南开大学

Nankai University

南　　开　　大　　学

计　算　机　学　院

并行程序设计作业报告

超级计算机体系结构调研

2013154 段辰睿

年级：2020 级

专业：计算机科学与技术

指导教师：王刚

2023 年 3 月 11 日

摘要

在本文中首先简要介绍超级计算机的发展历史，然后介绍超级计算机的体系结构迭代演变过程，之后从体系结构和技术两个方面分析超级计算机，并由此具体对目前最快的超级计算机富岳的体系结构进行了分析，最后对超算系统的未来发展做了一定的预测与评价。

关键字：超级计算、富岳、并行、体系架构

目录

一、 绪论

(一) 超级计算机介绍

超级计算机是 1929 年《纽约世界报》中最先报道出的一个名词，它是将大量的处理器集中在一起以处理庞大的数据量，同时运算速度比常规计算机快许多倍，是相对于大型计算机而言的一种运算速度更高、存储容量更大、功能更完善的计算机。它通常是指每秒中能运算 5000 万次以上、存储容量超过百万个字节的电子计算机。

超级计算机的性能评价目前以每秒钟浮点运算次数（flops）为主要衡量因素，目前最先进的超级计算机的每秒钟浮点运算次数均可达到十亿亿（10 17 ）次。根据 TOP-500 榜单，目前最新的数据显示（2022 年 11 月），目前超级计算机富岳实测峰值速度可以达到 442.01PFlops/s，理论峰值速度为 537212TFlops/s。而在 2010 年 10 月的榜单中，速度最快的超级计算机天河一号的实测峰值速度为 2566TFlops/s，理论峰值速度为 4701TFlops。根据以上数据可以看出，在 2010 年-2022 年间，超级计算机的速度增加了上百倍，超级计算机的发展迅速。而能够拥有如此快速的计算速度和如此强大的计算能力，这离不开超级计算机内部的并行体系结构。

本文第二章介绍了超级计算机体系结构的发展历史，第三章从体系结构和软件技术两方面分析了超级计算机的技术，第四章分析了超级计算机的发展趋势，第五章对于目前最快的超级计算机富岳的体系结构进行了具体剖析，第六章对两种不同架构的超算进行了对比分析，第七章做了总结。

二、 超级计算机发展历史

(一) 概述

总体上来说，超级计算机经历了三次革命性的发展：向量超级计算；多 CPU 上的并行超级计算；分层组织的、带有高速缓存的微处理器集群上的超级计算。

(二) 向量超级计算

1972 年，超级计算机设计师西摩·克雷创办了克雷公司并在 1976 年研制出具有流水结构的向量机 Cray-1。向量机是面向向量型并行计算，以流水线结构为主的并行处理计算机，采用先行控制和重叠操作技术、运算流水线、交叉访问的并行存储器等并行处理结构，大幅提高运算速度。这种向量机有着较高的性价比，持续计算能力成本与当时的成本性能冠军 Apple II 微型计算机的成本相当；不仅如此，它采用向量体系结构，其中浮点数的向量可以从存储器加载到向量寄存器中，并在算术单元中以流水线方式处理，速度比 CDC 6600 等前代标量运算数高得多，向量处理也成为后代超级计算机的理论基石。

(三) 多 CPU 上的并行超级计算

诞生于 1966 年的 ILLIAC IV 是第一台大规模并行计算机，它在设计之初具有 256 个 64 位浮点单元（FPU）和 4 个中央处理单元（CPU），提供高达 1 GFLOPS 的速度。80 年代，业界开始大规模转向大规模并行运算系统。大规模并行处理机（MPP）是由多个由微处理器，局部存储器及网络接口电路构成的节点组成的并行计算体系，节点间以定制的高速网络互联。大规模并行处理机是一种异步的多指令流多数据流，因为它的程序有多个进程，它们分布在各个微处理器上，每个进程有自己独立的地址空间，进程之间以消息传递进行相互通信。90 年代是 MPP 大规模爆发的时代，处理器个数由以前的几个迅速增长到几千个。

