并行计算技术的现状与未来

张明昆 2211585

2024 年 3 月 16 日

目录

1 引言

并行计算是一种计算模式，它利用多个处理单元同时执行计算任务，显著提高了计算速度和效率。这一概念的起源可以追溯到 20 世纪 60 年代早期，当时的科学家们开始探索使用多个处理器来解决大规模科学计算问题。自那以来，随着微电子技术的飞速发展，各种并行处理器和并行计算机系统相继诞生。并行计算的重要性体现在其对解决大规模、复杂问题的能力上。

在天气预报、气候变化模拟、生物信息学、材料科学、大数据分析等领域，传统的串行计算已经无法满足计算需求。并行计算技术的应用，不仅大大缩短了计算时间，还使得一些以往认为不可能解决的问题变得可行。

调研的目的在于：一方面，通过分析最新的并行计算技术，了解当前并行体系结构的技术水平和发展方向；另一方面，通过对比不同国家和地区在并行计算领域的发展历程，探讨并行计算技术对国家科技进步和经济发展的重要作用。

2 当前并行体系结构的概览

无论是在超级计算机的巨大规模还是个人计算设备的微观层面，并行计算技术都在不断推进计算能力的边界。接下来，我们将深入到这些技术在各个领域的具体应用和表现，从而更全面地理解并行体系结构的进步与发展趋势。

2.1 超级计算机

Frontier: 位于美国橡树岭国家实验室的 Frontier 超级计算机，是世界上首台达到 exaflop（每秒执行一百亿亿次浮点运算）水平的系统。它结合了高性能的 CPU 和 GPU，使用 Cray 的 Slingshot 网络，保证了高效的节点间通信。Frontier 利用了 AMD 的 Epyc CPU 和 Radeon Instinct GPU 的组合，整套系统包括 9,472 个 CPU 和 37,888 个 GPU，总计 CPU 内核数达 606,208 个，GPU 内核数达 8,335,360 个。[1] 通过这种 CPU 和 GPU 协同工作的并行计算模型，Frontier 实现了高效的数据处理和能源利用率，体现了异构并行计算技术的先进性。

Aurora: 位于阿贡国家实验室的 Aurora 超级计算机，自 2023 年以来，它一直是全球第二快的超级计算机。预计其性能优化后将超过 2ExaFLOPS，成为有史以来最快的计算机。[2] Aurora 配备了基于 Sapphire Rapids-SP 系列的 Xeon CPU 和基于 Ponte Vecchio 设计的 GPU，这些高性能处理单元可以提供强大的计算能力，适应各种高性能计算需求。特别是 GPU 的大量使用，标志着并行计算中对于特定任务（如图形处理、数据并行任务等）的处理能力的重视。

神威 • 太湖之光: 位于中国无锡的神威 • 太湖之光，是世界上首个使用中国自主研发的处理器的超级计算机。其主要特点是高能效比，适用于大规

模数值模拟、天气预报、生命科学等多个领域。神威 • 太湖之光使用的是国产的申威 26010 处理器，整套系统高达 40,960 个 SW26010 处理器，共有 10,649,600 个 CPU 核心。每个处理器为一个节点单元，一块主板上有两颗处理器，32 块这样的主板组成一架主机，每台主机作为一个“超级节点”，一共有 256 个这样的超级节点。[4] 基于片上网络的设计，处理单元之间可以高效地通信。

2.2 CPU 技术

Intel 酷睿 14 系列：代表了 Intel 在桌面和移动处理器市场的最新进展，酷睿 14 系列继续沿用并改进 Alder Lake 的混合架构设计，提供了更多的 P-cores 和 E-cores。该设计通过在同一芯片上集成不同类型的核心，实现了任务并行性的优化，使得处理器可以根据任务的需求动态调整，将计算负载高效地分配到适合的核心上。通过 Intel Thread Director 技术实现了硬件和操作系统的协作，优化 P-cores 和 E-cores 的使用。Windows 11 系统还对该系列 cpu 进行了优化，以更好地支持混合架构的线程调度，确保高性能核心处理更加需要计算能力的任务，而高效率核心则处理背景任务，提高了系统的整体效率。[5]

