Rocks, Rabbits and Snakes, Oh My!

LUG 2025

Cameron Harr
Lustre Operations

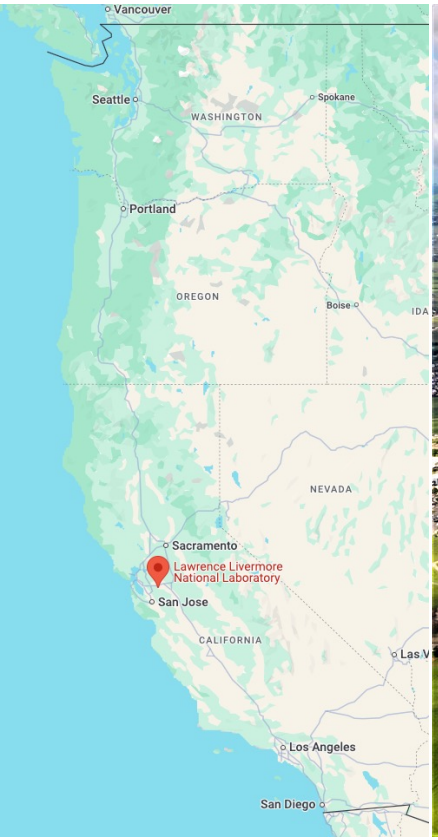
April 2, 2025

LLNL-PRES-2003913

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 Laboratory

Lawrence Livermore National Lab

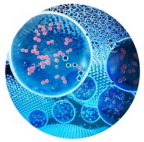


- Land commandeered for NAS in WWII
- Est. 1952 as UC Radiation Lab offshoot
  - Became LLNL in 1971
- Set up to compete with LANL
  - But with lots of collaboration
- ~ 9000 Employees

What do we do?

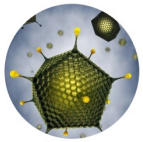
- *"Our mission is to enable U.S. security and global stability and resilience by empowering multidisciplinary teams to pursue bold and innovative science and technology."*

- US Dept. of Energy -> NNSA
- We're a science lab
  - Special nuclear security and design mandate
  - But so much more…



Advanced Materials and Manufacturing

Designing unique materials and fostering innovation to support our mission.



Bioscience and Bioengineering

Preventing and responding to current and future biological and environmental threats.

Earth and Atmospheric Science

Tackling climate challenges and understanding Earth processes to build energy and national security.

High Energy Density Science

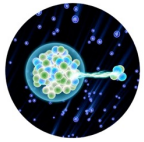
Understanding the behavior of materials at extreme temperatures and pressure.

HPC, Simulation and Data Science

Leveraging innovative computational and predictive solutions to support our mission.

Lasers and Optical Science and Technology

Advancing laser systems, optics and novel materials while working with next-generation technology.

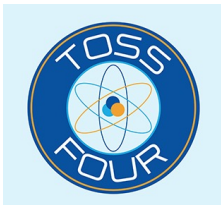


Nuclear, Chemical and Isotopic Science and Technology

Studying reaction pathways to strengthen national security and fundamental science.

Livermore Computing (LC) - Software

- Software
  - Part of ASCI PathForward effort funding Lustre creation
    - First production Lustre system: MCR in 2003
  - Slurm batch scheduler in 2002
    - And now Flux
  - ZFSonLinux
  - Tools like pdsh, conman, MPIFileUtils, IOR/mdtest, Spack and many more



