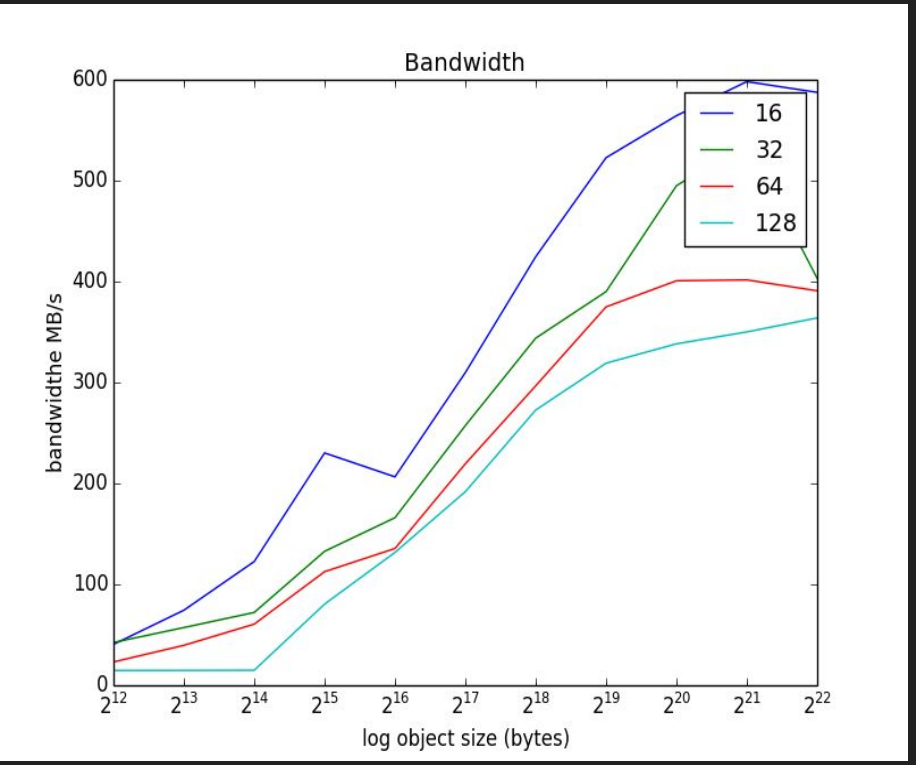
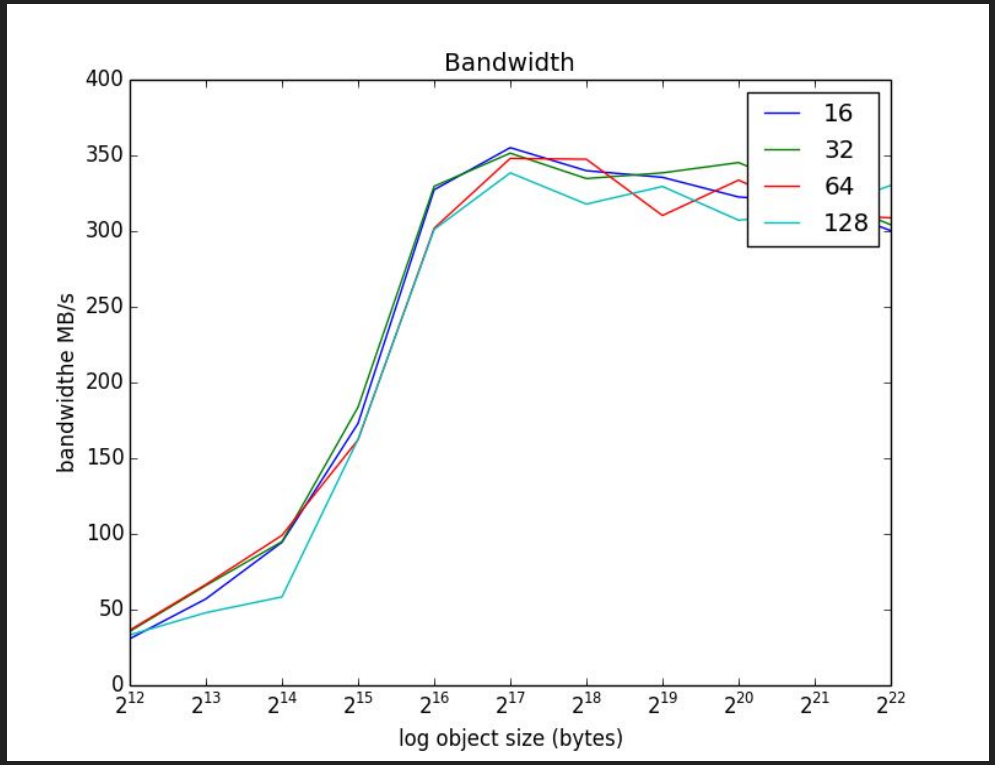
Distributed Systems P1: Intro to Ceph

Yuan-Ting Hsieh
Hsuan-Heng Wu

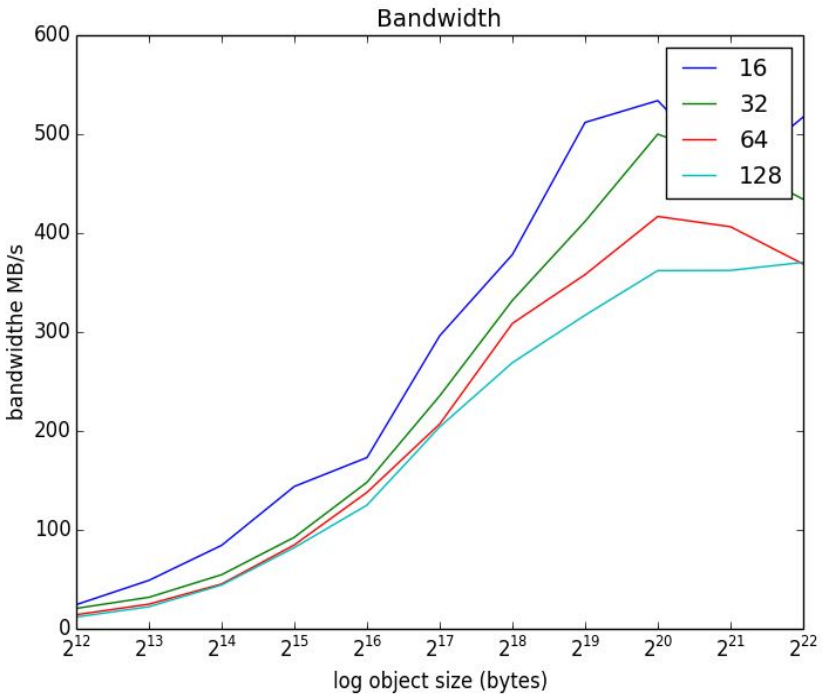
HDD Rand Bandwidth w.r.t Object Size

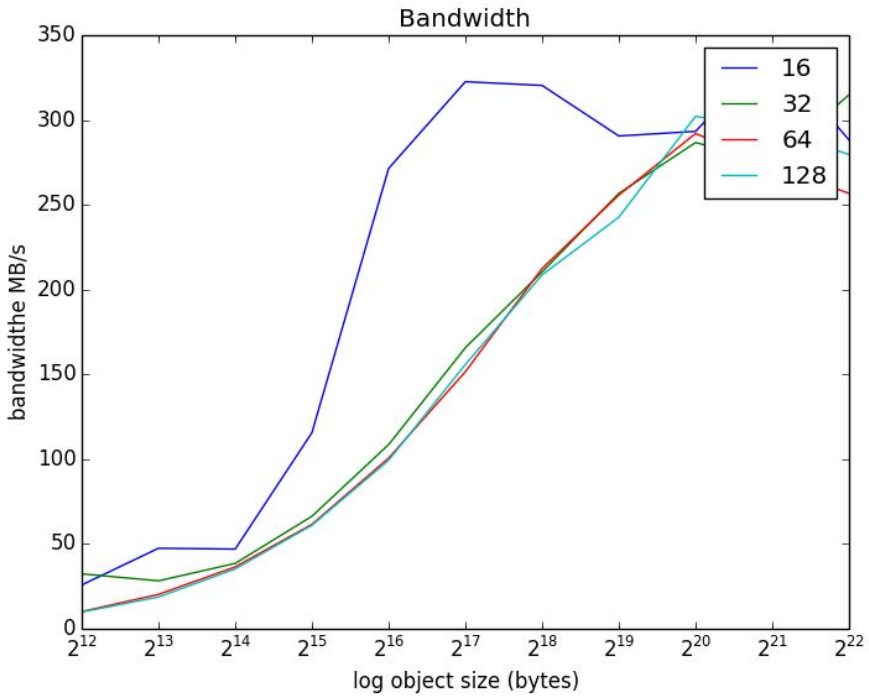
Read Bandwidth n = 1 , random left , sequential right



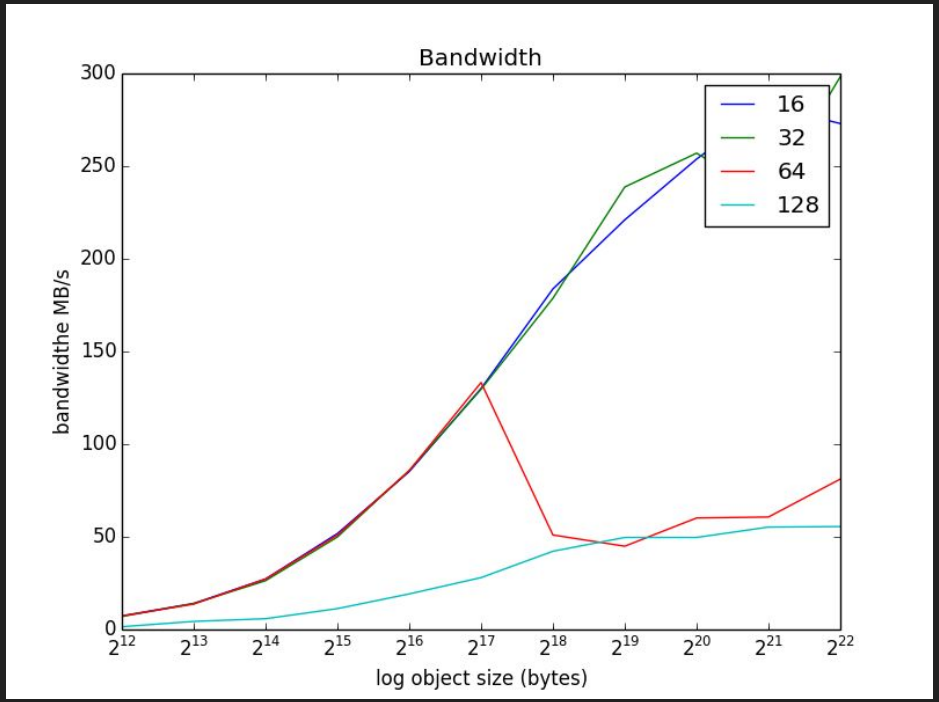


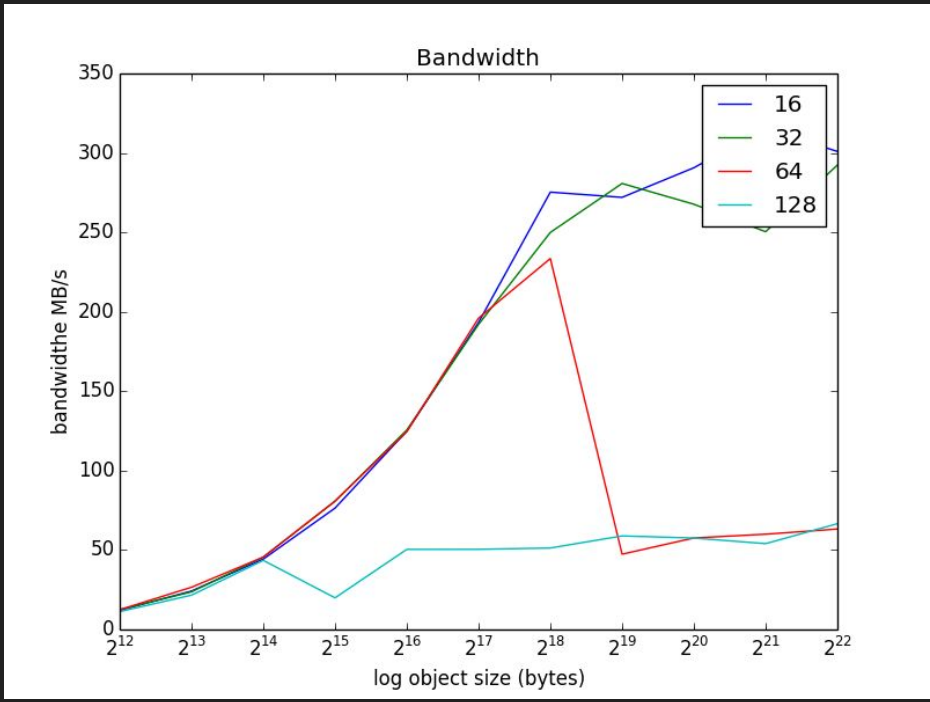
Read Bandwidth n = 2 , random left , sequential right





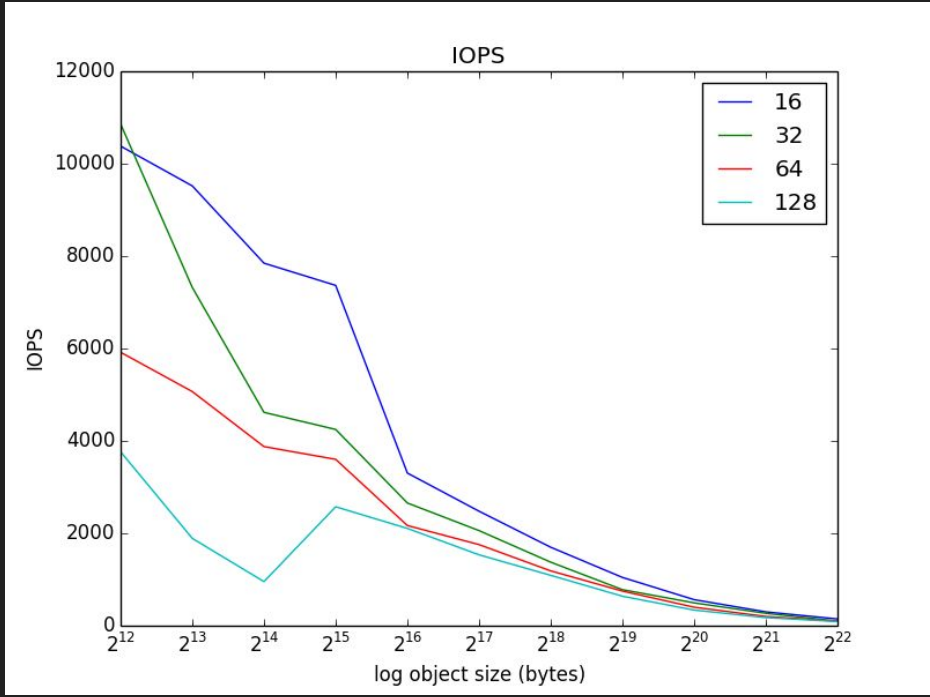
Read Bandwidth n = 3 , random left , sequential right

