FlashBlox: Achieving Both Performance Isolation and Uniform Lifetime for Virtualized SSDs

Jian Huang † Anirudh Badam Laura Caulfield

Suman Nath Sudipta Sengupta Bikash Sharma Moinuddin K. Qureshi †

† Georgia Institute of Technology

Microsoft

Flash Has Changed Over the Last Decade

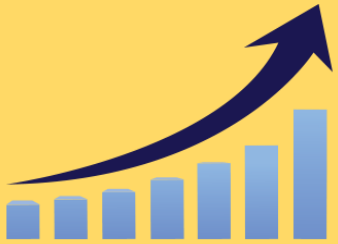
Performance Improvement

100x lower latency
5,000x higher throughput

Flash Has Changed Over the Last Decade

Performance Improvement

100x lower latency
5,000x higher throughput



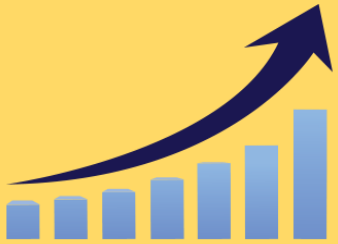
Increased Parallelism

Dozens of parallel chips

Flash Has Changed Over the Last Decade

Performance Improvement

100x lower latency
5,000x higher throughput



Increased Parallelism

Dozens of parallel chips

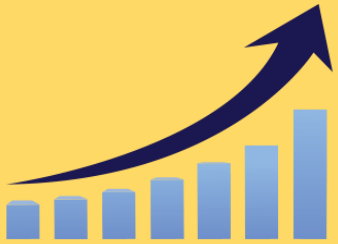
Became Commodity

Less than $0.3/GB

Flash Has Changed Over the Last Decade

Performance Improvement

100x lower latency
5,000x higher throughput



Increased Parallelism

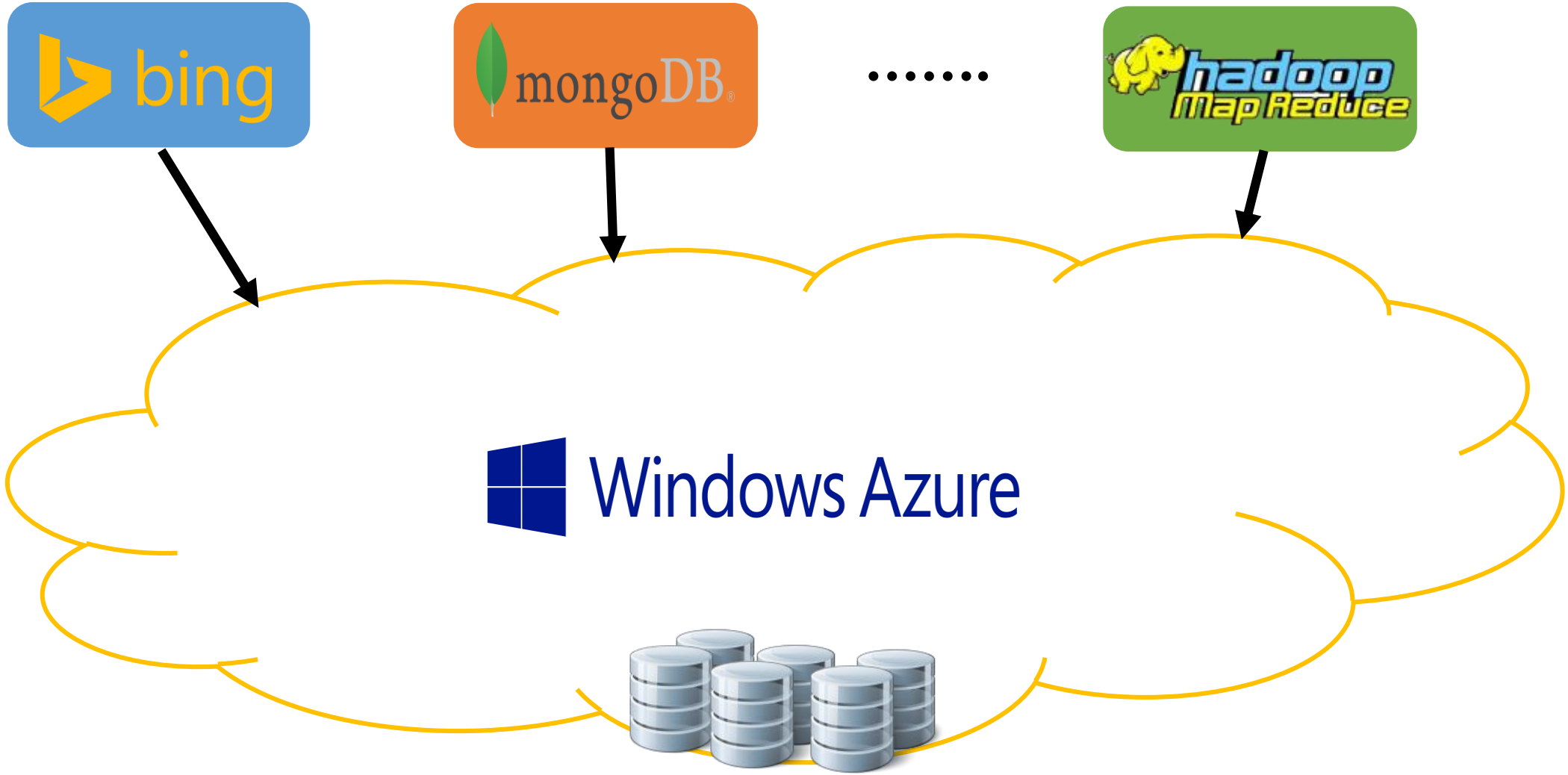
Dozens of parallel chips

Became Commodity

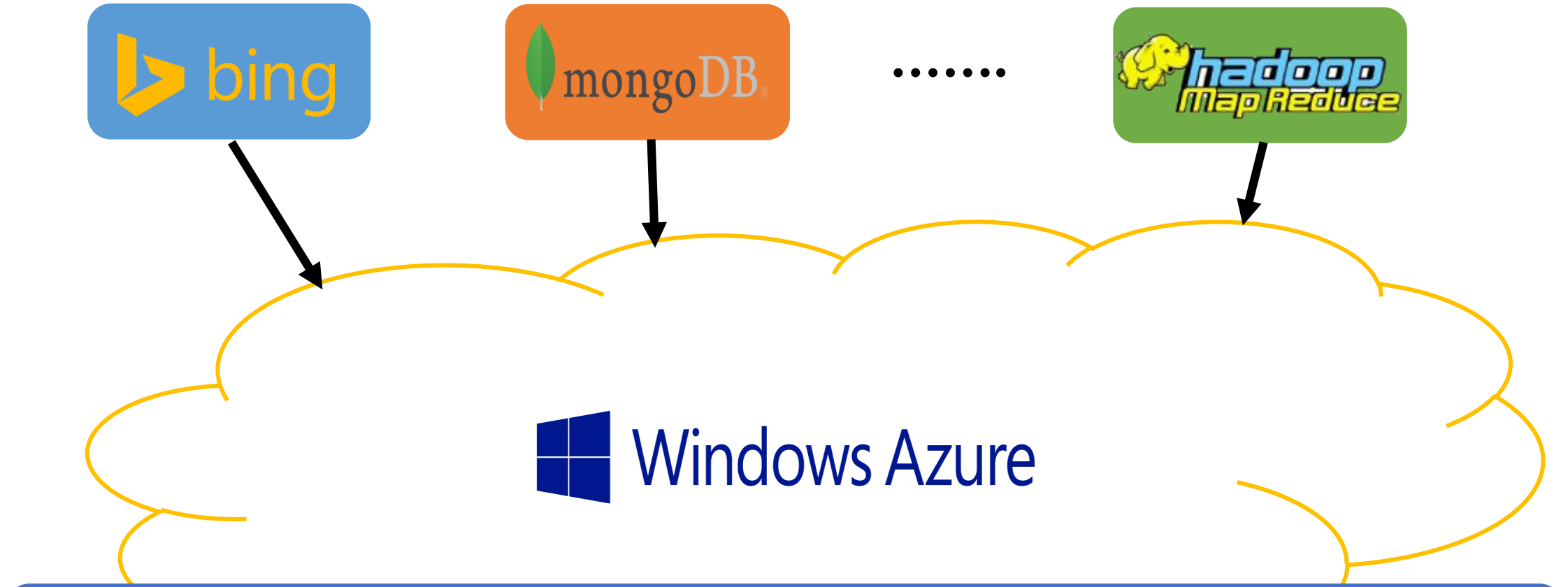
