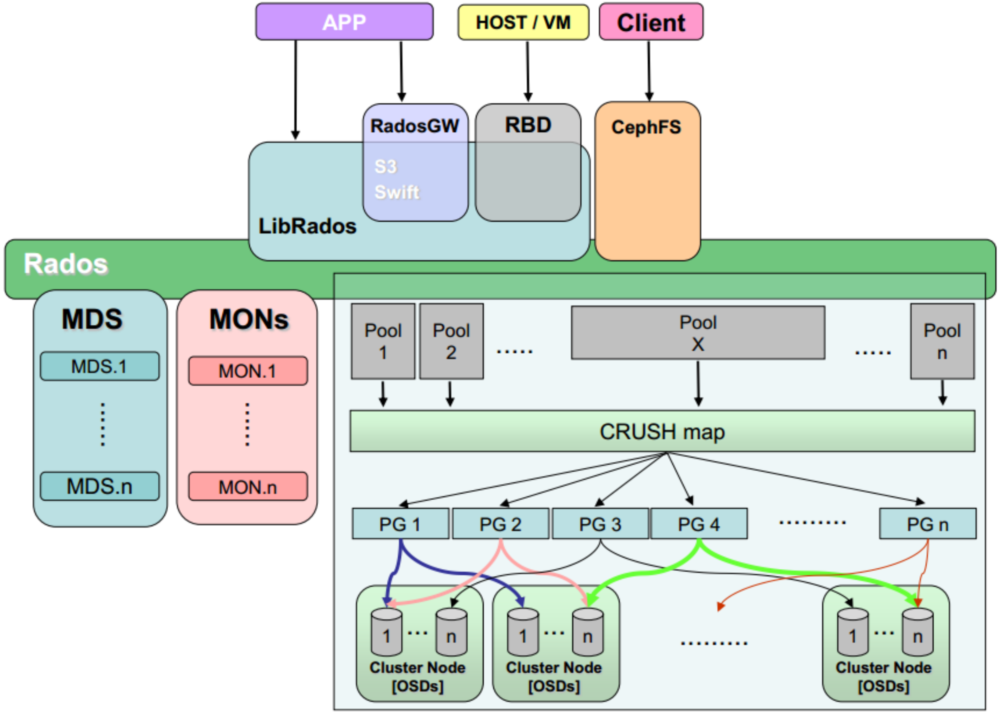
SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

Rethinking Ceph Architecture for Disaggregation using NVMe-over-Fabrics

Yi Zou (Research Scientist), Arun Raghunath (Research Scientist)
Tushar Gohad (Principal Engineer)
Intel Corporation

Contents

Ceph Refresher



- Open-source, object-based scale-out storage system
- Software-defined, hardware-agnostic – runs on commodity hardware
- Object, Block and File support in a unified storage cluster
- Highly durable, available – replication, erasure coding
- Replicates and re-balances dynamically

Disaggregation Refresher

- **Software Defined Storage** (SDS)**:** Scale-out approach for storage guarantees.
  - Disaggregates software from hardware
  - Numerous SDS offerings and deployments

Trend observed in both academia and industry

"Extreme resource modularity" *Gao, USENIX OSDI '16*

Open Compute Project; Intel RSD; HP MoonShot; Facebook disaggregated racks; AMD SeaMicro;

- **Disaggregation:** Separate servers into resource components (e.g. storage, compute blades)
  - Resource flexibility and utilization – TCO benefit
  - Provides deployment flexibility – pure disaggregation, hyper-converged, hybrid
  - Feasible now for SSD due to advancement of fabric technologies

