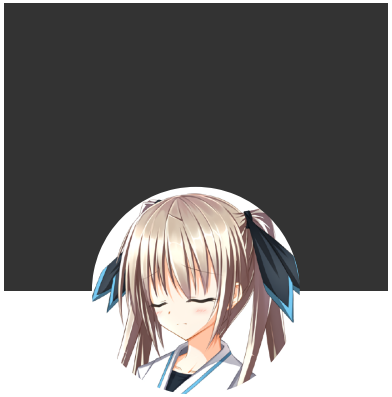


Rainmaker's Notebook

『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

Home

Archives

About

SiteXC

跨过 E 级计算的门槛之后

『如何评价 Frontier 成为首个达到 Exaflops 的超算并拿下 TOP500 第一名』

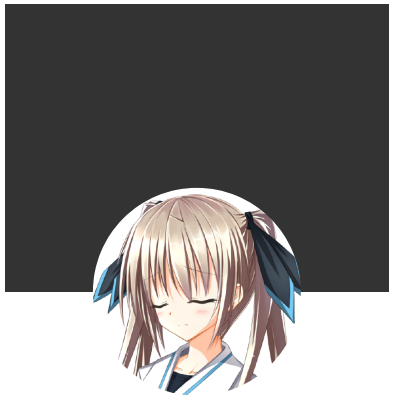
两天前，ISC 2022 更新了 TOP500 榜单，美国橡树岭国家实验室的 Frontier 首次突破 HPL 持续运行速度 1 Exaflops 的大关，成为名义上的第一台 E 级超算和新一任的 TOP500 冠军。仔细观察 TOP500 和 HPCG500 榜单，其实可以看出一些有意思的事情。

MI250X: 早熟的无情跑分显卡

这次 TOP500 前十名有三台来自 HPE Cray 的新机器（第一、三、十名），采用了相同的硬件配置：AMD 64-core EPYC 3rd gen, AMD Instinct MI250X, Slingshot-11 networking. 三台机器主要的区别是他们的大小。MI250X 有着相当恐怖的纸面性能：通用单 & 双精度、矩阵单 & 双精度、半精度理论峰值性能 47.9, 95.7, 383 TFLOPS, 128 GB HBM2e 显存，3276.8 GB/s 显存带宽。在这种恐怖性能的加持下，Frontier 得以跨过了 Exaflops 的门槛。

然而 Frontier 的登场有些出人意料，我觉得属于强扭的瓜，原因有二。首先是两个月前即有新闻指出，Slingshot 互联系统在而美国新超算上面可能遇到了问题（链接），甚至可能因此要延期大半年。说来美国人也是心大，三台旗舰级的新超算用的都是 Slingshot 互联系统，属实是鸡蛋一篮装。仅仅不到两个月，Frontier 就提交了 HPL 记录，不得不让人怀疑到底是 HPE 的工程师忽然醍醐灌顶，还是美国担心抢不到 E 级机一血而顶硬上。第二个原因是 Frontier 还没有交 HPCG 记录，但是采用相同硬件架构、排名第三的 LUMI 交了。凑巧，HPCG500 第二名是 Summit，它和 LUMI 的 HPL Rmax 性能几乎一样。然而 Summit HPCG 可以跑到 2.9 PFlops, LUMI 只能跑 1.9 PFlops，差了三成。再往下看，HPCG 排第四的 Perlmutter 也有 1.9 PFlops 的 HPCG 成绩，但是用的是 NVIDIA A100，只有 LUMI 一半的 HPL Rmax。我们正好可以对比一下 LUMI 和 Perlmutter。

LUMI (spec): 2560 节点，每节点 4 MI250X. 如果按 AMD 官方给出的纸面数据，LUMI 全系统的 GPU 应该有 490 PFlops, 32768 TB/s 显存带宽。不过按 LUMI 自己给出的数据，其 GPU 分区理论峰值速度只有 375 PFlops, 每个 MI250X 仅有 42.2 PFlops。我们假设 LUMI 只用了 2048 节点跑 HPCG，并且每个 MI250X 只按 42.2 TFlops 算，那么 2048 节点的总峰值速度和 GPU 显存带宽分别为 345 PFlops 和 26214 TB/s。



Rainmaker's Notebook
『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

Perlmutter (spec): 1536 节点，每节点 4 A100. 每块 A100 (spec): 9.7 TFlops 双精度理论峰值，2039 GB/s HBM2e 带宽，80 GB 显存。全系统的 GPU：59.6 PFlops，12234 TB/s 显存带宽。

这么一对比就很有意思了。LUMI 的理论峰值性能和理论峰值显存带宽是 Perlmutter 的 5 倍和 2 倍，但是只跑出了一样的 HPCG 性能。考虑到 HPCG 基本是个只看内存和显存带宽的程序（论文），这意味着 LUMI 上面的 HPCG 利用显存带宽的效率只有 Perlmutter 上面的一半。连 HPCG 这种高度抽象的基准测试都没有优化好就拿出来交成绩，如果这都不算赶鸭子上架，什么才算赶鸭子上架?

