Intel® Enterprise Edition for Lustre*

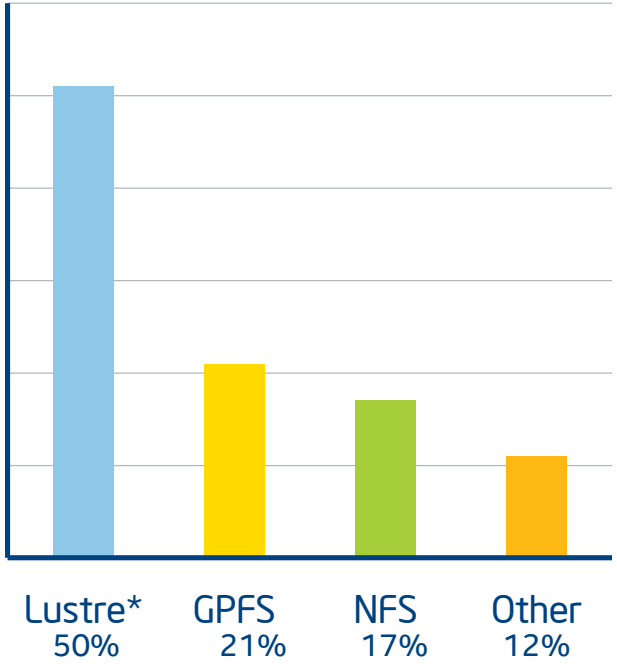
Zhang HongChao (hongchao.zhang@intel.com)
High Performance Data Division, Intel ® Corporation

日程

- 英特尔Lustre*企业版
- 英特尔Lustre集成化管理环境(IML)
-英特尔Lustre企业版推荐实施配置

HPC领域使用的主要文件系统

- Over 50% rely on Lustre* today….
- Over double the share of GPFS
- NFS performance and scaling limits appeal



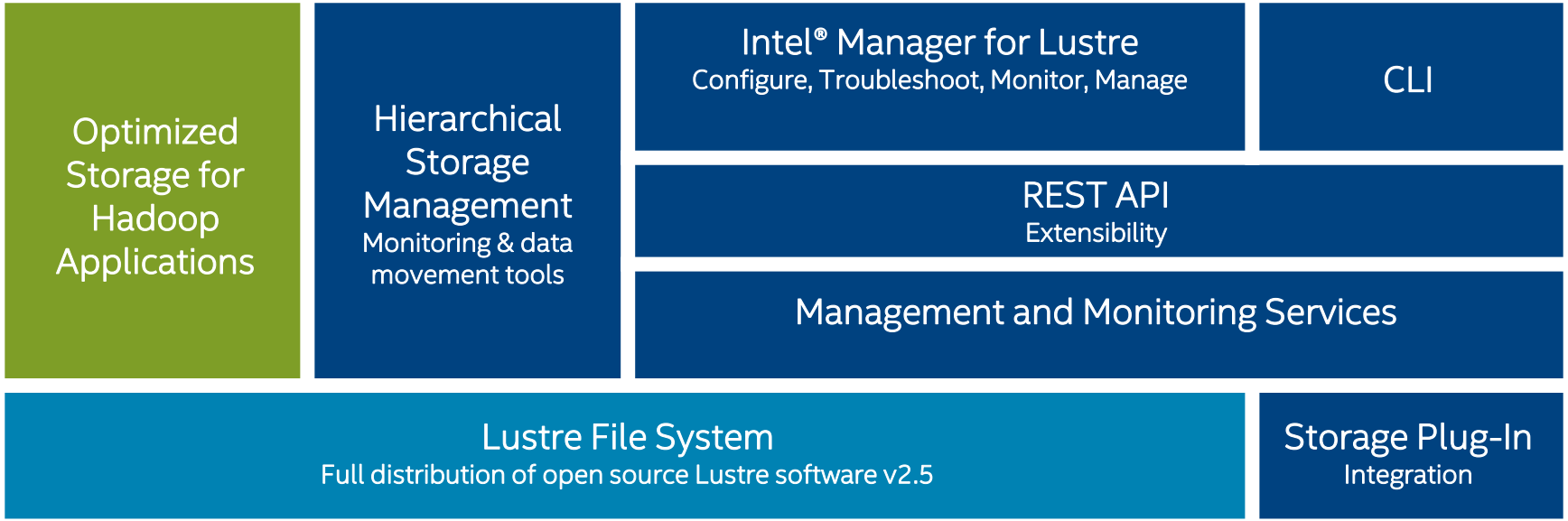
IDC survey research, May 2014

* Other names and brands may be claimed as the property of others

英特尔Lustre*企业版用户

- 能源，石油和天然气勘探
  - Schlumberger (北京)，Total/Fina, Shell, BP, Chevron
- 政府部门和教育科研机构
  - 上海天文台，北京防灾大学，北方交通大学，CEA, IU, ORNL, LLNL, ANU, SDSC, Purdue University, Harvard, Stanford
- 金融服务
  - 邮储银行，Morgan Stanley, Central Bank of Italy, National Bank of Poland
- 生命科学，制药，基因科学
  - Sanger Institute, Lund University, Cleveland Clinic
- 天气预报和气候建模
  - Australia Weather Bureau

英特尔Lustre*企业版 V2



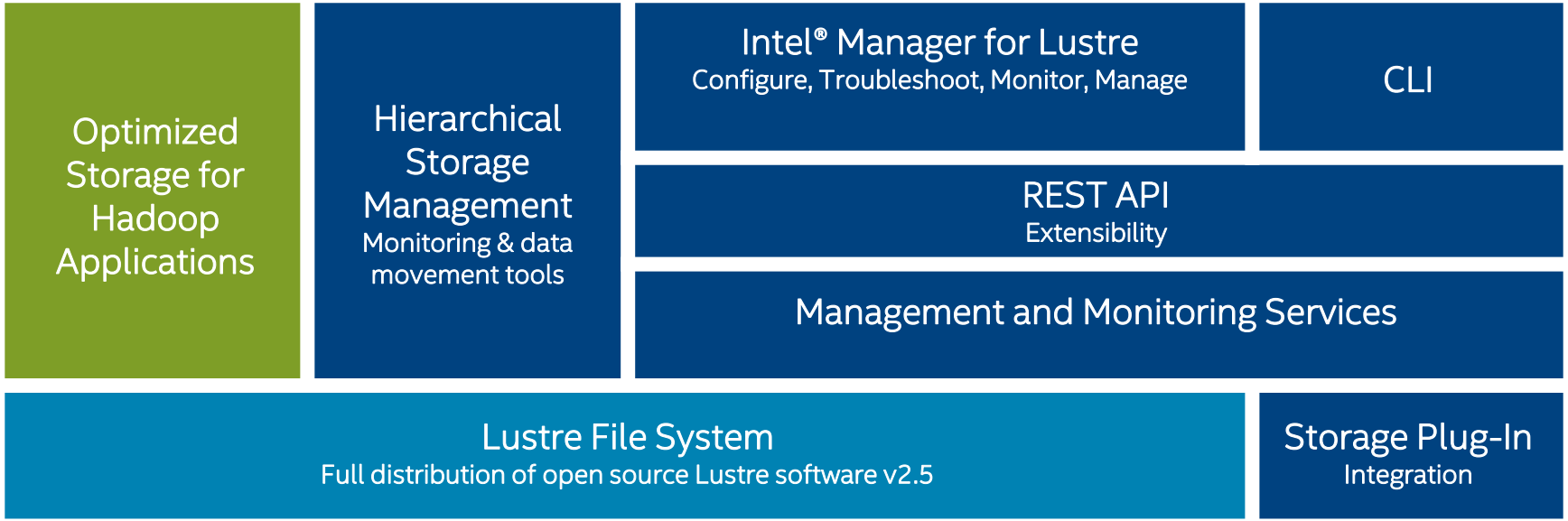
Open source base

Intel value-add for Lustre

Interoperability with Hadoop distributions for fast, shared, simple to manage storage for MapReduce applications

