inspur 浪潮

Lustre在HPC的实践与优化

王渭巍

浪潮 系统架构师

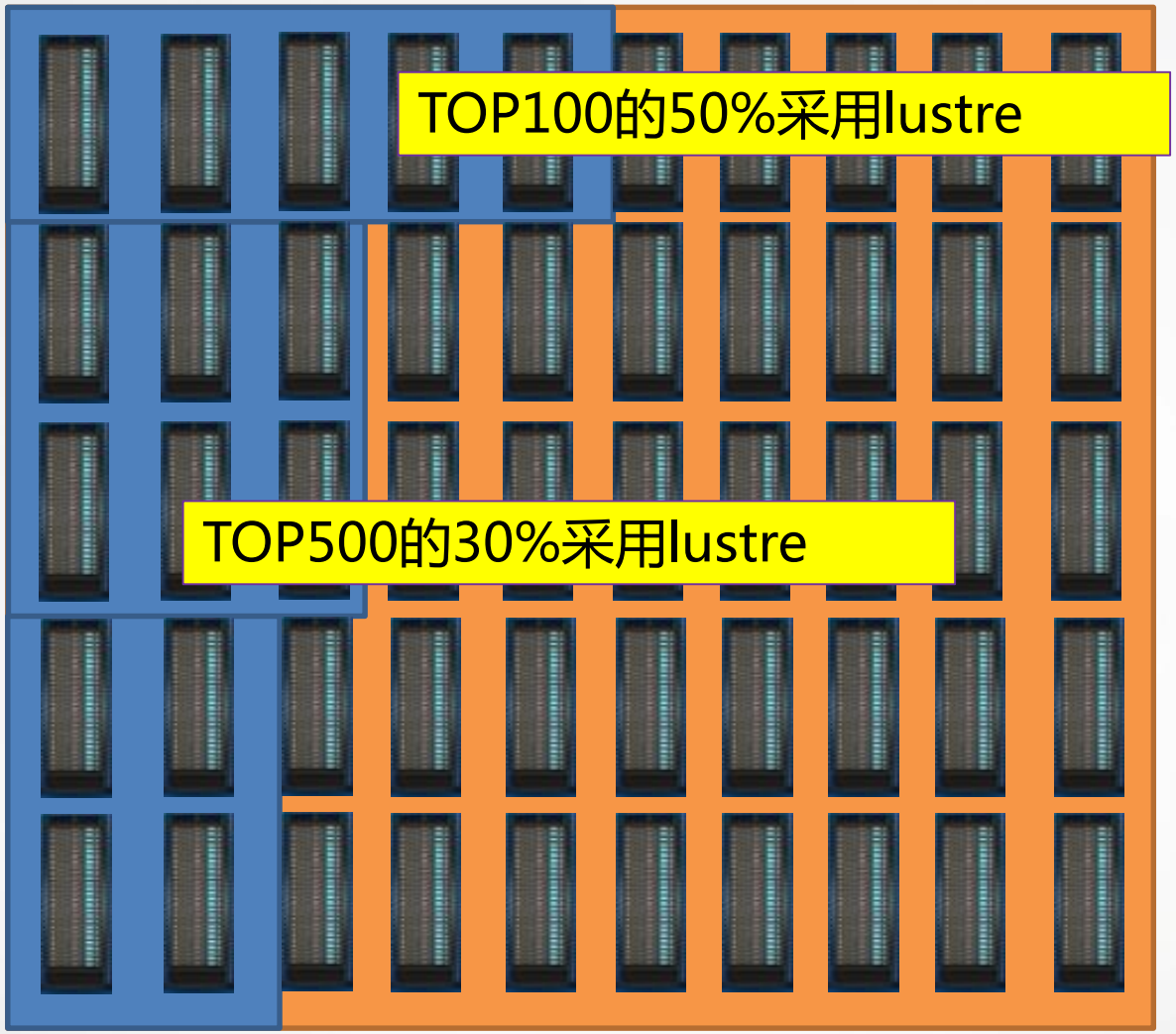
目录

1 Lustre的生产实践

2 Lustre的性能评估

3 Lustre的系统优化

Lustre文件系统广泛应用于HPC



- 存储系统是整个HPC系统的瓶颈
  - 高并发，容量大，带宽高，延迟低
  - Checkpoint要求瞬时带宽高
  - 随着系统规模增大，对存储系统的稳定性，可靠性要求越来越高
- Lustre在HPC领域应用广泛
  - 高效的并行文件系统
  - 开源，硬件无关，性能卓越
  - Lustre生态环境日益完善

Lustre如何应用于生产实践

作为开源软件，
Lustre应用于生产实践
还有很多工作要做

- 部署安装
- 管理维护
- 性能优化

大型

SKA(平方公里阵)

SKA的数据存储需求：

12Tb/s，12小时×30天产生的数据量为1944PB；如果采用目前容量最大的6T硬盘来存，需要324000块。一块6T硬盘的价格为3738元（亚马逊）， 324000块硬盘的总价格高达12.1亿。

大型 小型

某遥感卫星数据处理（试算）

研究人员的数据存储需求：

无力购置专业存储硬件

共10台服务器节点，每个节点两块硬盘

构建一个全局并行文件系统，提升IO带宽

提炼需求，设计方案

- 输入数据、输出数据和可视化数据在哪里？
- 这些数据如何共享
- 数据文件的大小

存储的硬件考虑

元数据服务器
- 高性能
- 低延迟

存储阵列
- 高带宽
- 低延迟

数据服务器
- 高带宽

如何部署Lustre

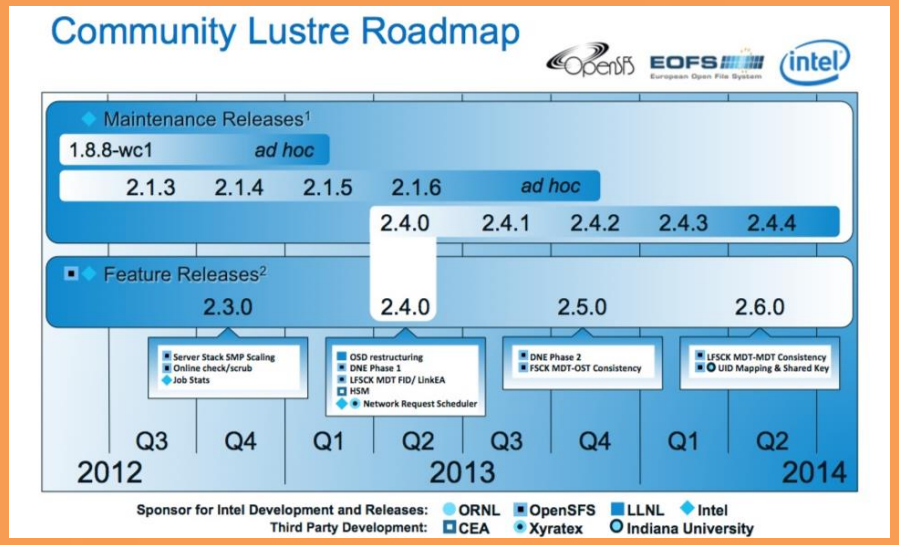
OS	存储网络	计算网络	存储阵列
• Kernel • 本地文件系统 • 驱动	• FC-SAN • IB-SAN	• TCP • IB	• 双控还是单控 • 是否采用本地硬盘