追不上硬件的算法

十年前，HPC 届刚刚进入 Petaflops 时代没多久。十年后的今天，Petaflops 似乎已经变得触手可及。我们来比较一下一套真实的 P 级系统和一套设想的新系统。

NERSC Cori (spec): 每节点 2 * Xeon E5 2697v3, 1.2 TFlops, 128 GB DDR4 2133，峰值内存带宽不会超过 136 GB/s。一个 1024 个节点的分区约有 1.23 PFlops, 128 TB DDR4, 136 TB/s 峰值内存带宽。

8 节点，每节点 4 个 MI250X：1.5 PFlops, 4 TB HBM2e, 102 TB/s 峰值内存带宽。

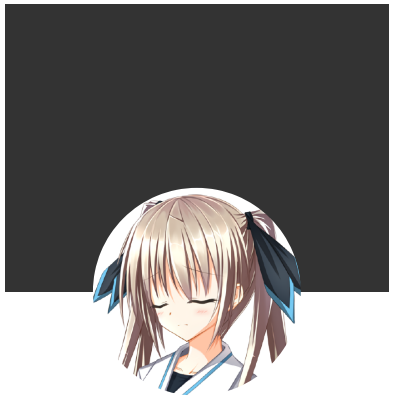
8 节点 * 4 MI250X 可以塞进一个标准机架里，而 1024 个 Cori Haswell 节点需要起码 7 个 cabinets。

为 P 级机设计的算法和程序，如果可以在 32 个 MI250X 上跑，大概率会比在 1024 个 Haswell 节点上跑的性能更好，毕竟通信开销大大降低了。然而，32 个 MI250X 只有 4TB 内存，而需要在 1024 个 Haswell 节点上跑的程序都是万核级的大程序，内存消耗肯定不小，未必能塞到 32 个 MI250X 上面。

另一方面，单卡 47.9 TFlops 的双精度性能，如果按 90% 效率计算，可以在 1 秒多一点的时间内算完一个两个 30000 * 30000 的双精度稠密矩阵相乘（如果用矩阵计算单元，只要半秒多一点）。而三个 30000 * 30000 的双精度稠密矩阵已经需要 20 GB 的存储空间了。实际应用中非常难见到如此大的稠密线性代数计算，其他计算类型的算术密集度更低，更难完全发挥出峰值计算性能的优势。以前各种计算-通信重叠来掩盖通信时间的算法，在新机器上面将变得聊胜于无，因为计算时间大幅度缩短了，而通信时间变化不大。

闷声发财的中国超算

上周我和老板聊的时候，老板提到了一个小道消息：Gordon Bell Prize 的组委会正在考虑，如果 ISC 上面没有新的中国超算上榜，那么应不应该把 GB 奖颁发给在中国新超算上跑的应用。老板因此怀疑中国的新超算不能长时间稳定运行，所以不交数据给 TOP500。



Rainmaker's Notebook
『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

Home

Archives

About

SiteXC

实际上，中国三台新超算之一的新神威，已经在 SC21 论文里透露了其规模和数据。根据论文，新神威每节点峰值速度约为 14 TFlops, 307 GB/s 峰值内存带宽，全系统有 108960 节点。照此推算，新神威全系统的理论峰值速度达到了 1.5 EFlops，理论峰值内存带宽为 32666 TB/s。按照同等峰值带宽换算，大约相当于 10000 块 MI250X 和 479 PFlops。论文里说新神威的 HPCG 可以跑到 5.91 PFlops，相当于 Perlmutter 的三倍，新神威的内存带宽是 Perlmutter 的不到三倍。

此外，今年四月有人在 GitHub 上面上传了新神威 HPL-AI 的运行记录和二进制文件（repo）。从运行记录上看，新神威最大的算例稳定运行了四千多秒，跑到了 5.5 EFlops 的半精度速度。如果按照 1:4 的半精度：双精度速度比例，那么新神威的 HPL 应该能稳定过 1 Exaflops.

