whamcloud

Berlin Buzzwords
June 2011

The Lustre Filesystem

- Eric Barton
  CTO
  Whamcloud, Inc
  eeb@whamcloud.com

Whamcloud Introduction

- Formed July 2010, California VC-backed corp.
- ~40 employees worldwide
- Committed to:
  - Open Source Lustre
  - Open Development, professional management
  - Vendor neutrality
- Maintains Lustre community resources
  - Wiki: http://wiki.whamcloud.com
  - All Lustre releases: http://downloads.whamcloud.com
  - Online bug database: http://jira.whamcloud.com
  - Git repositories: http://git.whamcloud.com
  - Jenkins build: http://build.whamcloud.com
  - Gerrit code review: http://review.whamcloud.com

Lustre Timeline

- 1999 – Lustre project starts
- 2003 V1.0 – Cluster File Systems
- 2007 V1.6 – Sun Microsystems
- 2009 V1.8 – Sun Microsystems
- 2010 V2.0 – Oracle Corp.

- V1.8.5 in widest use today – most robust

The Lustre file system

- Vendor neutral
- Open Source – GPLv2
- Community embraced and supported
  - EOFS
  - OpenSFS
- The industry leading FS for HPC

TOP 10 Systems - 11/2010

1	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C
2	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz
3	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU
4	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows
5	Hopper - Cray XE6 12-core 2.1 GHz
6	Tera-100 - Bull bullx super-node S6010/S6030
7	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband
8	Kraken XT5 - Cray XT5-HE Opteron 6-core 2.6 GHz
9	JUGENE - Blue Gene/P Solution
10	Cielo - Cray XE6 8-core 2.4 GHz

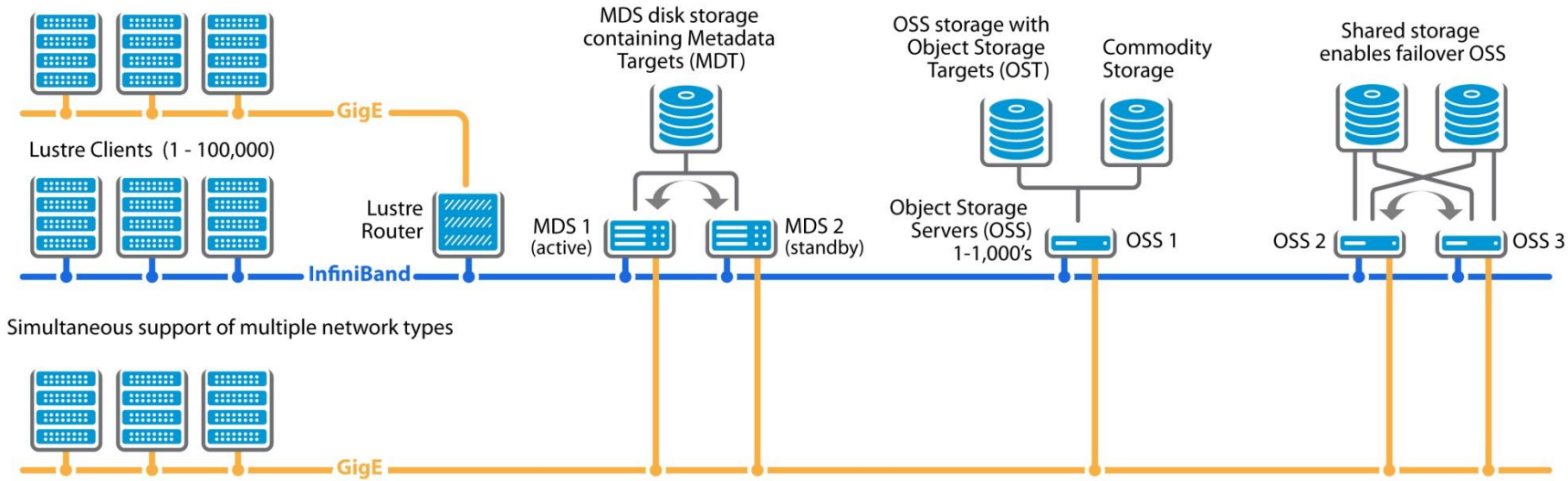
Why is Lustre so Popular

- Performance
- Scaling
- POSIX Compliance
- Performs in both LAN and WAN Environments
- Open Source
- Works with any block storage
- Wide and active development community
- Maturity and stability
- Wide and growing adoption

