十二年磨一剑：
三代架构演进，打造高性能、低成本块存储

张 伟 东

阿里云 - 块存储

2024/06/16

阿里云

CONTENT 目录

01 块存储以及 FAST'24 Best Paper 介绍

什么是块存储？What's the Story in EBS Glory?

02 块存储三代架构演进

架构要与时俱进，适应硬件进步和用户需求的变化

03 弹性、可用性、硬件卸载、 What if

没有一蹴而就的成功，唯有百炼成钢的坚持

04 未来架构演进和发展

EBSX 架构、 ESSD PLX、 EED (弹性临时盘)

阿里云

01 块存储以及 FAST’24 Best Paper 介绍

阿里云

● FAST'24 最佳论文

- ✓ FAST 是计算机存储领域国际最高水平的学术会议
- ✓ 论文《What's the Story in EBS Glory: Evolutions and Lessons in Building Cloud Block Store》
- ✓ 阿里云连续两年斩获 FAST 最佳论文，国内唯一！

What's the Story in EBS Glory:
Evolutions and Lessons in Building Cloud Block Store

Weidong Zhang, Erci Xu,* Qiuping Wang, Xiaolu Zhang, Yuesheng Gu, Zhenwei Lu, Tao Ouyang, Guanqun Dai, Wenwen Peng, Zhe Xu, Shuo Zhang, Dong Wu, Yilei Peng, Tianyun Wang, Haoran Zhang, Jiasheng Wang, Wenyuan Yan, Yuanyuan Dong, Wenhui Yao, Zhongjie Wu, Lingjun Zhu, Chao Shi, Yinhu Wang, Rong Liu, Junping Wu, Jiaji Zhu, Jiesheng Wu

Alibaba Group

Abstract

In this paper, we qualitatively and quantitatively discuss the design choices, production experience, and lessons in building the Elastic Block Storage (EBS) at ALIBABA CLOUD over the past decade. To cope with hardware advancement and users' demands, we shift our focus from design simplicity in EBS1 to high performance and space efficiency in EBS2, and finally reducing network traffic amplification in EBS3.

In addition to the architectural evolutions, we also summarize development lessons and experiences as four topics, including: (i) achieving high elasticity in latency, throughput, IOPS and capacity; (ii) improving availability by minimizing the blast radius of individual, regional, and global failure events; (iii) identifying the motivations and key tradeoffs in various hardware offloading solutions; and (iv) identifying the pros/cons of alternative solutions and explaining why seemingly promising ideas would not work in practice.

1 Introduction

Elastic Block Storage (EBS) service is a cornerstone in today's cloud [16, 18, 19]. In EBS, the storage service is in the form of virtual block devices with high performance, availability, and elasticity. The most outstanding characteristic of EBS architecture is the compute-to-storage disaggregation where the virtual machines (compute end) and disks (storage end) are not physically co-located but interconnected via datacenter networks.

In this paper, we start by revisiting the evolutions behind the three generations of EBS at ALIBABA CLOUD [16]. EBS1 marks our initial step in adopting the compute-to-storage philosophy. In EBS1, there are two notable design choices: in-place update from virtual disks (VDs) to physical disks, and the exclusive management of virtual disks. First, EBS1 directly maps a VD inside the virtual machine (VM) as a series of 64 MiB Ext4 files in the backend storage server. Moreover, EBS1 employs a fleet of stateless BlockServers to manage VDs where each VD is exclusively handled by a BlockServer. While EBS1 had been successfully deployed on more than 300 HDD-backed clusters, its limitations also unfolded. The straightforward virtualization led to severe space amplification and performance bottlenecks.

We then developed EBS2 with two significant changes: the log-structured design, and VD segmentation. First, we employed the Pangu [35] distributed file system as our storage backend, and redesigned the BlockServers to convert VDs' all writes to sequential appends. By switching to a log-structured layout, EBS2 still used three-way replication for incoming writes but could transparently perform data compression and erasure coding (EC) in the background during garbage collection (GC). Moreover, EBS2 split VDs into finer segments (32 GiB each), thus shifting the mapping between VDs and BlockServers from VD level to Segment level. With the above two changes, EBS2 was able to reduce the space efficiency from 3 (i.e., three-way replication) in EBS1 to 1.29 on average in the field. Moreover, supercharged with SSDs, an EBS2-backed VD can achieve up to 1 M IOPS and 4,000 MiB/s throughput with 100 μs-level latency on average. Unfortunately, EBS2 also faced a significant challenge. That is, the traffic amplification factor increased to 4.69, namely 3 (foreground replication write) plus 1 (background GC read) and 0.69 (background EC/compression write).

Hence, we built EBS3 to reduce traffic amplification using online (i.e., foreground) EC/compression via two techniques: Fusion Write Engine (FWE), and FPGA-based hardware compression. FWE aggregates write requests from different segments (if necessary) to meet the size requirement of EC and compression. Moreover, EBS3 offloads the compute-intensive compression to a customized FPGA for acceleration. As a result, EBS3 can reduce the storage amplification factor from 1.29 to 0.77 (after compression) and the traffic amplification factor from 4.69 to 1.59 while still maintaining performance similar to EBS2. Since release, EBS3 has been deployed on more than 100 clusters, serving over 500K VDs.

Figure 1 outlines the chronological progression of Alibaba EBS since 2012. We highlight the time of major releases (i.e., EBS1 to EBS3), the integration of key techniques (e.g., Luna, our user-space TCP stack [46]) and the adoption of advanced hardware (e.g., Persistent Memory in EBSX). The evolution of EBS demonstrates a shift in focus from performance to space

*Corresponding author.

阿里云 什么是块存储？

●弹性块存储（EBS）

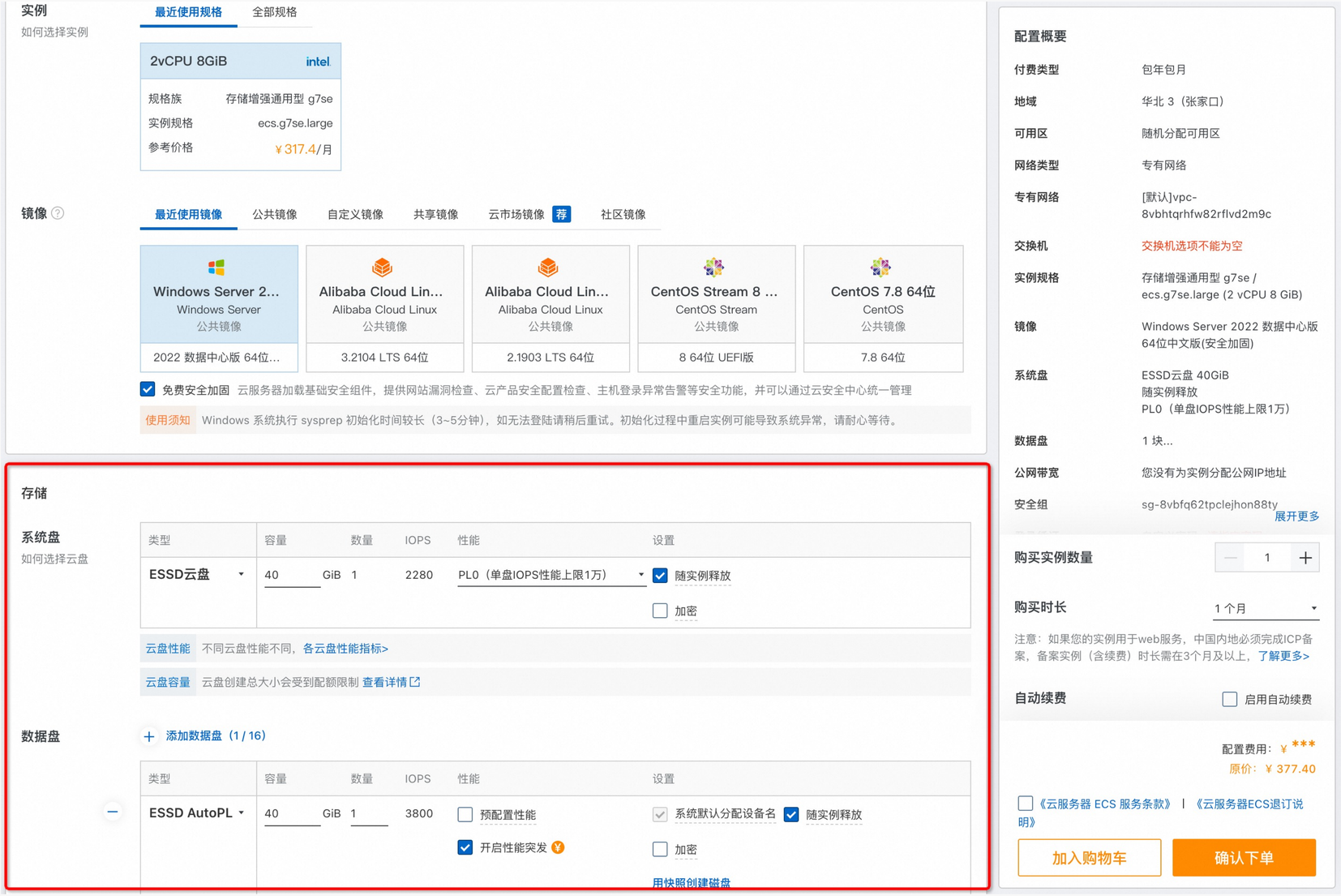
- ✓ 虚拟机: Virtual Machine (VM)
- ✓ 虚拟盘、云盘: Virtual Disk (VD)

