南开大学
Nankai University

南　　开　　大　　学

计 算 机 学 院

并行程序设计实验报告

并行体系结构调研

2011763 黄天昊

年级：2020 级

专业：计算机科学与技术

指导教师：王刚

2023 年 3 月 9 日

摘要

本文调研了当前超级计算机的发展历史、发展趋势以及面临的挑战，深入探究了曾经位列 TOP500 榜单第一的富岳超算的体系结构与核心技术。

关键字：超级计算机；富岳；体系结构

目录

一、 超级计算机——人类算力的巅峰

(一) 发展历史

随着科技发展，人类社会对计算的需求呈爆炸式增长，普通计算机已无法满足一些前沿科学研究的计算，超级计算机应运而生。1929 年,《纽约世界报》首次提出超算这一概念；1976 年，美国克雷公司开发出世界首台超级计算机 Cray-1，它的速度达到了每秒 1 亿次运算。1993 年，德国曼海姆大学教授汉斯、埃里克等人创建了全球超级计算机 TOP500 排行榜，为高性能计算机提供统计和排名。该排行榜用 Linpack(Linearsystem Package) 程序进行基准测试，以浮点运算实际测试值 HPL 为排序依据，每年 2 次发布榜单。

美国长期居于 TOP500 榜单之首，直到 2002 年日本推出超算“地球模拟器”。2011 年日本研制的超算“京”再次登顶世界第一。此外，欧洲通过多国联合，共同研制超算。2010 年“超级计算机合作平台计划”启动，由欧洲多国参与，旨在提升全欧洲的超级计算能力。中国研发的天河二号于 2013-2016 年登顶，随后研发的神威太湖之光于 2016 年-2018 年登顶。

超级计算机功能最强、运算速度最快、存储容量最大，因此被用于高精尖的国防科技领域。一个国家超级计算机的算力在某种程度上反映了该国的信息技术实力，是国家综合国力的重要标志。

名称	国家	位居世界第一的时间	Rmax(Tflops)	Rpeak(Tflops)
蓝色基因/L	美国	2004.11-2008.6	478.2	596.4
走鹃	美国	2008.6-2009.11	1105	1456.7
美洲虎	美国	2009.11-2010.11	1759	2331
天河一号	中国	2010.11-2011.6	2566	4701
京	日本	2011.6-2012.6	10510	11280.4
红衫	美国	2012.6-11	16324.8	20132.7
泰坦	美国	2012.11-2013.6	17590	27112.5
天河二号	中国	2013.6-2016.6	33862.7	54902.4
神威太湖之光	中国	2016.6-2018.6	93014.6	125435.9
顶点	美国	2018.6-2019.11	148600	200794.9
富岳	日本	2019.11 至今	442010	537212

图 1: 近 15 年 TOP500 榜首计算机

(二) 发展趋势及现状

图 11中，我们可以看到，Rmax 和 Rpeak 几乎以指数级增长，每次新的榜首出现，其速度都是上届榜首的近两倍；此外，当前世界已迈入 E(Exascale) 级计算时代。未来超级计算机的发展趋势如下:

1. 更高性能

- 多态体系结构
- 更多 CPU 核或 CPU+GPU
- 更快的工作主频

  以 Intel 为例，由于 MOS 电路物理限制的原因，CPU 的主频达到 3G 左右就难以提高。这是未来后摩尔时代需要解决的问题。

- 更快的存储系统

  存储芯片包括 DRAM、SRAM、FLASH、PRAM、RRAM、MRAM。目前 MRAM 应用前景较好，但在未来还需提高容量、性能和可靠性，以及降低功耗。

2. 更快互联

光电混合的高速互联技术是未来超算的发展趋势。由于光的传播速度快、干扰小，数据可以通过光电转化从一个计算节点到另一个计算节点，同时计算节点和存储节点间也会有更快的互联。

3. 更高可靠性

- 故障诊断、定位和隔离
- 智能容错例如使用多个节点进行计算时而发现有个别节点无法使用时，可以只使用剩下一半的节点继续计算。部分情况下，该方法比传统的重新分析的方法更节省时间。
- 任务高效迁移将出错任务的节点退回到某一固定点，然后迁移到另一节点继续计算的方法。

