Lustre User Group 2013 | China and Japan
Hosted by OpenSFS
Beijing - October 15 Tokyo - October 17
Sponsored by: intel

China LUG 2013

Our Work on Lustre at NUDT

董 勇
国防科学技术大学计算机学院计算机研究所
yongdong@nudt.edu.cn

- 主要内容
  - 关于我们
  - Lustre使用与优化
  - 当前的工作

关于我们

- 国防科学技术大学
  - 1953年，哈尔滨军事工程学院
  - 1970，南迁长沙，长沙工学院
  - 1978，国防科学技术大学
  - 1999，成立新的国防科学技术大学

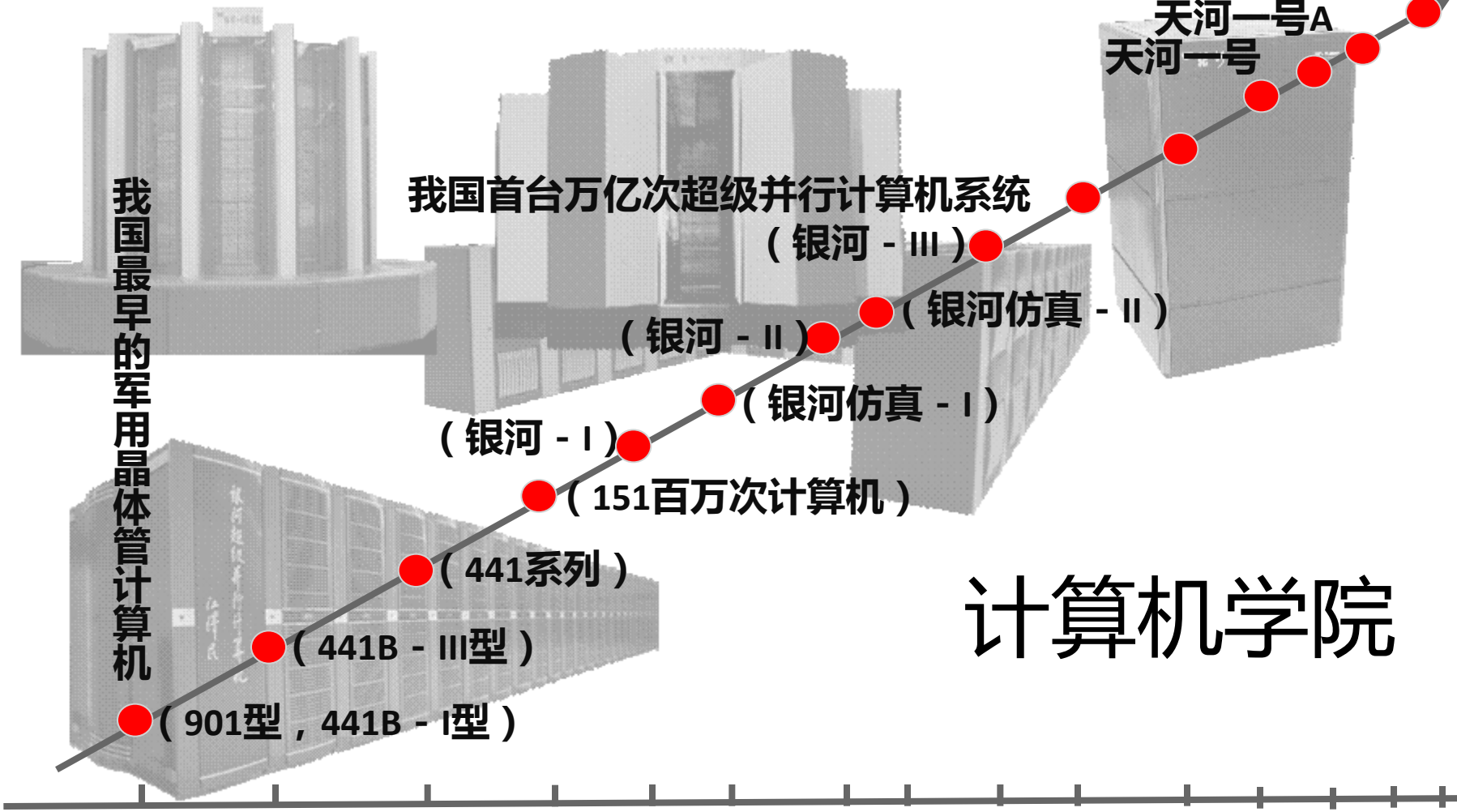
计算机学院

1958年计算机专业于哈军工起步

1966年成立全国第一个计算机系

1971年扩建成计算机系兼研究所

1999年成立计算机学院


我国最早的军用晶体管计算机
（901型，441B - I型）
（441B - III型）
（441系列）
（151百万次计算机）
（银河 - I）
（银河仿真 - I）
（银河 - II）
（银河仿真 - II）
我国首台万亿次超级并行计算机系统
（银河 - III）
天河一号
天河一号A
天河二号
1965
1970
1975
1979
1983
1985
1992 -93
1997
2000
05
07
09
10
13

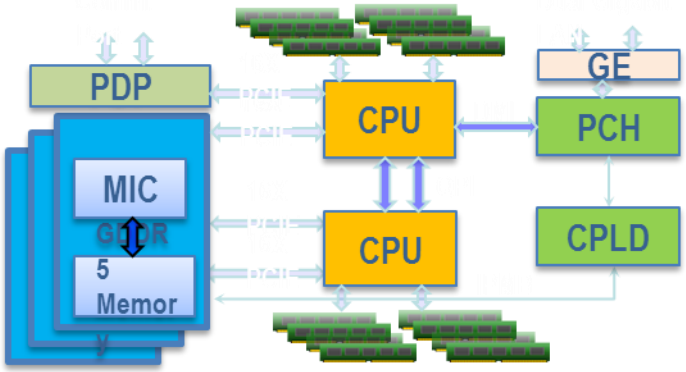
计算机学院

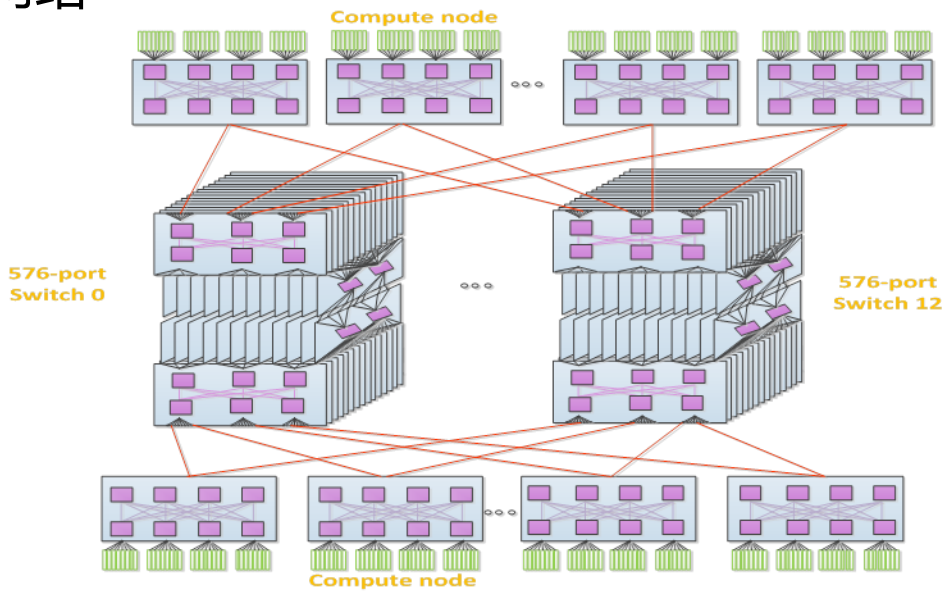
TH-2

天河

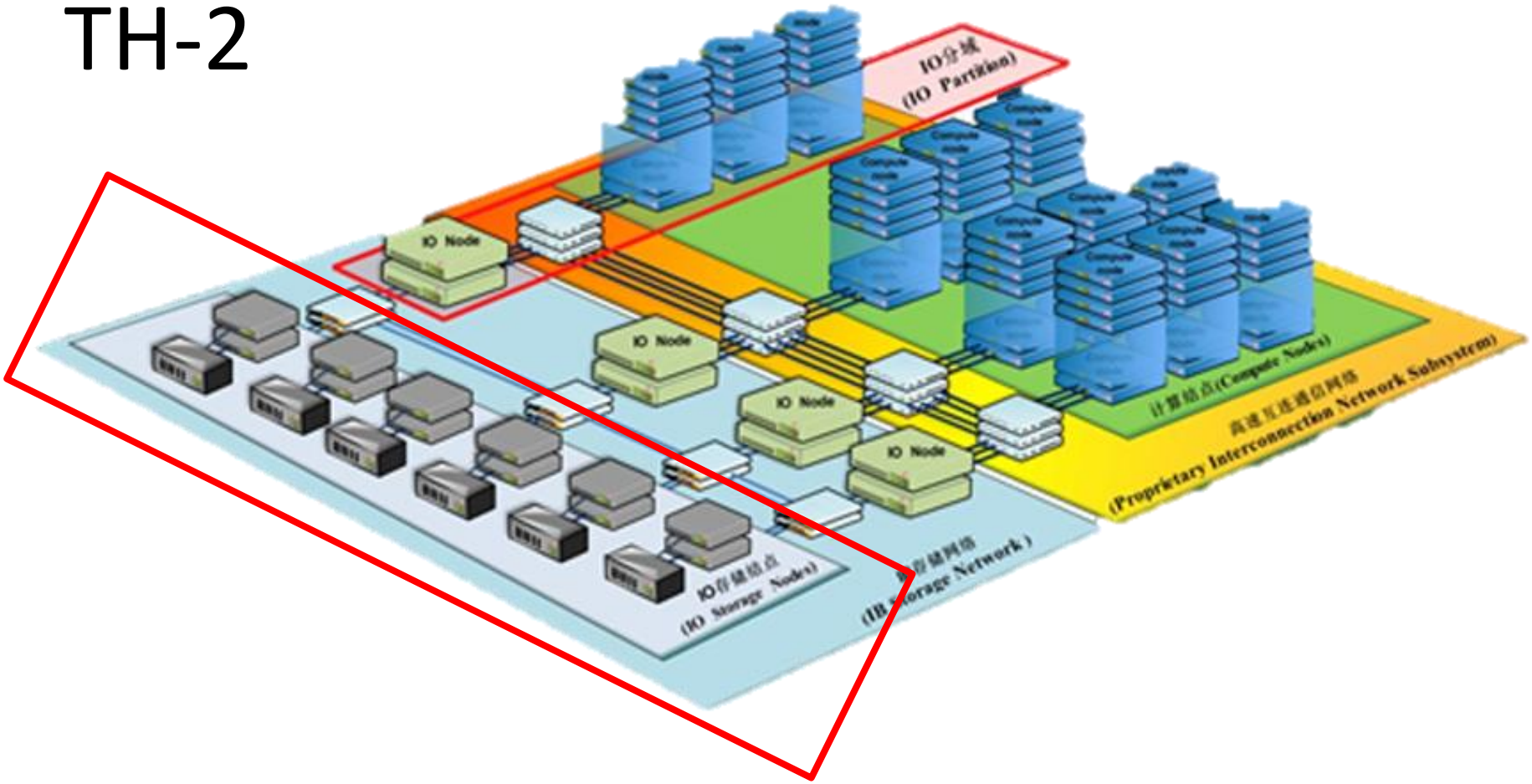
系统	配置
计算结点	16000个，双路12核CPU，3个MIC
处理器	32000 Xeon Ivy +48000 Xeon Phi +4096 FT1500 峰值 54.9PFlops, Linpack 33.86PFlops
高速互连网络	TH Express-2 10GB/s， 全局通信优化
内存容量	1.4PB
存储容量	Global shared parallel storage system, 12.4PB
机柜	125+13+24=162 compute/communication/storage
能耗	17.8 MW (1902MFlops/W)
制冷	密闭水风冷

- 微异体系结构（ Neo-Heterogeneous ）
  - ECC DDR3 DIMMs, 64GB
  - PDP接口
  - 结点峰值 3.432Tflops
  - 单机柜128结点，439.296TF
- TH Express-2 自主高速互连网络
  - 胖树拓扑结构
  - 13 个576口交换机
  - 光电混合传输





