SUSEcon digital '20

HPC Performance Tuning

BP-1123

David Byte

Sr. Technology Strategist, SUSE

SUSE

Agenda

1. Define Performance
2. How to Tune
3. Tools
4. Some Specific Items

What Is Performance?

Storage

- Individual or aggregate
- Latency (quickness of response)
  - IOPS
- Throughput
  - MB/s

Memory

- Bandwidth/core

Compute

- IPC
- FLOPs

Network

- Latency
- Throughput

The Process Of Tuning

How To Tune

Define the goal

What to measure and how

Incremental in nature

- Change one thing
- Measure
- Repeat

Document

Tools

Tools

FIO – simulate i/o patterns

Iostat – storage device i/o

HPL - Linpack TPP for floating point

SREAM – sustainable memory bandwidth and computation rate for simple vector kernel

Lower Level Tools

vmstat

top

iotop

iftop

numatop

cpupower

Specifics

Starting Point

SUSE Linux Enterprise for High Performance Computing

Updates

Toolchain Module

How To Tune - Start With Hardware

- CPU
  - Hardware performance bias
  - Memory Layout
- Disk
  - Caching Disk Controller
  - NVMe
- Network
  - Jumbo Frames
  - Bifurcated IB
- Firmware

How To Tune - OS

Look for network issues (buffers, soft IRQ issues, etc)

Can the PCIe read buffer be tuned

Security mitigations

Multi-queue block I/O

Nagle’s algorithm

Specific Things To Check - NUMA

NUMA Autotuning

- SUSE Linux Enterprise ships with auto-balance enabled
- Great for non-NUMA aware applications, but most HPC applications & schedulers are aware
- numa_balancing=disable
- https://documentation.suse.com/sles/15-SP1/html/SLES-all/cha-tuning-numactl.html

Specific Things To Check - THP

Transparent Huge Pages

- SUSE® default is enabled
  - May need to change to madvise or disable
- Check with ISV (if applicable)
- Does the software make use of madvise (we hope so)
- Benchmark

Specific Things To Check - *FED

OFED or MOFED

- OFED is upstream and open source
  - Help the community by driving vendors to keep current
- MOFED is vendor (Mellanox) specific

Specific Things To Check - Other

- Make sure to use the latest OS as new features will be enabled there.
- We recommend the toolchain module. Make sure that you specify the correct compiler and path(s)
- Select the right MPI stack:
  - openMPI
  - MPICH
  - MVAPICH
  - Vendor MPI

Virtual Memory Subsystem

Vm.page-cluster – try increase if your application swaps a lot

Swappiness – generally the less swapping the better, try set to zero

Lots of other tunables, but be careful, you can make things dramatically worse

- https://sysctl-explorer.net/vm/

Tune The Storage

HPC clusters can use NFS, Lustre, CephFS, etc

- Make sure it meets the performance requirements
- NFS can be fine for small clusters
- CephFS provides horizontal scalability with moderate single stream
- Most parallel file systems are very complex

Wrap Up

Conclusion

Lots of tools to use

Lots of tunables

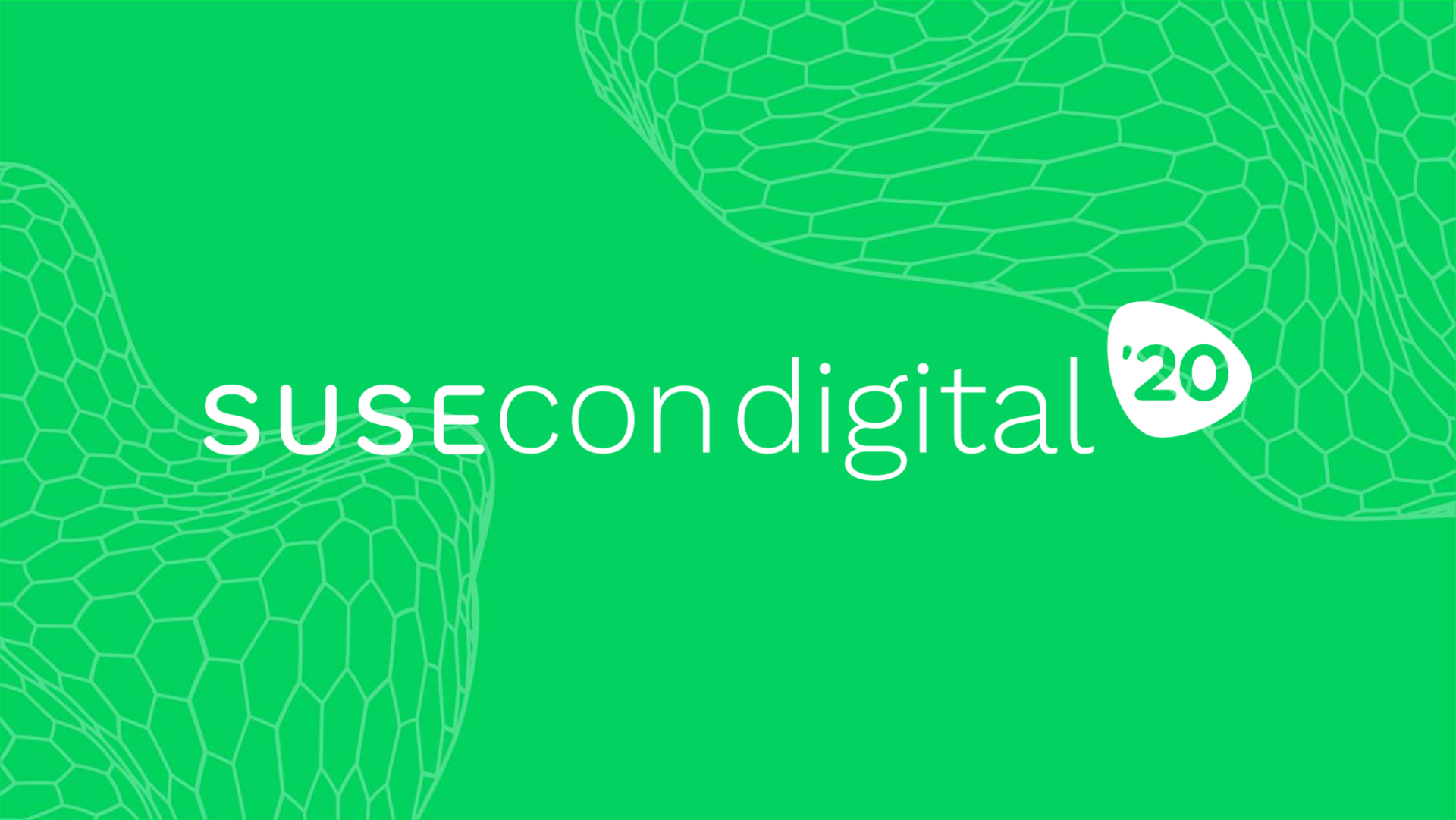
Be systematic in the approach

Measure, measure, measure

Be wary of synthetic benchmarks, they may not represent your environment

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