最后，考虑到美国已经如愿以偿拿到了 E 级机一血，下面贴一下我去年11月基于公开资料写的一份中国 E 级机分析，之后看看猜中了几个。

根据钱德沛教授2017年所作的报告，中国起码有三台 E 级机正在设计和制造：曙光公司基于国产 x86 处理器和 DCU 加速卡的新超算，基于申威26010升级版众核处理器的神威新超算，以及国防科大基于飞腾处理器和 MT 加速卡的天河新超算。这个报告里提到了如下关键的设计指标：

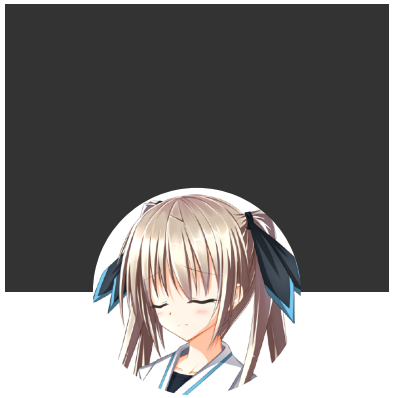
- HPL 运行峰值速度达到 EFLOPS 且并行效率超过 60%
- 10 PB 以上的内存
- 30 GFLOPS 以上的每瓦 HPL 性能
- 每节点 400 or 500 Gbps 的互联网络

下面的分析将会对应考察这几个指标。

曙光新超算

曙光新超算名字未知。曙光新超算将采用海光的 x86 Dhyana CPU 和海光 DCU (Deep Computing Unit). 海光 x86 是 AMD Zen 1 的授权；海光 DCU 并没有官方明确的消息，但是根据券商的中科曙光调研报告和海光 IPO 保荐书，海光 DCU 也是购买的 GPU 架构授权。而根据百度飞桨平台的安装说明，海光 DCU 使用的是 AMD ROCm 软件框架。因此可以肯定，海光 DCU 由 AMD GPU 授权而来。

钱德沛教授 2018 年在 SC18 的 PPT 指出，曙光原型机 512 节点（1024 海光 CPU + 512 海光 DCU）的理论双精度峰值速度是 3.18 PFLOPS。其中，海光 CPU 是 32 核型号。海光已上市的 CPU 中，C86 7185 为唯一的 32 核型号，其对应的 AMD CPU 为 AMD EPYC 7601，数据可参考这个 链接。根据 AnandTech 的 资料，海光 CPU 的向量浮点运算吞吐量遭到了严重的削弱，基本只有 EPYC Naples 架构的一半。2020 年来自上交的一篇 论文 对曙光原型机进行了基准测试，提供了宝贵的信息：双路 HPL 性能为 405.6 GFLOPS，STREAM Triad 内存带宽 253 GB/s。据此推测，单个 C86 7185 理论峰值性能



Rainmaker's Notebook
『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

Home

Archives

About

SiteXC

为 256 GLOPS，由 (256 bit / 64 bit) * 2.0 GHz * 32 core = 256 GFLOPS 给出；8 通道内存控制器，理论峰值内存带宽为 158.9 GB/s。

根据 3.18 PFLOPS 和 512 DCU 推测，单个海光 DCU 的理论双精度峰值速度约在 6 TFLOPS 的水平。从授权时间和 6 TFLOPS 理论双精度峰值推测，2018 年的海光 DCU 对应的 AMD GPU 可能为 GCN 5.1 架构的 MI50 (6.7 TFLOPS) 或者 MI60 (7.37 TFLOPS)。进一步，笔者查到飞桨在 2021 年与海光 DCU 系列完成的兼容性互认证中指出了海光 DCU-Z100 的存在。这个命名方式，笔者猜测 Z100 为 AMD MI100 的衍生版，后者具有 11.54 TFLOPS 的双精度理论性能，32 GB HBM2 显存，1229 GB/s 显存带宽，和 300W TDP（参考链接）。

MI100 的 flops/byte ratio 约为 9.38。若以 11.54 TFLOPS @ 300W 计算，每 W 性能约有 38 GFLOPS。

曙光原型机的互联系统每节点带宽是 200 Gbps，使用 6D Tours 结构（钱德沛 SC18 PPT）。

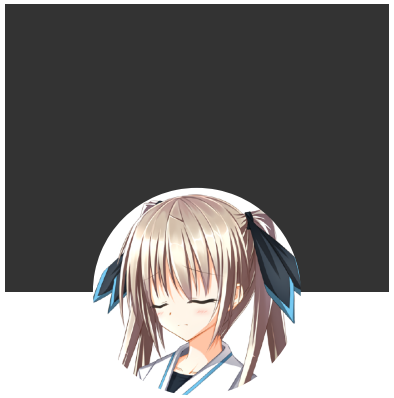
若以钱德沛教授 2017 年报告中提出的 DCU > 15 TFLOPS 性能，假设全系统 HPL 并行效率为 70%，则要达到峰值运行速度 1 EFLOPS，需要约 10 万个 DCU，对应 5 万个（每节点 2 DCU）或者 2.5 万个节点（每节点 4 DCU）。按 10 万个 DCU 估算，满载所需功耗为 100,000 * 300 W = 30 MW。加上其他系统，整机功耗可能超过 40 MW。按 10 万个 DCU 每个有 32 GB 显存估算，总显存为 100,000 * 32 GB = 3.2 PB，加上起码两倍的系统内存，总内存应该可以达到 10 PB。

