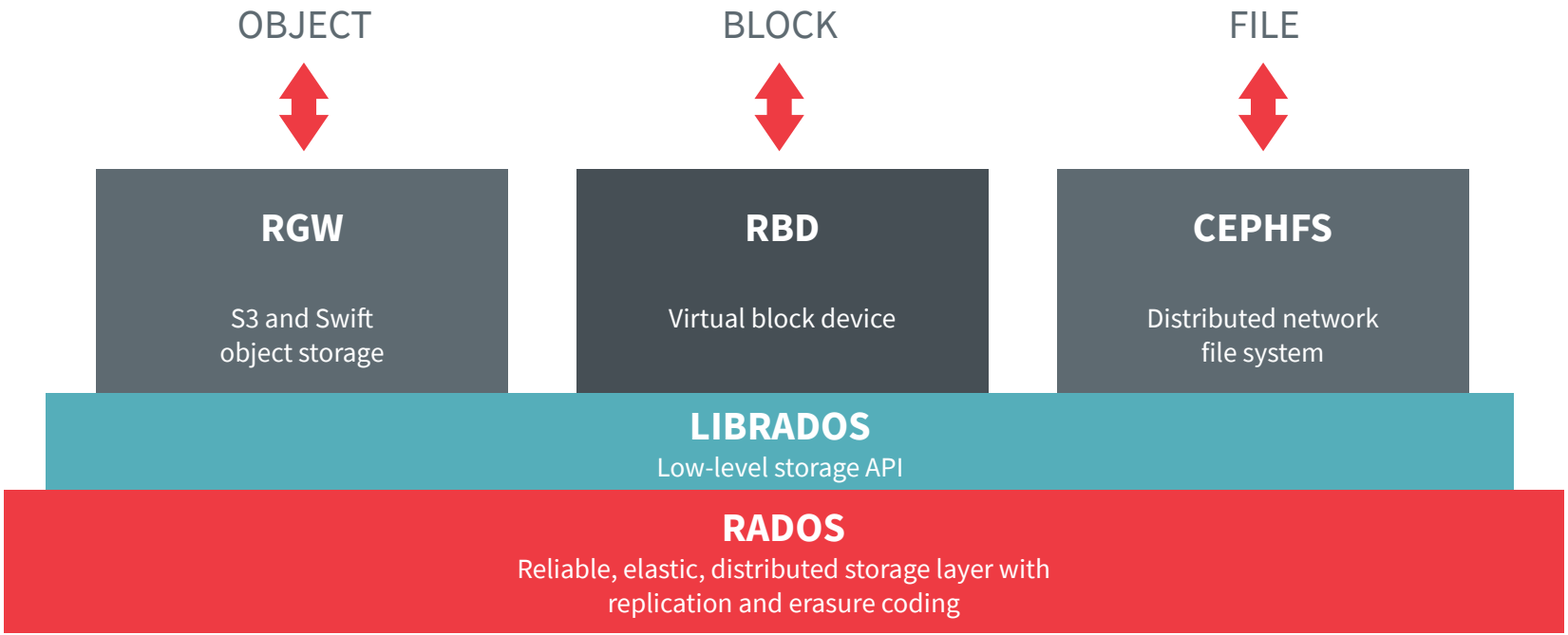
Update on Crimson

The Seatarized Ceph

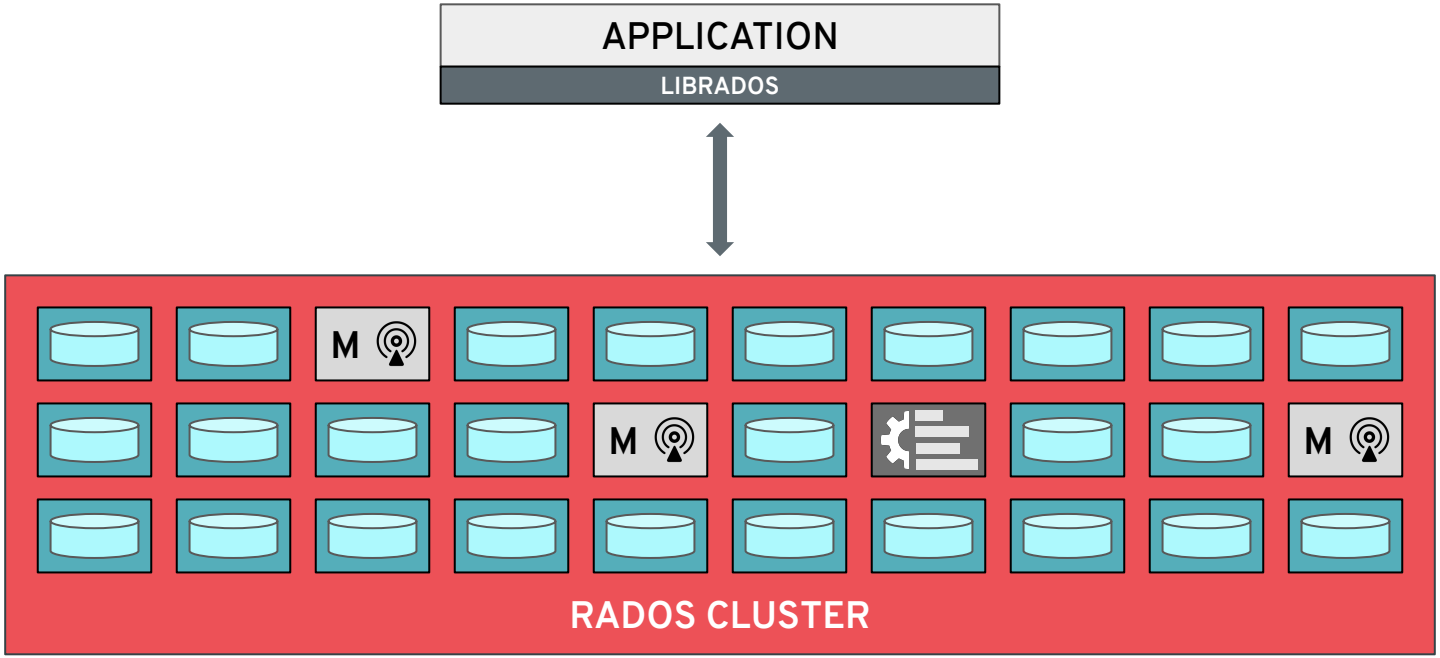
kchai@redhat.com

Seastar Summit 2019

