f(s) = 0.3033 \frac{1}{hw^2} \sqrt{\frac{1}{\rho E}} \frac{k_1/\lambda}{s^2 + 2\xi\omega_n s + \omega_n^2} a(s) \quad (4)$$

由式(4)可以看出影响频率变化量的影响因素。在敏感元件结构设计方案确定的情况下，可以通过调节系统的阻尼系数来改变不同振动频率下加速度计输出频率的变化量。

2.2 压膜阻尼模型

项目团队研制的石英振梁加速度计为基于精密微组装的三层结构，中间层为核心敏感层，上下两层为防护止挡层，如图2所示^[10]。由于敏感结构的尺寸远大于层间间隙 δ ，质量摆在外界加速度作用下在垂直方向上下运动，对间隙内的气体形成压缩作用，迫使气体产生流动形成附加压强，相对运动示意图如图3，间隙中的气体分子对运动的质量摆起到阻尼器的作用，其产生的阻尼称为压膜阻尼。

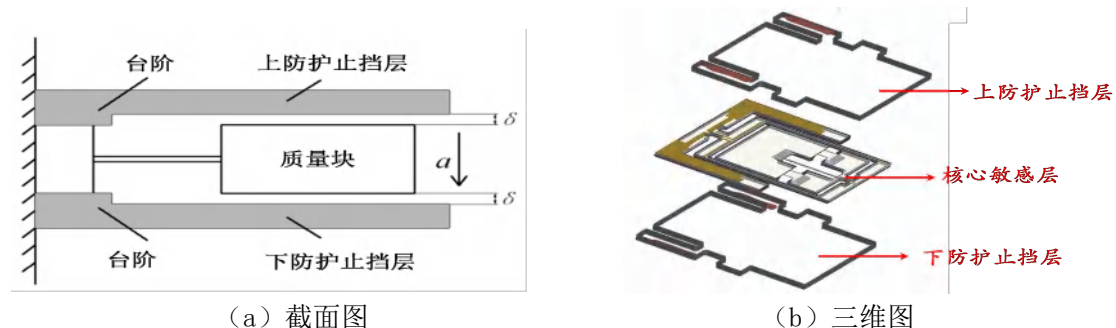


图2 精密微组装三层结构示意图

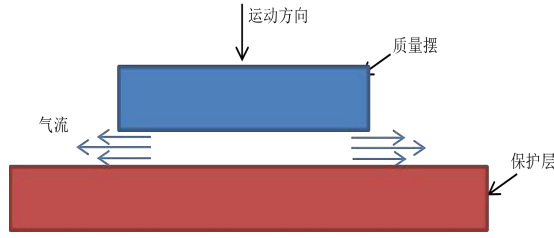


图 3 相对运动示意图

压膜阻尼一般表现为两种效应^[11-12]：压膜阻尼效应和压膜刚度效应。当振动频率较高，压膜系数较大时，压膜刚度效应显著，此时应考虑压膜刚度效应的影响。振动频率较低，压膜系数较小时，气体呈被挤压状态，主要表现为压膜阻尼效应^[13]。

将质量摆视为矩形平板结构，敏感层与止挡层间隙为 δ ，质量摆宽度为 L_x ，长度为 L_y ，则平板压膜阻尼力为^[6]：

$$P = \frac{\mu L_x^3 L_y}{\delta^3} \frac{d\delta}{dt} \beta \left(\frac{L_x}{L_y} \right) \quad (5)$$

式中： μ 为气体粘性系数。修正系数为：

$$\beta \left(\frac{L_x}{L_y} \right) = \left\{ 1 - \frac{192}{\pi^5} \left(\frac{L_y}{L_x} \right) \sum_{n=1,3,5,\dots}^{\infty} \left(\frac{n\pi L_x}{2L_y} \right) \right\} \quad (6)$$

结合具体的质量摆的结构设计参数，可以得到压膜阻尼系数为：

$$c_{squeeze} = \frac{\mu L_y L_x^3}{2\delta^3} \quad (7)$$

结合公式（4）和（7），在工程实践中可通过设计、工艺控制手段调整三层结构间的阻尼间隙，从而调节质量摆片的压膜阻尼力，实现振梁加速度计带宽特性的调节，抑制干扰模态的耦合影响，防止出现结构谐振放大的现象。

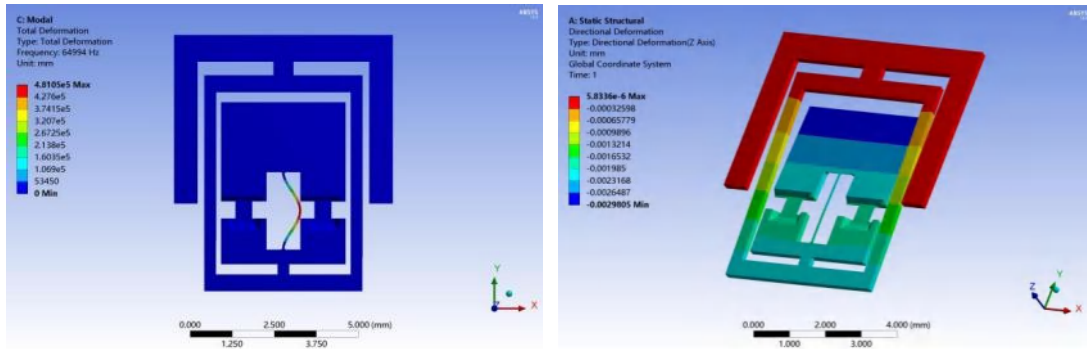
3 仿真实验

为了验证理论分析和结构设计的正确性和合理性，利用 ANSYS 有限元软件对石英振梁加速度计的核心敏感层进行建模，并进行有限元仿真分析。

建立有限元模型后，首先确定谐振梁结构变形和工作模态是否符合设计要求，即谐振梁的工作模态是否为一阶面内 X 向振动模态，仿真分析结果如图 4 所示。

由图 4（a）可知，石英谐振梁处于一阶面内 X 向振动模态，固有谐振频率为 64994Hz，从模态分布云图可见，振梁加速度计的内层隔振框、内外连接臂起到了很好的隔振效果，防止谐振梁的振动能量传递到外层安装框架，造成安装固支损耗，保证了良好的品质因数。图 4（b）为加速度计在受到 10g 加速度稳态载荷作用时的位移变形云图，可知位移变形最大值为 2.98um，位于质量块顶部。

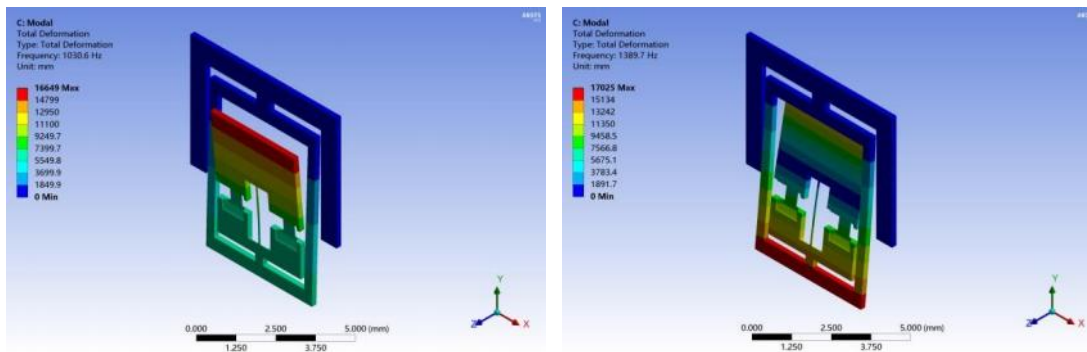
在设计多层精密微组装时，层间间隙应预留一定空间以满足在量程范围内质量块顶部的位移需求，避免质量块与止挡层发生接触而导致限位，无法实现全量程的测量。



(a) 工作模式振型图 (b) 10g 载荷作用时位移云图

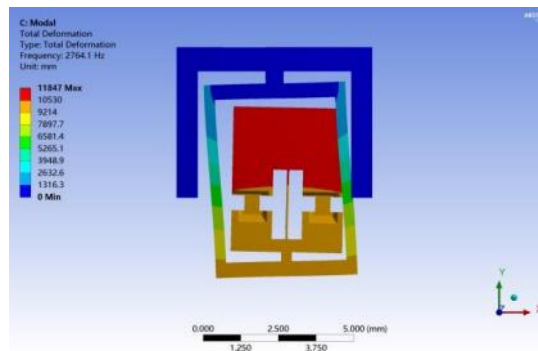
图 4 加速度计工作模式和载荷位移云图

在航天工程领域实际应用时经常存在 2000Hz 以内的随机振动环境，为防止在力学振动环境下加速度计结构发生谐振放大现象，尤其需要关注核心敏感层的低阶模态频率。图 5 为模态仿真分析获得的石英振梁加速度计前三阶模态。图 5 (a) 为一阶模态振型，固有频率为 1030.6Hz，其振型为质量摆片绕挠性轴在 Z 项摆动，也是加速度计的敏感模态。图 5 (b) 为二阶模态振型，固有频率为 1389.7Hz，其振型为内层隔振框 Z 向弯曲模态。图 5 (c) 为三阶模态振型，固有频率为 2764.1Hz，其振型为内层隔振框 X 向弯曲模态。



(a) 第一阶模态

(b) 第二阶模态



(c) 第三阶模态

图 5 加速度计前三阶模态振型图

由以上有限元模态仿真分析结果可知，石英振梁加速度计存在固有的一阶、二阶模态频率低于 2000Hz，且这两个模态振型都与质量摆片有关。石英振梁加速度计的动态响应可以近似地看成二阶响应系统。当激励频率接近或等于固有频率时，将引起谐振现象，放大质量摆片的位移变形。通过调节二阶响应系统中的阻尼系数，使其达到最佳阻尼，将使幅频特性曲线平坦段最大化，最大程度拓宽加速度计的带宽特性。振梁加速度计可以等效成弹簧-阻尼-质量块系统，由于质量摆片将加速度形成的惯性力通过力矩的方式作用于系统，由牛顿第二定律可得系统扭转角 φ 对时间的微分方程为。

$$J \frac{d^2\varphi(t)}{dt^2} + b \frac{d\varphi(t)}{dt} + k\varphi(t) = mg(t)L \quad (8)$$

式中 J 、 b 、 k 分别为系统转动惯量、黏性阻尼系数和扭转弹性刚度； m 为敏感质量； L 为质量摆片产生的惯性力矩的等效力臂。其中 J 、 k 、 m 、 L 由敏感单元的相关尺寸和材料属性可以完全确定， b 需要根据流体润滑相关理论，求解雷诺方程获得，一般来说矩形平板之间的法向相互运动挤压气体产生的阻尼可以求得，但质量摆片实际处于摆动状态，且与质量摆片相连的其他附属结构均会产生位移，雷诺方程很难获得精确解析解，因此需要采用有限元方法获得数值解更能准确反映运动状态。

利用多物理场耦合分析软件 COMSOL，对项目团队研制的基于多层精密微组装的石英振梁加速度计构建基于核心敏感层和防护止挡层的压膜阻尼有限元仿真模型。防护止挡层的主要作用是提供气体间隙，因此防护止挡层通过与间隙值一致的一层气体薄膜进行替代。

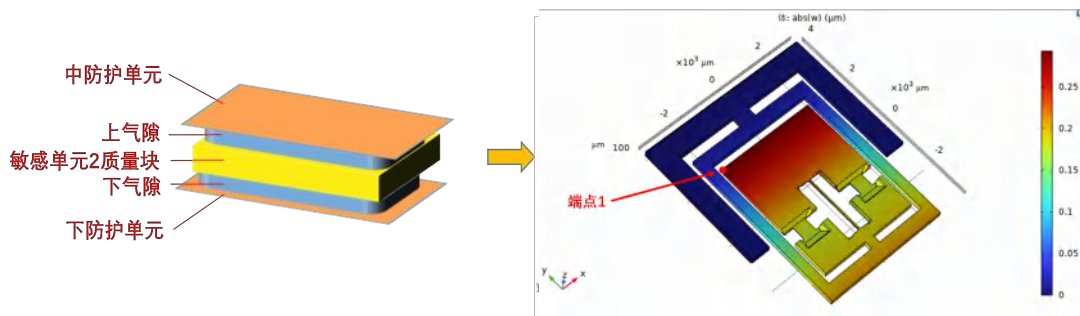
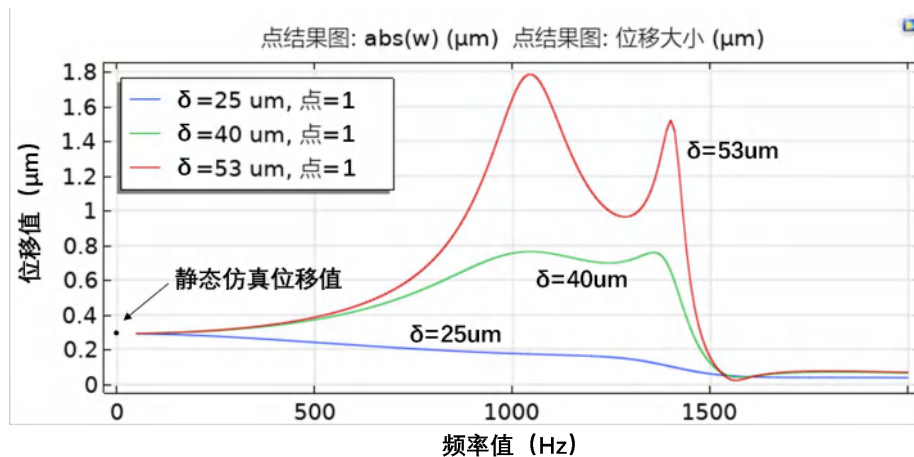
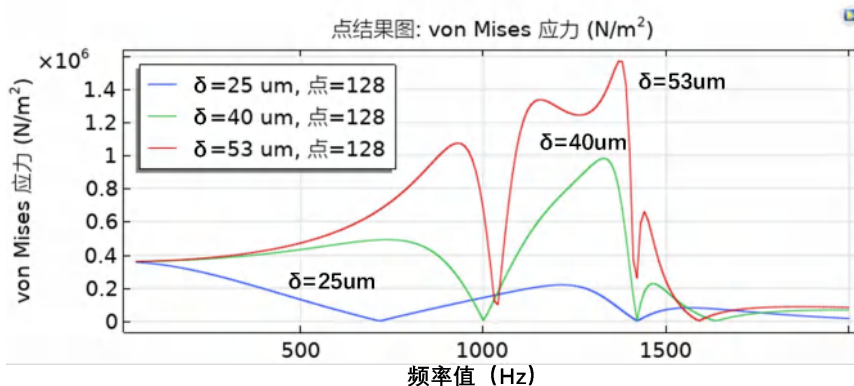


图 6 多物理场耦合仿真模型图

建立的多物理场耦合分析模型如图 6 所示，以质量摆顶端端点 1 为研究对象，输入载荷为 1g 正弦加速度，频率范围为 50Hz~2000Hz，频率步长为 10Hz，气压值为固定 1atm。仿真分析时，间隙参数值分别设定为 25 μm 、40 μm 和 53 μm 。



(a)位移幅频响应曲线



(b)应力幅频响应曲线

图 7 加速度计质量摆位移和应力幅频响应曲线

图 7 (a) 为端点 1 的位移幅频响应曲线。在低频段时，不同间隙下的位移一致。随频率增加，位移幅值随频率的变化趋势与层间间隙有直接关系，间隙越小，幅频曲线逐渐呈现出从欠阻尼到过阻尼的变化。随着频率的增加，气体受到质量摆片的快速挤压，来不及进入和排出间隙，呈现阻碍质量摆片运动的趋势，且间隙越小阻碍作用越明显。当间隙值为 $40\mu\text{m}$ 和 $53\mu\text{m}$ 时，在 1100Hz 和 1400Hz 附近均出现了谐振峰，位移幅值相比低频载荷下明显放大，表明阻尼效应不足以抑制低阶模态的耦合；出现谐振的频率点与前述模态分析结果基本一致。当间隙值为 $25\mu\text{m}$ 时，上述谐振峰消失，阻尼效果显著增加，系统的谐振放大效应被阻尼抑制，低阶模态的得到有效抑制，带外的高频振动得到有效衰减。

图 7 (b) 为模型最大应力的幅频响应曲线，与位移响应曲线规律相同。当间隙值为 $40\mu\text{m}$ 和 $53\mu\text{m}$ 时，出现明显应力放大现象；而间隙值为 $25\mu\text{m}$ 时，应力在高频段呈衰减状态，对低阶模态起到良好的抑制作用。

4 实验验证

为验证基于多层精密微组装结构的石英振梁加速度计对低阶频率模态抑制的实际效果，项目团队装配了不同控制间隙的加速度计样机，利用振动台进行实

验证测试。

多层微组装结构的控制间隙分别控制为 4 个不同值 ($h_1 > h_2 > h_3 > h_4$)，共 4 台样品。搭建的试验平台如图 8 所示，振动台 4808 产生标准的振动加速度，被测样品安装在振动台上，敏感轴向与振动方向一致。被测加速度计由 Keysight3649A 直流稳压电源提供稳定的工作电源，输出频率信号采用 Keysight53230A 通用频率计数器进行高速采样。实验测试频率范围为 30Hz~2000Hz，频率间隔为 50Hz，振动加速度为 9.8m/s^2 ，采样频率为 4kHz。



图 8 加速度计实验测试平台

对 4 台样品测试数据进行归一化处理，计算不同频率点时样品的灵敏度，以 30Hz 时灵敏度为参考点计算灵敏度的相对变化，实际测试得到加速度计灵敏度的幅频特性结果如图 9 所示。

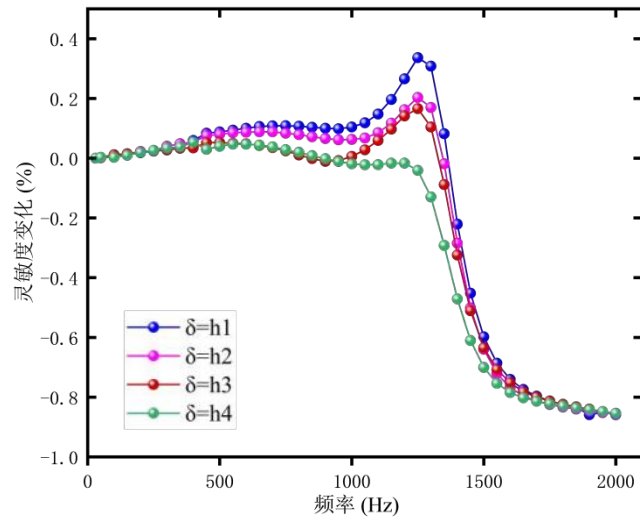


图 9 加速度计灵敏度变化幅频特性测试结果

从测试结果可知，加速度计灵敏度在 500Hz 后开始出现放大，在 1300Hz 时低阶模态处于谐振状态，导致灵敏度放大很多。随控制间隙的减小，加速度计在高频段的灵敏度变化减小，当控制间隙为 h_4 时，低阶频率模态得到有效抑制，加速度计灵敏度变化很小，使得加速度计具有良好的宽带特性，带宽达到 1000Hz 以上，能够满足一些高精度振动测量领域的需求。

5 结论

本文通过对基于多层精密微组装的石英振梁加速度计建立动力学模型,利用有限元软件进行模态分析,指出了低阶频率耦合模态的影响并提出抑制方法,研究不同层间控制间隙对加速度计动态特性影响,当层间间隙达到合理值时可以有效抑制低阶耦合模态,并通过实验测试验证了仿真分析的正确性,加速度计带宽可达到 1000Hz 以上并具有良好的平坦度。

受到工艺加工误差影响,层间间隙的精确控制是一项工艺难题,下一步通过优化石英湿法刻蚀工艺参数、设计高精度定位装配工装等措施保证精密装配的可控性,满足加速度计的动态特性要求。

参考文献:

- [1] Fall R H and Dipipi M L. Autonomous Ballistic Missile Inertial Guidance-A New Paradigm for the 21st Century [C]// AIAA Guidance, Navigation and Control Conference and Exhibit, 2008: 7299.
- [2] Traon O L, Janiaud D, Lecorre B et al. Monolithic Differential Vibrating Beam Accelerometer within an Isolating System between the Two Resonators [C]//Proceedings of IEEE Sensors, Irvine, CA, USA, 2005: 648-651.
- [3] Loret T, Hardy G. Navigation Grade Accelerometer with Quartz Vibrating Beam[C]//Inertial Sensors and Systems, 2014: P01.
- [4] 陈冲冲, 曾凡, 王毅飞等. 谐振式加速度计自激振荡电路性能评价方法的研究[J]. 遥测遥控, 2021,42(4): 112-118.
- [5] 李晓阳, 王伟魁, 汪守利, 彭泳卿等. MEMS 惯性传感器研究现状与发展趋势[J]. 遥测遥控, 2019,40(6): 1-13.
- [6] 杨挺, 杨贵玉, 李庆丰等. 石英振梁加速度计静态输入输出特性[J]. 中国惯性技术学报, 2014, 22(3): 386-390.
- [7] 胥梓夏. 关于改善谐振梁加速度计振动特性的方法研究[J]. 传感技术学报, 2019, 32(9): 1313-1317.
- [8] 彭泳卿, 陈巍, 陈青松. 高精度电容式微振动传感器的设计与实现[J]. 遥测遥控, 2012, 33(2): 67-72.
- [9] 毛志成, 张晗, 杨泽宇等. 石英振梁加速度计温度自补偿方法[J]. 仪表技术与传感器, 2024(7):19-24.
- [10] 杨贵玉, 刘平, 杨挺, 金小锋. 一体式石英振梁加速度计温度特性研究[C]. 第十五届全国敏感元件与传感器学术会议论文集. 郑州: 2018,401-405.
- [11] 杨挺, 刘平, 杨贵玉等. 一体式石英振梁加速度计工程化研究进展[J]. 遥测遥控, 2019,40(2): 39-44.
- [12] 王伟, 刘军洁. Z 轴微机械陀螺仪的空气阻尼分析[J]. 西安工业大学学报, 2010,30(3): 219-223.
- [13] 彭泳卿, 陈青松, 邹江波. 电容式加速度传感器的压膜阻尼分析与设计[J]. 仪器仪表学报, 2010,31(8):92-99.

作者简介

杨贵玉, 男, 1977 年, 高级工程师, 毕业于大连理工大学, 机械设计及理论专业硕士研究生, 主要从事 MEMS 传感器技术研究。

杨挺, 男, 1985 年, 高级工程师, 毕业于中科院半导体所, 材料物理与化学专业硕士研究生, 目前从事 MEMS 传感器研究及工艺开发。

陈家林, 男, 1994 年, 工程师, 毕业于中国地质大学, 机械工程专业硕士研究生, 目前从事传感器及 MEMS 技术研究。

焦海龙, 男, 1984 年, 研究员, 毕业于中科院电子所, 微电子学与固体电子学专业博士研究生, 目前从事 MEMS 传感器及微纳器件研究。

一种基于POD降阶和状态空间方程的MEMS热式流量传

感器宏模型提取方法

张浩东, 徐昊, 王贺, 赵临风, 周再发

东南大学 MEMS 教育部重点实验室

摘要: 针对复杂 MEMS 器件多物理场仿真和接口电路设计分离以及多场耦合导致系统级建模困难的问题, 本文提出了一种基于 POD 降阶与状态空间方程的宏模型提取方法, 利用多场耦合的 MEMS 热式流量传感器为实例对所提出的方法进行了验证。该方法通过分析非线性传热系统的矩阵方程, 将边界对流引起的非线性贡献项进行线性化分离, 并结合 POD 降阶和状态空间方程, 建立了 MEMS 热式流量传感器的宏模型。利用 Verilog-A 语言对建立的宏模型进行描述封装并插入到 Cadence Virtuoso 中进行系统级仿真验证, 稳态仿真结果与有限元相比误差 $<1.3\%$, 与实验结果相比误差 $<7.9\%$, 瞬态仿真时间仅 41.255ms, 较有限元缩短了 218 倍。

关键词: MEMS; 流量传感器; POD 降阶; 状态空间; 宏模型; 系统级仿真;

1 前言

MEMS 热式流量传感器凭借其微型化、低功耗、高灵敏度以及响应速度快等优势[1], 已在医疗监测[2]、智能建筑[3]、环境监测[4]、工业控制[5]等领域得到广泛运用, 并且凭借与 CMOS 工艺兼容的独特技术优势[6], 使得大规模、低成本生产多功能集成传感器成为可能。

然而, 面对复杂应用场景的多样化需求, 传感器设计仿真面临多重技术挑战: 首先, MEMS 器件的设计验证通常是利用有限元软件进行的, 微尺度多物理场耦合效应的精确模拟需要耗费极大的计算资源, 这限制了 MEMS 器件设计优化的效率; 其次, 有限元仿真通常无法反映大规模接口电路对 MEMS 器件的影响, 同时接口电路的设计也难以反映多物理场耦合的影响, 这意味着 MEMS 的设计与接口电路的设计是分离的, 需要建立 MEMS 器件的宏模型来实现与电路的系

统级联合仿真；此外，复杂微系统的多物理场耦合给 MEMS 器件系统级分析建模带来困难，尤其是多种因素导致的非线性效应，需要形成具有一定适用性和通用性的方法来建立 MEMS 器件的宏模型。

目前建立 MEMS 传感器宏模型的常用方法有等效电路法[7]，集总参数法[8]，模型降阶法[9]和神经网络[10]等。等效电路法通过将机械、热学中的物理量类比转化为 RLC 电路模型，但是这种方法只适用于简单 MEMS 器件，对于结构复杂的器件不适用。集总参数法是利用集总参数（如总热容、等效电阻）代替分布式特性对模型进行简化，但是对复杂几何结构和局部的物理场分析存在不足；神经网络具有优异的非线性拟合能力，但是过度依赖于数据集，建立的模型泛化能力需要进行额外的优化。模型降阶法多是基于线性系统降阶方法建立 MEMS 器件的宏模型，由于系统矩阵规模得到极大简化，使得建立的降阶模型求解准确而高效，具有很强的通用性。

基于前文的分析，本文提出一种基于 POD 模型降阶和状态空间描述的宏模型提取方法，该方法以 MEMS 热式流量传感器为例，实现了电-热-流三场耦合 MEMS 器件的宏模型提取。本文首先分析了非线性传热系统的矩阵方程，发现边界对流项同时对系统矩阵中的热传导矩阵 \mathbf{K} 和载荷向量 \mathbf{F} 有贡献。然后利用有限元软件获取了非线性传热系统的温度场快照，采用 POD 法对温度场进行 POD 基底的提取，利用提取的 POD 基底对系统矩阵进行了降阶。接着将对流换热系数 h 做线性近似，将边界对流引起的非线性贡献项进行了分离，并利用状态空间方程对降阶系统进行了描述，建立了 MEMS 热式流量传感器的宏模型。最后将提取的宏模型利用 Verilog-A 语言描述封装集成至 Cadence Virtuoso 进行系统级仿真验证，稳态结果较有限元仿真误差 $<1.3\%$ ，较实验结果误差 $<7.9\%$ ，瞬态仿真时间仅 41.255ms，较有限元仿真缩短了 218 倍。

本文章节安排如下：第一节为前言，第二节介绍本文提出的方法和原理，第三节为本文方法的仿真结果，第四节给出结论。

2 原理和理论基础

2.1 MEMS 热式流量传感器的工作原理

MEMS 热式流量传感器通常配备有一个或多个加热元件和测温元件[11]。如图 1 (a)和(b)所示为一维流量传感器的示意图：红色部分指代加热元件，位于传感器中心位置，绿色部分指代测温元件，对称分布在加热元件的两侧。当流体流速为 0 且加热元件处于工作状态时，传感器表面温度场以加热电阻为中心，呈现左右完全对称分布，如图 1(a) 中黑色曲线所示，此时，上下游测温元件的温度差为 0。当流体流速不为 0 时，流体的流动会破坏原先对称分布的温度场，如图 1(b) 中红色曲线所示，使得上下游测温元件之间产生温度差。通过图 1(c) 所示的接口电路将温度差转换为相应的电流信号后，即可得到流速信息[12]，其中 R_1 和 R_2 为对称分布的测温电阻。

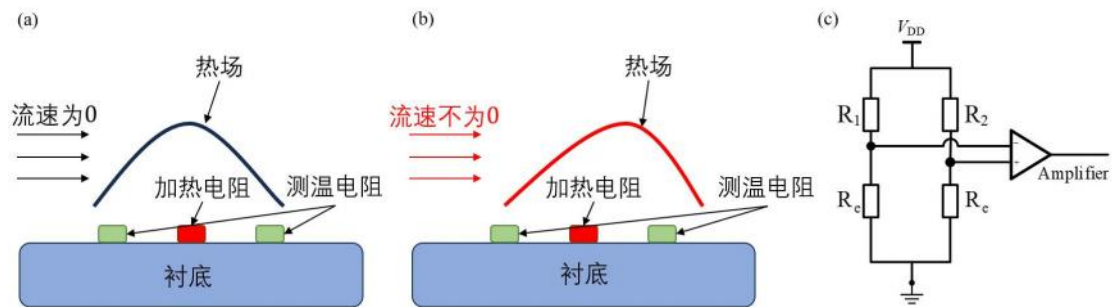


图 1 MEMS 热式流量传感器工作原理

根据测量原理来看，MEMS 热式流量传感器是一个电场、热场、流场耦合的器件。其中热场是由加热器施加一定功率产生的，流场是由流体流动引起并影响热场分布，电场是由外部控制电路引起的，用于提供电源输入并获得输出信号。在这个多物理场耦合的环境中，热场受到流场的影响，这是导致非线性的主要原因。本文以一维悬浮薄膜热式流量传感器为建模对象，说明了电场、热场、流场三场耦合器件宏模型提取的流程。图 2 展示了悬浮薄膜 MEMS 热式流量传感器的金相显微镜照片，传感器中心加热电阻与测温电阻之间有隔热槽，这样设计的目的是为了减少衬底传热对测温电阻的影响。图 3 展示了所建立的热式流量传感器有限元模型在不同流体流速下的温度分布情况，其中中心加热电阻的发热功

率恒定为 7mW。

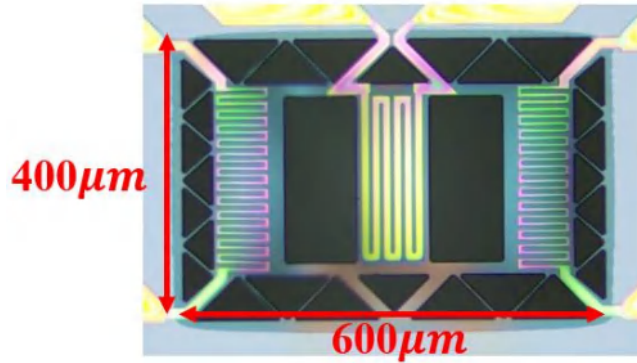


图 2 悬浮薄膜 MEMS 热式流量传感器

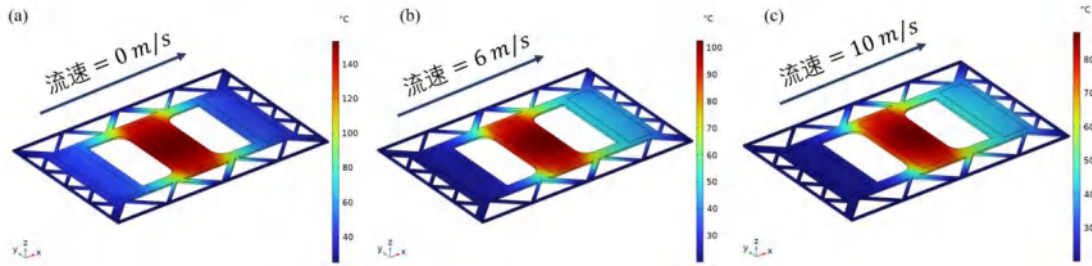


图 3 悬浮薄膜 MEMS 热式流量传感器有限元模型

2.2 含对流影响的非线性传热系统

对于线性传热系统的矩阵方程[13]为：

$$CT + KT = F$$

其中 C 为热容矩阵， K 为热传导矩阵， F 为载荷向量， T 为温度场分布， C, K, F 的规模与有限元模型的网格节点数 n 相关。而 MEMS 热式流量传感器明显是一个非线性传热系统，考虑流体引起的边界对流项同时对热传导矩阵 K 和载荷向量 F 都有贡献[14]，可以将热传导矩阵 K_{total} 写为热传导项 K_{cond} 和边界对流项 K_{conv} 的组合：

$$K_{total} = K_{cond} + K_{conv}$$

同理载荷向量 F_{total} 可以写为热源项 F_{source} 、边界热流项 F_{flux} 和边界对流项 F_{conv} 的组合：

$$F_{total} = F_{source} + F_{flux} + F_{conv}$$

在 MEMS 热式流量传感器中，加热电阻采用薄膜结构，其贡献可以看做面

热源，因此边界热流项引起的贡献可以合并到热源项中，那么就有

$$\mathbf{F}_{total} = \mathbf{F}'_{source} + \mathbf{F}_{conv}$$

那么对于 MEMS 热式流量传感器所体现的非线性传热系统的矩阵方程为

$$C\dot{\mathbf{T}} + (\mathbf{K}_{cond} + \mathbf{K}_{conv})\mathbf{T} = \mathbf{F}'_{source} + \mathbf{F}_{conv}$$

2.3 本征正交分解 (POD) 对非线性系统矩阵的降阶

本征正交分解 (POD) 是一种用于计算最优低维基函数 (即 POD 基底) 的方法。其主要作用是从高维系统数据中提取最能有效反映系统特征的正交基向量。这些基向量的构造原则是, 使得原始高维系统在该基底方向上投影的方差最大化 [15]。因此, 原始系统能够通过有限数量 POD 基底的线性组合来表征, 这显著降低了模型的维度, 可以达到模型降阶从而加快计算的目的 [16]。图 4 展示了该过程: 一个时域高维系统被映射到一个低维空间, 针对时域空间中任意时刻、任意空间节点的物理量都可以通过一组 POD 基底向量及一组时域相关的投影系数 α 表示。

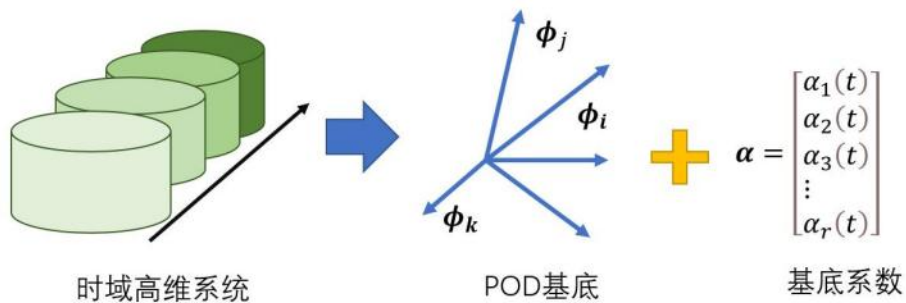


图 4 POD 降阶示意图 [14]

对于一个非线性传热系统, 温度场分布可以用 POD 基底 ϕ_j 的线性组合来表示:

$$\mathbf{T} = \Phi \alpha$$

其中 Φ 为基底矩阵, 表示为

$$\Phi = [\phi_1 \ \phi_2 \ \phi_3 \ \cdots \ \phi_r]$$

α 为系数矩阵, 表示为

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1(t) \\ \alpha_2(t) \\ \alpha_3(t) \\ \vdots \\ \alpha_r(t) \end{bmatrix}$$

将上式代入到系统方程中，得到：

$$\mathbf{C}\Phi\dot{\boldsymbol{\alpha}} + \mathbf{K}\Phi\boldsymbol{\alpha} = \mathbf{F}$$

两边同时乘以 Φ^T ，得到：

$$\Phi^T\mathbf{C}\Phi\dot{\boldsymbol{\alpha}} + \Phi^T\mathbf{K}\Phi\boldsymbol{\alpha} = \Phi^T\mathbf{F}$$

令

$$\mathbf{C}_r = \Phi^T\mathbf{C}\Phi$$

$$\mathbf{K}_r = \Phi^T\mathbf{K}\Phi$$

$$\mathbf{F}_r = \Phi^T\mathbf{F}$$

原系统方程变为：

$$\mathbf{C}_r\dot{\boldsymbol{\alpha}} + \mathbf{K}_r\boldsymbol{\alpha} = \mathbf{F}_r$$

如图 5 所示，这样就将原系统方程中待求解的变量 \mathbf{T} 转化为求解 POD 基底系数 $\boldsymbol{\alpha}$ ，未知量规模由 n 阶降为 r 阶。

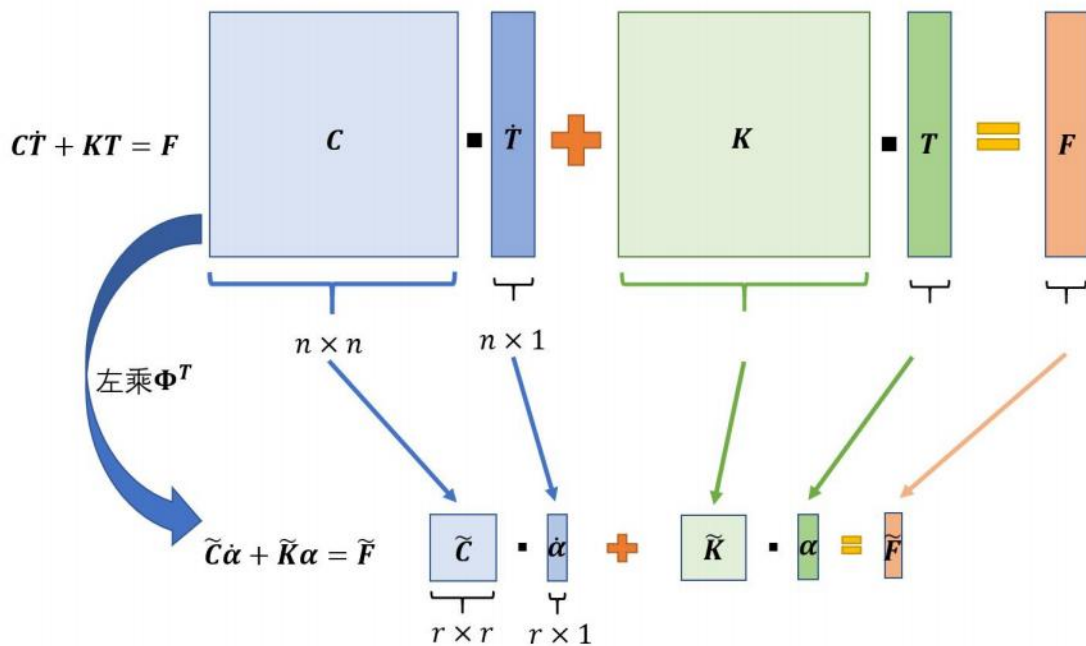


图 5 系统矩阵降阶

2.4 考虑流量影响的非线性传热系统方程

含对流影响的非线性传热系统方程为

$$C\dot{T} + (K_{cond} + K_{conv})T = F'_{source} + F_{conv}$$

降阶后的系统方程为

$$C_r \dot{\alpha} + K_{total_r} \alpha = F_{total_r}$$

其中

$$K_{total_r} = \Phi^T (K_{cond} + K_{conv}) \Phi$$

$$F_{total_r} = \Phi^T (F'_{source} + F_{conv})$$

这里 K_{cond} 表示没有对流边界（即 $h=0$ ）时，仅包含固体传导和固定边界条件的系统热导矩阵；

F'_{source} 表示在参考功率下，热源在边界或体内产生的载荷向量；

而热导矩阵中的边界对流项 K_{conv} 和载荷向量中的边界对流项 F_{conv} 可以写为：

$$K_{conv} = \bar{h} K_{conv_base}$$

$$F_{conv} = \bar{h} T_{\infty} F_{conv_base}$$

其中 K_{conv_base} 代表仅由对流边界对系统热导矩阵的贡献， F_{conv_base} 表示当 $h=1$ 且 T_{∞} 增加 1 度时，由对流边界条件对载荷向量的贡献， \bar{h} 代表传热薄膜上与流体接触部分的换热系数，这里做了线性近似， T_{∞} 代表流体温度。

那么它们的降阶矩阵为

$$K_{cond_r} = \Phi^T K_{cond} \Phi$$

$$K_{conv_base_r} = \Phi^T K_{conv_base} \Phi$$

$$F'_{source_r} = \Phi^T F'_{source}$$

$$F_{conv_base_r} = \Phi^T F_{conv_base}$$

因此降阶方程可以写为：

$$C_r \dot{\alpha} + (K_{cond_r} + \bar{h} K_{conv_base_r}) \alpha = F'_{source_r} + \bar{h} T_{\infty} F_{conv_base_r}$$

2.5 对 POD 时间系数的计算

前文通过对系统矩阵进行降阶，将待求解变量从温度 T 变为时间系数 α ，传统的对 POD 时间系数的求解采用差分算法，这导致求解时需要严格控制时间步长，但是对于宏模型来说，控制时间步长会导致额外的接口，并且这会导致宏模型不能完全反映加热电阻温度随时间的实时动态变化过程。本文利用状态空间方程对时间系数进行描述求解，这样可以保证解的时间连续性：

$$\frac{d\alpha}{dt} = \mathbf{C}_r^{-1} [(\mathbf{F}'_{source_r} + \bar{h}T_{\infty}\mathbf{F}_{conv_{base_r}}) - (\mathbf{K}_{cond_r} + \bar{h}\mathbf{K}_{conv_{base_r}})\alpha]$$

2.6 Verilog-A 宏模型实现

Verilog-A 是一种具有高级和模块化特性的硬件描述语言，用 Verilog-A 语言描述的模块可以用 Spectre 模拟器进行仿真[17]。图 6 所示是利用 Verilog-A 描述建立的 MEMS 热式流量传感器的系统级宏模型，该模型包含 5 个端口，其中 $heating$ ， $flow$ 是输入端口，分别表示加热电阻功率和流体流速； T_{up} ， $T_{heating}$ ， T_{down} 是输出端口，分别表示上游电阻温度、加热电阻温度和下游电阻温度。

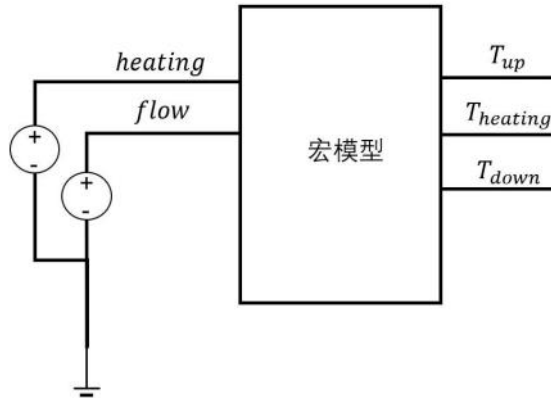


图 6 MEMS 热式流量传感器宏模型

3 仿真与实验结果

本文比较了 10 阶 POD 降阶宏模型系统级仿真结果和有限元仿真结果，比较了 MEMS 热式流量传感器在输入恒功率 7mW 条件下，不同流速(0 m/s, 1 m/s, 2 m/s, 4 m/s, 6 m/s, 10 m/s)流体中的响应，图 7 对比了加热电阻瞬态仿真结果，

其中电压是宏模型的结果，温度是有限元仿真结果；当 MEMS 流量传感器工作稳定后，宏模型和有限元仿真结果如图 8 所示，最大误差不超过 1.3%。表 1 对比了两种模型的仿真时间，在相同 CPU 型号下，全阶有限元瞬态仿真需要 9000 ms，而宏模型瞬态仿真时间仅 41.255 ms，相较于全阶模型仿真时间缩短了 218 倍。

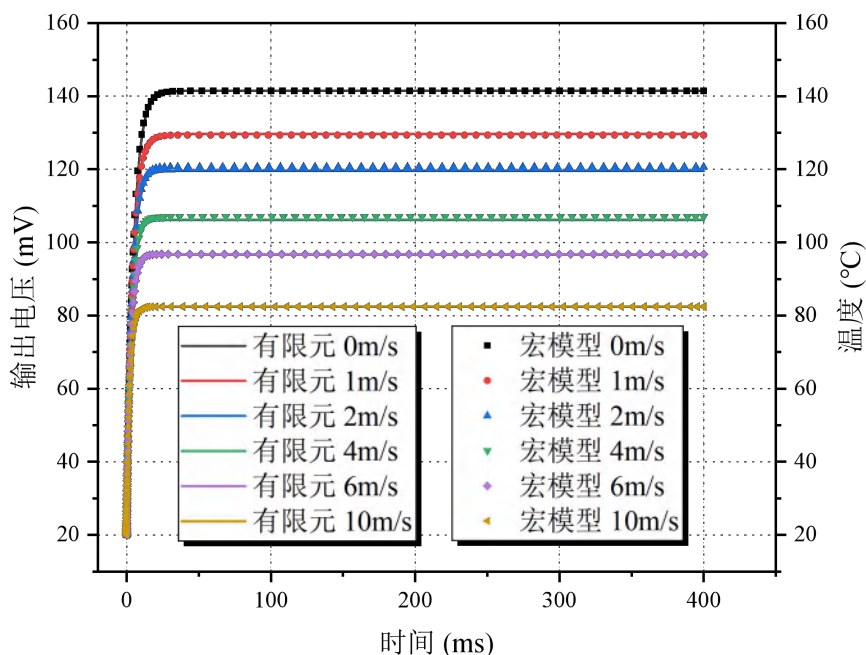


图 7 有限元与宏模型瞬态结果对比

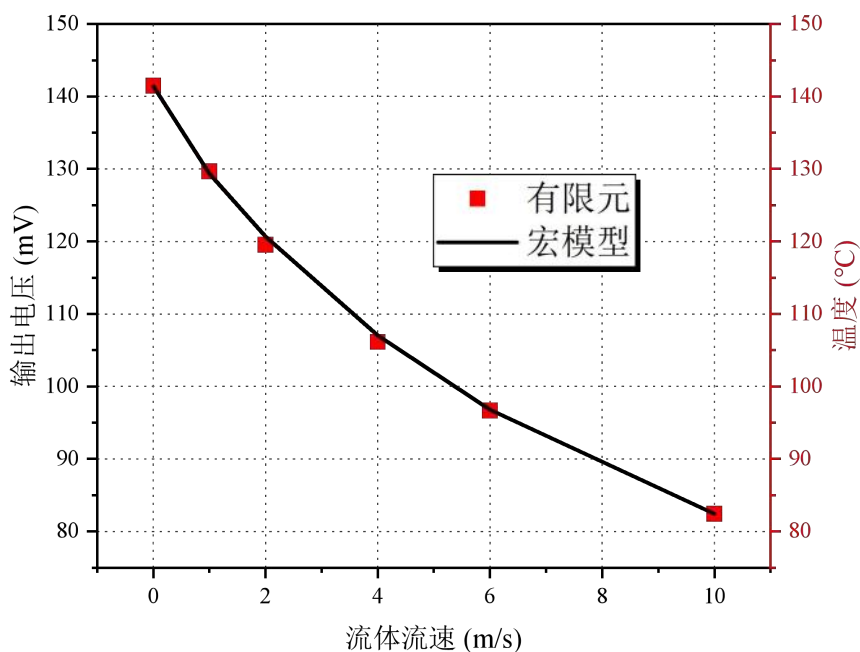


图 8 有限元与宏模型稳态结果对比

表 1 模型瞬态计算时间对比

模型类型	计算时间	加速比 (对比有限元)
有限元 (FEM)	9000 ms	1 ×
宏模型 (POD ROM)	41.255 ms	218.15×

本文还对所使用的 MEMS 热式流量传感器进行了实验测试，测量了恒功率 7mW 输入条件下，加热电阻在 0-10m/s 流体流速中的温度情况。图 9 展示了实验测试过程，图 10 对比了宏模型仿真结果与实验结果，最大误差小于 7.9%，表明本文所建立的宏模型具有较高精度。

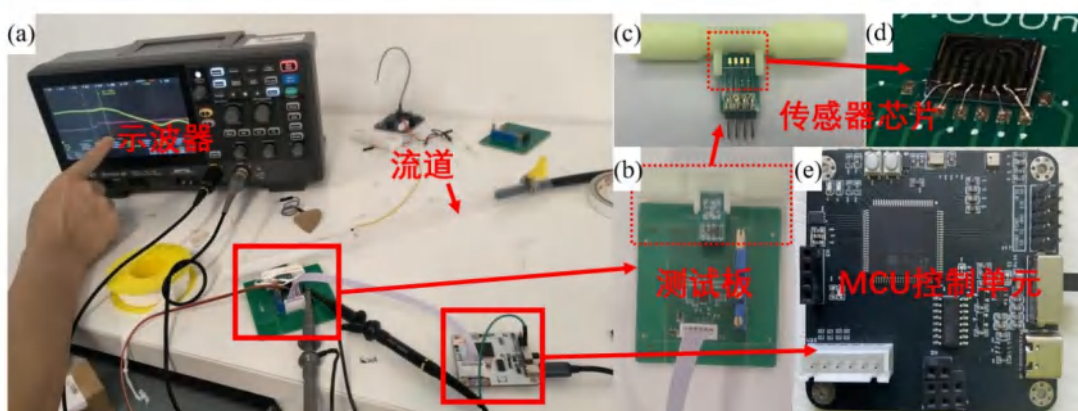


图 9 实验测试

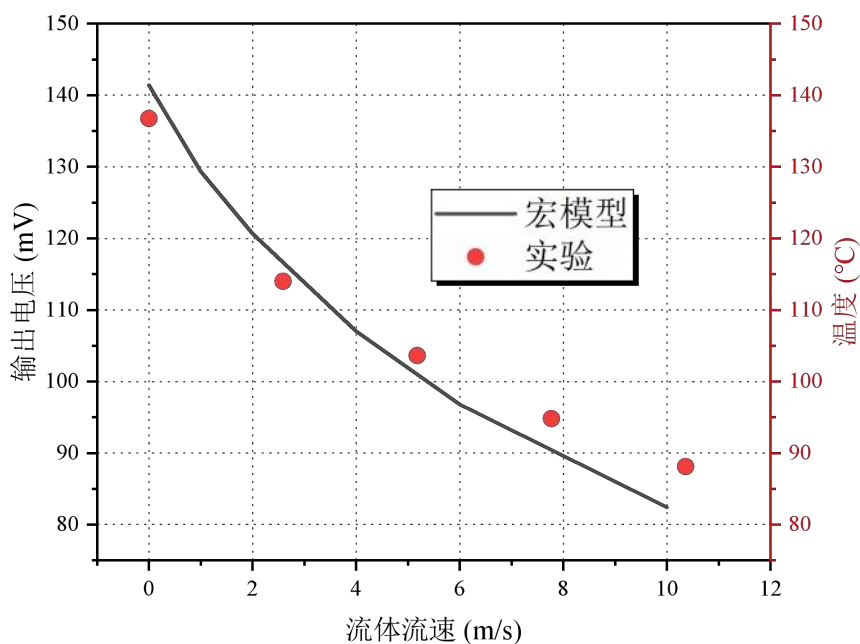


图 10 宏模型与实验测试稳态结果对比

4 结论

本文提出了一种基于 POD 降阶与状态空间方程的宏模型提取方法,该方法以 MEMS 热式流量传感器为例,实现了电-热-流三场耦合 MEMS 器件的宏模型提取。本文对比了 MEMS 热式流量传感器 10 阶 POD 宏模型系统级仿真结果和有限元结果,结果表明,系统级仿真与有限元相比最大误差不超过 1.3%,与实验结果相比最大误差不超过 7.9%,宏模型瞬态仿真时间仅 41.255 ms,相较于有限元模型仿真时间缩短了 218 倍。以上结果表明本文提供了一种高效准确且具有一定通用性的宏模型提取方法,该方法结合了 POD 方法的快速降阶模式和状态空间方程的连续时域描述,针对电-热-流三场耦合的 MEMS 器件形成了一套完整的系统级宏模型建模流程,这对于复杂 MEMS 设计以及系统级仿真有很好的参考价值。

参考文献

- [1] Yi, Z., Ye, Y., Qin, M., & Huang, Q. A. (2019). Modeling of packaged MEMS thermal wind sensor operating on CP mode. *IEEE Transactions on Electron Devices*, 66(5), 2375-2381.
- [2] Hong, L., Xiao, K., Song, X., Lin, L., & Xu, W. (2025). System-level modeling with temperature compensation for a CMOS-MEMS monolithic calorimetric flow sensing SoC. *Microsystems & Nanoengineering*, 11(1), 13.
- [3] Xu, W., Wang, X., Ke, Z., & Lee, Y. K. (2021). Bidirectional CMOS-MEMS airflow sensor with sub-mW power consumption and high sensitivity. *IEEE Transactions on Industrial Electronics*, 69(3), 3183-3192.
- [4] Y. Zhu, B. Chen, M. Qin and Q. -A. Huang, "2-D Micromachined Thermal Wind Sensors—A Review," in *IEEE Internet of Things Journal*, vol. 1, no. 3, pp. 216-232, June 2014, doi: 10.1109/JIOT.2014.2319296.
- [5] Ejeian, F., Azadi, S., Razmjou, A., Orooji, Y., Kottapalli, A., Warkiani, M. E., & Asadnia, M. (2019). Design and applications of MEMS flow sensors: A review. *Sensors and Actuators A: Physical*, 295, 483-502.

- [6] Li, Z., Fang, Z., Wang, B., Ahmed, M., Pan, X., Han, S. T., ... & Xu, W. (2022, May). System-level modeling and design of a temperature compensated CMOS MEMS thermal flow sensor. In 2022 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 2072-2076). IEEE.
- [7] Xu, W., Ma, S., Wang, X., Chiu, Y., & Lee, Y. K. (2019). A CMOS-MEMS thermoresistive micro calorimetric flow sensor with temperature compensation. *Journal of Microelectromechanical Systems*, 28(5), 841-849.
- [8] Wang, Z., Yi, Z., Tian, R., Qin, M., Huang, Q. A., Zhou, Z. F., & Long, K. (2021). Analysis and compensation of benchmark drift of micromachined thermal wind sensor caused by packaging asymmetry. *IEEE Transactions on Industrial Electronics*, 69(1), 950-959.
- [9] Xu, H., Wang, Z., Zhou, Z., Yi, Z., Qin, M., & Huang, Q. A. (2023). An efficient macromodel extraction strategy for packaged MEMS thermal wind sensor considering multiphysics. *IEEE Sensors Journal*, 23(6), 5879-5891.
- [10] Xu, H., Zhao, L. F., Zhou, Z. F., Yi, Z. X., Qin, M., & Huang, Q. A. (2024). A Novel Parametric System-Level Modeling Method for MEMS Devices Combining Artificial Neural Networks and Behavior Description. *Journal of Microelectromechanical Systems*.
- [11] Huang, Q. A., Chen, B., Zhu, Y. Q., & Qin, M. (2015). Modeling of temperature effects on micromachined silicon thermal wind sensors. *Journal of Microelectromechanical Systems*, 24(6), 2033-2039.
- [12] Cao, Z., Wang, Z., Yi, Z., Qin, M., & Huang, Q. A. (2022). Analytic model of dual-layer-structure MEMS thermal wind sensor with increased sensitivity. *IEEE Transactions on Electron Devices*, 69(3), 1341-1348.
- [13] 梁钰,郑保敬,高效伟,等.基于 POD 模型降阶法的非线性瞬态热传导分析[J].中国科学:物理学 力学 天文学,2018,48(12):36-45.
- [14] Rao, S. S. (2010). The finite element method in engineering. Elsevier.
- [15] 梁钰.基于自由单元法的瞬态非线性热传导问题模型降阶及参数辨识方法研究[D].大连

理工大学,2021.DOI:10.26991/d.cnki.gdllu.2021.003876.

- [16] Zhang, X., & Xiang, H. (2015). A fast meshless method based on proper orthogonal decomposition for the transient heat conduction problems. *International Journal of Heat and Mass Transfer*, 84, 729-739.
- [17] Dewey, A., Dussault, H., Hanna, J., Christen, E., Fedder, G., Romanowicz, B., & Maher, M. (1999). Energy-based characterization of microelectromechanical systems (MEMS) and component modeling using VHDL-AMS. *Proc. Modeling and Simulation of Microsystems, Semiconductors, Sensors, and Actuators (MSM 99)(San Juan, Puerto Rico, 19–21 April 1999)*.

应变下单层 2H-MoTe₂ 的振动响应

朱玲玉

上海交通大学, 上海 200240, 中国

摘要: 基于密度泛函理论 (DFT) 和密度泛函扰动理论 (DFPT) 计算, 我们系统研究了单层 2H-MoTe₂ 在等轴应变下的振动响应。研究表明, 在 Γ 点上, 拉曼活性模 (E' , A_1' 和 E'') 与红外活性模 (A_2'' 和 E') 的频率偏移呈现相似趋势: 拉伸应变下频率单调下降, 而压缩应变下频率先升高后迅速下降。在 K 点处, 一个特定的声学模式在压缩应变为 -11.27% 时发生软化并且频率降至零。分析发现, Mo 的 dz^2 轨道以及 Te 的 p_x 、 p_y 轨道中的电子占据削弱了 K 点表现为 Mo 原子的面内振动与 Te 原子的面外振动的振动模。此外, 压缩应变增强了费米能级附近的电子态嵌套, 导致该声学模式的振动频率急剧降低。这些结果为通过振动频率测量探测单层 2H-MoTe₂ 的应变状态提供了一条有效途径。本研究中, 该材料独特的应变敏感振动特性, 使其有潜力开发为高灵敏度的应变传感器元件, 或用于构建基于应变调控谐振频率的可调谐纳机电系统谐振器。

关键词: 单层 2H-MoTe₂; 振动响应; 等轴应变; 声学模式软化; 电子态嵌套; 纳机电系统

1 引言

过渡金属硫族化合物 (TMDs) 因其独特的二维结构及优异的电学、光学性质而受到广泛关注¹。其中, 2H 相的 MoTe₂ 由于其在光电设备和场效应晶体管中的潜在应用价值, 成为研究热点之一²。尤其是在纳机电系统 (NEMS) 领域, 二维材料因其原子级厚度、超高应变极限和低能耗特性, 正推动新一代谐振器向高频化、微型化发展³。然而, 尽管已有不少关于 2H-MoTe₂ 的研究探讨了其相变路径, 如化学修饰、栅控或静电掺杂以及单轴应变, 但对于等双轴变形条件下该材料的晶格振动特性尤其是声子模式频率变化规律尚不明确⁴。本研究首次量化揭示 MoTe₂ 声子模的应变依赖规律, 填补了二维 NEMS 谐振器核心材料基础参数的空白, 并探索如何利用这些振动特性来检测单层 2H-MoTe₂ 的应变状态, 并调控共振频率。

2 方法

采用 QUANTUM ESPRESSO 软件包进行平面波赝势方法计算, 使用

Perdew-Burke-Ernzerhof (PBE) 泛函, 设定截断能量为 40 Ry 以保证总能量精度优于 10^{-8} Ry。为了获取动态矩阵用于声子计算, 采用 $32 \times 32 \times 1$ 的 k 点网格对布里渊区积分电子态, 并用 $8 \times 8 \times 1$ 的 q 点网格。此外, 为验证计算结果的可靠性, 还采用了杂化泛函 (HSE06) 进行对比计算, 发现 PBE 和 HSE06 结果在声子频率的总体趋势上一致, 但 HSE06 在高频区域的频率值略高约 5-8%, 这与 TMDs 中常见的强电子关联效应相关。通过对比不同泛函的计算结果, 进一步验证了本研究中声子频率变化趋势的普适性, 为后续实验设计提供了理论支持。此外, 计算中还引入了范德华修正 (DFT-D3) 以更精确地描述层间相互作用, 尽管单层体系未直接涉及层间作用, 但这一修正有助于提高对原子弛豫过程的模拟精度。

3 结果与讨论

研究发现, 在 -5% 至 5% 中等应变范围内, 光学支的频率通常随压缩应变增加而上移, 随拉伸应变增加而下移。然而, 对于声学支而言, 特别是在 K 点存在一种异常模式, 即其频率随压缩应变减小而下降, 随拉伸应变增加而上升。特别地, 当施加的压缩应变达到临界值 -11.27% 时, 此模式频率降为零, 标志着所谓的“K 模式软化”现象。进一步分析表明, 这种现象与原子键长、键角的变化密切相关, 并且可以通过未来拉曼或红外实验验证。不同应变下 A''_2 、 E' 、 A'_1 和 E'' 模式的频率变化趋势, 其中 A''_2 和 E' 模式的频率变化斜率大于 A'_1 和 E'' 模式, 这可能归因于不同的有效质量导致的振动频率对应变的不同响应。随着等轴压缩应变的增加, A''_2 和 A'_1 模式的频率首先增加到某个临界应变值后开始迅速下降, 而对于 E' 和 E'' 模式, 这一转变发生在更早的应变水平。特别值得注意的是, E' 模在 -4% 压缩应变下呈现 7% 频率增益, 而在 +8% 拉伸应变时频率降低 12%, 这种非对称应变响应为设计双极可调谐振器提供了新思路。

K 模式软化的临界应变值 (-11.27%) 与单层 $2H-MoTe_2$ 的层间堆叠方式密切相关。K 点声学模软化临界值 (-11.27%) 可被确立为器件安全操作的

应变阈值。当施加压缩应变时，Mo-Te 键的键长缩短，而 Te-Te 键的键角增大，导致晶体对称性降低。这种结构畸变会显著改变声子色散关系，尤其是在 K 点附近。通过结合声子谱与电荷密度分布的分析发现，K 模式的软化与电子-声子耦合强度的增强直接相关。在压缩应变下，Mo 原子的 d 轨道与 Te 原子的 p 轨道之间的杂化程度增加，导致电子-声子耦合常数显著上升，从而抑制了声子的稳定性。这一机制与二维材料中的电荷密度波（CDW）相变类似，表明 2H-MoTe₂ 在极端应变条件下可能存在类似的集体电子-声子相互作用。

4 结论

综上所述，本工作不仅揭示了等双轴应变下单层 2H-MoTe₂ 的声子谱特征，还深入探讨了导致特定振动模式软化的物理原因。这些发现不仅丰富了我们对这类材料力学行为的理解，也为设计基于 TMDs 的柔性电子、热电及超导器件提供了新思路。未来的工作将进一步探索其他应变条件下的振动特性及其可能的应用前景，也为开发新型可重构 NEMS 谐振器提供了材料设计标准和工艺优化路径。

参考文献

- [1] Zhou, J. et al. A library of atomically thin metal chalcogenides. *Nature* 556, 355 – 359 (2018).
- [2] Sun, Y. et al. Elastic Properties and Fracture Behaviors of Biaxially Deformed, Polymorphic MoTe₂. *Nano Lett.* 19, 761 – 769 (2019).
- [3] Lee, J., Wang, Z., He, K., Shan, J. & Feng, P. X.-L. High Frequency MoS₂ Nanomechanical Resonators. *ACS Nano* 7, 6086 – 6091 (2013).
- [4] Qian, X., Liu, J., Fu, L. & Li, J. Quantum spin Hall effect in two-dimensional transition metal dichalcogenides. *Science* 346, 1344 – 1347 (2014).

Nonlinear Dynamic Characterizations of Few-Layer MoSe₂

Shuang Cai, Yalan Wang, Jiankai Zhu, Bo Xu, Juan Xia and Zenghui Wang

Abstract— Nanoelectromechanical systems devices hold great promise for fundamental research and applications. Here we present a contactless method to characterize the mechanical properties of few-layer MoSe₂ by analyzing the nonlinear dynamic responses. We extract the Young's modulus of a 7-layer MoSe₂ device ($E = 200$ GPa) and its pretension ($n_0 = 48.44$ mN/m). These findings provide critical insights into the mechanical behaviors of few-layer MoSe₂, enabling its further fundamental study and applications.

I. INTRODUCTION

Nanoelectromechanical systems (NEMS) devices, owing to their linear and nonlinear dynamical behaviors, serve as an excellent platform for fundamental research and applications, such as dynamical phonon softening [1] mode coupling [2], and exciton-optomechanical coupling tuning [3]. Notably, the nonlinear response is closely related to the physical properties of materials, especially their mechanical properties, thus offering a valuable opportunity for non-contact exploration of ultrathin materials [4]. We fabricate resonators using a novel transition metal dichalcogenides, MoSe₂ [5], and further characterize their linear and nonlinear behaviors in the frequency domain. By analyzing these responses, we extract the Young's modulus of the few-layer MoSe₂ and the pretension in the resonators. These results lay the foundation for MoSe₂-based NEMS device applications.

II. EXPERIMENT AND RESULTS

A. Experiment

Fully-clamped circular drumhead MoSe₂ resonators are fabricated via mechanical exfoliation and dry transfer technique. Resonant responses of the devices are measured using a custom-built 2D NEMS resonator measurement system [6] (Fig. 1 (a)). Device motion is characterized under both undriven and electrostatically driven conditions, using interferometric detection. Device thickness is determined by Raman spectroscopy (Fig. 1(b-c)). We detect the frequency response curves of a MoSe₂ resonator (inset of Fig. 1(a), with thickness $h = 4.9$ nm and radius $R = 3.5$ μ m) at $V_{DC} = -0.2$ V with different AC driving amplitudes (Fig. 1(d)).

B. Results

We can approximate the nonlinear response of the fundamental resonance mode by the Duffing equation:

$$m_{\text{eff}}\ddot{x} + c\dot{x} + k_1x + k_3x^3 = \xi F_d \cos(\omega t) \quad (1)$$

Zenghui Wang and Juan Xia are the corresponding author and with the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, P. R. China, zenghui.wang@uestc.edu.cn, juanxia@uestc.edu.cn. Shuang Cai, Yalan Wang, Jiankai Zhu and Bo Xu are with the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, P. R. China.

x is the deflection of the membrane center, c is the damping constant, k_1 and k_3 are the linear and cubic spring constants, $m_{\text{eff}} = \alpha m$ is the effective mass, and ξF_d represents the applied electrostatic force. $\alpha = 0.2695$ and $\xi = 0.432$ are the coefficients related to the mode shape. The parameters in equation (1) are related to the fundamental resonance frequency $f_0 = 4.14$ MHz, the quality factor by $Q = 2\pi f_0 m_{\text{eff}}/c$ and $k_1 = (2\pi f_0)^2 m_{\text{eff}}$. The f_0 is directly related to the pretension of the device: $n_0 = 0.69\pi^2 f_0^2 R^2 \rho h$, ρ is the mass density of the device (for the device, $n_0 = 48.44$ mN/m).

Through fitting analysis of the experimental data, we extract a cubic spring constant $k_3 = 2.44 \times 10^{14}$ N/m³, which achieves excellent agreement with measurements as demonstrated in Fig. 1(e-h) at $V_{DC} = -0.2$ V and Fig. 1(i-l) at $V_{DC} = 0.1$ V. We observe a reduction in the quality factors Q with the AC driving amplitudes increasing, consistent with previously reported on graphene mechanical resonators [4].

Based on the nonlinear membrane theory, the nonlinear mechanics of the device can be related to its material parameters. The Young's modulus E of isotropic materials can be determined from the k_3 by [4]:

$$E = \frac{(1.27 - 0.97\nu - 0.27\nu^2)R^2}{\pi h} k_3. \quad (2)$$

where ν is the Poisson's ratio of the material. Using this equation (2), we further extract the Young's modulus of few-layer MoSe₂ $E = 200$ GPa.

III. CONCLUSION

We measure the nonlinear dynamics of a 7-layer MoSe₂ resonator, extracting its Young's modulus ($E = 200$ GPa) and pretension ($n_0 = 48.44$ mN/m). These results establish a foundation for further exploring mechanical behaviors and developing functional devices based on MoSe₂ resonators.

REFERENCES

- [1] Yang R., et al. "Raman spectroscopic probe for nonlinear MoS₂ nanoelectromechanical resonators," *Nano Letters*, vol. 22, no. 14, pp. 5780-5787, July 2022.
- [2] Chen C., et al. "Direct observation of coherent energy transfer in nonlinear micromechanical oscillators," *Nature Communications*, vol. 8, no. 1, pp. 15523, Jan. 2017.
- [3] Xie H., et al. "Tunable exciton-optomechanical coupling in suspended monolayer MoSe₂," *Nano Letters*, vol. 21, no. 6, pp. 2538-2543, March 2021.
- [4] Davidovikj, D., et al. "Nonlinear dynamic characterization of two-dimensional materials," *Nature Communications*, vol. 8, no. 1, pp. 1253, Nov. 2017.
- [5] Qiu, H., et al. "Two-dimensional materials for future information technology: status and prospects," *Science China Information Sciences*, vol. 67, no. 6, pp. 160400, May 2024.
- [6] Zhu J., et al. "Frequency Scaling, Elastic Transition, and Broad-Range Frequency Tuning in WSe₂ Nanomechanical Resonators," *Nano Letters*, vol. 22, no. 13, pp. 6086-6091, May 2022.

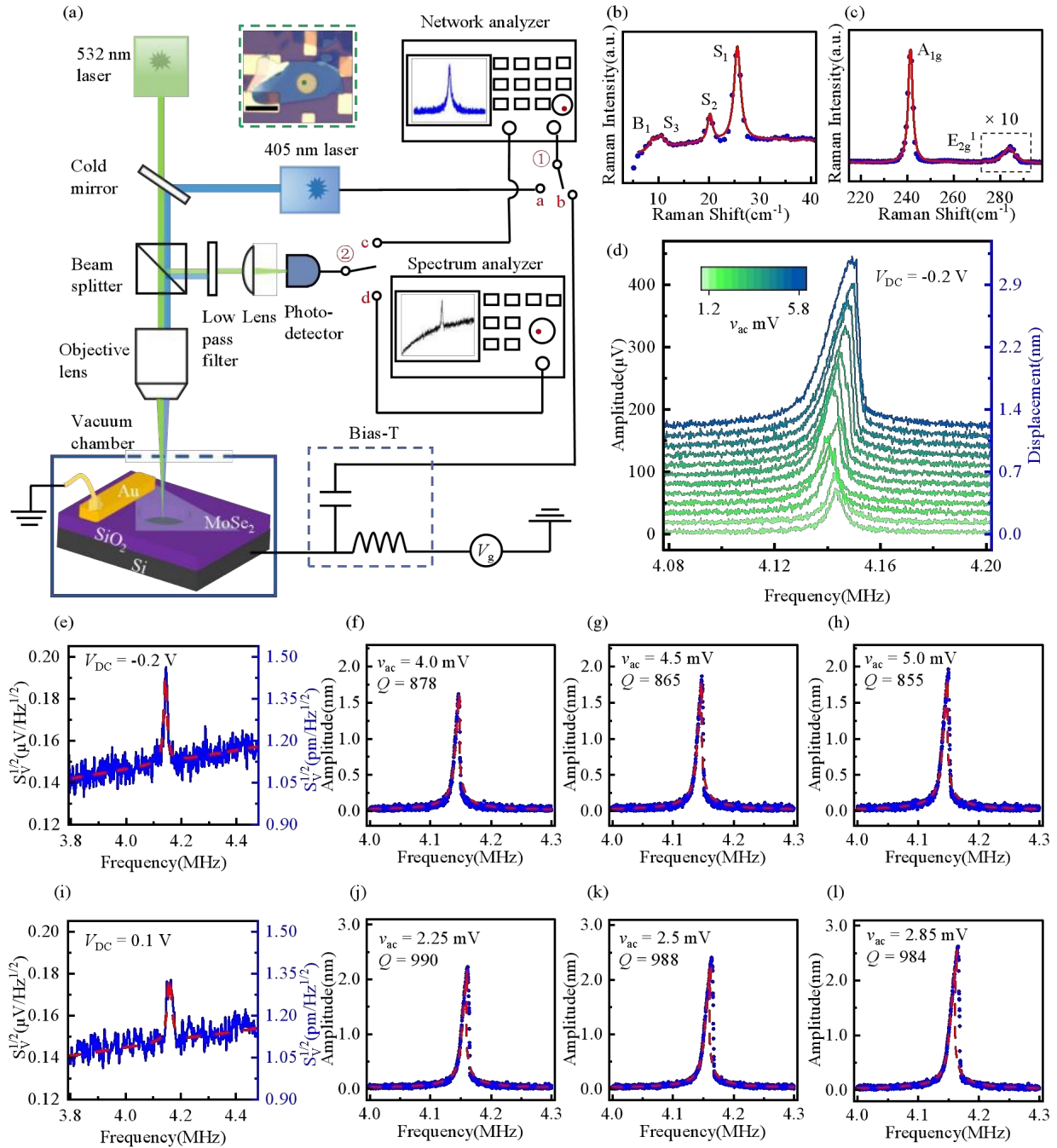


Figure 1. (a) Schematic of the custom-built 2D NEMS measurement system. Three different measurement configurations can be realized for detecting resonances with electrical excitation (switch ① is at “b”, and switch ② is at “c”), optothermal excitation (switch ① is at “a”, and switch ② is at “c”) and thermomechanical resonance without driven (only switch ② is at “d”). The optical image of the device is shown in the inset with a green dashed box, in which the green dot indicates the laser spot position, all scale bar: 10 μm . (b-c) Low-frequency Raman Spectrum and High-frequency Raman Spectrum (determined device thickness is 7 layers) of the fabricated MoSe₂ device, B₁ in (b) represents the interlayer breathing mode, S₁, S₂ and S₃ in (b) represent the interlayer shear modes. (d) Frequency response curves of the 7L MoSe₂ NEMS resonator, driven resonances measured at $V_{\text{DC}} = -0.2$ V with different AC driving amplitudes (1.2 mV to 5.8 mV), showing the transition from linear to nonlinear response for the device. (e) Thermomechanical (undriven) resonances measured from the device at $V_{\text{DC}} = -0.2$ V. (f-h) Measurements and fits of the nonlinear frequency response curves of the device at $V_{\text{DC}} = -0.2$ V with different AC driving amplitudes. The AC driving amplitude is 4.0 mV, 4.5 mV and 5.0 mV, respectively. (i) Thermomechanical (undriven) resonances measured from the device at $V_{\text{DC}} = 0.1$ V. (j-l) Measurements and fits of the nonlinear frequency response curves of the device at $V_{\text{DC}} = 0.1$ V with different AC driving amplitudes. The AC driving amplitude is 2.25 mV, 2.5 mV and 2.85 mV, respectively.

Few-layer MoSe₂ nanomechanical resonator with high power-to-frequency responsivity at room temperature

Yalan Wan, Shuang Cai, Bo Xu, Jiankai Zhu, Juan Xia and Zenghui Wang

Abstract— In this work, we experimentally study the laser power-dependent resonance behaviors of a tunable tri-layer (3L) MoSe₂ drumhead nanomechanical system (NEMS) resonator. Using the mechanical exfoliation and dry transfer techniques, we fabricate a 3L MoSe₂ NEMS device that exhibits robust vibrations in the very high frequency (VHF) range, with up to 191% frequency tunability. By varying the detection laser power, we further tune the resonance frequency and quality factor of this device, achieving a high power-to-frequency responsivity of -284 ppm/ μ W. Our work demonstrates the great potential of MoSe₂ resonators in laser power sensing, nanoelectronics and opto-mechanics.

I. INTRODUCTION

Recent studies show that molybdenum diselenides (MoSe₂) exhibit intriguing properties compared to sulfide compounds, including narrower bandgap, smaller linewidth, and tunable exciton charging effect, making it promising for the development of electronic devices[1][2][3]. Introducing mechanical degrees of freedom provides researchers unique opportunities to further explore material properties (e.g., thermal

properties [4][5]), and to develop novel functional devices (e.g., bolometers and laser power meters [7]). However, research on MoSe₂ nanomechanical devices or its NEMS-based sensors remains limited. Here, we demonstrate a tri-layer (3L) MoSe₂ drumhead resonator, with its device frequency electrostatically tuned from 14.21 to 41.36 MHz ($\sim 191\%$ tuning range ($\Delta f/f_0$)). Furthermore, we observe that both the resonance frequency (from 13.9 to 14.3 MHz) and quality factor (from 165 to 180) can be effectively tuned by varying the power variation (from 4.6 to 94.1 μ W). In addition, the high power-to-frequency response rate of -284 ppm/ μ W of this device surpasses most 2D NEMS devices.

II. DEVICE FABRICATION

The drumhead MoSe₂ resonator is fabricated via mechanical exfoliation and dry transfer techniques (Fig. 1a, c). And the sample thickness is confirmed as tri-layer (3L) by layer-dependent Raman and PL spectra (Fig. 1d, e).

III. RESONANCE DETECTION

We study the resonance response of MoSe₂ resonators using a custom-built two-dimensional (2D) resonator measurement system (Fig. 2a), with device motion recorded in two configurations[6], i.e., electrical excitation (Fig. 2b) and thermomechanical noise measurements (Fig. 2c). All devices are measured under vacuum ($\sim 1 \times 10^{-6}$ Torr).

Juan Xia and Zenghui Wang are the corresponding authors and with Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 610054, China, juanxia@uestc.edu.cn, zenghui.wang@uestc.edu.cn. Yalan Wang, Shuang Cai, Bo Xu and Jiankai Zhu are from Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 610054, China

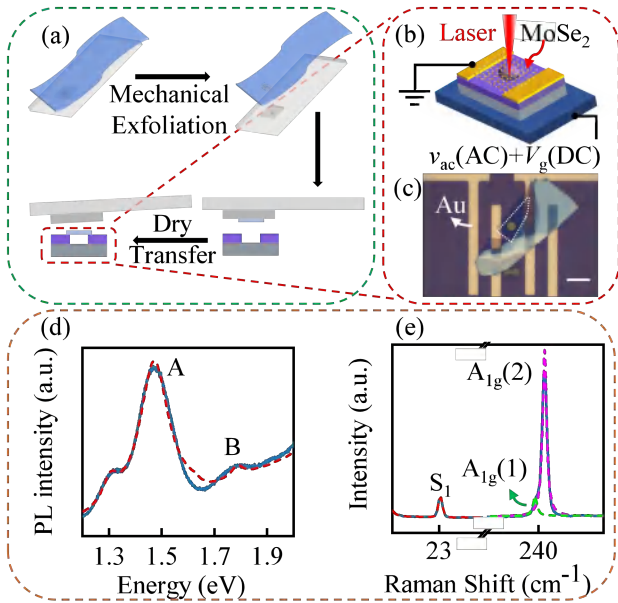


Figure 1. Schematic illustration of device fabrication and structural characterization of MoSe₂ atomic layers. (a) Mechanical exfoliation and dry transfer process. (b) A fully clamped circular drumhead MoSe₂ resonator with contact electrodes, showing the electrical connections. (c) Optical image of suspended MoSe₂ membrane with diameter $d = 4 \mu\text{m}$. Scale bar: $10 \mu\text{m}$. (d) PL spectrum of 3L MoSe₂ with 532 nm excitation, showing A and B excitons (red dashed line: fit). (e) Raman spectrum of 3L MoSe₂ showing the interlayer shear mode (S_1) and Davydov-splitting breathing modes ($A_{1g}(1)$ and $A_{1g}(2)$), fitted with red, green, and pink dashed lines, respectively.

IV. RESULTS AND DISCUSSION

A. Frequency Tunability

Fig. 2d-e demonstrates the electrostatic frequency tuning characteristics of the MoSe₂ resonator, achieving a wide tuning range ($\Delta f/f_0$) of 191% and excellent tuning efficiency of 12.25%/V. These values surpass those of other 2D NEMS resonators based on MoX₂, such as MoS₂ (154%, 5.1%/V) and MoTe₂ (12.3%, 0.8%/V)[8], highlighting the superior tunability of MoSe₂ for advanced NEMS-based functional devices.

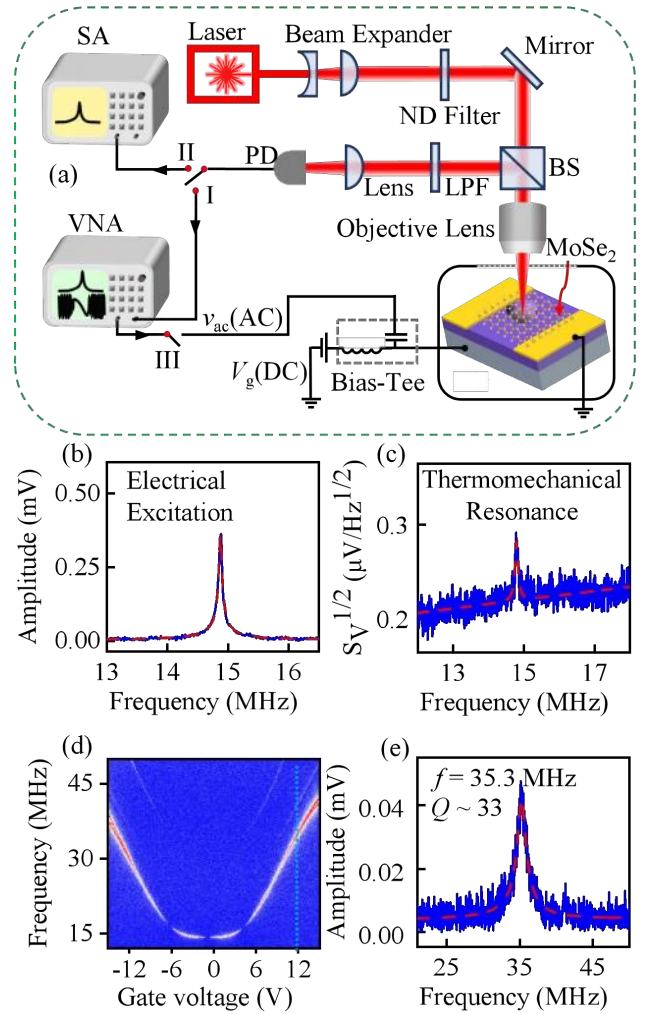


Figure 2. Measurement setup and results for the 3L MoSe₂ resonator. (a) Two electrical configurations: I and III for electrical excitation (b), and II for thermomechanical noise measurements (c). BS: beam splitter; LPF: long-pass filter; PD: photodetector; SA: spectrum analyzer; VNA: vector network analyzer. (d) Resonance amplitude at varying DC gate voltages with fixed AC drive (12.8 mV). (e) Resonance signal at $V_g = 12 \text{ V}$ (blue dashed line in (d)).

B. Laser Power Dependence

Fig. 3 shows the fundamental mode resonant response of the MoSe₂ device using electrical excitation under different 633 nm detection laser powers. In Fig. 3a, the device exhibits a significant frequency shift from 13.9 to 14.3 MHz with a laser power variation of just $89.46 \mu\text{W}$. As the laser power increases, a linear frequency decrease is observed, as shown in Figs. 3b-d,

demonstrating a high responsivity of -284 ppm/ μ W at room temperature, surpassing most drumhead NEMS, e.g., molybdenum disulfide resonator (MoS_2 , -41.16 ppm/ μ W) and highlighting its potential for high-performance light sensors. Additionally, the quality factor (Q) decreases from $Q \approx 180$ (Fig. 3b) to $Q \approx 165$ (Fig. 3c) with higher laser power, consistent with prior research on MoS_2 [7].

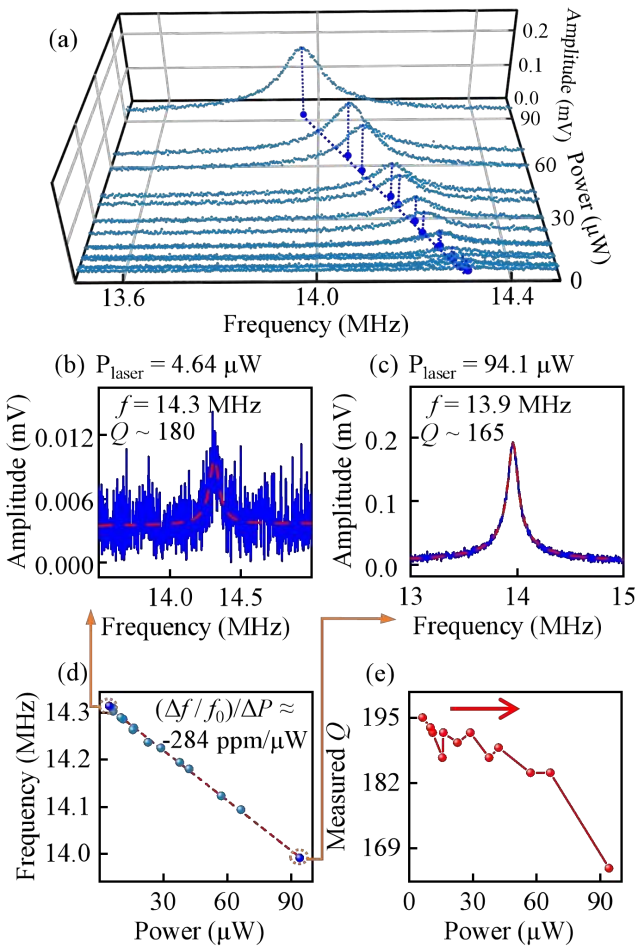


Figure 3. Measured resonances of the fully-covered MoSe_2 resonator under varying 633 nm laser powers. (a) Resonance response evolution with increasing laser power, with (b) and (c) showing results at minimum and maximum powers, respectively. (d) Frequency shift versus laser power, showing a responsivity of -284 ppm/ μ W (blue spheres: data; red dashed line: linear fit). (e) Q factor decreases with increasing laser power.

V. CONCLUSION

In summary, we report a 3L MoSe_2 resonator with 191 % frequency tunability and a power

frequency responsivity of -284 ppm/ μ W, showcasing its potential for sensitive laser power metering. This work lays the groundwork for device design, thermal stability, and novel applications in 2D nanoelectronics, optical sensing, and optomechanical systems.

REFERENCES

- [1] H. Qiu et al., “Two-dimensional materials for future information technology: status and prospects,” *Sci. China Inf. Sci.*, vol. 67, pp. 160400, May 2024.
- [2] S. Tongay et al., “Thermally driven crossover from indirect toward direct bandgap in 2D semiconductors: MoSe_2 versus MoS_2 ,” *Nano Lett.*, vol. 12, pp. 5576-5580, Oct. 2012.
- [3] H. Xie et al., “Tunable Exciton-Optomechanical Coupling in Suspended Monolayer MoSe_2 ,” *Nano Lett.*, vol. 21, no. 6, pp. 2538-2543, Mar. 2021.
- [4] B. Xu et al., “Nanomechanical resonators: toward atomic scale,” *ACS Nano*, vol. 16, pp. 15545-15585, Sep. 2022.
- [5] L. Wang et al., “Investigating thermal properties of 2D non-layered material using a NEMS-based 2-DOF approach towards ultrahigh-performance bolometer,” *Natl Sci Rev.*, vol. 11, pp. nwae248, Jul. 2024.
- [6] J. Zhu et al., “Analyzing electrostatic modulation of signal transduction efficiency in MoS_2 nanoelectromechanical resonators with interferometric readout,” *Sci. China Inf. Sci.*, vol. 65, pp. 122409, Jan. 2022.
- [7] R. Yang et al., “Calibrating temperature coefficient of frequency (TCf) and thermal expansion coefficient (α) of MoS_2 nanomechanical resonators,” presented at the 2015 Joint Conference of the IEEE International Frequency Control Symposium & the European Frequency and Time Forum, Denver CO, USA, April 12-16, 2015, pp. 198-201.
- [8] P. Prasad et al., “Gate Tunable Cooperativity between Vibrational Modes,” *Nano Lett.*, vol. 19, pp. 5862-5867, Sep. 2019.

Laser-Tunable Gold Nanomechanical Resonators

Jiaqi wu, Jiankai Zhu, Luming Wang and Zenghui Wang

Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China,
Chengdu, China

Ultrathin gold materials exhibit unique optical, thermal, and electrical properties, making them well-suited for biological imaging, hydrogen sensing, and photovoltaic applications. However, their mechanical properties and potential in NEMS-based sensing remain underexplored.

In this study, we synthesize single-crystal gold nanoflakes and further demonstrate that gold nanomechanical resonators can function effectively as laser sensors. Specifically, we employ a custom-built interferometry system to detect ultrathin gold circular drumhead resonators at room temperature and vacuum conditions. By varying the laser power from 2.31 μW to 478 μW over 18 cycles, we observe a periodic frequency shift ($\Delta f = 55.4$ kHz), corresponding to a relative responsivity of 11.9 ppm/ μW . Furthermore, our analysis reveals an excellent linear correlation between the resonance frequency and the laser power, with R^2 values exceeding 0.98. To broaden the operational frequency range of the gold resonators, we fabricate 35 gold nanomechanical resonators of varying sizes and derive the frequency scaling law for such devices. Through this analysis, we successfully extract the Young's modulus of gold $E_Y = 60 - 80$ GPa with the surface tension $\gamma = 0.05 - 0.5$ N/m. These findings highlight the unique advantages of 2D gold resonators and their exceptional potential for high-performance integrated sensing applications.

High-Precision Determination of Complex Refractive Indices in Ultrathin ReSe₂ Towards Enhanced Optomechanical Transduction Efficiency

苏子洛, 王璐明, 朱健凯, 王曾晖

电子科技大学基础与前沿研究院, 成都, 610054

摘要: We propose a spectroscopic approach to quantitatively determine the complex refractive indices of ultrathin ReSe₂, facilitating optimization of optical signal transduction efficiencies for advanced optomechanics and nanoelectromechanical systems.

关键词: complex refractive indices; ReSe₂; transduction efficiency

1 Introduction

Rhenium diselenide (ReSe₂) stands out due to its distorted 1T' phase, leading to anisotropic optical and electrical properties^[1]. Studying its light-matter interaction is essential for advancing nanoelectromechanical systems (NEMS) by optimizing light absorption or transmission efficiencies. However, accurately determining the complex refractive index ($n - j\kappa$) of ultrathin ReSe₂ remains challenging due to the microscale lateral dimensions and nanoscale thickness of the flakes. In this work, we propose a spectroscopy-based method for quantitatively determining complex refractive indices of ultrathin ReSe₂. Our approach, extendable to other 2D materials, unprecedentedly enable tailored design of optical characteristics, promoting the development of optomechanics based on nanoscale ReSe₂.

2 Theoretical analysis

We leverage the reflection spectrum of thin film for quantitatively analyzing the complex refractive indices. Specifically, a three-interface interferometric model based on the Fresnel's Law is established, yielding collective reflectance as follows: ^[2]

$$R = \left| \frac{r_1 e^{i(\varphi_2 + \varphi_3)} + r_2 e^{-i(\varphi_2 - \varphi_3)} + r_3 e^{-i(\varphi_2 + \varphi_3)} + r_1 r_2 r_3 e^{i(\varphi_2 - \varphi_3)}}{e^{i(\varphi_2 + \varphi_3)} + r_1 r_2 e^{-i(\varphi_2 - \varphi_3)} + r_1 r_3 e^{-i(\varphi_2 + \varphi_3)} + r_2 r_3 e^{i(\varphi_2 - \varphi_3)}} \right|^2 \quad (1)$$

Here, r_1 , r_2 and r_3 are reflection coefficients at air-ReSe₂, ReSe₂-SiO₂, and

SiO₂-Si interfaces, respectively: $r_i = (\mathbf{n}_i - \mathbf{n}_{i+1}) / (\mathbf{n}_i + \mathbf{n}_{i+1})$, while $\varphi_i = 2\pi\mathbf{n}_i d_i / \lambda$ is the associated phase shifts due to variations in optical path length, in which λ is wavelength, d_i is thickness of related layer, and $\mathbf{n}_i = n_i - j\kappa_i$ denotes the complex refractive indices of related material. Optical reflectance contrast of a ReSe₂ flake on a SiO₂/Si substrate can be defined as:

$$C = \frac{R - R_{\text{substrate}}}{R + R_{\text{substrate}}} \quad (2)$$

which effectively minimizes spectral variations due to passive optical components for ensuring stable spectroscopic analyses.

3 Experimental Results

We build a fine-engineered spectroscopy setup (Fig. 1a) to measure white-light reflectance of nanoflakes. A total of 16 ReSe₂ samples with varied thicknesses are prepared *via* mechanical exfoliation and dry transfer technique (Fig. 1b), forming an extraordinary degree of freedom – the thickness dimension – for spectroscopy analyses. Spatially-resolved reflectance mapping (Fig. 1c) visually reveals the morphological spectral features, with each pixel representing a local reflection spectrum. By analyzing all samples collectively, we extract contrast spectra correlated with ReSe₂ thickness at specific wavelengths (Fig. 1d). At 633 nm (Fig. 1e), contrast data is fitted to Eq. (2), yielding the complex refractive indices. By applying this procedure iteratively across different thicknesses, we obtain the quantitative spectra of complex refractive indices within 450-750 nm wavelength range, as proposed in Fig. 1f.

With experimentally determined optical constants, we propose effective guidelines for optimizing light-matter interactions in nanodevices, not only enhancing optomechanical device efficiency but also facilitating nanoflake identification during sample preparation. We calculate light absorption spectra of monolayer ReSe₂ as a function of oxide thickness, as shown in Fig. 1g. A valuable insight is that careful device structure design can more than double the absorbance of monolayer flake (from <20% to ~50%). Additionally, we calculate contrast spectra relative to oxide thickness (Fig. 1h), highlighting substantial improvement in nanoflake visibility – from negligible contrast to clearly discernible optical features (~0.5 contrast). This characteristic is evidenced by photomicrographs, in which showing enhanced visibility of a monolayer specimen under a yellow filter (Fig. 1j) compared to a blue filter (Fig. 1i). Both the absorption and contrast analyses provide solid guidelines for advancing the design and performance of ultrathin-material-based NEMS devices.

4 Conclusion

Our work quantitatively determines the complex refractive indices of ReSe₂ in the visible spectrum, highlighting the pronounced refractive characteristics of ReSe₂ even at monolayer thickness. Detailed technical discussions will be presented at the conference.

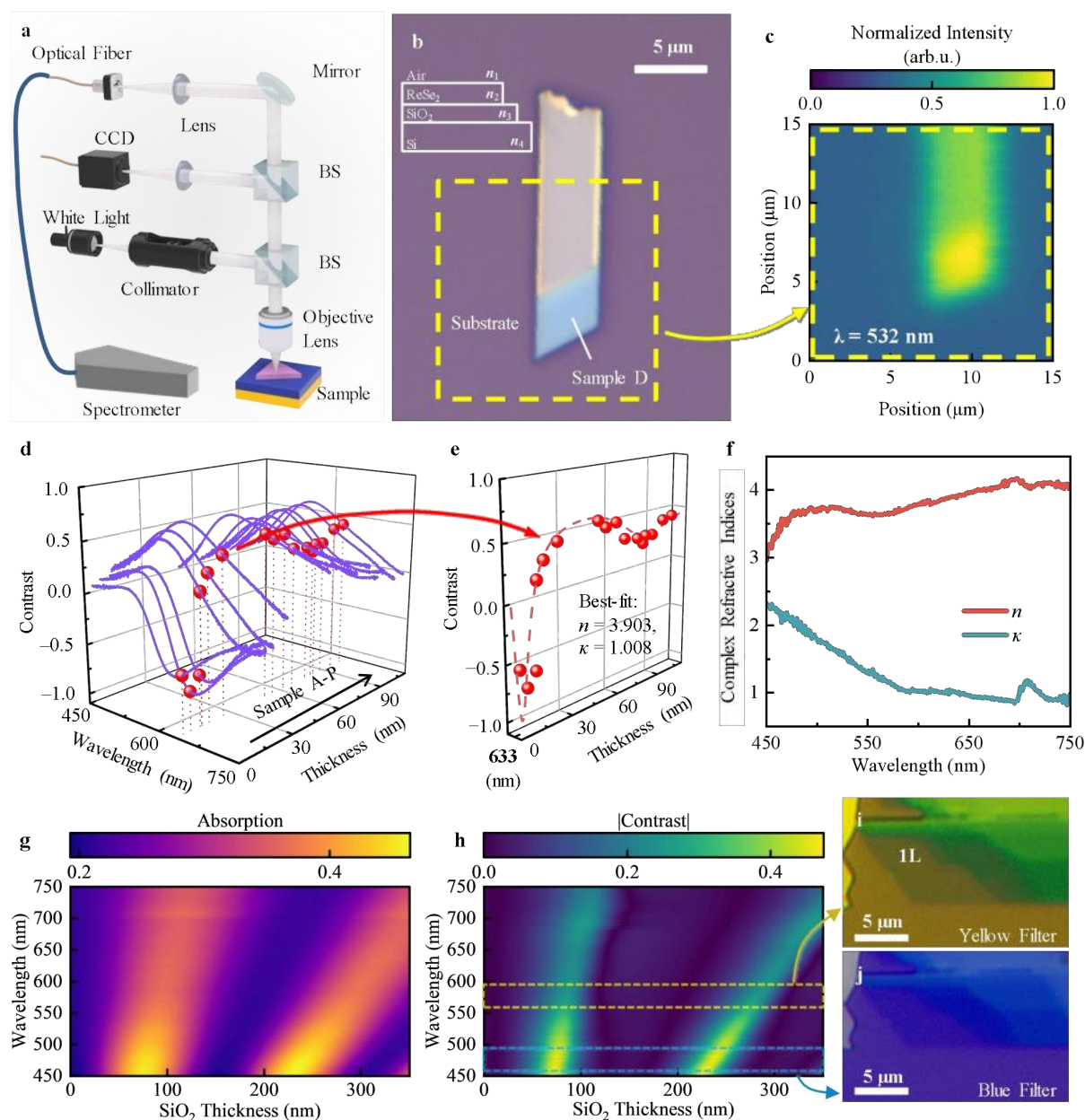


图 1 Determination of complex refractive indices, and analyses of optical

absorption characteristics for ultrathin ReSe₂. (a) Schematic diagram of the

custom-built experimental setup to measure the spatially-resolved optical contrast

spectra of ReSe₂ flakes on SiO₂/Si substrates. (b) Optical image of the studied ReSe₂

flake on silicon substrate with 290 nm oxide layer (inset: the geometric structure of

the three-interface air/ReSe₂/SiO₂/Si model) and (c) the reflectance mapping of 532

nm wavelength, extracted from hyperspectral measurement performed with a spatial resolution of $0.2 \mu\text{m}/\text{pixel}$ and a sampling area of 76×76 Cartesian grid matrix using a motorized stage. The yellow dashed box outlines corresponding scan area. **(d)** Contrast spectra of 16 ReSe_2 samples measured over the $450 - 750 \text{ nm}$ spectral range, each with a distinct thickness up to 110 nm . The red spheres represent the data points at 633 nm . **(e)** The extracted contrast of ReSe_2 as a function of thickness at a wavelength of 633 nm , along with the corresponding fitting curve (best-fit: $n = 3.903$, $\kappa = 1.008$). The red spheres represent the same value as those in (d). The red dashed line indicates the theoretical contrast evolution calculated by the complex refractive indices of ReSe_2 at 633 nm . **(f)** The complex refractive indices spectra of ReSe_2 obtained by curve-fitting the contrast-thickness relationship with Eq. (2) for incident wavelengths ranging from 450 nm to 750 nm . **(g)** The predicted optical absorption and **(h)** contrast spectra of different oxide thicknesses for monolayer samples. **(i-j)** Photomicrographs of a monolayer ReSe_2 sample captured with white-light illumination passed through a yellow and a blue filter, respectively. The approximated transmission band of the filters are depicted by dashed boxes in (h), while the oxide thickness (290 nm) is denoted by the vertical dashed line in (h).

References

- [1] LI X, CHEN C, YANG Y, et al. 2D Re-based transition metal chalcogenides: progress, challenges, and opportunities[J]. *Advanced Science*, 2020, 7(23): 2002320.
- [2] BLAKE P, HILL E W, CASTRO NETO A H, et al. Making graphene visible[J]. *Applied Physics Letters*, 2007, 91(6): 063124.

