“Reading Tea Leaves”

Understanding internal events and addressing performance issues within a CephFS/XRootD Storage Element.

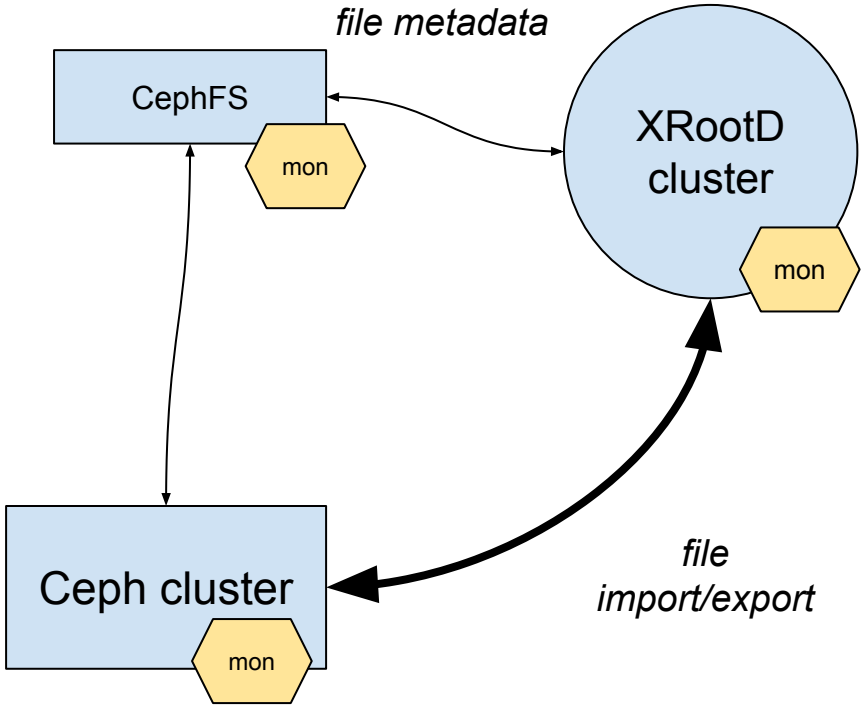
Gerard Hand, Matt Doidge, Steven Simpson
Lancaster University/GridPP
CHEP 2024, Krakow

Framing the Issue: Ceph(FS) and XRootD

The WLCG Grid Site at Lancaster has leveraged using XRootD to front a CephFS volume as the site Storage Element.

This is a complex system with many moving parts, problems can cause a cascade of knock on issues.

One “family” of issues commonly experienced are what are colloquially referred to as “Slow Ops”.



Details on the Lancaster Setup

CEPH:

- Ceph Reef installed using cephadm
- 32 OSD Nodes, 768 OSDs, 12.5PB (Raw), NVMe for Bluestore, ≈105 PGs per OSD
- 3 MON+MGR
- 1 Active + 1 Standby MDS
- 8+3 Erasure Coding

XRootD:

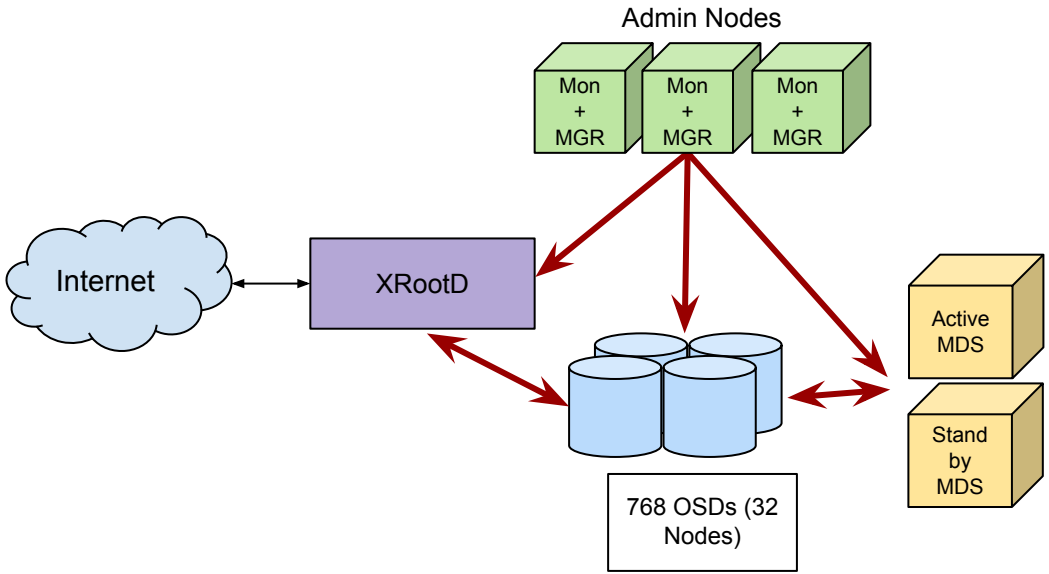
- XRootD 5.7.1
- 1 redirector and 6 gateways
- Ceph mounted using CephFS (kernel mount)

