

第X讲 国产超级计算机

目录

- 国产超算发展
- ARM架构及优势
- 华为鲲鹏HPC介绍

目录

- 国产超算发展
- ARM架构及优势
- 华为鲲鹏HPC介绍



国产超算发展

■ 世界超算发展史

- 20 世纪 70 年代，向量计算机
- 20 世纪 80 年代，对称多处理机
- 20 世纪 90 年代初期，大规模并行处理
- 20 世纪 90 年代中后期，将 SMP 的优点与 MPP 系统相结合形成了分布式共享内存结构
- 与 cc-NUMA 系统同时发展起来的还有集群系统
- 21 世纪，高性能计算机大多都采用了集群系统，而在计算节点内处理器则从单一的 CPU 变为了 CPU 与加速部件的混合结构



University of Electronic Science and Technology of China



国产超算发展

■ 国产超算发展史

- 1983年，“银河-I”巨型计算机
- 1992年11月，“银河-II”十亿次通用并行巨型机
- 1997年，“银河-III”等系列巨型机
- 2009年10月，“天河一号”千万亿次超级计算机
- 2013年6月，“天河二号” TOP500 榜首(2013-2015)
- 2016年6月，“神威·太湖之光”TOP500 榜首(2016-2017)
- 2017年，“天河-2A”技术升级和系统优化
- 至今，E级超算时代, 计划部署“天河三号”，“神威E级”“曙光E级”三台E级超算



国产超算发展

■ Top500, 2022.11

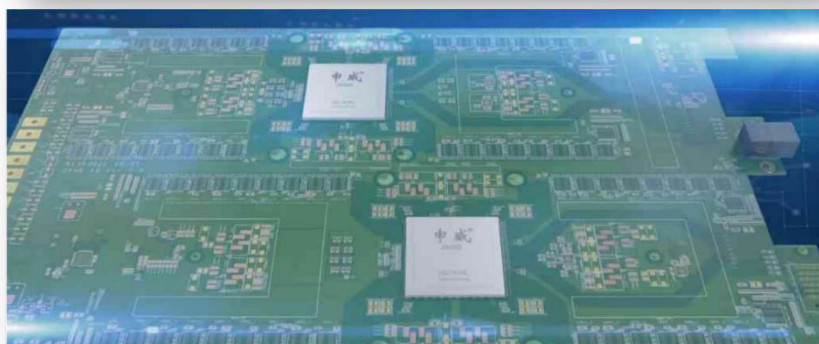
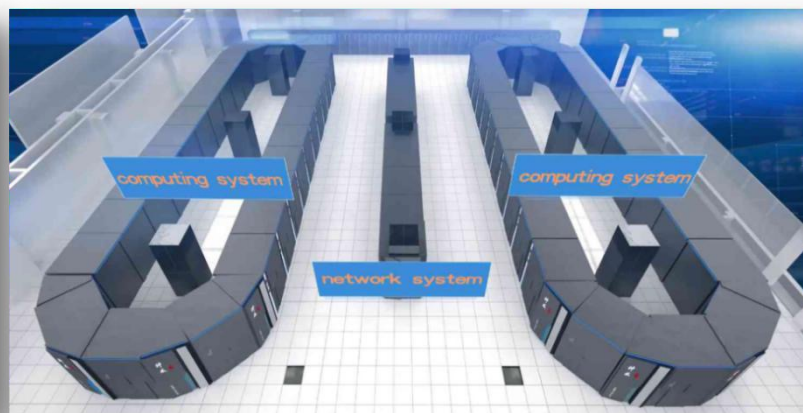


Rank	System	Cores	(PFlop/s)	(PFlop/s)	(kW)						
1	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100	6	Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
2	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899	7	Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
3	LUMI - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016	8	Perlmutter - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589
4	Leonardo - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610	9	Selene - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646
5	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096	10	Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61.44	100.68	18,482



国产超算发展

■ 神威·太湖之光

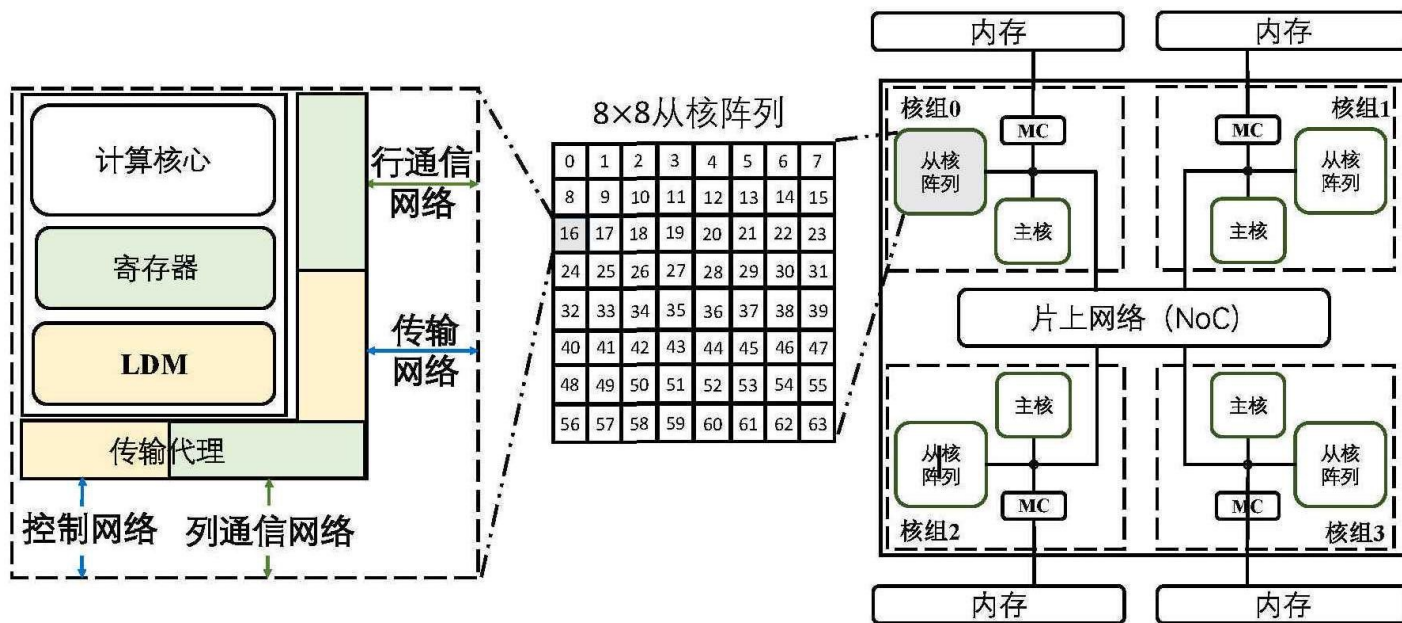


国产超算发展

■ 神威·太湖之光

➤ “申威 26010”处理器

- 完全我国自主技术研制的第四代高性能处理器
- 采用片上计算阵列集群和分布式共享存储相结合的异构众核体系结构

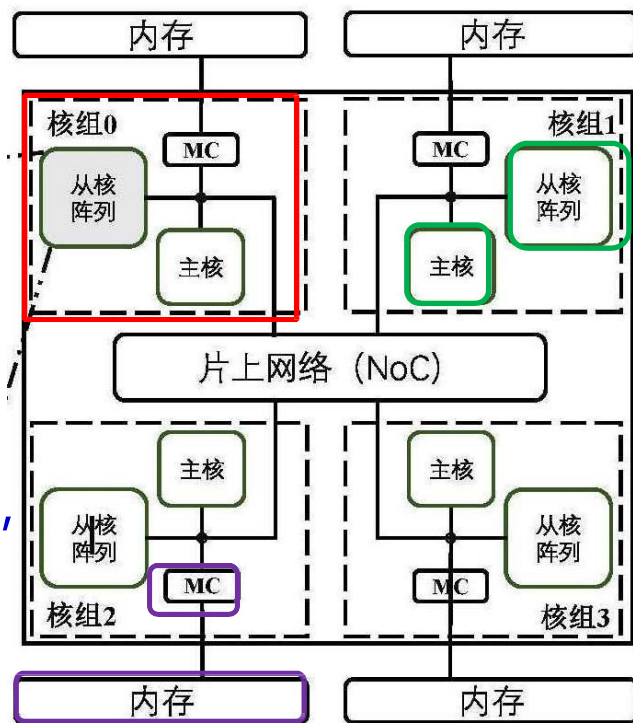


国产超算发展

■ 神威·太湖之光

➤ “申威26010”芯片：

- 四个运算核组，共260个运算核心。
- 每个核组采用主从异构的结构，包括一个主核和一个从核阵列。
- 每个核组还集成了8GB的DDR3内存，并由内存总线通过存储控制器与主核和从核阵列相连。
- 四个核组通过高速片上网络互连。
- 共计32GB的内存物理空间被统一编址，任意主核和从核均可以访问芯片上的所有主存空间。
- 片上配备有可提供16GB/s双向峰值带宽的8通道PCI-E 3.0 标准接口，用于和其他芯片互联。



国产超算发展

■ 神威·太湖之光

➤ “申威26010” 主核：

- 工作频率为1.45GHz。
- 采用64位的基于RISC的第四代申威指令集。
- 每个主核都包含两级Cache，一级Cache包含指令Cache和数据Cache两种，大小均为32KB。二级Cache被指令和数据共用，称之为SCache，大小为512KB。
- 为了处理共享存储器空间的Cache一致性访问，在每个核组还设置了一个二级Cache的标记副本，被称之为CTAG，与主核的Scache一一对应。
- 主核支持中断，同时可以运行操作系统和用户程序，进行计算、存储资源的管理，并提供消息、文件、调试、低功耗管理等服务。

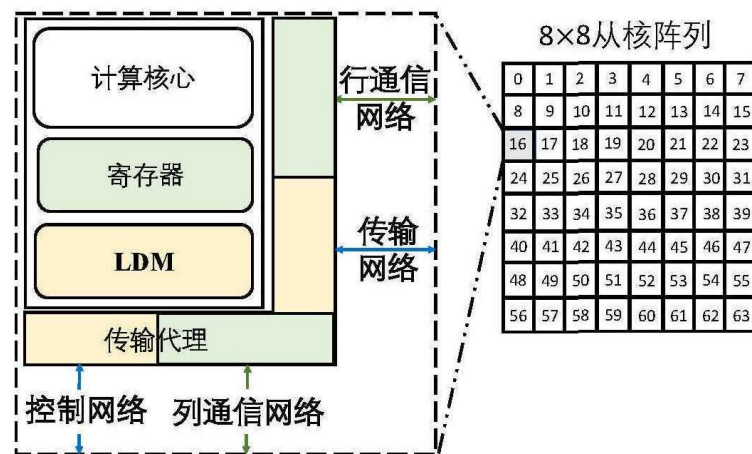


国产超算发展

■ 神威·太湖之光

➤ “申威26010”从核阵列：

- 64个相同的从核排布成8行、8列拓扑结构。
- 工作频率是1.45GHz。
- 其指令集兼容大部分主核指令，并添加了一些从核特有的寄存器通信相关指令。



- 独立的大小为16KB的一级指令Cache。大小为64KB的二级指令Cache被64个从核共享。每个从核拥有独立的64KB可重构局部数据存储作为数据Cache。
- 32个256-bit的通用寄存器，可供计算单元进行运算使用。
- 阵列数据传输网络，包括4个行向总线作为数据通路和指令装填通路，相邻两列16个从核共享一套这样的总线。



国产超算发展

■ 神威·太湖之光

➤ “申威26010”存储控制器：

- 每个核组的从核都需要经由同一个存储控制器访问内存，因此它们共享内存带宽资源。
- 整个芯片集成4块8GB DDR3S DRAM形式内存。
- 每块内存的数据接口位宽144位，理论最大主存数据带宽为134.4GB/s。