lustre®

nodediag

IOR

OpenZFS

powerman

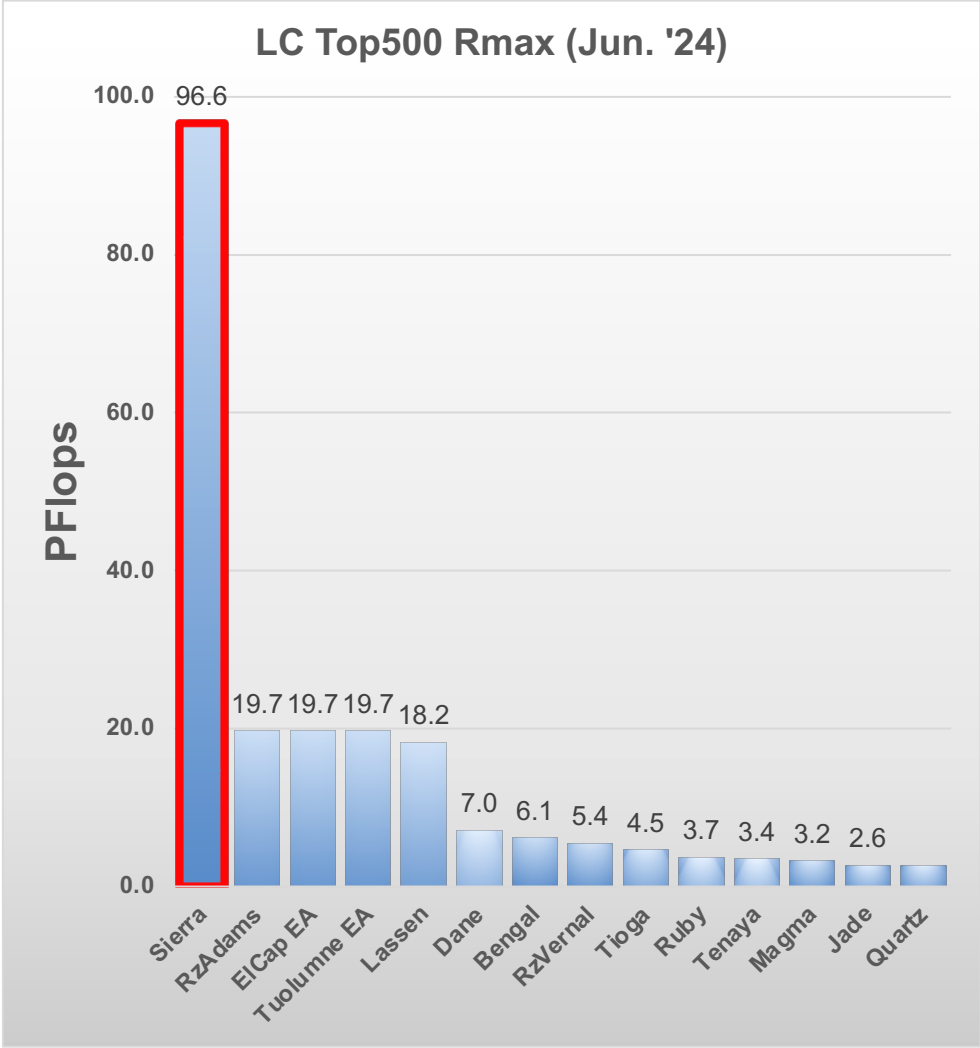
genders

pdsh

Livermore Computing (LC) - Hardware

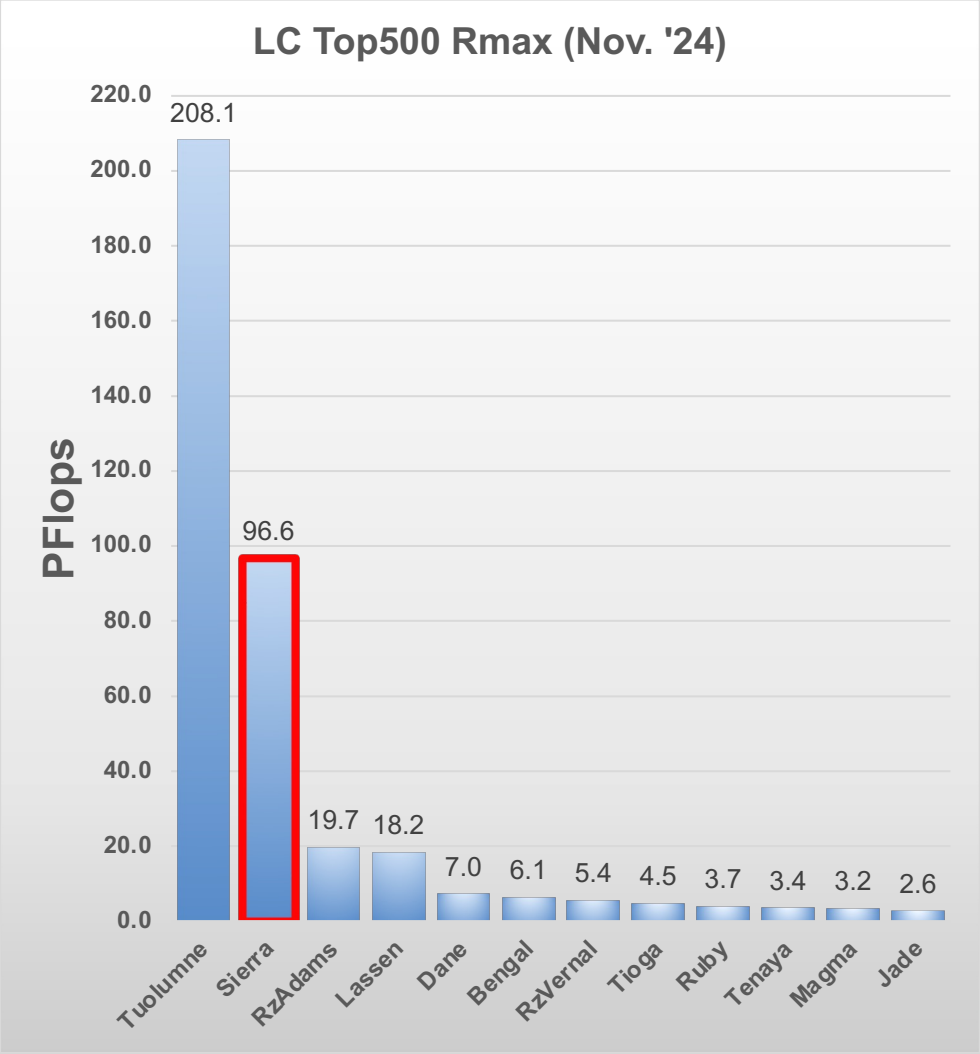
- Primary HPC Center @ LLNL
  - https://hpc.llnl.gov
- Multiple data centers and classification levels
  - Also manage HPC resources for other LLNL programs
- Hardware in LC
  - ~30 production compute clusters
    - 2 on Top10; 13 systems on Top500
    - 12 #1 Top500 entries
    - ~ 3 ExaFlops
  - Supercomputing back to 1953 with a Univac 1

June ‘24 Top500 Entries



- On June Top500, LLNL had 14
- *Sierra*
  - #2 (2018) -> #12 in June ‘24
  - Still dominated our other systems

