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。

目录

一、并行计算机发展历史简述

二、超算富岳体系结构分析

（一）富岳CPU「A64FX」处理器架构

（二）富岳TofuD互联结构

（三）富岳系统配置

三、总结

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。

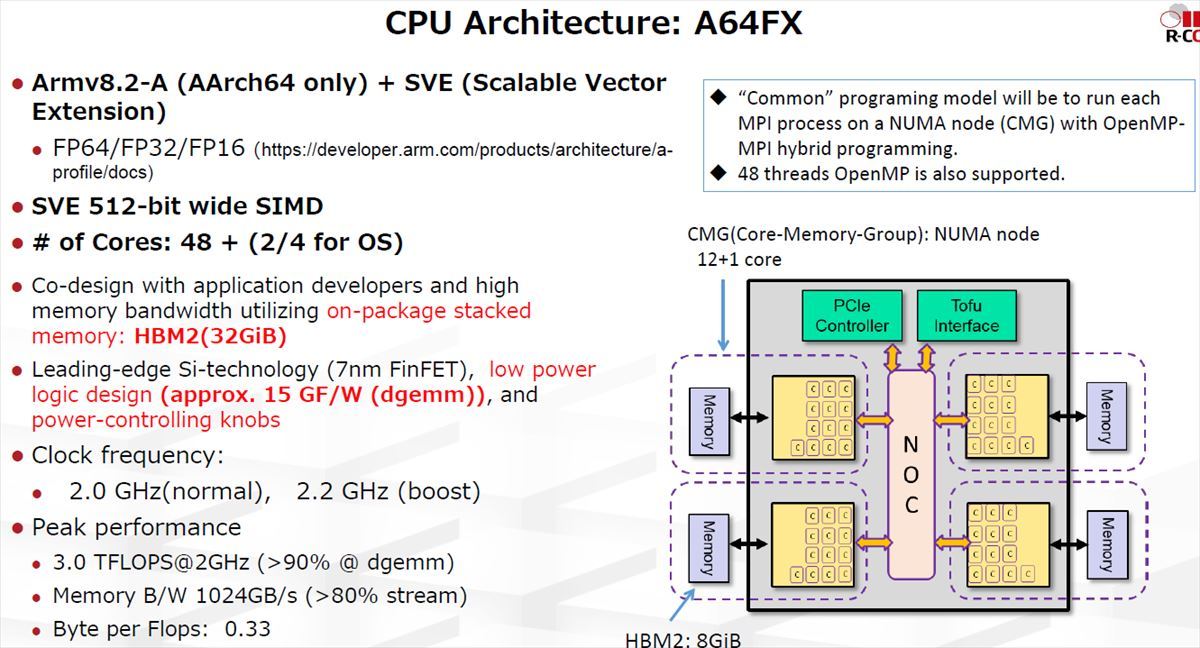


图1

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。

