超大规模集成电路基础文献阅读报告

郝淼 202328013229045

1 文献翻译

选择文章：Heterogeneous Multi-core Architectures

作者：Tulika Mitra

年份：2015

摘要：硅器件的晶体管数量遵循摩尔定律不断增加。但是登纳德缩放定律的失效将计算社区带到了一个十字路口，在这里，功率已经成为主要的限制因素。因此，未来的芯片可以有多个核心;但只有一小部分能在任何时间点被打开。在这个黑暗的硅时代，芯片领域的很大一部分仍然是黑暗的，这需要对架构设计进行根本性的反思。在这种情况下，异构多核架构结合了功能和性能方面的处理核心（CPU、GPU、专用加速器和可重构计算）的不同组合，提供了一个很有前途的选择。异构多核可以提供潜在的节能计算，因为只需要打开最适合当前计算的核心。本文概述了异构多核领域的最新技术。

关键字：异构多核，CPU, GPU，专用加速器，可重构计算

# 1 简介

多核和最终多核（单个芯片上的多个处理核心/CPU）的出现将计算机社区带到了十字路口。由于热/功率限制、可靠性问题、设计复杂性等原因，台式机、笔记本电脑和智能手机都不可逆转地向多核过渡。根据摩尔定律，随着晶体管密度继续呈指数级增长，芯片上核心的数量（而不是性能）预计每两年翻一番。然而，在未来，我们无法驾驭片上核心的增长曲线。主要的挑战是芯片上不断增加的功率密度，这阻止了所有核心同时打开。这种现象被称为暗硅，它正在推动异构多核架构的发展。

传统上，多核只是被设计为相同（但可能简单的）核的集合。这些同构多核设计简单，提供简单的芯片实现和常规的软件环境。不幸的是，来自不同应用程序领域的通用新兴工作负载具有非常不同的资源需求，很难用一组相同的核心来满足。相比之下，有许多证据表明，由不同核类型组成的异构多核解决方案在性能、功耗、面积和延迟方面具有显著优势。同时，异构多核非常适合暗硅系统，因为只需要打开适合应用程序的核。

本文将深入探讨异构多核体系结构及其相关的挑战和机遇。我们首先介绍技术、架构和应用趋势，这些因素导致了异构的出现。我们大致上将异构多核分为性能异构和功能异构两类，性能异构是指具有相同功能但不同功率性能特征的核心集成在一起；功能异构是指功能完全不同的核心集成在同一个芯片上。我们展示了性能异构多核架构和软件支持，以实现这种架构的全部潜力。然后，我们继续概述当前功能异构多核中存在的不同类型的处理元素，例如图形处理核心、专用加速器和可重构计算。

异构多核是一种不断发展的技术，目前面临着无数的挑战。这为计算栈的各个方面带来了有趣的研究问题，从设备、电路、体系结构一直到软件层，包括操作系统、编译器和编程范式。我们诚邀读者一同探索由异构多核提供的这些令人兴奋的机会。

# 2 背景

第一个微处理器于1971年以英特尔4004的形式推出，这是一个单芯片上的4位通用可编程CPU。从那时起，微处理器行业在多种因素的推动下实现了前所未有的性能增长：摩尔定律、登纳德缩放定律和微架构创新。**摩尔定律**[50]是戈登·摩尔在1965年基于观察提出的，即芯片中集成的晶体管数量大约每18到24个月翻一番，导致晶体管密度呈指数增长。由于处理速度与集成电路上晶体管之间的距离成反比，摩尔定律意味着微处理器的速度（时钟频率）也将每24个月翻一番。处理器速度的指数级增长直到2005年才有所减缓。例如，2005年推出的64位英特尔至强处理器运行频率为3.8 GHz，拥有1.69亿个晶体管。这与时钟频率为740千赫，只有2300个晶体管的英特尔4004形成鲜明对比。

虽然摩尔定律是时钟频率持续提高的原因，但由于一些微架构创新，包括处理器流水线、乱序执行、推测执行、存储层次结构，处理器性能进一步提高。这些进步使处理器能够通过利用指令级并行性（ILP）每周期执行多条指令，从而提高关键的每周期指令数（IPC）指标[29]。更重要的是，由于ILP是由底层架构从单线程程序中透明地提取出来的，因此软件开发人员无需任何额外的努力就能享受到性能上的好处。30年间，处理器性能伴随着时钟频率和IPC的增长而持续提升。

