北京大学

软件与微电子学院
集成电路与智能系统系

《集成电路前沿技术导论》

课程报告

报告题目：处理器前沿技术研究报告

姓 名：胡成成

学 号：2101210578

日 期：2021 年 12 月 01 日

处理器前沿技术研究报告

摘要

随着人工智能等领域的迅速发展，高密集的计算需求越来越明显，对高性能低功耗的处理器要求越来越高。本文针对处理器前沿技术与设计展开调研研究，得出针对独特场景设计的处理器是一个热门的趋势的结论，分析了当下处理器在多核设计下通信设计，人为因素以及功耗散热的挑战，以及 AI 处理器面临的大容量存储与高密集计算，面向特定领域的架构设计，终端和“云”的需求差异，片设计要求高、周期长、成本昂贵以及架构及工艺面临的挑战。对未来处理器的体系架构的设计与轻量级的硬件开发进行展望。

引言

随着物联网、云计算、大数据、人工智能的快速发展，在高性能计算和大数据应用的背景下，对性能的追求需要模块化、可扩展、节能、低成本的多核系统。其中，处理器技术直接推动着嵌入式计算机的发展。通用处理器是信息产业的基础部件，是电子设备的核心部件，也是计算机的运算与控制核心。嵌入式计算机设计很大程度上由处理器的发展推动的。另一方面，随着登纳德缩放定律和摩尔定律几近终结, 通过领域特定体系结构提升微处理器性能变得越来越重要, 迫切需要提升微处理器设计生产率来应对网络、智能、安全等领域特定需求。因此，处理器的研究也步入了嵌入式研究的热门方向，各大厂商与公司也越来越重视处理器的改进与优化，高性能、低功耗的处理器成为当今用户的首选标准。

研究背景

随着高性能计算的普及，处理器需要继续提供更高的性能，使用最新的工艺技术集成更多的内核，以及先进的芯片堆叠技术，以经济的解决方案实现广泛的性能目标。除了通用处理器外，图形处理器、AI 处理器以及张量处理器等也在不断发展，以适应日益增加的算力需求。如今，针对特定应用场景设计的处理器越来愈多，同时，用户不仅追求更高的性能，低功耗也是现在的追求的一大趋势。

除了处理器硬件上的设计之外，处理器的架构设计上也受到各大厂商的关注。如今，基于 ARM 架构的嵌入式芯片被广泛使用并大量生产，开源指令集 RISC-V 在国内的影响力也越来越大。其中，ARM 架构的推广得益于它的设计，使得处理器更加的节能、符合低功耗的设计需求。其次，处理器制作的工艺也在追求更加精细，从处理器工艺制作的历史来看，从 1976 年的 6 微米工艺到如今的 7 纳米工艺，可见处理器制作的工艺水平稳步提升，这也是摩尔定理的阐述。但如今，我们已经步入后摩尔时代，摩尔定理已经近乎终结，新工艺的发展进入瓶颈，这一点有待突破。

研究意义

处理器是计算机的核心，是算力和生产力的核心。处理器的类型如今多元化，对处理器的研究，包括性能，功耗，成本，兼容性等等，这些都是我们必须一次解决的问题，各大厂商不断追求高性能的芯片，更久的续航能力，能够被大众所接受，这便是研究处理器的重要意义。例如，微处理、微控制器与其他外设及接口一起构成了微控制器,以实现对各种智能电子产品的不同功能控制,从而确保电子产品在工业、国防、通信以及交通等领域为人们服务。处理器的性能优化,可以提高整个处理流程的速度,所以深入研究控制器并进行优化设计具有一定的理论与实际意义。除此之外，图形处理器(GPU)的应用已经从桌面计算系统、手持和便携电子设备、游戏机等领域扩展到高性能计算和人工智能等领域。GPU 架构从原来的图形专用加速器发展到现在的单指令流多数据流或单指令流多线程处理器。现代的 GPU 不仅仅是一个特殊用途的加速器或几何引擎,它也可以作为一个通用计算芯片。在集成电路工艺进步和应用发展的驱动下,当前的图形处理系统芯片正在发生变革。因此,针对新型图形算法和通用计算的需求,面向未来图形处理器芯片高级应用的不断发展,在体系结构方面应对长线、功耗和工艺缺陷等问题,研究新一代图形系统芯片的体系结构和关键技术,具有重要的科学意义,研究和设计 GPU 芯片,打破国外的垄断是我国社会和经济发展的迫切需要。

