Nine Months in the Life of an All Flash File System

Lisa Gerhardt*, Stephen Simms*, David Fox*, Kirill Lozinskiy*, Wahid Bhimji*, Ershaad Basheer* and Michael Moore†

*NERSC, Lawrence Berkeley National Lab

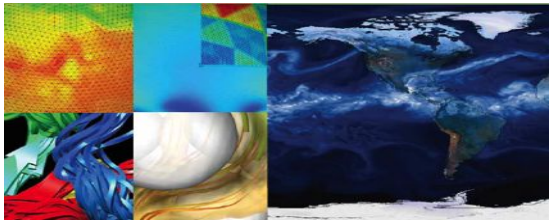
†Hewlett Packard Enterprise

May 8, 2024

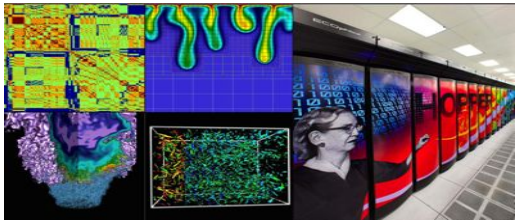
NERSC: Mission HPC for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

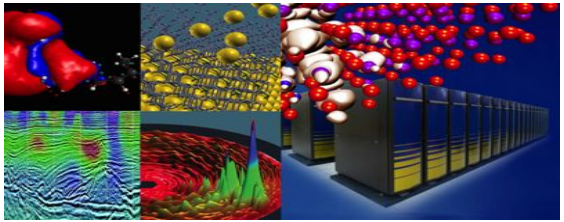
Largest funder of physical science research in the U.S.



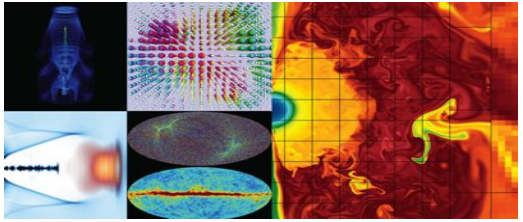
Biological and Environmental Research



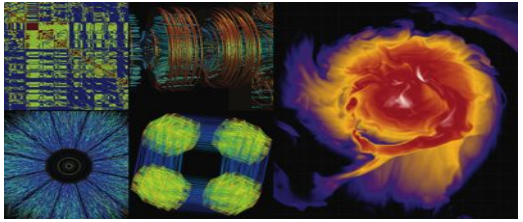
Computing



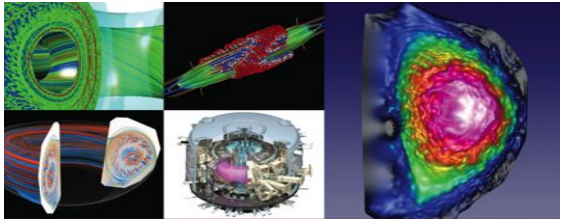
Basic Energy Sciences



High Energy Physics

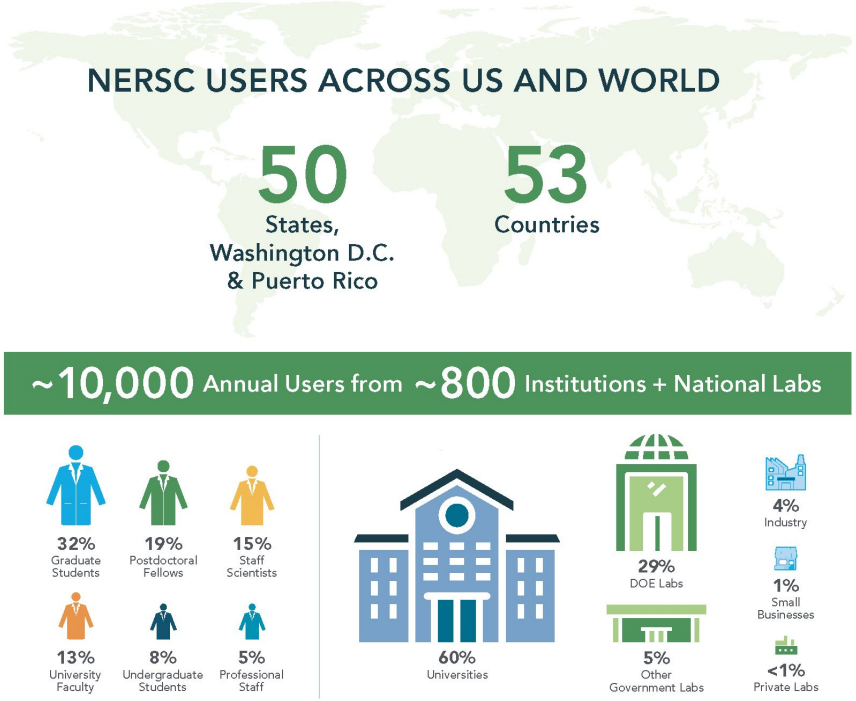


Nuclear Physics



Fusion Energy, Plasma Physics

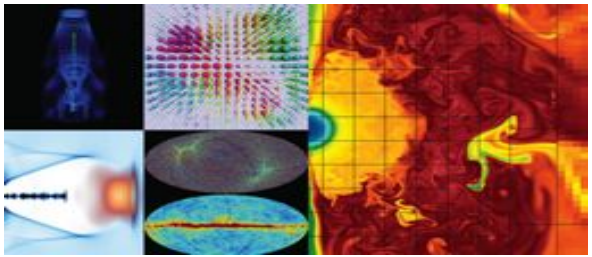
NERSC by the Numbers



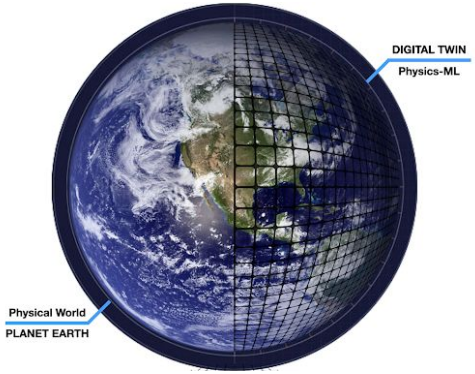
NERSC has been acknowledged in 5,829 refereed scientific publications & high profile journals since 2020