Modulating Resonance Mode Sequencing in Nanomechanical Resonators

王璐明, 巫佳琦, 朱健凯, 王曾晖

University of Electronic Science and Technology of China

Nanomechanical resonators that utilize suspended two-dimensional (2D) materials as moving components offer remarkable advantages for advanced sensing, actuation and signal processing applications. These delicate devices also exhibit exceptional performance at the device-level, including ultra-high sensitivity and broad tunability. Even more intriguing is that resonators based on mechanically anisotropic 2D materials exhibit enhanced multimodal behaviors with more distinct resonance modes, thus providing extra degrees of freedom for advanced nanoelectromechanical systems (NEMS). However, the controllable modulation of higher-order resonance modes remains largely underexplored, hindering the full exploitation of the extensive multimodal capabilities of NEMS resonators.

In this work, we present the first demonstration of on-site modulation of resonance mode sequencing, in NEMS resonators based on calcium niobate (CaNb_2O_6). By leveraging the material's intrinsic anisotropic mechanical and thermal elastic properties, we employ anisotropic strain engineering to achieve a swapping among higher-order resonance modes. This swapping is further experimentally validated by directly visualizing the corresponding resonance mode shapes. We also explore the role of thermal elastic anisotropy and propose a potential mechanism for frequency-based data encoding.

ATOMICALLY THIN NEMS FREQUENCY COMB WITH BOTH FREQUENCY TUNABILITY AND RECONFIGURABILITY VIA MECHANICAL MIXING OF TWO EXCITATIONS

Bo Xu, Zenghui Wang

Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China.

Novelty / Progress Claim(s)

We report and tune a frequency comb in a 2D Graphene NEMS resonator via mechanical mixing of two excitations. By applying a single-tone excitation and carefully setting the gate voltage V_g , 1:2 internal resonances can be realized and equidistant spectral lines around the fundamental mode frequency are observed. Its frequency response can be adjusted by introducing an additional excitation simultaneously. The frequency spacing and the number of spectral lines can be tuned by the driving amplitude and frequency of the mixing source. The combination of such resonant responses can be leveraged to realize NEMS frequency combs with high reconfigurability and fine tunability.

Background / State of the Art

Phononic frequency combs, like their optical counterparts, are playing an increasingly important role in fundamental research and practical applications, and have been widely demonstrated in MEMS devices [1-2], often leveraging nonlinear responses like internal resonance [3-4]. NEMS resonators based on 2D semiconductors exhibit multimode resonances, rich nonlinear dynamics, and wide frequency tuning range, which make them promising platforms for achieving internal resonance and thus realizing highly tunable phononic frequency comb [5-6]. In this work, we achieve phononic frequency comb by single-tone excitation (f_d) near the fundamental resonance and tune comb by mixing second driving source (f_m). By sweeping f_m and v_2 , both the frequency spacing and the number of spectral lines can both be continuously tuned which providing a new strategy for frequency control.

Description of the New Method or System

Graphene drumhead resonators with electrodes are excited electrostatically and detected using optical interferometry, and both the driving and detection signals can be routed to different instruments (*fig. 1*). The device is shown in the inset of *fig.3(a)*. All resonance measurements are performed in room temperature and under vacuum (10^{-3} mbar).

Experimental Results

We characterize multimode resonances in this Graphene resonator with Network Analyzer when sweeping the gate voltage. The frequency tuning data shown in *fig. 2* show that when $V_g=3$ V, 1:2 internal resonance conditions are met. We drive the device using Network Analyzer (v_1 increased from 2 to 60 mV) which its responses are shown in *fig. 3(a)*. Keeping device driven by single-tone excitation ($f_d=11.6$ MHz and $v_1 = 60$ mV), we observe the frequency comb in spectrum analyzer (shown in *fig. 3(b)*). Then, we apply another excitation (f_m and v_2) to device as well. By sweeping f_m from 11.1 MHz to 12.1 MHz with $v_2 = 10$ mV, the evolution of the frequency spectrum show mixed frequency comb behavior and the frequency spacing Δf clearly tunable by f_m (*fig.4*). Especially, there are frequency mixing around every comb spectral line when increase f_m from 11.4 MHz to 11.8 MHz (illustrated by red dotted box in *fig.4(a)*). Next, we also sweep f_m from 11.1 MHz to 12.1 MHz when fixing $v_2= 50$ mV, 80 mV and 160 mV (*fig.5(a-c)*). The evolution of frequency spectrum show which frequency spacing and the number of lines are both tuned by f_m . The frequency spacing is linearly correlated with the difference among f_d and f_m . Such rich dynamic behavior in the frequency domain is attributed to the simultaneous generation of 1:2 internal resonance and mechanical mixing among two excitations, and the interplay between them which can be leveraged to enable highly tunable and reconfigurable phononic frequency combs. Detailed analysis will be offered in the presentation.

Word count: 550

References

- [1] R. Wei, et al., *Proc. IEEE MEMS 2019*, pp. 165-168.
- [2] G. Gobat, et al., *Proc. IEEE MEMS 2021*, pp. 615-618.
- [3] X. Wang, et al., *Appl. Phys. Lett.*, 120 (2022), 173506.
- [4] D. A. Czaplewski, et al., *Phys. Rev. Lett.*, 121 (2018), 244302.
- [5] A Keşkekler, et al., *Nano Lett.*, 22 (2022), pp. 6048-6054.
- [6] J. Lee, et al., *Proc. IEEE MEMS 2022*, pp. 503-506.

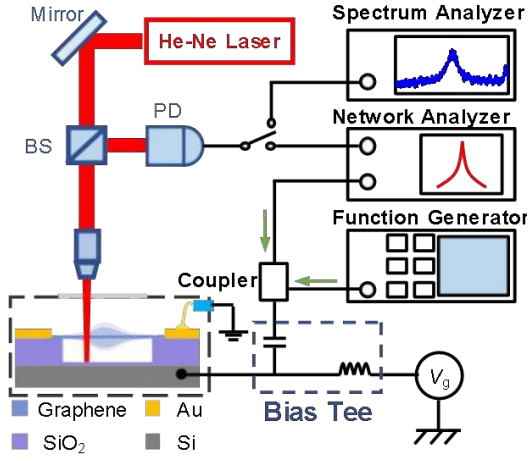


Fig. 1: The customized-built set-up with electrical excitation and optical detection.

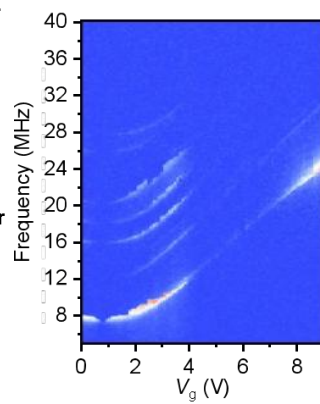


Fig. 2: The frequency tuning curve with $v_1 = 25$ mV when V_g is from 0 to 9 V.

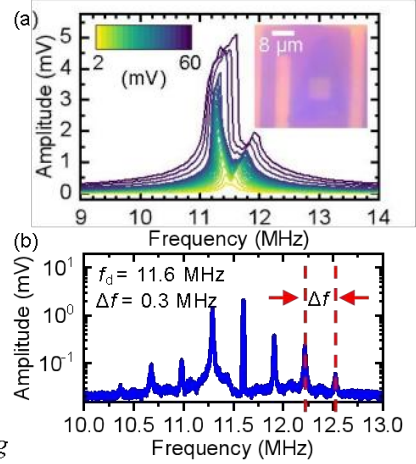


Fig. 3: The spectrum response at $V_g = 3$ V. (a): increasing v_1 from 2 to 60 mV. (b): $v_1 = 60$ mV

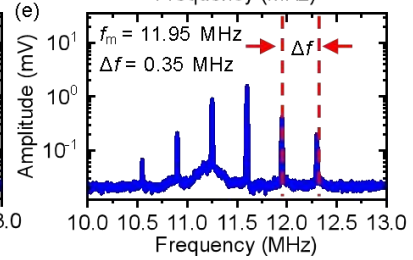
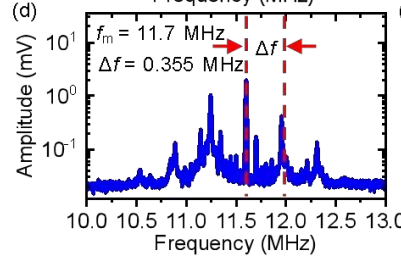
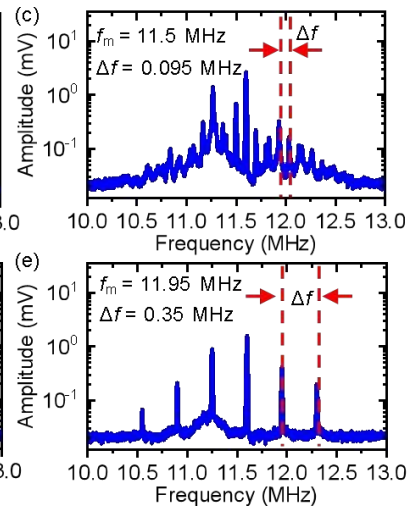
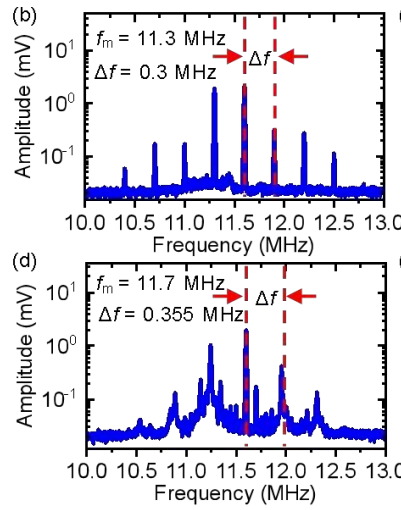
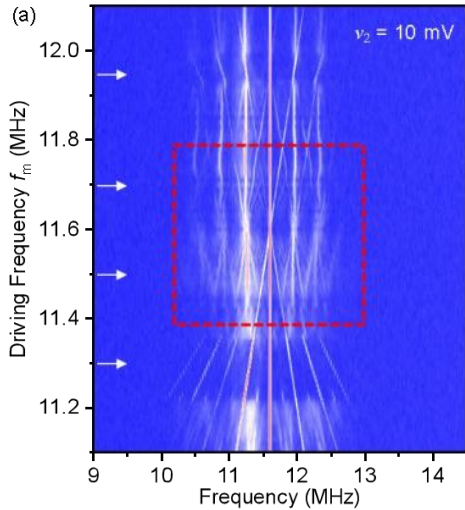


Fig. 4 (a): The frequency spectrum evolution with sweeping driving frequency f_m at $f_d = 11.6$ MHz. (b-e) : The spectrum response at $f_m = 11.3$ MHz, 11.5 MHz, 11.7 MHz and 11.95 MHz with $f_d = 11.6$ MHz are shown, respectively.

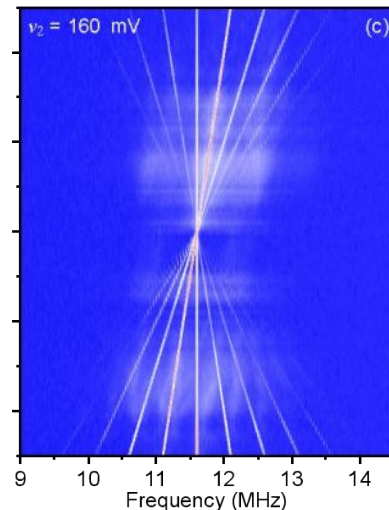
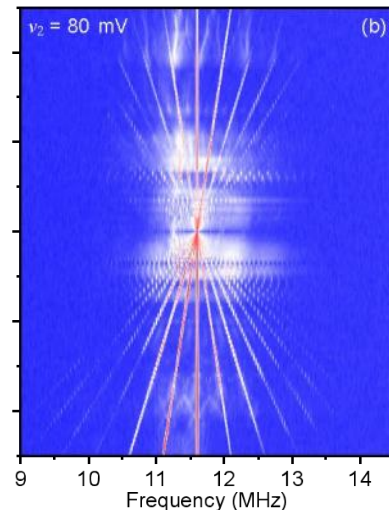
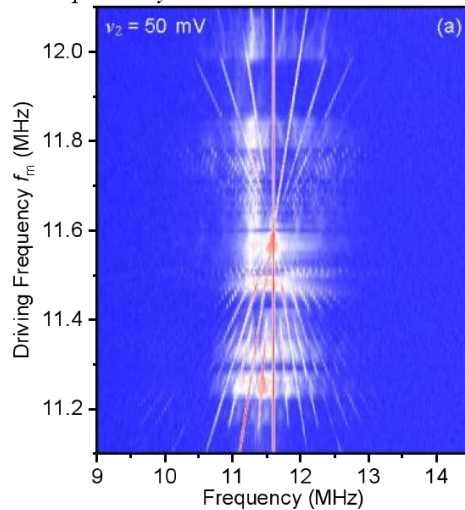


Fig. 5 (a-c): The frequency spectrum evolution with sweeping f_m at $v_2 = 50$ mV, 80 mV and 160 mV, respectively.

高频-甚高频 CrOCl 纳米机电谐振器

吕辰飞, 刘庭昊, 朱健凯, 王曾晖

电子科技大学基础与前沿研究院

摘要: 作为一种新兴的二维材料, CrOCl 因其独特的物理性质受到了广泛关注, 目前该材料已在垂直互补场效应晶体管、低温逻辑逆变器及多态数据存储器等领域展现出优异的性能和应用潜力[1-3]。然而, 目前关于二维 CrOCl 材料力学特性的实验研究仍相对缺乏。为此, 我们基于 CrOCl 二维材料构建了新型纳米机电器件 (NEMS Device), 通过对具有较大厚度和直径范围的 CrOCl 纳米机电谐振器的研究, 我们实现了对二维 CrOCl 力学特性 (杨氏模量, E_f) 的无损提取, 并揭示了这种二维半导体谐振器的频率随器件尺寸的变化规律。这一研究有望为基于二维 CrOCl 晶体的新型 NEMS 器件的开发和进一步的功能器件设计提供指导。

关键词: 纳米机电器件; 二维 CrOCl; 杨氏模量

1. 实验研究

范德华层状的 CrOCl 晶体具有正交结构(空间群 Pmmm, No.59), 其晶格常数为 $a=3.863 \text{ \AA}$ 、 $b=3.182 \text{ \AA}$ 和 $c=7.694 \text{ \AA}$ [4], 图 1(a)展示了单层 CrOCl 的三维晶体结构示意图。我们使用化学气相输运(CVT)技术制备块状 CrOCl 晶体, 通过机械剥离与干法转移制造了全钳位的圆形鼓膜谐振器[6,5], 并使用了原子力显微镜 (AFM)进行厚度测量并计算层数。

进一步, 我们在室温真空环境($\sim 1 \times 10^{-4}$ Torr)下对 CrOCl 谐振器的振动运动进行频域表征。由于 NEMS 器件的振幅极小 (pm 或更低量级), 因此我们开发了定制的二维 NEMS 测量系统, 该系统基于激光干涉进行谐振测量, 能分辨 $\text{fm}/\text{Hz}^{1/2}$ 级的振动[7,8]。具体地, 我们使用 405nm 激光作为激励光, 通过对光强的频率调制, 以周期性的光热效应实现对 NEMS 器件的光热驱动; 与此同时, 另一束 633nm 激光作为检测光, NEMS 器件振动会周期性地改变干涉条件, 从而将振动信息加载到检测激光的光强信号中, 之后被光电探测器转换为电信号用于后续处理。进一步, 通过将收集到的“幅-频”信号拟合到简谐振荡器模型(SHO)中, 我们测定了器件的谐振频率和品质因数(图 1(c))。

2. 数据分析

为了更好的研究频率随器件尺寸的关系, 我们总共制造并测量了 51 个不同厚度(0.76-159.17 nm)和不同直径($d=2, 3, 4, 5 \text{ \mu m}$)的器件, 这些器件在 2.39 MHz

到 75.26 MHz 的频率范围内表现出清晰的共振响应图像,涵盖了高频(HF)到甚高频(VHF)频段。

最后,为了建立二维 CrOCl NEMS 传感器的设计规则,我们引入了圆形鼓膜共振频率理论模型进行理论分析,以此研究其频率缩放规律。对于一个半径为 r 、厚度为 t 的全钳位圆形鼓面谐振器,基模谐振频率可以表示为[9,10]:

$$f_0 = \left(\frac{(kr)}{2\pi} \right) \sqrt{\frac{D}{\sigma r^4} \left[(kr)^2 + \frac{\gamma r^2}{D} \right]} \quad (1)$$

其中, (kr) 是数值求解的模态相关参数, σ (以 kg/m^2 为单位) 是鼓面的面质量密度 (对于 CrOCl, $\rho = \sigma/t = 3.50 \text{ g/cm}^3$), γ (以 N/m 为单位) 是二维张力, $D = E_Y t^3 / [12(1-\nu^2)]$ (其中 E_Y 和 ν 分别为杨氏模量与泊松比) 是弯曲刚度。当 $\gamma r^2/D \rightarrow 0$ 时, f_0 由材料杨氏模量 E_Y 通过弯曲刚度 D 决定,处于“板”状态,此时的基模谐振频率可以表示为:

$$f_0 \approx \beta_0 \sqrt{\frac{E_Y}{12\rho(1-\nu^2)}} \frac{t}{r^2} \quad (2)$$

其中 $\beta_0 = 10.215$ 。基于上述频率-尺寸关系,我们对所有器件的谐振频率数据进行拟合分析(图 1(d)),提取出 CrOCl 晶体的面内平均杨氏模量约为 $E_Y = 110\text{-}145 \text{ GPa}$,这与采用纳米压痕法所测得的结果非常一致[11]。此外,我们还观察到,对于较厚的器件(“板”状态),其谐振频率随着 t/r^2 线性增长,这与理论模型(式 2)吻合的很好。

3. 结论

综合上述研究,我们成功构建了基于范德华层状 CrOCl 晶体的新型二维 NEMS 谐振器,其频率范围为 2.39 至 75.26 MHz,处于高频(HF)和甚高频(VHF)频段。进一步地,我们通过一种无损测量方式提取出 CrOCl 晶体的面内平均杨氏模量约为 $E_Y = 110\text{-}145 \text{ GPa}$,并建立了 CrOCl 圆形鼓膜谐振器的频率与器件尺寸的关系。研究结果为未来基于二维 CrOCl 的谐振传感器件设计提供了关键设计依据与理论支撑。

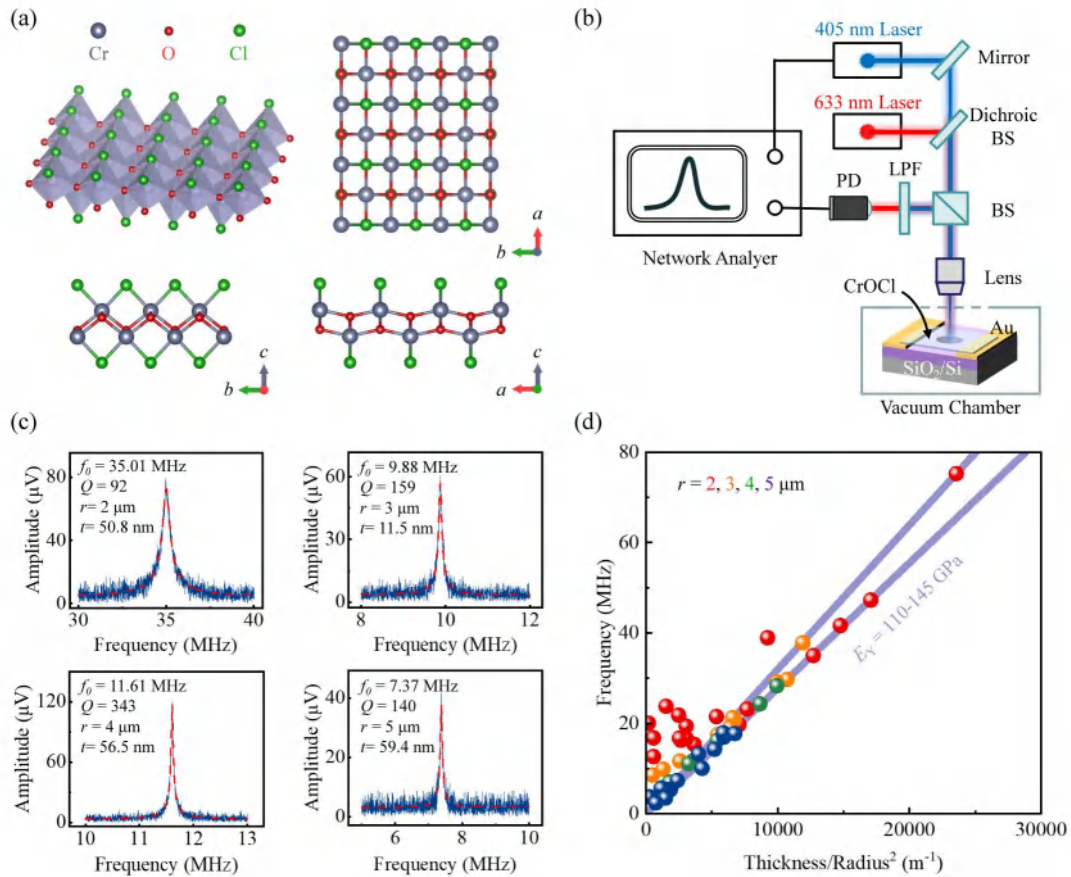


图 1 (a) CrOCl 的晶体结构及其三视图 (b) 自主搭建的谐振频率测试系统, 采用光热激励, 干涉检测 (c) 部分器件的谐振响应图像 (红色虚线为拟合曲线), 器件尺寸信息标注于图中 (d) 测得的 51 个不同直径的 CrOCl 谐振器的谐振频率, 以厚度除以半径的平方 (t/r^2) 作为横坐标绘制。基于器件的频率分布, 提取得出 CrOCl 的杨氏模量范围为 110–145 GPa。

参考文献

- [1] Guo Y, Li J, Zhan X, et al. *Nature*, 2024; **630**(8016): 346-352.
- [2] Gu P, Wang C, Su D, et al. *Nature Communications*, 2023; **14**(1): 3221.
- [3] Yang K, Gao X, Wang Y, et al. *Nature Communications*, 2023; **14**(1): 2136.
- [4] Noerlund Christensen A, Johansson T, Quezel S. *Chemischer Informationsdienst*, 1975; **6**(19): 1171.
- [5] Yang R, Zheng X, Wang Z, et al. *Journal of Vacuum Science & Technology B*, 2014; **32**(6): 061203.
- [6] Liang Y, Zhu J, Xiao F, et al. *IEEE Journal of The Electron Devices Society*, 2021; **9**: 1269–1274.
- [7] Zhu J, Zhang P, Yang R, et al. *Sci. China Inf. Sci.*, 2022; **65**(2): 122409.
- [8] Zhu J, Wang L, Wu J, et al. *Chinese Phys. Lett.*, 2023; **40**(3): 038102.
- [9] Suzuki H, Yamaguchi N, Izumi H. *Acoust. Sci. & Tech.*, 2009; **30**(5): 348–354.
- [10] Wang Z, Feng P, X,-L. *2D Materials*, 2015; **2**(2): 021001.
- [11] Li X, Sha X, Yan N, et al. *Magnetochemistry*, 2022; **8**(12): 170.

A switchable THz filter by integrating an H-shaped metamaterial and ultra-small MEMS switches

Ying Huang¹ and Yoshiaki Kanamori²

¹Sun Yat-sen University, 518066, China

²Tohoku University, 980-8577, Japan

E-mail: huangy888@mail.sysu.edu.cn, Tel: 13760374507

Abstract: We proposed ultra-fast switchable terahertz (THz) filters consisting of a metamaterial with ultra-small MEMS switches and fabricated static on- and off-state-filters offering a 55%-transmittance-difference at 0.61 THz. A 0.9 MHz-tuning-speed is expected.

Key words: MEMS; Optical filter; Terahertz, Finite element method

I. Introduction

Terahertz (THz) waves fall within 0.1–10 THz in frequency, have many potential applications owing to their spectral specificity. To further develop THz wave science and technologies, dynamic filters that actively regulate the transmission of waves are sought to. In 2020, an active THz filter by integrating an electromagnetically induced transparency metamaterial (MM) with micro-electro-mechanical systems (MEMS) was realized^[1]. However, their tuning speed is low due to the heavy mass of big MEMS actuators and the limit of temperature-change speed. In this study, a switchable THz filter integrating an H-shaped MM (H-MM) with ultra-small MEMS ohmic switches is proposed, fabricated, and discussed its characteristics. A calculated 0.9 MHz-tuning-speed is expected by using the ultra-small MEMS cantilevers. The shape of H-MM is mechanically changed by MEMS switches through applying drive voltage. Thus, transmittance of the filter is electromechanically switchable.

II. Design and fabrication

Figure 1 illustrates a schematic of a proposed active filter. Figure 2 shows fabrication processes of the filter. A quartz substrate is used. A 6- μm -linewidth lead array for applying drive voltage is aligned to the cantilever tip side and is embedded under a 0.4- μm -thick SiO_2 film to avoid electrical short-circuit. An H-shaped 2D gold MM^[3] is formed on the surface layer, which consists of two arms and a split-bar with periods P_x and P_y along x and y axes, a linewidth W , a gap G , a split length L , a bar space S , and a thickness T_1 . The MEMS switch array is composed of suspended cantilevers with a beam length L , a beam height H , and a thickness T_2 . Designed and fabricated values of parameters are list in Table 1. The MEMS switch array is arranged upon the split-bar of H-MM. By applying a potential between leads and H-MM, the suspended tip of the cantilever beam will pull down to touch H-MM. That means the split-bar of H-MM is electromechanically deformed into a connect-bar. Correspondingly, the filter switches from the off- ($H = 1.25 \mu\text{m}$) to the on-state ($H = 0 \mu\text{m}$). Figure 3 shows a bird's eye view of fabricated filter with downward cantilevers. By controlling the deposition condition in step (i) (see Fig. 2), the cantilever can be either straight or downward (i.e., in contact with H-MM) after sacrificial layer removal, which can be regarded as a static off- and on-state, respectively.

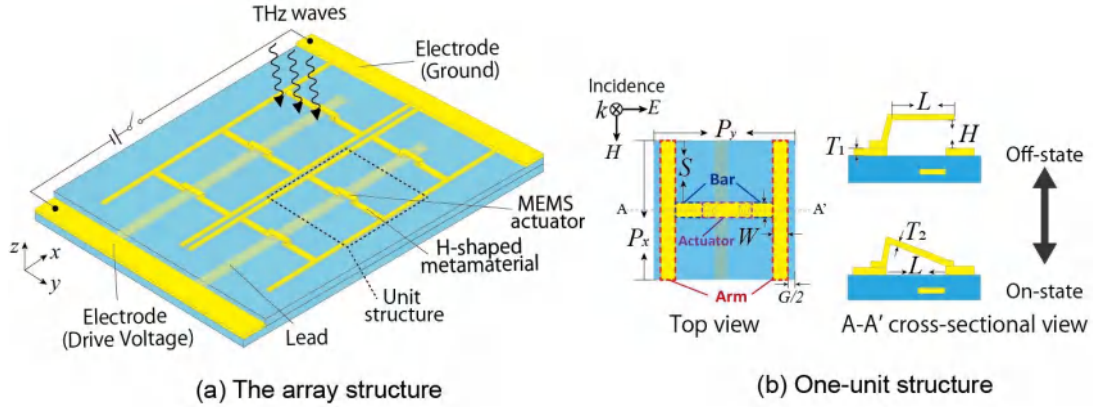


Fig. 1. Schematic of the proposed switchable THz filter. (a) Bird's eye view of the array structure. (b) Top view (left) of one-unit structure and its cross-sectional view (right) at on- and off-state.

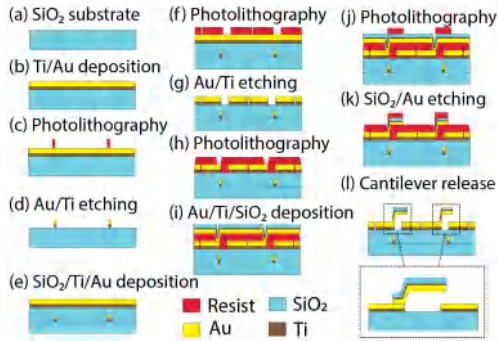


Fig. 2. Fabrication processes of the proposed device.

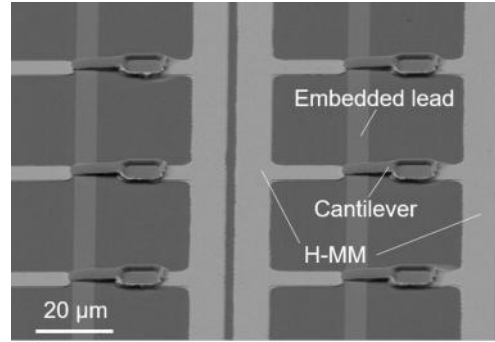


Fig. 3. An SEM image of fabricated static on-state filter.

Table I. Comparison of the design and fabrication values of dimensions of the device.

Parameters	Bar space	Gap	Line width	Beam/split length	Beam height	Thickness of anchor	Thickness of beam
Symbols	S	G	W	L	H	T_1	T_2
Design (μm)	28.50	3.00	9.00	12.00	1.25	0.20	0.40
Fabrication (μm)	28.61	2.56	9.38	11.80	1.34	0.23	0.38

III. Optical characteristics

Transmittance of static on-state and off-state filters under y -polarized normal incidences was evaluated. Numerical calculations were implemented via the finite element method. Measurements were performed by THz time-domain spectroscopy (THz-TDS). Calculated and measured transmittance of the filters was plotted in Fig. 4. For the static on-state ($H = 0 \mu\text{m}$) filter, it presents a broad valley with a central frequency of 0.61 THz and a minimum of -29 dB in the transmittance spectrum. For the static off-state ($H = 1.25 \mu\text{m}$) filter, it has a measured transmittance of 0.55 at the frequency of 0.61 THz. This result demonstrates that this filter can obtain a switchable transmittance by switching the on- and off-states.

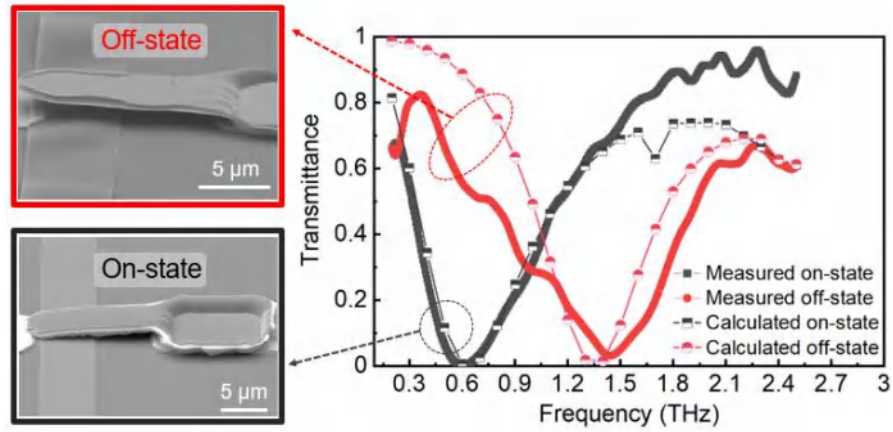


Fig. 4. Calculated and measured transmittance of static on-state filter and off-state filter. Insets show SEM images of the fabricated filters.

IV. Conclusions

An active THz filter was proposed by integrating an H-shaped MM and an ultra-small MEMS switch array. Fabricated static filters experimentally present a 55%-change in transmittance for 0.61 THz wave at static on- and off-state.

- [1] Y. Huang, K. Nakamura, Y. Takida, H. Minamide, K. Hane, and Y. Kanamori, "Actively tunable THz filter based on an electromagnetically induced transparency analog hybridized with a MEMS metamaterial," *Sci. Rep.* **10**, 20807(10pp) (2020).
- [2] Y. Huang, T. Okatani, and Y. Kanamori, "Broadband stop filters for THz waves using H-shaped metamaterials with dual electronic-plasmonic functionality," *Jpn. J. Appl. Phys.*, <https://doi.org/10.35848/1347-4065/ac55dd> (2022) [Accepted manuscript online].

Polymethyl Methacrylate (PMMA) Pyrolysis Assisted Transfer of 2D Materials for Large-Scale NEMS Resonator Arrays

Zuheng Liu^{1,†}, Jianyong Wei^{1,†}, Pengcheng Zhang¹, Yueyang Jia¹, Ying Chen²,

Hao Jia², Zenghui Wang³, and Rui Yang^{1,*}

¹University of Michigan – Shanghai Jiao Tong University Joint Institute,
Shanghai Jiao Tong University, Shanghai 200240, CHINA

²Shanghai Institute of Microsystem and Information Technology,
Chinese Academy of Sciences, Shanghai 200050, CHINA

³Institute of Fundamental and Frontier Sciences,
University of Electronic Science and Technology of China, Chengdu 610054, CHINA

[†]Equal Contribution

*E-mail: ruiyang@sjtu.edu.cn

ABSTRACT

We develop a novel transfer technique based on polymethyl methacrylate (PMMA) pyrolysis process for the fabrication of large-scale two-dimensional (2D) molybdenum disulfide (MoS₂) nanoelectromechanical resonator arrays. We experimentally demonstrate a consistent upward frequency tuning up to 160% and downward quality (Q) factor shift as DC gate voltage (V_{GS}) increases, and the measured resonance frequencies and Q among different devices show smaller variation compared with previous reports. The innovative transfer technique enables consistent and high-quality 2D resonator arrays, and paves the way towards scalable production of 2D NEMS resonators.

KEYWORDS

PMMA Pyrolysis, 2D NEMS, Large Scale Arrays, 2D Resonators

BACKGROUND

2D NEMS resonators are highly promising for a number of applications such as sensing, radio-frequency (RF) signal processing, and computing, thanks to their ultrahigh sensitivity and tunability, and ultralow power consumption [1]. Hence the large-scale integration of uniform 2D NEMS resonator array is essential. Monolayer graphene NEMS resonator arrays have been fabricated by releasing the suspended membranes

through etching of the sacrificial layer [2]. But the material experiences wet chemical processes, which could degrade the quality. In contrast, 2D material films can be dry-transferred onto pre-patterned substrates, also resulting in suspended NEMS arrays [3–5]. Towards even higher uniformity, polymers such as PMMA have been used for transferring 2D materials onto substrates followed by etching away the polymer using solvents [1]. However, solvents lead to the collapsing of 2D materials due to capillary force. In this manuscript, we demonstrate a highly scalable fabrication technique for uniform 2D NEMS resonator arrays, where the PMMA is removed by thermal decomposition in a vacuum furnace after the transfer of monolayer MoS₂ grown by chemical vapor deposition (CVD), resulting in circular “drumhead” resonators.

FABRICATION PROCESS

Fabrication of the large-scale 2D MoS₂ resonator arrays starts from the growth of centimeter-scale monolayer MoS₂ film on an oxidized silicon (Si) substrate. PMMA is then spin-coated onto the MoS₂ film on the growth substrate, and baked for 2 minutes at 180°C to form a solid film. Then the growth substrate is immersed in deionized water,

where the MoS₂–PMMA film can be separated. After that, the MoS₂–PMMA film is picked up with the assistance of deionized water, and then placed onto a target Si-SiO₂ substrate. On the substrate, we have etched the circular-shaped microtrenches with 2 μm diameters, and patterned the contact electrodes formed by 5 nm chromium (Cr) and 35 nm gold (Au). Then, the substrate with film is heated in a vacuum furnace at 390°C for ~240 minutes, during which the PMMA is decomposed, leaving behind the suspended 2D MoS₂ resonator arrays with microtrenches and contact electrodes (Fig. 1).

EXPERIMENTAL RESULTS

Atomic force microscopy (AFM) measurements show an almost flat membrane suspended on the microtrench, with a slight downward deflection of ~3 nm (Fig. 2a,b). We obtain the Raman peak separation of ~19.3 cm⁻¹ and high photoluminescence (PL) peak intensity (Fig. 2c,d), which match well with the distinct signatures of monolayer MoS₂ [3]. The statistics of Raman and PL peak positions as well as full width at half maximum (FWHM) show concentrated distribution (Fig. 2e–g), indicating high uniformity of the material. Resonant responses with V_{GS} sweeping from 0 to 12 V and AC excitation voltage (v_{ac}) fixed at 30 mV, as well as the data with V_{GS} fixed at 4 V and v_{ac} sweeping from 5 mV to 60 mV are recorded for 10 MoS₂ resonators, with results from 3 representative devices shown in Fig. 3. We observe consistent upward frequency tuning characteristics with similar tuning ranges (Fig. 3a,c,e), and clear transition from linear to nonlinear resonances as v_{ac} increases (Fig. 3b,d,f). We summarize the frequency and Q tuning properties under various V_{GS} (Fig. 4a,b), which shows high uniformity of frequency tuning with $\Delta f/f_0$ up to 160%, and Q tuning range $|\Delta Q/Q_0|$ of 47%. Summaries of resonance frequency and Q show much improved consistency compared with previous reports (Fig. 4c,d) [2,4,5]. The novel

fabrication technique paves the way towards scalable production of 2D NEMS resonators with high performance and high uniformity.

REFERENCES

- [1] B. Xu, *et al.*, *ACS Nano*, vol. 16, p. 15545, 2022.
- [2] A. M. van der Zande, *et al.*, *Nano Lett.*, vol. 10, p. 4869, 2010.
- [3] R. Yang, *et al.*, *J. Vac. Sci. Technol. B.*, vol. 32, 061203, 2014.
- [4] Z. Liu, *et al.*, *Proc. 36th IEEE Int. Conf. on Micro Electro Mechanical Systems (IEEE MEMS)*, Munich, Germany, Jan. 15–19, 2023, p. 281.
- [5] H. Jia, *et al.*, *Nanoscale*, vol. 8, p. 10677, 2016.

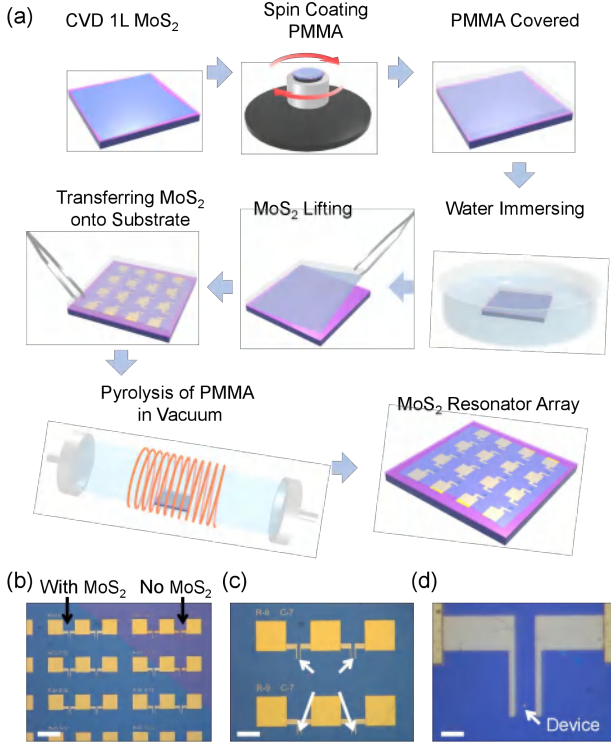


Figure 1: Fabrication process and optical images of the resonator array. (a) Illustration of the fabrication process. (b-d) Optical images of the fabricated resonator array. Scale bars: (b) 100 μm , (c) 50 μm , and (d) 10 μm .

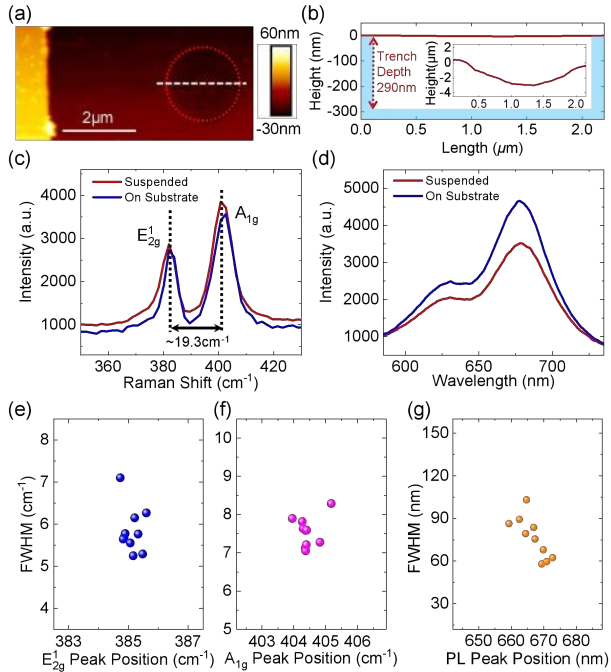


Figure 2: Material and structure characterization of the resonator array. (a-b) (a) AFM image, with the red dashed circle showing the resonator, and (b) line scan along the white dashed line in (a), showing the almost flat suspended membrane. (c-d) (c) Raman and (d) PL spectra of the suspended, and the substrate-supported MoS₂ for reference. (e-g) FWHM vs. peak position for E_{2g}, A_{1g}, and PL modes.

(e-g) Statistics of (e) E_{2g}¹ and (f) A_{1g} Raman modes, and (g) PL peak positions and FWHM obtained from fitting.

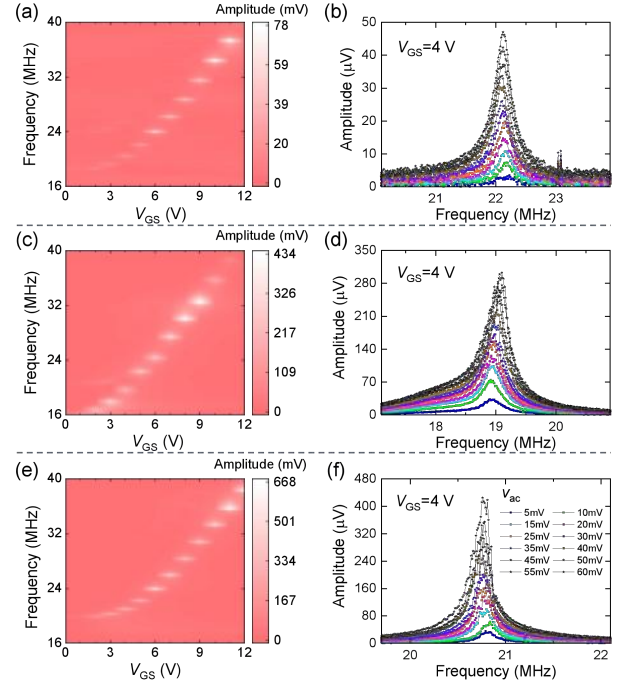


Figure 3: Measurements on the dependence of resonances on V_{GS} and v_{ac} for three typical resonators, with (a–b), (c–d), and (e–f) showing data from the same device, respectively. (a, c, e) Measured amplitude in color scale at different V_{GS} and frequencies, showing consistent resonance tuning, with v_{ac} at 30 mV. (b, d, f) Resonance spectra at increasing v_{ac} with V_{GS} at 4 V.

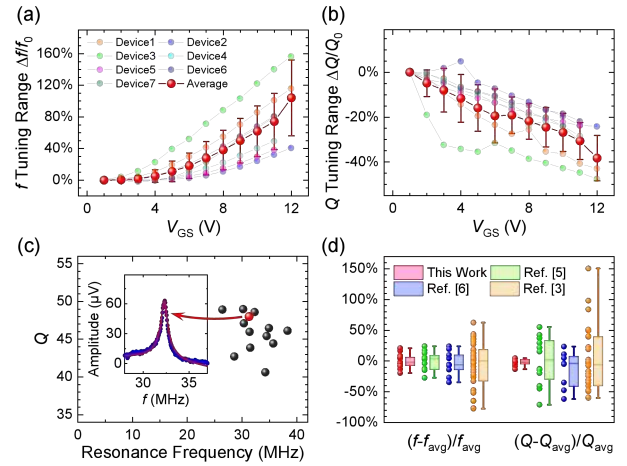


Figure 4: Summary of frequencies and Q for the resonator array. (a-b) Consistent frequency (a) and Q (b) tuning by V_{GS} for different resonators as shown by colors under electrical excitation, with v_{ac} at 30 mV. (c-d) Resonance frequency vs. Q of different resonators (c), with the inset showing the data for the red symbol, and (d) comparison of the frequency and Q distributions with published results under photothermal excitation

基于机器学习的先进CMOS电路单粒子软错误 SPICE仿真

卜建辉

中国科学院微电子研究所

摘要: 随着半导体器件特征尺寸的不断微缩,单粒子效应(SEE),特别是单粒子翻转(SEU)和单粒子瞬态(SET),对CMOS电路的可靠性构成了严峻挑战。近年来,单粒子效应的建模与模拟日益受到重视。

本论文提出了一种基于机器学习的SPICE建模方法,通过统一关键物理变量(漏极电压、线性能量传输(LET)、离子入射位置和温度)之间的非线性关系,精确预测SET引发的瞬态电流。该方法有别于传统的经验性或基于TCAD的方法,其提出的神经网络架构能够从辐射实验或校准的TCAD数据自主学习多参数间的相互依赖关系,实现了良好的预测精度,均方根误差(RMSE)低于5%。另外,开发了一款专用软件工具,用于CMOS电路的单粒子效应仿真,该工具将基于神经网络的SET模型集成到仿真流程中。输入文件包括GDSII版图、SET模型和仿真配置。该工具能够精确识别易受SET影响的敏感器件,从而实现针对性仿真,大幅降低计算开销。输出结果包括每次离子入射产生的瞬态电流波形,以及不同工作状态下的SEE截面数据。标准单元的分析结果还可用于评估大规模数字集成电路的SEE性能。

数据驱动的半导体器件建模方法

游海龙

西安电子科技大学

摘要： 半导体器件紧凑模型对于集成电路电路高效设计和仿真至关重要。基于物理驱动的传统紧凑模型开发需要大量的人工努力，并且可能耗时数年。此外，将新的物理特性（例如辐射效应）纳入现有的紧凑模型并非易事，可能需要从头重新开发。机器学习（ML）技术有潜力实现紧凑模型开发的自动化，提升模型精度与开发速度。本报告介绍器件建模的方法、问题，以及数据驱动的智能化工件建模研究进展及应用。

基于强化学习与图神经网络的晶圆级芯片自动化设计框架

张国和

西安交通大学

摘要：智能算法对计算吞吐量、存储容量及互连效率的需求呈指数级增长，传统芯片架构受光刻尺寸限制，在计算密度、内存带宽和系统扩展性上遭遇了严峻挑战。晶圆级芯片（Wafer-Scale Chips, WSCs）借助大规模异质集成技术，将多个芯粒密集集成到单晶圆上，成为突破上述瓶颈的关键途径，然而其设计优化（如任务分配、物理布局规划、互连延迟控制等）仍缺乏高效的自动化工具。报告将介绍一套端到端的晶圆级芯片自动化设计框架，可实现从任务分解到晶圆级物理布局的全流程优化。该框架包含三个核心模块：1) 动态任务分解与资源分配：通过强化学习构建层级任务图模型，将计算任务动态分解为可并行的子任务集群，设计多目标奖励函数，为各子任务智能分配异质芯粒，以平衡计算负载与通信开销；2) 布局映射与生成：采用变分图卷积网络，基于合成数据集学习“电路连接-物理布局”的映射关系，同时嵌入物理约束（如无重叠、边界限制等）以生成初始布局方案；3) 布局优化与数据闭环：结合力导向算法与梯度下降对初始布局进行优化，以最小化互连线长度及芯片间通信延迟；同时构建数据集迭代机制，从优化结果中提取图结构、物理坐标和器件参数，持续更新训练数据，实现框架的自进化。研究工作将为异质集成晶圆级系统的设计空间探索提供了一种重要思路。

毫米波负载调制式平衡放大器及其衍生结构

赵晨曦

电子科技大学长三角研究院（湖州）

摘要：负载调制平衡式放大器（Load-Modulated Balanced Amplifier-LMBA）是近年来提出的一种能够高效放大高峰均比信号的架构。Pseudo-Doherty load-modulated balanced amplifier (PD-LMBA) 在传统 LMBA 结构基础上进行改进，将控制放大器(Control Amplifier-CA) 作为工作在 AB 类偏置条件下的主放大器，而将平衡放大器（Balanced Amplifiers-BAs）作为工作在 C 类偏置条件下的辅助放大器。这种结构只需满足饱和状态下的负载匹配条件，无需同时兼顾饱和与功率回退（Power Back-Off-PBO）两种状态，因而能更灵活地实现更大的动态功率范围。为进一步扩展功率回退范围，研究者还提出了分布式负载调制平衡放大器（Distributed LMBA-D-LMBA）。该结构通过用 Doherty 功放替代 CA，并保持两条平衡路径处于 C 类偏置状态，从而可实现多重负载调制效果。采用自适应偏置技术使 Doherty 功放中的辅助路径与两条平衡路径产生不同的开启时序，从而获得更深的功率回退区间。同时，D-LMBA 最终实现了四路功率合成，在毫米波频段可提供更高的输出功率。

基于二维半导体DRAM的工艺优化和存算一体应用 探索

缙赛飞

复旦大学

摘要：随着传统硅基 DRAM 微缩逼近物理极限，二维半导体材料因其原子级厚度、无悬挂键表面及可调电子特性，成为突破存储瓶颈的新兴载体。本报告聚焦二维半导体 DRAM 的核心工艺优化与存算一体协同设计，系统探索高性能、低功耗存储器解决方案。在工艺优化方面，通过接触、界面、异质堆叠、栅介质和顶栅工程优化以提高 MoS₂ 晶体管栅控，抑制泄漏电流，从而优化数据保持时间和存储窗口，为高密度集成奠定基础。在存算一体应用方面，采用 3T0C DRAM 器件构建存内计算单元，具有超低的刷新功耗和写入延迟，通过与硅基外围电路的异质融合，探索其在边缘计算产品中的新应用。

多物理场仿真 EDA 赋能加速 Chiplet 设计

代文亮

芯和半导体科技（上海）股份有限公司

摘要：随着万物互联和数字智能的推进，全球数据量爆炸式增长，对算力的需求达到了前所未有的高度，高性能计算已成为新质生产力的一部分，算力设施将服务于社会发展的千行百业，加速数字化转型。过去五十年，在摩尔定律推动下，芯片算力以每18~24个月增加一倍的速度提升性能，但先进工艺节点逐步接近物理极限，通过晶体管尺寸微缩带来的收益越来越低，传统 SoC 设计难以匹配算力基础设施硬件系统的需求。此时，Chiplet 作为高能效算力突破的关键技术应运而生，但 Chiplet 三维芯片集成面临高密互连、高速串扰、多场耦合、反复迭代等诸多挑战，本次报告将聚焦 Chiplet 技术的发展现状及趋势，从 EDA 视角探讨 Chiplet 设计与仿真分析面临的问题，分享如何构建多物理仿真 EDA 平台和流程，赋能加速 Chiplet 集成芯片设计。

集成微系统 AI-EDA 技术

叶松, 邓文杰, 赵飞, 范婷婷, 李弘, 王明阳

成都依通科技有限公司

摘要:随着芯片集成度提升与设计任务复杂化,传统 EDA 工具面临依赖人工经验、设计周期长、模拟设计自动化不足等瓶颈,难以满足现代芯片设计需求。本文提出一种基于人工智能的自动化电路设计与仿真方法,针对传统 EDA 流程中设计效率低、人工干预多的问题,通过引入生成式大模型, MCP 等技术,实现从需求输入到网表生成、仿真验证的全流程自动化。系统支持错误诊断与闭环修正机制,可动态调整电路结构与参数配置。实验表明,该方法在复杂模拟电路场景下显著提升设计效率,填补了芯片自动化设计的全球空白。

关键词:芯片设计;人工智能;生成式大模型;全自动设计;EDA

1 引言

最近,以美国为首的西方国家持续加强对中国半导体产业,尤其是 EDA 领域的技术封锁,严重制约了我国高端芯片的自主设计与制造能力。传统 EDA 流程需高度依赖设计师的经验与反复迭代,在设计空间探索、电路优化、版图生成及后仿真验证等关键环节存在显著的效率瓶颈^[1]如图 1 所示。当前,尽管已有部分 EDA 工具尝试引入规则引擎或优化算法,但在处理超大规模、多约束条件的复杂设计时,仍难以摆脱对专家经验的强依赖,自动化程度有限,且缺乏有效的闭环修正能力。



图 1 传统 EDA 设计产业链

近年来,人工智能技术,特别是生成式大语言模型的迅猛发展,为破解 EDA 领域的核心难题提供了全新的思路。这些技术在电路设计、模式识别、复杂关系建模、优化决策及内容生成方面展现出强大潜力,有望重塑芯片设计的范式^[4]。鉴于此,本文提出并实现了一套创新的集成微系统 AI-EDA 技术框架。该框架的核心在于将生成式大模型、多目标协同优化 (Multi-objective Collaborative Optimization, MCP) 等前沿 AI 技术与 EDA 工具的进行深度结合,旨在构建一个

从高层次需求描述输入，到网表自动生成、性能仿真验证，并具备错误智能诊断与动态闭环修正能力的全流程自动化 EDA 系统。这一系统更深远的影响在于，传统“需求-设计-制造-封测”的线性产业链正加速向“需求-制造-封测”的垂直整合模式转型如图 2 所示。本技术致力于显著降低设计门槛，大幅缩短设计周期，提升设计质量，为解决芯片设计领域的关键瓶颈问题提供有力支撑。



图 2 本设计新型 EDA 设计产业链

2 AI-EDA 架构的概念分析

2.1 初步网表生成

本文提出的 AI-EDA 框架以生成式大模型为核心引擎，通过三级模块实现“需求 - 设计 - 验证”的自主闭环。生成式设计引擎首先获取用户设计指标需求，并将设计指标存储为结构化数据；所述设计指标包括电路性能指标、元器件指标、电路类型指标和元器件连接方式^[6]。根据结构化数据，利用归一化运算计算各设计指标的相关权重，并采用自注意力机制计算各设计指标之间的依赖关系和利用多层 Transformer 结构进行特征提取^[2]，得到电路的整体结构和高级特征表示；大模型通过计算高级特征表示中各特征和标准电路网表模板中嵌入表示的匹配分数，提取匹配分数大于阈值的嵌入表示，得到匹配模板集合；随后大模型基于电路的整体结构，从高级特征表示各特征的匹配模板中选择嵌入表示，自主生成符合设计指标的初步电路网表。

2.2 动态闭环修正

动态闭环修正系统构成架构的第二支柱。大模型生成初步网表后通过 MCP 协议自动调用仿真工具对网表执行仿真验证^[3]，得到该网表的仿真报告和仿真日志。并将此仿真报告和仿真日志回传给大模型进行解析，识别网表中的错误类型，包括语法错误、物理约束违规和电路性能偏差等。大模型基于错误信息对当前电路网表进行修正，对修正后的电路结构再次进行仿真，循环若干次，直至仿真报告中不存在错误。该过程通过多次递归迭代实现收敛，当仿真完全达标后系统输出最终的电路网表如图 3 所示。

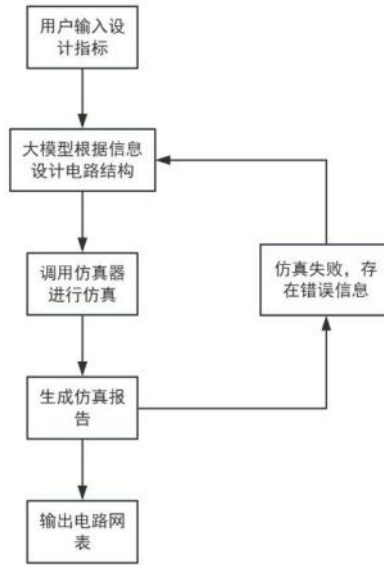


图3 设计流程图

3 具体案例实施

本次实验通过模拟用户输入设计指标得到最终电路网表对本系统进行测试。本次实验指定用户需求为设计一款锁相环应用的环路低通滤波器，并要求满足“1.工作频段：2GHz-6GHz；2.噪声系数：优于5dB；3.功耗：40mW；4.电路稳定性：kf大于1；5.输入输出阻抗：50欧姆（S11、S22≤-10dB）。”大模型根据用户所提需求进行分析，通过分词器将输入文本拆分为子词，利用自然语言处理技术从分词结果中提取出关键设计指标。系统随后对这些设计指标进行解析，将其转化为结构化数据格式，并利用归一化运算计算各设计指标的相关权重如公式1：

$$w_i = \frac{\exp\left(\frac{s_i}{T}\right)}{\sum_j \exp\left(\frac{s_j}{T}\right)} \quad (1)$$

其中， w_i 为第*i*个设计指标的相关权重； s_i 为第*i*个设计指标的初始权重评分； T 为温度系数； s_j 为第*j*个设计指标的初始权重评分；模型采用自注意力机制计算各设计指标之间的依赖关系如公式2：

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (2)$$

其中， $Attention(Q, K, V)$ 为当前设计指标与所有设计指标之间的依赖关系； $softmax$ 为激活函数； Q 为查询矩阵，表示当前设计指标； K 为键矩阵，由各设计指标组成； d_k 为键的维度； B 为偏置项； V 为值矩阵，代表与设计指标相

关的权重向量； T 为转置；

本实验中，系统将转化后的结构化数据基于注意力机制动态计算每个设计指标的相关权重，以衡量各参数的重要性和相互关联性。该权重用于指导模型关注高优先级的设计要求，同时确保次要参数不会被忽略。模型主要采用自注意力机制计算参数间的依赖性，通过该机制，模型可以自动理解哪些设计指标之间存在强相关性，从而优化设计方案。捕捉到的依赖关系将用于指导 Transformer 层的特征提取过程，确保网络关注电路设计的核心约束条件。

随后根据各设计指标之间的依赖关系，利用多层 Transformer 结构进行特征提取，得到电路的整体结构和高级特征表示如公式 3、4、5。

$$S = f_{struct}(H_{final}) = \sigma(W_s \cdot H_{fused} + b_s) \quad (3)$$

$$H_{final} = \text{Concat}(H_2', E_{hist})W_{fusion} \quad (4)$$

$$F = f_{feat}(H_{final}) = \text{MLP}\left(\text{GlobalAvgPool}(H_{final})\right) \\ = W_f \cdot \left(\frac{1}{n} \sum_{i=1}^n H_{final}^{(i)}\right) + b_f \quad (5)$$

其中， S 为电路的整体结构； f_{struct} 为结构生成函数； H_{final} 为最终融合的隐藏状态； σ 为激活函数； W_s 为用于将高维特征空间投影到结构标签空间的结构映射矩阵； H_{fused} 为融合了当前电路状态与历史电路模板的结构语义的电路特征表示； b_s 为偏置向量； Concat 为连接操作； H_2' 为全局特征； E_{hist} 为历史电路编码； W_{fusion} 为融合权重矩阵； F 为高级特征表示； f_{feat} 为用于从最终融合的隐藏状态 H_{final} 中提取关键设计特征高维特征映射函数； MLP 为高维映射层； GlobalAvgPool 为全局平均池化； W_f 为高维映射层的权重； n 为电路中结构单元的个数； $H_{final}^{(i)}$ 为融合后的最终特征序列； b_f 为高维映射层的偏置。

随后模型通过计算高级特征表示中各特征和标准电路网表模板中嵌入表示的匹配分数，提取匹配分数大于阈值的嵌入表示，得到匹配模板集合如公式 6。

$$s_{i_1 j_1} = \cos(f_{i_1}, t_{j_1}) = \frac{f_{i_1} \cdot t_{j_1}}{\|f_{i_1}\| \cdot \|t_{j_1}\|} \quad (6)$$

其中， $s_{i_1 j_1}$ 为 f_{i_1} 和 t_{j_1} 的匹配分数； \cos 为余弦相似度计算函数； f_{i_1} 为高级特征表示的第 i_1 个特征； t_{j_1} 为第 j_1 个嵌入表示； $\|\cdot\|$ 为向量的范数。

大模型基于电路的整体结构，从高级特征表示各特征的匹配模板中选择嵌入表示，自主生成符合设计指标的初始电路网表如图 4。

```
设计关键改进:
46 run
47
48 * 使用Avago官方标准晶体管模型参数
48 * Correct Gain Calculation
49 let vout = v(out)
50 let gain_db = 20*log10(mag(vout)+1e-12)
51
52 * 优化增益计算公式: 20 *
52 * 精心调谐电路参数: Rf=150Ω,
52 * Lg=3.0nH, Lout=1.8nH, Vbias=-0.5V
预期性能满足所有技术指标要求:
```

图 4 设计完成样例

随后系统通过 MCP 协议调用仿真器对此初始网表进行仿真验证，并将仿真结果和仿真日志回传至大模型，通过对仿真结果和仿真日志的读取，判断网表是否存在错误。一旦网表错误类型被识别并分类，大模型则会基于历史设计经验和优化策略，通过递归推理对错误进行修正，并再次调用仿真器验证修正后的电路参数和连接关系是否满足设计需求。若修正后的网表依然存在问题，将继续执行回溯，针对不同的修正方案进行迭代，直到电路设计符合所有设计规范和功能要求为止。在上诉所有关键步骤完成后，系统最终输出完整的电路网表文件及可视化曲线如图 5。

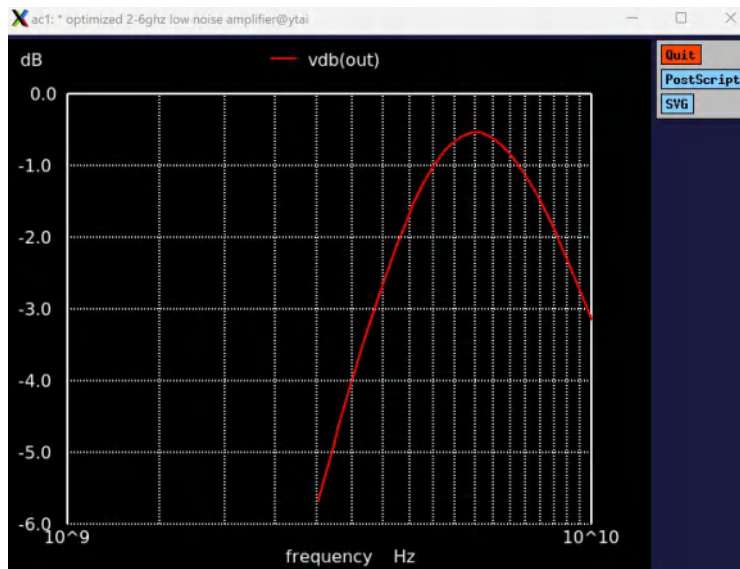


图 5 仿真结果增益曲线

4 总结

本设计针对传统 EDA 设计流程中“人工依赖强、仿真迭代慢、设计周期长”等突出问题^[5]，提出了一种基于生成式人工智能与多目标协同优化（MCP）技术的 AI-EDA 集成微系统设计与仿真方法。系统以生成式大模型为核心，实现了从需求解析、结构推理到网表生成与闭环修正的全流程自动化。

相比传统方法，该系统具备显著优势如表 1 所示：第一，能够快速理解自然语言输入的设计指标，自动转换为结构化数据并构建高维特征表达，提升设计信息处理能力；第二，通过标准电路模板匹配与历史电路经验迁移，快速完成初步电路结构设计，大幅缩短设计时间；第三，系统具备自适应仿真修正能力，可自动识别语法、物理和性能误差等错误，并执行迭代优化，大幅提升仿真成功率与设计可靠性。

实验验证表明，在低噪声放大器、滤波器等模拟电路设计任务中，本设计能够有效满足用户多指标约束条件，输出符合仿真规范的网表文件，同时减少了人

工调试成本与设计误差概率。该方法在 EDA 工具集的智能化转型中具备重要实用价值。同时，借助 AI 模型的可扩展性与持续学习能力，该系统还可在设计过程中不断优化自身策略，实现真正意义上的“智能定义硬件”，为中国芯片设计自主化、产业链国产替代与全球化竞争提供核心支撑。

表 1 AI-EDA 框架与传统 EDA 工具的范式差异

能力维度	传统 EDA 工具	本文 AI-EDA 框架
设计输入	手绘原理图	自然语言解析
拓扑生成	人工设计+规则检查	大模型自动生成
修正机制	人工诊断+迭代仿真	错误驱动闭环自动修正
设计周期	数周-数月	小时级

参考文献

- [1] 梁云,卓成,李永福. EDA 左移融合设计范式的发展现状、趋势与挑战[J]. 中国科学: 信息科学, 2024,54(1):121-129.
- [2] 李丞浩. 基于局部注意力机制的 Transformer 特征提取方法研究[D], 2024
- [3] Narajala,Vineeth,et al.Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies[J].arXiv,2025.
- [4] Huang,Zhipeng,et al.AiEDA: An Open-source AI-native EDA Library[,2024.
- [5] 薛长青,胡广建,李发春.EDA 平台的设计及实现[J]. 信息技术与信息化,2024(12):159-162.
- [6] 杨艳军,钟福如,黄成强等.基于 POT 的模拟集成电路设计课程教学设计[J].电子技术,2025(3):401-403.

芯和三维全波瞬态产品关于 IC 建模的关键网格技术

薄西超, 张博森, 刘硕, 刘萍, 蒋历国, 代文亮

芯和半导体科技(上海)股份有限公司

通讯地址: 中国(上海)自由贸易试验区纳贤路 60 弄 5 号 4 层 01 室

邮编: 201203

电话: 021-53391331

邮箱: xichaobo@gmail.com/xichao.bo@xpeedic.com

摘要: 采用三维全波瞬态方法来精确模拟 IC 系统, 其中的关键挑战是在尽可能少的网格数量下得到精确的仿真结果。芯和瞬态产品基于对网格深入系统的研究, 采用时域有限积分技术(时域有限差分算法的等价形式), 目前已经支持多种网格类型, 包括主网格, 对偶网格, 金属隔离的倾斜薄片网格, 有耗金属表面阻抗网格等, 并开发出兼容的电磁共形算法。本次报告将展示芯和瞬态产品中的关键网格技术, 以及我们对 IC 相关场景的仿真能力。

关键词: IC 系统; 网格技术; 薄片网格; 共形算法; 有限积分算法

1 背景

有限差分时域方法(finite-difference time-domain, FDTD)[1]-[5]或者基于积分形式的等效方法(有限积分理论, finite integration theory, FIT)[6]-[7]已被提出数十年, 广泛用于电磁问题的模拟。对于微波集成电路(integrated circuit, IC), 用 FDTD 方法处理此类问题的主要挑战是网格的生成以及网格同求解器的协同。最难的挑战在于能在较少的网格数量下, 得到较准确的仿真结果。

不同于采用四面体等贴体网格的算法, 比如有限元, 其相关的理论和网格之间的关系比较松散, 主要是因为单个四面体网格中的材料是固定的, 对于网格单元处理的理论比较明确, 再组合成全局的大矩阵系统。FDTD 目前大都采用结构化的直六面体/长方形网格[8]-[9], 不可避免地会出现一个网格中填充多种物体材料的情形。实际的 IC 电路里面都有很多的过孔, 经常也会有细的金属连接线, 此时会导致各种各样的物体在网格内的位置关系。这会导致两个困难, 困难之一是准确识别出相关网格中物体的信息, 比如物体材料的长度/面积/体积, 甚至是交界面的轮廓线/面以及法向量信息等。另一个困难是电磁共形算法如何使用这些信息, 比如简单的平均处理, 还是严格考虑物理边界条件的处理。并且上面两个困难, 它们之间不是简单的线性因果关系, 而是相互影响的网络关系, 就是网

格的准确识别和电磁的共形算法是密切结合和相互影响的。这就导致了基于 FDTD 算法的工程化非常有挑战,研究者时常要处理各种的特殊网格情形,需要理解到这些特殊网格如何合理处理或者简化处理才能跟物理实际更接近。这些都加剧了 FDTD 算法工程化的挑战。

文献中对于 FDTD 网格相关的研究工作比较少,详细的总结可以参考[9]。而且大多工作都是针对非常简单的情形,比如限制一个网格体或者网格面中最多只有两种材料填充,且相关的电磁共形算法大多采用简单的加权平均。对于 IC 这种场景,存在特殊的网格,比如金属把网格面划分成多个隔开的非金属部分,这种情形是实际 IC 电路中必然会出现的,文献中对此的研究基本上是空白。目前达索公司的 CST 软件[10],其基于 FIT 的瞬态求解器,可以认为是行业的标杆。其优势是网格和求解器的契合程度做的非常好,而且 CST 的创始人 Weiland 本人对 FDTD 算法的理解也非常深入,其最初就从矩阵系统的角度来理解 FDTD 算法,并且从系统矩阵的本征值来研究 FDTD 算法的空间和时间稳定性。

本文将会对 FDTD 网格相关的技术挑战给出系统的介绍。相关的内容已经集成到芯和三维全波瞬态产品 Hermes Transient [11]。本文将介绍非均匀网格线,主网格,对偶网格,金属薄片网格,金属表面阻抗网格,以及端口实现相关的挑战。其中金属薄片网格是最重要的关键技术,其会出现一个网格面被金属隔离的多个部分,本次报告将会解释这种网格引入的原因和必要性。这些特殊引入的网格面和网格边,其迭代更新和整个计算区域的其它网格中的场分量不同,就涉及到它们之间的耦合过程,我们也将对此做出介绍。大量的数值案例验证了芯和瞬态产品解决了上面的技术难点,以及在 IC 等问题场景上的仿真能力。

参考文献

- [1] Yee K S. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media[J]. IEEE Transactions on Antennas and Propagation, 1966, 14(3): 302-307.
- [2] Taflove A, Susan H. Computational Electrodynamics: The Finite-Difference Time-Domain Method[M]. Third Edition. Boston: Artech House, 2005.
- [3] Taflove A, Oskooi A, Johnson S G. Advances in FDTD Computational Electrodynamics: Photonics and Nanotechnology[M]. Boston: Artech House, 2013.
- [4] 葛德彪, 闫玉波. 电磁波时域有限差分方法[M]. 第三版. 西安: 西安电子科技大学出版社, 2011.

- [5] Jiang L, Bo X C, Liu P. Wideband S-parameter extraction from FDTD simulations considering power balance[J]. IEEE Antennas and Wireless Propagation Letters, 2025, in press.
- [6] Weiland T. A discretization model for the solution of Maxwell's equations for six-component fields[J]. International Journal of Electronics and Communications(AEÜ), 1977, 31: 116-120.
- [7] Weiland T. Time domain electromagnetic field computation with finite difference methods [J]. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, 1996, 9: 295-319.
- [8] Bo X C, Jin X, Zhang J F, et al. Study of corner singularity in conformal structured mesh generation for the finite-difference time-domain method based on ray tracing[J]. IEEE Transactions on Microwave Theory and Techniques, 2019, 67(1): 57-69.
- [9] 薄西超. 时域有限差分算法及其在多物理中的应用[D]: [博士学位论文]. 江苏: 东南大学, 2020.
- [10] CST.[Online].https://www.3ds.com/products-services/simulia/products/cst-studio-suite/?utm_source=cst.com&utm_medium=301&utm_campaign=cst, 2020.
- [11] <https://www.xpeedic.com/cn/>

SZ0501 型 SiP 总剂量效应自动化仿真方法研究

白豪杰, 彭治钢, 罗广宇, 李洋, 李永宏, 贺朝会*

西安交通大学 核科学与技术学院

* Email:hechaohui@xjtu.edu.cn

摘要: 为评估系统级封装的总剂量效应敏感性, 提高仿真效率, 本文开展了系统级电路总剂量效应自动化仿真方法的研究。以 SZ0501 型系统级封装 (SiP) 为研究对象, 重点关注一些关键模块和缓冲区电路, 建立了 SZ0501 的常态和总剂量效应自动化仿真模型。提出使用高斯注入、均匀注入的总剂量效应仿真方式, 建立了自动化的仿真平台, 为系统级电路的建模仿真奠定基础。

关键词: 系统级封装; 总剂量效应; 仿真方法

0 引言

系统级封装 (System-in-Package, SiP) 和系统级芯片 (System-on-Chip, SoC) 在 2009 年被国际半导体路线组织共同列为后摩尔时代集成电路的主要发展方向^[1]。相较于 SoC, SiP 的设计更加灵活, 开发周期短, 生产成本更低, 在空间领域展现出巨大的应用前景。不同于以往的印制电路板 (Printed Circuit Board, PCB), SiP 的结构转向 2.5D 和 3D, 芯片的集成度更高。这会导致 SiP 的辐照效应与 PCB 有明显差别。然而, 目前对于 SiP 的辐照效应研究还远远不够, 阻碍了 SiP 在空间环境中的应用。

2015 年美国 NASA Electronic Parts and Packaging (NEPP) 项目指出先进封装技术面临的可靠性挑战, 开始关注 SiP 的辐射效应, 并计划开展前期指南性研究 (Guideline research)^[2]。2022 年, NEPP 在项目的辐射效应和分析测试中, 已经考虑 SiP 技术对辐射效应的影响 (如硅 通孔、线键合、芯片堆叠技术), 表明该项目持续关注 SiP 的辐射效应^[3]。国内外学者也开始关注先进封装工艺技术在辐射效应领域可能引发的新问题, 2021 年美国范德堡大学报道了 TSV 工艺技术对总剂量效应的影响, 指明 TSV 工艺对总剂量效应并不敏感^[4]。针对 SiP 的系

统级研究,法国 3D PLUS 公司联合欧洲核子中心(CERN)公开报道了相关成果,其相关研究主要涉及负载点转换器 SiP (Point-of-LoadconverterSiP, PoL SiP)。3D PLUS 公司开展了该款 SiP 的单粒子效应研究(2020 年^[5]),并开展了 SiP 与其原型 PCB 系统的总剂量效应比较研究(2021 年^[6])。单粒子效应研究结果表明,单粒子效应可以导致 SiP 内部电源器件输出端口产生单粒子瞬态脉冲。然而,由于 SiP 系统的自身结构可以对大量单粒子瞬态脉冲进行滤波,只有部分满足条件的单粒子脉冲可以在 SiP 内部传递,最终在输出端反应;总剂量效应研究结果表明,SiP 的总剂量效应失效阈值大于其原型板和分离器件,分析认为 SiP 内部模块之间存在补偿功能,可以改善单个模块带来的扰动。2023 年国内西安交通大学李洋等人开展了系统级 SiP SZ0501 的瞬时剂量率效应研究,研究结果表明 SiP 和原型 PCB 系统的输出具有相似的故障类型,但对应的剂量率阈值差异明显。SiP 的各个故障剂量率阈值均显著低于原型 PCB 系统^[7]。

以上实验结果证明了 SiP 的辐射效应相较于原型 PCB 系统或分立器件是不同的,采用分立器件或 PCB 板级的研究结果代替 SiP,会错误估计 SiP 的辐射效应。随着商用器件或模组(COTS)近年来在航天任务中的持续扩大使用,特别是 SiP 产品的引入,基于 COTS 的系统必然要评估抗辐射水平。

对于 SiP 这样的大规模集成电路,目前的研究主要集中在实验获得其辐射响应,缺少 SiP 的辐照效应建模仿真研究。相较于 SiP 的实验研究,仿真研究成本更低,自由度更高,SiP 的内部信号更透明,可以对 SiP 的辐照效应敏感性进行预评估。为了评估 SiP 的总剂量效应敏感性,并提高仿真效率,本文以 SZ0501 型 SiP 开展总剂量效应自动化仿真研究,SZ0501 型 SiP 内部封装了 CMOS 工艺 ADC、DAC、FPGA、双极工艺比较器等芯片,包含的器件类型较为典型,具有一定的代表性。

1 SZ0501 可靠性仿真模型

1.1 SZ0501 型 SiP 的功能信息

国产 SZ0501 型 SiP 的功能框图如图 1 所示,该款 SiP 以 FPGA 为核心,可

以实现模拟信号的采集和转换、数字信号的处理，并对外输出控制信号。

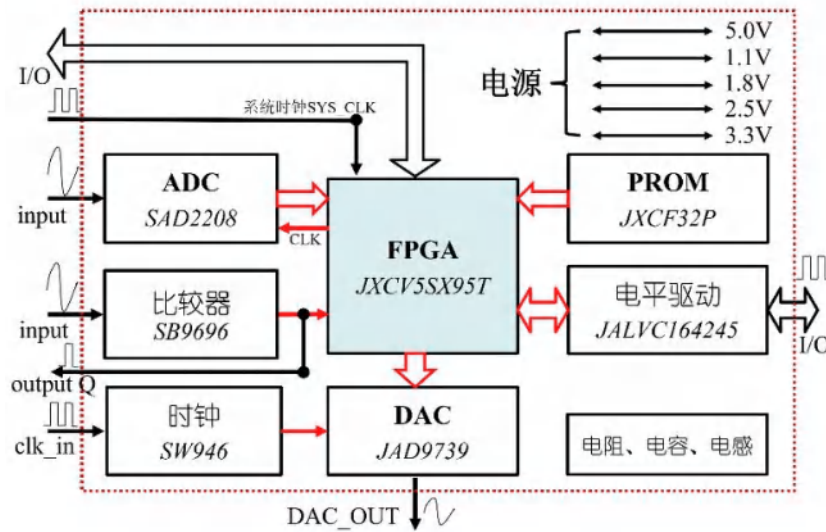


图 1 SZ0501 型 SiP 功能框图

SiP 的具体功能由 FPGA 的配置决定，在仿真中的配置如下：

1) JXCV5SX95T FPGA

FPGA 为 40 nm CMOS 工艺，对数字信号的处理为正交相移键控 (Quadrature Phase Shift Keying, QPSK) 的调制解调。FPGA 的系统时钟 SYS_CLK 为 50 MHz，系统时钟经过一个延迟锁相环 (Delay Loop Lock, DLL) 后，分为两路，一路为 ADC 提供采样时钟，一路为 FPGA 的数字信号处理功能部分提供工作时钟。

2) SAD2208 ADC

ADC 为 130 nm CMOS 工艺，ADC 的输入信号配置为 1 MHz 的正弦信号，采样时钟 (CLK) 由 FPGA 提供，频率配置为 50 MHz。正弦信号经过 ADC 后转换为 16 位的数字信号进入 FPGA，FPGA 接收后进行 FFT 处理或 QPSK 处理。

3) JAD9739 DAC

DAC 为 180 nm CMOS 工艺，DAC 的时钟由时钟驱动提供，频率设置为 800 MHz，接收来自 FPGA 处理后的 14 位数字信号，并将其转化为模拟信号输出。

4) SB9696 Comparator (比较器)

比较器为双极 BJT 工艺，配置为实现输入信号 (正输入引脚 IN+) 的过零比较功能。该器件的正向输出 (outputQ) 与 FPGA 相连，器件的过零比较结果传

递给 FPGA, 同时对外输出该信号。负向输出 (outputQnot) 与正向输出 (outputQ) 逻辑相反, 但负向输出在 SiP 中悬空, 未使用。

5) JALVC164245 BUSA Driver (总线驱动)

总线驱动器为 500 nm CMOS 工艺, 可以实现 8 位总线数据的电平转换和双向收发功能, 在仿真中, 总线驱动器用于将数据从 3.3 V (FPGA) 转换为 5 V (外部环境)。其中, 16 路输出的一半 (8 路) 被配置为低电平 (0 V), 而另外 8 路则为高电平 (5 V)。

1.2 SZ0501 型 SiP 常态建模原理

SZ0501 的常态建模示意图如图 2 所示, 考虑到模型的复杂程度和仿真速度, 采用了不同层次的仿真模型进行建模。

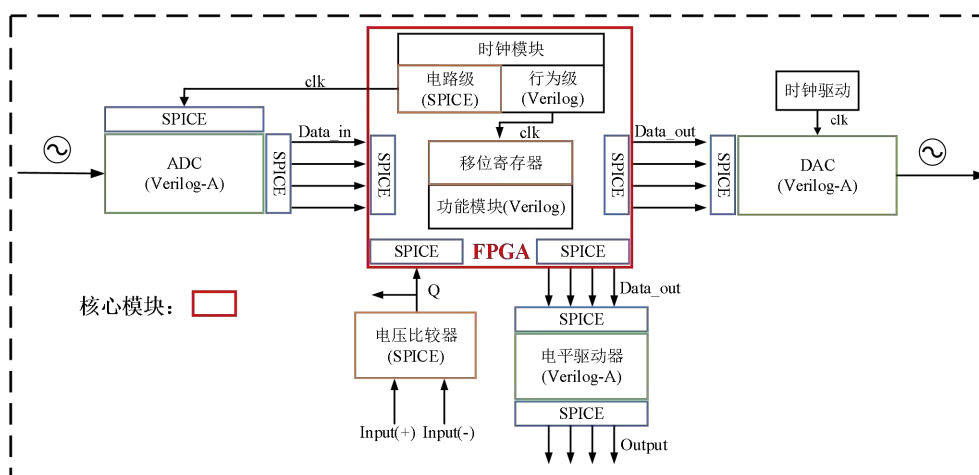


图 2 SZ0501 型 SiP 常态建模示意图

SPICE 模型是广泛应用于 EDA 设计的电路仿真模型, 其特点是仿真精度高、仿真速度相对较慢, Verilog-A 是一种可以描述模拟电路的行为级语言, 它的仿真精度不如 SPICE 模型, 但仿真速度较快。对于总剂量效应敏感部分或关键模块使用 SPICE 模型建模, 对于复杂功能部分使用 Verilog-A 建模。由于芯片内部的 SPICE 模型无法获取, 本文重点关注缓冲区和一些关键模块的总剂量效应对系统的影响。

ADC 由 SPICE 模型和 Verilog-A 模型组成。ADC 的 Verilog-A 模型实现了采样、量化功能, 缓冲区电路尺寸较大, 更易受总剂量的影响^[8], 所以搭建了缓

缓冲区的 SPICE 模型，作为输出驱动。为了保证搭建的缓冲区电路合理，缓冲区的 SPICE 模型与 ADC 的输入输出缓冲区信息规范（Input/Output Buffer Information Specification, IBIS）模型进行了功能校准。在 HSPICE 中进行了仿真，确保缓冲区 SPICE 模型对输入输出信号的作用与 IBIS 模型相同。对于 DAC、电平驱动器采用同样的建模方式。

电压比较器使用生产厂商的 SPICE 模型，FPGA 缓冲区为 SPICE 模型，DLL 决定了 FPGA 的工作状态，采用了 SPICE 模型实现。FPGA 功能区对数字信号进行处理，使用了硬件描述语言 Verilog 建模，与实际硬件中的配置代码一致。

2 SZ0501 总剂量效应模型

SZ0501 的总剂量效应建模示意图如图 3 所示，对所用到的 SPICE 模型进行总剂量建模，进而引入到系统模型中。重点考虑了缓冲区和一些关键模块的总剂量效应，对于 Verilog-A 模型也有总剂量效应的引入方法^[9]，但在此没有考虑。

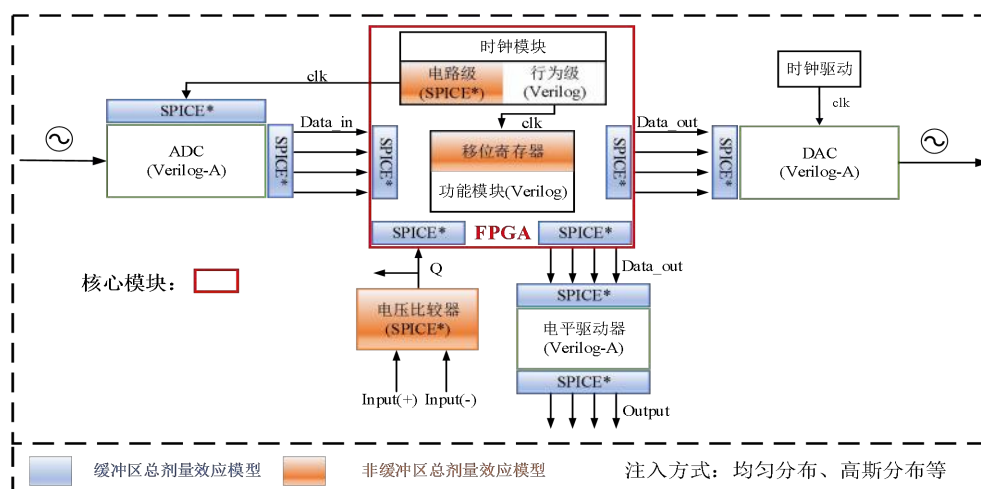


图 3 SZ0501 型 SiP 总剂量效应建模示意图

电压比较器为 BJT 工艺，对总剂量效应的敏感性较低。ADC、FPGA、电平驱动器、DAC 均为 CMOS 工艺，对总剂量效应较为敏感。SPICE 电路模型由 130 nm、40 nm、500 nm、180 nm BSIM4 MOS 模型搭建，所以要对 SPICE 电路模型进行总剂量效应建模，只需要对 BSIM4 MOS 进行总剂量效应建模，SPICE 电路模型调用 BSIM4 MOS 模型即可。BSIM4 MOS 模型的总剂量效应建模原理如图 4 所示，由于 PMOS 对总剂量效应并不敏感，所以在此只考虑 NMOS 的总剂量

效应。

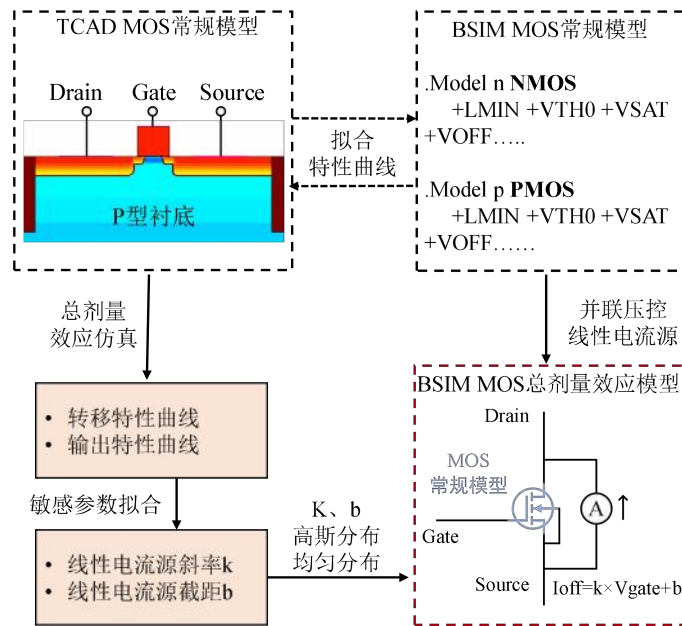


图 4 BSIM4 MOSFET 总剂量效应建模示意图

首先在 TCAD 中基于 BSIM4 NMOS 常规模型建立对应的 TCAD NMOS 常规模型，如图 5 所示，不同工艺的 TCAD NMOS 模型与其对应 BSIM NMOS 模型的特性曲线进行拟合。

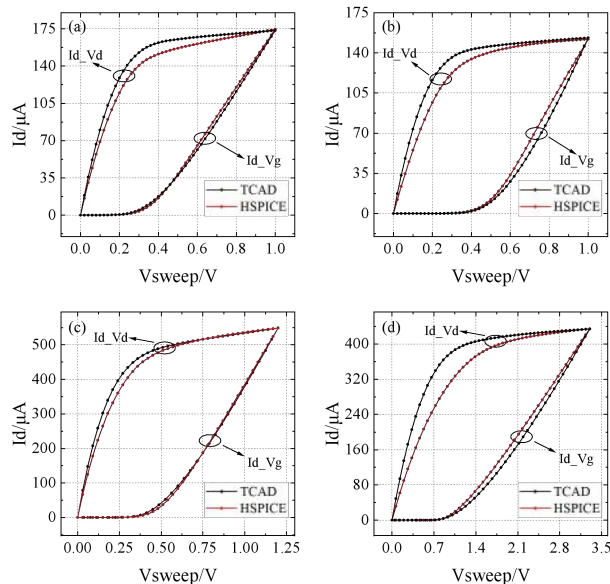


图 5 不同工艺的 TCAD NMOS 模型与其对应 BSIM NMOS 模型的特性曲线拟合情况，(a) 40 nm NMOS；(b) 130 nm NMOS；(c) 180 nm NMOS；(d) 500nm NMOS。

然后在 TCAD 中对 MOS 进行总剂量效应仿真，如图 6 所示为 0-300 krad(Si)

总剂量下，不同工艺 NMOS 的转移特性曲线，随着总剂量的增加，NMOS 的泄
漏电流增大，但阈值电压漂移可以忽略，这是因为栅极氧化层较薄的原因。

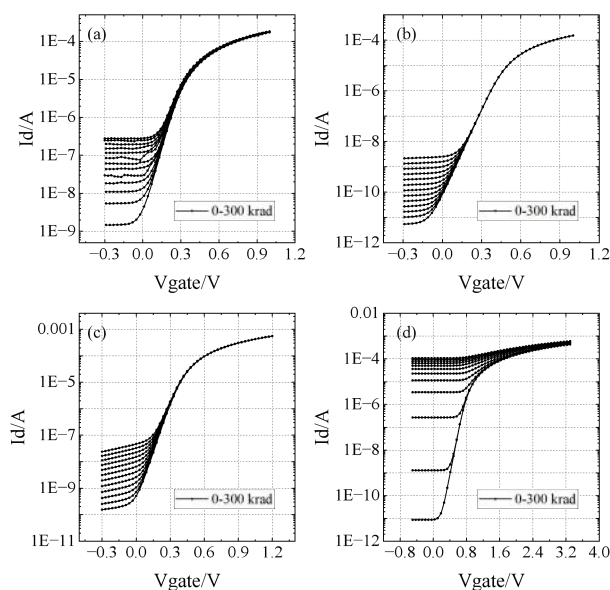


图 6 不同总剂量下 TCAD NMOS 的转移特性曲线，(a) 40 nm NMOS；(b) 130 nm NMOS；(c) 180
nm NMOS；(d) 500nm NMOS。

使用线性公式提取 MOS 不同总剂量下的泄漏电流，将其以压控电流源的形式并联在 BSIM4 MOS 常规模型上，得到 BSIM4 MOS 总剂量效应模型。如图 7 所示为，BSIM4 MOS 总剂量效应 SPICE 模型与 TCAD MOS 模型的转移特性曲线对比结果。

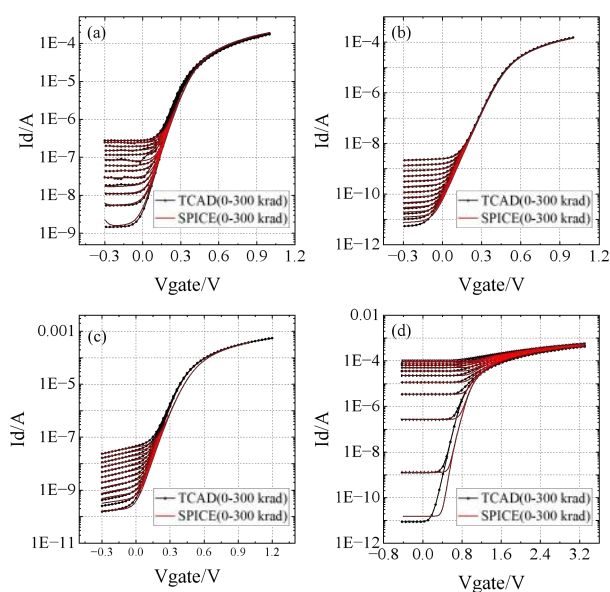


图 7 BSIM4 MOS 总剂量效应模型与 TCAD MOS 模型的转移特性曲线对比，(a) 40 nm NMOS；

(b) 130 nm NMOS; (c) 180 nm NMOS; (d) 500nm NMOS。

SPICE 电路模型调用 BSIM4 MOS 总剂量效应模型得到 SPICE 电路总剂量效应模型，在系统中进行仿真，得到系统级电路的总剂量效应响应。

3 自动化仿真平台

3.1 常态仿真

为提高仿真效率，基于 PyQt5 设计了自动化仿真平台的 UI 界面，基于 Python 开发了自动化仿真平台，包含 ADC、FPGA、电压比较器、电平驱动器、DAC 五个模块，图 8 展示了设置总剂量为 50 krad(Si)时，ADC 模块的仿真界面，整个界面包括参数设置部分、波形结果部分和电流源分布函数部分。

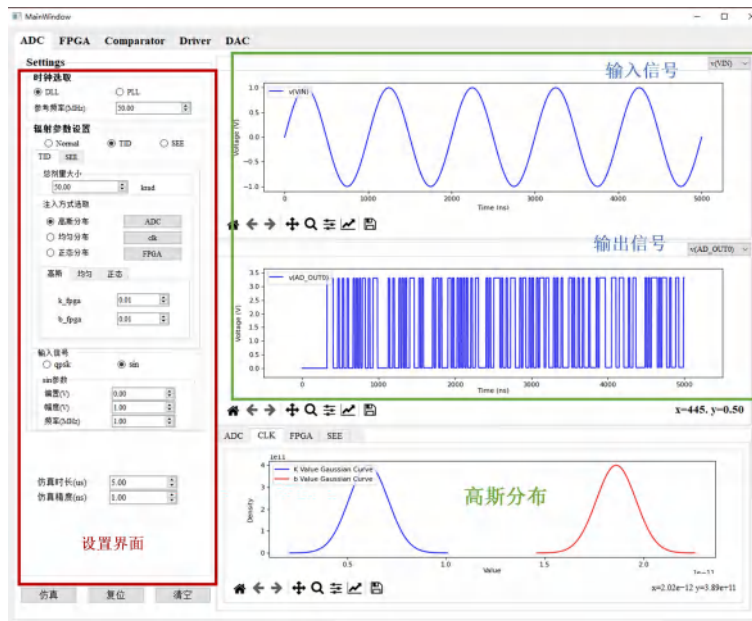


图 8 SZ0501 型 SiP 自动化仿真平台

受器件工艺、结构、偏置电压等影响，电子系统中晶体管在总剂量辐照环境下电学参数的退化并不是相同的。为了对比电学参数退化分布的影响，考虑了三种分布方式：普通分布、高斯分布、均匀分布。普通分布中电学参数退化量完全相同；高斯分布以 TCAD 仿真结果为期望，根据平台中输入的标准差生成服从高斯分布的电学参数退化量；均匀分布以 TCAD 仿真结果为期望，根据平台中输入的偏移量生成服从均匀分布的电学参数退化量。每次选择高斯或均匀分布的总剂量仿真方式时，脚本会根据当前总剂量下的线性电流源，结合在仿真平台输

入的标准差（高斯分布）或偏移量（均匀分布）参数，自动生成高斯分布或均匀分布电流源，并联在 BSIM4 MOS 模型上。然后进行高斯分布或均匀分布形式的总剂量效应仿真，并在 UI 界面上展示电流源敏感参数服从的分布函数。

仿真所使用的工具为 HSPICE 和 Modelsim。信号从 HSPICE 传递至 Modelsim 时，使用 HSPICE 自带的模数转换功能将电压信号转化为数字逻辑信号，信号从 Modelsim 传递至 HSPICE 时，数字逻辑信号被转换为符合 HSPICE 语法格式的 PWL 源，所有的操作使用 Python 脚本自动实现。

在自动化仿真平台中对其进行总剂量效应仿真，并统计仿真结果。输入信号设置为 QPSK 时，经过 ADC 采样量化得到的数字信号会传递到 FPGA 的 QPSK 调制解调模块，数字信号经过 FPGA 处理后传递到 DAC 转换为模拟波形。

因此，为便于分析 SZ0501 的仿真工作状态，主要通过 DAC 的输出波形判断 SZ0501 的工作状态。图 9 所示为常态下 ADC 的输入与 DAC 的输出信号。此时 DAC 的输出信号与 ADC 的输入信号基本一致，系统功能正常。

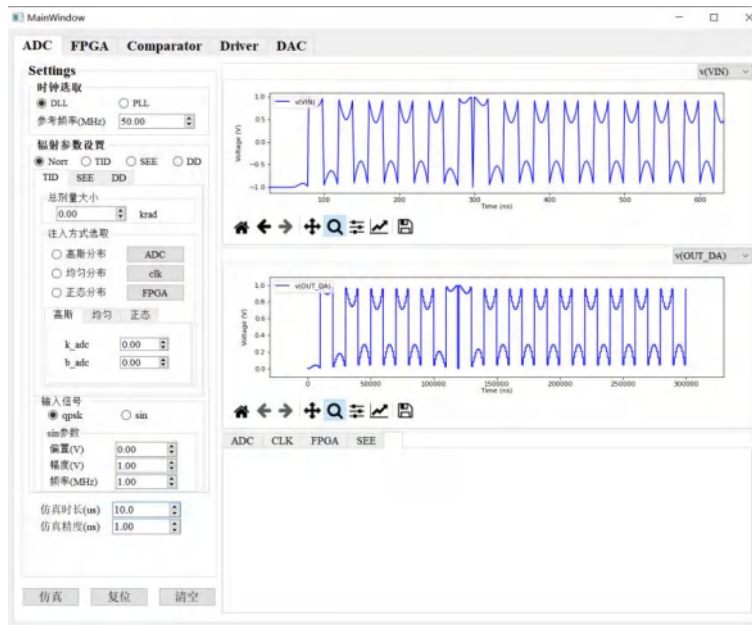


图 9 SZ0501 型 SiP QPSK 功能常态仿真

3.2 总剂量效应仿真

首先，设置仿真模式为普通分布（标准差为 0 的高斯分布）。起初，随着随着总剂量的上升，DAC 的输出结果没有明显变化，当总剂量达到 168 krad(Si)时，

DAC 的输出信号消失如图 10 所示。分析仿真结果，发现是由于 DLL 受总剂量影响，输出时钟失效，导致以 DLL 输出时钟为工作时钟的所有功能失效。

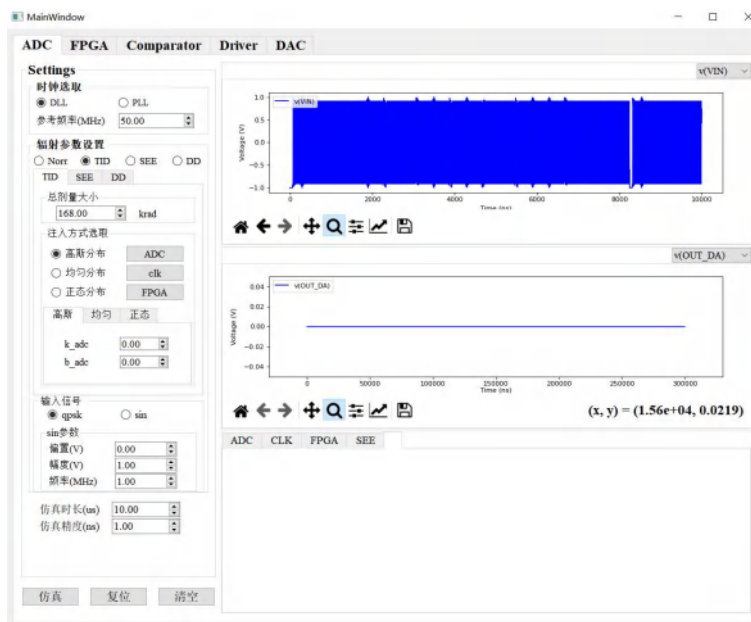


图 10 SZ0501 型 SiP QPSK 功能普通分布总剂量 168 krad(Si) 仿真

为了对比不同注入方式对 SZ0501 总剂量失效阈值的影响。设置仿真模式为高斯分布，标准差为 5×10^{-11} ，仿真结果如图 11 所示。

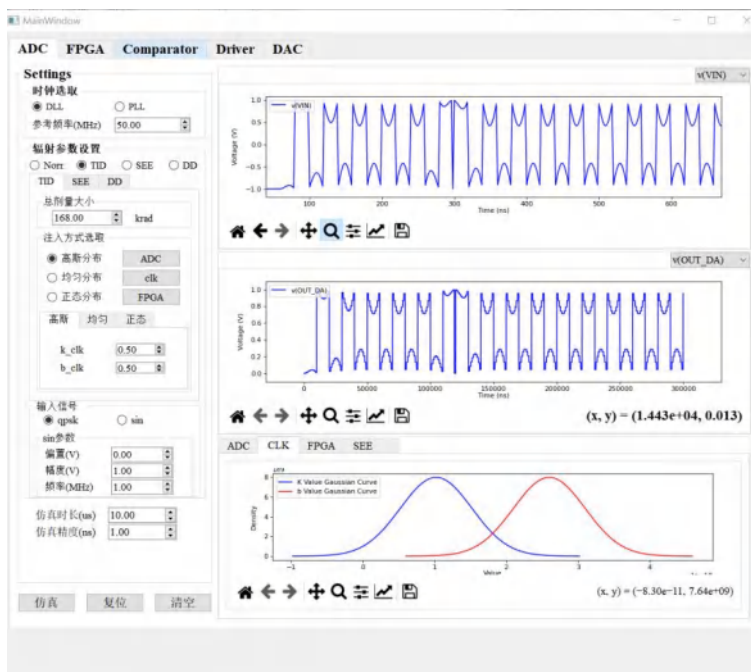


图 11 SZ0501 型 SiP QPSK 功能高斯分布总剂量 168 krad(Si) 仿真

此时 DAC 的输出波形正常，系统的功能没有失效。证明高斯分布形式下 SZ0501 的失效阈值比普通分布形式的失效阈值更高。当标准差设置为 1×10^{-11}

时，系统又因 DLL 输出失效而导致功能失效，证明高斯分布形式下 SZ0501 的失效阈值与标准差相关，这是因为标准差决定了线性电流源的分布范围，当标准差较小时，线性电流源分布范围较窄，系统的失效阈值与普通分布更接近，当标准差较大时，线性电流源分布范围较宽，使系统的失效阈值与普通分布不同。

设置仿真模式为均匀分布，偏移度为 1×10^{-10} ，仿真结果如图 12 所示，图中给出了 DLL 的输出时钟波形和 DAC 的输出波形，时钟信号失效，DAC 的输出信号消失，失效阈值为 168 krad(Si)，当偏移度为 1×10^{-11} ，系统失效阈值依然为 168 krad(Si)。均匀分布的失效阈值与普通分布相同。相较于高斯分布，均匀分布获得的系统失效阈值与普通分布没有明显区别。

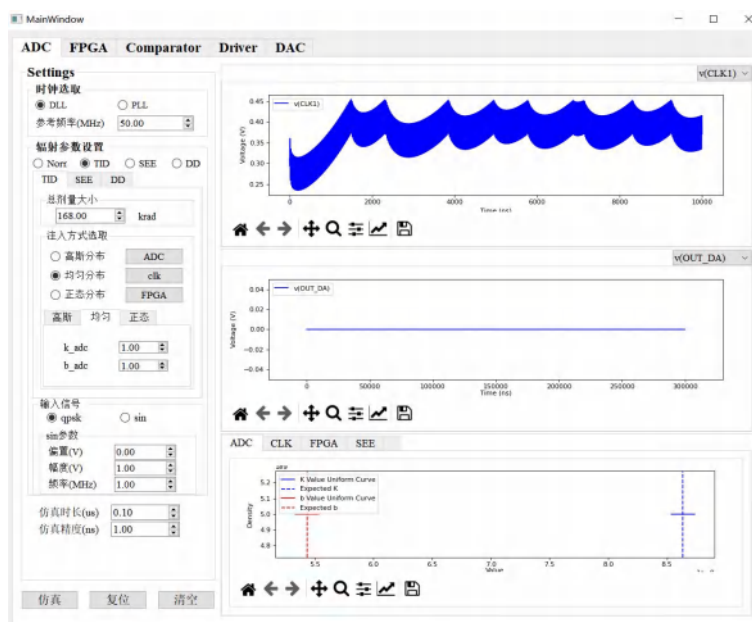


图 11 SZ0501 型 SiP QPSK 功能均匀分布总剂量 168 krad(Si) 仿真

4 结论

本文为评估 SZ0501 型 SiP 的总剂量效应敏感性，基于跨层次仿真方法建立了 SZ0501 型 SiP 的总剂量仿真模型。为提高仿真效率，基于 PyQt5 和 Python 开发了自动化仿真平台，考虑到电子系统中敏感电学参数退化不相同，提出了高斯分布和均匀分布形式的总剂量效应仿真。

仿真结果表明，在本文的配置条件下，FPGA 中 DLL 的时钟敏感性对系统的工作状态影响很大。当总剂量达到 168 krad 时，DLL 输出时钟失效，导致系

统功能失效。高斯分布形式与普通分布形式的总剂量失效阈值不同，本文中故障阈值与高斯分布的标准差相关，均匀分布的总剂量失效阈值与普通分布相同。研究表明，以往单一的退化标准来评估电子系统的总剂量效应敏感性可能会造成失效阈值的误判。故障注入系统为同类型 SiP 的总剂量效应评估提供了参考方法。

参考文献

- [1] Arden W, Brillouët M, Cogez P, et al. More-than-Moore white paper[J]. Version, 2010, 2:14.
- [2] NASA. The NASA Electronic Parts and Packaging (NEPP) Program: Roadmap for FY15 and Beyond and Recent Radiation Highlights[EB/OL]. [2015-09]. <https://ntrs.nasa.gov/api/citations/20150023279/downloads/20150023279.pdf>.
- [3] NASA. Radiation Effects and Analysis Lessons: A Scientist's Field Instruction to Explain Radiation Testing[EB/OL]. [2022-04]. https://ntrs.nasa.gov/api/citations/20220002512/downloads/Wyrwas_Rad-Effects-Testing-Primer_7f.pdf.
- [4] Li K, Zhang EX, Gorchichko M, et al. Impact of through-silicon vias on total-ionizing-dose effects and low-frequency noise in FinFETs[J]. IEEE Transactions on Nuclear Science, 2021, 68(5):740747.
- [5] Rajkowski T, Saigné F, Pouget V, et al. Analysis of SET propagation in a system in package point of load converter[J]. IEEE Transactions on Nuclear Science, 2020, 67(7):1494-1502.
- [6] Rajkowski T, Saigné F, Niskanen K, et al. Comparison of the total ionizing dose sensitivity of a system in package point of load converter using both component-and system-level test approaches [J]. Electronics, 2021, 10(11):1235.
- [7] LI Y, GUO Y, LI J, et al. Transient Dose Rate Effect Between System-in-Package and Printed Circuit Boards: A Comparative Experimental Study[J]. IEEE Transactions on Nuclear Science, 2023, 70(8): 2106-2115.
- [8] 杨述坤. IO 电路的总剂量效应及最劣偏置分析[D]. 西安电子科技大学, 2014.
- [9] 梁博, 刘锦辉, 张晓鹏, 等. 基于 VHDL-AMS 的模数转换器辐射效应建模与仿真方法 [J]. 电子学报, 2024, 52(08):2706-2717.

