It's all about the latency, not the bandwidth!

Ceph Days
Berlin
12-13 Nov 2025

Your.Online

ceph

12-11-2025

Public

01 Intro

Your.
Online

Who am I?

- Wido den Hollander (1986)
- Born and live in the Netherlands
- Two sons (2020 and 2022)
- CTO @ Your.Online
  - Strategic CTO, oversee our companies. No daily involvement
  - Started my own hosting company in 2003
  - Techie at 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*

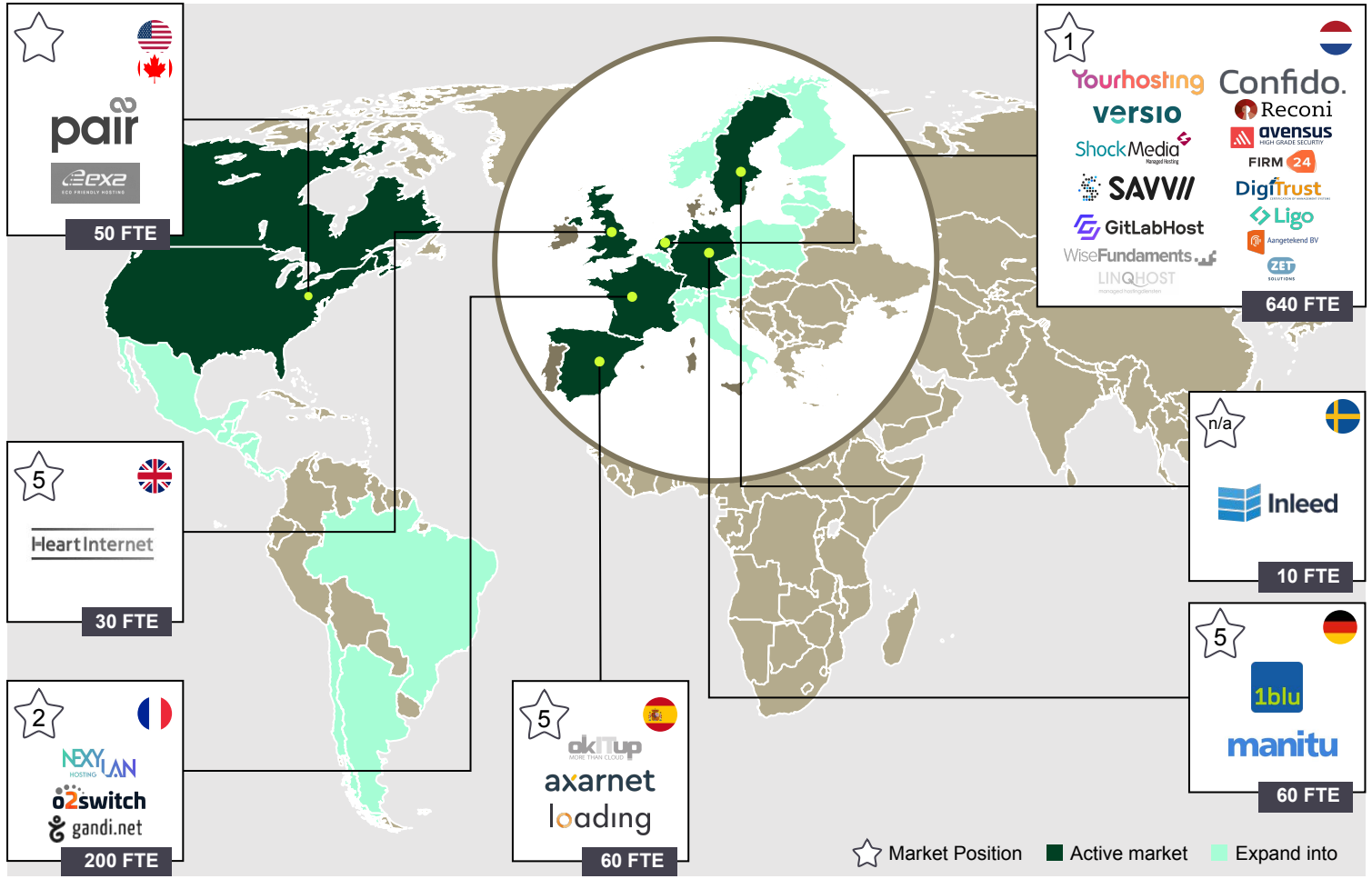
I ❤ 42on.