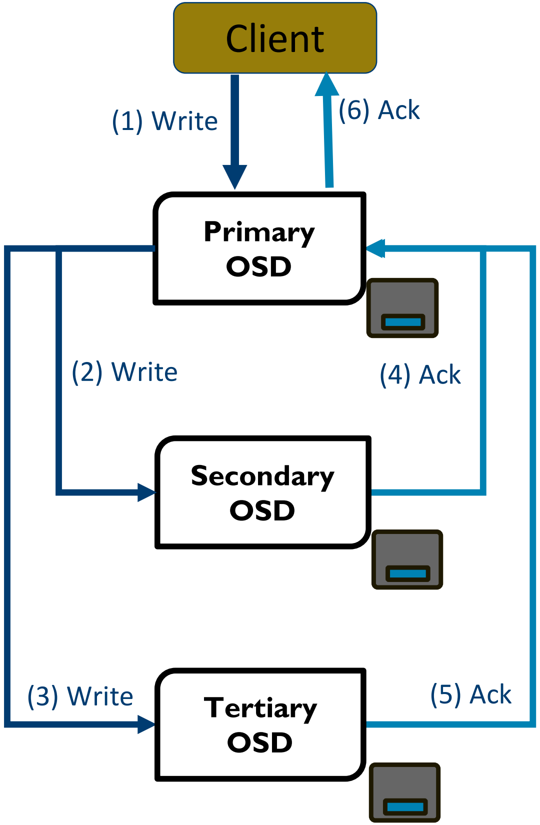
Intel Rack Scale Design

Ceph and NVMe-oF disaggregation options

- Rationale:
  - Share storage tier across multiple SDS options
  - Scale compute and storage sleds independently
  - Opens new optimization opportunities
- Approaches
  - Host based NVMeoF storage backend
    - NVMeoF volume replication in different failure domains
    - Not using Ceph for durability guarantees
  - Stock Ceph with NVMeoF storage backend
    - OSD directed replication
  - Decouple Ceph control and data flows

Ceph Replication Flow

- SDS reliability guarantees → data copies (replication / Erasure Coding)
- SDS durability guarantees → long running tasks to scrub and repair data
- We focus on replication flows in the rest of the talk

