Un-scratching Lustre

MSST 2019

Cameron Harr
(Lustre Ops & Stuff, LLNL)

May 21, 2019

LLNL-PRES-773414
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

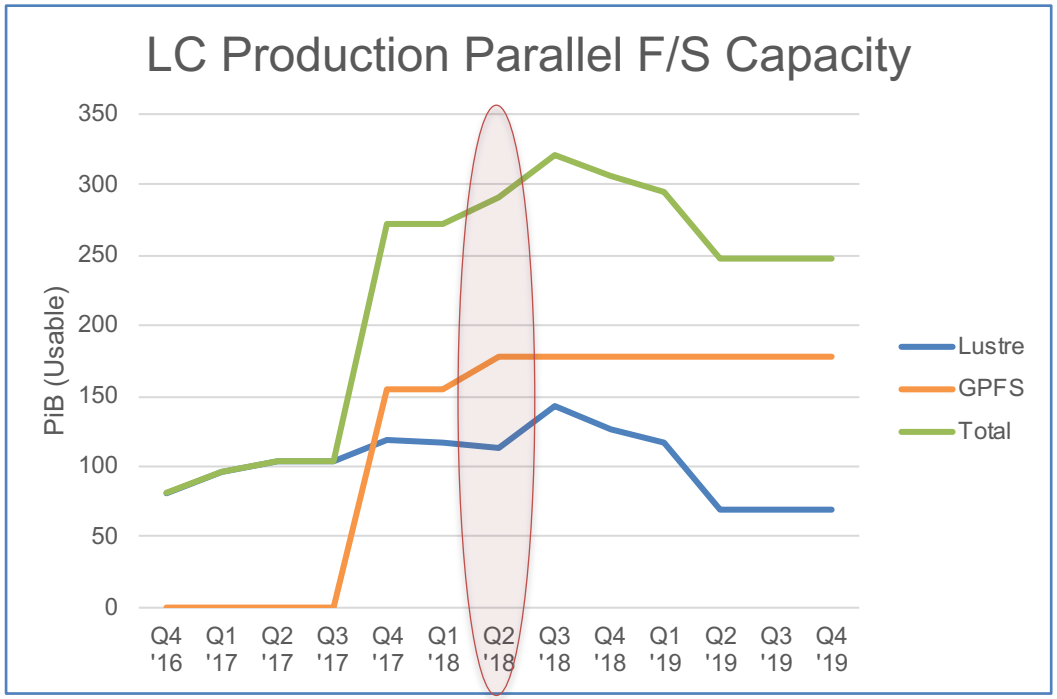
- US DoE / NNSA
  - Missions:
    - Biosecurity
    - Defense
    - Intelligence
    - Science
    - Counterterrorism
    - Energy
    - Nonproliferation
    - Weapons

Livermore Computing (LC)

- Compute
  - Classified: ~151 PF
    - Sierra: 126 PF_{pk}, #2
    - Sequoia: 20 PF_{pk}, #10
  - Unclassified: ~30 PF_{pk}
    - Lassen: 19 PF_{pk}, #11
- 4+ Data centers
  - TSF: 45MW -> 85MW
- 3 *Centers: CZ, RZ, SCF*

Parallel FS @ LC (2018)

