航天微电子
AEROSPACE MICROELECTRONICS

第2期 2022年6月
总第5期

目录 CONTENTS

军工领域嵌入式操作系统应用情况及我国自主发展的思考

周海洋，郭 涛

（北京微电子技术研究所，北京市 100076）

摘　要：本文以军工领域嵌入式操作系统为研究对象，搜集、整理了国内外应用情况，分析、说明了主要特点和国内外差距；之后，以新一代航天领域用处理器及配套软件为例，说明了嵌入式操作系统的未来发展趋势；最后，在此基础上说明了关于我国军工领域嵌入式操作系统自主发展的思考。

关键词：军工；嵌入式操作系统；自主发展；思考

中图分类号：TP316　　**文献标识码**：A

Application of Embedded Operating System in Military Industry and Suggestions to Chinese Autonomous and Controllable Development

Zhou Haiyang, Guo Tao

(Beijing Microelectronics Technology Institute, Beijing, 100076, China)

Abstract: Taking the embedded operating system in military industry as the research object , this paper collected and organizeed the domestic and foreign applications, analyzed and explained its main characteristics and the gap between domestic and foreign. Then, the future development of embedded operating system is illustrated by taking the new generation of aerospace processors and supporting software as an example. Finally, on the basis, the suggestions of Chinese autonomous and controllable development of embedded operating systems in military industry are explained.

Key words: military industry; embedded operating system; autonomous and controllable; suggestion

0 引言

嵌入式操作系统（Embedded Operating System）是一种运行在嵌入式系统硬件上的基础软件，其基本功能是对硬件进行有效管理并对硬件进行一定程度的抽象以便应用软件调用。在军工领域，嵌入式操作系统在具备基本功能的基础上，还需要具有实时性（Real-time）、安全性（Security & Safety）等特点。随着应用系统功能越来越复杂，军工产品应用嵌入式操作系统正逐渐成为设计必选项。

本文首先整理了国内外主要军用嵌入式操作系统应用情况，分析了国内外军用嵌入式操作系统的特点和差距；之后，在说明国外新一代嵌入式处理器软硬件特征的基础上，介绍了军工领域嵌入式操作系统的主要发展趋势；最后，基于上述分析，提出了我国军工领域嵌入式操作系统自主发展的一些思考。

1 国外主要嵌入式操作系统及应用情况

1.1 美国、欧洲、日本在军工领域嵌入式操作系统使用情况

美国、欧洲、日本在全球军工领域具有领先地位，多种嵌入式操作系统应用在多种飞行器、武器系统中，如表 1、表 2 和表 3 所示。

表 1 美国在军工领域主要嵌入式操作系统使用情况

Tab.1 The use of major embedded operating systems in the military industry in the United States

序号	名称	供应商	主要应用主体	典型应用领域
1	VxWorks	Wind River	NASA，Boeing，Northrup Grumman	InSight 火星着陆器[1]；NASA“好奇号”火星探测车；Boring787 安全系统；空中无人作战系统[2]
2	Integrity	Green Hills	Northrop Grumman	美国海军陆战队 UH-1Y 和 AH-1Z 直升机升级航电系统[3]；美军 GPS[4]；空客 A380[5]；美国陆军联合战术无线电系统（Joint Tactical Radio System，JTRS）[6]；美国陆军航空兵 CH-47 Chinook、AH-64 Apache 和 UH-60 Black Hawk[7]；美国空军电子战控制器 AN/ALQ-213(v5)[8]；美国空军 F-22 GPS 导航系统[9]
3	μC/OS	Micrium	NASA	“好奇号”火星车[10]
4	RedHawk Linux	concurrent	NASA	空间发射系统(Space Launch System，SLS)；B-1B 轰炸机武器控制系统；联合火箭发射联盟计划；陆基中段防御（Ground-based Midcourse Defense，GMD）计划[11]
5	ThreadX	Microsoft	NASA	深度撞击（Deep Impact）[12]
6	FreeRTOS	OpenSource	NASA	立方星[13]
7	RTEMS	OpenSource	NASA	高性能航天器计算（High-Performance Spaceflight Computing，HPSC）项目配套嵌入式操作系统[14]
8	Linux	OpenSource	SpaceX	Falcon 运载火箭；Dragon 太空舱[15,16]
9	LynxOS	LYNX	美国国防部，Boeing，NASA，Raytheon	武器[17]；Boeing777 客舱服务系统；NASA SLR2000 卫星测距系统；MK 57 发射系统[3]
10	Mbed OS	ARM	N/A	纳米卫星（Tel-USat）[18,19]
11	PikeOS	Sysgo	N/A	航电系统[20]
12	YoctoLinux	OpenSource	NASA，空军研究实验室（Air Force Research Laboratory，AFRL）	高性能航天器计算（High-Performance Spaceflight Computing，HPSC）项目配套嵌入式操作系统[14]
13	Deos	DDC-I	N/A	航电系统[21]

表 2 欧洲在军工领域主要嵌入式操作系统使用情况

Tab.2 The use of major embedded operating systems in the military industry in Europe

序号	名称	供应商	主要应用主体	典型应用领域
1	Deos	DDC-I	ESA	航电系统[21]
2	LithOS	fentISS	欧盟	De-RISC 项目[22]
3	PikeOS	Sysgo	N/A	航电系统[23]

表 3 日本在军工领域主要嵌入式操作系统使用情况

Tab.3 The use of major embedded operating systems in the military industry in Japan

序号	名称	供应商	主要应用主体	典型应用领域
1	Integrity	Green Hills	JAXA	H-IIA/H-IIB 运载火箭制导 / 控制系统[24]
2	TOPPERS/HRP	JAXA 和名古屋大学	JAXA	“瞳”卫星[25,26]
3	Esol RTOS	Esol	JAXA	太空立方体[27]
4	TRON	NEC 和无人空间实验自由飞行器研究所（Institute for Unmanned Space Experiment Free Flyer，USEF）	Japan	具有新观测系统架构的先进卫星（Advanced Satellite with New system ARchitecture for Observation，ASNARO）项目[28]

1.2 国外主要军工领域嵌入式操作系统特点

（1）应用嵌入式操作系统“百花齐放”，商业产品和开源产品并举

经统计，美国、欧洲、日本在军工领域使用的嵌入式操作系统不少于17种，其中商业产品13种（含商业公司维护的开源操作系统，表1中第1、2、3、4、5、9、10、11、13行所示，表2中第2行所示，表3中第2、3、4行所示），开源社区维护的产品4种（表1中第6、7、8、12行所示）。在商业产品中，Wind River公司的VxWorks和Green Hills公司的Integrity两者应用案例最多，分布于航天（如：NASA好奇号火星探测器）、航空（如：波音787）、空军（如：电子战控制器）、海军（如：AH-1Z武装直升机）、陆军（AH-64武装直升机）等诸多领域。在开源社区维护的产品中，FreeRTOS和RTEMS的应用案例较多。特别是RTEMS，其在NASA太阳动力观测站、LISA探路者、维纳斯快车等多项航天任务中得到应用。

除了大量使用具有实时性的嵌入式操作系统外，以SpaceX为代表的航天新势力也在尝试使用传统上认为不适合航天等高可靠场景应用的非实时嵌入式操作系统，这其中的代表是Linux。Linux作为极具代表性的开源项目，具有能够快速吸收并应用新技术、生态环境良好等显著特点。相比于VxWorks、RTEMS等其他嵌入式操作系统，Linux的开发资源要丰富许多，这降低了基于Linux开发应用程序的难度。良好的生态环境、大量的工程师，为Linux进入以航天应用为典型代表的军工领域奠定了基础。

同时，无论是已经大量应用的众多嵌入式操作系统，还是正在进行适应性修改以利使用的Linux，都依托于军工型号、元器件研制单位操作系统团队强大的自主研制能力。数量众多且高素质的嵌入式操作系统人才是国外军用领域嵌入式操作系统应用“百花齐放”、商业产品和开源产品并举的关键基础之一。

（2）制定行业标准，建立不同嵌入式操作系统和应用程序之间的良好互操作性

不同的嵌入式操作系统采用各自的方式管理计算机系统硬件，如果对上层应用程序也采用不同的抽象方法，将会导致在某一个嵌入式操作系统开发的应用程序无法直接在另一个嵌入式操作系统上运行，这既不利于不同厂商之间合作，也不利于新产品继承老产品的工作成果。为此，在国外嵌入式操作系统领域，制定了相关的行业标准。

常见的行业标准包括以下三个：

① POSIX[29,30]

可移植操作系统接口（Portable Operating System Interface of UNIX，POSIX）由IEEE制定并发布，是为在各种类UNIX操作系统上运行软件而定义的一系列API标准的总称，其正式称呼为IEEE 1003，而国际标准名称为ISO/IEC 9945。实现POSIX标准的库常被称作Pthreads，Linux、Solaris、Windows以及多数嵌入式操作系统都已实现。Pthreads定义了一套C语言的类型、函数与常量，它以pthread.h头文件和一个线程库形式存在。Pthreads中大致共有100个函数调用，全都以“pthread_”开头，并可以分为四类：

- 线程管理：创建线程、等待线程、查询线程状态等；
- 互斥锁（Mutex）：创建、摧毁、锁定、解锁、设置属性等操作；
- 条件变量（Condition Variable）：创建、摧毁、等待、通知、设置与查询属性等操作；
- 同步管理：基于互斥锁协调线程间运行。

② ARINC653[31-33]

航电应用软件标准接口653（Avionics Application Software Standard Interface 653，ARINC653）由美国航空电子工程委员会（Airlines Electronic Engineering Committee，AEEC）于1997年提出，是一种嵌入式操作系统应用程序接口标准，目前是国际上在飞行器软件方面比较通行的软件运行标准。其主要特点是时空分区，如图1所示。

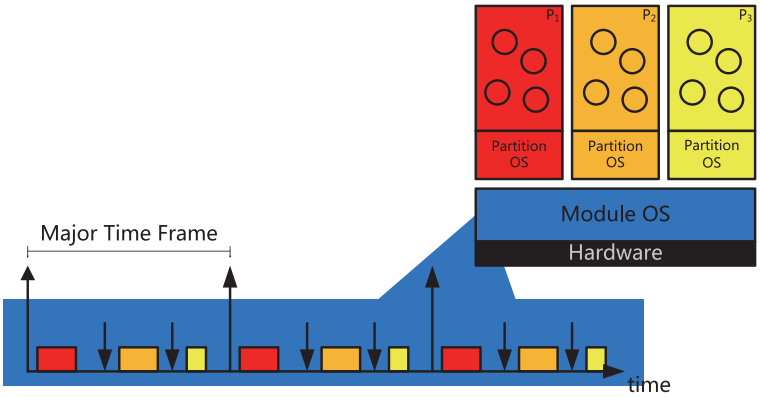


图 1 ARINC 653 时空分区示意图

Fig.1 ARINC653 space-time partition

在 ARINC653 标准中定义了一个主时间帧，再将主时间帧中分成多个小时间段，每个时间段分配给一个应用程序进程执行。随着航空软件系统的执行，主时间帧周而复始运行，使各个应用程序进程都能有效获得硬件资源。同时，由于每个进程只会在分配的小时间段中执行，从而避免了在时间上多个进程同时执行造成的相互影响。与小时间段相对应，利用处理器存储器管理单元（Memory Management Unit，MMU）、存储器控制器分区（bank）控制等硬件技术，每个应用程序进程运行时使用相互独立的一段存储器，从而避免了在存储器空间上多个进程同时执行造成的相互影响。ARINC653 通过使应用软件中的各进程在时间和空间上同时分开获得了较高的软件运行安全性，有效控制了进程发生错误的影响范围，避免了因为某一进程发生错误时威胁到整个航空软件系统运行，进而威胁到飞行器安全飞行的情况发生。

③ FACE[34-36]

未来机载能力环境（Future Airborne Capability Environment，FACE）在 2010 年由美国海军航空系统司令部发起、开源组织（OpenGroup）提出，其策略是在已安装好硬件的军用航电平台上建立软件通用操作环境，使 FACE 组件应用在不同平台上时可被重新部署，从而实现跨平台的可移植性和重用性。

FACE 采用“分段式”架构，自顶向下分为操作系统段、I/O 服务段、平台特定服务段、传输服务段和可移植组件段，每个段间的接口都进行了标准化定义，使得基于 FACE 标准的应用系统可以从任意一个段间接口开始设计具有自身特色功能段。相比 ARINC653 中只定义了应用程序分时分区使用硬件的软件接口，FACE 标准包含了应用程序从顶层通用服务到底层 IO 的全部内容，制定了应用程序各组件的标准化接口，为应用程序赋予了可移植性、开放性和灵活性，大幅提高了电子系统设计的便利性，为航电设备即插即用等应用场景提供了有效技术支撑。

（3）进行资质认证，从研制过程入手提高嵌入式操作系统安全性和可靠性

为使各种嵌入式操作系统在高可靠行业中应用时具有统一的标准，国外制定了相关资质认证标准。以航空领域为例，由美国航空无线电技术委员会（Radio Technical Commission for Aeronautics，RTCA）在 1982 年发布的 DO-178《机载系统和设备合格审定中的软件考虑》（Software Considerations in Airborne Systems and Equipment Certification）是一个典型代表[37-39]。

DO-178 是面向飞机适航性的认证标准，分别在 1985、1992 和 2011 年进行了三次改版，分别称为 DO-178A、DO-178B 和 DO-178C。DO-178 定义了一套飞机用软件开发的过程管理办法，飞机上使用的软件，包括嵌入式操作系统，必须按照 DO-178 规定进行研制，并在能够向飞机管理机构提供证据说明软件研制过程符合 DO-178 要求的情况下，才有可能通过飞机的适航性认证。DO-178 来源于软件工程，

给予了以下三方面的指导：第一，软件生命周期过程的目标；第二，为满足上述目标要进行的活动；第三，证明上述目标已经达到的证据，也即软件生命周期数据。DO-178C 主要过程如图 2 所示。

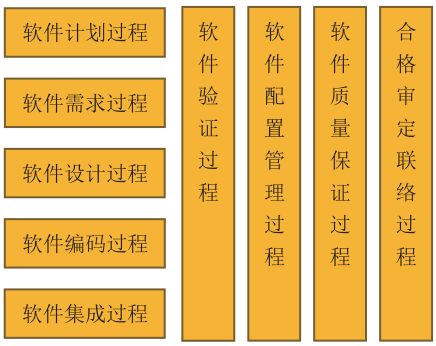


图 2 DO-178C 主要软件过程

Fig.2 DO-178C major software processes

需要说明的是，对嵌入式操作系统进行资质认证是其研制单位、使用客户、行业监管机构等多方意见共同决定的，并非所有应用在高可靠行业中的嵌入式操作系统都必须通过行业资质认证。

2 国内主要嵌入式操作系统及应用情况

2.1 国内自研嵌入式操作系统在军工领域使用情况

随着国内军工领域对嵌入式操作系统自主可控要求的不断提高，国内相关研制单位正在从使用国外引进的嵌入式操作系统转向开源或者国内厂商自研的嵌入式操作系统。目前能够从公开资料上查到的国产嵌入式操作系统不少于 14 个，其中已经明确应用在军工领域的嵌入式操作系统不少于 9 个，如表 4 和表 5 所示。

表 4 国内自研且有军工领域应用报道的嵌入式操作系统情况

Tab.4 Embedded operating system developed by domestic and reported in military field

序号	名称	供应商	典型应用领域
1	DeltaOS（道）	科银京成	工业控制、舰艇火控、舰载显示 [39-42]
2	SpaceOS（天卓）	航天科技 502 所	星载计算机 [43]
3	SylixOS（翼辉）	南京翼辉	工业控制、飞行器控制 [44,45]、遥感 35 号 C 星、行云二号 01/02 星、瓢虫一号卫星、双曲线一号运载火箭控制器、物联网通讯卫星、固定翼无人机 [46]
4	ReWorks（锐华）	电科 32 所	雷达信号处理、轨道交通、工业控制 [47,48]
5	AcoreOS（天脉）	中航 631 所	飞行器航电 [49,50]
6	Tyche（天熠）	航天科工 706 所	移动通信 [51-53]
7	Tadpole OS（科斗）	智慧海派	网络终端 [54]
8	望获 OS	国科环宇	星载、机载、弹载 [55]
9	RT-Thread	上海睿赛德	星载计算机 [56]、多款发射升空的航天设备 [57]

表 5 国内自研但暂没有军工领域应用报道的嵌入式操作系统情况

Tab.5 Embedded operating system developed by domestic but not reported in military field

序号	名称	供应商	典型应用领域
1	Hopen OS	凯思集团	工业控制、消费类电子、移动通信、智能家居 [58]
2	EEOS	中科院计算所	嵌入式 [59]
3	HBOS	浙江大学	信息家电、智能设备、仪器仪表 [59]
4	μTenux	大连悠龙	汽车电子、医疗电子、工业控制领域 [60]
5	aCoral	电子科大	嵌入式 [61,62]

2.2 国内主要嵌入式操作系统及其应用特点

（1）国产嵌入式操作系统“百家争鸣”

以航天科技、航天科工、航空工业、中国电科等国营大型企业和以南京翼辉、上海睿赛德等私营企业为代表的许多企业和单位都开展了国产嵌入式操作系统的研制和商用工作。技术上，国产嵌入式操作系统采用了各不相同的技术路线，例如：RT-Thread为了获得较快的运行速度和较小的体积而选用了微内核技术，望获 OS 为了获得较好的应用程序兼容性而选择在 Linux 内核的基础上增加硬实时能力实现实时嵌入式操作系统。应用上，国产嵌入式操作系统在IoT、工业控制、航空航天等领域均得到了较为广泛的应用，比如：RT-Thread 在 2021 年全球装机量已经超过 8 亿台，同时也已经应用在航天设备中。国产嵌入式操作系统，无论是依托于国家型号任务需求而研制，还是面向商业市场需要而开发，都在目标应用领域用中获得了比较好的使用，并通过使用得到不断完善，可谓“百家争鸣”。

（2）开放标准及资质认证方面尚有差距，需要追赶

国产嵌入式操作系统产品逐渐多样、应用日渐广泛，但是相比于国外嵌入式操作系统，在开发国家标准及资质认证方面的公开报道还非常罕见。国产嵌入式操作系统一般会定义一套自主的 API，通常部分兼容 POSIX 标准。这导致应用程序在不同嵌入式操作系统之间的可移植性较差，一定程度上封闭了某种嵌入式操作系统的生态环境，从长期发展看，不利于这种嵌入式操作系统向更多应用领域推广。此外，国产嵌入式操作系统中宣布支持 ARINC653 标准的产品较少，支持 FACE 标准的报道更少，同时，也少有国产嵌入式操作系统通过 DO-178 等类似国际认证的报道。虽然如前文所述，在军用领域应用嵌入式操作系统并非必须符合某种标准和通过某项认证，但是从行业发展和国际化的角度考虑，支持已有的国际标准和资质认证是有必要的。另外，制定具有我国特色的国外同类标准和资质认证也应当是国产嵌入式操作系统行业在未来发展中考虑的工作。

3 军用领域嵌入式操作系统应用发展

3.1 国外主要军用领域新一代嵌入式处理器

嵌入式操作系统的基本功能是对计算机系统进行管理和抽象，它的发展离不开新一代处理器的发展。以航天领域为例，国外典型的新一代处理器如下：

① HPSC[14]

NASA 从 2018 年主持进行的 HPSC 项目正在研制一种面向宇航应用的新型高性能处理器。该芯片将处理器核心划分为三个功能域，其中由 2 个分别包含 4 个 ARM Cortex-A53 核的簇负责高性能计算，由 2 个 ARM Cortex-R52 核和 1 个 ARM Cortex-A53 核负责实时处理，由三模冗余的 ARM Cortex-M4F 负责全片时钟、复位、配置和健康管理。HPSC 的硬件结构如图 3 所示。

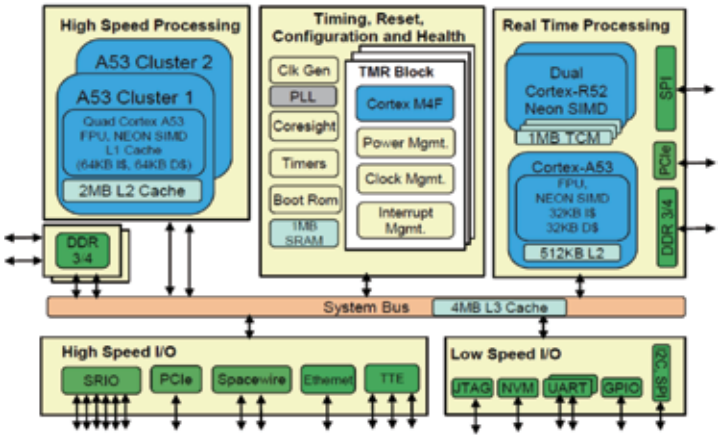


图 3 NASA HPSC 硬件结构图

Fig.3 NASA HPSC hardware architecture

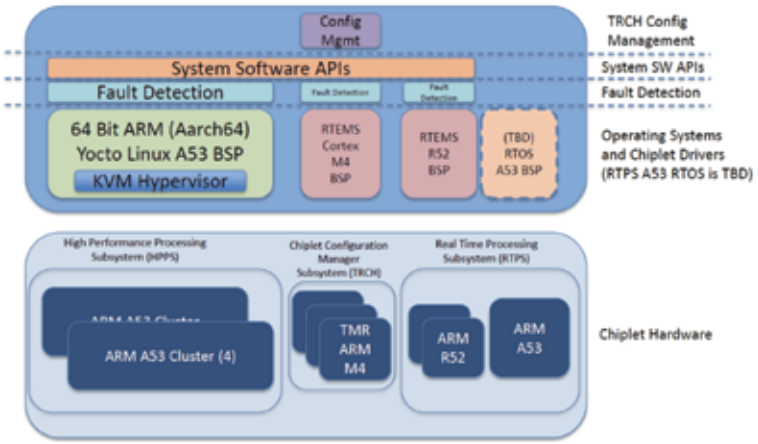


图 4 NASA HPSC 的配套软件系统

Fig.4 NASA HPSC software

与 HPSC 配套的软件系统如图 4 所示。从图中可以看出，HPSC 中运行了两类操作系统，一类是功能相对简单、代码体量较小的嵌入式操作系统，它有多个实体，如 RTEMS，分别运行在负责全片配置和实时处理的 M4、R52 和 A53 核上；另一类是功能相对复杂、代码体量较大的嵌入式操作系统，如 Yocto Linux，它通过 KVM Hypervisor 使用 2 簇 8 核 A53 提供的计算能力，并管理负责高性能计算的多个处理器核。

② Microchip PolarFire SoC FPGA[63,64]

2021 年，美国 Microchip 公司针对未来宇航应用需求，推出了新一代系统芯片 PolarFire SoC FPGA，其硬件模块如图 5 所示。这是一个同时包含处理器和 FPGA 的产品。处理器部分包括 5 个兼容 RISC-V 指令集的核，其中一个处理器核用于全片控制，另外四个处理器核用于高性能计算。芯片内部具有多层高速总线，用于互联处理器、FPGA 和其他片上外设。FPGA 部分除了具有可编程逻辑外，也具有 PCIe、GPIO 等高速接口，便于通过在 FPGA 中的定制逻辑进行功能扩展。

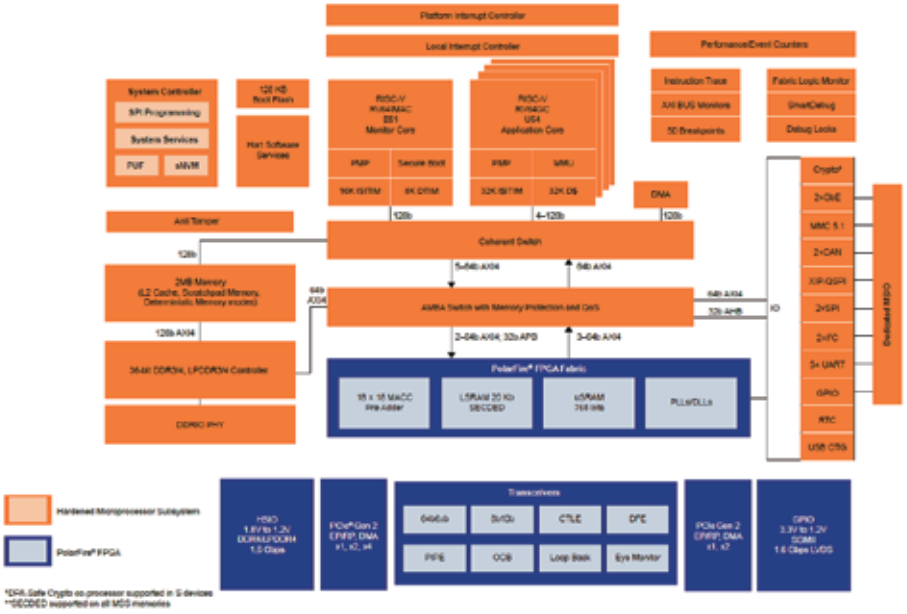


图 5 Microchip PolarFire SoC FPGA 功能框图

Fig.5 Microchip PolarFire SoC FPGA function

Microchip PolarFire SoC FPGA 中的处理器核可以以对称多处理（Symmetric Multi-Processing，SMP）和非对称多处理（Asymmetric Multi-Processing，AMP）两种模式使用。对于非实时应用场景，Microchip 公司建议使用 SMP 模式，此时每个处理器核上都运行非实时的 Linux；对于实时应用场景，Microchip 公司建议根据实际应用需要，在芯片中某几个核上运行包括实时操作系统在内的实时程序，而剩余的处理器核则可继续运行非实时的 Linux。同时，为了保证实时与非实时两类操作系统同时运行而互不影响，需要修改存储器系统设置，分别为两类操作系统配备不同功能的存储器。AMP 方式下 Microchip PolarFire SoC FPGA 软件运行如图 6 所示。

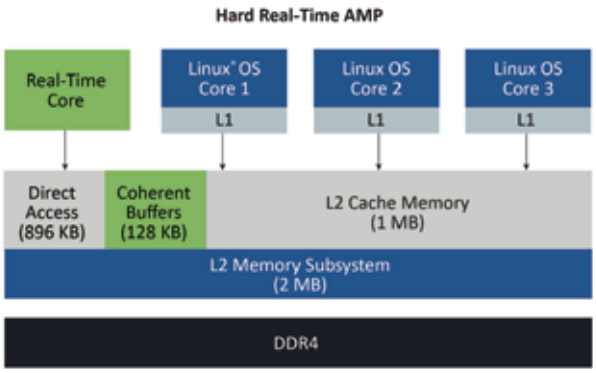


图 6 Microchip PolarFire SoC FPGA AMP 模式软件框图

Fig.6 Microchip PolarFire SoC FPGA software in AMP mode

③ De-RISC[22]

2020 年，欧洲启动“用于安全关键计算机系统的可靠实时基础架构”（Dependable Real-

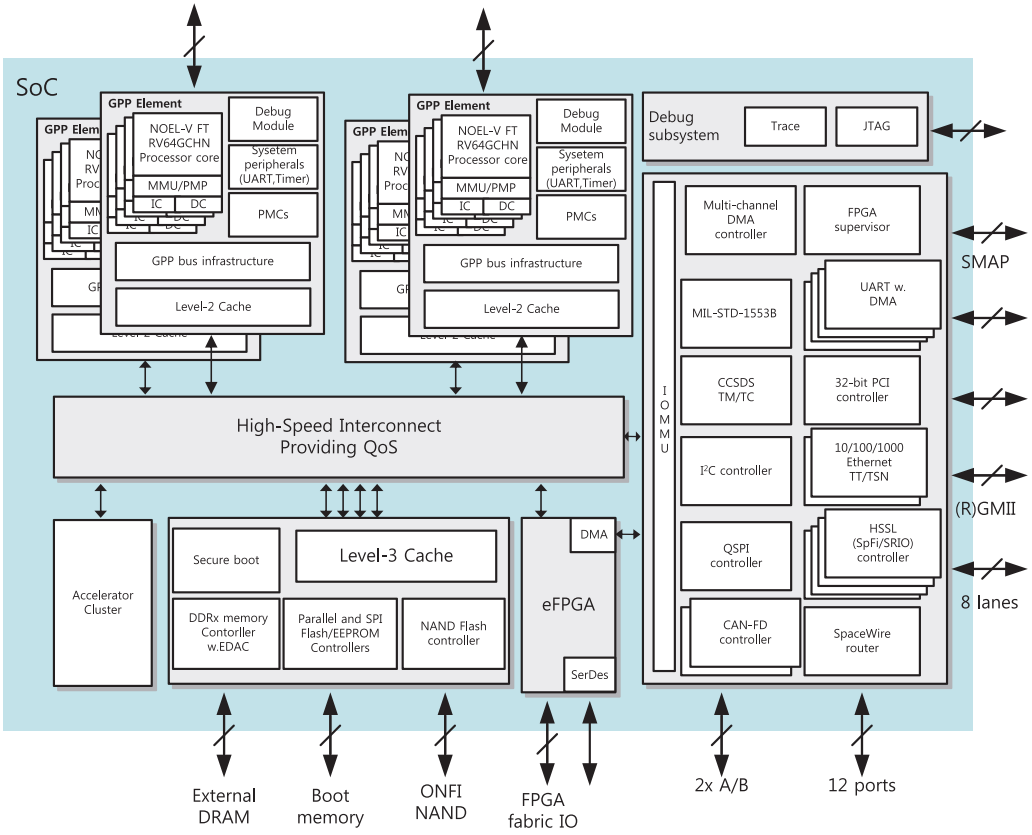


图 7 De-RISC 功能框图

Fig.7 De-RISC function

time Infrastructure for Safety-critical Computer Systems，De-RISC）项目。该项目旨在替换已经在航天领域成功应用的ERC32、AT697F、GR712RC、GR740、RAD750、RAD5545等处理器产品。De-RISC包含了4个处理器簇，每个簇中设计有4个RISC-V指令集兼容的处理器核。其内部采用具有QoS功能的高速互联总线将处理器簇、加速器簇、外设、嵌入式FPGA等模块连接在一起，如图7所示

De-RISC项目的配套软件如图8所示。采用XtratuM Hypervisor对芯片进行控制，以建立能够运行多个不同操作系统的软件环境。

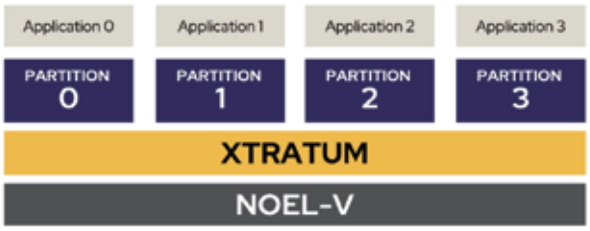


图8 De-RISC配套软件框图
Fig.8 De-RISC software

3.2 与国外主要军工领域新一代嵌入式处理器配套的嵌入式操作系统特点

综合分析上述三款军用领域新一代嵌入式处理器的软硬件特点，可以得到该领域嵌入式操作系统发展的三个趋势。

（1）单处理器上多个嵌入式操作系统同时运行

上述三个典型的军工领域新一代嵌入式处理器都采用了多处理器核设计。其中，HPSC和PolarFire SoC FPGA两者内部集成了不同微架构的处理器核，而De-RISC则集成了相同微架构的处理器核。为了充分发挥多处理器核的并行计算能力，每个处理器核上都部署了嵌入式操作系统。由于各个嵌入式操作系统之间相互独立运行，使得在进行电子系统设计时，可以将原本在多个分布式计算机上实现的应用软件搬移到这些嵌入式操作系统上。对于电子系统来说，在物理上，得益于微电子制造技术的提高，原本多个分布式计算机“集成”到一个多核嵌入式处理器上，单核运算速度、核间通信速率等性能都得以提高，也节省了电路板面积、降低了系统重量；在逻辑上，由于每个处理器核上依旧运行相互独立的嵌入式操作系统，使得原本运行在每个分布式计算机上的应用程序可以“平滑”的移植到这些嵌入式操作系统上。单处理器上同时运行多个嵌入式操作系统，较好的实现了应用软件的继承，降低了新研代码的工作量，同时有效提升了各个应用软件的执行性能。

（2）单处理器上实时、非实时嵌入式操作系统同时运行

传统认识上，出于安全性和可靠性的要求，军工领域使用的嵌入式操作系统通常都具有实时性，从而保证对时间关键任务具有良好的响应能力。然而，以PolarFire SoC FPGA为例，其多个处理器核上有的运行实时操作系统，有的运行非实时的Linux。如果说基于多核处理器的多个嵌入式操作系统同时运行可以实现多项应用程序同时、互不干扰的执行，那么，基于多核处理器的实时、非实时嵌入式操作系统组合运行可以实现时间关键任务和非时间关键任务在单芯片上的并行执行。将多处理器核心产生的强大并行算力同时提供给实时和非实时应用，这对于电子系统来说提供了相当的使用便利和设计自由度。

（3）使用基于多特权硬件的虚拟化技术

对于HPSC、PolarFire SoC FPGA和De-RISC三者来说，它们选择了ARM和RISC-V这两种比较新的指令集体系结构。得益于计算机领域的高速技术发展成果，ARM和RISC-V都定义了四级特权结构，从底向上分别是机器模式（M-mode）、Hypervisor模式（H-mode）、Supervisor模式（S-Mode）和用户模式（U-mode）。每一层特权模式都提供了对于处理器硬件资源不同范围的访问权限。在这种特权模式下，嵌入式操作系统通常运行在Supervisor模式。相比于SPARC V8等只有管理员和用户两种特权模式的指令集体系结构，运行于Supervisor模式的嵌入式操作系统不再直接控制硬件，而是通过Hypervisor模式上运行的虚拟化程序提供的运行时

（run-time）实现对硬件的调用。基于特权硬件的虚拟化技术能够有效利用硬件设计实现对不同特权模式中运行程序访问硬件权限的管控，具有较好的使用安全性，同时，由于每个特权模式中的软件各司其事，也能够使嵌入式操作系统和应用程序获得较好的执行性能。

4 我国军工领域嵌入式操作系统自主发展的思考

根据上述对国内外军工领域嵌入式操作系统应用及发展情况的说明和分析，对我国军用领域嵌入式操作系统自主发展提出如下三点思考：

（1）积极应用开源操作系统，夯实技术基础

从国内外操作系统在军工领域的应用情况看，虽然以 VxWorks、Integrity 等闭源嵌入式操作系统为主，但在以 SpaceX 为代表的航天新势力的不断探索下，开源操作系统应用也在逐渐增加，甚至传统上认为不适合航天领域应用的非实时开源操作系统 Linux，也成功使用在龙飞船、“机智”号火星直升机等宇航型号中。开源嵌入式操作系统具有获得容易、社区支持好等显著优势。特别是 Linux 操作系统，其技术发展在操作系统领域具有代表性和引领性，它的 API 接口遵循 POSIX 标准，具有良好的兼容性。开展开源操作系统应用，在为国产高性能处理器配套操作系统的同时，也能够利用开源操作系统全部代码可见的特点，开展操作系统任务调度、内存分配等核心机制研究，分析不同算法的特点，进而与高性能处理器硬件设计相结合，实现软硬件协同优化，提高嵌入式计算机核心算力。同时，基于开源操作系统形成的芯片驱动，既可以作为裸机上运行的基本操作函数，降低用户使用高性能处理器的难度，也可以作为移植其他闭源操作系统的参考，实现用户所需闭源操作系统的快速、正确应用。因此，积极开展开源操作系统应用，可以有效夯实嵌入式操作系统领域的技术基础，有效支持软硬件设计和应用系统研制。

（2）国内嵌入式操作系统厂商和元器件厂商应主动相互接触，共同拓展应用场景

近年，中兴遭罚款、华为被制裁的案例再次敲响了我国型号任务自主可控的警钟，国内军工型号承研单位越来越多的开始评估、使用国产嵌入式操作系统。除了传统军工企业研制的 SpaceOS、天脉等已经得到应用的嵌入式操作系统外，型号任务研制单位也在开始评估以南京翼辉、上海睿赛德为首的商业公司的嵌入式操作系统产品。技术上，国产嵌入式操作系统随着应用场景不断丰富、应用数量不断增长，其成熟度得到较快进步，以多个型号任务的成功证明了自身具备了替代国外嵌入式操作系统的能力；应用上，嵌入式操作系统作为连接用户应用程序和底层处理器硬件的关键软件，对于简化用户管理硬件难度、集中精力到应用程序开发上具有明显帮助。未来会有越来越多的型号任务选用国产嵌入式操作系统。对于嵌入式操作系统厂商和元器件厂商来说，两者的目标市场具有一定的重叠性，因此，双方应主动接触，尽早、尽可能多的实现国产嵌入式操作系统与国产高性能处理器的适配，为型号任务研制方提供软硬件一体的多种应用解决方案，既提供了国产化系统设计基础，也拓展国产高性能处理器和嵌入式操作系统的应用场景。

（3）开展 Hypervisor 等新技术研究，服务新一代高性能处理器应用

随着微电子技术的进步，单片集成多核处理器、FPGA 已经成为新一代军工高性能处理器的发展趋势。在此趋势下，嵌入式操作系统为了实现对复杂硬件的管理，功能向下延伸，产生了 Hypervisor，并通过 Hypervisor 的虚拟化功能，形成了在单芯片上同时运行多个操作系统的能力，这既拓展了系统设计的方式，也能充分发挥芯片中多处理器核并行计算的优势。因此，伴随新一代高性能处理器的研制，应开展 Hypervisor 等新技术研究，充分发挥新型处理器高算力的主要优点，为新一代型号任务提供更灵活、更多样、更强力的应用支撑。

5 结束语

本文对国内外军工领域嵌入式操作系统的应用情况和特点进行说明和分析，并以国外新一代航天领域处理器及配套软件为例，分析了军工领域嵌入式操作系统的发展趋势，在此基础上提出了对我国军工领域嵌入式操作系统发展的一些思考，可以作为相关领

域工作的参考。

参考文献（References）

[1] NASA JPL. Spacecraft[EB/OL]. [2021-12-21]. https://www.jpl.nasa.gov/news/press_kits/insight/launch/mission/spacecraft/.

[2] GARCIA L. Real-time operating systems case study ~ LynxOS vs. VxWorks[R]. Florida Atlantic University, 2017-12.

[3] Green Hills Software. New Green Hills Software for H-1Y/Z upgrades[EB/OL]. (2013-2-13) [2021-12-21]. https://www.helis.com/database/news/h1y_software/.

[4] Green Hills Software. Green Hills Software delivers RTOS for military M-code GPS ASIC[EB/OL]. [2021-09-09]. https://www.embedded.com/green-hills-software-delivers-rtos-for-military-m-code-gps-asic.html.

[5] Northrop Grumman Selects Green Hills Software's INTEGRITY®-178B RTOS For Use In Navigation System For Airbus And Other Airframe Manufacturers[EB/OL]. [2003-01-29]. https://www.ghs.com/news/230129n.html.

[6] Boeing Selects Green Hills Software's INTEGRITY RTOS for Software-Defined Radio Programs [EB/OL]. [2003-11-18]. https://www.ghs.com/news/20031118_boeing.html.

[7] U.S. Army Depends on INTEGRITY-178 tuMP RTOS for Improved Data Modem[EB/OL]. [2021-12-12]. https://www.ghs.com/news/20201112_us_army_idm_tump.html.

[8] Terma Selects INTEGRITY-178 tuMP RTOS for Next-Gen Electronic Warfare Controller[EB/OL]. [2021-12-19]. https://www.ghs.com/news/20210219_terma_warfare_controller_tump.html.

[9] Woodrow Bellamy III. Northrop Grumman's E-2D, F-22 Embedded GPS System Upgrade to Run on INTEGRITY-178 RTOS[EB/OL]. (2020-3-10) [2021-12-21]. https://www.aviationtoday.com/2020/03/10/northrop-grummans-e-2d-f-22-embedded-gps-system-upgrade-run-integrity-178-rtos/.

[10] nicholasldf. uCOS-II第一次离开地球吸引力，随Curiosity号登陆火星，负责火星样本分析实验室的控制[EB/OL]. (2012-12-18) [2021-12-21]. https://bbs.21ic.com/icview-404141-1-1.html?fromuser=.

[11] Concurrent Real-Time. Aerospace & Defense[EB/OL]. [2021-12-21]. https://www.concurrent-rt.com/industries/aerospace-defense/.

[12] WILLIAM LAMIE. Case study: NASA's "Deep Impact" employs embedded systems to score bullseye 80 million miles away[EB/OL]. (2006-1-13) [2021-12-21]. https://militaryembedded.com/comms/rf-and-microwave/case-bullseye-million-miles-away.

[13] QUIROS J O D, Duncan d' Hemecourt. Development of a flight software framework for student CubeSat missions[J]. Tecnología en marcha, Edición especial, 2019(Movilidad Estudiantil 6):180-197.

[14] Alan Cudmore. High-Performance Spaceflight Computing (HPSC) Middleware Overview[EB/OL]. (2019) [2021-12-21]. https://ntrs.nasa.gov/api/citations/20190001377/downloads/20190001377.pdf?attachment=true.

[15] Liam Tung. SpaceX: We've launched 32,000 Linux computers into space for Starlink internet[EB/OL]. (2020-6-8) [2021-12-21]. https://www.zdnet.com/article/spacex-weve-launched-32000-linux-computers-into-space-for-starlink-internet/.

[16] 贾浩楠，萧箫．人类刚给火星送去Linux系统，以及一款安卓手机芯片[EB/OL]. (2021-02-19) [2021-12-21]. https://www.163.com/dy/article/G36V8E030511DSSR.html.

[17] Securing connected embedded devices using built-in RTOS security[EB/OL]. [2015]. https://militaryembedded.com/cyber/cybersecurity/securing-connected-embedded-devices-using-built-in-rtos-security.html.

[18] HARFIAN A R, ARSENO D, EDWAR E, et al. An Investigation of RTOS-Based Sensor Data Management Performance for Tel-USat On Board Data Handling (OBDH) Subsystem[C]//2019 International Conference on

Information and Communications Technology (ICOIACT). 2019.

[19] PUTRA A C A Y, WIJANTO H, EDWAR. Design and Implementation RTOS (Real Time Operating System) as a Nano Satellite Control for Responding to Space Environmental Conditions[C]//2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob). IEEE, 2021.

[20] KOSMIDIS L, MAXIM C, J V, et al. Industrial experiences with resource management under software randomization in ARINC653 avionics environments[C]// the International Conference. 2018.

[21] Safety Critical RTOS for Avionics Applications requiring DO178C/ED-12C DAL A verification[EB/OL]. https://www.ddci.com/products_deos_do_178c_arinc_653.html.

[22] WESSMANX N, MALATESTAX F, et al. De-RISC: the First RISC-V Space-Grade Platform for Safety-Critical Systems[C].//2021 IEEE Space Computing Conference (SCC), IEEE, 2021:17-26.

[23] Sysgo. Avionics & Defense[EB/OL]. [2021-12-21]. https://www.sysgo.com/avionics.

[24] Japan Aerospace Exploration Agence(JAXA)[EB/OL]. https://ghs.com/customers/jaxa.html .

[25] 高信 性RTOSについて[EB/OL]. [2012]. https://rtos.jaxa.jp/top_rtos.html

[26] JAXA. 宇宙 搭 用リアルタイムOS ~ TOPPERS/HRP カーネル、Safety カーネル~[EB/OL]. (2011-12-15)[2021-12-21]. https://rtos.jaxa.jp/brochure.pdf.

[27] Space appliance development platform "Space Cube 2" [Satellite systems] [EB/OL]. https://www.esol.com/cn/successstory/case_36.html#.

[28] ASNARO (Advanced Satellite with New system ARchitecture for Observation)[EB/OL]. https://eoportal.org/web/eoportal/satellite-missions/a/asnaro.html.

[29] SERRA G, ARA G, FARA P, et al. ReTiF: A declarative real-time scheduling framework for POSIX systems[J]. Journal of Systems Architecture,2021,118.

[30] PETR R, ZUZANA B, JAN M, et al. Reproducible execution of POSIX programs with DiOS[J]. Software and Systems Modeling,2020(prepublish).

[31] 雷煜靓，胡宁，张磊．ARINC653实时系统可调度性验证综述[J]．信息技术与信息化，2021(06):25-27.

[32] 雷煜靓，胡宁，陈福，等．ARINC653实时任务可调度性验证方法[J]．单片机与嵌入式系统应用，2021,21(04):15-20.

[33] 杨静远，时磊，张前．一种ARINC653分区操作系统的设备空间配置管理模型[J]．信息技术与信息化，2021(02):94-96.

[34] 郝玉锴，吴姣，牛玥瑶，等．开放式软件架构的组织和构件管理方法研究[J]．信息技术与信息化，2021(11):144-146.

[35] 王亮，王璇，谢博琳．基于FACE的航电系统软件架构设计[J]．信息技术与信息化，2021(10):125-127.

[36] The Open Group. Future Airborne Capability Environment[EB/OL]. [2021-12-21]. https://www.opengroup.org/face.

[37] 谭莉娟，郑巍，刘友林，等．面向适航标准的机载软件测试验证方法综述[J]．计算机工程与应用，2021,57(15):9-22.

[38] 刘文，张道泽，王青，等．机载软件质量保证过程研究[J]．航空计算技术，2021,51(03):90-92+97.

[39] 武小舟，邹勇．嵌入式实时操作系统DeltaOS在导弹指挥计算机系统中的应用研究[J]．弹箭与制导学报，2008(05):201-204.

[40] 梁璐．基于DeltaOS系统的雷达动目标测量方法[J]．火控雷达技术，2021,50(02):36-39.

[41] 通用版操作系统[EB/OL]. [2018-11-17]. http://www.coretek.com.cn/index.php?m=book&f=read&articleID=11.html.

[42] 综合化版操作系统[EB/OL]. [2018-11-01]. http://www.coretek.com.cn/index.php?m=book&f=read&articleID=14.html.

[43] SpaceOS操作系统，中国航天造[EB/OL]. [2013-12-17]. http://zhuanti.spacechina.com/n561816/n562385/n563539/c605291/content.html.

[44] 焦进星．SylixOs的来龙去脉[J]．软件和集成电路，2018(07):68-69.

[45] 程文博，屈艺，吴盘龙，等．SylixOS平台下的火控实时解算与实现[J]．兵器装备工程学报，2020,41(10):29-34.

[46] 翼辉信息应用案例[EB/OL]. https://www.acoinfo.com/solution/case/case-aviation/?category=41&subCategory=24&curCategory=731.

[47] 仵引波，马俊臣．基于锐华和麒麟操作系统应用软件开发及测试系统设计[J]．中国科技信息，2021(21):89-90.

[48] 李浩正，罗利强，周游，等．基于锐华嵌入式实时操作系统雷达数据处理软件设计[J]．火控雷达技术，2018，47(01):54-57.

[49] 张斌．国产天脉1型操作系统的嵌入式软件开发配置[J]．单片机与嵌入式系统应用，2021,21(05):12-15.

[50] 张斌．基于天脉1型嵌入式操作系统光纤航姿软件开发[J]．电子产品世界，2021,28(03):70-73.

[51] “天熠”移动操作系统研制成功[J]．黑龙江科技信息，2012(11):10.

[52] 航天科工推出“三大三小”重器，国产信息技术产品再上新台阶[EB/OL]．[201-07-30].https://blog.csdn.net/Z1Y492Vn3ZYD9et3B06/article/details/81294499.html.

[53] 胡海明，周楠，龚宬，等．JFFS2文件系统在天熠操作系统中的实现[J]．计算机工程与设计，2017，38(12):3461-3467+3474.

[54] 老杳．ASR与航天科工通信技术研究院等达成合作意向，共同研发安全终端[EB/OL]．2019-2-2[2021-12-21]. https://zhuanlan.zhihu.com/p/56185827.

[55] 北京国科环宇科技股份有限公司．望获国产操作系统[EB/OL]．[2021-12-21]．http://www.ucas.com.cn/index.php?m=content&c=index&a=lists&catid=291.

[56] 上海睿赛德电子科技有限公司．RT-Thread[EB/OL]．[2021-12-21]．https://www.rt-thread.org/.

[57] 上海宇航系统工程研究所与RT-Thread签署战略合作[EB/OL]．[2022-04-11]．https://www.rt-thread.org/news/203.html.

[58] Hopen OS[EB/OL]．http://www.hopen.com.cn/a/product/hopen_os.html.

[59] cyhong826．国内著名的实时操作系统[EB/OL]．(2009-10-22) [2021-12-21]．https://blog.csdn.net/cyhong826/article/details/4713684.

[60] μTenux[EB/OL]．https://codingdict.com/os/software/89779.html.

[61] 姜良重．aCoral嵌入式操作系统多核调度机制优化设计与研究[D]．电子科技大学，2021.

[62] 李剑．嵌入式多核代码分析器研究与实现[D]．电子科技大学，2014.

[63] Microchip．PolarFire® SoC FPGAs[EB/OL]．[2021-12-21]．https://www.microchip.com/en-us/products/fpgas-and-plds/system-on-chip-fpgas/polarfire-soc-fpgas.

[64] Cannizzaro M J，Gretok E W，George A D．RISC-V Benchmarking for Onboard Sensor Processing[C]．2021 IEEE Space Computing Conference (SCC)，IEEE，2021:46-59.

作者简介：

周海洋(1981—)，男，北京人，硕士，研究员，长期从事计算机体系结构、系统芯片、嵌入式软件等领域的研究和工程实现工作。

众核处理器研究技术综述和分析

宋立国
（北京微电子技术研究所，北京市 100076）

摘 要：文章首先介绍了目前众核处理器的发展状况，指出集成的核数已经成为影响处理器性能的关键；然后对国外众核处理器最新研究成果，依据其有益效果，从能效、性能和可靠性三个方面，开展归纳与分析，对研究成果中涉及到的体系结构、片上存储和软件调度与编译等创新技术进行了重点阐述；文章最后结合后摩尔时代集成电路发展趋势，指出自适应技术、三维集成技术、异构集成技术将是未来众核处理器发展的重点。

关键词：众核处理器；片上网络；存储结构；软件调度
中图分类号：TP368 **文献标识码**：A

Summary and Analysis of Research Technologies on Many-Core Processor

Song Liguo
(Beijing Microelectronics Technology Institute, Beijing, 100076, China)

Abstract: Processors have been developing from single-core to many-core. The latest research results abroad on many-core are comprehensively analyzed. The development status quo is first introduced about many-core processors, and then the related recent papers are summarized from three aspects: architecture, on-chip storage and software scheduling. The main contributions and basic ideas of these papers are analyzed from the perspectives of energy efficiency, performance and reliability. Finally, combined with the development trend of integrated circuits in the post Moore era, three main technical directions which are the emerging adaptive architecture technology, three-dimensional integration technology and heterogeneous integration technology of many-core processors are expounded.

