异构多核人工智能 SoC 芯片的低功耗设计

颜 军 唐芳福 张志国 韩 俊 龚永红
珠海欧比特宇航科技股份有限公司,珠海 519080

摘 要 随着深空探测、载人航天、商业火箭和飞行器等各项航天任务的开展,各型号任务对硬件系统的智能化、可靠性、低功耗指标提出了更高的要求,作为系统"大脑"的 SoC 处理器亟需进行升级换代。本文综述了面向航天新任务应用的人工智能 SoC 芯片玉龙 810,介绍了新一代国产自主可控、高智能、高可靠、低功耗 SoC 芯片的功能特点、关键技术,重点描述了玉龙 810 芯片的低功耗设计方法和实现结果,通过优化技术玉龙 810 芯片动态峰值功耗达到了低于 5W 的指标。玉龙 810 芯片采用多核异构架构,主要由 4 个 SPARC V8 核、8 个 GPU 核和 8 个 NNA 核组成,片内通过 AMBA3.0 总线实现模块的互联互通,片上还集成 H.264/H.265, JPEC2000 等片上外设。

关键词 SPARC V8 处理器;GPU;NNA; AI 处理器单元;算力; 低功耗技术
中图分类号:TP332 **文献标识码**: A
文章编号:1006-3242(2020)02-0062-07

Low-Power Design for Heterogeneous Multi-Core AI SoC Chip

Yan Jun, Tang Fangfu, Zhang Zhiguo, Han Jun, Gong Yonghong
Zhuhai Orbita Aerospace Science & Technology Co., Ltd., Zhuhai 519080, China

Abstract *With the development of space missions such as deep space exploration, manned spaceflight, commercial rocket and aircraft, the requirements of intellectualization, integration, reliability and low power consumption for the control system on satellite, rocket and missile are increased.* SoC *processors need to be upgraded urgently as the brain of space control systems. The artificial intelligent* SoC *Yulong* 810 *is summarized in this paper, which is designed for new aerospace application. The heterogeneous multi-core design of a new generation of domestic autonomous controllable, high intelligence, high reliability and low power consumption* SoC *chip and the application scenario of the* SoC *chip is introduced in this paper, and the low power design method and implementation results of ultra-large scale Yulong* 810 *chip are also presented. The peak power consumption of chip design is less than* 5W. *The* 4 SPARC V8 *cores,* 8 GPU *cores and* NNA *cores is used to form a multi-core heterogeneous architecture in Yulong* 810 *chip. Modules are interconnected and interoperable on-chip through* AMBA3.0 *bus. On-chip peripheral devices such as* H.264/H.265 *and* JPEC2000 *are integrated.*

Key words SPARC *V8 Processor*; GPU; NNA; AI *processor unit*; *Computing power*; *Low-power*

收稿日期:2019-07-11

作者简介:**颜 军**(1962-),男,博士,高工,主要研究方向为智能控制、模糊控制、高可靠嵌入式控制器及 SoC 芯片的设计及产业化;**唐芳福**(1978-),男,工程师,主要研究方向为系统集成设计、高可靠 SoC 设计;**张志国**(1974-),男,工程师,主要从事 SoC 设计及仿真验证工作;**韩 俊**(1985-),男,中级工程师,主要从事 SoC 设计及仿真验证工作;**龚永红**(1977-),男,高级工程师,主要研究方向是高可靠 SoC 架构设计、计算机系统软件设计。

0　引言

人工智能是引领未来的战略性产业,是我国科技领域重要的发展战略[1],而人工智能(AI)芯片作为整个人工智能领域的关键技术环节,是我国人工智能产业的基础,是实现人工智能突破的重要关卡。

人工智能芯片主要解决的是对深度学习算法、卷积神经网络算法、自然语言处理(NLP)等的运算加速问题,需要具备足够大的算力。对于导引头、视频处理、遥感大数据等实时应用,通常会需要芯片提供 TOPS 级的运算能力,以智能导引头应用常见的 YOLO V3 算法为例,一个 416 × 416 图形输入理论上的计算量为 0.3TOPS 左右,导引头应用通常需要高于每秒 30 帧的处理速度,可知对处理器算力的要求为:不少于 9TOPS。

