“天河一号”异构通用计算环境

孟祥飞 博士

mengxf@nscc-tj.gov.cn

2011年12月2日 于中国科大

内容提要

- 超级计算发展回顾
- “天河一号”系统概况
- “天河一号”异构并行编程思想
- “天河一号”应用实践
- 大规模科学计算环境建设

超级计算发展回顾

- 科学研究的三大方法
  - 理论、实验、**计算**
- 是解决国家经济建设、社会发展、科学进步、国家安全和国防建设等领域一系列重大挑战性问题的重要手段
- 是国家综合国力、科技竞争力和信息化建设能力的重要体现
- 是国家创新体系的重要组成部分

超级计算发展回顾

CrayT3D

1993, 19Gflops

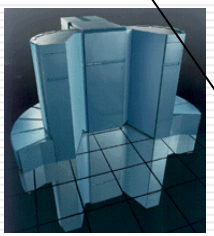
Cray T3E-1200

1998年1Tflops

IBM BlueGene/L

2005年，367Tflops

2007年，596Tflops



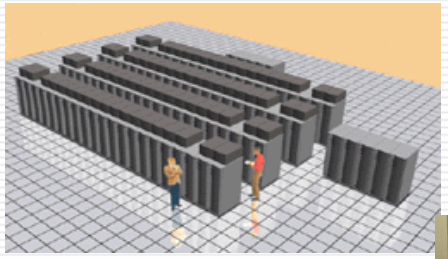
Cray-YMP

1988年，2.3Gflops

Cray-1

1976 160Mfl

30年前第一台商用巨型机问世



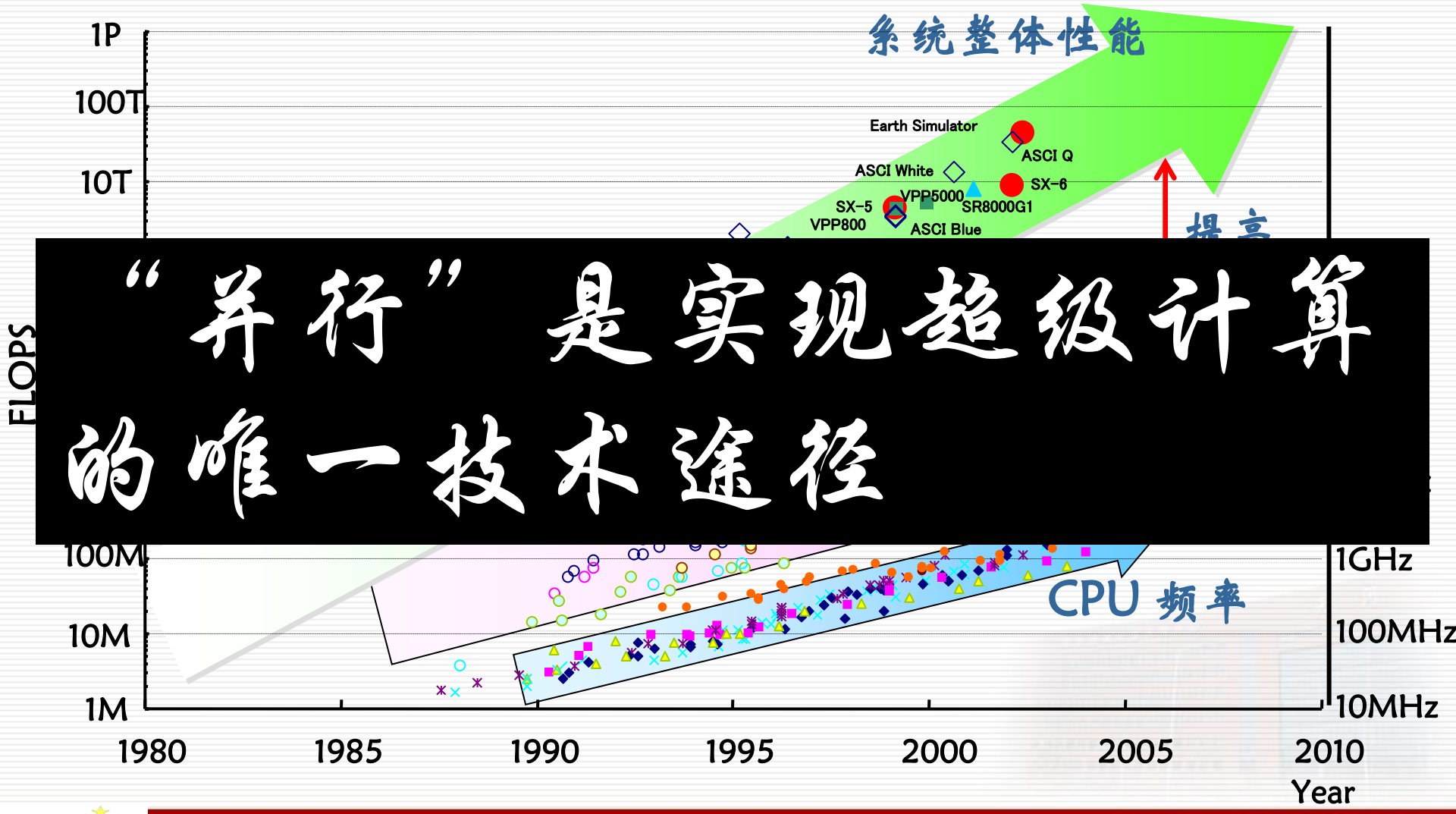
IBM红色选择

1999年，3万亿次

日本地球模拟器

30多年性能提高了3000万倍

超级计算发展回顾



“并行”是实现超级计算的唯一技术途径

超级计算发展回顾

- 千万亿次时代（PetaFlops）
  - 2008年 美国的Roadrunner，1.4PF(Top1 CPU并行)
  - 2009年 中国TianHe-1 1.2PF (Top5 异构并行)
  - 2010年 中国Tianhe-1A 4.7PF（2.566PF，Top1 异构并行)
  - ……

内容提要

- 超级计算发展回顾
- “天河一号”系统概况
- “天河一号”异构并行编程思想
- “天河一号”应用实践
- 大规模科学计算环境建设

“天河一号”超级计算机系统

- “863计划”信息技术领域“高效能计算机及网格服务环境”重大项目“千万亿次高效能计算机系统”研制成果
- 国家超级计算天津中心业务主机，中国国家网格主节点
- 国防科技大学牵头、滨海新区和浪潮公司参与研制
- 全系统分两期完成研制（2009.9；2010.8）

国防科学技术大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

TH-1A系统概述

- 异构结构：CPU + GPU
- 峰值性能：4.7PFlop/s
- 持续性能：2.566PFlop/s
- 功耗(满载)：4.04MW


TOP500 CERTIFICATE

Tinahe-1A, a NUDT YH Cluster System at the National Supercomputing Center in Tianjin, China

is ranked

No. 1

among the World's TOP500 Supercomputers

with 2.566 PFlop/s Linpack Performance

on the TOP500 List published at the SC10 Conference, November 16, 2010

Congratulations from the TOP500 Editors

Hans Meuer, University of Mannheim; Erich Strohmaier, NERSC/Berkeley Lab; Jack Dongarra, University of Tennessee; Horst Simon, NERSC/Berkeley Lab

