SUSEcon digital '20

Optimizing Ceph Deployments For High Performance

BP-1072

David Byte, Sr. Technology Strategist, SUSE

Jason Sparks, Sr. Systems Architect, SUSE

SUSE

Agenda

1. Define Performance
2. Process and Tools
3. OS
4. Ceph Cluster
5. Clients

What Is Performance?

For storage

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

What Is An IOP?

I/O per second

- What size?
- Read/Write?

Single Stream Or Aggregate

- 10GB/s
- Single Stream = 1 user needing 10 GB/s
- Aggregate = 100 users needing 100 MB/s

- Ceph is designed for aggregate throughput

Process And Tools

Process Characteristics

Define the goal

What to measure and how

Incremental in nature

- Change one thing
- Measure
- Repeat

Document

Tools

- fio
- iostat
- ceph daemon perf dump
- iotop
- iftop
- iperf3

Pick The Right Hardware

- CPU options
- Network
- Storage bus
- Node count

How To Tune - Start With Base Server Hardware

- CPU
  - Disable C-States?
  - Performance impacting mitigations
  - Hardware performance bias
- Disk
  - Check both latency constrained and unconstrained
  - Write-cache on/off
  - 4k, 8k, 64k, 256k, & 1M
  - Individual and all at once
- Network
  - Jumbo Frames
  - Buffers, kernel network tunables, etc

How To Tune – Then The Cluster

Core OSD performance

- Use krbd and ceph tell osd bench
- Watch iostat on nodes for distribution. Uneven may indicate too few pgs or a slow device
- Watch ceph counters (ceph daemon perf dump)
- Do single stream and aggregate for scaling check

Tuning The OS

There Is A Lot To Gain

In lab testing, OS tuning doubled (or better) performance in some situations

There are lots of places to tweak the OS:

- IO schedulers
- Network buffers
- TCP congestion
- Multi-queue
- NIC and HBA drivers
- etc

Tuning Script Used During Some Testing

```
#!/bin/bash
salt 'sr6*' cmd.run 'setpci -s 5b:00.0 68.w=5936'
salt 'sr6*' cmd.run 'setpci -s 5b:00.1 68.w=5936'
salt '*' cmd.run 'ip link set bond0 mtu 9000'

salt '*' cmd.run 'for j in `cat /sys/class/net/bond0/bonding/slaves`;do LOCAL_CPUS=`cat /sys/class/net/$j/device/local_cpus`;echo $LOCAL_CPUS >
/sys/class/net/$j/queues/rx-0/rps_cpus;done'
salt '*' cmd.run 'ethtool -G eth4 rx 8192 tx 8192'
salt '*' cmd.run 'ethtool -G eth5 rx 8192 tx 8192'
salt '*' cmd.run 'sysctl -w net.ipv4.tcp_timestamps=1'
salt '*' cmd.run 'sysctl -w net.ipv4.tcp_sack=1'
salt '*' cmd.run 'sysctl -w net.core.rmem_max=2147483647'
salt '*' cmd.run 'sysctl -w net.core.wmem_max=2147483647'
salt '*' cmd.run 'sysctl -w net.core.somaxconn=2048'
salt '*' cmd.run 'sysctl -w net.ipv4.tcp_low_latency=1'
salt '*' cmd.run 'sysctl -w net.ipv4.tcp_fastopen=1'
salt '*' cmd.run 'sysctl -w net.ipv4.tcp_rmem="10240 87380 2147483647"'
salt '*' cmd.run 'sysctl -w net.ipv4.tcp_wmem="10240 87380 2147483647"'
salt '*' cmd.run 'sysctl -w net.core.netdev_max_backlog=250000'
#salt '*' cmd.run 'mlnx_tune -p HIGH_THROUGHPUT'
salt '*' cmd.run 'echo never > /sys/kernel/mm/transparent_hugepage/enabled'

salt 'sr650-*' cmd.run 'for i in `lsblk -oname`;do hdparm -W 1 /dev/$i;done'
salt 'sr650-*' cmd.run 'for i in `lsblk -oname`;do hdparm -A 1 /dev/$i;done'
#salt 'sr650-*' cmd.run 'for i in `lsblk -o name,model|grep SSDSC|cut -f1 -d" "`;do echo 4096 >/sys/block/$i/queue/read_ahead_kb;done'
#salt 'sr650-*' cmd.run 'for i in `lsblk -o name,model|grep SSDSC|cut -f1 -d" "`;do echo 512 >/sys/block/$i/queue/nr_requests;done'
salt 'loadgen*' cmd.run 'ethtool -A eth0 rx on;ethtool -A eth1 rx on'
salt 'loadgen*' cmd.run 'ethtool -G eth0 rx 4078 tx 4078;ethtool -G eth1 rx 4078 tx 4078'
for i in `cat osdnodes.lst`;do ssh root@$i 'for i in `ls /sys/block|grep -v "dm-"`;do echo "none" >/sys/block/$i/queue/scheduler;done';done
```

