

Revisiting Ceph's Performance After 5 Years



04-12-2024
Public



ceph

Agenda

Agenda	
01	Who is who?
02	What is performance?
03	Benchmarking setup
04	L3 routing with BGP
05	The results

01 Who is who?

Wido den Hollander

Who am I?

- Wido den Hollander (1986)
- Born and live in the Netherlands
- Married, two sons (2020 and 2022)
- CTO @ Your.Online
 - Strategic CTO, oversee our companies. No daily involvement
 - Started my own hosting company in 2003
 - Techie in my heart
- Open Source & Tech
 - Ceph evangelist since 2008
 - Started to work with Ceph *before* version Argonaut
 - Founded **42on** and *used* to be Ceph Trainer & Consultant
 - Apache CloudStack developer and PMC member
 - IPv6 *fanatic*

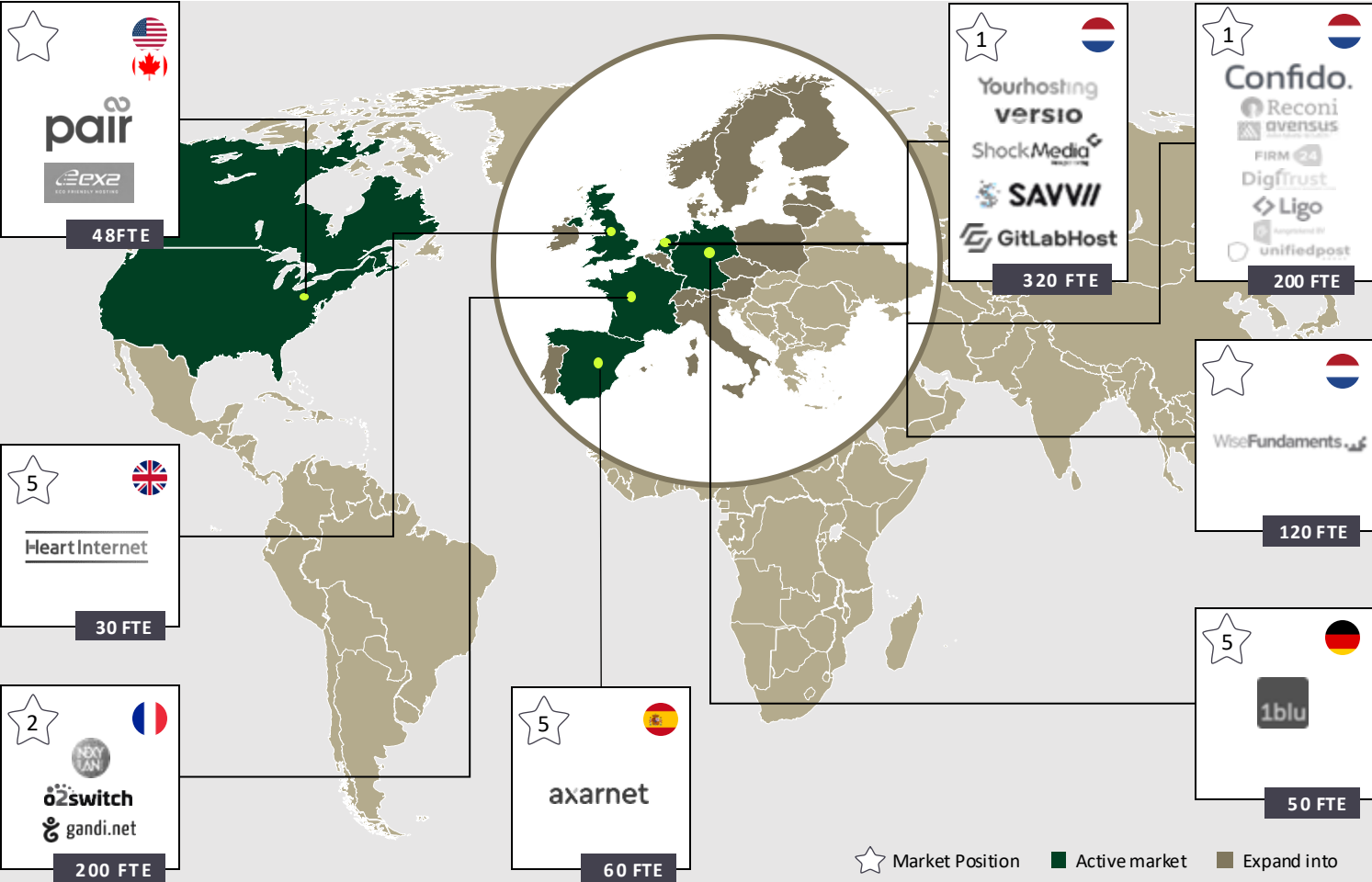


Who is Your.Online?