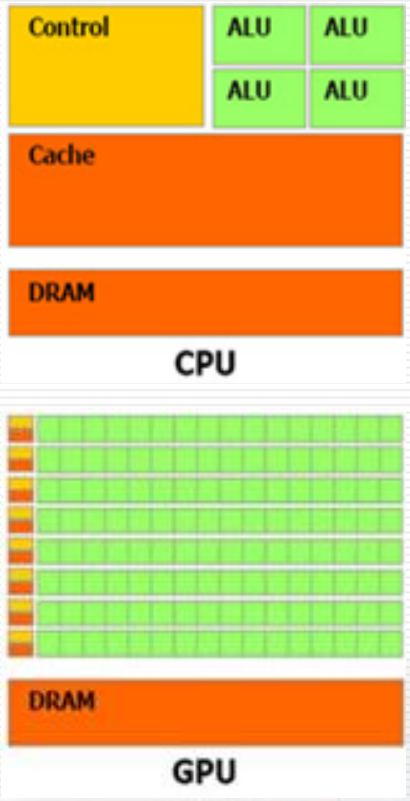
Items	Configuration
Processor	14336 Intel CPUs + 7168 nVIDIA GPUs + 2048FT CPUs
Memory	262TB in total
Interconnect	Proprietary high-speed interconnecting network
Storage	2PB
Cabinet	120 Compute，14 Storage，6 Communication

- 图形处理器(GPU)
  - graphics processing units
- 用于通用计算目的的图形处理器(GPGPU)
  - General-Purpose computation on GPUs
- Stream Computing (AMD)
- GPU Computing (NVIDIA)

- CPU发展规律
  - 基本上遵循“摩尔定律”（主频、多核、宽SIMD）
- GPU发展规律
  - 1993年开始，GPU的性能以每年2倍多的速度增长，近来趋势有所放缓，但每年1倍仍有希望，GPU浮点性能已比CPU高5~10倍

年份	CPU	GPU
2008	Intel 3.0 GHz Xeon E5472	NVIDIA Tesla C1060
	48Gflops	78Gflops
2010	Intel 2.93GHz Xeon X5670	NVIDIA Tesla M2050
	70Gflops	515Gflops

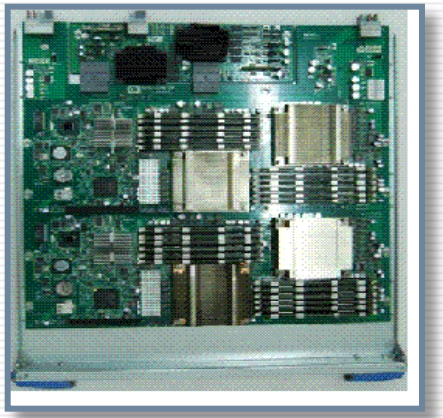
- CPU的大部分晶体管主要用于构建控制电路和Cache
  - CPU的5%是ALU，控制电路设计复杂
  - 访存延迟低，强调单线程能力
- GPU控制电路相对简单，而且对Cache的需求小，可以把大部分的晶体管用于计算单元
  - GPU的40%是ALU
  - GPU的内存带宽是CPU的5～10倍
  - 通过大量并发线程的运行来隐藏延迟

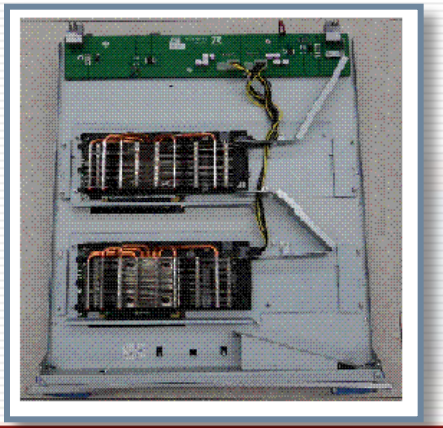



最 高 指 示
青年应该把坚定正确的政治方向放在第一位
时代不同了，男女都一样，男同志能办到的事情，女同志也能办得到。
结 婚 证
字第 号
自愿结婚，经审查合于
中华人民共和国婚姻法的规定，发此证。
敬祝毛主席万寿无疆！
一九六八年 月 日

“天河一号”技术突破

- 国际上首台CPU和GPU异构混合的千万亿次系统
  - 2007年，《64位流处理器体系结构研究》一文从学术上论证了流处理器用于高性能科学计算的可行性
  - 2009年，天河一号从工程实现上证明了GPU用于高性能科学计算的可行性
    - 流处理器研究成果发表在计算机体系结构顶级国际会议ISCA’2007上
    - 这是该会议近十年录取的第一篇来自中国研究机构由中国学者独立完成的
    - 被IEEE Transaction on Parallel and Distributed Systems 录取
  - 天河一号在国际上掀起了异构高性能计算的热潮





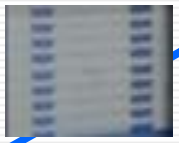
“天河一号”技术突破

- 国际上首台CPU和GPU异构混合的千万亿次系统

TH-1A System

Cabinet
(4 x rack)

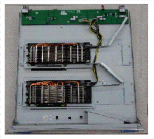
TH-Net

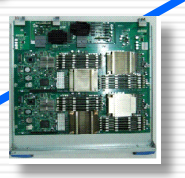


rack
(16 x cn)

Compute node
(4CPU+2GPU)

Twin GPU blade





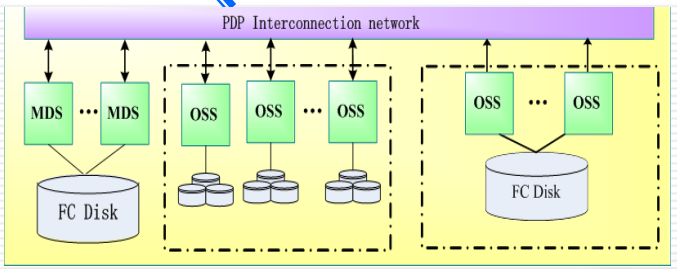
Quad cpu blade

M2050

X5670

FT-1000

Chips



On-line storage

“天河一号”技术突破

64位多核多线微处理器设计

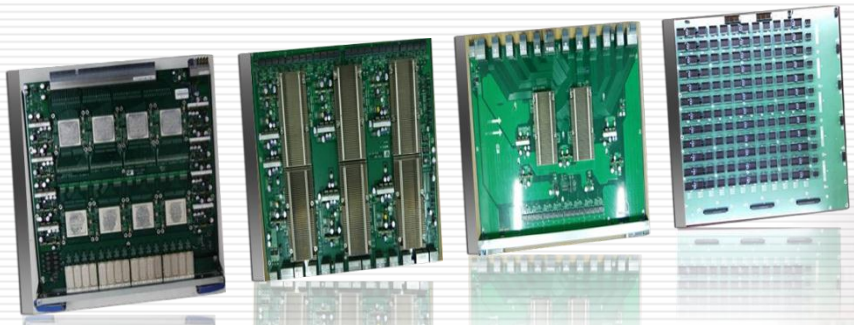
- 设计了自主高性能CPU：FT-1000
  - ✓8核64线程
  - ✓片上集成DDR3存控、PCIe 2.0 和 CPU直连接口
  - ✓主频1GHz，峰值性能8亿次
  - ✓SPEC实测性能达到2006年国际商用主流CPU水平
- 设计了4结点的自主CPU主板
- FT-1000被列入“国产软硬件推进计划”