●服务目标

- ✓ 高可靠
- ✓ 高性能
- ✓ 高弹性
- ✓ 高可用

●计算-存储分离架构

- ✓ 虚拟机和云盘分别在不同的物理集群



阿里云

02 块存储三代架构演进

●设计目标

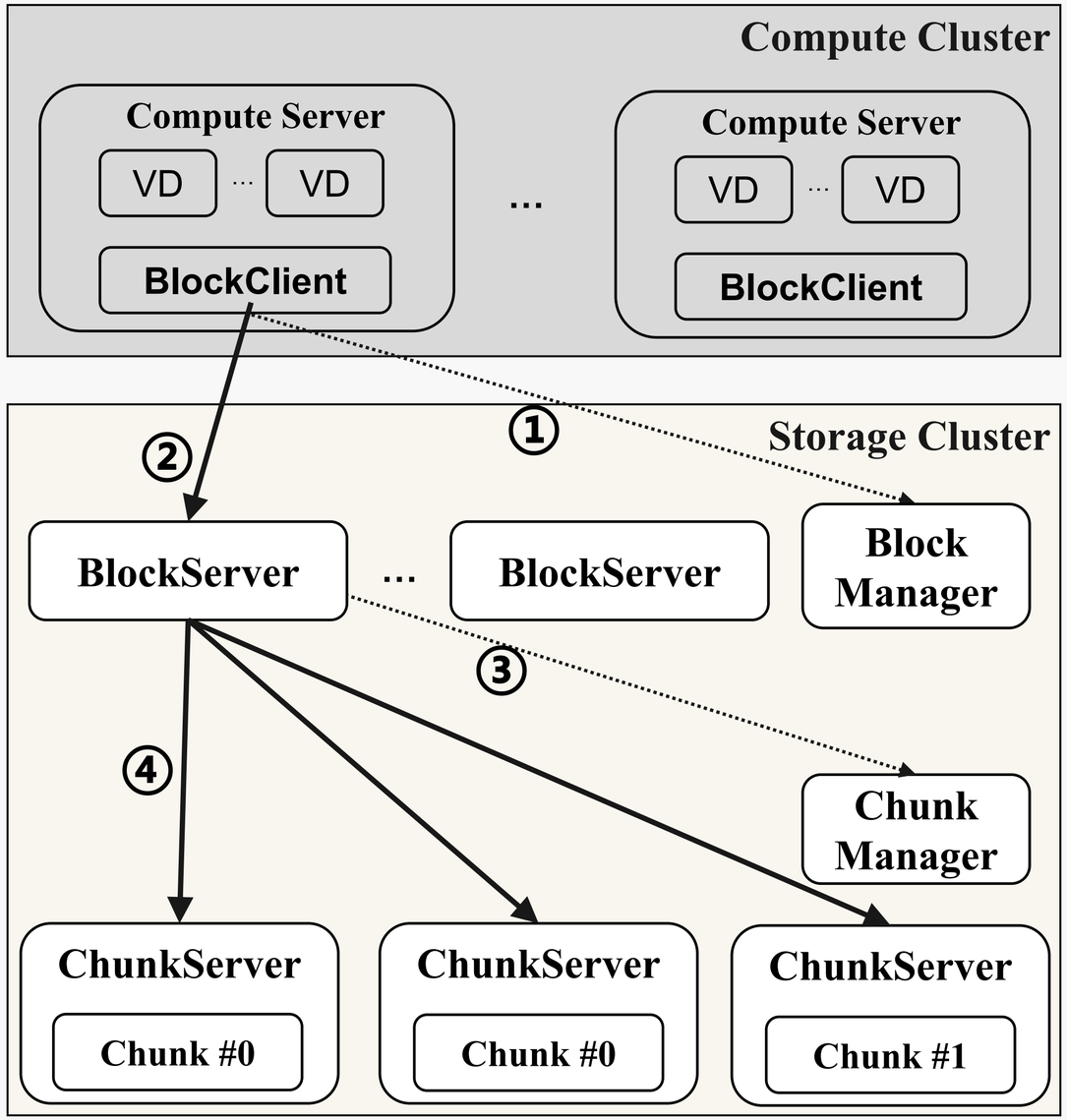
- ✓ 简单、直接，可以快速部署和快速上线

●架构

- ✓ 云盘空间被分为大小相同的 Chunks (64MiB)
- ✓ 两层架构：BlockServer + ChunkServer
- ✓ 每个 Chunk 都是一个 EX4 File

●特点

- ✓ 原地更新 (In-place updates) : VD = Ext4 files
- ✓ N(VDs)-to-1(BlockServer) mapping



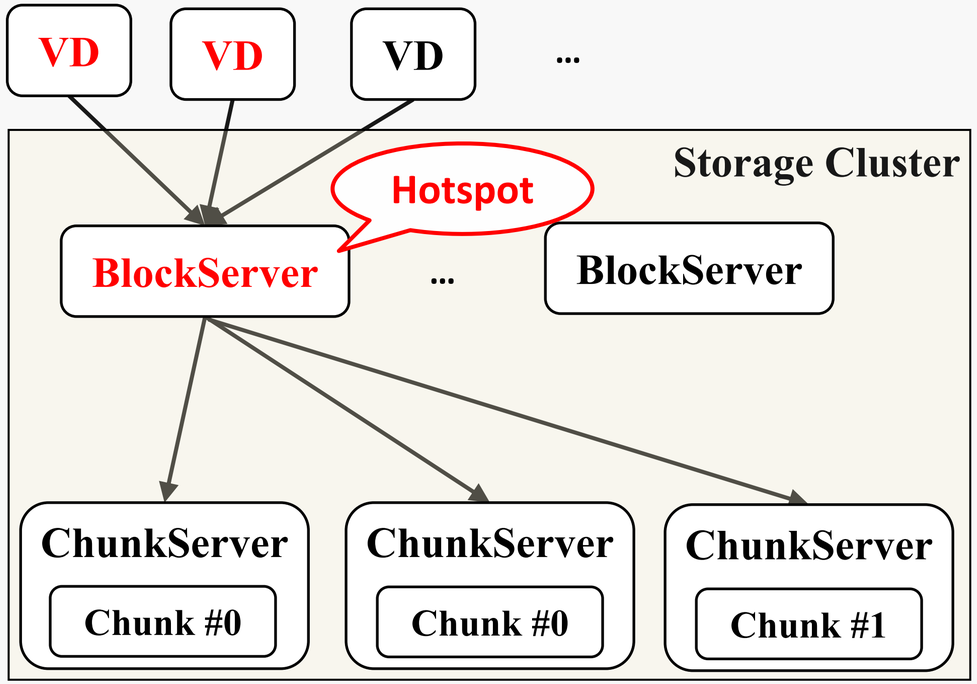
阿里云 EBS1: An Initial Foray

●部署情况

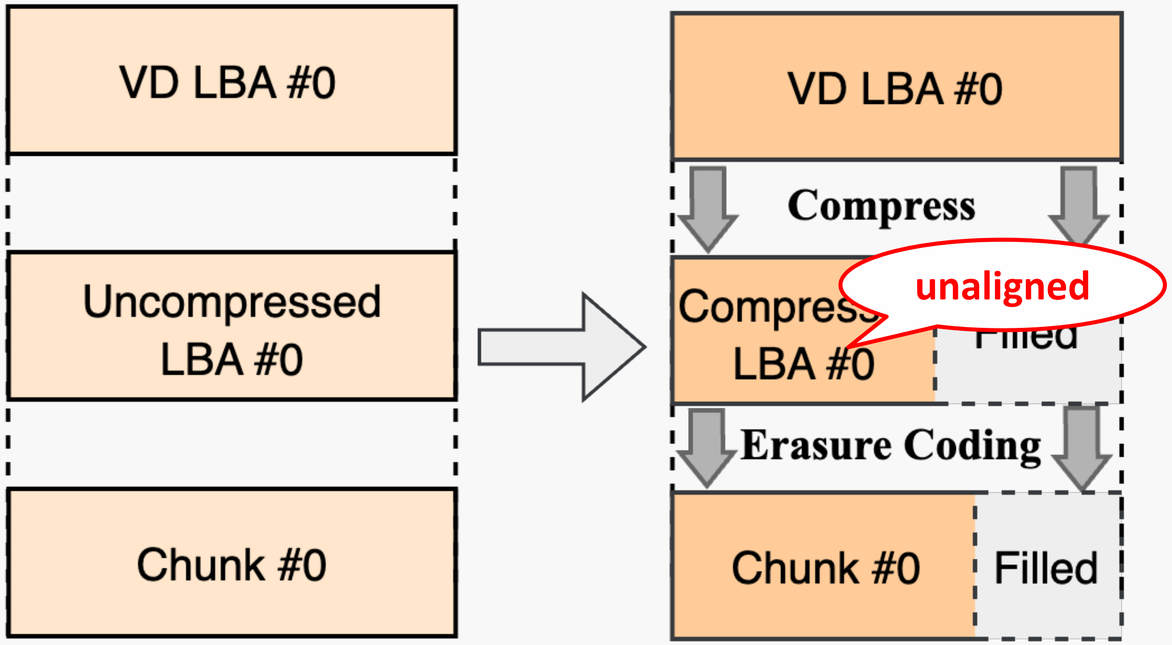
✓ 2012年上线，服务超过100万块云盘，累计部署超过数百个存储集群并存储数百 PB 数据

