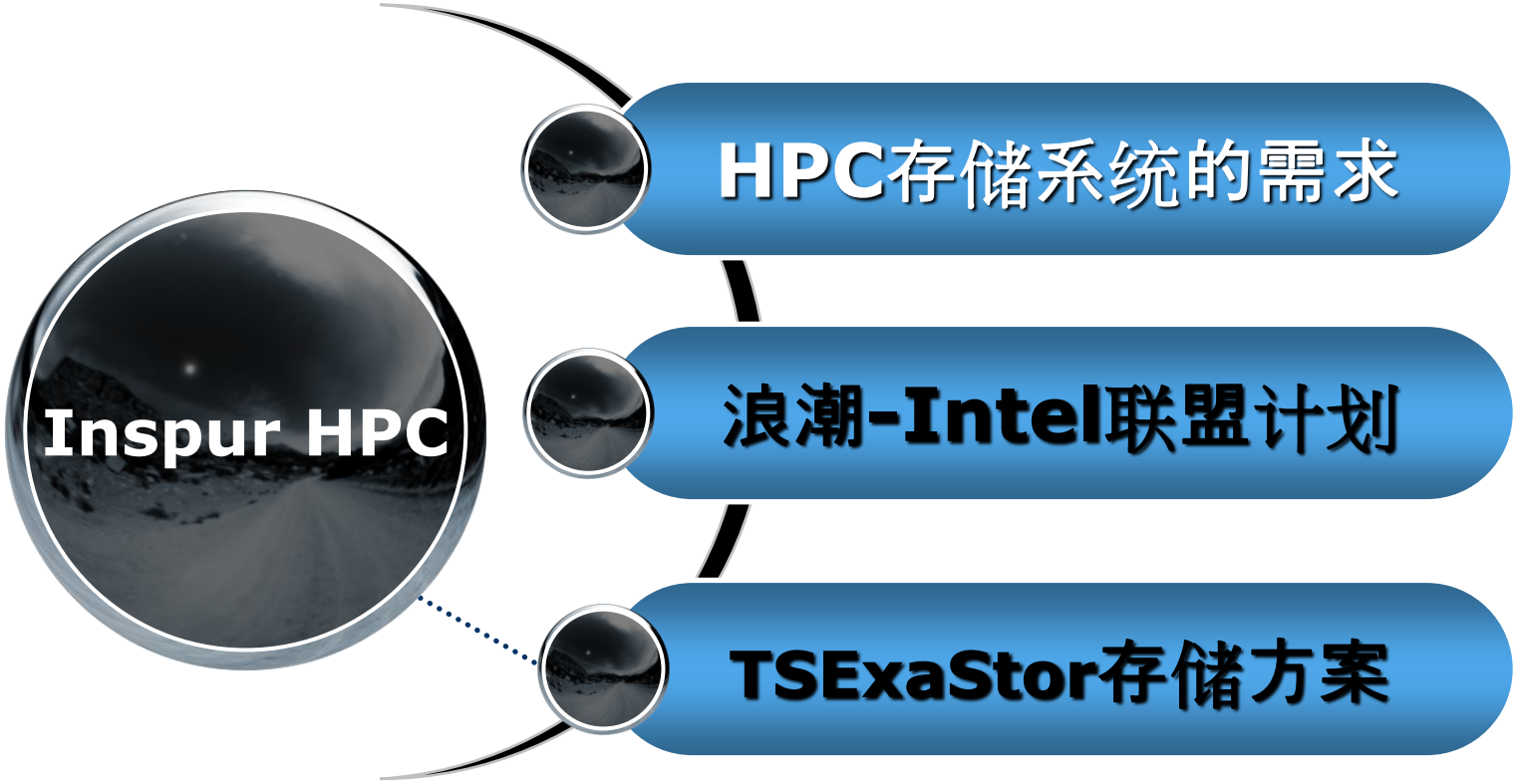
Lustre User Group 2013 | China and Japan
Hosted by OpenSFS
Beijing - October 15 Tokyo - October 17
Sponsored by:

浪潮高效能HPC存储TSExaStor

--基于IEEL

浪潮集团 HPC产品部 姜乐果

Content



PetaScale千万亿次时代……

2010.11 “天河1A” 以4.7 PetaFlops 峰值性能位列全球超级计算机第一. 国防科大与浪潮承担十一五863重大专项

2011.11 国家超算济南中心神威蓝光全国产处理器千万亿次超级计算机；浪潮承担通用处理器刀片计算集群、2 PB高带宽海量存储系统设计供应等

天河二号 再次问鼎全球超算冠军

- 2013年6月17日，由国防科学技术大学和浪潮集团共同研制的中国“天河2号”超算系统问鼎全球超算TOP500榜单，成为全球最快超级计算机。
- “天河2号” 是全球第一台峰值性能突破5亿亿次(50PFlops)的超级计算机，持续计算性能达到每秒3.39亿亿次，具备16,000节点、3,120,000个计算核心。与此前排名世界第一的美国“泰坦”系统相比，占地面积是它的85%，性能是它的两倍。

激活时间	2013年
承建商	中国国防科技大学、浪潮集团、863计划
作业管理者	中国国家超级计算中心
置放地点	中华人民共和国广东省广州市
架构	英特尔Xeon E5-2692， 英特尔Xeon Phi协处理器/运算加速卡， 麒麟操作系统
最大消耗功率	17.6MW（整机附带散热系统时为24MW）
容积、占地面积	720平方米
内部存储器	1,375TiB（1,000TiB为系统存储器，375TiB为协处理器独占）
外部存储器	12.4PB
运算速率	54.9PFLOPS（理论峰值） 33.86PFLOPS（实际峰值）
造价	一亿美元
用途	科学研究

高教超算进入百万亿次时代

- 上海交大超算中心, No.158@Top500
  - 目前高校最大的云超算中心
  - 采用CPU、GPU，Lustre等技术
  - ASC13 亚洲大学生超算竞赛东道主

- 清华大学超算中心，No.97@Top500
  - 高校第一套百万亿次集群
  - 采用CPU、GPU，Lustre等技术
  - 承办首届中国大学生超级计算机竞赛

- 居慧聪邓白氏调研，浪潮在高教行业整体占有量，排名第一
- 在整体高教TOP10系统排名中，占据40%份额
- 自2010年连续4年占据高校超算最大规模系统第1名。

"...Exascale ≠ Petascale x 1000..." Lucy Nowell, DOE

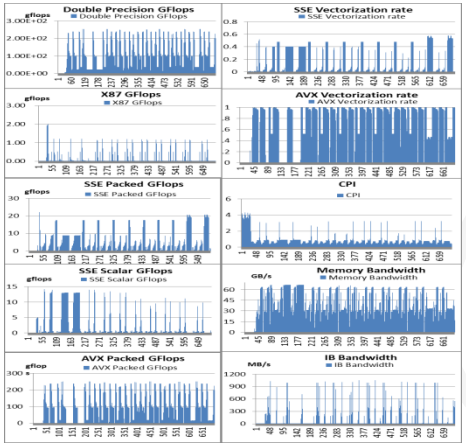
	2010	2018	Factor Change
System peak	2 Pf/s	1 Ef/s	500
Power	6 MW	20 MW	3
System Memory	0.3 PB	10 PB	33
Node Performance	0.125 Gf/s	10 Tf/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 cpus	1,000 cpus	83
Interconnect BW	1.5 GB/s	50 GB/s	33
System Size (nodes)	20 K nodes	1 M nodes	50
Total Concurrency	225 K	1 B	4,444
Storage	15 PB	300 PB	20
Input/Output bandwidth	0.2 TB/s	20 TB/s	100

