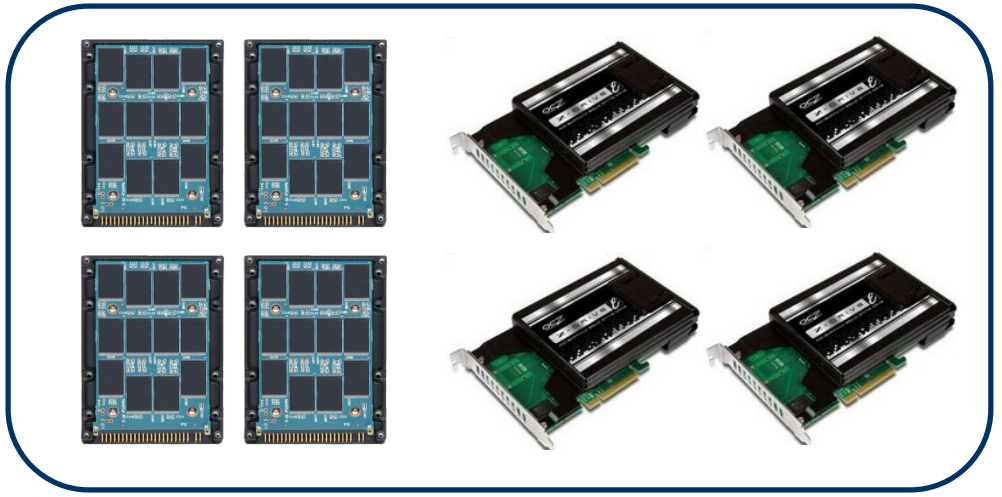
# RAID 遇上 SSD



# RAID for SSD

## SSDs are less array-friendly than hard disks

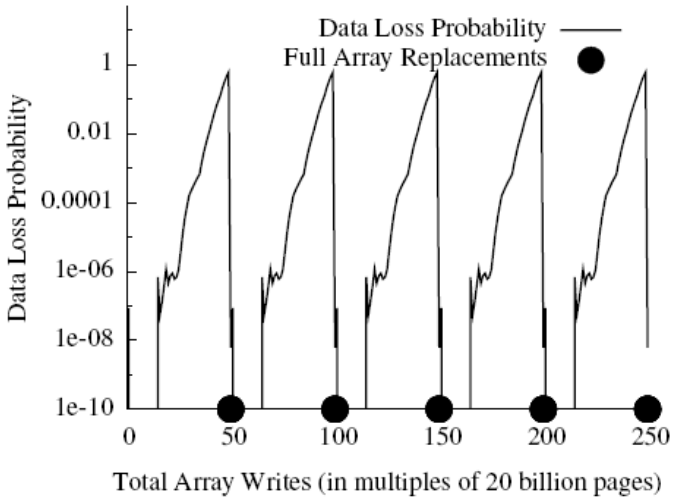
- SSD擦写次数受限，延迟波动大，尾延迟问题明显
  - SLC：100,000次， MLC：5,000-10,000次
- RAID中数据尽量均匀分布到各成员盘
  - RAID0， RAID5
- RAID5中校验数据块的擦写更频繁
  - N-disk RAID-5，任何一个成员盘的数据更新都会触发校验数据更新