➤ “申威26010”众核架构的主核从核特点：

	主频	向量长度	指令 Cache	数据 Cache	内存带宽
主核	1.45 GHz	256 bit	32 KB L1	32KB L1	33.6 GB/s
			256KB L2（数据指令共享）		
从核			16 KB L1 64KB L2 (64 从核共享)	64KB (SPM)	



国产超算发展

■ 神威·太湖之光

- “申威26010”与同时代超算采用的芯片架构对比：
 - **并行架构**：Intel KNL和NVIDIA Kepler属于SIMD架构，神威的核心计算部件——从核阵列属于MIMD并行架构，并行方式更加复杂。
 - **内存体系**：申威处理器256个从核都有独立的64KB LDM作为数据缓存。Kepler K40 的同一SM中的SP共享64KB L1 Cache，所有SM共享1.5MB 的L2 Cache。Intel KNL 的计算核心共享一个L2Cache。
 - **地址空间**：申威架构的主存地址空间被主核和从核共同编址并可以共同访问。Kepler 中GPU显存和CPU内存地址分离。KNL 有多种编址方式。
 - **访存计算比**：相比Kepler和KNL架构，申威提供了类似的浮点运算性能，但是在内存带宽方面则相形见绌。



国产超算发展

■ 天河二号

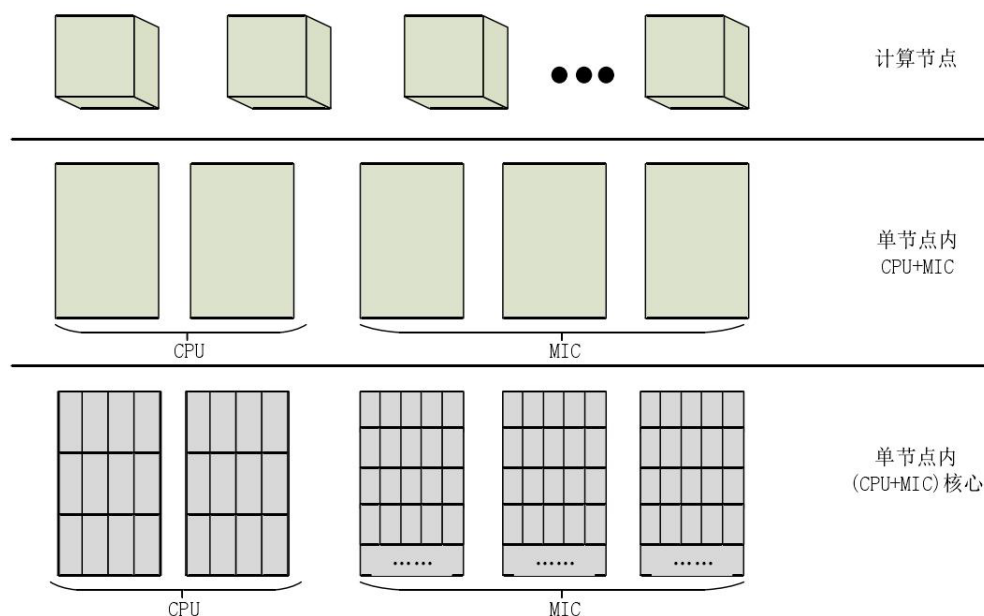
- CPU+MIC 混合异构架构。
- 共有 16,000 个运算节点，每节点配备两颗 Xeon E5 12 核心的 CPU、三个 Xeon Phi 57 核心的 MIC。累计 32,000 颗 Xeon E5 主处理器和 48,000 个 Xeon Phi 协处理器，共 312 万个计算核心。
- CPU 运作时钟频率为 2.2GHz 的 Xeon E5-2692v2 12 核心处理器，峰值性能 0.2112TFLOPS。
- MIC 运行时钟频率为 1.1GHz，拥有 57 个 x86 计算核心（实际上拥有 61 个核心），每个 x86 核心通过特殊的超线程技术处理，能运作 2 个线程，产生峰值性能为 1.003TFLOPS。



国产超算发展

■ 天河二号

- Cluster 系统结构
- 将多个可以独立运行的运算节点相互连接
- 节点间可以通过高速互连网络相互通信，节点内由 2 块 CPU 和 3 块 MIC 卡组合而成

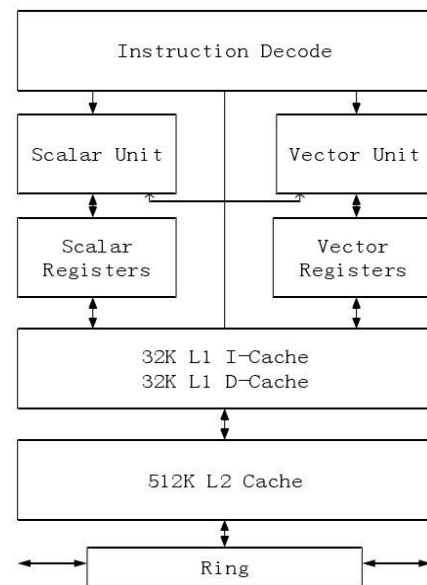
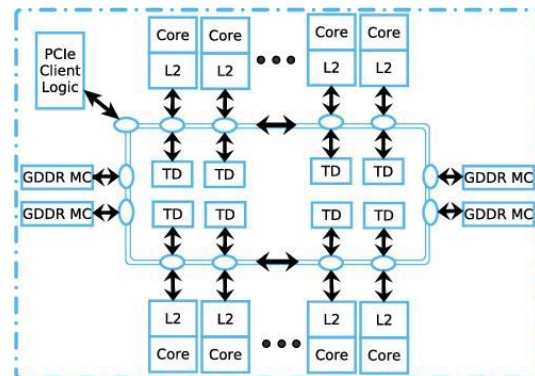


国产超算发展

■ 天河二号

➤ MIC体系架构：

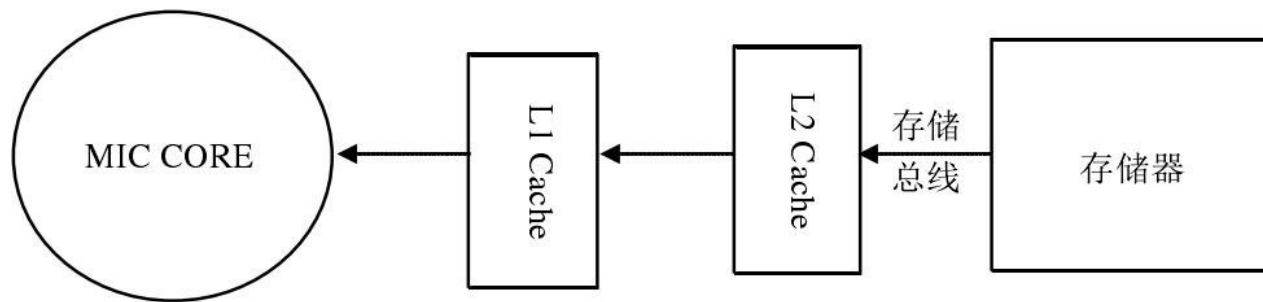
- 环形结构，环形总线连接 8 个片上高速互联的 MIC 架构计算核心、8 个分布式标签目录(Tag Directory, TD)和 8 个基于 GDDR5 规范的存储控制器(Memory Controller, MC)，支持 16 个传输通道。
- 每个MIC 核心都包含一个标量处理单元(Scalar Processing Unit, SPU)和一个向量处理单元 (Vector Processing Unit, VPU)。VPU 是一个建立在 512 位寄存器组上的 SIMD 引擎，一个时钟周期可执行 16 个单精度浮点运算或 8 个双精度浮点运算。
- MIC 核心上拥有两级 Cache，其中 L1 Cache 由 32KB 指令 Cache 和 32KB 数据 Cache 组成，用以满足 4 个硬件线程连续的高频度访存要求；L2 Cache 为全局缓存，大小为 512KB。



国产超算发展

■ 天河二号

- MIC 核心进行计算时对数据的访存结构：
 - MIC 在计算过程中对要使用的数据会先从主存中读出一段数据存放在 L2 Cache 中；
 - 再从 L2 Cache 中读出一段数据放在 L1 Cache，核心计算时直接使用 L1 Cache 数据。
 - 只有在 L1 和 L2 Cache 中找不到计算使用的数据时才会从主存中读入数据。



MIC 访存结构



国产超算发展

■ 天河二号

➤ CPU+MIC 编程模型：

- native 模式：包括 CPU 原生模式和 MIC 原生模式两种。
 - CPU 原生模式为纯 CPU 环境。
 - MIC原生模式是将 MIC 作为一个独立计算节点，将现有应用平顺移植到 MIC 上，即将程序和所需数据传输到 MIC 上，并直接在 MIC 卡上运行程序。
- offload 模式：通常采用了CPU 为主 MIC 为辅的模式，以 CPU 端为主，将部分加载到 MIC 进行计算，适用于程序中含有高并行度计算部分。
- symmetric 模式：CPU 与 MIC 对等模式。MIC 与 CPU 都当做相同的计算节点来看待，主控端可以是两者中的任意一个。



国产超算发展

■ 天河2A

➤ 天河2号升级为天河2A：

- 2015年美国政府颁布禁令，严禁Intel等公司向中国四家超算中心出口高性能计算芯片，因此天河2号不能使用Intel Xeon Phi加速卡
- 2017年广州超算中心使用国产加速器Matrix 2000取代原本的英特尔Xeon Phi加速器

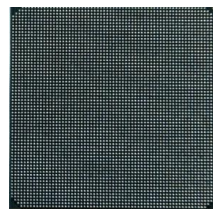
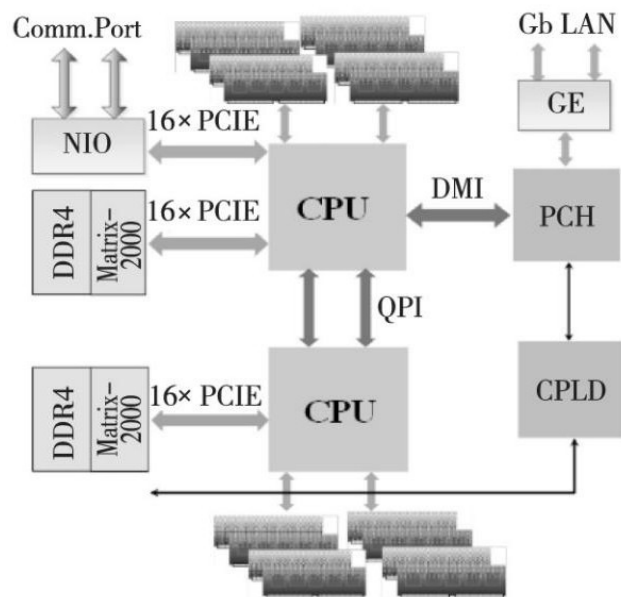
Components	TianHe-2	TianHe-2A
Nodes and performance	16,000 nodes with Intel CPUs + KNC	17,792 nodes with Intel CPUs + Matrix-2000
	54.9 Pflop/s	94.97 Pflop/s
Interconnection	10 Gbps, 1.57 us	14 Gbps, 1 us
Memory	1.4 PB	3.4 PB
Storage	12.4 PB, 512 GB/s	19 PB, 1 TB/s (upgrading, maybe larger)
Energy efficiency	17.8 MW, 1.9 Gflop/s per Watt	16.9MW, >5 Gflop/s per Watt (predicted)
Heterogeneous software	MPSS for Intel KNC	OpenMP/OpenCL for Matrix-2000



国产超算发展

■ 天河2A

- 每个节点由2颗Intel Xeon 微处理器器和2颗Matrix 2000加速器组成。
- 每个Intel Xeon微处理器 包含12核，工作频率为2.2GHz，采用英特尔 IvyBridge 微架构。
- 每个 Matrix 2000加速器包含128核，由4个超结点组成，每个超结点包含32个计算核，其中超结点支持64核超线程技术，有8个DDR4内存通，支持×16 PCIE 3.0 工作模式。



