DataDirect™
N E T W O R K S
I N F O R M A T I O N   I N   M O T I O N™

The State of the Lustre® File System and The Lustre Development Ecosystem

Lustre User's Group

April 16, 2013

Dave Fellinger

Chief Scientist, Office of Strategy & Technology

The Lustre File System Project is a Technological Success Story

- The winner is still winning.
- The primary file system of the petascale club.
- Lustre is the Top500 leader.
  - ~50% of the Top 100
  - 30% of the 'known' Top500
- Lustre has become a leading choice in research computing institutions worldwide.
- Also many commercial market successes:
  - Seismic Processing, Financial, Web Content

Success Driven by Focused Scalability Goals

- Early (2001) scalability design goals were;
  - Aggregate 100GB/s bandwidth
  - 1000 nodes as clients (2 sockets, one core per socket)
- Today, the latest deployments will require;
  - Aggregate 1TB/s bandwidth
  - Over 300,000 cores
- Early design goals revolved around data capture.
- Today, the roadmap addresses HSM and data management.
- Early designs were limited in their use of Linux tools, which inhibited the flexibility of the implementation.
- Today, the Lustre effort is more autonomous than ever
  - Eg. OSD APIs will allow for a great deal of flexibility, usability & reliability
  - DAOS extends this even further, blurs the line between storage & middleware

OOM Above the Goal…

OAK RIDGE
National Laboratory

SPIDER II
THE WORLD'S FASTEST FILE SYSTEM

DDN and ORNL Have Partnered To Build The World's Fastest Storage System; Supporting Titan

System Performance: 1TB/s+
Capacity: 40.3PB (raw)
File System: Lustre®
I/O Platform: 36 x DDN SFA12K-40
Media: 20,160 HDDs

Fun Fact:
ORNL Titan Is Designed With The I/O Bandwidth Equivalent To 80,000x the Amount of Tweets and Tweet Metadata per Second from Twitter.

A Retrospective

- 'It all started with a moose'
  - A trip to Canmore Canada, Spring 2002
  - And an appreciation for designing for scale.
- Continuing with meeting the team
- Working on one of the earliest NALs
- Talking on the phone with Eric… maybe too much ☺
- Delivering the first test hardware to Phil Schwan's Boston basement
- The first Lustre support agreement at 0.6
  - Tungsten @ NCSA, 2,500 Dual Processor Dell Servers; 15TF, 10GB/s!
- One of the first production deployments of failover