Less than $0.3/GB

Significant improvements on Flash

Shared Flash-Based Solid State Disk (SSD) in the Cloud

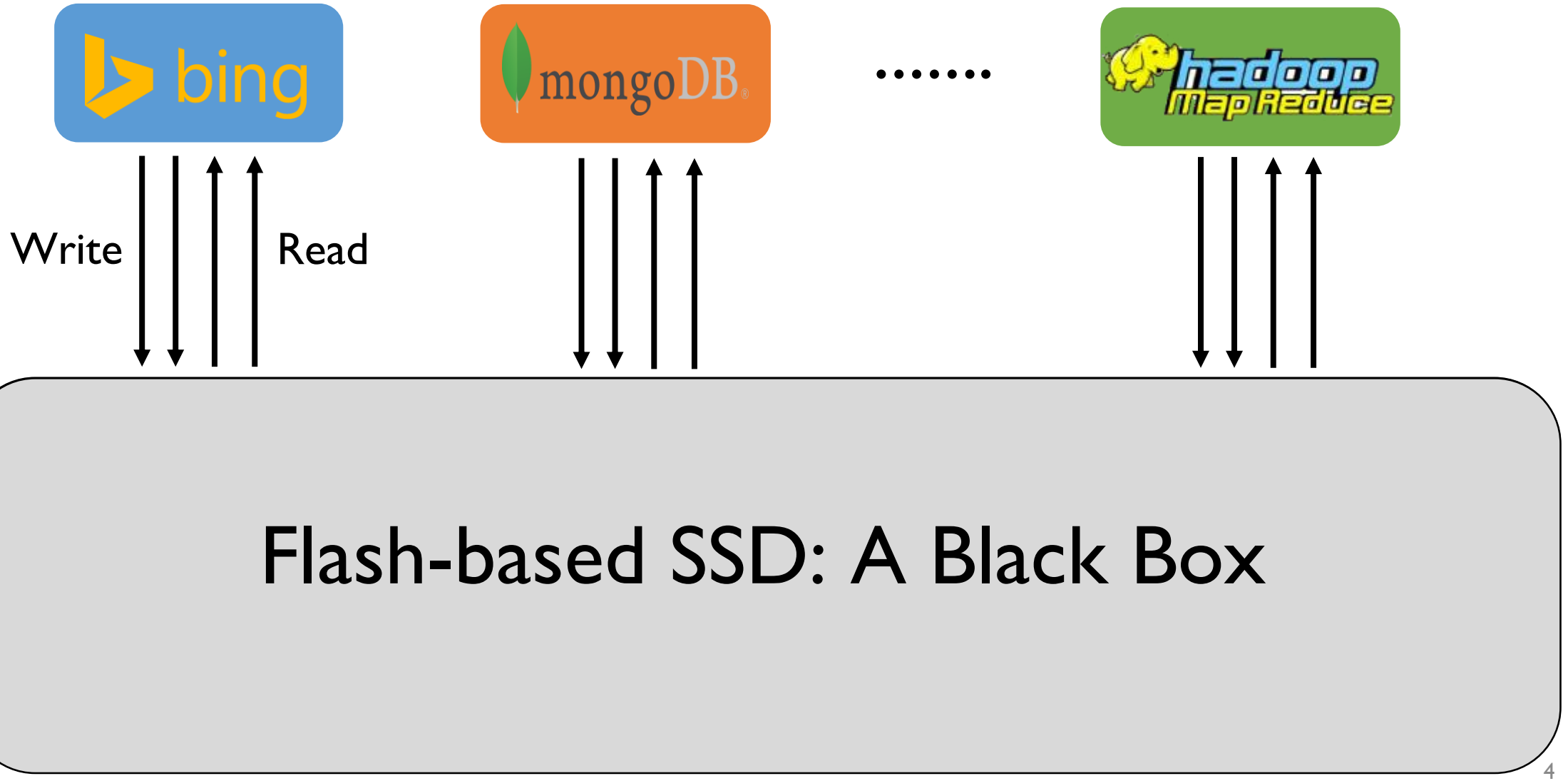


Shared Flash-Based Solid State Disk (SSD) in the Cloud

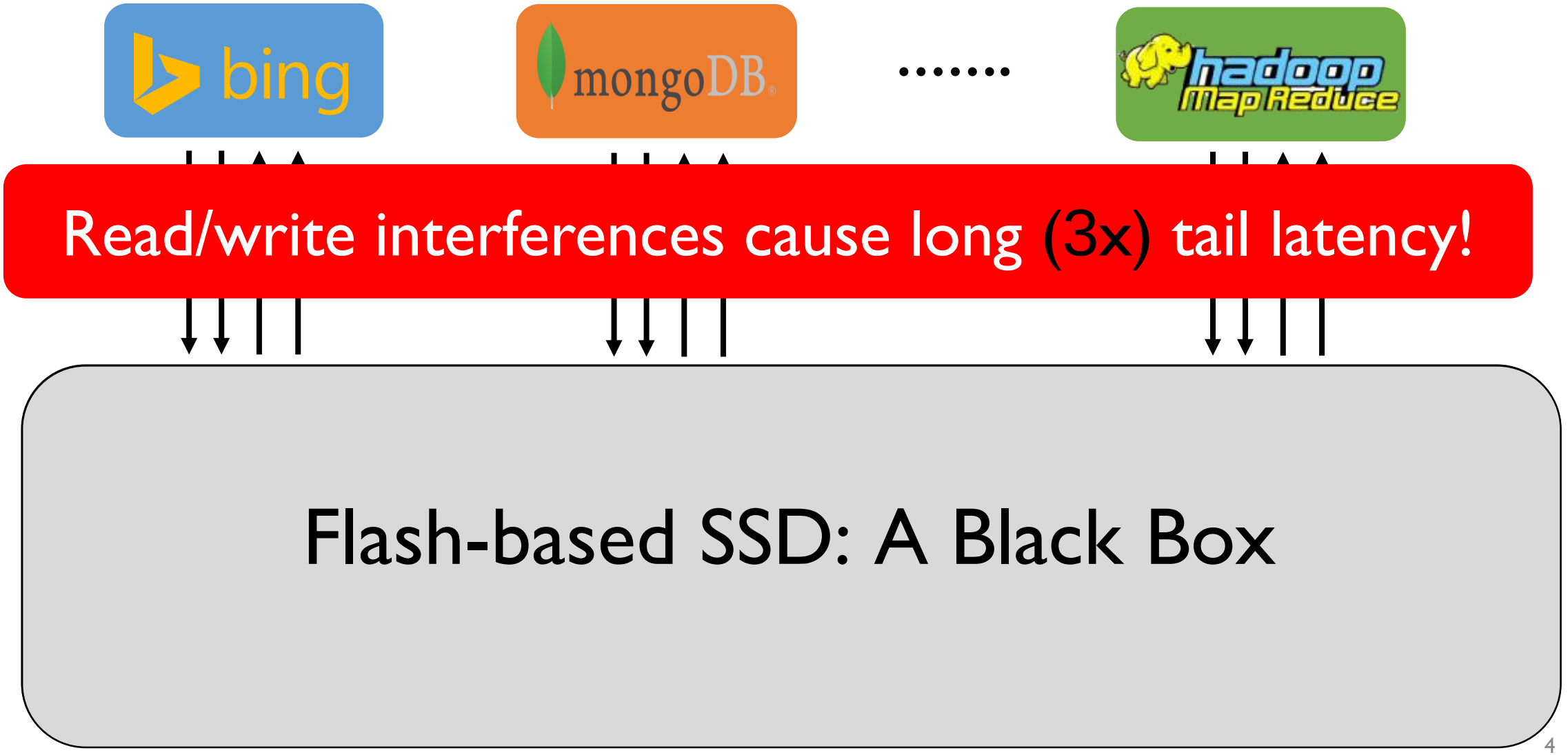


SSDs are virtualized and shared in data centers

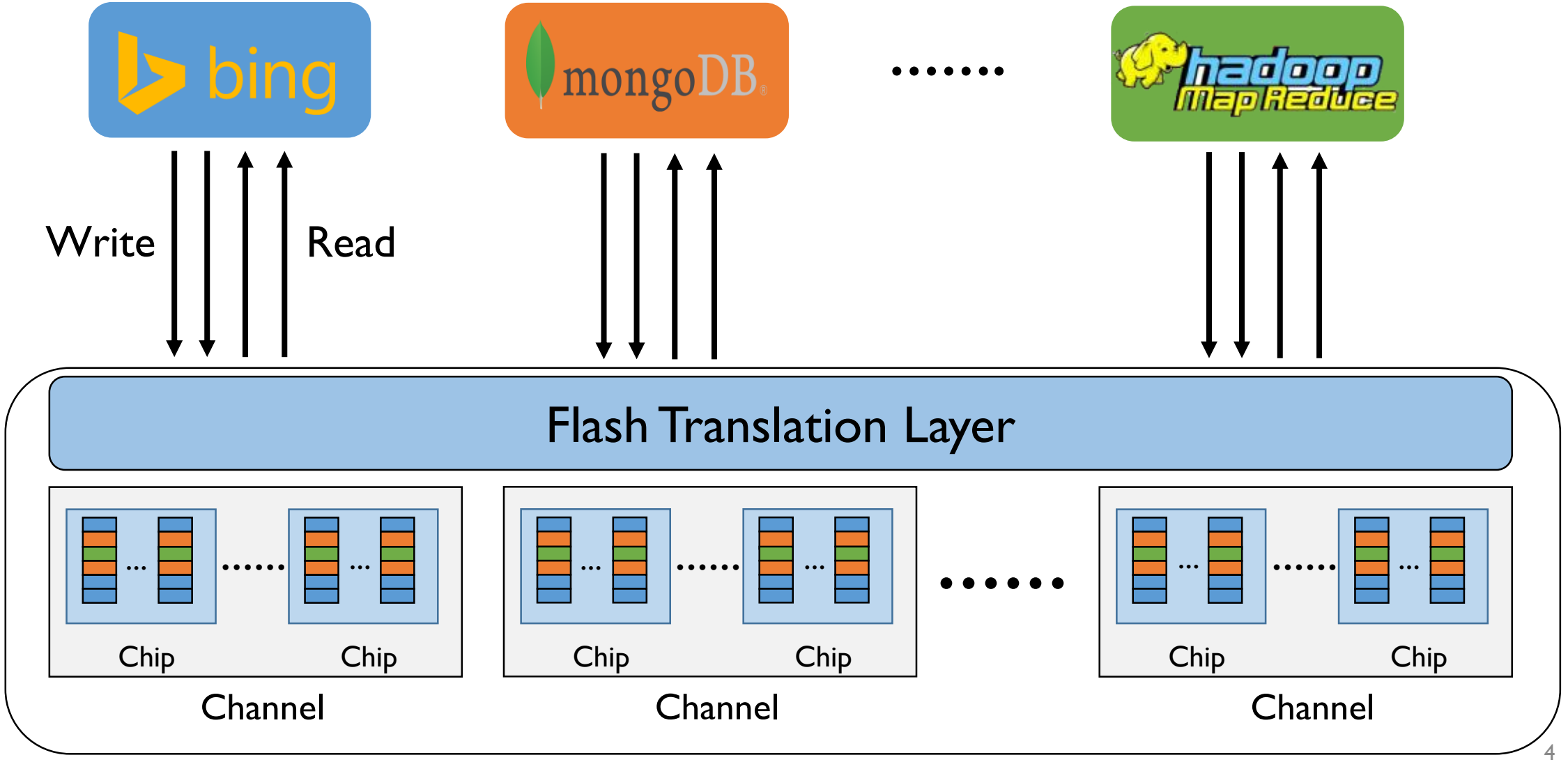
Performance Interference in Shared SSD



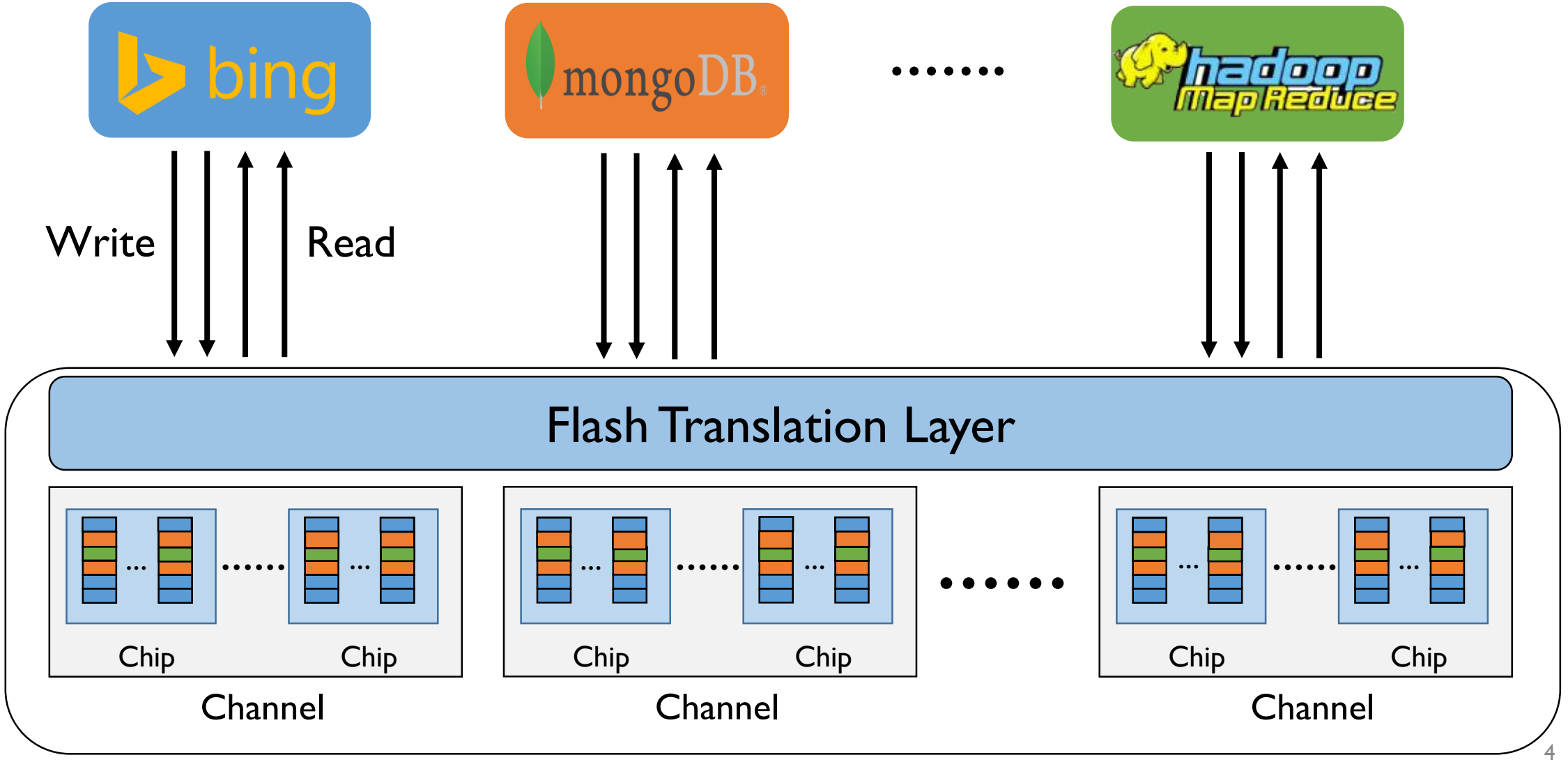
Performance Interference in Shared SSD



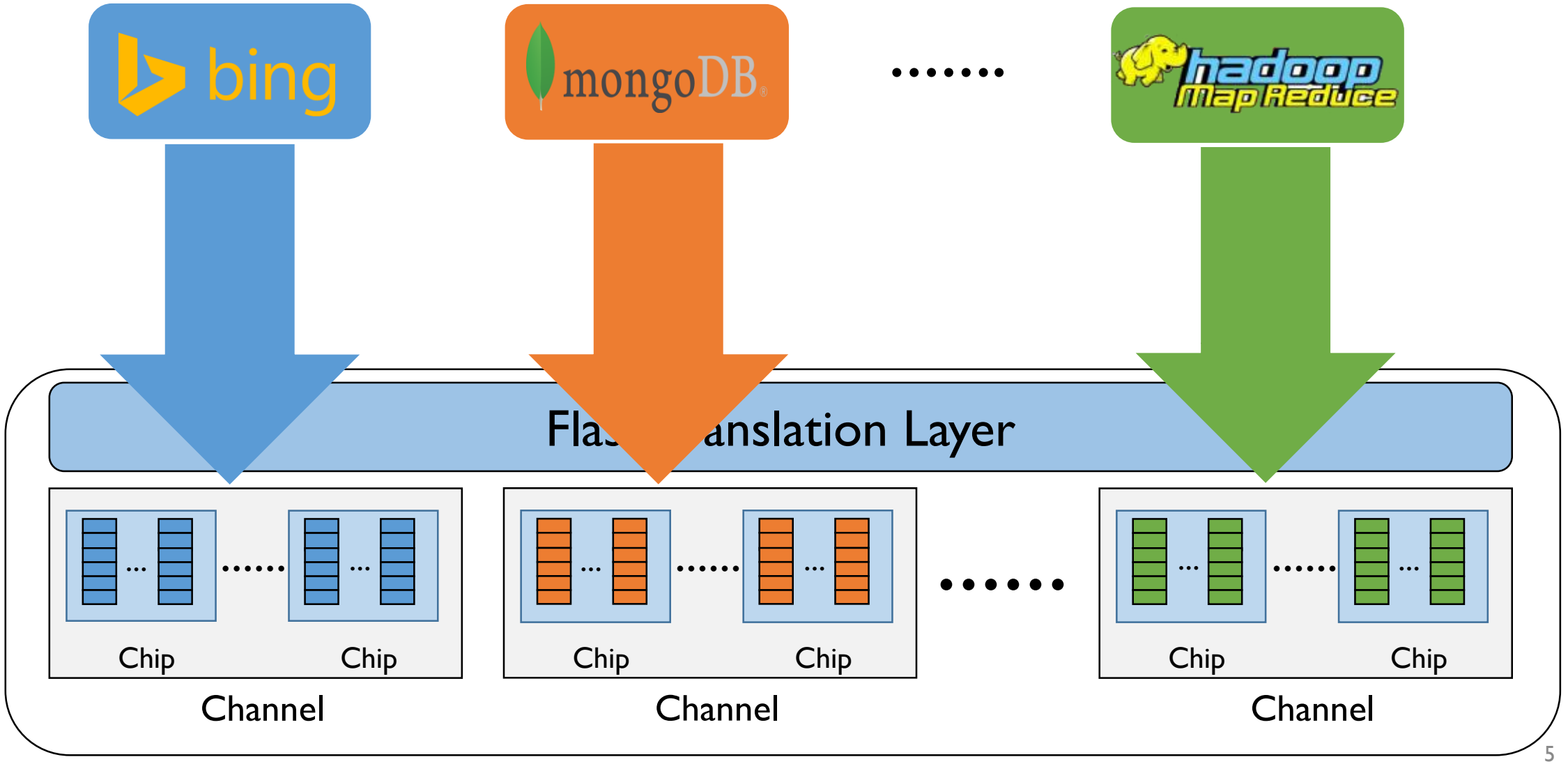
Performance Interference in Shared SSD



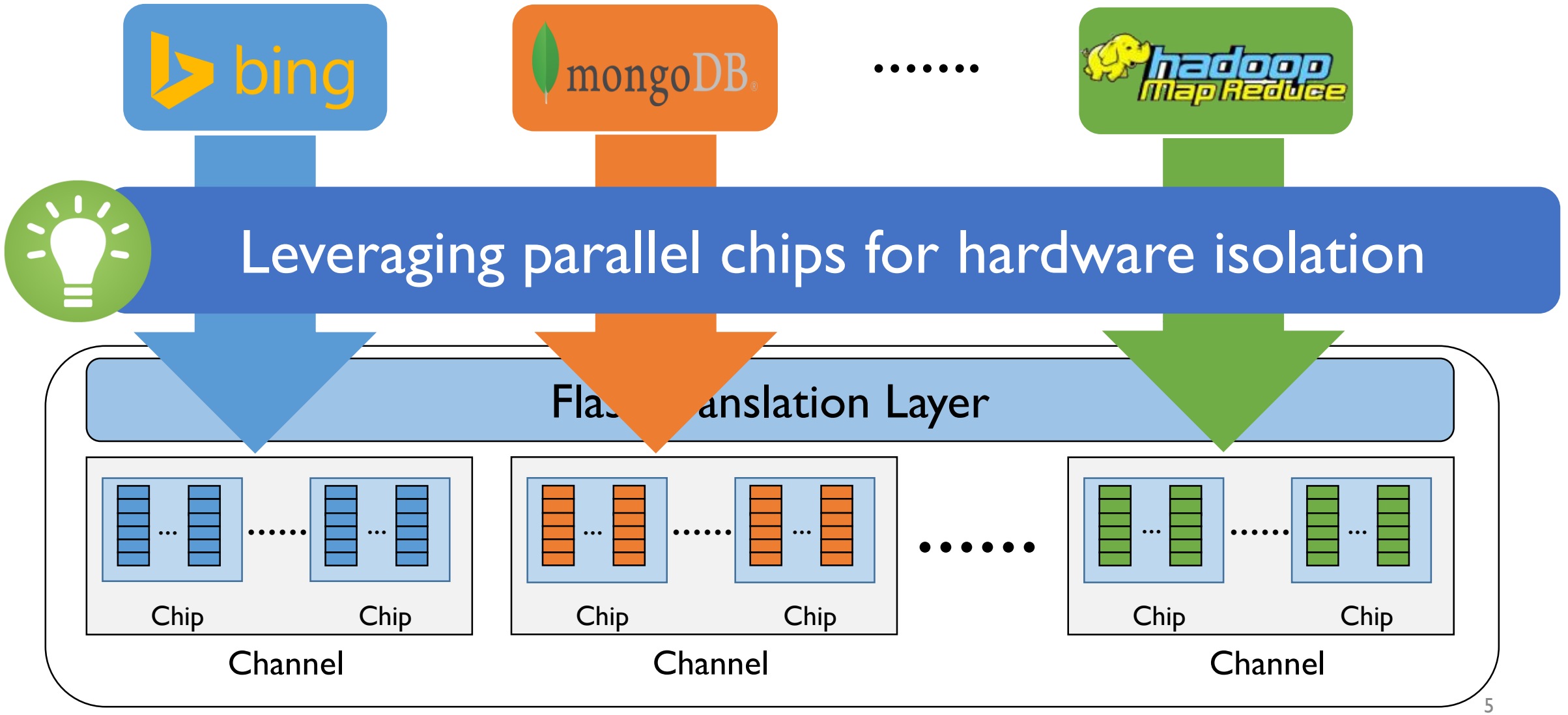
Performance Interference in Shared SSD



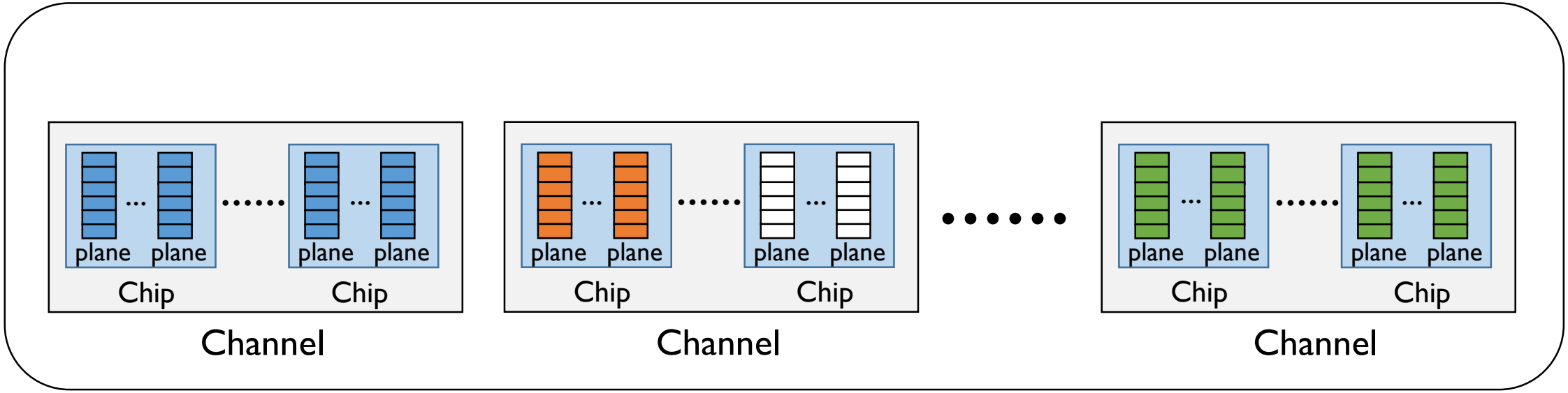
FlashBlox: Hardware Isolation in Cloud Storage



FlashBlox: Hardware Isolation in Cloud Storage

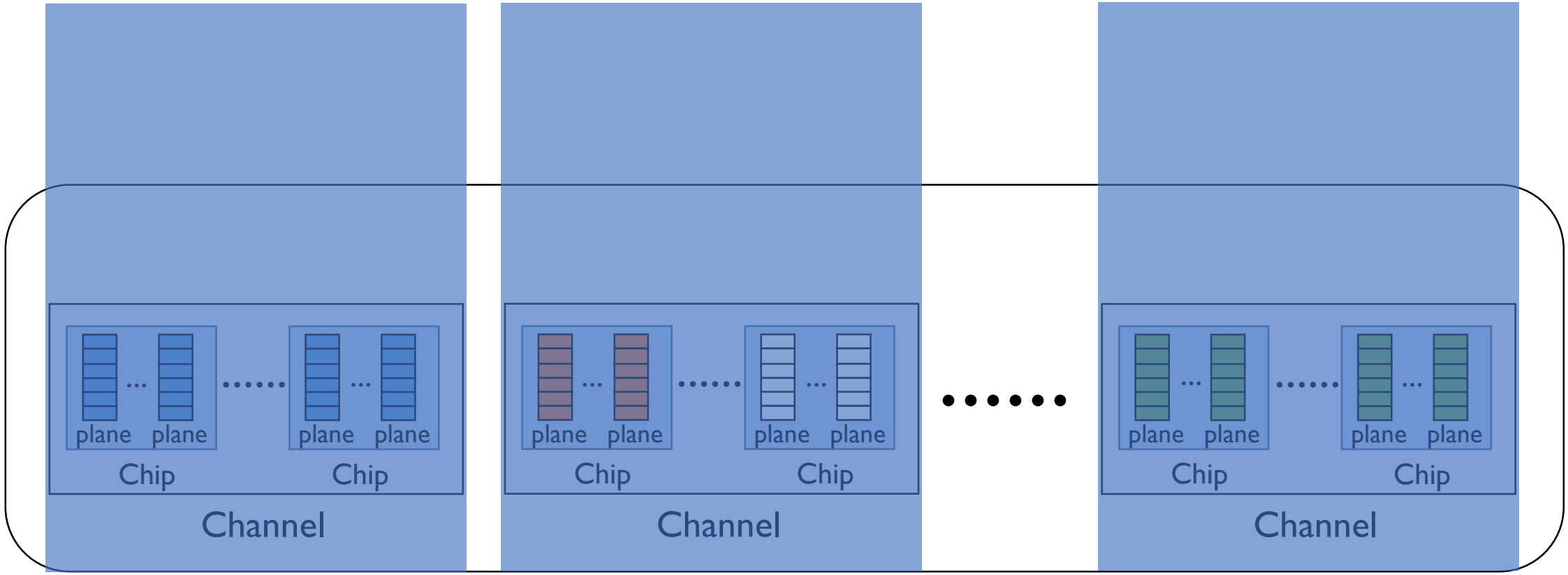


Internal Parallelism Enables Hardware Isolation

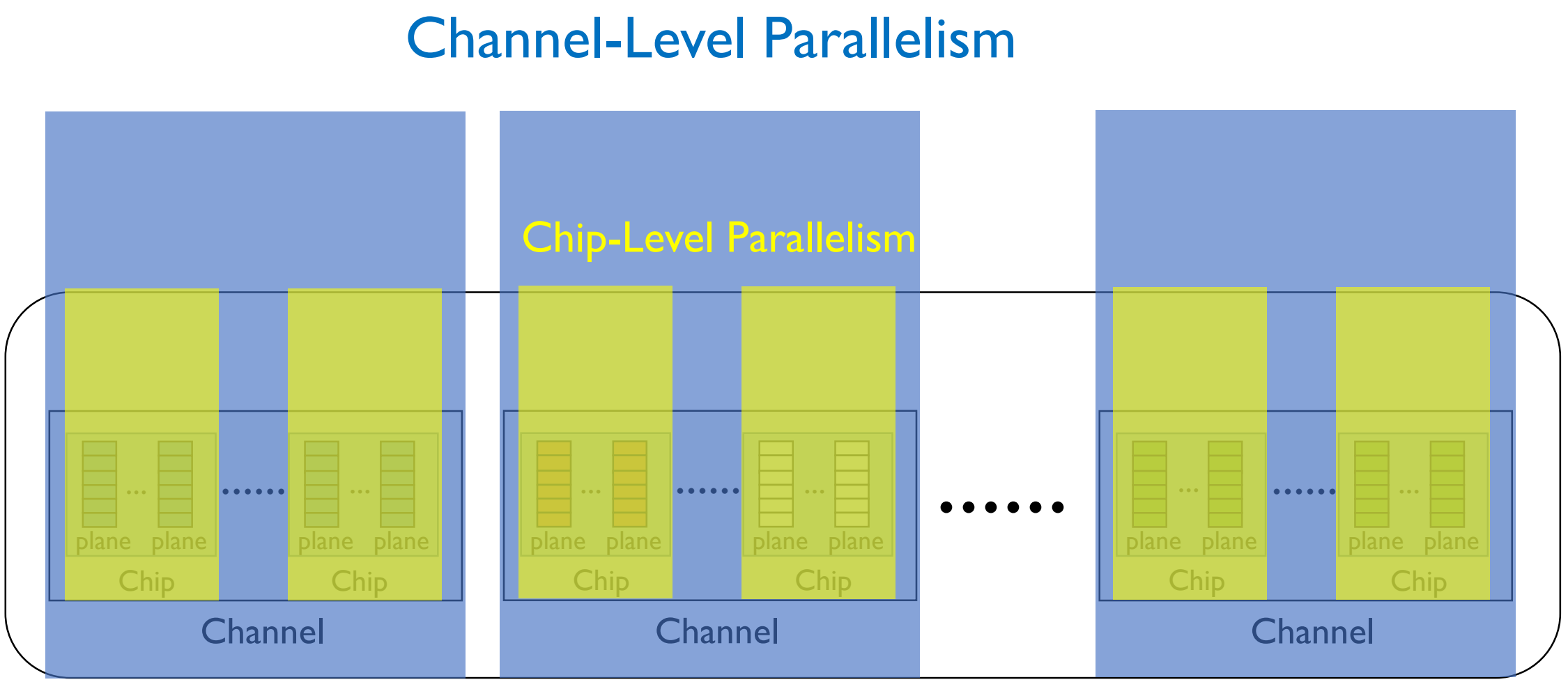


Internal Parallelism Enables Hardware Isolation

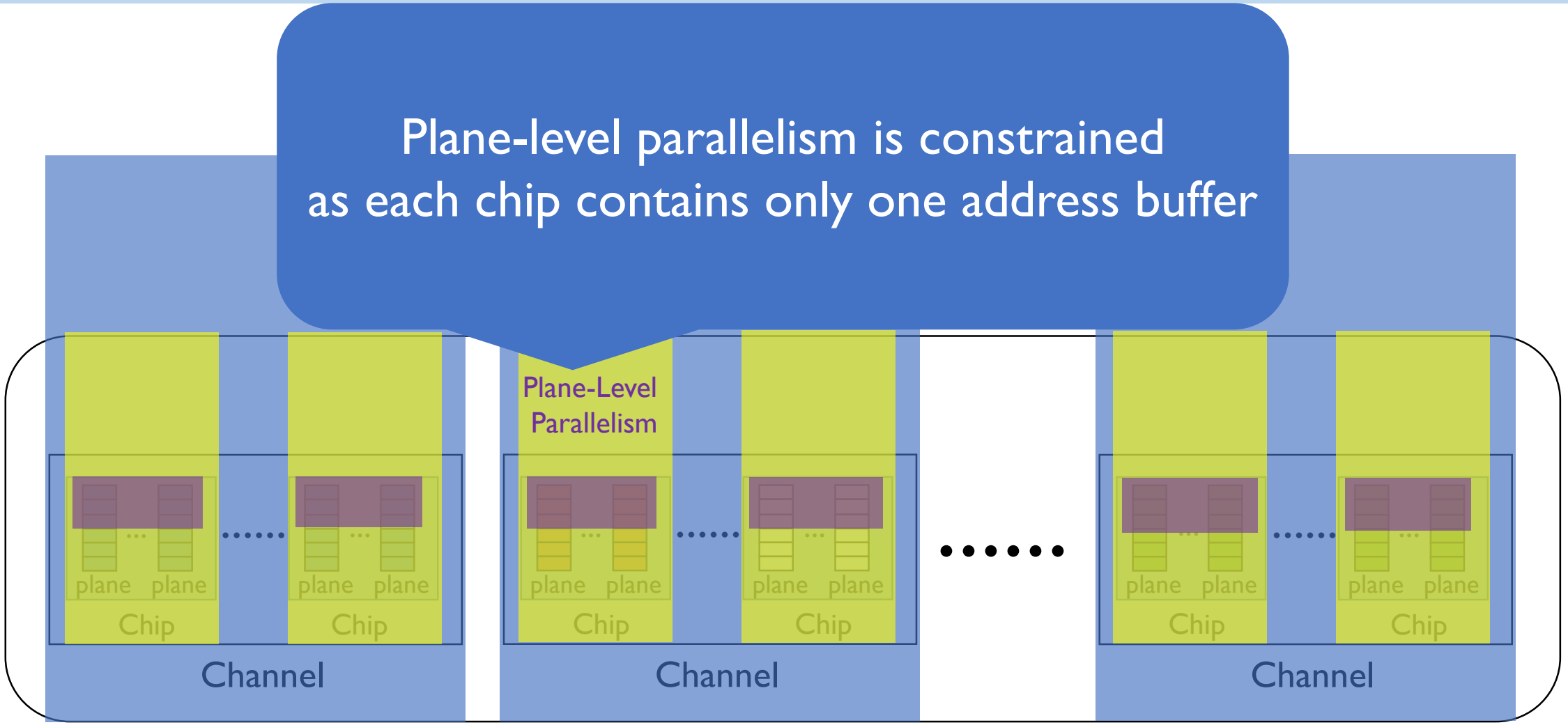
Channel-Level Parallelism



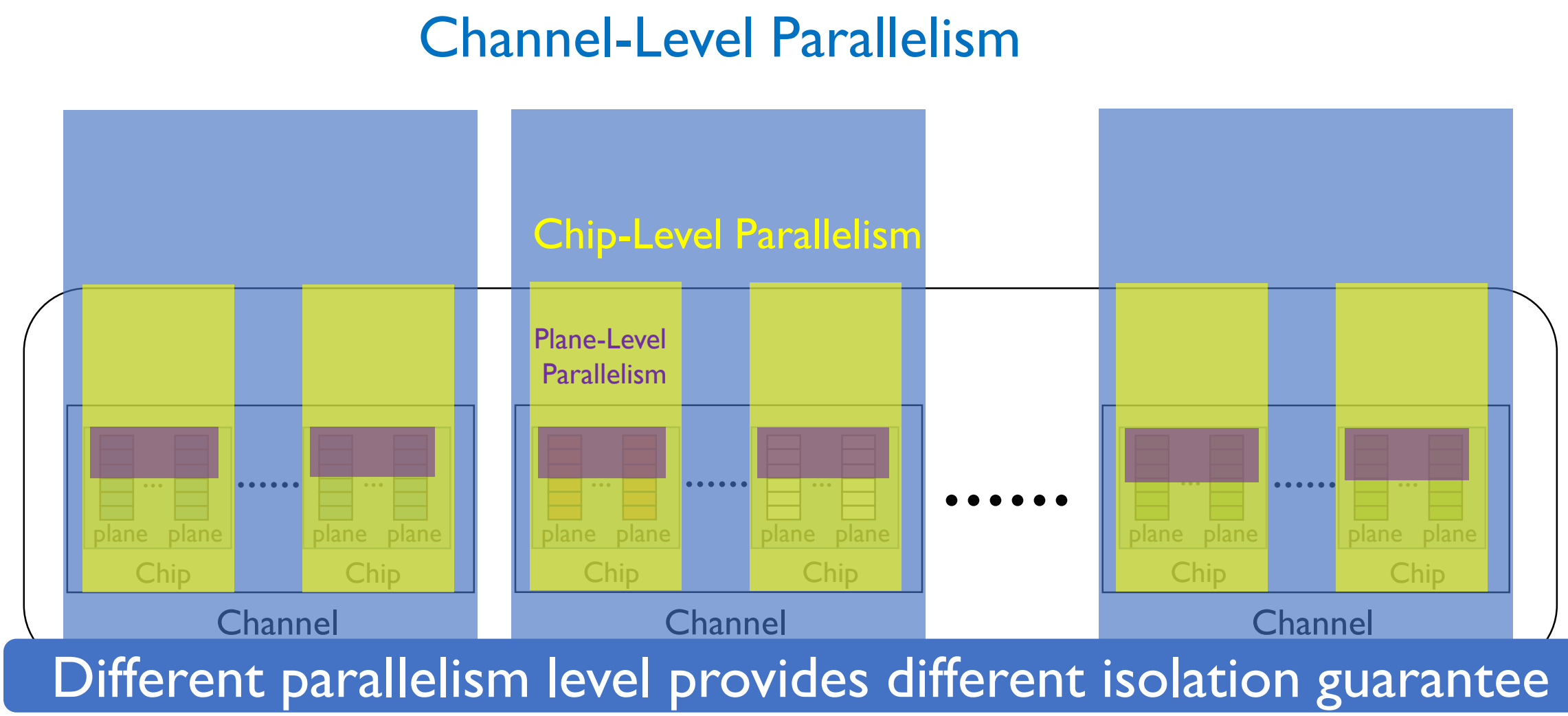
Internal Parallelism Enables Hardware Isolation