国产超算发展

■ E级超算

➤ 每秒可进行百亿亿次浮点运算

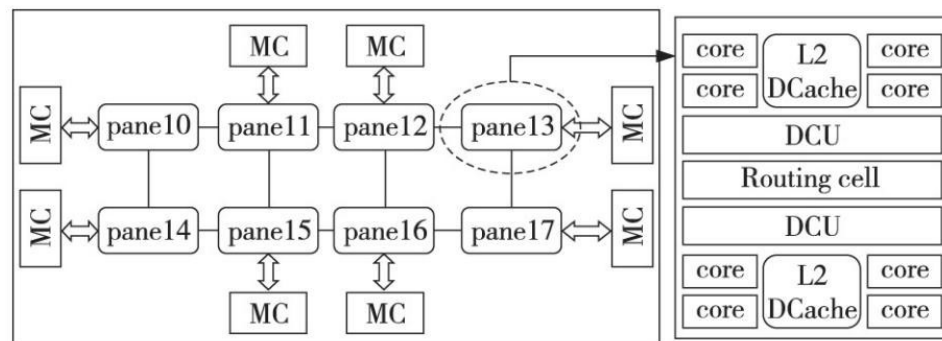
- 美国：“Aurora” 使用 Xeon Scalable CPUs 和 Intel Xe GPGPUs；“Frontier” 基于EPYC霄龙处理器（Zen 3或Zen 4）、Radeon Instinct加速卡；“El Capitan” 超算采用AMD CPUs和GPUs。
- 欧洲：“Vega” 使用 AMD Epyc 7H12 CPUs 和 Nvidia A100 GPUs。
- 日本：“Fugaku” ARM架构。
- 中国：“天河三号” ARM架构；“曙光E级”CPU + GPU；“神威E级”众核体系结构。



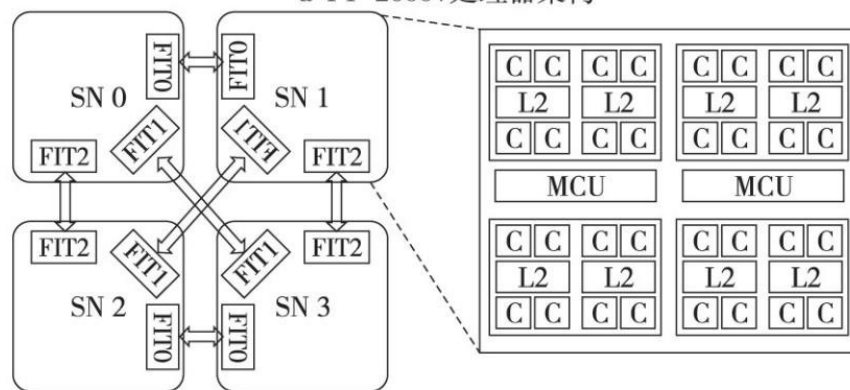
国产超算发展

■ 天河三号原型机

- 处理器包括 **FT-2000+(FTP)** 和 **MT-2000+(MTP)**
- **FTP** : 64个 **ARMv8** 架构的 **FTC662** 处理器核, 工作主频 2.2 ~ 2.4 GHz, 片上集成了 32MB 的二级 Cache, 可提供 204.8GB/s 访存带宽, 典型工作能耗约为 100W; FT-2000+ 兼容 ARMv8 指令集。
- **MTP** : 128个定制的处理核心, 被组织为 4个超级节点, 主频最高可达 2.0GHz, 整个处理器的消耗为 240W



a FT-2000+处理器架构



b MT-2000+处理器架构



国产超算发展

■ 天河三号原型机

- FTP和MTP都被划分成以32个核作为一个计算节点，目的是为了提供更多的计算节点以满足复杂的计算任务。计算节点由批处理调度系统管理和分配。
- 在FTP中，32个核共享64GB内存，而在MTP中，32个核共享16GB的内存。
- 由国防科技大学设计实现的原型集群互连技术提供 200Gbps 双向互连带宽。
- 使用Lustre进行分布式文件系统管理。提供多个版本的MPI编译实现。

表 1 天河三号原型机基本情况

Specifications		FT-2000+	MT-2000+
Hardware	Nodes	128	512
	Cores in a node	32	32
	Frequency	2.4 GHz	2.0GHz
	Memory	64 GB	16 GB
	Interconnect bandwidth	200 Gbps	
Software	OS	Kylin 4.0-1a OS with kernel v4.4.0	
	File system	Lustre	
	MPI	MPICH v3.2.1	
	Compiler	GCC v4.9.1/v4.9.3	
	Supported libraries	Boost, BLAS, OpenBLAS, Scalapack, etc.	



国产超算发展

■ 曙光E级原型机

- 采用 CPU+加速器的异构架构
 - CPU采用的是AMD授权的海光X86处理器
 - 加速器采用的是海光深度计算器 DCU (Deep Computing Unit) 加速卡

加速器测试平台对比

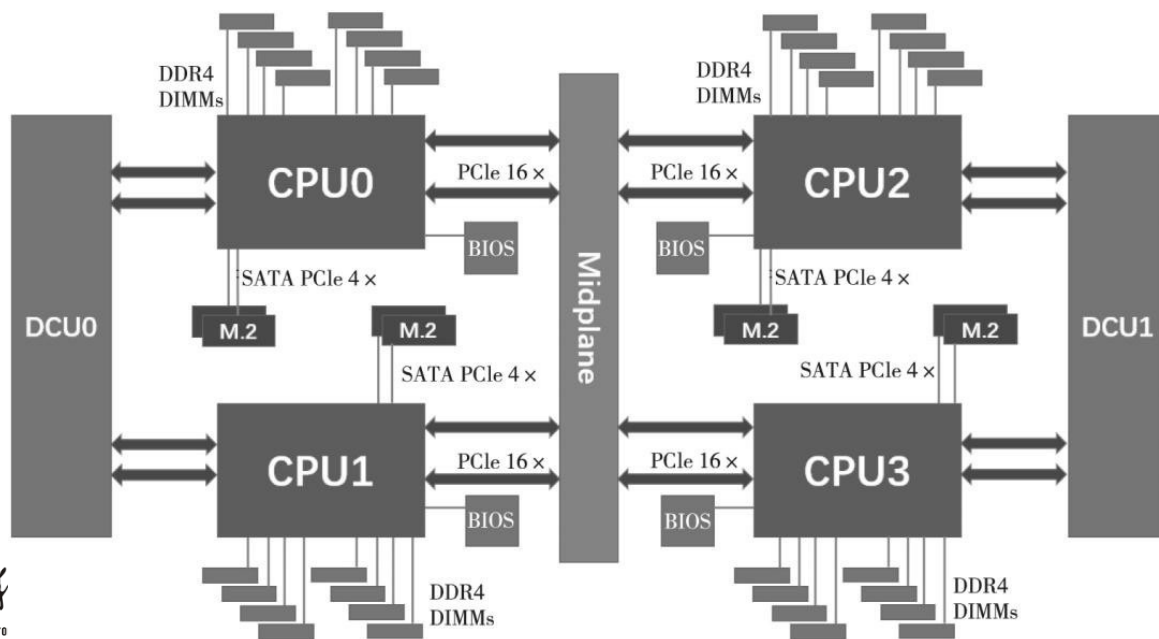
	微处理器数	时钟频率 /(MHz)	双精度性能 /(TFLOPS)	显存 /(GB)	编程模型	编译器
Hygon DCU	64	1 670	5.7	32	HIP1.8.5	hipcc
NVIDIA V100	80	1 530	7.8	32	CUDA9.2	nvcc



国产超算发展

■ 曙光E级原型机

- 共有512个节点，1024颗Hygon处理器和512块DCU加速卡。
- 各节点之间使用200Gbps的高速网络，采用6D-Torus的方式实现高维互连。
- 每个结点有2颗Hygon 7185处理器和1块DCU加速卡，256GB的DDR4内存，240GB的M.2 SSD硬盘。



目录

- 国产超算发展
- ARM架构及优势
- 华为鲲鹏HPC介绍



ARM架构及优势

■ 复杂指令集vs精简指令集

- 复杂指令集（ Complex Instruction Set Computing , CISC ）
 - 即冯·诺依曼结构（ 普林斯顿结构 ） , 指令与数据存储在同一存储器中；
 - 采用CISC结构的处理器，指令线与数据线分时复用；
 - 程序指令存储地址与数据存储地址指向同一个存储器的不同物理位置，程序指令和数据的宽度相同；
 - 取指令与取数据不能同时进行，速度受限。
- 精简指令集（ Reduced Instruction Set Computing , RISC ）
 - 即哈佛结构，指令与数据存储于两个不同的存储空间；
 - 程序存储器与数据存储器相互独立，独立编址，独立访问；
 - 分离的程序总线与数据总线在一个机器周期中，可同时获得指令字和操作数，提高执行效率；



ARM架构及优势

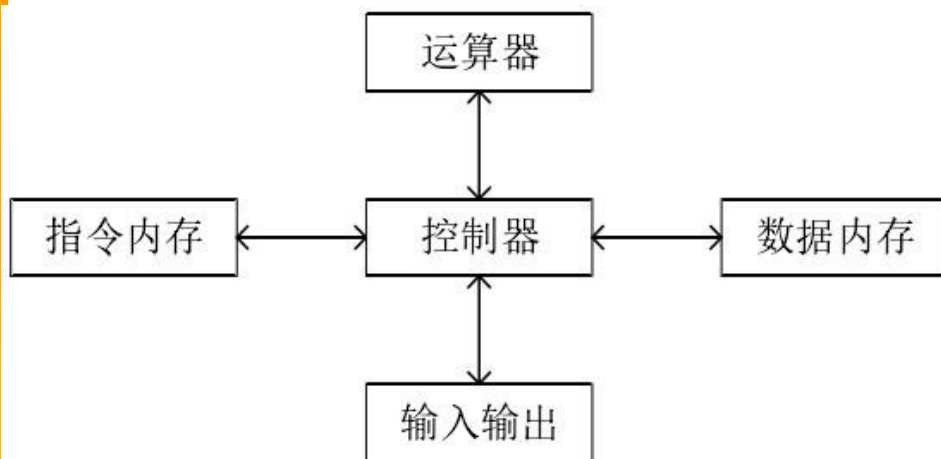
■ 复杂指令集vs精简指令集

项目	CISC	RISC
定义	指令集较多	指令集较少
内存单元	复杂指令执行内存单元	使用单独硬件实现指令
程序单元	固定程序单元	微程序
计算方式	计算缓慢	计算更快、更准确
指令译码	复杂	简单
指令周期	长短不一	固定
指令长度	不固定	固定
外部存储器	需要额外存储器	不需要额外存储
执行时间	长	短