**同构多核的兴起**

由于各种原因，单处理器的性能增长放缓并走向终结，分别为：**功耗墙**、**ILP墙**和**存储墙**[57]。主要的限制因素是所谓的功耗墙，由热设计功率（TDP）定义——微处理器芯片可以合理耗散的最大功率。在近30年的时间里，摩尔定律得到了**登纳德缩放定律**[19]的帮助，使得处理器的功率保持在限制范围内。根据登纳德的理论，当线性特征尺寸缩放比为（一种工艺技术到另一种）时，晶体管数量增加一倍（摩尔定律），频率增加40%，但每个晶体管的功率可以减少，同时保持芯片总功率不变[21]。每个晶体管功率的降低是通过按比例降低工作电压和电流来实现的，这些电压和电流与特征尺寸的缩放成正比。降低工作电压需要降低阈值电压（），这在2005年漏电功率最小的情况下是可行的。然而，在65nm及以下工艺技术下，漏电功率对芯片总功率的贡献显著，降低会导致漏电功率呈指数增长。因此，，或者说工作电压不能再缩放（因为当工作电压接近时，晶体管延迟迅速增加，导致时钟频率下降），导致登纳德缩放定律失效。在每一代中保持功率不变是不可能的。相反，后登纳德缩放定律导致对于相同的模具面积，功率每代增加1倍。保持功率在TDP预期内要求芯片以低于固有频率的频率工作。由于动态功率大致与频率的三次方成正比，保持较低的频率（而不是每代增加40%）可以补偿由于晶体管数量翻倍而增加的漏电功率。事实上，在过去的十年里，时钟速度一直停滞在4 GHz左右，尽管晶体管数量在摩尔定律之后出现了爆炸式增长（例如，英特尔2014年推出的Ivy Bridge-EX处理器中有43.1亿个晶体管）。

在后登纳德缩放定律时代，大量的晶体管可以用来构建更复杂的单处理器微架构，以进一步提高IPC，从而弥补时钟频率增加放缓的不足。但是处理器已经遇到了ILP墙，越来越难在单线程程序中找到更多的并行性[76]。进一步尝试挖掘ILP将导致处理器复杂性[53]和相关的功耗的超线性增加，而应用程序性能却没有线性加速。因此，这种架构级的优化很快就达到了收益递减的地步。此外，处理器和主存储器（DRAM）之间指数级增长的性能差距导致了存储墙[79]，其中系统性能由内存性能主导；使处理器更快（无论是IPC还是时钟速度）不会影响应用程序的执行时间。ILP墙和存储墙一起阻止了在单处理器设计中引入更多的复杂性。

在这一点上，为了有效地利用芯片上越来越多的晶体管，计算系统向多核架构进行了不可逆转的过渡[24]。多核处理器在单个芯片上包含两个或多个处理元件（处理器核）。随着晶体管密度从一个制程到另一个制程的增加，单个核心的频率和复杂性保持不变，甚至降低，以适应登纳德缩放定律的失效。因此，单核性能保持不变或下降。大量利用线程级并行性（TLP）的处理器核弥补了单核性能的不足，其中线程分布在各个处理器核上并并行执行。如果代码可以适当地并行化，即使每个核心都以较低的频率运行，应用程序的性能也会大大提高。第一个商用双核处理器是2001年推出的IBM Power 4，而桌面平台的第一个双核处理器（Intel Pentium Extreme Edition 840）是在2005年推出的。到目前为止，从智能手机到个人电脑再到企业服务器，多核在所有计算系统中都很普遍。

过去十年设计的大多数多核架构都是同构多核，即所有核心在指令集架构（ISA）和底层微架构（流水线、告诉缓存、分支预测配置）方面都是相同的。同构多核很容易设计和验证，因为我们只需要复制处理器核。因此有人预测，芯片上的核心数量将按照摩尔定律呈指数级增长。这一趋势一直保持到目前为止；例如，最新的英特尔至强处理器E7由15核组成。然而，就像单核性能的上升轨迹停止一样，简单地增加同构多核的核数也不会长期保持性能提升。