- Nature [32]
- Nature Communications [116]
- Proceedings of the National Academy of Sciences [55]
- Science [21]
- Nature family of journals [232]
- Monthly Notices of the Royal Astronomical Society [248]
- Physical Review B : Condensed Matter and Materials Physics [206]
- Physical Review D : Particles, Fields, Gravitation, and Cosmology [200]

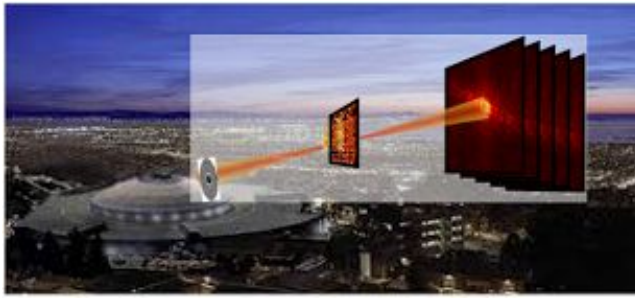
We Accelerate Scientific Discovery for Thousands of Office of Science Users with 3 Advanced Capability Thrusts



Large-scale applications for simulation, modeling and data analysis



Complex experimental and AI driven workflows



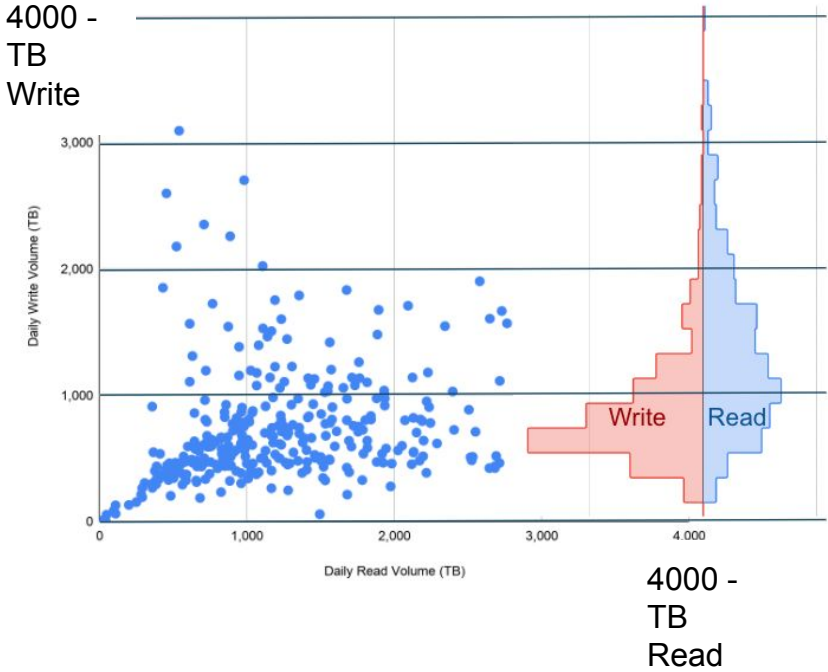
Time-sensitive and interactive computing

The NERSC workload is diverse with growing emphasis on integrated research workflows

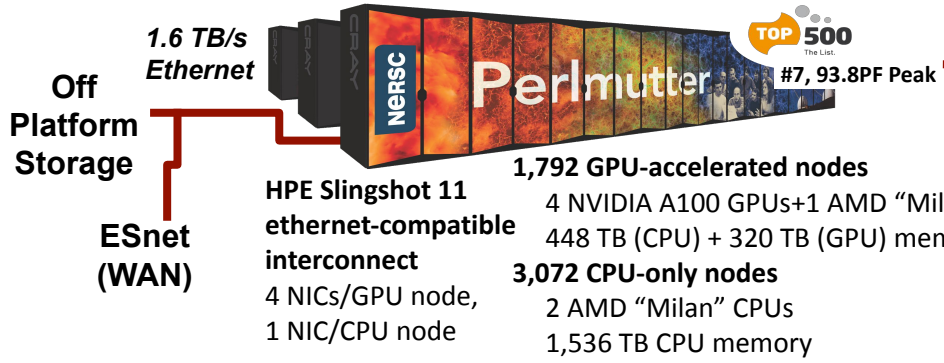
All-flash Solution for NERSC's Diverse Workload

- NERSC supports 10,000 users from 1,000 different projects
- Huge variety of I/O demands from
  - Large block checkpoint / restart
  - Small block random access across thousands of files
- Requirement for high performance and low performance variation
- All-flash can offer performance, productivity and sufficient capacity for this workload

