

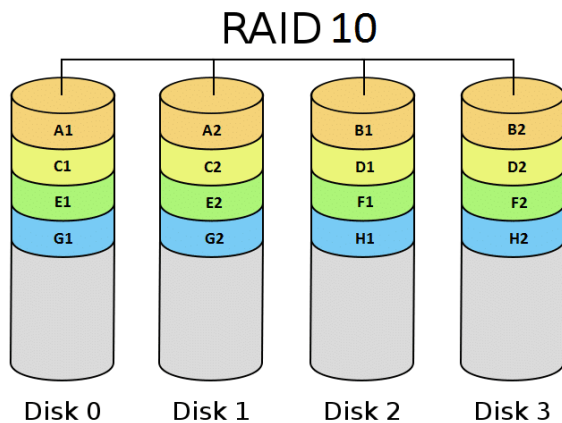
Re-think the Design of RAID upon the Arrival of Computational Storage Drives

Jiangpeng Li, ScaleFlux

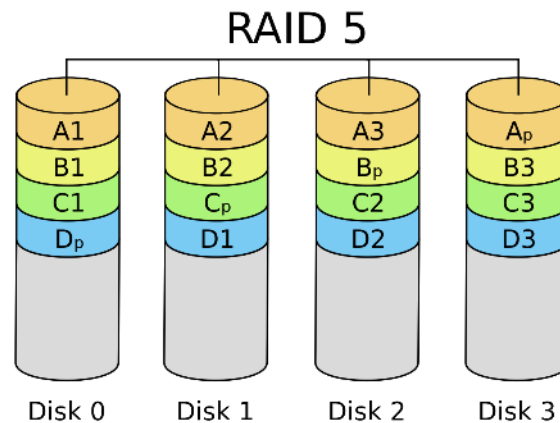
jiangpeng.li@scaleflux.com

RAID: Simple Idea with Big Impact

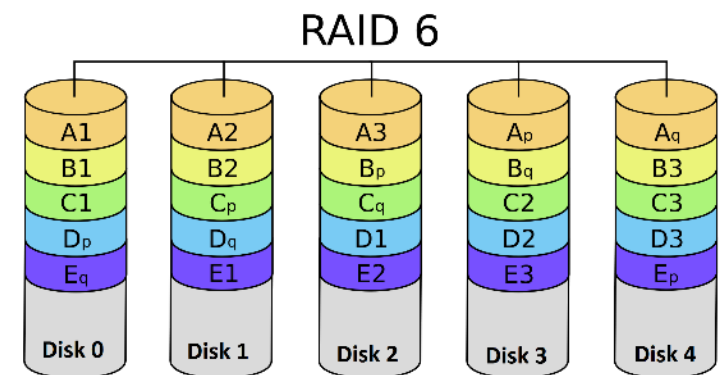
- ❑ A perfect example of “simplest ideas always work the best”
- ❑ Different RAID configurations with different trade-offs
- ❑ Current (painful) deployment practice
 - Users must choose and subsequently **stick** with one RAID configuration, after **painfully** deliberating the pros vs. cons of different options



VS



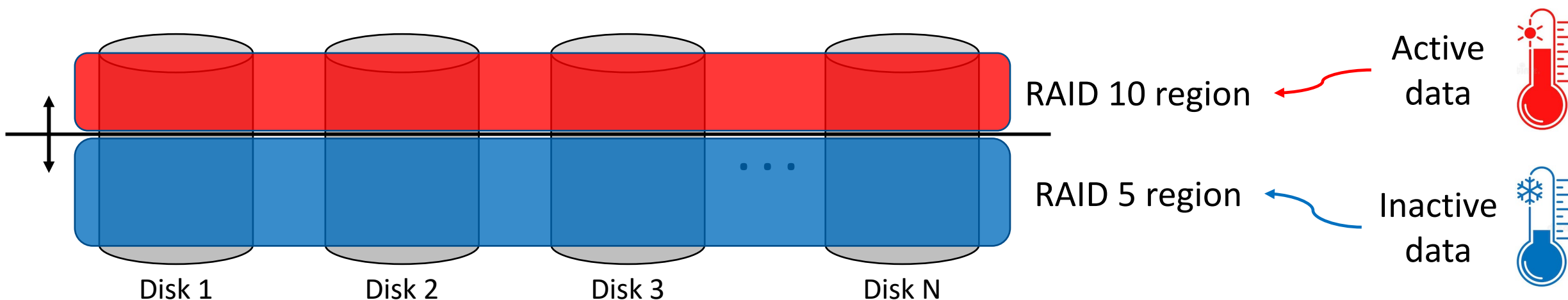
VS



Bring Flexibility into RAID

❑ The famous HP AutoRAID product (1990s~2010s)