11 Years later and we still work with the same team….

The Decentralization of the Community
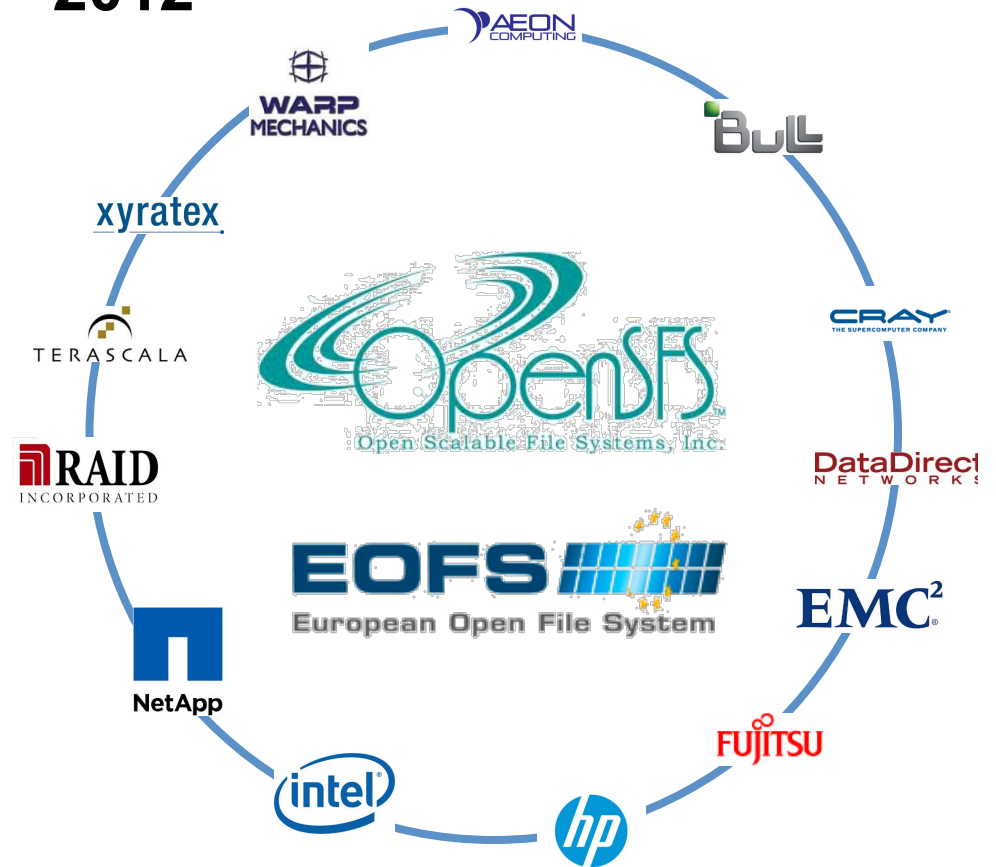

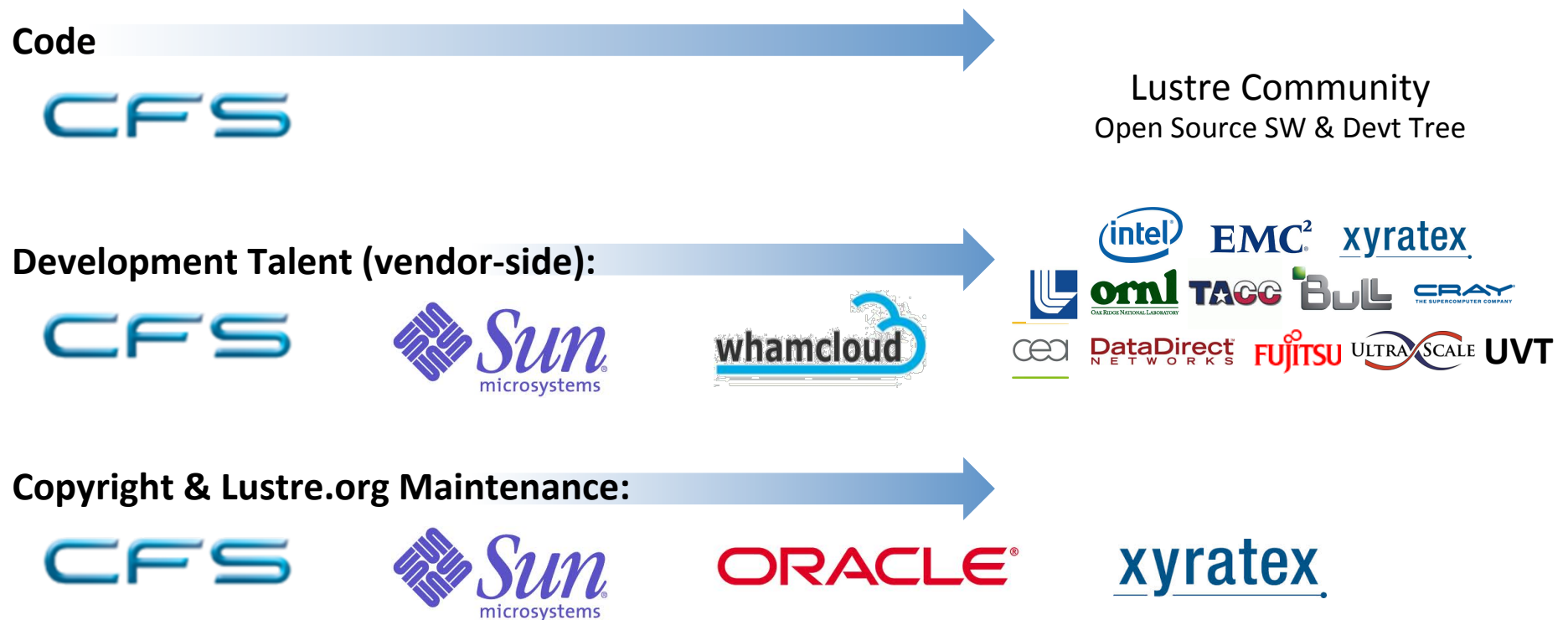**2003** → **2012**

Customers:

Developer:

Vendors:

Centralized Development;
Limited Contribution & Adoption

Community Driven by Users. Led by Contributors.
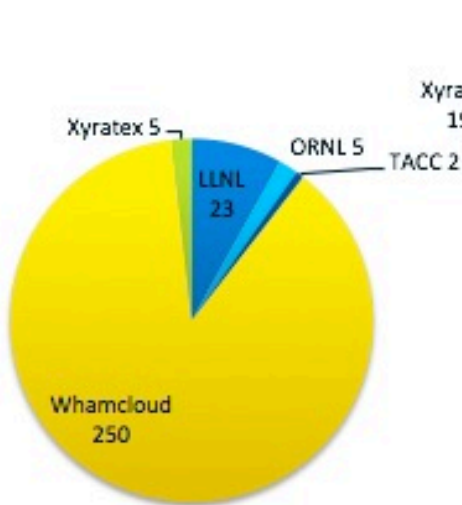Ecosystem Sponsored & Supported By Vendors.

