



RackBlox: A Software-Defined Rack-Scale Storage System with Network-Storage Co-Design

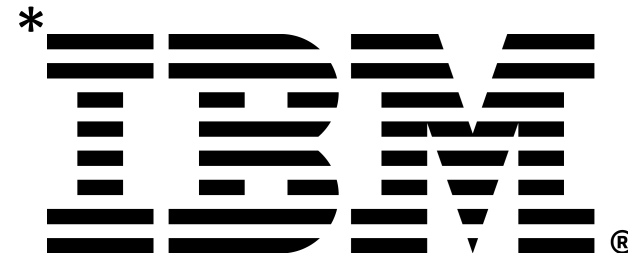
Benjamin Reidys

Yuqi Xue Daixuan Li Bharat Sukhwani*

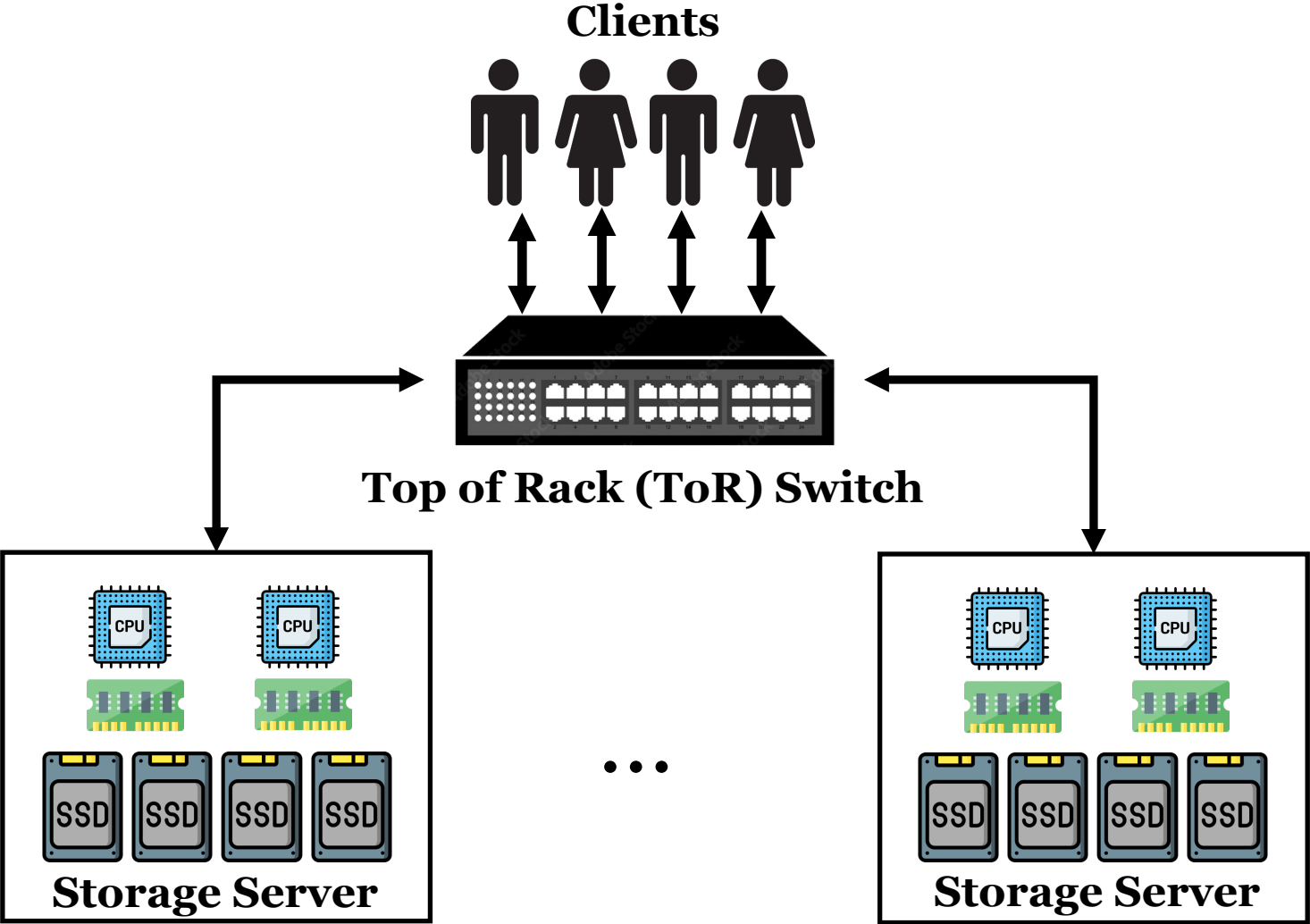
Wen-mei Hwu Deming Chen Sameh Asaad* Jian Huang