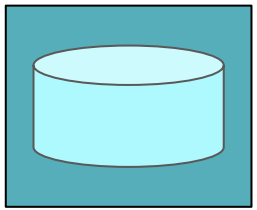
A unified storage system



RADOS -- The Cluster



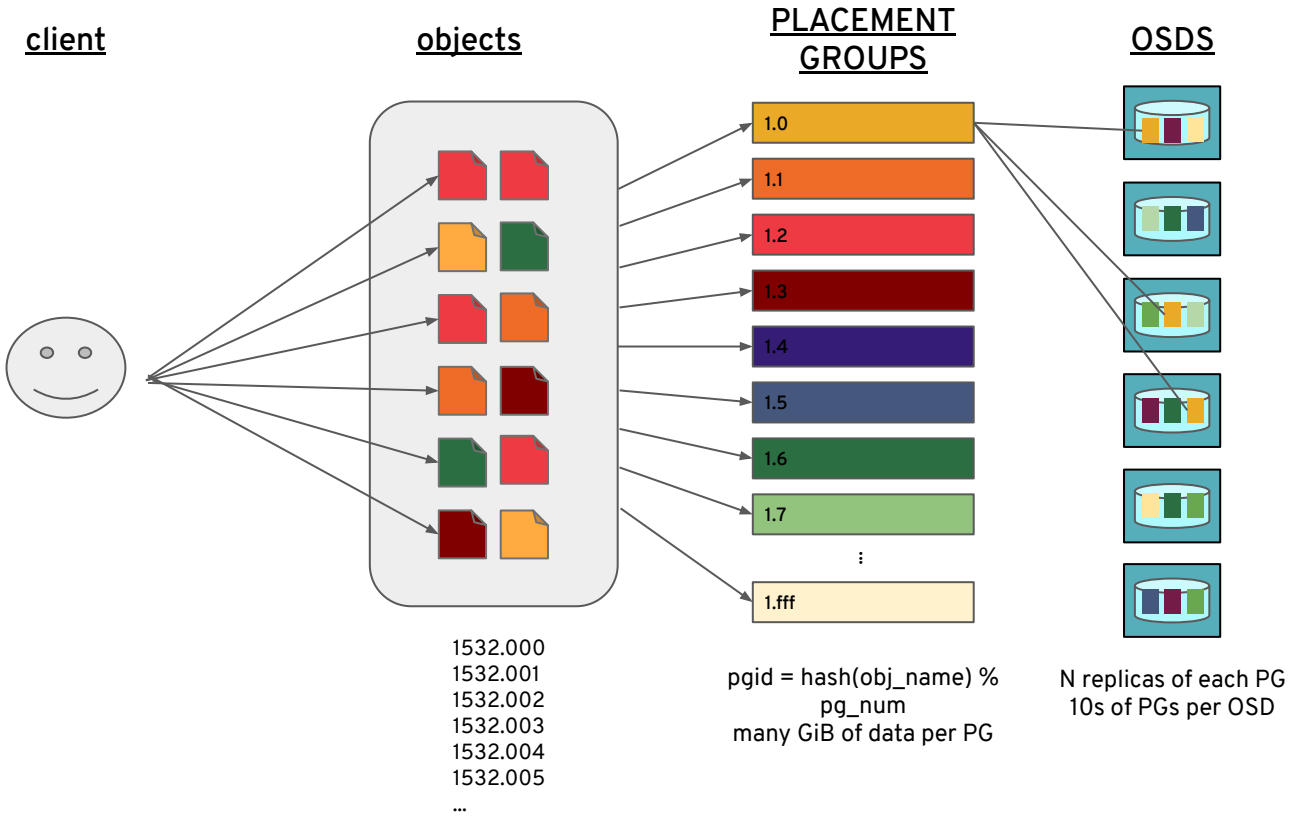
OSD (Object storage Daemon)



