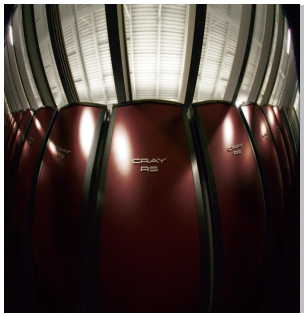


Profiling Platform Storage Using IO500 and Mistral

Nolan Monnier, **Jay Lofstead**,

Margaret Lawson, Matthew Curry

PDSW 2019

SAND2019-14085 C

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Astra Testbed

Sandia National Laboratories

- Prove viability of advanced technologies for NNSA integrated codes, at scale
- Expand the HPC-ecosystem by developing emerging yet-to-be proven technologies
- Is technology viable for future ATS/CTS platforms supporting ASC mission?
- Increase technology AND integrator choices
- Buy down risk and increase technology and vendor choices for future NNSA production platforms
- Ability to accept higher risk allows for more/faster technology advancement
- Lowers/eliminates mission risk and significantly reduces investment
- Jointly address hardware and software technologies
- First Prototype platform targeting Arm Architecture
- **#156 on top500 (June 2019) (2.2 PF)**

Astra Hardware

Sandia National Laboratories

HPE Apollo 70 Chassis: 4 nodes

18 chassis/rack

72 nodes/rack

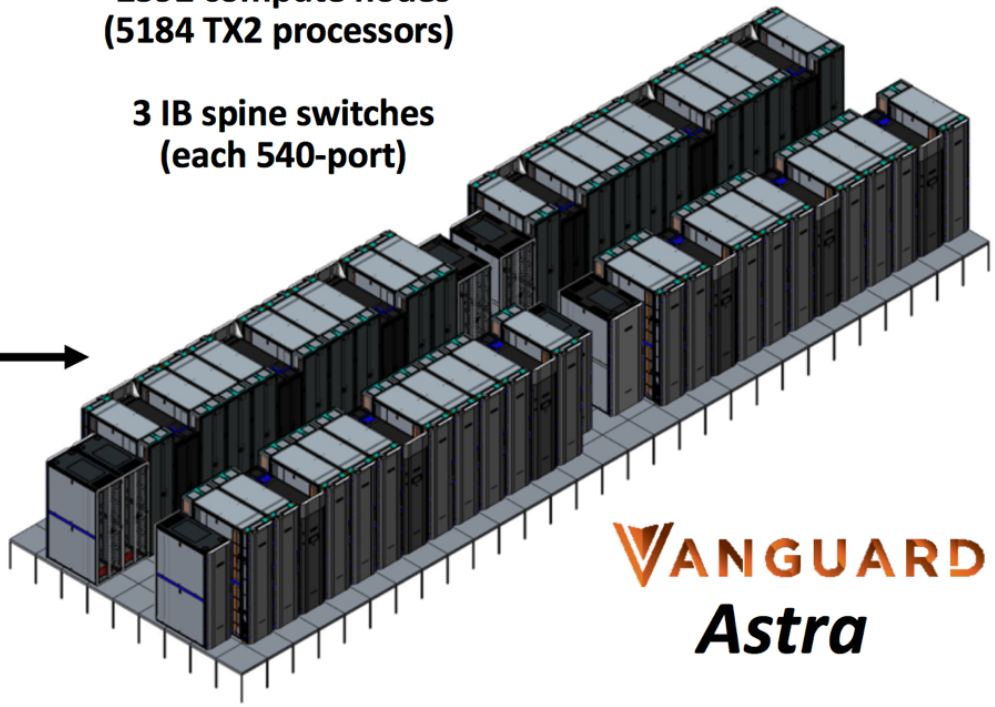
3 IB switches/rack
(one 36-port switch per 6 chassis)

HPE Apollo 70 Rack

36 compute racks
(9 scalable units, each 4 racks)

2592 compute nodes
(5184 TX2 processors)

