中国的高性能计算：挑战与进展

钱德沛
北京航空航天大学/中山大学
开放科学计算联盟学术年会
厦门，2019年11月30日

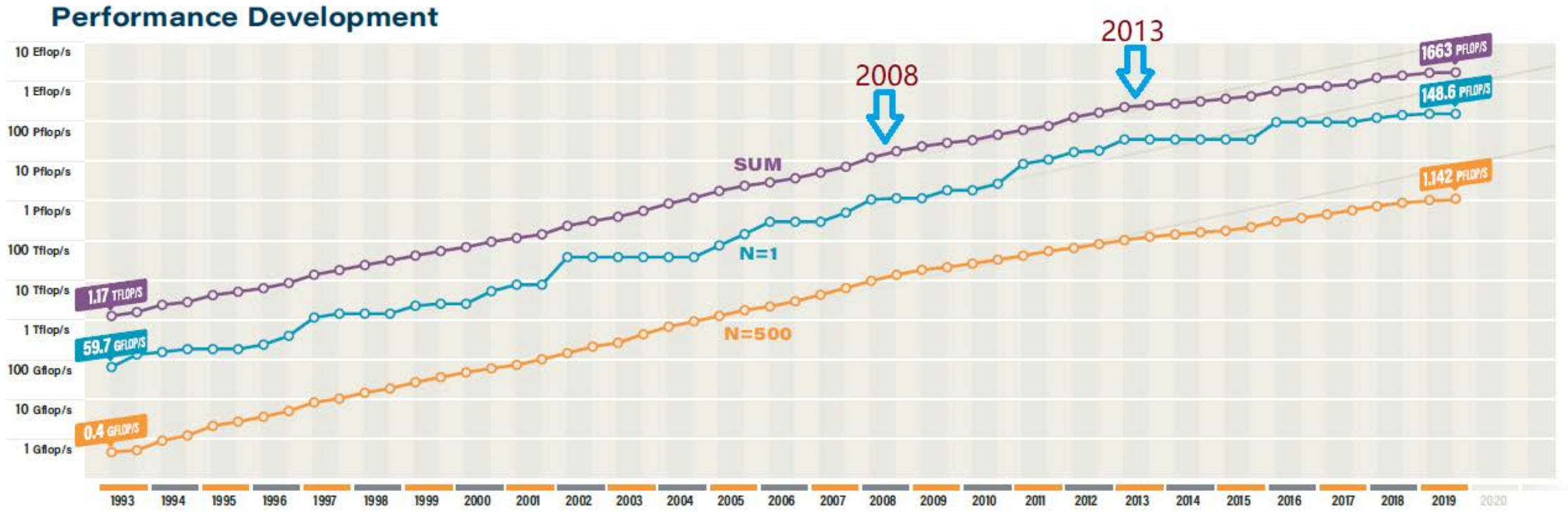
汇报内容

- 超级计算机发展的新动向
- 值得重视的几个问题
- 高性能计算重点专项进展

超级计算机发展的新动向

超级计算机发展遇到瓶颈

- 1993-2012年，超级计算机的性能以每10年提高1000倍的速率提高
- 从新的TOP500曲线看，从2013年起，上升速率变缓。2019年11月TOP500前十没有发生变化
- 如果没有大的突破，可能降低为每10年100倍左右



超级计算机发展遇到瓶颈（续）

- TOP500的数据说明超级计算机的发展遇到瓶颈，特别是
  - 能效指标的约束
  - 摩尔定律接近失效
  - 体系结构变化缓慢
  - 尚无颠覆性技术出现
  - 新原理器件缺少突破

超级计算领域竞争更趋激烈：美国

- 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，投入18亿美元，研制3台E级计算机，另外18亿美元研发应用
  - 持续性能1 EPF的A21在2021年上半年完成
  - 持续性能1EF的Frontier在2021-2022完成
  - Serra的后续E级将在2023完成，4-5EF
- 2024年达到8-12EF性能

US Exascale Systems

Computer Name	A21	Frontier (OLCF5)	El Capitan (ATS-4)	NERSC-10	NSF Exascale Phase 2
Overview	1st US Exascale System	Summit Follow-On	Sierra Follow-on	NERSC-9 Follow-On	TBD
Location	ANL	ORNL	LLNL	NERSC	TBD
Planned Delivery Date/ Estimated	2021	2021-2022	2023	2024	2024
Early Operation	2022	2023	2024	2025	2025
Planned/Est. Performance Pflops*	1,000	1500-3000	4000-5000	8000-12000	10X Phase One
Linpack/Peak Performance Ratio	70-80% (est.)	60-70% (est.)	50-60% (est.)	50-60% (est.)	TDB
Linpack Performance	700-800	900-2100	2000-3000	2000-3000	TDB
GF/Watt	40	60-100	134-200	266-480	TBD
Linpack GF/Watt	23.3-32.0	36-70	67-120	133-288	TBD

超级计算领域竞争更趋激烈：日本

- 日本的Fugaku（POST-K）全机将在2021年初完成
  - 性能100倍于K-Computer
  - 基于ARM处理器实现，方便软件的开发、移植
  - 新一代ARM处理器已经研制成功，扩展了512位的向量部件，能效指标高
  - **系统内存采用HBM2，内存带宽1TB/s，内容容量大（内存字节/Flops=0.4，太湖之光约0.01）**
  - 系统软件同步研发，研发了支持新处理器的微内核操作系统
  - 2019年11月，Fugaku的原型在Green500中排名第一（16.9GF/w），证明基于众核处理器的系统能效可以超过基于GPU的异构加速系统
  - 2020年上半年Fugaku系统提供试用，2021年初完成全规模系统
  - 日本还有后续E级系统
- NEC坚持向量路线，研发了Aurora Vector Engine处理器，以此研制超级计算机

超级计算领域竞争更趋激烈：欧盟

- 欧盟在2023年左右建立E级计算基础设施（3台左右）
  - 在目前的PRACE基础上发展
  - E级系统强调低功耗
- 欧洲处理器的研发策略
  - 自研欧洲处理器，Atos公司牵头
  - 非常重视开源处理器架构RISC-V，在欧盟支持下，依托巴塞罗那超算建立欧洲开放计算机体系结构实验室（LOCA）
- 欧洲高性能计算基础研究和应用基础好
  - 新的计算模型、语言、算法
  - 大规模数值模拟

超级计算领域竞争更趋激烈：中国

- 十三五重点研发专项“高性能计算”
- 专项目标
  - 突破E级计算机核心技术，依托自主可控技术，研制适应应用需求的E级（百亿亿次级）高性能计算机系统，使我国高性能计算机的性能在“十三五”末保持世界领先水平。
  - 研发一批关键领域/行业的高性能计算应用软件，建立国家级高性能计算应用软件中心，构建高性能计算应用生态环境。
  - 建立具有世界一流资源能力和服务水平的国家高性能计算环境，促进我国计算服务业发展。

值得重视的几个问题

我国E级机的指标

- 依托自主技术，研制成功E级高性能计算机，系统达到如下技术指标：
  - 系统峰值性能达到E级
  - 内存容量10PB，存储容量可支持EB级
  - 系统能效比达到30GFlops/W
  - 高速互连网络传输性能大于500Gbps，可扩展性好
  - 高效的大规模系统资源管理与调度系统
  - 方便易用的并行编程模型和开发环境
  - 全系统监控管理与容错机制
  - 高效支持大规模应用的可靠可扩展运行

E级机研制面临重大技术挑战

- E级计算机的研制面临重大技术挑战
  - 功耗(power)
    - EFlops/20MW (50GF/W)，还没有有效的技术途径达到
  - 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)
    - 巨大的系统规模使得系统的平均无故障时间大大缩短，甚至一小时以下
    - 如何完成长时间不间断运行的应用？
- 应对这些挑战需要
  - 体系结构的创新
  - 关键技术的突破
  - 软件硬件的协同

存在卡脖子技术

- 中美关系发生变化，美国遏制中国的思维占上风
  - 2015年对国防科大及相关超算中心禁运
  - 2018年对中兴公司全面禁运
  - 2019年5月贸易战升级，华为公司列入“实体名单”
  - 2019年6月将曙光及相关子公司、江南所列入“实体名单”
- E 级机研制存在瓶颈技术
  - 高性能处理器/加速器
  - 内存芯片，特别是3D内存
  - 新型存储系统/器件，NVM
  - 高速互连网，光传输和交换器件
  - IC设计EDA软件
  - 先进的芯片制造工艺
  - 工程计算软件

