whamcloud

LUSTRE MADE BRILLIANT

- Eric Barton

  CTO

  Whamcloud, Inc.

  eeb@whamcloud.com

Agenda

- Whamcloud introduction
  - Who we are and what we do
- Lustre current status
  - Community
  - Current development
  - Roadmap
- Looking forward
  - Exascale
  - Growing the Lustre market

Lustre Timeline

- 1999 – Lustre project startup
- 2001 – ASCI Pathforward
- 2003 V1.0 – CFS

- 2007 V1.6 – Sun
- 2009 V1.8 – Sun

- 2010 V2.0 – Oracle

- 2011 V2.1 – Whamcloud
- V2.2 underway

Whamcloud introduction

- VC-backed California Corp. Formed July 2010
- ~50 employees/contractors worldwide
  - Unique advantage: critical mass for Lustre technology
- ~100 supported sites worldwide
- Today's offerings:
  - Worldwide L3 support
  - NRE development
  - Chroma management

Whamcloud is widely recognized as the source for Lustre
We have the only HW vendor-neutral offering

Whamcloud management team

CEO Brent Gorda (> 25 yrs HPC)

- DOE program manager for Lustre, BlueGene, TLCC
- Entrepreneur: Myrias Research, Bonsai Software, MetaExchange

CTO Eric Barton (> 25 yrs in HPC)

- LNet developer since 2002, Lustre architect since 2008
- Entrepreneur: Meiko Scientific founder, Quadrics & consulting

Daniel Ferber – Cray/SGI/Sun/Oracle, bizdev

Peter Jones – Lustre support manager

Bryon Neitzel – Lustre development manager

Jessica Popp – Lustre project management

Robert Read – Lustre 2.0 lead engineer

Lustre technical dream team

- Andreas Dilger – Lustre founder, ext4 co-author
- Johann Lombardi – 1.6/1.8 lead engineer
- Alex Zhuralev – ext4 co-author, Orion lead
- And > 20 other experienced Lustre engineers
  - Mikhail Persion – Lustre recovery
  - Oleg Drokin – Lustre locking, Reiser F/S
  - Liang Zhen – SMP Scaling / LNet
  - Wang Di – DNE Lead
- Plus seasoned technical experts
  - Chris Gearing – Sr. manager of AMD support, troubleshooting
    - Lustre test automation
  - John Spray – product developer

Lustre community

- Whamcloud community membership

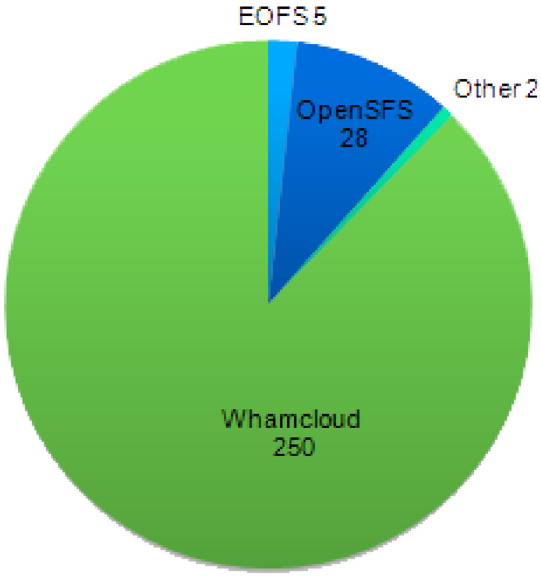
  OpenSFS
  Open Scalable File Systems, Inc.