Salt State Used To Apply Kernel Tuning

```
dmb_kern_tune:
  file.replace:
    - name: /etc/default/grub
    - pattern: quiet.*
    - repl: splash=silent intel_idle.max_cstate=0 processor.max_cstate=0
idle=poll scsi_mod.use_blk_mq=1 mitigations=off"

grub2_mkconfig:
  cmd.run:
    - runas: root
    - name: grub2-mkconfig -o /boot/grub2/grub.cfg
    - onchanges:
      - file: dmb_kern_tune
```

Tuning The Cluster

Things To Try

- Increase OSD RAM allocation
- Disable logging
- Tune RGW
- Increase MDS Cache RAM
- In some situations, disable cephx authentication
- Erasure Coding algorithms

Ceph.conf General Section

```
debug ms=0
debug mds=0
debug osd=0
debug optracker=0
debug auth=0
debug asok=0
debug bluestore=0
debug bluefs=0
debug bdev=0
debug kstore=0
debug rocksdb=0
debug eventtrace=0
debug default=0
debug rados=0
debug client=0
debug perfcounter=0
debug finisher=0
```

Ceph.conf OSD Section

```
[osd]

osd_memory_target = 6442450944

osd_memory_cache_min = 4294967296

bluestore_min_alloc_size = 4096

osd_op_num_threads_per_shard = 5
```

MDS

Multiple Active MDS

Manual Pinning

```
ceph.conf
[mds]
mds_cache_memory_limit=17179869184 #16GB MDS Cache

[client]
client cache size = 16384 #16k objects is default number of inodes in
cache
client oc max objects =10000#1000 default
client oc size = 209715200 #200MB default, can increase
client permissions = true #default true, disable in limited exposure
environment
```

RGW

SUSE® Enterprise Storage Administration Guide Section 16.3

```
rgw thread pool size = 512 #default 100, increase for heavier load
rgw max chunk size = 4194304# default is 128k, increase to 4MB helps with larger objects
rgw_obj_stripe_size = 4194304 # (default 4M for luminous)
rgw_list_bucket_min_readahead = 4000 #(default 1000)
rgw_max_listing_results = 4000
rgw_cache_expiry_interval = 1800 #(default 900s)
rgw_enable_usage_log = false
rgw_enable_ops_log = false
rgw dynamic resharding = false
rgw override bucket index max shards = 50 # alternatively we reshard the bucket manually after
creation
rgw bucket index max aio = 16 # default 8
rgw cache lru size = 50000 #number of entries in the cache
rgw_gc_obj_min_wait = 21600 #in seconds(default 2_hr), decreasing will more actively purge objects
rgw gc processor period = 7200 #(default 1hr, decreasing more actively purges deletion)
rgw objexp gc interval = 3600 # default 10_min, we dont run swift objexp. so no need to run this
objecter inflight op bytes = 1073741824 # default 100_M, this is 1GB
objecter inflight ops = 24576
```

Cache Tiering

SSD/NVME in-front of spinners

- Ensure the working set fits within tier
- Good for workloads like media editing

SUSE Enterprise Storage Administration Guide section 14

Client Tuning

Client Stuff

Bad hardware choices can’t be fixed by software

Performance Thieves:

- Anti-virus
- Resource Competition
- Routed Connections
- Power saving tech

RBD

- Client-side write caching (write-back, write-through)
- Client-side read-ahead caching
- SUSE® Enterprise Storage Admin Guide is a good reference
- https://documentation.suse.com/ses/6/single-html/ses-admin/#rbd-cache-settings
- multi-queue block I/O

