FusionRAID: Achieving Consistent Low Latency for Commodity SSD Arrays

Tianyang Jiang, Guangyan Zhang, Zican Huang, Xiaosong Ma,
Junyu Wei, Zhiyue Li, Weimin Zheng

Tsinghua University
Qatar Computing Research Institute, HBKU

All-Flash Arrays (AFAs) On Rise

- Widely used in recent years

Banks

Datacenters

Clouds

All-Flash Arrays (AFAs) On Rise

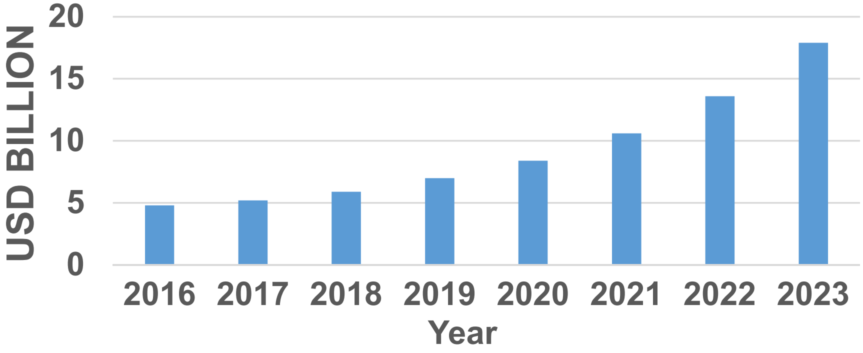
- Widely used in recent years

Banks

Datacenters

Clouds

- AFA market
  - Rapidly growing in past years
  - Growth projected to continue
  - Many products on market



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

All-Flash Arrays (AFAs) On Rise

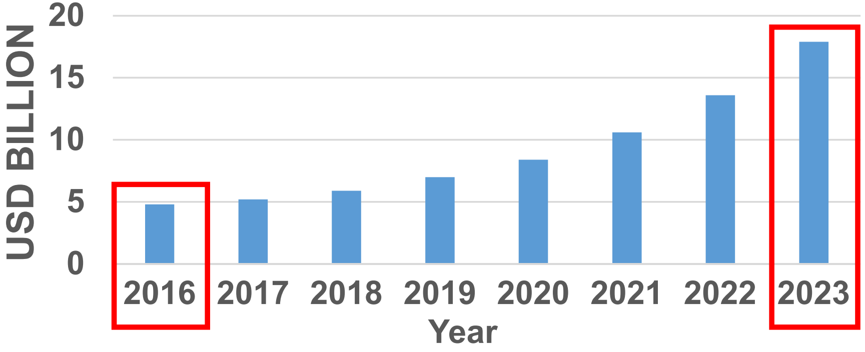
- Widely used in recent years

Banks

Datacenters

Clouds

- AFA market
  - Rapidly growing in past years
  - Growth projected to continue
  - Many products on market



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

All-Flash Arrays (AFAs) On Rise

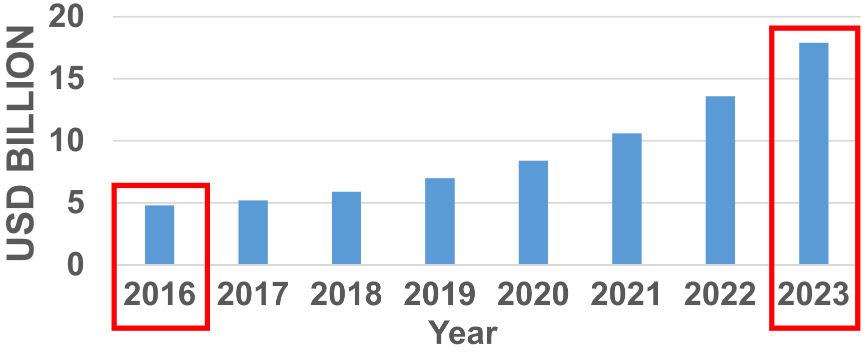
- Widely used in recent years

Banks

Datacenters

Clouds

- AFA market
  - Rapidly growing in past years
  - Growth projected to continue
  - Many products on market



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

DELL EMC VMAX

PureStorage FlashArray

SanDisk InfiniFlash

FUJITSU ETERNUS

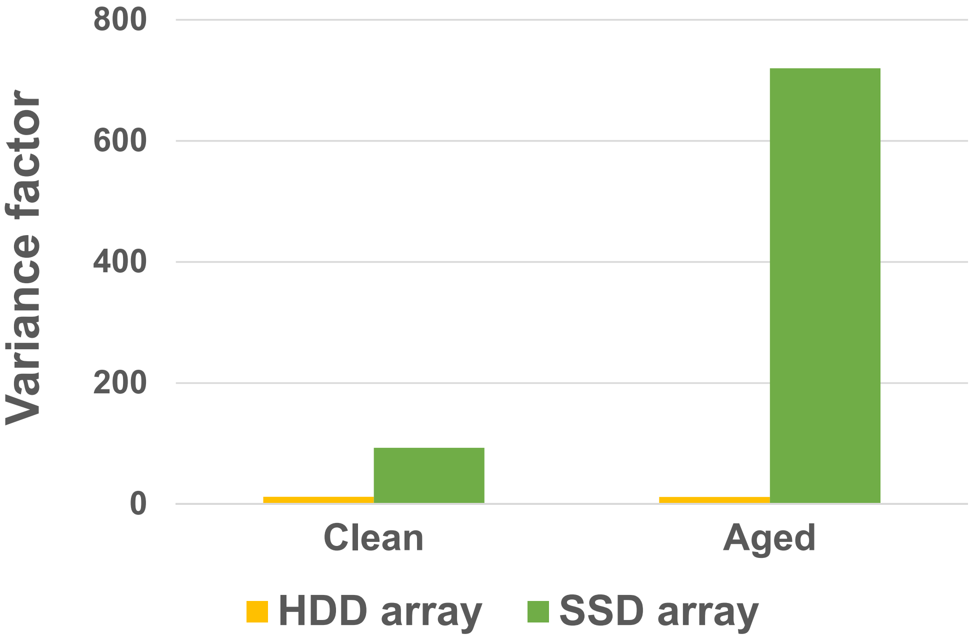
NetApp AFF

Severe SSD RAID Performance Problems

- Higher latency variability compared to HDD RAID
    - Tail deviate more from norm

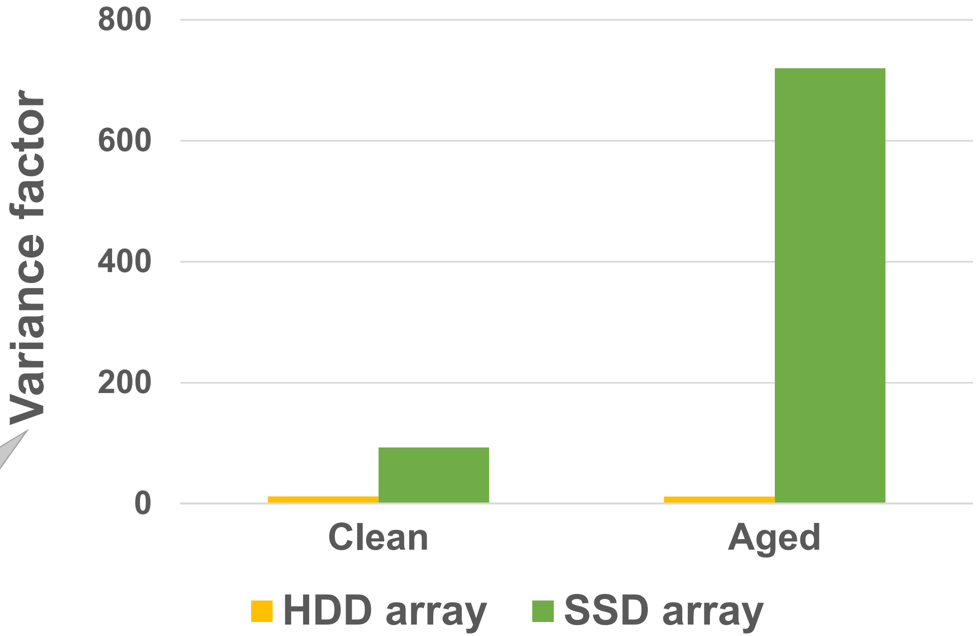
Severe SSD RAID Performance Problems

- Higher latency variability compared to HDD RAID
  - Tail deviate more from norm



Severe SSD RAID Performance Problems

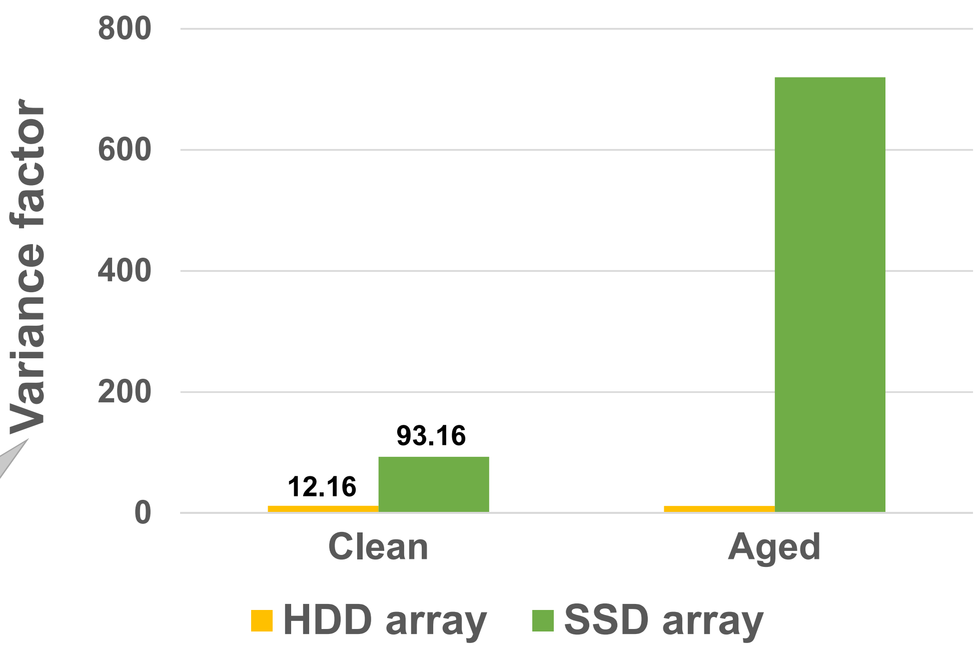
- Higher latency variability compared to HDD RAID
  - Tail deviate more from norm



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

Severe SSD RAID Performance Problems

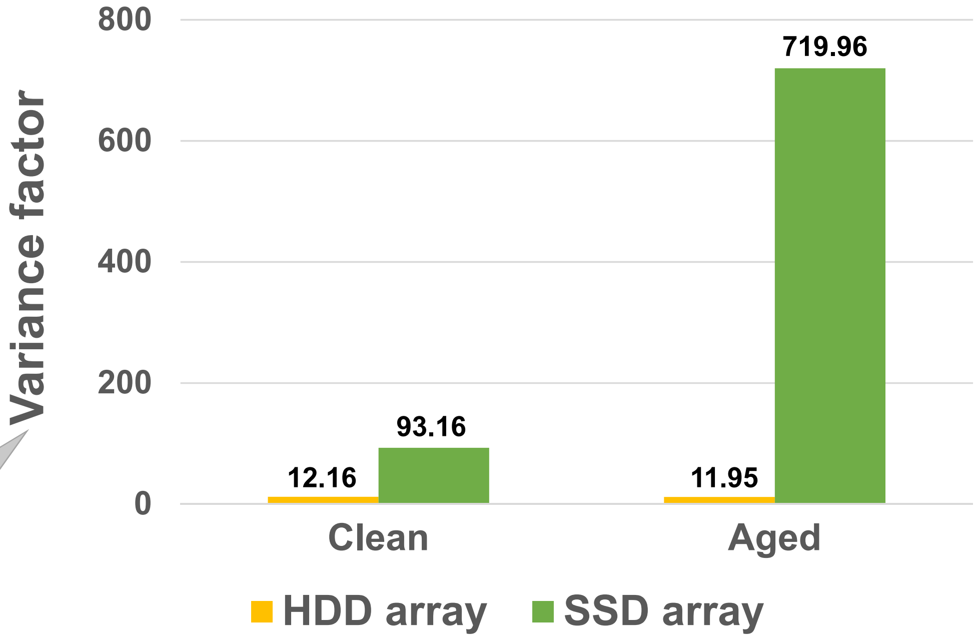
- Higher latency variability compared to HDD RAID
    - Tail deviate more from norm



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

Severe SSD RAID Performance Problems

- Higher latency variability compared to HDD RAID
  - Tail deviate more from norm
  - Further agitated by disk aging

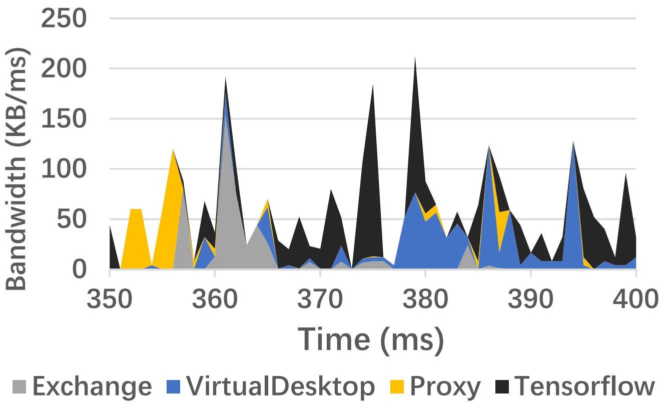


$$\frac{\textit{P99 latency}}{\textit{Median latency}}$$

Observations from Empirical Study

Observations from Empirical Study

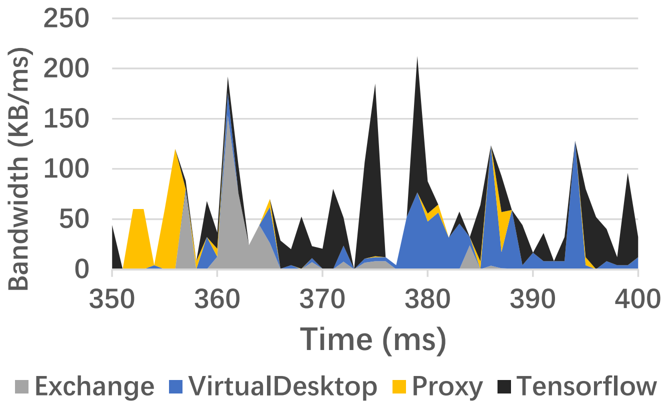
1. Workloads usually irregular, with interleaving bursts
   - **All-for-all model better than physically partitioning**



