Wednesday, May 6th 2026
USENIX NSDI 2026
Renton, WA, USA

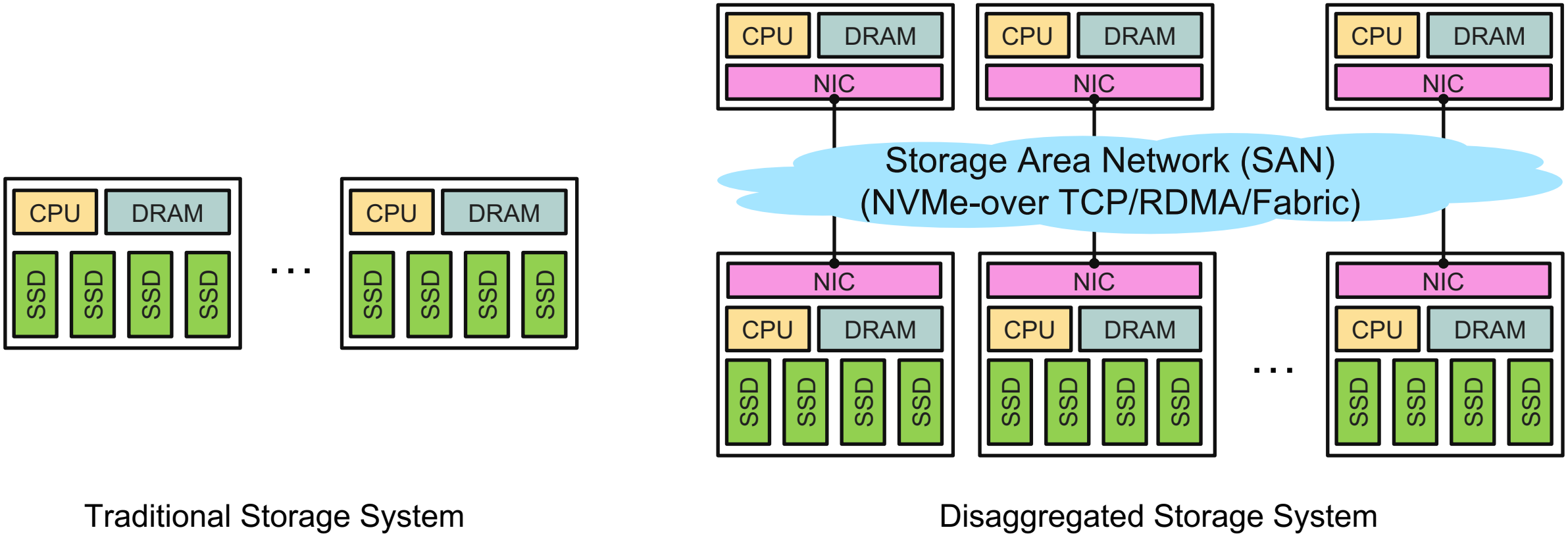
Co-Designing Traffic Control with NVMe-oF for Disaggregated Storage:

A Comparative Study of Switched and Switchless SAN Architectures

Chendong Wang, Joontaek Oh, Ming Liu
NetLab @ University of Wisconsin-Madison

Storage Disaggregation

Decoupling storage from compute

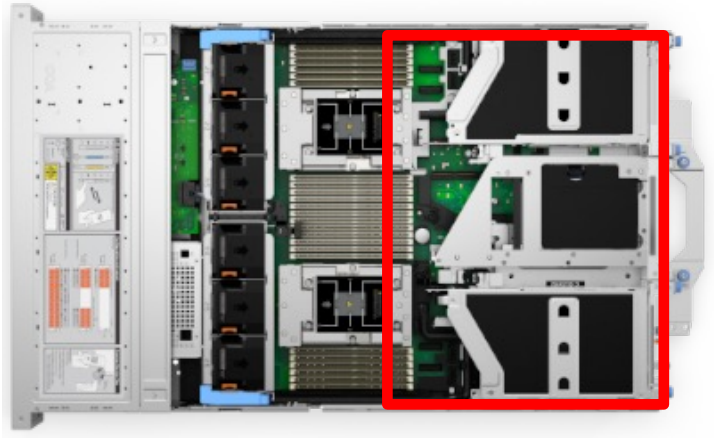


Traditional Storage System

Disaggregated Storage System

Evolution of Storage Node

Trend 1: PCIe Bandwidth Scaling



Dell R7725

- 256 PCIe Gen5 lanes
- Up to 1.0 TB/s Bandwidth
- PCIe Gen6/7 on the horizon

Trend 2: More Compact Form Factor

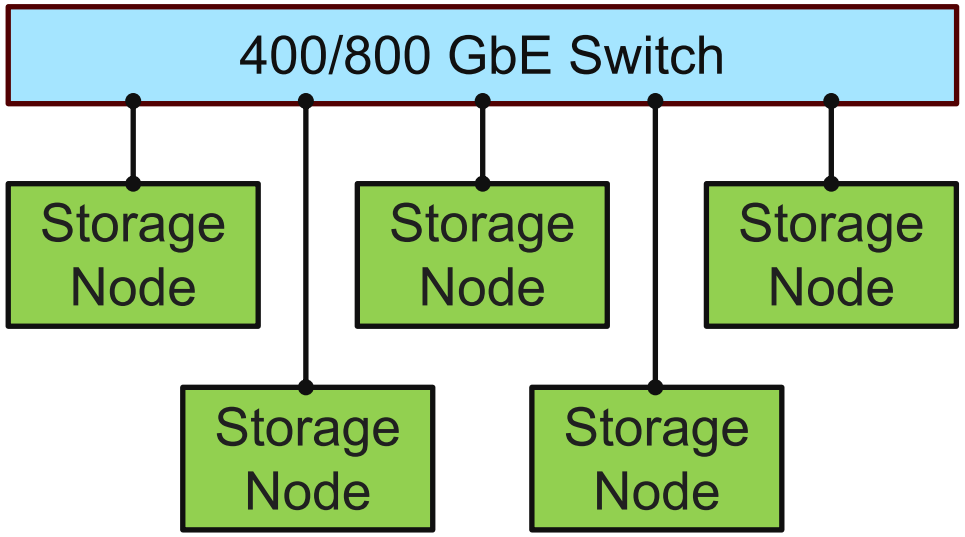
2.5” Form Factor

EDSFF Form Factor

- 1U: 10x2.5”
- 1U: 32 x E1.S **3X Density**

Two Industry Responses to Scale SAN

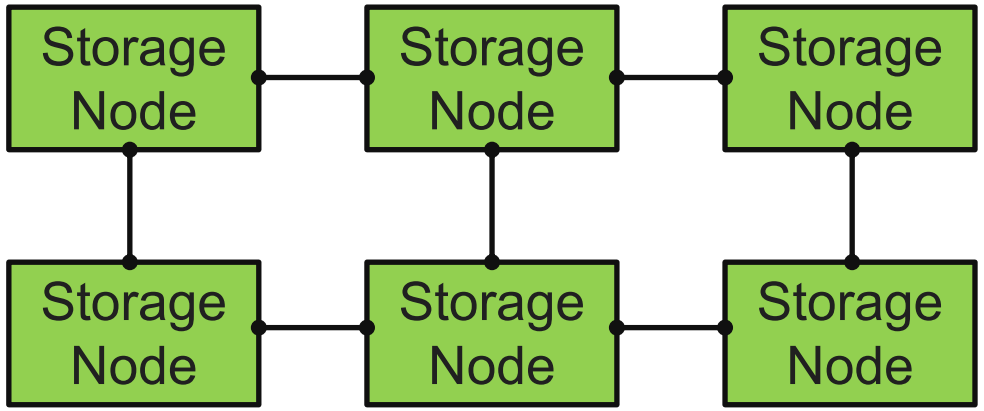
Switched SAN (Scale-Up)



✓ Simple, high-perf, easy to deploy

✗ Prohibitively expensive switches

Switchless SAN (Scale-Out)

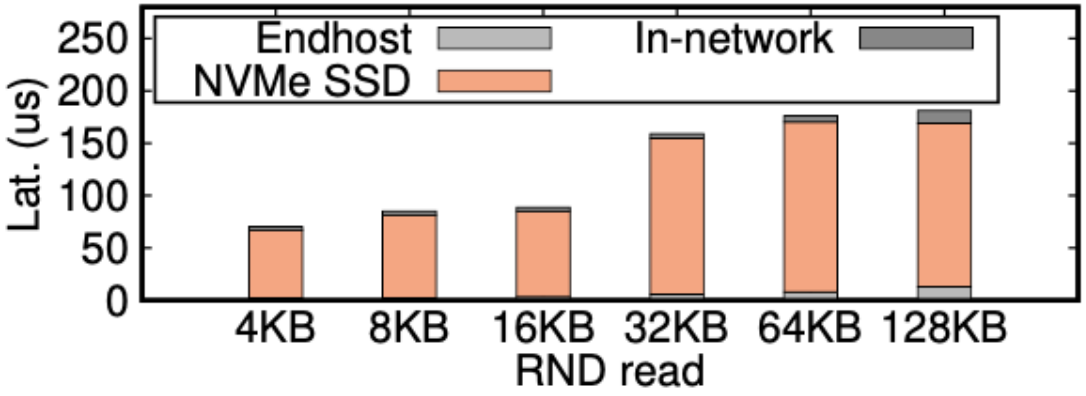


✓ Cheap, adaptive, multi-path

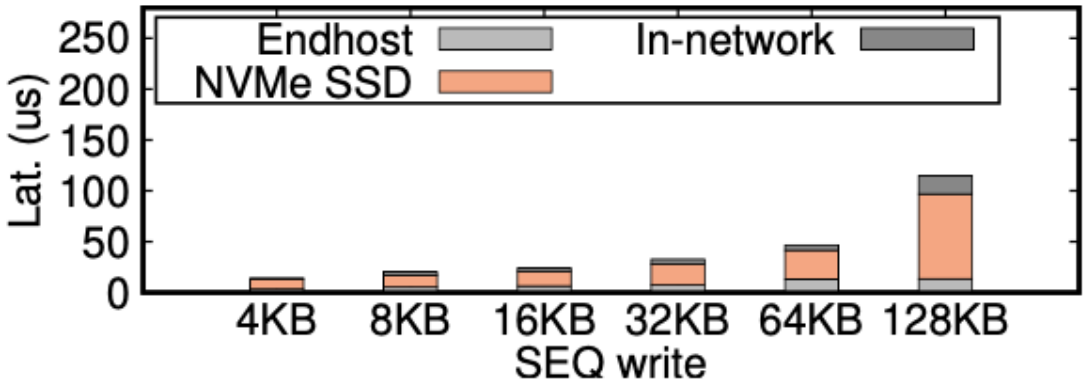
✗ Multi-hop, opaque to storage stack

Switched SAN vs. Switchless SAN?