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 Bandwidth

Your.
Online

What is Bandwidth?

Can anybody explain?

What is Bandwidth?

I asked Google and got an AI response

What is Bandwidth?

“Bandwidth is the maximum amount of data that can be transmitted over an internet connection in a given time, measured in bits per second (bps). It is often compared to the width of a highway, where higher bandwidth allows more data (cars) to travel at once, leading to a smoother online experience”

Oh, yes, I want bandwidth!

- We want a smooth experience, right?
- So you *want* this big highway!
- We just add one more lane (more bandwidth) if there is congestion!
- More lanes, more bandwidth! I want this!
  - Are you sure?

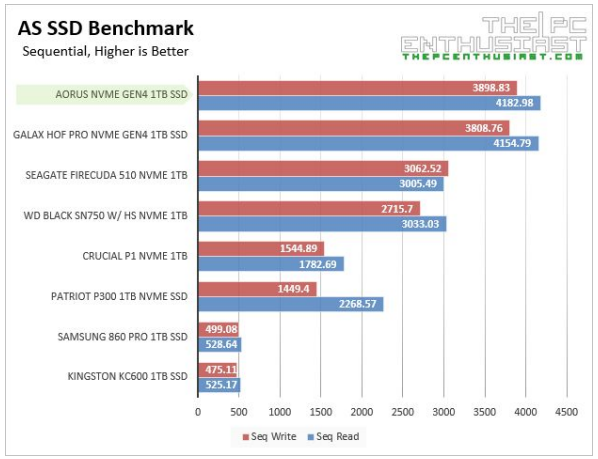
Oh, yes, I want bandwidth!

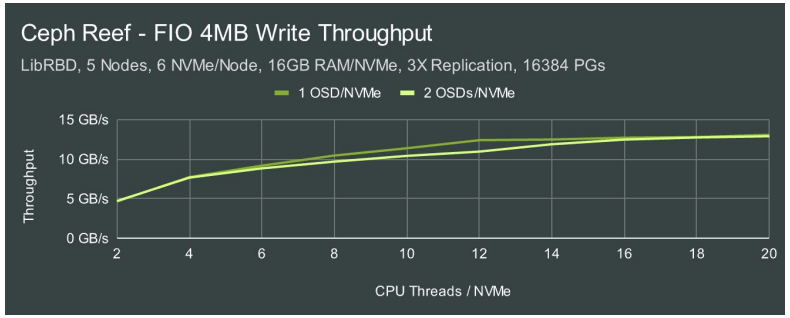
- More lanes will not get you to your destination faster
  - If you obey the speedlimit
- The total throughput of the highway will increase
  - The time it takes to travel does not decrease
- Are you still sure you want this?