“天河一号”技术突破

互联网络

- 互连通信对系统的实用性能具有至关重要的影响
  - ✓ 定制互连是国际上区分MPP和Cluster的主要技术特征
  - ✓ 美国在该领域对我国进行严格的技术封锁
- 层次胖树结构。8个结点通过16口交换板互连，全系统通过11个384口交换机互连
- 信号速度10Gbps，IB QDR的2倍；交换机吞吐率达61.44Tbps，IB QDR交换机的2.37倍

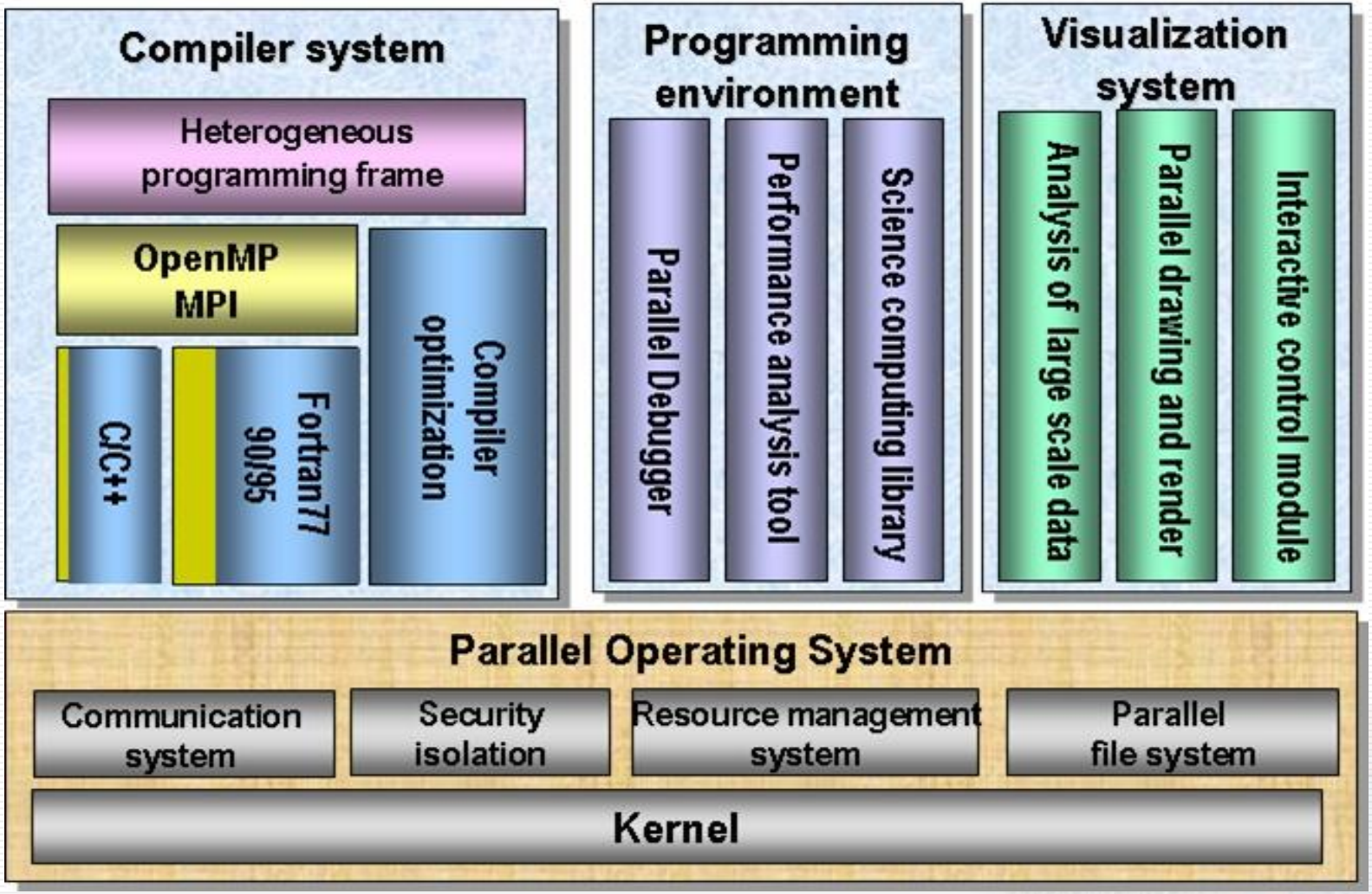
交换板，叶交换刀片，根交换刀片，交换机背板



NIC 与 NRC

交换机

“天河一号”超级计算机系统



TOP500 CERTIFICATE

Tinahe-1A, a NUDT YH Cluster System at the National Supercomputing Center in Tianjin, China

is ranked

No. 1

among the World's TOP500 Supercomputers

with 2.566 PFlop/s Linpack Performance

on the TOP500 List published at the SC10 Conference, November 16, 2010

Congratulations from the TOP500 Editors

Hans Meuer
University of Mannheim

Erich Strohmaier
NERSC/Berkeley Lab

Jack Dongarra
University of Tennessee

Horst Simon
NERSC/Berkeley Lab

“天河一号”重要意义

- “天河一号”显著提升了我国在超级计算机领域的国际地位，我国自主研制超级计算机水平进入世界先进行列
- “天河一号”为解决我国经济、国防、科技等领域的挑战性问题提供了重要手段，对于提升我国综合国力具有重要战略意义

- 计算、实验、理论：科学研究的三大“支柱”
- 提升我国的科技创新能力
- 推动我国的经济发展
- 增强我国的社会保障能力

内容提要

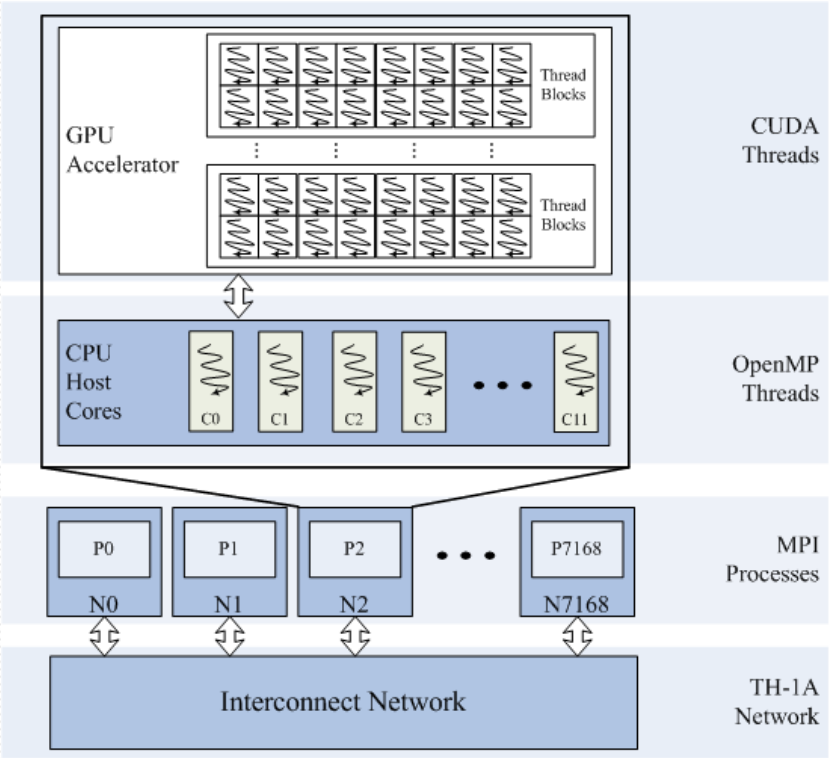
- 超级计算发展回顾
- “天河一号”系统概况
- “天河一号”异构并行编程思想
- “天河一号”应用实践
- 大规模科学计算环境建设