The Lustre file system

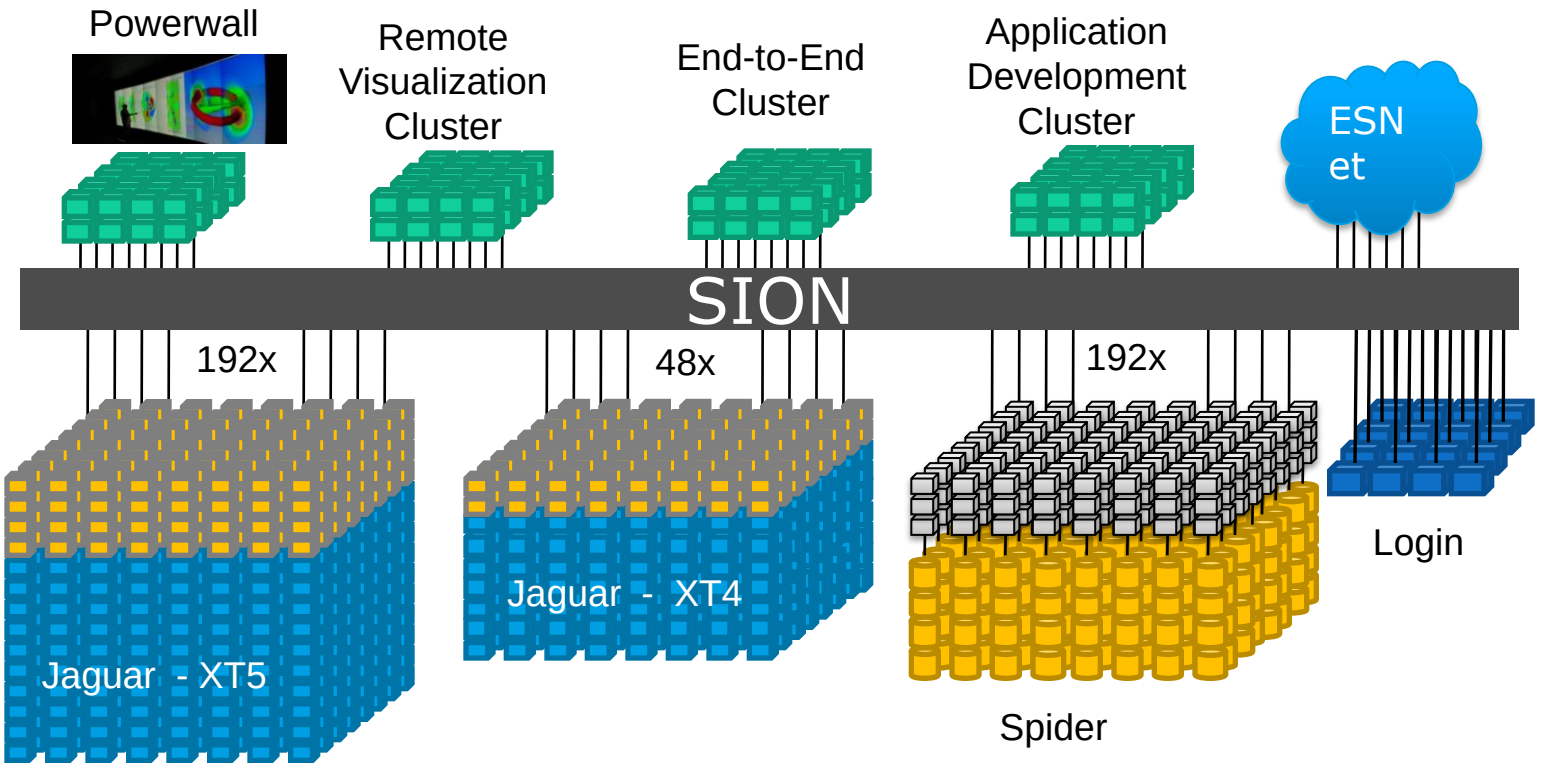
- Posix
  - Conventional application I/O model
- Strongly coherent
  - Client caching
    - Metadata: read-only
    - Data: read and write
      - Extent locks
  - Distributed Lock Manager
    - Callbacks delivered on locking conflicts
- HA
  - Transparent server failure
  - Replay on failover/restart