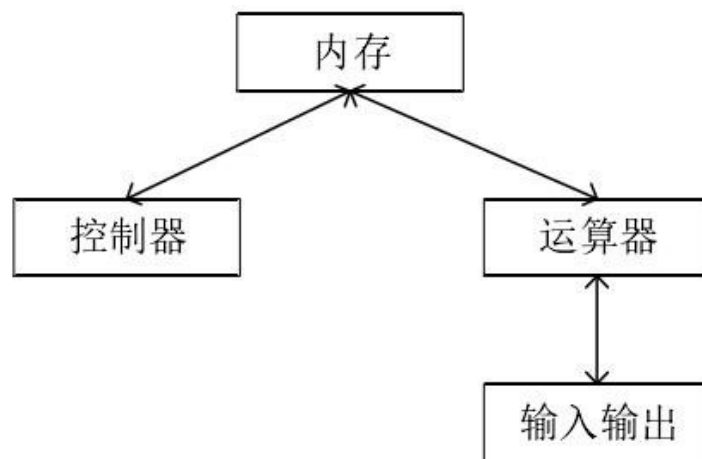


ARM架构及优势

■ 复杂指令集vs精简指令集



(a) 哈佛体系结构

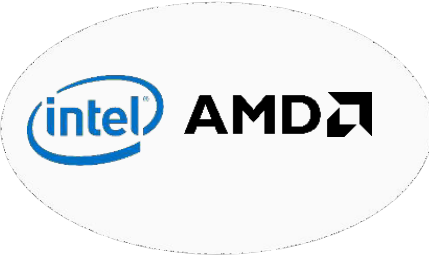

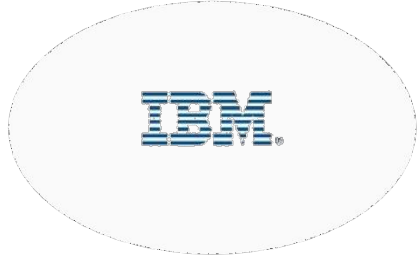


(b) 冯诺依曼体系结构



ARM架构及优势

■ 复杂指令集vs精简指令集

 X86	 ARM	 POWER
<ul style="list-style-type: none">• CISC, 复杂指令集	<ul style="list-style-type: none">• RISC, 精简指令集	<ul style="list-style-type: none">• RISC, 精简指令集
<ul style="list-style-type: none">• 重核架构, 高性能高功耗	<ul style="list-style-type: none">• 多核架构, 均衡的性能功耗比	<ul style="list-style-type: none">• 重核架构, 高性能内核
<ul style="list-style-type: none">• 14 nm, 摩尔定律放缓	<ul style="list-style-type: none">• 7 nm, 业界领先的制程工艺	<ul style="list-style-type: none">• 14 nm
<ul style="list-style-type: none">• 通用性强	<ul style="list-style-type: none">• 快速发展与完备	<ul style="list-style-type: none">• 聚焦大小型机和HPC
<ul style="list-style-type: none">• 封闭架构, 英特尔及AMD	<ul style="list-style-type: none">• 开放平台, IP授权的商业模式	<ul style="list-style-type: none">• 开放平台, IBM主导



ARM架构及优势

■ ARM 架构发展

Architecture	Core bit-width	Cores	
		Arm Ltd.	Third-party
ARMv1	32	ARM1	
ARMv2	32	ARM2, ARM250, ARM3	Amber, STORM Open Soft Core ^[60]
ARMv3	32	ARM6, ARM7	
ARMv4	32	ARM8	StrongARM, FA526, ZAP Open Source Processor Core
ARMv4T	32	ARM7TDMI, ARM9TDMI, SecurCore SC100	
ARMv5TE	32	ARM7EJ, ARM9E, ARM10E	XScale, FA626TE, Feroceon, PJ1/Mohawk
ARMv6	32	ARM11	
ARMv6-M	32	ARM Cortex-M0, ARM Cortex-M0+, ARM Cortex-M1, SecurCore SC000	
ARMv7-M	32	ARM Cortex-M3, SecurCore SC300	Apple M7
ARMv7E-M	32	ARM Cortex-M4, ARM Cortex-M7	
ARMv8-M	32	ARM Cortex-M23, ^[62] ARM Cortex-M33 ^[63]	
ARMv7-R	32	ARM Cortex-R4, ARM Cortex-R5, ARM Cortex-R7, ARM Cortex-R8	
ARMv8-R	32	ARM Cortex-R52	
ARMv8-R	64	ARM Cortex-R82 [®]	
ARMv7-A	32	ARM Cortex-A5, ARM Cortex-A7, ARM Cortex-A8, ARM Cortex-A9, ARM Cortex-A12, ARM Cortex-A15, ARM Cortex-A17	Qualcomm Scorpion/Krait, PJ4/Sheeva, Apple Swift (A6, A6X)
ARMv8-A	32	ARM Cortex-A32 ^[68]	
ARMv8-A	64/32	ARM Cortex-A35, ^[69] ARM Cortex-A53, ARM Cortex-A57, ^[70] ARM Cortex-A72 ^[71]	X-Genie, Nvidia Denver 1/2, Cavium ThunderX, AMD K12, Apple Cyclone (A7)/Typhoon (A8, A8X)/Twister (A9, A9X)/Hurricane+Zephyr (A10, A10X), Qualcomm Kryo, Samsung M1/M2 ("Mongoose") /M3 ("Meerkat")
ARMv8-A	64	ARM Cortex-A34 ^[79]	

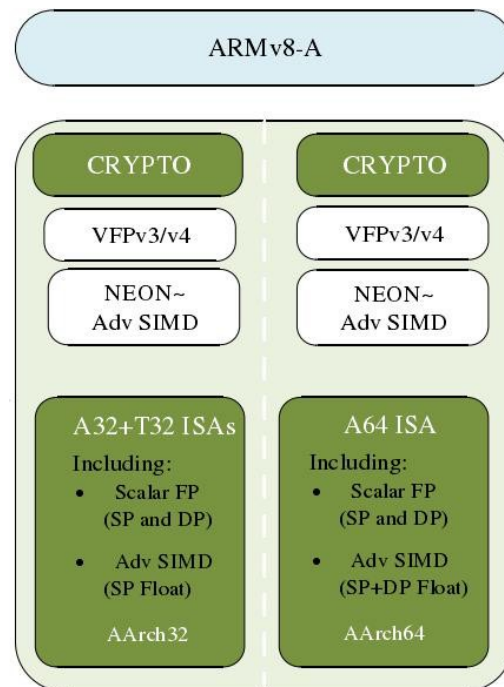
ARMv8.1-A	64/32	TBA	Cavium ThunderX2
ARMv8.2-A	64/32	ARM Cortex-A55, ^[81] ARM Cortex-A75, ^[82] ARM Cortex-A76, ^[83] ARM Cortex-A77, ARM Cortex-A78, ARM Cortex-X1, ARM Neoverse N1	Nvidia Carmel, Samsung M4 ("Cheetah"), Fujitsu A64FX (ARMv8 SVE 512-bit)
ARMv8.2-A	64	ARM Cortex-A65, ARM Neoverse E1 with simultaneous multithreading (SMT), ARM Cortex-A65AE ^[87] (also having e.g. ARMv8.4 Dot Product; made for safety critical tasks such as advanced driver-assistance systems (ADAS))	Apple Monsoon+Mistral (A11) (September 2017)
ARMv8.3-A	64/32	TBA	
ARMv8.3-A	64	TBA	Apple Vortex+Tempest (A12, A12X, A12Z), Marvell ThunderX3 (v8.3+) ^[88]
ARMv8.4-A	64/32	TBA	
ARMv8.4-A	64	TBA	Apple Lightning+Thunder (A13), Apple Firestorm+Icestorm (M1)
ARMv8.5-A	64/32	TBA	
ARMv8.5-A	64	TBA	Apple Firestorm+Icestorm (A14)
ARMv8.6-A	64	TBA	



ARM架构及优势

■ ARMv8-A架构

- 64位处理器架构，仍然支持ARMv7体系结构的32位“A32”指令集。
- 拥有更大的寻址范围、数量更多位数更宽的通用寄存器组、并发执行浮点计算的 128 b NEON 向量单元、SIMD 计算指令。
- 引入了两种执行状态（Execution state）：
 - “AArch64”（64位ARM体系结构）执行状态支持A64指令集，可以在64位寄存器中保存地址，并允许指令使用64位寄存器进行计算；
 - “AArch32”（32位ARM体系结构）执行状态则保留了与ARMv7-A体系结构的向后兼容性，使用32位寄存器保存地址，用32位寄存器进行计算。



ARM架构及优势

■ ARMv8-A架构

- 网络互联：
 - ThunderX2 和 X-Gene3 之类的处理器完全支持 **PCIe3.0**，典型的 Cray 公司开发的Aries 采用 Mellanox 的 **InfiniBand**（简称“IB”）网络进行高速互联，已成功应用于大规模的 ARM 集群。
- 架构安全性：
 - 采用**分支目标指示器**，使得间接分支可跳转到指示器所可接受的小部分代码中；
 - 使用**指针验证码**，确保函数回到程序所预设的位置；
 - 内存标签扩展功能，减少可利用内存安全错误的数量。



ARM架构及优势

■ ARMv8-A架构

➤ 架构性能：

- 用更大的物理地址，处理器可访问超过 4GB 的物理内存，此外 64 位的虚拟寻址空间，使得虚拟化功能在 ARM 服务器上得到增强。
- 采用更大的寄存器文件，减少了堆栈的使用。
- 增加新的异常模型，降低了系统和虚拟化软件管理的复杂性。
- 有效的缓存管理，通过用户空间缓存操作提高动态代码生成速度，并且能快速清除缓存。
- 用硬件对小文件快速加解密来提升性能。
- 利用 NEON 技术来实现双精度浮点计算并为 HPC 提供更大的向量宽度。

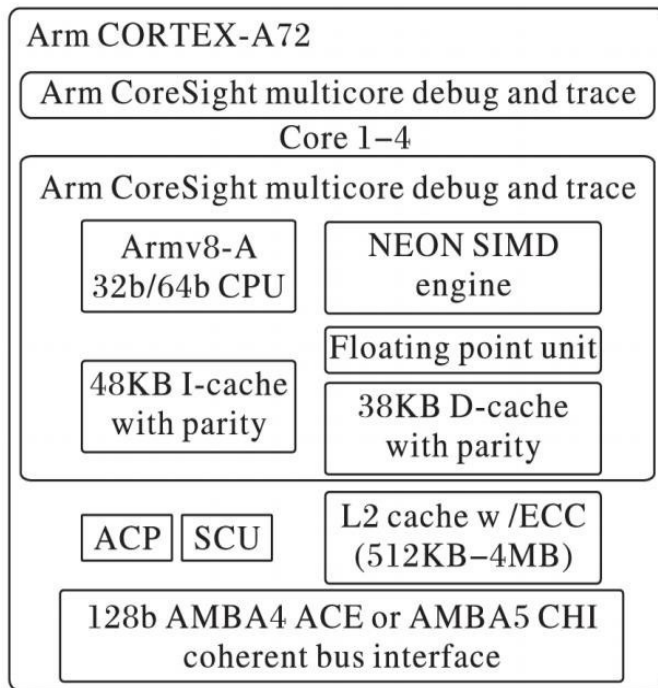


ARM架构及优势

■ ARMv8-A架构

➤ Cortex-A72 处理器

- 采用ARMv8-A架构；
- 同时兼容 32 b /64 b 指令集；
- 处理器内置 48 KB 的 3 路组相连 L1 指令 cache、32 KB 的二路组相连 L1 级数据缓存和 512 KB 到 4 MB 的 16 路组相连 L2 数据共享缓存；
- 一个电路板上集成 1 到 4 个 Cortex-A72 核心，核心之间通过 AMBA (Advanced Microcontroller Bus Architecture) 相连。
- 台积电 16 nm FinFET + 制程，使得核心主频最高可达到 2.5GHz。
- 3 发射超标量乱序流水线，每个核心内设 32 个 128 b 浮点寄存器 v0 ~ v31 专门用于SIMD计算指令。



