Alibaba Cloud
Worldwide Cloud Services Partner

SMRStore

A Storage Engine for Cloud Object Storage on HM-SMR Drives

Su Zhou, Erci Xu, Hao Wu, Yu Du, Jiacheng Cui, Wanyu Fu, Chang Liu, Yingni Wang, Wenbo Wang, Shouqu Sun, Xianfei Wang, Bo Feng, Biyun Zhu, Xin Tong, Weikang Kong, Linyan Liu, Zhongjie Wu, Jinbo Wu, Qingchao Luo, Jiesheng Wu

Alibaba Group

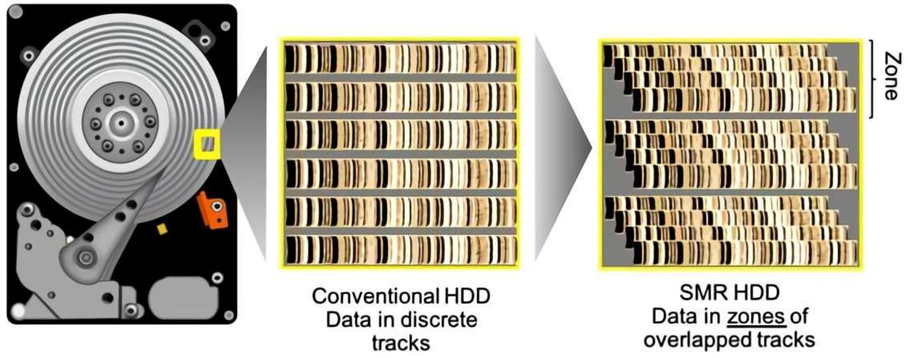
Reducing Cost is Important for OSS

Alibaba Cloud Object Storage Service (OSS)

OSS is an **exabyte-level** storage service based on **CMR drives**.

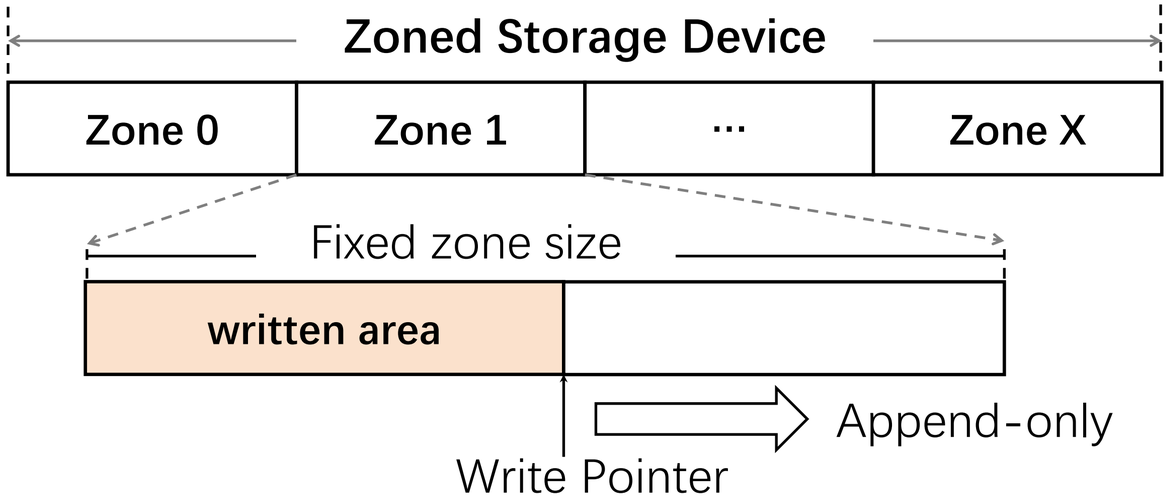
HM-SMR Drives (Host-managed)

- **~25%** higher areal density
- Better **cost-efficiency**



Backward-incompatibility

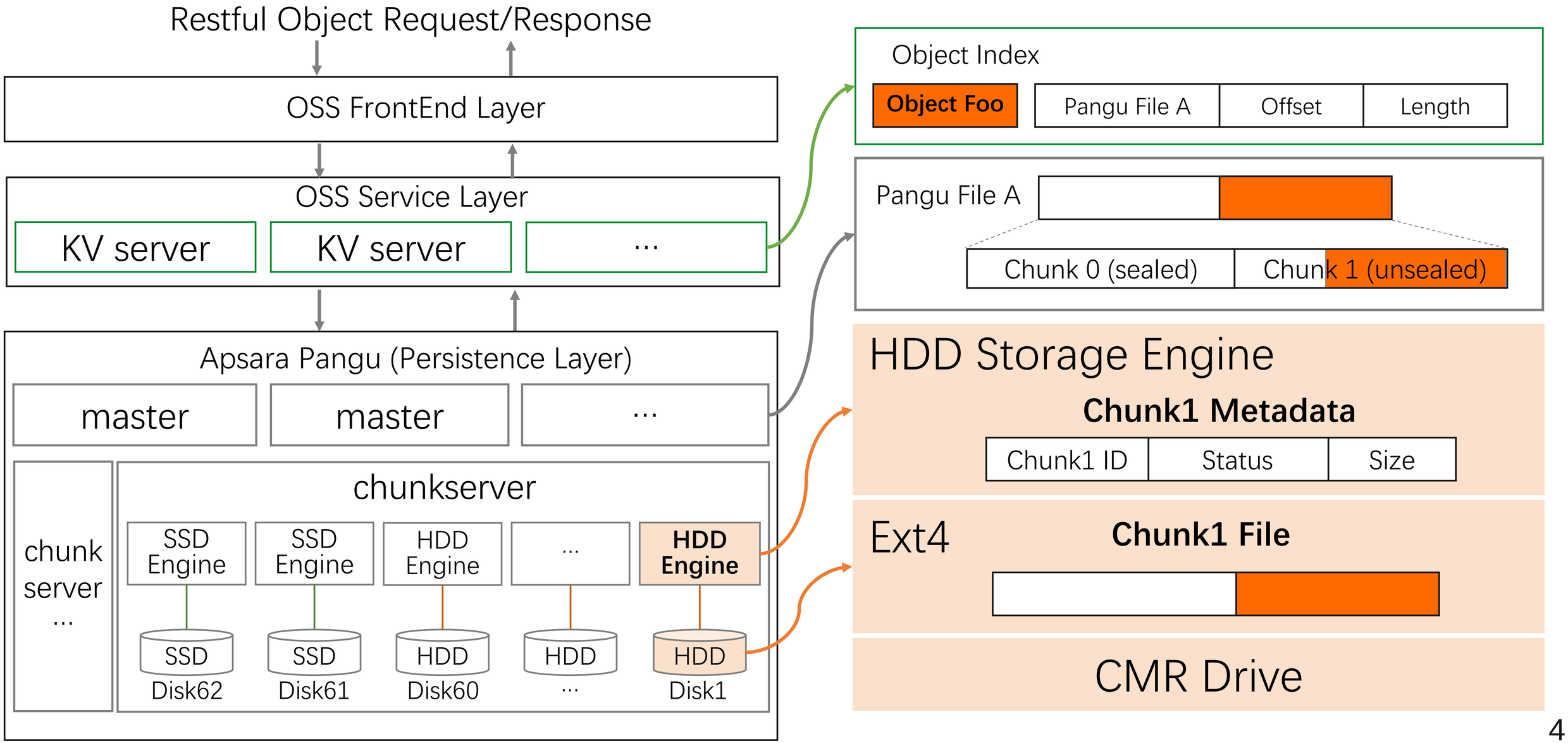
- Zone model (zone size 256MB)
- Sequential write constraint
- Open zone limit (128)



Goal

To improve cost-efficiency of OSS by replacing CMR drives with HM-SMR drives without compromising on performance.