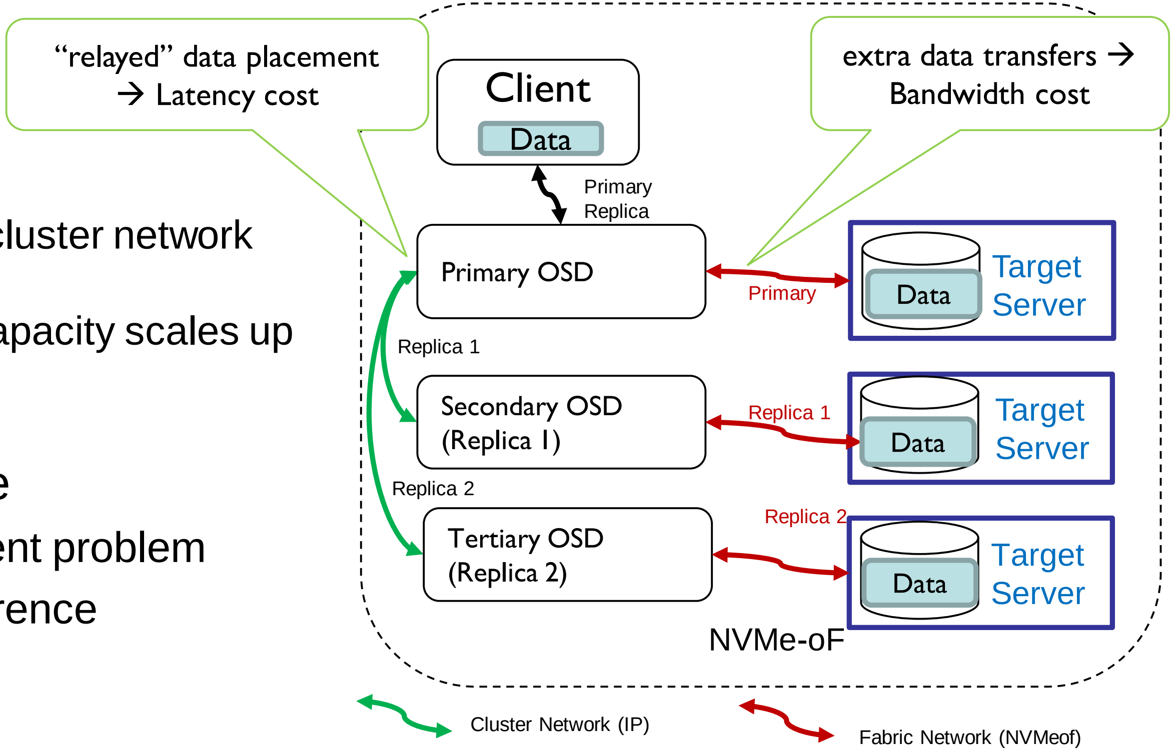


(1) Client writes to the primary OSD
(2) Primary identifies secondary and tertiary OSDs via CRUSH Map
(3) Primary writes to secondary and tertiary OSDs.
(4) Secondary OSD acks write to Primary
(5) Tertiary OSD acks write to Primary
(6) When writes are settled – Primary OSD Acks to the client

Stock Ceph disaggregation: Datacenter 'tax'

- Ceph Deployments Today
  - Common to provision separate cluster network for internal traffic
  - Network cost compounded as capacity scales up

- Disaggregating OSD storage
  - Exacerbates the data movement problem
  - Magnifies performance interference



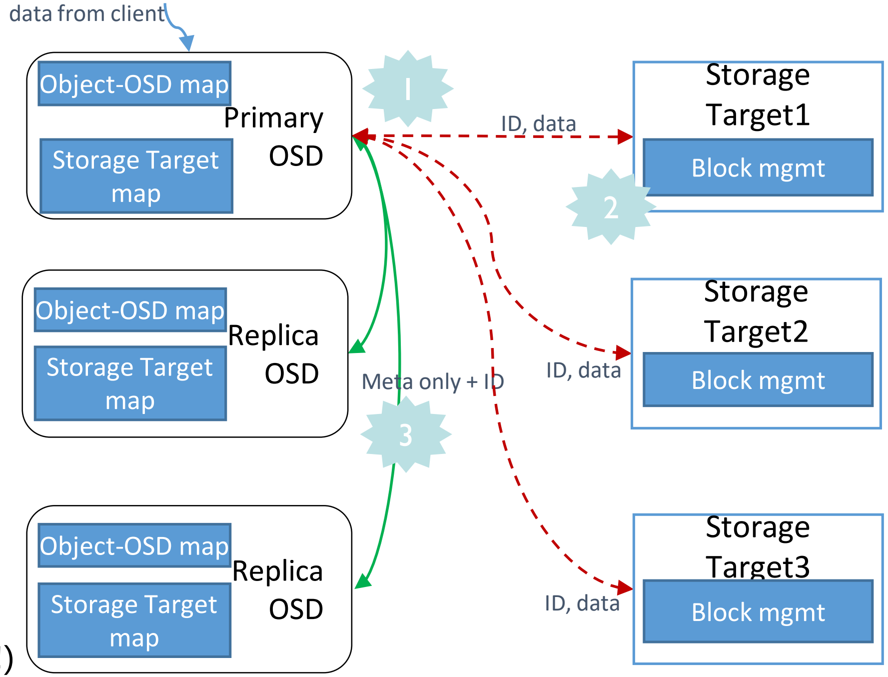
OSD: Object Storage Daemon

A Different Approach - Decoupling Data and Control Plane

1. Direct data copy to storage target
   - *Issue: Need final landing destination*
   - Current consistent hashing maps Object → OSD
   - Maintain a map of storage target assigned to each OSD
   - Consult map to find storage target for each OSD

2. Block ownership
   - *Issue: Currently the OSD host File-System owns blocks*
   - Move block ownership to remote target (details next slide)