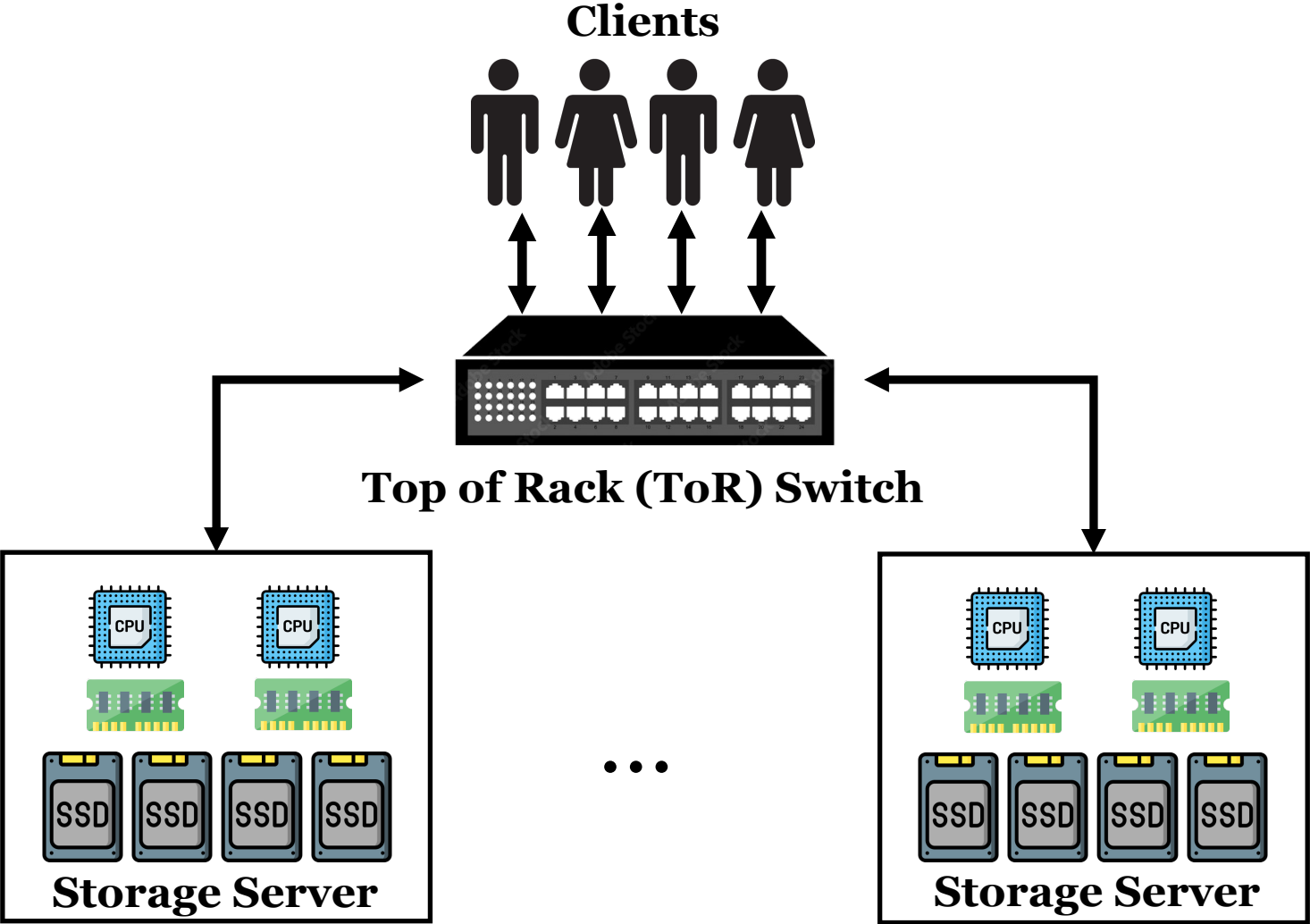
UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN



Software-Defined Infrastructure is Increasingly Popular

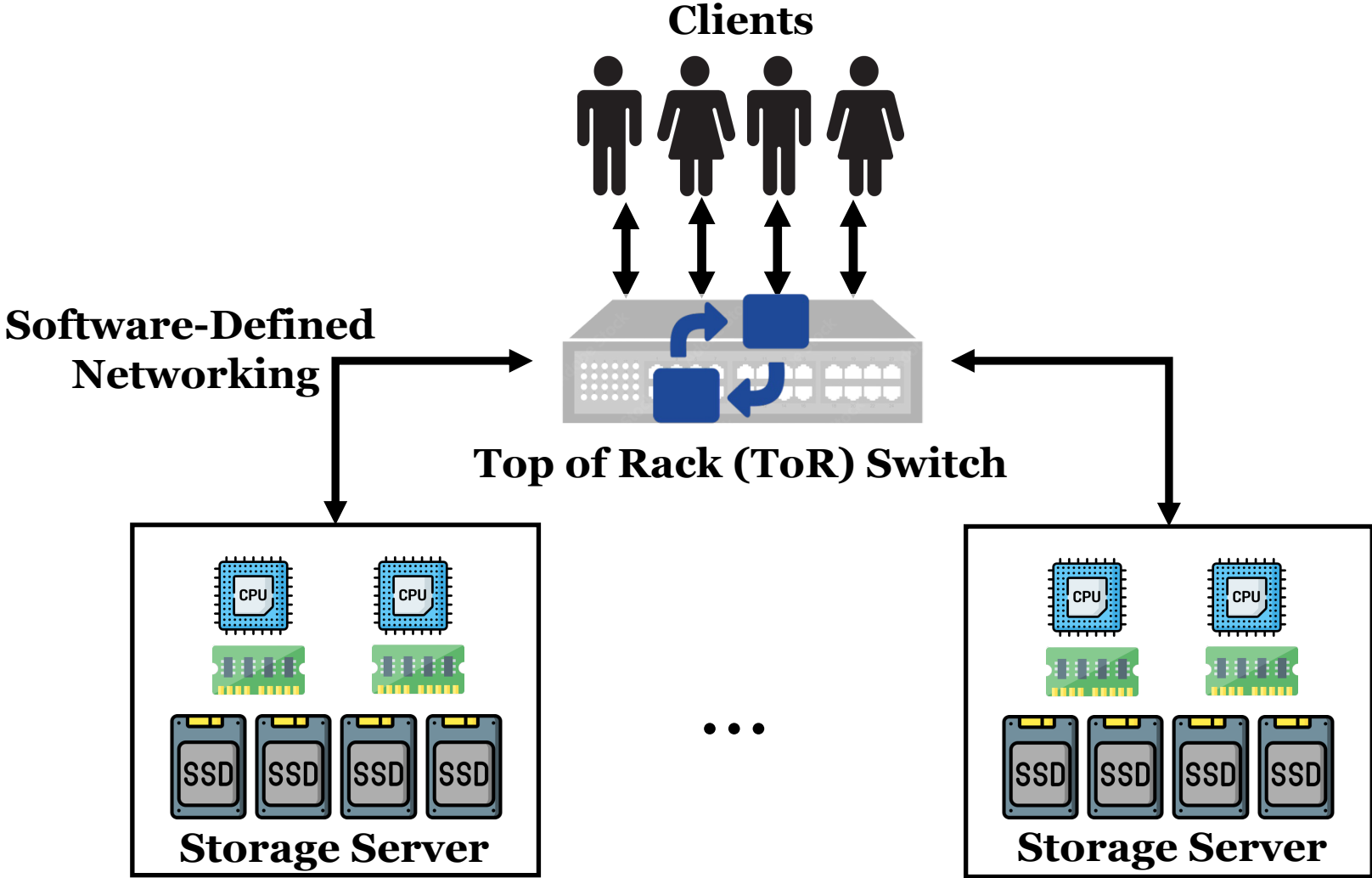


Software-Defined Infrastructure is Increasingly Popular



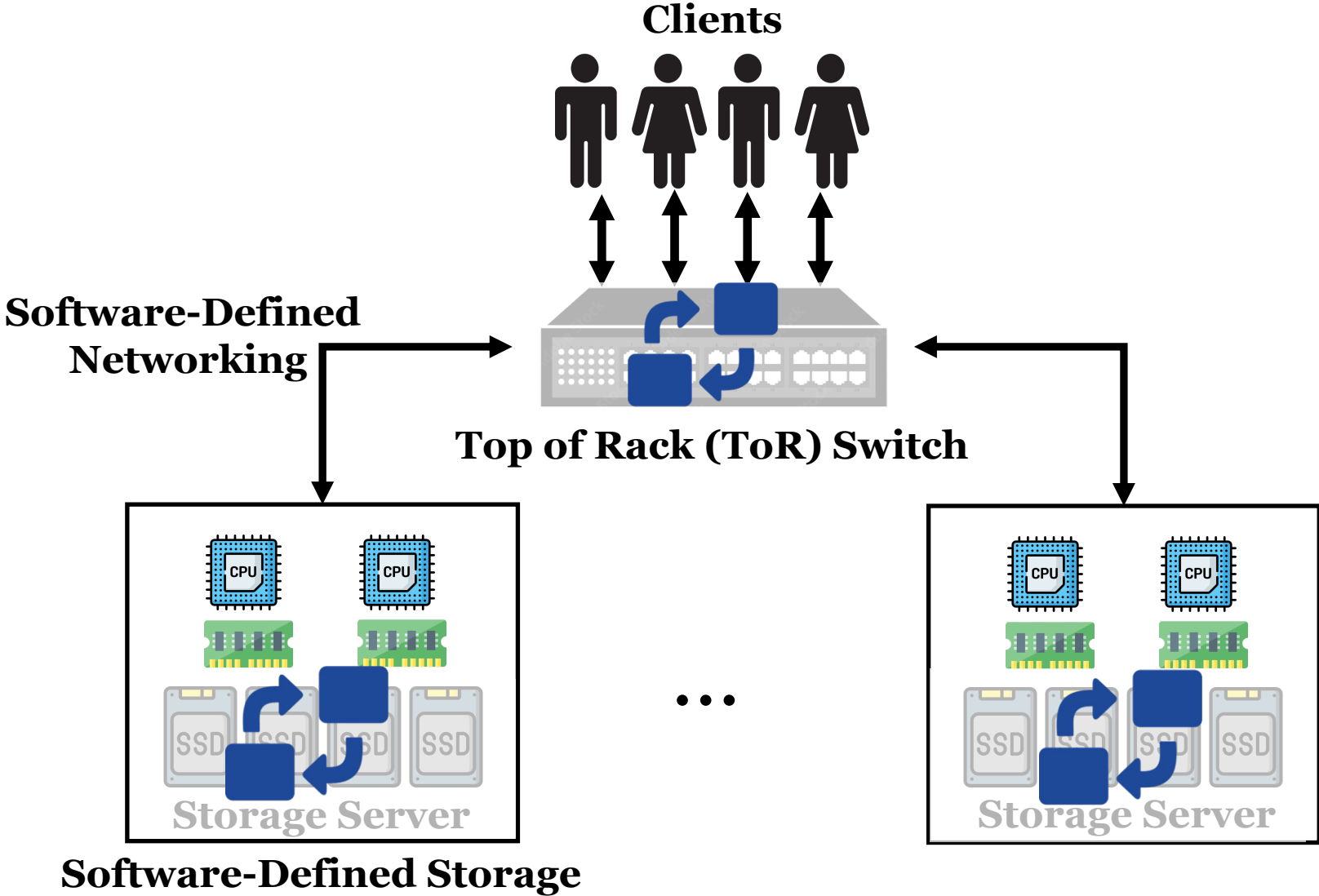