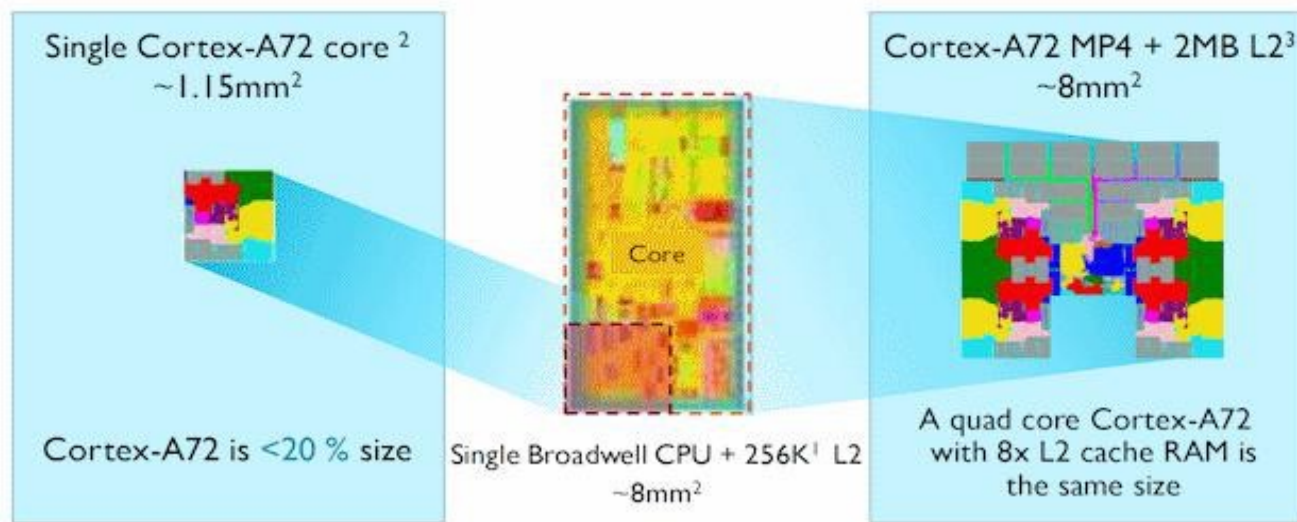
ARM架构及优势

■ ARM架构优势

➤ ARM的众核横向扩展空间优势明显

一个ARM核的面积仅为X86核的1/7，
同样的芯片尺寸下，ARM的核数是X86的4倍以上。

Cortex-A72: Ideal for dense compute environments



单个ARM核面积~1.15mm²

单个X86核面积~8mm²



ARM架构及优势

■ ARM架构优势

- 与X86架构相比的最大优势是**功耗**，ARM具有**高能耗比**。
- 通过测试，ARM架构处理器能耗平均比x86架构处理器低**20%~25%**。

Power Consumption Comparison				
	X86 (Totally)	ARM (Totally)	X86 (CPU)	ARM (CPU)
Idle	3.182W	2.474W	70.2mw	36.4mw
Cold boot	5.358W	3.280W	800mW	216mW
Sunspider0.9.1	4.775W	3.704W	722mW	520mW
Kraken	4.738W	3.582W	829mW	564mW
RIABench	3.962W	3.294W	379mW	261mW
WebXPRT	4.617W	3.225W	663mW	412mW
TouchXPRT (Photo Enhance)	4.789W	3.793W	913mW	378mW
GPU Workload	5.395W	3.656W	1432mW	488mW



ARM架构及优势

■ ARM架构优势

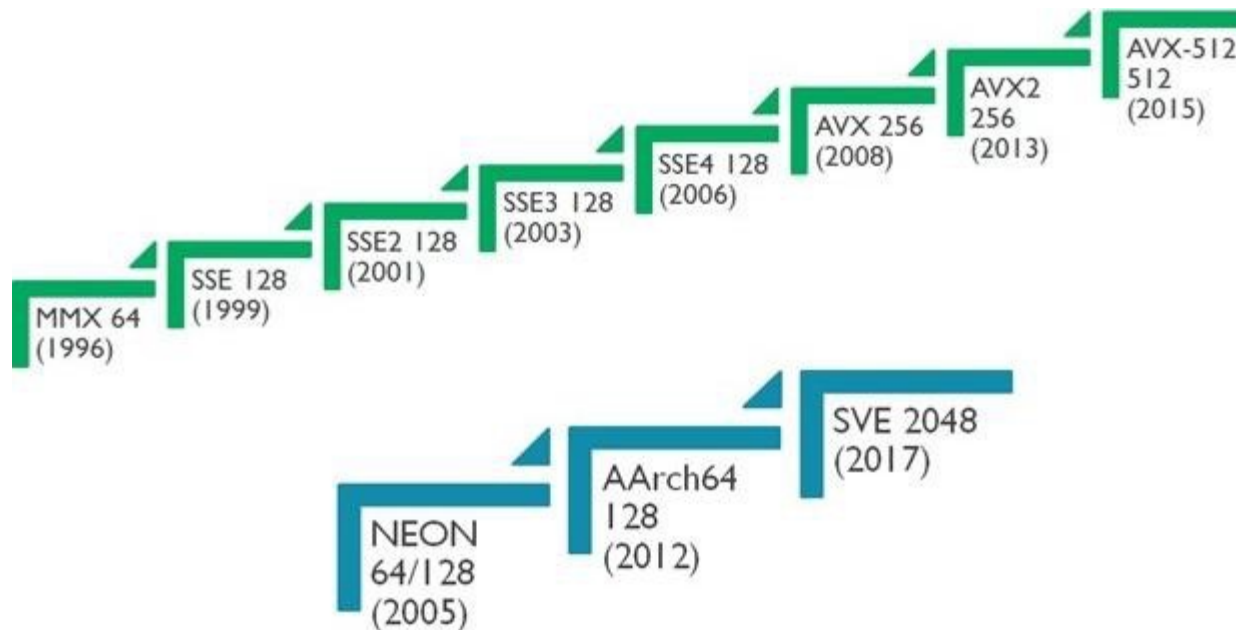
- ARM 面向高性能计算推出了可伸缩扩展矢量SVE(Scalable Vector Extension)。
- Top 500榜首日本超算富岳采用了SVE。
- SVE 是基于arm AArch64架构的下一代SIMD指令集，旨在加速高性能计算，SVE引入了很多新的架构特点：
 - 可变矢量长度：SVE提供32个向量寄存器，向量寄存器长度为128的整数倍，最低128位，最高可支持2048 位；
 - 每通道预测 (Per-Lane Predication) :SVE 提供16个预测寄存器预测寄存器每一位控制向量寄存器的一个字节，通过预测寄存器中每一位的有效状态来控制 向量寄存器中对应元素是否参与运算；
 - 聚集取和分散存 (Gather-Load , Scatter Store) ：支持非连续存储数据的高效访问。



ARM架构及优势

■ ARM架构优势

- SIMD指令发展史 intel vs arm :
 - SIMD指令总体趋势是向着越来越长的方向发展的，到了ARM SVE，最长可以支持2048位的矢量操作。



ARM架构及优势

■ ARM架构优势

- ARM HPC: 跟主流CPU FP应当算力挑战
- SVE是ARM的关键特性，显著提升在通用计算尤其HPC竞争力

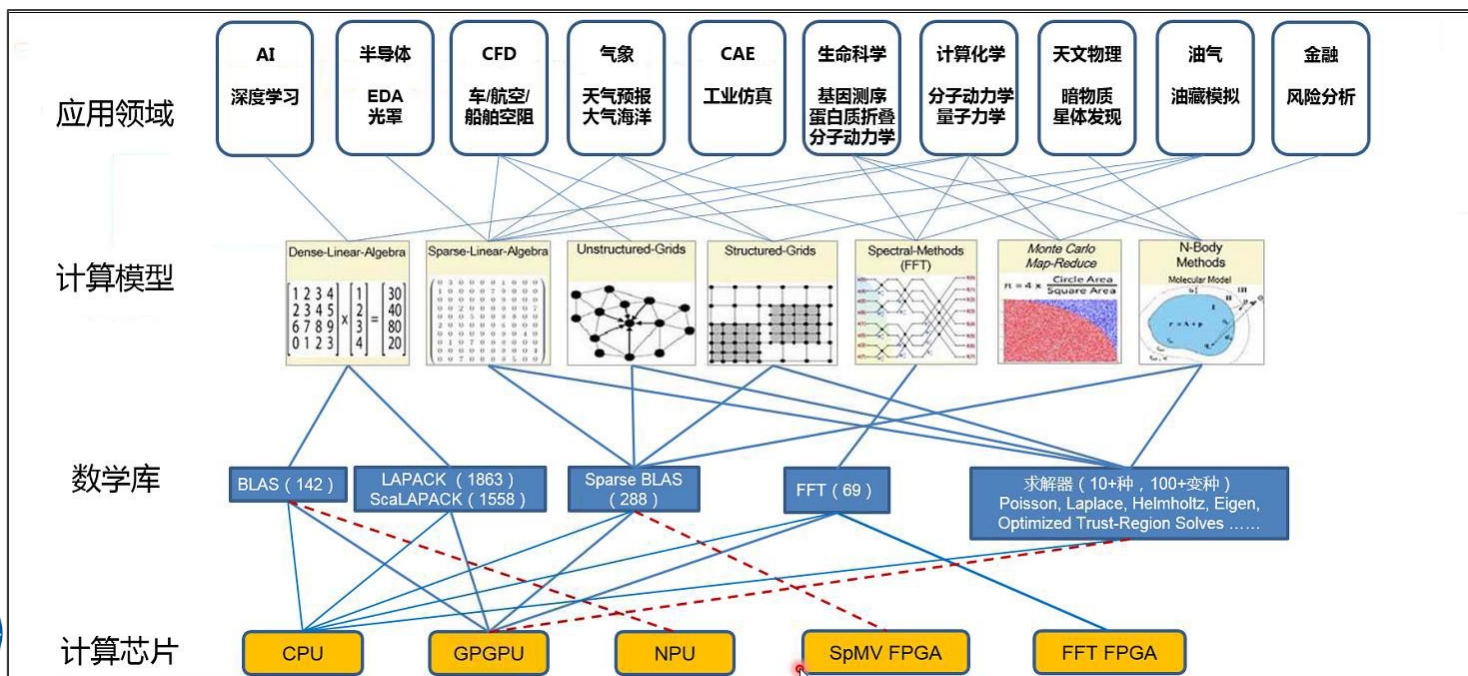
	#core	frequency (no turbo)	SIMD VL	SIMD unit cnt./core	TFLOPS(DP)
AMD zen3	64	2.25	256	3	~3.4/2.4
Intel V6 9282 (CascadeLake)	56	2.6	512	2	~3.0
A64FX(post-K)	48	1.8	512	2	~2.7
Intel V6 8280	28	2.7/1.8	512	2	~1.6
AMD 7702	64	2.25	256	2	~1.6
Intel V5 8180	28	2.5/1.7	512	2	~1.5
Intel V5 6148	20	2.4/1.6	512	2	~1.0
Hi1620	64	2.6	128	2	~0.6
Hi1616	32	2.4	128	1	~0.3



ARM架构及优势

■ ARM架构优势

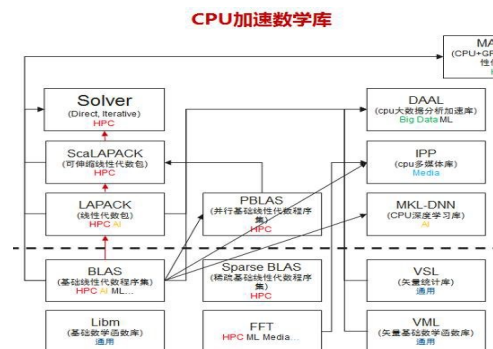
- ARM HPC: 通过SVE、算子、算法给HPC赋能
 - 通过计算能力的提升及主要算子库的优化，达到算力，能效的提升和领先。
 - 应用领域分析识别：算力需求+计算模型+数学库依赖；
 - 应用的计算效率优化：算法优化+算力分布+算子调优。