(四) 带有高速缓存的微处理器集群上的超级计算

具有大量处理器的系统通常采用两条路径，其一网格计算，是把处理器组成分布式系统，分别管理着计算机的计算能力，其二则是许多处理器彼此靠近使用，称为计算机集群。与网格计算机不同，计算机集群将每个节点设置为执行相同的任务，由软件控制和调度。集群的组件通常通过快速局域网相互连接，每个节点运行自己的操作系统实例。计算机集群的出现是许多计算趋势汇聚的结果，这些趋势包括低成本微处理器、高速网络以及用于高性能分布式计算软件的广泛使用。在这样的集中式大规模并行系统中，互连的速度和灵活性变得非常重要，现代超级计算机已经使用了从增强的 Infiniband 系统到三维环形互连的各种方法。

三、 超级计算机技术分析

(一) 体系结构方面

1. 原理

超级计算的基本原理是并行计算，可以节省计算时间，处理大型问题并提高精确度。整个问题被分解成若干份，每一部分由一个处理器计算。然而事实上问题通常不能很好的被分解成独立的部分，各个部分之间需要进行交互，包括计算中的数据传送和同步，因此超级计算的性能优化之一是提高并行可扩展性。

2. 架构

超级计算机可以按照并行计算方式是单指令多数据流(SIMD)还是多指令多数据流(MIMD)，存储器是共享还是分布进行分类。

早期的超级计算机系统以 SIMD 方式工作，由于系统中各处理器按阵列方式排布，所以又被称为阵列处理机。阵列处理机的专用性较强，一般只适合于求解某类算法，工作效率往往很高。

如今的超级计算机系统大多以 MIMD 方式工作。多向量机(Multi Vector Processor，MVP)系统中有多套向量部件，但存储器是共享的，因此属于 SM-MIMD 类型。对称多处理器(Symmetrical Mutil-Processing，SMP) 也属于这一类型。SMP 是指在一个计算机上汇集了一组处理器，多 CPU 各个之间共享内存子系统以及总线结构。MVP 与 SMP 又称为 UMA（Uniform Memory Access）系统，因为系统中所有处理器对任何存储单元有相同的访问时间。与 UMA 相对的是 NUMA 系统，在 NUMA 系统中，存储器是分布式的，各访问时间和处理器对同一存储单元的访问时间可能是不同的，依赖于处理器在系统中所处的具体物理位置。如果并行计算机系统中的处理器必须以消息传递的方式访问远程存储器，就称为 NORMA（No Remote Memory Access）系统。与 NUMA 系统不同，它有多个存储器地址空间，且系统中的每个处理器是一个独立的计算机。NORMA 系统按计算机间的互连紧密程度，又分为紧耦合和松耦合两种。集群系统是松耦合的典型代表，而 MPP 系统则是紧耦合的典型代表。MPP 一般以通用 64 位微处理器作业处理节点，多为分布存储方式，节点间通信用消息传递方式，其规模可扩展到数千节点。优点是峰值速度高，并有良好可扩展性，主要缺点是消息传递能力与节点运算能力难以匹配。集群系统中每个节点是一个完整的计算机，各节点服务器通过内部局域网相互通讯，每台服务器的操作系统和应用程序文件存储在其各自的本地储存空间上。

21 世纪后，多线程、多核技术应运而生，将异构并行计算架构引入超级计算机中。异构并行技术，需要有效开发计算任务的并行性，与机器不同部件支持的计算类型最佳匹配，以充分利用各种计算机资源，神威·太湖之光、天河二号与天河-2A 等顶尖超级计算机都采取异构并行的处理器架构。

(二) 编程技术方面

自 20 世纪末以来，基于超级计算机体系结构的变化，超级计算机操作系统发生了重大变化。虽然早期的操作系统是为每台超级计算机量身定制的，以提高速度，但趋势是从内部操作系统转向适应通用软件（如 Linux）的发展。由于现代大规模并行超级计算机通常通过使用多种类型的节点将计算与其他服务分开，因此它们通常在不同的节点上运行不同的操作系统。大多数现代超级计算机都使用基于 Linux 的操作系统，但由于硬件体系结构的差异也会针对硬件去设计优化操作系统。

超级计算机的并行体系结构通常要求使用特殊的编程技术来开发程序，从而完全发挥其高速计算的优势。通常用于分布式处理的软件工具有 MAPI、PVM、VTL 等标准 API 和 Beowulf 等开源软件。对于异构计算，则需要使用 CUDA 或 OpenCL 之类的编程模型进行编程。

