ZStore: A Fast, Strongly-consistent Object Store with ZNS SSDs

Shuwen Sun, Isaac Khor, Ji-yong Shin, Peter Desnoyers



Gap between hardware and system

Production Store: Ceph (Reef release) RGW

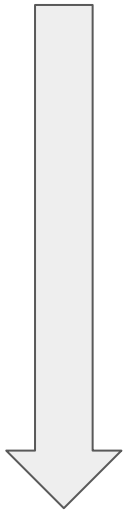
- 10 node, 60 NVMe drive with 3X replication
- 4KB: 312K IOPS for GET, 178K IOPS for PUT
  - 312K/60*3 = **15.6K IOPS for GET**
  - 178K/60*3 = **8.9K IOPS for PUT**

Hardware devices: Enterprise SSD: Samsung PM983

- Random Read (4KB): **540k IOPS**
- Random Write (4KB): **50k IOPS**

Consistency models (in dist. storage)

- Eventual consistency
- Read-after-write consistency
  - AWS S3: 2020 - now
- Linearizability



Weak

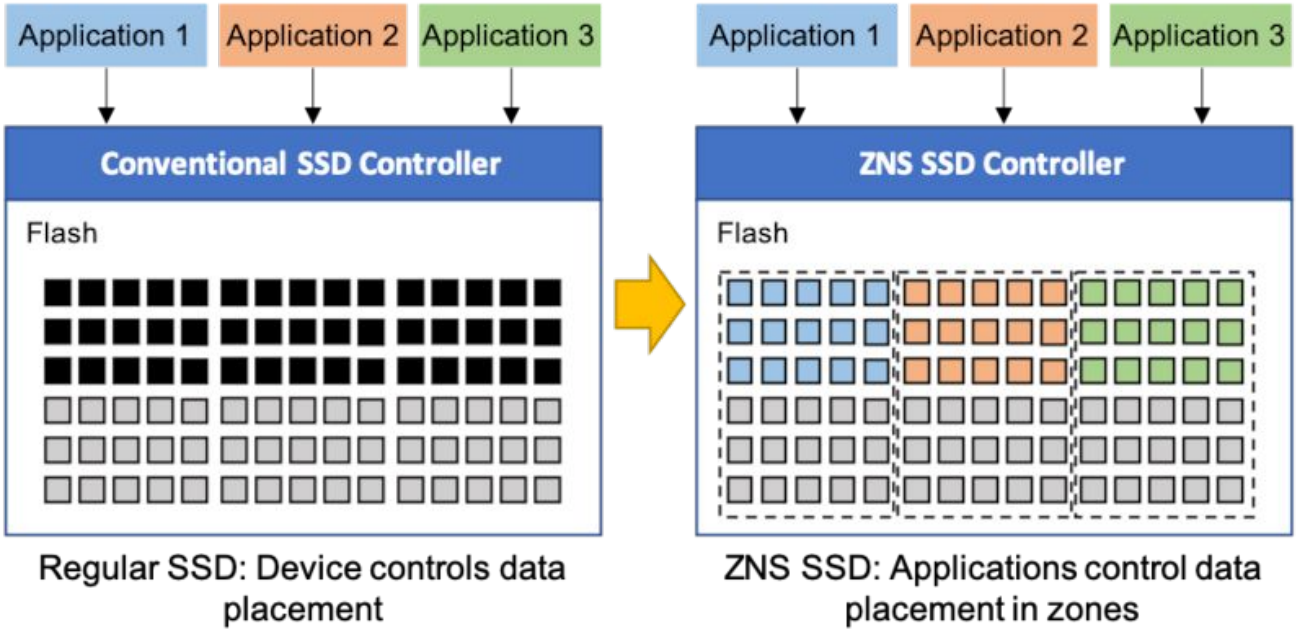
Strong

Explanation of The S3 Consistency Model

Research question

- Can we build a fast, efficient and strongly-consistent object store?
  - To answer that, we present ZStore, which achieves three goals
  - Machine learning and LLM
  - **NO!** We don’t need ML or LLM

Key technology: Zoned NameSpace SSD



Regular SSD: Device controls data placement

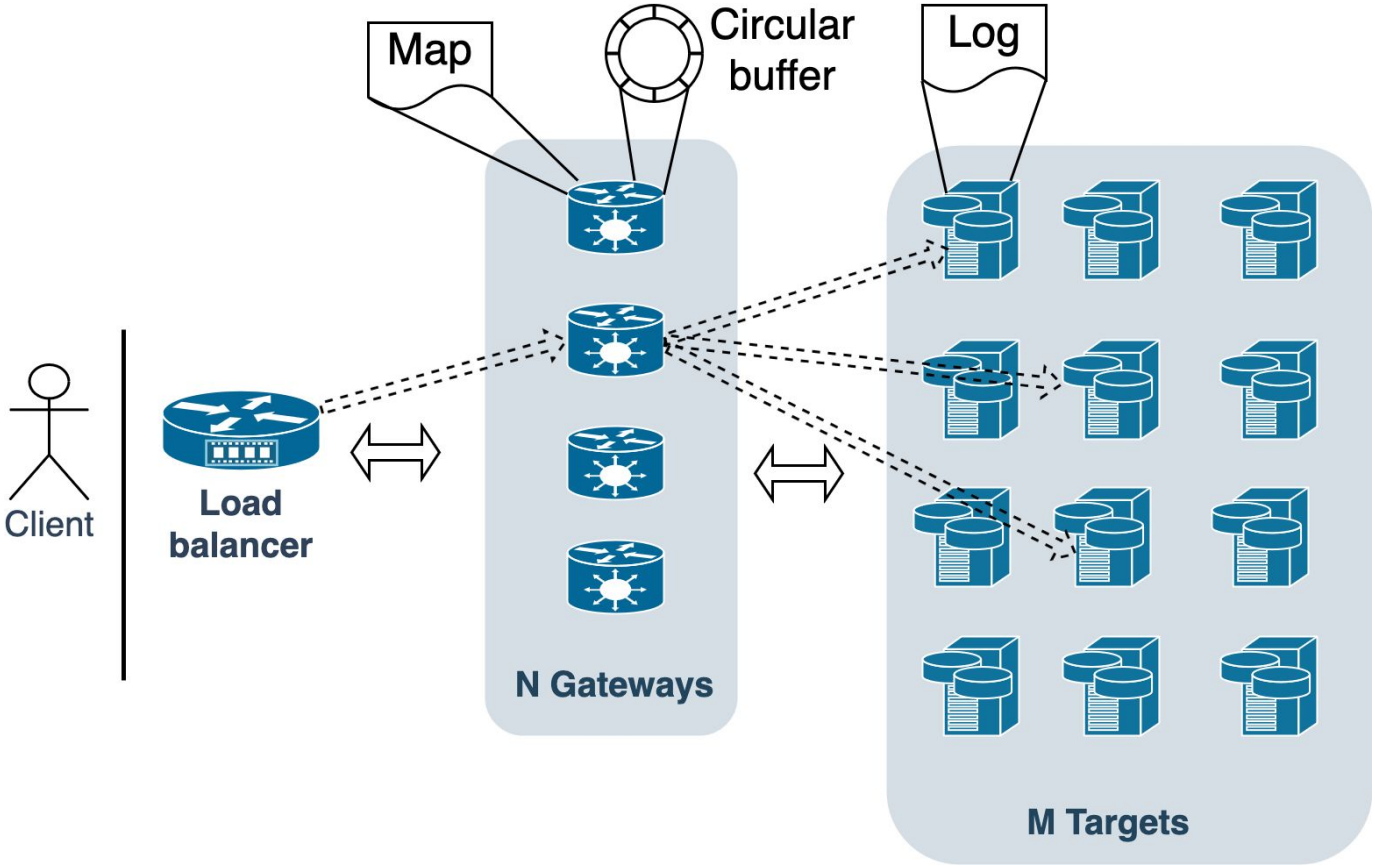
ZNS SSD: Applications control data placement in zones

- ZNS SSD follows a Zoned Storage model
- Translation layer implemented in host

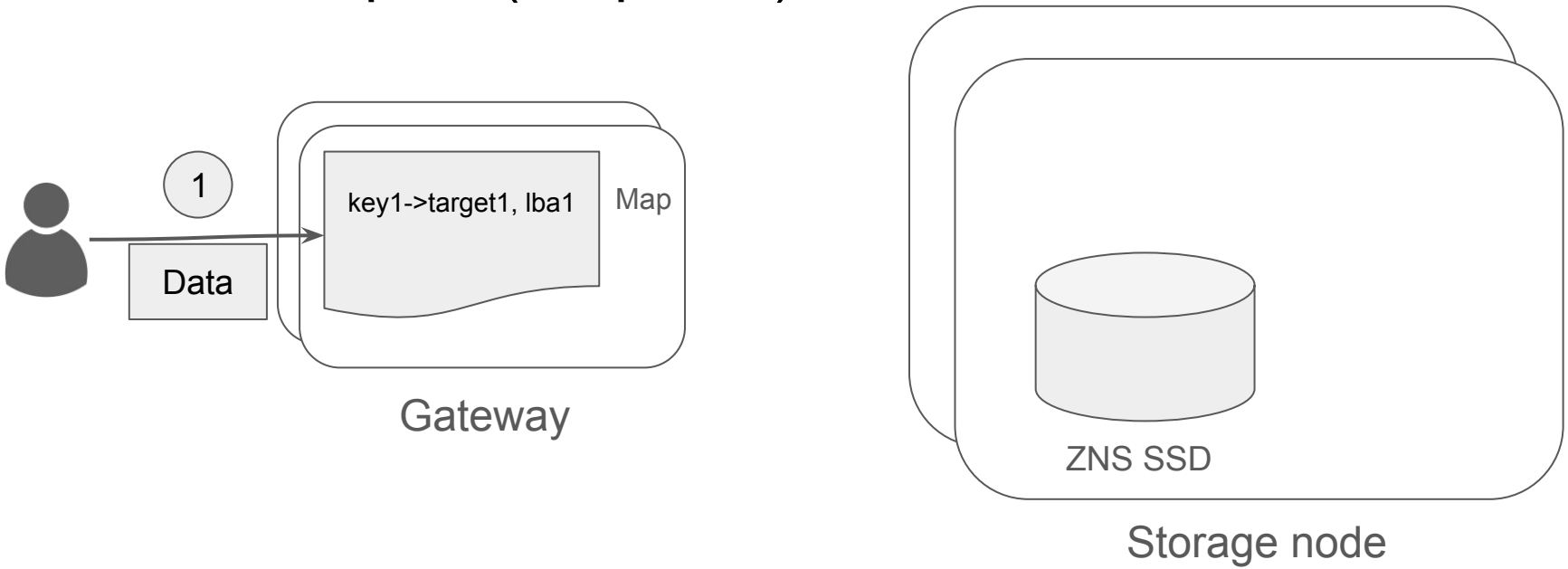
Key operation in ZNS: Zone Append

- **Zone Append**
  - Instead of specifying LBA write to, only specifying the **zone starting LBA**
  - Append data to a zone with implicit write pointer
  - Device/driver returns LBA where data was written in zone
- **Issues with Zone append:**
  - Reordering of sequence of writes, etc
  - Consistency challenge
- **Linearizability guarantee:**
  - RDMA to push information about ongoing writes without waiting for response

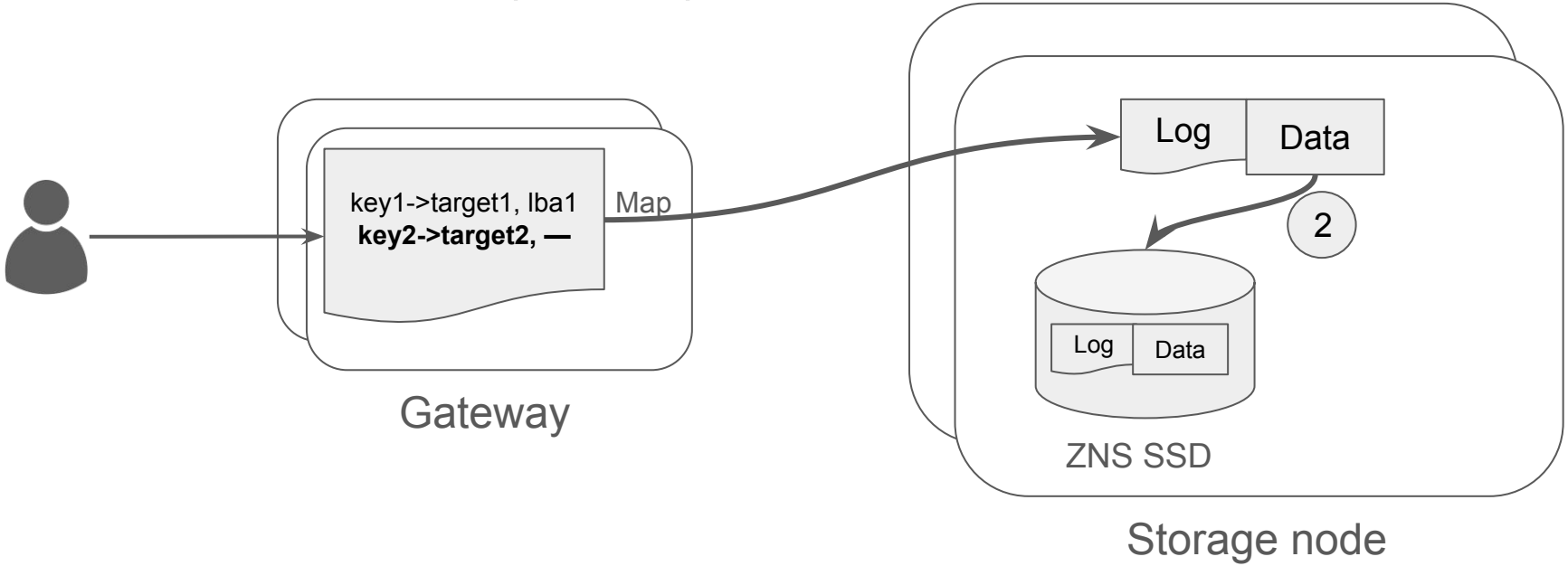
ZNS Object Store architecture



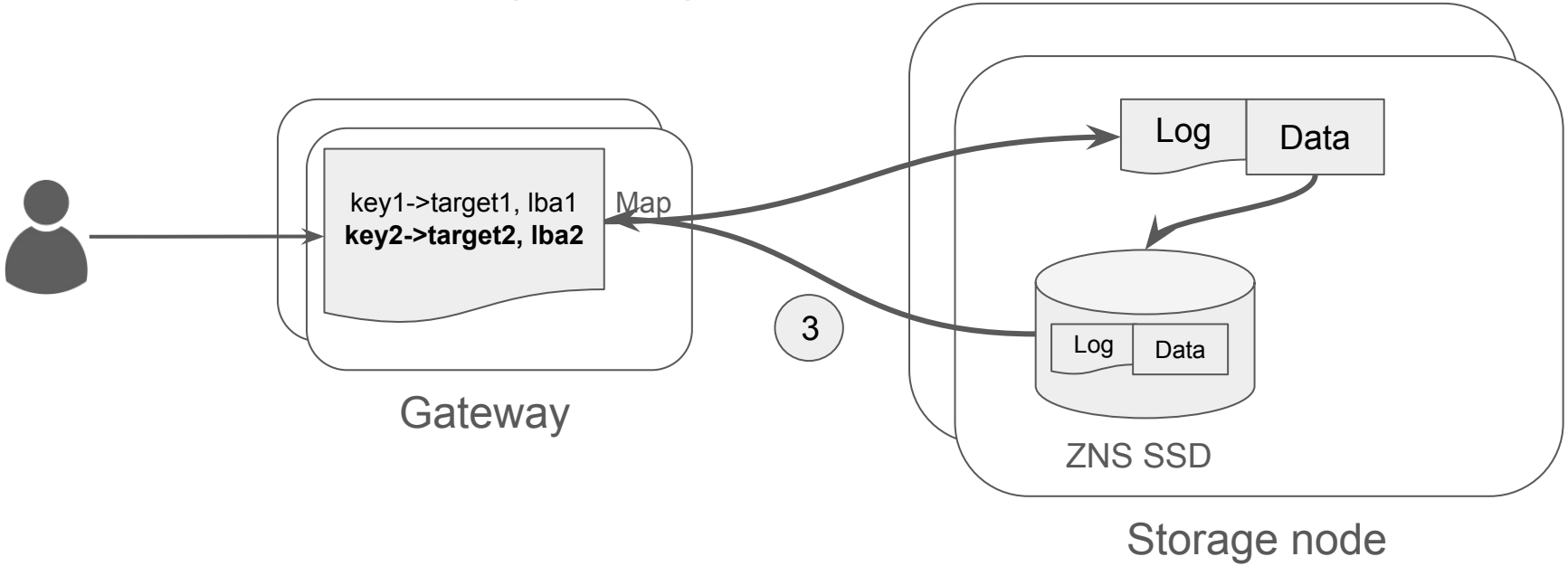
ZStore write path (simplified)



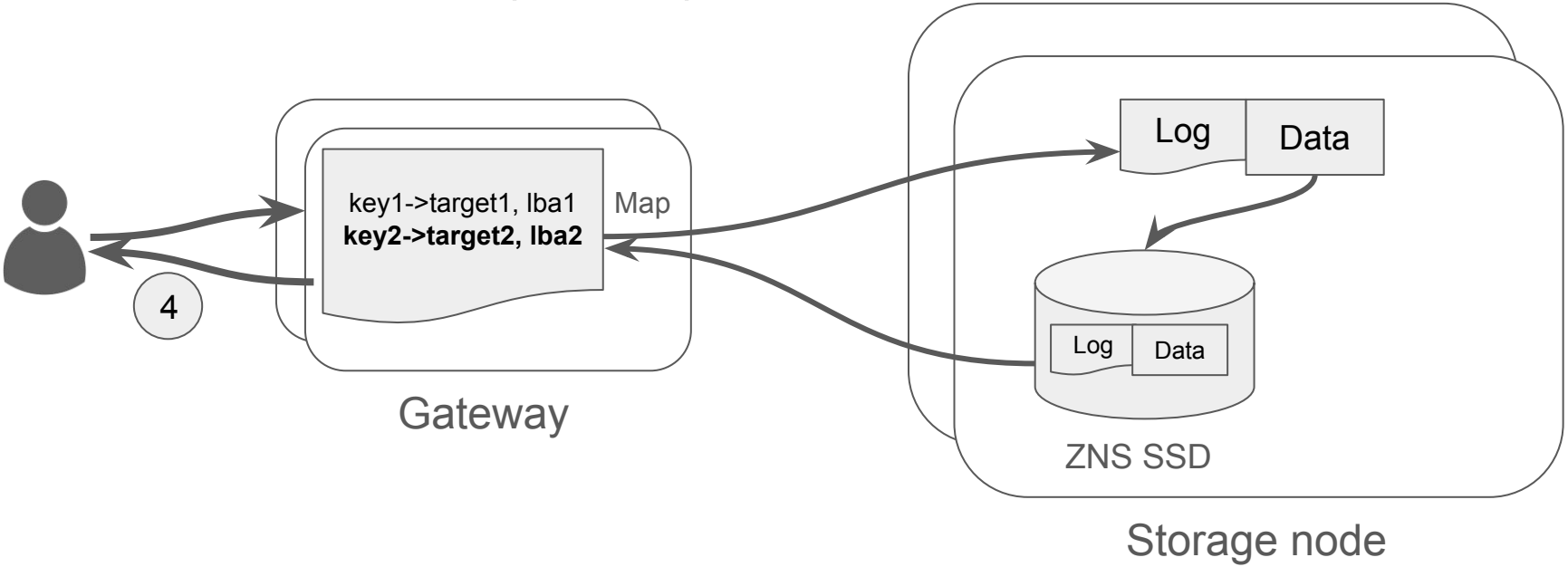
ZStore write path (cont'd)



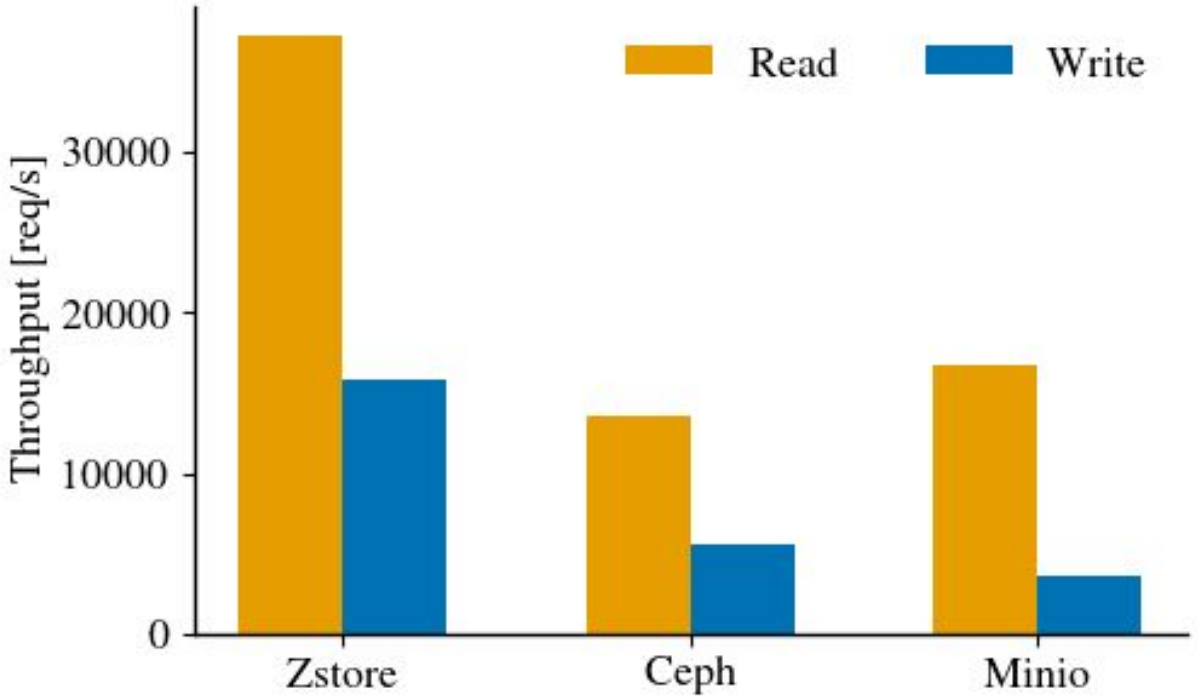
ZStore write path (cont'd)



ZStore write path (cont'd)