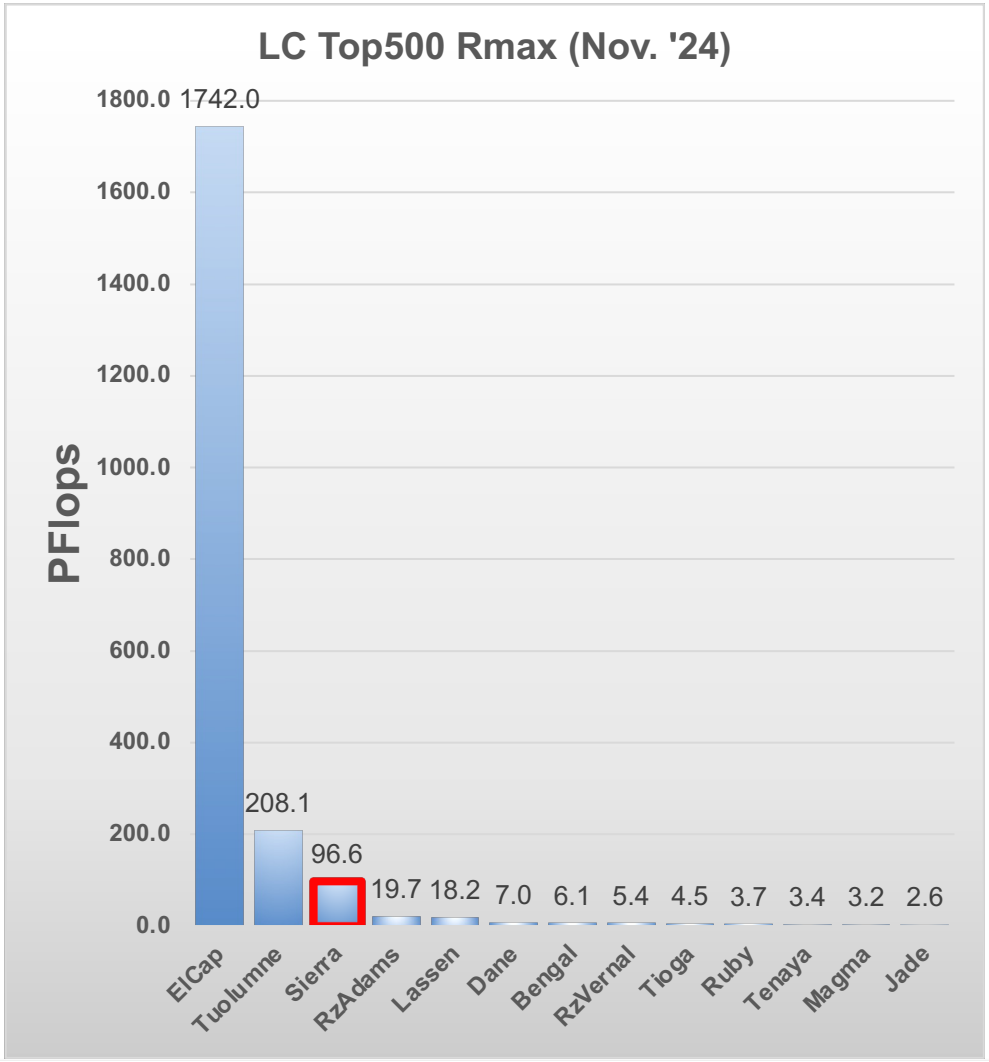
Top500 systems ++



- *Tuolumne*, a smaller sibling of *El Capitan took its place in Q3 '24*

- But was soon replaced with…

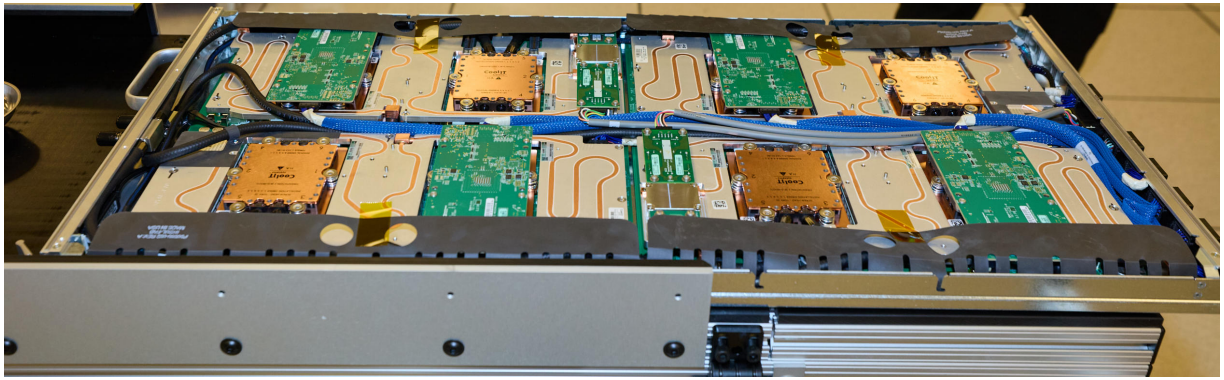
Rock-in' the Top500 in Nov. '24



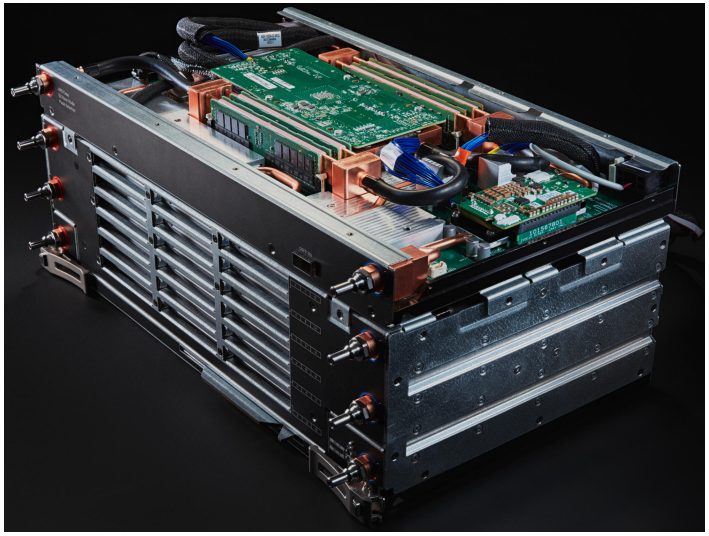
- *El Capitan*
  - *3rd US Exascale System*
  - *1st for NNSA*

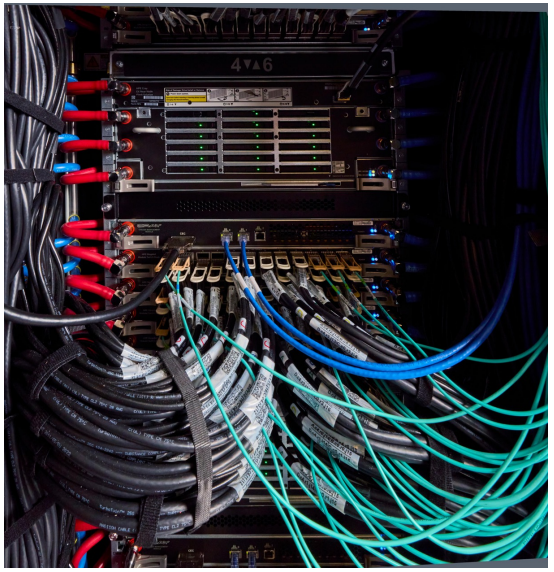
Rocks - El Capitan at a Glance

- ~7500 Ft^2
- 34.8 MW
- 11,936 Nodes
  - 11,104 compute
    - 4x AMD MI300A APUs
    - 512 GB HBM3
    - 2X Dual-port Slingshot
  - 696 Rabbits
    - 1x AMD Milan
    - ~ 30 TiB NVMe
- *CORAL 2* Contract



Rabbits – Near-node Flash





- 4U module
  - 16x 1.92TB NVMe
  - AMD 64c Milan CPU
  - PCIe to Compute chassis
    - XFS, GFS2
  - Slingshot to all Compute