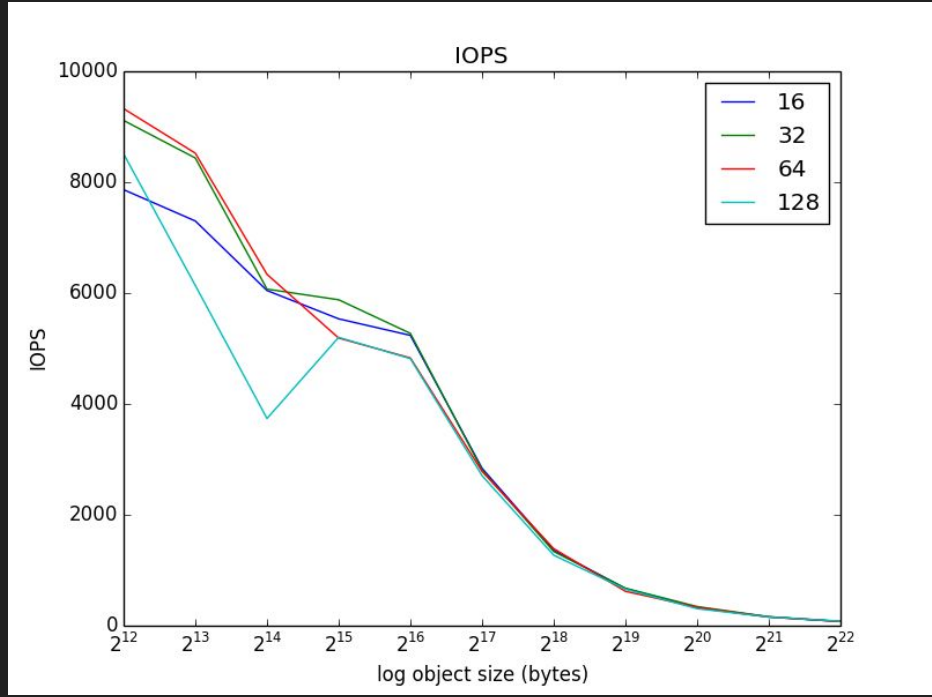




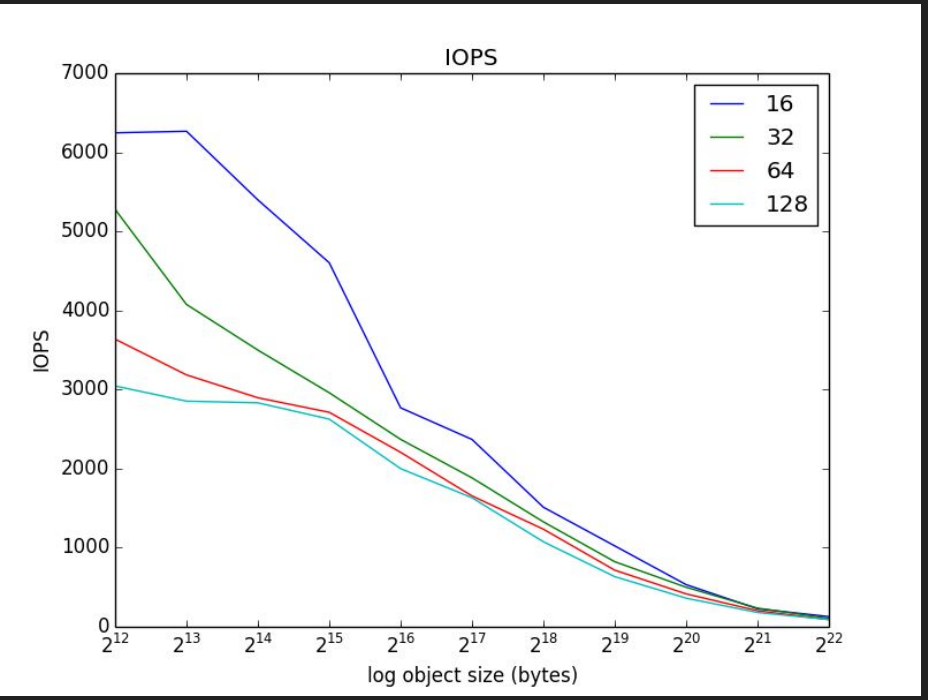
HDD Read IOPS w.r.t Object Size

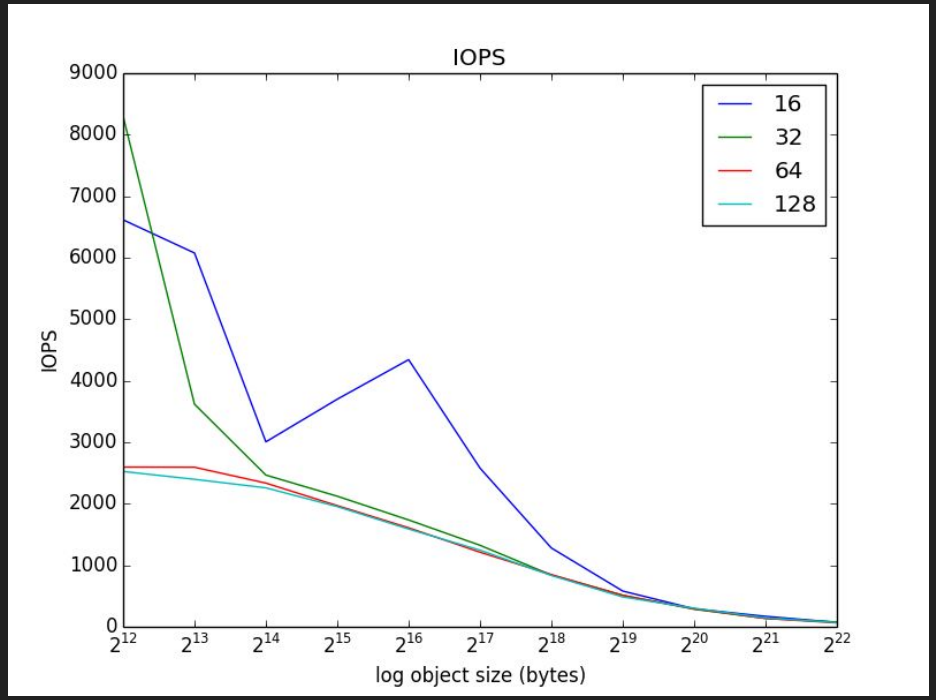
Read IOPS n = 1 , random left , sequential right



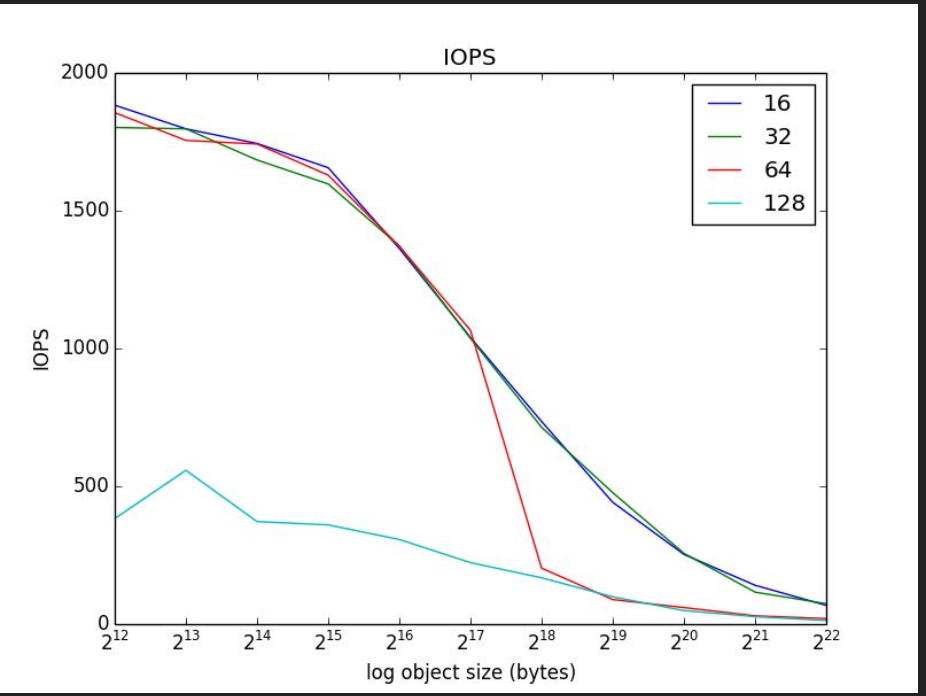


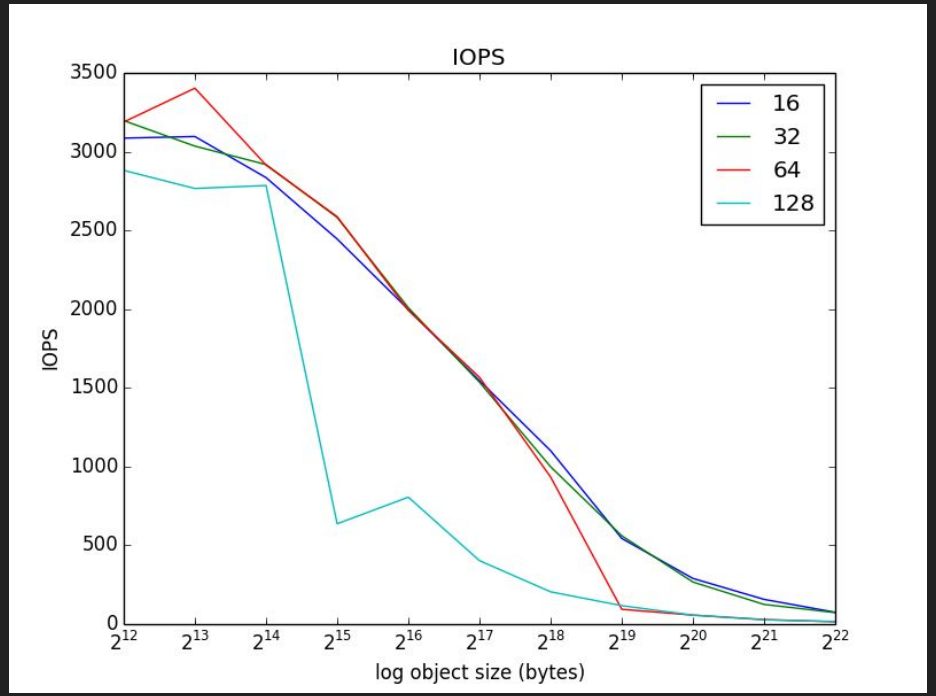
Read IOPS n = 2 , random left , sequential right





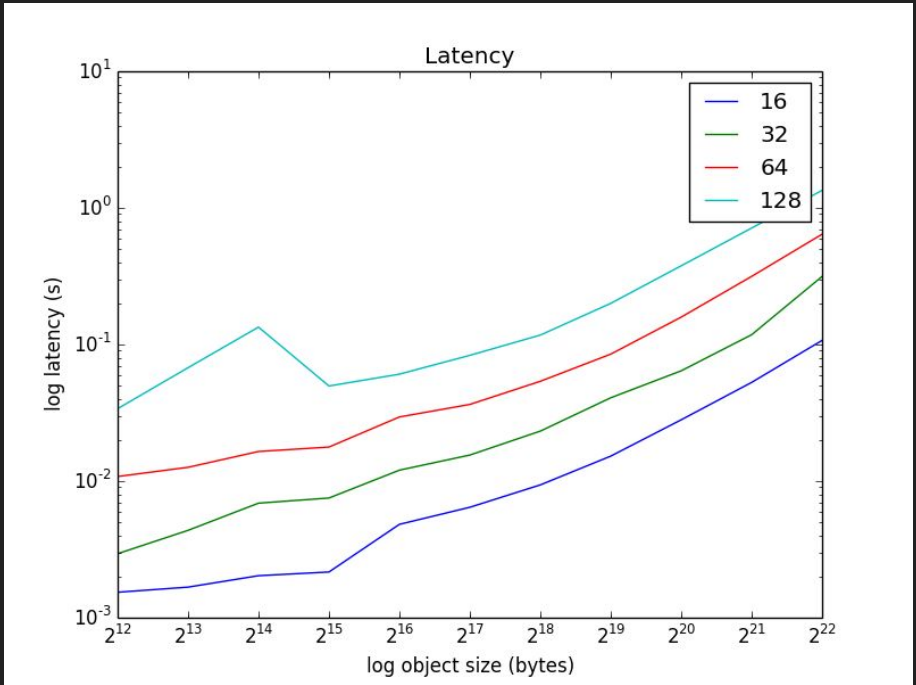
Read IOPS n = 3 , random left , sequential right