Intel Sapphire Rapids 系列：该系列处理器是针对高性能计算和数据中心市场推出的，Sapphire Rapids 采用的是 Intel 的新一代 Xeon Scalable 处理器架构，引入了 EMIB (嵌入式多芯片互连桥) 和 Foveros (3D 堆叠技术)，允许在单一处理器内部集成更多功能和高效率的数据传输。Sapphire Rapids 也是 Intel 首款支持 HBM 的 Xeon 处理器，该技术可以为处理器提供更高的内存带宽，特别适合于大数据量和高并行度的计算任务。Sapphire Rapids 还引入了更高级的矢量处理能力（AVX-512 扩展），以及针对人工智能应用优化的深度学习加速指令（DL Boost），特别是 AMX（Advanced Matrix Extensions）技术，为矩阵运算提供了专门的硬件加速，大大提高了深度学习和其他高性能计算应用的效率。[5]

AMD Zen 4 架构：Zen 4 架构继续使用多核心设计，支持高线程计数，通过增加核心数量和优化线程管理来提升并行处理能力。采用了改进的核心芯片设计（CCD）和集成输入/输出介质（IOD）。CCD 内包含多个 CPU 核心，而 IOD 则负责处理内存控制、PCIe 接口和其他 I/O 功能。使得数据在处理器内部的传输更加高效。该架构增加了 L3 缓存的大小，提高了缓存的命中率，从而减少了对主内存访问的需要。Zen 4 架构还使用了 TSMC

的 5 纳米制程技术，大幅提高了能效比。[6]

苹果 M3 系列处理器：在 M1 和 M2 系列的基础上，M3 系列处理器为苹果设备带来了更高的性能和能效。采用先进的 SoC 设计，集成了多核 CPU、GPU 和 AI 加速器等处理单元，这些单元可以高效协同工作，执行复杂的并行计算任务。特别在图形处理和机器学习等领域，M3 系列处理器的并行计算能力使苹果产品在性能和能效比上保持领先。

2.3 GPU 技术

英伟达 Ada Lovelace 架构：英伟达的 Ada Lovelace 架构采用了专为深度学习计算设计的第四代 Tensor 核心，提供了前所未有的 AI 推断和训练性能。该架构采用了第三代 RT 核心，专门用于光线追踪计算，这些核心可以并行处理大量的光线计算任务，极大提升了实时渲染性能。还增加了更多的 CUDA 核心，这些核心可用于广泛的并行计算任务，从图形渲染到科学计算等。采用了高带宽的 GDDR6X 内存，为并行计算任务提供了充足的数据传输速率。[7]

AMD RDNA3 架构：RDNA 3 架构进一步优化了计算单元（CU），提高了每 CU 的性能和效率，增加了更大的 Infinity Cache，减少了对外部内存的依赖，提高了数据访问速度，从而加速并行计算任务。还通过优化光线追踪单元，提高了并行处理光线追踪任务的能力。[6]

Intel Ponte Vecchio：该系列采用了 Xe-HPC 微架构，在浮点运算和矩阵计算方提供了更强大的并行计算能力。采用了高带宽内存技术，为大规模并行计算任务提供了巨大的内存带宽。通过 3D 堆叠技术，Intel Ponte Vecchio 实现了更高密度的核心集成。支持 OAM（Open Accelerator Module）标准，为 Intel Ponte Vecchio 提供了灵活的互连和扩展能力。[5]

3 并行体系结构的进步与发展趋势

3.1 技术进步

并行体系结构的进步主要体现在能效比、计算性能和集成度等方面。通过对比不同代的 CPU 和 GPU 架构，我们可以明显看到这些进步。

能效比：随着制程技术的进步，例如从 14nm 到 7nm 再到 5nm 甚至更先进的制程，CPU 和 GPU 的能效比得到了显著提高。这意味着在消耗相

同电量的情况下，新一代的处理器能完成更多的计算任务。例如，AMD 的 Zen 3 与 Zen 4 架构相比，后者在相同功耗下提供更高的性能。

计算性能：新一代的 CPU 和 GPU 通过增加核心数、提高时钟速度和优化架构设计，实现了计算性能的大幅提升。例如，苹果 M3 系列相比 M1 系列，在保持能效的同时，性能提升显著，这得益于其更多的处理核心和更高效的架构设计。

集成度：系统级芯片（SoC）的出现，将 CPU、GPU、AI 加速器等多种处理单元集成在一个芯片上，显著提高了集成度。这不仅减少了芯片间的数据传输延迟，也提高了整体系统的性能和能效。例如，苹果的 M 系列处理器就是一个典型的 SoC 设计，其集成了多种功能，为不同的应用场景提供了强大的并行计算能力。