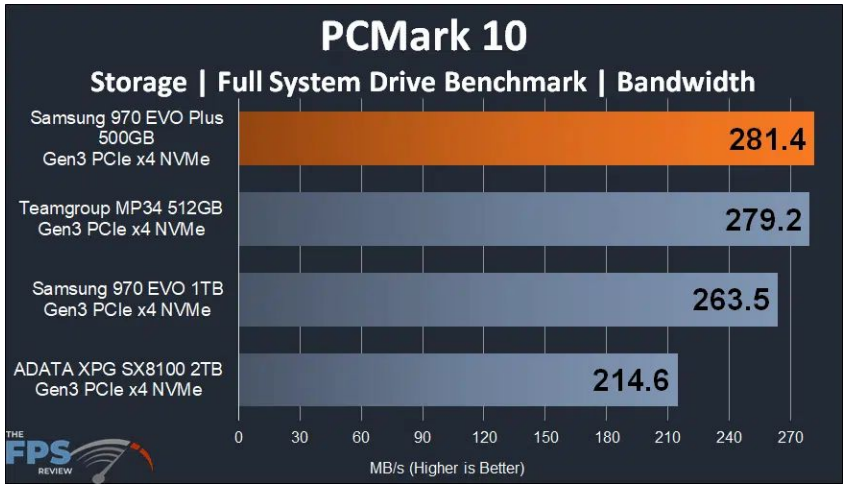
50 lane highway filled with traffic

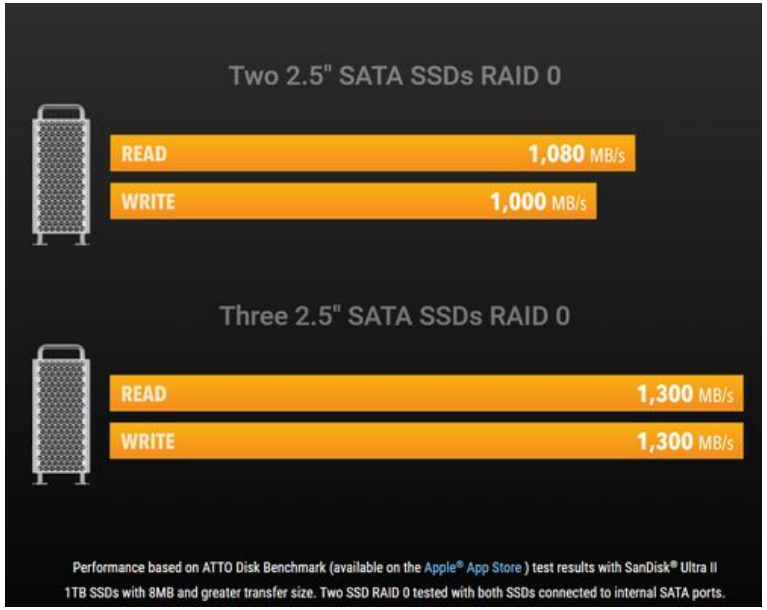
But I still need bandwidth, right?

Everybody **only** talks about bandwidth (Well, almost everybody….)









But I still need bandwidth, right?

Yes, you need bandwidth, but it’s not the most important and you should properly understand it

Yes, you need bandwidth!

Bandwidth is picking up this box full of items and moving it to somewhere else

Yes, you need bandwidth!

But….. I think that latency is far more important

03 Latency

Your.
Online

Time is the ultimate currency

Movie: In Time
(2011)

Filling out a form

Imagine you're waiting for someone to complete a form.

There are 30 people in the room, each working on a form that takes 15 minutes to finish.

The total **throughput** is high: 30×4 = **120** forms *per hour*.

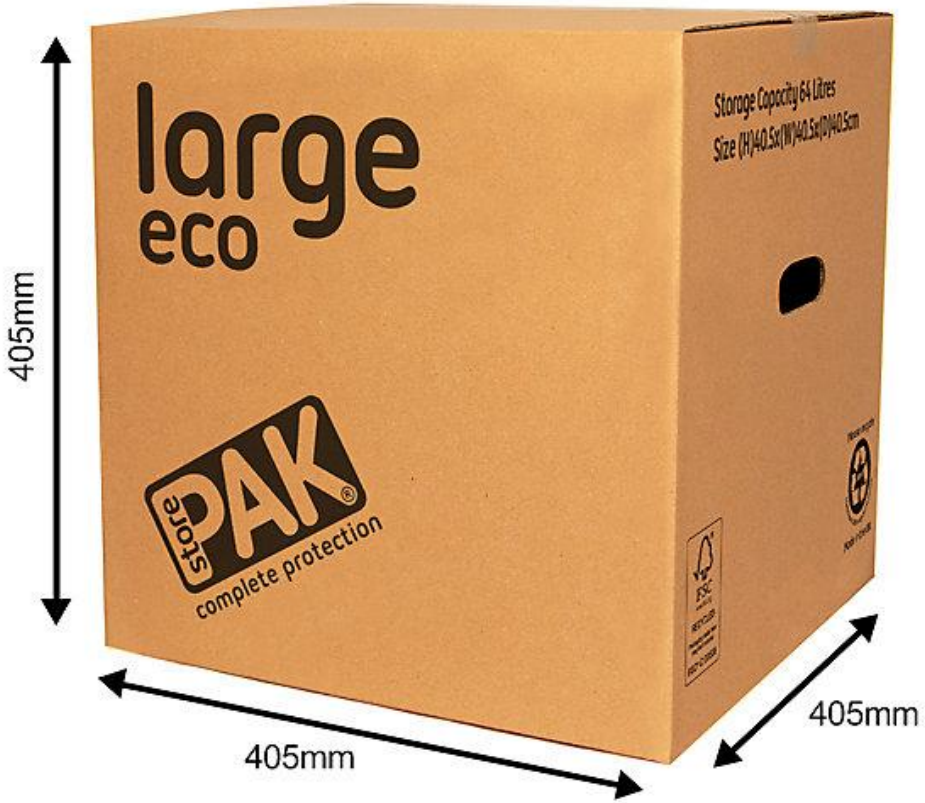
But the latency for any single form is still **15 minutes**.