要有底线思维

- 美国已经把中国三个超级计算机研制单位列入“实体名单”，实施禁运和封锁
- 如何在外部限制甚至封锁条件下保持我国的超级计算的持续发展?
- 在当前的国际形势下，自主可控不是可选项，而是唯一出路
- 依托自主可控和开放合作并不矛盾，只有自身强，才有合作的基础

要特别重视体系结构研究

- 摩尔定律渐近尽头，单靠主频提高、工艺改善就能获得性能增益的路走到头了
- 国际上提出体系结构的“寒武纪爆发”，体系结构将迎来黄金十年，虽有夸张，也不无道理
  - 能否出现“百花齐放、百家争鸣”的局面，类似于上一世纪80年代并行计算机的发展
  - 能否从以规模取胜的庞大系统，向灵巧、节能、应用高效的系统进化
- 体系结构的基本问题
  - 冯.诺伊曼结构的基本特点：存储程序，存储器是关键通路，程序决定执行次序，如何适应大规模并行执行
  - 问题分解、竞争冲突消解、通信与同步、存储一致性模型、激进与保守执行、投机执行…
  - 体系结构与计算模型的匹配
  - 计算与访存的匹配
- 没有一种体系结构能够覆盖所有应用的需求
  - 通用 vs 专用是长期争论的问题，未来是否会出现多样化、灵巧化、专用化的局面?

对现有流行体系结构的改进

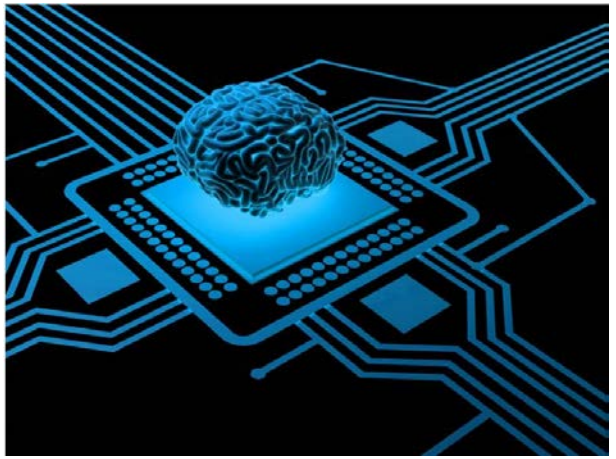
- 当前流行体系结构
  - **片内异构**。基于众核，（神威太湖之光）
    - 一个主核带众多计算小核
    - 与核数相比，内存容量偏小
    - 每核的内存带宽偏小
  - **节点内异构**。CPU+加速器结构，（天河二号）
    - 一个节点上一个CPU+一个或几个加速器，协处理器模式
    - CPU与加速器之间、加速器之间的数据传输是瓶颈
    - 对某些应用而言，加速器的利用率不高，不能充分利用
- 对现行体系结构的改进
  - **系统级异构**。软件定义的多态
  - CPU和加速器同等地位由互连网连接
  - CPU-CPU, CPU-加速器, 加速器-加速器之间直接通信
  - 软件定义系统配置：CPU only, accelerator only, CPU+accelerator, 不同组态在系统内共存，按需配置资源
  - 需要在研发过程中和实际使用中更细致地评价

追求计算与存储的匹配

- 计算和访存的匹配，减少数据访问和传输
  - 流式结构
    - 在数据流动中完成处理，减少内存的存取
  - 数据流结构
    - 发掘和利用应用中数据的内在并行性
    - 靠数据的可用来激发操作，但是目前的体系结构不能高效支持
  - 处理器和内存尽可能靠近
    - 处理器内置内存，提高访存带宽，降低时延，
    - 内存具有一定处理能力，就地完成一些操作

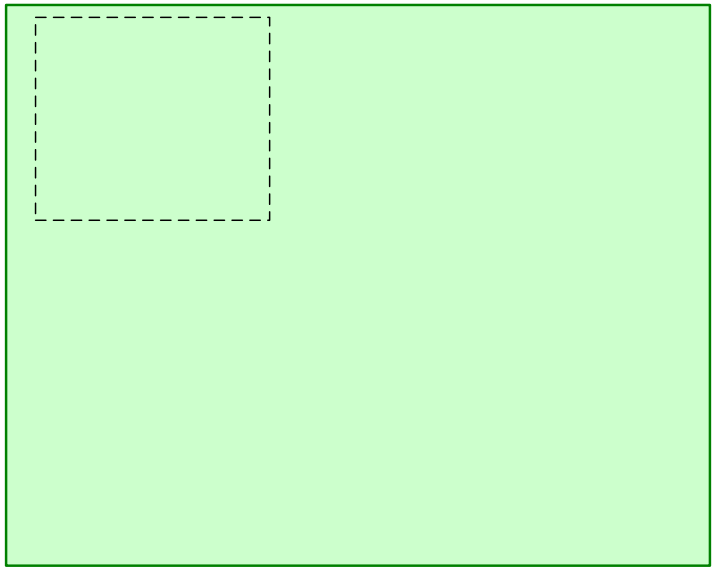
探索创新体系结构

- 新的体系结构
  - 面向领域的体系结构DSA
  - 深度可重构的柔性体系结构
  - 片内异构，集成高效的专用部件
    - 瑞士军刀 vs 专用工具集
  - 融合支持应用特征的专用部件
    - 卷积神经网络、脉冲神经网络，图计算
- 一个梦想：与制造晶体管一样方便地设计和制造处理器
  - 根据应用特点快捷地设计机器
  - 需要设计软件和芯片制造工艺流程的支持
    - 高层逻辑描述综合/系统级/寄存器传输级/硅片编译
    - 低成本、快响应的流片工艺



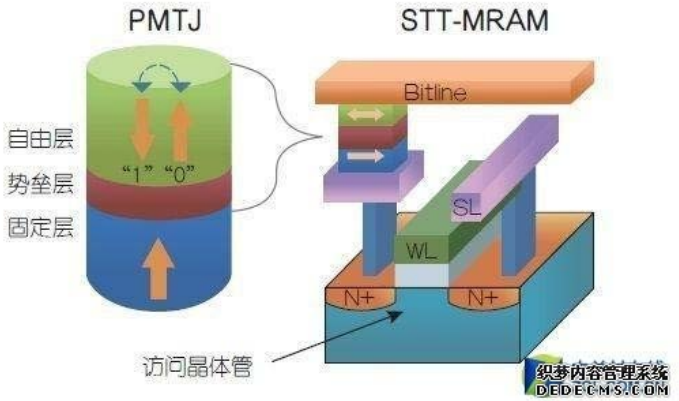
外部封锁条件下处理器的发展之路

- 半导体工艺趋近极限，限制处理器能效的提升，受功耗限制，靠提高主频的办法提高性能已不可能，在新原理器件出现之前，并行是唯一出路。
- 通用处理器
  - 降低核的复杂性来提高核数
  - 提升向量部件性能
  - 改进片上cache（容量和命中率）
  - 流式处理减少数据访问
  - 提升核间互连性能
  - 提高访存能力
  - 混合字长支持
- 专用加速器
  - 人工智能、大数据
- 片内异构
  - 片内多种专用部件，需要时激活
- 在国外目前围堵情况下中国处理器的路怎么走？
  - 中国处理器+RISC-V?
  - 能形成共识和合力吗？