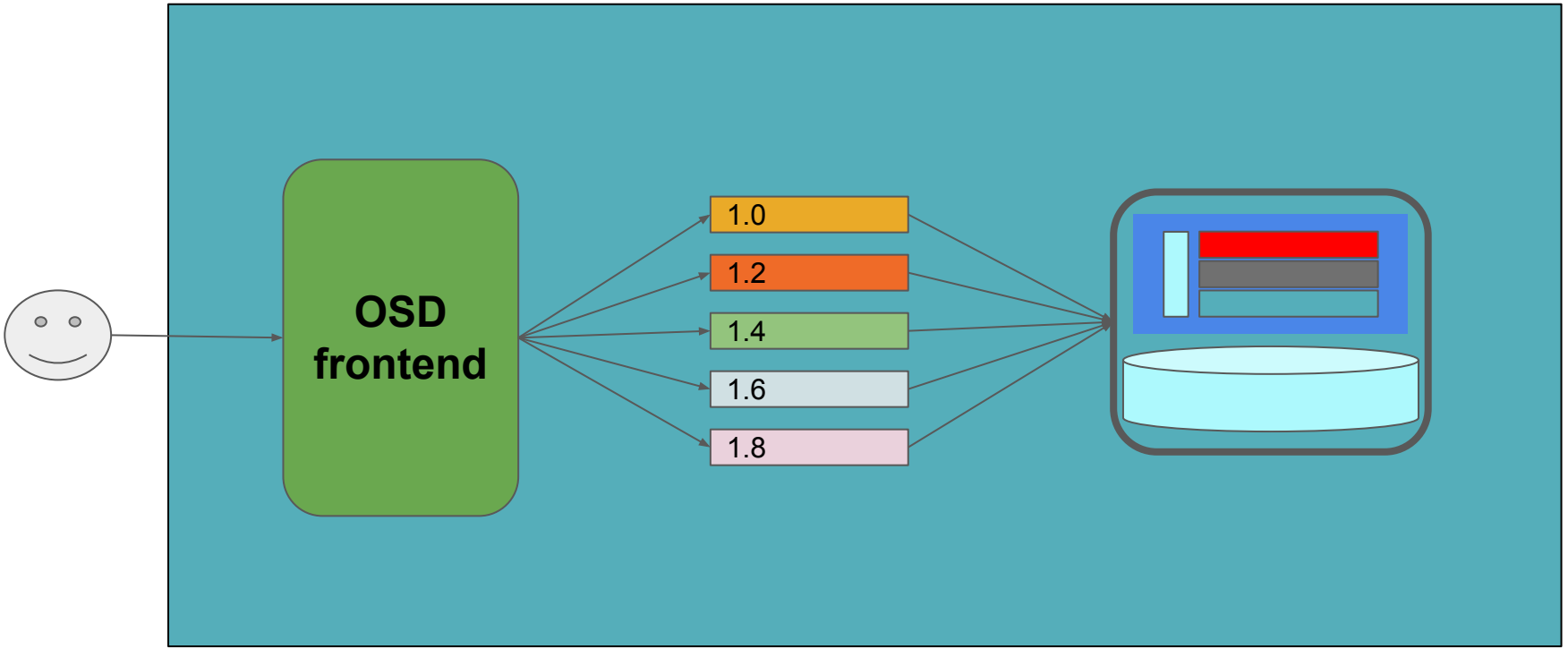
ceph-osd (crimson)

- Stores data on an HDD or SSD
- Services client IO requests
- Cooperatively peers, replicates, rebalances data
- Reports stats to manager daemons
- 10s-1000s per cluster

PG (placement groups)



A closer look



Crimson - a **faster** OSD

- Less overhead
  - Bypass kernel
  - Zero memcpy
  - Less context switches
- Understands modern storage devices

share something => share nothing

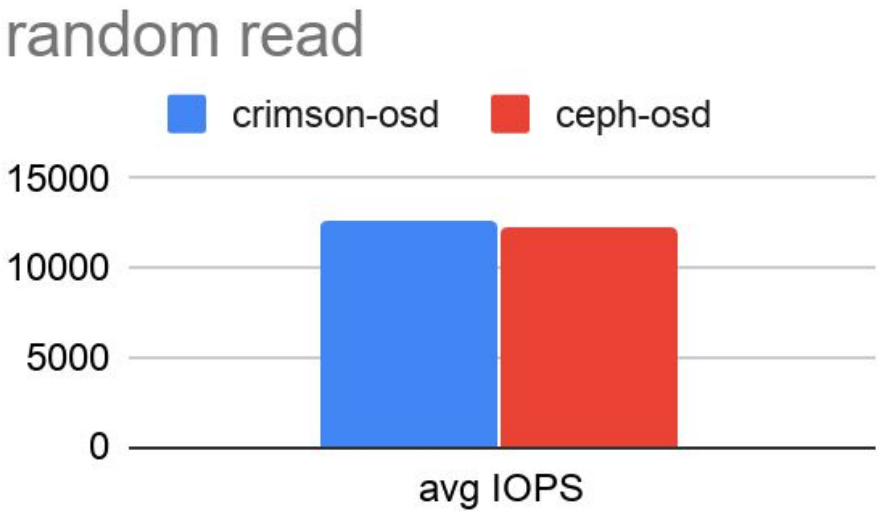
What we imaged:

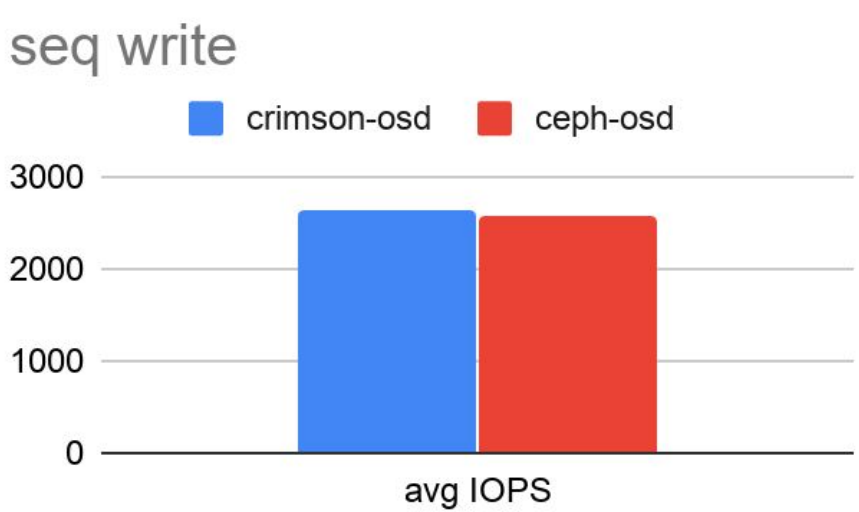
- Multi-reactor OSD
- Shared connections
  - Connections to manager daemons
  - Connections to peer OSDs
  - Connections to clients
- Shared io queue
- Shared metadata
  - Knowledge about the cluster

