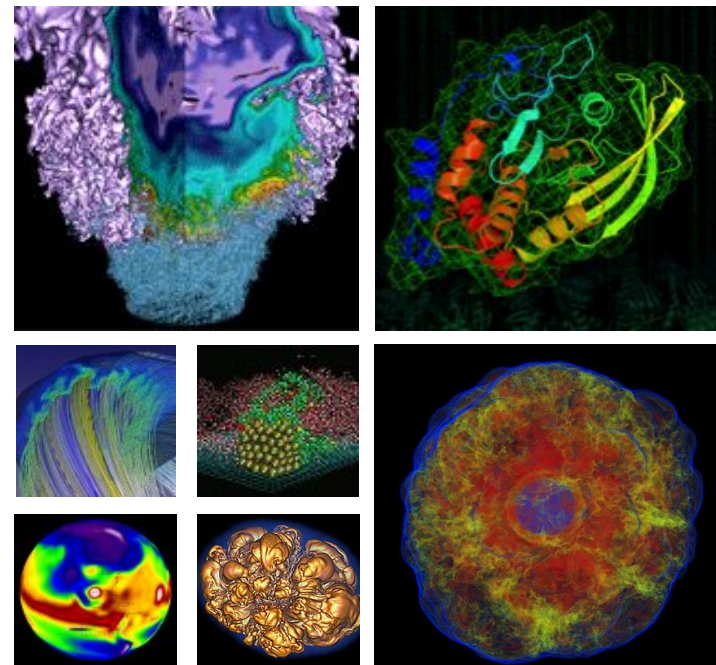


NUG Monthly Meeting



15 October, 2020



U.S. DEPARTMENT OF
ENERGY

Office of
Science



Today's plan



- Interactive - please participate!
 - Raise hand or just speak up
 - **NERSC User Slack** (link in chat), **#webinars** channel
- Agenda:
 - Win-of-the-month
 - Today-I-learned
 - Announcements/CFPs
 - Topic of the day: The cscratch1 crash
 - Coming meetings: topic suggestions/requests?
 - Last month's numbers

Win of the month



Show off an achievement, or shout out someone else's achievement, e.g.:

- Had a paper accepted
- Solved a bug
- A scientific achievement (maybe candidate for Science highlight, or High Impact Scientific Achievement award)
- An Innovative Use of High Performance Computing (also a candidate for an award) (<https://www.nersc.gov/science/nersc-hpc-achievement-awards/>)

Tell us what you did, and what was the key insight?

Today I learned



What surprised you that might benefit other users to hear about?
(and might help NERSC identify documentation improvements!)

Eg:

- Something you got stuck on, hit a dead end, or turned out to be wrong about
 - Give others the benefit of your experience!
 - Opportunity to improve NERSC documentation
- A tip for using NERSC
- Something you learned that might benefit other NERSC users

"If we knew what it was we were doing, it would not be called research, would it?" - Einstein

Announcements and CFPs



Check latest weekly email for these:

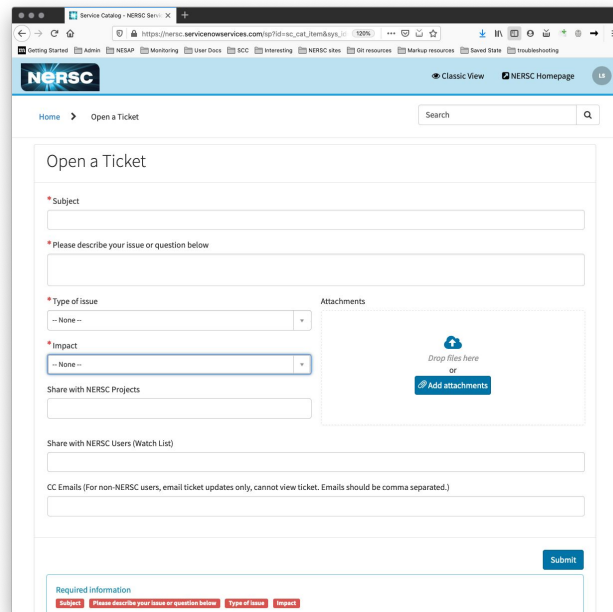
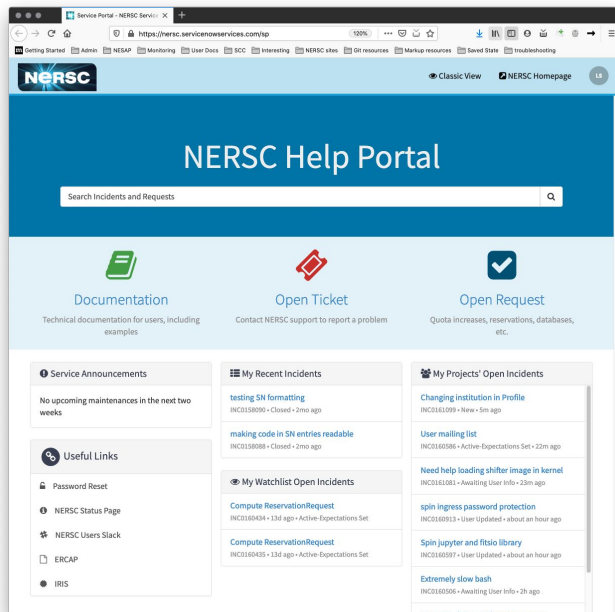
- Cori scheduled maintenance for October is cancelled
- Parallelware training
- First International Symposium on Checkpointing for Supercomputing (SuperCheck21)
- SC21 Call for Planning Committee Volunteers Is Now Open!