CephFS

There are the settings we did on the cluster side that affect clients

Mount options – understand the implications first

- noacl
- nocrc
- noatime
- nodiratime

Wrapping Up

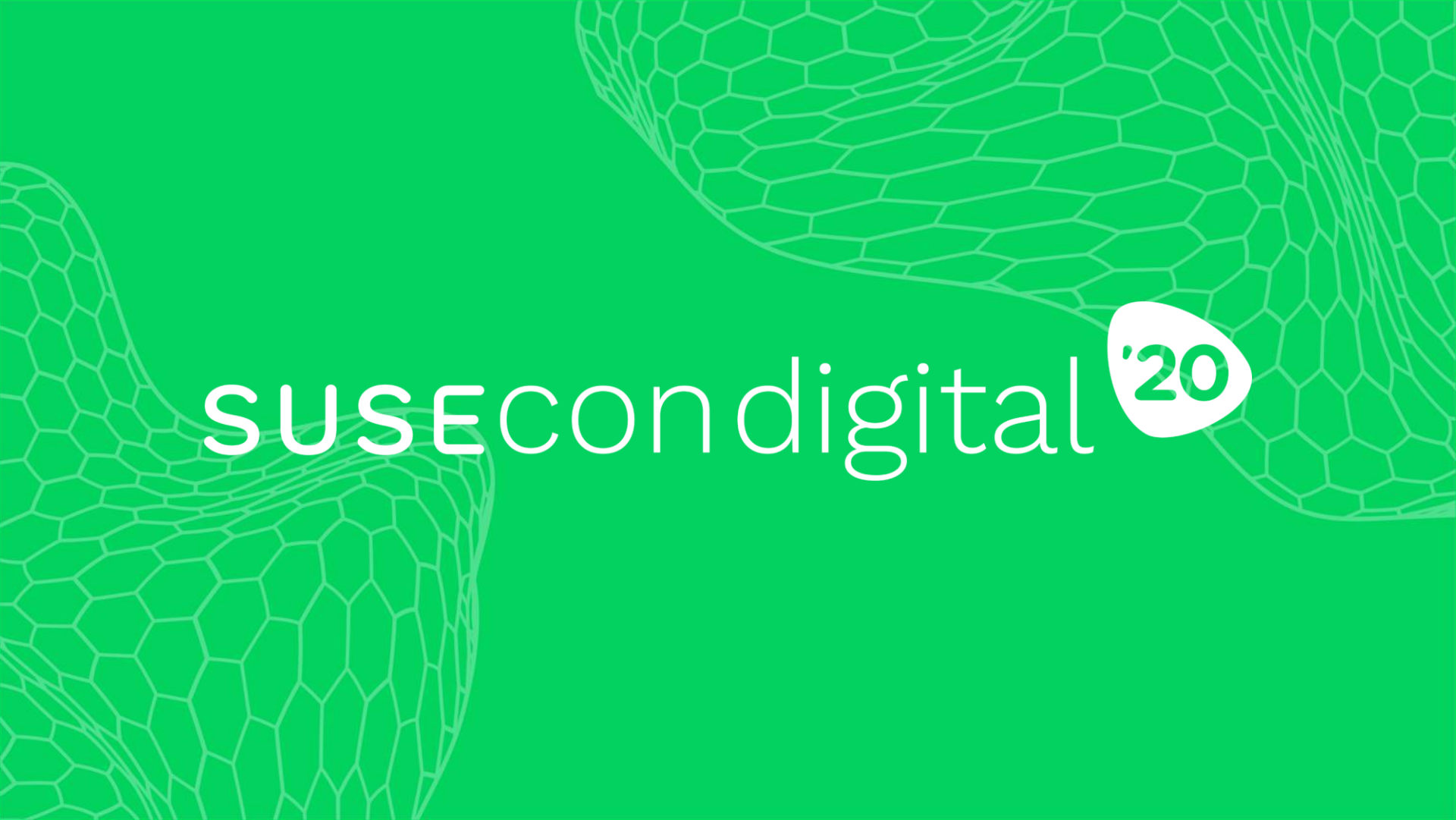
Many Knobs

There are lots of things to watch and many knobs to twist

- Be methodical
- Record your results
- Share the knowledge

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