3.2 发展趋势

更高的能效：随着技术的发展，未来的并行体系结构将继续追求更高的能效比。这意味着新一代的处理器将在消耗更少能源的同时，提供更强的计算性能。

AI 与传统计算的融合：AI 的广泛应用要求处理器具备强大的 AI 计算能力。因此，我认为未来的并行体系结构将更加注重 AI 计算与传统计算的融合，通过专门的 AI 加速器或优化的指令集来提升 AI 应用的性能和效率。

异构并行计算模型的普及：为了适应不同计算任务的需求，异构并行计算将成为一种趋势。这种模型结合了 CPU、GPU 以及其他类型的处理器的优势，能够针对特定任务提供最优的计算资源，从而提高计算效率和性能。

量子计算与并行计算的结合：随着量子计算技术的进步，未来或许会看到量子计算与传统并行计算的结合。量子处理器在特定类型的计算任务上展现出超越传统处理器的性能，其与传统并行计算技术的结合必将开启并行计算的新篇章。

4 中国与世界的超级计算机发展对比

4.1 中国的发展历史

中国的计算机工业始于 20 世纪 50 年代，通过借鉴苏联的技术，于 1958 年 8 月 1 日研发出国内首台数字电子计算机——103 机。[8] 随着时间的推

移，进入 1970 年代，中国对超级计算机的需求急剧增加，这主要是由于中长期天气预报、风洞实验模拟、三维地震数据处理以及新武器的开发和航天事业等领域对计算能力提出了更高的要求。

为了满足这些需求，中国开始致力于超级计算机的研发。1983 年 12 月 4 日，军方主导成功研制出银河一号超级计算机。[9] 随后，银河二号、银河三号、银河四号等系列的银河超级计算机相继问世，使中国成为全球少数几个能够发布 5 至 7 天中期数值天气预报的国家之一。1992 年，中国又成功研发了曙光一号超级计算机。随着时间的发展，人们发现向量型计算机由于自身的局限性难以继续发展，因此转向了并行型计算机的研发。基于这一新方向，中国研发了神威超级计算机及其后续型号神威蓝光超级计算机。[10] 2015 年，采用全国产 CPU 的神威 • 太湖之光超级计算机达成了世界第一的速度成就，并在软件设计方面获得了次年的戈登贝尔奖，成为继美国和日本之后，第一个获此殊荣的国家，打破了长达 30 年的获奖格局。[11]

在民间领域，2002 年联想集团研发成功的深腾 1800 型超级计算机及其系列产品标志着中国超级计算机技术的又一重大进步。同时，深圳大学自主研发了 KD 系列超级计算机，主要注重于减小体积和降低电力消耗，而非追求最高速度，这种设计使其可能被应用于军用或航天领域。

4.2 国际发展历史

国际上，超级计算机的发展可以追溯到上世纪 60 年代，美国和欧洲一直是这一领域的领导者。早期的控制数据公司机器可达十倍速于竞争对手，但仍然是比较原始的标量处理器。到了 1970 年代，大部分超级计算机就已经是向量处理器了，很多是新进者自行开发的廉价处理器来攻占市场。1980 年代初期，业界开始转向大规模并行计算系统，这时的超级计算机由成千上万的普通处理器所组成。1980 年代中叶，将适量的向量处理器（一般由 8 个到 16 个不等）联合起来进行并行计算成为通用的方法。1990 年代以后到 21 世纪初，超级计算机则主要互联基于精简指令集的张量处理器（譬如 PowerPC、PA-RISC 或 DEC Alpha）来进行并行计算。美国的 Cray-1（1976 年）、日本的富岳（2020 年）、美国的 Frontier（2021 年）等，都曾是或者是当前世界上最快的超级计算机。

