OPENFABRICS ALLIANCE

Update on Lustre and OpenSFS

John Carrier
Cray, Inc

3/27/2012

OpenSFS®
Open Scalable File Systems, Inc.

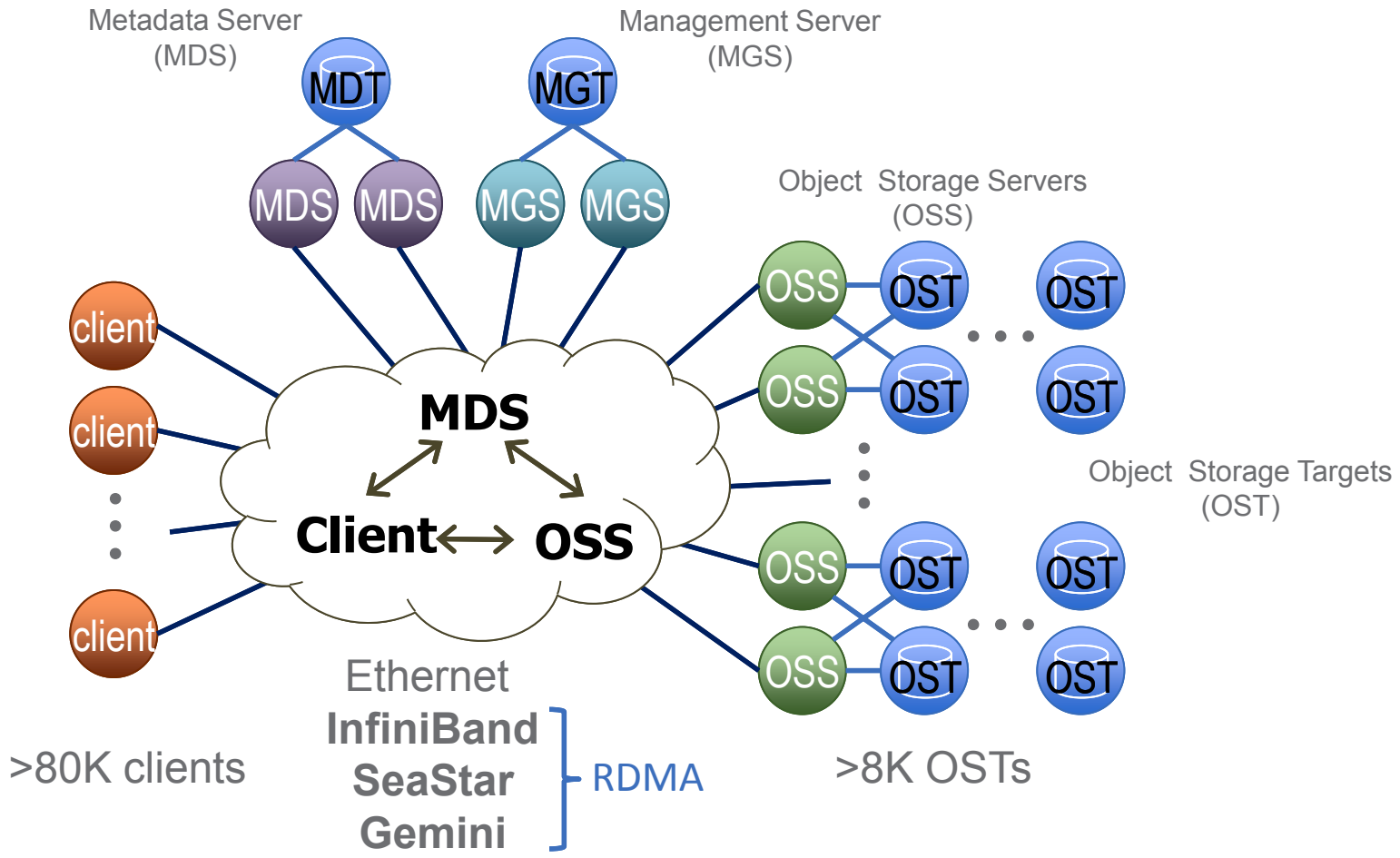
Outline

- Why is there a Lustre talk at an IB conference?
- How is OpenSFS ensuring Lustre's future?

