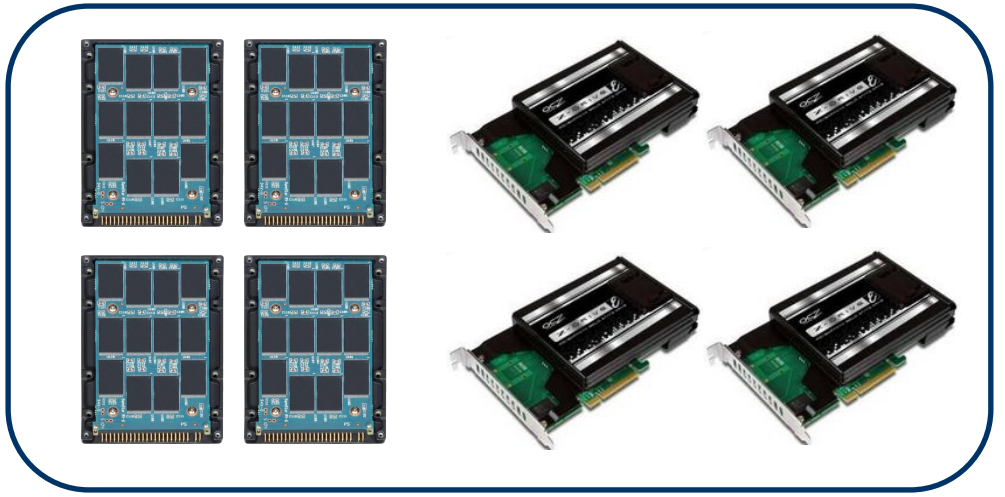
RAID 遇上 SSD



SSD优点

- 在便携式环境中的可靠性、无噪音
  - 无电机马达
- 快速启动
  - Does not need spin up
- 低读延迟（low read latency），随机读性能较好
  - 没有寻道时间(*x* us per page/4KB)
- 确定性的读性能(Deterministic read performance)
  - The performance does not depends on the location of data
- 静态功耗低、发热量低

SSD缺点

- 单位容量成本更高
  - 0.1$/GB vs. 0.05$/GB
- 有限的写入擦除次数(Lifetime)
  - 100000 writes for SLC (MLC is even fewer)
  - high endurance cells may have an 1-5 million
- 由于擦除块而导致的写入速度越来越慢
- 延迟不可预测性(Unpredictable Latency)
- 长尾延迟
- 写放大(Write Amplification)

写不友好设备

RAID for SSD

SSDs are less array-friendly than hard disks

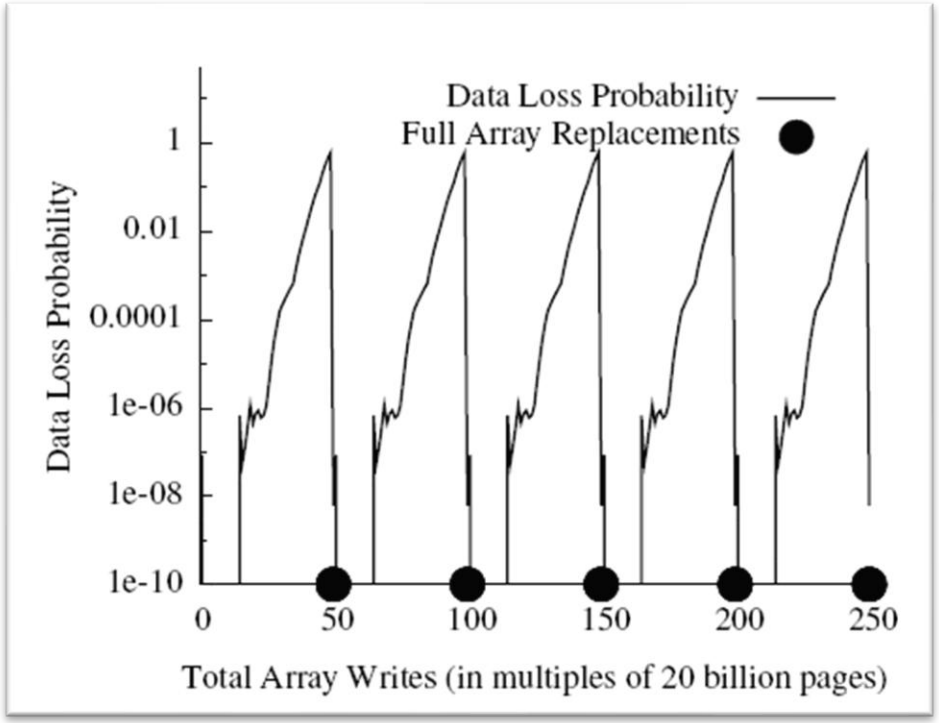
- SSD擦写次数受限，延迟波动大，尾延迟问题明显
  - SLC：100,000次， MLC：5,000-10,000次
- RAID中数据尽量均匀分布到各成员盘
  - RAID0， RAID5
- RAID5中校验数据块的擦写更频繁
  - N-disk RAID-5，任何一个成员盘的数据更新都会触发校验数据更新

当写不友好遇上写不友好……

Case 1： Differential RAID

问题：

- 成员盘故障集中爆发
  - 多个SSD会同期达到写入次数极限，发生不可恢复的bit err
- 数据不可恢复
  - RAID5在修复过程中再次发生SSD故障的概率高
- 不仅仅RAID5才会出现
  - RAID1、10、6同样会出现

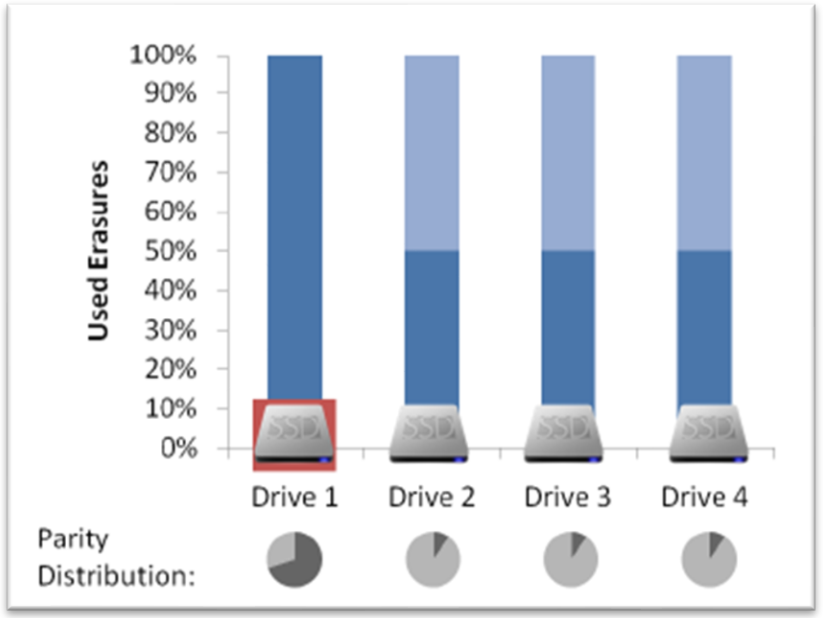


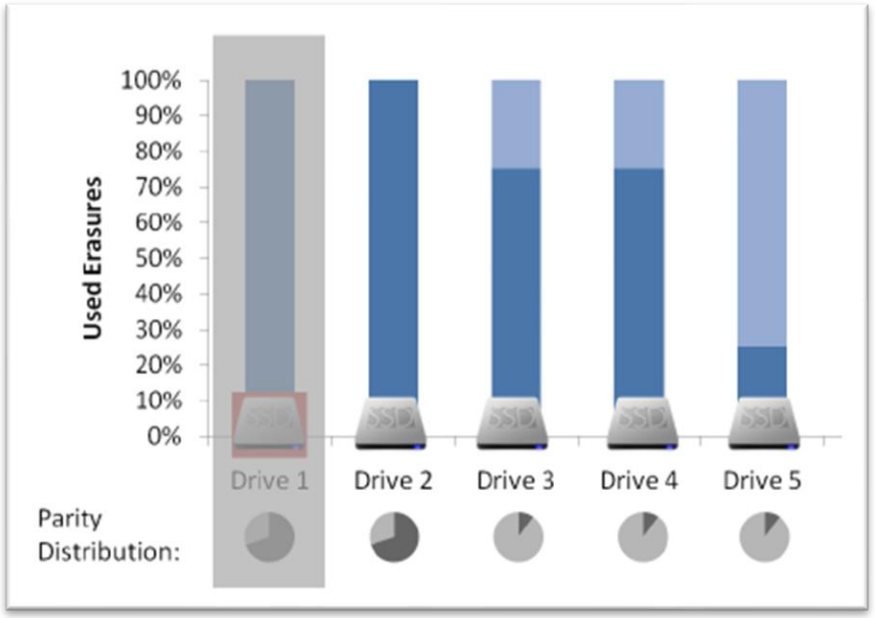
Case 1： Differential RAID

Solution： Differential RAID

- 打破平衡，让校验信息非均匀分布，造成寿命差异

在整个阵列中不均匀地分布奇偶校验块，利用它们较高的更新率控制不同设备的老化速率。





为了在旧设备被新设备替换时保持这种年龄差异，每次驱动器更换时重新调整奇偶校验分布。

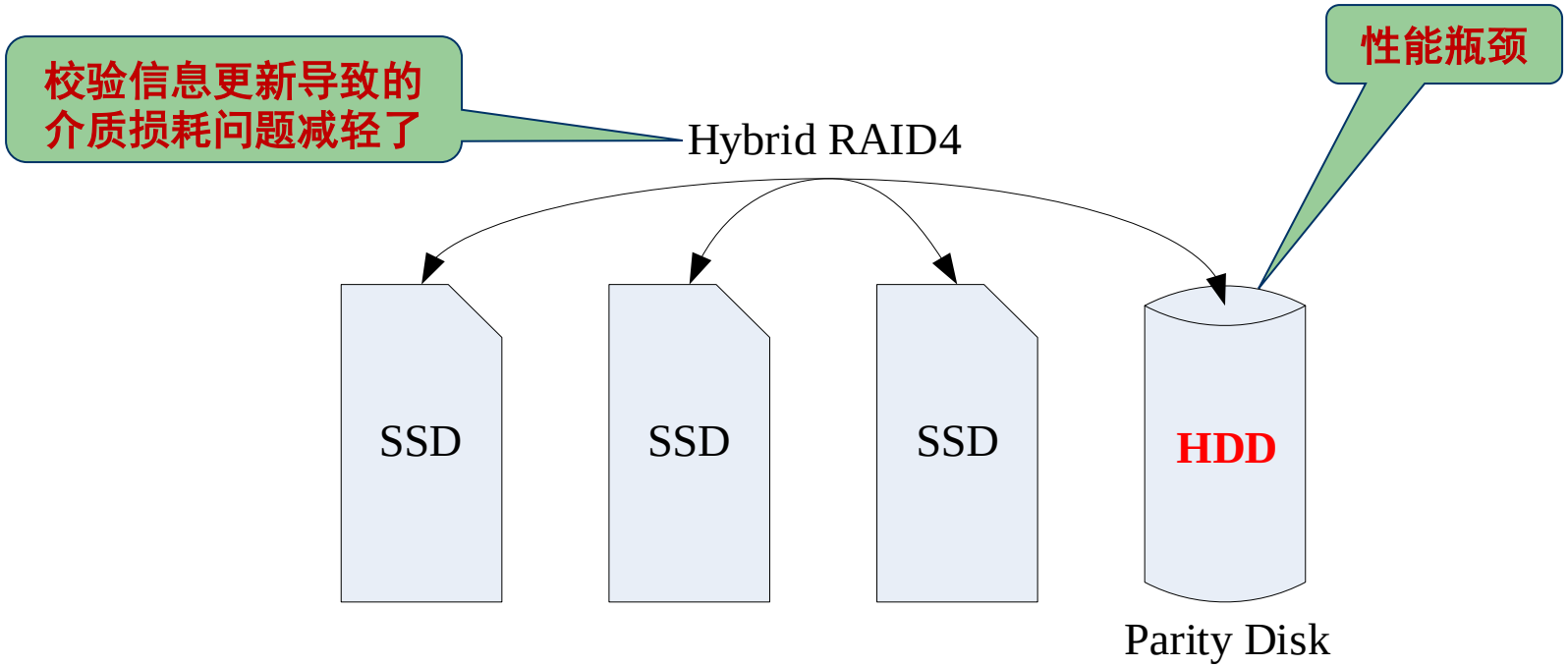
Mahesh Balakrishnan, Asim Kadav. Differential RAID: Rethinking RAID for SSD Reliability. ACM Transactions on Storage, Volume 6 Issue 2, July 2010.

Case2：混合式盘阵列

- 磁盘(HDD)
  - 高能耗
  - 非对称的顺序/随机性能
  - 介质损耗可忽略

- 固态盘(SSD)
  - 低能耗
  - 非对称读/写性能
  - 介质损耗问题

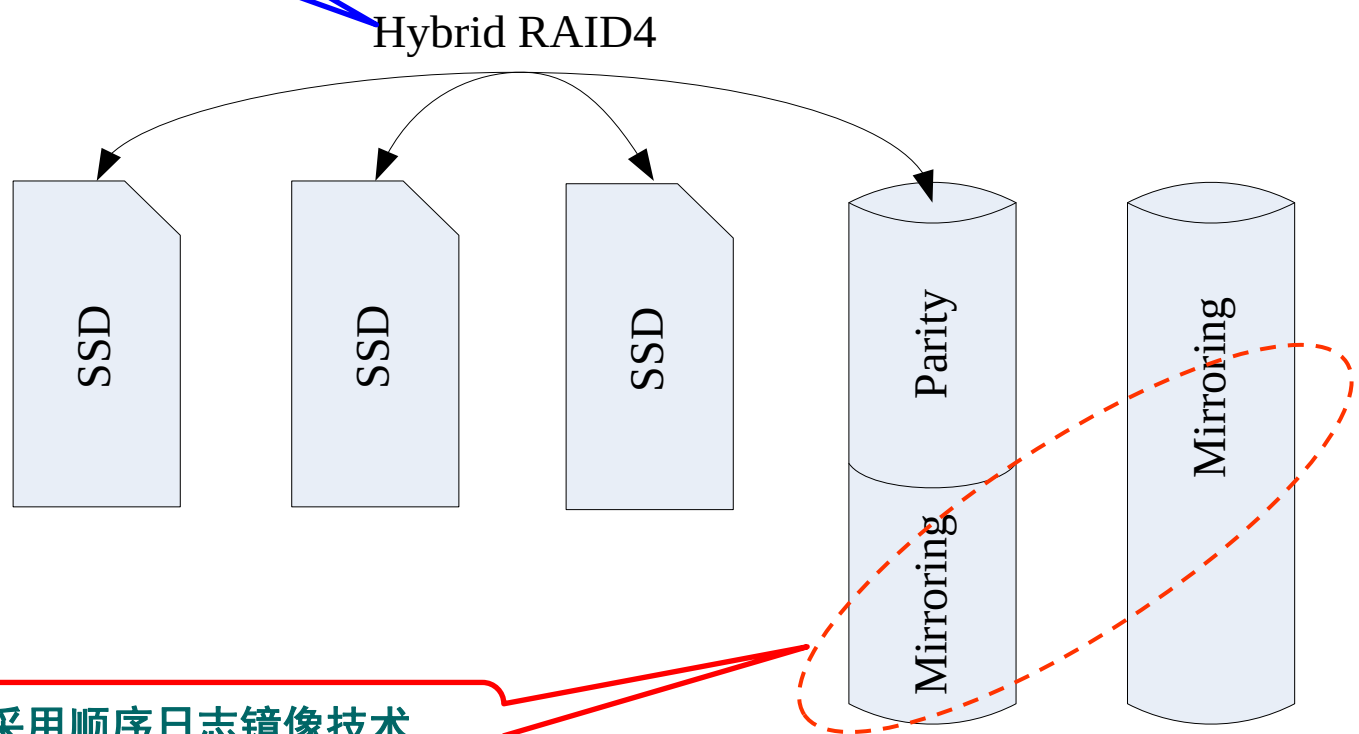
混合式盘阵列



- 随机小写性能问题依旧存在
- 数据恢复过程更为漫长（读取磁盘）

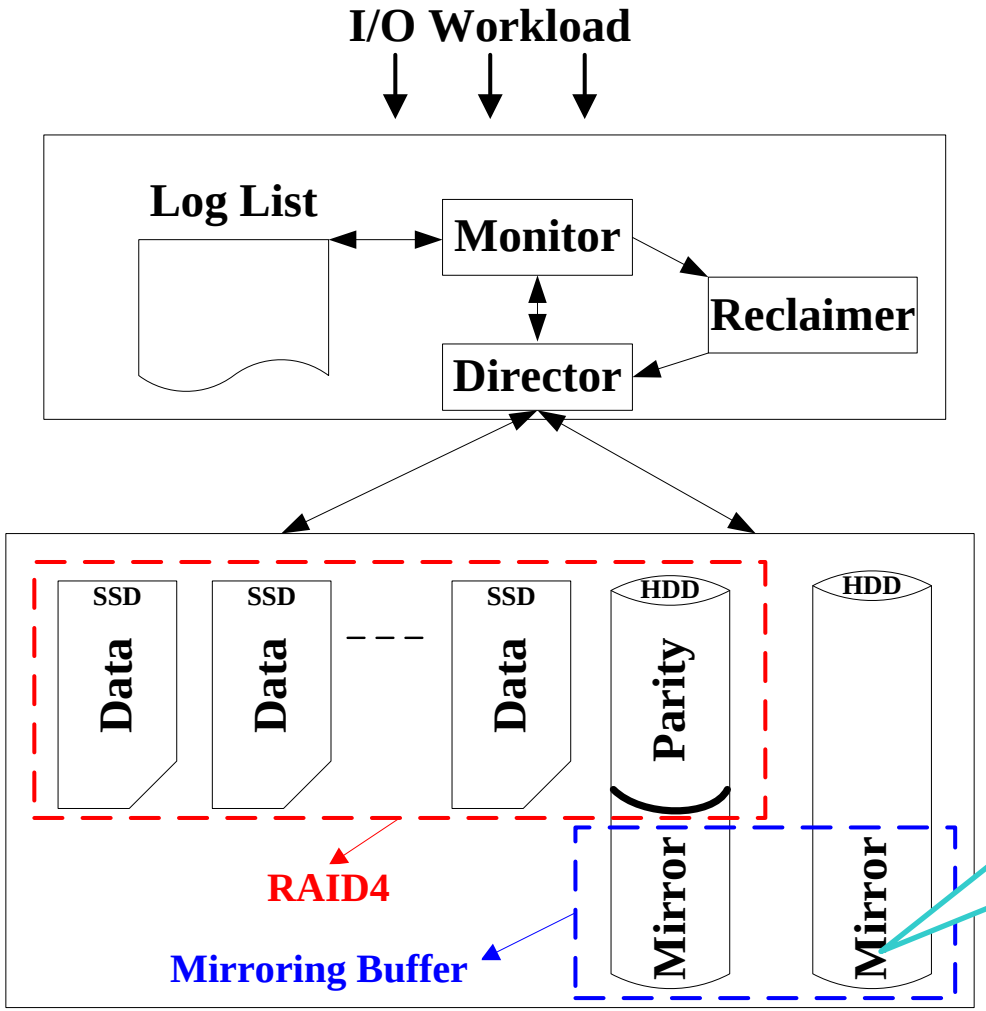
混合式盘阵列

校验信息更新导致的
介质损耗问题减轻了



由于采用顺序日志镜像技术，
减轻随机小写性能问题

混合式盘阵列：HPDA



暂时缓存来自用户的随机写请求数据，并在数据恢复过程中充当代理空间加快盘阵列重建

Bo Mao, et al. HPDA: A Hybrid Parity-based Disk Array for Enhanced Performance and Reliability. In the Proc. of the 24th International Parallel and Distributed Processing Symposium (IPDPS), 2010.

HPDA优点

校验块数据为热点更新数据，导致介质损耗问题更为严重

小写请求性能差伴随着校验更新导致小写性能更加严重

级别	可靠性	性能	价格
RAID0	低 （无冗余）	中 （小写性能问题）	低
RAID1/10	高 （双冗余）	低 （小写性能问题）	高（双冗余）
RAID5/6	低 （校验信息频繁更新）	低 （小写性能更加严重）	低
HPDA	高 （镜像和校验保护）	高 （基于日志的缓存）	低

Case 3： I-CASH

- **SSD：**
  - High performance for read, especially for random read
  - Small random write means low performance and wearing
- **HDD**
  - Good performance for sequential read and write
  - Very low performance for random read and write
  - Almost no wearing

	Random read	Sequent read	Random write	Sequent write	wearing
SSD	High	Very high	Low	High	Yes
HDD	Low	High	Low	High	No

Insight

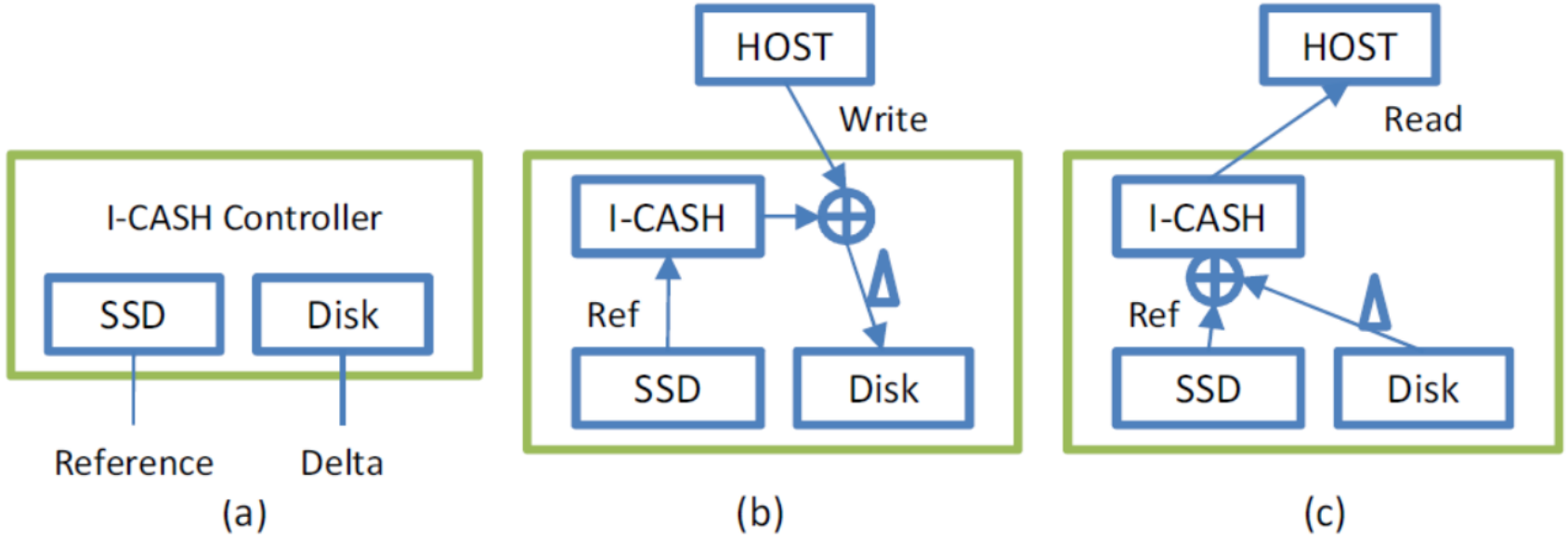
Content Locality：

Recent research literature has reported strong content locality in many data intensive applications with only 5% to 20% of bits inside a data block being actually changed on a typical block update operation

A Case of Hybrid RAID: SSD + HDD

I-CASH: Intelligently Coupled Array of SSD and HDD

- SSD stores seldom-changed and mostly read reference data blocks
- HDD stores a log of deltas between currently accessed I/O blocks and their corresponding reference blocks in the SSD



A Case of Hybrid RAID: SSD + HDD

I-CASH: Intelligently Coupled Array of SSD and HDD

- Random writes are not performed in SSD during online I/O operations
- High speed read performance of reference blocks stored in SSDs
- Potentially large number of small deltas packed in one delta block stored in HDD and cached in the RAM
- Exploit the fast read performance of SSDs and the high speed computation of modern multi-core CPUs to replace and substitute the mechanical operations of HDDs
- Avoid runtime SSD writes that are slow and wearing

以算代存

全闪存阵列

（AFAs，All-Flash Arrays）

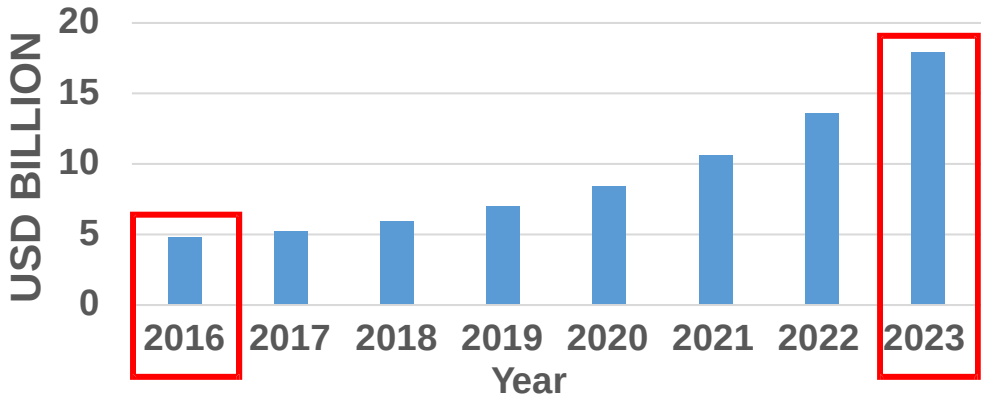
全闪阵列（ AFAs， All-Flash Arrays）

- 近年来应用广泛
- AFA 市场持续高速增长

Banks

Datacenters

Clouds



Data source: www.marketsandmarkets.com/Market-Reports/all-flash-array-market-41080938.html

DELL EMC VMAX

PureStorage FlashArray

SanDisk InfiniFlash

FUJITSU ETERNUS

NetApp AFF

全闪阵列研究聚焦

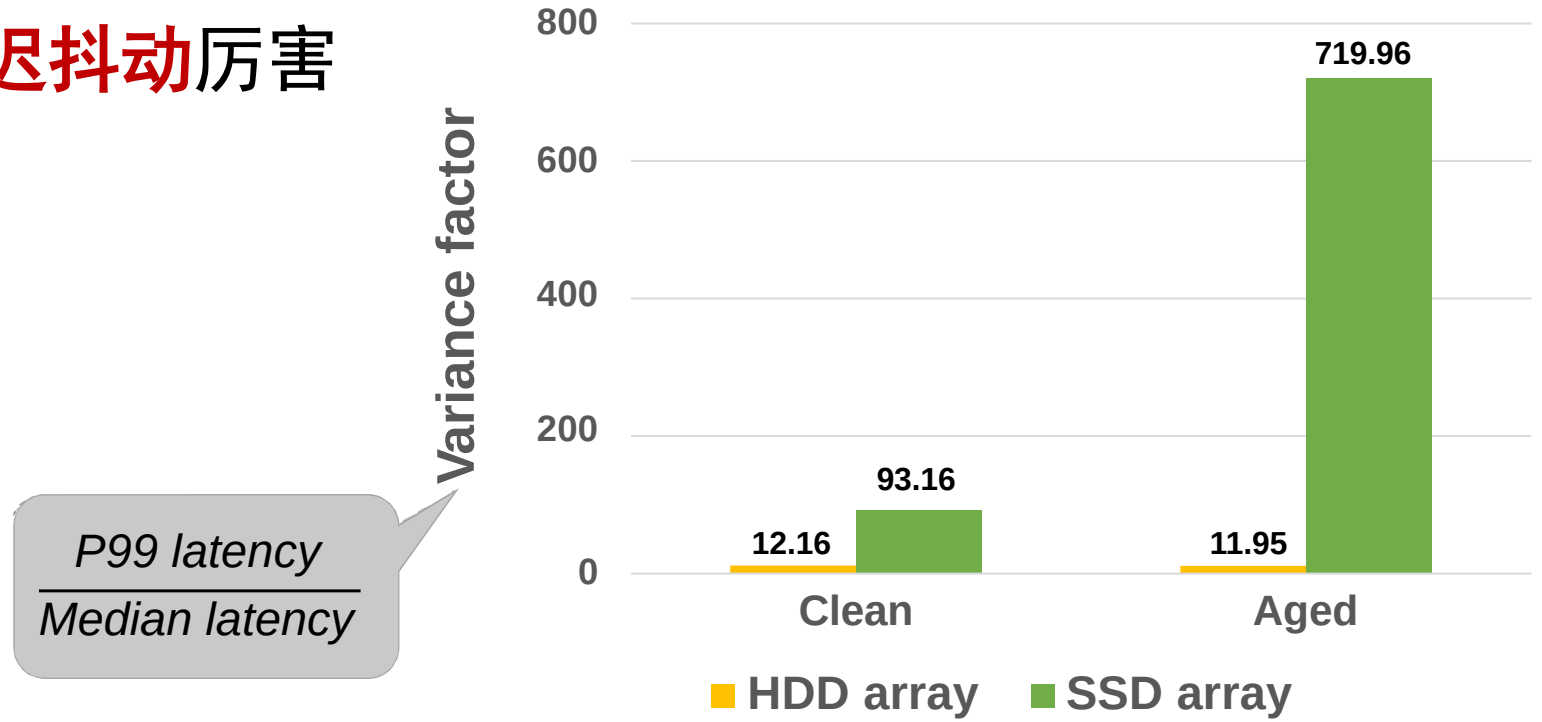
- 缓解尾部延迟（Tail-Latency）
- 提高性能（Performance）
- 改进可靠性（Availability）

Case 1：FusionRAID

Tianyang Jiang, Guangyan Zhang, et al,. **FusionRAID: Achieving Consistent Low Latency for Commodity SSD Arrays**. FAST 2021

SSD RAID 性能问题

- 与HDD RAID相比，**延迟抖动**厉害
  - 尾延迟突出
  - 随盘的老化问题加剧



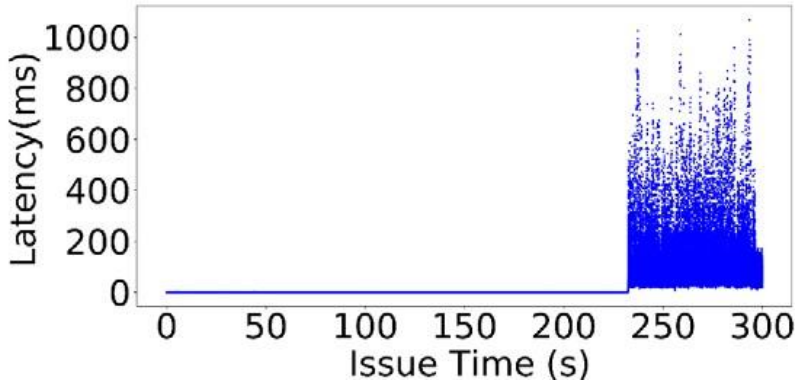
$$\frac{\textit{P99 latency}}{\textit{Median latency}}$$

	Median latency (ms)	Avg. latency (ms)	P99 latency (ms)	Variance factor
HDD RAID (clean)	68.67	134.37	835.35	12.16
HDD RAID (aged)	69.18	133.61	826.77	11.95
SSD RAID (clean)	0.275	3.57	25.62	93.16
SSD RAID (aged)	0.307	14.11	221.03	719.96

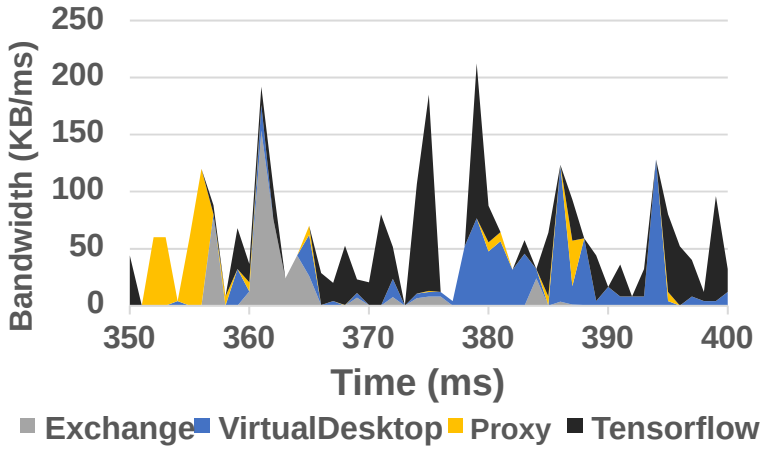
Table 1: Exchange latency, HDD vs. SSD RAID

实验观察：

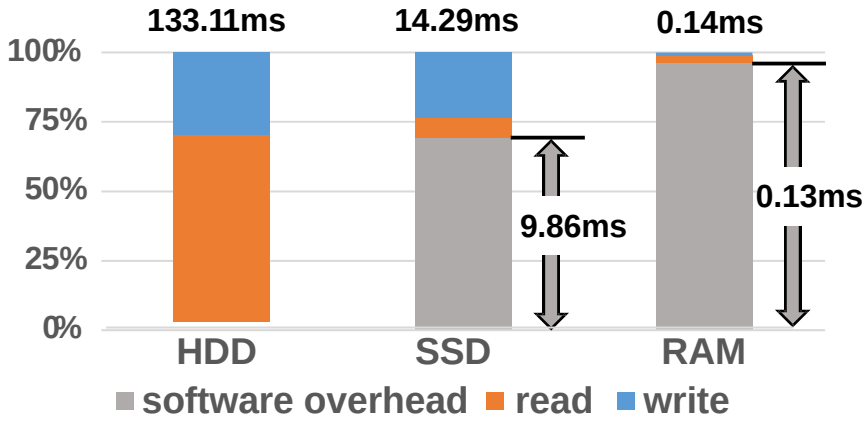
1. 工作负载通常不规则，交错突发（interleaving bursts）
   - All-for-all 映射模式优于物理分区
2. SSD RAID 写入操作**软件开销**严重
   - 相对开销比HDD RAID高，绝对开销远高于写RAM(Disk)
3. SSD性能异常常见，写尤其突出（幅度和持续时间）
   - 延迟峰值高且持久，足以在运行时感知



Datacenter SSDs with random writes



Bandwidth consumption in 4-workload mix

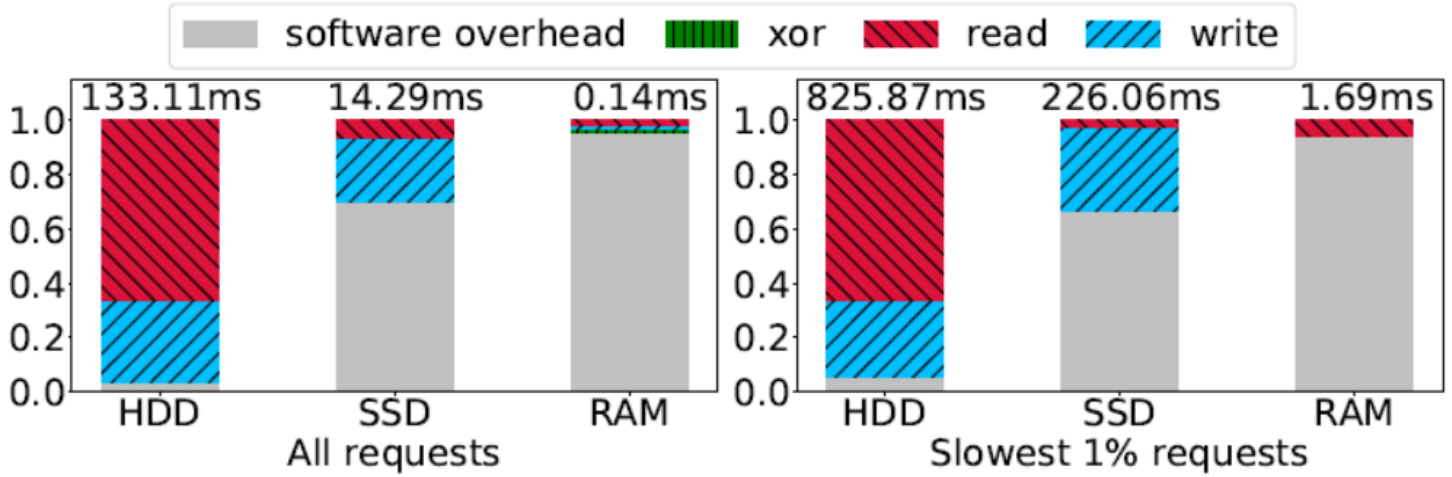


RAID write latency breakdown

Write Overhead in SSD RAID

1. 软件开销在HDD RAID写过程中对延迟的影响可以忽略不计
2. SSD RAID 写入操作**软件开销**是写盘时间的2. 9倍
   - 相对开销比HDD RAID高，绝对开销远高于写RAM（2个数量级）
   - 软件开销还使最慢的 1% 请求延迟增加为平均延迟的 10 倍
3. 在最慢的1% 请求中，SSD的写延迟对尾延迟贡献也很大，是平均写时延的20. 7倍

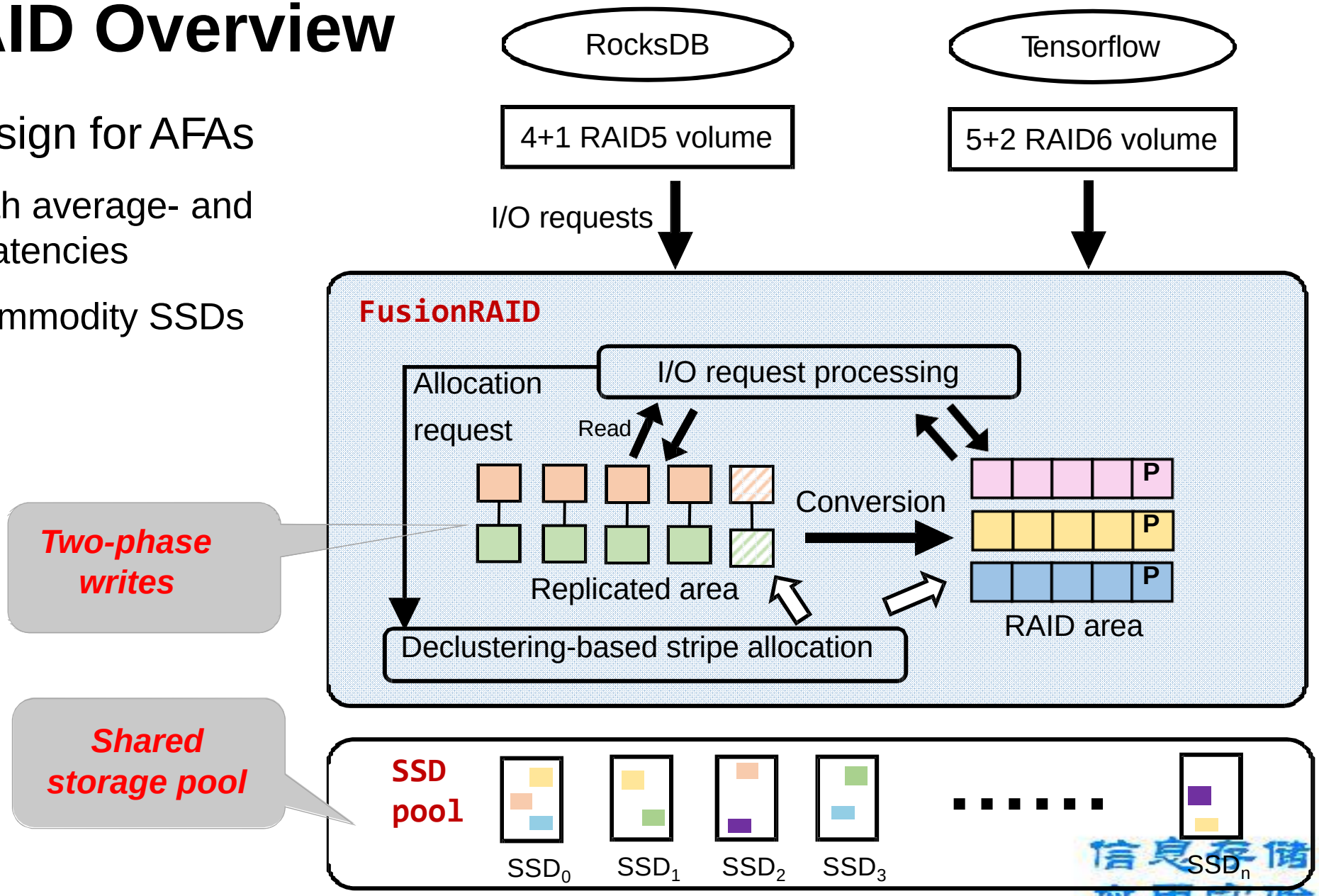
- 原因在于写同步
- 需要缩短写路径



缩短写操作路径，减少依赖关系，可大大降低SSD RAID访问延迟，包括平均延迟和最坏场景下的延迟.

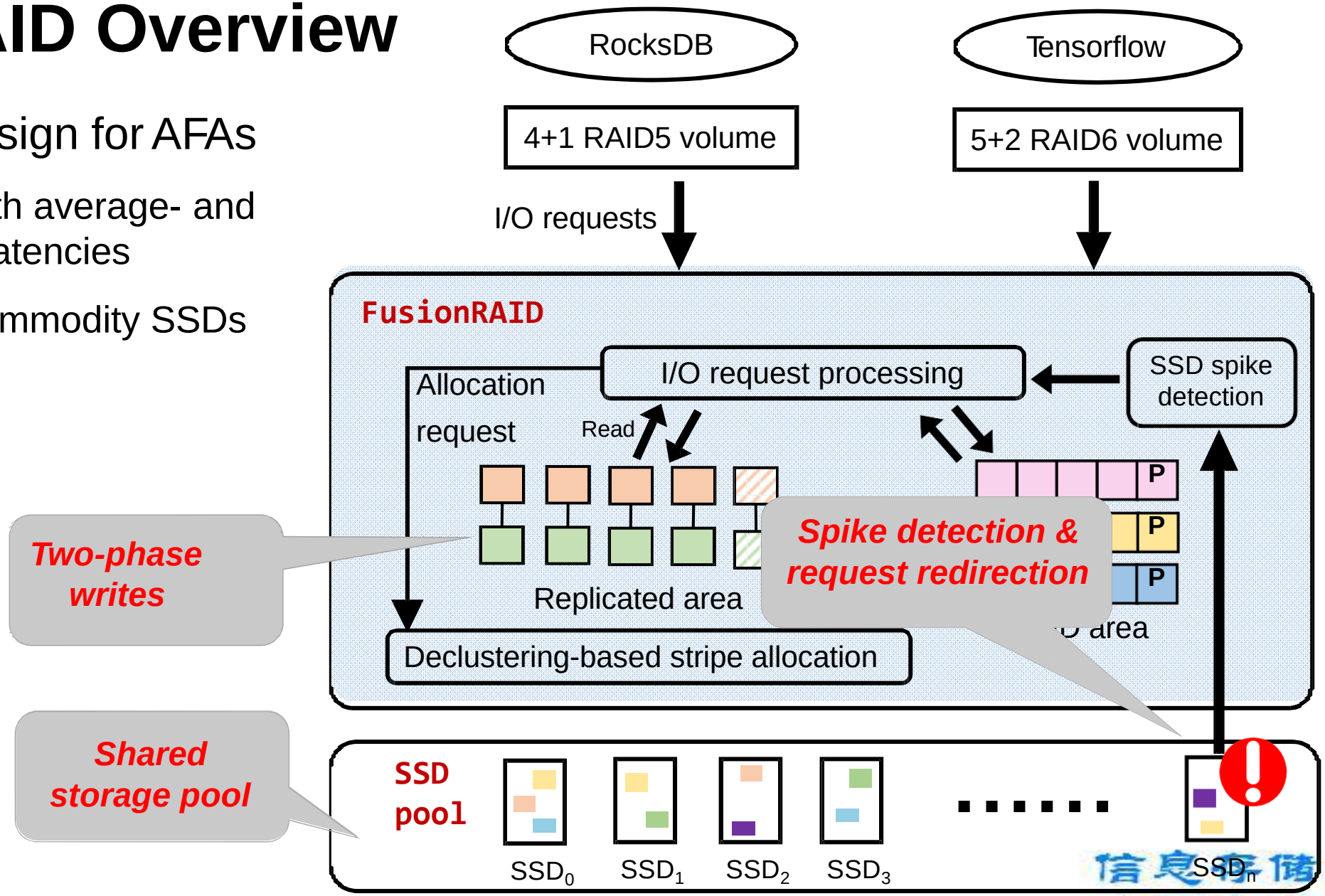
FusionRAID Overview

- New RAID design for AFAs
  - Reduces both average- and worst-case latencies
  - Works on commodity SSDs



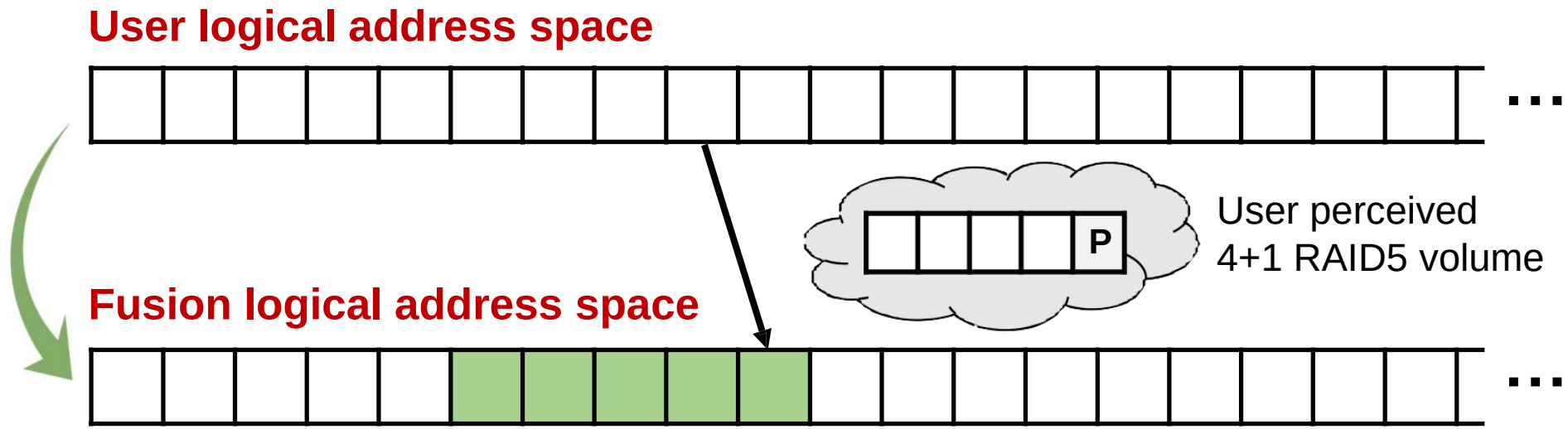
FusionRAID Overview

- New RAID design for AFAs
  - Reduces both average- and worst-case latencies
  - Works on commodity SSDs

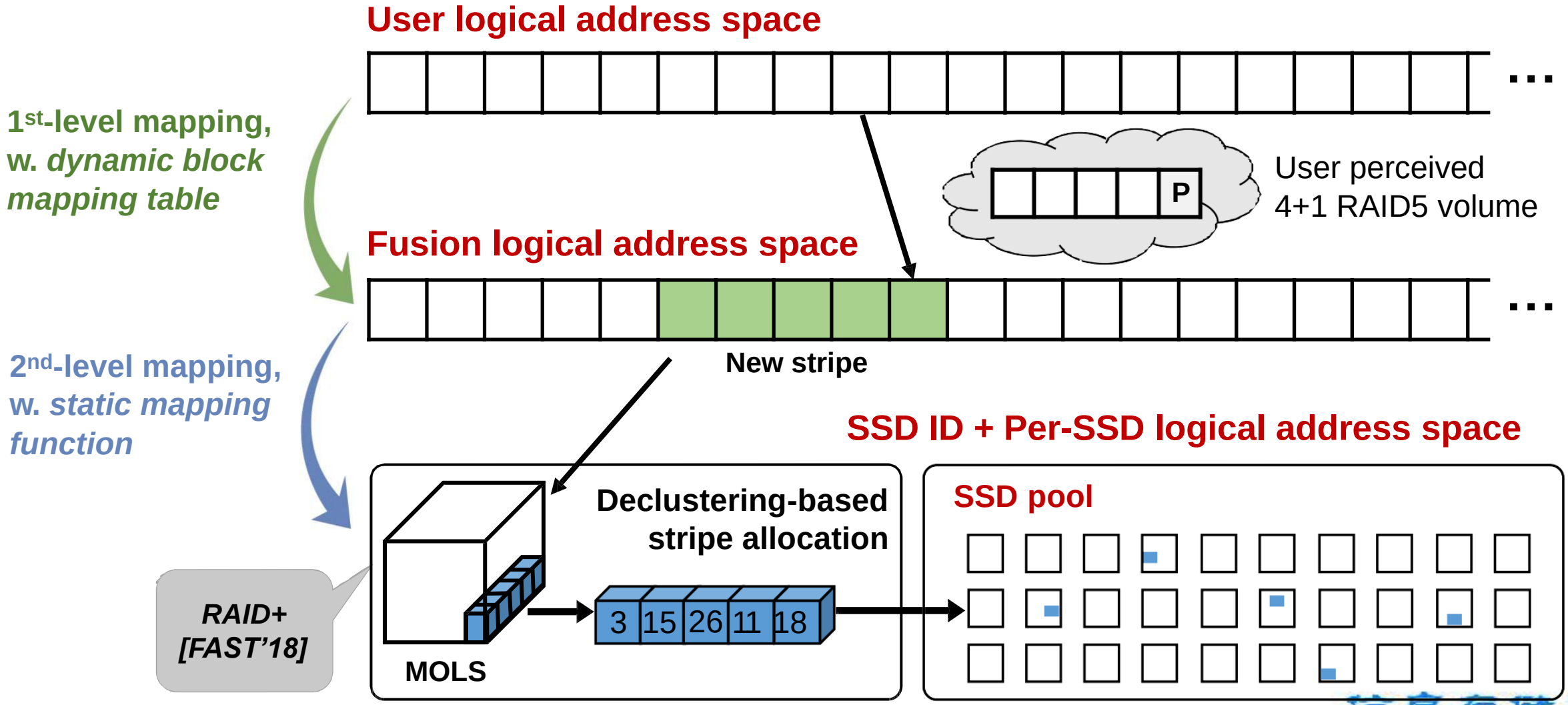


Shared Storage Pool

1st-level mapping, w. *dynamic block mapping table*



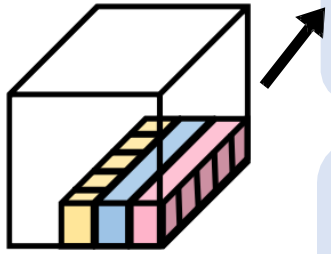
Shared Storage Pool



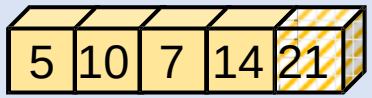
FusionRAID Optimized Writes

Large write request

(Direct RAID write)

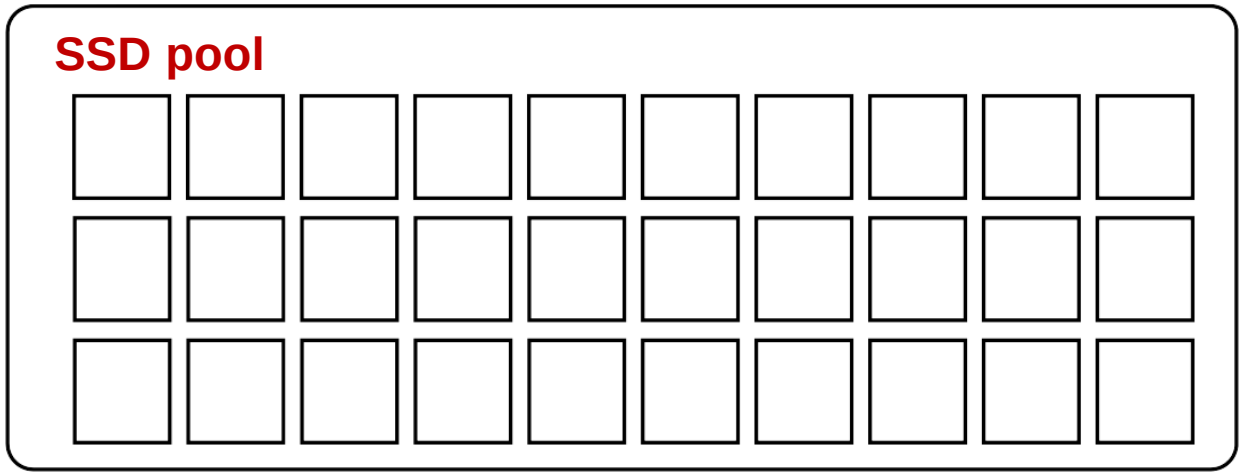


RAID storage



Replicated storage

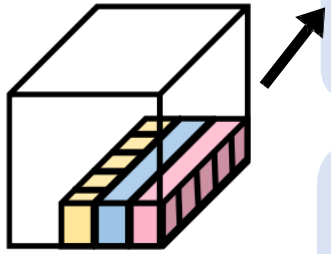
SSD pool



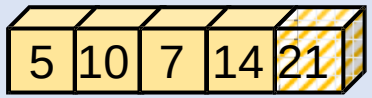
FusionRAID Optimized Writes

Large write request

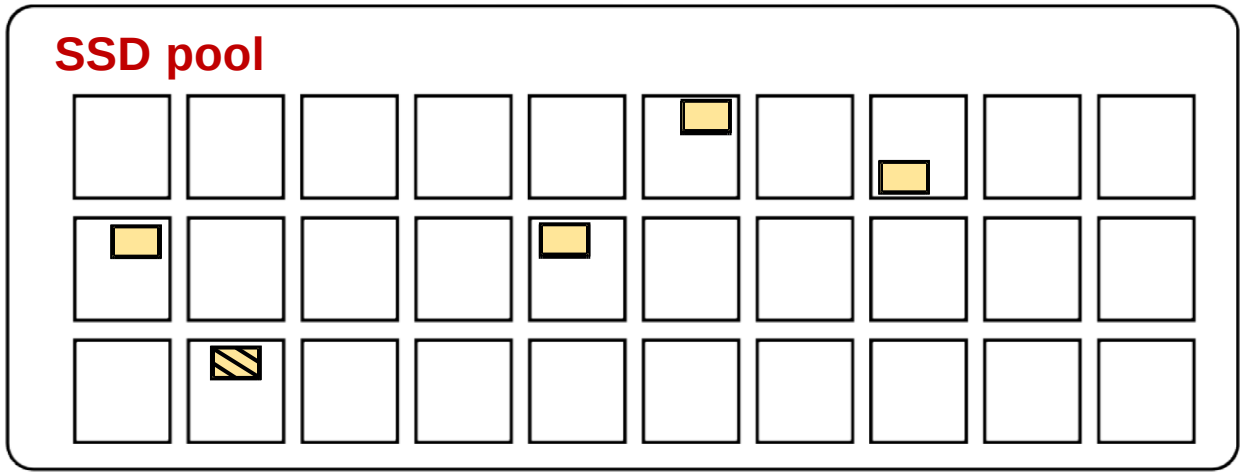
(Direct RAID write)



RAID storage



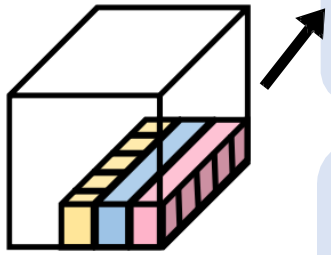
Replicated storage



FusionRAID Optimized Writes

Large write request

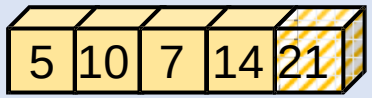
(Direct RAID write)



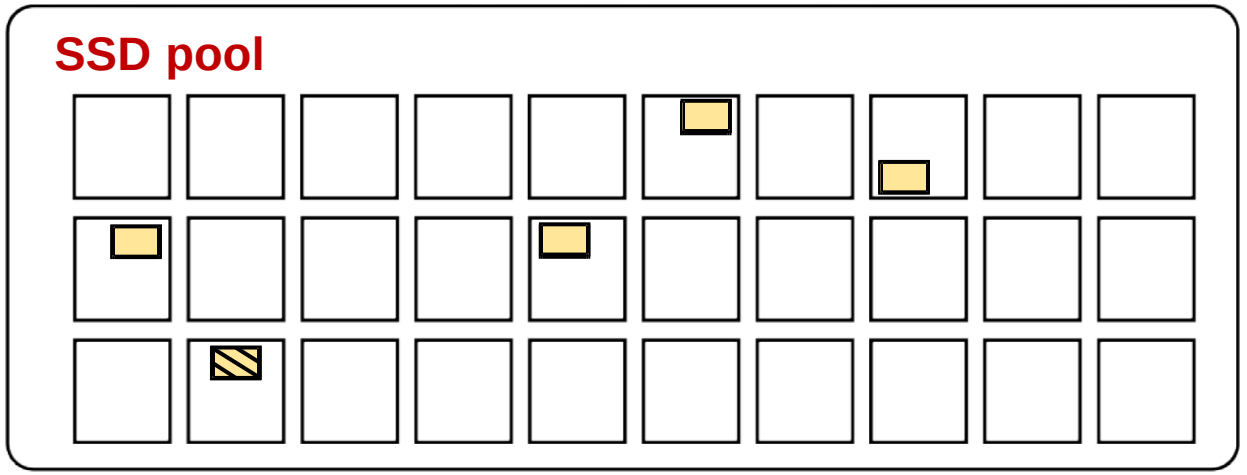
Small write request

(2-phase write)

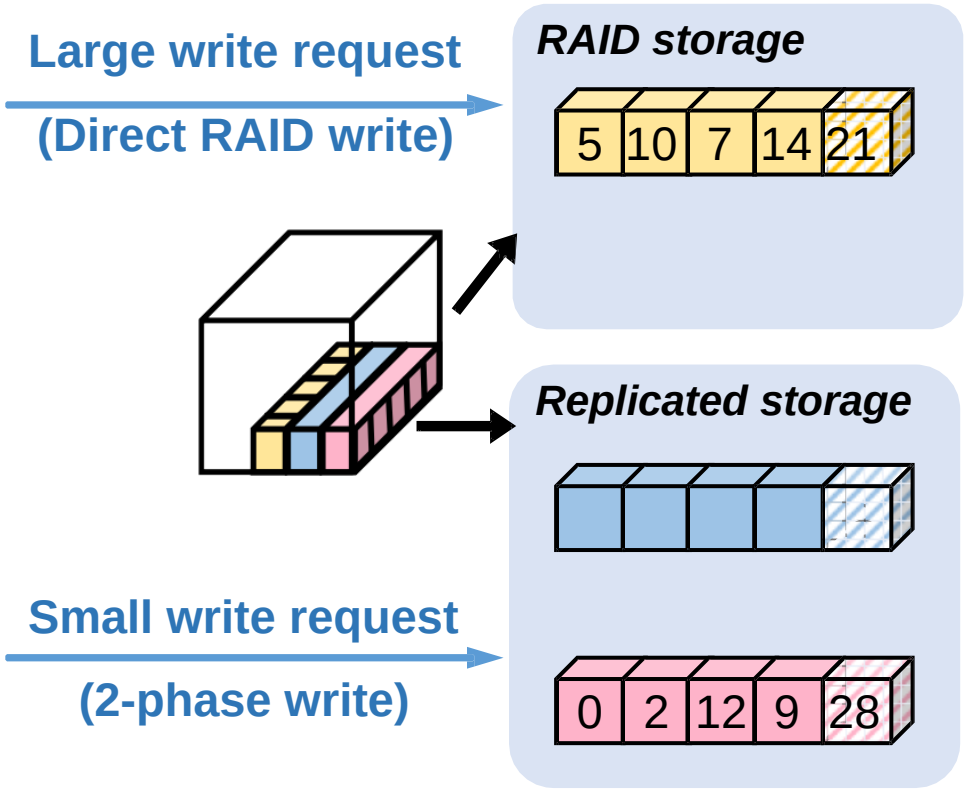
RAID storage

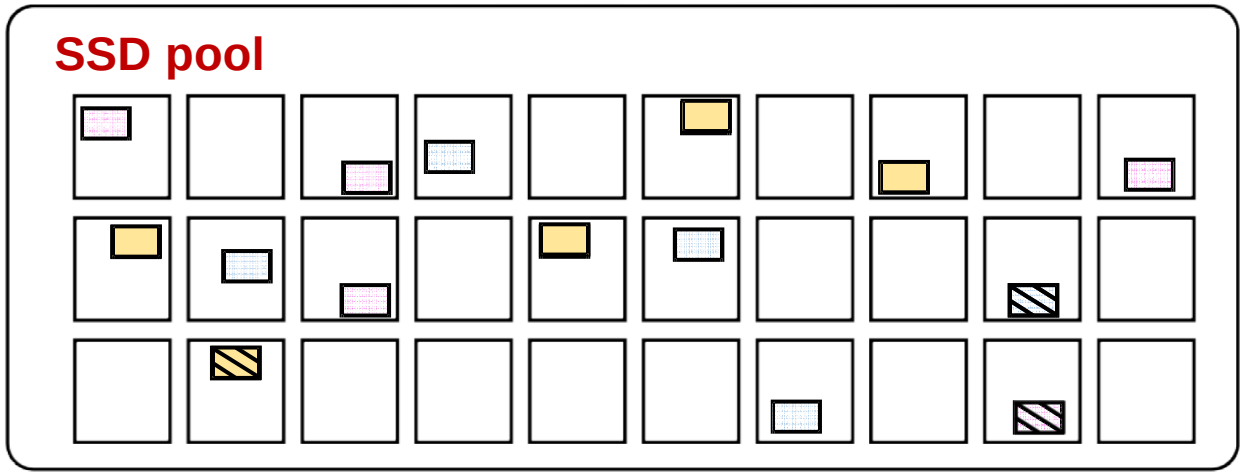


Replicated storage



FusionRAID Optimized Writes

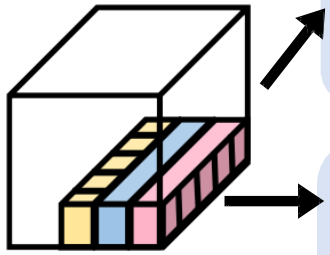




FusionRAID Optimized Writes

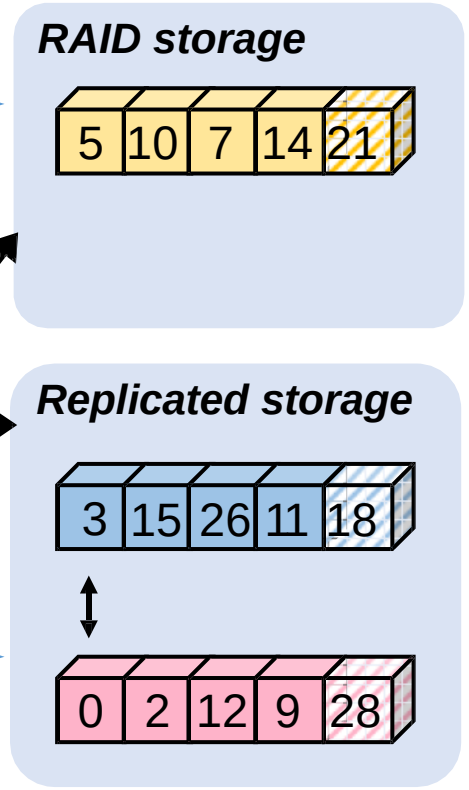
Large write request

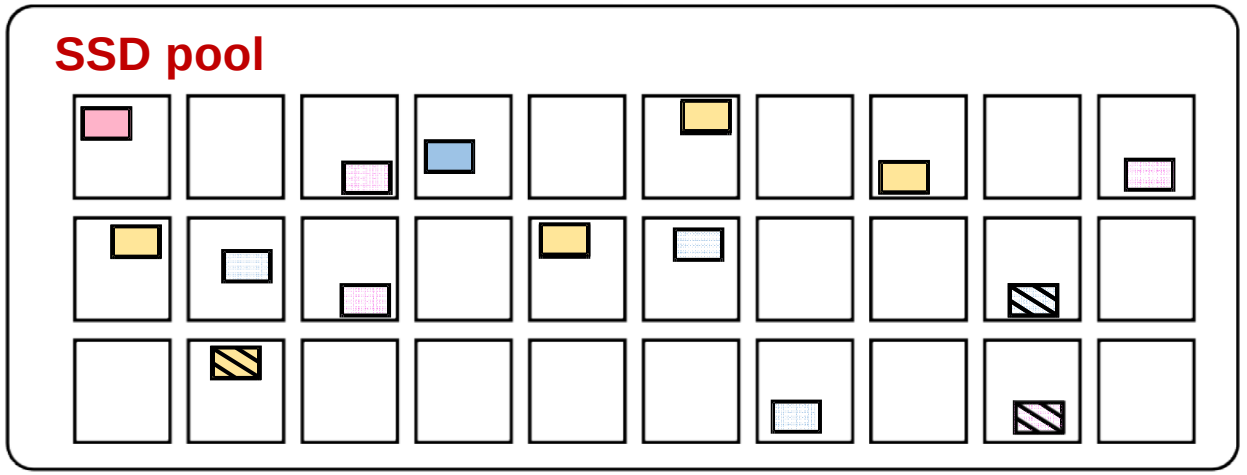
(Direct RAID write)



Small write request

(2-phase write)

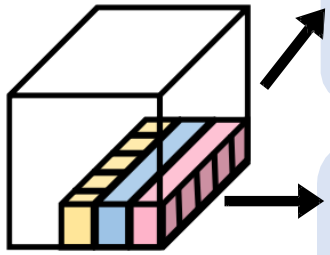




FusionRAID Optimized Writes

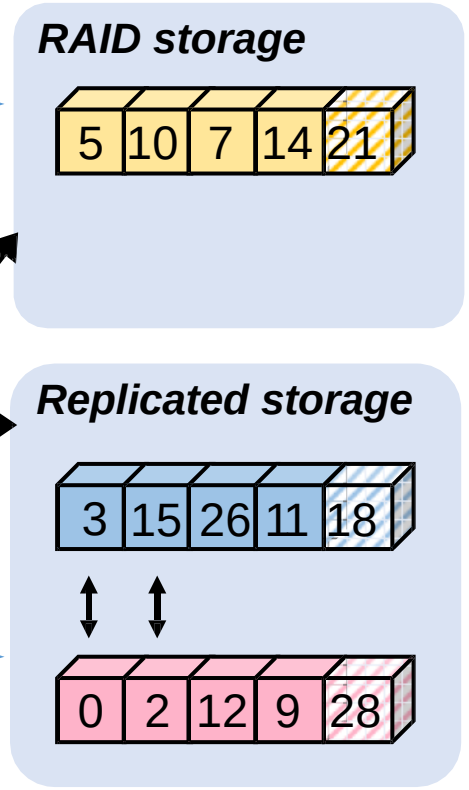
Large write request

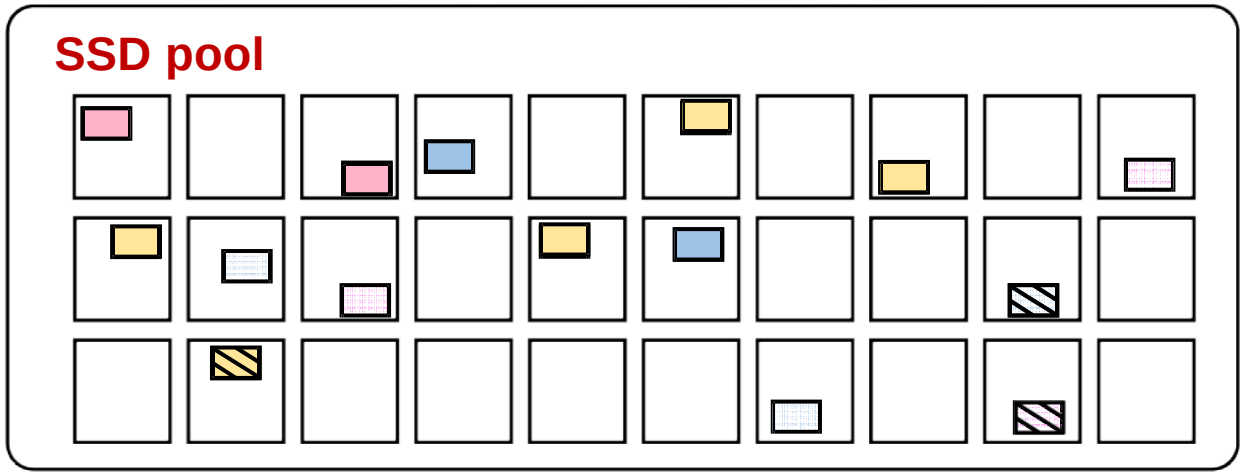
(Direct RAID write)



Small write request

(2-phase write)

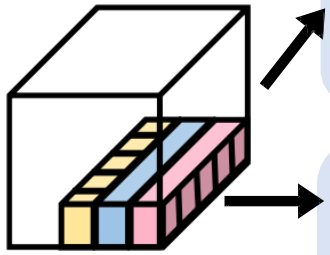




FusionRAID Optimized Writes

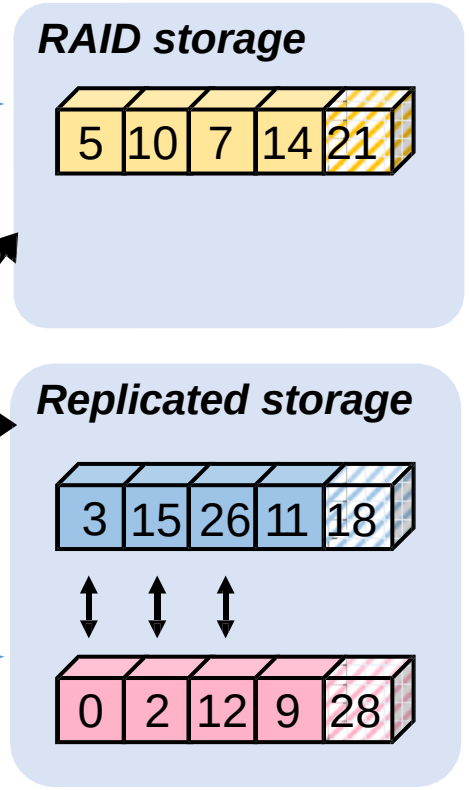
Large write request

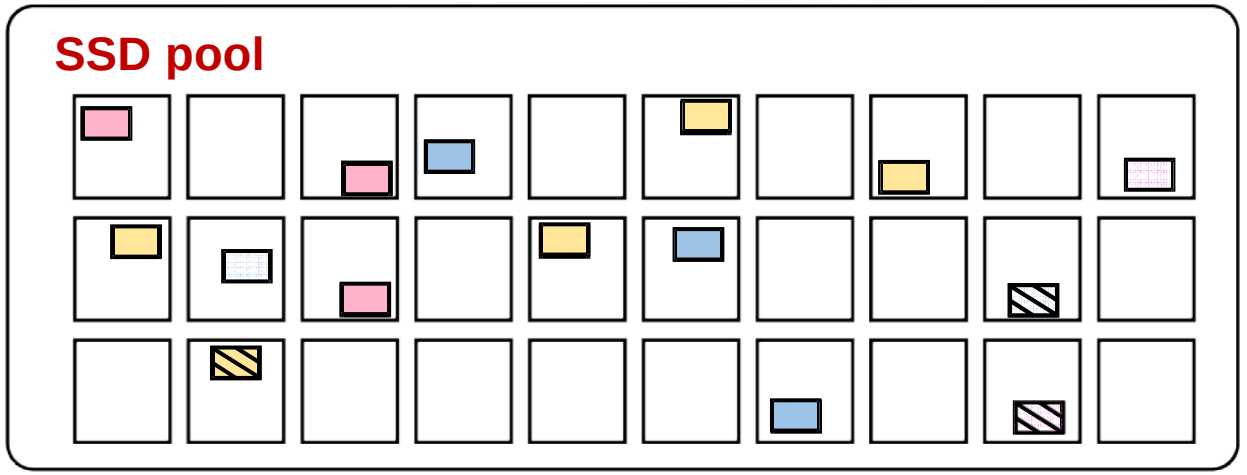
(Direct RAID write)



Small write request

(2-phase write)

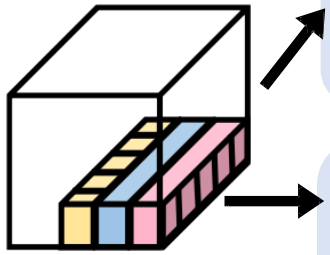




FusionRAID Optimized Writes

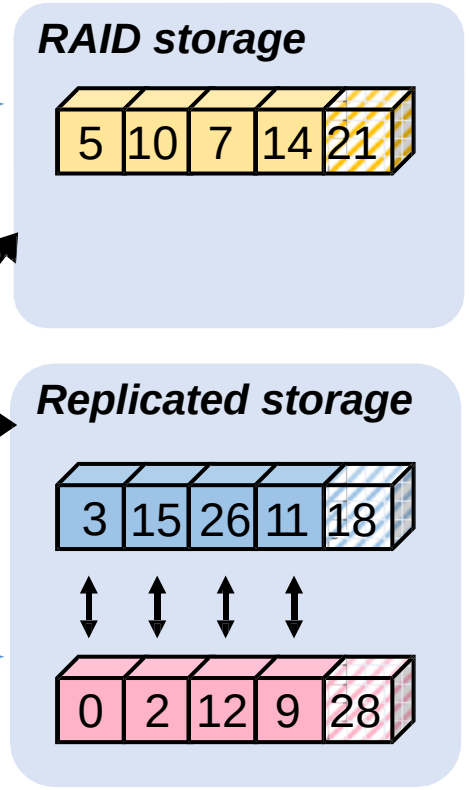
Large write request

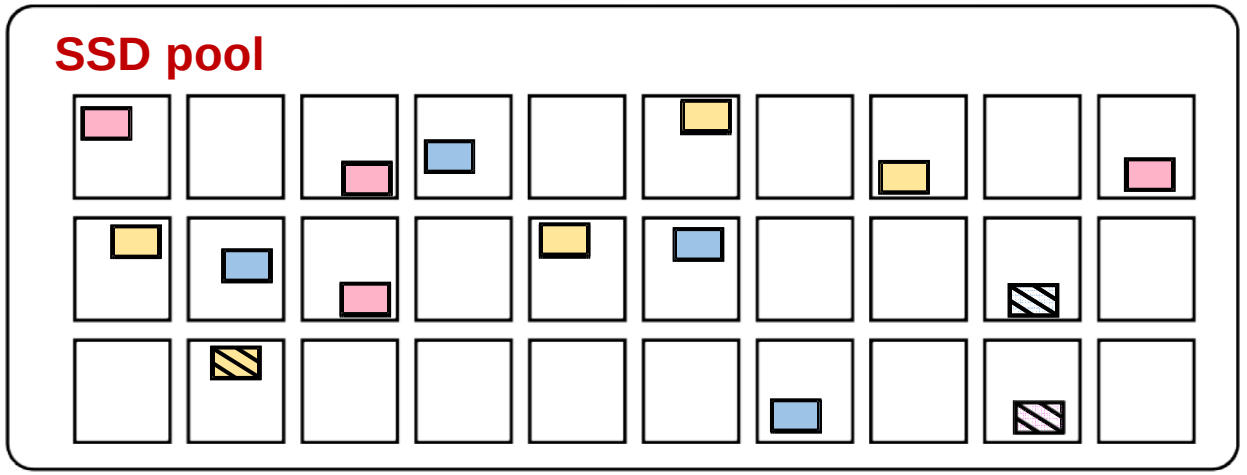
(Direct RAID write)



Small write request

(2-phase write)

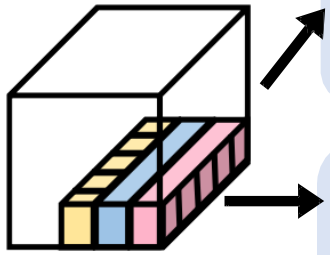




FusionRAID Optimized Writes

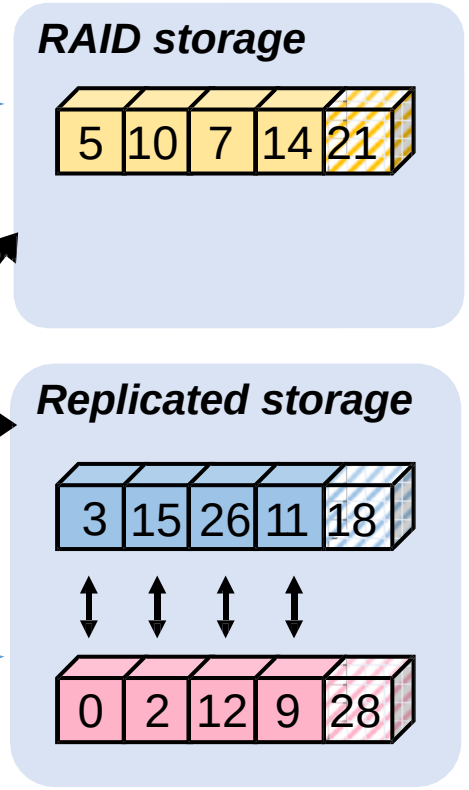
Large write request

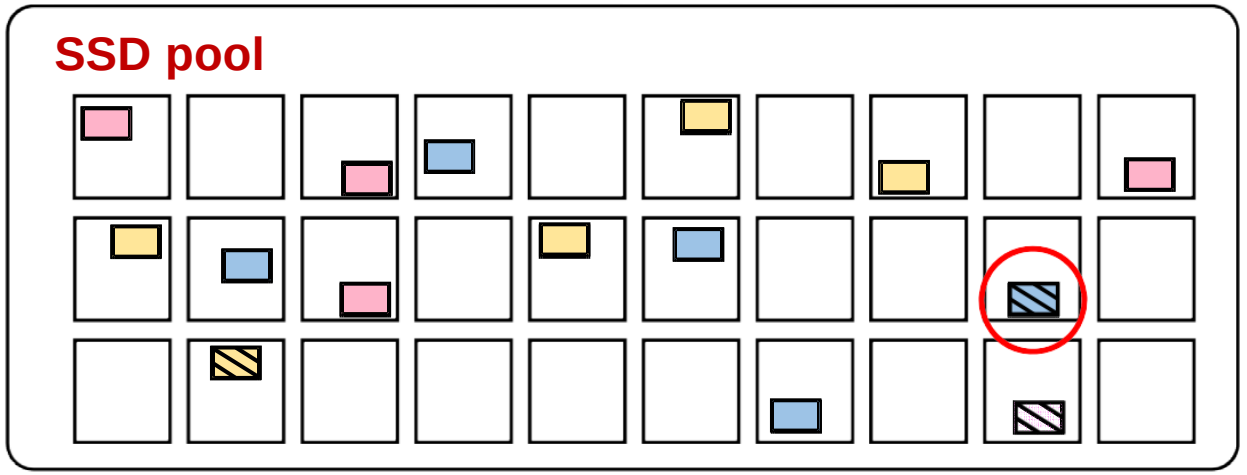
(Direct RAID write)



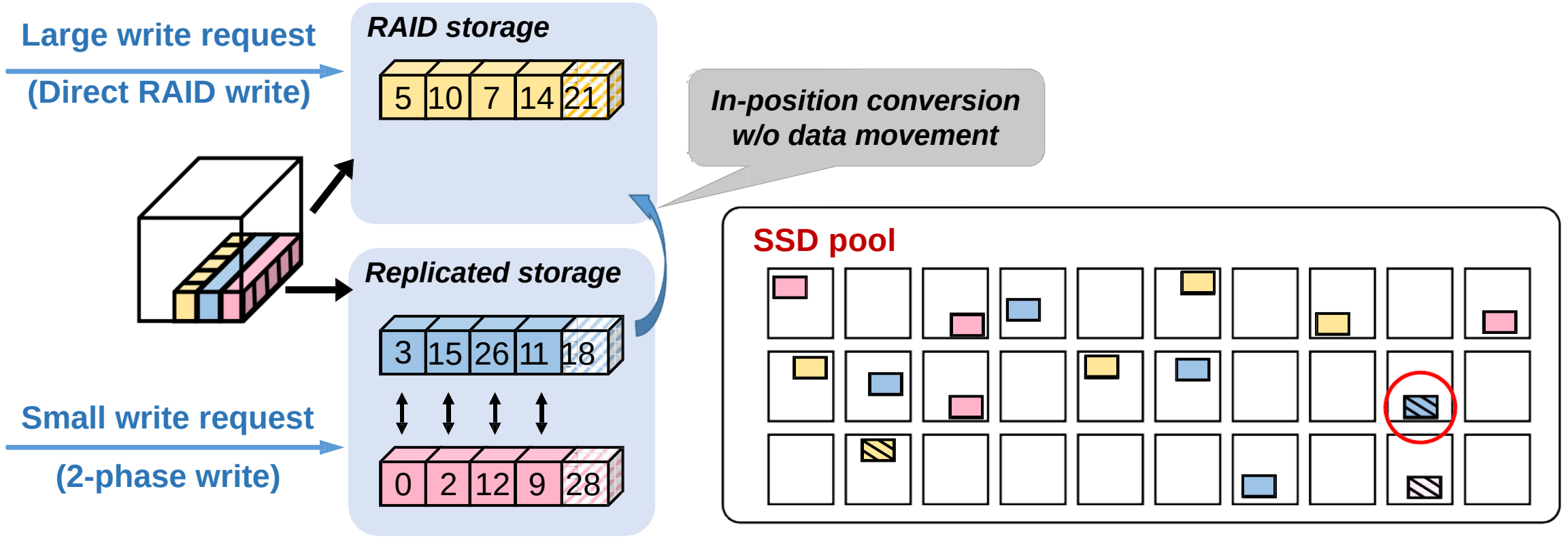
Small write request

(2-phase write)

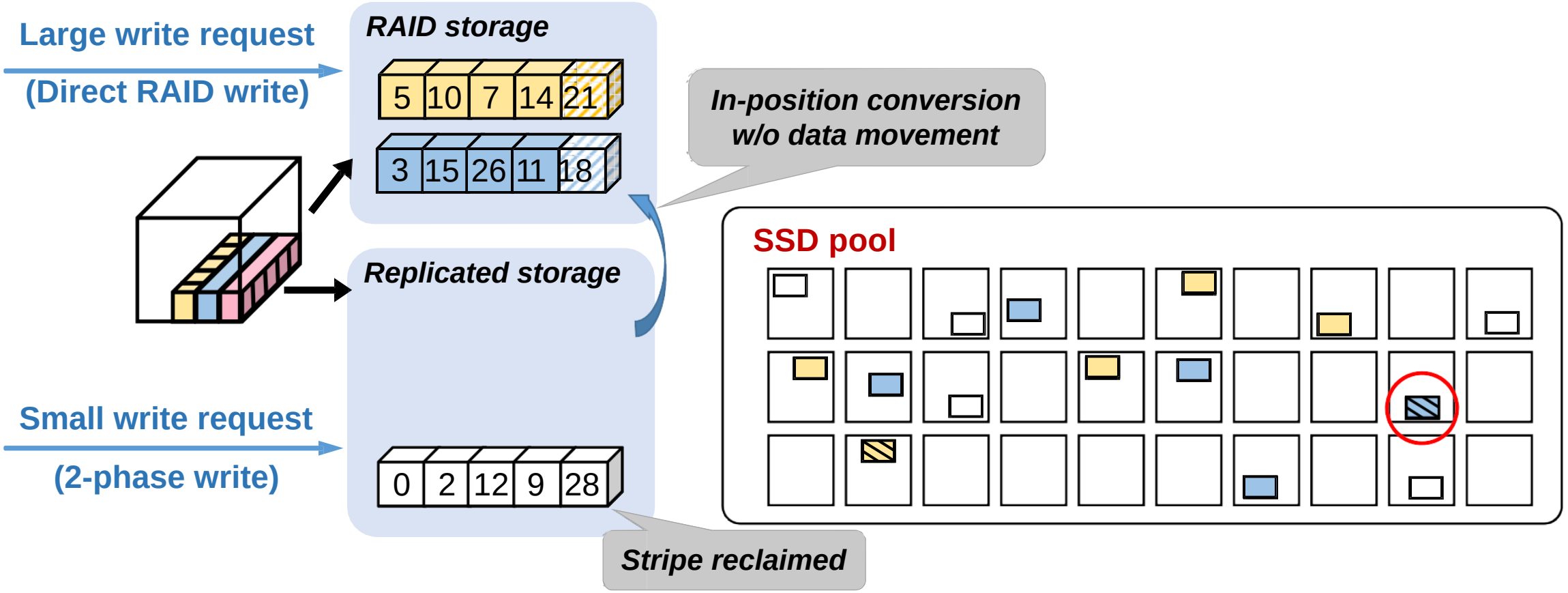




FusionRAID Optimized Writes



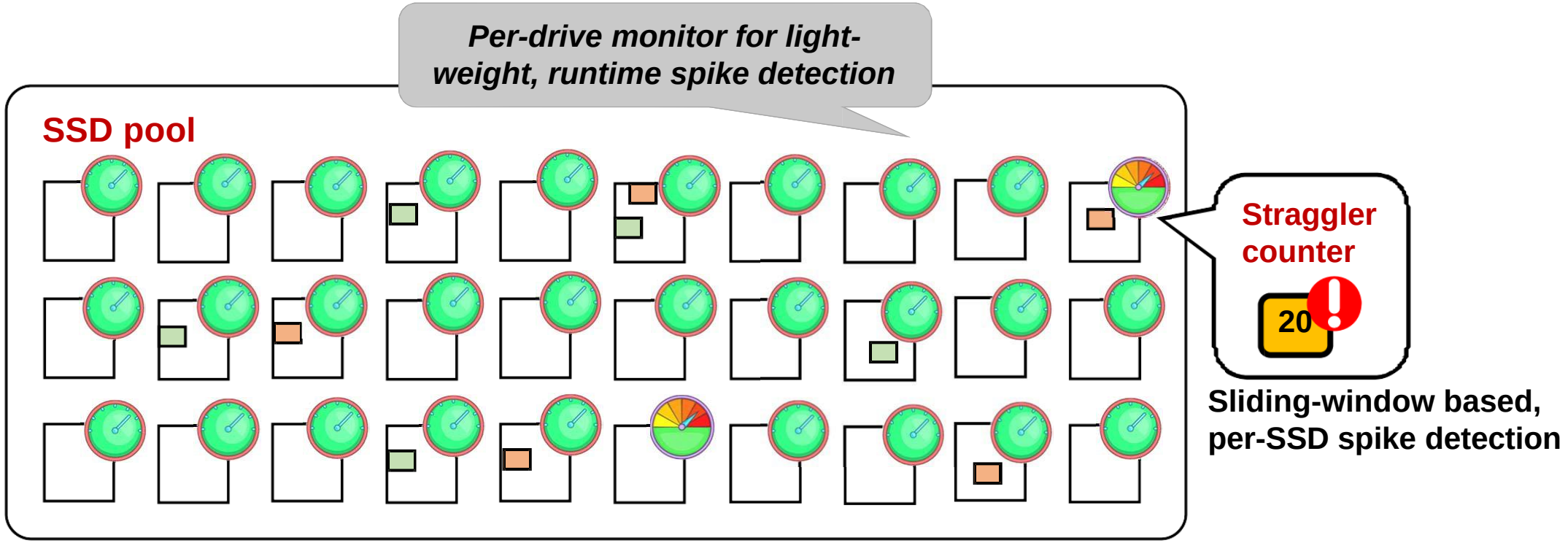
FusionRAID Optimized Writes



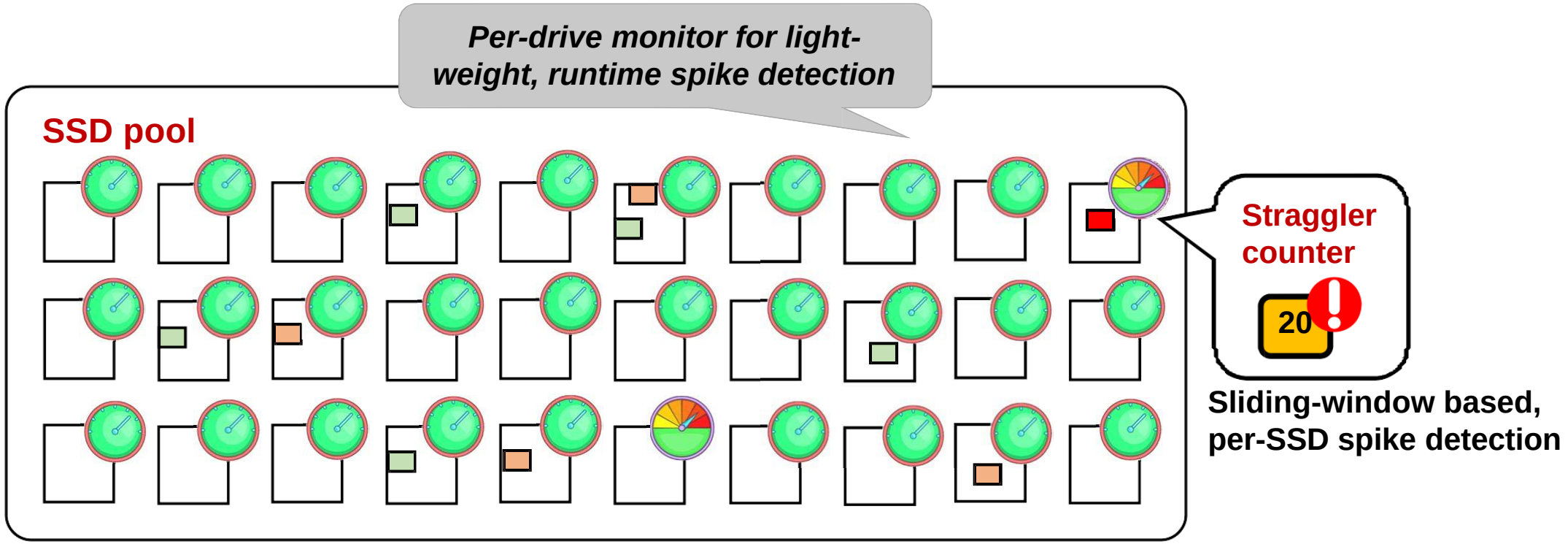
Extened Reading:

Guangyan Zhang, Zhufan Wang, et al. 2019. Determining Data Distribution for Large Disk Enclosures with 3-D Data Templates. ACM Trans. Storage 15, 4, Article 27 (December 2019),

Spike Detection and Request Redirection



Spike Detection and Request Redirection



Reactive request redirection

Case 2：Asym-RAID(非对称RAID)

Asymmetric RAID: Rethinking RAID for SSD Heterogeneity

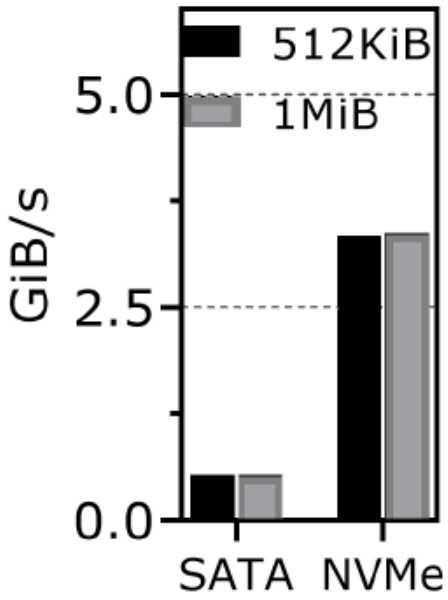
Authors: Ziyang Jiao, Bryan S. Kim | Authors Info & Claims

HotStorage '24: Proceedings of the 16th ACM Workshop on Hot Topics in Storage and File Systems • Pages 101 - 107
https://doi.org/10.1145/3655038.3665952

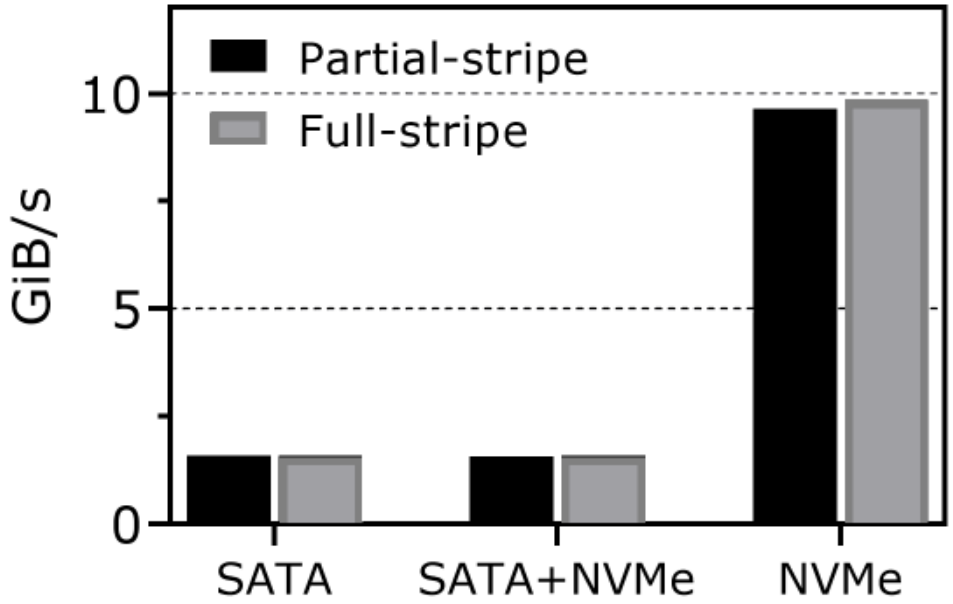
Published: 08 July 2024 Publication History

问题

- 传统RAID架构面对异质性SSD的性能瓶颈



(a) Performance comparison of SATA and NVMe SSDs under sequential read.

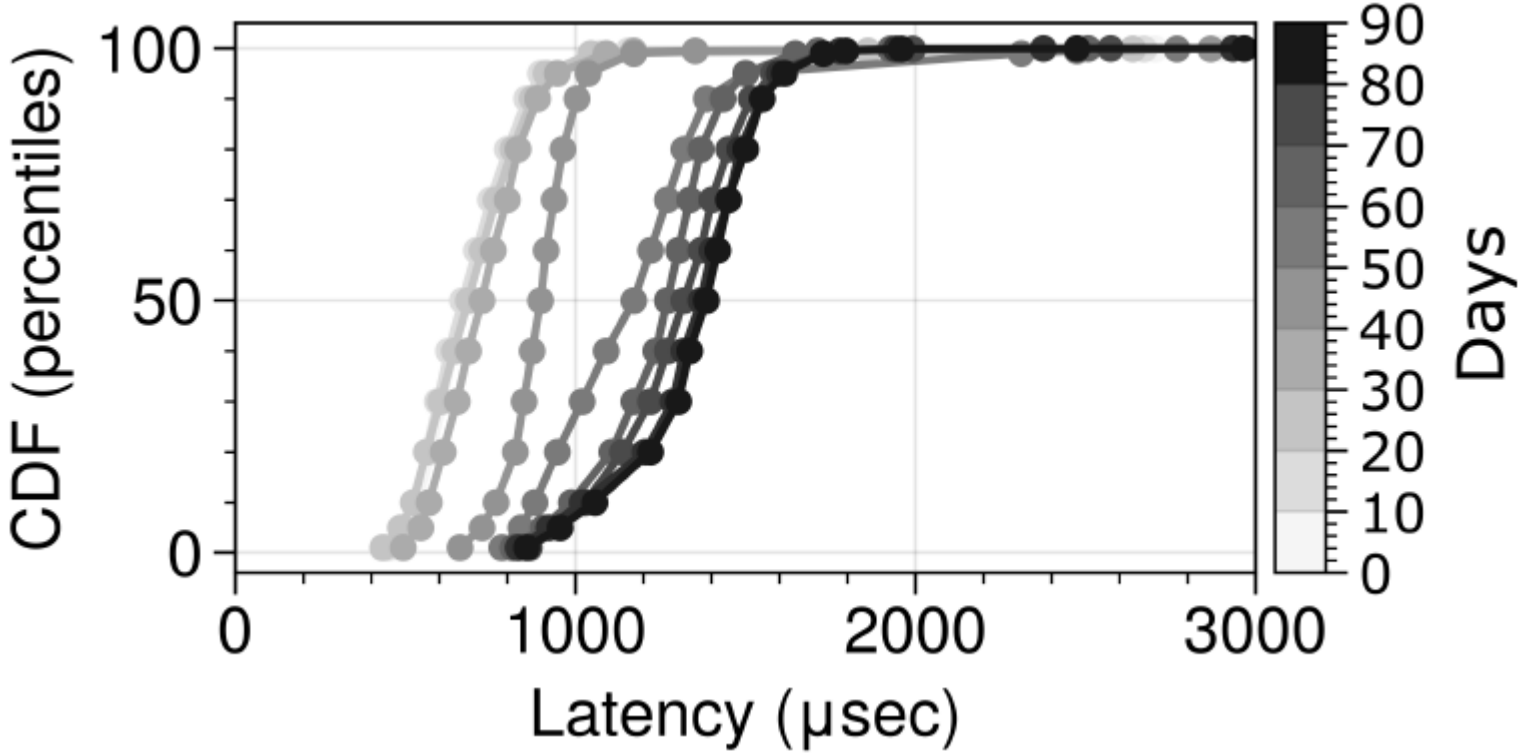


(b) Performance comparison of three RAID-5 systems. SATA (3 SATA SSDs) vs. SATA+NVMe (1 SATA SSD and 2 NVMe SSDs) vs. NVMe (3 NVMe SSDs).

木桶效应

问题

- 传统RAID架构面对异质性SSD的性能瓶颈



木桶效应

SSD无法在长时间内保持一致的性能，90天平均延迟增加了近1倍

问题

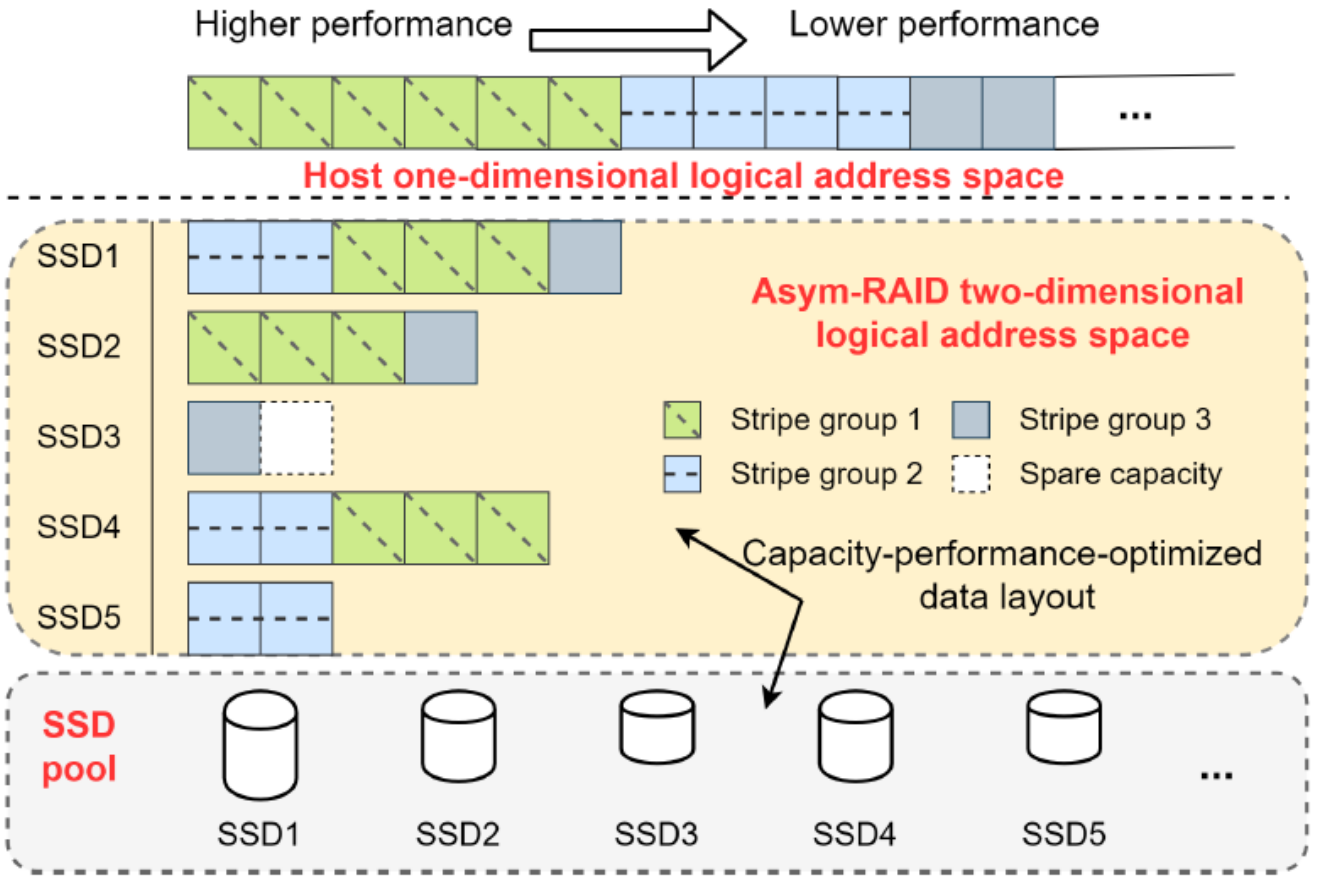
- 传统RAID架构面对异质性SSD的性能瓶颈

必须充分考虑SSD异构性对整个系统性能的影响

Asym-RAID

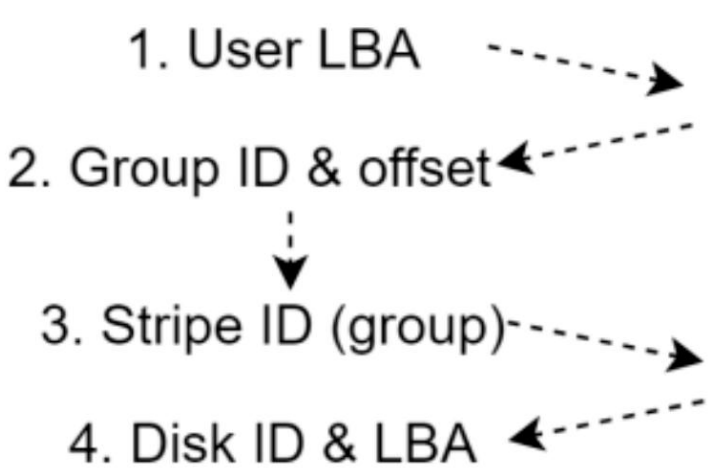
- 核心理念是异构性感知
- 不对称地分布数据，充分利用每个SSD的容量和性能

1. 最大化向主机导出的有效容量
2. 性能优化的放置



由5块物理盘构成2+1的RAID5示例

Asym-RAID



	Start LBA	Length	Disk components	Disk offsets
Stripe_group 1	0	6	(1, 2, 4)	(2, 0, 2)
Stripe_group 2	6	4	(1, 4, 5)	(0, 0, 0)
Stripe_group 3	10	2	(1, 2, 3)	(5, 3, 0)

地址映射表（初始化时完成）

Asym-RAID

Table 1. Comparison of existing approaches when managing heterogeneous SSDs in All-Flash Array storage.

	Disk heterogeneity	Device profiling reliance	Metadata overhead	Deployment cost	Disk utilization
Linux-MD [25]	Low	–	Low	Low	Low
SWAN (Log-RAID) [17]	Low	Low	High	Medium	Medium
IODA [19]	Low	Medium	Low	High	Medium
RAID+ [38]	Low	–	Low	Low	Medium
FusionRAID [8]	Low	High	Medium	High	Medium
StRAID [36]	Low	–	Low	Low	Medium
Diff-RAID [1]	Low	Low	Medium	Medium	Low
HeART [14]	Medium	Medium	High	High	Low
Pacemaker [13]	Medium	High	Medium	High	Low
Tiger [12]	Medium	High	Medium	Medium	Medium
Asym-RAID (proposed)	High	Low	Low	Low	High