●问题

✓ N-to-1 mapping 导致单点瓶颈，限制性能提升



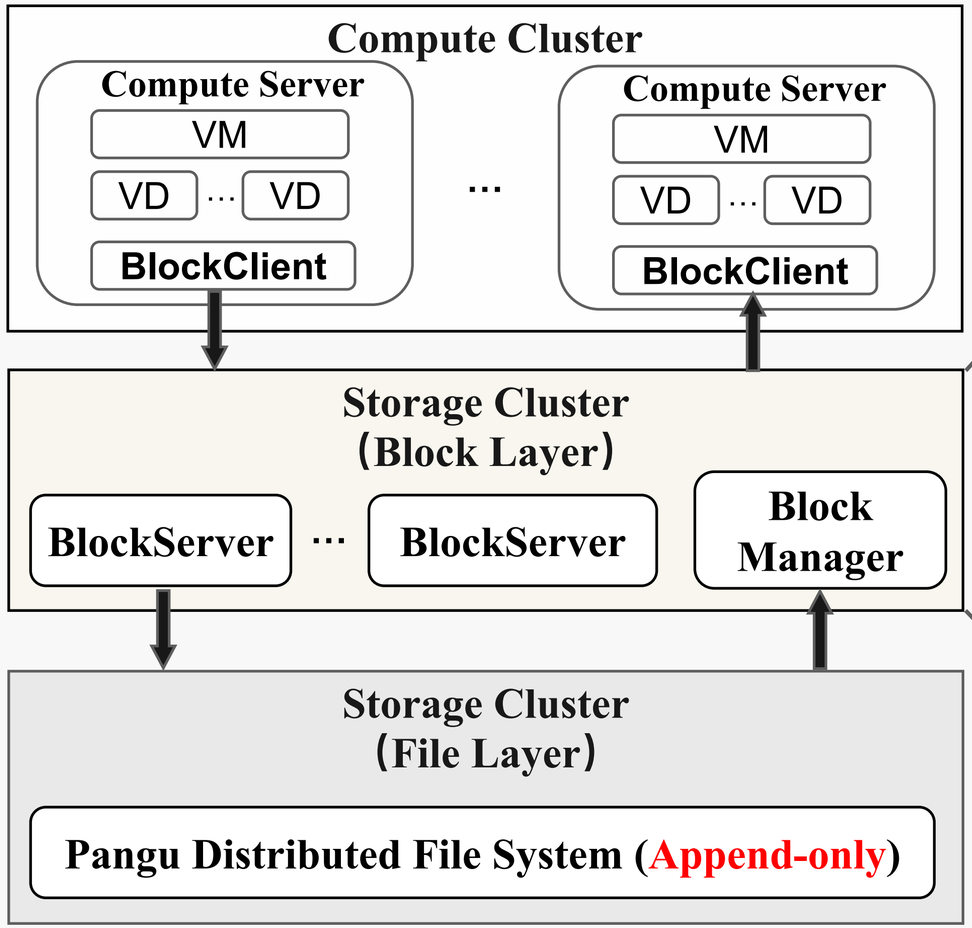
✓ 原地更新导致数据压缩和 EC 难以实现，阻碍了成本降低



阿里云 EBS2: Speedup with Space Efficiency

●设计目标

✓ 高性能 + 高空间效率

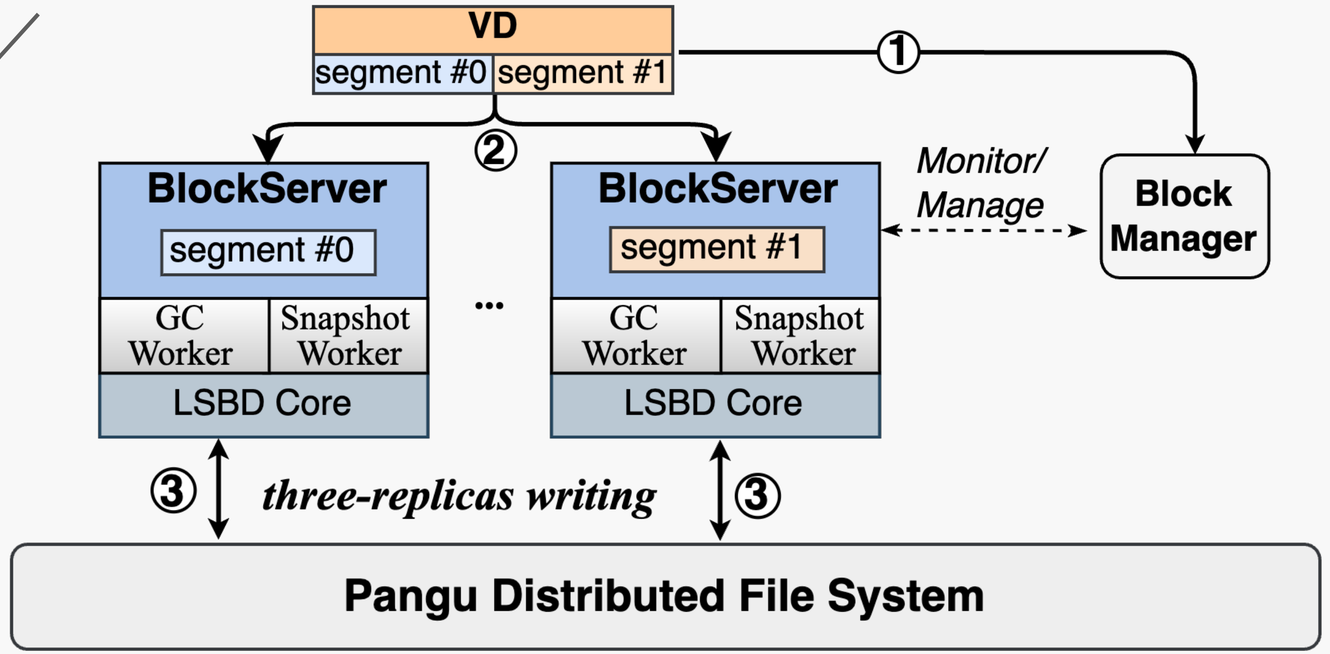


●关键设计

✓ 云盘条带化 (Disk segmentation)

✓ 日志结构 (Log-structured Block Device, LSBD)

