Flash Memory Summit

SW-HW Co-Design for High Performance Storage System on ZNS SSD

Wei Tang

08/08/2023

Overview

- **Background**
- Bytedance’s Storage System
- Results

Data Center Infrastructure Trend

Compute

- CPU bottleneck
- In-house AI Chip
- HW accelerator

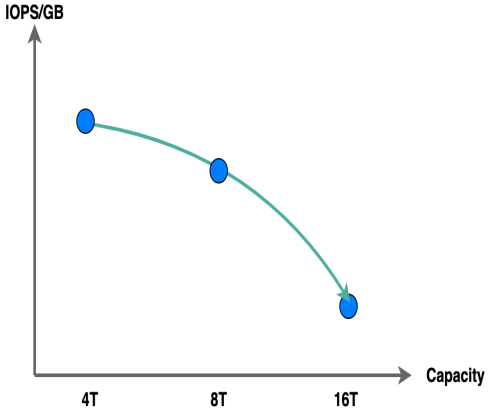
Network

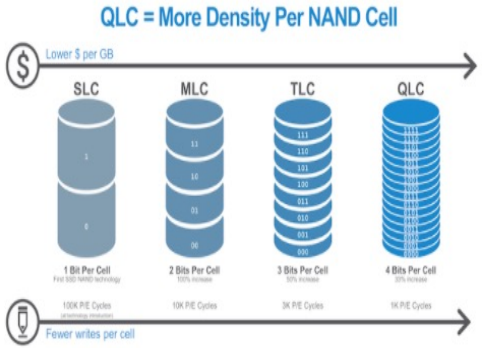
- 10G->100G->400G
- High Perf. Network Protocol
- Hardware Offload

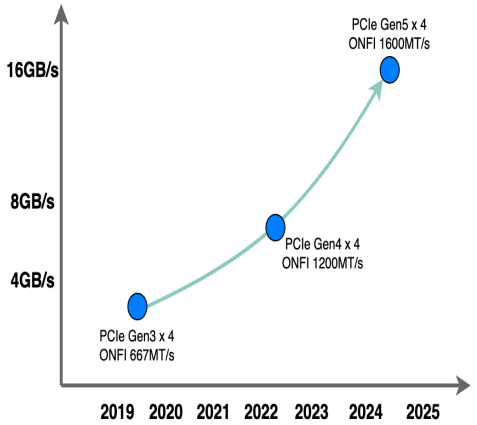
Storage

- OC, KVSSD, ZNS
- Smart SSD
- Computational SSD

Enterprise SSD Status & Trend







Drive Capacity Up-> IOPS/GB Down

NAND Endurance Down

RW BW increases with PCIe

Cloud Storage Pain Points

Performance

- Log-on-log problem: FS, FTL
- Heavy IO software stack
- SSD Garbage Collection

Cost

- Redundant Overprovision: FS, SSD OP
- Performance mismatch

Stability

- End-to-End performance analysis & debug
- Long problem solve cycle with SSD vendors

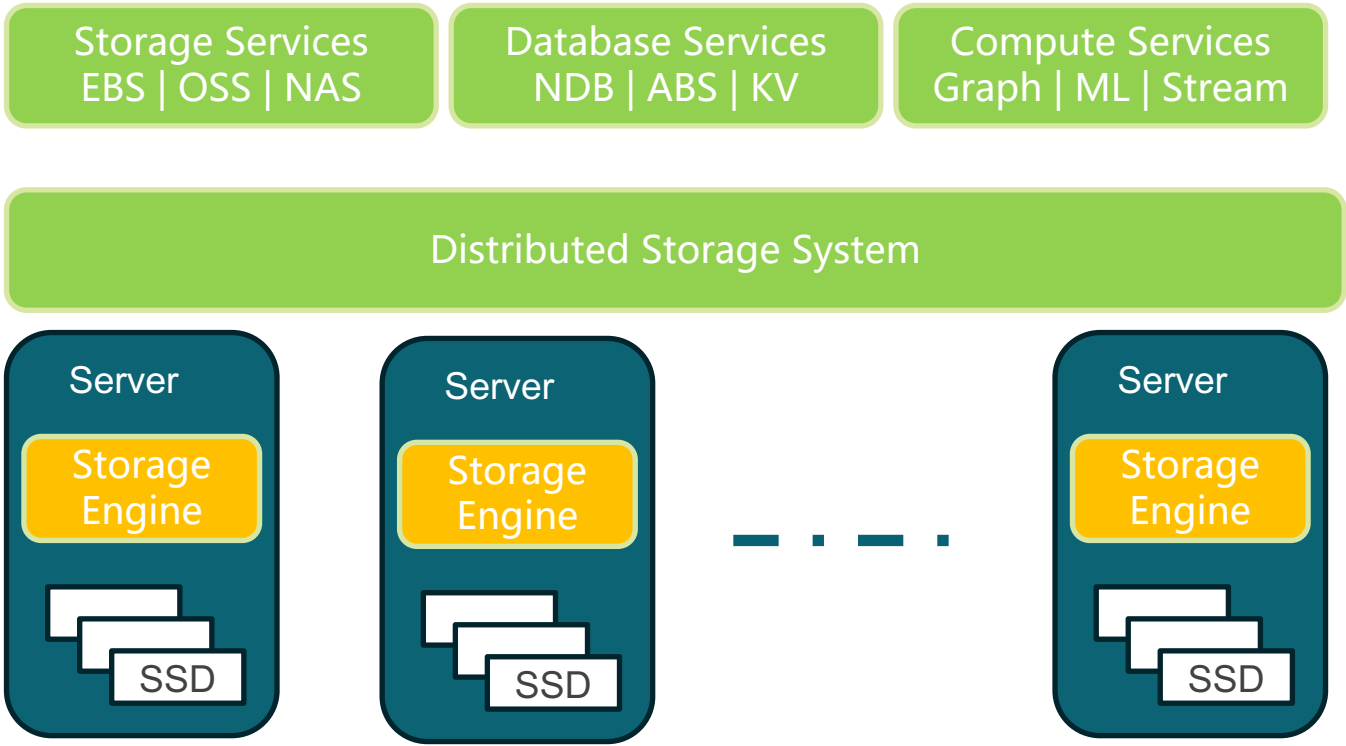
Customization

- SGL, CMB, PMR
- ZNS small zone support
- Workload specific optimization

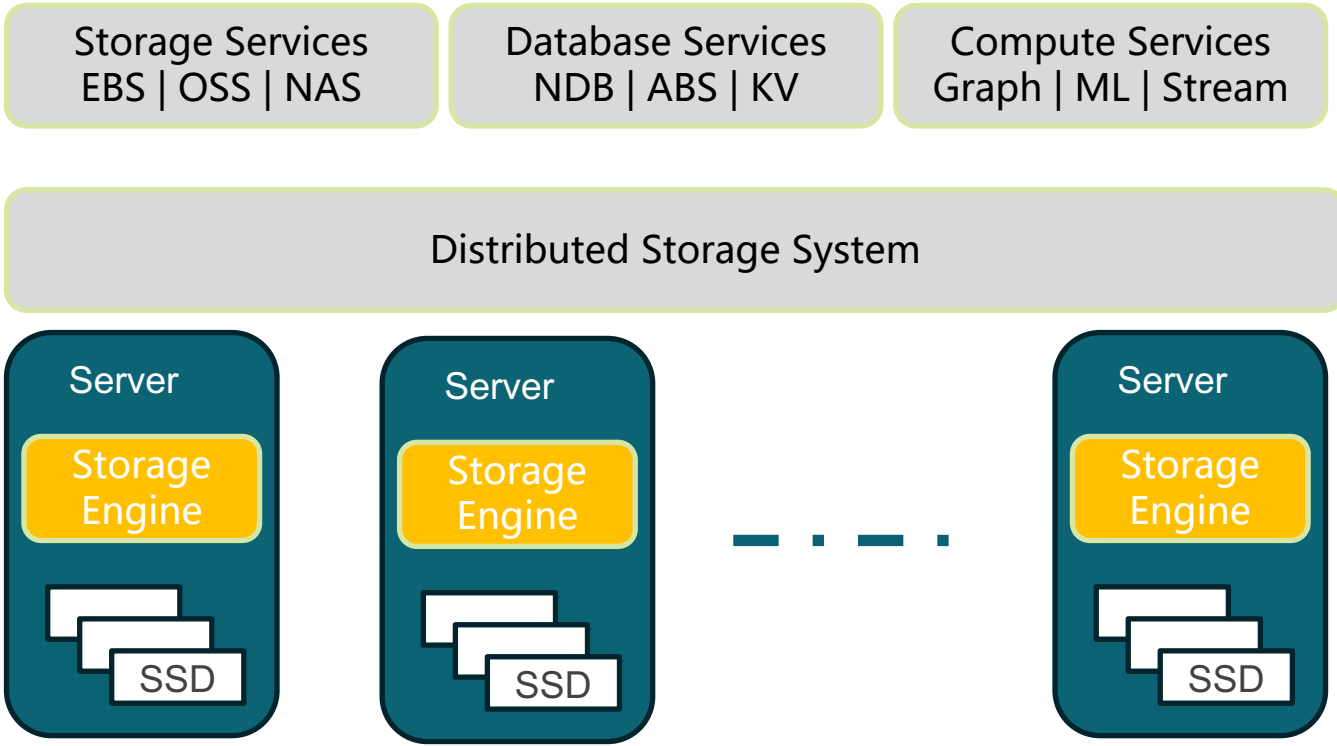
Overview

- Background
- **Bytedance's Storage System**
- Results

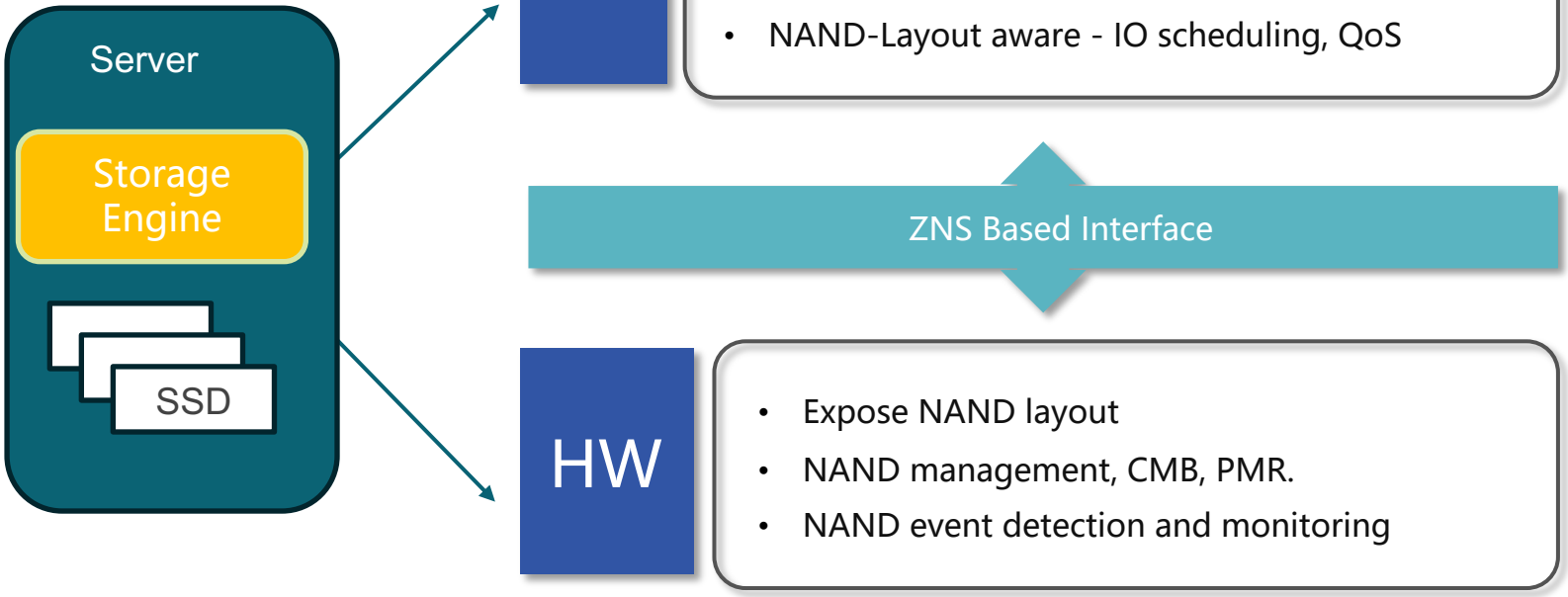
Storage System



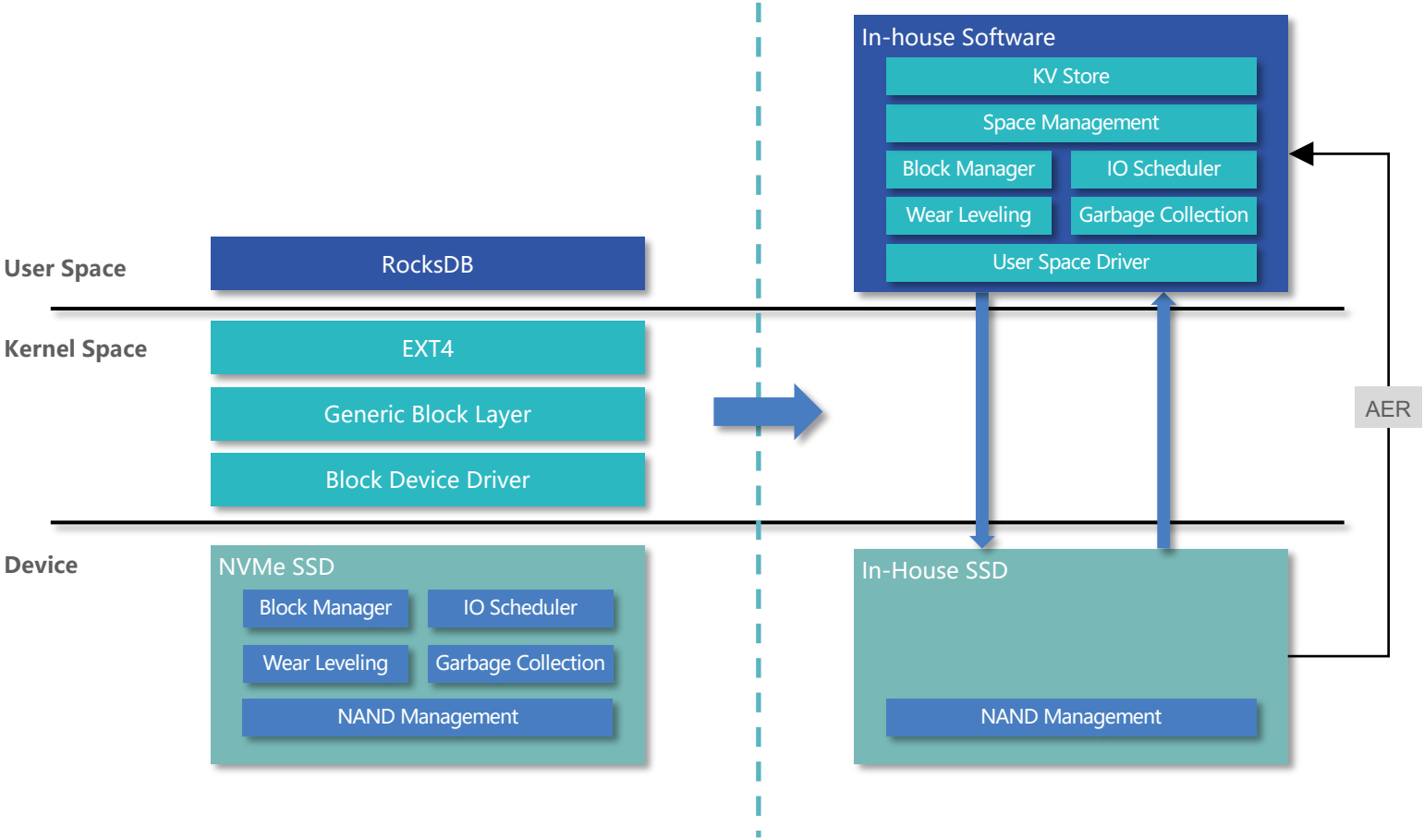
Storage System



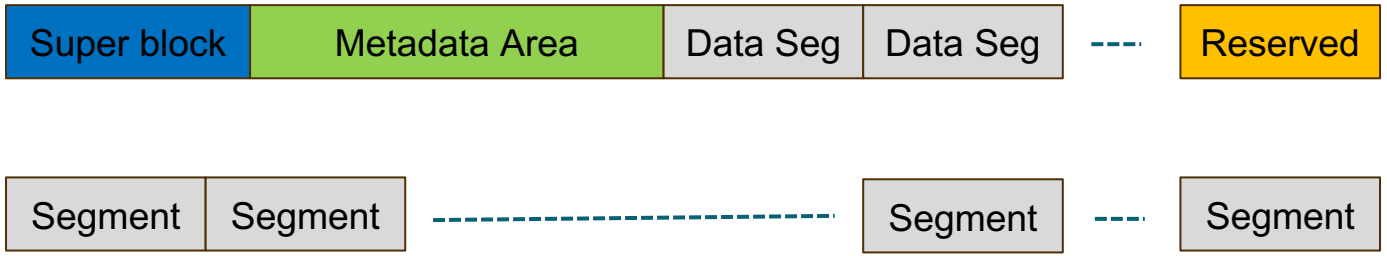
HW-SW Co-Design Storage Engine



Software Architecture



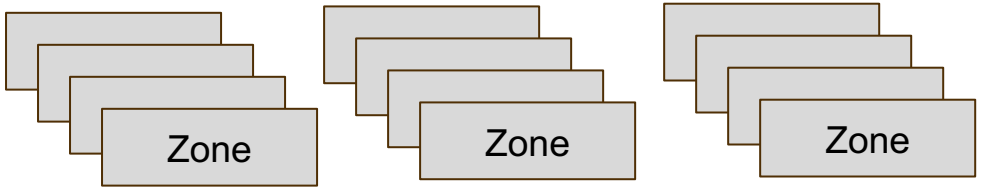
Space Management



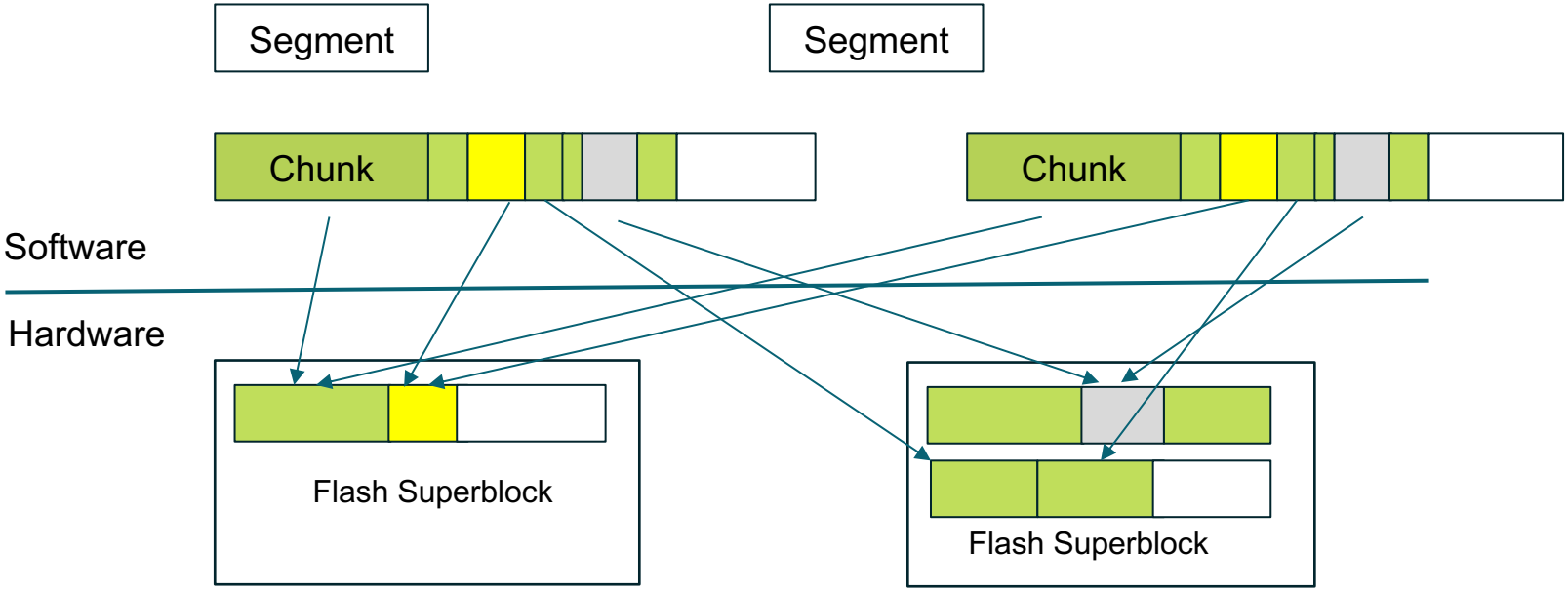
Software

Hardware

SSD

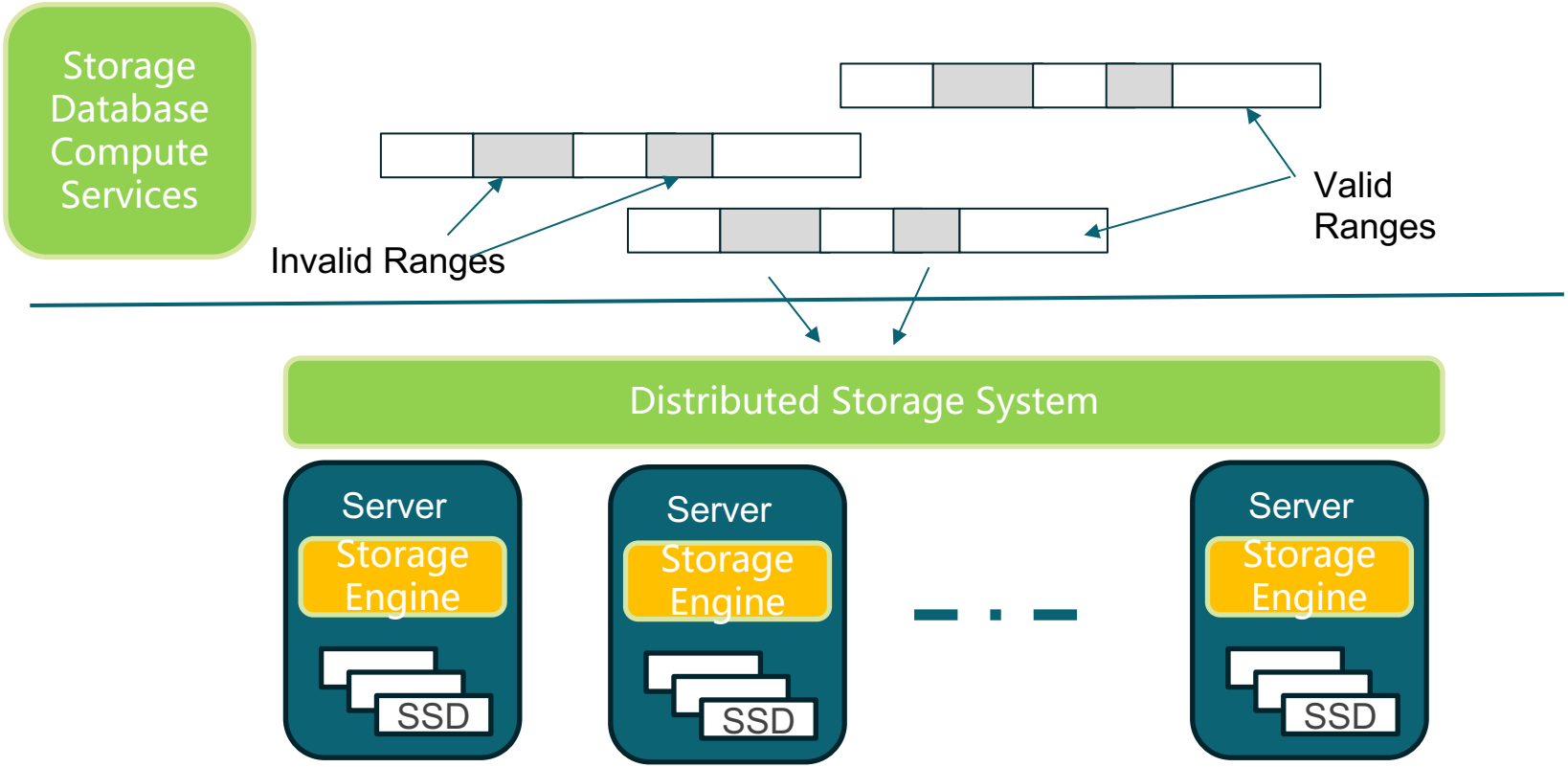


Data Model

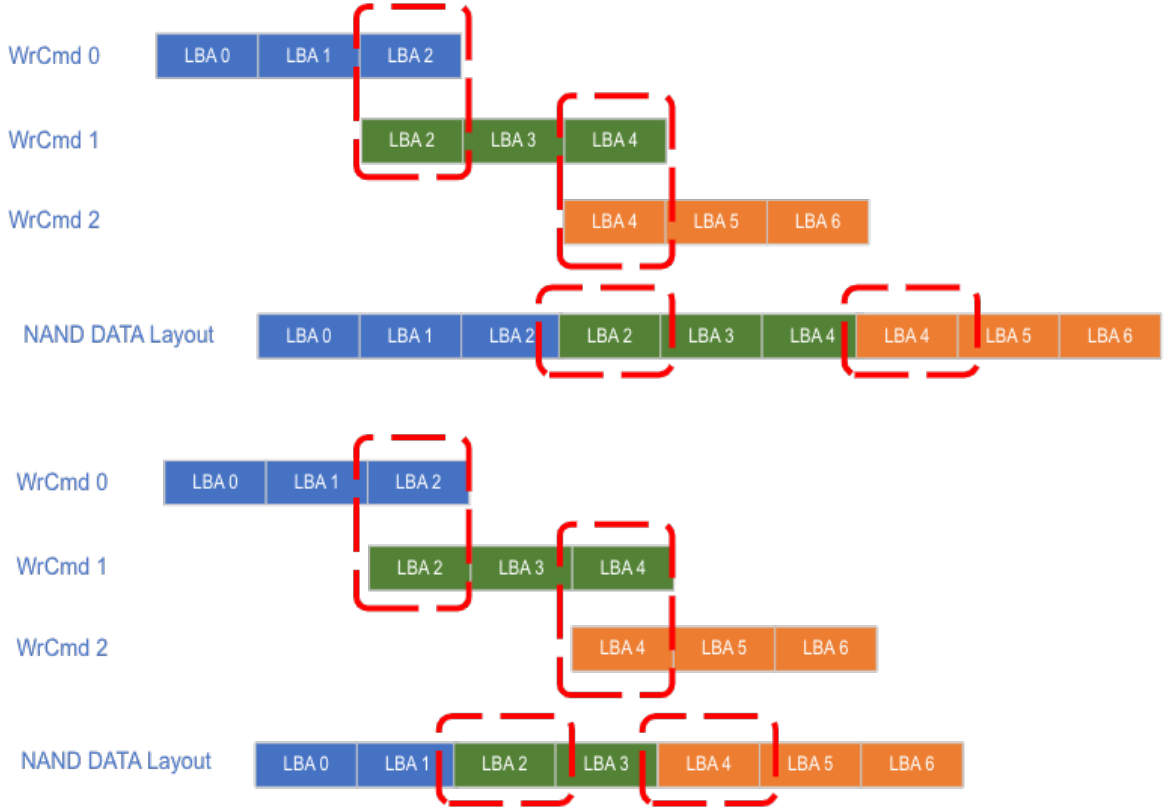


Chunk Logical address has three statuses:
1.Valid - contains valid data and occupies physical space
2.Invalid - contains invalid data and occupies physical space
3.Hole - does not occupy physical space

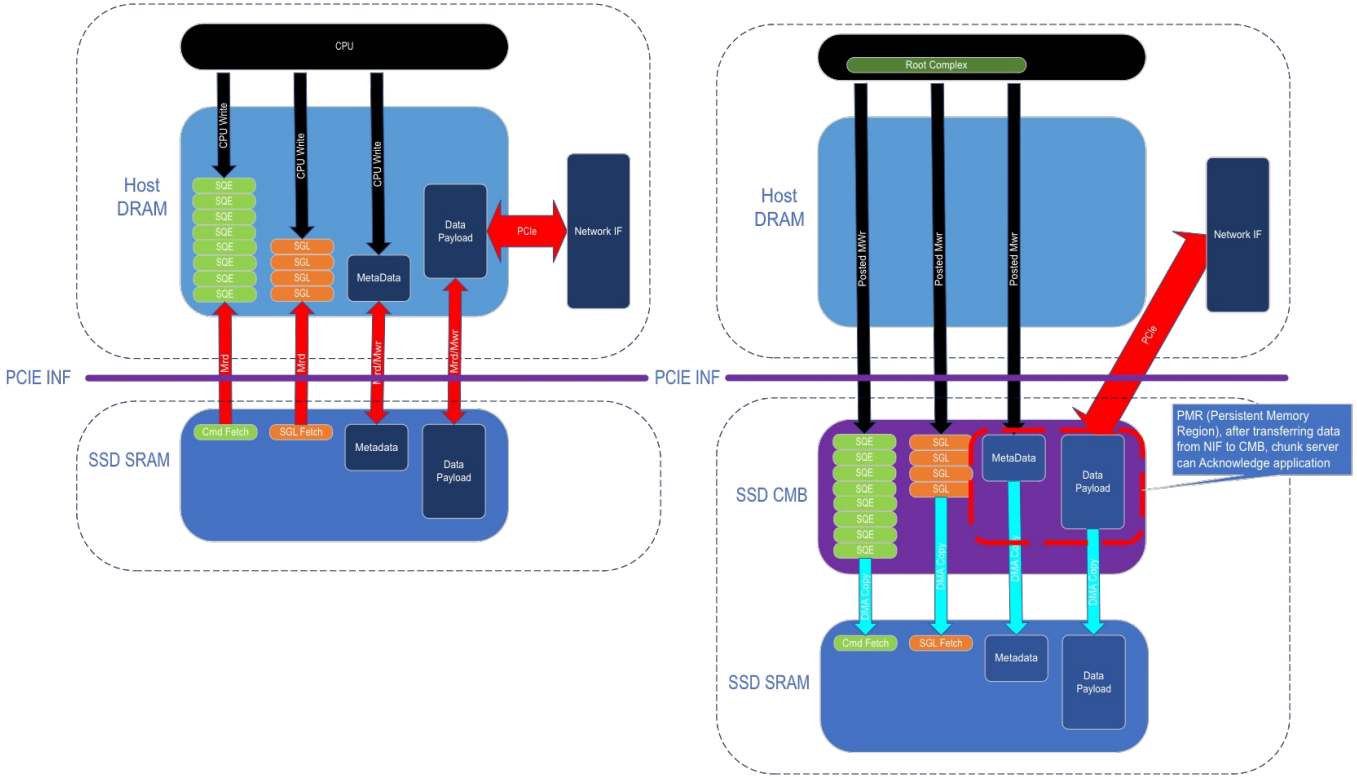
Collaboratory Garbage Collection



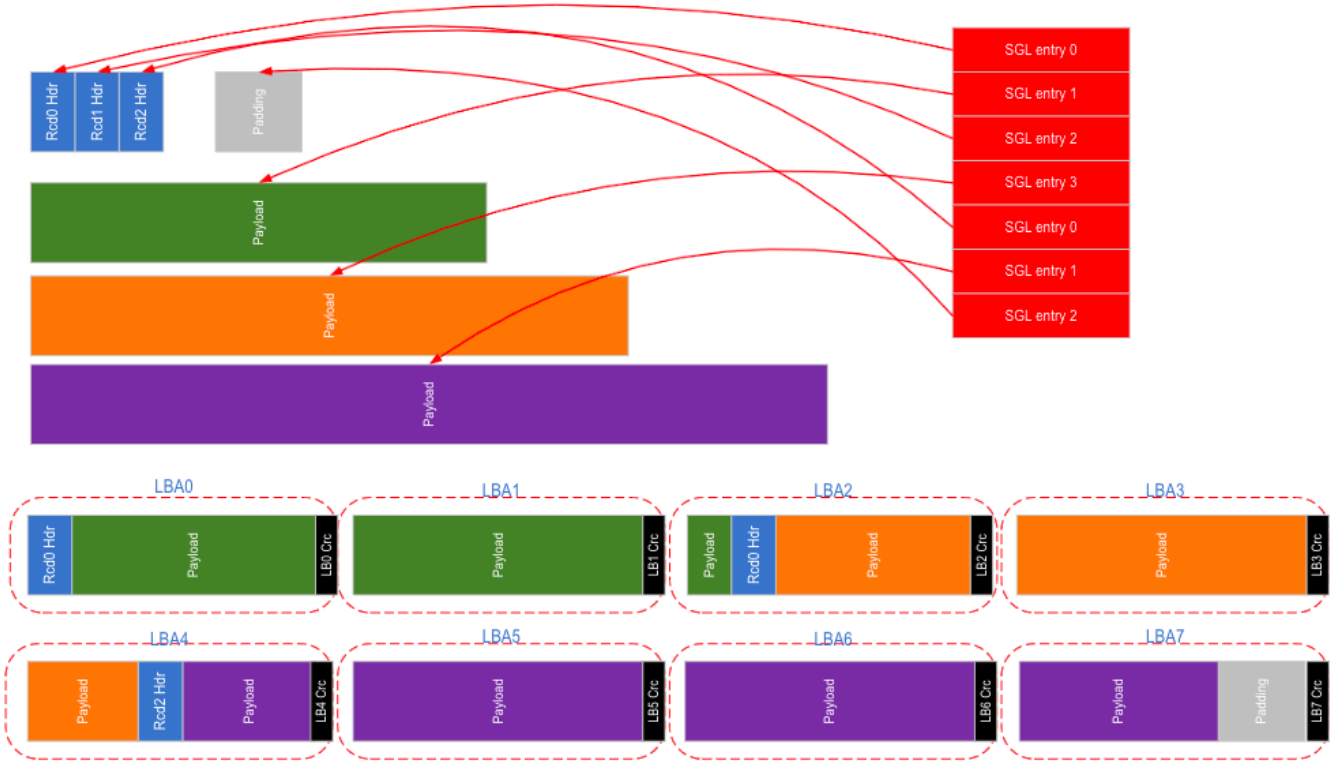
LBA Overwrite



Payload Passthrough - CMB/PMR



Zero Copy In IO Path - SGL



Cloud Storage Differential Requirements

LBA Overlap
Last LBA overwrite, reducing WA

SGL
Zero copy through SGL, improving perf. and QoS.

High QD Write
Compare QD=1 Zone Write, higher throughput and IOPs

RAID Ratio
Flexible Raid Configuration to save host space