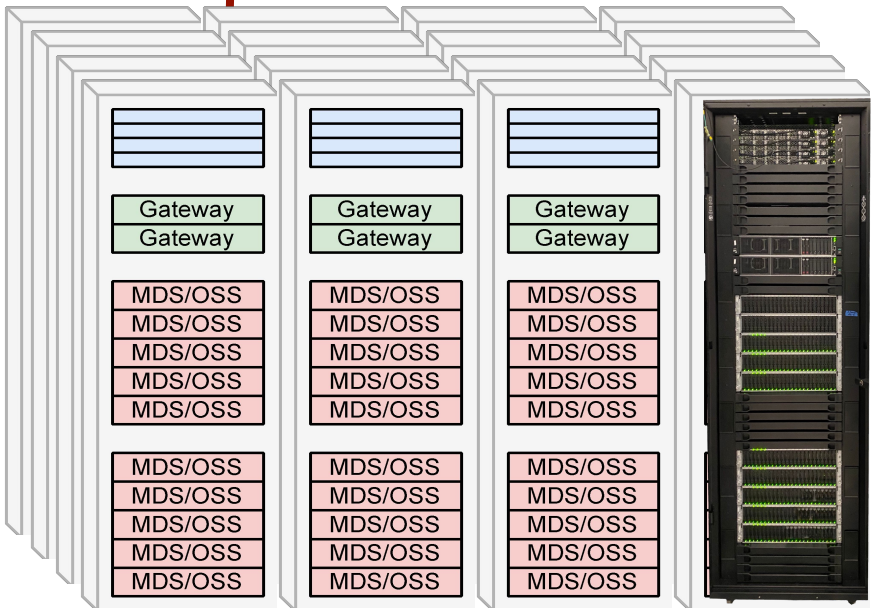
Daily read and write values (Cori) [TB]



Perlmutter's Scratch File System



kfabric Lustre network driver (kfilnd) to communicate with clients on computes



- >7k A100 GPUs, > 3k CPU-only nodes
- 33 PB usable (later expanded to 36PB), all-NVMe Lustre
  274 OSSes (later 298), 16 MDSes
- ClusterStor E1000 enclosures
- LDISKFS (based on EXT4) backend file system

All-flash Lustre is Blazingly Fast

- Perlmutter peak > 7 TB/s, previous Cori HDD Scratch ~750 GB/s
- Metadata on flash enables much faster listing and lookups
  - create 100,000 0-KiB files/sec – unlink 75,000 files/sec

# Clients	Transfer Size	I/O Pattern	I/O API	Number of Files	Required		Read					
					Read GB/s	Write GB/s	GB/s	# Clients	% Clients	PPN	JobID	% of Required
1	4 KiB	random	POSIX	N	0.3	0.3	1.66	1	0.0%	96	10223518	553%
1	4 KiB	random	POSIX	N	0.3	0.3	0.92	1	0.0%	96	10223522	305%
1	1 MiB	strided	POSIX	N	10.0	10.0	24.25	1	0.0%	32	10223520	243%
1	1 MiB	strided	POSIX	N	10.0	10.0	18.63	1	0.0%	32	10223533	186%
15%	4 KiB	random	POSIX	N	28.0	28.0	394.61	691	15.0%	48	10223534	1409%
15%	1 MiB	strided	POSIX	N	2,000.0	2,000.0	6,523.51	691	15.0%	4	10223508	326%
15%	128 MiB	strided	POSIX	N	2,000.0	2,000.0	5,347.32	691	15.0%	4	10223510	267%
15%	1 MiB	strided	MPI-IO	1	1,500.0	1,500.0	2,673.72	691	15.0%	32	10223562	178%
15%	128 MiB	N/A	HDF5	1	1,000.0	1,000.0	2,539.77	691	15.0%	4	10223805	254%
90%	1 MiB	strided	POSIX	N	6,000.0	6,000.0	7,961.45	4147	90.0%	2	10223480	133%
90%	1 MiB	strided	MPI-IO	1	2,000.0	2,000.0	3,554.25	4147	90.0%	2	10224064	178%
90%	128 MiB	N/A	HDF5	1	2,000.0	2,000.0	7,088.94	4147	90.0%	2	10223867	354%

NERSC Embraced Detailed Monitoring to Understand and Track Performance

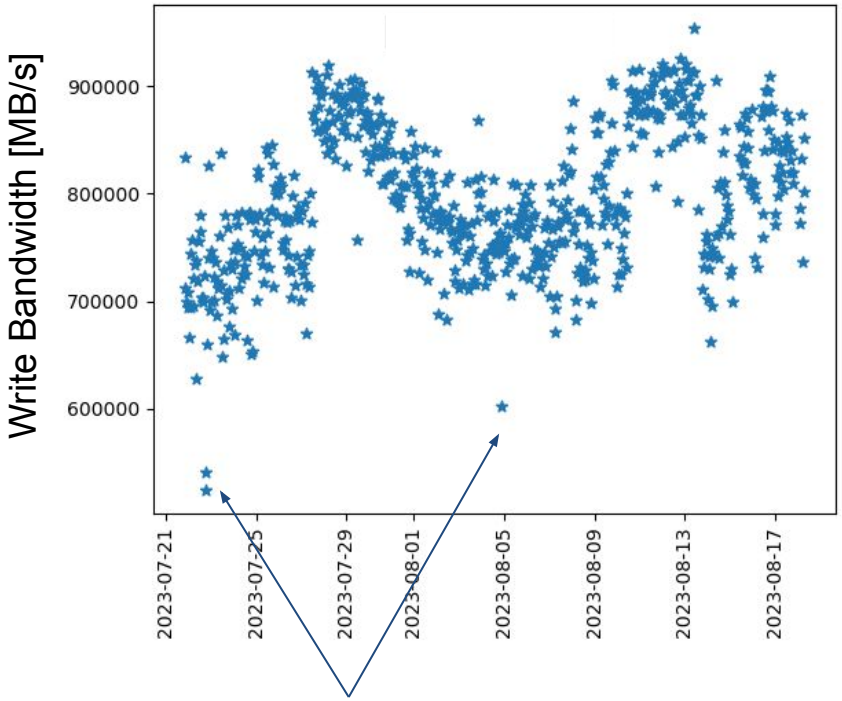
- IOR: Standard application-side parallel IO Benchmark
  - Since Fall 2023, NERSC has performed 5,937 IOR runs
  - Test parameters
    - POSIX, file-per-process, read and write
    - ~250TB from 32 compute nodes (64 processes / node)
    - 1MB xfer and block size,
    - Each file is striped to its own OST
- obdfilter-survey: Simulates Lustre client I/O
  - Tests performed daily on all OSTs
  - Evening test to reduce effects from interactive usage