异构并行编程

- 加速大规模、复杂应用，特别是处于开发状态的应用或者无法获得完整的源代码的应用
- 使用CPU和GPU的协同计算能力，同肘（部分）隐藏GPU编程
- 编程方法
  - 采用多级混合并行编程模型
  - 结点间使用同构并行编程，面向物理、数学、并行算法专家
  - 结点内使用异构并行编程，面向体系结构、编译、操作系统等计算机专家

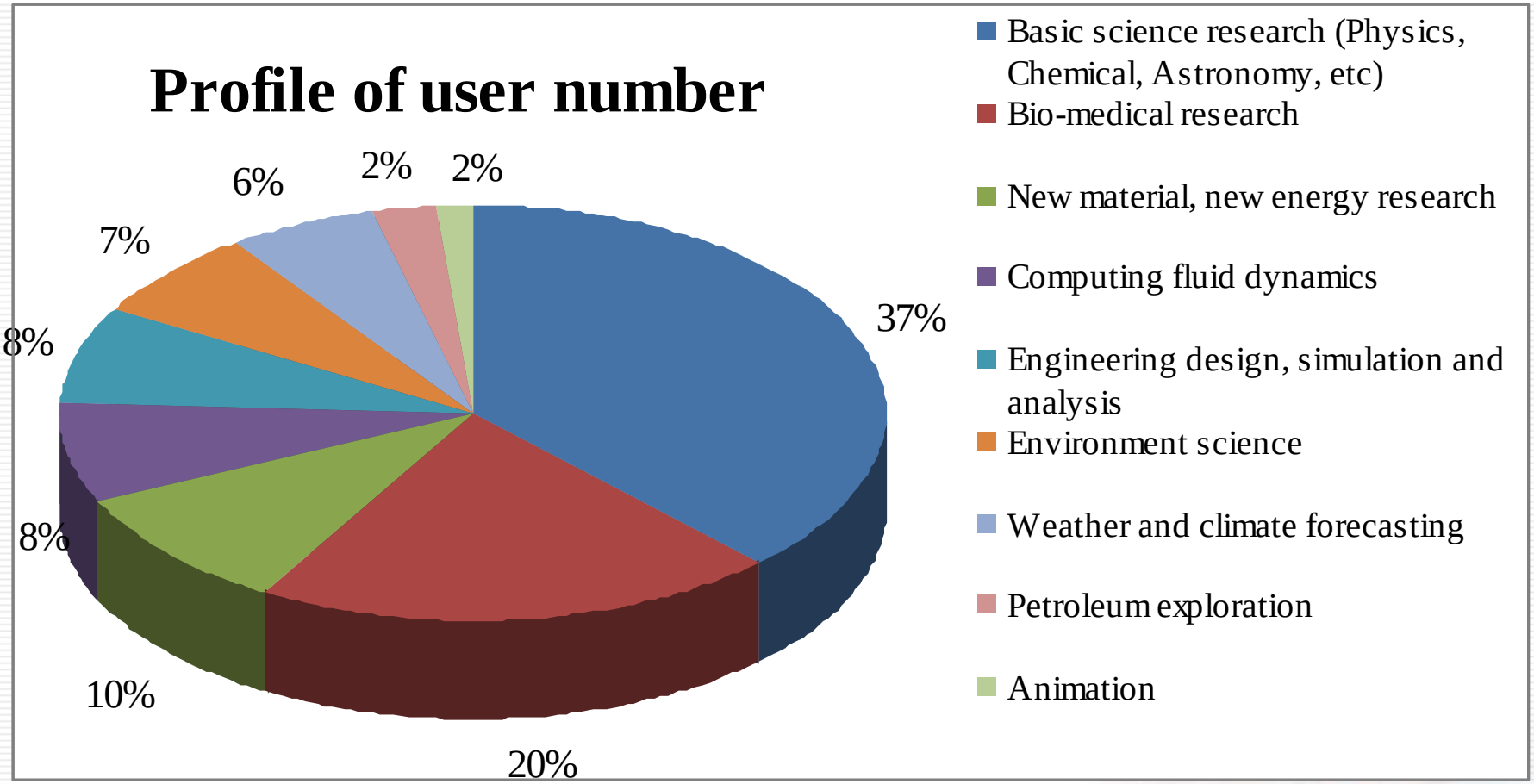
异构并行编程

- 多级混合并行编程模型
  - 体系结构
    - 结点间，对称结构
    - 结点内，异构结构
  - 编程模型
    - 结点间，消息传递
    - 结点内，共享存储
      - 纯CPU线程
      - CPU线程（控制GPU线程）

