

# Technical Working Group

John Carrier, Dave Dillow  
TWG Co-Leads



SC11 TWG F2F  
Seattle, WA, November 14, 2011

# Agenda

- ▶ TWG Overview/Status
- ▶ OpenSFS PAC update
- ▶ OpenSFS Roadmap
- ▶ 2012 Development Priorities
  -

# TWG OVERVIEW/STATUS

# TWG Mission Statement

- Work with the Lustre community to ensure that Lustre continues to support the stability, performance, and management requirements of the OpenSFS members as HPC compute platforms continue to scale
- Responsible for creating and managing the roadmap for the OpenSFS community:
  - Gather requirements from the Lustre HPC community
  - Prioritize and recommend development projects to the Board
  - Initiate RFPs for important features
  - Work with contractors to meet these requirements

# TWG Accomplishments 2011

- ▶ Mar 2011 Publish Community Lustre Requirements
- ▶ Apr 2011 Issue RFP to solicit improvements to metadata performance and scaling, OSD quotas, lfsck
- ▶ Aug 2011 Announce multi-year, multi-million dollar development contract with Whamcloud (see below)
- ▶ Sep 2011 Review and recommend IU's WAN Proposal for funding
- ▶ Nov 2011 OpenSFS agrees to fund IU's Lustre WAN improvements  
*(announcement is pending contract discussions)*

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

# PAC UPDATE

# Project Milestones

- ▶ Scope Statement
- ▶ Solution Architecture
- ▶ High Level Design
- ▶ Implementation
- ▶ Demonstration
- ▶ Delivery

# Single MDS Performance

- ▶ Parallel Directory Operations
  - Break up single lock per directory
  - Liang Zhen presenting at Lustre Pavilion, Tues 2pm
- ▶ Progress
  - Scope Statement -- delivered
  - Solution Architecture -- delivered
  - High Level Design -- delivered
  - Implementation -- in progress
- ▶ Next Phase
  - SMP Node Affinity (start ~Jan-Feb 2012)



# Distributed Namespace

## ▶ Remote Directories

- Multiple MDTs
- no striping of directories

## ▶ Progress

- Scope statement -- delivered
- Solution architecture -- in progress

## ▶ Next Phase

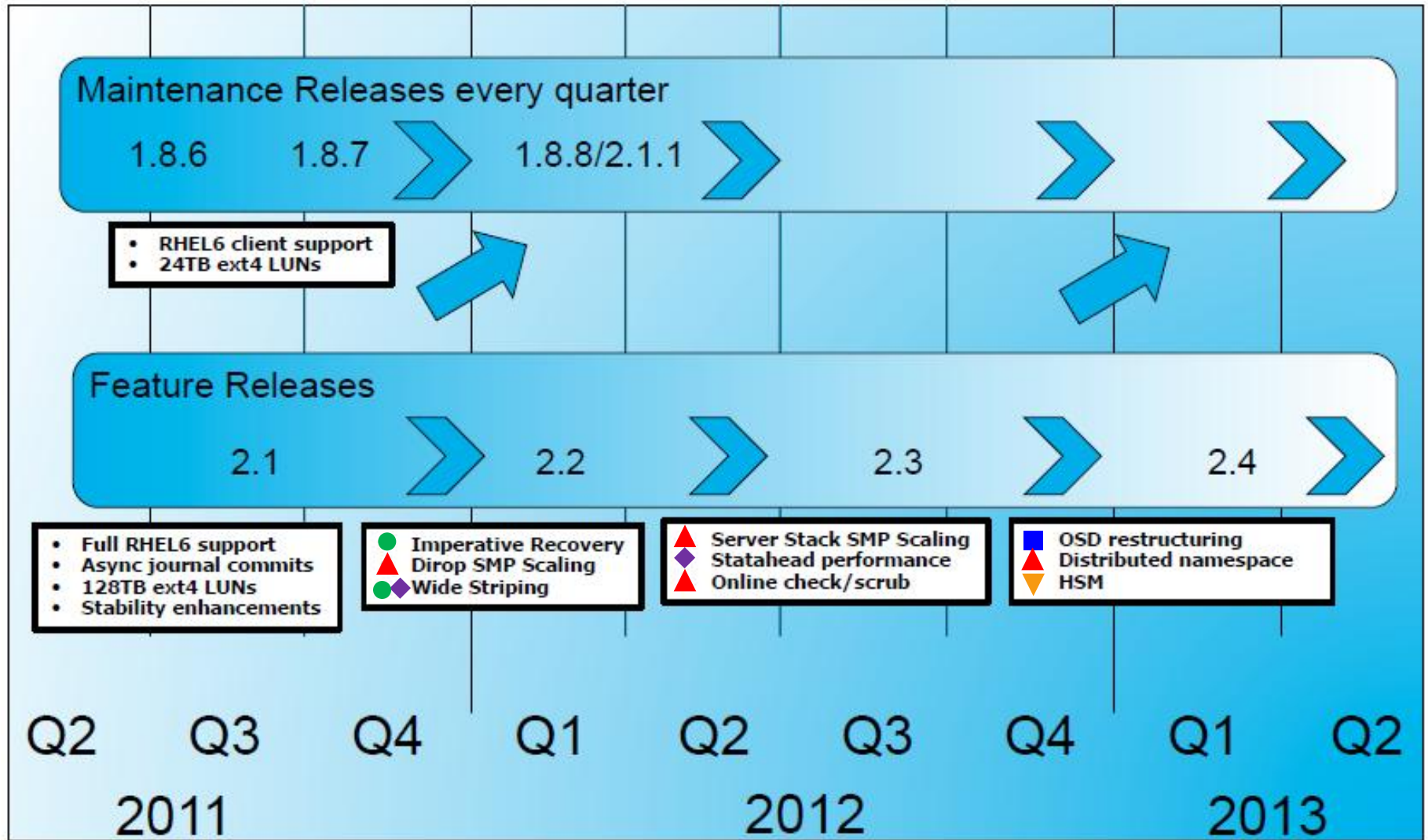
- Striped Directories will start after Remote Directories completes (Oct-Nov 2012)