DOE Exascale Initiative Roadmap, Architecture and Technology Workshop, San Diego, December, 2009

浪潮HPC应用特征自动提取器

天眼高性能应用特征提取器

科学家眼中超级计算机的速度计，让应用的性能直观、快速、简单、可见



应用软件开发

软件优化

集群性能评估

应用性能评估

浪潮HPC应用特征自动提取器

大规模集群系统支持，可提取规模大于4096 CPU物理核心特征数据

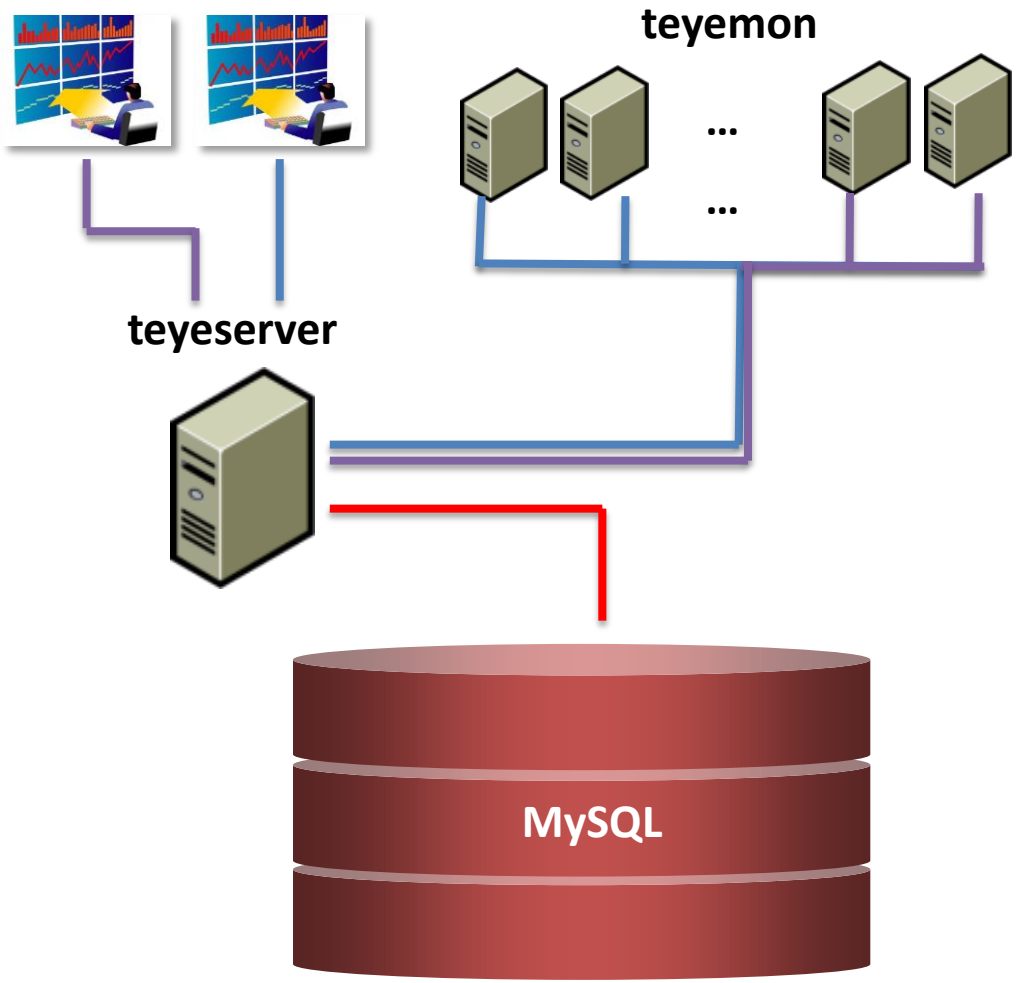
数据库支持，用户提取特征数据长久保留，以备历史查询

异步监控，不同用户可监控同一集群不同节点

低资源占用率，被监控节点资源占用低于千分之一，不影响应用运行

高特征数据提取频率，每秒刷新

软件小巧，安装使用简洁



浪潮HPC应用特征自动提取器

四十多项微架构级、系统级指标监控

处理器级

- 利用率：Usr%, sys%, idle%, iowait%
- 浮点性能：X87 GFLOPS, SP/DP SSE scalar/packed GFLOPS, SP/DP AVX scalar/packed GFLOPS
- 向量化率：SP/DP SSE VEC, SP/DP AVX VEC
- 执行效率：CPI

内存级

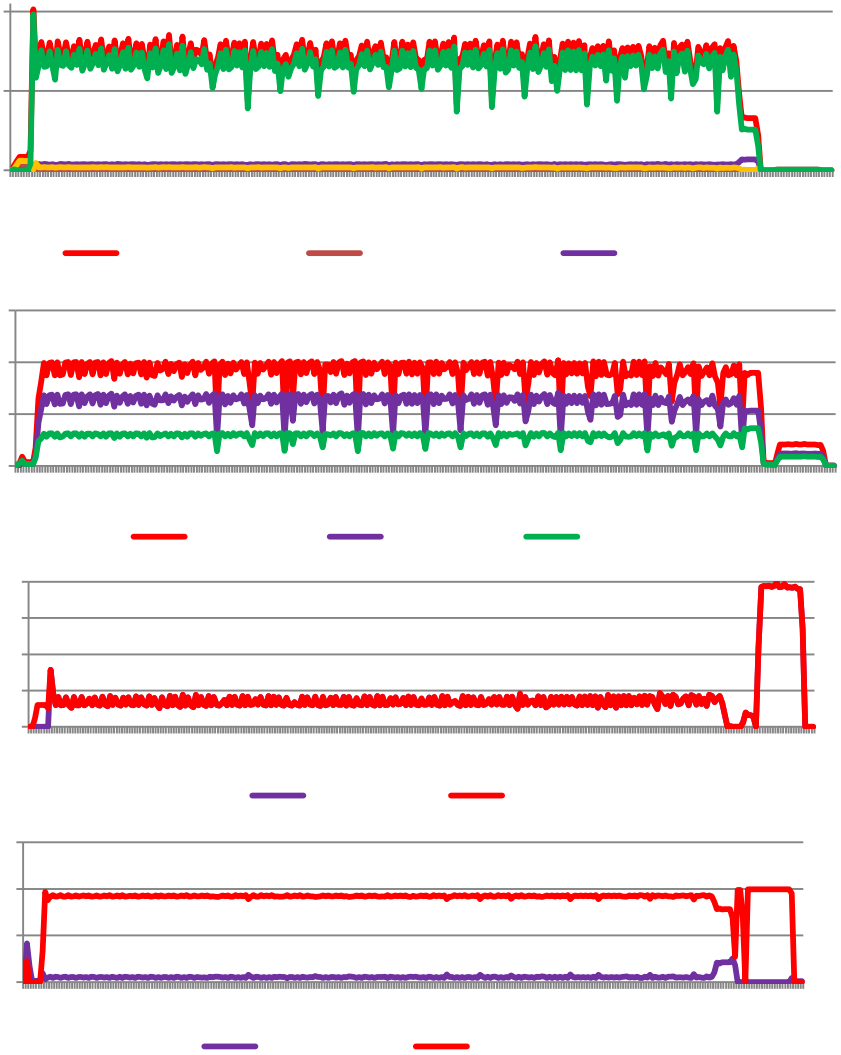
- 内存容量：总容量，used, cached, buffered
- 内存访问：内存读操作带宽、内存写操作带宽