* Other names and brands may be claimed as the property of others

英特尔Lustre*企业版 V2



Open source base

Intel value-add for Lustre

Interoperability with Hadoop distributions for fast, shared, simple to manage storage for MapReduce applications

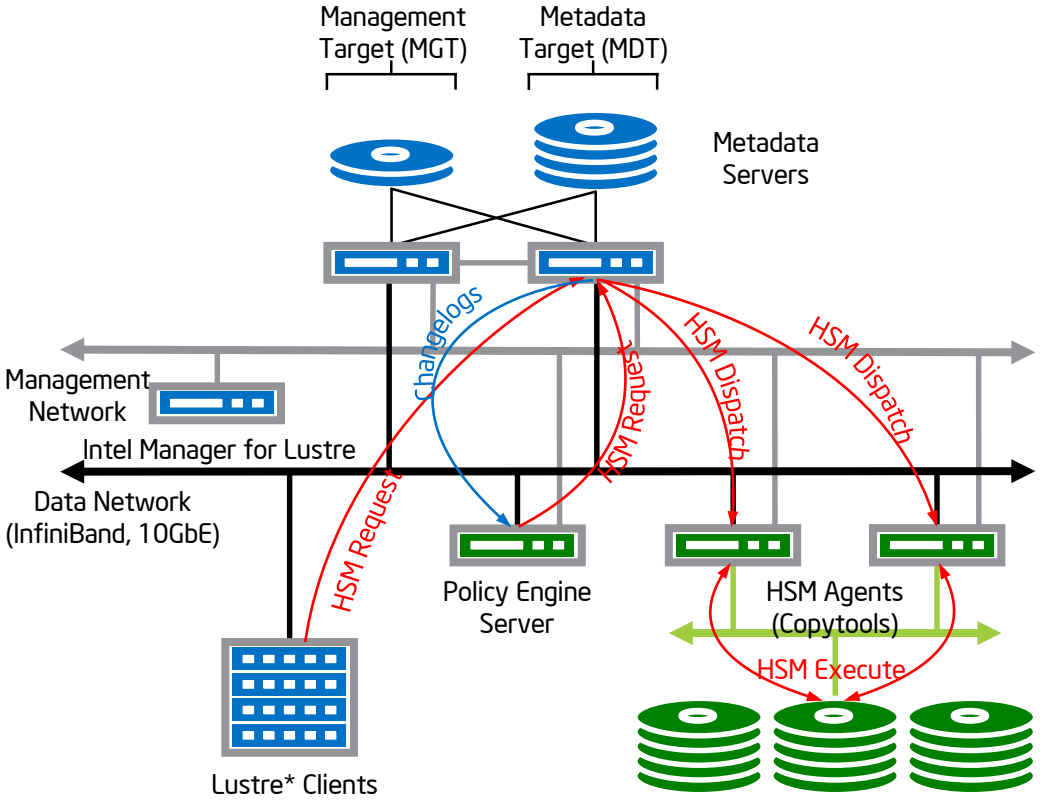
* Other names and brands may be claimed as the property of others

英特尔Lustre*企业版 V2

Feature or major enhancement	Benefit	Lustre 2.5	Intel® Enterprise Edition for Lustre* v2.0
Supported by Intel		✓	✓
CLI based management		✓	✓
Multi-vendor storage		✓	✓
Intelligent installation	Easily install software components from central server		✓
Intuitive configuration	Easily deploy proven, enterprise-grade storage		✓
In depth UI-based monitoring	Actively monitor previously installed Lustre solution		✓
Comprehensive UI-based management	Lowers complexity, raised productivity and overall ROII		✓
UI based configuration of hierarchical storage mgmt.	Enable HSM coverage for optimizing storage utilization		✓
Advanced and extensible charting	Detailed insights into storage in near-real time		✓
Automated configuration of high availability services	Easily configure server pairs for optimum availability		✓
Automated configuration of Lustre Networking	Ensure LNET services are configured correctly		✓
Intelligent log files	Quickly and easily understand important messages		✓
Storage plug-in infrastructure	Extends IML charting to include underlying hardware		✓
REST API	Integration with management tools		✓
Software integration layer (adapter) for Hadoop	Fast, shared, easily managed storage		✓

完整的分层存储体系结构(HSM)

- HSM提供高效动态数据迁移机制。在同一套文件系统中整合多种级别的存储设备，通过各种策略调整热数据和冷数据在存储系统中的配置，以达到最高的性价比。
- 图形化地进行配置和管理
- 提供完备的接口和ISV应用集成
- 好处：分层存储降低总体拥有成本和简化资源共享，优化整体存储使用效率。



* Other names and brands may be claimed as the property of others

Hierarchical Storage Management (HSM)

- HSM Coordinator (at MDT)

lctl set_param [-P] mdt.<fsname>-<MDT index>.hsm_control=enabled

lctl get_param mdt.<fsname>-<MDT index>.hsm_control

- HSM Agent (Copytools)

mount /dev/sdb /archive/demo

lhsmtool_posix --daemon --hsm_root /archive/demo --archive 1 /lustre/demo

- Policy Engine

lctl set_param mdd.<fsname>-MDT.changelog_mask="all-XATTR-MARK-ATIME"*

lctl --device <fsname>-<MDT index> changelog_register

sudo rbh-lhsm --scan –once -f /etc/robinhood.d/lhsm/demo-lustre-hsm.conf

sudo service robinhood-lhsm start

HSM Agent

- Using “Archive Identifier” to distinguish different high capacity storages
- Each agent binds with one or more “Archive Identifier”
- “0” means all archives (not recommended to use)
- HSM Coordinator at MDT will choose the best agents according to the work load of each agent (by the number of pending requests)

HSM Policy Engine Configuration

General –the mount point of Lustre*

Log – the information about various logs

ListManager – the configuration of MySql

ChangeLog – the MDT device name and the Changelog registration ID

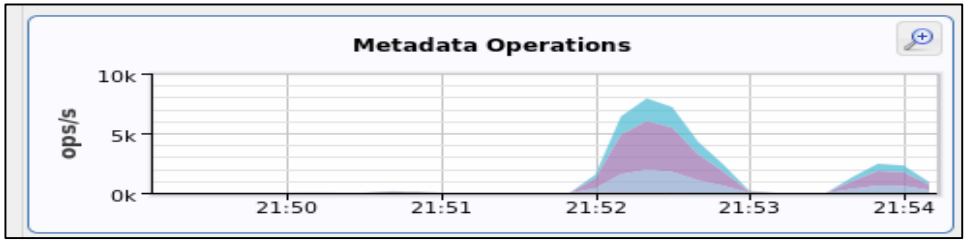
Migrate – copying files to the high capacity archive storage.