由于摩尔定律的限制,常规的提高主频的做法收效甚微。技术上,通常采用异构多核的架构来搭建 SoC 芯片,来提高整个 SoC 芯片的算力。

但算力的增加往往也意味着功耗的增加,而嵌入式多核处理器 SoC 芯片首先需要解决的就是在保证算力的情况下,功耗必须足够低。低功耗设计是一个系统工程,包含了电路级、结构级、算法级和操作系统级等多个方面的内容,需从多个方面进行综合性考虑[2]。而芯片的设计需要遵循平衡设计原则,需要在芯片的复杂度、内部结构、性能、功耗、扩展性等各个方面做一定的权衡,在设计过程中要坚持从整体结构的角度去权衡各个具体的结构问题。

1　玉龙 810 芯片设计指标、结构及应用

欧比特嵌入式人工智能处理器芯片玉龙 810,聚焦于前端图像处理和信号处理,具有对深度学习、神经网络算法的加速处理能力,算力要求达到 12TOPS、峰值功耗要求控制在 5W 之内。

芯片内部采用标准 AMBA3.0 总线协议的 AXI 总线,能够实现 SPARC V8 CPU 处理器[3]、GPU 及 NNA 等异构多核处理器的片内集成,如图 1 所示。芯片采用 FD - SOI 生产工艺,该工艺具有对单粒子锁定(SEL)天然免疫的特点;芯片外设接口丰富,具有 JPEG2000 编码器、CameraLink 数据接口、RapidIO 总线接口、1553B 总线接口[4]等航空航天专用处理单元和接口。

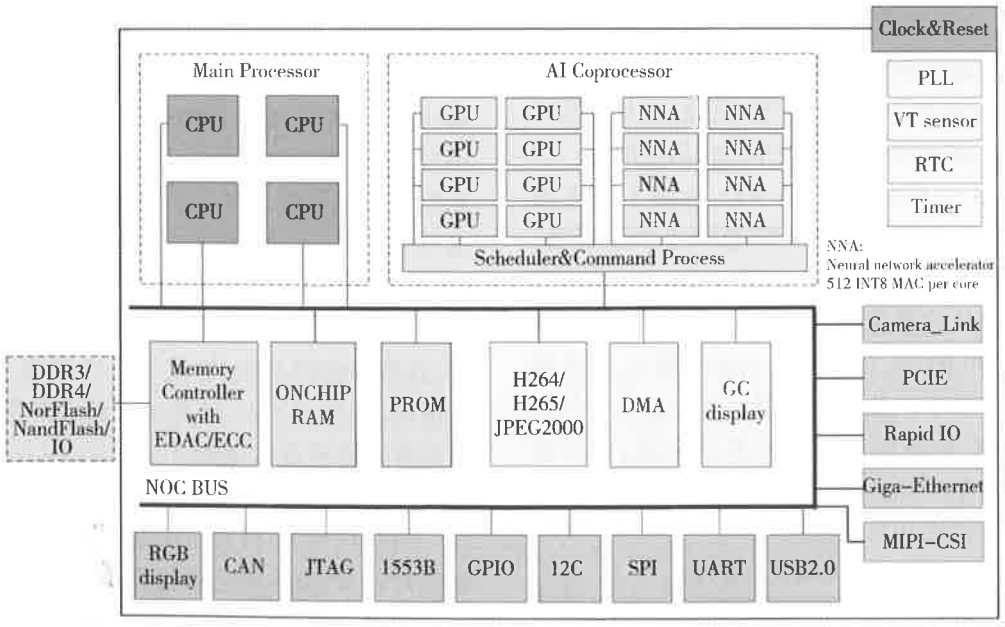

图 1　玉龙 810 SoC 芯片结构框图

芯片配套的软件开发框架中包含模型转换工具、软件开发环境等,能够实现与 TensorFlow,Caffe 等主流深度学习工具软件框架的无缝对接,支持绝大多数主流的深度学习网络模型,如 YOLO、SSD、RESNET、VGG、FastCNN 等,同时支持用户自定义网络模型。

玉龙 810 人工智能芯片的典型应用包括:

1)星上在轨情报提取:卫星上运行人工智能算

法，一边采集数据一边在轨实时提取情报信息，实时下传情报，大大提高情报获取的效率；

2）航天器在轨健康管理：航天器运行过程中，监测、采集运行数据，通过人工智能算法，自主完成故障分析、故障推理、故障处置以及故障预测，大大提高航天器的安全性；

3）飞行器智能制导：飞行器导引头使用人工智能芯片，运行人工智能算法，可以有效提高目标识别、目标跟踪的精度，并且通过人工智能技术有效排除诱饵干扰，提高飞行精度。

2 玉龙 810 芯片关键技术

2.1 适合超大数据吞吐量异构多核总线技术

玉龙 810 芯片内部功能模块通过片内 AXI3.0 总线互联。AXI3.0 具备高带宽、高传输速率性能[5]，其主要特点是：

1）单向通道体系结构。信息流只以单方向传输，简化时钟域间的桥接，减少门数量；当信号经过复杂的片上系统时，减少延时。

2）支持多项数据交换。通过并行执行猝发操作，提高数据吞吐能力，可在更短的时间内完成任务。

3）独立的地址和数据通道。地址和数据通道分开，能对每一个通道进行单独配置、优化，能根据需求控制时序通道，将时钟频率和效率进行最优配置。

芯片内部 SPARC CPU 是 AXI 总线上的主设备；AI 协处理单元作为也是 AXI 总线上的主设备，可以读写任何从设备的数据，但同时受 CPU 内核控制。总线上的从设备为：片内外设、片上存储器、片外存储器、片外 IO 等，这些从设备统一编址，被各处理器核心平等共享。总线控制器负责对总线访问进行仲裁和管理，仲裁管理逻辑和算法包括：固定优先、总线锁定、定时释放等。可通过对寄存器的设置选择仲裁管理逻辑和算法。

表 1 异构体系中模块分配列表

序列	模块	总线接口（主设备/从设备）
1	SPARC V8 内核	AXI（主）
2	AI 单元	AXI（主）/AHB（从）
3	H.264/H.265	AXI（从）
4	JPEG2000	AXI（从）
5	DDR4 控制器	AXI（从）
6	其他外设	APB（从）

2.2 AI 算法及对超大数据的运算支撑技术

芯片主要通过 GPU 核和 NNA 核来处理 AI 算法及超复杂数据运算，芯片内部配备了 8 个 GPU 核、8 个 NNA 加速器核。其中 GPU 核由标准 shader core 构成，可计算半精度、单精度、双精度浮点运算，也能处理定点运算；每个 NNA 单元由 768 个乘累加器（MAC）构成，可进行 8 位或 16 位定点运算，8 个 NNA 共同组成了 6144 个庞大的硬件计算阵列，在 1GHZ 主频的条件下可以提供 12TOPS 的定点运算算力。

表 2 AI 协处理器性能指标

	GPU	NN
Number of Cores	8	8
GFLOPS（32 - bit）@ 1GHz	64	NA
GFLOPS（16 - bit）@ 1GHz	256	NA
GFLOPS（64 - bit）@ 1GHz	8	NA
GOPS（16 - bit）@ 1GHz	256	3072
GOPS（8 - bit）@ 1GHz	512	12288

卷积神经网络（CNN）一般由卷积层、池化层、全连接层等组成，卷积层参数量小，计算量大，卷积运算在整个网路中的计算量占比一般超过 80%；NNA 核可以在 1 个或几个周期内完成大规模矩阵乘运算，从而实现对卷积层的加速。GPU 核使用浮点运算，可用于计算池化层、全连接层等，最大程度地保证系统精度。各层的分配由编译器事先指定，运行时 GPU 和 NNA 各自处理分配给自己的网络层，互不干扰。AI 协处理器除了 GPU 核和 NNA 核之外还包括：AXI 接口单元、内部 RAM、Cache 单元等，各部件协同工作，组成了一个完整、高效的处理子系统，也构成了对 AI 算法及超大数据提供高速算力的异构多核 SoC 架构。