**黑暗硅时代**

Esmailzadeh等人在一篇开创性的论文[21]中通过结合设备扩展、单核扩展和多核扩展来模拟多核扩展限制，以预测未来五代技术中并行工作负载的加速潜力。该论文的结论是，功率和并行性限制将在可实现的性能与摩尔定律（每个技术节点的性能加倍）预期的性能之间造成很大的差距（在8纳米技术节点上有ITRS预测[13]，用于设备缩放）。更重要的是，增加核心计数在某一点后达到功率预期。除此之外，还可以添加更多核心；但这些额外的核心必须关闭或显著降低时钟，以满足TDP约束。在任何时间点，很大一部分核心要么处于空闲状态（dark），要么处于低时钟状态（dim），这种现象被称为“暗硅”[49]。该研究声称，功率限制将严重限制可用芯片的比例：在22纳米，21%的芯片将是暗的，在8纳米，超过50%的芯片将不会使用ITRS器件缩放预测。参考文献[64]已经从硬件/软件协同设计的角度讨论了暗硅时代的架构决策、电源/热管理挑战和可靠性/可变性挑战。

有趣的是，在暗硅时代，并行性而不是功耗被证明是导致大多数工作负载性能差距的主要因素[21]。很少有应用程序具有足够的并行性，但由于功率限制导致的核数不足，仍然无法实现性能潜力。大多数当前和新兴的应用程序都不能完全并行化，无法利用同构多核架构提供的丰富的TLP。**阿姆达尔定律**[2]指出，多核上的加速将被限制在，其中是应用程序的串行分割数。即使对于具有99%并行代码的应用程序，剩余的1%顺序分数也会在内核数量无限的情况下将速度限制在100倍。因此，即使具有无限的功率预算，大多数应用程序的同构多核加速也有限。

为了面对即将到来的暗硅时代，Taylor[72]将潜在的架构上的应对方法分为四类。最明显和最悲观的方法是缩小芯片，芯片设计者制造更小（简单）的芯片。这种方法不仅阻碍了所有架构的进步，而且从成本、收入、功率和封装问题的角度来看也是有问题的。第二种可能是暗硅，其中逻辑被严重欠频以保持在功率预期以下。暗硅技术包括基于近阈值电压计算的多核[20]，更大的片上缓存而不是逻辑来吸收暗硅区域，通过动态电压频率缩放（DVFS）进行空间调光，以及通过计算冲刺[61]或英特尔Turbo Boost[63]进行时间调光，其中芯片允许暂时超过功率预算以提供短暂但显著的性能提升，随后是长时间的低功耗运行。“暗硅”仅对高度并行的工作负载有效；但是，由于缺乏并行性，而不是功率，是大多数工作负载实现高速加速的主要障碍，因此这种方法并不具有普遍的吸引力。更加面向未来的解决方案是超低功耗电路，其中晶体管由新兴技术构建，提供比CMOS器件更好的亚阈值特性[34]，因此可以大幅提高能源效率。最后一类，也是我们在本文中重点讨论的一种有前途的方法，是异构多核架构。

在暗硅时代，相对于功耗，硅面积成为一种指数级的廉价资源，架构师可以潜在地“花费”面积来“购买”能效[72]。这是异构多核架构背后的指导原则，其中不同类型的核心共存于同一芯片上。给定一个应用程序，只有最适合该应用程序的核心才能被激活，从而实现更快、更节能的计算。异构计算架构可以大致分为两类：性能异构和功能异构。

# 3 性能异构多核

性能异构多核架构由具有不同功耗-性能特征的核心组成，但它们具有相同的指令集架构。这种差异源于不同的微体系结构特性，比如顺序核心和乱序核心。复杂的核心可以以更高的功耗为代价提供更好的性能，而简单的核心则表现出低功耗行为和较低的性能。这也被称为单ISA异构多核架构[39]或非对称多核架构。这种方法的优点是，相同的二进制可执行文件可以根据上下文在所有不同的核心类型上运行，并且不需要额外的编程工作。然而，系统设计人员或运行时管理层必须为每个应用程序甚至同一应用程序中的不同阶段确定适当的核心类型。根据在运行时是否可以配置核的混合，非对称多核可以进一步分为静态非对称多核和动态非对称多核。