Purge – releasing files that have been archived in order to free capacity in Lustre*.

Remove – cleaning up archive copies of deleted files.

针对Hadoop应用优化的高性能并行存储

- Hadoop应用的读写模式与传统的文件系统操作完全不同，特点是同时打开/关闭海量文件
- 当运行大型Hadoop作业时元数据服务器会承受很大的读写压力，经常导致过载现象发生。
- 英特尔Lustre*企业版中提供新的扩展属性缓存机制--提高元数据服务器的性能和抗压能力。
- 好处：为Hadoop应用提供高效，高性能，高扩展性，易管理的存储解决方案

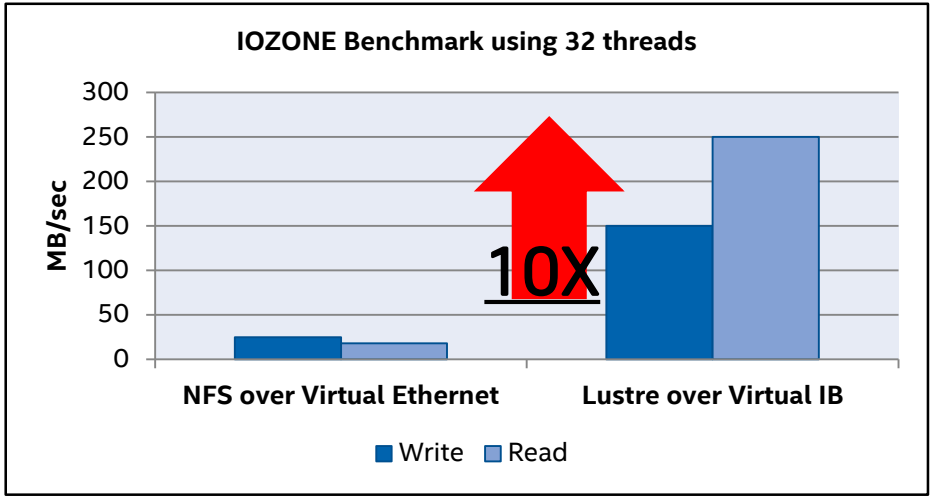


Metadata operations during a Terasort experiment[1] using a 3TB dataset and 240 Map-Reduce tasks

[1] Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

原生Intel® Xeon Phi™ Lustre*客户端

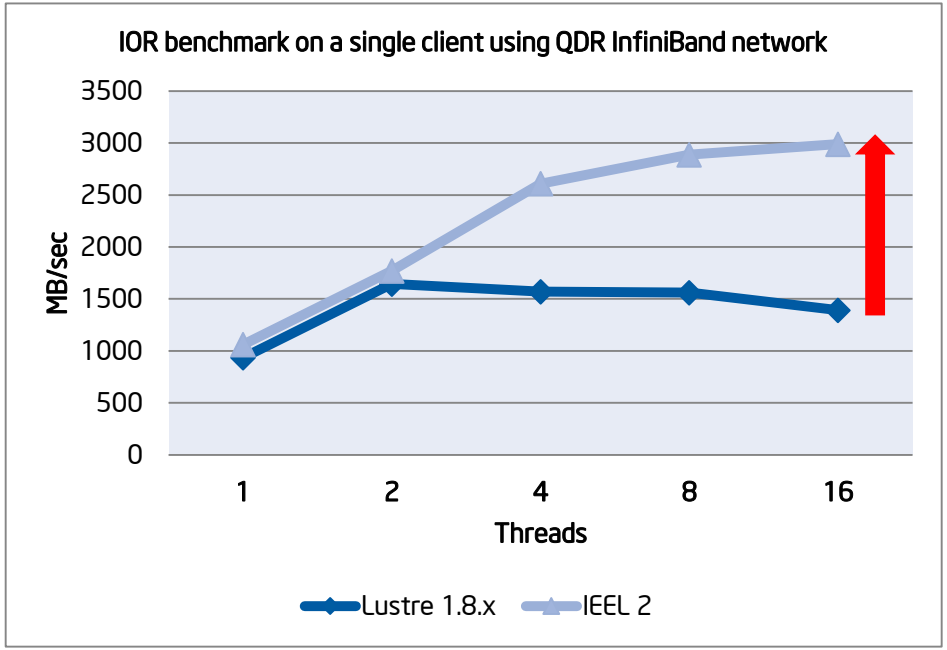
- 针对于多节点大型平行作业，例如每个节点均需要100+核
- 使在Xeon Phi™中运行的应用直接访问Lustre存储
- 好处：为Xeon Phi™应用提供更高性能的IO访问能力



[1] Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

优化单客户端读写性能[1]

- 在企业市场中经常会有需求为单线程单客户端的应用提供高IO服务
- 英特尔Lustre*企业版提供针对于单客户端数据流的优化。
- 好处：提高商业应用软件的读写性能，例如：MapReduce