Metrics for Progress and Success

- Performance variation is a barrier to users science
  - Planning and estimation important for production workflows
  - Major barrier to “realtime” computing - increasing importance at NERSC
- Characterize variation in performance data with Coefficient Of Variation (COV)
  - Defined as the standard deviation divided by the mean.
  - Averaged over 48 hours (twice wall time job limit) to ensure job mix
  - NERSC’s goal for COV is 8% or less
    - This is in line with our expectations for application benchmarks

All-flash Lustre Performance Challenges

- Early IOR runs revealed unexplained performance variation on a healthy, quiet system
- Ad-hoc obdfilter-survey runs showed intermittent write rates 25% to 50% slower than expected
- After considerable investigation, found this was due to inherent, unique challenges of solid state media



~60% drop

Solid State Media Has Unique Challenges

- **Garbage Collection (GC)**
  - SSD unit of access doesn't match the (often larger) unit of erasure
  - Unlike data on a spinning disk, solid state data are never overwritten, instead re-writes will
    - Be placed in new, unused blocks
    - Old locations will be marked invalid and erased for re-use
  - This is a background process managed by individual drives
  - GC can introduce performance variation with sustained writes
- **Trimming**
  - The device must be told if addresses are no longer in-use by the file system so those locations can be erased for future use
  - This is a manual process typically scheduled at an interval

Drive Utilization Amplifies Solid State Challenges

- Due to workload, utilization of drive capacity increased over time:
  - Expected consistent performance
  - Found write performance fell as utilization approached 75%
- Utilization created issues:
  - Garbage Collection: at around 75% utilization, GC activities during write workloads would impact performance
  - Linux Block Allocation:
    - A production file system at 75% full has fairly fragmented free space
    - High NVMe write rates place demands on block allocation
    - An issue with a specific block allocation loop (c1) was not effective and introducing a high CPU load, significantly reducing write rates
  - Write Block Throttling (WBT) - better to disable not throttle IO to drive
  - Reset OST allocation point (mb_last_group) to zero (lets drive manage GC)