Observation #1: Network RTT ≪ Storage RTT



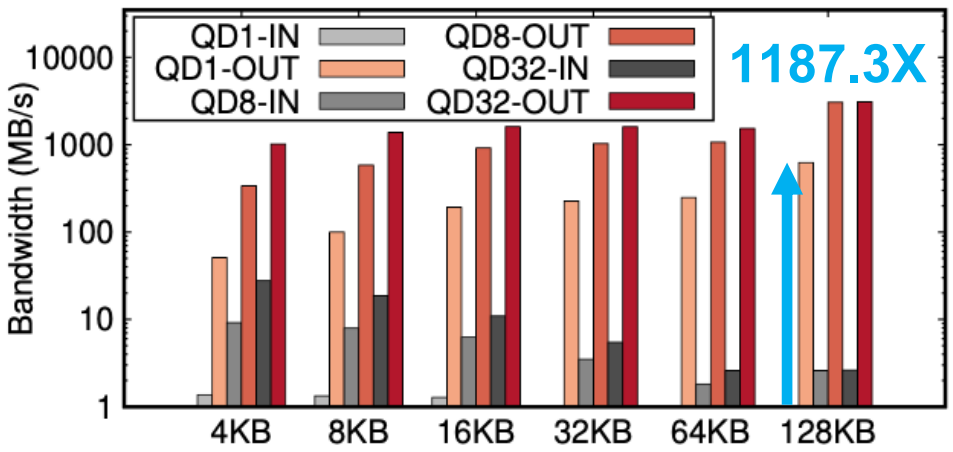
Network delay: 4.8%



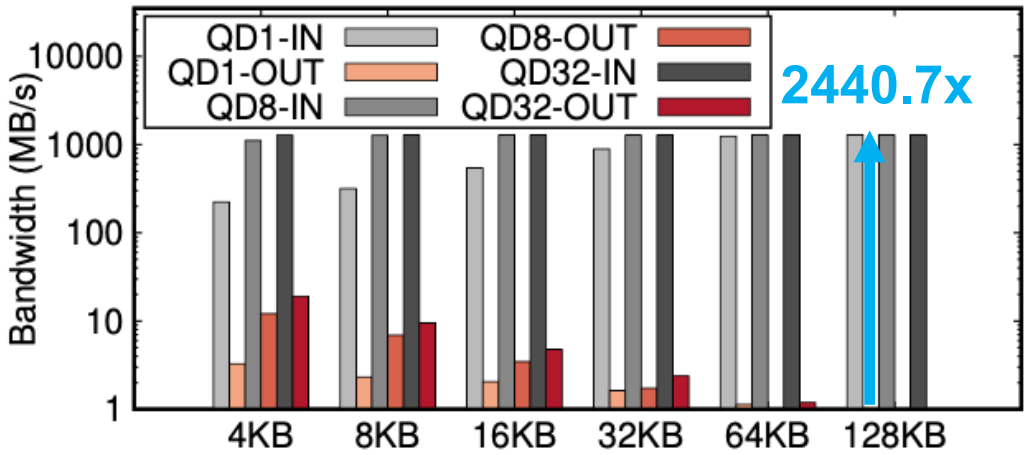
Network delay: 12.7%

Adding some network latencies will not adversely impact the NVMe-oF SAN performance

Observation #2: Pairwise and Asymmetry



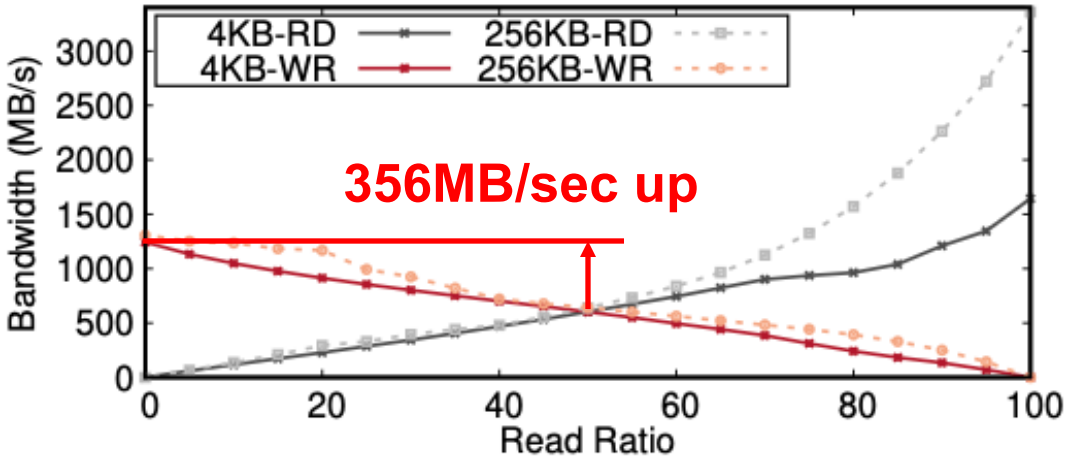
Random Read

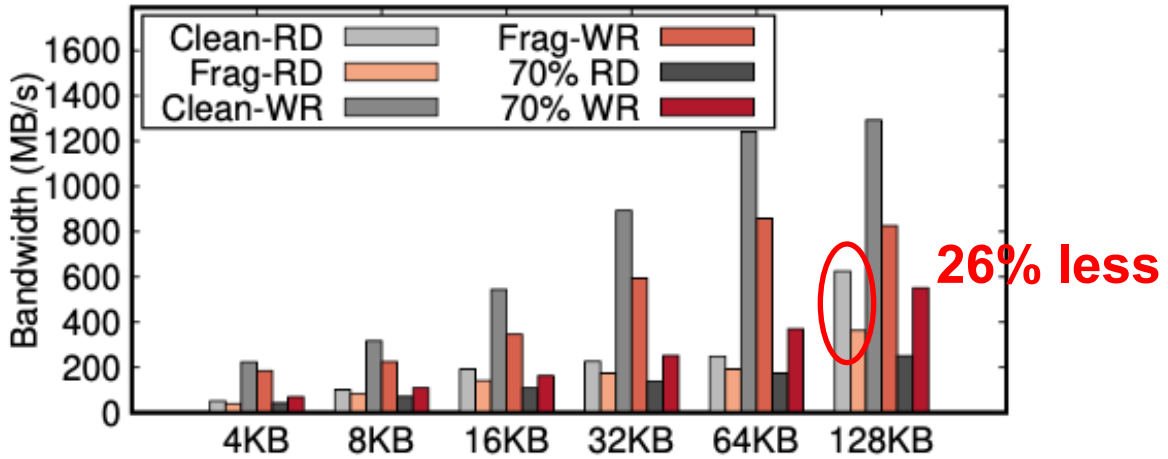


Sequential Write

Traffic is predictable. The BW requirement depends on the I/O read/write ratio

Observation #3: Backward-Propagated Queueing





NVMe-oF performance is determined by the minimum bandwidth of network and storage.