- Two-level RAID hierarchy: Active data on RAID 10 & inactive data on RAID 5



Dynamically varying data temperature

User storage capacity vs. performance trade-off

Complicated management & CPU/DRAM overhead

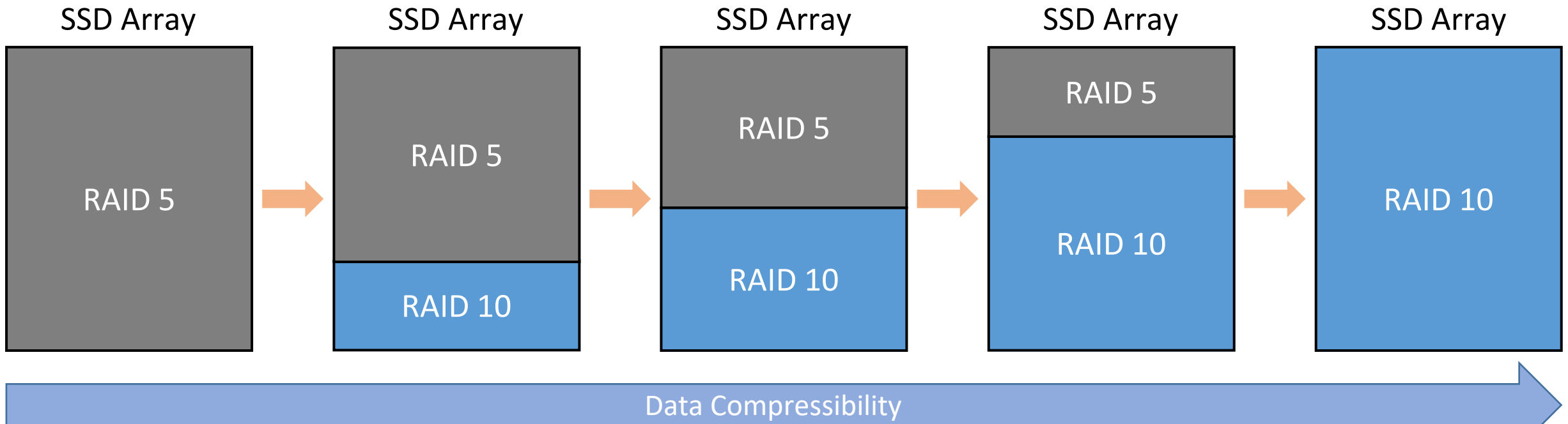
Wilkes, John, et al. "The HP AutoRAID hierarchical storage system." *ACM Transactions on Computer Systems (TOCS)* 14.1 (1996): 108-136.

Bring Flexibility into RAID



Abundant Data Compressibility in the Real World

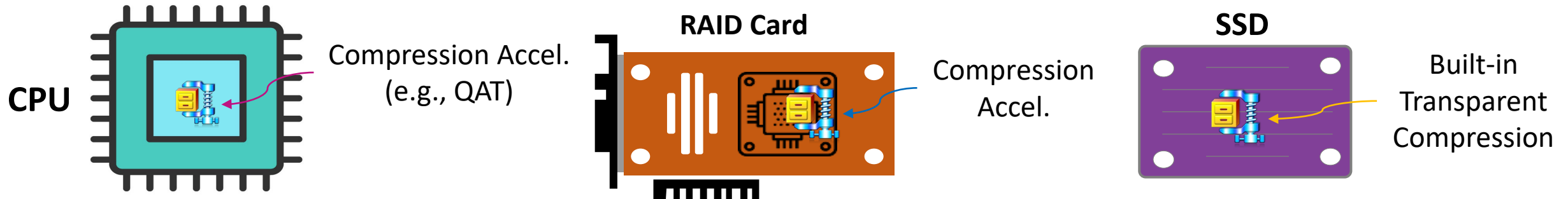
Exploit runtime data compressibility to enable opportunistic RAID 5 → RAID 10 conversion