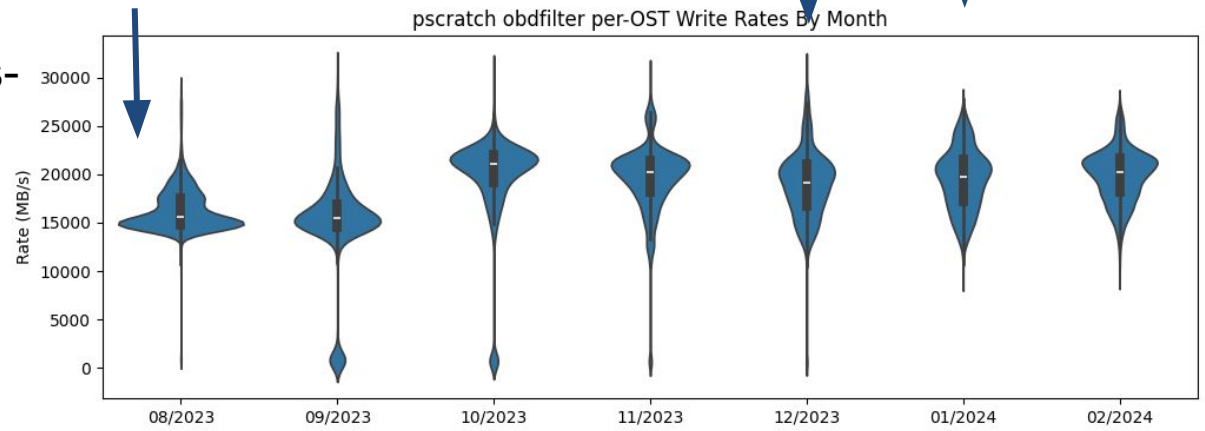
Overcoming The Challenges

Before Aug 2023 Trim before obdfilter

Dec 2023 WBT disabled

c1 loop skipped, Utilization reduced

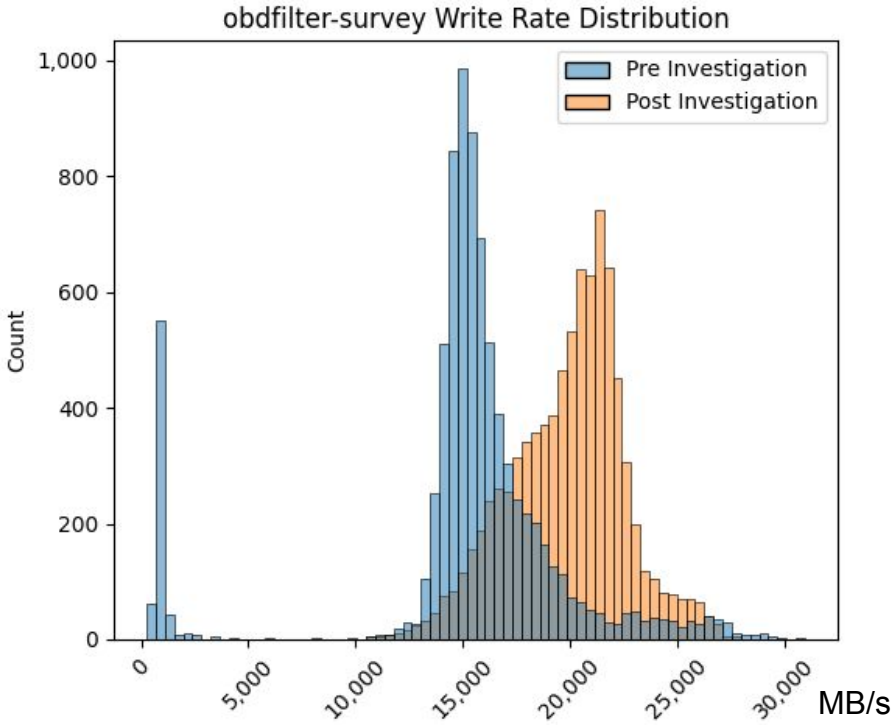
30 GB/s-Write



- Most issues with write. Long tails can be due to degraded OSTs, competing load, or failed over OSTs.
- Until Dec 2023 tails approach zero - very bad for user experience
- By Feb 2024 more OSTs achieve the ~20GB/s range

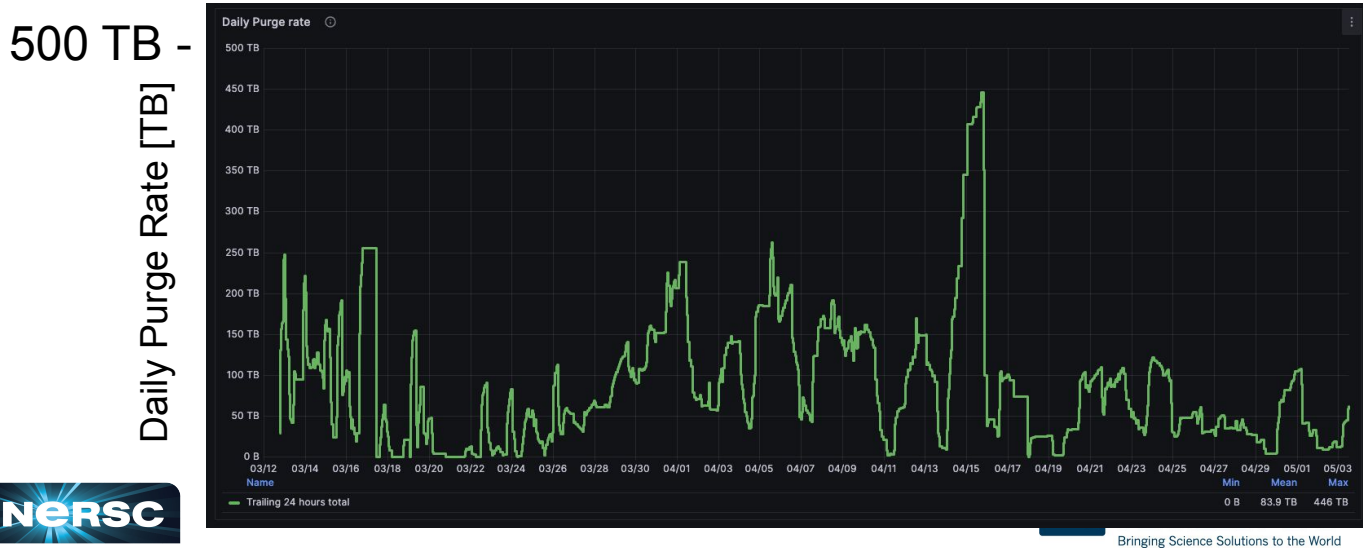
Overall Performance Improvements

- First and last month of investigation OST write performance
  - End-of-tail (<1GB/s) resolved
  - Median write bandwidth nearly 50% improved
- Cumulative effect of all improvements



Purging: Keeping Capacity in Check

- For optimal performance, need to keep OSTs less than 75% full
- This is a challenge on an active production file system
- NERSC Policy: all files not accessed in 8 weeks are eligible for purging
- Consume metadata created by the ClusterStor Data Services to generate lists of files to delete



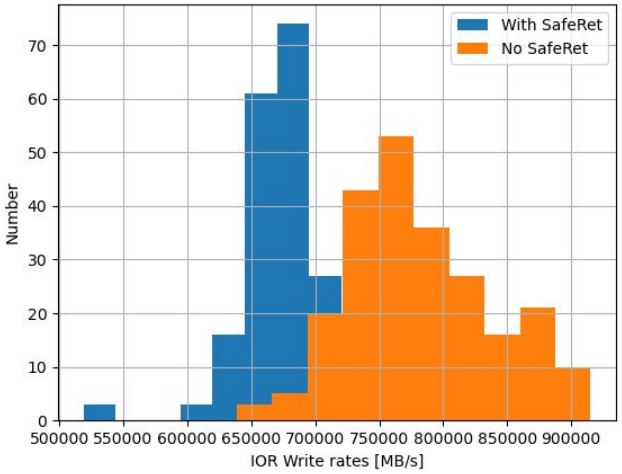
Daily purge rate - March to May 2024

Automated OST Monitoring and Management

- In addition to reducing overall capacity, want to reduce the likelihood a single OST will be above 75%
- Enabled Lustre's Weighted-Free Space allocator to strongly disfavor the fullest OSTs
  - Could be overruled by a user choosing a particular striping pattern, but most users just use the default striping of one OST
- Also automatically detect when an OST is fuller than 75% and disables new writes to that OST
  - Later added a limit to not disable too many at a time