    - Lustre
- Flexible, high-speed storage
  - High IOPs
  - Containerized applications
    - In-situ data analysis, reduction
  - Integrated with job scheduler
- **Q:** Will Rabbits save the PFS?

Merced

- 42 Racks
  - 2 MDS
  - 40 OSS
- 50 MDS
  - 1 MDT/MDS
  - 2x Slingshot
  - Raidz3
  - 2x Slingshot
- 240 OSS
  - 1 OST/OSS
  - 2x `draid2:11d:53c:1s`
- ~360 PiB
- Slingshot 11
- 10 LNet routers (ElCap nodes)
- HPE E1000 H/W Platform – but running TOSS

Snakes – ASP / Commodity Lustre

https://www.mdia.org/articles/alameda-whipsnake

https://en.wikipedia.org/wiki/File:ViperaAspis_1469AE.jpg

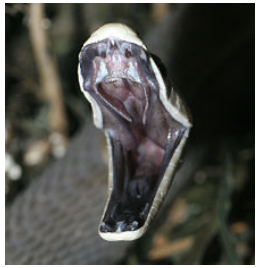
https://abcnews4.com/resources/media/fc1e865a-2163-4c1c-8be6-c18d8588cd5b-medium16x9_FilephotoofaboaconstrictorThinkstock.jpg