New: Job queue will be paused, and cscratch1 unavailable, for a few hours on Monday morning while we add an ADU (more soon) to help with debugging last week's crash - announcement message soon

New Help Portal



<https://nersc.servicenowservices.com/sp>



Announcements and CFPs



Check latest weekly email for these:

- Cori scheduled maintenance for October is cancelled
- Parallelware training
- First International Symposium on Checkpointing for Supercomputing (SuperCheck21)
- SC21 Call for Planning Committee Volunteers Is Now Open!
-

Others?



cscratch1	
Lustre Metadata (filenames, directories, access times, striping etc)	Actual file data. Can be single-striped (whole file on one OST) or striped over many OSTs (parallel I/O -> aggregate bandwidth)

The diagram illustrates the relationship between different components in a Linux filesystem. It shows a large icon labeled 'MDS' (Master Data Service) and a smaller icon labeled 'MDT' (Master Data Table). An arrow points from the 'MDS' icon to the 'MDT' icon. Another arrow points from the 'MDT' icon to a label 'Linux filesystem'. To the right, there is another set of icons labeled 'ADU' (Advanced Data Unit) and 'MDT' (Master Data Table).

The diagram illustrates the mapping of files to OSTs in a distributed storage system. At the top, four blue vertical rectangles represent files with `stripe_count=4`, and one purple horizontal rectangle represents a file with `stripe_count=1`. Below these, a row of storage nodes is shown, each consisting of an OSS (Object Storage System) icon and an OST (Object Storage Target) icon. Arrows indicate the mapping: the four blue files are mapped to the first four OSTs, and the purple file is mapped to the fifth OST. The nodes are labeled with an ellipsis (...) between the fourth and fifth, indicating a larger cluster. A green arrow points from the text 'Linux filesystem' to the first OST, and another green arrow points from the text 'Linux filesystem' to the fifth OST.

Actual file data was not involved (so no corruption). But if the directory entry was damaged, the file might be missing (moved to a "lost+found")

So what happened?



September						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			
October						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10

- **Thursday 24 Sept:** MDS crashes, its (local) file system journal was damaged, cscratch1 offline for check-and-repair (Cori effectively down) until Friday evening
- **Sunday 27 Sept:** Repeat of same crash. Probably not a coincidence, NERSC+HPE working to understand the cause
 - Insufficient information in logs, suspicion that a race condition or Linux kernel bug is involved.
- **Wednesday 30 Sept:** Exceptional measures:
 - MDS restarted with instrumented kernel to capture more data in the event of a crash.
 - Cori put into "special debug mode" with measures to isolate any damage from a future crash, and allow Cori to operate mostly independently of cscratch1
 - NERSC users assist efforts to reproduce crash
- **Friday 2 Oct:** Crash reproduced.
 - NERSC+HPE engineers analysing debug traces
 - Cori continues to operate in debug mode, without cscratch1.

So what happened?



September						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			
October						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10

- **Sunday 4 Oct:** Still debug mode, cscratch1 returned with more instrumentation and a guard that we hoped would catch the error and trigger a failover instead of a crash. A synthetic workload was introduced to reproduce the conditions that we think triggered the error. Error reproduced quickly, but guard fails to prevent crash. More data gathered.
- **Monday 5 Oct:** Cori briefly removed from service, verify that synthetic workload alone reproduces crash. Cori returned in debug mode with no cscratch1, while check-and-repair runs on MDS.
- **Thursday 8 Oct:** Cori rebooted to into normal production configuration, mitigations to prevent conditions triggering the crash. (Eg: users requested to limit striping). Verify that heavy load alone **does not** cause crash before return to service.