使用和维护Lustre

- Lustre的稳定版本
  - 1.8
  - 2.x
  - 根据需要选择
- 维护buglist
- 结合硬件使用与维护

2014

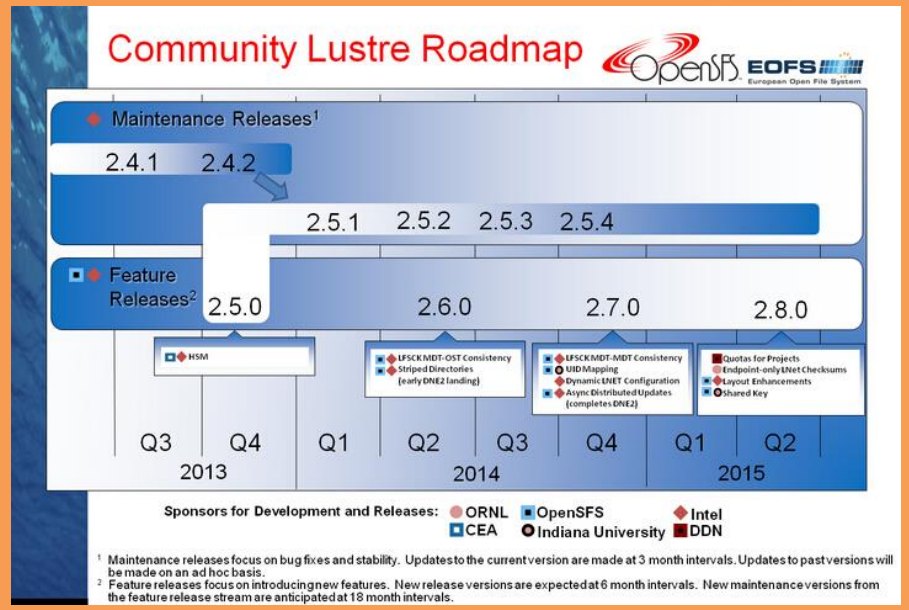
Lustre的有益特性



1.ZFS
2. Online check
3. Job status
4. Etc.