# Lustre File System Checker

- ▶ Inode Iterator & OI Scrub
  - Verifies inode->FID mapping in object index table
  - Allows for file-level backups of MDT
- ▶ Progress
  - Scope Statement -- delivered
  - Solution Architecture -- delivered
  - High Level Design -- non-required
  - Implementation -- in progress
- ▶ Next Phase
  - MDT-OST Consistency (start Jan-Feb 2012)
  - MDT-MDT Consistency (start after DNE)

# OPENSFS ROADMAP

# Community Lustre Roadmap



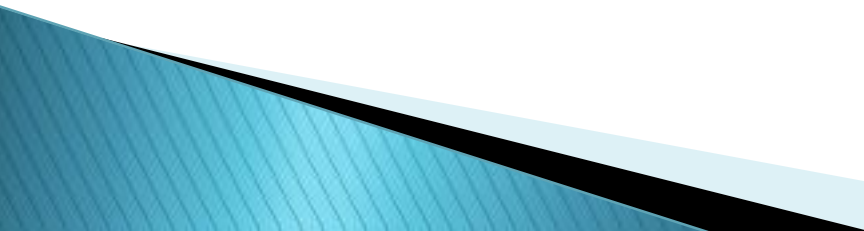
Contributor/Sponsor :    ◆ Whamcloud    ● ORNL    ▲ OpenSFS    ■ LLNL    ▼ CEA

# Roadmap Next Steps

- ▶ We need to show other contributions
- ▶ What other contributions are pending?
- ▶ How do we add to the Community Roadmap?

# OPENSFS REQUIREMENTS

# 2011 Prioritized Requirements

- ▶ Near-term requirements
    - Metadata server performance
    - Metadata server scaling
  - ▶ Long-term requirements
    - Support alternate storage backends
    - Start investigations of alternate storage backends
  - ▶ Improve the code base
    - Reduce maintenance effort
    - Reduce cost of new features
    - Pay-off our “technical debt”
- 

# Requirements Next Steps?

- Update the 2011 Requirements list (<http://goo.gl/cZSWG>)
- What other features should we drive?
- Are there foundational features we need for exascale?
- What other unsolicited proposals are lurking?



**For more information, e-mail  
discuss@lists.opensfs.org**



**BACKUP**

# 2011 Requirement Groups

- ▶ Performance Requirements
- ▶ Foundational Requirements
- ▶ Manageability and Administrative Requirements
- ▶ Application Interface Requirements

# 2011 Requirements List (1/3)

- ▶ Performance Requirements
  - ❖ Metadata server performance
  - ❖ Metadata server scalability
    - Single file performance
    - Quality of service
    - Locality and scalability
    - LNET channel bonding
- ▶ Foundational Requirements
  - Support for alternate backend file systems
  - Backend storage investigation
  - Scalable fault management

# 2011 Requirements List (2/3)

- ▶ Manageability and Administrative Requirements
  - Better support for newer kernels
  - Improved configuration of Lustre
  - Allowing for controlled partial-system maintenance
  - Balancing storage use
  - Adaptive storage layout
  - Arbitrary OST assignment
  - Better userspace tools
  - File system consistency checks
  - Snapshots
  - User Identity Mapping

# 2011 Requirements List (3/3)

- ▶ Application Interface Requirements
  - Improved storage semantics/interfaces
  - Better user tool API
- ▶ Other Requirements
  - Varying page-sizes
  - Mixed endian support



Open Scalable File Systems, Inc.