ARM架构及优势

■ ARM架构优势

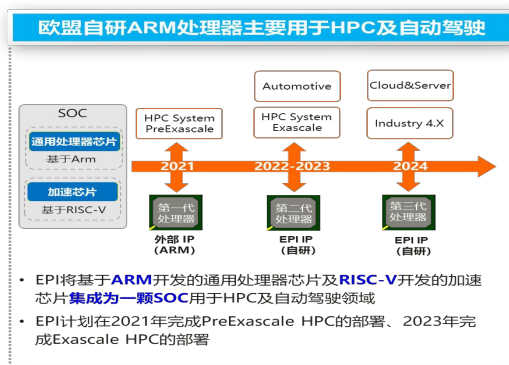
- ARM HPC: 充分利用SIMD(SVE)
 - SVE开发优化形式
 - **Assembly code**汇编代码：性能好但难度及工作量大
 - **C-Intrinsic/内置**: 难度较小，性能一般，无法精调
 - **Compiler auto-vect**: 优化难度大，但适用性广，性能case by case
 - 其他方式 pragma/directive: 性能一般
 - SIMD enabled platform主要应用形式
 - **提供高性能library**(MKL, APL, Eigen)供应用程序调用：性能及易用性均较好。
 - **提供优化compiler**：自动化程度高，性能不稳定，适用于对性能要求不高的场景。
 - 用户根据需求**自主优化**：难度大
 - **算子库**覆盖范围决定高性能应用范围



ARM架构及优势

■ ARM架构优势

➤ ARM HPC: 正成为主要超算中心的战略选择



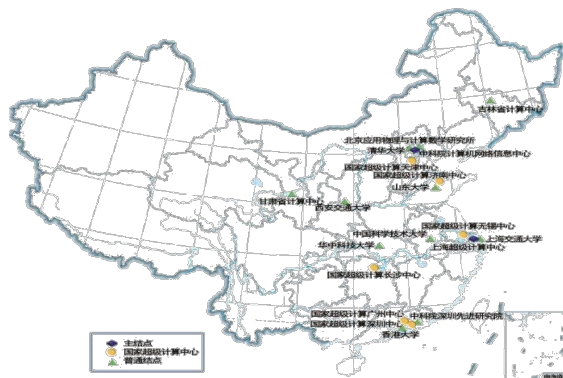
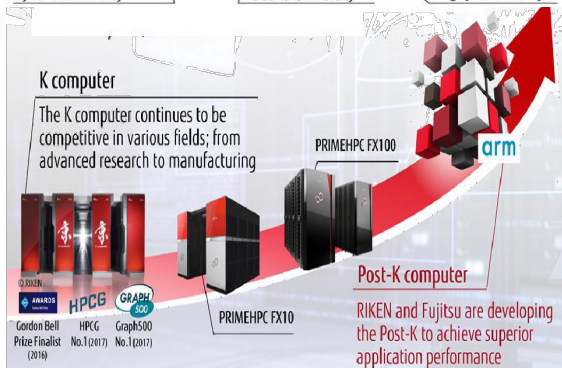
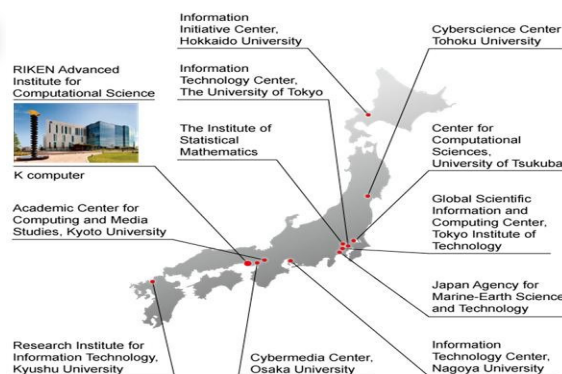
EuroHPC计划投入超35亿欧元建设HPC生态

项目时间	P级超算	PreE级超算	E级超算	预算合计 (M€)
2019-2020	Euro-IT4I 捷克 Meluxina 卢森堡 Deucalion 葡萄牙 Vega ESR 斯洛文尼亚 PetaSC-BG 保加利亚	Lumi-芬兰 BSC HE EuroHPC 西班牙 Leonardo-意大利	-	840
2021-2023	-	-	≥2套	2700

Euro HPC JU超算建设投入预算计划

• EuroHPC是**HPC的采购主体**，各区域超算中心作为**托管实体**向其申请托管预算并运营托管的HPC主机

• EuroHPC计划在2020年前投入**8.4亿欧元**建设P级、PreE级HPC，在2023年前再额外投入**27亿欧元**建设E级HPC



分类	节点	当前性能	提供商
主节点	中科院	2.3PF	曙光
	上海超算中心	399TF	申威、曙光
	无锡	126PF	申威
	天津	1PF CPU + 3.7PF GPU	飞腾
国家超算中心	济南	11.7PF	申威
	深圳	716TF CPU + 1.3PF GPU	曙光
	长沙	320TF CPU + 1.37PF GPU	国防科大
	广州	100PF	国防科大
	清华大学等	2.2PF	/
普通节点	清华大学等	2.2PF	/

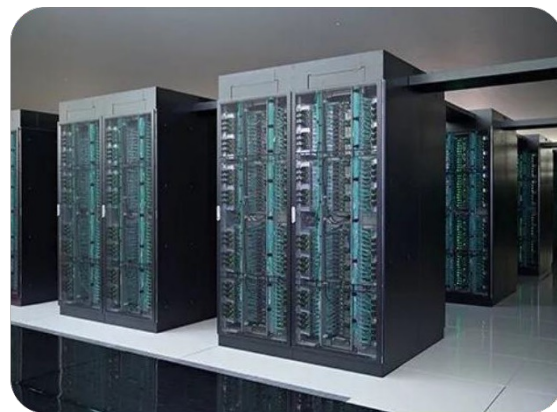


ARM架构及优势

■ ARM架构优势

- Top 500 榜首——Fugaku采用ARM架构
 - 由432个机架组成，其中396个机架各拥有384个节点，其余36个机架则各拥有192个节点，节点数总计158,976个；每个节点包含一个A64FX CPU。
 - 在Boost模式下，CPU频率最高可达2.2 GHz，整体双精度浮点运算理论峰值性能537 PFlop/s，同时还支持半精度浮点和整数运算。Fugaku总计拥有4.85 PiB存储，并提供163 PB/s总存储带宽。

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096

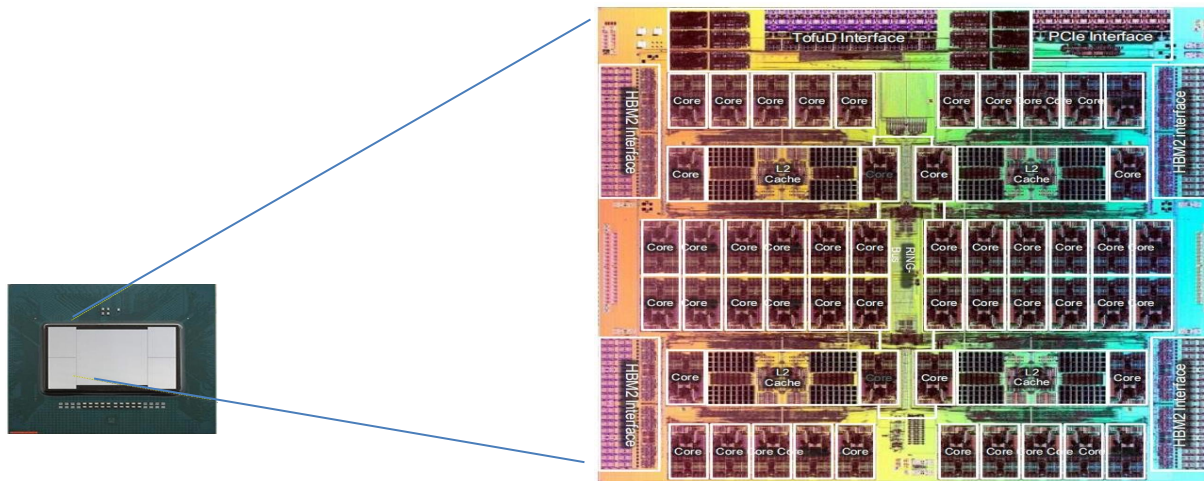


ARM架构及优势

■ ARM架构优势

➤ Fugaku

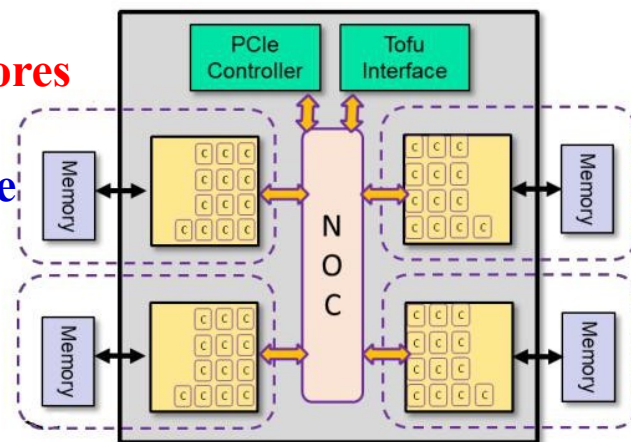
- 采用富士通ARM架构A64FX芯片：
- A64FX采用Armv8.2-A指令集，是世界上第一个采用SVE（Scalable Vector Extensions）扩展指令集的CPU。
- A64FX采用7 纳米FinFET工艺制程生产，内含87.86亿个晶体管，基础频率2 GHz，睿频可达2.2 GHz。



ARM架构及优势

■ ARM架构优势

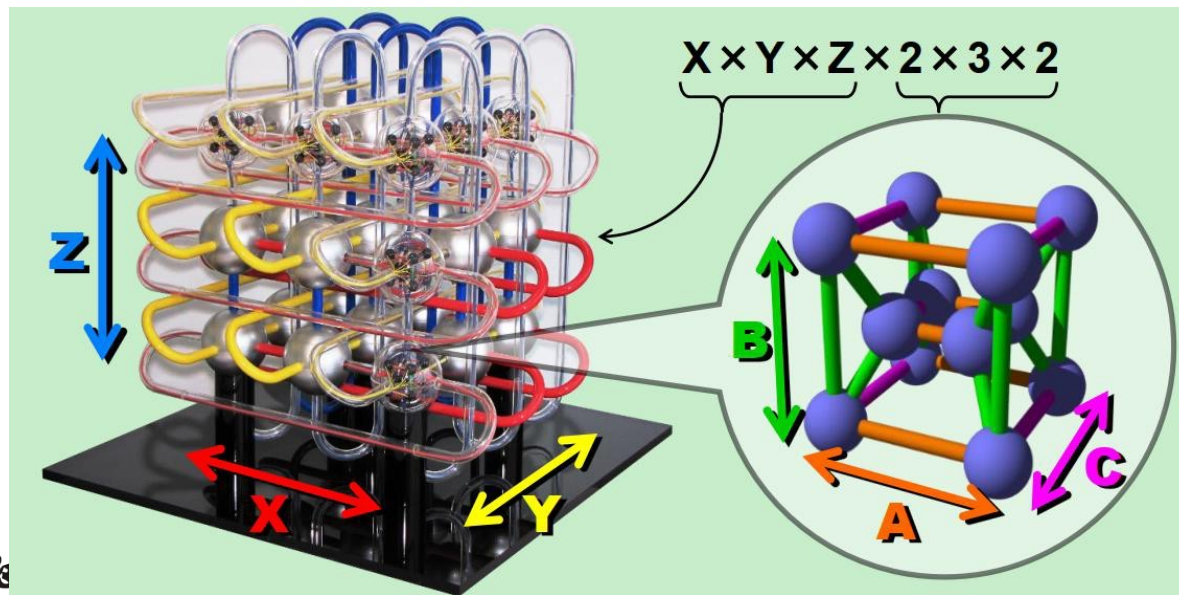
- Fugaku's FUjitsu A64fx Processor is...
- an **Many-Core ARM CPU**...
 - **48 compute cores + 2 or 4 assistant (OS) cores**
 - Brand new core design
 - Near Xeon-Class Integer performance core
 - **ARM V8** --- 64bit ARM ecosystem
 - **Tofu-D + PCIe 3** external connection
- ...but also an **accelerated GPU-like processor**
 - **SVE** 512 bit x 2 vector extensions (ARM & Fujitsu)
 - **Integer** (1, 2, 4, 8 bytes) + **Float** (16, 32, 64 bytes)
 - Cache + scratchpad-like local memory (sector cache)
 - HBM2 on package memory – Massive Mem BW (Bytes/DPF ~0.4)
 - Streaming memory access, strided access, scatter/gather etc.
 - Intra-chip barrier synch. and other memory enhancing features