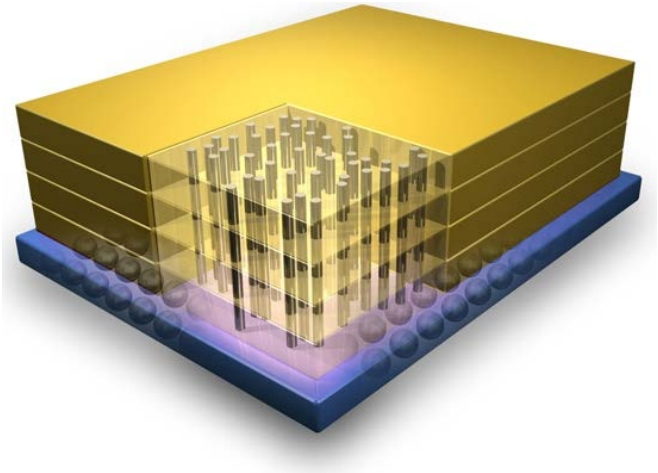
要特别重视内存系统

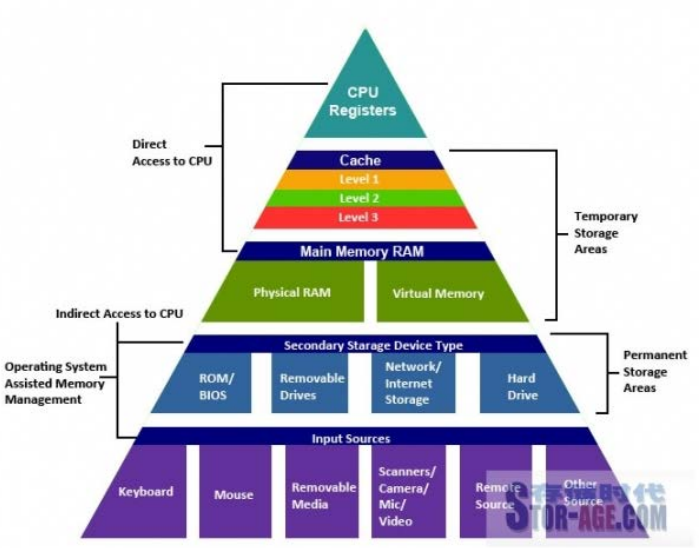
- 访存是计算机系统的性能瓶颈
- 访存成为系统能源消耗的主要来源
- 追求大容量、高带宽、低时延、低功耗
- DRAM+NVM的混合内存
  - NVM
    - 不需刷新，节能
    - 密度高，可实现大容量
    - 读取快，能耗低
    - 写入慢，寿命有限
  - 易失于非易失器件结合，既提高容量和能效，又缓解高写入开销和有限写次数问题
    - 只读数据在NVM，读写数据在DRAM
    - 按数据访问性质DRAM和NVM间切换



要特别重视内存系统

- 处理器和内存尽量靠近，缩短传输距离，提高传输带宽，降低访存时延
  - 3D内存芯片缓解访存墙
  - 处理器内置DRAM
  - 具有计算处理能力的内存
    - 哪些处理操作在内存完成?
- 提高数据复用，减少访存
  - 存储层次结构，cache结构和一致性协议
    - 适应NVM混合内存的特点
    - 适应异构加速系统的特点
- 适应异构的内存一致性，发掘异构系统的并行执行潜力，方便并行编程





要全面应对异构带来的挑战

- 异构系统影响多个层面
  - 问题的分解
  - 算法设计优化
  - 并行编程模型和语言
  - 并行软件实现
  - 资源管理
  - 任务调度
- 要从模型、语言、编译、库、调试、操作系统/运行时、资源管理、程序开发优化等多个层面提供支撑手段和工具

要构建高性能计算生态环境

- 构建我国高性能计算生态环境的任务十分紧迫
  - 要尽快围绕基于国产处理器的系统，研发系统软件、工具软件、应用软件，建立国产处理器应用的生态环境
    - 操作系统、语言、编译器
    - 调试器、性能优化器、能耗调优器
    - 应用软件开发环境
    - 应用软件开发
  - 特别要加强替代主流商业软件的自主应用软件的研发
  - 要尽快通过技术辐射，形成有一定市场份额的国产服务器系列，促使更多人有兴趣为其开发软件
  - 软件开发要走开源的路，让更多人为提高自主软件成熟度出力，“**高手在民间！**”

超算和人工智能、大数据的融合发展

- 超算、大数据、人工智能密切关联，相互支撑
  - 超算是大数据分析和基于深度学习的人工智能技术与应用的基础
  - 超算和大数据改变了人工智能研究和应用的方式
- 大数据和人工智能将深刻影响未来超级计算机体系结构和实现技术
  - 各类智能加速部件
  - 数据为中心的体系结构
  - 数据流体系结构
  - 神经态计算
- 人工智能会改变传统计算问题的求解方法
  - **Linpack-AI**，使用混合字长和改进算法获得同样问题3倍性能提高，对解决传统计算问题方法的启示（软件所孙家昶老师团队正在研究、实践）
  - 科学发现需要人工智能的帮助
- 超算、大数据、人工智能要协调、融合发展

高性能计算重点专项进展

已经部署的研究任务

	E级计算机系统研制	高性能计算应用软件研发	高性能计算环境研发
基础前沿	高性能互连 计算、编程及运行模型	E级计算的可计算建模与新型计算方法 面向E级计算的并行算法库	计算服务化模型及体系架构 虚拟数据空间
共性关键技术	E级机验证原型 E级计算机系统	并行编程框架 应用协同开发优化平台与工具	国家高性能计算环境服务化机制与支撑体系研究
应用示范		数值装置 领域应用软件	基于高性能计算环境的服务系统（集成业务平台、领域应用社区、HPC教育实践平台）

E级计算机系统研制

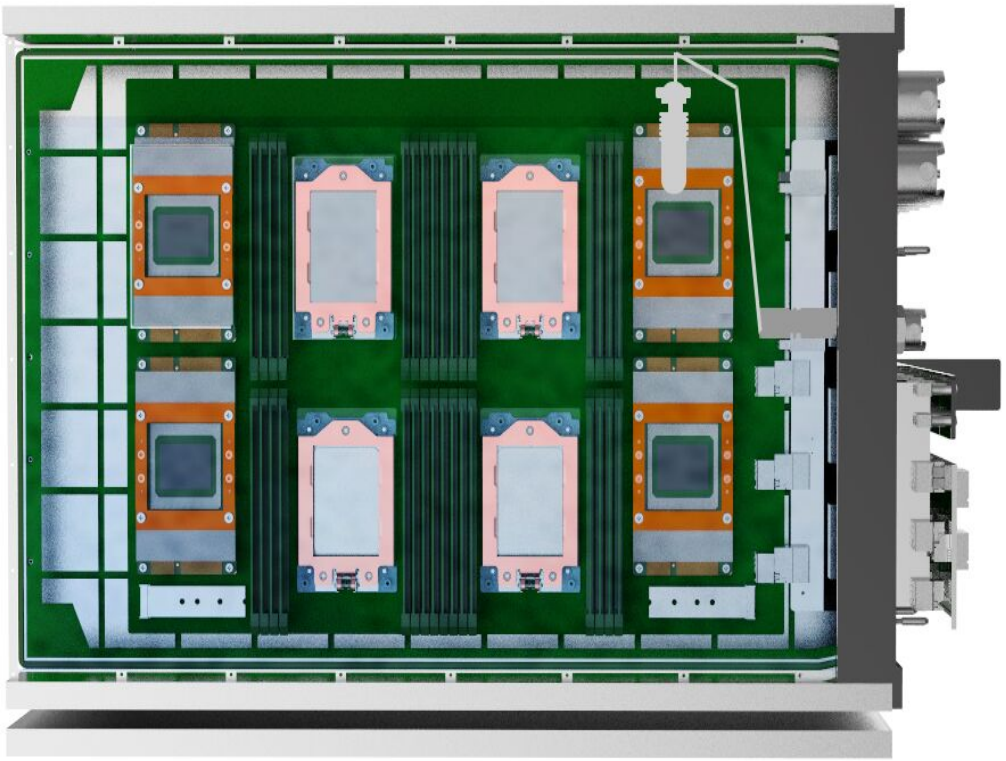
- 基础前沿研究
  - 面向E级计算的高性能互连
  - 新型高效能计算、编程和运行模型
- 共性关键技术研究
  - 总体技术与评测技术研究
  - E级计算机验证原型研制
  - E级计算机系统研制