2.3 低功耗优化技术

通过对各 IP 核的功耗参考数据的分析，可以得到芯片各 IP 核的理论功耗值，如表 3 所示。

如表 3 统计，如果功耗不加以控制，当主频在 1GHz 所有模块都通电运行的典型情况下，整个芯片的功耗将达到 8.83W，芯片功耗大，其弊端是：能源消耗大、芯片温度上升快、芯片寿命短。为了满足设计指标，整个芯片的功耗最好控制在 5W 以内。玉龙 810 芯片项目试图通过时钟门控、UPF 等技术来降低芯片整体功耗。

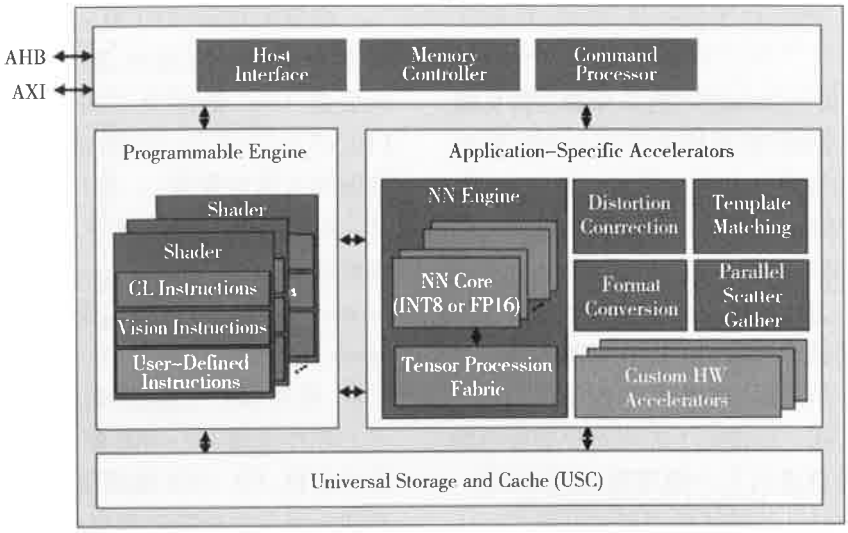

图 2　AI 协处理单元框架

表 3　各 IP 核理论功耗值

IP 核名称	关断电源下的功耗	运行状态下的功耗@1GHz
SPARC CPU	0.312μW	0.964W
GPU(8 核)	0.462μW	1.875W
NN(8 核)	0.375μW	1.951W
H.264	0.236μW	0.656W
H.265	0.312μW	0.737W
JPEG2000	0.294μW	0.368W
DDR4	不带控制功能	0.788W
其他模块	不带控制功能	1.491W
理论功耗		8.83W

3　低功耗设计及实现策略

CMOS 电路中的功耗由电路翻转时产生的动态功耗、P 管和 N 管同时导通时产生的短路功耗以及扩散区和衬底之间的反向偏置漏电路引起的静态功耗三部分组成[6]。

通常情况下静态功耗占总功耗的 1% 以下,系统非长时间处于休眠状态,则可以忽略不计。短路功耗在整个 CMOS 电路功耗中占比较小,与晶体管的转换速度有关,转换速度越快,其所占比例越小,短路功耗占总功耗的平均比例为 10% 左右。动态功耗占总功耗的比例约为 70% ~90%,而低功耗设计主要目的就是通过各种手段,实现降低动态功耗的数值[7]。

低功耗设计是一个系统的问题,需要在设计的各个层次上发展适当的技术,综合应用不同的设计策略,达到在降低功耗的同时维持系统性能的目的。研究证明在不同设计层次上的优化工作对功耗的改善程度不同,如表 4 所示,设计层次越高,改善效果越好[8]。

表 4　设计层次与改善程度关系表

设计层次	改善程度
行为级	50% ~90%
RTL 级	20% ~50%
门级	10% ~15%
晶体管级	5% ~10%
版图级	<5%

低功耗设计主要的策略有:

1)权衡面积和性能,使用并行、流水化和预计算等方法,用面积或时间换取低功耗;

2)关闭不用的逻辑和时钟;

3)使用专用电路代替可编程逻辑;

4)使用规则的算法和结构,以减少控制负荷;

5)采用新型的低功耗器件和工艺[9]。

3.1　预计算技术

预计算技术原理是:在第 t 个时钟周期内有选择性地预计算电路的输出逻辑值,然后在第($t+1$)个周期内或其后周期中,利用预计算的结果减少电路内部的跳变行为。预计算可分为单周期和多周期 2 种,综合多种情况的测试结果表明 2 种预计算技术均可降低功耗,部分情况下可降低 75%。预计算逻辑使得面积平均增加 3%,所引起的延迟增加通