Redefine
management policies

Software-Defined Infrastructure is Increasingly Popular



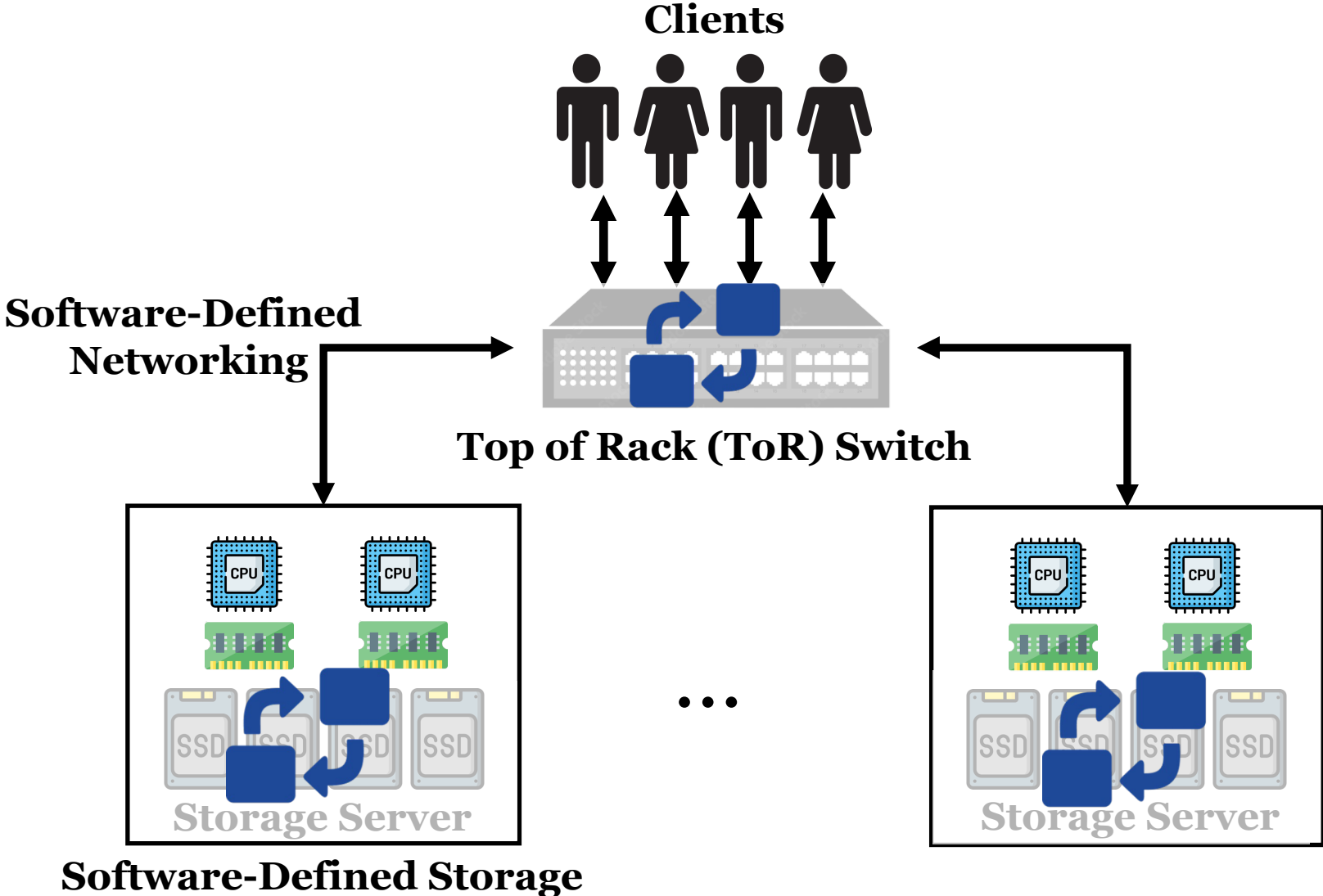
Redefine
management policies

