南开大学
Nankai University

南　　开　　大　　学

计　算　机　学　院

并行程序设计体系结构调研报告

简析超级计算机富岳体系结构

徐文斌

年级：2020 级

专业：计算机科学与技术

指导教师：王刚

2022 年 3 月 5 日

摘要

在多数超级计算机采用英特尔和 AMD 的芯片组的当下，日本研制的基于 ARM 架构的超级计算机富岳却取得了 TOP500 排行榜榜首的位置。同时，富岳也是 TOP500 排行榜里首个采用 ARM 芯片的超级计算机。从传统观念来看，ARM 是移动芯片，性能比较低。富岳因何可以凭借 ARM 处理器居于 TOP500 榜首，本文将对其体系结构进行较为详细的分析。

关键字：ARM、富岳、体系结构

目录

一、 并行计算机发展历史简述 **1**

二、 超算富岳体系结构分析 **1**

（一） 富岳 CPU「A64FX」处理器架构 1

（二） 富岳 TofuD 互联结构 2

（三） 富岳系统配置 3

三、 总结 **4**

一、 并行计算机发展历史简述

20 世纪 70、80 年代，以 Cray-1 为首的向量处理器问世，其可以在特定的工作环境中极大地提升性能，尤其是在数值模拟等领域。在此期间，向量处理器成为超级计算机设计的主导方向。现在的大多数 CPU 均支持某种形式的向量处理指令，即 SIMD。20 世纪 90 年代，是 MPP(Massive Parallel Processing) 大爆发的时代，处理器个数从原来的个位数开始迅速增长。MPP 架构组件大多是单独定制开发，每个节点使用定制 CPU、运行 OS 微内核，使用单独开发的专有网络连接。进入 21 世纪后，Cluster 得到了蓬勃的发展。Cluster 的节点是一台完整的商业服务器，运行通用操作系统，互联网络使用商业标准的 IB 和以太网设备连接。Cluster 的出现也打破了 MPP 超级计算机的单独定制门槛。

从上述历史简述中，我们可以看到，超级计算机大体的发展趋势为向着更通用、处理器数量更多的方向发展。在当今的超算排行榜 Top500 名单中，Cluster 集群架构占绝大多数，但仍有许多超算采用 MPP 大规模并行处理架构。富岳正是采用 MPP 架构设计的一台超算。

二、 超算富岳体系结构分析

(一) 富岳 CPU「A64FX」处理器架构

富岳的 CPU 为日本富士通公司自研的 A64FX 系列处理器。该处理器在 Arm v8.2-A 中增加了向量运算指令 SVE（Scalable Vector Extensions），成为世界上第一个采用 SVE 扩展指令集的 CPU。但该处理器不支持地址空间为 32 位的旧指令，严格来说不符合 Arm 的规范。处理器支持 FP64/FP32 和 AI 计算用的 FP16 浮点数计算。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。处理器集成了 48 个计算核心，此外还配备了 2 个或 4 个运行 OS 的辅助核心。同时，为了提供更高的内存带宽，富岳使用了堆叠内存芯片的 HBM2 内存。由于 HBM2 容量偏小的限制，每个 CPU 的内存容量固定为 32GB。

CPU 采用台积电 7nm FinFET 工艺制造，运行 dgemm（双精度普通矩阵乘法）时能达到 15GFlop/W 左右的高能效。时钟基础频率 2GHz，睿频可达 2.2GHz。并且 CPU 芯片在普通模式下峰值计算性能为 3TFlops，即使在执行 dgemm 时，计算性能也达到峰值的 90% 以上。内存带宽峰值为 1024GB/s，stream 性能为峰值的 80% 以上。Byte/Flops，即内存带宽与双精度浮点运算的计算性能之比，为 0.33，低于日本早先的曾位于 Top500 榜首的“京”超级计算机的 0.5，但在很多计算性能之比在 0.1 到 0.2 的超级计算机中，富岳具有更高的内存带宽。

如图1所示，四个 CMG（Core Memory Group）通过片上网络（NOC）连接。NOC 还将 GMG 与 Tofu 接口和 PCIe 控制器连接起来。

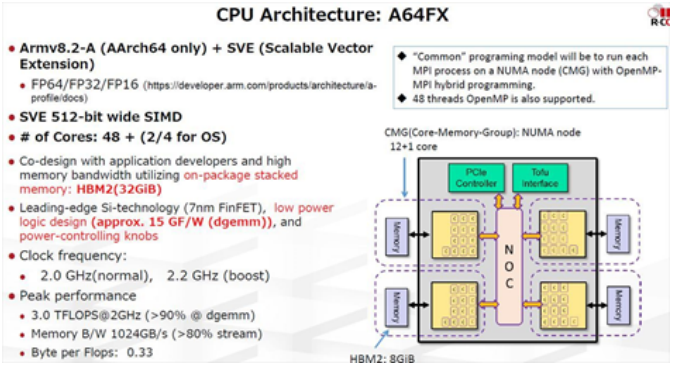

图 1: A64FX Architecture

图2左侧为搭载了 CPU 芯片和四个 HBM2 存储器的封装照片。右侧为 CPU 芯片的示例图，可以看到芯片上共有 52 个核心，分为 4 组，并留有 HBM2 接口、TofuD 接口和 PCIe 接口。