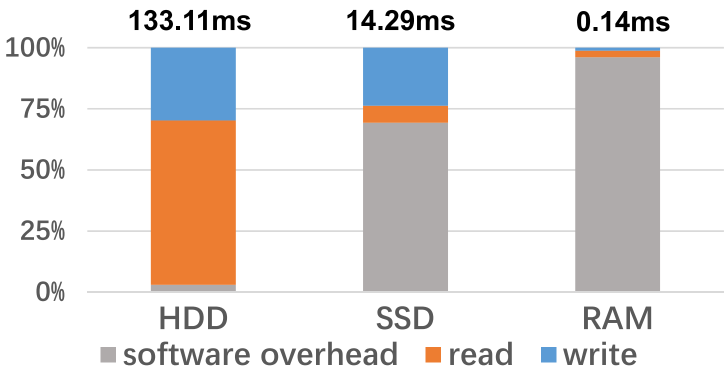
Bandwidth consumption in 4-workload mix

Observations from Empirical Study

1. Workloads usually irregular, with interleaving bursts
   - **All-for-all model better than physically partitioning**
2. SSD RAID writes suffer significant software overhead
   - Much higher relative overhead than w. HDD, and higher absolute overhead than w. RAM
   - Mainly caused by synchronization
   - **Shorter write path desirable**



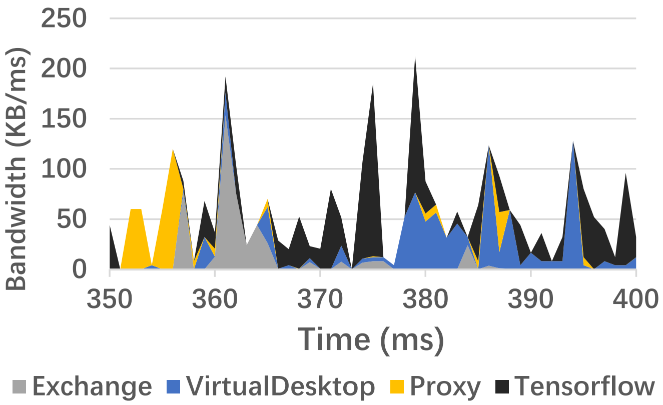
Bandwidth consumption in 4-workload mix



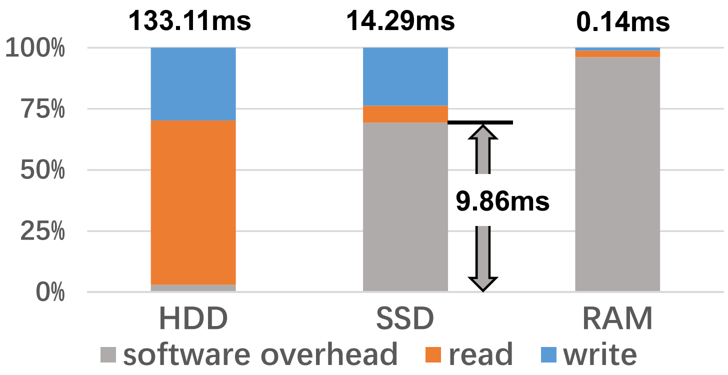
RAID write latency breakdown

Observations from Empirical Study

1. Workloads usually irregular, with interleaving bursts
   - **All-for-all model better than physically partitioning**
2. SSD RAID writes suffer significant software overhead
   - Much higher relative overhead than w. HDD, and higher absolute overhead than w. RAM
   - Mainly caused by synchronization
   - **Shorter write path desirable**



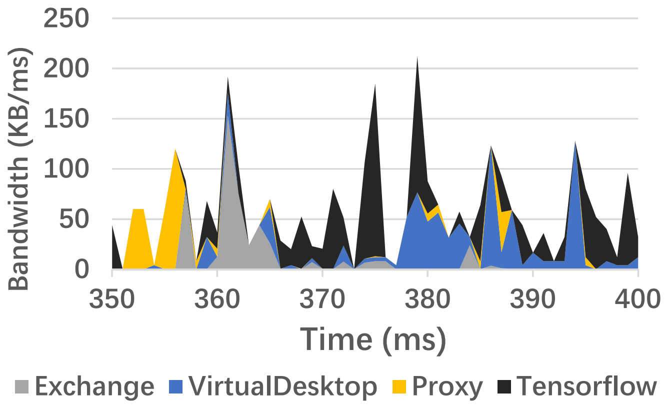
Bandwidth consumption in 4-workload mix



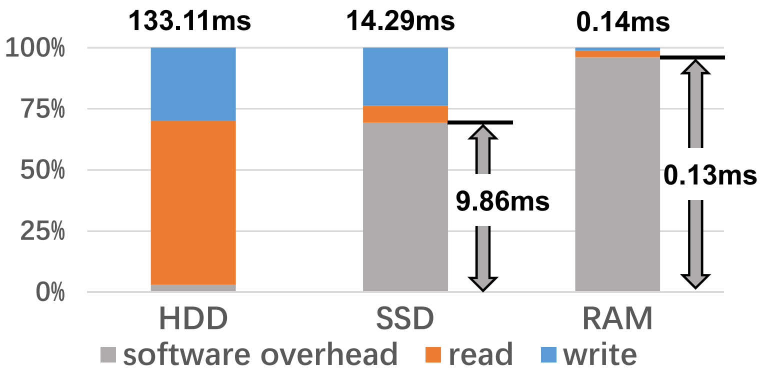
RAID write latency breakdown

Observations from Empirical Study

1. Workloads usually irregular, with interleaving bursts
   - **All-for-all model better than physically partitioning**
2. SSD RAID writes suffer significant software overhead
   - Much higher relative overhead than w. HDD, and higher absolute overhead than w. RAM
   - Mainly caused by synchronization
   - **Shorter write path desirable**



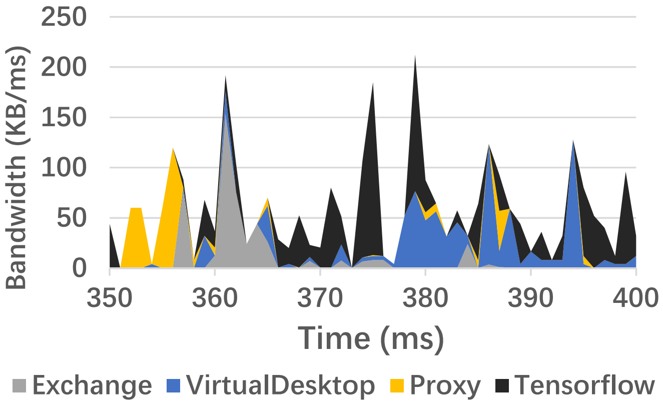
Bandwidth consumption in 4-workload mix



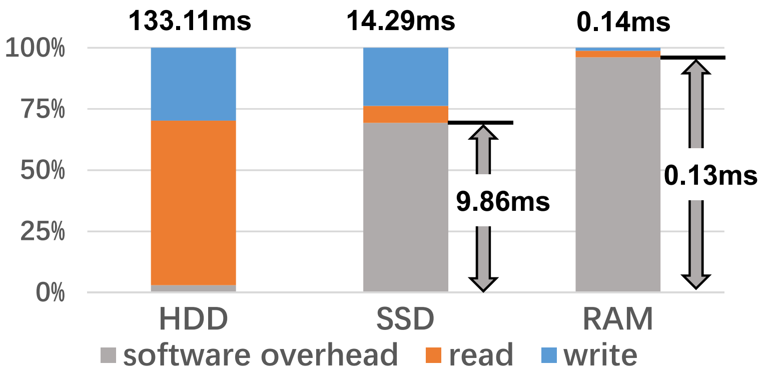
RAID write latency breakdown

Observations from Empirical Study

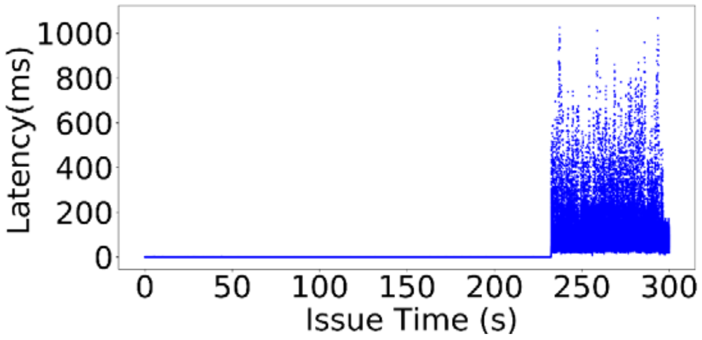
1. Workloads usually irregular, with interleaving bursts
   - **All-for-all model better than physically partitioning**
2. SSD RAID writes suffer significant software overhead
   - Much higher relative overhead than w. HDD, and higher absolute overhead than w. RAM
   - Mainly caused by synchronization
   - **Shorter write path desirable**
3. SSD performance anomaly common, w. significant magnitude and duration
   - Found in all 6 SSD models tested, both consumer and DC
   - **Latency spikes *tall and lasting enough* to be identified and sidestepped at runtime**



Bandwidth consumption in 4-workload mix



RAID write latency breakdown



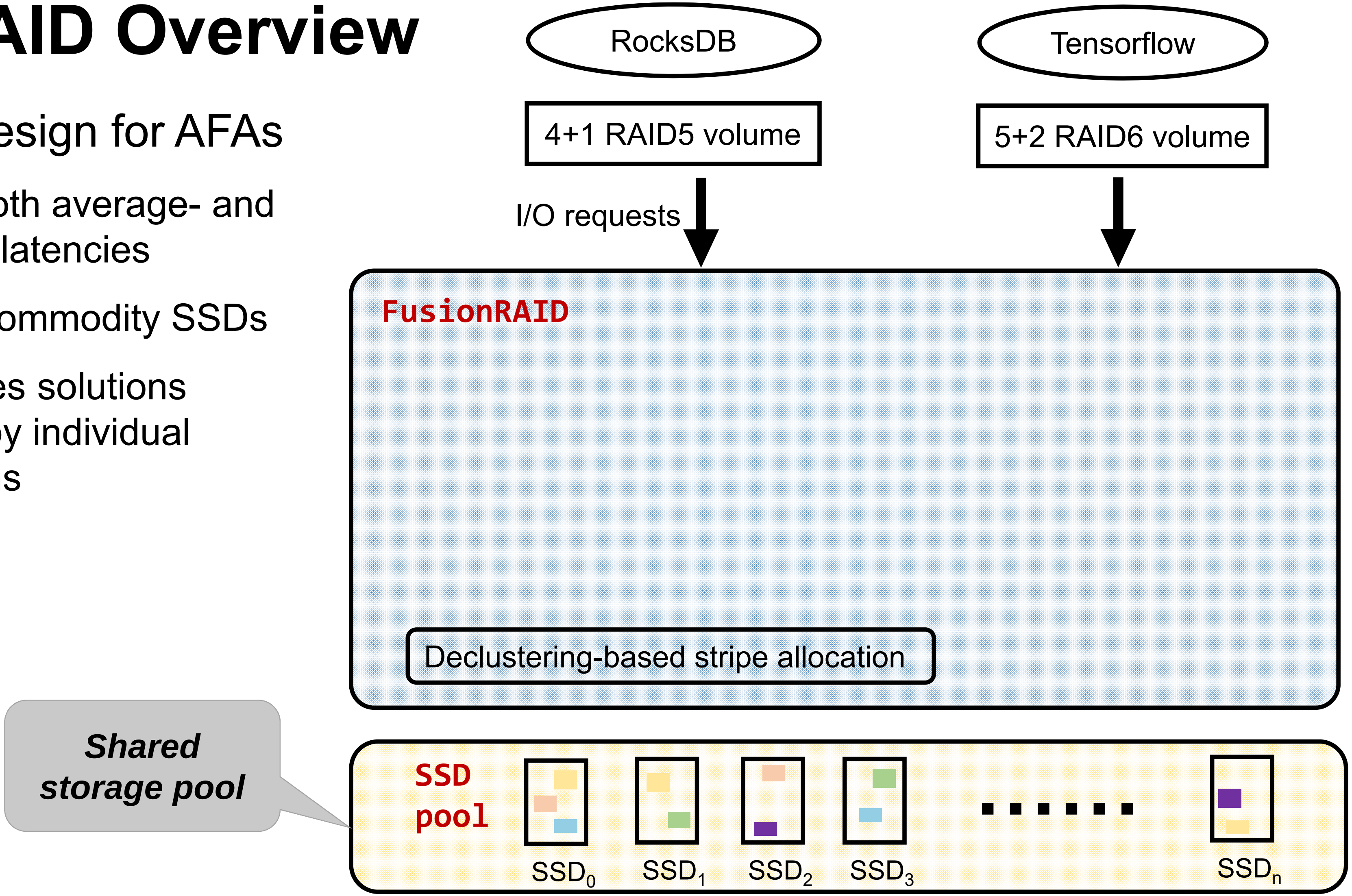
Datacenter SSDs with random writes

FusionRAID Overview

- New RAID design for AFAs
  - Reduces both average- and worst-case latencies
  - Works on commodity SSDs
  - Consolidates solutions motivated by individual observations