常很小[10]。

3.2 时钟门控

时钟门控(Clock - Gating)一直以来都是降低微处理器功耗的重要手段,主要针对寄存器翻转带来的动态功耗[11]。如何更加有效地设计时钟门控,对于最大限度地降低功耗,同时保证处理器的性能至关重要。多核多线程微处理器中,多个功能部件可能不是同时工作的,对于无执行任务的功能部件就可以将其时钟关闭,减少其随时钟翻转进行多余的内部寄存器翻转,从而降低产生功耗的浪费和热量聚集。对于需要控制的寄存器,在一定情况下关闭寄存器的传输功能,阻止无用的数据进入下一级逻辑,避免引起一连串不必要的逻辑翻转,达到降低功耗的可能[12]。

芯片在设计之初,就配置了多组时钟域,每组时钟都能够单独通过独立寄存器进行 PLL 倍频、分频控制,同时在综合阶段,根据应用场景的不同,及各个模块布局布线不同,分别插入了一级时钟门控单元和二级模块时钟门控单元,实现了当某个模块或是模块端口信号进入静止空闲状态时,模块的时钟将自动被钳制住,从而达到降低模块内部动态功耗的目的,当然为了适应用户习惯,时钟的门控功能也可以通过软件设置为无效状态。时钟门控电路结构框图如图 3 所示:

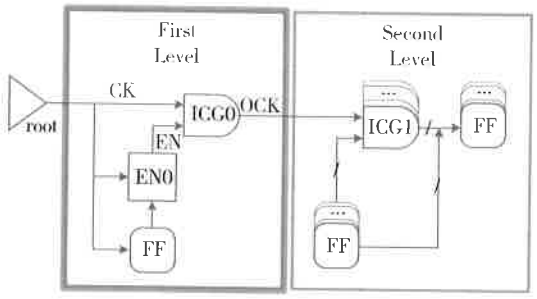

图 3 门控时钟树结构框图

3.3 多阈值单元库的应用

现在的工艺都会提供不同阈值 Vt 的单元库,同一工艺下不同阈值电压 Vtcell 特性不同,如表 5 所示[13]。

表 5 Vt cell 特性表

Vt 类型	速度	静态功耗	管脚兼容	优先级	场景表现
SVT	normal (1)	low (1)	yes	high	common
LVT	fast (1.5)	medium (1.5)	yes	medium	high frequency path
Ultra LVT	fastest (3)	high (4.5)	yes	low	very critical path

合理使用不同的 Vt cell 可以满足不同功耗性能需求,在使用过程中,应该优先使用 SVT 的 cell,而后是 LVT,最后万不得已的时候再使用 ULVT(ULVT 的 leakagecurrent 非常大,一般会达到 SVT 的四到五倍的量级)。设计工具支持 mix - Vt 的设计。在功耗优化的过程中,根据用户设定的 Vt 等价置换规则,在不影响 timing 的情况下,选择 leakagecurrent 小的 cell,这样在兼顾性能的时候可以满足 power 的需求。

3.4 采用 SEL 免疫的 FD - SOI 工艺

芯片采用 FD - SOI 制造工艺,与传统的块状硅技术相较,FD - SOI 能提供更好的晶体管静电特性,而埋入氧化层能降低源极(source)与汲极(drain)之间的寄生电容;此外该技术能有效限制源极与汲极之间的电子流动,大幅降低影响组件性能的泄漏电流,从而降低功耗。FD - SOI 22nm 工艺功耗比 28nmHKMG 降低了 70%,芯片面积比 28nm Bulk 缩小了 20%,光刻层比 FinFET 工艺减少约 50%,芯片成本比 16/14nm 低了 20%。除了低功耗与低成本,由于 FD - SOI 工艺的敏感体积更小,对闩锁效应(latch - up)免疫,具备更低的软错误率,以及更好的电磁兼容性,使其更适用于高可靠应用领域[14]。