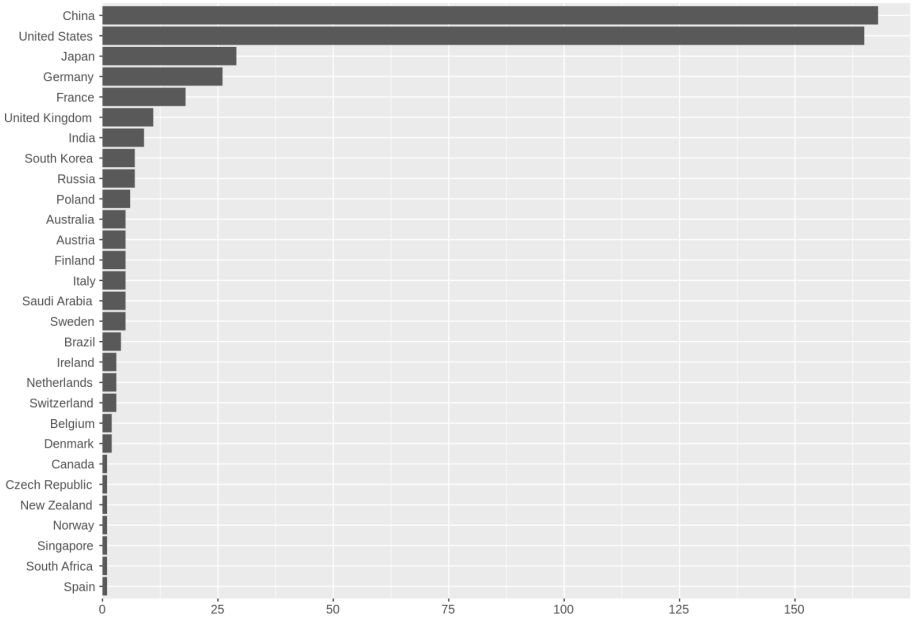

图 1: 2016 年各国超算算力柱状图

4.3 重要性与并行体系结构的发展趋势分析

超级计算机对国家发展的重要性不言而喻，它们在科学研究、国防安全、经济规划和社会管理等多个方面发挥着关键作用。

在气候模拟、生物信息学、材料科学和量子物理等领域，超算是进行高复杂度科学计算和数据密集型研究的关键工具。它们能够模拟复杂的自然和人工过程，加速新理论和技术的发展，从而推动科学进步和技术创新。

在国防领域，超算用于武器系统设计、加密与解密、战场模拟等关键任务，是维护国家安全的重要资产。通过模拟和分析，超算有助于提升武器性能，优化战略部署，增强防御能力。

并行体系结构是超级计算机发展的核心。随着科技进步，其发展呈现以下趋势：

多核和众核处理器：为了进一步提升计算性能，超算正向着使用更多核心的处理器发展。众核处理器可以处理更多的任务，提高计算效率。

异构计算：结合不同类型的处理器（如 CPU、GPU 和 FPGA）进行计算，可以针对不同的计算任务选择最合适的处理器，从而提高效率和能效

比。

高效能源管理：随着超算性能的提升，能耗问题日益突出。因此，开发高效的能源管理技术，降低功耗，是并行体系结构发展的一个重要方向。

软件和算法的优化：并行计算的软件和算法同样重要，它们需要与硬件发展同步，以充分利用并行体系结构的计算能力。

参考文献

[1] Charles Q. Choi, "The Beating Heart of the World's First Exascale Supercomputer," *IEEE Spectrum*, June 24, 2022, archived from the original on August 14, 2022, retrieved August 14, 2022.

[2] "Intel Data Center GPU Max Series Overview," Intel, retrieved November 14, 2023.

[3] 理化学研究所计算科学研究センター (R-CCS). スーパーコンピュータ「富岳」プロジェクト. [2019-06-03].（原始内容存档于 2020-05-14）[引用日期: 2019-06-03].（日语）

[4] "China Tops Supercomputer Rankings with New 93-Petaflop Machine." www.top500.org. [2016-06-20].（原始内容存档于 2019-05-31）.

[5] "Intel Corporation Official Website." www.intel.com. [访问日期: 2024-03-13].

[6] "Advanced Micro Devices, Inc Official Website." www.amd.com. [访问日期: 2024-03-13].

[7] "NVIDIA Corporation Official Website." www.nvidia.com. [访问日期: 2024-03-13].

[8] 中国科学院计算技术研究所. "103 机 - 中国科学院计算研究所官网". http://www.ict.ac.cn/kxcb/kxtp/200909/t20090922_2514368.html [2019-05-31].（页面存档备份，存于互联网档案馆）

[9] "银河到天河中国超级计算机发展大事记 - IT168 网". http://server.it168.com/a2009/1029/800/000000800121.shtml [2009-10-29].（页面存档备份，存于互联网档案馆）

[10] “全国产化超级计算机神威蓝光问世耗电量极低 - 东方网”. http://mil.eastday.com/m/20111031/u1a6179680.html [2011-10-31].（页面存档备份，存于互联网档案馆）

[11] “Chinese research team wins top award in supercomputing”. [2017-03-10].（原始内容存档于 2016-11-20）.