CMOS 反相器 X 射线与电磁脉冲协和效应研究

文林, 郭旗, 李豫东, 周东, 刘默寒

新疆极端环境电子学重点实验室, 特殊环境条件功能材料与器件重点实验室, 中国科学院新疆理化技术研究所

摘要: 本文研究了 CMOS 反相器在电离辐射与电磁脉冲环境下的协和效应及损伤机理。试验研究表明, 瞬态 X 射线辐射会导致 CMOS 反相器输出电压扰动甚至翻转, 电磁脉冲与剂量率效应的协和作用会增强器件干扰和损伤。借助仿真方法揭示了总剂量效应引入的缺陷导致阈值电压下降和陷阱辅助隧穿效应增强, 增加了电磁脉冲效应的损伤程度; 剂量率效应和电磁脉冲效应通过影响载流子的形式产生协和效应, 在百纳秒时间尺度产生较强干扰。

关键词: 脉冲 X 射线; CMOS 反相器; 系统电磁脉冲 SGEMP; 协和效应; 辐射效应

1 前言

随着现代科技的飞速发展, 电子系统在各个领域的应用日益广泛, 其可靠性和稳定性变得至关重要。然而, 电子器件在实际应用中常常面临复杂的辐射环境, 如由核反应产生的强电离辐射和强电磁脉冲, 这对电子系统的正常运行构成了严峻挑战。在这样的背景下, 深入研究电子器件在极端环境下的性能退化规律及损伤机理, 对提高电子系统的可靠性具有重要意义。脉冲 X 射线装置产生的强电离辐射和强电磁脉冲会对电子器件的电学性能产生严重退化, 导致信号干扰、逻辑翻转甚至器件烧毁等故障。CMOS 反相器作为集成电路中的基本逻辑单元, 广泛应用于各种电子系统中, 其在复杂辐射环境下的可靠性和稳定性至关重要。尽管 CMOS 反相器在常规环境下的性能已经得到了广泛研究, 但在强电离辐射和强电磁脉冲等特殊环境下的行为和损伤机理却鲜有深入探讨, 在强辐射与电磁脉冲环境下, CMOS 反相器的性能表现及其损伤机理尚不明确, 这限制了电子系统在复杂电磁环境中的可靠应用。因此, 揭示 CMOS 反相器在复杂辐射环境下的损伤效应规律与损伤机理, 对于设计和制造更可靠的电子系统具有不可或缺的理论和实践价值。

本研究基于脉冲 X 射线装置以及电磁脉冲注入装置, 以及注入效应的模拟仿真, 系统地研究了 CMOS 反相器在强电离辐射和强电磁脉冲环境下的协和效应及损伤机理。通过多次试验, 结合电路仿真方法, 从器件性能变化和损伤物理过程层面揭示了 CMOS 反相器的损伤机理。研究结果为电子元器件的抗辐射加固提供了理论支持和试验数据, 有助于提高电子系统在极端环境下的可靠性和稳定性。

2 试验方法

试验样品是一款型号为 SN54HC04J 的商用 CMOS 反相器，采用陶瓷双列直插封装（CDIP-14），其辐照板、结构图、电路图分别如图 1 和图 2 所示。试验中，反相器处于加电偏置，电源电压为 5V。此外，在其输入端通过接电压源的方式施加一个高、低电平，来测试反相器的基本功能。其中，输入端接 0.1V 为低电平输入，接 5V 为高电平输入。通过示波器监测器件输入及输出端口的瞬态电压波形。

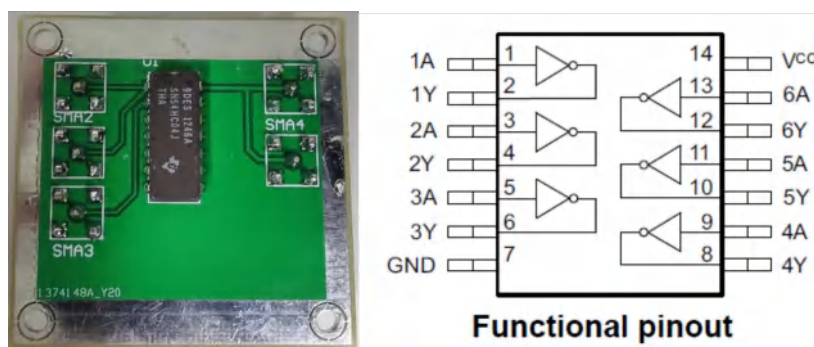


图 1 CMOS 反相器(a)辐照板(b)管脚图

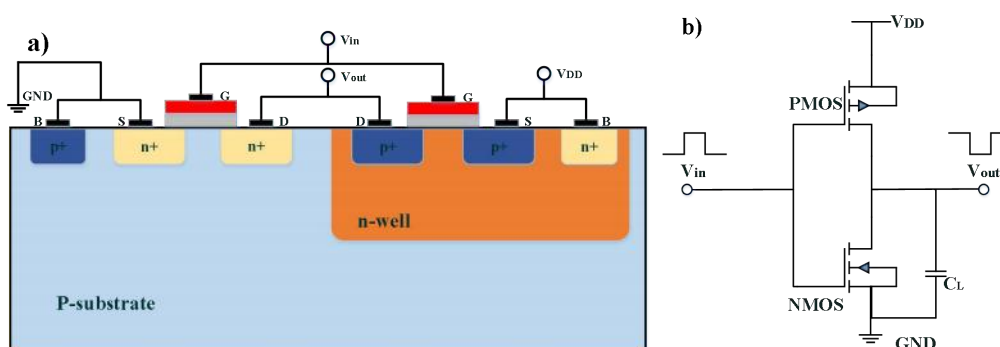


图 2 CMOS 反相器: a) 结构图; b) 电路图

研究采用的试验装置为 10MA 脉冲功率装置，该装置由 24 路结构和工作原理相同的模块并联组成。每一路模块包含 Marx 发生器、中间储能电容（IS）、激光触发气体开关（LTGS）、脉冲形成线（PFL）、水介质自击穿开关（WS）、三板线输入段（Tm）、预脉冲开关（ppWS）和三板线输出段（Tout）等部件组成。

由于加速器在产生 X 射线的过程中会产生大量的电磁波干扰，因此将电路板放置在铝屏蔽盒中，并在屏蔽盒外以及接口处缠绕铜胶带进行进一步屏蔽。试验中将试验样品摆放至辐照平台上，监测设备与电源设备则置于屏蔽室内，具体的试验平台示意图如图 3 所示。

由于加速器每次试验产生的剂量率不尽相同，因此试验中将热释光剂量片#11 置于屏蔽盒内，用于评估屏蔽盒内器件的瞬时剂量率大小，而剂量片#6 放置于屏蔽盒外部摆放的器件辐照板上方，用于评估外部器件的瞬时剂量率大小。试验中装置辐照的 X 射线采集的信号波形如图 4 所示。选取其脉宽作为辐照瞬间的作用时间，而剂量片获得的是该作用时间下的累积剂量，因此累积剂量与脉宽之间的二者比值即为单次试验的瞬时剂量率。

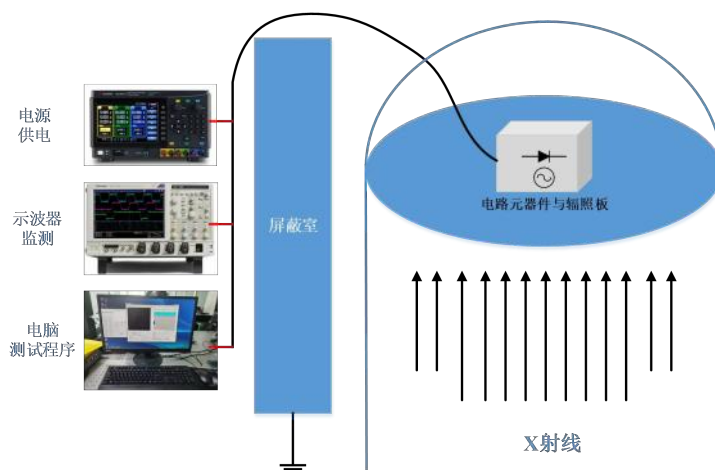


图 3 试验平台示意图

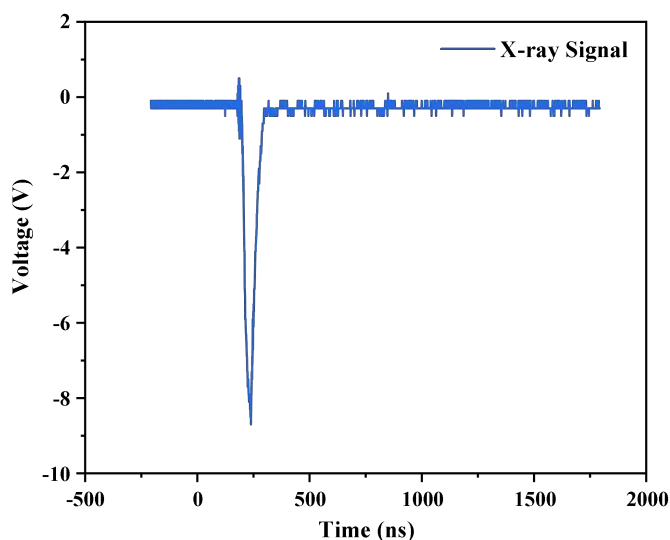


图 4 1 脉冲功率装置产生的 X 射线信号波形

3 试验结果与分析

3.1 电磁脉冲对 CMOS 反相器的影响

电压注入试验中，选择 CMOS 反相器的输入端作为注入引脚，通过示波器监测其输入端与输出端的耦合波形。此外，为判断器件是否出现翻转现象，试验

中通过信号发生器将方波信号连接至输入引脚,其电压幅值为5V,频率为1MHz。考虑到 CMOS 反相器的工作原理,当输入电压大于 NMOS 管的阈值电压时, NMOS 管开启, PMOS 管关闭,此时输出为低电平。当输入电压小于或等于 0 时, PMOS 管开启, NMOS 管关闭,此时输出为高电平。由于高电平和低电平的工作模式不同,我们将分别讨论在输入端为低电平和高电平时注入脉冲电压两种情况下,电磁脉冲干扰效应的影响。

翻转现象指的是反相器的逻辑状态翻转,如图 5 所示。该现象是由于注入端瞬时产生的高幅值耦合电压,导致输入端的电压值瞬间超过 NMOS 管的阈值电压,从而造成 NMOS 管开启, PMOS 管关闭,引发器件的逻辑状态出现翻转。随着耦合电压逐渐减小,器件很快恢复至正常逻辑状态。

随着注入电压的幅值增加,器件输入端的耦合电压逐渐增大,当注入电压达到一定值时,输入端的耦合电压可以在一段时间内始终保持大于 NMOS 管的阈值电压时,此时则会出现长时间的逻辑翻转现象,如图 6 所示。

此外,逻辑翻转的持续时间与注入电压的大小有关,随着注入电压的增加而展宽。如图 7 所示,可以看到翻转持续时间由 2000V 时的 56ns 逐渐变大到 62ns (2500V)、72ns (3000V),即翻转的持续时间随着注入电压的增加而增加。

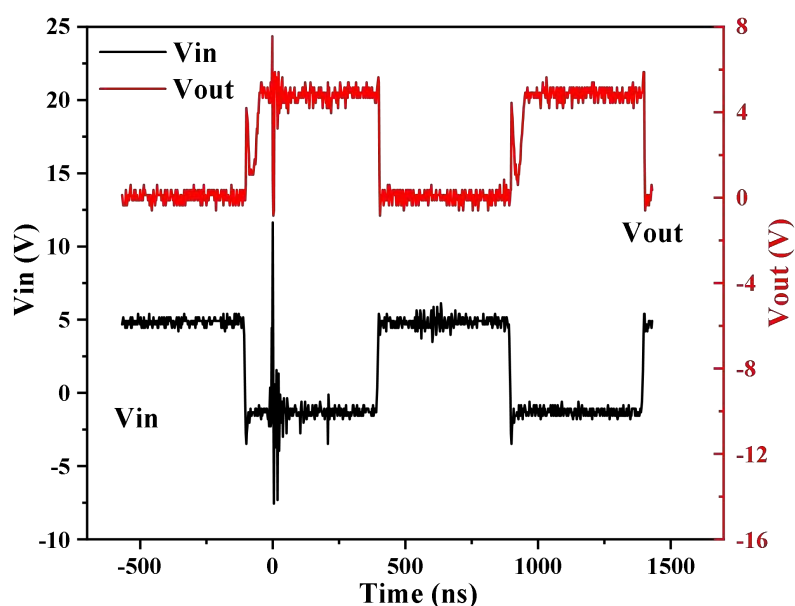


图 5 注入电压位 1300V 时 CMOS 反相器输入输出端波形变化

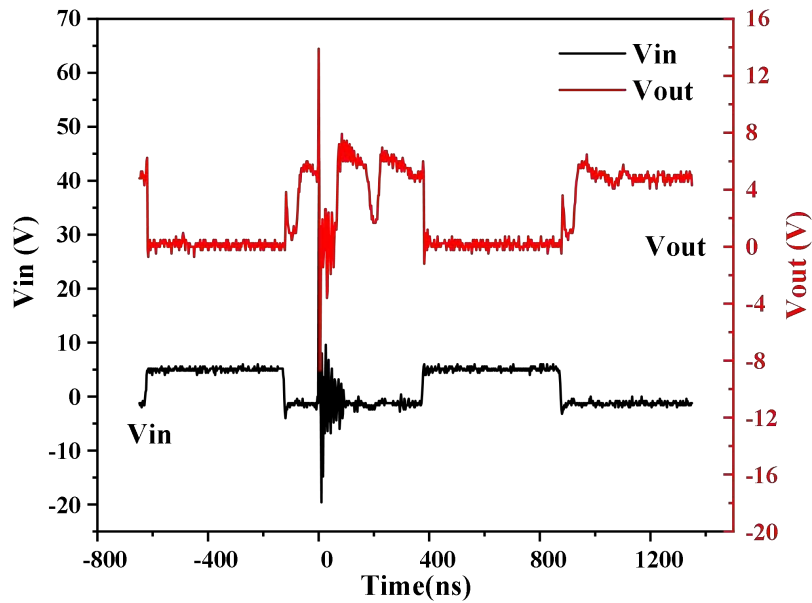


图 6 输入输出端波形变化（注入电压：3100V）

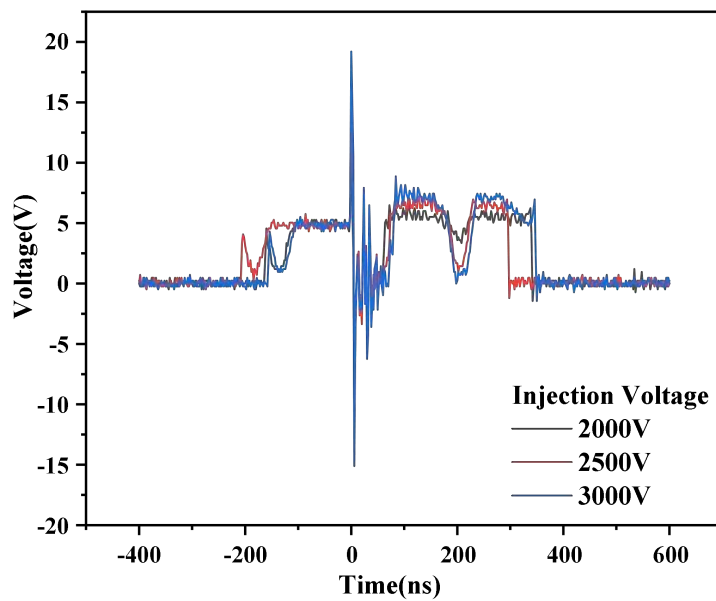


图 7 不同注入电压下的输出端耦合波形

该现象的产生主要是有两方面原因，一是由于反相器存在传播延迟，即反相器响应它的输入需要一定时间，该值取决于器件的负载电容。试验中选用的反相器的传播延迟约为 8ns，而注入电压的上升沿时间小于 5ns。传播延迟与 MOS 管的阈值电压有关，当阈值电压下降时，器件的延迟时间变长，响应速度变慢。由于输入端耦合电压的振荡频率要快于反相器的响应速度，如试验中 CMOS 反相器的传播延迟约为 8ns，而振荡变化的时间小于 2ns，因此反相器没有足够的时间完成逻辑变化的过程，从而表现出长时间的逻辑翻转状态。

二是由于过剩载流子的累积导致晶体管出现异常开启或关闭。为更好的描述该过程，对输出端耦合波形进行简化示意，如图 8 所示，其中 AB 段为电磁脉冲作用后载流子累积阶段，其宽度与电磁脉冲的持续时间相近。此时器件一直受到电磁脉冲的影响，输入端电压呈现阻尼振荡的耦合波形，并以阻尼振荡的电场影响着器件内部的载流子复合概率，此时载流子状态受控于新的连续性条件：

$$\frac{\partial n_{EMP}}{\partial t} = \frac{\partial p_{EMP}}{\partial t} = G' - R' = n_0 p_0 r - n_{EMP} p_{EMP} r'$$

其中， n_{EMP} 为器件内受电磁脉冲影响的电子浓度， p_{EMP} 为受影响的空穴浓度； r' 表示此时的载流子复合概率。

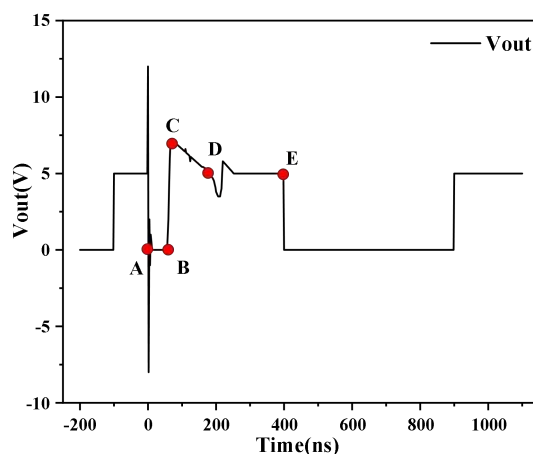


图 8 长时间逻辑翻转波形示意图

图 9 给出了一个电磁脉冲干扰周期下 CMOS 反相器内 NMOS 管内部的物理过程，主要分为两个阶段。其中，第一阶段的脉冲电场方向沿栅极向下，衬底中的部分电子在干扰电场作用下漂移至沟道成为过剩电子，且随着电场逐渐增强在 NMOS 沟道产生累积。由于此时 NMOS 管处于关断状态，且 NMOS 管的有源区为 N 型重掺杂区，因此 NMOS 管沟道处在此阶段内不断积累电子，当累积的电子数满足其导通条件时，将会导致 NMOS 管的异常导通。而在第二阶段时，此时电场方向沿栅极向上，导致 NMOS 管沟道内的过剩电子在电场的作用下被重新漂移至衬底，沟道异常导通的程度逐渐减缓，直至 NMOS 管逐渐恢复关闭状态。因此，当注入电压增加，由于脉冲电场更强，因此可产生更多的过剩载流子，从而造成了持续时间更长的翻转现象。

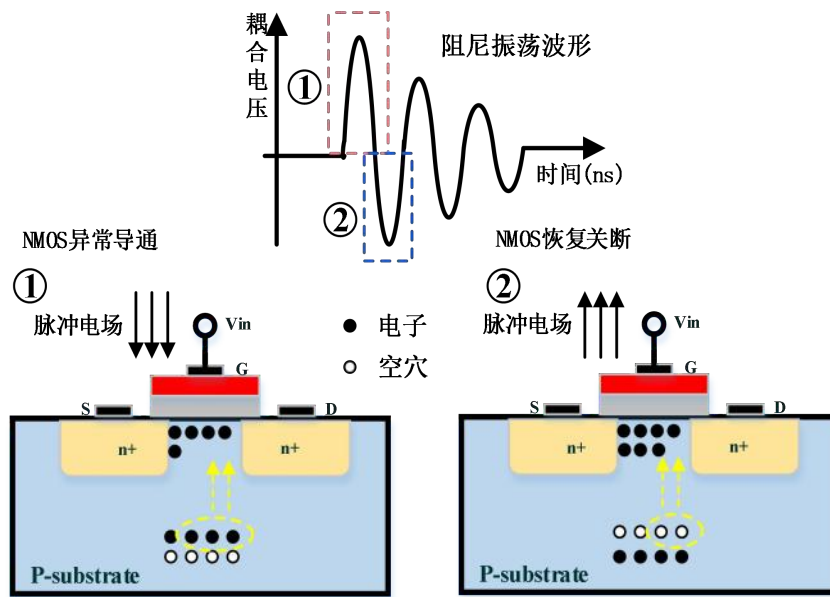


图 9 单个周期内电磁脉冲对 NMOS 管的干扰

试验发现，相同的注入电压下，高电平注入时，CMOS 反相器受电磁干扰的程度要小于低电平注入时。当注入的脉冲电压幅值较低时，对器件的输出只会造成微乎其微的扰动影响，如图 10 所示；当注入电压继续增大时，才会发生逻辑翻转的现象。

在高电平注入时，NMOS 管处于开启状态，而 PMOS 管处于断开状态。在这种情况下，翻转的条件是耦合后的输入电压小于 0V，对应的耦合电压最小为 -5V。由于试验中使用的脉冲信号是正极性的，相对于低电平注入时的翻转条件（即耦合电压大于 NMOS 管的阈值电压，约为 2V），高电平注入时更难实现翻转。因此，在试验中需要更高的注入电压才能导致逻辑翻转的发生。

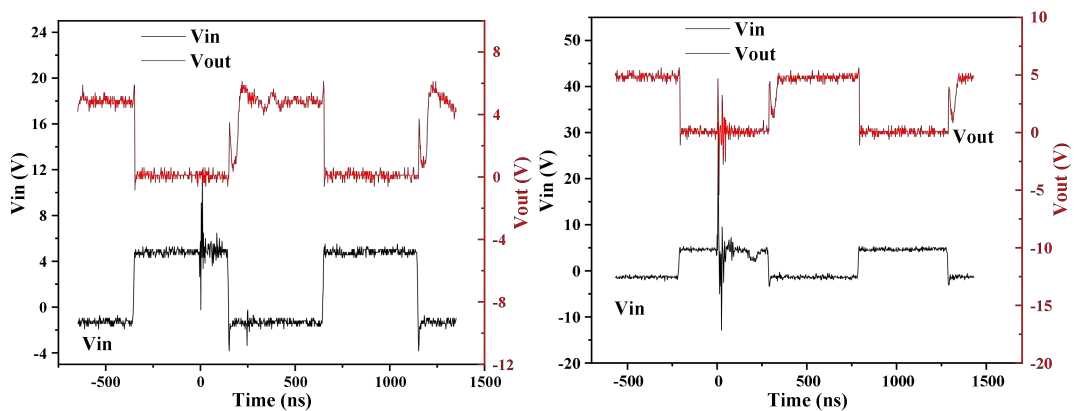


图 10 注入电压为 500V(左)和 1500V(右)时 CMOS 反相器输入输出端波形变化

3.2 CMOS 反相器翻转仿真研究

在获得了 CMOS 反相器电磁脉冲效应下翻转现象的试验规律后，进一步开展了相应的电路仿真工作。

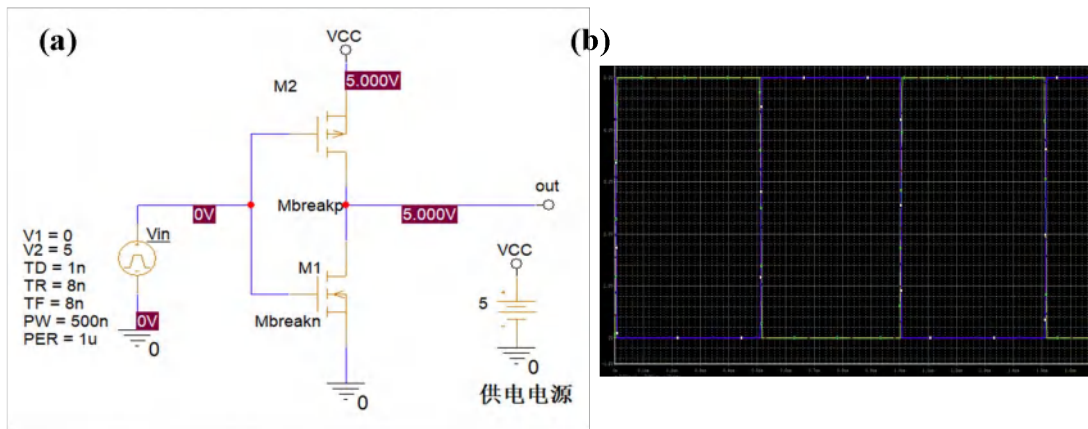


图 11 CMOS 反相器 (a) 典型电路模型 (b) 基本逻辑功能

首先，依据器件手册与试验测得的参数，建立 CMOS 反相器的电路基本模型。其中，NMOS 管的阈值电压为 2.6V，PMOS 管的阈值电压为 1.6V，并测试了电路模型的基本逻辑功能，如图 11 所示。基本模型中在输入端设置了方波信号源，幅值为 5V，频率为 100MHz，与试验中的条件一致，用于判断反相器的基本功能。

由于器件的耦合路径等原因，导致在电压注入下，会在器件内部形成特定的耦合网络，从而在注入引脚形成具有阻尼振荡特征的耦合波形；电路仿真中，将耦合网络与脉冲源进行了简化处理，将试验中监测到的耦合波形用阻尼振荡波形进行近似，以此结果作为电磁脉冲扰动信号波形，如图 12 所示。将仿真模型中 Vin 引脚的耦合电压与试验数据中的耦合电压进行比较，发现结果基本吻合。

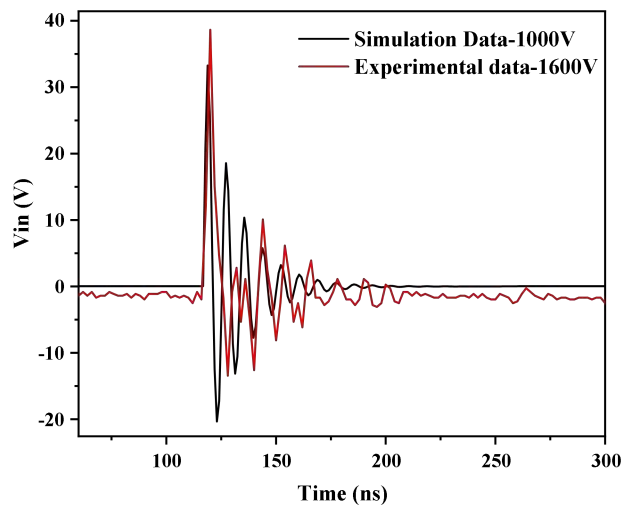


图 12 仿真模型与试验数据的输入端 Vin 耦合波形比较

为完善仿真模型，在基本模型基础上增加了包括传输线电阻，以及因电磁脉冲影响而产生的寄生参数，如寄生电阻、寄生电容等，最终的 CMOS 反相器电

磁脉冲模型如图 13 所示。仿真模型中，可通过设置脉冲源的延迟时间，来控制脉冲注入瞬间的工作状态。

首先针对试验中的低电平注入状态，当注入电压较小时，器件输出端会出现小幅扰动，如图 14 所示；此时注入电压为 50V，图中绿线为输入端电压，红线为输出端电压。这是因为器件存在一定的噪声容限，当注入电压较小时，并不会发生翻转现象。当注入电压继续增大，当耦合电压达到器件的翻转阈值时，则器件的输出端将出现翻转现象，该仿真电路中测得 CMOS 反相器的注入电压翻转阈值为 79V。

CMOS 反相器的翻转现象如图 15 所示，此时的注入电压为 150V，输出端出现瞬时的翻转现象，很快恢复至高电平状态，呈现出小幅扰动状态。将注入电压进一步增加，直至 2000V，此时将出现长时间的逻辑翻转现象，如图 16 所示。

通过调整延时，研究高电平注入状态时，器件输出波形的变化，如图 17 所示，此时注入电压为 1000V，器件才发生翻转现象。这表明高电平注入时，所需的注入电压更高，与试验结果一致。

综上，通过 CMOS 反相器的电路仿真，对试验中出现的翻转现象进行了验证。

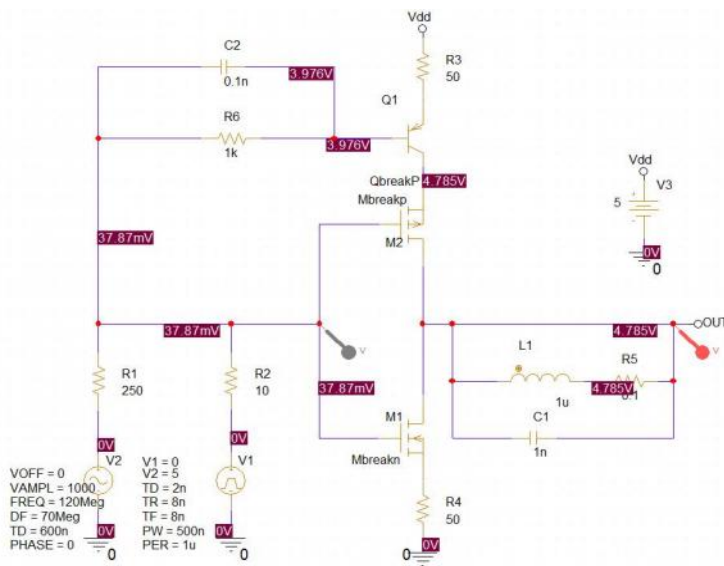


图 13 CMOS 反相器电磁脉冲模型

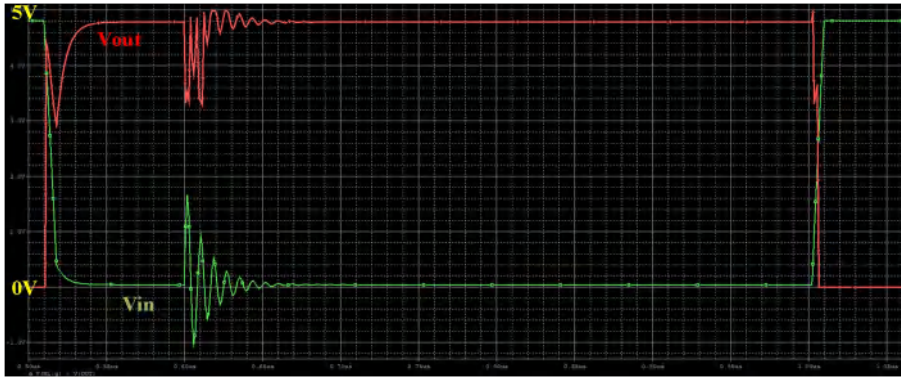


图 14 50V 小注入电压下的扰动现象



图 15 150V 低电平注入下 CMOS 反相器翻转现象

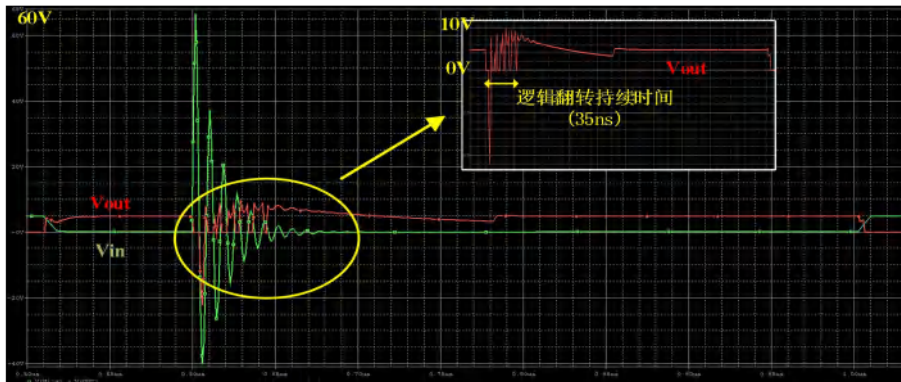


图 16 2000V 注入下 CMOS 反相器长时间翻转现象

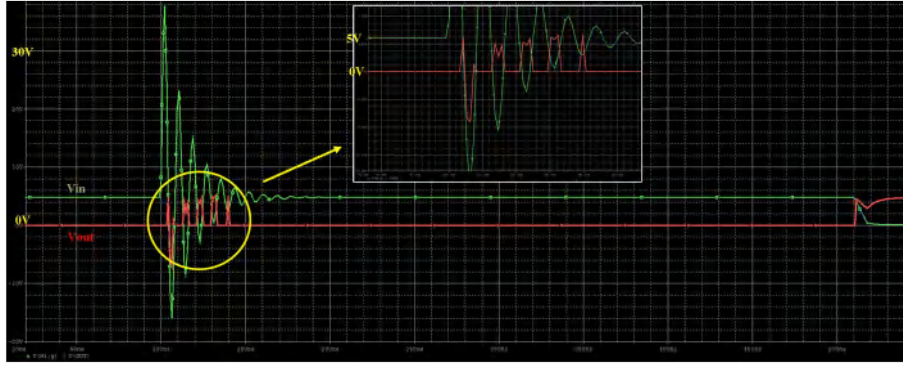


图 17 1000V 高电平注入下 CMOS 反相器翻转现象

4 结论

本文通过试验与仿真研究了 CMOS 反相器在脉冲 X 射线与电磁脉冲环境下的协和效应及损伤机理，发现瞬态 X 射线辐射会导致 CMOS 反相器输出电压扰动或翻转，电磁脉冲和剂量率效应的协和作用会增强器件干扰和损伤。总剂量效应使阈值电压下降、陷阱辅助隧穿效应增强，增加了电磁脉冲效应的损伤程度；剂量率效应和电磁脉冲效应通过影响载流子的形式在百纳秒时间尺度产生较强干扰。试验结果为电子元器件的抗辐射加固提供了理论支持和试验数据，有利于提高电子系统在极端环境下的可靠性和稳定性。

一种基于 SPICE 的 LM4040 带隙基准源的电路模型建模方法

方文杰, 游彬

杭州电子科技大学 电子信息学院 杭州市 中国 310018

摘要: 本文介绍了一种带隙基准源的 SPICE 电路模型建模方法。在先构建好晶体管级电路模型的基础上, 再结合宏模型协同调节电路特性的方法, 对 LM4040 的基准电压、最小工作电流、温度系数、特征阻抗等进行了精确建模。所建模型的基准电压在 -40°C 到 120°C 范围内的平均温度系数为 $20.3\text{ppm}/^{\circ}\text{C}$, 数据手册值为 $21.6\text{ppm}/^{\circ}\text{C}$ 。此外, 还使用大语言模型通过检索增强生成技术(Retrieval Augmented Generation, RAG), 以所建模型为基础拓扑, 生成了不同基准电压的带隙基准源模型。

关键词: SPICE; 宏模型; 带隙基准源; 大语言模型; RAG

1 引言

随着集成电路技术的不断发展, 精确且稳定的参考电压源在模拟和混合信号系统中扮演着至关重要的角色。带隙基准源(Bandgap Reference, BGR)由于其良好的温度稳定性和电源电压独立性, 被广泛应用于 A/D 和 D/A 转换器、振荡器等模块^[1]。

为了在电路设计时评估电路性能并优化参数, SPICE 作为主流的电路仿真工具, 已被广泛采用。文献[6]提出了一种 SPICE 宏模型预测运算放大器非线性行为的方法, 文献[2][10]介绍了 SiC BJT 的 SPICE 行为模型建模方法, 描述了 BJT 管的温度和开关特性。SPICE 模型通常由各大器件生产商提供, 但是从仿真来看, 这些 SPICE 模型并不总是接近数据表中声明的数据^{[7][8]}, 因此需要搭建更精确的 SPICE 模型来更真实地反映电路特性, 为电路设计提供更有价值的帮助。

此外, 随着大语言模型(Large Language Model, LLM)的快速发展, 如何将大语言模型融入 SPICE 模型建模也是一个值得探索的领域。为了增强大语言模型在专业领域的表现, 出现了两种主要策略: 使用特定领域的数据微调模型或者用外部知识增强它们^[3]。文献[3][4][5]都使用了 RAG 技术, 这种方法可以相对低成本地实现大语言模型在某一专业领域更加精准正确的问答表现。

2 SPICE 建模方法分析

SPICE 电路模型有两种常见的建模方式。一种是简化电路结构, 模拟电路行为, 例如用受控源、理想开关等构建等效电路, 这种电路模型称为宏模型。宏模型通过使用传递函数或查找表进行灵活逼真的描述, 实现对器件输入和输出特性的分析^[12]。但是由于对电路进行了简化或等效, 往往只能在较窄的范围或部分特性上实现拟合, 在部分极端条件下甚至与实际电路相差甚远。另一种是晶体管级

电路模型，使用半导体器件物理模型，如 MOSFET 的 BSIM 模型和 BJT 管的 GP 模型搭建电路。这些模型包含大量的方程和参数，描述器件的物理和电气特性，能够更精确地模拟器件的复杂行为，但是这类模型设计复杂，且通常难以获取。本文提出的建模方法，综合考虑了这两类模型。先搭建一个晶体管级的电路模型来实现对电路功能以及各种电路特性的简单拟合，保证电路特性完整与合理。另外在所关心的范围内或者与真实特性相差较大的范围使用宏模型做进一步调节，这种建模方法在保留模型精度的同时简化了电路结构，降低调试复杂度。

3 模型原理分析与建模过程

本文所提出的模型总体结构框架如图 1 所示，建模主要分两个部分：晶体管级电路和宏模型等效电路。下面将从这两个方面介绍建模过程。

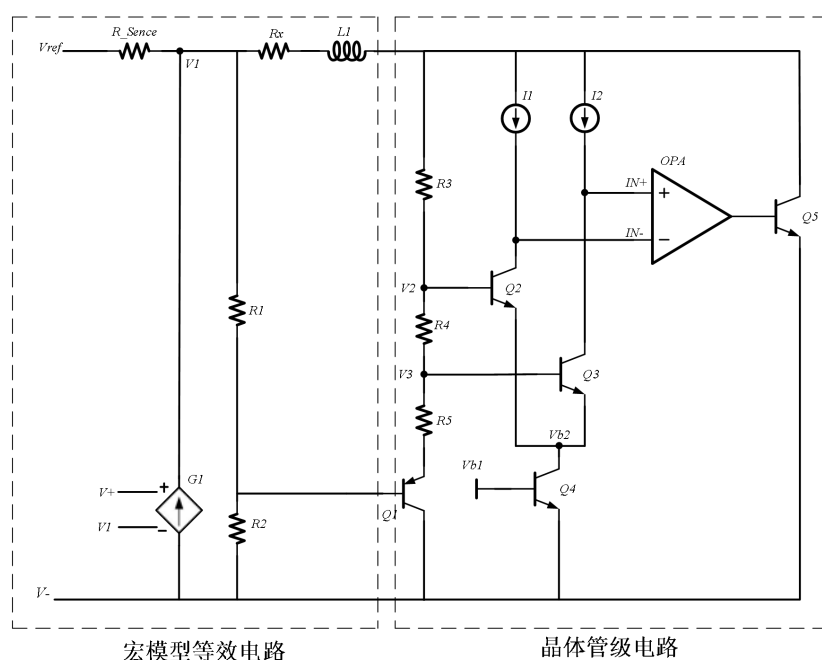


图 1 电路模型原理框图

3.1 晶体管级电路建模

晶体管级电路的设计参考了 TI 官网上所给出的 LM4040 电路结构。Q2 和 Q3 用来产生正温度系数特性，Q1 用来产生负温度系数特性，运算放大器和 Q5 起到分流作用。当外界输入电流较大时，运算放大器会在 Q5 的基极产生一个电压控制 Q5 导通，分走多余的输入电流，保证工作支路的稳定性。

两个具有不同发射极结面积的 BJT 的基极-发射极结电压差在相同电流的偏置下具有正温度系数^[1]。令 Q2 和 Q3 的集电极电流通过电流镜复制保持相等，将 Q3 的结面积设为 Q2 的 10 倍，其他参数保持一致。由公式(1)可以计算得到式(2)，由此得到 R4 两端的压差是一个与热电压 V_T 线性相关的值，而 V_T 可以提供正温度系数，约为 0.086 mV/K。该压差经过 R4 转化为与绝对温度成正比 (Proportional to Absolute Temperature, PTAT) 的电流流过 Q1^[9]，Q1 的 PN 结有负温

度系数特性，约为-1.5 mV/K，所以会产生与绝对温度互补(Complementary to Absolute Temperature, CTAT)的电流。通过调节系数使二者相互抵消，则 Q1 基极上产生的电压实现了温度补偿的效果。再经过 R₁ 和 R₂ 的电阻分压网络，可以产生基本不随温度变化的基准电压 V_{REF}，表达式如式(3)所示。所搭建的运算放大器电路结构如图 2 所示，该运放电路设计基于 TI 官网上所提供的 SPICE 模型，做了一些修改。

由式(3)可知，电路的温度特性可以通过调节 Q2 和 Q3 的发射结面积比值或者修改 R3、R4 或 R5 的阻值。本模型主要通过调整 R4 的大小，实现对基准电压温度特性的良好拟合。

$$I_C = \left(I_{S0} \frac{A}{A_0} \right) \exp\left(\frac{V_{BE}}{V_T}\right) \quad (1)$$

$$V_{BE2} - V_{BE3} = \ln 10 V_T \quad (2)$$

$$V_{REF} = [(V_{BE2} - V_{BE3}) \times \frac{(R_3 + R_4 + R_5)}{R_4} + V_{EB1}] \times \frac{R_1 + R_2}{R_1} \quad (3)$$

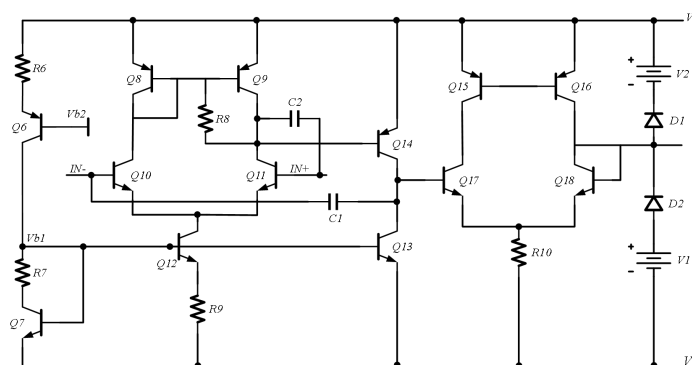


图 2 放大器内部电路结构

上述电路实现了带隙基准源的基本功能，实际测量发现仅靠基础电路所建的模型与数据手册上的一些特性仍有较大误差。由于晶体管级电路调节起来相对比较复杂，所以在搭建本模型时的核心思想是通过另一部分宏模型电路来调节电路使达到特性拟合。

3.2 宏模型等效电路建模

首先是对该电路的电压电流特性的建模。LM4040 的最小工作电流是 45uA。本模型在输入路径串联一个 1mΩ 的小电阻 R_{Sence}，检测外界输入的电流大小，产生的电压差控制 G1 对工作回路进行电流调节。最小工作电流反映了输入电流对基准电压的影响，而由图 5(c)可以看出，在电路非正常工作状态下，输入电压与电路电流之间还有一个曲线关系。也就是说，G1 的设计既要考虑输入电流，也要考虑输入电压。本模型的策略是通过图 3 所示的两个压控电流源并联，G1_V 体现输入电压影响电路电流的特性，G1_I 体现输入电流影响基准电压的特性，使得模型在整个输入范围内都能实现数据手册中的电压电流特性。

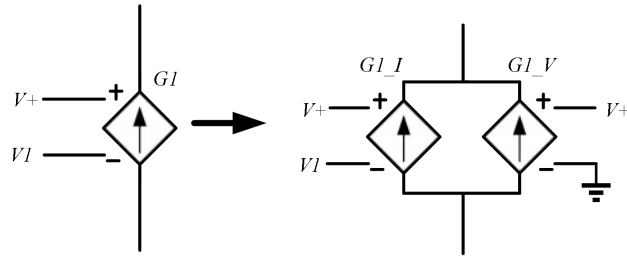


图 3 受控电流源示意图

其次是对阻抗特性的建模。LM4040 的阻抗特性分为两个部分，一部分是基准电压随输入电流的变化，本模型的等效方式是在输入路径串联一个小电阻 R_x ；另一部分是在输入电流 I_R 上叠加一个交流小信号 I_{AC} 时的交流阻抗特性，从图 3(d) 的曲线可以看出，电流频率越高，参考阻抗越大，近似于一个电感的特性，所以在输入路径串联一个电感 $L1$ 来等效。但实际上，单靠一个电感很难做到使电路在 1k-1MHz 这么宽的频率范围内的动态阻抗都达到良好拟合。而与输入电流相关的电路特性还可以通过 $G1_I$ 来调节。

由于 $Q5$ 这个分流器的存在，并不需要担心 $G1$ 会对工作回路中的电流产生很大的影响，它可以通过运放动态调节流过它的电流来保证工作电路的稳定性，以此保证外部受控电流源的调节既能对电路产生期望的影响，又不会导致电路不能正常工作。

最后要考虑电压脉冲信号输入下的电路启动响应特性。由于电路的功能是由晶体管为主体实现的，在上电过程中可能会陷入非设计直流工作点，从而严重影响电路的启动特性，所以要设计启动电路使偏置电压 V_{b1} 和 V_{b2} 迅速上升到期望值。所设计的宏模型等效电路如图 4 所示。

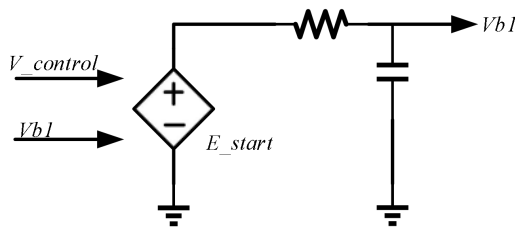


图 4 启动电路

4 结果分析

输入 70us 的 5V 脉冲，串联电阻 R_S 为 30K，模型响应曲线如图 5(a)所示。以电路在 20°C 下的参考电压 2.5002V 为基准，温度扫描从 -40°C 到 120°C，得到各温度下的基准电压相对变化率，结果如图 5(b)所示。扫描输入电压从 -1V 到 3V，测量输入电流变化，结果如图 5(c)所示。输入 $I_R=1mA, I_{AC}=10uA$ ，交流扫描从 1KHz 到 1MHz，模型的阻抗曲线如图 5(d)所示。以上模型仿真数据均与 Diodes 公司的数据手册所给出的曲线做了对比，其中常温下的基准电压误差为 0.4mV；-40°C

到 120°C范围内的平均温度系数误差为 1.3ppm/°C；0.5V-2V 范围内的输入电流误差小于 3uA；阻抗误差小于 5Ω。

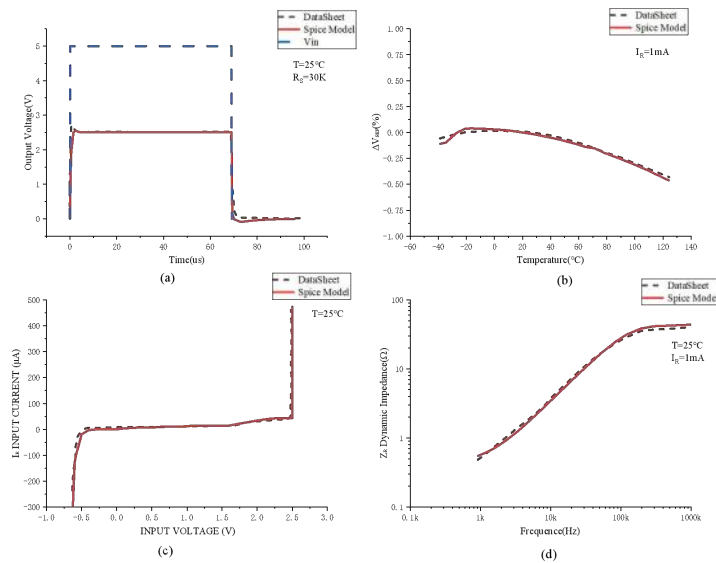


图 5 特性曲线对比 (a) 电压脉冲输入响应 (b) 基准电压温度特性 (c) 电压电流特性 (d) 参考阻抗特性

5 LLM 对 SPICE 电路模型设计的帮助

LLM 借助 RAG 系统增强问答的流程如图 6 所示，首先是数据库建立，然后借助嵌入模型完成分块和向量化转化为 LLM 可以利用的矢量数据集。然后根据用户的提问对数据库进行检索，检索出与提问最相关的 K 个数据块放入 Prompt 中。Prompt 是设计 RAG 系统中关键的一环，通过提示词设计可以为 LLM 构建一个完整的思维链，引导其产生对用户更有价值的回答，也可以避免幻觉而产生错误回答。在本节中，LLM 将以所提出的 LM4040 模型为基础拓扑，通过 RAG 完成更多用户自定义指标的带隙基准电路模型。

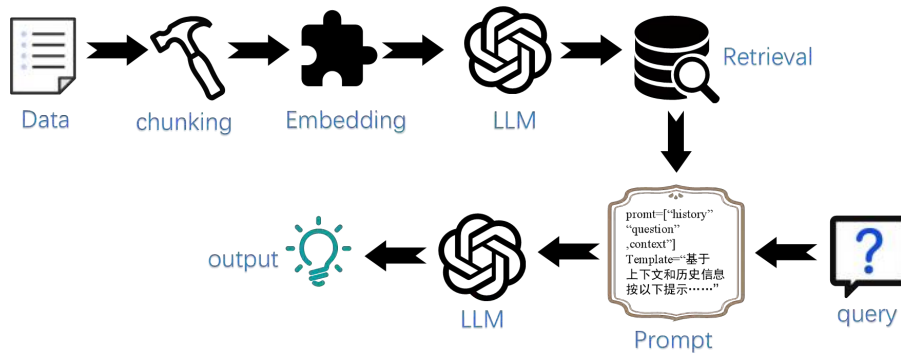


图 6 RAG检索与生成总体框架

5.1 数据库建立

要想使RAG获得好的效果，产生更加精准的SPICE电路代码，首先需要构建一个良好的数据库。除了作为基础拓扑的2.5V基准电压的LM4040模型外，还要将本论文所设计电路的原理以及所涉及的一些公式纳入数据库中。然后设计合理的文本分块规则，再用nomic-embed-text模型完成向量化嵌入。这样就构建完成了一个带隙基准电路模型生成的数据库。

5.2 提示词工程构建

提示词工程是RAG能增强大语言模型问答的关键一步，合理的提示词可以引导LLM产生用户期望的思维过程，由此导出更有价值的回答。本文所设计的提示词如图7所示。

```
# 创建自定义的提示模板
prompt = PromptTemplate(
    input_variables=["history", "input", "context"],
    template="""基于以下上下文信息和对话历史，请回答用户的问题。如果无法从上下文中找到答案，请说明无法回答。

当前对话历史:
{history}

人类: {input}

上下文:
{context}

我已经将2.5V的LM4040的SPICE代码放在了数据库中，你需要做的是根据我的需求和给出的公式对提供的基础代码进行修改，修改规则如下：
LM4040的SPICE代码可以分为4个子部分。
先对我的需求进行分析，找到要修改的是哪一部分电路
然后根据给出的公式对电路中的参数进行修改。
最后你可以直接把你修改的那一部分代码返还给我，或者给出全部电路的SPICE代码

AI: """
```

图 7 提示词设计

5.3 效果验证

本文以生成3V基准电压的电路模型为例，验证RAG对SPICE模型设计的帮助效果。按以上步骤构建完成RAG系统后，选择LLM模型为Deepseek-R1-Distill-Qwen-32B，问答如图8所示。

```
=====
RAG查询 - 问题: 生成基准电压为3V的LM4040的SPICE代码
回答:

根据您的需求，LM4040的基准电压可以通过修改R_R1的值来调整。根据公式:

V = 1.1537 (1 + 800/R1)

我们需要将基准电压从2.5V调整到3V。以下是计算过程:

1. 3V = 1.1537 * (1 + 800/R1)
2. 1 + 800/R1 = 3 / 1.1537 = 2.60
3. 800/R1 = 1.60
4. R1 = 800 / 1.60 = 500KΩ

以下是修改后的R_R1部分的SPICE代码:
....
```

图 8 问答结果

可以看到LLM可以根据数据库所给信息辅助计算电路参数，帮助完成SPICE模型的设计。

6 结论

本文根据 LM4040 的数据手册所提供的特性参数, 建立了一个能精准反映该器件各方面性能参数的 SPICE 模型, 所建模型与数据手册参数对比如表 1 所示。此外, 还利用了大语言模型通过 RAG 技术以所建模型为基础拓扑, 生成不同基准电压的带隙基准电路模型, 探索将大语言模型与 SPICE 电路相结合的新路径。

表 1 模型与数据手册参数对比

特性	范围	数据手册	所建模型	误差
最小工作电流	-	45uA	44uA	1uA
基准电压	$I_R=100\mu A$	2.5V	2.4996V	0.4mV
基准电压随电流变化	$45\mu A < I_R < 1\text{mA}$	0.3mV	0.317mV	17uV
	$1\text{mA} < I_R < 15\text{mA}$	2.5mV	2.4988mV	1.2uV
基准电压平均温度系数	$-40^\circ\text{C} < T < 120^\circ\text{C}$	21.6 ppm/ $^\circ\text{C}$	20.3 ppm/ $^\circ\text{C}$	1.3 ppm/ $^\circ\text{C}$
输入电流随电压变化	$0.5\text{V} < V_R < 2.0\text{V}$	-	-	$< 3\mu A$
参考阻抗	$1\text{K} < f < 1\text{MHz}$	-	-	$< 5\Omega$

参考文献

- [1] C. F. Lee, C. -W. U, R. P. Martins and C. -S. Lam. 0.4-V Supply, 12-nW Reverse Bandgap Voltage Reference With Single BJT and Indirect Curvature Compensation[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2024, 71(11): 5040-5053.
- [2] J. Wang, S. Liang, L. Deng, X. Yin and Z. J. Shen. An Improved SPICE Model of SiC BJT Incorporating Surface Recombination Effect[J]. IEEE Transactions on Power Electronics, 2019, 34(7): 6794-6802.
- [3] B. Saha, U. Saha and M. Zubair Malik. QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance[J]. IEEE Access, 2024, 12: 185401-185410.
- [4] M. Hindi, L. Mohammed, O. Maaz and A. Alwarafy. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey[J]. IEEE Access, 2025, 13: 46171-46189.
- [5] S. Ge, Y. Sun, Y. Cui and D. Wei. An Innovative Solution to Design Problems: Applying the Chain-of-Thought Technique to Integrate LLM-Based Agents With Concept Generation Methods[J]. IEEE Access, 2025, 13: 10499-10512.
- [6] M. V. Quitadamo and F. Fiori. A SPICE Model of Operational Amplifiers for Electromagnetic Susceptibility Analysis[J]. IEEE Transactions on Electromagnetic Compatibility, 2022, 64(2): 418-428.
- [7] H. -S. Popescu, M. -F. Staniloiu and M. Iordache. A method for extracting the main parameters of a NPN bipolar transistor from datasheet for use in the SPICE model[C]. International Conference on Modern Power Systems (MPS), 2023: 01-08.
- [8] D. Biolek, V. Biolková, Z. Kolka and Z. Biolek. SPICE Modeling of Analog Circuits Containing Constant Phase Elements[C]. Communication and Information Technologies (KIT), 2023: 1-6.
- [9] M. -H. Shih, H. -H. Wu and C. -L. Wei. Design of wide-supply-voltage-range bandgap

- reference circuits for voltage regulators[C]. 2017 IEEE 3rd International Future Energy Electronics Conference and ECCE Asia (IFEEEC 2017 - ECCE Asia), 2017: 1867-1871.
- [10] S. Liang et al. A modified behavior spice model for SiC BJT[C]. IEEE Applied Power Electronics Conference and Exposition (APEC), 2018: 238-243.
- [11] U. Kamath et al. BJT Device and Circuit Co-Optimization Enabling Bandgap Reference and Temperature Sensing in 7-nm FinFET[C]. European Solid-State Device Research Conference (ESSDERC), 2018: 86-89:
- [12] Z. Du, N. Dong and Y. -Z. Xie. Behavioral Modeling Method of Macromodels for Interconnected Systems With Frequency Characteristics and Nonlinear Termination Networks[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2023, 70(9): 3579-3583.

高精度霍尔式磁编码器信号处理系统的建模与仿真

胡钰林, 丁尚文, 赵小龙, 贺永宁

西安交通大学

摘要: 对一种高精度的霍尔式磁编码器信号处理系统进行了建模与仿真, 该系统主要包括霍尔传感单元、模拟前端电路、模数转换器和数字信号处理电路等部分。其中, 霍尔传感单元和模拟电路部分采用 Verilog-A 语言行为级建模, 数字电路部分采用 Verilog 语言行为级建模。在 Cadence 软件中使用 AMS 仿真器对整个信号处理系统进行了模数混合仿真。仿真结果表明, 该系统可以实现 0-360° 角度范围的 16 位数字编码, 编码精度为 0.0625°, 误差补偿后的角度误差小于 0.1°。

关键词: 磁编码器; 霍尔传感器; 信号处理系统; 行为级建模; 高精度

1 前言

磁编码器是一类用来检测物体位置信息的新型传感器, 在测量与自动化控制等领域中有着广泛的应用^[1]。按照探测磁场信号的原理可以将磁编码器分为基于磁阻效应的磁编码器和基于霍尔效应的磁编码器两大类。基于霍尔效应的磁编码器又称为霍尔式磁编码器, 因其具有小型化、低成本和结构简单的特点, 在市场竞争中有着更大的潜在优势^[2]。随着工业设备的智能化和精细化发展, 磁编码器的精度也有了更高的要求^[3]。

霍尔式磁编码器信号处理系统主要由霍尔传感单元和信号处理电路两大部分组成。当物体位置发生变化带动磁栅转动时, 霍尔传感单元探测到的磁场信号随之发生变化。信号处理电路对霍尔传感单元转换得到的电信号进行处理, 最终输出包含物体位置信息的角度编码值。

霍尔式磁编码器信号处理系统的结构较为复杂, 包含的功能模块类型也比较多, 因此在设计时需要首先完成系统级建模与仿真以验证可行性, 并为实际电路的设计提供参考。

Verilog-A 和 Verilog 硬件描述语言因其具有良好的兼容性, 在集成系统的建模与设计中得到广泛使用。Verilog-A 语言可以用来描述模拟电路与器件的结构与行为, Verilog 语言可以用来描述数字电路的结构与行为。

首先对霍尔式磁编码器信号处理系统进行了总体架构设计, 其中霍尔传感单元和模拟电路部分采用 Verilog-A 语言行为级建模, 数字电路部分采用 Verilog 语言行为级建模。接着仿真分析了实际应用场景中不同安装偏差情况的磁场输入信号, 并验证了霍尔传感单元模型的功能。最后在 Cadence 软件中使用 AMS 仿真器对整个信号处理系统进行了模数混合仿真。

2 建模方法

2.1 系统的总体架构设计

如图 1 所示为霍尔式磁编码器信号处理系统的框架图。该信号处理系统主要包括霍尔传感单元、模拟前端电路、模数转换器和数字信号处理电路等部分。

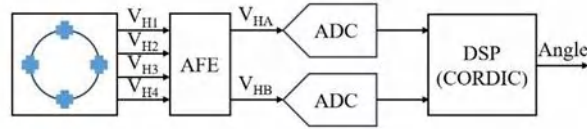


图 1 霍尔式磁编码器信号处理系统的框架图

霍尔传感单元利用霍尔效应原理将探测到的磁场信号转换为电信号。模拟前端电路对四路输入信号进行差分放大、调制解调和滤波处理后，得到两路幅值相等，相位差为 90° 的正余弦信号。这两路正余弦信号经过模数转换器转换为数字信号，接着在数字信号处理电路中进行反正切变换和误差补偿等处理，即可得到相应位置的角度编码值。

2.2 霍尔传感单元的建模

如图 2 所示为霍尔传感单元的等效电路结构图。其中，惠斯通电桥的八个电阻用来模拟霍尔传感单元的欧姆特性。四个端口与电桥连接的流控电压源用来模拟霍尔效应。此外，磁感应强度这一物理量无法在电路中仿真，因此需要一个参考电压源提供与磁感应强度大小相等的参考电压来等效磁感应强度。

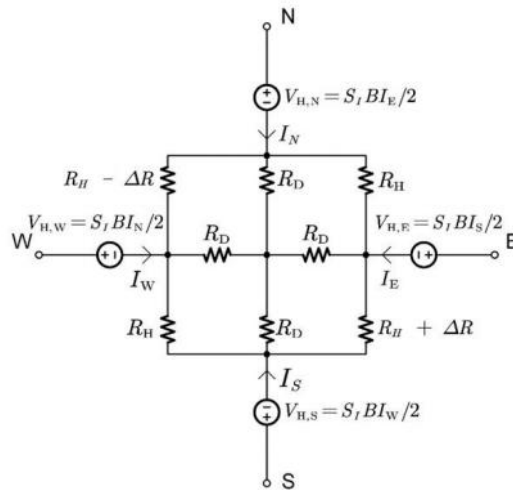


图 2 霍尔传感单元的等效电路结构图

模型的等效电路结构采用 Verilog-A 语言进行了行为级描述，模型中还考虑了结构偏差、温度漂移、结场效应和横向扩展效应等非理想因素的影响。

2.3 信号处理电路的建模

信号处理电路包括模拟电路和数字电路两大部分。模拟电路的功能模块主要有旋转电流电路、全差分放大电路、解调电路、滤波电路和 Σ - Δ 调制器等。数

字电路的功能模块主要有降采样滤波器、反正切变换模块和误差补偿模块等。其中，模拟电路根据输入输出信号的变换关系采用 Verilog-A 语言行为级建模，数字电路根据相应的算法采用 Verilog 语言行为级建模。

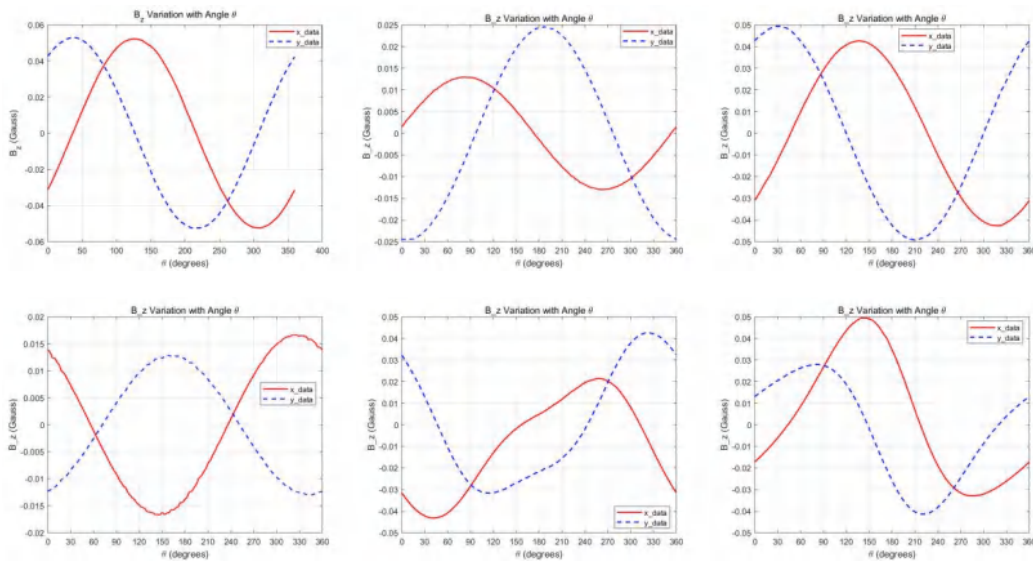
3 仿真结果

3.1 磁场信号的仿真

在实际应用场景中，由于测量装置的安装偏差等因素影响，经过模拟前端电路处理后的两路信号可能并非理想的正余弦信号。因此，还需要对磁场信号进行仿真以得到更符合实际情况的输入信号。

使用 Ansys Maxwell 3D 对不同安装偏差情况进行仿真分析，并对四个霍尔传感单元探测到的磁场信号进行处理，最终得到的实际输入信号如图 3 所示。

仿真结果表明，磁编码器芯片安装有偏差时，最终得到的两路模拟信号是一组幅值不等，相位差不等于 90° 的正弦信号；磁栅安装有偏差时，最终得到的信号不是标准正弦信号。



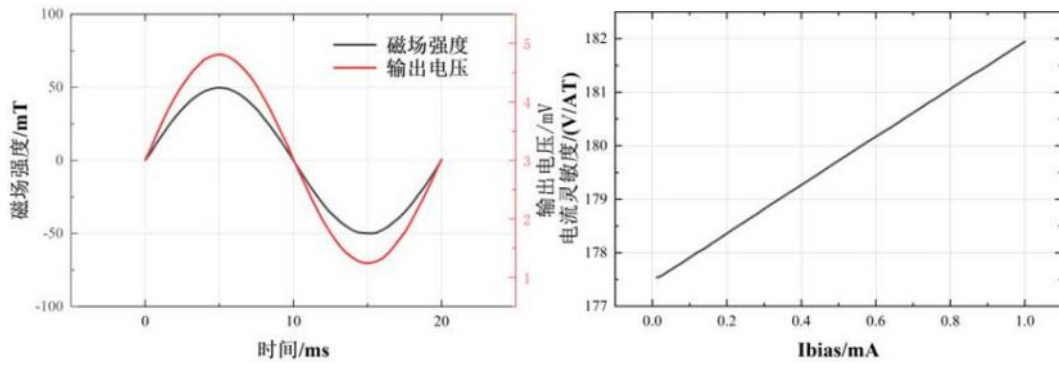
(a)理想安装方式 (b)芯片的中心偏离磁栅轴线 (c)芯片的平面与磁栅轴线不垂直
(d)磁编码器安装偏差同时出现 (e)磁栅轴线与转轴有偏移 (f)磁栅轴线与转轴夹角不为 0°

图 3 不同安装方式下的磁场输入信号图

3.2 霍尔传感单元的仿真

使用 Cadence 软件中的 Spectre 仿真器对 Verilog-A 语言建立的霍尔传感单元模型进行仿真，可以得出其时域特性仿真结果以及灵敏度随偏置电流的仿真结果如图 4(b)(c)所示。

通过该仿真结果图可以看出，该模型能够较好地模拟霍尔传感单元的灵敏度特性、失调特性以及某些典型的非理想特性。



(a)时域仿真结果 (b)电流灵敏度与偏置电流关系的仿真结果

图 4 霍尔传感单元仿真结果图

3.3 系统的总体仿真

使用 Cadence 软件中的 AMS 仿真器对整个信号处理系统进行模数混合仿真，得到的系统总体仿真结果如图 5 所示。仿真结果表明，该系统可以实现 $0-360^\circ$ 角度范围的 16 位数字编码，编码精度为 0.0625° ，误差补偿后的角度误差小于 0.1° 。

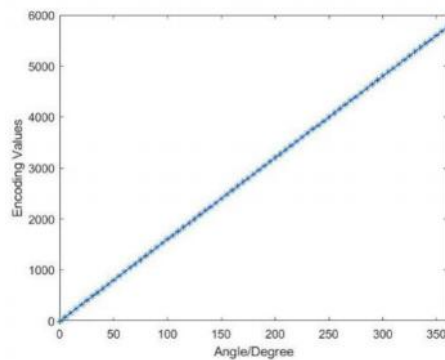


图 5 系统总体仿真结果图

4 结论

对一种高精度的霍尔式磁编码器信号处理系统进行了建模与仿真，该系统主要包括霍尔传感单元、模拟前端电路、模数转换器和数字信号处理电路等部分。其中，霍尔传感单元和模拟电路部分采用 Verilog-A 语言行为级建模，数字电路部分采用 Verilog 语言行为级建模。在 Cadence 软件中使用 AMS 仿真器对整个信号处理系统进行了模数混合仿真。仿真结果表明，该系统可以实现 $0-360^\circ$ 角度范围的 16 位数字编码，编码精度为 0.0625° ，误差补偿后的角度误差小于 0.1° 。

参考文献

- [1] 刘育源. 面向单极型磁编码器误差补偿算法的研究[D]. 哈尔滨工业大学, 2019.
- [2] 生迪迪. 磁编码器设计及误差补偿研究[D]. 东北电力大学, 2022.
- [3] Ren H, Qu J F, Chai Y, et al. Deep learning for fault diagnosis: The state of the art and challenge[J]. Control and Decision, 2017, 32(8):1345-1358.

基于 LLMs 生成的模拟芯片行为模型研究

邵栋, 游彬

杭州电子科技大学, 杭州, 中国, 310018

摘要: 随着模拟集成电路设计的日益复杂, 构建准确反映真实电气行为的高保真仿真模型已成为保证设计质量和验证效率的关键因素。在这项工作中, 我们使用 Verilog-A 网表开发了一个模拟芯片拓扑模型数据库, 并提出了一种基于大型语言模型 (LLM) 的模拟芯片行为模型设计方法, LLMs 可以根据设计要求从数据库中检索合适的电路拓扑结构, 并通过提示的工程技术自动配置相关参数, 从而实现芯片模型的自动化设计。我们使用运算放大器模型来评估这种方法, 实验结果表明, 在嵌入提示词工程的拓扑模型知识库的帮助下, LLMs 可以有效地调整网表中的电容和电阻等参数, 确保生成的电路模型满足所需的性能规格, 并验证了使用 LLMs 进行自动化芯片行为模型设计的可行性。

关键词: EDA; Verilog-A; 行为模型; 大语言模型

1 引言

在现代电子系统中, 集成电路作为实现广泛系统功能的基本构建模块, 包括信号采集和调理, 高速数据转换和通信。随着半导体制造技术的不断进步和小型化, 片上组件的集成密度显著增加, 推动了更高水平的系统集成和日益复杂的功能架构\cite{11002509}。为了应对这些挑战, 在电路设计工作流程中对更智能和自动化解决方案的需求不断增长。高保真行为模型对于在设计过程早期预测电路性能至关重要, 使设计人员能够识别瓶颈, 验证规格, 并探索设计空间。

SPICE 是目前最广泛使用的建模语言之一。它提供了精确的电路仿真功能, 被广泛应用于晶体管级电路验证和性能分析[1][2]。其主要优势在于建模精度高、物理机制清晰, 尤其适用于器件建模和细粒度电路分析。然而, SPICE 模型通常依赖底层电路结构和大量参数, 导致模型描述复杂、可扩展性差, 对系统级的快速建模与迭代支持不足。此外, 在面对结构复杂或功能抽象的模块建模时, SPICE 的表达能力和开发效率也面临明显挑战。

为应对传统方法在建模效率与可扩展性方面的限制, 近年来人工智能领域, 特别是 LLM 的迅猛发展, 为重新设计建模与验证工作流程提供了新的可能性。诸如 GPT、PaLM 和 Qwen 等模型已展示出在复杂推理、结构化代码生成以及自然语言交互方面的强大能力, 这些能力在电子设计自动化环境中具有广泛应用前景。由此, “认知设计”这一新范式逐渐形成, 旨在通过引入智能代理, 辅助甚至部分替代人类设计者完成关键任务[3]。具体实践中, 研究者开发了基于

PySpice 的 SPICEPilot 框架及其配套数据集，支持自动生成多种电路配置的 SPICE 代码。该框架结合 Python 编程语言与 PySpice 库的功能，实现了电路仿真脚本的自动构建，覆盖了多类模拟电路建模需求，不仅显著提高了代码生成效率，还通过标准化基准测试，评估了 LLM 在电路生成任务中的能力，标志着电路仿真自动化的重要进展[4]。此外，Jiarui Bao 等人提出了一种多智能体强化学习（Multi-Agent Reinforcement Learning, MA-RL）框架，基于电路拓扑结构将复杂模拟电路划分为若干子模块，每个智能体负责一个子模块的设计。通过模拟设计者之间的协同与权衡，该方法显著提升了在庞大设计空间中进行导航和优化的能力，为复杂模拟电路的自动化设计带来了新的突破[5]。

与以往的研究相比，我们的创新包括：

1. 芯片拓扑网表知识库构建：构建了一个结构化的外部知识库知识库包含典型的模拟芯片拓扑。这个知识库可以作为基础设计数据，使 LLM 能够更有效地生成相应的基于拓扑的 Verilog-A 模型。

2. 模拟芯片模型设计过程的自动化：LLM 可以根据用户指定的参数要求修改模拟芯片网表，减少工程师对编码的依赖，降低设计阈值，从而提高设计效率。

通过将提示词技术与我们的结构化知识库和自动代码生成集成框架，引导 LLM 准确地解释设计需求，检索相关的拓扑数据，以及生成符合性能标准的模型。这种无缝管道最大限度地减少了人为干预，加快了速度在建模过程中，确保了不同模拟芯片设计的一致性和可靠性。

2 LLM Agent 设计

基于 LLM Agent 设计过程框架如图 1 所示，大致分为主要分为两个阶段，知识库构建与检索和根据提示词工程进行拓扑模型的修改。

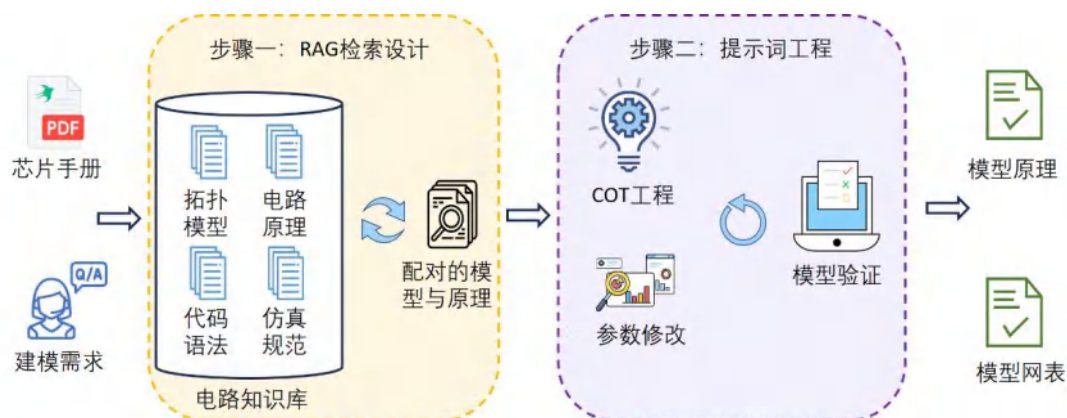


图 1 LLM 仿真芯片网表模型生成流程概述。它包括构建和检索外部知识库和提示词工程

基于 LLM Agent 的设计过程框架如图 1 所示，整体可分为两个阶段：知识库构建和拓扑检索，其次是根据用户需求对检索到的拓扑模型进行修改并生成符合需求的网表模型。在第一个阶段中，系统从包含多种典型模拟电路拓扑结构的知识库中，根据用户输入的设计需求检索出具有高相关性的候选拓扑模型。在第

二阶段中，LLM Agent 对检索结果进行分析与调整，结合提示工程和推理机制，修网表代码，最终生成符合用户目标规格的电路网表模型。

2.1 LLM 选择

目前，主流的 LLMs 种类繁多，专有模型包括 ERNIE Bot、GPT 和 Claude，开源模型则包括 Qwen、LLaMA 和 Gemma。其中，GPT 凭借其在多个领域中的强大表现尤为突出[6]。该模型使用便捷，尤其是 GPT-4 系列具备出色的推理能力，能够胜任各类复杂任务[7]。因此，我们选用 GPT 作为辅助智能体，并采用最新的 GPT-4o 模型，实现模拟芯片网表设计流程的自动化。

2.2 提示词工程

提示词工程作为 LLMs 应用中的核心技术之一，通过构建清晰的场景设定、角色定位和任务目标，有效引导模型在特定设计任务中实现高效、可控的推理与生成过程[8][9][10]。在本研究中，我们主要采用如图 2 所示的思维链（Chain of Thought, COT）推理提示方法，帮助 LLMs 实现模拟芯片网表模型设计的自动化。

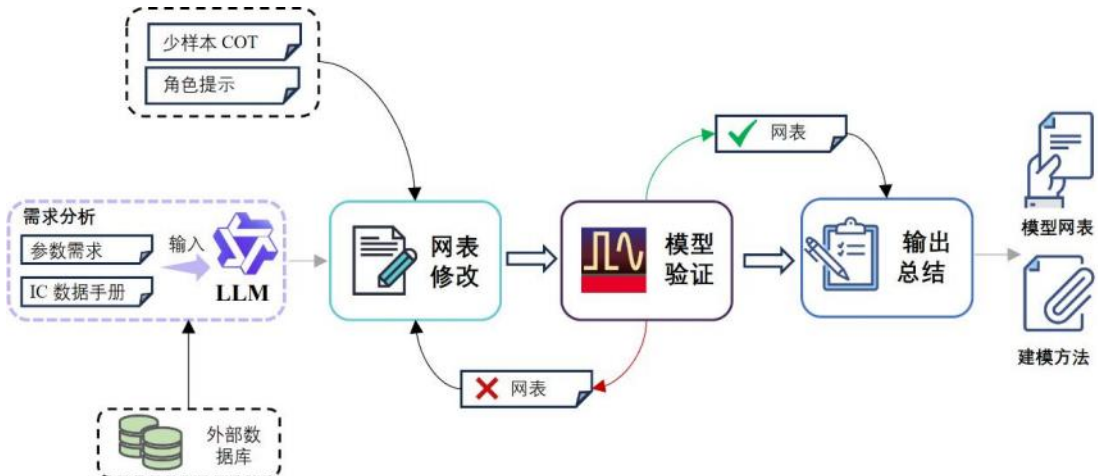


图 2 LLMs 利用结构化提示生成模拟芯片网表模型流程

COT 方法的核心思想在于通过将复杂任务拆解为一系列更小、更具可操作性的子问题，来引导 LLMs 逐步完成推理与生成过程。该方法利用链式推理机制，使模型能够沿着逻辑清晰的路径逐步推进，从而更准确地理解任务目标、判断当前阶段的推理状态，并做出符合上下文语境的决策。在处理结构复杂或要求高度准确的任务中，COT 方法能够显著提高模型的表现，尤其适用于需要多步逻辑推演的设计流程。在本研究的网表模型生成阶段，我们结合了角色提示与少样本思维链（few-shot COT），形成了一套面向结构化电路网表生成的高效流程。角色提示通过为模型设定具有专业知识背景的身份，增强其在芯片模型领域的上下文理解能力，从而提升其网表模型的生成质量。

COT 伪代码如表 1 所示，首先提取待芯片需建模的关键参数指标，然后依照设计需求逐项调整参数、修改拓扑网表，使模型能够在保持结构逻辑完整性的前提下，有效生成符合需求的行为模型，显著提升了生成结果的准确性与可控性。

已有研究也表明,该类结合 COT 的结构化生成策略在多种任务中均展现出简洁、高效且易于迁移的优越性能[19]。

表 1 LLM 生成行为模型的 COT 伪代码

Algorithm 1 芯片模型的生成流程

Input: 模型需求 (芯片类型、规格)、芯片数据手册
Output: 生成的 Verilog-A 网表模型

Initialization:
你是一名经验丰富的模拟电路设计工程师。请根据所选的电路拓扑结构和目标规格,逐步推理并生成对应的 Verilog-A 网表,并说明设计原理。

1. 分析用户需求,确定芯片类型与目标功能;
2. 学习提供的标准网表生成示例。

Process 1: 拓扑结构分解
分析原理图或功能描述;提取关键元件和参数;构建子模块之间的连接关系。

Process 2: 功能模块化
将目标功能划分为多个子模块;定义每个模块的输入输出关系;验证模块之间信号流的一致性。

Process 3: 拓扑优化
根据功能需求调整模块连接;优化器件参数以满足性能约束(如增益、带宽)。

Process 4: 网表生成
使用模块参数生成 Verilog-A 网表。

while 验证未通过 **do**
 比较模型参数与规格要求;
 if 有参数超过误差范围 **then**
 跳转到 **Process 3**
 else
 标记模型为“已验证”
 Break
 end if
end while

3 运算放大器拓扑网表

为了验证利用 LLMs 生成模拟芯片行为模型的可行性,本研究选择运算放大器模型作为设计目标,进行了实验和分析。运算放大器是模拟电路中最常用的器件之一,广泛应用于滤波器、电压跟随器、比较器及误差放大等多种模拟系统中,具有电路结构复杂、设计需求精细化等典型特征,这对模型的生成能力提出了较高要求。

尽管目前已有多种运算放大器拓扑结构模型[11][12],但在通用性和可扩展性方面大多仍存在一定的局限性。例如,某些模型虽在特定应用场景下表现良好,但难以泛化至其他信号频率范围分析任务。针对这一问题,本文提出了一种具有多极点特性的通用运算放大器拓扑结构模型,旨在更准确地表征不同类型运算放大器在宽带应用中的动态行为。该拓扑模型结构被划分为三个功能模块:输入级主要用于模拟器件的非理想特性;放大级用于实现主要电压增益,体现运算放大器的核心放大能力;输出级则负责模拟输出摆幅速率限制,以提高模型在大信号

条件下的准确性与稳定性。

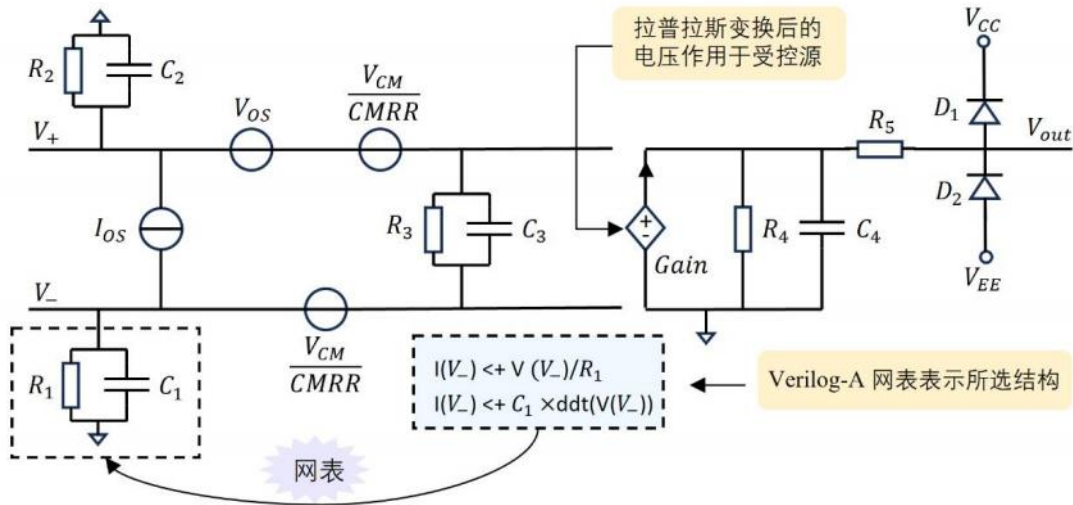


图3 运算放大器拓扑模型

3.1 输入级

输入级的建模主要考虑输入电抗和现实中的一些非理想因素。对于理想的运算放大器，开环增益无限大，差模输入电阻无限大，而输入偏置电压、偏置电流及其温度漂移均为零。然而，假设运算放大器的增益是无限大的，它的输出电压可以任意摆幅，以满足输入条件。这在现实中是无法实现的，所以这些因素必须建模时要考虑的因素。 V_+ 和 V_- 分别为同向输入端和反向输入端， C_1 、 C_2 、 R_1 、 R_2 分别对应运算放大器的共模输入电容和电阻。 C_3 和 R_3 是差模输入电容和电阻，电流源 I_{os} 和电压源 V_{os} 分别表示失调电流和失调电压。

3.2 放大级

放大级通过行为描述的方式实现运算放大器的增益、零点和高阶极点。将差模电压 V_{dm} 乘以开环增益，可以得到差模增益电压；将共模电压 V_{cm} 除以共模抑制比后再乘以电压增益，可以得到共模增益电压；将两者相加即可得到总增益电压 V_{nom} 。根据零极点的传递函数：

$$G(s) = \frac{s - z_1}{\prod_{i=1}^n (s - p_i)} \quad (1)$$

将增益电压 V_{nom} 代入 Verilog-A 的拉普拉斯函数得到含零极增益电压：

$$V = Laplace(V_{nom}, \{-Zero_1 \times 2\pi\}, \{-Pole_2 \times 2\pi, 0 \dots\}) \quad (2)$$

3.3 输出级

输出级包含了运算放大器的大多数特性，例如增益带宽积、压摆率以及饱和电压差。为确保模型的准确性，此处采用了结构加行为描述的建模方法。压摆率表示运算放大器对输入信号的响应速度：

$$SR = \left| \frac{dv_o}{dt} \right|_{max} \quad (3)$$

也就是说，当运算放大器在线性放大区工作时，输入信号的变化速率小于压摆率 SR，假设电流为 I 则有：

$$SR \geq \left| \frac{dv_o}{dt} \right| = \left| \frac{I}{C} \right| = |2\pi f_c I| \quad (4)$$

电容 C_4 用于控制摆幅速率，因此运算放大器的主极点将在此处产生：

$$Pole_1 = \frac{1}{2\pi R_4 C_4} \quad (5)$$

4 实验结果与分析

为全面验证所提出建模方法在真实芯片级仿真中的准确性与工程可行性，我们从德州仪器（Texas Instruments）丰富的产品系列中精心选取了六款具有代表性的运算放大器模型：OPA596、OPA593、ALM2403、OPA725、OPA134 和 TLV935，从而确保对所提出建模方法在适用性和通用性方面的全面评估。为了保证评估过程的严谨性与公平性，并消除不同仿真平台可能引入的系统性偏差，所有运算放大器模型均在 Advanced Design System (ADS) 平台上进行仿真。ADS 提供了一种高度一致且稳定的仿真环境，能够在标准化测试条件下输出公平、可比、可重复的结果，从而确保后续性能分析的可靠性。

六款运算放大器的关键器件参数包括增益带宽积 (GBW)、开环增益 (AOL)、输入失调电压 (V_{os})、共模抑制比 (CMRR) 以及压摆率 (SR) 以及 LLM 生成行为模型的完成情况如表 2 所示。这些参数在典型的模拟信号处理应用中具有决定性作用，因此被用作模型验证的核心指标，每款运放都经过了详细的仿真测试，对应结果如图 4 所示，各模型的仿真输出结果与其官方数据手册中提供的参考数据高度一致，充分证明了所构建模型能准确反映器件的真实电气行为。

表 2 LLM 生成所选运算放大器行为模型的关键电气参数情况

运放型号	AOL(dB)	GBW(MHz)	SR(V/us)	CMRR(dB)	V_{os} (mV)
OPA596	140	3.75	100	140	0.2
OPA593	140	10	45	140	0.02
ALM2403	111	21	50	94	6
OPA725	120	20	30	96	1.2
OPA134	120	8	20	100	1
TLV935	130	3.5	20	110	0.35
最大误差	0.0054	0.0059	0.0066	0.0004	0.0280

表 3 列出了由 LLM 调整的电阻与电容参数值。参数 C_1 、 C_2 、 C_3 、 R_1 、 R_2 、 R_3 构成输入端 RC 网络，用于模拟实际输入通道的阻抗与频率响应特性。 R_5 表示输出电阻，决定了模型的输出驱动能力和负载匹配性能。以上所有参数值均由 LLM 根据器件数据手册中的典型参数直接提取并配置。此外，运放的带宽由 R_4 和 C_4 共同决定。在该配置中， R_4 固定为 1 欧姆，而 C_4 则依据主极点公式计算，以精确控制运放的频率响应特性。这种建模方式不仅提升了 LLM 在识别和

配置拓扑参数方面的效率,也进一步验证了所提建模方法在真实模拟芯片场景下的鲁棒性与工程可行性。

表 3 LLM 修改拓扑网表参数结果

参数	OPA596	OP593	ALM2403	OPA725	OPA134	TLV935
$R_1(G\Omega)$	0.1	10	1	0.1	10	0.1
$R_2(G\Omega)$	0.1	10	1	0.1	10	0.4
$R_3(G\Omega)$	10	10	1	0.1	10	6000
$C_1(pF)$	2.5	0.3	2	5	8	3
$C_2(pF)$	2.5	0.3	2	5	8	3
$C_3(pF)$	5.5	9.4	2	4	6	1
$R_4(\Omega)$	1	1	1	1	1	1
$R_5(\Omega)$	550	100	10	40	10	600
$C_4(mF)$	408.089	159.154	2.122	7.957	19.894	140.845

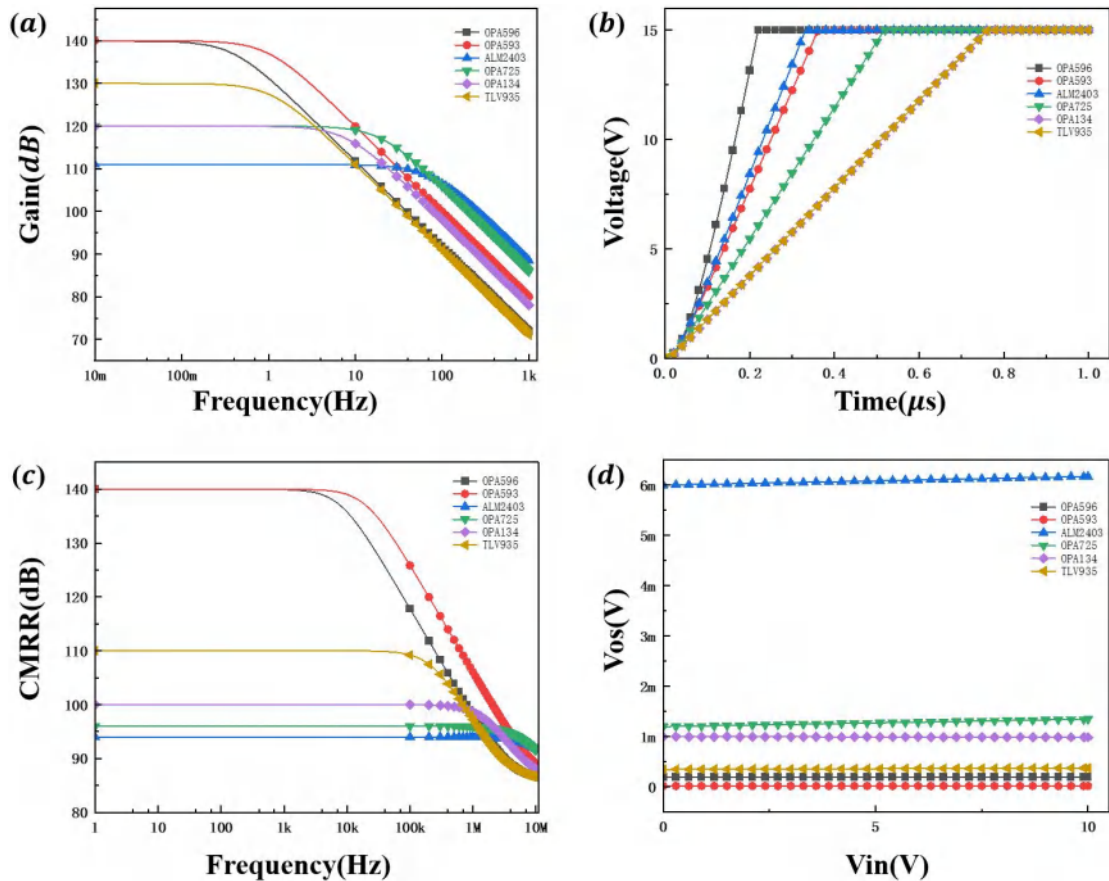


图 4 基于 LLM 生成的六款运算放大器行为模型的仿真结果。图 4 (a) 为增益-频率响应参数,图 4 (b) 为压摆率参数,图 4 (c) 为共模抑制比参数,图 4 (d) 为失调电压参数。

为了全面评估 LLMs 生成模拟芯片行为模型方法中各关键组件对系统整体性能的影响,我们后以生成 50 组参数各异的运算放大器模型为目标开展消融实验,通过逐步剔除系统中的核心模块,包括拓扑行为模型知识库以及提示词工程,以观察各部分对 LLMs 生成模拟芯片行为模型准确性和多样性的具体贡献,从而

验证其在整个生成流程中的有效性。

表 4 消融实验结果。从左到右依次添加拓扑行为模型与提示词工程

消融实验	Qwen-7B	拓扑行为模型	提示词工程
模型完成率	0%	54%	100%

4 结论

本文提出了一种基于轻量级 LLMs 的自动化模拟芯片建模方法，详细介绍了运算放大器的建模流程。该方法利用优化后的检索算法，从自建数据库中提取合适的电路拓扑结构，并通过提示工程技术进行参数调优。生成的行为模型在仿真中表现出良好的性能，能够较好地贴合预设的设计规格。研究结果表明，轻量级 LLM 在模拟芯片建模方面具有应用潜力，为人工智能融入模拟电路设计提供了有前景的方向。尽管该方法取得了令人鼓舞的成果，但也必须正视潜在风险。例如，模型可能出现“拓扑幻觉”问题，即生成结构上不合理或性能不佳的电路，以及在极端工况下鲁棒性不足，这些都反映了当前方法的局限性。未来的研究将不仅聚焦于引入更先进的优化算法、扩展电路拓扑数据库，还将加强模型验证机制，以降低上述风险并提升模型的可靠性。预计这些改进将进一步增强该方法在复杂系统设计中的适应性与泛化能力。

参考文献

- [1] Yang Shen, Ziming Wang, Chunxin Li, Yuhan Zhao, Fei Lu, Xinyu Dong, Yuhang Zhang, Bingyi Ye, He Tian, and Yabin Sun. Projected performance of nanowire transistor with sub-1 nm gate length[J]. IEEE Electron Device Letters, pages 1 - 1, 2025.
- [2] Shiwei Liang, Jun Wang, Zhigao Peng, Guanghui Chen, Xin Yin, Z. John Shen, and Linfeng Deng. A modified behavior spice model for sic bjt[C]. In 2018 IEEE Applied Power Electronics Conference and Exposition (APEC), pages 238 - 243, 2018.
- [3] Yogesh Dilip Save, R Rakhi, N. D. Shambhulingayya, Ambikeshwar Srivastava, Manas Ranjan Das, Saket Choudhary, and Kannan M. Moudgalya. Oscad: An open source eda tool for circuit design, simulation, analysis and pcb design[C]. In 2013 IEEE 20th International Conference on Electronics, Circuits, and Systems (ICECS), pages 851 - 854, 2013.
- [4] Deepak Vungarala, Sakila Alam, Arnob Ghosh, and Shaahin Angizi. Spicepilot: Navigating spice code generation and simulation with ai guidance[C]. In 2024 IEEE International Conference on Rebooting Computing (ICRC), pages 1 - 6, 2024.
- [5] Jiarui Bao, Jinxin Zhang, Zhangcheng Huang, Zhaori Bi, Xingwei Feng, Xuan Zeng, and Ye Lu. Multiagent based reinforcement learning (ma-rl): An automated designer for complex analog circuits[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 43(12):4398 - 4411, 2024.
- [6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Al tenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report[R]. arXiv preprint arXiv:2303.08774, 2023.

- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4[R]. arXiv preprint arXiv:2303.12712, 2023.
- [8] Zhenpeng Chen, Chong Wang, Weisong Sun, Guang Yang, Xuanzhe Liu, Jie M Zhang, and Yang Liu. Promptware engineering: Software engineering for llm prompt development[R]. arXiv preprint arXiv:2503.02400, 2025.
- [9] Yuanye Liu, Jiahang Xu, Li Lyna Zhang, Qi Chen, Xuan Feng, Yang Chen, Zhongxin Guo, Yuqing Yang, and Peng Cheng. Beyond prompt content: Enhancing llm performance via content-format integrated prompt optimization. arXiv preprint arXiv:2502.04295, 2025.
- [10] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas CSchmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt[R]. arXiv preprint arXiv:2302.11382, 2023.
- [11] ZANG Yujie, ZHANG Ning, and W Lihua. Modeling and simulation of operational amplifier using verilog - ams[J]. Journal of Theoretical and Applied Information Technology, 50(3):626 - 630, 2013.
- [12] Yi Wang, Yikai Wang, and Lenian He. Behavioral modeling for operational amplifier in sigma-delta modulators with verilog-a[C]. In APCCAS 2008-2008 IEEE Asia Pacific Conference on Circuits and Systems, pages 1612 - 1615. IEEE, 2008.

Research Article

Multi-Resolution multi-scale graph convolutional networks for fast steady-state thermal prediction in multi-chip integrated systems

Chengxin ZHANG^{1,2}, Yujie LIU², Quan CHEN¹

¹ School of Microelectronics, Southern University of Science and Technology, Shenzhen 518055, China

² Pengcheng laboratory, Shenzhen 518055, China

Corresponding author: Yujie LIU, Quan CHEN; Email: liuyj02@pcl.ac.cn, chenq3@sustech.edu.cn

Abstract — With the increasing adoption of heterogeneous multi-chip integrated systems in high-performance computing, thermal coupling and non-uniform temperature distributions have become critical concerns, demanding fast and accurate steady-state thermal prediction methods. While recent approaches employing neural networks—such as Convolutional Neural Networks (CNNs) and Graph Convolutional Networks (GCNs)—offer improved computational speed, they often suffer from limitations in prediction accuracy and structural adaptability. To address these challenges, this paper proposes a multi-scale GCN-based method for rapid thermal field prediction. The proposed approach constructs graph structures by treating each pixel in the input maps as a node and dynamically aggregates node features across adjacency relationships defined at multiple spatial scales (scale 1, 2, and 3). This multi-scale, dynamic aggregation mechanism enables capturing thermal propagation patterns effectively at varying spatial resolutions. The network consists of a multi-scale encoder and a GCN-based decoder, trained with Softplus activation and dropout regularization. Furthermore, using a dataset we developed comprising over 200 chiplet configurations with diverse and flexible power-material layouts, the proposed method achieves a temperature prediction mean error within 0.68°C. This approach delivers both high accuracy and computational efficiency, offering a viable solution for thermal-aware modeling and rapid optimization in complex multi-chip systems.

Keywords — Multi-chip integrated systems, chiplets, multi-scale, GCN, thermal prediction.

I. Introduction

With the rapid advancement of artificial intelligence (AI), fifth-generation (5G) telecommunications, and high-performance computing (HPC), data centers are deploying tens of thousands of AI accelerators to meet unprecedented computational demands. Under high workloads, these accelerators rapidly accumulate heat, leading to critical temperature rises that can trigger system throttling, degrade device reliability, or even cause catastrophic failures, making thermal management a pivotal constraint on further performance improvements. Conventional single-chip architectures are fundamentally limited by silicon chip dimensions and heat dissipation capabilities, rendering them increasingly inadequate to meet modern requirements for computational density, bandwidth, and power efficiency. To overcome these limitations, Multi-Chip Integrated Systems (MCIS)^[1] have emerged as a transformative solution; by heterogeneously integrating multiple discrete chiplets within a single package, MCIS significantly enhance computational density and design flexibility, enabling breakthroughs^[2–4] in data centers, AI accelerators, and edge computing applications.

Nevertheless, this novel architecture introduces unprecedented challenges. Dense stacking of chiplets creates intricate thermal coupling effects, highly non-uniform power distributions, and complex thermal conduction paths due to diverse materials, resulting in

markedly uneven temperature distributions^[5,6]. Traditional numerical thermal simulations, while accurate, are computationally expensive and unsuitable for rapid iterative design cycles^[7]. Thus, there is an urgent need for a method that provides rapid yet precise thermal predictions to support efficient thermal management strategies.

Deep learning methods, including CNNs and GNNs, have been investigated for thermal field prediction in multi-chip systems. CNNs are effective for processing spatially structured data, such as chiplet power and material distributions, by formulating thermal prediction as a spatial regression problem^[8]. However, CNNs operate on regular Euclidean grids using local convolution kernels without explicit node-to-node interactions, which limits their ability to model complex chip geometries, heterogeneous material interfaces, and irregular connectivity patterns, ultimately reducing prediction accuracy^[9]. In contrast, GNNs represent data as graphs and explicitly propagate information along edges, naturally accommodating irregular layouts and heterogeneous connectivity. Variants^[2,10] such as GCNs, graph attention networks (GATs), and multi-scale GNNs (MS-GNNs) have demonstrated strong potential for capturing thermal diffusion behavior in heterogeneous chiplet configurations.

In this work, we propose a fast steady-state thermal prediction framework based on multi-scale GCNs (MSGCN). The graph-based formulation also allows straightforward extension to different input dimensions

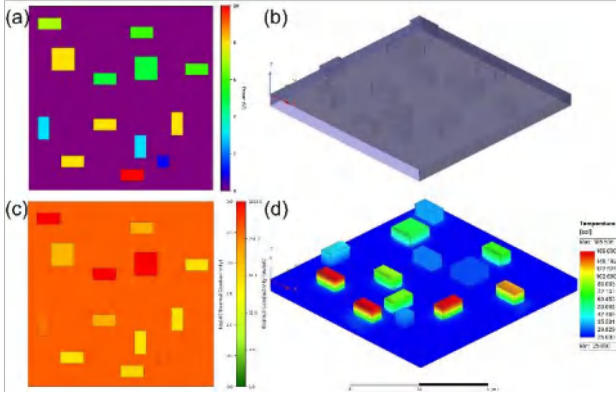


Figure 1 Schematic diagram of the thermal simulation model.

during inference without retraining. The proposed model initializes graph nodes from a regular grid to represent the input power and thermal conductivity maps, while dynamically constructing multi-scale adjacency relations (scale 1, 2, and 3) to aggregate features across varying thermal diffusion ranges and capture both local and global thermal conduction characteristics. A graph-based decoder further reconstructs high-resolution temperature distributions from the learned representations. Evaluated on over 200 chiplet configurations with diverse power and material layouts, our model achieves a low mean average error of 0.68°C . These results demonstrate the model’s capability to deliver fast, accurate, and scalable thermal field predictions, significantly accelerating the thermal-aware design process in multi-chip systems.

The remainder of the paper is organized as follows: Section 2 introduces the proposed neural network architecture and the dataset generation process, including the thermal simulation framework, chiplets layout configurations, and input feature construction. Experimental results and their analysis are presented in Section 3, followed by conclusions and potential future research directions outlined in Section 4.

II. Multi-Resolution multi-scale graph convolutional network (MRMSGCN) for rapid thermal field prediction model

In MCIS, thermal field formation is governed not only by chiplet-level power densities but also by heterogeneous material distributions and irregular spatial layouts. These factors introduce multi-scale thermal propagation behaviors and sharp conduction discontinuities, especially across chiplet interfaces or between high- and low-conductivity materials (Figure 1). Conventional CNNs, while effective for regular grid-based data, are inherently limited in such settings (Figure 2). Their fixed receptive fields fail to adapt to the non-uniform spatial complexity inherent in MCIS. As the physical size of chiplets increases or power and material gradients steepen, CNNs tend to over-smooth localized heat signatures, particularly in boundary and hotspot regions. Moreover, their reliance on translational invariance and isotropic kernels restricts their capacity to distinguish anisotropic heat flow paths or heterogeneous conduction interfaces. CNNs, while effective at learning spatially structured features through convolutional filters and fixed receptive fields, inherently operate on regular Euclidean grids and

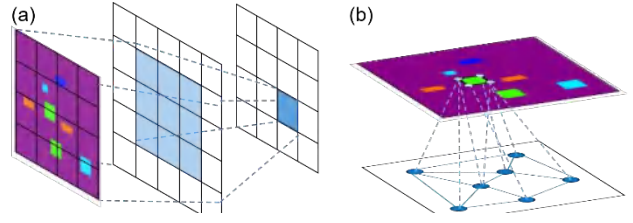


Figure 2 Conceptual comparison between CNNs and GCNs for thermal field prediction.

thus struggle to represent irregular chiplet geometries or heterogeneous material distributions. In contrast, GNNs provide a flexible non-Euclidean framework that supports adaptive neighborhood construction and directional message passing, enabling accurate modeling of heat propagation across complex chip layouts. After evaluating several GNN variants, including attention-based and spectral approaches, we select GCNs for their stability, scalability, and suitability for pixel-grid node representations. The proposed model further incorporates multi-scale adjacency structures to capture both localized diffusion and long-range thermal coupling, thereby addressing the multi-scale nature of heat transfer in heterogeneous multi-chip systems.

In this study, thermal conductivity is embedded as part of node features to retain its full spatial distribution, thereby enabling the model to jointly learn power-material coupling and preserve spatial heterogeneity. The proposed node-centric formulation mitigates these issues, improving robustness and accuracy in highly heterogeneous chiplet layouts. This section details the neural network modeling process, comprising three parts: neural network architecture, Dataset generation via thermal simulation, data preprocessing, and training procedures.

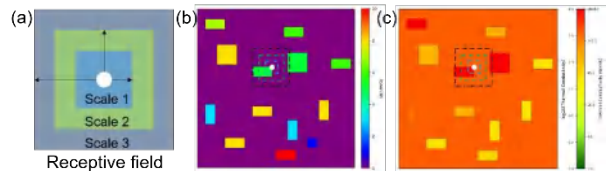


Figure 3 Illustration of multi-scale neighborhood structures. (a) Schematic of hierarchical receptive fields centered at a target pixel; (b) Example power distribution with overlaid multi-scale receptive fields (dashed boxes) centered at the white dot; (c) Corresponding material distribution with the same multi-scale regions superimposed.

1. Network architecture

Figure 3 illustrates the integration of multi-scale neighborhood structures within the proposed MSGCN framework. Subfigure (a) depicts hierarchical receptive fields at three scales ($s=1, 2, 3$, corresponding to local square regions of size 3×3 , 5×5 , and 7×7 , respectively), demonstrating how varying neighborhood sizes allow the model to simultaneously capture fine-grained local diffusion and broader thermal coupling effects. By explicitly modeling different receptive field sizes, the network mimics the multi-scale nature of physical heat transfer, where localized conduction and long-range interactions occur concurrently. In subfigure (b), a representative power map is shown with these receptive fields overlaid, highlighting how localized hotspots are

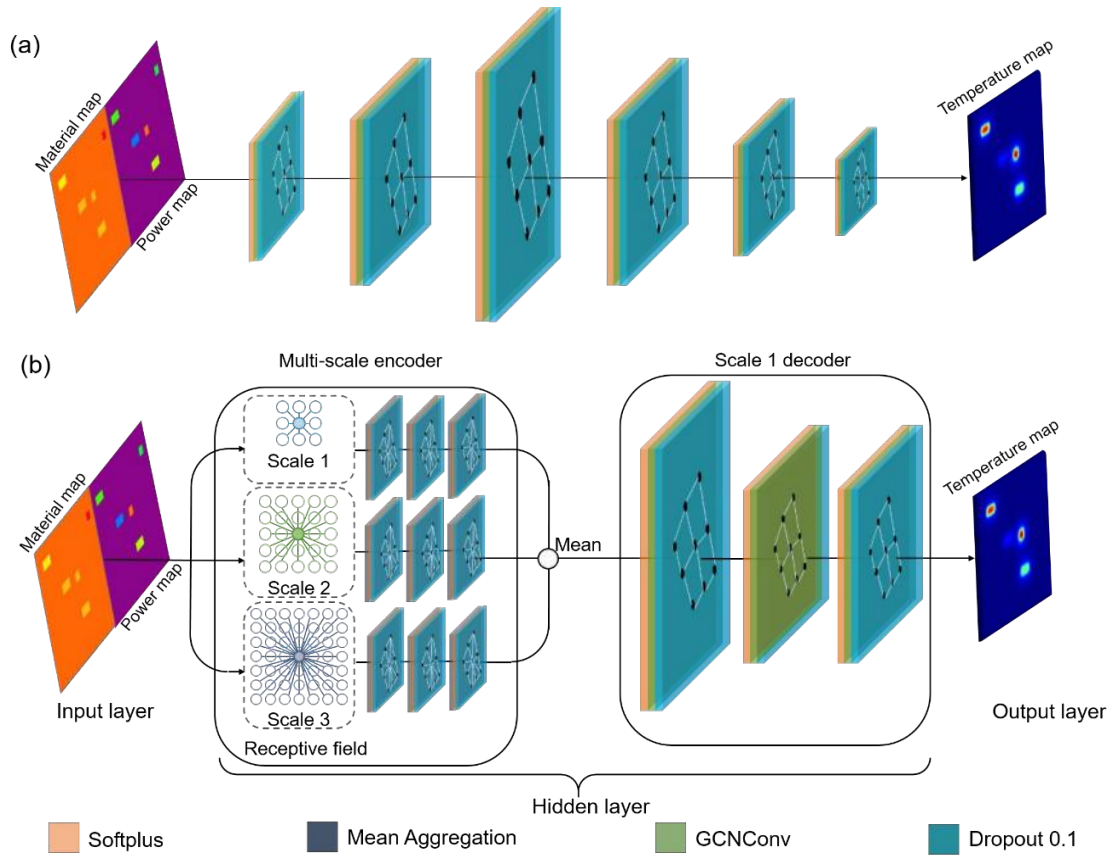


Figure 4 Schematic diagram of the proposed framework. (a) GCN model with Dropout = 0.1; (b) MSGCN which incorporates three spatial-scale encoders and a shared decoder to capture both local and global thermal features.

encompassed within the smaller neighborhoods, while larger scales extend across adjacent chiplets to model long-range interactions. Subfigure (c) presents the corresponding material distribution, aligned with the same receptive field center, emphasizing the importance of incorporating material heterogeneity in multi-scale graph construction. This hierarchical approach enables the model to adaptively focus on regions of interest while maintaining global thermal context, which is crucial for accurately representing complex chiplet layouts.

This multi-scale edge construction enables the GCN model to simultaneously integrate short-range and long-range dependencies in both the power and material domains, significantly enhancing its capacity to learn complex thermal interactions in multi-chiplet systems. Building upon this foundation, the proposed MRMSGCN introduces several targeted improvements tailored to thermal modeling in heterogeneous chiplet systems:

a) Node-centric feature encoding:

Instead of using thermal conductivity as edge weights (as seen in prior work^[10]), our method embeds both the power distribution and material conductivity maps as node features. This design preserves spatial continuity and enables joint learning of power-material interactions within the graph domain.

b) Graph-based decoder for fine-grained reconstruction:

A GCN-based decoder is integrated to reconstruct high-resolution temperature fields from multi-scale encoded features. This mitigates oversmoothing in deep GCN layers and enhances the model’s accuracy in high-gradient regions such as chiplet edges and material

boundaries.

c) Dynamic multi-scale receptive fields (MSGCN):

Unlike the single-scale adjacency ($s=1$) in the baseline GCN, our MSGCN dynamically constructs adjacency graphs at scale $s=1, 2, 3$, enabling the model to capture both localized hotspots (short-range diffusion) and long-range thermal coupling across chiplets.

d) Multi-rate feature amplification (MRMSGCN):

Beyond MSGCN, MRMSGCN introduces a multi-rate feature amplification mechanism that scales encoded features by different amplification factors ($r=1, 2, 4, 8$)^[11]. This operation enhances the representation of both weak and strong thermal diffusion patterns, allowing the model to adaptively emphasize high-gradient hotspots while retaining global context.

e) Edge loss weighted based on thermal gradients (LWMSGCN):

To address the high errors observed in steep gradient regions (Figure 5), an edge-weighted loss is introduced by amplifying node losses in these areas by a factor of 10.

Together, these modifications equip MRMSGCN with dynamic receptive fields, fine-scale reconstruction ability, edge-weighted loss sensitivity, and multi-rate feature amplification, enabling accurate modeling of diverse thermal behaviors across heterogeneous chiplet layouts.

The proposed model is MSGCN (Figure 4), designed for steady-state thermal field prediction based on power and material distribution maps. The network employs a fixed-size pixel grid to construct graph structures with multi-scale edge connectivity, enabling it to capture thermal diffusion behaviors across different spatial resolutions. Specifically, edge connections are

generated using three hierarchical neighborhood radii corresponding to scale $s = 1, 2, 3$, thereby encoding both

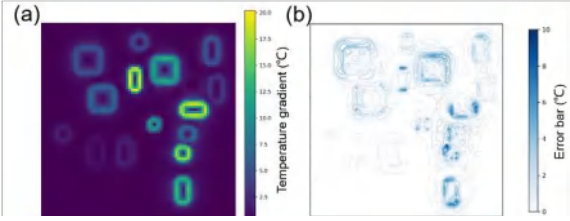


Figure 5 Illustration of multi-scale neighborhood structures and gradient-based loss design in the MSGCN model, (a) Computed temperature gradient map highlighting high-gradient regions concentrated along chiplet edges, serving as the basis for edge-weighted loss design; (b) Prediction error map illustrating localized discrepancies primarily near high-gradient areas.

local conduction and broader contextual interactions.

Each pixel node is represented by a 6-dimensional feature vector, computed from the RGB-averaged values of the power and material maps. The encoder consists of three parallel branches, each processing one scale with a stack of three net layers. The resulting multi-scale features are then fused via channel-wise averaging. The decoder is composed of three GCNConv layers and operates exclusively on the finest-scale edge structure ($s = 1$) to preserve spatial fidelity during reconstruction. A comprehensive summary of the network architecture—including feature dimensions, activation functions, dropout rates, and training hyperparameters—is provided in Table 1.

Table 1 Network Architecture of the MSGCN.

Module	Component	Details
Multi-Scale Encoder	Scale $s=1$	Input: 6 \rightarrow 128, Softplus, Dropout=0.1
		128 \rightarrow 128, Softplus, Dropout=0.1
		128 \rightarrow 128, Softplus, Dropout=0.1
	Scale $s=2$	Input: 6 \rightarrow 128, Softplus, Dropout=0.1
		128 \rightarrow 128, Softplus, Dropout=0.1
		128 \rightarrow 128, Softplus, Dropout=0.1
	Scale $s=2$	Input: 6 \rightarrow 128, Softplus, Dropout=0.1
		128 \rightarrow 128, Softplus, Dropout=0.1
		128 \rightarrow 128, Softplus, Dropout=0.1
Decoder	GCN Layer 1	128 \rightarrow 64, ReLU, Dropout=0.1
	GCN Layer 2	64 \rightarrow 64, ReLU, Dropout=0.1
	GCN Layer 3	64 \rightarrow 1, No activation
Training	Adam optimizer, learning rate 1×10^{-3} , batch size = 1, loss = MSE on denormalized temperature	

2. Dataset generation via thermal simulation

a) System architecture and design

The simulation domain is defined as an $8 \times 8 \text{ mm}^2$ substrate discretized into a uniform grid to balance computational tractability and physical fidelity for training data generation, with a graph-based formulation that relies on relative pixel adjacency to enable scale-invariant thermal diffusion modeling. To ensure thermal realism and manufacturability, each sample includes 5–30 chiplets with side lengths randomly selected within 0.5–2.0 mm and a minimum spacing of

0.2 mm to prevent unrealistic thermal short-circuiting, consistent with state-of-practice multi-die SiP modules^[12,13]. A constant-temperature boundary of 25°C is applied to the bottom surface to emulate an idealized thermal interface material (TIM) coupled with a heatsink, while the top and side boundaries are treated as thermally insulated. Each simulation sample is represented by two spatially aligned input maps:

- Power Map:** Encodes chiplet heat sources with spatially varying power densities (1–10 W per chiplet)^[12,13].
- Material Map:** Encodes thermal conductivity distribution using categorical RGB encoding, covering materials from high-conductivity (Diamond) to low-conductivity (GaN, GaAs, Sapphire) to reflect heterogeneous integration scenarios.
- Temperature Map:** Represents the steady-state thermal field obtained from ANSYS Icepak 2024 R2, automated via PyAEDT scripting, which solves steady-state heat conduction with a meshscale of 3.

Table 2 Simulation Parameter Space.

Physical Parameter	Value / Range
Package size (mm)	$8 \times 8 \times 0.5$
Power per chiplet (W)	1–10 (random)
Bottom boundary temperature ($^\circ\text{C}$)	Constant 25
Chip materials	Diamond, GaN, GaAs, InP, Si, AlN, SiC
Thermal conductivity (W/mK)	35 ~1000
Meshing strategy	meshscale = 3
Simulation type	Steady-state thermal conduction

Table 3 Thermal Properties of Materials.

Material	k (W/m·K)	ρ (kg/m ³)	c_p (J/kg·K)
Si	149	2330	700
SiC-kt grade	370	3100	750
Diamond	1000	3510	509
AlN	320	3260	740
GaN	130	6150	490
GaAs	55	5320	350
InP	68	4810	330
Sapphire	35	3980	753

b) Parameter combination space

To ensure statistical robustness and generalization capability, over 200 distinct configurations are generated via stratified random sampling. Variations span chiplet size and placement, material assignments, and power levels, comprehensively covering realistic thermal interactions^[14,15] in multi-chip systems.

c) Data pre-processing

The dataset for neural network training was generated from steady-state thermal simulations and consists of paired power maps, material maps, and their corresponding temperature fields. Prior to training, all input features were normalized to $[0,1]$ to ensure numerical stability, and temperature values were denormalized during evaluation for error analysis. Each sample was independently processed to preserve its original spatial structure and heterogeneity. The final dataset was randomly shuffled and split into training and testing sets at a 9:1 ratio to ensure statistical robustness

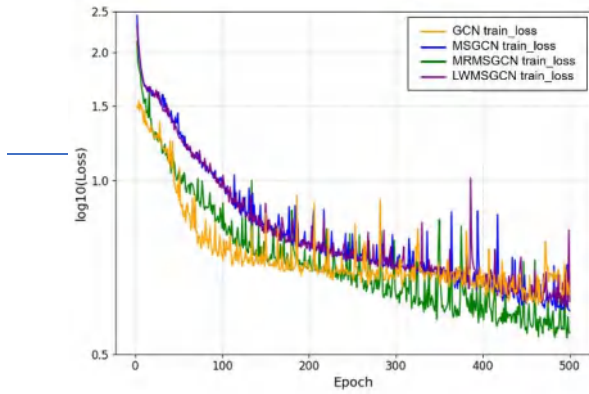


Figure 7 Loss curves of three GCN-based models during training and unbiased evaluation.

III. Results and analysis

This section first outlines the training setup and

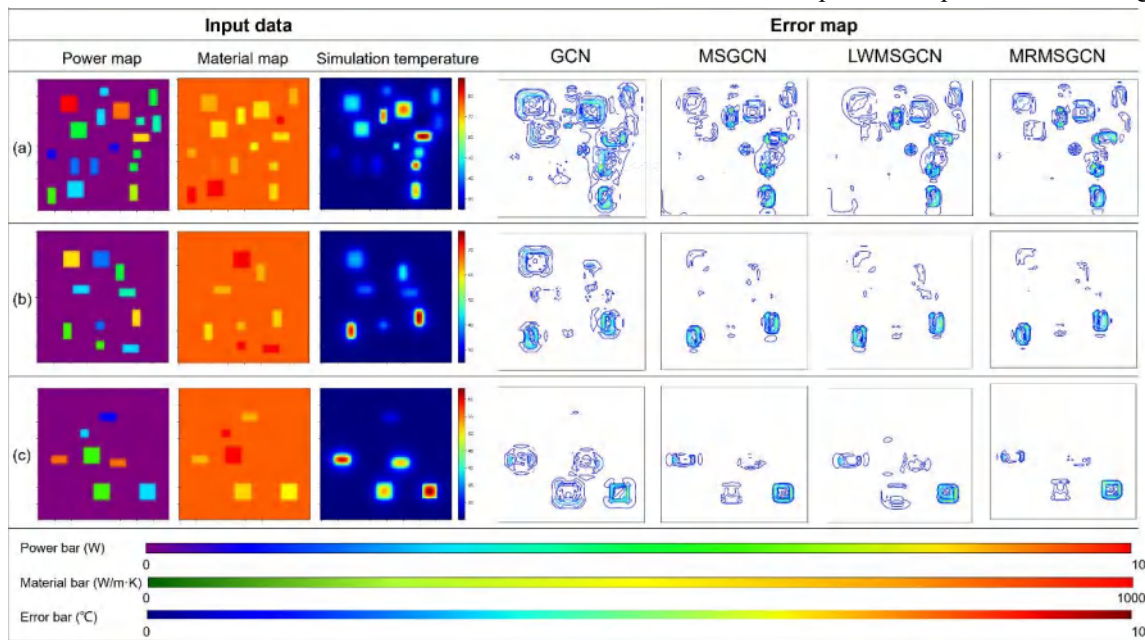


Figure 6 Visualization of a representative sample, including the input power map, material map, simulation temperature distribution, and error maps.

hyperparameter configurations, and then presents a comparative analysis of the experimental results. Three models were trained under identical dataset conditions to ensure fair evaluation. The first baseline, GCN, is a six-layer GCNConv network utilizing fixed single-scale adjacency. The second variant, MSGCN, extends the baseline by introducing multi-scale graph convolutions and applies dropout with a probability of 0.1 to enhance generalization performance. The third model, referred to as MRMSGCN, further builds upon MSGCN designed to amplify feature representations across scales, thereby improving the model's ability to capture fine-grained spatial and thermal variations.

Figure 7 presents the training and test loss curves (in logarithmic scale) across 500 epochs for all models. The GCN model exhibited fast convergence during early training but suffered from pronounced overfitting, as indicated by the divergence between training and test loss after 100 epochs. In contrast, both MSGCN and the proposed MRMSGCN demonstrate steadily decreasing loss throughout training, with MRMSGCN achieving the lowest and most stable test loss. This behavior highlights the benefits of introducing multi-scale graph connectivity, which enables the models to more effectively capture both local and global thermal diffusion patterns, thereby

reducing overfitting and improving generalization.

Table 4 Mean prediction errors ($^{\circ}\text{C}$) across models and test samples.

Model	Mean Error ($^{\circ}\text{C}$)				
	Testdataset	a	b	c	d
GCN	0.96	0.81	0.51	0.69	0.53
MSGCN	0.74	0.67	0.39	0.56	0.41
LWMSGCN	0.72	0.63	0.39	0.48	0.42
MSMSGCN	0.68	0.52	0.28	0.37	0.23

Figure 6 illustrates visual comparisons across four test samples for different models, where each row corresponds to an independent test case with inputs comprising a power density map and a material distribution map. The outputs include the ground-truth map.

steady-state temperature field, the predicted temperature distribution, and the corresponding error contours, while Table 4 summarizes the mean and maximum prediction errors. The baseline GCN, employing a single-scale adjacency ($s=1$), captures the coarse heat distribution but fails to resolve sharp thermal gradients near high-power regions and material discontinuities, leading to notable boundary errors. In contrast, the MSGCN leverages multi-scale neighborhood aggregation to better capture both local and global thermal interactions, and the error maps clearly show reduced boundary errors, particularly along chiplet edges where steep temperature gradients occur. Building upon this, the proposed MRMSGCN integrates multi-scale decoding that fuses features from $s=1,2,3$, further enhancing reconstruction fidelity and achieving the lowest mean error of 0.68°C , significantly outperforming GCN (0.96°C).

Table 5 Inference speed comparison of different models.

Model	parameters	Mean inference time (s)	Speed-up ratio
Ansys AEDT	--	62.5	1.0 (reference)
CNN	728577	0.36	174.6
GCN	329857	0.45	138.6
MSGCN	114241	1.76	35.45

LWMSGCN	114241	1.82	34.34
MRMSGCN	419521	7.62	8.2

Table 5 summarizes the inference speed and parameter scale of different thermal prediction models compared to Ansys AEDT, which is used as the reference for computational time. The baseline GCN achieves an inference time of 0.45 seconds per sample, offering a speed-up of approximately 138 \times over Ansys AEDT while

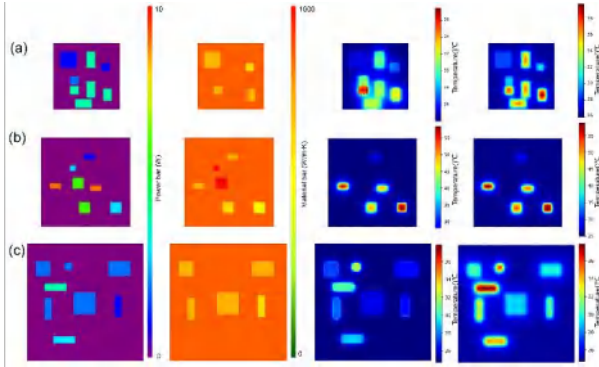


Figure 8 Prediction performance of the proposed MRMSGCN across different chip sizes. (a) 4 * 4 mm chip; (b) 8 * 8 mm chip; (c) 12 * 12 mm chip. The predicted temperature fields show close alignment with the simulated ground truth across all sizes, demonstrating strong size-invariant generalization.

maintaining a moderate parameter count. The MSGCN model, with its multi-scale feature aggregation, attains an inference time of 1.76 seconds and a speed-up ratio of 35 \times , reflecting the computational overhead introduced by parallel multi-scale processing. The proposed MRMSGCN, incorporating both multi-scale connectivity and multi-rate feature amplification, requires 7.62 seconds per sample—still over 8 \times faster than Ansys AEDT—while delivering the highest prediction accuracy. These results demonstrate that all GCN-based models achieve substantial efficiency gains compared to traditional simulation tools, with MRMSGCN balancing accuracy and computational cost, and GCN providing an extremely lightweight and rapid option suitable for early-stage design iterations.

Table 6 Prediction performance of the proposed MRMSGCN across different chip sizes. (a) 4*4 mm chip; (b) 8*8 mm chip; (c) 12*12 mm chip.

Chip Size (mm ²)	Mean Error (°C)	Tmin Error (°C)	Tmax error Error (°C)
4*4	2.02	0.47	3.33
8*8	0.68	0.30	0.92
12*12	1.23	0.30	10.55

To further validate the generalization capability of our model, we evaluated its prediction performance across chip sizes of 4 * 4 mm, 8 * 8 mm, and 12 * 12 mm. As illustrated in Figure 8, visual comparisons show that the predicted temperature fields closely with the ground-truth distributions across all chip sizes. The quantitative results, summarized in Table 6, indicate that the proposed MRMSGCN maintains consistently low mean errors (0.68–2.02 °C) across different chip dimensions, demonstrating the model’s generalization capability for chip layouts of varying scales.

IV. Conclusion

In this work, we present a novel thermal prediction framework based on a MRMSGCN tailored for heterogeneous multi-chip systems. By leveraging fixed-size pixel nodes and multi-scale graph edge constructions, the model effectively captures both local heat diffusion and global thermal coupling across diverse chiplet configurations. Extensive experiments on a dataset comprising over 200 physically simulated chip arrangements demonstrate that the proposed architecture significantly outperforms conventional GCN baselines in both accuracy and generalization. With an average prediction error of 0.68°C and inference time of 7.62 seconds per sample, our method achieves an optimal trade-off between computational efficiency and predictive fidelity. These results underscore the viability of multi-scale graph-based learning for real-time thermal-aware optimization in large-scale integrated system design.

Acknowledgement

The project is supported by the Peng Cheng Laboratory, Peng Cheng Cloud-Brain.

References

- [1] Marcin Szermer, Piotr Zajac, Lukasz Kotynia, Cezary Maj, Piotr Pietrzak, Michal Janicki, et al., “New methodology for thermal analysis of multi-core processors based on dedicated ASIC,” *Microelectronics Journal*, vol. 44, no. 7, pp. 623–630, DOI: 10.1016/j.mejo.2013.02.006, 2013.
- [2] Chen Wang, Qinghua Xu, Chao Nie, Haoran Cao, Junjie Liu, Dongkai Zhang, et al., “A Multiscale Anisotropic Thermal Model of Chiplet Heterogeneous Integration System,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 32, no. 1, pp. 178–190, DOI: 10.1109/TVLSI.2023.3321933, 2024.
- [3] Pengfei Zhang, Di-Wei Wang, and Wei-Sheng Zhao, “A Thermal and Power Integrity Co-Optimization Framework for 2.5-D Integrated Microsystem,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, in press, DOI: 10.1109/TCSI.2024.3454628, 2024.
- [4] Hao Sun, Di Cang, Zheng Dong, Jian Zhao, and Zhiwei Cai, “Enhanced Junction Temperature Prediction Model for CoWoS Packaging With Multiple Chiplets,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 14, no. 10, pp. 1783–1791, DOI: 10.1109/TCPMT.2024.3451136, 2024.
- [5] Yuxuan Chen, Dong Zhao, Feng Liu, Jun Gao, and Hao Zhu, “Thermal layout optimization for 3D stacked multichip modules,” *Microelectronics Journal*, vol. 139, Art. no. 105882, DOI: <https://doi.org/10.1016/j.mejo.2023.105882>, 2023.
- [6] Yifan Wang, Guojun Liu, Hao Wang, and Chuan Peng, “Analytical thermal resistance network model for calculating each mean die temperature of the multi-chips module combined with quad flat no-leads packaging,” *AIP Advances*, vol. 13, no. 8, Art. no. 085319, DOI: <https://doi.org/10.1063/5.0152571>, 2023.
- [7] Yifan Zhang, Yu Zhao, and Jun Wang, “Thermal interaction and cooling of electronic device with chiplet–2.5D integration,” *Applied Sciences*, vol. 14, 2024.
- [8] Wei Liu, Alessandro Calimera, Enrico Macii, Andrea Macii, Alberto Nannarelli, and Massimo Poncino, “Layout-Driven Post-Placement Techniques for Temperature Reduction and Thermal Gradient Minimization,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 3, pp. 406–419, DOI: 10.1109/TCAD.2012.2228267, 2013.
- [9] Yiyu Zhan and Sachin S. Sapatnekar, “High-Efficiency Green Function-Based Thermal Simulation Algorithms,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 9, pp. 1661–1675, DOI: 10.1109/TCAD.2007.895754, 2007.
- [10] Lin Chen, Weiyu Jin, and Sheldon X.-D. Tan, “Fast Thermal Analysis for Chiplet Design based on Graph Convolution Networks,” in

Proceedings of the 27th Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE, Tokyo, Japan, pp. 485–490, DOI: 10.1109/ASP-DAC52403.2022.9712583, 2022.

[11] Lizuo Liu, Bo Wang, and Wei Cai, “Linearized Learning Methods with Multiscale Deep Neural Networks for Stationary Navier-Stokes Equations with Oscillatory Solutions,” ArXiv e-Print, arXiv:2102.03293, DOI: N/A, 2021.

[12] Yu-Hao Lin, et al., “Power envelope analyses of chiplet module and system-on-chip,” in Proceedings of the International Microelectronics Assembly and Packaging Society (IMAPS), IMAPS, USA, Available at: <https://imapsource.org/api/v1/articles/94297-power-envelope-analyses-of-chiplet-module-and-system-on-chip.pdf>, 2023.

[13] Seong-Hwan Kim, Dong-Hwan Park, and Joon Cho, “Lightweight chiplet-based design framework for low-power edge systems,” *Integration*, vol. 98, Art. no. 102813, DOI: 10.1016/j.vlsi.2024.102813, 2024.

[14] “Die-to-wafer bonding solutions,” *Chip Scale Review*, Available at: N/A, 2022.

[15] John Doe, et al., “Manticore: A 4096-core RISC-V chiplet architecture for ultra-efficient floating-point computing,” ArXiv e-Print, 2020.

基于 Verilog_A 的系统级封装电路建模与仿真

喻依虎，张中，李靖，吴克军，宁宁，于奇

(电子科技大学，成都，611731)

摘要：本文基于 Verilog-A 硬件描述语言，针对系统级封装 (SiP) 中 ADC+FPGA+DAC 电路的电气性能仿真需求，提出了一种结合寄生参数建模与非理想因素分析的联合仿真方法。通过提取关键信号线的 RLC 寄生参数 (电阻、电感、电容) 并集成到 Verilog-A 模型中，实现了对 ADC 关键性能指标——无杂散动态范围 (SFDR) 的定量评估。仿真结果与实际测试对比显示，SFDR 预测误差仅为 3.43dB，误差范围小于 10dB，验证了模型的有效性和准确性，为模块系统级仿真提供理论支撑，有效减少 SiP 封装模组设计周期。

关键词：系统级封装 (SiP)；Verilog-A 建模；寄生参数；无杂散动态范围 (SFDR)

Modeling and Simulation of System-in-Package Circuits Based on Verilog-A

Yihu Yu, Zhong Zhang, Jing Li, Kejun WU, Ning NING, Qi YU

(University of Electronic Science and Technology of China, chengdu, 611731)

Abstract : This study presents an enhanced co-simulation framework leveraging Verilog-A hardware description language to address the electrical performance characterization challenges of System-in-Package (SiP) integrated ADC+FPGA+DAC circuits. Through systematic extraction and modeling of RLC parasitic parameters (resistance, inductance, capacitance) from critical interconnect traces, the proposed methodology enables precise prediction of the ADC's Spurious-Free Dynamic Range (SFDR). Cross-validation between simulation results and empirical measurements demonstrates exceptional agreement, with an SFDR prediction error limited to 3.43dB (within 10dB tolerance). The validated model provides a robust theoretical foundation for module-level system simulation, significantly reducing the design iteration cycle for SiP packaging modules.

Key words : System-in-Package (SiP); Verilog-A Modeling; Spurious-Free Dynamic Range (SFDR)

1 前言

系统级封装 (System in Package, SiP) 技术通过将多芯片/芯粒 (Chiplet) 集成于单一封装内，实现了高密度互连与异构集成。随着先进 Sip 技术的不断发展，

在同一个封装中集成的芯粒越来越多，同时走线密度也越来越高^[1]。在此情况下，对系统级封装电路的电气性能的仿真的难度和复杂度都达到了前所未有的程度。

在 SiP 电路设计中，通过高精度电气性能仿真能够有效验证系统功能完整性并优化关键性能指标，这对确保设计可靠性至关重要。传统基于 S 参数的评估方法因其便捷性成为主流，但其对无杂散动态范围（SFDR）、有效位数（ENOB）等关键电路性能指标的量化不足逐渐显现。本文创新性地采用 Verilog-A 硬件描述语言，构建了包含 RLC 寄生参数与器件核心非理想性（噪声、电容失配、时钟抖动等）的联合仿真模型。以 16 位 100MSPS ADC+FPGA+DAC 系统为验证平台，实现了 SFDR 仿真预测，其与实测结果的误差小于 10dB（具体为 3.43dB），充分验证了该模型的有效性和准确性。

2 系统级封装的 Verilog-A 建模方法

Verilog-A 是一种专门用于描述模拟电路系统和单元模块的硬件描述语言，通过模块化方式精确描述电路的结构、行为特性以及性能参数^[2]。本文研究的 SiP 电路核心由 16 位 100MSPS A/D 转换器（ADC）、FPGA 及 16 位 100MSPS 电流舵结构的 D/A 转换器（DAC）构成，其封装系统框图如图 1 所示。为精准评估 SiP 电路的关键性能指标并降低加工迭代成本，本文采用 Verilog-A 语言对模数转换器（ADC）核心模块及高速信号走线的寄生参数（如寄生电容、电阻、电感）进行精细化建模。在模型构建过程中，全面纳入了实际 ADC 架构中的噪声源（热噪声、量化噪声等）与非理想因素（如非线性失真、时钟抖动等）的作用机制。通过仿真与参数化扫描，实现对封装后电路实测性能（如动态范围）的高保真度还原，从而为 SiP 设计提供可量化的优化依据。

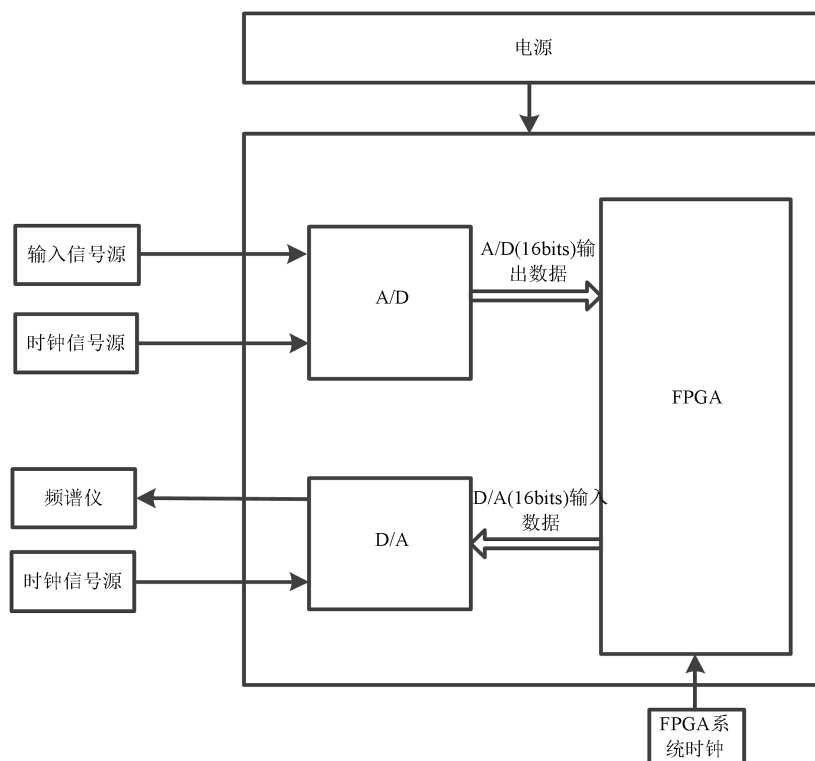


图 1 Sip 封装电路系统框图

2.1 ADC 建模与非理想因素分析

选用的目标 ADC 为 5+5+5+5+4 的五级流水线结构^[3]。其 ADC 整体电路架构如图 2 所示。每一级 MDAC 包含一个采保电路，一组子 ADC 与 DAC 和一个增益为 8 倍的闭环残差放大器。整体上，该 16 位 ADC 第一级(Stage1,STG1) MDAC 到第四级(Stage4,STG4)每级提供 5bit 码字输出，包含 1.5bit 冗余，用于校正各级比较器失调，并提供 3bit 有效二进制码字输出。第五级(Stage5,STG5)为 4bit 的 flash 模数转换器。五级流水线的输出通过数字码字对齐电路，最终实现 16bit 码字分辨率。

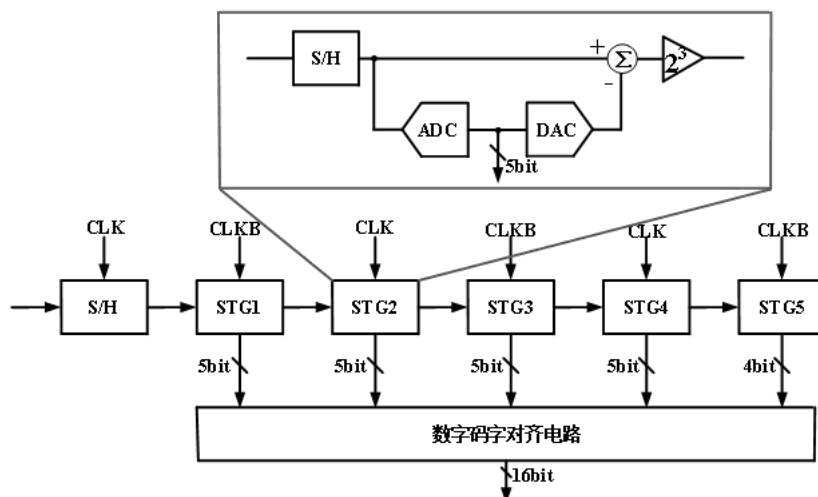


图 2 16 位流水线 ADC 整体电路架构

为精确表征 ADC 性能，在建立其基本功能模块模型的基础上，本文重点分析并建模了影响其性能的关键非理想因素，包括：噪声、电容失配及时钟抖动。

2.1.1 噪声建模

在模数转换器中，热噪声是系统噪声的主要来源。热噪声主要由采样开关网络和运算放大器器件热噪声产生,其等效电路如图 3 所示。

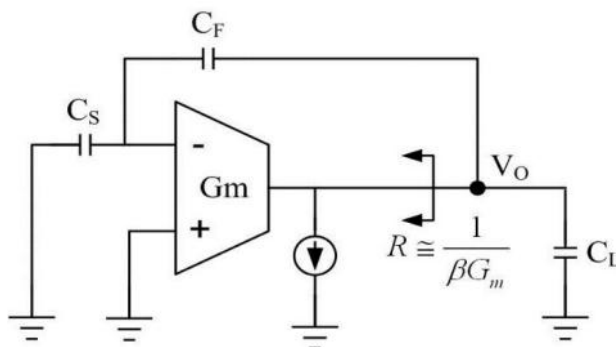


图 3 运放噪声等效电路

采样开关热噪声等效到输入端的噪声功率为 KT/C_s (C_s 为采样电容)。在保持阶段，量化器噪声主要来源于残差放大器，残差放大器的噪声主要受噪声源系数和热噪声系数影响。其噪声功率密度为：

$$P_{noise} = N_{op} \cdot 4KT\gamma G_m \cdot \Delta f \quad (1)$$

N_{op} 为残差运放的噪声源系数，与运放的结构和 MOS 管的跨导有关； γ 为热噪声系数，与工艺有关，在长沟道工艺中为 $2/3$ ； K 为波尔兹曼常数； T 为温度； β 为运放的反馈系数， Δf 为残差运放带宽。

在 Verilog-A 模型中，采样量化过程首先判断输入信号所在的电压区间，再根据传输特性曲线确定输出残差。为了模拟采样噪声影响，在输入信号上叠加满足正态分布，方差为采样噪声的随机电压量。同时采样噪声也会对残差信号产生一定影响，因此在传输特性曲线表达式中的输入信号变量上也要叠加采样噪声。

2.1.2 电容失配建模

电容失配是影响流水线模数转换器 (Pipeline ADC) 性能的一个关键因素。流水线 ADC 通常由多个级联的子级组成，每一级包括一个子 ADC 和一个残差放大器 (通常是由乘法 DAC 结构实现的 MDAC)。在这些结构中，电容是实现采样、数模转换和残差放大的核心元件。由于集成电路制造过程中的工艺波动，实际电容值与设计值会存在细微差异，即电容失配。这种失配虽微小，却会在高

精度 ADC 中引起显著的性能劣化。从作用机理来看，电容失配主要通过两个维度影响系统性能：

首先，电容失配会直接导致量化非线性。流水线 ADC 的 MDAC 利用电容网络来实现精确的数模转换，当电容比值偏离理想值时，DAC 输出电压出现误差。这种误差不仅导致转换结果的偏移，还会破坏每一级转换的线性特性，从而引起整个系统的 DNL（微分非线性）和 INL（积分非线性）恶化。尤其在高分辨率 ADC 中，哪怕极小的失配也足以造成严重的非线性问题。

其次，电容失配会影响残差的准确计算。MDAC 的作用是对每一级的输入信号减去已转换部分并放大剩余信号（即残差），传递给下一级进行进一步转换。如果电容失配，残差的计算和放大就会发生误差，这些误差会随着级数的增加而逐级放大，最终影响整个 ADC 的准确性。这不仅降低了系统的线性度，还可能引入额外的噪声，导致 SNR（信噪比）下降。

以 1.5bit 流水线级为例，假设 $C_1=C_2+\Delta C$ ，则

$$\frac{C_1}{C_2} = \frac{C_2 + \Delta C}{C_2} = 1 + \frac{\Delta C}{C_2} \quad (2)$$

带入传输函数可得：

$$V_{OUT} = \begin{cases} (2 + \frac{\Delta C}{C_2})V_{in} - (1 + \frac{\Delta C}{C_2})V_{ref} & (V_{in} > \frac{1}{4}V_{ref}) \\ (2 + \frac{\Delta C}{C_2})V_{in} & (-\frac{1}{4}V_{ref} < V_{in} < \frac{1}{4}V_{ref}) \\ (2 + \frac{\Delta C}{C_2})V_{in} + (1 - \frac{\Delta C}{C_2})V_{ref} & (V_{in} < -\frac{1}{4}V_{ref}) \end{cases} \quad (3)$$

由上式可以看出，电容失配不仅会引入增益误差，而且会引入子 DAC 参考电压失调。利用 verilog-A 中内建函数 `rdist_normal` 对电容失配进行建模，代码如下：

```
`define mis_cap(I) (0.001 * ($rdist_normal (I , 0.0 , 1.0) ) * mis_cap_en * c0)。
```

其中 `mis_cap_en` 为电容失配使能开关，`c0` 为选取的电容容值。

2.1.3 时钟抖动建模

时钟抖动（clock jitter）是指时钟信号边缘在理想时间位置附近的随机或周期性偏移。在流水线 ADC 中，时钟不仅用于控制采样保持电路（S/H）对输入

信号的采样，还用于控制各级之间的转换与数据传递。因此，时钟抖动会对流水线 ADC 的精度和性能造成重要影响，尤其在高速、高分辨率应用中表现尤为明显。

首先，时钟抖动对模拟输入信号的采样精度构成直接威胁。流水线 ADC 中的首级采样保持电路通常决定整个系统的信号捕获精度。如果采样时钟存在抖动，那么实际采样时刻将偏离理想值，导致输入信号在电压上被误采样。这种误差与输入信号的斜率 (dV/dt) 成正比，因此在高频或快速变化的输入信号下，时钟抖动造成的采样误差会显著上升，最终转化为系统噪声。其次，时钟抖动导致的采样误差会表现为随机噪声成分，进而影响流水线 ADC 的 SNR (信噪比)。SNR 的下降将直接影响到 ADC 的 ENOB (有效位数)，通常可通过下式估算抖动引起的 SNR 损失：

$$SNR_{jitter} = -20 \log(2\pi f_{in} t_{jitter}) \quad (4)$$

其中 f_{in} 是输入信号频率， t_{jitter} 是时钟抖动均方根值 (RMS)。可见，在高频输入信号条件下，即使抖动值很小，也可能造成多位 ENOB 的损失。此外，虽然流水线 ADC 中后续级的误差对系统性能影响相对较小，但如果其内部时钟或控制逻辑也受到抖动影响，可能导致级间数据传输或采样不稳定，进而引发非线性失真或系统性的失步错误。

使用 Verilog-A 的 `cross` 函数控制采样与量化时机 (在时钟越过阈值瞬间触发)。通过在时钟触发电平上叠加满足高斯分布的随机电压 V_{rand} 来等效时钟抖动。等效抖动值 ($jitter$) 计算为：

$$jitter = \frac{T_{rise} V_{rand}}{V_{clk}} \quad (5)$$

其中 T_{rise} 为时钟上升时间， V_{clk} 时钟幅值。Verilog-A 代码如下：

```
(cross(V(clk)-vtrans_clk+jitter*$rdist_normal(seed,0,1),1,1,clk.potential.abstol))
```

当 cross 函数表达式等于 0 时，即 $V(\text{clk}) = vtrans_clk$ 时执行后续程序。通过在 $vtrans_clk$ 上叠加高斯分布的随机数改变触发的时间间隔，模拟时钟抖动。

2.2 互连线 RLC 寄生参数建模

在超高速、高分辨率流水线 ADC 设计中，模拟前端与参考电压端的 RLC 寄生参数已成为制约性能提升的关键因素^[4-6]。RLC 寄生网络主要产生于 BGA 封装、PCB 走线和芯片级互连。当 ADC 工作频率较高或信号带宽较宽时，这些非

理想寄生参数将对 ADC 性能产生复杂而显著的影响。首先，RLC 寄生网络会对输入信号形成一个低通滤波器效应，限制信号的带宽和上升沿速度。对于流水线 ADC 而言，首级采样保持电路负责捕获外部输入信号。此时，输入端口的寄生电感和电容将与采样开关、电容网络共同构成一个复杂的二阶甚至更高阶网络，导致信号无法在期望的时间窗口内稳定采样，尤其在高速信号或高频正弦波输入条件下更为严重。这种信号畸变直接引起采样误差，进而表现为 SNR 下降和 THD（总谐波失真）上升。其次，输入寄生参数还会影响 ADC 的线性度。寄生电感可能在高速切换时引发电压尖峰（overshoot）或振荡，导致信号波形扭曲；寄生电阻则可能引起压降，降低输入信号幅度的准确性。特别是在多通道系统中，寄生参数差异还会造成通道间失配，进一步影响系统的一致性和可用性。金属层走线的 RLC 寄生参数的等效电路如图 4 所示：

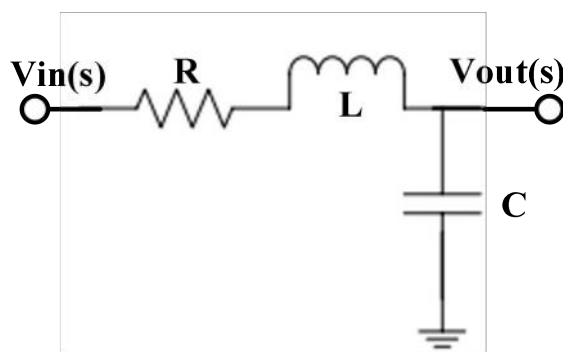


图 4 RLC 寄生参数等效电路

其 RLC 单元模型传输函数为：

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{1}{1 + RC_s + LC_s^2} \quad (6)$$

在 verilog-A 建模中，本文利用其自带的拉普拉斯函数进行建模，Verilog-A 代码如下：

```
V(out) <+ laplace_nd(V(in), {1,0,0}, {1,R*C,L*C}).
```

3 系统建模仿真与测试验证

通过 PowerDC 提取 ADC 的信号线级、时钟线从封装引脚到 ADC 管芯 PAD 上的寄生参数，其 RLC 寄生参数提取路径示意图如图 5 所示：

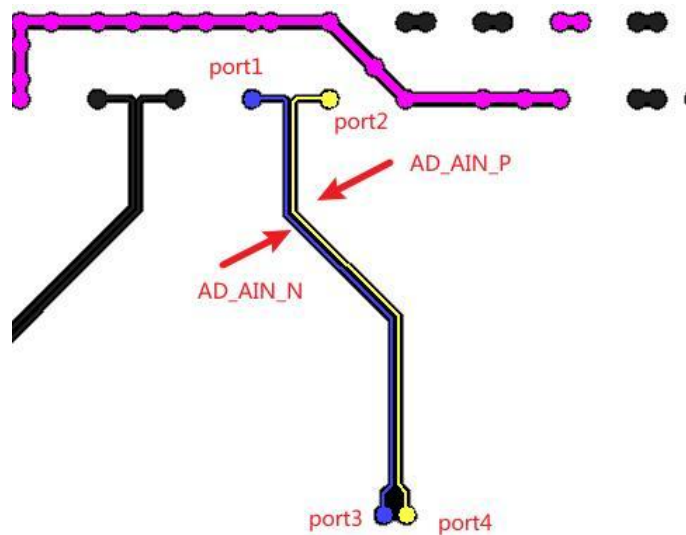


图 5 RLC 寄生参数提取路径示意图

将 RLC 寄生参数添加在系统级封装电路的仿真模型中，如图 6 所示。通过同位宽的理想 DAC 对 ADC 的输出信号进行转换并进行快速傅里叶变化测试。从而实现对系统级封装后的 ADC 的无杂散动态范围（SFDR）的仿真。



图 6 系统级封装电路仿真模型

在本文中实现对封装完成后的微系统实物样件进行测试，其测试环境如图 7 所示：

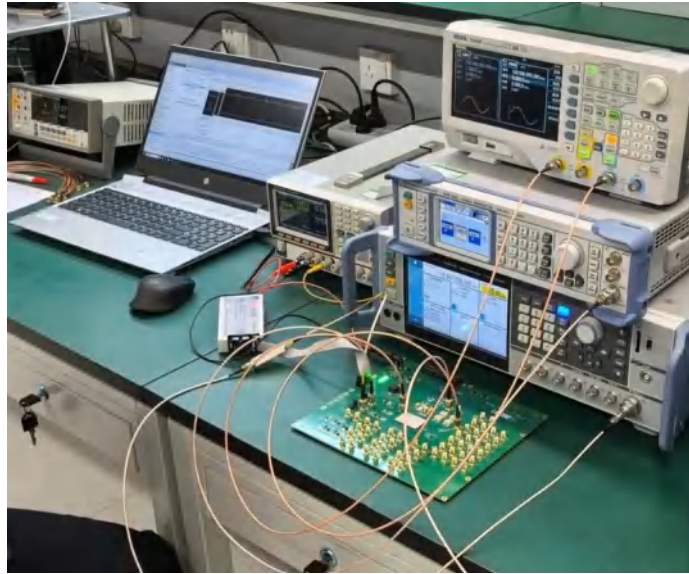
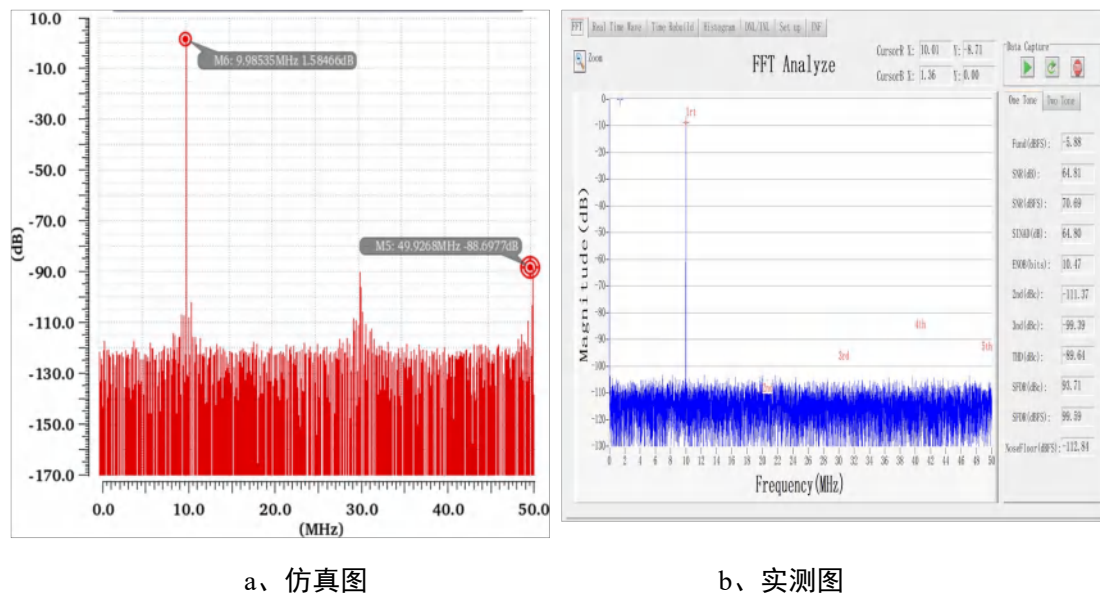


图 7 实物样件测试环境

信号源产生频率为 10MHz，功率为 10dBm 的模拟信号，并通过中心频率为 10MHz，带宽为 5MHz 的带通滤波器将信号输入到微系统中，通过微系统中的 FPGA 采集 ADC 的输出信号并进行快速傅里叶变化。其仿真值与实际测试值的对比，如图 8 所示：



a、仿真图

b、实测图

图 8 仿真与实测对比图

从图 8 中可以看出，包含 RLC 寄生参数的建模仿真的无杂散动态范围（SFDR）为 90.28dB；对系统级封装完成后的实物样件进行测试，其无杂散动态范围（SFDR）为 93.71dB，实测值与仿真值相差 3.43dB，误差范围小于 10dB。验证了该仿真建模的有效性。

4 结论

本文基于 Verilog-A 硬件描述语言，针对系统级封装（SiP）中 ADC+FPGA+DAC 电路的电气性能仿真挑战，提出并实现了一种有效的建模与仿真方法。该方法的核心在于：

(1) 联合建模：深入分析了影响 ADC 性能的关键非理想因素（噪声、电容失配、时钟抖动）并成功集成到 Verilog-A 模型中。

(2) 寄生参数集成：精确提取了 SiP 互连线（含关键信号线与时钟线）的 RLC 寄生参数，并将其纳入系统级仿真模型。

(3) SFDR 定量评估：通过构建包含非理想 ADC、RLC 寄生模型、理想 DAC 及 FFT 分析的完整仿真流程，实现了对封装后 ADC SFDR 的定量预测。

实验验证表明，基于该模型的 SFDR 仿真结果（90.28dB）与实物样件实测结果（93.71dB）的误差仅为 3.43dB，显著低于 10dB 的预期目标，充分证明了所提模型的有效性和准确性。

参考文献

- [1] Z. Zhang, M. Miao and X. Duan, "Compact Modeling and Analysis of a Typical Inter-Chiplet Serial High Speed Link on an Active Interposer," 2021 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), Zhuhai, China, 2021, pp. 65-68
- [2] V. Y. Ponce-Hinestroza and V. R. Gonzalez-Diaz, "System-Level Behavioral Model of a 12-Bit 1.5-Bit Per Stage Pipelined ADC Based on Verilog®-AMS," 2018 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), Prague, Czech Republic, 2018, pp. 301-304.
- [3] D. Osipov and Y. Bocharov, "Behavioral model of split capacitor array DAC for use in SAR ADC design," PRIME 2012; 8th Conference on Ph.D. Research in Microelectronics & Electronics, Aachen, Germany, 2012, pp. 1-4.
- [4] M. Waqas Chaudhary, A. Heinig and M. Dittrich, "Interposer based integration to achieve high speed interfaces for ADC application," 2016 IEEE International 3D Systems Integration Conference (3DIC), San Francisco, CA, USA, 2016, pp. 1-4.
- [5] Y. Zhou, B. Shi, T. Nguyen, H. Sun and J. E. Schutt-Ainé, "Signal and Power Integrity Co-Simulation of Chiplet-to-Chiplet Channel Based on Latency Insertion Method," 2024 IEEE 28th Workshop on Signal and Power Integrity (SPI), Lisbon, Portugal, 2024, pp. 1-4.
- [6] T. -L. Hsieh, H. -C. Kuo, M. -F. Jhong, C. -Y. Huang and C. -C. Wang, "Ultra High Density Package Design and Electrical Analysis in High Performance Computing Application," 2019 IEEE 21st Electronics Packaging Technology Conference (EPTC), Singapore, 2019, pp. 457-460.

领域专用处理器芯片技术及其产业化应用

李炎

复旦大学

摘要：领域专用处理器作为针对特定应用定制的可编程处理器，在能效比和性价比方面显著优于通用处理器，已成为学术界和产业界的技术热点。面向物联网、信息安全、工业控制等领域对芯片算力、功耗和安全性的迫切需求，本研究历时 20 余年，在架构-电路-芯片技术创新和产业化应用方面取得重要突破。在多核/众核处理器架构技术方面，针对通用微处理器效率低而专用芯片灵活性差的问题，创新提出"众多小核+可重构加速器"的领域专用众核处理器架构，发展了多机制融合的芯内/芯间互连技术和处理器核与加速器协同方案，突破异步互连设计及低功耗电路技术，有效解决多核处理器互连瓶颈。在复杂计算硬件加速技术方面，围绕领域专用计算的算力瓶颈，提出密码算法多模态高能效硬件加速器技术，创新基于随机计算的脉冲神经网络加速器，研发可重构低功耗神经计算芯片，开发多种动态防御机制保障信息安全芯片可靠性。在系统芯片集成优化技术方面，提出基于新型寄存器的标准单元库优化设计和双胞胎单元 SRAM 近阈值电压存储技术，显著提升 SoC 能效，创新延时调节技术在关键电路中成功应用，有力保障芯片产品可靠性和安全性。

基于图论的栅介质击穿的渗流模型研究

曲益明

华东师范大学

摘要：针对先进 MOSFET 中 TDDB 可靠性问题，传统渗流模型因假设缺陷均匀生成，无法解释高 k 金属栅（HKMG）的双模寿命分布及关态 TDDB 的非均匀电场效应。本文提出基于图论的矩阵渗流模型，通过邻接矩阵描述任意空间缺陷分布（ λ 矩阵可调），突破均匀性限制，经严格数学推导，成功应用于：1) HKMG 双层堆栈，复现缺陷速率比依赖的双模威布尔分布；2) 关态 TDDB，结合 TCAD 电场预测不同沟长器件寿命，均与实验/仿真吻合。该模型为复杂非均匀 TDDB 机制提供通用解析工具。

Research Article

Cross-Layer Error Modeling and Deployment of Compressed Sensing Object Detection on Compute-in-Memory Systems

Saike Zhu^{1,2}, Xiaoqian Huang³, Yong Gong³, Shifan Gao¹, Xianggao Wang^{1,2}, Hua Chen⁴, Cimang Lu⁵, and Yi Zhao^{1,2,6}

1. *The College of Electronic Engineering and Information Science, Zhejiang University, Hangzhou 310027, China*

2. *The International Joint Innovation Center, Zhejiang University, Haining 314400, China*

3. *China Nanhu Academy of Electronics and Information Technology, Jiaxing 314001, China*

4. *The School of Integrated Circuits, East China Normal University, Shanghai 200241, China*

5. *Flash Billion Semiconductor Company Limited, Shanghai 201210, China*

6. *HuaDa Semiconductor Company Limited, Shanghai 200120, China*

Corresponding author: Yi Zhao; Email: yizhao@zju.edu.cn.

Received June 20, 2025; Accepted March 22, 2022; Published March 22, 2022.

Abstract — Deploying deep neural networks (DNNs) on compute-in-memory (CIM) hardware is challenged by analog non-idealities such as device nonlinearity and circuit-level noise. This paper proposes a cross-layer framework that enables accurate and efficient deployment of a compressed sensing-based object detection model (CSDet) on a CIM platform built with programmable linear random-access memory (PLRAM). The framework incorporates device- and circuit-level error modeling into the training and deployment flow, enabling slope-aware compensation and statistical correction of residual errors. Experimental results show that the calibrated CSDet achieves up to 11.3% improvement in detection accuracy with minimal hardware overhead, demonstrating the practicality of robust inference on non-ideal CIM systems.

Keywords — Compute-in-Memory, NOR-Flash, Compressed Sensing, Error Calibration, Hardware-Aware Deployment.

I. Introduction

The growing adoption of artificial intelligence (AI) at the edge has raised critical demands for efficient and reliable inference under limited computational and energy resources. Traditional von Neumann architectures suffer from the “memory wall,” where frequent data movement between memory and processing units results in significant latency and energy overhead, limiting their suitability for edge AI applications [1–3]. Compute-in-memory (CIM) architectures have emerged as a promising alternative by performing core operations, such as matrix–vector multiplication (MVM), directly within memory arrays. This integration reduces data movement and enables parallel, energy-efficient computation [4–6].

Despite their architectural advantages, CIM systems remain challenged by hardware non-idealities. Device-level variations—including conductance nonlinearity, programming noise, and drift—along with circuit-level issues such as

sensing mismatch and ADC distortions, introduce significant errors during analog computation [7, 8]. These inaccuracies can accumulate and degrade the performance of deep neural networks (DNNs) when deployed directly on such platforms without dedicated compensation mechanisms[9].

Compressed sensing (CS) techniques offer a pathway to lightweight, efficient inference by reducing the dimensionality of input data while preserving essential features[10–13]. Building on this principle, CSDet—a compressed sensing-based object detection architecture has demonstrated strong potential for deployment in resource-constrained environments due to its compact design and reduced computation cost [14]. However, the analog variability inherent in CIM hardware poses a significant obstacle to the reliable deployment of CSDet in practice.

In this work, we propose a cross-layer deployment framework that integrates error modeling and calibration into the hardware mapping process of CSDet on a CIM system built with programmable linear random-access mem-

ory (PLRAM). By characterizing system-level error propagation and incorporating compensation into training and deployment, we enable robust inference on non-ideal hardware without modifying the model structure. This approach provides a practical and generalizable solution for deploying compressed sensing-based models on analog CIM platforms, bridging the gap between algorithm design and deployment on physical hardware.

II. Background and Motivation

1. Analog Non-Idealities in CIM Systems

Analog CIM systems offer substantial performance benefits through in-situ matrix-vector multiplication, yet they suffer from a wide range of hardware-level non-idealities. In particular, the analog nature of computation introduces deviations at multiple levels: device variability leads to nonlinear I-V behavior and unstable conductance states, array-level parasitic effects cause IR drop and signal distortion, and circuit-level mismatches in ADCs and sense amplifiers accumulate to degrade system-level precision. These factors collectively result in significant inference errors that propagate through deep neural networks.

variability, array parasitics, and circuit mismatch. These errors accumulate layer by layer, leading to accuracy degradation. As shown in Figure 1, such effects cause output shifts and reduced correlation between analog and ideal results. While prior work has addressed specific issues—such as IR drop[17], RC delay[18], or device nonlinearity[19]—these efforts remain isolated, lacking a unified modeling approach.

2. CSDet Architecture and Deployment Challenges

CSDet is a lightweight object detection architecture designed for resource-constrained environments. As illustrated in Figure 2, the system consists of two main components: the Compressed Sensing Detection Module (CSDM) and the Compressed Sensing Reconstruction Module (CSRSM). The CSDM performs object detection directly in the compressed domain using a joint multi-channel sampling matrix and a one-stage detection network. Compressed measurements are also forwarded to the CSRSM, which reconstructs high-quality images when needed. In practical applications, the CSDM is deployed on edge devices for real-time sensing and detection, while the CSRSM is typically used on back-end servers for image reconstruction and manual review.

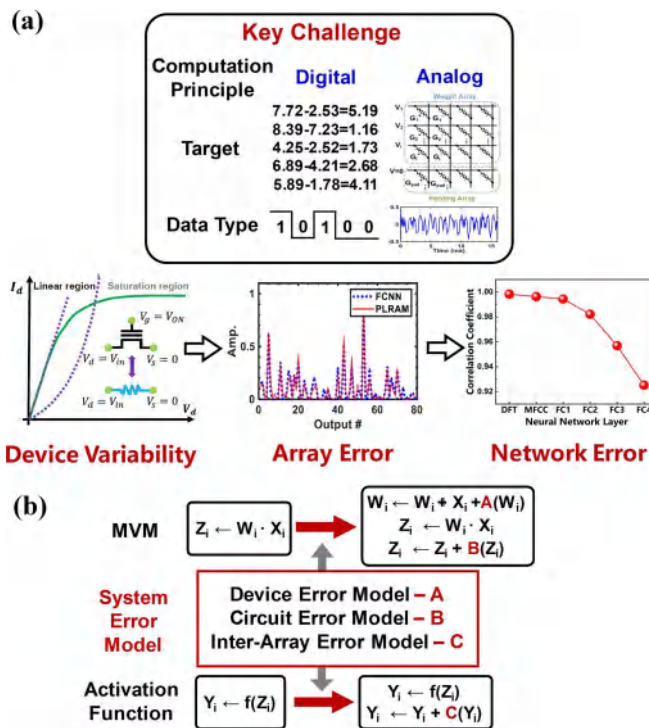


Figure 1 (a) Layer-wise error accumulation in analog CIM systems caused by device variability, array distortion, and circuit mismatch. (b) Hierarchical modeling of device, circuit, and array-level errors.

To address these challenges, it is essential to understand how hardware non-idealities propagate across the hardware–software stack. In PLRAM-based CIM systems[15, 16], deviations from ideal computation result from device

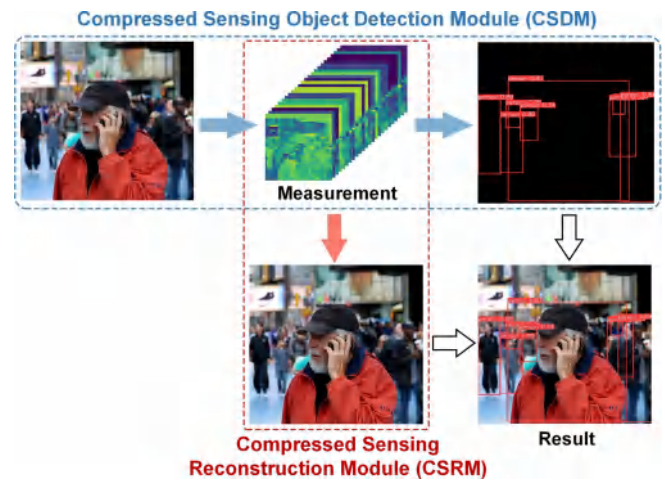


Figure 2 Overview of the CSDet architecture. The CSDM performs object detection in the compressed domain, while the CSRSM reconstructs high-quality images for visualization.

At the core of CSDet lies the compressed sensing formulation, where an input signal x is projected onto a low-dimensional measurement space through a sensing matrix ϕ . Assuming $x = \psi^{-1}s$, where ψ is a known basis and s is sparse in that domain, the measurement process is given by:

$$y = \phi x = \phi \psi^{-1} s \tag{1}$$

CSDet introduces a joint multi-channel sampling matrix, which can be interpreted as a convolution operation. As shown in Figure 3, the convolution process can be expressed as:

$$Y(i, j) = (X * K)(i, j) = \sum_{m=0}^{B-1} \sum_{n=0}^{B-1} X(i+m, j+n)K(m, n) \quad (2)$$

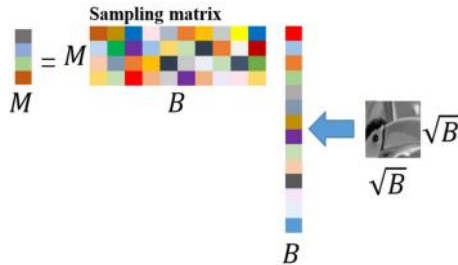


Figure 3 Compressed sensing via a joint multi-channel sampling matrix. $B \times B$ image patch is projected into an M -dimensional measurement vector.

Here, X denotes the input image patch, K the sampling kernel, and B the kernel size. The sampling matrix is trained to preserve task-relevant features. In this framework, the compression ratio R is defined as:

$$R = \frac{M}{B} \quad (3)$$

where M is the number of output channels and B is the number of pixels sampled in the input patch. This flexible design improves both compression efficiency and model performance.

This architecture offers several advantages for edge deployment. By operating in the compressed domain, CS-Det eliminates the computational overhead associated with conventional image reconstruction pipelines. Compared to image-domain detection models, it reduces parameter count by approximately 25% and improves mean Average Precision (mAP) by around 10%, making it highly suitable for deployment on source-limited hardware platforms.

III. Error Modeling for CIM Systems

1. Device-Level Weight Error Modeling

In PLRAM-based CIM systems, the memory device is a modified flash-type cell optimized for in-memory analog computing. For modeling purposes, each cell is represented using a standard MOSFET structure operating in the linear region, where the drain current I_d exhibits an approximately linear response to drain voltages. However, small weights correspond to high conductance values and large drain currents, which can push the device into saturation and introduce non-linearity. Moreover, variations in V_{ds} and V_{gs} further shift the device’s operating point, affecting the effective conductance. This results in a deviation between the programmed weight w and the actual weight w' , which accumulates during inference.

To characterize this behavior, we first model the linear-region current as:

$$I_{ds} = \frac{W}{L} C_{ox} \mu_n V_{gs} - V_t - \frac{m}{2} V_{ds} V_{ds} \quad (4)$$

Based on empirical measurement, we define an effective current–weight relationship:

$$I(w, V_{ds}) = a [f(w) - 0.6V_{ds}] V_{ds} \quad (5)$$

$$f(w) = \frac{kw}{a} + 0.24 \quad (6)$$

where a is a scaling constant capturing device transconductance, and k is the weight-to-conductance conversion coefficient.

Substituting $f(w)$ into the current equation (5) yields the analytical current model:

$$I(w, V_{ds}) = a \left(\frac{kw}{a} + 0.24 - V_{ds} - 0.6V_{ds} \right) V_{ds} \quad (7)$$

Defining the effective conductance $G = I/V_{ds}$ and solving for w' leads to:

$$w' = cw - b \cdot [V_s + 0.6(V_{ds} - 0.4)] \quad (8)$$

Here, c and b are fitted constants that represent the scaling factor of the original weights and the influence of voltage bias, respectively. As shown in Figure 4, the effective weight w' is affected by both the applied voltages V_s and V_{ds} . The model reveals that variations in operating voltages directly impact the actual conductance, and thus must be accounted for during system-level calibration.

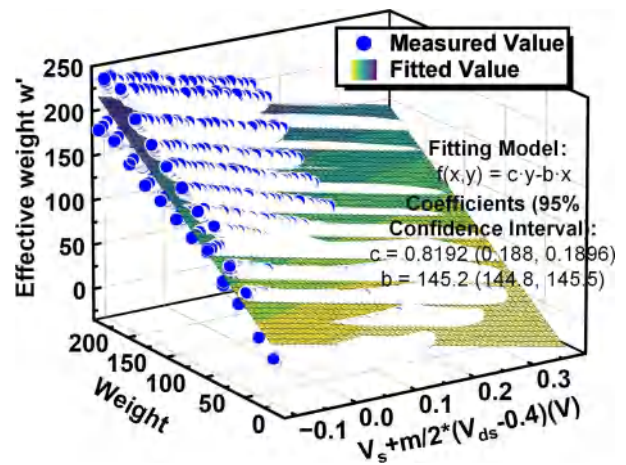


Figure 4 Fitted model of effective weight w' as a function of nominal weight w and bias voltage.

2. Circuit-Level Computation Modeling

In PLRAM-based CIM chips, the computation flow primarily includes three stages: input quantization, matrix-vector multiplication (MVM), and analog-to-digital conversion (ADC). To ensure numerical alignment between these stages, we introduce a circuit-level computation model that uses three scaling factors, denoted as α , β , and γ . These factors map the ideal input, weight, and output values to their corresponding circuit-level representations, facilitating accurate modeling of hardware-induced deviations for error analysis and compensation.

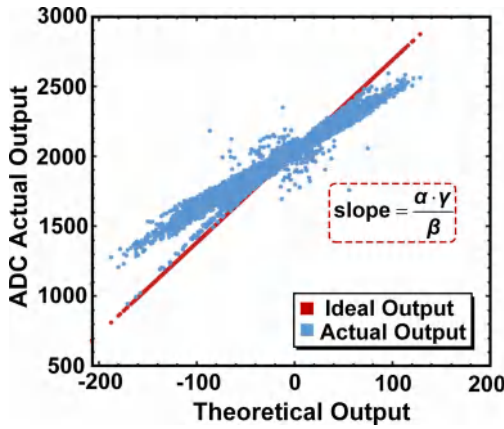


Figure 5 Comparison of ideal MVM outputs and ADC outputs in the PLRAM-based CIM chip.

1) Input Encoding

The quantized input X is mapped to an analog voltage within the chip's operating range, typically [0.1V, 0.75V]. This mapping is controlled by the input scaling factor α , defined as:

$$\alpha = \frac{V_{\max} - V_{\min}}{X_{\max} - X_{\min}} \cdot r \quad (9)$$

where X_{\max} and X_{\min} are the input data range, V_{\max} and V_{\min} are the voltage limits, and r is resolution scaling ratio. The encoded input voltage V_{in} is given by:

$$V_{in} = \alpha \cdot X + V_{\min} \quad (10)$$

2) MVM Computation

The weight matrix W is quantized to 8-bit values in [0, 255] and scaled by a factor β to map weights to conductance. The output voltage $V_{out,j}$ of column j is given by:

$$V_{out,j} = \frac{\sum_i (\alpha \cdot X_i + V_{\min}) \cdot k \cdot W_{ij}}{\sum_i k \cdot W_{ij} + \sum_m k \cdot W_{pad,mj}} \quad (11)$$

where k is the weight-to-conductance conversion factor, and W_{pad} represents padding weights for column balancing. The total conductance of column j is defined as $\beta = \sum_i G_{ij} + \sum_m G_{pad,mj}$, which simplifies (11) to:

$$V_{out,j} = \frac{\alpha}{\beta} \cdot \sum_i X_i \cdot W_{ij} \quad (12)$$

3) Output Decoding

The analog output voltage $V_{out,j}$ is amplified by a factor of 4 and digitized by an n -bit ADC, which maps [0, 1.5 V] to the digital range [0, 2^n]. The scaling factor γ , accounting for both amplification and ADC resolution, is defined as $\gamma = (2^n \cdot 4)/1.5$. The resulting digital output is:

$$ADC_{out} = \gamma \cdot V_{out} \quad (13)$$

To align hardware outputs with digital neural network computations, inverse scaling is applied using the three factors α , β , and γ , restoring the result to the theoretical value domain:

$$Y_j = \sum_i X_i \cdot W_{ij} = \frac{\beta}{\alpha \cdot \gamma} \cdot ADC_{out} \quad (14)$$

As shown in Figure 5, the actual ADC outputs exhibit a linear relationship with the theoretical MVM results, with a slope determined by the scaling factors $\alpha \cdot \gamma / \beta$. This confirms the validity of the proposed mapping model, while the observed deviations reflect inherent hardware-induced errors.

This circuit-level model captures the entire analog computation process, from input encoding to output decoding, and defines a numerical mapping between hardware-level signals and ideal operations. It serves as the basis for subsequent error modeling and hardware-aware inference.

IV. Error Compensation and CSDet Deployment on CIM System

1. Slope Compensation for ADC Output Alignment

Although the circuit-level model defines a linear mapping from theoretical MVM results to ADC outputs, hardware measurements reveal a consistent slope deviation due to mismatches in scaling factors and non-idealities in conductance programming.

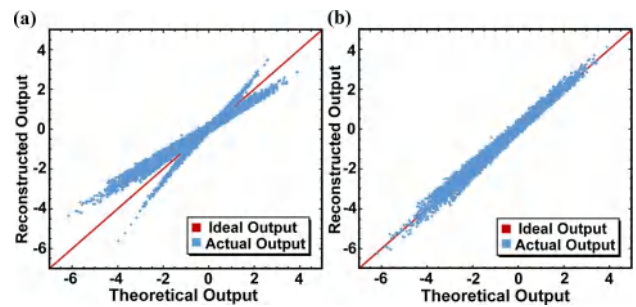


Figure 6 Comparison between chip outputs and theoretical MVM results. (a) Before slope correction. (b) After slope correction.

To correct this deviation, we introduce a slope factor ε , defined as the ratio of the measured slope to the theoretical slope:

$$\varepsilon = \frac{\text{slope}_{\text{chip}}}{\text{slope}_{\text{model}}} \quad (15)$$

Each ADC output y_j is scaled by the fitted slope factor ε to align with the theoretical model:

$$\tilde{y}_j = \varepsilon y_j \quad (16)$$

Here, y_j denotes the original ADC output from the chip, and \tilde{y}_j is the corrected value aligned to the model-predicted MVM result. As shown in Figure 6, this correction reduces systematic deviation and improves consistency between hardware outputs and ideal results.

2. Statistical Modeling of Residual Errors

Although slope correction significantly reduces systematic bias, residual deviations remain due to analog non-idealities in conductance mapping and ADC quantization. To characterize these errors, we analyze the distribution of output deviations after compensation.

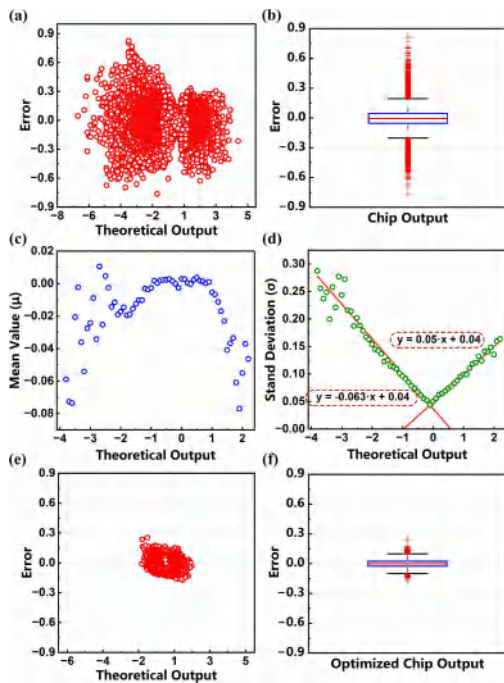


Figure 7 Statistical characteristics of residual errors after slope compensation. (a) Residual scatter plot. (b) Boxplot of residuals. (c) Distribution of residual means. (d) Distribution and fitting of residual standard deviations. (e) Residual scatter after optimization. (f) Boxplot after optimization.

Figure 7(a) and (b) show the scatter and boxplot of residual errors, revealing an approximately zero-mean and bounded distribution. Further analysis in Figure 7(c) and (d) shows that the mean remains close to zero across different output levels, while the standard deviation exhibits a piecewise linear relationship with the output.

Based on this, we model the residual error as zero-mean Gaussian noise with an output-dependent variance:

$$Y_j^{\text{train}} = Y_j + \mathcal{N}(0, \sigma(Y_j)) \quad (17)$$

where $\sigma(Y_j)$ is the standard deviation estimated from hardware measurements. This noise model is injected into the training process to simulate hardware uncertainty and improve deployment robustness. Experimental results show that this strategy reduces output bias by 63.6% and narrows the residual range, as illustrated in Figure 7(e) and (f).

3. CSDet Deployment and Performance Evaluation

To validate the proposed approach in a real hardware environment, the CSDet system is deployed on a PLRAM-based CIM platform with tasks distributed across the chip, host processor, and backend server, based on their computational roles.

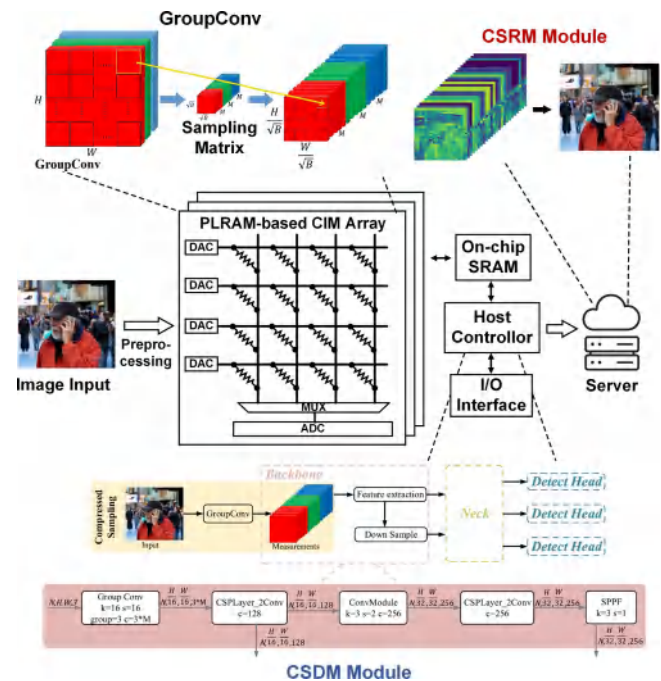


Figure 8 System-level deployment of CSDet on a PLRAM-based CIM platform.

As illustrated in Figure 8, raw image data is first captured by a CMOS image sensor and transferred to the CIM chip via a high-speed parallel interface. The chip performs the core compression stage using a joint multi-channel sampling matrix implemented as a group convolution. This module transforms the input into a compact sparse representation, signif-

icantly reducing the data volume and computational load for subsequent processing.

The compressed features are then transmitted to the host processor, which carries out the CSDM module. When needed, full-resolution images are reconstructed by the CSR module on a backend server to support visualization or manual inspection. This deployment scheme integrates analog compression with digital inference to support full-system execution.

To evaluate the performance of CSDet under real hardware deployment conditions, we conducted experiments on the PLRAM-based CIM chip. As shown in Figure 9, the proposed error modeling strategy from Section III was applied to correct deviations in ADC outputs. Figure 9 (a) shows improved correlation across RGB channels after compensation, while Figure 9 (b) presents a clear reduction in output error bias and variance, validating the effectiveness of the method.

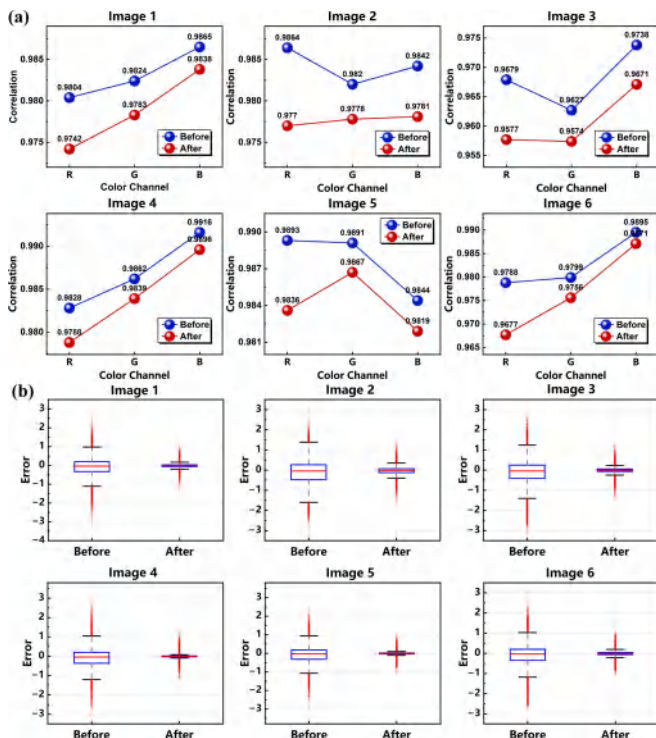


Figure 9 Comparison of PLRAM ADC outputs before and after error compensation. (a) Correlation with theoretical results across RGB channels. (b) Residual error distributions for six image samples.

We further conducted a full-system evaluation of CSDet inference on hardware. Figure 10 illustrates six different image samples, including the original input, reconstructed image, and compressed-domain detection results from the CSDM. Despite a low sampling rate of 0.1, CSDet successfully identified key object regions, demonstrating robust detection performance in resource-constrained settings. Quantitatively, the system achieved an average PSNR of 27.87 dB and SSIM of 0.8916 across 100 test images from the



Figure 10 Compressed-domain inference and image reconstruction results of CSDet on the PLRAM chip, showing six samples with original images, reconstructions (PSNR/SSIM), and detection outputs at a 0.1 sampling rate.

MSCOCO dataset, indicating effective preservation of visual information under compression.

Finally, we benchmarked detection accuracy across three deployment configurations: (1) image reconstruction followed by detection, (2) CSDet software inference on a PC, and (3) CSDet hardware inference on the PLRAM chip. As shown in Figure 11, the CIM-based deployment achieved AP50 of 70.77% and AP50-95 of 47.31%, with only a 2.83% average drop compared to the software baseline. Meanwhile, it delivered around 60% energy savings, demonstrating the efficiency and deployment potential of the proposed system in edge scenarios.

V. Conclusion

This paper presents a cross-layer deployment framework for CSDet, a compressed sensing-based object detection system,

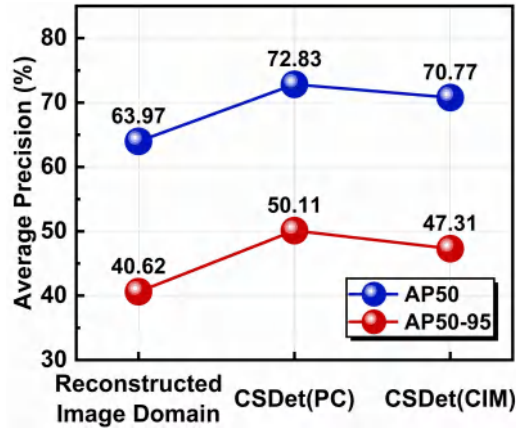


Figure 11 Detection accuracy under three deployment modes. Comparison of AP50 and AP50-95 for image-domain, PC, and CIM-based CSDet.

on a PLRAM-based CIM platform. To address the analog non-idealities challenges in such systems, we develop both device- and circuit-level error models to characterize deviations introduced during inference. A slope-aware correction method and statistical noise modeling are further proposed to compensate for systematic and residual errors.

Experimental results demonstrate that the proposed strategies significantly improve the accuracy of analog inference. The hardware-deployed CSDet achieves high detection performance with only a 2.83% drop in accuracy compared to software baselines, while reducing energy consumption by approximately 60%. These results confirm the effectiveness and practicality of our deployment framework for enabling robust and efficient inference on non-ideal CIM hardware.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0109001) and the ECNU/HDSC Integrated Circuit Engineering Technology Joint Laboratory.

References

- [1] C. Wolters, X. Yang, U. Schlichtmann, and T. Suzumura, "Memory is all you need: An overview of compute-in-memory architectures for accelerating large language model inference", *arXiv preprint arXiv:2406.08413*, in press, 2024.
- [2] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing", *Nature nanotechnology*, vol.15, no.7, pp.529–544, 2020.
- [3] J. Li, H. Abbas, D. S. Ang, A. Ali, and X. Ju, "Emerging memristive artificial neuron and synapse devices for the neuromorphic electronics era", *Nanoscale horizons*, in press, 2023.
- [4] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices", *Nature electronics*, vol.1, no.6, pp.333–343, 2018.
- [5] H.-T. Lue, C.-H. Hung, K.-C. Wang, and C.-Y. Lu, "Prospects of computing in or near flash memories", in *2024 IEEE International Electron Devices Meeting (IEDM)*, IEEE, pp.1–4, 2024.
- [6] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, *et al.*, "Neuromorphic

computing using non-volatile memory", *Advances in Physics: X*, vol.2, no.1, pp.89–124, 2017.

- [7] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, *et al.*, "A compute-in-memory chip based on resistive random-access memory", *Nature*, vol.608, no.7923, pp.504–512, 2022.
- [8] P. Yao, H. Wu, B. Gao, J. Tang, *et al.*, "Fully hardware-implemented memristor convolutional neural network", *Nature*, vol.577, no.7792, pp.641–646, 2020.
- [9] H.-T. Lue, P.-K. Hsu, M.-L. Wei, T.-H. Yeh, *et al.*, "Optimal design methods to transform 3d nand flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvcim) accelerator for deep-learning neural networks (dnn)", in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, pp.38–1, 2019.
- [10] D. L. Donoho, "Compressed sensing", *IEEE Transactions on information theory*, vol.52, no.4, pp.1289–1306, 2006.
- [11] Y. Jie, C. Guo, M. Li, and B. Feng, "Construction of compressed sensing matrices for signal processing", *Multimedia Tools and Applications*, vol.77, pp.30551–30574, 2018.
- [12] W. Shi, F. Jiang, S. Liu, and D. Zhao, "Image compressed sensing using convolutional neural network", *IEEE Transactions on Image Processing*, vol.29, pp.375–388, 2019.
- [13] W. Cui, S. Liu, F. Jiang, and D. Zhao, "Image compressed sensing using non-local neural network", *IEEE Transactions on multimedia*, vol.25, pp.816–830, 2021.
- [14] X. Huang, Y. Gong, W. Wu, S. Zhu, and Y. Zhao, "Csdet: A compressed sensing object detection architecture with lightweight networks", *IEEE Transactions on Circuits and Systems for Video Technology*, in press, 2024.
- [15] S. Gao, G. Yang, X. Qiu, C. Yang, *et al.*, "Programmable linear ram: A new flash memory-based memristor for artificial synapses and its application to speech recognition system", in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, pp.14–1, 2019.
- [16] S. Gao, Y. Cong, Z. Zhang, X. Qiu, *et al.*, "Superior data retention of programmable linear ram (plram) for compute-in-memory application", in *2020 IEEE International Reliability Physics Symposium (IRPS)*, IEEE, pp.1–5, 2020.
- [17] S. Gao, F. Yang, C. Lu, and Y. Zhao, "Fast algorithms for exact ir drop de-embedding in analog multiply-accumulate computing", *IEEE Transactions on Electron Devices*, vol.69, no.11, pp.6376–6383, 2022.
- [18] S. Zhu, C. Lu, X. Qiu, S. Gao, *et al.*, "Mitigating re-delay induced accuracy loss in analog in-memory computing: A non-compromising approach", *IEEE Transactions on Consumer Electronics*, in press, 2024.
- [19] J. Jin, S. Gao, C. Lu, X. Qiu, and Y. Zhao, "Device nonideality-aware compute-in-memory array architecting: Direct voltage sensing, iv symmetric bitcell and padding array", *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, in press, 2025.



Saike Zhu received the B.S. and M.S. degrees from the School of Communication Engineering, Southwest University, Chongqing, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in electronics and information engineering with Zhejiang University, Hangzhou, China. His research interests include flash-based computing-in-memory and hardware/software co-optimization in CIM architecture.



Xiaoqian Huang received the M.E. degree from Hunan University, Changsha, China, in 2021. She is currently an Engineer with China Nanhu Academy of Electronics and Information Technology. Her research interests include computer vision, compressive sensing, and AI compiler.



Cimang Lu received his bachelor's degree from Peking University and his Ph.D. with honor from the University of Tokyo. He previously conducted research at Qualcomm on emerging memory devices. He received the Best Paper Award at a premier VLSI conference.



Yong Gong received the B.S. and M.S. degrees from Air Force Engineering University, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in computer science from the PLA University of Science and Technology, Nanjing, China. He is currently a Senior Engineer with China Nanhu Academy of Electronics and Information Technology. His research interests include deep learning, AI compiler, and compute-in-memory architecture.



Yi Zhao (Senior Member, IEEE) received the B.S. degree from Nanjing University of Astronautics and Aeronautics, Nanjing, China, in 2000, the M.S. degree from Zhejiang University, Hangzhou, China, in 2003, and the Ph.D. degree from The University of Tokyo, Tokyo, Japan, in 2007. He is currently a Professor with Zhejiang University and East China Normal University. His recent research interests include advanced logic and memory technology, including process, device physical, and reliability.



Shifan Gao received his bachelor's degree in 2018 from the University of Electronic Science and Technology of China. He obtained his Ph.D. degree from Zhejiang University. His research interests focus on non-volatile memory and its computing-in-memory applications, particularly Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM) and Flash memory.



Xianggao Wang is currently pursuing the Ph.D. degree in electronics and information engineering with Zhejiang University, Hangzhou, China. His research interests include analog computing-in-memory and nonvolatile memories.



Hua Chen received the B.S. degree from Harbin Engineering University, Harbin, China, in 2004. He is currently pursuing the M.S. degree in Integrated Circuit Engineering at East China Normal University, Shanghai, China. He is also currently an Engineer with China Nanhu Academy of Electronics and Information Technology. His research interests include AI deployment, embedded systems.