四、 超级计算机发展趋势

20 世纪 80 年代到 90 年代中后期，超级计算机的研究中热度最高的是并行处理、高性能计算等技术基础；1996 年到 2004 年期间，评价类研究方向如基准问题测试、热度提高等后续部分研究热点陆续出现；2004 年以来，能源效率、程序设计模型、图形处理器等应用类方向成为研究重点。

目前的超级计算机发展呈现多极化趋势。MPP 系统、集群系统的应用进一步提高了超级计算机的性能,，各个国家和科研机构正在 E 级超级计算机的研制中激烈竞争。但随着超级计算机的速度不断发展，其功耗也不可避免的持续增加，如何在提示性能的同时减少功耗也成了研究的重点。

未来计算机的发展趋势可能会有以下几点:

1. **速度大幅提升**
   当前世界 TOP500 的超级计算机基本都可达到 P 级运算速度，目前各国普遍预计到 2023 年，E 级超级计算机将登上历史舞台，因此 E 级超级计算机是当前世界各国竞相角逐的战略制高点。从 P 级计算到 E 级计算不简单地是一个性能指标上的提升，E 级计算在能耗、性能、可扩展性、可靠性、生态环境、应用编程、应用效率与适应性、多领域应用融合等诸多方面面临着前所未有的挑战。
   速度的大幅提升需要硬件设施和体系结构的同时发展，目前 CPU 体系结构的发展趋势是向众核模式发展，同时核之间的快速连接方式也是持续发展的方向。

2. **与 AI 融合**
   随着全球移动互联网和物联网的发展，人类可利用的数据以爆炸的状态增长。这些海量的数据借助于深度学习可以为人类创造难以估量的价值。在未来，高性能计算与大数据融合的趋势会越来越明显，并会在很多领域得到广泛应用。因此面向应用优化的高性能计算系统研发、智能化的系统管理调度等将成为发展趋势。

五、 具体分析–富岳

富岳（Fugaku）是富士通与日本理化学研究所共同开发的超级电脑，于 2020 年 6 月以 415 PFLOPS 的计算速度成为 TOP500 排名第一的超级计算机，为第二名美国超算 Summit 速度的 2.8 倍。

(一) 节点组成

富岳也是全球首台获得 TOP500 榜首的 ARM 架构超算，采用富士通 48 核心 A64FX 芯片，是日本超算京（Kei）的后代产品。富岳由 432 个机架组成，其中 396 个机架各拥有 384 个节点，其余 36 个机架各拥有 192 个节点，节点数共计 158,976 个，每个节点包含一个 A64FX CPU。富岳共拥有两种节点，计算节点和计算和 I/O 节点，他们之间通过 6D Mesh/Torus, Fujitsu TofuD 连接。其中 Tofu 代表 Torus Fusion ， D 代表 Density 和 Dynamic ，意为高节点密度、动态分组切片及其带来的网络故障恢复能力。这种新的互联方式也是富岳除了 ARM 架构之外的另一个亮点。

(二) 处理器架构与性能

富岳在 Boost 模式下，CPU 主频可以达到 2.2GHz，双精度浮点运算理论峰值性能为 537 PFlop/s, 同时也支持半精度浮点和整数（8 位）运算。富岳共有内存 4.85PiB，并提供内存带宽 163PB/s。富岳的高性能离不开富士通开发的高性能处理器 A64FX。A64FX 是富士通于 2019 年发布 CPU，采用 Armv8.2-A 指令集，是第一个采用 SVE（Scalable Vector Extensions，ARM 指令集体系结构的矢量扩展）扩展指令集的 CPU。A64FX 采用 7nm FinFET 工艺制程生产，内含 87.86 亿个晶体管，基础频率 2.0GHz，Boost 模式下频率可以达到 2.2 GHz。他拥有 48 个计算核心和 2/4 个辅助核心（用于操作系统活动）组成，分为 4 个 CMG（Core Memory Group）单元，每个单元由 13 个核心，1 个 L2 Cache 和 1 个内存控制器组成。每个物理内存空间是分离的，并且由硬件隐式保证缓存一致性。富岳使用 8GB HBM 2（High Bandwidth Memory Gen2），带宽 256GB/s，共计 32GB 内存，带宽 1024GB/s。