What did we find?



- **No root cause yet**
 - May be a race condition (ie timing-dependent)
 - No "smoking gun" found in successive layers of Lustre, might be Linux kernel bug
- Memory dumps of crashes reveal an invalid value of a specific pointer in a very low-level IO data structure
 - Pages modified in those IO structures revealed a particular application as being correlated to the failure
- Application correlated to the failure was:
 - performing unlink (delete) operations
 - in parallel
 - files were striped over many (synthetic workload used all) OSTs (by accident)

(cont'd)



What did we find?



(cont'd)

- **The specific mechanism** of how this workload is causing the corrupted pointer **is unknown**. Given the immediate reproducibility in our testing, we suspect that the bug is somehow tied to **using many many OSTs**.
Speculation of reasons:
 - When stripe count is high (>160), extended attribute inode blocks must be used (possibly these are related)
 - There is increased complexity and messaging when deleting files, especially in parallel
- **Therefore:** we think using **stripe_large** is safe (sets stripe count to 72)
 - **BUT:** please don't stripe more than this



- Avoiding high stripe counts
 - General request to users
 - File system scan to detect high stripe counts, directly contact users
- A key workflow that exhibits the trigger conditions has been paused:
 - Modify workflow to avoid trigger conditions
 - Move workflow directory from main MDS to an ADU (minimize impact of any further crash)
- **Important: No user workflow *causes* or is at fault for the crash** - error is deeper in system
 - But avoiding conditions under which the crash occurs helps everyone

After-effects: Hanging Login Nodes



Login nodes observed "hanging" while accessing cscratch

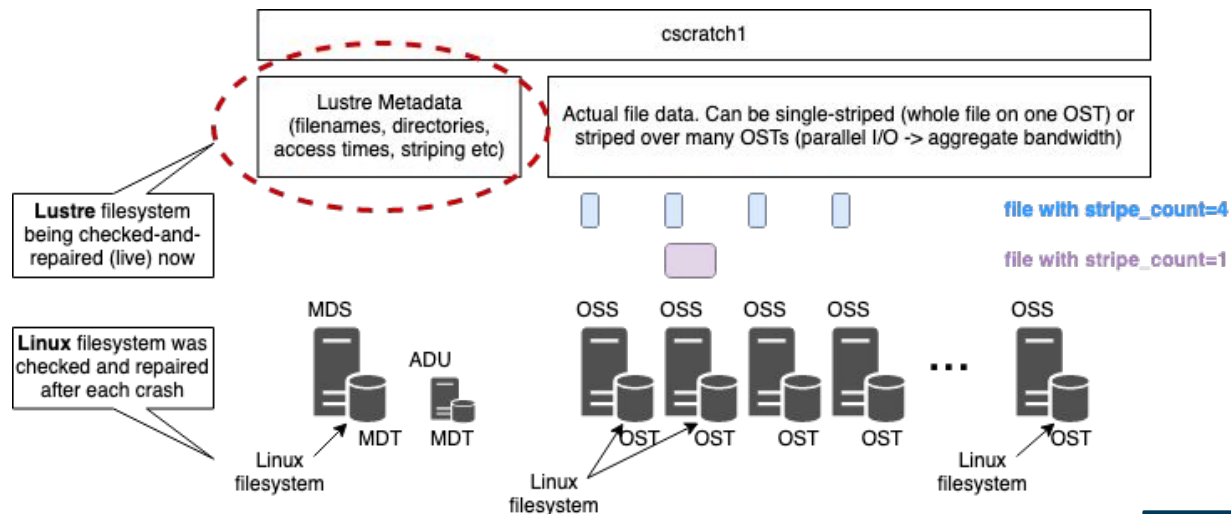
- Lustre client handles limited number of simultaneous metadata operations (RPCs in flight)
- Hang of an RPC or resource is possibly related to attempts to modify files with damaged metadata
- When enough such requests have hung, Lustre client freezes

Mitigation:

- Monitoring number of RPCs in flight, reboot login node at threshold

Fix (should complete soon):

- Check-and-repair of Lustre



After-effects: A few files w/corrupted metadata



- Metadata for small number of files may be corrupted
 - Manifests as files with “?????” for attributes when doing “ls -l”
 - Files can’t be deleted or moved
- NERSC is working with HPE to repair / recover these files
 - In the meantime, if they’re in your way, you can avoid them by renaming the directory they’re in
 - e.g. ``mv my_dir/file_w_md_issue.txt hold_my_dir/file_w_md_issue.txt``
- We are currently scanning the entire file system to find any undiscovered files with corrupted metadata
 - Scan will take about 2 weeks to complete (there’s a lot of data)
 - We will contact you if your data is on the list
- Feel free to contact us if you have questions (<https://help.nersc.gov>)

Debugging, and what next?