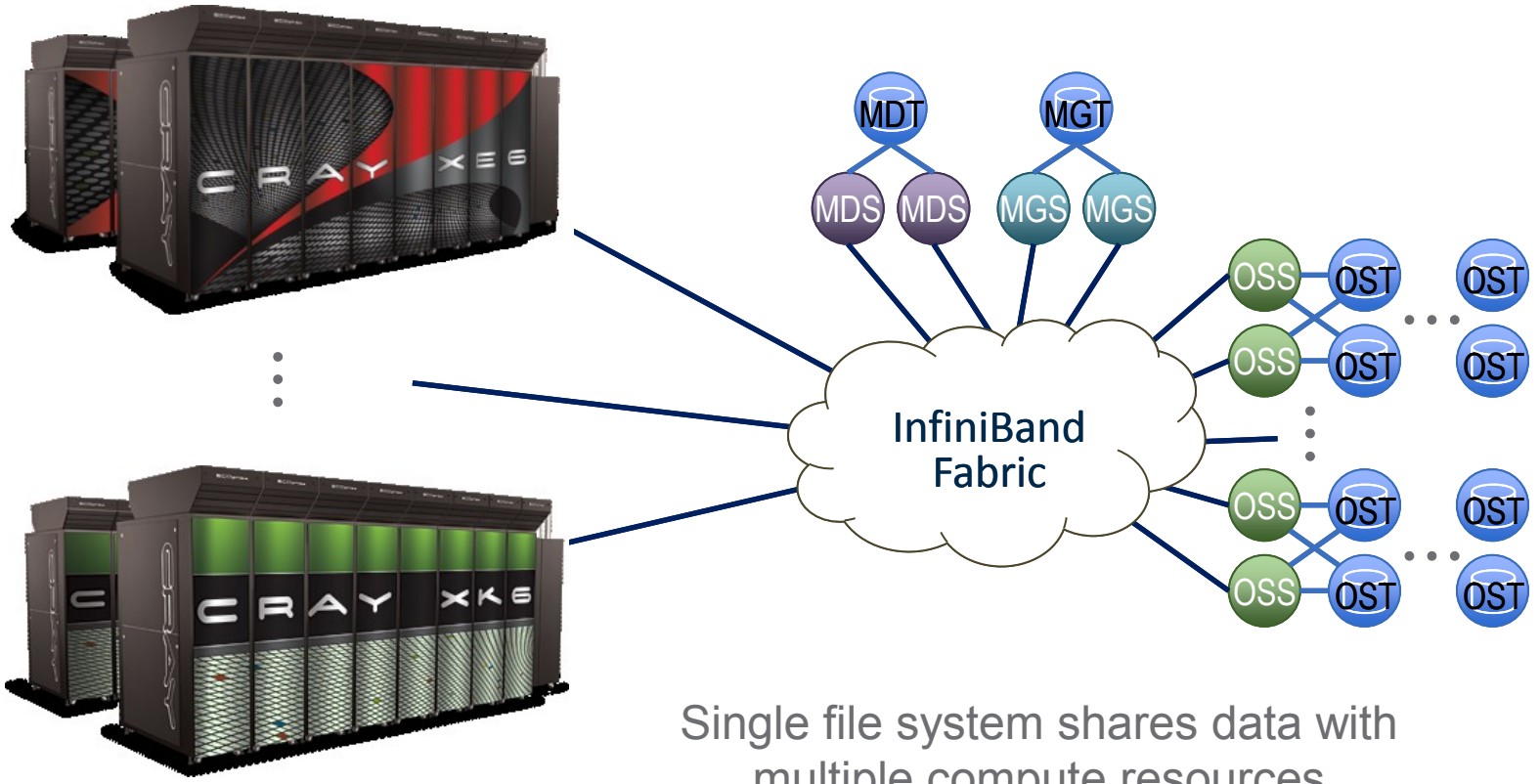
Lustre…

- …is a highly scalable, parallel file system
  - O(10,000) client nodes
  - supports both file per process (N-N) and shared file (N-1) access
- …runs on the world's fastest supercomputers
  - 64 of Top 100, 40 of Top 50, 8 of Top 10
- …is extremely fast
  - Two systems today over 240 GB/s
  - Three file systems to attempt 1TB/s this year!
- …is a major consumer of InfiniBand