  EOFS
  European Open File System

- Whamcloud maintains the community assets
  - Wiki + roadmap: http://wiki.whamcloud.com
  - All Lustre releases: http://downloads.whamcloud.com
  - Jira bug tracker: http://bugs.whamcloud.com
  - Git repositories: http://git.whamcloud.com
  - Gerrit code review: http://review.whamcloud.com
  - Build: http://build.whamcloud.com
- No copyright assignment on source contributions
  - Ensures no single entity can own whole copyright on Lustre
  - Has support of OpenSFS and EOFS

Lustre current status

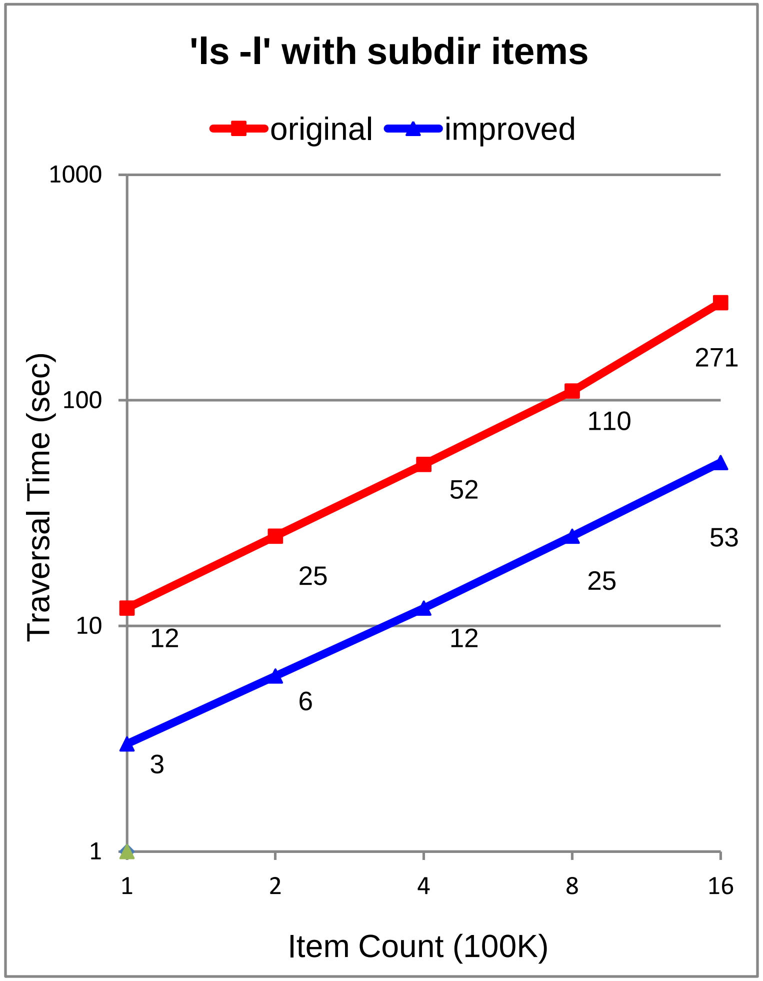
- Single community-wide source tree
  - Hosted at Whamcloud
  - Formally recognized by community
- Majority of development now at Whamcloud
  - Majority of the same engineers
- June 1.8.6-wc1 release
- September 2.1 release
  - 285 Patches Landed (image to right)
- November 1.8.7-wc1 release