- ASP Contract
  - Modular COTS storage systems
    - Flash, HDD, Infrastructure modules
  - 5-yr contract with SMC (‘20-’25)
  - 32+ systems purchased since 2020
  - Consolidate knowledge, inventory, support, procurement
    - Cross-team collaboration

https://www.mdia.org/articles/northern-pacific-rattlesnake

https://en.wikipedia.org/wiki/Bullsnake#/media/File:Pituophis_catenifer_sayi_007.jpg

https://www.animalspot.net/wp-content/uploads/2014/07/Black-Racer-Snake-300x164.jpg



https://en.wikipedia.org/wiki/Black_mamba#/media/File:Dendroaspis_polylepis_striking.JPG

https://inaturalist-open-data.s3.amazonaws.com/photos/607029/large.jpg?1544727453

http://utahherps.info/pics/thamnophis_e_vagrans_bjsl_041206_172.jpg

ZFS @ LC

- ZFS 2.2.x
  - Server pair + NVMe (+ 2 SAS JBODs)
    - NVMe MDTs
    - HDD + NVMe dRAID OSTs (since 2020)
    - Pacemaker-managed
  - One Pool/OST
    - One OST/OSS
    - One MDT/MDS
  - ZED
    - Detect/Fault sick drives
    - Auto-rebuild drives

  - ASP Config
    - MDT: 3x2-drive NVMe mirrors
    - OST: `draid2:8d:90c:2s`
      - 1 JBOD = 1 OST
      - 3x2-drive "Special" devices
        - `special_small_blocks=16K`
        - Use `` `zpool list -v` `` for usage
  - El Cap / Tuolumne Config
    - MDT: zraid2
      - NVMe has low failure rate
      - Didn't want to give up capacity for DoM
    - OST: 2x `draid2:11d:53c:1s`
      - 2 draids to reduce chance of parity loss
      - OST split across 2 JBODs (40% write perf)
      - 2-drive "Special" device for MD only

Lustre @ LC

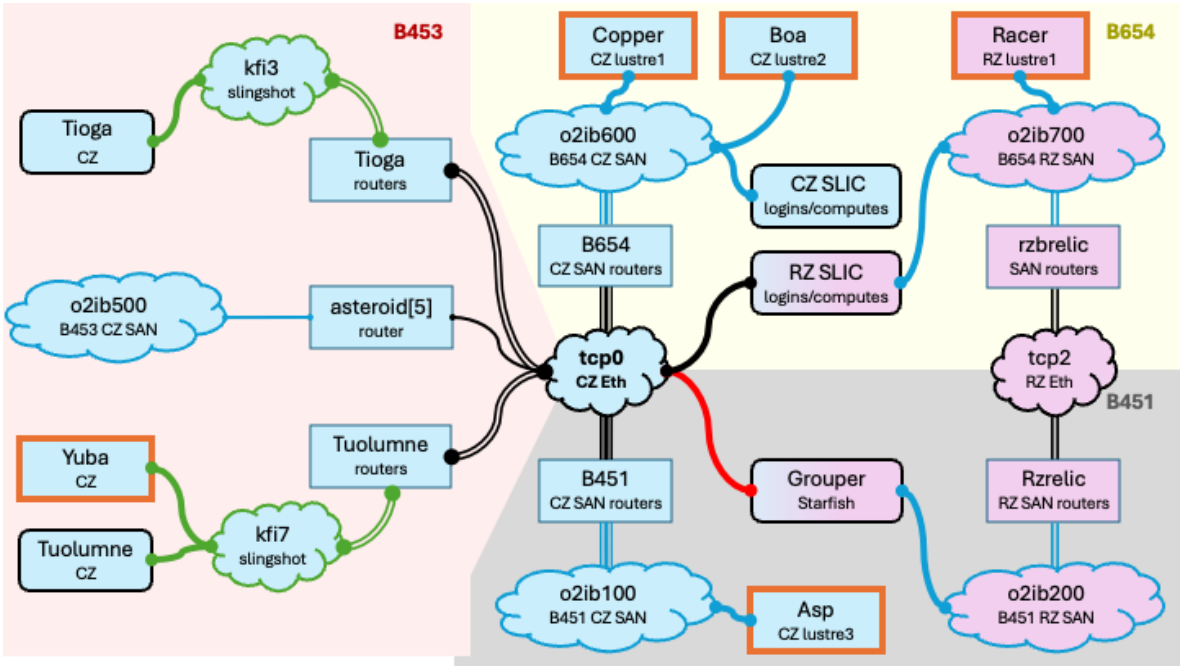
- LC Lustre
  - >500 PiB (usable)
  - **9 production file systems**
    - 4 networks
    - Most mounted Center-wide
      - Across multiple buildings
  - Not backed up, but not “scratch”
  - User quotas
    - No purging
    - User data persistent on non-CORAL Lustre
      - Must be migrated to new HW