Design Principles from NVMe-oF Characteristics

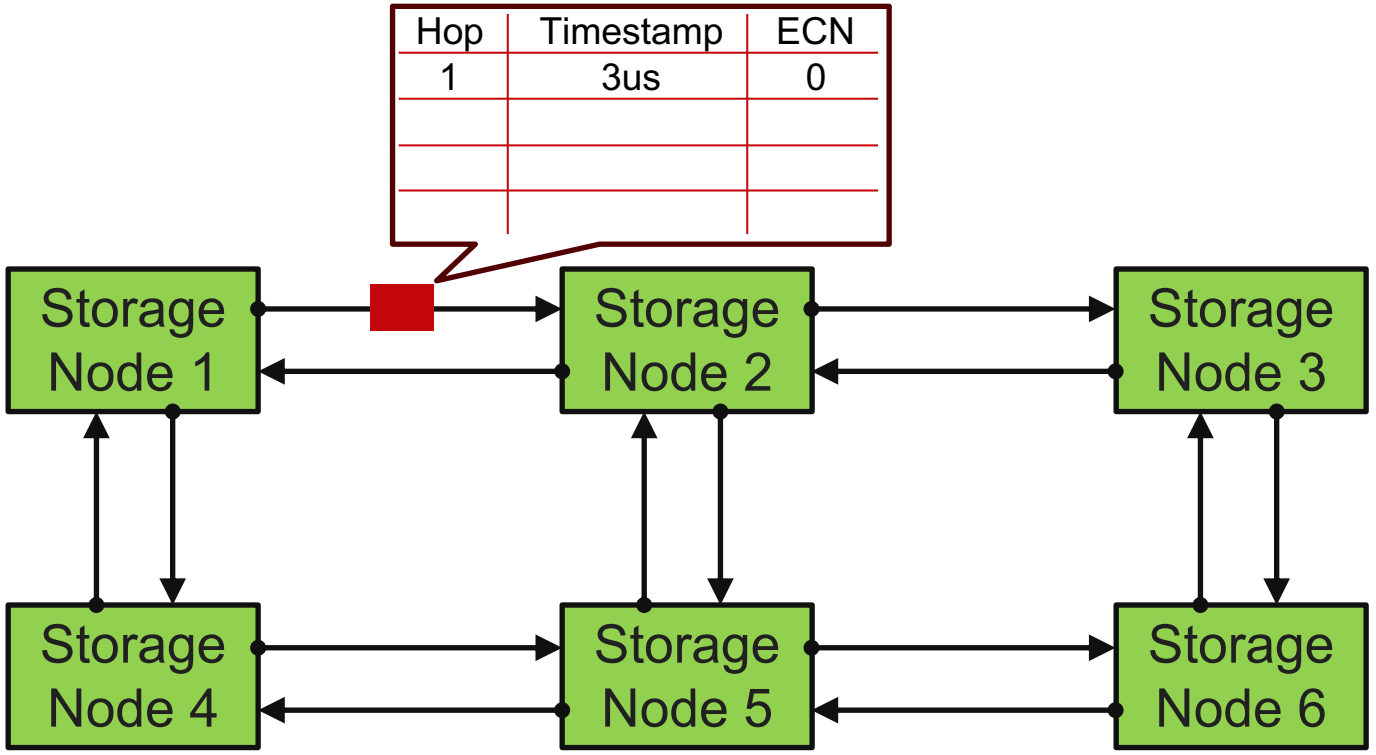
Observations	NVMe-oF Enabled Co-Design	
Network RTT ≪ Storage RTT	Greedy Routing	INT-assisted Symmetric Routing
Pairwise and Asymmetry	Predictable BW	Eager Bandwidth Reservation
Backward-Propagated Queueing	Consider Storage State	Storage-driven Traffic Scheduling

INT-Assisted Symmetric Routing (ISR)

(In-Network Telemetry)

Goal: Pick the least-congested path to the target SSD

Idea: Each completion packet carriers forward-path telemetry back, hop-by-hop.

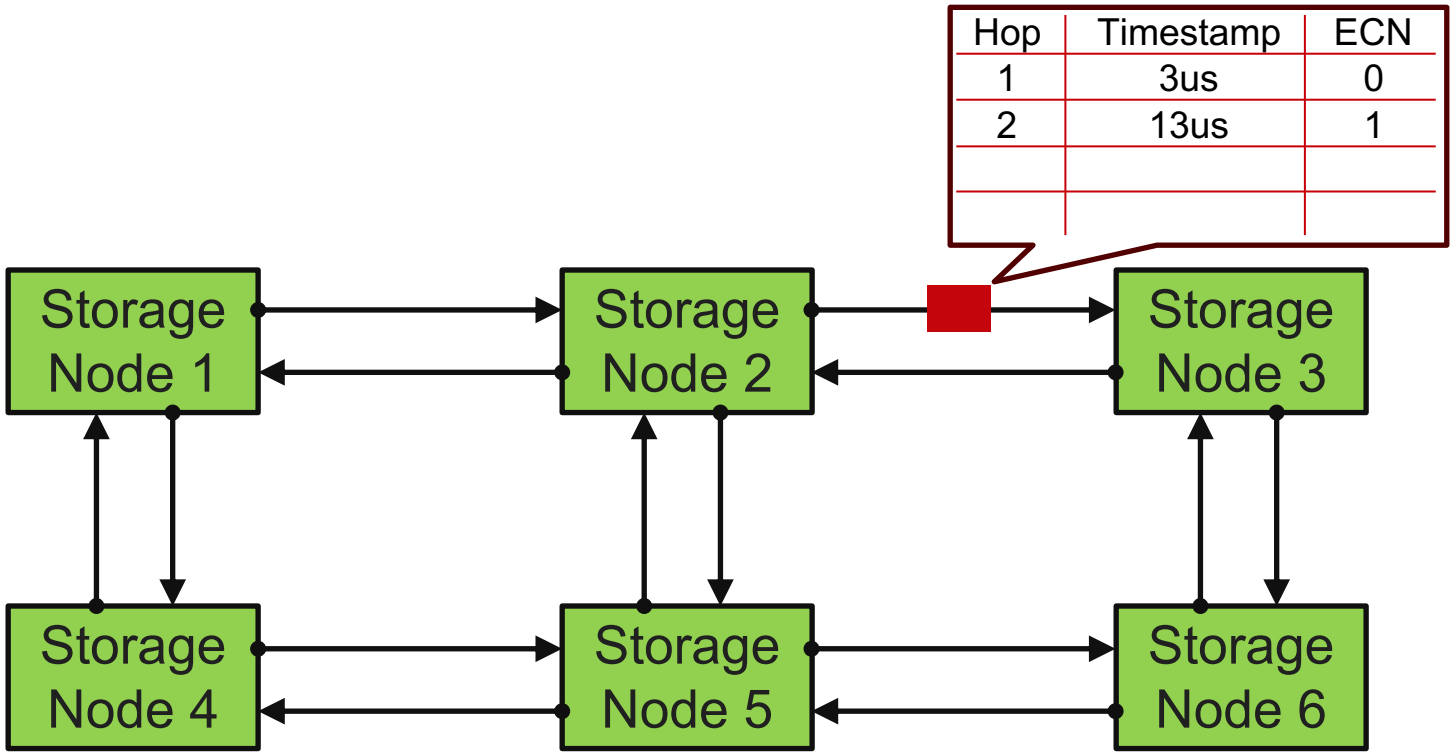


INT-Assisted Symmetric Routing (ISR)

(In-Network Telemetry)

Goal: Pick the least-congested path to the target SSD

Idea: Each completion packet carriers forward-path telemetry back, hop-by-hop.

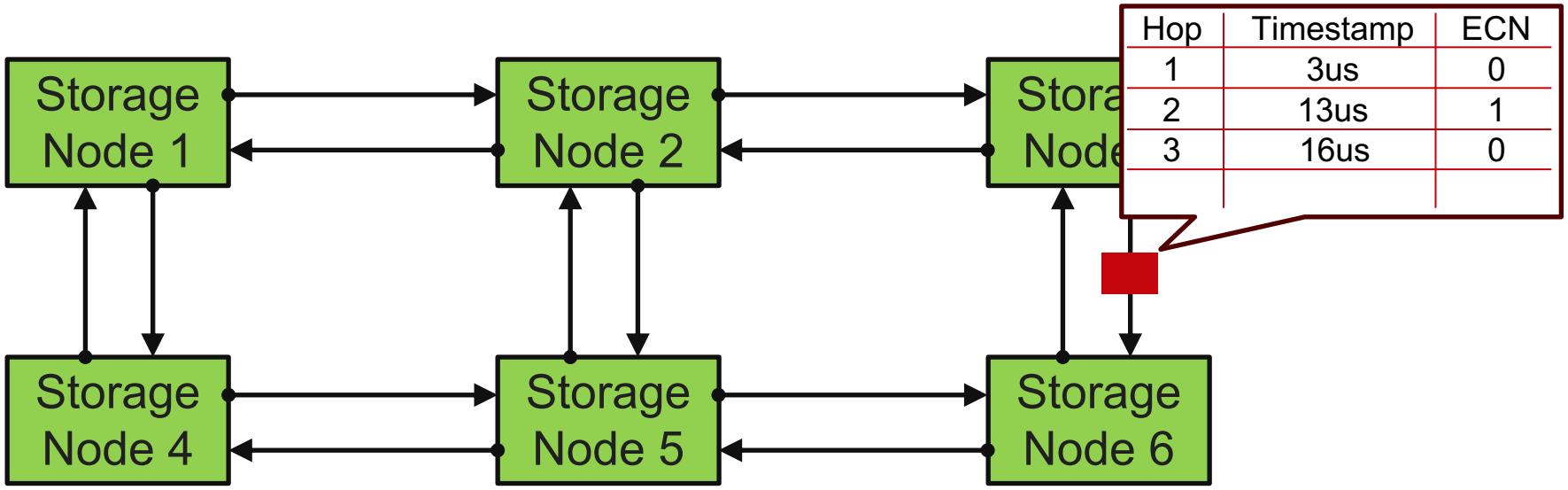


INT-Assisted Symmetric Routing (ISR)

(In-Network Telemetry)

Goal: Pick the least-congested path to the target SSD

Idea: Each completion packet carriers forward-path telemetry back, hop-by-hop.