网络级

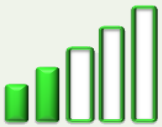
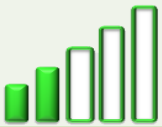
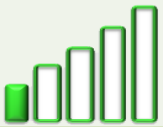
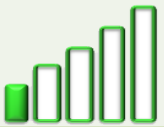
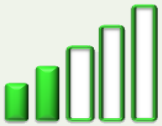
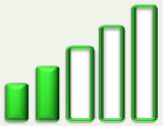
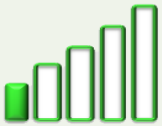
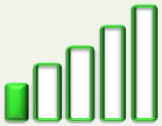
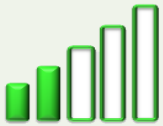
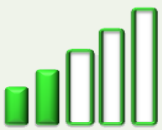
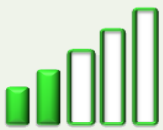
- 设备支持：Gigabit, Infiniband
- 协议支持：TCP/IP, UDP, RDMA, IPoIB
- 流量监控：千兆收、千兆发、IB收、IB发
- 包数量监控：千兆平均收/发包大小、IB收/发包数量

文件系统级

- 本地磁盘：本地读、本地写、本地读数据块大小、本地写数据块大小
- NFS文件系统：NFS客户端读、NFS客户端写、NFS服务端（总）读、NFS服务端（总）写



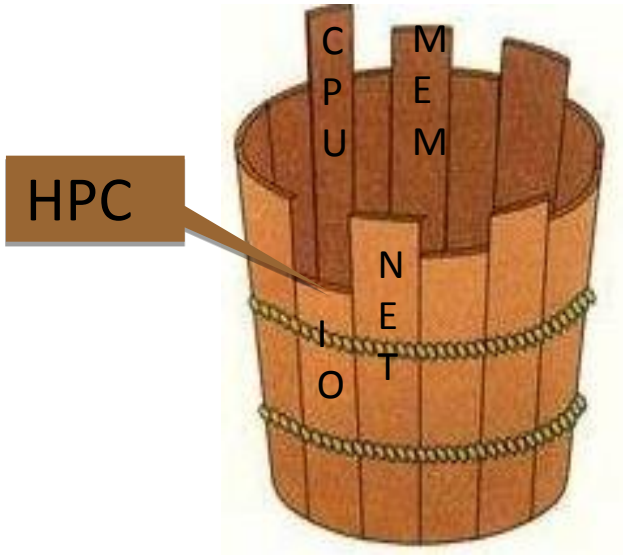
典型HPC应用特征需求

应用领域	典型应用	CPU	内存容量	内存带宽	存储	网络	扩展性
CFD	Fluent						
序列比对	BWA						
序列拼接	VELVET						
单颗粒重构	EMAN						
分子动力学	NAMD						
量子化学	GAUSSIAN						
材料科学	VASP						