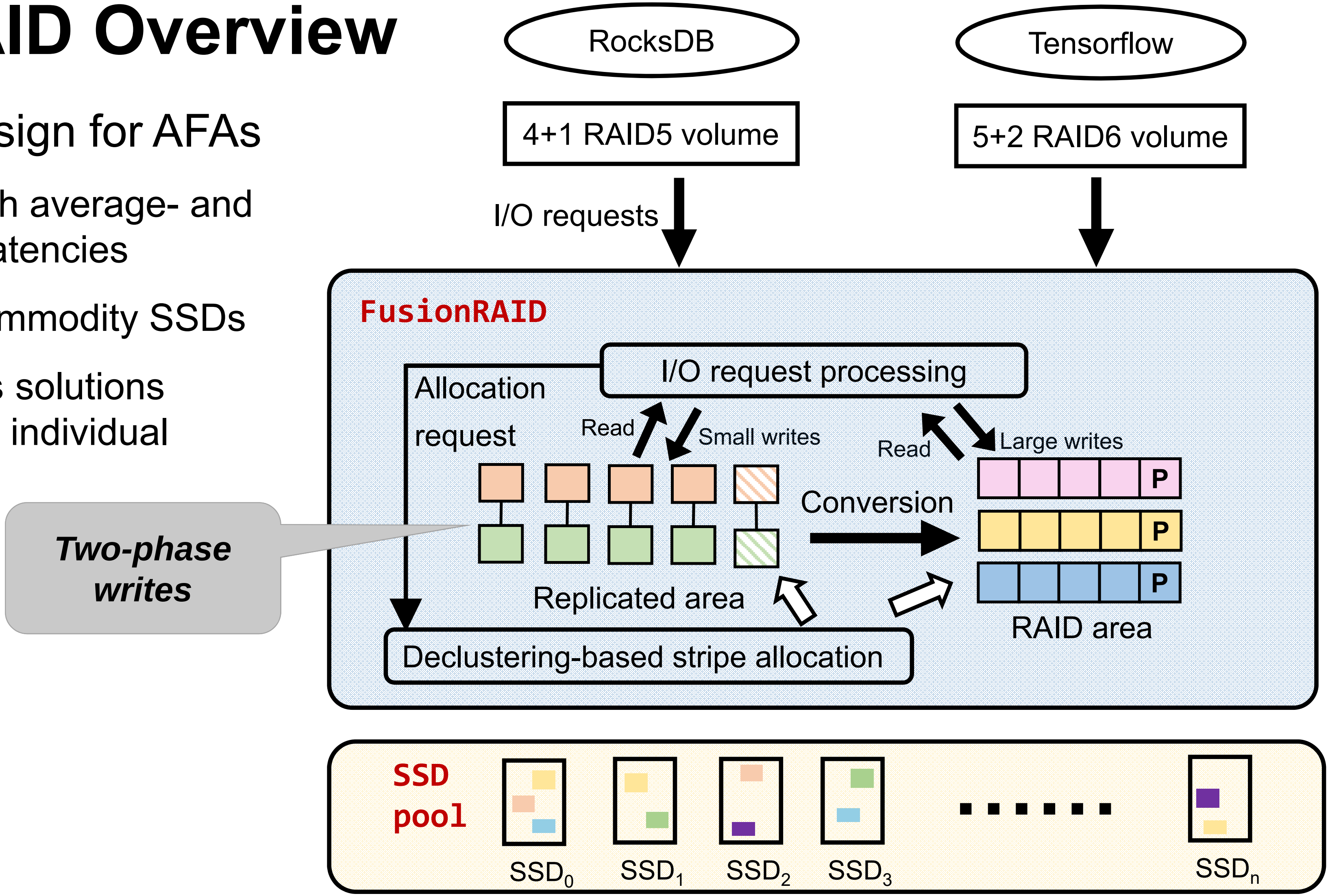
FusionRAID Overview

- New RAID design for AFAs
  - Reduces both average- and worst-case latencies
  - Works on commodity SSDs
  - Consolidates solutions motivated by individual observations



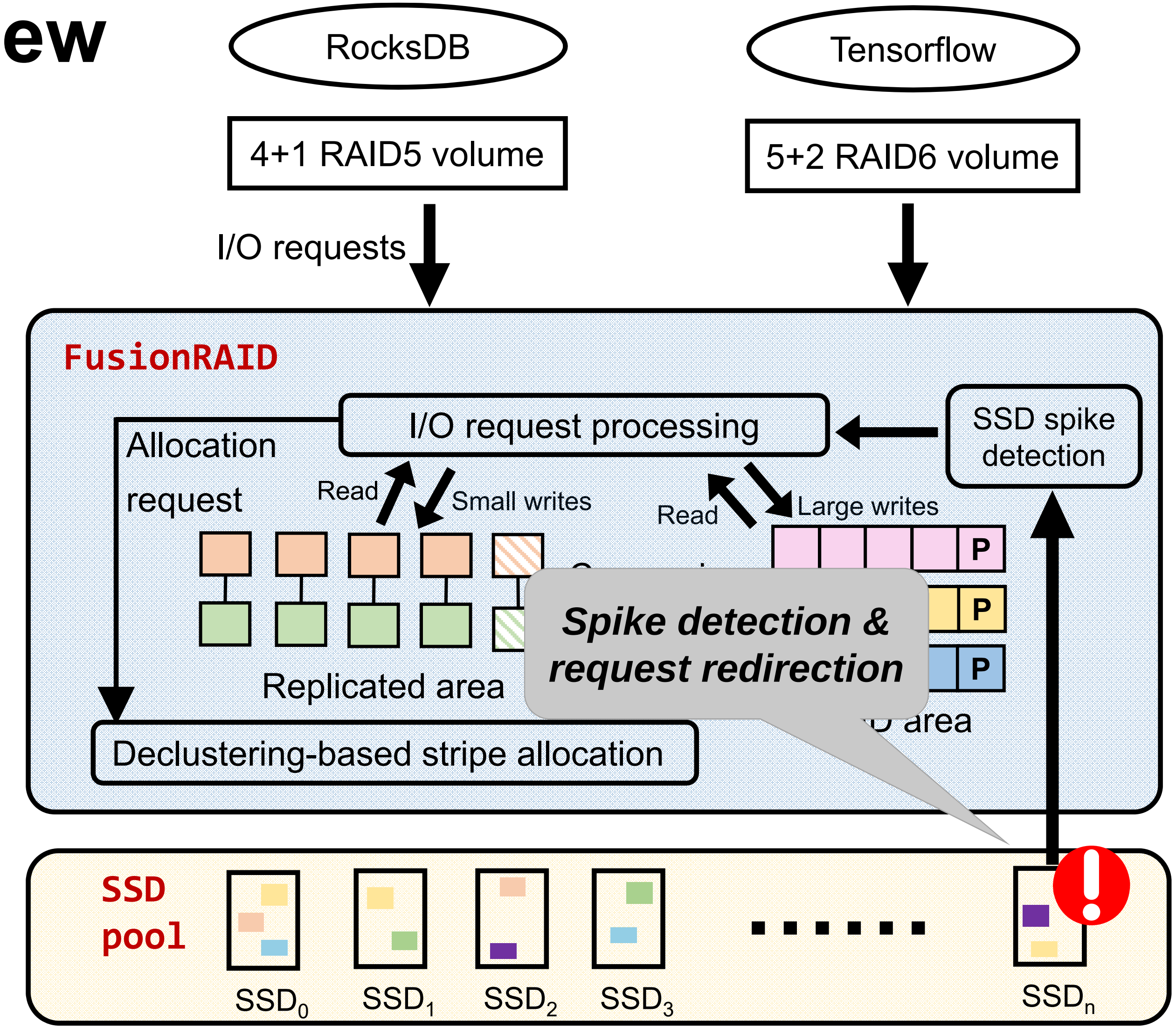
FusionRAID Overview

- New RAID design for AFAs
  - Reduces both average- and worst-case latencies
  - Works on commodity SSDs
  - Consolidates solutions motivated by individual observations



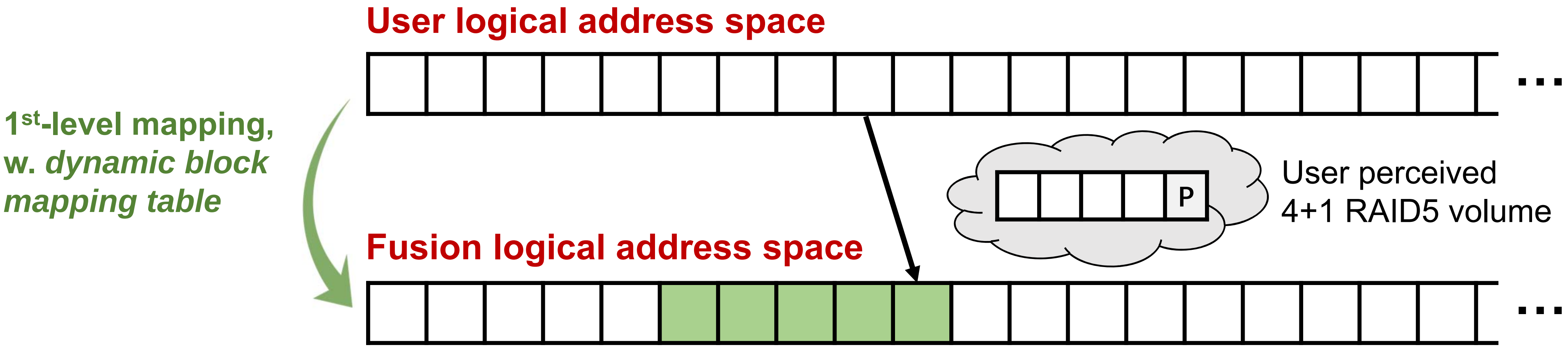
FusionRAID Overview

- New RAID design for AFAs
  - Reduces both average- and worst-case latencies
  - Works on commodity SSDs
  - Consolidates solutions motivated by individual observations

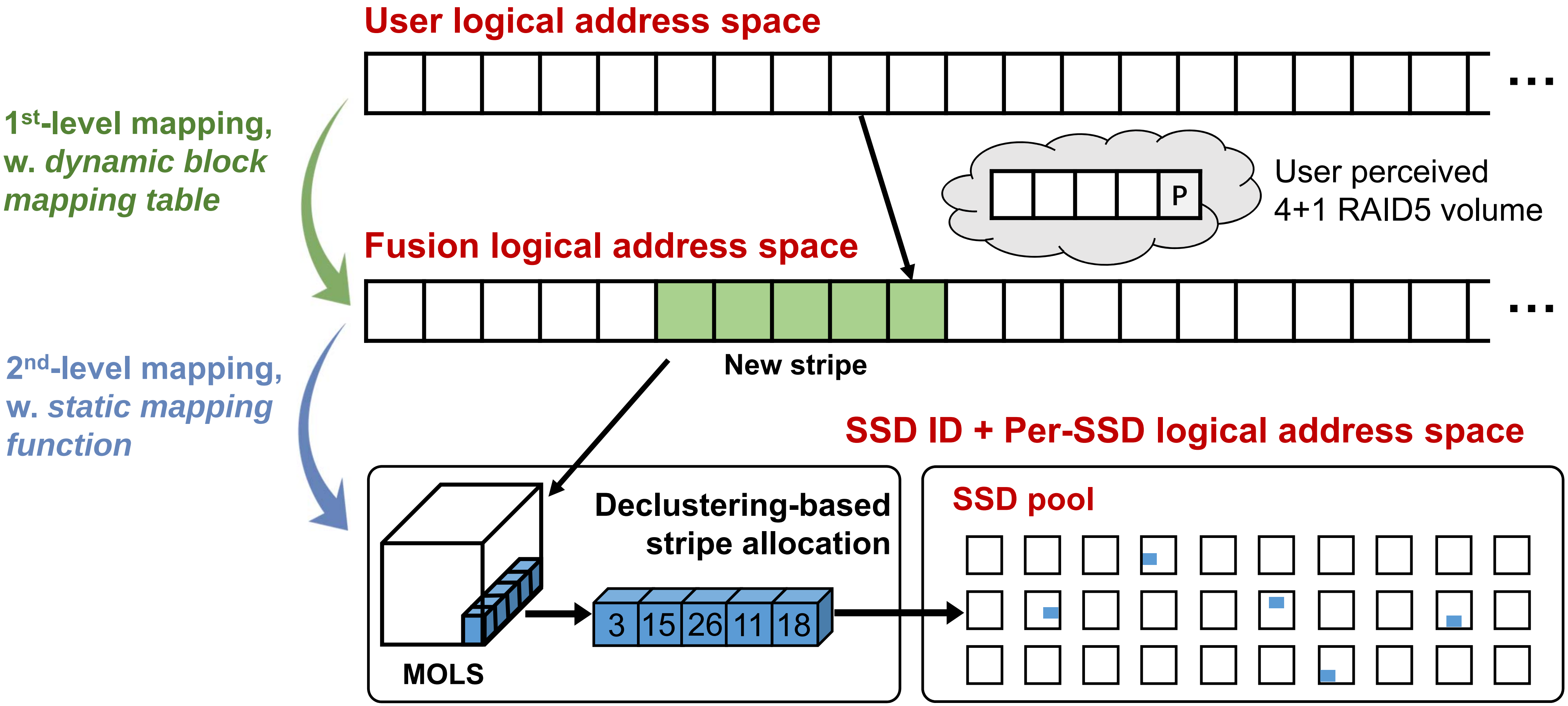


Shared Storage Pool

Shared Storage Pool

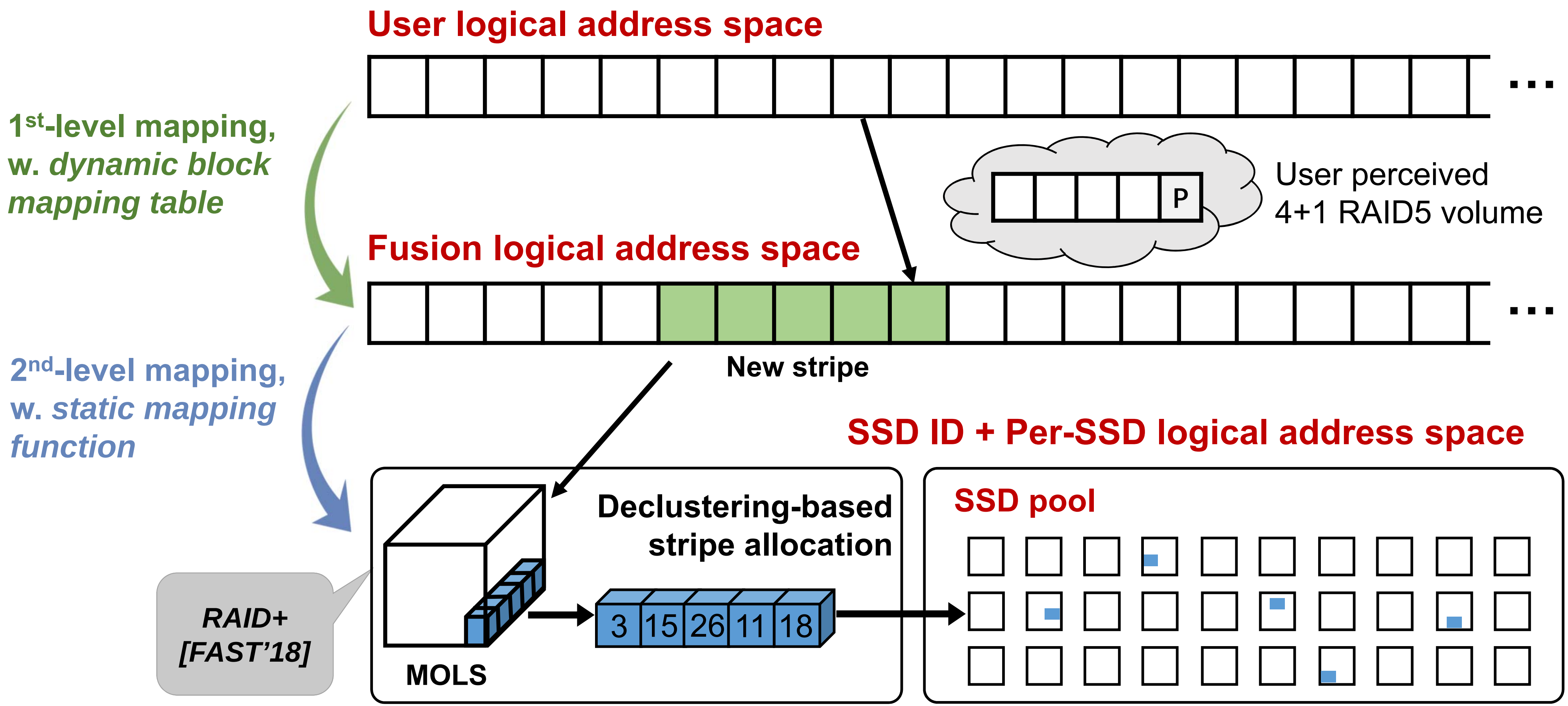

User logical address space
...
1st-level mapping,
w. *dynamic block mapping table*
P
User perceived
4+1 RAID5 volume
Fusion logical address space
...