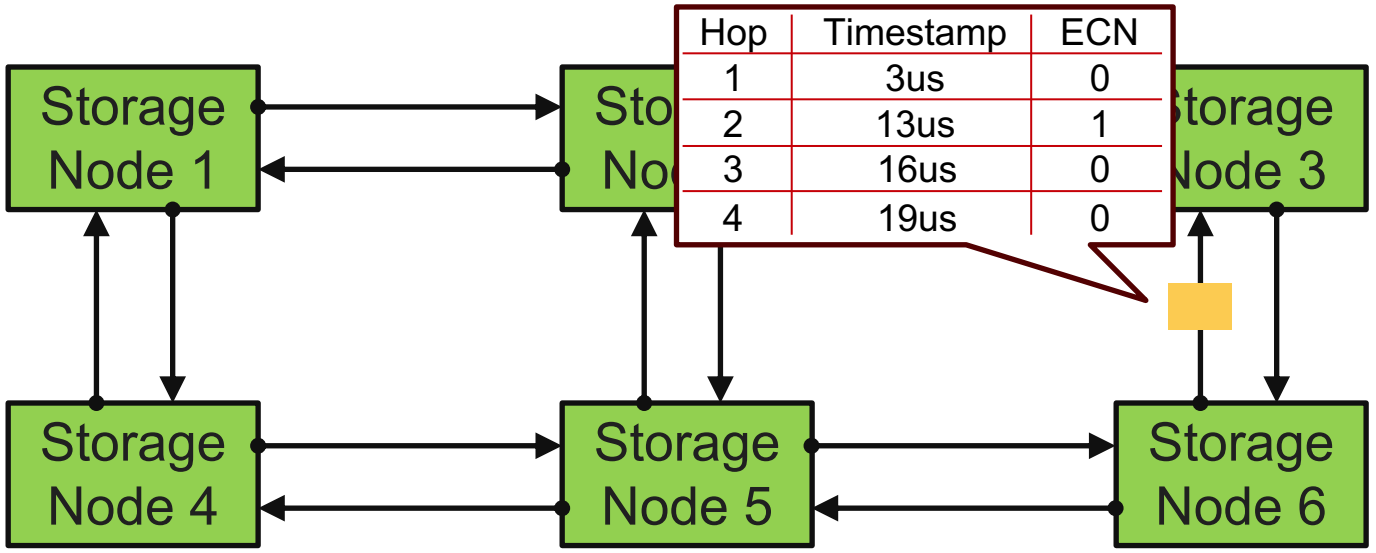


INT-Assisted Symmetric Routing (ISR)

(In-Network Telemetry)

Goal: Pick the least-congested path to the target SSD

Idea: Each completion packet carriers forward-path telemetry back, hop-by-hop.

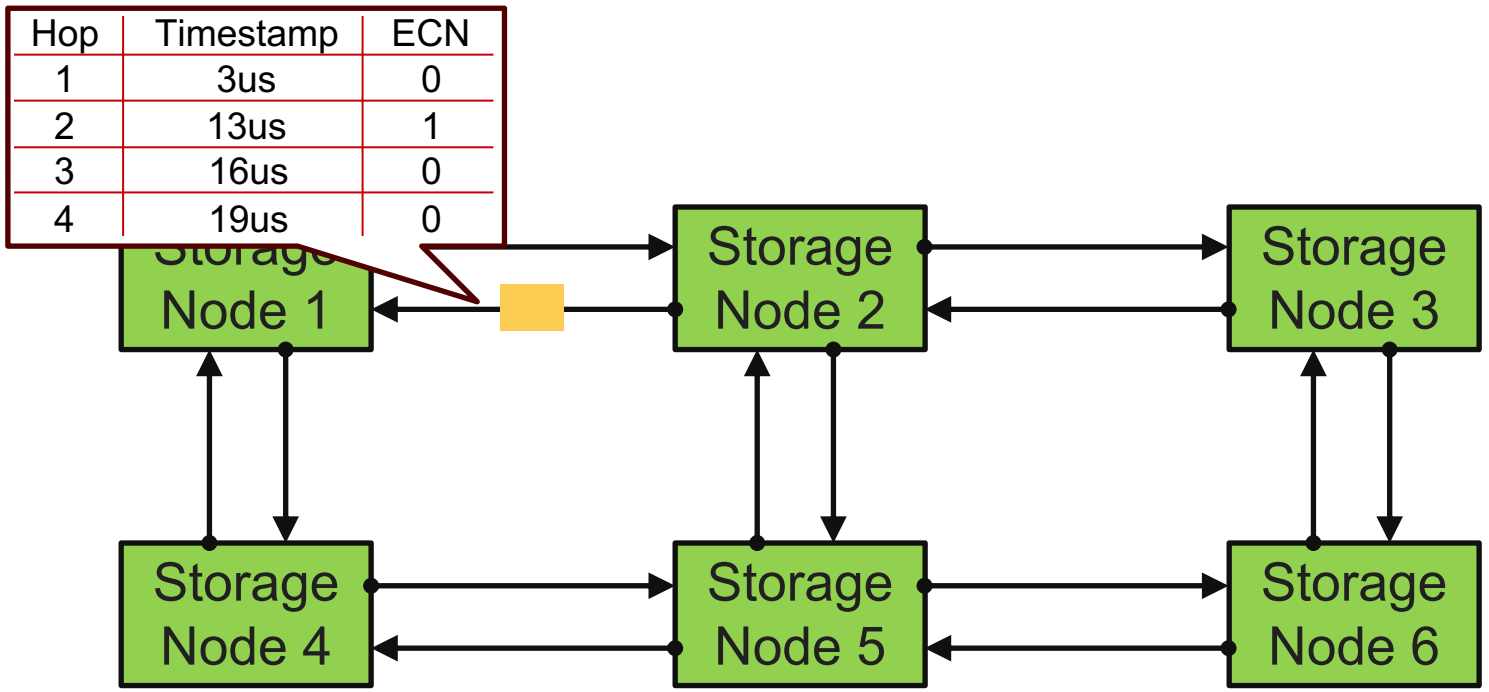


INT-Assisted Symmetric Routing (ISR)

(In-Network Telemetry)

Goal: Pick the least-congested path to the target SSD

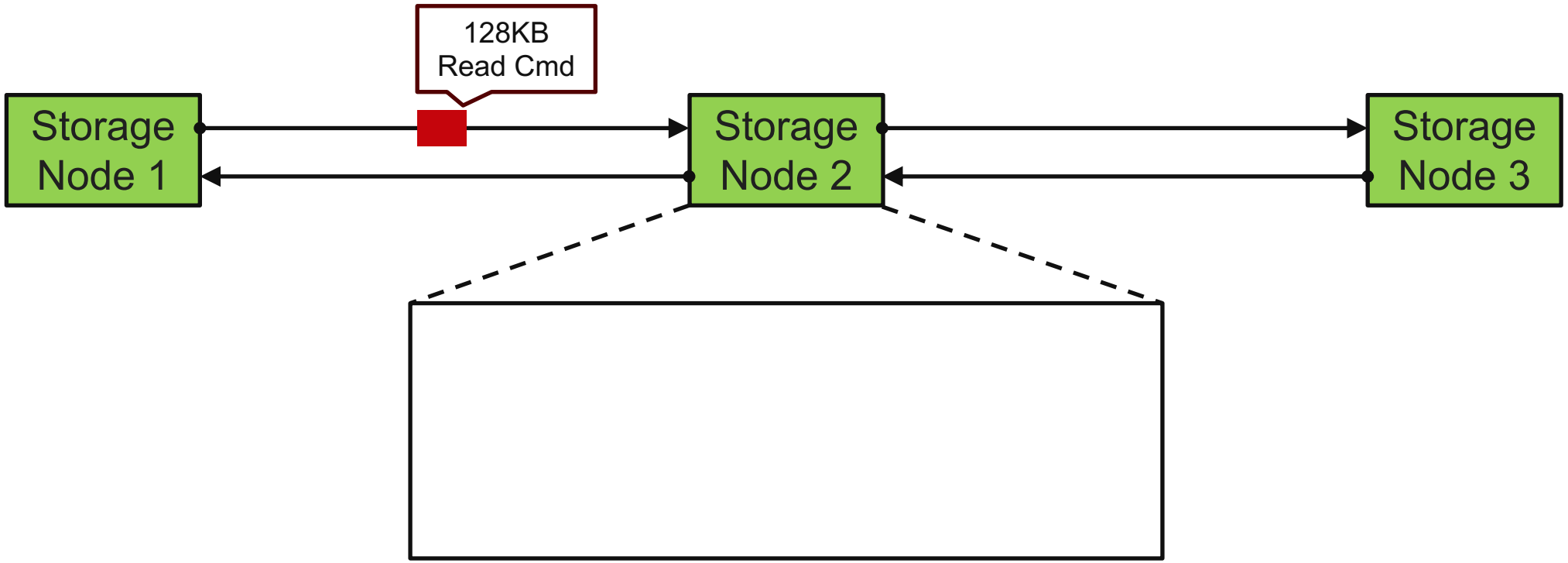
Idea: Each completion packet carriers forward-path telemetry back, hop-by-hop.



Eager Bandwidth Reservation

Goal: Pre-allocate BW for the response before it arrives

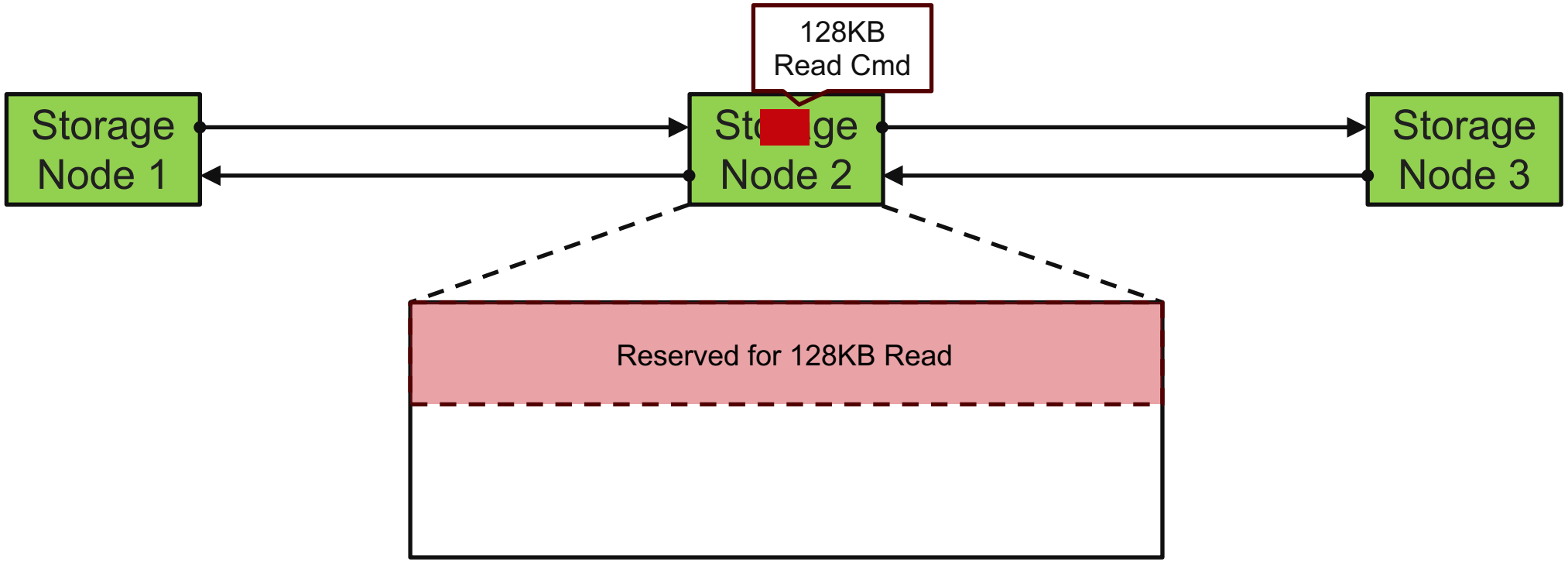
Idea: Exploit pairwise + asymmetric to predict the request response size



Eager Bandwidth Reservation

Goal: Pre-allocate BW for the response before it arrives

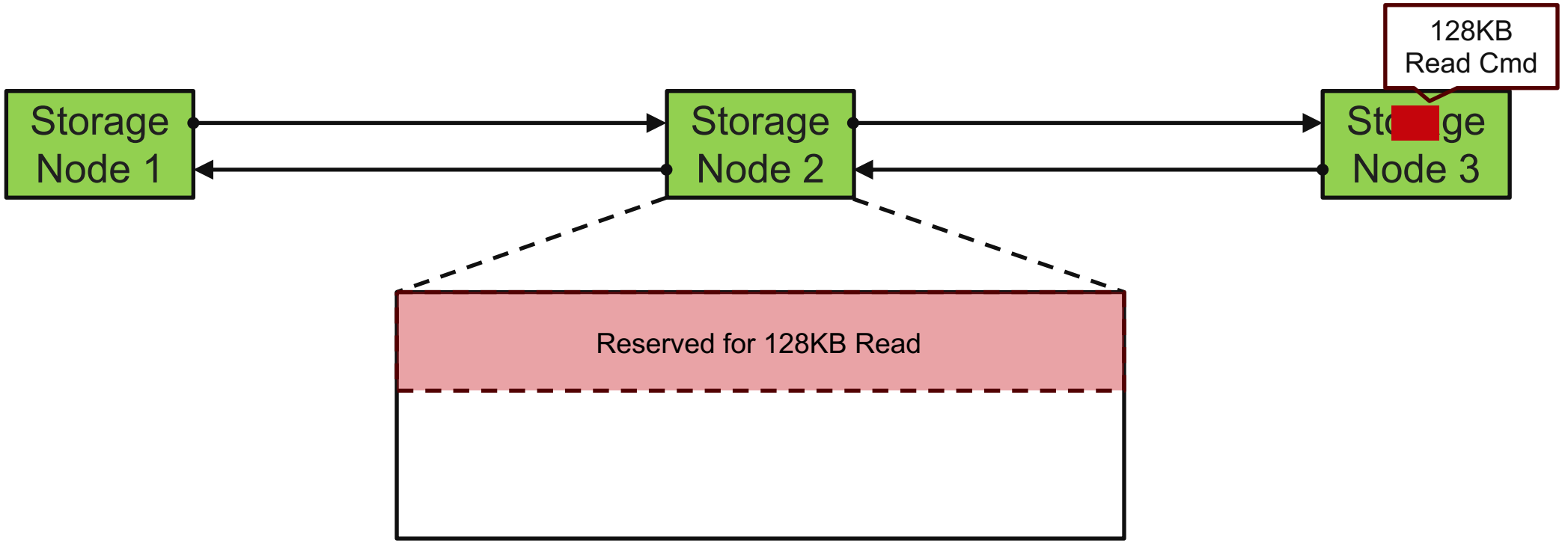
Idea: Exploit pairwise + asymmetric to predict the request response size



Eager Bandwidth Reservation

Goal: Pre-allocate BW for the response before it arrives

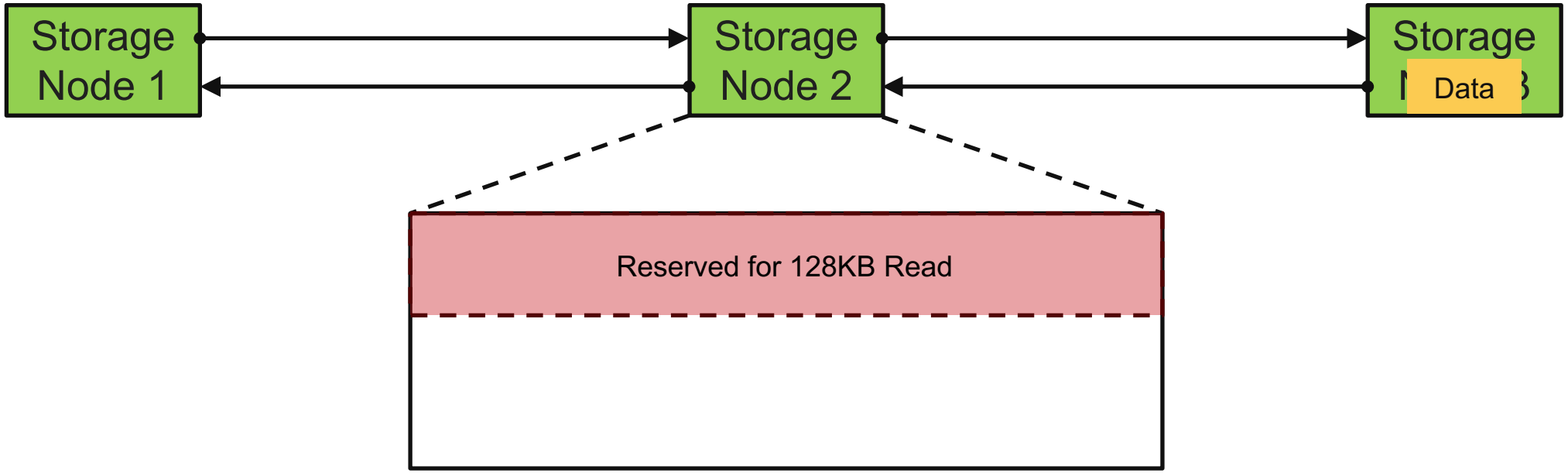
Idea: Exploit pairwise + asymmetric to predict the request response size



Eager Bandwidth Reservation

Goal: Pre-allocate BW for the response before it arrives

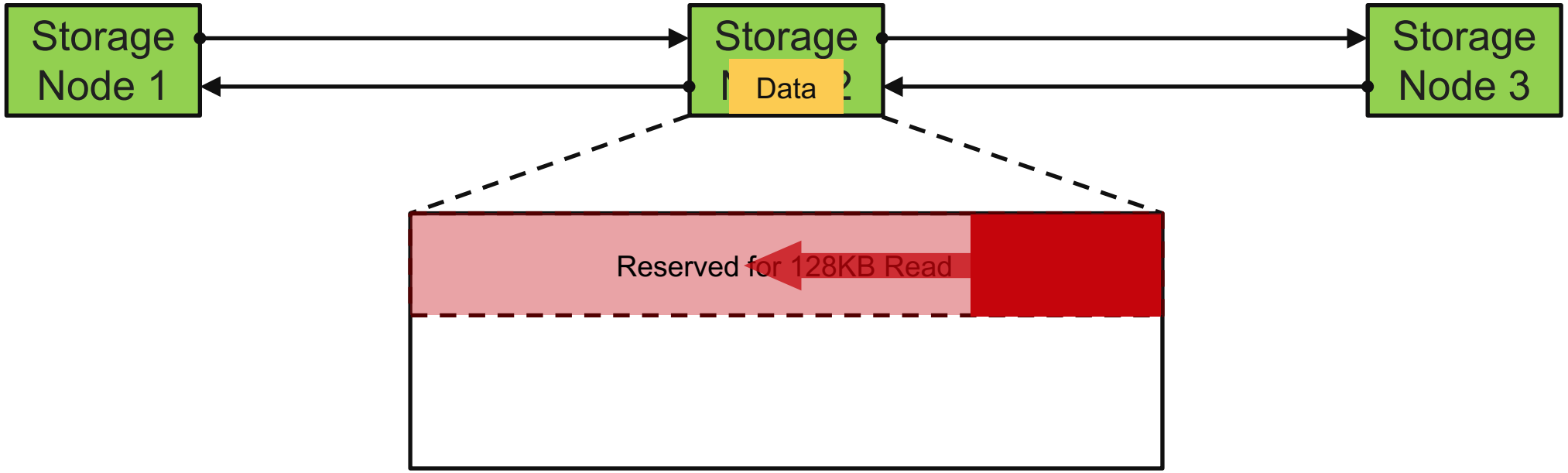
Idea: Exploit pairwise + asymmetric to predict the request response size