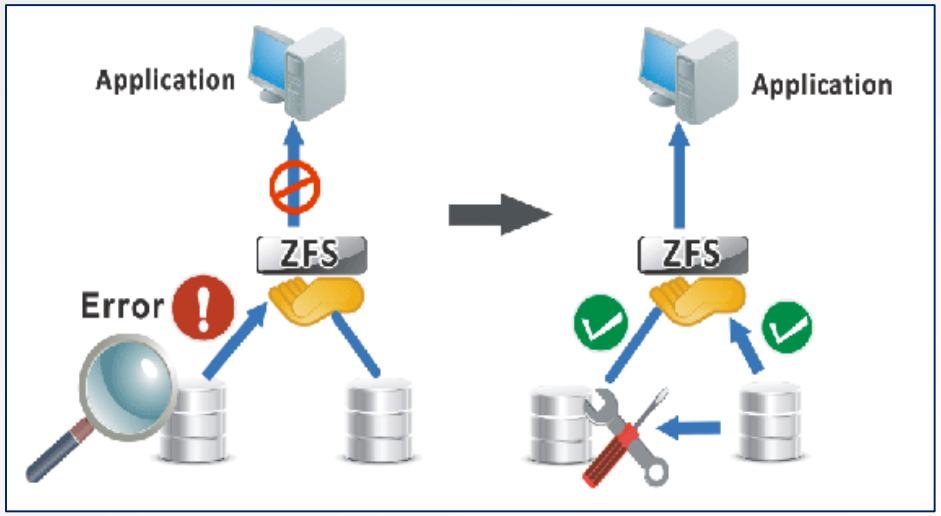
2014 2015

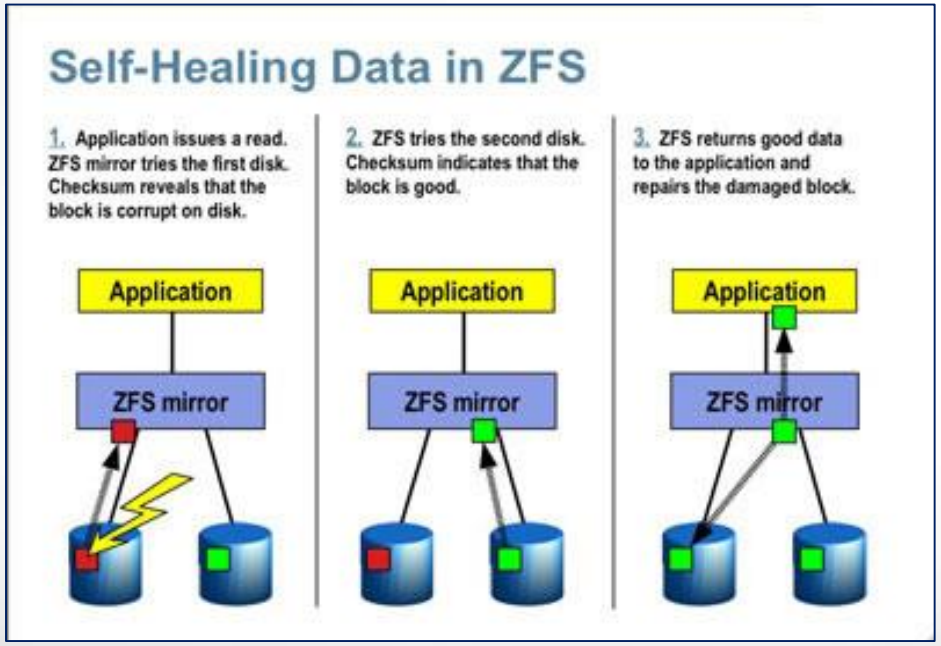
Lustre的有益特性



1. HSM
2. Etc.

强大的本地文件系统与Lustre结合



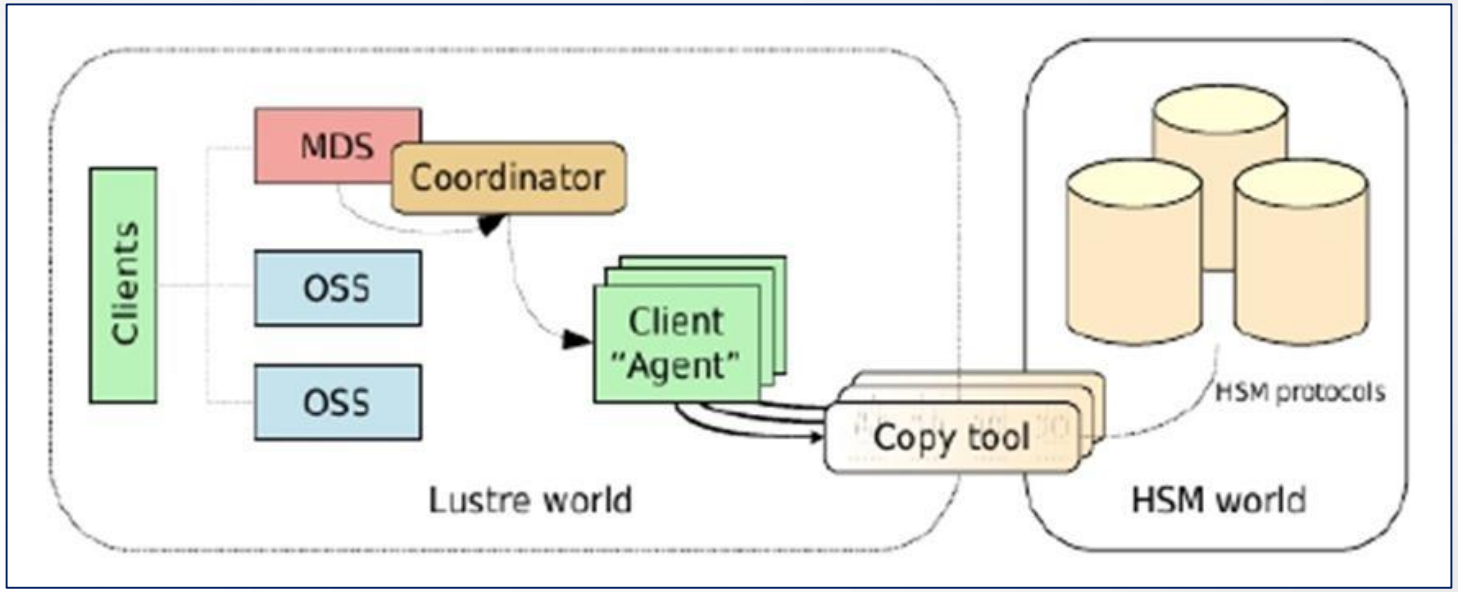


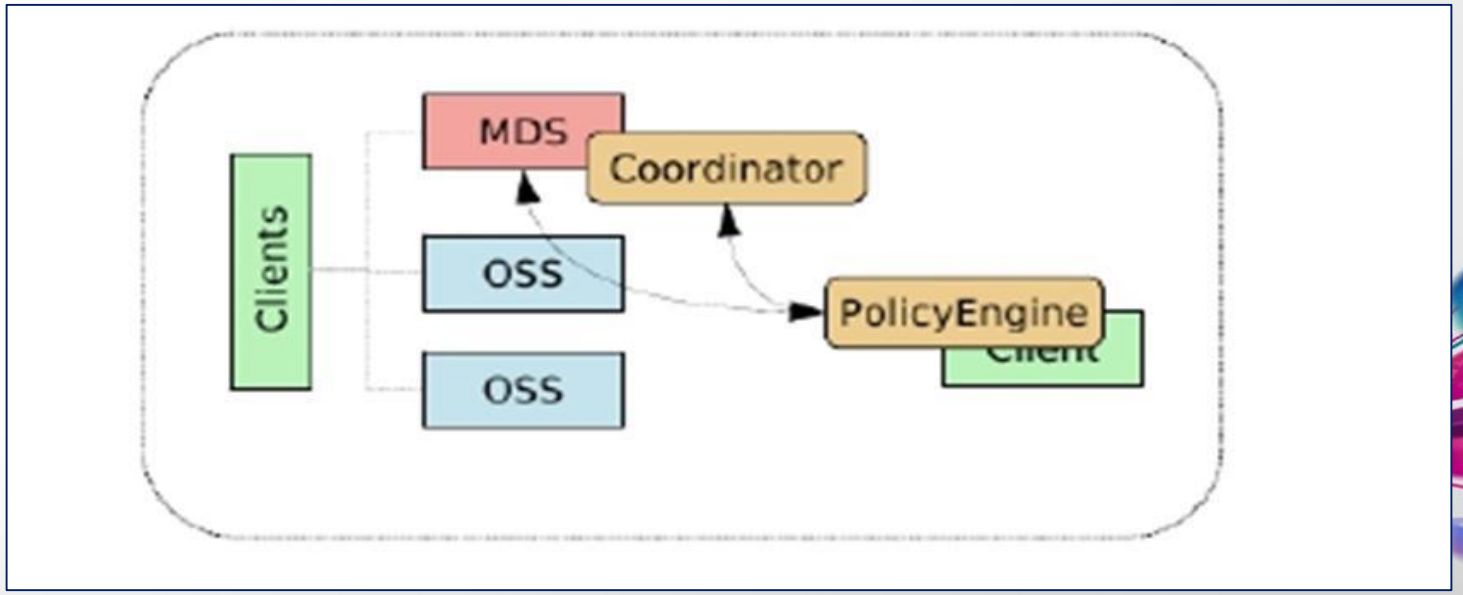
- ZFS 池存储
  ZFS 使用存储池的概念来管理物理存储。ZFS 可完全避免使用卷管理。
- 事务性语义
  ZFS 是事务性文件系统，这意味着文件系统状态在磁盘上始终是一致的。文件系统绝对不会因意外断电或系统崩溃而被损坏。
- 校验和与自我修复数据
  所有数据和元数据都通过用户可选择的校验和算法进行验证。ZFS 还会提供自我修复数据。
- 独一无二的可伸缩性
  该文件系统本身是 128 位的所有元数据都是动态分配的，因此在首次创建时无需预先分配 inode，否则就会限制文件系统的可伸缩性。
- ZFS 支持快照
- 简化的管理