Key words: many-core processor; network-on-chip; memory-on-chip; software scheduling

0 引言

众核处理器，芯片上集成了数十甚至更多的处理器核，在保持单个处理器核工作频率基本不变的情况下，处理器的理论计算性能随核数的增加而提高，既能够适应对高性能处理器的应用需求，又能够适应后摩尔时代延续摩尔（More Moore）和超越摩尔（More than Moore）的技术发展趋势，是处理器领域的未来发展方向。但目前的众核处理器相关研究论文，普遍是对某一专项技术进行阐述，缺乏综合性的文章对众核处理器国内外研究动态进行系统、全面的阐述，因此，本文首先对国内外最新众核处理器产品进行了分析；然后对国内外最新研究成果，从能效、性能和可靠性三个方面进行了综合分析；最后指出众核处理器技术未来发展方向。

1 众核处理器发展

众核处理器始于通用图形处理器（General-Purpose Graphics Processing Unit，GPGPU），实现浮点矩阵乘等矩阵算法并开始应用于传统的科学工程计算领域。同一时期，IBM 研发的 Cyclops-64 众核处理器、CELL 处理器，对业界产生了巨大的影响。随着众核处理器体系结构的持续改进，其适应性和好用性不断提高，目前已成为当前支持高性能技术、人工智能的关键核心器件。根据计算核心的结构复杂度和组织方式，众核处理器分为基于通用处理器核和基于计算簇的众核处理器两大类：

（1）基于通用处理核的众核处理器可以看作是多核结构处理器的进一步延伸。此时集成的处理器核一般由通用处理器简化而来，所有核心功能齐全，并保留通用处理器中的多级缓存（Cache）存储结构。典型代表产品包括：Intel 公司 Ice Lake-SP 架构的至强系列处理器，AMD 公司 Zen 架构的锐龙系列处理器，Kalray 公司 MPPA 系列处理器，富士通公司的 A64FX 处理器，我国飞腾公司基于自研 FTC662、FTC663 内核的 FT-2000+/64 处理器和 S2500 处理器，产品综合指标对比如表 1 所示。

表 1 基于通用处理器核的众核处理器对比

Tab.1 Comparison of many-core based on general processor core

处理器型号	工艺	核数	性能	功耗	发布时间
至强 8360	10nm	36	2.4GHz	250W	2021
酷睿 i9-9980	14nm	18	2.4GHz	45W	2019
锐龙 3990X	7nm	64	2.9GHz	280W	2020
MPPA3	16nm	80	1.2GHz	20W	2019
A64FX	7nm	52	2.2GHz	/	2019
S2500	16nm	64	2.2GHz	150W	2020

（2）基于计算簇的众核处理器，旨在通过简单运算单元的聚合提供超高计算性能。多个运算单元以组或簇的形式进行组织，计算簇内所有计算核心共用指令发射单元，共享一级 Cache 等存储资源；计算簇间则共享二级 Cache 和主存等。典型代表产品主要包括：NVIDIA 公司的 Ampere、Turing 架构系列 GPGPU，AMD 公司的 RDNA2 架构 GPU，日本 PEZY 公司的 PEZY-SC 系列处理器，我国的申威 26010 处理器，产品综合指标对比如表 2 所示。

表 2 基于计算簇的众核处理器对比

Tab.2 Comparison of many-core based on calculation cluster

处理器型号	工艺	核数	性能	功耗	发布时间
RTX3090	7nm	10496	285TFLOPS (FP16)	350W	2021 年
GA100	7nm	6912	624TOPS (INT8)	400W	2021 年
W6800	7nm	3840	35.6TFLOPS (FP16)	250W	2021 年
PEZY-SC3	7nm	8192	43.6TFLOPS	400W	2019 年
SW26010	28nm	260	3TFLOPS (FP32)	/	2016 年

从上面两个表中可以看到，随着工艺的进步，芯片中集成处理器核数量明显有不断增加的趋势，基于通用处理核的众核处理器，采用统一的指令集系统，兼顾应用的好用性和性能；基于计算簇的众核处理器集成的处理器核数量更多，峰值计算能力强大。集成的核数已经成为影响处理器性能的关键因素，众核架构是处理器发展的必然。

2 众核处理器重点优化方向分析

以图灵机理论为基础的冯·诺依曼体系结构是串行执行模型，而众核处理器则属于分布式并行结构，如何在解决二者不匹配的同时，降低处理器性能墙、功耗墙、存储墙、应用墙的制约，是众核处理器研究的关键，本节从能效、性能和可靠性三方面对近期众核相关研究论文进行综合分析。

2.1 众核处理器能效优化技术

众核处理器面临严峻的功耗墙挑战，伴随着 Dennard 缩放比例定律的失效，虽然晶体管的密度仍然会随工艺进步不断提升，但是每代晶体管能量优化的速率在快速降低。因此，众核处理器必须依据系统目标和关键应用需求，针对功耗问题，从多个层次上进行优化设计。

2.1.1 体系结构级能效优化

众核处理器体系结构主要包括处理器核、片上网络（NoC）两个主要部分，因此，众核处理器体系结构级能效优化主要围绕这两部分展开。

（1）处理器核能效优化

处理器的高能效技术按照作用的层次可分为系统级、结构级、电路级和工艺级。系统级主要通过软硬协同的方式，根据负载情况进行能耗管理，实现运行、休眠等不同运行等级状态的切换；结构级通过选择面向能效优化的算法和编码，在保证一定性能的前提下，控制芯片的峰值功耗和运行功耗；电路级主要是针对确定功能的部件，选择能效最优的电路实现；工艺级需要密切结合工艺情况，采用合适的晶体管和逻辑器件，优化后端设计流程，以降低功耗。动态电

压频率调节技术（DVFS）是系统级能耗管理中的一种先进技术。根据程序特征的实时变化，自适应地调节处理器核的电压和频率，在功耗和性能之间取得折中。对处理器核能效优化，通常采用两种策略：（1）将功耗控制和功耗分配解耦合，使两者独立地根据各自目标进行优化，降低复杂度；（2）建立频率和功耗的关系模型，通过反馈方式指导 DVFS 调节，并通过在线调整模型参数，提高众核处理器面对不同特征应用时的能效适应能力。文献 [1]–[4] 应用 DVFS 技术降低众核处理器功耗。

（2）片上网络能效优化

片上网络的功耗在处理器总功耗中具有很高的比重，如麻省理工学院的 RAW 处理器（16 核）[5]，片上网络功耗为 7.2W，占整个处理器功耗的 36%；Intel 公司的“万亿级芯片”（80 核），片上网络功耗占到总体功耗的 40% [6]。片上网络能效优化主要包括拓扑结构能效优化和路由单元能效优化。

片上网络的拓扑结构定义了网络内节点与链路的布局和互连方式，对网络的功耗、时延、面积等有至关重要的影响。文献 [7] 提出关闭与休眠核连接的片上网络路由器的方法，减小片上网络静态功耗，能够较好地降低片上网络功耗。由于缓存消耗的功耗又占了整个路由器的很大一部分，文献 [8]–[10] 致力于降低、实现无缓存路由器，研究表明无缓存路由器相对于传统的有缓存路由器最多能够节省 39%的片上网络功耗。WANG H 利用拓扑能耗模型 [11]，以二维网格环绕、高维网格 / 环绕等为例，探讨了不同制造工艺对片上网络拓扑能耗的影响，指出不同的制造工艺，对应能耗最优的网络拓扑也不同。PINTO A 研究了分簇对二维网格（Mesh）、二维环绕、胖树、蝶网等网络拓扑的影响 [12]，指出通过流量局部化技术，可减少 20%–40% 的能耗。SOUZA M A 对具有分布式和共享缓存机制的不同片上网络拓扑结构，开展设计空间探索和优化，能够降低 38% 的功耗 [13]。

路由单元是片上网络的重要组成部分，目前对路由单元功耗的研究涵盖了系统级和电路级两个层次：系统级主要是基于历史信息的动态调整电压 / 频率；电路级主要是缩减单元内部缓冲容量和修改交换开关结构。DOPPA J R 针对路由单元中的虚拟通道和链路开展细粒度控制 [14]，LEE D J 在路由单元中增加多功能自适应通道 [2]，根据数据流量动态分配链路和路由单元缓冲空间。BOKHARI H 提出 SUPERNET 结构 [15]，采用双电压 / 频率设计，设计两套路由链路，分别工作于不同的电压和频率，任务运行的过程中，能够根据应用需求，选择不同的路由链路传递信息。SCIONTI A 提出 SDNoC 结构 [16]，路由单元设计有正常、旁路和电源门控三种工作模式，通过动态配置，融合二维网络结构和环形网络结构，以有效降低单元功耗。

2.1.2 片上存储能效优化

片上缓存利用程序、数据的空间局部性和时间局部性，缓解主存与 CPU 处理速度不匹配的问题，因容量大、速度快、访问频繁，成为处理器功耗的主要来源，约占总功耗的 30%–60%[17,18]。在众核处理器中，L1 Cache 与处理器核紧密耦合，组织方式单一，设计的重点是高速；而 L2 及更高层次 Cache，则多为大容量 Cache，组织方式多样 [19]，因此 Cache 资源管理的研究多集中在 L2 及更高层次的 Cache 上。TITOSGIL R 基于路（Way）组合思想，提出缓存一致性目录结构 [20]，每个缓存条目设计有一个指针；对于那些需要多个指针的地址，可以从同一组（Set）中其它空路处获得额外的指针，在最大限度的减少存储开销的同时，又不丧失适应多个共享度的灵活性，提高缓存能效。

相较于 Cache 优化，集成软件控制的 SPM（便签存储器）是一个更为理想的选择。这是因为：（1）从面积角度，同 Cache 相比，由于缺少了用来存储地址的 TAG 存储器和比较地址的 TAG 比较器，硬件实现更加简单，在相同的制造工艺下，SPM 所占面积更小，约为 Cache 的 66%[21]；（2）从能耗角度看，SPM 访问能耗更低，相同工艺条件下约为 Cache 的 60%[22]；（3）从指令执行角度看，使用 Cache 无法预测实际最差工作情况，而 SPM 由于程序员可见，并受其直接控制，因此行为更加确定。正是由于在

面积、能耗、实时性方面的优势，ALVAREZ L针对众核共享存储结构，采用运行时库（Runtime Library）[23]，允许编译器自动产生管理SPM的代码，使得存储器访问操作能够自动转到有效数据地址空间。

非易失性新型存储器件（NVM）具有存储密度高、容量大、功耗低、非易失优点，很有希望取代传统存储器件。目前研究比较活跃的NVM有：铁电存储器（FeRAM）、磁性存储器（MRAM）、自旋转移力矩存储器（STT-RAM）、相变存储器（PCM）、阻变存储器（RRAM）、赛道存储器（DWM）等。然而NVM特有的属性问题，如写的寿命有限、写的功耗大和读写速度不一样等，阻碍NVM在处理器中的广泛使用。KIM N设计SRAM与NVM存储混合的末级缓存结构[24]，在缓存之间、片上存储之间进行优化设计，达到写访问的均衡性，填补片上缓存与片外存储访问之间不断扩大的鸿沟。

2.1.3 软件任务调度能效优化

能耗的本质来源即CMOS电路的开关电容活动，最终是由软件运行情况所决定，能效设计中存在相当大部分的节能空间是硬件级或者微体系结构级所无法涉足的，只有通过软件能效设计技术才能得到解决。现在，动态功耗管理（DPM）和动态电压频率调节（DVFS）等新型硬件能效技术的提出为软件能效管理技术提供了新的优化途径，已经产生大量研究成果。

并行任务调度算法的好坏和处理器的能效密切相关。Wang X Q提出一个可扩展的任务调度框架（WAANSO）[25]，基于小波聚类方法，根据运行给定应用程序的内核数量，将任务集自动映射到处理器核上，与蚁群优化算法、粒子群优化算法相比，WAANSO能够提高的能效在19%和65%之间。CAPOTONDI A对众核处理器上的计算簇进行空间划分[26]，分别负责运行OpenCL和OpenMP编写的核心程序，通过获取必要的硬件信息，建立运行时调度系统，实现对众核上运行程序的能效优化。LE T T基于流最小代价启发式算法[27]，通过任务映射和优化，最小化不同任务间信息传递的路由单元数量。上面的研究都是以处理器核为中心的调度，任务在核心上执行，直到由于共享资源访问而暂停执行，HUANG W H提出以共享资源为中心的调度策略[28]，计算任务首先向共享资源请求数据，直到请求操作挂起后，映射计算任务的处理器核才执行计算，以此提高任务执行的能效。

2.2 众核处理器性能优化技术

众核处理器核心数量的增长、性能的提高，对片上存储层次、片上互连和一致性协议的扩展性都提出了更高要求。目前的研究集中在异构集成、片上网络拓扑、片上缓存和软件编译与调度几个方面。

2.2.1 体系结构级性能优化

众核处理器体系结构级性能优化，主要从异构集成、资源动态组合和片上网络方面开展优化设计。

（1）异构集成与资源动态组合性能优化

异构集成是将不同类型处理器核心集成在一个芯片内，分别处理程序中具有不同特征的代码段，包括集成少量强大的管理核心；集成众多面向计算开发的精简运算核心，高效处理线程级并行、大幅提高芯片性能。DAVIDSON S在芯片结构中集成三种内核[29]，分别是5个支持操作系统的高性能RISC-V内核，496个支持大规模并行运算的精简RISC-V内核，和支持人工智能算法的神经网络加速引擎，性能相比NVIDIA公司的Jetson系列嵌入式GPU提高28倍。

利用众核内部丰富的处理器核资料，针对特定应用，构成高性能虚拟核或者虚拟加速单元，实现资源动态组合。CAPOTONDI A提出多编程模型运行时系统MPM-TRS[26]，对处理器核进行组合，分别负责运行OpenCL和OpenMP编写的虚拟加速事件，达到加速运行性能的目的。Wang X Q建立基于PyTorch学习库的密集和稀疏张量处理框架[25]，框架由三部分构成：①局部DAE机制（访问/执行解耦），芯片内处理器核划分为访问内核和执行内核，在访问内核与执行内核间通过SPM建立软件队列；②脉动DAE机制，利用数据重用，允许多个执行内核共享一个访问内核；③集成硬件访问加速模块，提

高数据吞吐量。访问／执行的解耦和硬件访问加速机制，使得处理器内部访问与执行达到了并行化、流水化，从而提高处理器性能。

（2）网络拓扑性能优化

片上网络是处理器计算核访问存储部件的通路，也是多个计算核协同工作的基础，片上网络性能直接决定了系统性能，关系着同步、通信及访存等开销，选择和设计合适的片上网络拓扑结构，是片上网络研究的关键。片上网络有效地分割和共享全局互连线，一方面提高片上通信的吞吐率，另一方面降低其功耗。例如在 SUN 的 Ultra Sparc T1 使用了一个交叉开关网络连接 8 个核，IBM 的 Cell 处理器使用四个报文交换的环形网络连接 9 个核，而 TILERA 的 64 核 TILE64 使用了五个 Mesh 网格型的网络。目前普遍都采用低阶路由器（环形的节点度是 2，Mesh 的节点度均为 4），能够很容易映射到单一金属层上，布线简单。DESHWAL A 基于数据驱动模型和工程师的经验知识，提出一种用于片上网络拓扑结构的多目标优化搜索框架 [30]，该框架是重复迭代的两阶段优化算法，在每次迭代中，首先基于数据驱动树模型，自动选择目标参数；之后，从选定的起始解开始执行局部搜索，并利用得到的帕累托最优组来更新数据驱动树模型，从仿真结果上看，片上网络性能可以提高 20%。

Mesh 网格拓扑简单、寻径方便、可扩展性好，成为最常用的片上网络互连结构，ABBAS E K 详细分析了二维网格及其扩展结构的片上互连性能 [31]，分析结果表明，Mesh 网格中所有结点在某一个方向（水平或垂直）上实际是一个线性阵列，因此在较大规模网络连接中网络直径较大，传输延时也较大。改善片上网络的平均延时，能够很大程度上提高系统的性能。OGRAS U Y 在 Mesh 网络中增加专用全局连线 [32]，以降低某些长距离全局通信的延时。

通过在路由节点添加流量控制器，建立传输反馈机制，从而降低网络拥塞，减小网络延时。瑞典皇家工学院的 Abbas 等人 [33] 阐述了片上网络性能的解析方法，对比分析排队论、网络演算、数据流分析等方法的特点和适用场合。BENO T 等人基于确定性网络演算技术，为保证 NoC 服务质量，从全局的角度，针对路径划分问题，将网络流量和突发度参数解耦，借助网络演算中的到达曲线和节点路由器的服务曲线，获得更好的网络流量和端到端传递延时边界 [34]。

（3）路由单元性能优化

路由单元由数据通路和控制通路两部分组成，降低数据通路延时，并且减少控制通路给数据流带来的停顿，是提高路由单元性能的关键因素。处理器核数目的增长，推动着处理器核间通信规模的持续增大，使得网络传输时所经历的平均跳步数不断增多，进而不可避免地导致片上通信延时的相应增大。众核微处理器的性能将越来越敏感于路由单元通信的性能，研究表明，当路由单元的流水线级数从 1 级提高到 5 级时，众核处理器的整体性能将下降约 10%[35]。

BOKHARI H 提出双层路由链路 [15]，分别传递读、写数据包，降低数据传递过程中的冲突，提高传递效率。DOPPA J R 在路由单元中增加 SMART 电路 [10]，允许跨路由、单周期多跳传递。LI Y 通过三种策略改进路由单元结构 [36]：①路由单元增加一条新的传递路径，使得处理器核与路由单元间存在两条传递路径；②路由单元中增加智能流量控制模块，降低全局通信竞争概率；③改进路由单元虚拟通道状态表、路由计算模式、虚拟通道分配算法、交叉条结构和分配算法，以降低局部通信竞争概率；改进后的路由单元，片上网络吞吐量会提高 51%，传递延时降低 38%。

2.2.2 片上缓存性能优化

众核处理器由于不均匀的数据访问延时和同一数据在多个处理器核上的不同拷贝，会导致严重的存储一致性问题。侦听与目录是存储一致性协议的 2 种重要的实现方式。由于侦听协议的可扩展性差，具有更高可扩展性的目录协议得到了广泛的使用，并且日趋成熟。

在此情况下，提高片上存储访问性能成为研究热点。HAN X 提出了一种可重构缓存系统 [37]，能够根

据需要将缓存行（Cache Line）配置为消息传递缓存区，提高片上存储的利用率。MASING L 开展基于区域可配置的缓存一致性研究[38]，保证区域内缓存的强一致性，区域间缓存的弱一致性，简化并加速核间一致性消息传递，从而提高整体性能。Burgio P 采用可预测执行模型（PRAM）[39]，基于内存感知方法，将任务分为存储和计算两个阶段，在存储阶段，显式控制从内存中检索数据、并将其复制到内核的本地缓存中，通过这种机制，降低内存争用而导致的延时可变性。CHEN K 为了发挥出 3D 堆叠 DRAM 的带宽，提出了历史辅助自适应粒度缓存机制[40]，能够根据存储访问的历史地址信息，对缓存访问进行前期预测，并对缓存粒度提供弹性映射，提升缓冲的命中率，有效降低存储访问延时。

缓存压缩技术能够在不增加缓存面积的条件下，增加缓存的有效容量，减少缓存失效，从而降低平均访存延时。与其它缓解存储墙问题的技术相比，缓存压缩技术不会增加系统访存带宽需求，这对于目前面向数据吞吐率的处理器系统很有吸引力。NGUYEN T M 提出 MORC 缓存压缩结构[41]，采用行间压缩技术，利用日志的方式将压缩后的数据组织在一起，结构中增加行映射表，实现缓存到日志的灵活映射。

2.2.3 软件编译与任务调度性能优化

固定的缓存策略难以适应程序中数据访存模式的多样性，容易造成缓存抖动，以致影响性能。为此，众核处理器动态地为程序中的数据对象分配 Cache 段，并且动态改变段容量、段内相联度、块大小，从而适应访存模式的多样性。为避免增加程序员的负担，众核缓存的软件管理工作主要由编译器来承担。TANG X L 提出一种编译器缓存优化框架[42]，采用代码重组和计算调度方法，在考虑缓存、内存中 BANK 数量的基础上，最大化缓存并行性（CLP）和内存并行性（MLP），自动优化 MLP 和 CLP 间的平衡，降低末级缓存未命中带来的延时，获得最佳的应用性能。KISLAL O 针对众核处理器中非均匀内存访问架构，在考虑内核、末级缓存、存储控制器的相对位置的基础上，提出了一种编译策略[43]，实现计算任务在处理器内核上的优化映射。众核处理器由于大部分采用分布式片上缓存，重用的数据需要在片上网络中移动，并且要移动的距离——空间重用距离（DIS）在决定应用程序性能方面的影响，与时间上的重用距离（DIT）一样大。KANDEMIR M T 首次尝试定义一个同时寻优 DIT 和 DIS 的编译器框架[44]，将数据重用分为四组：G1（低 DIT，低 DIS）、G2（高 DIT，低 DIS）、G3（低 DIT，高 DIS）、G4（高 DIT，高 DIS），论文提出一种重用传递策略，尽量将重用数据变换到 G1 或者 G4 组中，该策略能将应用程序的执行时间减少 19% 到 33%。

应用程序划分以及任务映射对网络流量起着决定性作用，将多个计算任务映射到处理器核上进行并行计算的过程，是一个 NP 完全问题，分为静态调度和动态调度。静态调度中任务划分主要利用了 EDF 动态优先级或 RM 固定优先级策略[45]，而任务分配问题则采用各种装箱问题的启发式方法，如首次适应（First-Fit）、下次适应（Next-Fit）、最佳适应（Best-Fit）、最坏适应（Worst-Fit）及按利用率递减排序等。

相较于事先确定的负载静态调度，运行时测量负载并划分的动态调度，能够获得 30%-40% 的性能提高[46]，具体技术包括包括遗传算法（GA）[47]，模因算法（MA）[48,49]，以及遗传算法与其它一些技术的融合，比如变邻域搜索[50]，神经网络和列表调度技术[51]。随着处理器集成核数的增加，具有数百或数千个核的众核处理器，经典的动态调度策略已不能充分利用众核多线程高度的细粒度并行性和众核集群的层次结构特征，甚至需要重新设计应用程序、库以及算法。因此，为众核处理器提供运行时支持，实现应用程序的自动优化和软件到硬件的动态映射，达到降低众核系统编程难度、提高性能的目的。MIOMANDRE H G 提出运行时管理器 SPIDER[52]，在计算簇上动态部署可重构数据流图，支持在非统一内存访问体系结构中执行可重构数据流图，运行时系统通过对计算内核和分配任务的调度和映射，达到提高性能的目的。ZHANG X 提出快速检测数据流处理瓶颈的方法[53]，通过细粒度运行时管理，快速响应程序工作负载和资

源（分配的内核数量）的变化，自适应扩展和收缩参与流处理的内核数量，提高流水处理的吞吐量。

2.3 众核处理器可靠性优化技术

众核处理器可靠性的问题非常突出，包括软错误和芯片老化两方面。

2.3.1 软错误容错优化技术

随着半导体工艺特征尺寸的缩小、工作电压的降低，引起电路翻转所需要的临界电荷也在不断降低，BORKAR S和SIVAKUMAR P分别指出每代工艺的进步会使数据位的软错误率（SER）增加8%，导致众核处理器的错误率急剧增大[54,55]。根据软错误发生的位置，可分为处理器核错误和路由单元错误。研究有效的众核处理器容错技术，应对软错误常态化挑战。

在体系结构方面，通过添加超过系统性能需求的资源，提高系统的复杂度来提高系统的容错能力。在这样的冗余设计上，一个部分出现了错误，不会影响整个系统的工作。有的系统具有重构功能，错误的冗余部分可以得到恢复从而恢复系统的容错能力，主要包括运算核重布局技术[56]，自适应路由技术[57]，冗余传输策略技术[58]。BOKHARI H在路由单元数据路径中增加纠检错模块[15]，能够对64位微片纠正8位错、检测16位错；在控制路径中，通过双模锁步模式提高传输的可靠性。DOPPA J R在路由单元中增加自我监测和配置电路[14]，防止死锁现象的发生。WANG K在路由单元中增加自适应错误纠正／检测模块和重新传输控制机制[59]，以提高可靠性。

在软件方面，BALBONI M针对故障的检测和规避，对受故障影响的区域进行界定，基于OSRLite运行时系统，动态重构路由功能[60]：①利用令牌快速建立、断开链路功能；②建立不同网络间隧道功能，将发生故障的路由所承担的功能，快速、且可控地切换到新的路由上，依靠令牌和隧道传输，实现新旧路由间的信息交换。

2.3.2 老化效应防护技术

器件老化导致的可靠性问题，主要包括：经时击穿（TDDB）、热载流子注入（HCI）、负栅压温度不稳定性（NBTI）、电迁移（EM）。同时，功耗密度的增加也使温度成为加剧器件故障的重要因素。HASELMAN M指出，在纳米工艺下，芯片中晶体管和连线发生故障的概率将超过15%[61]。工艺进步加剧了芯片老化，为了维持芯片的正常使用寿命，需要对电路进行抗老化设计，常用的方法有基于电路拓扑结构重构的老化防护技术[62,63]，基于输入向量控制（IVC）和内部节点控制（INC）的集成电路老化防护技术[64–66]和基于动态调整技术的集成电路老化防护[67–70]。

为降低电迁移对可靠性的影响，KIM T建立等效电流模型，提出一种系统级动态可靠性管理技术[71]，通过寻找最佳处理器核的电压和开／关状态，提高在接近阈值运行的低功耗众核微处理器的可靠性。TAEYOUNG K将自适应强化学习方法和混合整数线性规划方法相结合[4]，对处理器核上的计算任务进行实时调整或者迁移，结合DVFS技术，达到延长众核处理器寿命的目的。

为降低NBTI对可靠性的影响，RATHORE V提出一种面向性能的约束感知任务映射技术[72]，此技术由两部分构成：芯片设计阶段，采用阈值接受模拟退火算法达到帕累托最优，以最大化延长平均故障时间；在芯片运行阶段，通过工作负载的分配，对最脆弱的处理器核，降低工作负载。与性能贪婪和温度感知任务映射技术对比，此方法能够提高芯片寿命54%。

为降低功率密度增加导致的热力问题对众核可靠性的影响，RATHORE V为众核处理器设计了资源管理策略[73]：采用分层划分方法进行热缓解；使用强化学习技术降低芯片的老化效应，此资源管理策略，能够根据应用程序的负载数量和处理器核的数量进行扩展和动态改变，从仿真结果上看，MTTF提高0.33年，寿命提高69%。

3 众核处理器未来研究重点分析

当前，微电子发展处于技术变革的重要时期，延续摩尔定律和超越摩尔定律是其发展的两个重要方

面，到 2025 年，传统 CMOS 微缩面临终结，新原理、新结构将登上舞台。众核处理器未来关注的重点是体系结构的自适应性和三维集成的高效性。

3.1 众核处理器自适应性技术

众核处理器的设计理念与传统的单核和多核处理器不同，不再过度追求单个处理器核的计算能力，而是追求并行计算能力和处理器整体的计算能力。为适应变化的任务需求，众核处理器应具备软、硬件结构和计算模式的可变性和可编程性，研究众核处理器的自适应性。

现在的众核处理器很难高效适应不同的计算模式，为提高众核处理器适应性，DOPPA J R 提出了一种众核自适应体系结构[14]，自适应性体现在应用（任务分配和调度）、处理器核（DVFS 和电源门控）和互连网络（路由）多个方面。特别是在路由单元设计中，提出可配置中继器的概念，能够根据需要将多个短距离链路通过中继器连接起来，创建单周期长距离链路，允许单周期跨多个跃点传递数据。SCIONTI A 开展可扩展的软件定义片上网络（SDNoC）体系结构研究[16]，在保持二维网络拓扑固定的基础上，融合环形拓扑和动态配置能力，允许映射不同类型的拓扑；路由单元通过为每条链路配置计数器，用于跟踪数据传递的统计值，以动态调整链路资源分配，允许软件层控制并监控片上网络。ZHENG H 提出一种自适应 NoC 架构[74]，能够为并发运行的程序动态分配多个不相交的子网，每个子网区域能够配置成最适合映射程序特征的拓扑结构，例如网格、环形或树形，以提高性能。NAZARIAN S 提出了一个自优化、自适应的众核系统框架[75]，由代表计算模型、连接模型、内存模型和存储模型的四层图模式组成，该框架通过实时调整这些模型，将硬件木马和侧信道等攻击的影响降到最低。

除了众核片上网络的自适应能力，自适应性还体现在处理器核上。在任务执行过程中可以实时选取相应数量的物理处理器核动态组成与任务需求匹配的虚拟处理器核，实现针对不同任务、不同需求的资源最优匹配。因此，众核处理器的处理过程是一种不断“动态异构”的过程，具有面对不同应用的适应性、组织性和自主性。CAO Y J 通过对任务的拆分与合并，动态构建虚拟处理器核[76]，实现核资源的优化、隔离和访问。

3.2 众核处理器三维集成技术

受到电路性能改善的驱动，众核处理器向三维集成方向发展。三维结构有效减小全局和半全局互连线长度，在提高芯片集成度的同时，使互连延时和互连功耗明显降低。尽管三维集成具有上述优点，但也存在一些问题，热管理就是最严重的问题之一。这是因为垂直堆叠的方式增加了芯片的功率密度，而且层间材料的热导率相对于硅片和金属来说很小，导致处理器内部产生的热量很难散失。为了使三维处理器不会因为温度过高而失效，必须控制其温度在允许的范围之内。LEE D J 和 MUSAVVIR S 等人对单片三维众核处理器的性能和热平衡问题进行了研究[2,77]，建立热力学模型，提出了功耗 – 性能 – 热平衡的众核处理器架构设计原则，探索达到性能和热力平衡的单片三维众核处理器架构。CHATTERJEE A 基于系统计算和通信特征，考虑三维工艺的影响，提出了一种基于模拟学习算法的电源管理策略[3]，为各处理器核提供适当的电压／频率等级，以降低处理器核功耗，达到约束温度的目的。

在三维集成生产过程中，微观组织的特征尺寸相比于微焊点尺寸急剧增加，焊点中微空洞的形成、迁移、合并粗化，使得三维互连的可靠性面临更为严峻的挑战。SOURAV D 对硅通孔（TSV）的可靠性展开分析[1]，结合电源管理和自适应路由技术，达到提高三维众核处理器可靠性的目的。ARKA A I 讨论硅通孔（TSV）和近场感应耦合（NFIC）两种三维互连技术的可靠性以及各自适用的信号带宽[78]。

3.3 异构融合并行计算技术

众核处理器与 CPU、FPGA、ASIC 一起构成的异构计算系统，相比传统的对称处理器系统更有性能优势[79]，被视为继单核、多核之后的第三个时代，有

效解决能耗等问题，成为高性能计算领域的一种重要模式。

美国苹果公司的 M1 Ultra 芯片集成 20 个 CPU 核、64 个 GPU 核和 32 个神经网络引擎，性能达到 22TOPS[80]。美国 DARPA 资助的 Celerity 项目，集成 511 个不同的处理器核 [81]。赛灵思公司推出的 ACAP 异构计算架构，包括了 AI 引擎阵列、DSP 引擎阵列、CPU 核和可编程逻辑单元阵列。

由于异构众核内部的多种处理器核各自拥有不同的系统架构、指令集合以及编程模型，尤其是加速引擎具备自身特有的结构特征，异构系统往往有着不同于 CPU 的编程模型，缺乏一个标准化编程环境来统管异构系统内呈多样化发展态势的各种资源。虽然 OpenCL 提供了一个统一的编程模型，但是其编程抽象层次低，编程接口靠近底层，无法为用户屏蔽底层硬件和运行时的细节，导致编程逻辑复杂，编程困难易错。因此，屏蔽异构处理器并行编程模式的不同，提供一个统一的编程模型，并且能够保持异构处理器本身的并行效率，将是异构计算的重要研究方向。

Lee S 等 [82] 提出一种将通用并行编程语言 OpenMP 编写的标准应用程序转换成 GPGPU 代码的自动编译和优化框架。该编程框架的主要思路是让用户利用已有的通用并行编程模型来编写基于异构系统的并行程序，而由源到源编译器来完成具体的代码转换、性能优化及代码到具体硬件的映射工作，从而提高异构系统的可编程性。

吴树森 [83] 提出了并行编程架构 UPPA。架构中首先提出了数据关联计算编程模型，实现了不同层级不同模式并行性的统一描述；设计了数据关联计算描述语言，通过高层语义结构保留了应用的并行特征，指导编译和运行时系统实现向不同硬件架构的自动映射。

李雁冰 [84] 以 Sunway OpenACC 并行程序作为输出，提出了一种面向异构众核处理器的并行编译框架，将程序自动转换为异构并行程序。该框架主要包括 4 个模块：任务划分模块识别适合进行加速计算的程序段，实现了嵌套循环的多维并行识别方法；数据布局模块完成数据在主存和 SPM 之间的布局，实现了数组边界分析和指针范围分析；传输优化模块实现了数据传输合并、传输外提、打包传输、数组转置等多种数据传输优化方法；收益评估模块在构建代价模型的基础上实现了一种动静结合的收益评估。

以上通过设计编程模型或编程框架来解决异构系统可编程性问题的主要特点是：侧重于如何降低异构系统的编程复杂性，用户在编程时主要从算法层面考虑，不用考虑具体的底层硬件架构特点；具体性能优化的工作由编译器完成，对程序员透明。

4 结束语

综合来看，众核处理器的未来发展主流是融合创新。首先是多种拓扑网络的融合，不同拓扑结构决定了网络延时、带宽、吞吐率和系统功耗、芯片面积和容错能力上的不同，通过融合拓扑网络，优化出最适合应用的低直径拓扑、低功耗路由单元、容错路由算法及网络管理方式；其次是软件技术和硬件技术的创新融合，软件定义的范围和影响力将继续拓展，不仅可实现面向应用的处理器上资源的调度和管理，还针对系统架构、存储架构的个性化需求，实现软硬解耦和资源灵活配置，达到与算法或框架深度融合的硬件动态专用定制；再次，是经典设计与新器件、新工艺的融合，为了解决存储墙、I/O 带宽墙和功耗墙的问题，同时由于非易失存储、3D 堆叠等新技术的不断发展，众核处理器体系结构正发生着存储向计算靠近的融合结构变革。

参考文献（References）

[1] SOURAV D, JANARDHAN R D, PARTHA P P. Energy-Efficient and Reliable 3D Network-on-Chip (NoC): Architectures and Optimization Algorithms[C]//2016 IEEE/ACM international conference on Computer-Aided Design (ICCAD), 2016:1-6.

[2] LEE D J, DAS S, Doppa J R. Performance and Thermal Tradeoffs for Energy-Efficient Monolithic 3D Network-on-Chip[J], ACM Transactions on Design Automation of Electronic Systems, 2018,23(5):1-25.

[3] CHATTERJEE A, KIM R G, DOPPA J R. Power

Management of Monolithic 3D Manycore Chips with Inter-tier Process Variations[J]. ACM Journal on Emerging Technologies in Computing Systems, 2021, 17(2):1-19.

[4] TAEYOUNG K, SUN Z Y, CHEN H B, Energy and Lifetime Optimizations for Dark Silicon Manycore Microprocessor Considering Both Hard and Soft Errors[J]. IEEE transactions on very large scale integration systems, 2017,25(9):2561-2574.

[5] KIM J S, TAYLOR M B, MILLER J, et al. Energy characterization of a tiled architecture processor with on-chip networks[C]//Proceedings of the International Symposium on Low Power Electronics and Design,2003: 424-427.

[6] Intel Corp. From a Few Cores to Many: A Terascale Computing Research Overview[R]. 2006.

[7] SAMIH A, REN W, KRISHNA A, et al. Energy-efficient interconnect via router parking[C]//Proc of the 19th Int Symp on High Performance Computer Architecture(HPCA). Los Alamitos, CA: IEEE Computer Society, 2013:8-19.

[8] GÓMEZ C, GÓMEZ M E, LÓPEZ P, et al. Reducing packet dropping in a bufferless NoC[C]//Proc of the 14th Int Euro-Par Conf on Parallel Processing. Berlin: Springer, 2008:899-909.

[9] MOSCIBRODA T, MUTLU O. A case for bufferless routing in on-chip networks[C]//Proc of the 36th Annual Int Symp on Computer Architecture. New York: ACM, 2003:424-427.

[10] JIANG N, BECKER D U, Michelogiannakis G, et al. A detailed and flexible cycle-accurate network-on-chip simulator[C]//Proc of Performance Analysis of Systems and Software. New York: ACM, 2013:86-96.

[11] WANG H, PEH L S, MALIK S, A Technology-Aware and Energy-Oriented Topology Exploration for On-Chip Networks[C]//Proceedings of Design, Automation and Test in Europe, 2005:1238-1243.

[12] PINTO A, CARLONI L P, SANGIOVANNI VINCENTELLI A L, Efficient Synthesis of Networks on Chip[C]// Proceedings of the 21st International Conference on Computer Design,2003:146-150.

[13] SOUZA M A, FREITAS H C, MEHAUT J F, Design Space Exploration of Energy Efficient NoC-and Cache-Based Many-Core Architecture Using Distributed L2 and Adaptive L3 Caches[C]//2018 30th International Symposium on Computer Architecture and High Performance Computing, 2018:402-409.

[14] DOPPA J R, KIM R G. Adaptive Manycore Architectures for Big Data Computing[C]//2017 Eleventh IEEE/ ACM International Symposium on Networks0on-Chip, 2017:1-8.

[15] BOKHARI H, JAVAID H, SHAFIQUE M. SuperNet: Multimode Interconnect Architecture for Manycore Chips[C]//Proceeding of the 52nd Annual Design Automation Conference, 2015:1-6.

[16] SCIONTI A, MAZUMDAR S, PORTERO A. Software Defined Network-on-Chip for Scalable CMPs[C]//2016 international conference on High Performance Computing & Simulation (HPCS), 2016:112-115.

[17] GONZALEZ R, HOROW I . Energy dissipation in general purpose microprocessors[J]. IEEE Journal of Solid State Circuits, 1996,31(9):1277-1284.

[18] ISHMANN V, IRWINMJ K, et al. Energy-driven integrated hardware-software optimizations using simple-power[C]//Proceedings of the 27th Annual International Symposium on Computer Architecture, 2000:95-106.

[19] SUBRAMANIAN R, SMARAGDAKIS Y, LOH G H. Adaptive caches: effective shaping of cache behavior to workloads [C]//Proceedings of the 39th Annual IEEE/ ACM Int Symp on Microarchitecture, 2006:385-396.

[20] TITOSGIL R, FLORES A, FERNANDEZ-PASCUAL R. Way-Combining Directory: An Adaptive and Scalable Low-Cost Coherence Directory[C]//Proceedings of the International Conference on Supercomputing, 2017:1-10.

[21] BANAKAR R, STEINKE S, LEE B, et al. Scratchpad Memory: A Designed Alternative for Cache On-chip memory in Embedded System[C]//Proceedings of the tenth international symposium on Hardware/software

codesign(CODES 2002), 2002:73-78.

[22] BANAKAR R, STEINKE S, LEE B,et al. Comparison of cache and scratchpad-based memory systems with respect to performance, area and energy consumption[R]. University Dortmund.

[23] ALVAREZ L, VILANOVA L, MORETO M. Coherence Protocol for Transparent Management of Scratchpad Memories in Shared Memory Manycore Architectures[C]//Proceedings of the 42nd Annual International Symposium on Computing Architecture, 2015:720-732.

[24] KIM N, AHN J, CHOI K. Benzene: An Energy-Efficient Distributed Hybrid Cache Architecture for Manycore Systems[J]. ACM Transactions on Architecture and Code Optimization, 2018,15(1):1-23.

[25] WANG X Q, XI J J, WANG Y H, An Efficient Task Mapping for Manycore Systems[C]//2020 IEEE International Symposium on Circuits and Systems, 2020:1-4.

[26] CAPOTONDI A, HAUGOU G, MARONGIU A, BENINI L, Runtime Support for Multiple Offload-Based Programming Models on Embedded Manycore Accelerators[C]//Proceedings of the 2015 International Workshop on Code Optimisation for Multi and Manycores,2015:1-10.

[27] LE T T, ZHAO D, et al. Optimizing the Heterogeneous Network On-Chip Design in Manycore Architectures[C]//2017 30th IEEE International System-on-Chip Conference, 2017:184-189.

[28] HUANG W H, CHEN J J, REINEKE J, MIRROR: Symmetric timing analysis for real-time tasks on multicore platforms with shared resources[C]//2016 53rd ACM/EDAC/IEEE Design Automation Conference (DAC). IEEE, 2016:1-6.

[29] DAVIDSON S, XIE S, TORNG C, ALHAWAJ K. The Celerity Open-Source 511-Core RISC-V Tiered Accelerator Fabric: Fast Architectures and Design Methodologies for Fast Chips[J], IEEE Micro, 2018(3/4):30-41.

[30] DESHWAL A, JAYAKODI N K, et al. MOOS: A Multi-Objective Design Space Exploration and Optimization Framework for NoC Enabled Manycore Systems[J]. ACM Transaction on Embedded Computing Systems, 2019,18(5):1-23.

[31] ABBAS E K, AXEL J, et al. Mathematical formalisms for performance evaluation network on chip [J]. ACM computing Surveys, 2013,45(3):1-41.

[32] OGRAS U Y, MARCULESCU R. Application-specific network-on-chip architecture customization via long-range link insertion[C]//Proceedings of the 2005International Conference on Computer aided Design(ICCAD 05).Washington, DC:IEEE Computer Society, 2005:246-253.

[33] 王炜，乔林，杨广文，等．片上二维网络互连性能分析[J].计算机研究与发展，2009,46(10):1601-1611.

[34] BENO T DUPONT DE DINECHIN, GRAILLAT A. Network-on-Chip Service Guarantees on the Kalray MPPA-256 Bostan Processor[C]//Proceedings of the 2nd International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems, 2017:35-40.

[35] JERGER N E, PEH L S, LIPASTI M. Circuit-Switched Coherence[C]//Proceedings of the the 2nd International Symposium on Networks-on-Chip. 2008:193-202.

[36] LI Y, LOURI A. ALPHA: A Learning-Enabled High-Performance Network-on-Chip Router Design for Heterogeneous Manycore Architectures[J]. IEEE Transactions on Sustainable Computing, 2021,6(2):274-288.

[37] HAN X, FU Y, JIANG J. Reconfigurable MPB Combined with Cache Coherence Protocol in Many-core[C]//2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2016:385-388.

[38] MASING L, KRE F, SRIVATSA A, HERKERSDORF A, In-NoC circuits for low-latency cache coherence in distributed shared-memory architectures[C]//2018 IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip, 2018:138-145.

[39] BURGIO P, MARONGIU A, VALENTE P, et al. A memory-centric approach to enable timing-predictability

within embedded many-core accelerators[C]//2015 CSI Symposium on Real-Time and Embedded Systems and Technologies, 2015:1-8.

[40] CHEN K, LI S, AHN J H, MURALIMANOHAR N. History-Assisted Adaptive-Granularity Caches (HAAG$) for High Performance 3D DRAM Architectures[C]//Proceedings of the 29th ACM on International Conference on Supercomputing, 2015:251-261.

[41] NGUYEN T M, WENTZLAFF D. MORC: A Manycore-Oriented Compressed Cache[C]//Proceedings of the 48th International Symposium on Microarchitecture, 2015:76-88.

[42] TANG X L, KANDEMIR M T, KARAKOY M, et al. Co-optimizing Memory-Level Parallelism and Cache-Level Parallelism[C]//Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, 2019:935-945.

[43] KISLAL O, KOTRA J, TANG X L, et al. POSTER: Location-Aware Computation Mapping for Manycore Processors[C]//2017 26th International Conference on Parallel Architectures and Compilation Techniques, 2017:138-139.

[44] KANDEMIR M T, TANG X L, ZHAO H, RYOO J, Distance-in-Time versus Distance-in-Space[C]//Proceedings of the 42nd ACM SIGPLAN Conference on Programming Language Design and Implementation, 2021:665-680.

[45] DAVIS R I , BUMS A, A survey of hard real-time scheduling algorithms and schedule ability analysis techniques for multiprocessor systems[R]. University of York, Department of Computer Science Technical Report, YCS-2009-443, November, 2009.

[46] KAMEDA H, FA E S, RYU I, et al. A performance comparison of dynamic vs. static load balancing policies in a mainframe-personal computer network model [C]//Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia. 2000:1415-1420.

[47] WEN Y, XU H, YANG J D. A heuristic-based hybrid genetic-variable neighborhood search algorithm for task scheduling in heterogeneous multiprocessor system. Information Sciences, 2011,181(3):567-581.

[48] SANCHO S S, XU Y, YAO X. Hybrid meta-heuristics algorithms for task assignment in heterogeneous computing systems. Computers and Operations Research, 2006,33(3):820-835.

[49] CHITRA P, RAJARAM R, VENKATESH R. Application and comparison of hybrid evolutionary multi-objective optimization algorithms for solving task scheduling problem on heterogeneous systems. Applied Soft Computing, 2011,11(2):2725-2734.

[50] WU A S, YU H, LIN K C, et al. An incremental genetic algorithm approach to multiprocessor scheduling. IEEE Transactions on Parallel and Distributed Systems,2004,15(9):824-834.

[51] BRAUN T D, SIEGEL H J, MACIEJEWSKI A A, et al. Static resource allocation for heterogeneous computing environments with tasks having dependencies, priorities, deadlines, and multiple versions. Journal of Parallel and Distributed Computing,2008,68(11):1504-1516.

[52] MIOMANDRE H G, HASCOET J L, DESNOS K, et al. Embedded Runtime for Reconfigurable Dataflow Graphs on Manycore Architectures[C]//Proceedings of the 9th Workshop and 7th Workshop on Parallel Programming and RunTime Managemant Techniques for Manycore Architectures and Design Tools and Architectures for Multicore Embedded Computing Platforms, 2018:51-56.

[53] ZHANG X, JAVAID H, SHAFIQUE M, et al. ADAPT: An ADAptive Manycore Methodology for Software Pipelined Applications[C]//The 20th Asic and South Pacific Design Automation Conference, 2015:701-706.

[54] BORKAR S. Designing reliable systems from unreliable components: the challenges of transistor variability and degradation [J]. Micro IEEE. 2005,25(6):10-16.

[55] SIVAKUMAR P,et al. Modeling the effect of technology trends on soft error rate of combinatorial logic [C]//Proceedings International Conference on Dependable Systems and Networks. 2002:389-398.

[56] WONG R, LI J, FU A, et al. (α,k)-Anonymous data publishing[J]. Journal of Intelligent Information Systems, 2009,33(2):209-234.

[57] 杨高明，杨静，张健沛．半监督聚类的匿名数据发布[J]．哈尔滨工程大学学报，2011,32(11):1489-1495.

[58] 滕金芳，钟诚．基于聚类的敏感属性－多样性匿名化算法[J]．计算机工程与设计，2010,31(20):4378-4381.

[59] WANG K, LOURI A, KARANTH A, BUNESCU R. IntelliNoC: A Holistic Design Framework for Energy-Efficient and Reliable On-Chip Communication for Manycores[C]//Proceedings of the 46th International Symposium on Computing Architecture, 2019:589-600.

[60] BALBONI M, BERTOZZI D, et al. Synergistic Use of Multiple On-Chip Networks for Ultra-Low Latency and Scalable Distributed Routing Reconfiguration[C]//2015 Design, Automation & Test in Europe Conference & Exhibition,2015:806-811.

[61] HASELMAN M, HAUCK S. The Future of Integrated Circuits: A Survey of Nanoelectronics[C]//Proceeding of the IEEE, 2010,98(1):11-38.

[62] KAI-CHIANG W, MARCULESCU D, Joint logic restructuring and pin reordering against NBTI-induced performance degradation[C]. Proc. Design, Automation & Test in Europe Conference & Exhibition, 2009:75-80.

[63] BUTZEN P F, BEM V D, et al., Transistor network restructuring against NBTI degradation[J]. Microelectronics Reliability, 2010,50(9-11):1298-1303.

[64] ABELLA J, VERA X, GONZALEZ A, Penelope: The NBTI-Aware Processor[C]. Proc. Micro-architecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on, 2007:85-96.

[65] SONG J, et al., M-IVC: Using Multiple Input Vectors to Minimize Aging-Induced Delay[C]. Proc. Asian Test Symposium, 2009:437-442.

[66] WANG Y, LUO H, et al., Temperature-Aware NBTI Modeling and the Impact of Standby Leakage Reduction Techniques on Circuit Performance Degradation[J]. Dependable and Secure Computing, IEEE Transactions on, 2010:1-1.

[67] MINTARNO E, SKAF J, et al.. Self-Tuning for Maximized Lifetime Energy-Efficiency in the Presence of Circuit Aging[J].Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, 2011,30(5):760-773.

[68] LIDE Z, DICK R P, Scheduled voltage scaling for increasing lifetime in the presence of NBTI[C]. Proc. Design Automation Conference, 2009. ASP-DAC 2009. Asia and South Pacific, 2009:492-497.

[69] BASOGLU M, ORSHANSKY M, EREZ M, NBTI-aware DVFS: A new approach to saving energy and increasing processor lifetime[C]. Proc. Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on, 2010:253-258.

[70] KUMAR S V, et al., Adaptive Techniques for Overcoming Performance Degradation Due to Aging in CMOS Circuits[J]. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 2011,19(4):603-614.

[71] KIM T, SUN Z, COOK C, GADDIPATI J. Dynamic reliability management for near-threshold dark silicon processors[C]//Proceedings of the 35th International Conference on Computer-Aided Design, 2016:1-7.

[72] RATHORE V, CHATURVEDI V, SRIKANTHAN T, Performance Constraint-Aware Task Mapping to Optimize Lifetime Reliability of Manycore Systems[C]//Proceedings of the 26th edition on Great Lakes Symposium on VLSI, 2016:377-380.

[73] RATHORE V, CHATURVEDI V, SINGH A K, et al. Towards Scalable Lifetime Reliability Management for Dark Silicon Manycore Systems[C]//2019 IEEE 25th International Symposium on On-Line Testing and Roburst System Design, 2019:204-207.

[74] ZHENG H, WANG K, LOURI A. Adapt-NoC: A Flexible Network-on-Chip Design for Heterogeneous Manycore Architectures[C], 2021 IEEE International Symposium on High-Performance Computer Architecture, 2021:723-735.

[75] NAZARIAN S, BOGDAN P. S4oC: A Self-Optimizing, Self-Adapting Secure System-on-Chip Design Framework to Tackle Unknown Threats — A Network Theoretic, Learning Approach[C]//2020 IEEE International Symposium on Circuits and Systems, 2020:1-8.

[76] 曹仰杰，等．众核处理器系统核资源动态分组的自适应调度算法[J]．软件学报，2012,23(2):240−252.

[77] MUSAVVIR S, CHATTERJEE A, KIM R G, et al. Power, Performance, and Thermal Trade−offs in M3D−enabled Manycore Chips[C]//2020 Design Automation & Test in Europe Conference & Exhibition, 2020:1752−1757.

[78] ARKA A I, GOPAL S, DOPPA J R, HEO D. Making a Case for Partially Connected 3D NoC: NFIC versus TSV, ACM Journal on Emerging Technologies in Computing Systems[J]. 2020,16(4):1−17.

[79] 阳王东，王昊天，张宇峰，林圣乐，蔡沁耘．异构混合并行计算综述[J]．计算机科学，2020,47(8):5−16.

[80] 大势所趋的芯片异构，https://new.qq.com/omn/ 20220406/20220406A03IMQ00.html.

[81] DAVIDSON S. et al., the Celerity Open−Source 511−core RISC−V Tiered Accelerator Fabric: Fast Architectures and Design Methodologies for Fast Chips, in IEEE Micro, 2018,38(2):30−41.

[82] LEE S, EIGENMANN R. Open MPC:extended openMP programming and tuning for GPUs[A]. Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis[C]. IEEE, 2010:1−11.

[83] 吴树森，董小社，王宇菲，王龙翔，朱正东．UPPA：面向异构众核系统的统一并行编程架构[J]．计算机学报，2020,43(06):990−1009.

[84] 李雁冰，赵荣彩，韩林，赵捷，徐金龙，李颖颖．一种面向异构众核处理器的并行编译框架[J]．软件学报，2019,30(04):981−1001.

作者简介：

宋立国（1973—），男，河北省隆化县人，博士，研究员，研究方向为多核、众核处理器和系统集成芯片设计。

热载流子应力下脱氢和陷阱效应对 SiN/AlGaN/GaN MIS-HEMT 的电学退化影响研究

牛雪锐，马晓华，侯 斌，杨 凌，朱 青

（西安电子科技大学，陕西省 西安市 710071）

摘 要：这项工作研究了氮化镓基的金属－绝缘体－半导体高电子迁移率晶体管在不同栅极和漏极电压偏置下热电子效应诱导的退化机制。器件在热载流子应力过程中，跨导峰值异常增加，并且在去除电应力后跨导出现快速的部分恢复。提出了一个物理模型来解释由热载流子应力引起的器件异常的电学特性变化。通过使用密度泛函理论，计算了电子在热载流子应力过程中使氮化镓材料内存在的本征缺陷 $[N_{Ga}H_3]^{-1}$ 复合物脱氢的能量。缺陷的脱氢效应影响了器件的跨导峰值。同时，铝镓氮势垒层中施主陷阱的中性化也对跨导峰值的增加起到了重要影响。这部分缺陷的影响导致了跨导峰值的快速部分恢复。

关键词：氮化镓；跨导峰值；热载流子应力；脱氢效应

中图分类号：TN4　　**文献标识码**：A

Electrical Degradation of In Situ SiN/AlGaN/GaN MIS-HEMTs Caused by Dehydrogenation and Trap Effect under Hot Carrier Stress

Niu Xuerui, Ma Xiaohua, Hou Bin, Yang Ling, Zhu Qing

(Xidian University, Xi'an, 710071, China)

Abstract: The gate and drain bias dependence of hot electron-induced degradation in GaN-based metal-insulator-semiconductor high electron mobility transistors was investigated in this work. Devices exhibit an abnormal increase in peak transconductance during hot carrier stress and a partially quick recovery of that after removing the electrical stress. A physical model is proposed to explain the abnormal electrical characteristics caused by hot carrier stress. By using density functional theory, we calculated the energy for electrons to dehydrogenate pre-existing $[N_{Ga}H_3]^{-1}$complexes in GaN layer during stress. The dehydrogenation of defects affects the peak transconductance of devices. Meanwhile, the neutralization of donor traps in AlGaN barrier layer also plays a significant role in the increase of peak transconductance and the de-trapping effect of electrons from these traps after removing the electrical stress accounts for the partially quick recovery of peak transconductance.

Key words: GaN; peak transconductance; hot carrier stress; dehydrogenation effect

0 引言

GaN 基金属－绝缘体－半导体高电子迁移率晶体管 (MIS-HEMT) 由于高击穿电场 (3.1MV/cm)、低导通电阻和低栅极泄漏的特性在功率开关领域受到广泛关注 [1,2]。在工作期间，器件必须从导通状态切换到关断状态，在这过程中器件会进入一个半导通状态，此时漏极电压和电流同时处于高水平。因此，在半导通状态下，会发生热电子效应，并可能导致电荷俘获效应 [3,4]、新缺陷的产生 [5,6] 和原有缺陷的去俘获效应等 [7,8]。这些效应会限制器件工作的性能。因此，应该详细研究器件在半导通状态下由高漏极电压和电流引起的深层失效机制。

之前的一些研究报告称，最严重的退化可能发生在半导通偏置条件下，包括阈值电压的偏移，以及漏极电流和跨导的降低 [3-9]。然而，也有几项研究中跨导表现出相反的变化趋势。据文献 [6] 报道，在关

基金项目：国家重点基础研究发展计划（批准号：2018YFB1802100）、国家自然科学基金（批准号：62090014）资助的课题
作者或通信作者：马晓华；xhama@xidian.edu.cn

态应力下的器件的跨导峰值（$G_{m,max}$）会增加，这是由于 O_N-DX 中心的中和以及从 SiN/AlGaN 界面隧穿进入 GaN 沟道的额外电子的影响。O_N-DX 中心是氧原子替换氮原子的负电荷中心[10,11]，它可以将其电子发射到 AlGaN 的导带并变为中性 O_N，此过程需要的激活能为 0.25eV[12]。因此，可以改善载流子被散射的几率。此外，文献[7]报道了当器件在撤掉半开态应力后跨导的“超级恢复”现象，这归因于材料中存在的本征 O_N-H 缺陷的脱氢效应。O_N-H 向 O_N 的转换的过程可以提高电子迁移率并导致跨导的增加。虽然上述研究已经提出了一些与 AlGaN 层中的本征缺陷相关的模型来解释器件跨导的变化，但尚未考虑 GaN 沟道层附近的其他缺陷。此外，这些论文缺乏对不同栅极和漏极偏置应力条件下以及器件恢复过程中跨导特性的研究。因此，热载流子应力过程中器件的退化机制需要进一步研究。

本文研究了不同栅极和漏极偏置条件下热载流子应力对 GaN 基 MIS-HEMT 的影响。观察到器件在热载流子应力过程中跨导峰值的异常增加和撤销电应力后的跨导的快速部分恢复的现象。通过使用密度泛函理论（DFT），计算了 $[N_{Ga}H_3]^{-1}$ 脱氢成为 $[N_{Ga}H_2]^0$ 所需的能量。GaN 材料中预先存在的 $[N_{Ga}H_3]^{-1}$ 复合物被 GaN 沟道中高能电子脱氢是热载流子应力过程中 $G_{m,max}$ 异常增加的部分原因。此外，通过讨论 $G_{m,max}$ 的部分恢复特性，我们认为 AlGaN 势垒层中的施主陷阱也影响了器件跨导的变化。通过在不同温度下进行应力实验，我们验证了上述模型。

1 实验方法

本文使用的 SiN/AlGaN/GaN 异质结构是使用金属有机物化学气相沉积技术（MOCVD）在 6 英寸 Si 衬底上生长的，设计的外延层从下到上是 4-μm GaN 缓冲层、200-nm 非故意掺杂的 GaN(UID-GaN) 沟道层、25-nm $Al_{0.22}Ga_{0.78}N$ 势垒层和 30-nm 原位 SiN 层。器件制备从使用电感耦合等离子体刻蚀系统和基于 CF_4 的等离子体去除源极和漏极区域中的原位 SiN 开始。在欧姆接触形成和台面隔离之后，使用电子束蒸发沉积栅极金属。随后，沉积 90nm SiO_2 层作为第二保护层，随后去除电极区域中的 SiO_2 层。图 1(a) 是器件结构图，其中栅极长度、栅极到源极和栅极到漏极的距离分别为 5、3 和 30μm。

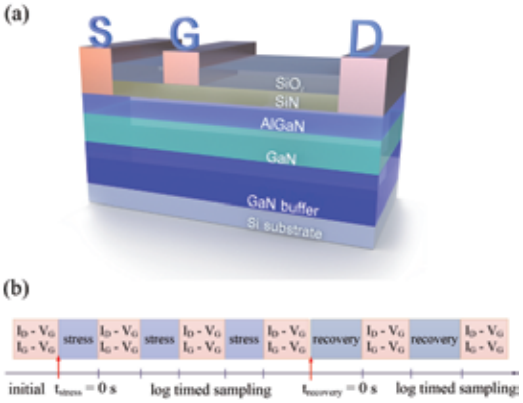


图 1 (a) SiN/AlGaN/GaN MIS-HEMT 结构图 (b) 应力实验过程图

Fig.1 (a) Schematic of SiN/AlGaN/GaN MIS-HEMT structure (b) Schematic of experimental process used in the analysis

应力测试是在不同栅极和漏极电压条件下进行的。此外，对器件进行了 300K 到 360K 不同温度下的应力实验。采用的应力实验过程如图 1(b) 所示，器件在恒定栅极偏置（$V_{Gstress}$）和恒定漏极偏置（$V_{Dstress}$）下进行长时间应力实验。而应力实验中引起的器件 V_{TH} 偏移和 G_m 的变化是通过反复中断应力实验得到的。器件转移特性在漏极电压 V_D=0.1V 下测试，其中栅极电压 V_G 从 -11V 到 -5V 扫描。I_G-V_G 曲线在源极电压 V_S=0V 和漏极浮空时测量。在应力测试结束时，撤掉应力电压并通过多次重复测量转移特性分析器件的恢复情况。这里，V_{TH} 定义为漏极电流达到 1μA/mm 时的 V_G。

2 实验结果和讨论分析

在 T=300K、$V_{Dstress}$=60V 或 100V、$V_{Gstress}$=$V_{TH}(0)$ + 4V 的偏置应力条件下进行了 1000s 的应力测试。$V_{TH}(0)$ 表示器件未施加应力前的阈值电压。图 2(a)-(c) 是器件的 I_D-V_G、跨导和 I_G-V_G 特性。在应力测试之后，器件出现三种退化现象，即 1)V_{TH} 的正向漂移，2) 栅极电流的降低和 3)$G_{m,max}$ 的增加。器件在 $V_{Dstress}$ = 100V 下应力实验后的变化比在 $V_{Dstress}$ =

60V 下应力实验的变化更严重，这表明更高的栅漏电场对器件的影响更大。$V_{Gstress}$ = $V_{TH}(0)$ + 4V 和 $V_{Dstress}$ = 100V 时的栅极应力电流如图 2(d) 所示。栅极应力电流随应力时间增加而降低是由于在 1000 秒的应力期间，电子在高电场下从栅极注入到 SiN 层。随着应力时间的增加，栅极下方被俘获的电子通过排斥力使后续电子更难注入到 SiN 层中，从而减缓了栅极应力电流的下降趋势。相应地，栅极下方的 SiN 层中捕获的电子也可以耗尽 GaN 沟道中的 2DEG，从而在 300K 应力期间使 V_{TH} 向正方向漂移，如图 2(a) 所示[2–13]。由于在 $V_{Dstress}$ =100V 时 SiN 层的电场较大，因此阈值电压正向偏移量更大。除了对 V_{TH} 和栅极电流的影响之外，SiN 绝缘层中的电子以及栅极下方的 SiN/AlGaN 界面处的电子也会因库仑散射而影响电子迁移率，从而导致 $G_{m,max}$ 降低[14]。然而，一些实验表明远程散射率随着电荷与沟道之间的距离变大而变小[15,16]。这些结论表明，本文中由于电荷和沟道之间的距离很大，至少为 AlGaN 势垒层的厚度，因此可以忽略远程散射。

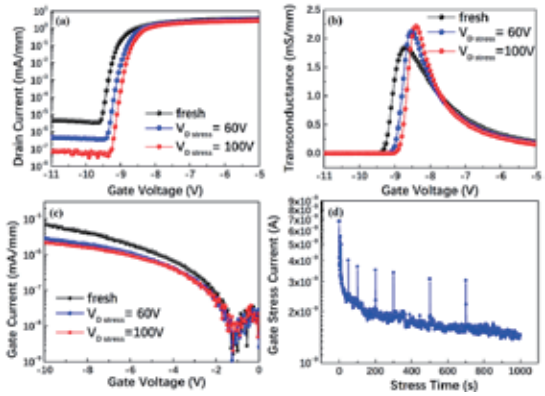


图 2 器件在 T=300K，$V_{Dstress}$ = 60V 和 100V，$V_{Gstress}$ = V_{TH} (0) +4V 偏置下进行应力实验 1000s 后，器件 (a) 转移特性 (b) 跨导特性 (c) 栅电流特性 (d)$V_{Gstress}$ = $V_{TH}(0)$ + 4V 和 $V_{Dstress}$ = 100V 时的栅极应力电流

Fig.2 (a) I_D–V_G (b) transconductance (c) I_G–V_G characteristics of devices with bias stress conditions of T = 300K, $V_{Dstress}$ = 60V or 100V, $V_{Gstress}$ = V_{TH} (0) +4 V for 1000s, where V_{TH} (0) denotes the pre–stress threshold voltage. The transfer measurements were carried out at V_D = 0.1V and the I_G–V_G curves were measured with V_S = 0V and drain floating (d) Gate current during stress test with $V_{Gstress}$ = V_{TH} (0) +4V and $V_{Dstress}$ = 100V

为了进一步了解跨导的变化机制，我们进行了多次应力实验。$V_{Dstress}$ 在所有应力测试中都固定为 100V，而 $V_{Gstress}$ 则不同。$V_{Gstress}$ 从 $V_{TH}(0)$ + 1V 变化到 $V_{TH}(0)$ + 7V。应力测试期间的温度为 300K。应力持续时间为 1000 秒。图 3(a) 是归一化 $G_{m,max}$ 与 $V_{Gstress}$ $-V_{TH}(0)$ 的关系。归一化 $G_{m,max}$ 首先出现增加，直到在 $V_{Gstress}$ $-V_{TH}(0)$ = + 3V 时达到 1.23，然后随着 $V_{Gstress}$ 的增大而减小。最后，它在 $V_{Gstress}$ $-V_{TH}(0)$ = + 7V 时达到 1.02。图 3(a) 中归一化 $G_{m,max}$ 与 $V_{Gstress}$ $-V_{TH}(0)$ 的趋势是非单调的“钟形”。

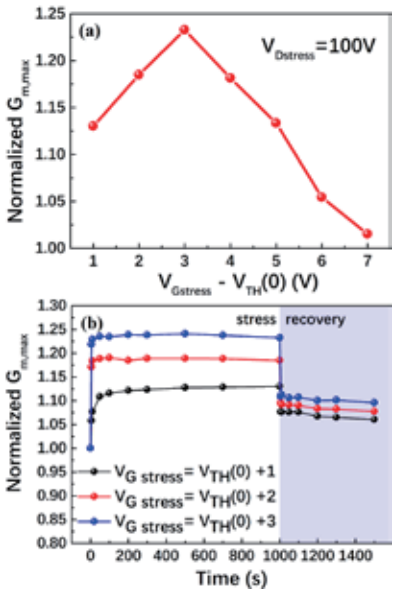


图 3 (a) 归一化跨导峰值特性 (b) 应力和恢复阶段跨导变化

Fig.3 (a) Normalized $G_{m,max}$ characteristic after 1000s of stress as a function of $V_{Gstress}$–$V_{TH}(0)$ (b) Variations of normalized $G_{m,max}$ during stress and recovery

由于与 $V_{Dstress}$ 相比，栅漏电压 (V_{GD}) 的变化很小，因此栅漏区电场的变化被认为可以忽略不计。随着 $V_{Gstress}$ 从 $V_{TH}(0)$ + 1V 增加到 V_{TH} (0) + 7V，电子的浓度变大。因此，电子被栅漏区的高电场“加热”，变成热电子。理论上，当 $V_{Gstress}$ 增加时，GaN 沟道层中热电子的浓度增加。然而，高 $V_{Dstress}$ 引起的高功率耗散使器件温度升高。$V_{Gstress}$ 越高，器件中的温度越高。由于载流子的迁移率随着温度的升高而降低，因此随着 $V_{Gstress}$ 的增加，载流子迁移率的劣化更加严重[17]。这导致当 $V_{Gstress}$ 增加时热电子的浓度降低。因此，热载流子应力 (HCS) 期间热电子的浓度是电