TH-2

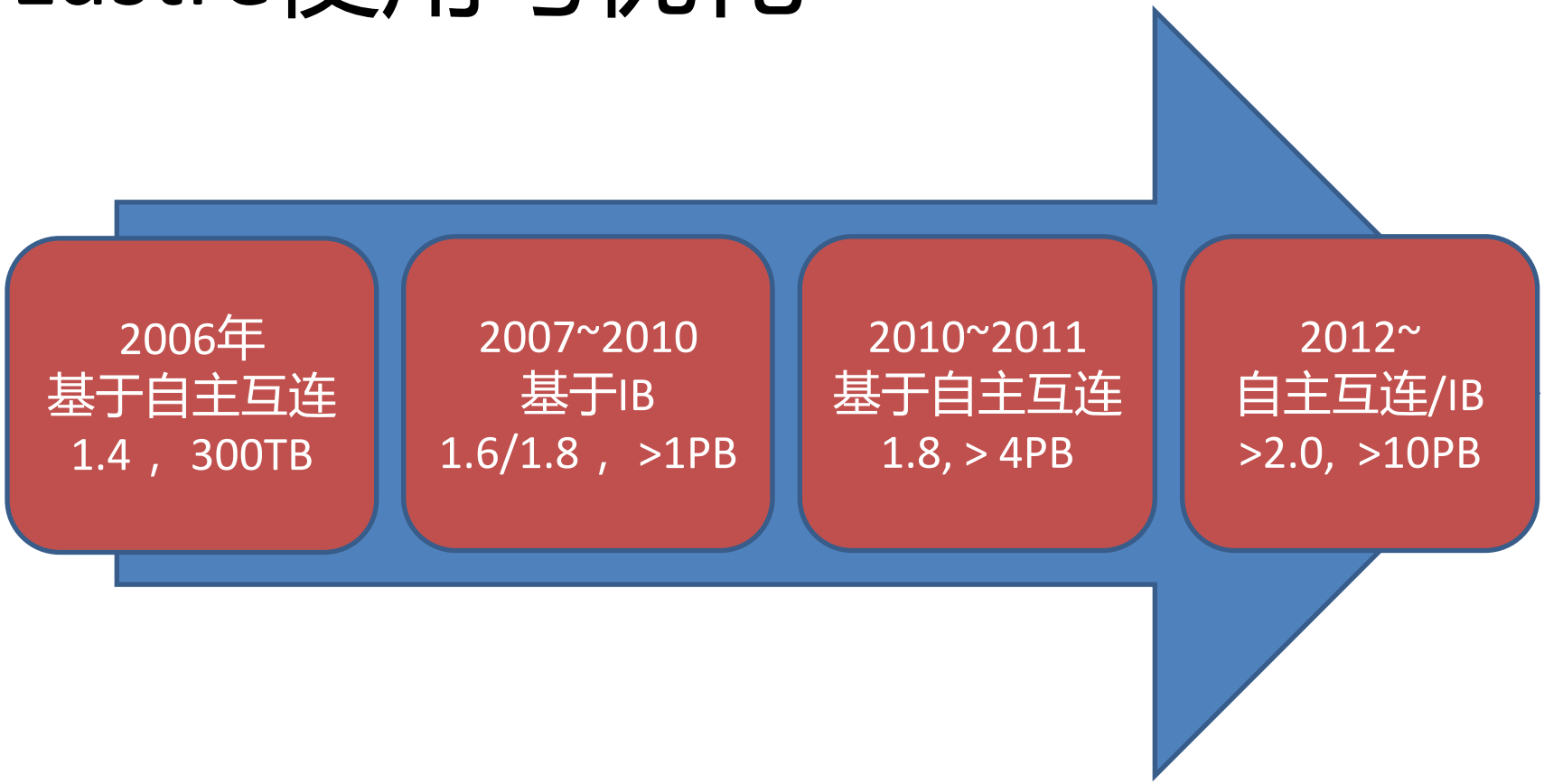

IO分域
(IO Partition)
IO Node
IO Node
IO Node
IO Node
计算结点(Compute Nodes)
(Proprietary Interconnection Network Subsystem)
IO存储结点
(IO Storage Nodes)
(IB Storage Network)

TH-2

- 并行文件系统研究组
  - 计算机研究所系统软件研究室
  - 教员+学生+工程师
  - 工程任务为主，兼顾学术研究
  - 主要为银河/天河系统定制并行文件系统
  - 2004年开始评估、分析、开发Lustre
    - 移植，体系结构适配
    - 优化，性能最大化
    - 新功能开发