Lustre & HSM

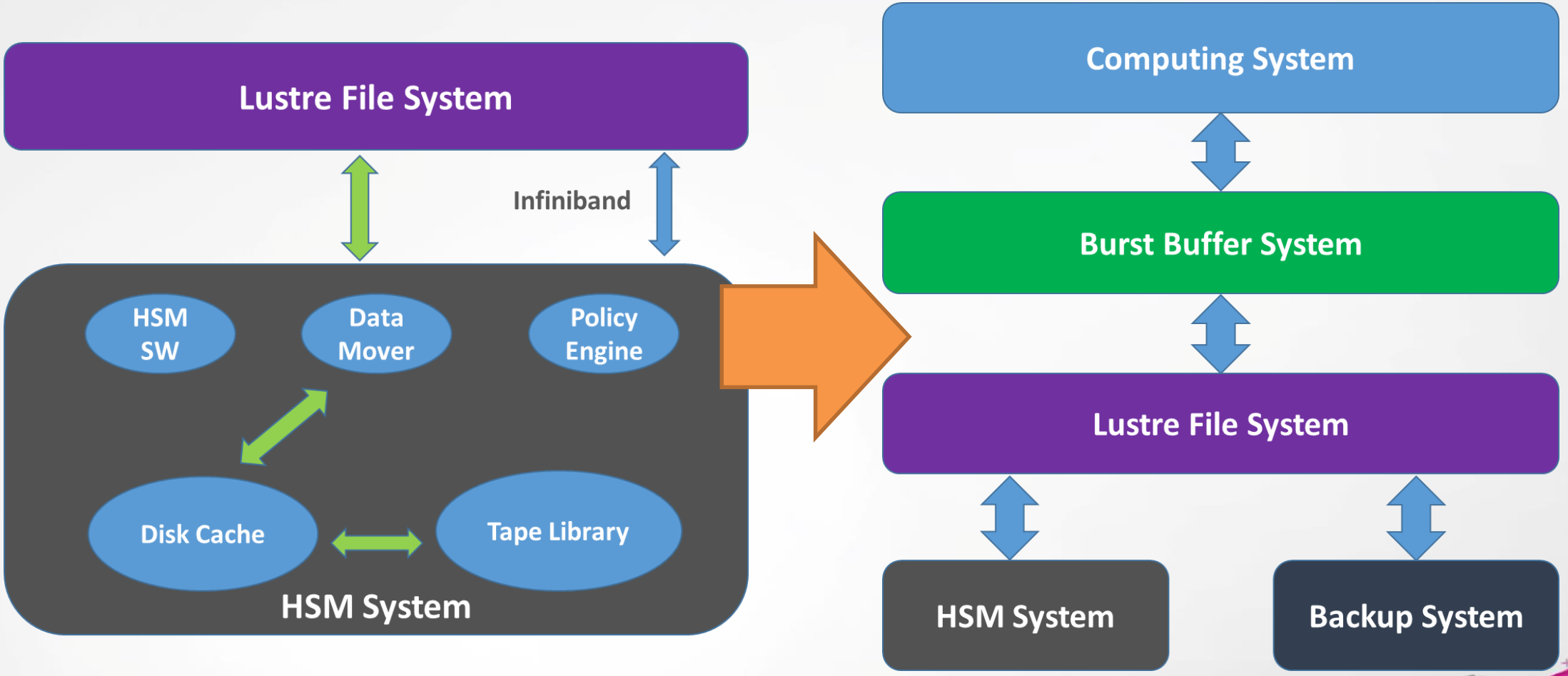
- HSM Architecture
- Coordinator 收集archiving需求并分分配agent节点
- Agent 指特殊的客户端，上面运行着copytool，在lustre和hsm之间迁移数据

- Policy Engine
- *robinhood-hsm*管理迁移策略和清理策略

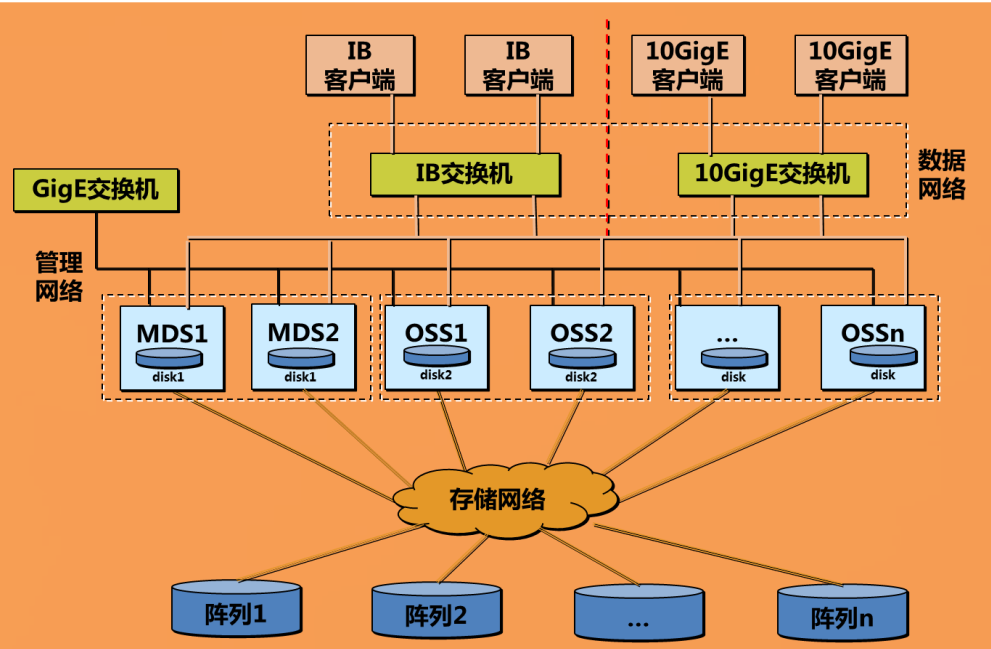




一个基本完善的Lustre应用方案

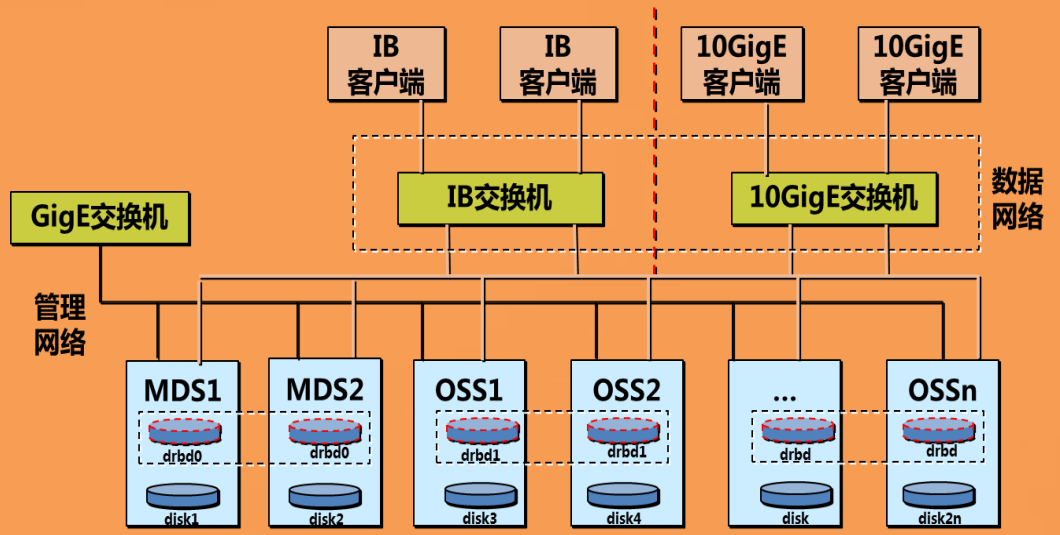


SKA(平方公里阵)