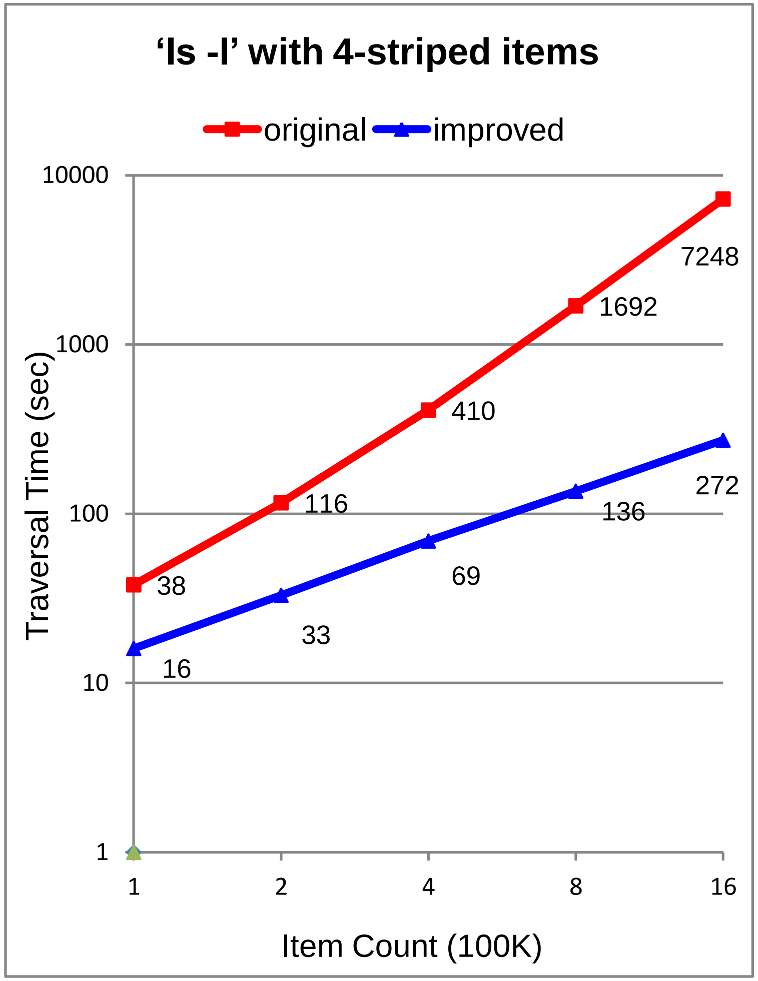


Accelerated single-client dir traversal

- Common to ls, du, find etc
- More efficient ldlm/object hash
  - Reduce hash bucket depth from ~3K to < 50
- Readdir
  - 1 page per RPC in current releases
  - 1 Mbyte per RPC reduces overhead
- Stat
  - MDS attributes
    - Getattr RPC fetches & locks UID, GID, nlink etc
    - Statahead pipelines RPCs and populates dcache and inode cache
  - OST attributes
    - Glimpse RPC fetches & locks size, blocks, mtime, ctime etc
    - Asynchronous glimpse pipelines RPCs