Monitoring and Alerting:

- Prometheus, Loki, Grafana
- Node/Ceph exporters; Promtail
- Custom XRootD metrics collectors

Other:

- CephFS also mounted RDONLY on the worker nodes.



Slow ops?

As one might expect, a Slow Op (Slow Operation) is an operation that takes longer than expected to complete. These operations can related to various sub-systems (MDS, OSD, MON). Lancaster is using the default warning threshold time of 30 seconds.

Slow Ops cover the majority of problems seen during the operation of our storage. These have mainly been slow OSD operations and some slow MDS operations.

An OSD operation can be considered as a collection of sub-operations/tasks. Ceph records each stage of an operation as it progresses: Message → PG → OSDs → Completion.

Details of historic OSD ops and the current OSD ops can be exported using:

`ceph daemon osd.<id> dump_historic_ops` or `ceph daemon osd.<id> dump_ops_in_flight`

These commands return a JSON-formatted list of operations and events which can be used to help identify where operations are taking the longest.

Problems they cause

When slow ops occur, there will be reduced cluster performance. At times of persistent slow OSD ops, we see:

- Possible CephFS client I/O timeouts.
- MDS daemons with slow metadata IO.
- Slow MDS ops.
- CephFS clients not responding to capability release.
- MDS_TRIM errors (pool-cache sync).
- XRootD client failures due to timeouts and high IOWAIT load.

The issues from Slow Ops arise from the complex interlocking chains of operations occurring within the Ceph(FS) system - any unexpected increase in latency between operations can cause compounding problems.

Identifying a “Slow Op”, fingerprints in the monitoring

The first place to see Slow Ops is in the Ceph Status from the command line:

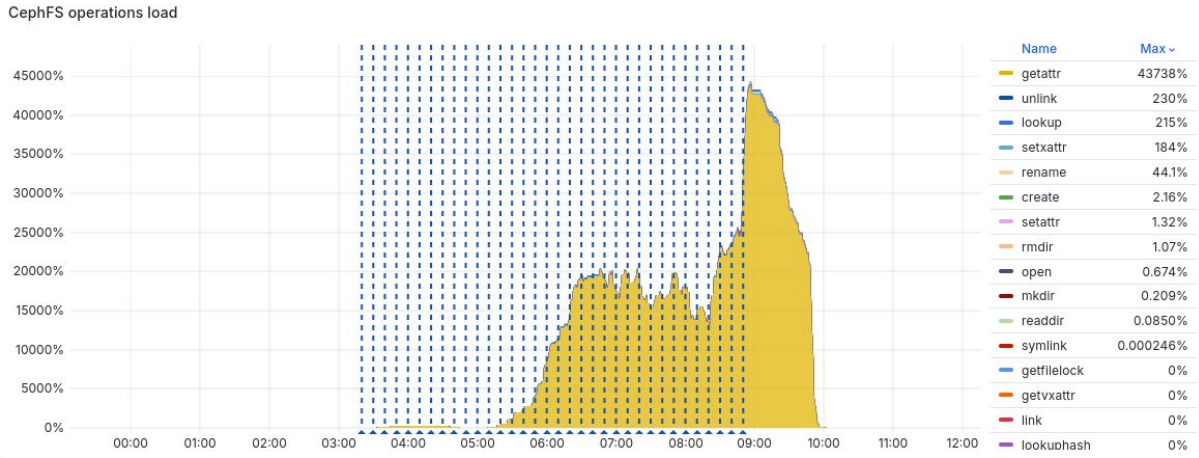
```
# ceph status
  cluster:
    id:     abcdef123-1111-22AA-334f-123456789012
    health: HEALTH_WARN
            7 slow ops, oldest one blocked for 223 sec, daemons [osd.302,osd.568,osd.87] have slow ops.
```