神威新超算

神威新超算名字未知。神威新超算使用 SW26010 处理器的升级版 SW26010-Pro，SC21 的这一篇 论文 讨论了在新一代神威超算上的 HPCG 优化，并提供了相当多的信息。

根据这篇论文，SW26010-Pro 有六个核组，每个核组有 64 个 CPE 从核，256 KB local device memory (LDM)。与前一代 SW26010 相比，Pro 版从四个核组变成了六个核组；每个核组仍有 64 个 CPE 从核；每个 CPE 的 local device memory (LDM) 从前一代的 64 KB 升级到了 256 KB，且从之前的只能由程序显式控制变成了可以由硬件控制 32 KB 或者 128 KB 作为数据缓存；CPE 的向量部件从前代的 256 bit 提升到了 512 bit。每个 SW26010-Pro 的理论峰值性能约为 14 TFLOPS，推测这一值由 (512 bit / 8 bit) * 2 (FMA) * 64 cores * 6 CG * 2.3 GHz 给出。

内存系统方面，SW26010-Pro 有 96 GB DDR4 内存，每个核组有一个内存控制器，连接到 16 GB 内存。内存到 LDM 的理论带宽峰值为 307 GB/s。考虑到 DDR4-3200 单通道内存带宽为 25.6 GB/s，推测每个核组的内存控制器为双通道 DDR4-3200。此外，核组之间使用 RMA 通信的峰值内存带宽有 460 GB/s。



Rainmaker's Notebook
『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

按 14 TFLOPS 和 307 GB/s 进行计算，SW26010-Pro flops/byte ratio 约为 45。功耗信息未知，若按 SW26010 为 300W 推测，SW26010-Pro 可能为 400W。若按 14 TFLOPS @ 400W 估算，每 W 性能约有 35 GFLOPS。

神威原型机使用的互联系统是 2-stage fat-tree, 每节点带宽 200 Gbps (钱德沛 SC18 talk)。

全系统层面，SC21 论文指出神威新超算有 108,960 个节点，峰值性能是神威太湖之光的12倍。假设全系统 HPL 并行效率为 70%，按每节点 14 TFLOPS 计算，108,960 个节点刚好可以使得峰值运行速度超过 1EFLOPS。同时，10 万个 SW26010-Pro 满载所需功耗推测为 100,000 * 400W = 40 MW。加上其他系统，整机功耗可能超过 50 MW。按 10 万个 SW26010-Pro 每个有 96 GB 内存估算，总内存约为 108,960 * 96 GB = 10.4 PB。

天河新超算

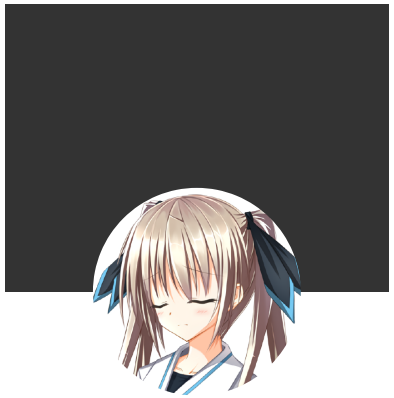
如无意外，天河新超算的名字是『天河三号』。天河三号将采用飞腾公司的 FT2000 系列 CPU 和国防科大的 Matrix-2000 系列加速卡，在天河三号原型机上试用的是 FT2000+ 和 MT2000+，不排除在天河三号上这两个芯片会再次升级。

根据论文 链接 和 链接，FT2000+ 是一个 64 核 ARMv8 处理器，每个核心有 32KB L1D cache，每四个核心共享 2MB L2 cache，工作频率 2.4 GHz，理论双精度峰值速度是 614.4 GFLOPS. 如果按 ARMv8 ASIMD 的 128 bit 位宽，则理论峰值速度应该由 (128 / 2) * 2 (FMA or 2 VFU) * 2.4 GHz * 64 cores = 614.4 GFLOPS 算得。FT2000+ 有 8 个 DDR4 内存控制器和 64 GB DDR4 内存，如果使用 DDR4 3200 内存则峰值内存带宽为 204.8 GB/s, flops/byte ratio 约为 3。天河三号原型机上每个 FT2000+ 最大功耗为 100 W。