Lustre Overview



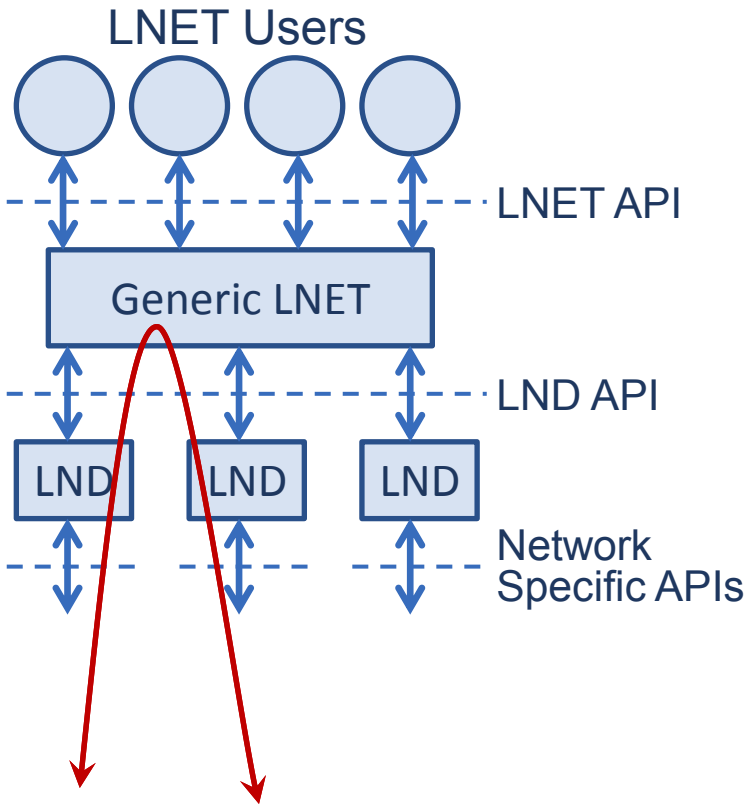
Shared Site-wide Lustre



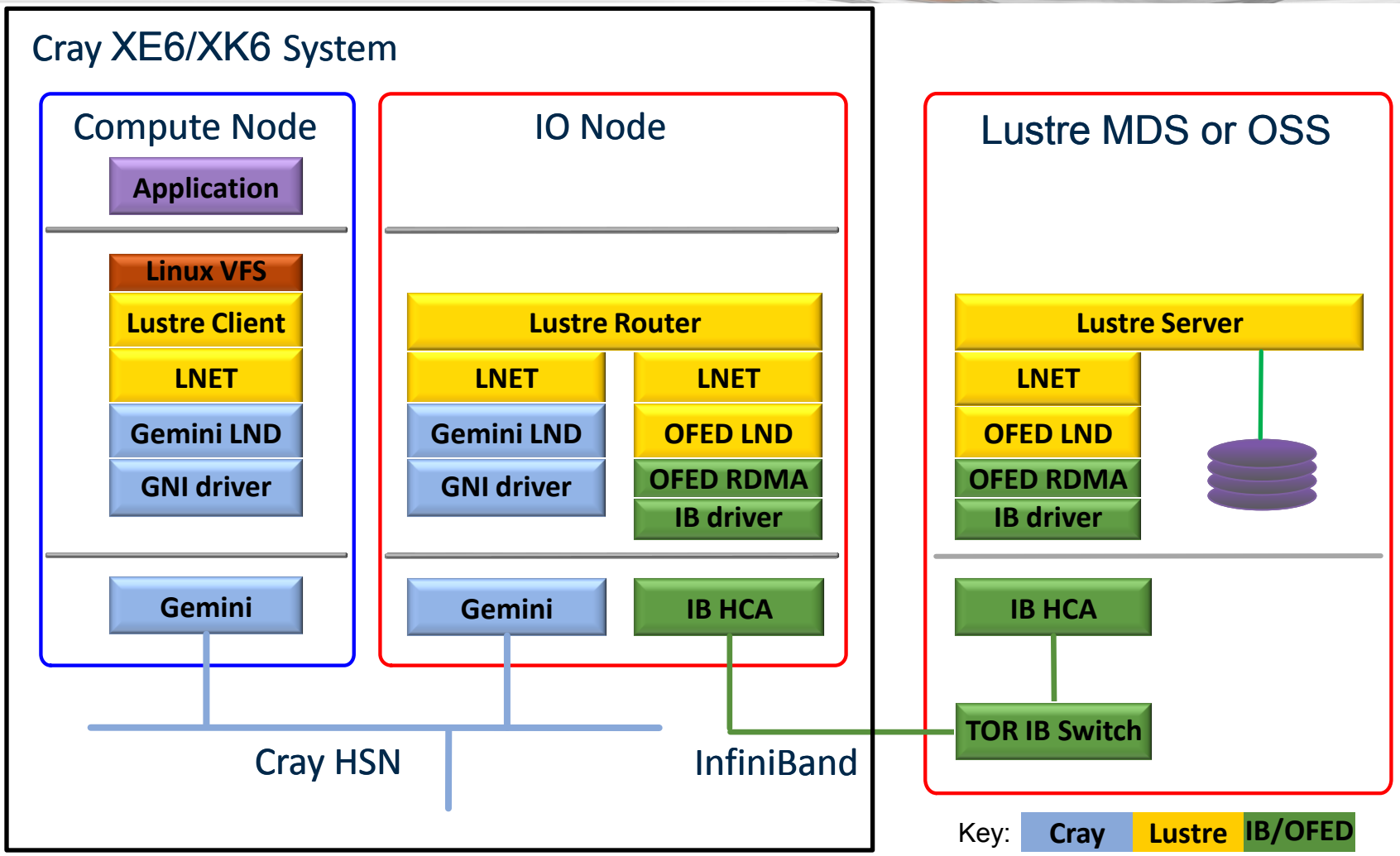
Single file system shares data with multiple compute resources

Lustre Networking

- Lustre remote procedure calls use the Lustre Network (LNET) Message Passing API for client-server communication
- Lustre Networking Drivers (LNDs) transport LNET messages across multiple network types
- LNDs make Lustre agnostic to the underlying transport
- LNET and LNDs can exploit RDMA for efficient, low CPU tranfsers
- Routing is possible at the LNET layer on nodes with two or more network interfaces