Do you want to wait *15 minutes* for just one form?

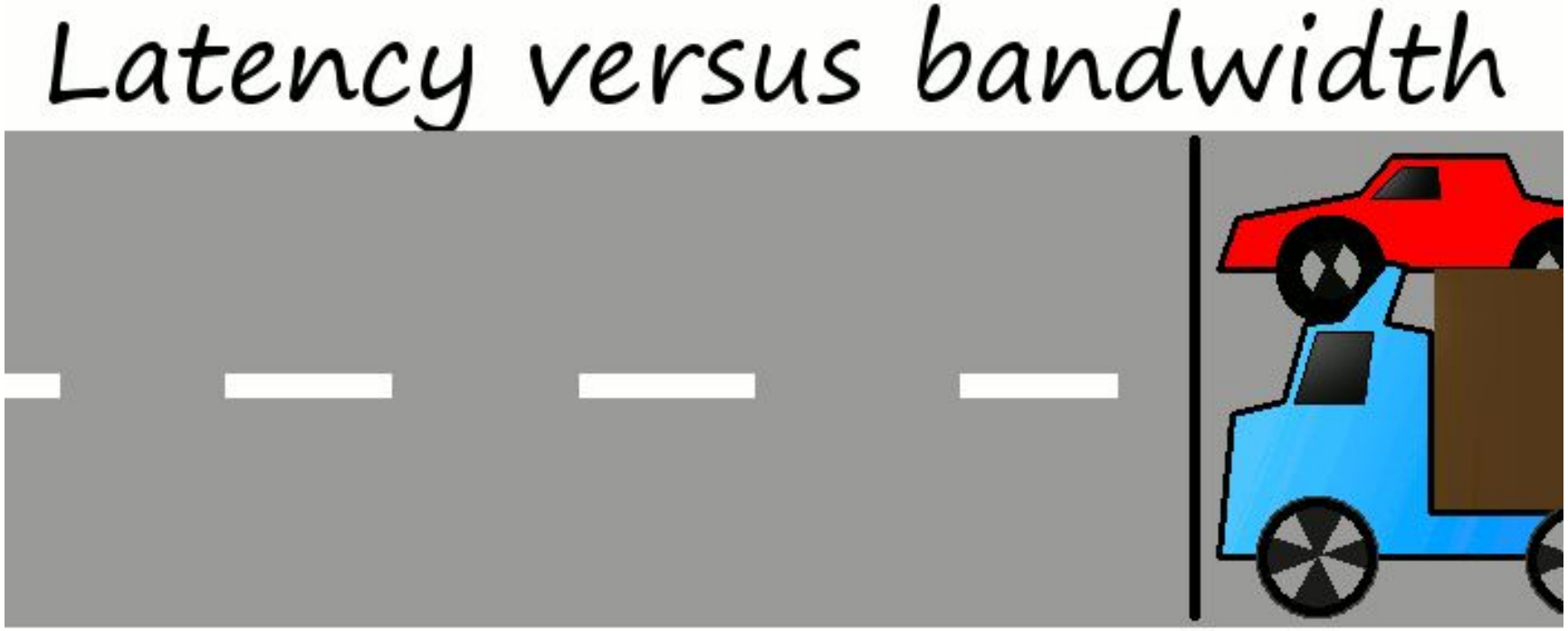
If you reduce the time it takes (*latency*) to fill out a single form, your throughput will go up!

Latency vs Bandwidth

Instead of moving a box full of papers, we move each sheet one by one. The time it takes to handle each piece of paper is the **latency**. You can start to read the first “page” once it arrives instead of having to wait for the whole box to arrive.



What would you choose?



