2018 LC User Meeting
Lustre Update

LC Users

Cameron Harr
LC SAG

February 8, 2018

LLNL-PRES-745892
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

- What's new?
  - Hardware
  - Software
- What's available now?
  - SCF
  - CZ
  - RZ
- What's different?
  - More PBs
  - More IOPs
  - More metadata management
- What's coming?

What's New: Hardware

- New generation of Lustre systems and IB
- Added 3 new Lustre file systems in 2017:
  - CZ: lscratchh (15 PB) in January
    - 15 PB
  - RZ: lscratchrza (7.5 PB) in April
    - 7.5 PB

- SCF: lscratch2 (15 PB) in October
  - 15 PB

What's New: Software

- Lustre 2.8
  - Previous: Lustre 2.5
  - Distributed metadata
    - Now 4 or 16 Meta Data Servers (MDSs) instead of 1
    - Data resides on multiple Meta Data Target (MDT) disks
  - Faster IOP/s

- ZFS 0.7
  - Large block support
    - LC default record size: 128K -> 1MB
  - Many administrative improvements

What's Available?

- Generational overlap = 1-3 years
- OCF: 43.2 PB

CZ F/S	Bandwidth (GB/s)	Capacity (PB)
lscratchd	90	5.7
lscratche	90	5.7
lscratchf	36	2.4
lscratchh	**108**	**15**
lscratchv	106	5.9

RZ F/S	Bandwidth (GB/s)	Capacity (PB)
lscratchrza	**54**	**7.5**
lscratchrzb	18	1.0

What's Available (cont.)?

- SCF: 85.1 PB

SCF F/S	Bandwidth (GB/s)	Capacity (PB)
lscratch1	850	53
lscratch2	**108**	**15**
lscratch3	90	5.7
lscratch6	90	5.7
lscratch7	90	5.7

https://hpc.llnl.gov/hardware/lustre-parallel-file-system

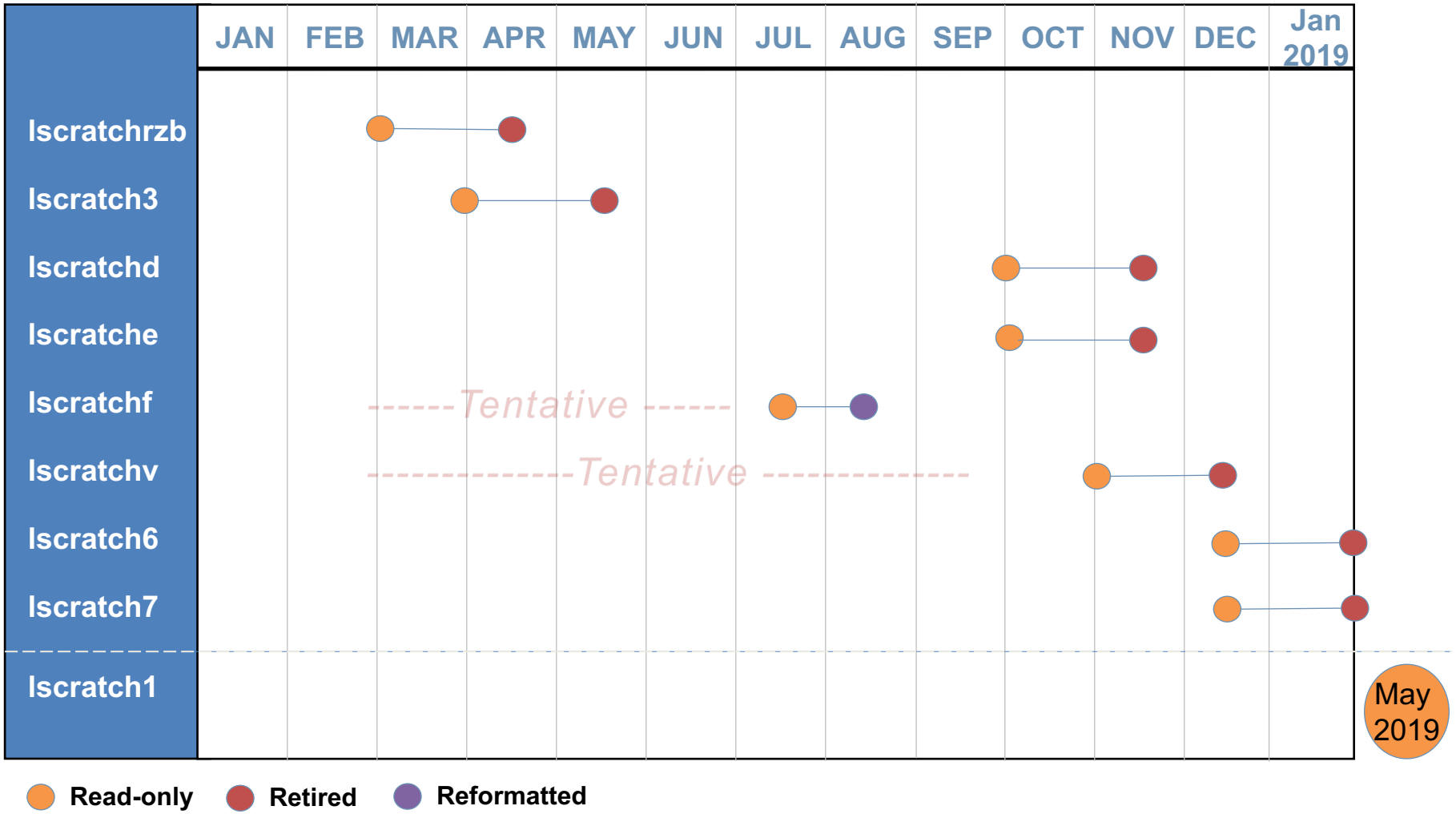
What's different?

- Added 22.5 PB to OCF, 15 PB to SCF
- More MDS nodes:
  - Faster IOPs
    - >100K
  - Resiliency: Single MDS failure won't take file system down
- Lustre DNE v1:
  - User directories randomly assigned to single MDT
  - Pro: Increased performance, resiliency
  - Con: Heavy users can fill up smaller MDT
    - Let LC know if you need to create 10s Millions of files
    - Use latest version of MFEM

What's Coming?

- New 15 PB SCF filesystem Q3 2018
- New 15 PB CZ filesystem Q3-Q4 2018
- Lustre 2.10
  - Progressive File Layout (PFL)
    - Let Lustre worry about striping your files!
- Retirement of Lustre 2.5
  - SCF
    - Lscratch[3,6,7] in 2018
    - Lscratch1 in 2019
  - OCF
    - Lscratch[d,e,v] in 2018
    - Lscratchf reformatted
    - Lscratchrzb in 2018

Lustre 2.5 Retirement



What's Coming?

- Retirement Process
  - File system will be Read-Only (RO) 6+ weeks before retirement
    1. File system mounted RO on compute clusters = 2 weeks
       - Jobs preventing remounts killed
    2. File system mounted RO **only** on SLIC nodes = 4 weeks
  - **File system will be destroyed upon retirement**
  - **Users responsible for migrating data**
    - Start now by using the new file systems already in production
      - Lscratch2 (SCF)
      - Lscratchh (CZ)
      - Lscratchrza (RZ)


Lawrence Livermore
National Laboratory