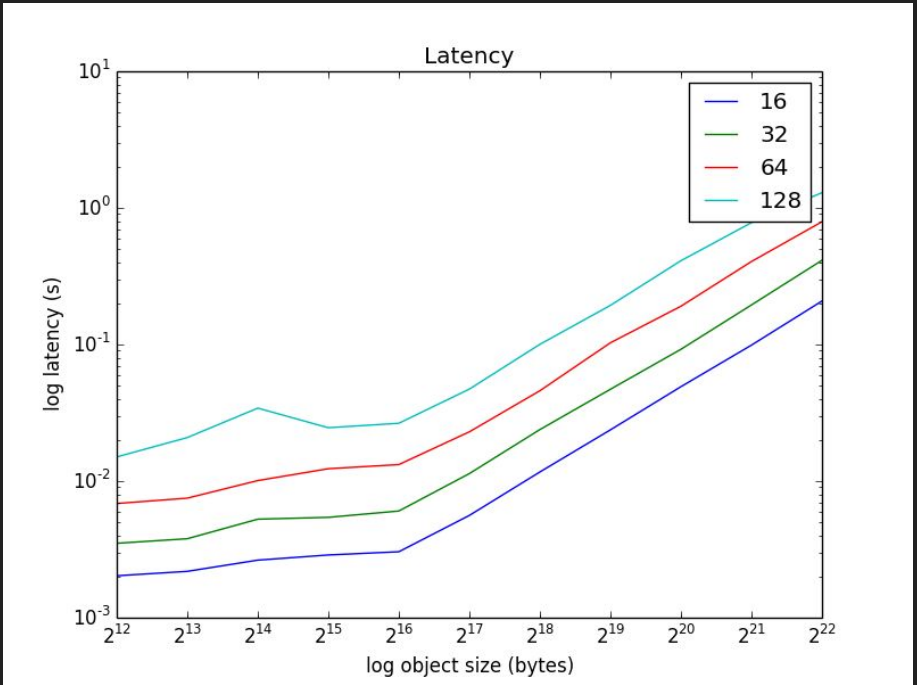




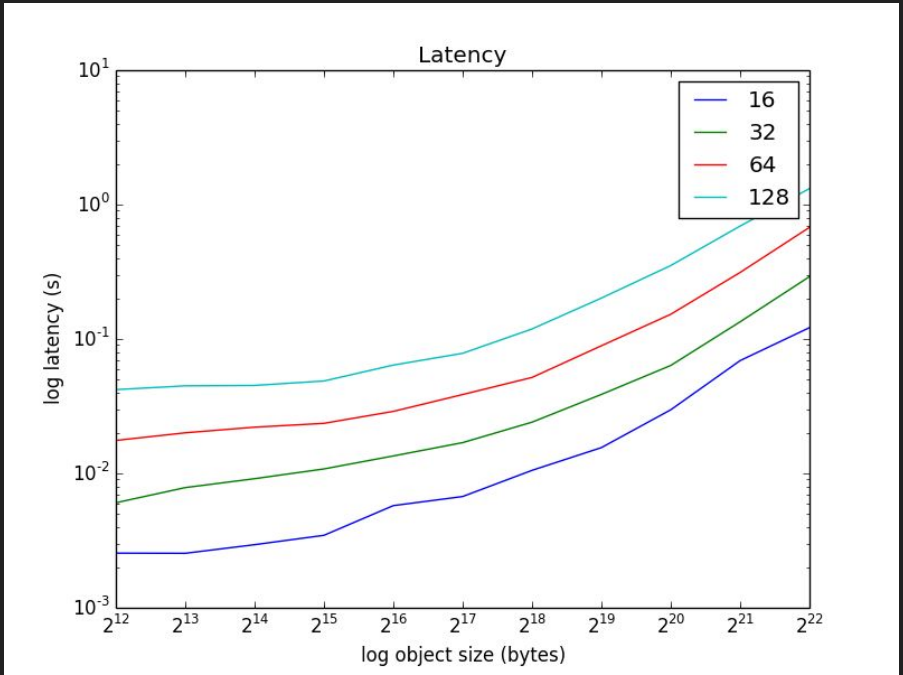
HDD Read Latency w.r.t Object Size

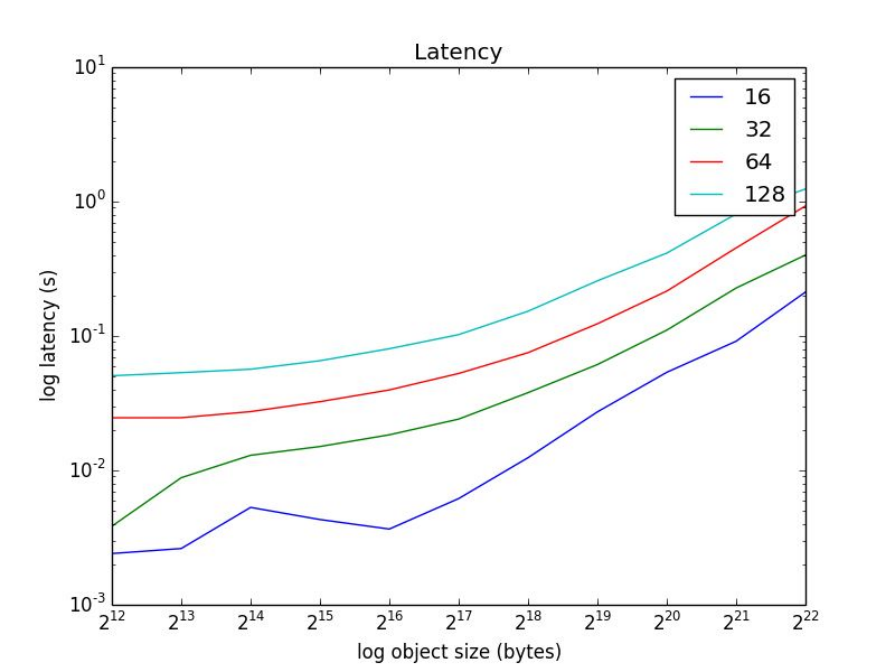
Read Latency n = 1 , random left , sequential right



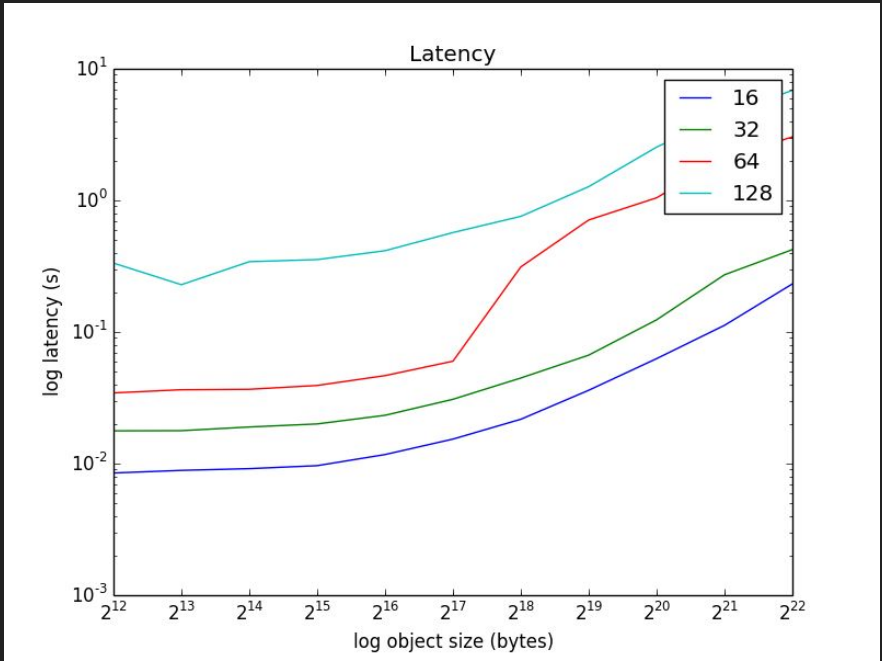


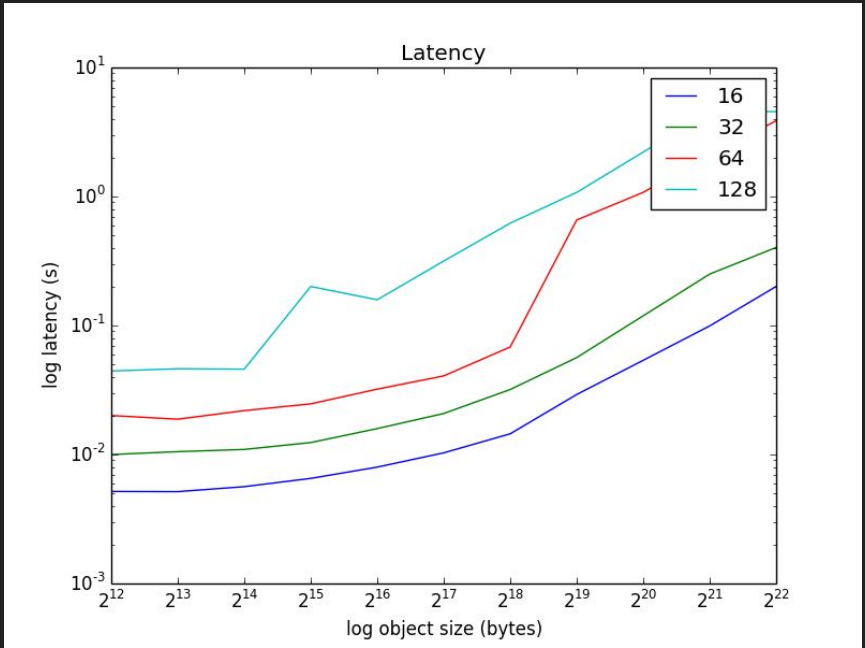
Read Latency n = 2 , random left , sequential right





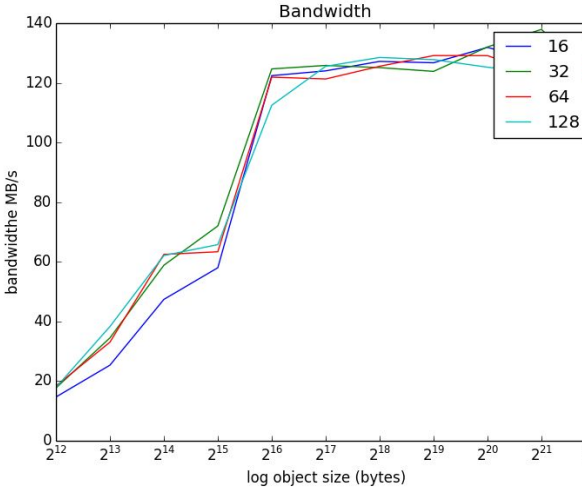
Read Latency n = 3 , random left , sequential right

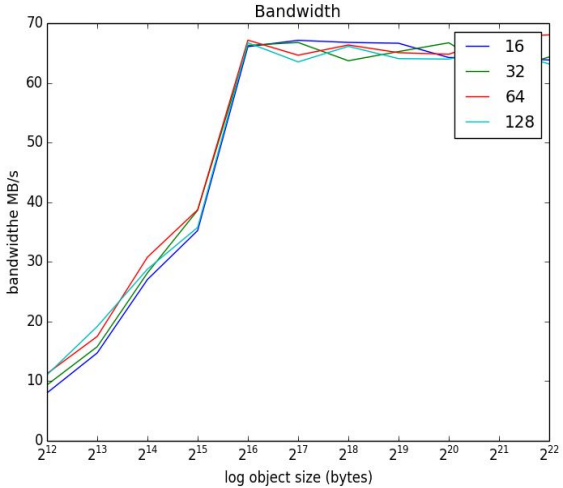


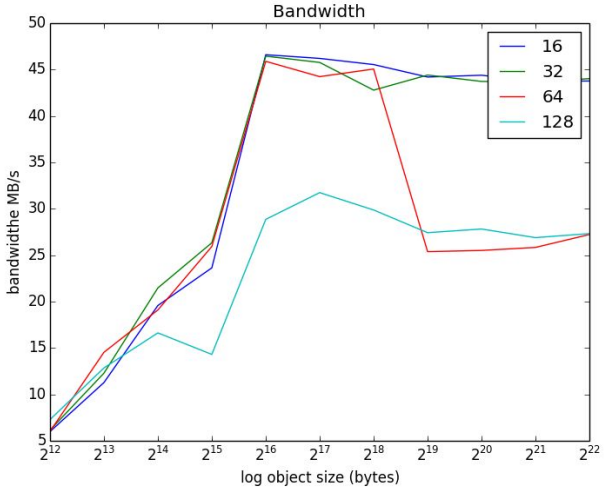


HDD Write Bandwidth w.r.t Object Size

Bandwidth, n = 1 , n= 2 , n = 3

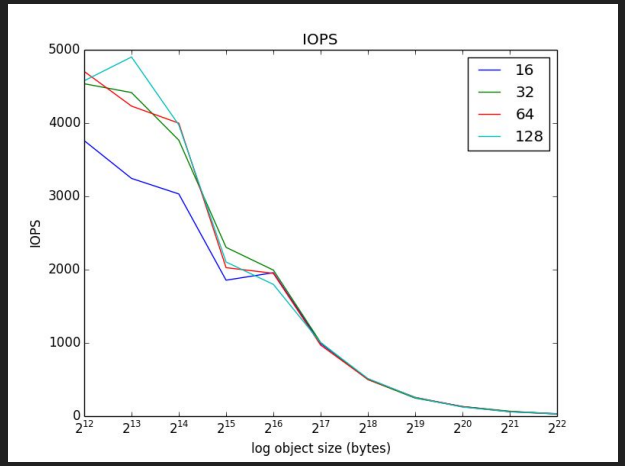


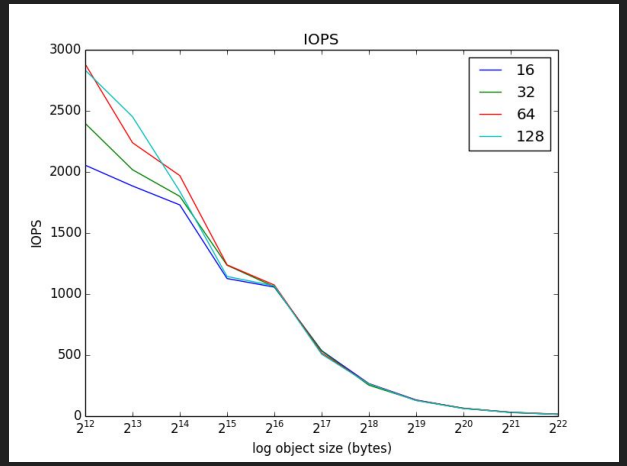


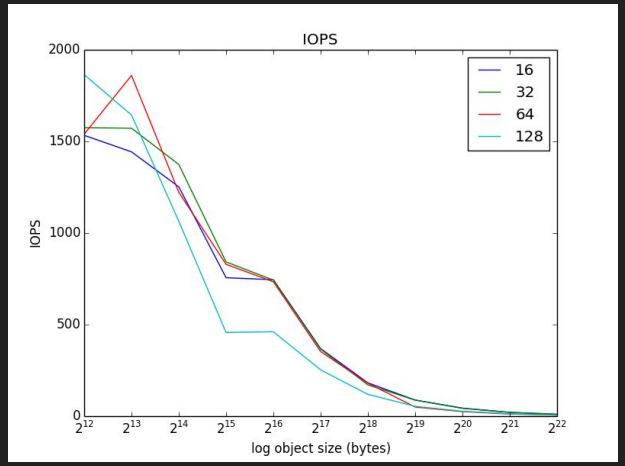


HDD Write IOPS w.r.t Object Size

IOPS, n = 1 , n= 2 , n = 3

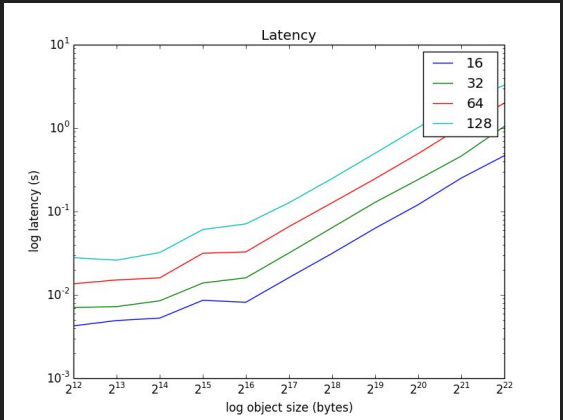

IOPS
16
32
64
128
IOPS
log object size (bytes)

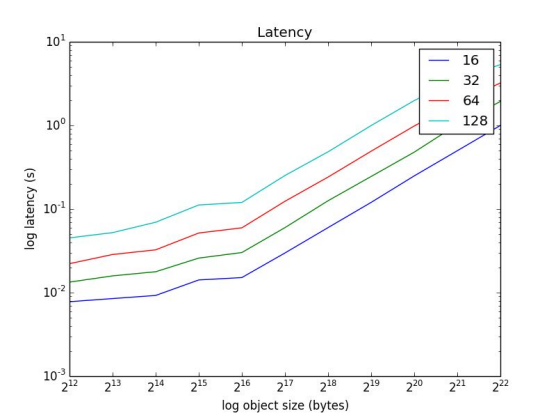

IOPS
16
32
64
128
IOPS
log object size (bytes)