[1] Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

* Other names and brands may be claimed as the property of others

Intel® 集成化Lustre*管理环境

配置，优化，管理Lustre*文件系统

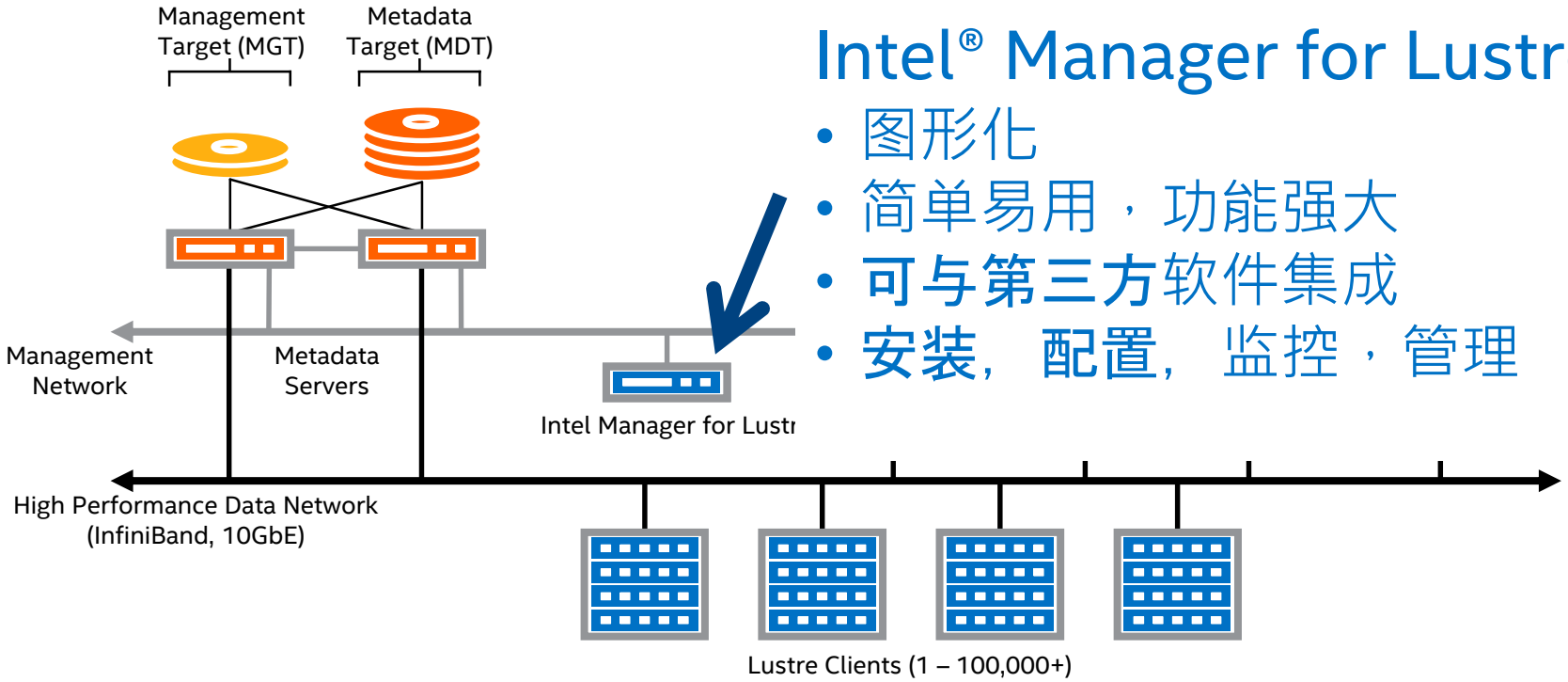
易用、高效、功能强大的图形管理平台，同时汇集了Intel开发支持团队的十多年最佳实践精华于其中。

- 利用图形化界面指导系统管理员，高效地配置、监控及管理整个Lustre存储平台以降低管理复杂度和管理成本
- 通过各种图表揭示存储系统内部的实时性能
- 自动配置存储服务器之间的互备机制，以实现文件系统高靠性。
- 智能化的预警设计及日志分析，使得系统管理员及时掌控文件系统的状态。

Intel® Manager for Lustre Makes Lustre Smarter, Simpler and More Productive

* Other names and brands may be claimed as the property of others

降低管理复杂度

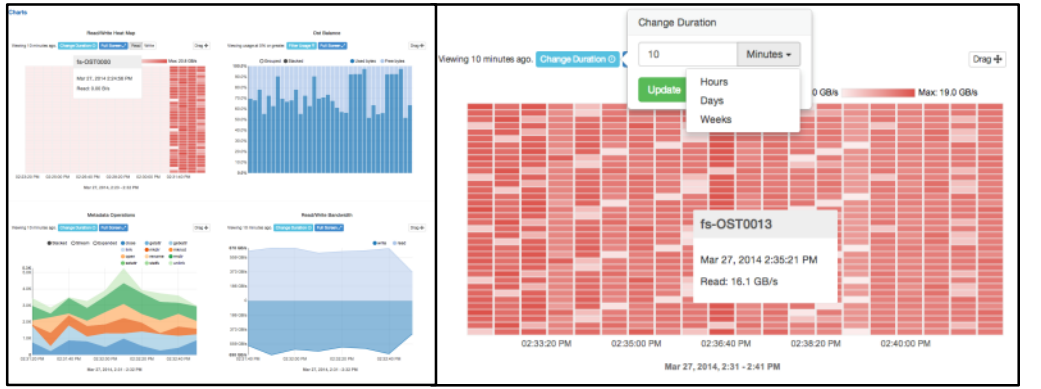


Intel® Manager for Lustre*

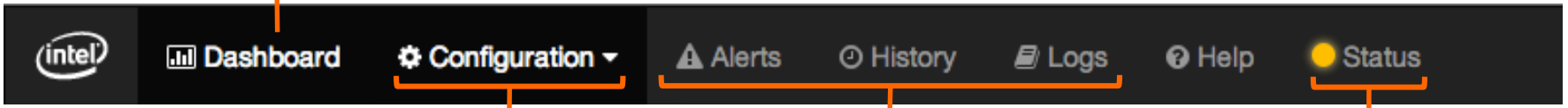
- 图形化
- 简单易用，功能强大
- **可与第三方**软件集成
- **安装，配置，监控，管理**