Software-Defined Infrastructure is Increasingly Popular



Redefine
management policies

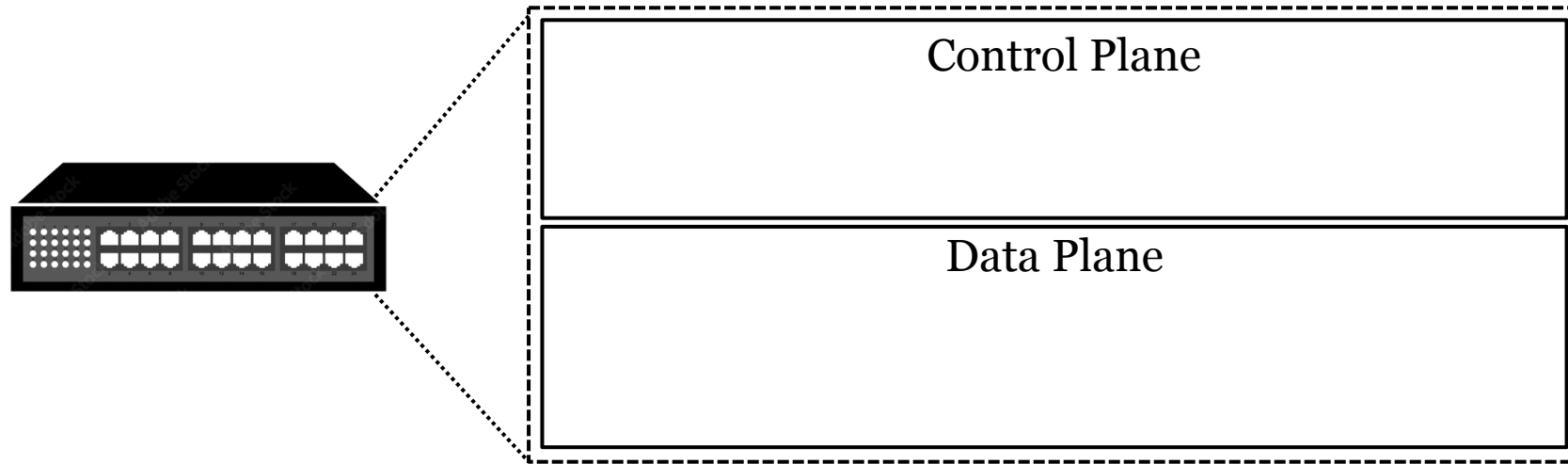
Software-Defined Infrastructure is Increasingly Popular