Your.Online is a global group specializing in **Online Presence**. Operating across Europe and North America, our companies provide a range of online services, enabling businesses and individuals to establish and grow their digital footprint.

Our portfolio includes services such as Domain Registration, Web Hosting, VPS, Dedicated Servers, and much more.

A significant number of our companies use Ceph as a component of their infrastructure, supporting deployments that range from small-scale setups to **multi-petabyte** environments.



02 What is performance?

Back in time

Let's go back to 2019

2019

- At Cephalocon 2019 I gave a presentation together with Piotr Dałek who worked at OVH at the time
- We tested with Ceph Luminous, Mimic and Nautilus back then
- Lots of things have changed since 2019, but not everything



```
Jobs: 1 (f=1): [w(1)][100.0%][r=0KiB/s,w=5348KiB/s][r=0,w=1337 IOPS][eta 00m:00s]
rbd_w_iodepth_1: (groupid=0, jobs=1): err= 0: pid=3059796: Tue Jul  2 09:32:23 2019
write: IOPS=1292, BW=5171KiB/s (5295kB/s)(303MiB/60001msec)
  slat (usec): min=2, max=115, avg= 7.33, stdev= 1.14
  clat (usec): min=545, max=48271, avg=765.03, stdev=422.48
    lat (usec): min=551, max=48278, avg=772.35, stdev=422.55
  clat percentiles (usec):
    | 1.00th=[ 652],  5.00th=[ 676], 10.00th=[ 685], 20.00th=[ 701],
    | 30.00th=[ 717], 40.00th=[ 725], 50.00th=[ 734], 60.00th=[ 742],
    | 70.00th=[ 750], 80.00th=[ 766], 90.00th=[ 791], 95.00th=[ 824],
    | 99.00th=[ 1745], 99.50th=[ 2999], 99.90th=[ 4752], 99.95th=[ 5211],
    | 99.99th=[11600]
  bw (  KiB/s): min= 3064, max= 5432, per=100.00%, avg=5171.29, stdev=272.57, samples=120
  iops       : min=  766, max= 1358, avg=1292.80, stdev=68.13, samples=120
  lat (usec)  : 750=68.22%, 1000=30.23%
  lat (msec)  : 2=0.67%, 4=0.64%, 10=0.23%, 20=0.01%, 50=0.01%
  cpu         : usr=32.71%, sys=66.81%, ctx=77832, majf=0, minf=8420
  IO depths   : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
    submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
    complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
    issued rw: total=0,77563,0, short=0,0,0, dropped=0,0,0
    latency   : target=0, window=0, percentile=100.00%, depth=1
```

0.772 ms

1,292 IOps

Conclusion in 2019

2019

- $\sim < 1\text{ms}$ write latency
 - Single thread
 - 4k object
 - 3x replication
 - 1000 – 1200 IOps
 - 10Gbit Base-T network (IPv6, MTU 9000, Juniper QFX5100 switches)
- Do not expect full NVMe performance
 - CPU and code are bottleneck
 - Consistency is #1 priority! not performance

Performance

queue depth = 1, block size = 4k

- I think that the single thread latency is the most important
- Many applications benefit from a low single thread latency
- Achieving a low latency is not easy nor cheap
- My background is from the webhosting industry
 - I am biased towards certain use-cases and applications



Performance

But what is performance?

- In the end it's up to you what works for your situation!
- In my case that's single thread low latency
- Always verify what your user experience is
 - Is the application running on top performing as expected?
- Is 200km/h sufficient while 400km/h just sounds cool?



Do you **need** one or do you **want** one?