4. 更高能效

能效为 FLOPS/WATT，是 Green500 榜单的排名标准。即使节能的天河二号功耗也达 17 兆瓦，神威太湖之光则至少需 30 兆瓦，而 E 级计算则需配备一个独立的发电站；若不改变方式，到 2024 年全球超算所需的电力据估计将超过全球发电量。因此，未来超算的另一发展趋势将是超导计算机。由于超导材料无电阻不发热，故所需电量只有传统计算机的 1/40-1/1000，而预算速度可以超越 E 级。

5. 面临挑战

- 存储访问墙：指处理器的处理速度和访存速度之间的不匹配导致并行系统计算效率的下降问题。
- 通信墙：指互联网络对计算性能的影响。
- 可靠性墙：由于并行度的不断扩大，系统可靠性越来越弱，系统经常出现故障，影响机器性能。
- 能耗墙：指功耗问题对超算发展的阻碍。

二、 富岳——曾经的 TOP1

富岳曾雄踞超算榜单榜首（现已被 Frontier 超越），富岳的前身是日本超级计算机“京”（K-Computer），由日本理化研究所和富士通共同研发。2014 年日本政府启动 Post-K 国家项目，历史 7 年建成投产。富岳即富士山，寓意对高性能和泛用性的追求，二者之间的矛盾给富岳的设计制造提出了巨大挑战。

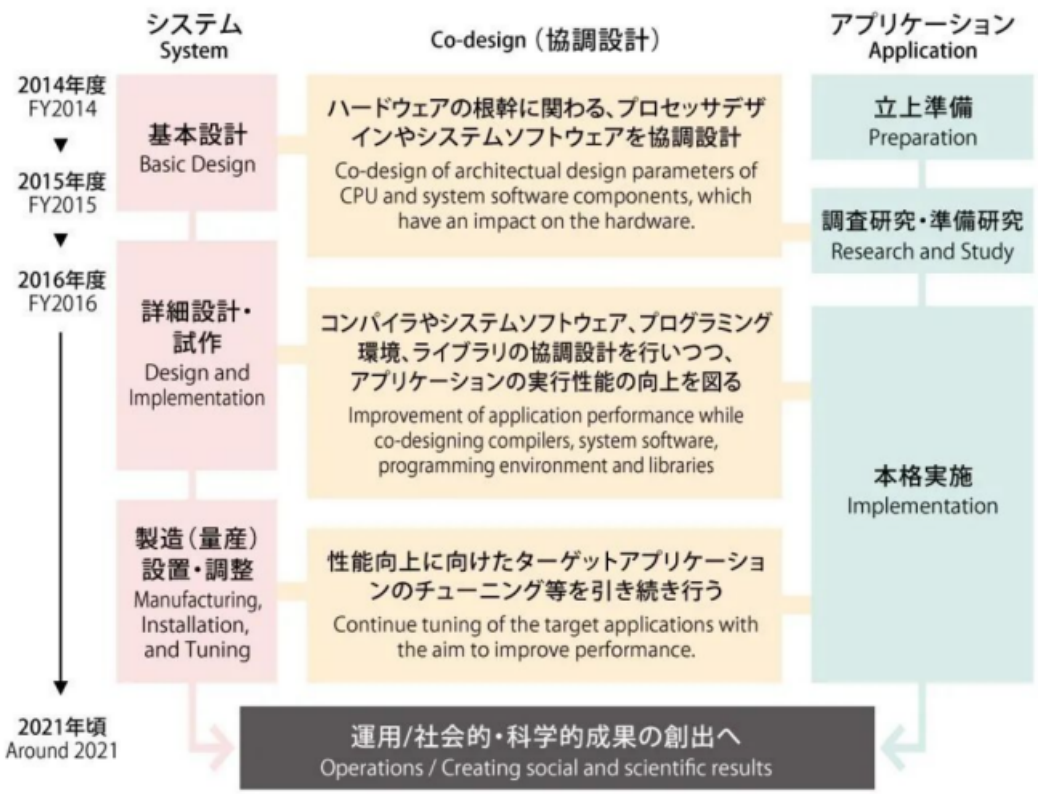

图 2: 富岳研发时间表

(一) 基本结构和技术介绍

1. 富士通 A64FX 处理器

A64FX 处理器集成了 48+4 个核心,配备 32GB HBM2 内存,带宽 1TB/s,浮点性 2.7TFLOPS,使用 7nm 工艺生产，平均是主流高性能服务器级 CPU 的三倍性能和能效。使用 Arm v8.2 指令集并首次使用 Arm SVE（可伸缩向量指令集)。