子浓度和应力测试期间器件温度之间平衡的结果。因此，我们可以假设图3(a)中$G_{m,max}$的非单调“钟形”特征与热电子效应有关。

为了理解器件归一化$G_{m,maxx}$的变化与热电子效应之间的关系，研究了$G_{m,max}$增量较大的区域。归一化$G_{m,max}$在$V_{Dstress}$ = 100V和$V_{Gstress}$在$V_{TH}(0)$ +1V到$V_{TH}(0)$ +3V下300K温度时1000秒应力的变化如图3(b)所示，其中紫色区域显示恢复期间归一化$G_{m,max}$的变化。归一化$G_{m,max}$表现出部分快速恢复，然后在接下来的时间内变化很小，这表明在应力期间有两种不同的机制共同起作用。由于载流子的迁移率是影响线性区工作的器件G_m的主要因素之一[18]，在电应力条件下陷阱态散射率的变化在接下来的一些段落中进行详细讨论。

由于氢可以在MOCVD生长过程中和器件制备过程中掺入到GaN中，因此器件中存在大量氢化缺陷[19,20]。尽管氢化Ga空位的形成能低于氢化N反位的形成能意味着氢化Ga空位很重要，但是当在富含NH_3的条件下通过MOCVD生长GaN时，高浓度的NH_3分子位于GaN表面。因此，Ga空位可以通过放热反应轻松捕获NH_3，这表明在富含NH_3的条件下，氢化N反位缺陷的浓度显然更为重要[8-21]。$[N_{Ga}H_3]^{-1}$通常充当库仑散射中心，会影响载流子的迁移率。在HCS过程中，$[N_{Ga}H_3]^{-1}$复合物可以在具有足够能量的热电子影响下释放一个氢原子，然后改变它的价态，这将影响器件的特性[22]。为了进一步了解$[N_{Ga}H_3]^{-1}$复合物在HCS下的变化以及对器件的影响，我们进行了如下DFT计算。

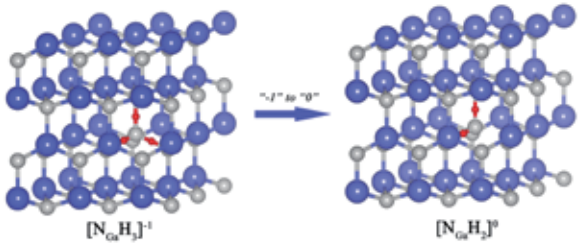


图4 $[N_{Ga}H_3]^{-1}$到$[N_{Ga}H_2]^0$球棍模型，紫色小球为Ga，灰色小球为N，红色小球为H

Fig.4 Dehydrogenation of $[N_{Ga}H_3]^{-1}$ into $[N_{Ga}H_2]^0$, purple spheres are Ga, gray spheres are N and red spheres are H

图4是氢化缺陷的球棍模型，包括$[N_{Ga}H_3]^{-1}$和$[N_{Ga}H_2]^0$。在HCS过程中，$[N_{Ga}H_3]^{-1}$会失去一个H原子，从而变成$[N_{Ga}H_2]^0$。为了获得$[N_{Ga}H_3]^{-1}$到$[N_{Ga}H_2]^0$脱氢所需的能量，使用DFT和Perdew-Burke-Ernzerhof广义梯度近似(PBE-GGA)[23]进行计算。PBE-GGA通常会导致GaN中缺陷的能量和特性不准确。为了确保计算出的缺陷状态更准确，我们根据文献[24]-[27]对特定原子使用了带有+U项的GGA。由于使用GGA+U方法且U(Ga) = 7.5eV和U(N) = 5eV的GaN能带与实验结果一致，我们使用U(Ga) = 7.5eV和U(N) = 5eV。在我们的计算中使用了具有64个N原子和64个Ga原子的4×4×2超胞。布里渊区使用由Monkhorst-Pack方法生成的5×5×3k点网格进行采样。

电荷态为q的氢化氮反位点[10,20,28]的形成能定义为：

$$E_{form}\left[(N_{Ga}H_n)^q\right]=E_{tot}\left[(N_{Ga}H_n)^q\right]-E_{tot}[GaN,bulk]+$$

$$\mu_{Ga}-\mu_N-n\mu_H+q\left[E_F+E_V+\Delta V\right]+E_{corr}^q \qquad (1)$$

这里，$E_{tot}[(N_{Ga}H_n)^q]$是从超胞计算得出的总能量，q是价态。$E_{tot}[GaN,bulk]$是体GaN超胞的总能量。μ_{Ga}、μ_N和μ_H分别对应于Ga、N和H原子的化学势。n是产生缺陷时超胞中的H原子数。E_F和E_V分别是GaN体中的费米能级和价带最大值的能量。ΔV表示将缺陷超胞中的参考电位与体中的参考电位对齐的校正值。$E_{corr}{}^q$对应于对超级单元有限大小的修正。化学势取决于所分析的缺陷[22,23]。在计算中，GaN块体的平衡提供了对Ga和N化学势的约束：

$$\mu_{Ga}+\mu_N=\mu_{GaN}^{bulk} \qquad (2)$$

其中$\mu_{GaN}{}^{bulk}$是体GaN纤锌矿相中每对Ga-N的能量。由于本工作中使用的AlGaN/GaN异质结构是通过MOCVD生长的，因此N化学势设置为

$$\mu_N=1/2\mu_{N_2} \qquad (3)$$

其中μ_N表示N_2气体的能量。因此，Ga化学势固定为：

$$\mu_{Ga} = \mu_{GaN}^{bulk} - 1/2\,\mu_{N_2} \quad (4)$$

然后，将 H 化学势设置为

$$\mu_H = \left(2\mu_{NH_3} - \mu_{N_2}\right)/6 \quad (5)$$

图 5 描绘了通过 MOCVD 生长的块状 GaN 中缺陷（$[N_{Ga}H_3]^{-1}$ 和 $[N_{Ga}H_2]^0$）的形成能与费米能量的关系。使用 Silvaco Atlas 确定应力之前和应力期间 GaN 中费米能级的位置。由于电子在 HCS[8,9] 下可以获得 1 ~ 3eV 范围内的能量，因此能量超过 1.55eV 的电子有可能将 $[N_{Ga}H_3]^{-1}$ 脱氢为 $[N_{Ga}H_2]^0$。为了更形象地展示脱氢效应，图 6 中用 GaN 的晶格结构代替了 GaN。如图 6(b) 所示，GaN 沟道中的高能电子可以将一个 H 原子从 $[N_{Ga}H_3]^{-1}$ 变为中性 $[N_{Ga}H_2]^0$。缺陷是从 $[N_{Ga}H_3]^{-1}$ 到 $[N_{Ga}H_2]^0$ 的转换可以降低 GaN 沟道层中的散射率，从而使 $G_{m,max}$ 在应力期间变得更高。

由于 $G_{m,max}$ 在撤销电应力后显示部分快速恢复的现象，如图 3(b) 所示，我们假设存在另一种影响 G_m 的机制并且是可恢复的。由于当 Al 摩尔分数小于 0.3 时，O_N 缺陷在 AlGaN 层中充当浅施主，因此在应力测试之前，它们是 AlGaN 层中带正电的电离缺陷 [10]。如图 6(b) 所示，在 HCS 期间，AlGaN 势垒层中漏极靠栅极边缘的电离施主陷阱可以被从沟道注入的高能电子填充，形成中性陷阱 [6,10,29]。因此，施主陷阱的中性化可以提高 GaN 沟道层中载流子的迁移率，导致应力期间 $G_{m,max}$ 的部分增加。当去除电应力时，电子可以通过热激发从 AlGaN 势垒层中的中性施主陷阱中逸出，再次形成电离施主陷阱，这是 $G_{m,max}$ 部分快速恢复的原因。由于脱氢更难恢复，脱氢引起的跨导增量几乎是恒定的。总之，如图 6 所示，GaN 沟道层中 $[N_{Ga}H_3]^{-1}$ 到 $[N_{Ga}H_2]^0$ 的脱氢效应与 AlGaN 势垒层中的施主陷阱的中和并行进行，以降低散射率，从而导致归一化 $G_{m,max}$ 的增加。

HCS 也在 300K 到 360K 的不同温度下进行，其中 $V_{Dstress}$=100V 和 $V_{Gstress} = V_{TH}(0) + 3V$，持续 1000 秒。图 7 比较了不同温度下 HCS 期间 V_{TH} 和归一化 $G_{m,max}$ 的变化。在更高的温度下，电子可以从 SiN 层中的陷阱中逸出，导致 V_{TH} 的正偏移减小 [2,30]。同时，归一化 $G_{m,maxx}$ 的增量随着温度升高而减小。表 1 显示了不同温度下应力测试期间的漏极电流。应力测试期间的漏极电流随着温度升高而降低。这说明载流子的迁移率降低。因此，热载流子的浓度降低，这削弱了 $[N_{Ga}H_3]^{-1}$ 到 $[N_{Ga}H_2]_0$ 的脱氢作用以及在高温应力测试期间对施主陷阱的中和作用。因此，归一化 $G_{m,max}$ 的增量随着温度的升高而减小。$G_{m,max}$ 随着温度升高的行为可以进一步证实上述机制。

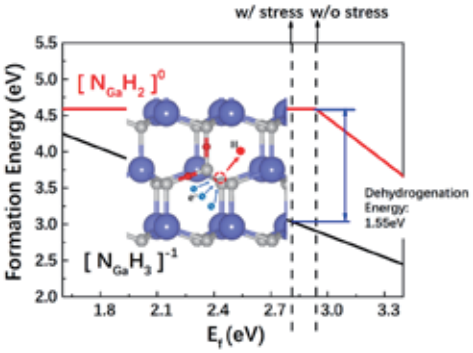


图 5 氢化 N 反位的形成能。虚线为有无应力时 GaN 费米能级的位置

Fig.5 Formation energy of hydrogenated nitrogen antisites. The dashed lines show the positions of the Fermi level with and without stress

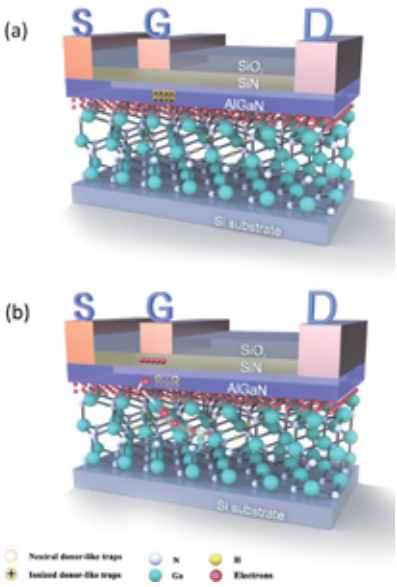


图 6 热载流子应力 (a) 前、(b) 后器件变化

Fig.6 Schematic cross-sections of SiN/AlGaN/GaN MIS-HEMTs (a) before and (b) after the HCS experiments

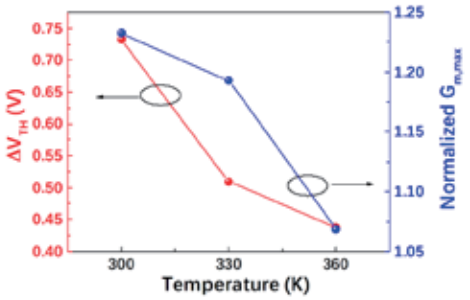


图 7 不同温度下应力后阈值电压和归一化峰值跨导特性

Fig.7 Variations of threshold voltage and normalized $G_{m,max}$ characteristics after stress tests for 1000s with bias stress conditions of $V_{Dstress}$ = 100V, $V_{Gstress}$ = V_{TH} (0) +3V as a function of temperature

表 1 不同温度下应力后漏极应力电流值

Tab.1 Drain stress current at various temperature conditions with $V_{Dstress}$ = 100V and $V_{Gstress}$ = V_{TH}(0)+3V

T (K)	$I_{Dstress}$ (mA/mm)
300	56.78
330	54.98
360	49.13

3 结论

我们已经讨论了 HCS 下 GaN 基 MIS-HEMT 的退化特性。在 300K 下进行 HCS 后，器件的 $G_{m,max}$ 出现异常增加。峰值跨导的增加可归因于 AlGaN 势垒层中的施主陷阱捕获从 GaN 沟道层注入的电子，以及来自 GaN 沟道层的高能电子将 $[N_{Ga}H_3]^{-1}$ 变至 $[N_{Ga}H_2]^0$ 的脱氢效应。$[N_{Ga}H_3]^{-1}$ 脱氢的能量由 DFT 计算，在热电子的能量范围内。还讨论了在不同温度下应力后器件特性，它们的变化现象也与上述机制一致。这说明降低材料中预先存在的缺陷密度对于 GaN 基 MIS-HEMT 的可靠性具有重要意义。

参考文献（References）

[1] MENEGHESSO G, MENEGHINI M, ROSSETTO I, et al. Reliability and parasitic issues in GaN-based power HEMTs: a review [J]. Semicond.Sci. Technol, 2016, 31(9): 093004.

[2] MENEGHINI M, ROSSETTO I, BISI D, et al. Negative Bias-Induced Threshold Voltage Instability in GaN-on-Si Power HEMTs [J]. IEEE Electron Device Lett, 2016, 37(4): 474-477.

[3] MENEGHESSO G, VERZELLESI G, DANESIN F, et al. Reliability of GaN high-electron-mobility transistors: State of the art and perspectives [J]. IEEE Trans.Device Mater. Reliab, 2008, 8(2): 332-343.

[4] MENEGHINI M, RONCHI N, STOCCO A, et al. Investigation of Trapping and Hot-Electron Effects in GaN HEMTs by Means of a Combined Electrooptical Method [J]. IEEE Trans. Electron Devices, 2011, 58(9): 2996-3003.

[5] TAPAJNA M, KILLAT N, PALANKOVSKI V, et al. Hot-Electron-Related Degradation in InAlN/GaN High-Electron-Mobility Transistors [J]. IEEE Trans. Electron Devices, 2014, 61(8): 2793-2801.

[6] JIANG R, SHEN X, FANG J, et al. Multiple Defects Cause Degradation After High Field Stress in AlGaN/GaN HEMTs [J]. IEEE Trans.Device Mater. Reliab, 2018, 18(3): 364-376.

[7] JIABG R, SHEN X, CHEN J, et al Degradation and annealing effects caused by oxygen in AlGaN/GaN high electron mobility transistors [J]. Appl. Phys. Lett, 2016, 109(2): 023511.

[8] ROY T, PUZYREV Y S, TUTTLE B R, et al. Electrical-stress-induced degradation in AlGaN/GaN high electron mobility transistors grown under gallium-rich, nitrogen-rich,and ammonia-rich conditions [J]. Appl. Phys. Lett, 2010, 96(13): 133503.

[9] PUZYREV Y S, MUKHERJEE S, CHEN J, et al. Gate Bias Dependence of Defect-Mediated Hot-Carrier Degradation in GaN HEMTs [J]. IEEE Trans. Electron Devices, 2014, 61(5): 1316-1320.

[10] Van de Walle C G and Neugebauer J. First-principles calculations for defects and impurities: Applications to III-nitrides [J]. J. Appl. Phys, 2004, 95(8): 3851-3879.

[11] WRIGHT A F.Substitutional and interstitial oxygen in wurtzite GaN [J]. J. Appl. Phys, 2005, 98(10): 103531.

[12] ROY T, ZHANG E X, Puzyrev Y S, et al. Temperature-dependence and microscopic origin of low frequency 1/f noise in GaN/AlGaN high electron mobility transis-

tors [J]. Appl. Phys. Lett, 2011, 99(20): 203501.

[13] MENEGHINI M ,BISI D, MARCON D, et al. Trapping in GaN-based metal-insulator-semiconductor transistors: Role of high drain bias and hot electrons [J] Appl. Phys. Lett, 2014,104(14): 143505.

[14] LIU Z H, NG G I, ARULKUMARAN S, et al. Improved two-dimensional electron gas transport characteristics in AlGaN/GaN metal-insulator-semiconductor high electron mobility transistor with atomic layer-deposited Al2O3 as gate insulator [J]. Appl. Phys. Lett, 2009, 95(22): 223501.

[15] HUNG T H, ESPOSTO M, AND RAJAN S. Interfacial charge effects on electron transport in III-Nitride metal insulator semiconductor transistors [J]. Appl. Phys. Lett, 2011, 99(16): 162104.

[16] JI D, LIU B, LU Y, et al. Polarization-induced remote interfacial charge scattering in Al2O3/AlGaN/GaN double heterojunction high electron mobility transistors [J]. Appl. Phys. Lett, 2012, 100(13): 132105.

[17] ZHU J, ZHU Q, CHEN L, et al. Impact of Recess Etching on the Temperature-Dependent Characteristics of GaN-Based MIS-HEMTs With Al2O3/AlN Gate-Stack [J]. IEEE Trans. Electron Devices, 2017, 64(3): 840-847.

[18] GRECO G, FIORENZA P, IUCOLANO F, et al. Conduction Mechanisms at Interface of AIN/SiN Dielectric Stacks with AIGaN/GaN Heterostructures for Normally-off High Electron Mobility Transistors: Correlating Device Behavior with Nanoscale Interfaces Properties [J]. ACS Appl. Mater. Interfaces, 2017, 9(40): 35383-35390.

[19] CHRIS G, VAN D W. Interactions of hydrogen with native defects in GaN [J]. Phys. Rev. B, 1997, 56(16): 10020-10023.

[20] WRIGHT A F. Interaction of hydrogen with gallium vacancies in wurtzite GaN [J]. J. Appl. Phys, 2001, 90(3): 1164-1169.

[21] PANTELIDES S T,PUZYREV Y, SHEN X, et al. Reliability of III-V devices - The defects that cause the trouble [J]. Microelectron. Eng, 2012, 90: 3-8.

[22] PUZYREV Y S, ROY T, BECK M, et al. Dehydrogenation of defects and hot-electron degradation in GaN high-electron-mobility transistors [J]. J. Appl. Phys, 2011, 109(3): 034501.

[23] GAO Y, SUN D, JIANG X, et al. Point defects in group III nitrides: A comparative first-principles study [J]. J. Appl. Phys, 2019, 125(21): 215705.

[24] LYONS J L and Van de WALLE C G. Computationally predicted energies and properties of defects in GaN [J]. NPJ Comput. Mater, 2017, 3: 12.

[25] ZAKRZEWSKI T and BOGUSLAWSKI P. Electronic structure of transition metal ions in GaN and AlN: Comparing GGA plus U with experiment [J]. J. Alloys Compd, 2016, 664: 565-579.

[26] MAGNUSON M, MATTESINI M, HOGLUND C, et al. Electronic structure of GaN and Ga investigated by soft x-ray spectroscopy and first-principles methods [J]. Phys. Rev. B, 2010, 81(8): 085125.

[27] LEI J, ZHU D P, XU M C , et al. First-principles simulations of two dimensional electron gas near the interface of ZnO/GaN (0001) superlattice [J]. Phys. Lett A, 2015, 379(38): 2384-2387.

[28] WRIGHT A F.Interaction of hydrogen with nitrogen interstitials in wurtzite GaN [J]. J. Appl. Phys, 2001, 90(12): 6526-6532.

[29] CAESAR M, DAMMANN M, POLYAKOV V, et al. Generation of traps in AlGaN/GaN HEMTs during RF- and DC-stress test. in Reliability Physics Symposium (IRPS), 2012, pp. CD. 6.1-6.5.

[30] RUZZARIN M, MENEGHINI M, ROSSETTO I, et al. Evidence of Hot-Electron Degradation in GaN-Based MIS-HEMTs Submitted to High Temperature Constant Source Current Stress. IEEE Electron Device Lett, 2016, 37(11): 1415-1417.

作者简介：

牛雪锐(1997—),女,陕西汉中人,在读博士，专注于 GaN 基器件可靠性以及互补逻辑器件研究。

采用富硅 SiN/Si_3N_4 双层钝化 AlGaN/GaN 高电子迁移率晶体管的功率特性及机理

刘捷龙，宓珉瀚，祝杰杰，侯 斌，杨 凌，王 宏，马晓华，郝 跃

（西安电子科技大学，陕西省 西安市 710071）

摘　要：本文研究了富硅 SiN/Si_3N_4 双层钝化的高电子迁移率晶体管 (HEMTs)。使用富硅的 SiN 插入层可以改善沟道传输性能、关态泄漏电流、电流崩塌、功率性能和随温度变化的稳定性。在 300K 到 420K 温度范围内，未进行富 Si SiN 钝化的器件泄漏电流增加 3 个数量级以上，电流崩塌量由 9.7% 增加到 24.7%；而有富硅 SiN 钝化的器件泄漏电流温度随温度变化非常小，电流崩塌量维持在 5% 左右。在 17GHz 时，具有富硅 SiN 插入层钝化的器件输出功率密度为 7W/mm，峰值功率附加效率 (PAE) 为 56%。采用含富硅 SiN 层钝化技术可以提高器件的功率性能，这归因于在高沟道温度下抑制了电流崩塌并且提高器件稳定性。

关键字：高电子迁移率晶体管；富硅 SiN；电流崩塌；沟道温度

中图分类号：TN325　　**文献标识码**：A

Improved Power Performance and the Mechanism of AlGaN/GaN HEMTs Using Si-rich SiN/Si_3N_4 Bilayer Passivation

Liu Jielong, Mi Minhan, Zhu Jiejie, Hou Bin, Yang Ling, Wang Hong, Ma Xiaohua, Hao Yue

(Xidian University, Xi'an 710071, China)

Abstract：AlGaN/GaN high electron mobility transistors (HEMTs) with Si-rich SiN/Si_3N_4 bilayer passivation were studied in this paper. The use of Si-rich SiN interlayer leads to improved channel transport property, off-state leakage current, current collapse, power performance, and temperature-dependent stability. HEMTs without Si-rich SiN interlayer passivation exhibit an increase in gate leakage current by over 3 orders of magnitudes, and an increase in current collapse from 9.7% to 24.7% with temperature increasing from 300K to 420K; while the devices with Si-rich SiN passivation exhibit weak temperature-dependency of leakage current and a constant current collapse about 5%. At 17GHz, devices with Si-rich SiN interlayer passivation exhibit an output power density of 7W/mm and a peak power-added efficiency (PAE) of 56%. The improved power performance of HEMTs using Si-rich SiN interlayer passivation is attributed to the suppressed current collapse and superior device stability under high channel temperature.

Key words： high electron mobility transistors; Si-rich SiN; current collapse; high channel temperature

0 引言

氮化镓基高电子迁移率晶体管 (HEMTs) 由于其优异的材料和异质结而被广泛应用于下一代射频功率和电力电子领域 [1]。然而，目前基于 GaN 的高电子迁移率晶体管还存在一些可靠性问题，如栅极泄漏电流和电流崩塌 [2]。电流崩塌在射频功率器件中起着至关重要的作用，它会降低输出功率和效率。以往的研究表明，电流崩塌可能是由于位于栅－漏区域的表面陷阱，在栅和漏偏置应力后产生带负电荷的虚栅。这种虚栅同时导致沟道电子耗尽和导通电阻增加 [3]，称为电流崩塌。通常，可以通过优化材料生长技术来缓解缓冲区中的陷阱。

有许多方法可以减轻表面陷阱的影响，如引入场板结构和表面钝化 [4]。其中 SiN 钝化可以有效地抑制电流崩塌，并且提高稳定性，进而成为最普遍的解决方案。通过优化 SiN 工艺条件，发现富硅的 SiN 极

基金项目：国家自然科学基金（批准号：11690042，61904135）和陕西省自然科学基金（批准号：2020JQ-316）资助的课题
通信作者：马晓华；E-mail:xhama@xidian.edu.cn

大地抑制了电流崩塌。富硅的 SiN 层钝化也被应用于 W 波段器件，在 96GHz 下实现了 3W/mm 的输出功率密度[5]。然而，富硅 SiN 技术有助于表面钝化和功率性能的原因在以往的工作中尚未得到进一步的研究。对于输出功率高的 GaN 器件，由于严重的功率损耗和热效应，沟道温度将远高于室温[6]。钝化层与氮化物界面之间陷阱态的捕获率和发射率是温度的指数函数，钝化技术将通过改善温度依赖特性和降低陷阱态对器件性能产生重大影响[7]。因此，有必要进一步分析不同钝化的 GaN 射频器件的温度依赖关系和机理[8]。

本文开发了一种用于 AlGaN/GaN HEMTs 的富硅 SiN/Si_3N_4 双层钝化技术。我们发现，传统 SiN 和 AlGaN 层之间的富硅 SiN 层可以有效的改善电流崩塌、表面漏电流、方块电阻和射频功率性能。在 17GHz 时，器件的功率密度为 7W/mm，峰值功率附加效率 (PAE) 为 56%。通过对比不同沟道温度下的钝化发现，采用富硅 SiN 插入层有助于抑制器件在高温下的退化，降低热效应造成的功率损耗，实现更高的射频输出功率和效率。

1 器件制作

采用金属有机气相沉积的方法在 SiC 衬底上生长外延异质结构。从下到上依次为 1.6μm Fe 掺杂 GaN 缓冲层、400nm 非故意掺杂 GaN 沟道层、1nm AlN 插入层、20nm $Al_{0.25}Ga_{0.75}N$ 势垒层和 2nm 非故意掺杂的 GaN 帽层。电子浓度和室温霍尔迁移率分别为 $1\times10^{13}cm^{-2}$ 和 $1900cm^2\cdot V^{-1}\cdot s^{-1}$。

该器件的制作从使用合金 Ti/Al/Ni/Au 金属堆栈形成欧姆接触开始，然后在 840 ℃ 的氮气环境下快速热退火 50 秒。器件隔离采用多次能量的氮离子注入。采用传输线模型得到欧接触电阻 R_c=0.37Ω · mm。在 SiN 钝化前，用有机溶液清洗去除表面有机污染物，用氨水去除表面天然氧化物。然后使用氮气稀释的硅烷（SiH_4）和氨气（NH_3）的等离子体增强型化学气相沉积（PECVD）设备沉积 SiN。采用 120nm 标准化学计量比 Si_3N_4（2%SiH_4:NH_3=100:2sccm）对传统的钝化结构进行钝化。对于双层氮化硅结构，第一层为厚度为 20nm 的富硅氮化硅（2%SiH_4:NH_3=450:2sccm），第二层为厚度为 100nm 的标准化学计量氮化硅。

采用电子束光刻和电感耦合等离子体 (ICP) 等离子体刻蚀技术对薄膜开孔形成了 T 形栅脚。在栅帽光刻之后，用 O_2 等离子体处理栅极区域，在栅极金属蒸发之前去除可能残留的光刻胶。采用电子束蒸发法制备 Ni/Au 多层栅电极，栅长为 150nm，栅帽长度为 0.60μm；栅源间距为 1.15μm，栅漏间距为 1.70μm；栅宽为 2×50μm。

含富硅 SiN 插入层的 AlGaN/GaN 器件的截面示意图如图 1（a）所示。扫描电子显微镜 (SEM) 显示富硅 SiN 与 Si_3N_4 有明显的分层现象，如图 1（b）所示。两种 SiN 薄膜的 FTIR 吸收光谱如图 1 (c) 所示，在 632nm 波长下，用椭偏仪测得的折射率分别为 1.9 和 2.4。傅里叶变换红外光谱（FTIR）测量范围从 $400cm^{-1}$ 到 $4000cm^{-1}$。富硅薄膜中硅 − 氢键相对含量约为 Si_3N_4 薄膜中硅 − 氢键相对含量的 3 倍。

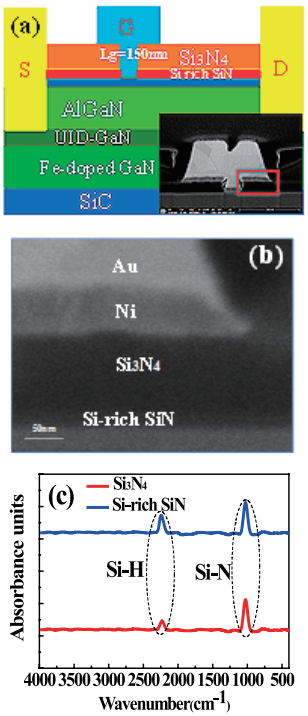


图 1 (a) 富硅 SiN 插入层的 AlGaN/GaN HEMTs 器件截面图及扫描电镜图 (b) 钝化层放大图 (c) 两种 SiN 薄膜的 FTIR 吸收光谱
Fig.1 (a) Cross section of AlGaN/GaN HEMTs with Si−rich SiN interlayer and the SEM of devices (b) Enlarged figure of passivation layer (c) FTIR absorption spectrum of Si_3N_4 and Si−rich SiN films

2 结果讨论

2.1 直流和脉冲特性

采用吉时利 4200SCS 半导体器件分析仪对直流和脉冲性能进行了表征。SiN 钝化后，室温下含富硅 SiN 和不含富硅 SiN 的器件 TLM 提取的方块电阻分别为 318Ω/sq 和 341Ω/sq，如图 2 所示。含富硅 SiN 钝化和常规 Si_3N_4 钝化的器件方块电阻随温度变化的斜率分别为 3.87Ω/(sq · K) 和 4.75Ω/(sq · K)。

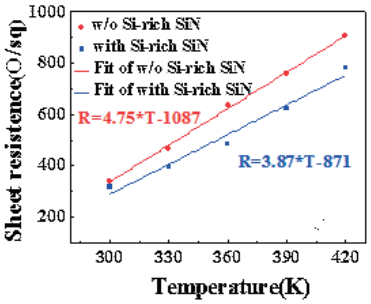


图 2 TLM 法提取金属宽度为 100μm 的温度依赖的方块电阻

Fig.2 Temperature-dependent Rsh extracted with 100μm metal pad width by TLM

为了确定差异的来源是由于迁移率还是载流子浓度，在室温下，我们通过霍尔测量得到了两个样品的迁移率和载流子密度，如表 1 所示。迁移率基本相同，但载流子浓度不同，说明导致不同的原因主要是载流子浓度的提高。

表 1 不同钝化后的霍尔测试结果

Tab.1 Hall test results after different passivation

参数	电子浓度	霍尔迁移率
	cm^{-2}	$cm^2 \cdot V^{-1} \cdot s^{-1}$
w/o Si-rich SiN	1.10×10^{13}	1917
with Si-rich SiN	1.25×10^{13}	1921

为了更好地理解钝化机理，我们测量了不同富 Si SiN 厚度的 AlGaN/GaN 异质结钝化后的室温霍尔，如图 3 所示。载流子浓度不随 SiN 厚度变化而变化，说明钝化层抑制了表面态到 2DEG 的损耗效应。因此，更低的 2DEG 薄片电阻是由于增加的 2DEG 密度，减轻了 2DEG 表面状态的损耗，而不是应力。

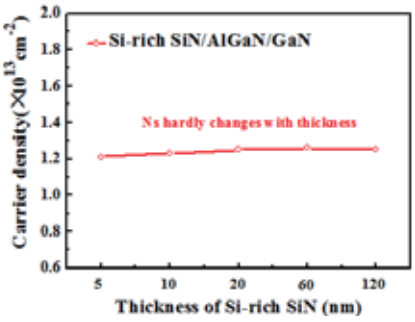


图 3 不同钝化层厚度的载流子浓度

Fig.3 Carrier concentration with differet passivation layer thickness

所测得的晶体管的转移特性如图 4（a）所示。含有富硅 SiN 和常规 Si_3N_4 钝化的器件关态电流（@V_g=−10V）分别为 2.1×10^{-5}mA/mm 和 7.4×10^{-5}mA/mm。对于含富硅 SiN 和常规 Si_3N_4 钝化的器件，其峰跨导分别为 282mS/mm 和 261mS/mm，器件跨导提升的原因是采用富 Si SiN 钝化后器件方块电阻减小。图 4（b）显示了输出特性，在栅极偏压为 2V 时，含富硅钝化器件的输出曲线的最大漏极电流为 1292mA/mm，大于无富硅钝化器件的 1229mA/mm。富硅器件的开态电阻为 2.39Ω · mm，常规 Si_3N_4 钝化的器件开态电阻为 2.85Ω · mm。

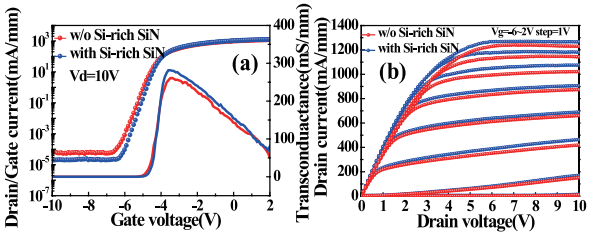


图 4 2×50μm 富硅 SiN 和常规 Si_3N_4 钝化器件的 (a) 转移特性 (b) 输出特性

Fig.4 (a) Transfer and (b) output characteristics of 2×50μm devices with Si-rich SiN and without Si-rich SiN

2.2 温度依赖的栅极漏电和电流崩塌

图 5 为不同温度下的 AlGaN/GaN 肖特基栅特性。常规 Si_3N_4 钝化器件的反向漏电流表现出强烈的温度和电压依赖关系。随着温度的升高，反向电流大大增加。当反向电压超过 2V 时，栅漏电流随着温度

的升高越来越明显。富硅 SiN 器件的反向漏电流则表现出微弱的温度和电压依赖关系。对于富硅 SiN 钝化的器件，观察到随着温度的升高栅极电流增加幅度较小。这是由于富硅的 SiN/Si_3N_4 双层钝化降低了表面的陷阱态密度。

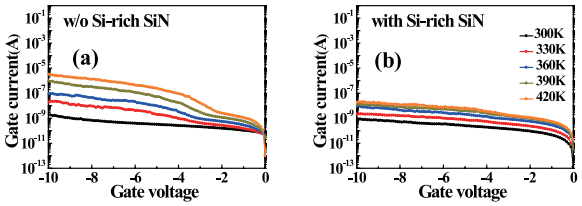


图 5 不同温度下 2×50μm 器件的反向肖特基栅漏电流曲线（a）无富硅 SiN（b）有富硅 SiN

Fig.5 Temperature-dependent reverse schottky-gate leakage current of 2×50μm devices (a) without and (b) with Si-rich SiN

为了评估表面陷阱对电流崩塌的影响，我们测量了两个样品在 300K 到 420K 范围内的脉冲特性，测量曲线及电流崩塌比较如图 6 所示。在 420K 时，无富硅 SiN 器件的电流崩塌为 24.7%，而富硅 SiN 钝化的器件的电流崩塌为 5.4%。结果表明，在高温条件下，富硅 SiN 钝化的器件的电流崩塌明显优于无富硅 SiN 钝化的器件。当栅漏电子注入到栅靠近漏区域时，表面虚栅效应更加明显，导致电流崩塌加剧。两个样品的电流崩塌差异如此之大的原因是富硅 SiN 钝化阻止了反向栅极泄漏，减少了栅极电子注入到栅极靠近漏端一侧，最终缓解了电流崩塌。

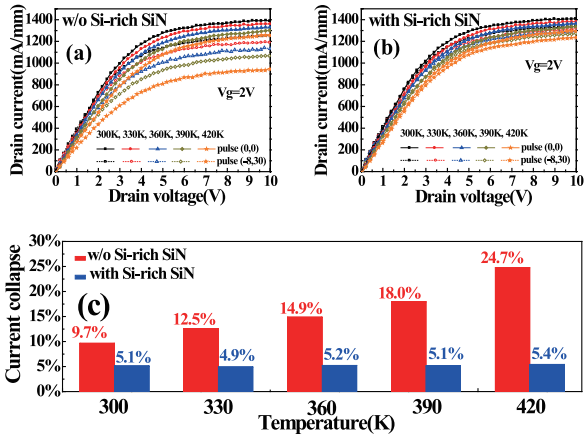


图 6 2×50μm 器件的温度相关脉冲输出曲线（a）无富硅 SiN 钝化（b）富硅 SiN 钝化（c）两种样品电流崩塌的比较

Fig.6 Temperature-dependent pulse output curves for 2×50μm devices (a) without and (b) with Si-rich SiN passivation (c) Comparison of current collapse of two samples

2.3 小信号特性

使用安捷伦 8363B 网络分析仪对器件在 100MHz 到 40GHz 范围内的小信号射频特性进行了表征，结果如图 7 所示。在 V_{ds}=10V 时，富硅 SiN 钝化器件的电流增益截止频率（f_T）为 71GHz，最大振荡频率（f_{max}）为 127GHz，无富硅钝化器件的电流增益截止频率（f_T）为 68GHz，最大振荡频率（f_{max}）为 117GHz。器件的小信号特性在加入 20nm 的富硅 SiN 后略有提高。

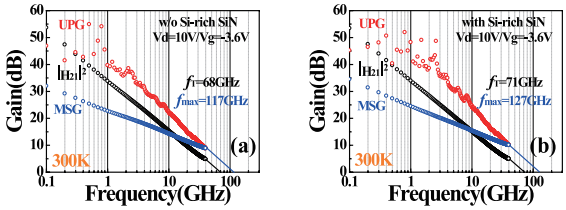


图 7 2×50μm（a）无富硅 SiN 钝化和（b）富硅 SiN 钝化器件的小信号特性

Fig.7 Small-signal characteristics of the 2×50μm devices (a) without and (b) with Si-rich SiN passivation

2.4 大信号特性

图 8 为 17GHz 连续波模式下器件的大信号功率性能。所有器件都偏置在 AB 类，源阻抗和负载阻抗分别匹配到最大增益和最大 PAE。在 V_{ds}=20V 时，具有富硅 SiN（无富硅 SiN）器件的饱和输出功率密度（P_{out}）为 5W/mm（4.4W/mm），PAE 峰值为 56%（51.2%）。在 V_{ds}=30V 时，P_{out} 为 7W/mm（5.8W/mm），PAE 峰值为 43.1%（37%）。

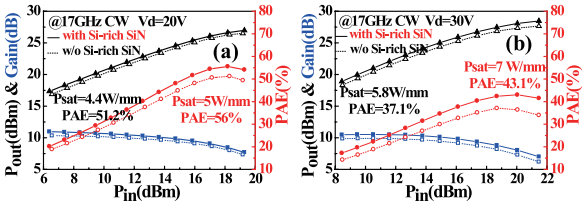


图 8 比较 2×50μm AlGaN/GaN HEMTs 在（a）V_{ds}=20V 和（b）V_{ds}=30V 时，没有和有富硅 SiN 钝化的情况下，在 17GHz 连续波功率性能

Fig.8 Comparison of CW power performance at 17GHz for 2×50μm AlGaN/GaN HEMTs without and with Si-rich SiN passivation at (a) V_{ds}=20V and (b) V_{ds}=30V, respectively

3 结论

本文开发了富硅 SiN（20nm）和 Si_3N_4（100nm）PECVD 工艺，有效钝化 AlGaN/GaN HEMTs。沟道输运特性、泄漏电流、电流崩塌、功率性能和温度依赖的稳定性通过插入薄的富硅 SiN 得到改善。在没有富硅插入层钝化的器件中，栅极漏电流随温度的升高从 300K 增加到 420K，增加超过 3 个数量级，电流崩塌从 9.7% 增加到 24.7%。然而，富硅钝化器件漏极最大电流对温度依赖性较弱，电流崩塌维持在约 5%。在 17GHz 下有富硅 SiN 钝化的器件输出功率密度为 7W/mm，PAE 峰值达 56%。

参考文献（References）

[1] WU Y F, KAPOLNEK D, IBBETSON J P, et al. Very-high power density AlGaN/GaN HEMTs [J]. IEEE Transactions on Electron Devices, 2001, 48(3): 586-590.

[2] WATANABE T, TERAMOTO A, NAKAO Y, et al. Low interface trap density and high breakdown electric field SiN films on GaN formed by plasma pretreatment using microwave-excited plasma-enhanced chemical vapor deposition [J]. IEEE Transactions on Electron Devices, 2016, 63(4): 1795-1801.

[3] WU Y F, SAXLER A, MOORE M, et al. 30-W/mm GaN HEMTs by field plate optimization [J]. IEEE Electron Device Letters, 2004, 25(3): 117-119.

[4] WANG X, HUANG S, ZHENG Y, et al. Robust SiNx/AlGaN interface in GaN HEMTs passivated by thick LPCVD-grown SiNx layer [J]. IEEE Electron Device Letters, 2015, 36(7): 666-668.

[5] MAKIYAMA K, OZAKI S, OHKI T, et al. Collapse-free high power InAlGaN/GaN-HEMT with 3 W/mm at 96 GHz [C]. IEEE International Electron Devices Meeting. Washington. USA, 2015: 9.1.1-9.1.4.

[6] JOH. J. ALAMO A DEL, CHOWDHURY U, et al. Measurement of channel temperature in GaN high-hlectron mobility transistors [J]. IEEE Transactions on Electron Devices, 2009, 56(12): 2895-2901.

[7] SUN H, WANG M, YIN R, et al. Investigation of the trap states and VTH instability in LPCVD Si3N4/AlGaN/GaN MIS-HEMTs with an in-situ Si3N4 interfacial layer [J]. IEEE Transactions on Electron Devices, 2009, 66(8): 3290-3295.

[8] HASHIZUME T, OOTOMO S, OYAMA S, et al. Chemistry and electrical properties of surfaces of GaN and GaN/AlGaN heterostructures [J]. Journal of Vacuum Science & Technology B Microelectronics & Nanometer Structures, 2001, 19(4): 1675-1681.

作者简介：

刘捷龙（1994—），男，陕西宝鸡人，在读博士，主要从事 GaN 微波器件研究。

基于浅槽刻蚀欧姆工艺的 AlGaN/GaN 器件功率特性提升技术

芦 浩，马晓华，杨 凌，侯 斌，霍 腾，司泽艳，张 濛，郝 跃

（西安电子科技大学，陕西省 西安市 710071）

摘 要：本文研究了采用 Ti/Au/Al/Ni/Au 金属叠层和浅槽刻蚀欧姆接触（STEOC）AlGaN/GaN 高电子迁移率晶体管的直流和射频性能。与传统的 Ti/Al/Ni/Au 欧姆接触相比，浅槽刻蚀欧姆工艺实现了 0.28Ω · mm 的低接触电阻和均方根（RMS）粗糙度仅为 6.3nm 的光滑表面形貌。并且，采用浅槽刻蚀欧姆工艺制造的晶体管显示出 1.52A/mm 的高输出电流和 2.09Ω · mm 的低导通电阻，晶体管实现了高电流截止频率（f_T）和最大振荡频率（f_{max}），为 60/150GHz。Sub-6GHz 连续波模式大信号测试显示出 67% 的高功率附加效率（PAE）。上述结果证明浅槽刻蚀欧姆工艺可以促进 GaN 在 5G 通讯功率放大器（PA）应用的巨大潜力。

关键词：铝镓氮 / 氮化镓；HEMT；射频；高电子迁移率晶体管

中图分类号：TN303 **文献标识码**：A

Shallow Trench Etching Ohmic Contact for Improved RF Power Performance of AlGaN/GaN Device

Lu Hao, Ma Xiaohua, Yang Ling, Hou Bin, Huo Teng, Si Zeyan, Zhang Meng, Hao Yue

(Xidian Univeristiy, Xi'an, 710071, China)

Abstract: In this article, we report the high dc and radio frequency(RF) performance of AlGaN/GaN high electron mobility transistors that ultilize Ti/Au/Al/Ni/Au metallization stack with shallow trench etching ohmic contact(STEOC). Low contact resistance of 0.28Ω·mm and highly smooth surface morphology with a root mean square(RMS) roughness of 6.3nm have been achieved simultaneously.. Consequently, the fabricated transistors with the STEOC process exhibit a high output current of 1.52A/mm and low ON-resistance(RON) of 2.09Ω·mm. In addition, a high current cutoff frequency(f_T) of 60GHz and maximum oscillation frequency(f_{max}) of 150GHz were obtained for the STEOC HEMT. Sub-6GHz continuous-wave mode power sweep measurements deliver a high power-added-efficiency(PAE) of 67%. These results show the significant potential of the STEOC process to facilitate the development of GaN-based power amplifier(PA) applied for 5G.

Key words: AlGaN/GaN; HEMT; radio frequency; high electron mobility transistor

0 引言

氮化镓凭借其卓越的频率和输出功率密度特性，已成为下一代无线基站功放器件极具竞争力的选择[1-3]。自从 1993 年美国南卡莱纳州立大学 Asif Khan 等人首次发明了 AlGaN/GaN 异质结场效应晶体管以来，国内外研究学者就一直致力于宽禁带氮化物的欧姆接触研究。随着工作环境要求的不断提高，欧姆接触技术已是制约 GaN 基微波器件性能的关键技术之一[4]。

目前 AlGaN/GaN 异质结的欧姆接触工艺是 Ti/Al/Ni/Au，工艺目前比较稳定，但是常规 Ti/Al/Ni/Au 金属堆栈在高温退火（HTA）后的过程中，会产生 Al-Au 或 Ni-Al 合金簇，将引入横向金属扩散和不均匀的欧姆表面粗糙度，这将恶化微波工作可靠性[5-7]。此外，源自欧姆金属边缘的金属尖锋可能会引入峰值电场，导致缓冲漏电流升高[8]。鉴于此，对于面向高频（如毫米波）应用，现有的 Ti/Al/Ni/Au 的高温退火欧姆接触工艺需要优化。

无论是面向高功率应用还是高频应用，具有低接触电阻（R_C）的欧姆接触、光滑的表面粗糙度以及可控的横向金属外扩是关键需求参数。针对此工艺难点，国际上提出了很多方案来解决此问题。2011 年，美国康奈尔大学提出欧姆凹槽刻蚀再生长的方法降低欧姆接触电阻，他们利用 MOCVD 的方法在 SiC 衬底上生长了 $In_{0.17}Al_{0.83}N$/AlN/GaN HEMT 结构，其中，InAlN 层的厚度为 5.6nm，然后在欧姆区刻蚀了 30nm 的深度后，利用 MBE 的方法在生长了 60nm 厚的 Si 掺杂的 N+ 重掺 GaN，最后淀积了 Ti 基金属叠层，经过 TLM 测试测得最终的欧姆接触电阻为 0.4±0.23Ω · mm；2013 年中国台湾交通大学 Yuen-Yee Wong 等人采用 Ti/Al/Ni/Cu 方案，使用 Cu 替代了 Au 帽层，接触电阻为 $1.35\times10^{-6}\Omega\cdot cm^2$，此方案可以解决 Au 帽层热退火过程中产生的 Al-Au 合金，优化表面粗糙度。2017 年，德国 IAF 实验室 Birte-Julia Godejohann 等人在 AlN/GaN 异质结上采用 Si 离子注入，并结合 1100℃的热退火激活来制备高 Al 组分的欧姆接触。经过接触电阻 TLM 测试，得到的接触电阻为 0.3Ω · mm，接触电阻率约为 $10^{-6}\Omega\cdot cm^2$。

尽管上述源漏 N+ 再生长 [9] 或 Si 离子注入 [10] 可以实现极低的欧姆接触，但过于复杂的工艺步骤和后期高昂的维护成本增加了产业推广的难度。通过传统的金属热退火方案来解决这个问题将仍是一个最优选择 [11]。

1 晶体管结构与制备流程

本工作采用的 AlGaN/GaN 异质结样品，材料结构示意图如图 1 所示。外延材料通过金属有机化学气相沉积（MOCVD）生长在 3 英寸半绝缘 4H-SiC 衬底。生长的外延层从下到上包括 AlN 成核层，Fe 掺杂缓冲层，非故意掺杂的 i-GaN 沟道层，22nm 未掺杂的 $Al_{0.25}Ga_{0.75}N$ 阻挡层和 3nm GaN 盖层。室温非接触霍尔测量显示方块载流子密度（n_s）为 $1.2\times10^{13}cm^{-2}$，迁移率为 $1760cm^2/V\cdot s$。

器件制造始于源漏欧姆图形光刻。通过欧姆光刻窗口，采用感应耦合等离子体（ICP）对样品进行预刻蚀，以减薄势垒厚度。随后，采用 Cl 基等离子体对样品进行 5 分钟的干法处理。如图 2 所示，Cl 基等离子体预处理可以去除表面氧化物并引入氮空位，导致表面施主态缺陷的增加。厚度为 20/20/140/55/45nm 的 Ti/Au/Al/Ni/Au 金属叠层通过电子束蒸发沉积并在 N_2 环境中 810℃温度下退火 30 秒。传统的 Ti/Al/Ni/Au(20/160/55/45nm) 金属方案样品作为对照组也进行了制备，但未进行浅槽刻蚀和干法等离子体处理。之后通过氮离子注入实现了器件的电隔离，并通过等离子体增强化学气相沉积（PECVD）沉积了 120nm 的 SiN 钝化层以抑制电流崩塌效应。然后使用电子束光刻（EBL）制作了栅长为 0.15μm，栅帽长度 0.6μm 的场板 T 型栅，并利用 F 基等离子刻蚀去除栅脚下方的 SiN。后通过 Ni/Au 栅极金属蒸发来完成肖特基接触。最后，蒸发 Ti/Au 金属用于器件互连。

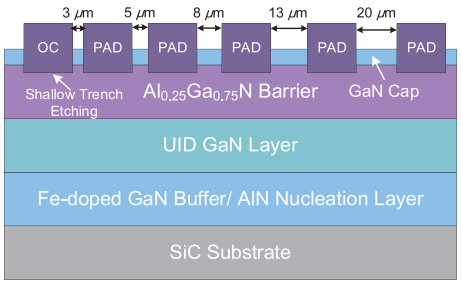


图 1 采用浅槽刻蚀工艺的 AlGaN/GaN 异质结材料结构示意图

Fig.1 Schematic diagram of the AlGaN/GaN heterostructure with the STE process

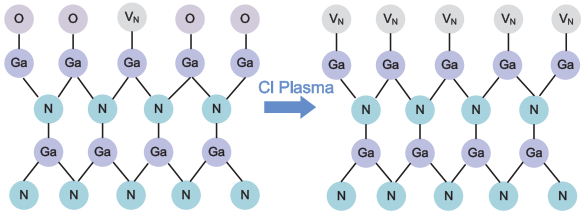


图 2 Cl 基等离子体处理工艺图

Fig.2 Schematic of the Cl-based plasma pretreatment process

3 器件特性与分析

3.1 欧姆接触特性

欧姆接触性能由传输线模型（TLM）进行评估。如图 3 所示，浅槽刻蚀样品的接触电阻

R_C 和比接触电阻率（ρ_C）分别为 0.28Ω · mm 和 2.1×10^{-6}Ω · cm^2，而常规方案的分别为 0.45Ω · mm 和 5.1×10^{-6}Ω · cm^2。需要注意的是，Ti/Al/Ni/Au 金属叠层在 810℃退火温度下为肖特基接触特性，因此后续不作为比较方案。因此，选择了 860℃退火温度的 Ti/Al/Ni/Au 样品与浅槽样品进行对比。如图 4(a) 和 (b) 所示，对退火后的 Ti/Al/Ni/Au 常规样品和 Ti/Au/Al/Ni/Au 浅槽刻蚀样品进行表面 AFM 对比，以比较两种方案退火后表面形貌。两种方案样品的粗糙度均方根值分别为 6.3nm 和 30.5nm，这意味着浅槽工艺显著改善了欧姆表面形貌。这可归因于浅槽欧姆工艺相对较低的退火温度（退火温度低于常规欧姆接触样品 50℃）。可以看到，常规样品有明显凸起结构，这是由于 Ni−Al 在高温退火过程中形成了金属球状合金簇，恶化了表面粗糙度，并且在实际 RF 工作中会导致传输电流不均匀以及明显的微波信号衰减，使得通讯质量劣化[7]。

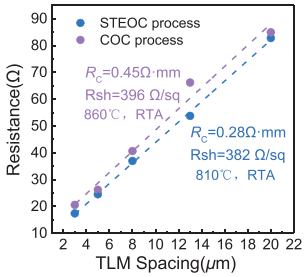


图 3 浅槽方案与常规方案 TLM 对比

Fig.3 The comparison of the TLM results between the STEOC−sample and COC−sample

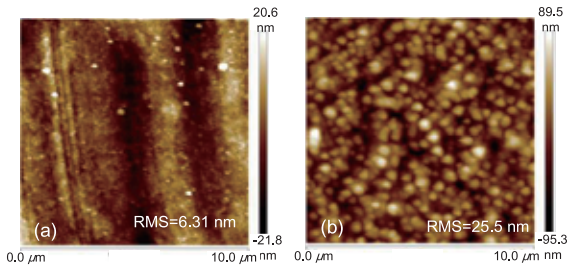


图 4 浅槽方案与常规方案 AFM 对比

Fig.4 The comparison of the AFM surface roughness between the STEOC−sample and COC−sample

表 1 针对含金欧姆接触，将本工作与国际主流报道进行对比。可以明显看到，采用浅槽刻蚀欧姆工艺的样品同时显示了高度光滑的表面形态和低的接触电阻。

表 1 关于含金欧姆接触特性国际报道对比

Tab.1 Comparison of Au−contained ohmic contact performance

参考文献	金属方案	RTA（℃）	RMS (nm)	R_C（Ω · mm）	ρ_C（Ω · cm^2）
[12]	Ti/Au/Al/Ni/Au	830	72	0.5	–
[12]	Ti/Al/Ni/Au	830	68	0.67	–
[7]	Ti/Al/Ni/Au	900	165.6	–	–
[14]	Ti/Al/Ni/Au	830	16.7	–	5.6×10^{-5}
[14]	Ti/Al/Ti/Ni/Au	870	27.6	0.28	–
本工作	STEOC	810	6.3	0.28	2.1×10^{-6}

3.2 三端器件特性

制备的栅极宽度 2×50μm 的器件样品，与常规样品的脉冲输出 I_D−V_{DS} 特性的比较如图 5(a) 和 (b) 所示。在 (0,0) 态静态偏置下，浅槽刻蚀欧姆接触可在 V_{GS}=+2V 和 V_{DS}=10V 时达到 1.7A/mm 的超高饱和漏极电流和低导通电阻，而传统样品的则为 1.58A/mm。如图所示，脉冲 I−V 测量是在 (V_{DS0}，V_{GS0}) = (10V，−8V) 下进行的，结果可以看到，浅槽刻蚀欧姆接触的饱和电流崩塌比为 3.3%，而常规样品饱和电流崩塌比为 3.5%，两种欧姆接触方案电流崩塌都较小。

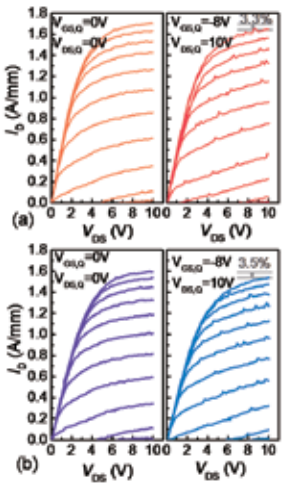


图 5（a）浅槽欧姆接触样品与（b）常规样品脉冲动态 IV 特性测试结果

Fig.5 The Pulse I−V characteristics of (a)STEOC HEMT and (b)COC HEMT

对两种类型的HEMT进行了频率范围为100MHz到40GHz的小信号测量对比。图6(a)和(b)显示了在V_{DS}=10V和V_{GS}=−6V的偏置下，浅槽刻蚀和常规欧姆器件的电流增益$|H_{21}|^2$和单边功率增益（UPG）随着频率的变化。可以看到，采用浅槽欧姆工艺的AlGaN/GaN HEMT的电流增益截止频率f_T和最大振荡频率f_{max}分别为60和150GHz，而常规欧姆接触工艺HEMT的f_T和f_{max}分别为53GHz和125GHz。

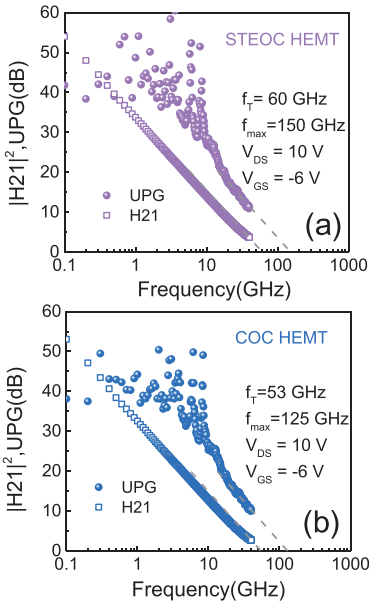


图6（a）浅槽欧姆接触样品与（b）常规样品的小信号频率特性对比

Fig.6 The small-signal frequency characteristics of (a)STEOC HEMT and (b)COC HEMT

3.3 大信号功率特性

功率测量在3.6GHz的连续波（CW）模式下进行，源阻抗和负载阻抗的调谐以实现最大PAE。如图7(a)所示，在V_{DS}=12V时，STEOC HEMT的峰值PAE为67%，附加输出功率密度(P_{out})达到1.6W/mm，而常规器件的峰值PAE和附加输出功率密度分别为62%和1.0W/mm。采用浅槽刻蚀欧姆接触对于器件输出功率及效率均有显著提升，这些结果表明浅槽刻蚀欧姆工艺在高性能射频应用方面具有巨大潜力。

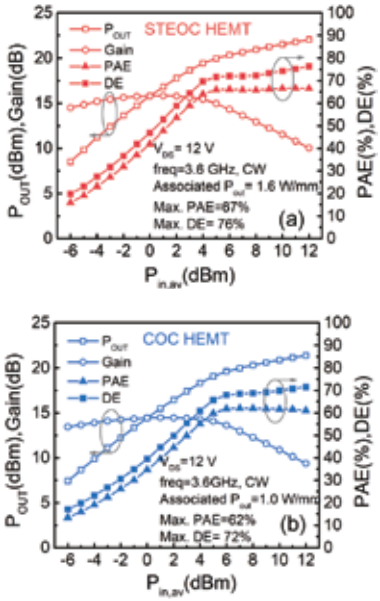


图7（a）浅槽欧姆接触样品与（b）常规样品的大信号功率特性对比

Fig.7 The large-signal power characteristics of (a)STEOC HEMT and (b)COC HEMT

4 结束语

综上所述，通过采用浅槽欧姆工艺的Ti/Au/Al/Ni/Au金属方案实现了低接触电阻(R_C=0.28Ω•mm)和高度光滑的表面形貌(RMS=6.3nm)。浅槽欧姆工艺的器件展现了优异的频率特性，电流截止频率f_T与最高振荡频率f_{max}分别为60/150GHz，在3.6GHz大信号功率特性测试表现出67%的高PAE。这些结果表明，浅槽刻蚀欧姆工艺对提升功率放大器（PA）应用性能具有重要意义。

参考文献（References）

[1] MISHRA U K, et al. GaN-based RF power devices and amplifiers [J]. Proc. IEEE, 2008, 96(2): 287-305.

[2] MISHRA U K, et al. AlGaN/GaN HEMTs—An overview of device operation and applications [J]. Proc. IEEE, 2002, 90(6): 1022-1031.

[3] HAO Y, et al. High-Performance Microwave Gate-Recessed AlGaN/AlN/GaN MOS-HEMT With 73% Power-Added Efficiency [J]. IEEE Electron Device

Letters, 2011, 32(5): 626-628.

