China LUG 2017

Dell Storage with Lustre for HPC/BigData/AI

Forrest Ling
HPC Enterprise Technologist, Dell China
2017.10.31

Agenda

- Storage Market in China
- Why Linux
- Why Lustre
- Dell Storage with Lustre
- Dell HPC Solutions for HPC/HPDA/AI
- Dell HPC Lab in Beijing China

HPC/Big Data/AI Storage Market in China

- HPC performance increasing 1000 times per 10 years since 1985 in China (reported in the keynotes in HPC China 2017)
- AI, Big Data and HPC are merging;
- HPC Storage CAGR is 9.4% from 2015 to 2019 report by IDC 2016
- Data explosion:
  - HiSeq X Ten sequencer generates 18Tb/3days;
  - Seismic Data : biggest single file size: 2389TB

Lustre in China Market

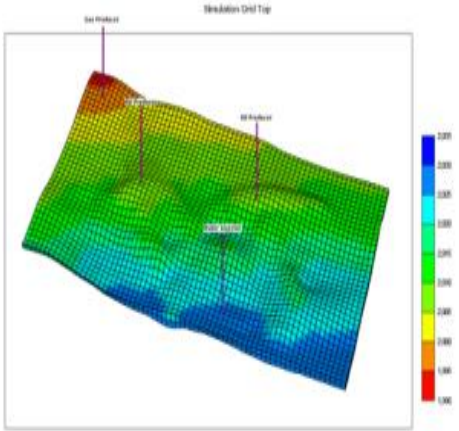
Life Sciences

- Genomics
- Cryo-EM

Super Computing Centers

Educations & Research

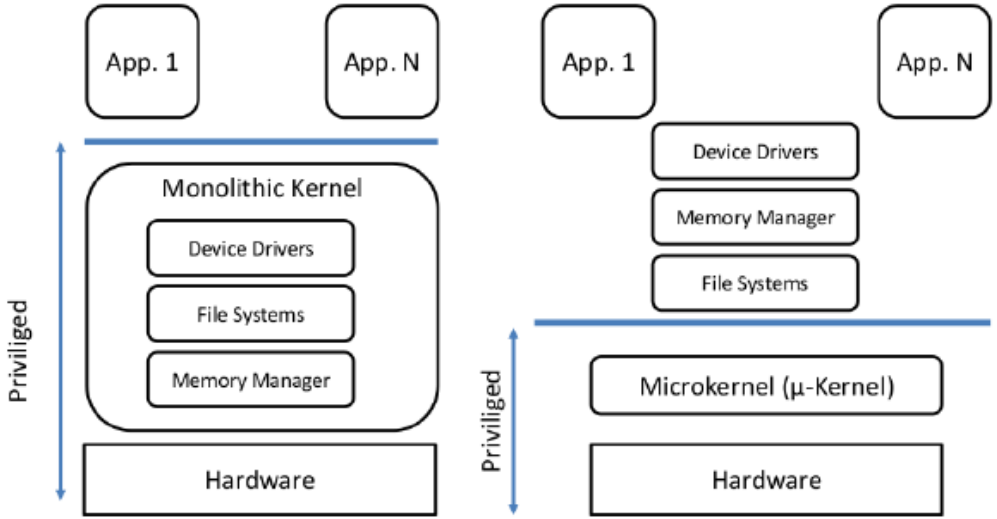
Seismic Data Processing



AI/DL

Linux Feature

- Kernel architect
  - Monolithic Kernel vs. Micro Kernel
- Kernel cgroup support containers
  - Process Namespaces;
  - Containers support mount File Systems
- VFS for IO
  - Global FS namespace
- HPC layers software & AI Framework
  - Open source on Linux
- Linux cost
  - Free of charge