Also in the system logs/journal:

```
2024-04-18T12:31:50.480977+0100 osd.87 [WRN] slow request osd_op(client.5834877.0:12345058 14.54s0
14:2a0a59dc:::100f38bdcfb.00000623:head [read 0~4194304] snapc 0=[] ondisk+read+known_if_redirected
e351030) initiated 2024-04-18T11:31:13.149690+0000 currently started
```

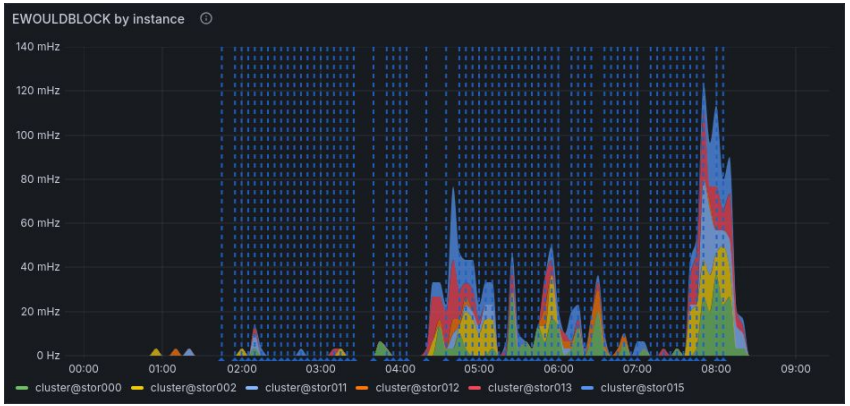
Impact of Slow Ops - getattr

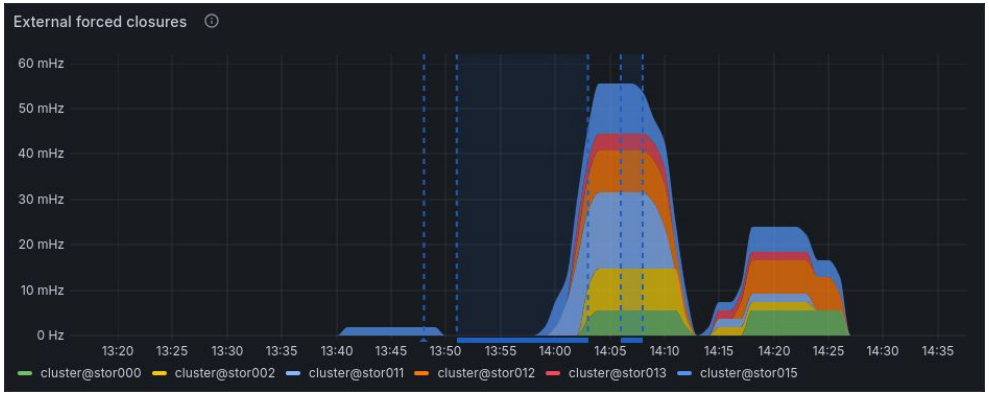
- Most used CephFS operation.
- Load/latency increase massively with Slow Ops the underlying cause.