处理器前沿技术研究现状

处理器设计的研究主要针对处理器架构、处理器的制作工艺、处理器的晶体管数目与相应所占的面积、处理器相关的缓存机制与技术、电源管理相关技术、

通信的 IO 设计与性能以及功耗、处理器正常工作与睡眠状态下的平均功率等等。这些都是评判一款处理器关键要点。各大厂商的研究人员针对不同场景设计不同的处理器，他们在不同的方面展现出自己的优势和一些先进的技术。

文献[5]设计了一款采用 x86-64 架构、采用节能型 TSMC 7nm FinFET 工艺制造的一款高效的 AMD 芯片，代号“Zen 2”。“Zen 2”设计比“Zen”有许多设计改进，包括平均单线程应用程序的每周期指令数（IPC）提高 15%。由于采用 7nm 的制作工艺，其在功率、性能和面积扩展上有显著的优势。但是，为了降低电流，团队将位线预充电从 VDDM 移动到 VDD。这一变化带来了一些新的危害，包括高 VDD/低 VDDM 下的位单元稳定性和低 VDD/高 VDDM 下的可写性。

文献[6]设计了一款基于 AMD Infinity 结构的第二代 SOC 芯片，使用 3 种独特的混合工艺技术芯片，分别针对服务器和客户端市场实现领先的性能、性能/美元和性能/瓦特。其设计具有很好的向后兼容性，降低了成本，功耗显著减小，提供了非常经济高效的性能。同时，芯片设计采用异构芯片体系结构，其在封装、互连、测试和电源管理基础设施的工程设计方面取得重大进展，但允许跨多个市场进行产品配置，这些配置的性能和成本效益远远高于其他方式。

文献[7]一种有源插入集成电路，1）用于片上电源管理的开关电容电压调节器（SCVR）；2）所有芯片之间灵活的系统互连拓扑，以支持可扩展的缓存一致性；3）用于密集层间通信的节能 3D 插头；4）用于套接字通信的内存 IO 控制器和 PHY。该电路是第一个基于芯片的多核架构，用于使用有源插入器进行高性能计算与集成。