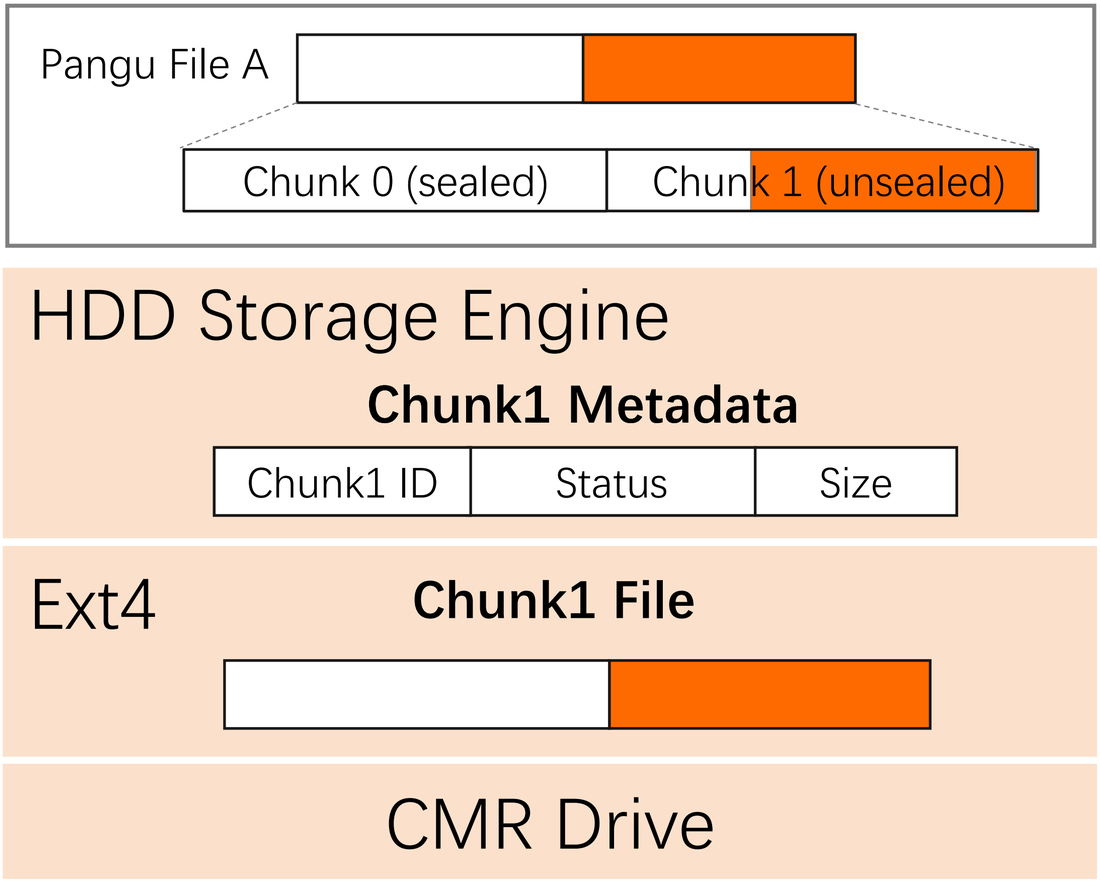
Alibaba Cloud Object Storage Service (OSS)



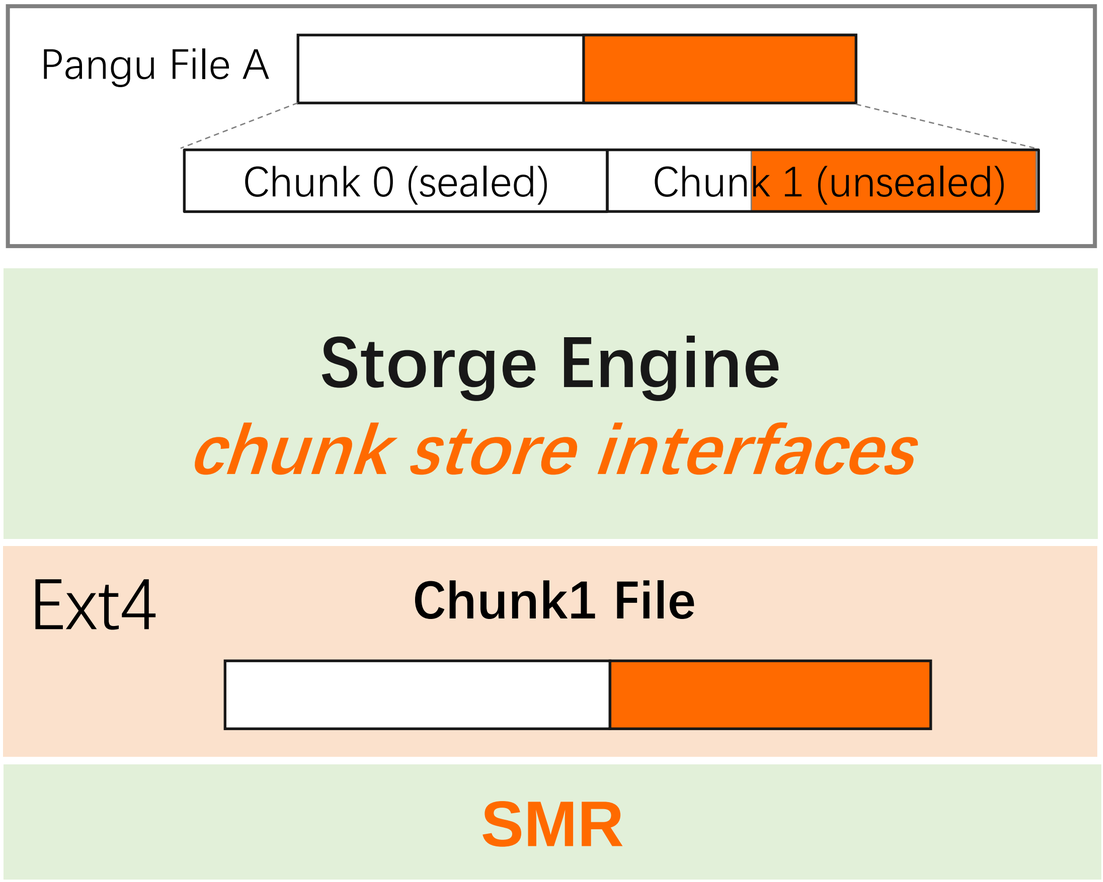
Pangu File & Chunk

- ✓ Pangu file is **append-only**.
- ✓ Chunk is **append-only**.
- ✓ **Configurable** limit of max chunk size.

- ✗ Chunk size can be **variable** when:
  - the Pangu file is smaller than the limit.
  - I/O failures occur (switch to a new chunk)

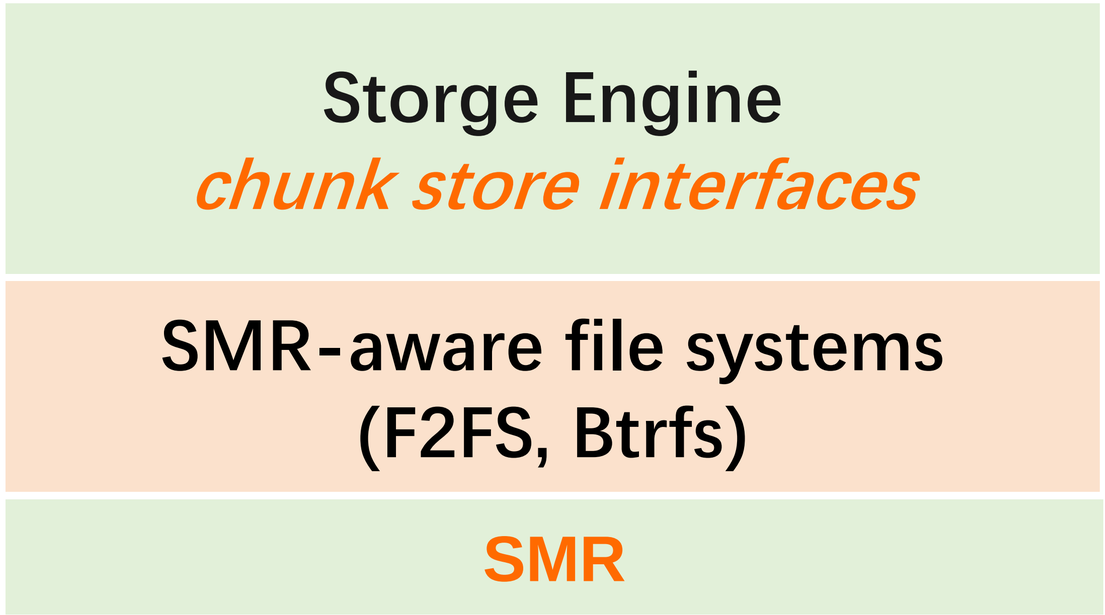


The key to adopt SMR drives



The key is to implement chunk store interfaces on SMR drives.

One choice is an SMR-aware file system.



Background

Existing Solutions

Design

Evaluation

Conclusion

How about F2FS?