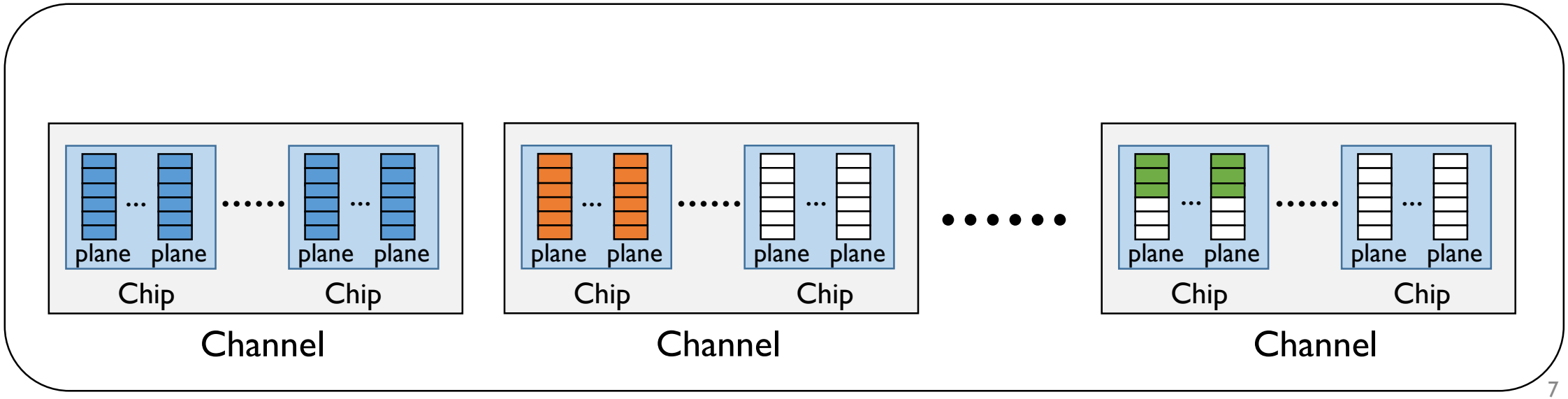
Internal Parallelism Enables Hardware Isolation



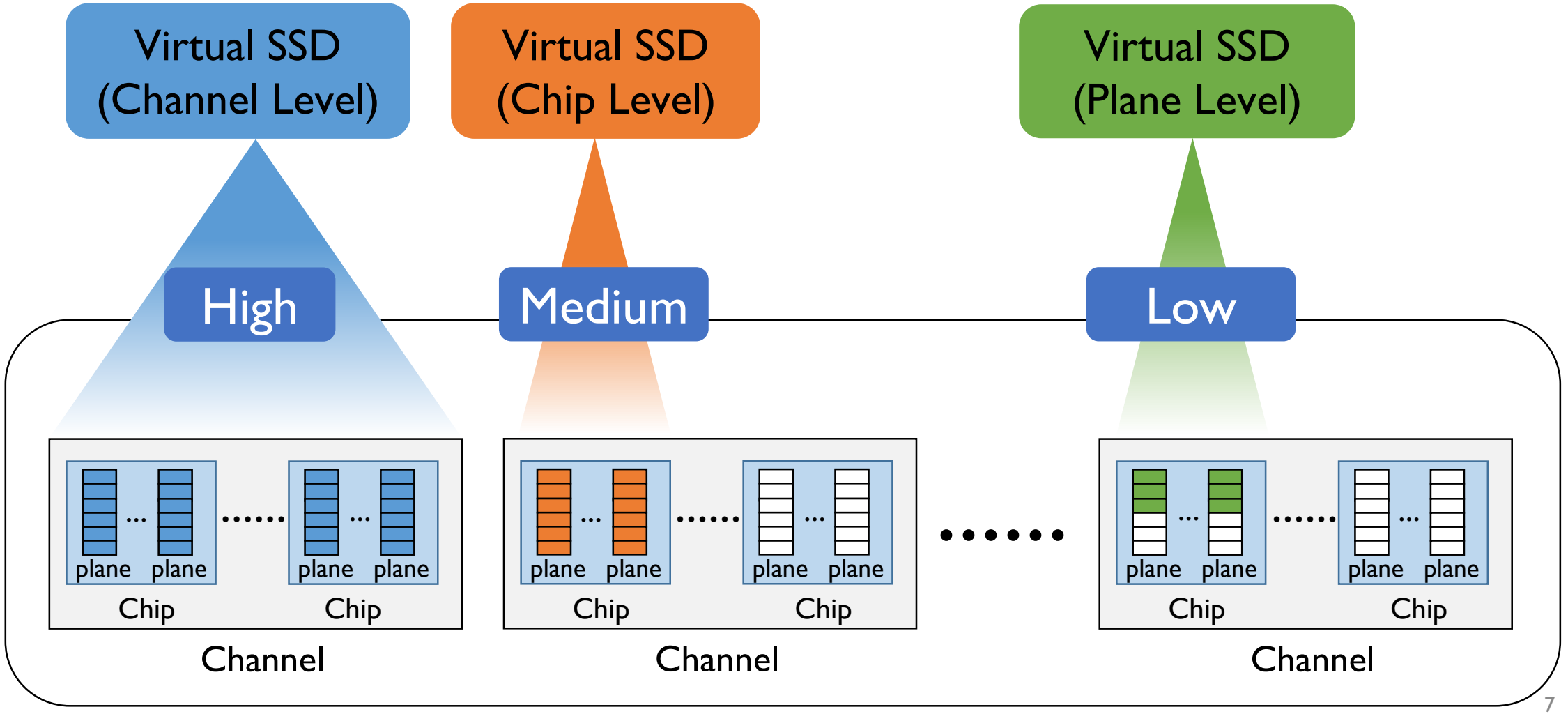
Internal Parallelism Enables Hardware Isolation



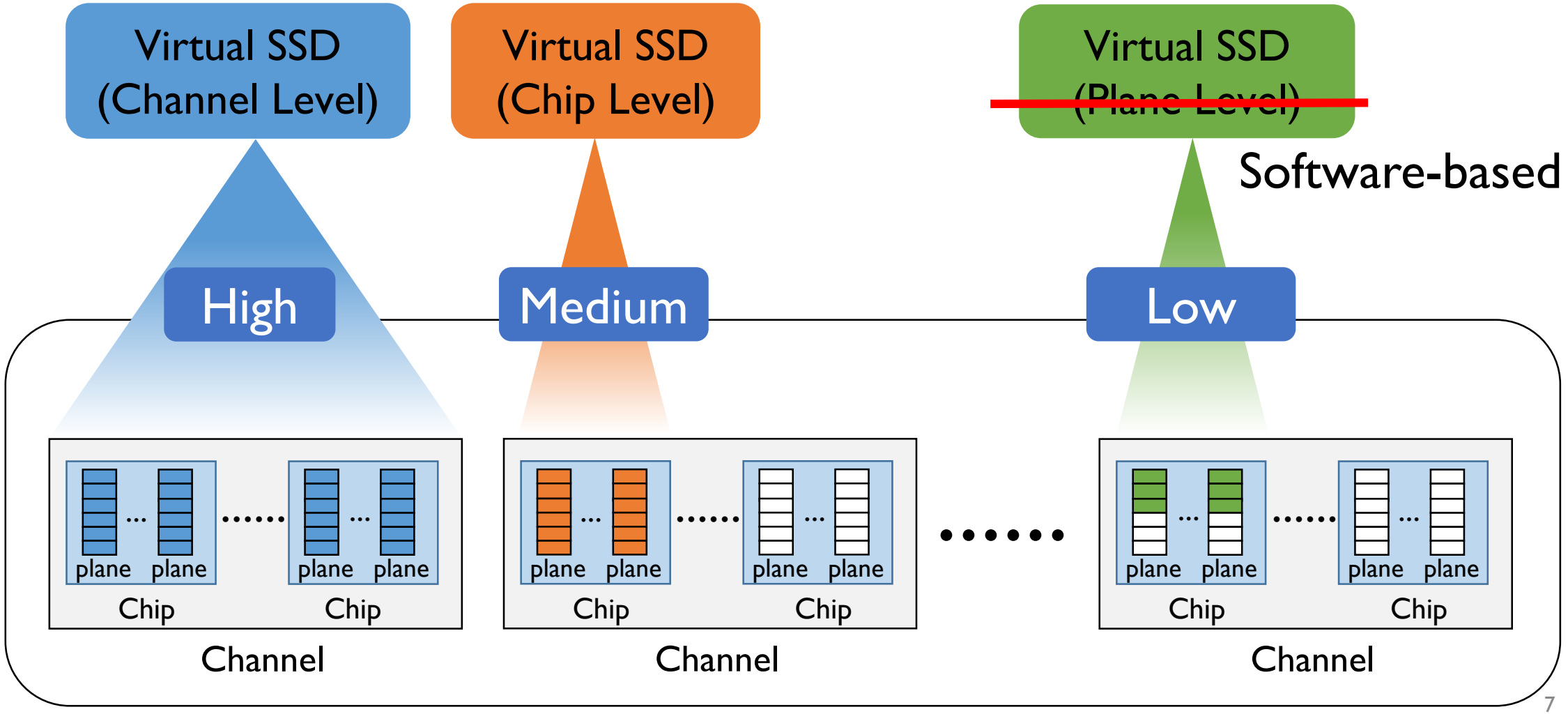
New Abstractions for Hardware Isolation



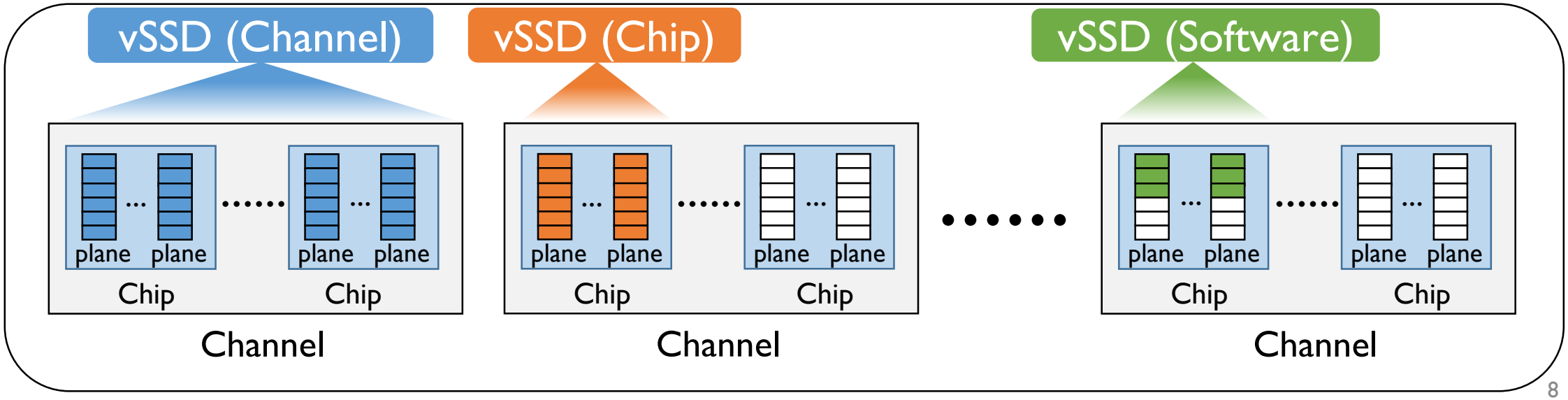
New Abstractions for Hardware Isolation



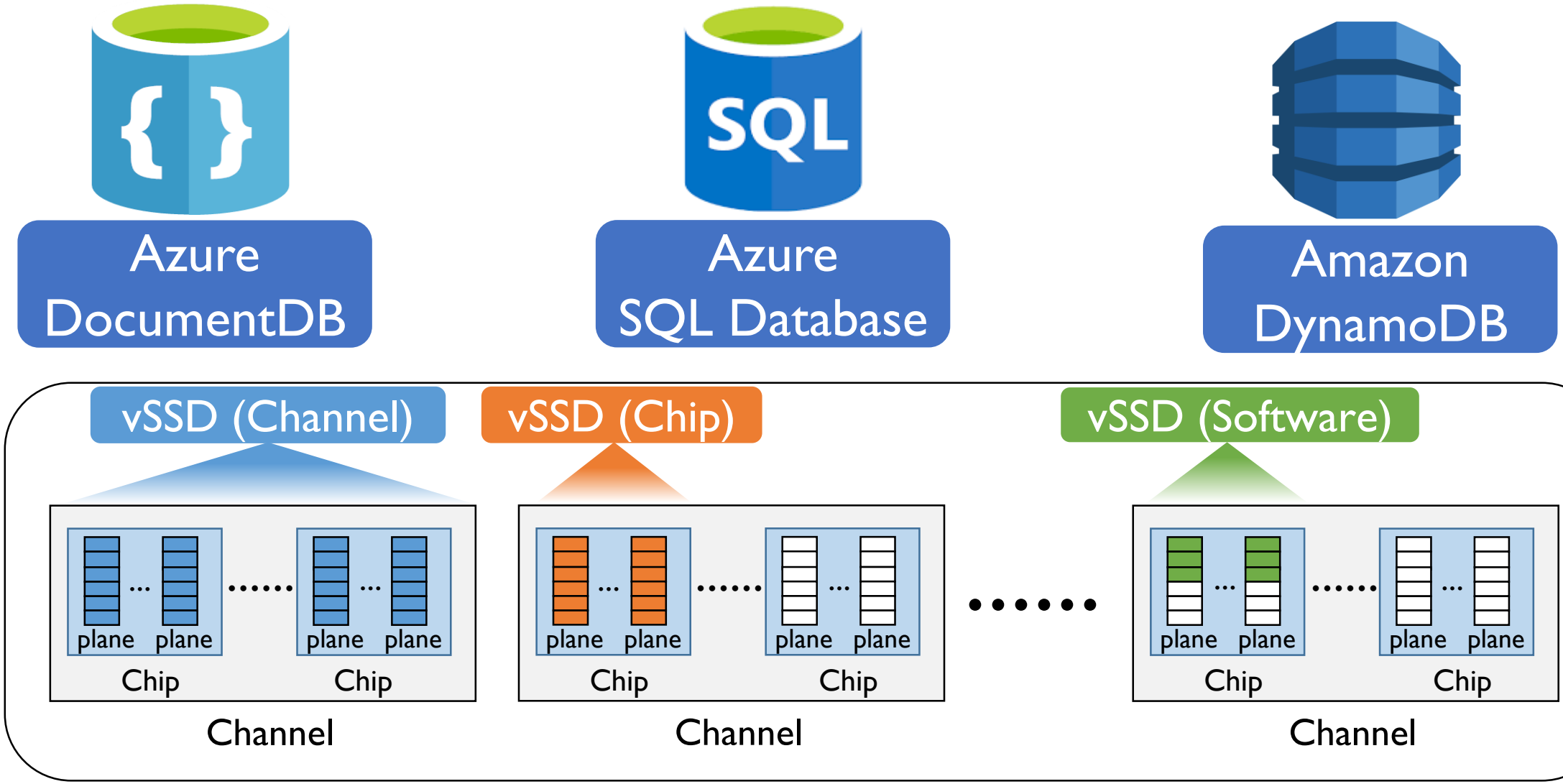
New Abstractions for Hardware Isolation



Hardware Isolation Meets the Pay-As-You-Go Model in Cloud

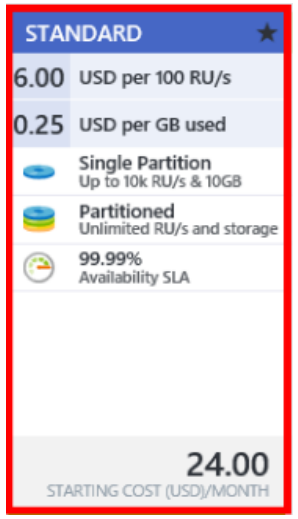


Hardware Isolation Meets the Pay-As-You-Go Model in Cloud

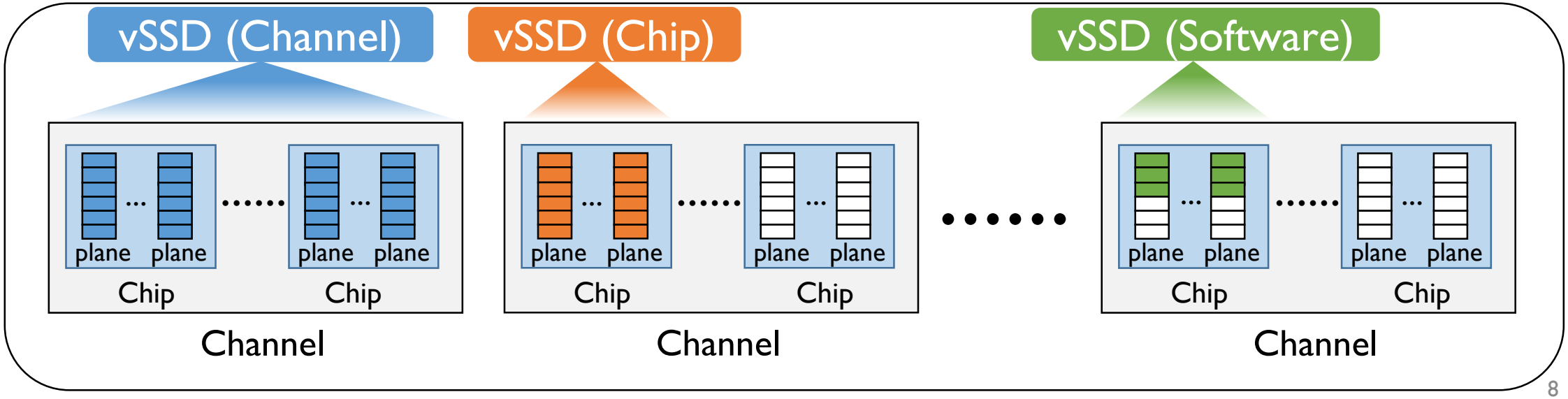


Hardware Isolation Meets the Pay-As-You-Go Model in Cloud

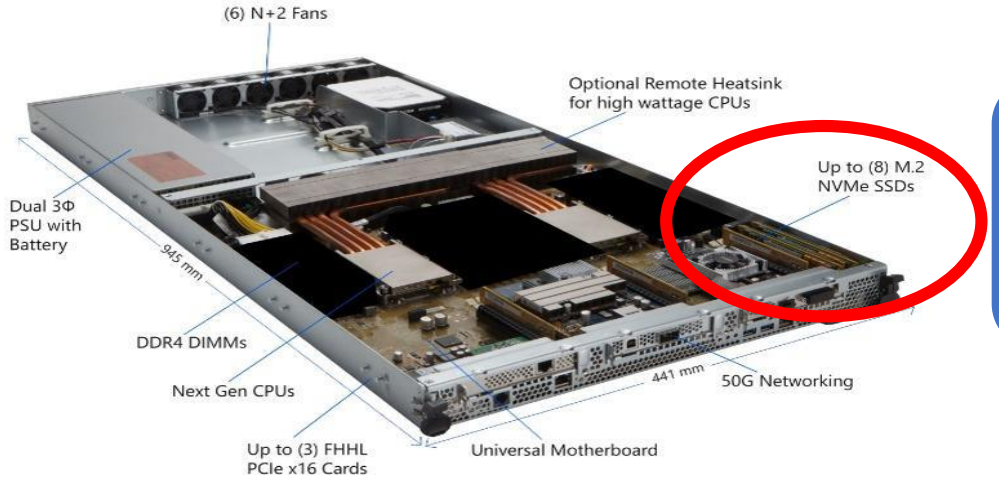
Azure DocumentDB



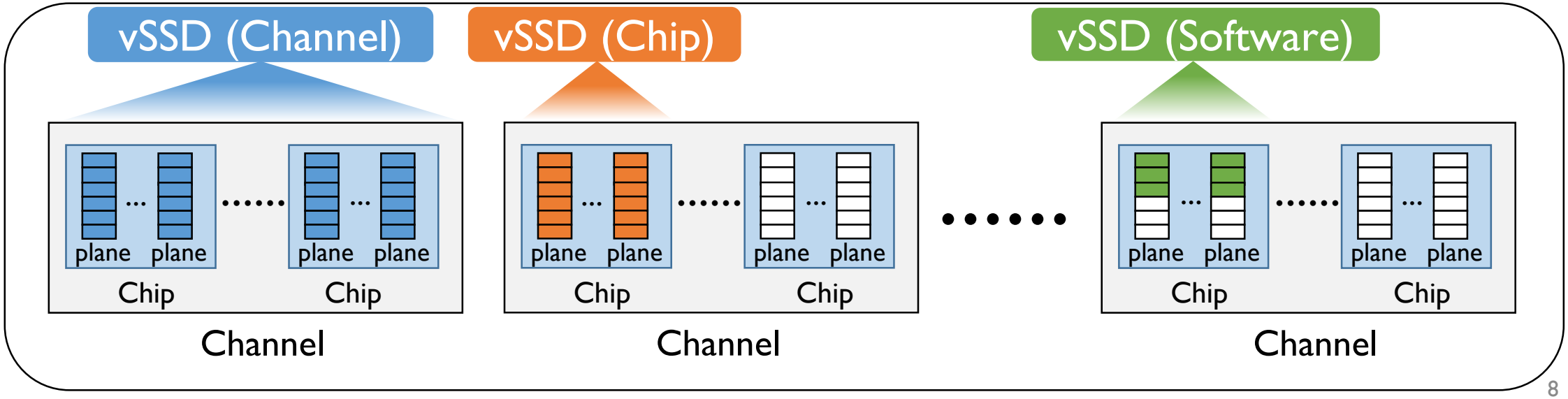
S1	S2	S3	
250 RU/s	1 K RU/s	2.5 K RU/s	Throughput
250 RU/s	1 K RU/s	2.5 K RU/s	
10 GB	10 GB	10 GB	Single Partition Size
$25 USD	$50 USD	$100 USD	Price