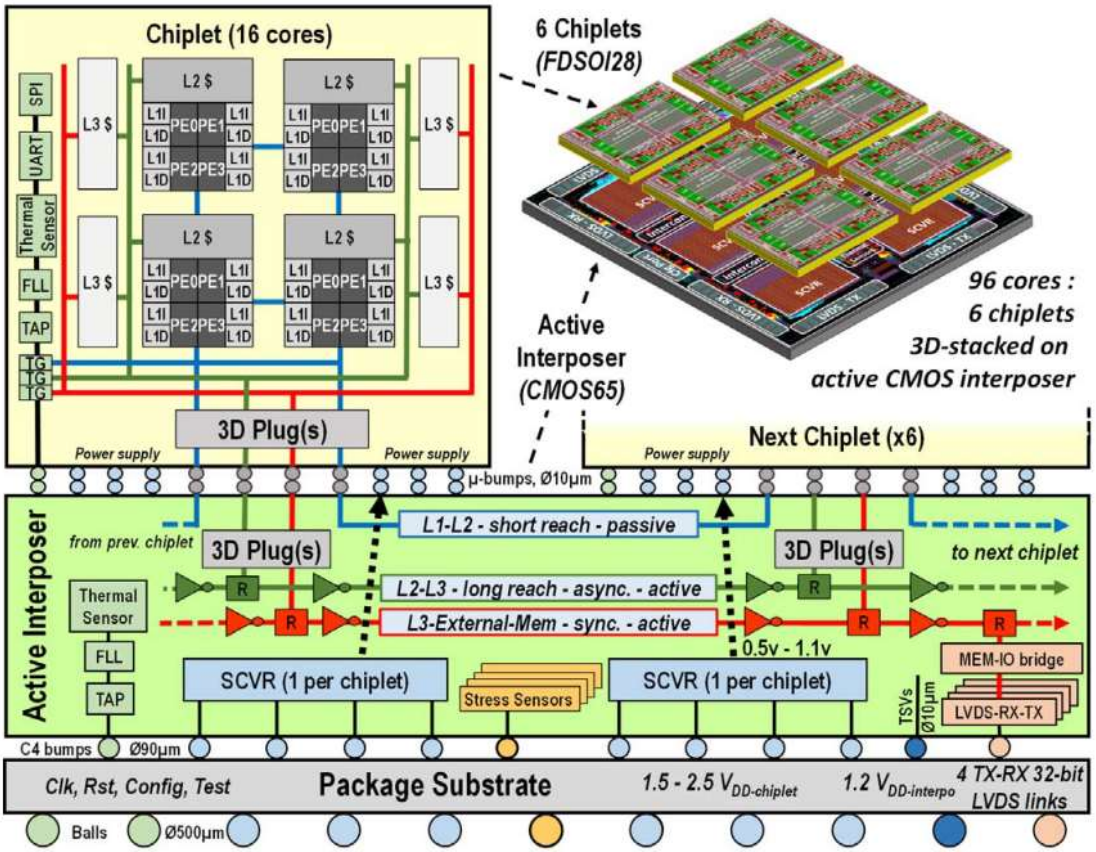

图 1 96-core architecture composed of 6 chiplets 3D-stacked onto an active CMOS interposer[7]

文献[8]设计了一种具有三集群 CPU、稀疏感知 NPU 和硬件自动时钟门控（HWCG）功能的低功耗高性能 7nm Exynos AP 处理器。用于移动应用，提供高性能以改善用户体验、更高的图形渲染性能、奇特的相机操作、更快的数据通信以及更长的电池寿命。

文献[9]介绍了一个完全集成的 5G 移动智能手机 SoC 的异构 CPU 复合体，该 SoC 采用 7nm FinFET 技术实现。该 SoC 集成了支持 4.7Gb/s 的下行速率和 2.5Gb/s 的上行速率的 5G 调制解调器，采用异构 CPU 复合体包含 8 个单线程内核，是一个完全集成的 5G 智能手机 SoC 的 CPU 复合体。