IB
客户端
IB
客户端
10GigE
客户端
10GigE
客户端
IB交换机
10GigE交换机
数据
网络
GigE交换机
管理
网络
MDS1
disk1
MDS2
disk1
OSS1
disk2
OSS2
disk2
...
disk
OSSn
disk
存储网络
阵列1
阵列2
...
阵列n

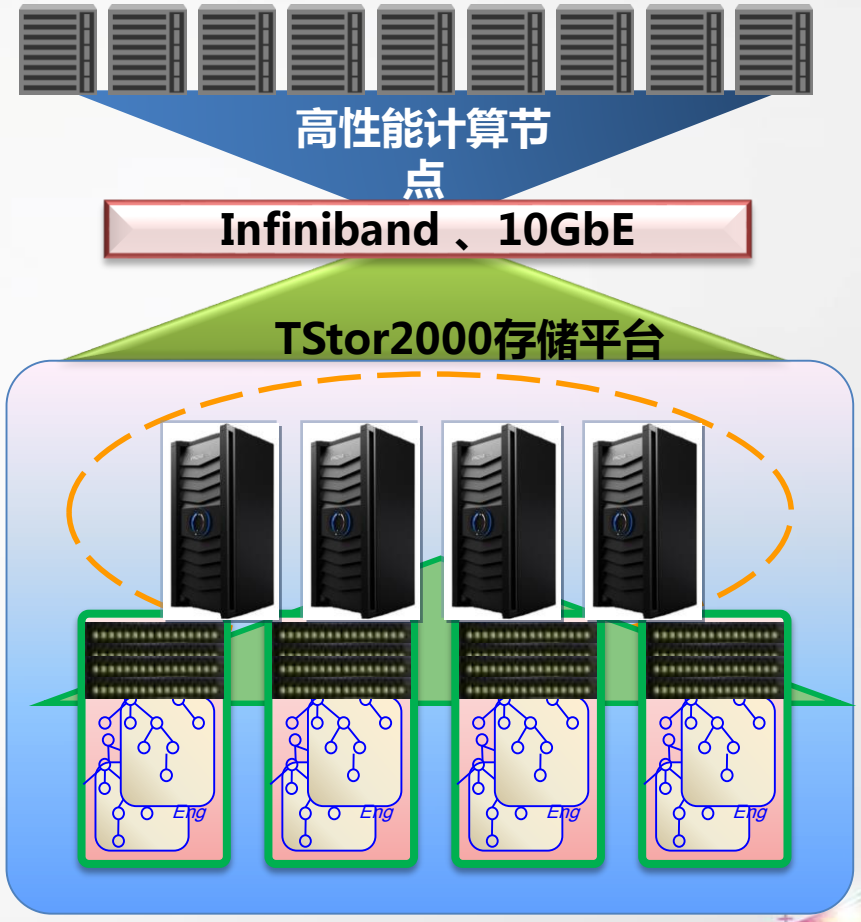
大型 小型

某遥感卫星数据处理（试算，研究性能）



TStor2000-基于IEEL的Inspur Lustre最佳实践

- 基于IEEL v2.x
- 面向HPC用户提供软硬一体化解决方案；
- 性能、容量在线按需扩展；
- 支持海量存储空间，高达512PB级存储容量，十亿级文件数量；
- 高达2.5TB/s并发访问能力，可支持数千并发访问；
- Active-Active高可靠冗余架构
- GUI统一监控、管理、部署
- IB、万兆网络数据传输
- 融合HPC与大数据计算应用
- All in one solution