Accelerated single-client dir traversal

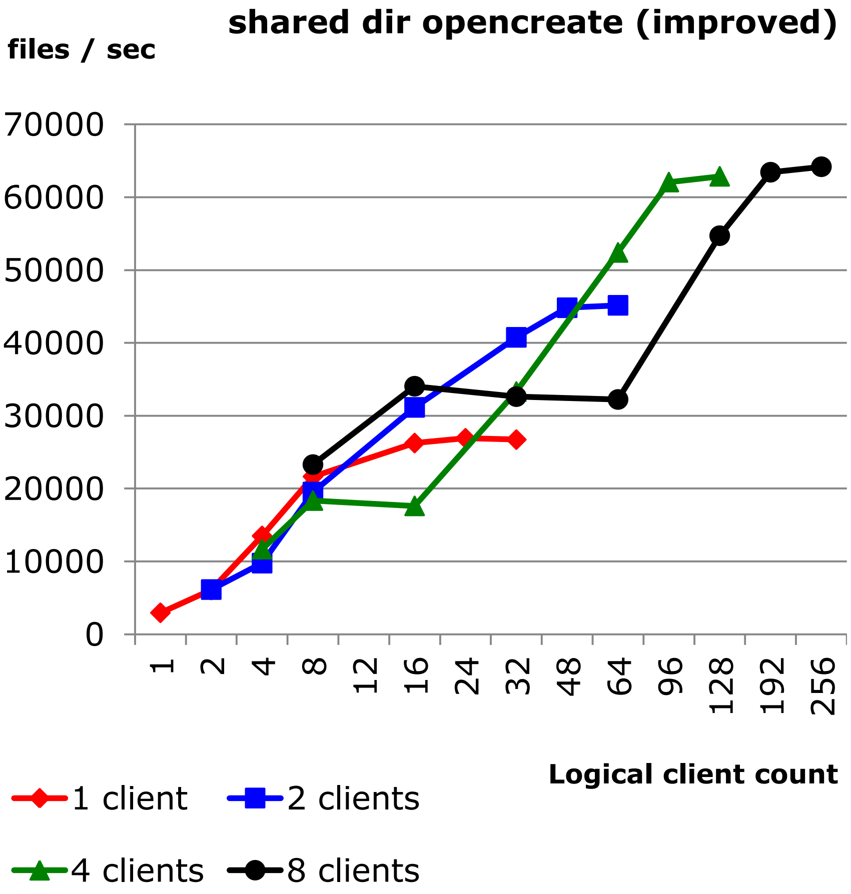


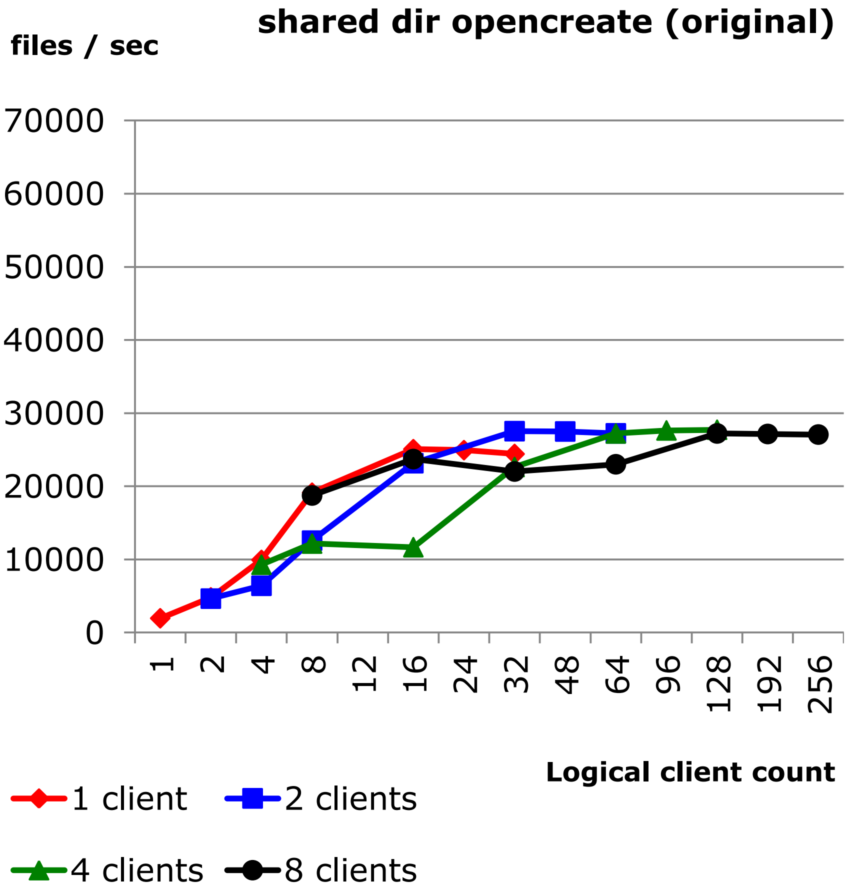


Improved MDS throughput

- MDS CPU bound
  - Poor affinity (cacheline pinging between CPUs)
  - Lock contention
- Request affinity
  - Define units of CPU affinity
    - Socket, core, hyper-thread
  - Separate RPC queue for each CPU unit
    - Reduced RPC queue lock contention
    - No data migration between CPU units
- Shared directory locking
  - Ldlm
    - Parallel directory operations support implemented in Lustre 2.0
  - Backend filesystem
    - IAM (incompatible with Lustre 1.8)
    - Ldiskfs/ext4
      - Hierarchical lock with shared modes
      - Contend only on leaf nodes if split not required