Eager Bandwidth Reservation

Goal: Pre-allocate BW for the response before it arrives

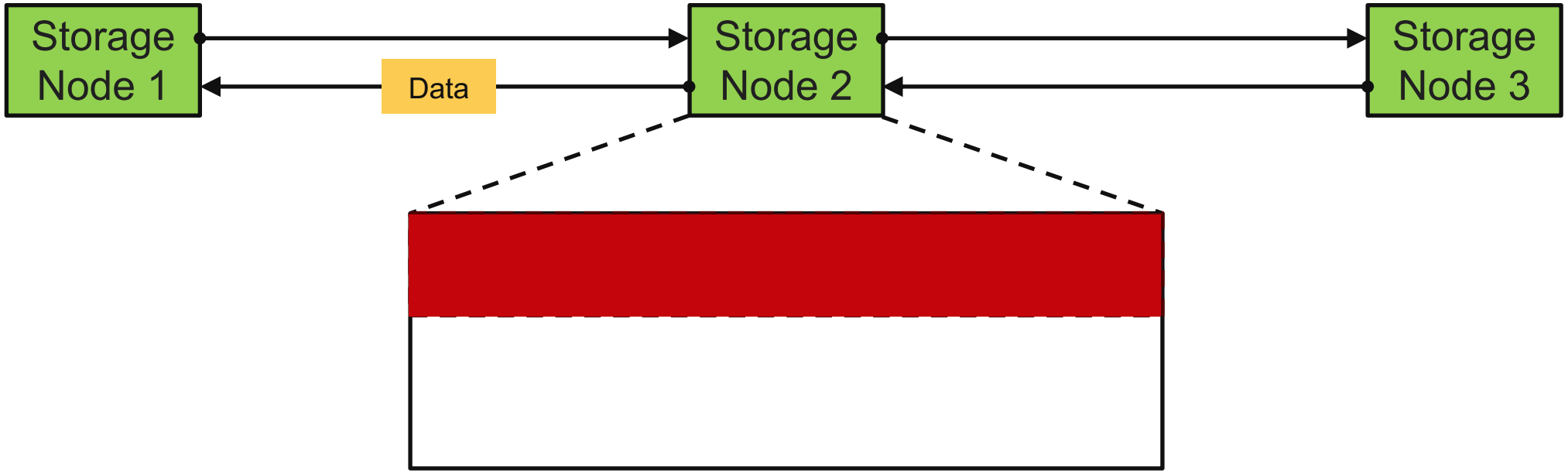
Idea: Exploit pairwise + asymmetric to predict the request response size



Eager Bandwidth Reservation

Goal: Pre-allocate BW for the response before it arrives

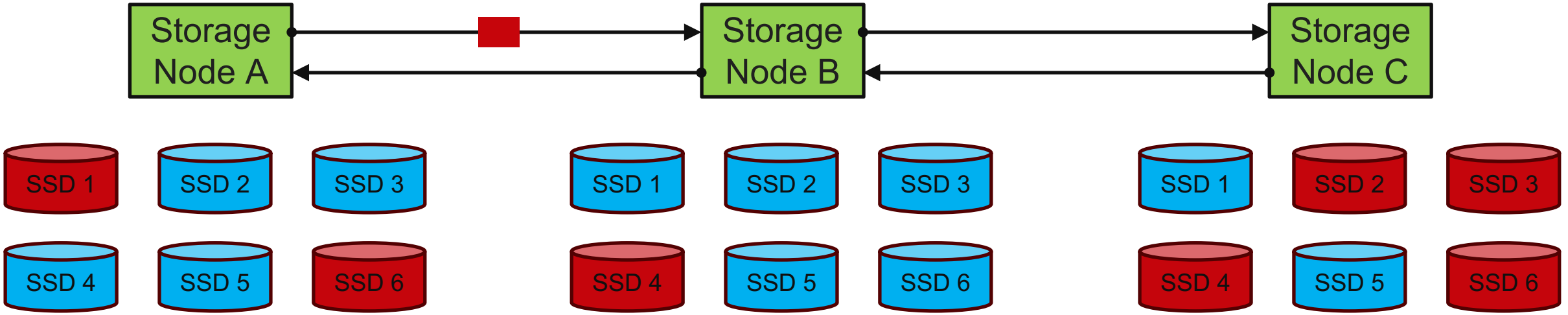
Idea: Exploit pairwise + asymmetric to predict the request response size



Storage-Driven Traffic Scheduling

Goal: Make the network aware of SSD load in real time

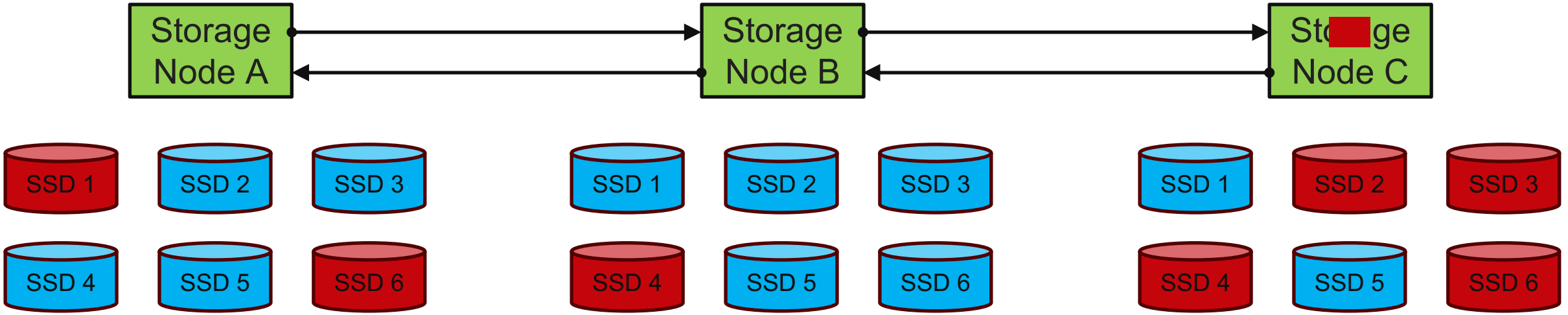
Idea: Storage encodes its bandwidth headroom as credits in completion packets



Storage-Driven Traffic Scheduling

Goal: Make the network aware of SSD load in real time

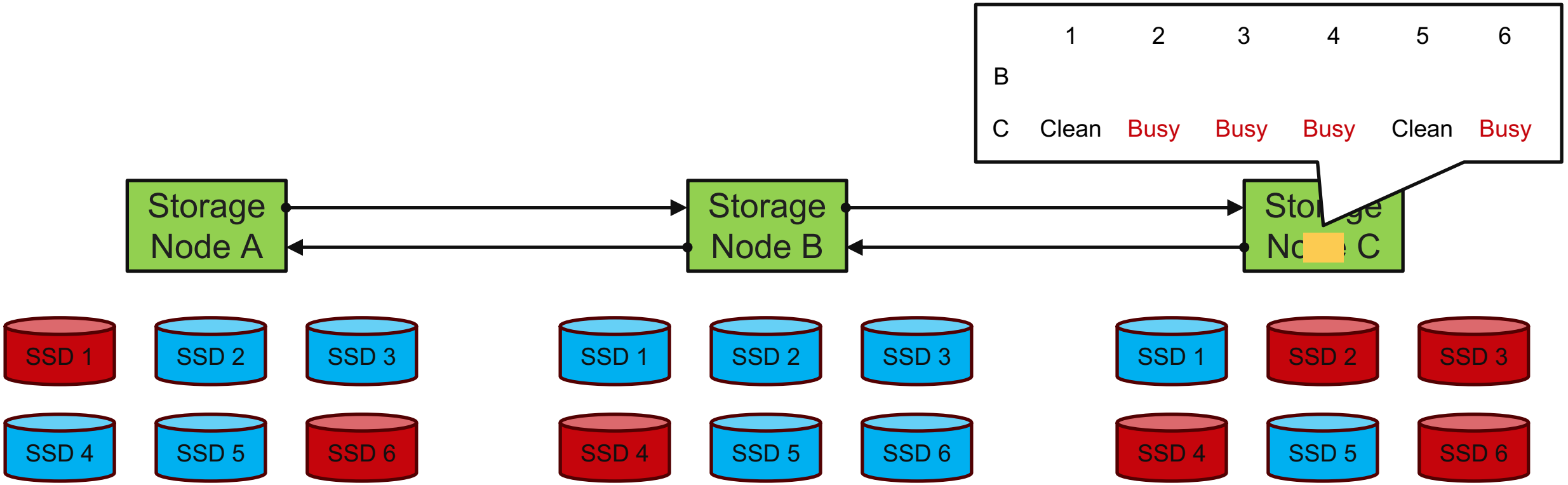
Idea: Storage encodes its bandwidth headroom as credits in completion packets