[4] LU Y, et al. High RF Performance AlGaN/GaN HEMT Fabricated by Recess-Arrayed Ohmic Contact Technology [J]. IEEE Electron Device Letters, 2018, 39(6): 811-814.

[5] BRIGHT A N, et al. Correlation of contact resistance with microstructure for Au/Ni/Al/Ti/AlGaN/GaN ohmic contacts using transmission electron microscopy [J]. J. Appl. Phys, 2001, 89(6): 3143-3150.

[6] ROCCAFORTE F, et al. Nanoscale carrier transport in Ti/Al/Ni/Au Ohmic contacts on AlGaN epilayers grown on Si(111) [J]. Appl. Phys. lett, 2006, 89(2): 022103.

[7] GONG R M, et al. Analysis of surface roughness in Ti/Al/Ni/Au Ohmic contact to AlGaN/GaN high electron mobility transistors [J]. Applied Physics Letters, 2010, 97(6): 062115.

[8] DORA Y, et al. Effect of ohmic contacts on buffer leakage of GaN transistors [J]. IEEE Electron Device Lett., 2006, 27(7): 529-531.

[9] GUO J, et al. MBE-regrown ohmics in InAlN HEMTs with a regrowth interface resistance of 0.05 Ω.mm [J]. IEEE Electron Device Lett, 2012, 33(4):525-527.

[10] RECHT F, et al. Nonalloyed ohmic contacts in AlGaN/GaN HEMTs by ion implantation with reduced activation annealing temperature [J]. IEEE Electron Device Letters, 2006, 27(4): 205-207.

[11] CHARAN V S, et al. Scandium-Based Ohmic Contacts to InAlN/GaN Heterostructures on Silicon [J]. IEEE Electron Device Letters, 2021, 42(4): 497-500.

[12] YADAV Y K, et al. Ti/Au/Al/Ni/Au Low Contact Resistance and Sharp Edge Acuity for Highly Scalable AlGaN/GaN HEMTs [J]. IEEE Electron Device Letters, 2019, 40(1): 67-70.

[13] SHRIKI A, et al. Formatio mechanism of gold-based and gold-free ohmic contacts to AlGaN/GaN heterostructure field effect transistors [J]. J. Appl. Phys., 2017, 121(6): 065301-1-065301-5.

[14] CHIU Y S, et al. Ti/Al/Ti/Ni/Au ohmic contacts on AlGaN/GaN high electron mobility transistors with improved surface morphology and low contact resistance [J]. J. Vac. Sci. Technol. B, Microelectron. Nanometer Struct. Process., Meas., Phenom., 2014, 32(1): 011216.

作者简介：

芦浩(1994—),男,山西垣曲人,博士研究生,研究方向为射频微波功率器件。

高压 p-GaN HEMTs 总剂量效应致动态阈值不稳定性研究

王 钊，周 锌，陈 辰，吴中华，舒 磊，乔 明，张 波

（电子科技大学，四川省 成都市 610054）

摘　要：首次对高压 p-GaN HEMTs 电离辐射总剂量（TID）效应致动态阈值电压（V_{TH}）不稳定性进行了研究。辐射后发现 V_{TH} 与动态栅极应力呈现非单调关系，辐射对 p-GaN/AlGaN/GaN 异质结造成了损伤。在 p-GaN/AlGaN 界面，辐射产生的新的界面陷阱捕获电子，导致 V_{TH} 正向漂移。在金属 /p-GaN 肖特基结，辐射损伤增大了正向栅极电流以及空穴注入，引入了可能与氮空位缺陷相关的施主陷阱。增强的空穴注入在引起更强的 OP（opticalpumping）效应的同时，更多的空穴被俘获在 AlGaN 势垒中，导致了 V_{TH} 负向漂移。TID 辐射改变了电子捕获与空穴注入之间的竞争关系，导致了 V_{TH} 随动态栅极应力非单调变化。被观测到的增大的栅极电流和漏极泄漏电流以及变化的栅极电容能够验证其机理。

关键词：高压 p-GaN HEMTs；TID；界面损伤；肖特基结损伤

中图分类号：TN386.1　　**文献标识码**：A

Total-Ionizing-Dose Radiation Induced Dynamic V_{TH} Instability in p-GaN Gate HEMTs

Wang Zhao, Zhou Xin, Chen Chen, Wu Zhonghua, Shu Lei, Qiao Ming, Zhang Bo

(University of Electronic Science and Technology of China, Sichuan, 610054, China)

Abstract: Total-ionizing-dose (TID) radiation induced dynamic threshold voltage (V_{TH}) instability in p-GaN gate HEMTs is studied for the first time. A nonmonotonic dependence of V_{TH} on dynamic gate stress is observed. Irradiation causes damages at metal/p-GaN/AlGaN stack and the mechanism underlying irradiation-induced V_{TH} shift is revealed. At the p-GaN/AlGaN interface, the new interface traps buildup by irradiation capture electrons, causing more positive V_{TH} shift. At the metal/p-GaN Schottky junction, irradiation damage related to possible nitrogen vacancies would enhance hole-injection with increasing gate current. Hole-injection enhancement gives rise to stronger optical pumping and more holes trapped in the AlGaN barrier, causing more negative V_{TH} shift. TID radiation changes competition relationship between electron trapping and hole-injection, which is responsible for nonmonotonic V_{TH} shift. Increase of gate current and drain leakage, and change of gate capacitance are present to verify the mechanism.

Key words: p-GaN gate HEMTs; TID; interface damage; Schottky junction damage

0 引言

空间电力电子系统正朝着体积小、重量轻的方向发展，对功率密度的要求也越来越高[1,2]。氮化镓基高电子迁移率晶体管（HEMTs）以其较高的开关频率、功率密度和热导率，在空间环境中具有广阔的应用前景。在空间辐射环境下，电子设备将面临严重的辐射效应，其中最显著的辐射效应之一就是电离辐射总剂量（TID）效应[3-5]。然而，目前 TID 效应对 GaN HEMTs 的影响尚未得到全面研究，一些损伤机制尚不清楚[6,7]。

p-GaN HEMTs 以其优良的性能、低廉的成本和相对简单的工艺在商业市场上占据主导地位[7]。由于栅极 p-GaN 层存在镁杂质和氮空位等缺陷，其对电子 / 空穴的捕获及释放行为将影响载流子动态输

基金项目：国家自然科学基金 62004034
通信作者：周锌；E-mail：zhouxin@uestc.edu.cn

运，导致动态阈值电压（V_{TH}）漂移[8,9]。V_{TH} 正向漂移会引起额外的导通损耗，V_{TH} 负向漂移则会导致器件误开启。因此，动态阈值不稳定性是一个常见的问题，对p-GaN HEMTs具有重要意义。然而，据我们所知，在 p-GaN HEMTs 中，TID 效应对动态 V_{TH} 的影响目前尚未报道。

本文报道了 p-GaN HEMTs 中 TID 辐射诱导的动态 V_{TH} 不稳定性。辐照后 V_{TH} 与动态栅应力的非单调关系被观测到。辐射引起的动态阈值漂移机理被揭示，辐照损伤位于 metal/p-Ga 肖特基结和 p-GaN/AlGaN 界面。通过栅极电流（I_G）、栅极电容（C_G）和关态漏电流（I_{off}）试验验证了该机理。

1 器件结构和实验条件

图 1 为 p-GaN HEMTs 的 TID 辐射实验装置示意图。本文使用的是 650V/7.5A p-GaN HEMTs 商用器件[10]。TID 辐射实验采用了剂量率为 70rad/s 的 ^{60}Co γ 辐射源。辐射过程中，器件的栅极偏置 6V，其他电极接地。V_{TH}、I_G 和 C_G 在 500krad 和 2.5Mrad 时进行监测，测试仪器为 Keithley 4200 半导体分析测试系统。V_{TH} 从不同动态栅极应力（V_{GSQ}）下的动态转移特性中提取，其定义为漏极电流为 1mA 时的栅极电压。动态传输特性所采用的双脉冲测试的周期为 100μs ~ 10ms，脉冲宽度固定为 10μs。

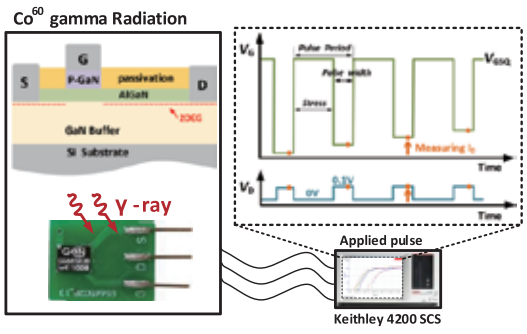


图 1 p-GaN HEMTs 的 TID 辐射实验装置示意图

Fig.1 Schematic of TID radiation experiment setup for p-GaN HEMTs

2 辐射实验结果

本文对偏置在栅极应力下的 7 只器件进行了辐照（施加 6V 栅极电压，漏极和源极接地）。辐照后，它们在相同测试条件下的阈值漂移特性基本相同，动态阈值随 V_{GSQ} 的增加均为先减小后增大。图 2(a) 显示了未辐射状态下 V_{TH} 与 V_{GSQ} 的关系。结果表明，V_{TH} 随 V_{GSQ} 的增加而单调增加。当 V_{GSQ} <5V 时，随着周期的增加，V_{TH} 正向漂移加剧。然而在 V_{GSQ}>5V 时，随着周期的增加 V_{TH} 正向漂移受到了抑制。图 2(b) 显示了不同 *TID* 下 V_{TH} 与 V_{GSQ} 的关系。与未辐射状态下 V_{TH} 与 V_{GSQ} 的单调关系明显不同，辐照后 V_{TH} 对 V_{GSQ} 呈现出非单调变化。辐照后，当不施加栅极电应力（V_{GSQ}=0V）时，V_{TH} 轻微地负向漂移。当 V_{GSQ}=3V 时，V_{TH} 显著增加。结果表明，随着 V_{GSQ} 从 3V 开始增加，V_{TH} 先减小后增大。在 *TID*=2.5Mrad 下，V_{GSQ}=3V 时 V_{TH} 为 1.82V，相比于 V_{GSQ}=0V 时的 V_{TH} 增加了 0.54V。随着 V_{GSQ} 的进一步增大，V_{TH} 在 V_{GSQ}=5V 时减小到最小值 1.55V，在 V_{GSQ}=7V 时增大到 1.92V。辐射结束器件静置退火 48 小时后，动态阈值曲线整体向上移动，1 个月后曲线又回落，如图 2(c) 所示。

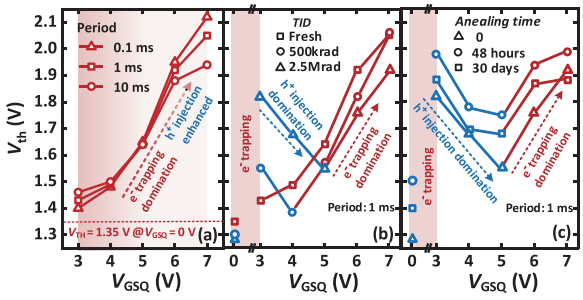


图 2 在 (a) 未辐射状态的不同周期下 (b) 不同 *TID* 下 (c) 不同退火时间下 V_{TH} 与 V_{GSQ} 的关系

Fig.2 V_{TH} as a function of V_{GSQ} (a) with different period for fresh state (b) with different *TID* and (c) with different annealing time

3 辐射诱导动态阈值不稳定性机理

图 3 显示了 p-GaN 栅极区域的辐射损伤能带示意图。当施加正向的栅极电压时，金属 /p-GaN 肖特基结反偏，而 p-GaN/AlGaN/GaN 结正向导通。来自 GaN 沟道的电子将穿过 AlGaN 势垒进入 p-GaN 层，其中部分电子在 p-GaN/AlGaN 界面被俘获[过

程 (i)]，导致 V_{TH} 正漂。同时，空穴翻越肖特基势垒从栅极金属注入 p-GaN 层。当获得足够的能量时，p-GaN 中的空穴将再次翻越 AlGaN 势垒注入 GaN 沟道，在这一过程中部分空穴会被 AlGaN 势垒层中的陷阱俘获 [过程 (ii)][11]。注入到 GaN 沟道的空穴与 2DEG 复合，产生电致发光 (EL)，其中包含的 3.4eV 紫外光子通过 OP(optical pumping) 效应可以有效地激发 p-GaN/AlGaN 界面处俘获的电子 [过程 (iii)][12]。由空穴注入引起的空穴捕获 [过程 (ii)] 和 OP 效应 [过程 (iii)] 可导致 V_{TH} 负漂。因此，V_{TH} 的漂移取决于空穴注入和电子俘获之间的竞争。

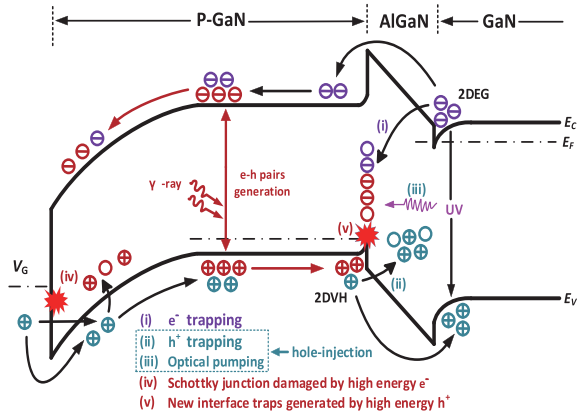


图 3 p-GaN 栅极区域的辐射损伤能带示意图

Fig.3 Schematic band diagram of p-GaN gate region with radiation damage

对于动态栅应力下的未辐射状态，V_{TH} 随 V_{GSQ} 的增加而不断增加，表明电子俘获始终占据主导地位。在 V_{GSQ}<5V 时，小的栅极泄漏电流限制了电子被空间局部陷阱捕获的速度，因此随着周期的增加，越来越多的电子被俘获，导致 V_{TH} 进一步正向漂移。在 V_{GSQ}>5V 时，随着栅极电流的增加，空穴注入逐渐增强，因此随着周期的增加 V_{TH} 的正向漂移受到了抑制。

在辐照过程中，p-GaN 层中会激发大量的电子-空穴对。在正向栅极偏压下，由于肖特基结空间电荷区内存在的高电场，电子会加速向金属 /p-GaN 界面移动并造成损伤 [过程 (iv)]。通过监测 I_G 可以证实以上推测，如图 4 所示。结果表明，随着 *TID* 的增加，正向 I_G 显著增加，反向 I_G 几乎不变。在 *TID*=2.5Mrad 时，观察到正向 I_G 增加了两个数量级以上。p-GaN 栅极结构可以建模为背对背串联的肖特基二极管 (D_{Sch}) 和 p-i-n 二极管 (D_{pin})，它们分别主要阻断正向栅极电流和反向栅极电流。因此，我们认为肖特基结在辐射后受损，而 p-i-n 结仍保持其功能。辐射激发的高能电子通过轰击金属 /p-GaN 界面在其附近形成陷阱，该陷阱可能与氮空位缺陷有关。陷阱辅助遂穿机制因此被引入，正向栅电流增大，空穴注入增强。空穴注入的增强一方面会导致更多的空穴被俘获在 AlGaN 势垒层中 [过程 (ii)]，另一方面也导致了更强的 OP 效应 [过程 (iii)]，这两种结果分别由图 5 和图 6 验证。

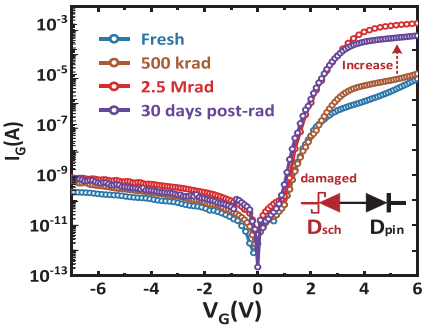


图 4 I_G 随 V_G 的变化关系

Fig.4 I_G as a function of the V_G(V_S=V_D=0V)

图 5 显示了静态阈值电压漂移 ($\Delta V_{TH,sta}$) 与静态栅极电压 ($V_{G,str}$) 的关系。结果表明，辐照后阈值曲线整体向下移动，即静态阈值电压负向漂移加剧。由于阈值负漂通常归结于 AlGaN 势垒中空穴陷阱电荷，这说明辐射后更多的空穴被俘获在 AlGaN 势垒层中。

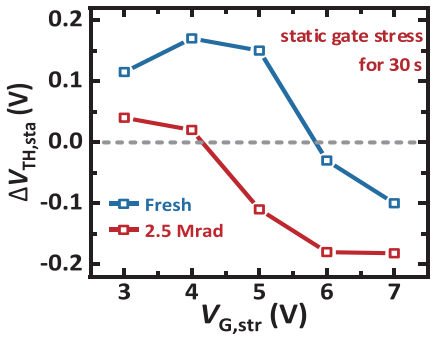


图 5 $\Delta V_{TH,sta}$ 随 $V_{G,str}$ 的变化关系

Fig.5 $\Delta V_{TH,sta}$ as a function of $V_{G,str}$ with a delay of 1s after applying $V_{G,str}$ for 30s

为了评估 OP 效应，一种有效的方法是在施加栅极应力后进行关态电流 I_{off} 测试[13]。图 6 显示了辐射前后不同 $V_{G,str}$ 下 I_{off} 随时间的变化关系。未辐射状态下，$V_{G,str}$=0V 时 I_{off} 最大值（对应于时间 0.1s 处）仅为 2.7×10^{-8}A。当 $V_{G,str}$ 增加到 7V 时，空穴注入诱导的 EL 通过 OP 效应可以有效降低栅下缓冲层的负空间电荷。GaN 沟道的势垒因此降低，导致 I_{off} 升高。辐射后，$V_{GSQ}\geqslant$ 7V 时 I_{off} 相比于辐射前增加显著，证实了辐射诱导的 OP 增强 [过程 (iii)]。增强的 OP 效应使得 GaN 沟道中的更多的电子陷阱电荷被释放，导致势垒进一步降低，I_{off} 增加。此外，如图 7 所示的微光显微测试也验证了 OP 作用的增强。与未辐射器件相比，辐射后的器件发光随着 *TID* 增加而更加强烈。

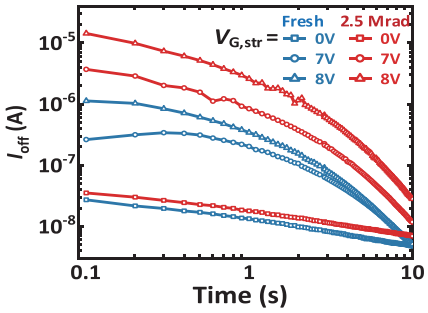


图 6 不同 $V_{G,str}$ 下 I_{off} 随时间的变化关系

Fig.6 I_{off} (at V_G=V_S=0V, V_D=50V) as a function of time with different $V_{G,str}$

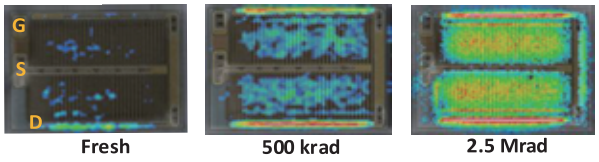


图 7 对于不同 *TID* 器件的微光显微测试

Fig.7 EMMI measurement of uncapped devices at V_G=6V with different *TID*

与电子运动相反，辐射激发的空穴在价带中加速向 AlGaN 势垒方向移动，并将能量释放到晶格中，从而在 p-GaN/AlGaN 界面产生新的缺陷 [过程 (v)]。因此，更多的电子将被增加的界面陷阱捕获，导致过程 (i) 在给定的正栅偏压下等效增强。为了验证辐射后电子捕获增强，在 1MHz 频率下测试了栅极电容 (C_G) 如图 7 所示。根据栅电容串联模型，总电容 $1/C_G=1/C_{Sch}+1/C_{pin}$，其中 C_{Sch} 为肖特基结的电容，C_{pin} 为 AlGaN 势垒的电容。当 V_G 超过阈值电压时，C_{pin} 上的压降会保持在固定值，其余电压将被加在 C_{Sch} 上[14]。辐射后，当 V_G<5V，C_G 下降，这是由于 C_{Sch} 下降所致。如前所述，在金属 /p-GaN 界面附近引入与氮空位相关的施主陷阱增多，p-GaN 区域的耗尽区扩展，导致 C_{Sch} 和 C_G 减少。当 V_G>5V，C_G 显著增加。我们认为这是由于辐射在 p-GaN/AlGaN 界面引入的电子陷阱 [过程 (v)] 导致了 C_{pin} 增加。随着 V_G 的增大，Femi 能级逐渐上升到能够响应 1MHz 频率的辐射诱导的浅能级界面陷阱位置处，导致了 C_{pin} 和 C_G 的增加。

因此，辐射损伤改变了电子捕获和空穴注入之间的竞争关系，这是辐射后 V_{TH} 与 V_{GSQ} 呈现非单调变化关系的原因。在图 2(b) 中，当 V_{GSQ}=3V 时，V_{TH} 显著增加是因为辐射诱导的陷阱捕获了更多的电子 [过程 (i) 和 (V)]。随着 V_{GSQ} 的增加，伴随着正向 I_G[过程 (iv)] 的增加和过程 (ii)、(iii) 的增强，空穴注入前期占据主导地位，导致了 V_{TH} 负向漂移。随着 V_{GSQ} 的进一步增大，正向 I_G 的增加趋于平缓，空穴注入效应饱和，电子捕获重新占据主导地位 [过程 (i)]。因此，V_{TH} 在后期正向漂移。当 *TID* 从 500krad 增加到 2.5Mrad 时，正向 I_G 的增加会引起更强烈的空穴注入，从而导致 V_{TH} 下降阶段的延长。退火一个月后，V_{TH} 与未辐射状态相比仍有较大的正漂，如图 2(c) 所示。这意味着辐射引入了深能级陷阱，导致 I_G 和 C_G 未完全恢复，如图 4 和图 6 所示。

值得注意的是，另外设置了两组对照实验。一组在辐射的时候电极全部浮空，另一组没有辐射，仅偏置在 6V 栅极电应力下 10 小时。两者均表现出单调的 V_{TH} 漂移，I_G 和 C_G 变化不大，说明辐射损伤损伤是辐射和电应力共同作用的结果，而不仅仅是辐射本身。

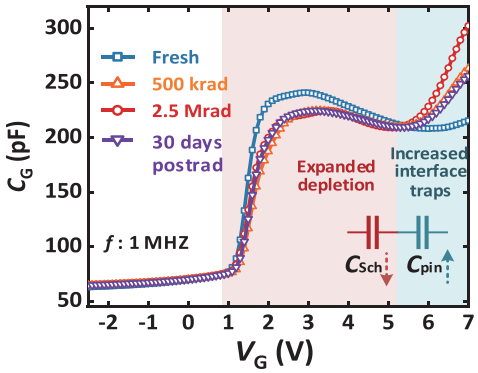


图 8 C_G 随 V_G 的变化关系

Fig.8 C_G as a function of V_G at frequency of 1MHz

4 结论

本文研究了 p-GaN HEMTs 中 TID 辐射诱导的动态 V_{TH} 不稳定性。揭示了辐射后 V_{TH} 与动栅应力非单调关系的机理。金属 /p-GaN 肖特基结附近引入了可能与氮空位相关的施主类陷阱，导致正向 I_G 增加以及空穴注入增强。辐射诱导 p-GaN/AlGaN 界面产生电子陷阱，导致更多的电子被捕获以及 C_G 变化。辐射改变了电子捕获和发空穴注入之间的竞争关系，最终决定了 V_{TH} 漂移。

参考文献（References）

[1] MCCLORY J W, PETROSKY J C, SATTLER J M, et al, An analysis of the effects of low-energy electron irradiation of AlGaN/GaN HFETs[J].IEEE Transactions on Nuclear Science, 2007. 54(6): 1946-1952.

[2] ZHOU X, YUAN Z, SHU L, et al, Total-ionizing-dose irradiation-induced dielectric field enhancement for high-voltage SOI LDMOS[J]. IEEE Transactions on Electron Devices, 2019. 40(4): 593-596.

[3] 王丹辉，赵元富，岳素格，等 . 高压 LDMOS 总剂量辐射效应研究 [J]. 微电子学与计算机，2015,(10):82-86.

[4] SHARMA C, MODOLO N, WU T, et al, Understanding γ -ray induced instability in AlGaN/GaN HEMTs using a physics-based compact model[J]. IEEE Transactions on Electron Devices, 2020. 67(3): 1126-1131.

[5] ZHENG X, FENG S, PENG C, et al, Evidence of GaN HEMT Schottky gate degradation after gamma irradiation[J].IEEE Transactions on Electron Devices, 2019. 66(9): 3784-3788.

[6] PEARTON J, REN F, PATRICK E, et al, Ionizing radiation damage effects on GaN devices[J]. ECS Journal of Solid State Science and Technology, 2015. 5(2): 35-60.

[7] SCHWANK J R, SHANEYFELT M R, Fleetwood D M, et al. Radiation Effects in MOS Oxides, IEEE Transactions on Nuclear Science, vol.55,no. 4, pp. 1833-1853,Sep. 2008, doi:10.1109/TNS.2008.2001040.

[8] UEMOTO Y, HIKITA M, Ueno H, et al, Gate injection transistor (GIT)—A normally-off AlGaN/GaN power transistor using conductivity modulation[J].Transactions on Electron Devices, 2007. 54(12): 3393-3399.

[9] WANG H, WEI J, XIE R, et al, Maximizing the performance of 650-V p-GaN gate HEMTs: Dynamic RON characterization and circuit design considerations, IEEE Transactions on Power Electronics[J], 2016. 32(7): 5539-5549.

[10] TALLARICO A N, STOFFELS S, POSTHUMA N, et al, Threshold voltage instability in GaN HEMTs with p-type gate: Mg doping compensation[J]. IEEE Electron Device letters, 2019. 40(4): 518-521.

[11] GaN Systems, GS66502B Datasheet. (2017). [Online]. Available: http://www.gansystems.com.

[12] SAYADI L, IANNACCONE G, SICRE S, et al, Threshold voltage instability in p-GaN gate AlGaN/GaN HFETs[J].Transactions on Electron Devices, 2018. 65(6): 2454-2460.

[13] TANG X, LI B, et al, Mechanism of threshold voltage shift in p-GaN Gate AlGaN/GaN transistors[J], IEEE Electron Device letters, 2018. 39(8): 1145-1148.

[14] ABDULSALAM A, KARUMURI N, DUTTA G, Modeling and analysis of normally-OFF p-GaN gate AlGaN/GaN HEMT as an ON-chip capacitor[J], IEEE Transactions on Electron Devices, 2020. 67(9): 3536-3540.

作者简介：

王钊（1996—），男，河南安阳人，博士研究生，主要从事高压 GaN 器件辐射机理与加固技术研究。

β－Ga_2O_3 肖特基势垒二极管低温退火界面特性优化研究

洪悦华，张翔宇，张 方，朱 甜，张 豪，郑雪峰，马晓华，郝 跃

（西安电子科技大学，陕西省 西安市 710071）

摘 要：利用低温退火技术，开展了垂直结构 β－Ga_2O_3 肖特基势垒二极管的性能提升研究。研究发现，经过低温退火，阳极金属镍扩散到 Ga_2O_3 中，在阳极和半导体界面处形成 NiO。通过变频电导法和 X 射线光电子能谱，发现 Ni/Ga_2O_3 界面的陷阱态密度以及表面与氧结合的碳的陷阱态密度降低。由于 NiO 的形成以及金属－半导体界面的优化，器件的关态电流密度（I_{off}）从 1.21×10^{-6}A/cm^2 降低至 5.12×10^{-9}A/cm^2，同时实现了更高的击穿电压（V_{br}）。研究结果表明，通过界面工程对垂直结构 β－Ga_2O_3 SBDs 进行低温退火技术能够提高器件性能。

关键词：氧化镓；肖特基势垒二极管；退火；界面工程

中图分类号：TN31 **文献标识码**：A

The Optimized Interface Characteristics of β-Ga_2O_3 Schottky Barrier Diode with Low Temperature Annealing

Hong Yuehua, Zhang Xiangyu, Zhang Fang, Zhu Tian, Zhang Hao, Zheng Xuefeng, Ma Xiaohua, Hao Yue

(Xidian University, Xi'an, 710071, China)

Abstract: A low temperature controlled annealing technique was utilized to improve the performance of vertical β-Ga_2O_3 Schottky barrier diodes in this work. The nickel diffuses into Ga_2O_3 and NiO was formed at the interface between the Anode and semiconductor generating p-n junction after low temperature annealing. Simultaneously, the trap state density of interface Ni/Ga_2O_3 as well as the carbon bonded with Oxygen on the surface was reduced, which was proved by the capacitance and conductance measurements and X-ray photoelectron spectroscopic analysis. Combined the decreased off-state current density(I_{off}) by 3 orders of magnitude from 1.21×10^{-6}A/cm^2 to 5.12×10^{-9}A/cm^2 and larger breakdown voltage(V_{br}) from 220V to 270V owing to optimized interface, and the formation of NiO, a low temperature annealing technique make certain of effective improvement for vertical β-Ga_2O_3 SBDs via interface engineering.

Key words: Gallium Oxides; Schottky barrier diodes; annealing; interface engineering

0 引言

β－氧化镓（β－Ga_2O_3）是一种具有超宽带隙（4.5eV－4.9eV）和高临界电场强度（8MV）的新型半导体材料。因此，氧化镓器件在超高耐压、抗辐照领域具有良好的应用前景。另外，β－Ga_2O_3 可以很容易地通过 Si 或 Sn 进行 N 型掺杂，是一种有希望在大规模衬底上生产的低成本半导体材料。

目前，β－Ga_2O_3 电子电力器件主要研究 p－n 二极管、肖特基势垒二极管（SBD）以及金属氧化物半导体场效应晶体管（MOSFET）。对于这些电子器件来说，研究其半导体和金属接触之间的界面工程是很有必要的，例如突变结的形成、功函数、钉扎效应[1]和界面态[2]。曾有学者利用阳极退火技术对 SiC[3] 和 GaN[4] 的 SBD 深有研究，但氧化镓上研究较少。MadaniLabed 等人提出在退火时，Ni 扩散到 Ga_2O_3 中，并形成一层 $(Ni_xGa_{1-x})_2O_3$ 以补偿界面上的缺陷[5]。Kim 等人表

基金项目：国家预研基金（批准号：31513020109）、国家自然科学基金（批准号：61974115，11690042，61634005，61974111）资助的课题
通信作者：郑雪峰；
E-mail：xfzheng@mail.xidian.edu.cn

明 Ni 扩散将替代 Ga，从而降低漏电流[6]。Hao 等人通过 350℃退火优化 NiO/Ga_2O_3 之间的界面，减少界面缺陷，有助于减少 I–V 回滞[7]。然而，这种与 β－Ga_2O_3 相关的界面研究仍存在许多问题。

本文制作了一种垂直 β－Ga_2O_3 SBD，并提出了通过低温退火的方法改善Ga_2O_3和Ni之间界面质量，从而提高器件性能。实验表明，通过 200℃退火，I_{off} 从 1.21×10^{-6}A/cm^2 下降到 5.12×10^{-9}A/cm^2，而击穿电压 V_{br} 从 220V 提高到 270V。其主要原因是在 Ni/Ga_2O_3 界面处形成了 NiO 薄膜，本文进一步优化了金属－半导体界面特性。

1 实验部分

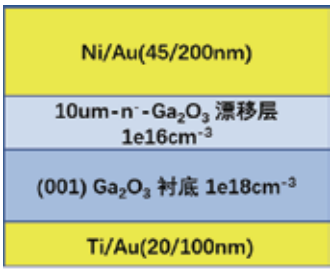


图 1 氧化镓肖特基二极管器件结构示意图

Fig.1 Schematics of vertical β－Ga_2O_3 Schottky barrier diode

图 1 展示了本文制备的 β－Ga_2O_3 SBD 的器件结构，由掺杂浓度为 1×10^{18}cm^{-3} 的 β－Ga_2O_3 (001) 衬底和厚度为 10μm 的轻度掺杂 (1×10^{16}cm^{-3}) 外延层组成。在样品背面通过磁控溅射制备阴极 Ti/Au(20/100nm)，然后在 N_2 环境中在 440℃下快速热退火。通过光刻制备出阳极区域，后通过电子束蒸发沉积 Ni/Au(45nm/200nm) 作为阳极金属。器件制备后利用快速退火炉在 N_2 气体中分别在 100℃和 200℃下进行低温退火 5min。

为了深入研究阳极和氧化镓之间的界面，在室温下以 AlKα 作为 X 射线辐射源进行 X 射线光电子能谱 (XPS) 测量。在测试 XPS 前，使用 KI 溶液和稀盐酸除去阳极金属 Au 和 Ni 后，真空干燥存放 24 小时。采用 Keithley4200 分析仪来测量正向电流－电压正向特性和变频电导特性，1505A 半导体分析仪测量反向电流－电压特性。

2 讨论部分

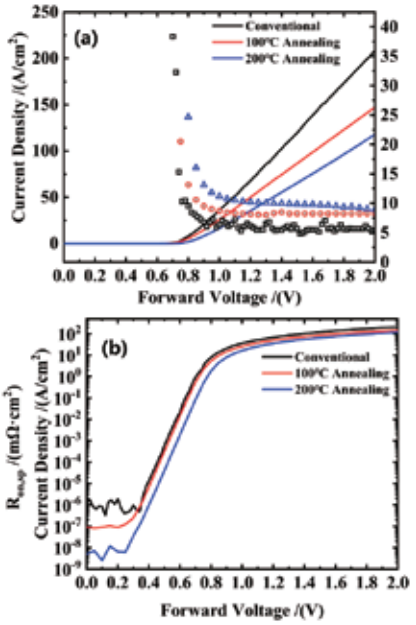


图 2 （a）肖特基二极管正向特性曲线（线性坐标）及导通电阻（b）正向特性曲线（半对数坐标）

Fig.2 (a) Linear forward I–V characteristics and extracted $R_{on,sp}$ as a function of forward bias and (b) Semi-logarithmic forward I–V characteristics of three types of devicesfor conventional，100℃ ，and 200℃ annealing

图 2（a）分别显示了常规 SBD、100℃和 200℃退火后的 SBD 的典型正向 I–V（线性坐标）和提取的 $R_{on,sp}$–V 特性。常规 SBD、100℃和 200℃退火后的 SBD 在正向 V=2V 处的电流密度 (I_F) 峰值分别为 208、147、118A/cm^2，导通电阻 ($R_{on,sp}$) 值分别为 4.81、7.98 和 8.96mΩ·cm^2。200℃退火的 SBD 的 $R_{on,sp}$ 与文献报道的 NiO/β－Ga_2O_3 p–n 二极管[8]的 $R_{on,sp}$ 相似，这将在下面进一步深入讨论。对于常规 SBD，100℃和 200℃退火后的 SBD，V_{on} 分别为 0.70、0.72 和 0.80V。结果表明，经过退火后，器件的开启电压发生正漂移。退火后的 SBD 具有较高的 V_{on} 值可能主要是因为 NiO 的耗尽效应和 pn 异质结的形成，这将在下面进一步讨论。

Ga_2O_3 SBD 的半对数正向 I–V 特性曲线如图 2（b）所示。常规的 SBD 的测得 1.21×10^{-6}A/cm^2，整流比为 1.24×10^8。而 100℃和 200℃退火后的 SBD 的 I_{off} 分别为 9.27×10^{-8}A/cm^2 和

5.12×10^{-9}A/cm^2，整流比分别为1.59×10^{9}和2.31×10^{10}。结果表明，退火后器件的I_{off}得到明显的降低，降低了3个数量级。基于热电子发射(TE)模型，对于常规SBD、100℃和200℃退火后的SBD，理想因子(η)分别提取为1.03、1.03、1.04。

$$I = I_S \cdot \{\exp[q\cdot(V-1\cdot R_S)/\eta kT]-1\} \quad (1)$$

公式（1）中R_s、T、k、η和I_s是串联电阻、绝对温度、玻尔兹曼分别为常数、理想因子和反向饱和电流。η值越接近1，表明半导体和金属之间存在肖特基界面质量越高。

$$\phi_B = (kT/q)\cdot\ln(AA^*T^2/I_S) \quad (2)$$

此外，由公式（2）得到肖特基势垒高度（ϕ_B），其中A和$A*$分别是阳极面积和理查德常数[9]。常规的SBD、100℃、200℃退火后SBD的肖特基势垒高度分别为1.07eV、1.08eV、1.14eV。表1总结了上述三种SBD的正向特性曲线的参数。显然，经过200℃退火后，器件的ϕ_B在从1.07eV上升到1.14eV，增幅达到6%。V_{on}增加主要源自于ϕ_B的增加。而ϕ_B的增加则是因为Ni和Ga_2O_3之间的界面在退火后发生了变化。同时，由于ϕ_B增加，I_{off}从1.21×10^{-6}A/cm^{2}降低到5.12×10^{-9}A/cm^2，整流比增加了3个数量级。

表1 肖特基二极管基本正向特性参数

Tab.1 The electrical performance parameters extracted from $I-V$ characteristic curves

Device	$R_{on,sp}$ /mΩ · cm^2	V_{on} /V	n	ϕ /eV
Conventional	4.81	0.70	1.03	1.07
100℃ Annealing	7.98	0.72	1.03	1.08
200℃ Annealing	8.96	0.80	1.04	1.14

为了进一步研究Ni和半导体之间的界面，在10kHz到1MHz的频率范围内进行了变频电导测试。图3展示了三种SBD的电导的变频电导特性。测试的偏压是选择在开启电压0.8V附近。陷阱态能量E_T可以从公式（3）的拟合过程中计算出来[10]：

$$\tau_T = (\sigma N_c v_T)^{-1}\exp(E_T/kT) \quad (3)$$

其中σ是陷阱态的捕获截面，N_c是导带中的态密度，v_T是载流子的平均热速度。此处$N_c=2\times10^{18}$cm^{-3}，$v_T=2\times10^{7}$cm/s[11]。通过变频电导法测试，陷阱时间常数随电压关系曲线如图4（a）所示。陷阱态的时间常数与偏置有关，其中E_t为Ga_2O_3的费米能级。随着偏置电压的增加，能级较浅的陷阱首先被检测到，陷阱的时间常数则越小。因此，随着偏置电压增加，陷阱的时间常数就会减小。值得注意的是，在相同的偏置下，经过退火的SBD器件其陷阱时间常数总小于常规的SBD，这表明退火后SBD的陷阱态能级比常规SBD的要浅。同时，Ni和Ga_2O_3之间的界面在退火后发生了变化，陷阱态则变得不同。

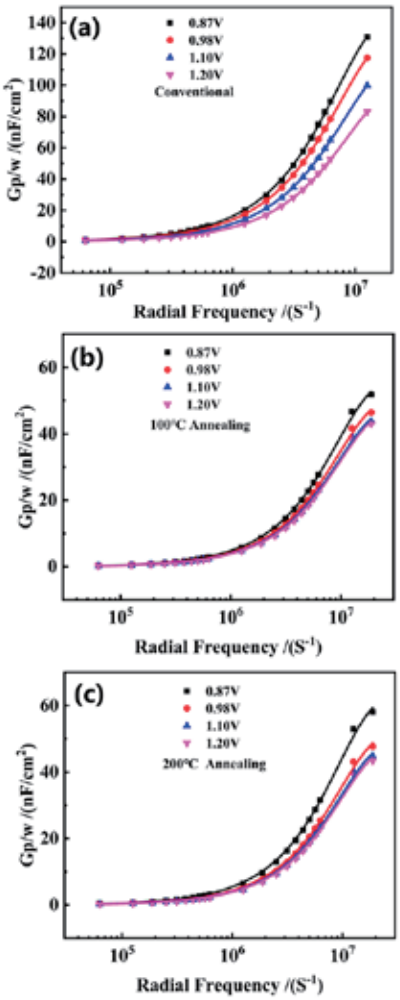


图3 变频电导特性曲线（a）常规SBD（b）100℃退火的SBD（c）200℃退火的SBD

Fig.3 Conductance as a function of radial frequency at selected bias near *Von* for samples (a) conventional (b) 100℃ annealing (c) 200℃ annealing, respectively

陷阱态密度随陷阱能级的变化关系如图 4（b）所示。常规的 SBD 中的陷阱态密度从 $2.24\times10^{12}cm^{-2}eV^{-1}$ 下降到 $1.53\times10^{12}cm^{-2}eV^{-1}$，能级从 0.080eV 下降到 0.075eV，而 100 ℃退火的 SBD 和 200 ℃退火的 SBD 陷阱态密度从 8.63×10^{11} $cm^{-2}eV^{-1}$ 下降至 $7.26\times10^{11}cm^{-2}eV^{-1}$ 和 $9.69\times10^{11}cm^{-2}eV^{-1}$ 降至 $7.36\times10^{11}cm^{-2}eV^{-1}$，能级从 0.074eV 降低到 0.072eV。当能级（E_T）降低时，陷阱态密度降低。值得注意的是，可以观察到常规 SBD 的能级总是略大于退火后的 SBD 的能级。特别是退火后的 SBD 的陷阱态密度始终低于常规的 SBD 器件，这意味着退火使得 Ni 和半导体之间的界面得到了优化。这与退火后 SiO_2/Ga_2O_3 界面的陷阱态密度降低现象类似。该现象与 I_{off} 的降低有关，这归因于陷阱状态的减少。这可能是由于退火过程中的 Ni 扩散通过填充 Ga 空位来减少缺陷 [5]。

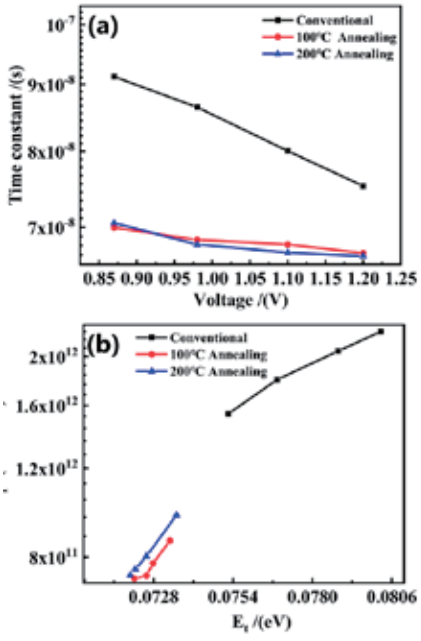


图 4 （a）陷阱时常数随电压关系曲线（b）陷阱态密度随陷阱能级的变化关系

Fig.4 (a) Trap state time constant as a function of forward bias (b) Trap state density as a function of their energy for three types of devices with different annealing conditions

为了进一步研究 Ni 和 Ga_2O_3 的界面情况，对常规 SBD 和 200℃退火的 SBD 进行了 XPS 测量。图 5（a）（b）和（c）（d）分别是 Ni−2p 和 Ga−3d 的光电子能谱。图 5（a）中 200℃退火后的 SBD 的 Ni−$2p_{3/2}$ 核心级光谱被分解为 4 个分量，集中在 860.59eV、855.65eV、853.31eV[12,13] 和 852.17eV[13]，这分别是卫星峰、Ni^{3+}、Ni^{2+} 和 Ni 金属。这里 Ni 金属的出现则是因为 Ni 没有被彻底去除。相比之下，图 5(b) 中的常规 SBD 中没有看到 Ni−$2p_{3/2}$。应该注意的是，Ni^{2+}、Ni^{3+} 和卫星峰是 NiO 中典型的 XPS 特征。这证实了 Ni 扩散到 Ga_2O_3 中，并且在界面处形成了 NiO。NiO 天然的是 P 型半导体。这意味着界面形成了 pn 结（NiO/Ga_2O_3）。这就解释了退火后 ϕ_B 的增加。常规 SBD 和 200℃退火的 SBD 的 O_{1s} 图谱分别如图 5（c）和（d）所示。两个光谱都显示了 530.8eV 处的强峰，是 Ga−O 峰。Ni^{3+} 和 Ni^{2+} 的峰值分别为 530.47eV 和 531.67eV。需要注意的是，531.91eV 的峰值则对应的是与氧键合的不定碳，仅出现在常规 SBD 中。这说明在退火后可以钝化表面上与氧结合的碳等陷阱态，并通过形成 NiO 来优化界面。这可以解释退火后 I_{off} 的降低。

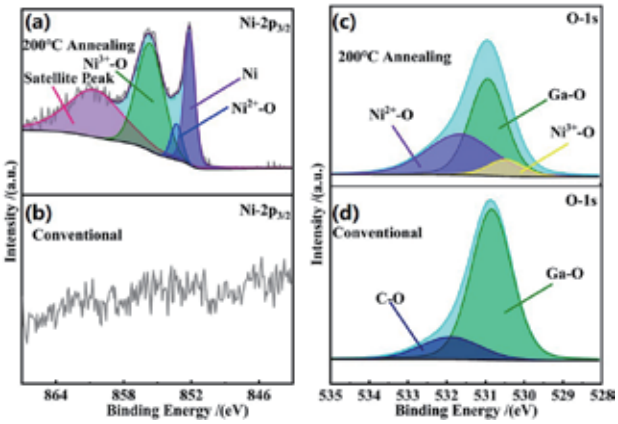


图 5 X 射线光电子能谱（a）200℃退火后的 SBD 的 Ni−$2p_{3/2}$ 核心级光谱（b）常规 SBD 光电子能谱（c）200℃退火的 SBD 的 O_{1s} 图谱（d）常规 SBD 的 O_{1s} 图谱

Fig.5 XPS spectra of Ni−$2p_{3/2}$ for (a) 200℃ annealing and (b) conventional and of O_{1s} for (c) 200℃ annealing and (d) conventional

图 6 是三种器件的反向 $I-V$ 特性。进行反向击穿测量时，阴极接地，阳极的偏置从 0 反向增加到 V_{br}。常规 SBD 的 V_{br} 达到 220V，而 100℃和 200℃退火后器件的 V_{br} 分别为 234V 和 270V。结果表明，

200℃退火后击穿电压提高了23%。这表明使用低温退火技术可以实现更大的击穿电压，这可能是由于退火后形成了NiO。NiO可以降低峰值电场强度，这是由于异质结构引起的耗尽效应。

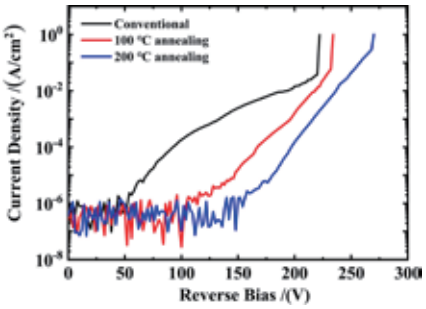


图6 肖特基二极管反向特性曲线

Fig.6 Reverse $I-V$ characteristics of three types of devices

为了确认NiO有助于降低阳极边缘的电场（E）峰值电场效应，对常规的SBD和退火的SBD进行了TCAD模拟。模拟的掺杂浓度和结构设置为与制造的器件相同。假设退火后的NiO厚度为2nm。施加的击穿电压为270V。常规的SBD和200℃退火的SBD的模拟电场分布分别如图7（a）和（b）所示。需要注意的是，最大电场出现在阳极边缘。可以观察到，退火后的SBD峰值电场较传统的明显降低。从阳极边缘向3μm深度的Ga_2O_3提取的电场分部如图7（c）所示。常规的SBD和退火后SBD的模拟峰值电场分别达到4.89MV/cm和3.83MV/cm。峰值电场强度的抑制则是在退火后形成P型NiO层。这使得电场扩展终止于NiO层，而不是会聚集到Ga_2O_3区域的阳极边缘。此仿真结果与实验获得的反向$I-V$特性非常吻合。

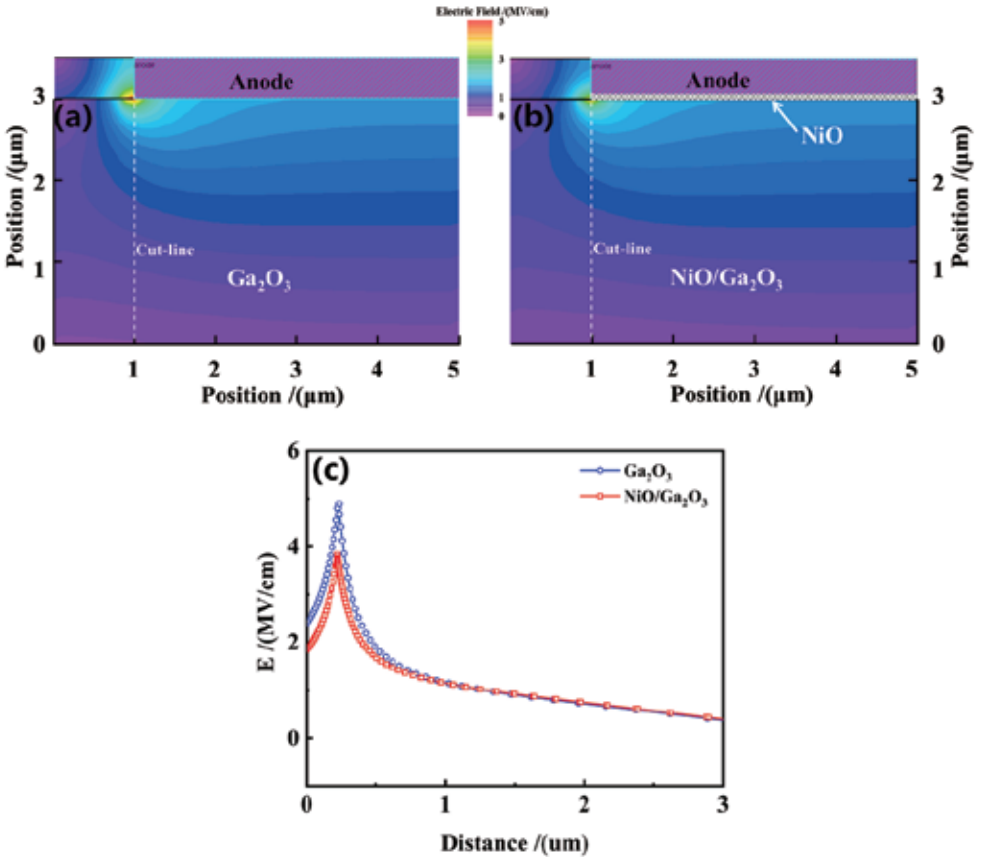


图7（a）常规SBD（b）退火后SBD在270V反向偏压下的二维电场分布（c）两种器件在阳极边缘区域的垂直电场分布

Fig.7 Two-dimensional electric filed distributions at a reverse bias of 270 V for device (a) conventional (b) post-annealing (c) vertical E distribution in the anode edge region for two types of devices

3 结论

综上所述，本文利用从100℃到200℃的低温退火技术能够优化垂直结构 β－Ga_2O_3 SBD 界面特性。经过200℃退火后，能够使器件的I_{off}从1.21×10^{-6}A/cm^2降低到5.12×10^{-9}A/cm^2。这是因为在退火过程中，阳极金属镍扩散到Ga_2O_3中并形成NiO所导致。本文采用了XPS能谱确认了NiO的形成。基于变频电导法测试，表明退火后金属－半导体界面的陷阱态密度降低，继而改善阳极Ga_2O_3的界面质量。此外，NiO可以有效地缓解电场，使得击穿电压从220V提高到270V，提高了23%。

参考文献（References）

[1] BHATTACHARYYA A，RANGA P，SALEH M，et al.Schottky Barrier Height Engineering in β－Ga2O3 Using SiO2 Interlayer Dielectric [J]. IEEE J. Electron Devices Soc. 2020 (8)：286－294.

[2] LINGAPARTHI R，THIEU Q T，KOSHI K,etal.Surface states on (001) oriented β－Ga_2O_3 epilayers，their origin，and their effect on the electrical properties of Schottky barrier diodes [J]. Appl. Phys. Lett. 2020 9(116)：092101－092105.

[3] NAKAMURA T，MIYANAGI T，KAMATA I，et al. A 4.15 kV 9.07－m Omega center dot cm^2 4H－SiC Schottky－barrier diode using Mo contact annealed at high temperature [J]. IEEE Electron Device Lett. 2005 26 (2)：99－101.

[4] ZHANG T，WANG Y，ZHANG Y，et al. Comprehensive Annealing Effects on AlGaN/GaN Schottky Barrier Diodes With Different Work－Function Metals [J]. IEEE Trans. Electron Devices，2021 68(6)：2661－2666.

[5] MADANI L，NOUREDINE S，MOHAMED，et al. Modeling and analyzing temperature－dependent parameters of Ni/β－Ga_2O_3 Schottky barrier diode deposited by confined magnetic field－based sputtering [J]. Semicond. Sci. Technol. 2021 3(36)：035020－035025.

[6] HOJOONG K，SINSU K，TAIYOUNG K，et al. Effective surface diffusion of nickel on single crystal β－Ga_2O_3 for Schottky barrier modulation and high thermal stability [J]. J. Mater. Chem. C 2019 35(7)：10953－10960.

[7] HAO W B，HE Q M，ZHOU K，et al. Low defect density and small I－V curve hysteresis in NiO/β－Ga_2O_3 pn diode with a high PFOM of 0.65 GW/cm^2 [J]. Applied Physics Letters，2021,118(4):043501－043505.

[8] GONG H H，CHEN X H，XU Y，et al.A 1.86－kV double－layered NiO/β－Ga_2O_3 vertical p－n heterojunction diode [J]. Appl. Phys. Lett. 2020 2(117)：022104－022110.

[9] SCHRODER D. K. in Semiconductor Material and Device Characterization [M]. 2005：127－184.

[10] ZHANG K，XUE J S，CAO M Y，et al. Trap states in InAlN/AlN/GaN－based double－channel high electron mobility transistors [J]. J. Appl. Phys. 2013 17(113)：174503－174510.

[11] KALYGINA V，ALMAEV A，YU P，et al. Anomalous temperature dependence of the electrical conductivity in metal/β－Ga_2O_3/n－Si structures [J]. Superlattices Microstruct. 2020 (141)：106491－106496.

[12] MATTHEA A. PECK MA. A. LANGELL，Comparison of Nanoscaled and Bulk NiO Structural and Environmental Characteristics by XRD，XAFS，and XPS [J]. Chem. Mater. 2012 23(24)：4483－4490.

[13] YAN X D，TIAN L H，CHEN X B. Crystalline/amorphous Ni/NiO core/shell nanosheets as highly active electrocatalysts for hydrogen evolution reaction [J]. J. Power Sources 2015 (300)：336－343.

作者简介：
洪悦华(1996—)，男，广东汕尾人，博士，学生，主要从事宽禁带半导体器件研究。

新一代宽带卫星数字透明处理器研究

陈 战，魏 星，乐立鹏，安印龙
（北京微电子技术研究所（西安分部），陕西省 西安市 710119）

摘 要：作为新一代宽带卫星通信系统的关键技术之一，数字透明处理器技术可以准确提取上行链路中各子带信号并重构出下行链路信号。研究和设计了一种高效数字透明处理器及信道化技术，搭建了一种高效数字透明处理器实现结构，对功能、性能进行了仿真验证。研究成果可为新一代宽带通信卫星的星上柔性转发器中数字透明处理器提供参考。

关键词：柔性转发器；数字透明处理；数字信道化；多相滤波
中图分类号：TN91 **文献标识码**：A

Study on Digital Transparent Processor of New Generation Broadband Satellite

Chen Zhan, Wei Xing, Yue Lipeng, An Yinlong
(Beijing Microelectronics Technology Institute (Xi'an Branch), Xi'an, 710119, China)

Abstract: As one of the key techniques of new generation broadband satellite communication system, the digital transparent processing technique can extract the sub-band signal of the uplink channel effectively and reconstruct the required downlink signal. A high efficiency digital transparent processing and digital channelization technique is investigated and designed in this paper. An efficient structure of digital transparent processor has come up, the functions and performances of the system are simulated. The research findings can be used reference of the digital transparent processor in the flexible transponder of new generation broadband communication satellite.