03 Benchmarking setup

Hardware

Dell R6615 (3x)

- AMD Epyc 9124 16C/32T
 - Base: 3Ghz
 - Boost: 3.6Ghz
- 128GB DDR5 Memory
- 2x Samsung PM9A3 MZQL23T8HCLS-00W07 3.84TB
- Separate NVMe for Boot/OS
- Mellanox ConnectX-5 2x25Gb SFP28

Benchmarking performed in the summer of 2024



Special thanks to **Ynvoke**
for providing the hardware

Ynvoke, based in *the Netherlands*, is a *global* circular systems integrator providing sustainable IT lifecycle solutions focused on enterprise hardware, refurbishment, and tailored support.

www.ynvolve.com

Benchmark setup

Software

Ubuntu Linux 22.04

- Kernel 5.15.0-94-generic
- Ceph 18.2.4
- Fio as a benchmarking tool
 - runtime=60
 - rw=randwrite
 - bs=4k
 - iodepth=1
 - pool=rbd
 - rbdname=fio_test



Benchmarking performed in the summer of 2024

Fio

queue depth = 1
block size = 4k



Ceph & Tuning

- Ceph config is almost default
 - Don't tune because of tuning!
 - 3x replication
- CPU tuning
 - kernel option 'processor.max_cstate=1' to pin C-State
 - Set CPU governor to 'performance' (cpufrequtils)



Ceph Health

```
[root@osd-138-c13-34:~# ceph -s
cluster:
  id:      24b1c9b4-4b51-11ef-9911-c916cce9e9d4
  health: HEALTH_OK

services:
  mon: 3 daemons, quorum osd-138-c13-34,osd-138-c13-36,osd-138-c13-38 (age 2w)
  mgr: osd-138-c13-34.jheaoe(active, since 2w), standbys: osd-138-c13-36.wkbopd
  osd: 6 osds: 6 up (since 2w), 6 in (since 2w)

data:
  pools:      2 pools, 129 pgs
  objects:    25.61k objects, 99 GiB
  usage:      123 GiB used, 21 TiB / 21 TiB avail
  pgs:        129 active+clean

[root@osd-138-c13-34:~# █
```

3 nodes, 2 OSDs per node

04 L3 routing with BGP

L3 network

- Dell S5248F-ON 48x25Gb Top-of-Rack
 - Cumulus Linux 4
- Each Ceph node connected with 2x 25GbE
- MTU set is 9216
- BGP towards each Ceph node
 - Each node announces a **/128** IPv6 address
 - Top-of-Rack announces **::/0** (default) downstream
- 100% L3, **no** VLANs!



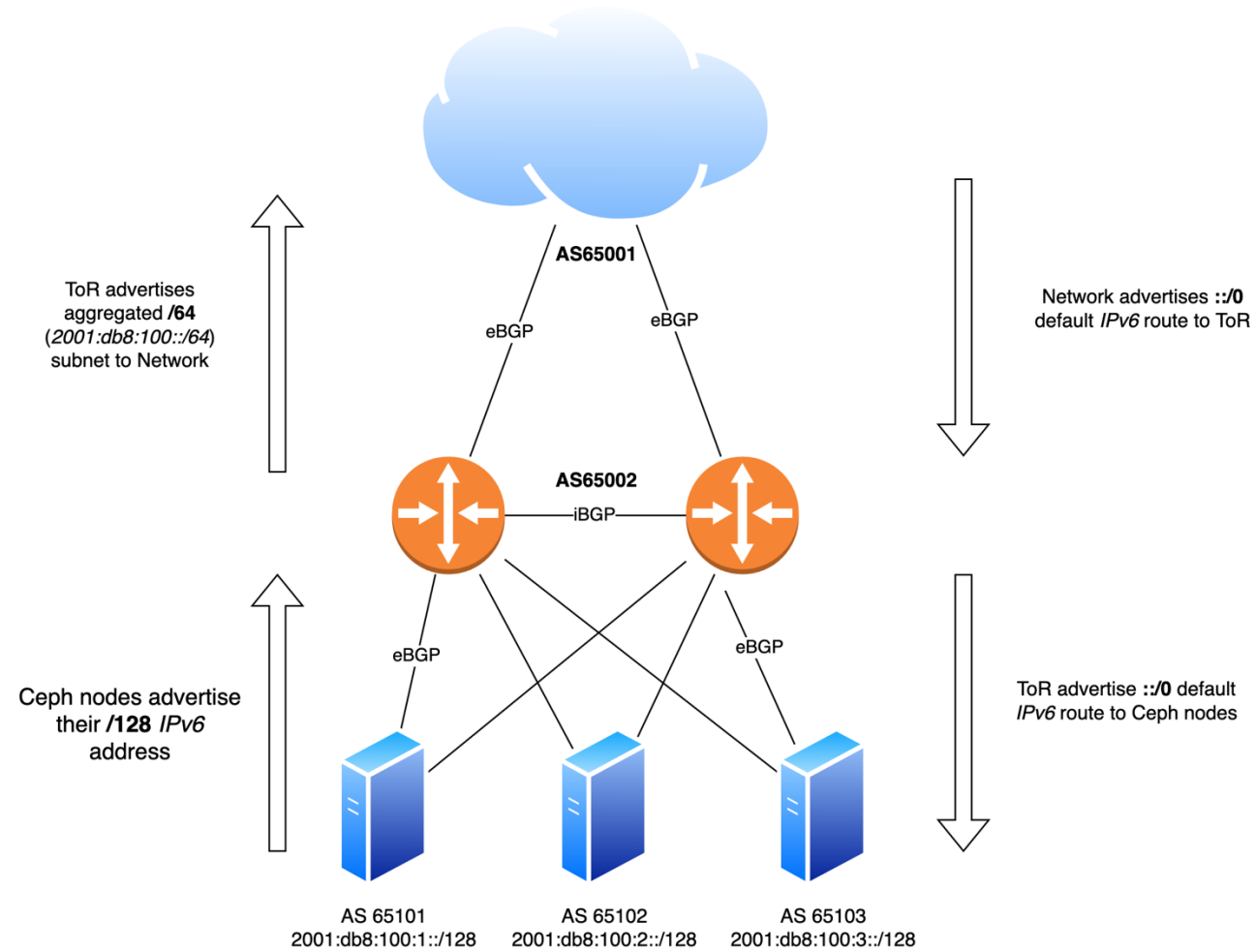
<https://blog.widodh.nl/2024/05/using-l3-bgp-routing-for-your-ceph-storage/>