Research Article

A Point-to-Point Feature Processing Approach for Current Detection in CIM SoC

Yu Zetao¹, Qiu Xiang¹, Lu Cimang²

¹ School of Integrated Circuits, East China Normal University, Shanghai 200241, China

² Flash Billion Semiconductor Co. Ltd., Shanghai 201210, China

Corresponding author: Qiu Xiang; Email: xqiu@ic.ecnu.edu.cn.

Manuscript Received June 20, 2025; Accepted March 22, 2022; Published Online March 22, 2022.

This work was supported by the National Science and Technology Major Project(2020AAA0109001) and the ECNU/HDSC Integrated Circuit Engineering Technology Joint Laboratory.

Abstract — Computing-in-Memory (CIM) System-on-Chip (SoC) solutions based on Programmable Linear Random-Access Memory (PLRAM) demonstrate strong performance in deploying neural networks for applications like speech and image recognition. However, significant accuracy degradation occurs when deploying specific load detection models. To enable efficient on-chip deployment of such models while improving accuracy, this paper proposes a spiking neural network-inspired point-to-point (P2P) feature processing scheme. This approach achieves 99% accuracy for neural network deployment on the CIM chip, representing a 4.5% improvement over conventional implementations.

Keywords — Computing-in-Memory (CIM), point-to-point (P2P) processing, Current detection, PLRAM.

I. Introduction

CIM technology based on emerging memory devices, particularly SoC solutions utilizing PLRAM architectures, has emerged as a highly promising approach for artificial intelligence edge computing. These chips significantly reduce the energy consumption and latency bottlenecks caused by frequent data movement between processors and memory by embedding computing units within or tightly coupling them to memory arrays. Consequently, they demonstrate distinct advantages in efficiently deploying mainstream neural network models (e.g., speech recognition, image recognition), achieving high-throughput and low-power processing capabilities[1-4].

However, as application scenarios continue to expand—especially in critical domains such as Industrial IoT, smart grids, and battery management systems—the demand for load signal detection and analysis models has grown substantially. These models, typically used for equipment condition monitoring, fault diagnosis, and energy efficiency optimization, impose extremely high requirements for real-time performance and accuracy when deployed on edge embedded devices. Yet, existing PLRAM-based CIM chips exhibit a noticeable decline in accuracy compared to traditional architectures when deploying such specific current detection models. This performance gap severely limits CIM technology's application potential in broader, more critical scenarios. The root cause likely stems from the unique dynamic characteristics of current signals (e.g., transients, noise patterns), where existing CIM architectures struggle to capture transient features during processing, leading to degraded model generalization and compromised

inference accuracy after hardware mapping.

To achieve efficient and high-accuracy deployment of specific neural network models like load current detection on CIM chips, inspired by Spiking Neural Network (SNN) [5], this paper innovatively proposes a P2P feature processing scheme. This solution is not a strict biological SNN but adapts its core principles—event-driven processing, sparse computation, and spatiotemporal information handling—optimized for CIM hardware characteristics. Through this SNN-inspired mechanism, we aim to more effectively capture dynamic features and temporal dependencies in current signals while fully leveraging CIM's inherent advantages of high parallelism and low power consumption to achieve accurate load detection.

II. Related Works

1. CIM SoC Architecture

The computing-in-memory chip utilized in this study integrates a crossbar switch circuit shown in Figure 1. Its memory cells employ PLRAM, delivering exceptional programming/erasure capabilities [8-10]. Input signals undergo analog-to-digital conversion into voltage vectors mapped onto rows of the memristor-based crossbar array. Since memristor terminals connect to both array rows and columns, current flow through each memristor can be modulated by adjusting row or column voltages. Within the crossbar array, the current through individual memristors is determined according to Ohm's Law. By tuning memristor conductance to emulate neural network synaptic weights, each device's conductance becomes proportional to its current during signal application.

Column current measurements yield computational results of input vector-weight matrix operations, while Kirchoff's Current Law — as shown in Equation (1) establishes that each output node's current equals the summation of these column currents, representing the neural network's final output[11].

$$I_{out} = \sum_{i=1}^n \frac{V_i}{R_i} = \sum_{i=1}^n V_i G_i \quad (1)$$

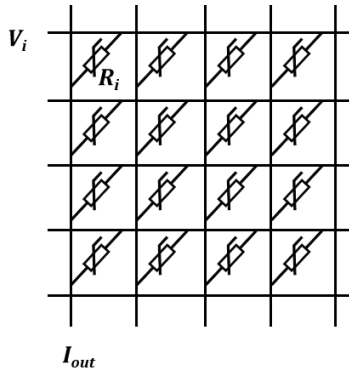


Figure 1 4×4 memristor crossbar array architecture with PLRAM cells.

2. Load Feature Extraction

In power system load feature extraction, three feature categories serve distinct purposes — steady-state characteristics (e.g., power variations, V-I trajectories), transient signatures (e.g., inrush currents, voltage spikes), and non-conventional approaches (e.g., EMI patterns, recurrence plots)[7]—yet transient and non-conventional features incur prohibitive computational costs and exhibit poor computing-in-memory compatibility due to dimensionality complexity and noise sensitivity. Steady-state features conversely provide exceptional stability and engineering practicality critical for smart grid monitoring, deriving from their minimal parameter fluctuations ensuring training data reliability, mature power-flow-based quantization enabling efficient processing, and inherent sensitivity to gradual anomalies facilitating early fault detection.

While Biansoongnern et al. [8] utilized low sampling rates to detect power variations using artificial neural networks (ANNs) trained on steady-state active/reactive power features, and Chang et al. [9] extracted power characteristics via genetic algorithms, both approaches demonstrate limitations in distinguishing low-wattage appliances due to insufficient feature discrimination. This work employs 6400 Hz high-frequency sampling to capture complete appliance operational cycles, generating temporally optimized steady-state waveforms that directly enable point-to-point feature processing where regularized inputs enhance dynamic characteristic extraction.

III. Point-to-Point Methods

1. Feature Processing Algorithm

In power system load detection, raw current and voltage waveforms typically contain substantial low-frequency redundancy and environmental noise interference. Their high-dimensional nature further complicates effective extraction of temporal dynamic features, ultimately constraining real-time performance and detection accuracy. To address these limitations, this paper introduces a P2P feature processing method inspired by SNN. The P2P approach captures instantaneous signal variations by computing temporal differences between adjacent data points as expressed in Equation (2):

$$X_{p2p} = X_i - X_{i-1} \quad (2)$$

The workflow begins by truncating raw waveforms into 128-timestep segments, starting precisely at voltage zero-crossing points. These truncated voltage and current data segments are then concatenated into a single input sequence: the first 128 timesteps contain voltage values, immediately followed by the next 128 timesteps containing current values, as depicted in Figure 2. This concatenated structure enables the simultaneous utilization of both voltage and current characteristics within a unified input. Subsequently, the concatenated signals are converted into rate-of-change feature matrices through differencing operations. Local state transfer is then applied to these derived features to detect abrupt data mutations.

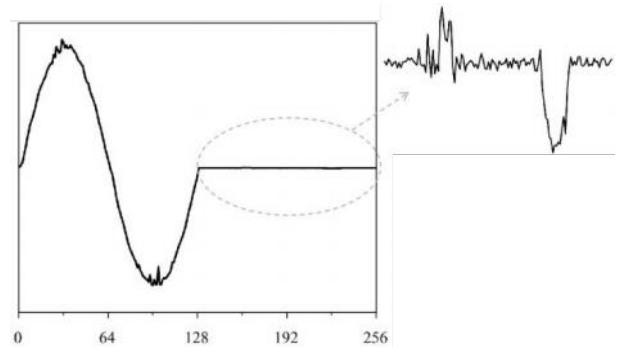


Figure 2 Zero-crossing triggered voltage-current concatenation for P2P input vector formation

2. Network Architecture

Neural network models are extensively applied in data processing, classification, and prediction tasks due to their efficient learning capability for data features[12-14]. Employing Python 3.8.5, an NVIDIA Tesla P40 GPU, and the PyTorch 2.1.2 framework[15], this work developed a neural network model whose architecture is illustrated in Figure 3. The model integrates a one-dimensional convolutional layer for temporal feature extraction with three sequentially connected fully-connected layers (FC1-FC3) for progressive dimensionality reduction. A rectified linear unit (ReLU) activation function follows FC1 to enhance generalization capability, complemented by Dropout regularization to mitigate overfitting. Optimization was performed using the Adam optimizer

with an initial learning rate of 1×10^{-4} and negative log-likelihood loss (NLLoss) objective function. Final classification is achieved through Softmax normalization of FC3 outputs[16], where current categories are determined by maximum probability values.

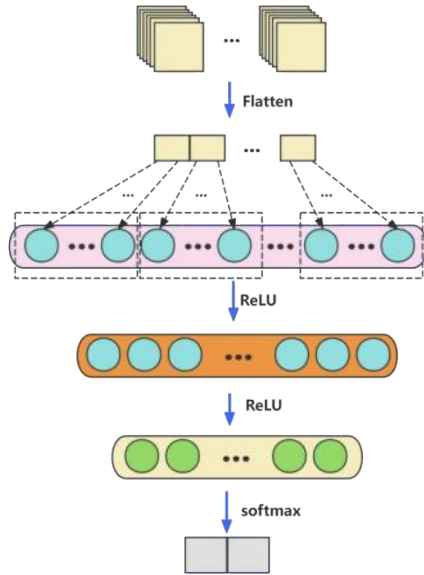


Figure 3 P2P-enhanced neural network with 1D convolutional and three fully-connected layers

3. Result

The model was trained using PyTorch across a multi-GPU environment, employing both raw and P2P-processed datasets comprising 114,646 non-EB (electric bicycle) load samples and 151,758 EB-charging samples. The 266,404 total samples partitioned into 80:20 training-validation splits. Training spanned 60 epochs with early stopping protocols to prevent overfitting. As shown in Figure 4, the P2P-enhanced model achieved 99.90% peak accuracy, surpassing the raw data model's 98.93% accuracy while exhibiting accelerated convergence that reached optimal performance in 5 epochs. This convergence rate was four times faster than the baseline model's 20 epochs. These results demonstrate the P2P algorithm's dual capability to enhance classification accuracy and substantially improve training efficiency.

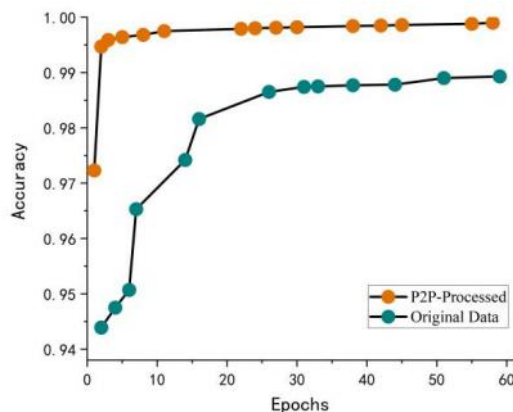


Figure 4 Model training accuracy with different datasets .

To validate the deployment effectiveness of neural network models on the computing-in-memory (CIM) chip, we randomly selected 200 test samples - 100 for none-EB load and 100 for EB-charging. The trained weights from the 1D convolutional layer and three fully-connected layers, along with their bias parameters, were deployed onto the CIM chip's crossbar array. Row and column selectors activated sub-arrays containing weights to perform multiply-accumulate (MAC) operations. During multi-layer computation, each layer's outputs were temporarily stored in on-chip SRAM buffers for subsequent layer processing until final control signals were output externally.

We evaluated the hardware performance by comparing correlation between each layer's outputs from the CIM chip and theoretical software training results, and matching final classification results against ground truth. Figure 5 quantifies the layer-wise output correlation between CIM hardware computations and reference software outputs for comparable layers (excluding FC3 due to its 2-dimensional output space preventing valid correlation analysis). The P2P-optimized model demonstrates significantly enhanced correlation coefficients across all evaluated layers - ConvId (0.967 vs 0.94), FC1 (0.947 vs 0.93), and FC2 (0.995 vs 0.96) - compared to the original model.

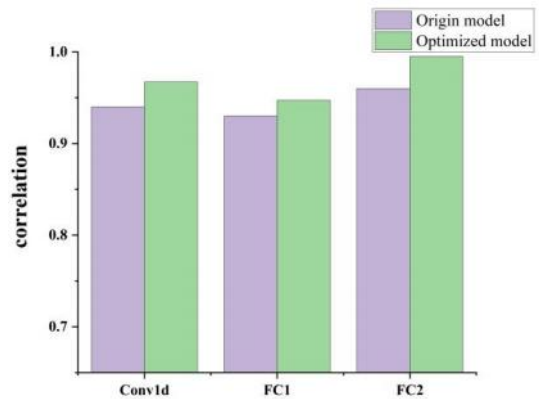


Figure 5 Layer-wise Output Correlation to Theoretical Ground Truth for Different Models and accuracy of different model deployed on CIM SoC.

Figure 6 presents the final deployment classification accuracy against ground truth labels. The inherent precision loss from floating-point to fixed-point conversion during on-chip computation results in lower hardware accuracy compared to software results for both models. Despite this challenge, the P2P-optimized model achieves a hardware accuracy of 99%, significantly outperforming the baseline model's 94.5%. This substantial 4.5% accuracy gain directly corresponds to the suppressed layer-wise output variation quantified in Figure 5, conclusively demonstrating P2P's capability to suppress hardware-induced noise and enhance the robustness of neural network inference on the CIM platform.

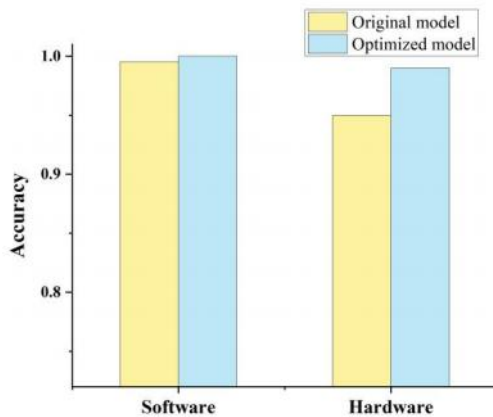


Figure 6 Accuracy Comparison Across Deployment Platforms.

IV. Conclusions

To address accuracy degradation during load current detection model deployment on CIM SoC chips, this study designs a neural network model integrated with P2P feature processing for steady-state load identification. The model captures subtle current/voltage waveform features during stable appliance operation by computing instantaneous differentials between adjacent data points, while incorporating a spiking neural network-inspired event-driven mechanism that enhances feature robustness through periodic dynamic adjustments. Combined with iterative parameter optimization, this approach significantly improves appliance state recognition accuracy. Experimental results demonstrate 99% post-deployment accuracy on the CIM chip, representing a 4.5% improvement over traditional methods, alongside reduced single-sample inference latency of 4 ms. These findings validate the combined benefits of P2P processing and CIM architecture for load current detection.

References

- [1] Gao, S.; Yang, G.; Qiu, X.; Yang, C.; Zhang, C.; Li, B.; Gao, C.; Jiang, H.; Wang, Z.; Hu, J.; et al. Programmable Linear RAM: A New Flash Memory-based Memristor for Artificial Synapses and Its Application to Speech Recognition System. In Proceedings of the 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 7–11 December 2019; pp. 14.1.1–14.1.4.
- [2] Gao, S.; Cong, Y.; Zhang, Z.; Qiu, X.; Lee, C.; Zhao, Y. Superior Data Retention of Programmable Linear RAM (PLRAM) for Compute-in-Memory Application. In Proceedings of the 2020 IEEE International Reliability Physics Symposium (IRPS), Dallas, TX, USA, 28 April–30 May 2020; pp. 1–5.
- [3] Zhao, L.; Gao, S.; Zhang, S.; Qiu, X.; Yang, F.; Li, J.; Chen, Z.; Zhao, Y. Neural Network Acceleration and Voice Recognition with a Flash-based In-Memory Computing SoC. In Proceedings of the 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), Washington, DC, DC, USA, 6–9 June 2021; pp. 1–5.
- [4] Wang KANG, Jing KOU, Weisheng ZHAO, In-memory computing technology: development status, trends and challenges, SCIENTIA SINICA Informationis, Volume 54, Issue 1, 2024, Pages 16-, ISSN 1674-7267.
- [5] Maass, W. Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, vol. 10, pp: 1659–1671, 1997.
- [6] Jin, J., Qiu, X., & Lu, C. Edge artificial intelligence for electrical anomaly detection based on process-in-memory chip. *Electronics*, 2024, 13(21), 4255.
- [7] P. A. Schirmer and I. Mporas, "Non-Intrusive Load Monitoring: A Review," in *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 769-784, Jan. 2023.
- [8] S. Biansoongnern and B. Plangklang, "Nonintrusive load monitoring (NILM) using an Artificial Neural Network in embedded system with low sampling rate," 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand, 2016, pp. 1-4.
- [9] Hsueh-Hsien Chang, Po-Ching Chien, Lung-Shu Lin, and Nanming Chen, "Feature Extraction of Non-Intrusive Load-Monitoring System Using Genetic Algorithm in Smart Meters " , Eighth IEEE International Conference on e-Business Engineering, pp. 299–304, 2011
- [10] Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete Cosine Transform. *IEEE Trans. Comput.* 1974, C-23, 90–93.
- [11] Prezioso, M.; Merrih-Bayat, F.; Hoskins, B.D.; Adam, G.C.; Likharev, K.K.; Strukov, D.B. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* 2015, 521, 61–64.
- [12] Murat, Y.S.; Ceylan, H. Use of artificial neural networks for transport energy demand modeling. *Energy Policy* 2006, 34, 3165 – 3172.
- [13] Omitaomu, O.A.; Niu, H. Artificial intelligence techniques in smart grid: A survey. *Smart Cities* 2021, 4, 548–568.
- [14] Hallmann, M.; Pietracho, R.; Komarnicki, P. Comparison of Artificial Intelligence and Machine Learning Methods Used in Electric Power System Operation. *Energies* 2024, 17, 2790.
- [15] Imambi, S.; Prakash, K.B.; Kanagachidambaresan, G.R. *PyTorch*; Springer: Cham, Switzerland, 2021; pp. 87–104.
- [16] Yuan, B. Efficient hardware architecture of softmax layer in deep neural network. In Proceedings of the 2016 29th IEEE International System-on-Chip Conference (SOCC), Seattle, WA, USA, 6 – 9 September 2016.



(Email: 51265900008@stu.ecnu.edu.cn)

Yu Zetao he received the B.S. degree in electronic engineering from the School of Communication and Artificial Intelligence, Nanjing Institute of Technology, Nanjing, China, in 2022. He is now pursuing the M.S. degree in the School of Integrated Circuits, East China Normal University, Shanghai, China. His interests include Computing-in-Memory Architecture.



(Email: xqiu@ic.ecnu.edu.cn)

Qiu Xiang (corresponding author) he received the B.S. degree in Electronic Science and Technology and the M.S. degree in Microelectronics from Tsinghua University, Beijing, China, in 2006 and 2008, respectively, and the Ph.D. in Computer Engineering from the University of California, Santa Barbara, Santa Barbara, CA, USA, in 2013. He is now a Full time researcher at the School of Integrated Circuits, East China Normal University, Shanghai, China. His interests include Electronic Design Automation and Computing-in-Memory Architecture.

基于 STT-MRAM 模拟存内计算的阵列规模分析框架

王向杲¹, 赵毅^{1,2,*}

¹浙江大学信息与电子工程学院, 杭州 310027

²华大半导体有限公司, 上海 201203

摘要: 模拟域 STT-MRAM 存内计算在大规模神经网络加速计算任务中具有突出的优势, 然而现有研究中尚缺少存算阵列规模设计方法。本文从多个角度建立了 STT-MRAM 模拟存算阵列规模分析框架, 并通过 SPICE 仿真比较了不同存算单元阵列规模对计算准确率的影响。

关键词: 自旋转移矩磁性随机存储器 (STT-MRAM); 存内计算; SPICE 仿真

1 前言

传统计算设备普遍采用的冯诺依曼架构存在数据传输功耗高、速度慢等不足, 对边缘侧设备部署高性能 AI 应用造成严重制约^[1]。近年来“存内计算”范式的提出打破了存储和计算的边界, 为缓解“存储墙”限制提供了新的思路。以 STT-MRAM 为代表的新型非易失存储器具有计算逻辑简单、数据无需重复写入等优势, 因此在模拟存内计算领域具有广阔的应用潜力, 近年来也受到诸多研究人员的关注^[2-4]。然而在现有的报道中, 存算单元阵列的规模通常是结合外围电路或任务需求确定, 其设计缺少理论依据或分析框架的支撑。如何在满足计算精度的前提下确定存算阵列的合适规模, 目前的研究仍缺少评估方法。

2 分析框架与仿真结果

通常而言, 为了提高计算密度和并行度, 存算核设计时会尽可能增大单个阵列的规模。然而由于模拟存储器件有限开关比以及电路非理想因素的限制, 过大的阵列规模会带来感知裕度偏小以及信噪比降低的问题, 因此实际设计中需要在阵列规模和计算精度之间做出权衡。因此本文从感知裕度、输入输出线性度、信号衰减和电流串扰等多个角度, 建立一种通用型模拟存算阵列规模的分析框架。

为了方便说明与对比, 以 256×256 、 512×256 和 256×512 三种规模的存算单元阵列为例展开讨论。每个磁隧道结 (MTJ) 器件仅考虑高阻态 (反平行态, AP) 和低阻态 (平行态, P) 两种存储状态, 器件电学特性来自于商用 28 nm CMOS 工艺库, 下面基于 SPICE 仿真的结果进行评估并讨论。

2.1 感知裕度

对于存内计算阵列而言，感知裕度是指在读取阵列输出时，能够可靠区分不同结果所对应模拟量之间的最小差值范围。在绝大多数神经网络应用中，8-bit 输出量化可以实现极低的准确率损失。因此为了验证阵列感知裕度能否满足要求，可以通过改变输入或权重使得差分输出理论值改变满幅范围（Full-scale Range, FSR）的 $1/256$ 并检查输出改变量的实际值能否超过 ADC 的最小识别精度。这里通过改变最远端行的输入或单个权重改变理论结果，图 1 展示了不同列改变前后实际差分电压的变化量。如果记 9-bit ADC 最小识别范围为 Δ ，可以看到对于 256×256 规模的阵列输出改变量可以达到 1.1Δ 。而另两种阵列最差情况的改变量分别仅为 0.5Δ 和 0.6Δ ，意味着计算结果无法被 ADC 正确转换。

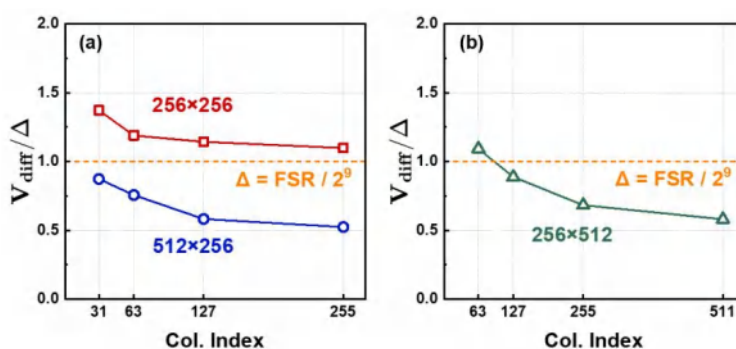


图1 理论输出改变前后实际计算结果的差值

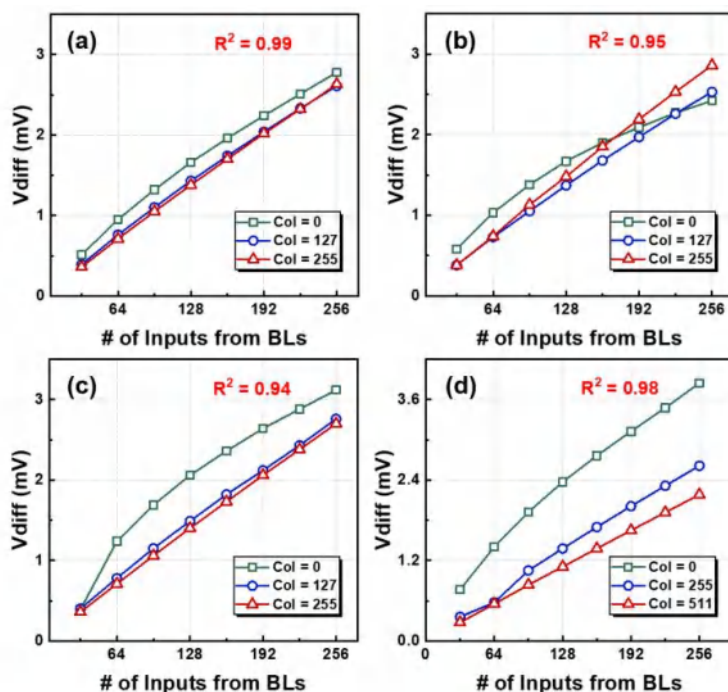


图2 三种规模阵列的输出差分电压随输入变化曲线。其中 (a) 为 256×256 规模，(b-c) 分别为 512×256 规模从近端和远端输入，(d) 为 256×512 规模

2.2 输入输出线性度

矩阵向量乘法运算是一种典型的线性运算，只有输出信号与输入信号保持高度的线性关系，才能确保经转换与后处理得到的最终结果是准确的。为了验证几种规模阵列的输入输出线性度，简单起见这里选择存储权重相同且输入电流幅值固定，通过改变使能位线的个数来改变总输入大小。

上图 2 分别展示了三种规模阵列的输入输出曲线，为保持输入并行度一致这里将 512×256 规模的阵列按使能行分为近端和远端两组。可以看到 256×256 规模的阵列输入输出线性度最高，而另外两种阵列的线性度较差，且近端列与远端列之间的计算均一性显著降低。

2.3 信号衰减

IR-drop 效应是指电流信号经过金属走线时，由于电压衰减导致工作单元两端的实际电压小于理论值的现象。为了探究三种结构阵列中的 IR-drop 效应，这里仍通过电流幅值固定、改变使能位线个数的方式控制输入总电流相等，其中 512×256 规模的阵列采用了奇偶交替方式使能了一半位线，另两种结构则全部位线使能。阵列中 MTJ 器件两端的压差随行列序号的分布情况图 3 所示，可以看出阵列的行数增加至 512 后 IR-drop 效应最为严重，器件两端电压衰减最大幅度为 55.7%，而在 256×512 的阵列中最大衰减为 41.1%。

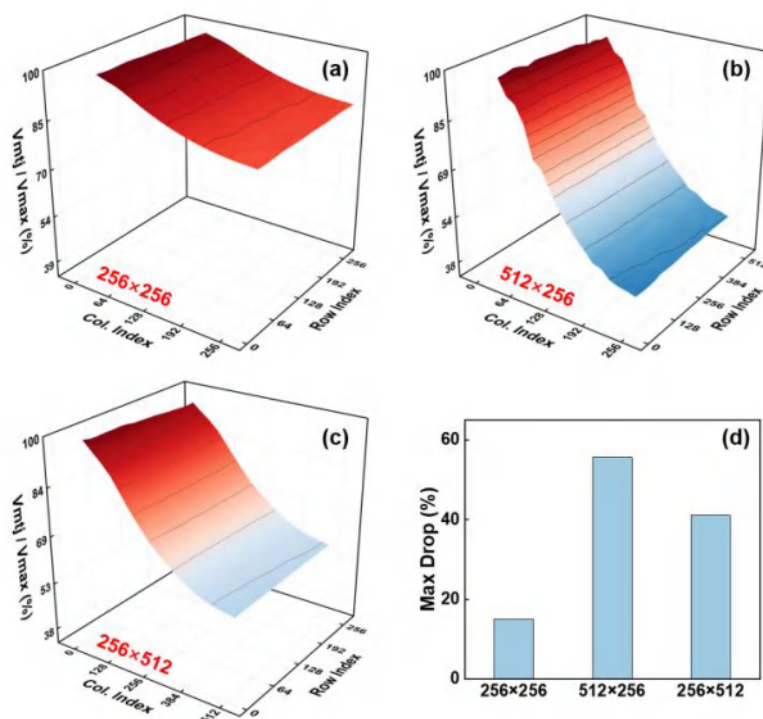


图3 三种规模阵列中 (a-c) 不同坐标MTJ两端电压衰减和 (d) 最大衰减对比

2.4 电流串扰

在大规模阵列中如果提高计算并行度，电流在被使能的行列中发生串扰是无法避免的，因此必须考虑其对计算结果带来的影响。阵列中的电流串扰可以分为行串扰和列串扰。本节提出的表征阵列行串扰方法如图 4 (a) 所示。图 4 (b-d) 展示了行串扰给输出带来的变化。可以看到在规模 256×256 的阵列中行串扰造成的最大输出改变为 1.4%，而在另两种更大规模的阵列中最大误差则超过了 6%。

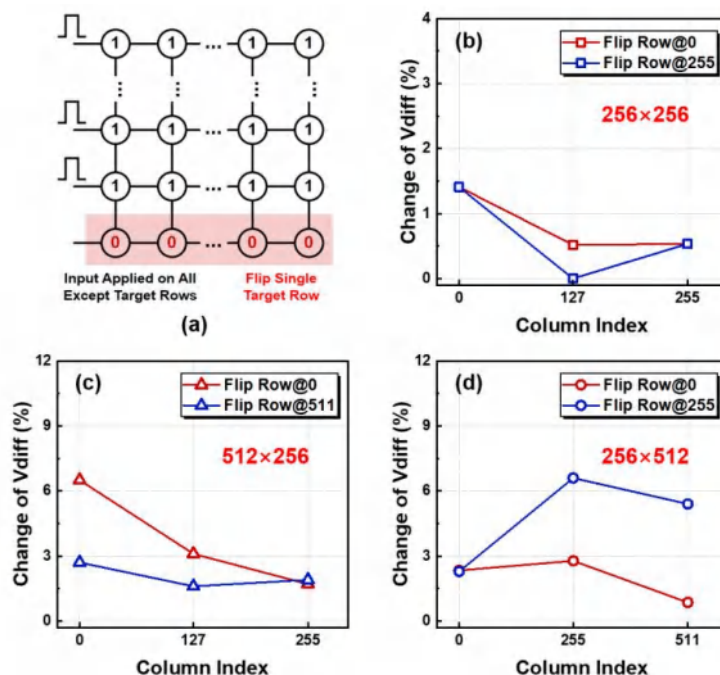


图4 (a) 行串扰测试激励示意图 (b-d) 三种规模阵列的行串扰影响大小

类似的，下图 5 (a) 展示了列串扰的表征方法。图 5 (b) 显示了在相邻列权重改变前后的输出结果变化，可以看出三种规模的阵列受到列串扰的影响最大为 0.65%，规模 256×256 的阵列变化量甚至小于 0.1%。该结果表明电流列串扰对计算准确率的影响远小于行串扰。另外考虑到计算中神经网络输入数据的稀疏性，实际的电流串扰对计算结果读取的影响应当更小。

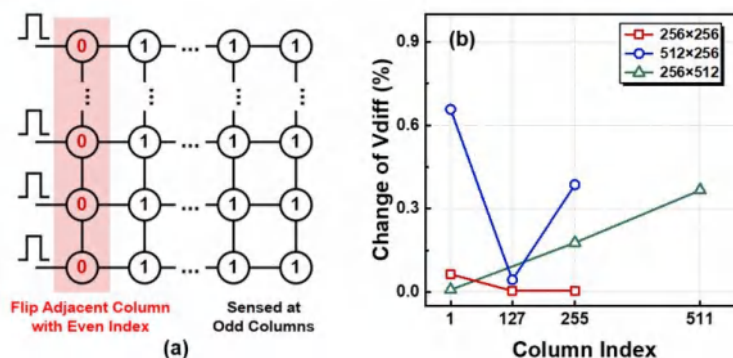


图5 (a) 列串扰测试激励示意图 (b) 三种规模阵列的列串扰影响大小

3 总结

本文从感知裕度、输入输出线性度、信号衰减和电流串扰等角度建立了 STT-MRAM 模拟存算阵列规模分析框架，并通过 SPICE 仿真确定了基于所用工艺的最优存算单元阵列规模为 256×256 。以上分析模式可以应用到各种模拟存算技术中，因此本节所建立的模拟存算阵列规模分析框架为后续类似的研究提供了一种通用范式。

致谢

本研究受到“新一代人工智能国家科技重大专项（2020AAA0109001）”的资助

参考文献

- [1] O. Mutlu, "Processing data where it makes sense in modern computing systems: Enabling in-memory computation," 2018 7th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2018, pp. 8-9, doi: 10.1109/MECO.2018.8405955.
- [2] A. Yusuf, T. Adegbija and D. Gajaria, "Domain-Specific STT-MRAM-Based In-Memory Computing: A Survey," in IEEE Access, vol. 12, pp. 28036-28056, 2024, doi: 10.1109/ACCESS.2024.3365632.
- [3] P. Deaville, B. Zhang and N. Verma, "A Fully Row/Column-Parallel MRAM in-Memory Computing Macro With Memory-Resistance Boosting and Weighted Multi-Column ADC Readout," in IEEE Journal of Solid-State Circuits, vol. 60, no. 5, pp. 1856-1866, May 2025, doi: 10.1109/JSSC.2024.3512360.
- [4] Jung, S., Lee, H., Myung, S. et al. "A crossbar array of magnetoresistive memory devices for in-memory computing." Nature 601, 211–216 (2022), doi: 10.1038/s41586-021-04196-6.

面向多芯片异构集成的 SiP 封装技术

赵国强^{1,2}, 赵毅^{1,2,*}

¹浙江大学信息与电子工程学院, 杭州 310027

²华大半导体有限公司, 上海 201203

摘要: 系统级封装 (SiP) 作为实现多芯片异构集成的核心技术体系, 目前已获得业界广泛关注并在高性能电子设备中实现规模化应用。本文基于多芯片三维集成拓扑, 对 SiP 技术进行体系化分类, 并深入剖析各封装形态的关键工艺技术特征。

关键词: 多芯片; 异构集成; SiP; 系统级封装

基金项目: 新一代人工智能国家科技重大专项 (2020AAA0109001)

1 前言

在后摩尔时代, 通过先进封装技术而非单纯依赖制程微缩实现电子系统的性能提升, 已成为半导体行业的技术共识并在工程实践中广泛应用。相较于片上系统 (System-on-Chip, SoC), 系统级封装 (System-in-Package, SiP) 展现出更强的异构集成灵活性: SoC 将各类功能模块高度集成于单一芯片载体, 要求所有模块遵循统一设计规则并采用同构工艺制造; 而 SiP 技术可对各模块芯片进行异构集成, 允许不同功能单元基于差异化工艺制程独立研发, 在实现性能优化的同时, 显著降低设计成本并压缩研发周期。

2 SiP 技术分类

在 SiP 技术体系中, 多芯片的空间排布可采用平面排列、垂直堆叠或二者兼有形式, 并通过中介层 (Interposer) 实现与基板的电气连接。根据多芯片排列形式及中介层是否存在, SiP 技术可系统划分为 2D、2.5D、3D 及 3.5D 封装类型。以下将针对各封装类型展开典型案例分析与技术特征阐释。

2.1 2D 封装

如图 1 所示, 该封装采用双芯片平面布局方式, 基于倒装芯片 (Flip Chip) 技术通过凸点 (Bump) 实现与基板的电气互连。当凸点间距处于 100-400 μm 区间时, 可选用 C4 凸点 (Controlled Collapse Chip Connection) 工艺, 凸点高度约 50-150 μm ; 若需提升互连密度, 可通过 C2 凸点 (Chip Connection) 将间距缩小至 50-150 μm , 凸点高度相应调整为 100-300 μm 。除倒装焊凸点互连外, 亦可采用引线键合 (Wire Bonding) 技术实现芯片与基板的低密度电气连接。

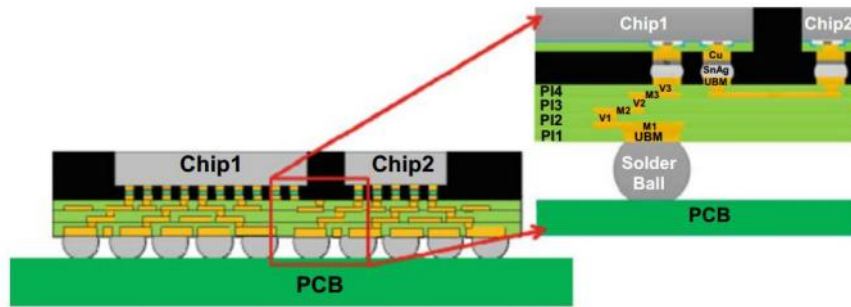


图1 IME展示的扇外型晶圆级封装（Fan-out Wafer Level Packaging, FOWLP)技术方案[1]

2.2 2.5D 封装

台积电 TSMC 的 CoWoS（Chip on Wafer on Substrate）[2]与英特尔 Intel 的 EMIB（Embedded Multi-Die Interconnect Bridge）[3]封装技术均为 2.5D 封装技术的典型代表。其技术特征以中介层的引入为标志，中介层通过硅通孔（Through-Silicon Via, TSV）与重分布层（Redistribution Layer, RDL）技术的协同应用，可有效降低信号传输延迟并大幅提升带宽密度。TSV 技术通过在硅中介层中刻蚀深孔并填充导电材料（如铜），构建起芯片间的垂直电气互连通道。而 RDL 技术则通过溅射、电镀等工艺在中介层表面形成多层金属布线，将芯片原生 I/O 端口重新分布至更密集的区域，实现布线密度的提升。当前主流中介层类型除硅中介层外，还包括玻璃中介层，后者常应用在毫米波通信等高频场景中。

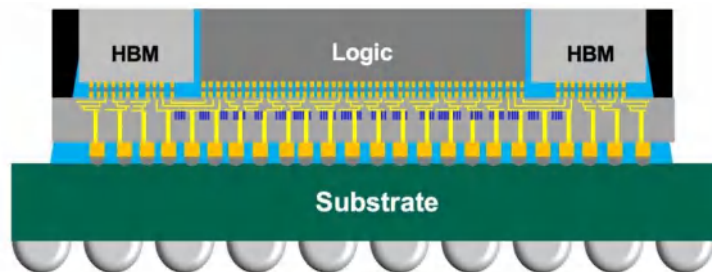


图2 TSMC提出的基于硅中介层的CoWoS封装技术[4]

2.3 3D 封装

对芯片集成密度的极致追求，正驱动功能模块从二维平面布局向三维立体集成演进。图 3 展示了英特尔 Intel 在 2019 年架构日活动中展示的首款基于 Foveros 3D [5]封装技术的“混合 x86 处理器”。该处理器将 10nm 工艺的计算芯片与 22nm FFL 工艺的 I/O 芯片通过微凸点（Microbump）和硅通孔垂直堆叠，在 12×12mm 的封装面积内集成了高性能核心、低功耗核心、核显、内存控制器等模块。随着混合键合（Hybrid Bonding）技术的产业化成熟，后续迭代方案逐步以混合键合替代微凸点互连。这种无凸点键合工艺将互连间距从微凸点的 50-100 μm 缩小至 3-5 μm 。



图3 英特尔展示的首款基于Foveros 3D封装技术的“混合x86处理器”[6]

2.4 3.5D 封装

3.5D 封装技术通过融合 3D 垂直堆叠与 2.5D 中介层集成技术，构建起异构芯片的混合维度集成范式。AMD 推出的高性能计算与人工智能加速器 MI300A，正是该技术体系的典型应用案例。采用铜铜混合键合工艺将 GPU/CPU 模块进行三维垂直堆叠，再借助硅中介层技术，将堆叠后的计算模块与 HBM 高带宽内存、无源器件实现水平互连，形成“3D 堆叠+2.5D 中介层”的复合封装架构。该技术已应用于美国 EI Capitan 超级计算机，展现出 3.5D 封装在 HPC 与 AI 融合场景中的技术优势。

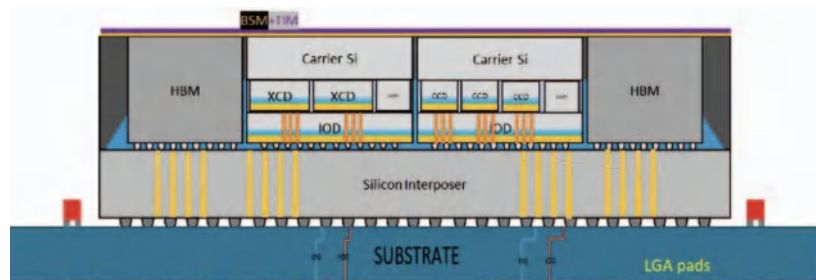


图3 AMD推出的基于3.5D封装的MI300A加速器[7]

3 总结

异构集成的核心在于通过先进封装架构创新弥补摩尔定律演进放缓带来的性能提升瓶颈，实现跨制程、跨功能异构芯片组件的高效协同运行。系统级封装 SiP 通过模块化架构实现功能部件的异构集成，在显著压缩研发成本与缩短产品迭代周期的同时，已成为构建高性能计算单元的最优成本效益解决方案之一。本文系统梳理了 SiP 技术的多元封装形态，各类封装范式各具技术特性与应用边界，实际工程中常呈现多形态封装与键合技术的协同应用特征。

参考文献

- [1] V. S. Rao et al., "Development of High Density Fan Out Wafer Level Package (HD FOWLP) with Multi-layer Fine Pitch RDL for Mobile Applications," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 2016, pp.

1522-1529.

- [2] Y. -C. Hu et al., "CoWoS Architecture Evolution for Next Generation HPC on 2.5D System in Package," 2023 IEEE 73rd Electronic Components and Technology Conference (ECTC), Orlando, FL, USA, 2023, pp. 1022-1026.
- [3] R. Mahajan et al., "Embedded Multi-die Interconnect Bridge (EMIB) -- A High Density, High Bandwidth Packaging Interconnect," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 2016, pp. 557-565.
- [4] <https://3dfabric.tsmc.com/schinese/dedicatedFoundry/technology/cowos.htm>.
- [5] D. B. Ingerly et al., "Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 19.6.1-19.6.4.
- [6] <https://www.ednchina.com/news/201812131421.html>.
- [7] C. S. Mandalapu et al., "3.5D Advanced Packaging Enabling Heterogenous Integration of HPC and AI Accelerators," 2024 IEEE 74th Electronic Components and Technology Conference (ECTC), Denver, CO, USA, 2024, pp. 798-802.

55nm 工艺存算一体芯片的架构创新与能效优化

金健孜¹, 赵毅^{1,2,*}

¹华东师范大学集成电路科学与工程学院,上海 200241

²浙江大学信息与电子工程学院, 杭州 310027

摘要: 随着人工智能和边缘计算的快速发展,对高性能、低功耗的芯片需求日益增长。55nm 工艺因其成熟的制程、较低的功耗和较高的性价比,成为存算一体芯片的理想选择。本文系统梳理了 55nm 工艺存算一体芯片的研究现状,深入分析其技术特点、创新点及性能表现,并对未来发展趋势进行了展望。

关键词: 存内计算; 异构集成; 能效优化; 55nm 工艺

资助: 本文受新一代人工智能国家科技重大专项(2020AAA0109001)和华东师范大学/华大半导体有限公司集成电路工程技术联合实验室资助。

1 引言

近年来,随着人工智能(AI)在边缘计算和物联网(IoT)领域的广泛应用,催生了对低功耗、高能效的芯片的迫切需求。传统的冯·诺依曼架构的存储墙,处理器无法快速获取海量数据,内存带宽和能耗成为瓶颈,严重影响了神经网络在芯片上计算的效率^[1]。存算一体架构通过整合存储与计算单元成为解决这一问题的有效方案^[2]。在众多半导体工艺节点中,55nm 工艺凭借其成熟的制程技术、较低的制造成本以及适中的功耗表现,成为存算一体芯片的理想选择,为边缘计算提供了高性价比的解决方案。相较于更先进的制程(如 7nm 或 5nm),55nm 工艺在性能、功耗和成本之间实现了更好的平衡,尤其适合资源受限的终端设备,如智能传感器、可穿戴设备和嵌入式系统。

本文首先介绍了存算一体芯片技术的提出以及 55nm 工艺的优势,之后介绍了 55nm 存算一体芯片的发展现状,如存内计算(Computing-in-Memory, CIM)、异构集成和混合精度设计等创新型芯片架构,以在该工艺上实现更高的能效比和计算密度。最终分析了发展意义及展望未来发展方向。

2 发展现状

当前,存算一体技术的研究与应用主要聚焦于 AI 专用计算领域,尤其在深度神经网络加速方面展现出显著优势。而基于 55nm 工艺的存算一体芯片研究呈现出多元化的发展态势,数字、模拟及混合信号计算范式各具特色。

2.1 数字计算与模拟计算的创新

在数字计算领域，研究者通过优化计算架构和存储层次来突破性能瓶颈。如图 1 (a) 所示，清华大学团队开发的异构 CIM 系统级芯片 (System on Chip, SoC) 采用了生产者-消费者指令依赖控制器 (PCIDC) 和 8 端口共享 SRAM (Static Random-Access Memory, 静态随机存取存储器) 架构，实现了多核并行处理时的零延迟数据共享，将外部数据传输减少了 45.7%，SoC 级能效提升至 33.0 TOPS/W^[3]。

模拟计算方向则聚焦于提升精度和能效的平衡。图 1 (b) 中展示了 Intel 提出的 8 位有符号 AciM (Analog compute-in-memory) 加速器采用混合差分电容梯度和稀疏感知 DAC 技术，在保持 8 位精度的同时实现了 42.6 TOPS/W 的能效^[4]。其嵌入式 SAR-ADC 通过 LSB 跳过机制进一步将 ReLU 运算的功耗降低 50%，展现了模拟计算在低功耗场景中的潜力。

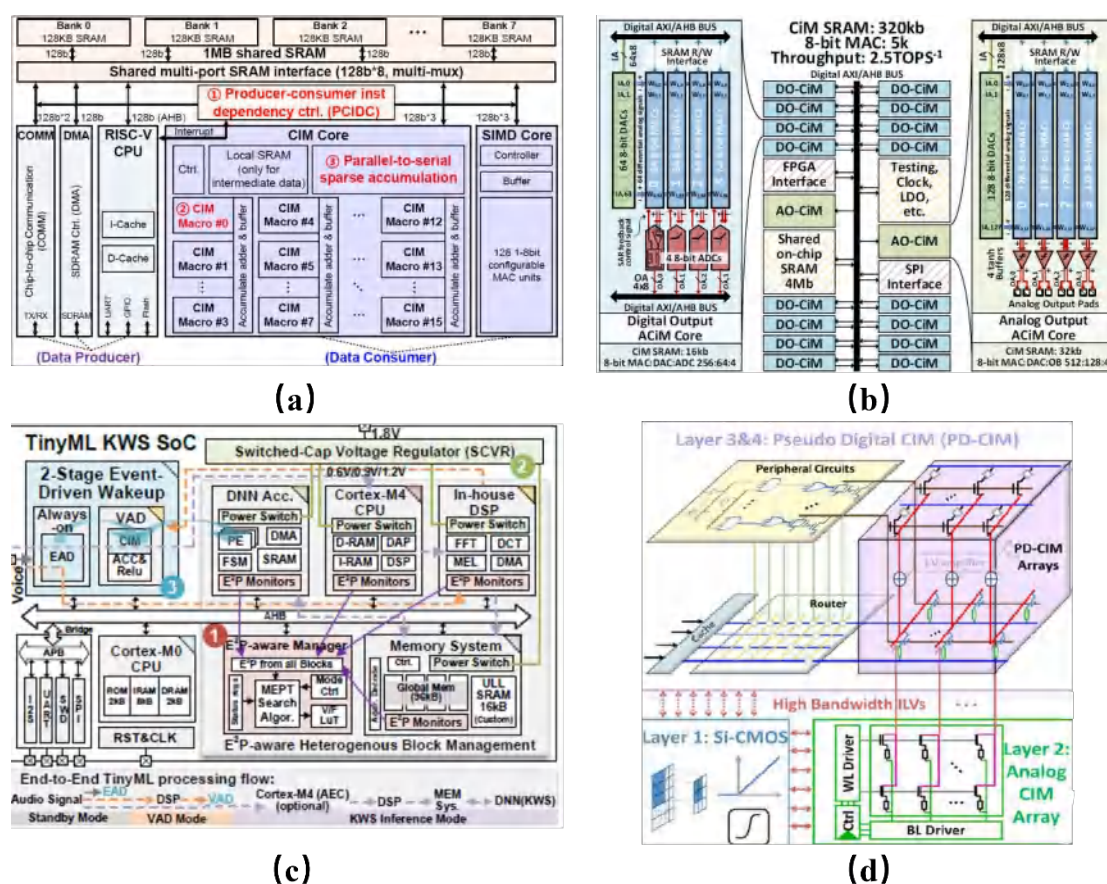


图1 4种技术的结构示意图。(a) 清华大学团队开发的异构CIM系统级芯片结构图^[3]；(b) Intel提出的8位有符号AciM (Analog compute-in-memory) 加速器结构图^[4]；(c) 北京大学团队提出的TinyML SoC结构图^[5]；(d) 清华大学实现的M3D-MP4芯片结构图^[6]。

2.2 系统级优化与 3D 集成

在系统级优化方面，当前研究主要围绕数据搬运瓶颈、能效动态管理和 3D

集成展开突破。如图 1 (c) 所示, 北京大学团队在 TinyML SoC 中采用的两级事件驱动唤醒机制, 通过 CIM-based 语音活动检测将“常开”功耗降低 87%, 待机功耗仅为 $3.5\mu\text{W}$ ^[5]。此外, 全数字合成的 E²P (Energy-event-performance) 监视器仅占用 3% 的芯片面积, 却能实现 >95% 的能耗预测精度, 为动态电源管理提供了高效工具。

3D 集成技术为 55nm 工艺注入了新的活力。清华大学实现的 M3D-MP4 (Monolithic 3D intergration chip featuring Mixed-Precision analog-digital CIM architecture with 4 functional layers) 芯片 (图 1 (d)) 通过单片集成四层功能, 在 55nm 工艺下达到了与 GPU 相当的推理精度 (ResNet-32 在 CIFAR-10 上 90.15% 准确率), 同时能效提升 16.92 倍^[6]。这些进展充分展现了 55nm 工艺在边缘 AI 计算中的巨大潜力。

表 1 性能对比

	[3]	[4]	[5]	[6]
工艺	55nm	55nm	55nm	55nm + 3D集成
能效 (TOPS/W)	33.0 (SoC)	42.6	3.5 μW 待机	16.92 \times 优于GPU
精度	4-8bit	8bit有符号	INT8	混合精度 (GPU 等效)

3 结论

基于 55nm 工艺的存算一体芯片的发展具有重要的技术价值和产业意义。从技术层面看, 这些创新成果突破了传统冯·诺依曼架构的限制, 通过存内计算、异构集成等新范式, 在成熟工艺节点上实现了接近先进制程的性能表现。从产业应用角度, 这些技术进步推动了 AI 向边缘端的快速渗透。而 55nm 工艺的成熟性和经济性则加速了 AI 技术在工业控制、智能家居等领域的落地应用。未来 55nm 的存算一体芯片的发展将呈现两个主要趋势。一方面, 3D 集成技术将继续深化, 通过垂直堆叠不同计算单元和存储器件来进一步提升性能密度。另一方面, 智能化电源管理和自适应计算架构将成为研究重点, 通过更精细的运行监测和动态调节, 在保证计算精度的前提下进一步降低功耗。

综上所述, 55nm 工艺的存算一体芯片通过持续创新展现了强大的生命力。这些技术不仅为边缘 AI 计算提供了高性价比的解决方案, 也为后摩尔时代的芯片发展探索了新路径。随着架构优化、3D 集成和智能管理技术的不断进步, 55nm 工艺有望在边缘计算领域持续发挥重要作用。

参考文献

- [1] ZHANG W Q, GAO B, TANG J S, et al. Neuro-inspired computing chips[J]. Nature Electronics, 2020, 3(7): 371-382. DOI: 10.1038/s41928-020-0435-7.
- [2] 杨兴华, 张静, 杨子翼, 等. 基于 SRAM 的感存算一体化技术综述[J]. 电子与信息学报, 2023, 45(8): 2828-2838.
- [3] Yue J, Zhan M, Wang Z, et al. A 5.6-89.9TOPS/W Heterogeneous Computing-in-Memory SoC with High-Utilization Producer-Consumer Architecture and High-Frequency Read-Free CIM Macro[J]. VLSI Symposium on Technology and Circuits Digest, 2023: C21-3.
- [4] Wang H, Liu R, Dorrance R, et al. A PVT Robust 8-Bit Signed Analog Compute-In-Memory Accelerator with Integrated Activation Functions for AI Applications[J]. VLSI Symposium on Technology and Circuits, 2024: 000639.
- [5] Dong Y, Liu X, Bai K, et al. A Heterogeneous TinyML SoC with Energy-Event-Performance-Aware Management and Compute-in-Memory Two-Stage Event-Driven Wakeup[J]. VLSI Symposium on Technology and Circuits, 2024: 000721.
- [6] Zhang Y, Tang J, Li Y, et al. Monolithic 3D Integration of Multi-layer CNT-CMOS/RRAM Macros for Mixed-Precision Analog-Digital Computing-in-Memory Architecture[J]. IEDM Technical Digest, 2024.

面向存算一体架构的 AI 编译器研究进展

龚勇¹, 赵毅^{1,2}

¹ 华大半导体有限公司, 上海 201203

² 浙江大学信息与电子工程学院, 杭州 310027

摘要: 存算一体架构通过打破“内存墙”, 为人工智能计算提供了显著的能效提升潜力。然而, 其非冯·诺依曼特性、硬件异构性与计算精度挑战, 对传统 AI 编译器提出了革新需求。本文综述了近年来面向存算一体架构的 AI 编译器技术进展, 包括硬件抽象、数据映射策略、精度补偿技术以及典型编译框架, 并指出了标准化抽象、大模型支持、多层次协同优化等未来研究重点和方向。

关键词: 存算一体; AI 编译器; 内存计算; 神经网络加速; 近存计算; 硬件软件协同设计
基金项目: 新一代人工智能国家科技重大专项 (2020AAA0109001)

1 前言

传统冯·诺依曼架构中数据在处理器与存储器间的频繁搬运造成了巨大的能耗开销 (“内存墙”问题), 严重制约了人工智能 (AI), 特别是深度学习模型的能效比。存算一体架构通过在存储器单元内部或近旁执行计算操作, 从根本上减少了数据移动, 为高效 AI 计算提供了革命性路径[1]。ReRAM、PCM、MRAM、FeFET 等新型非易失性存储器 (NVM) 以及基于 SRAM/DRAM 的存内/近存计算方案在近年取得了显著进展, 相关原型芯片与系统 (如 IBM 的 Analog AI Core[2]、清华大学 Thinker[3]等) 不断涌现。

然而, 存算一体硬件具有高度异构性 (数字/模拟、单元结构、数据位宽)、独特的非理想特性 (如模拟计算的漂移、噪声、非线性) 以及颠覆性的执行模型, 使得通用 AI 编译器 (如 TVM[4]、MLIR[5]) 难以直接高效利用其优势。面向 CIM 的专用编译器研究应运而生, 旨在弥合高层的神经网络模型与底层的存算一体硬件特性之间的语义鸿沟, 实现软硬件协同优化。本文分析该领域的核心挑战, 总结近年来关键技术进展, 梳理优化方法, 并展望未来趋势。

2 存算一体架构特性与编译器挑战

存算一体架构主要分为近存计算 (Computing near Memory, CNM)、存内处理 (Processing in Memory) 和存内计算 (Computing in Memory, CIM), 存内计算又可分为数字存内计算 (Digital CIM) 和模拟存内计算 (Analog CIM)。其架构特性与编译器挑战主要体现在:

(1) 计算位置与范式改变: 计算发生在存储阵列内部 (存内) 或紧邻存储

单元（近存），打破了传统“计算单元-存储单元”的分离范式。编译器需重新定义计算抽象和调度策略。

（2）高度异构性：不同存算一体技术在数据表示（模拟电压/电流、数字脉冲）、精度（1-8bit 常见）、支持操作（MAC 为主，部分支持非线性计算）上差异巨大。编译器需要强大的硬件抽象能力。

（3）模拟非理想特性：Analog CIM 受器件/电路非理想因素（非线性、噪声、漂移、工艺偏差）影响显著，导致计算误差。编译器需集成误差建模与补偿技术。

（4）数据映射与移动约束：数据（权重、输入）需预先映射到特定存储单元，片上缓存有限，跨阵列/核数据移动代价高，编译器需精细优化数据布局与通信。

（5）探索空间庞大：将神经网络映射到存算一体硬件涉及权重/激活量化、切片（Tiling）、阵列映射、调度等多个维度的联合优化，搜索空间巨大[6]。

3 面向存算一体架构的 AI 编译关键技术

3.1 硬件抽象

为捕捉存算一体的硬件特性（如阵列结构、模拟域操作等），研究者提出了相关的硬件抽象和仿真工具。

PIMulator[7]是一个快速且灵活的存内处理（PIM）仿真平台，可用于 PIM 架构评估和原型设计，支持用户自定义内存配置和 PIM 逻辑单元的结构与分布。MNSIM[8][9]提出一个行为级仿真的建模工具，支持混合精度的推理仿真，可用于描述数字和模拟存内处理，支持基于忆阻器的神经形态计算系统仿真。

3.2 数据映射与布局优化

CIM 中权重通常预编程（烧录）到 NVM 单元，而激活动态加载，编译器需联合优化权重布局、激活调度和输入输出数据流管理。

权重布局：考虑数据模式（如权重分布、稀疏性）与阵列非理想特性（如电导变化范围）的关系，优化权重在阵列中的物理位置分配，以最小化计算误差或最大化并行度[10]。激活调度：通过优化激活向量在多个阵列间的分配、加载顺序和重用策略，减少对有限带宽的片外存储访问和片上网络拥塞。输入/输出数据流管理：设计高效的数据缓冲（Buffer）方案和 DMA 调度策略，匹配 CIM 核的计算吞吐率，避免成为瓶颈。

PIMCOMP-NN[11]提出了一个面向存算阵列架构的神经网络通用编译框架。通过对硬件架构和算子的统一抽象构建指令集架构[12]，PIMCOMP-NN 实现了从神经网络结构描述到指令流的自动化生成，并在编译过程中优化任务映射和调度。

3.3 精度感知与误差补偿编译

在编译流程中（如 Cost Model）嵌入器件/电路级非理想模型（如电导漂移

模型、ADC/DAC 噪声模型)，使优化决策考虑精度影响，在编译时进行误差补偿。

编译时误差补偿：通过权重微调、操作重排序/选择、混合精度分配等方法进行误差补偿。权重微调，即编译器指导在量化/映射后，基于硬件误差模型对模型进行微调，补偿系统误差[13][14]；操作重排序/选择，即选择对误差更不敏感的操作序列或近似算法[15]。混合精度分配[16]，即根据网络层/算子对误差的敏感度和 CIM 阵列支持的不同精度档位，进行细粒度的混合精度映射策略优化。

4 代表性编译框架与系统

部分代表性的面向存算一体架构的 AI 编译器框架/系统如表 1 所示：

表 1 代表性存算一体 AI 编译框架

框架/系统	发表年份	目标硬件	核心贡献/关键技术
PRIME Compiler[17]	2016	ReRAM CIM	首个通过编译实现神经网络映射、数据布局、数据流控制等。
OCC[18]	2022	通用 CIM	基于 MLIR 进行多层 IR 重写,实现自动化的端到端编译。
PIMCOMP[11]	2023	通用 CIM	提出了一个面向存算阵列架构的神经网络通用编译框架。
CIM-MLC[19]	2024	通用 CIM	提出多级调度策略,同时支持粗粒度到细粒度的算子,可充分释放 CIM 架构的加速潜力
C4CAM[20]	2024	基于 CAM 的 CIM	第一个支持 CAM 配置和从 Torch 脚本端到端生成加速器代码的编译器框架。
CINM(Cinnamon)[21]	2024	异构 CNM+CIM	可为近存计算 (CNM) 和存内计算 (CIM) 异构架构提供统一的编程模型和优化支持。

5 未来趋势展望

为促进存算一体的实际推广应用，面向存算一体的 AI 编译器需要在多方面进行研究突破。

构建标准化抽象层：CIM 硬件碎片化严重，缺乏统一的编程模型和硬件抽象层（类似 CUDA 之于 GPU），导致编译器移植成本高、优化通用性差。定义开放、分层的 CIM 抽象接口（如指令集、内存模型）是生态建设的关键。

模拟计算误差的高效建模与补偿：现有器件模型与补偿策略仍需提高精度、降低开销。探索基于学习的轻量级在线误差预测与补偿机制，以及与训练框架的深度联合设计是重要方向[22]。

大模型支持：当前优化多聚焦于 CNN 骨干或小规模 MLP/ViT。高效编译支持

Transformer、图神经网络（GNN）、大语言模型（LLM）等复杂模型在多层/多核 CIM 系统上的映射是前沿课题[23]。

工具链成熟度与评估基准：相比成熟的 CPU/GPU 编译生态，存算一体编译器工具链仍处于早期。建立涵盖精度、能效、面积、吞吐等多维度的标准化基准测试集（如扩展 MLPerf Tiny Benchmark[24]）至关重要。

未来研究还将向多层次协同优化（算法-编译-架构-器件）、支持新兴存算一体计算范式（存内逻辑、存内搜索、存内训练）、与先进封装集成（Chiplet、3D IC）紧密结合的编译器设计，以及探索存算一体与类脑计算（神经形态计算）编译[25][26]等方向发展。

参考文献

- [1] Ielmini D, Wong H S P. In-memory computing with resistive switching devices[J]. *Nature electronics*, 2018, 1(6): 333-343.
- [2] Ambrogio S, Narayanan P, Tsai H, et al. Equivalent-accuracy accelerated neural-network training using analogue memory[J]. *Nature*, 2018, 558(7708): 60-67.
- [3] Guo R, Liu Y, Zheng S, et al. A 5.1 pJ/neuron 127.3 us/inference RNN-based speech recognition processor using 16 computing-in-memory SRAM macros in 65nm CMOS[C]//2019 Symposium on VLSI Circuits. IEEE, 2019: C120-C121.
- [4] Chen T, Moreau T, Jiang Z, et al. TVM: An automated End-to-End optimizing compiler for deep learning[C]//13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18). 2018: 578-594.
- [5] Lattner C, Amini M, Bondhugula U, et al. MLIR: Scaling compiler infrastructure for domain specific computation[C]//2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO). IEEE, 2021: 2-14.
- [6] Wang J, Ge M, Ding B, et al. NicePIM: Design Space Exploration for Processing-In-Memory DNN Accelerators With 3-D Stacked-DRAM[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023, 43(5): 1456-1469.
- [7] Mosanu S, Sakib M N, Tracy T, et al. PiMulator: A fast and flexible processing-in-memory emulation platform[C]//2022 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2022: 1473-1478.
- [8] Xia L, Li B, Tang T, et al. MNSIM: Simulation platform for memristor-based neuromorphic computing system[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017, 37(5): 1009-1022.
- [9] Zhu Z, Sun H, Xie T, et al. Mnsim 2.0: A behavior-level modeling tool for

- processing-in-memory architectures[J]. IEEE transactions on computer-aided design of integrated circuits and systems, 2023, 42(11): 4112-4125.
- [10] Chen P Y, Lin B, Wang I T, et al. Mitigating effects of non-ideal synaptic device characteristics for on-chip learning[C]//2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2015: 194-199.
- [11] Sun X, Wang X, Li W, et al. PIMCOMP: A Universal Compilation Framework for Crossbar-based PIM DNN Accelerators[C]//2023 60th ACM/IEEE Design Automation Conference (DAC). IEEE, 2023: 1-6.
- [12] Chen X. Instruction set architecture (isa) for processing-in-memory dnn accelerators[J]. arXiv preprint arXiv:2308.06449, 2023.
- [13] Joshi V, Le Gallo M, Haefeli S, et al. Accurate deep neural network inference using computational phase-change memory[J]. Nature communications, 2020, 11(1): 2473.
- [14] Rasch M J, Mackin C, Le Gallo M, et al. Hardware-aware training for large-scale and diverse deep learning inference workloads using in-memory computing-based accelerators[J]. Nature communications, 2023, 14(1): 5282.
- [15] Duan C, Qi Y, Yang J. Emerging Trends and Advances in Computing-in-Memory: A Software-Hardware Co-Design Perspective[C]//2025 Conference of Science and Technology of Integrated Circuits (CSTIC). IEEE, 2025: 1-3.
- [16] Sun S, Bai J, Shi Z, et al. CIM²PQ: An Arraywise and Hardware-Friendly Mixed Precision Quantization Method for Analog Computing-In-Memory[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024, 43(7): 2084-2097.
- [17] Chi P, Li S, Xu C, et al. Prime: A novel processing-in-memory architecture for neural network computation in rram-based main memory[J]. ACM SIGARCH Computer Architecture News, 2016, 44(3): 27-39.
- [18] Siemieniuk A, Chelini L, Khan A A, et al. OCC: An automated end-to-end machine learning optimizing compiler for computing-in-memory[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021, 41(6): 1674-1686.
- [19] Qu S, Zhao S, Li B, et al. Cim-mlc: A multi-level compilation stack for computing-in-memory accelerators[C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. 2024: 185-200.
- [20] Farzaneh H, De Lima J P C, Li M, et al. C4CAM: A Compiler for CAM-based In-memory Accelerators[C]//Proceedings of the 29th ACM International Conference on Architectural

Support for Programming Languages and Operating Systems, Volume 3. 2024: 164-177.

- [21] Khan A A, Farzaneh H, Friebe K F A, et al. Cinm (cinnamon): A compilation infrastructure for heterogeneous compute in-memory and compute near-memory paradigms[C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4. 2024: 31-46.
- [22] Eldebiky A, Zhang G L, Böcherer G, et al. Correctnet: Robustness enhancement of analog in-memory computing for neural networks by error suppression and compensation[C]//2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023: 1-6.
- [23] Wolters C, Yang X, Schlichtmann U, et al. Memory is all you need: An overview of compute-in-memory architectures for accelerating large language model inference[J]. arXiv preprint arXiv:2406.08413, 2024.
- [24] Banbury C, Reddi V J, Torelli P, et al. Mlperf tiny benchmark[J]. arXiv preprint arXiv:2106.07597, 2021..
- [25] Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing[J]. Nature nanotechnology, 2020, 15(7): 529-544.
- [26] Davies M, Wild A, Orchard G, et al. Advancing neuromorphic computing with loihi: A survey of results and outlook[J]. Proceedings of the IEEE, 2021, 109(5): 911-934.

面向脉冲神经网络的存算一体芯片综述与展望

陈华^{1,2}, 夏红杰², 朱赛克³, 赵毅^{1,3*}

1. 华东师范大学集成电路科学与工程学院, 上海 200241

2. 中国电子科技南湖研究院, 嘉兴 314000

3. 浙江大学信息与电子工程学院, 杭州 310027

摘要: 脉冲神经网络 (Spiking Neural Network, SNN) 通过模拟生物神经元的脉冲通信机制, 在低功耗和实时处理领域展现出独特优势。存算一体 (Computing-in-Memory, CIM) 技术则通过将计算功能嵌入存储单元, 有效减少了数据搬运, 两者结合催生了兼具高能效与低延迟的新型计算架构。本文回顾了面向脉冲神经网络的存算一体芯片研究进展, 并讨论了未来发展趋势。

关键词: 脉冲神经网络; 存算一体; 忆阻器; 神经形态计算

1 前言

脉冲神经网络 (Spiking Neural Network, SNN) 作为第三代神经网络^[1], 通过模拟生物神经元的脉冲通信机制, 在低功耗和实时处理领域展现出独特优势。存算一体 (Computing-in-Memory, CIM) 技术则通过将计算与存储深度融合, 有效减少了传统冯·诺依曼架构的数据搬运。两者结合催生了兼具高能效与低延迟的新型计算架构, 成为物联网、低功耗人工智能应用的重要技术方向。本文介绍了面向脉冲神经网络的存算一体芯片研究进展, 并展望了未来趋势。

2 脉冲神经网络和存算一体技术概述

SNN 是一种模仿人类大脑神经元工作方式的神经网络, 它以脉冲信号进行信息传输和处理, 具有异步、稀疏等特性。与传统神经网络相比, SNN 中的神经元仅在接收到输入脉冲时才进行计算, 这种事件驱动的方式避免了大量无效计算, 大大降低了能耗。在实时处理方面, SNN 能够快速响应输入信号的变化, 适用于对实时性要求较高的应用场景, 如物联网边缘设备的实时监测。

存算一体概念最早再 1969 年提出^[2], 该技术通过将存储单元与计算单元融合, 使数据在存储单元直接进行计算, 避免了数据在存储和计算单元之间的频繁搬运, 从而降低功耗并提升计算效率。随着人工智能等算力密集型应用的兴起, 传统计算机架构因性能瓶颈无法满足需求, 存算一体技术被提出, 并逐渐从学术研究走向实际应用。

根据计算范式的不同, 存算一体可分为数字式和模拟式两类, 如图 1 所示。数字式存算一体采用全数字逻辑实现, 对工艺、电压和温度变化具有较强的鲁棒性。模拟式存算一体的核心是利用忆阻器阵列实现乘累加操作, 且数据输入输出需要进行数模和模数转换, 一般而言, 其计算吞吐量高于数字式存算一体, 具有

基金项目: 新一代人工智能国家科技重大专项 (2020AAA0109001); 国家自然科学基金资助项目 (No. U23B2040).

更出色的能效。根据存储介质的不同，包括基于传统存储器，如 Flash、SRAM，以及新型非易失性存储器，如 RRAM、FeRAM、MRAM 等。

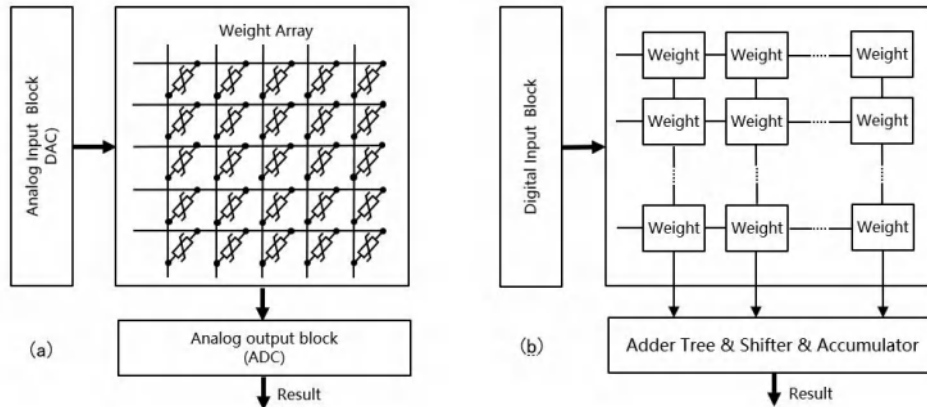


图 1 存算一体计算范式示意：(a) 模拟式存算一体结构；(b) 数字式存算一体结构

3 面向脉冲神经网络的存算一体芯片研究进展

目前，SNN 主要使用积分-发放（Integrate - and - Fire, IF）神经元模型或泄露积分-发放（Leaky Integrate - and - Fire, LIF）神经元模型，其工作原理如图 2 所示。IF 神经元模型将神经元的运算过程分为积分和放电两个阶段，在积分阶段，神经元接收来自其他神经元的输入，并对这些输入进行加权求和以获得一个总输入信号，当总输入信号达到神经元的阈值时，神经元进入放电阶段，输出一个脉冲信号，然后将其电压重置为静息电位。相较于 LF 模型，LIF 模型增加了神经元电压泄漏的模拟，即“泄漏”过程。面向 SNN 的存算一体芯片架构核心在于将神经元的突触权重存储与脉冲信号的乘累加运算集成于同一物理单元，通过硬件层面的优化高效实现上述神经元模型。

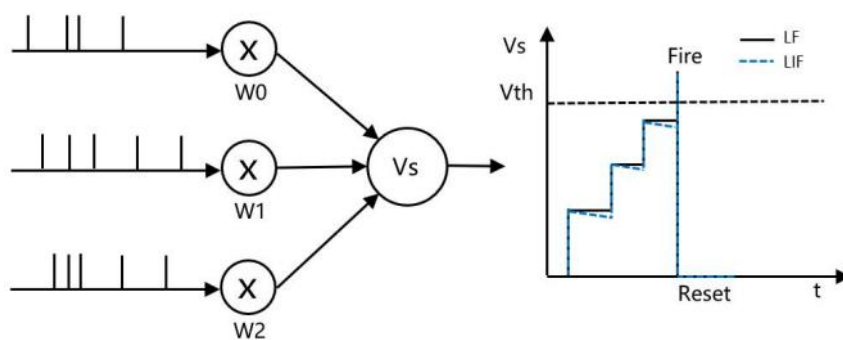


图 2 IF和LIF神经元模型工作示意图

SRAM 由于其成熟可靠，往往是 SNN 存算一体设计的主流选择。文献^[3]提出了一种低泄漏的 8T SRAM 单元（LL8T SRAM），用于实现神经元的突触，并生成神经元的电流型树突信号，电路如图 3 所示。该设计同时提出了一种低功耗、紧凑且全数字的数字到脉冲转换器，该转换器由四个 XOR 门和一个 AND 门组成，通过一个全局的 4 位计数器将数字输入转换为脉冲信号，芯片在 90 纳米 CMOS 技术下

制造，其实现了 SNN 的完整输出层，测量结果显示该设计具有 196 纳秒的平均推理延迟和 81.4% 的推理准确率，功耗为 242 微瓦，能量效率为 4.74 皮焦耳/推理/神经元。另一种电荷域 SRAM-CIM^[4] 架构利用金属-氧化物-金属 (MOM) 电容实现电荷共享计算，将 MNIST 分类的单次推理能耗降至 15.09nJ，延迟仅 0.46 μ s。基于 9T1C 单元的时间域 SNN-CIM (TS-CIM)^[5,6] 通过基于 RC 延迟的时域内位单元计算方法和多比特累加方法，实现 701.7 TOPS/W 的能效比，较传统电流域设计提升 1.58 倍。

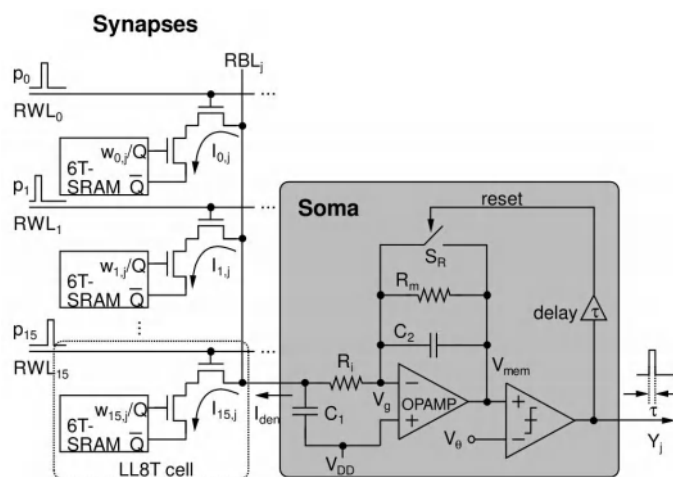


图 3 基于 LL8T SRAM 的 SNN 神经元电路

基于 DRAM 构建面向 SNN 的存算一体芯片方面，SNN eDRAM-CIM^[7] 提出了一种高密度的可重构神经单元阵列 (ReNCA) 和动态可调神经连接 (DAN Link)，通过这两种关键特性来实现高效能。具体来说，1T1C eDRAM 架构通过以下方式提高神经网络计算的系统效率：利用充电泵和重用 1T1C 单元阵列，显著减少了 41% 的面积和 90% 的功耗；通过双模 ReNCA 和 DAN Link，系统能够根据不同层的配置动态调整内存和外围逻辑的比例。该芯片达到了 1531.3 TOPS/W 的效率，并且在系统效率上比之前的 SNN CIM 提高了 10 倍。

DS-CIM^[8] 提出了一种基于 MRAM 的无 ADC、双脉冲 CIM 宏单元，在保持极低能耗的同时，实现了很高的准确率。在数字分类、图像分类、手势识别和动作识别任务中，其准确率分别达到了 96.99%、82.87%、90.00% 和 85.97%，能耗仅为 8.07nJ、71.26nJ、729.3nJ 和 369.82nJ，为边缘设备的低功耗推理提供了方案。

铁电场效应晶体管 (FeFET) 也可用来实现面向 SNN 的存算一体芯片，该设计^[9] 提出了一种基于铁电场效应晶体管 (FeFET) 的神经形态计算的存算一体器件和电路设计，其既可用于人工神经网络的乘加运算，也可以支持脉冲神经网络的积分运算。FeFET 的高可扩展性和高开关比有助于实现大容量的神经网络计算存储一体，其在神经形态计算有较大潜力。

此外，异构融合架构也具有独特的优势。异构 RRAM-SRAM 混合架构^[10] 通过 RRAM-CIM 处理卷积神经网络 (如 MobileNet)，利用其非易失性实现权重存储，

同时通过 SRAM-CNM 处理 SNN 的脉冲神经网络计算，在目标跟踪任务中节省了 91.8% 的功耗。这种架构通过两层功率门控机制，使 RRAM 阵列在非工作状态时能耗趋近于零。

4 结论

面向脉冲神经网络的存算一体芯片通过存储与计算的融合，在低功耗、能效、延迟方面有着独特的优势，为边缘人工智能和物联网提供了可行的技术路径。虽然，现阶段仍有诸多问题需要解决，比如模拟计算的精度和可靠性、软件生态的不完善限制了推广和应用等，未来，随着新型存储器件、软硬件协同设计方法的发展，该架构有望成为下一代低功耗人工智能计算的核心范式。

参考文献

- [1] I. Garg, S. S. Chowdhury, and K. Roy, DCT-SNN: Using DCT to distribute spatial information over time for low-latency spiking neural networks, in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct.2021, pp. 4651 – 4660.
- [2] KAUTZ W H. Cellular logic-in-memory arrays[J]. IEEE Transactions on Computers, 1969, C-18(8): 719 – 727. doi:10.1109/T-C.1969.222754.
- [3] Chao-Yu Chen, Yan-Siou Dai, Hao-Chiao Hong. A Neuromorphic Spiking Neural Network Using Time-to-First-Spike Coding Scheme and Analog Computing in Low-Leakage 8T SRAM[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2024, 32(5): 848-859.
- [4] Yuchao Zhang, Zihao Xuan, Yi Kang. A 28nm 15.09nJ/inference Neuromorphic Processor with SRAM-Based Charge Domain in-Memory Computing[C]. IEEE ASICON 2023, 2023.
- [5] Keonhee Park, Hoichang Jeong, Seungbin Kim, et al. A 701.7 TOPS/W Compute-in-Memory Processor With Time-Domain Computing for Spiking Neural Network[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2025, 72(1): 25-35.
- [6] Keonhee Park, Hoichang Jeong, Kyuho Lee. A 701.7 TOPS/W Time-Domain Spiking Neural Network Compute-in-Memory Processor with 9T1C Bitcell[C]. IEEE MWSCAS 2023, 2023.
- [7] Seryeong Kim, Soyeon Kim, Soyeon Um, et al. A Reconfigurable 1T1C eDRAM-based Spiking Neural Network Computing-In-Memory Processor for High System-Level Efficiency[C]. IEEE ISCAS 2023, 2023.
- [8] Haotian Fu, Yulong Huang, Tingran Chen, et al. DS-CIM: A 40nm Asynchronous Dual-Spike Driven, MRAM Compute-In-Memory Macro for Spiking Neural Network[J]. IEEE Transactions on Circuits and Systems—I: Regular Papers, 2024, 71(4): 1638-1650.
- [9] Chihiro Matsui, Kasidit Toprasertpong, Shinichi Takagi, et al. FeFET Local Multiply and

-
- Global Accumulate Voltage-Sensing Computation-In-Memory Circuit Design for Neuromorphic Computing[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2024, 32(3): 468-479.
- [10] Ashwin Sanjay Lele, Muya Chang, Samuel D. Spetalnick, et al. A Heterogeneous RRAM In-Memory and SRAM Near-Memory SoC for Fused Frame and Event-Based Target Identification and Tracking[J]. IEEE Journal of Solid-State Circuits, 2024, 59(1): 52-64.

通过噪声感知训练提升存算芯片的推理精度

Youngseo Kim¹, 鲁辞莽², 赵毅^{1,3*}

1 华东师范大学集成电路科学与工程学院,上海 200241

2 上海闪易半导体有限公司,上海 201210

3. 浙江大学信息与电子工程学院, 杭州 310027

摘要: 随着对高性能低功耗神经网络部署需求的增长, 处理器内存一体化 (Processing-In-Memory, PIM) 架构因其高效性而受到关注。然而, 其模拟计算引入的噪声易导致推理精度下降。为此, 本文提出噪声感知训练 (Noise-Aware Training, NAT) 方法, 在训练阶段引入基于真实 PIM 芯片误差建模的噪声, 从而提升模型对噪声的鲁棒性。实验表明, NAT 在不影响模型原始精度的前提下, 显著提升了 PIM 芯片上的推理表现, 优于传统训练与量化感知训练, 具备广泛应用潜力。

关键词: 深度学习; 模型优化; 存算一体; 量化感知训练; 语音识别

资助: 本文受新一代人工智能国家科技重大专项 (2020AAA0109001) 和华东师范大学/华大半导体有限公司集成电路工程技术联合实验室资助。

1 引言与背景

近年来, 随着实际应用需求增长及半导体技术进步, 神经网络算法取得了显著发展。其中, 处理器内存一体化 (Processing-In-Memory, PIM) 架构作为一种高效的芯片技术, 在神经网络部署中表现突出。相比传统方法, PIM 在速度、功耗与成本方面具有明显优势, 如图 1(a)所示。其高效性主要源于利用欧姆定律 ($I=V \times 1/R$) 在存储单元内直接执行乘加 (MAC) 运算, 既降低能耗, 又减少了外部权重数据的传输开销^[1-7]。

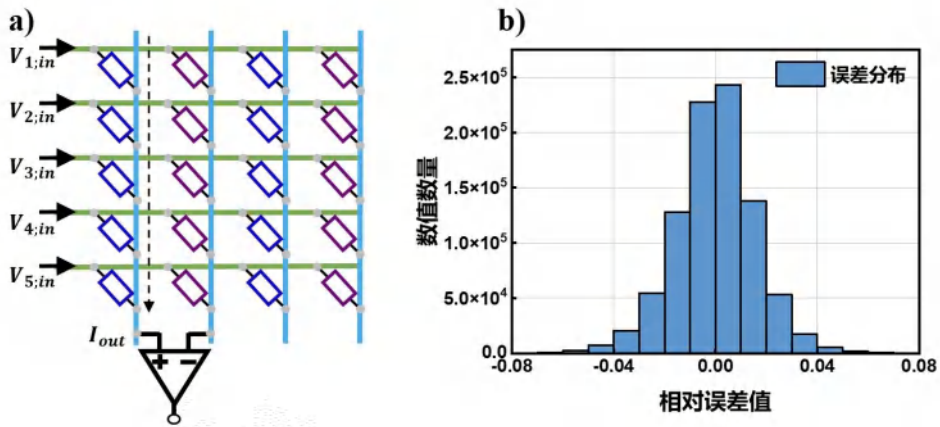


图1 (a) 基于欧姆定律的PIM架构。输入电压加在字线，权重由存储单元电导表示，输出电流从位线读取。(b) 本PIM芯片执行MAC操作的误差分布，相较于GPU的FP32结果。误差已按MAC结果幅度归一化。

尽管如此，PIM 在计算过程中仍存在精度损失问题。实验显示，其 MAC 结果与 32 位浮点（FP32）结果之间存在相对误差，虽与现有 PIM 芯片相当，但对精度敏感任务仍构成限制，如图 1(b)所示。为应对该挑战，研究者已提出包括量化优化、噪声补偿及非线性失真缓解等多种策略，并探索了如 UPMEM、Mensa 与 SIMDRAM 等先进 DRAM-PIM 架构，及 CHIME 用于降低内存端读放大。然而，模拟运算噪声问题依然难以根除。

针对上述问题，本文提出一种噪声感知训练方法，在训练阶段引入 PIM 芯片中观测到的 MAC 噪声特性，使模型在部署后能更好适应 PIM 平台的固有限制，从而有效提升其在精度敏感任务中的推理可靠性与性能。

2 实验方法

为验证本文提出的噪声感知训练方法在 PIM 芯片部署中的有效性，我们基于 TensorFlow 平台构建了一个全连接（Fully-Connected, FC）层的关键词语音识别模型。

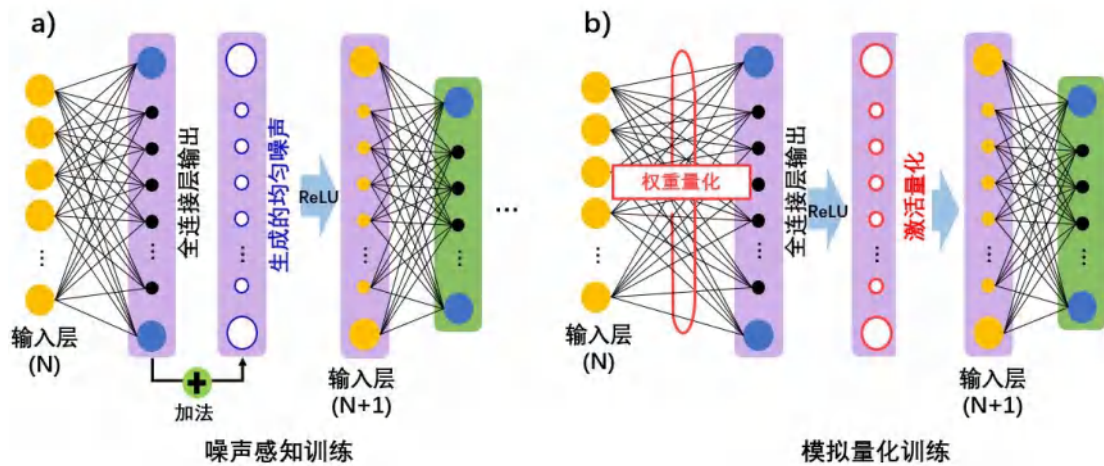


图2 (a) NAT模型的算法示意图。每个神经元加入均匀噪声。(b) 量化训练 (QAT) 流程。权重和激活值均被量化，以适配边缘芯片中的量化MAC运算。

NAT 训练方法通过在模型训练阶段，于每个 FC 层的输出中引入均匀分布的噪声，以模拟 PIM 芯片中乘加 (MAC) 运算所带来的误差，如图 2(a)所示。该噪声来源于 PIM 与 FP32 计算结果之间的相对误差，其幅度设定在 4% 至 8% 之间。噪声注入的计算公式如下：

$$\text{Noise} = (\text{UniformNoise}(\text{size}) - 0.5) \times (\text{valmax} - \text{valmin}) \times \text{Noiseintensity}$$

其中，噪声遵循 $[-0.5, 0.5]$ 区间的均匀分布，并根据当前批次张量的动态范围及设定噪声强度进行缩放。所有 FC 层均采用该策略处理输出，再输入下一层网络。

为对比模型在低比特运算下的鲁棒性，我们同时采用了，量化感知训练 (Quantization-Aware Training, QAT) 方法，如图 2(b)所示。QAT 是广泛应用于边缘设备部署的训练技术，通过在前向传播中引入“伪量化”步骤来模拟权重与激活值的量化过程，而反向传播仍保留 FP32 精度。这使得模型可在训练阶段学习量化噪声影响，从而提高低位精度运算下的表现。此外，PIM-QAT 方法还可有效缓解由模数转换器 (ADC) 误差、热噪声和非线性效应带来的精度下降。

为评估三种训练策略 (NAT、QAT、基线) 的有效性，我们采用统一数据集，对相同结构的五层全连接神经网络进行训练与测试，并分别在 PIM 芯片与 FP32 GPU 上进行推理精度对比。所有模型均在 Visual Studio Code 开发环境下使用 TensorFlow Keras 2.2.0 版本训练完成。QAT 模型训练结束后，采用 TensorFlow Lite (TFLite) 进行量化，并提取模型权重以进行推理测试。

3 实验结果和分析

为评估噪声注入对芯片端推理精度的影响，本文比较了四种训练配置下的模

型表现: 无噪声注入的基线五层模型、噪声强度为 4% 的 NAT 模型(4%-NAT)、噪声强度为 8% 的 NAT 模型 (8%-NAT)，以及量化感知训练 (QAT) 模型。

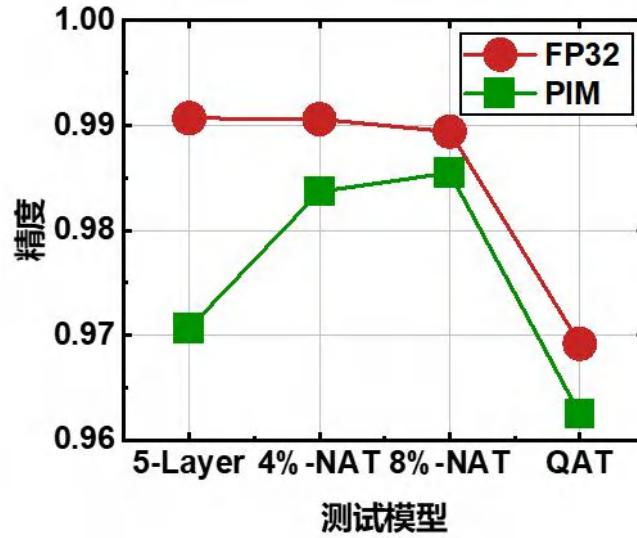


图3 基于FP32和PIM芯片的推理精度对比，测试对象为不同5层模型。

图 3 展示了上述模型在 FP32 GPU（即训练环境）与 PIM 芯片上的推理精度对比。结果表明，NAT 模型在 FP32 环境下的推理准确率与基线模型一致，均保持在 99%，说明 NAT 方法不会降低模型的固有精度。然而，在 PIM 芯片部署中，NAT 模型相较于基线模型表现出更低的精度损失。这一现象合理地反映出，NAT 在训练阶段已融入了模拟噪声特性，从而提升了模型对芯片误差的鲁棒性。

相比之下，QAT 模型在从 FP32 到 PIM 芯片的迁移过程中虽然准确率下降较小，但其在 FP32 下的本身精度即低于基线与 NAT 模型，约低 2%。这揭示了 QAT 方法的固有限制：量化操作虽可提升硬件兼容性，却牺牲了模型的整体性能。其精度下降主要源于每个 FC 层输出值呈长尾分布，导致有效动态范围难以充分利用。

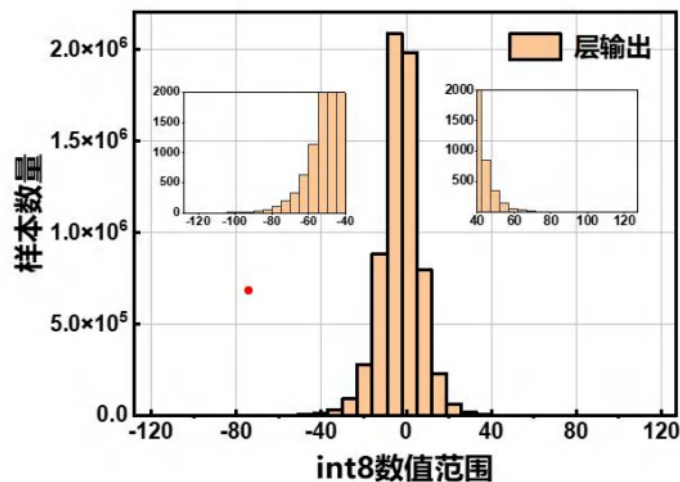


图4 QAT模型中全连接层量化输出的直方图结果。

图4展示了QAT模型第一个FC层量化输出的直方图。量化过程基于测试集中输出的最大与最小值设定范围，但结果显示，绝大多数样本集中于较窄的区间（如-10至10），远未充分覆盖整个量化区间，导致量化误差显著积累，进而在多层计算传播中造成较大精度损失。相比之下，NAT模型对PIM器件误差表现出更强鲁棒性，更适用于存在模拟噪声的芯片推理场景。

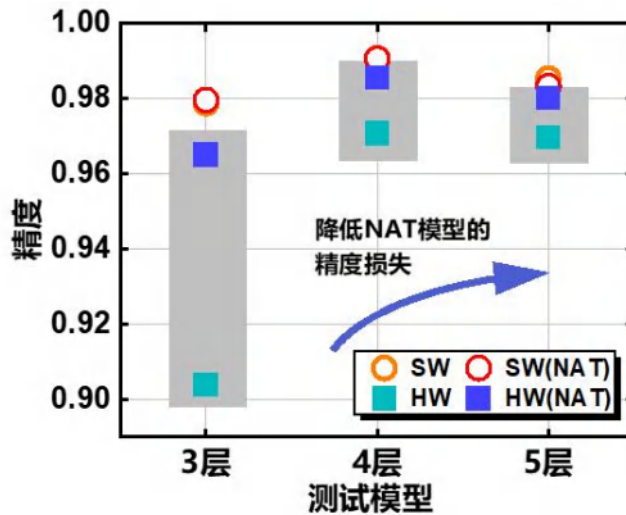


图5 在不同层数模型中应用噪声后，各模型精度的比较结果。

为验证NAT方法在不同网络结构中的适应性，我们进一步测试了在4层至6层网络结构中引入噪声后的推理表现。图5展示了不同模型结构下在FP32与PIM芯片上的推理精度。实验结果表明，NAT模型在所有结构中均展现出相较于未加噪声模型更小的精度损失，验证了NAT方法的通用性和适配性。

本实验进一步指出，NAT尤其适用于以大规模矩阵运算为主的神经网络任务，如语音识别与信号处理等应用场景。而在图像识别等领域，QAT依然被广泛采用。未来，若将NAT与QAT方法结合，有望在边缘计算、嵌入式AI以及实时处理等任务中进一步提升深度学习模型的精度与性能。

4 结论

在本研究中，我们提出了一种**噪声感知训练（Noise-Aware Training, NAT）**方法，用于缓解PIM芯片部署中常见的精度下降问题。该方法在训练阶段向每个MAC层的输出添加均匀分布的噪声（噪声幅度基于真实PIM芯片测得的数据建模），有效减轻了精度退化，同时不影响模型本身的准确性。实验结果表明，NAT方法在性能上优于传统训练方式和量化感知训练（Quantization-Aware

Training, QAT), 在 PIM 芯片上实现了显著更小的精度损失。这一改进为 PIM 架构在高性能、能效敏感的实际应用中的广泛应用提供了有力支持。

参考文献

- [1] Angizi, Shaahin, et al. "Accelerating deep neural networks in processing-in-memory platforms: Analog or digital approach?." 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). IEEE, 2019.
- [2] Jhang, Chuan-Jia, et al. "Challenges and trends of SRAM-based computing-in-memory for AI edge devices." IEEE Transactions on Circuits and Systems I: Regular Papers 68.5 (2021): 1773-1786.
- [3] Chi, Ping, et al. "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory." ACM SIGARCH Computer Architecture News 44.3 (2016): 27-39. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
- [4] Ahn, Junwhan, et al. "A scalable processing-in-memory accelerator for parallel graph processing." Proceedings of the 42nd Annual International Symposium on Computer Architecture. 2015.
- [5] Chen, Jia, et al. "Multiply accumulate operations in memristor crossbar arrays for analog computing." Journal of Semiconductors 42.1 (2021): 013104.
- [6] Chou, Teyuh, et al. "Cascade: Connecting rrams to extend analog dataflow in an end-to-end in-memory processing paradigm." Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture. 2019.
- [7] Zou, Xingqi, et al. "Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology." Science China Information Sciences 64.6 (2021): 160404.

基于注意力的磁随机存储器缺陷分类检测

陈晖*, 赵毅

浙江大学信息与电子工程学院, 杭州 310027

摘要: 随着磁随机存储器 (MRAM) 制造工艺的不断进步, 对高精度缺陷检测的需求日益迫切。然而, 传统检测方法在处理大规模 MRAM 缺陷数据时, 存在识别效率低、准确率不足等问题。为此, 本文提出一种基于注意力增强的 MRAM 缺陷检测算法: 首先, 在高分辨率网络 (HRNet) 架构中引入全局注意力模块以强化特征提取能力; 其次, 在特征融合阶段采用注意力引导机制, 通过注意力分数筛选关键特征; 最后, 在自制 MRAM 缺陷数据集和公开 NTU-CLS 数据集上进行对比实验。结果表明, 该算法在 MRAM 缺陷数据集和 NTU-CLS 数据集上的平均识别准确率均显著优于基线算法, 展现出优异的检测性能和泛化能力。

关键词: STT-MRAM; 缺陷检测; 注意力机制; HRNet

基金项目: 新一代人工智能国家科技重大专项 (2020AAA0109001)

1 引言与背景

在半导体工艺逼近物理极限的背景下, 自旋转移矩磁随机存储器 (STT-MRAM) 因其非易失性、高耐久性和 CMOS 兼容性成为下一代存储器的候选技术。MRAM 的核心结构磁隧道结 (MTJ) 制造工艺复杂, 缺陷检测对产品质量至关重要。深度学习技术为 MRAM 缺陷分类提供了高效解决方案。近年来, Transformer 模型如 Swin Transformer^[1]及其改进版本 (如 Zhu^[2]和 Zhang^[3]提出的方法) 通过注意力机制增强了全局特征提取能力, 但计算复杂度较高。当前, 其应用于图像缺陷识别的主要挑战在于如何平衡局部与全局特征提取的效率与精度, 以进一步提升缺陷分类的性能。

Transformer^[4]最初为自然语言处理而设计, 其自注意力机制能有效捕捉长距离依赖关系, 后被引入计算机视觉领域。多尺度融合技术是计算机视觉中处理目标尺度变化的关键方法, 通过生成和融合多尺度特征提升模型鲁棒性。特征生成通常采用两种策略: (1) 不同步长的卷积核, 以差异化的采样频率提取特征, 计算效率较高; (2) 输入多分辨率图像, 分别提取特征。特征融合方法包括简单操作^[5]和基于注意力的自适应方法^[6], 后者能动态加权关键特征。

针对缺陷分类, 本研究提出基于高分辨率网络 (HRNet)^[7]的改进模型, 利用其多尺度结构保持高分辨率特征, 有效捕捉不同尺寸缺陷的细节; 通过池化下采样和全局自注意力模块建模长程依赖, 结合关联上采样和上下文机制增强关键特征定位, 从而显著提升分类性能。

2 实验方法

2.1 全局注意力模块改进

全局特征提取模块是提取全局特征的关键，直观的方法是使用视觉Transformer（Vision Transformer, ViT）的全局注意力。然而，HRNet 在每个阶段都需保持高分辨率输出，导致使用基于 ViT 的全局注意力时计算负担较大。此外，ViT 在处理小规模图像数据集时缺乏足够的归纳偏差。因此，本研究设计了一个全新的全局注意力模块（图 1），通过池化下采样（Pooling Downsampling）将过程输入进行下采样，然后执行多头自注意力（Multi-head self-attention, MS）以捕捉下采样空间中的全局依赖关系，最后通过 AU 算法将下采样空间映射回输入图像的大小。

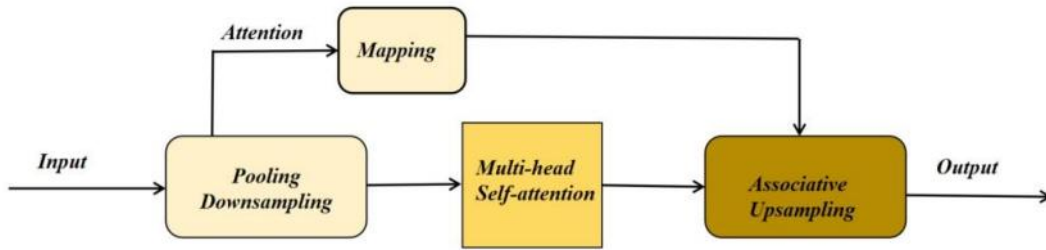


图 1 全局特征提取模块的结构图

2.2 多尺度特征融合模块改进

HRNet 在每阶段特征融合后，多分支交互会引入冗余信息，增加计算负担并干扰关键特征提取。为此，我们提出注意力引导的提纯模块（Attention-Guided Block, AGB），以抑制冗余响应并增强关键特征表示。该模块的核心功能在于对融合后的特征进行提纯优化，精准识别并剔除那些冗余信息，从而将注意力高度集中在关键特征上。具体定义如下：

$$X_A = X_F + \text{Soft max}\left(\frac{X_O X_K^T}{\sqrt{d}}\right) X_V \quad (1)$$

$$Y = X_A + \text{FFN}(\text{LN}(X_A)) \quad (2)$$

其中， X_F 表示融合后的特征， $X_O = \text{LN}(X_F)$ ， $X_K = \text{GAP}(X_O)$ ， $X_V = \text{MLP}(X_K)$ 。 $\text{LN}(\cdot)$ ， $\text{GAP}(\cdot)$ ， $\text{MLP}(\cdot)$ ， $\text{FFN}(\cdot)$ 分别表示层归一化、全局平均池化、多层感知器和前馈神经网络。 $\text{MLP}(\cdot)$ 包含三层，隐藏层神经元的数量等于 C/r ，其中 r 为缩减比例。

3 实验结果与分析

为验证改进 HRNet 算法的性能，采用改进的算法对磁随机存储器缺陷数据集进行缺陷识别实验，使用数据集中的训练集对改进算法进行训练，使用数据集中的验证集进行预测，验证改进 HRNet 算法的性能。选用精确度(J)、准确率(A)、召回率(R)、F1 分数(F1-score)指标评估算法的性能，结果如表 2 和表 3 所示。

在 MRAM 数据集的实验结果中能够发现本研究所构建的模型无法准确识别 Defect、Oxide_scratch 这两类缺陷，该问题将在后续的研究中进行改进优化。而在 NTU-CLS 数据集上，除了 inclusion 和 scratches，其余类别都已经达到了 100% 准确率。

表 1 OurModel 在 MRAM 数据集上与其他模型的对比结果

模型	准确率 (%)	精确率 (%)	召回率 (%)	F1 分数 (%)
HRNet	92.0	75.8	79.2	76.8
UNet	91.2	73.8	75.6	74.2
SwinTransformerV2	90.9	72.0	71.3	71.1
OurModel	92.9	78.1	81.0	78.0

注：表现最突出的用加粗字体显示。

表 2 OurModel 在 NEU-CLS 数据集上与其他模型的对比结果

模型	准确率 (%)	精确率 (%)	召回率 (%)	F1 分数 (%)
HRNet	97.2	97.3	97.2	97.2
UNet	98.3	98.4	98.3	98.3
SwinTransformerV2	98.3	98.3	98.3	98.3
GAPC-KLS	97.8	-	-	-
ResNet50+MFN	97.7	-	-	-
OurModel	98.6	98.6	98.6	98.6

注：表现最突出的用加粗字体显示，‘-’表示原论文并没有提供实验数据。

通过将本研究的改进模型与其他方法进行比较以验证该改进模型的有效性。从实验结果中，可以看到 OurModel 的模型效果要明显好于其他对比模型，与基准模型 HrNet 相比，准确率要高 0.9%，精确率要高 2.3%，召回率要高 1.8%，F1 值要高 1.2%。在 NUT-CLS 数据集中，我们各项指标也要好于其他基准模型。

4 结论

在本研究中，我们提出了一种全局注意力缺陷分类模型，用于解决工业缺陷识别中的关键挑战。实验结果表明，该方法在性能上显著优于传统 CNN 和 Transformer 模型，在磁随机存储器缺陷数据集和 NTU-CLS 数据集上分别实现了 92.90%和 98.60%的平均识别准确率。这一改进为工业缺陷检测提供了高效可靠的解决方案，为智能制造的实际应用奠定了重要基础。

参考文献

- [1] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows[J]. arXiv e-prints, 2021. arXiv:2103.14030.
- [2] Wei Zhu, Hui Zhang, Chao Zhang, et al. Surface defect detection and classification of steel using an efficient Swin Transformer[J]. Advanced Engineering Informatics, 2023. 57:1474-0346.
- [3] Zhang, N., Liu, Z., Zhang, E. et al. An ESG-ConvNeXt network for steel surface defect classification based on hybrid attention mechanism[J]. Sci Rep, 2025. 15, 10926.
- [4] Vaswani Ashish, Shazeer Noam, Parmar Niki, et al. Attention is all you need[J]. arXiv e-prints, 2017. arXiv:1706.03762.
- [5] K. He, X. Zhang, S. Ren, et al. Deep Residual Learning for Image Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, 770-778.
- [6] X. Li, W. Wang, X. Hu, et al. Selective Kernel Networks[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, 510-519.
- [7] K. Sun, B. Xiao, D. Liu, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, 5686-5696.

GaN功率HEMT器件SPICE模型研究进展

刘斯扬

东南大学

摘要：功率半导体器件在智能电网、工业控制、航空航天等领域有广泛的应用背景，近年来，作为第三代半导体器件的功率 GaN 高电子迁移率晶体管（HEMT）器件凭借其栅电荷小、开关速度快、无反向恢复等优点受到了业界的广泛关注，成为研究和产业化热点。高精度的器件模型是功率系统设计的基础和核心环节。本报告探讨了 GaN HEMT 器件的电学特性 SPICE 模型，包括经验模型和物理模型的构建与应用，特别介绍了基于物理机理的增强型 GaN HEMT 器件模型（QSE-HEMT）。此外，报告还介绍了宽禁带功率器件寿命预测模型及其 EDA 仿真工具的研发与应用，展示了其在功率器件和功率系统可靠性评估中的重要作用。

SiC光控功率器件的建模与仿真

王曦

西安理工大学

摘要：碳化硅（SiC）作为第三代半导体材料，具有禁带宽度大、临界击穿电场强度高、载流子饱和漂移速度高、热导率高以及化学稳定性好等诸多优点。凭借材料的性能优势，SiC功率器件具有比硅器件更高的耐压、更大的电流密度、更快的开通速度以及更优的耐高温性能。倘若能以光信号替代电信号实现SiC功率器件的直接驱动与控制，将不仅能简化驱动电路、提高系统功率密度，还能大幅提升SiC功率器件极端环境耐受能力，使SiC功率器件的应用场景获得进一步拓展。本文以陕西省功率半导体器件及装备创新团队在SiC功率器件光控技术方面的研究为基础，主要围绕SiC光控晶体管、光触发晶闸管等光控功率器件的建模与仿真工作进行介绍与讨论，为下一步组建小型化、模块化的新一代SiC光控功率集成系统提供理论依据。

Research Article

A Nonsegmented Behavioral Modeling for SiC MOSFETs in PSpice based on Neural Network

Ai-Jun Zhang¹, Yu-Ming Zhang¹, Xiao-Yan Tang¹, Qing-Wen Song¹, Zhi-Yuan Qi¹

¹ The Key Laboratory of Wide Band-Gap Semiconductor Materials and Devices, Xidian University, Xi'an 710071, China
Affiliation, City name and postal code, Country

Corresponding author: Xiao-Yan Tang; Email: xytang@mail.xidian.edu.cn

Manuscript Received March 22, 2022; Accepted March 22, 2022; Published Online March 22, 2022.

Abstract — This article introduces an advanced modeling method for the Silicon Carbide (SiC) metal-oxide-semiconductor field-effect transistors (MOSFETs). This method based on artificial neural network (ANN) to address the problem of existing modeling being unable to describe the current-voltage characteristics (I-V) of the first and third quadrant in a nonsegmented model. Additionally, the voltage-dependent capacitance characteristics (C-V) of intrinsic parasitic elements—gate-source capacitance (C_{GS}), gate-drain capacitance (C_{GD}), and drain-source capacitance (C_{DS})—are modeled using the same ANN framework, enabling comprehensive representation of the device's dynamic behavior. In order to verify the effectiveness of the proposed model method, the SiC MOSFET E3M0075120D (1200 V/75 mΩ) is modeled by ANN. ANN training utilizes I-V and C-V data from datasheet. The optimized ANN architecture comprises a three-layer feedforward network for both the I-V and C-V. Then, the obtained ANN structure is described using the circuit description language in PSpice to realize the MOSFET characteristics, thus completing the establishment of the model. To verified the accuracy of the model, the modeling results is compared with the measured results. Based on the comparison results, the modeling results are largely consistent with the measurement results which means this ANN-based modeling approach is suitable for power SiC MOSFETs.

Keywords — Silicon carbide (SiC), MOSFET, Third quadrant (3rd-quad) characteristics, ANN.

I. Introduction

Silicon Carbide (SiC) has emerged as a promising material for power electronic devices due to its superior properties such as high breakdown voltage, high thermal conductivity, and low on-resistance [1]. SiC metal-oxide-semiconductor field-effect transistors (MOSFETs) have shown great potential for applications in high-power, high-frequency, and high-temperature environments [2]. To fully exploit the advantages of SiC MOSFETs, accurate and reliable device modeling is essential for understanding device behavior and optimizing circuit performance.

Multiple models have been proposed, which can be mainly classified into three categories: physics-based model, behavioral model, and analytical model [3]. The physics-based model is the most accurate and complex model for the device, which need to understand the professional knowledge of semiconductor physics [4],[5]. The establishment of various related physical parameters in the modeling process increases the time for model establishment. The behavioral model is widely used by using mathematical fitting methods and datasheet to build model. This model, solely reflecting device functional characteristics without considering semiconductor physical processes, enhances design efficiency and accelerates simulation speed. An analytical model refers to the ability to use analytical methods such as calculus. Analytical models have advantages in terms of accuracy and reliability, however, the applicability of the models is limited and the mathematical difficulty is high.

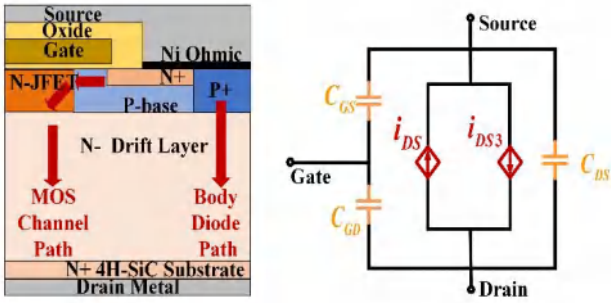
To increase the power density, the third quadrant

(3rd-quad) characteristics of MOSFET instead of the externally paralleled Schottky diode is used for freewheeling during the deadtime. It is well known that the third quadrant characteristics include not only body diode characteristics but also MOS channel characteristics, and parasitic bipolar transistor characteristics, which should be taken into account in the modeling process [6],[7]. There are several ways of modeling the third quadrant with existing models, including the Break model that comes with PSpice, the circuit model that is a combination of a resistor and a current source, and the Model that is built by summing different current paths [8],[9],[10],[11]. The first two models are simpler and do not reflect the third quadrant characteristics of the device well, while the third one involves more physical parameters, which is more accurate but also complex and time-consuming.

Conventional modeling methods have struggled to keep pace with the rapid technological advancements in emerging materials and device architectures [12]. In response, researchers have developed an artificial neural network (ANN) based modeling methodology that enables rapid and precise characterization of device nonlinearities [13],[14]. Therefore, in this paper, a modeling method based on ANN is proposed to complement the existing models. To model the MOSFET, a neural network is first used to describe the characteristics, and then the resulting network structure is described using the circuit description language in PSpice to establish a nonlinear electronic device model.

II. Modeling Framework

The cross-sectional view of the SiC MOSFET and the comment subcircuit model is shown in Figure 1 . As illustrated in Figure 1 (a), SiC MOSFETs exhibit two current pathways in the third quadrant: one through the reverse conduction of the MOSFET channel, and another through the parasitic conduction of the body diode. Both pathways influence the device's switching characteristics during third-quadrant operation. In power switching applications, the third-quadrant behavior of silicon carbide (SiC) MOSFETs holds critical significance, particularly during freewheeling operations. Consequently, when modeling the device's I-V characteristics, the third-quadrant I-V behavior requires



equal attention to ensure accurate performance prediction.

Figure 1 (a) Cross-sectional view of the SiC MOSFET. (b) subcircuit model of the SiC MOSFET.

As shown in Figure 1 (b), the MOSFET subcircuit model can be divided into two main components: I-V characteristics and C-V characteristics. The I-V characteristics are further categorized into first-quadrant behavior (I_{DS}) and third-quadrant behavior (I_{DS3}). For the C-V characteristics, the modeling primarily focuses on three capacitance components: gate-source capacitance (C_{GS}), gate-drain capacitance (C_{GD}), and drain-source capacitance (C_{DS}).

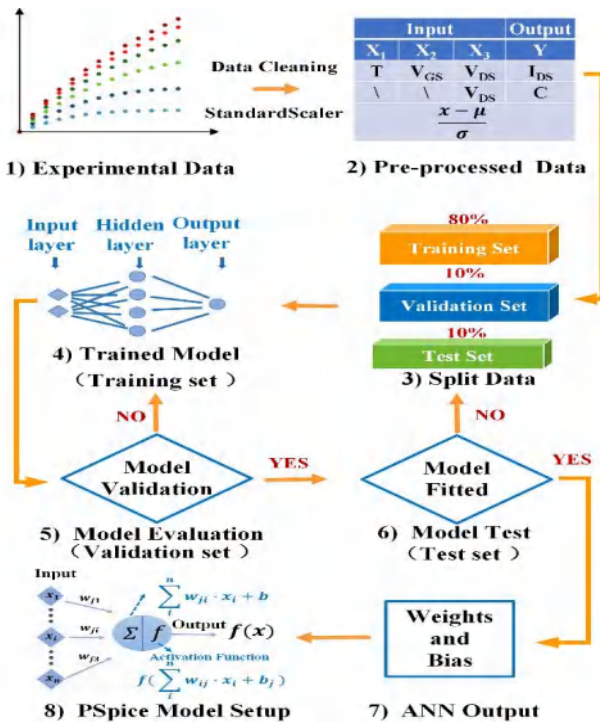


Figure 2 Schematic flowchart of ANN modeling

1. Construction of ANN

The ANN flow can be seen in Figure 2 . The machine learning modeling process begins with collecting experimental data including T , I_{DS} , I_{DS} , I_{DS} , and capacitances (C_{GS} , C_{DS} , C_{DS}). Modeling data is acquired by reading data tables via the GetData software.

In this work, we employ temperature (T), gate-source voltage (I_{DS}), drain-source voltage (I_{DS}) as inputs, while predicting the drain-source current (I_{DS}) and intrinsic parasitic capacitances (C_{GS} , C_{DS} , C_{DS}) as outputs. This raw data undergoes preprocessing where features are normalized using StandardScaler to achieve $N(0,1)$ distribution. The processed dataset is then split into training (80%), validation (10%), and test (10%) subsets. An ANN model is trained on the normalized training set with early stopping to prevent overfitting. The trained model is evaluated using the validation set to tune hyperparameters and assess performance metrics like accuracy and loss. Subsequently, the final model is tested on the untouched test set to measure generalization capability. The ANN outputs predicted parameters including I_{DS} , C_{GS} , C_{DS} , and C_{DS} . Finally, these ANN-generated parameters are integrated into PSpice circuits for circuit simulations, enabling direct comparison between predicted and simulated results.

StandardScaler standardization is a critical step in machine learning data preprocessing, with the core purpose of eliminating dimensional disparities among features and transforming the data into a standard normal distribution, thereby enhancing the efficiency and stability of model training. The formula is expressed as:

$$x_{input} = \frac{x - \mu}{\sigma} \quad (1)$$

where μ represents the mean value of the feature, and σ denotes the standard deviation. μ can be expressed as:

$$\mu_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (2)$$

μ_j denotes the mean value of parameter j . Here, j can represent various variables such as temperature, V_{DS} , and V_{GS} , etc. n indicates the total number of parameters i , and x_{ij} represents the value of variable j at the i -th level. σ can be expressed as:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_j)^2} \quad (3)$$

This process ensures that the model focuses on the relative relationships between features rather than their absolute numerical values, serving as a standard preprocessing procedure in most machine learning workflows.

2. Training of ANN

The TensorFlow package in Python is used to construct the ANN model. The adaptive moment estimation ("Adam") optimization is chosen for training with a learning rate around $6E-5$ and sigmoid as the activation function. This network has been preliminarily selected as a structure with two hidden layer and one output layer. After extensive experimentation, the hidden layer neuron

counts for the I-V and C-V models were optimized to 164 and 32, respectively. The suitable network is selected and trained to achieve the error target. After the training is completed, record the weight (W) and threshold (B) of the function. Finally, describe the constructed network in PSpice and build the device model accordingly.

3. Modeling Accuracy Verification

The developed SiC MOSFET model is implemented in the commercial PSpice simulator. The comparisons between the circuit simulation results and datasheet values are given in this section.

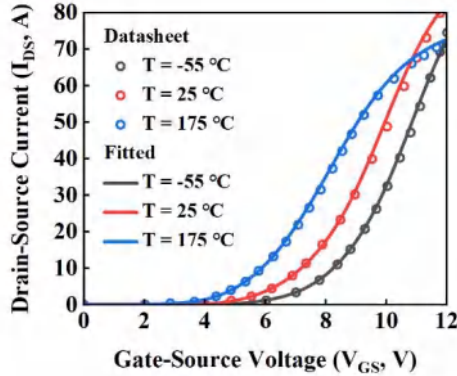
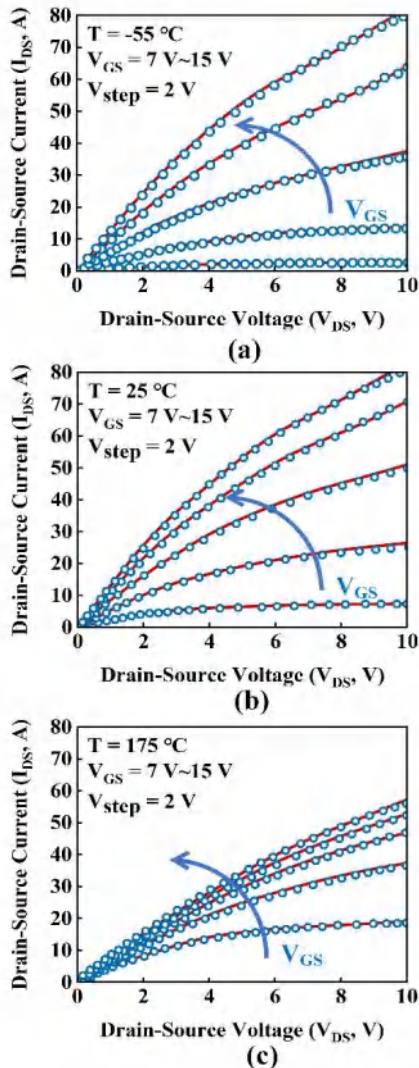


Figure 3 Comparisons of the modeling results (solid lines) and datasheet (scatter lines) in terms of the first quadrant transfer characteristics



at different temperatures. $V_{DS} = 20\text{ V}$.

Figure 4 Comparisons of the modeling results (solid lines) and datasheet (scatter lines) in terms of the first quadrant output characteristics at different temperatures and V_{GS} . (a) $-55\text{ }^\circ\text{C}$. (b) $25\text{ }^\circ\text{C}$. (c) $150\text{ }^\circ\text{C}$.

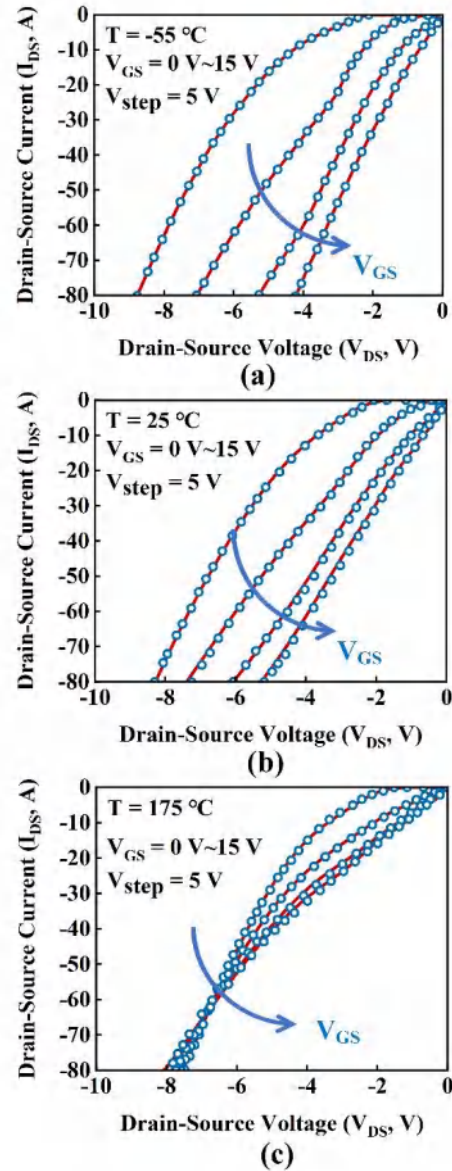


Figure 5 Comparisons of the modeling results (solid lines) and datasheet (scatter lines) in terms of the third quadrant output characteristics at different temperatures and V_{GS} . (a) $-55\text{ }^\circ\text{C}$. (b) $25\text{ }^\circ\text{C}$. (c) $150\text{ }^\circ\text{C}$.

Figure 3, Figure 4 and Figure 5 show the fitting results of transfer and output characteristics. It can be seen that the artificial neural network model demonstrates good fitting capabilities for the first and third quadrant current-voltage (I-V) characteristic data. Across varying temperature conditions, the modeling results fit the device datasheet well. This data-driven approach effectively circumvents the complex device physics analysis required by conventional models, significantly reducing the computational burden associated with fitting parameter extraction.

Figure 6 shows the comparisons between the proposed model and datasheet values in describing the device capacitances. The results validate the model's high precision in characterizing voltage-dependent capacitance behavior, showing excellent agreement between predicted curves and measured data across the full voltage range.

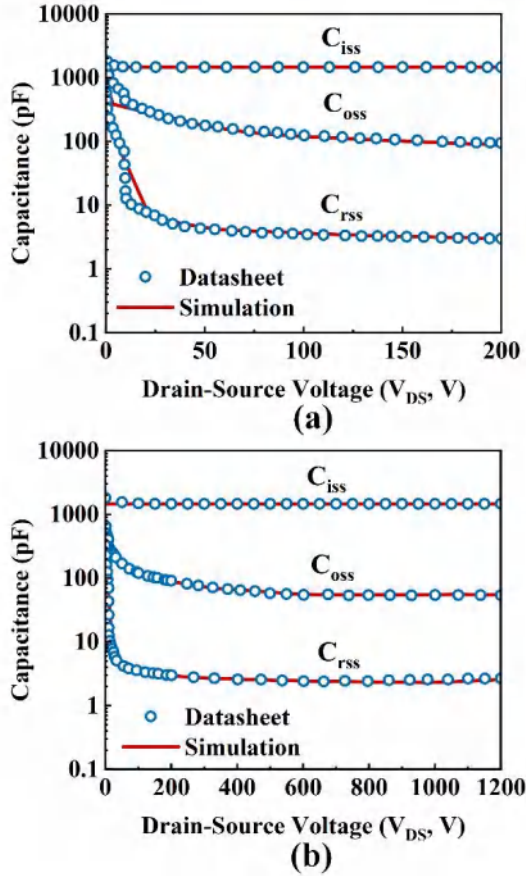


Figure 6 Comparisons between the modeling results (solid lines) and datasheet (scatter lines) in terms of the device capacitances.

III. Conclusion

This paper presents a simple ANN-based modeling approach for 1.2 kV SiC MOSFETs, featuring a nonsegmented model that simultaneously captures first and third quadrant I-V characteristics along with C-V behaviors. Comprehensive validation was performed using a commercial 1200 V/75 mΩ SiC MOSFET (E3M0075120D). Through seamless integration of ANN-derived models into PSpice circuit simulations, the proposed methodology achieves accurate characterization of static and dynamic device behaviors. Comparative analysis confirms the model's capability to precisely replicate SiC MOSFET operational characteristics across different temperature.

Acknowledgement

This work was supported by the National Key R&D Program of China (2023YFB3609503).

Reference

- [1] S. Safari, A. Castellazzi and P. Wheeler, "Evaluation of SiC power devices for a high power density matrix converter", *2012 IEEE Energy Conversion Congress and Exposition (ECCE)*, Raleigh, NC, USA, pp. 3934-3941, doi: 10.1109/ECCE.2012.6342297, 2012.
- [2] X. Deng, X. Xu, X. Li, et al., "A Novel SiC MOSFET Embedding Low Barrier Diode With Enhanced Third Quadrant and Switching Performance", in *IEEE Electron Device Letters*, vol. 41, no. 10, pp. 1472-1475, doi: 10.1109/LED.2020.3017650, Oct. 2020.
- [3] J. Wang et al., "Characterization, Modeling, and Application of 10-kV SiC MOSFET", in *IEEE Transactions on Electron Devices*, vol. 55, no. 8, pp. 1798-1806, doi: 10.1109/TED.2008.926650, Aug. 2008.
- [4] R. Kraus and A. Castellazzi, "A physics-based compact model of SiC power MOSFETs", in *IEEE Transactions on Power Electronics*, vol. 31, no. 8, pp. 5863-5870 doi: 10.1109/TPEL.2015.2488106, Aug. 2016.
- [5] L. Wu, L. Xiao, J. Zhao, et al., "Physical analysis and modeling of the nonlinear Miller capacitance for SiC MOSFET", in *Proc. IECON43rd Annu. Conf. IEEE Ind. Electron. Soc.*, pp. 1411-1416, doi: 10.1109/IECON.2017.8216240, Oct. 2017.
- [6] L. Tang et al., "Analyzing the Changes in the Third Quadrant Characteristics of SiC MOSFET Induced by Threshold Drift", in *IEEE Transactions on Electron Devices*, vol. 71, no. 4, pp. 2342-2348, April doi: 10.1109/TED.2024.3362773, 2024.
- [7] L. Tang et al., "Investigation Into the Third Quadrant Characteristics of Silicon Carbide MOSFET", in *IEEE Transactions on Power Electronics*, vol. 38, no. 1, pp. 1155-1165, doi: 10.1109/TPEL.2022.3202705, Jan. 2023.
- [8] K. Sun, H. Wu, J. Lu, et al., "Improved Modeling of Medium Voltage SiC MOSFET Within Wide Temperature Range", in *IEEE Transactions on Power Electronics*, vol. 29, no. 5, pp. 2229-2237, May 2014, doi: 10.1109/TPEL.2013.2273459.
- [9] A. P. Arribas, F. Shang, M. Krishnamurthy, et al., "Simple and Accurate Circuit Simulation Model for SiC Power MOSFETs", in *IEEE Transactions on Electron Devices*, vol. 62, no. 2, pp. 449-457, doi: 10.1109/TED.2014.2384277, Feb. 2015.
- [10] D. Bavi and S. Khandelwal, "Self-consistent Compact Modeling of First- and Third Quadrant I-V behavior in SiC MOSFETs," *2022 IEEE Latin American Electron Devices Conference (LAEDC)*, Cancun, Mexico, pp. 1-4, doi: 10.1109/LAEDC54796.2022.9908186, 2022.
- [11] T. Yang, X. Li, S. Yin, et al., "A Datasheet-Driven Nonsegmented Empirical SPICE Model of SiC MOSFET With Improved Accuracy and Convergence Capability", in *IEEE Transactions on Electron Devices*, vol. 70, no. 1, pp. 4-12, doi: 10.1109/TED.2022.3220481, Jan. 2023.
- [12] W. Liu, W. Na, W. Zhang, et al., "A Review of Recent Neural Network Approaches to the Modeling of Nonlinear Microwave Devices", *2020 13th UK-Europe-China Workshop on Millimetre-Waves and Terahertz Technologies (UCMMT)*, Tianjin, China, pp. 1-3, doi: 10.1109/UCMMT49983.2020.9296017, 2020.
- [13] S. B. Kutub, H. -J. Jiang, N. -Y. Chen, et al., "Artificial Neural Network-Based (ANN) Approach for Characteristics Modeling and Prediction in GaN-on-Si Power Devices", *2020 32nd International Symposium on Power Semiconductor Devices and ICs (ISPSD)*, Vienna, Austria, pp. 529-532, doi: 10.1109/ISPSD46842.2020.9170110, 2020.
- [14] B. Kim and M. Shin, "A Novel Neural-Network Device Modeling Based on Physics-Informed Machine Learning", in *IEEE Transactions on Electron Devices*, vol. 70, no. 11, pp.6021-6025, doi: 10.1109/TED.2023.3316635, Nov. 2023.

1200V SiC 凹状基区光控晶体管的特性仿真与分析

刘乐晨, 王曦, 熊娟, 汪雅馨, 李姚, 蒲红斌

1) 西安理工大学电子工程系, 陕西省 西安市 710048

2) 西安市电力电子器件与高效电能变换重点实验室, 陕西省 西安市 710048

摘要: SiC 光控功率器件具有驱动简单、抗电磁干扰能力强等优点。本文通过 Sentaurus TCAD 建模, 仿真研究了 1200V SiC 凹状基区晶体管的特性, 并与传统结构进行了对比。结果显示, 在击穿电压相同 (1476V) 时, 凹状基区光控晶体管的电流增益可以达到 345.9A/W, 较传统 SiC 光控晶体管提升了 38.91%。在电流增益相同 (249A/W) 时, 凹状基区光控晶体管的击穿电压可以达到 1775V, 相较传统 SiC 光控晶体管提升了 20.35%。

关键词: SiC 光控晶体管; 凹状基区; 电流增益

1 前言

碳化硅(SiC)材料由于其禁带宽度大、临界雪崩击穿电场强度高特点[1-2], 在高压、高频、大功率等领域展现出巨大的应用潜力[3]。为了解决 SiC 双极器件驱动电路的复杂性问题, 采用光控替代电控模式, 如 SiC 光控晶体管[4]。传统光控晶体管作为功率开关应用时面临着增益与耐压之间相互制约的显著矛盾。

本文基于 SiC 光控晶体管常规结构, 提出具有凹状基区结构的新型 SiC 光控晶体管, 在满足器件 1200V 耐压的前提下, 扩大基区窗口, 有效提高电流增益。

2 方法

在 Sentaurus TCAD 器件仿真平台建立一个凹状基区结构的 SiC 光控晶体管的元胞结构如图 1 (b) 所示。通过优化设计厚基区单元的宽度、个数、间距以及掺杂浓度等结构参数, 可以在维持器件高耐压性能的同时, 有效提高电流增益。

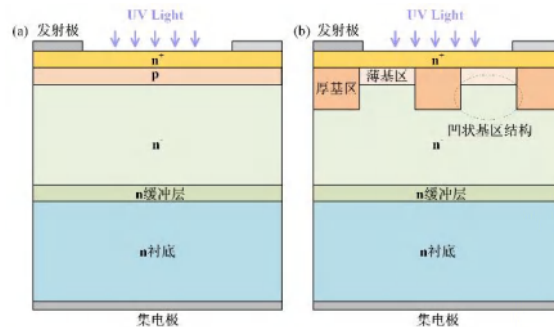


图 1 SiC 光控晶体管结构示意图 (a) 传统结构 (b) 凹状基区结构

3 结果分析

如图 2 所示, 仿真结果表明凹状基区结构的引入显著改善了 SiC 光控晶体管的增益特性与耐压之间相互制约的矛盾。从图 2(a)可以看出在击穿电压同为 1476V 的条件下, 传统 SiC 光控晶体管电流增益为 249A/W, 而具有凹状基区结

构的 SiC 光控晶体管的电流增益可以达到 345.9A/W，较前者提升了 38.91%。这是由于凹状基区结构的存在，薄基区厚度和掺杂浓度可以进一步减小，而无需牺牲器件的耐压能力，这使得注入到基区的载流子更快地穿越基区到达集电区，减少了基区内的复合几率，从而显著提升了电流增益性能。两种结构的阻断特性对比如图 2(b)所示，在电流增益相同（249A/W）的条件下，传统 SiC 光控晶体管击穿电压为 1476V，而具有凹状基区结构的 SiC 光控晶体管的击穿电压可以达到 1775V，相较前者提升了 20.35%。相比于传统结构，凹状基区结构电场分布更加均匀，峰值电场有所降低，从而提高了器件的击穿电压。同时厚基区与 n-集电区之间形成的反偏 pn 结进一步分担了电压，提高了整体的耐压能力。图 2(c)表明在整个电流密度范围内，凹状基区结构 SiC 光控晶体管的静态导通功耗始终低于传统 SiC 光控晶体管。当电流密度为 100A/cm² 时，传统 SiC 光控晶体管导通功耗为 51.6W，凹状基区 SiC 光控晶体管导通功耗为 46.9W，相比降低了 9.1%。

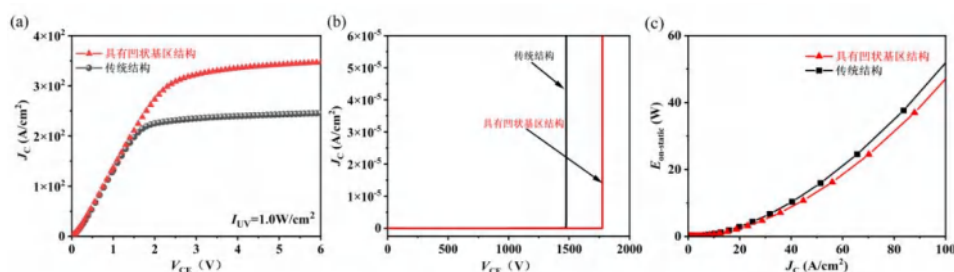


图 2 两种结构的 (a) 输出特性 (b) 阻断特性 (c) 静态导通功耗

4 结论

为了改善传统光控晶体管作为功率开关应用时所面临的增益与耐压之间相互制约的显著矛盾，本文提出了一种在电流增益和耐压性能上相较于传统光控晶体管表现出明显优势的凹状基区光控晶体管。在击穿电压相同（1476V）时，凹状基区光控晶体管的电流增益可以达到 345.9A/W，较传统 SiC 光控晶体管提升了 38.91%。在电流增益相同（249A/W）时，凹状基区光控晶体管的击穿电压可以达到 1775V，相较传统 SiC 光控晶体管提升了 20.35%。

参考文献

- [1] Nguyen N K, Nguyen T, Nguyen T K, et al. Wide-band-gap semiconductors for biointegrated electronics: recent advances and future directions[J]. ACS Applied Electronic Materials, 2021, 3(5): 1959-1981.
- [2] Z. Guo, C. Wang, L. Guan and W. Yang, "Research on Breakdown Mechanism and Characteristics of 4H-SiC BJT with Trench-JTE Composite Termination," 2024 3rd International Symposium on Semiconductor and Electronic Technology, 2024, 80-83,
- [3] Shiwei Liang, Jun Wang, Fang Fang, Linfeng Deng, and Z. John Shen. Optimization Design of Isolation Rings in Monolithically Integrated 1200V SiC Transistor and Antiparallel Diode for Improved Blocking Voltage[J]. IEEE Journal of Emerging and Selected Topics in Power Electronics, 2019, 7(3): 1513-1518.
- [4] 张颖颖.一种具有阶梯变掺杂基区的 SiC 光控晶体管[J]. 半导体技术,2018,43(8):598-602.

An Accurate 3.3 kV SiC MOSFET Behavioral Model Considering Temperature Effects

Jie Wang, Qingwen Song, Aijun Zhang, Yibo Zhang, Xiaoyan Tang, Hao Yuan, and Yuming Zhang

(School of Microelectronics, Xidian University, Xi'an 710071, China)

Corresponding author: Qingwen Song; Email: qwsong@xidian.edu.cn.

Abstract — Silicon Carbide (SiC) MOSFETs show significant advantages in diverse application scenarios due to their superior performance. However, the existing SPICE models developed for low-voltage devices cannot be effectively applied to high-voltage (> 3.3 kV) SiC MOSFETs in practical applications, which may cause unreliable circuit simulations. This paper presents a behavioral model for 3.3 kV SiC MOSFETs that considers temperature effects. The model focuses on static and dynamic characteristics using non-segmented equations, ensuring smooth and accurate predictions of the device's output and transfer characteristics. Besides, temperature-dependent parameters are incorporated into the equations to correctly reflect the correlation between characteristics and temperature. For the capacitance model, the new compensation equation is added to the commonly used capacitance fitting equations, and the accuracy of this enhanced capacitance model is verified by fitting the capacitance experimental results of the device using the new capacitance equations. A comparison of simulation results with experimental results indicates that the proposed model accurately reflects both the static and dynamic characteristics of the device. The adoption of non-segmented effectively avoids convergence problems inherent in traditional segmented models, while the integration of compensation equations brings simulation results closer to real operating conditions. Moreover, the incorporation of temperature-dependent parameters expands the model's applicability, enabling it to better support related design processes.

Keywords — SiC MOSFET, Behavioral model, Temperature effect, Parasitic Capacitance.

I. Introduction

With the development of power electronics technology, the performance of power electronic devices has been continuously improving. Among them, silicon carbide metal-oxide-semiconductor field-effect transistors (SiC MOSFETs) are one of the silicon carbide power devices that have been widely applied across diverse scenarios. In contrast to traditional silicon-based MOSFETs, they exhibit advantages such as high current density, low on-resistance, excellent heat dissipation performance, and fast switching speed, making them a major research focus worldwide^[1].

Compared with low-voltage MOSFET devices, the technology for high-voltage (> 3.3 kV) SiC MOSFETs is still immature. A key challenge hindering the application of high-voltage devices is the lack of a matching simulation model for high-voltage SiC MOSFETs. In order to fully leverage the advantages of high-voltage SiC MOSFETs in power systems, it is necessary to develop a simple and effective device model^[2]. However, there are few device models applicable to high-voltage SiC MOSFETs currently, and most of the proposed models fail to accurately predict the performance of such devices^[3]. Some models focus on internal physical mechanisms, exhibiting high accuracy, however, due to the high complexity and slow simulation speed, their application in circuit simulations is limited^[4,5]. In addition, the static characteristics of some models are built through segmented equations, which can easily cause non-convergence in practical simulation and thus cannot accurately predict device characteristics^[6,7]. Among

different types of device models, behavioral model aims to abstract the actual electrical characteristics of devices and use simple mathematical models to predict changes in various characteristics under different conditions. With its efficiency and practicality, behavioral model has irreplaceable advantages in power electronic system design.

In this paper, a behavioral model considering temperature effects for the self-developed 3.3 kV SiC MOSFETs is established. Non-segmented equations are used for describing static characteristics, where the transfer curve model and the output curve model are independent of each other, and temperature-dependent parameters are added through equation fitting. For the dynamic characteristics model, the existing capacitance model is adjusted to better fit the experimental results of high-voltage devices. The simulation results of the established model can well reflect the static and dynamic characteristics of the device under different temperatures.

II. Model Construction

The basic equivalent circuit model structure of the 3.3 kV SiC MOSFET is shown in Figure 1. This model includes a voltage-controlled current source I_{DS} , which is used to express the transfer characteristic and output characteristic of the device, three parasitic capacitances (gate-source capacitor C_{GS} , gate-drain capacitor C_{GD} , drain-source capacitor C_{DS}) and a reverse-parallel body diode D_{Body} , all of which can be used to describe the dynamic characteristics of the device.

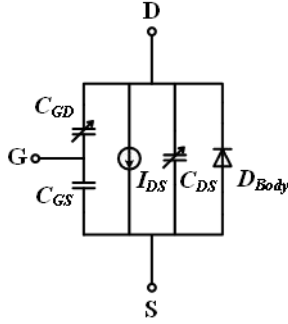


Figure 1 3.3 kV SiC MOSFET equivalent circuit model

1. Modeling of the Static Characteristic

A voltage-controlled current source is utilized to model the static characteristics. In this model, the drain-source current I_{DS} is characterized by the non-segmented equation. It can be divided into two parts: an independent output characteristic equation I_{D_out} (Equation (1)) and a transfer characteristic equation I_{D_yth} (Equation (2)), which is a simplified version of the Angelov model^[8].

$$I_{D_out} = p \cdot \left(1 - \frac{1}{1 + q \cdot V_{DS}} \right) \quad (1)$$

$$I_{D_yth} = l \cdot \{ 1 + \tanh [a \cdot V_{GS} + b \cdot (V_{GS} + c)^2 + d] \} \quad (2)$$

where p and q are fitting parameters in the output characteristic equation, while l , a , b , c , and d are fitting parameters in the transfer characteristic equation.

In order to characterize both the output and transfer characteristics of the device at the same time, Equation (1) and Equation (2) are multiplied to obtain the final static characteristic equation^[9], as shown in Equation (3).

$$I_{DS} = I_{D_out} \cdot I_{D_yth} = p \cdot \left(1 - \frac{1}{1 + q \cdot V_{DS}} \right) \cdot l \quad (3)$$

$$\cdot \{ 1 + \tanh [a \cdot V_{GS} + b \cdot (V_{GS} + c)^2 + d] \}$$

Since the output characteristic curve of the SiC MOSFET is dependent on both the gate-source voltage V_{GS} and the drain-source voltage V_{DS} , while the equation only describes the variation of the current I_{D_out} with V_{DS} , it is necessary to set parameters p and q as variables that change with V_{GS} . Their variation relationships with V_{GS} are characterized by Equation (4) and Equation (5).

$$p(V_{GS}) = \frac{p_1}{1 + \exp[-p_2 \cdot (V_{GS} - p_3)]} \quad (4)$$

$$q(V_{GS}) = q_1 \cdot \exp(-q_2 \cdot V_{GS}) + q_3 \quad (5)$$

where p_1 , p_2 , p_3 , q_1 , q_2 , and q_3 are all fitting parameters, which reflect the dependence relationship between parameters p , q and V_{GS} .

1) Temperature Parameters to the Transfer Curve

Equation (2), which describes the transfer characteristic, contains five model parameters. To avoid excessive complexity in the parameter extraction process, we first select parameters with higher temperature sensitivity through fitting iterations. Ultimately, parameters a , b , and c are fixed, while parameters d and l are defined as temperature-dependent parameters and described by new equations.

$$d(T) = d_1 \cdot T^3 + d_2 \cdot T^2 + d_3 \cdot T + d_4 \quad (6)$$

$$l(T) = l_1 \cdot T^2 + l_2 \cdot T + l_3 \quad (7)$$

where d_1 , d_2 , d_3 , and d_4 are temperature-dependent parameters for parameter d ; l_1 , l_2 , and l_3 are temperature-dependent parameters for parameter l . The extracted parameter values are shown in Table 1.

Table 1 Transfer characteristic fitting parameters

Parameters	Values	Parameters	Values
a	0.4806	l_3	0.6736
b	-0.0164	d_1	1.716e-07
c	-1.5864	d_2	-0.0001
l_1	-7.962e-07	d_3	0.0105
l_2	-0.0012	d_4	-4.0129

2) Temperature Parameters to the Output Curve

Considering that the output characteristic model and the transfer characteristic model are independent of each other, it is necessary to separately add temperature-dependent parameters to the output characteristic model. Equation (8) and Equation (9) are used to characterize the dependence of output parameters p and q on temperature, and the parameter values obtained after fitting with experimental data are shown in Table 2 and Table 3.

$$p_i(T) = m_{i1} \cdot e^{m_{i2} \cdot T} + m_{i3} \cdot e^{m_{i4} \cdot T}, \quad i=1,2,3 \quad (8)$$

$$q_j(T) = n_{j1} \cdot e^{n_{j2} \cdot T} + n_{j3} \cdot e^{n_{j4} \cdot T}, \quad j=1,2,3 \quad (9)$$

Table 2 Temperature-dependent parameters table (p)

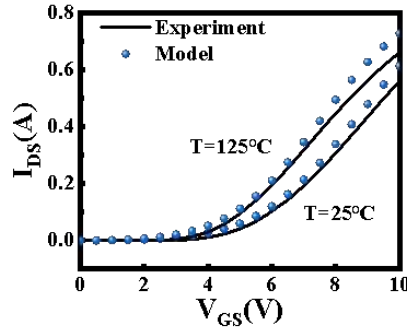
Parameters	Values	Parameters	Values
m_{11}	2439.5	m_{31}	16.223
m_{12}	-0.0424	m_{32}	-0.0338
m_{13}	301.86	m_{33}	21.499
m_{14}	0.0005	m_{34}	-0.0015
m_{21}	0.2523	m_{23}	-0.0546
m_{22}	0.0022	m_{24}	-0.0194

Table 3 Temperature-dependent parameters table (q)

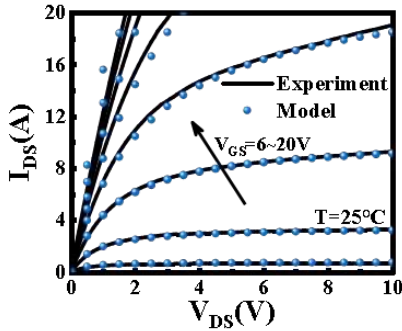
Parameters	Values	Parameters	Values
n_{11}	0.0142	n_{31}	0.1267
n_{12}	0.0267	n_{32}	-0.0095
n_{13}	17.474	n_{33}	0.002
n_{14}	0.0018	n_{34}	0.0077
n_{21}	0.0003	n_{23}	0.3301
n_{22}	0.0227	n_{24}	0.0008

A static characteristic test circuit was built in PSpice to verify the accuracy of the model after parameter extraction. The simulation results of the established model and the experimental results of the device are shown in Figure 2. From the comparison of transfer curves at different temperatures, it can be seen that under low V_{GS} , the simulation results of the transfer characteristic model are in good agreement with the actual results. When V_{GS} is relatively high, the simulation results of the model are slightly higher than the actual test results, but the effect can be neglected. Figure 2(b) and Figure 2(c) show the output characteristic curves at different temperatures respectively. By comparing the simulation results and experimental results in the figures, it can be observed that the established static model can

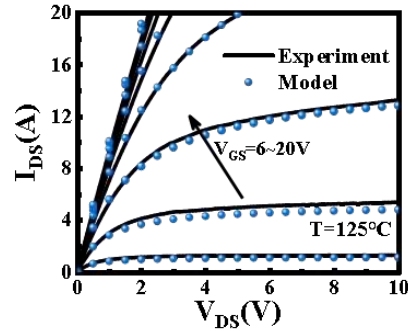
well reflect the output characteristic curves of the device under different temperatures and different V_{GS} .



(a) Transfer characteristic curve



(b) Output characteristic curve (T=25°C)



(c) Output characteristic curve (T=125°C)

Figure 2. Comparison of static characteristics simulation considering temperature effect

2. Modeling of the Dynamic Characteristic

The dynamic characteristics of SiC MOSFETs mainly refer to their switching characteristics, which can reflect various transient behaviors of the device during switching processes. These characteristics directly affect the switching loss, efficiency, and reliability of the circuit. Since the essence of switching characteristics is closely related to various parasitic capacitances inside SiC MOSFETs, this section focuses on modeling these parasitic capacitances.

As shown in Figure 1, there are mainly three parasitic capacitances, namely the gate-source capacitance C_{GS} , the drain-source capacitance C_{DS} , and the gate-drain capacitance C_{GD} . However, the capacitances characterized in the capacitance test curves are actually C_{iss} , C_{oss} , and C_{rss} . The dependency relationships between the above-mentioned capacitances are shown in Equation (10).

$$\begin{cases} C_{iss} = C_{GS} + C_{GD} \\ C_{oss} = C_{DS} + C_{GD} \\ C_{rss} = C_{GD} \end{cases} \quad (10)$$

1) Parasitic Capacitance C_{GD}

The gate-drain capacitance C_{GD} is actually formed by the series combination of the oxide capacitance C_{OX} at the overlap region between the gate and drain and the depletion layer capacitance C_{GDj} . The composition of C_{GD} can be expressed by Equation (11). As the voltage changes, the depletion layer capacitance C_{GDj} varies, as shown in Equation (12). Therefore, the capacitance C_{GD} is not a constant value but a variable capacitance that exhibits a strong voltage dependence.

$$C_{GD} = \frac{C_{OX} \cdot C_{GDj}}{C_{OX} + C_{GDj}} \quad (11)$$

$$C_{GDj} = \frac{\epsilon_s \cdot S_{GD}}{W_{GDj}} \quad (12)$$

where ϵ_s is permittivity of the SiC material, S_{GD} is the capacitance area, and W_{GDj} is the width of the depletion layer capacitance.

The modeling of the variable capacitance C_{GD} is achieved by using a voltage-controlled current source to model the current I_{GD} flowing through the capacitance C_{GD} . The current I_{GD} flowing through the capacitance C_{GD} can be expressed by Equation (13):

$$I_{GD} = C_{GD} \frac{dV_{GD}}{dt} \quad (13)$$

where V_{GD} is the voltage across this capacitance. Since there is a dependency relationship between the capacitance value of the variable capacitor and the applied voltage, it can be represented by $f(V_{GD})$. Therefore, the capacitance value of C_{GD} can be expressed as:

$$C_{GD} = C_0 \cdot f(V_{GD}) \quad (14)$$

where C_0 is a constant capacitance (set according to actual requirements). The expression for the current flowing through the capacitance C_{GD} is shown in Equation (15):

$$I_{GD} = C_{GD} \frac{dV_{GD}}{dt} = f(V_{GD}) \cdot \left(C_0 \frac{dV_{GD}}{dt} \right) \quad (15)$$

In addition, the relationship between the capacitance C_{GD} and $f(V_{GD})$ is shown in Equation (16):

$$f(V_{GD}) = \frac{C_{GD}}{C_0} \quad (16)$$

The equivalent circuit model of the variable capacitor established based on the capacitance C_{GD} is shown in Figure 3, and the constant capacitance C_0 is set to 1 pF.

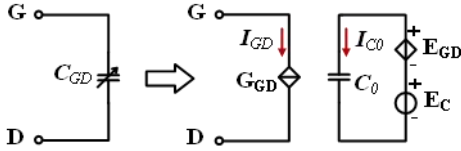


Figure 3 Equivalent circuit model of gate drain capacitor

According to existing research results, Equation (17) is first used to fit the measured capacitance C_{GD} ^[10,11]. Here, C_{GD0} is the value of C_{GD} when V_{GD} is 0 V; C_{GDmin} is the value of C_{GD} when the voltage increases continuously until the capacitance change becomes extremely small; and g_1 is a fitting parameter.

$$C_{GD} = \frac{C_{GD0} \left[\frac{\pi}{2} + \arctan\left(-\frac{V_{GD}}{g_1}\right) \right]}{\pi/2} + C_{GDmin} \quad (17)$$

After simulation verification, it was found that the model based on the above fitting equation cannot accurately match the experimental results. Due to limitations inherent in the fitting equation itself, there are differences between the capacitance variation trends in the model simulation and experiment. Therefore, this paper proposes a compensation method applied to the original model, using the following compensation equation:

$$\omega(V_{GD}) = g_2 \cdot \frac{1}{(1+V_{GD})^{g_3}} \quad (18)$$

where g_2 and g_3 are fitting parameter values. The fitting equation for C_{GD} after incorporating the compensation equation can be expressed as:

$$C_{GD} = C_{GD0} \frac{\frac{\pi}{2} + \arctan\left(-\frac{V_{GD}}{g_1}\right)}{\pi/2} + C_{GDmin} - \omega(V_{GD}) \quad (19)$$

The relevant parameter values extracted with Equation (19) are shown in Table 4.

Table 4 C_{GD} fitting parameters

Parameters	Values	Parameters	Values
C_{GD0}	4222	g_2	7559
C_{GDmin}	18.81	g_3	1.163
g_1	2.691	/	/

2) Parasitic Capacitance C_{DS}

The drain-source capacitance C_{DS} is dominated by the PN junction capacitance formed by the P-type well region and N-type drift region inside the SiC MOSFET. In most cases, it is equivalent to the PN junction depletion capacitance. The relationship between the PN junction depletion layer width and C_{DS} can be expressed as:

$$C_{DS} = \frac{\epsilon_s \cdot S_{DS}}{W_{DS}} \quad (20)$$

where S_{DS} is the capacitance area, and W_{DS} is the width of the depletion layer capacitance.

The equivalent circuit model composition of the drain-source capacitance C_{DS} is the same as that of the capacitance C_{GD} (Figure 3), and the basic modeling progress is also the same. It is necessary to replace the relevant node voltages in Figure 3 with the node voltages corresponding to the capacitance C_{DS} . Additionally, the commonly used fitting equation is shown in Equation

(21):

$$C_{DS} = \frac{C_{DS0}}{\sqrt{1+g_4 \cdot V_{DS}}} \quad (21)$$

where C_{DS0} is the value of C_{DS} when V_{DS} is 0 V, and g_4 is a fitting parameter.

During the fitting process of the capacitance C_{DS} , the same compensation method as for C_{GD} is adopted. Equation (22) is used for difference compensation of the original fitting equation, where both g_5 and g_6 are fitting parameters. After incorporating the compensation equation, the fitting equation for the capacitance C_{DS} is shown in Equation (23). The relevant parameter values extracted are shown in Table 5.

$$\omega(V_{DS}) = g_5 \cdot \frac{1}{(1+V_{DS})^{g_6}} \quad (22)$$

$$C_{DS} = \frac{C_{DS0}}{\sqrt{1+g_4 \cdot V_{ds}}} - \omega(V_{DS}) \quad (23)$$

Table 5 C_{DS} fitting parameters

Parameters	Values	Parameters	Values
C_{DS0}	3401	g_4	1.106
g_5	868.31	g_6	0.475

3) Verification of Parasitic Capacitance Model

The gate-source capacitance C_{GS} is mainly determined the overlap between the gate electrode and both the N⁺ source and P-well regions. Since both the N⁺ source and P-well region are heavily doped regions, the depletion layer width is extremely small, this capacitance is not sensitive to changes of V_{DS} . During the modeling process, the capacitance C_{GS} is set as a constant capacitance with the value of 5235 pF.. It can reduce the complexity of the model, simplify the parameter extraction process, and effectively improve convergence.

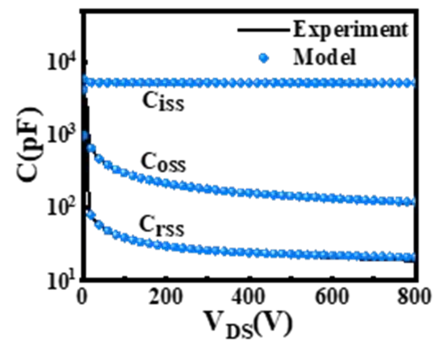


Figure 4 Comparison of Junction Capacitance

The output results of each capacitance model were converted using Equation (10), and the comparison between the simulation results and the experimental results is shown in Figure 4. It can be observed that after compensating and adjusting the commonly used capacitance fitting equations, the fitting errors caused by using traditional equations are effectively remove. It enables the capacitance model to more accurately reflect the variation trend of parasitic capacitances with V_{DS} , which is conducive to fitting the current and voltage curves of the device during switching processes and

improving the overall accuracy of the model.

III. Conclusion

In this paper, an equivalent circuit model for SiC MOSFETs is proposed, which incorporates behavioral modeling for the self-developed 3.3 kV SiC MOSFETs with consideration of temperature effects. It mainly consists of two parts: a static characteristic model and a dynamic characteristic model. The static characteristic model is established to reflect the output characteristic I_{D_out} and transfer characteristic I_{D_vth} of the device through continuous mathematical expressions, and systematically considers the influence of temperature on key parameters. By fitting the static characteristic experimental results under different temperature conditions, the corresponding model parameters are extracted to complete the static characteristic modeling. For the dynamic characteristic model, the junction capacitance model of the device is mainly considered. After compensating the commonly used capacitance fitting equations, the new capacitance equations are used to fit and extract parameters from the experimental results.

The comparison between the model simulation results and experimental results shows that the established model can well reflect the static and dynamic characteristics of the device. The use of non-segmented equations can effectively avoid convergence issues caused by traditional models, and the application of compensation equations makes the model simulation results closer to actual working conditions. In addition, the inclusion of temperature parameters enhances the applicability of the model, providing enhanced support for power electronics design.

Acknowledgement

This work was supported by Project supported by the National Key R&D Program of China (2023YFB3609503).

References

- [1] J. Millán, P. Godignon, X. Perpiñà, et al. "A Survey of Wide Bandgap Power Semiconductor Devices", *IEEE Transactions on Power Electronics*, vol.29, no.5, pp.2155-2163, doi: 10.1109/TPEL.2013.2268900, 2014.
- [2] J. Xie, L. Wang, P. Sun, et al. "An Accurate and Effective Spice Model of 6.5kV SiC MOSFET with Parasitic Parameters Analysis in Medium-Voltage Power Module", in *2024 IEEE 10th International Power Electronics and Motion Control Conference*, Chengdu, China, pp.3487-3492, doi: 10.1109/IPEM-ECCEAsia60879.2024.10567836, 2024.
- [3] W. Li, Y. Xu, Q. Li, et al. "Behavior Modeling and Analysis of High-Voltage SiC MOSFET Considering Temperature Effect", in *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*, Singapore, Singapore, pp.1-7, doi: 10.1109/IECON51785.2023.10311758, 2023.
- [4] R. Kraus, A. Castellazzi. "A Physics-Based Compact Model of SiC Power MOSFETs", *IEEE Transactions on Power Electronics*, vol.31, no.8, pp.5863-5870, doi: 10.1109/TPEL.2015.2488106, 2016.
- [5] L. Di Benedetto, G. D. Licciardo, T. Erlbacher, et al. "A Model of Electric Field Distribution in Gate Oxide and JFET-Region of 4H-SiC DMOSFETs", *IEEE Transactions on Electron Devices*, vol.63, no.9, pp.3795-3799, doi: 10.1109/TED.2016.2584218, 2016.
- [6] Z. Duan, T. Fan, X. Wen, et al. "Improved SiC Power MOSFET Model Considering Nonlinear Junction Capacitances", *IEEE*

- Transactions on Power Electronics*, vol.33, no.3, pp.2509-2517, doi: 10.1109/TPEL.2017.2692274, 2018.
- [7] Y. Zhang, Z. Meng. "Modeling Analysis and Simulation of SiC MOSFET Considering Temperature Effects", in *2023 3rd International Conference on Electronic Information Engineering and Computer*, Shenzhen, China, pp.199-202, doi: 10.1109/EIECT60552.2023.10442892, 2023.
- [8] I. Angelov, K. Andersson, D. Schreurs, et al. "Large-signal modelling and comparison of AlGaIn/GaN HEMTs and SiC MESFETs", in *2006 Asia-Pacific Microwave Conference*, Yokohama, Japan, pp.279-282, doi: 10.1109/APMC.2006.4429422, 2006.
- [9] P. Shen, Y. Jiang, X. Zhang, et al. "A Temperature-Dependent SPICE Model of SiC Power Trench MOSFET Switching Behavior Considering Parasitic Parameters", *IEEE Journal of the Electron Devices Society*, vol.13, pp.98-105, doi: 10.1109/JEDS.2024.3498008, 2025.
- [10] M. Riccio, G. D. Falco, P. Mirone, et al. "Accurate SPICE Modeling of Reverse-Conducting IGBTs Including Self-Heating Effects", *IEEE Transactions on Power Electronics*, vol.32, no.4, pp.3088-3098, doi: 10.1109/TPEL.2016.2578363, 2017.
- [11] M. Riccio, V. d'Alessandro, G. Romano, et al. "A Temperature-Dependent SPICE Model of SiC Power MOSFETs for Within and Out-of-SOA Simulations", *IEEE Transactions on Power Electronics*, vol.33, no.9, pp.8020-8029, doi: 10.1109/TPEL.2017.2774764, 2018.

Research Article

High-Temperature Operation of Ga₂O₃ SBDs with 2.33 mV/°C Sensitivity: Enabling Real-Time Junction Temperature for Power Electronics

Mingzhe Li¹, Jun Yuan^{1,2}, Shaodong Xu¹, Ruoshi Peng¹, Liyang Zhu¹, Kuan Wang¹, Qiangmin Wei¹, Rong Zhang², Guoqing Xin² and Zhiqiang Wang²

¹ Dept of Integrated Power Systems and Device Technology, JFS Laboratory, Wuhan 430206, China

² School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Jun Yuan, Email: yuanjun@jflslab.com.cn
Manuscript Received XXX; Accepted XXX; Published Online XXX.

Abstract — Vertical β -Ga₂O₃ Schottky barrier diode (β -Ga₂O₃ SBD) is fabricated by using a single-step slope field plate (SSFP) architecture. Subsequently, the temperature-dependent characteristics of the SBD are systematically investigated. The results show the possibility of Ga₂O₃ Schottky barrier diodes (SBDs) as temperature sensor for temperature monitoring applications. It is found that n and ϕ_b values increases and decreases with the increase of temperature, respectively. Furthermore, there is a linear relationship between n and ϕ_b , implying a typical barrier inhomogeneous behavior. And the current-voltage characteristics of the SBD measured range from 25 to 175°C present a zero-temperature coefficient (ZTC) bias point of approximately 1.19V. At the voltage below the ZTC bias point (sub-threshold region), the forward voltage decreases linearly with the increase of temperature at a fixed current (eg. 10nA~10 μ A), and the fitted lines of sensitivity show good linearity ($R^2 > 0.9945$) in the range from 10nA to 10 μ A, exhibiting a maximum R^2 of 0.9974 at 100nA, causing a sensitivity of approximately 2.33mV/°C. Furthermore, when the sensitivity was area-normalized to account for diode active area dimensions, the resultant value of 1.92×10^{-4} mV/(°C· μ m²) represents a 41% enhancement over reported state-of-the-art counterpart. Those results can be interpreted by the thermionic emission model.

Keywords — β -Ga₂O₃, Schottky barrier diode (SBD), temperature-dependent characteristics, temperature monitor.

I. Introduction

With the booming development of new energy vehicles, 5/6G communications, automotive electronics, and other fields, people have put forward higher and higher requirements for power devices. The performance of traditional silicon power devices has been close to the theoretical limit determined by its material properties, people have set their sights on wide bandgap (WBG) semiconductor materials including gallium nitride (GaN) [1]-[3] and silicon carbide (SiC) [4]-[6], even the ultra-wide bandgap (UWBG) semiconductor material — gallium oxide (Ga₂O₃). Owing to its excellent material properties, such as ultra-wide bandgap ~4.8eV, high critical breakdown field 8MV/cm and high saturation velocity ~2×10⁷ cm s⁻¹ [7]-[10], Ga₂O₃ has emerged as a prospective material for the next generation power electronic devices. And that, UWBG Ga₂O₃ power devices have been shown many great advantages including adjustable n-type doping concentration, available melt-grown substrates and remarkable bandgap property [10]-[13].

However, because of lower thermal conductivity of Ga₂O₃ material and the heat dissipation during on/off switching process in various higher power applications, the self-heating effect (SHE) of Ga₂O₃-based devices has been highlighted, and will inevitably result in the increasing junction temperature. The higher junction

temperature also could be a critical issue, which degenerates the temperature-dependent electrical parameters (eg. R_{on} and I_r), hinders the high-performance power module, even is detrimental to the reliability of power devices [14]-[16]. According to a large number of previous circuit research, once accurate junction temperature has been detected and recorded, we can design over-temperature protection circuit to control the temperature of device in a safe range, and protect device work properly. So, the in situ real-time measurement of the accurate junction temperature is necessary for devices.

Many methods based on the in-situ or external temperature sensors have been proposed to monitor the junction temperature. Among these technologies, semiconductor devices especially diodes, such as Schottky barrier diode (SBD) [17]-[20] and p-n diode (PND) [21]-[23], show the simplest structure and easily integrate with the power devices in ICs by the co-package or monolithic, which can provide a real-time monitoring to detect and predict the failures. Generally, the forward voltage (V_F) at a fixed current shows a strong dependence on temperature. Good linearity is convenient to complete the pre-calibration procedure by making only use of two temperature points. At present, vertical GaN and SiC diodes have been fabricated as temperature sensors show a good linearity and sensitivity [18][19]. However, the temperature sensors with Ga₂O₃-compatible is rarely reported.

错误! 未定义样式。

In this work, we propose a real-time monitoring method for junction temperature of Ga_2O_3 devices, and a vertical Ga_2O_3 SBD temperature sensor was realized and evaluated in a range from 25 to 175°C. The sensitivity was deduced by the linear voltage shift versus temperature at a constant current in the current-voltage (I - V) characteristics. Following the thermionic emission (TE) theory, the feasibility of temperature sensing was analyzed in the sub-threshold region.

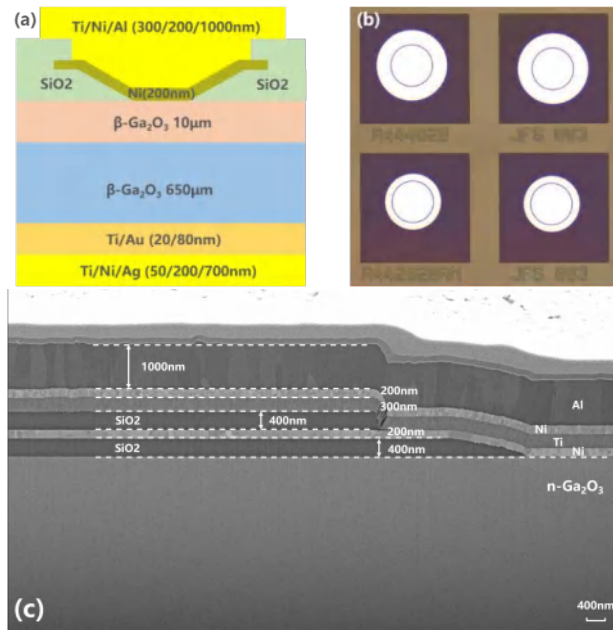


Figure1 (a) The cross-sectional schematic and (b) top view of the vertical $\beta\text{-Ga}_2\text{O}_3$ SSFP-SBD with a diameter of 110 μm . (c) SEM of the SBD active area.

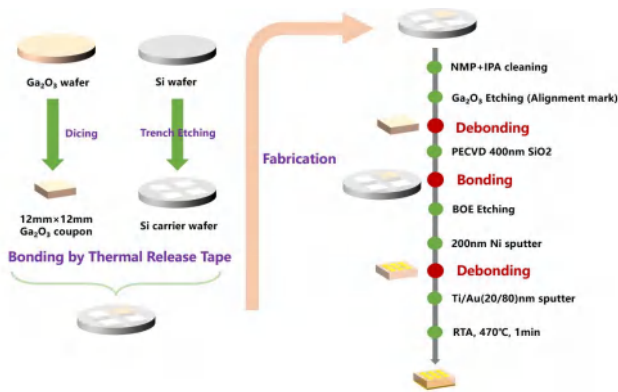


Figure2 Fabrication process of the vertical $\beta\text{-Ga}_2\text{O}_3$ SSFP-SBD by adopting coupon to wafer technique.

II. Device Structure and Fabrication

The sample utilized in this study consisted of a 650 μm thickness commercially available (001) $\beta\text{-Ga}_2\text{O}_3$ substrate and a 10 μm thickness Si-doped N- Ga_2O_3 epitaxial layer by halide vapor phase epitaxy (HVPE) purchased from Novel Crystal Technology, Inc. Fig.1(a) shows the schematic of $\beta\text{-Ga}_2\text{O}_3$ single-step slope field plate (SSFP) SBD, and the top view and SEM of the SBD are shown in Fig.1(b)(c). Device fabrication started with dicing 2inch wafer into 12mm \times 12mm pieces, which

adopted coupon to wafer technique to carry out, in other words, $\beta\text{-Ga}_2\text{O}_3$ pieces bonding onto trenches of 6-inch Si carrier wafer by Thermal Release Tape (shown in Fig.2), and it is a temporary bonding method. Coupon to wafer technique provides a solution which can help irregular Ga_2O_3 wafer tape out based on the 4/6/8inch pilot production line platform.

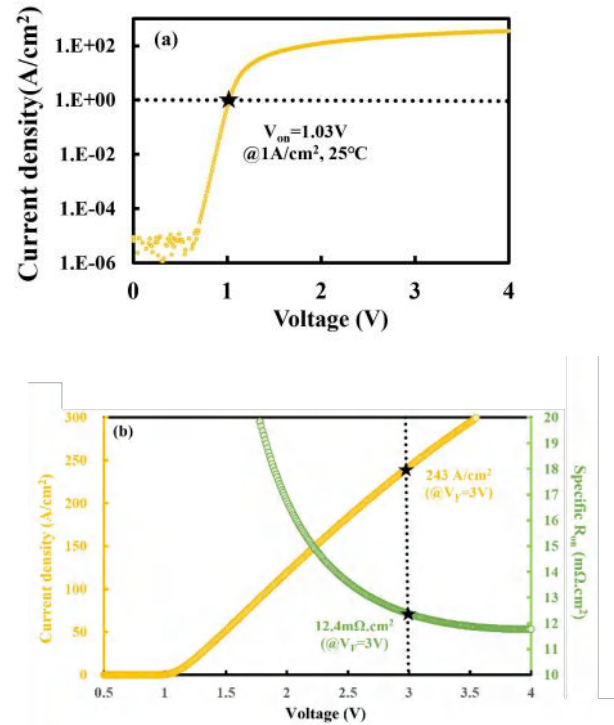


Figure3 (a)The semi-log plot of forward I - V . (b) The forward I - V characteristics and corresponding $R_{on,sp}$ of the SSFP-SBD.

The front of sample was cleaned several times in N-Methyl-Pyrrolidone (NMP) and Isopropanol (IPA), separately. Alignment mark was defined by photo-lithography followed by inductively coupled plasma (ICP) etching with etch depth of 550 nm. The Ga_2O_3 etch was carried out by BCl_3/Ar gas combinations with RF power of 400 W and gas flow rate of 20/10 sccm at a etch rate of 100nm/min. A 400 nm SiO_2 layer was deposited at temperature of 300°C by plasma enhanced chemical vapor deposition (PECVD). The single slope dielectric open area of SiO_2 layer was described by photo-lithography and formed by being immersing in BOE solution ($\text{HF}:\text{NH}_4\text{F} = 1:7$) for 3 minutes. Circular Ni (200nm) Schottky metal was deposited above photo-resist by electronic beam evaporation (EBE) and lifted off to form anode electrodes. Then, the Ti/Au (20/80nm) metal stack was also deposited by EBE on the backside of the substrate as ohmic cathode electrodes followed by a rapid thermal annealing (RTA) in N_2 ambient for 1 minutes at 470 °C. Another 400nm SiO_2 layer was also deposited at temperature of 300 °C by PECVD, and the dielectric open area was described by photo-lithography and dry etching. Subsequently, the anode electrodes were thickened by Ti/Ni/Al (300/200/1000nm) and cathode electrodes by Ti/Ni/Ag (50/200/700nm) to improve the contact between probe and electrodes. Finally, the Ga_2O_3 single-step slope field

plate Schottky barrier diode (SSFP-SBD) has been achieved. The slope metal field plate was beneficial to suppress the electric-field crowding effect and enhance the breakdown voltage. To evaluate the temperature sensing ability of the SBD, current-voltage (I - V) characteristics were recorded at 25, 50, 75, 100, 125, 150 and 175°C, respectively.

III. Results and Discussion

Fig.3(a) shows the forward I & $R_{on,sp}$ - V characteristics of vertical β -Ga₂O₃ SSFP-SBD with a diameter of 110 μ m at room temperature. The turn-on voltage (V_{on}) is determined to be 1.03V defined at the forward current density (I_F) of 1A/cm². At forward bias (V_F) of 3V, a maximum IF of 243 A/cm², and the $R_{on,sp}$ value of the SBD is 12.4 m Ω .cm², correspondingly.

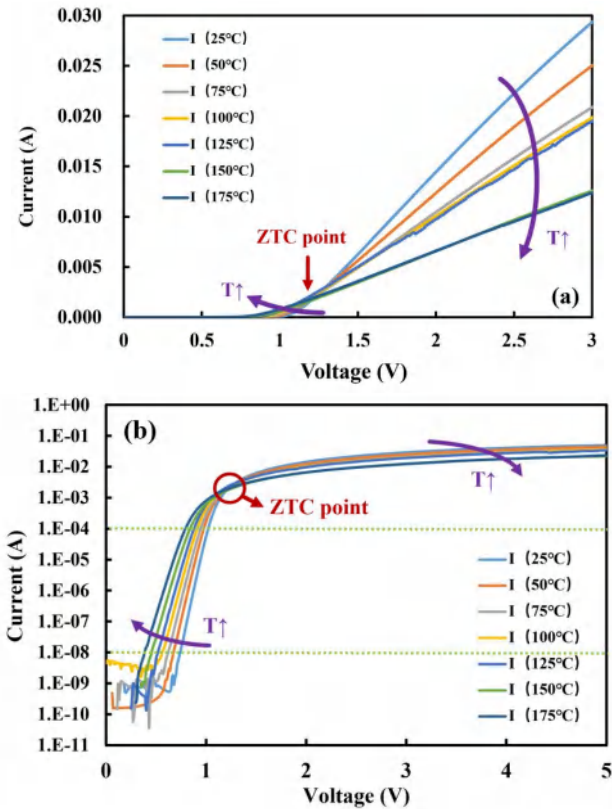


Figure4 The temperature dependent I - V characteristics of SBD in (a) linear and (b) semi-log scale.

The temperature dependent I - V characteristics of SBD were recorded at various temperature ranging from 25 to 175°C with a step of 25°C (as shown in Fig. 4). The turn-on voltage of the traditional lateral SBD translationally decreases to a smaller value with the increase of temperature. While the I - V characteristics of vertical SBD exist a zero-temperature coefficient (ZTC) bias point (about 1.19V in our measurement) with the increasing temperature. $V_F < ZTC$ bias point (sub-threshold region), the current increases gradually with the increase of temperature because of electronics gain enough energy and the decrease of turn-on voltage. While $V_F > ZTC$ bias point (fully turn-on region), the

forward current decreases with the increasing temperature due to the carrier mobility degradation and the increasing series resistance. The ZTC bias point is usually used to design voltage reference of over temperature protection structure in power integrated circuit (IC) because the thermal compensation is realized at only a constant voltage. Then, the voltage reference can be re-configured easily to work at different output voltage without extra circuit design. In addition, the sub-threshold region of I - V characteristics in the semi-logarithmic plot are linear and shift negatively with the increase of temperature (Fig.4 (b)), suggesting that the carrier transportation follows the TE model.

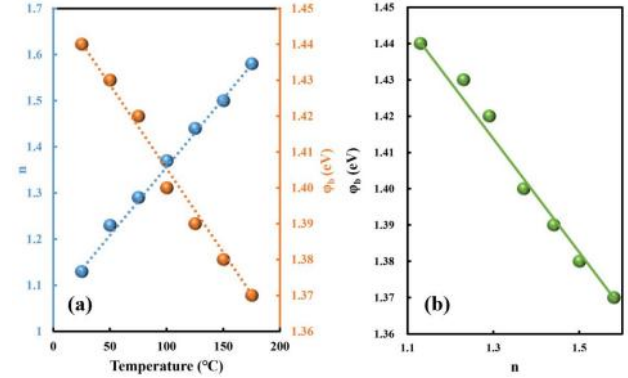


Figure5 (a) The temperature dependent n and ϕ_b from 25 to 175°C. (b) the relationship between n and ϕ_b .

To reveal the temperature dependent variation in detail, the I - V curves are fitted by using the TE model. When $V_F > 3kT/q$, the forward current (I_F) can be represented as [24], [25]

$$I_F = A_e A^* T^2 \exp\left(-\frac{q\phi_b}{kT}\right) \exp\left(\frac{q(V_F - I_F R_S)}{nkT}\right). \quad (1)$$

where ϕ_b , A_e , T , q , n , k , V , A^* , and R_S are the Schottky barrier height (SBH), Schottky contact area, Kelvin temperature, electronic charge, ideality factor, Boltzmann's constant, bias, Richardson constant, and series resistance, respectively. $I_F R_S$ is the voltage drop across series resistance. The forward voltage is solved as

$$V_F = n\phi_b + \frac{nkT}{q} \left[\ln\left(\frac{I_F}{A_e A^*}\right) - 2\ln(T) \right] + I_F R_S \quad (2)$$

The obtained n and ϕ_b values, as shown in the Fig.5a, increases and decreases with the increase of temperature, respectively. It is note that there is a linear relationship between n and ϕ_b (Fig.5b), implying a typical barrier inhomogeneous behavior. The barrier inhomogeneous assumes that some relatively lower barriers with different cross-section areas are embedded into a uniform barrier [26]. At low temperature, the current of diodes are dominated by the low barrier regions, while at high temperature, carriers gain adequate power to cross through the barrier regions, and can think of that the barrier height goes down relatively, which could be the reason why the measured effective ϕ_b at low temperature is bigger than that at high temperature.

In addition, the corresponding R_S increases clearly with the increase of temperature, showing a positive slope

(dR_S/dT) of approximately $0.74\Omega/^\circ\text{C}$. As reported in the previous references [20],[26], the value of vertical SBD is much higher than that of lateral SBD due to the different current conduction mode in vertical and lateral diodes. For a diode temperature sensor at the fully-turn-on region, the mobility of the drift layer would degrade further with the increasing temperature due to the scattering, leading to an evident incremental of resistance. It was reported that the series resistance would decrease the coefficient of determination and the root mean square error of temperature, leading to the degradation of temperature error and linearity [19][23]. Hence, the fully-turn-on region of the vertical diode is not adaptable to temperature sensing application.

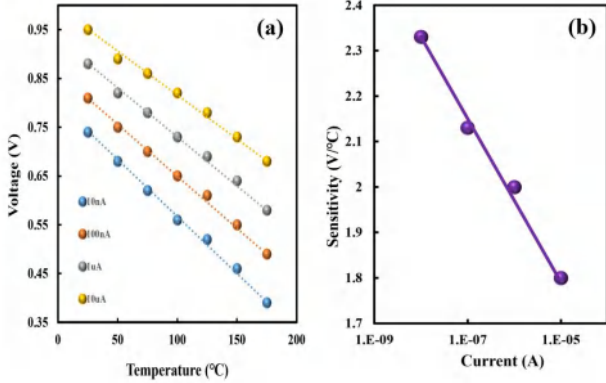


Figure6 (a) The forward voltage versus temperature at sub-threshold state with IF levels of 10nA, 100nA, 1µA and 10µA. (b) the sensitivity versus current.

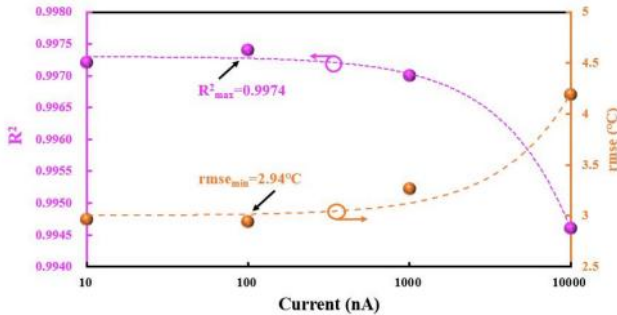


Figure7 R^2 and $rmse(T)$ versus I_F at the given current.

So, while at the sub-threshold region, the proportion of $I_F R_S$ of (2) shows a so small value resulting from the relatively small current (μA level) and can be negligible. Then, the product of n and ϕ_b could be seemed a constant as a result of that these two values have opposite trend versus temperature [26] (Fig.5b). And that, the effect of temperature on the nonlinear part $\ln(T)$ of (2) is vanishingly small until very high temperature. Therefore, the V_F at a fixed current drop linearly versus the temperature with a sensitivity relative to the current density (I_F/A_e) as

$$\frac{dV_F}{dT} \propto \ln\left(\frac{I_F}{A_e}\right) \quad (3)$$

The forward voltage versus temperature (V_F-T) at different current are plotted in Fig.6a based on the best-calculated linear fitting lines model ($f_L(T)$) [27]. The temperature sensitivities (S) at four current levels of 1×10^{-8} A, 1×10^{-7} A, 1×10^{-6} A and 1×10^{-5} A are calculated

from the slope of the fitting lines, to be approximately 2.33, 2.13, 2.00 and 1.80 $\text{mV}/^\circ\text{C}$, respectively. Usually, analyst use the coefficient of determination (R^2) to quantify the sensor linearity, which can be calculated by the following equation:

$$R^2 = \frac{\sum_{i=1}^m [f_L(T_i) - \bar{V}_F(T_i)]^2}{\sum_{i=1}^m [V_F(T_i) - \bar{V}_F(T_i)]^2} \quad (4)$$

where $V_F(T_i)$ is the measured forward voltage of each temperature, $\bar{V}_F(T_i)$ is the average value of measured forward voltage, $f_L(T_i)$ is the calculated voltage value on the fitted line at a given temperature (T_i), m is the number of the temperature points and is equal to 7 in this article; Hence, $\sum_{i=1}^m [f_L(T_i) - \bar{V}_F(T_i)]^2$ is the regression sum of squares and $\sum_{i=1}^m [V_F(T_i) - \bar{V}_F(T_i)]^2$ is the total sum of squares.

The obtained R^2-I_F plot for the sub-threshold region is shown in Fig.7. It proves that all the fitted lines show good linearity ($R^2 > 0.9945$) in the range from 10nA to $10\mu\text{A}$, exhibiting a maximum R^2 of 0.9974 at 100nA. Although the linearity showed mildly degradation in the higher current, the R^2 values are still larger than 0.9945.

$$rmse(T) = \sqrt{\frac{\sum_{i=1}^m [V_F(T_i) - f_L(T_i)]^2}{m}} \quad (5)$$

where m is the number of measurement data and equal to 7 in this article. The plot of $rmse(T)-I_F$ shows opposite tendency with R^2 , is also shown in Fig.7. The $rmse(T)$ is below 4.2°C in the range from 10nA to $100\mu\text{A}$, and the minimum value is 2.94°C at 100nA. Although the $rmse(T)$ showed an increase in the higher current, the $rmse(T)$ values are still smaller than 4.2°C .

The sensitivity shows a decreasing trend with the increasing current level in the sub-threshold region. After in consideration of the diode area, the sensitivity is proportional to the logarithms of current (Fig.6b). Based on formula (3), a large current at a certain diameter will decrease the absolute value of $\ln(I_F/A_e)$ and lead to a smaller sensitivity. Hence, a large device area (diameter) also can be used to obtain a higher sensitivity at a specific current. For a better comparison, the key parameters of temperature sensors with lateral and vertical structures are listed in TABLE I.

Table 1 The comparison of temperature sensors fabricated with lateral and vertical structures

Sensor type	Sensitivity (mV/ $^\circ\text{C}$)	Diode area(μm^2)	Area-normalized Sensitivity (mV/ $^\circ\text{C} \cdot \mu\text{m}^2$)	Ref
Vertical Diamond SBD	1.60	200×200	4.0×10^{-5}	[17]
Vertical SiC SBD	1.24	200×200	3.1×10^{-5}	[18]
Lateral GaN SBD	1.13	200×200	2.8×10^{-5}	[19]
Lateral Series GaN SBD	2.25	150×150	1.0×10^{-4}	[28]
Vertical PTAT SiC SBD	5.11	$150 \times 150 \times 2$	1.13×10^{-4}	[29]
Vertical $\beta\text{-Ga}_2\text{O}_3$ SBD	2.33	110×110	1.92×10^{-4}	This work

As systematically investigated and discussed earlier, the sensitivity parameters exhibit strong correlation with diode area. To enable direct comparison, we have conducted area-normalized analysis in the following Table I. Area-normalized sensitivity of our temperature sensor demonstrates a 41% enhancement compared to the state-of-the-art counterpart [29], confirming that the sensitivity of the sensors in this paper has a significant advantage.

IV. Conclusion

In summary, a kind of vertical β -Ga₂O₃ SBD was fabricated and the temperature-dependent characteristics of the SBD were systematically investigated. The forward I-V characteristics at different temperatures could be explained by TE model with good agreements. It was noted that there is a linear relationship between n and ϕ_b , indicating a typical barrier inhomogeneous behavior. The temperature-dependent current-voltage characteristics of the SBD measured range from 25 to 175°C present a zero-temperature coefficient (ZTC) bias point of approximately 1.19V. At the voltage below the ZTC bias point (sub-threshold region), the forward voltage decreases linearly with the increase of temperature at a fixed current, and the fitted lines of sensitivity show good linearity ($R^2 > 0.9945$) in the range from 10nA to 10 μ A, exhibiting a maximum R^2 of 0.9974 at 100nA, causing a sensitivity of approximately 2.33mV/°C. Furthermore, when the sensitivity was area-normalized to account for diode active area dimensions, the resultant value of 1.92×10^{-4} mV/(°C $\cdot\mu$ m²) represents a 41% enhancement over reported state-of-the-art counterpart.

Acknowledgement

This work was supported by the Major Program of Hubei Province under Grant No. 2023BAA009.

References

- [1] R. Vetry, N. Q. Zhang, S. Keller, and U. K. Mishra, "The impact of surface states on the DC and RF characteristics of AlGaIn/GaN HFETs," *IEEE Trans. Electron Devices*, vol. 48, no. 3, pp. 560–566, Mar. 2001, doi: 10.1109/16.906451.
- [2] B. J. Baliga, "Gallium nitride devices for power electronic applications," *Semicond. Sci. Technol.*, vol. 28, no. 7, Jun. 2013, Art. no. 074001, doi: 10.1088/0268-1242/28/7/074011.
- [3] S. Fujita, "Wide-bandgap semiconductor materials: For their full bloom," *Jpn. J. Appl. Phys.*, vol. 54, no. 3, Feb. 2015, Art. no. 030101, doi: 10.7567/JJAP.54.030101.
- [4] T. Kimoto, "Material Science and Device Physics in SiC Technology for High-Voltage Power Devices," *Jpn. J. Appl. Phys.*, vol. 54, no. 4, Jan. 2015, Art. no. 040103, doi: 10.7567/JJAP.54.040103.
- [5] T. Nakamura, "Novel developments towards increased SiC power device and module efficiency," *IEEE Energytech, Cleveland, OH, USA*, pp. 1–6 May 2012, doi: 10.1109/EnergyTech.2012.6304633.
- [6] X. She, "Review of Silicon Carbide Power Devices and Their Applications," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 10, pp. 8193–8205, Oct. 2017, doi: 10.1109/TIE.2017.2652401.
- [7] M. Higashiwaki, "Recent progress in Ga₂O₃ power devices," *Semicond. Sci. Technol.*, vol. 31, no. 3, Jan. 2016, Art. no. 034001, doi: 10.1088/0268-1242/31/3/034001.
- [8] N. Ma, "Intrinsic electron mobility limits in β -Ga₂O₃," *Appl. Phys. Lett.*, vol. 109, no. 21, Nov. 2016, Art. no. 212101, doi: 10.1063/1.4968550.
- [9] M. Higashiwaki, "Gallium oxide (Ga₂O₃) metal-semiconductor field-effect transistors on single-crystal β -Ga₂O₃ (010) substrates," *Appl. Phys. Lett.*, vol. 100, no. 1, 2012, Art. no. 013504, doi: 10.1063/1.3674287.
- [10] M. Higashiwaki, "Guest editorial: The dawn of gallium oxide microelectronics," *Appl. Phys. Lett.*, vol. 112, no. 6, Feb. 2018, Art. no. 060401, doi: 10.1063/1.5017845.
- [11] H. Zhou, "A review of the most recent progresses of state-of-art gallium oxide power devices," *J. Semicond.*, vol. 40, no. 1, Jan. 2019, Art. no. 011803, doi: 10.1088/1674-4926/40/1/011803.
- [12] Z. Galazka, "On the bulk β -Ga₂O₃ single crystals grown by the Czochralski method," *J. Crystal Growth*, vol. 404, pp. 184–191, Oct. 2014, doi: 10.1016/j.jcrysgro.2014.07.021.
- [13] A. Kuramata, "High-quality β -Ga₂O₃ single crystals grown by edge-defined film-fed growth," *Jpn. J. Appl. Phys.*, vol. 55, no. 12, 2016, Art. no. 1202A2, doi: 10.7567/JJAP.55.1202A2.
- [14] Z. Jiang, "Experimental Investigation on Threshold Voltage Instability for β -Ga₂O₃ MOSFET Under Electrical and Thermal Stress," *IEEE Transactions on Electron Devices*, vol. 69, no. 9, pp. 5048–5054, Sept. 2022, doi: 10.1109/TED.2022.3188584.
- [15] Z. Cheng, "Thermal transport across ion-cut monocrystalline β -Ga₂O₃ thin films and bonded β -Ga₂O₃-SiC interfaces," *ACS Appl. Mater. Interface*, vol. 12, no. 40, pp. 44943–44951, Sep. 2020, doi: 10.1021/acami.0c11672.
- [16] H. Zhang, "Progress of ultra-wide bandgap Ga₂O₃ semiconductor materials in power MOSFETs," *IEEE Trans. Power Electron.*, vol. 35, no. 5, pp. 5157–5179, May 2020, doi: 10.1109/TPEL.2019.2946367.
- [17] G. Perez, "Integrated temperature sensor with diamond Schottky diodes using a thermosensitive parameter," *Diamond Rel. Mater.*, vol. 78, pp. 83–87, Sep. 2017, doi: 10.1016/j.diamond.2017.08.008.
- [18] G. Pristavu, "Characterization of non-uniform Ni/4H-SiC Schottky diodes for improved responsivity in high-temperature sensing," *Mater. Sci. Semicond. Process.*, vol. 94, pp. 64–69, May 2019, doi: 10.1016/j.mssp.2019.01.018.
- [19] X. Li, "Correlation between anode area and sensitivity for the TiN/GaN Schottky barrier diode temperature sensor," *IEEE Trans. Electron Devices*, vol. 67, no. 3, pp. 1171–1175, Mar. 2020, doi: 10.1109/TED.2020.2968358.
- [20] L. Li, "Vertical GaN-Based Temperature Sensor by Using TiN Anode Schottky Barrier Diode," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1273–1278, 15 Jan. 15, 2021, doi: 10.1109/JSEN.2020.3018330.
- [21] S. Rao, "Highly linear temperature sensor based on 4H-silicon carbide p-i-n diodes," *IEEE Electron Device Lett.*, vol. 36, no. 11, pp. 1205–1208, Nov. 2015, doi: 10.1109/LED.2015.2481721.
- [22] N. Zhang, "Temperature sensor based on 4H-silicon carbide pn diode operational from 20°C to 600°C," *Appl. Phys. Lett.*, vol. 104, no. 7, Feb. 2014, Art. no. 073504, doi: 10.1063/1.4865372.
- [23] X. Li, "P-NiO/n-GaN heterostructure diode for temperature sensor application," *IEEE Sensors J.* vol. 20, no. 1, pp. 62–66, Jan. 2020, doi: 10.1109/JSEN.2019.2939045.
- [24] L. Boukhris, "Electrical resistivity, structure and composition of d.c. sputtered WN_x films," *Thin Solid Films*, vol. 310, no. 1, pp. 222–227, Nov. 1997, doi: 10.1016/S0040-6090(97)00348-9.
- [25] S. M. Sze, "Metal-semiconductor contacts," in *Physics of Semiconductor Devices*, 2nd ed. Hoboken, NJ, USA: Wiley, 2007, pp. 79–134.
- [26] L. Li, "Temperature sensor using thermally stable TiN anode GaN Schottky barrier diode for high power device application," *Superlattices Microstruct.*, vol. 123, pp. 274–279, Nov. 2018, doi: 10.1016/j.spmi.2018.09.007.
- [27] N. J. D. Nagelkerke, "A note on a general definition of the coefficient of determination" [*J. Biometrika*], 1991, 78(3):691–692, doi: 10.2307/23737038.
- [28] X. Li, "Enhanced sensitivity of GaN-based temperature sensor by using the series Schottky barrier diode structure," *IEEE Electron Device Lett.*, vol. 41, no. 4, pp. 601–604, Apr. 2020, doi: 10.1109/LED.2020.2971263.
- [29] S. Rao, "High performance temperature sensor based on 4H-SiC Schottky diodes," *IEEE Electron Device Lett.*, vol. 36, no. 7, pp. 720–722, Jul. 2015, doi: 10.1109/LED.2015.2436213.



Mingzhe Li received the M.S. degree in integrated circuit engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2023. His current research interests include design, fabrication and characterization of compound semiconductor power devices at JFS Laboratory. (Email: limingzhe@jfslab.com.cn)



Jun Yuan received the B.S. degree in School of Physics and Technology from Wuhan University, Wuhan, China, in 2005, the M.S. degree in

错误! 未定义样式。

condensed matter physics from the Beijing Normal University, Beijing, China, in 2010. He is currently studying for a Ph.D. in Electrical and Electronic Engineering at Huazhong University of Science and Technology. At the same time, he focus research on wide-band gap semiconductor materials and chip technology at JFS Laboratory.
(Email: yuanjun@jfslab.com.cn)

Research Article

A Comprehensive Study on the Fabrication of a Full-P Encapsulated Trench MOSFET with N-P-N Sandwich Epitaxial Wafers for 4H-SiC Devices

Zhijie Cheng¹, Jun Yuan^{1,2}, Kuan Wang¹, Wei Chen¹, Yangyang Wu¹, Rong Zhang², Guoqing Xin², and Zhiqiang Wang²

¹ Dept of Integrated Power Systems and Device Technology, JFS Laboratory, Wuhan 430206, China

² School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Jun Yuan; Email: yuanjun@jfslab.com.cn.
Manuscript Received xxx; Accepted xxx; Published Online xxx.

Abstract — In this study, we have designed and fabricated a novel 1.2 kV full-P encapsulated trench MOSFET with N-P-N sandwich epitaxial wafers for 4H-SiC devices. It only needs to be controlled by design and no additional process steps are required. The entire preparation process and results have been fully presented. By altering the width of different N⁺ conductive path and the concentration of the P⁺ buried layer, three different devices were fabricated. The gate leakage current of all devices is guaranteed to be around 20 nA. Due to better gate protection, the breakdown voltages of both device B and device C are above 1500 V. All the devices have normal conduction characteristic. The device A sacrifices some of its breakdown characteristic, thus having a lower on-resistance. The device B has a better compromise between the breakdown characteristic and the conduction characteristic. The threshold voltages of all the devices remain at around 4.8 V. The devices B and C have demonstrated that the proposed structure is feasible. In the future, the trade-off relationship between the breakdown characteristic and the conduction characteristic of the devices will be further improved by optimizing the design parameters of the N⁺ conductive path and the P⁺ buried layer. This stable and reliable design creates unlimited possibilities for the future development of 4H-SiC power devices.

Keywords — trench MOSFET, sandwich epitaxial wafers, gate leakage current, breakdown characteristic, conduction characteristic.

I. Introduction

Silicon carbide (SiC) has the advantages of high electron mobility, high critical electric field strength and high thermal conductivity, so it is widely used in new energy vehicles, photovoltaic inverter, rail transit [1]-[3]. Due to the fact that the stacking energy between different polytypes of SiC is very close, it is easy to produce other polytypes during the growth process [4]. Therefore, the SiC is a crystalline compound with more than 170 polytypes [5], [6]. As the 4H-SiC technology becomes increasingly mature, its role becomes increasingly important. It is reasonable to believe that it is a promising semiconductor for the next generation of high power, high frequency and high temperature applications [7], [8].

4H-SiC trench MOSFET has lower on-resistance compared to the planar MOSFET. This is because the trench MOSFET has a higher cell density and lower JFET region resistance [9]-[11]. Therefore, an increasing number of research institutions and enterprises are beginning to delve deeply into the design and process of the trench MOSFET [12]-[15]. However, some of the device structures are quite complex in design, and the production costs are high. Moreover, there might be similar problems to those of existing commercial devices. That is to say, currently, designing a 4H-SiC trench MOSFET with independent device design and balanced

performance is a very challenging work.

In this paper, we designed and fabricated a novel 1.2 kV full-P encapsulated trench MOSFET (FPE-TMOS) with N-P-N sandwich epitaxial wafers for 4H-SiC devices. In addition, this design does not add additional process steps, greatly reducing the fabrication cost. Furthermore, the detailed process flow for the preparation of the proposed structure is described, and the fabricated cell structure and termination structure are also presented. Three different devices were fabricated through the design of experiments. Further studies were conducted on the gate leakage current, breakdown characteristic, conduction characteristic and threshold voltage of these three devices. Most importantly, an effective method for further improving the trade-off relationship between the breakdown characteristic and conduction characteristic of the devices was proposed.

II. Device Structure and Mechanism

As shown in Figure 1(a), the section A of the proposed structure is presented, which is the main conductive region, occupying over 90% of the chip area. And Figure 1(b) shows the section B of the proposed structure, which mainly ensures that the P-shield region is in a grounded state. This design occupies less than 10% of the chip area. The main device parameters used in simulation are listed in Table 1. Furthermore, the SRH

and auger recombination models, CVT model, high field saturation, aniso model, bandgap model, fermi model, incomplete ionization model are exploited in the simulation [16]-[18].

All SiC wafers are composed of a N-P-N epitaxial structure consisting of the N-drift region, P+ buried layer and N+CSL (current extension layer). Compared with the traditional process, this design does not require additional process steps. At the same time, as the epitaxial cost continues to decrease, this design scheme is feasible. The thickness and concentration of P+ buried layer are 0.3 μm and $1 \times 10^{17} \text{ cm}^{-3}$, respectively. This is a reduction from the previous P+ buried layer of 0.6 μm and $3 \times 10^{17} \text{ cm}^{-3}$ [13], in order to completely deplete the P+ buried layer under reverse bias, so that the FLRs (field limiting rings) play the main role of withstand voltage. In the design of this structure, the parameters of the N+ conductive path and the P+ buried layer are of great significance. They need to be optimized through a combined process of simulation and experimentation.

Table 1 Main device parameters used in simulation

Structure parameters	FPE-TMOS
N+buffer region concentration $N_b \text{ (cm}^{-3}\text{)}$	1×10^{18}
N+buffer region thickness $T_b \text{ (}\mu\text{m)}$	0.5
N-drift region concentration $N_d \text{ (cm}^{-3}\text{)}$	1×10^{16}
N-drift region thickness $T_d \text{ (}\mu\text{m)}$	11
P+ buried layer concentration $N_p \text{ (cm}^{-3}\text{)}$	$1 \& 2 \times 10^{17}$
P+ buried layer thickness $T_p \text{ (}\mu\text{m)}$	0.3
N+CSL concentration $N_c \text{ (cm}^{-3}\text{)}$	4×10^{16}
N+CSL thickness $T_c \text{ (}\mu\text{m)}$	1.5
Gate trench width $W_{gt} \text{ (}\mu\text{m)}$	1.5
Gate oxide thickness $T_{ox} \text{ (nm)}$	60
Number of FLRs	15
Initial ring spacing of FLRs $\text{(}\mu\text{m)}$	1.2
Ring width of FLRs $\text{(}\mu\text{m)}$	1.5
Trench width of termination $\text{(}\mu\text{m)}$	3

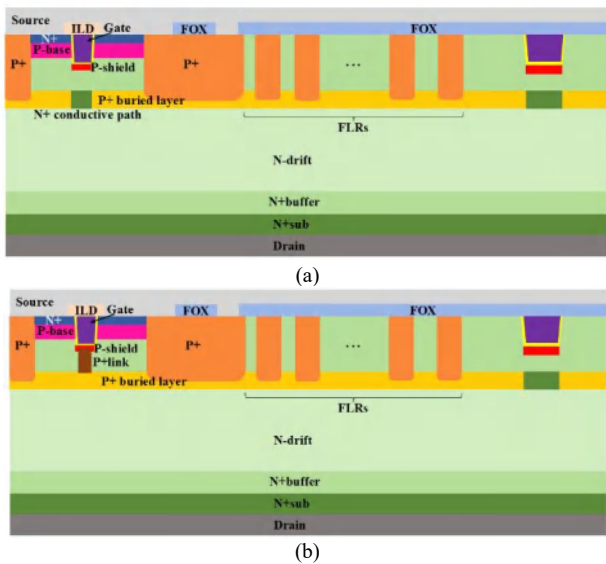


Figure 1 A full-P encapsulated 4H-SiC trench MOSFET with N-P-N sandwich epitaxial wafers. (a) Section A. (b) Section B.

III. Device Fabrication

The specific process flow of the proposed structure is shown in Figure 2. On the N-P-N sandwich epitaxial layers, the P-base region and the N+ region are first formed by ion implantation respectively. Then, the P+ region and FLRs are formed simultaneously through Al ion implantation. Further, the gate trench and the trench of the termination are formed through dry etching. Ion implantation is carried out at the bottom of the trench to form the P-shield region. After this, the next step is to start preparing the fabricate processes for section A and section B respectively. In section B, the P+ link region needs to be formed through ion implantation to ensure that the P-shield is grounded. In section A, the N+ conductive path needs to be formed through ion implantation to ensure the conduction characteristic of the device. Next growing the gate oxide layer, and the polysilicon is deposited. Finally, the ohmic contacts are formed and the metal is deposited.

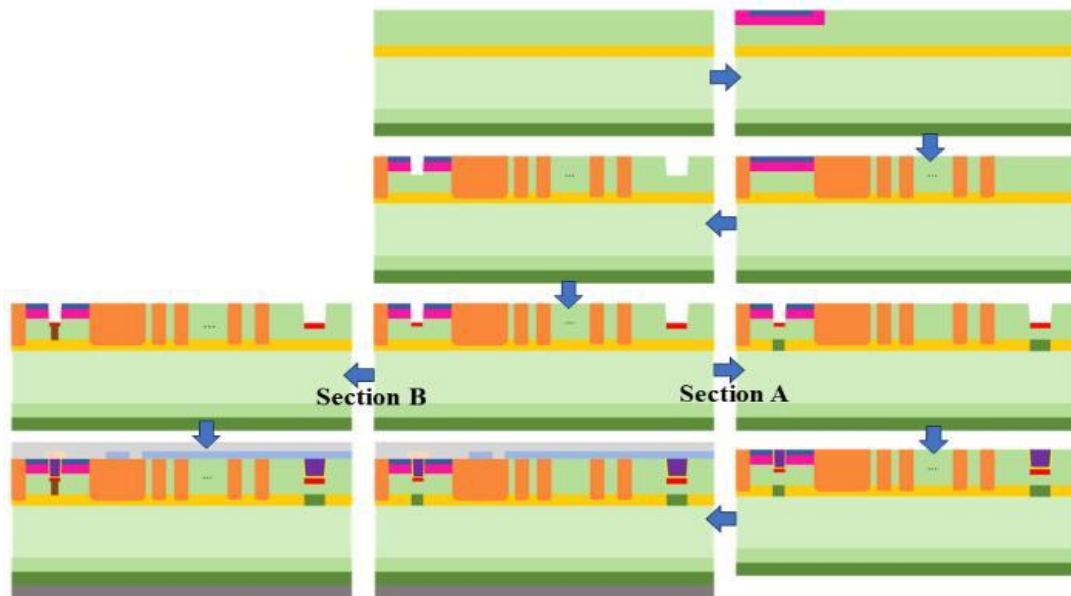


Figure 2 The complete process flow of the proposed 4H-SiC trench MOSFET cell and termination structure.

As shown in Figure 3, actual SEM micrographs of the process obtained according to the process flow are presented, including section A and section B. Furthermore, Figure 4 also presents the SEM micrograph of the termination structure. The presented results are based on the conditions where the thickness and concentration of the P+ buried layer are $0.3\ \mu\text{m}$ and $1.0 \times 10^{17}\ \text{cm}^{-3}$ respectively, and the ion implantation mask for the N+ conductive path also follows the baseline conditions.

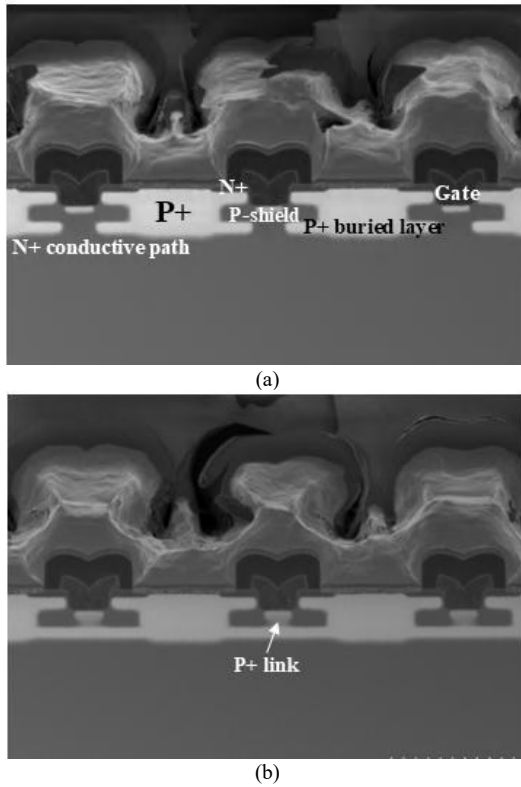


Figure 3 SEM micrographs of the fabricated FPE-TMOS cell structure. (a) Section A. (b) Section B.

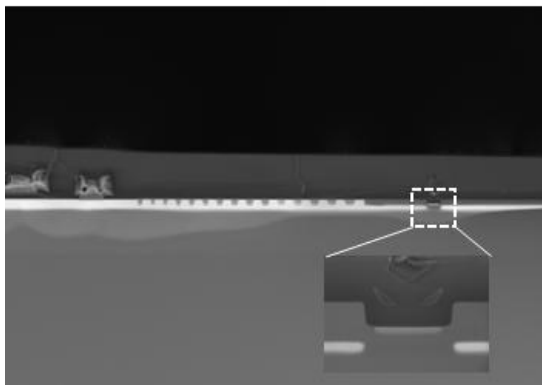


Figure 4 SEM micrograph of the fabricated FPE-TMOS termination structure.

IV. Results and Discussion

The fabricated cell structure and termination structure are presented. The most important point is that this is the first time that the influence of the N+

conductive path and the P+ buried layer parameters on the electrical performance of the cell structure and termination structure has been studied through actual process experiments. The main research focused on the effects of different widths of N+ conductive path and different concentrations of P+ buried layer on the electrical performances of the devices. Specifically, the design of three devices in fabrication are presented in Table 2. The device A represents the initial design parameters. The difference between device C and the other devices lies in the different concentrations of the P+ buried layer. While device B increased the thickness of the sidewall mask by 100 nm during the ion implantation process of the N+ conductive path, thereby reducing the width of the N+ conductive path. As shown in Figure 5, the difference between the baseline mask for the N+ conductive path and the 100nm sidewall mask is presented. Furthermore, the different electrical performances of the three devices were studied in detail.

Table 2 The design of different devices in fabrication

device	P+ buried layer	N+ conductive path
Device A	$0.3\ \mu\text{m}$ and $1.0 \times 10^{17}\ \text{cm}^{-3}$	Baseline mask
Device B	$0.3\ \mu\text{m}$ and $1.0 \times 10^{17}\ \text{cm}^{-3}$	100 nm sidewall
Device C	$0.3\ \mu\text{m}$ and $2.0 \times 10^{17}\ \text{cm}^{-3}$	Baseline mask

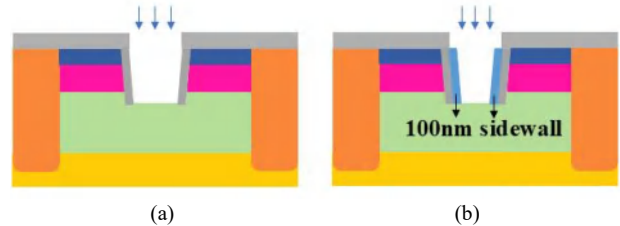


Figure 5 Mask for ion implantation of N+ conductive path. (a) Baseline mask. (b) 100 nm sidewall.

1. Gate Leakage Current

Figure 6 shows the gate leakage current of the three devices when $V_{gs}=22\ \text{V}$. It can be seen that the gate leakage current is around 20 nA, which indicates that the devices have excellent performance.

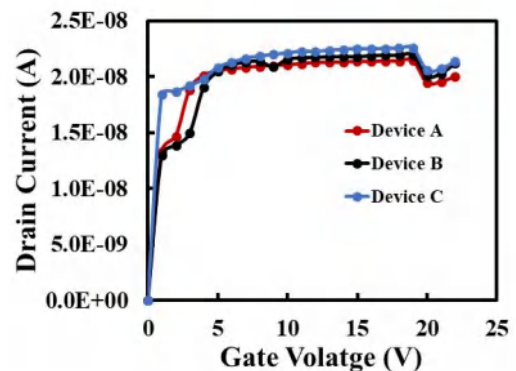


Figure 6 The measured gate leakage current for device A, device B and device C.

2. Breakdown Characteristic

Figure 7 shows the measured breakdown voltages of

the three devices. The breakdown voltages of devices A, B, and C are 1237 V, 1523 V, and 1622 V, respectively. As shown in Figure 8, the maximum gate oxide electric field distribution at $V_{ds}=1200$ V in cell structure of the three devices during the simulation is respectively 2.64 MV/cm, 2.57 MV/cm, and 1.97 MV/cm. The device B has a narrower width of the N+ conductive path compared to the device A. That is to say, the P+ buried layer provides better protection for the gate, which can also be seen from the fact that the gate oxide electric field of the device B is lower than that of the device A. Therefore, the breakdown voltage of device B is higher than that of device A. Furthermore, since the device C provides a higher concentration of P+ buried layer, this will significantly reduce the gate oxide electric field of the device. Therefore, the device C has the highest breakdown voltage.

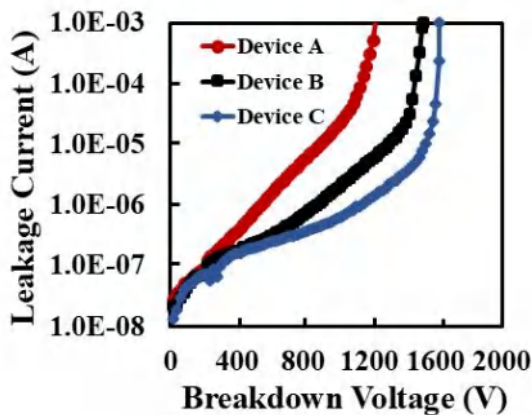


Figure 7 The measured breakdown voltages for device A, device B and device C.

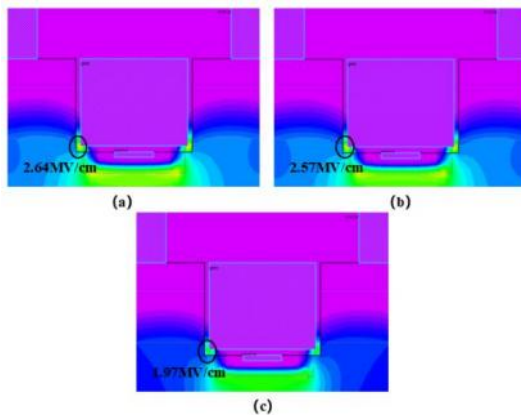


Figure 8 The simulated gate oxide electric field for the three devices. (a) Device A. (b) Device B. (c) Device C.

However, the breakdown voltage of the device is jointly determined by the cell structure and the termination structure. As shown in Figure 9, it presents the simulated electric field distribution of the proposed termination for device A, device B, device C and the device with a P+ buried layer concentration of $3.0 \times 10^{17} \text{ cm}^{-3}$. It can be clearly observed that if the concentration of the P+ buried layer is high or its thickness is large, the electric field of the termination end will continuously increase. The electric field above the P+ buried layer may even drop to 0 MV/cm, making it difficult for the FLR to

function effectively. That is to say, the electric field can more easily penetrate the P+ buried layer with lower concentration and thickness, enabling the FLRs to play an important role. This termination design requires that the concentration or thickness of the P+ buried layer should not be too high. The specific mechanism is as follows: The P+ buried layer first acts as a JTE to modulate the electric field, and then the FLRs modulate the electric field, thereby forming a more reliable and stable termination.

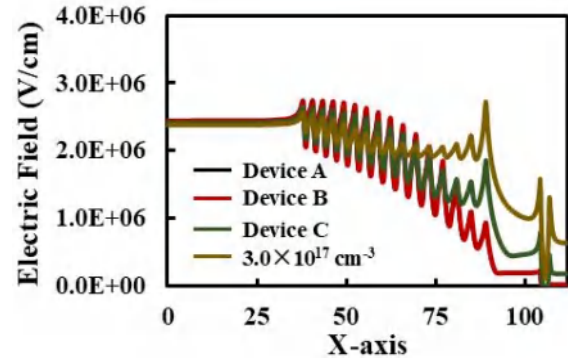


Figure 9 The simulated electric field distribution of the proposed termination for device A, device B, device C and the device with a P+ buried layer concentration of $3.0 \times 10^{17} \text{ cm}^{-3}$.

3. Conduction Characteristic

Figure 10 shows the measured on-resistance at $V_{gs}=20$ V of the three devices. The on-resistances of devices A, B, and C are 86.9 mΩ, 100.1 mΩ, and 102.6 mΩ respectively. Figure 11 shows the simulated current density at $V_{gs}=20$ V of these three devices. It can be clearly observed that when the N+ conductive path becomes narrower, the on-resistance of the device will increase sharply. Moreover, since a higher concentration of P+ buried layer leads to a wider depletion region, resulting in a narrower middle conduction region, the device B and device C have higher on-resistance.

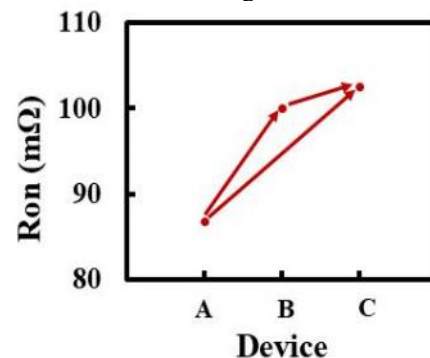


Figure 10 The measured on-resistance for device A, device B and device C.

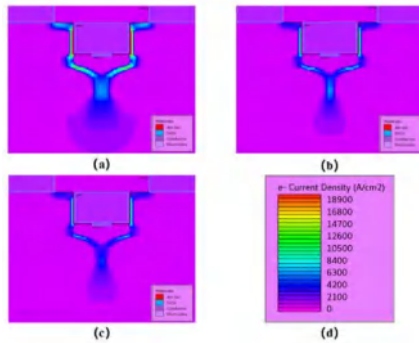


Figure 11 The simulated current density for the three devices. (a) Device A. (b) Device B. (c) Device C. (d) Scaleplate.

4. Threshold Voltage

Figure 12 shows the measured threshold voltage of the three devices. The threshold voltages are all around 4.8V. This indicates that the device has excellent conducting capability and stable consistency.

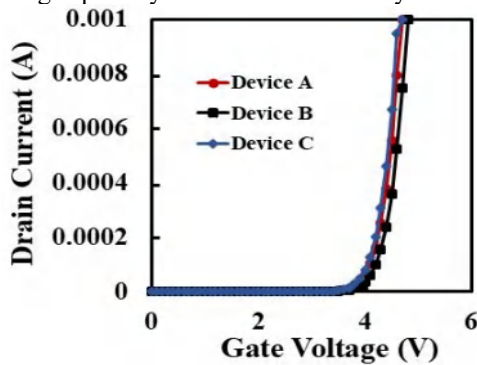


Figure 12 The measured threshold voltage for device A, device B and device C.

Based on the above research, it was found that device B and device C have excellent breakdown characteristic and conduction characteristic, proving that the proposed full-P encapsulated trench MOSFET with N-P-N sandwich epitaxial wafers is feasible. Subsequently, based on this foundation, the trade-off relationship between the breakdown characteristic and the conduction characteristic of the devices will be further improved by adjusting the width and concentration of the N⁺ conductive path as well as the design parameters of the P⁺ buried layer.

V. Conclusions

In this paper, a novel 1.2kV full-P encapsulated trench MOSFET with N-P-N sandwich epitaxial wafers for 4H-SiC devices has been designed and fabricated. Through the design of experiments, three different devices were studied. All three devices have low gate leakage current and stable threshold voltage. Since a narrower N⁺ conductive path or a higher P⁺ buried layer concentration can provide better protection for the gate, devices B and C have higher breakdown voltages.

However, an excessively high concentration of the P⁺ buried layer may cause the termination of the bottom charge modulated field limiting rings to fail. All the three fabricated devices have good conduction characteristic. In conclusion, the device B has a better trade-off relationship between breakdown characteristic and conduction characteristic. Subsequently, the device performance will be further optimized by continuously optimizing the design parameters of the N⁺ conduction path and the P⁺ buried layer.

References

- [1] T. Kimoto, "Material science and device physics in SiC technology for high-voltage power devices," *Jpn. J. Appl. Phys.*, vol. 54, no.4, p.040103 (2015).
- [2] X. C. Deng, Y. Wen, X. D. Wang, et al., "Design and optimization of linearly graded field limiting ring termination for high-voltage SiC diode," *2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology*, 2014: 1-3.
- [3] Y. Zhang, H. Yuan, H. Yang, et al., "Long-Term Lifetime Evolution Mechanism of 4H-SiC MOSFETs Under Nitric Oxide Annealing," *IEEE Transactions on Electron Devices*, vol. 71, no. 12, pp. 7682-7688, Dec. 2024.
- [4] G. J. Hu, G. L. Zhong, et al., "Formation Mechanism of Two Types of Polytype Transformation in off-axis 4H-SiC Boules," *2021 18th China International Forum on Solid State Lighting & 2021 7th International Forum on Wide Bandgap Semiconductors (SSLChina: IFWS)*, Shenzhen, China, 2021, pp. 74-77.
- [5] E. V. Ramos-Fernández, J. Narciso, "Manufacture of SiC: Effect of Carbon Precursor," *Materials*. 2023; 16(5):2034.
- [6] W. Taha, "Comparative Study on Silicon Carbide (SiC) Polytypes in High Voltage Devices," *2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET)*, Hyderabad, India, 2021, pp. 1-6.
- [7] Z. Y. Tang, X. Y. Tang, Y. M. Zhang, et al., "4H-SiC integrated circuits for high-temperature applications," *Journal of Crystal Growth*, Volume 605, 2023, 127060.
- [8] V. F. La, D. Alquier, F. Giannazzo, et al., "Emerging SiC Applications beyond Power Electronic Devices," *Micromachines*. 2023; 14(6):1200.
- [9] X. She, A. Q. Huang, Ó. Lucía, et al., "Review of Silicon Carbide Power Devices and Their Applications," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 10, pp. 8193-8205, Oct. 2017.
- [10] J. Zhang, Z. Chen, Y. Tu, X. Deng and B. Zhang, "A Novel SiC Asymmetric Cell Trench MOSFET With Split Gate and Integrated JBS Diode," *IEEE Journal of the Electron Devices Society*, vol. 9, pp. 713-721, 2021.
- [11] J. Yuan et al., "Design and Optimization of a Novel Full-P Encapsulated 4H-SiC Trench MOSFET with Extremely Compact Cell Pitch," *2024 3rd Asia Power and Electrical Technology Conference (APET)*, Fuzhou, China, 2024, pp. 453-456.
- [12] Z. Han et al., "A Novel 4H-SiC Trench MOSFET Integrated With Mesa-Sidewall SBD," *IEEE Transactions on Electron Devices*, vol. 68, no. 1, pp. 192-196, Jan. 2021.
- [13] J. Yuan, Z. J. Cheng, F. Guo, et al., "A Trench and Field Limiting Rings Co-assisted JTE Termination With N-P-N Sandwich Epitaxial Wafers for 4H-SiC Devices," *IEEE Electron Device Letters*, doi: 10.1109/LED.2024.3416959.
- [14] H. Wang, B. Wang, et al., "4H-SiC Trench Gate Lateral MOSFET with Deep-Shallow Source Trench for Improved RESURF Dose Window and Reduced Oxide Field," *2022 IEEE Workshop on Wide Bandgap Power Devices and Applications in Europe (WIPDA Europe)*, Coventry, United Kingdom, 2022, pp. 1-4.
- [15] J. Ding et al., "A Low-Loss Diode Integrated SiC Trench MOSFET for Improving Switching Performance," *IEEE Transactions on Electron Devices*, vol. 69, no. 11, pp. 6249-6254, Nov. 2022.
- [16] C. Qian, H. Guo, C. Han, H. Yuan, Y. Zhang and Y. Zhang, "Design of High-Efficiency SiC Betavoltaic Battery Structures With Reduced Impact of Near-Surface Recombination Based on Accurate Modeling," *IEEE Transactions on Electron Devices*, vol. 69, no. 12, pp. 7141-7146, Dec. 2022.
- [17] Y. Jiao, J. Q. Zhang, Y. Duan, Y. Chang and P. Liu, "Investigation on Single Pulse Avalanche Failure of 1200V SiC MOSFETs at Different Temperatures," *2023 20th China International*

Forum on Solid State Lighting & 2023 9th International Forum on Wide Bandgap Semiconductors (SSLCHINA: IFWS), Xiamen, China, 2023, pp. 104-108.

[18] K. Takahashi, H. Tanaka, M. Kaneko and T. Kimoto, "Carrier Trapping Effects on Forward Characteristics of SiC p-i-n Diodes Fabricated on High-Purity Semi-Insulating Substrates," *IEEE Transactions on Electron Devices*, vol. 69, no. 4, pp. 1989-1994, April 2022.



Zhijie Cheng received the M.S. degree from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2023. He is currently working at JFS Laboratory, Wuhan, China. His research interest focuses on semiconductor power devices, mainly concentrating on the study of silicon carbide trench MOSFET.
(Email: chengzhijie@jfslab.com.cn)



Jun Yuan received the B.S. degree in School of Physics and Technology from Wuhan University, Wuhan, China, in 2005, the M.S. degree in condensed matter physics from the Beijing Normal University, Beijing, China, in 2010. He is currently studying for a Ph.D. in Electrical and Electronic Engineering at Huazhong University of Science and Technology. At the same time, he focuses research on wide-band gap semiconductor materials and chip technology at JFS Laboratory.
(Email: yuanjun@jfslab.com.cn)

20kV SiC n-IGBT P 型集电区仿真优化设计

叶子轩^{1,2}, 夏云^{1*}, 陈刚¹, 万玉喜¹, 邓小川²

(1 深圳平湖实验室, 深圳市龙岗区平湖街道新厦大道, 518111;

2 电子科技大学, 四川省成都市高新区(西区)西源大道 2006 号, 611731)

(Email*: xiayun@phlab.com.cn)

摘要: 由于目前无高质量 P 型 4H-SiC 衬底, SiC n-IGBT 的 P 型集电区通常采用外延生长工艺形成, P 型集电区浓度及厚度不仅影响外延质量也对所制备的 SiC n-IGBT 性能影响较大。本文基于 TCAD 仿真平台, 系统研究了 P 型集电区结构参数对 20kV SiC n-IGBT 器件开关特性的影响规律。研究表明, P 型集电区浓度、厚度参数直接影响器件电导调制效应, 从而影响器件的导通压降和关断损耗。当 P 型集电区浓度为 $1 \times 10^{19} \text{cm}^{-3}$ 、厚度为 $10 \mu\text{m}$ 时, 器件获得导通压降 ($V_{\text{on}}=5.4\text{V}$) 和关断损耗 ($E_{\text{off}}=8.7\text{mJ}$) 的最佳折中。本研究为超高压 SiC IGBT 器件设计提供了参考和理论依据。

关键词: SiC IGBT; P 型集电区; 导通压降; 关断损耗

1 前言

作为第三代半导体材料, 碳化硅 (SiC) 具有宽禁带、高临界击穿电场、高热导率等优异特性^[1]。相较于传统硅 (Si) 基器件和 SiC MOSFET 器件, SiC 绝缘栅双极型晶体管 (Insulated Gate Bipolar Transistor, IGBT) 兼具超高阻断电压与低导通电阻的特性, 在智能电网、电气化交通等超高压应用场景中具有广阔应用前景^{[2][3]}。

目前超高压 SiC IGBT 器件性能优化相关研究多集中在器件表面结构、漂移区结构以及缓冲层结构的参数优化^{[4][5][6]}。然而, 集电区结构参数对器件电学特性的影响尚未得到系统研究。由于目前尚无高质量 P 型 4H-SiC 衬底, SiC n-IGBT 器件一般基于 N 型衬底制备, P 型集电区则基于 N 型衬底采用外延工艺来制备。P 型集电区的浓度或厚度过高, 一方面会增加 N 型外延层的缺陷密度, 降低器件可靠性^[7], 另一方面会导致增加晶圆翘曲度, 影响后续工艺加工。因此, 研究 P 型集电区浓度和厚度对 SiC n-IGBT 性能的影响, 对于 SiC n-IGBT 的材料开发以及器件性能优化均具有重要意义。本文利用 TCAD 仿真软件, 系统仿真分析了 P 型集电区浓度和厚度对 20kV SiC n-IGBT 器件导通压降 (V_{on}) 和关断损耗 (E_{off}) 的影响, 为超高压 SiC IGBT 器件设计提供了参考和理论依据。

2 器件结构与仿真模型

本文仿真设计的 SiC n-IGBT 器件结构如图 1 所示, 采用常规平面栅结构^[8]。为实现 20kV 耐压, 器件 N 型漂移区厚度为 $180 \mu\text{m}$ 、掺杂浓度为 $2 \times 10^{14} \text{cm}^{-3}$; N

型缓冲层厚度为 $3\mu\text{m}$ 、掺杂浓度为 $3\times 10^{17}\text{cm}^{-3}$ 。为实现 20A 通流能力，器件有源区面积设置为 0.42cm^2 。

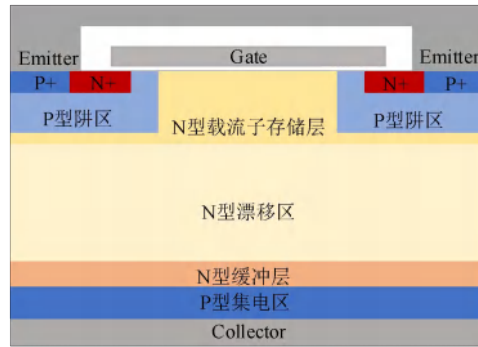


图 1 SiC n-IGBT 器件仿真结构示意图

在正向导通时，N+发射区通过沟道向漂移区注入电子，P型集电区向N型漂移区注入空穴，电子空穴对在漂移区内积累而发生电导调制效应。电导调制效应越强，漂移区电阻越低，器件导通压降越低。P型集电区的厚度和浓度会影响空穴的注入效率，从而影响器件的电学特性^[9]。集电区厚度增大、浓度增加，正向导通时集电极空穴注入效率越高，漂移区空穴浓度越高，电导调制效应增强，导通压降降低，导通特性越好。然而，当IGBT器件关断时，需要将漂移区内电子空穴对抽取或复合完，器件才能完全关断。因此导通时注入的空穴越多，在关断时需要抽取的空穴也越多，关断时间越长，关断损耗增加。本文将通过TCAD仿真方法具体分析P型集电区浓度、厚度对IGBT器件电学特性的影响。

本文使用Sentaurus TCAD对SiC n-IGBT的正向导通特性、阻断特性、关断特性进行仿真分析，所使用的物理模型包括SRH复合模型、载流子高低电场迁移率模型、杂质的不完全电离模型、碰撞电离模型等。

3 仿真结果分析

3.1 P型集电区浓度对器件电学特性的影响

通过正向特性仿真，得到P型集电区浓度对器件导通特性的影响如图2，可以看出随着集电区浓度从 $1\times 10^{18}\text{cm}^{-3}$ 增加到 $1\times 10^{19}\text{cm}^{-3}$ 时，器件导通压降 V_{on} 迅速降低，从 7.7V 降低至 4.8V。而当集电区浓度超过 $1\times 10^{19}\text{cm}^{-3}$ 时，导通性能的提升趋于饱和，增加浓度带来的静态特性优化效果有限。

相同导通电流下，P型集电区掺杂浓度对漂移区内空穴浓度分布影响如图3所示。随着P型集电区掺杂浓度增加，空穴注入效率增强，漂移区内分布的空穴浓度越高，电导调制效应因此越强。当掺杂浓度在 $1\times 10^{19}\text{cm}^{-3}$ 时，漂移区内的空穴已经达到较高水平，继续增加掺杂浓度，对提高空穴浓度的效果不明显。

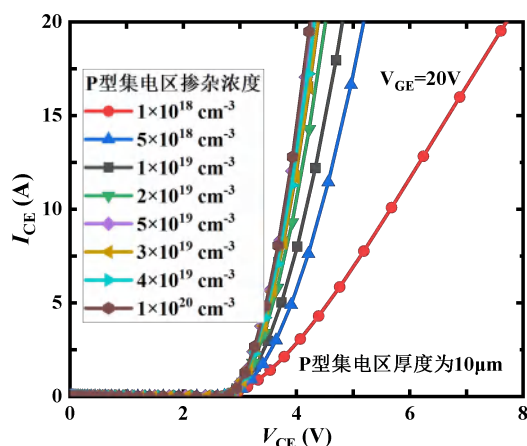


图 2 P 型集电区掺杂浓度对导通特性的影响

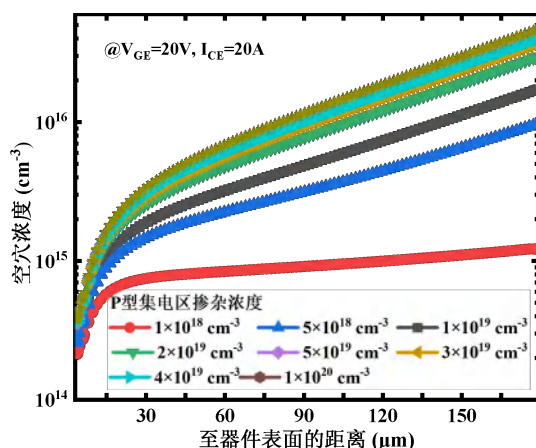


图 3 P 型集电区掺杂浓度对漂移区空穴分布影响

通过动态特性仿真得到器件 $E_{off} - V_{on}$ 折中曲线，如图 4 所示。器件关断损耗随着 P 型集电区掺杂浓度的增加而不断增大。P 型集电区浓度从 $1 \times 10^{18} \text{ cm}^{-3}$ 提升到 $1 \times 10^{19} \text{ cm}^{-3}$ 时，关断损耗小于 20 mJ；而一旦浓度超过 $1 \times 10^{19} \text{ cm}^{-3}$ ，器件关断损耗成倍增长。考虑关断损耗与导通压降的折中关系，同时考虑到工艺实现难易程度，P 型集电区浓度过高会导致外延缺陷增加、外延质量下降。因此，在本论文中，20kV SiC n-IGBT 的 P 型集电区掺杂浓度设置为 $1 \times 10^{19} \text{ cm}^{-3}$ ，此时器件具有折中优化的关断损耗及导通压降性能。

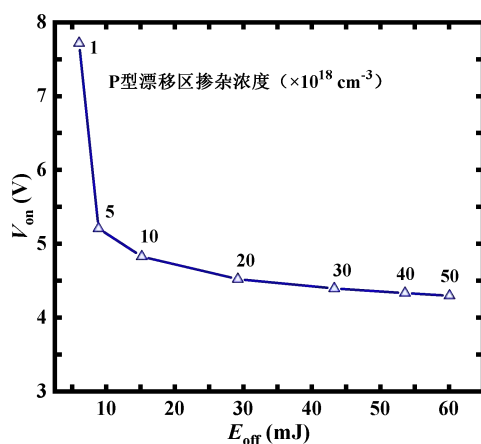


图 4 P 型集电区掺杂浓度对 $E_{off} - V_{on}$ 的影响

3.1 P型集电区厚度对器件电学特性的影响

在P型集电区浓度为 $1 \times 10^{19} \text{ cm}^{-3}$ 的基础上, 研究P型集电区厚度变化对器件的关断及导通性能的影响。如图5所示, P型集电区厚度 $1 \mu\text{m}$ 增至 $10 \mu\text{m}$ 时, 器件导通性能明显提升, V_{on} 从 24.6 V 降至 5.4 V 。继续增加集电区厚度, V_{on} 下降十分有限。

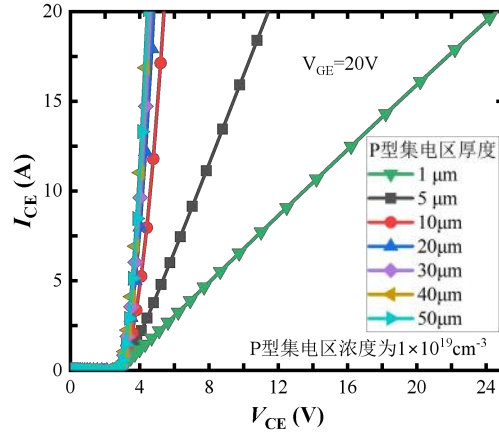


图5 不同P型集电区厚度下I-V特性曲线

图6截取了相同集电极电流下器件漂移区内少数载流子分布情况。集电区厚度 $1 \mu\text{m}$ 时空穴注入效率低, 电导调制效应较差, 导通压降过高。随着P型集电区厚度增加, 集电极载流子注入效率增强。当厚度增至 $10 \mu\text{m}$ 时, 漂移区载流子浓度已经达到较高水平, 导通压降也降至 5.4 V , 继续增加厚度对空穴注入效率的提升十分有限。

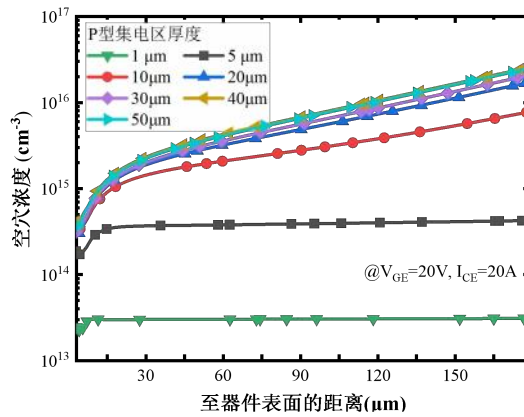


图6 P型集电区厚度对漂移区空穴分布影响

通过动态特性仿真得到P型集电区对 $E_{\text{off}} - V_{\text{on}}$ 折中特性影响, 如图7所示。随着P型集电区厚度增加, 正向导通过程中漂移区内存储的载流子浓度增加, 器件关断时需要抽取的载流子数量增多, 关断时间延长, 关断损耗增大。考虑关断损耗与导通压降的折中关系, 20 kV SiC IGBT 集电区厚度位 $10 \mu\text{m}$ 时最优, 此时器件的关断损耗为 8.7 mJ , 导通压降为 5.4 V 。

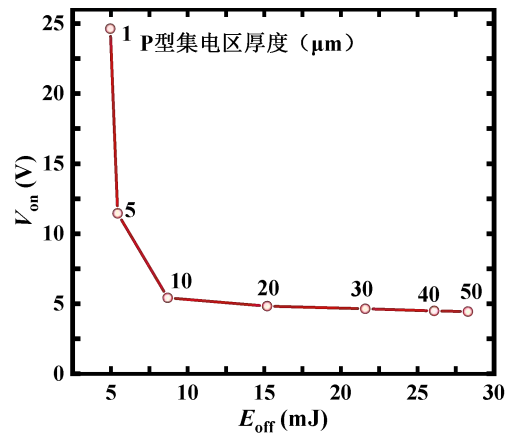


图 7 P 型集电区厚度对 $E_{off} - V_{on}$ 的影响

4 总结

本文研究了 20kV SiC n-IGBT 的 P 型集电区参数（浓度、厚度）对器件正向导通特性、关断特性以及 $E_{off} - V_{on}$ 折中特性的影响，并进行了理论分析。20kV SiC n-IGBT P 型集电区浓度为 $1 \times 10^{19} \text{ cm}^{-3}$ 、厚度为 10 μm 时，器件表现出最优电学特性，实现了 5.4V 的低导通压降，同时关断损耗为 8.7mJ。仿真研究结果明确了 P 型集电区参数对器件静态/动态特性的影响，为超高压 SiC IGBT 集电区参数设计提供关键设计准则和理论依据。

参考文献

- [1] R. Luo, B. Sun, X. Hou, W. Shi, G. Zhang and J. Fan. Optimal Design of 100–2000 V 4H–SiC Power MOSFETs Using Multi-Objective Particle Swarm Optimization Algorithms[J]. IEEE Electron Device Letters, 2024, 45(5): 786-788.
- [2] L. Han, L. Liang, Y. Kang and Y. Qiu. A Review of SiC IGBT: Models, Fabrications, Characteristics, and Applications[J]. IEEE Transactions on Power Electronics, 2021, 36(2): 2080-2093.
- [3] Abdalgader IAS, Kivrak S, Özer T. Power Performance Comparison of SiC-IGBT and Si-IGBT Switches in a Three-Phase Inverter for Aircraft Applications[J]. Micromachines. 2022, 13(2): 313.
- [4] 张莉, 陈致宇. 18 kV 4H-SiC ESC-IGBT 结构设计与特性研究[J]. 现代电子技术, 2023, 46(18): 47-52.
- [5] 吴栋, 姚登浪, 郭祥, 等. 具有可控空穴抽取路径的低损耗 4H-SiC 双沟槽超结 IGBT[J]. 电子元件与材料, 2024, 43(09): 1071-1080.
- [6] 刘国友, 高云斌, 陈喜明, 等. SiC IGBT 正向导通特性研究[J]. 电力电子技术, 2017, 51(08): 42-43+57.
- [7] 郭钰, 刘春俊, 张新河, 等. 碳化硅同质外延质量影响因素的分析与综述[J]. 人工晶体学报, 2024, 53(02): 210-217.
- [8] 杨晓磊, 李士颜, 赵志飞, 等. 超高压碳化硅 N 沟道 IGBT 器件的设计与制造[J]. 电子与封装, 2022, 22(04): 14-19.
- [9] Wu L, Liu M, Zhang M, et al. Low on-state voltage and EMI noise 4H- SiC IGBT with self-biased split - gate PMOS [J]. IEEE Transactions on Electron Devices, 2023, 70(2): 647-652.

温度升高对碳化硅快速离子器件特性影响的分析

张宏伟¹⁾, 张庶鑫¹⁾, 蒲红斌^{1, 2)}

¹⁾ 西安理工大学电子工程系, 陕西 西安 710048

²⁾ 西安市电力电子器件与高效电能变换重点实验室, 陕西 西安 710048

*Email: puhongbin@xaut.edu.cn

摘要: 快速离子器件 (Fast Ionization Dynistor, FID) 是一种新型全固态半导体脉冲功率开关, 广泛应用于超快大功率脉冲领域。硅 (Silicon, Si) FID 是脉冲功率开关的传统研究方向, 而开展碳化硅 (Silicon Carbide, SiC) FID 则具有重要的理论意义和应用前景。温度对半导体开关的正常工作至关重要, Si FID 的研究成果多是基于室温条件, SiC FID 也有 100°C 下稳定输出的报道, 但关于温度对其特性影响的研究仍然非常有限。本研究采用数值仿真的方法, 研究了温度对 SiC FID 的静态与动态特性影响的差异, 探讨了与温度相关的动态和静态击穿物理机制, 并据此解释了温度对两类特性影响差异的原因。研究表明, 随着温度升高, SiC FID 的静态特性有所提高, 其主要原因是载流子的漂移速度随温度升高而减低; 而动态特性则呈现下降趋势, 这是由于电子和空穴浓度对碰撞电离率的影响超过了电场变化带来的作用。在单次触发模式下, 温度升高对 SiC FID 特性的影响较小; 但在重复频率模式下, 温度变化的影响不能忽略, 需要更多分析和研究。

关键词: 碳化硅; 快速离子器件; 碰撞电离; 特性

1 引言

快速离子器件 (Fast Ionization Dynistor, FID)^[1] 是一种类晶闸管结构的新颖半导体全固态脉冲功率开关, 其基本工作原理是俄罗斯约飞物理研究所发现的可逆延迟雪崩击穿效应^[2]。由于普通商用晶闸管在高压触发脉冲下也会出现类似 FID 的可逆延迟雪崩击穿效应并实现超快开通, 所以近年来对 FID 的研究对象由最初自行研发的器件转变为直接采用普通商用晶闸管^[3]。相应地, 该类器件的名称也由逐渐演变为深能级晶闸管 (Deep Level Dynistor, DLD) 或高功率晶闸管 (High-Power Thyristor, HPT)。

硅 (Silicon, Si) 是制备 FID 的传统半导体材料, 但 Si FID 在耐压水平、开关速度、工作温度、器件尺寸以及可靠性等方面受到材料极限得限制^[4]。近年来, 随着碳化硅 (Silicon carbide, SiC) 材料和工艺的不断成熟, 这种宽禁带半导体材料在超高压、高压功率器件研制领域展现出显著优势^[5]。例如, 科锐 (Cree, Inc.) 公司于 2013 年率先报道了 22 kV 的 SiC

GTO^[6]; 西安理工大学于 2016 年报道了 20 kV SiC 晶闸管及其终端结构^[7]; 东南大学则于 2020 年报道了 6kV 的 SiC FID^[8]。与 Si FID 相比, SiC FID 具有明显优势, 但目前国内外对 SiC FID 的研究报道却相对较少。

大多数的半导体器件对其使用温度范围都有严格规定, 一旦超过该范围, 便存在器件失效或被损坏的风险。温度对 FID 的正常工作也同样重要, 然而目前相关研究较为有限。在有直流偏压条件中, Si FID 在室温下均可正常工作, 但高温下的实验结果尚未见报道^[9-14]; 而在无直流偏压的条件下, 已有研究表明当温度达到 200°C, Si FID 仍可观察到超快开通现象^[15]。对于碳化硅 FID 温度测试表明, 在工作温度达到 100°C 时, 该器件仍可实现稳定电流脉冲输出^[8]。

尽管已有部分文献表明温度对 FID 的工作特性存在一定影响, 但这些研究大多停留在实验现象的描述层面, 缺乏对温度变化趋势下 FID 开关特性演变规律的系统分析, 更未深入探讨其背后的物理机制。为此, 为此, 本研究采用数值仿真方法, 系统研究温度变化对 SiC FID 特性的影响, 获取了不同温度下器件的静态和动态特性参数, 并结合温度

对仿真模型的影响，深入分析了器件静态和动态特性随温度变化产生差异的原因。

2 器件结构和仿真方法

图 1 是 SiC FID 的仿真器件的元胞示意图，器件整体结构参考晶传统 SiC 闸管缓冲层设计的方案，采用 P⁺⁺/N⁺/N⁻/P⁻/N⁺⁺ 五层结构的穿通结构设计。各层的掺杂浓度和厚度参数参考文献 [7] 和文献 [6]，以确保其具备承受 3.3 kV 正向阻断电压的能力。此外，在 SiC FID 的世纪制备过程中，需要单独设计器件的终端结构，考虑采用二氧化硅或氮化硅作为终端钝化层材料。

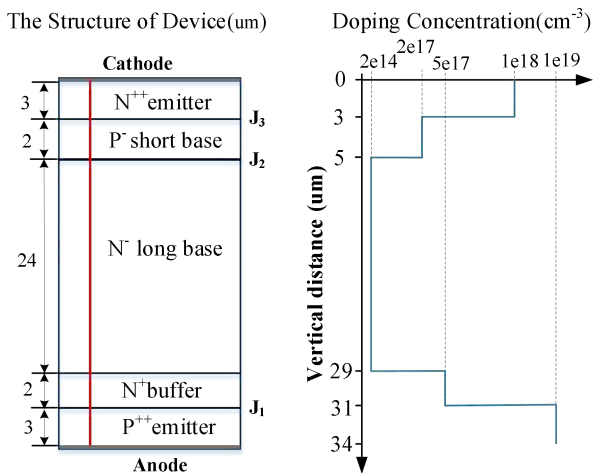


图 1 SiC FID 的仿真器件的元胞示意图。

为了研究温度变化对 SiC FID 的动态特性的影响，本研究使用 Sentaurus-TCAD 仿真软件^[16]进行建模分析。仿真中选用了禁带窄化模型、不完全电离模型、速度饱和模型、载流子与载流子散射模型、俄歇复合模型、SRH 复合模型、雪崩模型等物理模型等。器件所采用物理模型及其参数见文献[17]。器件的电学特性是通过耦合求解泊松方程、连续性方程、输运方程和基尔霍夫定律获得。

图 2 是 SiC FID 的仿真电路示意图。该电路包括三个支路。左侧为触发回路(TC)，由脉冲电压源 U_s ，隔离电容 C_1 ，FID 构成，其主要功能提供触发脉冲电压。中间为主回路(MC)，由储能电容 C_0 ，负载电阻 R_L ，隔离电感 L_0 和 FID 组成，用于在 SiC FID 被触发脉冲电压开通后，通过 C_0 放电在负载电阻上产生高压脉冲电流。右侧为充电回路(CC)，

由直流电源 U_0 ，限流电阻 R_1 ，开关 S_1 ，隔离电容 L_0 ，负载电阻 R_L 和储能电容 C_0 组成，其作用是给 SiC FID 提供初始偏压，并为 C_0 进行预充电。

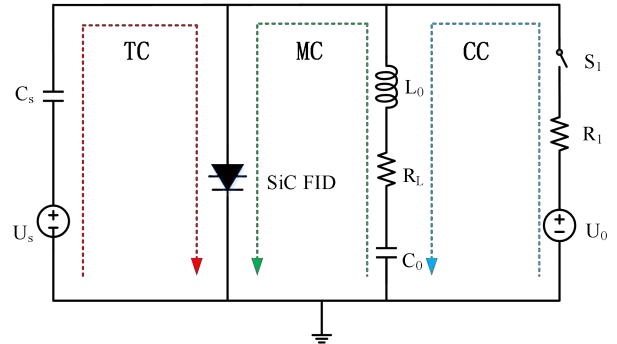


图 2 SiC FID 的仿真电路示意图

在仿真过程中，器件有源区的面积设定为 0.16 mm^2 ，电路中各元件参数设置如下：直流电源 $U_0=3.5\text{kV}$ ，限流电阻 $R_0=2 \text{ k}\Omega$ ，隔离电容 $C_s=0.1 \text{ nF}$ ，隔离电感 $L_0=200 \text{ nH}$ ，储能电容 $C_0=60 \text{ nF}$ ，负载电阻 $R_L=0.25 \text{ }\Omega$ ，外部触发脉冲 U_s 的脉冲幅值变化率 $A=15 \text{ kV/ns}$ 。

3 结果和讨论

3.1 温度对 SiC FID 静态特性的影响

SiC FID 的静态特性主要是器件的正向阻断能力，通常对静态特性的研究集中于不同参数变化对静态击穿电(U_{bv})的影响。图 3 展示了 300K, 400K, 500K 和 600K 温度条件下，SiC FID 的正向阻断特性曲线。从图中可以看出，随着温度的升高，SiC FID 的 U_{bv} 呈现持续上升的趋势。

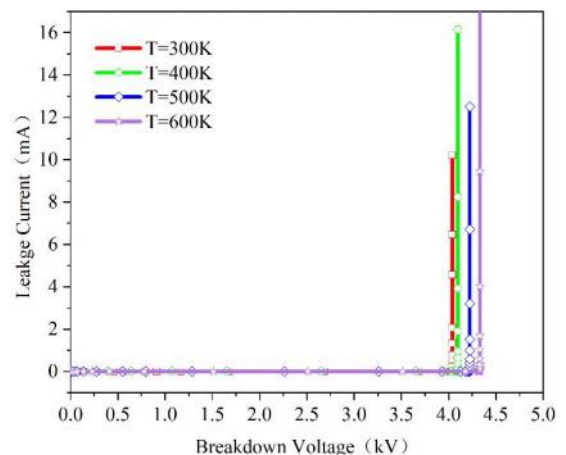


图 3 不同温度下 SiC FID 的静态击穿电压

表 1 不同温度下 SiC FID 的静态特性参数比较

T	300K	400K	500K	600K
U_{bv}	3909	4097 (4.80) higher	4224 (8.06) higher	4332 (10.82) higher

表 1 是 300K, 400K, 500K 和 600K 温度条件下的 U_{bv} 。与 300K 的 U_{bv} 相比, SiC FID 在 400K, 500K 和 600K 时的 U_{bv} 分别提升了 4.80%, 8.06% 和 10.82%。由上述静态仿真结果可知, 随着温度升高, SiC FID 的静态特性表现出增强的趋势。

3.2 温度对 SiC FID 动态特性的影响

FID 动态特性一般是指在器件阳-阴极施加直流正向阻断偏压和触发脉冲电压后 FID 的电压随时间变化的特性。通常, 研究 FID 的动态特性参数有阻断电压峰值(U_{max})、残余电压(U_{res})和开通时间(T_{on})等。

图 4 是 FID 是开通特性示意图。如图所示, 阻断电压峰值 (U_{max})是器件在触发脉冲下阳-阴极的最大阻断电压。残余电压是器件在被外部脉冲触发后阳-阴极电压的最小值。开通时间(T_{on})是器件的阻断电压峰值从 $90% * U_{max}$ 下降到 $10% * U_{max}$ 的时间间隔。

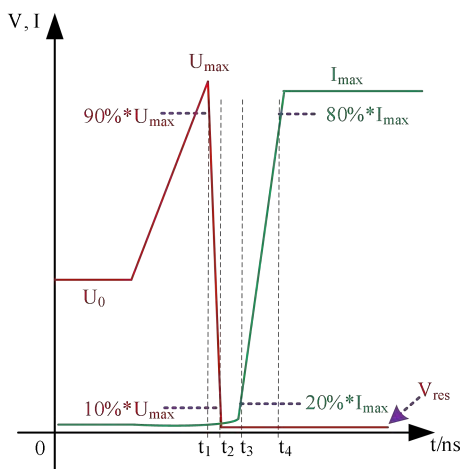


图 4 FID 的开关特性示意图

不同温度下 SiC FID 的开通特性曲线如

图 5 所示, 由图可见, 在温度升高过程中, SiC FID 的 U_{max} , U_{res} 和 T_{on} 等动态参数均呈现上升趋势。表 2 是 300K, 400K, 500K 和 600K 温度下的 U_{max} 、 T_{on} 和 U_{res} 。与 300K 的 U_{bv} 相比, SiC FID 在 400K, 500K 和 600K 时的 U_{max} 分别提高了 6.00%, 12.16% 和 16.67%; T_{on} 分别提高了 2.04%, 9.69% 和 21.92%; U_{res} 分别提高了 24.23%, 56.37% 和 102.01%。

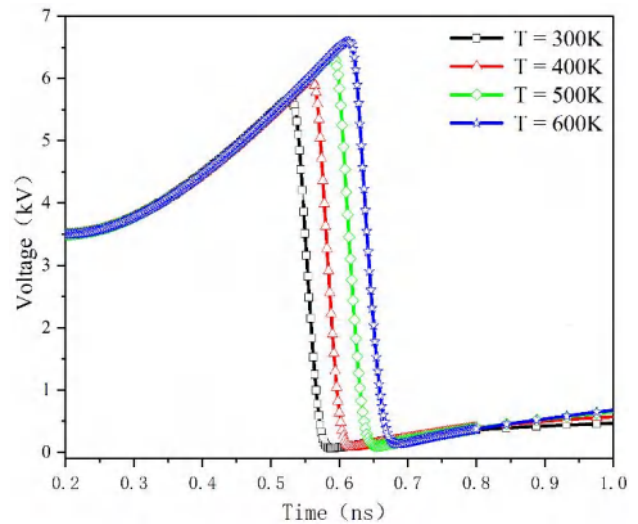


图 5 不同温度下 FID 的开通特性曲线

表 2 SiC FID 的动态特性参数比较

T	U_{max}	T_{on}	U_{res}
300K	5651	29.93	65.74
400K	5990 (6.00%) higher	30.54 (2.04%) higher	81.67 (24.23%) higher
500K	6338 (12.16%) higher	32.83 (9.69%) higher	102.8 (56.37%) higher
600K	6598 (16.76%) higher	36.49 (21.92%) higher	132.8 (102.01%) higher

备注: ps = picosecond, V = voltage

4 温度对器件特性的影响因素分析

4.1 温度对物理模型的影响分析

对于半导体器件而言, 温度升高主要影响仿真过程器件基区本征载流子浓度、少子寿命、迁移率、饱和速度和碰撞电离率等五个物理模型。首先分析本征载流子浓度随温

度升高影响。在 SiC 材料的器件中，器件基区本征载流子浓度随温度变化的关系式为：

$$n_i = 1.70 \times 10^{16} \cdot T^{3/2} \cdot e^{-(2.08 \times 10^4)/T} \quad (1)$$

式 (1) 中 n_i 是本征载流子浓度， T 是温度。由该式可知，随着温度的升高，器件的本征载流子浓度会增大。结合掺杂浓度、少数浓度与本征载流子的关系可知，本征载流子浓度的增加会使器件基区少数浓度增加。

其次是关于少数寿命随温度升高的影响分析。根据 SRH 复合模型，载流子寿命随温度变化的关系式为：

$$\tau(T) = \frac{\tau_{n_0,p_0}}{1 + \left(\frac{N_{D,A}}{3 \times 10^{17}}\right)} \cdot \left(\frac{T}{300}\right)^{1.72} \quad (2)$$

式 (2) 中 $\tau(T)$ 是少数寿命， τ_{n_0,p_0} 是电子和空穴在 $T=0$ 时的寿命， $N_{D,A}$ 是施主和受主掺杂浓度， T 是温度。由该式可知，随着温度的升高，器件基区少数载流子寿命将会随之增加。在其他因素不变的情况下，器件长基区少数寿命增加会减弱载流子的复合效应。

再次是关于迁移率和饱和速度随温度升高影响分析。在低场条件下，载流子的漂移速度与迁移率是线性关系，即：

$$v = \mu_{n,p} \cdot E \quad (3)$$

式 (3) 中 v 是载流子的漂移速度， E 是电场强度， $\mu_{n,p}$ 是电子和空穴的迁移率。由该式可知，在低场条件下，载流子的漂移速度随着电场增大。但在高电场条件下，载流子的迁移率受到载流子速度饱和效应的影响，其迁移率的表达式为：

$$\mu_{high} = \frac{\mu_{low}}{\left[1 + \left(\frac{\mu_{low} \cdot E}{v_{sat}}\right)^\delta\right]^{1/\delta}} \quad (4)$$

式 (4) 中 μ_{high} 是高场下的迁移率， μ_{low} 是低场下的迁移率， E 是电场强度， δ 是常数， v_{sat} 是饱和速度。其中 v_{sat} 与温度的关系式为：

$$v_{sat}(T) = v_{sat,0} \cdot \left(\frac{300K}{T}\right)^{v_{sat,exp}} \quad (5)$$

式 (5) 中 $v_{sat}(T)$ 是载流子的饱和速度， $v_{sat,0}$ 是 $T=0K$ 时的饱和速度， $v_{sat,exp}$ 是修正常数， T 是温度。由该式可知， $v_{sat}(T)$ 随着温度的升高而减小。

最后碰撞电离率随温度升高影响分析。仿真中碰撞电离率的表达式为：

$$G = \alpha_n \cdot n \cdot v_n + \alpha_p \cdot p \cdot v_p \quad (6)$$

式 (6) 中 α_n 和 α_p 是电子和空穴的碰撞电离系数， n 和 p 是电子和空穴浓度， v_n 和 v_p 是电子和空穴的漂移速度。其中电子和空穴碰撞电离系数 α_n 和 α_p 有下面的表达式得到：

$$\alpha_{n,p} = a [1 + c(T - T_0)] E^\gamma e^{-\left(\frac{b[1+d(T-T_0)]}{E}\right)^\delta} \quad (7)$$

式 (7) 中 a, b, c 和 d 是常数， E 是电场强度， $T_0=0K$ ， T 是温度。由该式可以得出，温度升高会导致电子和空穴的碰撞电离系数减小，基区电子空穴浓度也会随之减少，两者共同作用导致碰撞电离率减少。

4.2 温度对器件静态和动态特性的影响分析

通过上述温度对器件仿真物理模型的分析可知，温度升高会导致本征载流子浓度增大、少数寿命增大、迁移率减小、饱和速度减小，碰撞电离率减小。然而，对于 SiC FID 静态和动态过程而言，最主要的物理模型是碰撞电离模型。从本质上讲，SiC FID 静态和动态过程都是载流子碰撞电离的过程，其差别在于：静态过程是非平衡在低场通过碰撞电离逐渐积累至开通的过程。而动态过程则是非平衡载流子在高场下通过强烈碰撞电离产生大量电子空穴对而实现开通的过程。

在 SiC FID 静态过程中，直流电压是从小到大逐渐加载到器件阴阳极两端的，电子和空穴的碰撞电离系数随着电场的增强而逐渐增大，器件基区在非平衡载流子的漂移速度也随着电场逐渐增加。但在静态条件下，温度升高导致载流子的迁移率减小。迁移率的减小导致非平衡载流子的漂移速度减小，

该速度的减小进而会导致碰撞电离产生率下降,从而导致在静态仿真中出现低温下 U_{vb} 略低于高温下的 U_{vb} 的现象。

图 6 展示是不同温度下 SiC FID 静态击穿时碰撞电离率的差异。可以看出,当温度升高时,器件的碰撞电离率存在明显差异,较低温度下的碰撞电离率高,较高温度的碰撞电离率低。这一结果与前述分析结果一致,表面在静态条件下, SiC FID 的 U_{vb} 随着温度的升高而增大,主要与漂移速度随温度升高而减小密切相关。

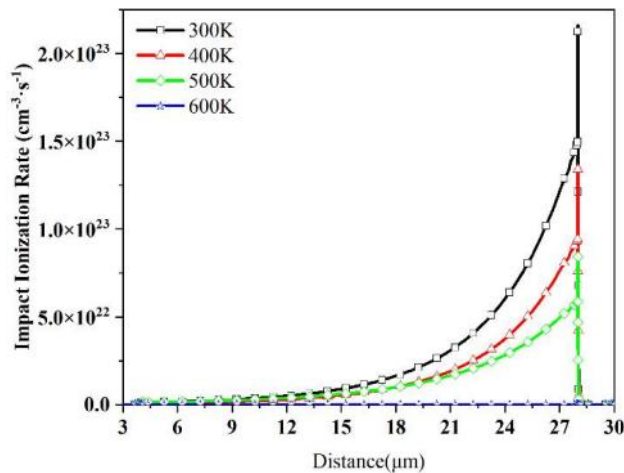


图 6 不同温度 SiC FID 静态击穿时碰撞电离率的差异

在 SiC FID 动态过程中,由于器件两端存在初始偏压,在触发脉冲电压到达之前,器件基区已完全耗尽且处于高电场状态。当触发脉冲开始施加,器件长基区的非平衡载流子在高场下发生延迟雪崩击穿效应,在极短时间内产生大量等离子体而超快开通。

因为是器件基区处于高电场,非平衡载流子在进入基区后漂移速度迅速达到饱和速度,电子空穴的碰撞电离系数也较大。因此,在 SiC FID 动态过程中,长基区发生的碰撞电离要比静态击穿过程中的偏置电离剧烈,所以产生电子空穴对的浓度和生成速度也远高于比静态条件,从而使得器件在开通后残留电压更低。

图 7 展示了不同温度下 SiC FID 动态击穿时碰撞电离率的差异。可以看出,在动态击穿过程中,不同温度下的碰撞电离率仍表现出低温高、高温低的趋势,与静态击穿趋势一致。而,与静态过程不同的是,动态过

程中的碰撞电离在整个基区内呈现准均匀分布,而静态击穿时的碰撞电离则主要集中在 J_2 结附近。

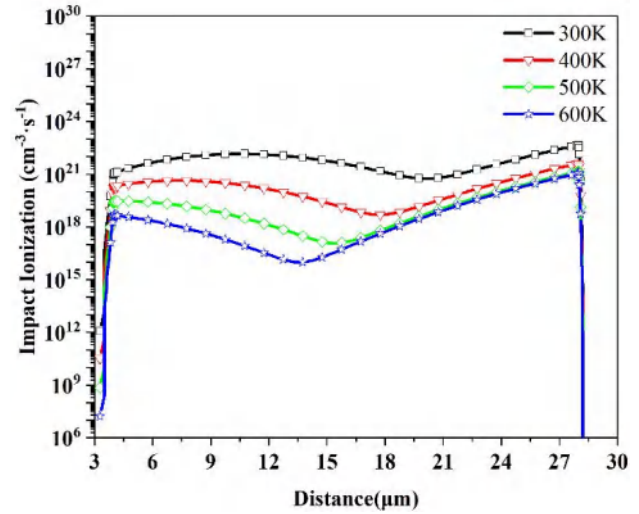


图 7 不同温度 SiC FID 动态击穿时碰撞电离率的差异

这两种碰撞电离率在器件基区的空间分布上的差异,一方面导致新生成的电子与空穴浓度在基区内的分布存在差别;另一方面,这些新生载流子也会对基区的电场分布产生影响。电子与空穴浓度在基区中空间分布的差异,决定了 SiC FID 在静态与动态击穿过程中的行为不同,开通时间及阳极电压峰值的差异也与此一致。

新生载流子浓度对基区电场分布的影响主要体现在 SiC FID 的动态过程中。由于低温下的碰撞电离率高于高温下的数值,因此低温条件下基区中新产生的电子-空穴对浓度也高于高温条件。新生载流子浓度对碰撞电离过程的影响要强于电场变化对其的作用,从而导致在高温下, SiC FID 的开通时间比低温下更长,残余电压也更高。

5 结论

本文采用仿真的方法研究了温度对非对称结构 SiC FID 静态和动态特性的影响。研究表明:随着温度升高, SiC FID 的静态特性有所提升,而其动态特性则呈现下降趋势。然而,作为一类用于脉冲功率系统的开关器件, SiC FID 动态性能的下降在实际应用中是不利的,因为温度升高会增加器件开通过程中的能量损耗。

进一步分析了温度对本征载流子浓度、少子寿命、迁移率、速度饱和和碰撞电离等模型的影响。在静态条件下，SiC FID 长基区的电场从低到高的逐渐变化，温度对物理模型的影响于常规半导体类似。而在动态的过程中，电子和空穴浓度对碰撞电离率的影响大于电场对碰撞电离率的影响，进而导致器件的动态特性下降。

尽管目前尚缺乏关于温度对 SiC FID 特性影响的实验报道，但从本研究的结果来看，温度升高确实对 SiC FID 的动态性能产生不利影响，具体表现为开通时间 T_{on} 和残余电压 U_{res} 的增加，这将显著提高器件在开通过程中的功耗。对于单次开关而言，温度升高的影响相对较小，这当然是碳化硅器件的材料特性优势。然而，这类器件通常应用于重复频率（重频）工作模式下，在该模式下温度升高将成为不可忽视的问题。因此，未来有必要针对 SiC FID 在重频模式下的热效应开展更深入的研究。

参考文献

- [1] V. Efanov, V. Karavaev, A. Kardo-Sysoev, et al. Fast ionization dynistor (FID)-a new semiconductor superpower closing switch. Digest of Technical Papers 11th IEEE International Pulsed Power Conference Baltimore, MA, USA, 1997, 988-991.
- [2] I.V. Grekhov, A.F. Kardo-Sysoev, L.S. Kostina, et al. high-power subnanosecond switch[J]. ELECTRONICS LETTERS, 1981, 17 (12): 422-423.
- [3] A.I. Gusev, S.K. Lyubutin, S.N. Rukin, et al. High power thyristors triggering providing a subnanosecond closing time[J]. International Power Modulator and High Voltage Conference (IPMHVC), 2014, 31-34.
- [4] E. Shahriari, T. Maysonnave, A.I. Gusev, et al. A Study of a 5.2-kV/100-mm Thyristor Triggered in Impact-Ionization Mode Toward Fast High-Current Applications[J]. IEEE Transactions on Plasma Science, 2022, 50 (10): 3452-3458.
- [5] P. Rodin, P. Ivanov, I.V. Grekhov. Performance evaluation of picosecond high-voltage power switches based on propagation of superfast impact ionization fronts in SiC structures[J]. Journal of applied physics, 2006, 99 (4): 04503.
- [6] Q. Liu. Research on structure and performance improvement of SiC GTO Thyristor. Xi'an Shanxi: Xi'an University of Technology, 2022.
- [7] X. Wang. Study on novel structure of 4H-SiC light triggered thyristor. Xi'an Shanxi: Xi'an University of Technology, 2019.
- [8] X. Zhao. Design of Silicon carbide High Power Fast Ionization Device and Test Characterization Circuit. Nanjing Jiangsu: Southeast University, 2020.
- [9] A. Gusev, S. Lyubutin, S. Rukin, et al. A thyristor switch with a subnanosecond switching time[J]. Instruments and Experimental Techniques, 2015, 58 376-380.
- [10] V.I. Brylevskiy, I.A. Smirnova, A.V. Rozhkov, et al. Picosecond-range avalanche switching of high-voltage diodes: Si versus GaAs structures[J]. 2016, 44 (10): 1941-1946.
- [11] A. Gusev, S. Lyubutin, S. Rukin, et al. High-current pulse switching by thyristors triggered in the impact-ionization wave mode[J]. Instruments and Experimental Techniques, 2017, 60 (4): 545-550.
- [12] X. Huang. Study on optimal design and Failure analysis of Fast Ionization Dynistor: Huazhong university of Science and Technology, 2022.
- [13] E. Shahriari, A.I. Gusev, T. Maysonnave, et al. Experimental and Simulation Study of Parallel-Connected Thyristors in Impact-Ionization Switching Mode[J]. IEEE T PLASMA SCI, 2024, 1-9.
- [14] E. Shahriari, T. Maysonnave, A.I. Gusev, et al. Series-Parallel Connection of Thyristors Triggered in Impact-Ionization Wave Mode[J]. IEEE T PLASMA SCI, 2024, 1-8.
- [15] O. Perminova, S. Tsyanov. Study of process of avalanche switching of silicon thyristors without bias voltage. Journal of Physics: Conference Series, 2018, 022021.
- [16] TCAD Sentaurus Device Manual. Synopsys Inc. Mountain View, CA, USA, 2018.
- [17] H. Zhang, H. Pu, Z. Zhang, et al. Comparative Study on the Normal Avalanche Breakdown and Delayed Avalanche Breakdown for 10kV 4H-SiC FID. 3rd ISSET, Xi'an, China, 23-25 Aug. 2024 2024, 638-642.

短路瞬态应力下不同平面栅结构对 SiC MOSFET 内多物理场行为的影响

许士康, 赵汉青, 李轩*, 邓小川, 张波

电子科技大学, 集成电路科学与工程学院 (示范性微电子学院), 电子薄膜与集成器件全国重点实验室, 成都, 611731

通讯作者: 李轩, xuanli@uestc.edu.cn

摘要: 随着 SiC MOSFET 在新能源汽车电驱系统、光伏逆变装置等新型能源领域内的规模应用, 但由于较差的可靠性限制了其广泛应用, 尤其是短路可靠性。本文对具备三种栅极结构的平面栅 SiC MOSFET 进行了建模, 并利用有限元仿真软件分析比较了其在短路瞬态应力下的电-热-力多物理场演化过程。结果表明, 不同栅极结构对电场、温度和机械应力分布均有显著影响, 特别是对层间介质处电场和机械应力的影响。本研究对于平面型 SiC MOSFET 短路瞬态应力下的栅极加固和高可靠性设计提供了参考。

关键词: SiC MOSFET; 电-热-力演化; 短路瞬态应力; 栅极结构;

1 前言

在全球能源结构转型加速推进与“双碳”战略深入实施的双重驱动下, 电力电子系统对高效化和小型化的需求日益迫切[1]。碳化硅(Silicon Carbide, SiC)金属-氧化物半导体场效应晶体管(Metal-Oxide Semiconductor Field Effect Transistor, MOSFET)因耐压高、导通损耗低、开关速度快等优势, 可显著提升系统效率并实现设备轻量化, 已规模化应用于新能源汽车电驱系统、光伏逆变装置及数据中心供电模块等新型能源领域[2]。然而, 由于较高的功率密度, SiC MOSFET 的可靠性相较于传统硅基绝缘栅型场效应晶体管 (Silicon based Insulate Gate Bipolar Transistor, Si IGBT) 仍较差。其中, 短路(short circuit)可靠性问题尤为突出, 其短路最大可耐受时间较短, 对于外部保护电路的响应时间要求更高, 严重限制了其广泛应用。因此急需对 SiC MOSFET 在短路失效应力下的多物理场演化过程和失效机理进行全面而深入的研究, 以突破其可靠性瓶颈。

一些研究工作已将 SiC MOSFET 的失效模式总结为两类: 栅失效和热失效。其中, 栅失效通常是由于栅氧附近电场过大, 引发大量载流子注入使得栅氧结构发生不可恢复损伤[3]; 而热失效是由于器件结温在高压大电流下迅速升高, 使得器件源极金属铝融化或本征激发占据主导引发热烧毁[4]。因此在 SiC MOSFET

的短路研究中，多关注其中的电-热场及其演化过程。然而，由于 SiC 材料独特的机械性能，SiC 基器件内的机械应力会显著高于 Si 基器件。一些研究发现，短路过程中机械应力可能诱发介质层裂纹并直接导致器件失效，尤其是栅极相关区域[5]。因此不同的栅极结构不可避免会对器件内部电热力演化过程产生显著的影响。

因此，本文选取了 SiC MOSFET 中三种典型的平面栅极结构，综合研究了栅极几何构型对短路瞬态过程中电场-温度-机械应力演化行为的影响。

2 研究方法

2.1 研究对象

本次研究选取的三种平面栅极结构分别为：传统栅极、倒角栅极、阶梯栅极，如图 1 所示。其中，相较于传统栅极结构，阶梯栅极在 JFET 区上方淀积了厚场氧以降低该区域氧化层电场并降低密勒电容 C_{gd} 以提升开关速度。倒角栅极则通过重构层间介质几何形貌消除金属-介质界面拐角，但同时导致多晶硅栅与源极金属间距缩减。除上述差异外，其余区域结构完全相同。

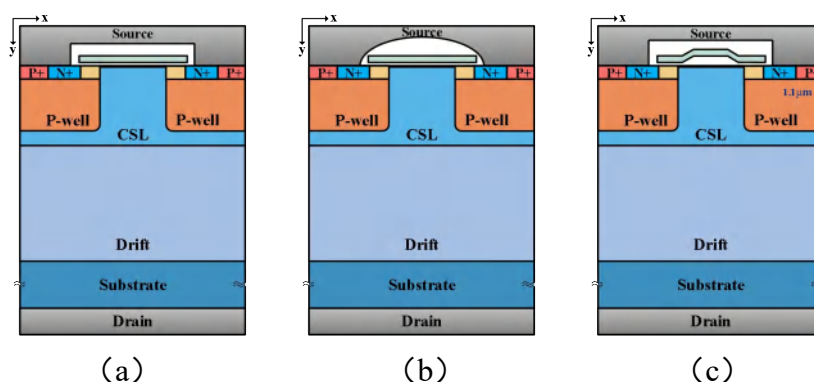


图1 具有不同平面栅极结构的SiC MOSFET。(a) 传统栅极MOSFET (b) 倒角栅极MOSFET (c) 阶梯栅极MOSFET

2.2 研究过程

为探究三种栅极结构的 SiC MOSFET 在短路瞬态应力下的电场-温度-机械应力的演变过程，首先采用了 Sentaurus TCAD 仿真软件对传统栅极 MOSFET、倒角栅极 MOSFET、阶梯栅极 MOSFET 进行了模型构建，如图 2 所示。然后采用电-热-力顺序耦合策略，分别对电场-温度和温度-机械应力两个演化阶段进行了仿真分析。在母线电压 400 V、短路持续时间 10 μ s 工况下，对三种器件进行短路瞬态仿真。在极高电场作用下，器件内大量载流子由源极流向漏极，短路峰值电流迅速升高至 400 A 左右。进一步在高压大电流作用下，器件内瞬时功耗迅速增加，结温随之在 10 μ s 内激增至 1300 K 左右，如图 3 所示。

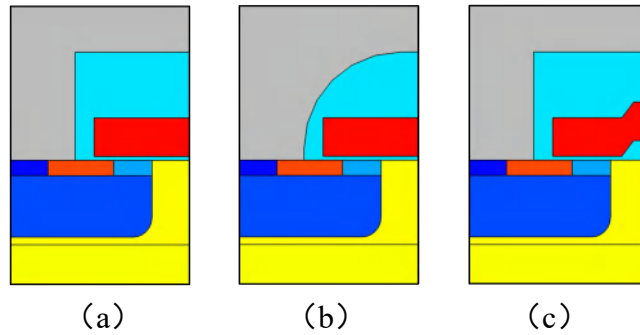


图 2 不同栅极结构建模。(a) 传统栅极MOSFET (b) 倒角栅极MOSFET (c)

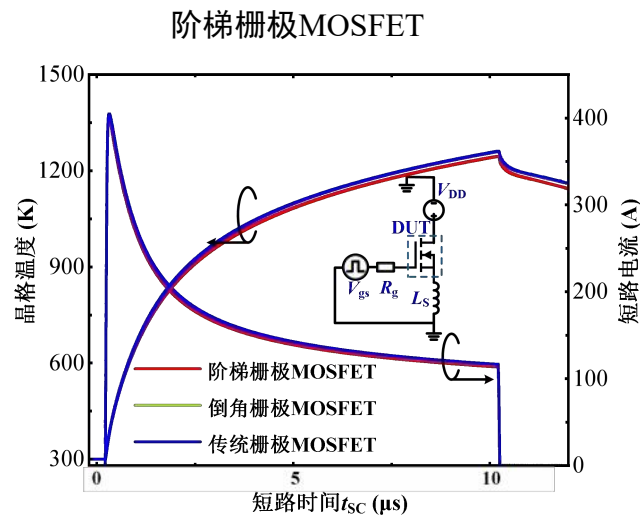


图 3 不同栅极结构短路电流和温度

3 结果与讨论

3.1 结果

为解析不同栅极结构对器件电-热-力特性的影响，本研究提取短路开始 $6 \mu\text{s}$ 后器件内部电场、温度、机械应力二维分布进行对比分析。图 4 展示了不同栅极结构 MOSFET 内的电场分布，相比其他两种结构，阶梯栅极 MOSFET 在 JFET 区上方的电场强度得到了有效降低，该局域处电场峰值仅为 1.6 MV/cm ，较传统栅极 MOSFET 与倒角栅极 MOSFET 的 3.0 MV/cm 电场峰值降低约 50%。另外值得注意的是，倒角栅极 MOSFET 由于源极金属与栅极多晶硅之间的距离减小，多晶硅左上方层间介质出的电场由 1.2 MV/cm 显著增加至 1.7 MV/cm ，表明几何形貌优化可能引发电场局部恶化风险。

图 5 展示了不同栅极结构 MOSFET 内的温度分布。由于未影响半导体内载流子输运过程，传统栅极 MOSFET 和倒角栅极 MOSFET 中温度分布基本相同，峰值温度均为 1135 K 。而阶梯栅极 MOSFET 峰值温度为 1121 K ，略有下降。这一差异源自阶梯栅极 MOSFET 中 JFET 下方积累层电子浓度下降，导致相同母线电压下，短路等效电阻增加，短路电流略有下降。因此，阶梯栅极 MOSFET

的晶格温度相对较低,这也与图 3 中的短路电流和晶格温度随时间演变的结果保持一致。

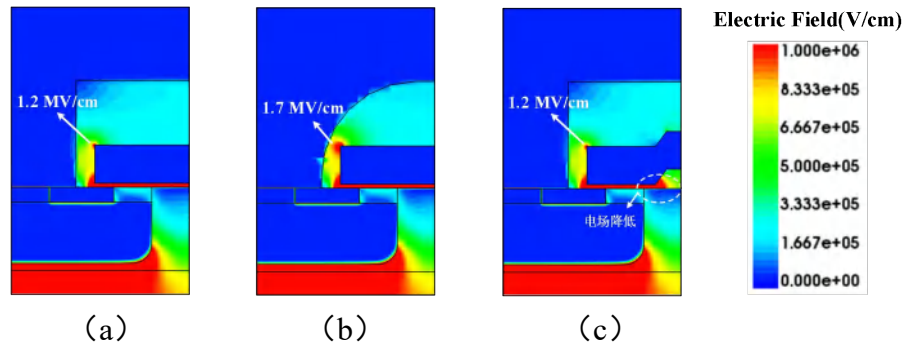


图 4 不同栅极结构的在短路开始6 μs后的电场分布。(a) 传统栅极MOSFET (b) 倒角栅极MOSFET (c) 阶梯栅极MOSFET

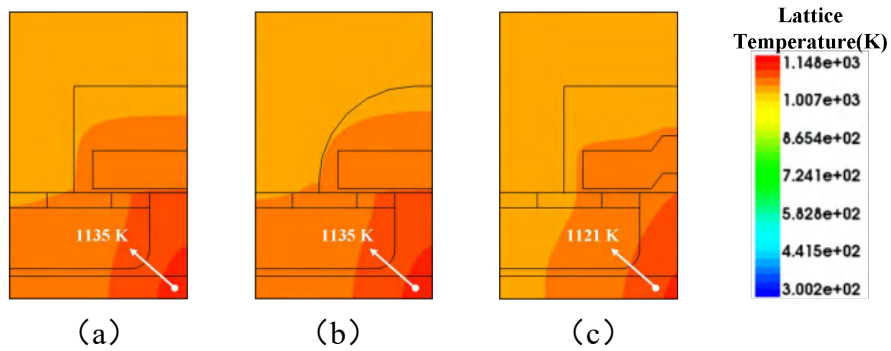


图 5 不同栅极结构的在短路开始6 μs后的温度分布。(a) 传统栅极MOSFET (b) 倒角栅极MOSFET (c) 阶梯栅极MOSFET

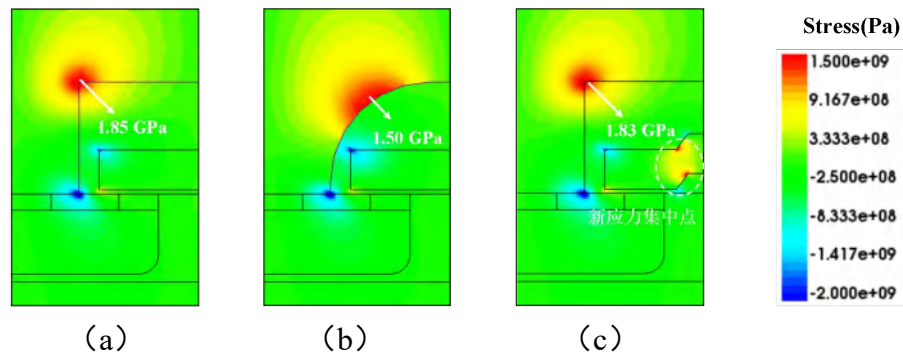


图 6 不同栅极结构的在短路开始6 μs后的机械应力场分布。(a) 传统栅极 MOSFET (b) 倒角栅极MOSFET (c) MOSFET

图 6 展示了不同栅极结构 MOSFET 中的机械应力分布。相较于传统栅极 MOSFET, 倒角栅极 MOSFET 通过消除层间介质处的几何突变, 使其此处机械应力峰值显著下降 19% (由 1.85 GPa 下降至 1.5 GPa)。然而, 由于源极金属和层间介质之间的接触面积的增加, 倒角栅极 MOSFET 在该处的应力集中区域面

积也随之增加。另外值得注意的是,阶梯栅极 MOSFET 内更低的温度导致层间介质处机械应力峰值略有下降。但在其多晶硅拐角处,层间介质和栅极场氧间的距离缩减又引入了新的应力集中点,在短路瞬态应力下可能成为新的失效风险点。

3.2 讨论

本文对比了三种平面栅极结构 SiC MOSFET 在短路瞬态应力下的电场、温度和机械应力的分布,重点关注其对电-热-力特性的差异。首先在电场方面,阶梯栅极 MOSFET 有利于缓解栅氧处电场,而倒角栅极 MOSFET 显著提高了层间介质处的电场峰值。在温度方面,不同结构间温度分布差异较小,仅阶梯栅极 MOSFET 因其积累层电阻的增大使得短路电流和温度略有降低。在机械应力方面,倒角栅极 MOSFET 对于应力峰值的下降效果最优,而阶梯栅极 MOSFET 则因引入了新应力集中点存在失效风险。研究表明,对于 SiC MOSFET 栅极设计需应当综合考虑多物理场的影响,进行折中设计和协同优化。

致谢

本文受国家自然科学基金项目(基金号:62434002)支持。

参考文献

- [1] 汪春艳. “双碳”目标下新型能源电力系统的关键技术和发展前景[J]. 通讯世界, 2024, 31(09): 91-93.
- [2] 张波, 邓小川, 张有润, 等. 宽禁带半导体 SiC 功率器件发展现状及展望[J]. 中国电子科学研究院学报, 2009, 4(02): 111-118.
- [3] X. Li et al., “An in-depth investigation into short-circuit failure mechanisms of state-of-the-art 1200 V double trench SiC MOSFETs,” *IEEE Trans. Power Electron.*, pp. 1–8, 2024, doi: 10.1109/TPEL.2024.3431296.
- [4] Deng X, Li X, Li X, et al. Short-Circuit Capability Prediction and Failure Mode of Asymmetric and Double Trench SiC MOSFETs[J]. *IEEE Transactions on Power Electronics*, 2021, 36(7): 8300-8307.
- [5] J. Liu, G. Zhang, B. Wang, W. Li, and J. Wang, “Gate failure physics of SiC MOSFETs under short-circuit stress,” *IEEE Electron Device Lett.*, vol. 41, no. 1, Art. no. 1, Jan. 2020, doi: 10.1109/LED.2019.2953235.

基于单粒子辐照损伤的碳化硅 MOSFET 栅极 泄漏电流退化模型研究

李春铨, 徐涛, 曾行, 严栩柯, 李轩, 邓小川*

电子科技大学 集成电路科学与工程学院, 成都 611731

(Email*: xcdeng@uestc.edu.cn)

摘要: 针对空间单粒子效应引发的 SiC MOSFET 栅极泄漏电流退化问题, 本文提出了一种基于单粒子辐照损伤的器件栅极泄漏电流退化物理模型。通过 TCAD 仿真模拟并修正单粒子辐照下 SiC MOSFET 的栅氧化层电场分布, 基于不同线性能量传输 (LET) 及偏置电压 (V_{DS}) 下器件栅极泄漏电流实验数据, 并结合 Fowler-Nordheim 隧穿理论, 建立了相应辐照条件下器件栅极泄漏电流退化模型。与实验测试结果的对比表明, 该模型能有效表征不同 LET 辐照下栅极泄漏电流随偏置电压的变化规律。

关键词: 碳化硅; 单粒子效应; 栅氧化层退化; 泄漏电流模型

1 前言

碳化硅 (SiC) MOSFET 具有高击穿电压、低导通损耗和高开关频率等优良特性, 为高效高功率密度航空航天电源提供了关键支撑^[1], 但是空间辐射环境中的单粒子效应对器件可靠性构成了严峻挑战^[2]。高能离子入射器件时, 会诱发瞬态电荷聚集导致局部电场突变, 最终使泄漏电流异常增大, 严重威胁器件的长期稳定性^[3]。为了有效评估器件的抗辐照能力并保障其可靠运行, 亟需对辐照引发的泄漏电流进行精确量化。因此, 建立能够准确表征辐照损伤下泄漏电流的物理模型至关重要。

本文通过 TCAD 仿真修正了高能离子诱发的空穴聚集效应在栅氧化层内产生的附加电场, 并基于 Fowler-Nordheim 隧穿理论, 提出了一种评估器件在单粒子效应下栅极泄漏电流退化的物理模型。该模型通过考虑辐照导致的局部电场强度, 实现了基于偏置电压及 LET 值等关键参数的栅极泄漏电流定量分析。

2 模型建立

2.1 Fowler-Nordheim 隧穿理论

SiC MOSFET 栅氧化层的厚度、介电常数及界面质量是决定器件栅氧化层瞬态电场的关键参数, 直接影响了器件的抗辐照能力^[4]。当高能离子轰击 SiC MOSFET 栅氧化层时, 会诱发瞬态电荷堆积与缺陷生成, 导致局部电场增强, 进而使栅极泄漏电流增大^[5]。

Fowler-Nordheim 隧穿理论建立在强电场下电子隧穿三角形势垒的量子力学机制, 其经典表达式如下:

$$I = AE_{ox}^2 \cdot \exp\left(-\frac{B}{E_{ox}}\right) \quad (1)$$

其中, I 为栅极泄漏电流, E_{ox} 为栅氧化层电场, 而参数 A 、 B 与器件几何尺寸及势垒高度有关。

2.2 单粒子效应下栅氧化层电场修正

图 1 为基于 Silvaco TCAD 平台构建的 SiC MOSFET 二维器件结构。图 2 展示了 $^{131}\text{Xe}^{+35}$ 重离子入射器件后, 在 $V_{DS}=300\text{V}$ 的偏置条件下, 栅氧化层电场达到峰值时刻的器件内部电场分布。

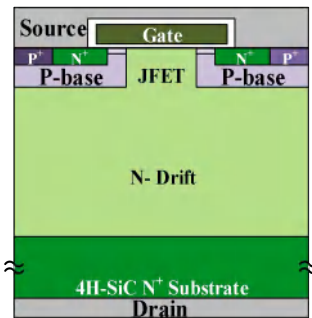


图 1 平面型 SiC MOSFET 器件结构

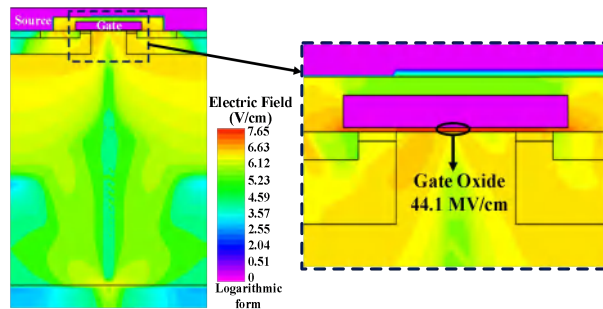


图 2 $^{131}\text{Xe}^{+35}$ 入射器件后在 20ps 时刻栅氧化层的电场分布 ($V_{DS}=300\text{V}$)

当高能离子穿过栅氧化层时, 其径迹电离效应会导致电子空穴对倍增, 形成柱状电离区。由于空穴迁移率低且易被栅氧化层陷阱俘获, 导致电离区内空穴聚集和局部电场突变。表达式 (2) 可用来表征电离区内空穴聚集产生的栅氧化层电场 E_{ox_hole} ^[6]:

$$E_{ox_hole} = \frac{1}{2\epsilon_{SiO_2}\epsilon_0} \int_0^{l_{ox}} dz \int_0^{+\infty} qP_s \frac{zr}{(z^2+r^2)^{\frac{3}{2}}} dr \quad (2)$$

其中 ϵ_{SiO_2} 是二氧化硅的相对介电常数, ϵ_0 是真空介电常数, z 和 r 分别表示圆柱区域的高度和半径, 函数 $P_{S(r,t)}$ 表示在半径 r 和时间 t 下未复合的空穴浓度。

通过求解表达式 (2) 获得约 7MV/cm 的重离子辐照下空穴聚集产生的栅氧化层电场, 将该附加值与 TCAD 仿真提取的平均电场叠加, 从而获得辐照条件下器件栅氧化层内的实际电场分布。

2.3 模型建立

Corinna Martinella 等人报道了不同重离子辐照及 V_{DS} 偏置条件下, SiC MOSFET 栅极泄漏电流的实验测试结果^[7], 其测试条件如表 1 所示。

表 1 单粒子辐照测试条件

辐照条件	对应的 V_{DS} 值(V)
$^{131}\text{Xe}^{+35}$ @LET=62.39(MeV·cm ² /mg)	120 150 170 200 250 300 350 400
$^{56}\text{Fe}^{+15}$ @LET=20.05(MeV·cm ² /mg)	300 340 370 400 420

基于修正后的栅氧化层实际电场分布, 应用 Fowler–Nordheim 隧穿理论建立重离子辐照下 SiC MOSFET 栅极泄漏电流退化模型。通过对表达式 (1) 取对数变换, 得到方程 (3):

$$\ln \frac{I}{E_{ox}^2} = \ln A - \frac{B}{E_{ox}} \quad (3)$$

将栅极泄漏电流实验数据与修正后的电场代入方程 (3), 利用最小二乘法进行线性回归分析, 对不同 LET 下的材料参数 A 与 B 进行拟合, 并将拟合参数代入方程 (3) 就可以获得不同 E_{ox} 下的栅极泄漏电流。

3 结果分析和讨论

图 3 和图 4 分别展示了两种重离子辐照下栅极泄漏电流的实验数据与模型预测结果对比。图中实心点表示测量值, 红色实线为本模型拟合曲线。图 3 表明在高 LET 辐照下, 模型预测值与实验数据的拟合程度较高, 决定系数 R^2 达 0.945, 验证了模型的有效性。而图 4 中低 LET 辐照的预测曲线虽能反映实验数据变化趋势, 但决定系数 R^2 较低 (约 0.779), 表明低 LET 辐照下, 诸如缺陷分布等因素的影响仍需深入研究。

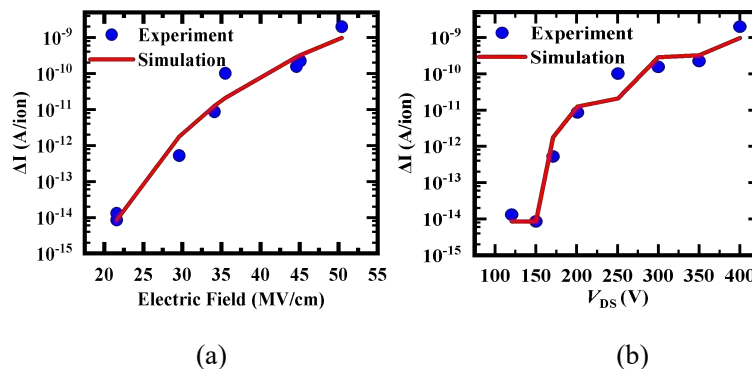


图 3 $^{131}\text{Xe}^{+35}$ 辐照后栅极泄漏电流拟合。(a) 与 E_{ox} 关系; (b) 与 V_{DS} 关系

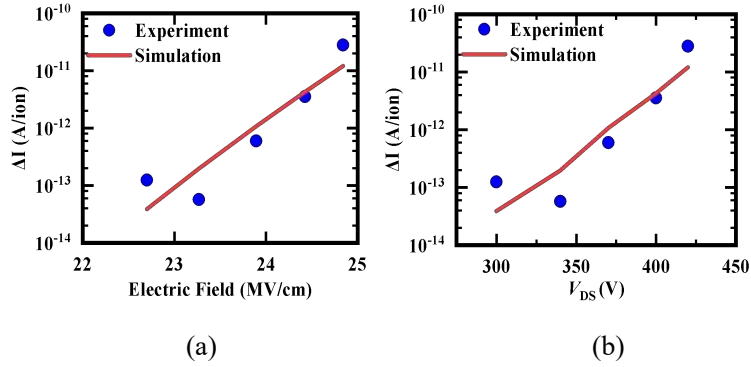


图4 $^{56}\text{Fe}^{+15}$ 辐照后栅极泄漏电流拟合。(a)与 E_{ox} 关系；(b)与 V_{ds} 关系

综合对比不同辐照条件的拟合曲线，表明该模型对高 LET 辐照场景的泄漏电流预测精度显著较优，而低 LET 条件下虽然决定系数较低，但仍然有较好的拟合程度，表明修正电场的引入是提升模型预测精度的关键。

4 结论

本文通过结合实验数据与栅氧化层修正电场，建立了基于 Fowler-Nordheim 隧穿理论的 SiC MOSFET 单粒子辐照下栅极泄漏电流退化物理模型。该模型通过量化辐照导致的栅氧化层电场，实现了对不同重离子辐照条件及偏置电压下栅极泄漏电流的定量计算。分析表明，模型决定系数 R^2 较高，为宇航用功率器件的抗辐照加固设计、屏蔽方案优化及可靠性评估提供了理论工具与仿真依据。

参考文献

- [1] Yuan X, Laird I, Walder S. Opportunities, Challenges, and Potential Solutions in the Application of Fast-Switching SiC Power Devices and Converters[J]. IEEE Transactions on Power Electronics, 2021, 36(4): 3925–3945.
- [2] Mizuta, Eiichi, Satoshi Kuboyama, Hiroshi Abe, Yoshiyuki Iwata and Takashi Tamura. Investigation of Single-Event Damages on Silicon Carbide (SiC) Power MOSFETs[J]. IEEE Transactions on Nuclear Science, 2014, vol. 61, no. 4: pp. 1924–1928.
- [3] Yan, X.Y., Z. He, Q.Y. Chen, P.P. Hu, S. Gao, S.W. Zhao, C. Cai. An Investigation of Angle Effect on Heavy Ion Induced Single Event Effect in SiC MOSFET[J]. Microelectronics Reliability, 2022, Volume 138.
- [4] Zhou, Jian-Cheng, Ying Wang, Xing-Ji Li, Jian-Qun Yang, Meng-Tian Bao and Fei Cao. Simulation Study of Single-Event Effects for the 4h-Sic Vdmosfet with Ultralow on-Resistance. IEEE Transactions on Electron Devices, 2022, vol. 69, no. 6: pp. 3283-3289.
- [5] engupta A, Ball D R, Witulski A F, et al. Impact of Heavy-Ion Range on Single-Event Effects in Silicon Carbide Power Junction Barrier Schottky Diodes[J]. IEEE Transactions on Nuclear Science, 2023, 70(4): 394–400.
- [6] C. Abbate, G. Busatto, D. Tedesco, A. Sanseverino, F. Velardi, and J. Wyss, Gate damages induced in SiC power MOSFETs during heavy ion irradiation—Part II[J]. IEEE Transactions on Electron Devices, 2019, vol. 66, no. 10, pp. 4243–4250.
- [7] Martinella, C., R. Stark, T. Ziemann, R. G. Alia, Y. Kadi, U. Grossner and A. Javanainen. Current Transport Mechanism for Heavy-Ion Degraded SiC MOSFETs[J]. IEEE Transactions on Nuclear Science, 2019, 66(7) : 1702-1709.