3 IB spine switches
(each 540-port)

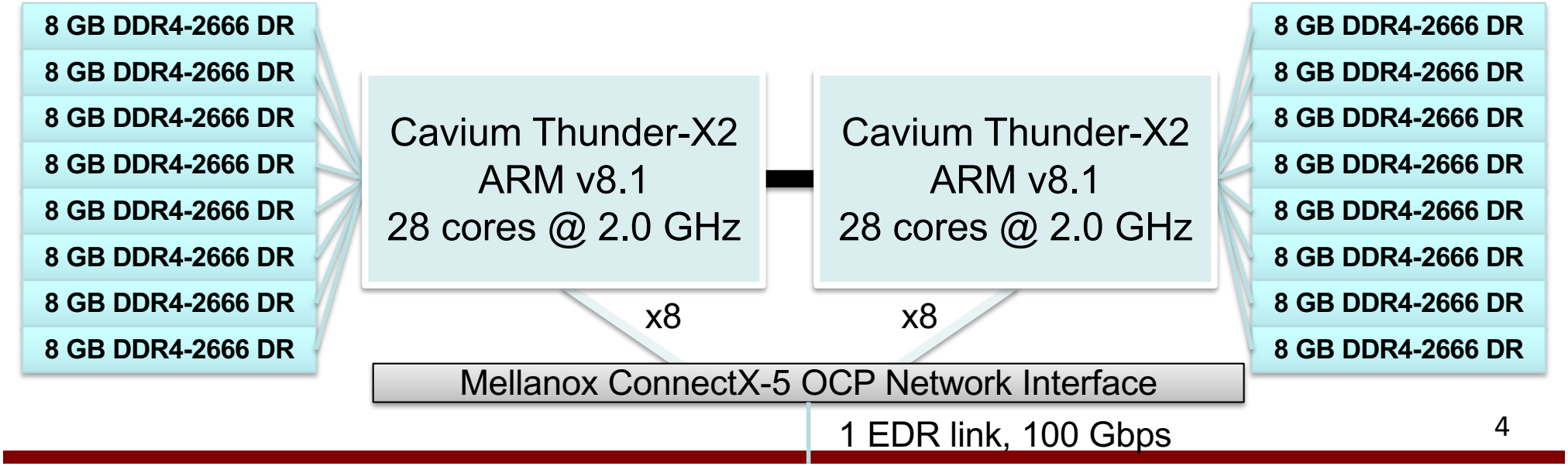


VANGUARD
Astra

7

Astra System Specs

- 2,592 HPE Apollo 70 compute nodes
- Cavium Thunder-X2 Arm SoC, 28 core, 2.0 GHz
- 5,184 CPUs, 145,152 cores, 2.3 PFLOPs system peak
- 128GB DDR Memory per node (8 memory channels per socket)
- Aggregate capacity: 332 TB, Aggregate Bandwidth: 885 TB/s
- Mellanox IB EDR, ConnectX-5
- HPE Apollo 4520 All–flash storage, Lustre parallel file-system
- Capacity: 403 TB (usable)
- Bandwidth 244 GB/s



Astra Storage Specs

Sandia National Laboratories

- `/lustre` filesystem
  - 40 OSSes
  - 2 MDSes
  - 21 1.6TB NVMe devices per OSS spread across 3 ZFS pools per node using raidz
  - 240 GB/sec peak bandwidth
  - 990 TB usable storage
- `/oscratch` filesystem
  - 8 OSSes, each serving 10 OSTs
  - 2 MDSes, 1 MDT each

- Storage has full bandwidth to interconnect core (4xEDR 100 Gpbs Infiniband)

IO500

Sandia National Laboratories

- Virtual Institute for IO (https://www.vi4io.org) hosted
- http://io500.org
- Collect details information about storage and platforms
- Provide a balanced way to compare storage systems
  - No, it is not perfect
- MDTest, IOR, and find
- “hard” tests for worst case scenarios
- “easy” tests user configurable to showcase a system’s potential
- Find to represent walking the file tree for purge or similar
- Geometric mean of all values to get resulting score
- Published at ISC and SC every year (Tuesday, 12:15, 205-207)

Profiling IO

Sandia National Laboratories

- Darshan is available, but we wanted to see how this Ellexus' Mistral tool worked and then compare
  - Future work
- Mistral offers node-level statistics
  - Breeze is the per process tool
- Collect on a per-second basis
- Ideal configuration is Elasticsearch and Grafana

- How well is our ARM-based Lustre client and the back end Lustre system working?
- Can we learn using IO500's setup instead of apps or app proxies?

Tuning IO500 to Push the System

Sandia National Laboratories

1) Obtain system information and theoretical characteris- tics.
2) Set test directories' stripe size based on test files' size and number of storage targets.
3) Determine number of nodes to use.
4) Increase the cores per node to maximize bandwidth, until the bandwidth for ior easy reasonably approaches a theoretical limit.
5) Adjust the cores per node to balance bandwidth and metadata results

Tuning Final Configuration

Sandia National Laboratories

- Configured striping to best suit the ”hard” tests
- Set nodes used == to number of storage targets
  - `/lustre` used 58 procs (on the 58 cores) on 121 nodes
- 10 node challenge ran proportional configurations
- Metadata did not require special stripe configurations

IO500 Scores

Sandia National Laboratories

Nodes	Mount Point	Bandwidth	IOPS	Score
121	/lustre	84.8118 GB/s	35.9847 kiops	55.2443
10	/lustre	28.4097 GB/s	45.7227 kiops	36.0412