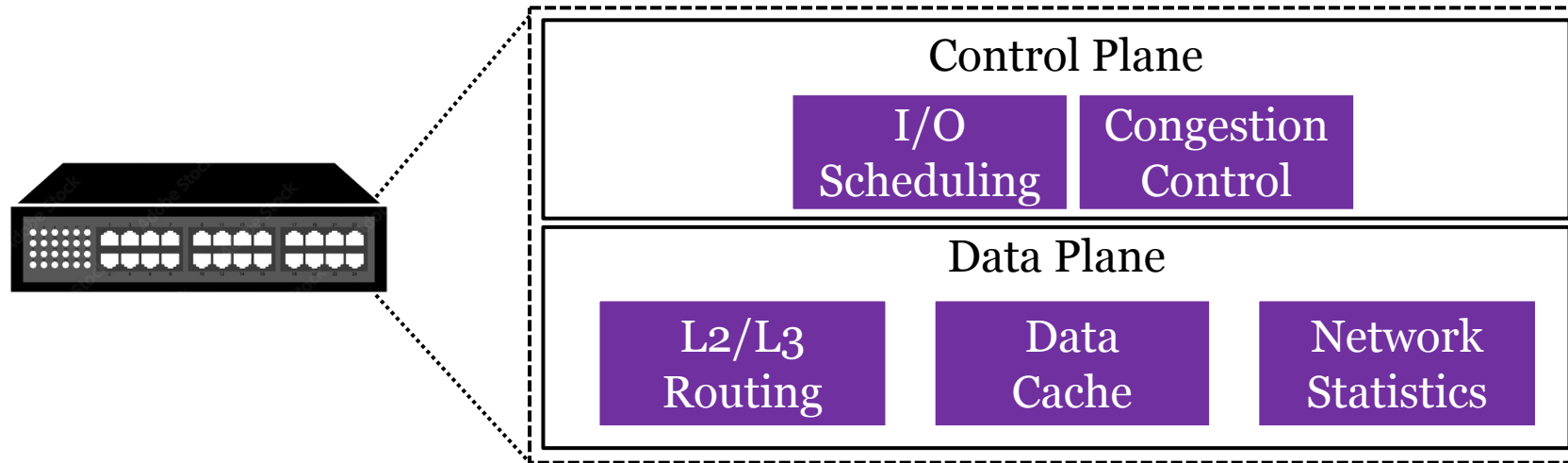
Redefine
management policies

Adopted in major data centers

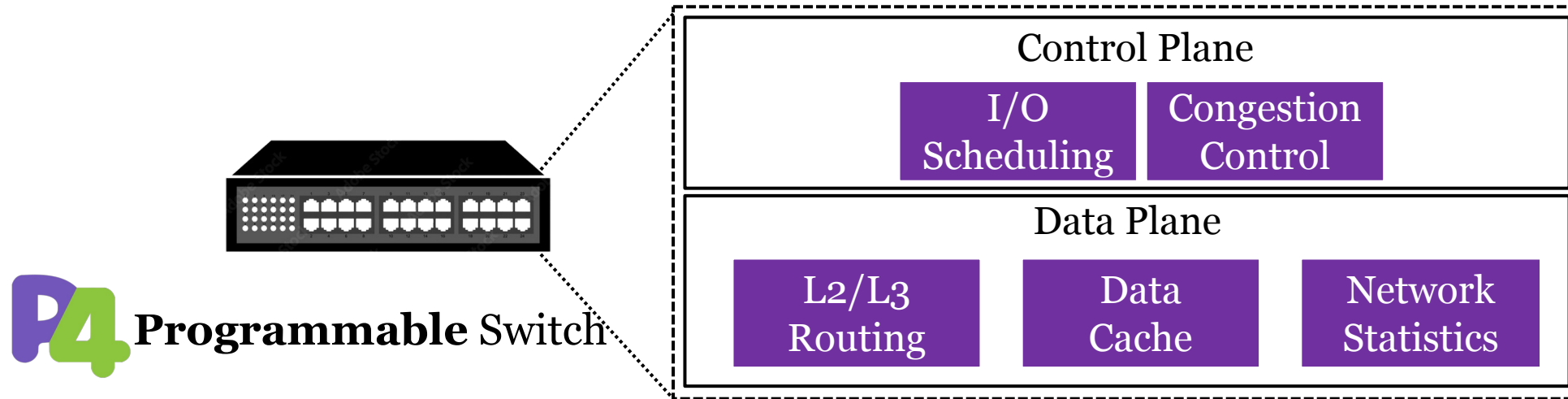
What is Software-Defined Networking?



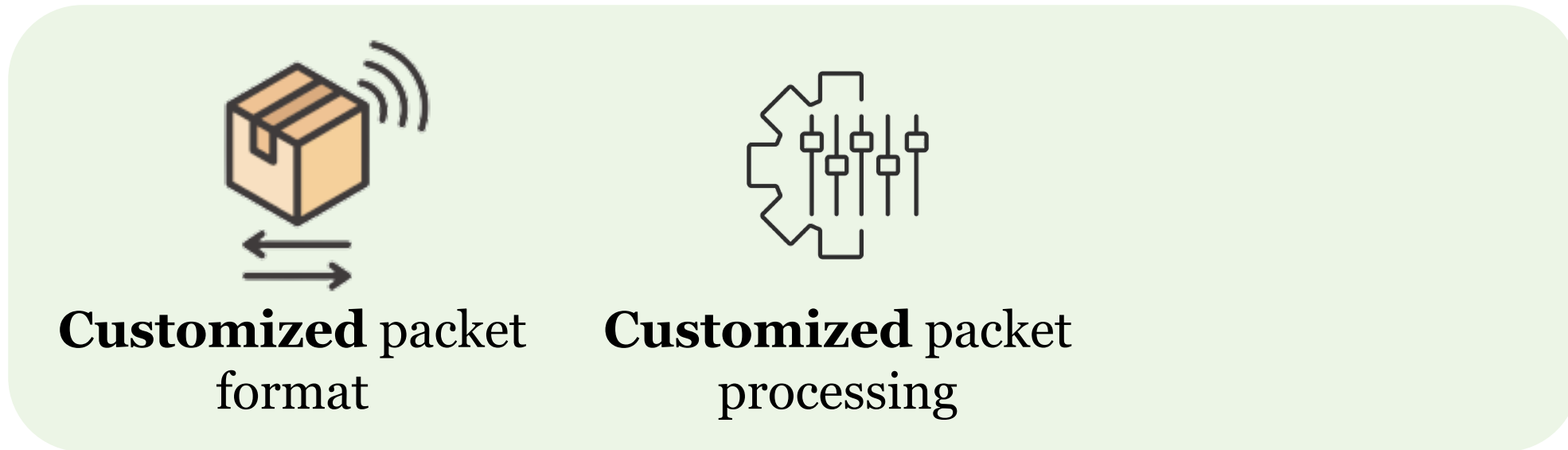
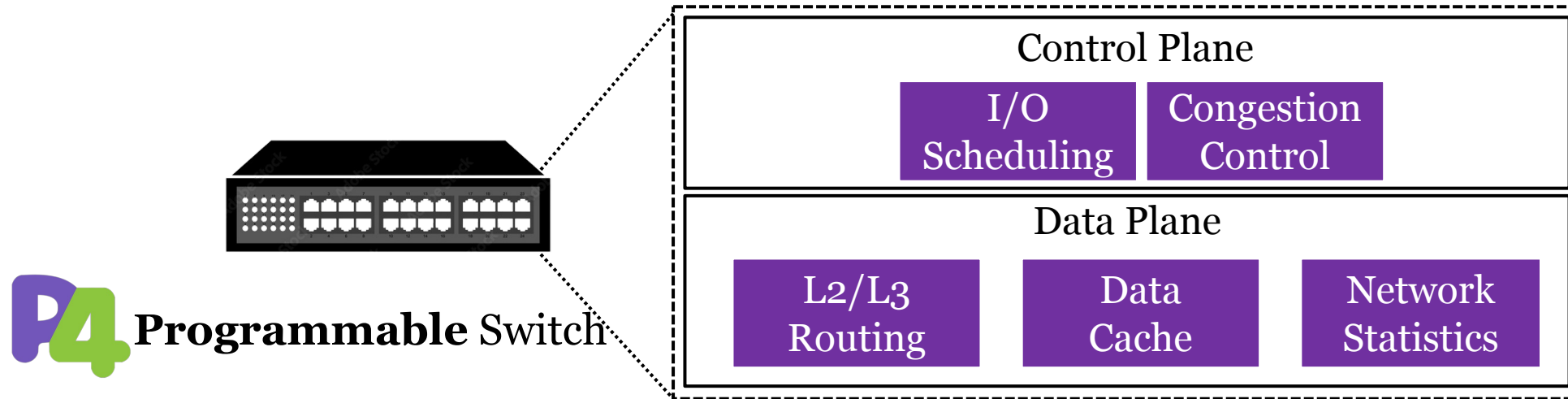
What is Software-Defined Networking?