## 3.1 静态非对称多核

在静态非对称多核中，不同核类型的混合在设计时是固定的。例如，**图1**显示了一个由12个小核和1个大核组成的静态非对称多核。商用非对称多核的例子，包括ARM big.LITTLE[26]，集成了高性能乱序核和低功耗顺序核，nVidia Kal-El（品牌名称为Tegra3）[52]由四个高性能核和一个低功耗核组成，而最近来自Ineda的可穿戴处理器（WPU）由具有不同功耗-性能特征的核组成[33]。如**图2**所示，ARM big.LITTLE的一个实例集成四核ARM Cortex-A15（大核）和四核ARM Cortex-A7（小核），出现在三星Exynos 5 Octa SoC上，并应用于三星Galaxy S4智能手机。

性能异构的非对称多核架构有望解决当今两个相关但关键的问题：Amdahl定律[2]造成的性能限制和暗硅时代多核计算的能源效率。Hill和Marty在一篇开创性的论文[30]中提出，异构多核可以将顺序代码片段映射到一个复杂的核上——例如，通过乱序执行，能够利用指令级并行性（ILP），从而加速顺序部分的执行——极大地提高了应用程序的加速。这是因为代码的并行部分可以通过提供TLP的简单核心阵列来加速，而顺序部分可以通过利用复杂核心的ILP来加速。

异构计算的能效优势非常明显。在任何时候，我们都只需要在不影响性能的情况下打开当前计算需求中最节能的核心。例如，在智能手机中，低功耗的小核心可以处理简单的任务，如电子邮件客户端，网页浏览等，从而节省能源，而复杂的核心必须开启计算密集型任务，如3D游戏，浏览基于flash的网站等，从而牺牲能源。这种计算模型非常适合暗硅时代，因为热约束无论如何都限制了可以在任何时间点打开的核心的比例；因此，打开合适的核心以获得更好的能源效率是有益的。请注意，除了微架构差异之外，这些架构还通过内核的动态电压-频率缩放（DVFS）在功耗-性能权衡曲线中提供了额外的设计点。

**图3**显示了商用ARM big.LITTLE架构处理器异构多核架构的功耗-性能异构性[60]。我们使用的评估平台是Versatile Express开发平台，该平台由一个带有两个Cortex-A15核心和三个Cortex-A7核心的原型芯片组成，采用45纳米技术。所有核心都实现了ARM v7A ISA。虽然每个核心都有私有的L1指令和数据缓存，但L2缓存在集群中的所有核心之间共享。跨集群的L2缓存通过缓存一致互连保持无缝一致，以便应用程序可以轻松地从一个集群迁移到另一个集群。该体系结构为每个集群提供DVFS特性。但是集群中的所有核心应该在相同的频率级别上运行。此外，如果需要，可以关闭空闲集群的电源。该芯片配备了测量每个集群的频率、电压、功率、能耗的传感器和性能计数器。

**图3**分别绘制了Cortex-A7和Cortex-A15集群上基准测试应用程序的指令周期（IPC）和平均功耗（瓦特）。在这个实验中，我们为两个集群设置了相同的电压（1.05伏）和频率（1 GHz）水平，并且一次只使用一个核心来运行基准测试。注意，我们只能测量集群级别的功率，而不能测量单个核心级别的功率。因此，即使只有一个核心在运行基准测试应用程序，而其他核心处于空闲状态，该图中报告的功耗也对应于集群中的功耗。显然，与A7相比，A15具有更好的IPC（平均加速比1.86），但功耗行为要差得多（平均功率是A7的1.7倍）。

**软件支持**

利用异构多核提供的独特机会来实现节能计算需要复杂的运行时管理技术。这包括（a）确定最适合应用或应用阶段的核心类型，（b）通过任务迁移将应用移动到适当的核心，以及（c）为核心设置适当的电压频率水平，以便以最小的能量满足应用的性能要求，同时不超过热设计功率（TDP）限制。