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

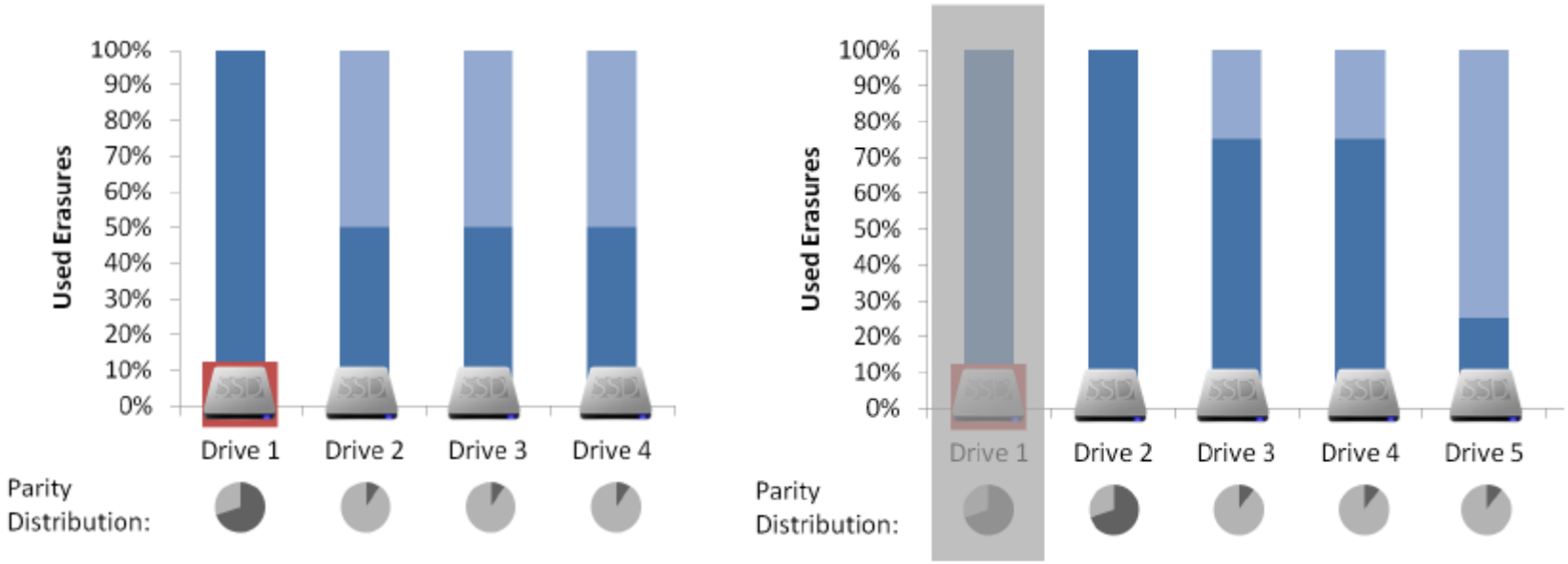


# Case 1： Differential RAID

## 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以不同的速率老化设备。



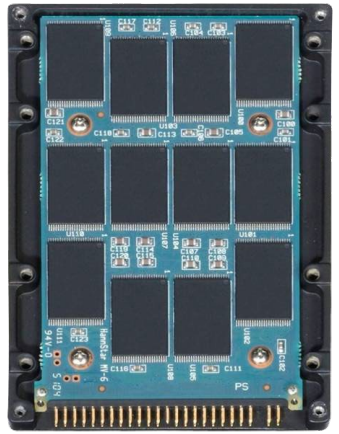

为了在旧设备被新设备替换时保持这种年龄差异，每次驱动器更换时重新调整奇偶校验分布。

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

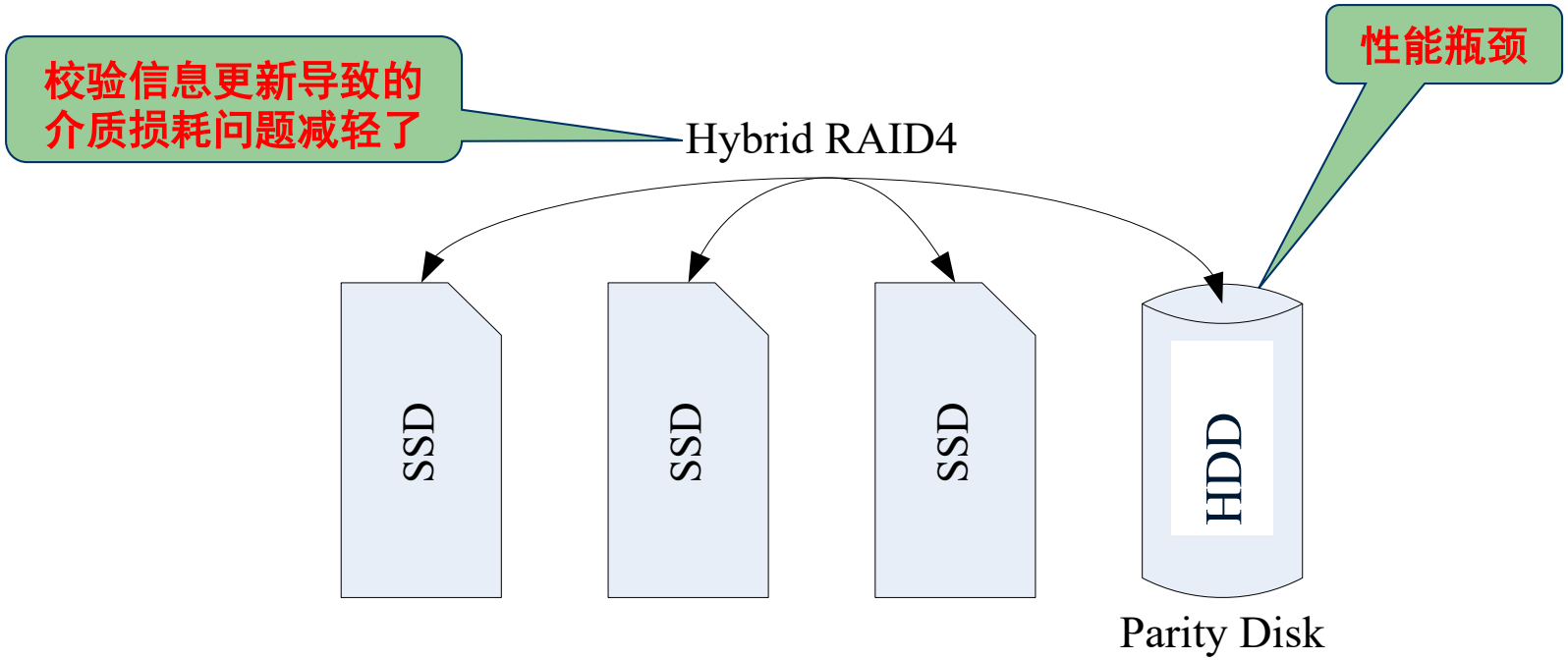
# Case2：混合式盘阵列

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

- 固态盘
  - 低能耗
  - 非对称读/写性能
  - 介质损耗问题



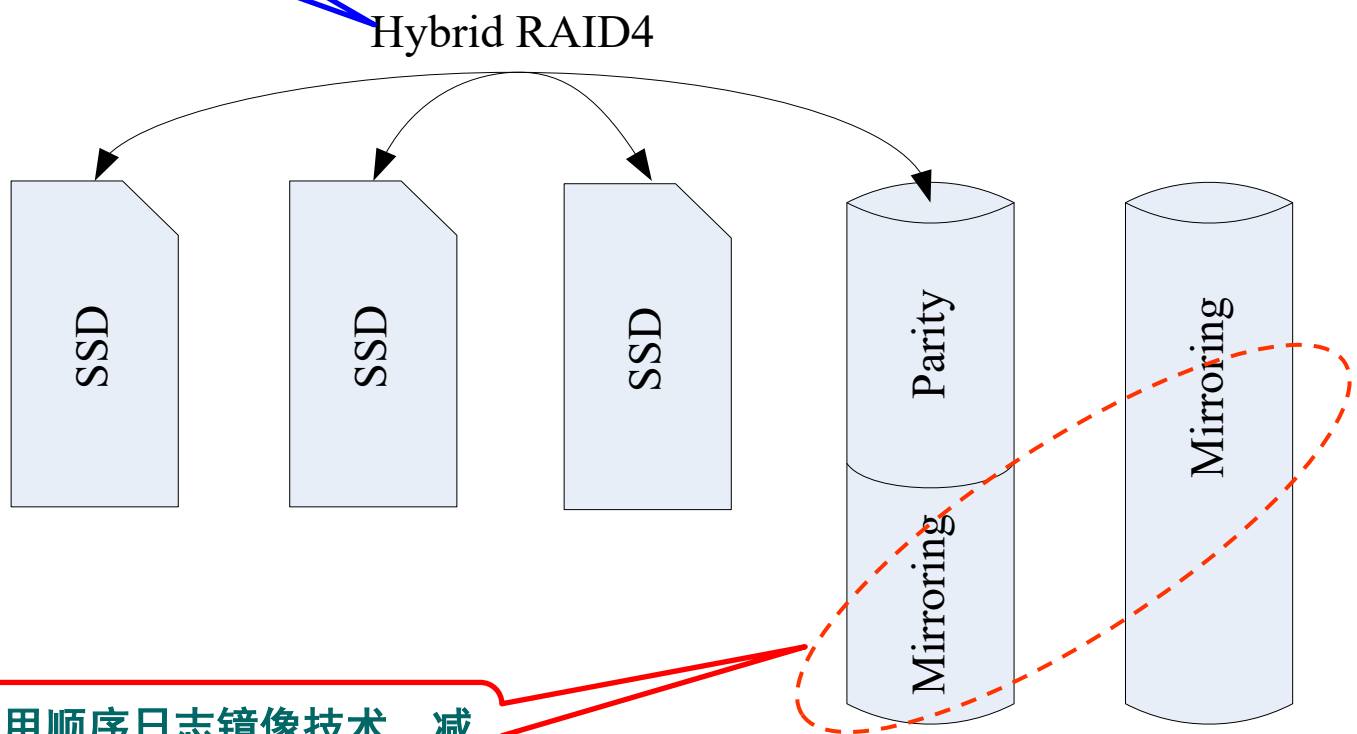
# 混合式盘阵列




- 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）

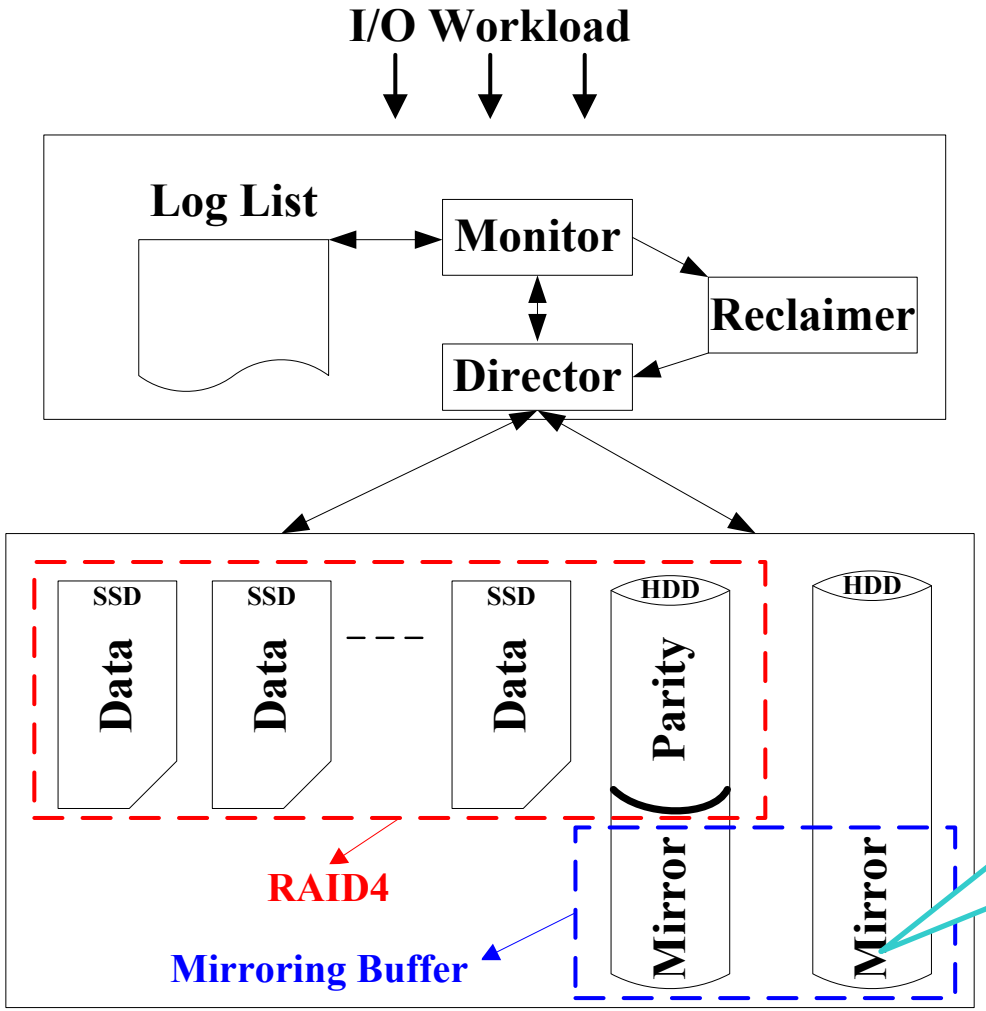
# 混合式盘阵列

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，减轻了随机小写性能问题

# 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 |

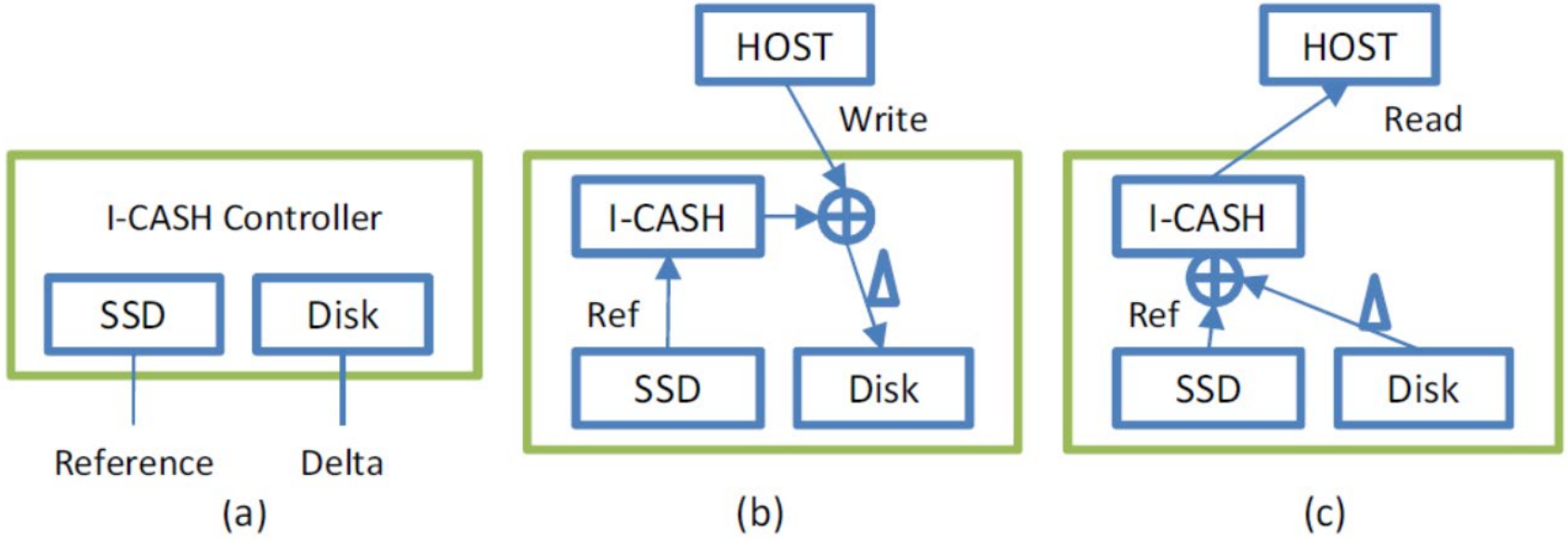
**Content Locality：**

**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 write 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

## Selected Publications

ISCA
- "TRAP-Array: A Disk Array Architecture Providing Timely Recovery to Any Point-in -time" in The 33rd Annual International Symposium on Computer Architecture, 2006 (ISCA'06). Qing Yang, Weijun Xiao, and Jin Ren

ISCA
- ``DCD---Disk Caching Disk: A New Approach for Boosting I/O Performance," The 23rd Annual International Symposium on Computer Architecture, Philadelphia PA May, 1996. (ISCA'96). Y. Hu and Qing Yang

ISCA
- ``Caching Address Tags: A technique to reduce chip area cost for on-chip caches," The 22nd Annual International Symposium on Computer Architecture,Santa Margherita Ligure, Italy, June, 1995. (ISCA'95). H. Wang, T. Sun, and Qing Yang

ISCA
- ``A novel cache design for vector processing," The 19th International Symposium on Computer Architecture, May 1992, pp. 362-37 1. Gold Coast, Australia. (ISCA'92). Qing Yang and Liping W. Yang

HPCA
- "I-CASH: Intelligently Coupled Array of SSD and HDD" in The 17th IEEE International Symposium on High Performance Computer Architecture, 2011 (HPCA'11), San Antonio, TX, Feb 2011. Jin Ren and Qing Yang