文献[10]设计了一个 SoC 平台，该平台集成了简化电源管理的技术，并通过芯片微体系结构解决方案支持成本敏感系统。

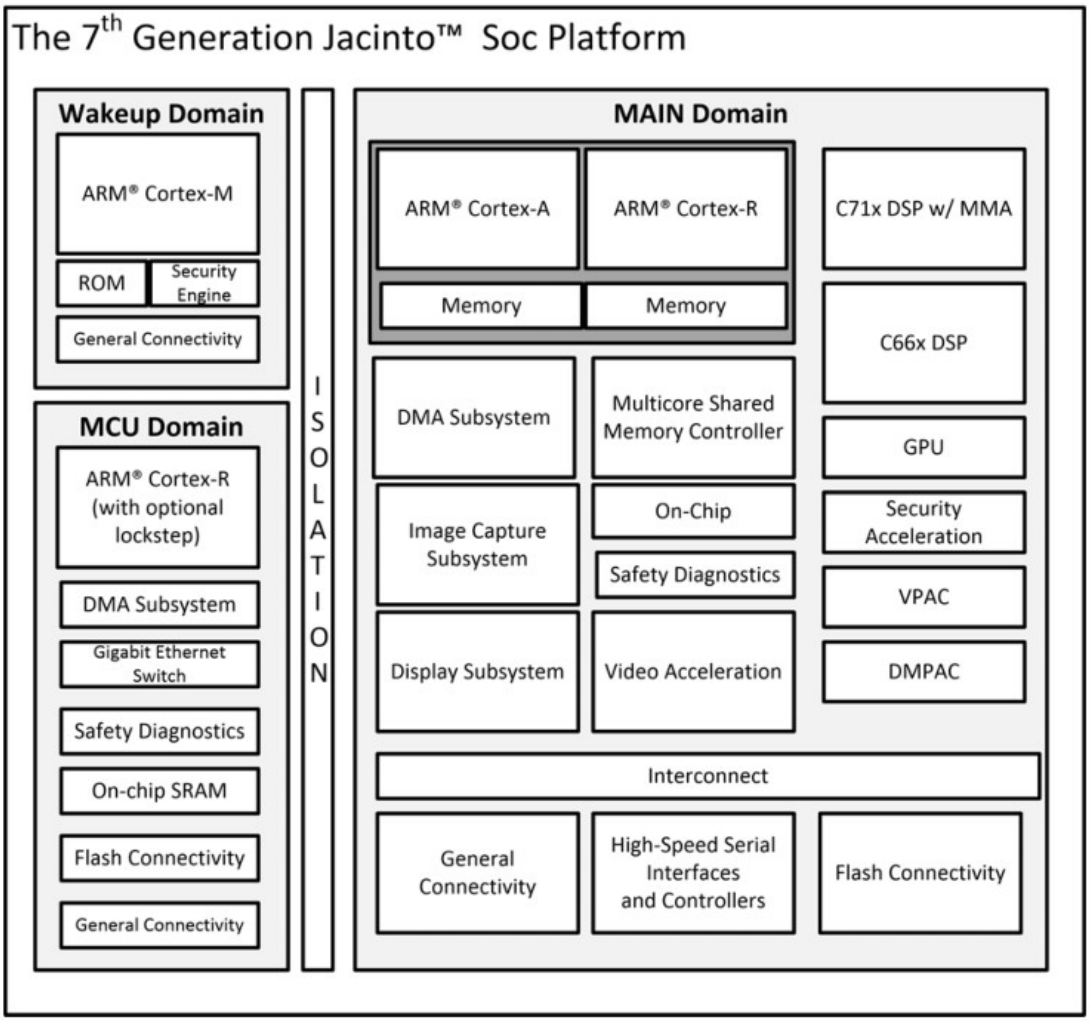

图 2 The 7th generation Jacinto™SoC platform block diagram.[10]

文献[11]重新设计了 z15 系统中最新的 IBM Z 微处理器，相比 z14 系统相比，性能、系统容量和安全性都有所提高，主要得益于 FinFET 技术和 17 层铜互连技术。z15 中的系统拓扑发生了变化，其最大配置为 5 个抽屉，每个抽屉包含一个 SC 芯片和四个 CP 芯片，总共 240 个物理核心。

文献[12]针对自动驾驶系统的应用场景，实现了一种高效的针对该场景下多的处理器，为产品落地提供了一种实用的解决方案。针对计算机视觉的六个核心问题出发，大大提升了 CNN 计算效率，为自动驾驶的图像处理落地打下基础。这是一款针对 AI 特定场景设计的 SOC。

文献[13]针对深度神经网络越来越依赖密集的算术计算模式，这不适合于通用处理器的问题，设计了一款采用 TSMC 16nm FinFET 制造的 SOC 芯片，该系统包含八个应用核心、一个系统管理核心、八个串行链路和三级缓存，应用核心包括一个标量 RISC-V 处理器和一个解耦矢量加速器。