Why L3?

- All the hyperscalers use it!
- I think using VLANs is something from the past
- Modern networks should be built on Layer 3 routing
 - Using IPv6

Why use “The Future of Storage”, but use “The Past of Networking”?

BGP



BGP sessions between ToR and Ceph nodes are established using **BGP Unnumbered** (RFC5549)

BGP Neighbors

```
ceph01# sh bgp summary
```

IPv6 Unicast Summary (VRF default):

BGP router identifier 1.2.3.4, local AS number 65101 vrf-id 0

BGP table version 10875

RIB entries 511, using 96 KiB of memory

Peers 2, using 1448 KiB of memory

Peer groups 1, using 64 bytes of memory

Neighbor	V	AS	MsgRcvd	MsgSent	TblVer	InQ	OutQ	Up/Down	State/PfxRcd	PfxSnt	Desc
enp196s0f0np0	4	65002	487385	353917	0	0	0	3d18h17m	1	1	N/A
enp196s0f1np1	4	65002	558998	411452	0	0	0	01:38:55	1	1	N/A

Total number of neighbors 2

```
ceph01#
```

BGP Routes

```
ceph01# sh ipv6 route ::/0
```

```
Routing entry for ::/0
```

```
Known via "bgp", distance 20, metric 0
```

```
Last update 01:42:00 ago
```

```
fe80::e29:eff:fed7:4719, via enp196s0f0np0, weight 1
```

```
fe80::7686:e2ff:fe7c:a19e, via enp196s0f1np1, weight 1
```

```
ceph01#
```

/etc/frr/frr.conf (FRRouting daemon)

```
frr defaults traditional
hostname ceph01
log syslog informational
no ip forwarding
no ipv6 forwarding
service integrated-vtysh-config
!
interface enp196s0f0np0
  no ipv6 nd suppress-ra
exit
!
interface enp196s0f1np1
  no ipv6 nd suppress-ra
exit
!
interface lo
  ipv6 address 2001:db8:100:1::/128
exit
!
```

```
router bgp 65101
  bgp router-id 1.2.3.4
  no bgp ebgp-requires-policy
  no bgp default ipv4-unicast
  no bgp network import-check
  neighbor upstream peer-group
  neighbor upstream remote-as external
  neighbor enp196s0f0np0 interface peer-group upstream
  neighbor enp196s0f1np1 interface peer-group upstream
!
  address-family ipv6 unicast
    redistribute connected
    neighbor upstream activate
  exit-address-family
exit
!
end
```

