Update on Lustre Filesystems at LC

LC User Meeting

Ned Bass
Software Development Group Leader

September 8, 2016

LLNL-PRES-703009
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Agenda

- Summary of current and planned Lustre filesystems
- New hardware
  - RAID Inc supplying new filesystem hardware
  - ZFS Software RAID in place of hardware controllers
- New software
  - Lustre 2.8
  - ZFS on Linux 0.7 (Lustre's backend filesystem)
  - New features, performance enhancements, and bug fixes

Current Parallel File System Summary (OCF)

OCF File Systems	Bandwidth	Capacity	OSS	OSTs
lscratchrzb (Stout)	18GB/s	1.2PB	16	16
lscratchf (Cider)	36GB/s	2.4PB	16	32
lscratchd (Pilsner)	90GB/s	5.7PB	80	80
lscratche (Porter)	90GB/s	5.7PB	80	80
lscratchv (Vesta)	106GB/s	6.7PB	96	96

* Multiple MDS nodes will be utilized in the future when LC stability requirements.

Current Parallel File System Summary (SCF)

SCF File Systems	Bandwidth	Capacity	OSS	OSTs
lscratch1 (Grove)	850GB/s	53PB	768	768
lscratch7 (Lambic)	90GB/s	5.7PB	80	80
lscratch3 (Marzen)	90GB/s	5.7PB	80	80
lscratch6 (Bock)	90GB/s	5.7PB	80	80
lscratchs1 (SNSI)	4.5GB/s	300TB	4	4

* Multiple MDS nodes will be utilized in the future when LC stability requirements.

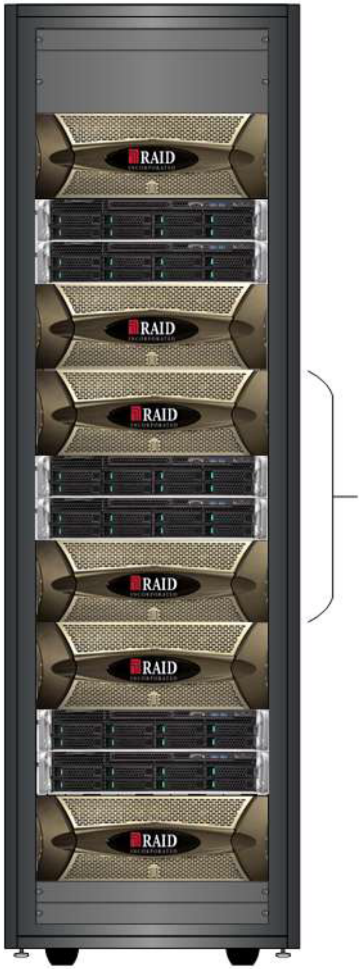
Upcoming Parallel Filesystems for CTS-1 systems

OCF File Systems	Mounted on	Bandwidth	Capacity	OSS	MDS
lscratchh (Zinc)	Quartz	60 GB/s	18PB	36	1 (eventually up to 16)
lscratchrzj (Brass)	RZTopaz	30 GB/s	9PB	18	1 (eventually up to 4)

These will not be mounted on existing TOSS 2 clusters!

Multiple MDS nodes will be utilized when that feature meets LC stability requirements.

New filesystems from RAID Inc. use JBODs and ZFS software RAID for OST storage



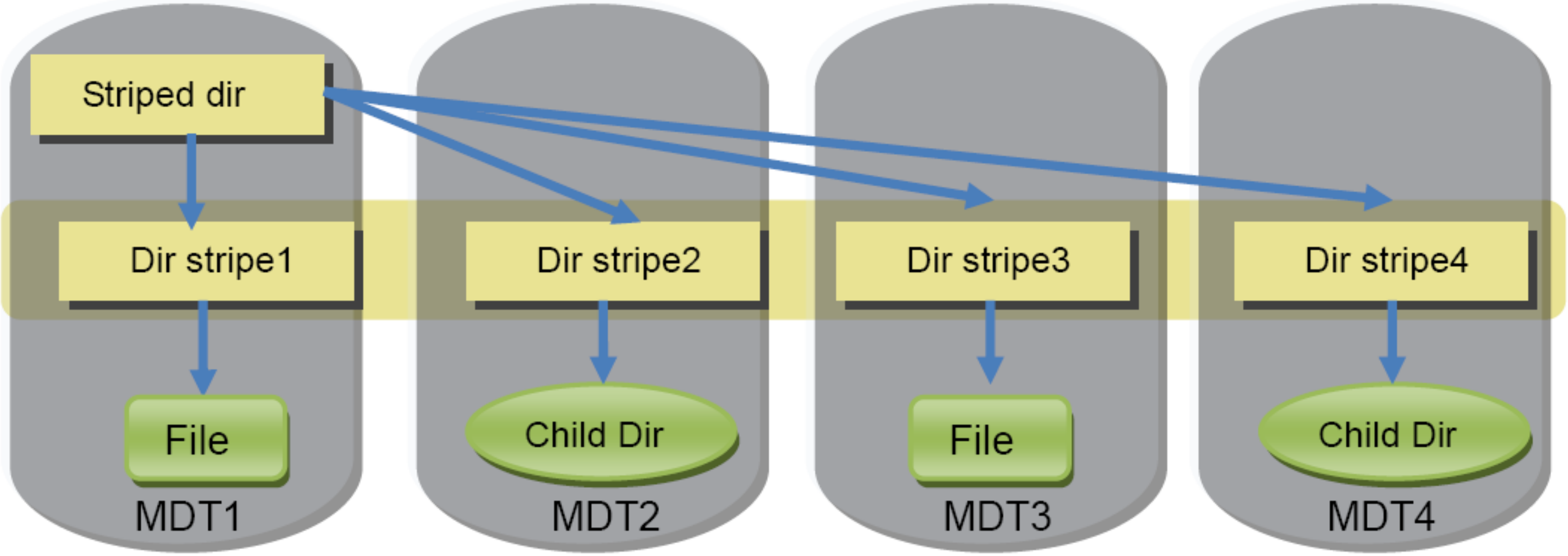
Each Storage Scalable Unit (SSU) fills one rack and contains:

- 6 OSS server nodes
- 6 JBOD disk enclosures
- 480 NL-SASdisks
- 2.8 PB Usable storage

ZFS RAID offers better value, resiliency, and performance than hardware controllers

- Ability to automatically self-heal data corruption
- Superior capacity to cost ratio
- Vendor agnostic storage software stack
- ZFS performs best when managing individual disks

Lustre 2.8 introduces support for fully distributed metadata



New lfs subcommands to manage directory striping:

```
lfs {setdirstripe|getdirstripe} …
```

Striped directories offer better performance of metadata operations such as file creates, removes, and stats.

LC does not yet consider distributed metadata in Lustre 2.8 production ready

- Development/QA finding and working through bugs
- Optimistic forecast: ready early 2017
- Lustre 2.8 filesystems will be deployed with a single MDS in the meantime
- Cannot easily add MDS nodes to an existing filesystem
- LC will provide migration path to multiple-MDS Lustre filesystems when the feature is ready

Other reasons LC is moving to Lustre 2.8

- Many bug fixes since Lustre 2.5
- SELinux support
- Performance improvements in ZFS and Lustre
- Better support from our vendor and the community
- Support for RHEL 7 / TOSS 3

Lawrence Livermore
National Laboratory