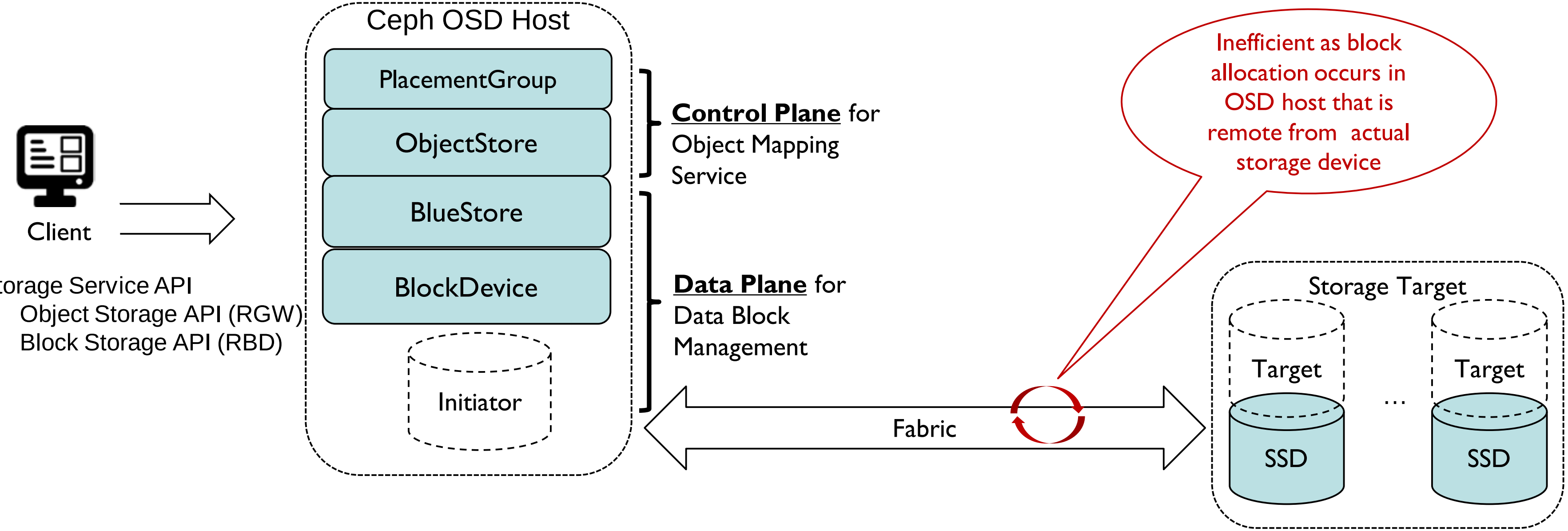
3. Control plane
   - *Issue: Metadata tightly coupled with data*
   - Send only metadata to replica OSD (eliminates N-1 data copies!)
   - Unique ID to correlate meta with data



Typical 3-way replication, total 4 hops here vs 6 hops in stock Ceph E2E from client to target!

* OSD: Object Storage Daemon

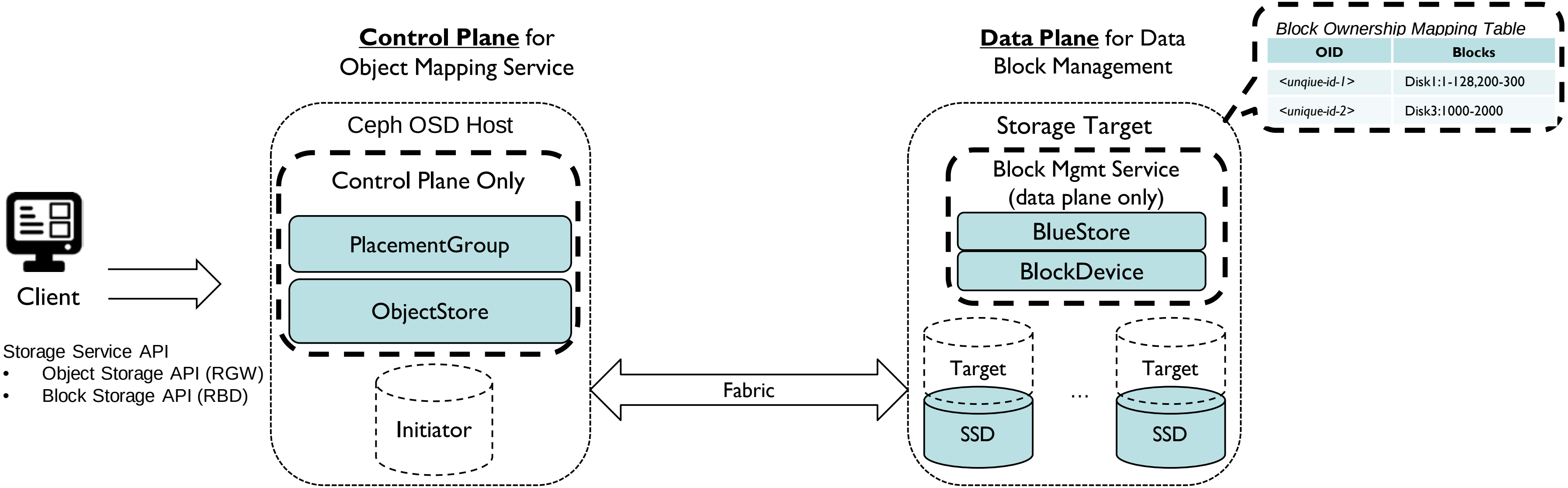
Stock Ceph Architecture – Control and Data Plane



