True, Unbiased Story to Affordable, Reliable and FAST Ceph on NVMe

Wido den Hollander, 42on
Piotr Dałek, OVH
Agenda

- Is Ceph a right tool for you?
- Plan ahead
- Defining KPIs
- Talking with your vendor
- Benchmarking
- Tweaking
- Sample systems
Wido den Hollander

- Netherlands
- Working with Ceph since 2010
- Ceph Trainer and consultant
- 42on.com

Piotr Dałek

- Poland
- Working with Ceph since 2014
- Software developer, local Ceph consultant
- OVH

The speakers
Is Ceph a right tool for you?
From the performance perspective

THE BAD

• Absurdly high CPU overhead
  • RocksDB
  • Java-like memory management
• Large network overhead
• Lots of tunables, many poorly documented or changing behavior between versions
• Lots of things you can break by accident
• Sensitivity to hardware
  – One bad RAM stick can ruin your performance
  – One I/O error can kill OSD
  – Doesn’t work with consumer SSD/NVMe

THE GOOD

• Automatic replication
• Data safety is a very high priority
• Self-healing (not invulnerability)
• Easy up/downscaling
• Fail-fast approach
  – If something is broken, you’ll get notified fast
• Not tied to particular HW vendor
  • Mixing Intel NVMe with Memblaze NVMe and AMD CPUs is an option!
Plan ahead

• Take your time, don’t make rushed decisions (super important!)
• Consider all options, even those that seem unrealistic
• What’s expected load?
• What’s expected capacity?
• What’s expected performance (you’ll get it wrong anyway)?
• Talk to vendors
Defining KPIs
Some example KPIs you can adopt

• IOPS
• Average latency, 99th percentile, 99.99% percentile
• Bandwidth
• Clients served simultaneously with acceptable performance
• Recovery impact
• SLA
• Usable capacity
Defining KPI
What can go wrong and why

• You’ll overestimate performance
  – We picked 240k randrw IOPs. We got 200k at best because we were CPU bound.

• You’ll underestimate clients/customers/internal users
  – „WOO! NVMe storage! Let’s benchmark it and see if it’s real NVMe! :D”
  – „What do you mean – we can’t take and remove snapshot every 15 seconds? Is that a joke?”

• Managers won’t listen. They never listen.
  – „I know 10 NVMe devices is way too much, but we need capacity”

• You skip on testing locally before going to prod
  – „Vendor did this for us, they can’t be wrong or much off”

• You get a lot of IOPS, but clients complain
  – You forgot about latencies!
Talking to vendors

- State your needs clearly
  - IOPS, latency, capacity, node count, etc
  - Price!
- Talk to few vendors in parallel, never stick to one
- Bring them back from Neverland
- Verify claims
- Verify claims
- Verify claims
- Verify claims (repeating 4th time just to be sure you remember)
Bring them back from Neverland
This is PROD, not vendor showroom

- Replica factor of 3
  - “NVMe drives are more reliable than HDD, no need for 3x”
  - 2x replication is very dangerous, even with NVMe. Don’t do it. Don’t.
- Insane Ceph configuration
  - Reliability and component wearout is irrelevant
- Adding SPOFs with caching drives
  - Little performance gains at greater cost and risks
- You don’t need 100Gbit of bandwidth per node
  - Latency is what matters! Not bandwidth. Spend you money on something else
Bring them back from Neverland

Replica count

- Vendors often use Replica factor of 2, equivalent of RAID-1 with two disks
- We had a power outage that irreversibly killed two of our NVMe drives
- They were hosting two replicas of the same data group
- We weren’t screwed only because we had last (third) replica left
- If you don’t need 3 replicas, do you need any replicas at all?
- If you don’t need replicas, do you need Ceph at all?
Bring them back from Neverland

Replica count

IOPS

avglatency [ms]
Bring them back from Neverland
Insane Ceph configuration

• Don’t copy vendor config, use what you really need and only if you understand what and why you need
• If you don’t know or understand something, DON’T USE IT EVEN IF ASKED BY VENDOR
• Vendors care about performance, not reliability or security
  – Excessively large buffers/caches
  – Tons of worker threads
  – Replication 2x
  – Disabled throttles
  – Disabling reliability-assuring functionality (CRC)
  – Disabling CephX
  – Favoring NVMe wearout over CPU utilization
  – Soon: disabling on-wire encryption
  – Bonus: nonexistant options taken from vendor’s development branches
Bring them back from Neverland

Caching drive – what can go wrong?