Storage-Driven Traffic Scheduling

Goal: Make the network aware of SSD load in real time

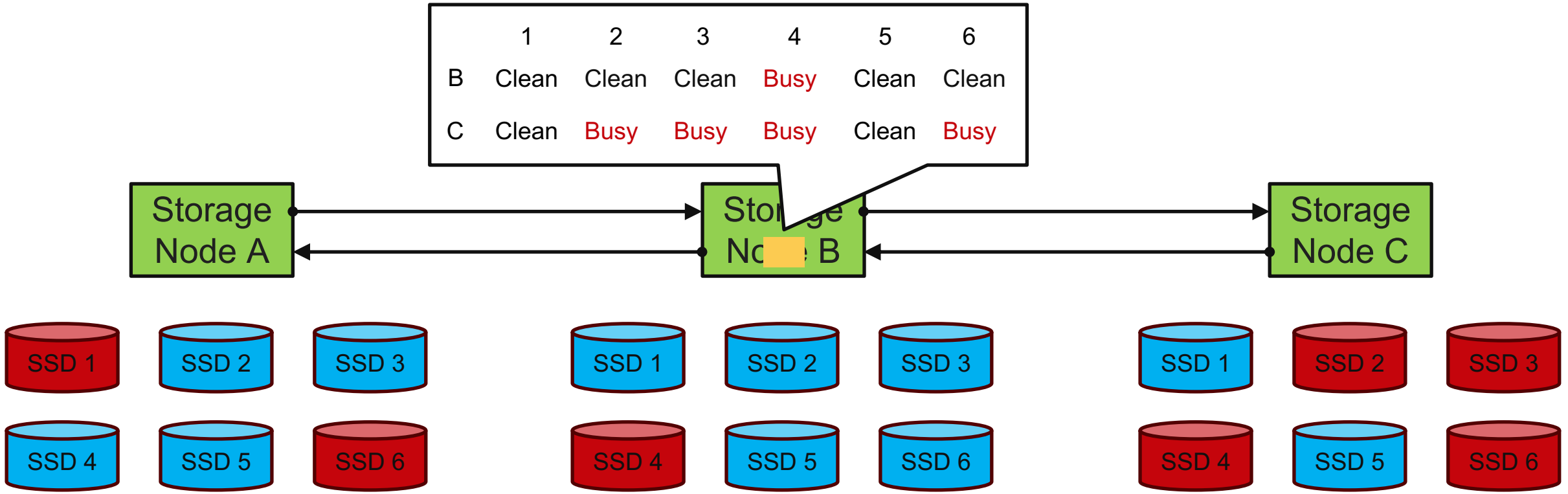
Idea: Storage encodes its bandwidth headroom as credits in completion packets



Storage-Driven Traffic Scheduling

Goal: Make the network aware of SSD load in real time

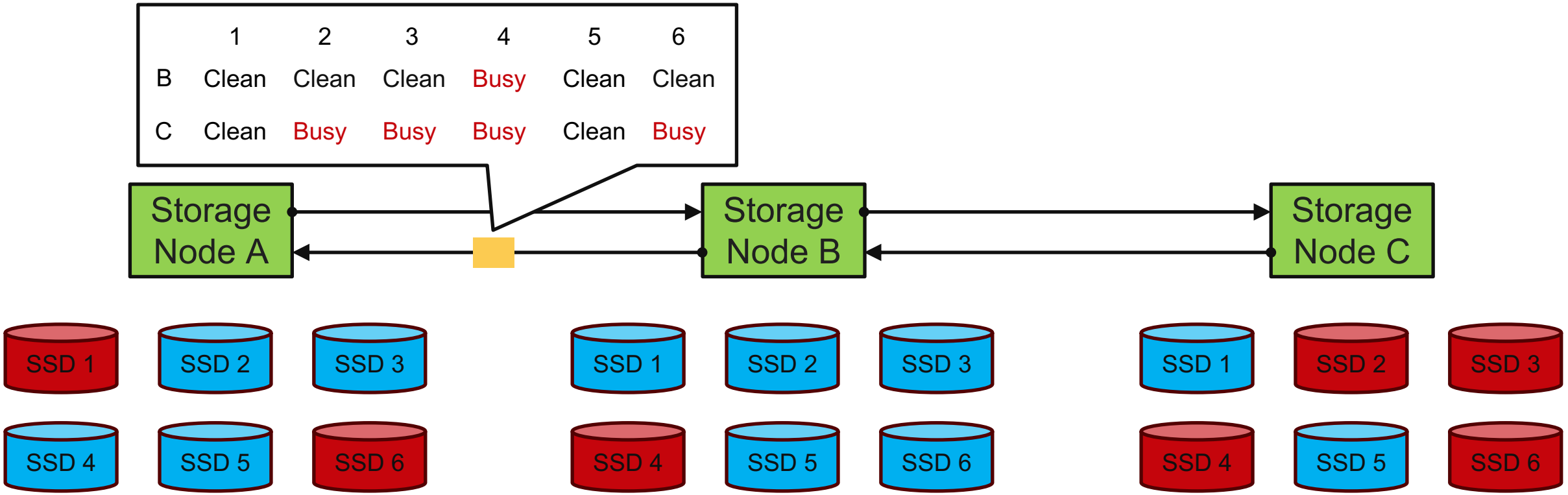
Idea: Storage encodes its bandwidth headroom as credits in completion packets



Storage-Driven Traffic Scheduling

Goal: Make the network aware of SSD load in real time

Idea: Storage encodes its bandwidth headroom as credits in completion packets



Evaluation (Switched vs. Switchless) Setup

Common Setup

- 4 storage nodes
- Storage server: 2× Intel Xeon, 256 GB DDR4
- NIC: dual-port 100GbE ConnectX-6
- Drives: 4× Samsung PM9A3 960 GB NVMe SSD
- Protocol: NVMe-over-TCP

Switched

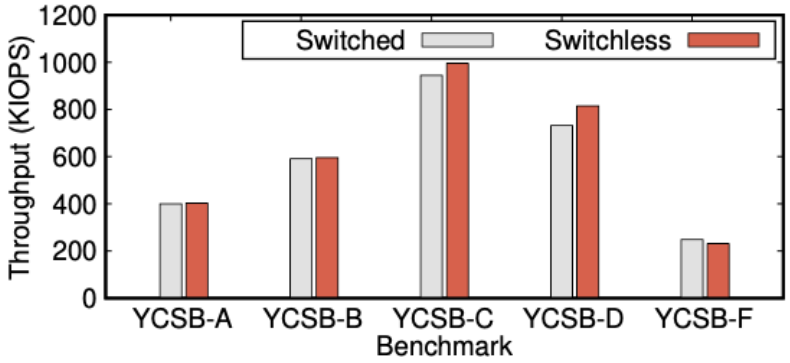
- Dell Z9100-ON 32-port 100GbE ToR
- Aurora 710 Tofino P4 for ISR / PIFO
- Single client–ToR path

Switchless

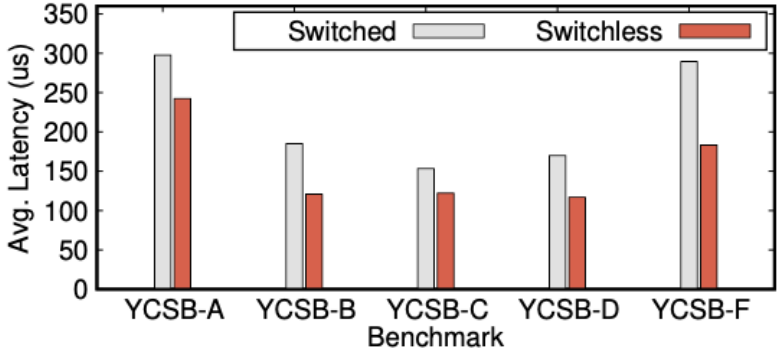
- BlueField-2 SmartNIC adapter
- Same NIC substrate as ConnectX-6
- 2D Torus, 4 disjoint paths

Performance: Throughput and Latency

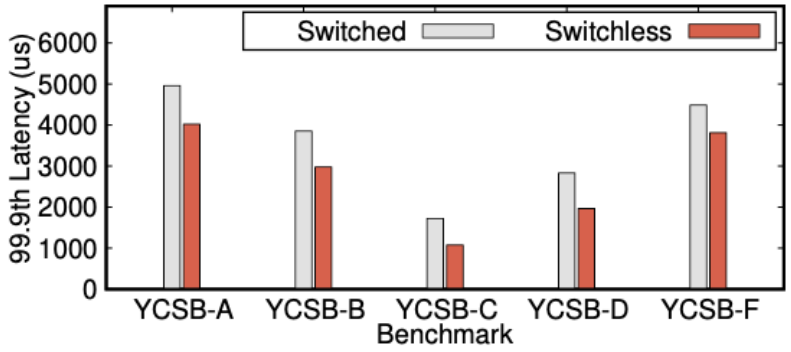
- Switchless SAN matches the switched one in terms of throughput on all YCSB workloads.
- Switchless SAN reduces average read latency by 28.3%.
- Switchless SAN cuts 99.9th-percentile read latency by 25%.



(a) Throughput.