3.5 UPF 技术

UPF 技术是由 Synopsys 公司提出,基于 IEEE1801 标准 Unified Power Format 的完整低功耗实现的设计流程标准[15]。

玉龙 810 芯片中 SPARC CPU、AI 协处理器、H.264/H.265、JPEG2000 以及外设的功耗较大,为了进一步降低功耗,对上述模块分别用独立电源域实现(switch - offdomain),以减小漏电,其余逻辑位于常开电源域(always domain)。采用成熟的 UPF 标准设计方法,如图 4 所示,采用不同电源给不同模块供电,插入电源开关控制,插入隔离器件,实现不同处理模块供电的单独控制方法。在某些功能不使用的时候,就把 switch - offdomain 关掉,这个时候,switch - offdomain 里的 power - gating cell 的输出会呈现出一个无限接近电源(header power - gating)或者地(footer power - gating)的状态,从而理论上确保了 switch - offdomain 的 leakagecurrent 为 0(由于 power gating cell 本身会有漏电的问题,所以 0 的漏电只是理论上的)[16]。

UPF 原理如图 4 所示。

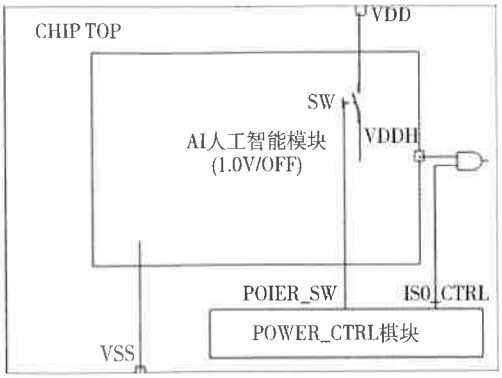

图 4　UPF 原理图

1)添加电源开关控制

```
create_power_switch PD_01_sw \
  -domain PD_01 \
  -output_supply_port{VDD_OUT VDD_01} \
  -input_supply_port {VDD_IN VDD} \
  -control_port {PSW_CTRL psw_en_01} \
  -on_state{PSW_ON VDD_IN {PSW_CTRL}}
  -off_state {PSW_OFF {! PSW_CTRL}}
```

2)插入隔离器件

```
set_isolation PD_07_ISO_IN -domain PD_07 -no_isolation -applies_to inputs

set_isolation PD_07_ISO_OUT_LOW -domain PD_07 -isolation_power_net VDD -isolation_ground_net VSS -clamp_value 0 -applies_to outputs

set_isolation_control PD_07_ISO_OUT_LOW -domain PD_07 -isolation_signal ios_en_07

-isolation_sense high -location parent
```

3)综合时导入 UPF 文件

```
Load_upf  top.upf
```

4　改进后的功耗结果

采取以上方法和策略后,采用 PTPX[16] 功耗分析工具,VCLP 低功耗检查工具[17],并利用激励文件 testbench 和仿真工具 VCS 产生 VCD 波形文件,然后使用 Power Compile[18] 工具将 VCD 文件转换成 SAIF 文件,并设置相关参数,产生功耗报告结果如下:

```
Net  Switching Power = 0.993W(20%)
Cell  Internal  Power = 3.283 W(66%)
Cell Leakage Power = 0.680W(13%)
Total Power = 4.96W(100.00%)
```

从功耗报告可以看出,芯片整体功耗降低到了约 4.96W,达到设计指标。同时通过仿真结果可以看到,芯片的处理能力没有降低,主频在 1GHz,浮点处理能力 64GFLOPS,定点处理能力 12TOPS,芯片最关键的能耗比指标为 2.4TOPS/W。

5　结论

功耗是 AI SoC 芯片的重要指标,功耗过高将极大地限制 AI SoC 芯片的应用。玉龙 810 人工智能芯片通过时钟门控、UPF 等技术成功降低了整体功耗,使芯片在具备高可靠、高性能指标的同时,达到了功耗小于 5W 的指标,远低于市场同类产品。在航空、航天领域核心元器件要求完全自主、可控的大背景下,玉龙 810 芯片的投产能够为型号项目的人工智能算法及超大数据高速处理及应用提供一个理想的 AI SoC 芯片平台。

参　考　文　献

[1] 新一代人工智能发展规划的通知[Z]. 国务院, 2017(7).