Shared Storage Pool

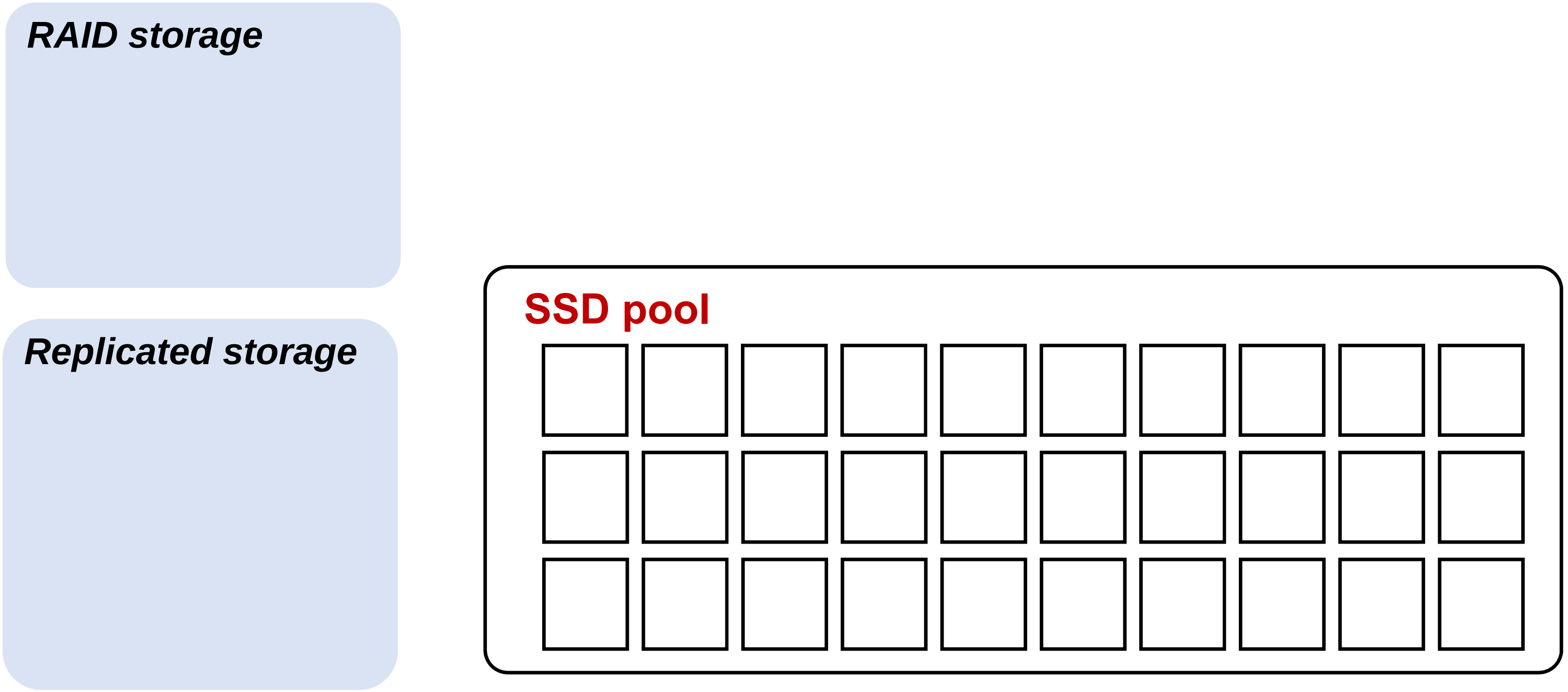


User logical address space
…
1st-level mapping,
w. *dynamic block mapping table*
User perceived
4+1 RAID5 volume
P
Fusion logical address space
…
New stripe
2nd-level mapping,
w. *static mapping function*
SSD ID + Per-SSD logical address space
Declustering-based
stripe allocation
MOLS
3
15
26
11
18
SSD pool

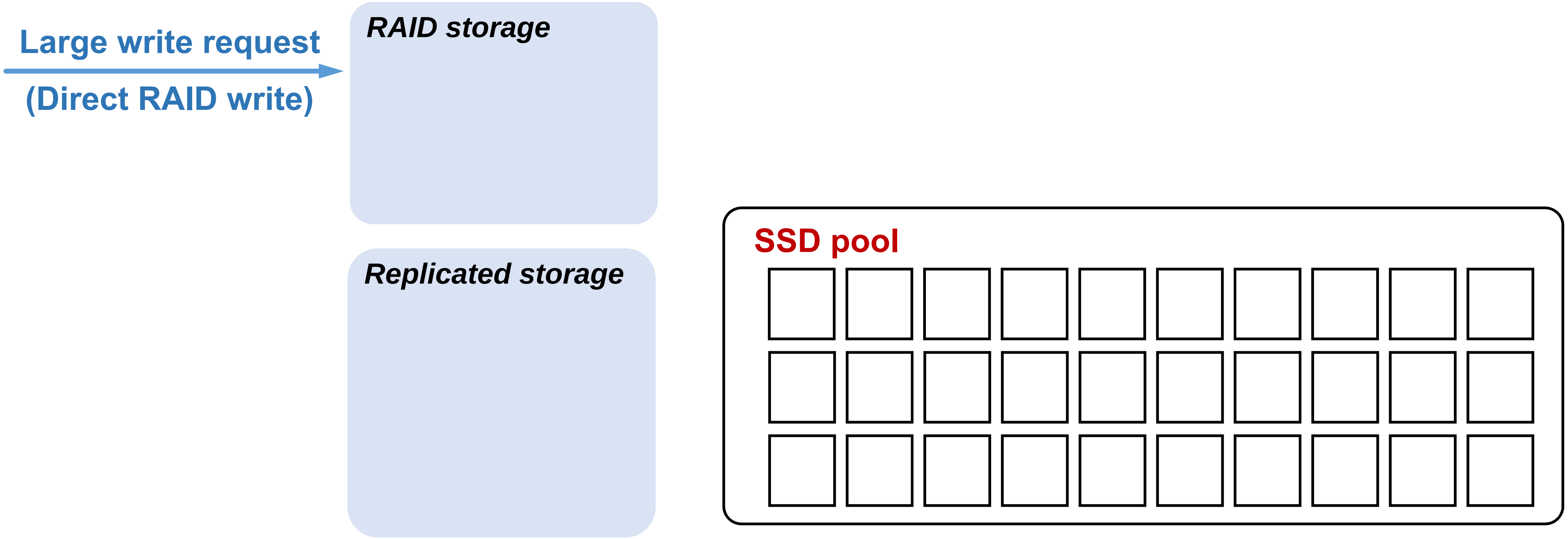
Shared Storage Pool


User logical address space
...
1st-level mapping, w. *dynamic block mapping table*
P
User perceived 4+1 RAID5 volume
Fusion logical address space
...
New stripe
2nd-level mapping, w. *static mapping function*
SSD ID + Per-SSD logical address space
Declustering-based stripe allocation
SSD pool
RAID+ [FAST'18]
3
15
26
11
18
MOLS

FusionRAID Optimized Writes



FusionRAID Optimized Writes



Large write request

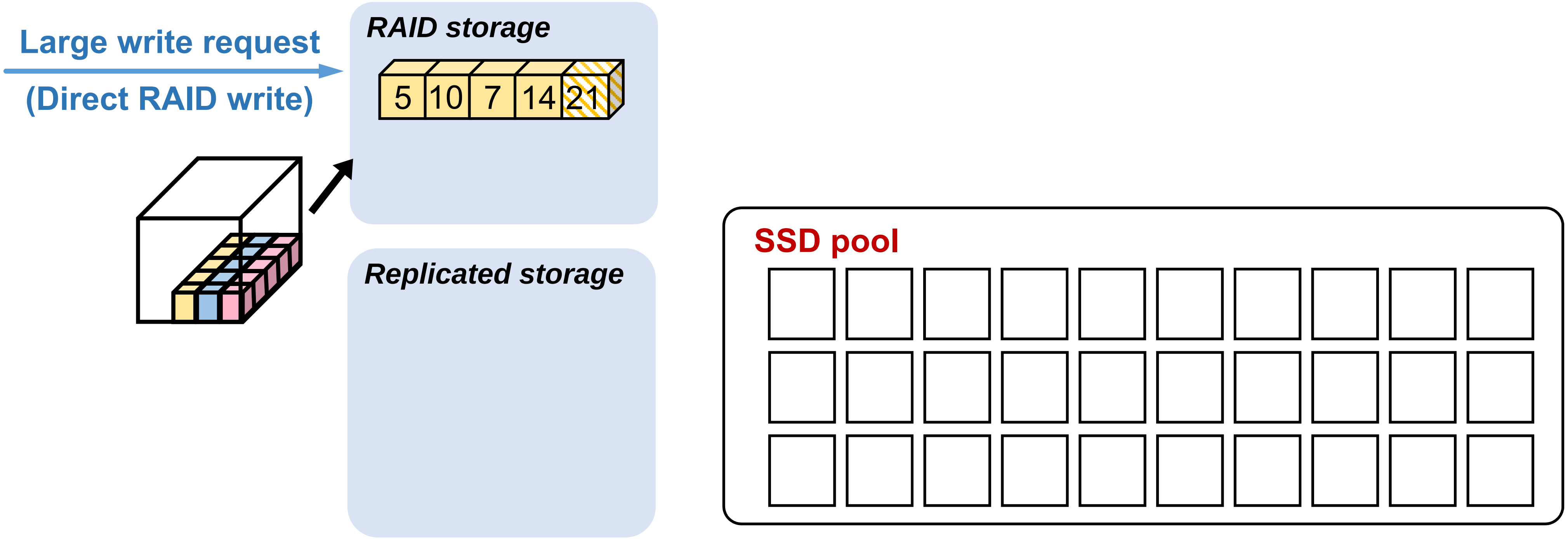
(Direct RAID write)

RAID storage

Replicated storage

SSD pool

FusionRAID Optimized Writes



Large write request

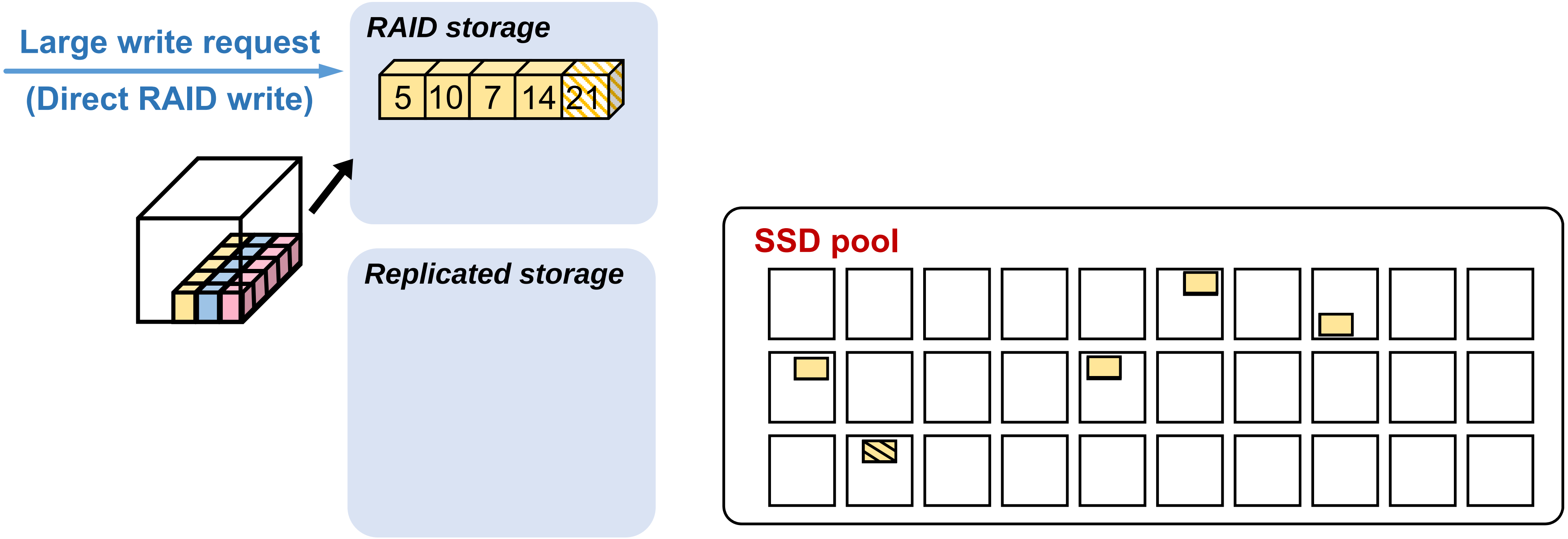
(Direct RAID write)