CMB
Hosts to use CMB, reducing PCIe and DRAM BW

Layout
Host-aware layout for better QoS

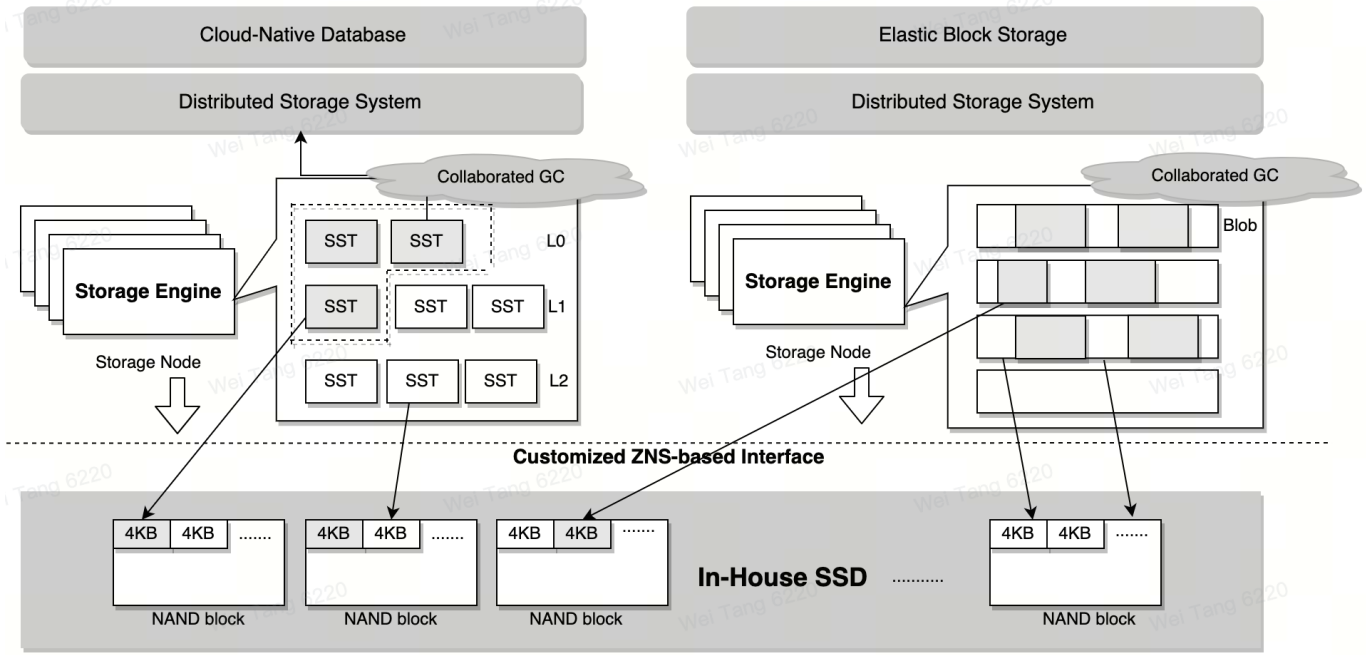
PMR
Persistent in-device storage

Logging
Host-aware logging and monitoring for system reliability

Copy Command
Batch Copy in GC, improving performance.

Simulator
Software stack verification and new functionality

Potential Return Of Interests



Usable space Increase

30%+

Bandwidth Increase

3X+

Storage Cost

20%+

THANKS