What we have now:

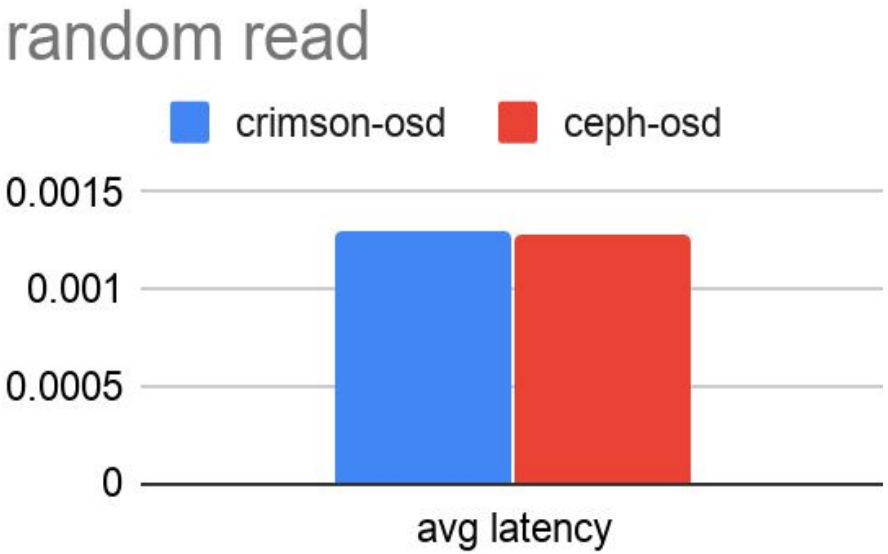
- Single threaded OSD
- Fully connected network
- Monitor’s load increases