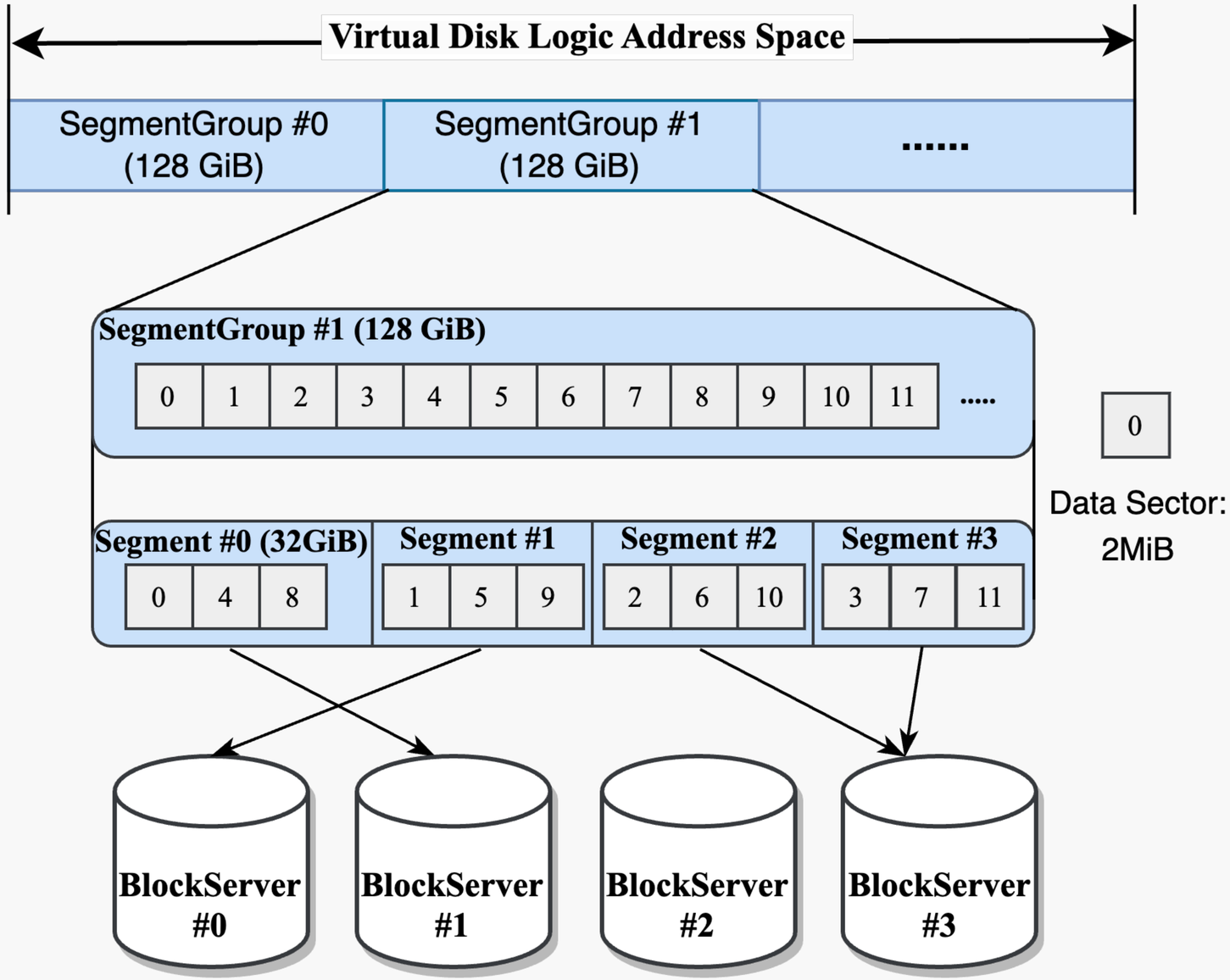
✓ 带 数据压缩和 EC 的垃圾回收



阿里云 EBS2: Speedup with Space Efficiency

●云盘条带化

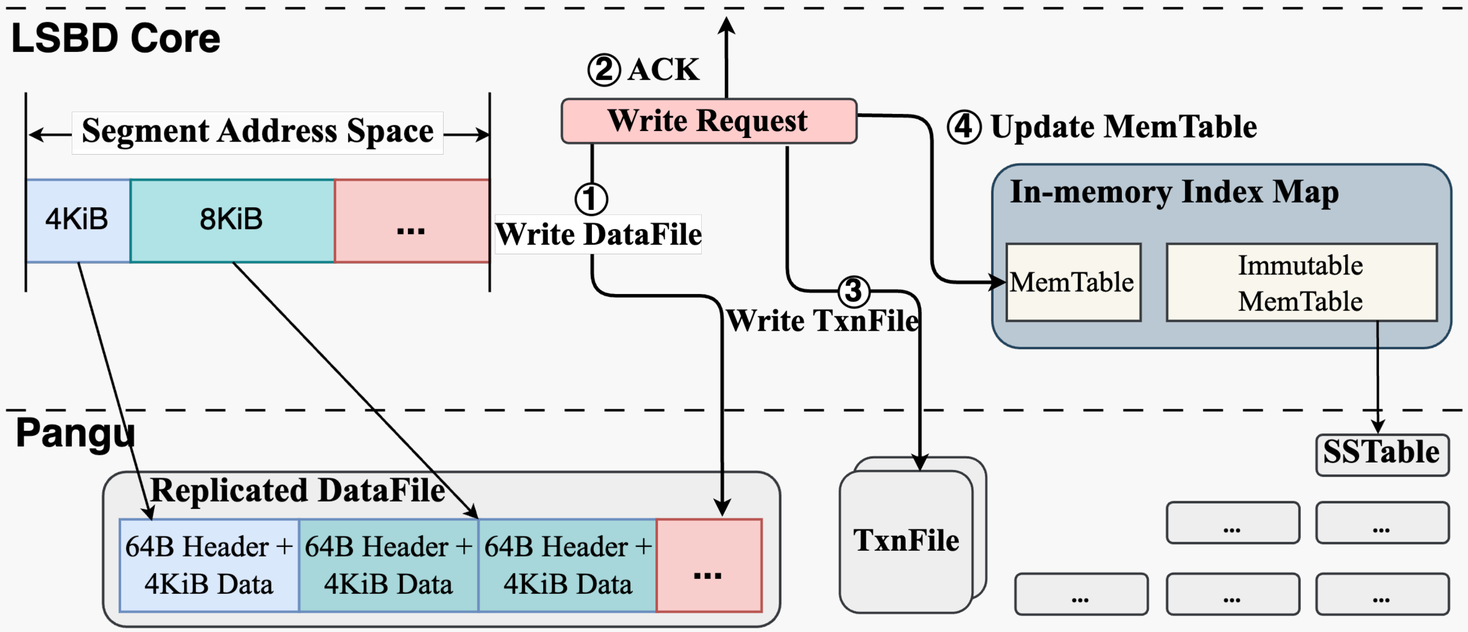
- ✓ 整个云盘的逻辑空间被划分为多个连续的 SegmentGroup
- ✓ 每个 SegmentGroup 被组织为一系列Data Sectors
- ✓ Data Sectors 以 Round-Robin的方式分配给 Segments
- ✓ Segments 是 BlockServer 运行的最小粒度



阿里云 EBS2: Speedup with Space Efficiency

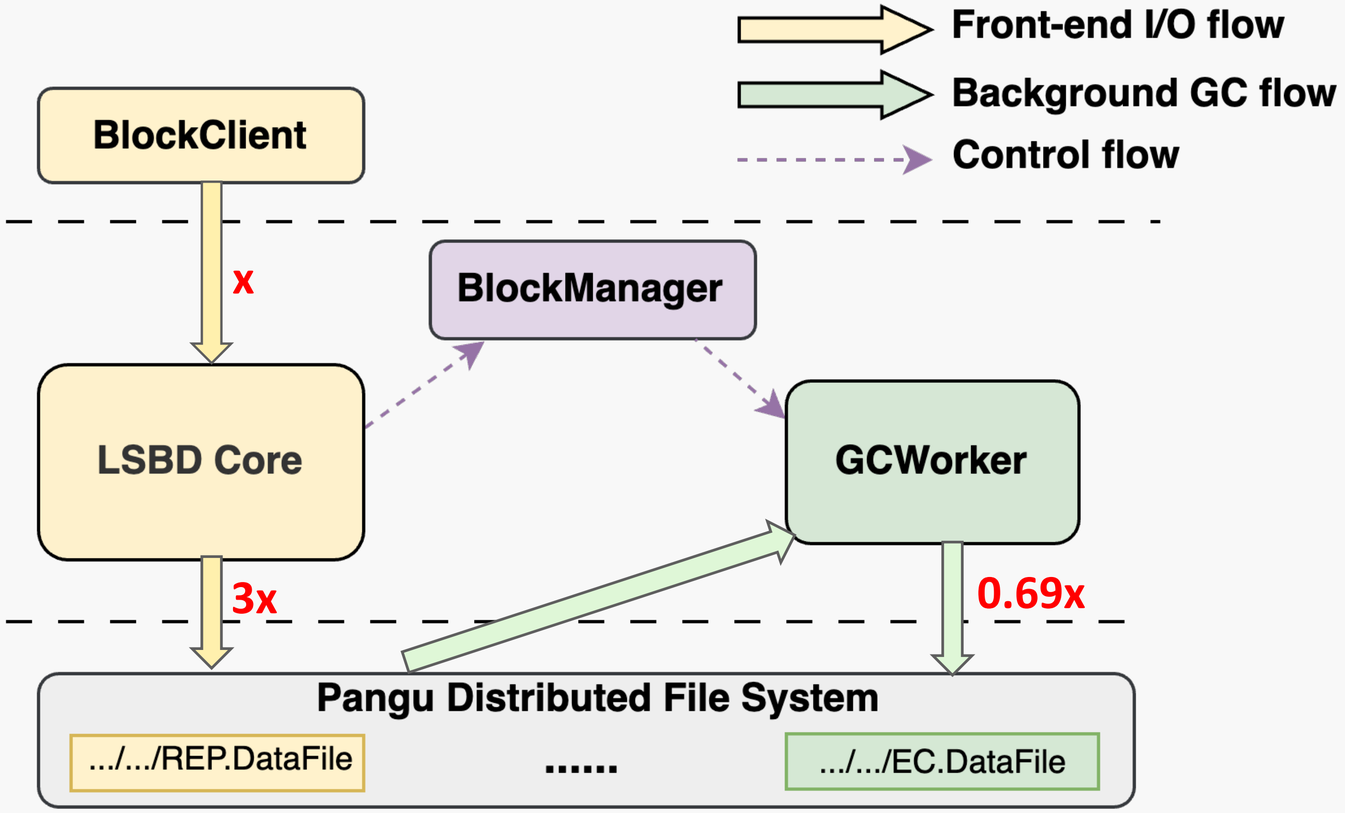
● 日志结构（LSBD）

- ✓ DataFile = (4KB data +64B Header) x N
- ✓ Txnfile 用来加速故障恢复
- ✓ 内存中的 Index Map 用来加速读请求