Other performance issues: Security Mitigation for Subprocess: aka safe-ret

- August 2023 AMD CVE
- BIOS patch and enabled a safe-ret kernel feature
- Fix added an overhead to each operation
- Caused a 14% decrease in IOR write bandwidth
- No current solution that's compatible with NERSC's security posture



Return Address Security Bulletin

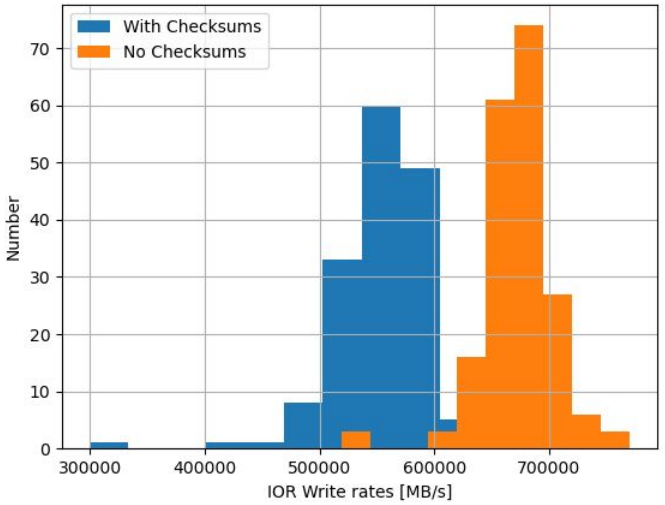
Bulletin ID: AMD-SB-7005
Potential Impact: Data Confidentiality
Severity: Medium

Summary

AMD has received an external report titled 'INCEPTION', describing a new speculative side channel attack. The attack can result in speculative execution at an attacker-controlled address, potentially

Other performance issues: Checksums

- November 2023 HPE bulletin advised enabling checksums for systems using Lustre over kfilnd to avoid data corruption
- Previously no reports of data corruption, but since enabling checksums, two incidents were observed where potential corruption was caught
- Data integrity is assured, but comes with a 17% drop in mean IOR write bandwidth

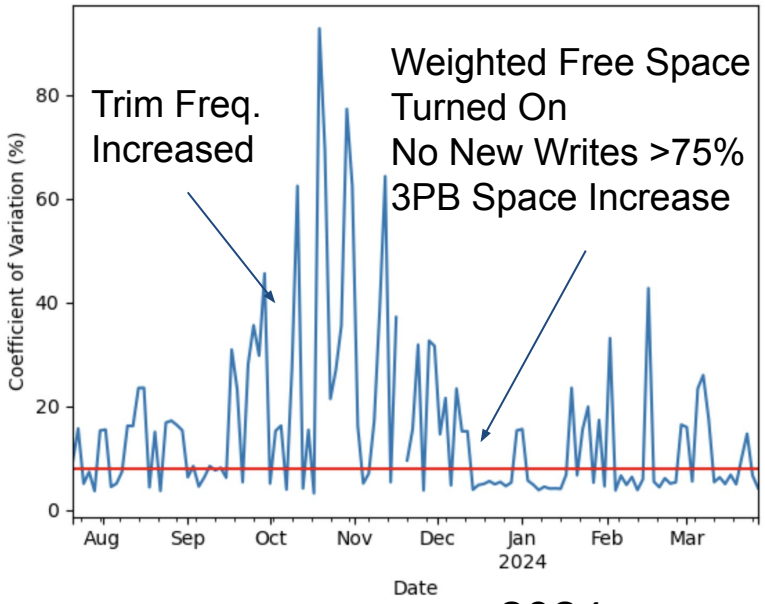


See Significant Improvements in Stability

- Changes to trim frequency, lustre block allocation, and OST usage sufficient to dramatically improve stability by the beginning of 2024