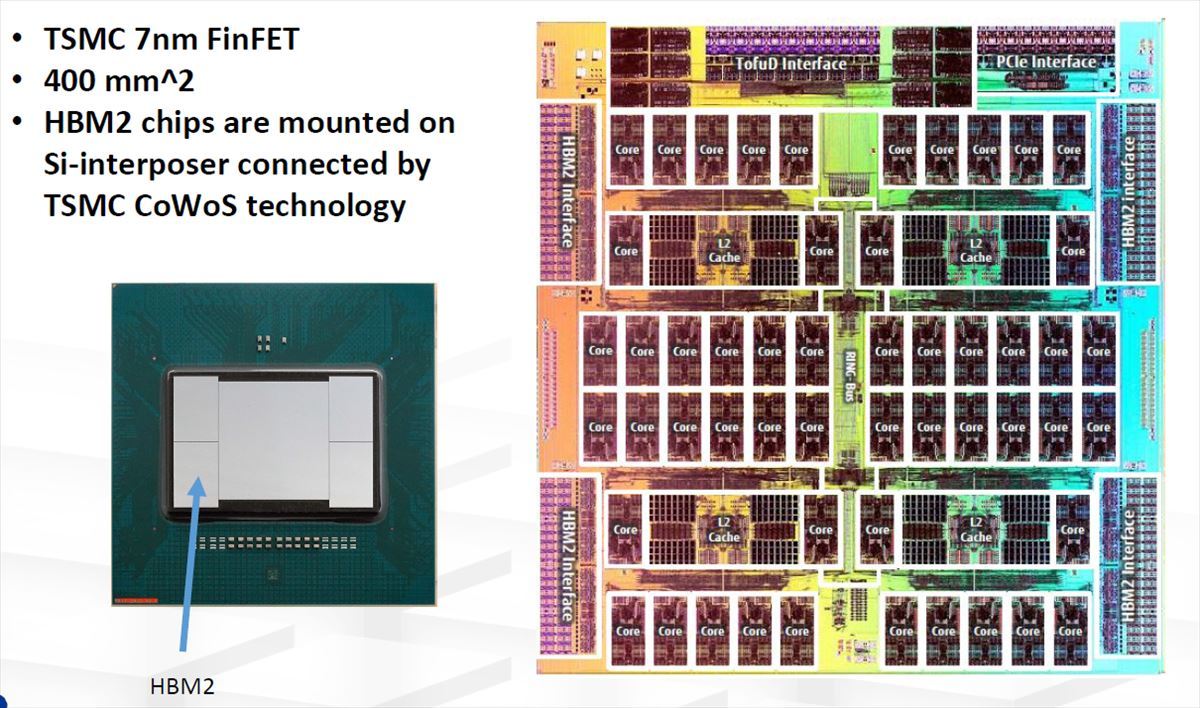


图2

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。

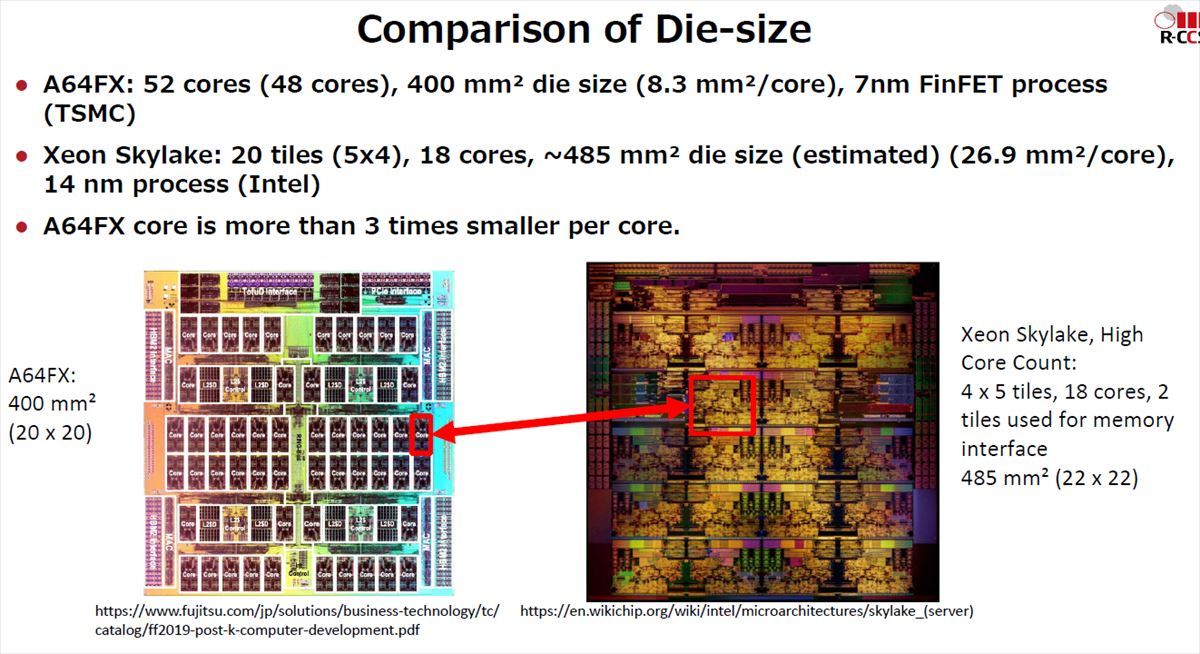


图3

（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。

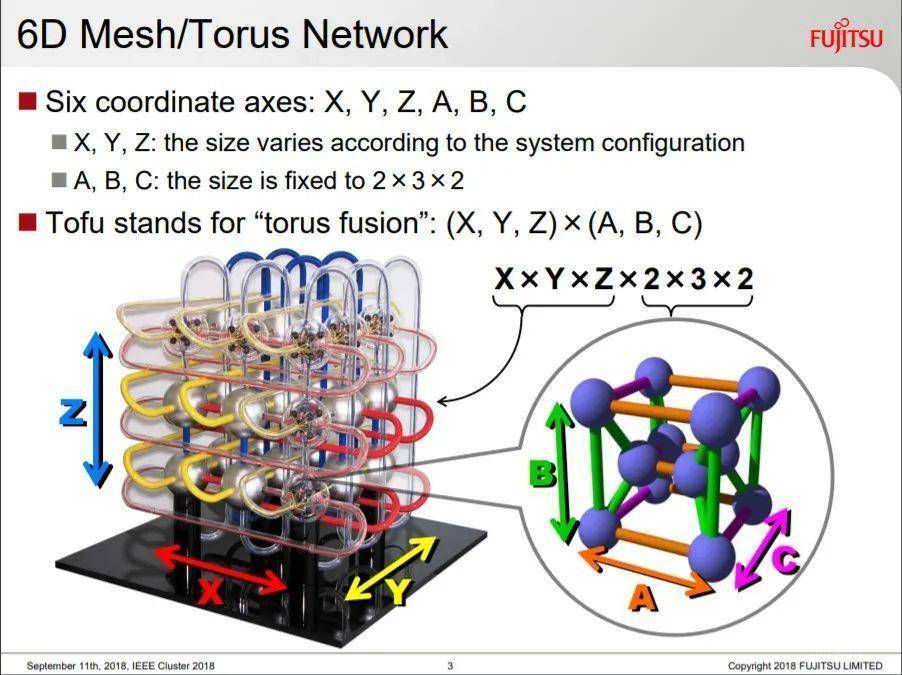


图4

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：