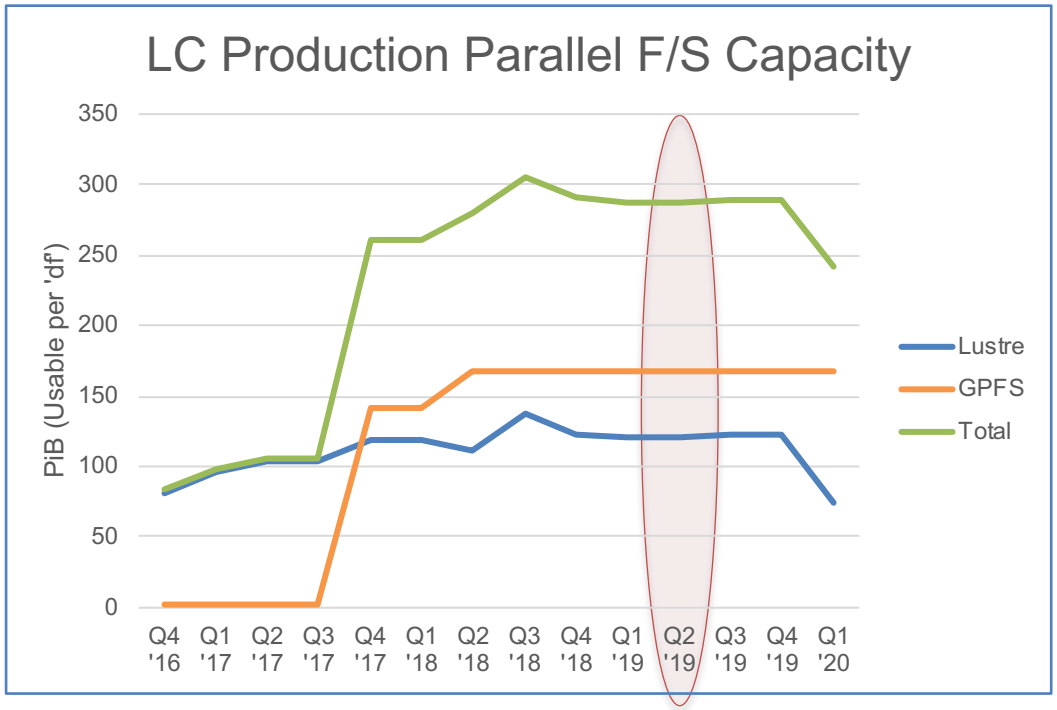
- Production Lustre
  - 13 production file systems
  - >118 PiB (useable)
  - ~15B files
- Multi-generation
  - Lustre 2.5 (NetApp/Cray)
    - 1 MDS
    - ZFS 0.6
  - Lustre 2.8 (RAID Inc.)
    - JBODs
    - 4-16 MDS
      - DNE v1
    - ZFS 0.7



filesystem	Used Space in TB	Percent Full	Millions of files	Average File Size in KB
/p/lscratchd	4294	78%	1035	4456
/p/lscratche	3759	69%	1130	3573
/p/lscratchf	1679	77%	809	2229
/p/lscratchh	6807	41%	3363	2173
/p/lscratchrza	5734	69%	1207	5099
/p/lscratchrzb	461	42%	361	1371
/p/lscratchv	3659	56%	1638	2399

Parallel FS @ LC (2019)

- Production Lustre
  - 8 production f/s
    - 13 - 8 + 3
  - ~120 PiB (useable)
- Multi-generation
  - 3x NetApp
    - 2x 2.5
    - 1x 2.10
  - 5x RAID Inc.
    - 3x 2.10
    - 2x 2.8
  - 2.8/2.10 clients

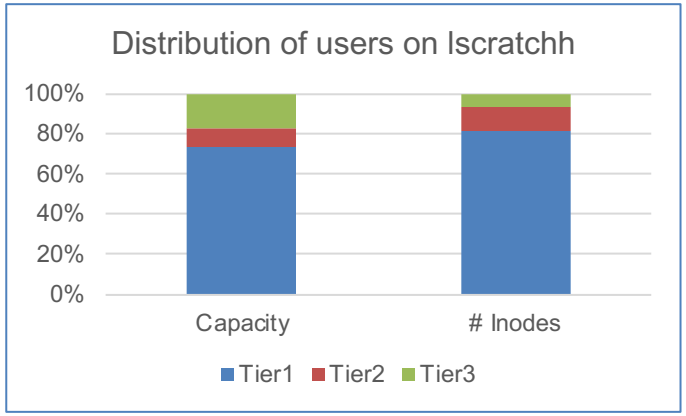


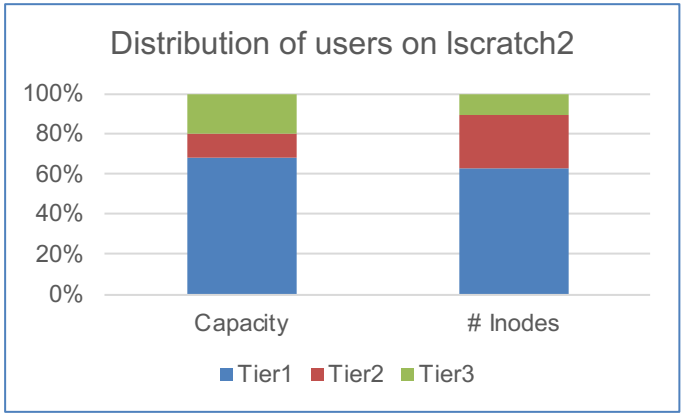
filesystem	Used Space in TB	Percent Full	Millions of files	Average File Size in KB
/p/czlustre1	2739	17%	310	9492
/p/czlustre2	4058	25%	733	5943
/p/czlustre3	1417	26%	546	2788
/p/lustre1	1955	24%	322	6516