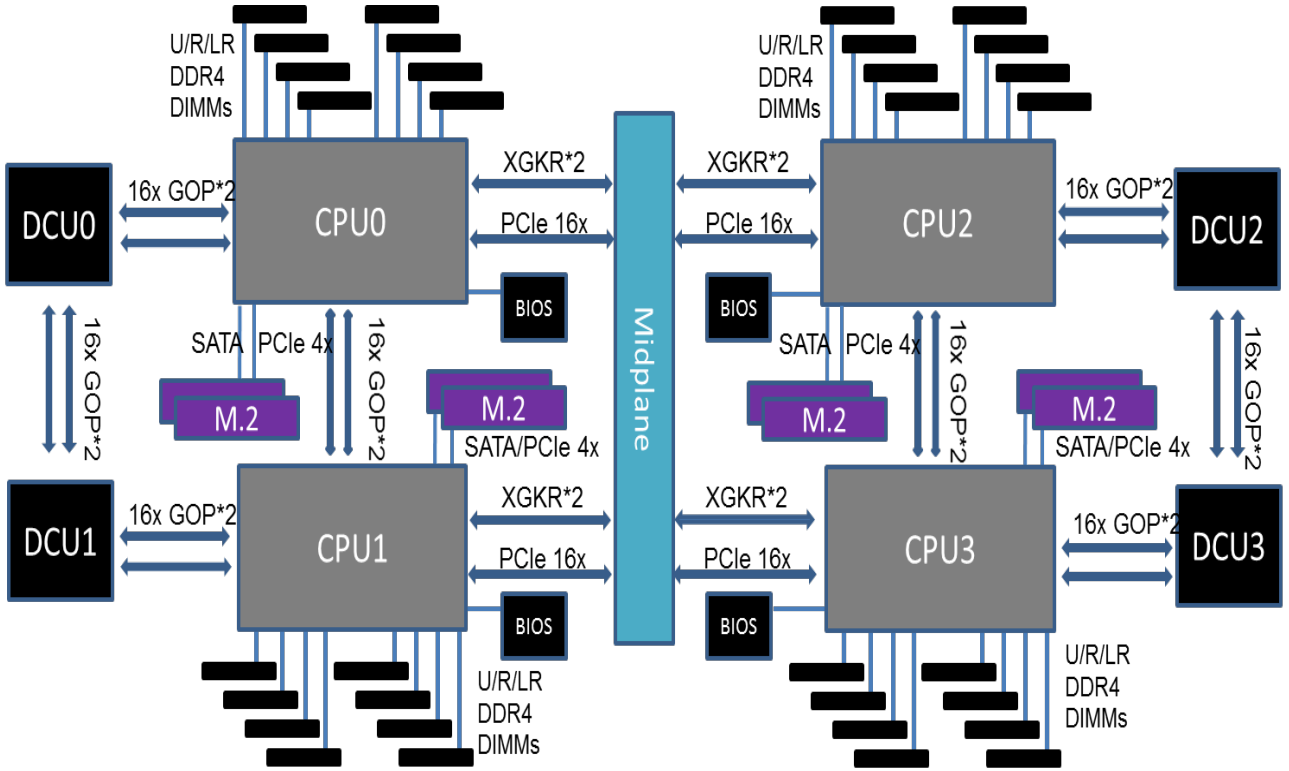
曙光E级原型系统

- 加速体系结构（**节点内异构**）
  - 采用 x86 处理器和加速器
  - 保护现有软件资产
- 512 节点，1024 海光x86处理器，512 海光DCU加速器
- 6D Tours互连网，200Gbps/node
- 峰值性能：3.18PFlops，Linpack性能：2.274PFlops，效率71.5%

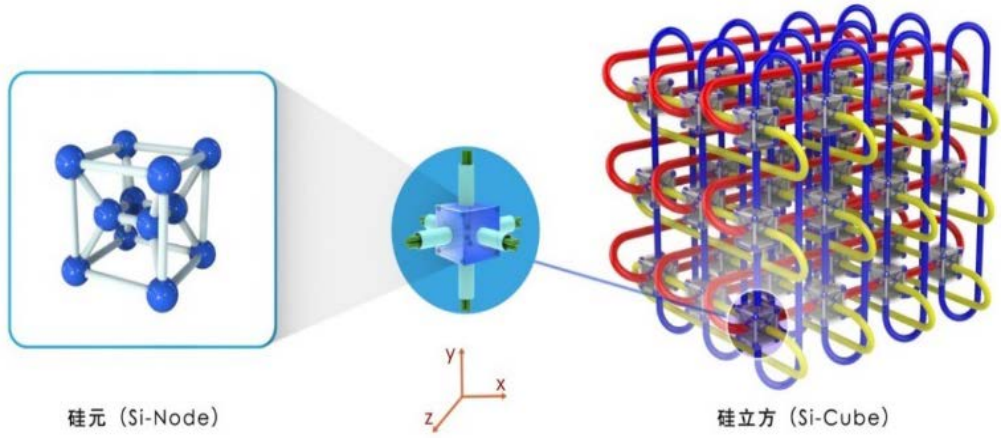
曙光E级原型计算节点

- 2 CPU + 2 DCU，通过GOP高速总线互连
- 内存带宽：2667 Mbps, DDR4
- 内存容量： ≥128G DDR4
- 互连：200Gbps





曙光E级原型互连与冷却

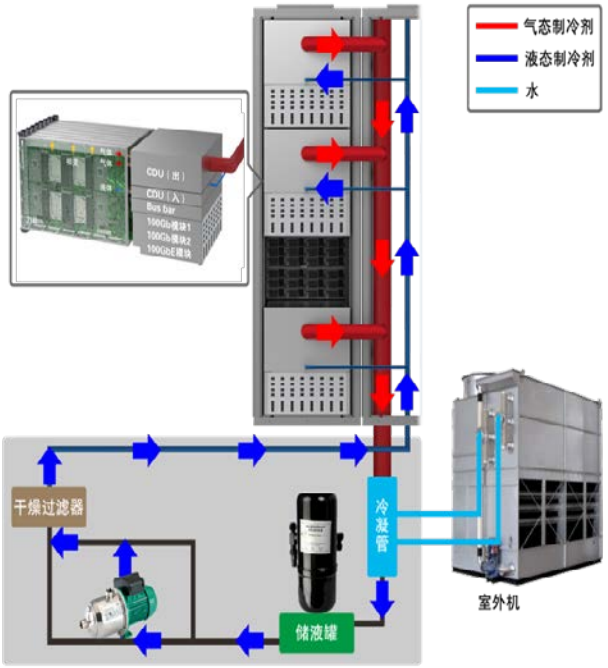


6D-Torus高速互连网

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互联
- **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

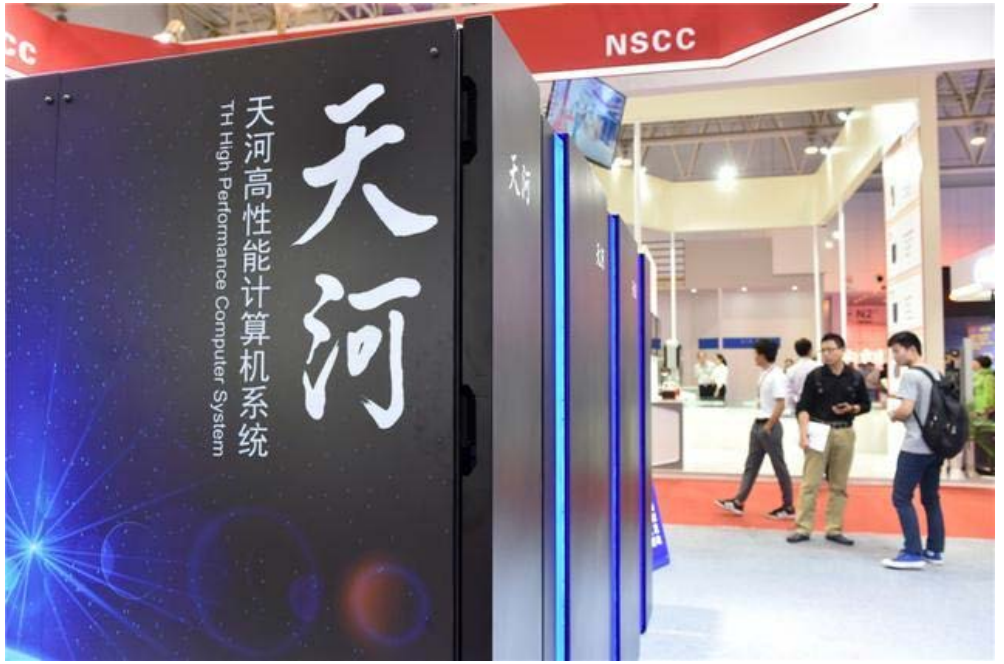
Imm058是一种物理兼容性和化学稳定性优良、绝缘性好、**无闪点**、不燃烧、无毒无害的氟化物，在常压下的沸点为**50℃**左右，兼容性好，服务器的各种元器件均可以在其中**长期**稳定运行。