The Lustre file system



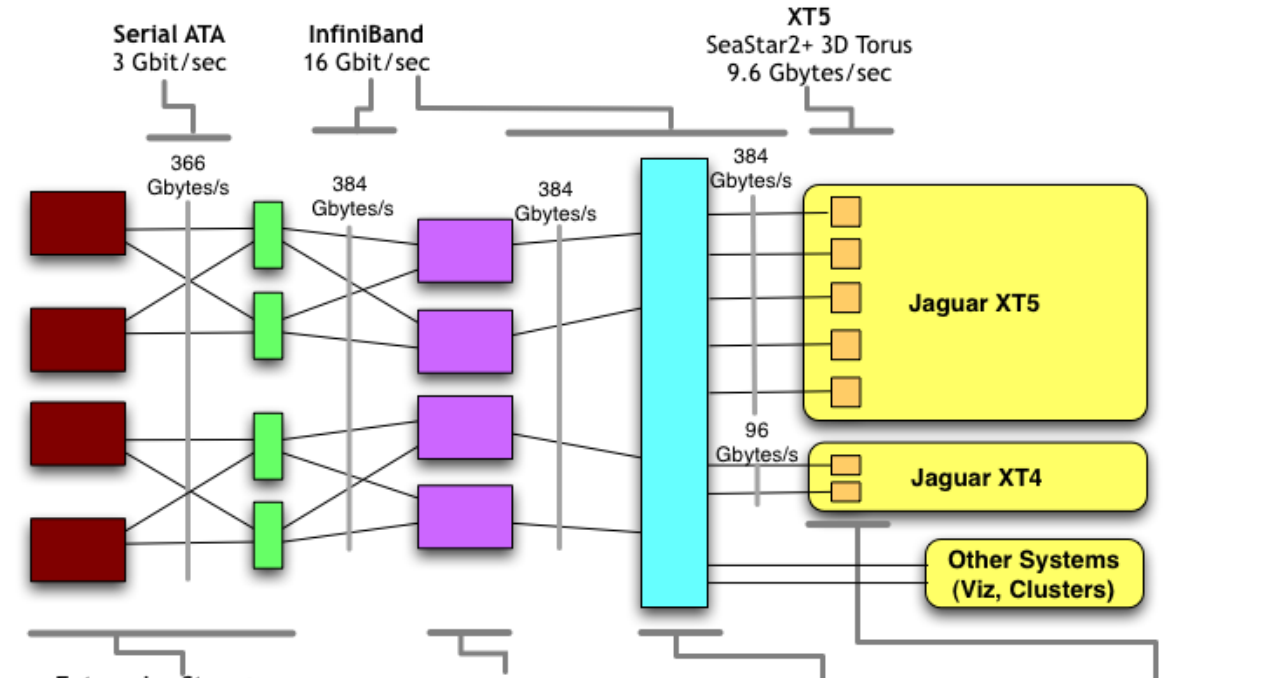
- Native networking
  - RDMA
  - Redundant routers
- High performance streaming
- Resilient Object Storage
  - Failover server pairs
    - OSS – active-active
    - MDS – active-passive
- Data striped over OSTs

ORNL Spider



- Site-wide filesystem
  - Decouples compute/storage
  - Over 26,000 client nodes
- High Performance
  - Over 282,000,000 files
  - Multiple petabytes of data
  - 240GBytes/s