Debugging this sort of issue is challenging:

- Hard (impossible?) to reproduce at smaller scale
 - Huge file system plus heavy I/O workload to drive it
- Long debug cycle
 - Check-and-repair (fsck) process takes ~1 full day, can be more
- Complex system
 - Concurrency
 - Multiple layers of software

Multi-team effort:

- HPE engineers
- NERSC engineers
- NERSC users (thank you again!)

Use kernel instrumentation to log actions, and analyse memory dump at crash to detect conditions at time of crash (eg, engineers found a pointer with bad value, tracing it back to find root cause)

Debugging, and what next?



Debugging efforts continue

- Test systems *just* large enough in dimensions that we think matter
- Modify synthetic workload to exercise same code paths at smaller scale
- Tests with isolated ADU
 - Limit "blast radius" of crash (other cscratch1 users not affected)
 - To integrate the new ADU, we'll have a brief (< 4 hours) outage on 10/19 starting at 7AM

Update to come!

- Doug Jacobsen: Group Lead, Computational Systems
- Lisa Gerhardt: Data Analytics Services
- Alberto Chiusole: Data Analytics Services
- Steve Leak: User Engagement

Coming up



Next meeting will return to "3rd Thursday" schedule

Topic requests/suggestions?

Tip: We'd love to hear some lightning talks from NERSC users about the research you use NERSC for!

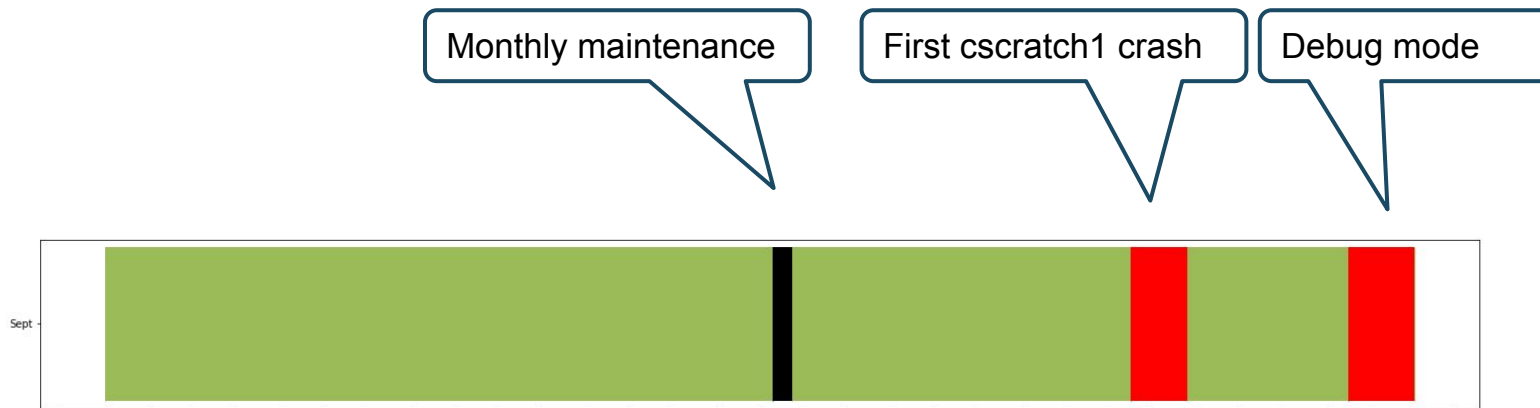
Last month's numbers - September



Scheduled and overall availability:

	Scheduled	Overall
Cori	86.5%	85.0%
HPSS	100.00%	100.0%
CFS	100.00%	100.0%

Cori:



Last month's numbers - September



Cori Utilization: **97.2%**

Large jobs: **37.1%**

New Tickets: **647**

Closed Tickets: **496**

Backlog at 1 Oct: **636**



Thank You



U.S. DEPARTMENT OF
ENERGY

Office of
Science