Key words: flexible transponder; digital transparent processing; digital channelization; polyphase filter

0 引言

在宽带卫星移动通信系统中，系统的通信容量主要取决于星载转发器的性能。目前，星载转发器主要分为再生式转发器和透明式转发器两种[1，2]。其中，再生式转发器具有频谱利用率高、通信质量好等优点，且可进行星上交换处理，但是会大大增加星上设备复杂度和卫星功耗；透明式转发器具有容量大、结构简单、可靠性高等优点，但是频带划分由用户终端决定，且抗干扰能力低及不适用于大容量系统。

柔性转发器作为新一代宽带卫星移动通信系统中的关键组成部分，规避了再生式转发器和透明式转发器的弊端，兼具了传统转发器高效性和灵活性的优点，借助于数字信道化技术的发展，可以实现均匀带宽信号或非均匀带宽信号处理[3，4]。数字透明处理器为柔性转发器的基本组成单元之一，实现对特定波束覆盖区域内接收信号的分路与交换后信号的合路功能。它通过与交换模块的结合，实现各波束覆盖区域内所有用户信号的交换。本文给出了一种具有高效实现特点的数字透明处理器结构，该结构能够以较低的计算复杂度实现子信道的分离与合成以及交换过程。该数字透明处理结构主要由分路单元、电路交换单元、合路单元三部分组成，其中分路单元和合路单元中通道滤波器组均由原型滤波器加 FFT 的模型来实现。

1 星载数字透明处理器及信道化模型

1.1 数字信道化基本原理

传统信道化处理采用低通滤波器组将宽带信道分成多个子带输出，对于实信号的频谱做信道划分，表达式如下：

$$\omega_k = \left(k - \frac{2D-1}{4}\right) \cdot \frac{2\pi}{D} \quad k = 0,1,2\cdots,D-1 \quad (1)$$

将实信号和复本振 $\exp(j\omega_k n)$ 相乘后对实信号做 I/Q 正交化处理。每路通过乘以不同的复本振把原始的宽带信号对应的不同中心频率信号移到零频，然后通过低通滤波器，每路的复信号带宽是实信号的一半，然后进行 D 倍数的抽取。调制余弦调制滤波器组和复指数调制滤波器组[5]，本文采用复指数调制滤波器组。

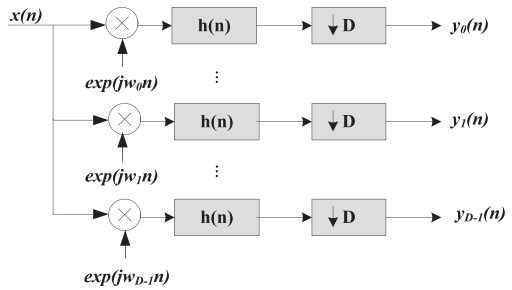


图 1 低通滤波器组的信道化结构

Fig.1 The architecture of channelization with lowpass filter groups

本文研究的数字透明处理器中数字信道化处理部分是基于多相滤波器和 FFT 相结合实现的，推导过程如下：

根据图 1，第 k 信道的输出为

$$\begin{aligned} y_k(m) &= x(n)e^{(j\omega_k n)} * h(n) \\ &= \sum_{i=-\infty}^{+\infty} x(n-i)e^{j\omega_k(n-i)} \cdot h(i) \\ &= \sum_{p=0}^{D-1} s(m) \cdot e^{-j\omega_k p} \end{aligned} \quad (2)$$

将式（1）代入式（2）得到

$$y_k(m) = \sum_{p=0}^{D-1} [s(m) \cdot e^{j\frac{(D-1)\pi}{D}}] \cdot e^{-j\frac{2\pi}{D}kp}$$

$$= \sum_{p=0}^{D-1} x_p'(m) \cdot e^{j\frac{\pi}{2D}p} = DFT\left(x_p'(m)\right) \quad (3)$$

由上可知，整个数字信道化处理器的计算可以简化为多相滤波器和 D 点的 DFT 运算，DFT 运算可以用快速算法 FFT 代替，这样处理速度更快，通常为了计算方便可以选取信道数 D 尽量为 2^n。数字信道化器对应的输入为单路 $x(n)$ 信号，输出为 D 路子信号 $y(m)$，其结构如图 2 所示。

运算复杂度方面，低通滤波器组需要的乘法次数为：$N_1=D\cdot(1+N)$，多相结构需要的乘法次数为：$N_2=N+2\cdot D\log_2 D$。假设原型滤波器 4608 阶，信道数 512 计算得到 N_1=2359808，N_2=10240，因此多相滤波器结构实现信道化可以大大降低硬件资源，提高运算效率。

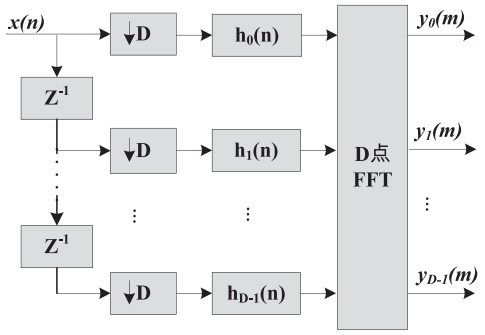


图 2 多相滤波器的信道化结构

Fig.2 The architecture of channelization with polyphase filter

1.2 宽带数字信道化器结构设计

1.2.1 系统的参数设定

在均匀宽带信道化体制中，上行链路被划分为若干个等带宽子信道，每个用户根据带宽需求占用一定数量的子信道。各个子信道信号经过电路交换单元、重构单元后重构得到完整的下行链路信号，则数字信道化器完成了均匀带宽信号跨波束、跨频段的交换。

本设计的数字信道化器均分有效子带数为 1600 个，子带带宽为 1.25MHz，为了方便 FFT 运算，子信道数倍扩展到 2048 个，从低频到高频依次编号为 1 ～ 2048，具体参数如下：

（1）输入信号带宽：2000MHz；

（2）子带数：2048 个；

（3）子带带宽：1.25MHz。

1.2.2 数字信道化器高效结构设计

由 1.1 节的推导得到分路器的高效结构如图 2 所示，合路器的推导与分路器类似，推导过程此处不再赘述，可以认为合路器是分路器的逆过程，这里直接给出合路器的高效结构如图 3 所示：

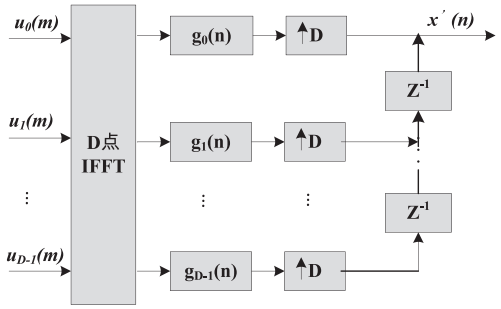


图 3 合路器的高效结构

Fig.3 Efficient architecture of combiner

完整的数字信道化器由分路单元、数字信号交换和处理单元以及合路单元组成，如图 4 所示，数字信道化器中采用的关键技术包括数字信道化技术、多相分解技术及多抽样率数字信号处理技术。

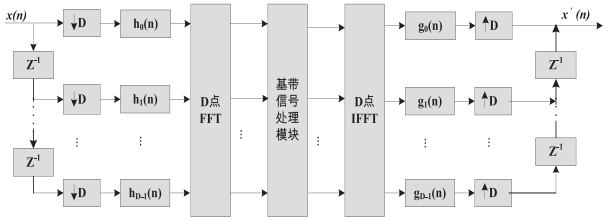


图 4 数字信道化器高效结构

Fig.4 Efficient architecture of digital channelization

2 数字透明处理器设计与仿真

根据系统参数设计的数字透明处理器结构框图如图 5 所示：

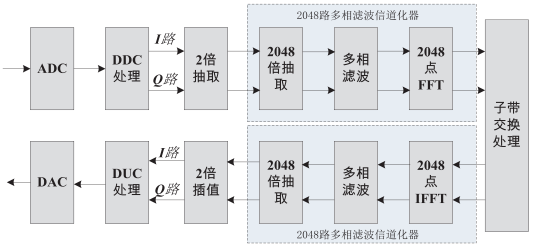


图 5 数字透明处理器原型结构

Fig.5 Prototype architecture of digital transparent processor and channelization

系统的工作流程为：下行将基带信号划分为 2048 个子带，通过下行信道化处理器以及数字上变频合为一路信号经 DAC 发送。上行链路对接收信号 ADC 采样后数据进行下变频，经过上行信道化处理器将基带信号恢复出来。

其中原型滤波器是整个信道化处理器中的关键部分，对于本文中的信道化处理器，输入信号采样率 5120MHz，子带数目 2048 个，每个子带的采样频率为 2.5MHz，频谱 50% 交叠。根据上述指标利用 MATLAB 的 fdatool 工具进行原型滤波器的设计，其幅频率响应如图 6 所示

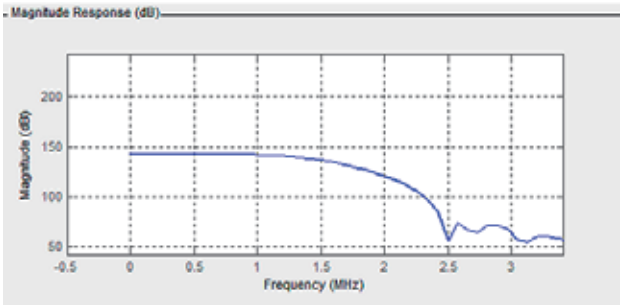


图 6 原型滤波器的频谱

Fig.6 Spectrum of prototype filter

可以看出，原型滤波器通带截止频率约为 1.25MHz，阻带截止频率为 2.5MHz。阻带衰减接近 70dB。将原型滤波器进行 2048 路信道化分解，其频谱如图 7 所示，滤波器特性完全符合系统要求。

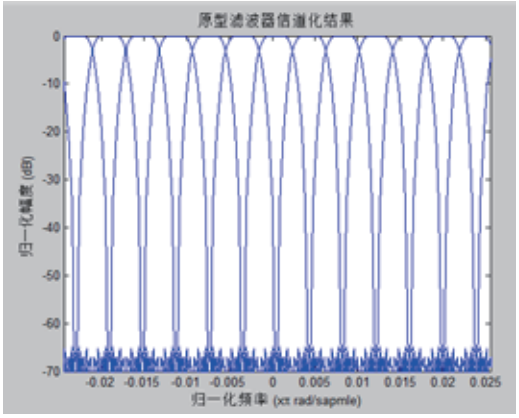


图 7 原型滤波器 2048 路信道化分解频谱

Fig.7 2048 channelized spectrum of prototype filter

MATLAB 环境下，在下行（合路）发送端第

400 个子带（IFFT 输入处）加频率为 1.25MHz 的单频复信号（图 8(a)），上行接收到的信号（FFT 输出）归一化后如图 8(b) 所示，可以看出接收端对发送信号进行了较为精确重构，并且在邻道几乎没有泄露。

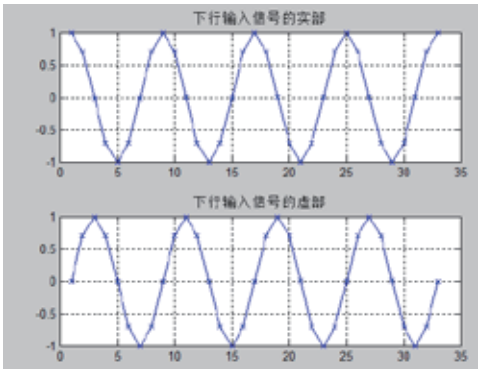


图 8 (a) 发送端信号

Fig.8 (a) Transmitting signals

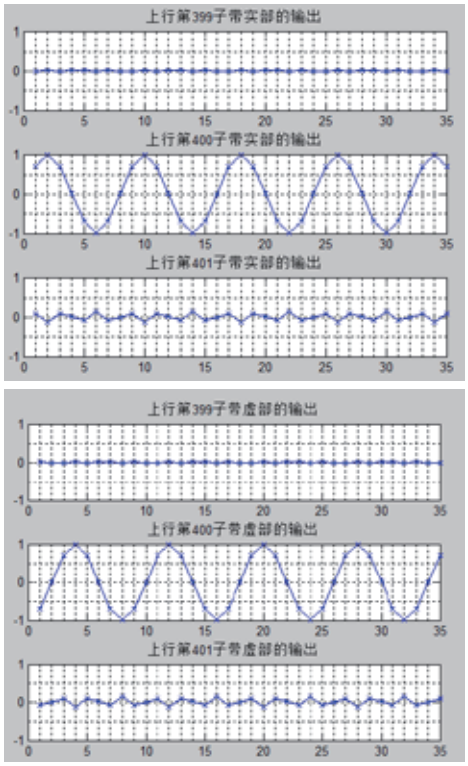


图 8 (b) 接收端信号

Fig.8 (b) Receiving signals

3 数字透明处理器的实现

根据图 5 所示的数字透明处理器原型架构，整个系统的具体实现框图如图 9 所示。

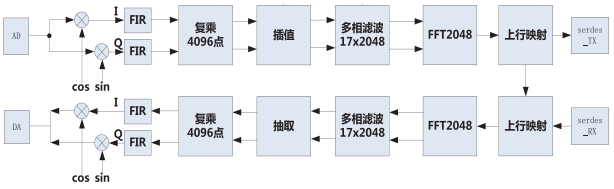


图 9 数字透明处理器硬件实现结构

Fig.9 Hardware architecture of digital transparent processor

3.1 各模块具体实现

整个处理流程的模块划分包括：

（1）DDC 模块：对 AD 来的实信号进行正交分解，变成 I/Q 两路复信号，分别进行 FIR 低通滤波，对滤波结果进行 2 倍抽取。

（2）多相滤波器：上行多相滤波器对 DDC 输出的两路数据按照列的方向写入 2048 个数据并插入 2048 个 0，然后按照行的方向进行多相滤波；下行多相则是对 IFFT 输出的结果进行多相滤波，数据方向与上行一样，多相滤波后按列方向进行 2048 倍抽取，输出一列，丢弃一列。

（3）复乘实现子带的频谱搬移，上行复乘在多相滤波的输出进行，下行复乘在多相滤波之前进行，复乘因子是长度为 4096 点整周期的正余弦。

（4）FFT/IFFT 模块：点数 2048，采用 1 级基 2、5 级基 4 的分裂基结构，较基 2 更节省运算资源。

（5）前后向映射：后向映射是根据映射表地址进行有效子带的挑选，前向映射则是将有效子带还原到原来所在位置。

（6）DUC 模块：将下行多相滤波输出的数据做 2 倍的插值并进行 FIR 低通滤波，然后将 I/Q 路合为实信号输出到 DAC。

（7）发送与接收数据采用 204B 协议的高速 SerDes 接口，数据有效传输速率为 5Gbps。

3.2 处理速率及硬件资源

由于系统时钟高达 5120MHz，22nm 工艺下无

法单路处理。因此，在实际硬件实现中采用了 8 路并行处理的结构，每路时钟速率为 640MHz，此可完全满足速率要求。

整个处理器使用 SRAM 14 种，主要用于数据缓存和行列之间的交织转换，总容量大约 11.48Mbit。FIR 低通滤波、复乘、FFT 和多相滤波器总共使用乘法器和加法器共约 4000 个。

在 VCS 仿真环境下对上述实现进行了仿真，同样在下行第 400 个子带加频率为 1.25MHz 的单频复信号，观察上行第 400 个子带的输出，其结果如图 10 所示，与图 8 (b) 中 MATLAB 接收端的结果一致，验证了实现的正确性。

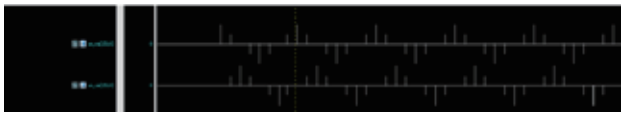


图 10 数字透明处理器仿真结果

Fig.10 The simulation result of digital transparent processor

4 结论

星上载荷对宽带通信信号的传输、处理及存储投入了越来越多的时间和资源。本文设计的星载数字透明处理器性能优良，处理带宽较上一代提升 4 倍至 2GHz，子带数提升至 2000 个，可以有效提高卫星通信的处理带宽和传输质量，提升频谱利用率。在保证高传输速率的前提下，可实现复杂频率信号任意频段子信号间的灵活交换。

参考文献（References）

[1] 郭道省，张邦宁，甘仲民．透明转发器卫星通信系统在干扰条件下的性能 [J]．通信学报，2003，2：118-124．

[2] 朱子行，赵尚弘，李勇军，等．再生式通信卫星转发器的研究进展 [J]．电讯技术，2009，8：47-153．

[3] 易克初，孙永军．数字通信理论与系统 [M]．北京：电子工业出版社，2013．

[4] 张飞，张更新，王可青，等．卫星通信中柔性转发技术研究 [J]．空间电子技术，2012，9(3)：13-23．

[5] BELLANGER M G，BONNEROT G，COUDREUSE M．Digital filtering by polyphase network：Application to sample-rate alteration and filter banks[J]．IEEE Transactions on Acoustics， Speech and Signal Processing，1976，24(2)：109-114．

作者简介：

陈战（1985—），男，陕西省商洛市人，硕士，高级工程师，主要研究方向是通信、雷达信号处理器设计与实现。

Leon3 多核处理器 AMP 模式下并行计算

王 月，李 杰，伍攀峰

（中国空间技术研究院 山东航天电子技术研究所，山东省 烟台市 264000）

摘　要：非对称多处理（Asymmetric Multiprocessing，AMP）运行模式能够实现多核处理器在各自的存储空间独立运行，具有应用灵活的特点。研究 AMP 模式下多核处理器的并行运算可充分利用系统资源，缩短程序运行时间。首先采用 Sparc V8 的原子指令设计自旋锁，实现了多核间对共享资源的互斥访问，然后制定了基于共享内存的通信机制，实现多核间的信息交互。在此基础上，开发了 AMP 模式下多核矩阵乘法并行运算框架。最后在基于 FPGA 的 Leon 3 四核处理器开发板进行了矩阵乘法仿真实验，实现了 AMP 模式下的矩阵拆分，数据分发及并行计算，并验证所提出并行框架的可行性及有效性。

关键词：AMP；并行计算；矩阵乘；自旋锁

中图分类号：TP311　　**文献标识码**：A

Parallel Computing in AMP Mode of Leon3 Multicore Processor

Wang Yue, Li Jie, Wu Panfeng

(China Academy of Space Technology, Yantai, 264000, China)

Abstract: AMP operation mode can realize the independent operation of multicore processors in their respective storage space, and has the characteristics of flexible application. Studying the parallel operation of multicore processor in AMP mode can make full use of system resources and shorten the program running time. Firstly, the atomic instruction of SparcV8 is used to design the spin lock to realize the mutual exclusive access to shared resources between multicores. Then, the communication mechanism based on shared memory is formulated to realize the information interaction between multicores. On this basis, a multicore matrix multiplication parallel computing framework in AMP mode is developed. Finally, the simulation experiment of matrix multiplication is carried out on the Leon3 four core processor development board based on FPGA. The matrix splitting, data distribution and parallel computing in AMP mode are realized, and the feasibility and effectiveness of the proposed parallel framework are verified.

Key words: AMP; parallel computing; matrix multiplication; spin lock

0 引言

多核处理器运行模式主要分为 SMP（Symmetric Multiprocessing）模式和 AMP（Asymmetric Multiprocessing）模式两种[1]。在 SMP 模式下，单个操作系统管理着一个处理器中所有的核。从用户的角度看，操作系统似乎是运行在一个单核处理器上。与之相对应的是在 AMP 模式下，每个处理器核可以拥有自己的操作系统或用户程序，不同的处理器核具有相同或不同的功能，并在各自的存储空间独立运行。这种模式下，同一处理器各核之间既可配合进行并行加速，也可运行相同程序作为备份，同时还为不同的核及其软件之间提供了一定程度的隔离。一般情况下，SMP 模式由操作系统提供相应的管理和调度机制，管理和调度对开发而言可以是透明的，无需人为介入，也可以由开发者根据需要调用相关函数自行管理并行计算过程。AMP 模式则需要由开发者根据具体的应用场景，规划多核各自运行的地址空间，设计多核之间的数据和信息交互通道，制定多核间的交互策略，实现多核间的同步等，进而产生各核所需镜像文件，由各核载入并运行。其开发过程相比操作系

统下的 SMP 模式要复杂，但对于应用扩展相对灵活。

并行计算是多核处理器应用的主要形式之一。而矩阵乘法由于其运算数据量较大，计算密度较高，因而是研究和评估多核并行计算能力的主要方法之一[2]。通过设计 AMP 模式下多核矩阵并行乘法运算方法，缩短矩阵乘法时间，实现矩阵乘法加速，对于研究 AMP 模式下多核处理器资源管理、任务调度机制及优化并行等方面具有重要意义。本文研究并提出一种 AMP 模式下矩阵乘法并行运算框架，设计了基于共享存储的自旋锁及信息交互方法，实现了 AMP 模式下的矩阵乘法并行计算。在基于 FPGA 的 Leon3 开发板上进行了仿真实验，通过实验验证了所提出并行框架的可行性及有效性。

1 基于共享存储的自旋锁与信息交互

1.1 基本架构

处理器多核采用主从工作模式。其中，核 0 为主核，其余为从核。加电启动后，只有主核进入正常工作状态，运行初始化和管理程序，所有从核均为 power-down 状态。主核完成初始化工作后，根据程序设计在适当时机由主核逐一唤醒从核。被唤醒的从核运行自己的应用程序，提供相应功能。

为实现并行运算，多核之间需要在数据分发、结果收集、启动配合等方面进行数据和信息的交互。一般可通过核间的消息传递，或共享存储区这两种方式来实现[3]。本文采用共享存储区实现信息交互。基本方法如下图 1 所示。

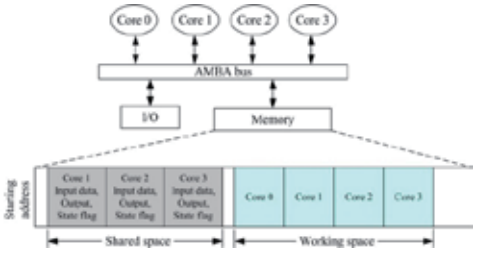


图 1 Leon3 四核处理器共享存储架构示意图

Fig.1 Schematic diagram of Leon3 quar core processor shared storage architecture

1.1 自旋锁

当多个处理器核同时访问共享资源时，会产生访问冲突问题。为保证多核间临界资源的同步互斥访问，以自旋锁为最常用的方法[4]。

自旋锁维护一个锁变量值 flag，其原理示意图如图 2 所示。进行初始化时将锁变量值初始化为 data1，表示该自旋锁处于空闲状态。进行加锁操作时，申请锁处理器核循环判断锁变量值，当其空闲时成功获取锁，并原子性修改锁变量值为 data2，表示自旋锁正在被占用，其它申请自旋锁的处理器核读取到该变量为 data2 则自旋等待。进行解锁操作时，将锁变量值重新置为 data1。

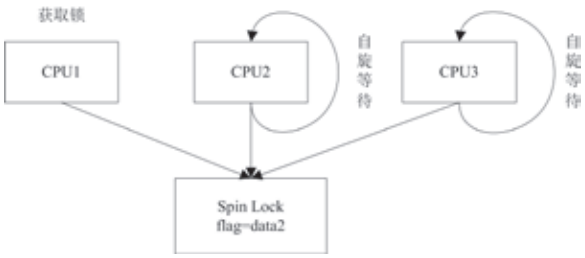


图 2 自旋锁示意图

Fig.2 Schematic diagram of spin lock

锁的实现主要考虑两方面，一是与锁代码相关的指令，二是处理器核检查锁时其状态的可见性[5]。锁是通过系统内存中特定位置的一块数据来实现。在最简单的情况下，数据是内存中的一个字节。锁机制的实现需要处理器原语指令的配合，大多数处理器都提供一定形式的 test-and-set 指令，即在一个原子操作中完成对内存的读和写操作，以保证这个过程不会被打断。不同处理器核提供不同原语访问指令，在基于 SPARC V8 的处理器中，使用 ldstub 指令、casa 指令等。另一方面，当锁的值发生改变时，该变化需要对所有正在运行的处理器核可见，首先高速缓存实现 write-through 模式，使占用锁的处理器核写入到高速缓存的数据也会同步更新到内存中[6, 7]，以及通过处理器硬件总线协议保证高速缓存的一致性，使其它申请锁的处理器核读取到修改后的最新数据[8, 9]。

1.2 信息交互

在处理器的存储空间中，开辟出一个共享区域，

如图 1 所示。共享存储区主要用于主核与从核间的数据交互、状态信息交互，所有的核都可访问共享存储区，每个从核各自有三个区域：输入数据区、输出数据区、状态标志区。数据与信息交互的基本流程如下：

(1) 主核完成初始化，并将各核计算所需的源数据分别写入到各核的输入数据区，供各从核读取；

(2) 从核被主核唤醒后，从各自的输入数据区读走源数据，开始各自的运算；

(3) 从核将各自的计算结果写入输出数据区；

(4) 从核修改状态标志区的标志位，表明已经完成运算，进入等待状态；

(5) 主核查询从核的状态标志，待所有核都完成运算后，将所有从核的结果数据取走；

(6) 主核将所有状态标志清零。

2 AMP 模式下矩阵并行运算框架

2.1 矩阵分块

矩阵乘法运算的并行化主要分为两种情况，第一种情况为存在多个阶数相对较少的矩阵进行乘法运算。各处理器核同时运行一次相应的矩阵乘运算，第二种情况为当矩阵的阶较大时，进行一次矩阵乘运算需要消耗大量的时间，将一次矩阵乘运算拆分到多个处理器核并行运行。第一种情况为第二种情况实现的基础。

第二种情况中，$m\times m$ 阶方阵 A 乘 $m\times m$ 阶方阵 B 得 $m\times m$ 阶方阵 C，即 $A_{mxm}\times B_{mxm}=C_{mxm}$（其中 $m\geqslant 1$），根据多核处理器中处理器核数（处理器核数为 n），对方阵进行分块运算。本并行化框架适用于处理器核数为 $n=2^k$ 的情况下，当 k 为奇数时，方阵 A 和 B 分为 2^{k+1} 块，即 m 阶方阵 A 和 B 分为 $(k+1)\times(k+1)$ 个大小为 $p\times p$ 的矩阵子块 C_{ij}；当 k 为偶数时，方阵 A 和 B 分为 2^k 块，即 m 阶方阵 A 和 B 分为 $k\times k$ 个大小为 $q\times q$ 的矩阵子块 C_{ij}，其中，$p=m/(k+1)$，$q=m/k$（当 p，q 不为整数时，使用 0 元素对方阵 A 和 B 进行填充）。矩阵分块乘运算中 $C_{ij}=\sum_{a=1}^{k}A_{ia}B_{aj}$，当 k 为奇数时，每个处理器核进行四个矩阵子块 C_{ij} 运算，即 C_{ij}，$C_{i+1,j}$，$C_{i,j+1}$，$C_{i+1,j+1}$；当 k 为偶数时，每个处理器核进行一个矩阵子块 C_{ij} 运算。

2.2 并行运算框架

2.2.1 系统描述

在 4 核 Leon3 处理器中，每个处理器核都用于各自的指令缓存，数据缓存和 AHB 总线接口，并在 AHB 总线上设置总线窥探，保持数据缓存的缓存一致性。多核处理器配置为 AMP 模式，框架程序为每个处理器核分配一块连续的内存空间，处理器核在其中运行对应矩阵乘法运算程序。由于内存空间重叠会导致程序运行出错，在编码阶段之前需要仔细地进行内存空间划分，并控制各个处理器核运行空间可以容纳对应矩阵子块的大小。在 AMP 模式下各个处理器核除运行各自程序外，不同处理器核还需要进行核间通信，需要在存储器空间中分配共享内存区域，不同处理器核通过共享内存传输入数据，输出结果以及其它用于交互的数据信息。同时在 Leon3 多核处理器中，处理器核 0 是唯一一个在系统启动后处于工作模式的核心，其它核心在 power down 模式下保持静止。在系统初始化后，其它处理器核可以通过中断控制器进行启动。

2.2.2 框架设计

本文并行框架实现第二种情况，即在多核处理器上的矩阵乘并行处理的执行模式，流程如图 3 所示。矩阵乘法运算步骤如下：

(1) 在系统初始化后，只有处理器核 0 启动运行，开始程序；

(2) 处理器核 0 程序首先初始化共享内存中存放的公共变量，如运算完成标志变量、矩阵数据等；

(3) 处理器核 0 将矩阵 A 和 B 矩阵数据存入共享区域后，启动其它处理器核；

(4) 各处理器核启动完成后，将共享存储中所需的分块矩阵数据读取到其运行空间内；

(5) 处理器核对读取的矩阵数据进行相应的矩阵

乘和矩阵加法运算；

(6) 各处理器核将求得的分块矩阵 C_{ij} 的结果通过自旋锁机制互斥地存入共享区域，并保证数据的完整性；

(7) 处理器核完成矩阵运算并存放数据后，将运算完成标志中对应位置 1；

(8) 处理器核 0 完成对应运算后，判断运算完成标志中是否全部处理器核完成运算，若是则进行步骤 9，若否则继续循环判断至全部完成；

(9) 将分块矩阵结果组合为矩阵 C，并输出运行结果，程序运行结束。

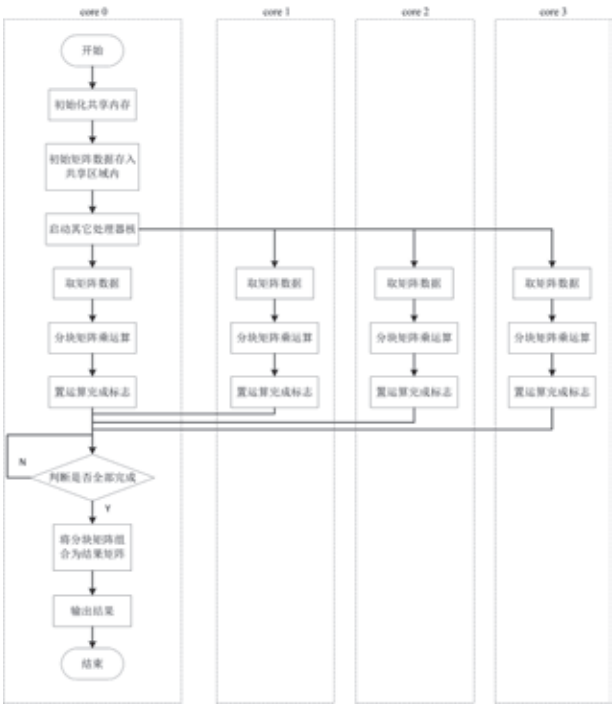


图 3 矩阵并行运算流程图

Fig.3 Flow chart of matrix parallel operation

3 实验仿真与分析

3.1 实验设置

使用 Xilinx 的 Nexys 开发板作为开发平台，配置了 Leon3 四核处理，其中每个核的数据缓存和指令缓存均为 4×4KB，没有设置二级缓存，处理器单核工作频率为 100MHz。

使用 C 语言对本文提出的矩阵并行乘运算框架进行了编程。无需操作系统支持，程序编译采用了 Conham Gaisler 提供的交叉编译器 bcc 2.0。通过设置处理器核编号、运行空间起始地址、堆栈地址等参数，编译获得各核自己的可执行文件。调试工具使用 Grmon3，在 Grmon3 中，将各核的可执行文件分别加载，加载完成后运行。

处理器核空间分配及堆栈地址如下表所示。

表 1 处理器核空间分配及堆栈地址

Tab.1 Processor core space allocation and stack address

序号	项目	空间分配	运行空间大小（MB）	堆栈地址
1	共享存储空间	0x40100000−0x401fffff	1	−
2	处理器核 0	0x40200000−0x40ffffff	13	0x40fff000
3	处理器核 1	0x41000000−0x41ffffff	15	0x41fff000
4	处理器核 2	0x42000000−0x42ffffff	15	0x42fff000
5	处理器核 3	0x43000000−0x43ffffff	15	0x43fff000

3.2 矩阵乘法实验结果

在仿真实验中，各处理器核均运行矩阵大小为 32×32，48×48，64×64，200×200，300×300，400×400，500×500，600×600 共 9 组的矩阵数据进行并行乘法测试（运行时间测试不包括输出结果），并将程序在并行模式运行时间与串行模式运行时间的比例（加速比）作为衡量加速效果的依据 [10]。由于相同程序运行时间可能存在微小差异，仿真实验中对每组测试数据运行 20 次，最终加速比取 20 次加速比的平均值。不同大小矩阵乘并行运算时间及加速比如表 2 所示。

表 2 小维度矩阵乘法运算时间及加速比

Tab.2 Multiplication time and speedup ratio of small dimension matrix

矩阵大小	串行处理时间/μs	并行处理时间/μs	加速比
32×32	5105	3697	1.381
48×48	16604	8385	1.980
64×64	43013	18084	2.518
100×100	203693	52511	3.879
200×200	1745500	562278	3.104
300×300	6313794	2026165	3.116
400×400	19172536	5715837	3.354
500×500	42959760	11419078	3.762
600×600	91636171	20619449	4.444

3.3 结果分析

各处理器核在访问 AMBA 总线、在共享存储空间存取数据、在内存空间存取数据时，需要争夺自旋锁的控制权，只有抢到锁的处理器核才能控制总线进行存取操作，其他处理器核只能等待锁被释放后再尝试加锁。由于 AMBA 总线同一时刻只能有一个控制者，所以各核抢锁并控制总线这个过程实际上是串行的过程。这个过程极大影响了并行的效果。根据表 2 中数据可以看到，本文实验中这个串行过程主要受到两个影响因素：一是各处理器核从共享存储区搬运源数据的时间，一是计算过程中各处理器核与内存间搬运临时数据的时间。

在矩阵维度较小时，源数据搬运过程在各处理器核矩阵处理过程中所占时间比例相对较大。同时，矩阵小，过程中中间变量较小，需要到内存搬运数据的次数也相对较少，数据搬运开销相对小，所得加速比结果较小。

当矩阵维度增大时，源数据搬运过程所需要的时间虽然增大，但其在整个处理过程所占比例下降，同时由于矩阵增大，运算过程访存次数大大增加，在处理器核缓存数据有限的情况下，数据量增大造成访存次数增加对运行时间的影响更大。由于串行模式下用于运算的数据量是并行模式的 4 倍，因此串行处理时间随矩阵维度增长迅速。因此，随矩阵维度的扩大加速比逐渐增加。

由于受到各处理器核数据缓存大小以及处理器核运行空间大小的影响，当矩阵数据量比较大时，各核缓存中数据更新次数大大增加，频繁地抢占锁和占用总线导致各核在进行各自矩阵处理的时间都在增加。串行模式下需要更为频繁在处理器核与内存间交换数据。在这种情况下，串行模式访存次数的增加与并行模式访存次数的增加已不再是线性关系，其增长更快，导致加速比大幅度上升，如表 2 中 600×600 矩阵的运算情况。

4 结束语

本文研究了一种 AMP 模式下多核处理器矩阵乘法并行运算框架，实现了一种基于共享存储的自旋锁，设计了多核之间信息交互方法，在此基础上实现了 AMP 模式下的矩阵乘法并行计算。在基于 FPGA 的 Leon3 四核处理器开发板上进行矩阵乘法仿真实验，通过实验验证了所提并行运行框架的可行性及有效性。在本文设定的矩阵维度范围内，在 AMP 模式下 Leon3 四核处理器进行矩阵乘法并行计算，相比串行计算其加速比最高可达到 4.44，但本文所涉及的运算比较简单。在实际应用中，如果考虑到系统内外各种交互，事务处理等因素造成的开销，实际加速比可能还会有所变化。

实验结果表明，并行运算时，核间信息交互与数据交互、核与内存间数据的搬运对加速效果有较大的影响。在开发并行算法时，应对此加以考虑，使得并行运算所需数据量与各处理器核缓存大小相适配。相关研究在后续开展。

参考文献（References）

[1] 董延军，项涛．多核嵌入式操作系统及板级结构探讨[J]. 信息通信，2018(12)：145-147.

[2] FIALKO S. Parellel Direct Solver for Solving Systems of Linear Equations Resulting from Finite Element Method on Multi-core Desktops and Workstations[J]. Computers & Mathematics with Applications, 2015,70(12)：2968-2987.

[3] 韩乐．多核体系结构通信机制的研究与优化[D]. 中国科学技术大学，2014.

[4] 王海峰，蒋晓华．基于 SPARC 多核 SOC 的 Linux 操作系统研究[J]. 航天控制，2017,35(03)：49-53.

[5] 虞保忠，郝继锋．多核操作系统自旋锁技术研究[J]. 航空计算技术，2017,47(04)：115-117.

[6] GANG H, ZENG H, MD N, et al. Experimental Evaluation and Selection of Data Consistency Mechanisms for Hard Real-Time Applications on Multicore Platforms[J]. IEEE Transactions on Industrial Informatics, 2014, 10(2)：903-918.

[7] CHWA H S, LEE J, PHAN K M, et al. Global EDF Schedulability Analysis for Synchronous Parallel

Tasks on Multicore Platforms[C]//Real-Time Systems (ECRTS), 2013 25th Euromicro Conference on. 2013.

[8] VIJAY N, et al. A Primer on Memory Consistency and Cache Coherence[M]. San Rafael: Morgan & Claypool Publishers, 2020.

[9] AKSHAY S, et al. DynaCo: Dynamic Coherence Management for Tiled Manycore Architectures[J]. International Journal of Parallel Programming, 2021:1-30.

[10] 肖汉，肖诗洋，李彩林，周清雷．异构平台上基于 OpenCL 的矩阵乘并行算法[J].西南大学学报(自然科学版),2020,42(11):147-153.

作者简介：

王月(1970—)，女，吉林通化人，硕士研究生，主要研究方向为航天器嵌入计算机系统，多核空间技术应用等。

一种高安全可抵御 StarBleed 漏洞攻击的 FPGA 硬件防护设计方法

杨佳奇，陈 雷，李学武，孙华波

（北京微电子技术研究所，北京市 100076 ）

摘　要：本文针对国外披露的 Xilinx 7 系列 FPGA 存在的 StarBleed 漏洞，开展了 Xilinx 7 系列 FPGA 电路漏洞机理分析，搭建了加密码流安全性仿真验证平台，设计攻击码流进行了 FPGA StarBleed 漏洞验证，创新性地提出了一种高安全可抵御 StarBleed 漏洞攻击的硬件电路防护设计策略，通过仿真验证了修改后的硬件电路可抵御恶意码流的攻击。最后，FPGA 在电路回片后经过板级加密安全性测试，证实了电路已从硬件上解决了 StarBleed 漏洞，提升了 FPGA 的安全性与可靠性。

关键词：7 系列 FPGA；StarBleed 漏洞；加密；鉴权

中图分类号：TN47　　**文献标识码**：A

A High Security FPGA Hardware Defence Method against StarBleed Vulnerability Attack

Yang Jiaqi, Chen Lei, Li Xuewu, Sun Huabo

(Beijing Microelectronics Technology Institute, Beijing, 100076, China)

Abstract: Considering StarBleed vulnerability existed in Xilinx 7 Series FPGA disclosed abroad, by analysing vulnerability mechanism of 7 Series FPGA circuit, building encrypted bitstream security simulation platform, designing attack bitstream and simulating StarBleed vulnerability attack, this paper proposes innovatively a high security FPGA hardware defence method against StarBleed vulnerability attack. The redesigned circuit has the ability against StarBleed vulnerability attack by simulation. Finally, after chip returning FPGA circuit passes the test of encrypted bitstream security on board, confirming StarBleed vulnerability solved from hardware circuit and improving the security and reliability of FPGA.

Key words: 7 Series FPGA; StarBleed vulnerability; encryption; authentication

0 引言

作为电子信息系统中的核心元器件，FPGA 所扮演的角色越来越重要。目前高性能高可靠 FPGA 已广泛应用于深空探测、宽频通信卫星、高分辨率对地观测、预警侦察卫星、科学实验卫星等航天工程和新一代武器装备中 [1]。这些领域都要求系统具有极高的稳定性与安全性，若 FPGA 芯片中的设计信息被恶意窃取，将会带来不可估量的影响。

2020 年 4 月，来自德国波鸿鲁尔大学和 Max Planck 网络安全与隐私研究所的研究人员披露了一个名为“StarBleed”的安全漏洞，它存在于 Xilinx 公司的 Virtex7、Kintex7、Artix7、Spartan7 等全部 7 系列 FPGA 产品，而 UltraScale 和 UltraScale+ 等高端系列 FPGA 则不会受到影响。利用这个漏洞，攻击者可以同时攻破 FPGA 码流文件的加密和鉴权，并由此可以随意修改 FPGA 中实现的逻辑功能。更严重的是，该漏洞并不能通过软件补丁的方式修复，一旦芯片被攻破，就只能更换芯片 [2]。

漏洞的攻击者利用 7 系列 FPGA 内部 WBSTAR 寄存器的特殊性，以及鉴权过程晚于加解密过程的弱点，彻底攻破了 7 系列 FPGA 的安全保护机制。

为了从硬件电路上解决 StarBleed 漏洞，本文分析了 FPGA 加密策略和漏洞攻击机理，搭建了 FPGA 加密安全仿真验证平台，并对 Xilinx 7 系列 FPGA 进行模拟攻击仿真验证，结果表明 Xilinx 7 系列 FPGA 确实存在 StarBleed 漏洞。随后本文基于 FPGA 的 WBSTAR 寄存器的特性提出了鉴权控制回读的安全策略，重新设计后的电路经过恶意码流攻击验证，仿真结果表明，更改设计后的电路具备可抵抗 StarBleed 漏洞攻击的特性。本文最后进行了芯片回片测试，实测结果证明 FPGA 电路已不存在 StarBleed 漏洞。

1 背景介绍

1.1 FPGA 介绍

与 CPU、ASIC 等芯片相比，FPGA（现场可编程门阵列）芯片最大的特点是可配置性，它内部包含了大量的可编程逻辑阵列、可编程输入输出单元以及可编程布线资源。随着 FPGA 在数据中心、人工智能以及网络通信等领域的应用，它内部还嵌有数字信号处理（DSP）、大容量存储器（BRAM）、时钟管理模块（CMT）以及高速接口等丰富逻辑资源 [3]。

FPGA 所有的设计信息都包含在码流文件中，对于 SRAM 型 FPGA，码流文件中的 bit 信息都需要下载存储到 SRAM 阵列中。为了满足不同应用的需要，FPGA 支持多种配置模式，如 SelectMAP、JTAG、串行、SPI、BPI 等 [4]。FPGA 的配置功能主要包括以下 3 种：

①初次配置功能。该过程可以分为单片配置和多片配置，明文配置与加密配置。多片配置包含从串菊花链配置、从并菊花链配置，JTAG 菊花链配置等。密文配置支持 AES-GCM 码流加密鉴权机制 [5]，可支持 JTAG、SelectMAP 等方式。

②重配置功能。FPGA 芯片的重新配置有三种触发方式，一是通过 PROGRAM 引脚对芯片进行复位，重新启动配置进程；二是将重配置命令通过 ICAP 端口发送或是将其嵌入到配置码流中从而触发重配置功能；三是当配置或工作过程出错时，FPGA 本身会触发 FallBack 的重配置功能。

③回读功能。通过回读，用户可以检测 FPGA 内已经写入的配置数据是否保持正确。用户可以自行通过 SelectMAP、ICAP 以及 JTAG 端口回读数据，然后按照掩码文件与原配置文件逐位对比来确定是否相同。

1.2 加密配置

为了保护 FPGA 中的设计信息或知识产权，7 系列 FPGA 引入了两种码流保护机制，一种是加密，另一种是鉴权。

1.2.1 加密

加密是指采用特定的算法对码流文件进行处理，将其转换为密文，使得其中的内容对外不可见 [6]。在 Xilinx 7 系列 FPGA 中，使用 CBC-AES-256 算法进行码流加密。7 系列 FPGA 支持 BBRAM 以及 FUSE 存储密钥两种方式 [7]，这两种方式各有优势，可根据用户的具体情形来确定。密钥通过 JTAG 端口写入，密钥加载完成后若尝试回读密钥，BBRAM 的内容以及整个 FPGA 的 SRAM 都将清零 [8]。

1.2.2 鉴权

鉴权指的是对码流文件进行身份验证，防止对其进行篡改和删减，这类似于日常生活中的身份验证，如果加密的码流被修改，势必会导致错误的鉴权结果。如果将修改后的码流下载到 FPGA 中，会导致鉴权错误而配置失败，从而避免被攻击的可能。在 Xilinx7 系列 FPGA 中，使用了基于 SHA-256 的 HMAC（散列消息认证码）方式进行鉴权。非加密码流与 HMAC 密钥经过 SHA-256 运算，生成一个 MAC，它被嵌入到 AES 加密码流中。在加密配置过程中，FPGA 的 SHA-256 引擎会计算 AES 解密后的数据，新生成一个 MAC，并将与码流中的 MAC 进行比较，如果两个 MAC 一致，配置电路将进入启动时序，完成配置。若不同，FPGA 将启动失败，配置出错。图 1 是 HMAC 鉴权的运算流程图 [9]。

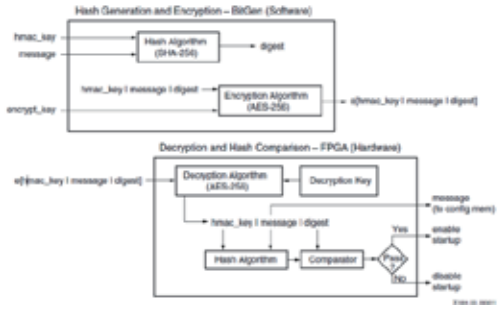


图 1 HMAC 鉴权

Fig.1 HMAC authentication

可以想象，如果码流的加密机制被破解，攻击者可以读出码流文件中的全部信息，从而进行反向工程，IP 破解，信息收集等工作 [10]。如果鉴权机制被破解，攻击者可以对码流文件进行任意修改，如修改系统功能、木马注入等。所以，这两种保护方式缺一不可。

1.3 Multiboot 多重引导

7 系列 FPGA 的 Multiboot 与 Fallback 特性主要用来系统更新，码流版本可以动态升级。如果在 Multiboot 进程中检测到错误，FPGA 将会触发 Fallback 特性来确保下载一个安全（golden）码流到 FPGA 中 [11]。

当发生 Fallback 错误时，FPGA 内部将产生一个脉冲复位整个配置电路，然而特定的 Multiboot 逻辑以及 WBSTAR 寄存器都不会复位清零。WBSTAR 寄存器为可读可写寄存器，保存着下一段码流的起始地址，图 2 是 WBSTAR 寄存器的功能描述。

Table 5-38: WBSTAR Register

Description	RS[1:0]	RS_TS_B	START_ADDR
Bit Index	31	30	29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
Value	0	0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Table 5-39: WBSTAR Register Description

Name	Bit Index	Description
RS[1:0]	[31:30]	RS[1:0] pin value on next warm boot. The default is 00.
RS_TS_B	29	RS[1:0] pins 3-state enable. 0: 3-state enabled (RS[1:0] disabled) (default) 1: 3-state disabled (RS[1:0] enabled)
START_ADDR	[28:0]	Next bitstream start address. The default start address is address zero.

图 2 WBSTAR 寄存器

Fig.2 WBSTAR register

2 StarBleed 漏洞机理分析与安全解决方案

StarBleed 漏洞整个攻击过程分为两个大部分，首先是对加密码流进行破解，读出全部的明文信息，其次是破解原码流的鉴权机制，达到控制整个码流的目的。根据漏洞攻击机理以及 FPGA 鉴权特性，本文提出了一种鉴权控制回读的安全解决方案。

2.1 加密破解

图 3 为 7 系列 FPGA 正常加密码流结构。

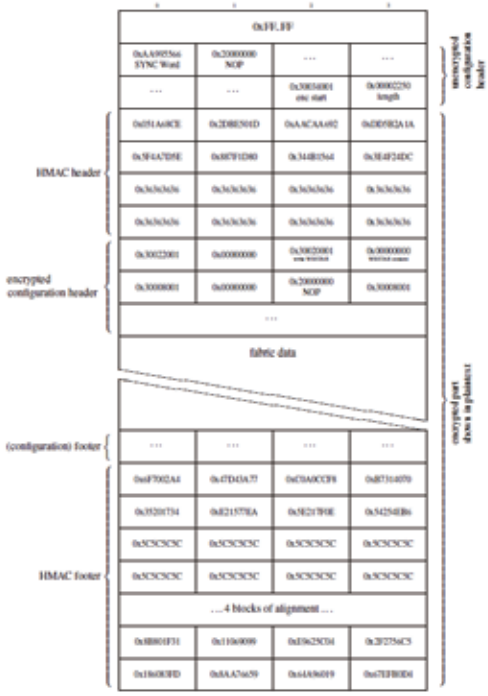


图 3 加密码流结构

Fig.3 Structure of encrypted bitstream

从图 3 中可以看到加密码流结构包含未加密配置包头与密文，其中未加密配置包头包含同步字、码流宽度检测、控制信息以及初始 CBC 值，密文包含 HMAC 包头、加密配置命令包头、加密配置数据、配置包尾以及 HMAC 包尾。

攻击者对加密码流的破解分为五步：

第一步：攻击者对正常加密码流进行篡改，生成攻击码流如图 4，图 4 中一行包含 4 个字，每个字为 32 个 bit，篡改方式为在码流中定位写 WBSTAR 命令的位置（码流生成位置固定），利用 WBSTAR 寄存器写多个字只保存最后一个字的特性，构造攻击码流。而写入 WBSTAR 寄存器的值为解密后的数据，由于鉴权采用 SHA-256 算法，所以需要 512bit 作为

一个运算块。除了图 4 中标①的这一行，另需要构造 3 行，最后一行标④为待解密的数据，假设先解密最后一行的最后一个字，那就需要在 WBSTAR 寄存器中写 13 个字，用 16 进制表示为 D，写 WBSTAR 命令将改变为 0x3002000D（明文），后利用 CBC 的延展性将明文转换为密文，即更改本行的明文，只需更改上一行的密文，然后再构造剩余的两行，图 4 中标②的一行只起填充作用，可以为任意数据，标③的一行需要作为④这一行的解密初始向量，所以应为待解密文的上一行密文。若解密④行中其他 3 个字，只需要改写 WBSTAR 寄存器的字数即可，同时剩下的字改成 0x20000000，防止载入未知的命令，再根据 CBC 的延展性将标③的一行修改为相对应的解密初始向量。由此，攻击码流构造完成。

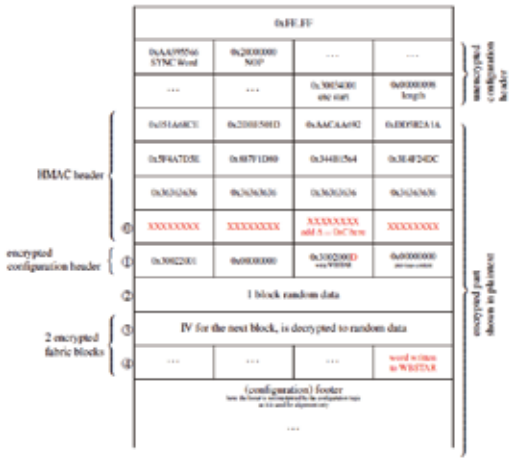


图 4 攻击码流

Fig.4 Attack bitstream

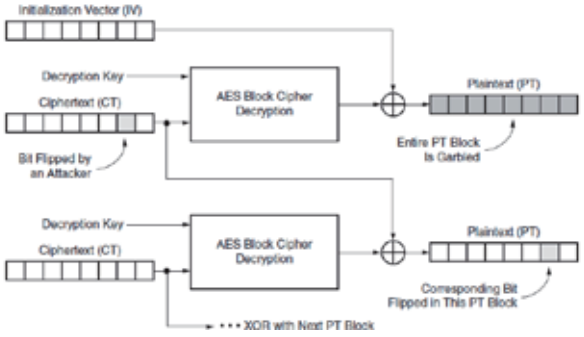


图 5 CBC 延展性

Fig.5 CBC malleability

第二步：将攻击码流下载到 FPGA 中，此时，FPGA 解密攻击码流，解密后的 13 个 32 位字依次写入 WBSTAR 寄存器，但后面写入的数据会覆盖掉之前的数据，所以 WBSTAR 寄存器保存的是最后一次写入的 32 位字，即解密后的原码。

第三步：加载完毕后，由于码流被修改过，所以 HMAC 鉴权失败，触发 FPGA Fallback 机制，开始系统复位。

第四步：载入一个未加密的码流文件，用于读取 WBSTAR 寄存器的内容。由于 WBSTAR 寄存器在发生 Fallback 复位时，其内容不会被清除。因此，攻击者使用一个未加密的码流文件，可以回读出解密后码流中的一个 32 位字。回读码流文件如图 6。

LISTING 1: Readout Bitstream

```
0xFF, 0xFF, 0xFF, 0xFF,
0xFF, 0xFF, 0xFF, 0xFF,
0xFF, 0xFF, 0xFF, 0xFF,
0xFF, 0xFF, 0xFF, 0xFF,
0xFF, 0xFF, 0xFF, 0xFF,
0xFF, 0xFF, 0xFF, 0xFF,
0x00, 0x00, 0x00, 0xBB,
0x11, 0x22, 0x00, 0x44, #BUS Size Detect
0xFF, 0xFF, 0xFF, 0xFF,
0xFF, 0xFF, 0xFF, 0xFF,
0xAA, 0x99, 0x55, 0x66, #SYNC Word
0x20, 0x00, 0x00, 0x00, #NOP
0x30, 0x00, 0x80, 0x01,
0x00, 0x00, 0x00, 0x04,
0x20, 0x00, 0x00, 0x00,
0x20, 0x00, 0x00, 0x00,
0x20, 0x00, 0x00, 0x00,
0x28, 0x02, 0x00, 0x01, #read reg WBSTA
0x20, 0x00, 0x00, 0x00,
0x20, 0x00, 0x00, 0x00,
0x20, 0x00, 0x00, 0x00,
0x20, 0x00, 0x00, 0x00
```

图 6 回读码流文件

Fig.6 Readout bitstream

第五步：手工复位，然后再重复上述步骤，直到整个加密码流全部窃取完毕。

可以看到，攻击者通过“蚂蚁搬家”的方式完成了对加密码流的完全解密。更可怜的是，FPGA 本身也沦为了帮助解密的工具。

2.2 鉴权破解

在攻击者对加密的码流破解完毕后，进行第二部分，即破解码流鉴权。当整个加密码流破解后，攻击

者得到了整个明文，同时也知道原加密密文，由此攻击者可以通过反向提取原明文，插入木马或者修改原设计功能，再生成全新的攻击者设计的密文码流。利用原密文、原明文、新明文以及 CBC 延展性，通过公式（1）逐行计算就可以得到新的加密密文，新的密文可以通过 HMAC 鉴权，成功下载到 FPGA 中，从而实现攻击者的目的。

$$C'_{n-1} = P_n \oplus C_{n-1} \oplus P'_n \quad (1)$$

公式中 C'_{n-1} 为新的加密块，P'_n 为攻击者期望的明文块，C_{n-1} 为原加密块，P_n 为原明文块。公式从最后一行往前算，最后的 C'_0 为 CBC 的初始值。

2.3 安全解决方案

StarBleed 漏洞利用的是 FPGA 内部的 WBSTAR 寄存器的特殊性，以及鉴权过程晚于加解密过程的缺陷，从而攻破了 Xilinx 7 系列 FPGA 的加密与鉴权。因此，相应的解决措施主要包含两个方面，一是增加 WBSTAR 寄存器安全可读特性，二是更新加密鉴权机制，即先鉴权再解密。本文从第一个方面入手，基于 FPGA 的 WBSTAR 寄存器与鉴权特性提出了鉴权控制回读的安全策略。

鉴权控制回读策略，即通过检测鉴权结果信号来判断是否开启 WBSTAR 寄存器的回读权限。若码流鉴权成功后，则执行 FPGA 的启动命令，FPGA 进入正常启动时序。若码流鉴权失败，配置状态机检测到鉴权失败信号会锁死 FPGA 的全部配置接口（如 SelectMAP、JTAG 等），以阻止载入外部恶意数据，同时鉴权失败指示信号使能鉴权控制回读电路，记录当前电路的 HMAC 错误，关闭 WBSTAR 寄存器的回读路径，使 WBSTAR 寄存器成为只写寄存器，从而读不出解密数据，而在其他情形下，WBSTAR 寄存器仍为可读可写寄存器，不影响用户的正常使用。若用户设置开启 FALLBACK，鉴权失败指示信号也会触发 FPGA 的 FALLBACK 特性，复位 FPGA，但不会复位鉴权控制回读电路，保证数据不会泄露。如图 7 为修改后的鉴权检测电路图。

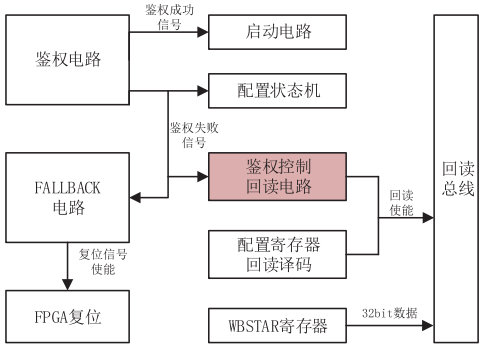


图 7 修改后的鉴权检测电路

Fig.7 Modified authentication detection circuit

3 仿真验证

仿真验证主要包含模拟攻击者窃取码流的过程，验证修改方案的可行性，即是否可以有效阻止攻击，以及修改后的电路对用户正常使用 Multiboot 功能的影响。仿真主要分为 3 个部分：

3.1 模拟攻击者窃取码流

StarBleed 漏洞是按照一字一字依次窃取完整个码流，所以仿真以窃取其中的一个字作为例子，窃取的字选为 0x30008001，码流中的位置如图 8 所示，按照本论文第三章的方法构造攻击码流，如图 9。将构造好的攻击码流载入到 FPGA 中，从图 10 中可以看到 FPGA 将解密后的数据载入到 WBSTAR 寄存器，保留在 WBSTAR 寄存器的值是最后一次写进去的数，即 0x30008001。由于原码流被修改过，载入攻击码流导致 FPGA 鉴权失败，HMAC_ERROR 信号拉高，FPGA 进入 fallback 进程，从图 11 可以看到在清零过程中 WBSTAR 寄存器不会清零，保持 0x30008001。在清零结束后，通过 JTAG 载入回读 WBSTAR 寄存器码流，如图 12 装载回读命令后 WBSTAR_R_EN 使能信号拉高，回读总线上装载了 WBSTAR 寄存器的值 0x30008001，最后窃取到的数据从 TDO 读出，如图 13。从仿真验证结果可知 Xilinx 7 系列 FPGA 确实存在 StarBleed 漏洞，且该漏洞可窃取到用户设计的明文信息，FPGA 已沦为攻击者用来解密的工具。

图 8 窃取码流位置

Fig.8 Position of stolen bitstream

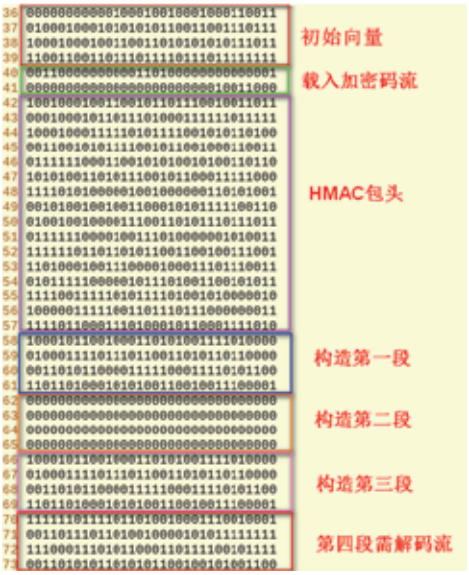


图 9 构造的攻击码流部分

Fig.9 The part of created attack bitstream

3.2 阻止攻击者窃取码流

采用鉴权控制回读策略，仿真结果如图 14，加载与本章 3.1 小节相同的攻击码流，此时 WBSTAR 寄存器的值为 0x30008001，开始 fallback 进程，电路清零完毕后载入回读码流，图 15 中 WBSTAR_R_EN 信号未拉高，回读总线上也未载入数据，TDO 端口没有读出数据，实现无法窃取原码的目的。

3.3 Multiboot 功能验证

除需验证安全方案可有效解决漏洞外，还需验证是否会影响用户正常使用 Multiboot 功能。采取的仿真方案为：加载非加密码流，码流中 IDcode 发生错误，触发 Fallback，通过 Multiboot 功能载入正确码流，观察仿真是否配置成功，且配置结束后 WBSTAR 寄存器是否可回读出值。为了方便观测读出 WBSTAR 寄存器的值，在 Vivado 中设置了 WBSTAR 寄存器的值为 0x11223344。图 16 为加载 IDcode 错误码流，ID ERROR 标志位拉高，触发 Fallback，图 17 为载入正确码流，配置结束，DONE 拉高，完成启动时序，图 18 为载入回读 WBSTAR 码流，WBSTAR_R_EN 使能信号拉高，回读总线也载入数据，回读数据正常从 TDO 移出，如图 19，所以修改方案未影响用户非加密码流 Multiboot 功能。

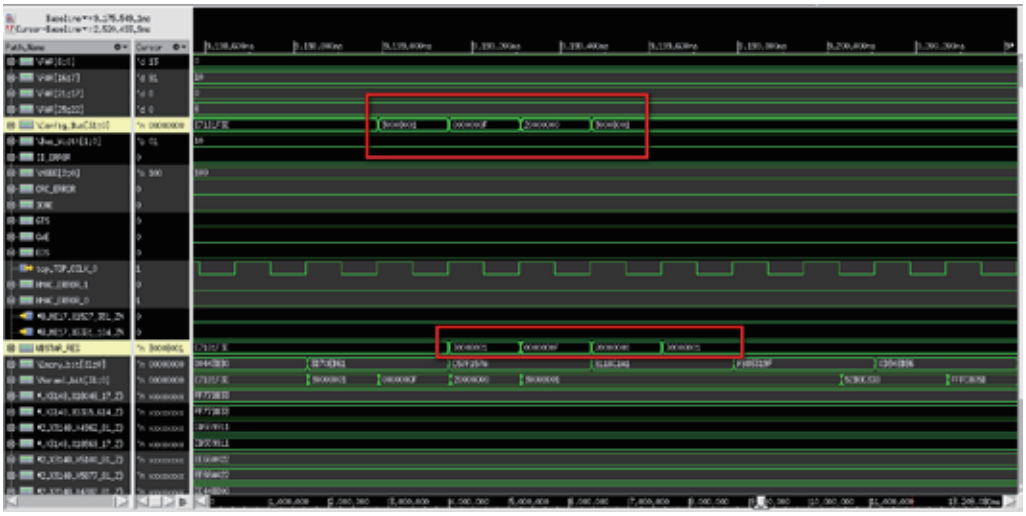


图 10 解密字送入 WBSTAR 寄存器

Fig.10 WBSTAR register of decrypted data

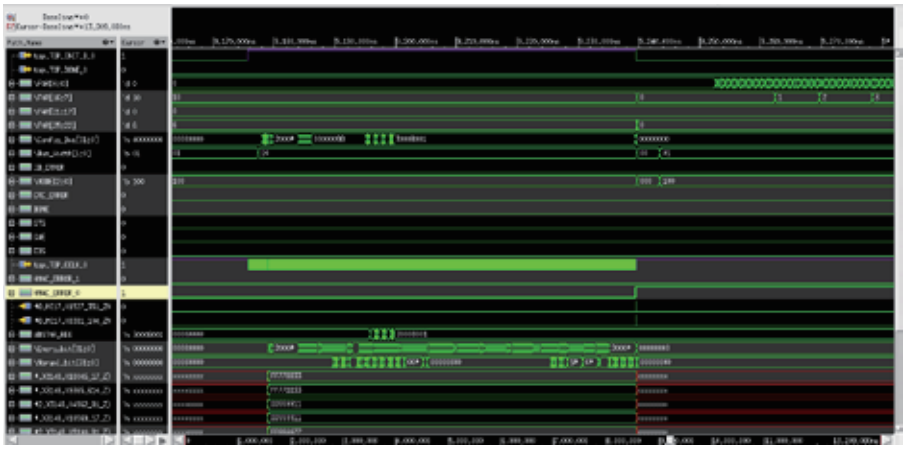


图 11 fallback 进程

Fig.11 Fallback process

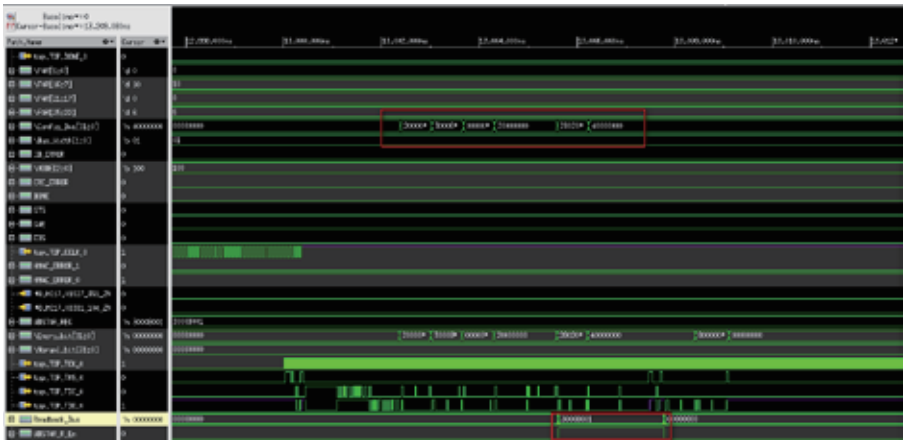


图 12 读出数据

Fig.12 Readout data

```
00110000000000001000000000000001
00110000000000001000000000000001
~
```

图 13 JTAG TDO 读出数据

Fig.13 Data from JTAG TDO

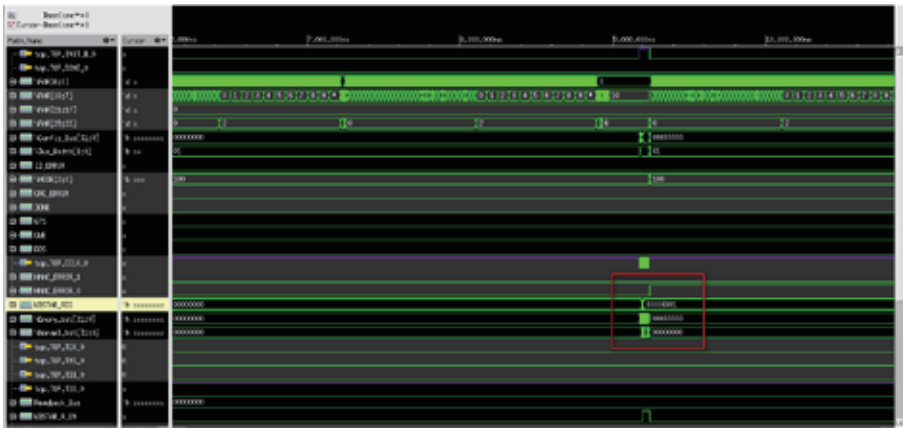


图 14 载入攻击码流

Fig.14 Downloading attack bitstream

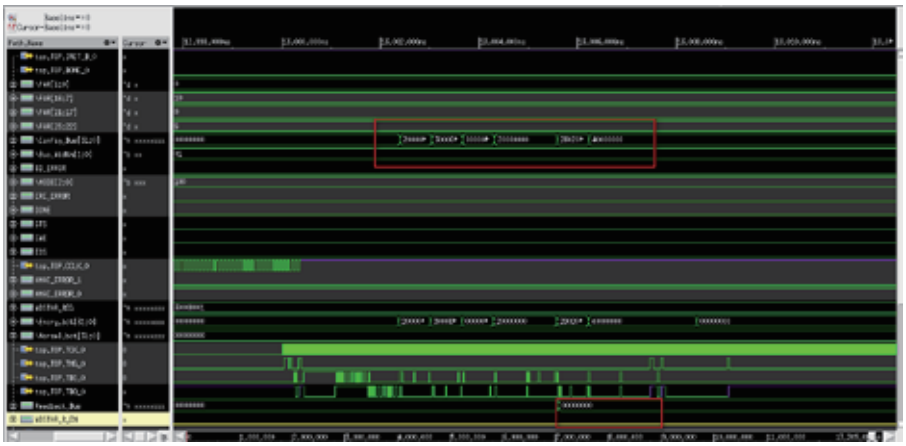


图 15 未读出数据
Fig.15 Failure of readout data

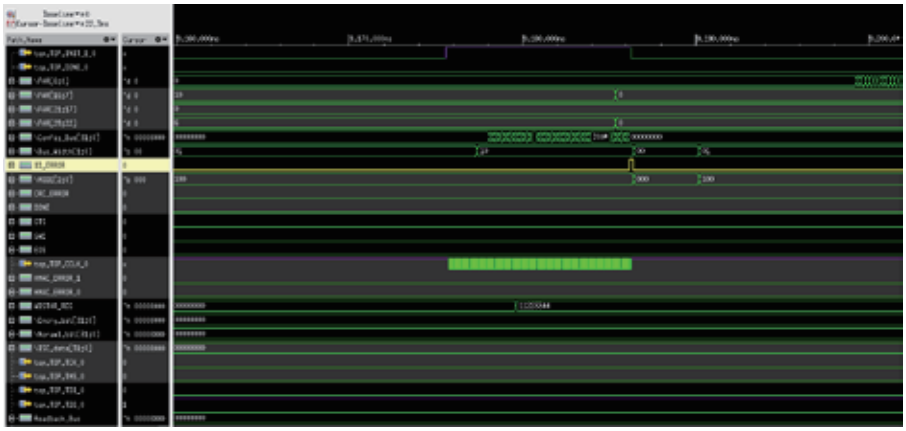


图 16 触发 Fallback
Fig.16 Triggering Fallback

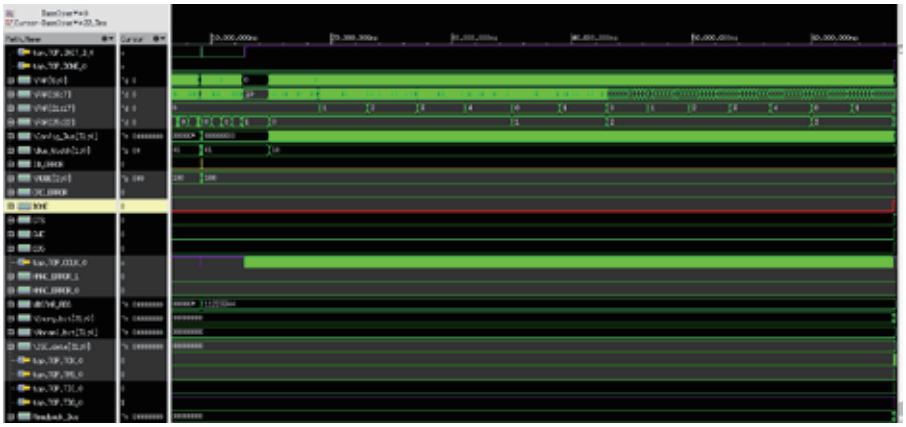


图 17 DONE 信号拉高
Fig.17 Pullup the signal of DONE

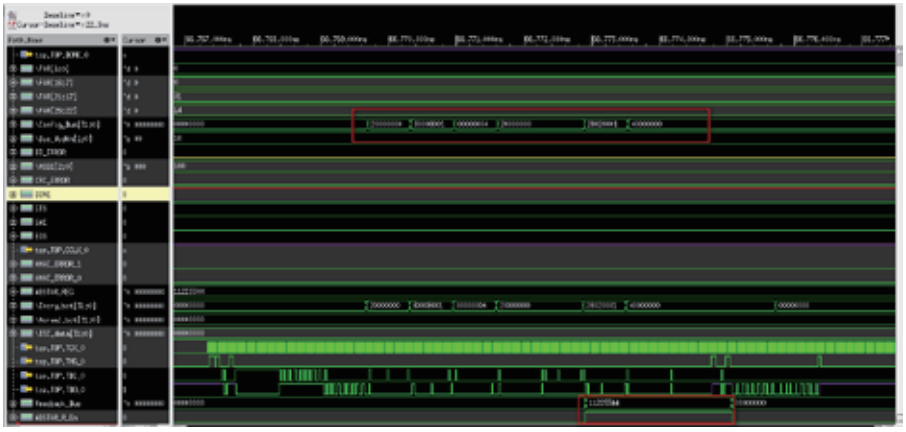


图 18 回读数据正常读出

Fig.18 Readout data normally

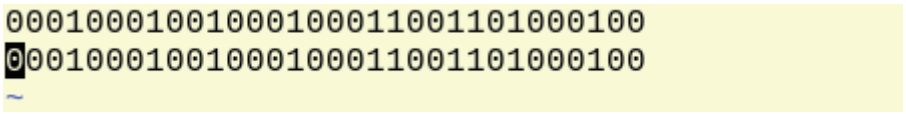


图 19 JTAG TDO 端口回读的数据

Fig.19 The data from JTAG TDO

从上述 3 部分仿真验证可知，Xilinx 7 系列 FPGA 确实存在 StarBleed 漏洞，采用鉴权控制回读策略的电路不仅可以解决 StarBleed 漏洞，同时不会影响用户正常使用 Multiboot 功能。

4 板级实测

为了验证修改后的电路已不存在 StarBleed 漏洞，在 BQ7V 和 BQ7K 芯片回片后，本文进行了 BQ7V、BQ7K 系列芯片与 XQ7K325T 的码流安全对比测试验证。

（1）按照本文第三章的方法设计攻击码流，即通过修改加密码流，将加密码流载入 FPGA 进行解密，攻击者想要窃取的原码字将被写入到 WBSTAR 寄存器中。实测以窃取码流中选定的一个字为例（本次实测选择的是 0x30008001−0x00000000− 0x20000000−0x30008001 这一行的最后一个字 0x30008001），为了说明该字是被窃取到，而不是 WBSTAR 已存的值，在 Vivado 码流选项中，将 WBSTAR 配置为 0x11223344，并且通过回读读出了 WBSTAR 的值为 0x11223344，如图 20 所示。

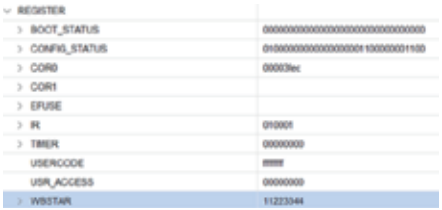


图 20 回读 WBSTAR 寄存器的值

Fig.20 Readout the data from WBATSR register

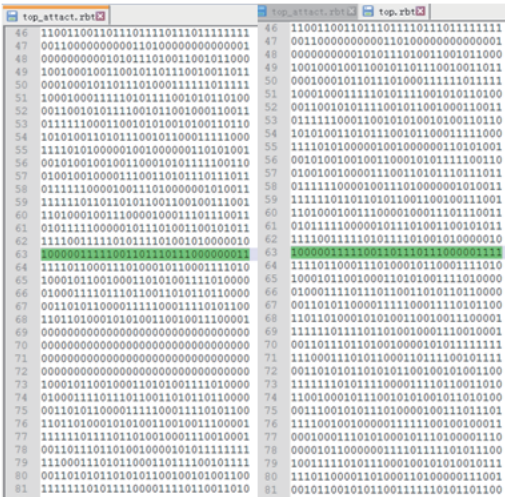


图 21 攻击码流（left）和原码流（right）

Fig.21 Attack birstream (left) and normal bitstream (right)

（2）使用 Vivado 软件对 XQ7K325T 载入密钥以及攻击码流如图 21，之后发生 Fallback 复位 FPGA，再通过软件回读 WBSTAR 寄存器，如图 22 所示，此时 FPGA 记录发生过 HMAC 错误，且 WBSTAR 的值为 0x30008001，表明窃取 FPGA 数据成功。

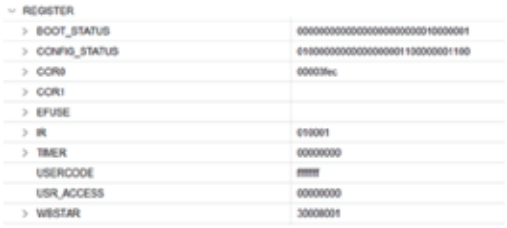


图 22 窃取数据成功，存在 StarBleed 漏洞

Fig.22 Stealing data successfully and exiting StarBleed vulnerability

（3）验证 BQ7V、BQ7K 系列芯片中 WBSTAR 寄存器的可读可写特性，即通过 JTAG 先载入 WBSTAR 写命令与数值（0x00112233），再载入 WBSTAR 回读命令，读出 WBSTAR 值（0x00112233），表明 WBSTAR 寄存器的可读可写特性并未受到影响。

（4）使用 Vivado 软件对 BQ7V、BQ7K 系列芯片载入密钥以及攻击码流，之后发生 Fallback 复位 FPGA，再通过软件回读 WBSTAR 寄存器，如图 23 所示，此时 FPGA 记录发生过 HMAC 错误，但 WBSTAR 的值为 0x00000000，表明窃取 FPGA 数据失败，电路中已不存在 StarBleed 漏洞。

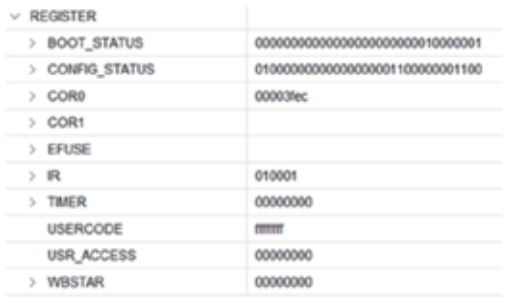


图 23 窃取数据失败，无 StarBleed 漏洞

Fig.23 Stealing data unsuccessfully and non-existent StarBleed vulnerability

（5）验证 BQ7V、BQ7K 系列芯片的 Multiboot 功能，即进行了 SPI 非加密码流与加密码流 Multiboot 功能的实测，实测结果可通过 Multiboot 引导 Update 码流，如图 24 所示，通过 Multiboot 加载 SPI 的跑马灯码流正常运行，表明 BQ7V、BQ7K 系列芯片的 Multiboot 功能正常运行并未受到影响。

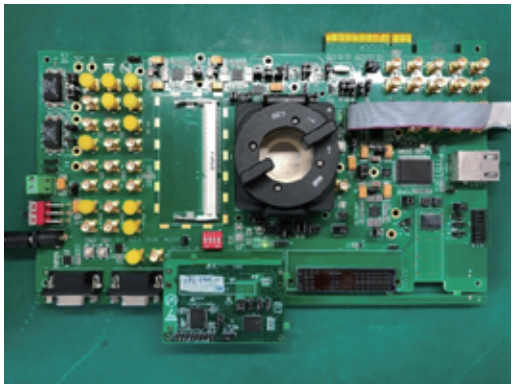


图 24 Multiboot 功能正常

Fig.24 Achieving Multiboot function

通过上述 5 方面的板级回片测试，验证了鉴权失败禁止回读安全方案的正确性，同时也证实了经过电路硬件防护设计，BQ7V、BQ7K 系列芯片已解决了 StarBleed 漏洞，同时未影响 FPGA 的 Multiboot 功能。

5 结论

随着 FPGA 在数据中心、通信基础设施、AI 加速、医疗设备、边缘计算等多个领域的广泛使用，对 FPGA 安全性的研究已经成为了一个热点。本次曝出 StarBleed 漏洞危害严重，它已经将 FPGA 沦为一个解密的工具。本文针对这一漏洞，分析了其工作机理，通过仿真验证证实了它存在于 Xilinx 7 系列 FPGA 中；随后基于 WBSTAR 寄存器与 FPGA 鉴权的特点提出了鉴权失败禁止回读的安全解决方案，通过仿真验证与回片测试表明 BQ7V、BQ7K 系列芯片已不存在 StarBleed 漏洞，同时未影响 FPGA 的 Multiboot 等特性，提升了 FPGA 的安全性，保障了用户的设计安全与知识产权。StarBleed 漏洞已经给业界敲响了警钟，这将提升人们对 FPGA 安全性的重视，并重新考虑未来 FPGA 安全性的设计。

参考文献（References）

[1] DORSCH J. 现场可编程门阵列 FPGA 芯片及其应用[J]. 集成电路应用，2018，35(1):77-79.

[2] ENDER M，MORADI A，PAAR C. The Unpatchable Silicon：A Full Break of the Birstream Encryption of Xilinx 7-Series FPGAs[C]. 29th USENIX Security Symposium. Boston：USENIX Association. 2020:1803-1819.

[3] 揭应平 . FPGA 芯片设计及其应用分析[J]. 集成电路应用，2017，34(12):37-41.

[4] Xilinx Inc. 7 Series FPGAs Configuration User Guide[Z]. UG470(v1.13.1),August 20,2018.

[5] Xilinx Inc. Using Encryption and Authentication to Secure an UltraScale /UltraScale+ FPGA Bitstream[Z]. XAPP1267(v1.3)，October 12,2018.

[6] WANG P，ZHANG Y M，YANG J. Research and Design of AES Security Processor Model Based on FPGA [J]. Procedia Computer Science，2018,6:249-254.

[7] Xilinx Inc. Internal Programming of BBRAM and eFuses Application Note [Z]. XAPP1283(v1.2)，July 31,2020.

[8] Xilinx Inc. Developing Tamper Resistant Designs with Xilinx Virtex-6 and 7 Series FPGAs[Z]. XAPP1084(v1.4)，June 13,2017.

[9] Xilinx Inc. Using Encryption to Secure a 7 Series FPGAs Bitstream[Z]. XAPP1239(v1.1)，July 16,2018.

[10] Xilinx Inc. UltraScale Architecture Configuration User Guide[Z]. UG570(v1.13)，July 28,2020.

[11] Xilinx Inc. MultiBoot with 7 Series FPGAs and SPI[Z]. XAPP1247(v1.1)，February 28,2017.

作者简介：

杨佳奇（1992—），男，山西省大同市人，硕士，工程师，目前从事超大规模FPGA设计及验证。

用于 TMR 的低噪声斩波仪表放大器

张文博，陈伟平，尹 亮
（哈尔滨工业大学，黑龙江省 哈尔滨市 150000）

摘 要：本文提出了一款用于隧穿式磁阻传感器（TMR）的低噪声单端输出 CMOS 斩波仪表放大器。该仪表运放具有低噪声低功耗高带宽等优点，可支持从几赫兹到几十万赫兹频率范围内微弱磁场信号的检测，适用于微弱磁场信号的测量。为了得到更高的直流增益和线性度，仪表放大器采用三级密勒补偿搭配电流反馈结构。为了得到更高的共模抑制比（CMRR）和电源抑制比(PSRR)，运放的前两极采用全差分结构。为了提高运放稳定性，降低功耗，运放采用跨导电容前馈补偿拓扑结构。仪表运放采用斩波技术和连续时间纹波抑制回路降低闪烁噪声并抑制高频纹波。芯片采用 0.35μm BCD 工艺加工制造，总面积 1mm^2。测试表明，该款仪表放大器在 1Hz 处等效输入噪声为 11nV/Hz$^{1/2}$。在放大倍数 65 倍，负载电容 20pF 时，该款仪表放大器带宽大于 50kHz。5V 电源电压下，总功耗为 300μA。

关键词：仪表放大器；斩波技术；隧穿式磁阻传感器
中图分类号：TN432 **文献标识码**：A

A Low Noise CMOS Instrumentation Amplifier for TMR-effect-based Magnetic Sensors

Zhang Wenbo, Chen Weiping, Yin Liang
(Harbin Institute of Technology, Harbin, 150000, China)

Abstract: This paper presents a low 1/*f* noise CMOS single-ended output instrumentation amplifier (IA) for tunneling magnetic resistance (TMR) sensors. The instrumentation amplifier has the advantages of low noise, low power consumption and high bandwidth, and can support the detection of weak magnetic field signals in the frequency range from a few hertz to hundreds of thousands of hertz, which is suitable for measurement of weak magnetic field signals. For high DC gain and linearity, the amplifier adopts three-stage current-feedback topology. For high CMRR and PSRR, the first two stage adopts fully differential input. To maintain stability and lower the power dissipation, the amplifier adopts trans-conductance with capacitance feedback compensation (TCFC) topology. The amplifier uses chopping technology and continuous-time AC-coupled ripple reduction loop to reduce 1/*f* noise and chopping ripple. The whole chip is fabricated using 0.35μm CMOS-BCD technology and the total area is 1mm^2. Test result shows an input-referred noise power spectral density (PSD) of 11nV/Hz$^{1/2}$ is achieved with 1Hz 1/*f* corner. A bandwidth larger than 50kHz(65x magnification) with 20pF load capacitor. The total current is 300μA at 5V supply.

Key words: instrumentation amplifier; chopping technology; tunneling magnetic resistance sensors

0 引言

TMR 传感器是一种以隧穿效应为原理加工出来的磁阻传感器。由于隧穿式磁阻采用隧道结构，相比于其他结构的磁阻传感器，TMR 传感器具有电阻率高，能耗小，稳定性强等优点[1]。TMR 传感器广泛应用于军事和民用领域，如监控地球和空间磁场的变化[2]，检测磁性生物信号[3]，测量高自旋弹丸的侧倾角等[4]。这些应用的带宽范围在几赫兹到几十万赫兹之间。所以，为满足 TMR 应用需求，所设计的前级检测电路既要具有极小的低频噪声又要具备较大的带宽。目前 TMR 传感器前端检测通常采用如 AD620、AD623 等商用仪表运放，这些仪表运放通常仅对低

频噪声、功耗、带宽等单项指标进行优化，难以实现 TMR 应用范围内各性能指标的折中。本文结合 TMR 磁阻传感器的工作特点，针对性的设计出了一款低噪声斩波仪表放大器，实现了其应用领域内各项指标的综合优化。

本文设计了一款电流反馈式斩波仪表运算放大器 [5,6]。仪表运算放大器的主体电路采用跨导电容前馈补偿（TCFC）拓扑结构。该补偿结构可使运放在相同的功耗下获得更大的带宽 [7]。仪表运放采用斩波技术消除低频闪烁噪声，并采用连续时间纹波抑制回路消除斩波带来的高频纹波的影响。芯片采用 0.35μm BCD 工艺加工制造。该款仪表放大器在 1Hz 处等效输入噪声为 11nV/Hz$^{1/2}$。在放大倍数 65 倍，负载电容 20pF 时，该款仪表放大器带宽大于 50kHz。纹波抑制回路可以将斩波带来的纹波抑制到忽略不计。

1 斩波仪表放大器拓扑结构

仪表放大器整体结构如图 1 所示。仪表运放整体结构由电流反馈斩波仪表放大器和纹波抑制回路组成。G_{m1} 到 G_{m4}，R_1 到 R_2 形成电流反馈仪表放大器主体结构。G_{mt}，G_{mf}，C_1，C_2 构成的密勒补偿回路用以确保仪表运放稳定性。CH_1–CH_3 为斩波开关用以消除闪烁噪声影响。G_{m5}，G_{m6}，C_3，C_4，CH_4 构成连续时间纹波抑制回路用以消除斩波带来的高频纹波。

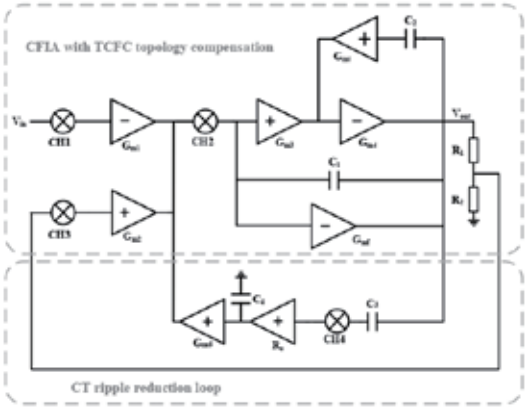


图 1 斩波仪表放大器拓扑结构

Fig.1 Topology of the proposed IA

1.1 电流反馈斩波仪表运算放大器

1.1.1 结构选择

隧穿式磁阻的阻值与磁场的方向和大小有关，根据这一特点，可以利用惠斯通电桥结构将磁场的变化转换为电压的变化并输出。仪表放大器用于检测并放大隧穿式磁阻传感器输出的电压信号。仪表放大器主要有三种结构：三运放结构、电容耦合结构和电流反馈结构 [8]。三运放结构功耗高且对电阻匹配度要求较高。电容耦合结构很难适应低频应用，因为设计中为消除低频噪声往往会采用斩波技术，该结构下采用斩波技术会严重降低输入阻抗。电流反馈仪表放大器输入阻抗可以做到很高。并且输入管容易匹配，共模抑制比更高。由于隧穿式磁阻电阻率高，为实现更好的阻抗匹配，仪表放大器的等效阻抗需要很高，因此本设计采取电流反馈式仪表放大器结构。

电流反馈仪表放大器结构已由图 1 给出。该仪表运放闭环增益为：

$$Gain = \frac{G_{m1}(R_1 + R_2)}{G_{m2}R_2} \tag{1}$$

输入管精确匹配后，仪表运放的闭环增益仅由电阻比例决定。

1.1.2 仪表放大器主体运放结构

电流反馈仪表运算放大器核心运放结构是采用 TCFC 补偿方式的三级运放。在相同的带宽要求下，该结构所需功耗较其他拓扑结构更低。图 2 展示了 TCFC 三级运放的晶体管级电路和等效小信号分析模型。根据小信号模型可得出运放传递函数为：

$$\frac{V_{out}}{V_{in}}(s) = \frac{NUM(s)}{DEM(s)} \tag{2}$$

其中：

$$\begin{aligned} NUM(s) = {} & s^3 G_{m1}C_1C_2C_p + s^2 G_{m1}C_2(G_{my}C_1 + G_{mf}C_2) \\ & - sG_{m1}(G_{m3}G_{m4}C_2 + G_{mt}G_{mf}C_p) \\ & - G_{m1}G_{m3}G_{mt}G_{m4} \end{aligned} \tag{3}$$

$$DEM(s)=s^4C_1C_2C_pC_L+s^3C_1C_p(G_{mt}C_L+G_{mt}C_2+G_{mf}C_2)+s^2C_1(G_{m3}G_{m4}C_2+G_{mt}G_{m4}C_2+G_{mt}G_{mf}C_p)+sC_1G_{m3}G_{mt}G_{m4} \quad (4)$$

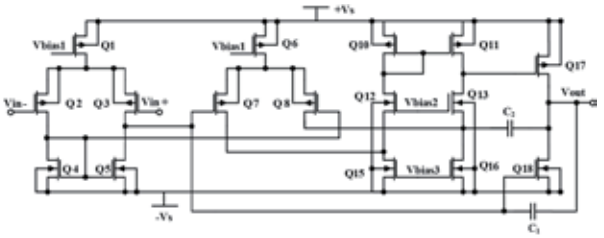
式中，C_p 为第二级输出等效电容，C_L 为负载电容，G_{m1}，G_{m3}，G_{m4}，G_{mt}，G_{mf} 分别是 Q_3，Q_8，Q_{17}，Q_{13}，Q_{18} 的等效跨导。根据式（3）和式（4），可以计算出系统的四个极点和三个零点。假定 $G_{m4} \approx G_{m4}$，可解得零极点为：

$$\begin{aligned}
\omega_{ug} &= -\frac{G_{m1}}{C_1} \\
\omega_{pd} &= -\frac{1}{C_1R_{o1}A_{2,3}} \\
\omega_{p1} &= -\frac{G_{m3}G_{mt}}{G_{m3}C_2+G_{mt}C_2+G_{mt}C_p} \\
\omega_{p2} &= -\frac{G_{mt}C_L+G_{mt}C_2+G_{m4}C_2}{C_2C_L} \\
\omega_{p3} &= -\frac{G_{m4}(G_{m3}C_2+G_{mt}C_2+G_{mt}C_p)}{C_p(G_{mt}C_L+G_{mt}C_2+G_{m4}C_2)} \\
\omega_{z1} &= -\frac{G_{m3}G_{mt}}{G_{m3}C_2+G_{mt}C_p} \\
\omega_{z2} &= -\frac{G_{mt}C_1+G_{m4}C_2}{C_1C_2} \\
\omega_{z3} &= \frac{G_{m4}(G_{m4}C_2+G_{mt}C_p)}{G_{mt}C_1C_p+G_{m4}C_2C_p}
\end{aligned} \quad (5)$$

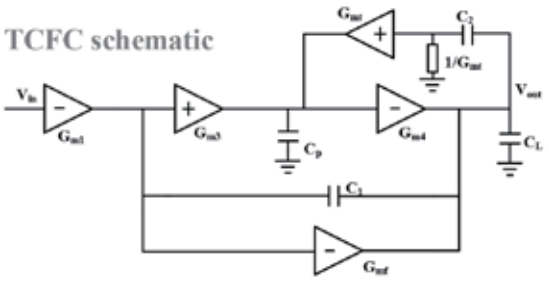
其中 ω_{ug} 为运放的增益带宽积，Ro_1 是第一级输出阻抗，ω_{pd} 为运放主极点 $A_{2,3}$ 运放二三级直流增益。非主极点 ω_{p1}，ω_{p2} 可以分别与零点 ω_{z1}，ω_{z2} 抵消，第三个零点 ω_{z3} 位于 s 平面右半部分，但由于寄生电容 C_p 远小于 C_2，ω_{z3} 将远大于增益带宽积，其影响可以忽略不计。因此非主极点 ω_{p3} 是我们所需考虑的对象。当 C_2g_{m17} 远大于 C_pg_{m13} 时，可保证该非主极点远大于增益带宽积。因此 TCFC 补偿方式很容易实现运放的稳定。

1.2 连续时间纹波抑制回路

CMOS 电路不可避免的会受到闪烁噪声影响。因此，为提高其检测精度需要用到低频噪声消除技术。常用的低频噪声消除技术有自动调零和高频斩波两种。由于自动调零电路是采样系统，会不可避免的遭受噪声混叠的影响。因此斩波技术为噪声抑制首选。然而斩波会带来高频纹波，严重时将影响电路工作。因此必须设计纹波抑制回路消除其影响。



(a) 运放晶体管级电路
(a) Operational amplifier transistor-level circuit



(b) TCFC 三级运放小信号模型
(b) TCFC three-stage operational amplifier small-signal module

图 2 TCFC 三级运放晶体管级电路及等效小信号模型
Fig.2 Transistor circuit of the amplifier and its equivalent diagram

高频纹波产生及抑制机理如下。如图 3 所示，第一级运放 G_{m1} 的直流失调电压 V_{os} 经过 G_{m1} 后被 CH_2 调制到高频，形成高频方波 I_{chop}。C_1、G_{m3} 和 G_{m4} 等效为积分电路，I_{chop} 通过积分电路耦合到输出，产生输出高频纹波 $V_{out,ripple}$。如果没有纹波抑制回路，纹波的开环及闭环幅度可以表示为：

$$V_{out,ripple,open}=\frac{G_{m1}V_{os}}{2C_1f_{ch}} \quad (6)$$

$$V_{out,ripple,close}=\frac{V_{os}}{R_2/(R_1+R_2)+2C_1f_{ch}/G_{m1}} \tag{7}$$

由式（6）可知，在小倍数放大时，输出纹波大小正比于失调与放大倍数的乘积，在大倍数放大时，输出纹波大小由开环参数 G_{m1}，f_{ch}，C_1 决定。

当纹波抑制回路工作时，高频纹波 $V_{out,ripple}$ 通过抑制回路中的电容 C_3 转换为高频类方波电流 I_{AC}，再被斩波器 CH_4 解调回低频，形成低频电流信号 I_{DC}，此信号经过 G_{m5} 和 C_{int} 组成的积分器后转换为补偿电压 V_o。补偿电压 V_o 经过 G_{m6} 后形成补偿电流，此电流用以补偿失调在 A 点形成的失调电流，以此来达到抑制高频输出纹波的目的。可见，整个电路中，纹波抑制回路相当于一个带通滤波器，它必须只让高频纹波信号通过以形成失调补偿电流，而阻止低频输出信号的通过以避免影响电路正常的功能。

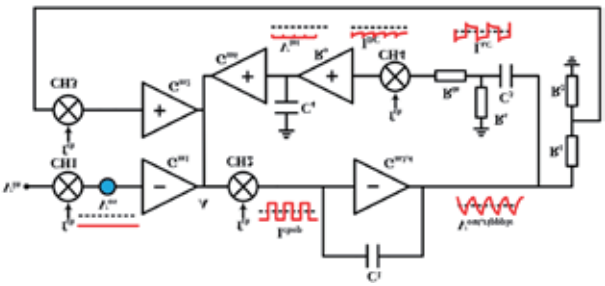


图 3 纹波抑制回路工作原理
Fig.3 Schematic of RRL

图 3 中电阻 R_s 和 R_{in} 的大小由纹波抑制回路的实际电路得到。晶体管级电路如图 4 所示。纹波抑制回路的开环增益可以表示为：

$$\begin{aligned}L(s)&=\frac{C_3R_sR_oG_{m6}}{(sC_3R_sR_{in}+R_s+R_{in})(sC_4R_o+1)C_1}\\&+\frac{R_2G_{m2}}{sC_1(R_1+R_2)}\\&\approx\frac{C_3R_sR_oG_{m6}}{(sC_3R_sR_{in}+R_s+R_{in})(sC_4R_o+1)C_1}\end{aligned} \tag{8}$$

式（8）中的约等号基于 $G_{m6}R_o$ 远大于仪表运算放大器闭环放大倍数的事实。闭环后纹波幅度可以表示为：

$$\begin{aligned}V_{out,ripple,RRL}&=\frac{V_{out,ripple,open}}{1+L(0)}\\&=\frac{G_{m1}V_{os}(R_s+R_{in})}{2C_1f_{ch}(R_s+R_{in})+2C_3R_sR_oG_{m6}f_{ch}}\\&\approx\frac{G_{m1}V_{os}}{2G_{m6}R_oC_3f_{ch}}\end{aligned} \tag{9}$$

式（9）中的约等号是基于 R_{in} 远小于 R_s 的事实。从式中可以看出，纹波幅度将被抑制为原来的 $C_3/G_{m6}R_oC_1$。由于 $G_{m6}R_o$ 很容易达到 80dB，输出纹波将被抑制到毫伏级别。根据式（8），我们可以推导出纹波抑制回路的主极点 $\omega_{d,RRL}=1/C_4R_o$ 和非主极点 $\omega_{nd,RRL}=1/C_3R_{in}$。主极点与直流增益的乘积为纹波抑制回路的增益带宽积：

$$\omega_{ug,RRL}=\omega_{d,RRL}L(0)\approx\frac{G_{m6}C_3}{C_1C_4} \tag{10}$$

式（10）决定了纹波抑制回路的收敛速度，通常情况下增加回路收敛速度会将纹波抑制回路中更多的噪声等效到输入。因此设计时需对精度与速度进行折中处理。

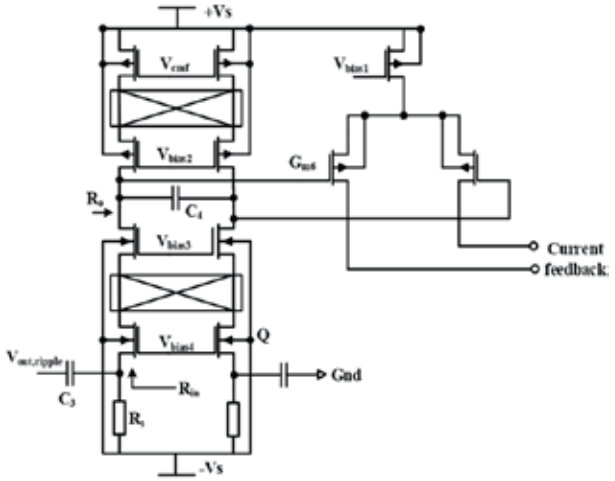


图 4 纹波抑制回路晶体管级实现
Fig.4 Transistor circuit of RRL

2 电路及测试

2.1 电路实现与纹波抑制回路仿真结果

所设计的仪表运算放大器晶体管级电路如图5所示。鉴于芯片设计时，纹波抑制回路上电就开始工作，且纹波被抑制到噪声之下，难以观测到纹波抑制回路工作过程。这里以仿真结果验证上电时纹波抑制回路的工作过程。加入10mV失调电压，斩波频率为150kHz，仿真结果如图6所示。结果表明上电后输出纹波在100μs之内被有效抑制。

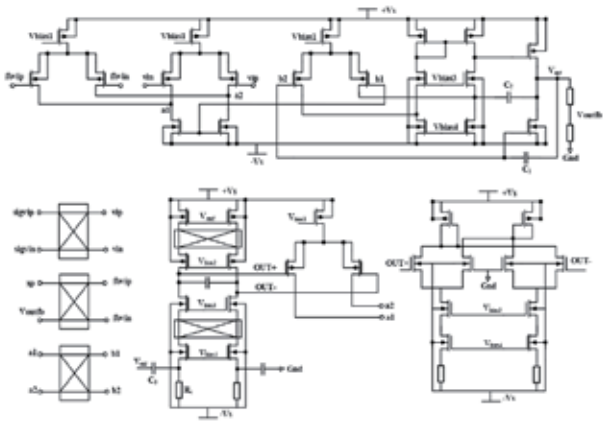


图5 低噪声斩波仪表放大器晶体管级实现
Fig.5 Circuit realization of the proposed IA

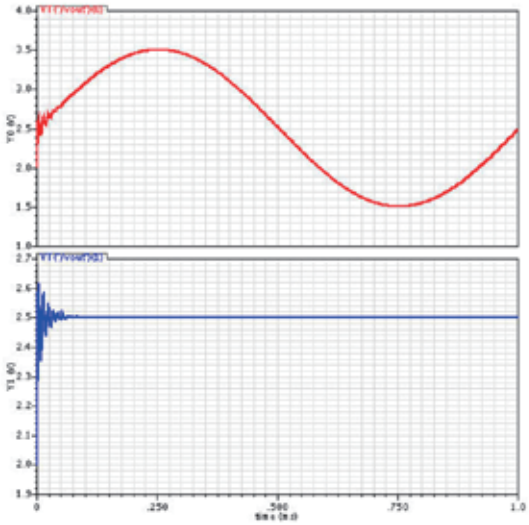


图6 纹波抑制回路仿真结果
Fig.6 Simulation results of RRL function

2.2 芯片照片及测试结果

芯片采用0.35μm BCD工艺加工制造，总面积约为1mm^2。芯片照片如图7所示。仪表运放放大倍数调至65倍时，电路差分信号幅频特性、共模信号幅频特性、正负电源抑制幅频特性与噪声频谱实测结果如图8到图12所示。测试结果表明在该条件下仪表运放带宽仍大于50kHz。共模抑制比为131dB（差模增益与共模增益之比），正负电源抑制比分别为104dB与107dB（差模增益与电源增益之比）。1Hz处输出噪声谱密度为624nV/Hz$^{1/2}$等效输入噪声为9.6nV/Hz$^{1/2}$（50Hz，150Hz处谱线为工频干扰）。等效输入失调8μV，5V电源电压下功耗为300μA。图13与图14分别为纹波抑制回路关断和开启后仪表运放输出波形，实测表明输出纹波得到了有效抑制。表1总结了本款芯片与常用商用仪表放大器性能对比图。对照表明，在TMR磁阻测量常用工作频率范围（几赫兹到几十万赫兹之间），本款芯片在功耗、低频噪声、带宽等方面更具优势。

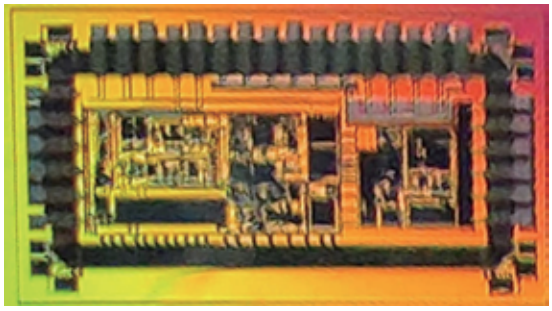


图7 仪表放大器芯片照片
Fig.7 The photograph of the chip

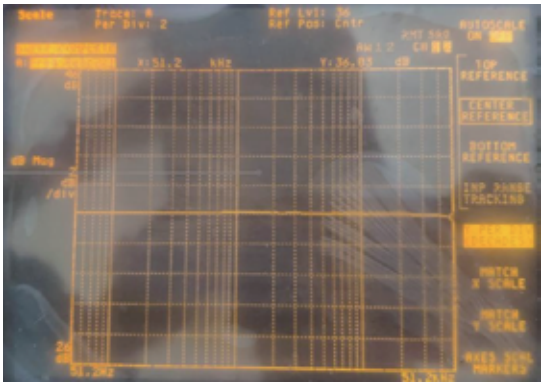


图8 仪表放大器带宽测试图
Fig.8 Frequency response of the chip

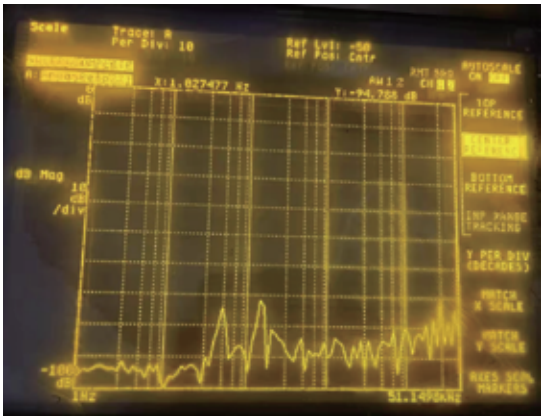


图 9 共模抑制比测试图
Fig.9 CMRR frequency response of the chip

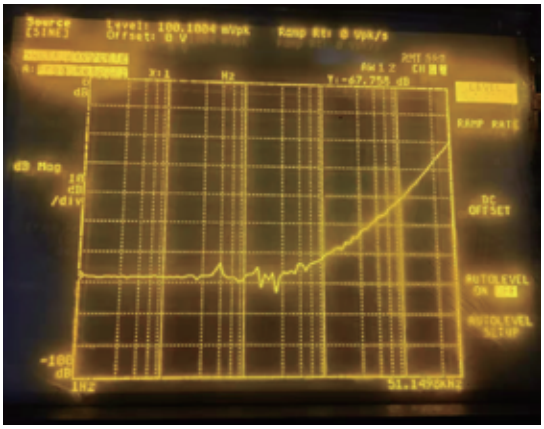


图 10 正向电源抑制比测试图
Fig.10 Positive PSRR frequency response

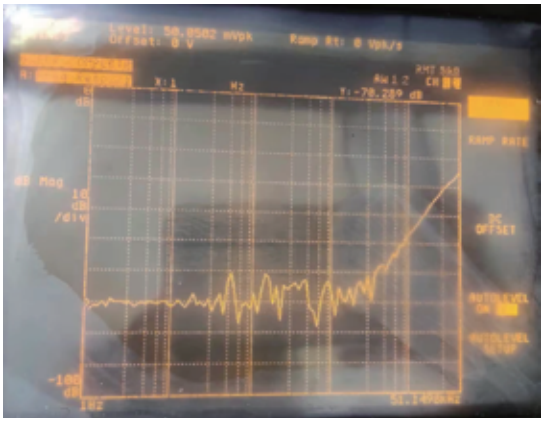


图 11 负向电源抑制比测试图
Fig.11 Negative PSRR frequency response

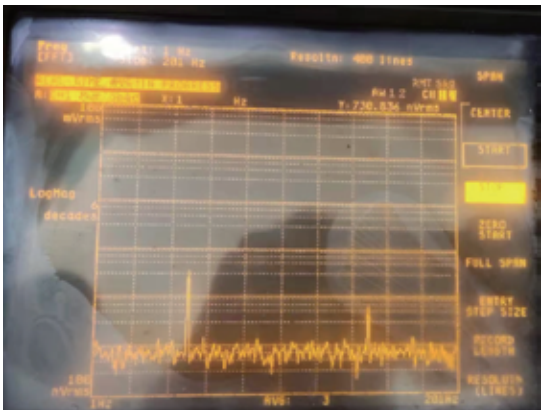


图 12 仪表放大器噪声测试图
Fig.12 Output noise power spectral density

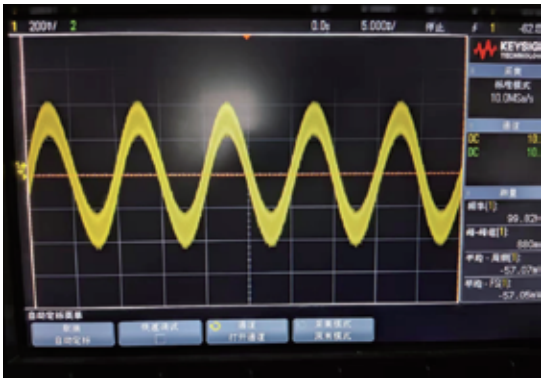


图 13 纹波抑制回路关断后的瞬态输出
Fig.13 Transient output without RRL loop

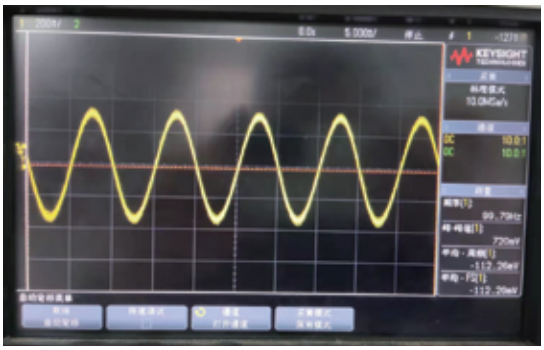


图 14 纹波抑制回路打开后的瞬态输出
Fig.14 Transient output with RRL loop

表 1 常用仪表运放性能对比
Tab.1 Performance comparison of instrumental amplifiers

指标	本文	INA821[9]	AD623[10]	AD620[11]
输入失调（μV）	8	10	200	50
等效输入噪声（nV/Hz$^{1/2}$@1Hz）	11	15	130	20

共模抑制比（dB）	130	125	120	130
电源抑制比（dB）	100	130	120	140
增益带宽积（Hz）	4M	4.7M	1M	12M
静态电流（mA）	0.3	0.65	0.3	0.9

3 结论

本文提出了一种应用于TMR磁阻传感器的CMOS斩波仪表运算放大器，并详细介绍了其中相位补偿回路和纹波抑制回路的工作原理。两路反馈回路的加入降低了仪表运算放大器的功耗并解决了斩波带来的纹波问题。整款芯片采用0.35μm BCD工艺加工制造，5V电源电压供电下，静态电流小于300μA。斩波技术使得仪表运放1Hz处噪声仅为11nV/$Hz^{1/2}$，失调仅为8μV。相比于其他商用仪表运放，本款芯片在较小的功耗下实现了更低的低频噪声和较大的增益带宽积，更加适用于微弱磁场信号测量领域的应用。

参考文献（References）

[1] RIPKA P, JANOSEK M. Advances in magnetic field sensors [J]. IEEE Sensors Journal, 2010, 10(6): 1108-1116.

[2] LI X, HU J, CHEN W, YIN L, et al. A Novel High-Precision Digital Tunneling Magnetic Resistance-Type Sensor for the Nanosatellites Space Application [J]. Micromachines, 2018, 9 (3).

[3] WANG M, WANG Y, PENG L, et al. Measurement of Triaxial Magnetocardiography Using High Sensitivity Tunnel Magnetoresistance Sensor [J]. IEEE Sensors Journal, 2019, 19(21): 9610-9615.

[4] CAO P, WANG X M, JIA F X, et al. Circuit Design and System Error Analysis Based on MR/GPS Combination Measuring Projectile Roll Angle [C]// Proceedings of Instruments, Measurement, Electronics and Information Engineering. 2013:1059-1062.

[5] MAHMOUD S A, ALHAMMADI A A. Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization [J]. Proceedings of the IEEE. 1996, 84(11): 1584-1614.

[6] WU R, MAKINWA K A A, HUIJSING J K. A chopper current-feedback instrumentation amplifier with a 1mHz 1/f noise corner and an AC-coupled ripple-reduction loop [J]. IEEE Sensors Journal, 2009, 44(12): 3232-3243.

[7] PENG X, SASEN W. Transconductance with capacitances feedback compensation for multi-stage amplifiers[J]. IEEE J. Solid-State Circuits, 2005, 40(7): 1514-1520.

[8] DOOL B J, HUIJSING J K. Indirect current feedback instrumentation amplifier with a common-mode input range that includes the negative roll[J]. IEEE Journal of Solid-State Circuits,1993,28(7): 743-749.

[9] Texas instruments. INA821 data sheet, 2018, https://www.ti.com/product/INA821.

[10] Analog Devices Inc. AD623 data sheet, 1999, https://www.analog.com/en/products/ad623.html.

[11] Analog Devices Inc. AD620 data sheet, 2003, https://www.analog.com/en/products/ad620.html.

作者简介：
张文博（1990—），男，黑龙江哈尔滨人，研究生，博士研究生在读，研究方向为传感器接口芯片设计。

基于改进轻量化网络的空间非合作目标部件识别算法

郝 强，李 杰，王 路，张 曼

（上海航天电子技术研究所，上海市 201109）

摘　要：在使用深度学习方法进行空间非合作目标部件识别时，由于神经网络参数量大且嵌入式设备计算能力不足，难以将神经网络有效地部署在嵌入式平台上。针对该问题提出一种改进的轻量化目标检测网络，在保证检测精度的同时，有效降低网络参数量提升网络检测速度。提出的网络模型在 YOLOv3 的基础上借鉴深度可分离卷积的思想，引入 Bottleneck 模块降低模型参数量提升检测速度，同时引入 Res2Net 残差模块来增加模型的接受域尺度丰富性和结构深度，提高网络对于小目标的检测能力，设计了一个新的轻量化特征提取主干网络 Res2−MobileNet，并结合多尺度检测方法进行空间非合作目标部件识别。实验结果表明，相比于 YOLOv3，本模型在参数量上降低了 55.5%，检测速度由 34fps 提高到 65fps，同时对于小目标的检测效果也有显著提升。

关键词：目标识别；轻量化；YOLOv3；空间非合作目标

中图分类号：TP391　　**文献标识码**：A

Spatial Non-cooperative Target Components Recognition Algorithm Based on Improved Lightweight Network

Hao Qiang, Li Jie, Wang Lu, Zhang Man

(Shanghai Aerospace Electronics Technology Research Institute, Shanghai, 201109, China)

Abstract: Due to the large amount of neural network parameters and insufficient computing power of embedded devices, it is difficult to effectively deploy neural networks on embedded platforms when using deep learning methods to identify spatial non-cooperative target components. Aiming at this problem, an improved lightweight target detection network is proposed in this paper. On the basis of YOLOv3, we designed a new lightweight feature extraction backbone network Res2-MobileNet, drawing on the idea of Depth Separable Convolution, introducing the Bottleneck module to reduce the amount of model parameters to improve the detection speed, and introducing the Res2Net residual module to increase the sensitivity of network to small targets by increasing the model's receptive field scale richness and structural depth, and combines multi-scale detection methods to recognize spatial non-cooperative target components. The experimental results show that compared with the YOLOv3 model, the size of this model is reduced by 55.5%, the detection speed is increased from 34fps to 65fps, and the detection effect for small targets is also significantly improved.