在浮点运算性能上，A64FX 每核拥有双流水线 SVE 512 位 SIMD，而每个 SIMD 可以同时执行两条 FMA 指令，因此单核每周期可提供 2 pipelines × 512 bit × 2 FMAs / 64 bit = 32 FLOPS 的双精度浮点性能。若以 2.2Ghz 的频率运行，则每颗 CPU 最高可提供 32 FLOPS × 2.2 GHz × 48 cores = 3379.2 GFlop/s 双精度浮点性能。

SVE 是 ARM 指令集体系结构的矢量扩展。SVE 定义了最多 2,048 位的矢量长度作为指令集，具有从 128 位的倍数中选择和实现硬件矢量长度的能力。A64FX 支持长度为 128、256 和 512 位的向量。这种独特的扩展指令也为 A64FX 的高性能贡献了巨大力量。

(三) 网络

富岳的独特连接网络 TofuD 是 Tofu 系列最新产品。其采用 6D Mesh 网络，六个维度为 X, Y, Z, A, B, C，其中 (A, B, C) 根据系统配置确定，(A, B, C) 固定为 $(2, 3, 2)$。TofuD 还采用了虚拟 3D-Torus Rank-mapping 技术、Tofu Barrier 等。相比于前代，具有更先进的评估环境、更低的延迟，并增加了 Tofu Barrier 资源，在注重吞吐量和注入率的同时也保持高效率。

(四) 编程环境

富岳采用 Red Hat Enterprise Linux 8 和 McKernel (provided by RIKEN) 操作系统，MPI 支持 Fujitsu MPI (based on Open MPI), RIKEN MPI (based on MPICH)，文件 I/O 系统采用 LLIO (provided by Fujitsu) 和 Application-oriented file IO libraries (provided by RIKEN)，编程语言支持 Fortran、C++、OPENMP、Java、Python 等。

（五） 基于 ARM 架构的典型

富岳作为第一台登顶 TOP-500 的 ARM 架构的超级计算机，其计算速度是第二名 Summit 的近三倍，而与此同时其功率也是 Summit 的近三倍，这一反我们通常对于 ARM 架构低功率的认知。在 x86 指令集中也存在这与 SVE 类似的扩展指令集 AVX，而根据英特尔官方消息，AVX 在带来更高性能的同时，CPU 的峰值功耗也提高了。因此根据 AVX 与 SVE 的相似性，富岳的高功耗很有可能也是 SVE 造成的结果。

值得一提的是，富岳并没有采用 GPU 来进行加速。过去数年，超级计算机大多使用英特尔和 AMD 的 x86 架构处理器并配合 GPU 加速，并且排行榜的第二、第三、第四均采用了此架构。这也足以说明 ARM 架构和 SVE 的强大。但 x86 和 GPU 作为传统超算架构，很有可能也可以达到富岳的计算速度，甚至超过富岳。富岳若以最高性能运行，其每秒浮点运算的次数可超过 1000 PFlops。世界上各超算大国均制订了向 E 级超算进军的目标，预计近年内将会由多台 E 级超级计算机诞生。

富岳的高性能是 ARM 和 SVE 处理器 A64FX、TofuD 等众多新技术的结合而得到的，作为唯一一台登顶的 ARM 架构处理器，其也证明了 ARM 架构不仅仅只能用于嵌入式系统，在桌面平台甚至是超算领域也能大放异彩，但也付出了大功耗的代价。x86 架构与 GPU 结合的设计思路仍然是超算领域的设计主流，预计各国都将于今年内推出新一代的 E 级超算。

六、 不同体系架构超级计算机对比

（一） 神威太湖之光与富岳的整体对比

超级计算机	Cores	Processor	Memory	Interconnect
富岳	7630848	A64FX 48C 2.2GHz	5087232GB	Tofu interconnect D
神威太湖之光	10649600	Sunway SW26010 260C 1.45GHz	1310720GB	Sunway