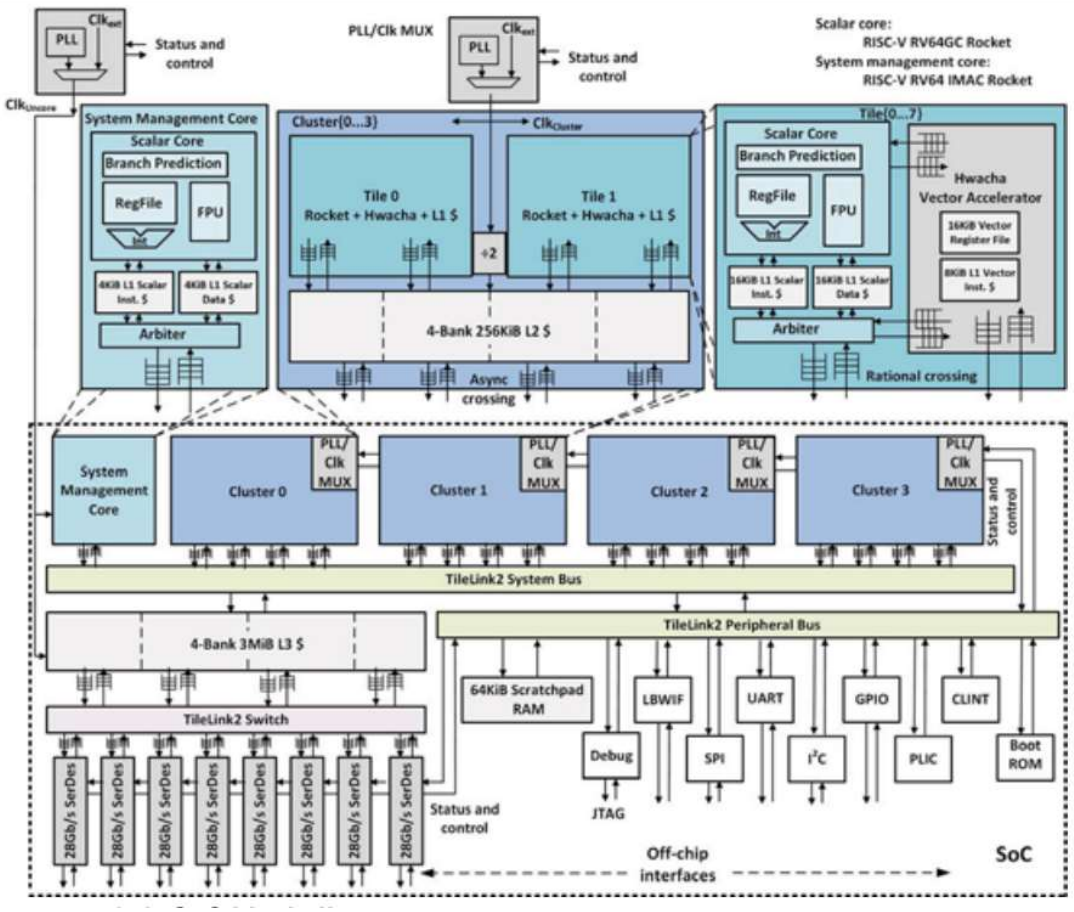

图 3 SoC block diagram[13]

文献[14]针对物联网要求终端节点具有超低功耗、长电池寿命、高性能、能源效率和极端灵活性的特点，设计了一款叫 Vega 的 SOC，满足 NSAAs 算法的性能和灵活性要求，SoC 具有 10 个 RISC-V 核：一个用于 SoC 和 IO 管理的核，以及一个支持多精度 SIMD 整数和浮点计算的 9 核集群；两个可编程机器学习（ML）加速器分别在睡眠和活动状态下提高能效。