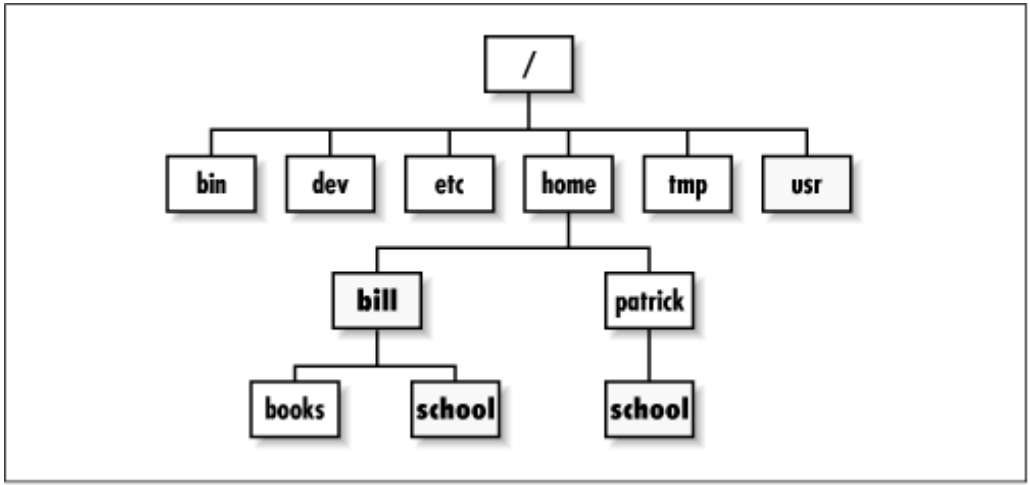


A: Monolothic kernel design

B: Microkernel design

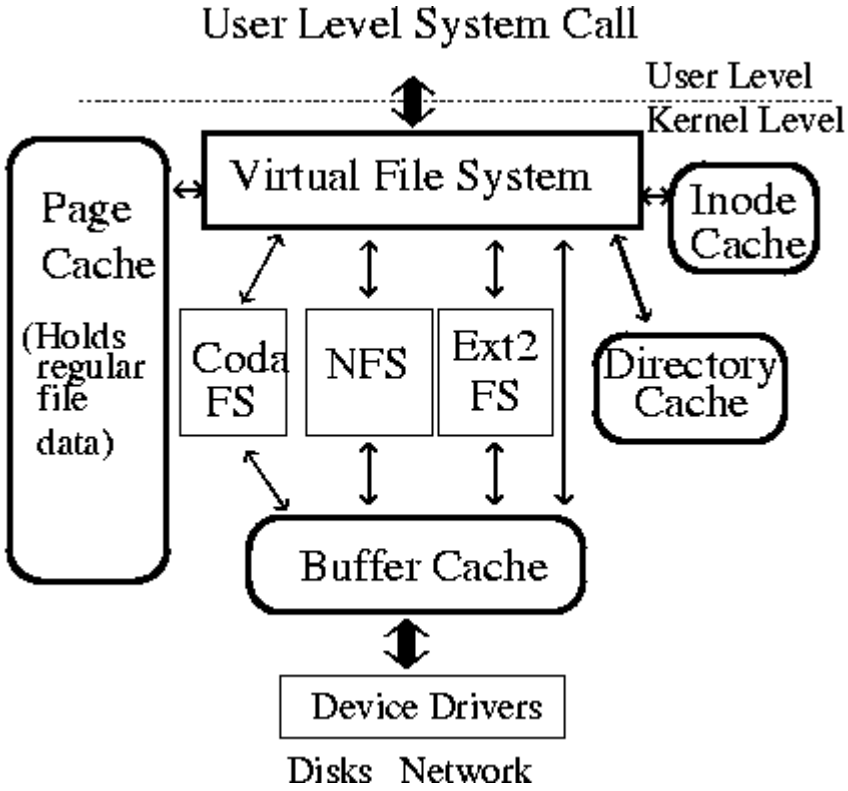
Linux VFS

- Global and unified namespace of files and directories
- Can mount any Filesystem on a point in the global VFS namespace