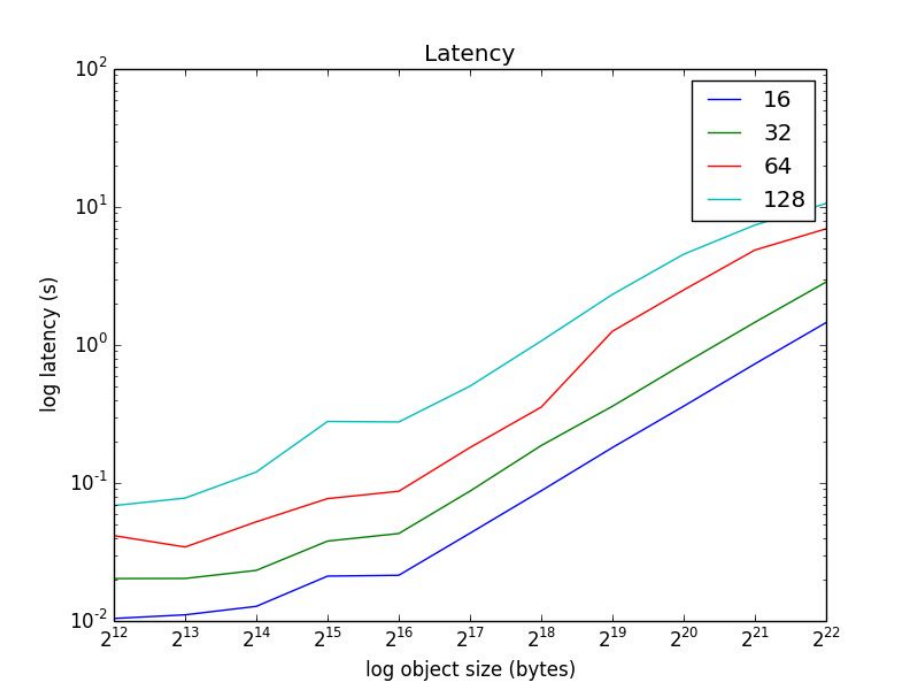

IOPS
16
32
64
128
IOPS
log object size (bytes)

HDD Write Latency w.r.t Object Size

Latency, n = 1 , n= 2 , n = 3

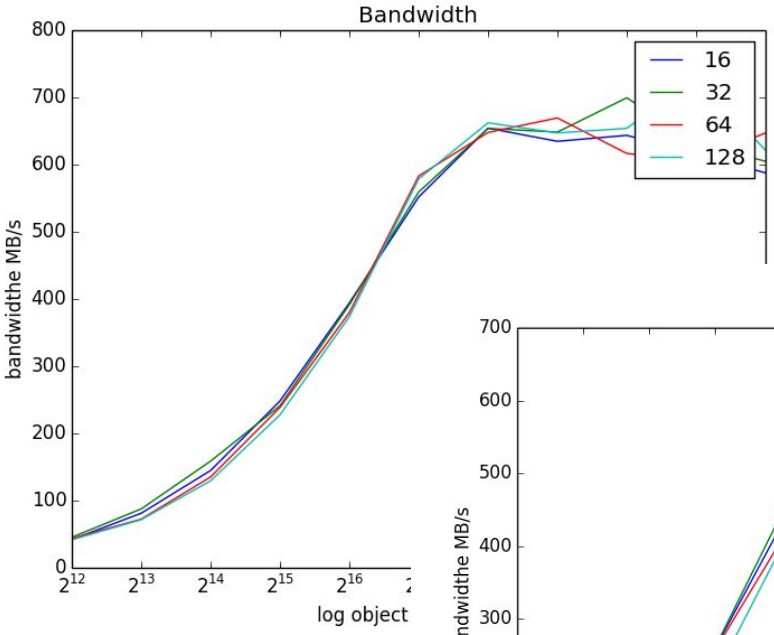


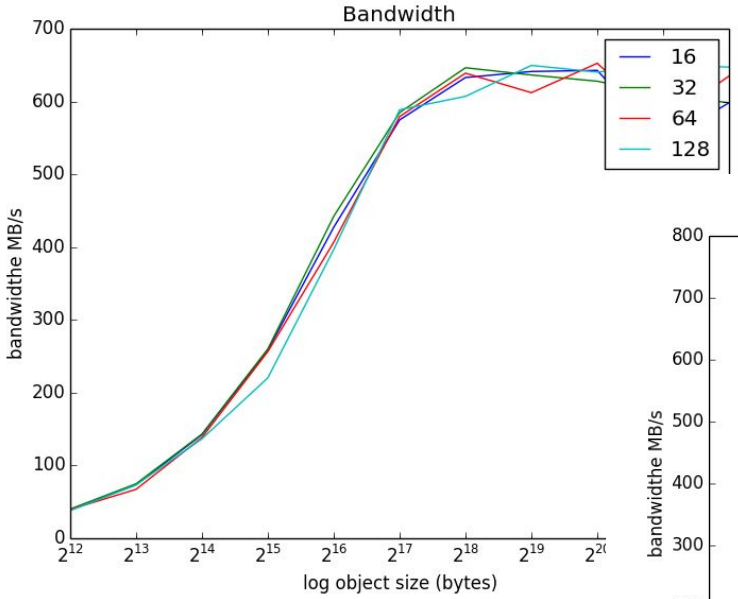


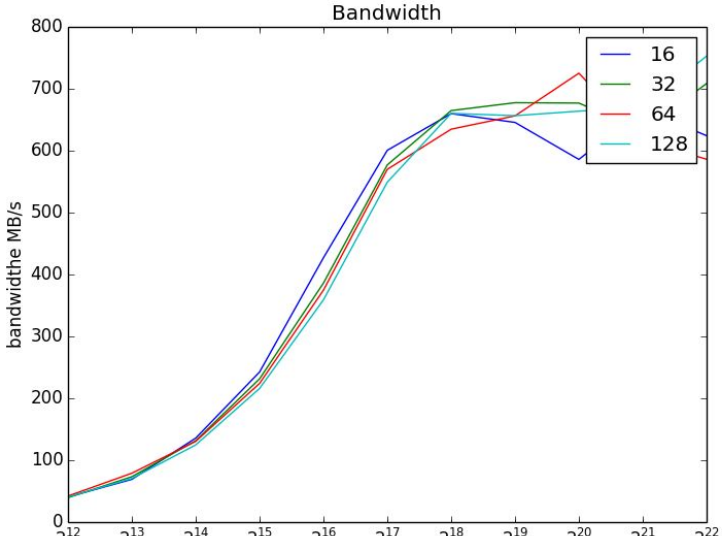


SSD Results - Read Rand

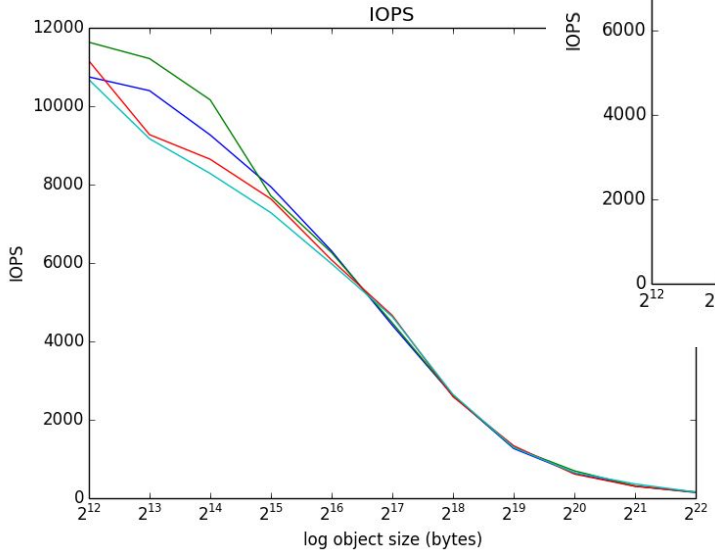
Bandwidth n = 1,2,3

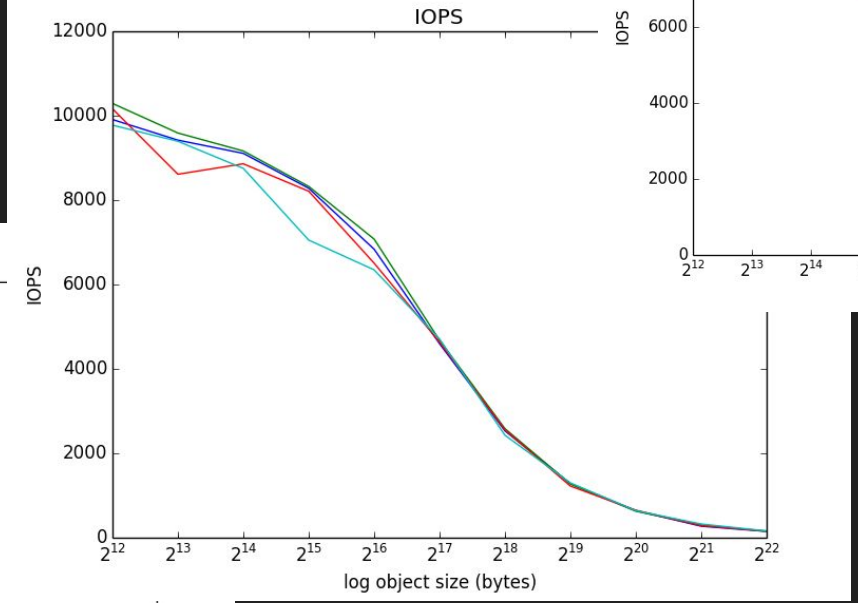


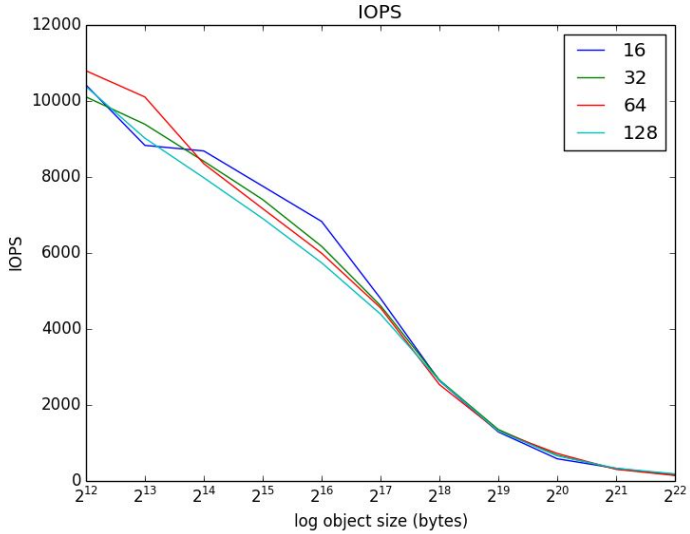




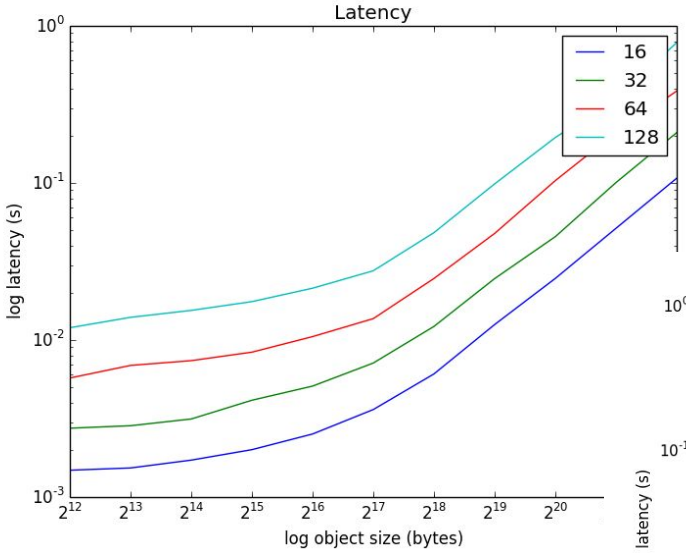
IOPS n = 1,2,3

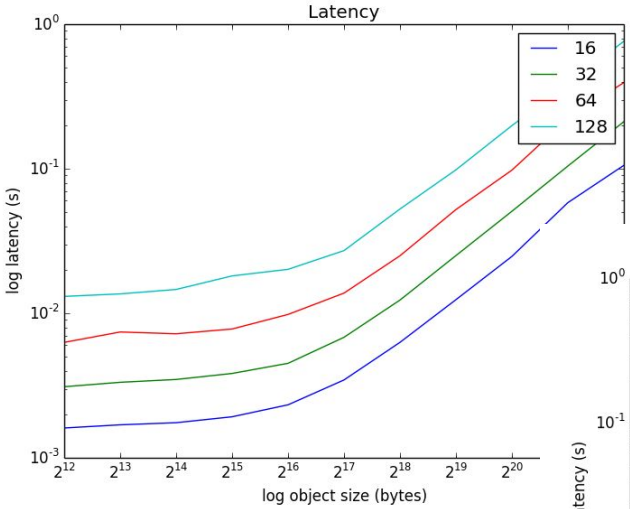

IOPS
12000
10000
8000
6000
4000
2000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

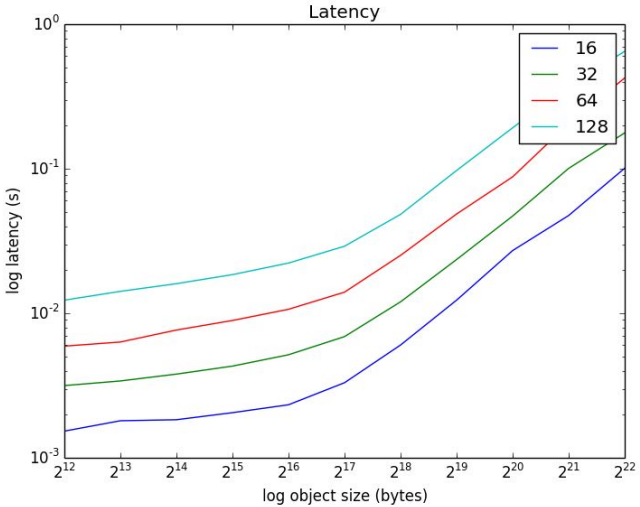

IOPS
12000
10000
8000
6000
4000
2000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)


IOPS
16
32
64
128
12000
10000
8000
6000
4000
2000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

