Rebuilding a Faster OSD with Future
– the crimson project update

Radoslaw Zarzynski & Kefu Chai
Today’s storage

4K Read IOPS

- NVMe SSD: 500000
- SAS SSD: 190000
- SATA SSD / 6Gbit/s: 10000
- SAS HDD / 10k: 150
- SATA HDD / 7200: 75

* Data collected from https://en.wikipedia.org/wiki/IOPS
Today’s ethernet

Evolution of Ethernet

2010

2003 10000

1998 1000

1995 100

1980 10

* the date of creation of corresponding standard/specification.
Today’s CPU

42 Years of Microprocessor Trend Data

- Transistors (thousands)
- Single-Thread Performance (SpecINT x 10^3)
- Frequency (MHz)
- Typical Power (Watts)
- Number of Logical Cores

Year


Problems with OSD
Problems with OSD (cont.)
Problems with OSD (cont.)
Problems with OSD (cont.)
Problems with OSD (cont.)
Problems with OSD (cont.)
Problems with OSD
Solutions

- Lockless, share-nothing
  - Seastar
- Kernel bypass
  - DPDK and SPDK
- Minimal memcpy()
A proof of concept

crimson
Crimson OSD

Data flow diagram showing the interaction between messenger, PG, and Object Storage.
repeat([conn, this] {
    return conn.read_request().then([this](auto req) {
        return handle_request(req);
    }).then([conn] (auto result) {
        return conn.send_response(result);
    });
});
an (over) simplified OSD

- single replica
- memory-backed store
- only basic read/write supported
- no peering
random read (depth=16)
random read (depth=16)
random read (depth=16)
random write (depth=16)

https://docs.google.com/spreadsheets/d/1yoH_T3KXHHTTF6Eu4v5X2oehIluD7rQhtDoYAelbzI/edit?usp=sharing
Next Steps

Push forward the frontier
● Performance Test as part of CI
● Port more features
  – BlueStore
  – Scrubbing, recovery
● Seastore
  – Integrate SPDK into seastar
more about crimson

- 123 PRs since Jan 27 2018
- 7 developers from Red Hat and Intel
- under heavy development
  - Crimson/Seastar-OSD Weekly playlist in Ceph Channel on YouTube
  - crimson project kanban at https://github.com/ceph/ceph/projects/2