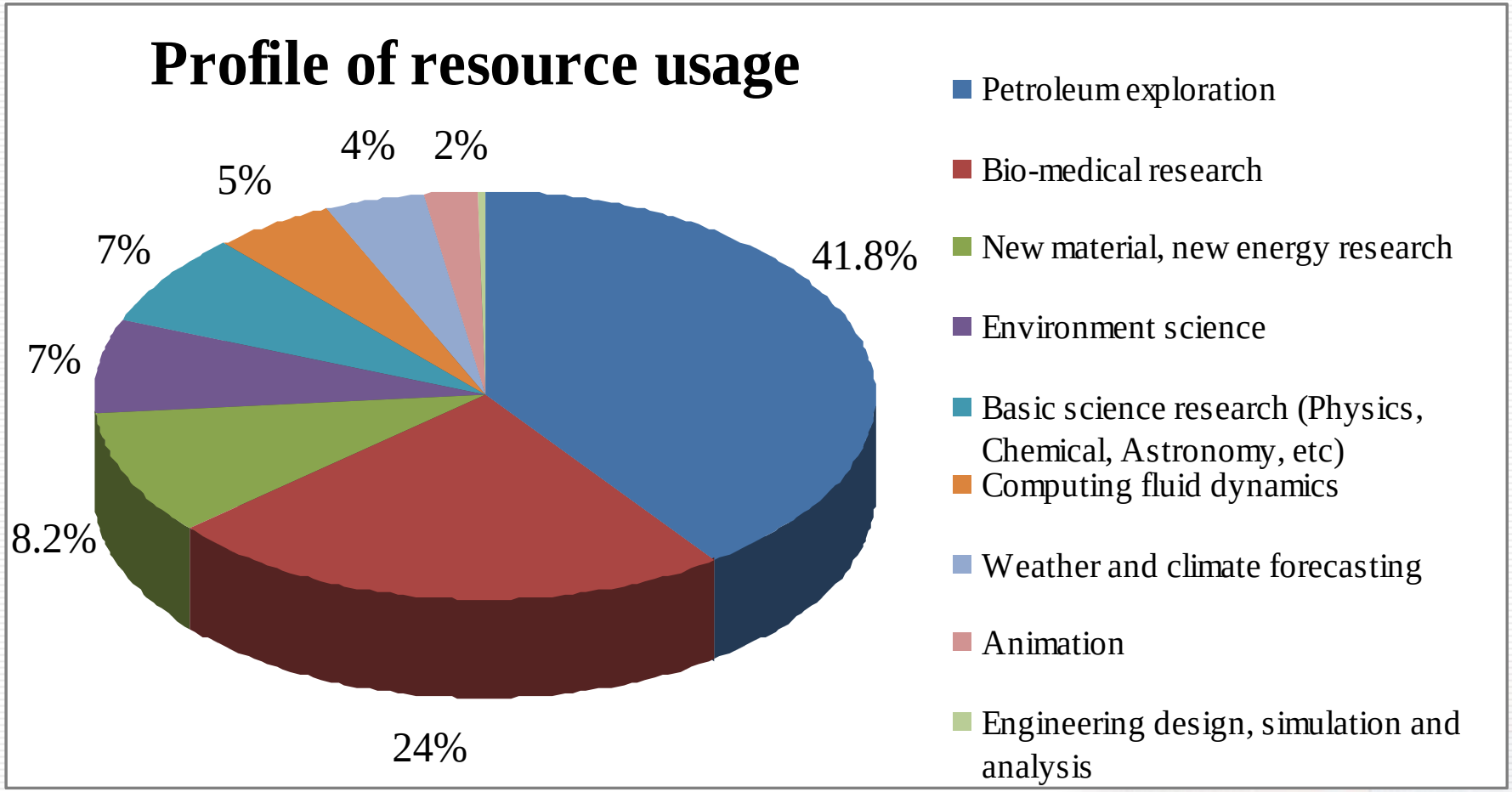


内容提要

- 超级计算发展回顾
- “天河一号”系统概况
- “天河一号”异构并行编程思想
- “天河一号”应用实践
- 大规模科学计算环境建设

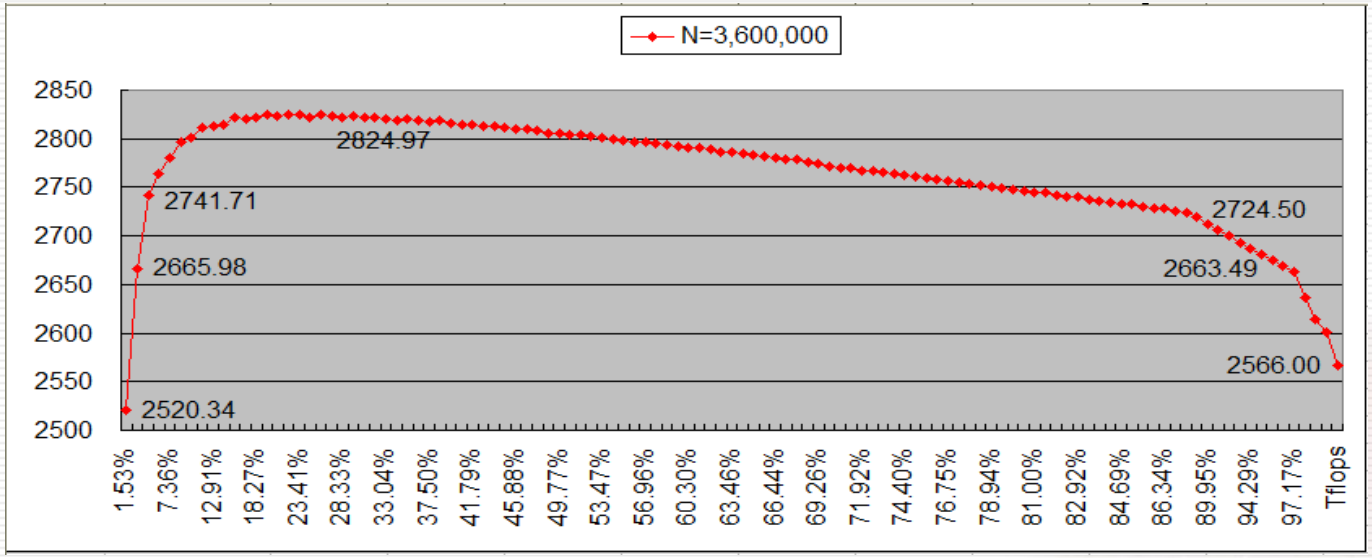


——NSCC-TJ (Nov.2010 – Aug. 2011)



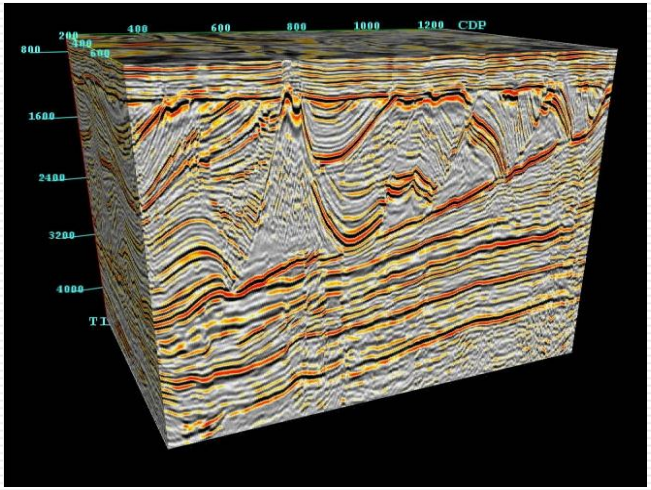
——NSCC-TJ (Nov.2010 – Aug. 2011)

- Tianhe-1A，HPLinpack
  - 求解稠密线性系统
  - 计算复杂度 $(2/3)N^3 + 2N^2 + N$
  - 参数：N=3,600,000，NB=512，Tflops=2566



石油地震勘探数据处理

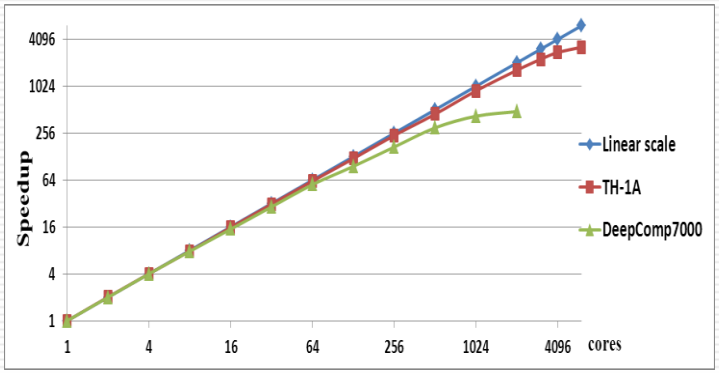
- 面积1060平方公里石油勘探地震资料三维逆时叠前深度偏移处理，是最大规模的资料
- 中石油公司自主设计的“叠前深度偏移”软件，在“天河一号”全系统上用时仅不到16小时
- 在国内已有的大型计算机上需要计算近30天，在“天河”上提高速度约30倍



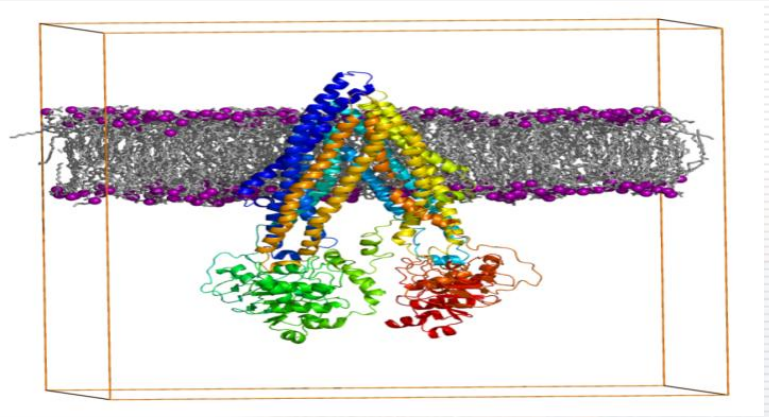
地下地质结构示意图，32 x 32 x 5 (km)，1060平方公里

生物分子动力学模拟

- 采用NAMD软件，对于百万原子规模的生物体系，最大测试规模12288核。采用6144个核计算，模拟速度大约为20纳秒/天
- 上海药物所采用Gromaccs软件对大规模的模拟体系（约四十万个原子），截断半径为12 埃 ，采用2048个核，模拟速度为85纳秒/天