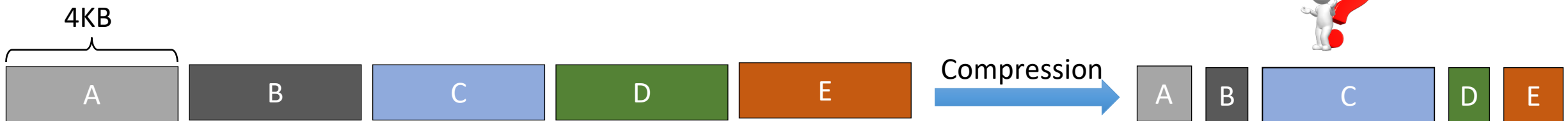
Compression-assisted Elastic RAID

? Realization of data (de)compression

- Ensure highest RAID IOPS → per-4KB (de)compression HW acceleration



? Storage management of compressed variable-length blocks





Compression-assisted Elastic RAID

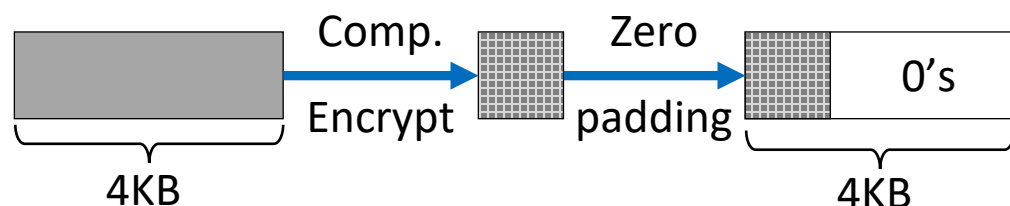
? Storage management of compressed variable-length blocks