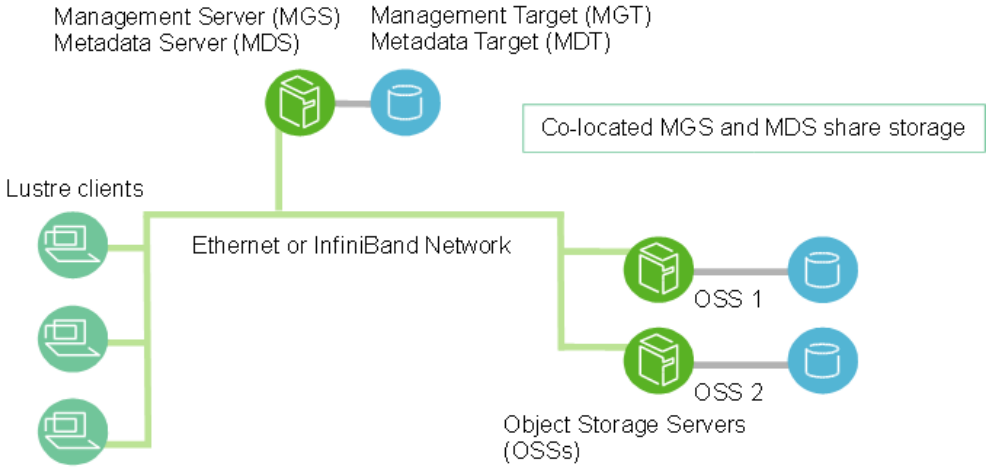


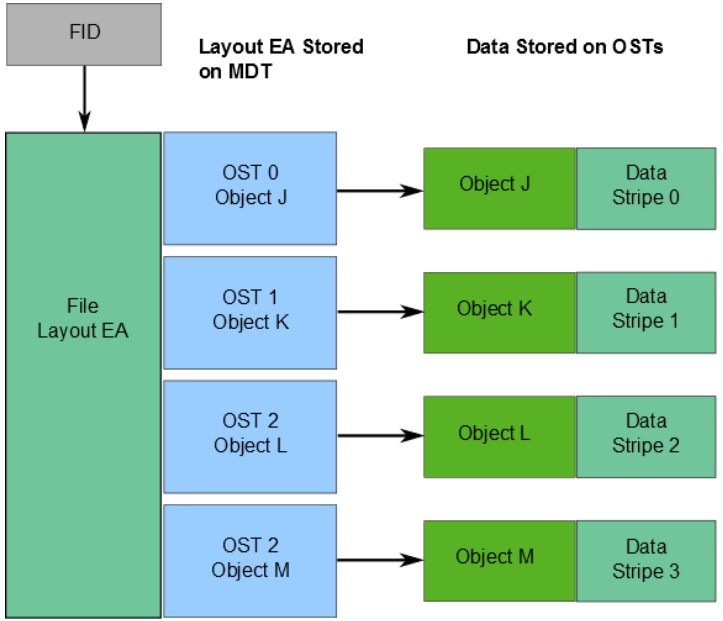
Linux VFS

- Global VFS Namespace
  - Mount any Filesystem on a point
- Page size (memory management unit)
  - 4KB
- Block size (VFS management unit)
  - 512B, 1KB, 4KB
- Block Group (ext2/3)
  - 1x superblock + n* block of inode + m* block of data …
- Sector size (disks management unit)
  - 512B