- 14th on the ISC19 overall list
- 13th on the ISC19 10 node challenge

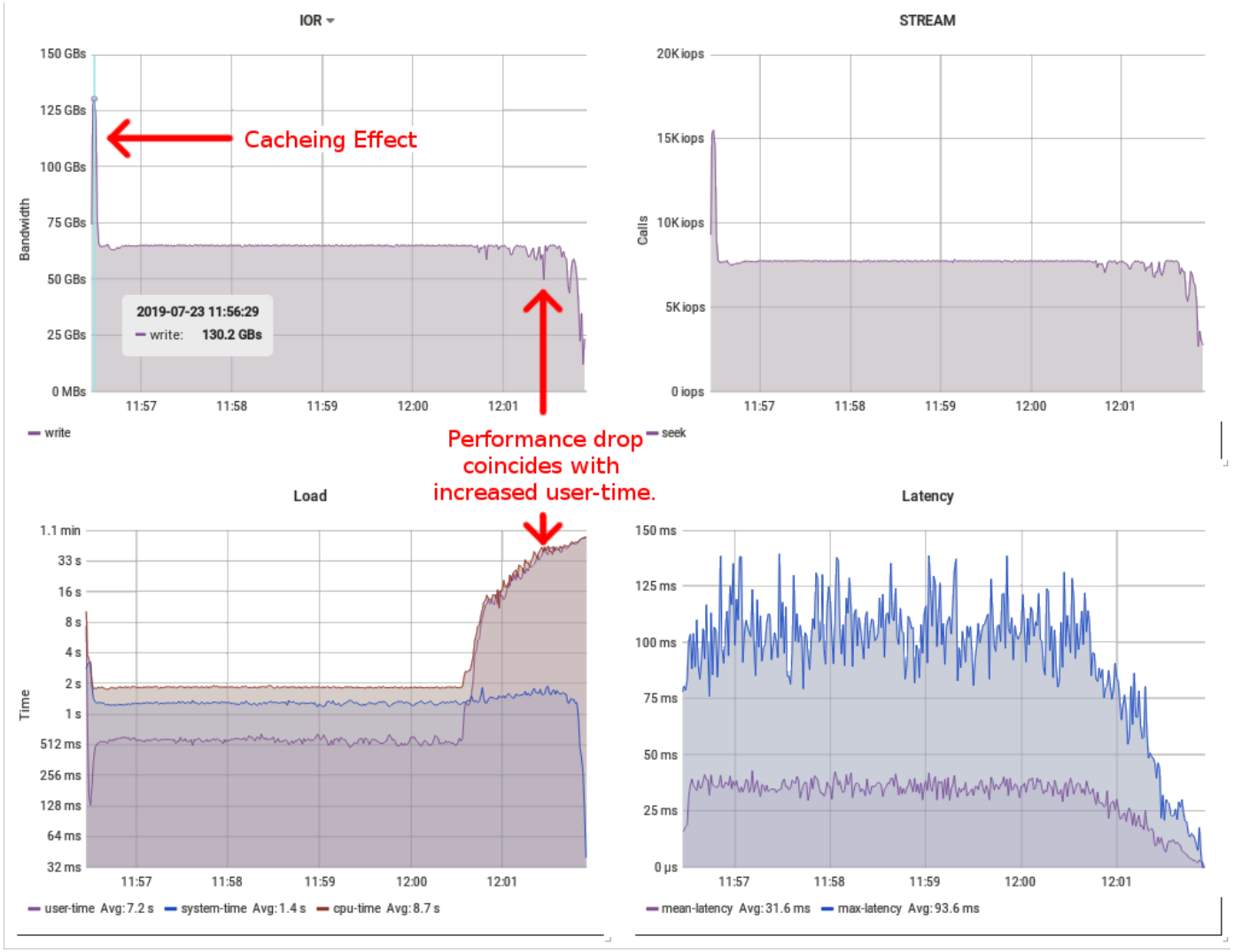
- Plug: (see new results at IO500 BoF Tuesday 12:15, 205-207)

Problems!