Latency & Performance

```
root@osd-138-c13-34:~# qperf 2a05:1500:ff01:100::38 tcp_lat
```

```
tcp_lat:
```

```
    latency = 13.3 us
```

```
[root@osd-138-c13-34:~# ping -s 8192 -c 3 2a05:1500:ff01:100::38
```

```
PING 2a05:1500:ff01:100::38(2a05:1500:ff01:100::38) 8192 data bytes
```

```
8200 bytes from 2a05:1500:ff01:100::38: icmp_seq=1 ttl=63 time=0.091 ms
```

```
8200 bytes from 2a05:1500:ff01:100::38: icmp_seq=2 ttl=63 time=0.106 ms
```

```
8200 bytes from 2a05:1500:ff01:100::38: icmp_seq=3 ttl=63 time=0.118 ms
```

```
--- 2a05:1500:ff01:100::38 ping statistics ---
```

```
3 packets transmitted, 3 received, 0% packet loss, time 2042ms
```

```
rtt min/avg/max/mdev = 0.091/0.105/0.118/0.011 ms
```

```
root@osd-138-c13-34:~#
```

```
root@osd-138-c13-34:~# iperf -c 2a05:1500:ff01:100::38
```

```
-----  
Client connecting to 2a05:1500:ff01:100::38, TCP port 5001
```

```
TCP window size: 325 KByte (default)  
-----
```

```
[ 1] local 2a05:1500:ff01:100:: port 45962 connected with 2a05:1500:ff01:100::38 port 5001
```

```
[ ID] Interval      Transfer      Bandwidth
```

```
[ 1] 0.0000-10.0075 sec 28.8 GBytes 24.7 Gbits/sec
```

```
root@osd-138-c13-34:~#
```

I perform the 'ping' test with 8k packets to ensure Jumbo Frames work

GOOD

05 The results

Results

Results

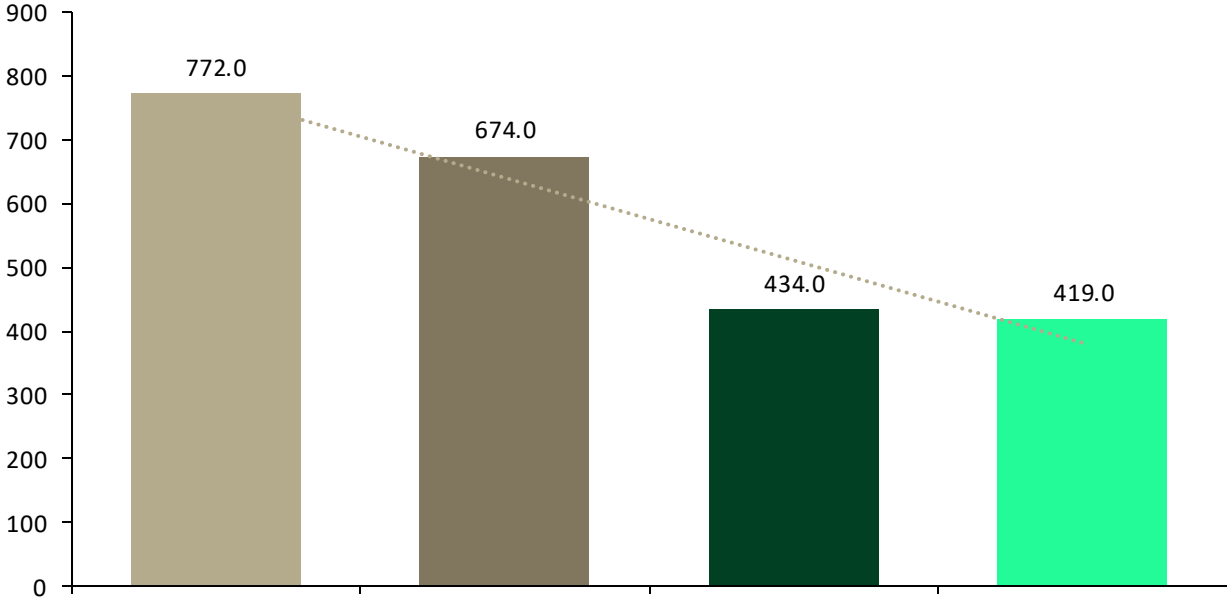
This is what you all came for, right?