RAID storage

Replicated storage

SSD pool

FusionRAID Optimized Writes



Large write request

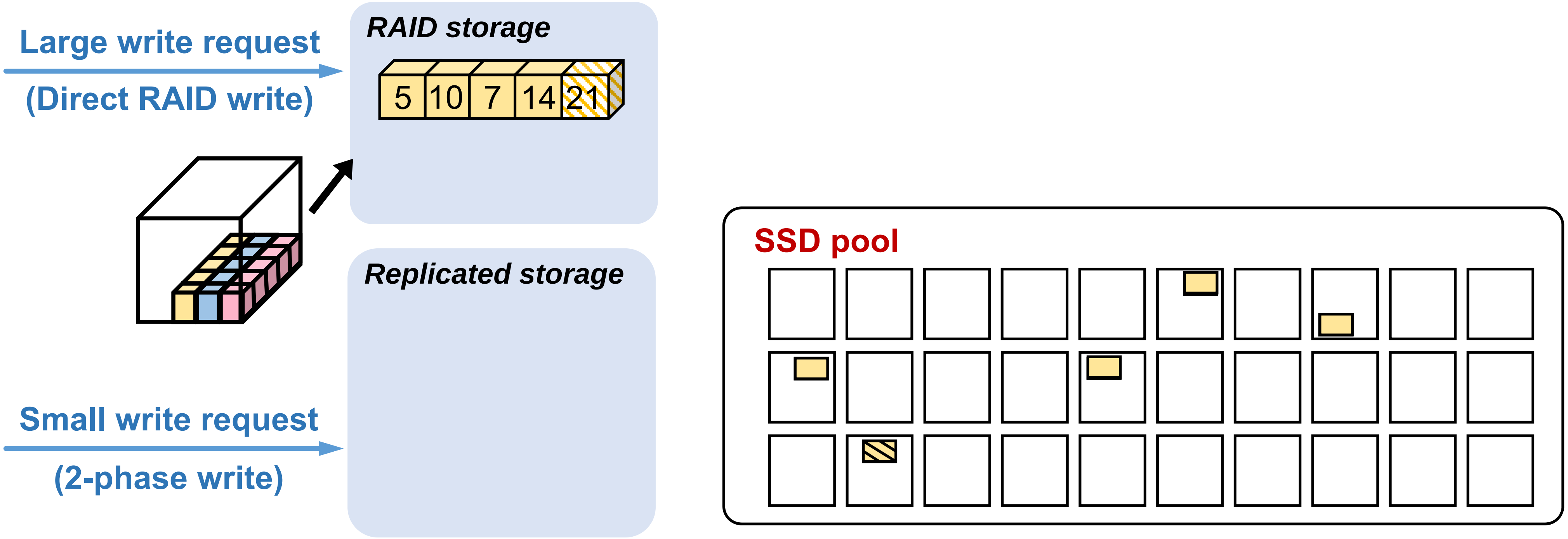
(Direct RAID write)

RAID storage

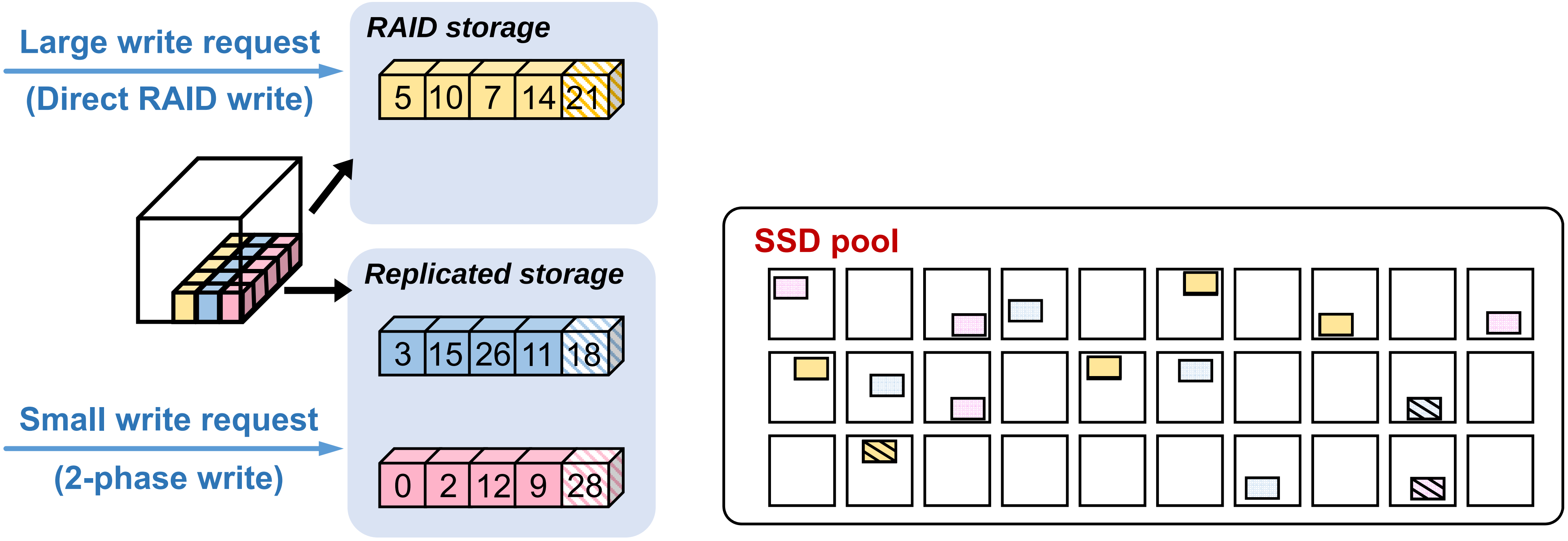
Replicated storage

SSD pool

FusionRAID Optimized Writes



FusionRAID Optimized Writes



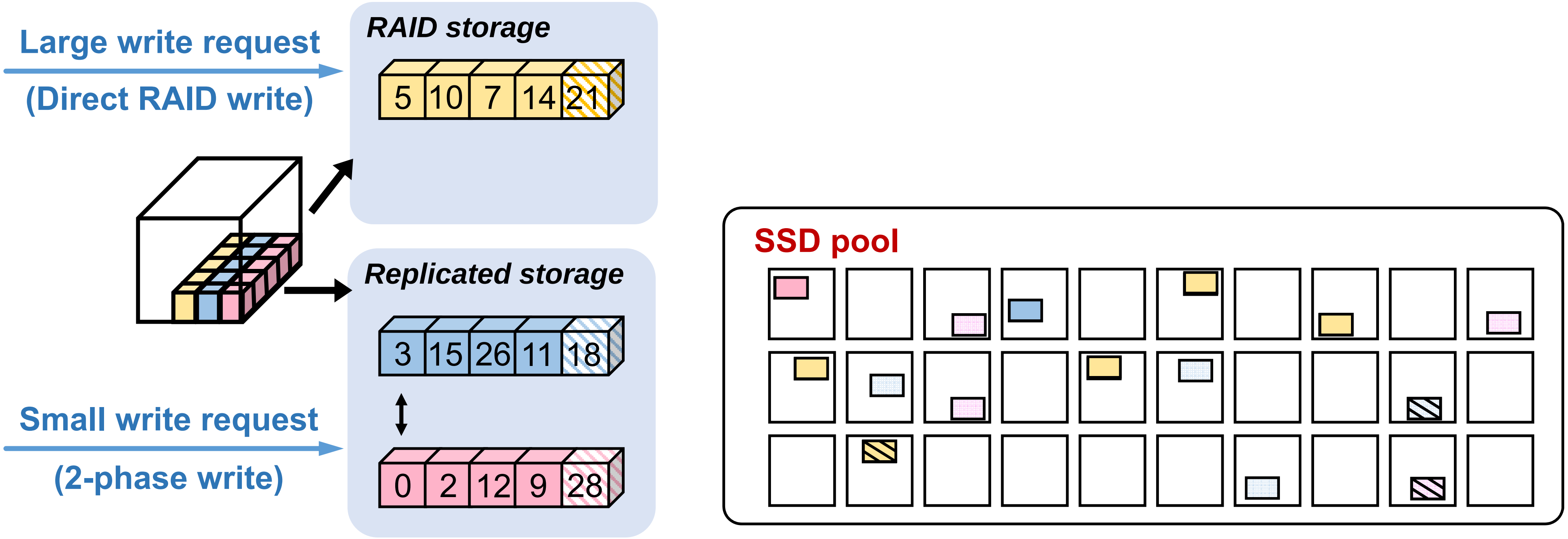
Large write request

(Direct RAID write)

Small write request

(2-phase write)

FusionRAID Optimized Writes



Large write request

(Direct RAID write)

RAID storage

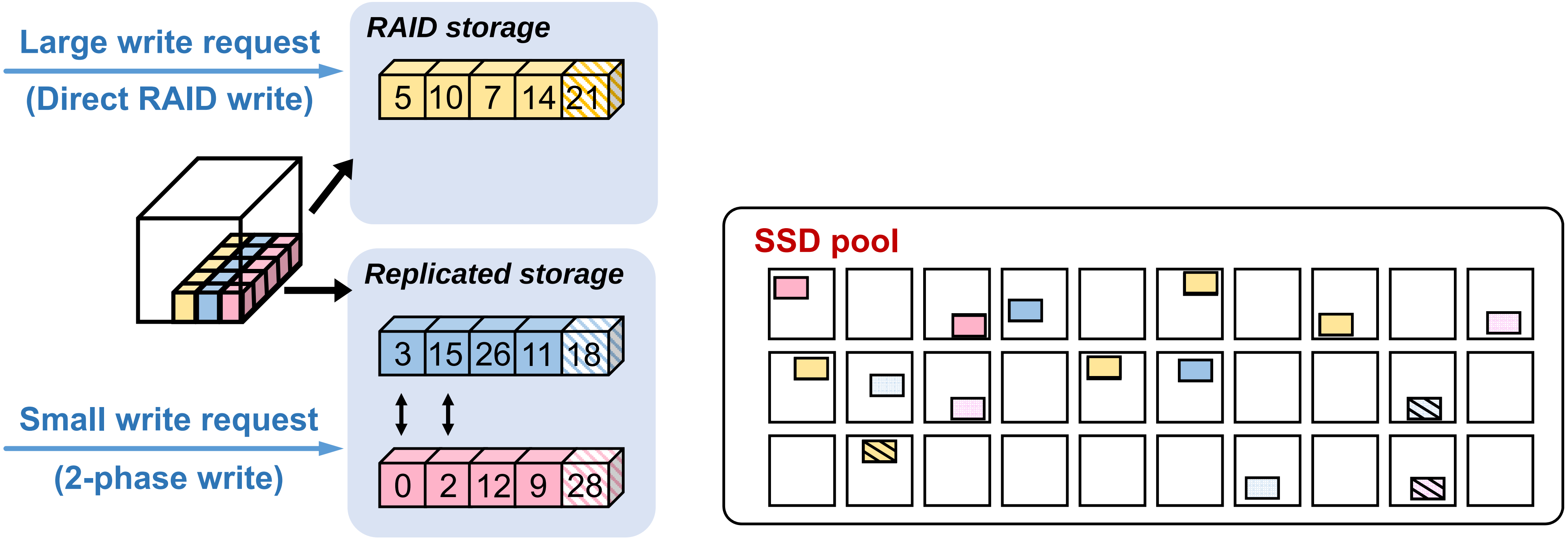
Replicated storage

Small write request

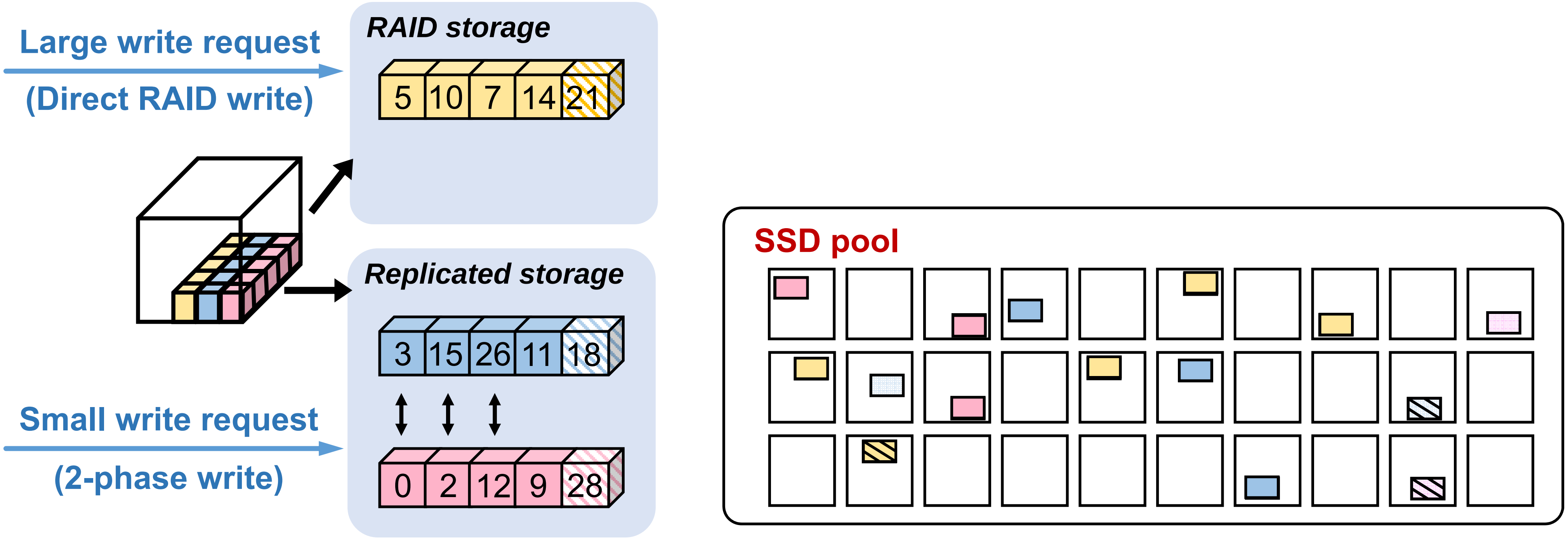
(2-phase write)