6

The Evolution of Ownership

Code

CFS → Lustre Community
Open Source SW & Devt Tree

Development Talent (vendor-side):

CFS → Sun microsystems → whamcloud → intel EMC² xyratex
ornl TACC Bull CRAY
cea DataDirect NETWORKS FUJITSU ULTRA SCALE UVT

Copyright & Lustre.org Maintenance:

CFS → Sun microsystems → ORACLE → xyratex

The State of Lustre Development

Contribution By Number of Patches.

2.1

- Xyratex 5
- LLNL 23
- ORNL 5
- TACC 2
- Whamcloud 250

5 Contributors

2.2

- Xyratex 19
- Bull 1
- CEA 3
- Cray 5
- DDN 3
- LLNL 10
- NICS 1
- ORNL 15
- TACC 3
- Whamcloud 258

10 Contributors

2.3

- Ultrascale 2
- UVT 1
- Xyratex 28
- Bull 4
- CEA 3
- Cray 4
- DDN 1
- Fujitsu 1
- TACC 5
- ORNL 17
- LLNL 19
- EMC 38
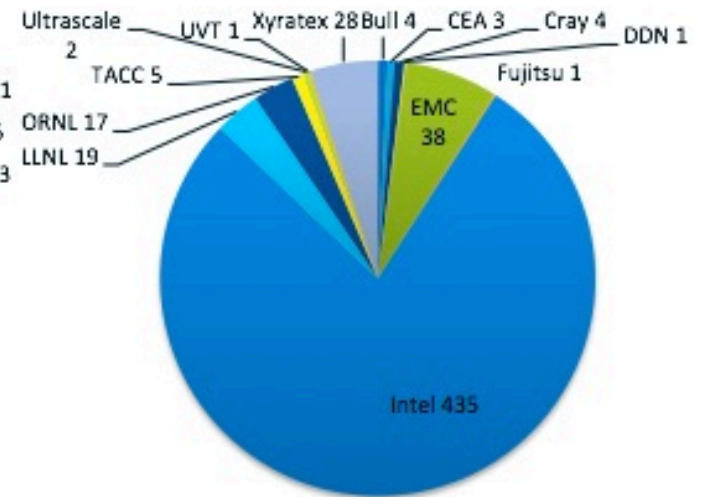- Intel 435
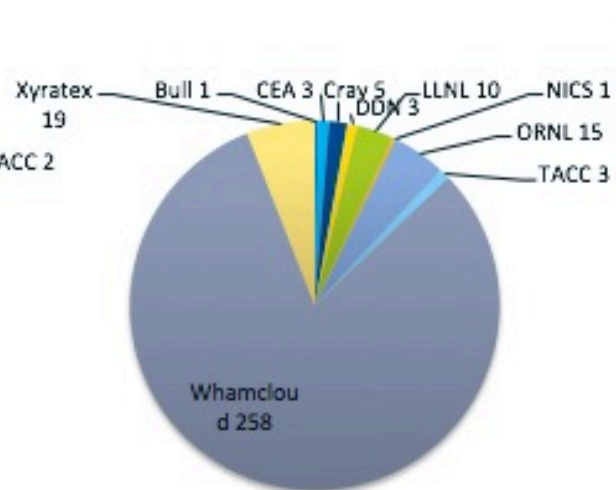
13 Contributors

How are we doing?

- We ran a survey to pulse check the ecosystem
- Why?
  - The world is changing rapidly.
  - Lustre is a moving target… but other technologies and trends are also moving
  - Always good to know how we're doing.

Survey Results

- Over 50% of the respondents use Lustre for scratch and >95% are satisfied.

- Over 20% use it for home directories but almost half of those are dissatisfied.

- 80% plan to expand Lustre usage.

- Over 40% purchase external support today.

- Respondents report 6-7 people within their organizations involved in Lustre deployment and support

- >50% indicated that they will expand their involvement in the Lustre community in the next 12 months

Survey Results

- The top developments for organizations to CONTINUE using Lustre include;
  - Performance
  - Stability
  - Heterogeneity – more OS's, faster support of new releases
  - Tools for logging, HSM and, replication
- The developments for organization to EXPAND the use of Lustre include;
  - Performance
  - Heterogeneity – more OS's, faster support of new releases
  - Tools for usability, HSM, logging, simplified support

Conclusion

- Thank you, Larry. The community is thriving.

- The Users have spoken:
  - 95% satisfaction for a scratch file system

- At a time of upheaval, 2 releases were delivered in 2012;
  - 2.2 with improved performance, faster recovery, larger stripes
  - 2.3 with object index scrub and, improved SMP performance

- A roadmap with designated contributors though 2014 including HSM, better consistency, exascale foundation.

- A proven history of **shared** focus and effort.

Congratulations on a technical, business & community success, Exascale Computing… here we come.

… but don't forget.

Disruption Is EVERYWHERE

LUG & OpenSFS Must Keep Our Shared Efforts Focused

Thank You

Questions?