ARM架构及优势

■ ARM架构优势

- A64FX: Tofu interconnect D
 - Tofu代表Torus Fusion，环形融合；
 - 最后的一个D代表，High Density的节点和Dynamic packet slicing for Dual-rail（双导轨）transfer。
 - 6D网络使用六个坐标系表示，X,Y,Z,A,B,C，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。



ARM架构及优势

■ ARM架构优势

➤ ARM HPC 总结

- SVE作为ARM新推出的SIMD ISA，能显著提升ARM生态在HPC领域竞争力。
 - 更强的算力潜力；
 - 更丰富的指令支持；
 - 预留了足够的后续扩展空间。
- SVE相关生态（硬件，软件，工具）建设还处于起始阶段。
 - 软硬协同优化，尽早发现硬件设计上的不足；
 - 基本算子库的性能优化，提升SVE的可用性；
 - 更多算子库的优化开发需要激活社区更多用户参与；
 - 工具链的完善还有很长的路要走：编译器功能，向量优化等。



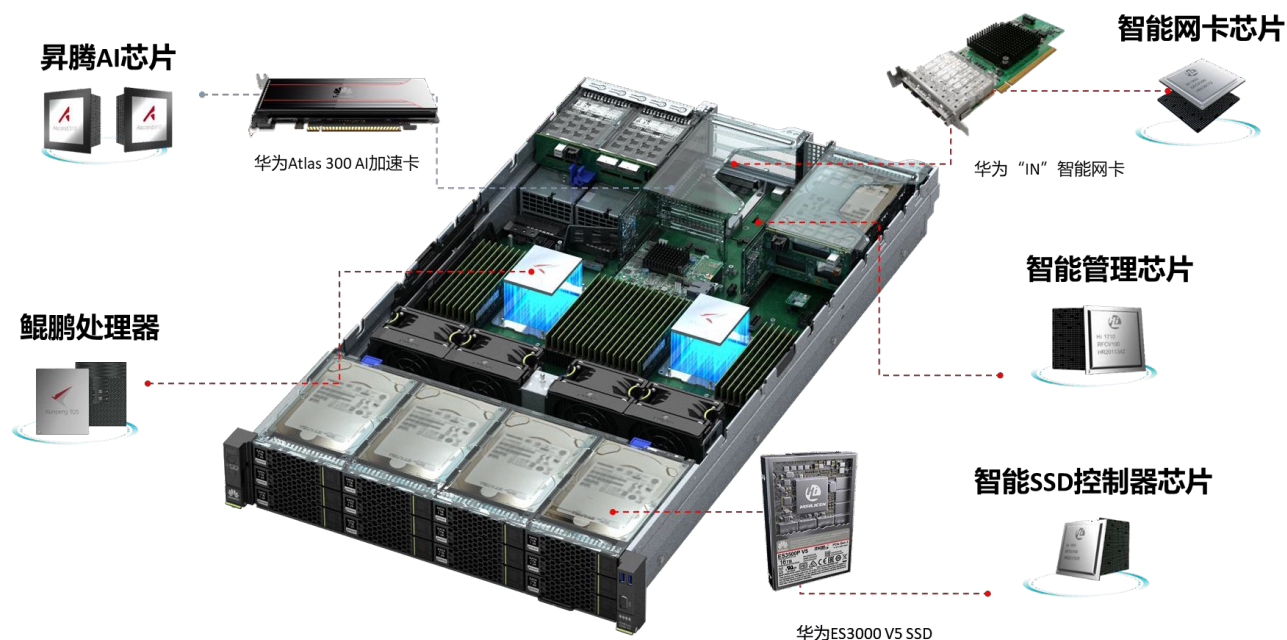
目录

- 国产超算发展
- ARM架构及优势
- 华为鲲鹏HPC介绍



华为鲲鹏HPC介绍

■ 计算平台——华为TaiShan服务器



计算能力**强**

64核2.6G H z高性能处理器

内存带宽**高**

8内存通道

I/O吞吐**高**

PCIe Gen4

2倍于PCIe Gen3带宽

散热能耗**低**

板级液冷

软硬件**协同优化**

华为编译器、M P I 、 数学库、华为调度器、集成管理软件



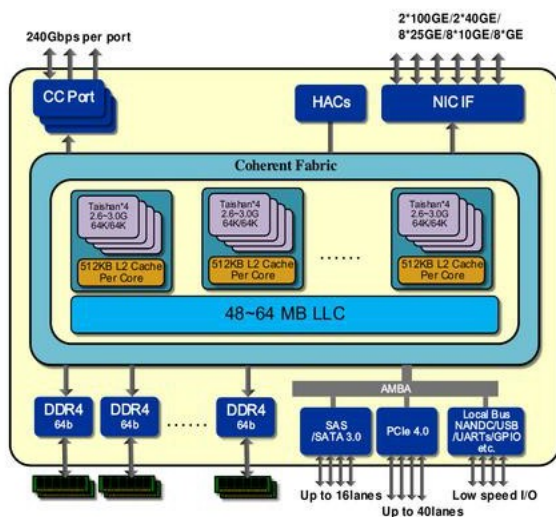
电子科技大学
University of Electronic Science and Technology of China

华为鲲鹏HPC介绍

■ 华为TaiShan服务器

➤ 采用鲲鹏 920(Hi1620)ARM芯片

Hi1620 Specifications Overview



CPU core	Up to 64 ARMv 8.2 cores, 3.0 GHz, 48-bit physical address 4 issue OoO superscalar design 64 KB L1 I Cache and 64 KB L1 D cache
L2 cache	512 KB private per core, 24 MB total
L3 cache	48 MB shared for all (1 MB/core), Partitioned
Memory	8-channel DDR4-2400/2666/2933/3200 16 ranks/channel, 1DPC and 2DPC configurations x4/x8 support ECC, SDDC, DDDC
PCIe	40 lanes of PCIe Gen4.0 16x
Integrated I/O	8 lanes of ETH, Combo MACs, supporting 2 x 100GE, 2 x 40GE, 8 x 25GE/10GE, 10 x GE, supporting SR-IOV RoCEv2/RoCEv1 x4 USB 3.0 x8 SAS 3.0 x2 SATA 3.0
Crypto engine	AES, DES/3DES, MD5, SHA1, SHA2, HMAC, CMAC Up to 100 Gbit/s
Compression	GZIP, LZS, LZ4 Up to 40 Gbit/s (compress)/100 Gbit/s (decompression)
RAID	RAID5/6, DIF, XOR, PQ acceleration
CCIX	Cache coherency interface for accelerator, like Xilinx FPGA World's 1st CCIX solution
Scale-up	Coherent SMP interface for 2P/4P 3*240Gbps bandwidth
Power	TDP ~150 W (48C 2.6 GHz)



华为鲲鹏HPC介绍

■ 华为TaiShan服务器

➤ 更强的计算能力

高性能

930+

SPECint®_rate_base2006 评估跑分

高吞吐

内存带宽: 2.4x

I/O 总带宽: 1.7x

网络带宽: 10x

高集成

1 颗 = 4 颗芯片

(CPU, 南桥、网卡、SAS 控制器)

高效能

35% ↑

*基于鲲鹏920-6426 vs 鲲鹏916处理器的华为实验室测试对比数据, 结果在不同环境中可能有偏差



工艺: 7nm | 多核: 64核 | 内存: 8通道
接口: PCIe 4.0 & 100GE



电子科技大学
University of Electronic Science and Technology of China

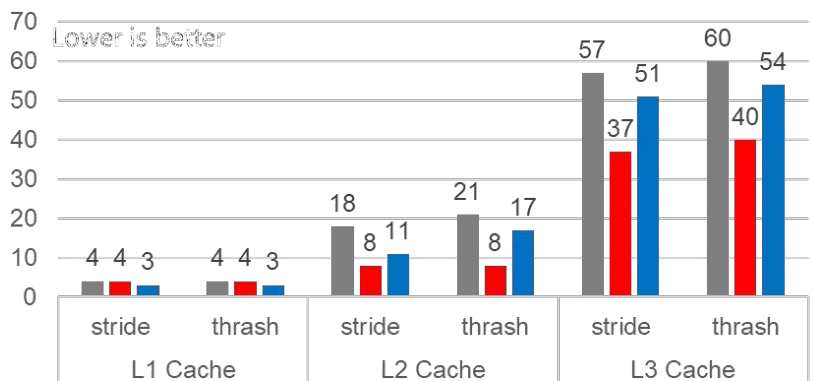
华为鲲鹏HPC介绍

■ 华为TaiShan服务器

➤ 更高的内存带宽

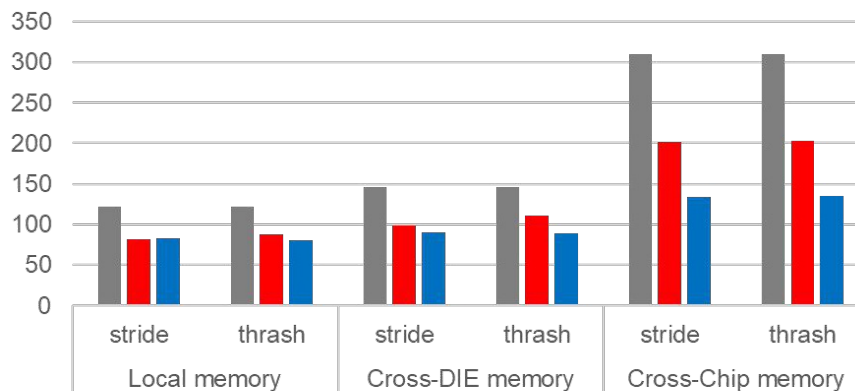
	2P Kunpeng 920 (64 cores, 2.6 GHz, DDR4-2933)	2P Skylake 6148 (20 cores, 2.4 GHz, DDR4-2666)
STREAM	284 GB/S with 75.64% efficiency	197 GB/S with 76.95% efficiency

Imbench Latency (L1/L2/L3)



■ Kunpeng 916 1core ■ Kunpeng 920 1core ■ Skylake Gold 6148 1core

Imbench Latency(DDR)



■ Kunpeng 916 1core ■ Kunpeng 920 1core ■ Skylake Gold 6148 1core

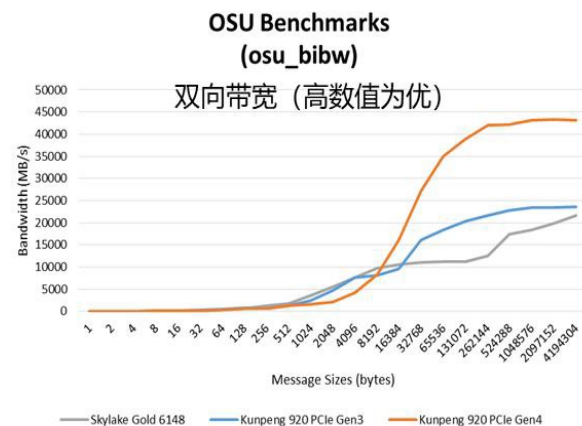
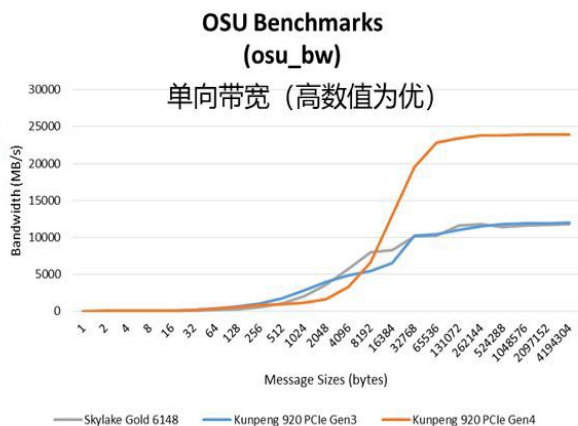
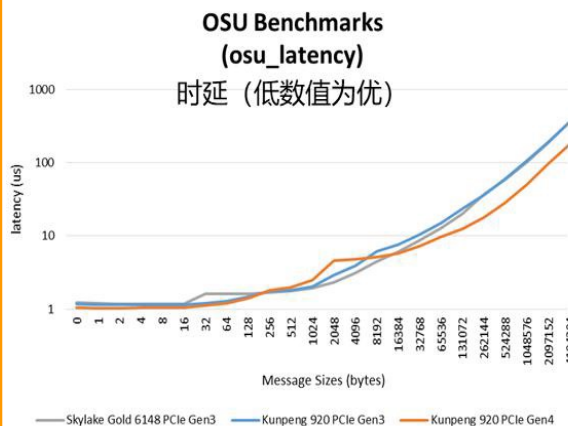