Storage Service API

- Object Storage API (RGW)
- Block Storage API (RBD)

Architecture Change – Remote Block Management

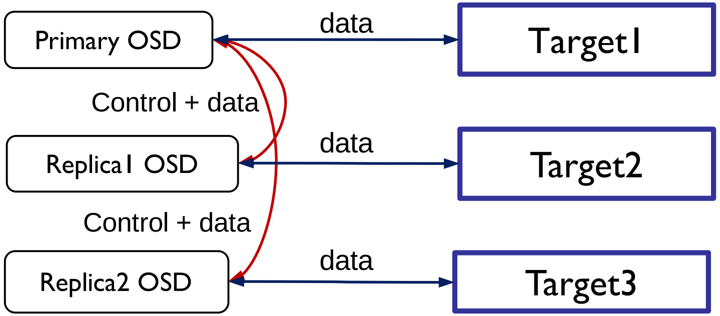


OID	Blocks
<unqiue-id-1>	Disk1:1-128,200-300
<unique-id-2>	Disk3:1000-2000

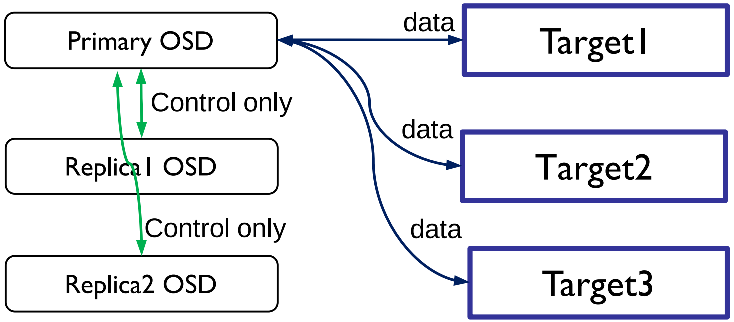
Storage Service API
- Object Storage API (RGW)
- Block Storage API (RBD)

Bandwidth benefits: Remote Block Management

Stock Ceph with NVMe-oF



Ceph optimized for NVMe-oF

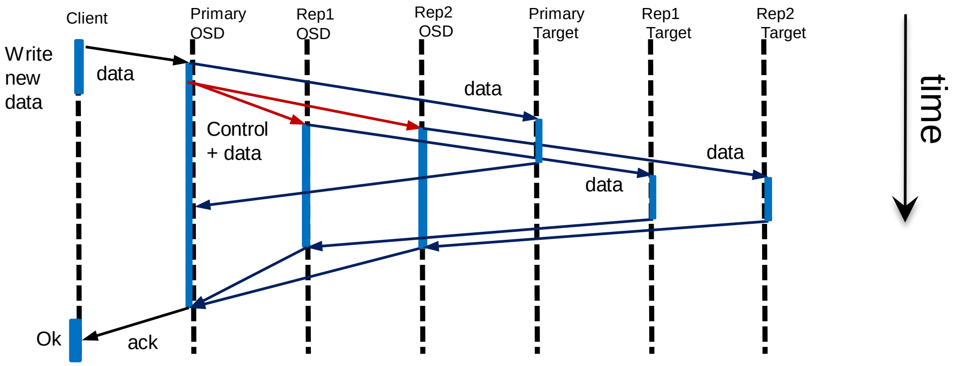


Estimated Reduction in Bandwidth consumption
$Reduction\ (bytes) = (r - 1) \times (M_d - M_m)$