MT2000+ 是一个 128 核 ARMv8 处理器，工作频率 2.0GHz，理论双精度峰值速度是 4096 GFLOPS。天河二号-A 上面的 MT2000 每个处理器有 2 个 256 位的向量部件且每个时钟周期可以进行 16 FLOPS（来自报告），则应该是每时钟周期进行 2 * (256 / 64) * 2 (FMA) FLOPS. 若 MT2000+ 仍然采用 2 个 256 位向量部件，则理论峰值速度 2 * (256 / 2) * 2 FMA * 2 GHz * 128 cores = 4096 GFLOPS. 每个 MT2000+ 只有 8 个 DDR4 内存控制器和 16 GB 内存，推测其峰值内存速度仍然不会超过 204.8 GB/s。

按 4096 GFLOPS 和 204.8 GB/s 推算，MT2000+ flops/byte ratio 约为 20。MT2000+ 最大功耗为 240 W，每 W 性能只有约 17 GFLOPS，不到 30 GF/W 设计指标的 60%。

天河三号原型机的互联系统每节点带宽是 200 Gbps，正式系统应该会到 400 Gbps，网络架构为 3D butterfly network (钱德沛 SC18 PPT & talk)



Rainmaker's Notebook
『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

Home
Archives
About
SiteXC

天河三号原型机每节点有 2 个 MT2000+ 和 2 个 FT2000+。若天河三号仍为每节点 2 CPU + 2 加速卡，使用的芯片性能是原型系统的 1.5 倍（比较乐观估计），全系统 HPL 并行效率为 70%，则要达到峰值运行速度 1 EFLOPS，需要约 10 万个节点。如果 MT2000+ 可以独立使用，按 20 万个 MT-2000+ 估算，满载所需功耗为 200,000 * 240 W = 48 MW。加上其他系统，整机功耗可能超过 60 MW。若按 10 万个节点估算，则总内存为 200,000 * 80 GB = 16 PB。

2022年10月更新

经朋友提醒，天河新超算使用的应该是 MT-3000, 今年五月的时候有一篇论文已经挂了出来（还有一篇高度相关的论文在 arxiv). 根据论文透露，MT-3000 有 16 个通用核心，四个 Acceleration Zone (AZ), 每个 AZ 有 24 个控制核心和 384 个计算核心（合计 96 个控制核心和 1536 个计算核心），AZ 运行在 1.2 GHz 频率。MT-3000 的 AZ 用的是 VLIW 架构，每个计算核心有 3 个乘加单元（MAC），一个整数执行单元，两个存储单元。按此计算，1536 cores * 1.2 Ghz * (3 * 2) MAC Flops = 11 TFlops. 很不幸的是，MT-3000 的内存系统和 MT-2000+ 的相比反而倒退了：只有 4 个 DDR4 module, 每个容量 32GB，带宽 25.6 GB/s, 合计带宽 102 GB/s. 这使得 MT-3000 的 flops/byte ratio 高达 110, 对程序是一个极大的挑战。功耗方面，论文指出 MT-3000 的能耗比为 45.4 GF/W, 即芯片功耗约为 250W. 论文按一台超算 80% 的功耗为处理器功耗进行估计，则可以得到 10 万节点规模超算的能耗约为 30MW. 这一比例和我上文的估算基本接近。

发布于 2022-06-02 / tags: { Supercomputing }

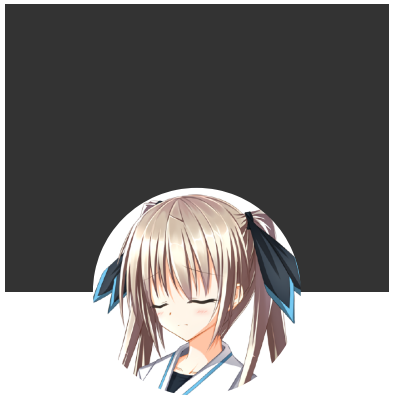
2 comments Anonymous

Leave a comment

Markdown is supported Login with GitHub Preview

ruixueqingyang commented about 2 months ago

MT-3000的DDR带宽数据似乎和论文里的有些出入。“MT-3000: a heterogeneous multi-zone processor for HPC”



Rainmaker's Notebook

『求雨巫师的神奇之处在于他总是躲着不见你，却总说刚下完的雨是拜他所赐。』——《天真的人类学家》

Home

Archives

About

SiteXC

LfieLike commented 26 days ago

看论文mt3000能跟A100拼一拼