SSD pool

FusionRAID Optimized Writes



FusionRAID Optimized Writes



Large write request

(Direct RAID write)

Small write request

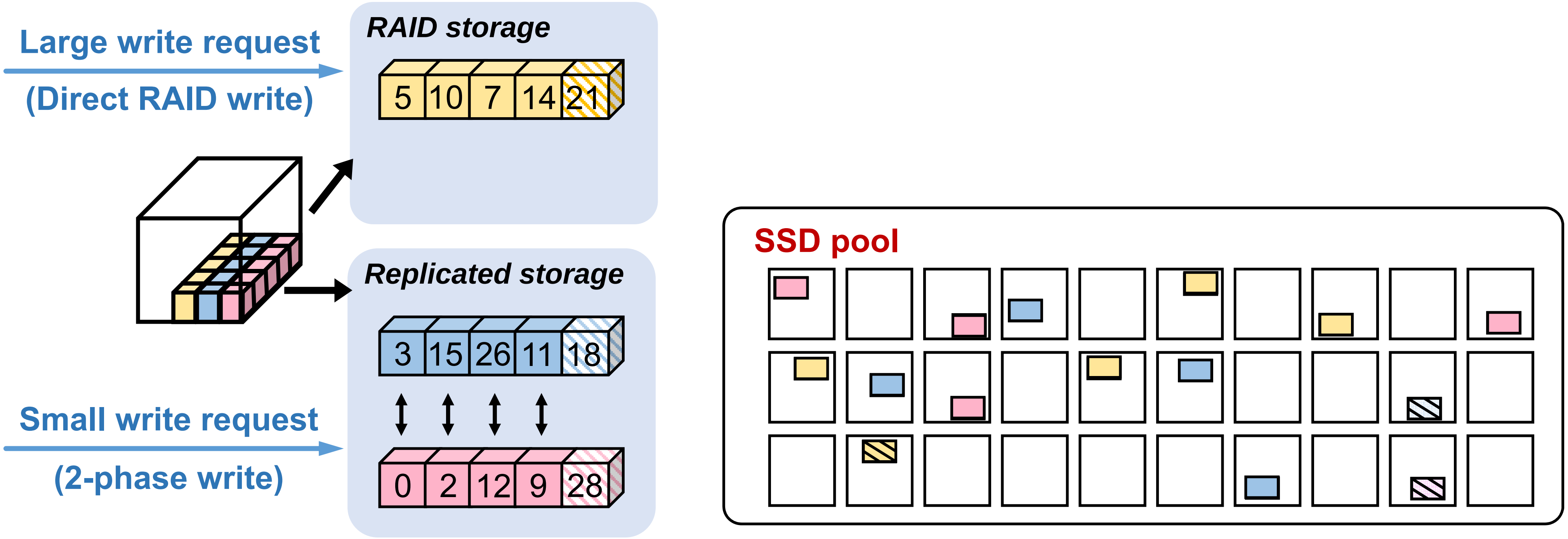
(2-phase write)

RAID storage

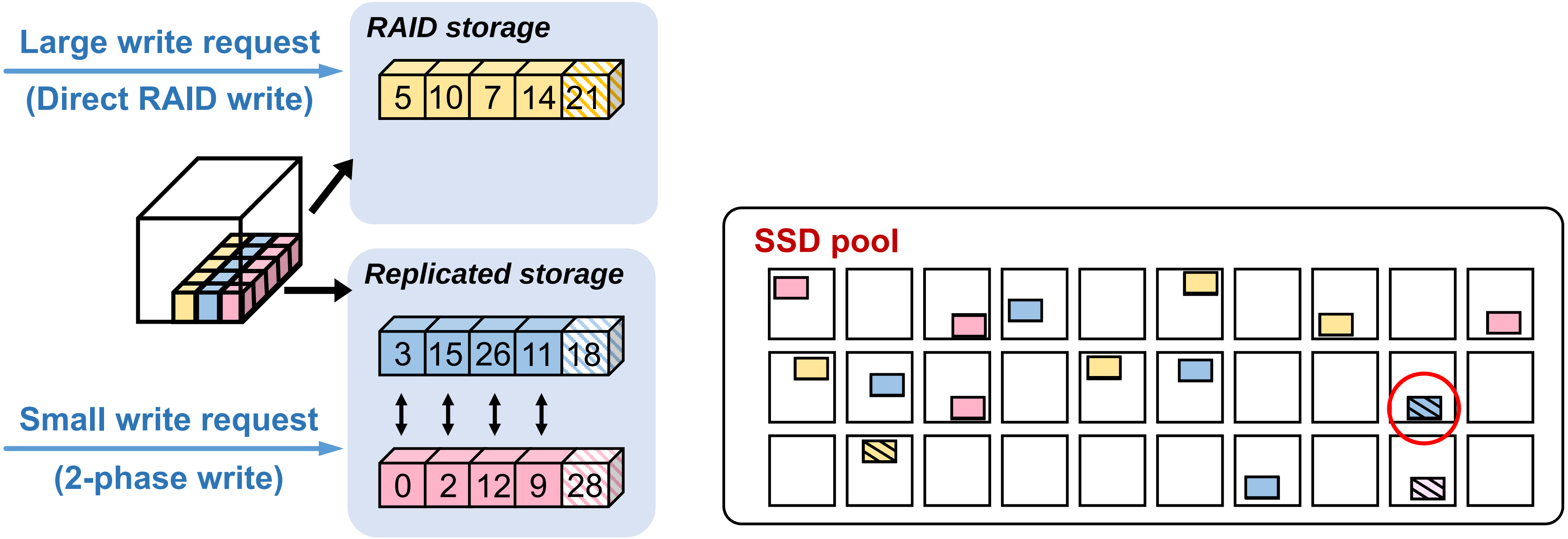
Replicated storage

SSD pool

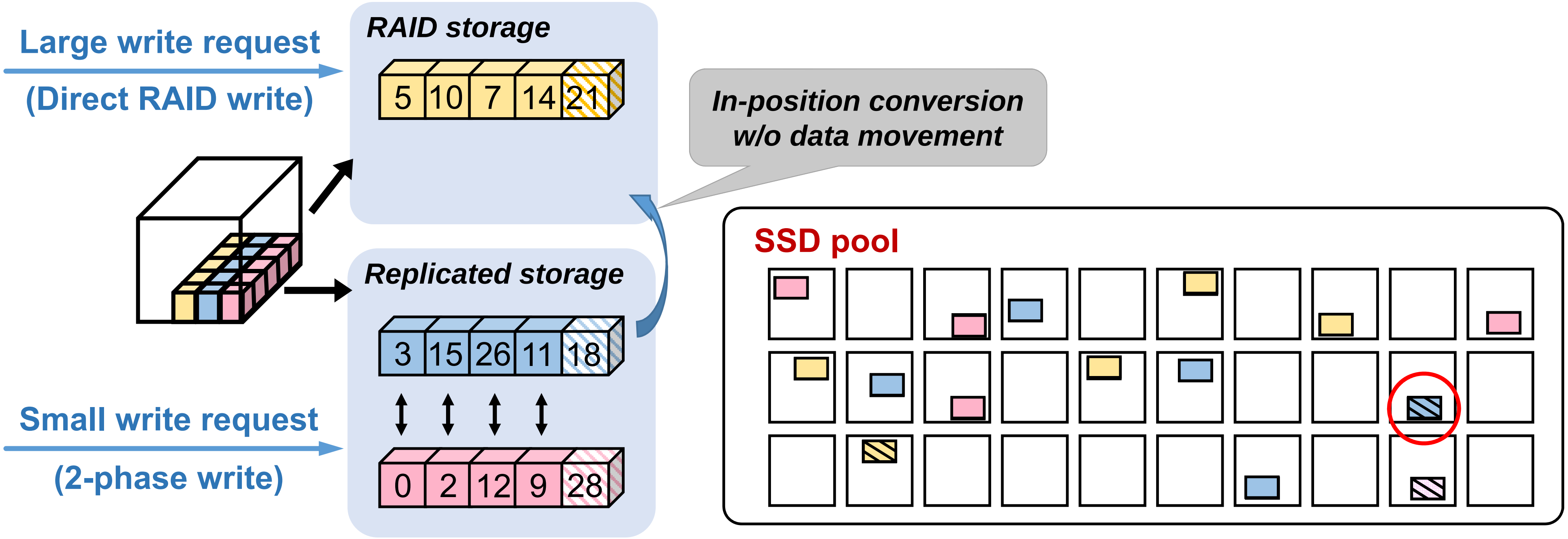
FusionRAID Optimized Writes



FusionRAID Optimized Writes



FusionRAID Optimized Writes



Large write request

(Direct RAID write)

Small write request

(2-phase write)

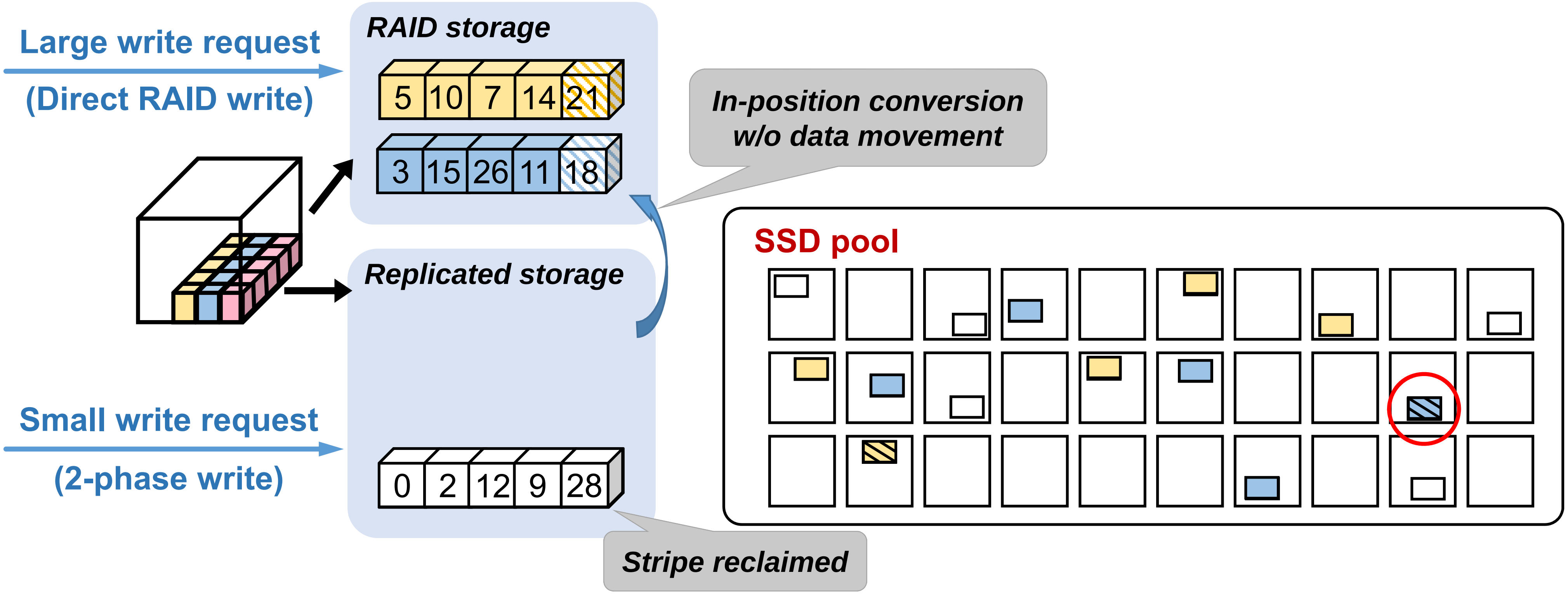
RAID storage

Replicated storage

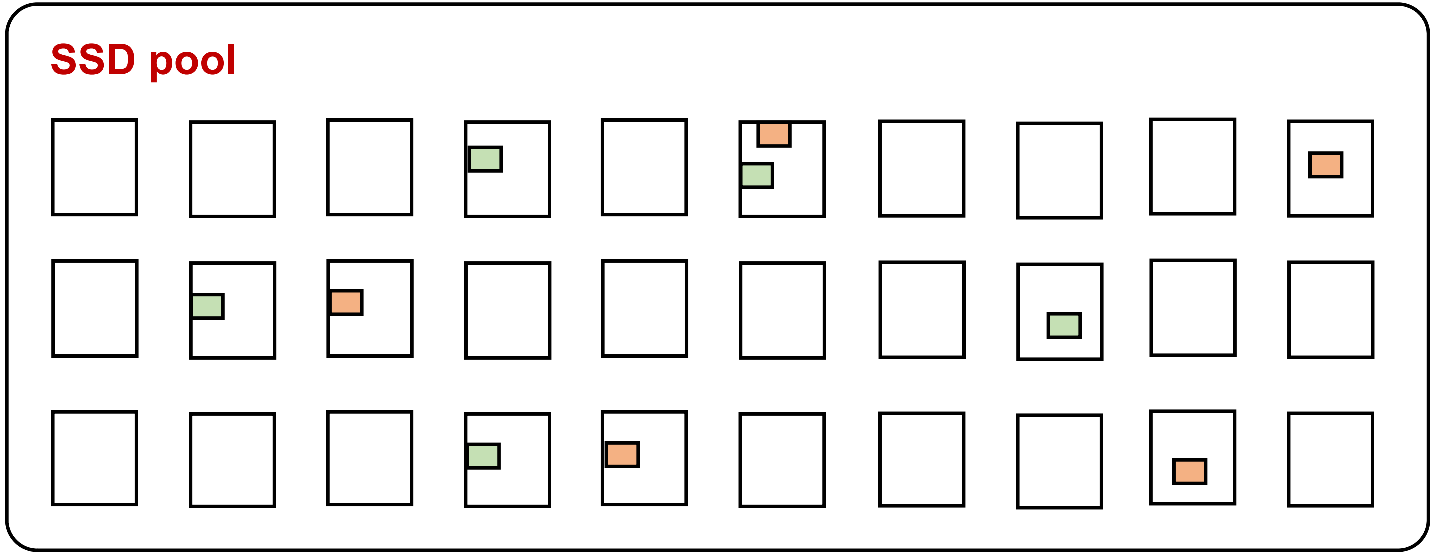
In-position conversion w/o data movement