采用分布式结构，CPU(特别是二级缓存)、内存 (HBM2) 和网卡都通过带有多个直接内存访问控制（DMAC）的片上网络连接。这允许任何处理器节点通过远程直接内存存取（RDMA）访问系统中 160K 个节点的的任何存储器区域，并以小于微秒级别的延迟与二级内存交换数据。核心能实现任意拆分和组合。

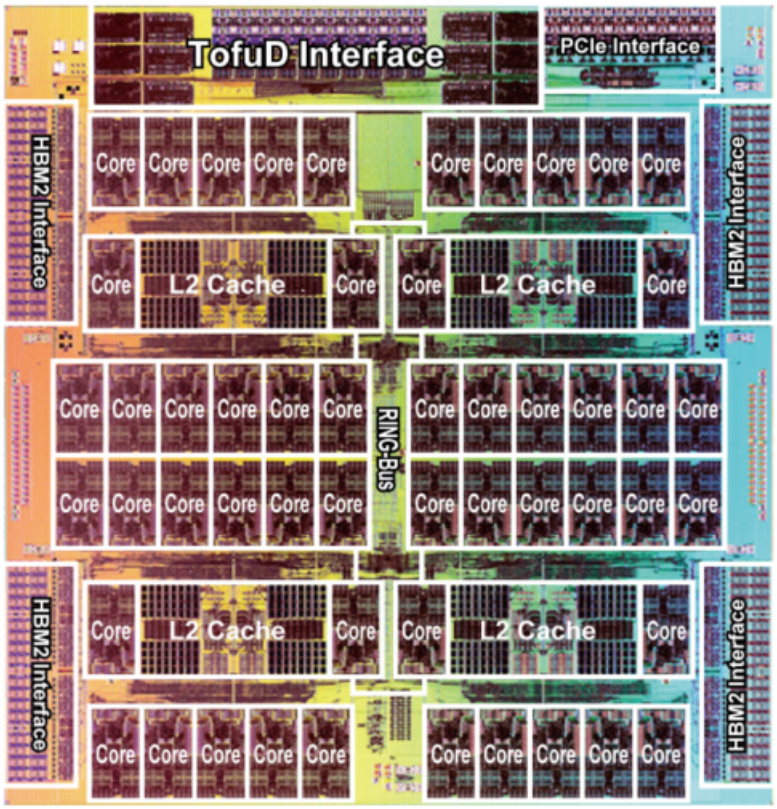

图 3: A64FX 结构

A64FX 基本上是替换指令集的 SPARC64 XIfx，从核心微架构到基础结构皆大同小异，系统存储器改用包在一起的 32GB HBM2，大幅精简空间。

与应用于“京”和 PRIMEHPC FX100 上的前代 CPU 相比，如图4所示，A64FX 在向量计算（加宽的 SIMD 指令集）和 AI 计算（FP16 和 INT16/INT8）上分别有提升和新技术应用，硬件屏蔽技术（Hardware Barrier）在硬件层面实现了并行计算同步问题。

A64FX Features

FUJITSU

- Collaboration with Arm to develop and optimize SVE for a wide range of applications
  - FP16 and INT16/8 dot product are introduced for AI applications

	A64FX (Post-K)	SPARC64 XIfx (PRIMEHPC FX100)	SPAR64 VIIIfx (K computer)
ISA	Armv8.2-A + SVE	SPARC-V9 + HPC-ACE2	SPARC-V9 + HPC-ACE
SIMD Width	512-bit	256-bit	128-bit
Four-operand FMA	✓ Enhanced	✓	✓
Gather/Scatter	✓ Enhanced	✓	
Predicated Operations	✓ Enhanced	✓	✓
Math. Acceleration	✓ Further enhanced	✓ Enhanced	✓
Compress	✓ Enhanced	✓	
First Fault Load	✓ New		
FP16	✓ New		
INT16/ INT8 Dot Product	✓ New		
HW Barrier* / Sector Cache*	✓ Further enhanced	✓ Enhanced	✓

* Utilizing AArch64 implementation-defined system registers

图 4: A64FX 与前代 CPU 特性对比

2. Tofu-D 互联结构

节点之间的数据交换常常会有竞争的情况，且一个节点发生错误需要隔离并排除故障。为了解决这一问题，富岳采用了 Tofu-D 的高维互联结构，组间采用超立方体结构，而组内采用 2*3*2 的网格结构，最低能实现 0.5 s 的点到点延迟和 38 GByte/s 带宽。如图5所示：