Opencreate under shared dir

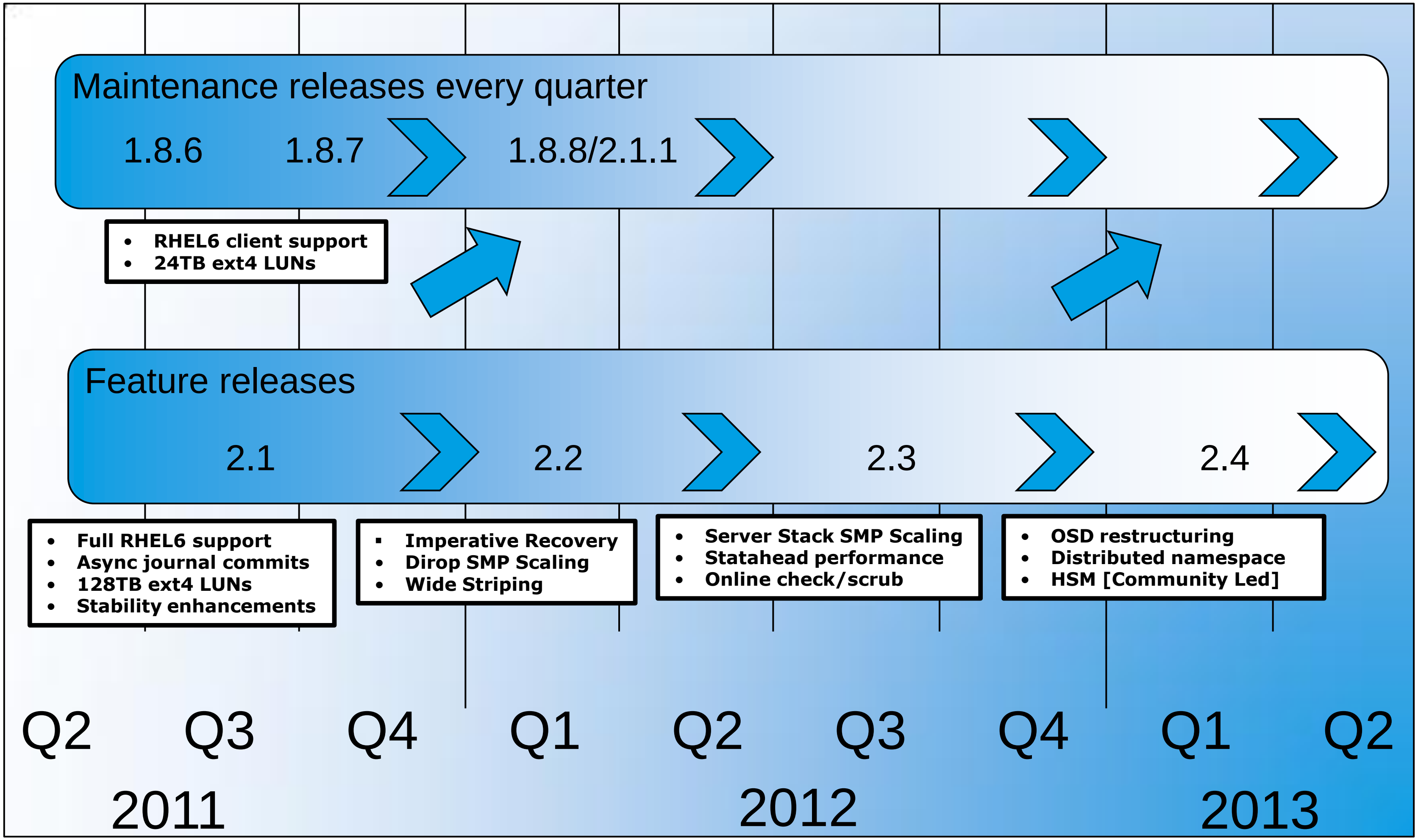




Distributed namespace

- Inodes on same MDT as parent dirent by default
  - Create scalable namespace using distributed (slower) operations
  - Use scalable namespace with non-distributed (fast) operations
  - Scale aggregate throughput
- Phase 1 – remote directories
  - Home/project dirs scattered over all MDTs
  - Home/project subdirs constrained to same MDT
- Phase 2 - striped directories
  - Directory entries hashed over directory stripes
  - O(n) speedup for shared dir ops (e.g. file-per-process create)
- Funded by OpenSFS

Whamcloud Lustre roadmap



Maintenance releases every quarter

1.8.6 1.8.7 1.8.8/2.1.1

- RHEL6 client support
- 24TB ext4 LUNs

Feature releases

2.1 2.2 2.3 2.4

- Full RHEL6 support
- Async journal commits
- 128TB ext4 LUNs
- Stability enhancements