	Commodity	Tuolumne	El Capitan
Tier 1	20TB/1M	50TB/5M	100TB/10M
Tier 2	75TB/25M	100TB/50M	500TB/100M
Tier 3	Custom Justification		

Raw Lustre Stats

filesystem	Used Space in TB	Percent Full	Millions of files	Average File Size in KB
/p/czlustre1	9996	59%	1385	7752
/p/czlustre2	9710	37%	1314	7935
/p/czlustre3	1992	23%	690	3100
/p/czlustre4	345446	86%	105919	3502
/p/czlustre5	2094	5%	6	399977
/p/lustre1	8171	37%	1902	4614

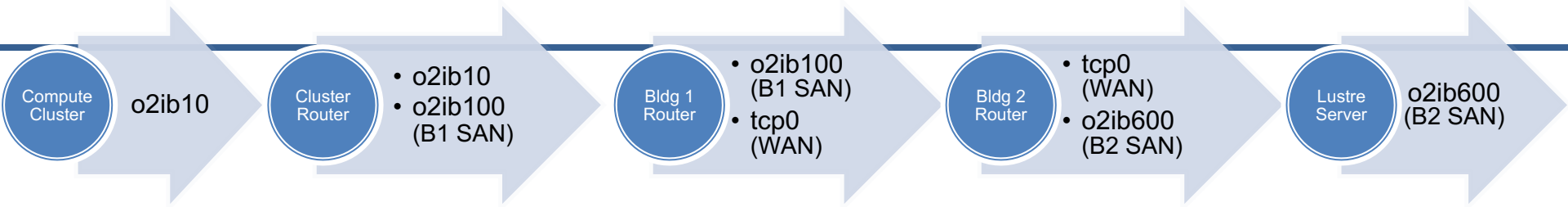
Routing in LC



Graphic courtesy of Jason Kim @ LLNL

- LNet Routing Facts
  - 4 Fabrics
    - Slingshot, OPA, IB, TCP
  - 3 LNet types
    - KFI, O2IB, TCP
  - 3 Buildings
  - 61 Lustre Networks
    - 8 SAN
    - 53 Cluster
  - Fine-grained Routing
  - Discovery off

Routing Configs




```
options lnet networks="o2ib19      (hsi0)  "

options lnet routes="o2ib100 1 192.168.58.[ 47-54]@o2ib19;
 tcp0    192.168.58.[ 47-54]@o2ib19; \
 o2ib500 192.168.58.[ 47-54]@o2ib19; \
 o2ib600 192.168.58.[ 47-54]@o2ib19"
```

Which networks can I talk to?