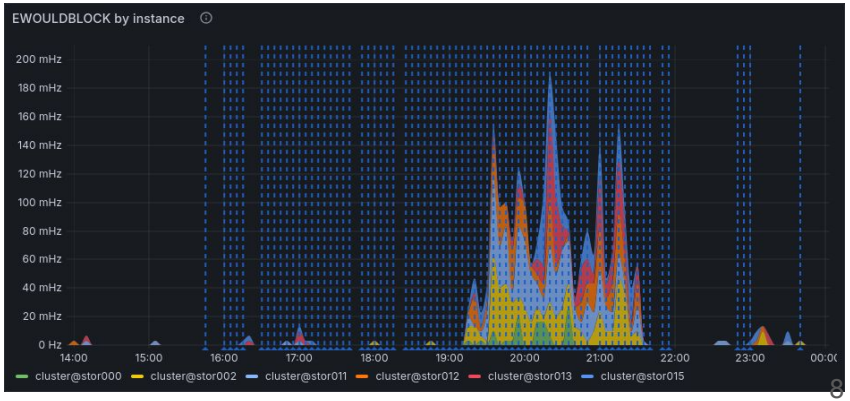


Impact of Slow Ops - XRootD timeouts, forced closures

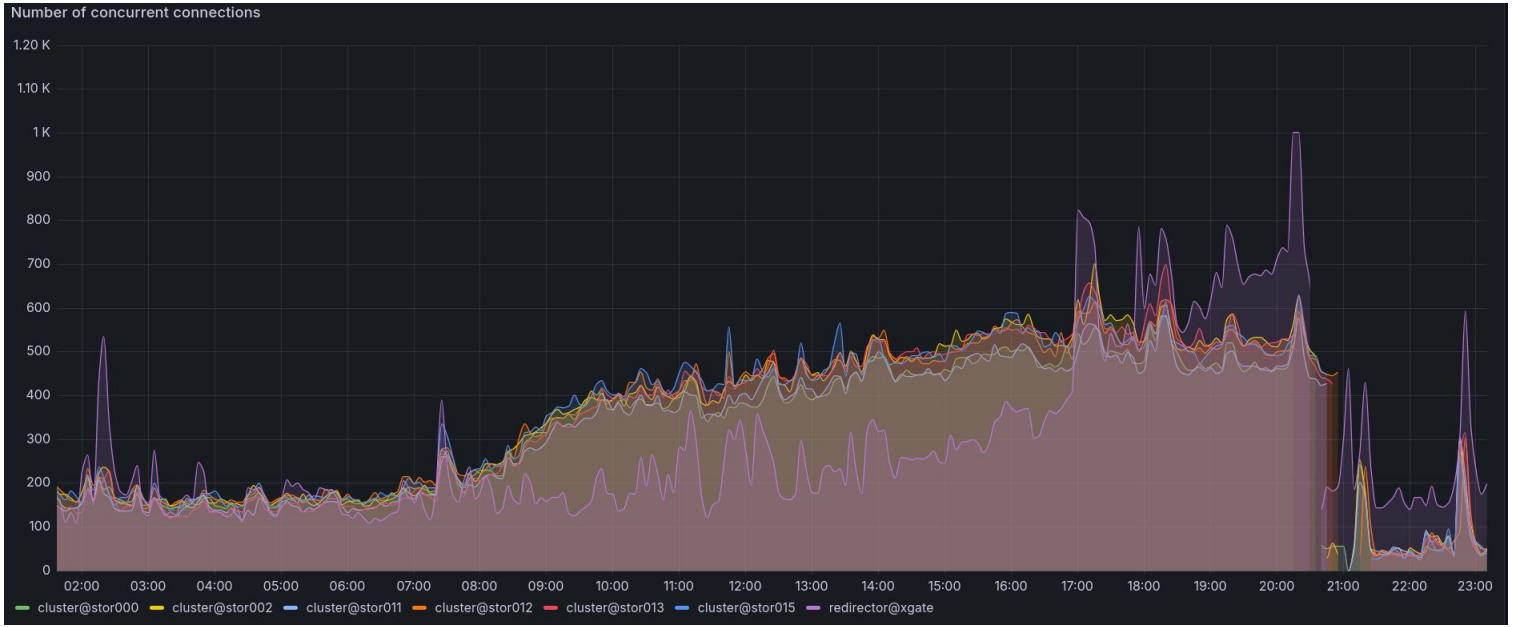
- Mostly send and endsess timeouts (right plots).
- Transfer failures (below).