Offload the storage management task to SSDs

NVMe SSD

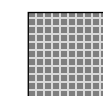
CPU or RAID Card



nvm
EXPRESS

Controller

Zero pruning
& Agile FTL

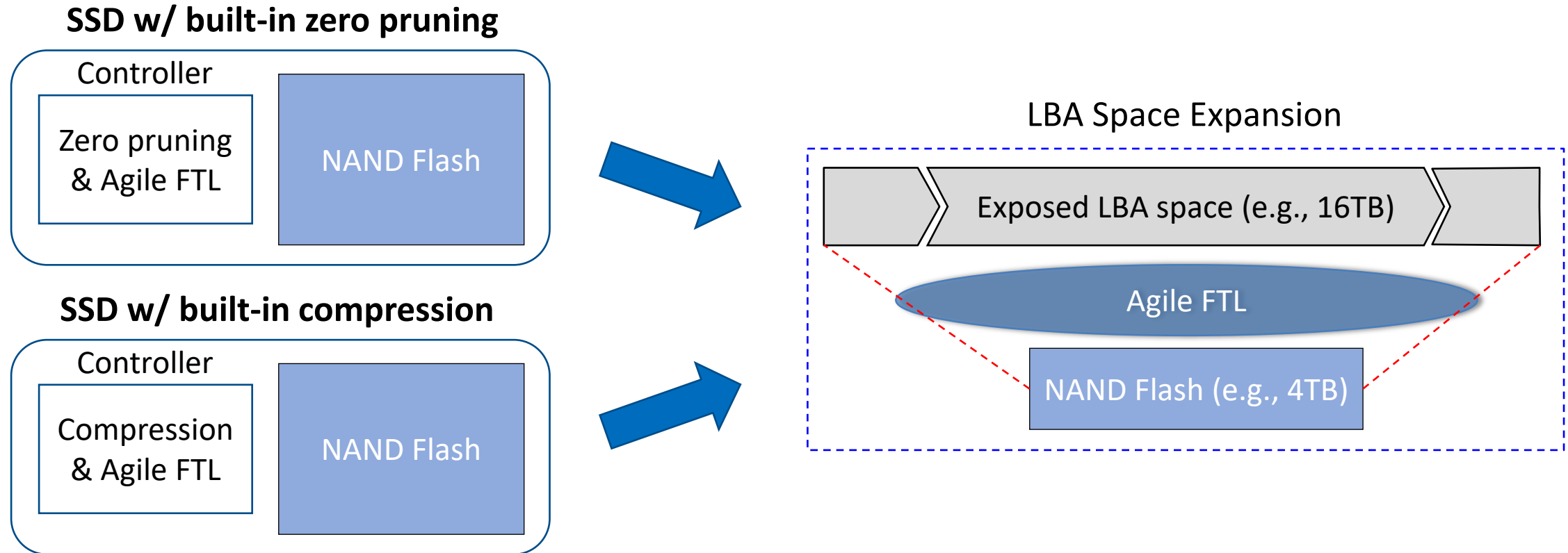


NAND Flash

Compression-assisted Elastic RAID

? Realization of RAID level conversion

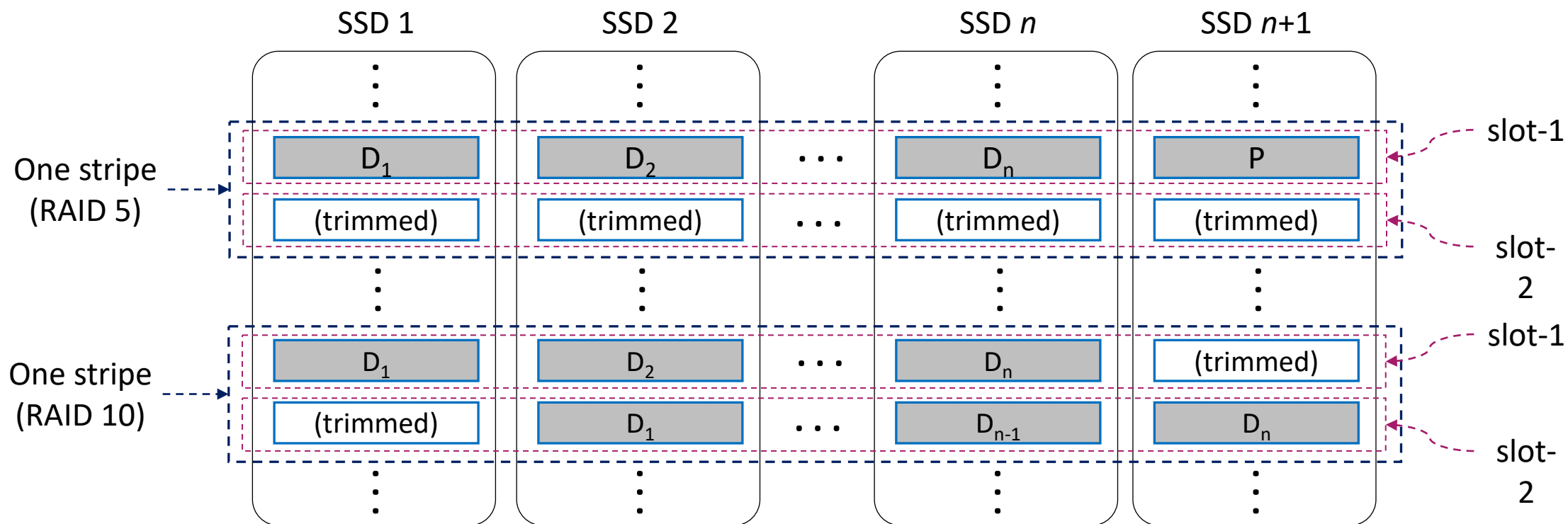