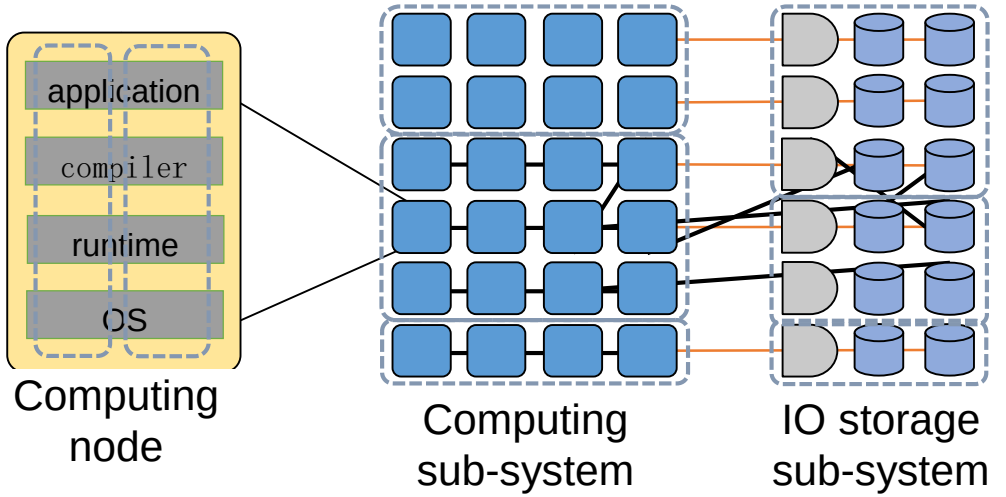


高效浸没式冷却

天河E级原型验证系统

- 满足不同应用需求的柔性体系结构（**系统级异构**）
  - 所有处理器由互连网对等连接
  - 软件定义系统的组态
    - CPU-only, 加速器-only, CPU+加速器
- 基于128核迈创处理器
- 高速互连
- 512节点
  - 峰值性能3.14PF
  - Linpack效率：78.5%
- 安装在国家超算天津中心，投入运行

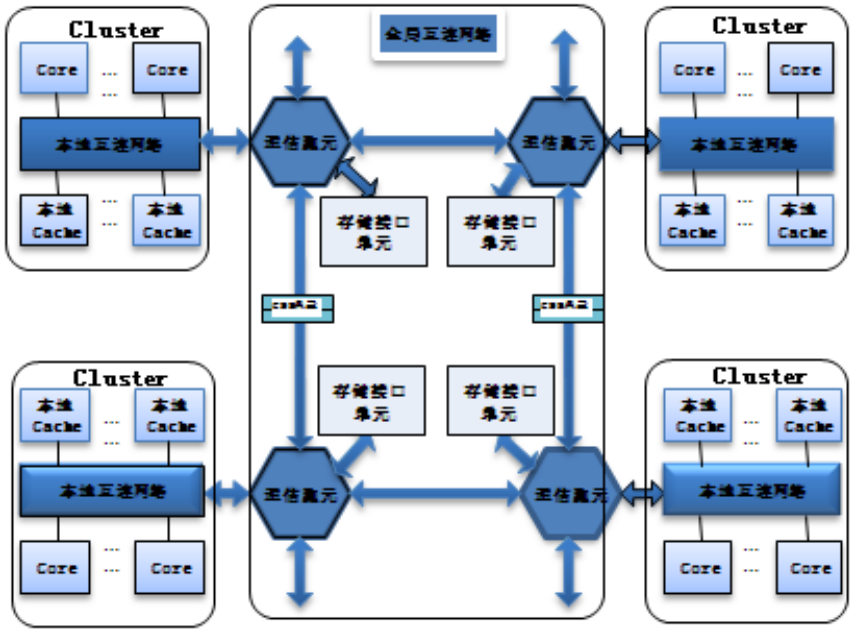




天河E级原型节点

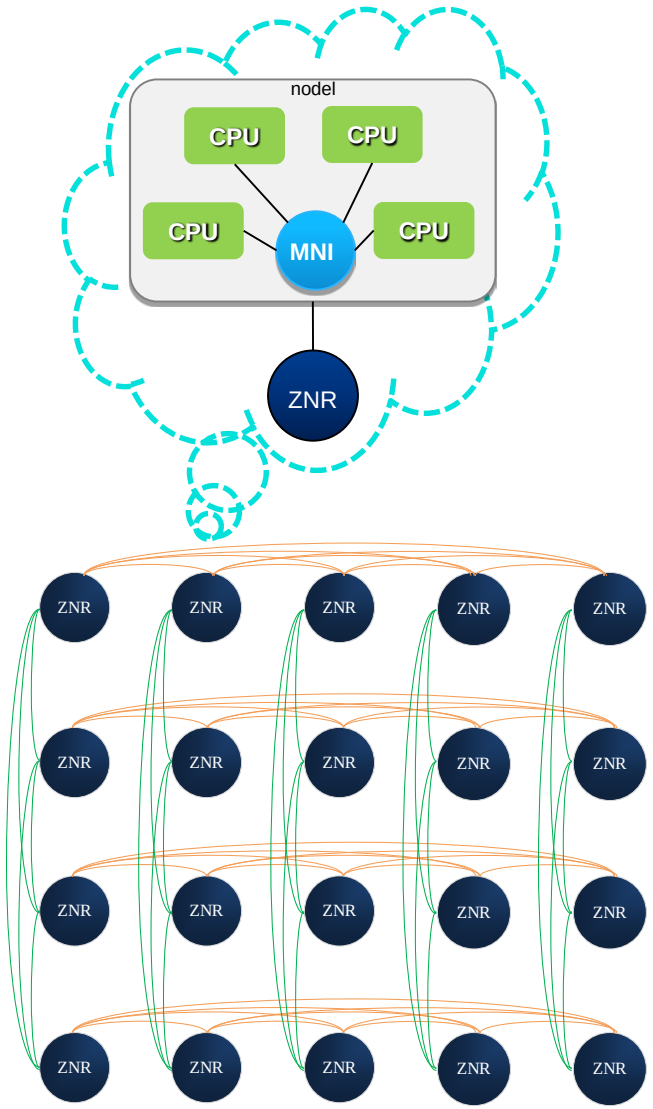
- 众核处理器MT-2000+
  - 基于MT-2000做了能效优化
    - 流水线、存储系统优化等
  - 16nm FinFET工艺
  - 2.0 GHz核心工作频率
  - 128核，峰值2.048Tflops
  - 典型功耗130W
  - 能效比达到15Gflops/W以上

- 计算节点
  - 三个MT-2000+/节点
  - 节点性能：>6TFlops



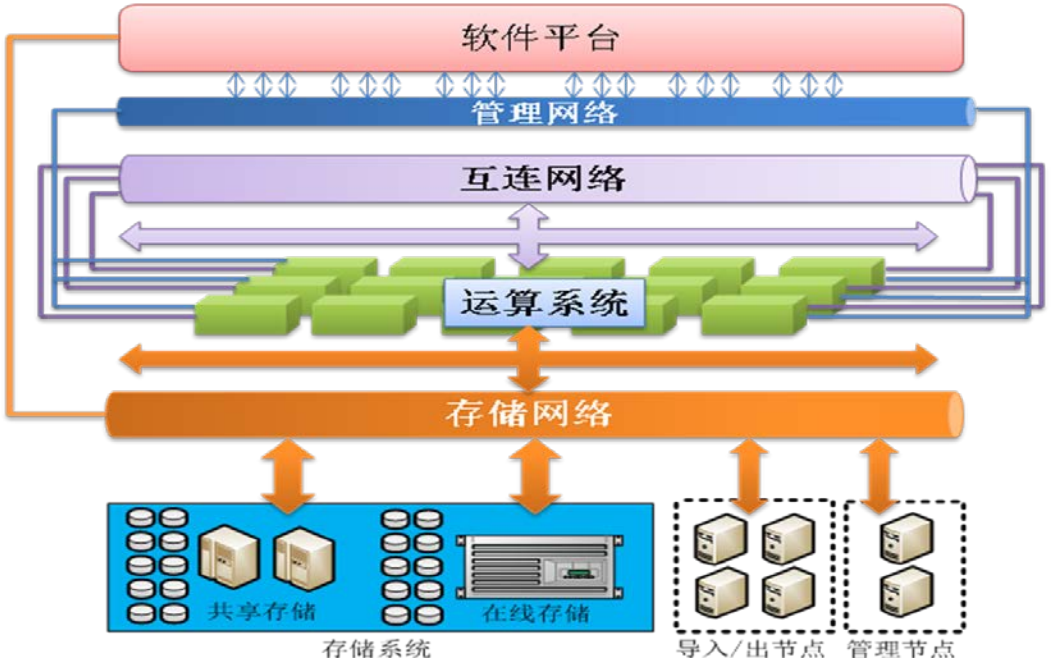
天河E级原型互连网

- 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级原型验证系统

- 面向多目标优化的多态多尺度自适应体系结构（**片内异构**）
  - 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
  - 总计算性能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
- 部署在国家超算济南中心，投入使用