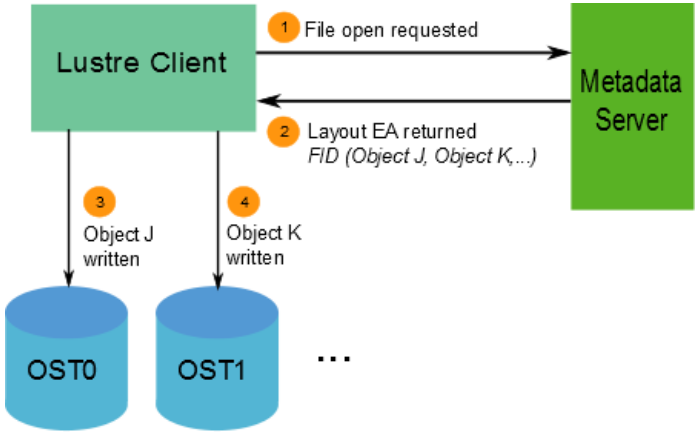


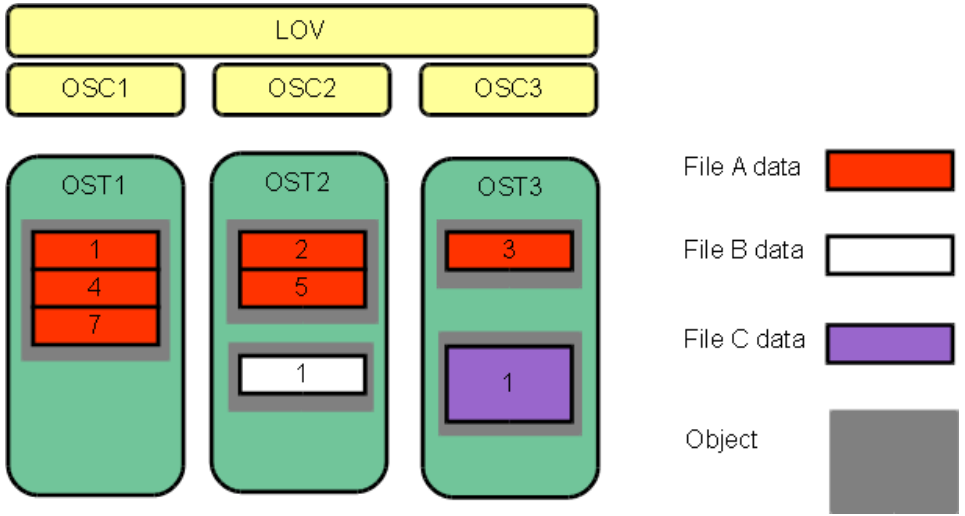
Lustre Architecture





Lustre FS Access





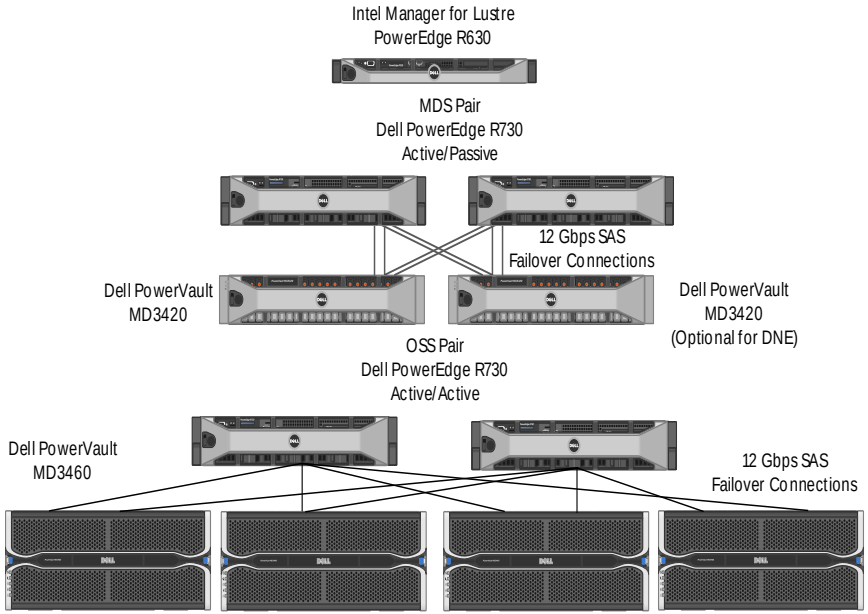
Default stripe_count: 1
Default stripe_size: 1MB

Lustre Feature

Feature	Current Practical Range
Client Scalability	100-100000
Client Performance	Single client: I/O 90% of network bandwidth Aggregate: 10 TB/sec I/O
OSS Scalability	Single OSS: 1-32 OSTs per OSS Single OST: 300M objects, 128TB per OST (ldiskfs) 500M objects, 256TB per OST (ZFS) OSS count: 1000 OSSs, with up to 4000 OSTs
OSS Performance	Single OSS: 15 GB/sec Aggregate: 10 TB/sec

Feature	Current Practical Range
MDS Scalability	Single MDS: 1-4 MDTs per MDS Single MDT: 4 billion files, 8TB per MDT (ldiskfs) 64 billion files, 64TB per MDT (ZFS) MDS count: 1 primary + 1 standby Introduced in Lustre 2.4 256 MDSs, with up to 256 MDTs
MDS Performance	50000/s create operations, 200000/s metadata stat operations
File system Scalability	Single File: 32 PB max file size (ldiskfs) 2^63 bytes (ZFS) Aggregate: 512 PB space, 1 trillion files

A scalable building block design



- **Designed to scale to the Exabyte, with a Petabyte of throughput**

Solution benefits & Dell differentiation

- Single file system namespace scalable to high capacities and performance
- Engineered by Dell HPC Engineering to provide maximum throughput per building block with on-the-fly storage expansion
- Solution design for Big Data workloads using Intel Hadoop Adapter for Lustre (HAL)
- Dell Networking 10/40GbE, InfiniBand, or Omni-Path