- How to seamlessly accommodate different storage space usage of different RAID levels?



Compression-assisted Elastic RAID

? Realization of RAID level conversion

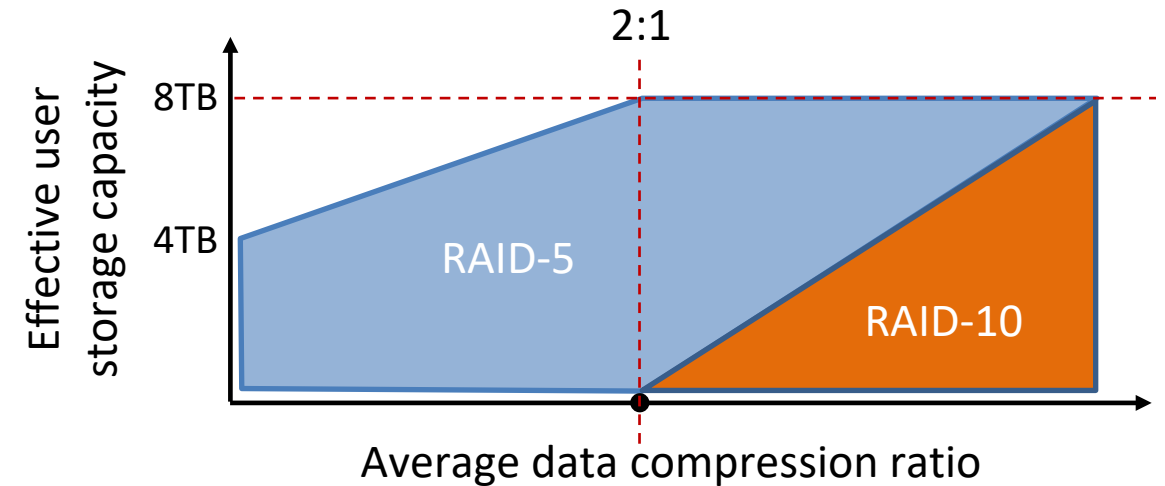
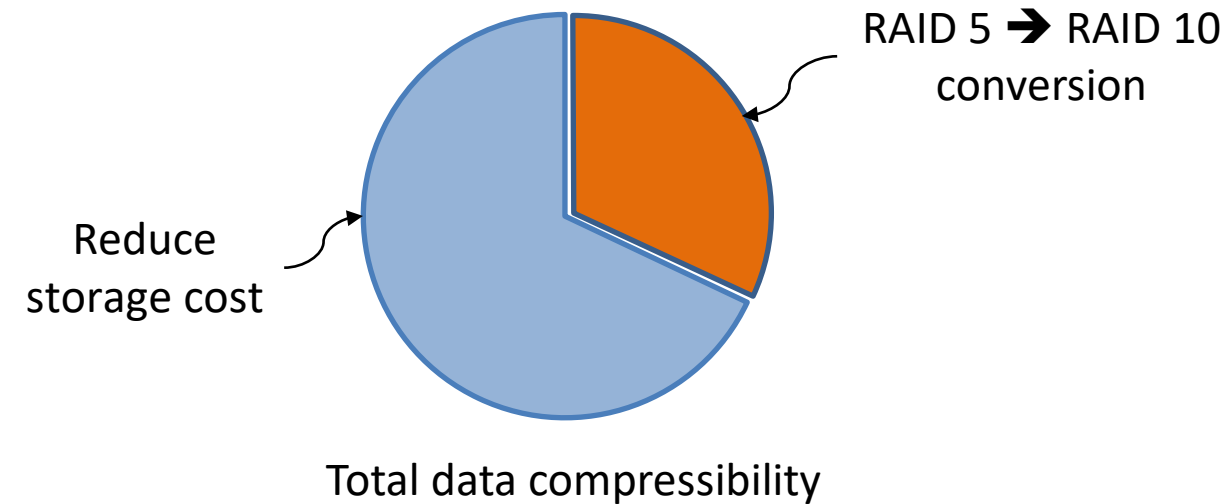
- How to seamlessly accommodate different storage space usage of different RAID levels?