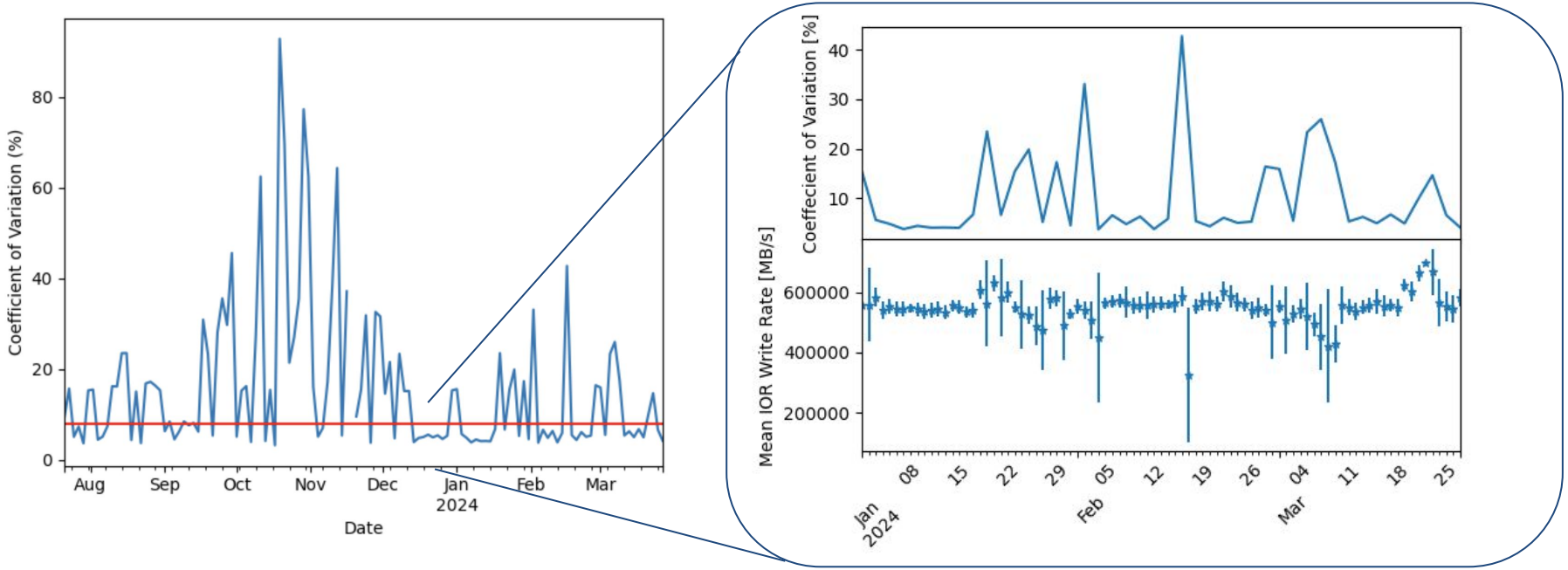
80%--

COV Metric (%)



2024 →

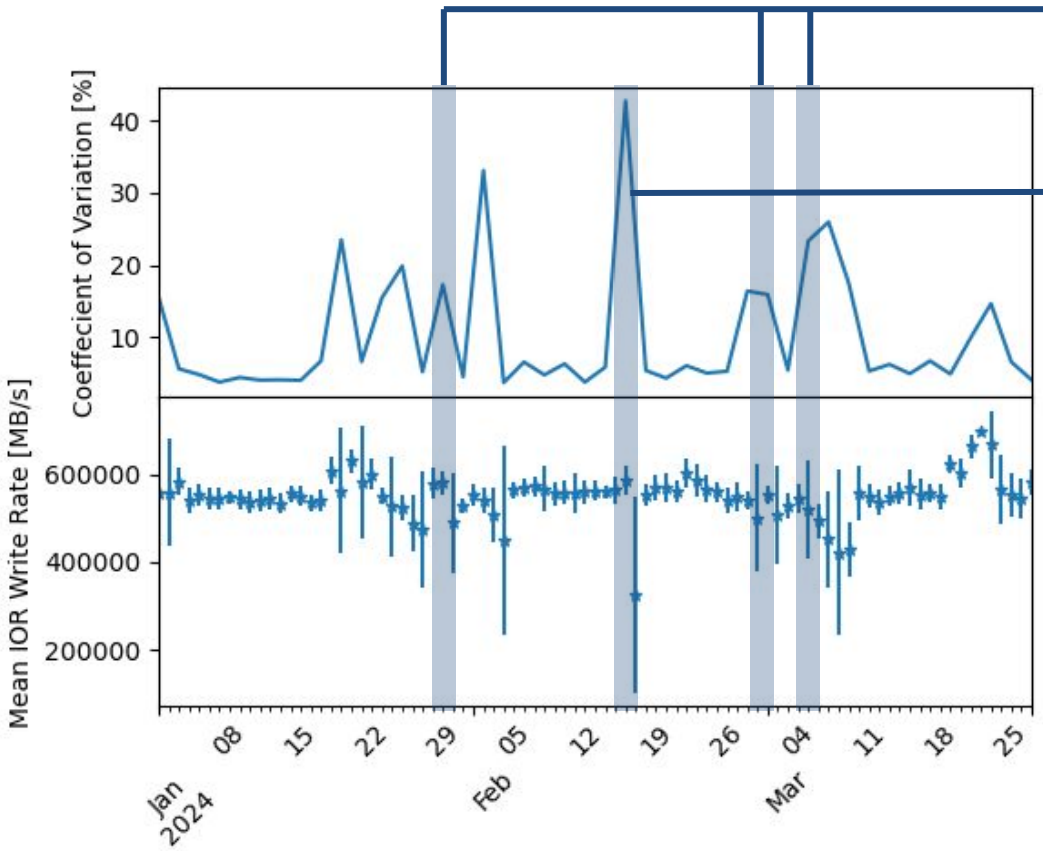
Current Performance of Active File System



In 2024 there have only been 13 instances where COV > 8%

- Due to operational rather than SSD inherent issues.
- Can analyse each case to drive further operational understanding