华为鲲鹏HPC介绍

■ 华为TaiShan服务器

➤ 更高的IO吞吐量

- Kunpeng 920 支持 **PCIe 4.0**
- PCIe 4.0**双口卡**能带来**两倍带宽和更低时延**
- 华为与Mellanox公司联合对PCIe Gen4进行**深度性能优化**



华为鲲鹏HPC介绍

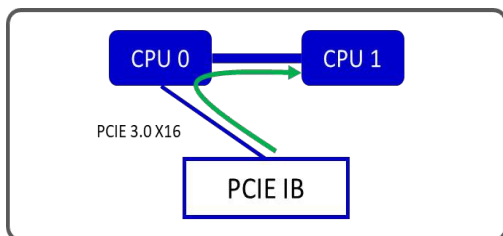
■ 华为TaiShan服务器

➤ 更高速的网络接口

InfiniBand网络

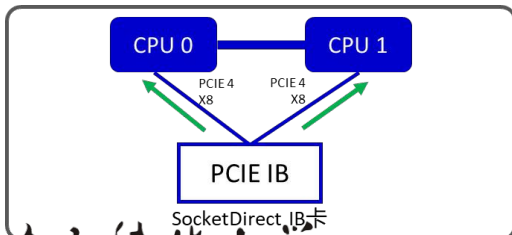
传统设计：

PCIe Gen3 网卡挂在单CPU上，
CPU1对外时延相比CPU0高



创新设计：

单插槽PCIe Gen4 x8 + x8分别连接到两个CPU
CPU1和CPU0对外时延相同



SocketDirect IB卡

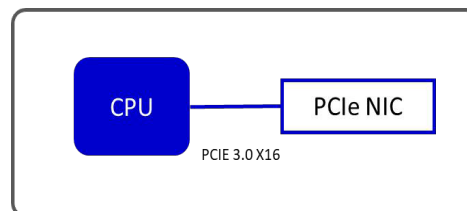


电子科技大学
University of Electronic Science and Technology of China

RoCEv2低时延计算网络

传统设计：

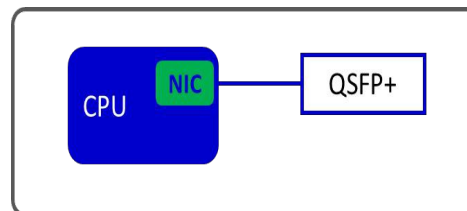
板载10G或者通过PCIe插槽插入RoCEv2网卡



创新设计：

外部信号直接到达CPU内部网卡

1. 减少PCIe信号处理，降低链路时延
2. 免网卡，降低网络投入，



华为鲲鹏HPC介绍

■ 华为TaiShan服务器

➤ 更低的散热能耗

➤ PUE (Power Usage Effectiveness) : 电力使用效率。

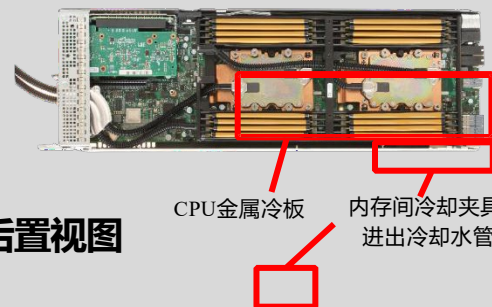
- $PUE = \text{数据中心总能耗} / \text{IT设备能耗}$ 。
- PUE越接近1表明非IT设备耗能越少，即能效水平越高。
- 国外先进的数据中心机房PUE值通常小于2，而我国的大多数数据中心的PUE值在2-3之间。

➤ 华为鲲鹏HPC

➤ 板级液冷/全液冷，液冷占比高达95%，PUE降至1.05

1. HPC场景CPU节能10%
2. 最高支持50°C进水
3. 高可靠、高抗压设计
4. 单节点插拔，易维护

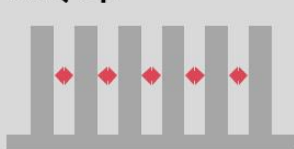
内部视图



CPU冷却



内部铲齿微通道设计



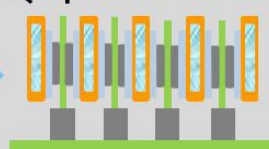
优化铲齿间距和通道流阻

优化铲齿设计，能效提升10%↑

内存冷却



传统内存冷板设计



内存间走水设计

内存间进水，热传导路径缩短热阻减小65%
栅栏式夹具设计，减少风冷接触面积80%

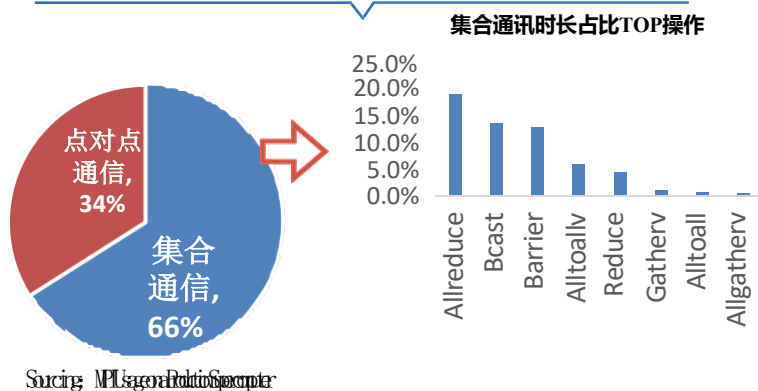


华为鲲鹏HPC介绍

■ 华为TaiShan服务器

- 华为MPI：针对MPI通信TOP瓶颈，采用**集合通信**加速创新方案，性能逼近并逐渐超越业界最佳

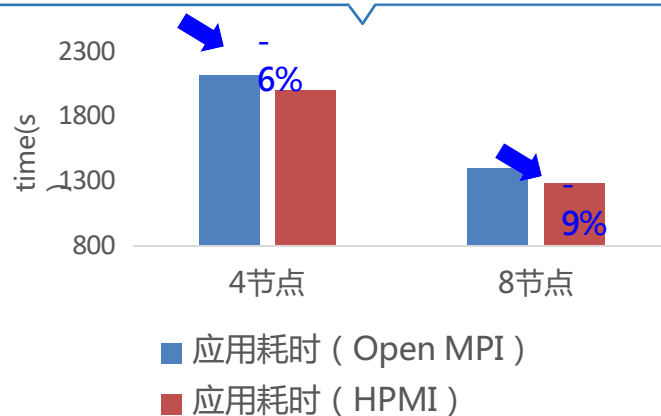
集合通讯在MPI中占比**66%**，显著大于点对点通信



HMPI创新方案

- 交换机在网计算
 - 部分计算“卸载”到交换机，减少计算节点间通信次数。
- 集合通信算法优化
 - 拓扑感知，减少跨节点通信。
- 通信架构优化
 - 优化通信框架，减少通信初始化时间。

针对TOP场景，华为MPI性能逼近并逐渐超越业界最佳



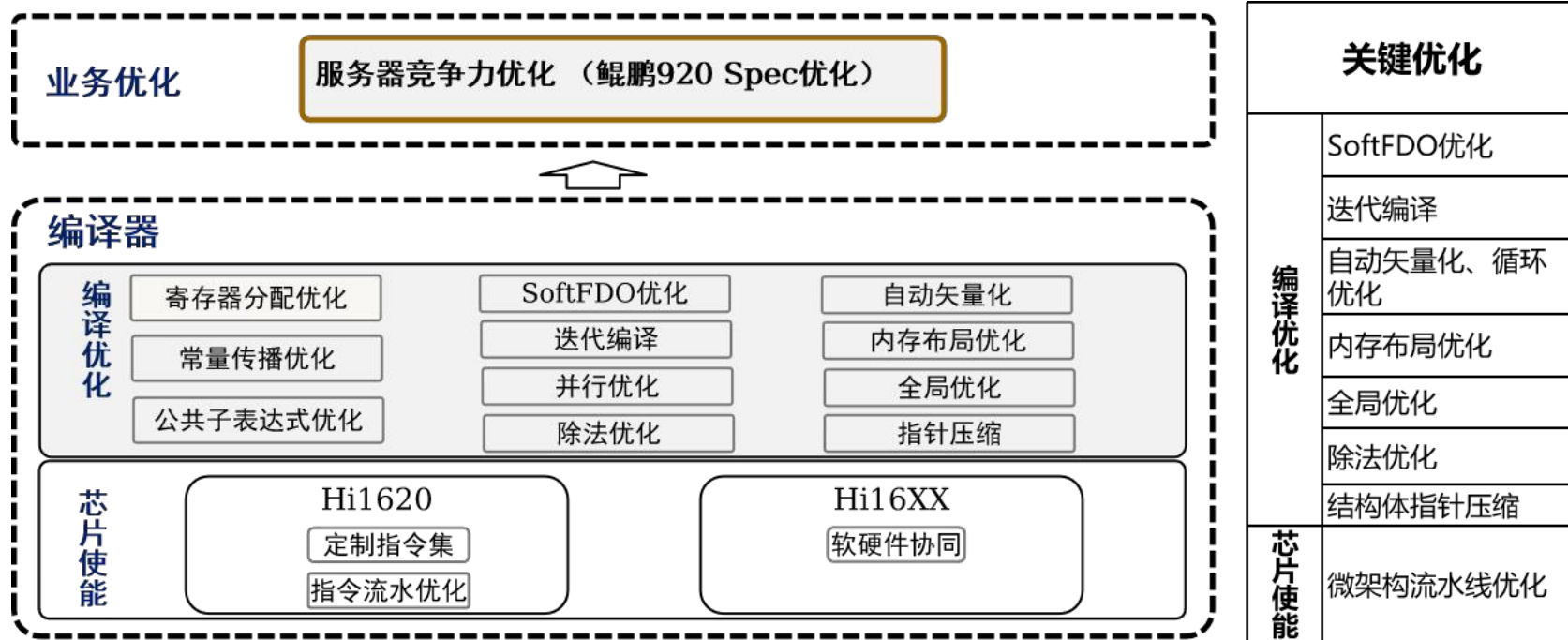
- 与开源Open MPI相比，HMPI促使应用耗时代下降5%~10%（典型应用GRAPES基线测试）。
- 2020年底，HMPI针对集合通信典型场景，基线性能持平或超越商业MPI。



华为鲲鹏HPC介绍

■ 华为TaiShan服务器

➤ HCC编译器优化技术全景图



- 自动并行化
编译器通过分析循环迭代间数据访存依赖关系，对无依赖的循环并行到多核执行，加速程序运行。
- 自动矢量化
编译器通过分析串行指令，将合适的指令序列编译成SVE向量指令。