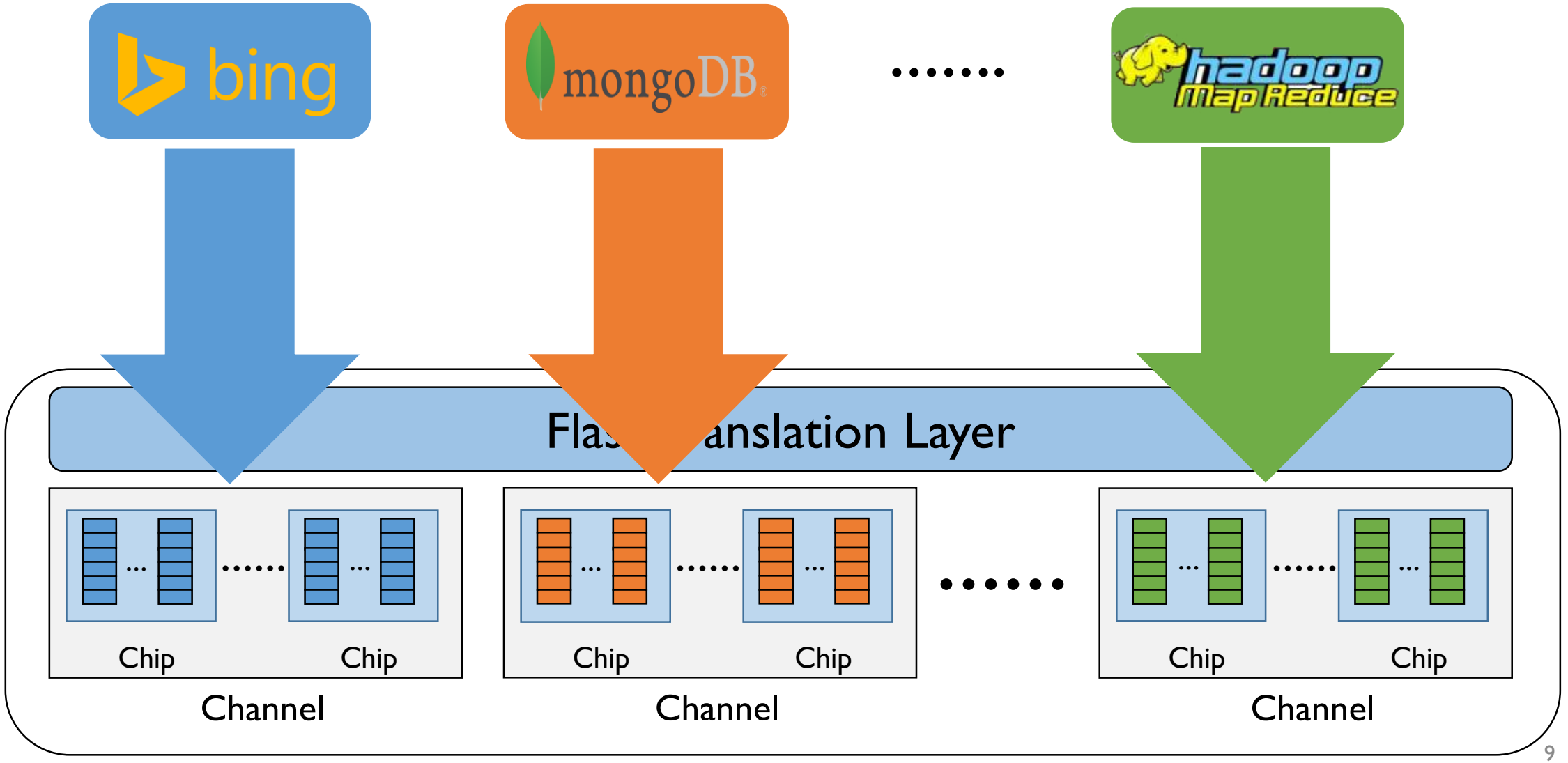
Hardware Isolation Meets the Pay-As-You-Go Model in Cloud



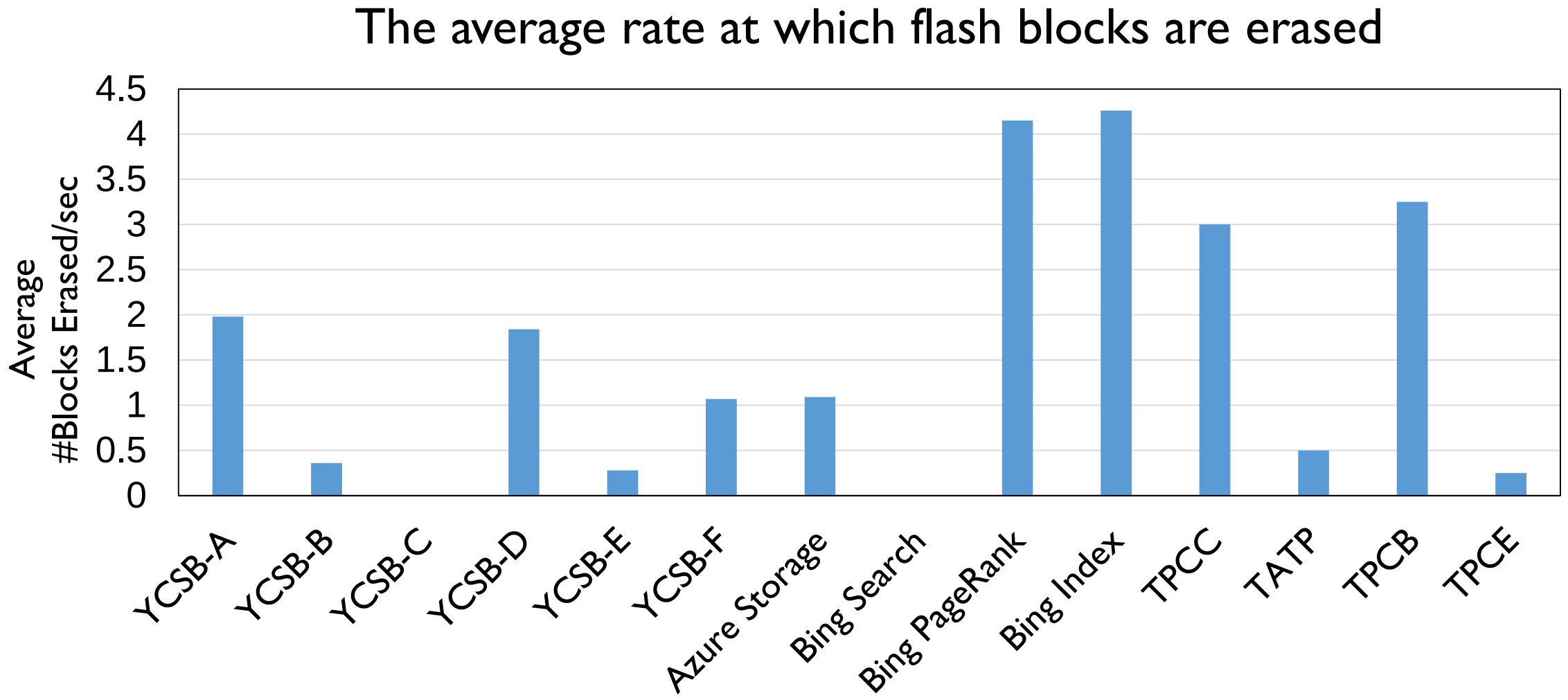
Hundreds of vSSDs can be supported in a single server



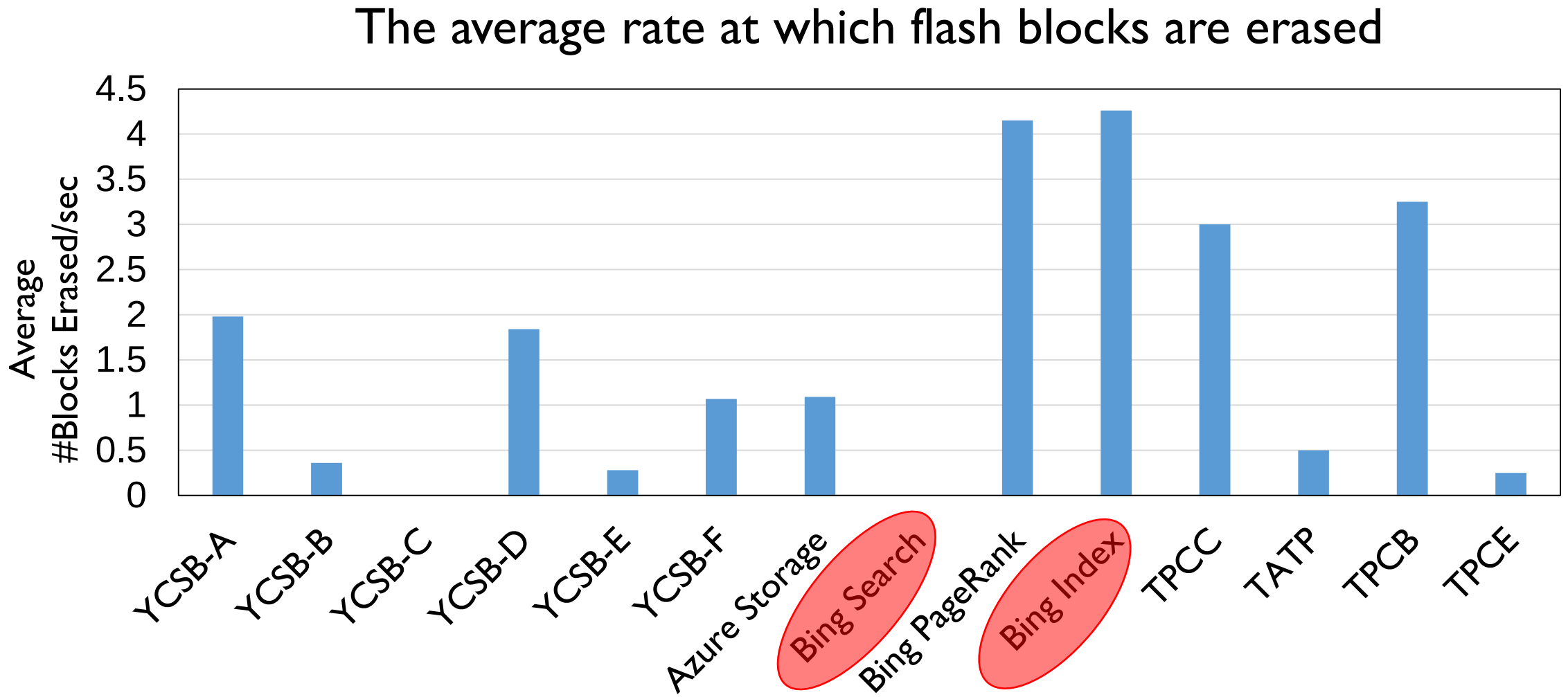
Impact of Hardware Isolation on SSD Lifetime



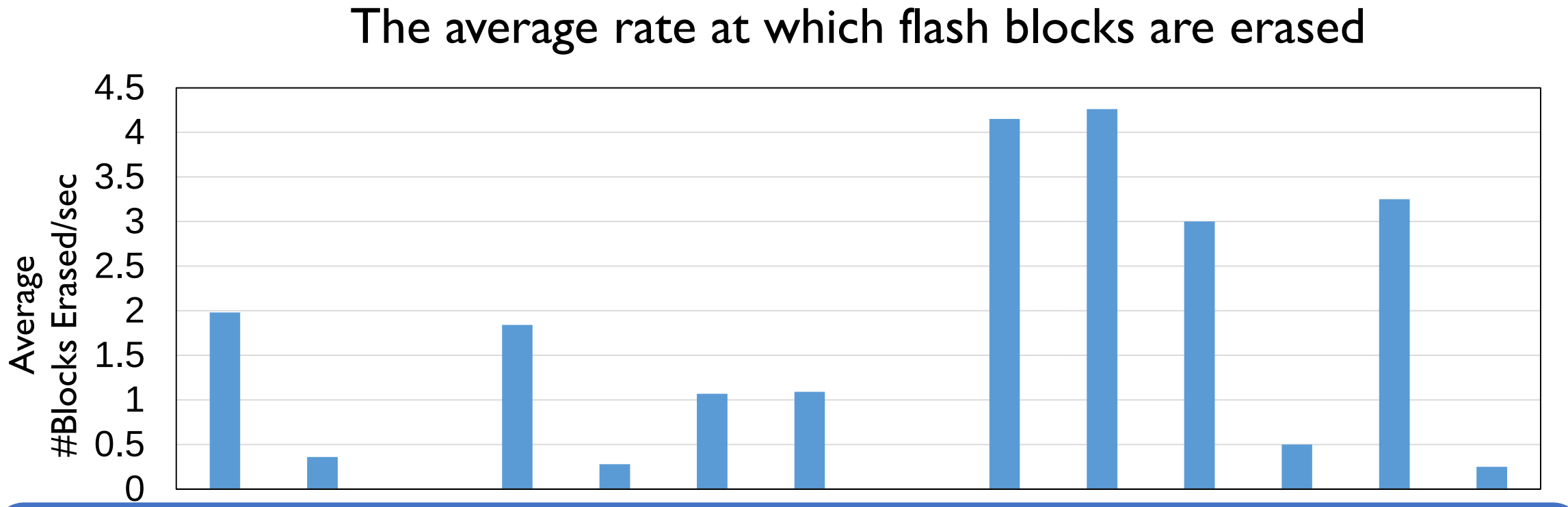
Impact of Hardware Isolation on SSD Lifetime



Impact of Hardware Isolation on SSD Lifetime

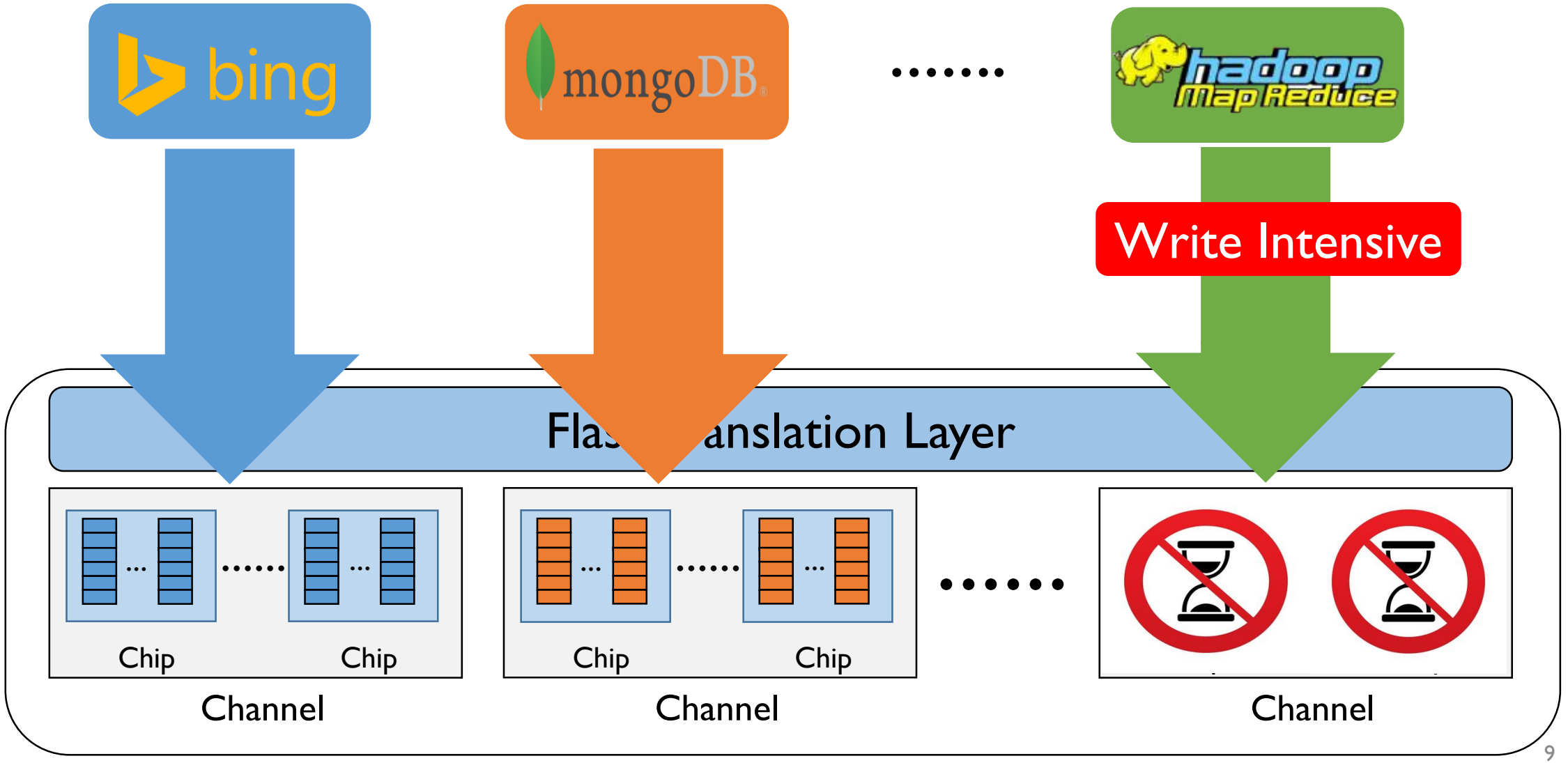


Impact of Hardware Isolation on SSD Lifetime

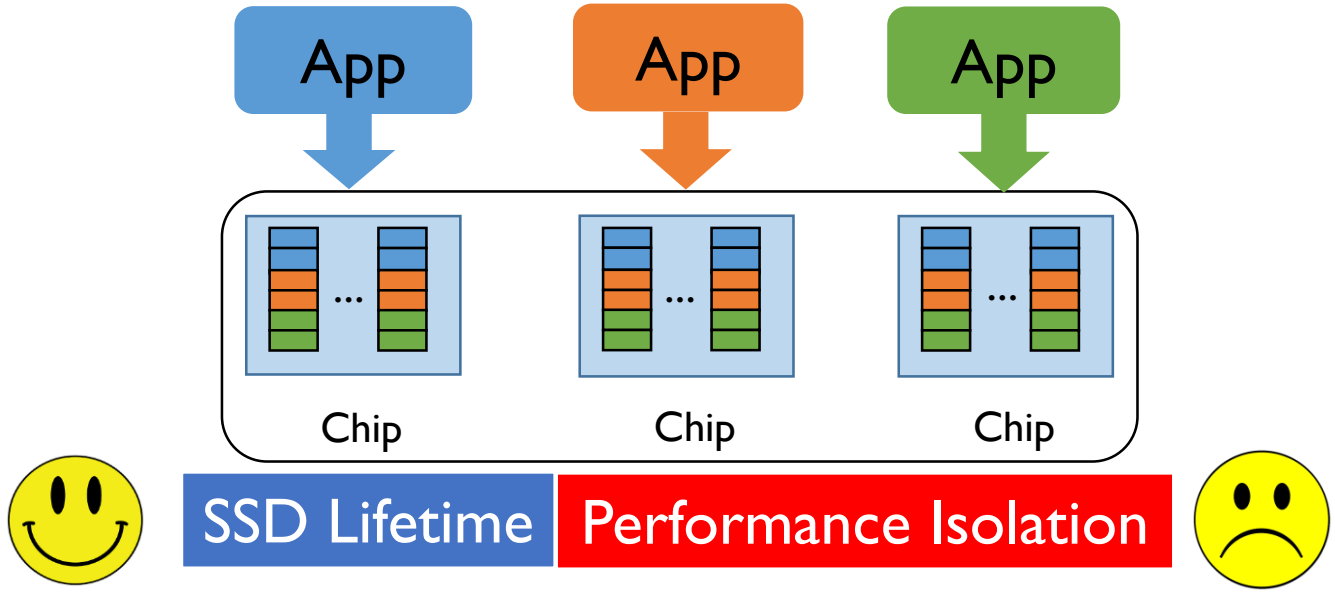


Flash blocks wear out at different rate with different workload

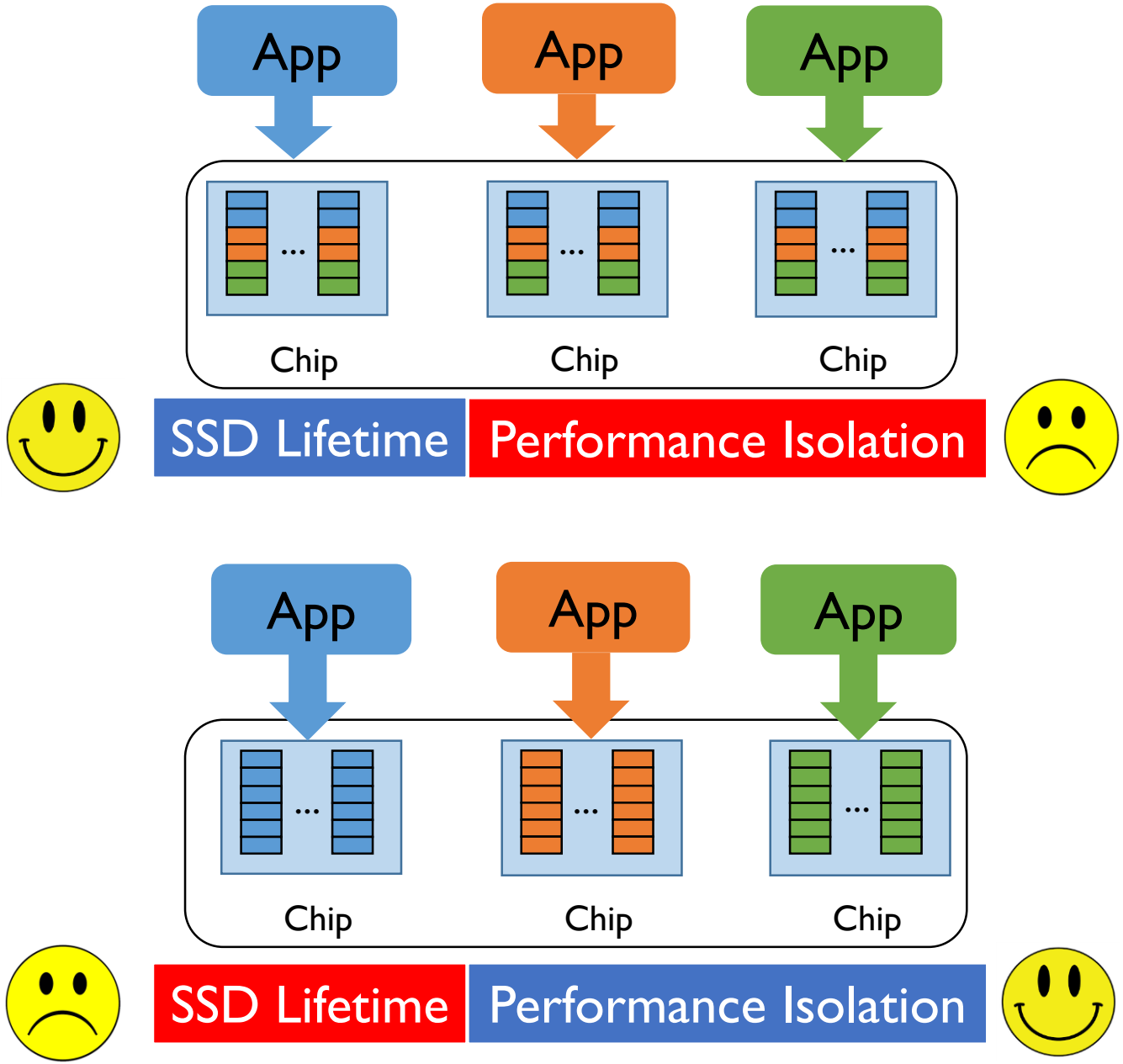
Impact of Hardware Isolation on SSD Lifetime