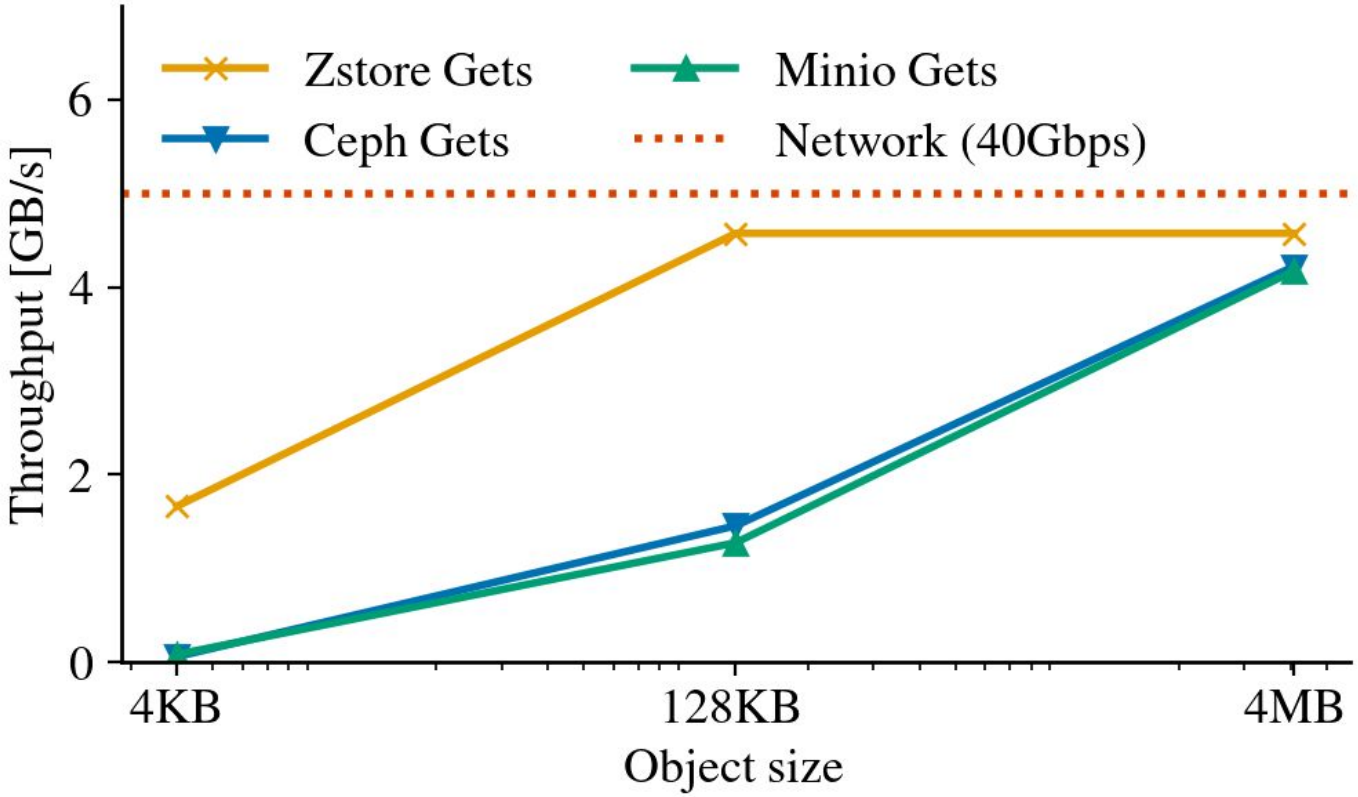


S3 Bench evaluation



- Single gateway
- Three targets
- 4KB object

Wrk evaluation



- Zstore achieves 400k RPS at 4K
- Saturates network at 8KB

Takeaway

- ZStore is a new object store built on ZNS SSDs
- ZStore is efficient:
  - k-way replication: k NVMe writes
  - Read: 1 NVMe read of object data and metadata
- Close the gap between
  - hardware performance and
  - the performance of consistent storage service