注： 信号满格表示对此项需求很高

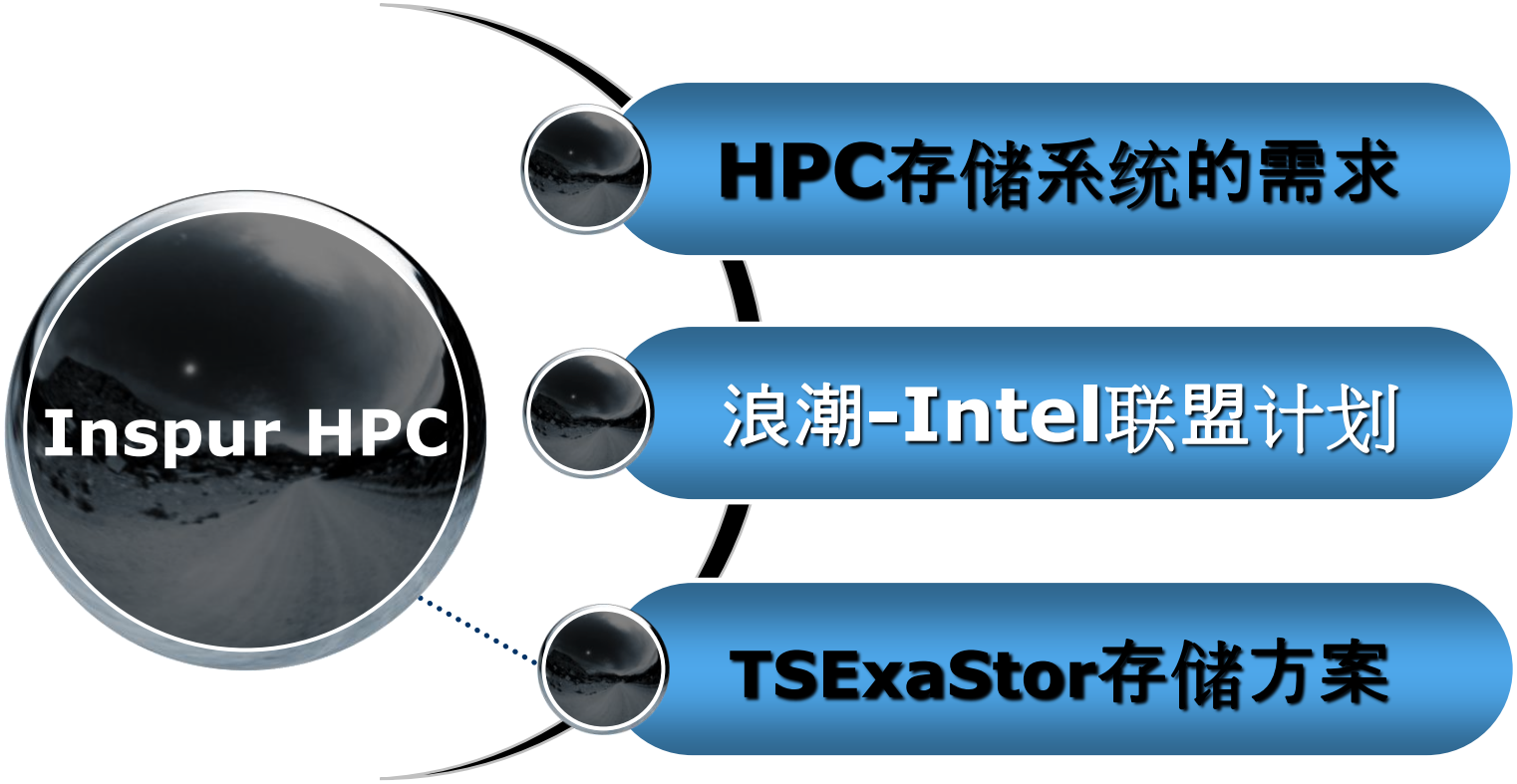
HPC系统中对于存储IO的需求

- 稳定性、可靠性要求越来越高
- 存储IO成为系统主要瓶颈
  - 预取数据量大、结果数据巨大
  - 计算过程频繁IO交互



计算模块 + 存储模块 + 网络模块 + 软件系统 + 基础环境 = HPC系统

Content



合作背景

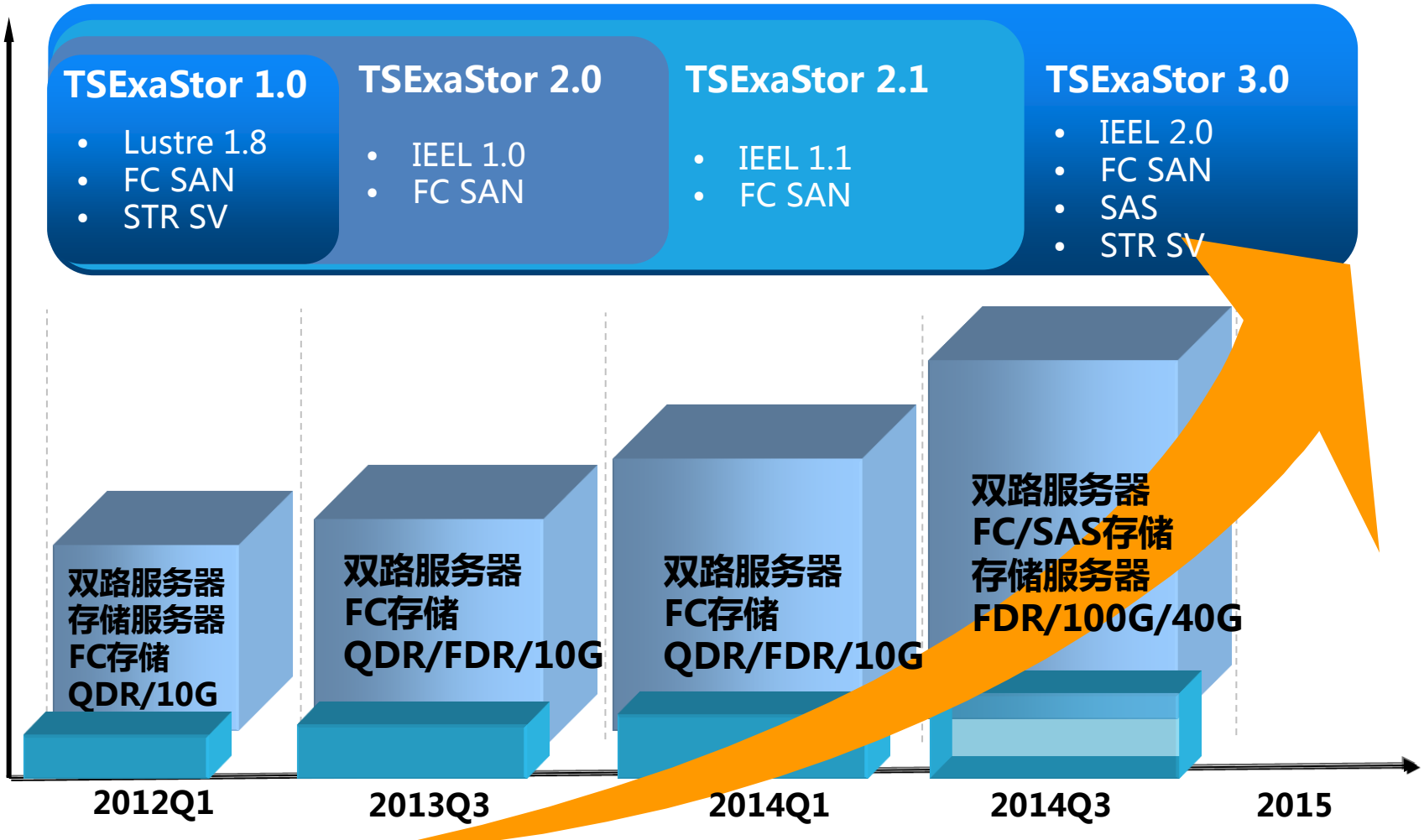
- HPC集群对于存储的高可靠、高带宽、高IOPS需求迫切；
- 浪潮拥有丰富的HPC系统研发经验；
- 浪潮拥有自主研发的服务器、存储等硬件产品；
- Intel一直致力于HPC系统建设的推进工作；
- IEEL是非常适合于高性能计算的集群并行文件系统；
- 浪潮与Intel一直保持着良好的合作伙伴关系。

浪潮-Intel联盟高效能HPC存储推进计划

- 目标：打造适合高性能应用的存储一体化系统,解决HPC对于存储的高可靠、高带宽、高IOPS需求；
- 产品推进：基于IEEL构建TSExaStor高效能HPC存储系统；
- 推进计划：
  - 搭建DemoCenter，为各行业用户提供测试平台；
  - 浪潮团队进行系统的实施与后续服务工作；
  - Intel团队提供深度的技术支持；