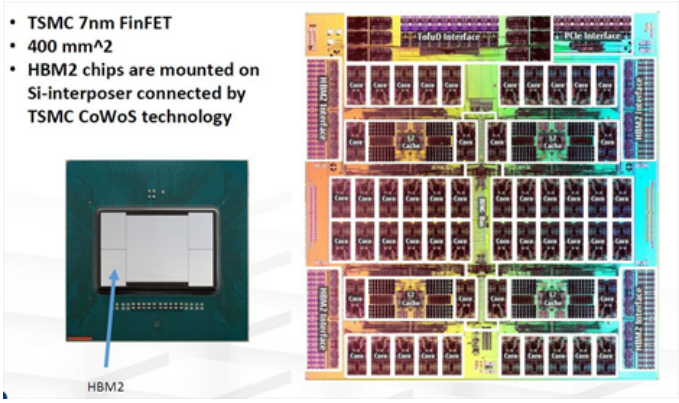

图 2: A64FX Picture

图3右侧为 Intel 的 Xeon Skylake 处理器架构。Skylake 芯片共集成了 18 个核心，而 A64FX 有 48 个核心。A64FX 的核心密度大约是 Skylake 的 3 倍。也可以看出 A64FX 架构更加注重处理器的吞吐量。

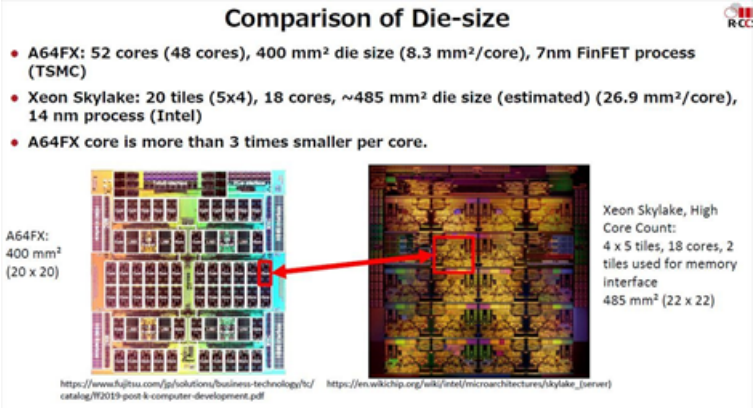

图 3: Comparasion of A64FX and Xeon Skylake

(二) 富岳 TofuD 互联结构

富岳的另一个亮点是互联方式采用了富士通自研 Tofu Interconnect 系列中的 Tofu Interconnect D (TofuD)，其中 Tofu 代表“Torus Fusion”，环形融合；D 代表 High Density 的节点和 Dynamic packet slicing for Dual-rail（双导轨）transfer，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。物理 6D 网络中的节点使用六维坐标 X、Y、Z、A、B、C 表示。其中，A、C 坐标可以是 0 或者 1，B 坐标可以是 0、1、2，X、Y、Z 的坐标值取决于系统的规模。具体如图4所示。

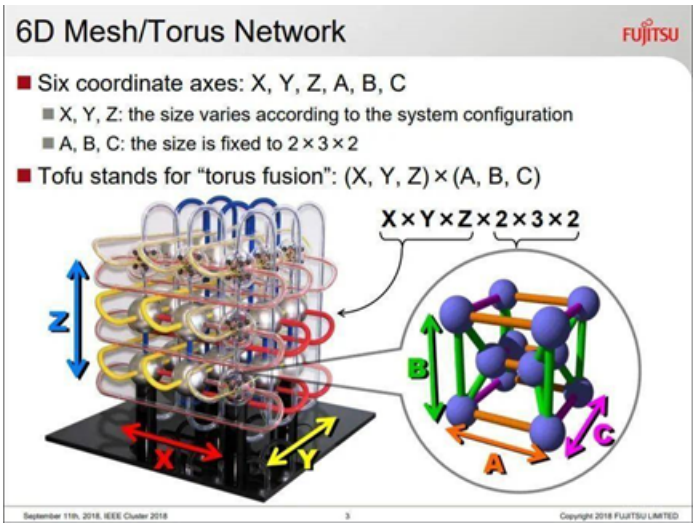

图 4: TofuD Network

6D 网络中，每个节点拥有 10 端口。X 轴、Y 轴、Z 轴和 B 轴各使用 2 个端口，A 轴和 C 轴各使用 1 个端口。每个端口对应的链路提供 5GB/s 的峰值吞吐量。每个链路有 8 条高速差分 I/O 信号通道，数据速率为 6.25 Gbps。每个节点共有 20 个信号通道（Lane），每个 Lane 的数据速率可达 28Gbps。则单个 Link 的带宽为：