- It makes sense when in front of HDD
- Somewhat reasonable with SATA/SAS SSD
  - Limited max BW – just ~550MB/s per device
- Dubious when in front of NVMe
- Failure of caching drive kills all OSD behind
  - Tons of recovery!
- Failure of single drive can trash caches, reducing performance for ALL drives
Verify claims
Why this is important

• We asked vendor to provide performance data of particular CPU
• They provided performance of other CPU with few cores disabled to align with the CPU we asked
• We re-did tests and got totally different – worse - results!
  – Different TDP budget and in-CPU caches affected results
  – We weren’t able to get to their level of performance no matter what
  – We asked them again to re-do it on the CPU we asked in the first place
  – They achieved worse results than us!
• Vendors usually benchmark with the most expensive hardware
  – For example, Intel uses Xeon Platinum CPUs in their benchmarks which are over 10k USD/EUR per CPU. Not realistic for most deployments
Verify claims
Why this is important

- Vendor claimed 1048576x performance upgrade when used ExpensiveDrive 9900FTW as metadata storage
- That’s what we got: (load/qd/images). 8% perf increase at best, in single case.
Verify claims
Benchmark components

- We got very unstable IO performance
- No clue why, nothing obvious
- Direct FIO with QD=1 on drive revealed sad truth
- Eventually we switched to other drive
Benchmarking
What to use

- **FIO**
  - Very flexible and very complex

- **CBT**

- **Rados bench for smoke testing only**
  - Rados bench operates on entire objects, which adds additional latency and skews results
  - Rados bench results are unreliable on large writes/reads
  - Rados bench operates on AIOs, meaning it gets bottlenecked by CPU fast and you can’t workaround it
  - You can run only single rados bench per pool on single host
  - But it’s great to verify whether your cluster is operating at more-or-less expected performance, or what performance expectations you should have
Benchmarking

How to use

• One run at least 10 minutes, some warmup recommended
• Repeat at least 4 times, calculate average of last 3 runs (discard warm-up run)
  – Because tcmalloc sucks
• Try different op sizes, queue depths, client count
• Use several client **machines**, not just clients
  – They easily get overloaded
• Use single image for single client (exclusive lock kills performance)
• Always precondition (fill with data) RBD images
• Make sure your networking can handle the load
Benchmarking
That’s how 539K IOPS look from 5 client machines perspective

Also, that’s why you need several client machines for serious perf testing.
Tweaking
Getting more IOPS from your existing hardware

• MTU 9000
  – Less IRQs, less packets, less TCP overhead
• Disable powersaving / Enable C-State 0 pinning (kernel parameter)
• Use decent kernel
  – AMD Epyc works best with 4.15 or newer, with 4.10 – not so well
• Make sure your RAM configuration is correct
  – 2x less IOPS and 5x worse latency when RAM sticks were improperly placed!
• Make sure your BIOS/UEFI is set up correctly
• Did you enable “Above 4G decoding” and “NUMA mode enable”?
• Disable Spectre/Meltdown/Foreshadow/SPOILER/TLBleed/Zombieload/RIDL/Fallout mitigations on Intel platforms if feasible
• Disable Spectre mitigations on all platforms if feasible
HT/SMT impact
AMD 7551p (32 cores, 64 threads)

Avglat perf diff
IOPS perf diff

qd16 perf diff
qd64 perf diff
Tweaking

„logging”

• Disable logging for various subsystems
  – debug_osd=0
  – debug_bluestore=0
  – debug_bluefs=0
  – debug_ms=0
  – debug_auth=0
  – debug_rocksdb=0

• This saves the CPU spending time on generating loglines and reduces latency
Tweaking
„osd op threads“

- `osd_op_num_threads_per_shard * osd_op_num_shards`
- Keep at `number_of_threads_of_your_cpu_can_handle – async_msgr_op_threads – 3..5`
  - For example, „osd op num shards = 8“, „osd op num threads per shard = 2“ and „ms async op threads = 3“ for 22-core CPU with HT/SMT (2*8+3 = 19, 3 threads left for Bluestore, RocksDB, etc.)
- Increase in case of slower NVMe to improve IOPS and latency in random reads/writes
  - Offset NVMe processing time (iowait) by using other thread to do CPU-consuming work in the meantime
- Don’t set too high, or context switches will kill your performance
- Too low values will cause your OSDs to stall
- Change requires OSD restart
Tweaking

„osd op threads“ – note the 16t curve vs 20t (default)
Tweaking
„ms async op threads“

- Async messenger threads process all incoming and outgoing data
- Default number (3) is a safe bet
- Increase on slower CPUs to reduce latency
- Decrease to reduce context switch impact
- Too low and too high values impact performance
- Change requires OSD restart
Tweaking