Flexible per-stripe RAID level conversion

Compression-assisted Elastic RAID

- ❑ Leverage compression to **simultaneously** reduce storage cost and realize elastic RAID



Flexible user-defined exploitation of total data compressibility

Compression-assisted Elastic RAID

❑ One interesting observation



RAID 5 is **not** necessarily always more storage efficient than RAID 10

XOR tends to destroy data entropy → RAID 5 parity is (much) less compressible than user data

Experiments w/ 3+1 RAID 5

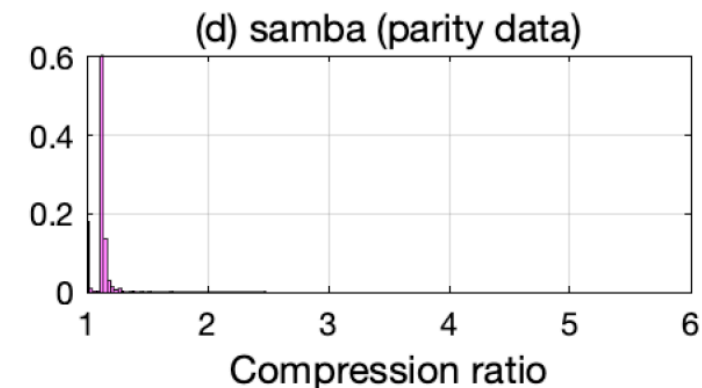
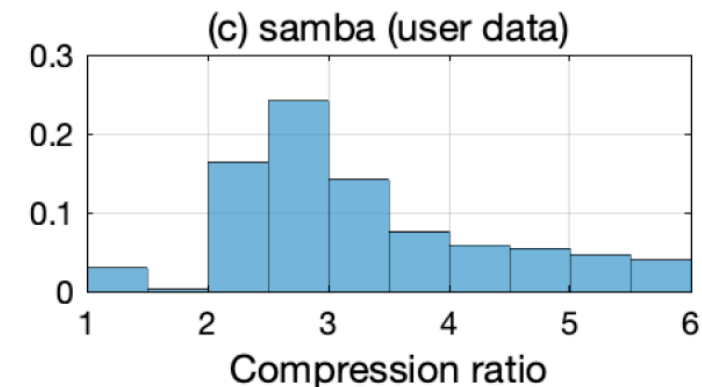
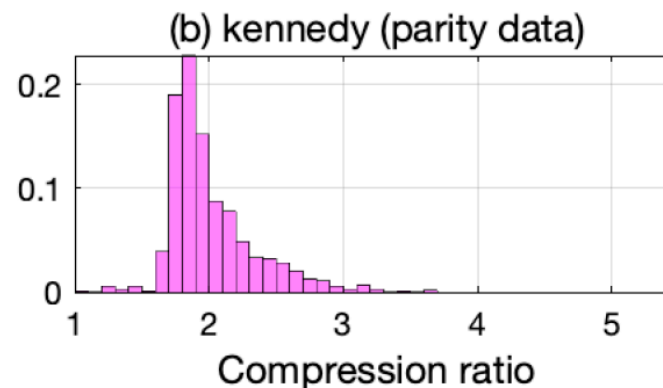
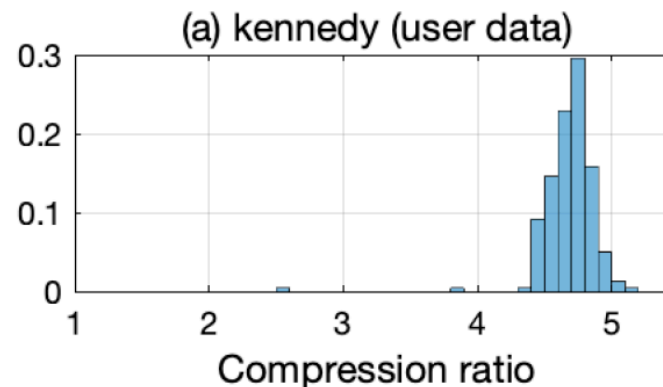
➤ Test data