* Other names and brands may be claimed as the property of others

Intel® Manager for Lustre*图形化管理界面



集成显示面板整合各种图表显**示性能和**资源利用率。系统管**理**员可以很容易地浏览文件系统，监测作业的资源利用状态**以及性能**。



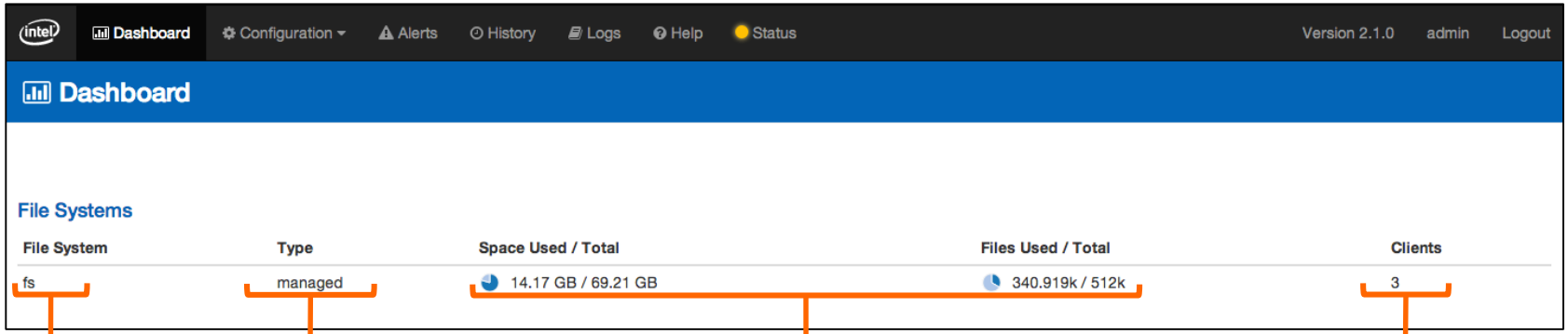
配置服务器，存储卷，电源管理，进一步激活**并配置HSM服**务

智能化的报警和**日志**监控机制

文件系*统全局****状****态指示灯*

* Other names and brands may be claimed as the property of others

Intel® Manager for Lustre* 显示面板



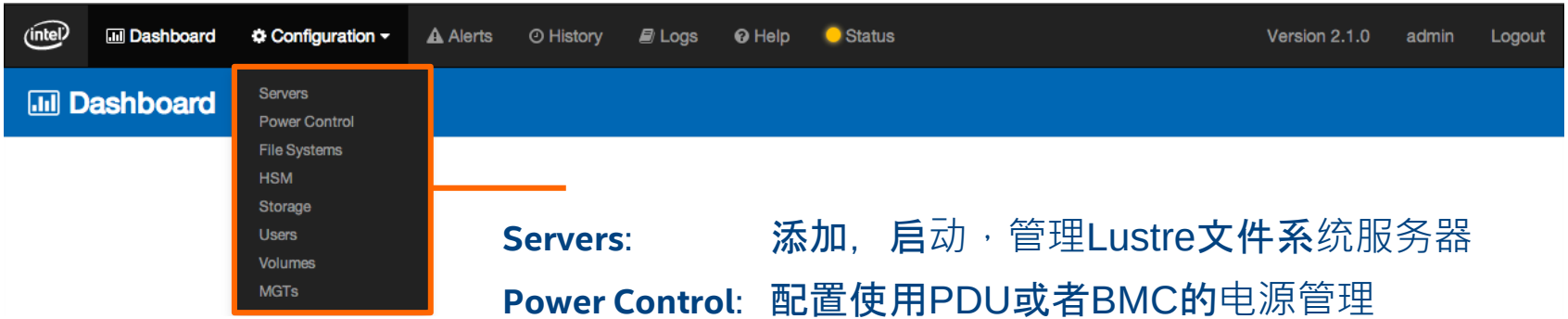
文件系统

管理/监视模式

文件系统使用率，及接入文件系统的客户端数目

* Other names and brands may be claimed as the property of others

简化配置流程



Servers: **添加，启**动，管理Lustre**文件系**统服务器

Power Control: **配置使用**PDU**或者**BMC**的**电源管理

File Systems: 创建、删除、启动、停止文件系统

HSM: **激活分**层存储服务

Storage: **集成存**储硬件插件

Users: **管理**账号

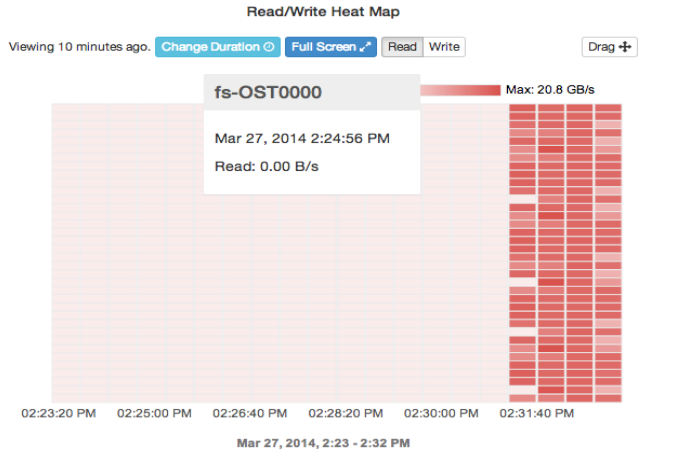
Volumes: **配置存**储卷及高可靠**性互**备机制

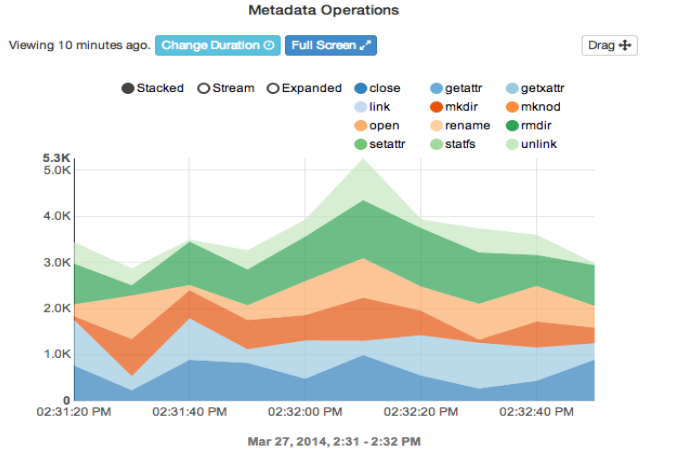
MGTs: 设置管理存储设备**的参数**

* Other names and brands may be claimed as the property of others

热点和元数据操作

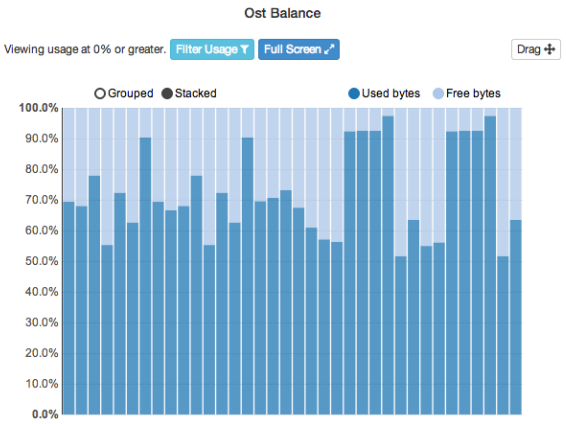
- 动态显示存储资源的使用分布效果，帮助系统管理员迅速确定热点区域并进一步分析到作业以及底层存储设备读/写操作。
- 统计整个文件系统的全部元数据IO操作，通过各种图表以展示从全局一直细化到单类型操作的状态