- 双方参与人员
- 浪潮产品、系统研发、硬件研发、项目实施人员；
- Intel IEEL相关产品、研发等人员

TSExaStor产品RoadMap



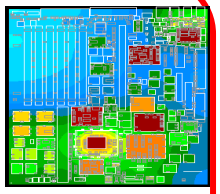
TSExaStor产品适合应用领域

云计算中心/
省市计算中心

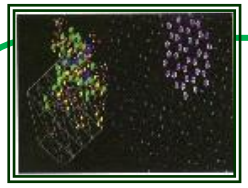
金融分析

影视渲染

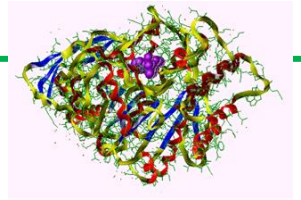
基因信息



芯片设计



生物物理

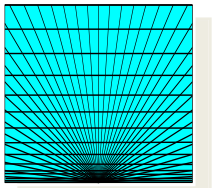


药物设计

航空航天

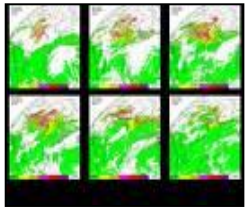
国防军事

船舶制造

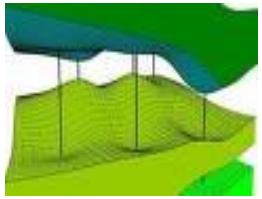


数值计算

汽车设计



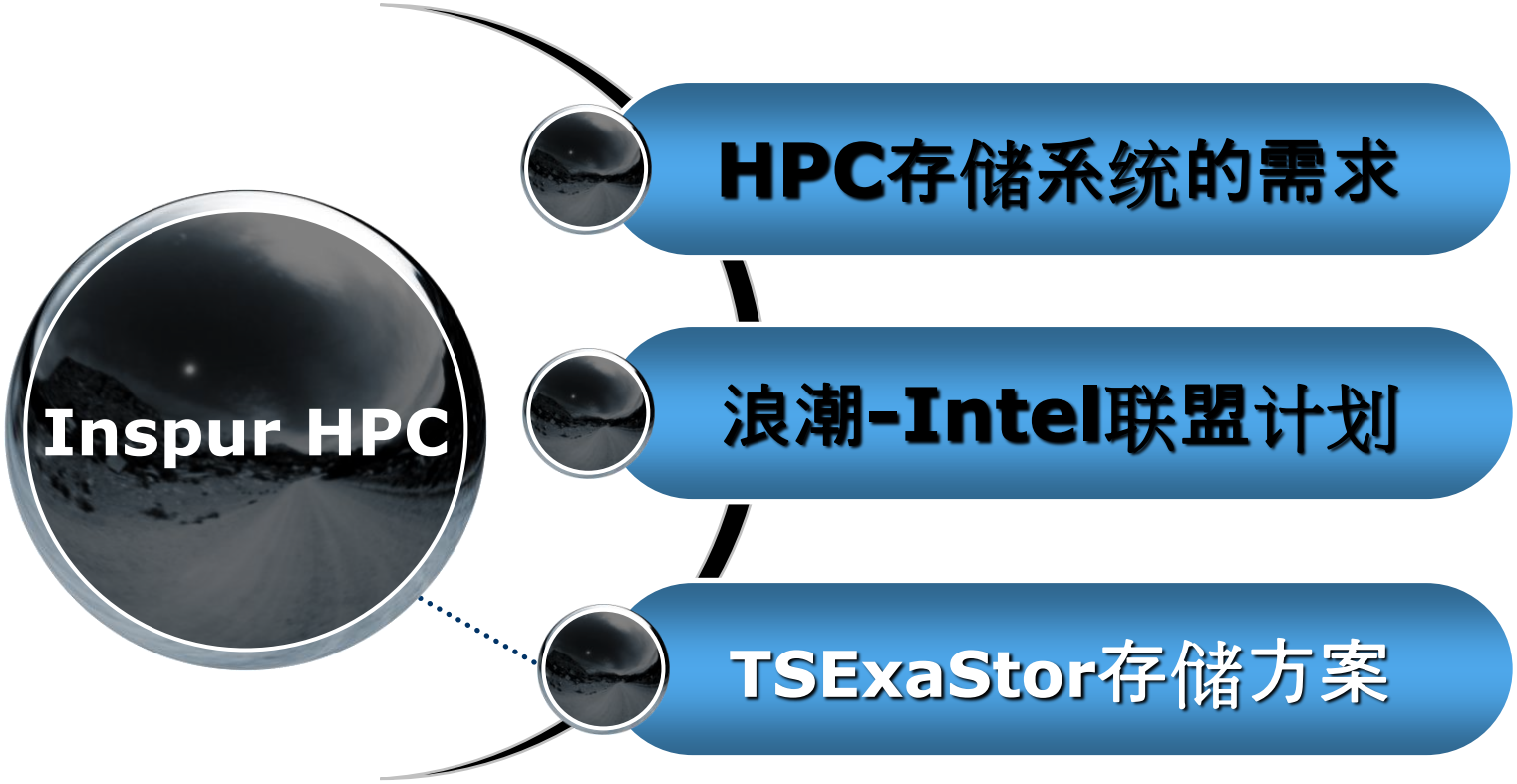
气象预报



石油勘探

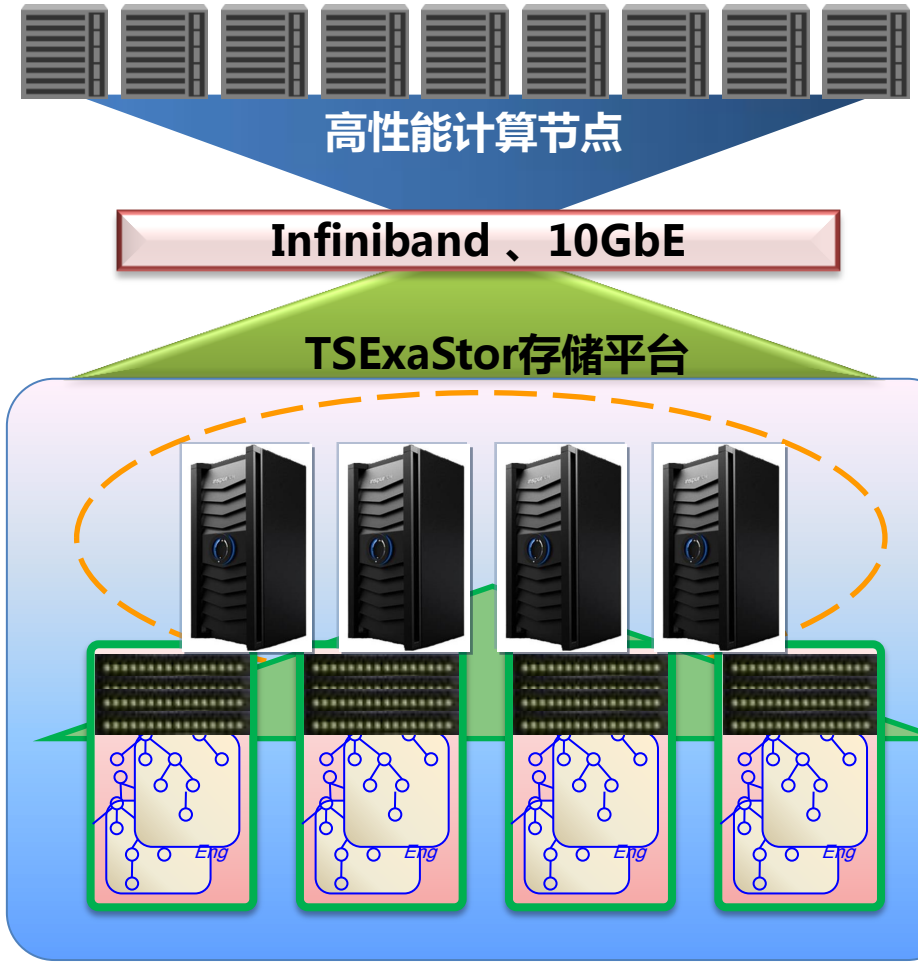
生命科学

Content



TSExaStor存储产品-产品定位

- 面向高效能HPC存储IO需求，提供软硬一体化解决方案；
- 横向扩展型（Scale-Out）存储；
- 扩展能力：扩展大于128个IO；
- 海量存储空间支持：支持PB级存储容量，十亿级文件数量；
- 并发访问能力：支持数十GB聚合带宽，可支持数千并发访问；
- 数据传输能力：IB、万兆网络传输