- *Kennedy* from the Canterbury corpus
- *Samba* from the Silesia corpus

➤ Compression ratio is defined as

$$\frac{\text{Original data size}}{\text{Compressed data size}}$$

RAID parity is 2~4x less compressible than user data



Prototype Implementation

Linux block layer software elastic RAID

- ✓ Support RAID 10 and RAID 5, and per-stripe RAID 5 ↔ RAID 10 conversion
- ✓ RAID 5 write logging to mitigate the well-known *write hole* issue and streamline writes
- ✓ Multi-threaded background data migration to fully exploit SSD IOPS
- ✓ Compressibility-adaptive RAID level conversion
- ✓ Operate over SSDs with built-in transparent compression

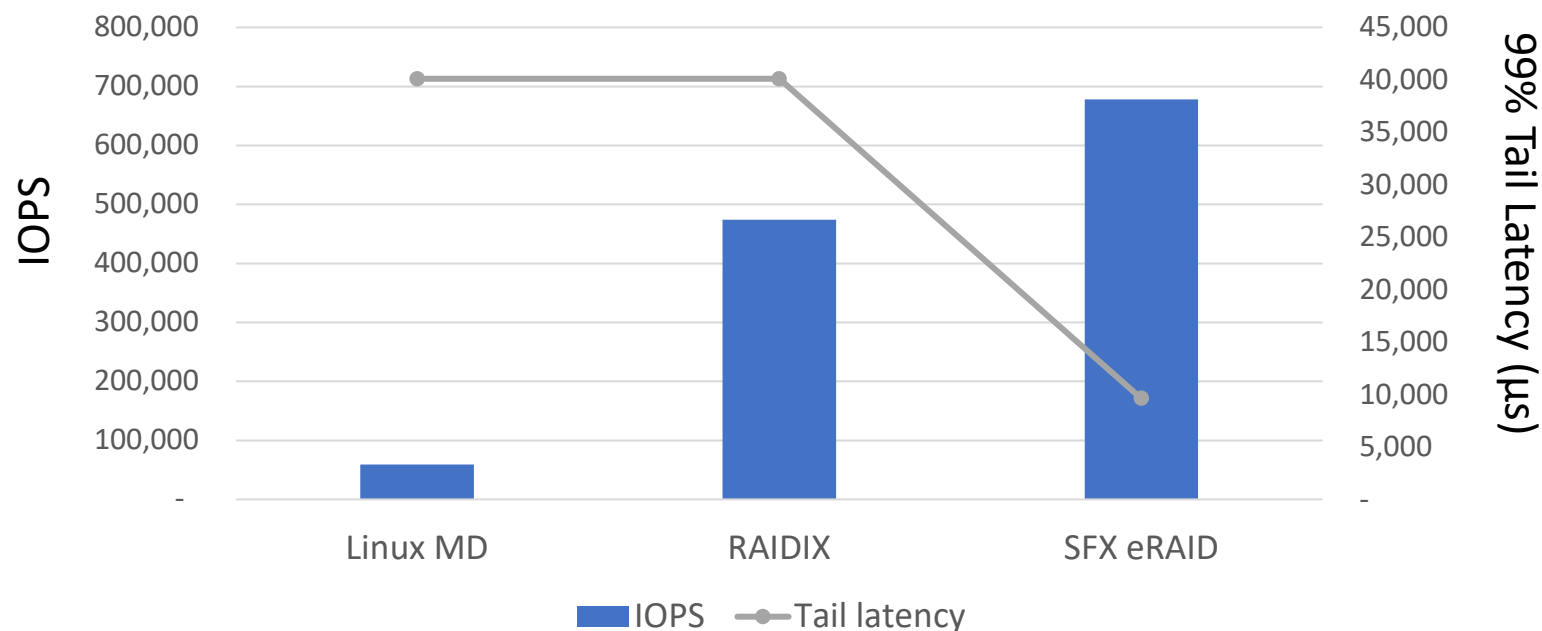
Experimental setup

- Intel(R) Xeon(R) Platinum 8269CY CPU @ 2.50GHz, 565GB DDR4 DRAM
- CentOS Linux release 7.6.1810, FIO 3.13
- ScaleFlux 4TB CSD 2000 with built-in transparent compression

Experiment Results

Baseline 3+1 RAID5 & FIO 4KB random write (16 jobs, QD=128)

❑ Under incompressible data, our elastic RAID (eRAID) operates in the RAID 5 only mode



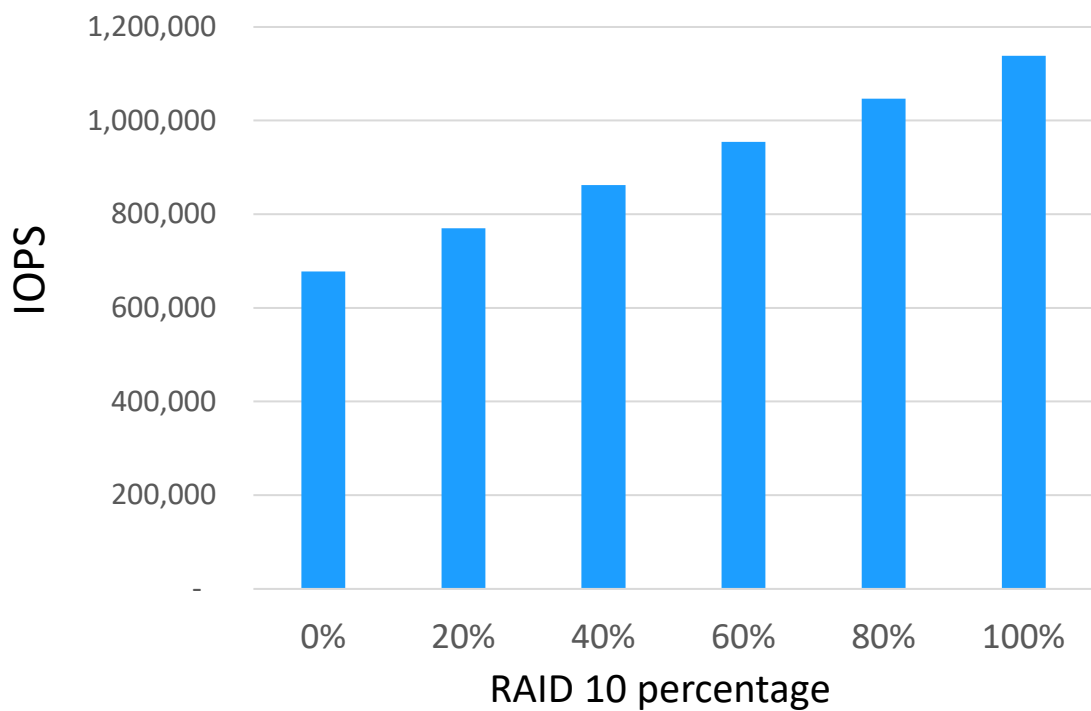
- ✓ 11x higher IOPS and 4x shorter 99% tail latency compared with Linux MD
- ✓ 1.4x higher IOPS and 4x shorter 99% tail latency compared with RAIDIX

Why? ➔ Efficient implementation of write logging and multi-threaded background data migration