神威E级原型节点

- 节点
  - 两个SW26010处理器
  - 4路DDR4内存
  - 峰值性能：6.12TFlops
  - 节点能效：11GFlops/W

Node (2 CPU)

Node board(8 CPU+4 NI)

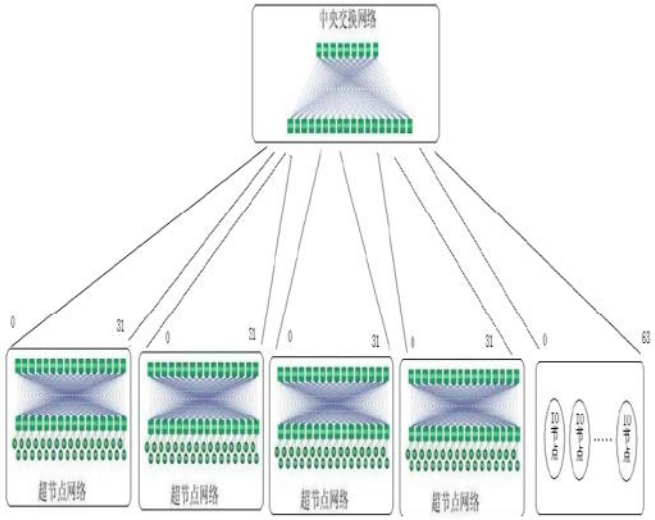
- 超节点
  - 256节点，256X256全连接
  - 单点上网：2路25Gbps X4
  - 点对点单向带宽：200Gbps

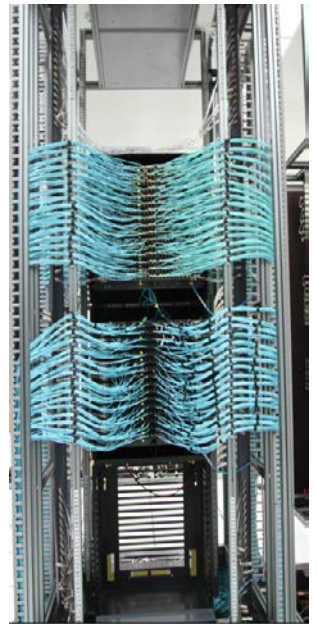
Supernode (32 boards)

Cabinet (whole prototype)

神威E级原型互连网

- 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

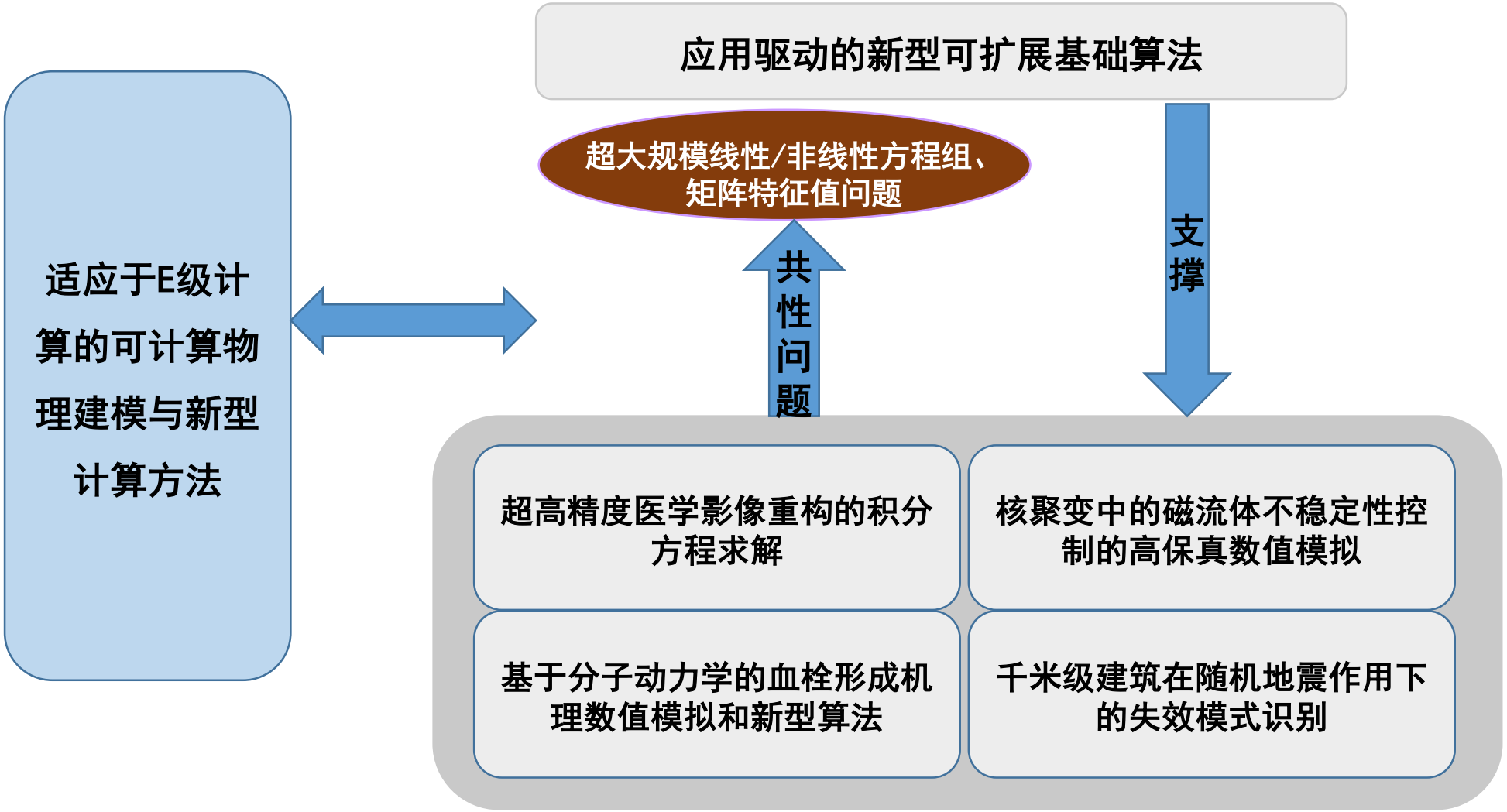




高性能计算应用软件研发

- 基础前沿
  - E级计算的可计算建模与新型计算方法
  - 面向E级计算的并行算法库
- 共性关键技术
  - 并行编程框架
  - 应用协同开发优化平台与工具
- 应用示范
  - 4个数值装置
  - 14个领域应用软件