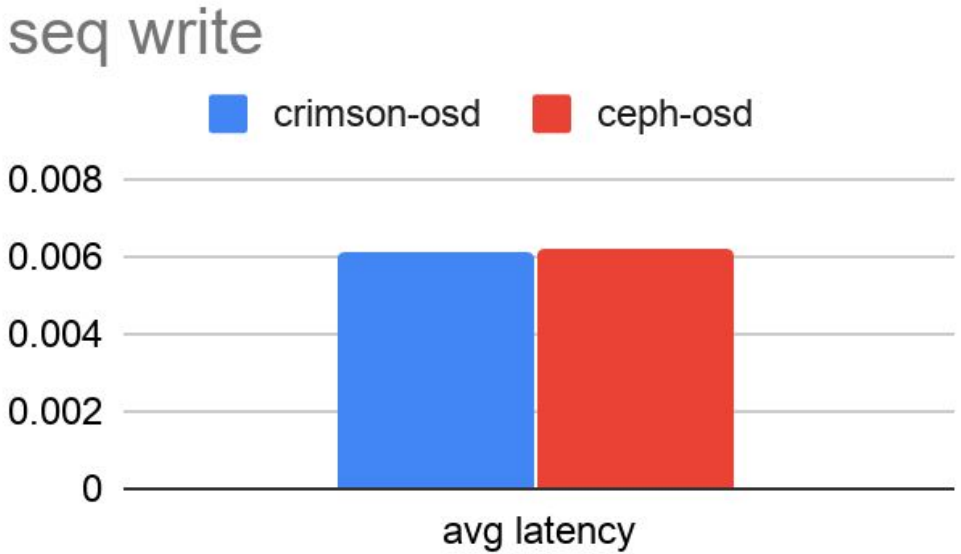
Average IOPS



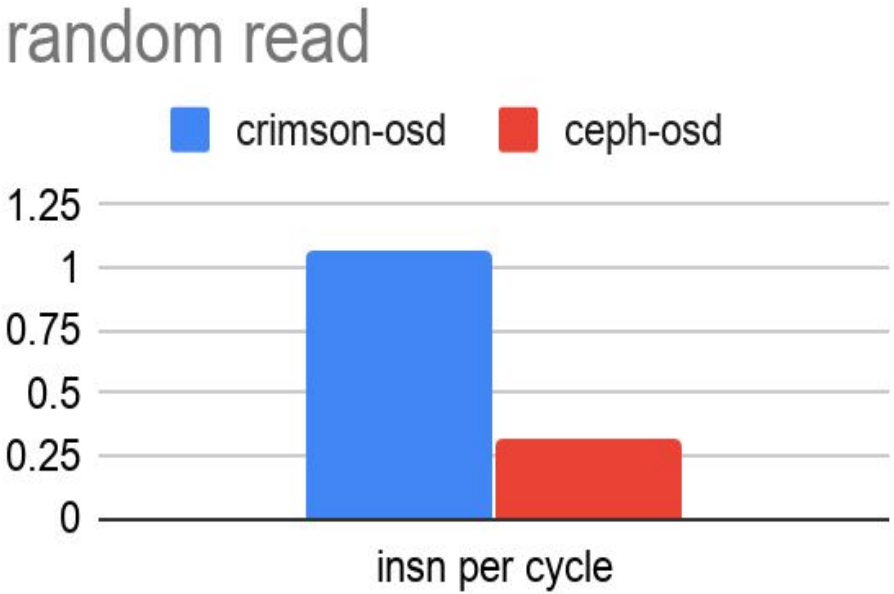


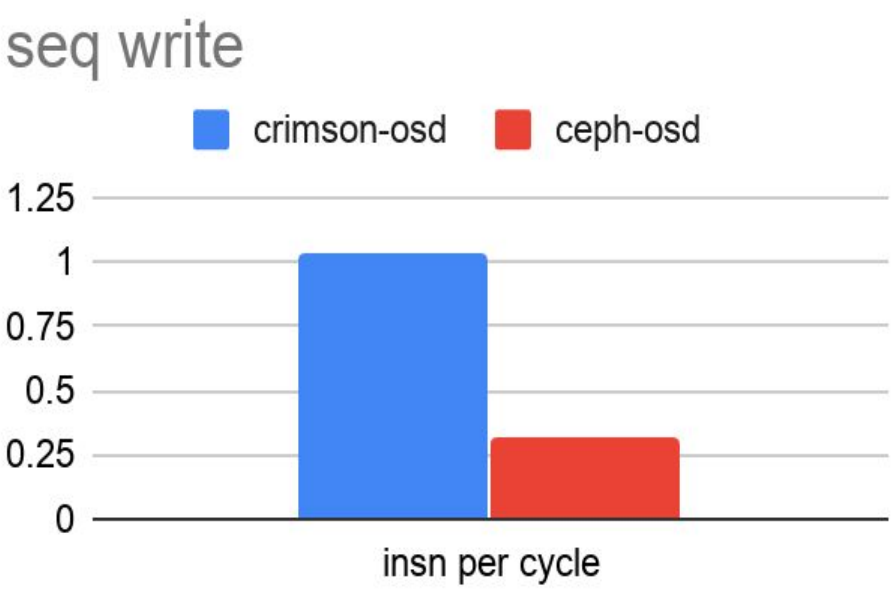
Average latency