Key words: target recognition; lightweight; YOLOv3; spatial non-cooperative target

0 引言

随着空间目标数量的不断增加和空间环境的日益复杂，空间态势感知成为太空安全防护与太空控制的重要基础[1]。空间目标识别是空间态势感知的重要组成部分。目前，目标数据的来源为地基和天基探测，包括光学和雷达设备等探测数据。空间目标识别就是利用这些数据的特征信息，对目标的身份、姿态、状态做出有效的判断和识别[2]。在保障太空行动和太空飞行安全等方面，对在役空间设备的维修保养需要对于其部件的精准识别。因此，对于空间非合作目标部件识别技术的研究具有重要价值。

对于空间非合作目标部件的识别方法主要分为两种：一是传统的目标识别算法，利用人工设定的识别目标的特征提取算子，如 Canny 边缘检测算子，Hough 直线检测算子等对目标进行特征提取[3]，传统算法精度低，对于空间多目标的识别需要同时运行多

个识别算法，泛化能力差，鲁棒性弱。二是基于深度学习的目标识别算法，利用标注好的图像训练卷积神经网络 (Convolutional Neural Networks，CNN)，通过减小损失函数值，得到最优的卷积核的权重。基于深度学习的方法可以克服传统方法的缺点，在识别的正确率和实时性方面具有良好表现 [4]。

目前，基于深度学习的光学图像目标检测方法主要有以下两种：两阶段 (two-stage) 法和一阶段 (one-stage) 法。两阶段检测方法以 R-CNN 系列 [5-7] 网络为代表，主要原理为首先对输入的图像进行区域划分，将得到的候选框分类，合并相同的类别，最终回归得到每个目标的候选框。两阶段法检测精度较高，但区域划分使得网络的检测速度变慢；一阶段检测法的典型代表有 SSD[8-10] 系列和 YOLO[11-13] 系列网络，其原理是通过神经网络直接回归出目标的坐标框、置信度和类别。虽然相较于一阶段法检测精度有所下降，但在检测速度方面取得了明显提升。YOLOv3 的出现使得在快速检测出目标的同时保证了检测精度。

然而，基于深度学习的检测算法对硬件的计算能力要求很高。在航天领域的应用中，由于功耗、体积、存储量等限制，现有嵌入式设备很难满足深度学习方法对于计算能力的要求。大多数的网络都难以在这类设备上进行有效部署，因此迫切需要设计轻量化的网络模型。针对这种情况，近些年有一系列用于图像分类的轻量化卷积神经网络被设计出来，例如 Xception[14]、MobileNet 系列 [15] 和 ShuffleNet[16] 系列。虽然这类模型可以降低网络的参数量和计算量，但其主要针对图像分类任务，缺少对于目标检测任务的专门优化设计。

本文在广泛应用的目标检测模型 YOLOv3 的基础上结合 MobileNetv2 进行改进，设计了一种轻量化的主干网络 Res2-MobileNet，以卫星仿真模型为样本，对其本体及部件进行检测识别，大大降低原模型参数量，明显提升检测速度，且具有良好的检测精度。

1 空间目标特性分析

以卫星为代表的空间非合作目标典型结构部件包括卫星本体、太阳帆板、天线、星敏感器、喷管、对接环等。各部件的大小、几何形状呈现多样化特点，并且根据拍摄角度和卫星姿态的不同，各部件在图像中存在遮挡以及明暗变化。

卫星本体结构形状一般较为规则、轮廓清晰，有利于在图像中识别。太阳帆板外部轮廓大多为规则矩形，内部为太阳能电池阵列，且安装位置与卫星主体有一定距离，对其检测识别也相对容易 [17]。卫星根据其执行任务的类型不同会携带不同类型的天线，且数量不止一个，例如鞭状天线、螺旋天线、反射面天线、相控阵天线等。这些天线的结构差异较大，因此可利用卫星的部分先验知识辅助天线识别。对接环和喷管结构相对固定，对于识别效果的提升有一定帮助。

卫星一般还会搭载与其功能相关的有效载荷，各功能卫星和其搭载的载荷可分为以下几种：(1) 科学探测和实验类卫星携带载荷：电离探针、磁强计、质谱计等；(2) 信息获取类卫星携带载荷：光学相机、多谱段扫描仪、微波辐射计、合成孔径雷达、无线电侦察接收机等；(3) 信息传输类卫星携带载荷：通信转发器和通信天线；(4) 信息基准类卫星携带载荷：无线电信标机、激光反射器等 [18]。

2 相关网络算法与模型

2.1 YOLOv3 网络简述

YOLO 系列网络首次将回归思想用在目标识别上，将目标识别转化为回归问题，通过预测目标框的坐标变量，与标定目标框的坐标比较，通过减小损失函数值实现模型迭代优化。YOLOv3 相较之前版本，在目标识别的速度和精度方面都有所提升。

YOLOv3 的主干网络使用 Darknet-53 进行特征提取，网络引入残差模块，有效解决深层网络的梯度问题，每个残差模块由两个卷积层和一个 shortcut 连接组成。网络结构里没有池化层和全连接层，通过设置卷积步长为 2 实现下采样。每个卷积层的实现又包括标准卷积 +BN+Leaky Relu。YOLOv3 采用了多尺度融合 (Feature Pyramid Networks，FPN) 的思想，分别在 8 倍、16 倍、32 倍下采样输出 3 种不同尺度的特征图，将 32 倍下采样输出做一次目标检测，

将 32 倍下采样的输出通过上采样与 16 倍下采样的结果进行通道拼接后再做一次目标预测，8 倍下采样的输出与拼接结果的上采样再进行一次通道拼接后做一次目标检测，从而输出三个不同尺度的预测结果，分别进行三次独立目标检测，即可以保证其检测速度，同时也可以进一步提升目标检测效果，尤其是对于小目标的检测效果。

YOLOv3 将输入网络的图片像素缩减至 416×416，将整幅图像划分为 S×S 个单元格，分别在 13×13、26×26 和 52×52 三种不同大小的特征图上进行预测，在每个特征图的每个单元格都会预测三个不同尺寸的边界框。每一个边界框产生一个 5+C 维向量，其中 C 表示数据类别数，5 维向量包括边框中心坐标、宽高以及边框置信度。置信度(Confidence)公式如式(1)所示：

$$Confidence = \Pr(object) \times IOU_{pred}^{trouth} \tag{1}$$

其中$\Pr(object)$表示单元格中是否由目标物体存在，若目标存在则为 1，不存在为 0；IOU_{pred}^{trouth}表示预测框与真实框的交并比(Intersection Over Union，IOU)。当网络检测完毕时，会将特征图重新映射会原图中绘制预测框，预测框(DetectionResult)与真实框(GroudTruth)在原图中以像素为坐标值计算二者的 IOU，这一过程可表示为：

$$IOU_{pred}^{truth} = \frac{DetectionResult \cap GroudTruth}{DetectionResult \cup GroudTruth} \tag{2}$$

如果单元格内包含目标，则该单元格还需要预测此目标属于第 i 类的概率$\Pr(class_i|Object)$，即目标分类条件概率。最后，对预测框进行非极大值抑制(Non-Maximum Suppression，NMS)处理，选择最好的预测框最为最终的预测框输出。

2.2 MobileNet v2

轻量化网络 MobileNet v2 使用的卷积与标准卷积计算方式不同，主要运用了深度可分离卷积，它是许多高效神经网络的关键结构块，可以减少运算量和参数量，标准卷积与深度可分离卷积结构比较如图 1 所示，深度可分离卷积由两部分组成，第一部分是深度卷积层(Depthwise)，它对特征图的各个通道应用单个卷积核进行卷积操作，第二部分是 1×1 的逐点卷积。深度卷积与标准卷积的计算量比较如式(3)所示：

$$\frac{H \times W \times M \times k \times k + H \times W \times M \times N}{H \times W \times M \times N \times k \times k} = \frac{1}{N} + \frac{1}{k^2} \tag{3}$$

其中 H、W、M 表示输入特征图的宽、高和通道数，N 表示输出通道数，$k \times k$ 表示卷积核大小。可见这种深度卷积的方式可以大大减少卷积计算量。

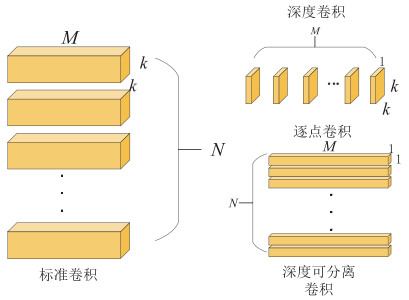


图 1 标准卷积和深度可分离卷积比较

Fig.1 Comparison of standard convolution and depth separable convolution

3 改进后的算法与模型

基于卷积升级网络的目标检测模型普遍存在参数量大、占用资源多、计算速度慢，无法满足空间场景下的嵌入式目标检测任务。本文在 YOLOv3 的基础上进行网络轻量化改进，基于 MobileNet v2 对骨干网络重新设计来进行特征提取，引入 Bottleneck 结构，如图 2 所示，分别为卷积步长 Stride 为 1 和 2 时的 Bottleneck 结构。

该模块将输入的低维特征图通过 1×1 的 Expansion Layer 从低维空间映射到高维空间，之后输入深度卷积层，这样既可以有效减小网络的计算量同时又可以提取到整体的足够多的信息来保证网络的精度，之后通过 Projection Layer 进行降维，使网络重新变小。

为进一步提升模型对于多尺度目标检测的准确度，引入 Res2Net 残差模块，通过增加模型的感受野尺度丰富性和结构深度来提高网络对于小目标的敏感性和特征提取能力，该模块采用一组级联的 3×3 卷积层，使得结构内部也具有了一定的深度及残差性，

通过将同一层内不同的通道之间的特征图建立连接，不仅重用了该层的部分信息，还使得同一层的特征图包含了不同的感受野的特征，增强了轻量化骨干网络的多尺度表达能力。如图 3 所示为 Res2Net 残差模块结构。

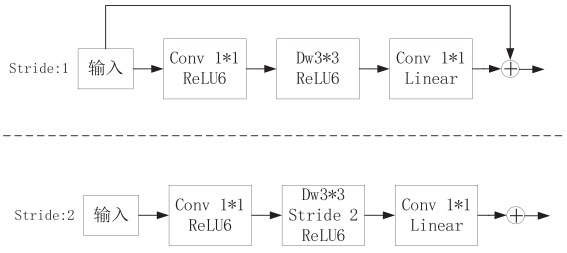


图 2 卷积步长为 1 和 2 的 Bottleneck 结构

Fig.2 Bottleneck structure with convolution stride of 1 and 2

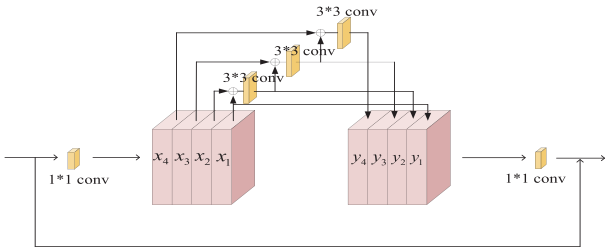


图 3 Res2Net 残差结构

Fig.3 Res2Net residual structure

其中第一层 1×1 卷积层输出特征图通道均分为 s 组，每一组特征用 X_i 表示，$i=\{1,2,3,\cdots, s\}$，对于分组后的每一组特征，除了第一组特征，其他组特征都会将上一组的输出特征y_{i-1}与当前组的x_i进行残差连接，通过卷积操作后生成新的特征图。每一组的输出y_i用公式表达如式 (4) 所示：

$$y_i=\begin{cases}x_i & i=1\\ conv(x_i+y_{i-1}) & 1<i\le s\end{cases} \tag{4}$$

其中 conv 表示 3×3 的卷积。

改进后 Res2−MoblieNet 主干网络结构如表 1 所示，整体网络模型结构如图 4 所示。

表 1 改进后 Res2−MobileNet 网络结构

Tab.1 Improved Res2−MobileNet network structure

输入	操作	步长	重复	输出
416×416×3	Conv2d	2	1	208×208×32
208×208×32	Bottleneck	1	1	208×208×16
208×208×16	Bottleneck	2	2	104×104×24
104×104×24	Res2Net Construction	1	2	104×104×24
104×104×24	Bottleneck	2	3	52×52×32
52×52×32	Res2Net Construction	1	2	52×52×32
52×52×32	Bottleneck	2	4	26×26×64
26×26×64	Bottleneck	1	3	26×26×96

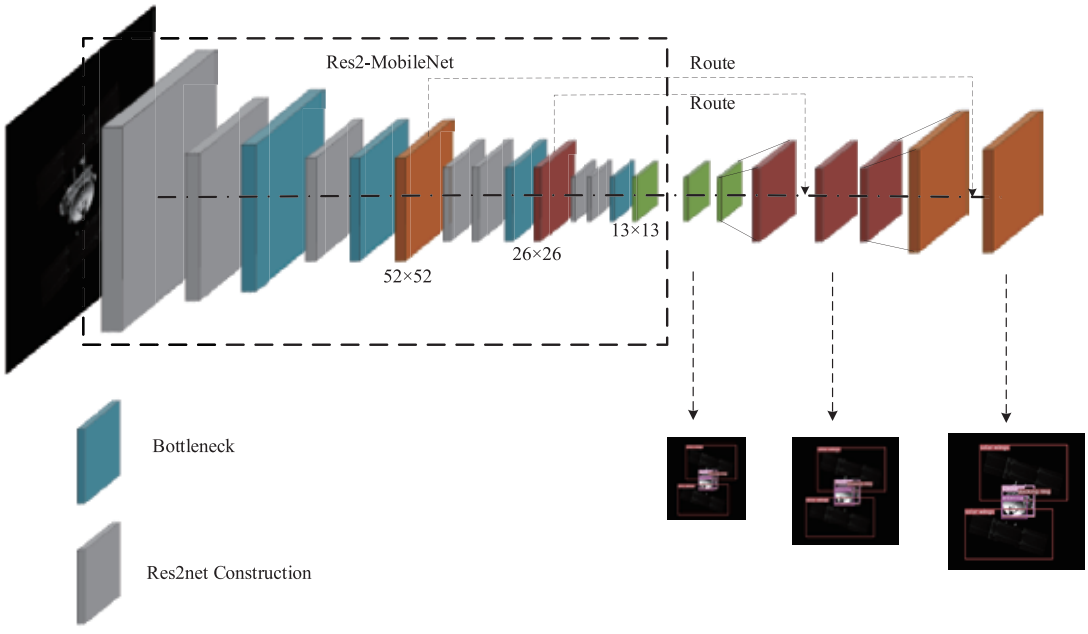


图 4 网络模型整体结构

Fig.4 Whole structure of the network

26×26×96	Res2Net Construction	1	2	26×26×96
26×26×96	Bottleneck	2	3	13×13×160
13×13×160	Bottleneck	1	1	13×13×320
13×13×320	Res2Net Construction	1	2	13×13×320
13×13×320	Conv2d	1	1	13×13×1280

4 实验过程及结果分析

本文实验仿真使用 Pytorch 框架，操作系统为 Ubuntu 16.04.6，GPU 为 Nvidia GeForce RTX 2080Ti。Cuda 版本为 11.0。

4.1 实验数据集

本实验采用自建卫星仿真数据集，包括 CGRO、GPS、quadsat 等 10 种卫星在不同光照条件下各种姿态的仿真图片，通过旋转变换、对比度变换、噪声变换等数据增强处理后，共得到 10000 张图片，如图 5 所示。分别标记图片中的卫星主体、帆板、天线、对接环、喷管。按 6：1：3 比例将本数据集分为训练集、验证集和测试集。

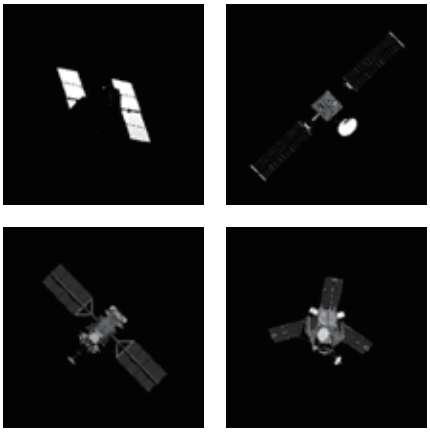


图 5 样本示例

Fig.5 Sample example

4.2 实验过程

本文实验采用平均精度(Average Precision，AP)评估每一种部件的检测精度，采用均值平均精度(Mean Average Precision，mAP)衡量多类部件的平均检测精度。采用每秒帧数(Frame Per Second，FPS)衡量模型检测速度。 输入图像尺寸为 1024×1024，RGB 通道的 JPEG 图像，使用单 GPU 训练；Batch size 设为 8，Epoch 设为 100，共进行 12500 次迭代，学习率设置为 0.0001，使用模拟退火策略调整网络学习率。设置 IoU 阈值为 0.5，置信度阈值为 0.6。同时使用 K-means 聚类算法生成 9 个适用于本数据集的 anchor 大小。训练过程中损失函数变化情况如图 6 所示，可以看到网络的收敛情况较好。

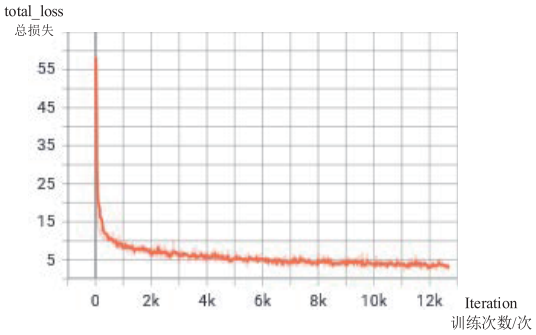
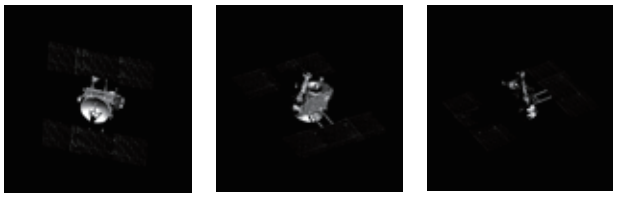


图 6 训练损失

Fig.6 Loss of training

4.3 实验结果及分析

本文网络模型测试结果示例如图 7 所示。



(a) 测试原图

(a) Original image for test

(b) 本文网络模型检测效果

(b) Detection effect of network model in this article

图 7 本文网络模型测试结果示例

Fig.7 Example of network model test results in this article

对改进后的网络模型和YOLOv3进行对比分析，

为了更清晰地反应本网络的优化效果，列出了卫星主体及每一类部件的识别精度，如表 2 所示。

表 2 卫星主体及不同部件的识别精度

Tab.2 Recognition accuracy of satellite body and different parts

类别	YOLOv3	Ours
主体	97.2%	97.4%
帆板	98.5%	98.6%
天线	87.2%	87.9%
对接环	83.3%	86.2%
喷管	81.2%	85.5%

从表 2 可以看出，由于光照和遮挡等原因，对接环和喷管这类小目标识别精度较低。本文模型在对于对接环、喷管的识别相比 YOLOv3 有明显的提升，表明本文所使用的 Res2Net 残差模块能够显著提升小目标的检测能力。其次对于主体、帆板、天线等目标体积较大、易于识别的部件，两种模型都有较高的识别精度，且本文模型精度稍高。在网络模型的大小和网络检测速度方面，将本文模型和 YOLOv3 进行对比得到结果如表 3 所示。

表 3 网络模型对比

Tab.3 Network model comparison

模型	主干网络	模型大小 (MB)	检测速度 (FPS)	mAP(%)
Yolov3	Darknet 53	236	34	89.5
Ours	Res2-MobileNet	105	65	91.1

通过模型对比可以看到，本文改进后的网络模型由于引入 Bottleneck 结构，在模型大小和检测速度方面相比 YOLOv3 有着明显的优势，同时 mAP 值也提升了 1.6%，对于后续模型在嵌入式设备上的部署及其实时性检测方面都有重要意义。

5 结束语

在使用深度学习方法进行空间非合作目标识别时，为解决在太空环境中使用的嵌入式设备计算能力不足神经网络模型难以部署的问题，本文提出了一种基于 YOLOv3 的轻量化网络模型，通过引入 Bottleneck 结构以及 Res2Net 残差结构，同时通过 K-means 聚类生成先验框。实验结果表明，本文模型相比于 YOLOv3 网络大小明显降低，同时提升检测速度和多尺度目标检测精度，为下一步在嵌入式设备上的部署奠定基础。

参考文献（References）

[1] 王柳，基于深度学习的空间多目标识别方法研究 [J]. 无人系统技术，2019，2(03)：49-55.

[2] 杨明冬．空间非合作面目标跟踪技术研究 [D]. 中国科学院研究生院（上海技术物理研究所），2015.

[3] 刘礼城，周如好，张丽敏，等．基于视觉的航天器的特征识别算法 [A]. 中国指挥与控制学会空天安全平行系统专业委员会．第二届中国空天安全会议论文集 [C]. 中国指挥与控制学会空天安全平行系统专业委员会：中国指挥与控制学会空天安全平行系统专业委员会，2017:8.

[4] 化嫣然，张卓，龙赛，等．基于改进 YOLO 算法的遥感图像目标检测 [J]. 电子测量技术，2020，43(24)：87-92.

[5] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proc. of the Computer Vision and Pattern Recognition，Columbus，2014：580-587.

[6] GIRSHICK R. Fast R-CNN[C]//Proc. of the International Conference on Computer Vision，2015：1440-1448.

[7] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Trans. on Pattern Analysis & Machine Intelligence，2017，39(6)：1137-1149.

[8] LIU W，ANGUELOV D，ERHAN D，et al. SSD：single shot MultiBox detector[C]//Proc. of the European Conference on Computer Vision，Amsterdam，2016：21-37.

[9] Li Z，Zhou F. FSSD：feature fusion single shot multibox detector[J]. arXiv preprint arXiv：1712.00960，2017.

[10] FU C，LIU W，RANGA A，et al. DSSD：deconvolutional single shot detector [EB/OL]. http://arxiv.org/abs/1701.06659，2020-07-16.

[11] REDMON J，DIVVALA S K，GIRSHICK R，et al. You only look once：unified，real-time object detection[C]//Proc. of the Computer Vision and Pattern Recognition，2016：779-788.

[12] REDMON J，FARHADI A. YOLO9000：Better，faster，

stronger[C]//Proc. of the Computer Vision and Pattern Recognition,2017：6517-6525.

[13] REDMON J，FARHADI A. YOLOv3:An incremental improvement [EB/OL]. http://arxiv.org/abs/1804.02767，2020-07-16.

[14] CHOLLET F. Xception：deep learning with depthwise separable convolutions[C]//IEEE conference on computer vision and pattern recognition. Honolulu，HI：IEEE，2017：1800-1807.

[15] 孔英会，朱成诚，车辚辚. 复杂背景下基于 Mobile-Nets 的花卉识别与模型剪枝 [J] . 科学技术与工程，2018，18(19)：84-88.

[16] ZHANG X，ZHOU X，LIN M，et al. ShuffleNet：an extremely efficient convolutional neural network for mobile devices[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City，UT：IEEE，2018：6848-6856.

[17] PAYNE T，GREGORY S，LUU K. SSA analysis of GEOS photometric signature classifications and solar panel offsets[C]//The Advanced Maui Optical and Space Surveillance Technologies Conference. 2006：E73.

[18] WALKER G K，TAYLOR J K. Satellite identification and antenna alignment [J] .Molecular Ecology Notes，1994，6 (3) ：882-885.

作者简介：

郝强（1996—），男，内蒙古自治区乌兰察布市人，硕士研究生，研究方向为星上智能处理技术。

基于正则化多元逻辑回归的 GNSS／INS 组合导航系统非完整约束算法

吕 冰，刘肖姬，郭 权，倪 枫，李 楠，李文杰
（北京微电子技术研究所，北京市 100076）

摘 要：全球导航卫星系统 (GNSS) 和惯性导航系统 (INS) 卫星惯性组合导航接收机在独立模式工作时，位置、速度和姿态方面的误差会大幅增加。在传统的紧组合导航方案中，系统在 GNSS 信号中断时采用预测模型，定位精度由惯性导航的精度决定。同时，由于卫星信号丢失，缺乏观测使得利用 GNSS 信息估计惯性导航误差的可靠性较低。提出了一种改进的基于正则化多元逻辑回归的非完整约束 (NHC) 方法，目的是在可见星数量不足时提高导航精度。速度约束条件可以用来简化基于微机电系统 (MEMS) 的惯性导航系统的系统计算方程。此外，通过训练基于收集数据的正则化多元逻辑回归模型可以来识别车辆运动模式，从而实现深层次约束。仿真和现场测试结果表明，有效降低了导航误差，有利于提高低成本 GNSS/INS 组合导航接收机的精度。

关键词：多元逻辑回归；组合导航；非完整约束
中图分类号：TN967.2 **文献标识码**：A

Using Regularized Softmax Regression in the GNSS/INS Integrated Navigation System with Nonholonomic Constraints

Lv Bing, Liu Xiaoji, Guo Quan, Ni Feng, Li Nan, Li Wenjie
（Beijing Microelectronics Technology Institute, Beijing, 100076, China）

Abstract: The integration of global navigation satellite system(GNSS) and Inertial navigation system (INS) is widely implemented in land-vehicle navigation applications. However, the satellite signal is vulnerable in some special urban scenarios, consequently errors in terms of position, velocity and attitude grow rapidly in stand-alone mode especially forlow-cost MEMS-based INS. In the conventional tight combination navigation schemes, system works on predicting model during the GNSS signal outage and the positioning accuracy is determined by the precision of the inertial navigation. Besides the lack of observation makes the estimate of inertial navigation error with GNSS information less reliable due to the satellite signal loss. In this paper, an improved non-holonomic constraints (NHC) method based on regularized softmax regression is proposed to enhance navigation precision when the number of visible satellite is insufficient. The velocity constraint condition is applied to simplify the system calculating equations of MEMS-based INS. Furthermore, a regularization softmax regression model based on the collected data is trained to recognize the vehicle motion pattern so as to realize deeper constraints. Simulation and field-test results indicate that the method is beneficial to raise the precision of low-cost GNSS/INS integrated navigation receiver by efficiently reduce the navigation errors.

Key words: FPGA; softmax regression;integrated navigation; nonholonomic constraints

0 引言

基于全球导航卫星系统和惯性导航系统（INS）的组合导航接收芯片已广泛应用于陆上车辆导航（LVN）等各个领域[1]。GNSS 和 INS 的组合导航利用了这两种导航方法，由于独立工作的 INS 可以提供完整的六自由度解决方案，具有高更新率，弥补了卫星导航在抗干扰方面的不足。MEMS 惯性传感器在功耗、重量和成本方面占主导地位，完全满足 INS 的规范和要求。然而，低成本、低品质的惯性传感器输出的MEMS惯性系统的广泛使用使系统噪声较高，

不确定性较大，成本降低导致了整体精度的下降[2]。因此，使用低成本惯性传感器的惯性导航系统只能在非常有限的时间内用于独立导航。尽管系统可以对传感器偏差进行估算，但 GNSS 信号中断期间的位置误差大幅增大，导航解决方案的参考价值会急剧下降。

导航精度可以通过提高硬件结构或导入其他传感器数据来提高[3]。系统性能越高，所需成本和系统复杂度也越高。在大多数组合方案中，GNSS 作为校正惯性传感器系统误差的准确参考。组合导航系统可以工作在不使用 GNSS 的环境中，但是缺乏参照物使会使 GNSS 信息惯性导航误差估计的可靠性较低[4]。对陆上车辆导航应用来说，非完整约束（NHC）是最常见的辅助信息类型之一。

本文提出了一种基于低成本 INS 和 GNSS 接收机的 NHC 紧组合导航系统，详细描述了 INS 伪距方程和伪距率方程。同时，组合滤波器导入了 NHC 条件用于 LVN 应用。此外，本文使用了正则化多元逻辑回归来识别车辆运动模式，从而可以实现更深层次的约束。经过仿真分析，验证了该方法的有效性，现场测试结果表明在卫星信号中断期间，系统性能得到了显著提高。

1 实现方法

1.1 紧耦合方案总体设计

图 1 描述了组合导航方案的简要结构图。通过惯性导航结果和卫星星历信息计算出与卫星的相对距离，作为惯性导航系统的伪距和伪距率。GNSS 观测量和惯性导航系统的伪距和伪距率做差作为扩展卡尔曼滤波器的观测量输入[5]。注意数据预处理和初始对准不在本文的讨论中。

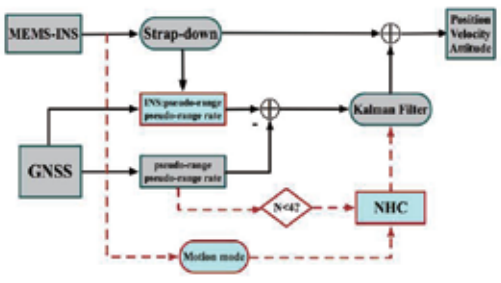


图 1 GNSS/INS 整体方案

Fig.1 The overall scheme of GNSS/INS

公式（1）定义了基于伪距和伪距率的紧耦合 INS/GNSS 组合滤波器更新的状态向量。

$$X = [\phi_e \ \phi_n \ \phi_u \ \Delta\lambda \ \Delta l \ \Delta h \ \Delta v_e \ \Delta v_n \ \Delta v_u \ \delta t_u \ \delta t_{ru}]^T \quad (1)$$

其中下标 e、n、u 代表地心坐标系的三轴，λ 、l、h 分别代表经度、纬度和高度坐标。ϕ 表示平台的误差角，速度、时钟和频率的误差分别表示为是 Δv，δt_u和δt_{ru}。

1.2 伪距和伪距率方程

令惯性单元数据计算得到的位置为$(x_I \ y_I \ z_I)^T$，从星历信息中提取的准确的卫星位置为$(x_S \ y_S \ z_S)^T$。GNSS 接收机的伪距表示为ρ_G。从 INS 到 GNSS 卫星的实际距离如公式（2）所示：

$$r_j = [(x_{sj} - x)^2 + (y_{sj} - y)^2 + (z_{sj} - z)^2]^{\frac{1}{2}} \quad (2)$$

实际坐标方向为$(x \ y \ z)^T$，INS 测量距离的泰勒级数展开式可表示为公式（3），取一次项。因此，通过计算位置值误差，系统的伪距如公式（5）所示。

$$r_j = [(x_{sj} - x_I)^2 + (y_{sj} - y_I)^2 + (z_{sj} - z_I)^2]^{\frac{1}{2}} - H \quad (3)$$

$$H = \frac{\partial r_j}{\partial x}\delta x + \frac{\partial r_j}{\partial y}\delta y + \frac{\partial r_j}{\partial z}\delta z \quad (4)$$

$$\rho_{Ij} = r_j - e_{j1}\delta x - e_{j2}\delta y - e_{j3}\delta z \quad (5)$$

$$\frac{\partial \rho_j}{\partial x} = -\frac{x_{sj} - x_I}{r_j} = -e_{j1} \quad (6)$$

$$\frac{\partial \rho_j}{\partial y} = -\frac{y_{sj} - y_I}{r_j} = -e_{j2} \quad (7)$$

$$\frac{\partial \rho_j}{\partial z} = -\frac{z_{sj} - z_I}{r_j} = -e_{j3} \quad (8)$$

类似地，可以推断伪距率如公式（11）所示。GNSS 接收机的伪距和伪距率的表达式在公式（9）和公式（10）中给出。 $\upsilon_{\rho j}$代表卫星 j 的电离层误差。

$$\rho_{Gj} = r_j + \delta t_u + \upsilon_{\rho j} \quad (9)$$

$$\dot{\rho}_{Gj} = e_{j1}(\dot{x}_{sj} - \dot{x}) + e_{j2}(\dot{y}_{sj} - \dot{y}) + e_{j3}(\dot{z}_{sj} - \dot{z}) + \delta t_{ru} + \dot{\upsilon}_{\rho j} \quad (10)$$

$$\dot{\rho}_{Ij} = \dot{r}_j - e_{j1}\delta \dot{x} - e_{j2}\delta \dot{y} - e_{j3}\delta \dot{z} \quad (11)$$

因此，伪距和伪距率偏移的观测方程可以表示为：

$$\delta\rho_j = e_{j1}\delta x + e_{j2}\delta y + e_{j3}\delta z + \delta t_u + \upsilon_{\rho j} \quad (12)$$

$$\delta\dot{\rho}_j = e_{j1}\delta \dot{x} + e_{j2}\delta \dot{y} + e_{j3}\delta \dot{z} + \delta t_u + \dot{\upsilon}_{\rho j} \quad (13)$$

可以得出，紧耦合组合系统将伪距和伪距率偏移作为组合滤波器的观测量来更新导航信息。

1.3 LVN 的非完整约束

当可见星数小于 4 时，组合滤波器虽然仍然可以工作，但由于缺乏观测量，导航结果的精度和可靠性无法保证。NHC 方法实际上多用于陆地车辆的导航。对于城市环境中的陆地车辆，可以假设载体的高度在短时间内不会突然变化，那么丢失前一个时刻之的高度可以用作 GNSS 中断过程中一个高度约束的测量。此外，NHC 假设在 LVN 的情况下，除非车辆离开地面或在地面上滑动，车辆在垂直于前一方向的速度几乎为零。速度约束如公式（14）所示。

$$\left(v_x^b, v_z^b\right)^T \sim N\left(0, R_v\right) \quad (14)$$

其中R_υ代表由于角度未对准引起的扰动误差。对于高度值，滤波器工作在约束状态下时，系统加入了一个测量方程，如公式（15）所示。

$$h_{INS} - h_{const} = \delta h + \upsilon \quad (15)$$

其中h_{INS}和h_{const}分别表示当前时刻 INS 解算得到的高度值和上一卫星信号良好时刻 GNSS 解算的高度值。假设h_{const}是系统丢星时间内的高度真实值，那么导航解算高度与不丢星时高度值相同，否则的话系统的测量值会存在偏差。根据标准卡尔曼滤波方程，测量值的偏差会由状态估计方程计算到最后的滤波结果中，当h_{const}与真实值相差过大时，高度约束条件会导致更大的误差。然而，对于陆地车载系统来说，在运行过程中高度不会大幅变动，即使假定高度不准确，定位解决方案仍然优于不加约束时的定位解决方案。在速度约束时，相当于再添加两个假设的真值。与高度约束类似，如果车辆处于高动态情况下，例如转弯或方向变化很大，测量误差会更大。因此，对于 NHC 方法仍需要引进更深层次的约束条件。

2 深层次约束改进方法

2.1 运动模式识别中的正交多元逻辑回归

当车辆改变姿态或方向时，位置和速度的约束条件误差较大，换句话说，NHC 的性能受运动状态的影响。惯性传感器可以反映载体的运动情况，可以根据惯性导航系统的输出确定当前的运动状态引进不同的约束条件 [6]。

多元逻辑回归是逻辑回归在多分类问题上的扩展，属于有监督学习模型 [7]。多元逻辑回归的概率函数和损失函数定义如下所示。

$$h_\theta\left(x^{(i)}\right) = \frac{1}{\sum_{j=1}^{3}\exp\left(\theta_j^T x^{(i)}\right)}\begin{bmatrix}\exp\left(\theta_1^T x^{(i)}\right)\\ \exp\left(\theta_2^T x^{(i)}\right)\\ \exp\left(\theta_3^T x^{(i)}\right)\end{bmatrix} \quad (16)$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{3}\left\{y^{(i)} = j\right\}\ln\left(\frac{\exp\left(\theta_j^T x^{(i)}\right)}{\sum_{n=1}^{3}\exp\left(\theta_n^T x^{(i)}\right)}\right)\right] + \lambda\|\theta\|^2 \quad (17)$$

其中m代表数据样本的数量，θ代表分类器的权重参数。此外，引入正则项作为调节项，以避免参数过拟合和冗余。因此，损失函数由交叉熵和正则项组成，可用于获得最优解 [8]。针对车辆的运动特点，运动状态大致可以概括为静止、直行和转弯。基于实际数据的训练方法如图 2 所示。

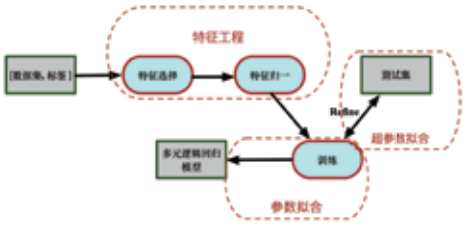


图 2 训练方法
Fig.2 Training method

数据样本和校准是真实的路测数据。由于 MEMS-INS 固定在车辆上，在行驶过程中，扰动也会在垂直方向产生加速度。考虑到陆地车辆的运动特性，本文提出了一种实用的特征选择方法。特征表达式如下所示。

$$x_1 = \sum_{t_i=t_k-n}^{t_k}\omega_{norm}(t_i) \quad (18)$$

$$x_2 = \sum_{t_i=t_k-n}^{t_k} \left| a_{norm}(t_i) - a_{norm}(t_i - i) \right| \quad (19)$$

$$x_3 = \frac{1}{m} \sum_{t_i=t_k-m}^{t_k} \omega_z(t_i) \quad (20)$$

其中 $\omega_{norm} = \sqrt{\omega_x^2 + \omega_y^2}$ 代表水平角速度，$a_{norm} = \sqrt{a_x^{b2} + a_y^{b2} + a_z^{b2}}$ 代表总的加速度。同时，进行了尺度归一化处理。样本中每个维度的特征范围均归一化为 [0，1]。我们可以通过基于当前特征向量的多元逻辑回归得到运动状态概率，如图 3 所示。

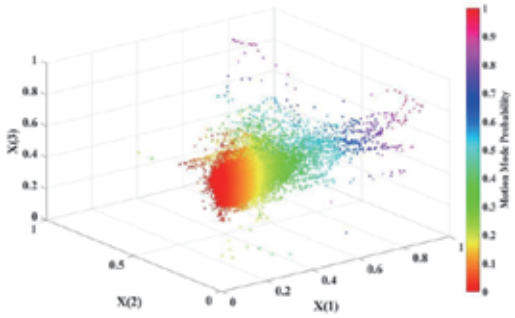


图 3 正则化多元回归的概率分布

Fig.3 The probability distribution by regularized softmax regression

运动状态可以由最优概率估算得到。在道路实测时，识别性能如图 4 所示。在大多数情况下，结果与实际运动状态一致。

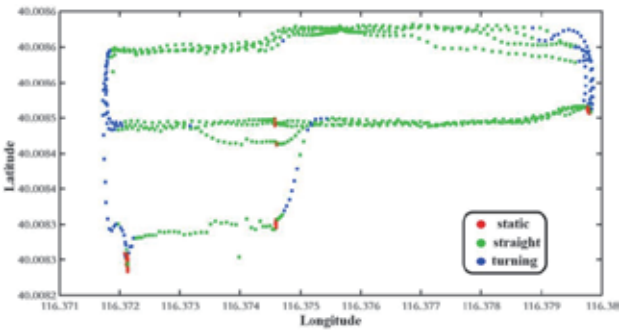


图 4 正则化多元回归的运动模式解决方案

Fig.4 The motion mode solution by regularized softmax regression

2.2 NHC 的更深层次限制

对运动模式进行识别后，可以根据不同的运动状态引入不同的运动约束条件。首先对于静止状态，需要进行零速度校正，同时通过加速度计校正修改俯仰角和横滚角，此外还需要保持航向角不变且重新计算陀螺仪零偏值。约束条件如下。

$$v = 0 \quad (21)$$

$$\theta = \sin 1^{-1}\left(\frac{a_y}{g}\right) \quad (22)$$

$$\phi = \sin^{-1}\left(\frac{-a_x}{\sqrt{g^2 - a_y^2}}\right) \quad (23)$$

$$B_\omega = \omega \quad (24)$$

对于直线运动，横滚角可以通过 x 方向的加速度计进行校正，如公式（25）所示。对于转弯运动，速度可以通过向心加速度与角速度之间的关系来计算得到，如公式（26）所示。

$$\phi = \sin^{-1}\left(\frac{-a_x}{\sqrt{g\cos\theta}}\right) \quad (25)$$

$$v = \frac{a_x + g\sin\phi\cos\theta}{\omega_z} \quad (26)$$

3 实验与讨论

3.1 使用改进的 NHC 模拟 GNSS/INS

上文已详细说明了该方法在车辆运动状态识别方面的性能。接下来将带有正则化多元逻辑回归的 NHC 模型导入到自主研发的 GNSS/INS 组合导航接收机中，如图 5 所示。

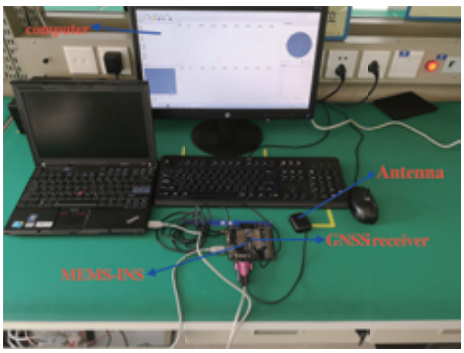


图 5 组合导航测试平台

Fig.5 The self-developed GNSS/INS

可见星不足的模拟环境由信号模拟器生成，当可见星数量少于 4 颗时，该系统在改进后的 NHC 基础上仍能高精度工作。图 7 显示了模拟静止状态的结果

误差。

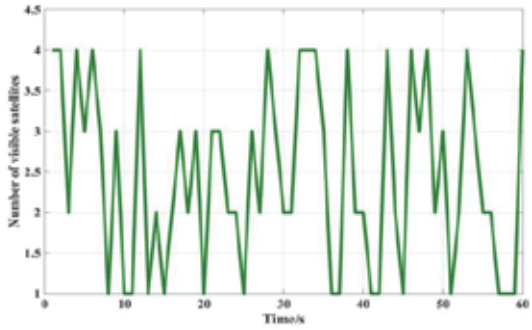


图 6 可见卫星数量

Fig.6 The number of visible satellites

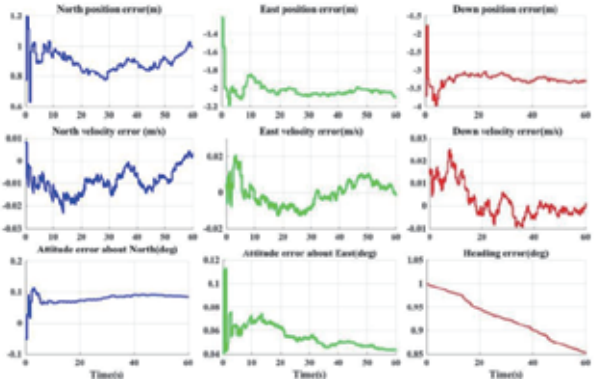


图 7 静态仿真中的导航误差

Fig.7 The navigation error in static simulation

表 1 位置误差的均方根误差

Tab.1 RMSE of location error

RMSE of location error	Static Simulation
East position (m)	2.87
North position (m)	0.95
Down position (m)	3.37
East velocity (m/s)	0.02
North velocity (m/s)	0.01
Down velocity (m/s)	0.12
Pitch (°)	0.18
Roll (°)	0.12

如表 1 所示，在模拟静止状态场景中，该方法可以进一步减小误差，NHC 显著地提高了系统整体精度。系统在识别为静止状态时，将在零速度更新模式下工作。

3.2 路测试验

现场试验沿图 8 所示路线进行实地测试，该路线为城市场景。进一步测量导航误差后，得到该方法的性能如图 9 所示。

图 8 路试路线

Fig.8 The road-test route

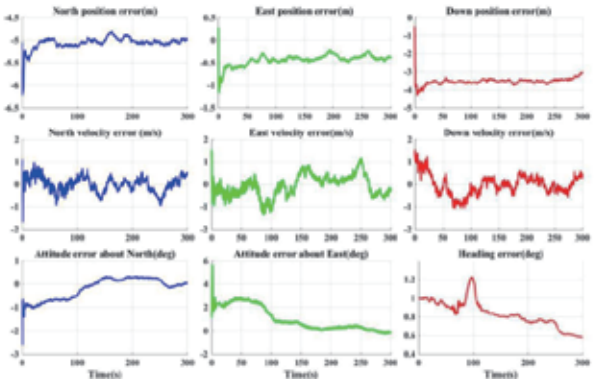


图 9 路试导航误差

Fig.9 The navigation error of road test

表 2 位置误差的均方根误差

Tab.2 RMSE of location error

RMSE of location error	Static Simulation
East position (m)	5.47
North position (m)	0.75
Down position (m)	3.89
East velocity (m/s)	0.98
North velocity (m/s)	1.12
Down velocity (m/s)	0.87
Pitch (°)	1.12
Roll (°)	1.84

如图 9 和表 2 所示，由于现场试验情况复杂，导航误差相对静态模拟有一定程度差别，但精度得到了提高。此外，向下的位置误差是由初始对齐引起的。

正常行驶在不丢星的情况下，GNSS-INS 工作在正常模式，位置定位精度能够达到 3 轴 10m，速度定位精度能够达到 0.1m/s，姿态精度能够达到 1°/s，满足车辆导航的精度需求。

在丢星情况下，GNSS-INS 工作在约束条件下，位置定位精度能够达到单轴 10m，速度定位精度能够达到 1.5m/s，姿态精度能够达到 2°/s，由于加入约束，误差没有迅速积累，验证此方法可以满足车辆导航的精度需求。

可以证明，针对 LVN 应用中基于 MEMS 的低成本导航系统，通过 NHC 中的正则化多元逻辑回归实现更深层次的约束，提高了导航性能，并验证了其有效性。

4 结论

本文实现了一种基于低成本 INS 和 GNSS 接收机的自主开发 NHC 紧组合导航系统的应用正则化多元逻辑回归算法。该算法应用正则化多元逻辑回归来识别车辆运动模式，从而实现更深层次的约束。经过详细的分析和现场测试结果推断，本文所述方法可以显着提高 GNSS 信号中断期间系统的性能。

参考文献（References）

[1] NOURELDIN A ,KARAMAT T B,GEORGY J .Fundamentals of Inertial Navigation, Satellite-based Positioning and their Integration|Cld[J].2013,10.1007/978-3-642-30466-8:125-166.

[2] SHIN E H . Accuracy Improvement of Low Cost INS/GPS for Land Applications[J]. sheimy, 2001.

[3] 董明．卫星／惯性／视觉组合导航信息融合关键技术研究[D]. 解放军信息工程大学，2014.

[4] 高帅和．分布式 GPS/SINS 超紧组合架构下的信号处理和信息融合技术研究 [D]. 哈尔滨工程大学 ,2012.

[5] SHIN E H . Estimation techniques for low-cost inertial navigation.[D]. University of Calgary (Canada).2005.

[6] 路丹晖．融合视觉与惯性导航的机器人自主定位 [D]. 浙江大学 ,2012.

[7] PEREIRA F, MITCHELL T, BOTVINICK M. Machine learning classifiers and fMRI: a tutorial overview.[J]. Neuroimage,2009,45(1):199-209.

[8] 陈灿．IMU 辅助 GPS 接收机载波环路跟踪算法研究 [D]. 重庆大学 ,2015.

作者简介：

吕冰（1993—），女，黑龙江省齐齐哈尔市人，硕士，工程师，目前从事软件设计。

新型双栅隧穿场效应晶体管电特性增强工艺对比仿真研究

王倩琼，赖晓玲，巨 艇，张 健，朱 启
（中国空间技术研究院（西安分院），陕西省 西安市，710100）

摘 要：本文对新型掺杂型双栅隧穿场效应晶体管(double gate TFET, DGTFET)器件的直流特性和频率特性进行了对比仿真研究。利用Silvaco-Atala仿真工具构建了源区和体区之间加入重掺杂势垒pocket层的DGTFET器件，和基于在器件前、背栅均使用两种功函数的栅材料共同完成栅极驱动的三栅DGTFET（Triple gate material DGTFET, TGM-DGTFET)，并与传统DGTFET器件的电特性进行了对比研究。研究结果表明，应用于TGM-DGTFET器件的栅极靠近源区一侧和靠近漏区一侧使用较低功函数的栅材料方法，可有效提升带带隧穿几率，拥有最好的开态电流特性，大大提高了器件性能。当栅压为1.5V且漏压为0.5V时，TGM-DGTFET器件较传统双栅隧穿场效应晶体管的开态电流提高了12倍，电流开关比数量级可达到10^{11}。此外，又通过跨导、输出电导、截止频率和增益带宽积等频率表征参数进一步验证了使用多种功函数构成的栅电极方法较在源区和体区之间增加势垒层提高器件频率特性更为有效。

关键词：双栅隧穿场效应晶体管；直流特性；频率特性；重掺杂势垒层；功函数
中图分类号：TL99 **文献标识码**：A

Analog/RF and DC Performance of Novel Double Gate Tunneling FETs with Improved Processes: A Comparison Study

Wang Qianqiong, Lai Xiaoling, Ju Ting, Zhang Jian, Zhu Qi
(China Academy of Space Technology(Xi'an Branch), Xi'an, 710100, China)

Abstract: The DC characteristics and analog/RF performance comparison of several novel double gate TFETs (DGTFETs) are investigated. The DGTFET with pocket layers and the Triple gate material DGTFET (TGM-DGTFET) are investigated by using Silvaco-Atalas simulation tool. Among all the considered devices, TGM-DGTFET has the best on-current due to the most efficient band-to-band tunneling (BTBT) rate. When V_{gs}=1.5V and V_{ds}=0.5V, the on-state current of TGM-DGTFET is 12 times larger than that of convential DGTFET (Co-DGTFET), and the magnitude of on/off current ratio can reach 10^{11}. In addition, the RF performance of the above three devices is researched by simulating the transconductance, output conductance, cut-off frequency, and gain bandwidth product. The results suggest that triple gate electrodes of TGM-DGTFET has the most outstanding characteristic. Thus, according to the analysis above, the triple gate electrode would improve RF performance more effectively than the technique in which a pocket layer is added between source and channel region for double gate TFET.

Key words: Double Gate Tunnel Field-Effect Transistor(DGTFETs); DC characteristics; analog/RF performance; pocket layers; work function

0 引言

随着集成度不断上升，短沟道效应、漏致势垒降低、热载流子效应以及阈值电压的降低导致关态泄漏电流急剧增大等可靠性问题会使得MOSFET的功耗问题变得更为严重[1]。除此之外，由于物理机制的限制，MOS器件亚阈值摆幅在室温下无法低于60mV/dec，使得其很难满足未来集成电路低功耗设计的应用要求。因此，迫切需要研究新型结构器件或使用新型半导体材料降低集成电路的功耗问题。为解决以上问题，国内外学者已通过大量研究提出诸如：隧穿场效应晶体管(Tunnel Field-Effect Transistor，TFET)[2,3]、碰撞电离MOS器件[4,5]、铁电器件[6,7]、

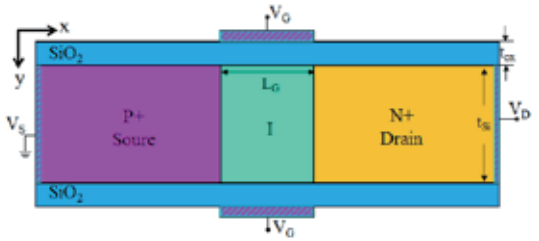
悬栅器件[8]等超低功耗新型器件。其中，TFET 由于可在下一代纳米集成电路应用中替代 MOS 器件而备受瞩目。TFET 通过源极至沟道间的带带隧穿这种新型载流子输运机制，实现了在室温下实现远小于 60mV/dec 的亚阈值摆幅以及较低的关态漏电流，具有更为陡峭的开关特性。

然而，由于较低的开态电流和双极效应使得 TFET 器件在数字电路应用中受到了极大限制，为了提高器件的性能，学者们通过不懈的研究，提出了许多新型栅结构 TFET 器件，如多栅[9,10]、Ω 形栅[11]、环形栅[12]、堆叠栅[13]、栅向衬底嵌入[14]等结构器件，实现了隧穿晶体管良好的性能提高，且均突破了 60mV/dec 亚阈值摆幅极限。目前研究发现，非平面栅结构会增加器件工艺复杂度。例如，环形栅结构会增加工艺步骤，同时其计量学会面临挑战。Ω 形栅器件研制采用的是二维材料，而当前缺乏将二维材料引入硅半导体工厂产线的解决方案。对于双栅、堆叠栅、栅向衬底嵌入等结构器件的工艺继承性存在优势。其中，对称双栅可有效增强栅控能力，是目前主流的 TFET 器件结构。但若采用这类简易平面架构，则需同时配备增强工艺技术来实现开态电流和频率特性的提升。然而，目前针对此类器件的各种电特性（直流特性和频率特性）增强工艺的提升能力量化评估对比研究较少。因此，本文对用于提高双栅隧穿场效应晶体管电特性的有效工艺方法进行对比仿真研究。其中一种增强工艺技术是，利用在源区和体区之间加入 pocket 层[15]，来构建器件，这一势垒层的增强工艺技术在工艺流片得到了验证，对制成的 TFET 器件性能有显著提升[16]，可利用 CVD 技术生长出一层仅几个纳米厚度的超薄重参杂区。而另一种增强工艺技术，是在器件前、背栅均使用两种功函数的栅材料共同完成栅极驱动方法构造新型 DGTFET，这在工艺的实现主要是栅极金属的选择，即采用不同功函数的金属作为不同功能的栅电极[17]。同时，将以上所述器件与传统DGTFET的频率特性进行了量化对比研究。在对比中发现，在栅极靠近源区一侧和靠近漏区一侧使用较低功函数的栅材料可提高器件频率特性并有效抑制双极电流。

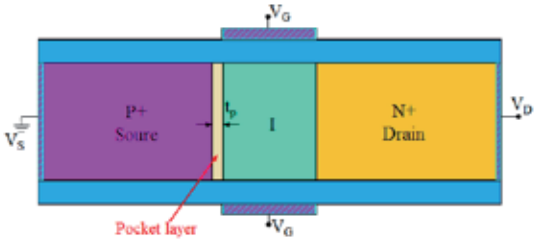
1 器件结构及仿真方法

1.1 器件模型

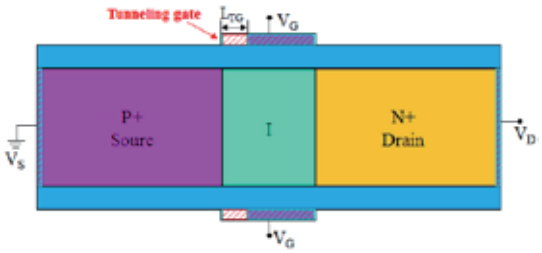
图 1 为本文仿真的几种双栅隧穿场效应晶体管器件剖面结构图，其中，图 1(a) 为传统双栅隧穿晶体管 (Convential double gate TFET，Co-DGTFET)，当器件的源极接地，漏极接正电压时，器件中的隧穿结位于源 / 体结附近；图 1(b) 则在 Co-DGTFET 器件的源区和体区之间增加一层 5nm 宽的 N 型重掺杂 pocket 层[18]，用来提供隧穿时的电子，并增大源至体区的势垒，以此抑制器件的双极电流；图 1(c) 为在 Co-DGTFET 器件的栅极靠近源区一侧的部分换为较低功函数的栅材料，构成隧穿控制栅，用以提高电子 BTBT 隧穿产生率，可有效提高器件开态电流及频率特性，将该器件命名为两种栅电极的双栅隧穿晶体管 (Dual gate material DGTFET，DGM-DGTFET)；图 1(d) 中的优化器件则是在图 1(c) 的器件基础上在栅极靠近漏区的部分换为与 TG 极相同功函数的栅材料，构成辅助栅，有助于降低漏 / 体结处横向电场，从而抑制热载流子效应和双极电流，为了与 DGM-DGTFET 器件区分，在此命名为三种栅电极的双栅隧穿晶体管 (Triple gate material DGTFET，TGM-DGTFET)。以上仿真器件均采用硅材料构成源区、体区和漏区，且掺杂分布选择均匀掺杂，而氧化层则由 SiO_2 材料构成。仿真器件的具体参数如表 1 所示[19-20]。



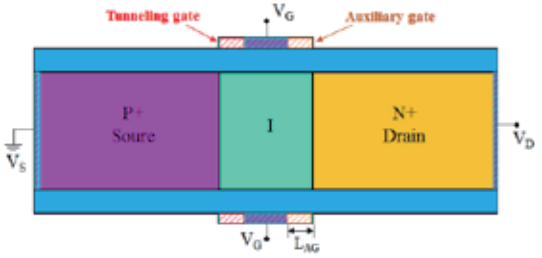
(a)



(b)



(c)



(d)

图 1 几种双栅隧穿场效应晶体管结构示意图：(a) 传统双栅隧穿场效应晶体管 (b) 含 pocket 层双栅隧穿场效应晶体管 (c) 两种栅电极的双栅隧穿场效应晶体管 (d) 三种栅电极的双栅隧穿场效应晶体管

Fig.1 Cross sectional view of simulated devices: (a) Co-DGTFET (b) DGTFET with pocket (c) DGM-DGTFET and (d) TGM-DGTFET

表 1 器件仿真的工艺参数

Tab.1 Device parameters used for the simulations

参数	值	参数	值
体硅厚度 (t_{Si})	10nm	源区掺杂浓度 N_S	$1\times10^{20}cm^{-3}$
前、背栅氧化层厚度 (t_{ox})	2nm	漏区掺杂浓度 N_D	$1\times10^{18}cm^{-3}$
Pocket 层厚度 (t_p)	5nm	Pocket 掺杂浓度 N_P	$1\times10^{19}cm^{-3}$
总栅长 (L_G)	50nm	体区掺杂浓度 N_B	$1\times10^{15}cm^{-3}$
隧穿栅长 (L_{TG})	5nm	控制栅功函数 Φ_{CG}	4.6eV
辅助栅长 (L_{AG})	15nm	隧穿栅 (Φ_{TG})/ 辅助栅 (Φ_{AG}) 功函数	4.0eV

1.2 仿真模型和方法

本文的研究工作均通过 silvaco 的 TCAD 软件仿真得到。与基于热载流子注入的 MOSFET 器件不同，TFET 是一种新型载流子输运机理的器件，其载流子注入机理是通过源区与沟道间的带间量子隧穿实现的，因此仿真中选用了动态非局域性隧穿模型，为了更准确的计算均匀电场下隧穿的产生率 G，根据大多数文献的结论 [21-23]，本文选用了 Kane 模型，该模型用公式可描述为：

$$G = A(F/F_0)^P \exp(-E/F) \quad (1)$$

其中，F_0=1V/cm，本文中所用的硅材料是间接带隙半导体，选用间接隧穿机制 P=2.5，并且根据文献 [24] 中实验测得的参数 A 为 $4.0\times10^{14}cm^{-1}s^{-1}$，$B$ 为 $9.9\times10^{6}V/cm$，F 为电场强度。

由于源区、漏区和 pocket 的重掺杂分布，仿真模型还需添加禁带变窄模型和费米 - 狄拉克统计分布；由于界面陷阱对带带隧穿机制存在一定的影响，陷阱辅助隧穿模型也需考虑；而对于载流子的复合，本仿真中考虑了 SRH 复合和俄歇复合，此外还包括迁移率随掺杂浓度和电场的变化模型。由于该模型的硅膜厚度为 10nm，量子效应可以忽略，因为量子限制模型只有当硅膜厚度低于 5nm 时才会考虑 [25]。而器件频率特性则通过在 Silvaco 交流小信号分析的 Y 参量矩阵计算得到，其中工作频率为 1MHz。

2 新型隧穿场效应晶体管的电特性

2.1 隧穿场效应晶体管的直流特性对比研究

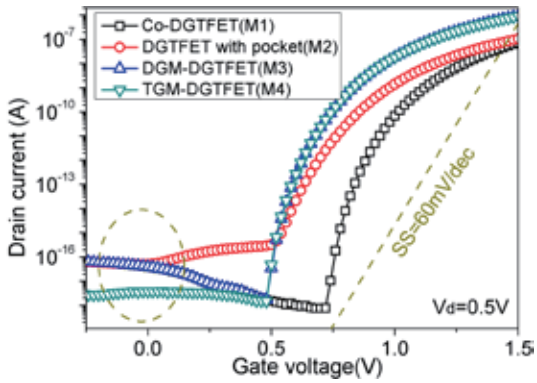


图 2 漏压为 0.5V 时双栅隧穿场效应晶体管转移特性曲线

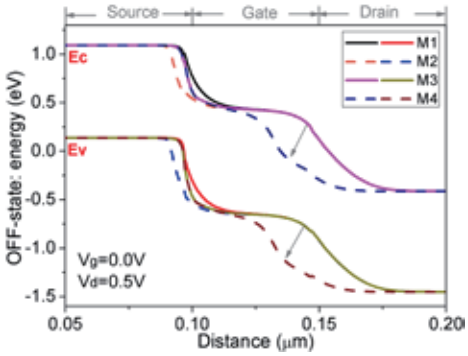
Fig.2 Simulated transfer characteristics for the four TFETs at V_d=0.5V

图 2 所示为漏压为 0.5V 时本文建立的以上四种双栅隧穿场效应晶体管仿真计算所得到的转移特性曲线。从图中可以看到，以上 DGTFETs 器件的最小亚阈值摆幅都远小于 60mV/dec，且三种改进器件的阈值电压较传统双栅隧穿晶体管 (M1) 均负方向移动。

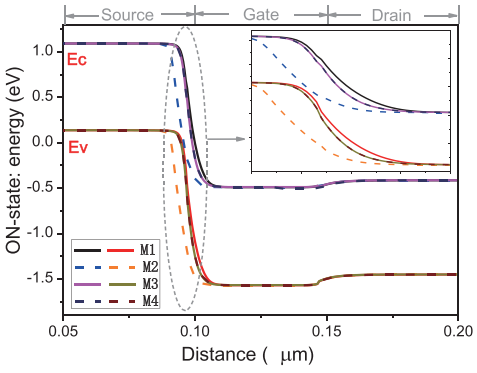
其中，通过在源区和体区间引入 N^+pocket 这一势垒层，使得 M2 器件 (DGTFET with pocket) 在栅压为负值时可有效抑制双极电流，如图中虚线圈内数据所示，在较低栅压 (≤ 1.0V) 时漏电流也有明显的提高。但在含有此类改进方法的器件中，随着栅压的增大，漏电流增长幅度却在逐渐减小，当栅压从 1.0V 增至 1.5V 时，M2 器件的漏电流由 M1 器件的 22 倍降为仅仅提高了 39.4%，且器件的关态电流也有所提高，这是设计高性能 TFET 器件所不希望的。

对于在栅极靠近源区一侧引入低功函数隧穿栅电极的 DGM-DGTFET（M3 器件）和 TGM-DGT-FET（M4 器件）器件而言，当栅压≥ 0.5V 时，两者的漏电流随栅压的增长幅度大致相同，且漏电流在栅压为 1.5V 时较 M1 器件提高了 12 倍。此外，通过在 M4 器件栅极靠近漏区一侧引入低功函数辅助栅电极，可有效抑制双极效应；而对于无此结构的 M3 器件，在栅压负向增长时，漏电流也随之增大，产生了双极电流，如图 2 中虚线圈内数据所示。由此可以说明，对于双栅结构的隧穿场效应晶体管，采用重掺杂 pocket 层虽然抑制了双极效应，提高了开态电流，但随着栅压的增长，这一优势逐渐减弱，同时关态电流提高，降低了器件电流开关比。而在栅极同时引入隧穿栅电极和辅助栅电极，可以在大幅度提高器件开态电流和降低亚阈值摆幅的同时有效抑制双极效应，其电流开关比数量级达到了 10^{11}。

理解带带隧穿机制和观测隧穿路径及窗口变化的最有效方法为分析器件能带分布变化，如图 3 所示，为以上四种 DGTFET 器件在不同电压偏置下由器件源区至漏区的能带分布图，其中切线位置为体硅区域的纵向中心。如图 3(a) 所示，为关态 (OFF-state：V_g=0V，V_d=0.5V) 条件下器件的能带沿 x 轴的分布曲线，从图中可以看出 M2 和 M4 器件均有抑制双极效应的功能。这是由于，在 M2 中引入 pocket 层区域，使得源区靠近体区一侧能带有明显的弯曲，在此增大的了该处势垒。而在 M4 中引入辅助栅电极，使得在体区与漏区之间的能带向下弯曲，如图中箭头所示，当栅压负向增长时，这一能带弯曲会增大漏至体间的势垒距离，阻碍电子由漏区逆向输运，抑制了双极效应。图 3(b) 为开态 (ON-state：V_g=1.5V，V_d=0.5V) 条件下器件的能带图分布，从图中可以看出，三种改进器件在源 - 体结附近的能带均有明显的弯曲，缩短了隧穿距离，如图 (b) 中小图所示，促使更多的电子由源区的价带隧穿进入体区的导带，可有效增大开态电流。