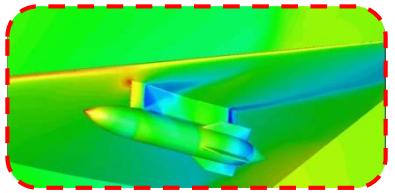
E级计算的可计算建模与新型计算方法

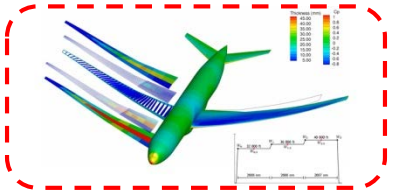


数值装置：数值飞行器


大型飞机


先进战斗机


复杂多体分离


气动/结构综合优化

牵引

100P级超级计算机

依托

数值飞行器原型系统

突破

- 超大规模网格生成技术
- 异构并行计算方法
- 非线性流固耦合计算方法
- 气动/结构综合优化算法

目标

自主知识产权

- 空气动力学数值模拟软件
- 结构强度力学分析软件
- 非线性流固耦合数值模拟软件
- 气动/结构综合优化设计软件

意义

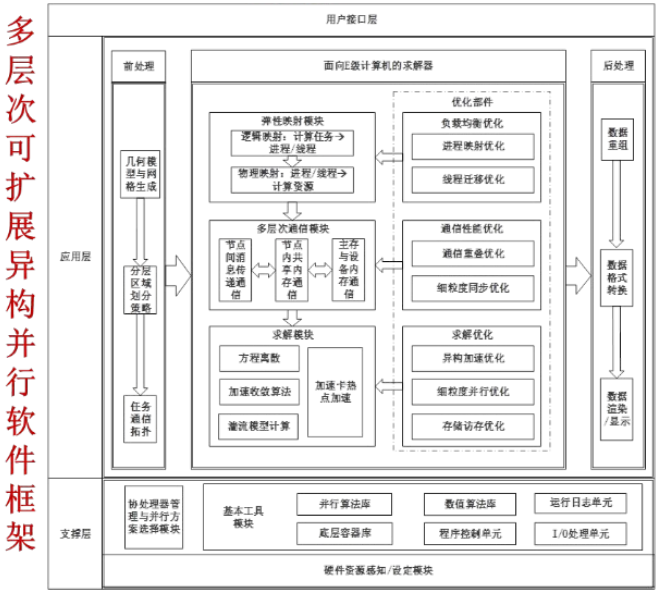
- 提升数值模拟技术工程应用水平
- 挖掘高性能计算机的应用潜能
- 提供高性能计算机持续发展的技术支撑

大型流体机械并行计算软件

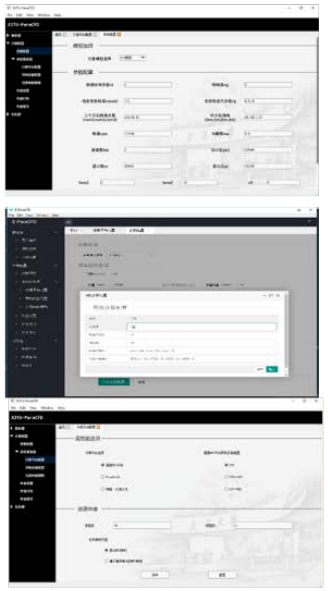
- 研发了多层次可扩展异构并行软件
- 创新：设计了高可扩展并行CFD软件框架，在国产机器上实现了三套实例

- 建立了国内首个10万等级空分主压缩机全尺寸性能测试台位，完成了全速全压全负荷气动性能试验
- 开展了10万等级空分主压缩机的多排单叶道混合平面法的定常计算，轴流段多变效率87.6%，离心段多变效率86.4%