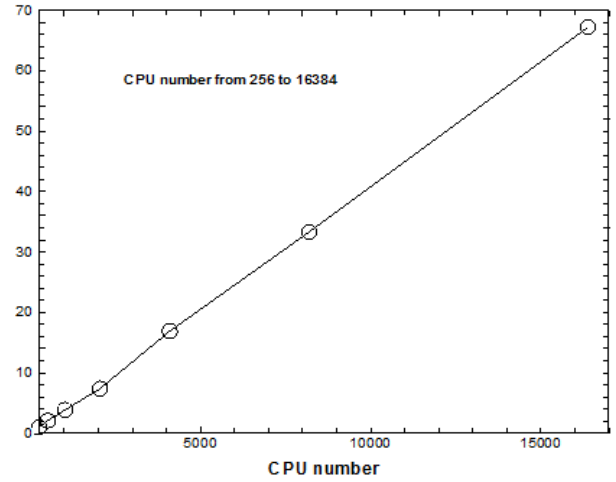


NAMD百万原子体系测试结果

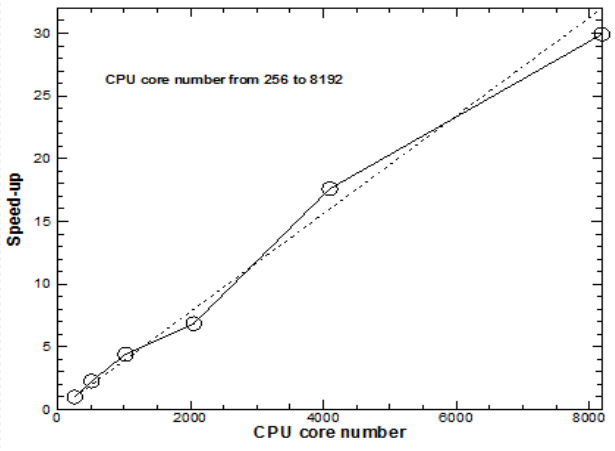


航空--飞行器流场模拟仿真

- 中航第一飞机设计研究院，中科院力学所等
- 应用规模：最大规模16384CPU核
- CCFD应用测试，可压缩平板绕流数值模拟，来流马赫数等于8，来流温度等于169.44K（壁面温度等于1.9倍来流温度），雷诺数等于2000000
- RAE2822绕流模拟和强冷壁平板绕流模拟测试，马赫数等于8的平板绕流做了直接数值模拟并取得良好的测试结果，之前世界其他仿真结果的马赫数都在6以下

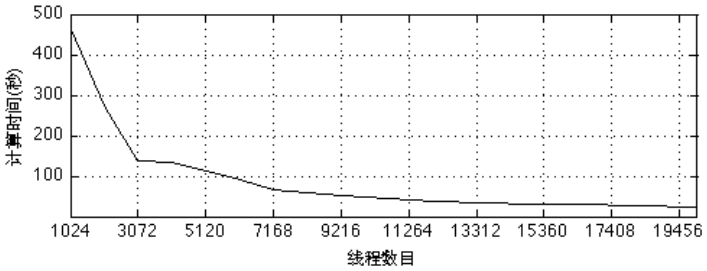


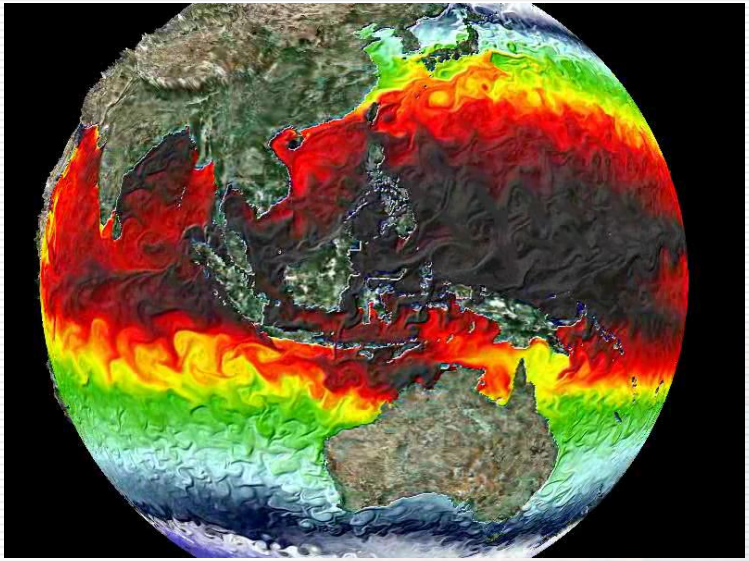
(b)



全球海洋环境与气候变化研究

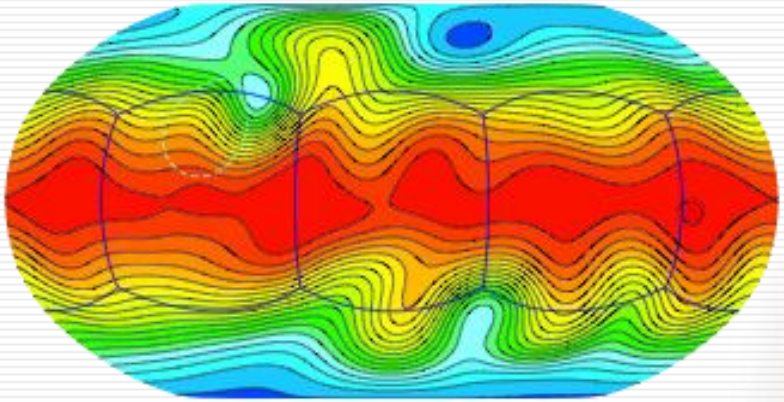
- 国家海洋局、北京师范大学、中科院大气所、加拿大海洋所与天津科技大学等
- 研究内容：在“天河一号”上开展了全球气候变化及地球科学系统研究。利用“天河一号”的超级计算能力进行地球系统多模式耦合研究
- 应用规模：国家海洋局一所自主海浪模式并行应用规模最大20000核，性能优异，加速比保持线性甚至超线性。各研究机构业务规模基本都在数千核