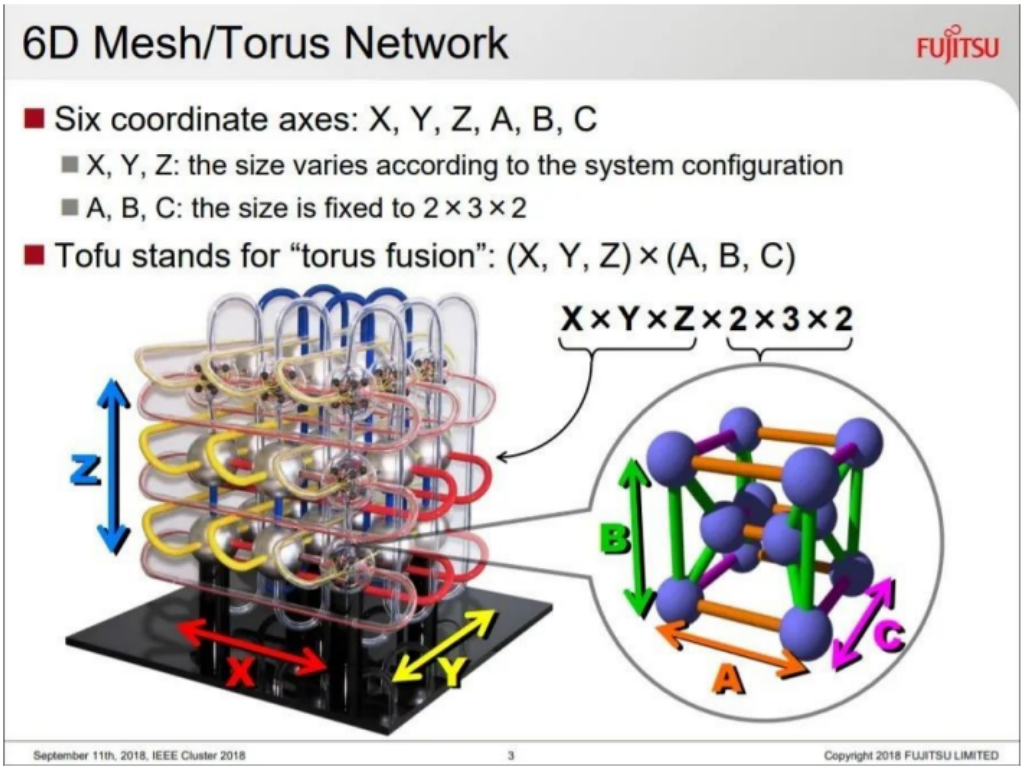

图 5: Tofu-D 互联结构

这样的结构不仅提升了节点之间的互连程度，而且通过将组间环路连接和组内网格连接虚拟化，在处理器发生故障时，可以从虚拟三维空间中排除故障位置，减少数据流量拥塞，并最大限度地减少维护替换的分区隔离，提高系统的可用性。

3. 协处理

富岳采用了协处理（Co-design）设计模式，在靠近数据的地方进行计算可以减少延迟。富岳在 2014 年设计之初就采用了这样的设计，将 A64FX 分为四个计算内存组（Core Memory Group），每个组由 12 个计算核心和 1 个辅助核心组成，与边缘计算有异曲同工之妙。如图6所示：

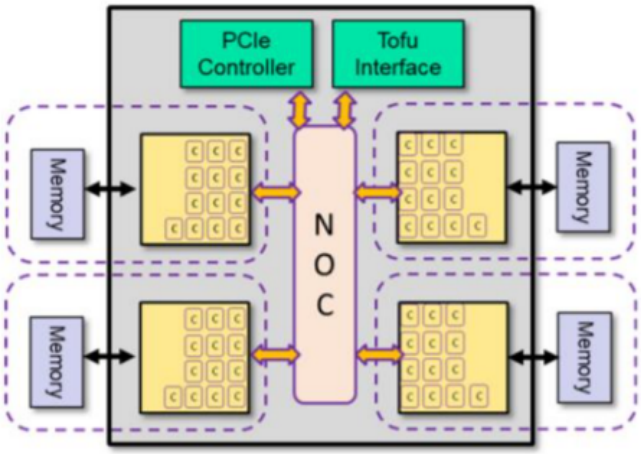

图 6: A64FX 架构——12+1 协处理单元

4. SVE 伸缩向量指令集

SVE (Scalable Vector Extensions) 扩展指令集是专门为高性能计算科学工作负载的向量化而开发，允许将可变向量长度实现为 128 到 2048 位。在 A64FX 中，每个计算核心都拥有双流水线 (2 pipelines)，支持 SVE 的 512 位 SIMD，大幅增强了浮点计算能力。