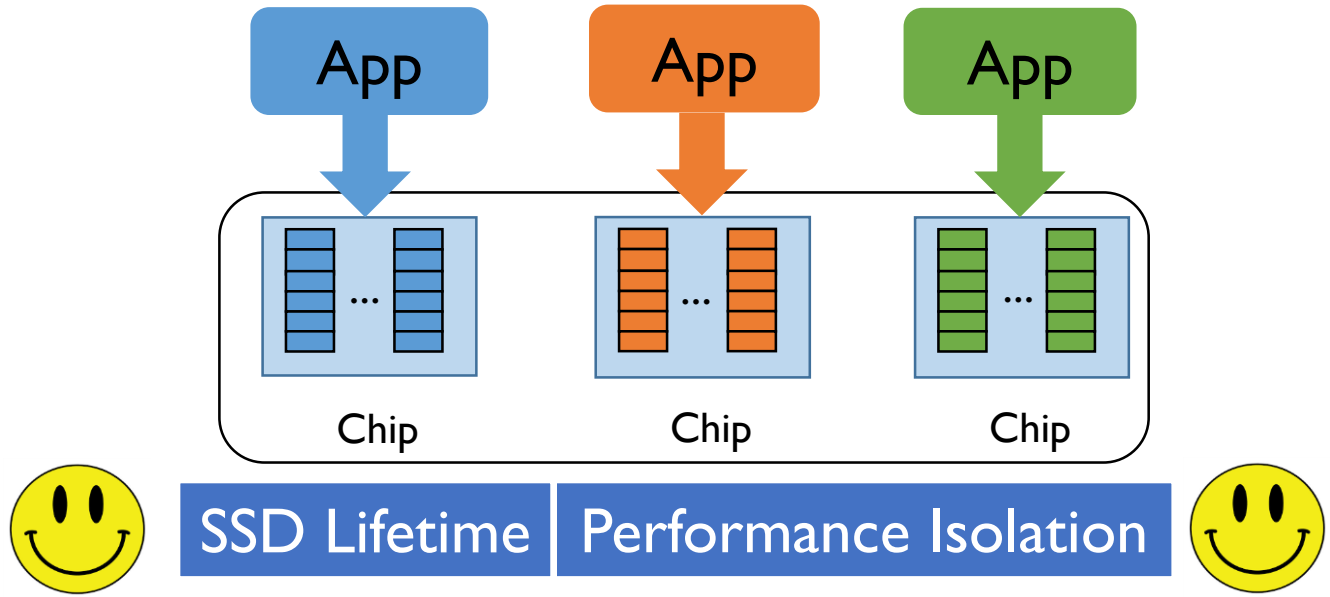
FlashBlox Challenges



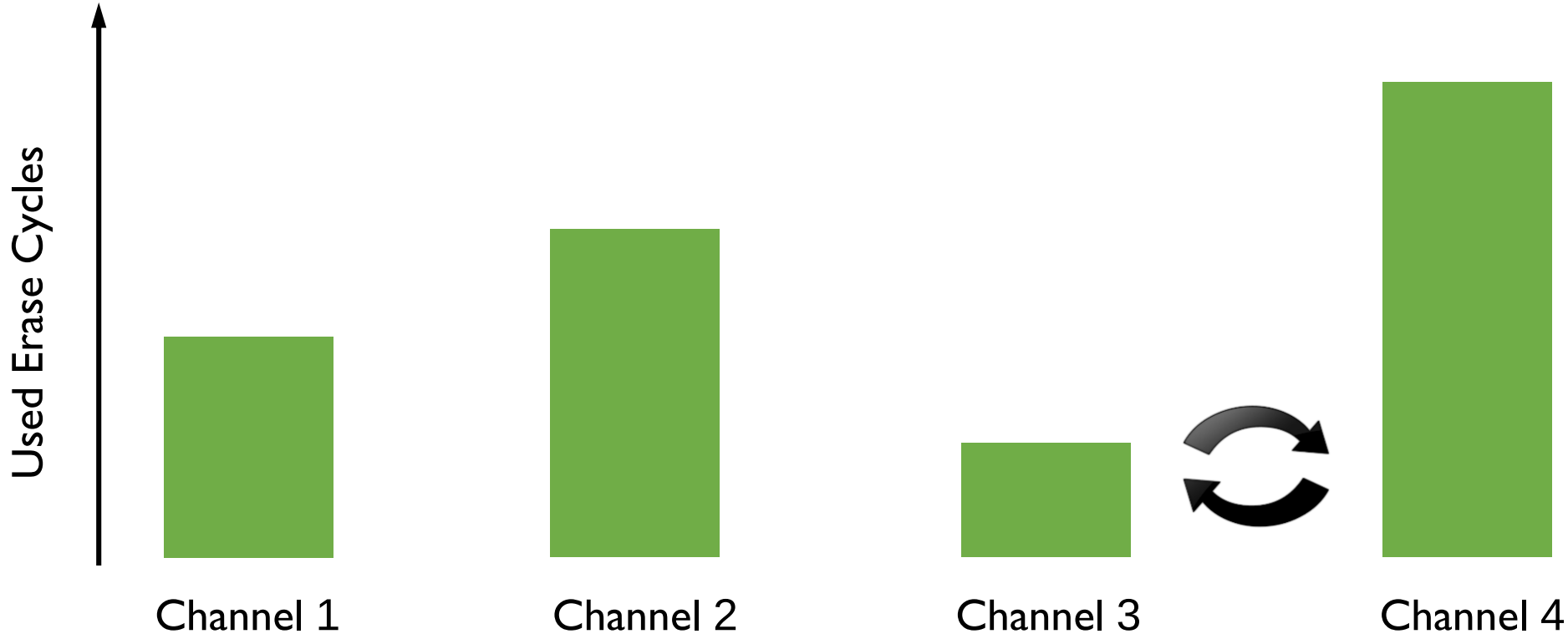
FlashBlox Challenges



FlashBlox Challenges

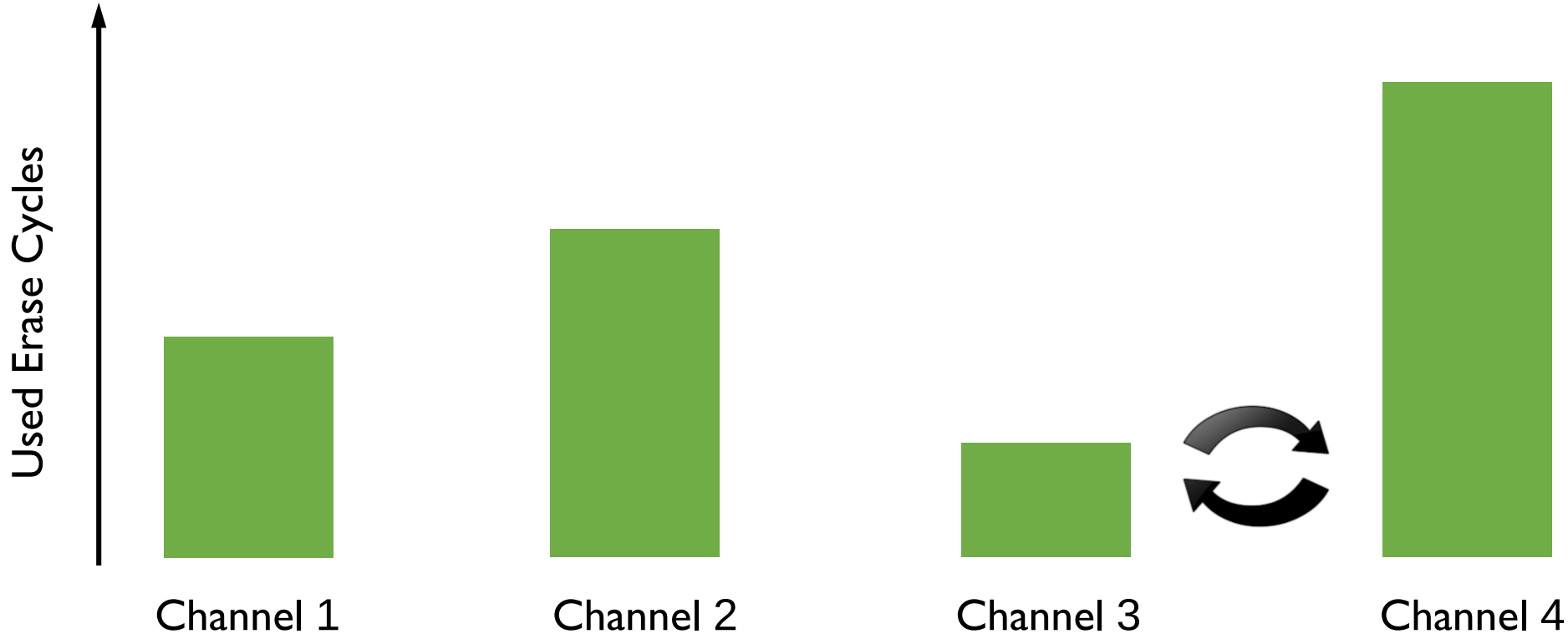


FlashBlox: Swapping Channels for Wear Balance



Adjusting the wear imbalance at a more coarse time granularity can achieve near-ideal SSD lifetime

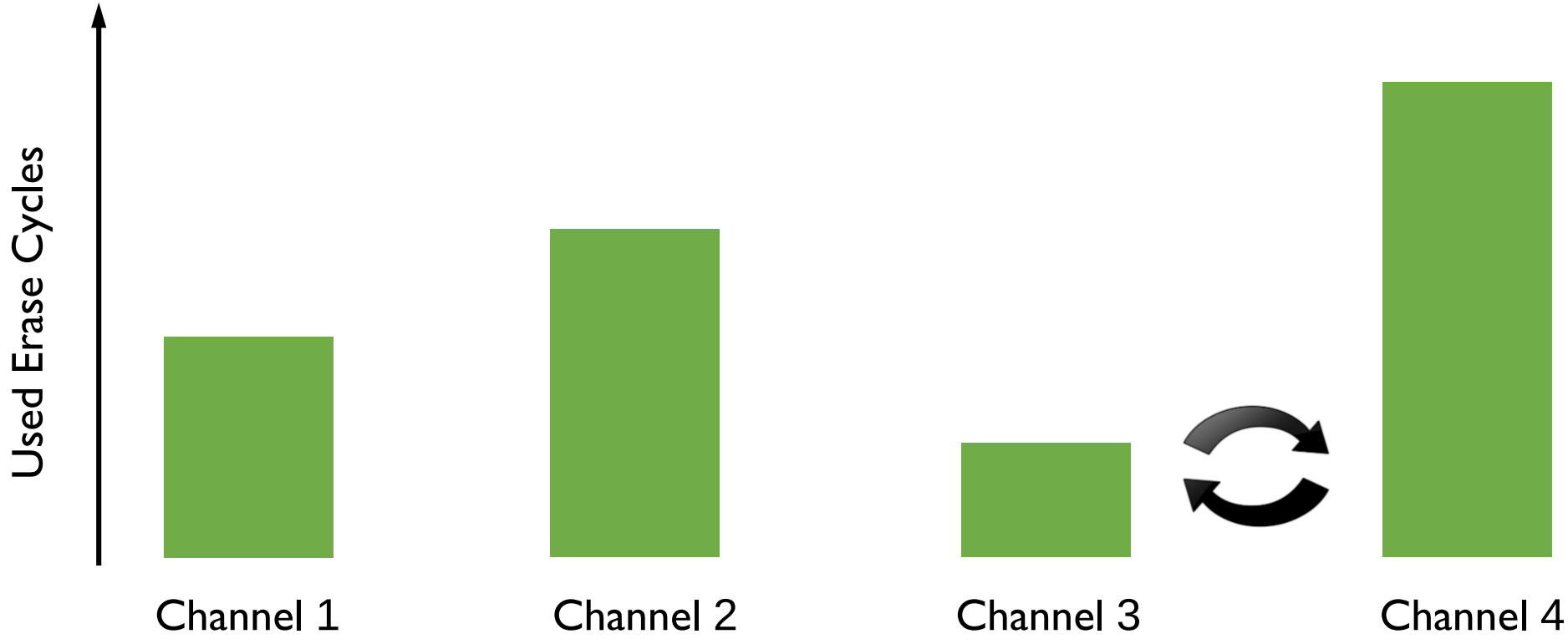
FlashBlox: Swapping Channels for Wear Balance



The channel that has incurred the maximum wearout

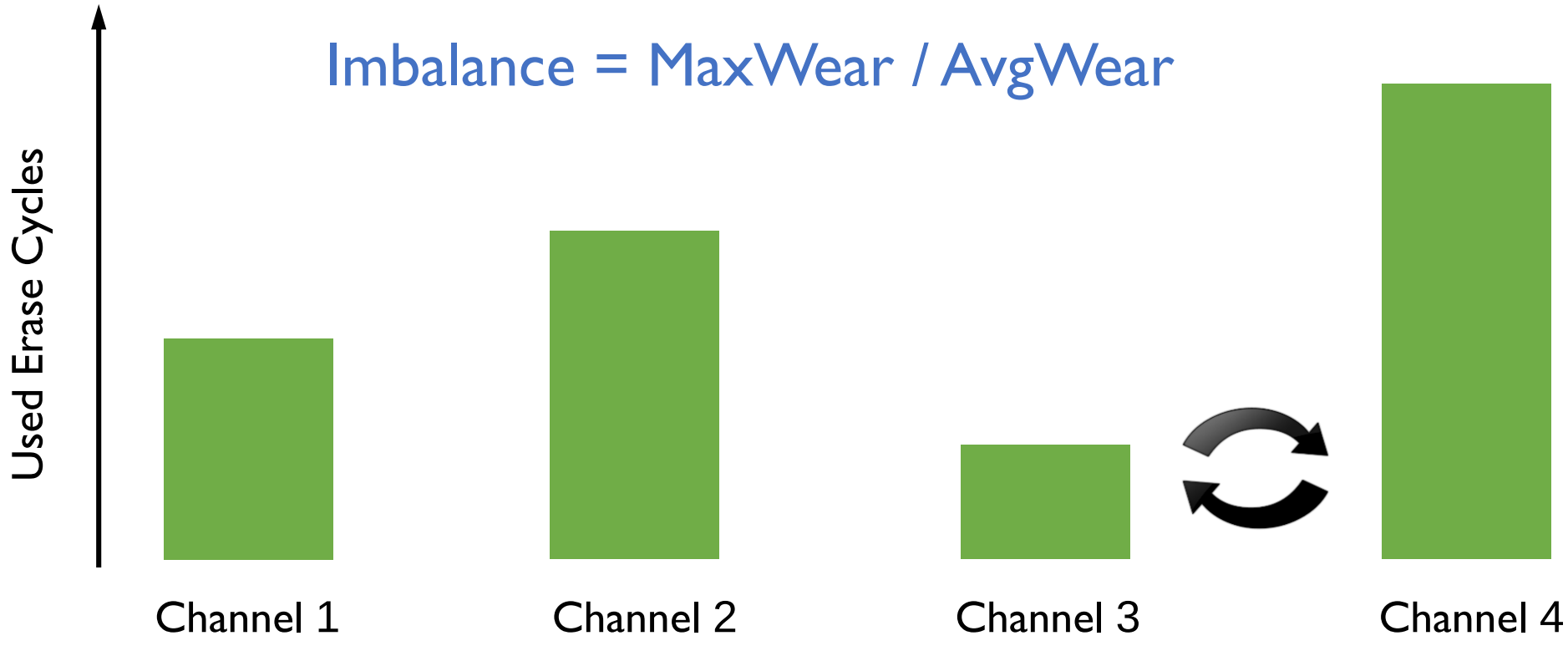
The channel that has the minimum rate of wearout

FlashBlox: Swapping Channels for Wear Balance

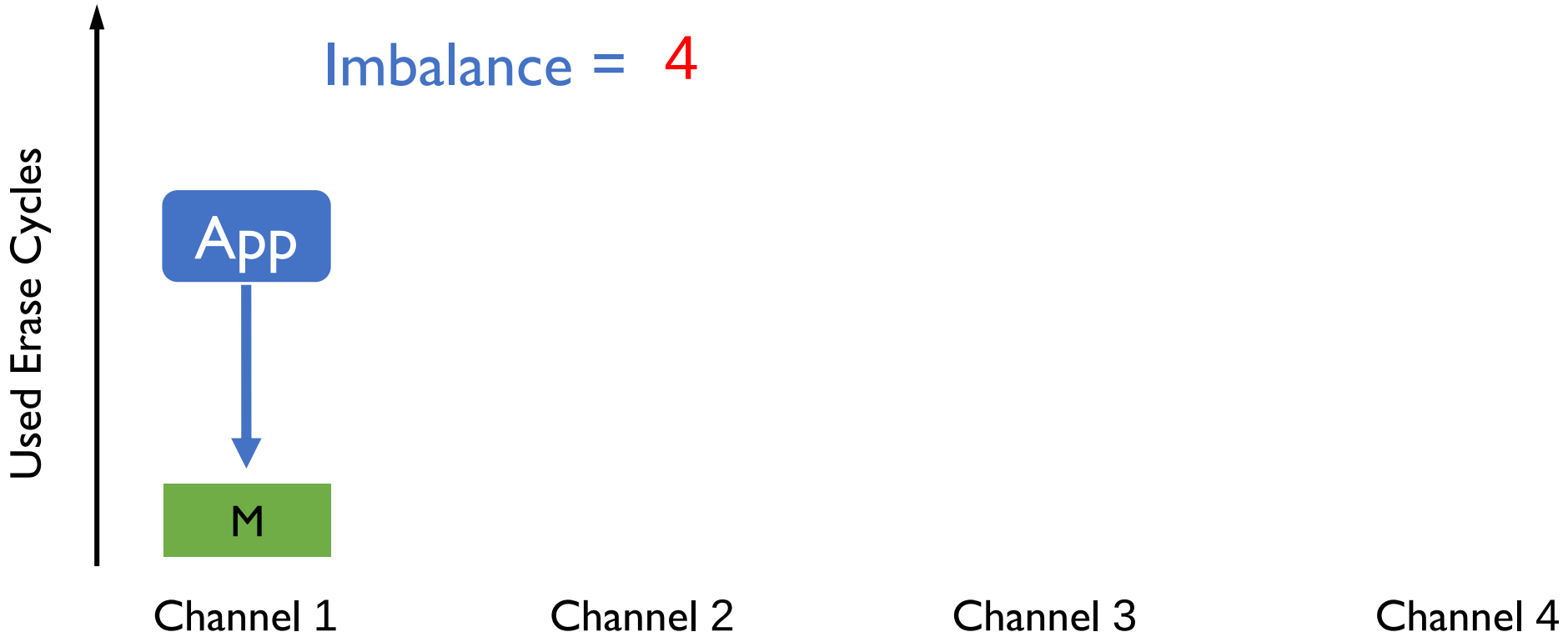


Channel migration takes 15 minutes, once per 19 days
Overall performance drops only for 0.04% of all the time

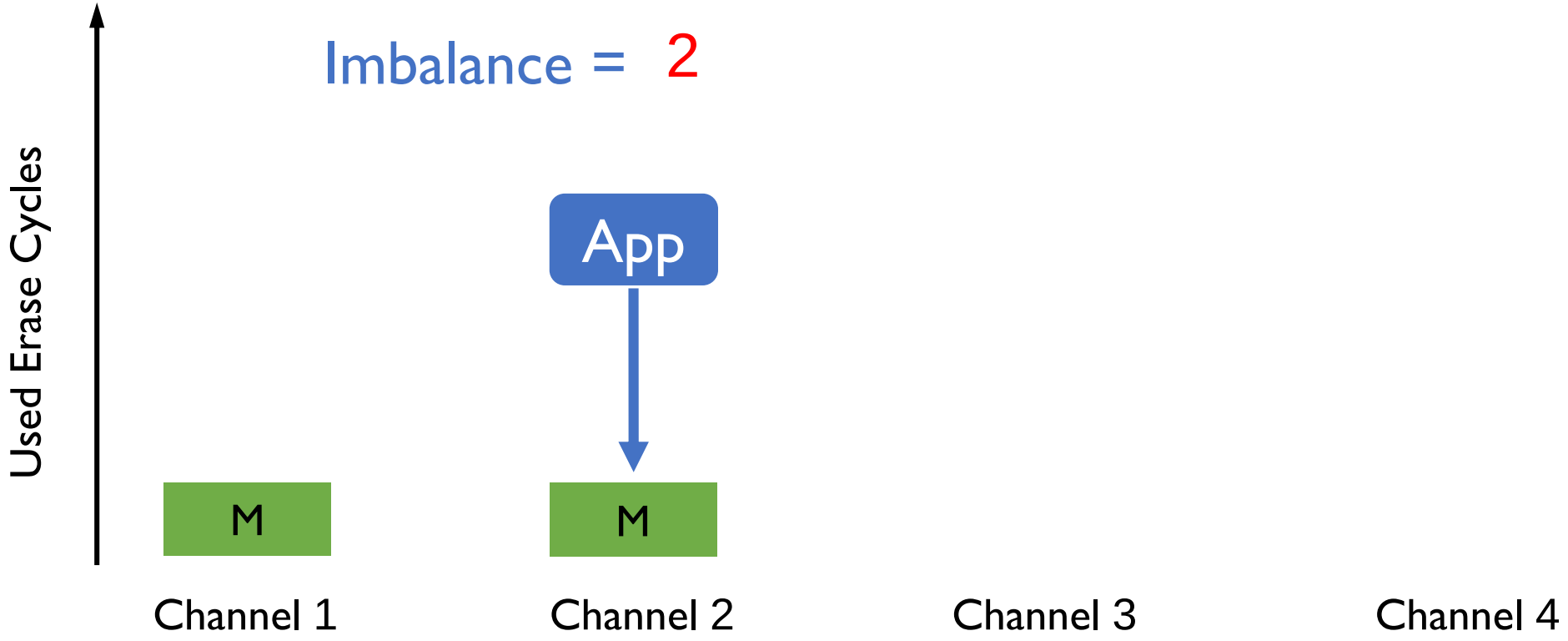
How Frequently Should We Swap?