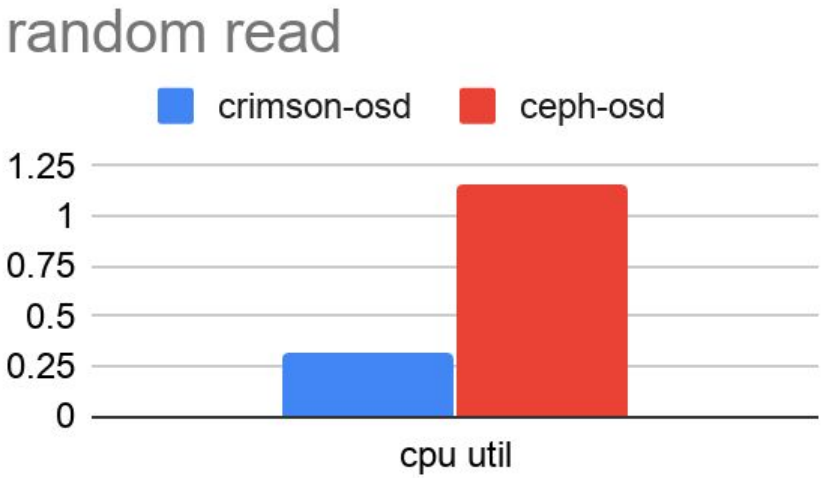


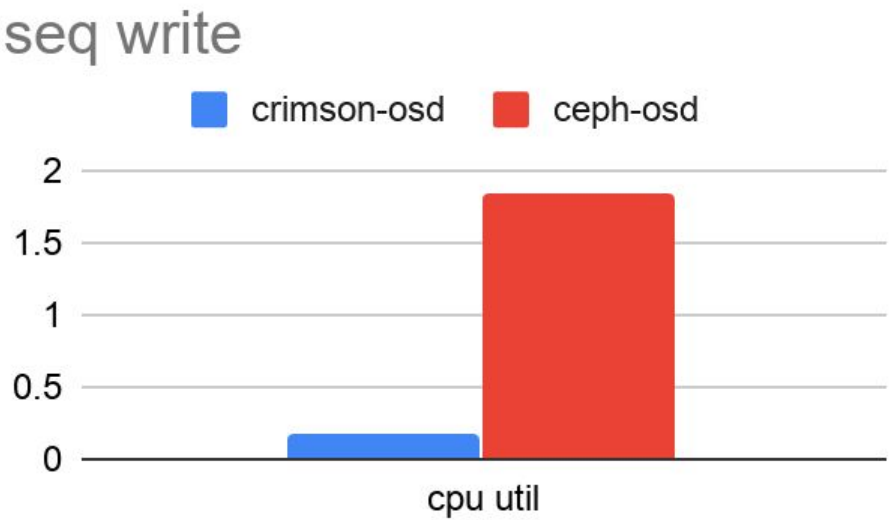
Instructions per cycle



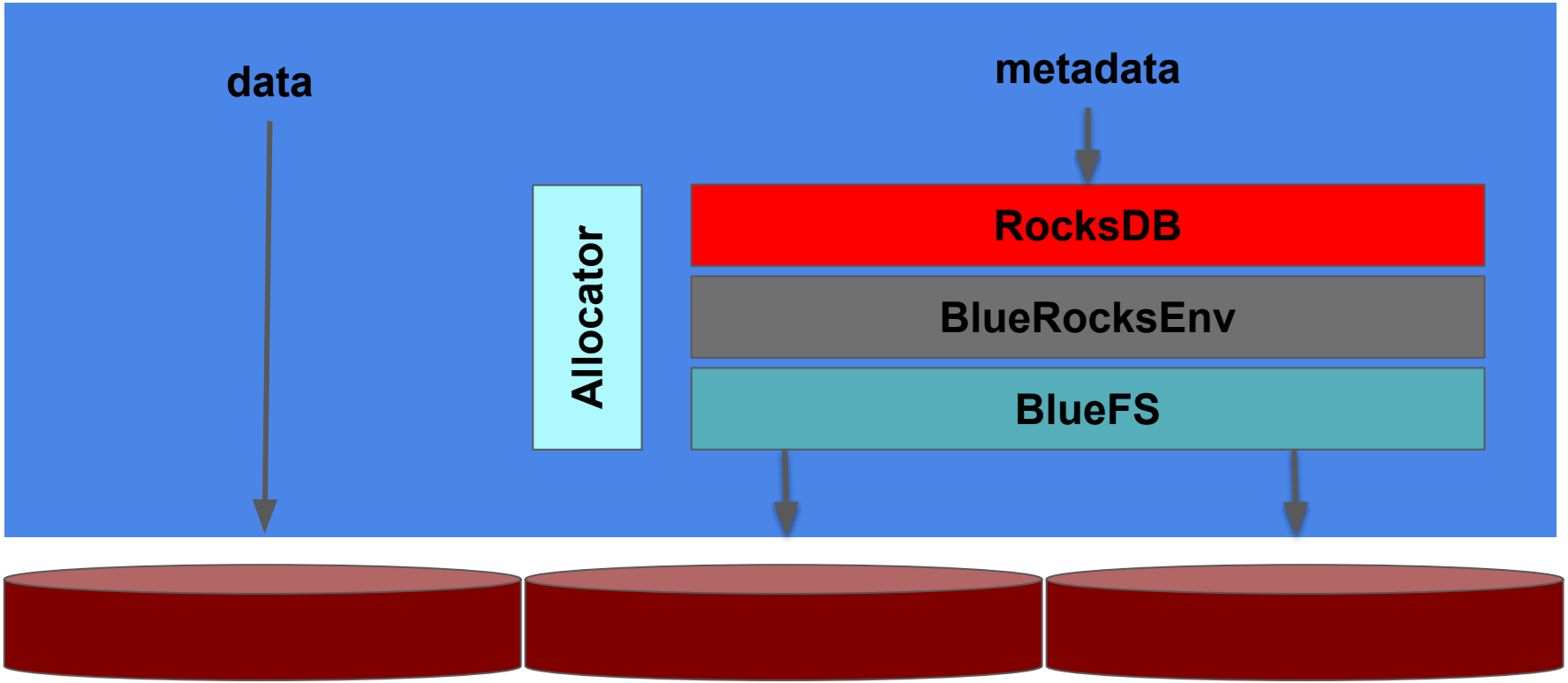


CPU util

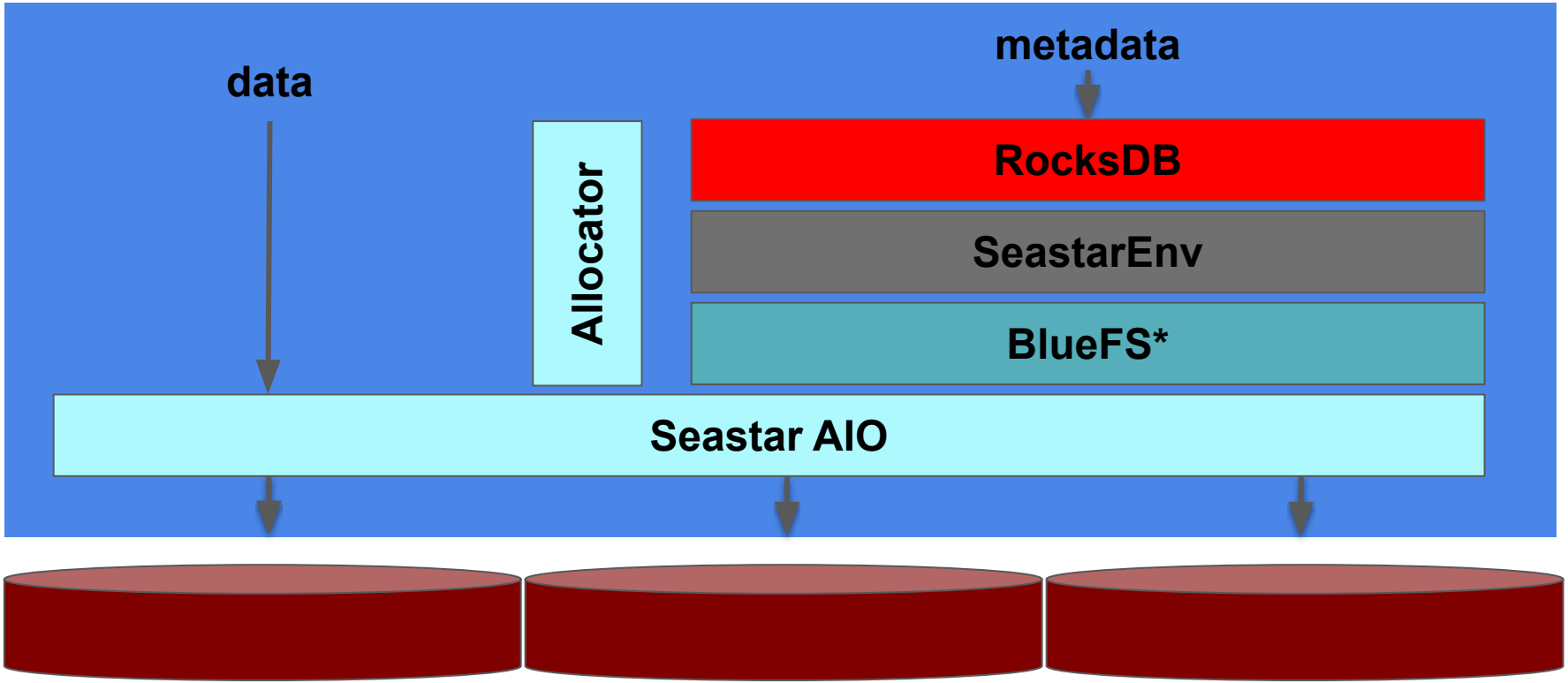