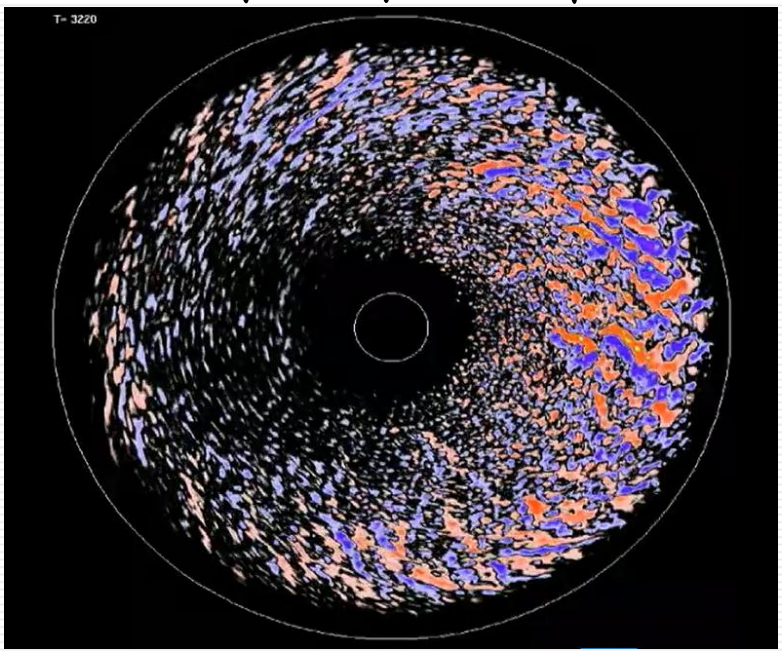
全球大气浅水波全隐式数值模拟

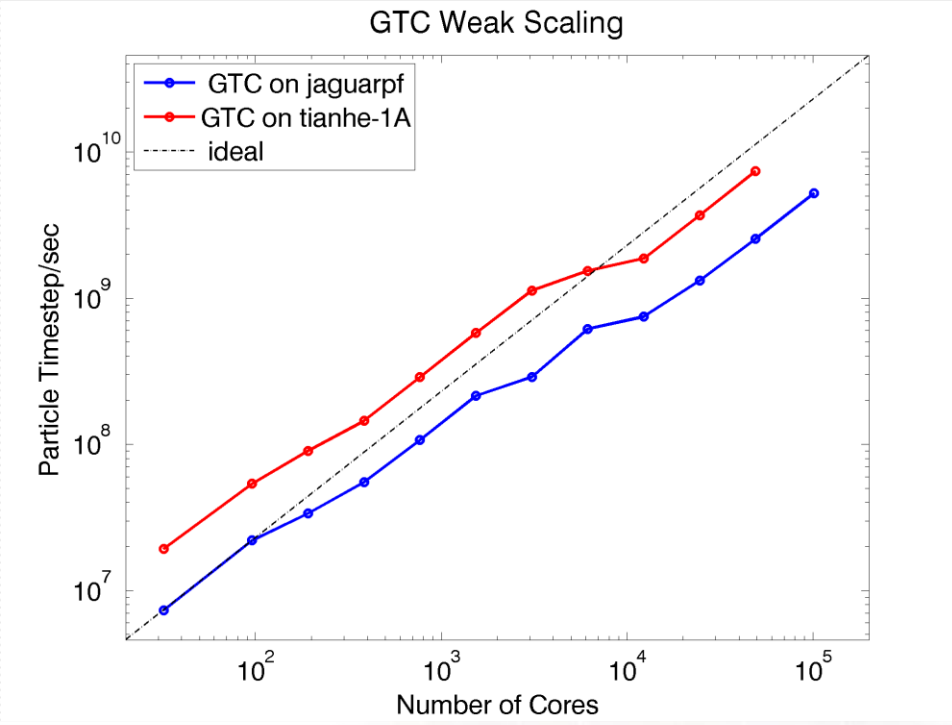
- 中科院软件所自主研发
- 全球数值气候预报耦合了大气、海洋、陆地、海冰四个模式，其中大气模式的计算占据了大量的时间
- 程序从1152核扩展至82944核，并行效率为60%
- 下图为采用Williamson标准测试集中的Isolated mountain算例，第15天的模拟结果



磁约束聚变应用研究

- ITER-CN计划
- 与美国能源部“美洲豹”系统比较
- 与北大、浙大聚变中心联合开发GPU版本程序，并完成第一阶段工作

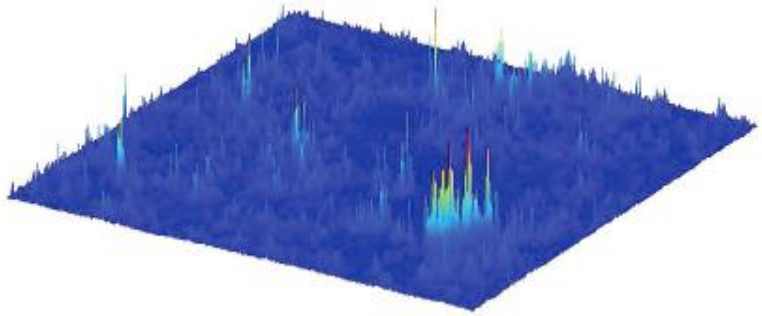




世界最大规模湍流模拟

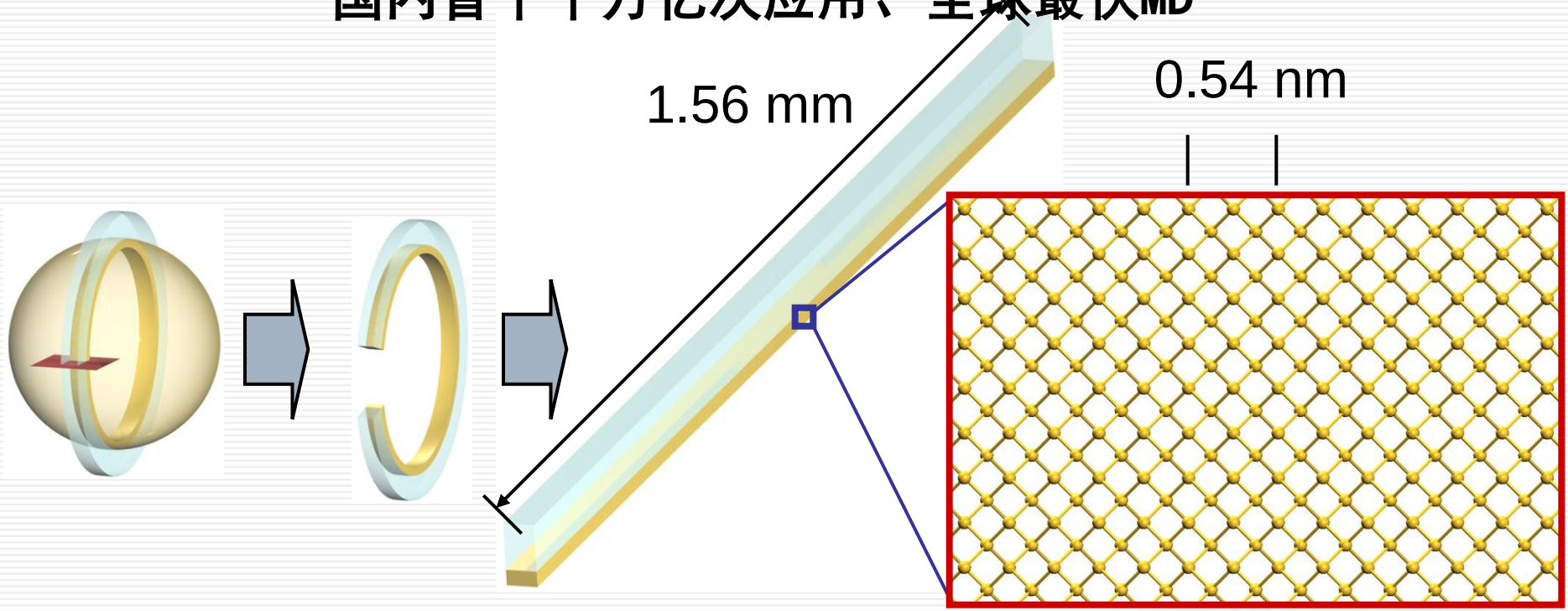
- 中国北京大学工学院，自主开发CPU+GPU程序
- 在“天河一号”7168个节点上进行了14336立方数据规模的湍流数值模拟，其流动参数已经与自然界中的实际湍流相当
- 分析结果可广泛应用于能源、化工、航空、造船等领域