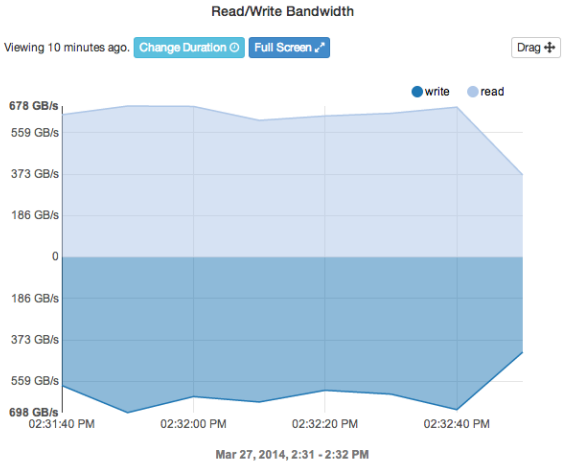




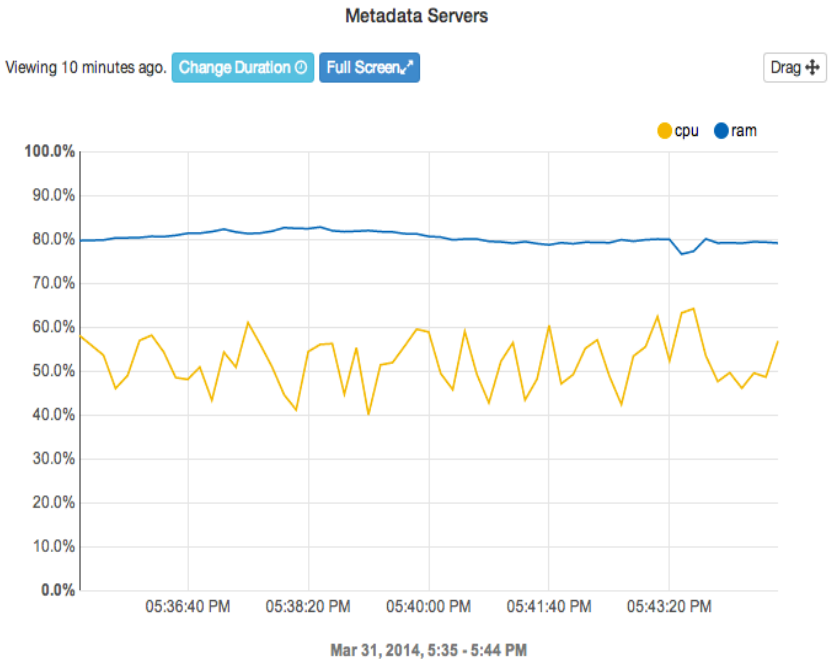
OST分布和IO统计

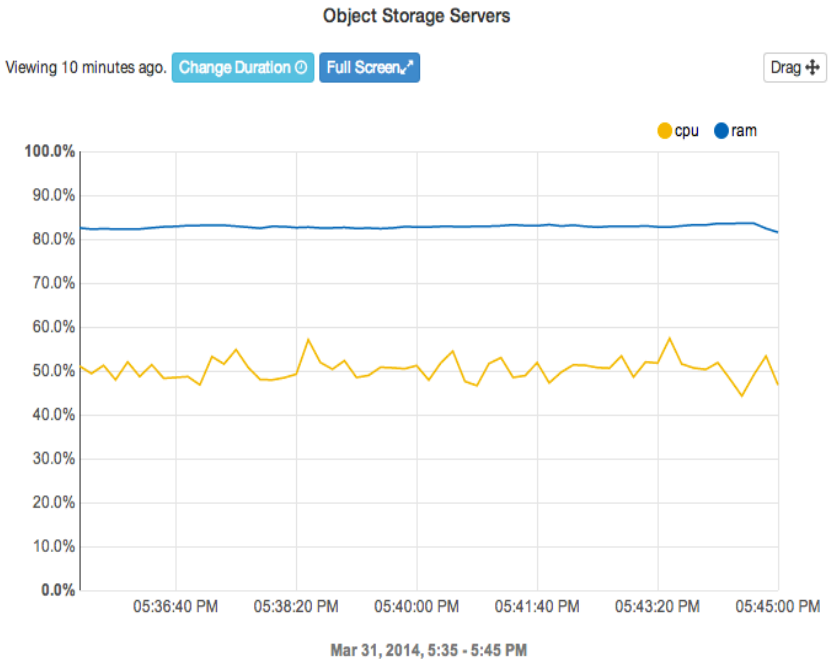
- **揭示底**层对象存储设备使用率**分布状**态，**帮助系**统管理员直观地确认出是否需要对文件系统中的存储空间做再均衡操作。**同**时提供每一个OST**的**详细信**息**。
- 统计显示读写带宽，**展示从整个文件系**统和细化到单个OST**的**读写吞吐量。





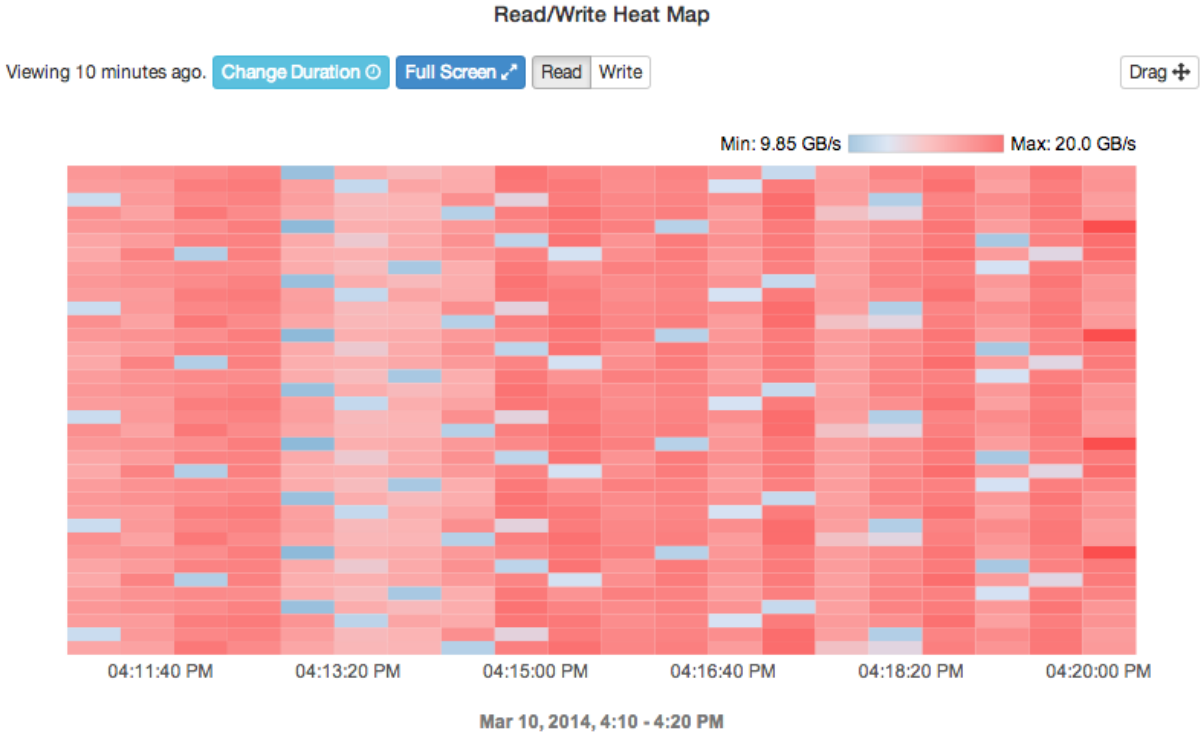
CPU和内存利用率


Metadata Servers
Viewing 10 minutes ago.
Change Duration
Full Screen
Drag
cpu
ram
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
05:36:40 PM
05:38:20 PM
05:40:00 PM
05:41:40 PM
05:43:20 PM
Mar 31, 2014, 5:35 - 5:44 PM

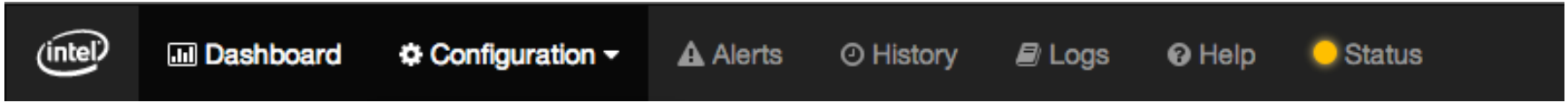

Object Storage Servers
Viewing 10 minutes ago.
Change Duration
Full Screen
Drag
cpu
ram
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
05:36:40 PM
05:38:20 PM
05:40:00 PM
05:41:40 PM
05:43:20 PM
05:45:00 PM
Mar 31, 2014, 5:35 - 5:45 PM