random read
crimson-osd
ceph-osd
1.25
1
0.75
0.5
0.25
0
cpu util


seq write
crimson-osd
ceph-osd
2
1.5
1
0.5
0
cpu util

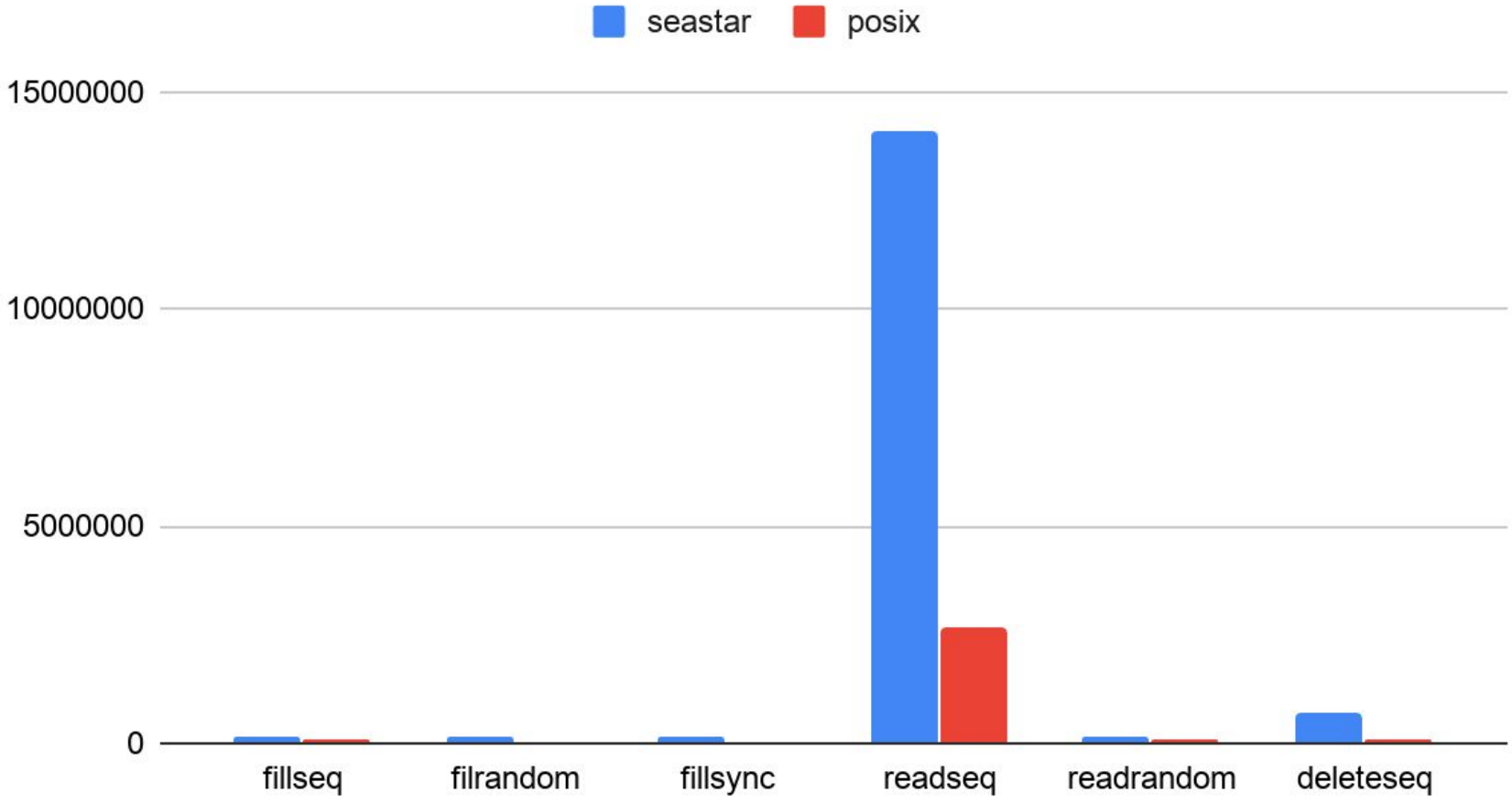