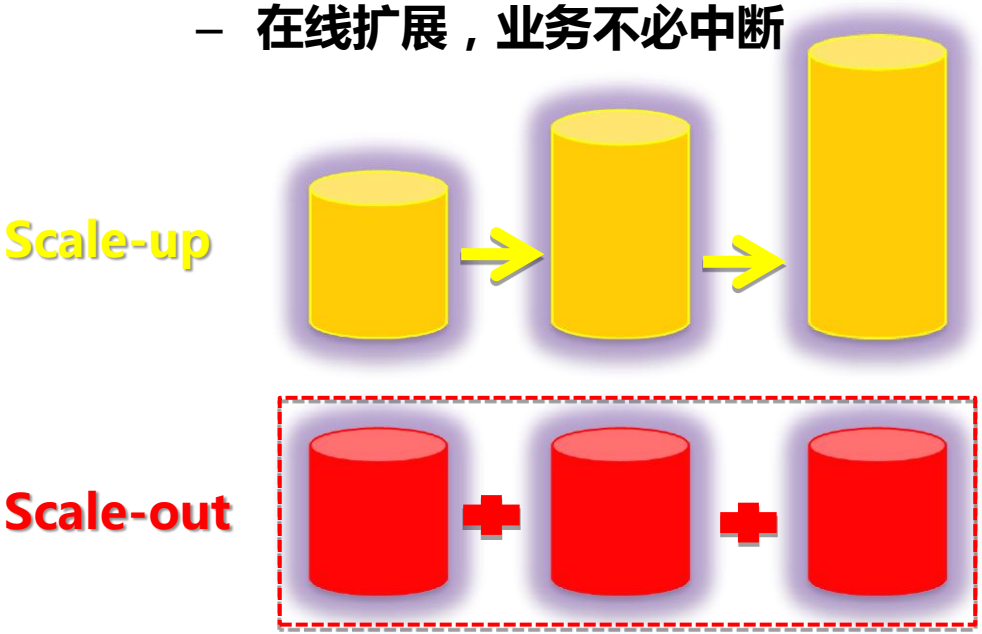


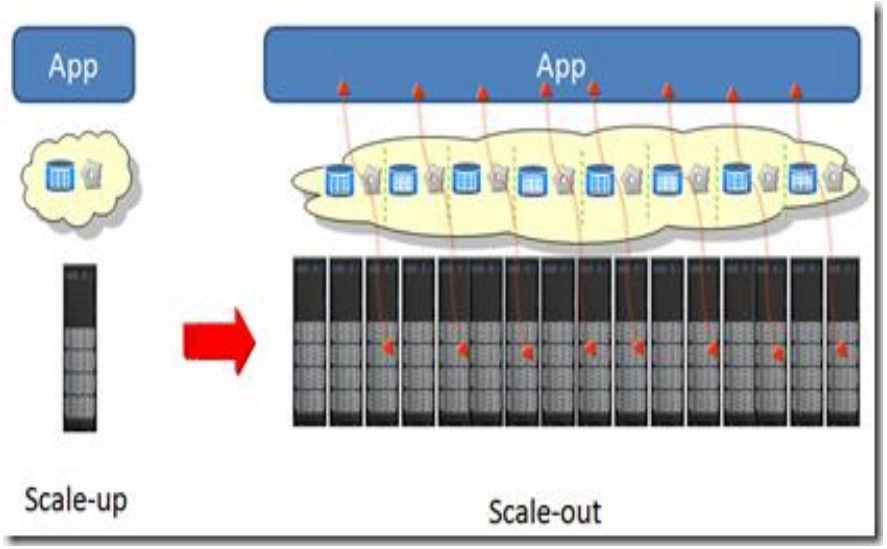
TSExaStor存储产品-主要规格

- 系统主要规格如下
  - **PB级存储空间，最大可扩展至64PB**
  - 文件总数支持**10**亿量级，单目录可有效支持千万量级
  - 支持扩展到**128**个**IO**控制器
  - **FDR IB/QDR IB/10GbE**主机接口
  - 支持**Web**、**CLI**管理方式
  - 支持图形化性能及状态监控
  - 支持快速部署
  - 支持故障预警、邮件、**SNMP**通知等功能
  - 冗余配置，允许任一硬件故障