Let’s compare quickly

	2019	2024
Ceph version	14.2.2	18.2.4
Operating System	Ubuntu 18.04	Ubuntu 22.04
Linux kernel	4.15	5.15
CPU	AMD Epyc 7351P 16C	AMD Epyc 9124 16C
Storage	Samsung PM983	Samsung PM9A3

Fio benchmark results in microseconds (usec) of latency (qd=1, bs=4k)



	2019	Test 1	Test 2	Test 3
C-State pinning	Yes	No	Yes	Yes
CPU governor	performance	ondemand	performance	performance
Ceph logging	All to 0/0	defaults	default	All to 0/0
Ceph	14.2.2	18.2.4	18.2.4	18.2.4
CPU	Epyc 7351P	Epyc 9124	Epyc 9124	Epyc 9124

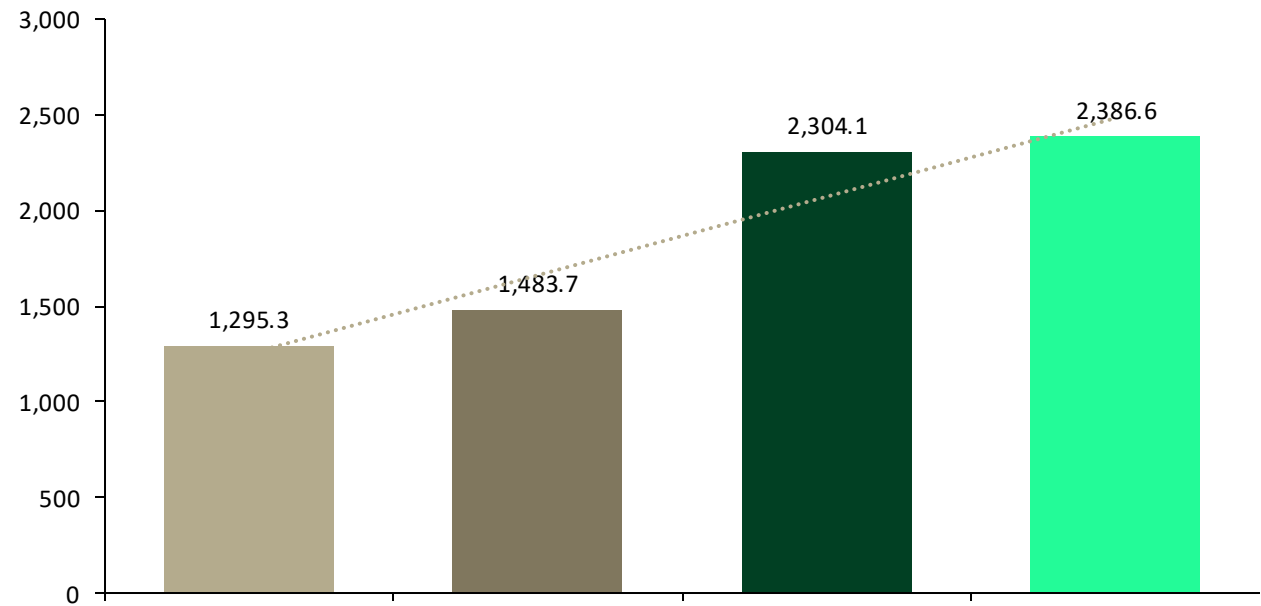
CPU tuning

- This still seems to make the biggest difference
- Newer generation of AMD Epyc has big impact
- Could not properly determine if Ceph had improved or if it was only the CPU

Ceph logging

- Disabling logging makes a *small* difference

Fio benchmark results in I/O operations per second (IOPS) (qd=1, bs=4k)



	2019	Test 1	Test 2	Test 3
C-State pinning	Yes	No	Yes	Yes
CPU governor	performance	ondemand	performance	performance
Ceph logging	All to 0/0	defaults	default	All to 0/0
Ceph	14.2.2	18.2.4	18.2.4	18.2.4
CPU	Epyc 7351P	Epyc 9124	Epyc 9124	Epyc 9124

CPU tuning

- This still seems to make the biggest difference
- Newer generation of AMD Epyc has big impact
- Could not properly determine if Ceph had improved or if it was only the CPU

Ceph logging

- Disabling logging makes a *small* difference

That was it!

Thank you!

Any questions or comments?

@widodh

wido@denhollander.io

blog.widodh.nl