目录

1 Lustre的生产实践

2 Lustre的性能评估

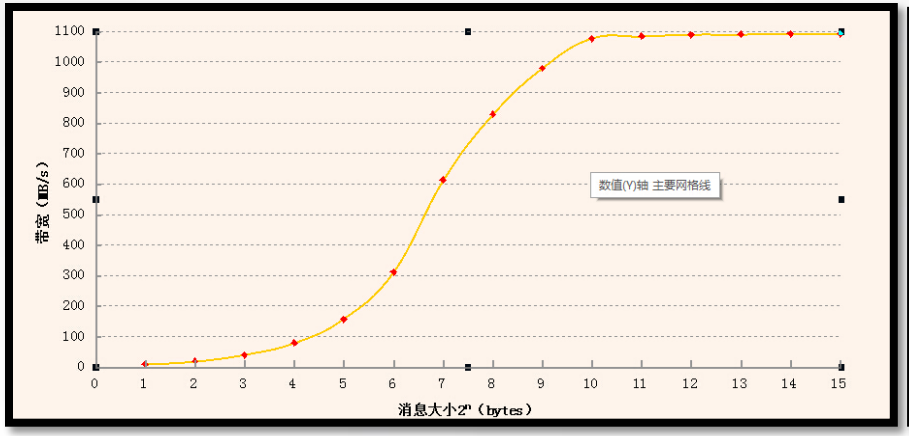
3 Lustre的系统优化

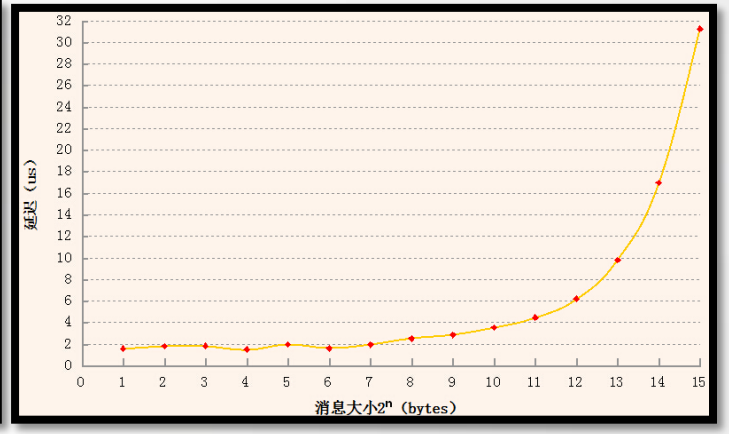
Lustre的性能评估

裸设备	文件系统	应用测试
• 网络 • 存储	• 带宽，IOPS • 元数据	• 具体应用具体分析

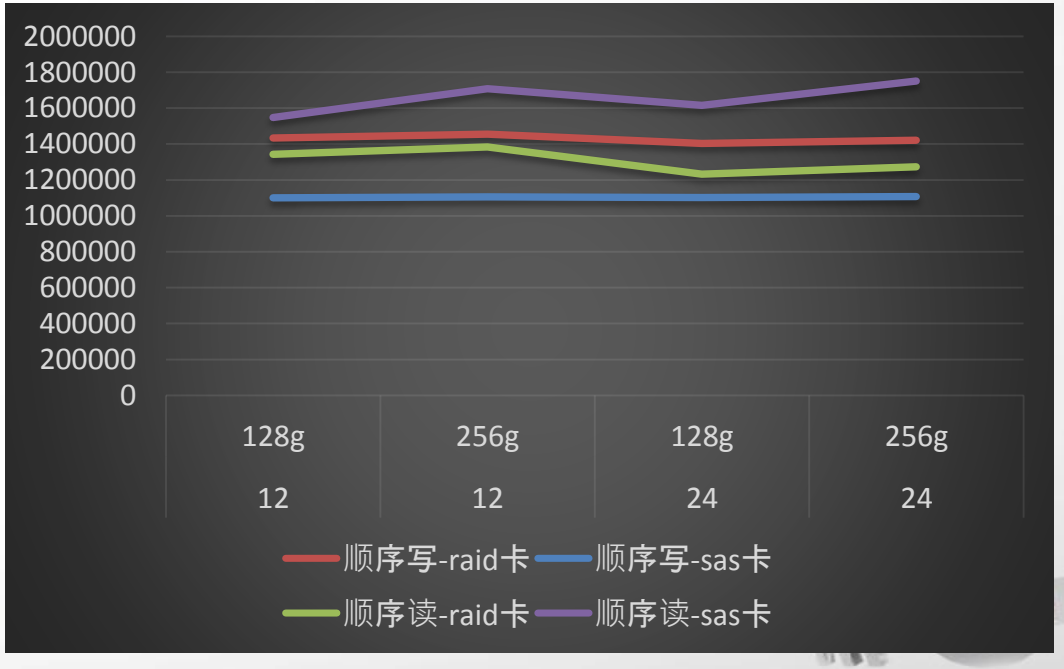
Lustre的性能评估之裸设备测试

- 网络性能测试
  - Iperf
  - Ib_write_bw
  - Ib_read_bw
- 确保带宽和延迟正常





- 存储设备性能测试
  - 硬盘的类型：SATA，SAS，SSD
  - RAID卡，sas卡
  - RAID类型，RAID参数
  - 存储阵列，缓存设置，多路径策略
- 确保存储设备裸设备性能正常



Lustre的性能评估之文件系统测试

带宽测试

- 测试工具
  - Iozone
  - IOR
  - 测试参数的设置
    - 块大小
    - 线程数
    - 测试模式
    - …

TIPS

- 条带设置
- 每个ost逐个测试
- 性能监控工具

	Transfer Size			
	4 KB		1 MB	
Projected performance at full scale	Max Write	Max Read	Max Write	Max Read
POSIX File-Per-Process				
POSIX Shared File				
MPI/IO File-Per-Process				
MPI/IO Shared File				
HDF5				

Lustre的性能评估之元数据测试

- 元数据性能测试
  - 测试工具
    - mdtest
    - 测试参数的设置
      - 文件数量
      - 测试模式
      - 进程数
  - 性能监控工具

文件/目录数	进程数	Directory creation	Directory removal	File creation	File removal
199936	1	1298.274	1394.526	925.236	1295.24
	2	2106.233	2291.21	2007.347	1536.416
	4	2011.61	1491.463	1771.859	1550.27
	8	2039.049	1499.855	1835.466	1515.984
	16	1855.025	1857.192	1887.362	1601.704
	32	2039.595	2035.842	1902.935	2140.435

Lustre的性能评估之应用IO性能评估

分类	模式	介绍
大气数值模式	WRF	美国国家大气研究中心（NCAR）研制的新一代用于有限区域气象预报模拟的中小尺度数值模式
	MM5	由美国宾州大学和NCAR联合开发的用于有限区域气象预报模拟的中小尺度数值模式，目前正在向wrf做替换
	GRAPES	中国气象科学研究院研制的新一代全球/区域多尺度统一数值预报模式
大气环流模式	GAMIL	中国科学院大气物理研究所大气科学和地球流体力学数值模拟国家重点实验室（LASG）研制的大气环流格点模式
	SAMIL	LASG研制的大气环流谱模式
气候模式	CCSM	NCAR研制的Community Climate System Model共同气候系统模式
	RegCMS	美国NCAR研制的区域气候及全球季节预测模式
海气耦合模式	LICOM	LASG研制的气候系统海洋模式
	ROMS	中尺度海气耦合模式
空气质量模式	CMAQ	美国国家环保局（EPA）第三代空气质量模式(Models-3)，气象模式系统、排放模式系统以及公共多尺度模式系统
	RAMS	多尺度空气质量模式
暴雨预报模式	AREM	LASG研制的对中国区域暴雨有较强预报能力的暴雨数值预报模式