$$28Gbps * 2Lane/8 = 7.125GB/s \tag{1}$$

每个节点可同时通信的 Link 数为 6，则 6 个 Link 的带宽一共为：

$$6.8GB/s * 6 = 40.8GB/s \tag{2}$$

TofuD 网络中，6D mesh/torus 网络实现了计算节点的高扩展性，而虚拟的 3D torus rank mapping scheme 则同时提供了高可用和 topology-aware 的可编程性。

(三) 富岳系统配置

如图5所示，富岳系统共有 396 个满配的 Rack 和 36 个半配的 Rack，一个 Rack 有 384 个节点，那么总的节点数目就是 396Full*384+36Half*192=152064+6912=158976。与此相比，“京”计算机有 88,128 个节点，几乎为富岳系统总节点数的一半。

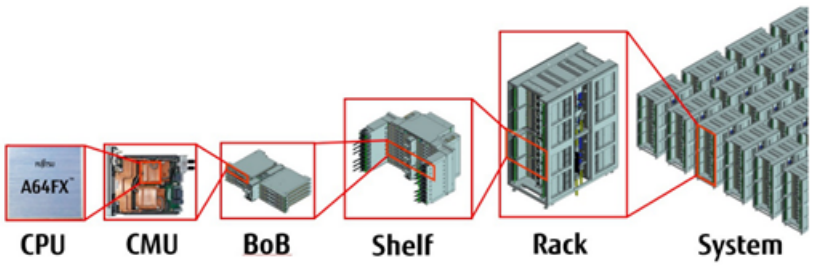

图 5

最后，讨论富岳的存储系统。富岳的存储分为三层，第一层是连接到 16 个计算节点之一的 1.6TB SSD 的存储，这个存储是整个文件系统的缓存。该存储还用于存储临时文件，例如计算

节点的本地文件系统。第一层存储的吞吐量为写 125MB/s/node 和读 293MB/s/node。第二层存储是富士通的 FEFS 存储，容量约 150PB。第二层存储的吞吐量为写 220GB/s/volume 和读 211GB/s/volume。而第三层的存储使用云存储服务。

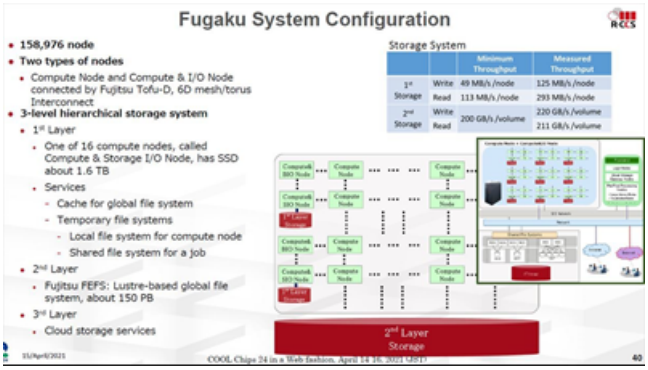

图 6: 富岳系统配置

三、 总结

在并行计算机你追我赶快速发展的当代，富岳超级计算机能够数次居于 Top500 榜首，固然是有许多值得我们学习的方面。比如首次将 Arm 架构用于超级计算机的架构设计，其设计的 TofuD 互联网络也令人眼前一亮。也许在未来会出现出更多的使用 Arm 架构的超级计算机的产生，会有更多新颖高效的互联网络被设计出来。富岳已经达到了 0.5E 级的计算性能。相信在不久的将来，真正的 E 级超算将会问世。[1] [2] [3]

参考文献

[1] Hisa Ando. スパコン「富岳」の cpu「a64fx」. [EB/OL]. 2021[2022-3-2]. https://news.mynavi.jp/techplus/article/coolchips24_fugaku-6/.

[2] Jack Dongarra. Report on the fujitsu fugaku system. [J/OL]. 2021[2022-3-2]. https://www.icl.utk.edu/files/publications/2020/icl-utk-1379-2020.pdf.

[3] TOP500. November 2021. [EB/OL]. 2021[2022-3-2]. https://www.top500.org/lists/top500/2021/11/.