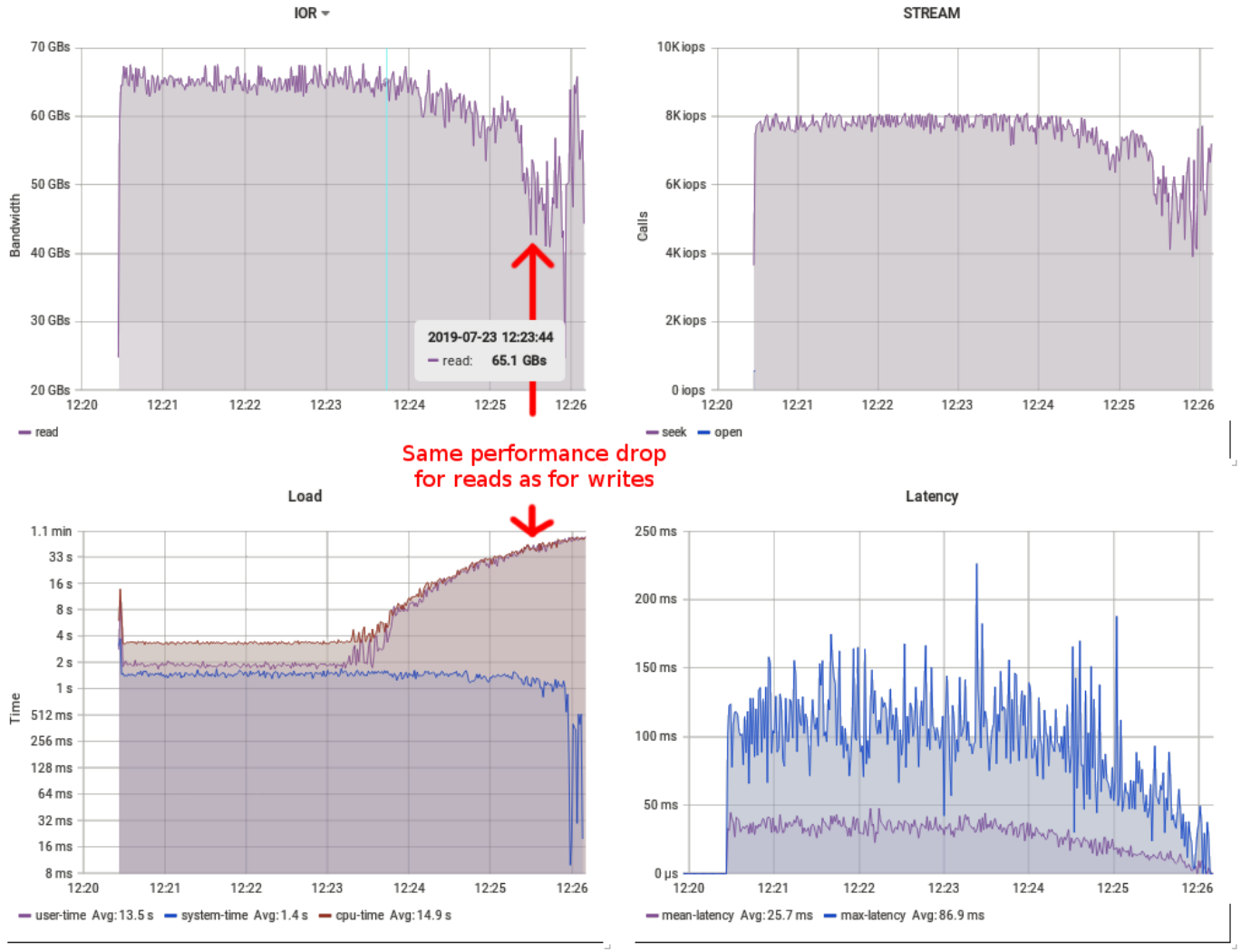
Sandia National Laboratories

- On `/oscratch` filesystem, we ran out of space in directory when creating files
  - Lustre has an option to extend capacity, but we were not allowed to change it.
- This prevented getting “official” `/oscratch` results