作业状态统计

- 针对指定作业做**深度分析**
- 针对指定OST的**作业统计**



事件、报警及日志



简洁化地集中
管理界面

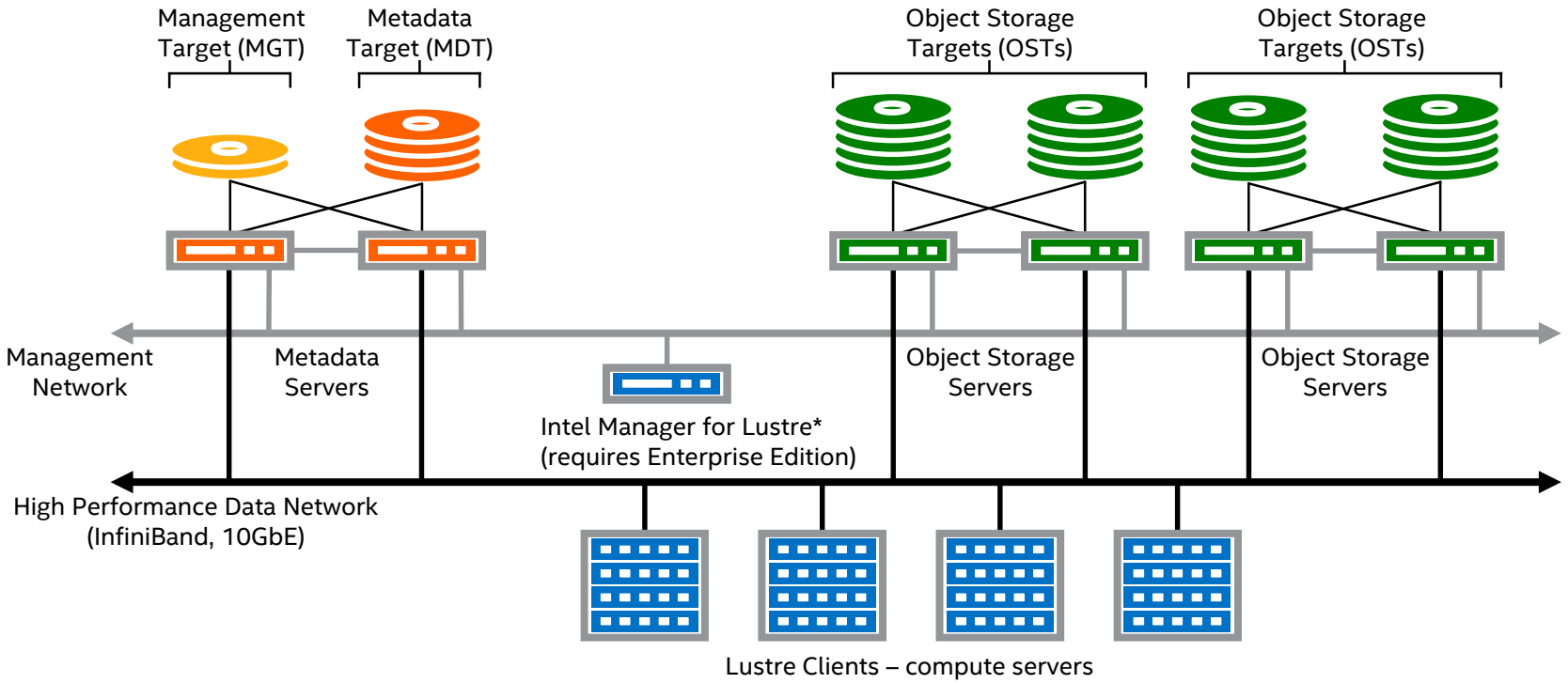
整合整个Lustre*存储集群

- 预警
- **事件及操作历史（可设置查询时间段）**
- **日志（解析过的）**

* Other names and brands may be claimed as the property of others

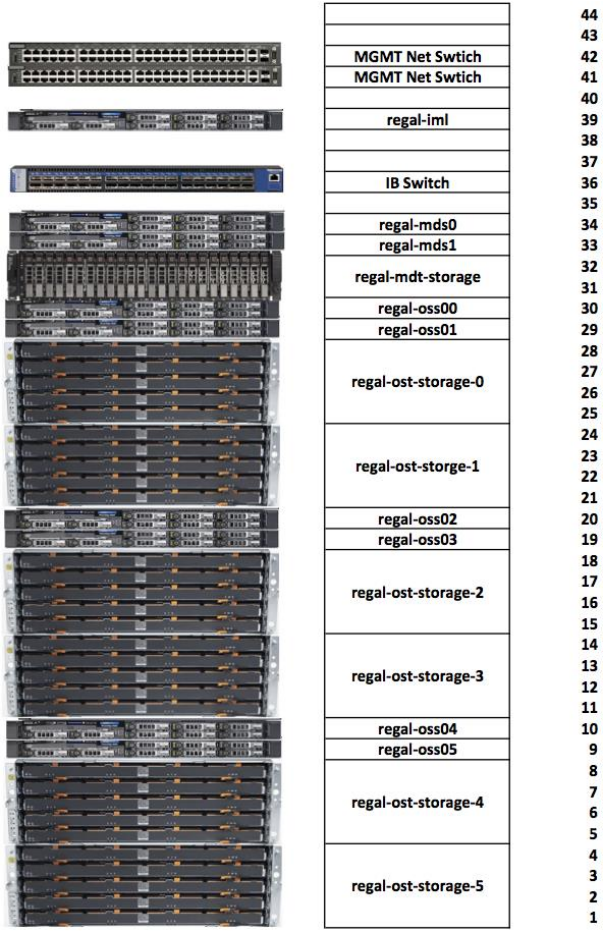
英特尔Lustre*企业版推荐实施配置

IEEL推荐架构图



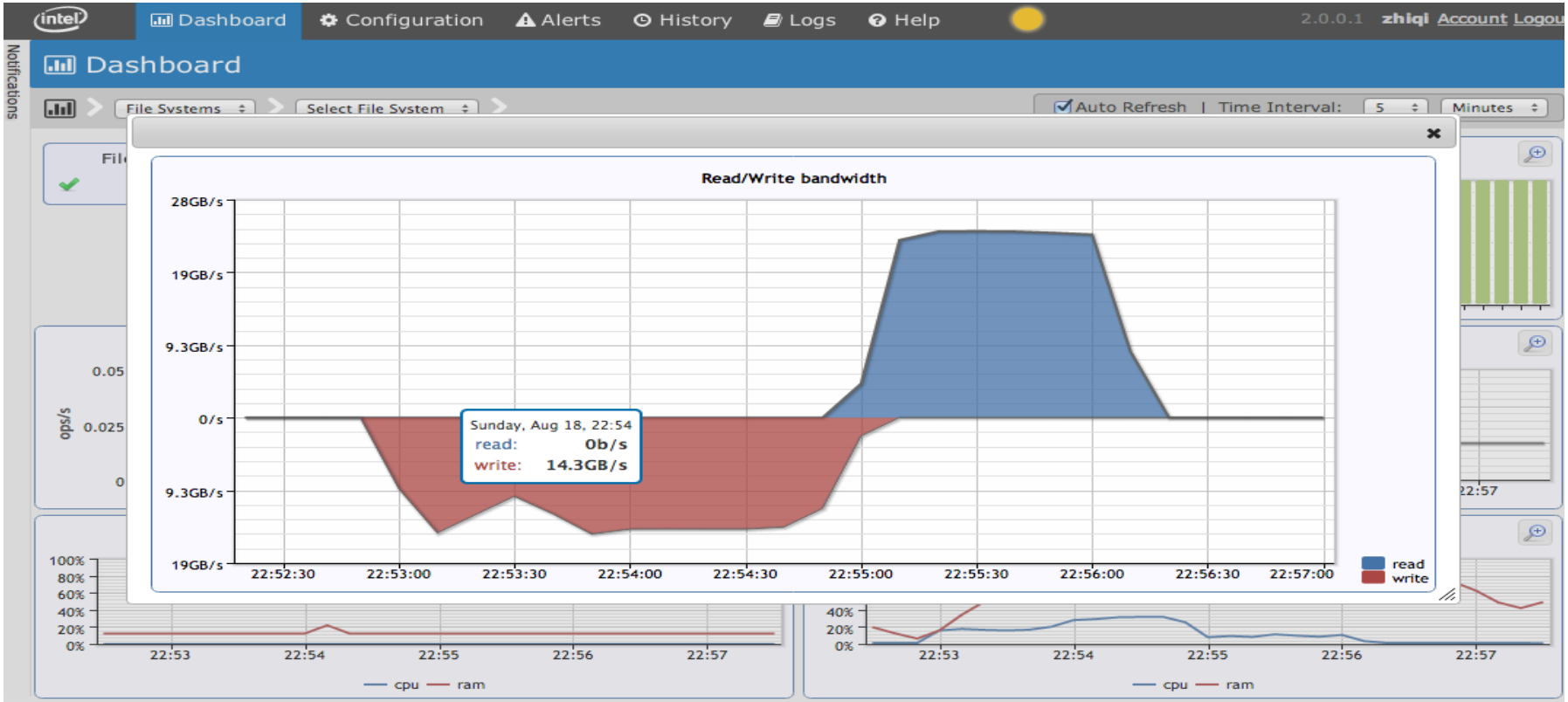
* Other names and brands may be claimed as the property of others