一个框架



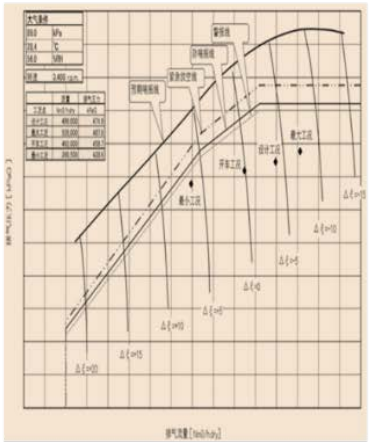
三套实例



性能测试方法

- JB/T3165《离心和轴流式鼓风机和压缩机热力性能试验》标准中的一类试验要求。
- 通过测试分别获得了10万空分压缩机的段及整机性能曲线。

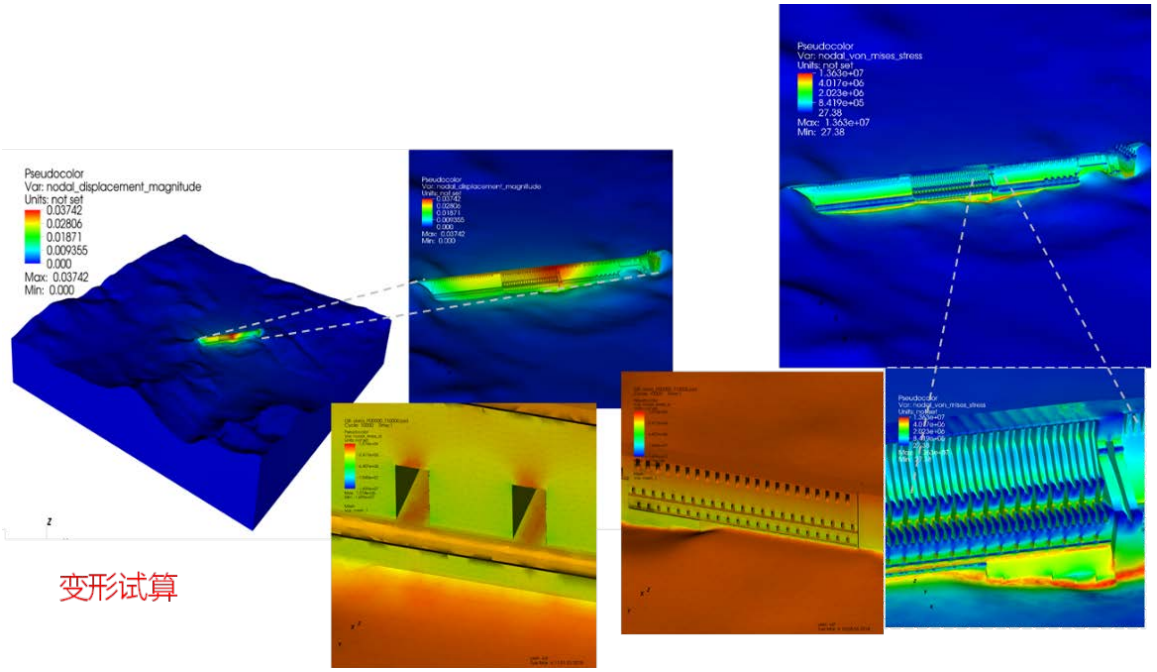
性能测试结果

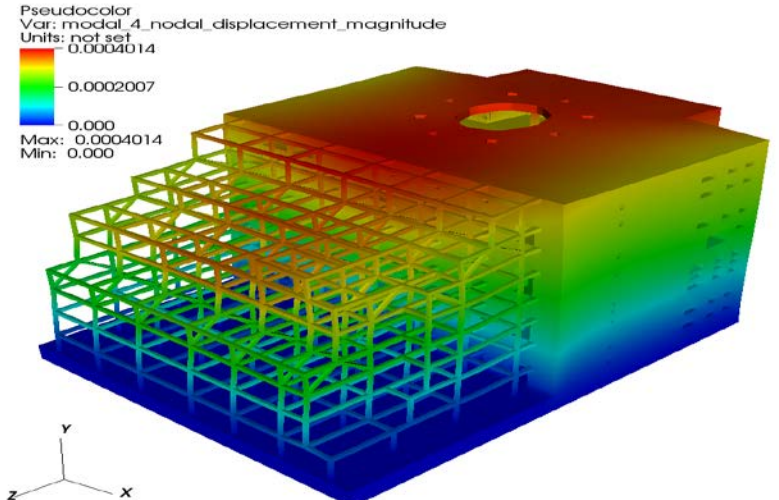


成果鉴定会

复杂工程力学高性能应用软件

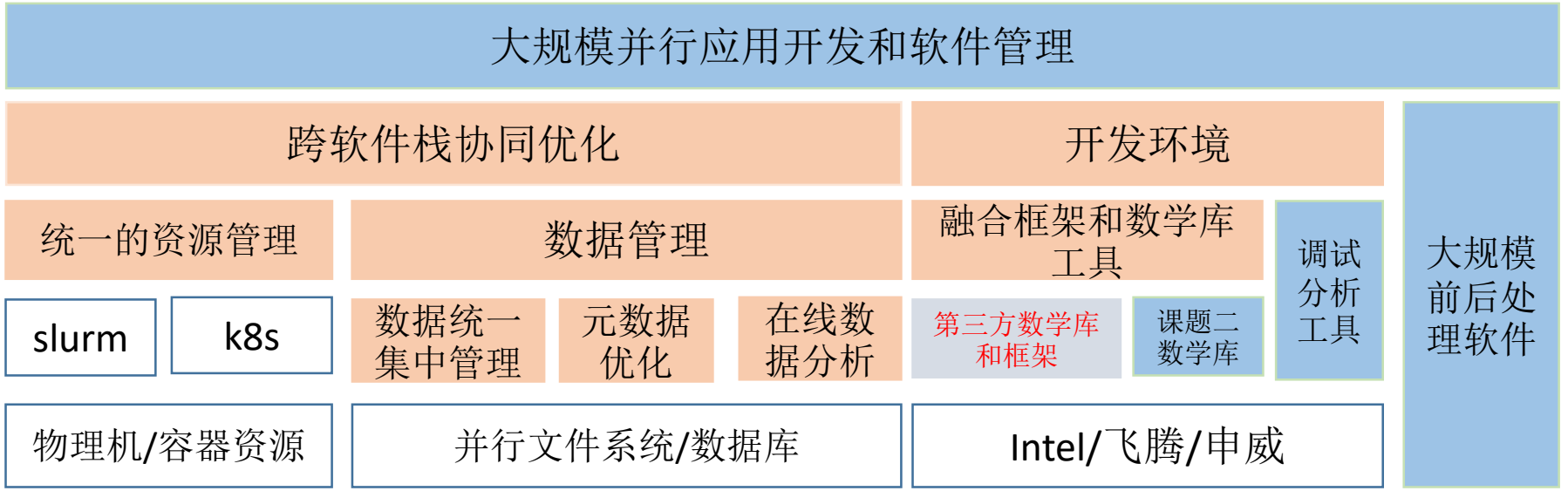
- 发展了高精度应力单元算法
- 发展了面向复杂几何的“100亿单元”非结构网格高可扩展建模技术和复杂构造建模的块体切割与重构技术，并应用于工程实践
- 示范应用
  - 首次完成了三峡大坝101亿网格、50亿自由度规模的结构静力计算
  - 完成了神光III光机4.2亿自由度地震破坏模拟计算。为工程分析和评价提供支撑





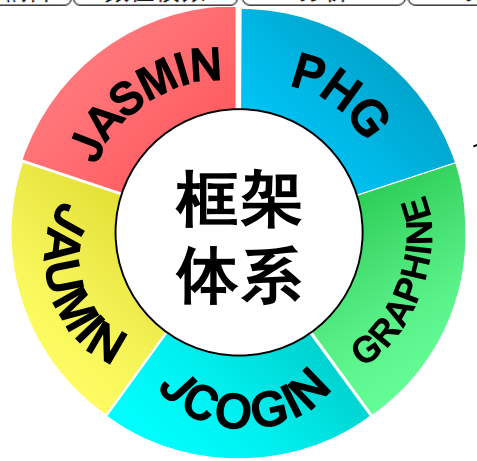
应用软件协同开发工具与环境

- 国产超算开发工具+融合框架和数学库的开发环境
- 跨软件栈的综合优化软件
- 大规模前后处理可视化工具
- 性能与能效调优工具
- 大规模并行应用软件资源库



并行编程框架

辐射流体耦合中子输运	流体力学耦合粒子输运	多介质弹塑性流体力学	ICF二维总体程序	激光等离子体粒子模拟	激光等离子体流体力学
等离子体模拟	关联电子体系模拟	分子动力学模拟	位错动力学模拟	流体力学界面不稳定性	三维蒙卡辐射输运模拟
冲击动力学粒子模拟	三维分子动力学	超高速碰撞动力学模拟	平台级时域电磁模拟	平台级频域电磁模拟	器件级全电磁粒子模拟
弹塑性流体力学模拟	光滑粒子动力学模拟	欧拉流体动力学	三维ALE流体力学	结构网格并行Sn软件	三维静态蒙卡粒子输运
全球大气环流模式	全球海洋环流模式	全球海冰模式	冲击动力学分析	非结构网格粒子输运	二维动态蒙卡粒子输运
区域数值天气预报模式	地球系统高分辨率耦合	地下水流动数值模拟	结构静力学分析	模态与振动分析	中子光子耦合粒子输运



数值模拟的网格规模达千亿、粒子数达万亿、自由度达数万亿。

200万处理器核
并行效率达30%

应用成果

原始创新

基础创新

高效能实现关键技术体系
超级并行应用软件研制方法

高性能计算环境研发

- 基础前沿研究
  - 计算服务化模型及体系架构
  - 虚拟数据空间
- 共性关键技术研究
  - 国家高性能计算环境服务化机制与支撑体系研究
- 应用示范
  - 基于高性能计算环境的服务系统
    - 集成业务平台
    - 领域应用社区
    - HPC教育实践平台

国家高性能计算环境

双运行中心（北京/合肥）

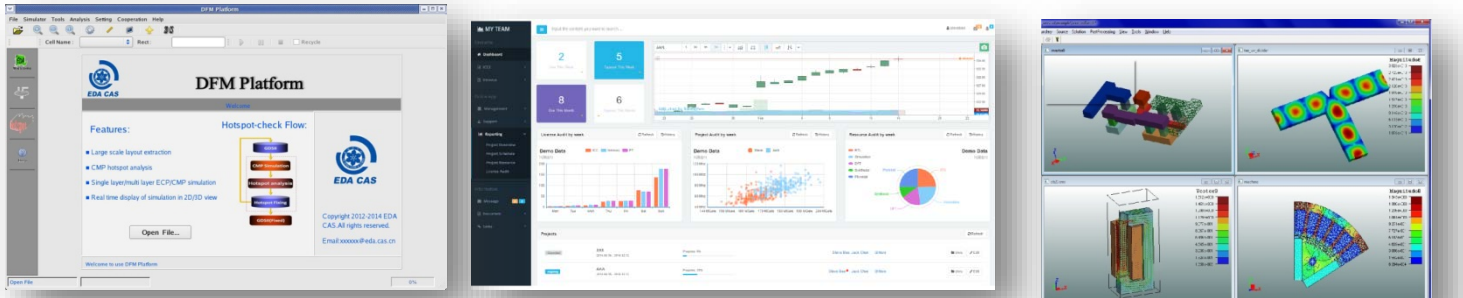
19个结点（200PF+162PB）

互联带宽1000Mb
（北京/合肥/无锡/广州/上海）

基于微服务结构的计算门户
基于应用的全局调度与预测

《资源评价标准白皮书》
《环境综合评价指数》

基于HPC的EDA平台



Tianhe-2 HPC

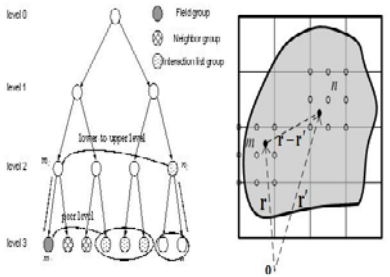
100个以上用户可以同时在平台上设计IC， 千万门级电路的仿真加速2-10倍

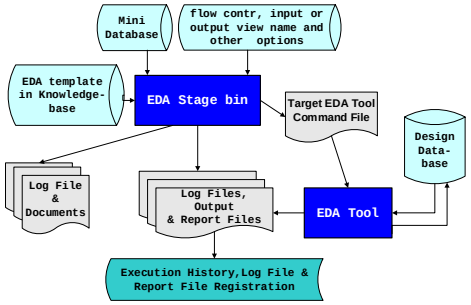
Transfer EDA toolkits to HPC env

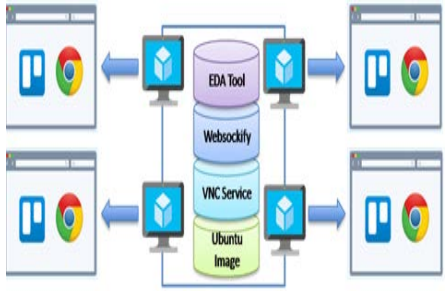
EDA design flow development

HPC resource management

Visualize remote user management







谢 谢!