- Imperative Recovery
- Dirop SMP Scaling
- Wide Striping

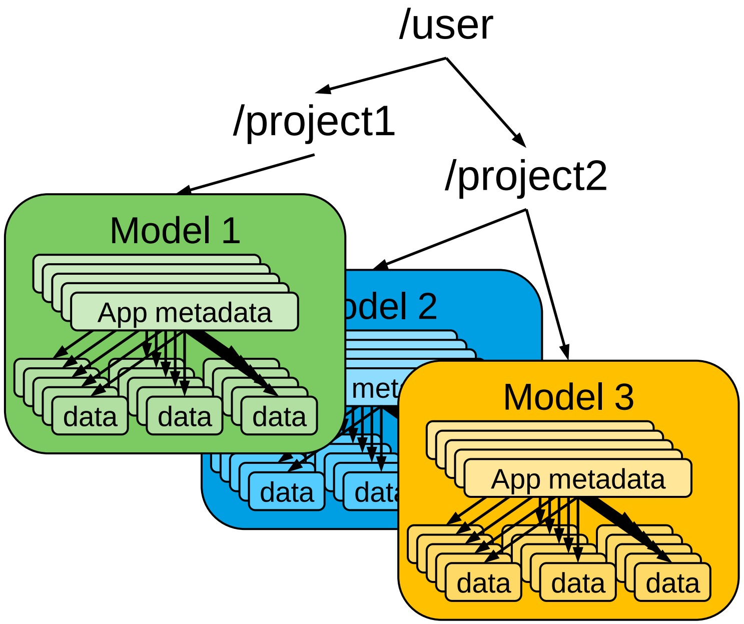
- Server Stack SMP Scaling
- Statahead performance
- Online check/scrub

- OSD restructuring
- Distributed namespace
- HSM [Community Led]

Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2

2011 2012 2013

Exascale challenges

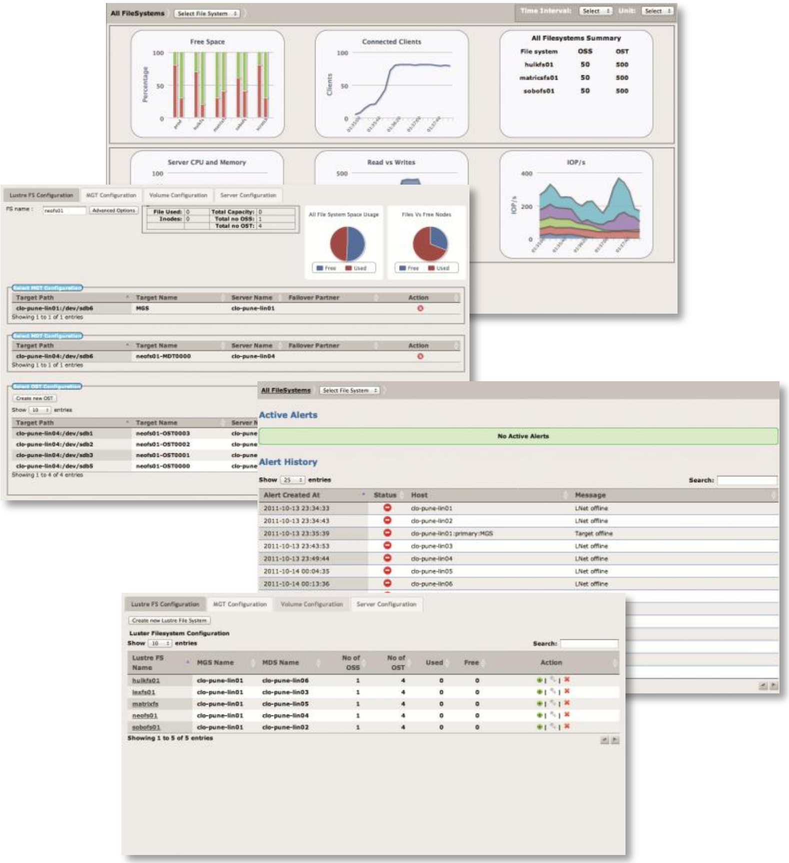


Application data + metadata

- Explosive growth
  - Large, sophisticated models
  - Uncertainty Qualification
  - Billions – trillions of “Leaf” data objects
  - Complex analysis
- Filesystem namespace pollution
  - Keep filesystem namespace for storage management / administration
  - Separate namespace for application data + metadata
    - Distributed Application Object Storage (DAOS) containers
- Preserve model integrity in the face of all possible failures
  - Very large atomic, durable transactions
  - Integrity APIs at all levels of the I/O stack
- Search / query / analysis
  - Non-resident index maintenance & traversal / non-sequential data traversal
  - Move query processing to global storage
    - Same programming model as apps?