扩流型 SiC MOSFET 动态电阻的机理研究

宋怡慧, 汪志刚

西南交通大学, 集成电路科学与工程学院

项目编号: 2025-YF11-00007-HZ, 2024-YF08-00041-GX

摘要: 本文提出了一种新型扩流型 SiC MOSFET 结构, 其核心创新在于引入阶梯形 P-Well, 构建多级复合 JFET 区, 实现了电流横向扩展与空间电场分布的协同优化。仿真结果表明, 该结构在略微提升器件耐压 (从 1450 V 提高至 1480 V) 的同时, 将比导通电阻从传统 SiC MOSFET 的 $3.2 \text{ m}\Omega \cdot \text{cm}^2$ 降低至 $1.7 \text{ m}\Omega \cdot \text{cm}^2$, 降幅达 46.88%。动态特性分析进一步验证了该结构在降低动态导通电阻、抑制栅极下方缺陷俘获及提升开关速度等方面的优势, 为高性能 SiC MOSFET 设计提供了新的思路与参考。

关键词: 动态电阻; 电子俘获; 碳化硅晶体管

1 前言

碳化硅 (Silicon Carbide, SiC) 作为典型的宽禁带半导体, 25°C 下击穿电场强度可达 2.5 MV/cm , 是硅 (Silicon, Si) 的 10 倍^[1], 且导热系数高、饱和电子漂移速度快、禁带宽度大^[2-4]。因此, SiC 功率器件在高温、高功率场景下相比传统硅器件更具优势, 目前已在工业电机驱动、光伏逆变器、新能源汽车等电力电子领域得到广泛应用^[5-7]。

尽管 SiC MOSFET 已实现商业化, 传统方形元胞结构依然存在限制: 单一垂直电流路径需穿过高阻 JFET 区, 且为防止相邻 P 基区耗尽层连通, JFET 区通常被迫加宽, 导致导通电阻升高、功率密度受限^[8-10]。如何在保证耐压能力的同时优化电流路径、提高元胞紧凑度, 已成为制约 SiC MOSFET 性能进一步提升的关键问题。

针对这一问题, 本文提出了电流扩展型 SiC MOSFET (以下简称扩流型 MOS), 采用阶梯式 P-Well 结构结合外侧注入, 形成横向电流扩散路径, 绕过高阻 JFET 区, 实现更低阻的电流导通和更高效的电流注入。基于 Sentaurus TCAD 仿真结果, 所提出结构在不影响阻断特性的前提下, 显著降低了动态导通电阻, 并有效改善了比导通电阻和栅漏电容, 综合性能更优, 适用于更高功率密度和更苛刻的开关应用。

2 结构和性能

图 1 给出了 SiC MOS 传统结构、二阶扩展型 MOS 结构以及三阶扩流型 MOS 结构, 其主要区别在于 JFET 区是否存在电流扩展功能。具有扩流功能的 JFET

区是在阶梯状 P-Well 外，引入相对高的 N 型离子注入区，实现扩流功能。不同于传统方形结型场效应晶体管（JFET）区，扩流型 MOS 的 JFET 区沿阶梯轮廓延展，形成反向阶梯结构，使电流可在阶梯内同时实现横向与纵向传输。阶梯数量可根据需求调节，综合两阶和三阶扩流型 MOS 结构特征，综合分析其静态特性，重点讨论了其动态电阻的物理机理。

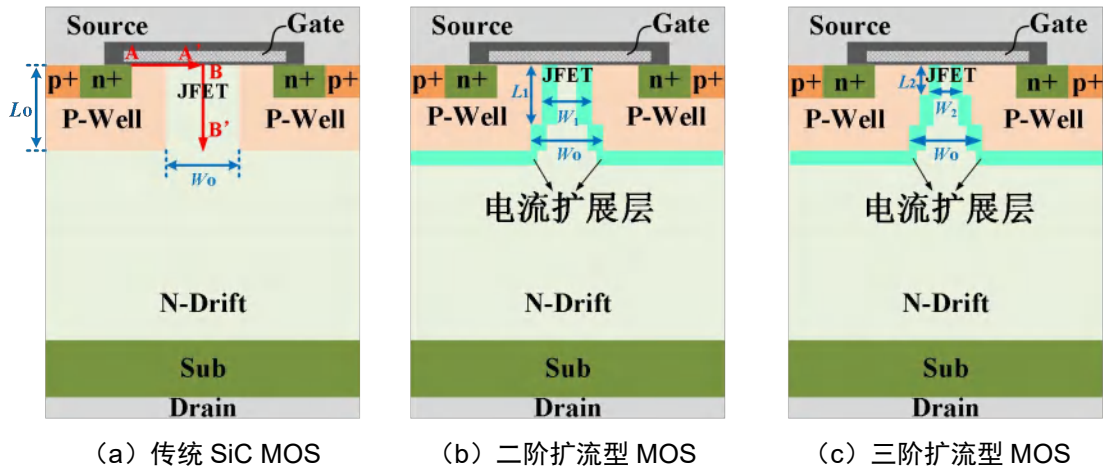


图 1 器件结构示意图

为了设计结果的准确性，此论文的相关仿真根据已发表文章进行了校准，Sentaurus TCAD 仿真中采用的 SiC 器件物理模型包括：不完全电离模型、迁移率模型、Shockley-Read-Hall（SRH）寿命模型、俄歇复合模型、碰撞电离模型、雪崩模型（Okuto）、各向异性模型、费米模型、高场速度饱和模型等多种物理模型。用于仿真的主要结构参数如表 1 所示。

表 1 用于仿真的主要结构参数

	传统 SiC MOS	二阶扩流型 MOS	三阶扩流型 MOS
单元间距 (μm)	5.2	5.2	5.2
JFET 区掺杂浓度 (cm^{-3})	1×10^{16}	1×10^{16}	1×10^{16}
P-Well 区掺杂浓度 (cm^{-3})	5×10^{17}	5×10^{17}	5×10^{17}
N-Drift 区掺杂浓度 (cm^{-3})	1×10^{16}	1×10^{16}	1×10^{16}
N-Drift 区长度 (μm)	10	10	10
L_0 (μm)	1	1	1
L_1 (μm)	-	0.4	0.4
L_2 (μm)	-	-	0.2
W_0 (μm)	1.0	1.0	1.0
W_1 (μm)	-	0.6	0.6
W_2 (μm)	-	-	0.2

器件的 $I-V$ 特性如图 2 所示。传统 SiC MOS、二阶扩流型 MOS 及三阶扩流型 MOS 的比导通电阻 ($R_{on,sp}$) 分别为 $3.2 \text{ m}\Omega/\text{cm}^2$ 、 $2.9 \text{ m}\Omega/\text{cm}^2$ 和 $1.7 \text{ m}\Omega/\text{cm}^2$ 。相较于传统结构，二阶扩流型 MOS 通过引入双层电流扩展区，改善了 JFET 区的电流引导和分流能力，使 $R_{on,sp}$ 降低了 9.38%，展现出良好的导通性能优化效果；在此基础上，三阶扩流型 MOS 进一步增加了额外的阶梯级数，形成更丰富的电流扩展通道和更均匀的电场分布，使得电子在 JFET 区内得以更充分地横向扩散，并高效导入 P-Well 区下方，极大减少了局部电流拥堵和低电离区的电阻。得益于这种多阶梯的电流整流和区域利用率提升，三阶扩流型 MOS 的 $R_{on,sp}$ 相比传统结构降低了 46.88%，导通性能得到显著增强。

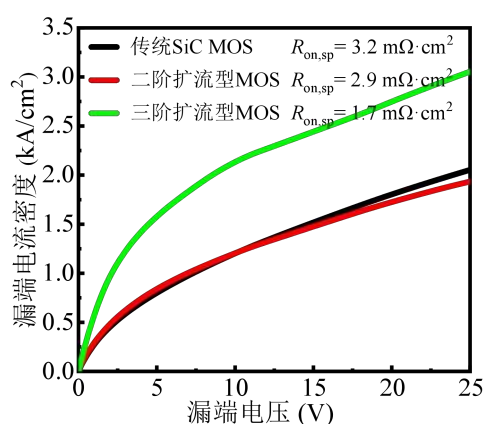


图 2 $I-V$ 特性

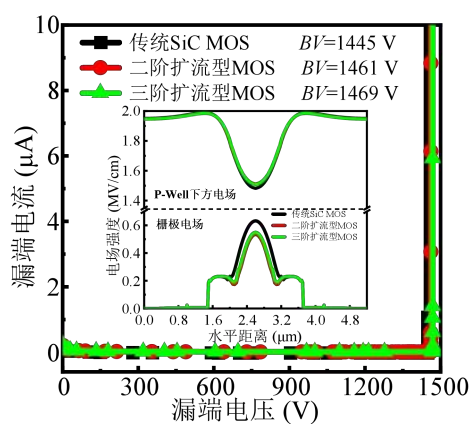


图 3 电压击穿特性

图 3 展示了三种器件的击穿特性曲线。传统 SiC MOS、二阶扩流型 MOS 及三阶扩流型 MOS 的击穿电压 (Breakdown Voltage, BV) 分别为 1445 V、1461 V 和 1469 V。通过对器件纵向剖面进行电场分布仿真，并截取 P-Well 下方及栅极氧化层界面处的局部电场强度，可以看出扩流型 MOS 的栅极区域电场峰值较传统结构有所降低。这表明引入阶梯式电流扩展层能够有效分散局部电场尖峰，从而在显著降低 $R_{on,sp}$ 的同时，依然保持甚至提升器件的 BV ，实现了导通性能与耐压能力的双重优化与平衡。

图 4 显示了传统 SiC MOS、二阶扩流型 MOS 和三阶扩流型 MOS 的栅极电荷 (Q_g) 特性仿真结果，重点对比了不同阶梯结构对栅极充放电行为的影响。结果表明，三种器件的米勒电荷平台 (Q_{gd}) 长度基本一致，均约为 $23.34 \text{ nC}/\text{cm}^2$ ，说明导通过程中由栅漏电容 (C_{gd}) 放电形成的米勒区间所需电荷量相近，阶梯式设计并未在该阶段引入额外电荷损耗。然而在米勒平台结束后的栅电压上升阶段，扩流型 MOS 的 Q_g 增长速率略微高于传统 SiC MOS，表明多阶梯结构在有效降低 $R_{on,sp}$ 的同时，对栅极驱动能力提出了更高要求，尤其在快速开通时需要更大的电荷注入速率，这对高频开关性能具有一定影响。进一步比较可知，在

$V_d=2\text{ V}$ 条件下, 传统 SiC MOS、二阶扩流型 MOS 和三阶扩流型 MOS 的 C_{gd} 分别为 262.31 pF/cm^2 、 155.56 pF/cm^2 和 88.79 pF/cm^2 , 表明二阶与三阶扩流型 MOS 的电容特性均优于传统结构, 且三阶扩流型 MOS 表现最优, 兼具更低的栅漏电容和更强的载流子调控能力。

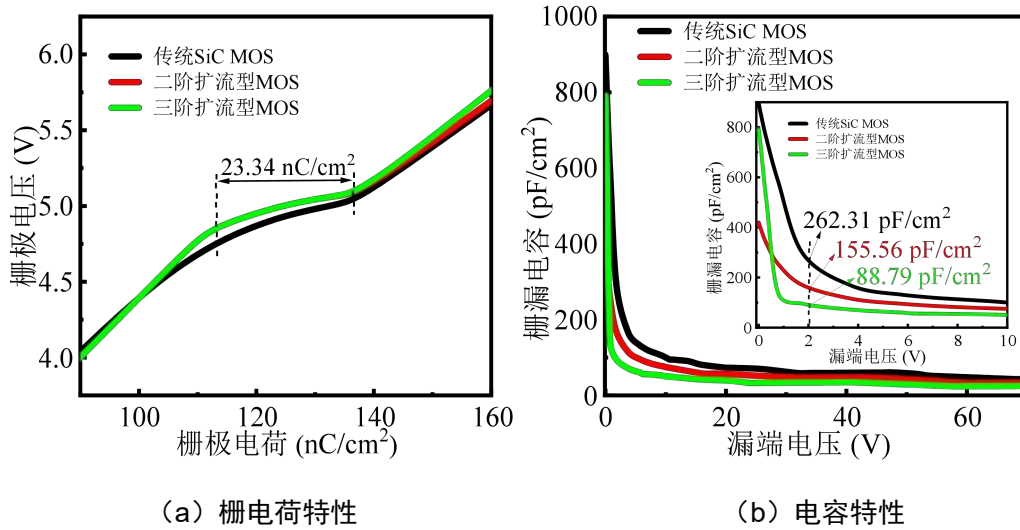


图 4 栅电荷特性与电容特性

为深入分析 JFET 区结构动态特性, 以优化开关性能, 本文开展了双脉冲仿真试验: 在 $3\text{ }\mu\text{s}$ 时对器件关断, $12\text{ }\mu\text{s}$ 时重新开启, 开关阶段的电流电压波形如图 5 所示。仿真结果表明, 阶梯型结构在开关特性上与传统 MOS 呈现明显差异: 关断阶段, 扩流型 MOS 的关断速度得到显著提升, 且阶梯数越多效果越优, 三阶扩流型 MOS 表现最突出; 而在开启阶段, 二阶扩流型 MOS 和三阶扩流型 MOS 的开启速度均略低于传统 MOS。其核心机制在于, 尽管扩流型 MOS 保持了与传统器件相同的 Q_{gd} , 但通过优化的 P-Well 几何结构和额外的离子注入工艺, 有效降低了等效 C_{gd} , 并在关断瞬态提供了更强的电荷分流效应, 缩短了米勒平台后的电荷抽取时间, 从而加快了栅极电压上升速率, 实现更快关断。相比之下, 开启阶段的轻微延迟则主要来自扩流型 MOS 的阶梯结构引入的多层载流子通道, 导致开启初期电流聚集路径相对延长, 尽管三阶扩流型 MOS 在米勒平台后的栅电荷上升速率 (dQ_g/dt) 最高, 但复杂的电流引导路径在通道完全形成前存在瞬时瓶颈, 略微抑制了导通初始的响应速度。这一结果说明, 扩流型 MOS 通过优化电荷调控显著提升关断性能的同时, 在导通动态中需要平衡路径复杂性与载流子注入效率。

在对漏极和栅极施加图 6 (a) 所示电压波形的条件下, 仿真比较了三种器件的动态电阻特性, 结果如图 6 (b) 和 (c) 所示。与传统结构相比, 扩流型 MOS 在动态导通电阻方面表现出显著优势, 且随阶梯级数增加, 性能进一步提

升，其中三阶扩流型 MOS 效果最优。这一性能提升得益于阶梯式 P-Well 结构对 JFET 区上方栅氧层的部分覆盖，有效屏蔽了高缺陷区对载流子的俘获，并通过电流扩展层优化电流分布，将电流从栅极下方高陷阱密度区域引导至 P-Well 下方低阻区，降低陷阱活化水平。同时，阶梯形几何设计使 JFET 区电场分布更均匀、俘获截面更小，有效抑制了动态导通过程中的陷阱填充和释放效应。

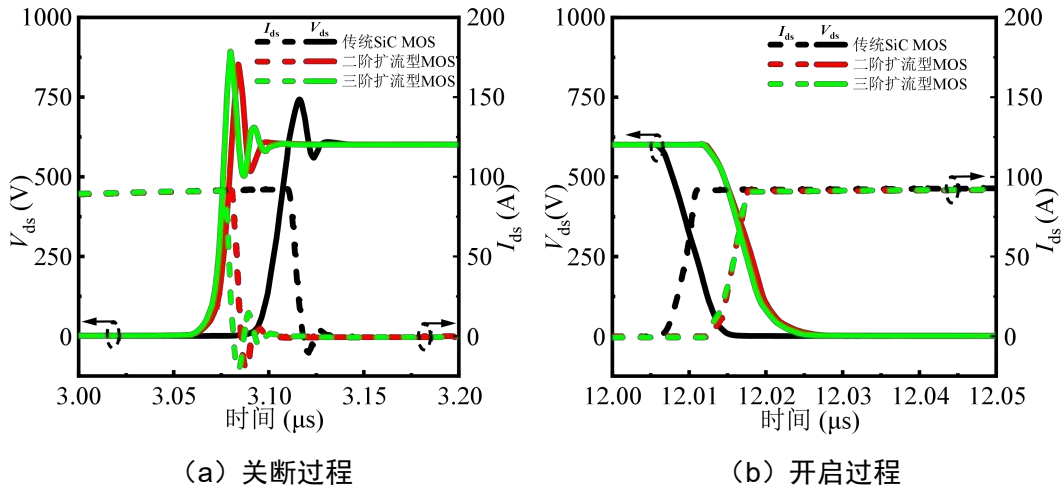
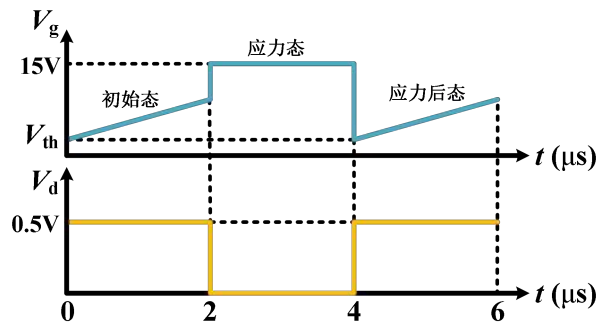
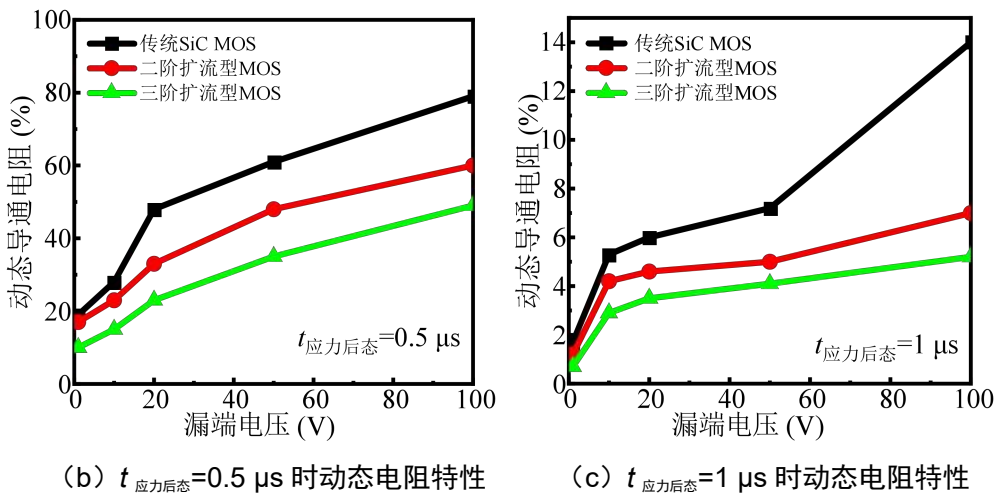


图 5 双脉冲仿真电流电压图



(a) 栅极和漏极所加电压



(b) $t_{\text{应力后态}}=0.5 \mu\text{s}$ 时动态电阻特性

(c) $t_{\text{应力后态}}=1 \mu\text{s}$ 时动态电阻特性

图 6 动态电阻仿真

在 SiC 材料中设置了 $1.5 \times 10^8 \text{ cm}^{-3}$ 的受主型缺陷，氧化层中添加 $1.5 \times 10^{12} \text{ cm}^{-3}$ 的受主型缺陷，并在 SiC/SiO₂ 界面处引入 $1 \times 10^{11} \text{ cm}^{-2}$ 的界面受主缺陷，同时通过调整电子与空穴的俘获截面，分析不同结构下的缺陷占据特性。在施加 15 V 栅极电压条件下对器件进行充电，图 7 和图 8 展示了栅极下方区域在不同俘获截面条件下的受主缺陷占据浓度分布。结果表明，扩流型 MOS 由于引入了阶梯式 JFET 区域，局部俘获截面相对更小，可有效降低载流子被陷阱俘获的概率。随着俘获截面的增大，栅极下方的缺陷占据浓度明显升高，在相同条件下，阶梯扩流型 MOS 对缺陷占据的抑制效果更为显著，特别是三阶扩流型 MOS 在 JFET 区被阶梯型 P-Well 覆盖的区域内，受主缺陷活化水平最低，有助于动态导通电阻降低并增强器件可靠性。

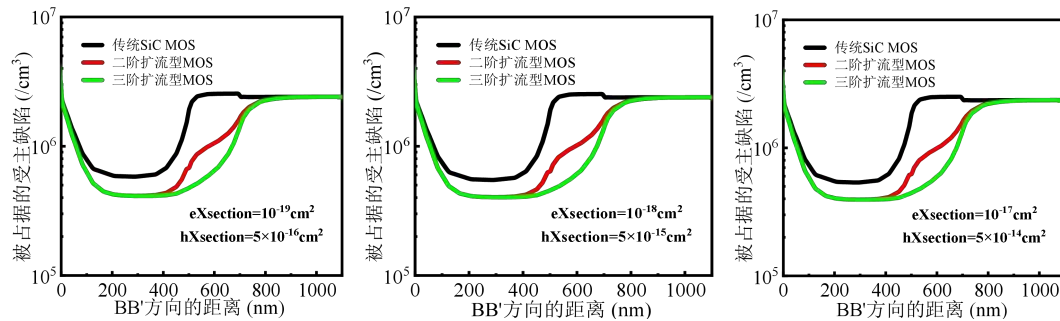


图 7 不同俘获截面的栅极下方缺陷

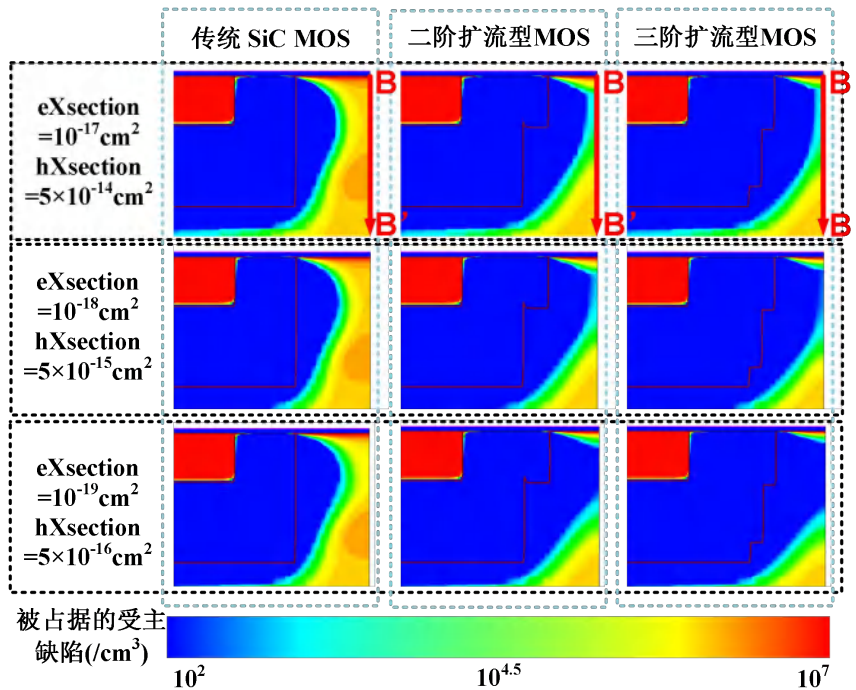


图 8 栅极下方区域被占据的受主缺陷密度分布

为实现扩流型 MOS 的成功流片，设计了一套针对性工艺流程。该流程在沿用传统 MOS 制备工艺的基础上，无需额外增加光刻掩模层，显著降低了制造成

本，同时避免了工艺复杂度的增加。如图 9 所示，以三阶扩流型 MOS 为例，其核心制备步骤包括：(a) 在 N⁺型衬底上外延生长漂移区；(b) 进行第一次 P-Well 区离子注入；(c) 使用尺寸略小于首次 P-Well 区的光刻掩模，执行第二次 P-Well 区离子注入，并依靠离子的横向扩散形成功能扩展层；(d) 重复使用尺寸进一步缩小的光刻掩模进行第三次 P-Well 区离子注入，形成多级阶梯式 P-Well 结构；(e) 通过离子注入形成包围阶梯型 P-Well 的 JFET 区及其高掺杂区；(f) 分别完成 N⁺区和 P⁺区的注入；(g) 通过热氧化生长栅氧化层，并沉积多晶硅以构建栅极；(h) 淀积层间介质 (ILD)；(i) 最后通过金属化工工艺形成源极与漏极的欧姆接触。

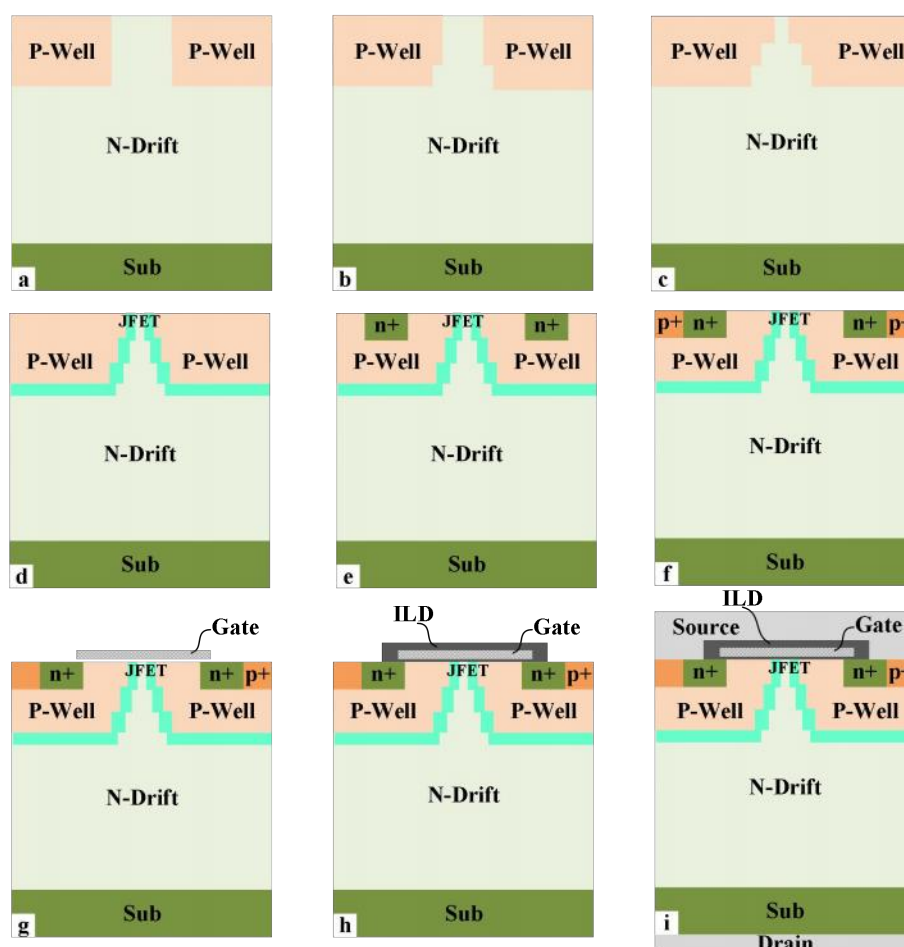


图 9 工艺流程图

3 结论

本文提出的扩流型 MOS 结构通过在 JFET 区引入阶梯式 P-Well，有效引导电流横向扩展，提升了 P-Well 下方的电流利用率，并压缩了 JFET 窄区电阻。仿真结果表明，该结构在略微提升击穿电压的同时，将导通电阻最大降低 46.88%，其中动态导通电阻的改善尤为显著，主要归因于载流子在导通过程中的分布更均匀，电流拥堵得到有效抑制，同时器件的寄生电容得到协同优化，减小了充放电

损耗。相比传统结构，该设计在高压大功率条件下，可显著降低开通过损耗并提升开关速度。

参考文献

- [1] Chen Y, et al. Investigation of threshold voltage instability of SiC MOSFETs under different gate voltage sequences[J]. IEEE Transactions on Electron Devices, 2024, 71(4): 2536-2542.
- [2] Lyu G, Ali H, Tan H, et al. Review on Short-Circuit Protection Methods for SiC MOSFETs[J]. Energies, 2024, 17(17): 4523.
- [3] Nel B, Perinpanayagam S. A brief overview of SiC MOSFET failure modes and design reliability[J]. Procedia CIRP, 2017, 59: 280-285.
- [4] Rødal G L, Pefitsis D. Gate-drive circuits for adaptive operation of SiC MOSFETs[J]. IEEE Transactions on Power Electronics, 2024, 39(7): 8162-8186.
- [5] Wang J, Jiang X. Review and analysis of SiC MOSFETs' ruggedness and reliability[J]. IET Power Electronics, 2020, 13(3): 445-455.
- [6] Wang J, et al. Characterization, modeling, and application of 10-kV SiC MOSFET[J]. IEEE Transactions on Electron Devices, 2008, 55(8): 1798-1806.
- [7] Xue P, Davari P. A temperature-dependent analytical transient model of SiC mosfet in half-bridge circuits[J]. IEEE Transactions on Power Electronics, 2024.
- [8] Alonso A R, Diaz M F, Lamar D G, et al. Switching performance comparison of the SiC JFET and SiC JFET/Si MOSFET cascode configuration[J]. IEEE transactions on power electronics, 2013, 29(5): 2428-2440.
- [9] Xu F, Han T J, Jiang D, et al. Development of a SiC JFET-based six-pack power module for a fully integrated inverter[J]. IEEE Transactions on Power Electronics, 2012, 28(3): 1464-1478.
- [10] Zhang M, Li H, Yang Z, et al. Short circuit protection of silicon carbide MOSFETs: Challenges, methods, and prospects[J]. IEEE Transactions on Power Electronics, 2024.

SiC 双沟槽 MOSFET 短路特性与失效机制分析

汪宇浩, 冯文洁, 赵佩芝, 付祥和, 张钧辉, 张圆圆, 贺永宁

西安交通大学微电子学院,

陕西省电子器件与高端芯片重点实验室, 西安 710049

摘要: 随着 Si 基功率器件性能已接近“硅极限”, 宽禁带半导体 SiC 功率 MOSFET 作为有望在中高压领域替代 Si IGBT, 其短路可靠性问题研究是目前的重点之一。本研究针对罗姆第 4 代 SiC 双沟槽 MOS 展开单脉冲短路测试, 并通过器件开封显微镜分析和短路建模仿真研究其短路失效机制, 结果表明器件短路失效是由漂移区发生热积累导致的, 为器件优化设计及可靠性提升提供了关键依据。

关键词: SiC 双沟槽 MOS; 短路失效分析; 单脉冲短路测试

1. 引言

随着半导体几十年的发展, Si 基功率器件的性能已经接近“硅极限”。半导体器件以不同的输出功率和输出频率已经衍生出(晶闸管、IGBT、MOSFET 等)各种结构, 并应用于不同领域^[1]。但受制于工艺制造难度带来的成本和可靠性问题, 以 SiC 为代表的第三代半导体材料飞速发展, 而 SiC MOSFET 凭借高频、高效、耐高温等优势, 在新能源汽车、储能系统等高压大功率领域应用广泛^{[2][3]}。短路可靠性是衡量器件极限工况下稳定性的核心指标, 直接影响电力电子系统的安全运行^[4]。

本研究针对商用罗姆第 4 代 1200V 62mΩ 车规级 SiC 双沟槽 MOS 器件^[5]展开短路可靠性的研究, 利用相应的短路测试平台对该器件进行了不同漏极电压下的单脉冲短路特性研究模拟器件在实际应用的工况, 并通过 Sentaurus TCAD 仿真软件对 SiC 双沟槽 MOS 进行仿真设计和器件短路瞬态过程的建模分析, 结合激光开封、光学显微等手段分析得出器件的短路退化机制。

2. 器件的单脉冲短路测试及短路失效分析

2.1 单脉冲短路测试

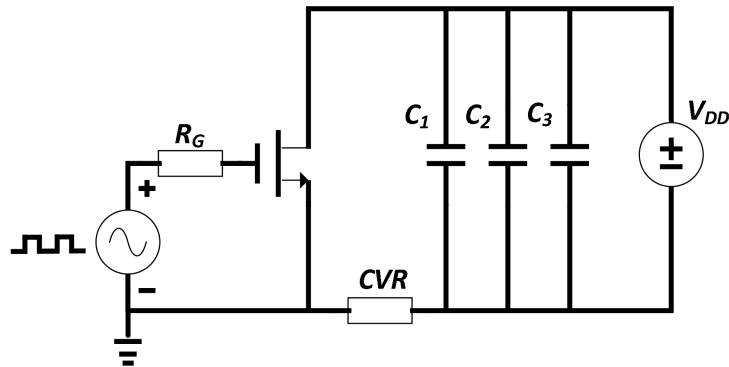


图 1 短路测试电路原理图

单脉冲短路测试的原理是通过在 SiC 双沟槽 MOSFET 导通状态下人为制造短路故障，以研究其动态响应和失效机制，原理图如图 1 所示。测试时，首先在漏极施加 400V-800V 的高电压，然后施加一个 -5V-18V 的栅极脉冲信号（脉冲宽度则是器件的短路时间）使器件导通和关断，导通时电流急剧上升，器件同时承受高电压和大电流，产生瞬时高功耗，引发结温骤升。通过监测短路电流波形和电压塌陷时间及热分布分析器件的短路耐受时间（SCWT）和失效机理（如热失控、栅极损伤或结构熔毁）[6]。

本研究在 25°C 常温条件下，对 SiC 双沟槽 MOSFET 器件进行了单脉冲短路特性测试，分别考察了 400V 和 800V 两种漏极电压下的动态响应特性。通过实时监测器件的短路电流波形和栅极电压变化，系统分析不同漏极偏置电压对器件短路行为的影响机制。

2.1.1 单脉冲短路测试结果

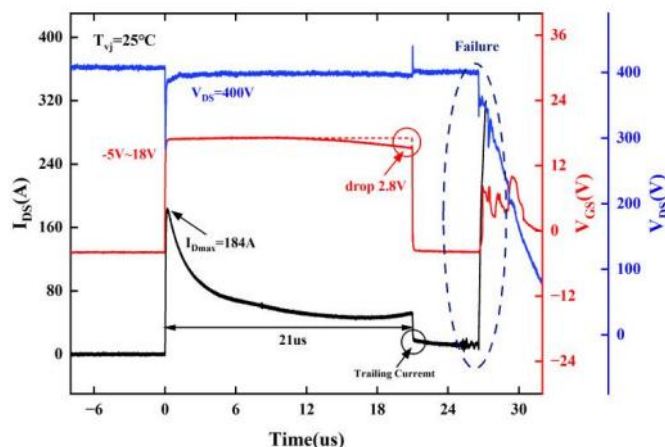


图 2 $V_{DS}=400V$ 时短路失效曲线

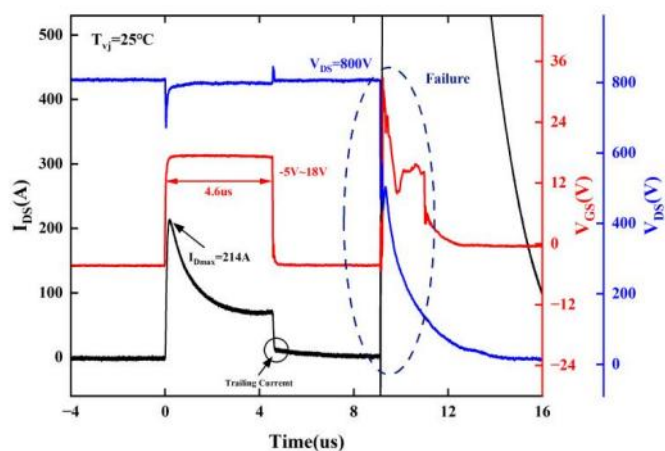


图 3 $V_{DS}=800V$ 时短路失效曲线

从图 2 和图 3 测试结果可以知道， $V_{DS}=400V$ 时器件的短路峰值电流为 184A、

SCWT 为 21 μ s, $V_{DS}=800$ 时器件的短路峰值电流为 214A、SCWT=4.6 μ s; 器件关断时有很明显的拖尾电流, 持续几微秒不能完全关断, 导致器件短路失效漏极电流骤升; $V_{DD}=400V$ 时器件随着短路时间的增加 V_{GS} 有一个明显的下降, 而 $V_{DD}=800V$ 时则维持稳定。

2.2 短路失效分析

2.2.1 短路失效物理分析

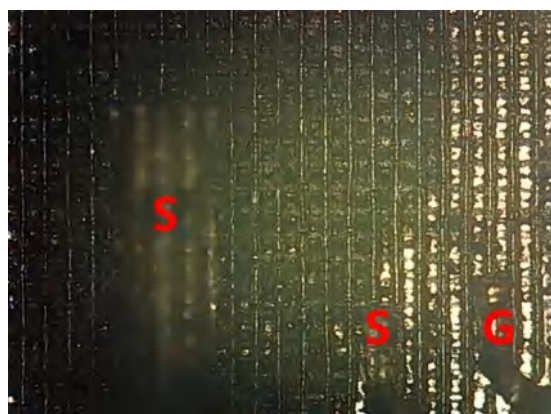


图 3 $V_{DS}=400V$ 失效器件开封表层

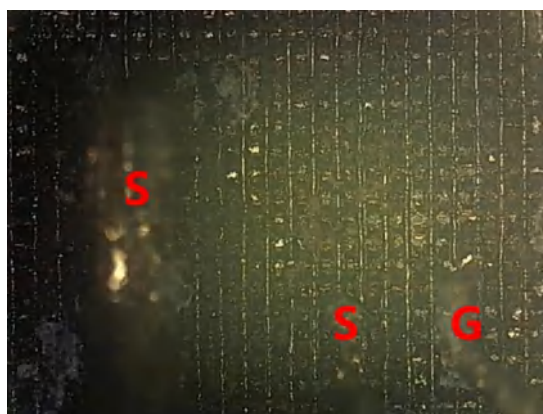


图 4 $V_{DS}=800V$ 失效器件开封表层

通过对失效器件进行激光开封及光学显微分析, 可以清晰观察到不同偏置电压下的损伤形貌特征: 在 400V 漏极电压条件下, 源极金属层呈现明显的熔融烧蚀现象, 而栅极金属结构保持相对完整; 当漏极电压升高至 800V 时, 源极与栅极金属层均发生严重烧蚀。因此, 可以推断, 高电流密度导致源漏极间形成局部热点并首先失效; 随着热量的持续积累, 热失控效应逐步扩展至栅极区域, 最终引发器件三端整体性热损毁。

2.2.2 短路失效仿真分析

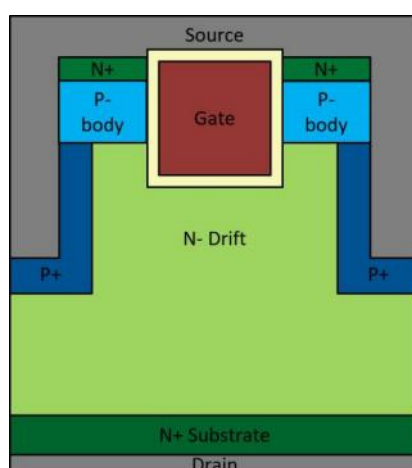


图 5 SiC 双沟槽 MOSFET 结构

用 Sentaurus TCAD 软件设计了一款 1200V, 20.6m Ω 的 SiC 双沟槽 MOS, 其结构如图 5 所示。通过软件自带的混合电路仿真功能搭建短路测试平台并进行仿真, 仿真器件的短路特性如图 6 所示, 当器件在短路时间 $T_{sc}=4\mu$ s 时尝试关断,

可以观察到明显的拖尾电流现象，这表明器件在该工作点下已失去正常的关断能力。所以，该器件的 SCWT 为 $4\mu\text{s}$ 。

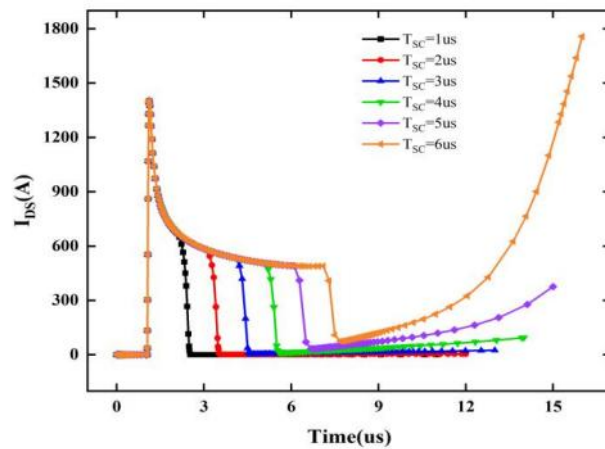


图 6 器件短路特性仿真曲线

通过仿真对短路耐受时间为 $4\mu\text{s}$ 、实际短路时间 $5\mu\text{s}$ 的器件进行热分析，可以清晰地观察到器件内部的动态热演变过程：在关断后 $2\mu\text{s}$ 时，热分布主要集中在漏极-源极之间的漂移区，最高晶格温度达到 1806K ，并形成明显的导电沟道。随着时间推移至 $10\mu\text{s}$ ，热效应进一步加剧，漂移区最高温度上升至 1912K ，且热量向栅极方向扩展。这一过程中，器件的耗尽区结构发生显著畸变，最终导致器件功能完全丧失。因此，可以得出 SiC 双沟槽 MOSFET 的短路失效机制，在短路过程中，漏极-源极间电流急剧增大，功耗飙升导致晶格温度快速累积，器件本征载流子被大量激发，形成电子-空穴对的“正反馈循环”，集中在漂移区的热量先后经历漏极-源极熔通、栅氧化层热击穿，最终形成贯穿性热失控通道，造成不可逆的永久性热失效。

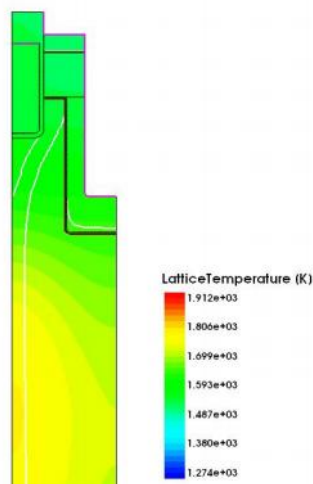


图 7 $T_{\text{sc}}=5\mu\text{s}$ ，关断 $2\mu\text{s}$ 内部晶格温度示意图

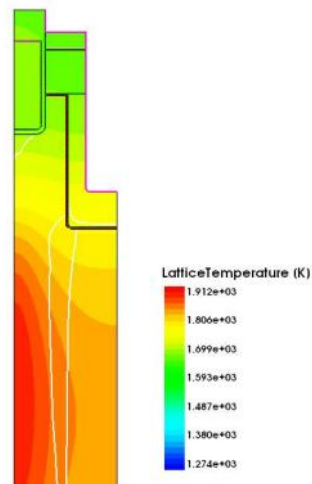


图 8 $T_{\text{sc}}=5\mu\text{s}$ ，关断 $10\mu\text{s}$ 内部晶格温度示意图

3.结论

开封后通过显微观察发现,漏极-源极区域存在明显的局部过热痕迹,印证了高温导致的物理损伤是失效的直接表现;结合 TCAD 仿真与热-电耦合模型,明确失效核心机理为:短路过程中,漏极-源极间电流急剧增大,功耗飙升导致晶格温度快速累积,器件本征载流子被大量激发,形成电子-空穴对的“正反馈循环”,最终,温度超过材料耐受极限,而 SiC 双沟槽 MOS 器件热量主要集中在漂移区先后经历漏极-源极熔通、栅氧化层热击穿,最终形成贯穿性热失控通道,造成不可逆的永久性热失效。

参考文献:

- [1] J. Millán, P. Godignon, X. Perpiñà, et al. "A Survey of Wide Bandgap Power Semiconductor Devices,"[J] in IEEE Transactions on Power Electronics, 2014, vol. 29, no. 5, pp. 2155-2163.
- [2] Bhalla, A. Silicon Carbide Semiconductors with Wide Bandgap for Electric Vehicles, [J] ATZ Electron Worldw, 2021, 16, 18–21.
- [3] L. F. S. Alves et al. "SiC power devices in power electronics: An overview," [C] 2017 Brazilian Power Electronics Conference (COBEP), Juiz de Fora, Brazil, 2017, pp. 1-8.
- [4] S. Yang, A. Bryant, P. Mawby, et al. "An Industry-Based Survey of Reliability in Power Electronic Converters," [J]in IEEE Transactions on Industry Applications, 2011, vol. 47, no. 3, pp. 1441-1451.
- [5] ROHM Semiconductor, SCT4062KRHR Datasheet [EB/OL], <https://rohmfs-rohm-com-cn.oss-cn-shanghai.aliyuncs.com/en/products/databook/datasheet/discrete/sic/mosfet/sct4062krhr-e.pdf>, 2023.
- [6] G. Romano et al., "A Comprehensive Study of Short-Circuit Ruggedness of Silicon Carbide Power MOSFETs," [J] in IEEE Journal of Emerging and Selected Topics in Power Electronics, 2016, vol. 4, no. 3, pp. 978-987.

基于电-热-力多物理场的 SiC 功率模块封装结构优化

陈朝兴, 李明阳, 李轩*, 韩久鹏, 邓小川, 张波

电子科技大学, 集成电路科学与工程学院 (示范性微电子学院), 电子薄膜与集成器件全国重点实验室, 成都, 611731

通讯作者: 李轩, xuanli@uestc.edu.cn

摘要: SiC 功率模块在运行时会呈现出电场、热场和应力场相互影响、相互作用的多物理场耦合特性。本文基于 COMSOL Multiphysics 软件展示了 SiC 功率模块的电-热-力多物理场耦合有限元仿真模型的详细建模过程, 根据仿真结果分析了功率模块的电流分布、温度分布和应力分布特征, 并进一步采用响应面法和粒子群算法对模块进行多目标协同优化。相较于原始封装结构, 优化结构尺寸后, 芯片结温下降 1.41°C, 芯片最大等效应力下降 37.98MPa。本研究为 SiC 功率模块在多物理场耦合作用下的设计和优化工作提供了参考, 具有一定的应用价值。

关键词: SiC 功率模块; 多物理场; 有限元仿真; COMSOL; 结构优化

1 引言

功率半导体器件是电能转换的核心组件, 在新能源汽车、轨道交通、智能电网和航空航天等领域发挥着关键作用^[1-2]。随着对功率器件工作频率、功率密度、电流/电压等级以及耐高温性能的要求不断提高, 传统硅基功率器件在高频、高压和高温等严苛工况下的应用严重受限。因此, 近年来, 以碳化硅(Silicon Carbide, SiC)为代表的第三代宽禁带半导体材料, 以其禁带宽度大、击穿电场强度高、热导率高以及电子饱和漂移速度快等显著特征, 获得了广泛关注^[3-4]。与硅基功率器件相比, SiC 功率器件具有耐高压、耐高温、开关速度快、热导率高和导通电阻低等优点, 能够满足功率器件向更高性能方向发展的需求。

在大功率应用场景下, 通常将多个 SiC 功率芯片封装为功率模块进行使用。SiC 功率模块在实际工作时会同同时存在电场、热场、应力场等多个物理场, 处于多物理场叠加耦合的状态。为充分发挥 SiC 功率模块的优异性能, 确保其安全稳定地工作, 需合理设计其封装结构。封装结构的设计在很大程度上依赖于仿真软件来准确表征相应的物理量, 如芯片结温和应力等。基于 SiC 功率模块的实际工作状态, 有必要实现考虑复杂耦合关系的多物理场协同仿真, 以保证仿真精度^[5]。

据此, 本文以 SiC 功率模块为研究对象, 选取 CAS120M12BM2 全 SiC 半桥模块为参考, 通过 COMSOL Multiphysics 软件建立该模块的有限元仿真模型, 开展电-热-力多物理场耦合仿真实验, 并进一步基于响应面法和粒子群算法对功率模块进行多目标协同优化, 降低了模块内芯片的结温和应力。该方法对复杂多物理场下的功率模块封装结构设计具有一定的借鉴和指导意义。

2 SiC 功率模块多物理场仿真模型的建立

2.1 SiC 功率模块封装结构三维模型的建立

本节以 CAS120M12BM2 全 SiC 半桥模块为参考建立 SiC 功率模块封装结构的三维模型,如图 1 所示,其中包含 2 颗型号为 CPM2-1200-0080B 的 SiC MOSFET 芯片和 2 颗型号为 CPW4-1200S020B 的 SiC Diode 芯片。封装结构的几何尺寸从上往下依次列于表 1。

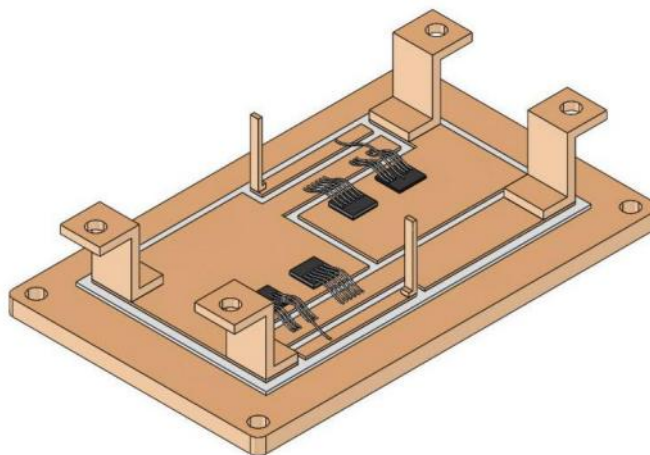


图 1 SiC 功率模块封装结构的三维模型

表 1 SiC 功率模块封装结构的几何尺寸

结构	尺寸(长×宽×高)/mm
SiC MOSFET 芯片	3.36×3.1×0.18
SiC 二极管芯片	3.08×3.08×0.377
SiC MOSFET 芯片焊料层	3.36×3.10×0.20
SiC 二极管芯片焊料层	3.08×3.08×0.20
DBC 上/下铜层	39.00×22.00×0.30
DBC 陶瓷层	41.00×24.00×0.48
DBC 焊料层	39.00×22.00×0.20
铜基板	51.00×32.00×2.00

2.2 仿真材料参数设置

在进行有限元仿真时,各封装材料的属性会直接影响仿真的准确性、收敛性及模型的性能。通过查阅相关文献^[6]及芯片 datasheet 将功率模块封装结构的各层材料参数输入 COMSOL 软件材料数据库,各封装结构的材料参数如表 2 所示,各种材料都假设为各向同性材料。

表 2 SiC 功率模块封装结构的材料参数

材料	密度 (kg/m ³)	热导率 (W/(m·K))	比热容 (J/(kg·K))	热膨胀系数 (10 ⁻⁶ ·K ⁻¹)	杨氏模量 (GPa)	泊松比
SiC	3200	370	710	4.5	410	0.14
Cu	8960	400	390	17	110	0.34
AlN	3300	180	745	4.8	310	0.24
Nano-Ag	8580	240	235	19.5	9	0.38
Al	2700	238	900	23.2	70	0.33

2.3 仿真边界条件设定

完成 SiC 功率模块封装结构三维模型的建立及材料参数设定后,需添加电流场、热场和应力场三个物理场及电-热、热-力物理场耦合接口。各物理场边界条件设置如下:

(1) 电流场边界条件设定

由于 DBC 陶瓷层材料 AlN 的电导率很低,因此将其设置为电绝缘条件,如图 2 所示。电流从终端 (DC+) 流入,经 DBC 上铜层进入 SiC 芯片,流经键合引线后通过另一侧的 DBC 上铜层、SiC 芯片和键合引线,最终到达接地端(DC-)。终端 (DC+) 设定输入电流为 20A。

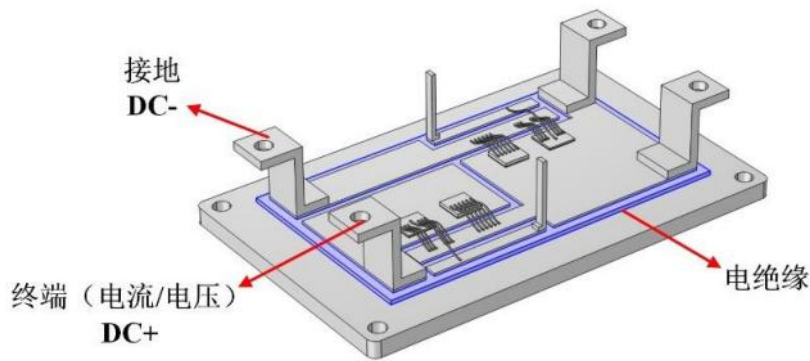


图 2 SiC 功率模块电流场边界条件设定

(2) 热场边界条件设定

图 3 为热场边界条件设置示意图,由于建模时忽略了铜基板下方的散热器,故将铜基板的底面设为对流热通量的边界,传热系数为 5000W/(m²·K),模拟强制水冷对流换热。实际功率模块经塑封填充后除基板外的其余面无额外散热条件,因此对基板四侧面设置对流换热系数为 10W/(m²·K),模拟现实中空气对流,其余面不设置散热。仿真的环境温度设置为 25°C。

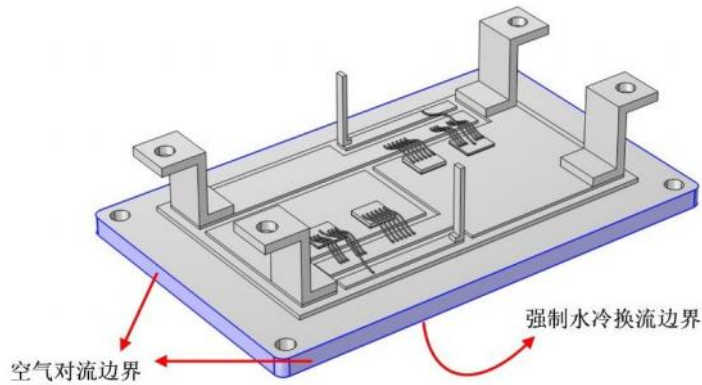


图 3 SiC 功率模块热场边界条件设定

(3) 应力场边界条件设定

将铜基板底面设为固定约束，其他部分视为自由边界，初始位移场为零。由于模块的每层相邻材料热膨胀系数都不同，故热膨胀边界包括了所有区域，如图 4 所示。

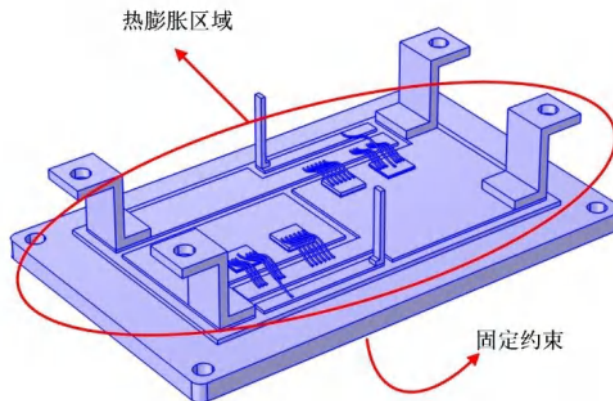


图 4 SiC 功率模块应力场边界条件设定

3 多物理场仿真结果分析

根据以上给出的封装结构尺寸、材料参数以及边界条件完成有限元仿真模型的设置，划分网格后进行了仿真分析。

3.1 SiC 功率模块封装结构电流、温度及应力分布

图 5 展示了 SiC 功率模块封装结构电流密度分布情况。由图可知，电流密度在模块内呈现非均匀分布，且最大电流密度为 $1.48 \times 10^8 \text{ A/m}^2$ 。键合引线区域的电流密度明显高于模块内其他部分，表明此处为电流汇聚和流经的关键路径，也是潜在的热点和可靠性薄弱环节。

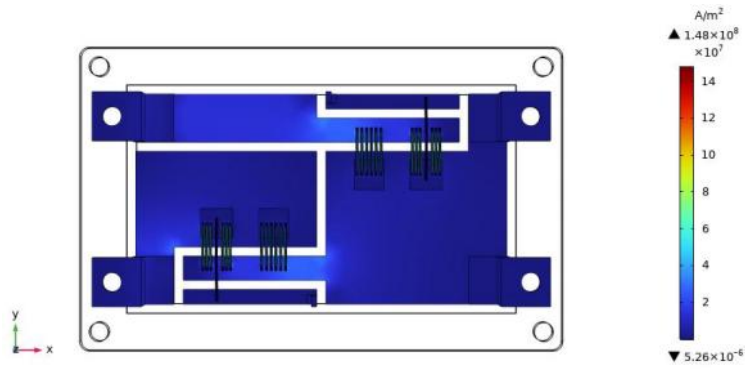


图 5 SiC 功率模块封装结构电流密度分布

图 6 为 SiC 功率模块封装结构温度分布图，从图中可以看出模块最高温度出现在芯片与键合引线连接处，且最高温度为 168°C，这主要是由于键合引线上的电流密度较大，导致其发热量较大，从而造成此处温度显著高于其他区域。

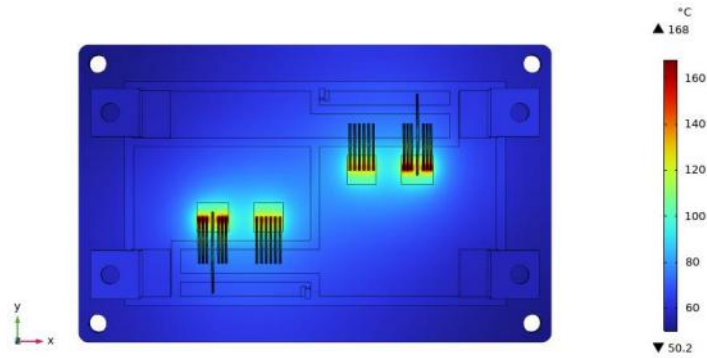


图 6 SiC 功率模块封装结构温度分布

图 7 为 SiC 功率模块封装结构应力分布图，由图可知模块封装结构的最大等效应力为 397MPa，也出现在芯片与键合引线连接处。这是由于键合引线与芯片的热膨胀系数差异较大，在模块工作温升过程中，热膨胀系数失配导致二者界面发生约束变形，进而产生较大的热应力。

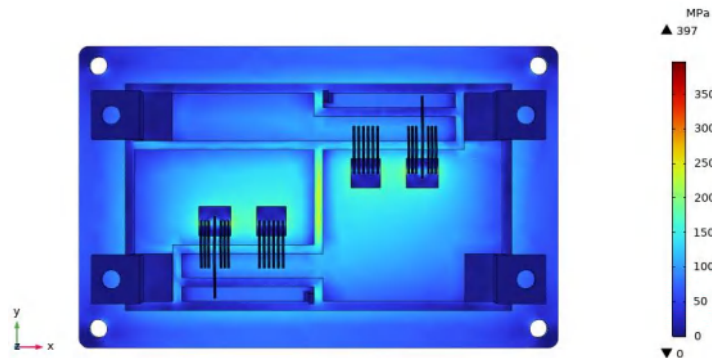


图 7 SiC 功率模块封装结构应力分布

4 SiC 功率模块封装结构优化

模块中各主要结构层的厚度是影响模块性能的重要参数^[7]，以 DBC 陶瓷层厚度 x_1 、芯片焊料层厚度 x_2 、DBC 基板焊料层厚度 x_3 作为优化变量，采用响应

面法和粒子群算法对模块封装结构进行多目标协同优化,构建了表 3 所示的试验因子数值表,选取芯片结温与最大等效应力为优化目标,得到优化前和优化后的结果如表 4 所示。从表 4 可以看出,优化后芯片结温下降 1.41°C,芯片最大等效应力下降 37.98MPa。

表 3 响应面法试验因子数值表

最值	x_1 (mm)	x_2 (mm)	x_3 (mm)
最小值	0.38	0.10	0.10
最大值	0.98	0.40	0.40

表 4 SiC 功率模块封装结构优化结果

对比	DBC 陶瓷层 厚度 (mm)	芯片焊料层 厚度 (mm)	DBC 基板焊 料层厚度 (mm)	芯片结温 (°C)	芯片最大 等效应力 (MPa)
优化前	0.48	0.20	0.20	172.02	279.81
优化后	0.475	0.158	0.381	170.61	241.83

6 结论与展望

本文采用电-热-力多物理场有限元仿真方法,研究了多物理场耦合作用下模块内的温度、应力和电流密度分布。通过响应面法和粒子群算法对封装结构进行优化,得到最优的结构尺寸参数组合,与初始未经优化的结构相比,模块中芯片结温降低 1.41°C,最大等效应力降低 37.98MPa。该研究对于降低功率模块的最高工作温度及整体应力提供了参考方案,有助于提升功率模块的工作寿命。

参考文献

- [1] 盛况,唐苇羽,吴赞.碳化硅功率模块封装及热管理关键技术[J].机车电传动,2023,(05):1-9.
- [2] 王广来.基于石墨嵌入式 SiC 功率模块封装的结构设计与优化[D].桂林电子科技大学,2024.
- [3] 程书博,张金利,张义政,等.SiC 功率模块封装材料的研究进展[J].科技创新与应用,2024,14(03):106-109.
- [4] C. Chen, F. Luo and Y. Kang. A review of SiC power module packaging: Layout, material system and integration[J]. CPSS Transactions on Power Electronics and Applications,2017,2(3): 170-186.
- [5] Y. Yang, Z. Wang, Y. Ge, G. Xin and X. Shi. An Automated Field-Circuit Coupling Simulation Method Based on PSpice-MATLAB-COMSOL for SiC Power Module Design[J]. IEEE Transactions on Power Electronics,2023,38(10): 12634-12647.
- [6] R. Khazaka, L. Mendizabal, D. Henry and R. Hanna. Survey of High-Temperature Reliability of Power Electronics Packaging Components[J], IEEE Transactions on Power Electronics, 2015,30(5): 2456-2464.
- [7] 张一鸣,邓二平,赵志斌,等.压接型 IGBT 器件封装内部多物理场耦合问题研究概述[J].中国电机工程学报,2019,39(21):6351-6365.

超结功率 MOSFET 反向恢复特性仿真研究

郭思睿, 付祥和, 张园园, 郭子印, 张钧辉, 汪宇浩, 贺永宁

西安交通大学微电子学院,

陕西省电子器件与高端芯片重点实验室, 西安 710049

摘要: 超结 (SJ) MOSFET 因低导通电阻应用广泛, 但高频应用需求增长使其反向恢复特性变得重要。本文完成了 900V SJ MOSFET 建模, 通过提取反向峰值电流、恢复时间和软度系数仿真研究了缓冲层厚度、工作温度及深能级缺陷对反向恢复特性的影响规律。其中, 深能级缺陷对反向恢复的影响占主导, 深能级缺陷浓度较大达到 10^{15}cm^{-3} , 位置更靠近禁带中央至 $E_c - 0.350 \text{eV}$ 时, 反向峰值电流减小达 62.7%, 恢复时间减小 40.1%, 软度提升 165%。但实际设计需考虑与导通电阻、反向漏电流等参数折衷。

关键词: 超结 MOSFET; 反向恢复特性; 深能级缺陷; 反向峰值电流; 反向恢复时间; 软度系数

1 前言

超结 (SJ) 功率 MOSFET 基于电荷补偿原理, 在保证击穿电压 (BV) 不变的情况下具有更低的特征导通电阻 ($R_{\text{on,sp}}$), 打破了传统“硅限”, 大幅度降低了器件导通功耗^[1], 广泛应用于电机驱动、工业控制、汽车电子等领域, 控制着功率的流向和转换。然而, 由于 SJ MOSFET 器件内源极和漏极间寄生体二极管的存在, 带来了反向恢复的问题^[2], 当体二极管从导通态切换至关断态时, 需要抽取其在导通阶段存储的大量电荷, 从而产生大的反向电流, 直接影响器件的开关损耗、电磁干扰 (EMI) 水平和可靠性。随着现代社会与电力电子技术的发展, 对器件工作频率不断提出更高的要求, 导致在逆变电路、双向 DC-DC 转化器等应用场景中, SJ MOSFET 体二极管高频率地续流与关断, 使得反向恢复特性的研究变得重要^[3]。

因此, 本文针对高压高频率的应用场合, 完成了 900V SJ MOSFET 的设计与建模, 并通过反向恢复特性的仿真, 提取了不同条件下反向峰值电流 I_{rm} 、反向恢复时间 t_{rr} 和软度系数 S 三项参数进行对比, 研究了器件缓冲层厚度的结构参数、工作温度以及深能级缺陷对于反向恢复特性的影响规律, 为 SJ MOSFET 的反向恢复特性优化提供一定的参考。

2 器件建模与仿真方法

2.1 超结 MOSFET 器件建模

本文所采用器件元胞结构如图 1 所示，为硅基平面栅型 SJ VDMOS，漂移区中 n 型条带和 p 型条带交替排列形成超结结构，并且在超结柱与衬底之间具有一层 n 型缓冲层，降低实际工艺制造难度。

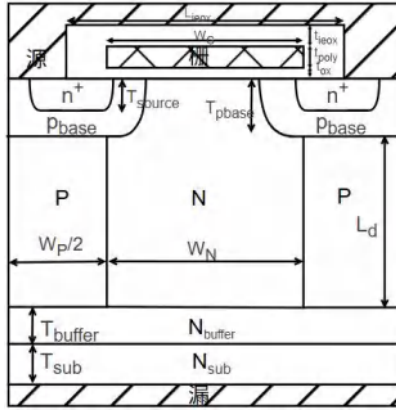


图 1 SJ VDMOS元胞结构

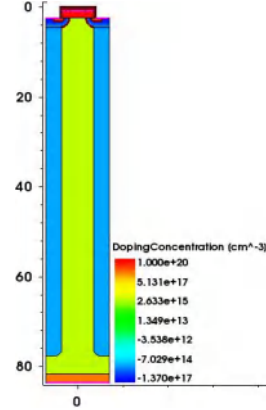


图 2 SJ VDMOS元胞建模

SJ VDMOS 设计参数的初步计算基于章文通等人所提出的非全耗尽（NFD）模式和全耗尽（FD）模式的超结器件模型及相关理论^[4]，对于 NFD 模式下的 SJ VDMOS 器件，能够实现更优的导通电阻与击穿电压间的关系，具有以下计算关系：

$$N_d W_d = 2.19 \times 10^{12} W_d^{-0.267} V_B^{0.145} (\text{cm}^{-2}) \quad (1)$$

$$L_d = 2.31 \times 10^{-2} V_B^{7/6} (\mu\text{m}) \quad (2)$$

$$R_{on,sp} = 1.17 \times 10^{-3} W_d^{1.16} V_B^{1.03} (m\Omega \cdot \text{cm}^2) \quad (3)$$

式中， N_d 为漂移区掺杂浓度，且认为超结柱掺杂浓度与漂移区掺杂浓度相同， W_d 是超结柱宽度， V_B 是击穿电压， L_d 为超结柱长度。

根据最小击穿电压 900V 的设计目标，考虑一定的终端和工艺等方面容差，选定元胞击穿电压的设计目标为 $V_B = 1000\text{V}$ 。选取的超结柱宽 $W_n = W_p = W_d = 7\mu\text{m}$ ，带入式（1）、（2）和（3），即可得到初步设计参数。

利用 Sentaurus TCAD 软件进行器件结构建模如图 2 所示，通过仿真优化可得到 SJ VDMOS 的阈值电压为 3.506V，击穿电压达到 1115V，特征导通电阻 $19.16 m\Omega \cdot \text{cm}^2$ 。总结最终 SJ VDMOS 元胞建模参数于表 1。

表 1 900V SJ VDMOS 元胞建模参数

符号	参数	选取数值	单位
W_N	N 柱宽度	7	μm
W_P	P 柱宽度	7	μm
L_d	超结柱长度	73	μm
T_{buffer}	缓冲层厚度	4	μm
N_{pillar}	N 柱掺杂浓度	5×10^{15}	cm^{-3}
P_{pillar}	P 柱掺杂浓度	5×10^{15}	cm^{-3}
N_{sub}	衬底掺杂浓度	1×10^{19}	cm^{-3}

2.2 反向恢复特性仿真方法

反向恢复特性的仿真基于双脉冲测试电路，在 Sentaurus 中利用 SPICE 模型搭建系统电路如图 3 所示，其中 SJ MOSFET 即为本文建模 SJ VDMOS，仿真设置开关管栅脉冲电压 10V，高电平时间 10 μ s，低电平时间 5 μ s。

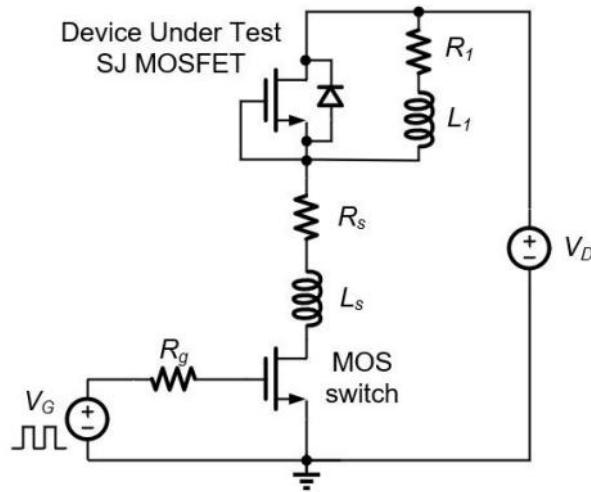


图 3 双脉冲仿真测试电路

对于反向恢复特性的评判，通过提取其反向恢复时间 t_{rr} 、反向峰值电流 I_{rrm} 和软度系数 S 三项参数进行比较。更短的 t_{rr} 、更小的 I_{rrm} 和更大的 S 是我们所期望的，能够使电路在快速关断的同时，具有更小的损耗和更小的震荡与噪声^[5]。

3 结果与讨论

3.1 缓冲层厚度 T_{buffer} 的结构参数对反向恢复特性的影响

SJ MOSFET 的缓冲层在体二极管导通后同样会存储电荷，从而影响器件的反向恢复特性。为探究其具体影响，选取不同缓冲层厚度 T_{buffer} 参数，仿真其反向恢复特性如图 4 所示，同时提取其特性参数如表 2 所示。

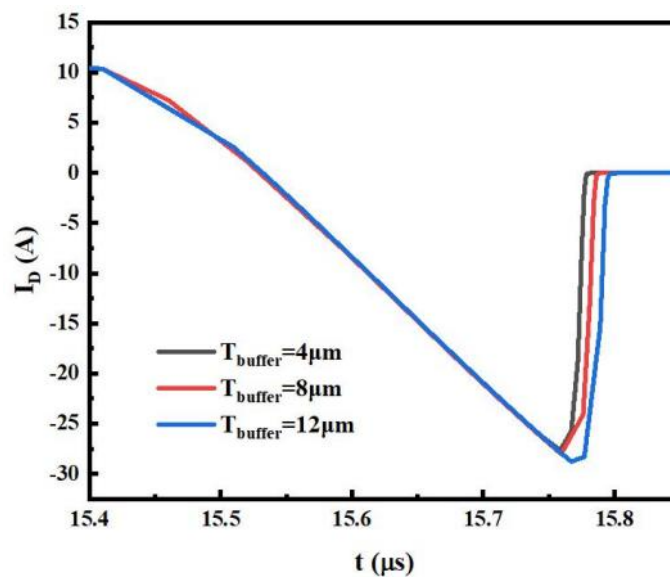


图 4 SJ MOSFET 不同缓冲层厚度下反向恢复特性对比

可以看出，当缓冲层厚度增大时，SJ MOSFET 的反向峰值电流 I_{rrm} 增大，且

反向恢复时间 t_{rr} 增加，是因为体二极管关断后同样需要抽取增加的缓冲层中存储的电荷。但同时，由于抽取缓冲层中电荷时器件的阻断电压已经达到施加的反向偏压，因此缓冲层厚度的增加会使反向恢复的软体有所提升。同时，从表 2 可以看出，当 T_{buffer} 增加到 $8\mu m$ 时，软度系数 S 增加，但当 T_{buffer} 增加到 $12\mu m$ 时，软度系数 S 则几乎不变，可能与导通时注入的少数载流子达到饱和有关。且本文建模中，缓冲层掺杂浓度与 n 柱掺杂浓度一致，而在实际应用中，缓冲层的掺杂浓度往往需要更改，以实现更优的软度。此外，缓冲层厚度的增加，也会使器件的导通电阻 R_{on} 增加。

因此，实际设计中，缓冲层厚度增加能够优化反向恢复的软度，但需要与 I_{rrm} 和 t_{rr} 进行折衷，同时考虑其对 R_{on} 带来的不利影响。

表 2 SJ MOSFET 不同缓冲层厚度下反向恢复特性参数提取

$T_{buffer}/\mu m$	I_{rrm}/A	$t_{rr}/\mu s$	S
4	27.6	0.267	0.192
8	28	0.291	0.299
12	28.3	0.312	0.297

3.2 SJ MOSFET 工作温度对反向恢复特性的影响

SJ MOSFET 在实际环境中不同的工作温度也会对反向恢复特性带来影响，因此本节设置了不同的工作温度进行仿真，得到反向恢复特性如图 5 所示，并提取其特性参数如表 3 所示。

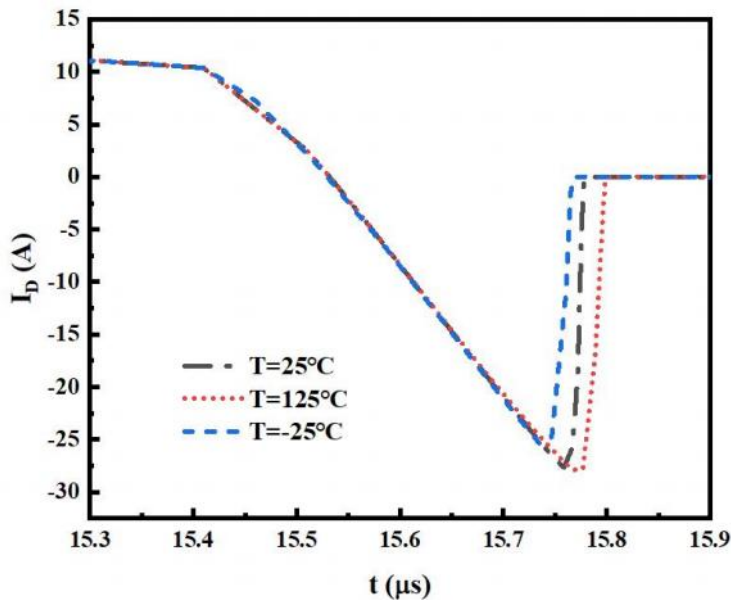


图 5 SJ MOSFET 不同工作温度下反向恢复特性对比

可以看出，随着温度的升高，SJ MOSFET 的反向恢复特性发生恶化，反向峰值电流 I_{rrm} 增大，反向恢复时间 t_{rr} 增加，并且软度系数 S 也有所减小。这是因为随着温度的升高，载流子寿命增加，使需要抽取的电荷增加。

因此，在以反向恢复特性为目标进行器件设计时，需要确保高温下反向恢复特性的恶化处于可接受区间。

表 3 SJ MOSFET 不同工作温度下反向恢复特性参数提取

T/°C	I _{rm} /A	t _{rr} /μs	S
-25	25.4	0.242	0.21
25	27.6	0.267	0.192
125	27.9	0.275	0.185

3.3 深能级缺陷对反向恢复特性的影响

现今工艺生产中，人们往往通过金、铂等重金属掺杂^[6]，或电子、质子等高能粒子辐照技术来改善器件的反向恢复特性^[7]，其本质是在器件内引入深能级缺陷，即在禁带中引入深能级。为了探究深能级缺陷对 SJ MOSFET 反向恢复特性的具体影响，本文在仿真中加入郭春生等人通过深能级瞬态谱（DLTS）测得的电子辐照在硅中产生的深能级缺陷能级进行仿真^[8]。

首先以 E_c-0.350eV 位置处深能级为例，仿真不同缺陷能级浓度对反向恢复特性的影响如图 6 所示，并提取其特性参数如表 4 所示。

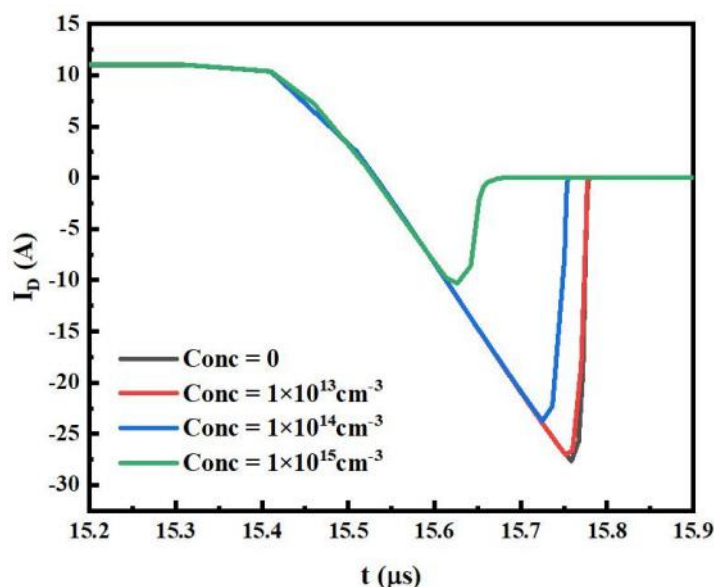


图 6 E_c-0.350eV 处不同缺陷能级浓度下反向恢复特性对比

可以看出，随着深能级缺陷能级浓度的增大，SJ MOSFET 的反向恢复特性改善越大，反向峰值电流 I_{rm} 逐渐减小，反向恢复时间 t_{rr} 同样逐渐减小，同时软度系数 S 增大，反向恢复软度也得到较大改善。这是因为引入深能级缺陷能级增加了少数载流子的复合率，减小了载流子寿命，当引入浓度越大时，对于载流子寿命的影响也越大。

表 4 E_c-0.350eV 处不同缺陷能级浓度下反向恢复特性参数提取

Conc/cm ⁻³	I _{rm} /A	t _{rr} /μs	S
0	27.6	0.267	0.192
1e13	27	0.262	0.213
1e14	23.7	0.252	0.326
1e15	10.3	0.16	0.509

接下来在 1 × 10¹⁵ cm⁻³ 这一对反向恢复特性改性较为有效的能级浓度下，仿真研究了不同位置处深能级缺陷能级对反向恢复特性的影响如图 7 所示，同时提

取了其特性参数如表 5 所示。

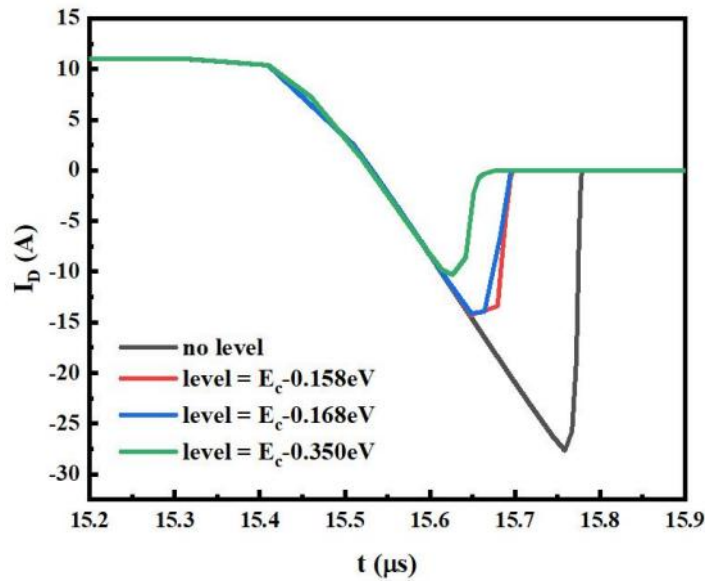


图 7 $1 \times 10^{15} \text{cm}^{-3}$ 浓度下不同位置处缺陷能级反向恢复特性对比

可以看出，随着深能级缺陷能级位置更靠近禁带中心处，其对于反向恢复特性的改善越有效，反向峰值电流 I_{rrm} 越小，反向恢复时间 t_{rr} 越短。这是因为当深能级越靠近禁带中心时，对于电子和空穴两种载流子捕获概率接近，更能作为有效的复合中心。但是不同位置深能级缺陷能级对于软度的影响较小，软度系数 S 几乎不变。

表 5 $1 \times 10^{15} \text{cm}^{-3}$ 浓度下不同位置处缺陷能级反向恢复特性参数提取

位置	I_{rrm}/A	$t_{\text{rr}}/\mu\text{s}$	S
无	27.6	0.267	0.192
$E_c - 0.158\text{eV}$	14.3	0.172	0.513
$E_c - 0.168\text{eV}$	14.1	0.17	0.494
$E_c - 0.350\text{eV}$	10.3	0.16	0.509

因此，在实际器件引入的深能级缺陷能级中，浓度更大、位置更靠近禁带中心的能级对于反向恢复特性的改善更有利。但同时，这些能级往往也会使器件的反向漏电流增大，因而在实际生产中需要在深能级缺陷能级对反向恢复特性的改善和对反向漏电流的恶化之间折衷。

同时，将缓冲层厚度、工作温度和深能级缺陷对反向恢复特性的影响进行对比，可以看出，深能级缺陷对于反向恢复特性的影响最为重要，占主导作用。当深能级缺陷浓度较大达到 $1 \times 10^{15} \text{cm}^{-3}$ ，位置更靠近禁带中央至 $E_c - 0.350\text{eV}$ 时，对反向峰值电流 I_{rr} 改善可达 62.7%，反向恢复时间 t_{rr} 改善 40.1%，软度提升 165%。

4 结论

本文完成了 900 SJ MOSFET 的建模，并针对其反向恢复特性展开仿真研究。通过提取其反向峰值电流 I_{rrm} 、反向恢复时间 t_{rr} 和软度系数 S 三项特性参数，研究了器件缓冲层厚度的结构参数、工作温度以及深能级缺陷对于反向恢复特性的

影响规律：缓冲层厚度越大，对于反向恢复的软度优化越好，但会使反向峰值电流和反向恢复时间增大；工作温度的升高会恶化反向恢复特性，反向峰值电流越大，反向恢复时间越长，并且反向恢复软度越硬；深能级缺陷能级浓度越大、位置更靠近禁带中心，反向恢复特性越好，反向峰值电流越小，反向恢复时间越短，但仅浓度增大可使反向恢复软度变软，缺陷位置对于反向恢复软度影响不明显。其中，深能级缺陷对反向恢复特性的影响占主导作用，深能级缺陷浓度较大达到 $1 \times 10^{15} \text{cm}^{-3}$ ，位置更靠近禁带中央至 $E_c - 0.350 \text{eV}$ 时，对反向峰值电流 I_{rr} 改善可达 62.7%，对反向恢复时间 t_{rr} 改善 40.1%，软度提升 165%。此外，在实际设计与生产中，还需要将反向恢复特性与导通电阻、反向漏电流等参数进行折衷考虑，以实现器件整体性能的优化。

参考文献

- [1] Udrea F, Deboy G, Fujihira T. Superjunction power devices, history, development, and future prospects[J]. IEEE Transactions on Electron Devices, 2017, 64(3): 713-727.
- [2] Jahdi S, Alatisse O, Bonyadi R, et al. An analysis of the switching performance and robustness of power MOSFETs body diodes: A technology evaluation[J]. IEEE Transactions on Power Electronics, 2014, 30(5): 2383-2394.
- [3] Chimento F, Musumeci S, Raciti A, et al. Super-junction MOSFET and SiC diode application for the efficiency improvement in a boost PFC converter[C]//IECON 2006-32nd Annual Conference on IEEE Industrial Electronics. IEEE, 2006: 2067-2072.
- [4] Zhang W, Zhang B, Li Z, et al. Theory of superjunction with NFD and FD modes based on normalized breakdown voltage[J]. IEEE Transactions on Electron Devices, 2015, 62(12): 4114-4120.
- [5] 潘永涛. 快恢复二极管在高频电路中的选型应用[J]. 电子测试, 2022, (19): 48-51. DOI:10.16520/j.cnki.1000-8519.2022.19.035.
- [6] Khaderbad M A, Houadef A, Rogers C, et al. Reducing Reverse Recovery Charge in Trench Power MOSFETs Using Platinum Traps Introduced by Implantation and PVD[C]//2024 36th International Symposium on Power Semiconductor Devices and ICs (ISPSD). IEEE, 2024: 410-413.
- [7] Ivanov P A, Kudoyarov M F, Potapov A S, et al. Correction of the Reverse Recovery Characteristics of High-Voltage 4 H-SiC Junction Diodes Using Proton Irradiation[J]. Semiconductors, 2019, 53(6): 850-852.
- [8] Guo C S, Wang R M, Zhang Y W, et al. Identifying defect energy levels using DLTS under different electron irradiation conditions[J]. Nuclear Science and Techniques, 2017, 28(12): 183.

超结 MOSFET 器件 EMI 优化仿真分析

张园园^{1,2}, 张钧辉², 汪宇浩², 贺永宁²

1. 西安龙飞电气技术有限公司, 西安 710021

2. 西安交通大学微电子学院, 西安 710049

摘要: 超结 MOSFET 因其导通电阻低、开关速度快被广泛应用于高频高功率场景, 随着器件不断迭代及应用升级, EMI 问题逐渐凸显。本文通过器件本身参数的优化来降低 EMI, 基于 TCAD 仿真工具进行超结 MOSFET 器件建模, 仿真了不同工艺条件对器件 EMI 特性的影响, 发现通过增加 JFET 区注入剂量可以在不影响 BV 及 R_{dson} 的情况下将 di/dt 降低 52%; 通过适当改变栅氧层厚度也可以将 di/dt 降低 40%。以上工艺调控措施不需要改变器件版图, 现有产线工艺易于实现, 对超结 MOSFET 器件的设计、制造具有很大实用价值。

关键词: 超结 MOSFET; EMI; di/dt

1 引言

功率 MOSFET 由于易驱动、关态阻抗高、开关速度快, 广泛用于各类电能转换系统中。超结 MOSFET 结构最早由电子科技大学的陈星弼院士提出^[1], 基于电荷平衡的原理打破了硅材料的“硅限”^[2], 在相同的芯片面积上, 相同额定电压下, 超结 MOSFET 的导通电阻会比普通的 VDMOS 降低 3-10 倍, 大大降低了电源变换系统的损耗, 提高了电能的使用效率。同时, 由于有横向电场的辅助耗尽, 超结 MOSFET 的开关速度更快, 在相同厚度的外延层上可以实现更高的电压, 非常适用于高压、高频、高功率应用场景^[3]。但是随着器件的升级迭代, 系统的频率不断提高, 电磁兼容(EMC)问题开始突显, EMC 包括了 EMI 和 EMS 两部分^[4]。EMI 指的是电路中变化的电流和电压产生的电磁场与周围的电路和设备相互作用, 导致电路中的其他信号变得不稳定或失真, 影响设备或系统正常运行的现象。EMS 是指在特定环境下, 设备或系统能够承受满足标准规定范围内的电磁能量干扰, 从而保持正常运行的能力。

EMI 会影响电子设备的正常运行也会对设备使用人员造成电磁辐射污染, 国家对各类电子设备的电磁干扰发射限度有明确的要求^[5, 6]。因此, 应用于高频高功率场景的超结 MOSFET 器件, 在设计及制造中要特别注意 EMI 特性的评估, 以满足系统对 EMI 指标的要求。

2 超结 MOSFET 开关过程中的 EMI 分析

图 1 展示了超结 MOSFET 的等效电路图, 器件主要寄生参数包括了体二极

管、寄生电容、栅极电阻、寄生电感四个部分。其中体二极管寄生于 P-Body 和 N 型漂移区。寄生电容来自器件内部，主要为器件各极间的耦合电容^[7]。栅极电阻是指器件内部的栅极总电阻 R_g 。寄生电感则来自于封装键合线，不同封装形式（如贴片式和直插式封装）需要的键合线的长度、数量和材质都有所不同，从而产生较大差异。这些非理想寄生参数在器件的不同工作状态时，会与外电路构成谐振，引发电流和电压振荡，从而产生严重 EMI 噪声。

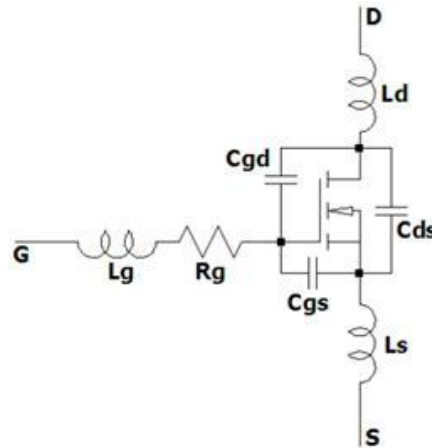


图 1 MOSFET 等效电路

2.1 超结 MOSFET 寄生电容引发的 EMI

器件的开关速度过快会导致电流和电压在短时间内产生突变，形成电流和电压的尖峰，产生振荡，从而引起严重的 EMI 噪声。具体来说，以器件开启过程为例，存在两个电流电压快速变化的时间段，分别是电流上升时间 t_{ri} 和电压下降时间 t_{fv} ， t_{fv} 时的电压变化率可以表示为公式（1）：

$$\frac{dv}{dt} = \frac{V_{PK}}{t_{fv}} = \frac{V_{PK}}{R_g C_{gd} (V_{DS}) \frac{V_{PK} - V_{ON}}{V_P}} \quad (1)$$

同理可将 t_{ri} 时电流变化率表示为公式（2）：

$$\frac{di}{dt} \sim \frac{I_D}{R_g [C_{gs} + C_{gd} (V_{DS})] \ln \left(\frac{V_P}{V_{TH}} \right)} \quad (2)$$

由公式（1）及公式（2）可见，器件开关过程的 di/dt 和 dv/dt 主要与 R_g 、 C_{gs} 和 C_{gd} 有关。其中 R_g 和 C_{gs} 主要与器件本身结构有关，二者越小，超结 MOSFET 开关速度越快， di/dt 和 dv/dt 越大，产生的 EMI 噪声越严重。

2.2 超结 MOSFET 密勒电容与栅源电容引发的 EMI

超结 MOSFET 的饱和区漏电流 I_D 在开启初态与栅压的关系为^[8]：

$$I_D = \frac{1}{2} \mu_n C_{ox} \frac{W}{l} (V_{GS}(t) - V_{th})^2 \quad (3)$$

MOSFET 为栅控器件，栅压的变化可以影响电流 I_D 的变化，故 dI_D/dt 可以表示为：

$$\frac{dI_D}{dt} = K \frac{dV_{GS}(t)}{dt} \quad (4)$$

其中 $K = \mu_n C_{ox} \frac{w}{l} (V_{GS}(t) - V_{th})$ ，代入开通过程可以得出：

$$\frac{dI_D}{dt} = \frac{K \cdot \frac{V_{drive} - V_{GS}(t)}{R_G}}{(C_{GS} + C_{GD,OX}) \left[1 - KR_{CH} \frac{1}{\frac{C_{GS}}{C_{GD,OX}} + 1} \right]} \quad (5)$$

由公式（5）中可知，增大 C_{GS} 和增大 $C_{GS}/C_{GD,OX}$ 都可以使 di/dt 减小，从而降低 EMI。

由开关过程的分析可见，优化 EMI 特性的关键在于降低开关过程中的 di/dt 及 dv/dt ，降低 di/dt 及 dv/dt 的关键在于增大 R_g 、 C_{gs} 、 $C_{GS}/C_{GD,OX}$ 及降低 C_{GD} 随 V_{DS} 的变化速率，与之相关的工艺参数主要为栅极长度、栅氧化层厚度、JFET 区大小及 JFET 区注入剂量、buffer 层厚度及电阻率等。

3 超结 MOSFET 的 EMI 优化仿真

首先基于 TCAD 对器件进行建模，部分器件参数信息见表 1，器件结构见图 2。

表 1 器件部分基础参数

JFET	60k 2e12	GOX	1000A
Ronsp	27.11mΩ·cm ²	poly	4000A
BV	790.0V	VT	3.15V

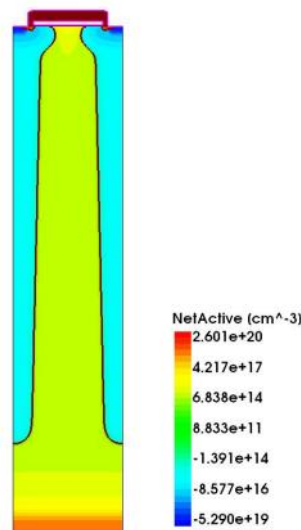


图 2 超结MOSFET器件结构示意图

3.1 改变氧化层厚度仿真

基于建模器件将栅氧厚度从 0.04 μm 拉偏到 0.17 μm 进行开关过程仿真, 关断过程 di/dt 随栅氧厚度变化如图 3 所示, 开通过程中 I_d - V_g 随栅氧厚度拉偏的变化如图 4 所示。

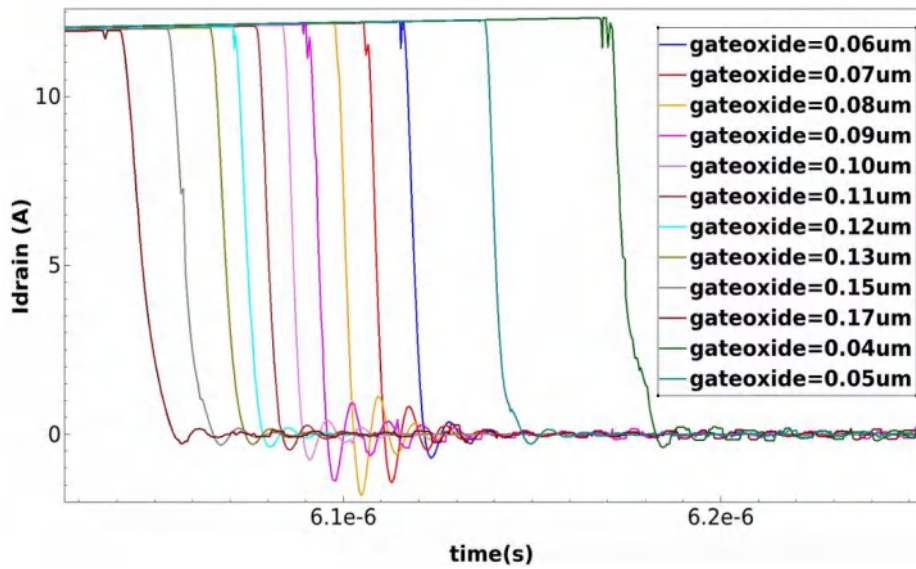


图 3 关断过程漏电流随栅氧化层厚度的变化

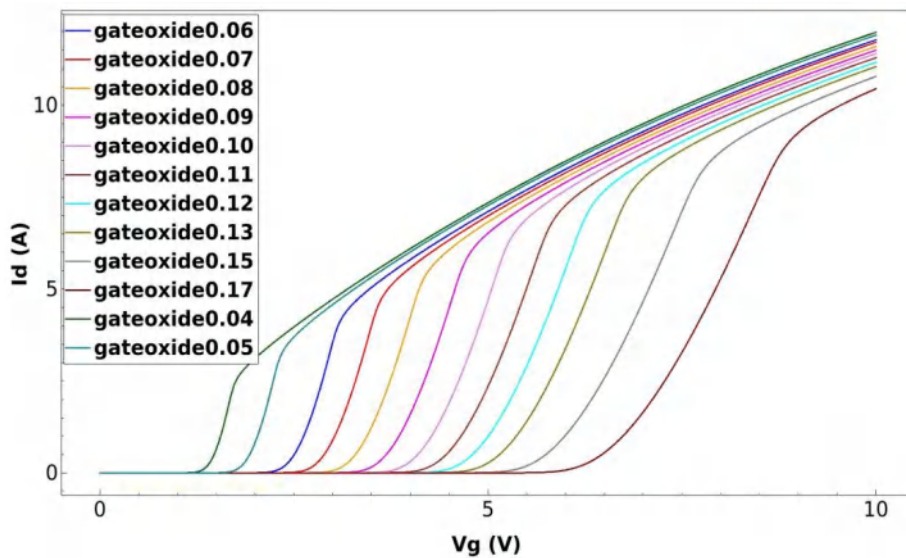


图 4 I_d - V_g 随栅氧化层厚度的变化

提取不同栅氧厚度条件的器件在开关过程中的 di/dt 、 C_{gs} 等参数, 如表 2 所示, 其中 $t_{ox} \cdot C_{GS}$ 表征了 $C_{GS}/C_{GD,ox}$ 。

由仿真结果可以看到, 随着氧化层厚度的增大, $t_{ox} \cdot C_{GS}$ 的值在变小, 使得 di/dt 降低; 同时 C_{GS} 的值也在变小, 使得 di/dt 增加。当氧化层厚度大于 0.07 μm 时, $t_{ox} \cdot C_{GS}$ 占据主导地位, 导致了 di/dt 的降低, 当氧化层厚度小于 0.07 μm 时, C_{GS} 的值占据主导地位, 使 di/dt 降低。

表 2 不同氧化层厚度器件性能参数

氧化层厚度(um)	VT(V)	di/dt(A/us)	Cgs(F,Vd=200V)	tox·C _{GS}
0.04	0.99	2558	1.051×10^{-9}	0.4204
0.05	1.36	4956	9.165×10^{-10}	0.4583
0.06	1.84	5510	8.065×10^{-10}	0.4839
0.07	2.15	6299	7.423×10^{-10}	0.5196
0.08	2.48	6094	6.847×10^{-10}	0.5478
0.09	2.81	5716	6.42×10^{-10}	0.5778
0.10	3.15	5879	6.083×10^{-10}	0.6083
0.11	3.44	4720	5.729×10^{-10}	0.6302
0.12	3.71	4137	5.455×10^{-10}	0.6546
0.13	4.0	3640	5.216×10^{-10}	0.6781
0.15	4.47	3041	4.792×10^{-10}	0.7188
0.17	5.07	2220	4.478×10^{-10}	0.7613

改变氧化层厚度后，会导致阈值电压漂移，减小和增大氧化层厚度都会使 di/dt 减小，但减小氧化层 di/dt 的变化幅度较小，工艺上也比较难制造，因此选择增大氧化层厚度，同时减小 pwell 离子注入浓度来控制阈值电压的漂移，结果如表 3 所示。当氧化层厚度为 0.15um, pwell 注入剂量为 $3.5 \times 10^{13} \text{ cm}^{-2}$ 时, VT 3.1V 满足器件参数要求, di/dt 由原本的 5879 A/us 降低到 3542A/us, 降低 40%, 极大的优化了器件的 EMI 特性。

表 3 不同 pwell 离子注入浓度器件性能参数

氧化层厚度(um)	pwell 离子注入浓度(cm^{-2})	VT (V)	di/dt(A/us)
0.11	5.5×10^{13}	3.15	5118
	5×10^{13}	2.97	4924
0.12	5×10^{13}	3.21	4192
	4.5×10^{13}	3.0	5223
0.13	5×10^{13}	3.45	4041
	4.5×10^{13}	3.23	3714
	4×10^{13}	3.0	3957
0.15	3.5×10^{13}	3.11	3542

3.2 改变 JFET 注入剂量仿真

基于建模器件将 JFET 注入剂量从 $1 \times 10^{12} \text{ cm}^{-2}$ 拉偏到 $2 \times 10^{13} \text{ cm}^{-2}$ 进行开关过程仿真，关断过程 di/dt 随 JFET 注入剂量变化如图 5 所示。提取 di/dt 参数及其

他器件关键特性列于表 4。可以看到，随着 JFET 剂量的增加， di/dt 逐渐降低，当 JFET 剂量从 $2 \times 10^{12} \text{ cm}^{-2}$ 增加到 $1 \times 10^{13} \text{ cm}^{-2}$ 时， di/dt 从由原本的 4995 A/us 降低到 2419A/us，降低 52%，极大的优化了器件的 EMI 特性，同时 BV 未退化， R_{onsp} 降低了 5%，略有优化。

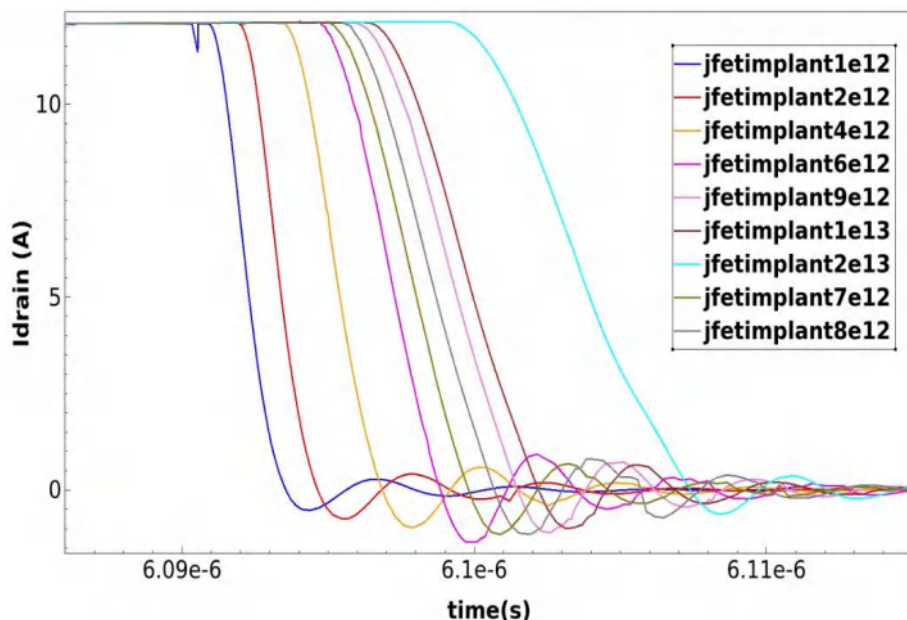


图 5 关断过程漏电流随JFET注入剂量的变化

表 4 不同 JFET 注入剂量器件性能参数

JFET 离子注入剂量 (cm^{-2})	BV (V)	R_{onsp} ($\text{m}\Omega \cdot \text{cm}^2$)	di/dt (A/us)
1×10^{12}	789.5	28.15	5159
2×10^{12}	790.0	27.11	4995
4×10^{12}	791.1	26.35	4045
6×10^{12}	792.1	25.95	3356
7×10^{12}	792.5	25.91	2873
8×10^{12}	793.0	25.83	2683
9×10^{12}	793.6	25.67	2606
1×10^{13}	794.0	25.63	2419
2×10^{13}	448.3	25.19	1796

除了栅氧厚度及 JFET 注入剂量，还仿真了超结器件 buffer 层厚度及电阻率、P+注入剂量及能量等工艺参数对开关过程中 di/dt 的影响，发现变化率均不足 10%，可见栅氧厚度及 JFET 注入剂量为影响器件 EMI 特性的关键工艺参数。另外，本文仿真过程中未改变器件的结构，即原胞尺寸、栅极宽度、p-pillar 间距等，因此可以不用变更光刻版，仅通过很方便的工艺拉偏达到 di/dt 50%的优化。

4 小结

EMI 特性对超结 MOSFET 的应用非常关键，需要在器件的设计及制造阶段充分考虑及优化，以保证器件可以满足终端模块及系统的 EMI 指标要求。可以通过开关过程中的 di/dt 或 dv/dt 来表征 EMI 特性， di/dt 或 dv/dt 越小，则 EMI 特性越好。

本文基于 TCAD 器件建模，仿真了不同工艺条件下超结 MOSFET 开关过程中的 di/dt 参数及其他关键器件特性参数。发现在不改变器件版图设计的前提下，栅氧厚度及 JFET 注入剂量为影响器件 EMI 特性的关键工艺参数。当氧化层厚度从 $0.1\mu\text{m}$ 增加到 $0.15\mu\text{m}$ ，通过将 p_{well} 注入剂量从 $5.5 \times 10^{13} \text{ cm}^{-2}$ 降低到 $3.5 \times 10^{13} \text{ cm}^{-2}$ 使 VT 参数满足要求，同时 di/dt 由原本的 5879 A/us 降低到 3542 A/us ，降低 40%；当 JFET 剂量从 $2 \times 10^{12} \text{ cm}^{-2}$ 增加到 $1 \times 10^{13} \text{ cm}^{-2}$ 时， di/dt 从由原本的 4995 A/us 降低到 2419 A/us ，降低 52%，同时 BV 未退化， R_{onsp} 降低了 5%，略有优化。仅通过工艺优化，极大的降低了 di/dt ，达到优化 EMI 特性的目的，对超结 MOSFET 器件的设计、制造具有很大的实用价值。

参考文献

- [1] Chen X. Semiconductor power devices with alternating conductivity type high-voltage breakdown regions: U.S. Patent 5,216,275[P]. 1993-6-1.
- [2] Deboy G, Marz N, Stengl J P, et al. A new generation of high voltage MOSFETs breaks the limit line of silicon[C]. International Electron Devices Meeting 1998. Technical Digest (Cat. No. 98CH36217). IEEE, 1998: 683-685.
- [3] Musumeci S, Portoghese F, Raciti A, et al. Super junction MOSFET in power factor correction converters[C]. 2017 AEIT International Annual Conference. IEEE, 2017: 1-6.
- [4] Fang Lin L, Hong Y. Investigation of EMI, EMS and EMC in power DC/DC converters[C]. The Fifth International Conference on Power Electronics and Drive Systems, 2003 PEDS 2003, 2003:572-577 Vol.571.
- [5] 颜伟.复杂电子系统电磁干扰噪声机理建模与抑制方法研究[D].南京:南京师范大学, 2014.
- [6] 吴晓, 霍宏艳, 丁少华. 新版 GB13837 标准解析及项目验证[J].电视技术, 2012, 36(20):102-104+108.
- [7] Caka N, Zabeli M, Limani M, et al. Impact of MOSFET parameters on its parasitic capacitances[C]. Proc. 6th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications, Stevens Point, Wisconsin, USA. 2007: 55-59.
- [8] Baliga B J. Fundamentals of power semiconductor devices[M]. Springer, 2018.

主办单位:



中国仿真学会
China Simulation Federation
集成微系统建模与仿真专业委员会
Committee of Integrated Microsystem Modeling and Simulation

协办单位:

电子科技大学
电子科技大学先进毫米波技术集成攻关研究院
杭州电子科技大学射频电路与系统教育部重点实验室
江苏省多维感知信息技术联合实验室
宽带微波电路高密度集成四川省工程研究中心
雷达探测感知全国重点实验室
微组装技术创新中心
浙江大学信息与电子工程学院
浙江省半导体行业协会

合作企业:



微波学报



宽带微波电路高密度集成
四川省工程研究中心