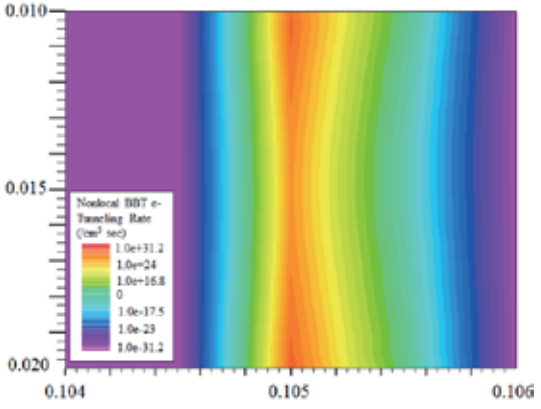
(a)



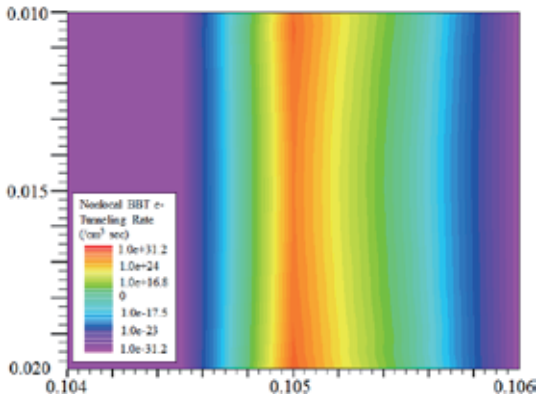
(b)

图 3 四种双栅隧穿场效应晶体管模型的能带图：(a) 关态 (V_d=V_g=0V) (b) 开态 (V_d=0.5V，V_g=1.5V)

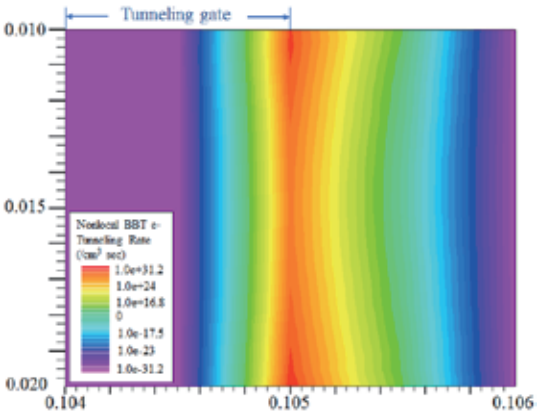
Fig.3 The energy band diagrams in the four TFETs from body to drain region: (a) OFF-state(V_d=V_g=0V) (b) ON-state(V_d=0.5V，V_g=1.5V)



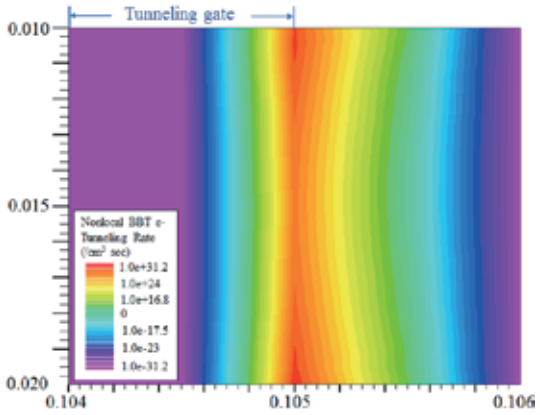
(a)



(b)



(c)



(d)

图 4 器件处于开态（V_d=0.5V 且 V_g=1.5V）条件时体硅中电子带带隧穿二维几率分布：(a) Co-DGTFET (b)DGTFET with pocket (c) DGM-DGTFET (d) TGM-DGTFET

Fig.4 Simulated diagrams of electron BTBT generation rate at V_d=0.5V and V_g=1.5V: (a) Co-DGTFET (b)DGTFET with pocket (c) DGM-DGTFET (d) TGM-DGTFET

为了进一步解释 M2、M3 和 M4 器件对开态电流提高能力的差异，本文分析了器件内部电子隧穿几率的分布。如图 4 所示，为四种 DGTFET 器件在偏压条件为 V_d=0.5V，V_g=1.5V 时电子隧穿几率的二维分布图，图中显示隧穿几率的峰值范围位于器件体区靠近源区一侧，且由于双栅增强了器件前背沟道处电场强度，使得高隧穿几率分布区域可以贯穿整个纵向体区。从图中还可以明显看出，pocket 层和隧穿栅均会促进隧穿几率的增强及高隧穿几率范围的加宽，但对比 M2 与 M3 和 M4 器件的隧穿几率峰值发现，隧穿栅极导致隧穿几率峰值更大，其位于隧穿栅电极和控制栅电极之间，x=0.105μm 处靠近前背栅区域。这一现象是由于该区域处电场强度被有效增强，如图 5 所示，为沿器件 x 方向沟道中心处电场强度分布图，其中偏置状态与图 4 一致。从图中可以清晰的看出，由于功函数的降低，隧穿电极有效了增强了源 - 体结处电场分布，如图 5 中小图所示的 M3 和 M4 所对应分区曲线。而 pocket 层会使得电场峰值转移进入源区，受栅极电压偏置的影响会减弱。因此，对于双栅结构 TFET 器件而言，提高器件直流特性的方法，在体区靠近源区一侧引入低功函数栅电极比在源 - 体之间加入势垒层更为有效。

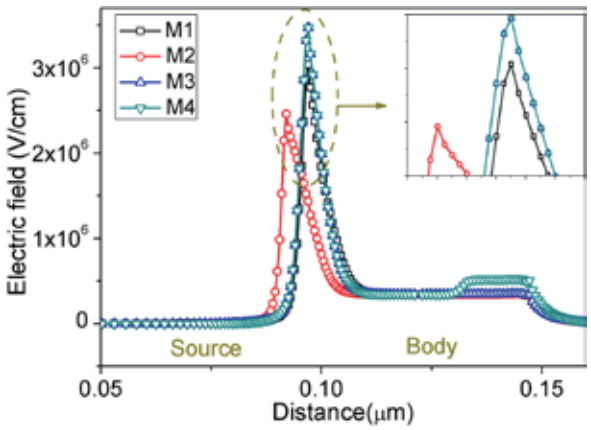


图 5 四种双栅隧穿晶体管模型处于开态（V_d=0.5V 且 V_g=1.5V）条件时电场分布图

Fig.5 1-D profile of the electric field for the four TFETs in a cut-line along the x axis at the center of channel region in the ON-state(V_d=0.5V and V_g=1.5V)

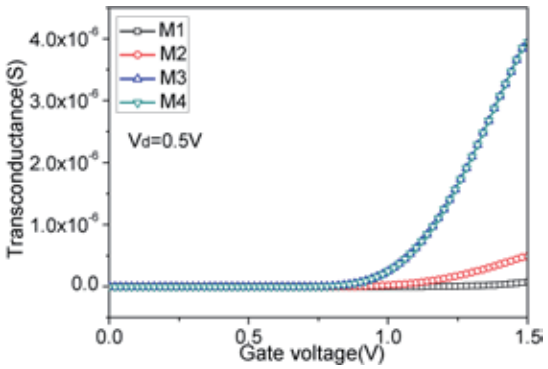
2.2 隧穿场效应晶体管频率特性的对比研究

在高速运算时代，频率特性是评价器件性能的重要指标。本小节通过跨导（g_m）、输出电导（g_{ds}）、截止

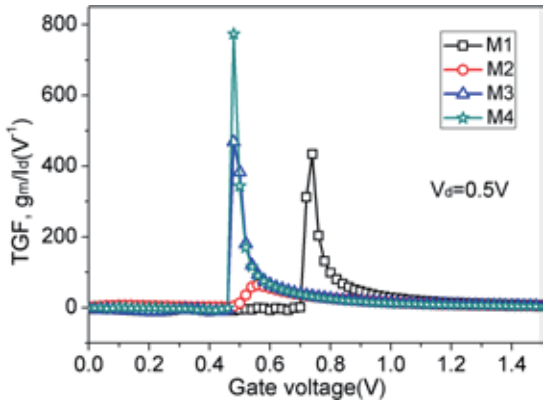
频率 (f_T) 和增益带宽积 (gain bandwidth product, *GBP*) 等频率特性表征参数的对比分析，研究以上四种双栅隧穿场效应晶体管的频率特性。

器件跨导是模拟电路设计中运算放大器和运算跨导放大器的关键参数，可影响 DC 增益、频率特性、以及噪声特性等，跨导的提高可以改善器件的放大能力和截止频率 [26]。跨导可以反映栅压对器件漏电流的控制能力，可在给定漏极偏置下，通过漏电流对栅－源电压的一阶偏导求得，即

$$g_m = dI_{ds} / dV_{gs} \tag{2}$$



(a)



(b)

图 6 漏极偏置为 0.5V 时四种双栅隧穿晶体管 (a) 跨导和 (b) 放大系数随栅压的变化曲线

Fig.6 Simulated (a) transconductance and (b) amplification factor as a function of gate voltage for the four TFETs

漏极偏置为 0.5V 时四种 DGTFET 器件的跨导随栅压的变化曲线如图 6(a) 所示。随着栅压的增大，漏电流会呈指数增长，这会导致跨导也随之指数增大，且 M2、M3 和 M4 器件的跨导均较传统双栅隧穿场效应晶体管的有所提高。其中，当栅压为 1.5V 时，较传统器件的跨导，使用 pocket 层的 M2 器件可将跨导提高 6.5 倍，使用隧穿栅极的 M3 和 M4 则可将跨导提高两个数量级。

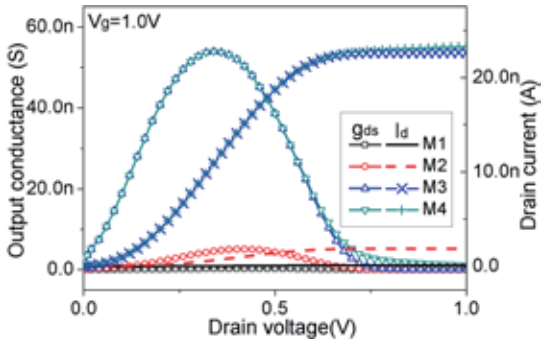
频率特性的另一个关键参数，跨导产生因子 (transconductance generation factor, *TGF*) 可通过跨导和漏电流比 ($=g_m/I_{ds}$) 得到，用于表征器件直流特性的漏极电流有效转化为频率参数跨导，因此，提高 *TGF* 意味着改善单位漏电流的放大能力，更适合于低功耗和高增益的模拟电路；而较低的 *TGF* 则表示降低了输入信号驱动能力，增大电路功耗。如图 6(b) 所示，为以上四种器件的跨导产生因子随栅压的变化曲线。从图中不难看出，在栅压较低（对应于器件转移特性曲线的亚阈值区域）时，会出现 TGF 的极大值，其中，M4 器件拥有最大的 TGF_{max}=773.6V^{-1}，而 M2 器件的 TGF_{max}(=63.5V^{-1}) 最小。随着栅压的进一步增大，跨导品质因子减小，这是由于亚阈值摆幅逐渐增大导致的。因此，有着隧穿栅电极和辅助栅电极的 M4 器件更适合于低功耗和高增益的模拟电路应用。

如图 7(a) 和 (b) 所示，分别为栅极偏置为 1.0V 和 0.6V 时，四种双栅隧穿场效应晶体管漏电流和输出电导随漏压的变化曲线，其中，输出电导可通过公式 (3) 计算得到，

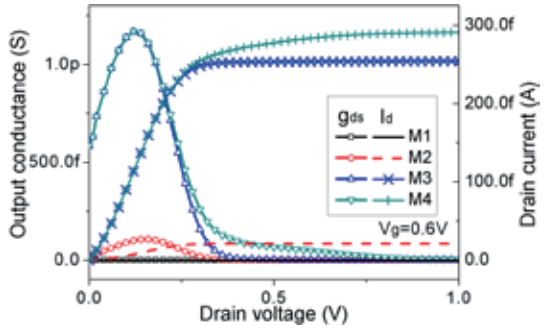
$$g_{ds} = dI_{ds} / dV_{ds} \tag{3}$$

从图 7(a) 中可以看到，改进后的新型器件的漏电流较传统双栅隧穿晶体管均有明显的提高，且 M3 和 M4 器件的输出特性要优于 M2 器件，有着最大饱和漏电流和夹断电压。对于隧穿晶体管而言，沟道电阻的增大，亦可说沟道中电子浓度的减小，会使得器件的工作状态更早进入饱和区，此时 TFET 器件的漏电流主要由源至沟道间带带隧穿机制产生，降低了漏电压对漏电流的控制能力。正如前文分析可知，使用隧穿栅电极后器件的隧穿几率的提高要优于在源区和体区间插入势垒层，因此，M3 和 M4 器件的沟道电子浓度是高于 M2 器件的，有着较小沟道电阻的 M3 和 M4 会推迟进入饱和区。同时，增强的隧穿几率致使 M3 和 M4 的漏电流最大。当栅压减小，器件会更

早进入饱和区，如图 7(b) 所示，这是因为隧穿几率强烈依赖电场，当栅压减小，隧穿结处电场减弱，隧穿机制的削弱会使得沟道电阻提高。当栅压由 1.0V 减小至 0.6V，器件的饱和区域在向左移动。



(a)



(b)

图 7 四种双栅隧穿晶体管漏电流和输出电导随漏压的变化曲线：(a) 栅压为 0.6V (b) 栅压为 1.0V

Fig.7 Simulated drain current and output conductance as a function of drain voltage for the four TFETs: (a) V_g=0.6V (b) V_g=1.0V

从图 7(a) 和 (b) 中还可以看到，随着栅压的提高，输出电导增大，这与前文分析得到的电流变化相一致。当输出特性处于线性区时，输出电导的增大是由于输出特性曲线斜率的增大；当处于饱和区时，漏压对漏电流的控制能力减弱，输出电导逐渐减小，当漏压进一步增大，输出电导并无明显变化。从图中器件的表征参数变化对比不难看到，在相同的栅极和漏极偏置下，由于增强的带带隧穿机制，M3 和 M4 的输出电导是大于 M2 和 M1 器件的。

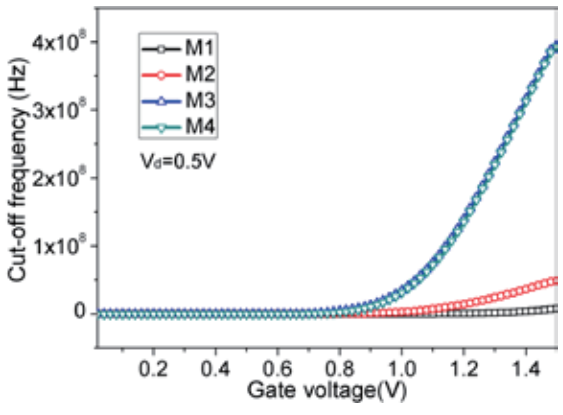
在无线通信应用领域中，器件的截止频率、增益带宽积和渡越时间 (τ) 是重要的频率表征参数。截止频率可通过公式 (4) 计算得到，

$$f_T = g_m / 2\pi C_{gg} \tag{4}$$

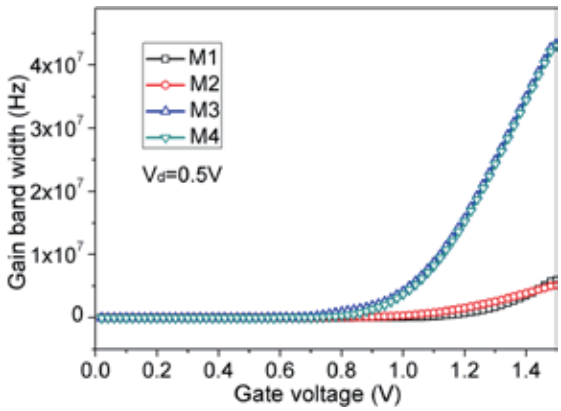
其中，C_{gg} 是器件栅电容，可近似为栅 − 源电容和栅 − 漏电容之和。如图 8(a) 所示，为漏极偏置 0.5V 时，四种双栅隧穿场效应晶体管的截止频率随栅压的变化曲线。从图中可以看到，随着栅压的增大，截止频率逐渐增大，这主要受器件跨导随电压的变化趋势所影响。从图中还可以看到，当栅压增长至 1.5V 时，传统双栅 TFET 器件的截止频率仅为 7.7×10^6Hz，使用 pocket 层的 M2 器件的截止频率提升至 4.9×10^7Hz，而使用隧穿栅电极的 M3 和 M4 器件则将截止频率提升至了 3.9×10^8Hz，大大改善了传统双栅 TFET 的频率特性。通过公式 (5)，本文还计算得到了增益带宽积随栅压的变化情况，

$$f_A = g_m / 2\pi 10 C_{gd} \tag{5}$$

其中，C_{gd} 为栅 − 漏电容。如图 8(b) 所示，器件的偏置条件与截止频率的计算情况相同，M3 和 M4 器件的增益带宽积与截止频率的变化趋势一致，当 $V_g \geqslant 1.0$V 时，该值随着栅压呈指数增长。当栅压为 1.5V 时，M3 和 M4 器件的增益带宽积要远大于另外两个器件。联合以上对这四种器件的直流特性对比研究可知，使用隧穿栅电极可有效提高器件的电特性和频率特性，对于双栅结构的 TFET 而言，较在源区和体区之间增加势垒层更为有效。



(a)



(b)

图 8 漏极偏置为 0.5V 时四中双栅隧穿晶体管 (a) 截止频率和 (b) 增益带宽积随栅压的变化曲线

Fig.8 The (a) cut-off frequency and (b) gain bandwith products as functions of gate voltage for the four TFETs

3 结论

本文仿真研究了新型掺杂型双栅隧穿场效应晶体管器件的直流特性和频率特性。通过在源区和体区之间加入重掺杂势垒层（与源区掺杂类型相反）的改进措施，或基于在器件前、背栅均使用两种功函数的栅材料共同完成栅极驱动的方法，构造出两种新型 DGTFET，并与传统 DGTFET 器件的电特性进行了对比研究。首先，通过器件的关态电流、开态电流和电流开关比等直流特性表征参数对三种结构器件的进行了研究，发现在栅极靠近源区一侧和靠近漏区一侧使用较低功函数的栅材料方法，大大提高了器件性能，当栅压为 1.5V 且漏压为 0.5V 时，该器件较传统双栅隧穿场效应晶体管的开态电流提高了 12 倍，电流开关比数量级可达到 10^{11}。这主要是由于隧穿栅电极较势垒层的引入可有效提高源－体界面处的电场强度，增大能带弯曲，缩小隧穿路径，促进带带隧穿机制所导致的。此外，又通过跨导、输出电导、截止频率和增益带宽积对以上三种结构器件的频率特性进行了仿真研究。研究结果表明，三种栅电极的器件 TGM-DGTFET 频率特性最为突出，且该器件的截止频率和增益带宽积均较传统 DGTFET 器件提高了 50 倍。以上分析表明，对于双栅结构的 TFET 而言，使用多种功函数构成栅电极的方法较在源区和体区之间增加势垒层提高器件频率特性更为有效。

参考文献 (References)

[1] IONESCU A M, AND RIEL H. Tunnel field-effect transistors as energy-efficient electronic switches[J]. Nature, vol. 479, pp. 329 337, Nov. 2011, doi: 10.1038/nature10679.

[2] JAIN G, SAWHNEY R S, KUMAR R, et al. Analytical modeling analysis and simulation study of dual material gate underlap dopingless TFET[J]. Superlattices and Microstructures, 2021, 153:106866.

[3] AVINASH L, AND MAMIDALA J K. The charge plasma n-p-n impact ionization MOS and FDSOI technology: proposal and analysis[J]. IEEE Transactions on Electron Device, 2017, 64(1): 3-7.

[4] GAURAV M, SHUBHAM S, RAGHVENDRA S S, et al. An impact ionization MOSFET with reduced breakdown voltage based on back-gate misalignment[J]. IEEE Transactions on Electron Device, 2019, 62(2): 868-875.

[5] ALOKK K, AND JAWARDS. Simulation-based ultralow energy and high-speed LIF neuron using silicon bipolar impact ionization MOSFET for spiking neural networks[J]. IEEE Transactions on Electron Device, 2020, 67(6): 2600-2606.

[6] OTA H, MIGITA S, HATTORI J, et al. Structural advantages of silicon-on-insulator FET over FinFETs in steep subthreshold-swing operation in ferroelectric-gate FETs[J]. Japanese Journal of Applied Physics, 2017, 56 (4s): 04CD10.

[7] JINDAL S, MANHAS S K, GAUTAM S K, et al. Investigation of gate-length scaling of ferroelectric FET[J]. IEEE Transactions on Electron Device, 2021, 68(3): 1364-1368.

[8] KASHYAP R, BAISHYA S, AND TAYE J. Design and simulation of a compact low-stiffness MEMS-gate for suspended-gate MOSFET[J]. International Journal of Advancements in Technology, 2014, 5(2): 126-136.

[9] SANGEETA J M, BUDHADITYA M, KARUMBAIAH N C, et al. Performance analysis of the diagonal tunneling based dielectrically modulated tunnel FET for bio-sensing application[J]. IEEE Sersors Journal, 2021,

doi：10.1109/JSEN.2021.3103998.

[10] PANDEY C K, SINGH A, CHAUDHURY S. Effect of asymmetric gate-drain overlap on ambipolar behavior of double-gate TFET and its impact on HF performances[J]. Applied Physics A, 2020, 126:225(1-12).

[11] KNOLL L, SCHMIDT M, ZHAO Q T, et al. Si tunneling transistors with high on-currents and slopes of 50 mV/dec using segregation doped NiSi2 tunnel junctions[J]. Solid-State Electronics, 2013, 84:211-215.

[12] ANAND I V, SAMUEL T, VIMALA T, et al. Modelling and simulation of hetero-dielectric surrounding gate TFET[J]. Jouranl of Nona Research, 2020, 62: 47-58.

[13] MITRA S K, GOSWAMI R, BHOWMICK B. A hetero-dielectric stack gate SOI-TFET with back gate and its application as a digital inverter[J]. Superlattices and Microstructures, 2016, 92: 37-51.

[14] HYUN W K, DAEWOONG K. Steep switching characteristics of L-shaped tunnel FET with doping engineering. IEEE Journal of the Electron Devices Society[J], 2021, 9: 359-364.

[15] DASH S, SAHOO G S, AND MISHRA G P. Improved cut-off frequency for cylindrical gate TFET using source delta doping [J]. Procedia Technology, 2016, 25: 450-455.

[16] DHEERAJ M, SAURABH M, ASHISH A, et al. Experimental Staggered-Source and N+ Pocket-Doped Channel III V Tunnel Field-Effect Transistors and Their Scalabilities.

[17] SAFA S, NOOR S L, AND KHAN M Z R. Triple material double gate TFET with optimized Si film thickness [C]. Dhaka: Electrical Engineering and Information Communication Technology (ICEEICT), 2016: 22-24.

[18] ABDI D B, KUMAR M J, In-built N+ pocket p-n-p-n tunnel field-effect transistor [J]. IEEE Electron Device Letters, 2014, 35(12):1170-1172.

[19] CUI N, LIANG R, WANG J, et al. Lateral energy band profield modulation in tunnel field effect transistors based on gate structure engineering [J]. AIP Advances, 2012, 2(2): 022111.

[20] RAAD B R, NIGAM K, SHARMA D, et al. Performance investigation of bandgap, gate material work function and gate dielectric engineered TFET with device reliability improvement[J]. Superlattices and Microstructures, 2016, 94: 138-146.

[21] LEE I Y, IM D. Low-power SOI CMOS antenna switch driver circuit with RF leakage suppression and fast switching time[J]. Electronics Letters, 2017, 53(5): 293-294.

[22] DENNARD R H, GAENSSLEN F H, Rideout V L, et al. Design of ion-implanted MOSFET's with very small physical dimensions[J]. IEEE Journal of Solid-State Circuits, 1974, 9(5): 256-268.

[23] BOUCART K, IONESCU A M. Double-gate tunnel FET with high-k gate dielectric[J]. IEEE Transactions on Electron Devices, 54: 1725-1733.

[24] KNOCH J, MANTL S, AND WASER R. Nanoelectronics and Information Technology[M]. Wiley-VCH, 3rd edition, 2012: 347.

[25] QUERLIOZ D, SAINT-MARTIN J, HUET K, et al. On the ability of the particle Monte Carlo technique to include quantum effects in nano-MOSFET simulation[J]. IEEE Transactions on Electron Devices, 2007, 54(9): 2232-2242.

[26] KONDEKAR P N, NIGAM K, PANDEY S, et al. Design and analysis of polarity controlled electrically doped tunnel FET With bandgap engineering for analog/RF applications[J]. IEEE Transactions on Electron Devices, 2017, 64(2): 412-418.

作者简介：

王倩琼（1987—），女，陕西西安人，博士，工程师，从事纳米级新型元器件可靠性分析、数模混合集成电路设计、ASIC/SoC 设计。

面向通用高性能数字处理平台的电源启动时序控制电路

马 婷，龚 科，刘 洁，李文琛，王江涛，高玉龙，戴 璐，邢建丽

（中国空间技术研究院（西安分院），陕西省 西安市 710100）

摘　要：针对复杂高速数据处理系统数字处理器与数模混合器件的多电源电压启动时序设计特殊要求，提出一种面向高性能数字处理平台的启动时序电源管理方法，设计了分立式无源延迟网络，建立电源启动自反馈网络，实现各电源模块使能端精确时序控制，优化了数模混合类处理系统的整体供电拓扑，相比传统利用可编程逻辑器件实现电源时序控制的方法，该方法电源启动时序控制精确，电路简单可靠，降低系统性能对单一控制器的依赖，对星上通用高性能计算平台、软件定义硬件平台等复杂高速数据处理系统具有很高的工程应用价值。

关键词：电源管理；上电时序；软启动；延迟电路

中图分类号：U262.7+3　　**文献标识码**：A

A Control Circuit of Powerup Sequence for General High-performance Digital Processing Platform

Ma Ting, Gong Ke, Liu Jie, Li Wenchen, Wang Jiangtao, Gao Yulong, Dai Lu, Xing Jianli

（China Academy of Space Technology(Xi'an Branch), Xi'an, 710100, China）

Abstract: Aiming at the special requirements of multi power supply voltage start-up timing design of digital processor and digital analog hybrid devices in complex high-speed data processing system, a start-up timing self-control power management method for high-performance digital processing platform is proposed, a discrete delay network is designed, a power supply start-up self-feedback network is established, and the accurate timing control of each power module is realized, The overall power supply topology of the digital analog hybrid processing system is optimized. Compared with the traditional method of using programmable logic devices to realize power supply timing control, this method has accurate power supply startup timing control, simple and reliable circuit, reduces the dependence of system performance on a single controller, and has high engineering application value for general high performance computing platform, software definition hardware platform and other complex high speed data processing system on board.

Key words: power management; power up sequence; soft start; delay circuit

0 引言

出于对产品性能和成本的追求，近年来地面设备及航天载荷电子系统均呈现出“软件化”的发展趋势，硬件实现最大范围覆盖，由软件体现差异化服务，具有通用性高的特点。同时也加速了硬件系统复杂性升级，对硬件系统的可靠性保障、噪声管理、功耗控制等方面提出了设计挑战。电源管理模块作为系统的核心组成，与系统可靠性及产品功耗息息相关，电源管理技术成为通用硬件系统设计的关键技术之一[1-4]。

现有电源管理方法主要分为两类，有源数字控制法和软启动调节法。有源数字控制法是指通过可编程逻辑器件产生具有特定时序的数字电平信号，控制各个电源模块使能端，精确控制各电源模块启动时序。该方法需要增加额外的可编程器件，以及专用供电电路，形成系统控制拓扑的单点，降低系统可靠性。软启动调节法通常选用具有专门软启动功能的供电模块，控制该功能引脚电压上升斜率，调节各输出电压上升速度，实现各电源模块的启动时序控制。该方法需要依赖电源芯片的软启动功能引脚，使设计的器件

选型受限，同时过度延缓启动时间容易使电源启动受到干扰[5-7]。

1 多电源启动时序

大规模高速数字处理系统电路包含FPGA、CPU、ASIC等，它们之间供电电压以及上电时序均不同，且同一器件的核电和IO电之间上电时序也有严格要求。图1为一款ASIC芯片上电时序图，供电电压为IO的VDD33，3.3V SerDes的AVDH，1.2V SerDes的AVDL，1.2V内核的VDD，延迟间隔大于1ms，各电压之间严格的上电时序对电源设计提出更高要求。

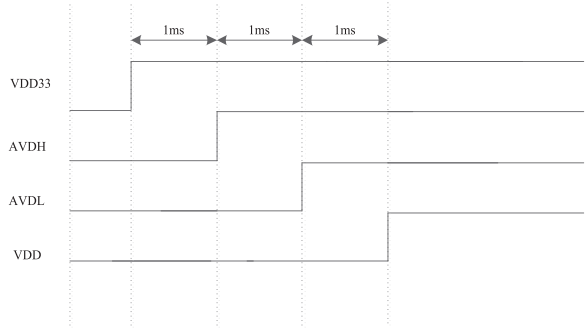


图1 ASIC上电时序图
Fig.1 Power up sequence of ASIC

1.1 有源数字控制法

有源数字控制法是指通过CPLD可编程逻辑器件产生具有特定时序的数字电平信号，控制各个电源模块使能端，达到精确控制系统各电压上电时序的要求，如图2所示。

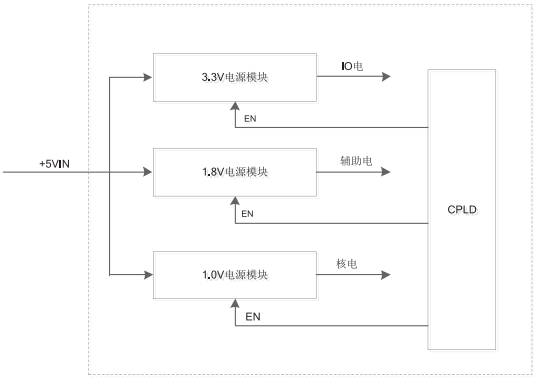


图2 CPLD控制电源上电图
Fig.2 Power up by CPLD

1.2 软启动调节法

软启动调节法通常选用具有专门软启动功能的供电模块，控制该功能引脚电压上升斜率，调节各输出电压上升速度，实现各电源模块的启动时序控制。如图3所示，选择具有软启动功能的电源模块，设计延迟电路控制电源模块的软启动引脚SS，调节各输出电压的上升速度，控制各电源模块的启动时间，实现系统各电压上电时序的精确控制。

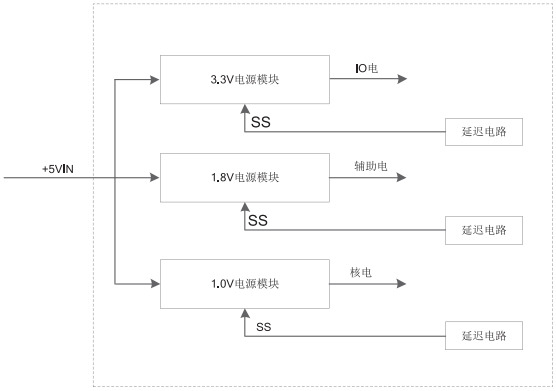


图3 延迟电路控制电源上电图
Fig.3 Power up by soft start circuit

2 启动时序电源管理方法

数字处理器电源种类多、功耗大，采用高转换效率的DC/DC电源模块为其供电；高速数模混合器件A/D、D/A工作需要同时提供数字电和模拟电，数字电和模拟电需单独供电。A/D、D/A器件模拟电对电源噪声异常敏感，为保证不受电源噪声影响，采用线性稳压器LDO为其提供稳定低噪声模拟电压。

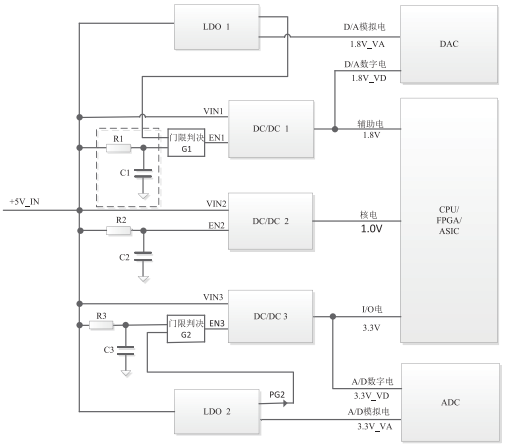


图4 电源供电系统拓扑图
Fig.4 Power supply system topology

为简化电源管理电路设计，本文提出一种启动时序电源管理方法。如图 4 所示，数字处理器辅助电、核电以及 I/O 电分别通过 DC/DC1、DC/DC2、DC/DC3 提供。D/A 数字电直接采用数字处理器辅助电，模拟电通过线性稳压器 LDO1 转换后提供；A/D 数字电直接采用数字处理器 I/O 电，模拟电通过线性稳压器 LDO2 转换后提供。通过设置 RC 延迟电路的延迟时间间隔控制 DC/DC 电源模块电使能信号，同时在 LDO 与门电路之间建立电源启动自反馈网络，实现各电源模块精确时序控制，满足处理器上电时序要求。

DC/DC 电源模块输入电压为 +5V，DC/DC1 输出 1.8V，DC/DC2 输出 1.0V，DC/DC3 输出 3.3V，分别为处理系统提供辅助电、核电电以及 I/O 电。每个 DC/DC 电源模块可通过控制使能端 EN 电平调整电源模块输出电压，当 DC/DC 使能端电压不低于使能端门限电压时，器件开机正常工作，否则器件停止工作，无电压输出。对多个电源模块使能端引脚引入 RC 延迟电路，根据处理系统上电时间间隔要求设计 RC 延迟电路，满足不同处理系统上电时序要求。

根据公式 1 计算各电源上电时延：

$$T=-RC\cdot\ln((E-V)/E) \qquad (1)$$

其中 E 为电阻 R 和电容 C 之间的电压，V 为电容要达到的电压。通过上述计算，R、C 可以选取合适的值，满足不同处理系统多种电压正常工作电源时序要求。

A/D 数字电 3.3V_VD 由 DC/DC3 转得到换得到，模拟电 3.3V_VA 由 LDO2 转换得到。A/D 内部自身存在潜通路，3.3V_VA 上电后，器件内部的数字电路导通，造成 A/D 数字电 3.3V_VD 电压上升曲线不单调，该电压与处理系统的 I/O 电为相同电源网络，导致处理系统的 3.3V_IO 电压上升曲线不单调，不满足数字处理器上电要求。

D/A 数字电 1.8V_VD 由 DC/DC1 转换，模拟电 1.8V_VA 由 LDO1 转换。D/A 内部自身存在潜通路，1.8V_VA 上电后，器件内部的数字电路导通，造成 D/A 数字电 1.8V_VD 电压上升曲线不单调，该电压与处理系统用 1.8V 电为相同电源网络，导致处理系统的 1.8V 电压上升曲线不单调，不满足数字处理器上电要求。

为解决该问题，将判决电路 G1 和 G2 分别接入到开关电源 DC/DC1 和 DC/DC3 使能端，根据开关电源使能端门限电平调节 RC 延迟网络中 R 和 C 的参数，确保 RC 延迟网络输出电压上升时间满足数字处理器电源启动时序要求。将 LDO 输出电压正常的状态信号 PG，作为反馈引入到 DC/DC 使能端门电路，LDO2 反馈信号 PG2 与 RC 电路输出通过门电路 G2 进行逻辑转换，再引入 DC/DC3 的使能引脚 EN3，确保 LDO2 输出的 A/D 模拟电稳定后，再输出 A/D 数字电；同理，将 LDO1 的 PG1 信号与 DC/DC1 的 RC 延迟电路通过门电路 G1 进行逻辑转换，再引入到 DC/DC1 的使能引脚 EN1，确保 LDO1 输出的 D/A 模拟电稳定后，再输出 D/A 数字电。通过该反馈网络有效解决因潜通路导致数模混合器件数字电电压上升曲线不单调而引起处理系统电源电压不满足上电时序的问题。

3 实验结果

本文设计的启动时序电源管理方法已成功应用于一款高速数模混合数据处理系统产品电源管理电路。

该系统选用 ASIC 处理器，三种供电电压上电时序为：1.2V、1.8V、3.3V，各电压启动间隔不少于 3ms。DC/DC2 上电后延时 0ms，DC/DC1 上电后延时 4.7ms，DC/DC3 上电后延时 8.3ms，满足三种电启动间隔不小于 3ms 的要求，实现各电源模块上电时序精确控制，如图 5 所示。各电压启动上升曲线单调有序，同时解决因潜通路导致数模混合器件数字电电压上升曲线不单调问题，如图 5 所示。

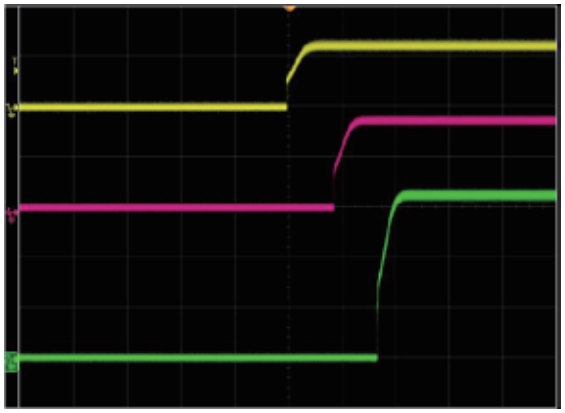


图 5 ASIC 电压上电时序
Fig.5 Power up timing of ASIC voltage

4 结束语

综上所述，本文提出一种启动时序电源管理方法，由分立式无源延迟网络，实现各电源模块使能端精确时序控制，降低了对电源模块软启动引脚的依赖性；优化了数模混合类处理系统的整体供电拓扑，通过自反馈的方式，避免了各类供电间的启动干扰，提高了系统的可靠性。该方法电路简单可靠，具有很高的工程应用价值。

参考文献（References）

[1] Dynamic Power Management for Embedded Systems (Ver1.1)[Z]. IBM.monta Vista Inc. 2002-11-09.

[2] 王洁．电源管理技术在新环境下程序新态势[J]. 电子技术应用，2015，41(5)：19-20.

[3] 赵霞，陈向群，郭耀等．操作系统电源管理研究进展．计算机研究与发展，200，45(5)：817-824.

[4] 钱威，张国勇，董房等．上电时间对卫星单机射击的影响分析．航天器环境工程，2016，33(1)：82-85.

[5] PENZIN S H，CRAWFORD K B，et al. The SEU pulse width modulation controllers with soft start and shutdown circuits. Radiation Effects Data Workshop，IEEE，Issue，1997：73.

[6] Texas Instruments.TPS6527x Single-Chip PMIC for Battery Powered Systems [DB/OL]. (2011-11-17) [2017-6-28]. http://www.ti.com/cn/lit/gpn/tps6527.pdf.

[7] Powering Cool Runner-II CPLDs [Z]. Xilinx Inc. 2003-05-19.

[8] 赵秋明，丁云，欧阳宁等．一种软件无线电平台的电源系统设计与实现[J]. 电源技术研究与设计，2012，36(4)：561-563.

[9] 何允灵，秦捐，王佳等．SoC 处理器电源管理系统设计[J]. 计算机工程，2008，34(16)：262-264.

作者简介：
马婷（1983—），女，工程师，主要研究方向为卫星数字通信技术、高速数模混合电路设计技术等。

NAND Flash 抗辐射加固措施

朱新忠[1]，吴振广[1]，王 琴[2]，白 郁[1]，杨伟东[2]

（1. 上海航天电子技术研究所，上海市 201109；2. 上海交通大学，上海市 200240）

关键词：NAND Flash；配置区刷新；单粒子；抗辐射加固

中图分类号：TN47　　**文献标识码**：A

宇航存储产品是卫星关键产品，随着载荷种类的增多，数据速率的增加，对宇航存储的容量和吞吐速率要求也越来越高。NAND Flash 具有非易失、读写速度快、存储密度大和较好的抗震性等优点，广泛应用于宇航存储器[1]。

美国国家地球物理数据中心统计，空间辐射效应是使得航天器发生故障的主要因素。同时存储器在一个芯片中占据了 30% 的面积，而在集成电路系统级芯片中存储器所占面积更是超过了 60%[2]。随着宇航技术的飞速发展，对宇航系统可靠性的要求越来越高。如何提高宇航存储的可靠性一直都是宇航领域的关注焦点[3]。

3D-PLUS 公司的立体叠装技术把筛选后的半导体器件叠装到一个高度微型化的封装里，能突破容量及体积限制，在体积上至少缩小十倍，而容量上至少增加十倍。本文以 3D-PLUS 公司的 128Gb NAND Flash 器件 3DFN128G08 为例，其内部由 8 片经过筛选的 Micron 工业级 NAND Flash 芯片 MT29F16G08 叠装而成。3DFN128G08 具有 10 万次编程 / 擦周期，10 年数据可靠驻留时间，其抗辐射指标见表 1。

表 1 3DFN128G08 器件抗辐射指标

Tab.1 3DFN128G08 device radiation hardness

序号	辐照参数	指标
1	TID	60Krads(Si)
2	SEL	>62.5MeV · cm²/mg
3	SEU	1.3MeV · cm²/mg
4	SEFI	1MeV · cm²/mg

MT29F16G08 采用 32nm 工艺尺寸，存储单元为 SLC 型晶体管复合栅结构，如图 1 所示。

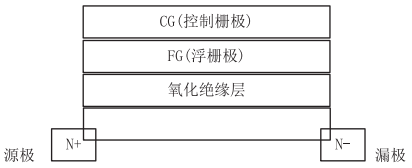


图 1 NAND Flash 存储单元结构

Fig.1 NAND Flash storage unit structure

在源极和栅极的基础上增加浮栅极 FG，在使用过程通过浮栅极的电子含量来判断晶体管的状态。

为了保证在轨可靠性、编程的灵活性，宇航存储产品普遍采用抗辐射 FPGA 完成存储芯片控制和数据管理。以宇航存储产品设计为例，如图 2 所示。

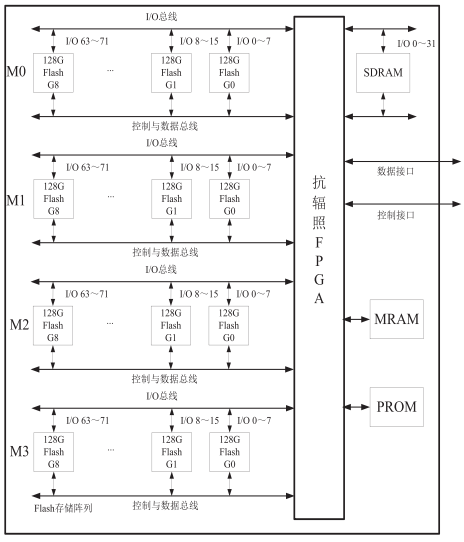


图 2 宇航存储产品

Fig.2 Space storage product

其中控制单元采用抗辐射的 A54SX72A 反熔丝 FPGA，主要完成对存储单元的各种底层接口控制和纠错编译码；存储单元共分为四个子模块，每个子模块由 9 片 128Gb 的 3Dplus 公司三维封装的 Flash 芯片并行级联而成，其中 8 片作为数据存储、1 片作为校验，级联设置开关可以设置四组存储子模块的级联关系（深度扩展或位宽扩展），从而实现存储容量和存储速度的扩展；每组存储子模块可支持 640Mbps 存储速率，容量为 1Tb（不含校验），因此该宇航存储产品最大可支持 2.56Gbps 存储速率，存储容量最大可达 4Tb。

单粒子翻转(SEU)效应是影响 Flash 存储器错误的主要问题之一。当翻转发生在关键位置时，可能导致整个存储器件错误。因此，必须对 Flash 的单粒子翻转效应进行充分测试分析和研究。

Flash 基片 MT29F16G08 上电需进行复位操作，并增加了状态配置功能[4]：时序配置、OTP（one time program）设置等。NAND Flash 芯片内部主要由指令控制电路（指令锁存和指令译码电路）、地址译码电路、数据接口电路（数据锁存和驱动控制电路）和数据存储阵列等功能块组成。具有配置功能的 Flash 芯片还有功能配置电路，以调整指令控制电路和数据接口电路部分设置。

在空间环境中，指令锁存电路、地址锁存电路、数据锁存电路和数据存储阵列均有可能发生单粒子翻转，但对电路的影响不同。

表 2 功能电路单粒子翻转影响分析
Tab.2 Analysis of single particle flip effect of functional circuit

电路类型	单粒子故障影响	纠正措施
指令锁存电路	指令翻转，错误操作或不执行操作	自动恢复或状态重置
地址锁存电路	地址位翻转，读写错误页地址，或擦除错误区块地址	自动恢复；误码较多，采用强校验方能纠正
数据锁存电路	存储数据翻转，导致误码	自动恢复，校验纠正
数据存储阵列	数据翻转，导致误码	自动恢复，校验纠正

由表 2 对比结果可知，单粒子翻转发生在指令锁存电路上将给整个芯片带来严重的影响。按照 MT29F16G08 芯片资料配置指令的定义如表 3 所示。

表 3 MT29F16G08 配置命令定义
Tab.3 MT29F16G08 configuration command definition

配置功能地址	定义
00h	预留
01h	时序模式
02h−0Fh	预留
10h	可编程输出驱动强度
11h−7Fh	预留
80h	可编程输出驱动强度
81h	可编程 RB# 信号下拉强度
82h−8Fh	预留
90h	数组操作模式
91h−FFh	预留

从表 3 可以看出，配置位主要有 0x01h，0x10h，0x80h，0x81h 和 0x90h 这几个地址，每个地址的配置操作时序如图 3。

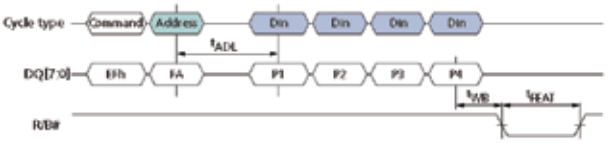


图 3 芯片配置操作时序
Fig.3 Chip configuration operation timing

在 20MHz 频率下，通过模拟基片 MT29F16G08 配置命令翻转测试如表 4。

表 4 模拟测试结果
Tab.4 Simulation test results

配置模式	操作	结果
地址 01h	模拟子功能参数 DQ[2:0] 翻转	误码
	正常擦，更改 DQ[0] 写，正常读	影响写操作
地址 90H	正常擦，正常写，更改 DQ[0] 后读	影响读操作
	更改 DQ[0] 后擦，直接读数	影响擦除操作
其它配置位	更改后读配置位信息	保留区未被更改

由表 4 可知，影响 Flash 芯片操作的主要因素为：时序模式命令地址 01h 的子功能参数 P1 中的数据位 DQ[2:0] 和 OTP 模式命令地址 90h 的子功能参数 P1 中的数据位 DQ[0]。这 4bit 其中某一位发生翻转均会导致该芯片功能异常，因此针对 MT29F16G08 配置

命令翻转的抗辐射加固措施十分必要。

宇航 NAND Flash 存储产品的可靠性设计可以采取多种措施防止空间辐射效应的影响或者降低其影响的概率。例如在硬件上，使用具有抗辐射能力的器件设计宇航存储产品，同时采用双机冷备或者热备的工作模式；在 FPGA 设计上，对 Flash 进行坏块管理、EDAC 校验等。根据 Flash 的单粒子翻转效应分析结论：单粒子翻转发生在关键位置时尤其是芯片的状态模式配置命令上，可能导致 Flash 致命错误。基于 Flash 配置命令翻转导致芯片功能异常的问题，提出了一种 Flash 配置命令区的频繁刷新的抗辐射加固措施。Flash 芯片读和写基本操作是按页进行的，最频繁的刷新操作便是每次页写和页读前均进行一次刷新配置，保证配置命令的可靠性。改进后的配置刷新流程如图 4 所示。

增加配置刷新操作后的时间开销为 3μs，每次页操作时间从 0.4096ms 增加到 0.4126ms，因配置刷新时间开销增加 0.73%，对页操作的影响很小，增加页操作前的刷新配置流程不影响整机的读写速率，方案可行。

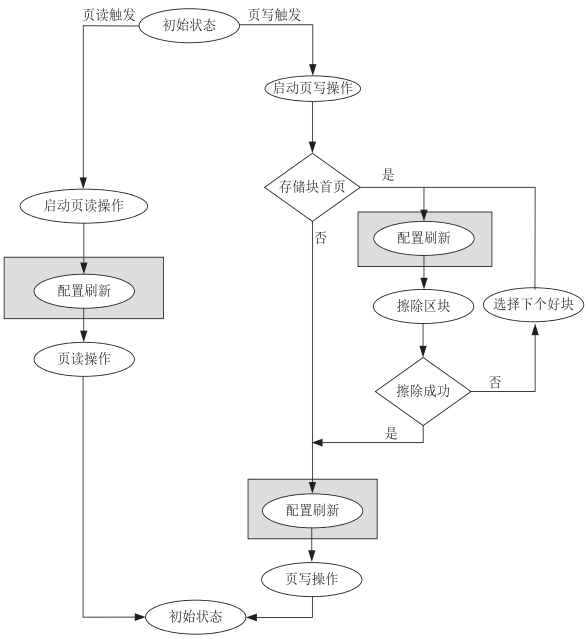


图 4 改进后的配置刷新流程

Fig.4 Improved configuration refresh process

单粒子翻转率是指器件每天每位发生单粒子翻转或错误的概率。单粒子翻转率是评价宇航产品可靠性的主要指标之一。国际上普遍采用 FOM 法 [6] (Figure Of Merit，FOM)，FOM 可以由重离子实验数据得到，也可以由质子实验数据计算得到，本文以位翻转截面为单粒子翻转极限截面，计算公式如下：

$$\begin{aligned} FOM &= \sigma_h / L_{0.25}^2 \\ &= 4.5\times10^4 \times \sigma_p(\infty) \end{aligned} \quad (1)$$

其中σ_h为重离子单粒子翻转饱和截面，$L_{0.25}^2$为单粒子翻转饱和截面的 25% 处所对应的重离子的 LET 值，$\sigma_p(\infty)$为质子单粒子翻转极限截面，C 为轨道翻转率系数。由此可以计算单粒子翻转率 R。

$$\begin{aligned} R &= \mathrm{C}\times FOM \\ &= 4.5\times10^4 \times \mathrm{C}\times\sigma_p(\infty) \end{aligned} \quad (2)$$

MT29F16G08 基片中的 Flash 配置命令区共 160bits，相关状态翻转可能影响读写功能的配置位只有 4bits，因此基片的翻转概率 R_G 为：

$$\begin{aligned} R_G &= 4R/160 \\ &= 1.125\times10^3 \times \mathrm{C}\times\sigma_p(\infty) \end{aligned} \quad (3)$$

该卫星运行在太阳同步轨道 700 ~ 900km，C 取值 490，$\sigma_p(\infty)$为质子单粒子翻转极限截面，取值 2.04×10^{-10}，代入式（3）得 R_G=0.0001125 upsets/day。

3DFN128G08 由 8 片 MT29F16G08 叠装而成。图 2 中的宇航存储模块共使用 4 组，每组 9 个 3DFN128G08，其发生翻转的概率 R_{4T} 为：

$$R_{4T} = 4\times9\times8\times R_G \quad (4)$$

计算可得 R_{4T}=0.0324upsets/day，即 4Tb 宇航存储模块大约 31 天内发生一次翻转，发生翻转后其错误状态不能自动恢复，影响产品可靠性。因此，必须针对配置命令采取单粒子翻转加固技术。

采用改进后的配置刷新流程，单粒子翻转加固技术在页操作前增加配置刷新操作，保证配置翻转不会影响下次页操作，而只影响本级芯片（共 9 片 Flash 芯片）的单页数据。此时整机翻转概率$\overline{R}_{4T}$为：

$$\overline{R}_{4T} = 4\times9\times R_G \quad (5)$$

Flash 芯片读写是突发进行的，发生单粒子翻转时芯片有可能处于空闲状态，此时单粒子翻转不影响读写，因此整机的翻转概率考虑页操作占空比σ，计算单写、单读、同时读写状态下 4Tb 存储的翻转概率分别为：

$$\overline{R}_{4\mathrm{Tw}}=\overline{R}_{4\mathrm{T}}\times\sigma_{W} \tag{6}$$

$$\overline{R}_{4\mathrm{Tr}}=\overline{R}_{4\mathrm{T}}\times\sigma_{\mathrm{R}} \tag{7}$$

$$\overline{R}_{4\mathrm{Trw}}=\overline{R}_{4\mathrm{T}}\times\sigma_{\mathrm{RW}} \tag{8}$$

以输入 40Mbps、输出 225Mbps、内部读写速率为 640Mbps 的同时读写情况进行计算得到同时读写页操作占空比$\sigma_{\mathrm{RW}}=0.375$，单写页操作占空比$\sigma_{W}=0.0625$，单读页操作占空比$\sigma_{\mathrm{R}}=0.3125$。代入公式（5）、（6）、（7）、（8）得到的 4Tb 存储的翻转概率如表 5 所示：

表 5 翻转概率对比表

Tab.5 Flip probability comparison table

	未采用抗辐射加固	单写	采用抗辐射加固后单读	同时读写
翻转概率（upsets/day）	0.0324	0.00025312	0.00126562	0.00151875
翻转一次的天数（day）	31	3951	790	658

对比表 5 中数据可知，采用 NAND Flash 存储器抗辐射加固后，4Tb 宇航存储翻转概率由约 31 天／次改进为约 658 天／次，针对 Flash 配置命令区的频繁刷新的抗辐射加固技术使得单板 4Tb 存储器单粒子翻转的功能性失效概率降低了约 21.23 倍，提高了宇航存储器的可靠度。

本文通过对 NAND Flash 存储器的单粒子翻转效应进行深入研究和分析，提出一种针对 Flash 配置命令区的频繁刷新的抗辐射加固措施。通过抗辐射加固性能评估的结果表明，该措施使得单板 4Tb 存储器单粒子翻转的功能性失效概率降低了约 21.23 倍，提高了宇航存储产品可靠性。

参考文献（References）

[1] 王世元．NAND Flash 错误特性模型及应用研究 [D]. 哈尔滨工业大学，2016.
WANG S. Research on Error Characteristics Modeling of NAND Flash And Applications[D]. Harbin Insititute of Technology, 2016.

[2] RAJSUMAN R. Design and test of large embedded memories: an overview[J].IEEE Design and Test of Computers,2001,18(3):16-27.

[3] 施宇根，李少甫，齐艺轲．存储器抗辐射加固的矩阵纠错码研究 [J]. 中国空间科学技术，2019，39(1):67-72.
LI S ,QI Y.A study on matrix error correction code menory hardening[J].Chinese Space Science and Technology, 2019,39(1): 67-72.

[4] 赵元富，王亮，岳素格等．纳米级 CMOS 集成电路的单粒子加固技术及抗辐射加固设计平台 [J]. 第六届航天电子战略研究论坛论文集，2019，(65):1-8.
ZHAO Y, WANG L, YUE S, et al. Single Particle Reinforcement Technology and anti-radiation Reinforcement design platform for nanoscale COMS Integrated Circuits.[J]. Proceedings of the sixth Aerospace Electronics Strategic Research Forum, 2019,(65):1-8.

[5] CHEN D, WILCOX E, LADBURY R L, et al. Heavy Ion and Proton-Induced Single Event Upset Characteristics of a 3-D NAND Flash Memory[J]. IEEE Transactions on Nuclear Science, 2018, 65(1):19-26.

[6] 任学明，贺朝会．空间轨道单粒子翻转率预估的小样本法 [J]. 原子能科学技术，2009，(02):165-169.
REN X M, HE C H. Small Sample Method for Predicting Single Event Upset Rate in Space Orbits[J] Atomic Energy Science and Technology, 2009, (02):165-169.

作者简介：
朱新忠，男，宁夏中卫人，硕士研究生，研究员，主要研究方向为卫星综合电子。

《航天微电子》征文通知

《航天微电子》是由北京微电子技术研究所主办，由中国航天科技集团有限公司科技委微电子及元器件应用专业组作学术指导的一份专业性科技期刊。

本刊的宗旨是综合反映宇航和军用微系统、集成电路、半导体分立器件在材料与器件、设计与制造、测试与验证、质量与可靠性、集成与应用等方面进行前沿探索、理论研究、技术创新、工程实践的成果；为航天和军用微电子及元器件应用相关技术的学者、工程师、管理人员和学生提供一个交流的平台，进一步促进微电子技术与航天工程各专业技术领域的融合与创新。

本刊遵循“博采众长，百花齐放”的方针，以开放的态度广纳同业研究成果，恪守科学精神，弘扬学术民主，积极发挥好学术交流平台的作用，使之成为宣传和展示航天微电子技术和学术成果的一个窗口。

《航天微电子》长期面向广大从事微电子及其应用相关的科技工作者征文，欢迎积极踊跃投稿，一经录用稿酬从优。

征稿须知

征稿范围

《专家视角》栏目以特邀稿件形式报道专家的宏观视野、发展展望、回顾思考、理论见解、专业评论;

《战略前沿》栏目主要报道面向技术创新发展的战略性研究、前瞻性探索、趋势性分析、可行性研判、策略性建议;

《研究论坛》栏目主要报道具有原创性和一定学术价值的理论、技术创新研究成果;

《应用在线》栏目主要报道新技术和新产品在工程应用方面的方法创新、数据分析、经验总结;

《技术通讯》栏目主要报道具有重要意义的阶段性研发成果和技术动态。

来稿要求

稿件版面与格式要求遵照“航天微电子论文版面要求”的具体条件，信息完整，无泄密内容。同时需附作者单位保密部门出具的论文内容保密审查证明。投稿文章要求未曾在正式出版物上发表过、且不在其他刊物或会议的审稿过程之中。作者需保证投稿文章无抄袭和侵权等非法行为。

审　核

来稿将送相关领域专家审阅，作者在收到修改意见后，须在 1 周内修改完成并发往编辑部。是否刊登收录稿件最终由编委员会审定，在收到修回稿 2 周内由编辑部通知作者。不适合刊登之稿，会尽快通知作者（电子邮件形式）。若投稿后三个月内未收到编辑部任何通知，作者有权改投其他刊物。

录　用

经审核确定录用的论文，编辑部有权做必要的技术性和文字性修改。论文一经刊登，将酌致稿酬，并赠送当期。

版　权

经作者签字并在本刊发表之论文，表明作者已经认可其版权以下使用权（含数字版权）转让给本刊编辑部。本刊在与国内数字出版机构（文献数据库或检索系统）交流及合作时，不再征询作者意见。考虑到本刊目前为行业内交流资料，非正式出版物，本刊在论文刊登后，允许作者再次投稿其他刊物，并同意作者转让其版权。

投　稿

目前以电子投稿为主，不接收纸稿。文件以 doc 和 pdf 格式文件为宜。

投稿邮箱：内网 htwdz@mx.catec.casc　　外网 htwdz@mxtronics.com

联系人：范隆　　电话：68198371