F2FS began to support SMR drives with kernel 4.10

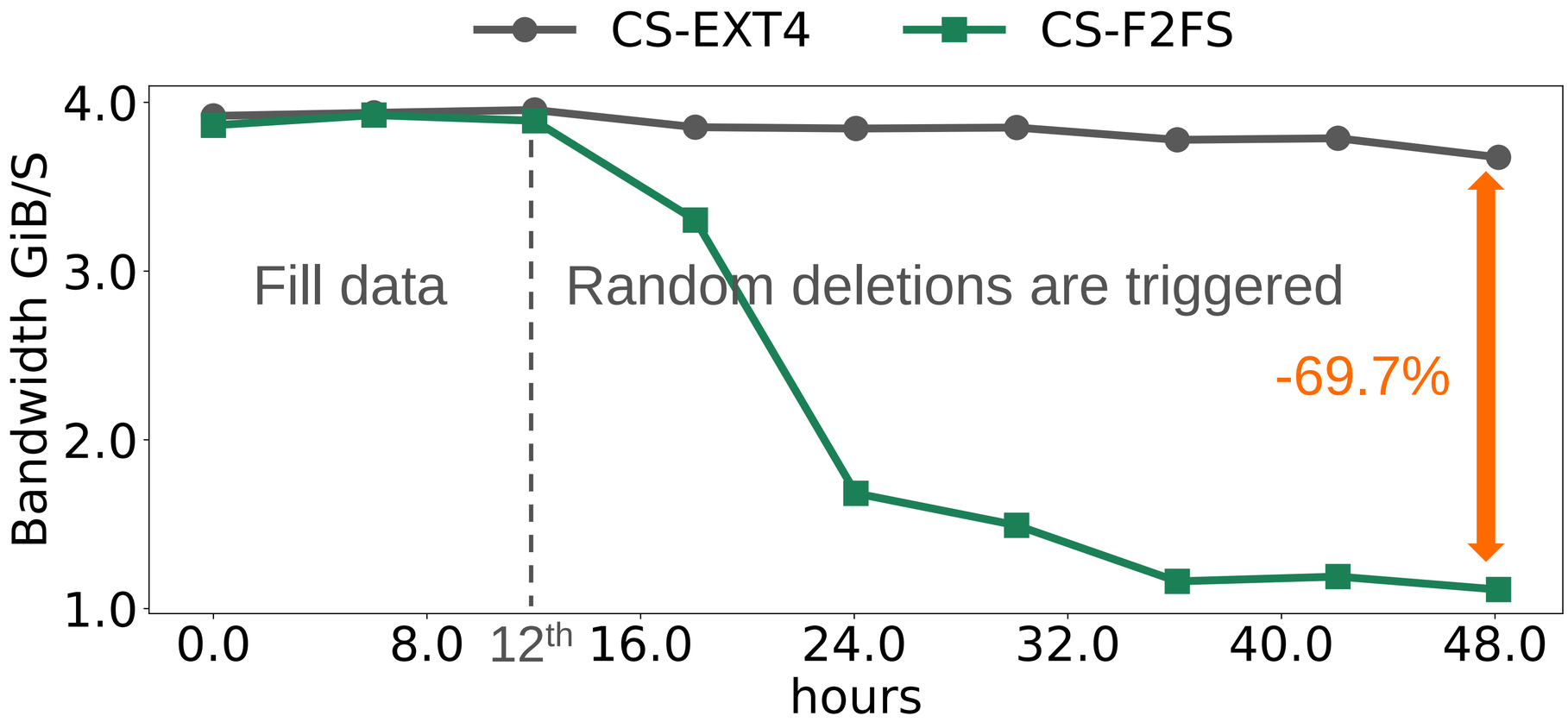
Workload Generator

- **Fio * 4** using Pangu APIs

One chunkserver

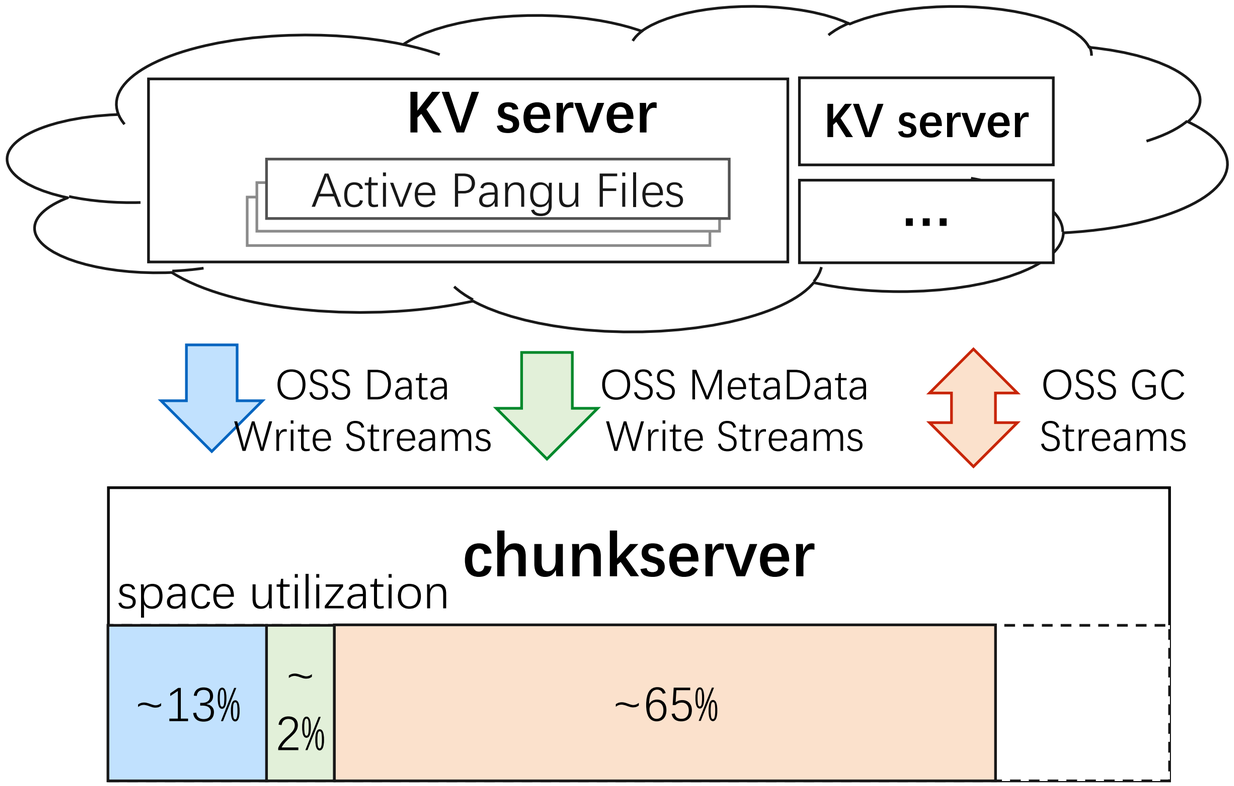
- **60** HDDs, **2** SSDs

80% Space utilization



Why does F2FS suffer a 70% throughput drop?

Observation 1: OSS Workloads



OSS Data Write Streams
- Lifespan of chunks is **Short** (<7Days)

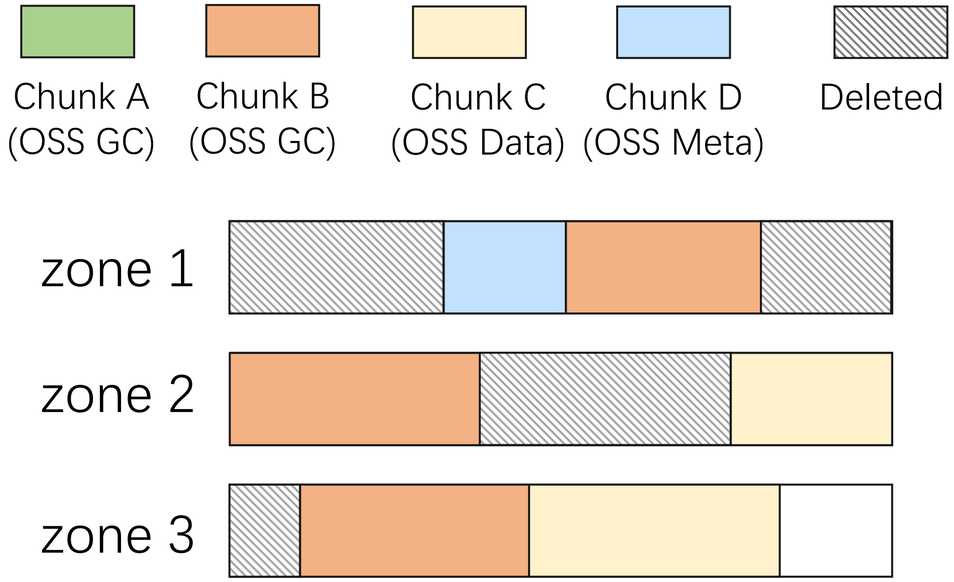
OSS MetaData Write Streams
- Chunks are usually **Small** (< 16MB)