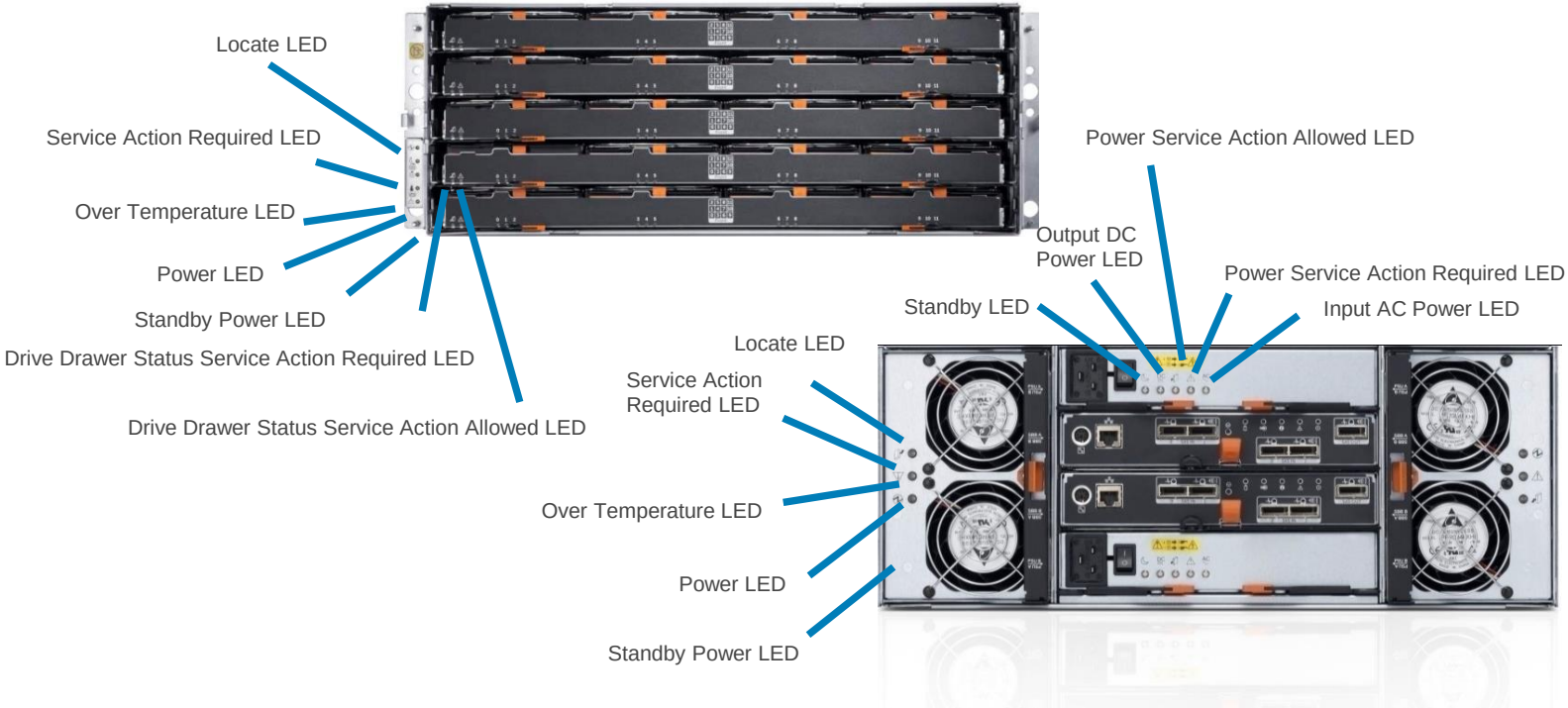
MD3 Dense Array

Dense array-save space, power and cooling

- High-density 4U disk shelf supporting 60 SAS drives
- Controlled drawer movement allows each 12-drive drawer to be extended while its drives remain active
- Individual drawer extension and front access enables safer drive replacement
- Up to 10% reduction in power and cooling
- High-speed 6Gb SAS interconnects with expansion enclosures
- Optimized for redundancy and reliability with hot-swappable drives and other components