在这个过程中需要的第一步是一个准确的功耗-性能估计机制。当应用程序在一种核心类型上运行时，需要预测它在其他核心类型和不同电压频率水平上的功耗和性能行为，从而决定是否应该迁移应用程序以及迁移到哪里。文献[60]中已经开发了对于ARM big.LITTLE架构的这种功耗-性能模型。由于各种原因，这种建模对于真实的体系结构来说是具有挑战性的。首先，大核心和小核心不仅在流水线组织中，而且在缓存层次结构和分支预测器方面都有很大的不同。因此，给定一种核心类型的缓存缺失率或分支错误预测，我们必须估计另一种核心类型的缓存缺失率或分支错误预测。其次，我们受到核心上可用的性能计数器的限制，不能假设根本不可用的额外分析信息，例如指令间依赖关系。这些挑战可以通过静态程序分析（识别指令间依赖关系）、基于对底层架构的理解构建分析模型的机制建模（例如，由于指令间依赖关系和资源约束而导致的流水线阻塞的影响，而不是错过性能统计事件），以及使用统计推断技术（如回归）创建分析模型（用于核间缺失事件估计）的经验建模。

然后可以利用这些预测为应用程序或应用程序的阶段及其DVFS选择合适的核心类型。为了优化性能，Craeynest等人[73]提出了一种非对称多核调度技术，使用跨不同核类型的在线性能估计。类似地，Koufaty等人[38]提出了一种动态异构感知调度器，该调度器在复杂核心上调度具有非常低内存延迟的任务以获得更高的性能。但是这些工作并没有考虑到功耗问题。Winter等人的一项研究[77]评估了异构多核的各种调度和电源管理技术，并特别考虑了这些方法的可扩展性。他们提出了一种称为“最陡下降”的线程调度算法，该算法开销很小，但没有考虑核心的频率变化。

文献[51]提出了一种基于控制理论的方法，该方法协同集成多个控制器（处理不同的约束或优化目标），以实现运行在异构多核系统上的多个应用程序的能源效率。然而，这种方法存在可扩展性问题，这是由于在严格的TDP约束下，关于任务迁移和核心之间的功耗分配的集中决策。这种可扩展性问题是通过一种分布式方法来解决的，该方法基于参考文献[68]中经济学价格理论的坚实基础。资源分配、DVFS、任务映射和迁移都是通过虚拟市场进行控制的，在虚拟市场中交易的商品是使用虚拟货币的处理单元。该框架被实现为称为代理的自治实体的集合，每个代理对应一个任务、核心、集群和整个芯片。将性能要求建模为需求，而将处理能力建模为供应（取决于核心类型和频率）。价格理论原则指出，市场只有在价格均衡时才稳定，价格均衡是指供给与需求相等的价格，因此对应于最小的能源消耗。在一系列工作负载中，基于价格理论的电源管理框架将平均功耗降低到2.96 W，而Linux异构感知调度器（做出简单的任务迁移决策）和按需调控器（用于DVFS）在相同甚至更好的性能水平上的平均功耗为5.99 W。

## 3.2

尽管静态非对称多核在适应软件多样性(ILP和TLP工作负载的混合)方面明显优于同构多核，但它们仍然不是理想的解决方案。由于在设计/制造期间必须冻结简单核和复杂核的混合，因此静态不对称多核缺乏调整自身以适应工作负载动态特性的灵活性。应用程序需求的任何变化都会对生产成本产生很大的影响。支持多样化和动态工作负载的下一个合乎逻辑的步骤是设计动态非对称多核，它可以在运行时根据应用程序定制自己[30]。这种自适应架构在物理上是由一组简单的、同构的核心组成的。在运行时，两个或更多这样的简单核心可以合并在一起，以创建一个更复杂的虚拟核心。类似地，参与复杂虚拟核心的简单核心可以在任何时间点分离。一个典型的例子是形成两个2路外序(0)核的联盟，以创建一个4路外序核。换句话说，我们希望通过简单的重构动态创建静态非对称多核。图1显示了一个动态非对称多核体系结构，它由16个基本核组成，在运行时配置为创建一个中核和一个大核。下一节介绍了文献中提出的一些动态非对称多核架构。 核心融合架构[35]提供了一个详细的架构解决方案来支持运行时核心合并。物理基板由相同的、相对高效的2问题无序内核组成。在运行时，这些核心可以融合在一起，以创建更大的(8期或4期)乱序核心。所提出的体系结构具有可重构的分布式前端和指令缓存组织，可以利用单个核心的前端结构来提供积极的融合后端，同时最小化单个前端的过度供应。 Kumar等人对联合核芯片多处理中的核心联盟提供了另一种解释[40]，其中相邻的核心可以共享复杂的结构(例如，浮点单元、交叉端口、指令缓存、数据缓存等)，从而节省了大量的面积。