XRootD Concurrent Connections - A Useful Canary

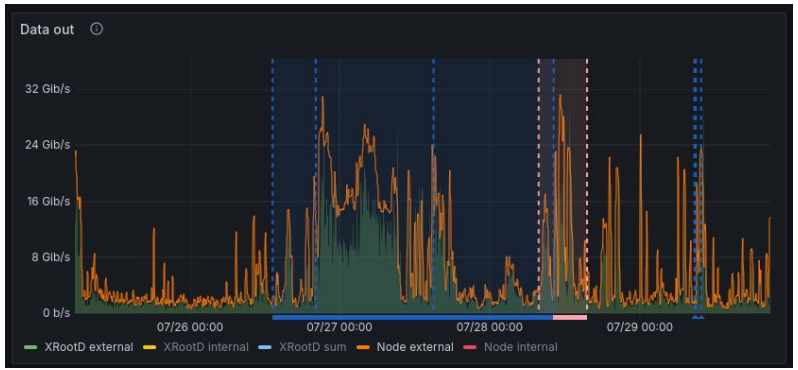


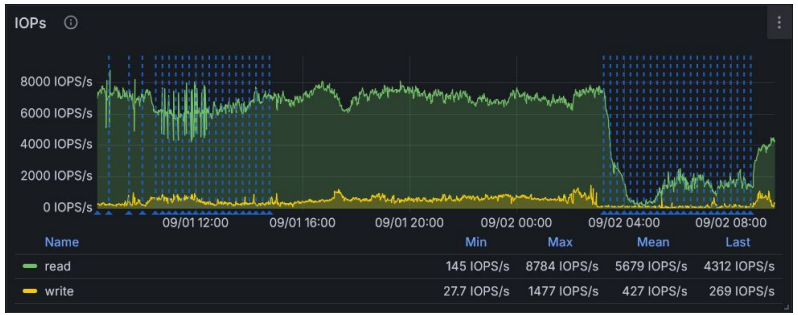
Client issues cause by problems with the CephFS mounts can often be seen in build ups of concurrent connections, caused for example connections fail to close cleanly.

Things that may cause Slow Ops

- Full OSDs or Cluster as a whole.
- Overloaded systems:
  - CPU maxing out or overheating and throttling.
  - Network congestion.
  - Insufficient RAM.
- Failing hardware (particularly spinning disks).
- Exceptional high load from clients.
- RocksDB compaction taking too long.
- Bios configuration e.g., latency introduced by CPU c-state transitions.
- Ceph misconfiguration.
  - PGs per OSD too high - Increased memory usage, PG log length and PG stat updates.
  - Backfilling and Recovery settings too high.
- System activity: Scrubbing, recovery, backfilling.

Causes: gateway load?

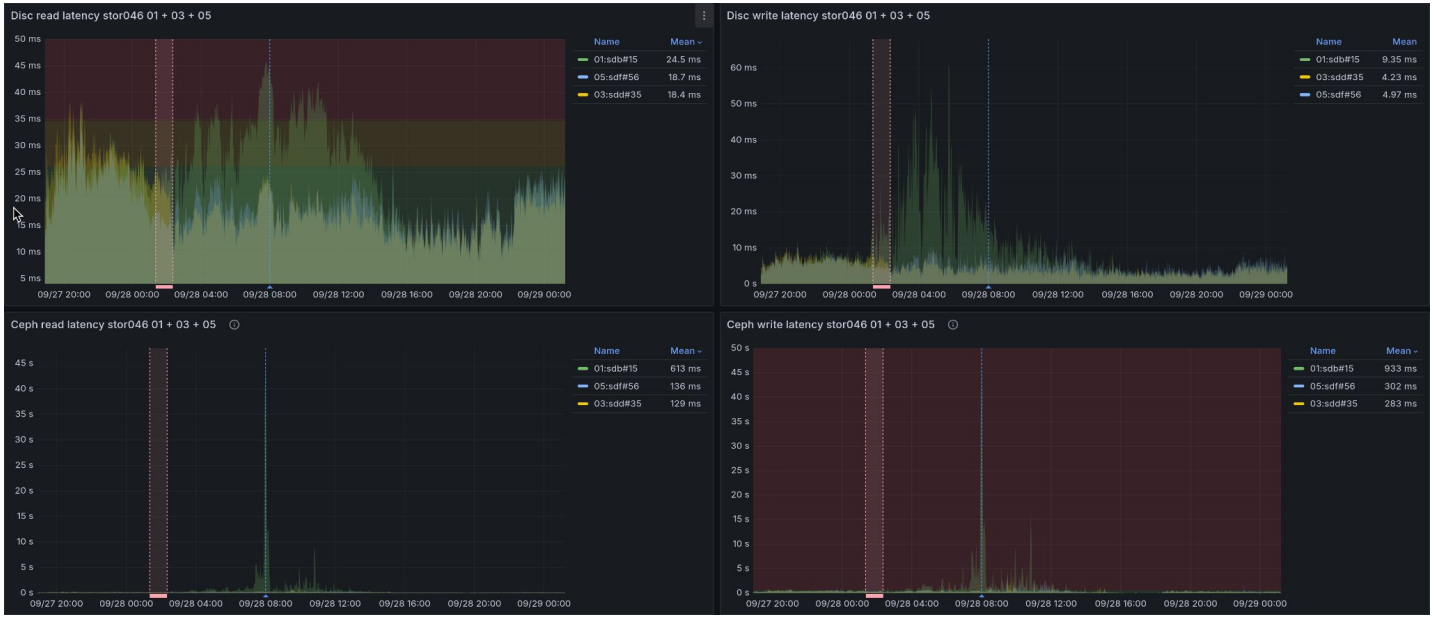




- XRootD load looked like it triggered Slow Ops.
- Larger area ⇒ local reads via XRootD (rare and unwanted)
- Smaller area ⇒ Slow Ops
- Doesn't always occur on high load
- Happens without high load