文献[15]针对生物医学 AI 处理的场景存在的一些冗余设计造成的功耗损失的问题，设计了一种具有自适应学习的可重构生物医学 AI 处理器。它具有以下关键特征：1）具有可重构神经网络和生物医学处理引擎的可重构生物医学 AI 处理架构，以支持多功能生物医学 AI 处理。2） 事件驱动的生物医学 AI 处理架构和近似数据压缩技术，以降低功耗。3） 一种基于人工智能的自适应学习架构，用于解决患者与患者之间的差异。4） 一种可重构 FIR 引擎，可重用神经网络引擎以减少硬件开销。

文献[16]针对一种新的计算机体系结构，即退火处理器（AP）的高速发展，设计了一种具有两项关键技术的可扩展 CMOS-AP：1）基于触发器（FF）的自旋

电路，通过复制 Metropolis 算法（固定温度的 SA）允许扩展比特宽度，以及 2）芯片间接口（i/F）采用一种利用退火特性的数据压缩方法，在不降低退火速度和精度的情况下实现多芯片操作。CMOS-AP 演示了 9×16k 自旋系统的多芯片操作，退火速度快 233 倍，计算能量比在 CPU 上运行 SG3 低 972 倍。

文献[17]针对超分辨率技术支持高分辨率视频流、图像放大和远目标识别的技术问题，提出了基于双三次插值，采用预学习滤波器提高图像质量的新的处理器算法。

文献[18]针对介绍了一种通过多格式视频解码器（MFD）支持 AV1 标准的视频处理器，其采用最先进的 5nm CMOS 工艺制造，具有 5835K NAND2 等效门计数和 518kB SRAM。对于 AV1，解码 AV1 7680×4320p 视频时，MFD 消耗 116mW，解码速度为 30fps（0.12nJ/像素），核心工作电压为 0.7V。

分析近两年 ISSCC 会议上关于处理器的研究成果，主要包括：一些大厂商在自己之前研究的基础上展开新的研究并取得一定性能的提升与优化；一些研究者针对特定的人工智能场景设计 SOC，例如自动驾驶，医疗等方面；还有一些研究者主要针对一些独特的体系结构或者特定的使用工具展开设计。

处理器前沿技术挑战与解决方案

多核设计

如今，处理器的设计趋向于更高性能与低功耗。单个处理器的能力在固定尺寸下的性能有限，我们知道，如果设计一款两倍于原来大小的处理器，性能并会不增加到原来的两倍，性能大概只会作平方根的增长，所以尺寸加倍并不会带来同等数量级的性能增长，只会带来大概 1.4 倍于原来的性能，即增长 40%。这个被称为 Pollack 定律。因此，多核处理器的设计成为当下的主流设计，也是面临一大挑战。主要在以下三个方面：

1）多核设计中核心之间的通信也许是最关键的问题，因为在一些系统中会起到至关重要的作用。一般通信量的增加与核心的数量 n 是一个二次方程 (n+1)*n/2 的关系，当下可能的解决办法是采用中心存储器来缓冲，不过当所有的处理器核都与存储器进行内部通信时就会出现问题。一般会采用信息包开关网络或其他架构。

2）人为设计因素同样会影响多核系统的架构。将设计分解为一个一个的模块和模块之间的互联。同样，连接的数目也同模块数是一个二次方程的关系。解决复杂问题的方法是应用和归纳概念的能力。

3）降低多核系统的功耗和散热处理是工程师面临的另一个主要的问题。降低功耗可以通过限制每个核心的运算来实现，称为"voltage scaling" ——每个核心的频率和供电电压可以根据每个处理核的任务运算两来进行功耗优化。

AI 处理器设计

在人工智能以及针对各种截然不同的终端市场和系统而设计的机器学习芯片快速发展的推动下，人们可选择的存储器/体系架构数量呈现爆炸式增长。AI 芯片的设计面临的问题主要在以下几个方面：