Lustre Scratch Purge Policy (2018)

- Official policy: files > 60 days can be purged
  - Bad for users as losing one file can destroy a large dataset
  - Small users and early-alphabet users purged disproportionately
- Effective policy: purge @ ~80% after cleanup
  - Target top-10 users (files or capacity)
  - Ask users to clean up, then use *lpurge* as last resort on select users
  - Pros
    - Saves small users from suffering from the actions of power users
    - Enables greater utilization of f/s
  - Cons
    - Still requires overhead/time from admins and LC Hotline
    - Delays from users can cause uncomfortable levels of usage
    - Users don't clean up unless forced to

Lustre Quota Policy (2019)





Quota Tier	Capacity (TB)		# Files		Grace Period (days)
	Soft	Hard	Soft	Hard	
1	18	20	900K	1M	10
2	45	50	9M	10M	10
3	Levels set per justification				10

- Per-file system
- Tier 3:
  - Custom # inodes, TB
  - Max duration: 6 months

Auto-delete

- AutoDelete directories
  - Users would `rm –rf <dir>`
    - And wait
    - … and wait
    - … and wait
  - Now they can `mv <dir> …`and get on with life
  - drm job, as <user>, removes the files quickly
  - https://github.com/hpc/mpifileutils

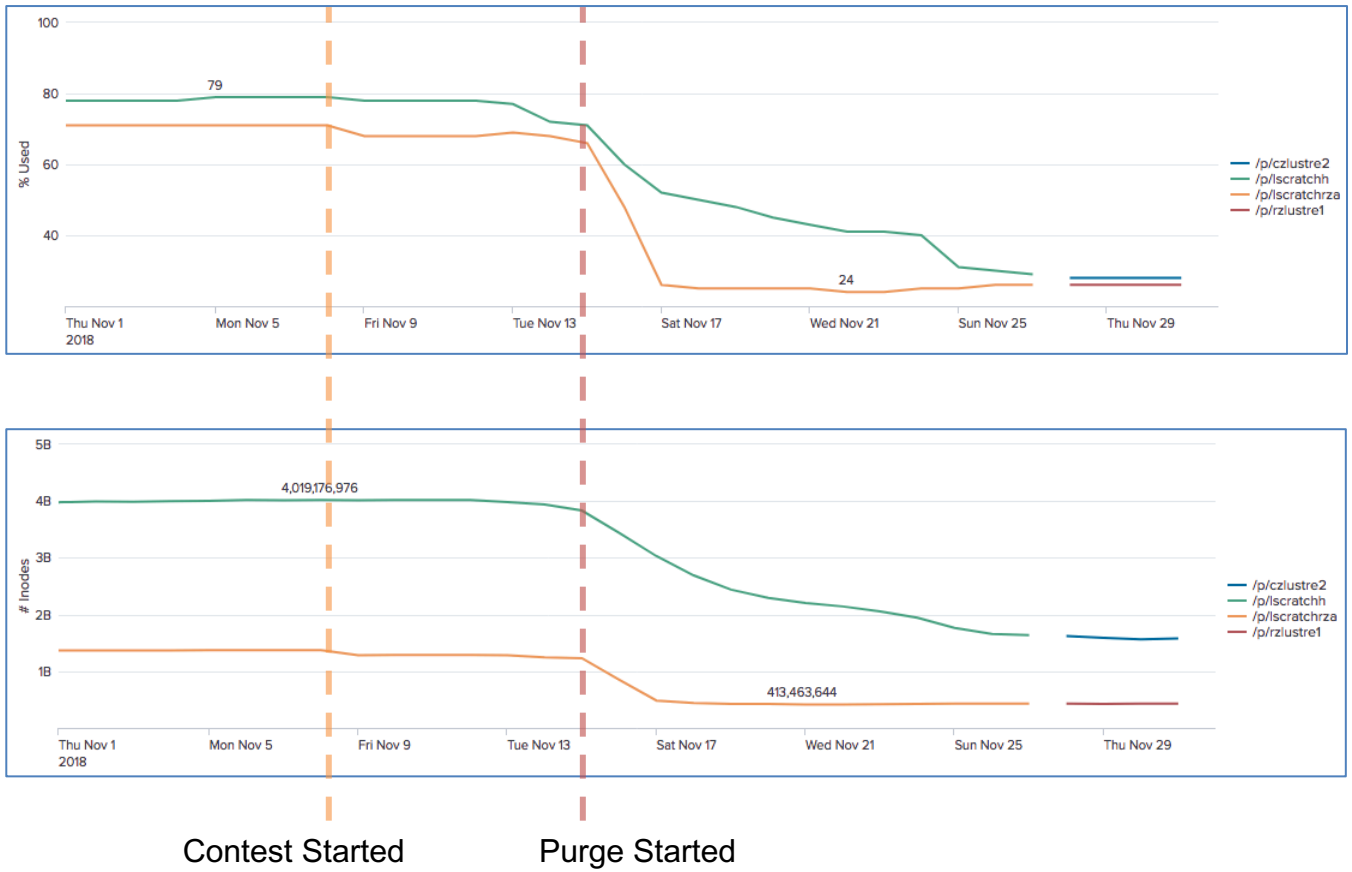
How We Did It

- Stand up new file systems with new policy
- Incentivize clean-up on existing file systems
  - Gift card
  - Exemptions
- One-and-done big purge

https://www.bulldozer.in/images/solid_waste_%20machine/sd7n_solid_waste_blade_dozer.jpg

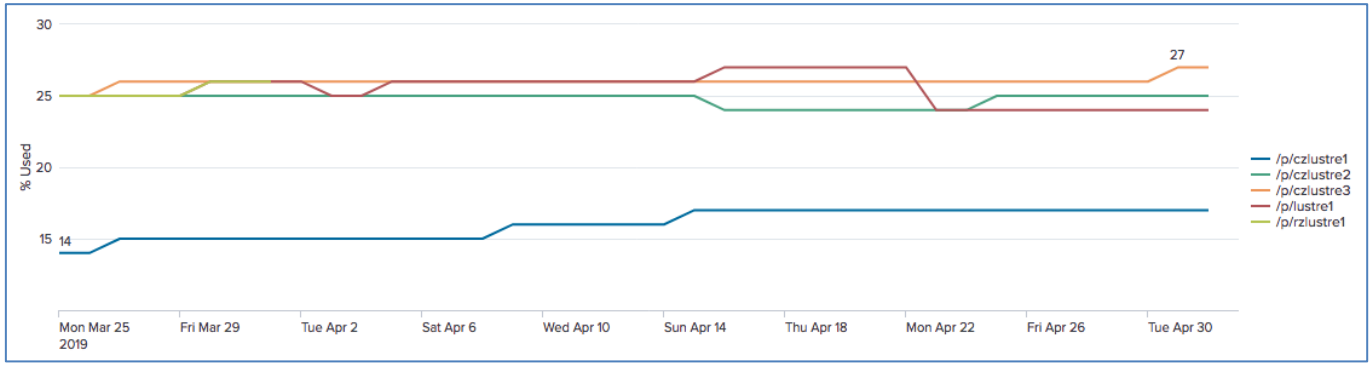
The Purge

- Before Cleanup
  - Capacity:
    - 79% full
    - 13.2 PB
  - Inodes:
    - 4 Bi