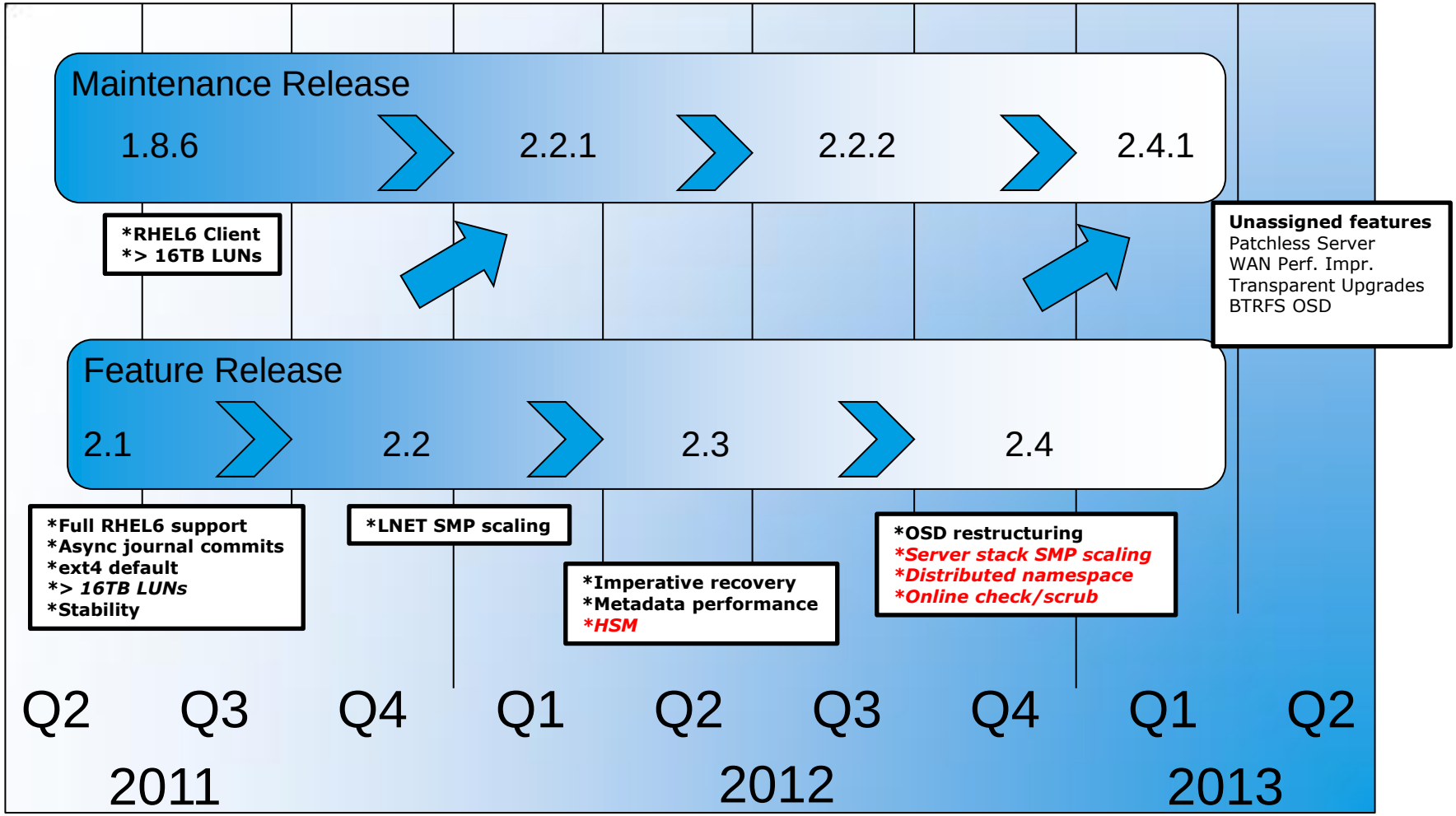
ORNL Spider: feeds & speeds



Status today

- Lustre available from several sources
  - A common source tree:
    - 1.8.* release from Oracle/Whamcloud
    - 2.1 community release from Whamcloud
    - Other parties doing their own releases
- Whamcloud actively developing
  - Roadmap presented at LUG
  - Metadata performance enhancements
  - Imperative recovery
  - ZFS/alternate infrastructure
  - Other development activities in planning

Lustre Roadmap (wiki.whamcloud.com)



MapReduce on Lustre

Experiments…

- Where to Map?
  - On OSS nodes good for I/O intensive?
  - On dedicated client nodes for CPU intensive?
  - OSS nodes typically tuned to stream between disk and network
- How to distribute input/output files
  - Many single-stripe files distributed over all OSSes?
  - Few files striped over all OSSes?
- How to handle tmp files
  - Dedicated local filesystem?
  - Lustre
    - Careful placement v. scatter?

Thank You

- Eric Barton
  CTO, Whamcloud Inc.
  eeb@whamcloud.com