1）大容量存储和高密度计算。当神经深度学习网络的复杂度越来越高，参数也会越来越多。

2）面向特定领域的架构设计。AI 应用场景越来越丰富，这些场景的计算需求是完全不一样的。比如当我们面临语音场景的时候，图像处理的知识能够重复使用的部分是非常有限的。如何通过对于不同的场景的理解，设置不同的硬件架构变得非常重要。

3）终端和“云”的需求差异。云端和终端的设计完全不同，云端需要对海量数据进行处理，要进行存储、训练，要高并行、高带宽；终端上首先要采集，然后做终端的推理，还要做一些训练，但是更关心安全性、低能耗、低延时等等的处理。

4）片设计要求高、周期长、成本昂贵。从芯片规格设计芯片结构设计、RTL 设计、物理版图设计、晶圆制造、晶圆测试封装，需要两到三年时间，正常的时间里软件会有一个非常快速的发展。

5）架构及工艺面临的挑战。 随着我们的工艺不断的提升，从 90 纳米到 10 纳米，逻辑门生产的成本到最后变得饱和。我们也许在速度上、功耗上会有提升，但单个逻辑生产的成本不会再有新的下降。这种情况下如果仍然用几千甚至上万个晶体管去做一个比较简单的深度学习的逻辑，你会发现到最后在成本上是得不偿失的。

针对处理器设计的思考与见解

处理器的设计在当下人工智能热潮下越来越多元化，通过本次调研学习，了解了处理器研究领域大方向下的一些细节，当下处理器 SOC 设计是一个比较热门的方向，针对不同场景设计的需求日渐明显，这对当下的工艺制造提出挑战。这让我对“Processor”一词有了新的理解，专用处理器的也许是未来的趋势，除了如今的 PC，手机等移动端采用的是通用处理器，一些物联网设备的普及，将要求专用处理器的大量普及，从近两年的顶会论文也可以看出，专用处理器的设计越来愈多，尤其是 AI 领域的设计，针对不同的场景进行独特设计。我想，这对专有处理器的工艺提出巨大挑战，设计成本大大增加，寻求一个模块化可复用的单元设计方案也许是未来的专有处理器设计的趋势。同时，我们知道，处理器的流水线设计，并行处理以及多核设计的普及，大大提升了计算性能，基本能满足当下 AI 领域的算力需求，但是面临的功耗问题很大，现如今在算法层面上有 SNN 的设计来降低神经网络训练的功耗，但是在密集的计算下硬件的设计我想更是关键，可持续发展是我们面临的最大问题，低功耗的需求将被我们越来越重视，相信在不久的未来，低功耗处理器将普及，同时性能可以被大众所接受。

总结与展望

本次调研从处理器的前沿研究出发，针对近两年的芯片顶会的论文出发，分析了 ISSCC 会议处理器专栏下的所有论文，主要包括：一些大厂商在自己之前研究的基础上展开新的研究并取得一定性能的提升与优化；一些研究者针对特定的人工智能场景设计 SOC，例如自动驾驶，医疗等方面；还有一些研究者主要针对一些独特的体系结构或者特定的使用工具展开设计。并且针对顶会论文的设计方法以及他们的研究对当下的处理器面临的技术挑战进行分析，调研其可能的解决方法，针对多核设计和 AI 处理器设计的几点挑战进行分析，最后针对这些问题与挑战，提出鄙人的一些简介与看法。本次调研学习，看到了处理器研究领域的前沿技术及其进展，主要是从处理器本身的设计与工艺出发，也许，在未来对处理器体系架构的独特设计是新的机遇，同时开放性的计算机体系架构，面向处理器的操作系统的实现，轻量级的硬件开发等等，这些都是未来的一些方向和机遇，值得我们去关注探索。相信，不久的未来，处理器的高性能低功耗的设计将

普及到大众的日常生活中，国产的芯片设计也能迅猛发展，步入国际前沿水平，打破国外的技术封锁！

参考文献

[1]. Singh T, Naffziger S, Vivet P. Session 2 Overview: Processors[J], 2020.