bluestore



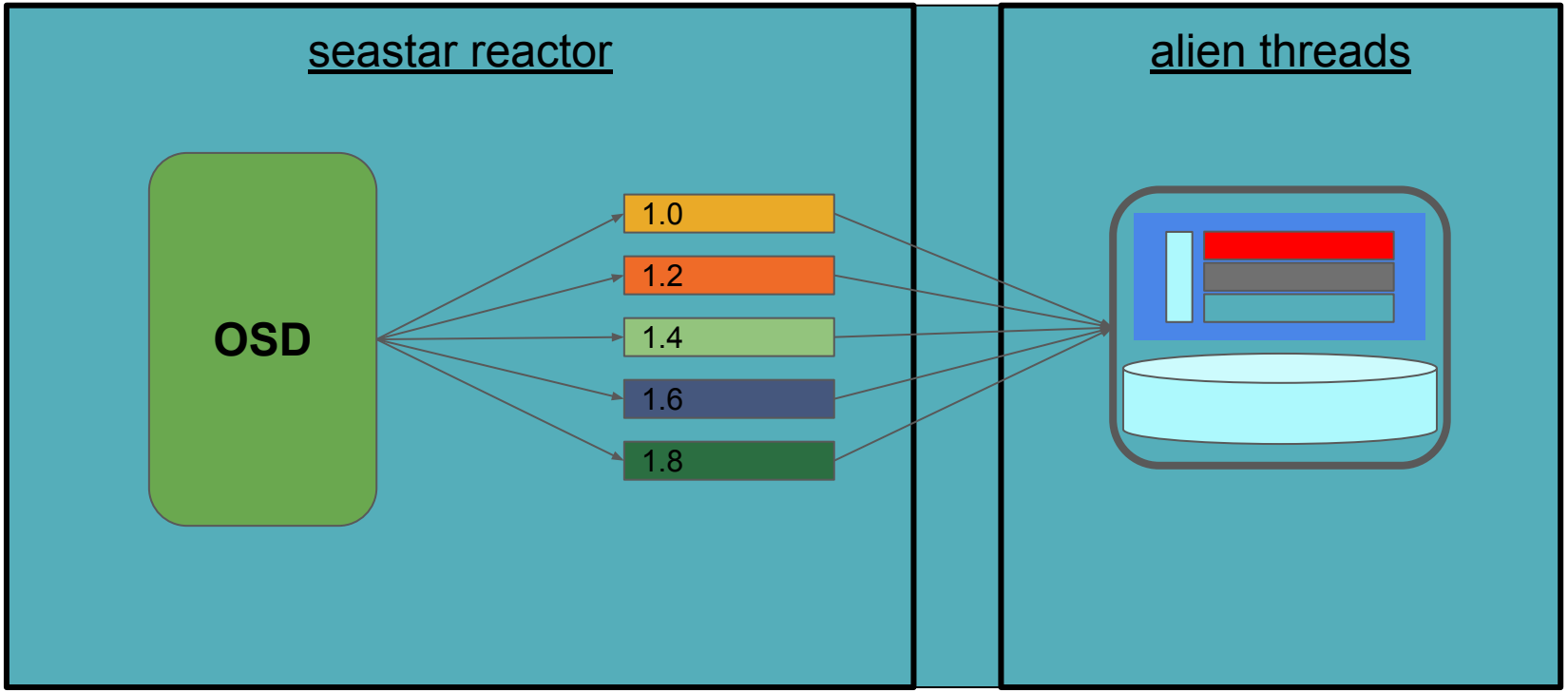
Seastarized bluestore



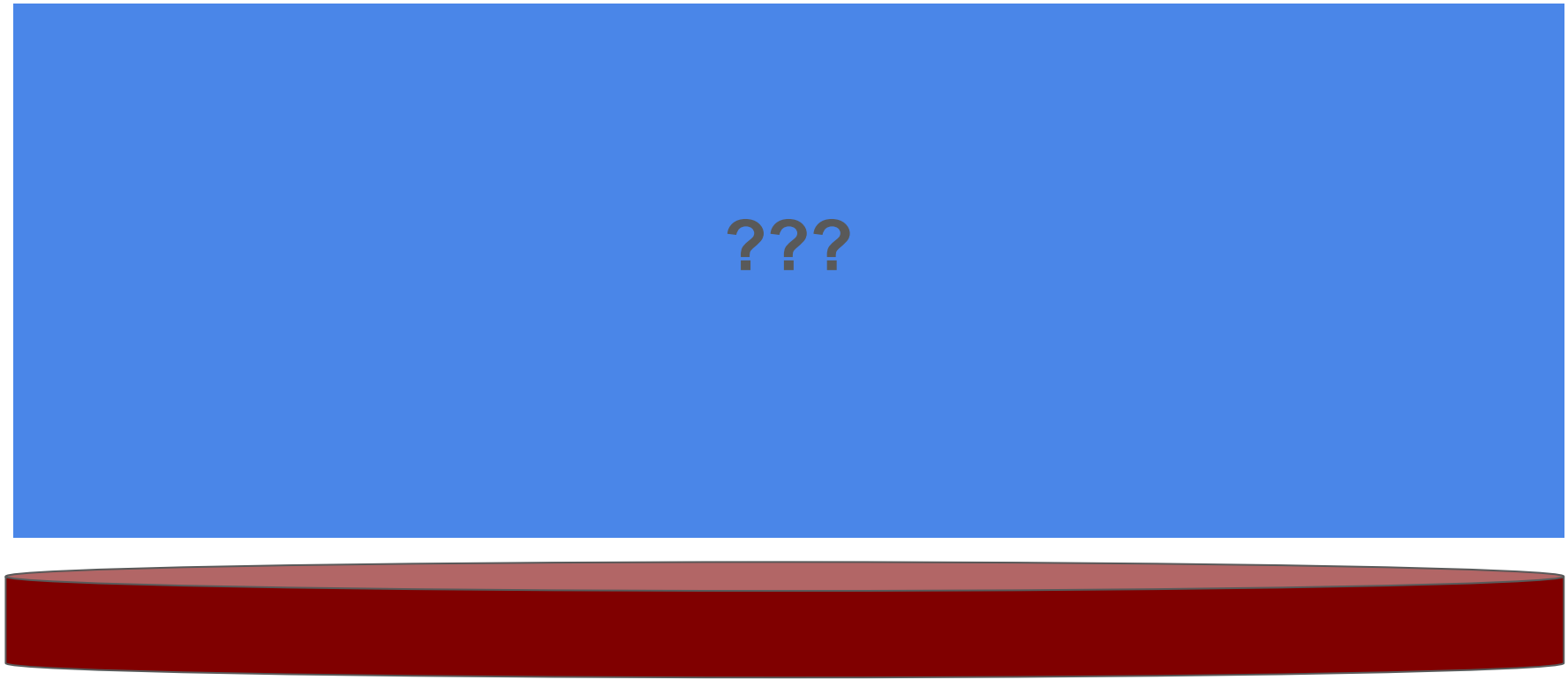
db_bench (ops/secs)



Alienized bluestore



SeaStore



Q & A