SUSEcon digital '20

Best Practices For HPC Storage

BP-1122

David Byte, Sr. Technology Strategist

Darren Soothill, Storage Technical Strategist EMEA, SUSE

SUSE

Agenda

1. Understanding HPC workload I/O
2. Storage options and specific thoughts

HPC Workload I/O Patterns

Understanding I/O needs can save you $$$ or prevent a project failure

- Steady
- Bursty
- Batch
- Low Latency
- High Single Stream Throughput
- High Aggregate Throughput

Understand Your Current Workload

General Info:

- Majority of HPC workloads e.g., climate and physics, perform large sequential writes.
- Astronomy and Genomics tend to have small, random I/O.
- Particle Physics tends to have large, sequential read I/O.

BUT, you should still measure your environment to understand I/O needs.

Direction The Codes And Applications Are Going?

Are your codes changing the way they handle storage in light of exascale needs?

What about your vendors?

There’s a lot of interest in object storage vs POSIX file storage.

Ask your vendors and code communities for their direction.

Storage Options

Tons Of Options

NFS

CephFS

Lustre

Spectrum Scale

BeeGFS

Weka

OrangeFS

DAOS

Etc

Hardware Thoughts

To RAID or not to RAID

SATA vs SAS

SSD vs NVMe

Don’t let the network get you down

IB vs Ethernet

Complexity vs Performance vs Scalability vs Cost

Does it make a difference if it is up and serving data in 3 days vs 3 months?

I know it’s fast, but will you really make use of the performance available?

How many nodes, how much data, and how long do you want to keep it online?

What’s the real TCO that includes soft costs, professional services, etc?

Lustre Thoughts

- Low latencies and high throughput ability
- Complex
- Tune stripe count to I/O type (small number of large files vs large number of small files
- Support
- Still lingering uncertainty about ZFS

NFS Thoughts

- Good for small clusters.
- Quick and easy to setup.
- Locking can quickly become a problem.
- Doesn’t scale well.

Ceph Thoughts

- Offers both POSIX file system and Object Interfaces.
- Designed to scale horizontally.
- HDF Group has done some work on a VOL driver for RADOS.
- CephFS on IO-500.
- Not well known in the HPC world, yet.
- Can work as a tier for GPFS, Weka, and Lustre (via DMF).

Ceph/RADOS Better Than Lustre?

When it comes to scalabilty, recent research by Lawrence Berkeley Nation Laboratory, The HDF Group, and Intel indicates this is the case.

https://sdm.lbl.gov/pdc/pubs/201811_PDSW2018-ObjEval.pdf

Work uses HDF5 VOL plugin for RADOS

DAOS

- Designed for Server Class Memory (SCM) and NVMe tech.
- VERY fast.
- Still very early in lifecycle.

Some Best Practices

While you may tune your cluster to take advantage of C-states, you may not want to do this on your storage.

Watch out for drive controllers that choke under pressure.

- Find by testing the devices individually, then as a group. Aggregate should be ~ # of devices x throughput of 1 device.

Use multi-queue block I/O.

Nearline SAS vs SATA.

More Best Practices

Similar To General HPC Tuning

- I/O Scheduler
- NUMA Auto-balance
- THP
- CPU Mitigations
- 1 Socket vs 2+
- Tune the transport
- Make sure hardware layout is optimal

Wrap Up

Grand Finale

What you did yesterday may not work for tomorrow.

Know where your codes are going for storage.

Look at the number of available options.

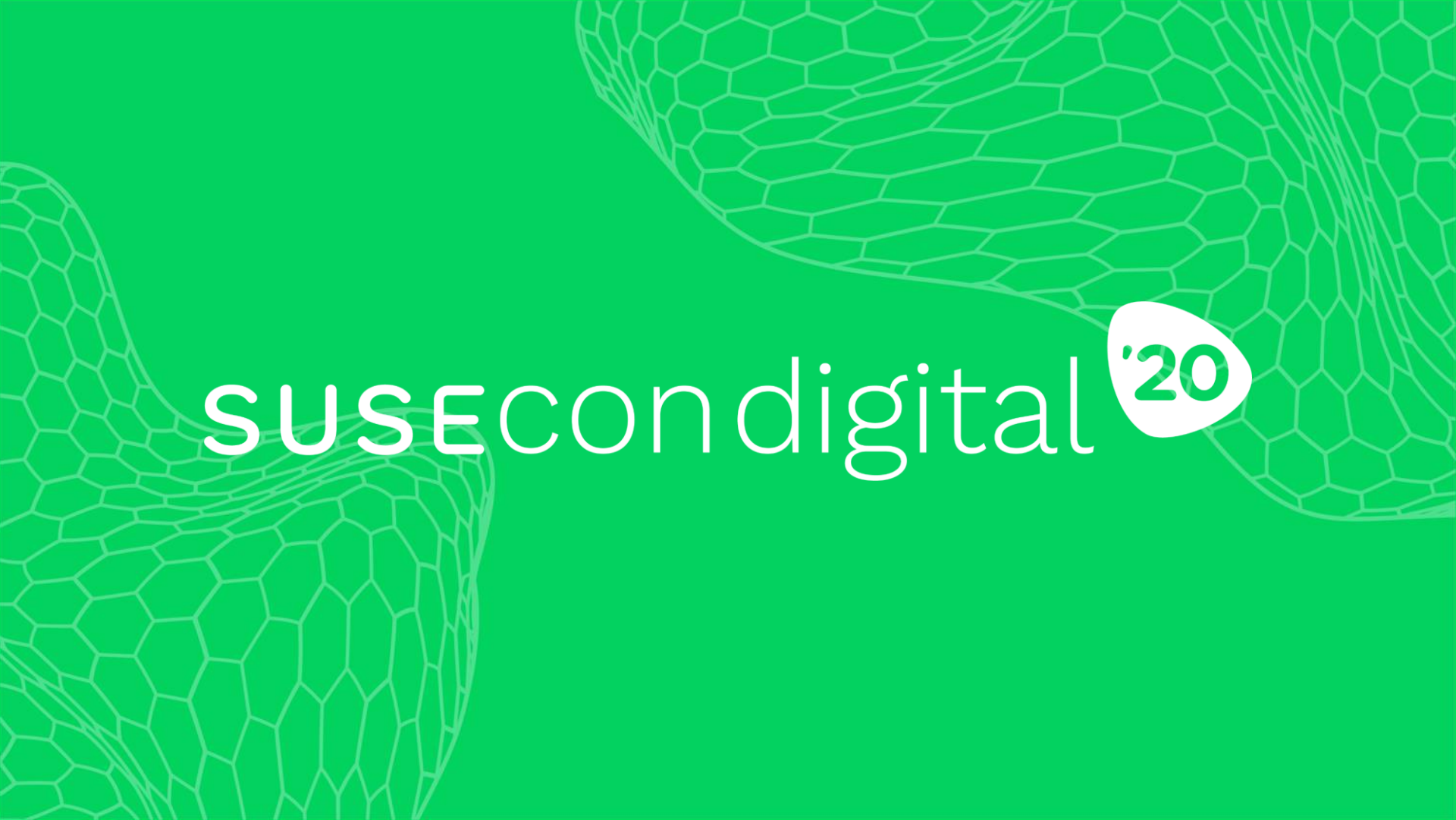
Tune the environment appropriately.

Ensure that you can manage the storage for the longer term.

Don’t lock yourself into a technical nightmare.

General Disclaimer

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All SUSE marks referenced in this presentation are trademarks or registered trademarks of SUSE, LLC, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.


SUSEcon digital '20