Lustre使用与优化

Lustre使用与优化



Lustre使用与优化

- Lustre是HPC系统并行文件系统的首选
  - 最大限度利用硬件带宽
    - 磁盘
    - 互连网络
  - 可接受的可靠性
  - 对设备的广泛支持
  - 开源
  - 版本推进
- 实际应用效果证明了上述结论

Lustre使用与优化

- 我们关注
  - 性能
  - 可靠性
  - 可管理性
  - 易用性
  - 故障诊断
  - 数据备份
  - …

- Lustre性能
  - 基于自主互连网络的通信模块
    - GLND
    - 基于自主互连网络Glex[1]
    - 采用RDMA协议
    - 1.4，1.6，1.8，2.0～
    - 运行在自主系统中，银河/天河系列
    - 具有比IB QDR更高的通信带宽
  - 多IB设备绑定通信

[1] Xie M, Lu Y, Wang K, et al. Tianhe-1A Interconnect and Massage-Passing Services[J]. IEEE MICRO, 2012, : 8～20

- Lustre性能
  - 小文件读写性能优化
    - OSS小文件专用cache优化
      - 基于对象命名关联性预取
    - 优势
      - 通过cache降低小文件的磁盘访问开销
      - 减少大文件读写对小文件访问引发的干扰
      - 多client共享OSS cache
  - 面向应用的性能调优
    - 参数配置
      - Cache，client/OSS
      - RPC
      - 网络拓扑优化

- Lustre可靠性
  - 数据可靠性设计
    - 客户端基于LOV实现数据文件RAID0+1
    - 针对关键数据
    - 对读性能优化
  - 根据不同的配置采用不同的可靠性方案
    - Lustre自带HA功能
    - 存储节点自有RAID设备

- Lustre可管理性
  - 优化监控与管理
    - MDS和OSS采用单映像管理
      - 网络引导、自动配置
      - 成员管理，文件系统配置自动化
      - 参数调整，OST分区和屏蔽，容量扩展，HA
    - OSS监控
      - 状态监控
      - OST和盘阵联动
      - IPMI
    - 对client采用主动监控
      - OSS连接数

- 故障诊断
  - dmesg
  - Hard work
  - 不断总结，经验很重要
- 数据备份
  - copy
  - tar

- 对Lustre的需求
  - 可扩展性
    - 支持数十万结点
  - 可靠性
    - 后端FS的可靠性
      - Super block group corrupted…
      - How about Zfs ?
  - 性能
    - 元数据集群？

➢ 可管理性
- dmesg更易懂
- 监控、管理工具
- 版本问题
  - kernel，ofed，lustre
- 备份问题

目前的主要工作

- HPC系统性能稳步提高，I/O瓶颈问题无法回避
  - E级系统体系结构需求
    - TB/s以上带宽，>100PB以上空间
    - 高并发性要求，100K以上
    - 管理能力需求，满足效率、可靠和成本多项要求
  - 应用需求
    - 传统应用的I/O需求： 计算数据、检查点、 …
    - 数据密集计算的I/O需求：地震数据处理、可视化…
    - Big data类应用： graph、基因数据、社会网络…

- 多类存储资源具备不同的特性，相互之间无法替代
  - 本地存储、全局存储
  - SSD、 disk
  - Single disk 、 RAID
- 多类系统架构具有不同的特点，拥有各自的优势

- 面向未来新系统的文件系统设计H^2FS
  - 全新的设计，用户态实现
  - 基于自主互连网络的通信优化
  - 更好的可扩展性
  - 提高应用可获得的I/O性能
  - Lustre作为重要支撑组件

- 主要思想
  - 分层
    - 全局存储 Lustre
    - 本地存储
  - 计算与I/O的融合
    - 融合使用视图
    - 统一管理
  - 提高数据访问性能
    - 放松I/O语义
    - 挖掘应用特点

- 目前状态：
  - 已经在地震数据处理应用中成功应用
    - 典型的计算密集型应用，同时具有I/O密集型的特点，集中存储存在严重的可扩展性问题
    - 输入输出总的数据量巨大
    - 各节点处理数据相对独立
    - 与检查点应用、大规模可视化应用特点类似

- 总结
  - Lustre是HPC系统的首选
  - Lustre可以更好
    - 可扩展性
    - 可靠性
    - 易用性
    - …
  - 面向未来新系统H²FS

谢谢！