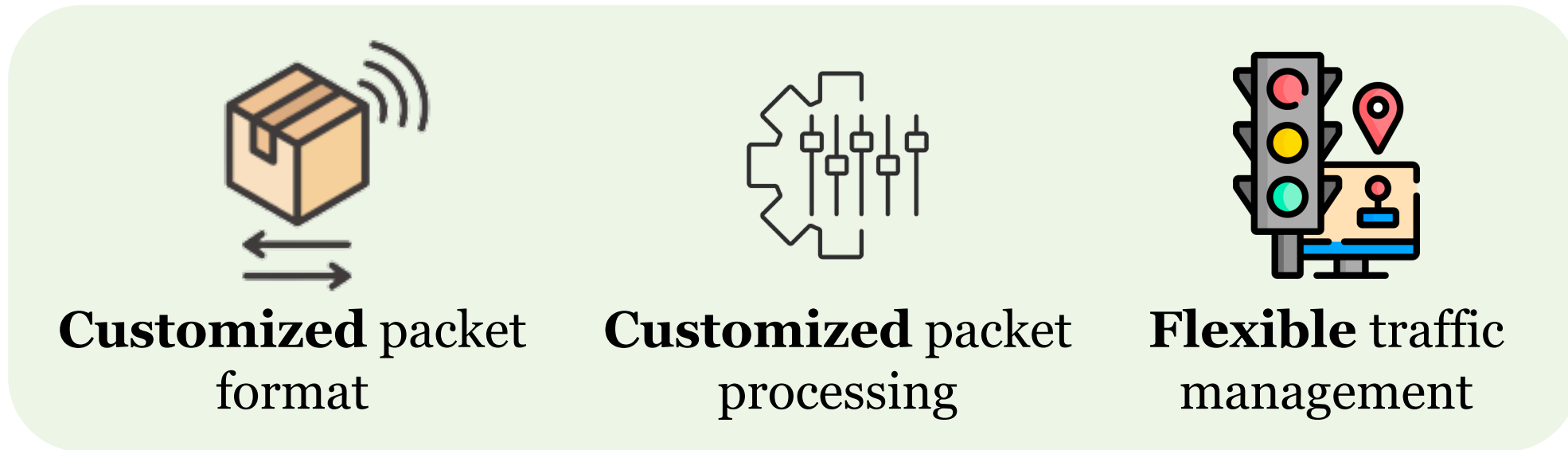
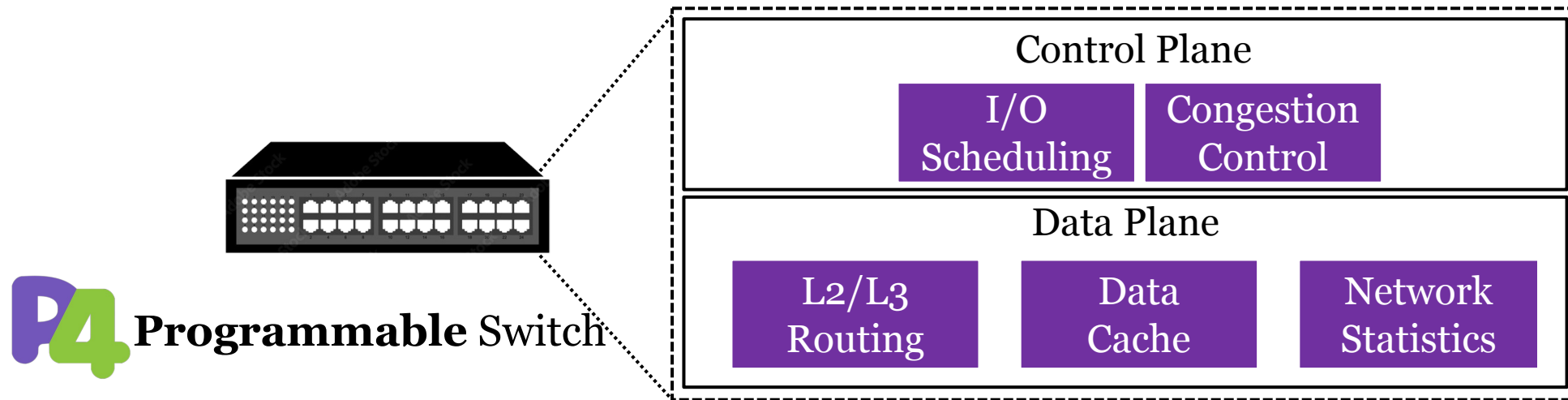
What is Software-Defined Networking?



