stoney cloud

# Benchmarking Ceph's BlueStore

Niklaus Hofer
Open Cloud Day 2018
2018-05-30

# stoney cloud

# Where we are coming from

- Small cloud company
  - Mostly PaaS
- FOSS-Cloud.org based cloud
  - libvirt based
  - No longer actively maintained
  - GlusterFS storage backend

# State of the (OpenStack) cloud

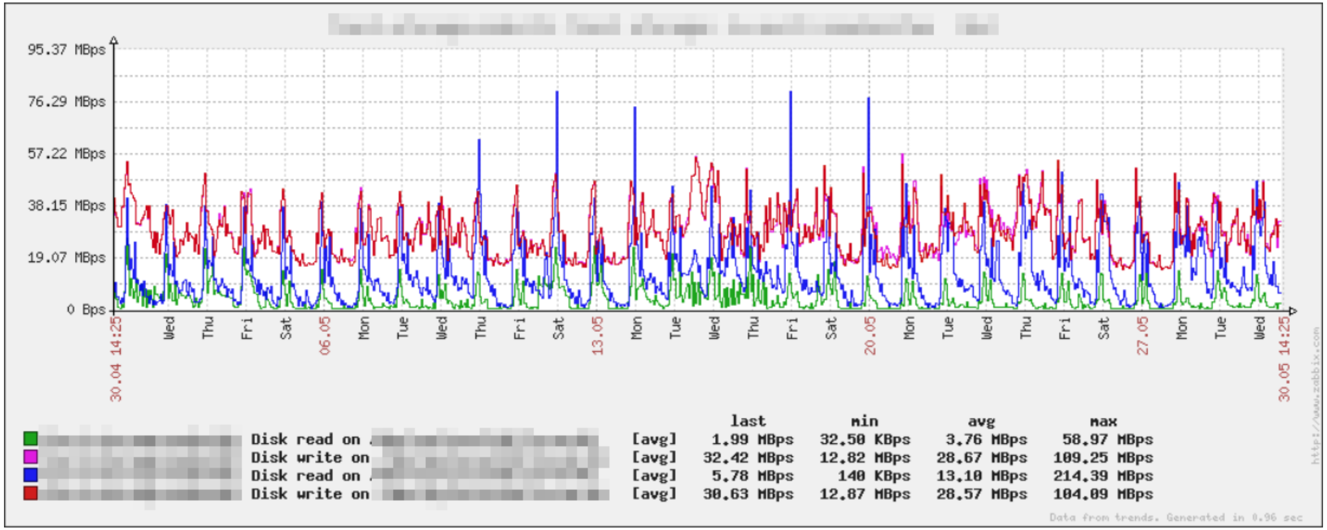
- OpenStack based
- Live: next month
  - A lot of expensive hardware is currently idly
  - Benchmarks!

# Early benchmarks

- Direct SSD vs Ceph vs GlusterFS
- Later: GlusterFS is NOT an option
- Variations within Ceph
  - BlueStore vs Filestore

# Ceph hardware

- Dedicated nodes
- 3 nodes
- SATA SSDs
  - Micron m5100 max
- NVME SSDs Test



[redacted]

| | | | last | min | avg | max |
|---|---|---|---|---|---|---|
| 🟩 [redacted] | Disk read on [redacted] | [avg] | 1.99 MBps | 32.50 KBps | 3.76 MBps | 58.97 MBps |
| 🟪 [redacted] | Disk write on [redacted] | [avg] | 32.42 MBps | 12.82 MBps | 28.67 MBps | 109.25 MBps |
| 🟦 [redacted] | Disk read on [redacted] | [avg] | 5.78 MBps | 140 KBps | 13.10 MBps | 214.39 MBps |
| 🟥 [redacted] | Disk write on [redacted] | [avg] | 30.63 MBps | 12.87 MBps | 28.57 MBps | 104.09 MBps |

Data from trends. Generated in 0.96 sec

# Ceph setup

- Ceph Luminous
- BlueStore
- Replica 3

# Ceph

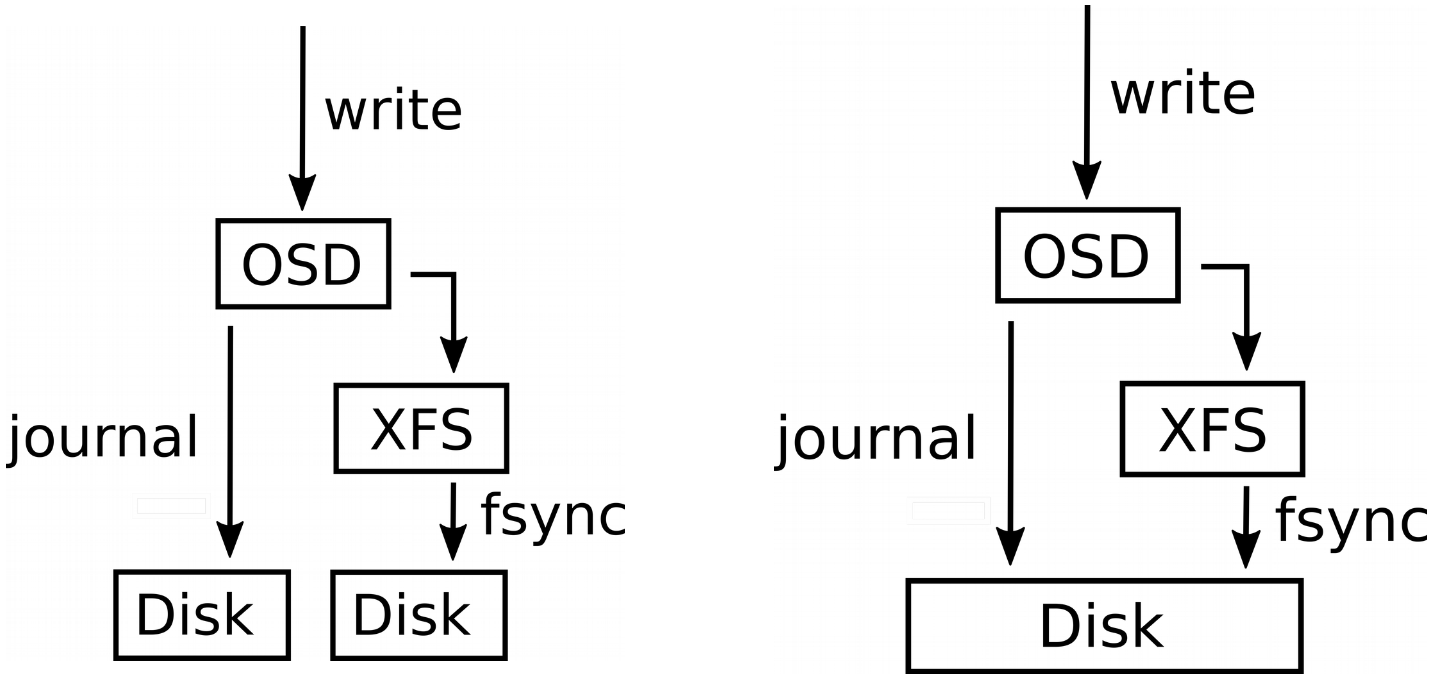
# Ceph journal

- Limitations of POSIX filesystem
- All data gets written twice
- Optimization: Journal on separate disk

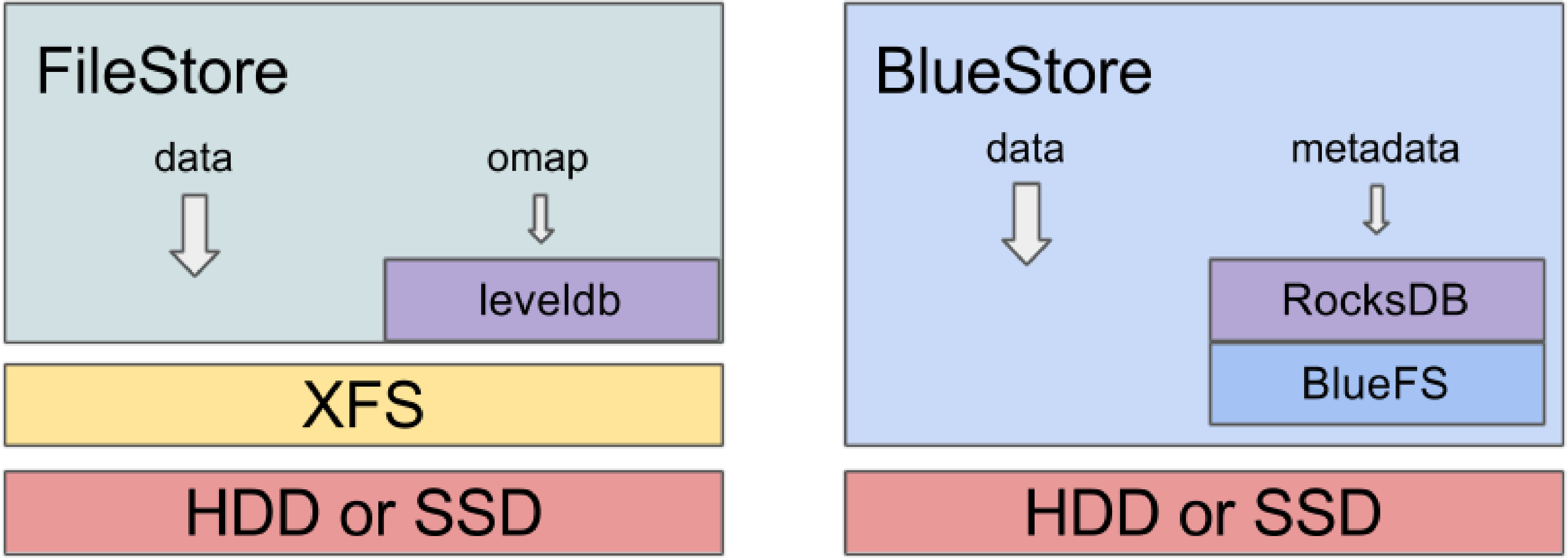
```
#         Start          End    Size  Type            Name
 1     10487808   3750748814    1.8T  Ceph OSD        ceph data
 2         2048     10487807      5G  Ceph Journal    ceph journal
```

```
/dev/sdd :
 /dev/sdd1 ceph data, active, cluster ceph, osd.0, journal /dev/sdd2
 /dev/sdd2 ceph journal, for /dev/sdd1
```

# Ceph journal optimization



# Ceph BlueStore



# Healing performance

```
--dsk/sdc-- --dsk/sdd-- --dsk/sde--
 read  writ: read  writ: read  writ
3905k 4397k:3469k 3666k:3380k 5890k
   0   193M:   0     0 : 111M    0
   0   264M:   0     0 : 123M    0
   0   241M:   0     0 : 114M    0
   0   255M:   0     0 : 114M    0
   0   251M:   0     0 : 121M    0
   0   241M:   0     0 : 116M    0
   0   247M:   0     0 : 120M    0
   0   245M:   0     0 : 122M    0
   0   244M:   0     0 : 122M    0
   0   246M:   0     0 : 110M    0
   0   236M:   0     0 : 118M    0
   0   253M:   0     0 : 124M    0
   0   251M:   0     0 : 126M    0
   0   251M:   0     0 : 122M    0
   0   247M:   0     0 : 114M    0
```

```
--dsk/sdc-- --dsk/sdd-- --dsk/sde--
 read  writ: read  writ: read  writ
   0   132M: 128M    0 :   0     0
   0   128M: 132M    0 :   0     0
   0   128M: 128M    0 :   0     0
   0   112M: 108M    0 :   0     0
   0   125M: 128M    0 :   0     0
   0   128M: 128M    0 :   0     0
   0   120M: 117M    0 :   0     0
   0   111M: 112M    0 :   0     0
   0   125M: 128M    0 :   0     0
   0   128M: 127M  652k:   0     0
   0   130M: 129M    0 :   0     0
   0   132M: 132M    0 :   0     0
   0   131M: 132M    0 :   0     0
   0   130M: 128M    0 :   0     0
   0   131M: 132M    0 :   0     0
   0   129M: 128M    0 :   0     0
```

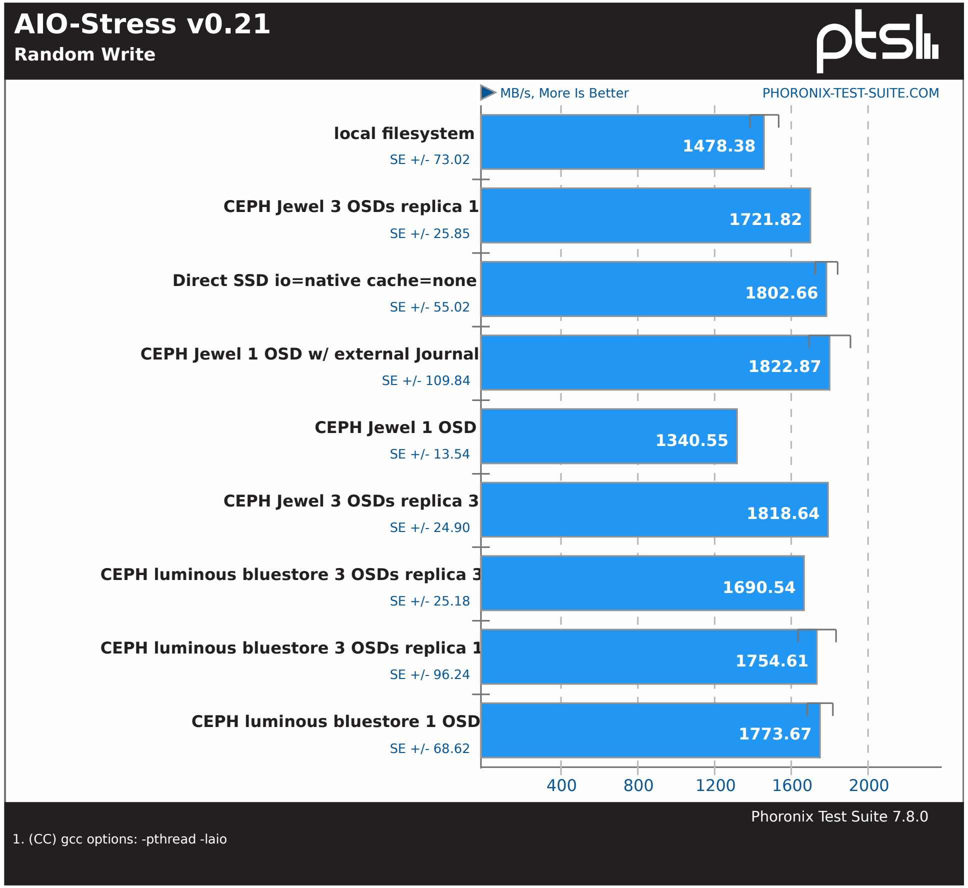
# Benchmarking

# Objectives

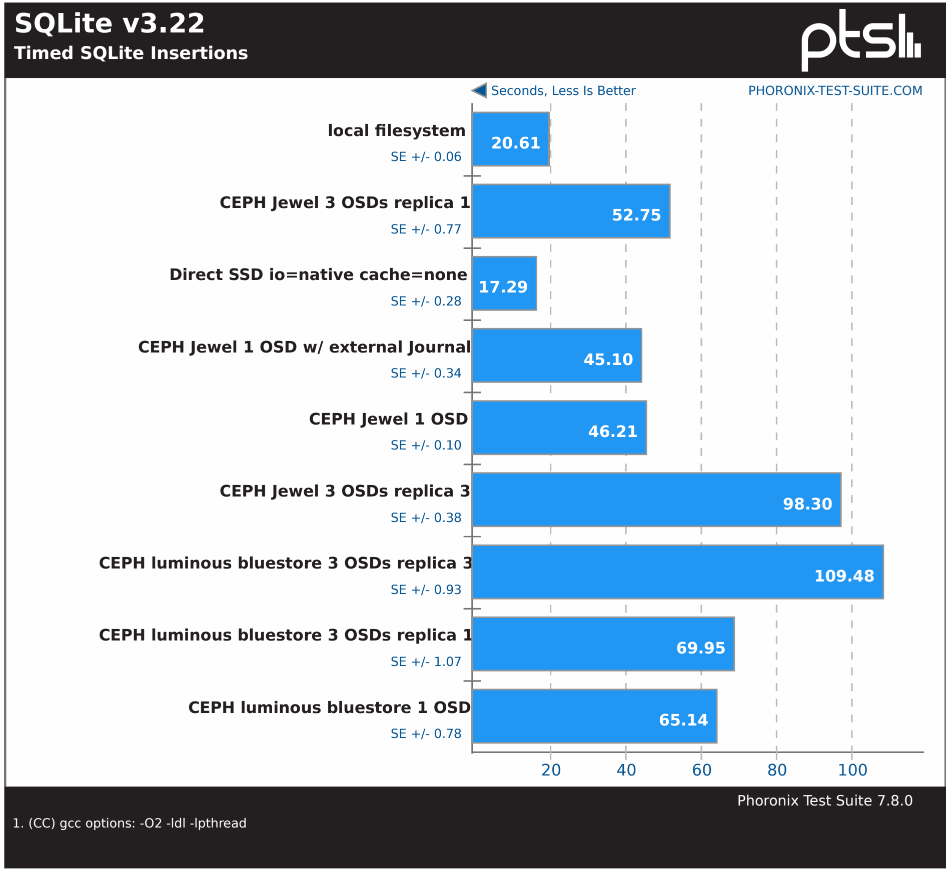
- Single VM
- VM located on separate storage
- Phoronix Test Suite
  - pts/disk

# Benchmarking results

# AIO-Stress

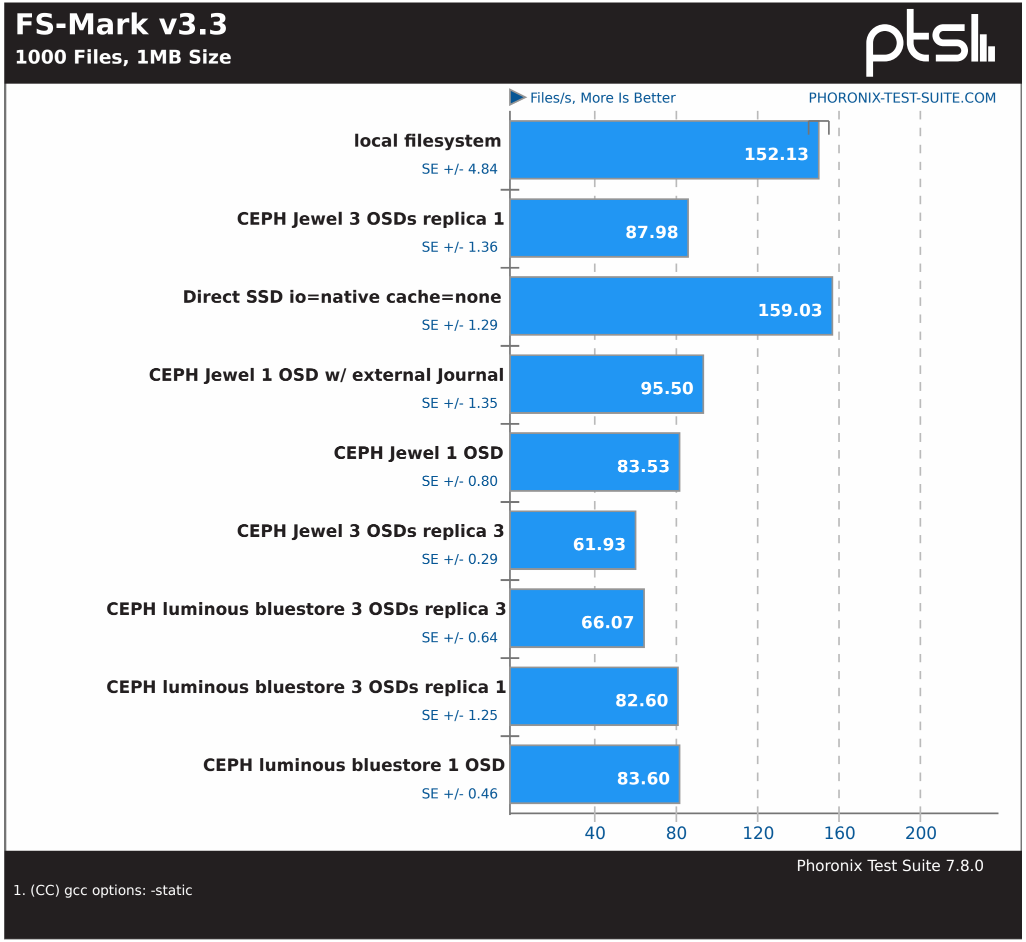

AIO-Stress v0.21
Random Write
MB/s, More Is Better
PHORONIX-TEST-SUITE.COM
local filesystem
SE +/- 73.02
1478.38
CEPH Jewel 3 OSDs replica 1
SE +/- 25.85
1721.82
Direct SSD io=native cache=none
SE +/- 55.02
1802.66
CEPH Jewel 1 OSD w/ external Journal
SE +/- 109.84
1822.87
CEPH Jewel 1 OSD
SE +/- 13.54
1340.55
CEPH Jewel 3 OSDs replica 3
SE +/- 24.90
1818.64
CEPH luminous bluestore 3 OSDs replica 3
SE +/- 25.18
1690.54
CEPH luminous bluestore 3 OSDs replica 1
SE +/- 96.24
1754.61
CEPH luminous bluestore 1 OSD
SE +/- 68.62
1773.67
400
800
1200
1600
2000
Phoronix Test Suite 7.8.0
1. (CC) gcc options: -pthread -laio

# SQLite



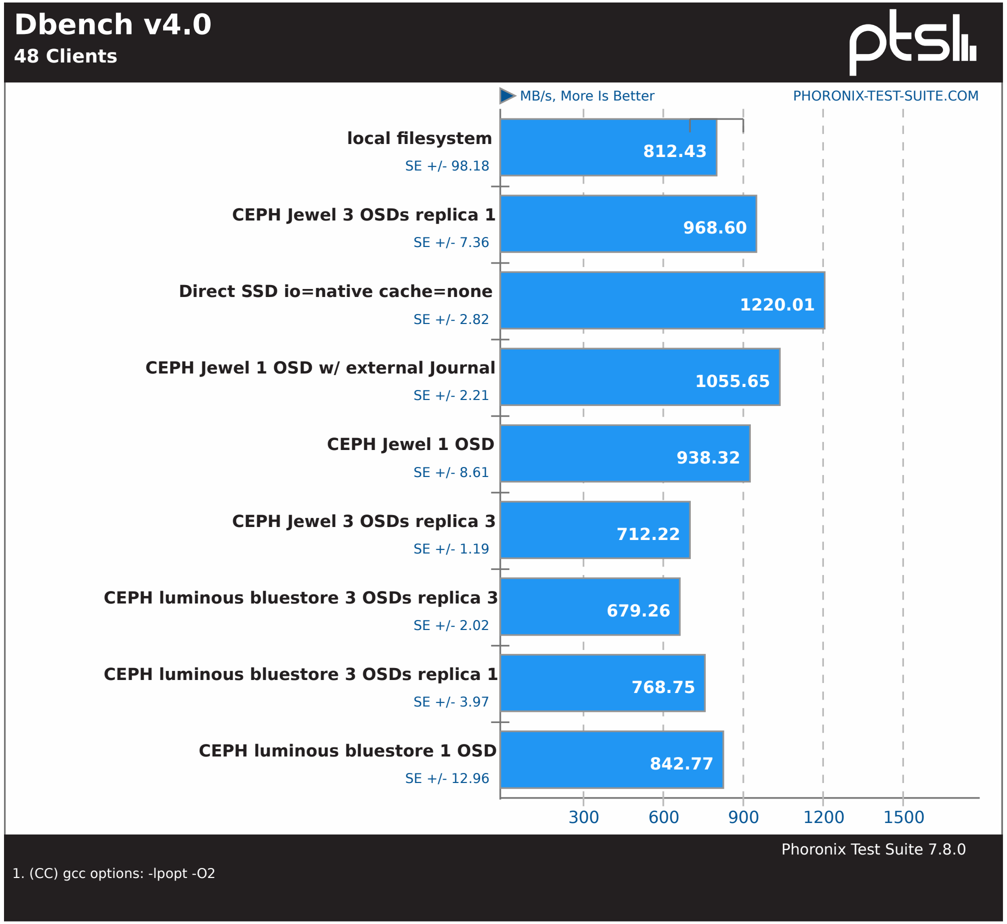
SQLite v3.22
Timed SQLite Insertions
ptsh
Seconds, Less Is Better
PHORONIX-TEST-SUITE.COM
local filesystem
SE +/- 0.06
20.61
CEPH Jewel 3 OSDs replica 1
SE +/- 0.77
52.75
Direct SSD io=native cache=none
SE +/- 0.28
17.29
CEPH Jewel 1 OSD w/ external Journal
SE +/- 0.34
45.10
CEPH Jewel 1 OSD
SE +/- 0.10
46.21
CEPH Jewel 3 OSDs replica 3
SE +/- 0.38
98.30
CEPH luminous bluestore 3 OSDs replica 3
SE +/- 0.93
109.48
CEPH luminous bluestore 3 OSDs replica 1
SE +/- 1.07
69.95
CEPH luminous bluestore 1 OSD
SE +/- 0.78
65.14
20
40
60
80
100
Phoronix Test Suite 7.8.0
1. (CC) gcc options: -O2 -ldl -lpthread

# FS-Mark

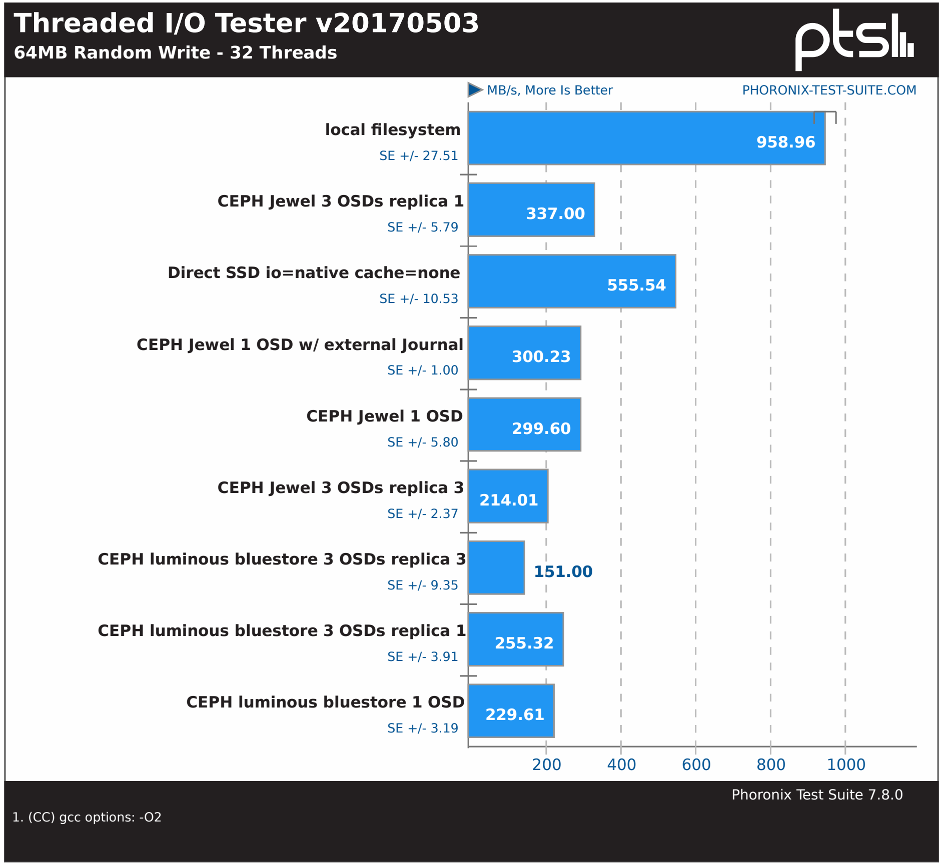


stoney cloud

# Dbench


Dbench v4.0
48 Clients
ptsl
MB/s, More Is Better
PHORONIX-TEST-SUITE.COM
local filesystem
SE +/- 98.18
812.43
CEPH Jewel 3 OSDs replica 1
SE +/- 7.36
968.60
Direct SSD io=native cache=none
SE +/- 2.82
1220.01
CEPH Jewel 1 OSD w/ external Journal
SE +/- 2.21
1055.65
CEPH Jewel 1 OSD
SE +/- 8.61
938.32
CEPH Jewel 3 OSDs replica 3
SE +/- 1.19
712.22
CEPH luminous bluestore 3 OSDs replica 3
SE +/- 2.02
679.26
CEPH luminous bluestore 3 OSDs replica 1
SE +/- 3.97
768.75
CEPH luminous bluestore 1 OSD
SE +/- 12.96
842.77
300
600
900
1200
1500
Phoronix Test Suite 7.8.0
1. (CC) gcc options: -lpopt -O2

# Threaded I/O tester



# Analysis and conclusion

# What went wrong?

- Largely guesswork for now
- Spectre patches?
- Regression?
- Configuration problem?

# Future work

- Optimization
- NVME vs SATA
  - Especially for database setups
- More precise IOPS measurements
  - fio, highly workload dependant

# Questions?

stoney cloud

**stepping stone GmbH**
Wasserwerkgasse 7
CH-3011 Bern

Telefon: +41 31 332 53 63
www.stepping-stone.ch
info@stepping-stone.ch