Lustre的性能评估之应用IO性能评估

减少IO占时比例！

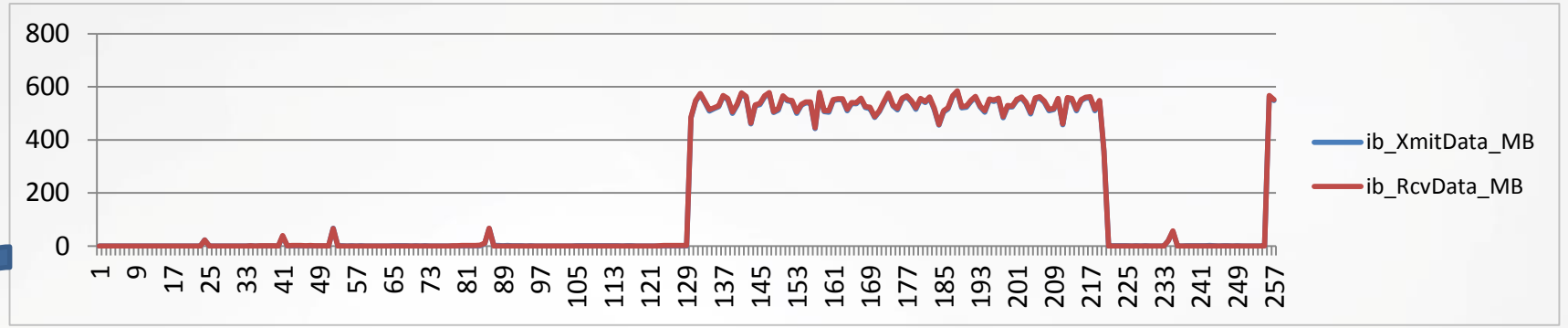
分析： 对WRF运行的IO方面优化主要有以下三种途径：

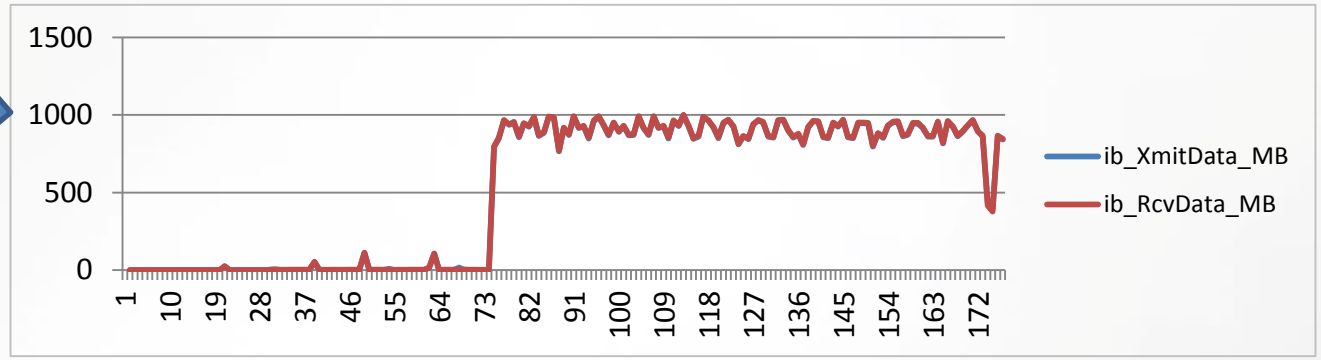
- Lustre性能：提升IO速度
- Quilt IO：将IO时间隐藏
- Pnetcdf：提升IO机制