Latency n = 1,2,3

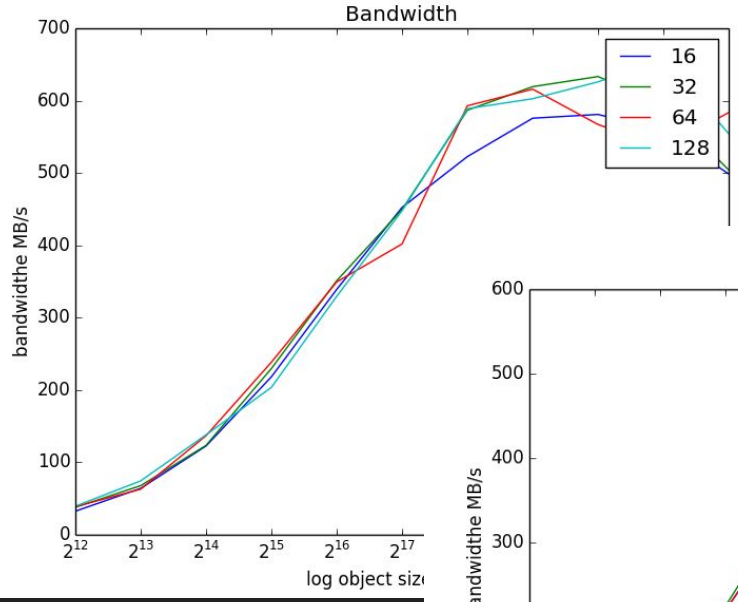


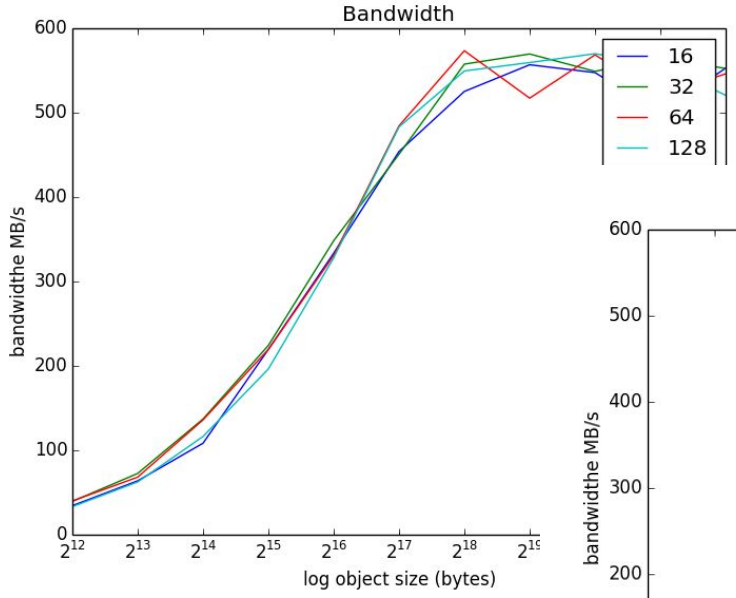


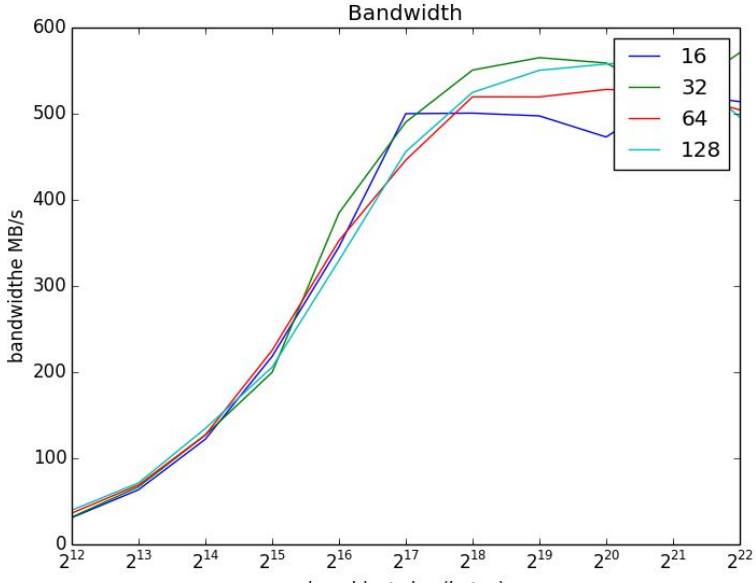


SSD Results - Read Seq

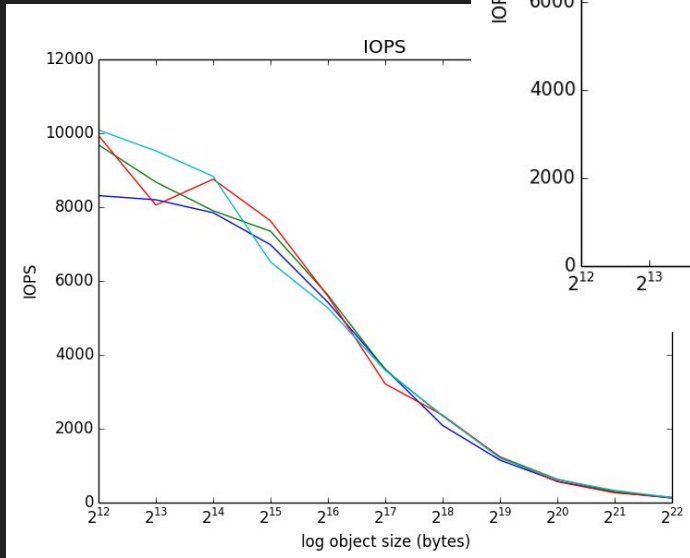
Bandwidth n = 1,2,3

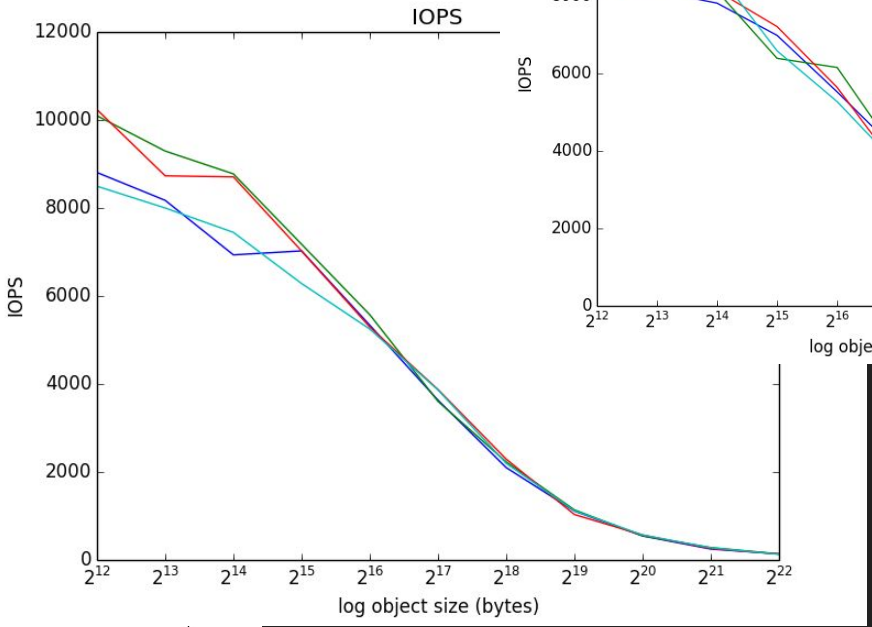


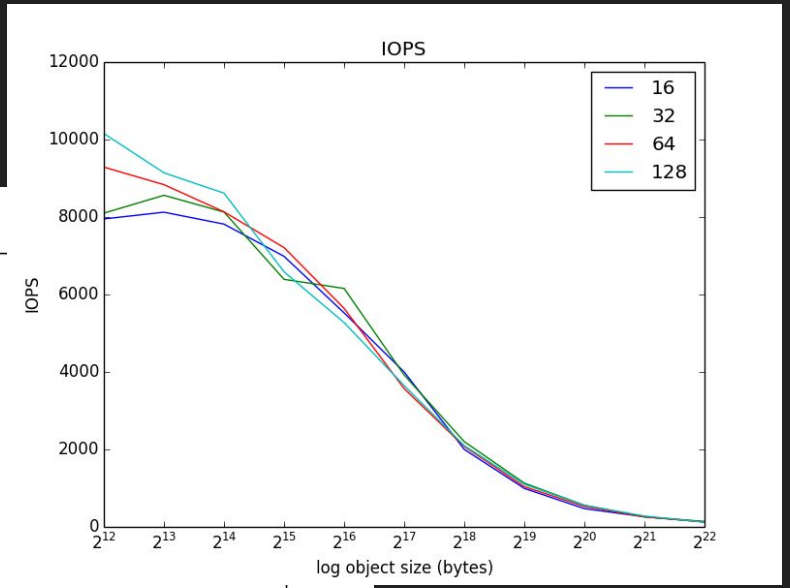




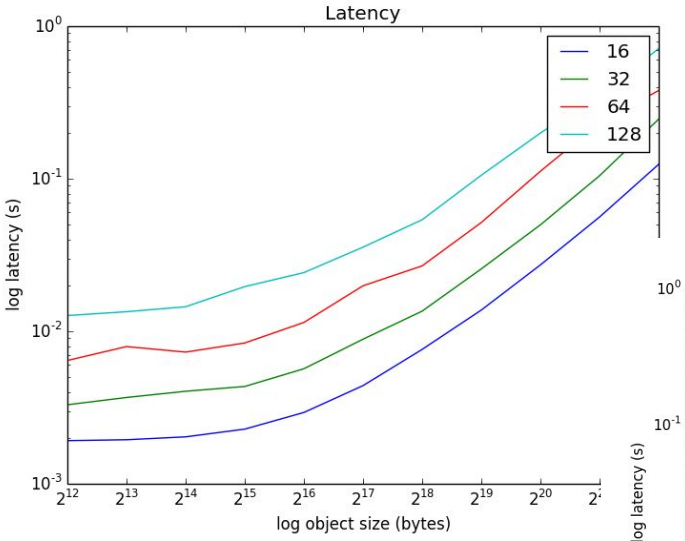
IOPS n = 1,2,3

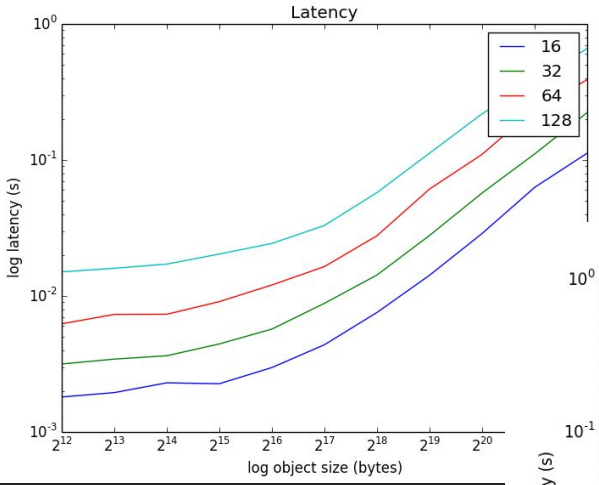

IOPS
12000
10000
8000
6000
4000
2000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

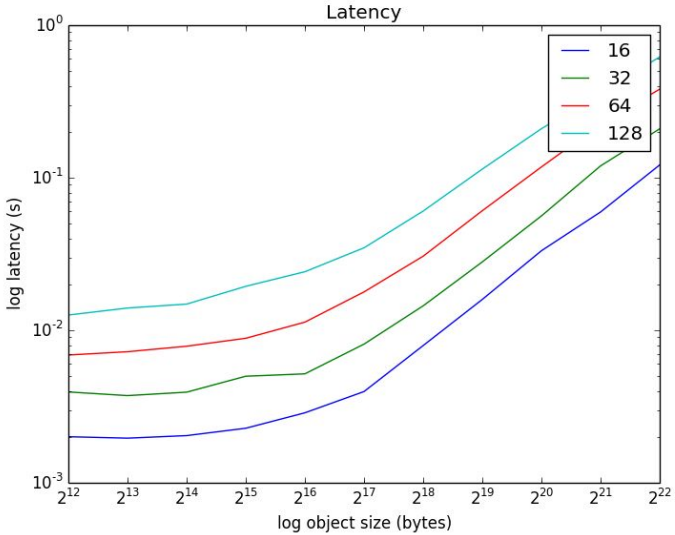

IOPS
12000
10000
8000
6000
4000
2000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)


IOPS
16
32
64
128
12000
10000
8000
6000
4000
2000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

Latency n = 1,2,3

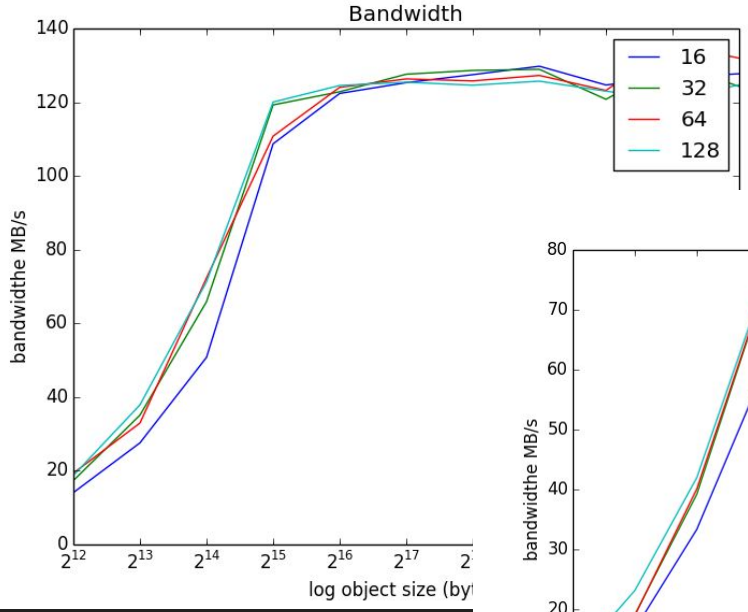


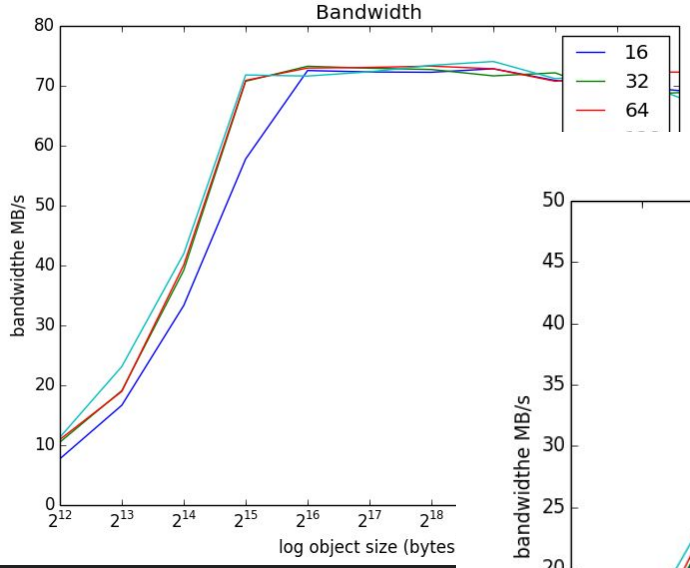


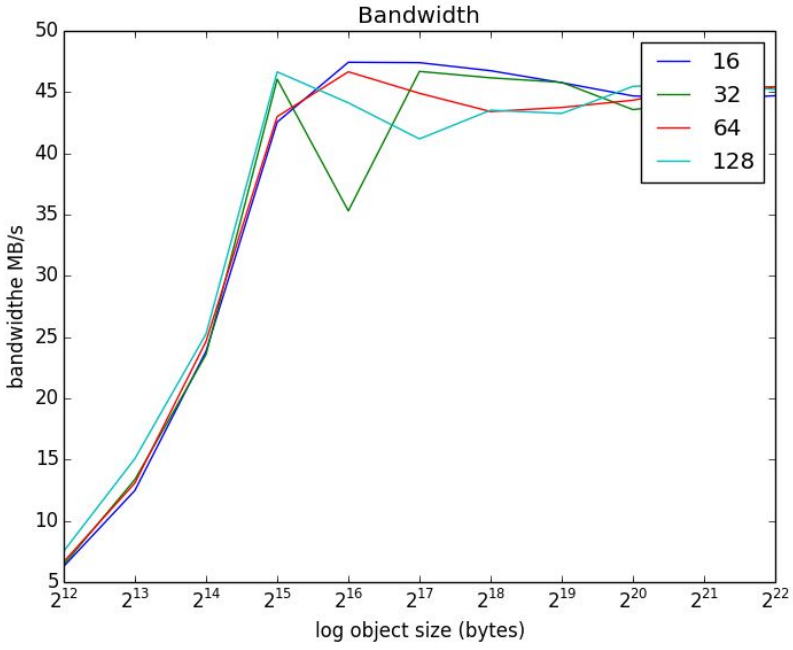


SSD Results - Write

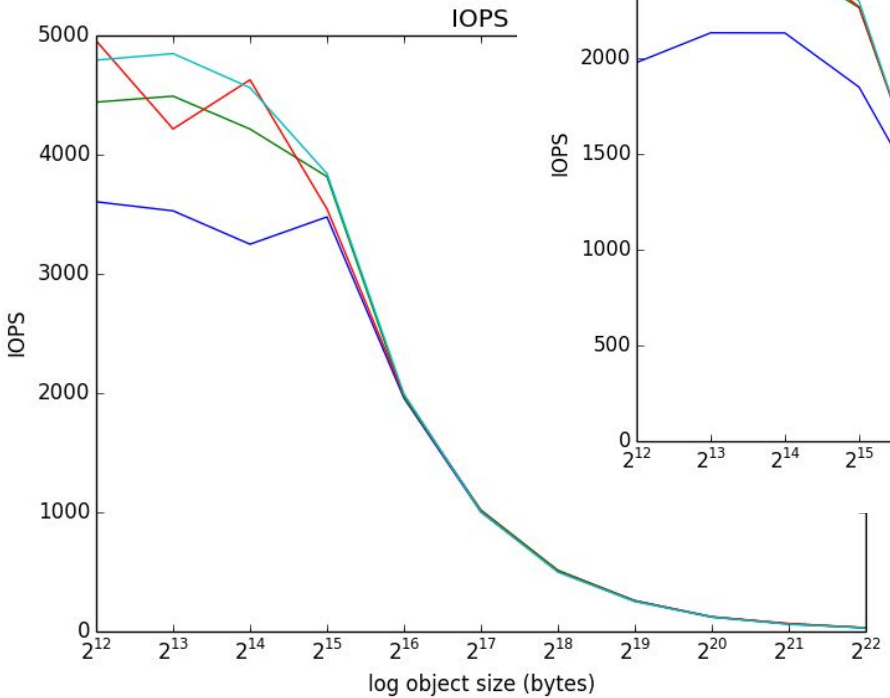
Bandwidth n =1,2,3

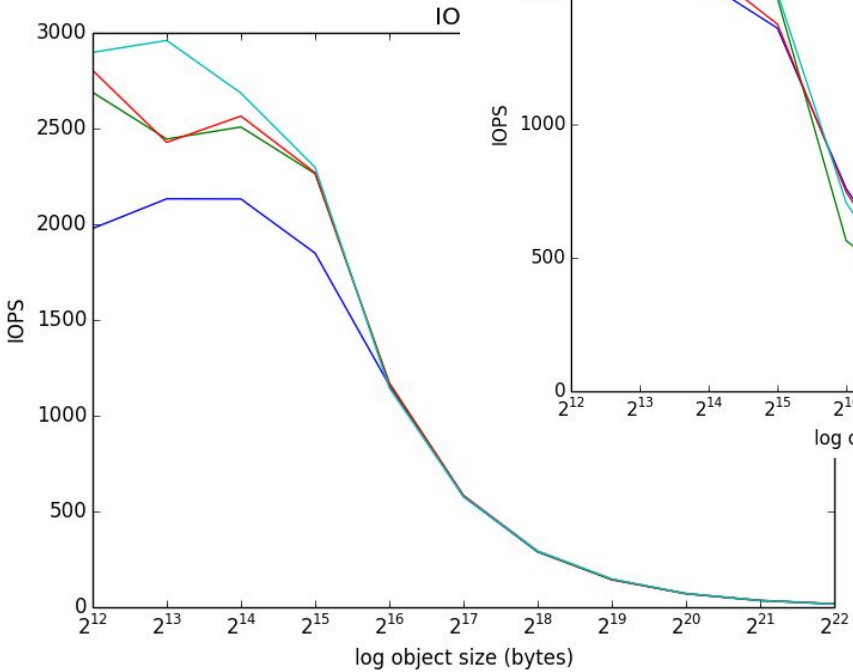


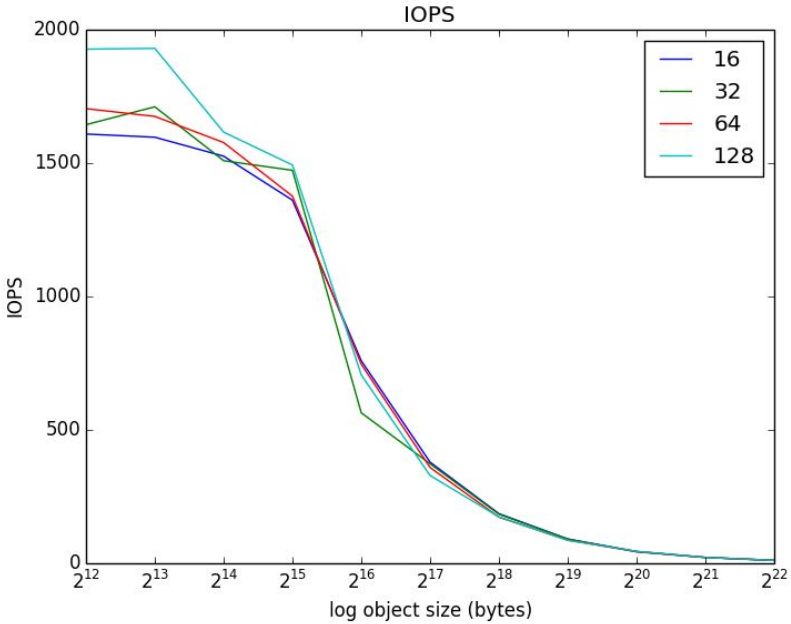




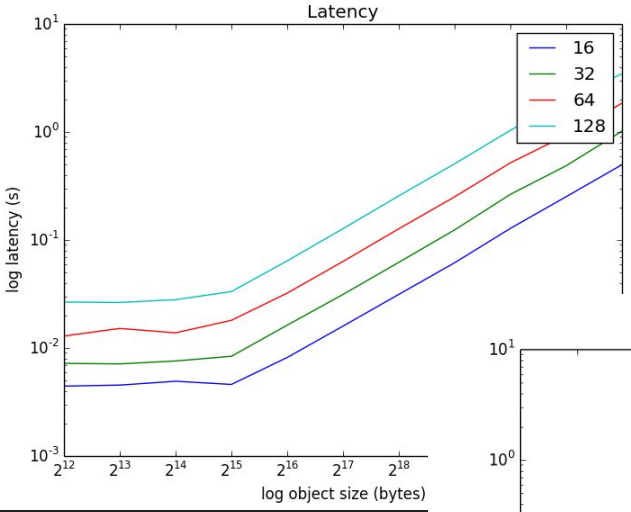
IOPS n = 1,2,3

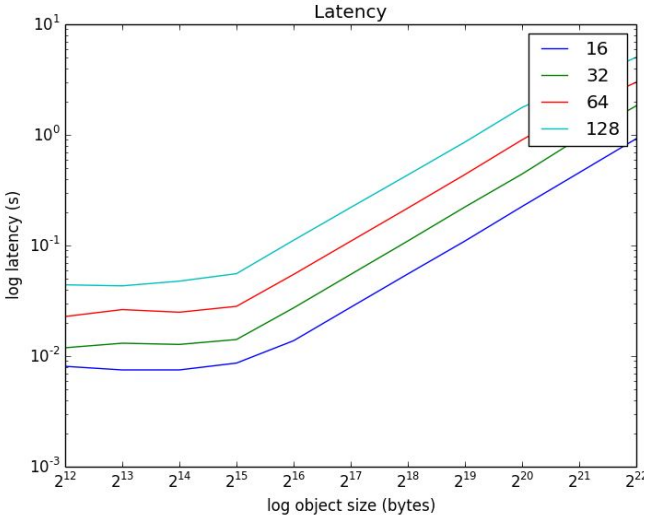

IOPS
5000
4000
3000
2000
1000
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

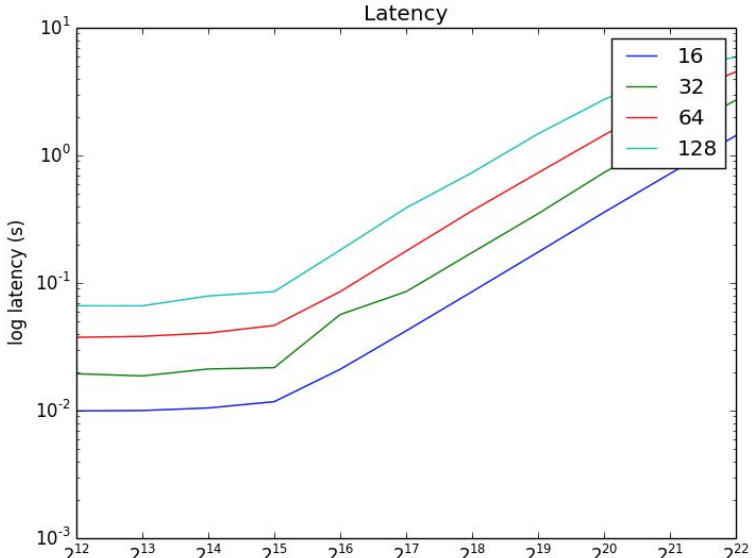

IO
3000
2500
2000
1500
1000
500
0
IOPS
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)


IOPS
2000
1500
1000
500
0
IOPS
16
32
64
128
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

Latency n = 1,2,3,

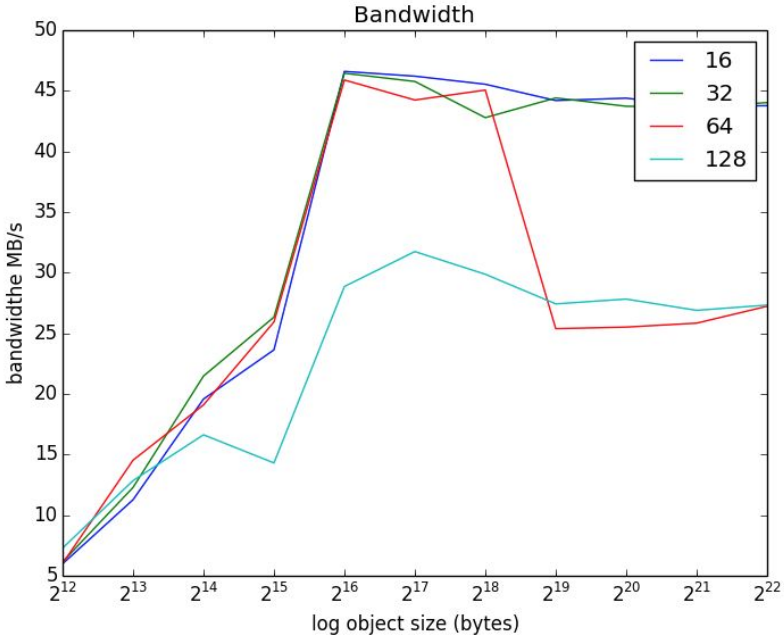


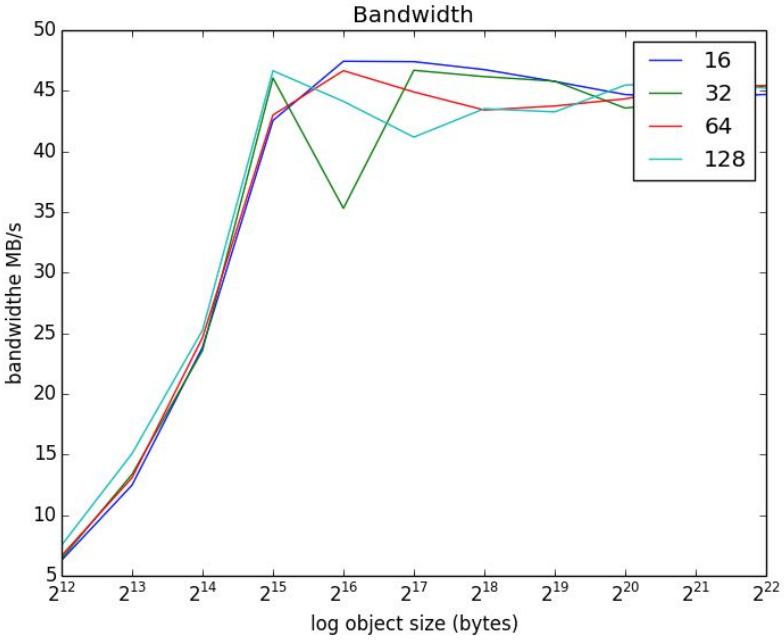




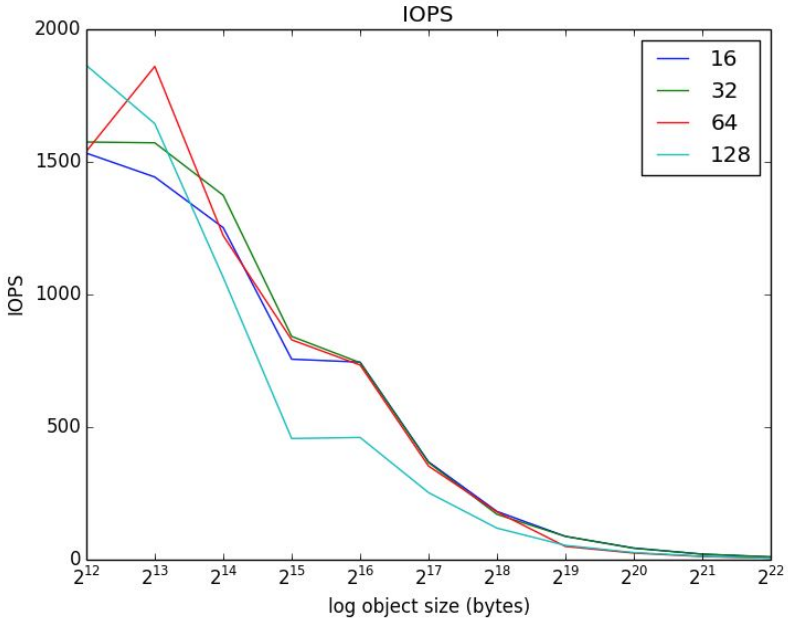
HDD vs SSD - Write

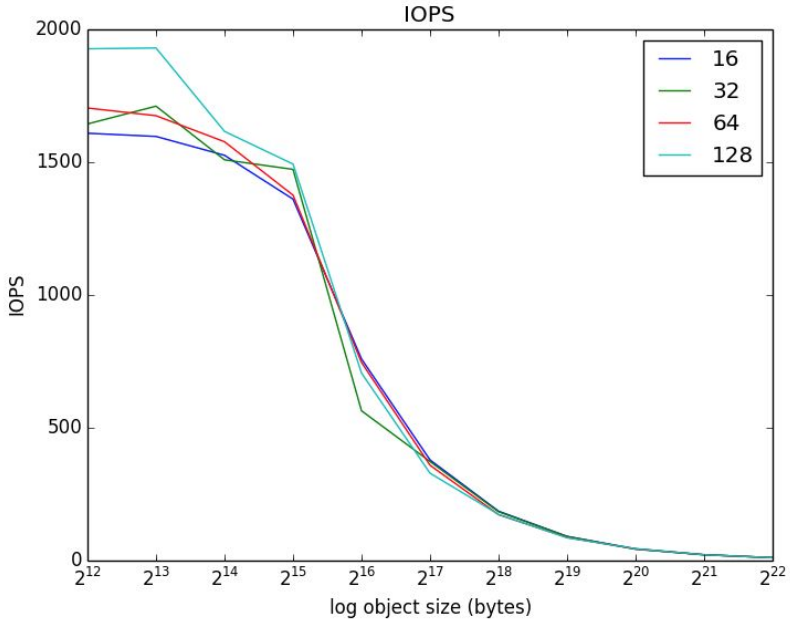
Bandwidth n =3



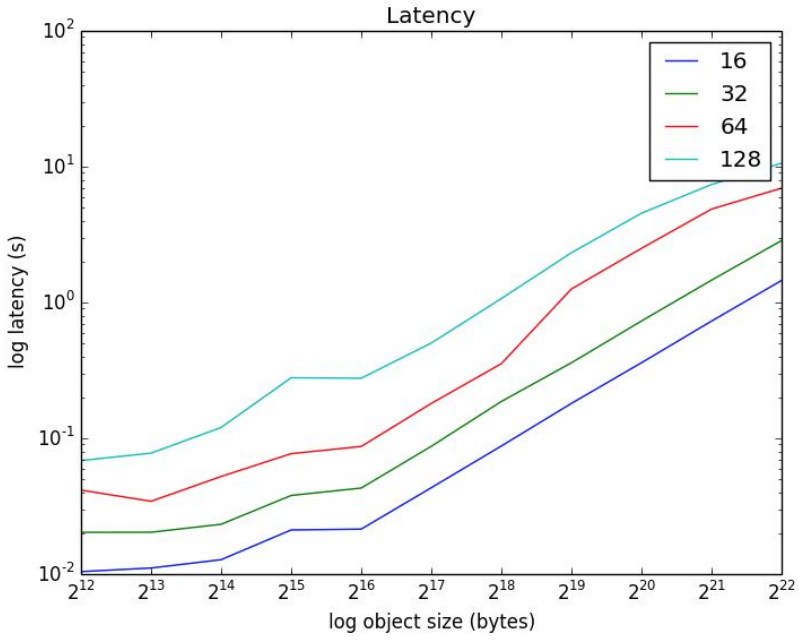


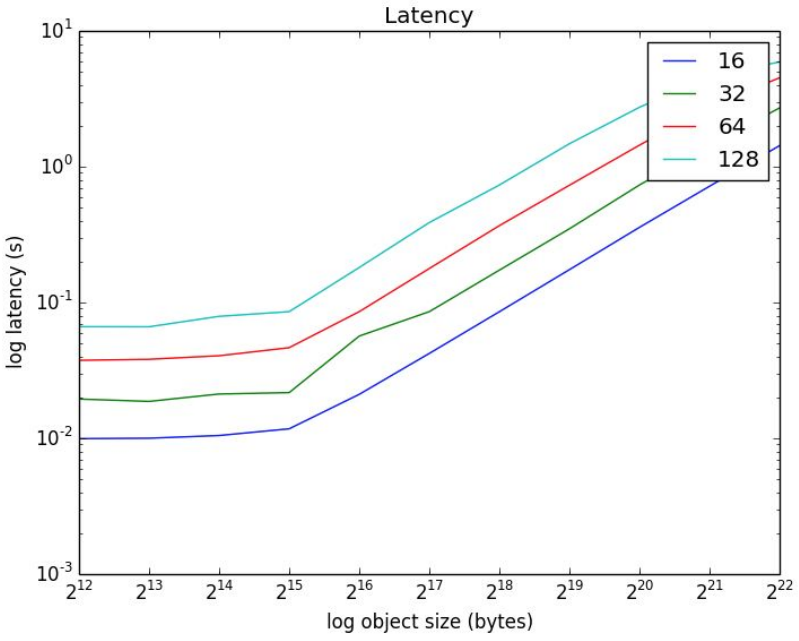
IOPS n=3


IOPS
2000
1500
1000
500
0
IOPS
16
32
64
128
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)


IOPS
2000
1500
1000
500
0
IOPS
16
32
64
128
2^{12} 2^{13} 2^{14} 2^{15} 2^{16} 2^{17} 2^{18} 2^{19} 2^{20} 2^{21} 2^{22}
log object size (bytes)

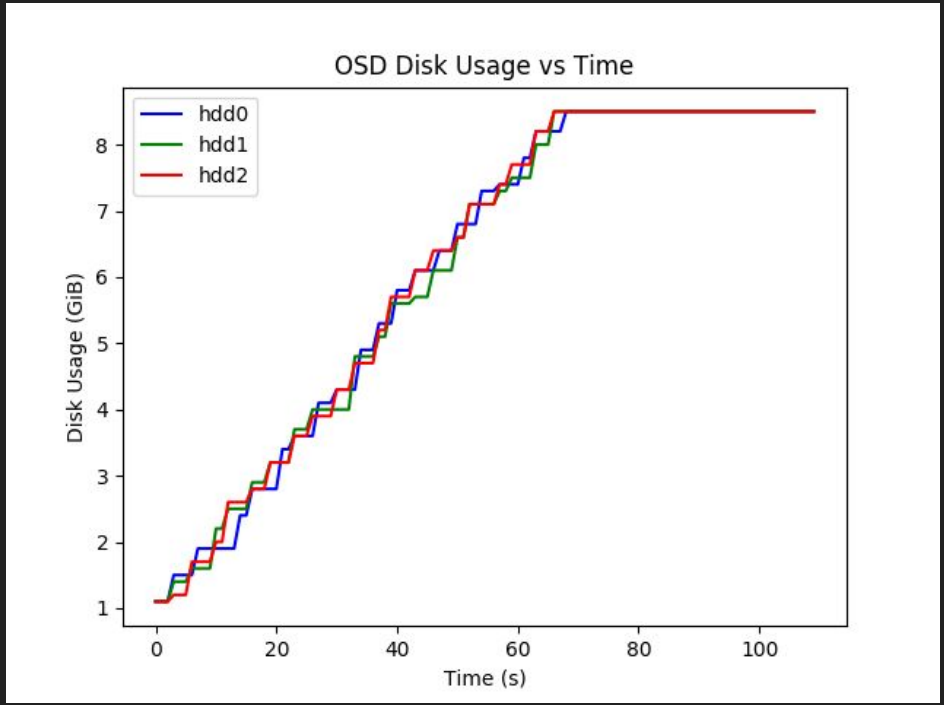
Latency n = 3

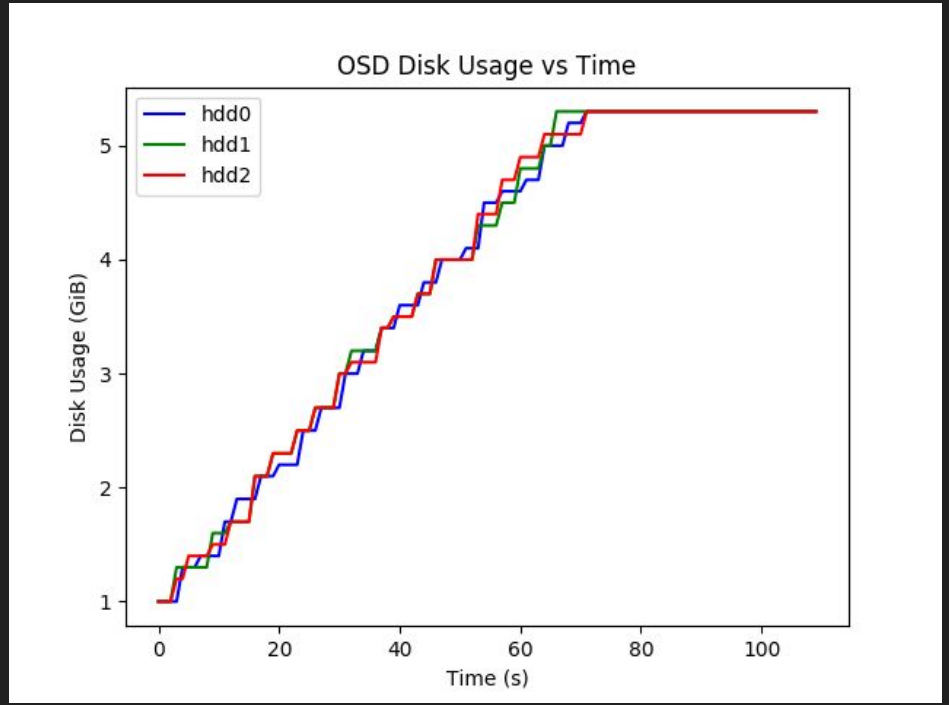

Latency
16
32
64
128
log latency (s)
log object size (bytes)


Latency
16
32
64
128
log latency (s)
log object size (bytes)

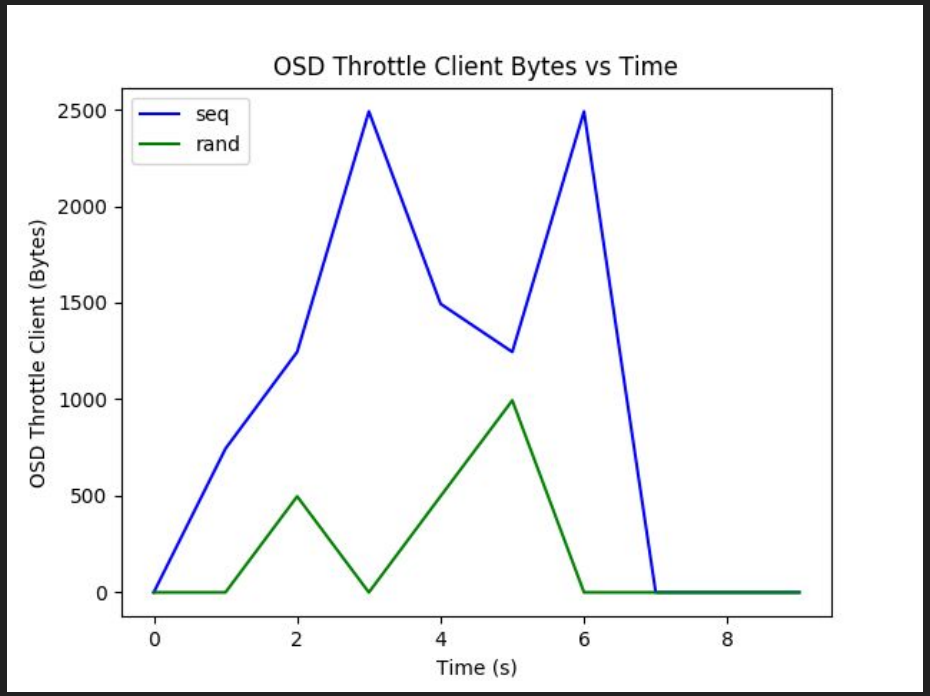
Workload Balancing of Ceph

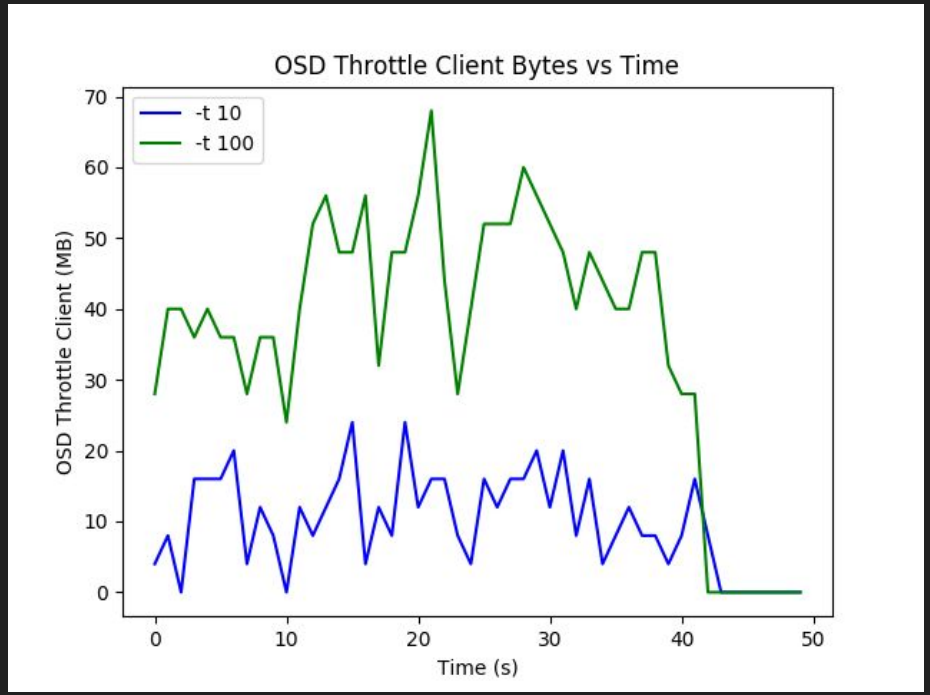
Pretty good at homogeneous environment (3 HDD with 200G)





OSD Throttle Bytes



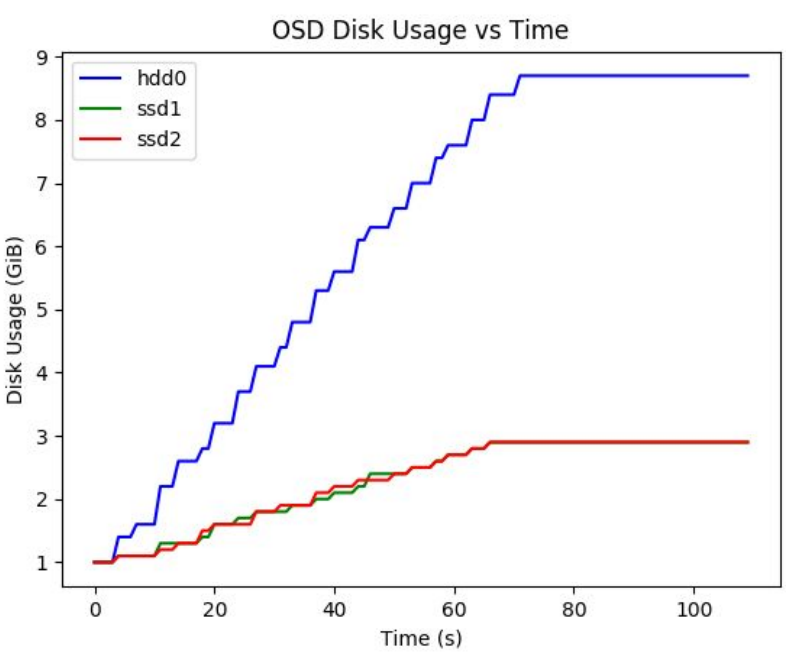


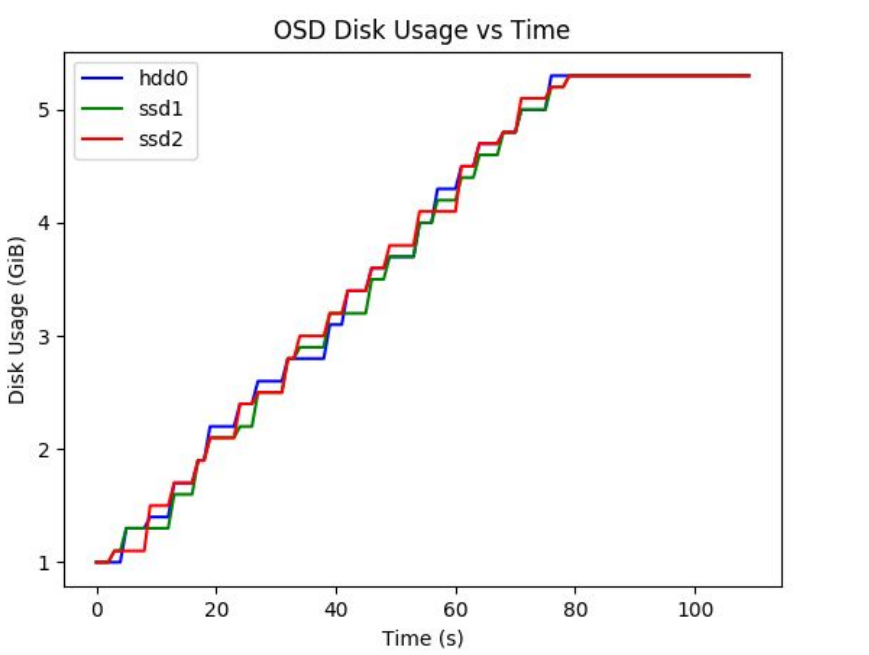
Ceph Placement Groups

- Each object will be map to a placement group
- Placement groups are assigned to OSDs using CRUSH
- Can use ceph pg dump_stuck to see bad pgs
  - Inactive: cannot process reads or writes because they are waiting for an OSD with the most up-to-date data to come up and in
  - Unclean: Placement groups contain objects that are not replicated the desired number of times. They should be recovering
  - Stale Placement groups are in an unknown state - the OSDs that host them have not reported to the monitor cluster in a while (configured by mon_osd_report_timeout)

Workload Balancing in Heterogeneous Environment

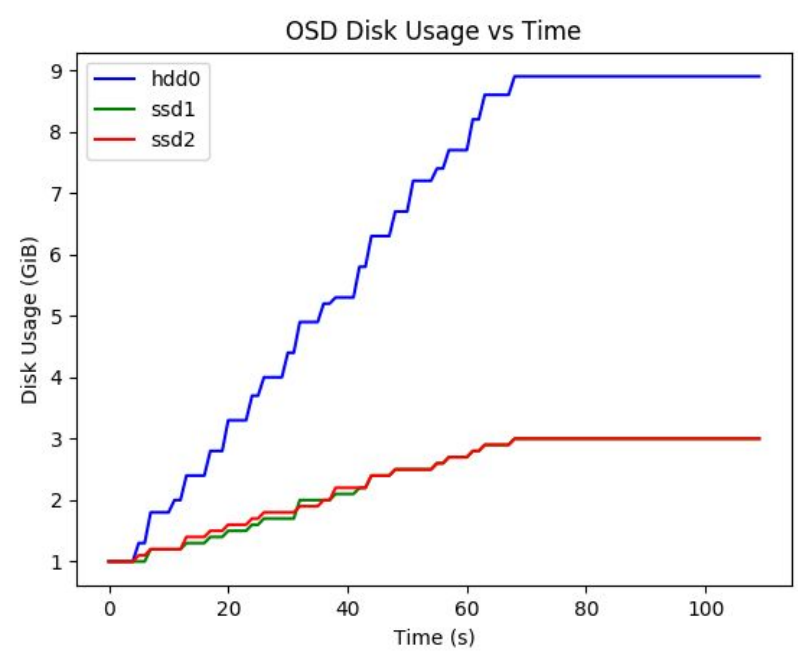
Note that SSD is 100G while HDD is 200G

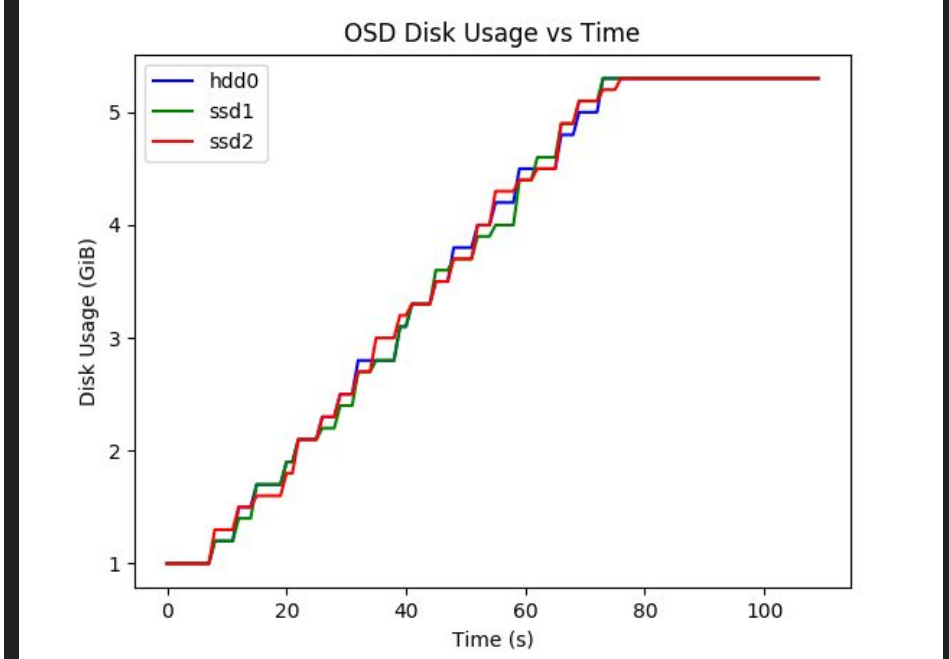




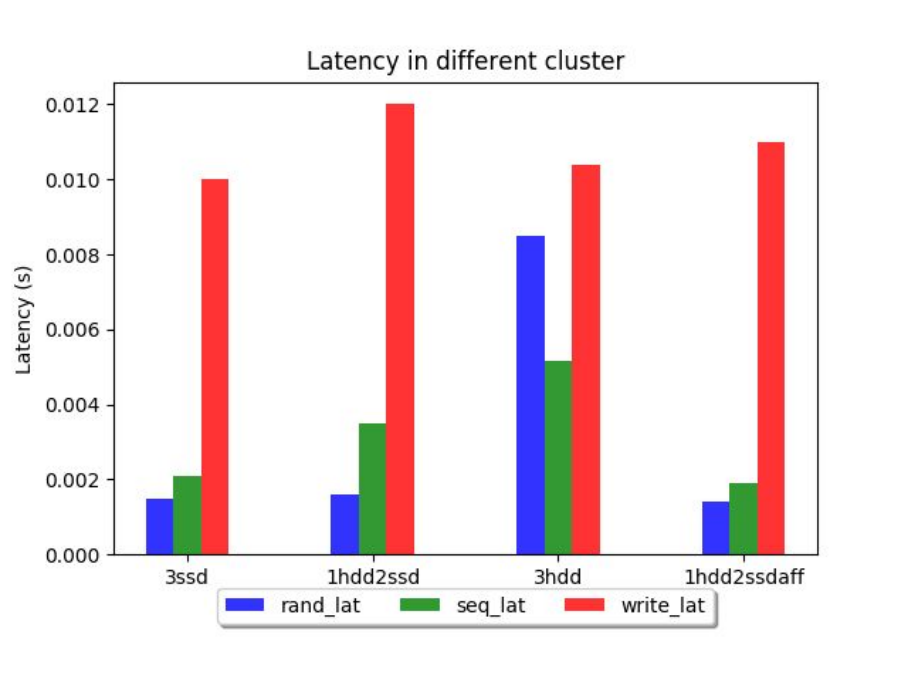
Workload Balancing in Heterogeneous Environment

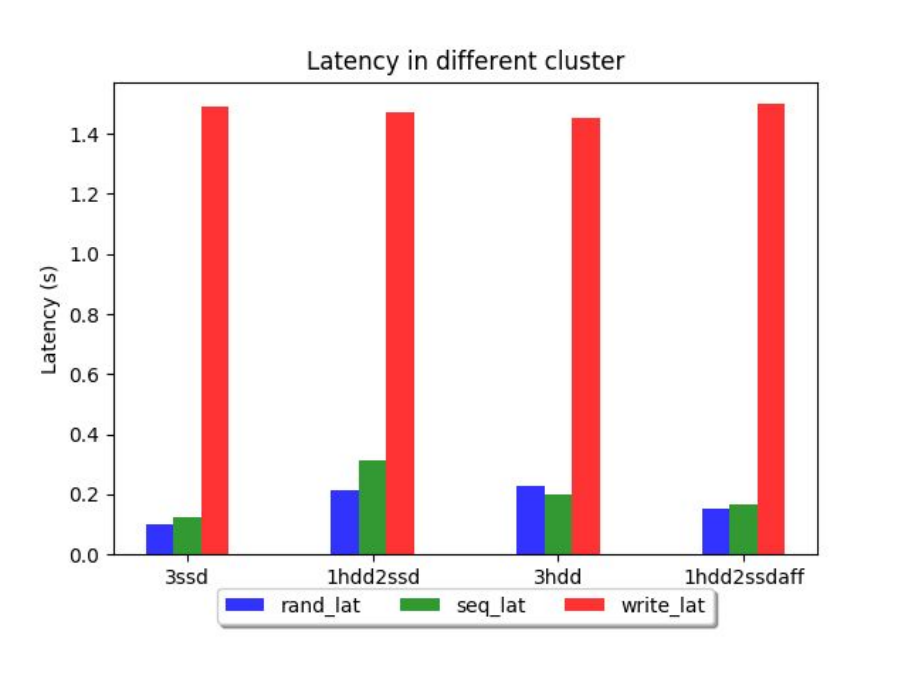
If we set affinity of HDD to 0, i.e. it won't be primary for some placement group



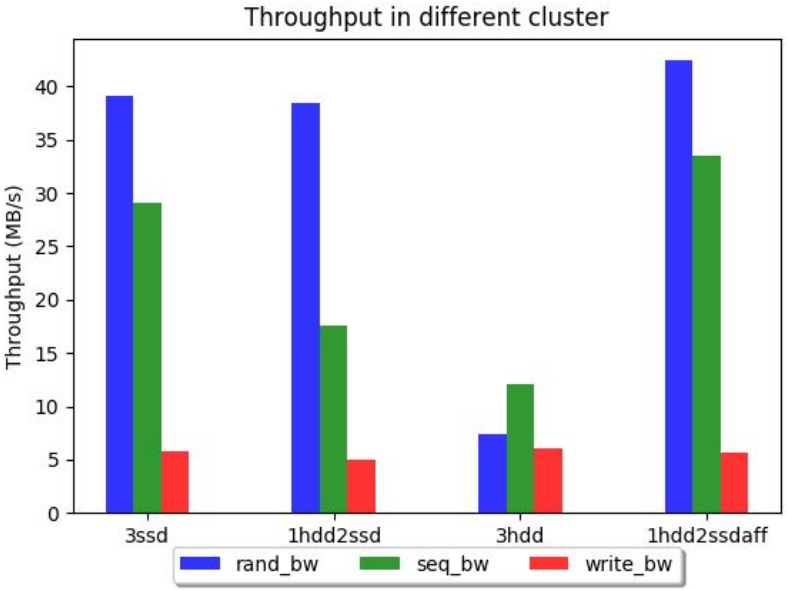


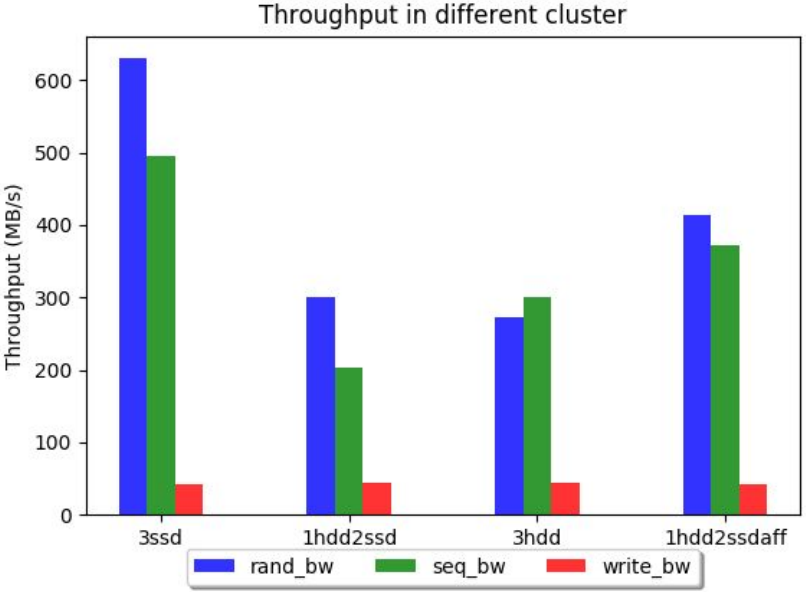
Latency Comparison





Throughput Comparison





Problems

1. 3 HDD small object size, read performance is very bad. Why?
2. When size and min_size of OSD pool grows, rand read bandwidth and latency gets worse, Why?
3. A good way to change system config, right now we just remove all and install all again.

Reference links

- Benchmark: https://tracker.ceph.com/projects/ceph/wiki/Benchmark_Ceph_Cluster_Performance
- Some interesting metrics: https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/1.3/html/administration_guide/performance_counters#mon-throttle-table
- Primary Affinity: https://ceph.com/geen-categorie/ceph-get-the-best-of-your-ssd-with-primary-affinity/
- Pools: http://docs.ceph.com/docs/jewel/rados/operations/pools/
-