HPCA
- ``RAPID-Cache --- A Reliable and Inexpensive Write Cache for Disk I/O Systems", in The 5th International Symposium on High Performance Computer Architecture (HPCA-5). Orlando, Florida. Jan. 1999. Y. Hu, Qing Yang, and T. Nightingale

# Case 4：FusionRAID

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

# 全闪阵列（AFAs，All-Flash Arrays）

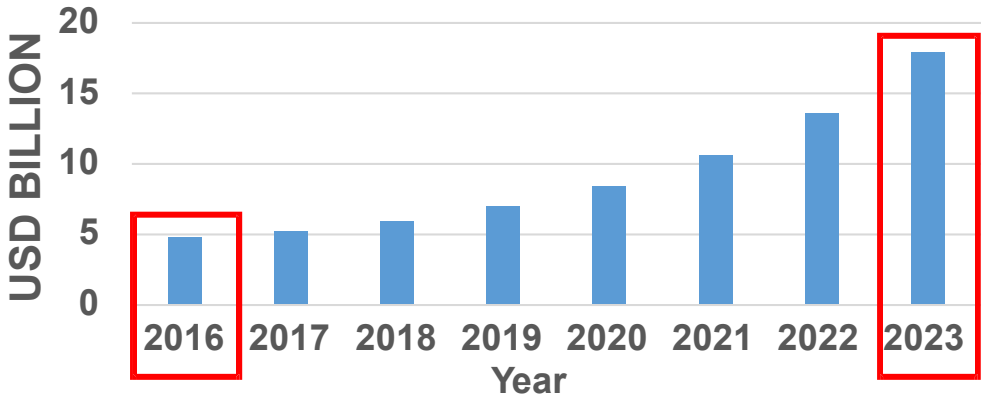
- 近年来应用广泛


Banks

Datacenters


Clouds

- AFA 市场持续高速增长




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


DELL EMC VMAX


PureStorage FlashArray

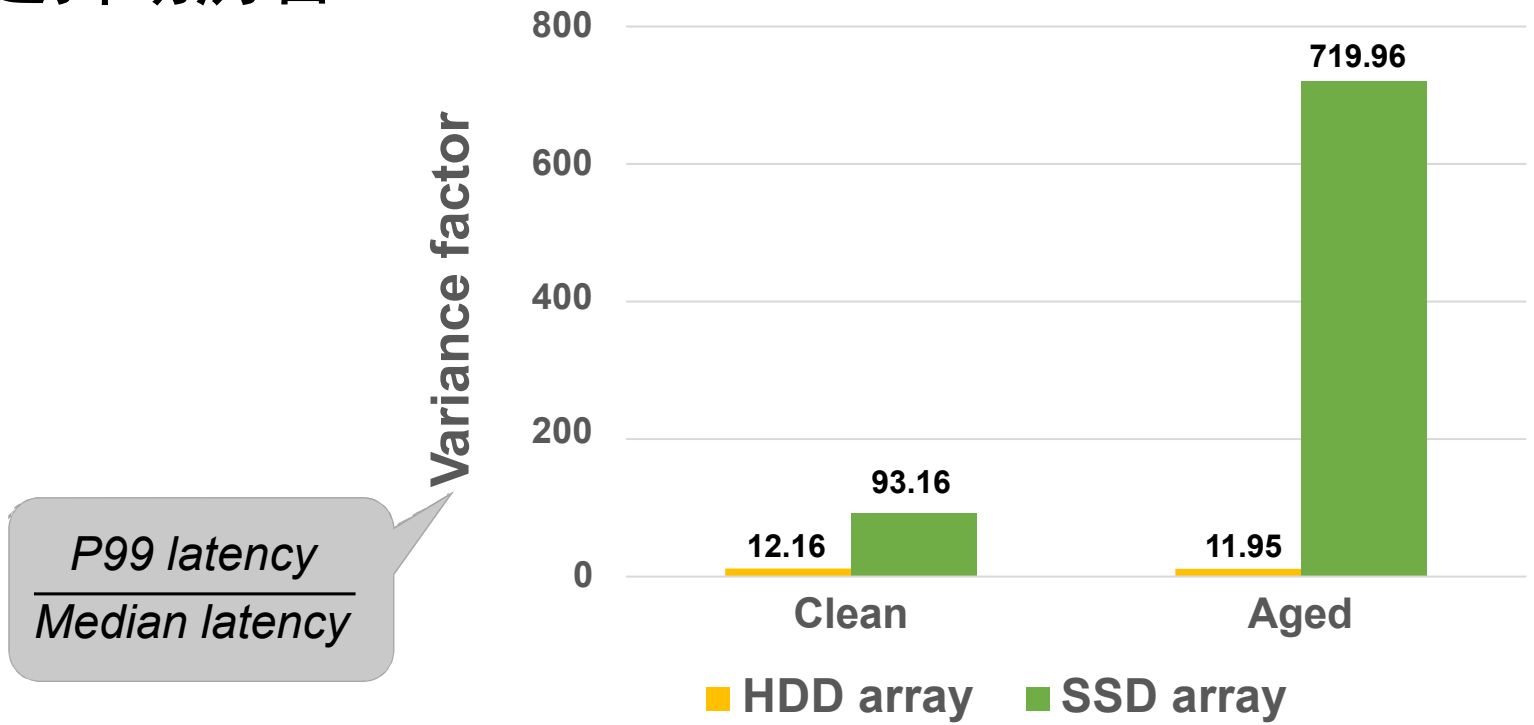

SanDisk InfiniFlash


FUJITSU ETERNUS


NetApp AFF

# 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

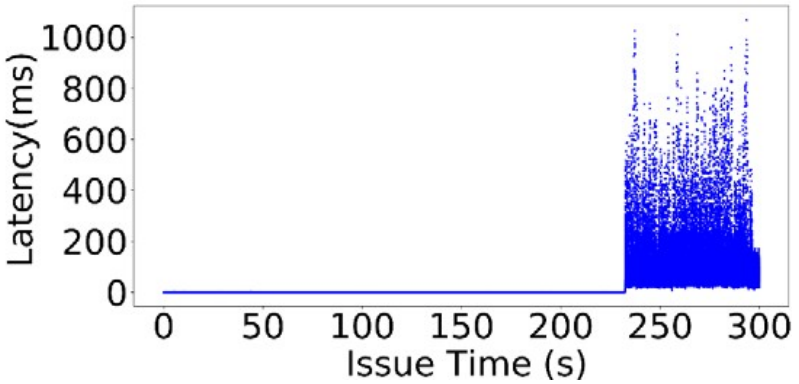



| | Median latency (ms) | Avg. latency (ms) | P99 latency (ms) | Variance factor |
|---|---|---|---|---|
| *HDD RAID (clean)* | 68.67 | 134.37 | 835.35 | 12.16 |
| *HDD RAID (aged)* | 69.18 | 133.61 | 826.77 | 11.95 |
| *SSD RAID (clean)* | 0.275 | 3.57 | 25.62 | 93.16 |
| *SSD RAID (aged)* | 0.307 |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

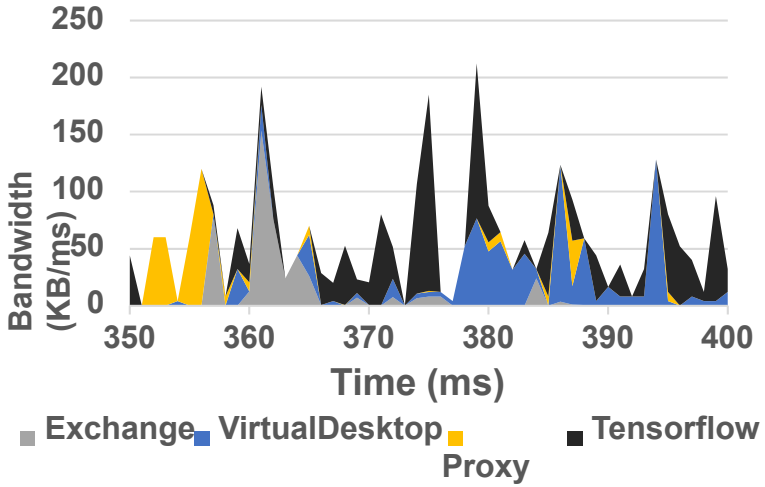
Table 1: Exchange latency, HDD vs. SSD RAID

# 实验观察：

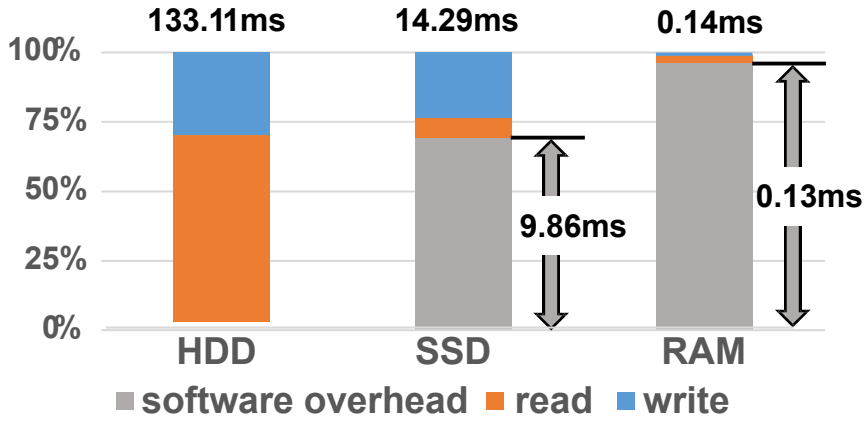
1. 工作负载通常不规则，交错突发（interleaving bursts）
   - All-for-all 映射模式优于物理分区
2. SSD RAID 写入操作**软件开销**严重
   - 相对开销比HDD RAID高，绝对开销远高于写RAM(Disk)
   - 原因在于synchronization
   - 需要缩短写路径
3. SSD性能异常常见，写尤其突出（幅度和持续时间）
   - 延迟峰值高且持久，足以在运行时感知




Datacenter SSDs with random writes



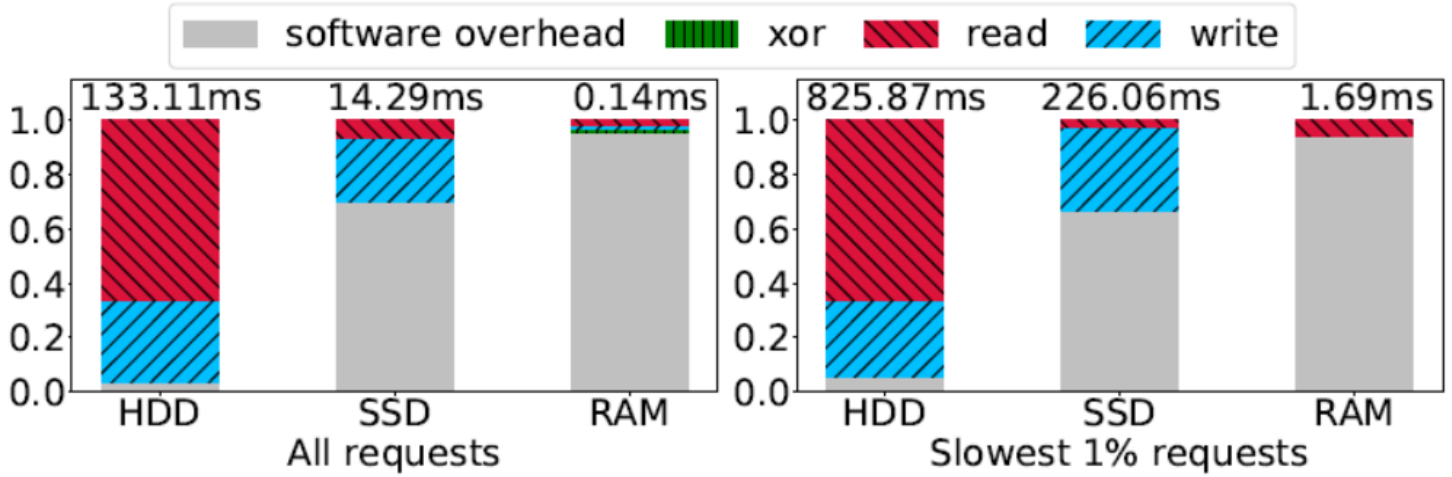

Bandwidth consumption in 4-workload mix




RAID write latency breakdown

# Write Overhead in SSD RAID

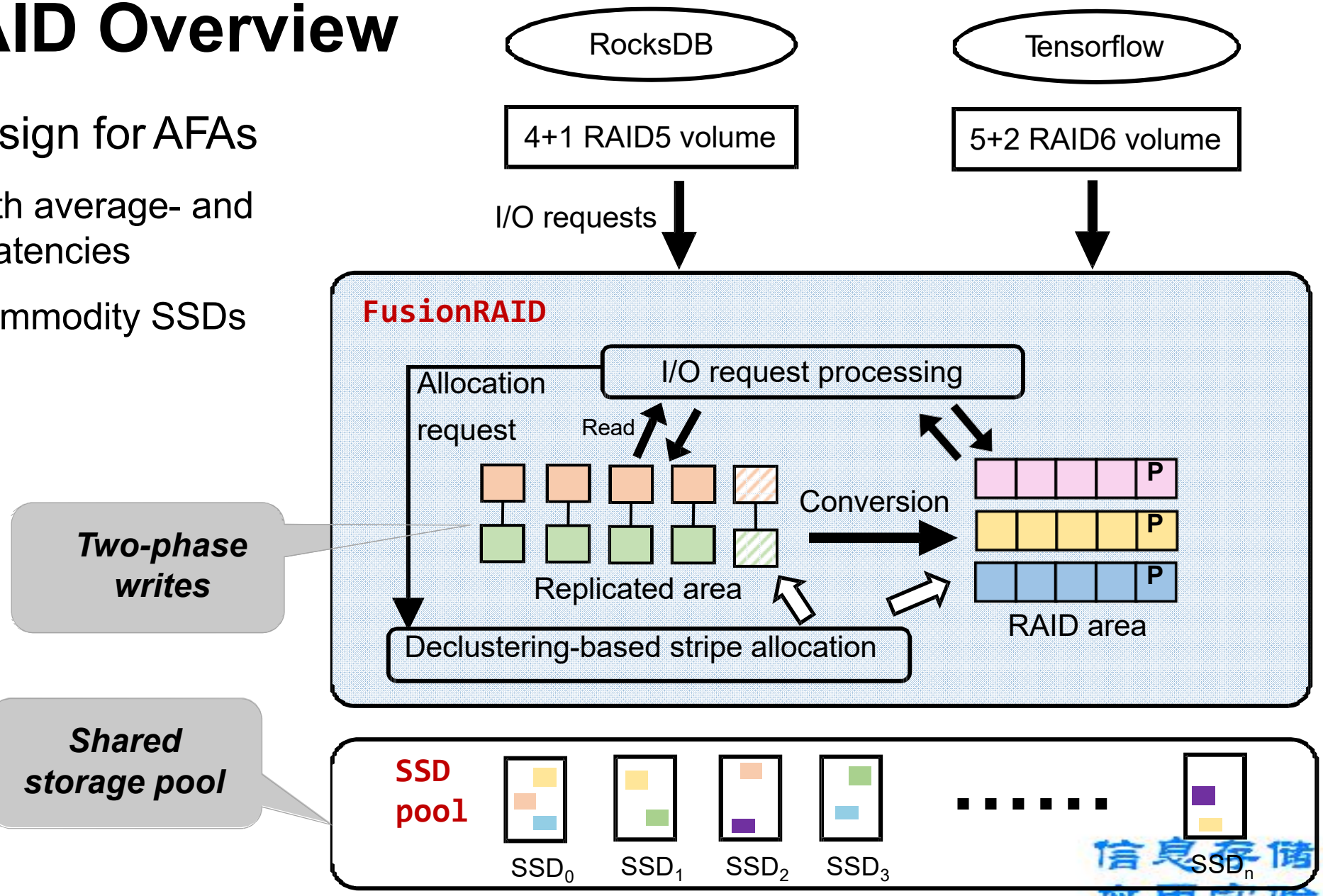
1. 软件开销在HDD RAID写过程中对延迟的影响可以忽略不计
2. SSD RAID 写入操作**软件开销**是写盘时间的2. 9倍
   - 相对开销比HDD RAID高，绝对开销远高于写RAM（2个数量级）
   - 软件开销还使最慢的 1% 请求延迟增加为平均延迟的 10 倍
3. 在最慢的1% 请求中，SSD的写延迟对尾延迟贡献也很大，是平均写时延的20. 7倍




**A shorter write path, with fewer dependencies**, may greatly reduce SSD RAID latency, both under average and worst-case scenarios.

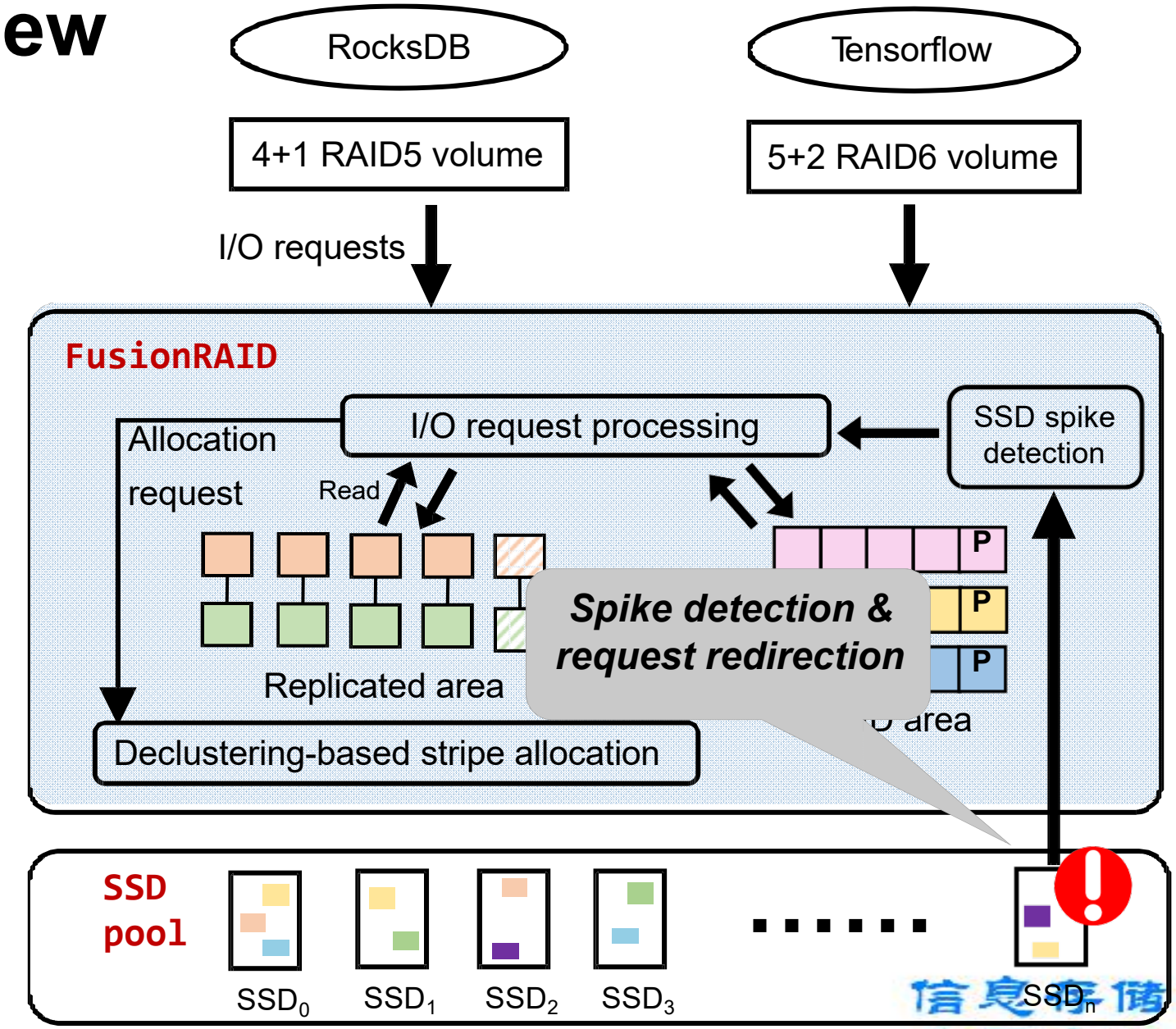
# FusionRAID Overview

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



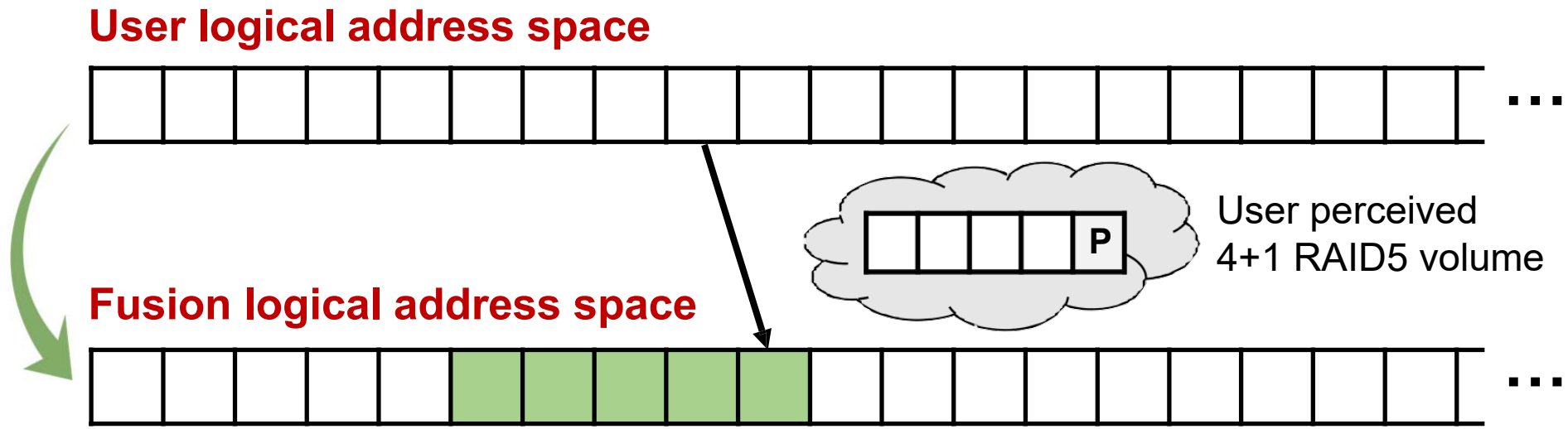
# FusionRAID Overview

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

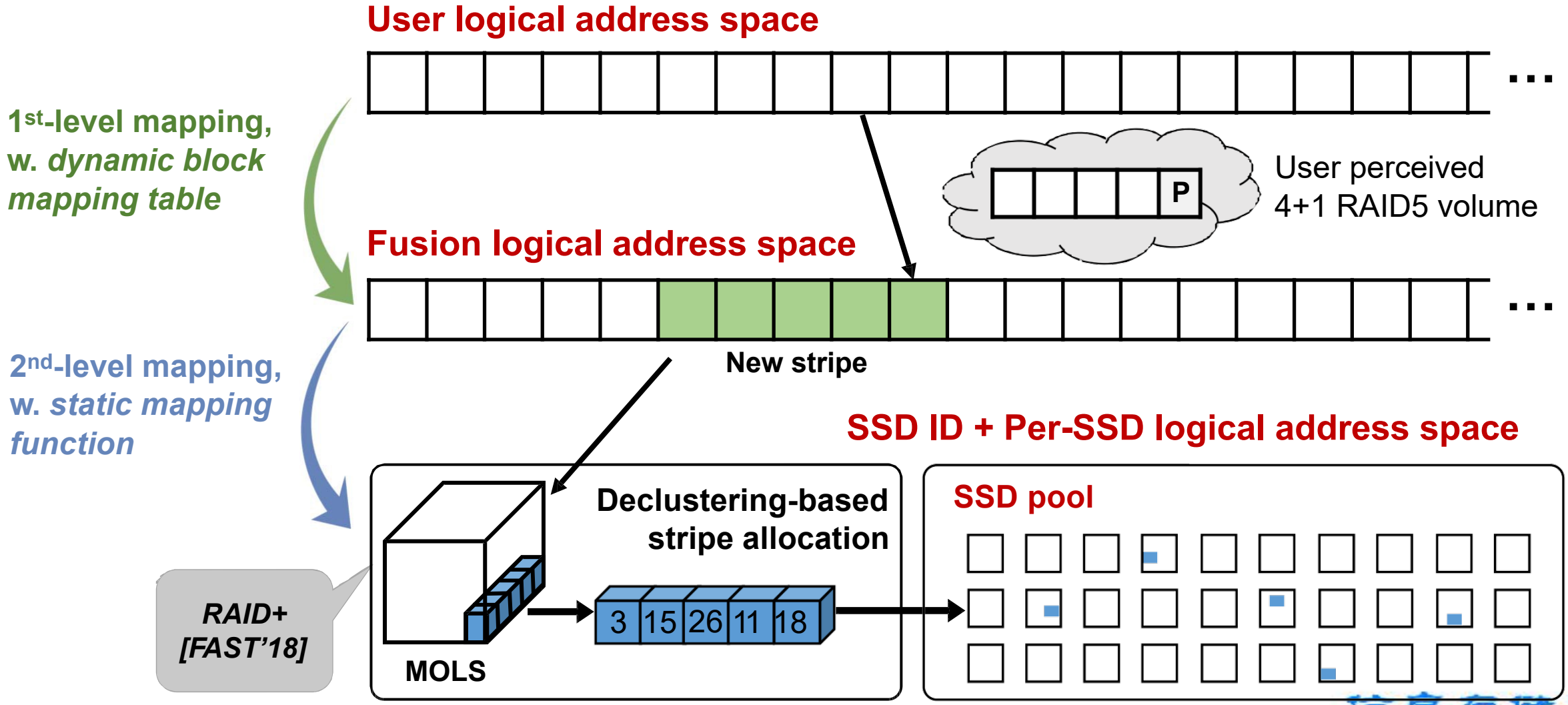


# Shared Storage Pool

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



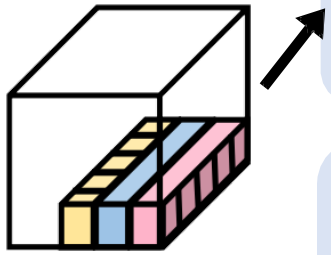
# Shared Storage Pool



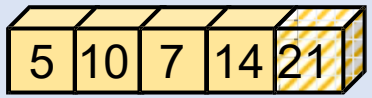
# FusionRAID Optimized Writes

Large write request

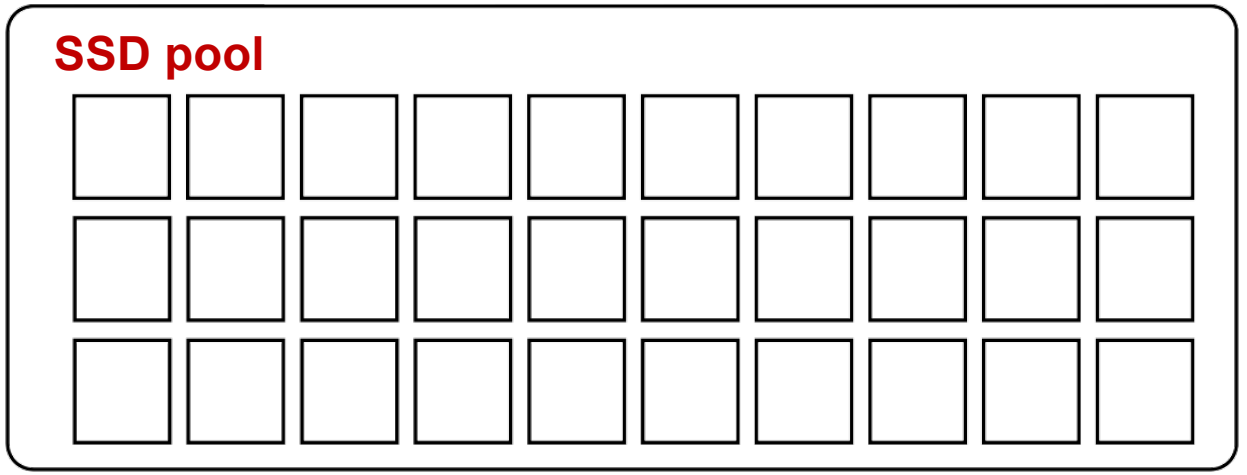
(Direct RAID write)



*RAID storage*



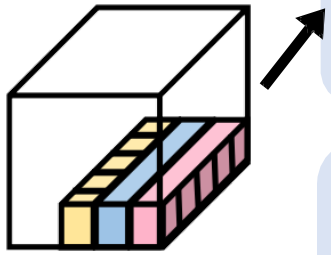

*Replicated storage*



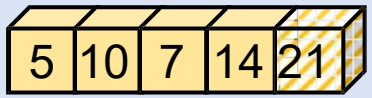
# FusionRAID Optimized Writes

Large write request

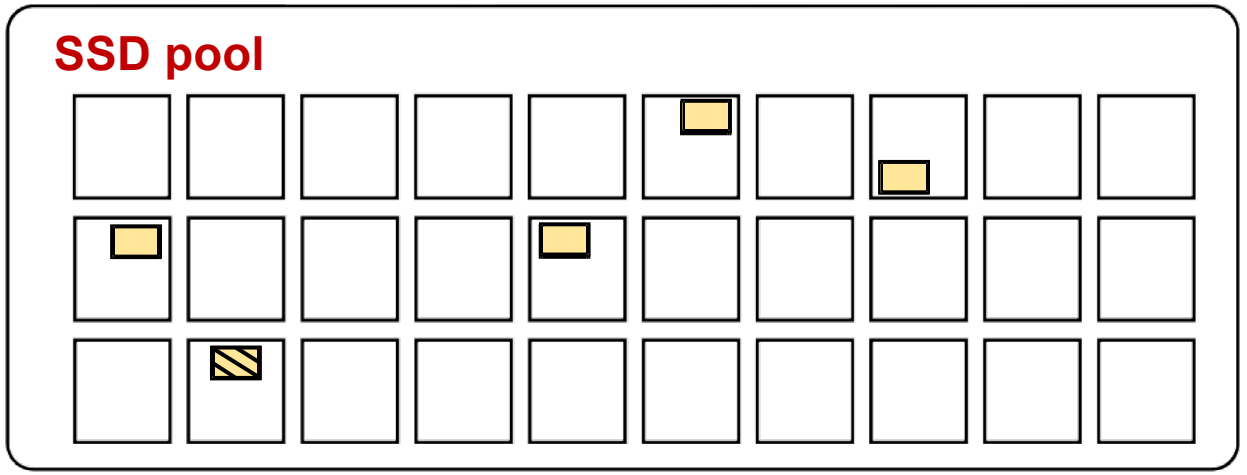
(Direct RAID write)



*RAID storage*



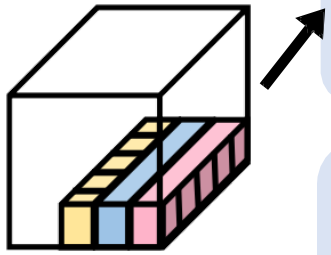

*Replicated storage*



# FusionRAID Optimized Writes

Large write request

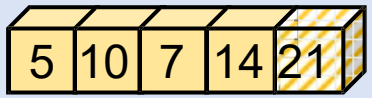
(Direct RAID write)



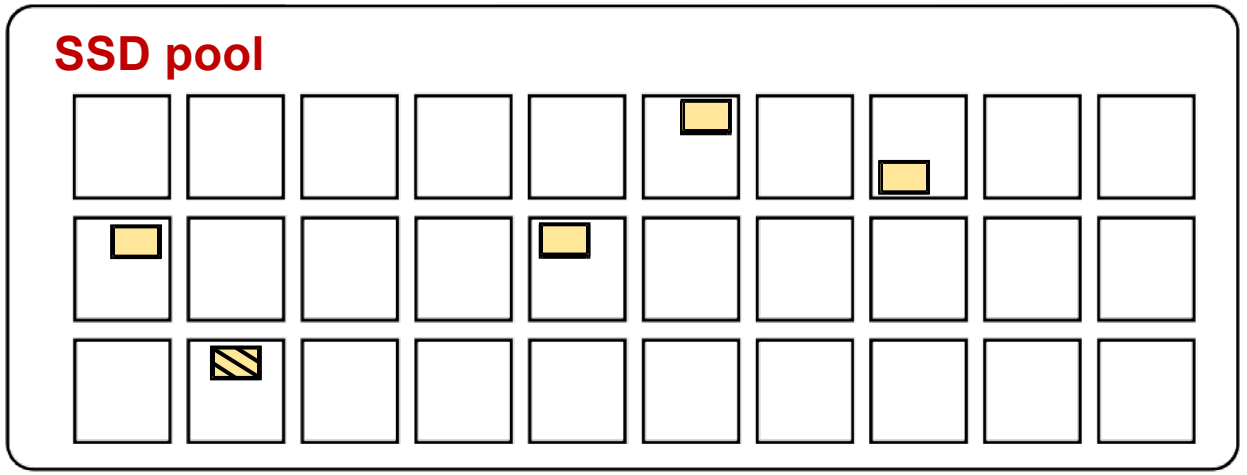
Small write request

(2-phase write)

*RAID storage*



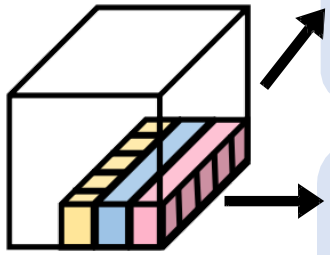

*Replicated storage*



# FusionRAID Optimized Writes

Large write request

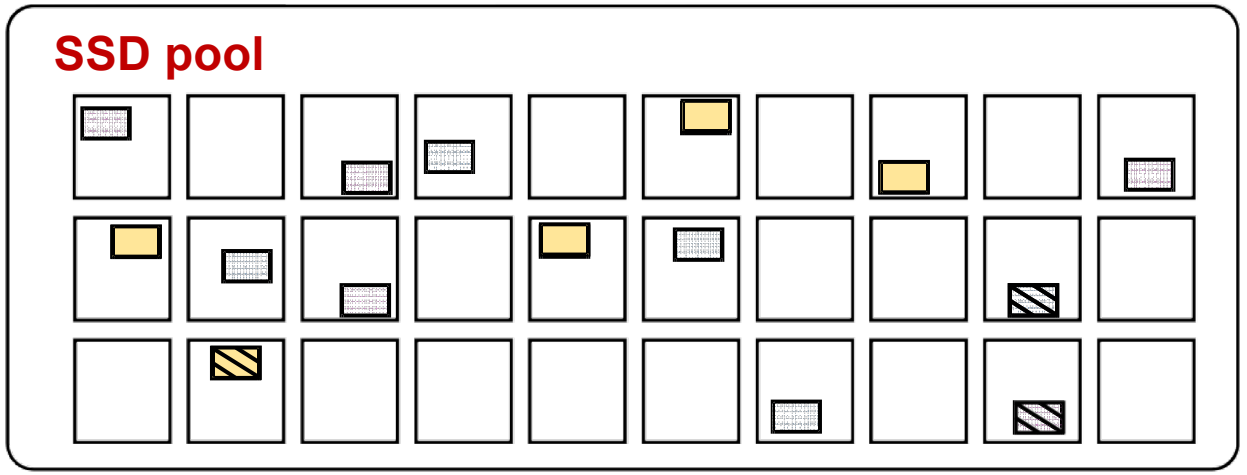
(Direct RAID write)



Small write request

(2-phase write)

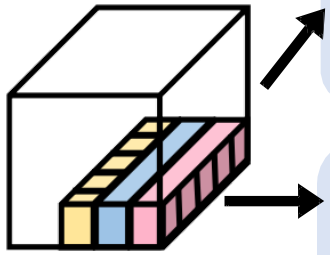


# FusionRAID Optimized Writes

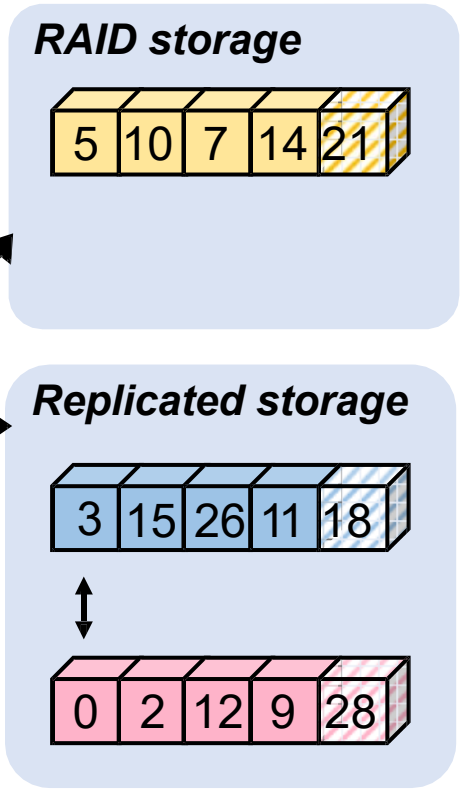
Large write request

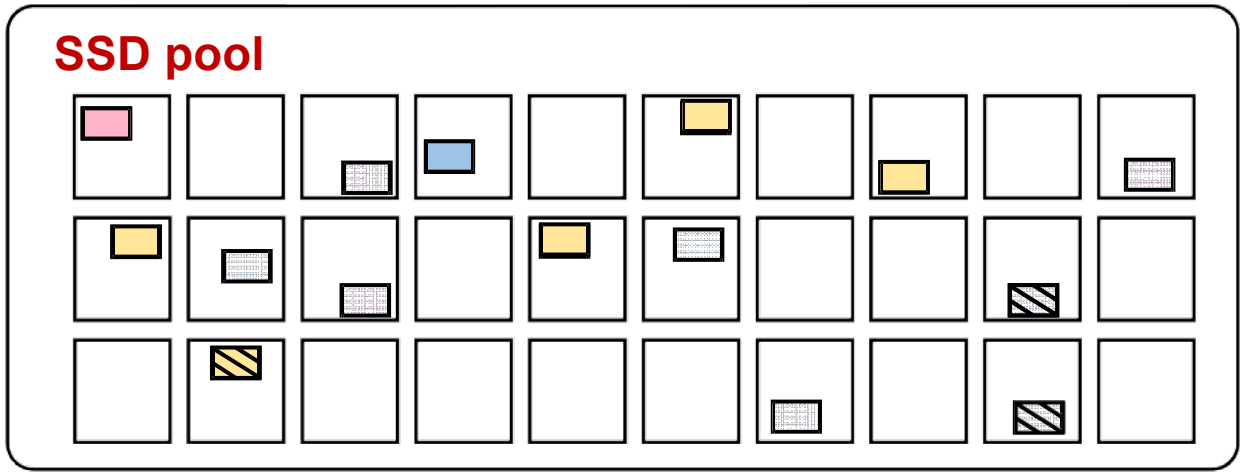
(Direct RAID write)



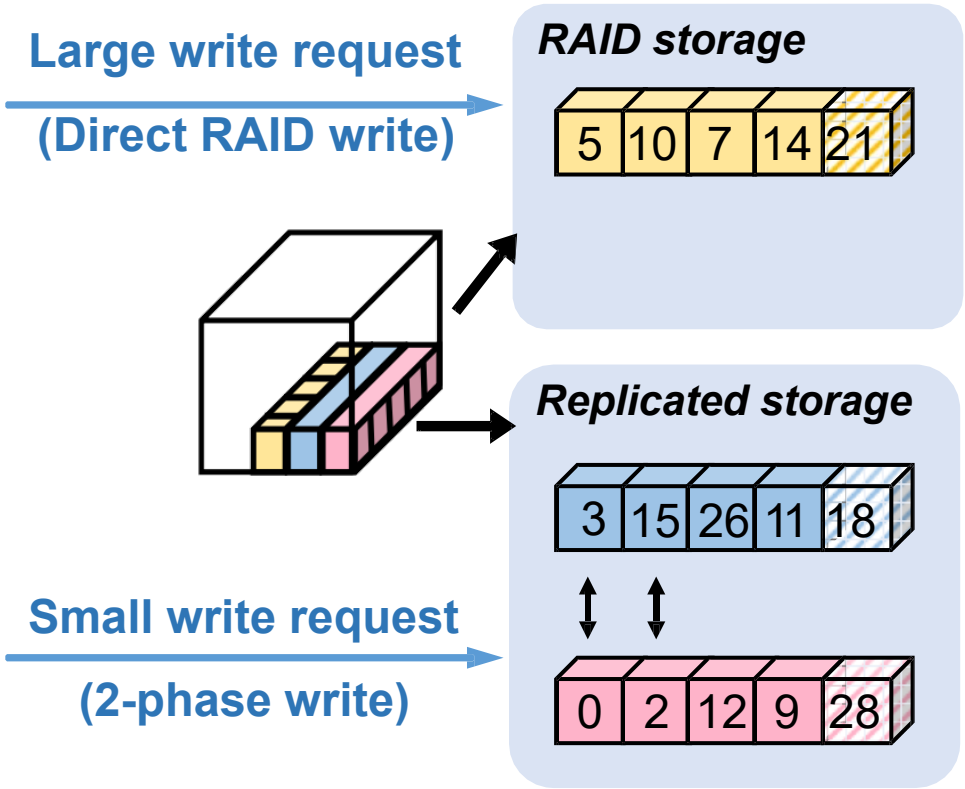
Small write request

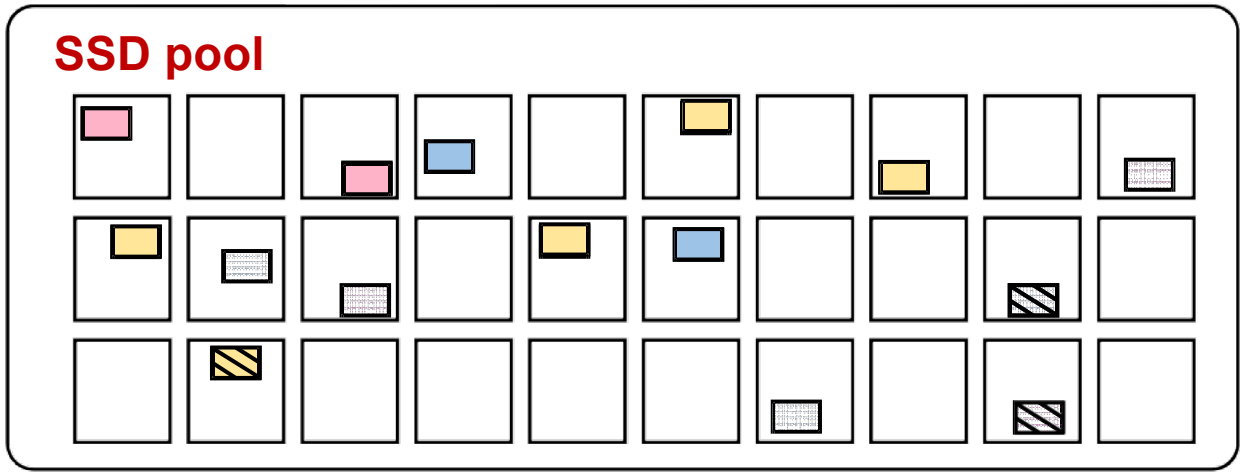
(2-phase write)



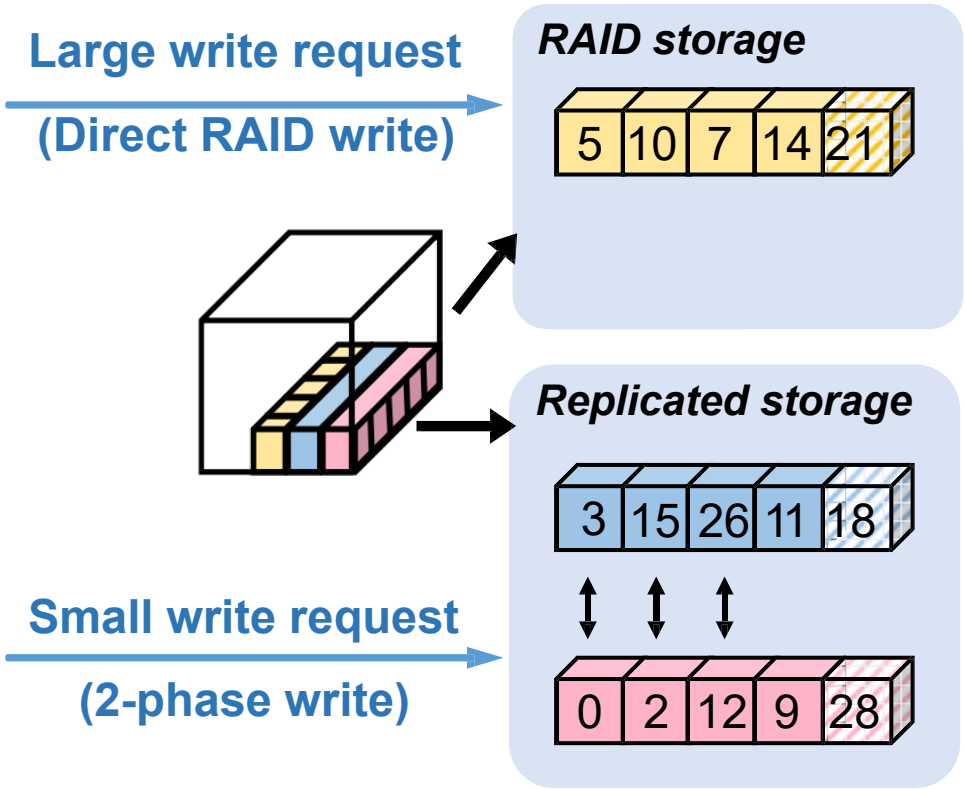



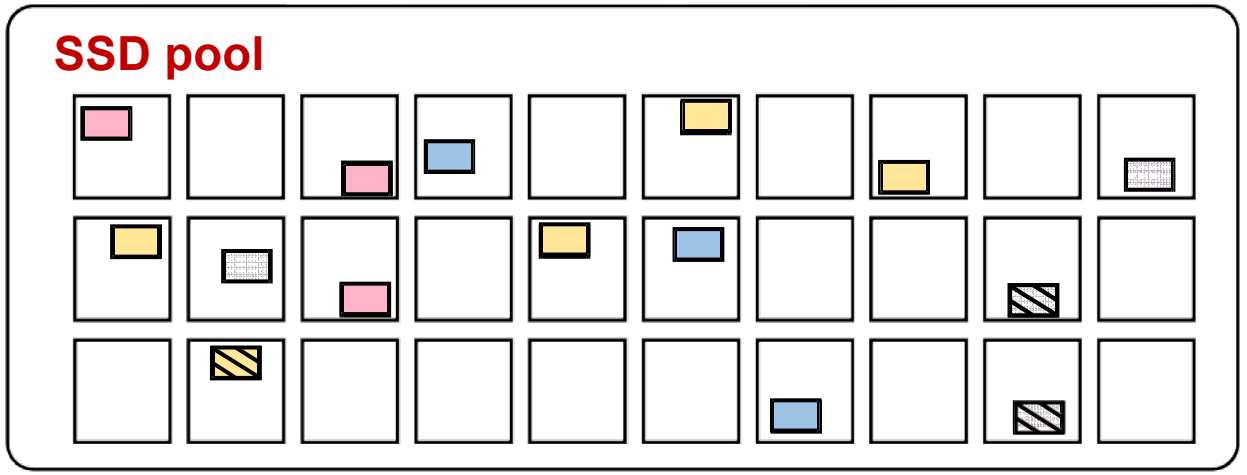
# FusionRAID Optimized Writes



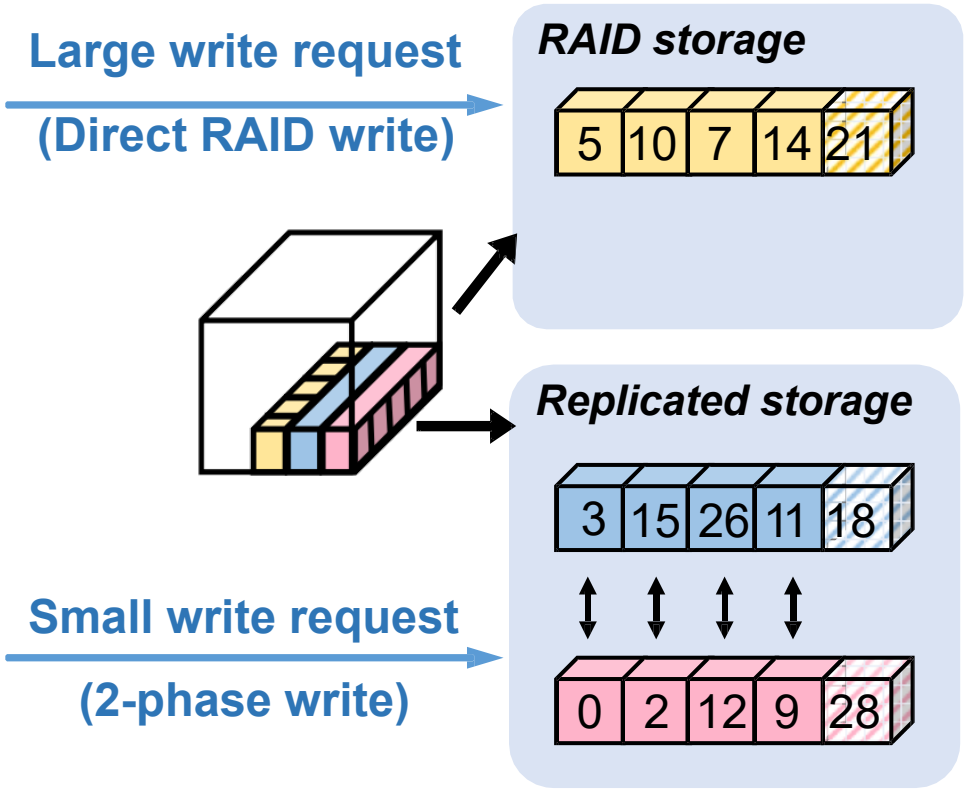



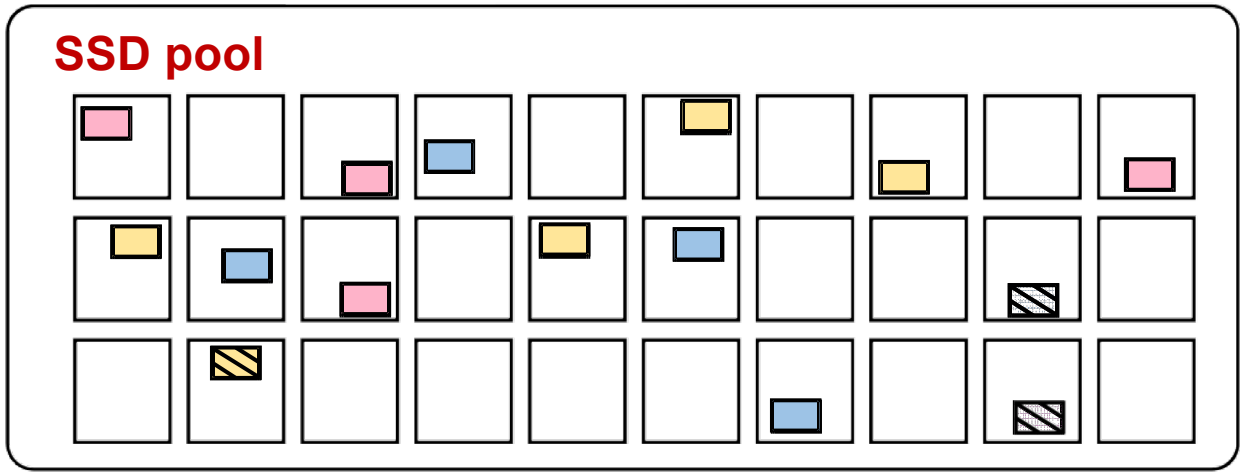
# FusionRAID Optimized Writes



# FusionRAID Optimized Writes

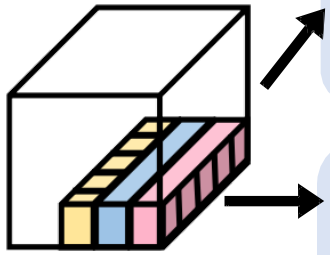


# FusionRAID Optimized Writes

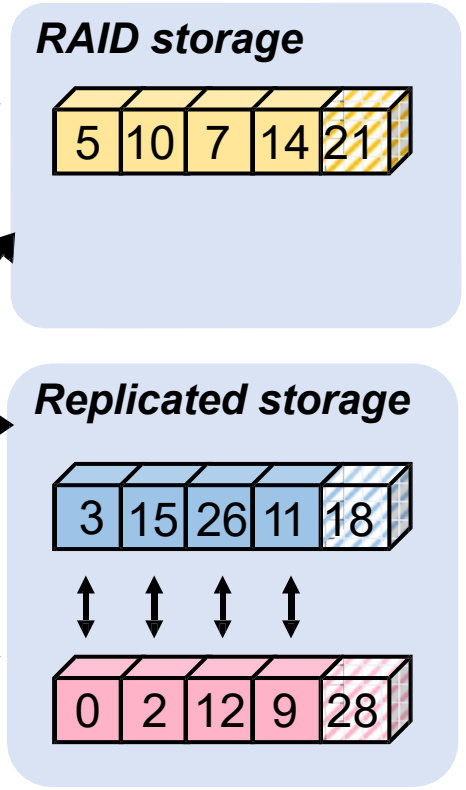
Large write request

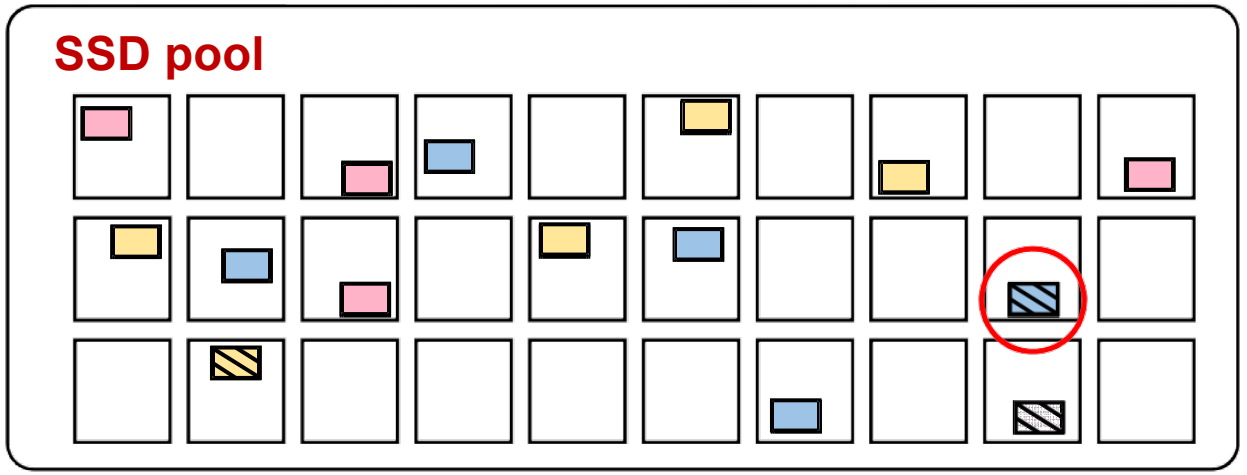
(Direct RAID write)



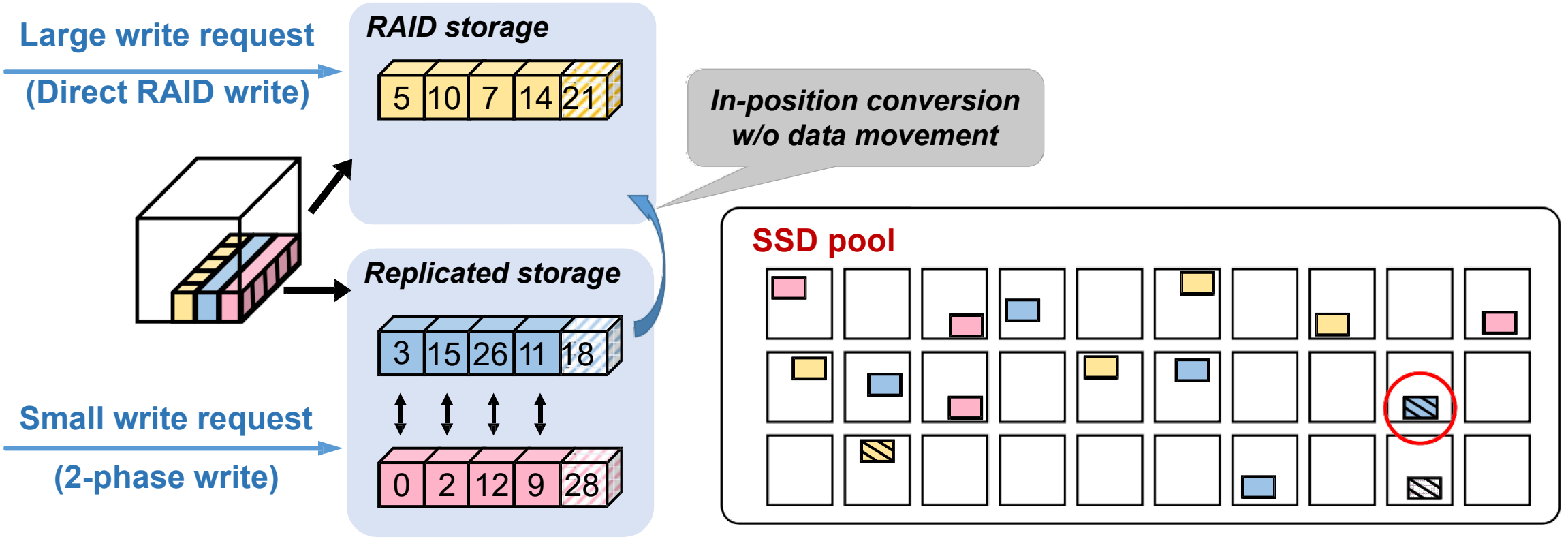
Small write request

(2-phase write)

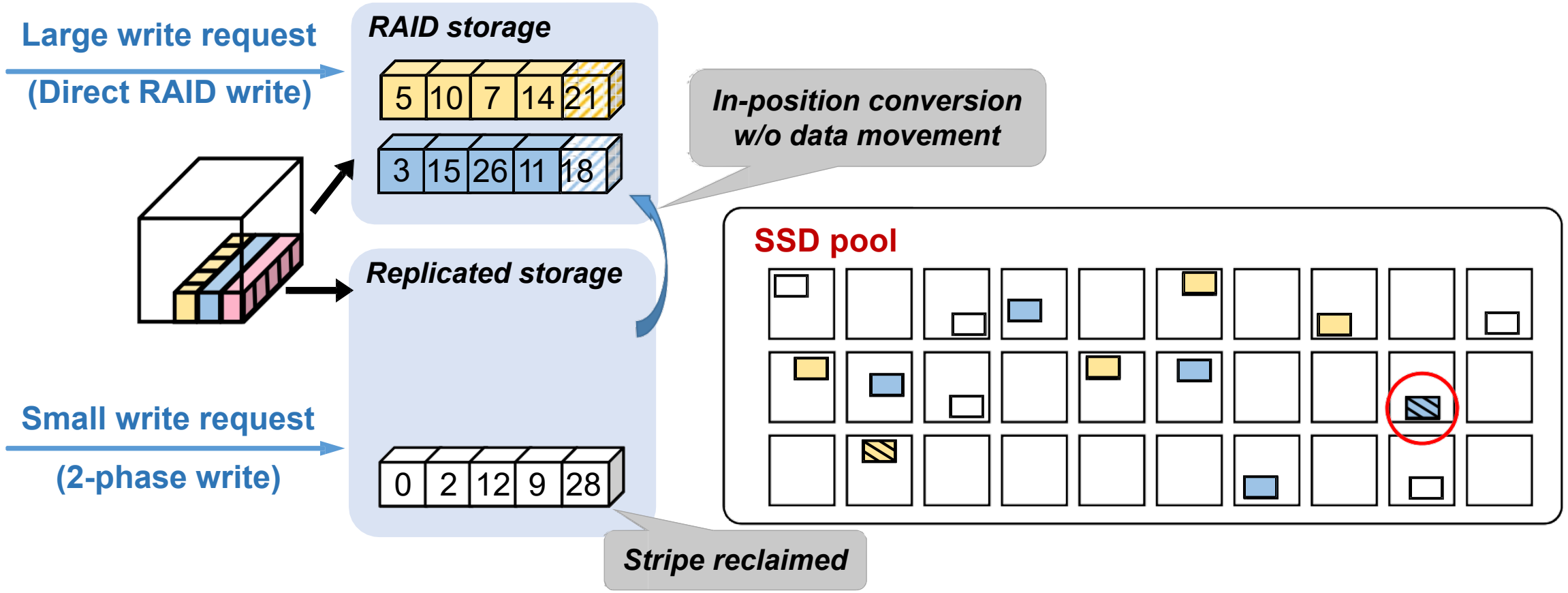


# FusionRAID Optimized Writes



# 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),

# The End