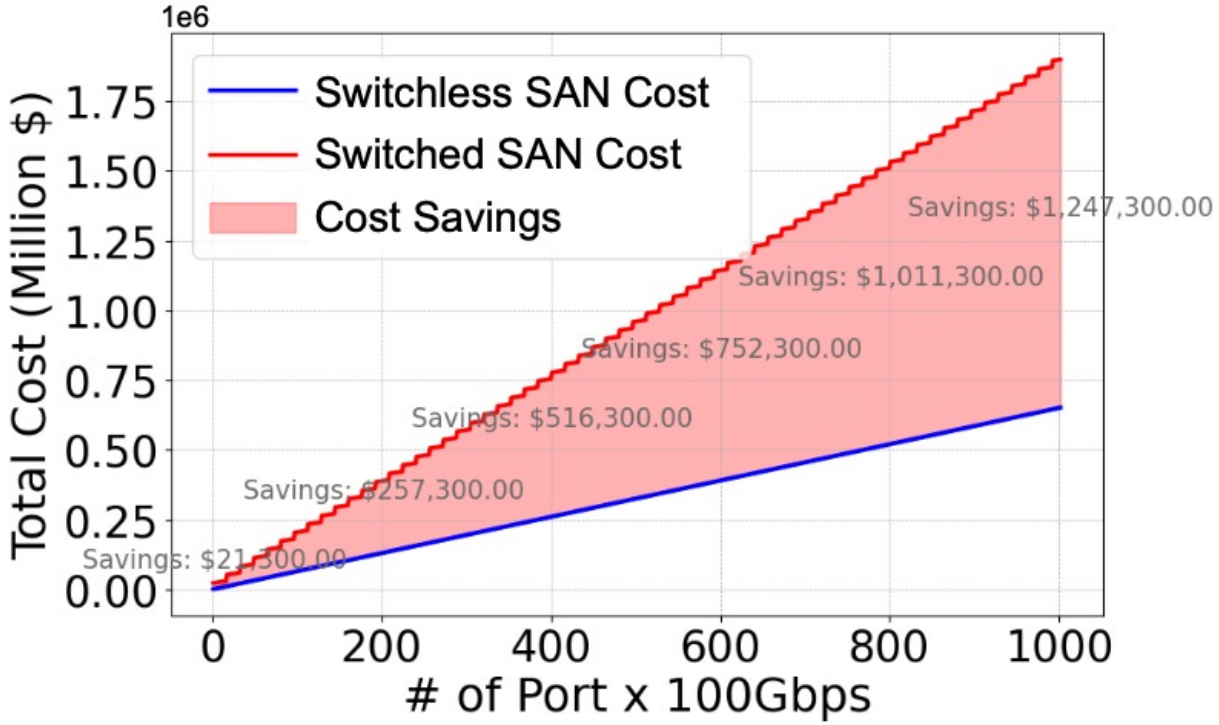
(b) Average read latency.



(c) 99.9th read latency.

Cost Analysis

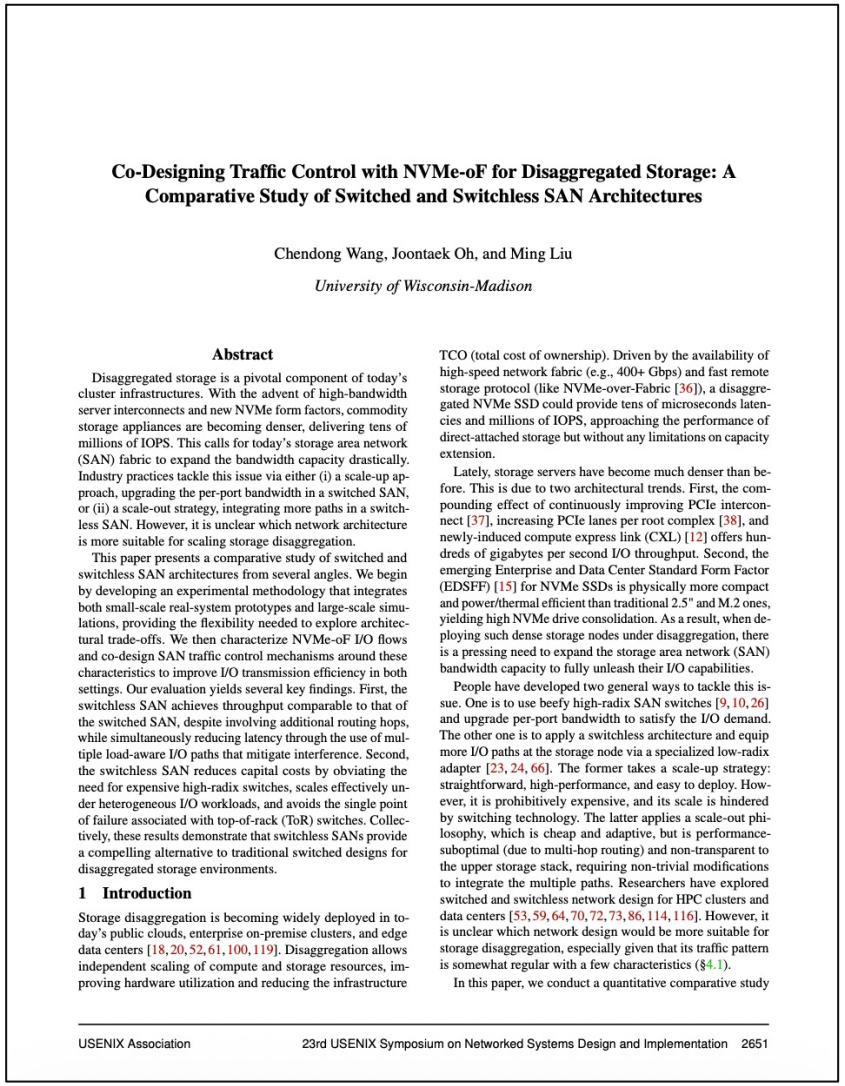
- Gap widens linearly with cluster scale.
- Switchless SAN achieves 50.5% capital savings.



More Evaluations

- For more details,

  please check the paper.



Co-Designing Traffic Control with NVMe-oF for Disaggregated Storage: A Comparative Study of Switched and Switchless SAN Architectures

Chendong Wang, Joontaek Oh, and Ming Liu

University of Wisconsin-Madison

Abstract

Disaggregated storage is a pivotal component of today's cluster infrastructures. With the advent of high-bandwidth server interconnects and new NVMe form factors, commodity storage appliances are becoming denser, delivering tens of millions of IOPS. This calls for today's storage area network (SAN) fabric to expand the bandwidth capacity drastically. Industry practices tackle this issue via either (i) a scale-up approach, upgrading the per-port bandwidth in a switched SAN, or (ii) a scale-out strategy, integrating more paths in a switchless SAN. However, it is unclear which network architecture is more suitable for scaling storage disaggregation.

This paper presents a comparative study of switched and switchless SAN architectures from several angles. We begin by developing an experimental methodology that integrates both small-scale real-system prototypes and large-scale simulations, providing the flexibility needed to explore architectural trade-offs. We then characterize NVMe-oF I/O flows and co-design SAN traffic control mechanisms around these characteristics to improve I/O transmission efficiency in both settings. Our evaluation yields several key findings. First, the switchless SAN achieves throughput comparable to that of the switched SAN, despite involving additional routing hops, while simultaneously reducing latency through the use of multiple load-aware I/O paths that mitigate interference. Second, the switchless SAN reduces capital costs by obviating the need for expensive high-radix switches, scales effectively under heterogeneous I/O workloads, and avoids the single point of failure associated with top-of-rack (ToR) switches. Collectively, these results demonstrate that switchless SANs provide a compelling alternative to traditional switched designs for disaggregated storage environments.

1 Introduction

Storage disaggregation is becoming widely deployed in today's public clouds, enterprise on-premise clusters, and edge data centers [18, 20, 52, 61, 100, 119]. Disaggregation allows independent scaling of compute and storage resources, improving hardware utilization and reducing the infrastructure TCO (total cost of ownership). Driven by the availability of high-speed network fabric (e.g., 400+ Gbps) and fast remote storage protocol (like NVMe-over-Fabric [36]), a disaggregated NVMe SSD could provide tens of microseconds latencies and millions of IOPS, approaching the performance of direct-attached storage but without any limitations on capacity extension.

Lately, storage servers have become much denser than before. This is due to two architectural trends. First, the compounding effect of continuously improving PCIe interconnect [37], increasing PCIe lanes per root complex [38], and newly-induced compute express link (CXL) [12] offers hundreds of gigabytes per second I/O throughput. Second, the emerging Enterprise and Data Center Standard Form Factor (EDSFF) [15] for NVMe SSDs is physically more compact and power/thermal efficient than traditional 2.5" and M.2 ones, yielding high NVMe drive consolidation. As a result, when deploying such dense storage nodes under disaggregation, there is a pressing need to expand the storage area network (SAN) bandwidth capacity to fully unleash their I/O capabilities.

People have developed two general ways to tackle this issue. One is to use beefy high-radix SAN switches [9, 10, 26] and upgrade per-port bandwidth to satisfy the I/O demand. The other one is to apply a switchless architecture and equip more I/O paths at the storage node via a specialized low-radix adapter [23, 24, 66]. The former takes a scale-up strategy: straightforward, high-performance, and easy to deploy. However, it is prohibitively expensive, and its scale is hindered by switching technology. The latter applies a scale-out philosophy, which is cheap and adaptive, but is performance-suboptimal (due to multi-hop routing) and non-transparent to the upper storage stack, requiring non-trivial modifications to integrate the multiple paths. Researchers have explored switched and switchless network design for HPC clusters and data centers [53, 59, 64, 70, 72, 73, 86, 114, 116]. However, it is unclear which network design would be more suitable for storage disaggregation, especially given that its traffic pattern is somewhat regular with a few characteristics (§4.1).

In this paper, we conduct a quantitative comparative study

Conclusion

- Three NVMe-oF characteristics
  - Per-IO: Network RTT ≪ Storage RTT
  - Per-Command Pair: Pairwise and Asymmetry
  - Per-Storage Session: Backward-Propagated Queueing
- Co-designed traffic control with NVMe-oF
  - INT-assisted symmetric routing
  - Eager bandwidth reservation
  - Storage-driven flow scheduling.
- Perform a comparative study of switched and switchless SAN architecture

Thank you