IEEL配置案例

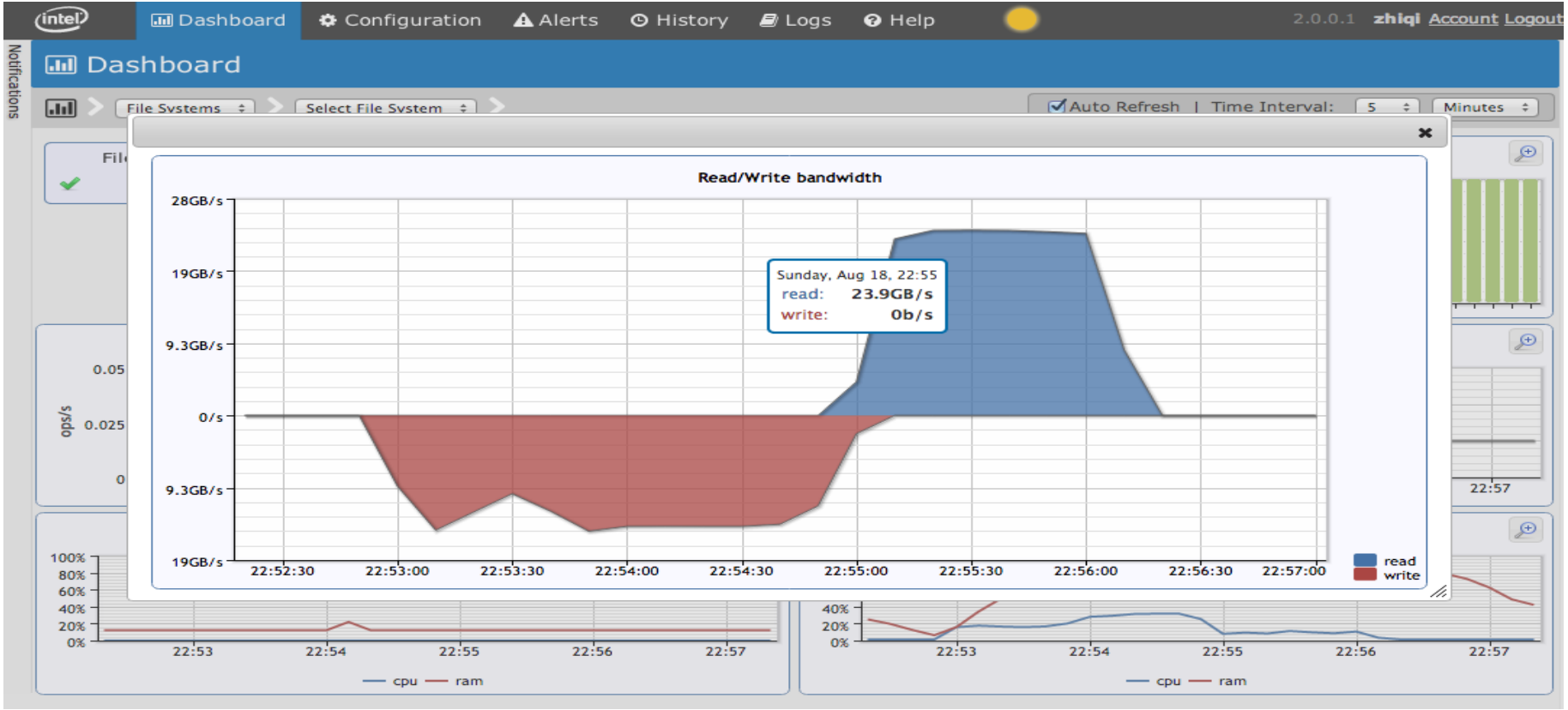


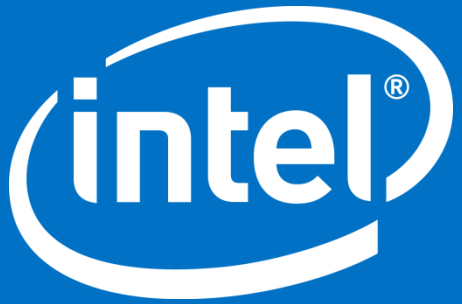
- 容量
  - 1050 TB **可用空间**
  - 11亿文件
- 性能
  - 写 13GB/s
  - 读 20GB/s
- 元数据操作速率
  - Create 60K
  - Lookup 220K
  - Getattr 98K
  - Setxattr 13K
  - Destroy 33K
- 高可靠性
  - Active-Standby MDS
  - Active-Active OSS
  - Active-Active Controller
  - RAID10/RAID6

IEEL配置案例 – 写性能展示



IEEL配置案例– 读性能展示





Lustre.Intel.Com