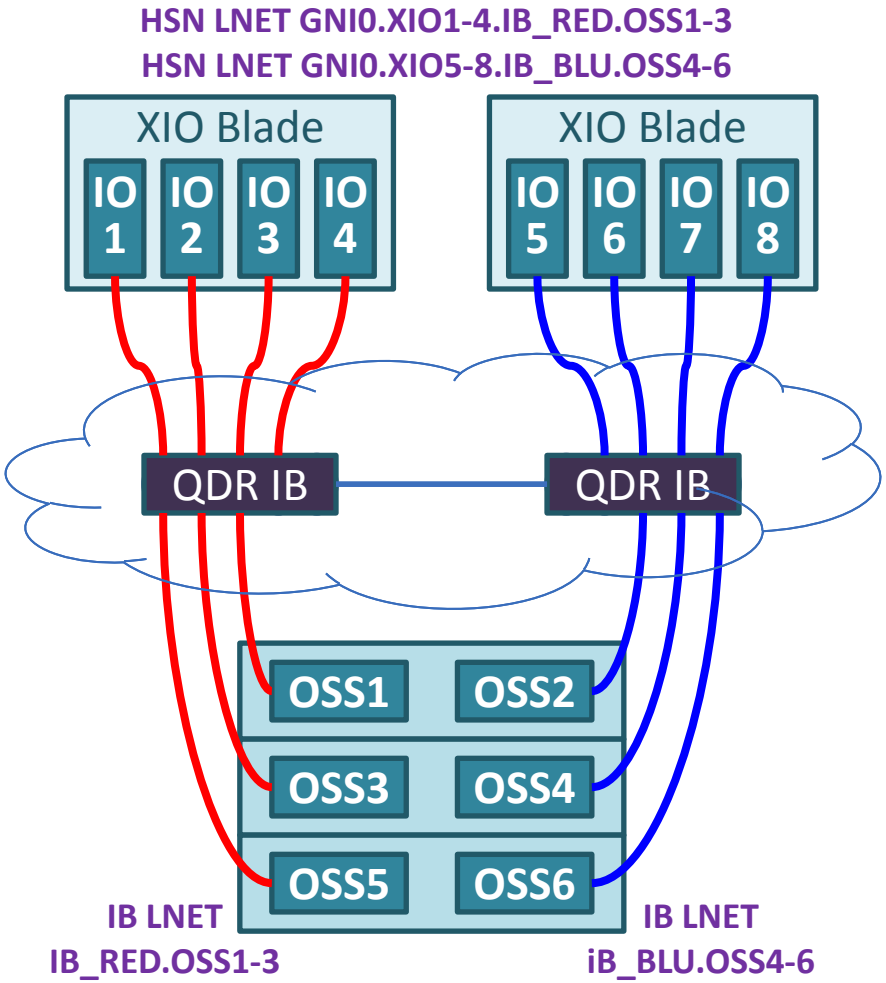


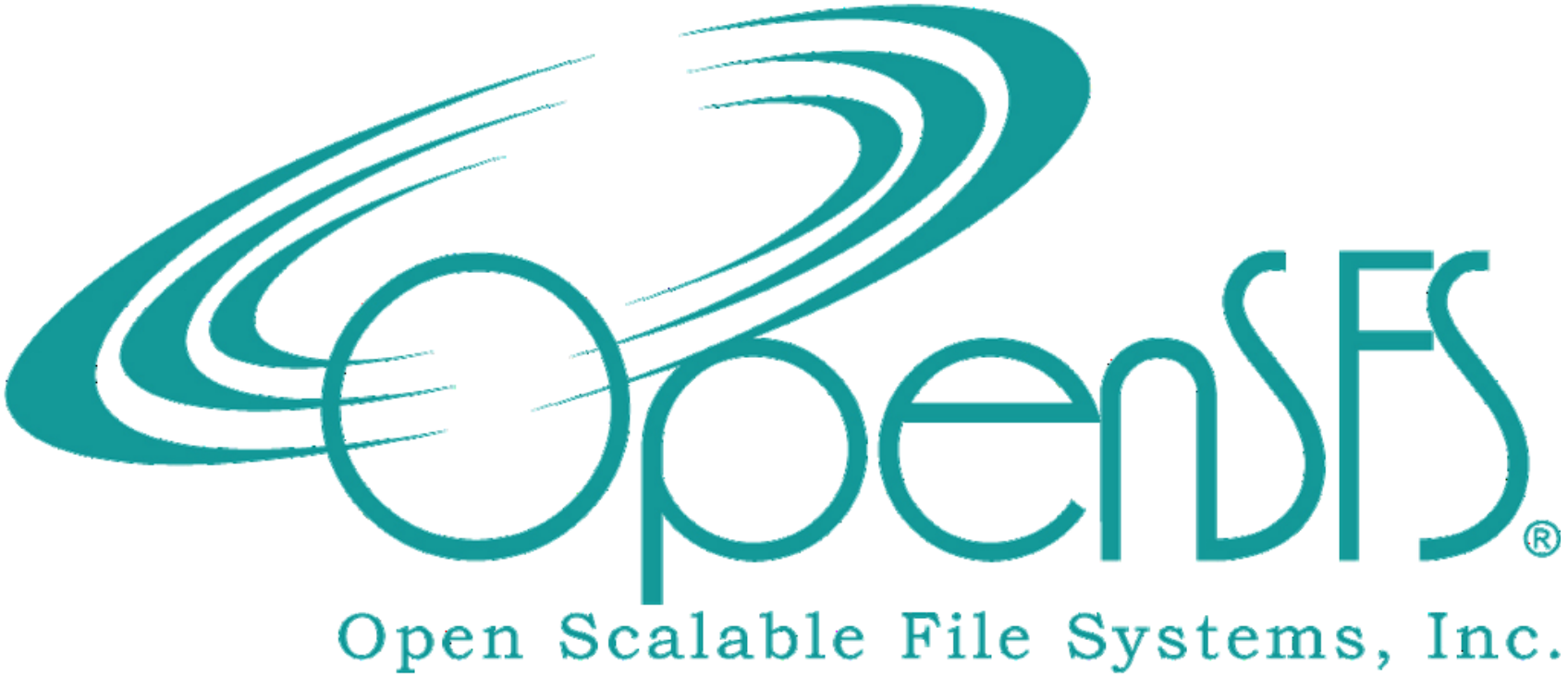
Lustre Router Software Stack



LNET Fine Grain Routing

- Lustre networking can create logical subnets within an IB fabric
- Define multiple LNETs to isolate I/O to specific physical paths through the fabric
- LNET Fine Grain Routing groups routers and OSSes connected through the same leaf switch
- Without link aggregation, failover and load balance at the router, not the HCA





http://www.opensfs.org/
http://lists.opensfs.org/

Lustre Community Transition

- In 2010, the Lustre Community moved to support Lustre on its own
  - Two companies—Whamcloud & Xyratex—hire Lustre developers and offer level three support to organizations
  - Three community groups—OpenSFS, HPCFS, EOFS—form to offer venues for collaboration and support for end-users
- In 2011, the Community consolidated
  - The three community groups quickly aligned their efforts to create a unified, global effort to support Lustre
  - OpenSFS gathered community requirements, solicited proposals, then funded a 2 year development contract to improve Lustre metadata performance and scalability
  - Whamcloud created Lustre 2.1 as the first community release and OpenSFS agreed to fund future releases
- In 2012, the community grows …

What is OpenSFS?

- OpenSFS is a vendor-neutral, member-supported non-profit organization bringing together the open source file system community for the high performance computing sector
- Our mission is to aggregate community resources and be the center of collaborative activities to ensure efficient coordination of technology advancement, development and education

The end goal is the continued evolution of robust open source file systems for the HPC community

Who is OpenSFS?