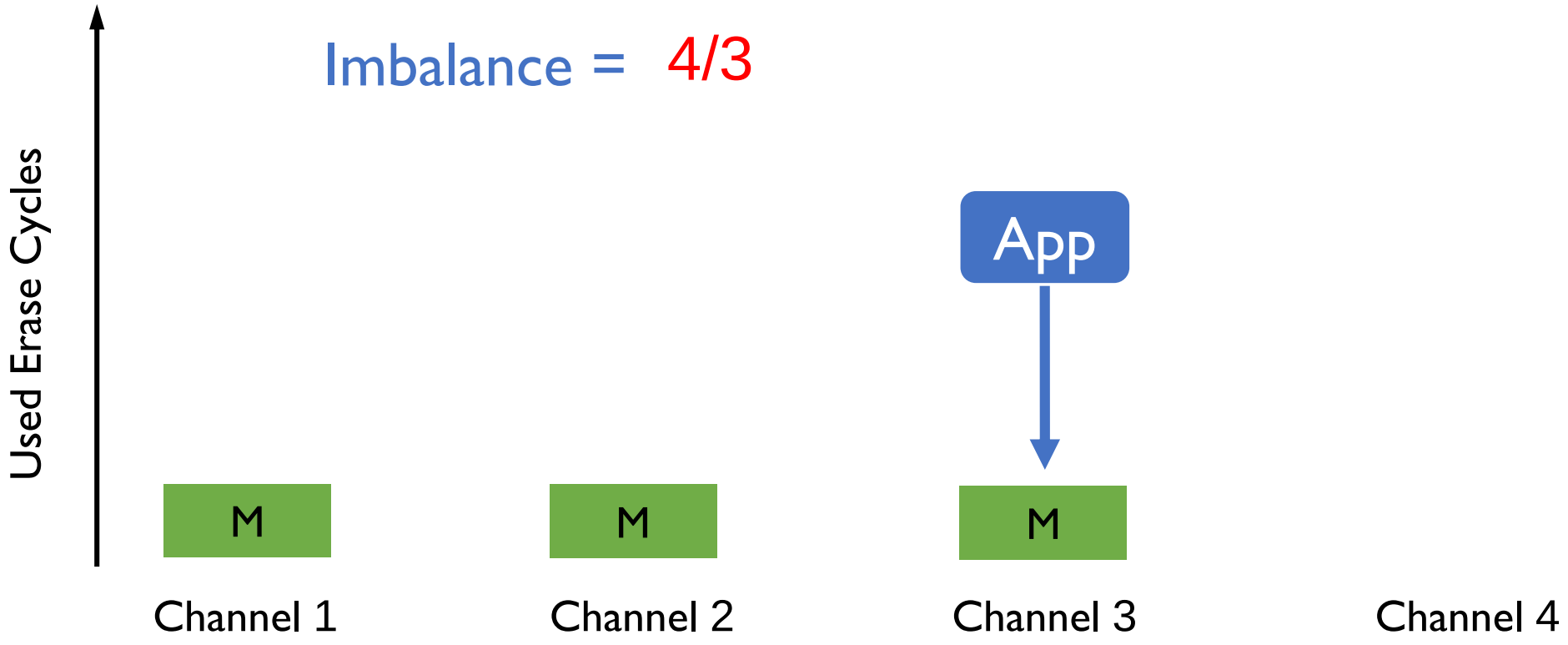
How Frequently Should We Swap?



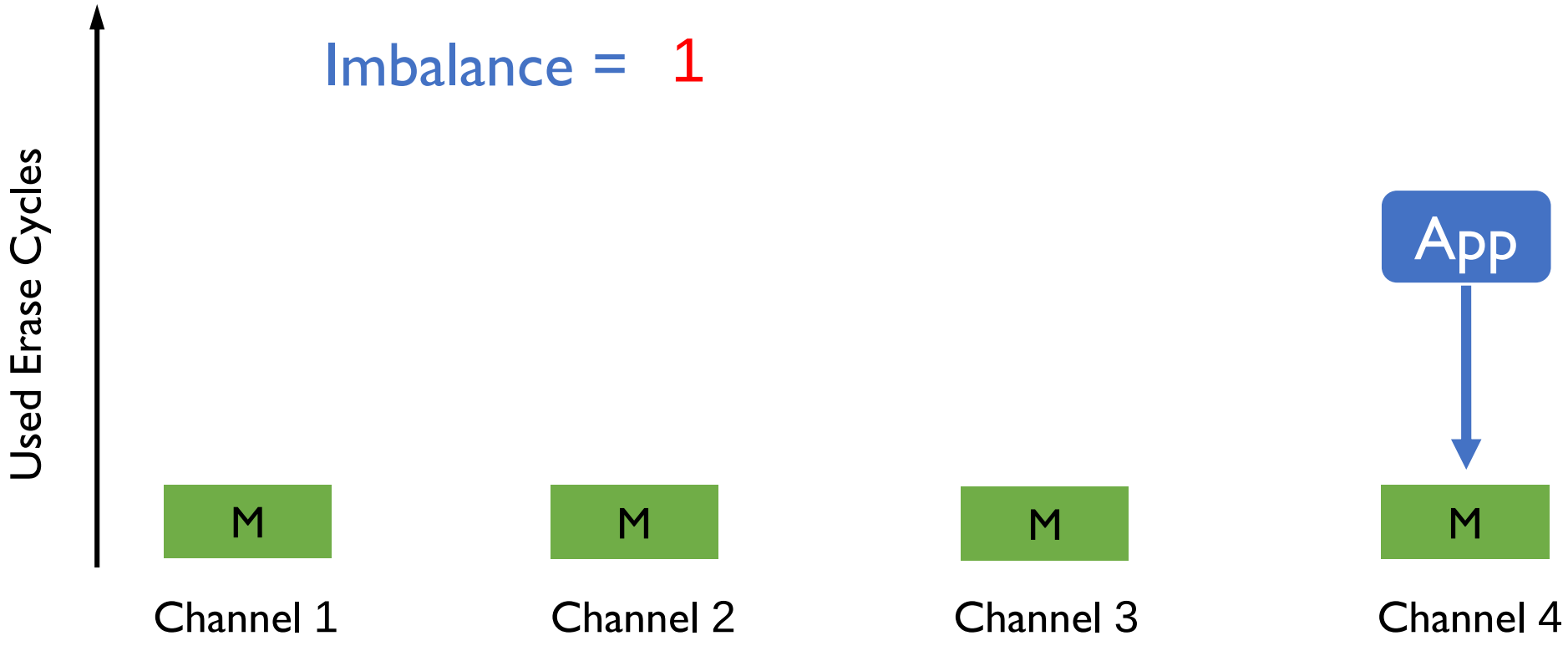
How Frequently Should We Swap?



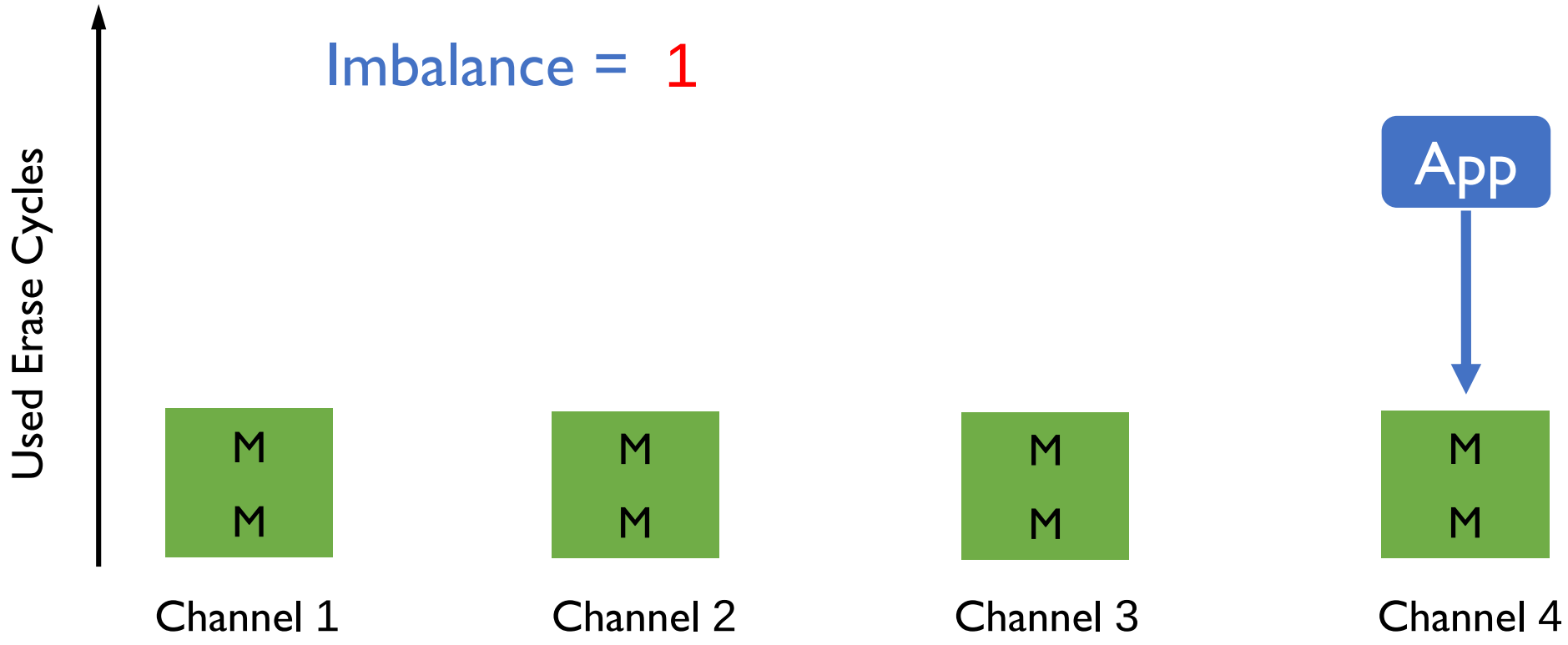
How Frequently Should We Swap?



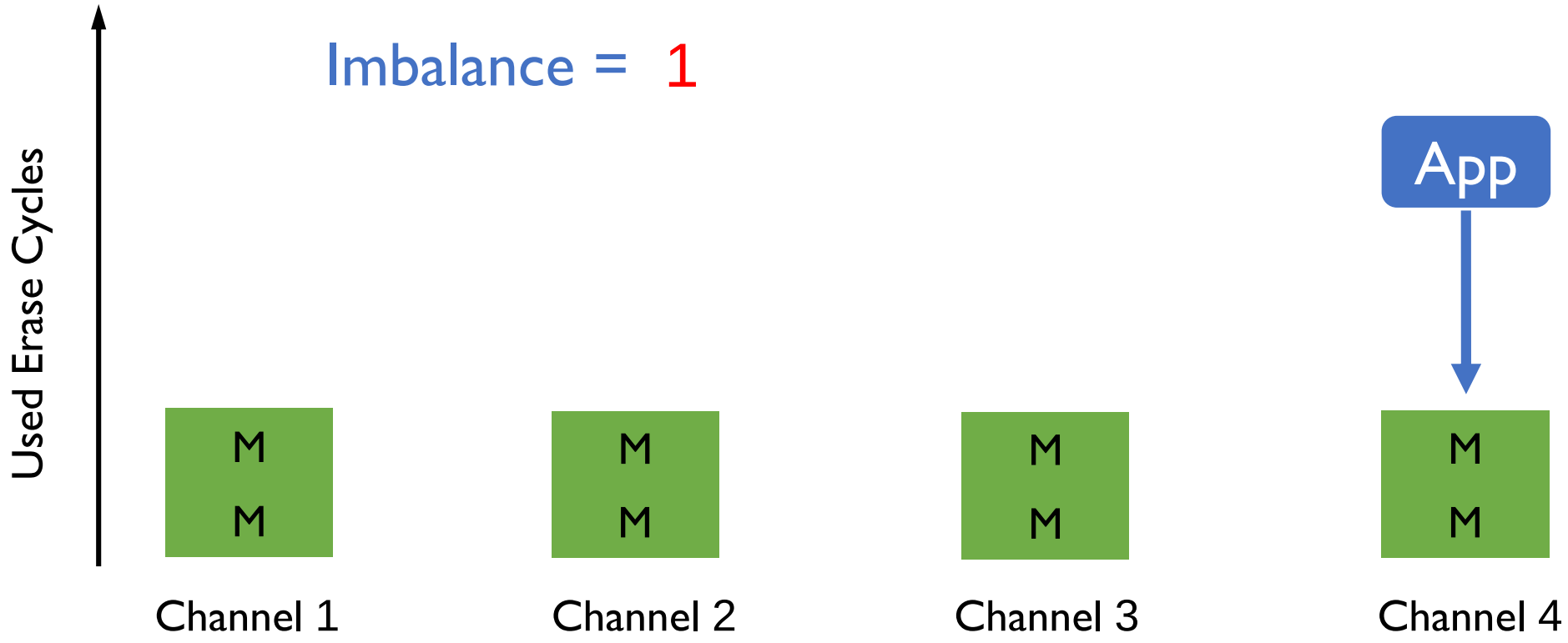
How Frequently Should We Swap?



How Frequently Should We Swap?



How Frequently Should We Swap?



How many times should we swap within SSD lifetime?

Quantifying the Swapping Frequency

Assume there are N channels,

wear imbalance target: 1+x

after K rounds of cycling:

$$\text{Wear Imbalance} = (MK + M)/(MK + M/N) = (K + 1)/(K + 1/N) \leq (1 + x)$$

Maximum Wearout

Average Wearout

Quantifying the Swapping Frequency

Assume there are N channels,

wear imbalance target: $1+x$

after K rounds of cycling:

$$\text{Wear Imbalance} = (MK + M)/(MK + M/N) = (K + 1)/(K + 1/N) \leq (1 + x)$$

$$K \geq (N - 1 - x) / (Nx)$$

Quantifying the Swapping Frequency

Assume there are N channels,

wear imbalance target: 1+x

after K rounds of cycling:

$$\text{Wear Imbalance} = (MK + M)/(MK + M/N) = (K + 1)/(K + 1/N) \leq (1 + x)$$

$$K \geq (N - 1 - x) / (Nx)$$

If N = 16, x = 0.1, then K = 9, which means after swap NK = 148 times,

we can guarantee the wear imbalance is bounded in 1.1

Quantifying the Swapping Frequency

Assume there are N channels,

wear imbalance target: $1+x$

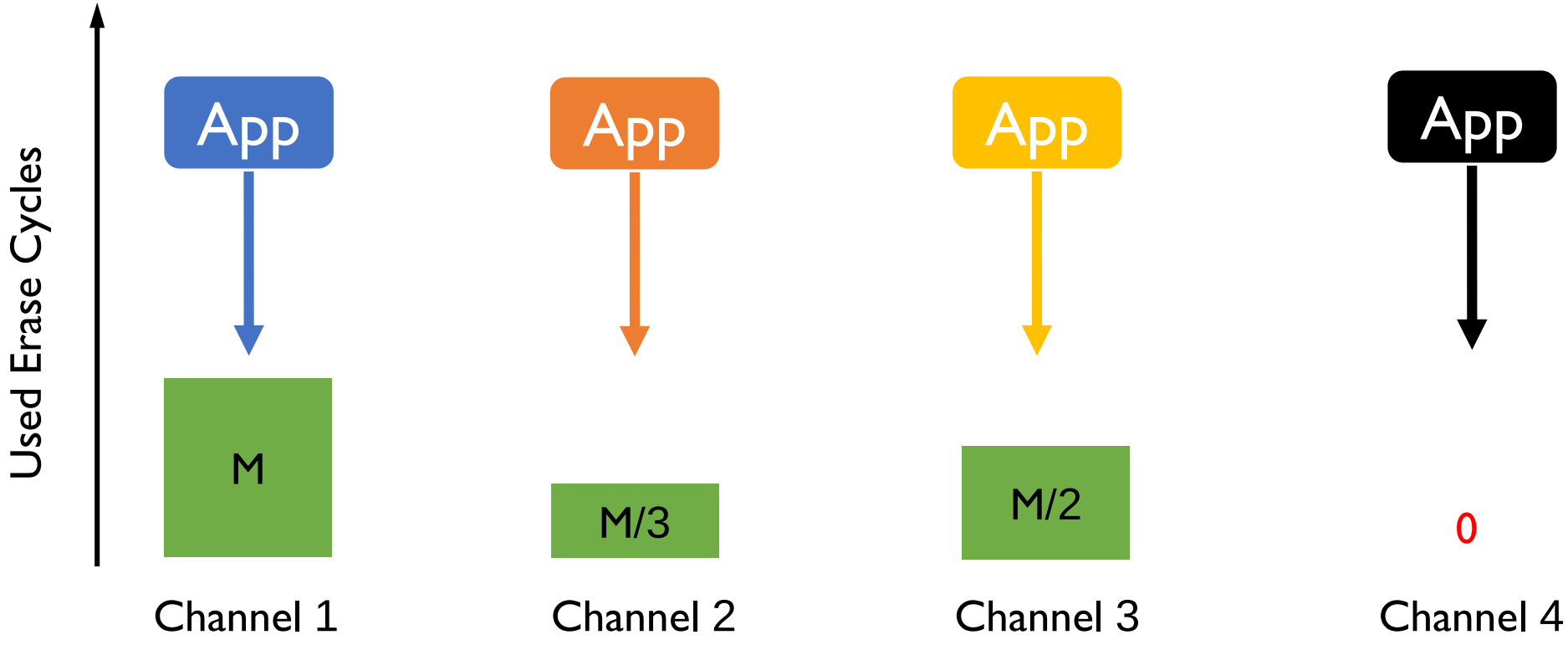
after K rounds of cycling:

$$\text{Wear Imbalance} = (MK + M)/(MK + M/N) = (K + 1)/(K + 1/N) \leq (1 + x)$$

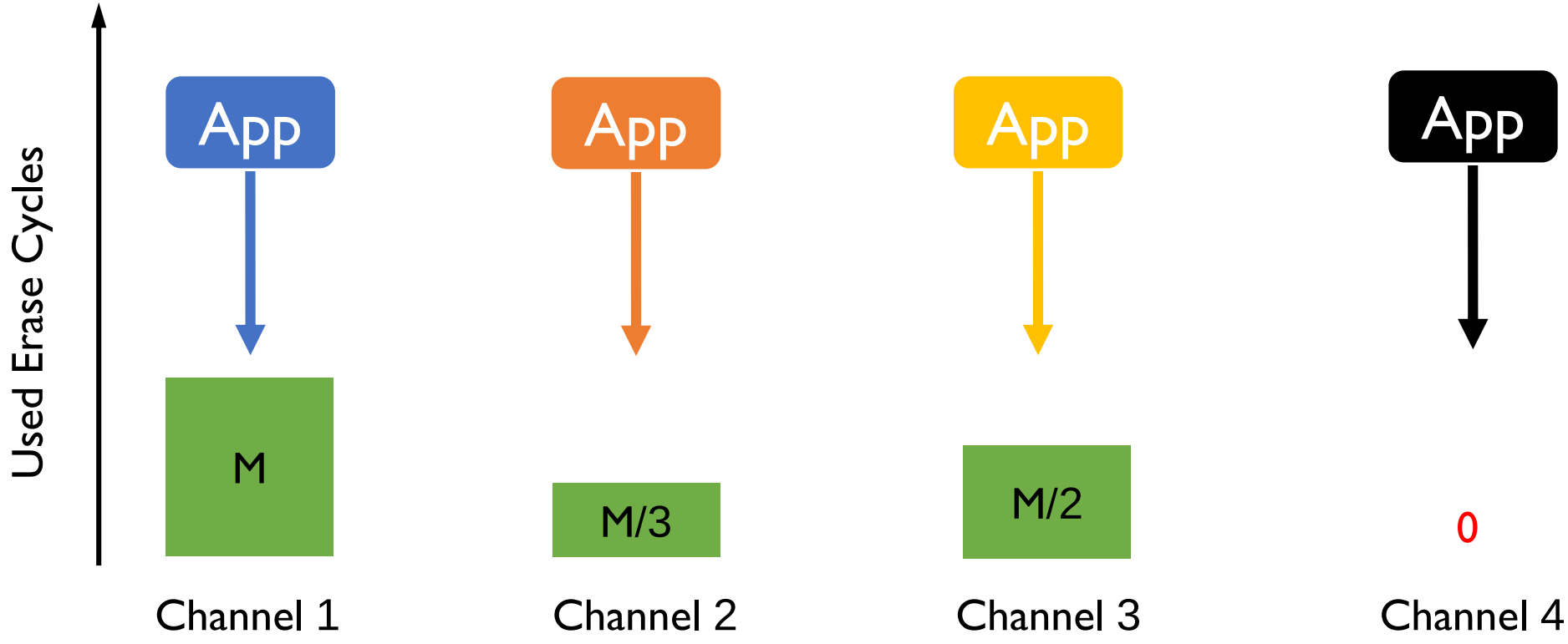
$$K \geq (N - 1 - x) / (Nx)$$

For an SSD with 5 years lifetime, swap once per 12 days can guarantee the channels are well balanced for worst case