„ms async op threads“

IOPS qd64

12at  6at  5at  4at  3at  2at  1at

12  1
Tweaking

"ms async op threads"

avglat_avg qd64[msec]

12at
6at
5at
4at
3at
2at
1at

1
12
Tweaking
Use Mimic if possible

• Luminous is LTS
• Mimic is not
• Nautilus is new, too new
• Mimic contains a lot of perf improvements vs Luminous
  – For example, Op tracker overhead is much lower – up to 5% performance lost vs much more on Luminous
  – CephX is also much faster now
  – Impact of disk log writes is reduced, though still large
Tweaking

Luminous IOPS disadvantage vs Mimic

-28%
-26%
-26%
-12%
-19%
-21%
-15%
0%
CPU frequency trap
MOAR GIGAHERTZ!!!111!!11oneoneoneeleven

![CPU frequency diagram with processor GHz on the x-axis and Passmark CPU score single thread on the y-axis. Data (c) cpubenchmark.net](image-url)
CPU frequency trap
MOAR GIGAHERTZ!!!111!!11oneoneoneeleven

- CPU performance depends on
  - Architecture
  - Cache sizes and levels
  - Clockspeed
  - Thermal budget
  - Turbo speed
  - Software

Single thread score, 2,8GHz CPUs

- Netburst: 632
- Wolfdale: 1143
- Nehalem/Lynnfield: 1231
- Sandy Bridge (i7): 1807
- Skylake (i5): 1899
- Coffe Lake (i5): 2460

Data (c) cpubenchmark.net
Choosing components
What to look for, and why

• Fast CPU reduces op latency
  – Faster write path execution
• Many cores improve cluster op capacity (more parallel clients)
  – More clients can be served before cluster performance breaks down
  – Reduced impact of recovery/backfill/snaptrim
• Fast but small NVMe might be slower than slow and large NVMe
  – Less flash chips = less parallelism
  – Less flash chips = less TBW

• Ceph’s write path is CPU-bound, so getting fast NVMe to improve writes won’t work as well as getting fast CPU
  – Offset this by putting more than 1 OSD on one NVMe
  – Consider increasing Placement Groups
• Get a powerful switch
  – Client that does 5Gbit/s writes will require 15Gbit/s of switching performance (3x replication)
• Get plenty of fast memory
  – Bluestore has its own, private disk cache that’s not shared thorough entire machine
  – 12GB per OSD is a safe bet (fast memory!)
Sample systems
Internet prices in EUR for components that aren’t common, not including any discounts

<table>
<thead>
<tr>
<th>Component</th>
<th>Intel:</th>
<th>Intel 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>QCT D52B-1U</td>
<td>1881</td>
<td>1881</td>
</tr>
<tr>
<td>Xeon 6152</td>
<td>3797,42</td>
<td>3797,42 x2</td>
</tr>
<tr>
<td>Intel dc p4510 8TB</td>
<td>2058,81 x3</td>
<td>2058,81 x3</td>
</tr>
<tr>
<td>Total per host</td>
<td>11854,85</td>
<td>15652,27</td>
</tr>
<tr>
<td>Total per cluster</td>
<td>35564,55</td>
<td>46956,81</td>
</tr>
</tbody>
</table>
Sample systems
Internet prices in EUR for components that aren’t common, not including any discounts

- **AMD:**
  - TYAN GT62FB8026: 990
  - Epyc 7551p: 2189.72
  - Intel dc p4510 2tb: 413.15 x6
  - Total per host: 5658.62
  - Total per cluster: 16975.86

- **AMD 2:**
  - TYAN GT62FB8026: 990
  - Epyc 7551p: 2189.72
  - Intel dc p4510 4tb: 756.53 x6
  - Total per host: 7718.9
  - Total per cluster: 23156.7
SuperMicro Sample systems
Sample pricing for SuperMicro systems known by working as consultant

• Intel:
  – SuperMicro 1029U-TN10RT
  – 2x Intel Xeon Silver 4112
  – 96GB Memory
  – 10x Samsung PM983 3.84TB
  – Total per host 9450

• AMD:
  – SuperMicro AS-1113S-WN10RT
  – AMD Epyc 7351P
  – 128GB Memory
  – 10x Samsung PM983 3.84TB
  – Total per host 8525
Sample systems

IOPS

At qd64, AMD 1 overtook Intel 1

At qd16, AMD 1 overtook Intel 1
Sample systems

Latency

 Avg latency qd16 [ms]

 Avg latency qd64 [ms]

Intel 1 qd16  Intel 2 qd16  AMD 1 qd16

Intel 1 qd64  Intel 2 qd64  AMD 1 qd64