		后	节省时间
初始读网格、分配任务	130s	75s	**175s**
wrfout写结果	40s	1-2s	
Nudge写过程	100s	24s	

Lustre的性能评估之应用IO性能评估

优化IO





Lustre的性能评估之应用IO性能评估

一些一般的调整思路

- Lnet debug关闭
- 条带设置
- Ost pool 设置
- 客户端缓存
- max_rpcs_in_flight

目录

1 Lustre的生产实践

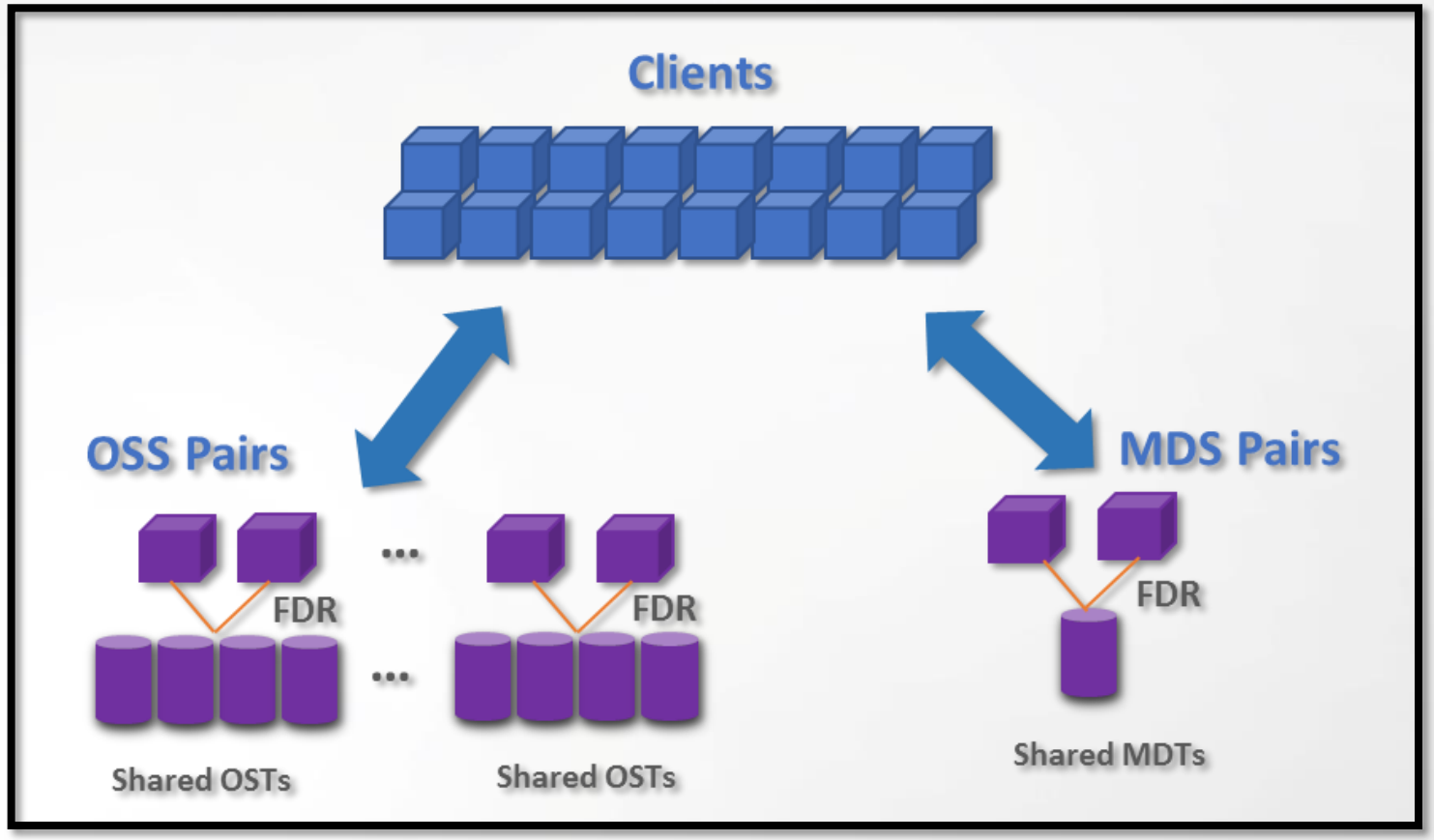
2 Lustre的性能评估

3 Lustre的系统优化

Lustre自动部署脚本

Lustre双机方案

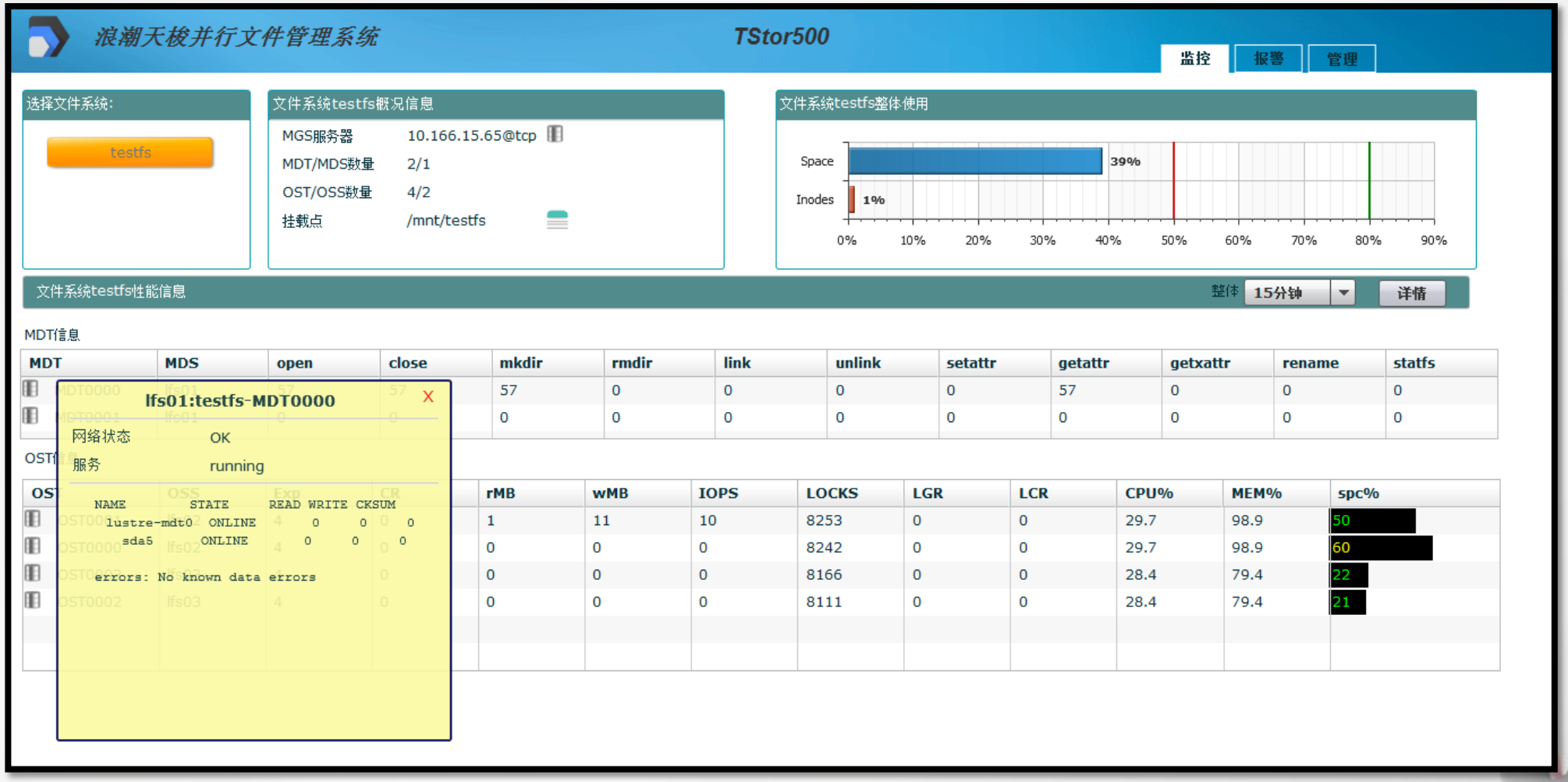
- HA软件是一个集群资源管理者。他用资源级别的监测和恢复来保证集群服务的最大可用性，可以管理以下资源：ping，ipmi，node，磁盘，文件系统…
- Active-Active



Lustre图形Web监控管理

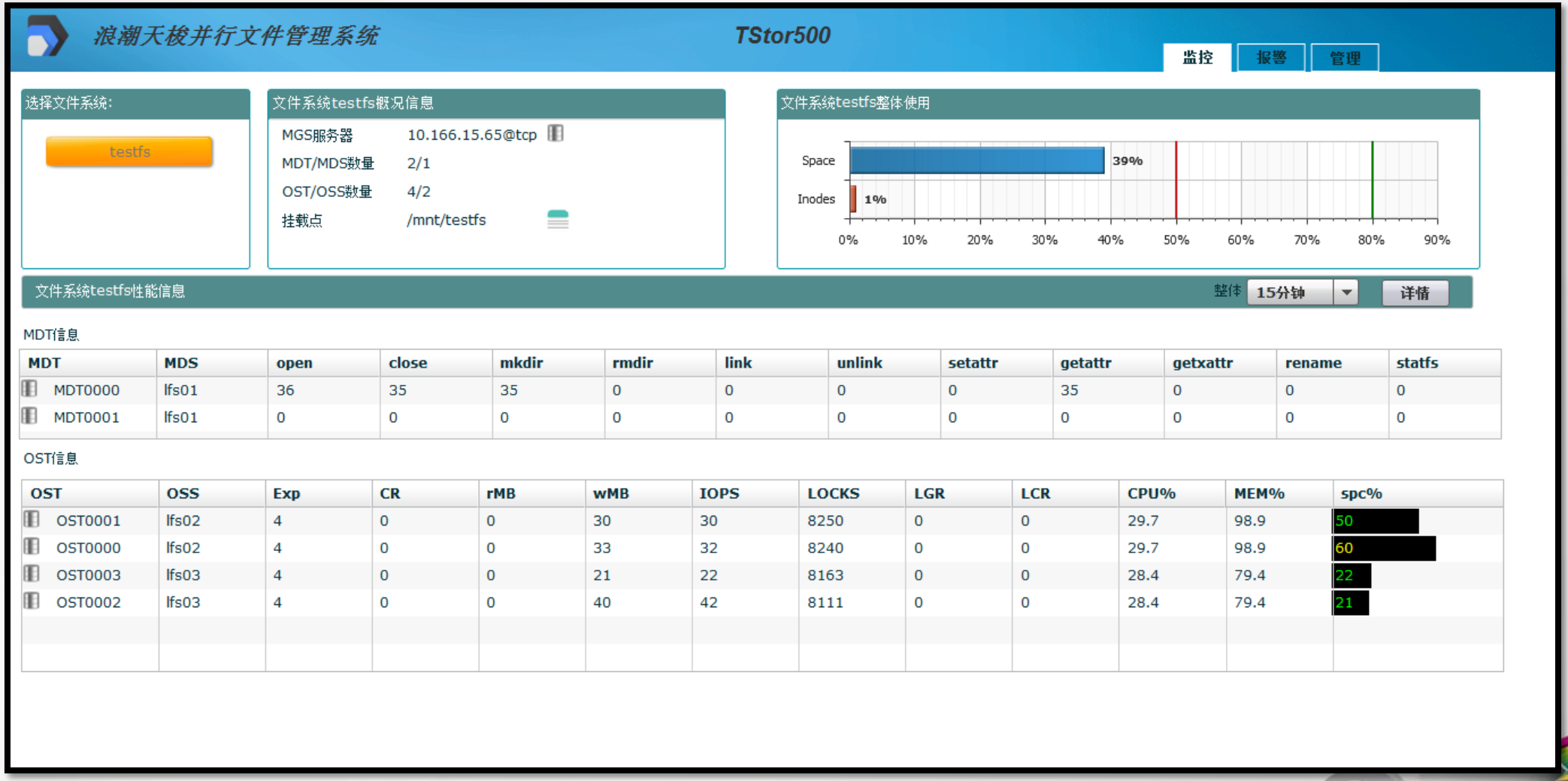
系统整体界面

Lustre图形Web监控管理



Target观察界面

Lustre图形Web监控管理



详细监控界面

Lustre图形Web监控管理

告警界面

inspur 浪潮

THANKS