Adaptive Wear Leveling in Practice

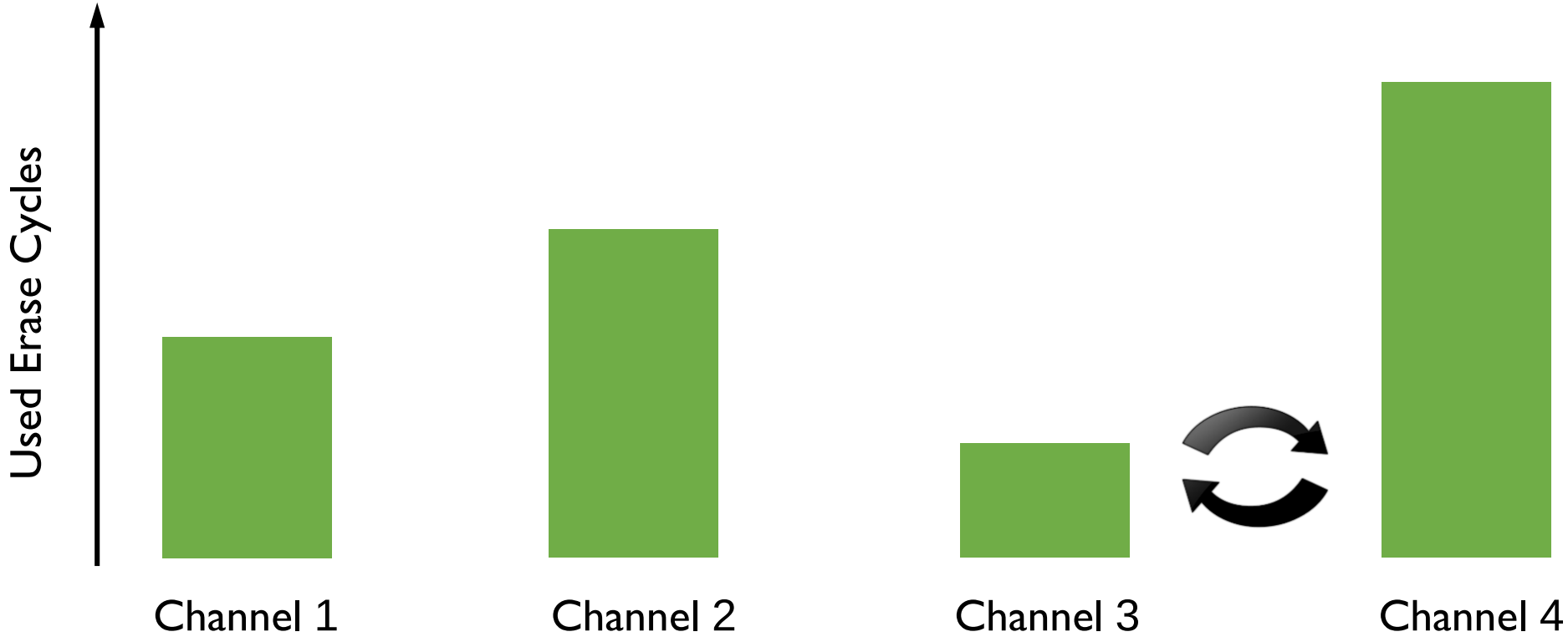


Adaptive Wear Leveling in Practice

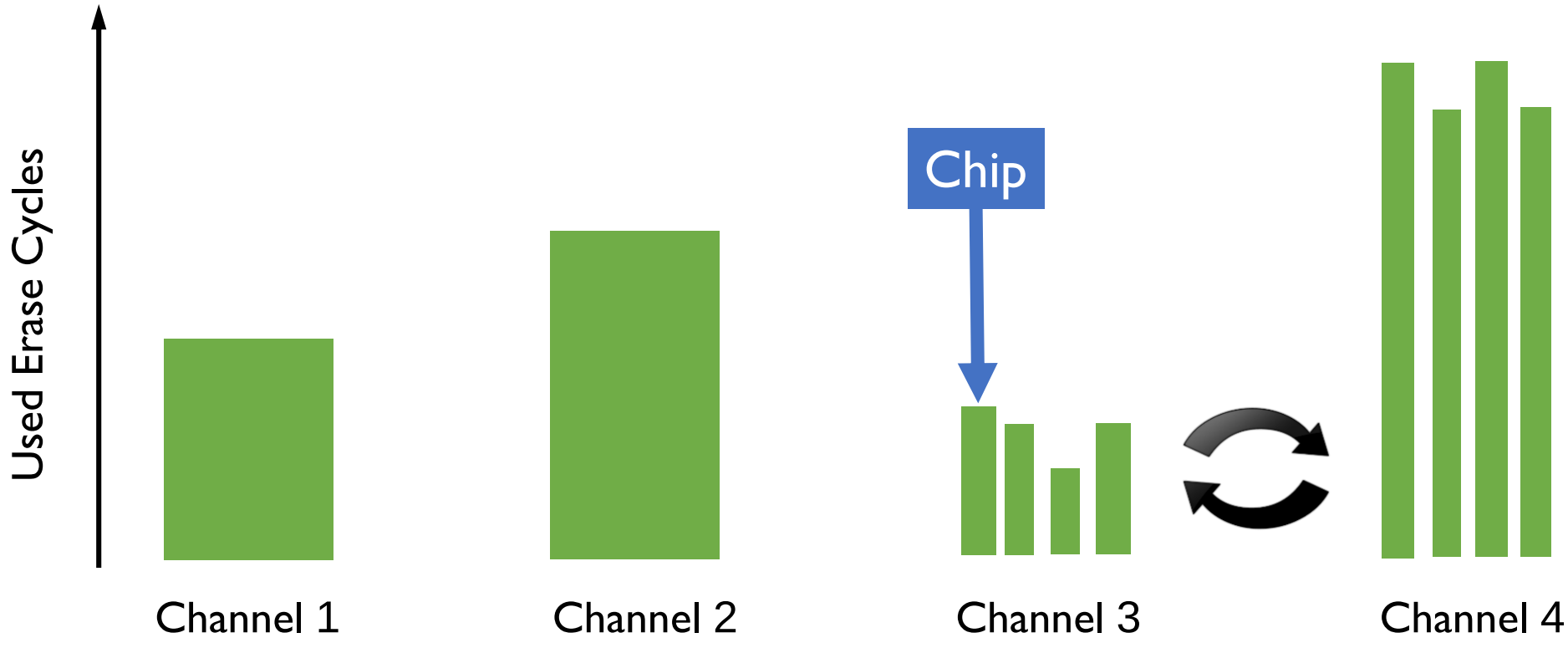


Using erase rate as the trigger condition for swapping

Intra Channel Wear Leveling

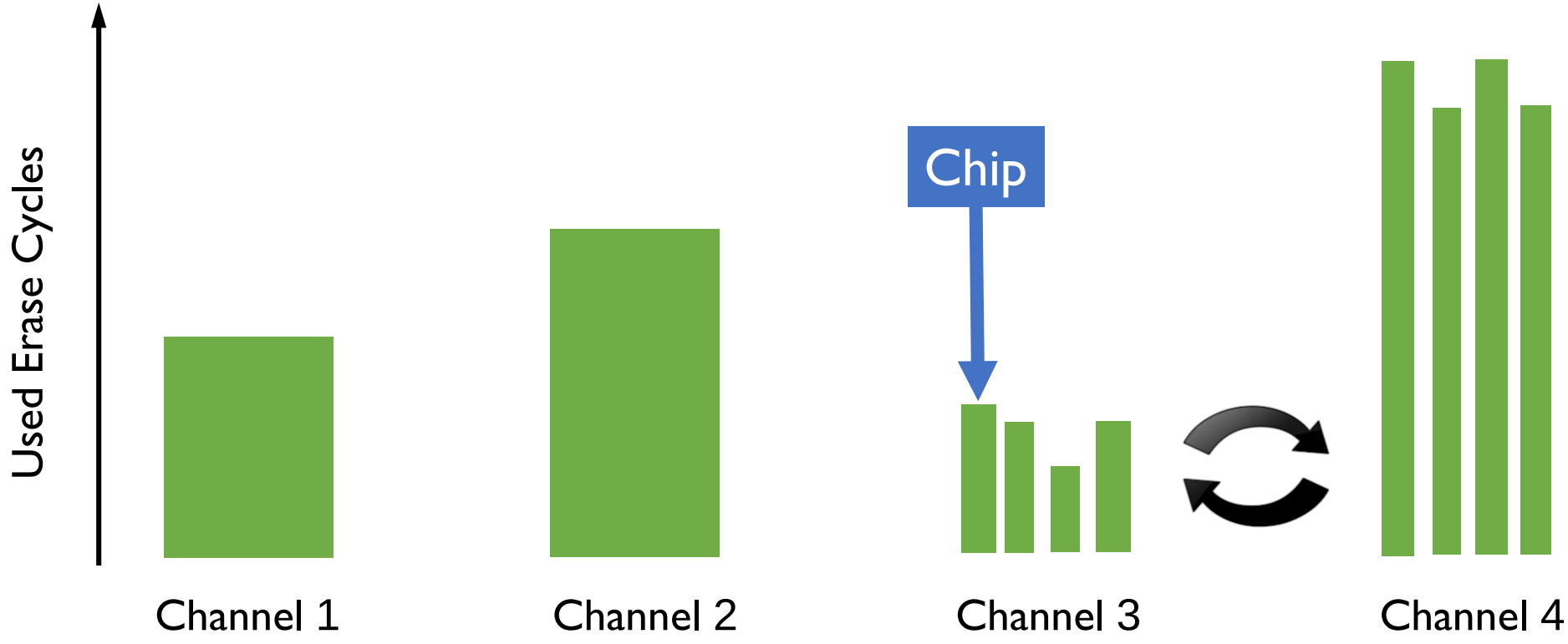


Intra Channel Wear Leveling



Chips will be swapped along with the channel migration

Intra Channel Wear Leveling

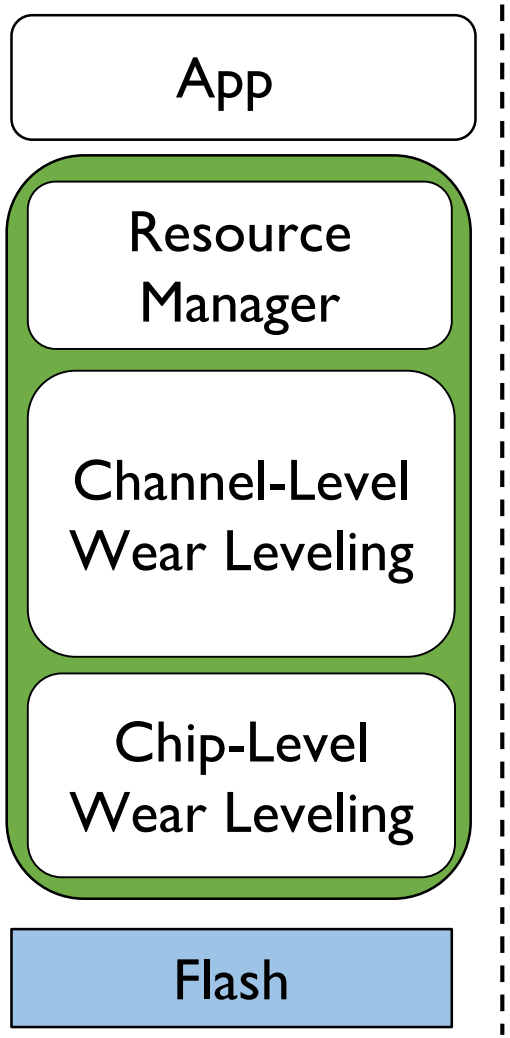


Chips will be swapped along with the channel migration

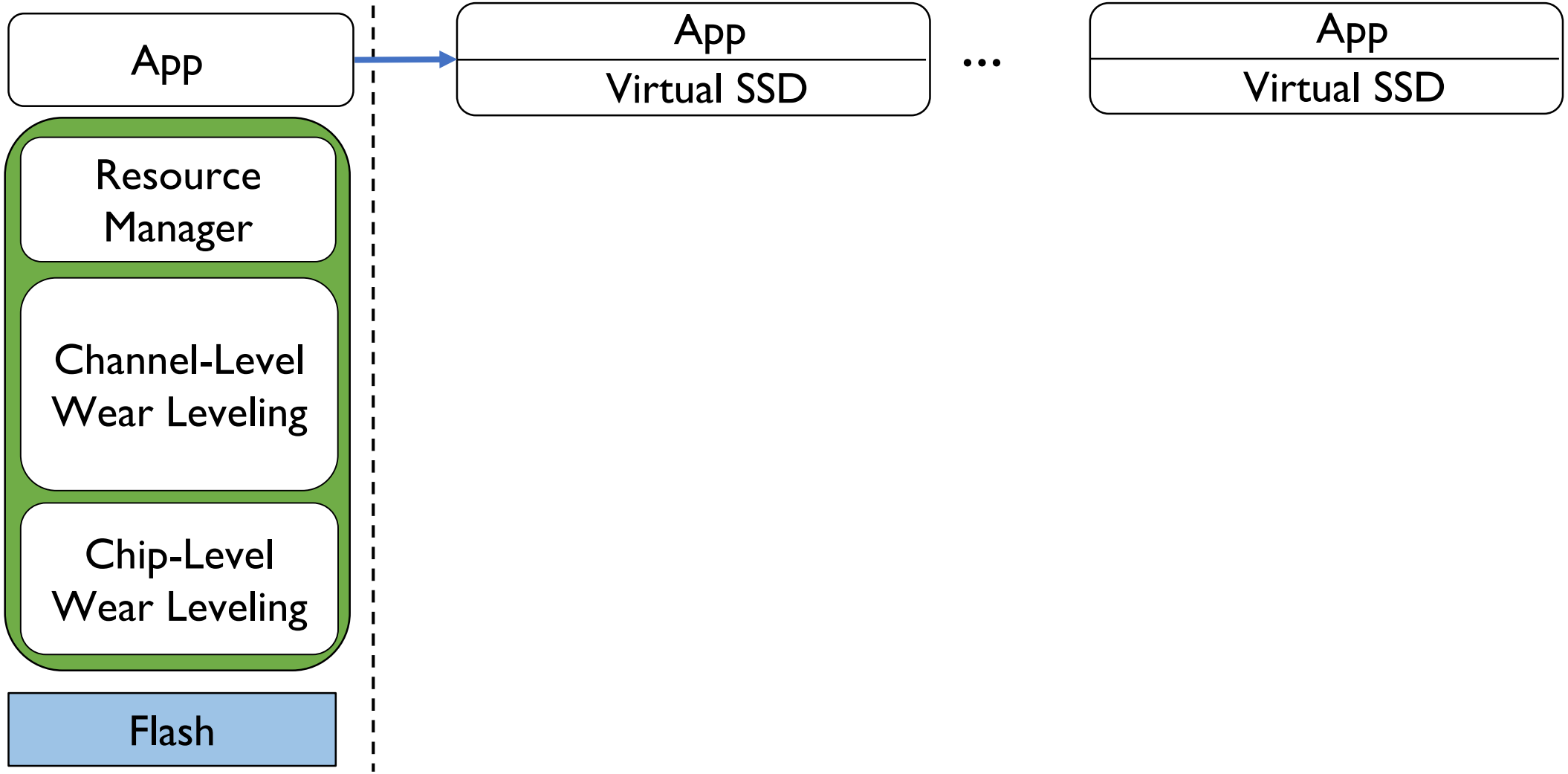
+

Intra-chip wear leveling mechanisms

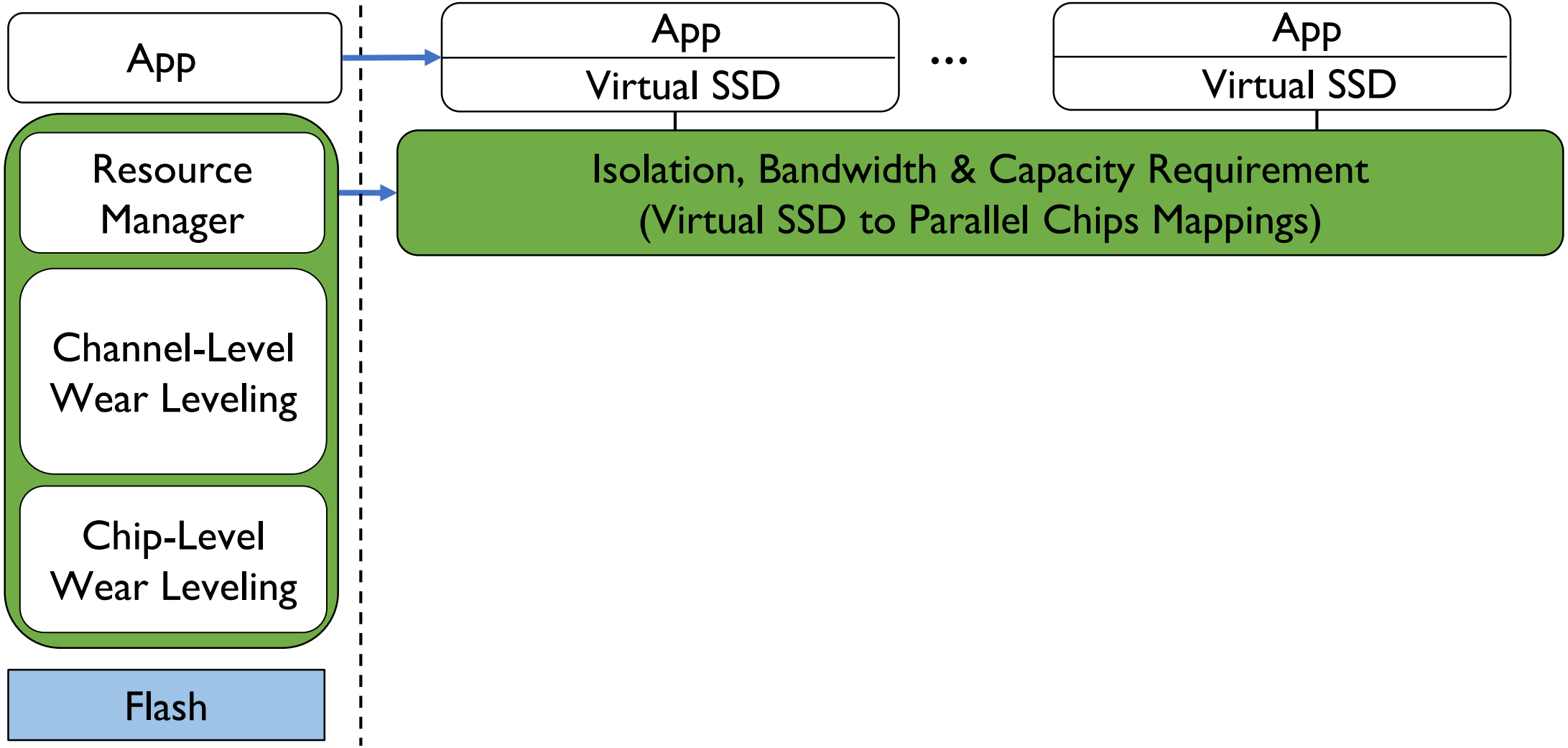
FlashBlox Architecture



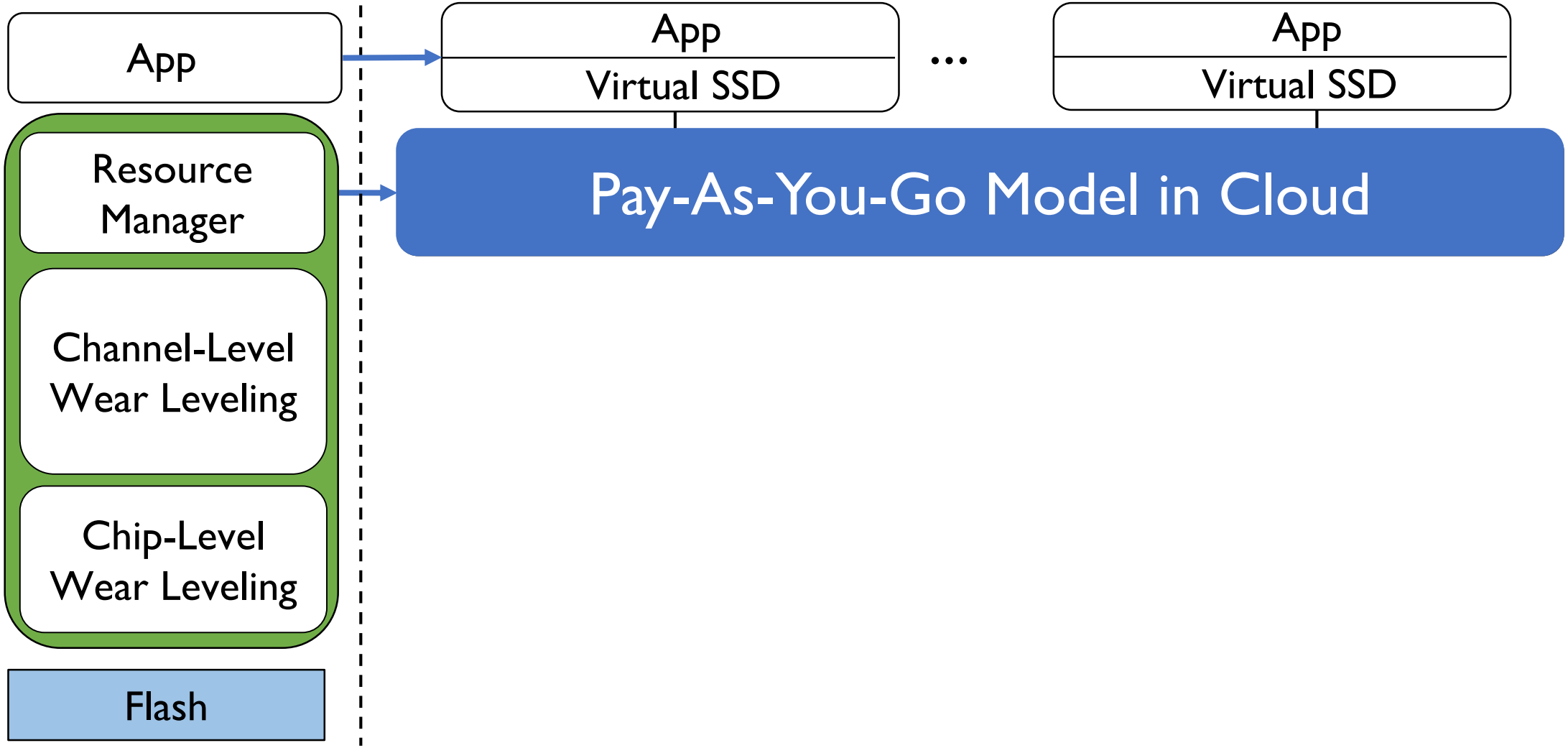
FlashBlox Architecture



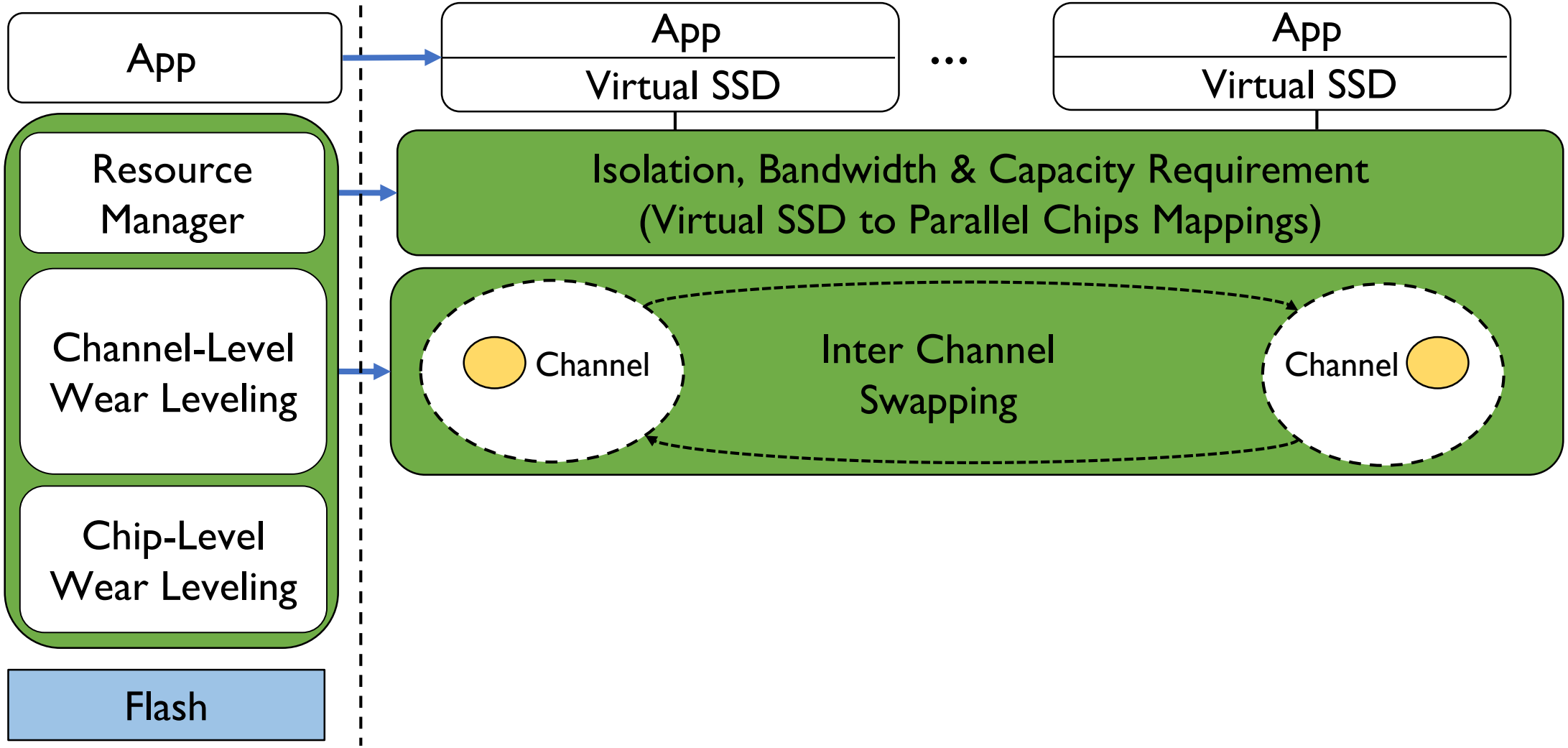
FlashBlox Architecture



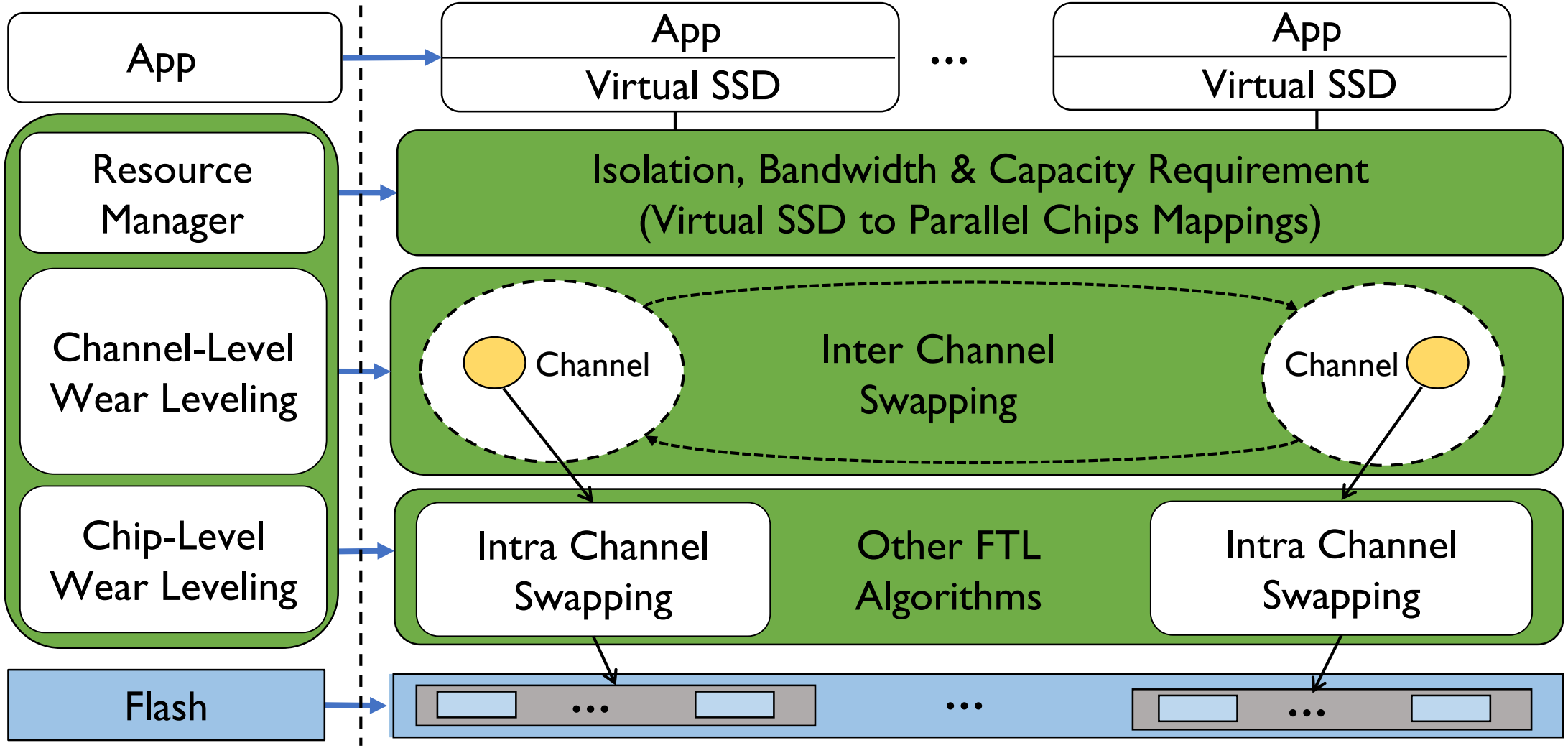
FlashBlox Architecture



FlashBlox Architecture



FlashBlox Architecture



FlashBlox Experimental Setup

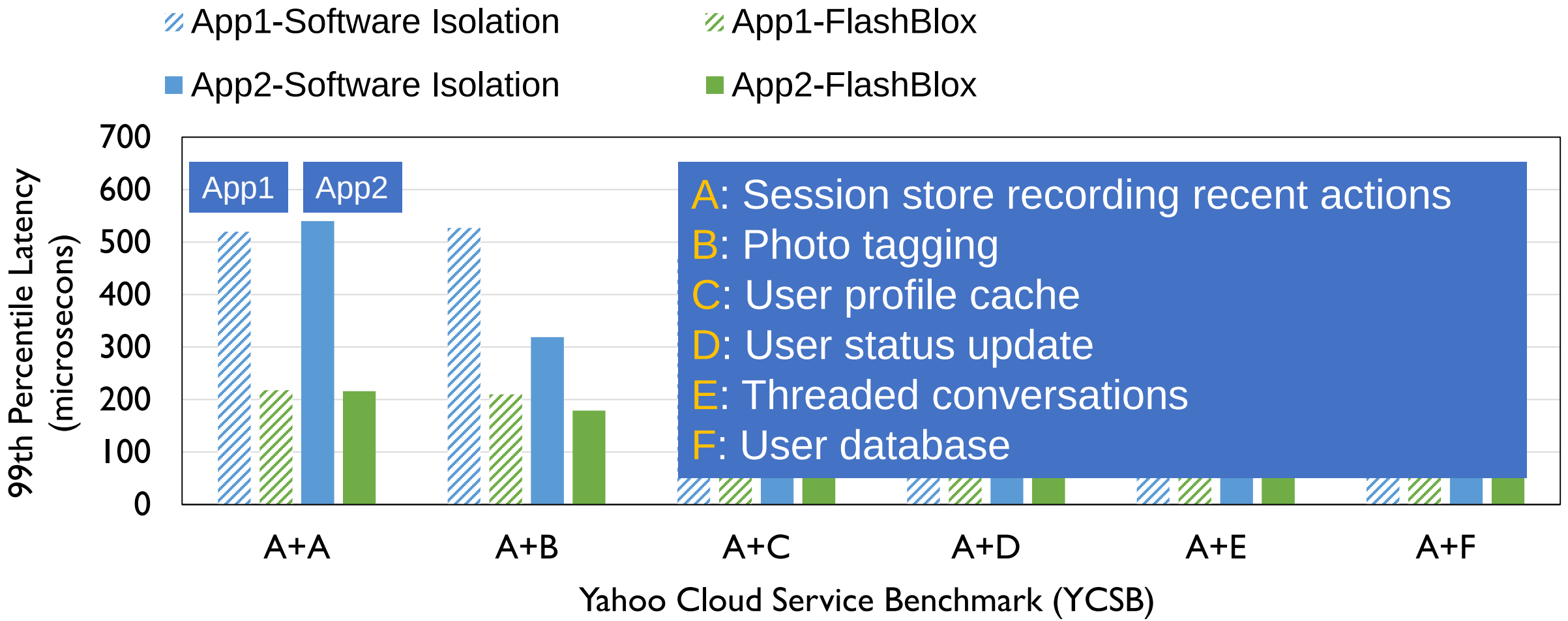
CNEXLABS

16 channels
4 chips
4 planes