(二) 同构与异构

1. 计算单元

根据是否使用相同指令集和体系结构的计算单元组成，可以将超算分为同构和异构两种体系结构。计算单元包括 CPU，GPU，FPGA，DSP 等。

2. 理论性能、效率与能耗

PU 天然适合并行编程，具有较高的浮点计算能力，而 CPU 适合计算控制与任务分配，二者组合的异构模式广泛应用于目前的超级计算机上。以曾经的 TOP500 第二名 SUMMIT 为例，使用 CPU+GPU 的异构结构。Summit 使用了 IBM Power 9（PowerPC 架构，RISC 指令集）及 NVIDIA Tesla V100（GPU，单精度性能 15.7TFLIPS，双精度性能 7.8TFLOPS，功耗 250W）。同功率下，GPU 提供的浮点运算能力大于 CPU，因此 CPU+GPU 的架构具有较高的理论性能。

然而，计算富岳与 Summit 的性能功耗比之比，却只得到约 1.004，即二者的性能功耗比大致相同。这是因为加速单元（如 GPU）的性能会直接影响到超算的性能，如天河 2 号的 Xeon-Phi 在运行 Linpack 测试时，实际性能仅仅相当于理论最大运算性能的 65%-70%

3. 技术与编程

CPU 适合做串行，逻辑复杂度高的任务，GPU 适合做简单，并行度高的任务，CPU 和 GPU 的编程模型是不一致的；所以异构超算在编程方面就不太容易，例如 GPGPU 编程就需要协调 GPU 和 CPU，就不如直接在 CPU 上编程容易。另外，GPU 和 CPU 不共享内存，显式拷贝内存会导致性能损失。

在 Summit 超算上，为了提高效率，使用了如下的技术思想：简化 CPU 线程和 GPU 之间的交互；在 GPU 上实现计算内核并减少 CPU-GPU 内存传输；允许异步 GPU 通信；通过组合线性代数运算来增加计算强度。显然，异构计算单元之间的需要有效的通信并且应该尽量在本地处理数据。

4. 专用与通用

CPU 的通用计算能力强，因此同构超算的适用面广。GPU 长于浮点运算，有较高运算性能和资源，目前的 GPGPU 技术便旨在将 GPU 运算能力推广到一般计算中。另外，由于人工智能计算的需求，NPU 等适用于此类计算的单元也在涌现，并配置在设备中，起到异构架构的作用。

5. 总结

正如大小核的出现是为了适应不同计算量的需求，面对如今多种多样的计算需求，专用性和通用性并存的局面，使用异构技术能在理论上发挥各单元的计算优势，达到提高计算能力并节省功耗的目的，但本质上存在着效率降低的限制，需要通过多种技术减少 CPU-GPU 间的内存传输并优化计算单元间的通信以提高效率。相比之下，同构架构的兼容性、可编程性、效率都优于异构架构。如何在合适的场景权衡性能、效率、功耗是设计的难题。

三、 结语

本文调研了当前超级计算机的发展历史、发展趋势以及面临的挑战，深入探究了 TOP500 榜单第一的富岳超算的体系结构，从 CPU 的内核分布和较前代的技术进步、互联结构、住协处理、可伸缩向量指令集角度介绍。最后结合 Summit 等异构超算的优劣特性，提出对同构和异构模式的思考。

通过对并行计算机体系结构的调研，我开拓了对于并行计算机的认知（计算单元、存储结构、互联方式、指令集、编程方式等角度），也在计算机如何权衡性能和功耗等方面产生了新的思考，收获良多。

参考文献

[1] 司宏伟, 冯立升. 世界超级计算机创新发展研究 [J]. 科学管理研究,2017,35(04):117-120

[2] 张云泉, 袁良, 袁国兴, 李希代.2020 年中国高性能计算机发展现状分析与展望 [J]. 数据与计算发展前沿,2020,2(06):1-10.

[3] 方粮. 超级计算机发展现状及趋势分析 [J]. 智能物联技术,2020,3(05):1-8.