[2] 黄智伟. 低功耗系统设计[M]. 北京:电子工业出版社,2011,9.

[3] 颜军,龚永红,许怡冰,等. S698PM 宇航芯片的软件支持及信息处理性能测试[J]. 航天控制, 2019, 37(2):60-65. (Yan Jun, Gong Yonghong, Xu Yibing, et al. Software support and information processing performance test of S698PM aerospace chip[J]. Aerospace Control, 2019, 37(2):60-65.)

[4] 颜军,蒋晓华,唐芳福,等. 面向宇航应用的高性能多核处理器 S698PM 芯片的设计[J]. 航天控制, 2016, 34:89-94. (Yan Jun, Jiang Xiaohua, Tang Fangfu, et al. Design of high-performance multi-core S698PM for

space applications[J]. Aerospace Control, 2016, 34(4):89-94.)

[5] ARM Corporation. AMBA AXI and ACE protocol[Z]. Specification. 2017,11.

[6] Christian P. 低功耗处理器及片上系统设计[M]. 北京:科学出版社,2012,41.

[7] 梁宇,韩奇. 低功耗数字系统设计方法[J]. 东南大学学报,2000, 30(5):136-138. (Liang Yu, Han Qi. Low power design methodology[J]. Journal of Southeast University,2000,30(5):136-138.)

[8] 单长虹. 低功耗双边沿触发计数器的设计[J]. 计算机工程与应用,2004,13:126-127. (Shan Changhong. Design of low power double edge triggered counter[J]. Computer Engineering and Applicatons, 2004. 13:126-127.)

[9] 陈春章,王国雄. 数字集成电路物理设计[M]. 科学出版社,2008.

[10] Sung Mo Kang, Yusuf Leblebici. CMOS 数字集成电路分析与设计(第三版)[M]. 北京:电子工业出版社, 2004.

[11] 王延升,刘雷波. SoC 设计中的时钟低功耗技术[J]. 计算机工程,2009,35(24):257-261. (Wang Yansheng, Liu Leibo. Clock low power consumption technique in SoC design[J]. Computer Engineering, 2009,35(24): 257-261.)

[12] Michael K, David F, et al. Low power methodology manual for SOC design[M]. Springer, 2008,5.

[13] 浅论芯片低功耗的设计实现[EB/OL]. https://blog.csdn.net/i_chip_backend/article/details/90486964. 2019,05.

[14] 格罗方德专家深度揭秘 FD-SOI 工艺四大优势[EB/OL]. 电子创新网,2016.

[15] 孙铁群. 数字集成电路低功耗物理实现技术与 UPF[C]. Synopsys SNUG,2011:20-24.

[16] Synopsys Corporation. PrimeTime PX user guide[M]. Version L-2016.06, June 2016.

[17] Prance Zhang. Dynamic power optimization with IC compiler and prime time PX[C]. Synopsys SNUG, 2013, 13.

[18] Synopsys Corporation. PowerCompiler User Guide[M]. Version L-2016.06-SP2, September, 2016.

(上接第 55 页)

[6] 黄成,陈兴林,王岩,等. 基于气浮台的交会对接模拟及姿态跟踪控制[J]. 中国惯性技术学报,2016, 230(1): 172-188. (Huang Cheng, Chen Xinglin, Wang Yan, et al. Rendezvous and docking simulation and attitude tracking control based on air-bearing table[J]. Journal of Chinese Inertial Technology. 2016, 230(1): 172-188.)

[7] 许剑,任迪,杨庆俊,等. 五自由度气浮仿真试验台的动力学建模[J]. 宇航学报,2010,31(1):60-64. (Xu Jian, Ren Di, Yang Qingjun, et al. Dynamic modeling for the 5-DOF air-bearing spacecraft simulator[J]. Journal of Astronautics,2010,31(1):60-64.)

[8] 何兆伟,师鹏,葛冰,等. 航天器地面实验的相似性分析方法[J]. 北京航空航天大学学报,2012,38(4): 502-508. (He Zhaowei, Shi Peng, Ge Bing, et al. Similitude investigation for ground experiment of spacecraft[J]. Journal of Beijing University of Aeronautics and Astronautics, 2012,38(4):502-508.)