16 KB page size

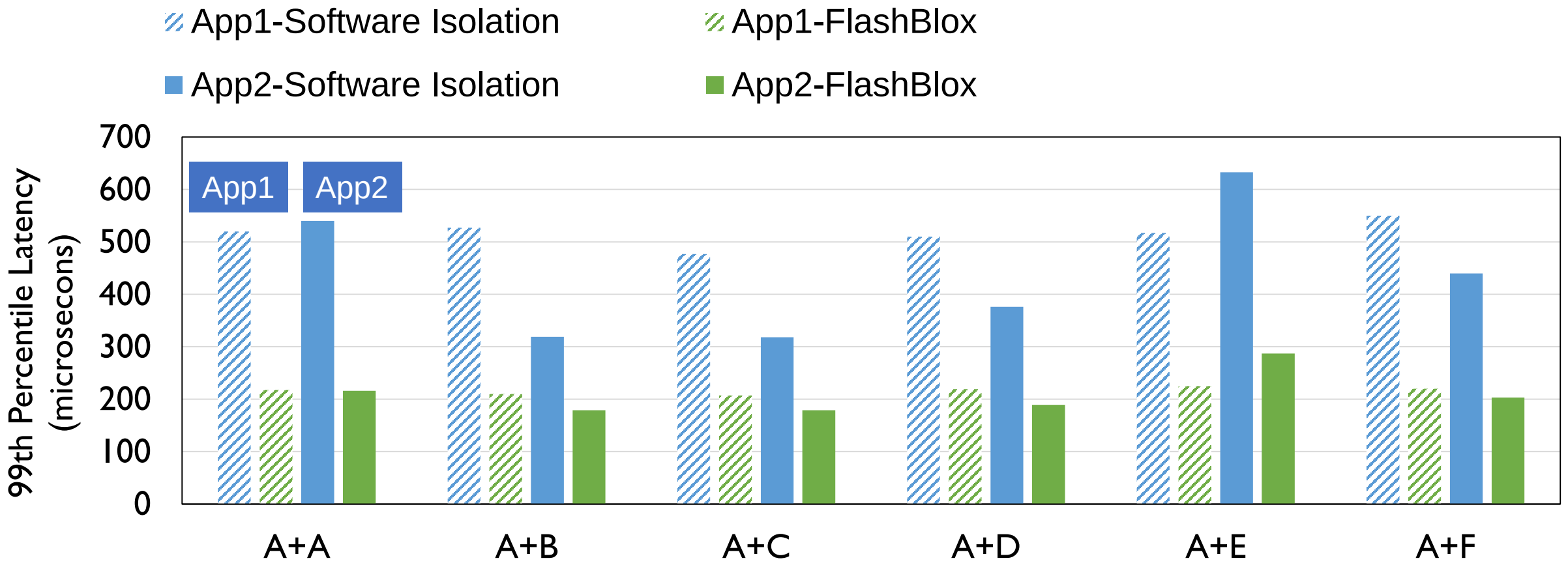
14 data center workloads

Yahoo Cloud Service Benchmark
Bing Search / Index / PageRank
Transactional Database
Azure Storage

Tail Latency Reduction with FlashBlox

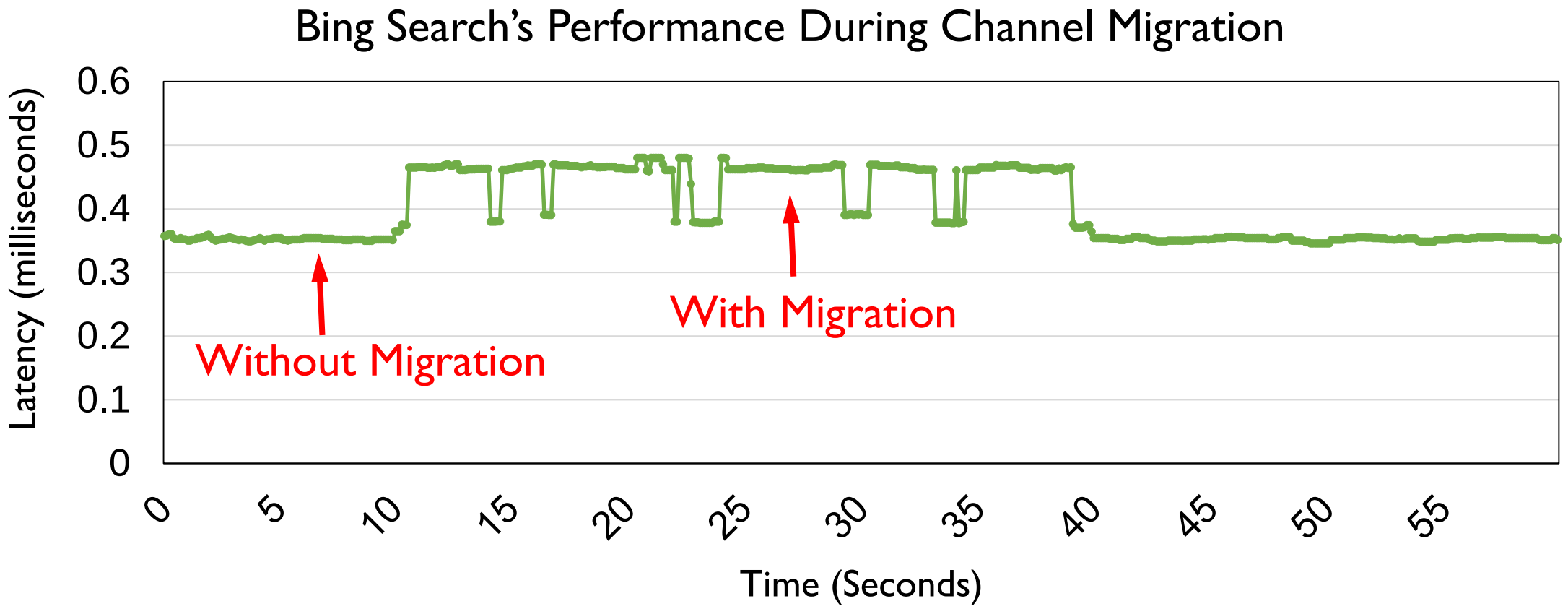


Tail Latency Reduction with FlashBlox

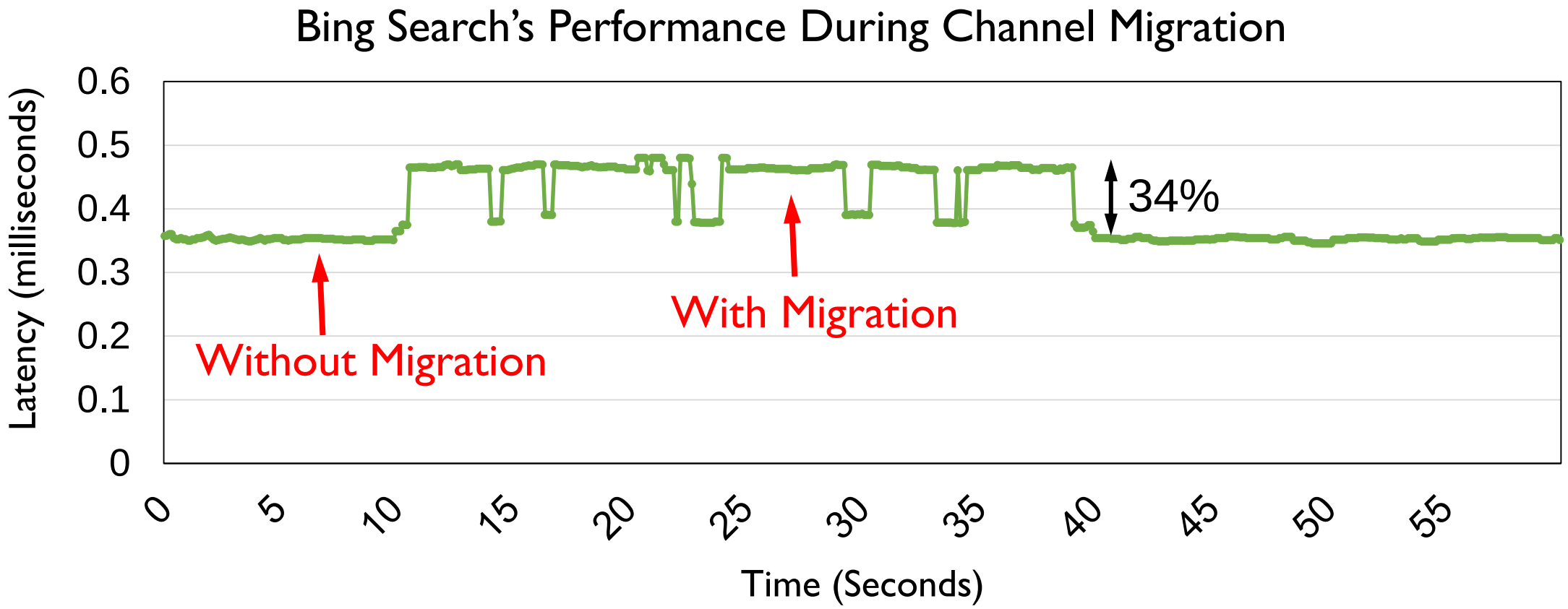


Tail latency reduction: 2.6x, average latency reduction: 1.4x

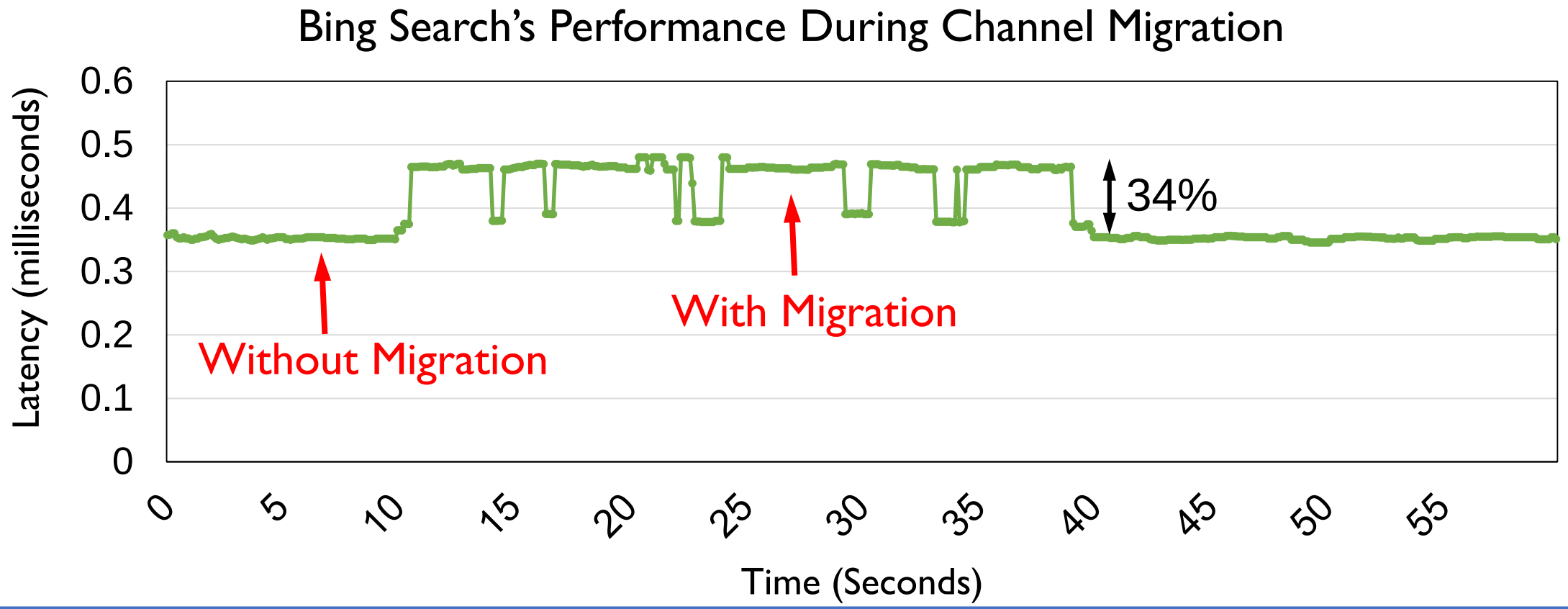
Impact of Channel Migration on Application Performance



Impact of Channel Migration on Application Performance



Impact of Channel Migration on Application Performance



Channel migration takes 15 minutes, once per 19 days
Overall performance drops only for 0.04% of all the time

FlashBlox Summary

CNEXLABS

2.6x reduction on tail latency

Near-ideal SSD lifetime

Swap once per 19 days

Thanks!

Jian Huang†
jian.huang@gatech.edu

Anirudh Badam Laura Caulfield Suman Nath
Sudipta Sengupta Bikash Sharma Moinuddin K. Qureshi †

† Georgia Tech Microsoft

Q&A