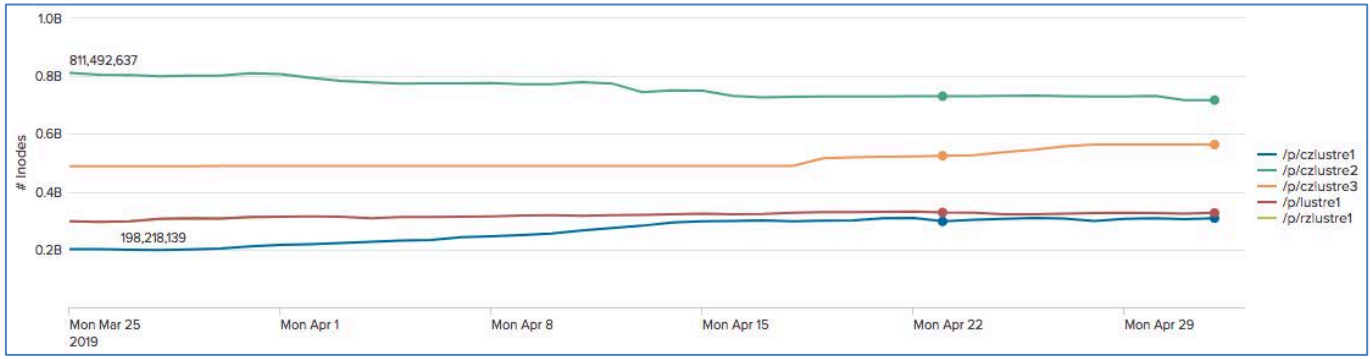


Long-term Results

- Current utilization
  - Capacity:
    - < 30% full

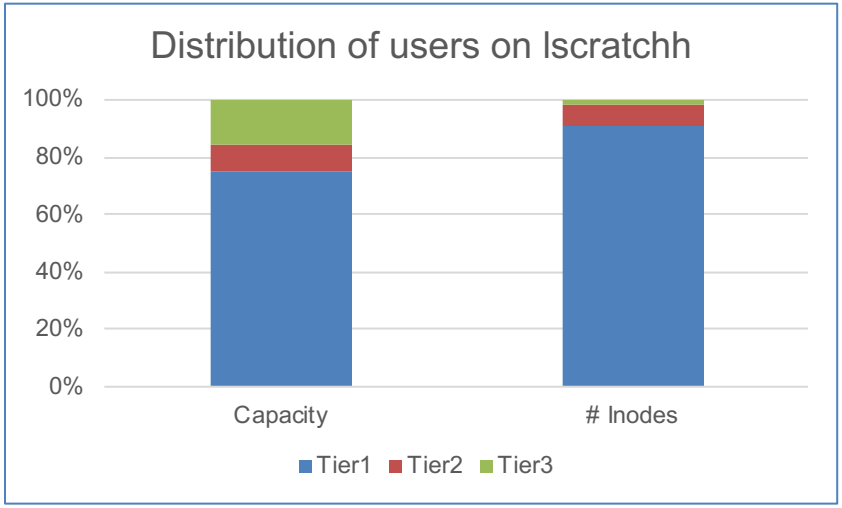


  - Inodes:
    - < 1B files



Long-term Results (cont.)

- Current status
  - Tier 3 allocations (aggregate):
    - 65 users on CZ/RZ
    - 21 users on SCF



- Lessons learned
  - More increases requested than anticipated
    - Enabled LC Hotline to effect the changes
    - Inodes more in demand than expected
      - Bumped Tier 1 to 1M from 500K files
  - Created system to track/check/set/remove Tier 3 allocations

Users' Thoughts

- Current status (cont.)
  - Users mostly pleased with the change
    - Only one user vocally unhappy
    - Paraphrased user responses (per user coordinator):
      - WHAT?!? My files aren't going to disappear?!? That is wonderful! Why didn't I hear about this?
      - 20TB is toooo small for me. Why can't I get more? I can get more?!? You're the best!
      - Ugh. Now I have to figure out what to delete? Why can't LC do that for me based on these rules <insertruleshere>? But they better never delete file X - that is the exception to those rules. Oh. I see now what you mean. That autodelete directory is super nice!
      - Wait. I know you said my files weren't going to disappear, but did you really mean it? I figured that once the system got to a certain point, they would.
      - I reallllllyyyyy like that my files aren't going to disappear.
      - THANK you for emailing me that I am reaching my quota. I wish that it came <more/less> often.
      - While I hate having to clean up after myself, it is WONDERFUL that I am not going to lose any files.
  - Lustre soft quota grace period expiration isn't liked
    - "Why can't I use all my allocated storage?"
    - Would like to set infinite grace period

Thank you!