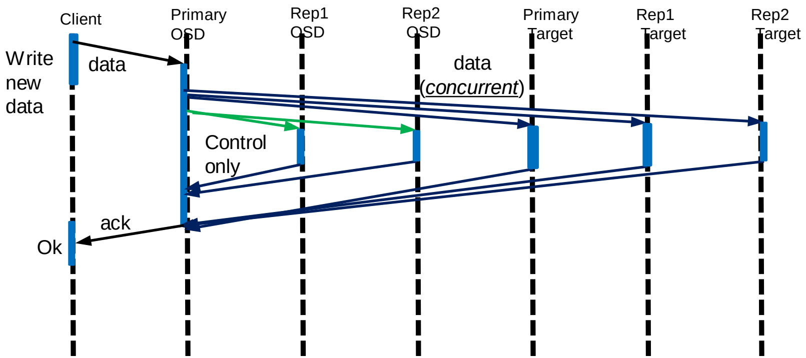
40% reduction for 3-way replication!

Latency benefits: Decouple control and data flows

Stock Ceph with NVMe-oF



Ceph optimized for NVMe-oF

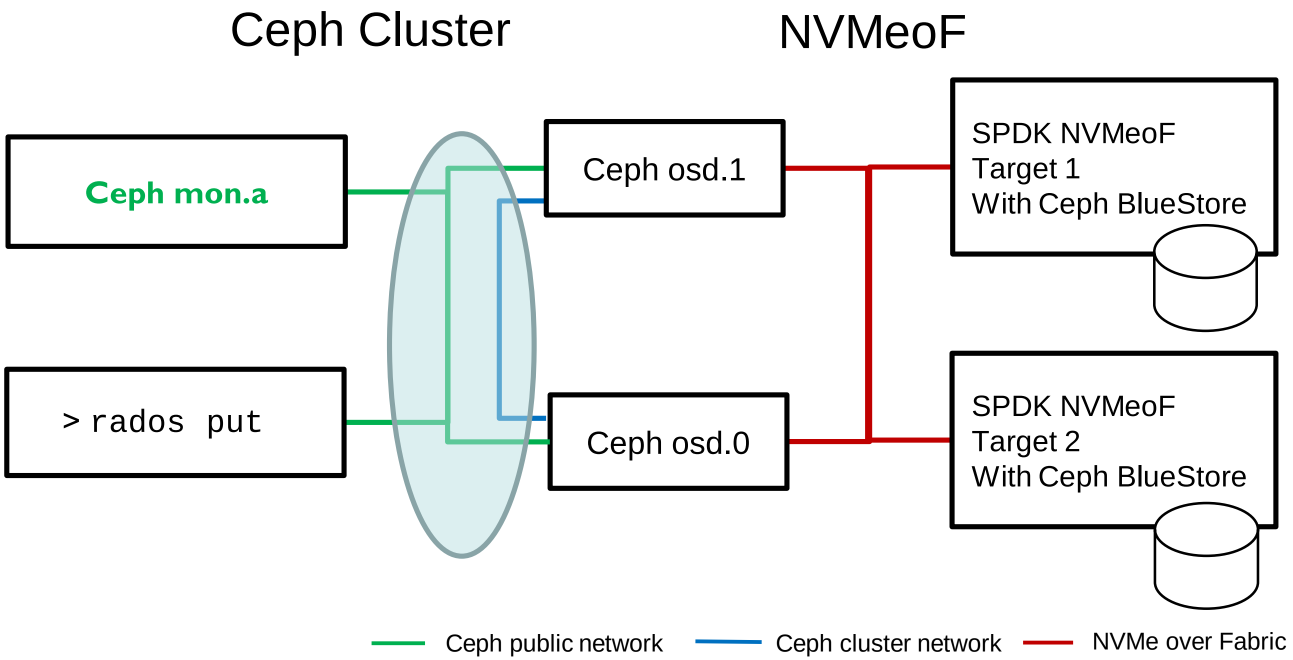


Estimated Latency Reduction
$Reduction\ (usec) = N_d + m + N_a$

1.5X latency improvement !