●带数据压缩和 EC 的垃圾回收

- ✓ LSBD 将流量分为前端（用户 I/O）和后端（GC ）
- ✓ 垃圾回收以 DataFiles 的粒度运行
- ✓ 垃圾回收使用 EC(8, 3) 和 LZ4/ZSTD 压缩算法将“REP.DataFiles”转换为“EC.DataFiles”



$$SpaceCost_{EBS1} = 3$$

$$SpaceCost_{EBS2} = 1(\text{original}) \times 0.5(\text{compressed}) \times \frac{8+3}{8}(\text{EC}) = 0.69$$

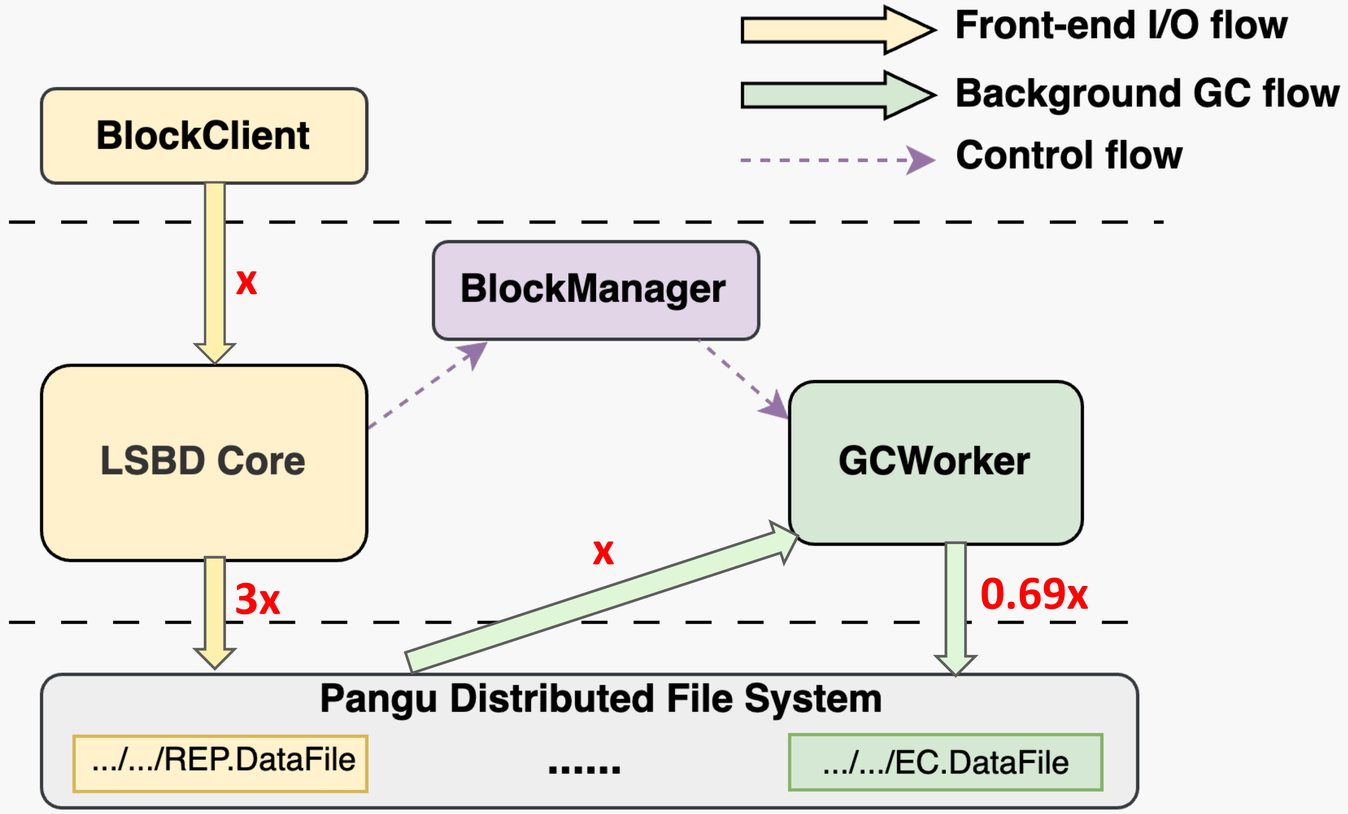
阿里云 EBS2: Speedup with Space Efficiency

●部署

- ✓ 100μs 平均写延迟 和 1 百万 IOPS per VD.
- ✓ 500个存储集群 并且 服务超过 2 百万块云盘.
- ✓ 每份用户数据实际占用空间低至 1.29.

●问题

- ✓ 流量放大比高达 4.69.
- ✓ 随着 SSD 每 GiB 成本的下降，云存储已经从 空间敏感型 转向 流量敏感型.



$$TrafficAmplification_{EBS1} = 3x \div x = 3$$

$$TrafficAmplification_{EBS2} = (3x + x + 0.69x) \div x = 4.69$$

阿里云 EBS3: Foreground EC/Compression

●设计目标

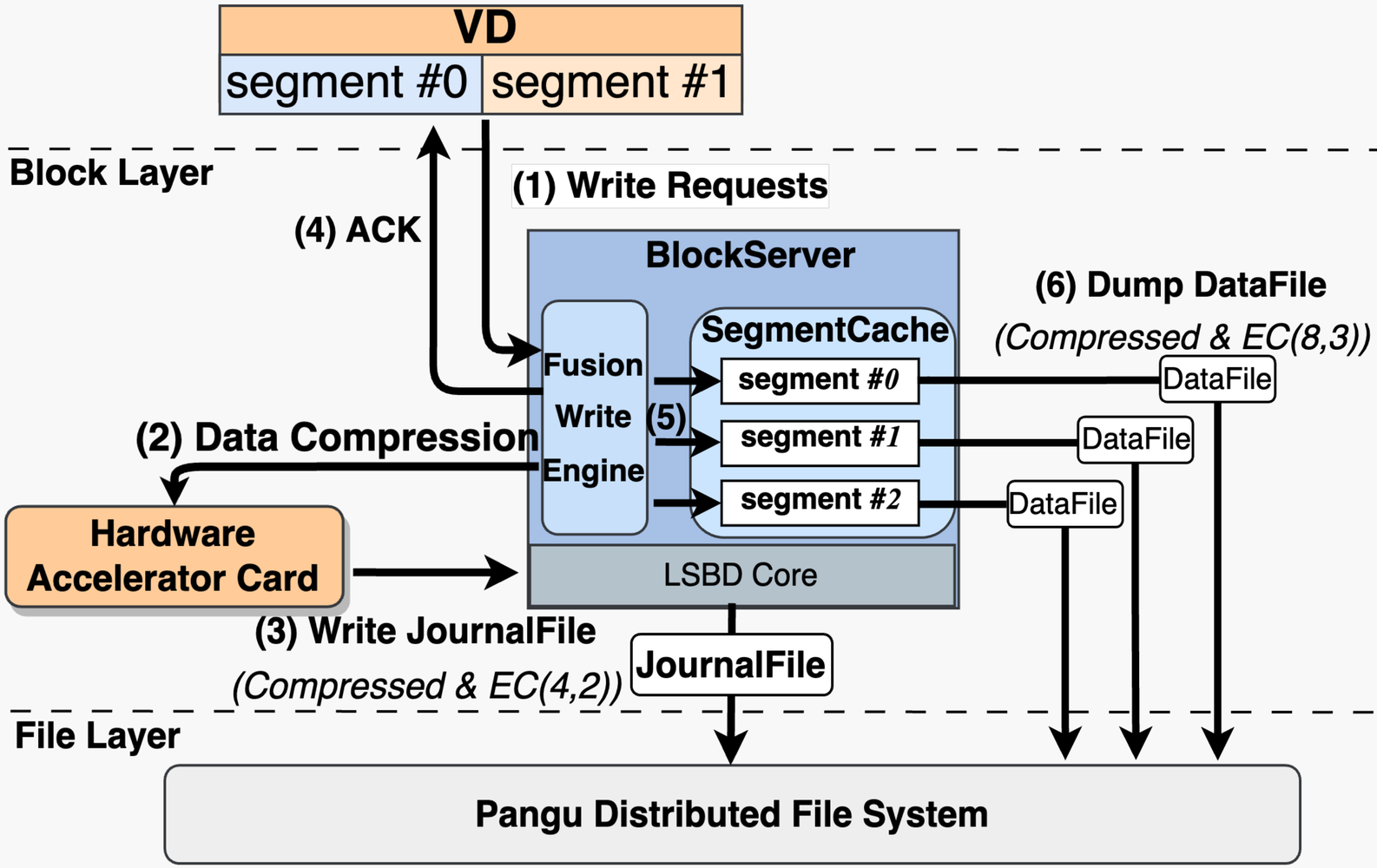
- ✓ 降低流量消耗和存储空间成本
- ✓ 无性能损失

●关键设计

- ✓ 写入路径分叉（Bifurcated Write Path）
- ✓ 融合写入引擎（Fusion Write Engine）
- ✓ 硬件卸载压缩（FPGA-based compression offloading）