Talk to local cluster routers

```
options lnet forwarding="enabled"

            options lnet networks="o2ib19      (hsi0)  ,o2ib100      (san0)  "

options lnet routes="tcp0    1 172.6.2.[2-5]@o2ib100; \
 o2ib500 172.6.2.[2-5]@o2ib100; \
 o2ib600 172.6.2.[2-5]@o2ib100"
```

```
options lnet forwarding="enabled"

            options lnet networks="tcp0      (aci.2364)  ,o2ib100      (san0)  "

options lnet routes="o2ib16  1 172.6.1.[50-51]@o2ib100; \
 o2ib31  1 172.6.1.99@o2ib100; \
 o2ib36  1 172.6.2.[26-28]@o2ib100; \
 o2ib38  1 172.6.2.[37-38]@o2ib100; \
 o2ib19  1 172.6.2.[39-46]@o2ib100; \
 o2ib44  1 172.6.1.[106-108]@o2ib100; \
 o2ib45  1 172.6.1.[91-92]@o2ib100; \
 o2ib56  1 172.6.1.49@o2ib100; \
 o2ib58  1 172.6.2.[35-36]@o2ib100; \
 o2ib500 1 172.4.94.10@tcp; \
 o2ib600 1 172.4.70.[12-15]@tcp0; \
 o2ib28  172.4.70.[12-15]@tcp0"
```

```
options lnet forwarding="enabled"

            options lnet networks="tcp0      (aci.2315)  ,o2ib600      (san0)  "

options lnet routes="o2ib28  1 172.6.3.56@o2ib600; \
 o2ib100 1 172.4.66.[2-5]@tcp0; \
 kfi3  1 172.4.94.[18-19]@tcp0; \
 kfi7  1 172.4.94.[20-21]@tcp0; \
 o2ib16  172.4.66.[2-5]@tcp0; \
 o2ib20  172.4.66.[2-5]@tcp0; \
 o2ib31  172.4.66.[2-5]@tcp0; \
 o2ib34  172.4.66.[2-5]@tcp0; \
 o2ib36  172.4.66.[2-5]@tcp0; \
 o2ib38  172.4.66.[2-5]@tcp0; \
 o2ib19  172.4.66.[2-5]@tcp0; \
 o2ib44  172.4.66.[2-5]@tcp0; \
 o2ib45  172.4.66.[2-5]@tcp0; \
 o2ib56  172.4.66.[2-5]@tcp0; \
 o2ib58  172.4.66.[2-5]@tcp0; \
 o2ib60  172.6.4.[39-42]@o2ib600"
```

```
options lnet networks="o2ib600      (san0)  "

options lnet routes="tcp0    1 172.6.3.[12-15]@o2ib600; \
 o2ib28  1 172.6.3.56@o2ib600; \
 o2ib37  1 172.6.3.[59-60]@o2ib600; \
 o2ib57  1 172.6.3.[20-51]@o2ib600; \
 o2ib60  1 172.6.4.[39-42]@o2ib600; \
 kfi3  172.6.3.[12-15]@o2ib600; \
 o2ib16  172.6.3.[12-15]@o2ib600; \
 o2ib21  172.6.3.[12-15]@o2ib600; \
 o2ib31  172.6.3.[12-15]@o2ib600; \
 o2ib34  172.6.3.[12-15]@o2ib600; \
 o2ib36  172.6.3.[12-15]@o2ib600; \
 o2ib38  172.6.3.[12-15]@o2ib600; \
 o2ib19  172.6.3.[12-15]@o2ib600; \
 o2ib45  172.6.3.[12-15]@o2ib600; \
 o2ib56  172.6.3.[12-15]@o2ib600; \
 o2ib58  172.6.3.[12-15]@o2ib600; \
 o2ib100 172.6.3.[12-15]@o2ib600"
```

Challenges (General)

- Router tuning
  - Frequent hangs due to misaligned buffers and credits
  - Initially hard to find LNet tuning documentation
    - https://wiki.whamcloud.com/display/LNet/LNet+Routing+Setup+Verification+and+Tuning
  - Huge improvement in stability after calculating proper values

- LU-14555
- LU-16106
- LU-16244
- LU-17440

- Migration
  - `lustre_fssync` tool based on `dsync` from MPI File Utils
    - Bug with `--delete` would remove already synced files on destination
      - Caused months-long delays
      - Use MFU 0.12+ !
    - New SAS card changed enumeration of vdevs unknowingly
    - Bug in our script forced DIO, resulting in large file syncs to choke

Challenges (ElCap/Merced)

- Non-CORAL2 clusters use central Pacemaker + pacemaker_remote on nodes
  - Single point of control and view
  - Too serial to scale to 290 servers!
  - Adopted “pair-wise” configuration w/ all nodes running Pacemaker
    - Much faster
    - Still miss niceties of centralized management
- Slingshot
  - New interconnects always have growing pains!
  - Lack of common tuning and debugging knowledge
  - Frequent fabric-wide interruptions
  - Couldn’t do striping across all OSTs
  - Occasional checksum errors
    - `lctl get_param osc.*.checksums` (Enabled by default)

Thank you!

Lawrence Livermore
National Laboratory