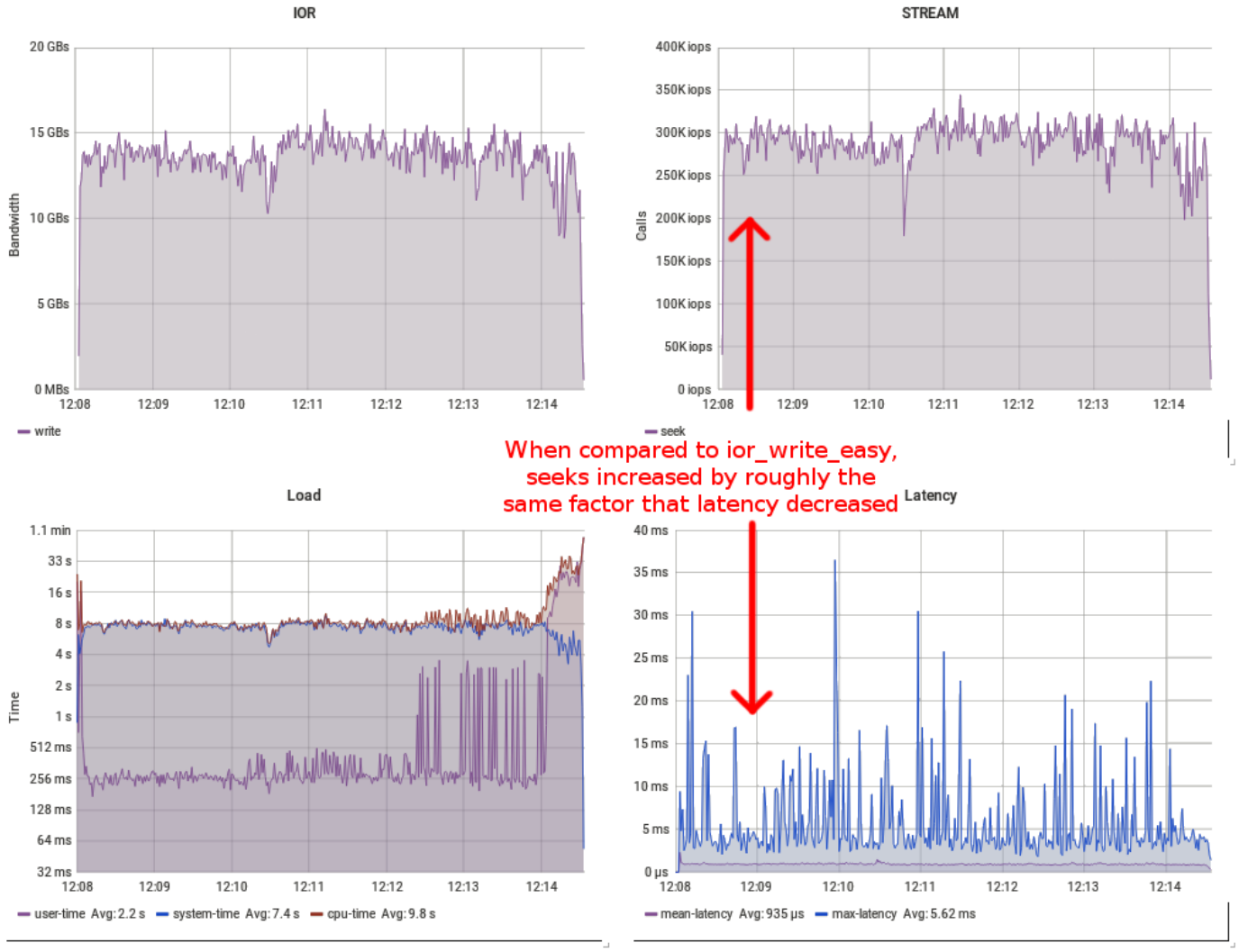
IOR Write Easy



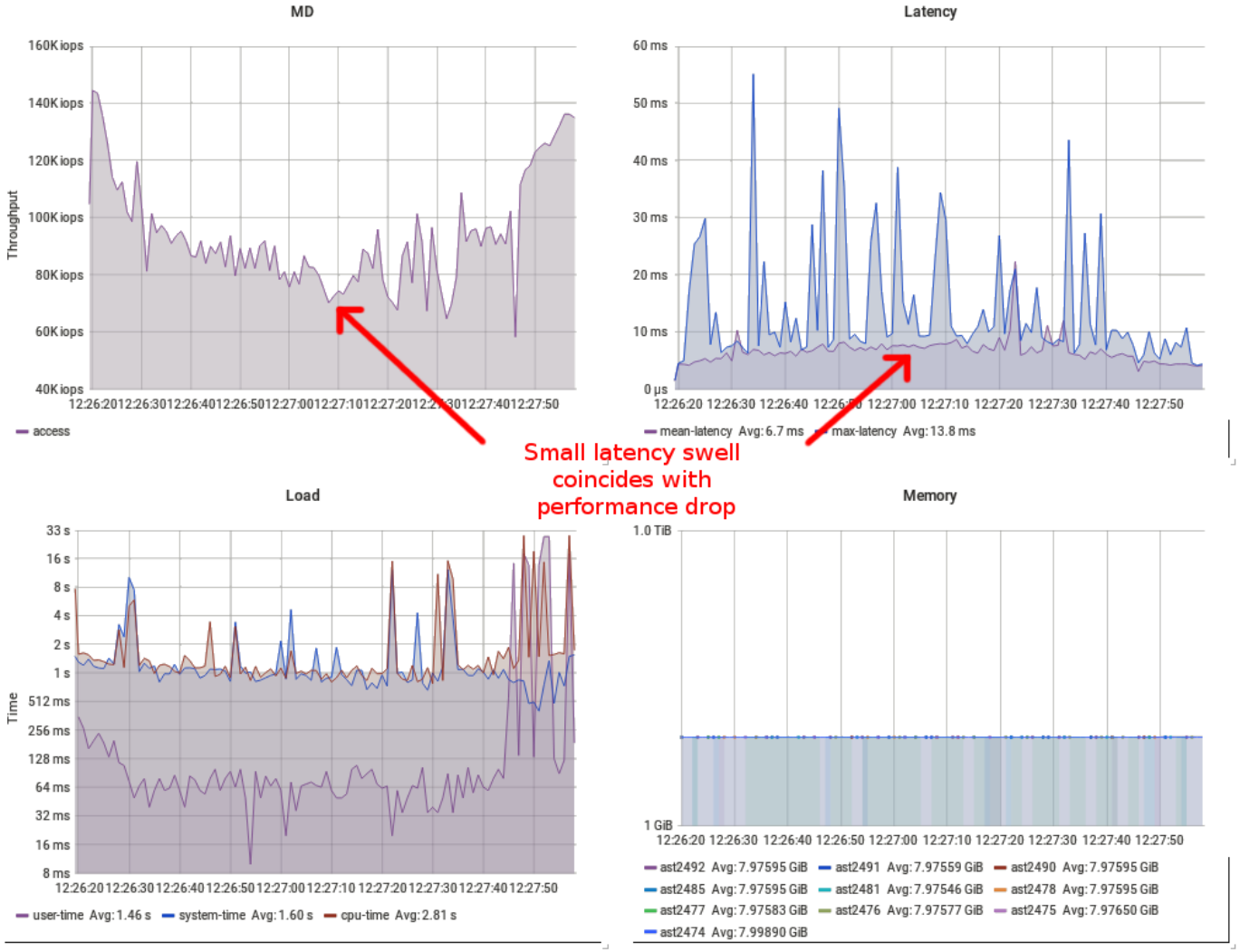
IOR Read Easy



IOR Write Hard



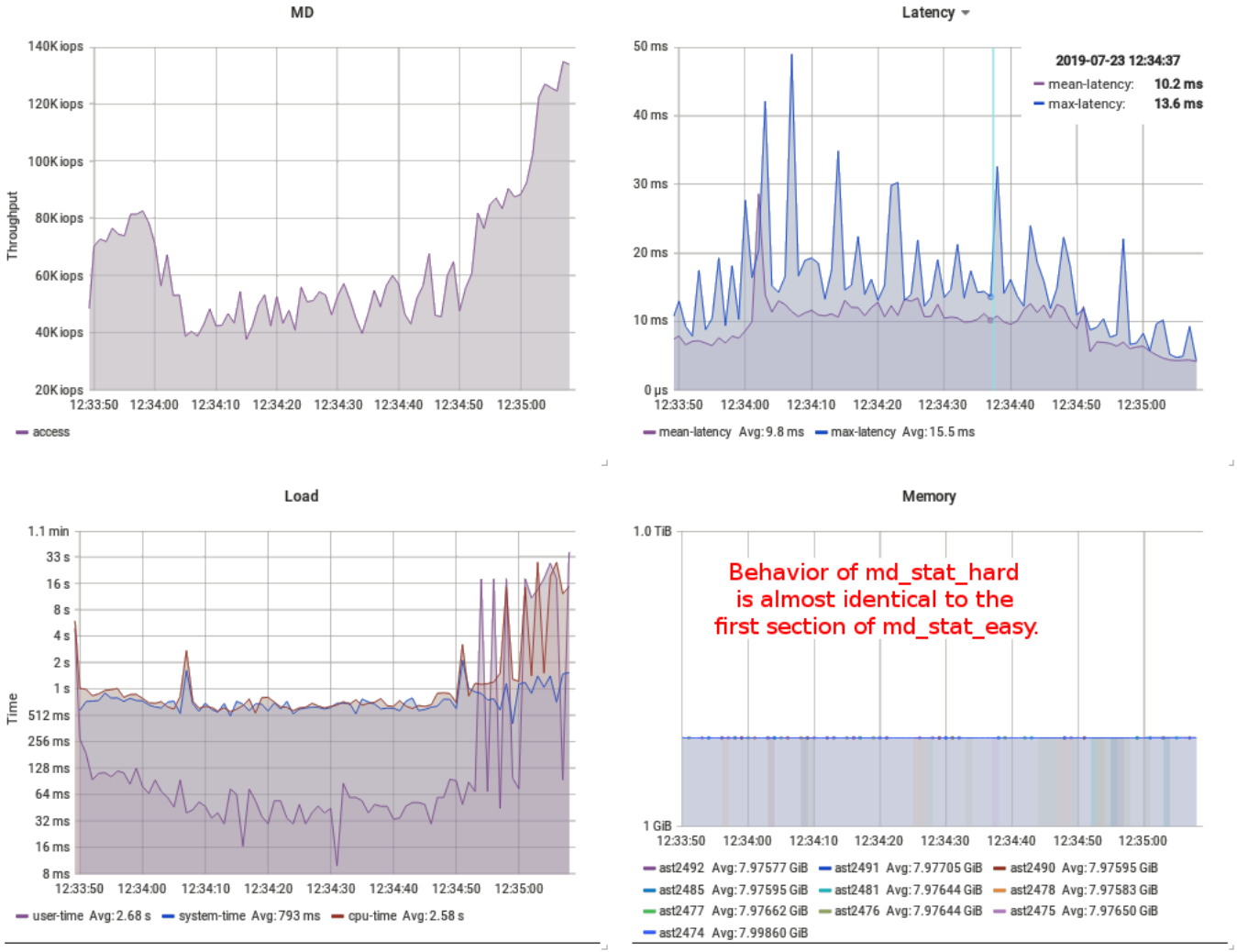
MD Stat Easy


MD
160K iops
140K iops
120K iops
100K iops
80K iops
60K iops
40K iops
Throughput
access
Latency
60 ms
50 ms
40 ms
30 ms
20 ms
10 ms
0 µs
mean-latency Avg: 6.7 ms
max-latency Avg: 13.8 ms
Small latency swell coincides with performance drop
Load
33 s
16 s
8 s
4 s
2 s
1 s
512 ms
256 ms
128 ms
64 ms
32 ms
16 ms
8 ms
Time
user-time Avg: 1.46 s
system-time Avg: 1.60 s
cpu-time Avg: 2.81 s
Memory
1.0 TiB
1 GiB
ast2492 Avg: 7.97595 GiB
ast2491 Avg: 7.97559 GiB
ast2490 Avg: 7.97595 GiB
ast2485 Avg: 7.97595 GiB
ast2481 Avg: 7.97546 GiB
ast2478 Avg: 7.97595 GiB
ast2477 Avg: 7.97583 GiB
ast2476 Avg: 7.97577 GiB
ast2475 Avg: 7.97650 GiB
ast2474 Avg: 7.99890 GiB