PoC Setup

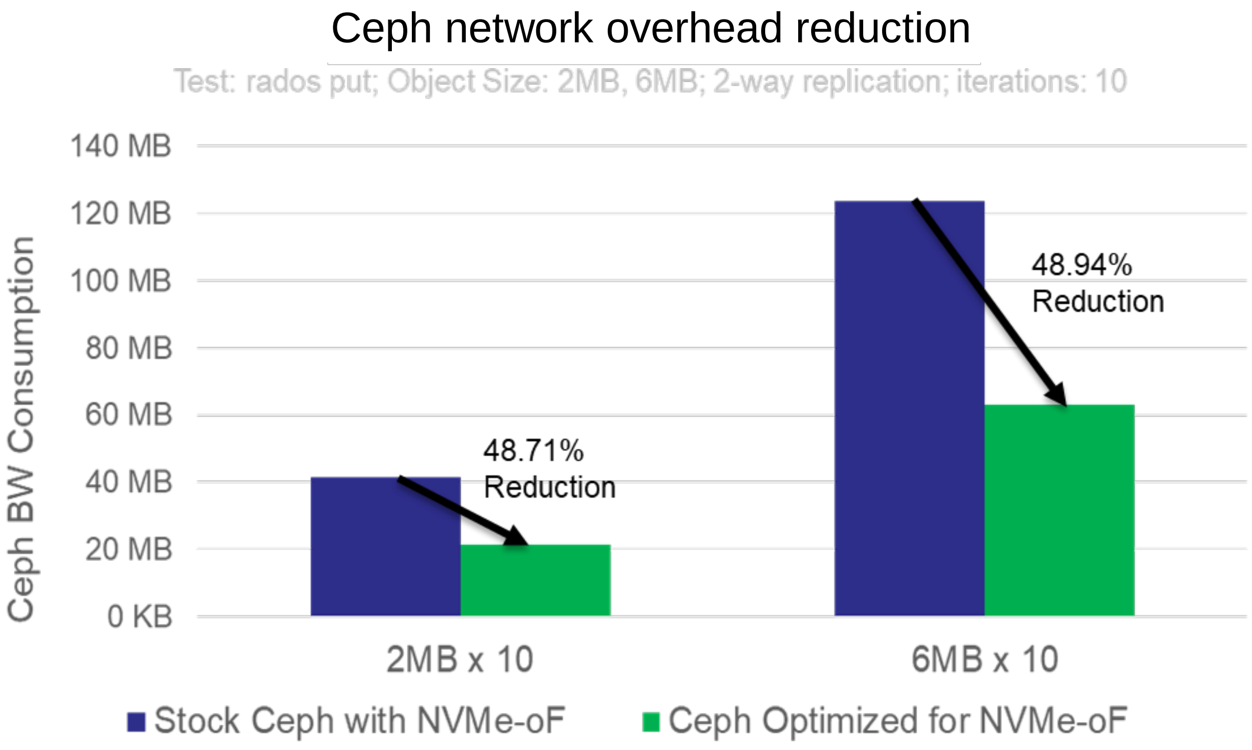
- Ceph Luminous
- 2-Way Replication
- Ceph ObjectStore as SPDK NVMe-oF Initiator
- SPDK RDMA transport
- SPDK NVMe-oF target
- SPDK bdev maps requests to remote Ceph BlueStore
- Linux Soft RoCE (rxe)



Metric: Ceph cluster network rx/tx bytes

Preliminary Results*

- Test: rados put
- 10 iterations
- Measure Ceph network rx/tx bytes
- Derive reduction in bandwidth consumption



* Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Intel is the trademark of Intel Corporation in the U.S. and/or other countries. Other names and brands may be claimed as the property of others. See Trademarks on intel.com for full list of Intel trademarks or the Trademarks & Brands Names Database

Summary & Next Steps

Summary

- Eliminate datacenter ‘tax’
  - Decouple control/data flows
  - Remote block management
- Preserve Ceph SDS value proposition
- Reduce TCO for Ceph on disaggregated storage
- Bring NVMe-oF value to Ceph users

Future work

- Validate new architecture with Ceph community
- Integrate storage target information with crush-map
- Evaluate performance at scale
- Mechanism to survive OSD node failures
- Explore additional offloads for replication

Q&A