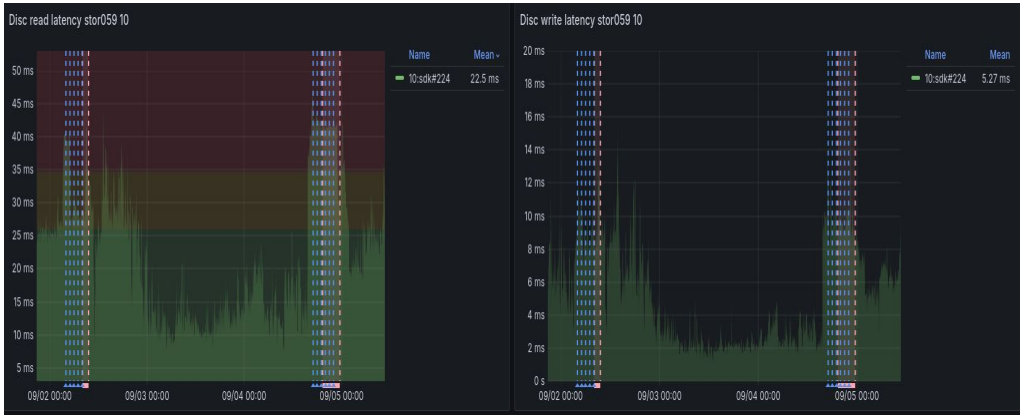
Causes: disc latency?

- Disc I/O latency (top row).
- Slow reads and writes precede brief moment of Ceph Slow Ops.
- But doesn't always happen when discs are slow.
- Happens when discs are running well.



Causes: post-drain?

- OSD switches from draining (right).
- Initial struggle to fill?



- Slow discs are seen here too (left).

Attempting to replicate a slow op manually

Initially the slow ops seemed to happen when data rates were high and there was a significant increase in the write rate. We attempted to to recreate the slow ops so that we could effectively determine if system changes improved the situation.

- Read/wrote/deleted data via multiple CephFS mounts.
- Tried various combinations of reading, writing and deleting.
- Tried various file sizes and number of files. Used file sizes from 4KB to 40GB.
- Total dataset size was approximately 1PB.

Despite pushing the IOPs and data rates higher than seen in normal day to day operations, we were unable to reliably recreate slow OSD ops. There were two instances of slow ops but multiple re-runs of the test failed to reproduce the slow ops.

Triggering alerts on what we have learned

- Slow Ops (warnings within Ceph).
- High OSD usage (>87%).
- High total usage (>90%).
- High rate of XRootD forced closures.
- High objects-per-PG.
- High PGs-per-OSD deviation.
- Inconsistent PGs.
- Disc defects.
- High Ceph r/w (process) latency.
  - read/write process > 0.16s
  - getattr > 0.35s
- High disc latency.
  - read > 0.0336s

Mitigating Slow Ops

Short term:

- Manual intervention to take steps to resolve slow ops.
- Identify bad OSDs/nodes/clients early and deal with them.
- Increase the amount of free storage.

Longer term

- Implement a manager module which monitors and determines a cause of slow ops and takes action. (Stop scrubbing/backfilling, restart OSDs if necessary, etc).
- “In-flight” XRootD checksumming to reduce the load from transfers,
- Hope that Ceph implements improvements that reduce causes of slow ops, or improves the logging to ease identifying the causes of slow op alerts.

Wrap-up/Conclusion

- Slow ops can be the root cause of many issues in a CephFS system.
  - Negatively impacts operations.
- Understanding the causes is difficult, in part because they are emergent and difficult to replicate artificially.
  - A combination of heavy writes/reads and background processes.
- Early alerting and extensive monitoring can enable soothing the issues early, reducing impact.
  - But predicting and pre-empting is still out of our reach.