Chroma: Lustre management for all

- Administration
  - Provisioning
  - Maintenance
  - HA Setup
  - Fault diagnosis
- Management information
  - Performance
  - Utilization
  - Alerts
- Intuitive interfaces
  - GUI (single pane of glass)
  - Scriptable CLI (automation)
- System integration
  - Multi-vendor storage management
  - Multi-vendor cluster/site management
  - Partners can build their own appliance



All Filesystems Summary

File system	OSS	OST
hulkfs01	50	500
matricsfs01	50	500
sobofs01	50	500

Lustre FS Configuration | MGT Configuration | Volume Configuration | Server Configuration

FS name : neofs01 Advanced Options

File Used:	0	Total Capacity:	0
Inodes:	0	Total no OSS:	1
		Total no OST:	4

Select MGT Configuration

Target Path	Target Name	Server Name	Failover Partner	Action
clo-pune-lin01:/dev/sdb6	MGS	clo-pune-lin01		⊗

Showing 1 to 1 of 1 entries

Select MDT Configuration

Target Path	Target Name	Server Name	Failover Partner	Action
clo-pune-lin04:/dev/sdb6	neofs01-MDT0000	clo-pune-lin04		⊗

Showing 1 to 1 of 1 entries

Select OST Configuration

Create new OST

Show 10 entries

Target Path	Target Name	Server N
clo-pune-lin04:/dev/sdb1	neofs01-OST0003	clo-pune
clo-pune-lin04:/dev/sdb2	neofs01-OST0002	clo-pune
clo-pune-lin04:/dev/sdb3	neofs01-OST0001	clo-pune
clo-pune-lin04:/dev/sdb5	neofs01-OST0000	clo-pune

Showing 1 to 4 of 4 entries

All FileSystems Select File System

Active Alerts

No Active Alerts

Alert History

Show 25 entries Search:

Alert Created At	Status	Host	Message
2011-10-13 23:34:33	⊖	clo-pune-lin01	LNet offline
2011-10-13 23:34:43	⊖	clo-pune-lin02	LNet offline
2011-10-13 23:35:39	⊖	clo-pune-lin01:primary:MGS	Target offline
2011-10-13 23:43:53	⊖	clo-pune-lin03	LNet offline
2011-10-13 23:49:44	⊖	clo-pune-lin04	LNet offline
2011-10-14 00:04:35	⊖	clo-pune-lin05	LNet offline
2011-10-14 00:13:36	⊖	clo-pune-lin06	LNet offline

Lustre FS Configuration | MGT Configuration | Volume Configuration | Server Configuration

Create new Lustre File System

Luster Filesystem Configuration

Show 10 entries Search:

Lustre FS Name	MGS Name	MDS Name	No of OSS	No of OST	Used	Free	Action
hulkfs01	clo-pune-lin01	clo-pune-lin06	1	4	0	0	⊕ | ✎ | ✖
lexfs01	clo-pune-lin01	clo-pune-lin03	1	4	0	0	⊕ | ✎ | ✖
matrixfs	clo-pune-lin01	clo-pune-lin05	1	4	0	0	⊕ | ✎ | ✖
neofs01	clo-pune-lin01	clo-pune-lin04	1	4	0	0	⊕ | ✎ | ✖
sobofs01	clo-pune-lin01	clo-pune-lin02	1	4	0	0	⊕ | ✎ | ✖

Showing 1 to 5 of 5 entries

Thank You

- Eric Barton

CTO
Whamcloud, Inc.
eeb@whamcloud.com