OPENFABRICS ALLIANCE

- Promoters

- Adopters

- Supporters

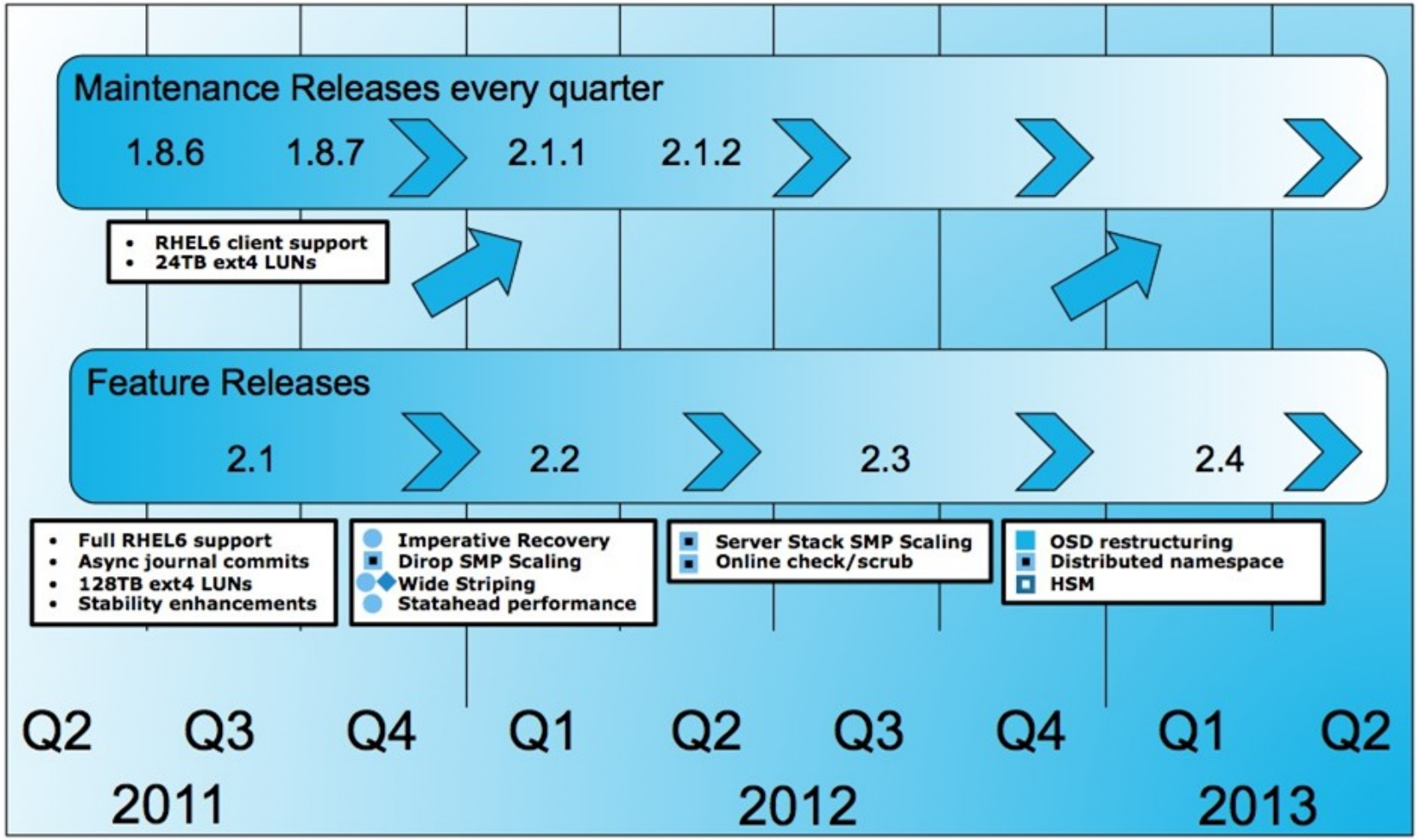
OpenSFS Working Groups

- Technical Working Group
  - Gather requirements from community
  - Propose and manage development projects
  - Generate Lustre feature roadmap
- Community Development Working Group
  - Manage Lustre releases
  - Coordinate release roadmap for new features
- Benchmarking Working Group
  - Investigate HPC user workloads
  - Define tests to evaluate file system scalability
- WAN Working Group
  - Coordinate use cases and features for wide-area Lustre

OpenSFS Development Projects

Feature	Purpose	Project
Single Server Metadata Performance Improvements	Scale-up strategy to remove MDS processing bottlenecks	SMP Node Affinity
		Parallel Directory Operations
Distributed Namespace	Scale-out strategy to enable multiple MDS per file system	Remote Directories
		Striped Directories
Lustre File System Checker	Monitor, validate, and repair file system state on-line	Inode Iterator & OI Scrub
		MDT-OST Consistency
		MDT-MDT Consistency
WAN authentication	Improve support for Lustre as a wide-area file system	UID Mapping
		GSSAPI Extensions

Lustre Community Roadmap



Source:http://wiki.whamcloud.com/display/PUB/Community+Lustre+Roadmap

Upcoming Activities

- Lustre Users Group (LUG)
  - Annual meeting is April 23-25, 2012 in Austin, TX
  - See http://www.opensfs.org/lug for more info
- OpenSFS Working Groups
  - Community Release WG
    - Whamcloud & OpenSFS to release Lustre 2.2 this spring
    - Seeking contributions for future releases
  - Technical WG
    - Gathering requirements for new development contracts
  - Benchmarking WG
    - Defining workloads, investigating tests
  - WG F2F meetings Sun 4/22 & Weds 4/25 in Austin

Summary

- Lustre uses RDMA natively and drives deployment of large scale IB fabrics
- OpenSFS members are funding Lustre development and future releases
- If your organization uses Lustre, contribute to the next release by joining OpenSFS

http://www.opensfs.org/
http://lists.opensfs.org/

Thank You