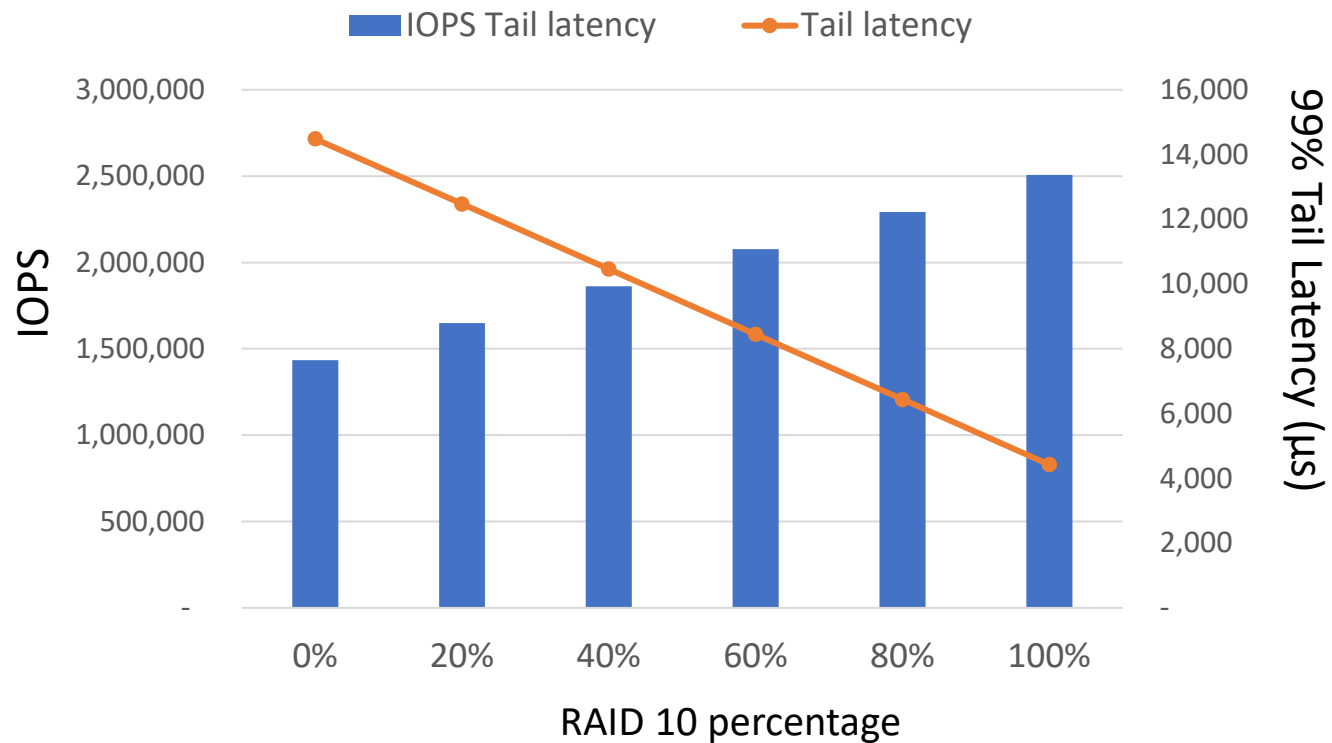
RAIDIX: A leading-edge software RAID product, <https://www.raidix.com>

Experiment Results

Elastic RAID (eRAID) IOPS under different RAID 10 percentage

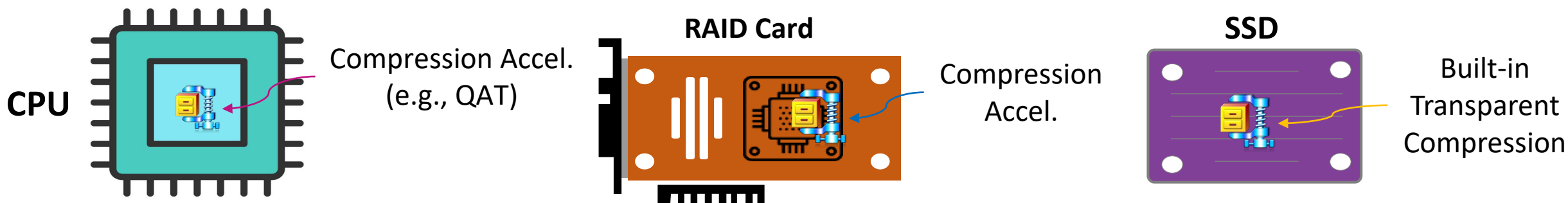


4KB random write (16 jobs, QD=128)



4KB random read under 1 SSD failure (16 jobs, QD=128)

Conclusion



- ❑ Widely available hardware compression acceleration opens a door towards **elastic RAID**
 - Exploit abundant data compressibility to approach RAID 10 perform on RAID 5 setup
- ❑ Key ideas to practically implement elastic RAID
 - Offload compressed data block management to SSD FTL
 - Per-stripe RAID level conversion by leveraging SSD LBA space expansion
- ❑ Applicable to both software RAID and hardware RAID



Questions

Please take the Session Survey Thank you!



Tuesday, August 2



Wednesday, August 3



Thursday, August 4