SSD pool

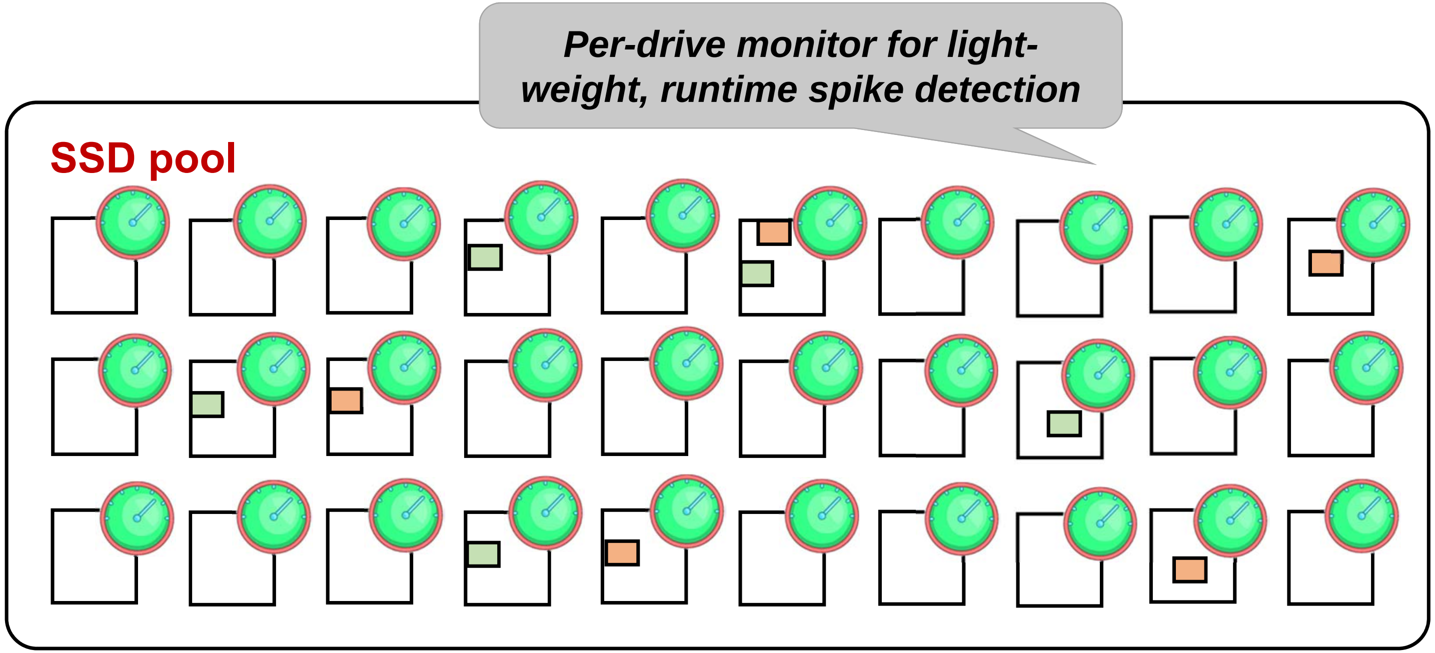
FusionRAID Optimized Writes



Spike Detection and Request Redirection

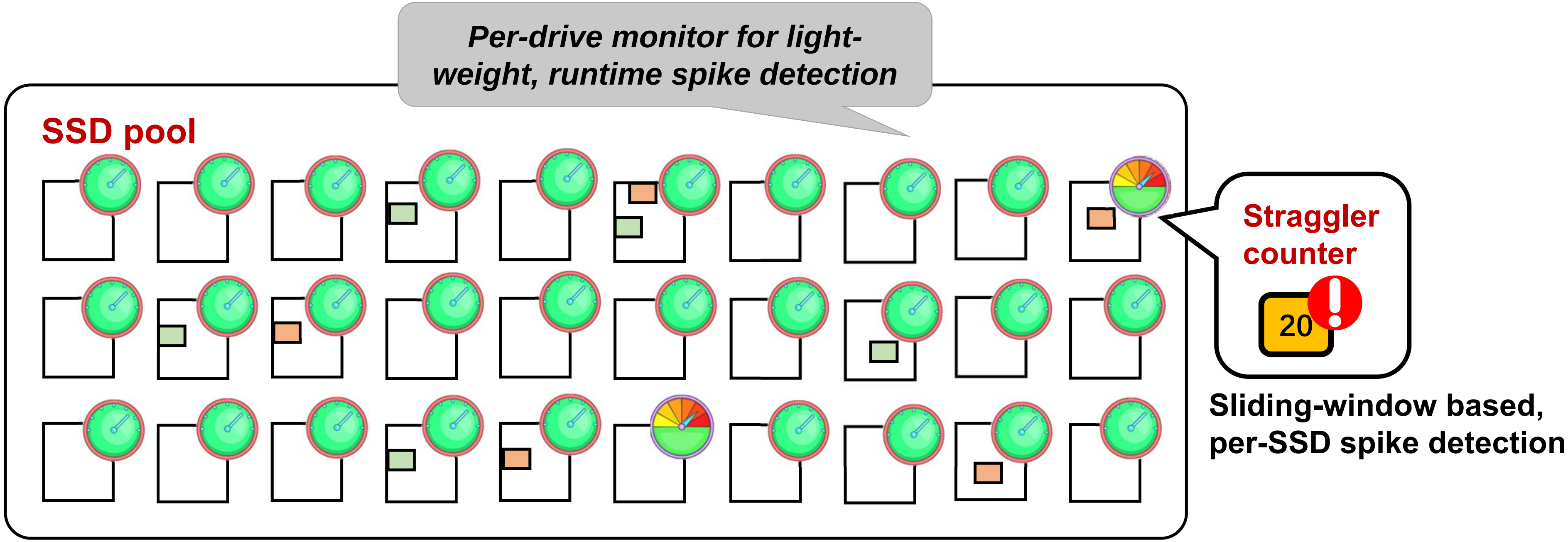


Spike Detection and Request Redirection



Per-drive monitor for light-weight, runtime spike detection

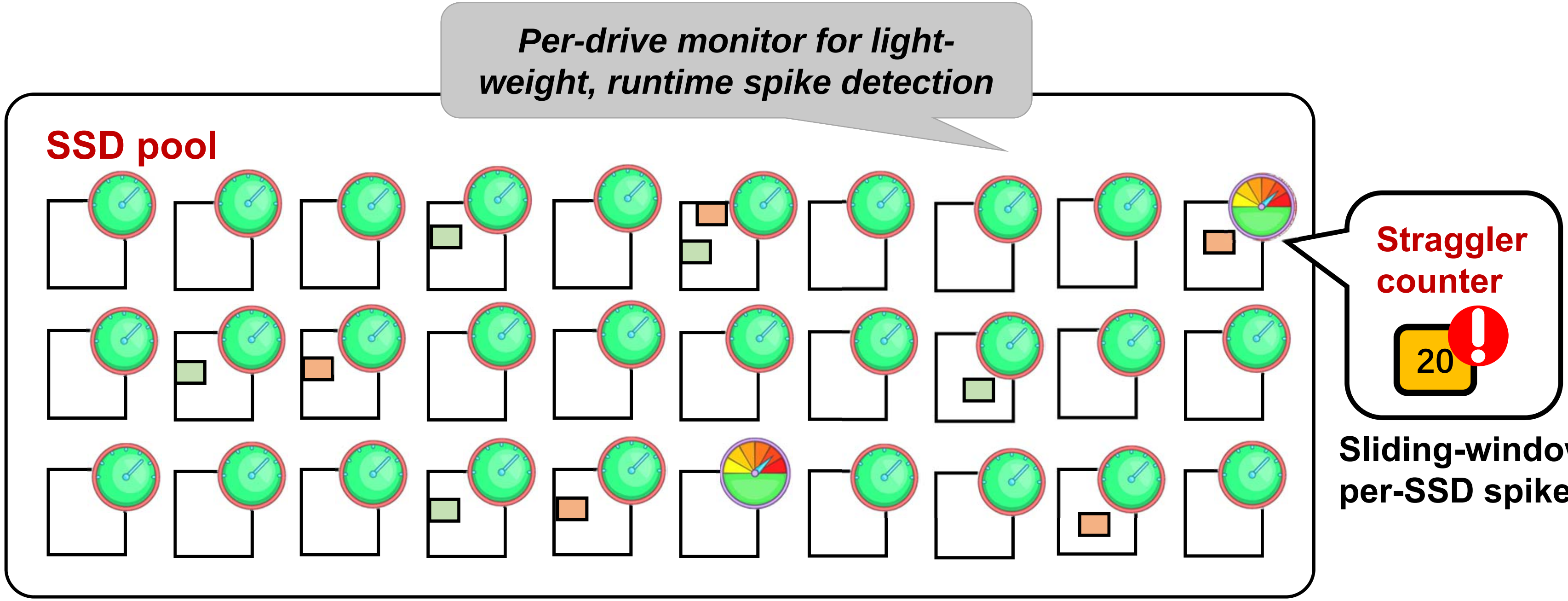
Spike Detection and Request Redirection



Per-drive monitor for light-weight, runtime spike detection

Sliding-window based, per-SSD spike detection

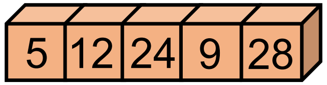
Spike Detection and Request Redirection



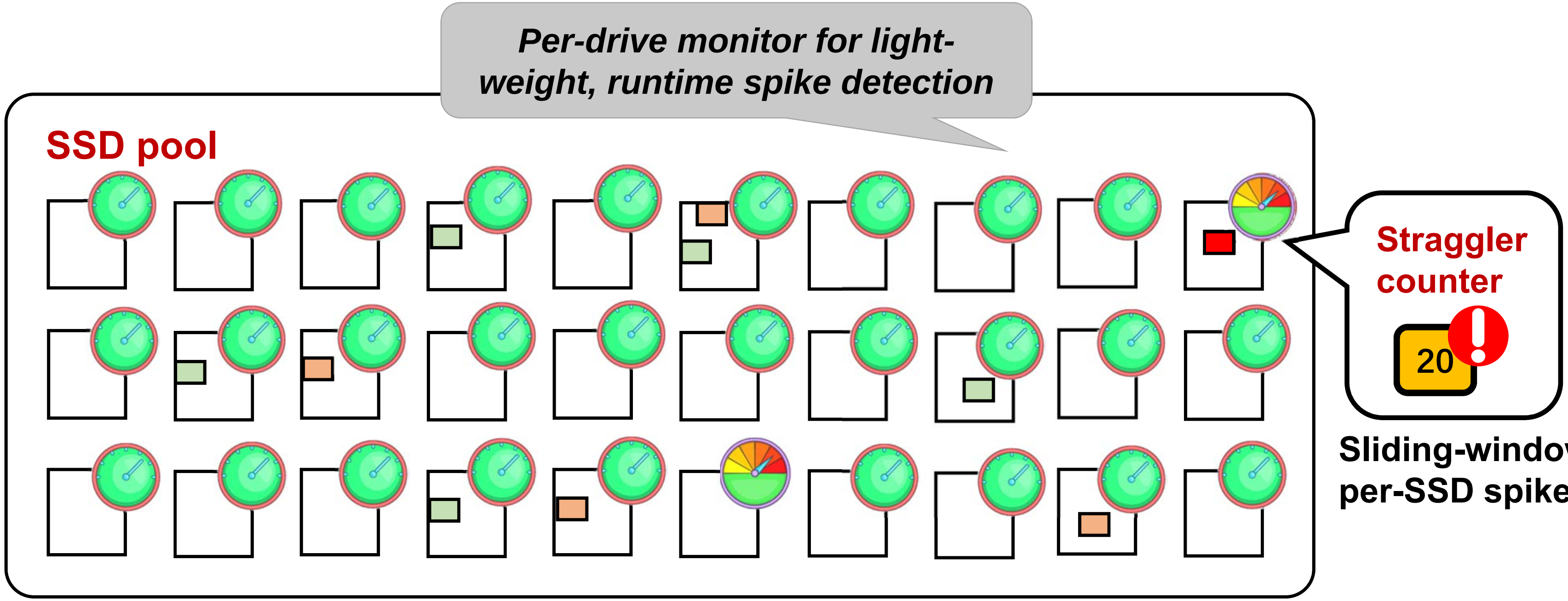
Per-drive monitor for light-weight, runtime spike detection

Sliding-window based, per-SSD spike detection

Reactive request redirection



Spike Detection and Request Redirection



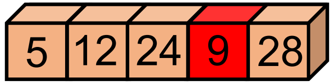
Per-drive monitor for light-weight, runtime spike detection