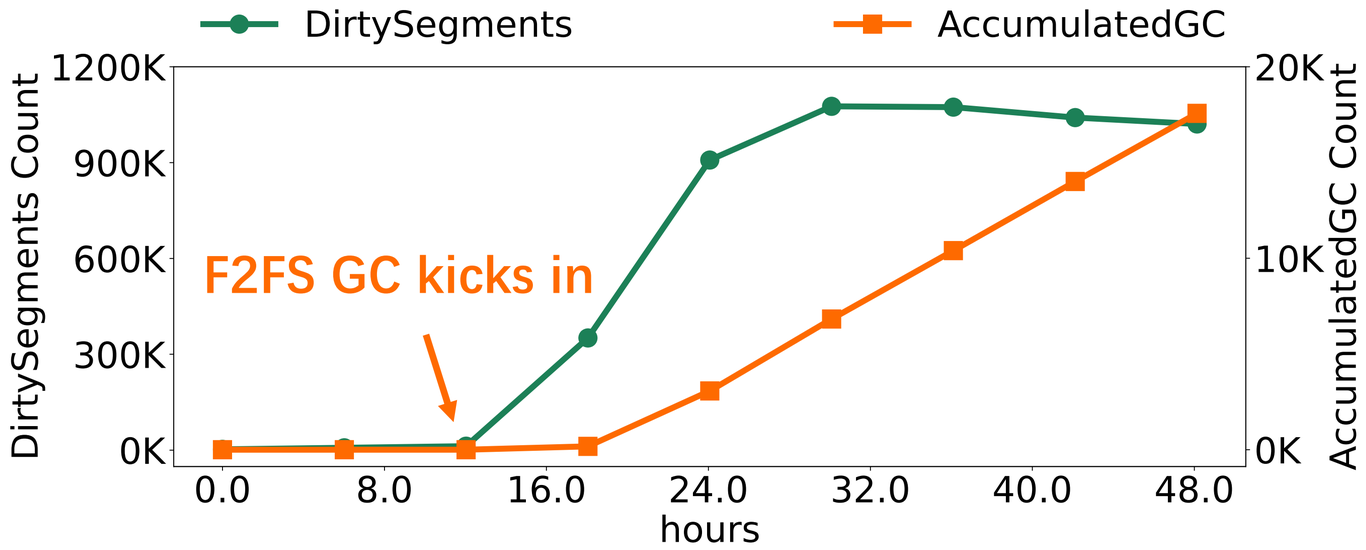
OSS GC Streams
- **Most** chunks are large (>90%).
- Stream concurrency is **Low** (~100 per chunkserver).

All Streams
- **Random** Deletions

OSS has quite different workloads
hot vs. cold, small chunks vs. large chunks…

Observation 2: F2FS Placement



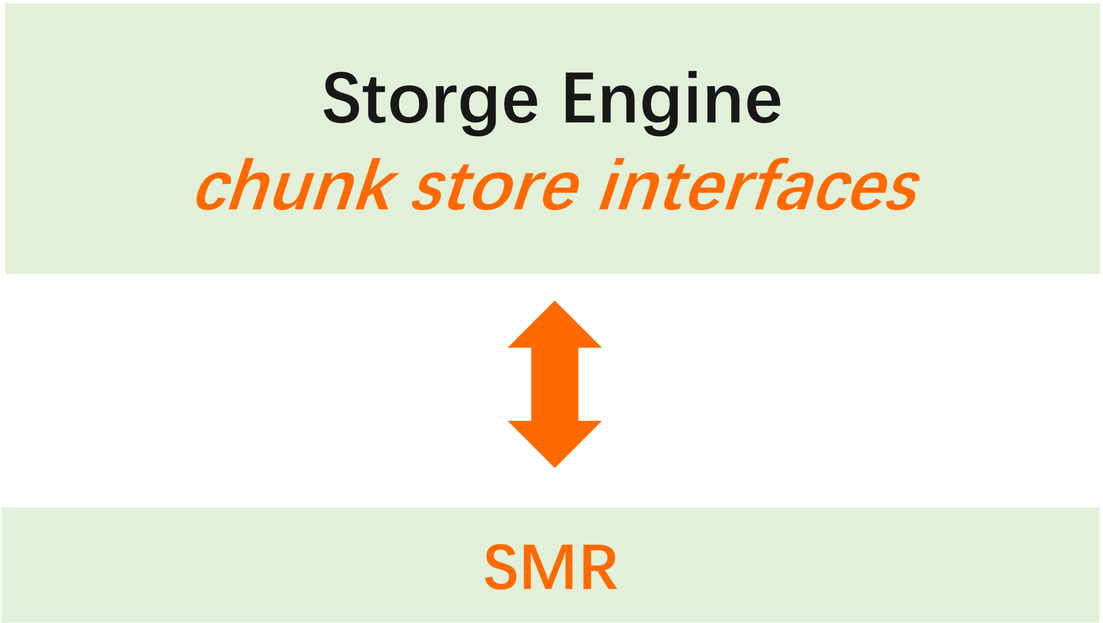


Three zones should be reclaimed when chunk A is deleted.

Mixing chunks results in heavy F2FS GC under OSS workloads.

Our Choice

Build a new user-space storage engine on HM-SMR drives co-designed with OSS.