每个节点可同时通信的Link数为6，则6个Link的带宽一共为：

。

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。

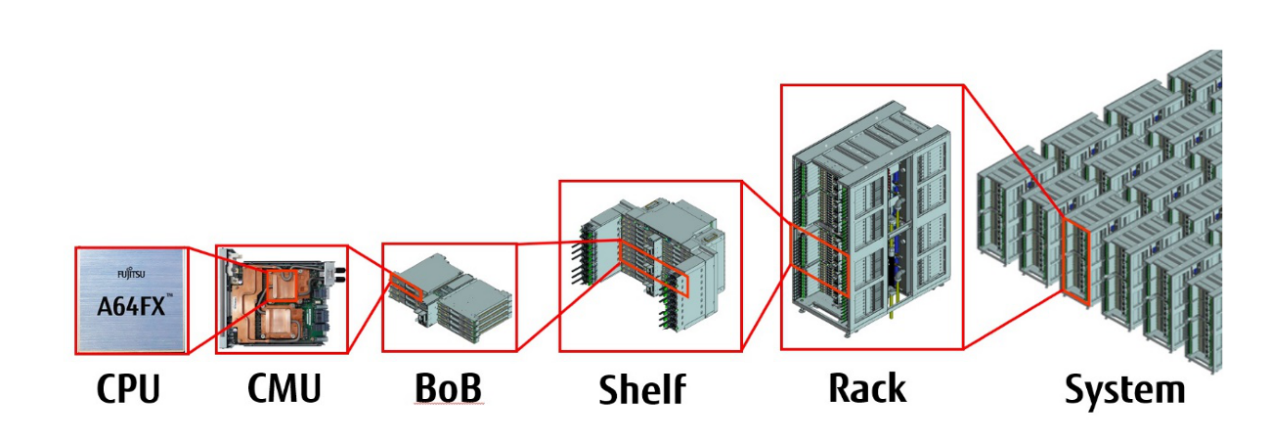


图5

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。

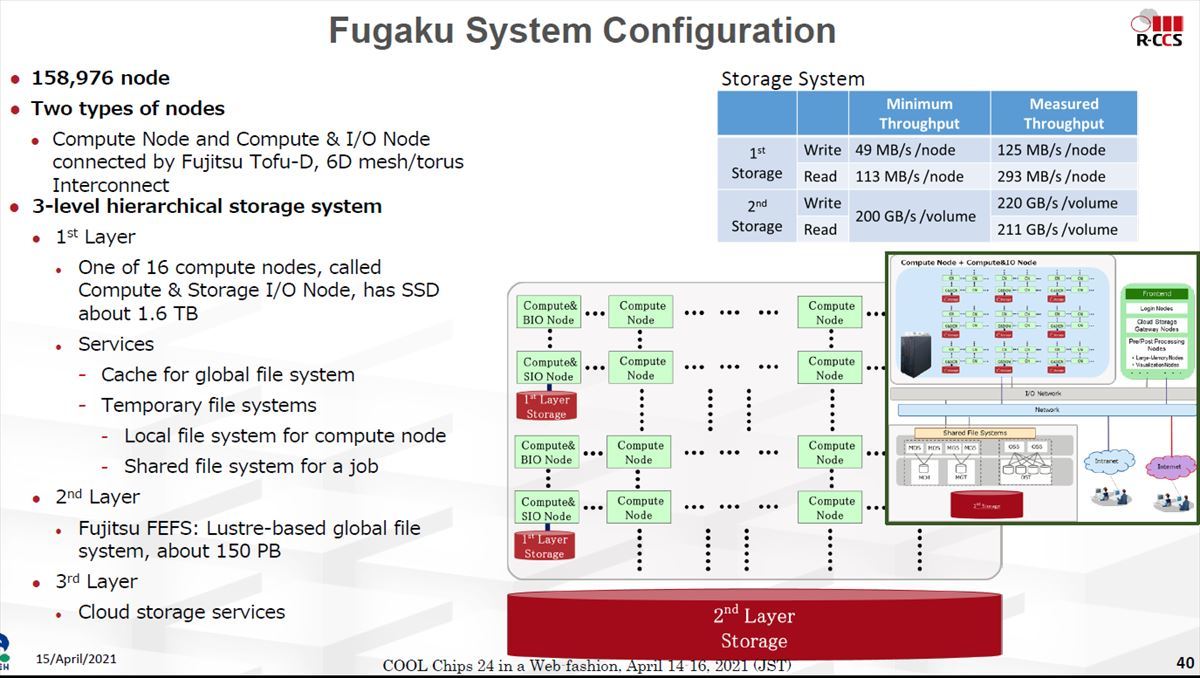


图6

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。