If you are doing light web browsing and want the websites to be snappy, you want low latency. Bandwidth is only important to a certain extent.
If you are downloading a large game from Steam, you want your bandwidth to be as high as possible. Latency is not really a factor in this case.

What about IOps?

- IOps are often used to showcase the performance of a storage system
- They are the result of the latency of a single I/O thread and the number of threads (queue depth) you are using

	System 1	System 2
Threads / Queue Depth	1	1,000
Latency (ms)	1	1,000
IOps	1,000	1,000

Both systems have a performance of 1,000 IOps, but which one do you think performs best for most use-cases?

It’s all about the IOps!

When you see a benchmark, ask these questions:

- What was the block size?
- What was the queue depth?
- Was it read or write?
- What was the **latency** for a single I/O?

I only test writes when I benchmark

queue depth = 1

block size = 4k

Let's go back to 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
- We looked at qd=1, bs=4k IOps of a Ceph cluster.

```
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 rwt: total=0,77563,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1
```

201
0

5.2MB/
s

1,292 IOps x 4096k =
5,2MB/s

0.772 ms

1,292
IOps

1,000ms / 0.772 = 1,292
IOps

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
  - Waiting for those files to be written to the filesystem and *fsync()* to complete
  - An *INSERT* into a SQL database and waiting for *COMMIT* to finish
- 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
  - Millions and millions of small files
    - (Waiting for *rsync* in the middle of the night…..)

MariaDB SQL INSERT

```
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 1576
Server version: 10.6.22-MariaDB-0ubuntu0.22.04.1 Ubuntu 22.04

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [example]> INSERT INTO events (name, city, country, start_date, end_date) VALUES ('Ceph Days Berlin', 'Berlin', 'DE', '2025-11-12', '2025-11-13');
Query OK, 1 row affected (0.002 sec)

MariaDB [example]>
```

Your storage **latency** determines how long this takes, not the *bandwidth*

04 Performance

Your.
Online

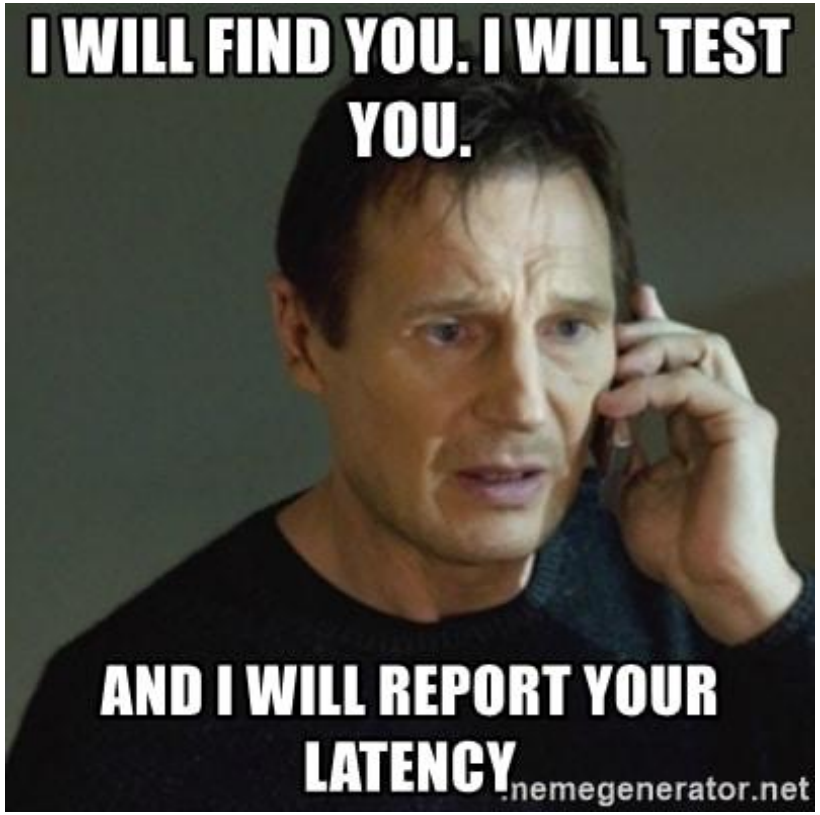
Why you should care

- Latency is what users experience
  - When you hit ”Play” in Netflix and your show starts
  - If you click “Add to Cart” and the product is in your shoppingcart
  - Click “next” to browse through the images of a product and image loads instantly

- All the things above benefit from a low(er) latency

What can you do (with Ceph?)

- **Accept** the fact that network storage has latency implications
- Replication 3x with Ceph takes time
- Choose your applications to run on Ceph wisely
  - Let applications do the data replication instead of Ceph:
    - Don’t run a MariaDB Galera cluster on top of a Ceph cluster
    - Redis can replicate on it’s own, no need for Ceph to do it
- Things that influence the latency of Ceph
  - Storage (NVMe, HDD)
  - CPU clock speed
    - Single core performance, clockspeed
  - Network
    - Do not stretch Ceph over long(er) distances
    - You don’t need 2x100Gb per Ceph node
- **TEST! Benchmark!**



But what is good? And what's bad?

- In the end it's up to you what works for your situation!

- In my case that's single thread low latency
  - Most applications benefit from this

- 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 co

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

05 Benchmarking

Your.
Online

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 **Ynvolve** for providing the hardware

Ynvolve, 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

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

06 The results

Your.
Online

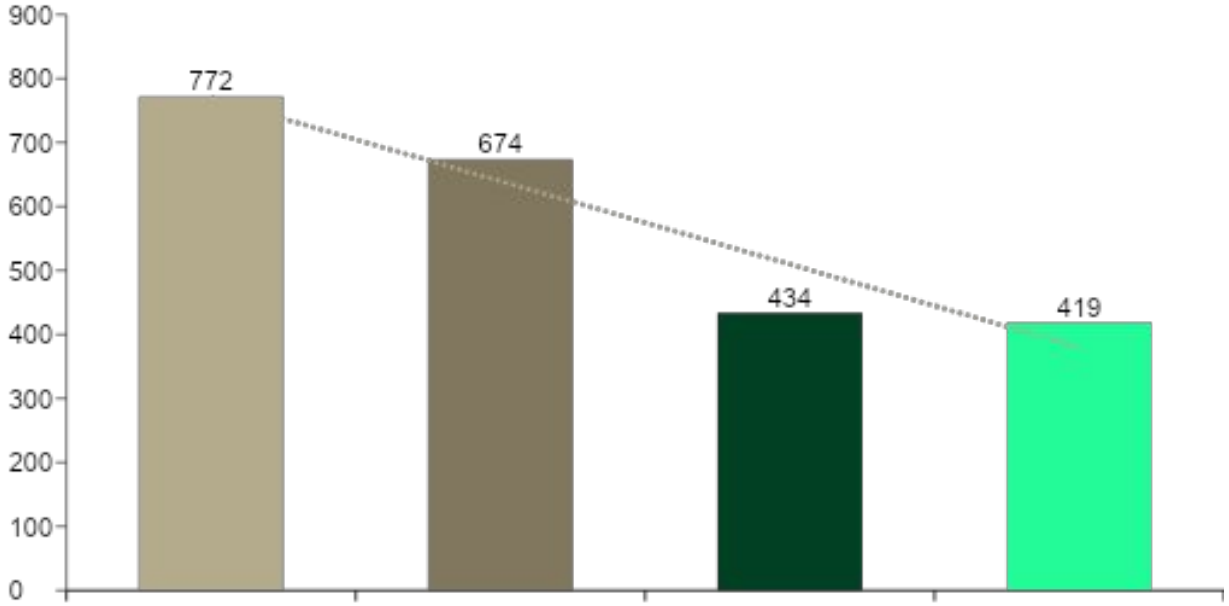
Results

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

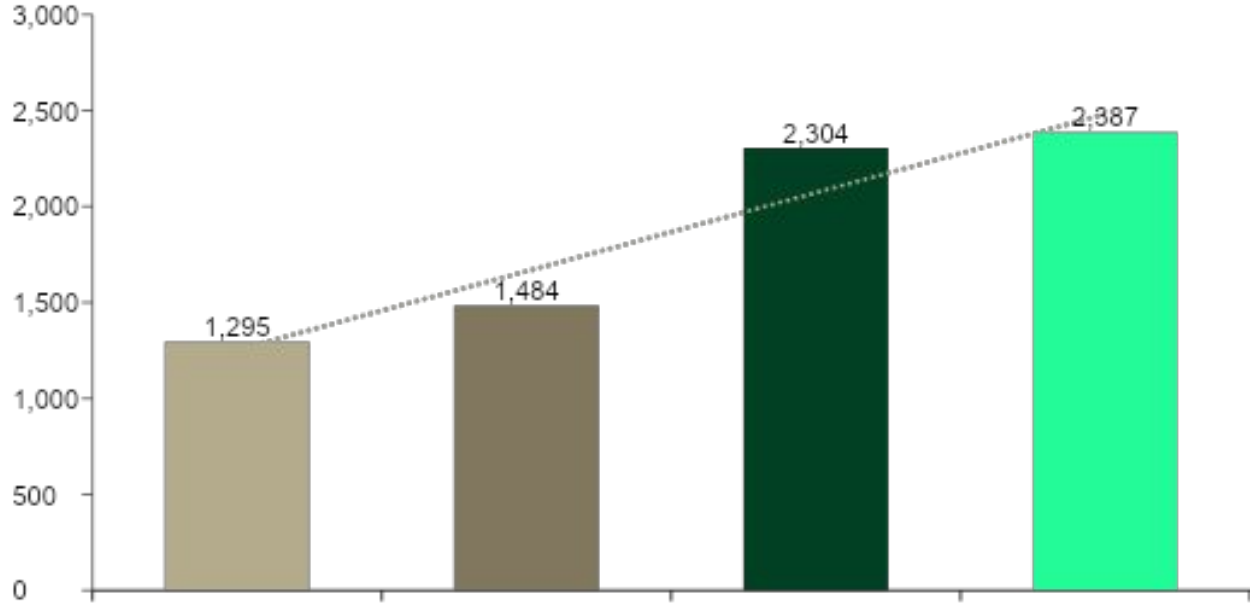
CPU tuning

- 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

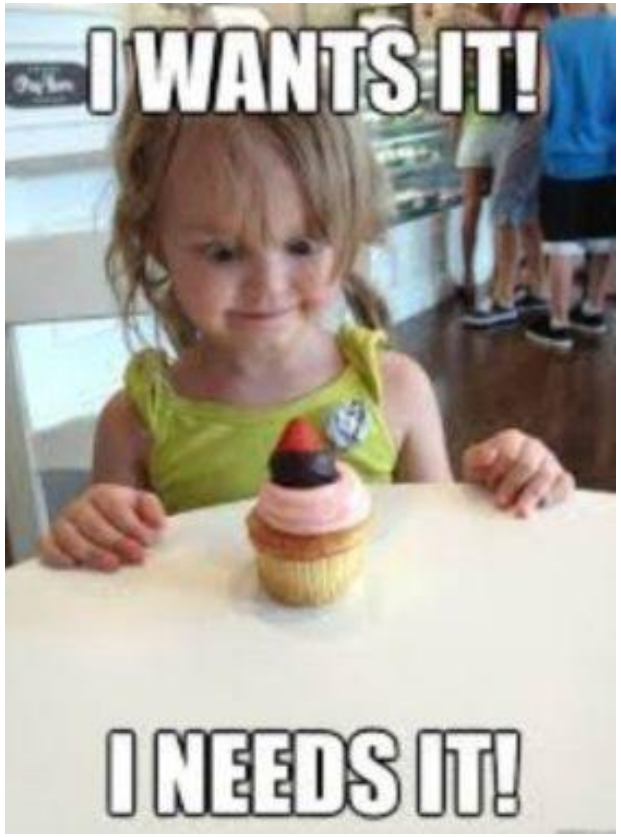
- This still seems to make the biggest difference
- Newer generation of AMD Epyc has big impact
- Could not properly determine if Ceph had improved or if it was only the CPU

Ceph logging

- Disabling logging makes a *small* difference

Famous last words

- Bandwidth is nice, latency is more important
  - You don’t *need* 100Gb networking per Ceph node
- Focus on the (write) latency of a single I/O first
  - Most applications write small pieces of data to the disk
- Latency is what user experience, not bandwidth
  - People want snappy applications, that’s latency
- Remember that I/O is expensive and storage is cheap



That was it!

Thank you!

Any questions or comments?

@widodh

wido@denhollander.io

blog.widodh.nl

Your.
Online