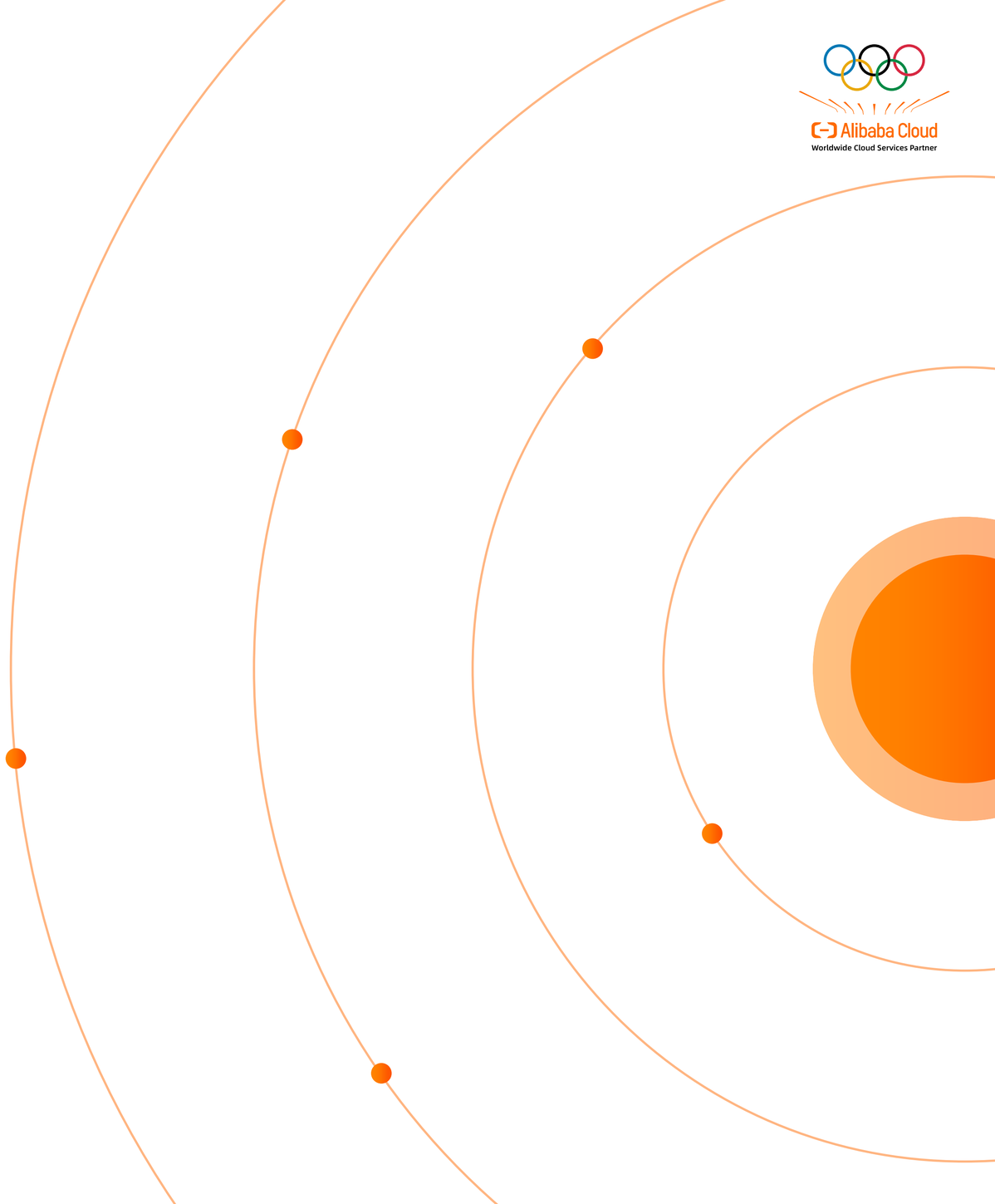
Background

Existing Solutions

Design

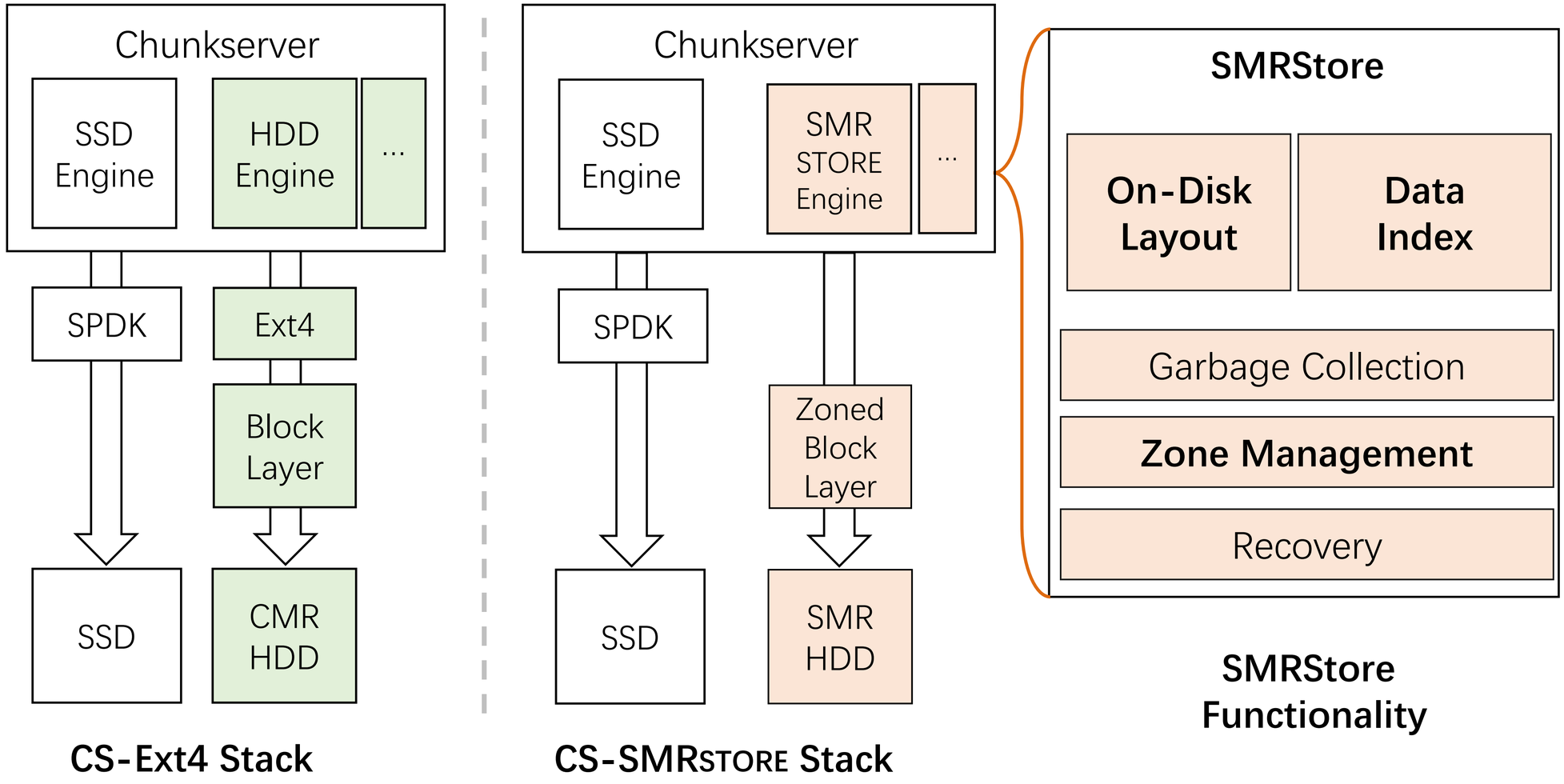
Evaluation

Conclusion



Traditional Engine vs. SMRStore

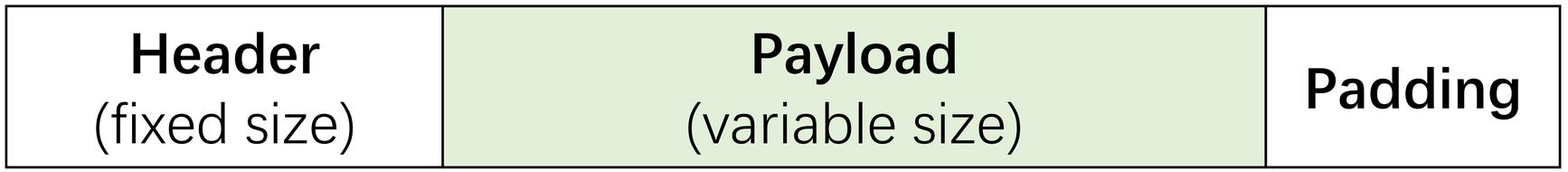
architecture comparison



On-Disk Data Layout

log structured design

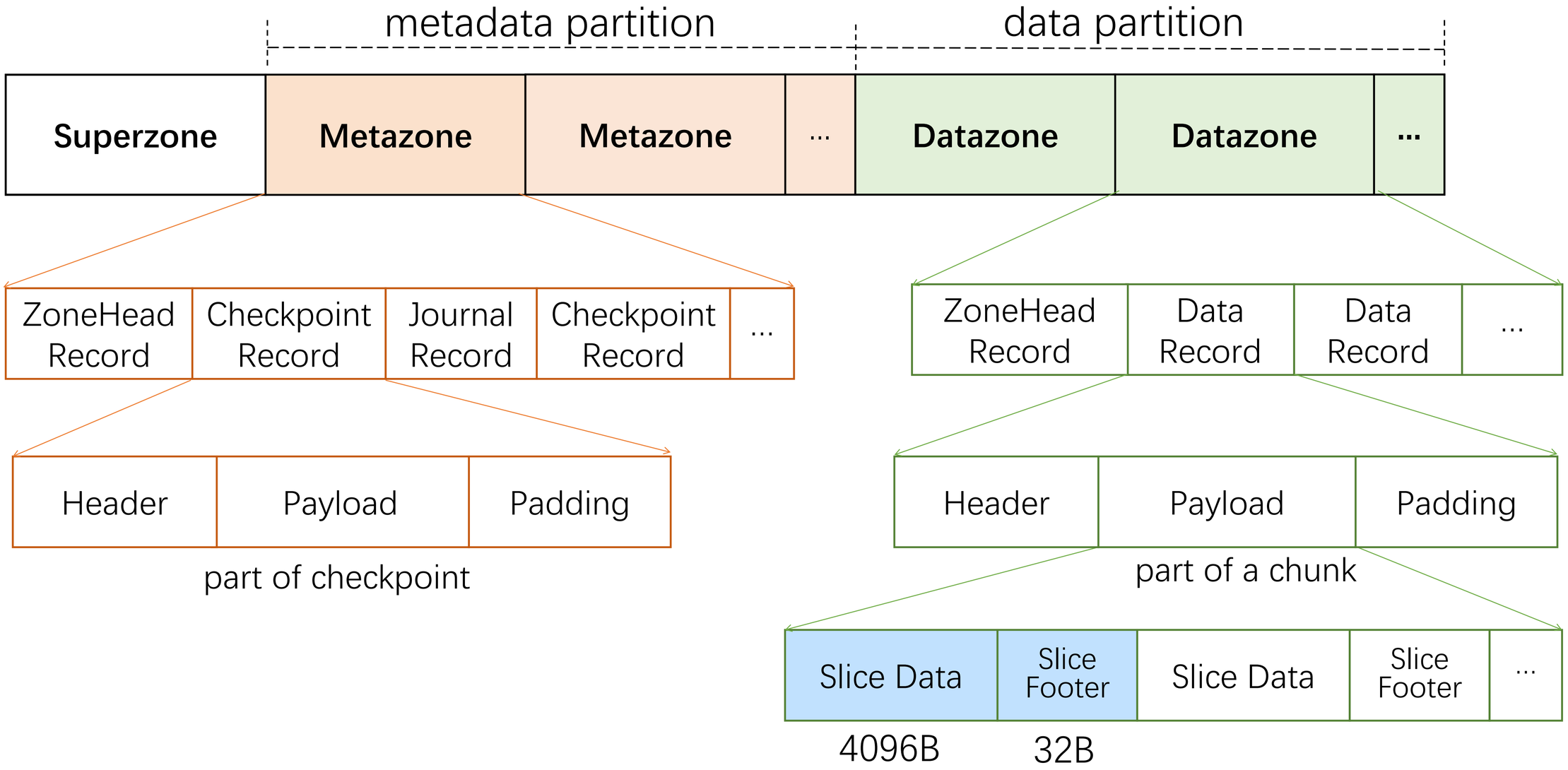
Everything is a **Record.**



record type
record size
info of payload

for 4KB-alignment

On-Disk Data Layout

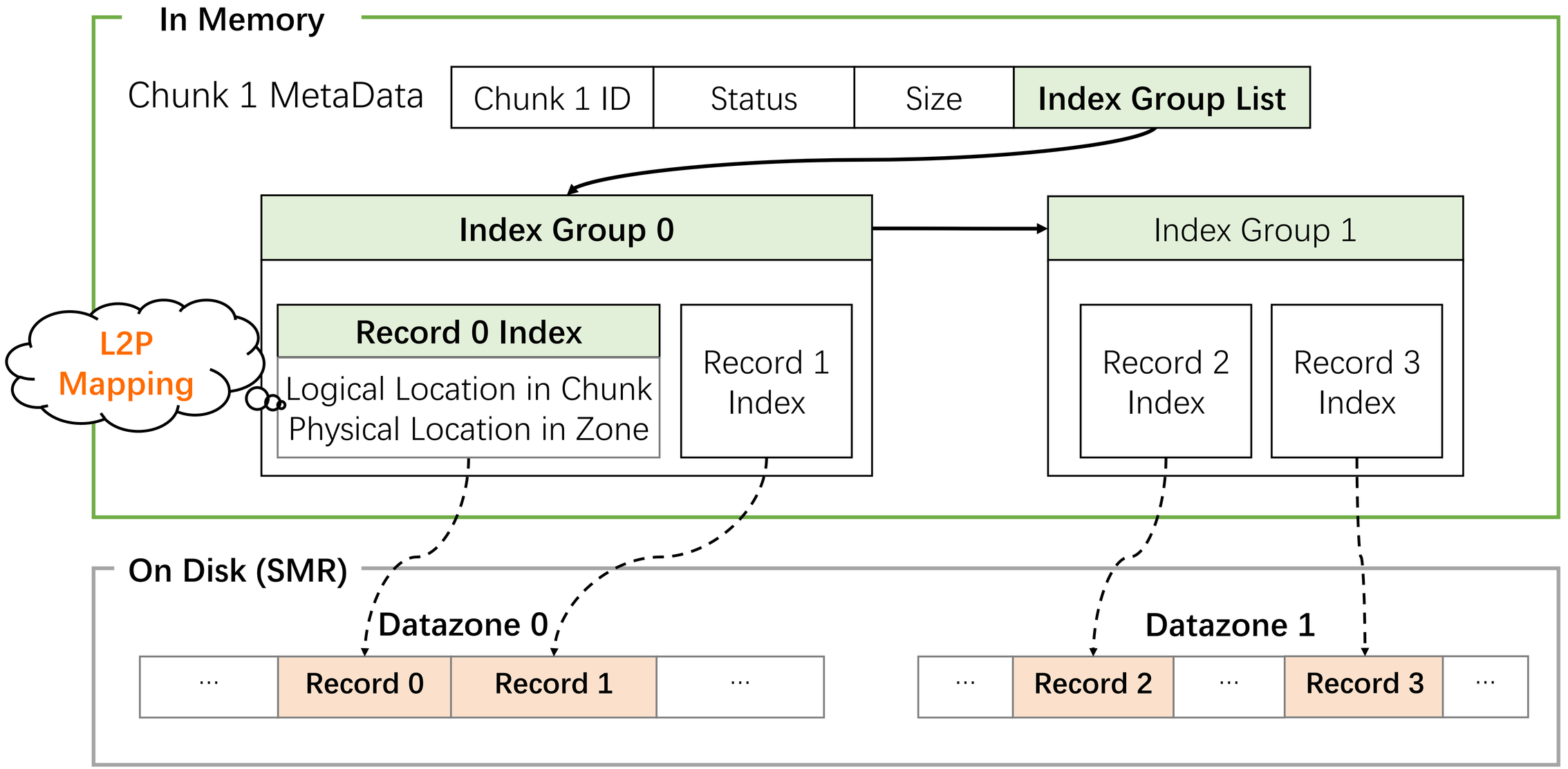


With such a layout…

How does SMRStore organize in-memory structures?

How does SMRStore optimize data placement?

In-Memory Data Index



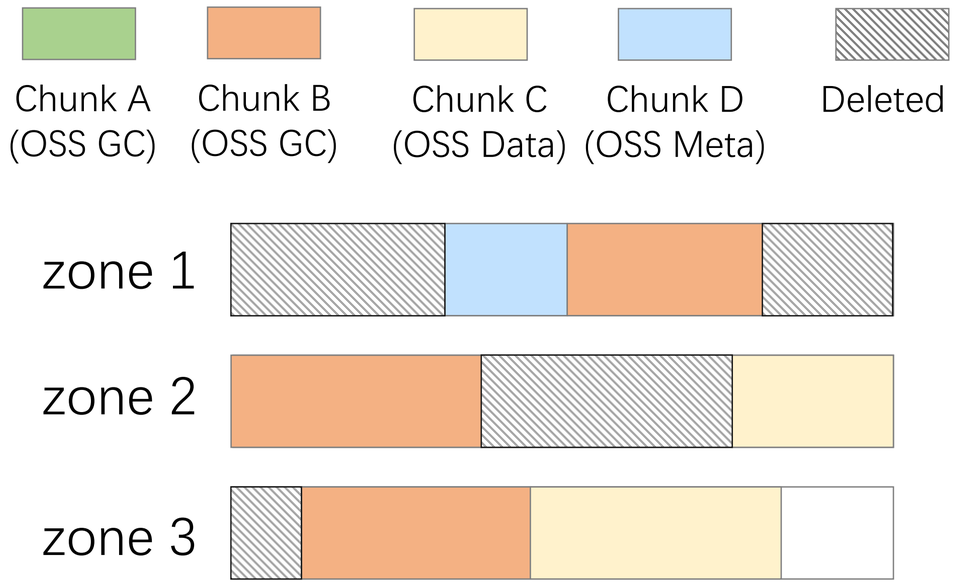
Workload-aware Data Placement

to reduce SMRStore GC

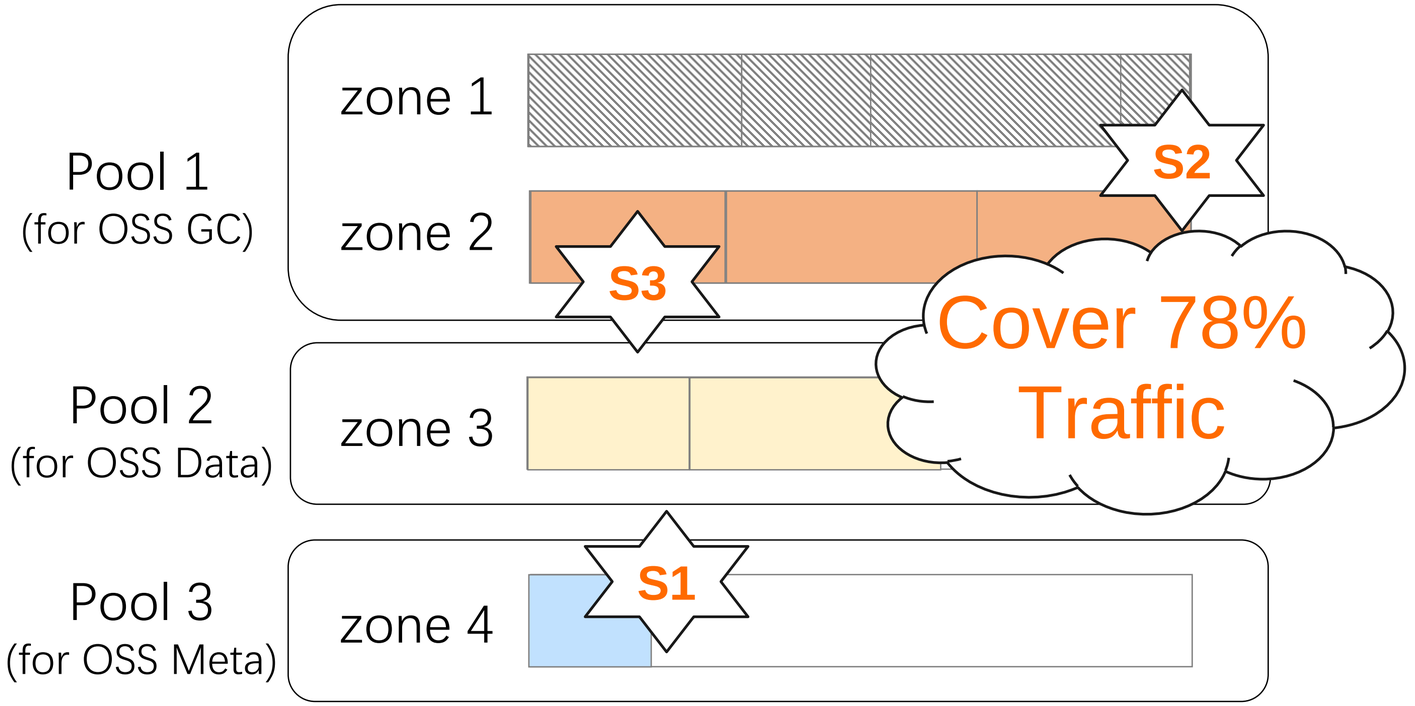
Different streams. → Strategy 1: Separating streams by types.

Most chunks can be large. → Strategy 2: Adapting chunk size limit for datazone.

Low stream concurrency. → Strategy 3: Zone pool & round-robin allocation.



a) No optimization



b) With SMRstore strategies

Alibaba Cloud
Worldwide Cloud Services Partner

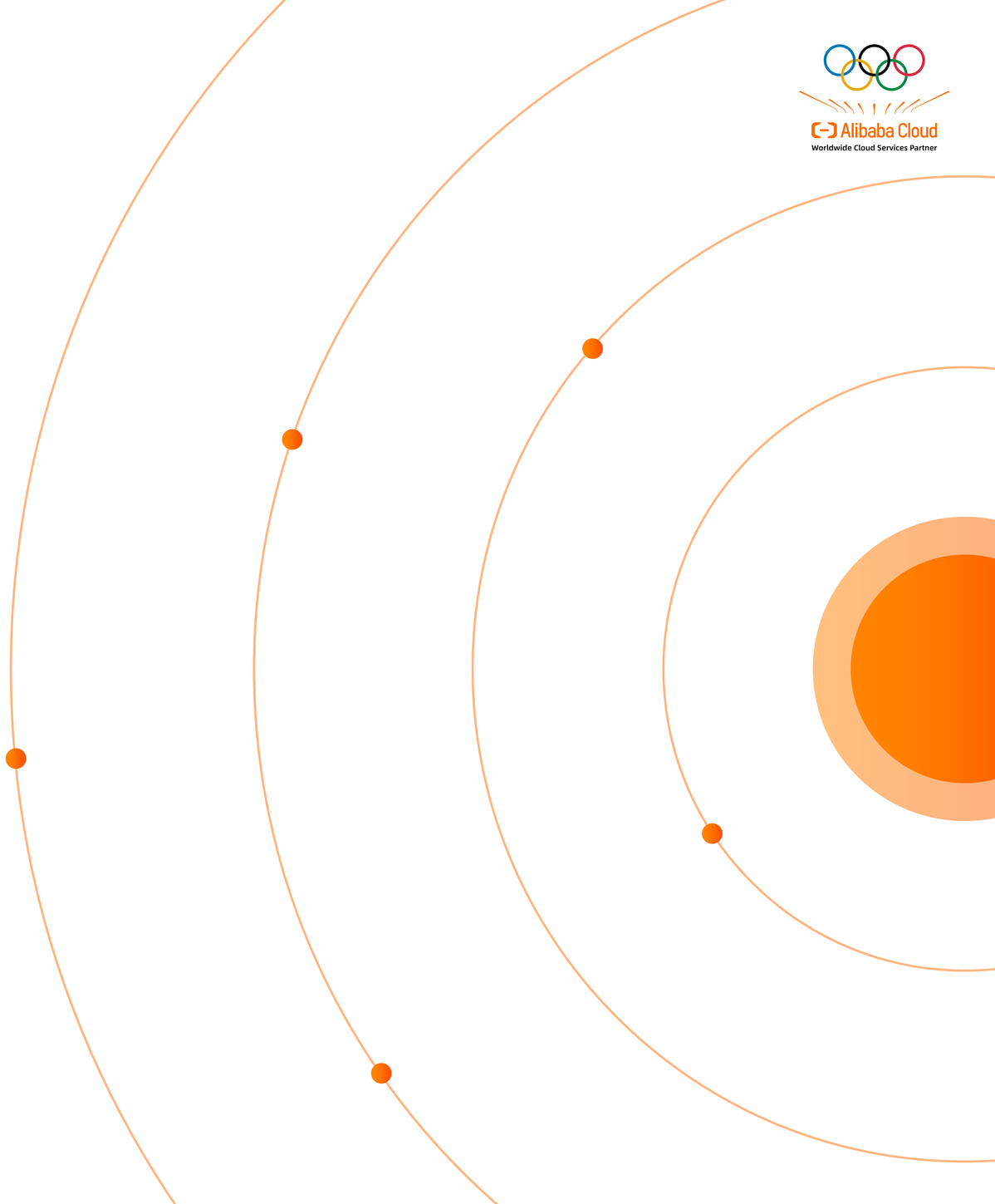
Background

Existing Solutions

Design

Evaluation

Conclusion



Setups

Ext4 (CMR) vs F2Fs (SMR) vs SMRStore (SMR)

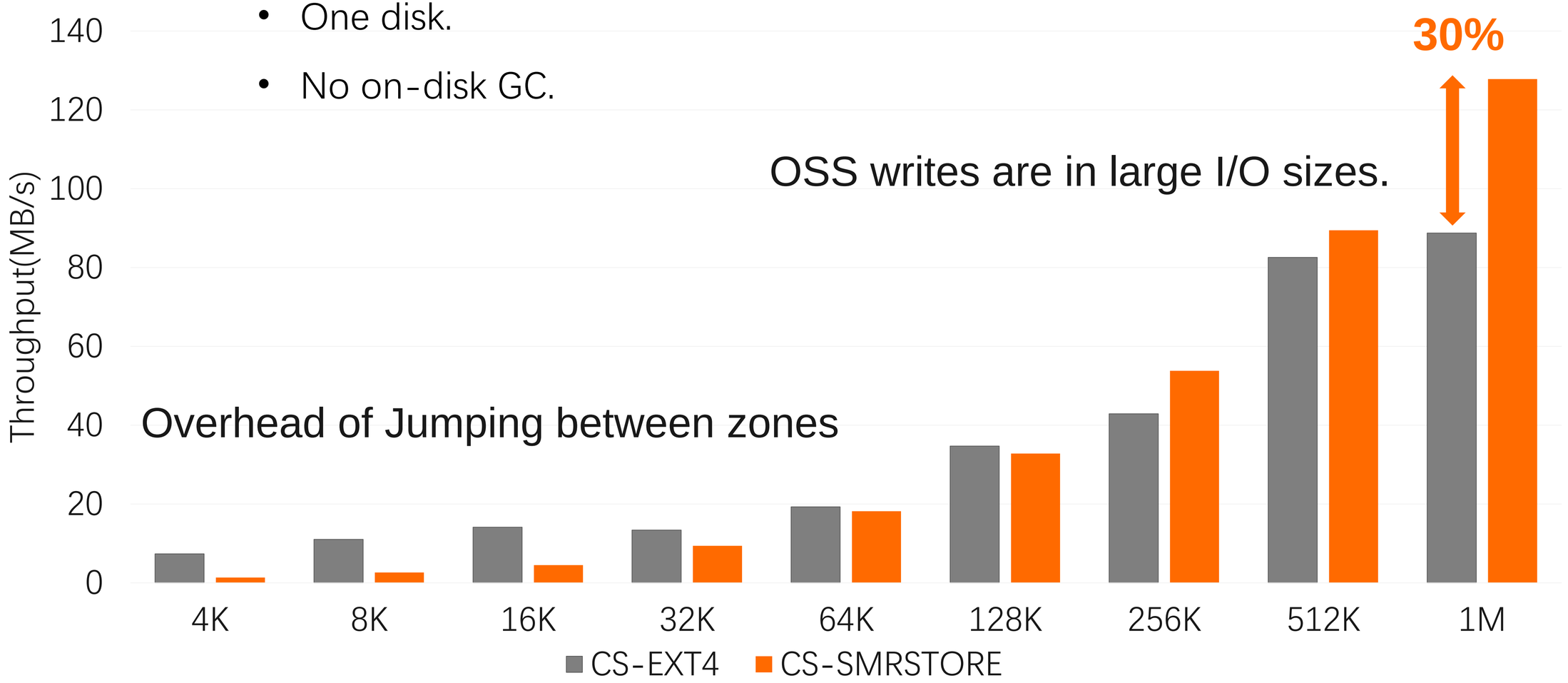
- Microbenchmark - Write throughput (one HDD, **no GC**)
- Macrobenchmark – Consistent Performance (one server, **with GC**)
- Effectiveness of Placement Strategies

	CMR Server	SMR Server
OS	Linux 4.19.91	
CPU	2 * Intel(R) Xeon(R) Platinum 8331C CPU@2.50GHz 48 Cores 96 Threads	
SSD	2 * INTEL SSDPF21Q800GB	
Mem	512G	
HDD	**CMR HDD** **60 * 16T** Rand. 4KB(IOPS): 113 Seq. 512KB(MB/s): 254.8(W) 254.5(R)	**HM-SMR HDD** **60 * 20T** Rand4KB(IOPS): 121 Seq. 512KB(MB/s): 255.7(W) 255.6(R)

up to **30%** Higher Write Performance

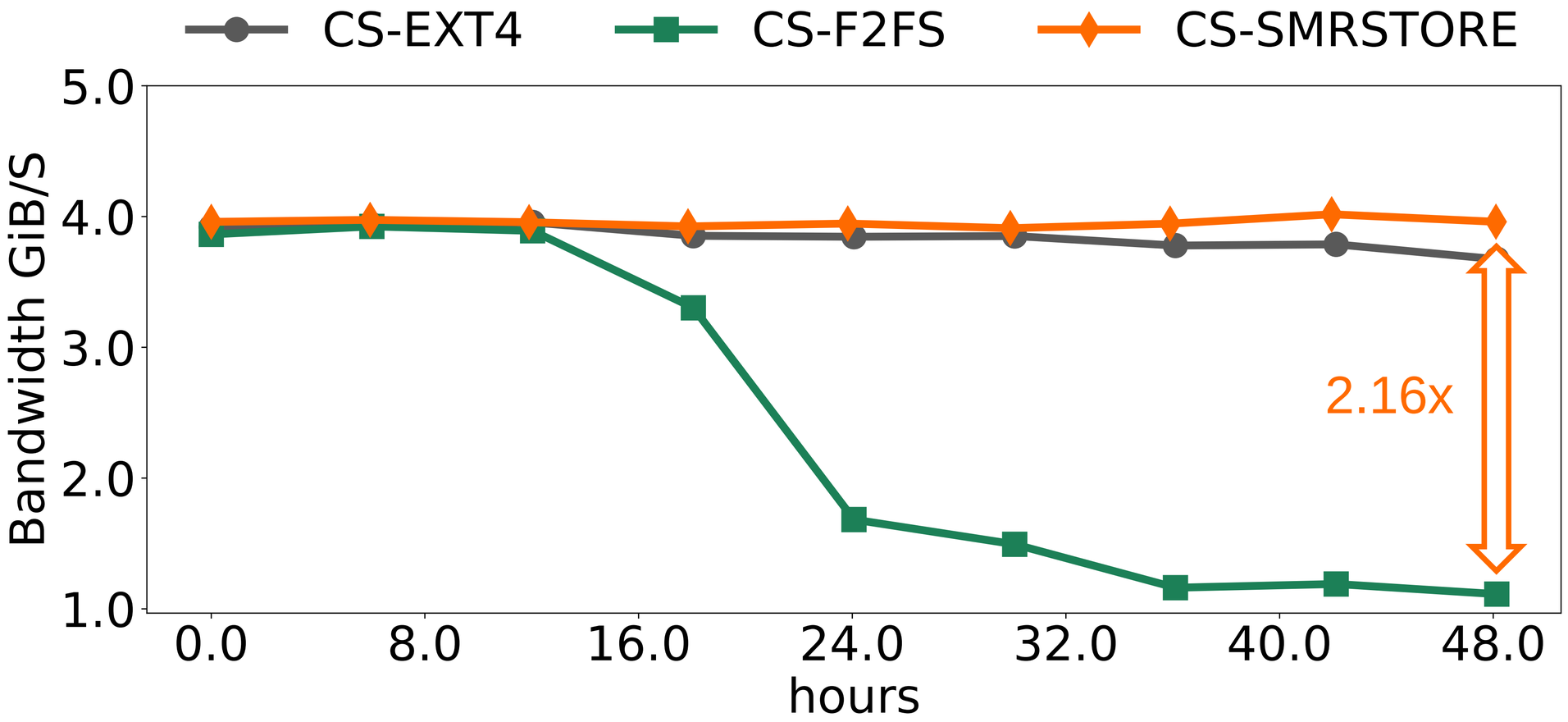
benefits from user-space design

- Fio setups: 4 numjobs, 4 iodepth, 128 nrfiles
- One disk.
- No on-disk GC.

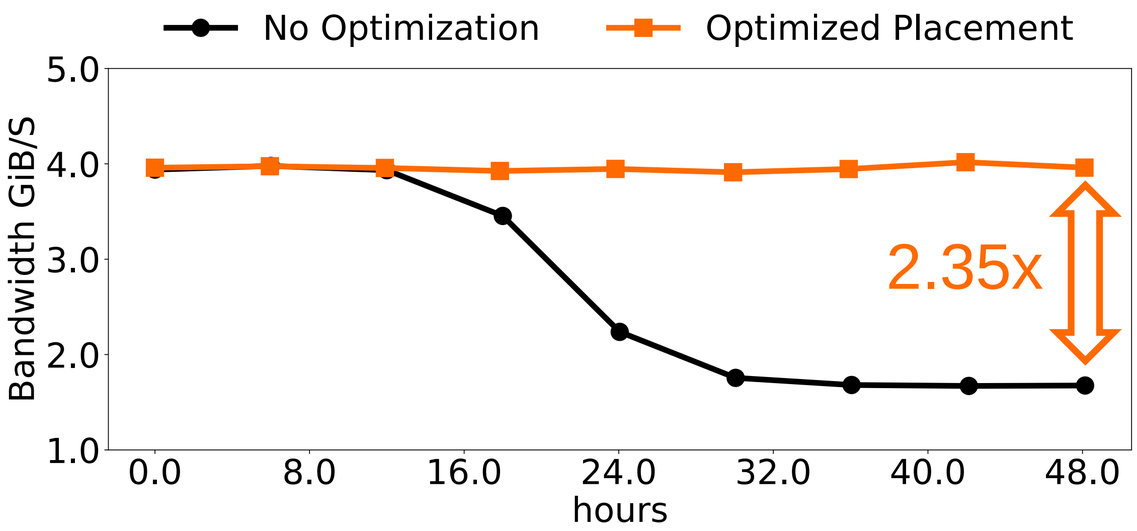


SMRStore **2.16x** faster than F2FS

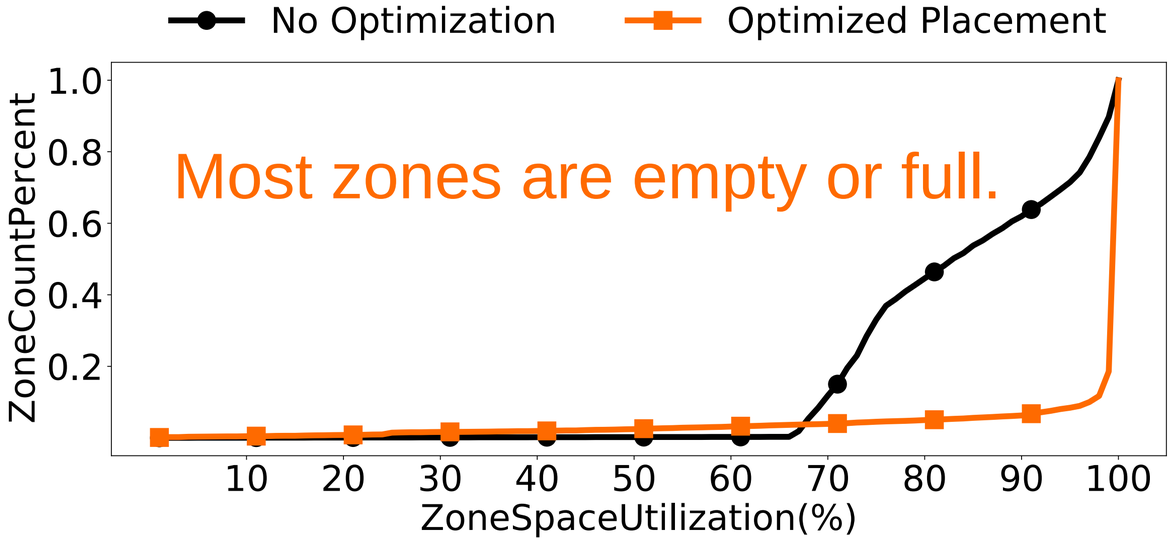
macrobenchmark to simulate OSS workloads



Effectiveness of Placement Strategies



Write throughput comparsion



CDFs of zone space utilization

2.35x higher throughput with optimized placement.

Conclusion

What have we done?

- Build a new **user-space** storage engine to adopt HM-SMR drives.
- **Workload-aware** data placement strategies.

What has SMRStore achieved?

- **Comparable performance** with Ext4 on CMR drives.
- **2.16x faster** than F2FS on SMR drives under OSS workloads.
- Cost-Efficiency: **15%** cost reduction.

Alibaba Cloud
Worldwide Cloud Services Partner

Thanks for Listening!

Q&A

Email: zhousu.zs@Alibaba-inc.com