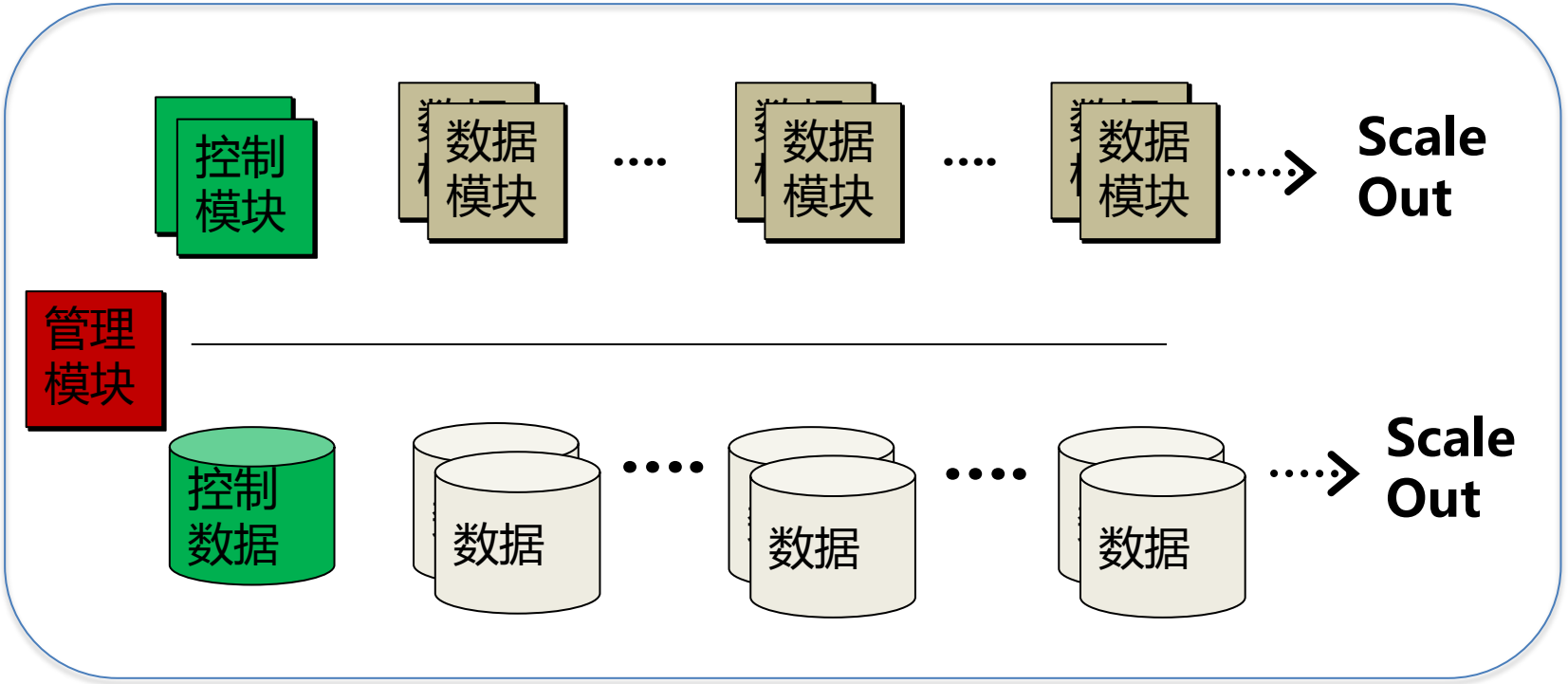
TSExaStor存储产品-系统架构

- **基于全新Scale-Out架构设计，摒弃Scale-UP架构缺点**
  - **容量和性能：通过增加功能模块进行无限扩容**
  - **可用性：软硬件层次全冗余设计，消除单点故障**
  - **在线扩展，业务不必中断**