●部署

- ✓ 超过 100 个集群，服务超过 500,000 块云盘
- ✓ 每份数据的占用空间降低至 0.77



阿里云 EBS3: Foreground EC/Compression

●设计目标

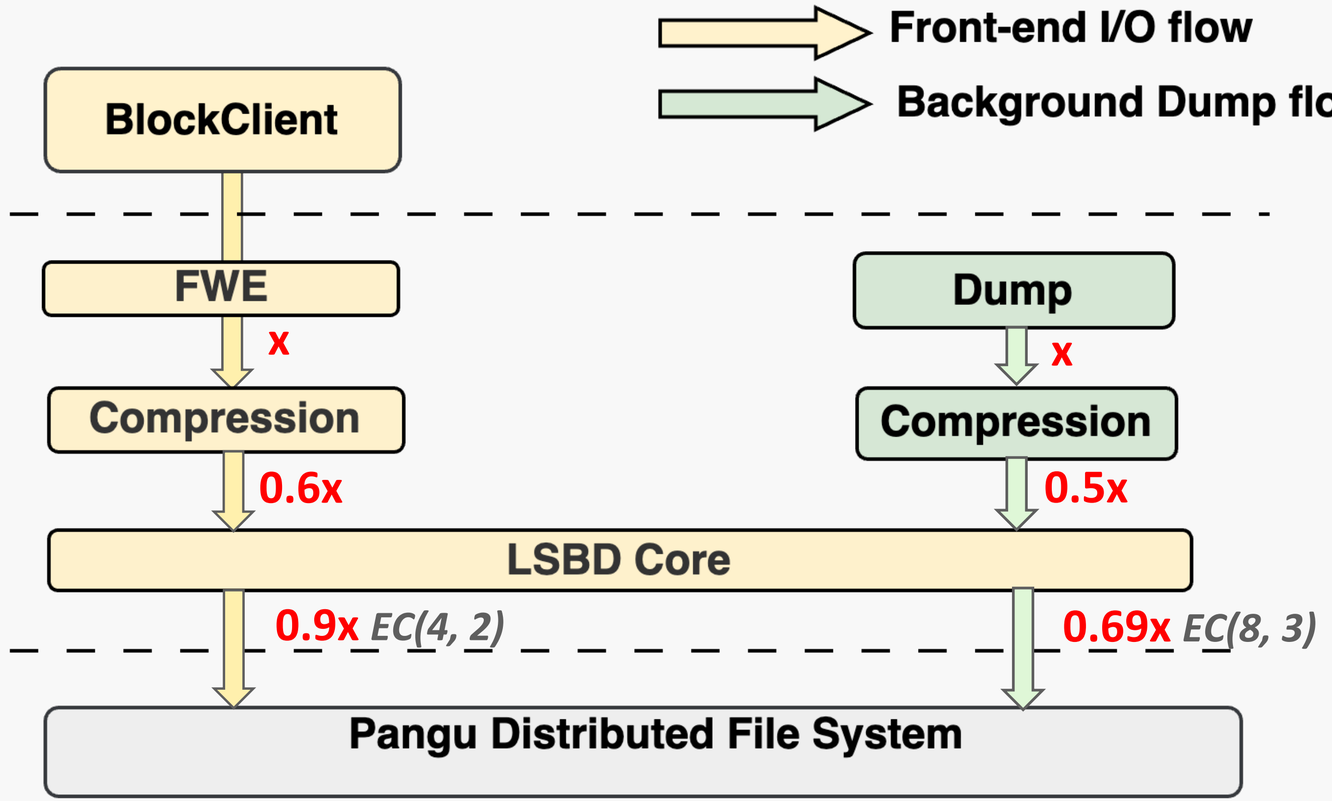
- ✓ 降低流量消耗和存储空间成本
- ✓ 无性能损失

●关键设计

- ✓ 写入路径分叉（Bifurcated Write Path）
- ✓ 融合写入引擎（Fusion Write Engine）
- ✓ 硬件卸载压缩（FPGA-based compression offloading）

●部署

- ✓ 超过 100 个集群，服务超过 500,000 块云盘
- ✓ 每份数据的占用空间降低至 0.77



$$TrafficAmplification_{EBS2} = (3x + x + 0.69x) \div x = \mathbf{4.69}$$

$$TrafficAmplification_{EBS3} = (0.9x + 0.69x) \div x = \mathbf{1.59}$$

阿里云 三代 EBS 架构对比

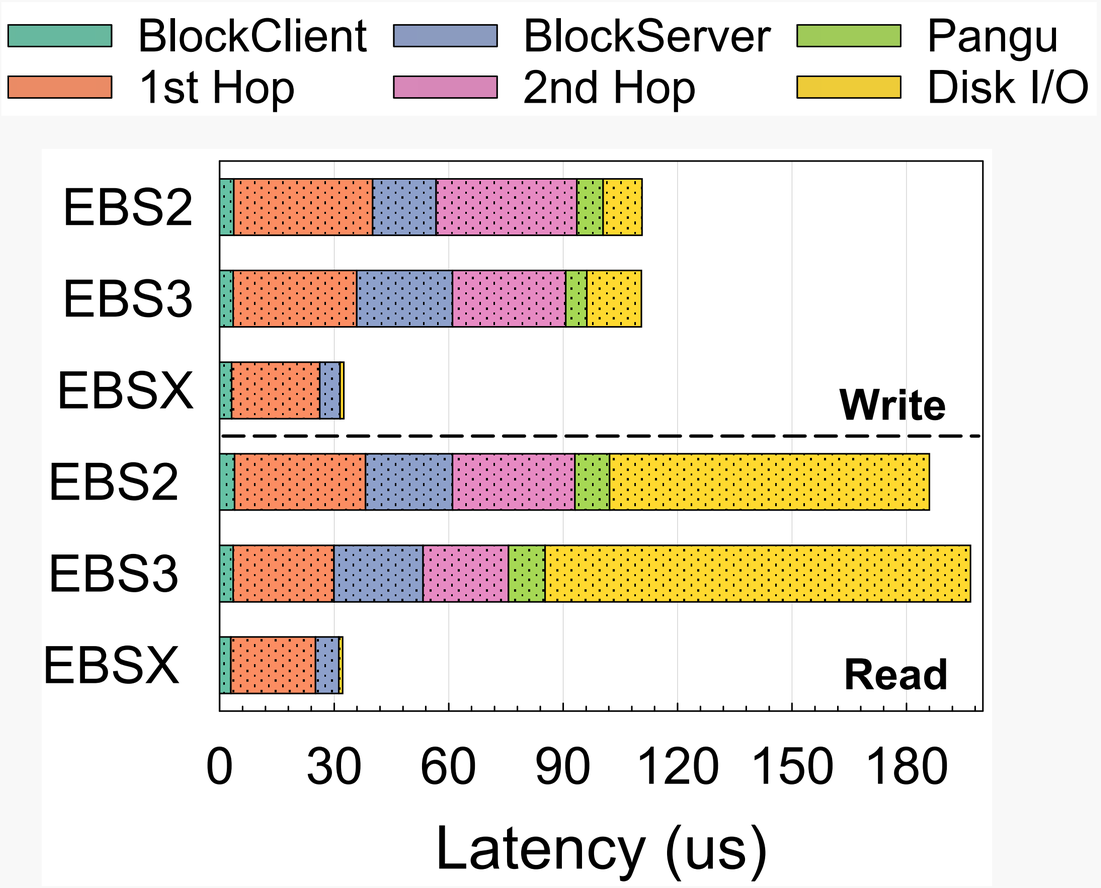
	EBS1	EBS2	EBS3
平均延迟	Millisecond Level	Hundred-microsecond Level	Hundred-microsecond Level
最大 IOPS / Throughput	25,000	1,000,000	1,000,000
关键特征	In-place updates N-to-1mapping	**后台** **EC & Compression**	**前台** **EC & Compression**
空间成本 **（ Replicas per Data ）**	3	1.29	0.77
流量放大比	3	4.69	1.59