Understanding Usage and Tuning Allocators

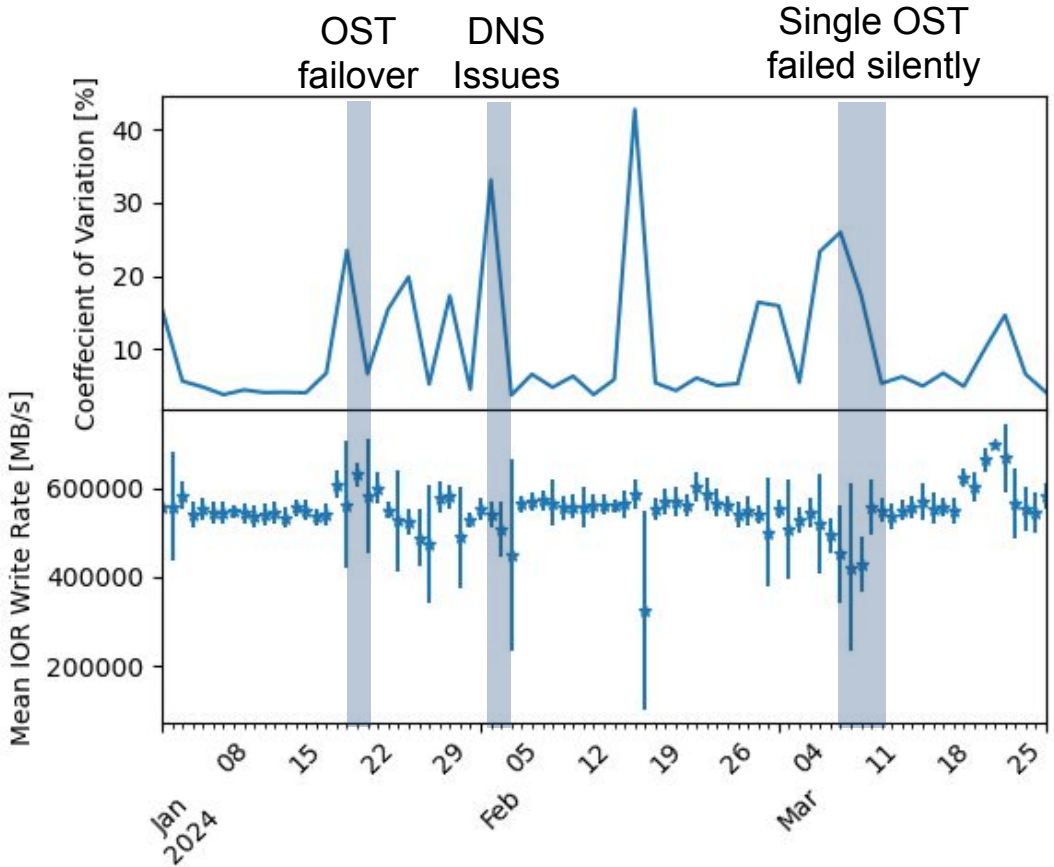


Too many OSTs disabled for >75% full

OST filled up

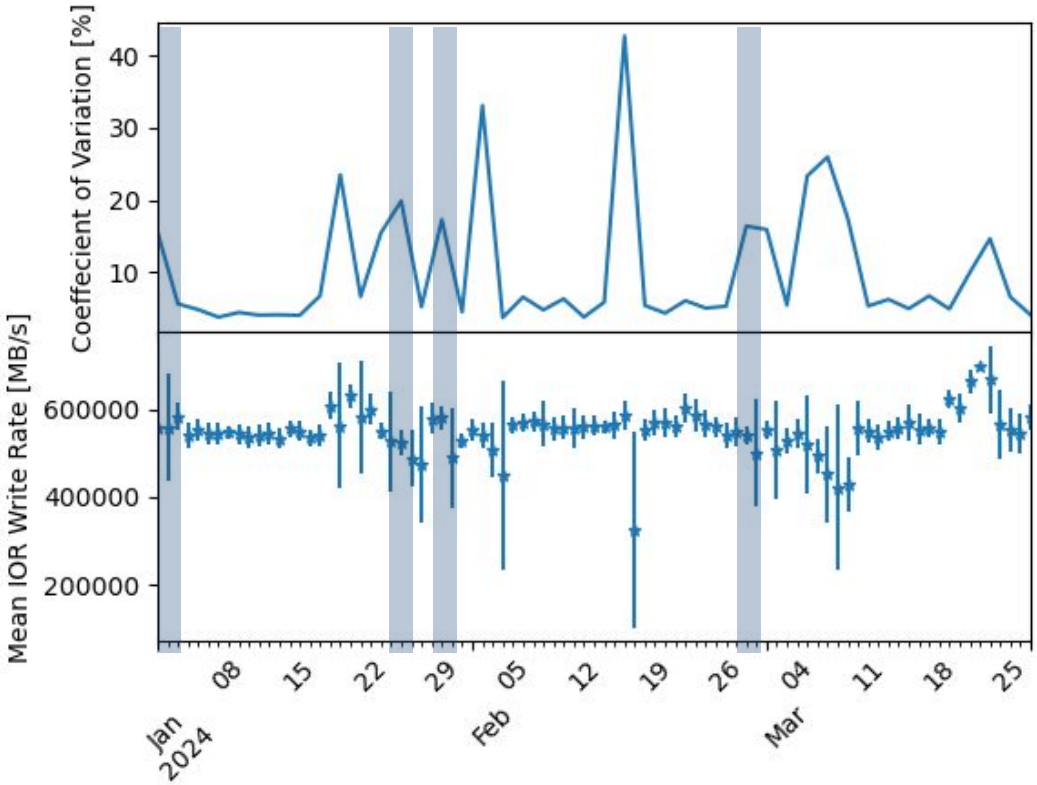
Needed to build operational experience and tuning to balance filesystem usage behavior and keep individual OST usage below 75%

Monitoring Also Reflects Other System Issues



With all this analysis only 4 incidents remain unexplained

Remaining Issues



- Four incidents remain to be further understood
  - Significant improvement over 2023

Conclusion

- NERSC and HPE have worked together to deliver a high-performance, and low-variation, scratch file system
  - Comprehensive test suite can quickly identify issues
- Even with extensive testing and monitoring, diagnosing issues on a busy production file system is a challenge
- File system capacity plays a larger than expected role in All-SSD Lustre filesystem
  - Current optimal OST fullness is less than 75%
- We are continuing to work to improve performance and stability