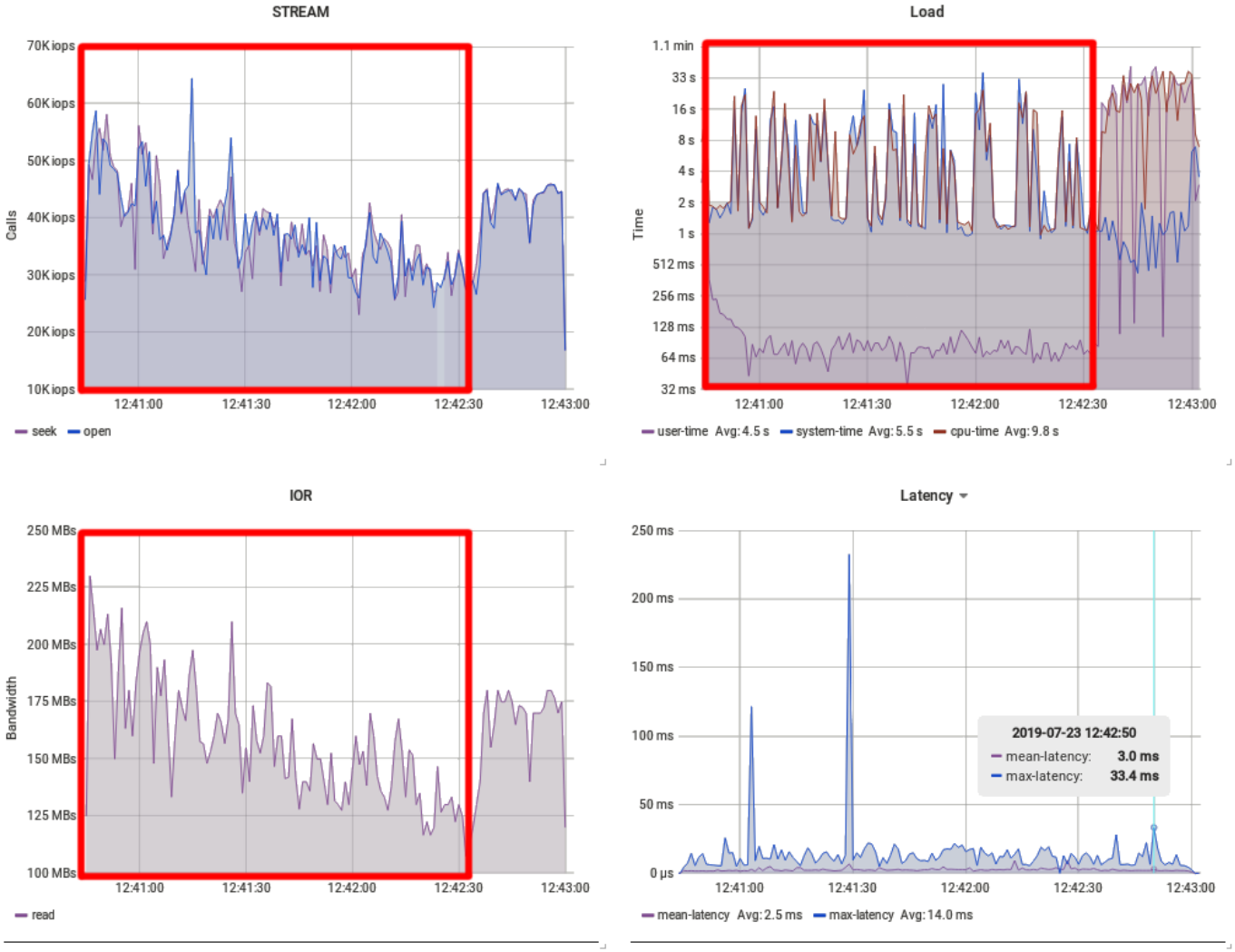
MD Stat Easy


MD
200K iops
150K iops
100K iops
50K iops
0 iops
Throughput
12:26:30
12:27:00
12:27:30
12:28:00
12:28:30
access
Latency
2.5 s
2.0 s
1.5 s
1.0 s
500 ms
0 µs
mean-latency Avg: 31 ms
max-latency Avg: 43 ms
Load
1.1 min
33 s
16 s
8 s
4 s
2 s
1 s
512 ms
256 ms
128 ms
64 ms
32 ms
16 ms
8 ms
Time
user-time Avg: 10.2 s
system-time Avg: 2.2 s
cpu-time Avg: 9.5 s
Memory
1.0 TiB
1 GiB
ast2492 Avg: 7.993 GiB
ast2491 Avg: 7.976 GiB
ast2490 Avg: 7.974 GiB
ast2485 Avg: 7.994 GiB
ast2481 Avg: 8.015 GiB
ast2478 Avg: 7.976 GiB
ast2477 Avg: 7.994 GiB
ast2476 Avg: 7.963 GiB
ast2475 Avg: 7.916 GiB
ast2474 Avg: 7.962 GiB