阿里云

03 弹性、可用性、硬件卸载、What if

●延迟弹性是粗粒度的

✓ 由硬件架构决定

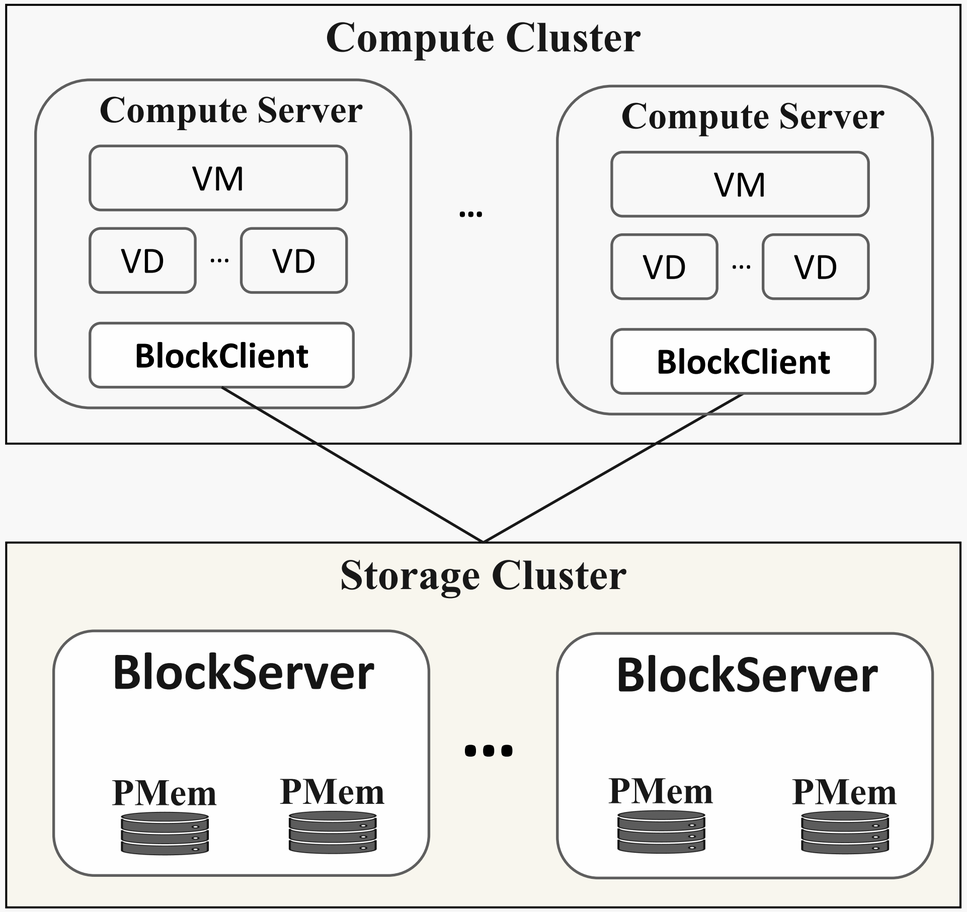


● 延迟弹性是粗粒度的

- ✓ 由硬件架构决定

● EBSX

- ✓ 缩短路径 *(例如，减少一跳网络)*
- ✓ 使用更快的存储介质*(例如，用 PMem 替代 SSD)*
- ✓ 简单高效的数据一致性协议



弹性：延迟

●延迟弹性是粗粒度的

✓ 由硬件架构决定

● EBSX

✓ 缩短路径 *(例如，减少一跳网络)*

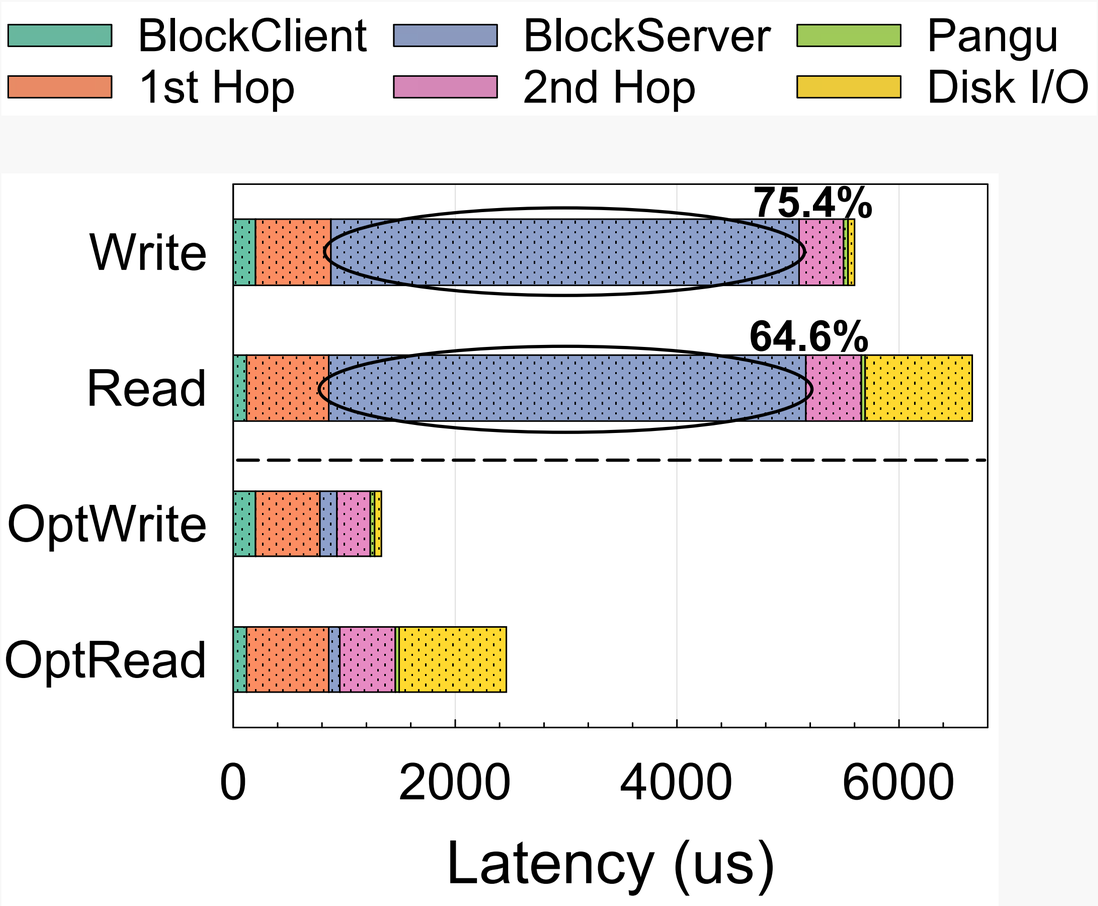
✓ 使用更快的存储介质*(例如，用 PMem 替代 SSD)*

✓ 简单高效的数据一致性协议

●长尾延迟

✓ 软件可能是造成长尾延迟的主要原因

✓ 有效手段：将 IO 与后台任务（例如, GC 和周期性扫描）分开

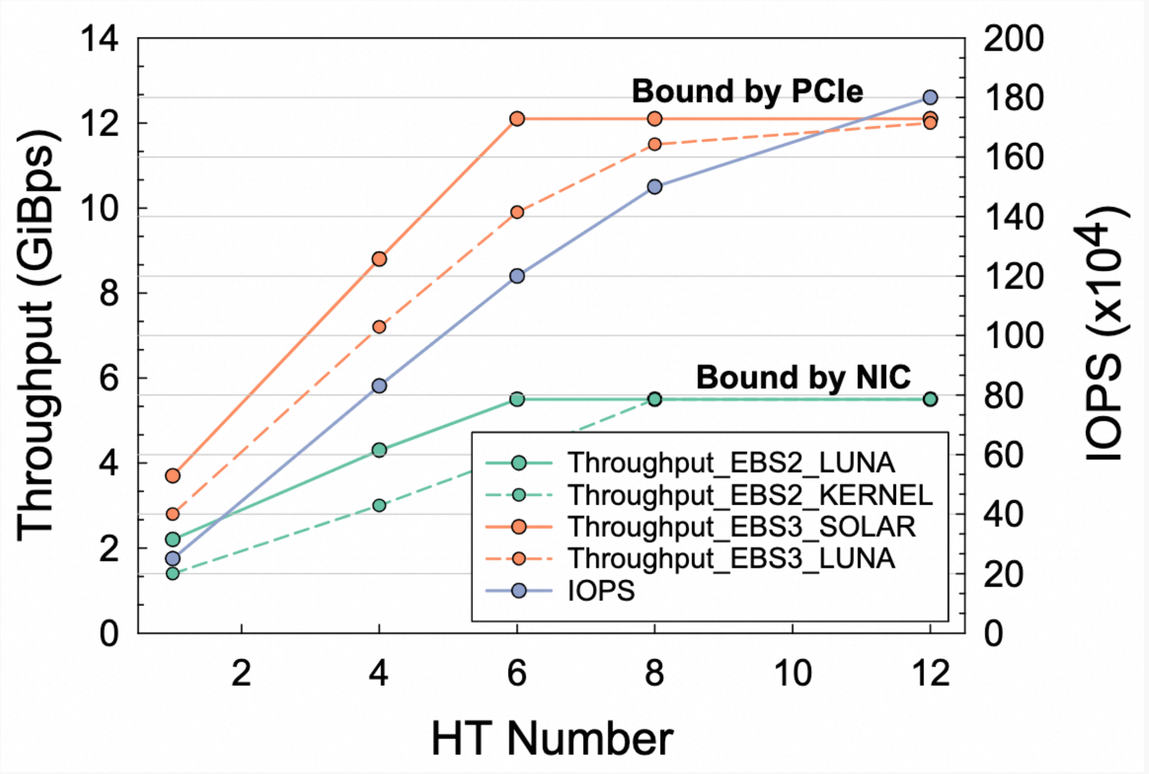


99.999th 长尾延迟

阿里云 弹性：IOPS/Throughput

● 上限由 BlockClient 决定

- ✓ 后端可以轻松扩展
- ✓ BlockClient受处理和转发能力约束
- ✓ 从内核空间到用户空间，然后到硬件卸载

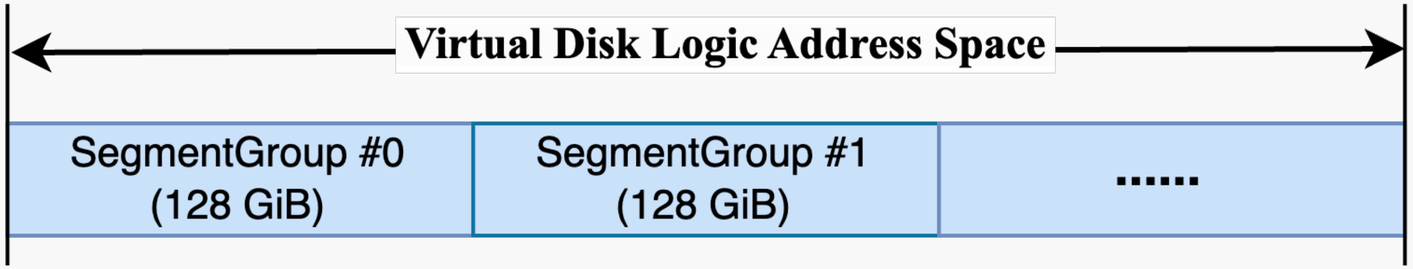


● High IOPS/Throughput is often desired but not always needed

- ✓ Auto performance level (AutoPL) 云盘：按需使用，无需改变容量
- ✓ Base + Burst strategy：高效地将 IOPS/吞吐量分配给 VD
- ✓ Base IOPS/Throughput 意味着绝对可以满足
- ✓ Burst IOPS/Throughput 意味着尽力满足

● 灵活调整空间大小

- ✓ 通过添加或删除 SegmentGroup 实现云盘容量调整
- ✓ 虚拟磁盘大小高达 64 TiB



● 快速云盘克隆

- ✓ Pangu 文件支持*硬链接（Hard Link）*
- ✓ 1 分钟内支持最多克隆 10,000 块云盘（每个 40 GiB）

可用性、硬件卸载、What if

● 可用性挑战和解决方案

(See Section 4)*

- ✓ **Challenge 1:** a BlockServer crash impacts more VDs
  Solution: Federated BlockManager (Two-layer control nodes)
- ✓ **Challenge 2:** Segment migration leads to cascading failures
  Solution: Logical Failure Domain (Limited migration)

● 硬件卸载

(See Section 5)*

形式遵循功能