[2]. Goswami K, Mondal H, Sen M. A review on all-optical logic adder: Heading towards next-generation processor[J]. Optics Communications, 2021, 483: 126668.

[3]. Chang L, Li C, Zhang Z, et al. Energy-efficient computing-in-memory architecture for AI processor: device, circuit, architecture perspective[J]. Science China Information Sciences, 2021, 64(6): 1-15.

[4]. Asghar M N. A Review of ARM Processor Architecture History, Progress and Applications[J]. Journal of Applied and Emerging Sciences, 2020, 10(2): pp 171-179.

[5]. Singh T, Rangarajan S, John D, et al. 2.1 Zen 2: The AMD 7nm Energy-Efficient High-Performance x86-64 Microprocessor Core[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 42-44.

[6]. Naffziger S, Lepak K, Paraschou M, et al. 2.2 AMD chiplet architecture for high-performance server and desktop products[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 44-45.

[7]. Vivet P, Guthmuller E, Thonnart Y, et al. 2.3 A 220GOPS 96-Core Processor with 6 Chiplets 3D-Stacked on an Active Interposer Offering 0.6 ns/mm Latency, 3Tb/s/mm 2 Inter-Chiplet Interconnects and 156mW/mm 2@ 82%-Peak-Efficiency DC-DC Converters[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 46-48.

[8]. Kim Y D, Jeong W, Jung L, et al. 2.4 A 7nm High-Performance and Energy-Efficient Mobile Application Processor with Tri-Cluster CPUs and a Sparsity-Aware NPU[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 48-50.

[9]. Mair H, Wang E, Nayak A, et al. 2.5 A 7nm FinFET 2.5 GHz/2.0 GHz Dual-Gear Octa-Core CPU Subsystem with Power/Performance Enhancements for a Fully

Integrated 5G Smartphone SoC[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 50-52.

[10]. Venkatasubramanian R, Steiss D, Shurtz G, et al. 2.6 A 16nm 3.5 B+ Transistor> 14TOPS 2-to-10W Multicore SoC Platform for Automotive and Embedded Applications with Integrated Safety MCU, 512b Vector VLIW DSP, Embedded Vision and Imaging Acceleration[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 52-54.

[11]. Berry C, Bell B, Jatkowski A, et al. 2.7 IBM z15: A 12-Core 5.2 GHz Microprocessor[C]//2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020: 54-56.

[12]. Matsubara K, Hanno L, Kimura M, et al. 4.2 A 12nm Autonomous-Driving Processor with 60.4 TOPS, 13.8 TOPS/W CNN Executed by Task-Separated ASIL D Control[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 56-58.

[13]. Schmidt C, Wright J, Wang Z, et al. 4.3 An Eight-Core 1.44 GHz RISC-V Vector Machine in 16nm FinFET[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 58-60.

[14]. Rossi D, Conti F, Eggiman M, et al. 4.4 A 1.3 TOPS/W@ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7 μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 60-62.

[15]. Liu J, Zhu Z, Zhou Y, et al. 4.5 BioAIP: A Reconfigurable Biomedical AI Processor with Adaptive Learning for Versatile Intelligent Health Monitoring[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 62-64.

[16]. Takemoto T, Yamamoto K, Yoshimura C, et al. 4.6 A 144Kb Annealing System Composed of 9× 16Kb Annealing Processor Chips with Scalable Chip-to-Chip Connections for Large-Scale Combinatorial Optimization Problems[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 64-66.

[17]. Shen H Y, Lee Y C, Tong T W, et al. 4.7 A 91mW 90fps Super-Resolution Processor for Full HD Images[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 66-68.

[18]. Kim T S, Lee S, Lee K, et al. 4.8 An Area and Energy Efficient 0.12 nJ/Pixel 8K 30fps AV1 Video Decoder in 5nm CMOS Process[C]//2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021, 64: 68-70.