计算规模（天河节点）	湍流规模	说明
2048	4096^3	之前世界最大湍流规模
4096	8192^3	
7168（含GPU）	14336^3	世界最大规模，流动参数与自然界实际湍流相当



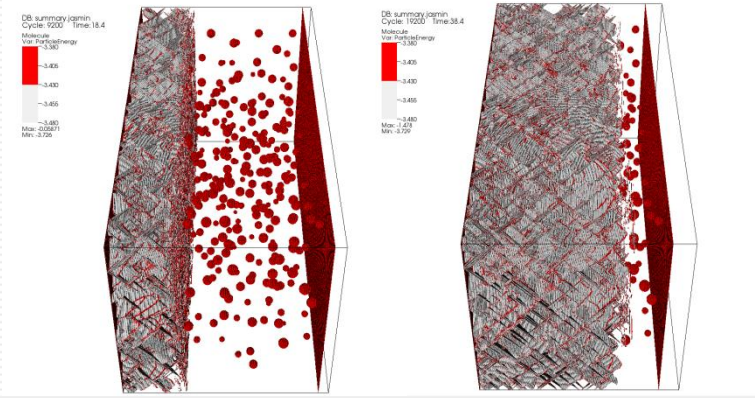
硅晶体原子模拟与尺度效应分析—分子动力学

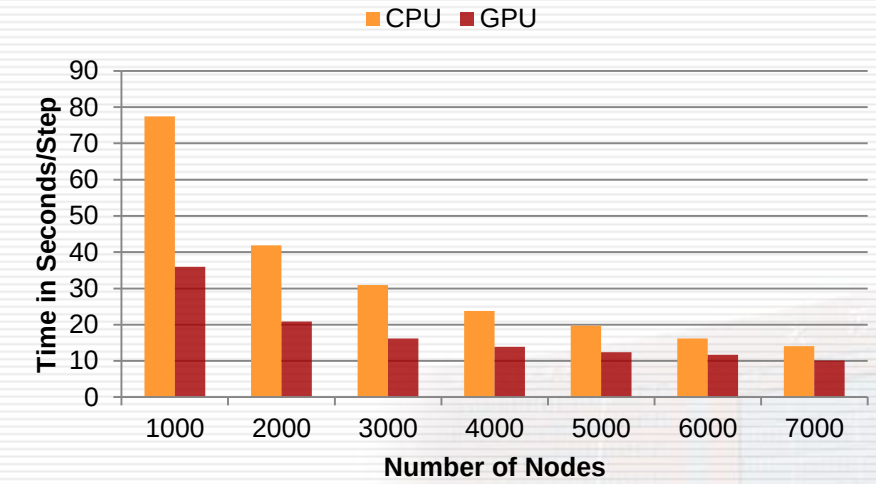
在 Petaflops 机器上的 Petaflops 应用
国内首个千万亿次应用、全球最快MD



1100 亿原子 (文献<15 亿)
1.87 Petaflops， Tianhe-1A 的7168 个GPUs

- Tianhe-1A，冲击波在金属介质里的传播
  - 粒子受力计算通过GPU加速（EAM Potential）
  - 单结点中，1个GPU相当于30个CPU核的性能
  - 7000个计算结点（182,000个CPU/GPU核）上模拟有良好的扩展性





内容提要

- 超级计算发展回顾
- “天河一号”系统概况
- “天河一号”异构并行编程思想
- “天河一号”应用实践
- 大规模科学计算环境建设

大规模科学计算环境建设

大规模高性能计算平台

大规模硬件设施

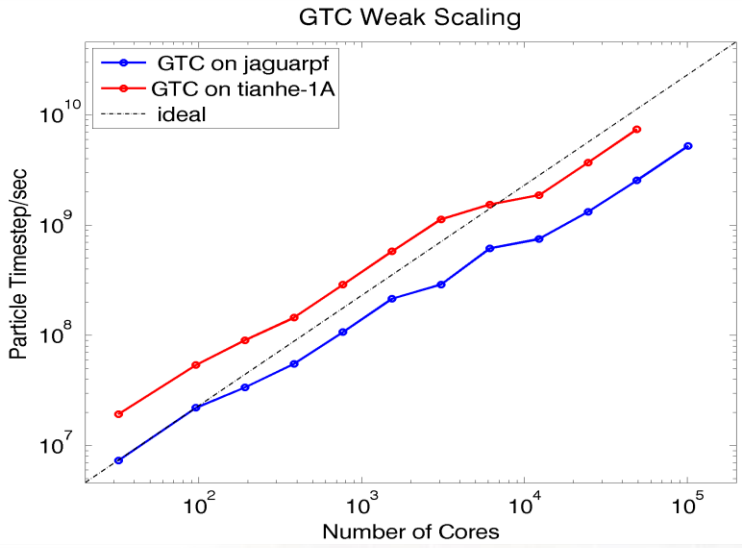
- 基于“天河一号”，提供万核量级大规模硬件平台
- 高速互联通信网络

访问网络环境

- 目前使用联通100M光线接入，北方用户的访问速度M量级
- 100M电信网络已于上月底启用

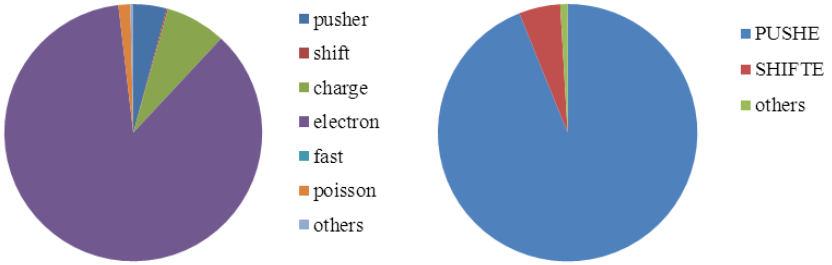
支持软件

- 编译、开发、科学计算软件、工程仿真软件等



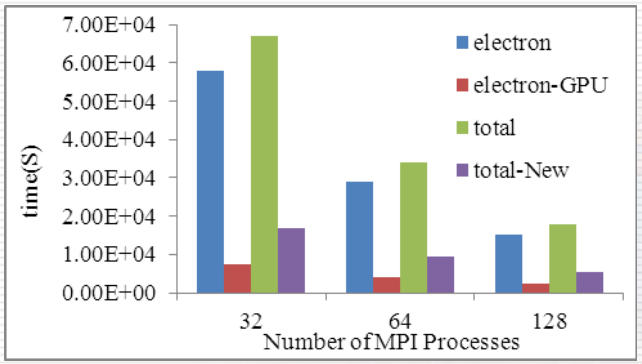
技术与研发

- 应用程序向“天河一号”系统移植
  - 程序部署
  - 程序优化：性能、I/O
- 高性能应用开发
  - 大规模并行开发
  - GPU应用开发
  - 异构并行开发
- 技术支持团队
  - 组建30-50人的技术支持、开发团队等



(a) Time proportion of modules (b) Time proportion of subroutines

Fig.2 Execution time proportion of GTC



合作申请国家支持

合作优势

- 中科大：中国著名的高等学府
- 天河：国家科技创新标志、服务平台

合作方式

- 合作申请各部委重点、重大项目，如科技部、基金委、能源局等
- 以协同创新为指导，探讨新的合作方式和争取多方支持

感谢您的支持与合作