Designing Toward NVMe-aware Distributed Storage System

Samsung Electronics

Sungmin Lee

OpenInfra
Community Day
Korea 2021

Contents

Agenda

Recent Datacenter Trends

What Are We Focusing on?

Global Deduplication

Storage Disaggregation

Summary

Contents

Agenda

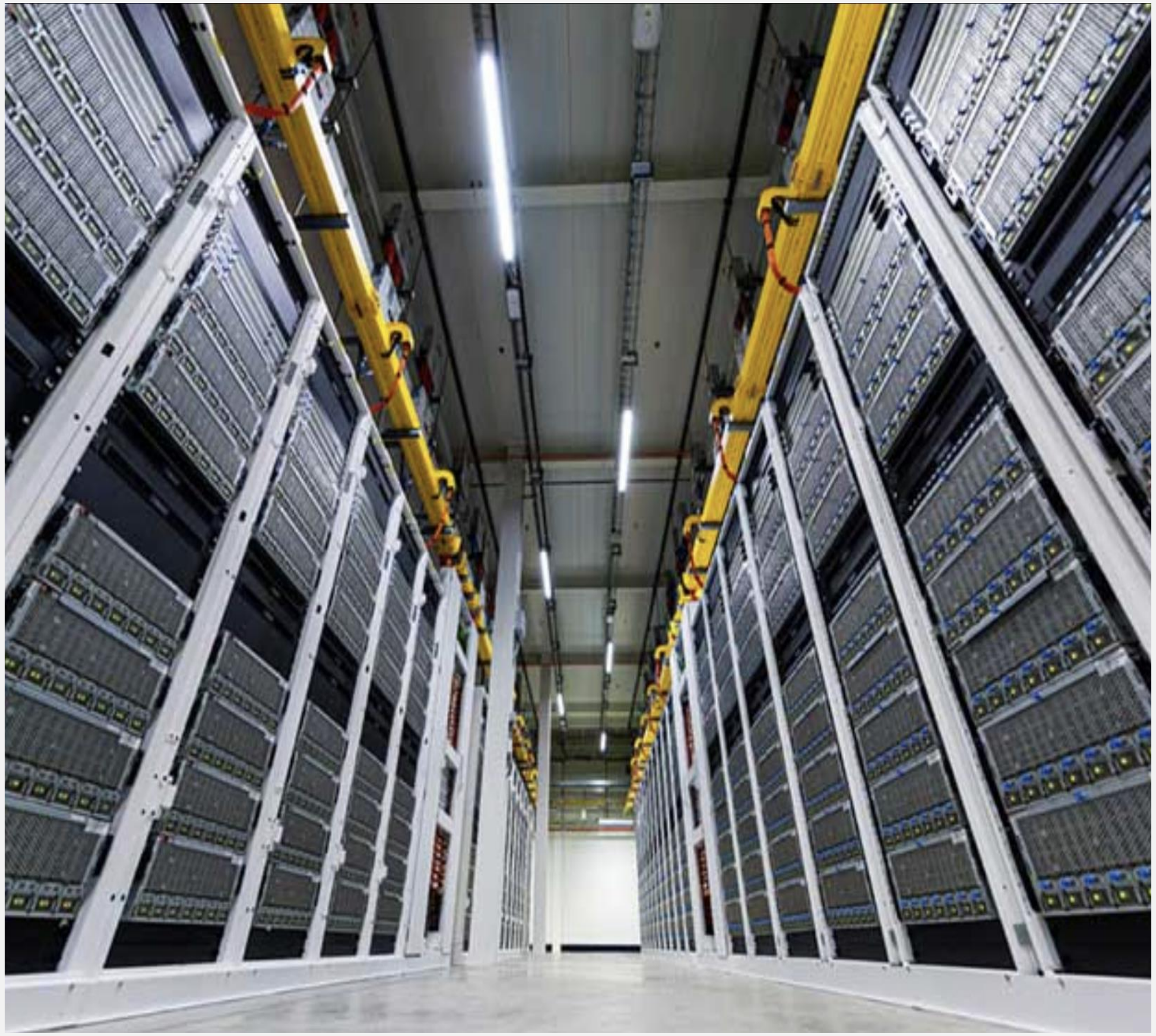
Recent Datacenter Trends

What Are We Focusing on?

Global Deduplication

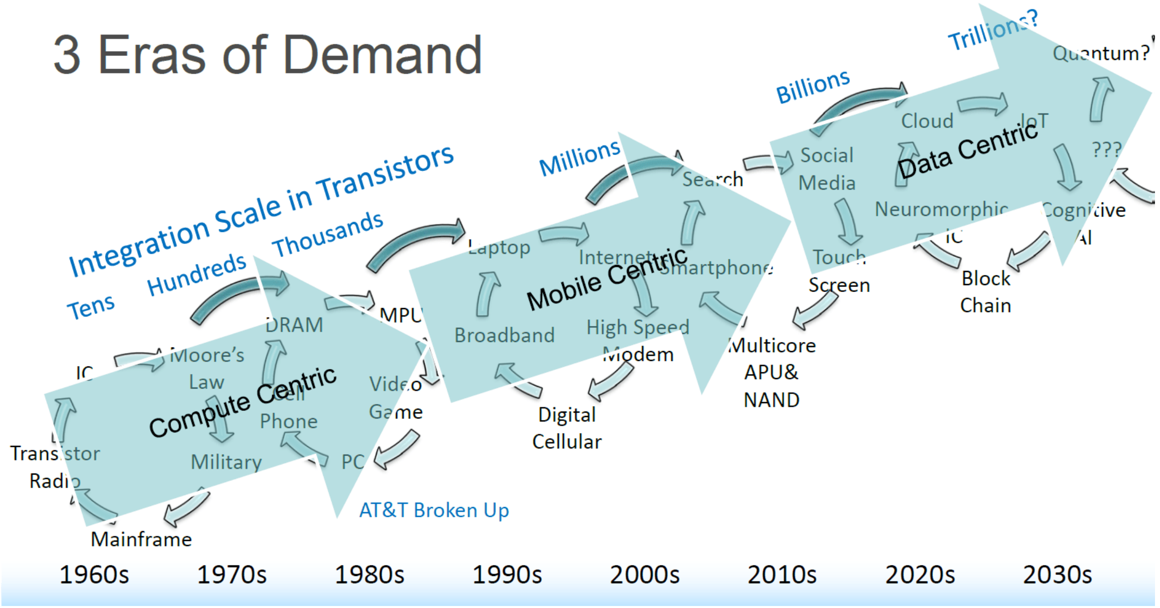
Storage Disaggregation

Summary

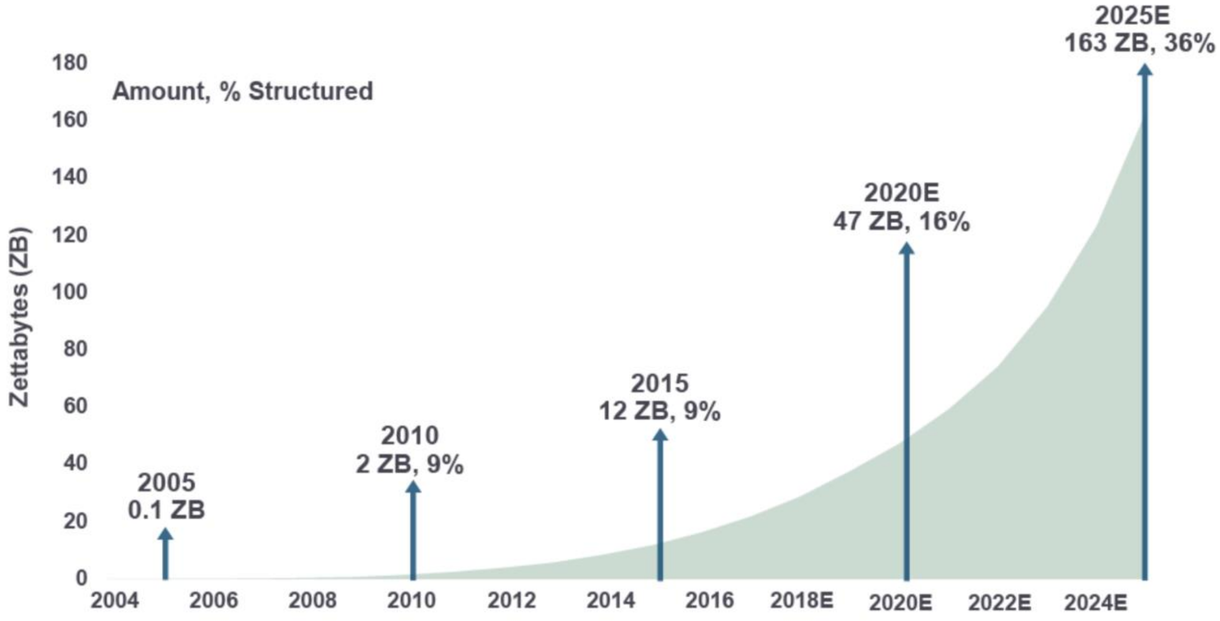


Source: Microsoft

Data-Centric Era



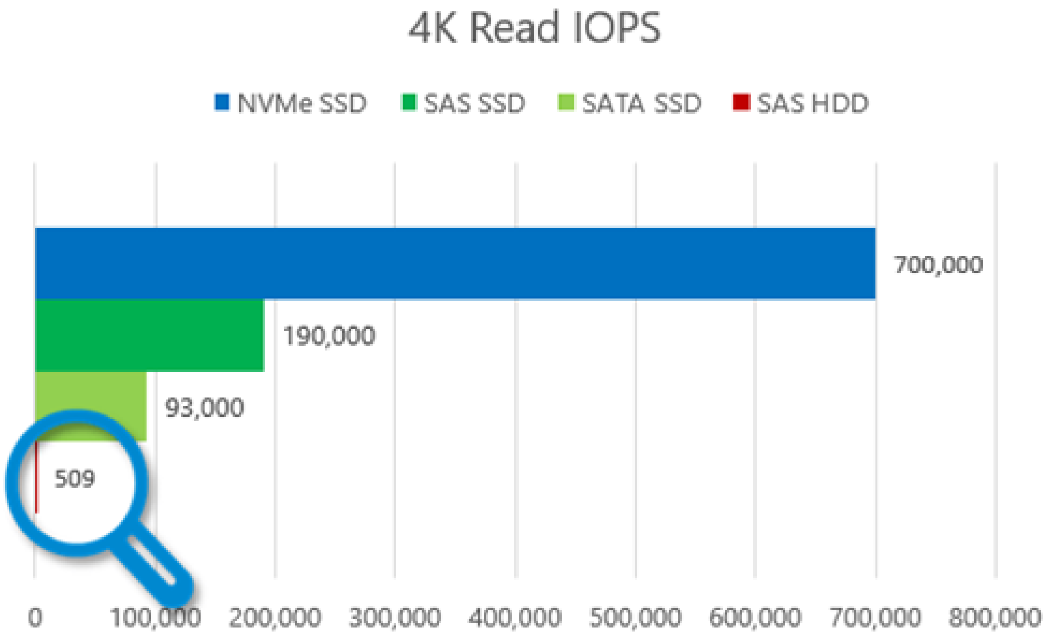
Source: VLSI Research, ISS US, January 2018



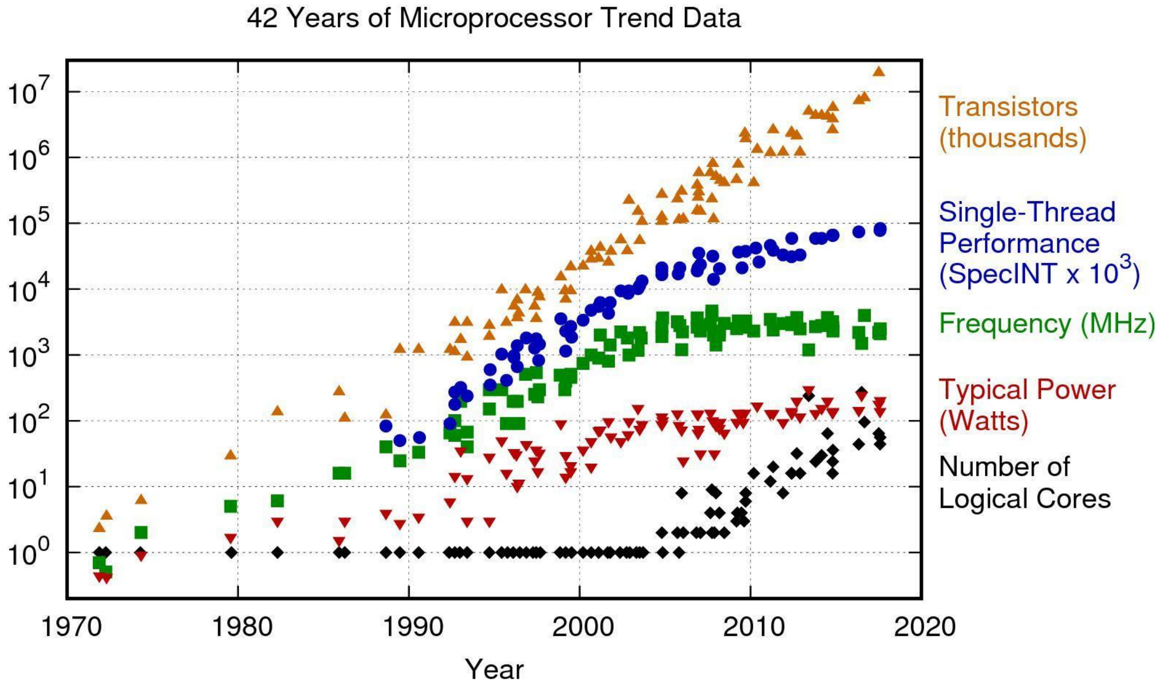
Source: https://www.monsoonblockchainstorage.com/data-growth/

- **Spending towards cloud storage is growing beyond $1 Trillion in each year**
- **Information Created Worldwide Expected to reach 163 Zettabytes by 2025**

CPUs are Beyond it's Power



Source: https://www.micron.com/about/blog/2017/july/the-business-case-for-nvme-pcie-ssds



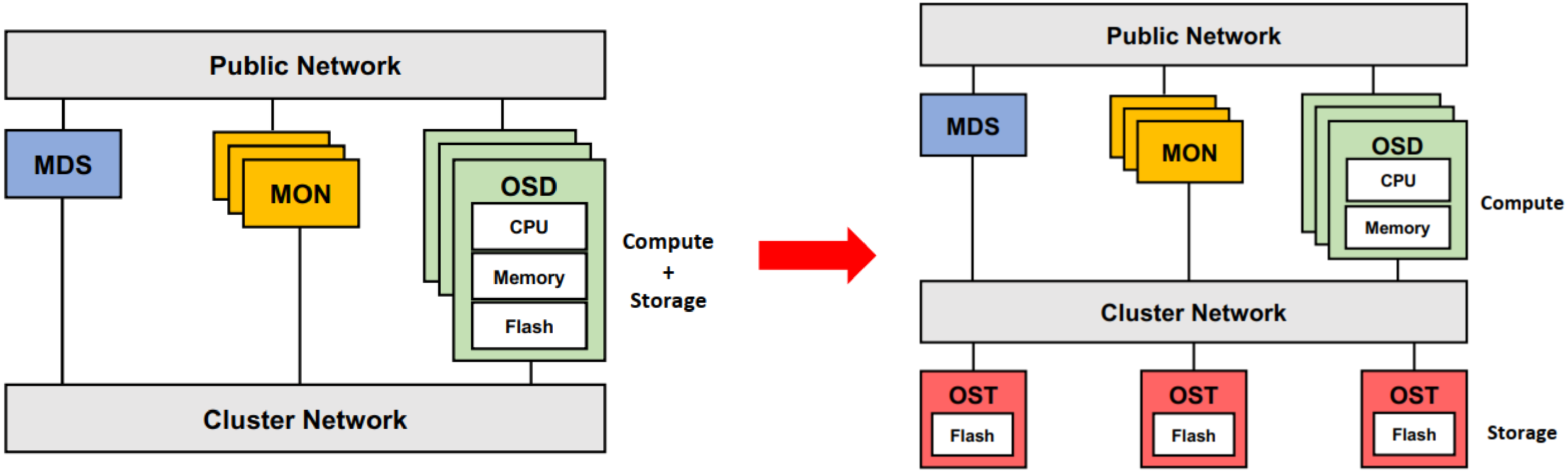
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2017 by K. Rupp

- **Storage is getting faster rapidly but, CPU isn't**
- **Advances in CPU technology slowed down due to the power wall**
- **Single core performance ⬇ vs. Single storage device performance ⬆**

Storage Disaggregation

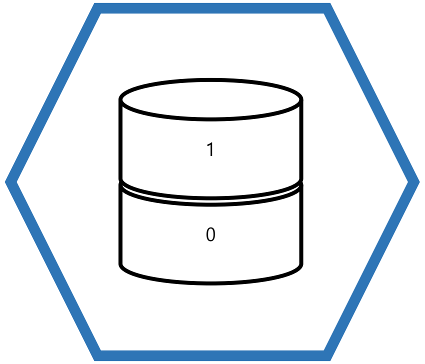
Storage Disaggregation with NVMe-oF

- Separates servers into compute and storage nodes components
- Any-to-any access among components
- Independent resource scaling
- NVMe-oF enables remote I/O operation with line speed



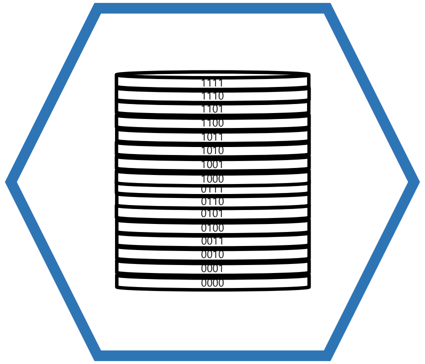
Storage Device Diversification

- **Storage media is going more diverse**



Fast NVMe

- ZSSD, Optane
- Low latency
- Fast

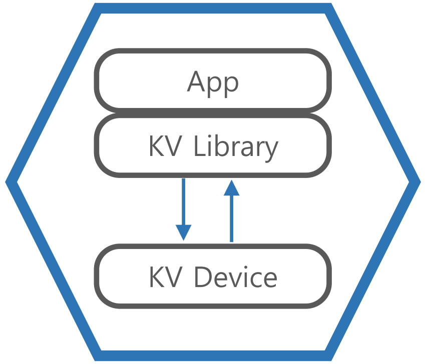


Large-density

- QLC SSD
- Large capacity
- Low price
- Slow

Zoned-Namespace

- ZNS SSD
- Separate write by zone
- Append-only write
- No GC, WAF ↓, Lifetime ↑



Key-Value

- Key Value SSD
- Enable direct KV I/F
- Shorten SW stack

Contents

Agenda

Recent Datacenter Trends

What Are We Focusing on?

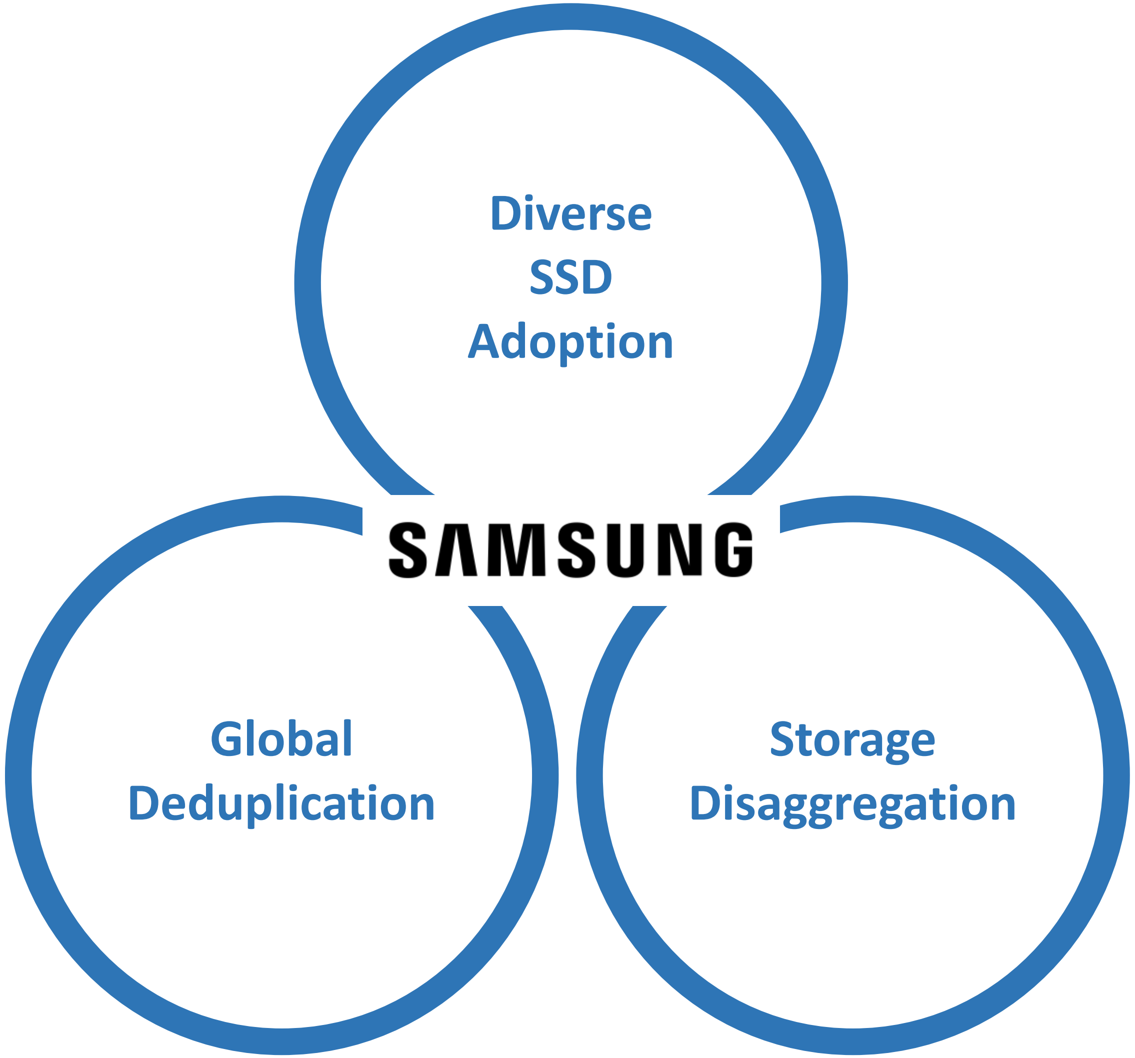
Global Deduplication

Storage Disaggregation

Summary

Source: Ceph Pacific Release

Ceph with Datacenter Trends



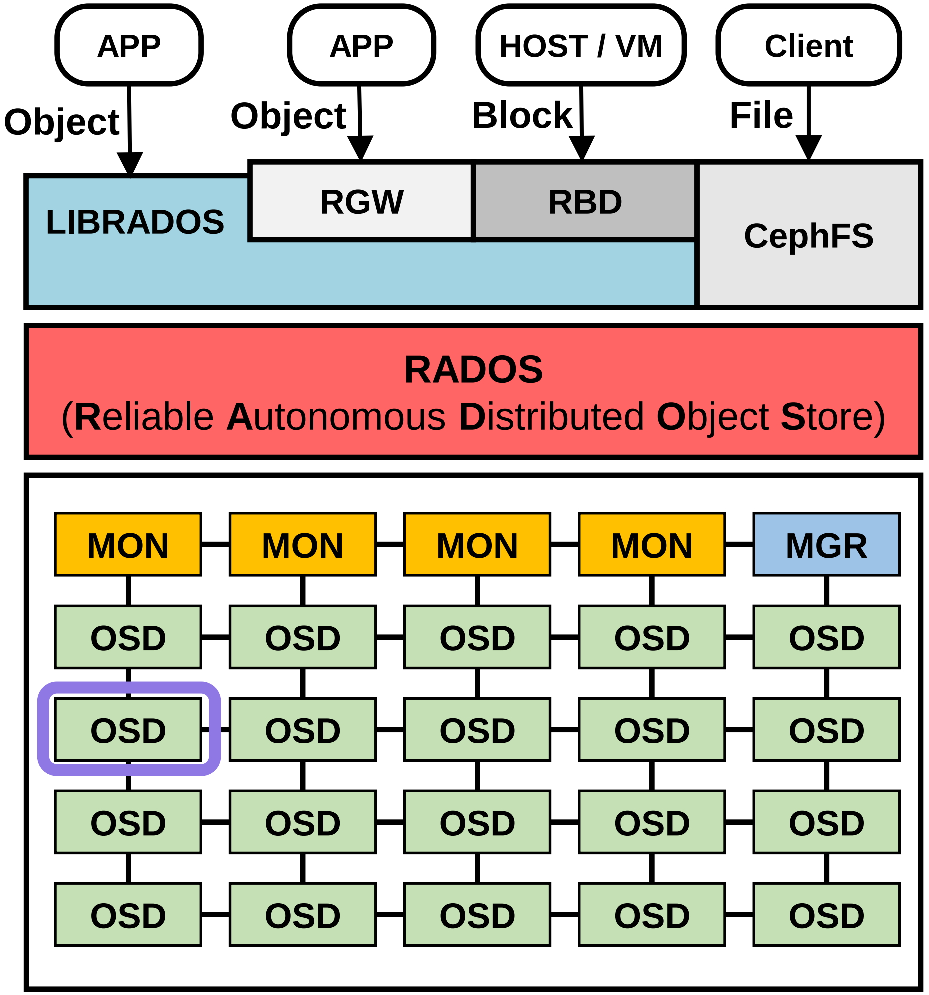
Diverse SSD Adoption

Global Deduplication

Storage Disaggregation

Ceph

- **Open-source software-defined object storage system**
  - Reliable storage service out of unreliable components
    - No single point of failure
    - Data durability via replication or erasure coding
    - Fault tolerance
  - Scalable storage service
- **Provides 3-in-1 interfaces:**
  - Object-
  - Block-
  - File-
- **RADOS is the core component**
  - Monitor (MON)
  - **Object Store Daemon (OSD)**
  - Manager (MGR)
  - Metadata Server (MDS) for CephFS



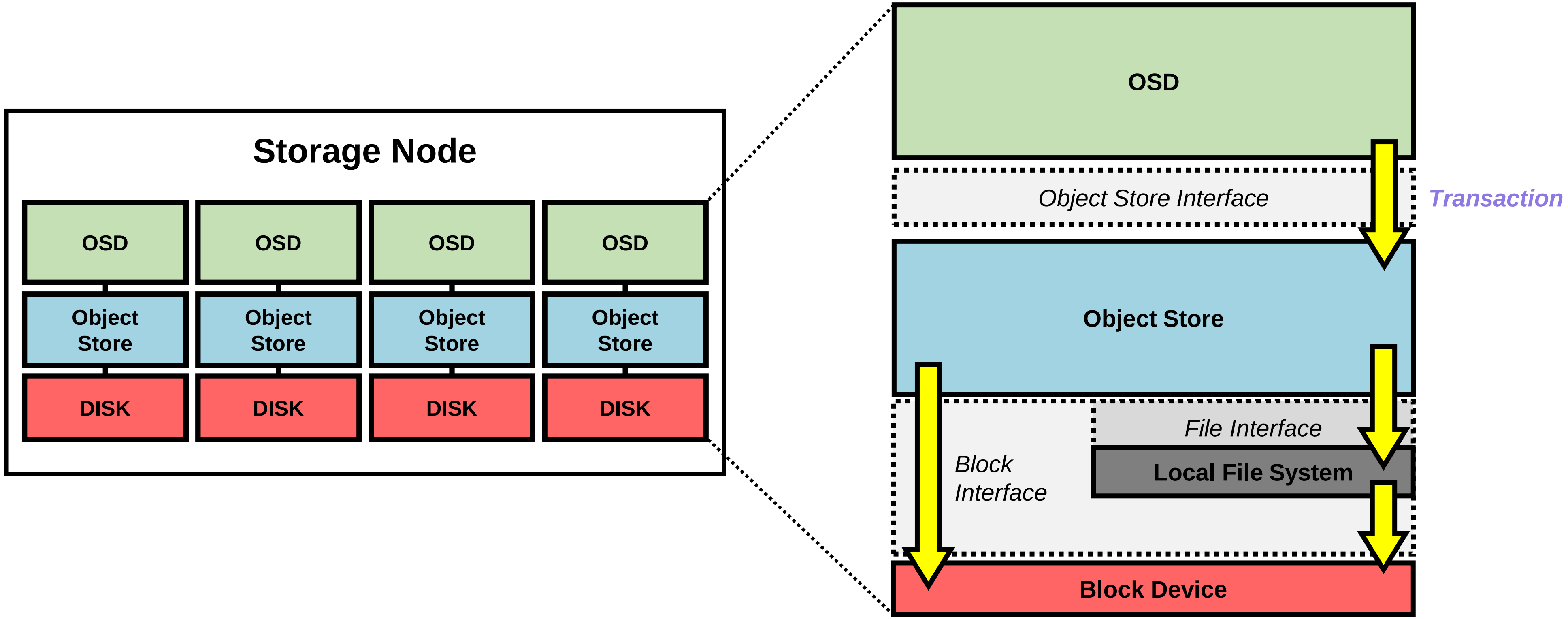
OSD and Object Store

OSD

- Responsible for
  - Storing and retrieving objects
  - Providing access to them over network
  - Peering, Replication, Recovery, etc.

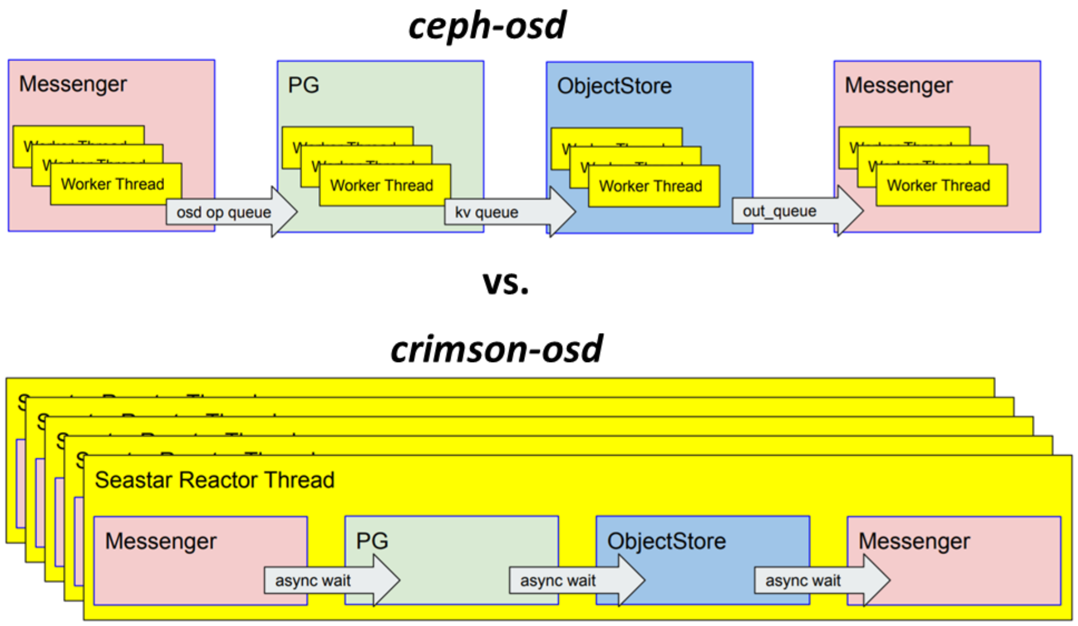
Object Store

- Storage backend for OSD
  - Storing and retrieving objects in the storage device attached to physical machine
  - Transaction support



Crimson

- **Minimize CPU overhead**
  - Minimize cycles/iop
  - Minimize cross-core communication
  - Minimize copies
  - Bypass kernel, avoid context switch
- **Enable emerging storage devices**
  - Zoned Namespace
  - Persistent Memory
  - Fast NVMe



Source: Vault 20, Crimson: A New Ceph OSD for the Age of Persistent Memory and Fast NVMe Storage

Contents

Agenda

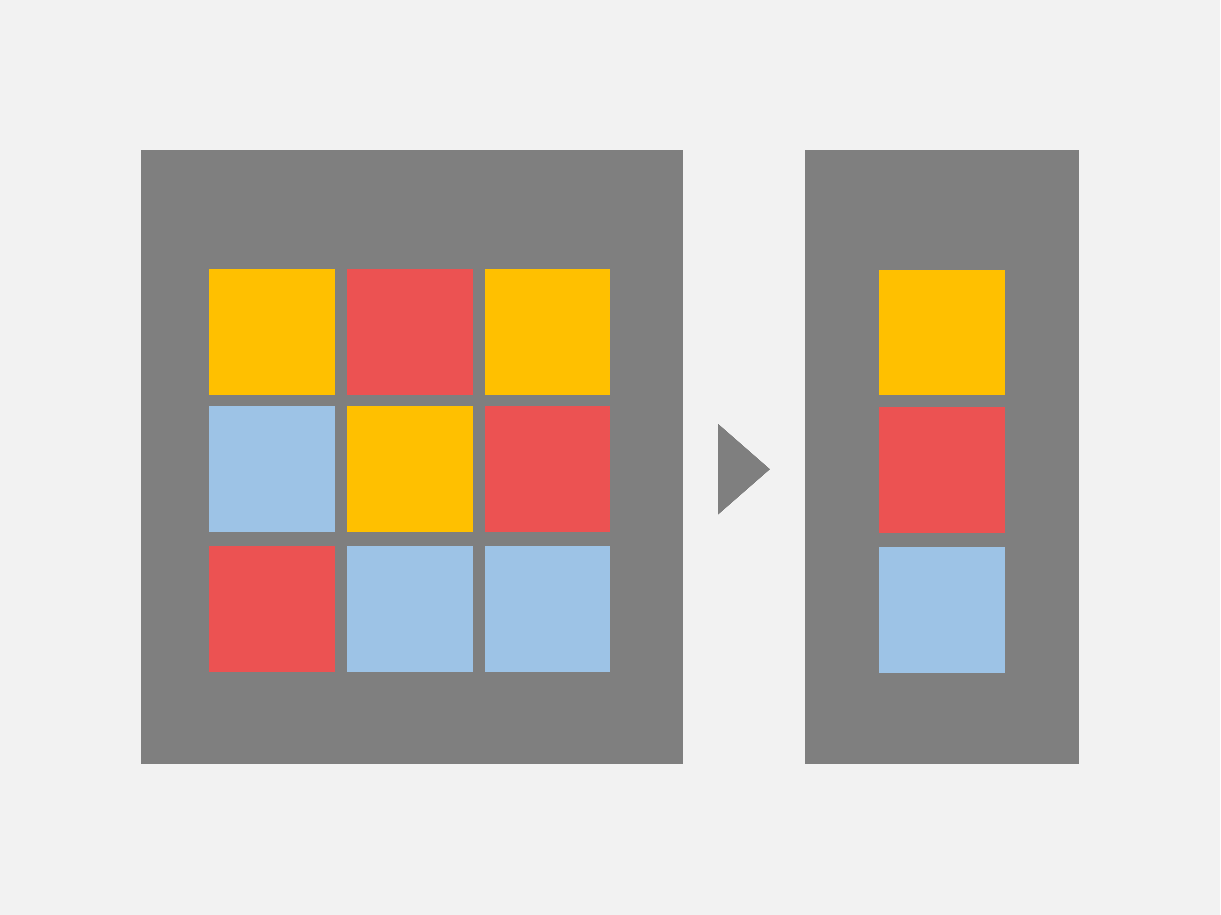
Recent Datacenter Trends

What We are Focusing on?

Global Deduplication

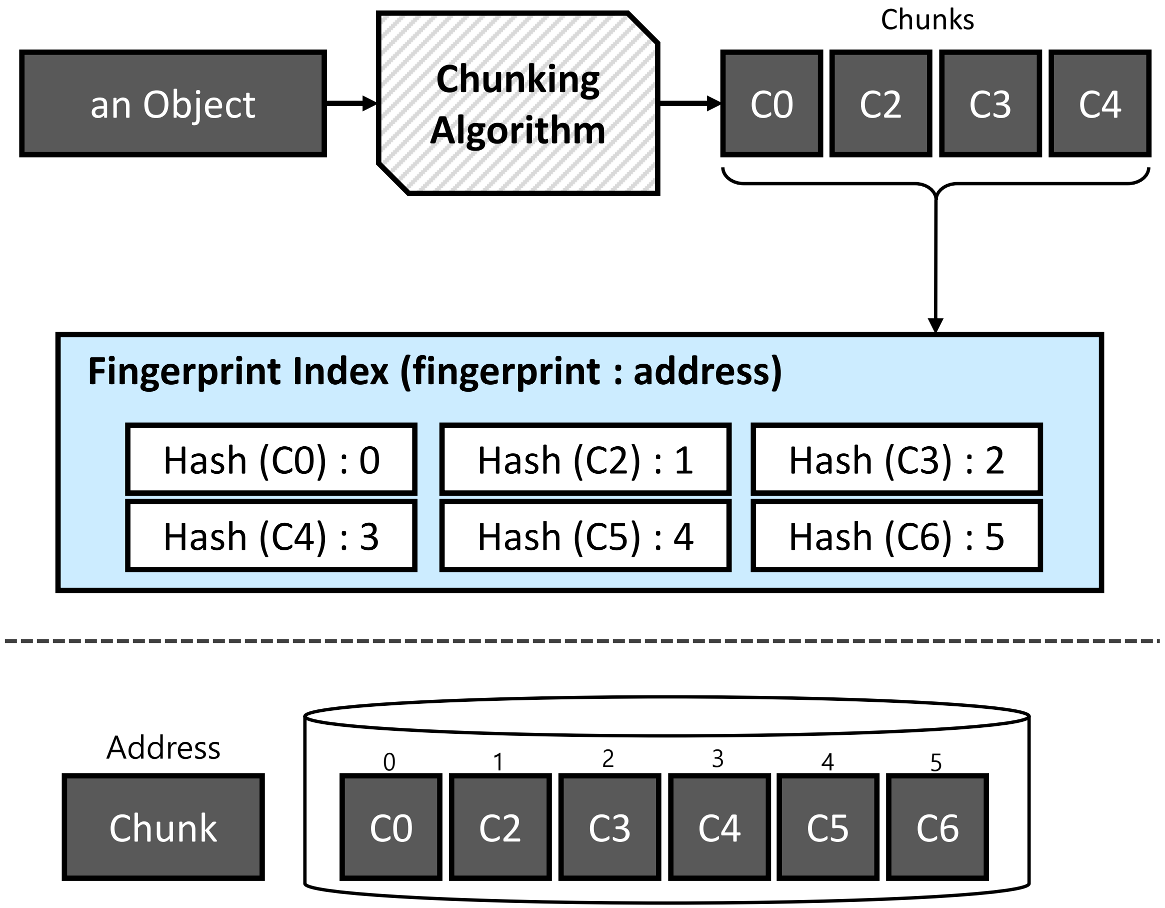
Storage Disaggregation

Summary



Deduplication

- **Save storage capacity by eliminating redundant data**
  - Chunking
    - Divide a data stream into smaller chunks
  - Fingerprinting
    - Generate a representative value using a hash algorithm
  - Comparing
    - If matched, chunk is considered as redundant



Double Hashing

Combine two mismatched input value

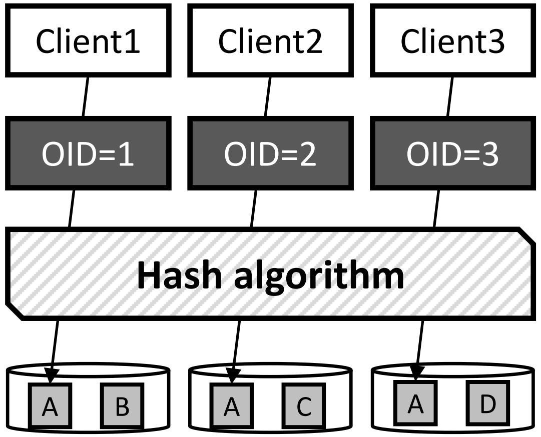
- Hash value of chunk for a deduplication system
- Object ID of chunk for a distributed storage system

Advantages

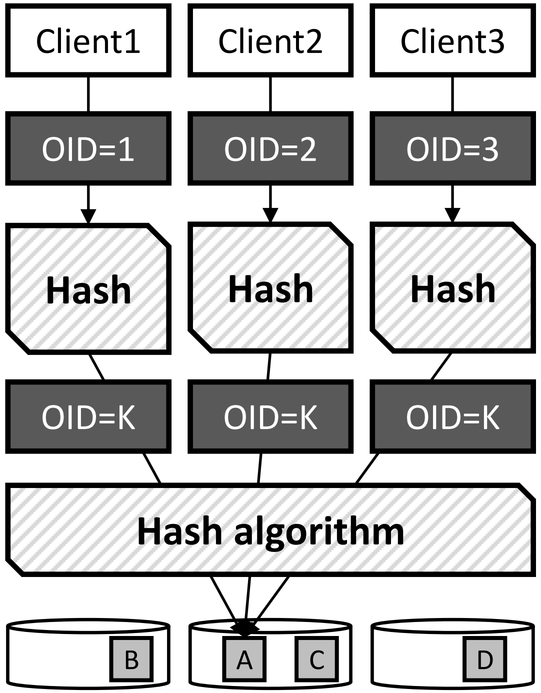
- Remove the fingerprint index
- Preserve the scalability of the underlying storage system
- No modification is required

Obj. ID	1	2	3	4	5	6
Content	A	A	A	B	C	D

Obj. ID – Content relation

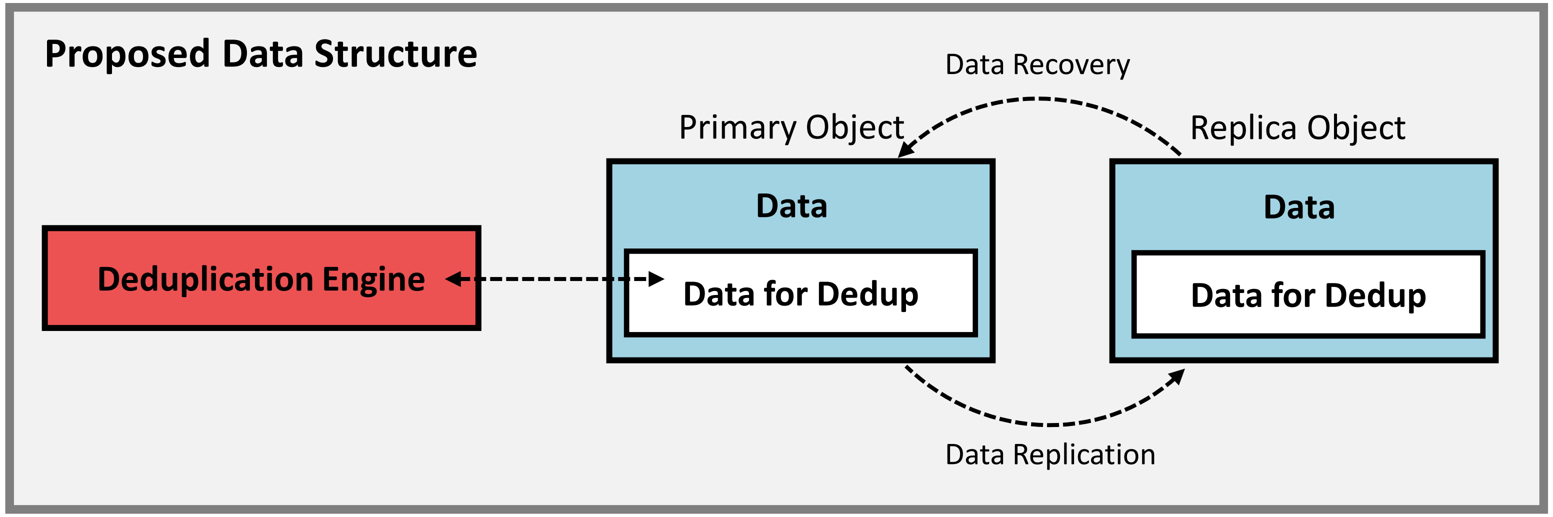


An ordinary OID-based distributed Storage



A content-hashed OID-based distributed Storage

Self-contained Metadata Structure



Source: Design of Global Deduplication for a Scale-Out Distributed Storage System, ICDCS 18

- **Design dedup system without any external component**
- **Extend the underlying storage's metadata to contain deduplication information**
- **Enable to exploit existing storage features while using dedup**

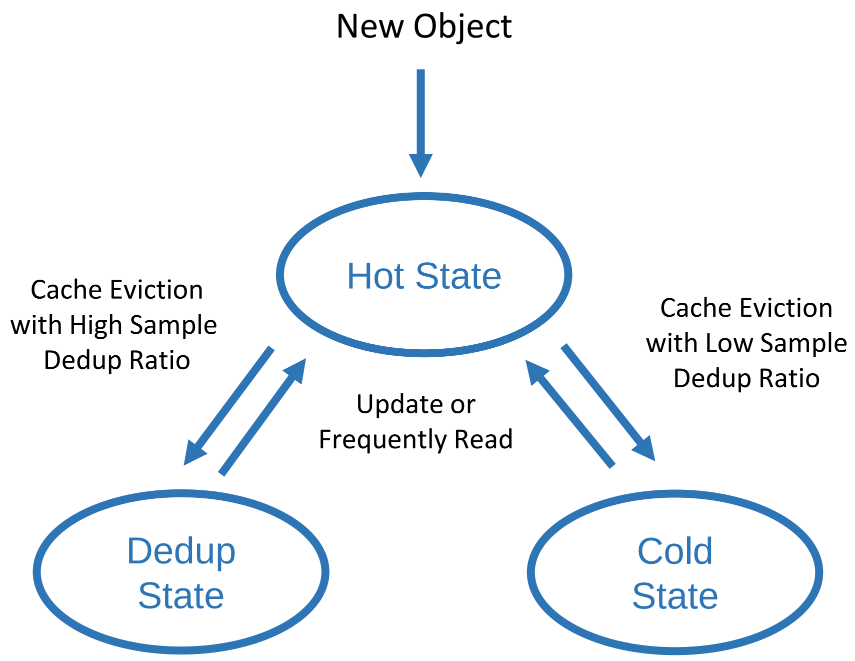
Global Deduplication with Tiering

Distributed system dedup challenges

- Additional data processing and I/O overhead
- Metadata management

How can we do better?

- Tiering based distributed storage design
  - *Hot / Cold / Dedup*
  - Deduplication-aware replacement policy
- Dedup ratio awareness


New Object
Hot State
Cache Eviction with High Sample Dedup Ratio
Cache Eviction with Low Sample Dedup Ratio
Update or Frequently Read
Dedup State
Cold State

Three states that represent the state of each object

reference: Design of Global Data Deduplication for a Scale-Out Distributed Storage System, ICDCS'18

Dedup Ratio Awareness

Dedup information *Crawling*

- Random sampling method
- Selective cluster-level crawling → Low overhead
- Shallow mode
  - Choose a small number of objects
  - Save CPU and memory overhead
  - Lower accuracy
- Deep mode
  - Higher accuracy
  - Consumes more time

Post-processing with rate control and selective deduplication

- Periodically conduct a deduplication job (background I/O) through rate control
- Maintain the object's hotness
  - Hot object is not deduplicated until its state is changed

Benefits

- Guarantee constant throughput
- Give a chance that frequently modified object does not need to be deduplicated

Contents

Agenda

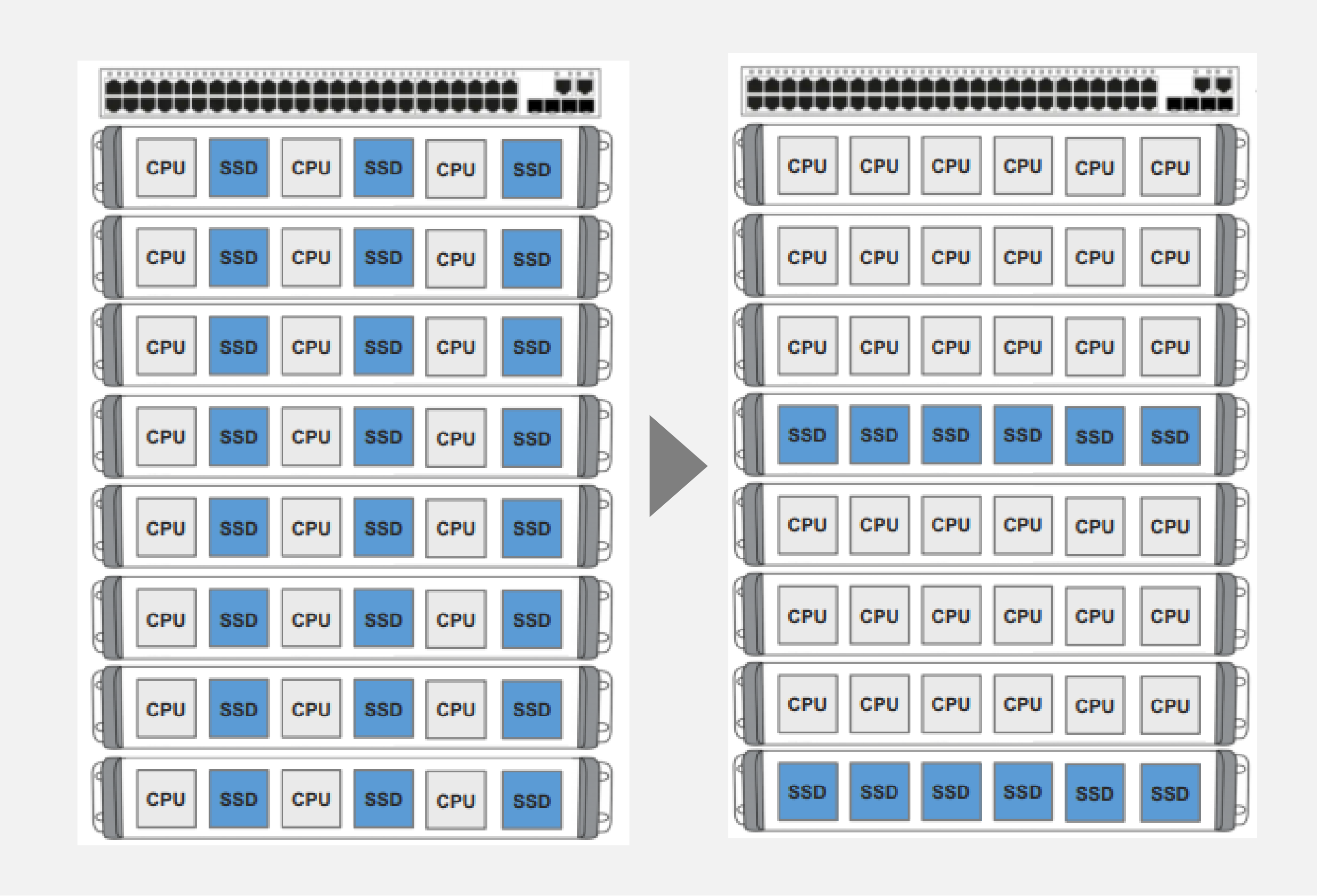
Recent Datacenter Trends

What Are We Focusing on?

Global Deduplication

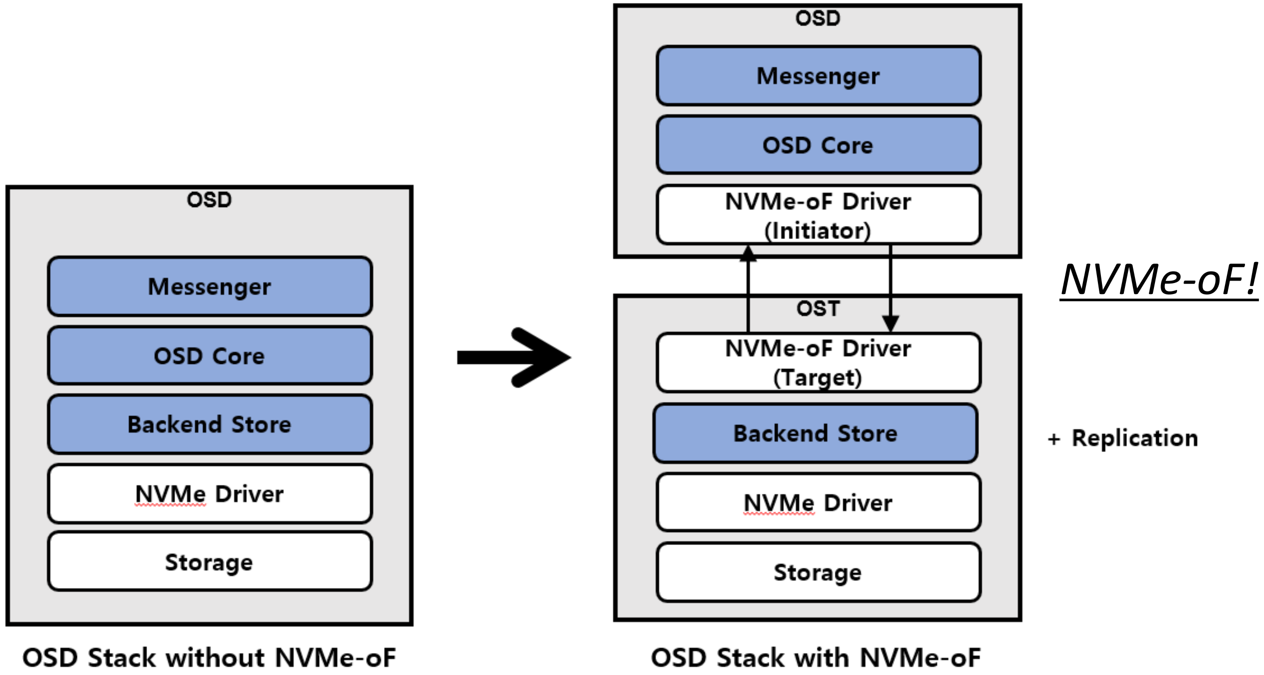
Storage Disaggregation

Summary



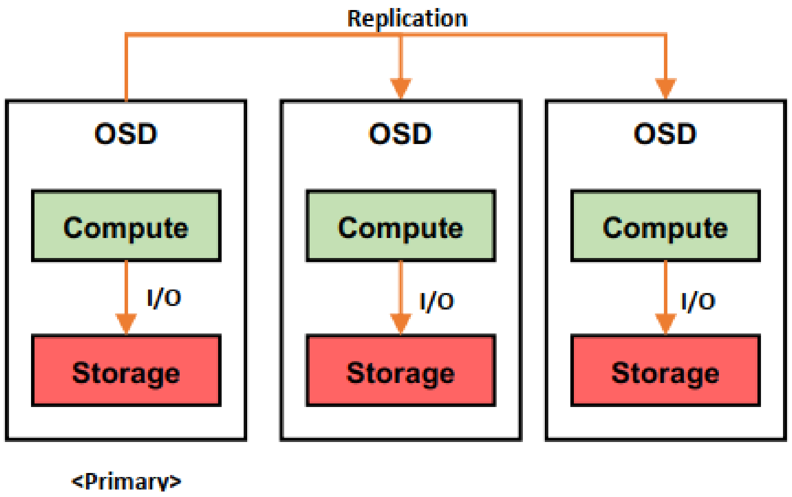
Storage Disaggregation in Ceph

- **Ceph itself does not support storage disaggregation**
  - Ceph does not aware of OST
  - OSD and OST is tightly coupled → Cannot share storage devices
  - Additional latency

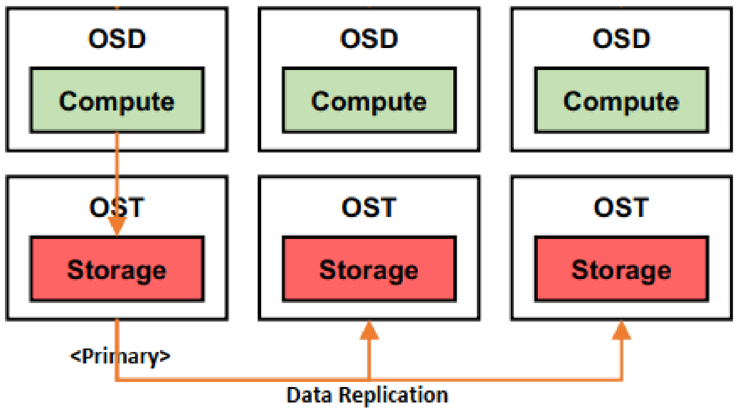


Storage Disaggregation Approach

- **Pursue low network traffic & OSD CPU consumption**



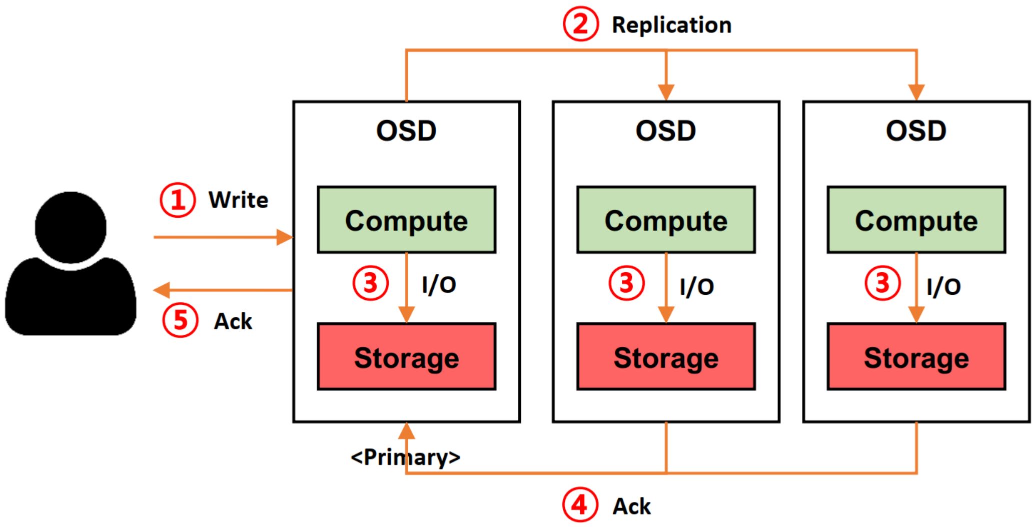
Legacy replication



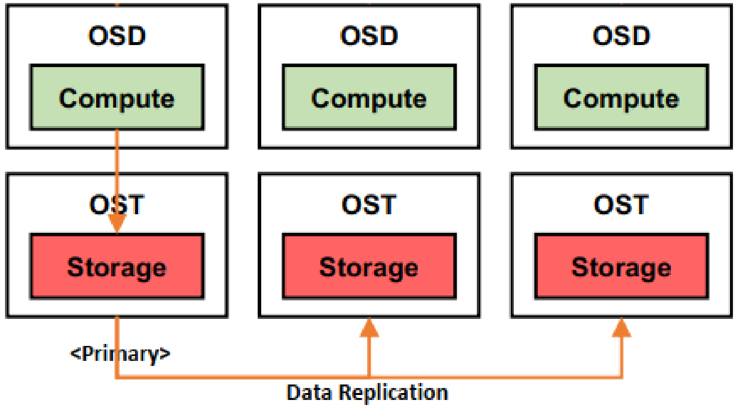
Replication with NVMe-oF

Storage Disaggregation Approach

- **Pursue low network traffic & OSD CPU consumption**



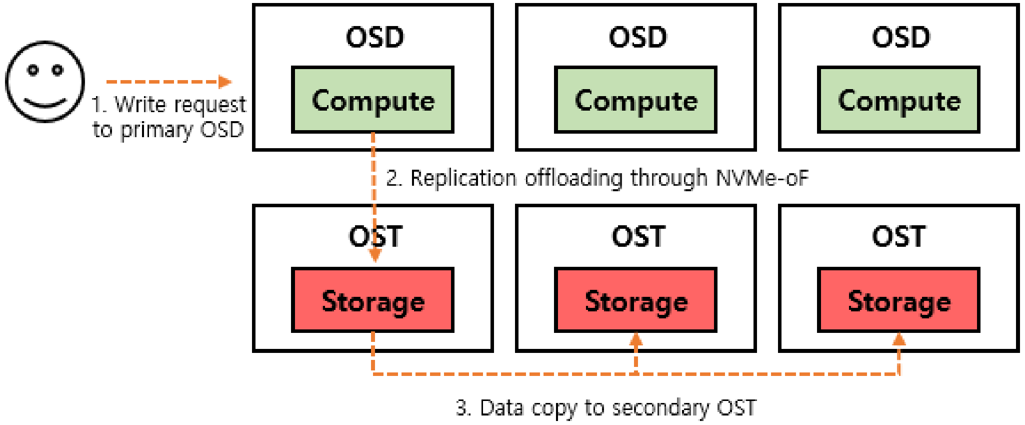
Legacy replication

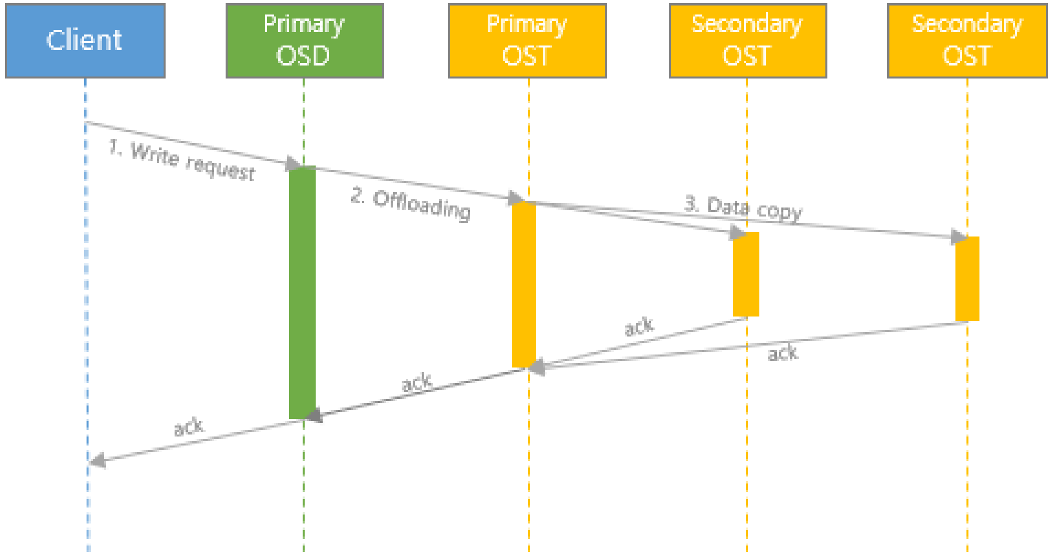


Replication with NVMe-oF

How Replication Offloading Works

- **The OSD hands the replication authority over to the OST**


1. Write request to primary OSD
OSD
Compute
OSD
Compute
OSD
Compute
2. Replication offloading through NVMe-oF
OST
Storage
OST
Storage
OST
Storage
3. Data copy to secondary OST



Benefits

- **Lessen CPU burden of OSD nodes**
- **Write speed improvement while taking the same level of consistency and reliability**
- **Fault tolerant & enables fast recovery**

Contents

Agenda

Recent Datacenter Trends

What Are We Focusing on?

Global Deduplication

Storage Disaggregation

Summary

Summary

- **Storage devices have been diversified by user needs**

- **Storage Disaggregation can be a solution to get over CPU limitation**

- **Deduplication for distributed storage system can manage storage more efficiently**

THANK YOU :D