SSD pool

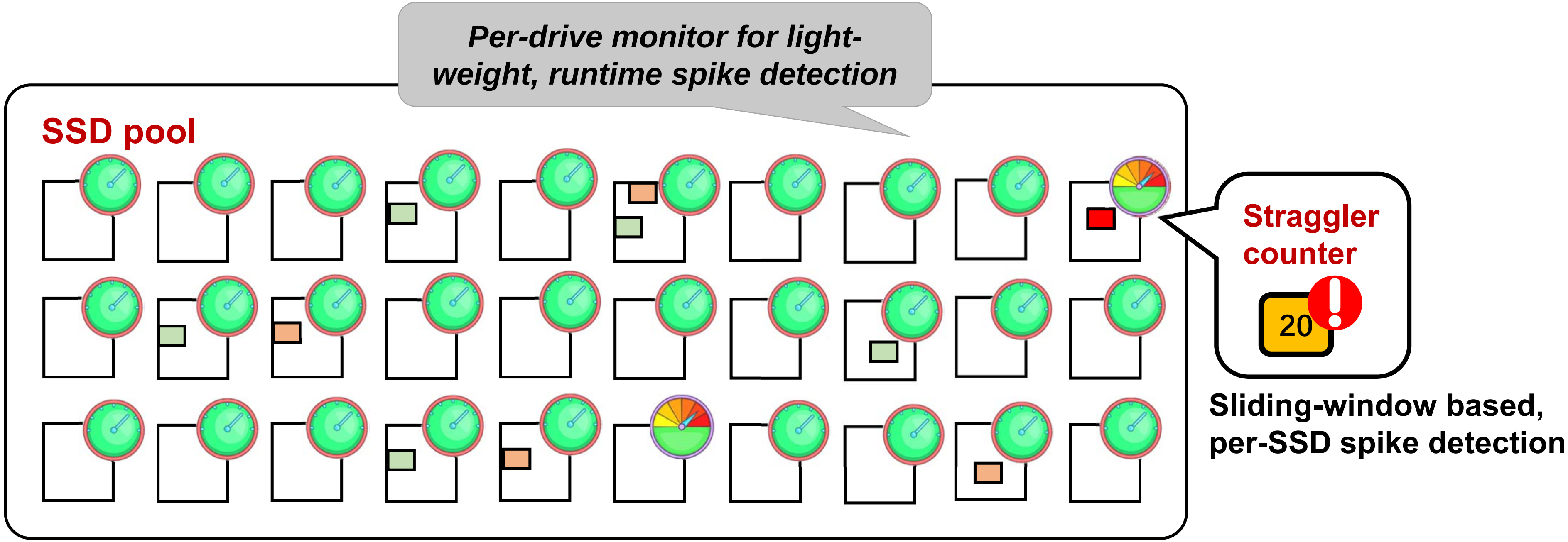
Straggler counter

Sliding-window based, per-SSD spike detection

Reactive request redirection



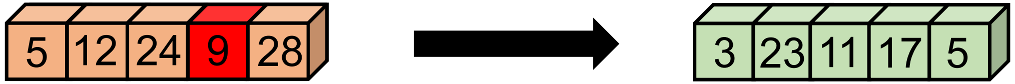
Spike Detection and Request Redirection



Per-drive monitor for light-weight, runtime spike detection

Sliding-window based, per-SSD spike detection

Reactive request redirection



Evaluation Overview

Evaluation Overview

- Testbed

CPU	2 Intel Xeon E5-2650 V4
DRAM	128 GB
SSD	30 Intel D3-S4510
OS	Ubuntu 16.04, Linux kernel 4.15.0

Evaluation Overview

- Testbed

CPU	2 Intel Xeon E5-2650 V4
DRAM	128 GB
SSD	30 Intel D3-S4510
OS	Ubuntu 16.04, Linux kernel 4.15.0

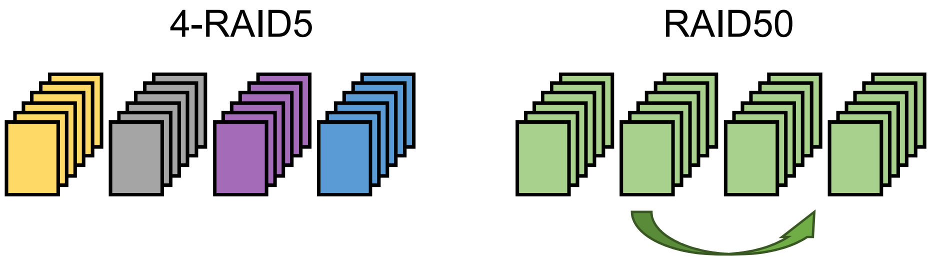
- Benchmark
  - **Trace-driven** workloads
  - **Real application** (YCSB + RocksDB)

Evaluation Overview

- Testbed

CPU	2 Intel Xeon E5-2650 V4
DRAM	128 GB
SSD	30 Intel D3-S4510
OS	Ubuntu 16.04, Linux kernel 4.15.0

- Benchmark
  - **Trace-driven** workloads
  - **Real application** (YCSB + RocksDB)



- Systems
  - **Commercial RAID**: 4-RAID5, RAID50
  - **Latest RAID in paper**: ToleRAID (FAST’16), LogRAID (SYSTOR’14, ATC’19)

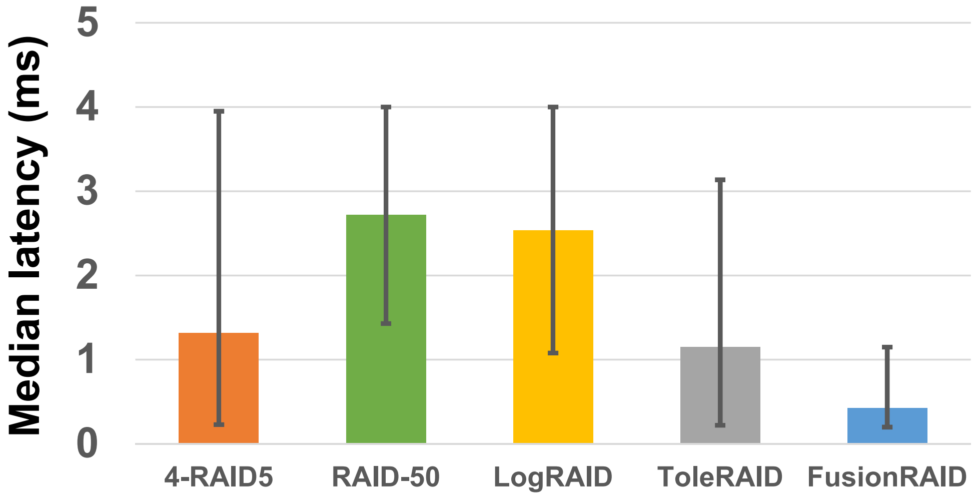
Evaluation: Trace-driven Workloads

Evaluation: Trace-driven Workloads

- Running 4-workload mixes on compared RAID systems
- Randomly selected 20 mixes from 8 storage workloads

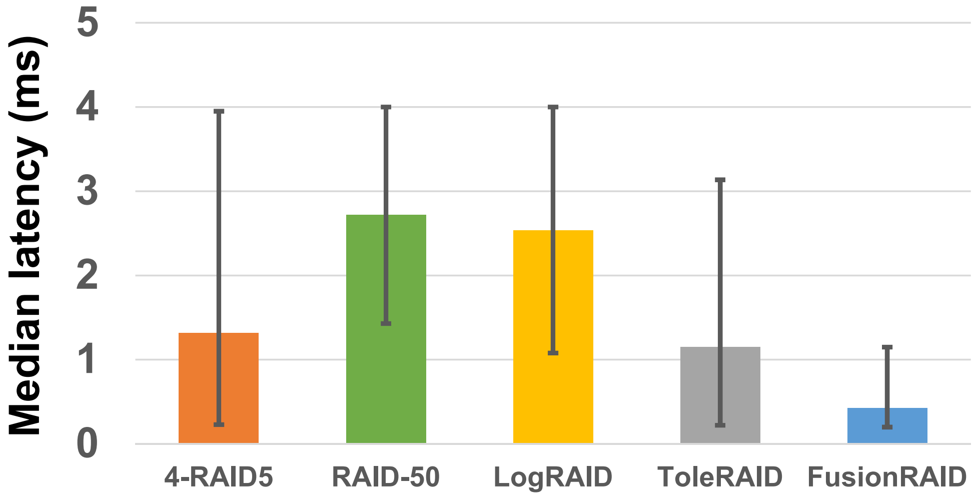
Evaluation: Trace-driven Workloads

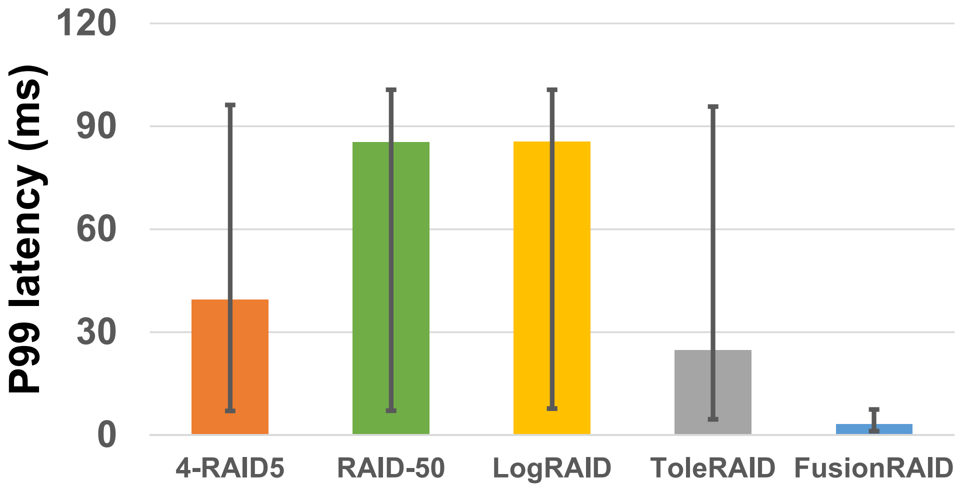
- Running 4-workload mixes on compared RAID systems
- Randomly selected 20 mixes from 8 storage workloads



Evaluation: Trace-driven Workloads

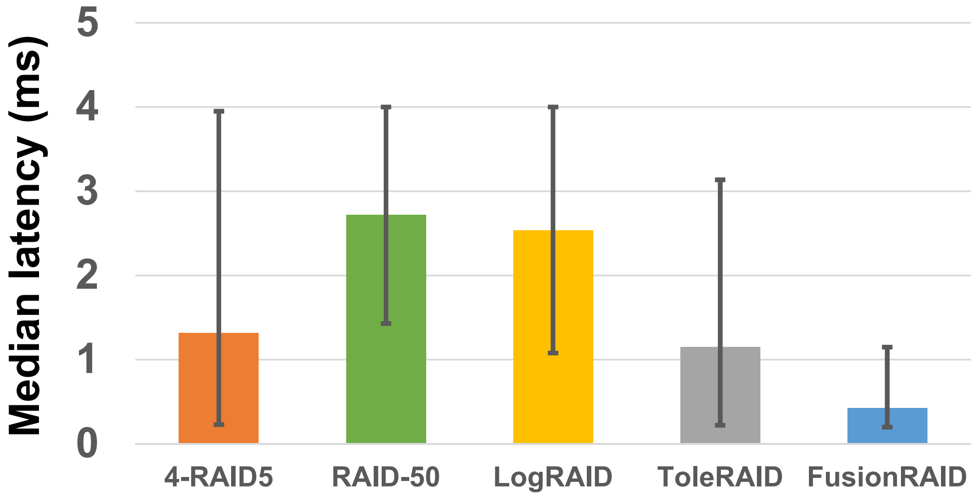
- Running 4-workload mixes on compared RAID systems
- Randomly selected 20 mixes from 8 storage workloads

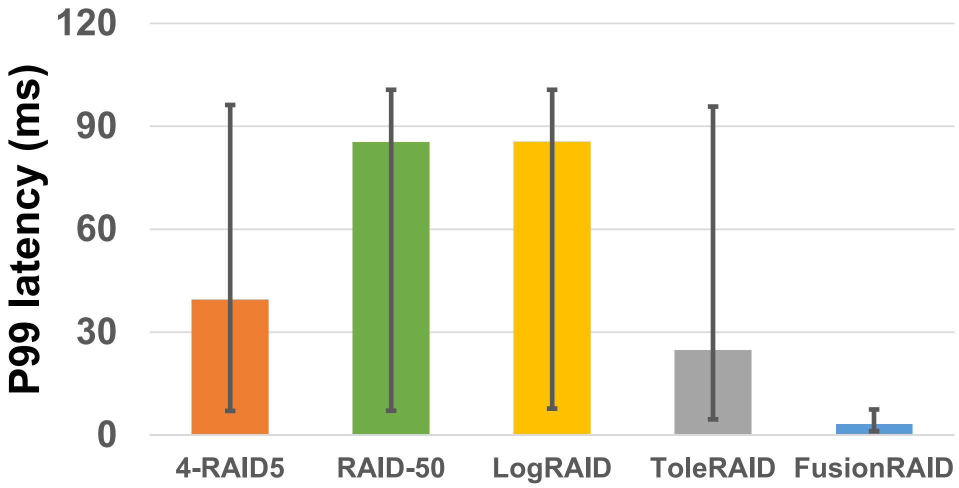




Evaluation: Trace-driven Workloads

- Running 4-workload mixes on compared RAID systems
- Randomly selected 20 mixes from 8 storage workloads





FusionRAID reduces **median latency by 45%~81%**
and **P99 latency by 8.3×~35×!**

Evaluation: Applications and FusionRAID Overhead

Evaluation: Applications and FusionRAID Overhead

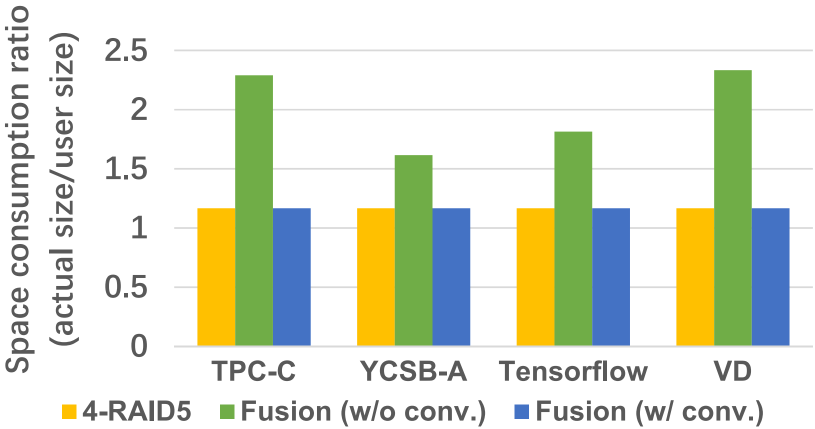
- Real application results
  - Running RocksDB on FusionRAID and RAID50
  - FusionRAID reduces tail latency by **4.1×**

Evaluation: Applications and FusionRAID Overhead

- Real application results
  - Running RocksDB on FusionRAID and RAID50
  - FusionRAID reduces tail latency by **4.1×**
- Conversion only brings **18% increase** in tail latency

Evaluation: Applications and FusionRAID Overhead

- Real application results
  - Running RocksDB on FusionRAID and RAID50
  - FusionRAID reduces tail latency by **4.1×**
- Conversion only brings **18% increase** in tail latency
- FusionRAID without conversion consumes **2×** space within running, and decrease to **1.17×** if conversion on



Space overhead

FusionRAID: Achieving Consistent Low Latency for Commodity SSD Arrays

Thank you!

清華大學 TSINGHUA UNIVERSITY 1911

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY