CAICT 中国信通院

TRUCS2019

TRUSTED CLOUD SUMMIT

可 信 云 大 会

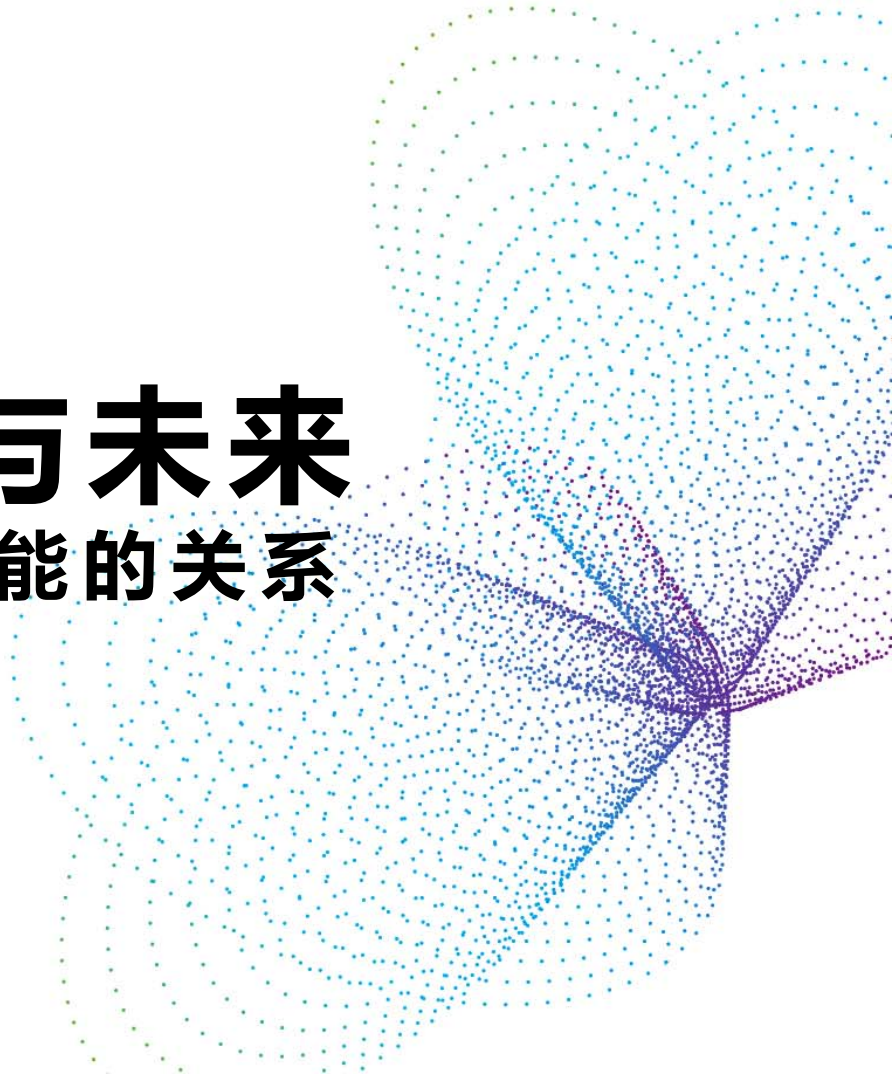
中国·北京　2019.7.2-3

高性能计算的现状与未来

--兼谈与人工智能的关系

刘 轶

北京航空航天大学 计算机学院

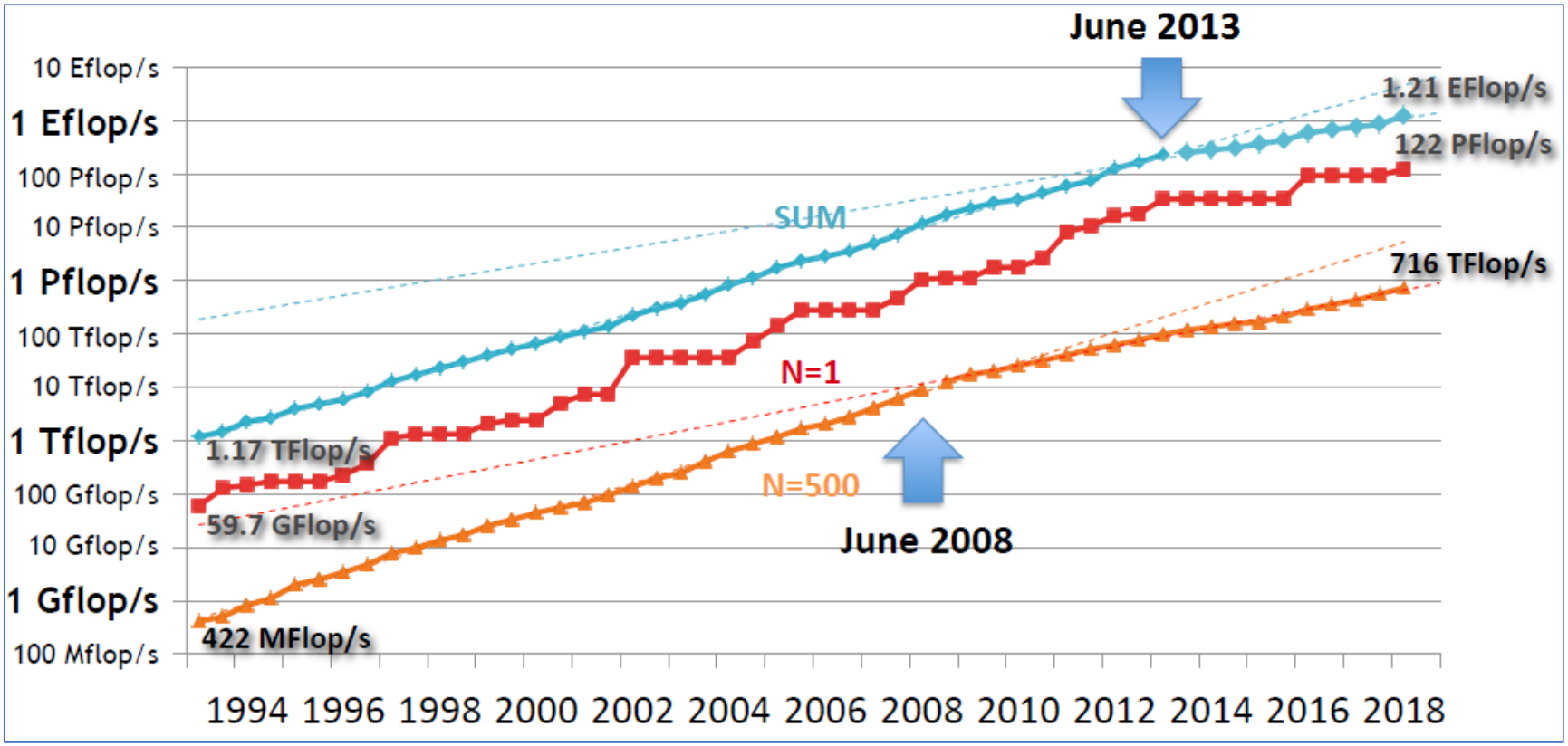


高性能计算机发展现状

高性能计算机发展现状

- 高性能计算机性能长期维持“十年千倍”的增长规律(超摩尔定律)
- 从2013年起，性能增长速度趋缓
  - 可能降至“十年百倍”，且随摩尔定律减缓还有进一步减缓趋势

TOP500性能增长
(http://www.top500.org)



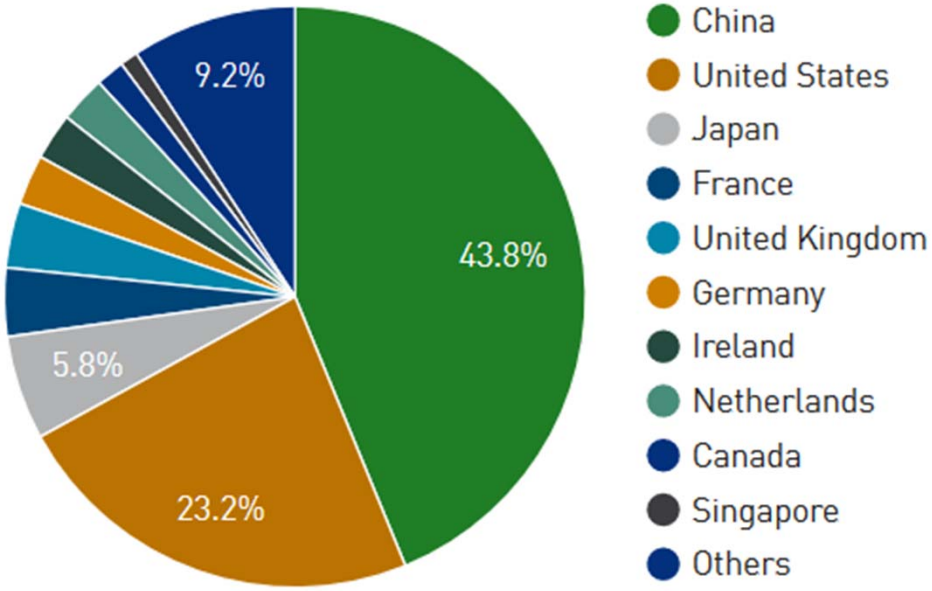
TOP500排名前10的超级计算机(2019年6月)

(http://www.top500.org)

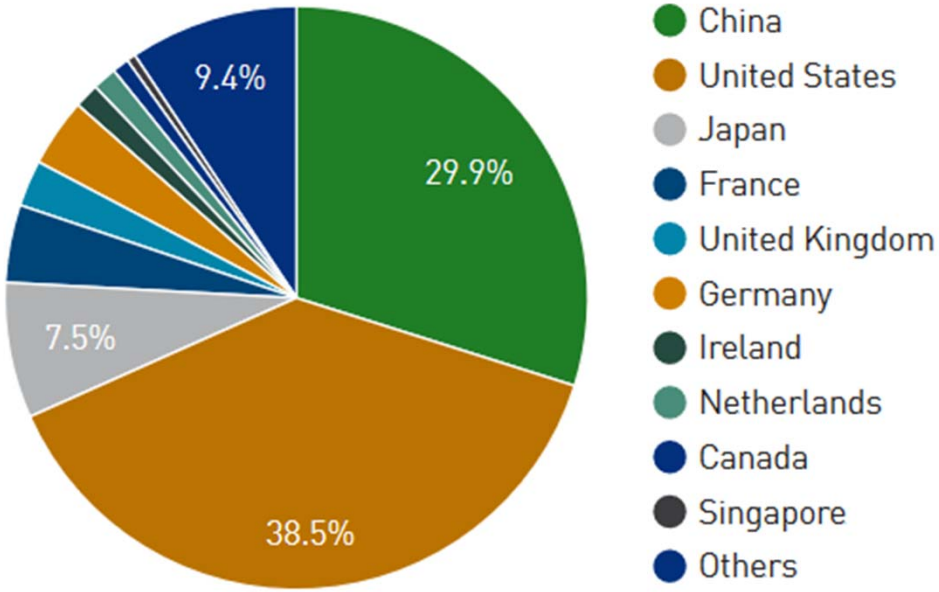
排名	地点	系统/制造商	处理器	核数	Linpack性能 (TFlops)	峰值性能 (TFlops)	功耗 (KW)
1	美国	Summit / IBM & Nvidia	IBM Power9 + Nvidia GV100	2414592	148600	200795	10096
2	美国	Sierra / IBM & Nvidia	IBM Power9 + Nvidia GV100	1572480	94640	125712	7438
3	中国	Sunway TaihuLight / NRCPC	Sunway SW26010	10649600	93014	125436	15371
4	中国	Tianhe-2A / NUDT	Intel Xeon + Matrix2000	4981760	61444	100678	18482
5	美国	Frontera / Dell	Intel Xeon	448448	23516	38745	---
6	瑞士	Piz Daint / Cray Inc.	Intel Xeon + Nvidia P100	387872	21230	27154	2384
7	美国	Trinity / Cray Inc.	Intel Xeon + Phi	979072	20158	41461	7578
8	日本	AI Bridging Cloud Infrastructure (ABCI) / Fujitsu	Intel Xeon + Nvidia V100	391680	19880	32576	1649
9	德国	SuperMUC-NG / Lenovo	Intel Xeon	305856	19476	26873	---
10	美国	Lassen / IBM & Nvidia	IBM Power9 + Nvidia V100	288288	18200	23047	---

TOP500按国别统计(2019年6月)

(http://www.top500.org)



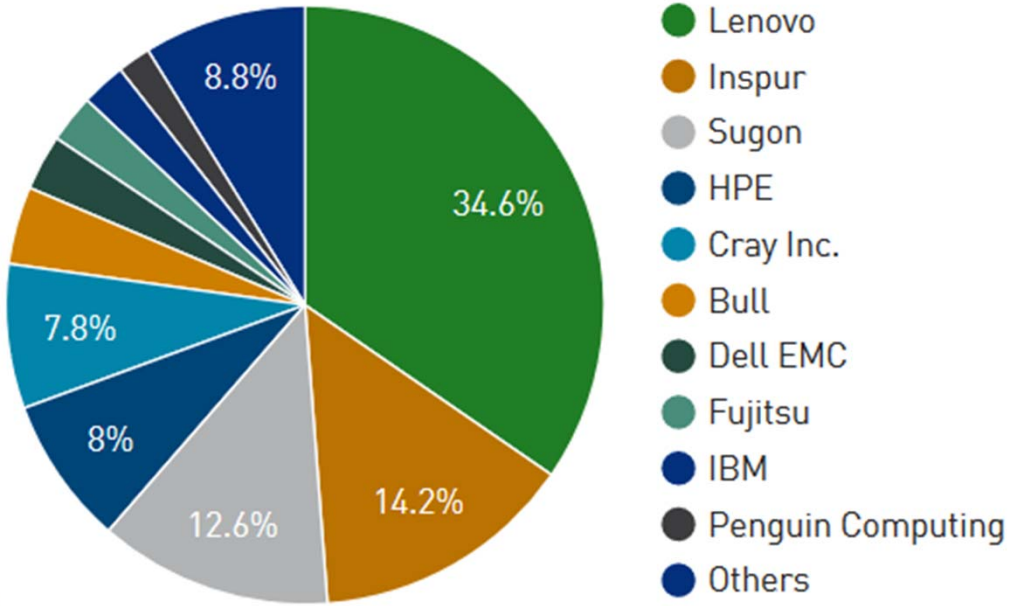
系统数量



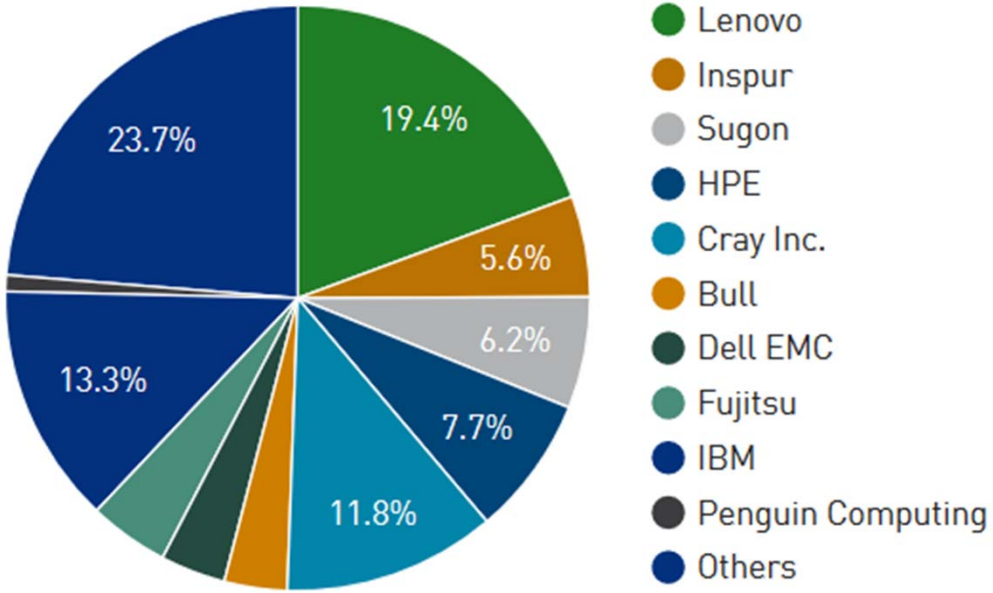
性能合计

TOP500按厂商统计(2019年6月)

(http://www.top500.org)



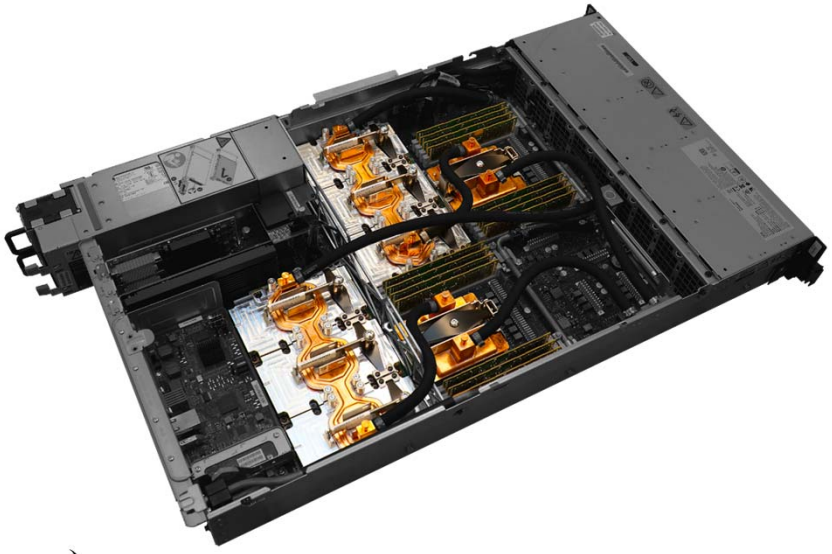
系统数量



性能合计

Summit系统简介

- 节点：峰值性能42TFlops
  - 处理器：2个IBM POWER 9
  - 加速部件：6个NVIDIA VOLTA GPU
  - 内存：512GB DDR4 + 96GB HBM
  - GPU显存：1600GB
- 节点个数：4608
  - 9216个CPU + 27648个GPU
- 互连网络：
  - 节点间互连：InfiniBand(Mellanox EDR 100G)
  - 节点内GPU间互连：NvLink
- 操作系统： Red Hat Enterprise Linux
- 并行文件系统：GPFS
  - 存储容量250 PB, 聚合I/O带宽2.5 TB/s
- 性能
  - 峰值性能200PFlops；Linpack性能148PFlops
- 安装地点
  - 美国能源部橡树岭国家实验室(Oak Ridge National Laboratory)



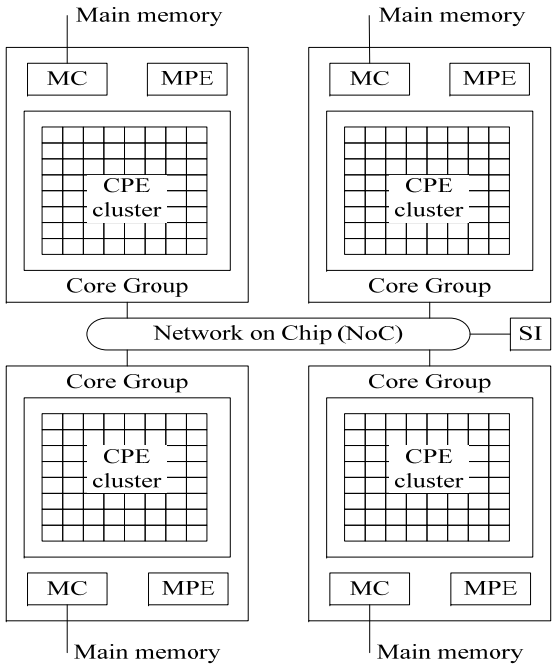
神威•太湖之光(Sunway TaihuLight)

内 容	指 标
系统峰值性能	125.4PFlops
Linpack性能及效率	93.0PFlops，效率74.1%
系统能效比	6.05GFlops/W
操作系统	符合POSIX和LSB标准的操作系统
编程语言	C、 C++、 Fortran
并行语言及环境	MPI、OpenMP、OpenACC

- 采用自主申威众核处理器SW26010
- 运算节点
  - 1个SW 26010众核处理器，32GB 2133Mbps DDR3主存和8X PCIe 3.0
- 高密度紧耦合弹性超节点
  - 256个SW26010众核处理器
- 2016年部署于无锡超算中心
- 连续四次TOP500排名第一
  - 2016年6月～2017年12月

SW26010处理器简介

- 260个处理器核：
  - 4个核组，每个核组包含1主核+64从核
  - 主核(管理核)运行操作系统，从核(运算核)负责并行计算
- 核组间通过片上网络互连，主核可直接访问主存内存
- 每个从核拥有64KB局部存储，与主核通过DMA进行数据传输
- 双精度浮点峰值性能：3.06TFlops



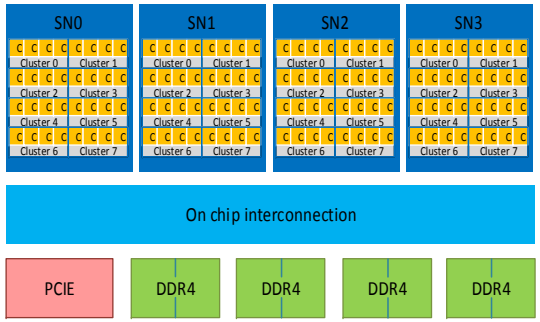
天河二号(Tianhe-2A)

内 容	指 标
系统峰值性能	110.85PFlops
Linpack性能及效率	61.4445PFlops，效率61.03%
系统能效比	5.45GFlops/W
操作系统	符合POSIX和LSB标准的操作系统
编程语言	C、 C++、 Fortran 、Java
并行语言及环境	MPI、OpenMP、OpenCL

- 一期系统
  - 2013年6月，峰值性能54.9PFlops
  - Intel x86处理器 + Intel Xeon Phi
- 二期系统(美国禁运后)
  - 2017年，峰值性能110.85Pflops
  - 基于自主加速器Matrix2000，部分节点CPU升级为FT2000
- 2013年部署于广州超算中心
- 连续六次TOP500排名第一
  - 2013年6月～2015年11月

Matrix2000众核加速器

- 128核，4个超结点，32个核/超节点
- 自定义256位向量指令集
- 每核每拍16个双精度浮点计算结果
- 峰值性能2.4576Tflops@1.2GHz
- 8 DDR4-2400 channels
- 支持X16 PCIE 3.0 EP 工作模式
- 能耗240W，能效比10GFlops/W



E级超级计算机的技术挑战

E级：Exa-scale、百亿亿次

E级计算机面临的主要技术挑战

功耗(Power)

- 业内最初设定E级系统功耗≤20MW
- 能效比须达到50GFlops/W，目前还没有有效的技术途径

应用性能(Performance)

- 追求应用可获得的性能而不是峰值性能，应用性能经常在10%甚至5%的峰值以下

可编程性 (Programmability)

- 大规模并行和异构体系结构给并行编程带来巨大困难
- 并行程序编程难，调试难，性能不确定

可靠性 (Resilience)

- 系统规模庞大→系统平均无故障时间缩短，甚至小于1小时
- 如何完成长时间不间断运行的应用？

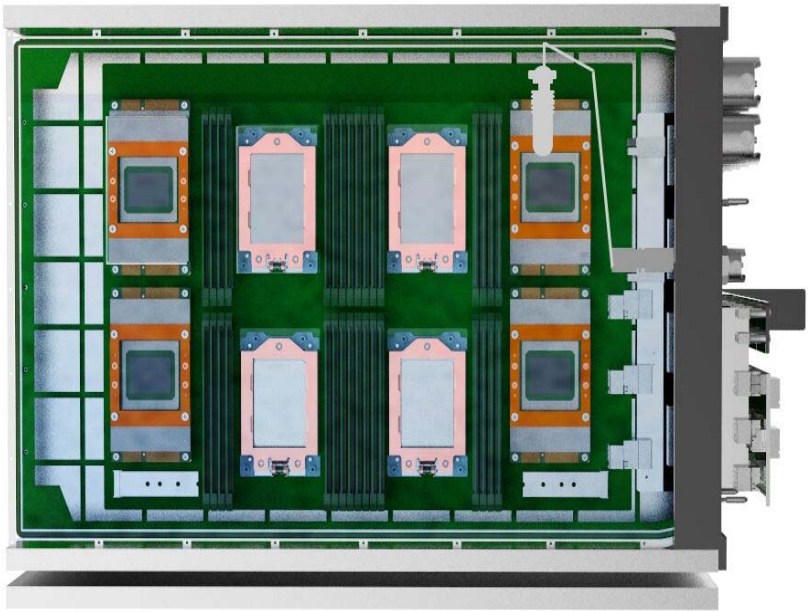
我国的三台E级原型系统

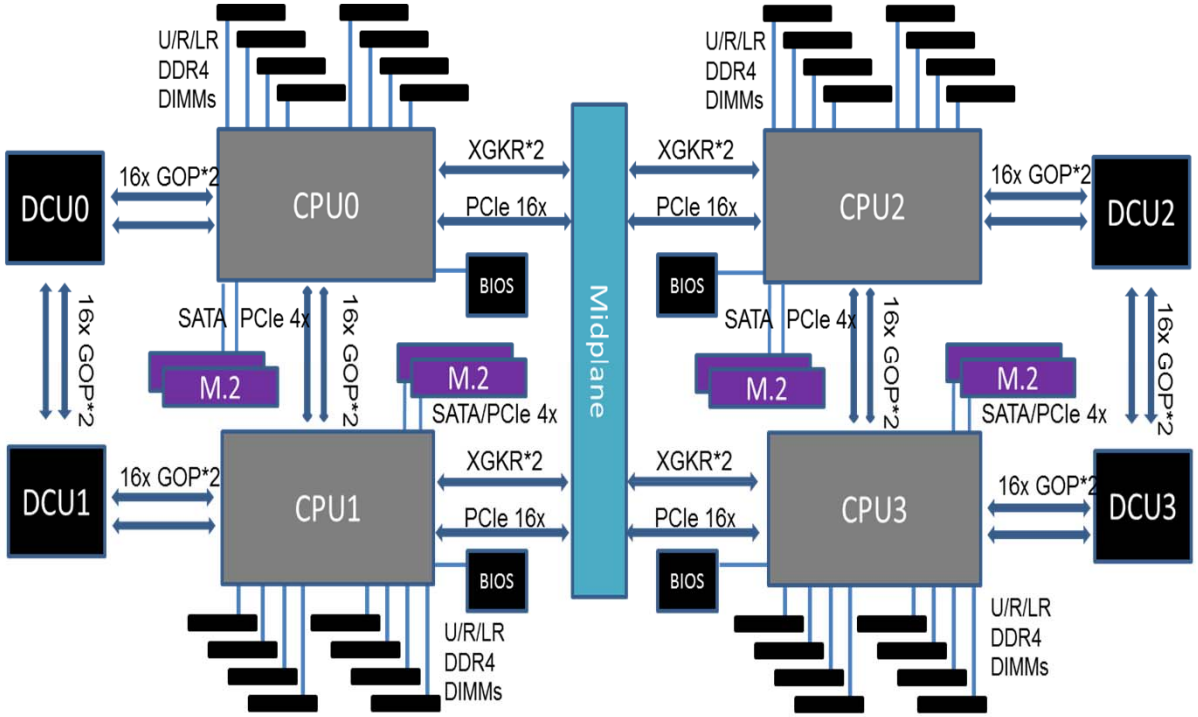
曙光E级原型系统(曙光公司)

- 处理器+加速器结构
  - 使用X86处理器以获得较好的软件兼容性
- 512节点、1024个海光x86处理器、512个海光DCU加速器
- 互连网络：6D Tours、200Gbps
- 峰值性能3.18PFlops、Linpack性能2.274PFlops、Linpack效率71.5%

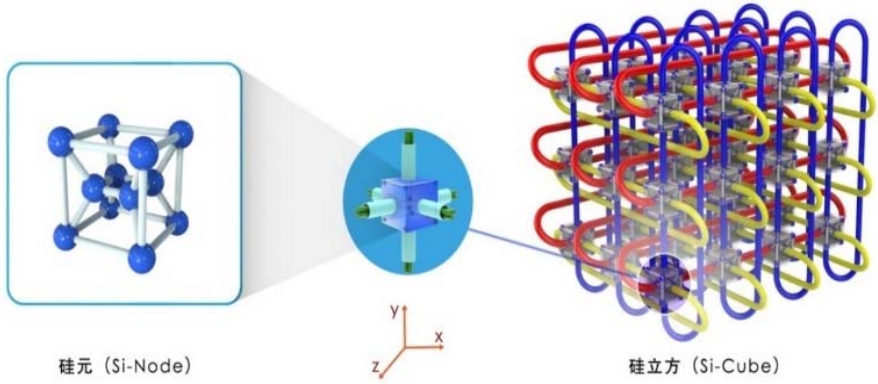
曙光E级原型系统(曙光公司)

- 节点结构
  - 2 CPU + 2 DCU，通过GOP高速总线互连
  - CPU：64位x86架构，32核
  - DCU：类GPU加速器
  - 内存：128GB DDR4



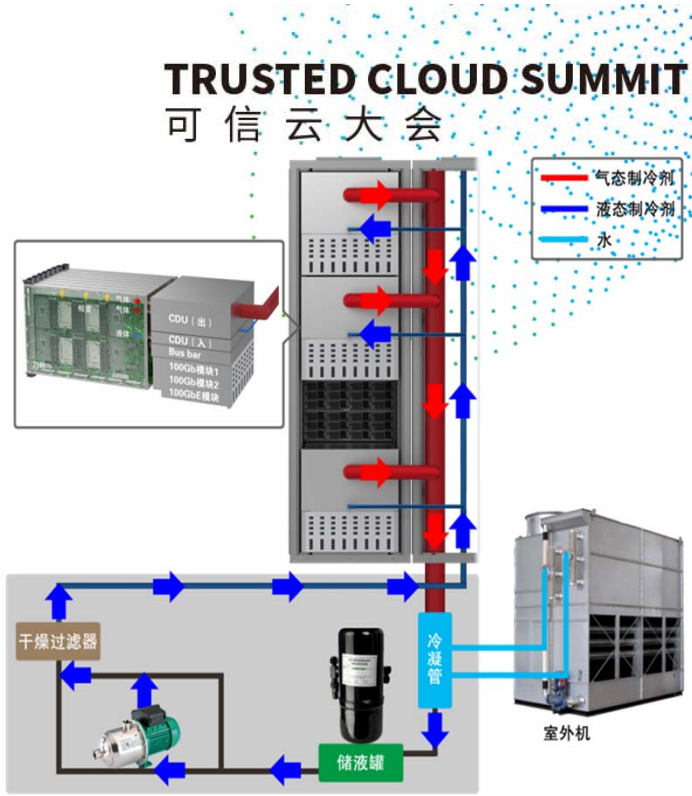


曙光E级原型系统(曙光公司)



基于 6D-Torus 的层次化的高速网络结构

- **第一层**：超节点（Super Node）内全线速交换
- **第二层**：超节点间基于局部 a-b-c 坐标的 3D-Torus互连
- **第三层**：硅元（Silicon-Node）间基于全局 X-Y-Z 坐标的 3D-Torus互联
- 光交换快速通路，解决Torus网络在网络跳步数、网络全局通信性能方面的问题



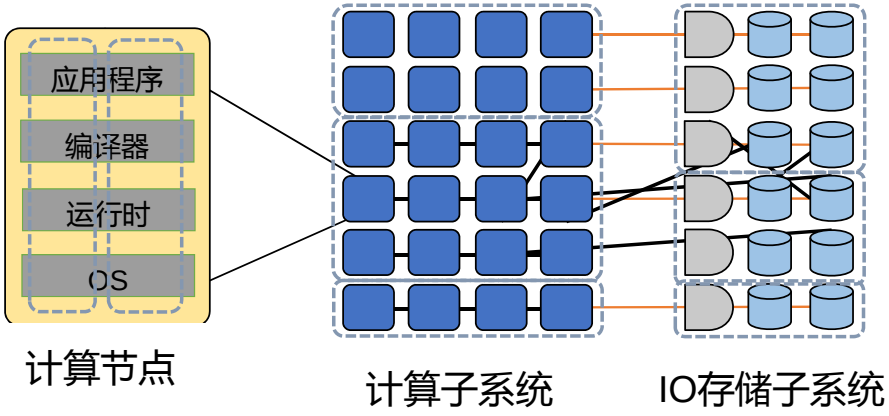
全浸泡式蒸发冷却技术

- **Imm058制冷剂、常压沸点50℃**

更节能	PUE<1.1
更可靠	降低CPU核温20℃
更低噪音	无风扇设计
更高性能	CPU超频性能提高5%

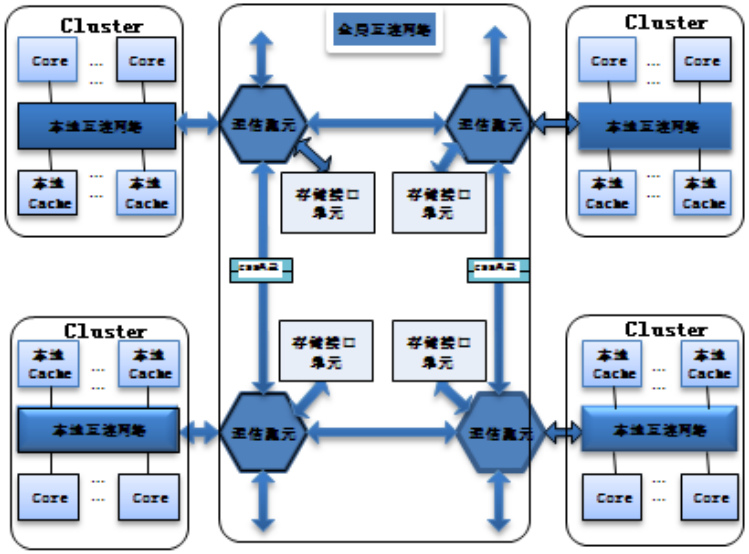
天河E级原型系统(国防科大)

- 可配置柔性系统架构，以适应不同应用需求
  - 可配置的计算环境
  - 软件定义互连
- 高速互连网络
- 512节点
  - 峰值性能：3.14PFlops
  - Linpack效率：78.5%
- 部署于国家超算天津中心



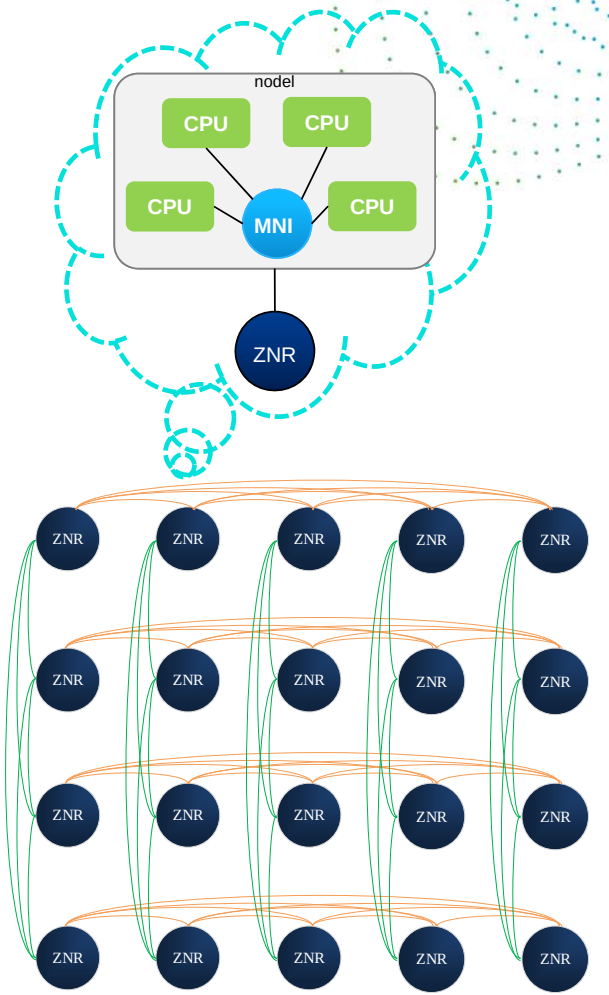
天河E级原型系统(国防科大)

- 众核处理器MT-2000+
  - 天河2A中MT-2000的优化版
  - 128核，峰值2.048Tflops
  - 16nm FinFET工艺
  - 2.0 GHz核心工作频率
  - 典型功耗130W
  - 能效比达到15Gflops/W以上
- 节点
  - 3个MT-2000+
  - 节点峰值性能 >6TFlops



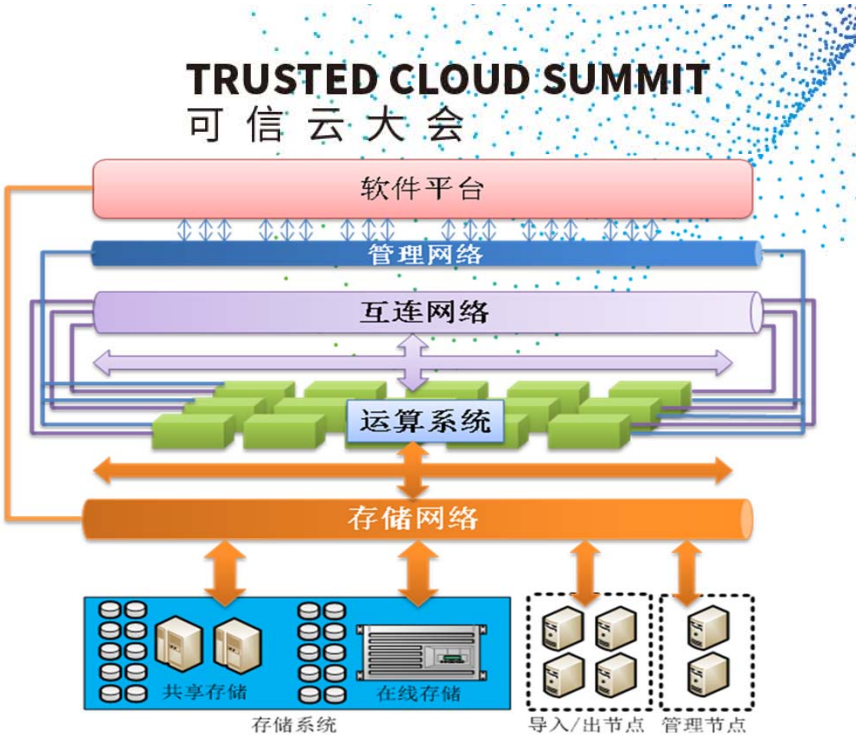
天河E级原型系统(国防科大)

- 高可扩展三维蝶形网络结构
  - 第一级互连：机框内计算结点互连
  - 第二/第三级互连：采取二维蝶形网络拓扑结构，实现每个维度点到点直接相连
- 特点
  - 性能高：全系统结点间最大步长为4
  - 易容错：支持网络流量软件控制和网络平面容错备份
  - 可扩展：多个维度扩展支持10万结点以上规模



神威E级原型系统(江南计算所)

- 面向多目标优化的多态多尺度自适应体系结构
  - 基于国产申威众核处理器
  - 高密度弹性超节点
  - 高流量复合网络架构
  - 512个节点
  - 峰值性能3.13PFlops
  - Linpack效率81.51%
- 从硬件层、软件层到应用层，全面验证未来E级计算机关键技术
- 部署于国家超算济南中心



神威E级原型系统(江南计算所)

- 运算节点
  - 2个SW26010处理器
    - 260核处理器：4大核+256小核
  - 峰值性能： >6TFlops
  - 单节点能效：约11GFlops/W
- 运算超节点
  - 规模：256节点，256X256全交叉互连
  - 单点上网：2路25Gbps X4
  - 点对点单向带宽：200Gbps

运算节点板(双CPU)

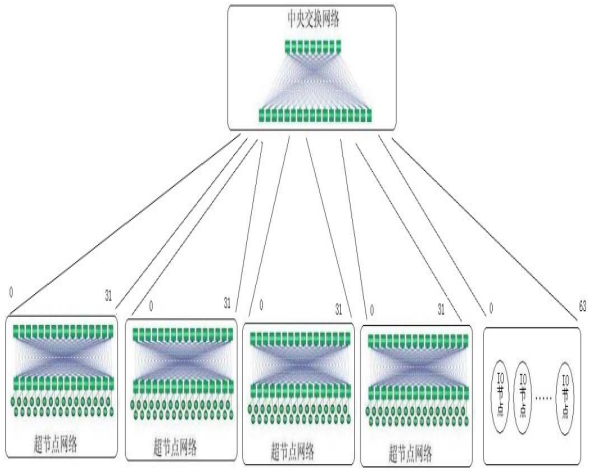
运算插件(8 CPU+4 NI)

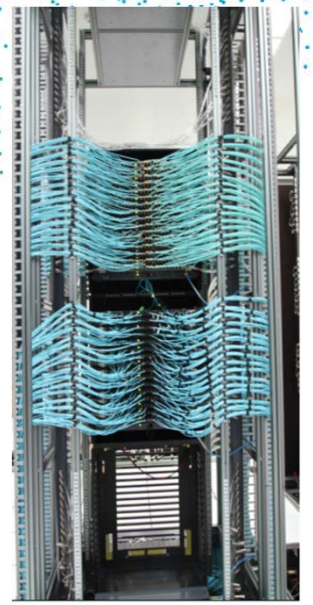
运算超节点(32个运算插件)

原型系统机仓

神威E级原型系统(江南计算所)

- 互连网络系统
  - 采用高流量可扩展复合网络结构和自研网络芯片组
  - 二级胖树全交叉互连结构
  - 规模：512节点+64 I/O节点
  - 链路传输速率: 25 Gbps
  - 网络延迟：<1.5us
  - 可扩展性：支持10万节点以上互连

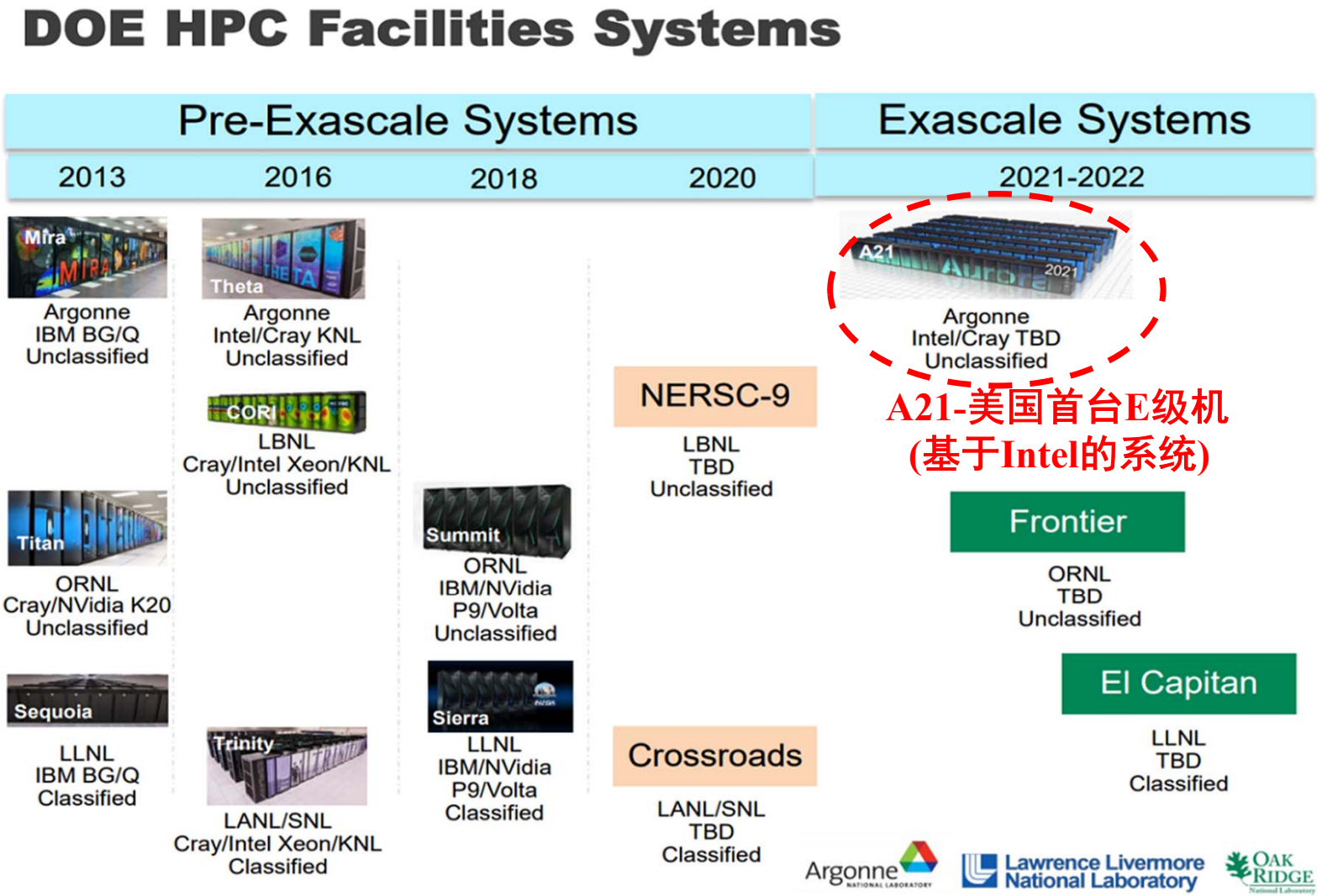




美、日的E级超级计算机情况

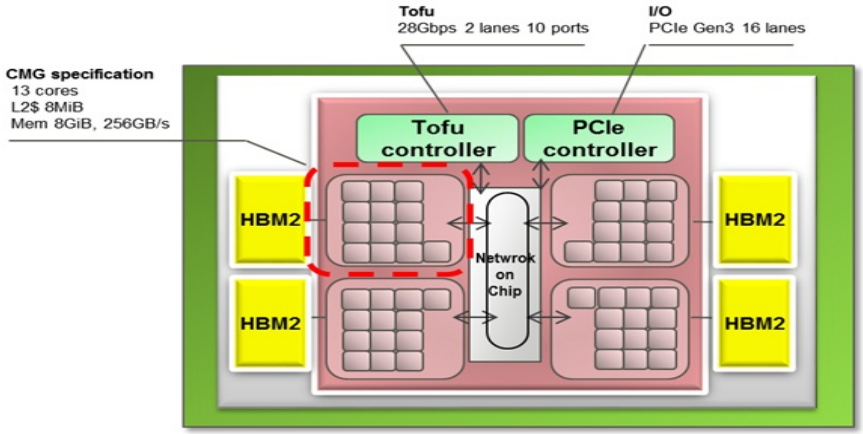
美国的E级机研制

- 美国提出NSCI计划，多个政府部门协同发展超级计算
- DoE实施ECP计划
  - 2021年上半年：A21，峰值1EFlops
  - 2021-2022年：Frontier，持续性能1EFlops



日本Post-K (Fugaku, 富岳)

- Fujitsu公司和日本理化研究所(RIKEN)联合研制
- 系统峰值性能>400PFlops
  - 号称应用性能达到E级
- 采用ARM架构处理器A64FX
  - 48个计算核 + 2个辅助I/O核
  - 向量部件SVE
  - 双精度浮点性能≥2.7TFlops
  - 高访存带宽
- 节点数：150,000，1处理器/节点
- 2019年3月开始生产



CPU	Name	A64FX™
	Instruction set architecture	Armv8.2-A SVE
	Number of cores	Computational node: 48 cores + 2 assistant cores I/O and computation node: 48 cores + 4 assistant cores
	Theoretical computational performance	Over 2.7 TFLOPS (double precision)
Nodes	Architecture	1 CPU/node
	Memory capacity	32 GB (HBM2, 4 stacks)
	Memory bandwidth	1,024 GB/s
	Interconnects	Tofu Interconnect D
Racks	Maximum number of nodes	384 nodes/rack
Software	OS	Linux
	HPC middleware	A successor to the Fujitsu Software Technical Computing Suite

- 高性能计算系统的特点和趋势小结
  - 以“CPU+加速部件”、“通用核+计算核”为代表的异构系统已成为主流
  - 从面向工程/科学计算 → 兼顾人工智能、大数据等应用
  - 不断增大的系统规模带来若干技术挑战
    - 功耗(Power)、性能(Performance)、可编程性(Programmability)、可靠性(Resilience)
  - 世界主要强国围绕E级超级计算机的竞争激烈
  - 摩尔定律(Moore’s Law)的终结将给高性能计算乃至整个计算机产业带来巨大影响
    - 大约2023年前后
  - 我国高性能计算发展的主要短板
    - 基础技术较为薄弱
    - 基于国产处理器的软件生态环境
    - 高性能计算应用软件(几乎所有的大型商用计算软件均为进口)

需要正确的策略和长期持续的努力

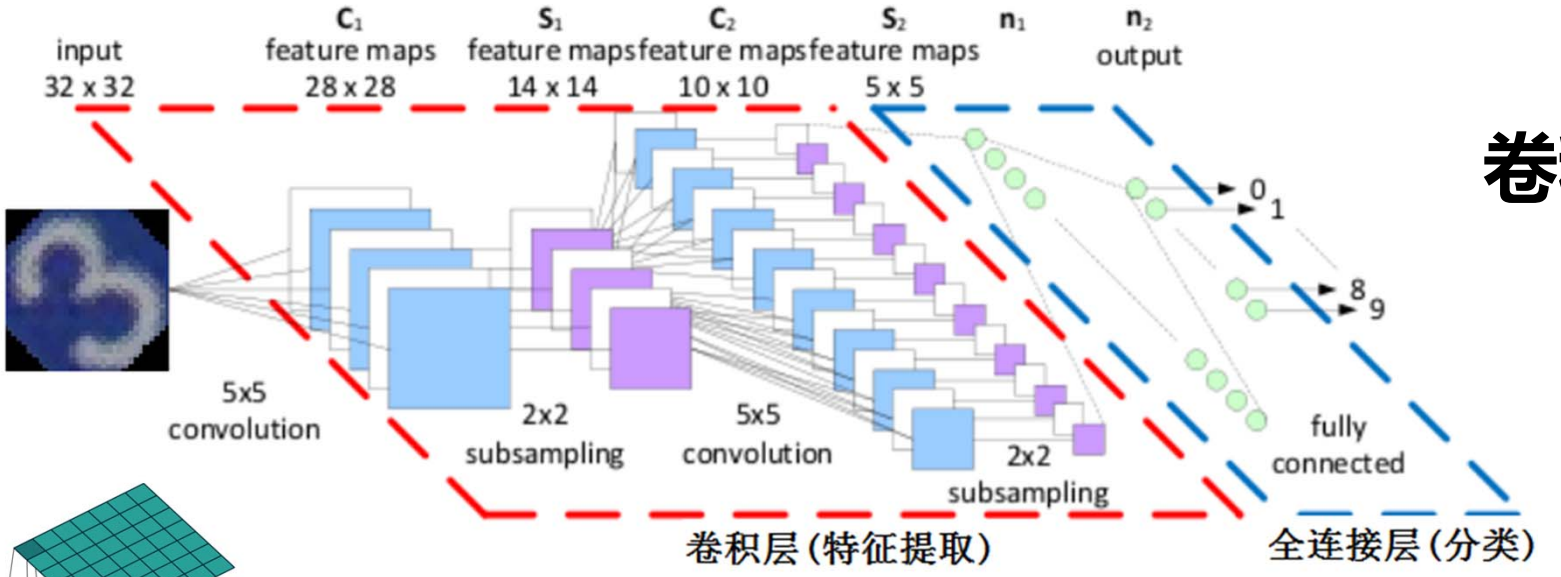
“十二五”高性能计算重大项目经费投入

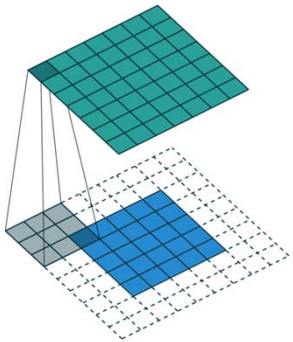
类　别	占　比
高性能计算机研制	91%
软件及应用研发	6%
关键技术研究及评测	2%
应用服务环境	1%

TRUSTED CLOUD SUMMIT
可 信 云 大 会

高性能计算与人工智能

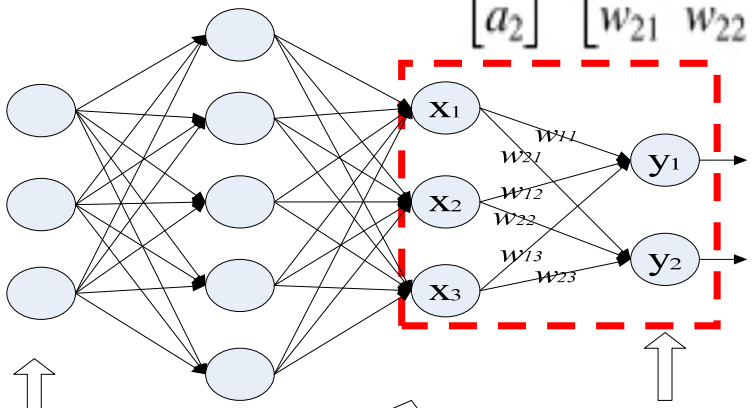
卷积神经网络示例





本例中卷积计算可
归结为3x3矩阵相乘

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$



神经元数量
一般在千万
以上

- 深度神经网络的核心是包含多个隐层的神经网络
- 卷积层实现图像的特征提取，作为神经网络的输入
- 经过大量训练样本的训练，得到神经元之间连接权值，并在推理计算时实现图像的识别(分类)
- 卷积层和全连接层的计算量占比达到90%以上
- 卷积层和全连接层的核心计算均为向量/矩阵相乘

- **高性能计算技术是新一代人工智能的重要支撑**
  - 新一代人工智能技术的基础核心是深度学习(deep learning)
  - 多层神经网络模型训练和推理需要庞大的计算量，适合高性能计算平台
  - 人工智能应用已成为高性能计算机上的重要应用
    - 美国的高性能计算机支持了一系列AI相关研究项目
    - 国产超级计算机上已实现了深度学习框架的移植，并支持了诸多深度学习应用

Summit(排名第1)
logo和介绍语

system Titan. Summit is providing scientists with incredible computing power to solve challenges in energy, artificial intelligence, human health, and other research areas, that were simply out of reach until now. These discoveries will help shape our

About ABCI

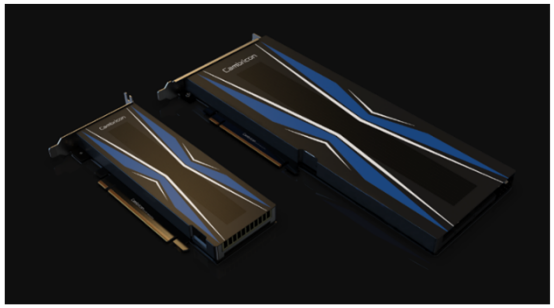
AI Bridging Cloud Infrastructure (ABCI) is the world's first large-scale Open AI Computing Infrastructure, constructed and operated by National Institute of Advanced Industrial Science and Technology (AIST).

World's Largest, Super Energy Saving, Open AI Infrastructure

ABCI(排名第8)
介绍语

- **高性能计算与人工智能相互融合**
  - 人工智能应用的快速发展反过来影响着高性能计算
  - 表现一：人工智能技术被应用于高性能计算
    - 基于机器学习的任务调度、程序性能优化、。。。
    - 深度学习技术用于特定领域应用的优化：蛋白质折叠/新药发现、优化求解、…
    - 机器学习技术用于科学计算大数据分析
  - 表现二：高性能硬件从面向科学/工程计算 → 兼顾人工智能
    - Nvidia GPU的Tensor core
    - 深度学习处理器/加速部件：Google TPU、寒武纪、FPGA、ASIC…

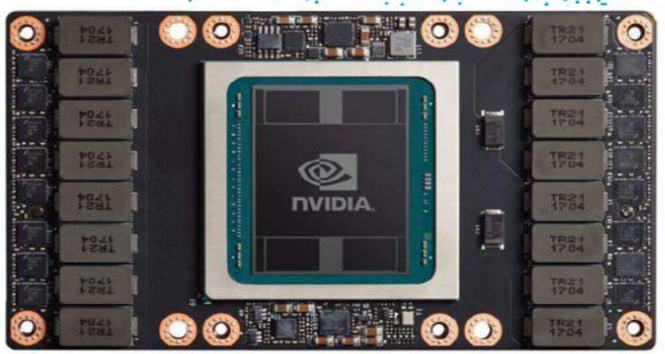
Google TPU



寒武纪智能云服务器加速卡

- **Nvidia Tesla V100**
  - **5120个CUDA核**
  - **640个Tensor核**

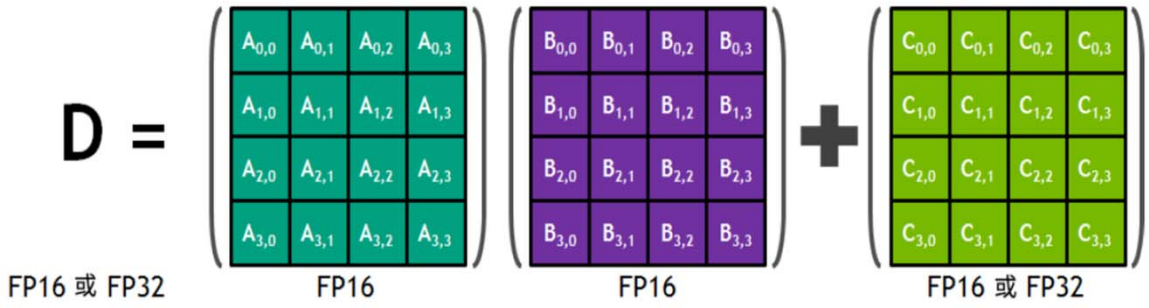
指 标	NVLink 版本	PCIe 版本
双精度浮点性能	7.8TFlops	7TFlops
单精度浮点性能	15.7TFlops	14TFlops
半精度浮点性能 (深度学习)	125TFlops	112TFlops
互连带宽	300GB/s (NVLink)	32GB/s
显存及带宽	32/16GB HBM2； 900GB/s	
功耗	300W	250W

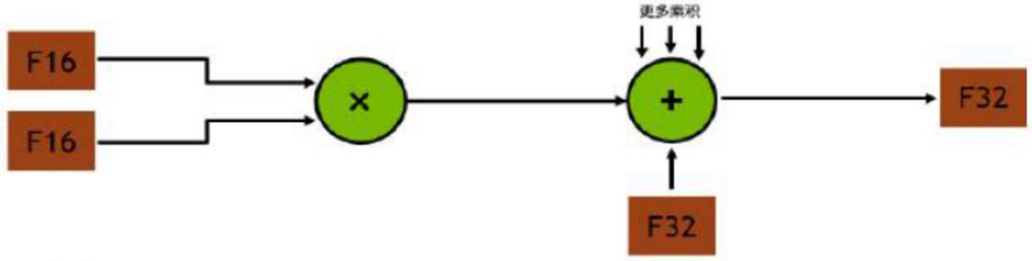


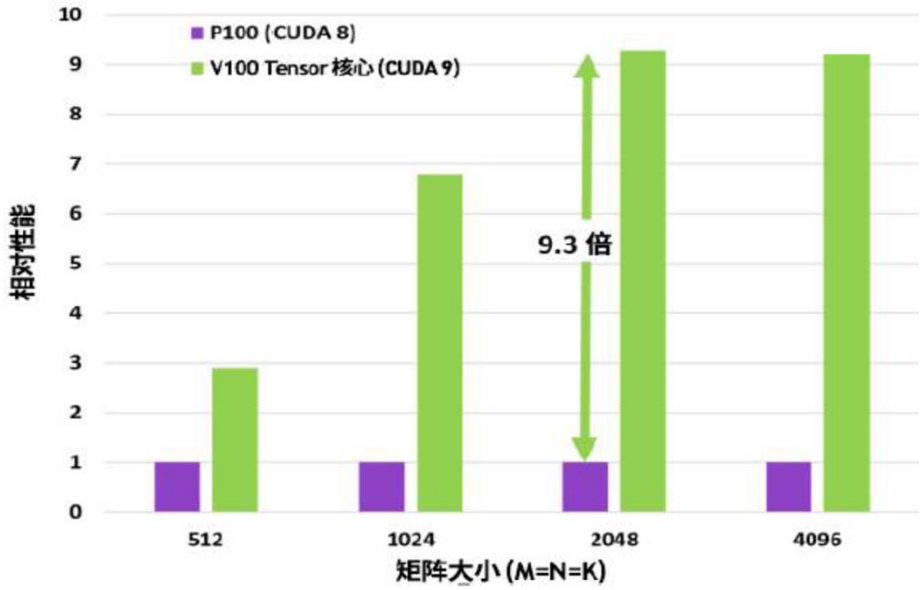
Nvidia GPU面向深度学习的加速机制

- Tensor core
- 半精度/混合精度计算

浮点数类型	位数	取值范围	U=2^{-t}
半精度 (Half precision)	16 bits	$10^{\pm5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
单精度 (Single precision)	32 bits	$10^{\pm38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
双精度 (Double precision)	64 bits	$10^{\pm308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
四精度 (Quad precision)	128 bits	$10^{\pm4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

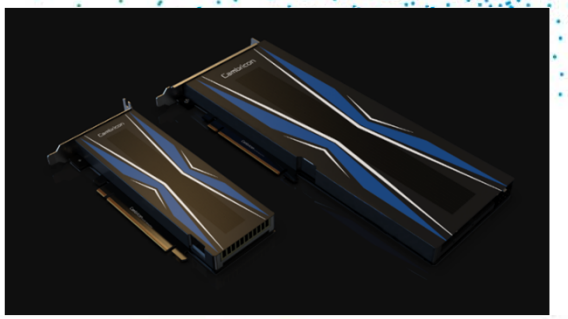






寒武纪MLU智能云服务器芯片/处理卡

- 半精度浮点和8位整数计算
- 稀疏神经网络性能优化
- 高能效



产品型号	MLU100-C	MLU100-D
核心架构	Cambricon MLUv01	
核心频率	1 GHz	
半精度浮点运算速度（FP16）	16 TFLOPS（关闭稀疏模式时峰值）	
	64 TFLOPS（打开稀疏模式时峰值）	
整数运算速度（INT8）	32 TOPS（关闭稀疏模式时峰值）	
	128 TOPS（打开稀疏模式时峰值）	
内存容量	8GB/16GB	
内存位宽	256-bit	
内存带宽	102.4 GB/s	
系统接口	PCI Express 3.0 x16	
外形	全高全长，单slot	半高半长，单slot
是否支持解码	支持解码	不支持解码
典型/最大功耗	110W	75W
ECC保护	是	
散热方式	被动散热	

CAICT 中国信通院

TRUCS2019

TRUSTED CLOUD SUMMIT

可 信 云 大 会

中国·北京　2019.7.2-3

THANKS