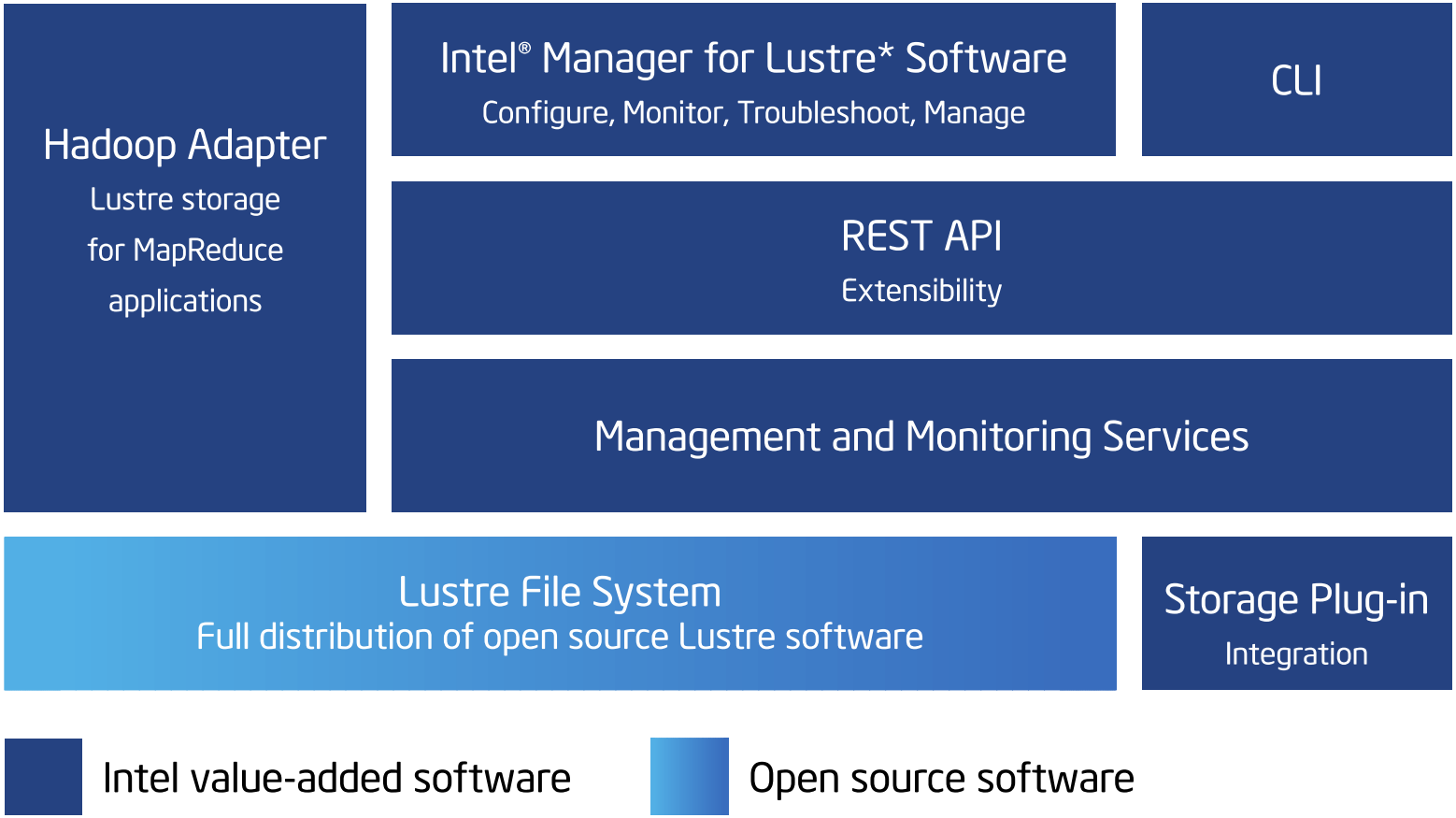


TSExaStor存储产品-系统架构



- 高度模块化设计：管理模块、控制模块、数据模块、存储模块相分离

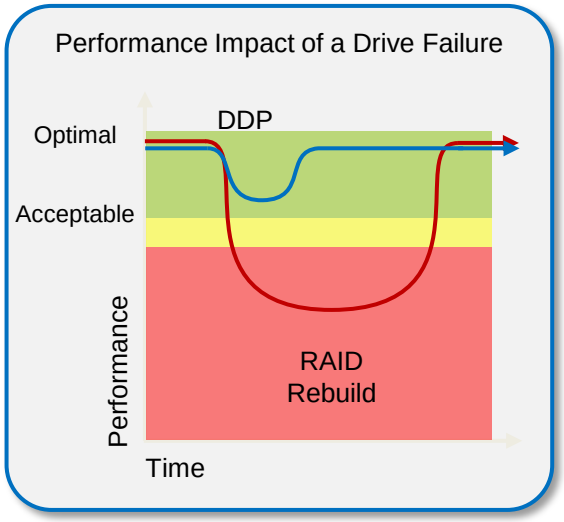
TSExaStor存储产品-软件架构



TSExaStor特色硬件技术：动态磁盘池DDP技术

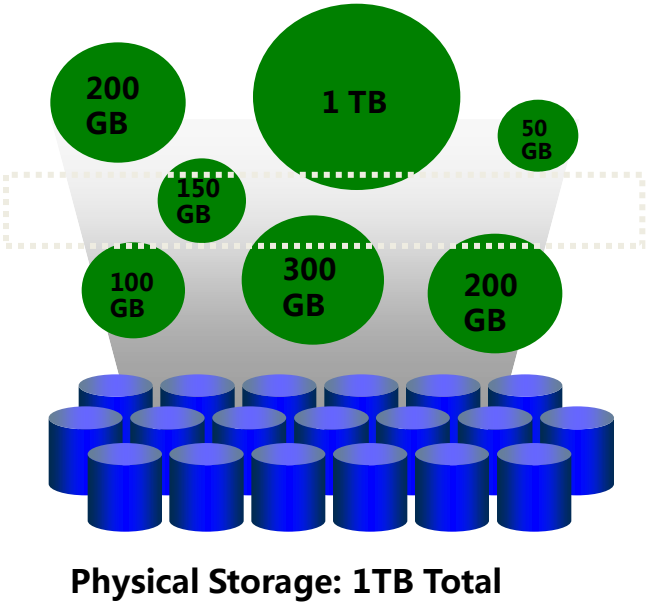
动态磁盘池DDP

- **系统提供持续不受影响的性能**
- **系统性能保持在 “绿色区域”**
  - **硬盘故障对系统的性能影响最小**
  - **显著加快系统恢复时间**
  - **10倍于传统RAID的恢复速度**
  - **加速数据重建**
- **磁盘池规避硬盘热点**
  - **所有的卷空间分布在磁盘池中全部的硬盘中**
  - **降低硬盘故障率**
- **动态的数据分布和再分配由后台持续进行**



TSExaStor特色硬件技术：自动精简

自动精简



价值优势：

允许用户创建灵活卷 ，以大于实际物理空间的容量，可在后期追加物理容量，使容量规划的效率更高

不浪费空间，仅当写入数据才占用空间，降低存储系统的采购成本

节省电能和机房空间，降低热量的排放，高效低碳

用最低的成本存储最多的数据

TSExaStor特色硬件技术：SSD缓存加速技术

SSD硬盘做缓存

描述：

基于控制器的读缓存使用SSD，可扩展至5T；

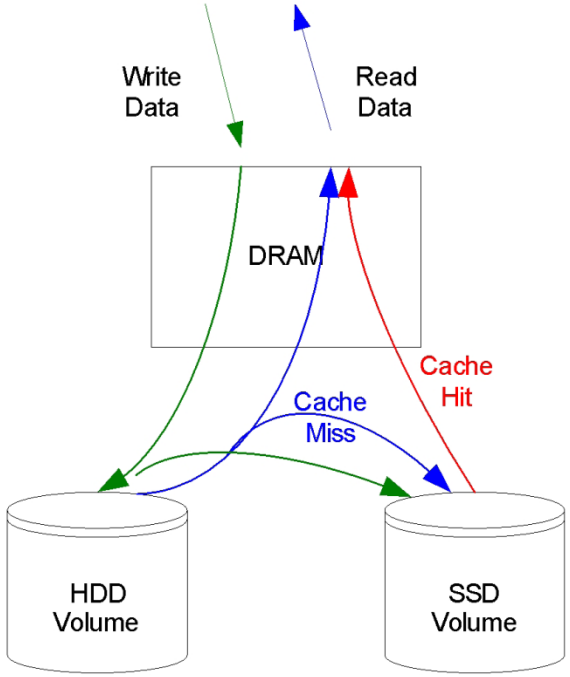
减少驱动器的数量，满足IOPS性能要求，显著提高应用程序的读取性能；

优势：

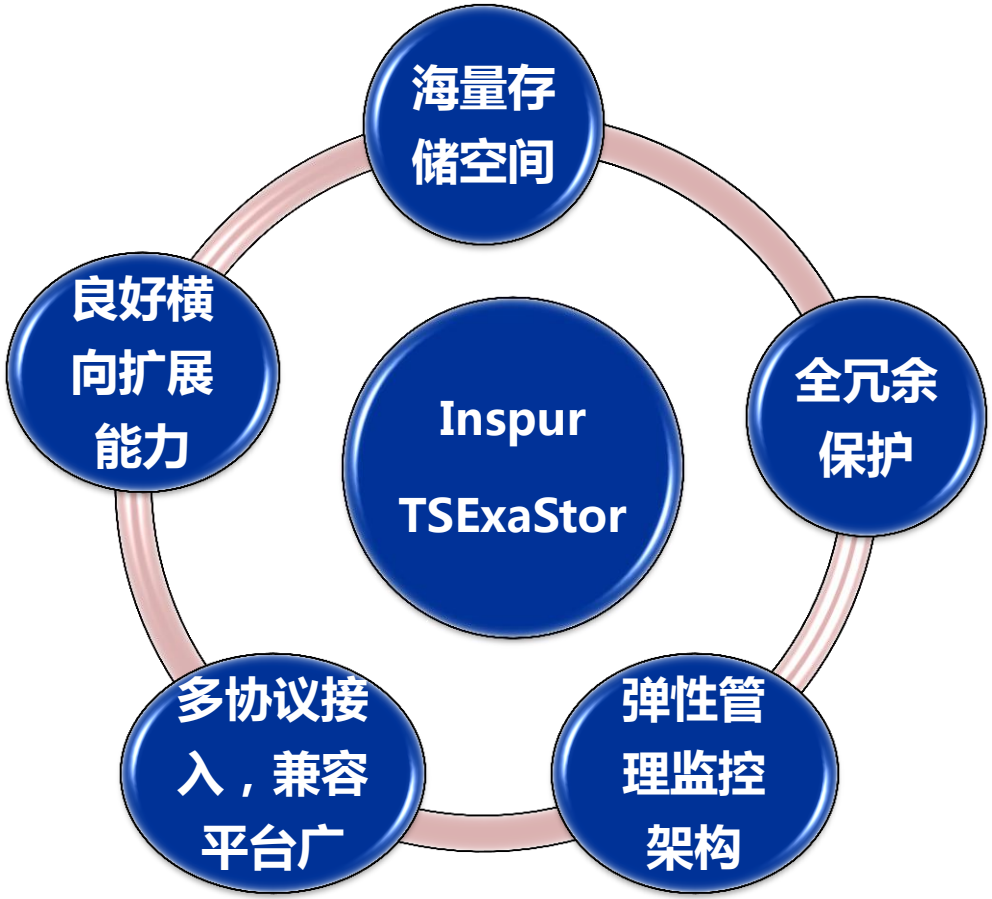
自动为热点数据进行分层管理；

通过SSD硬盘改善存储效率及成本；

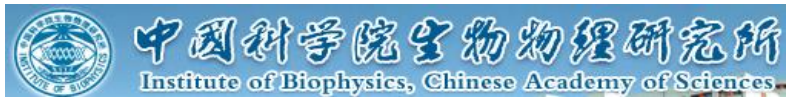
通过SSD硬盘配置提高IOPS，提高随机数据读速率；

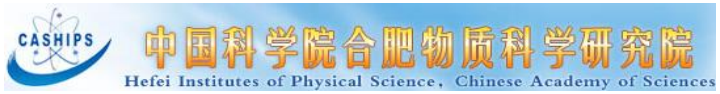


TSExaStor存储产品-产品特色总结



部分浪潮Lustre存储方案用户





Thank You !