Form follows function

- ✓ **FPGA is not ideal: expensive, high failure rates**
- ✓ **Blockclient offloading: FPGA → ASIC**: 1. cost-friendly 2. a fixed set of functions.
- ✓ **BlockServer offloading: FPGA → Many-core ARM**: 1. cost-friendly 2. comparable performance

● What if?

(See Section 6)*

- ✓ **Q1: W/o log-structured design? Both cost and performance cannot move forward.**
- ✓ **Q2: EBS with open-source software? Co-design will be never possible.**
- ✓ **Q3: Not separating Pangu? Slow down the development of EBS.**

更多细节请参阅论文 * Zhang W, Xu E, Wang Q, et al. What's the Story in EBS Glory: Evolutions and Lessons in Building Cloud Block Store[C]// FAST 24.

阿里云

04 未来架构演进和发展

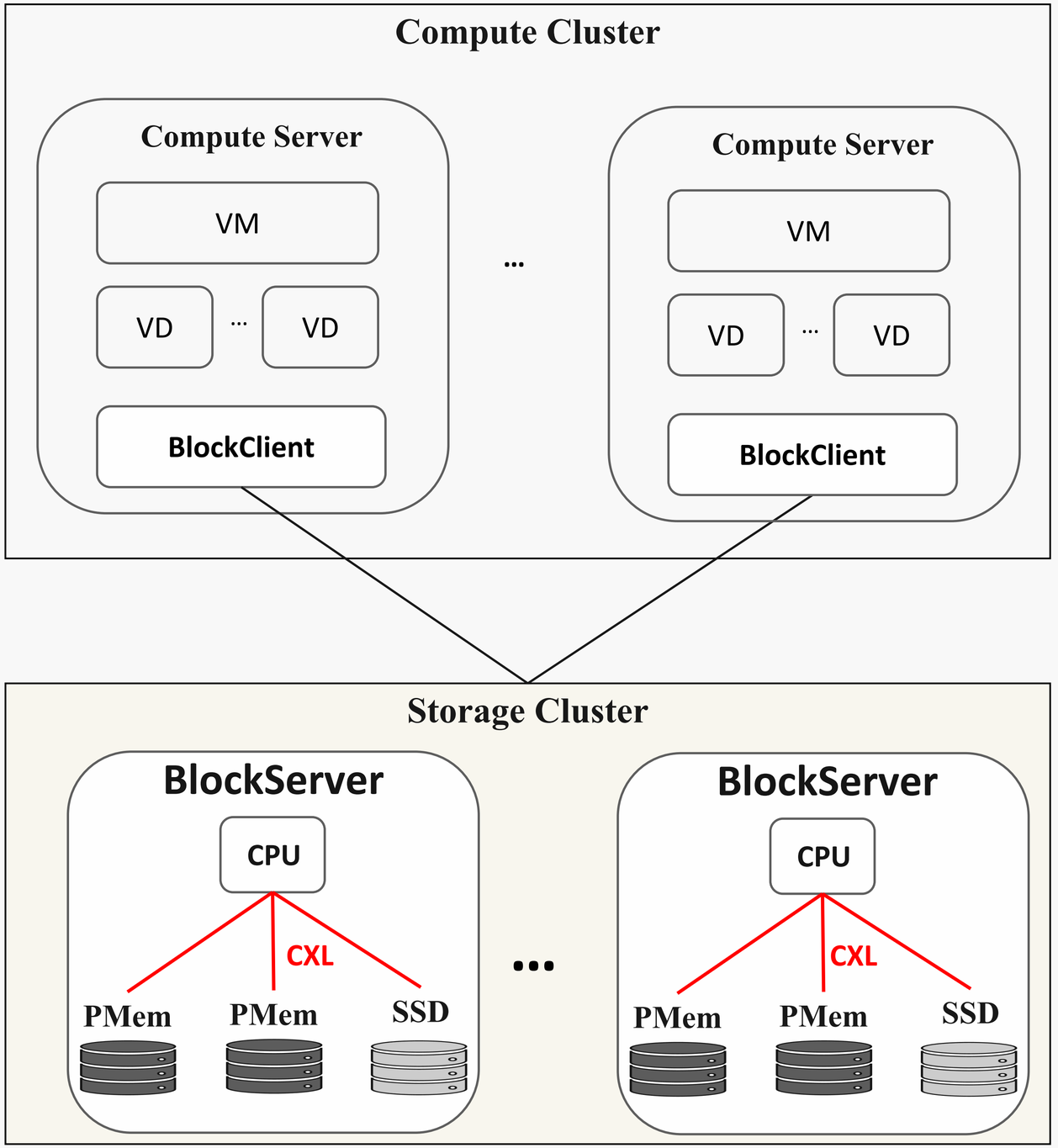
阿里云 EBSX 架构

EBSX 架构的新特性

- ✓ **超低延迟**：4KB 读/写的平均延迟 **35us**
- ✓ **超高 IOPS 和吞吐**：单云盘支持最高 **3,000,000 IOPS**，**12GiB/s** 吞吐
- ✓ 单机故障对用户**无影响**（On the way）
- ✓ 成本可控

ESSD PLX、弹性临时盘（EED）

- ✓ **基于 EBSX 架构**
- ✓ 既有**本地盘的性能**，又有**云盘的弹性和可用性**
- ✓ 适用于高性能数据库、数据库缓存、大数据 shuffle 等



阿里云

弹性临时盘，阿里云新一代本地盘产品

具备弹性容量能力的“本地盘”

相同点：

- 高性能、低价格
- 无可靠性和企业级特性

不同点：

- 性能随容量线性增长
- 按需使用，可随实例创建，也可以单独创建挂载到指定实例
- 可随业务需求快速释放，降低成本
- 创建时可指定容量，也可以在线扩容
- 故障后可快速恢复，无需等待更换硬件

功能点	弹性临时盘	本地盘实例 物理盘
存储容量按需选择	✓	×
灵活创建和释放	✓	×
故障快速更换	✓	×
加密	×（即将支持）	×
在线扩容	✓	×

计算，为了无法计算的价值

E-mail: zhangweidong.zwd@alibaba-inc.com /
iszhangwd@hotmail.com