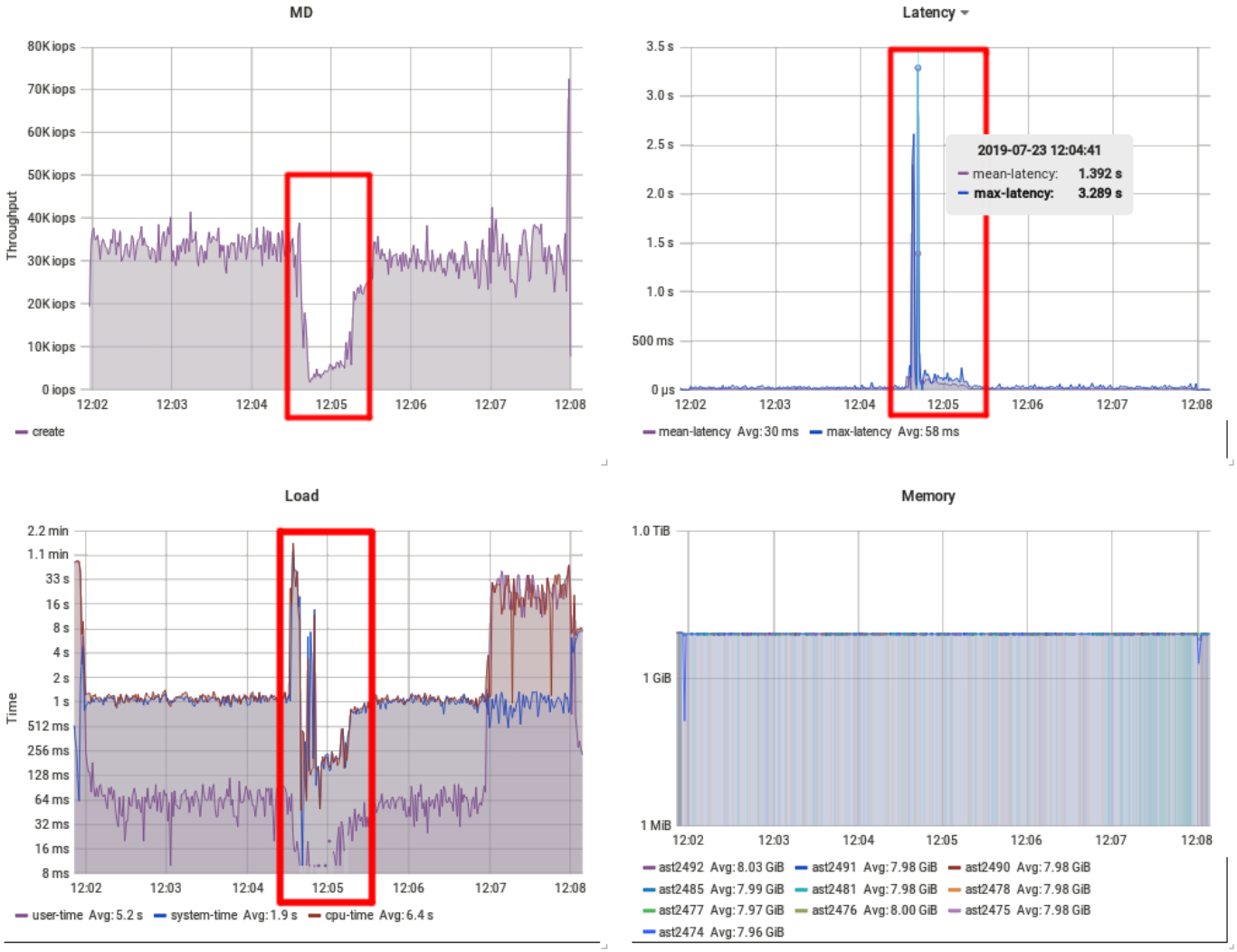
MD Stat Hard


MD
Throughput
140K iops
120K iops
100K iops
80K iops
60K iops
40K iops
20K iops
12:33:50 12:34:00 12:34:10 12:34:20 12:34:30 12:34:40 12:34:50 12:35:00
access
Latency
50 ms
40 ms
30 ms
20 ms
10 ms
0 µs
2019-07-23 12:34:37
mean-latency: 10.2 ms
max-latency: 13.6 ms
mean-latency Avg: 9.8 ms
max-latency Avg: 15.5 ms
Load
Time
1.1 min
33 s
16 s
8 s
4 s
2 s
1 s
512 ms
256 ms
128 ms
64 ms
32 ms
16 ms
8 ms
user-time Avg: 2.68 s
system-time Avg: 793 ms
cpu-time Avg: 2.58 s
Memory
1.0 TiB
1 GiB
Behavior of md_stat_hard
is almost identical to the
first section of md_stat_easy.
ast2492 Avg: 7.97577 GiB
ast2491 Avg: 7.97705 GiB
ast2490 Avg: 7.97595 GiB
ast2485 Avg: 7.97595 GiB
ast2481 Avg: 7.97644 GiB
ast2478 Avg: 7.97583 GiB
ast2477 Avg: 7.97662 GiB
ast2476 Avg: 7.97644 GiB
ast2475 Avg: 7.97650 GiB
ast2474 Avg: 7.99860 GiB

MD Read Hard


STREAM
70K iops
60K iops
50K iops
40K iops
30K iops
20K iops
10K iops
Calls
12:41:00
12:41:30
12:42:00
12:42:30
12:43:00
seek
open
Load
1.1 min
33 s
16 s
8 s
4 s
2 s
1 s
512 ms
256 ms
128 ms
64 ms
32 ms
Time
user-time Avg: 4.5 s
system-time Avg: 5.5 s
cpu-time Avg: 9.8 s
IOR
250 MBs
225 MBs
200 MBs
175 MBs
150 MBs
125 MBs
100 MBs
Bandwidth
read
Latency
250 ms
200 ms
150 ms
100 ms
50 ms
0 µs
2019-07-23 12:42:50
mean-latency: 3.0 ms
max-latency: 33.4 ms
mean-latency Avg: 2.5 ms
max-latency Avg: 14.0 ms

MD Write Easy



Conclusion

Sandia National Laboratories

- Hit 33% of peak (midpoint of 20%-50% of peak expected)
- Reproducibly issues with the Lustre client and MPI we had not isolated before
- Found gradual performance degradations and odd performance fluxuations to debug

Acknowledgements

- Thank you to Ellexus for supporting this work

- Supported under the Vanguard project from NNSA

- Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA0003525.