表 1: 硬件数据

超级计算机	Linpack Performance(Rmax)	Theoretical Peak(Rpeak)	Nmax	HPCG[TFlop/s]
富岳	442010 TFlop/s	537212 TFlop/s	21288960	16004.5
神威太湖之光	93014.6 TFlop/	125436 TFlop/	12288000	480.848

表 2: 性能

超级计算机	Power	Power measurement Level
富岳	29899.23 kW(Optimized:26248.36kW)	2
神威太湖之光	15371.00kW(Submitted)	2

表 3: 功耗

从数据上来讲，相较于富岳，神威大湖之光在内存、CPU 主频、运算速度、峰值理论性能等等方面均略逊一筹。

从架构上来讲，神威-太湖之光的核心数要大于富岳，看单核心性能，神威比较来讲很羸弱，28nm 2Ghz 大约是 2Ghz ARM A75-A76 的水平，与 ARM 阵营像苹果 A12、A13 使用的 CPU 核心结构落后很多。但神威强大在于架构先进，核心之间互联通讯的开销小，这也说明了中国仍需提升工艺与单核心性能。同时，SW26010 采用改进型 alpha 指令集，64 位 RISC（精简指令集）架构，专为高性能计算研发。神威太湖之光专为提升运算速度设计，简单的架构正是它运算速度胜过其他高能耗 HPC 系统的原因。不过，申威芯片采用定制 64 位指令集，频率处于中等水平（1.45GHz），而且每个核心只能执行一个线程（不支持超线程），软件支持地没有 Intel 那么丰富。

从发展趋势上讲，全球仍在等待 Exascale，而人 2017 年神威-太湖之光登顶到 2021 年富岳登顶，超级计算机呈现出高速发展趋势，预计在不久之后，运算性能就可以达到 Exaflop 级别的提升。

(二) 编程及软件分析

1. 富岳

富岳采用 FUJITSU Software Technical Computing Suite V4.0 編降器，支持 Fortran 2008 and Fortran 2018、带有 GNU 和 Clang 扩展的 C11、带有 GNU 和 Clang が展的 C++14 和 C++17, OpenMP 4.5 and OpenMP 5.0 以及 java。

富岳的并行編程采用 XcalableMP 和 FDPS。

XcalableMP (XMP) 是由 XMP SpecWG 提出的一种对分布式内存的 PGAS 編程模型和語言，用手提高并行編程的生戶力和性能。XMP 規范 1.4 版本的新特性：混合了 OpenMP 和 OpenACC，具有用于集体通信的库。该语言具有全局视图编程与全局视图分布式数据结构的数据并行性、基手指令的 Fortran 和 C 的 PGAS 模型的語言展的特点。FDPS, 全称 Framework for Developing Particle Simlator，是框架开发粒子模拟器，FDPS 的 api 提供动态负载平衡、节点间的通信和力的计算。

富岳采用 Red Hat Enterprise Linux 操作系统。

2. 神威·太湖之光

神威·太湖之光采用 Raise Linux 操作系统，支持 C、C++、Fortran 编程语言，适用于 MPI、OpenMP、OpenACC 等并行语言及环境。

七、 总结

超级计算机自 1960 年诞生，至今只有 60 年的时间，却从无到有，实现了 E 级运算速度的突破。从单 CPU 到向量机，再到多 CPU 并行运算，再扩大成为带有高速缓存的微处理器集群，最后结合 GPU 的异构计算模式。超级计算机体系结构的一代一代的变迁使得其运算速度越来越快。富岳的诞生又向全世界展示了 ARM 架构在超算领域的实力。超算的发展离不开并行计算，针对并行体系结构开发的并行软件才可以充分利用超算的算力，发挥出超算真实的水平。超算目前的发展现状还是以 CPU+GPU 的异构计算体系结构为主，但也存在如富岳只利用 CPU 的超算。超算未来的发展趋势会向着速度的提升，与 AI 的融合，甚至是其他类型计算机的产出发展。目前各国的 E 级超算研究均已快完成，近年来可能将会有多款 E 级超算诞生。

参考文献 [2] [1] [3]

参考文献

[1] Fujitsu a64fx cpu microarchitecture manual. https://github.com/fujitsu/A64FX.

[2] Supercomputer fugaku. https://www.r-ccs.riken.jp/en/fugaku/project. 2022.

[3] top500. https://www.top500.org/.