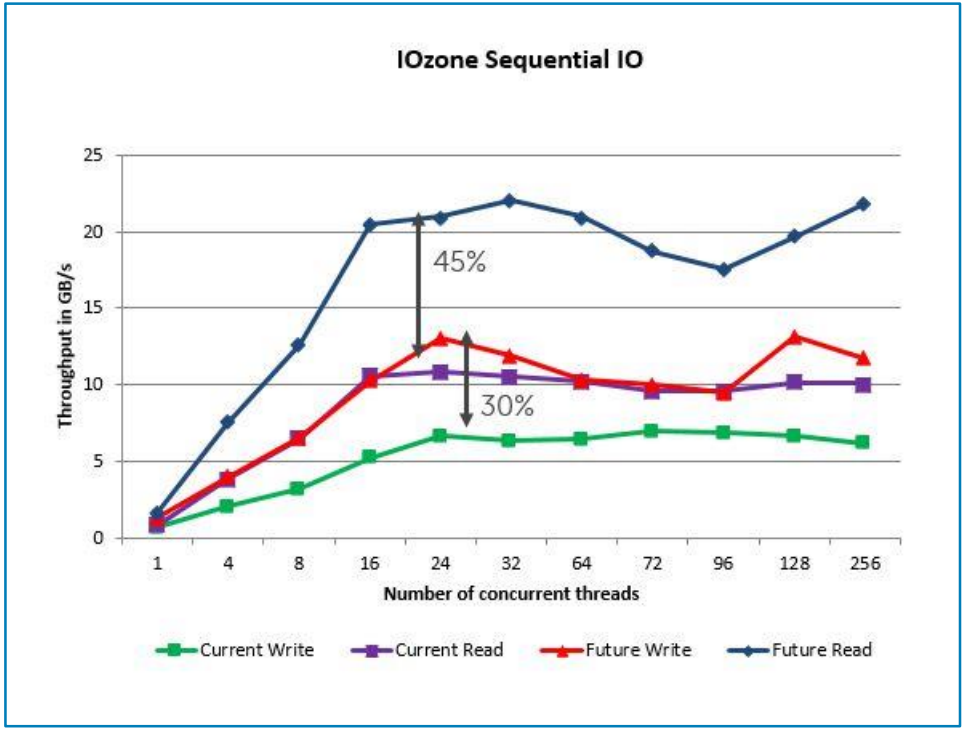
MD3 Dense Array

Front and back views



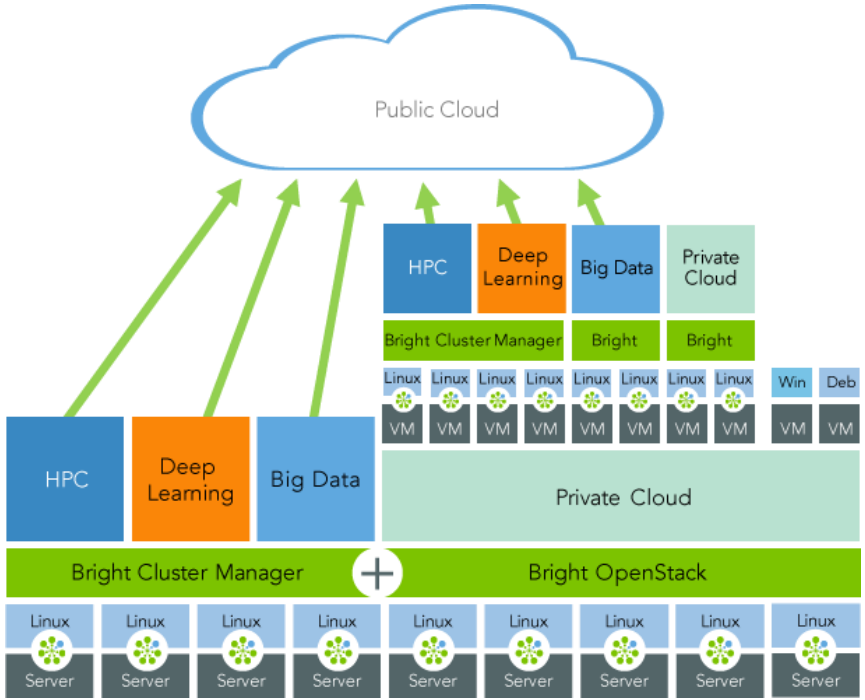
Increase metadata performance with Luster Distributed Namespace feature (Lustre DNE)

- Lustre DNE Phase 2 and ZFS support
- Lustre sub-directories can now be distributed across multiple MDTs to increase metadata capacity capabilities and performance.



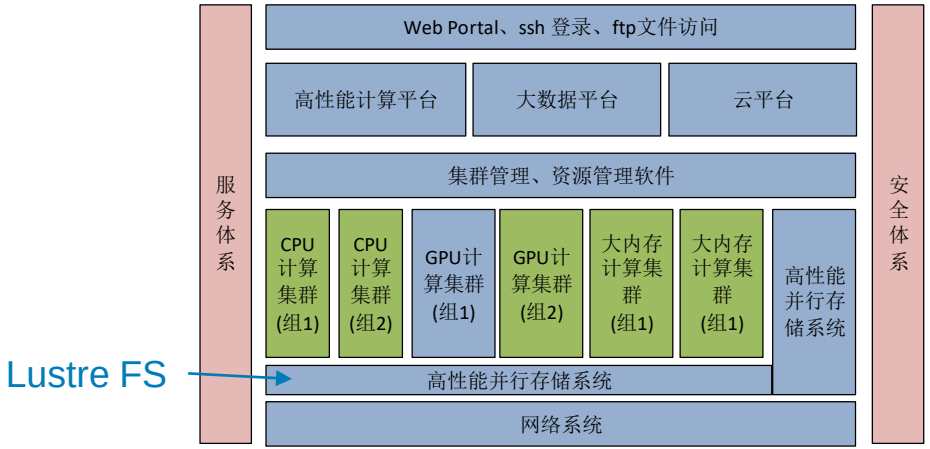
Dell HPC/AI Solution

- **Deep Learning framework**
  - Caffe, Torch, Tensorflow,Theano
- **Machine Learning libs**
  - MLPython, cuDNN, DIGITS, CaffeOnSpark
- **Supported hardware drivers**
  - CUDA drivers
  - CUB (CUDA building blocks)
  - NCCL (library of standard collective communication routines)
- **Support Hadoop**
- **Support OpenStack/containers**
- **Support Lustre (including Hadoop on Lustre)**



CASAI Dell Advanced HPC/AI Platform

Dell公司和中科院脑科学和智能技术
IT技术支撑平台



热烈祝贺：

戴尔（中国）有限公司

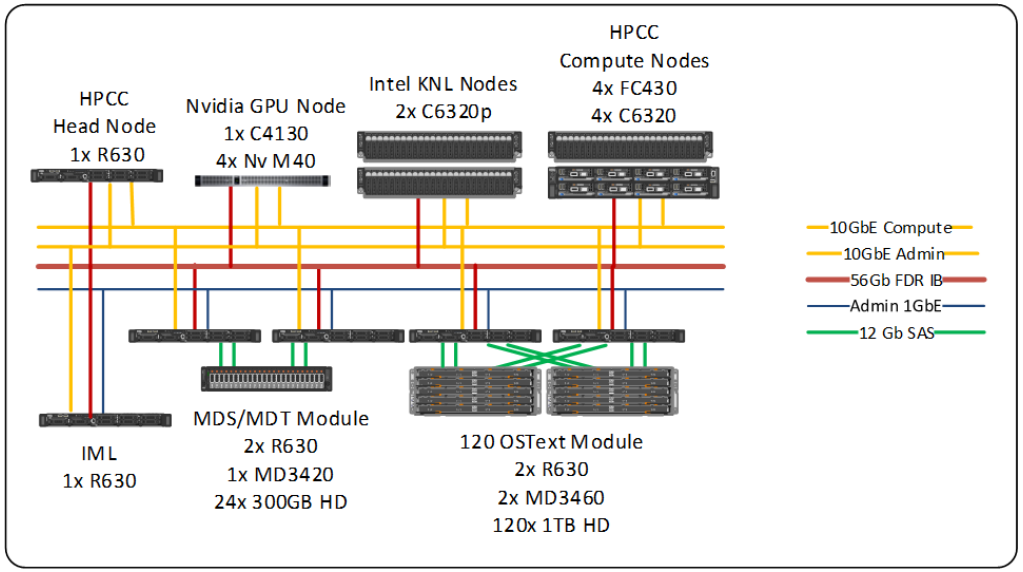
荣获
人工智能技术实践
创新奖

HPC China 2016大会

Dell Intel HPC Lab in Beijing China

Tested Software：

- CentOS/Redhat Linux
- Bright Cluster Manager
- Altair PBSworks
- Intel Enterprise Lustre
- 联科集团 - CHESS
- 并行科技 - Paraplus
- 蓝海彤翔 – COMS
- OpenHPC



Dell Intel HPC Lab in Beijing China
戴尔北京HPC高性能计算创新实验室
欢迎客户访问、测试、验证和创新


Thank You!