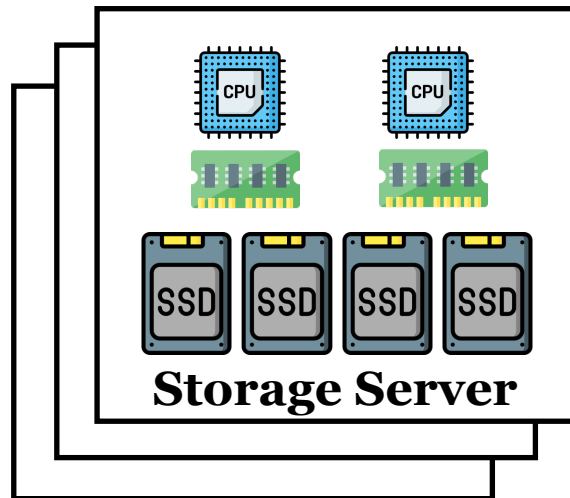
What is Software-Defined Networking?



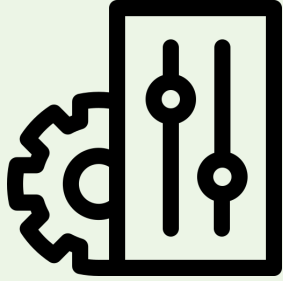
What is Software-Defined Networking?



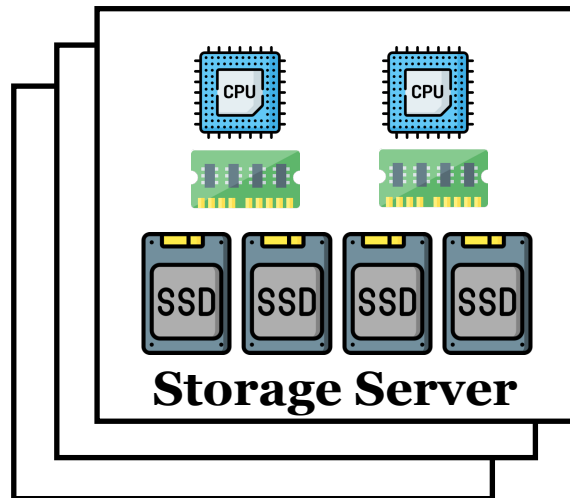
What is Software-Defined Storage?



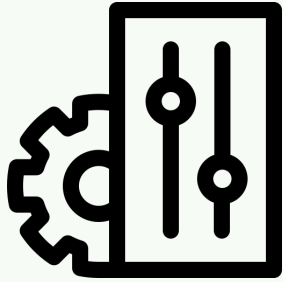
What is Software-Defined Storage?



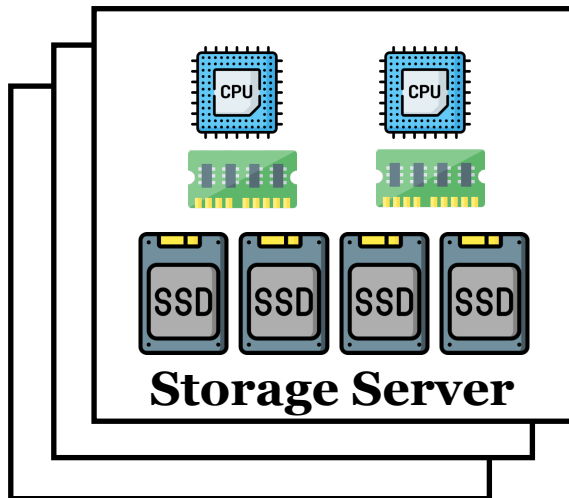
Allow software to **manage** storage chips



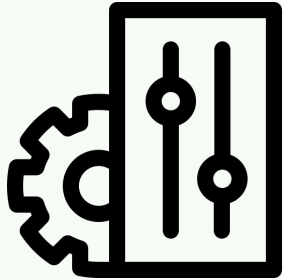
Background on Solid-State Drives



Allow software to **manage** storage chips



Background on Solid-State Drives



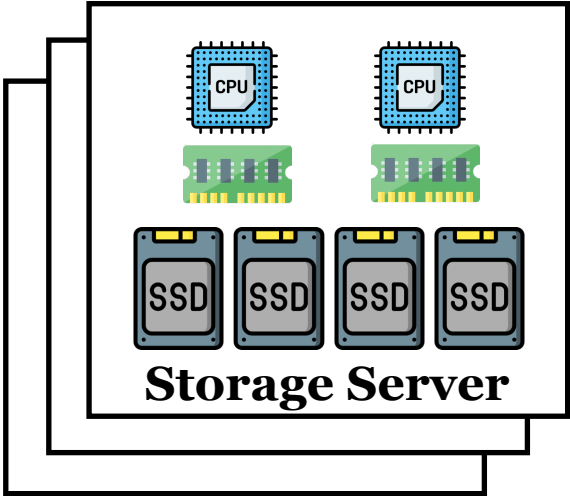
Allow software to **manage** storage chips



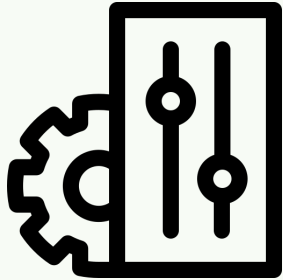
Increasing performance



Decreasing Cost



Background on Solid-State Drives



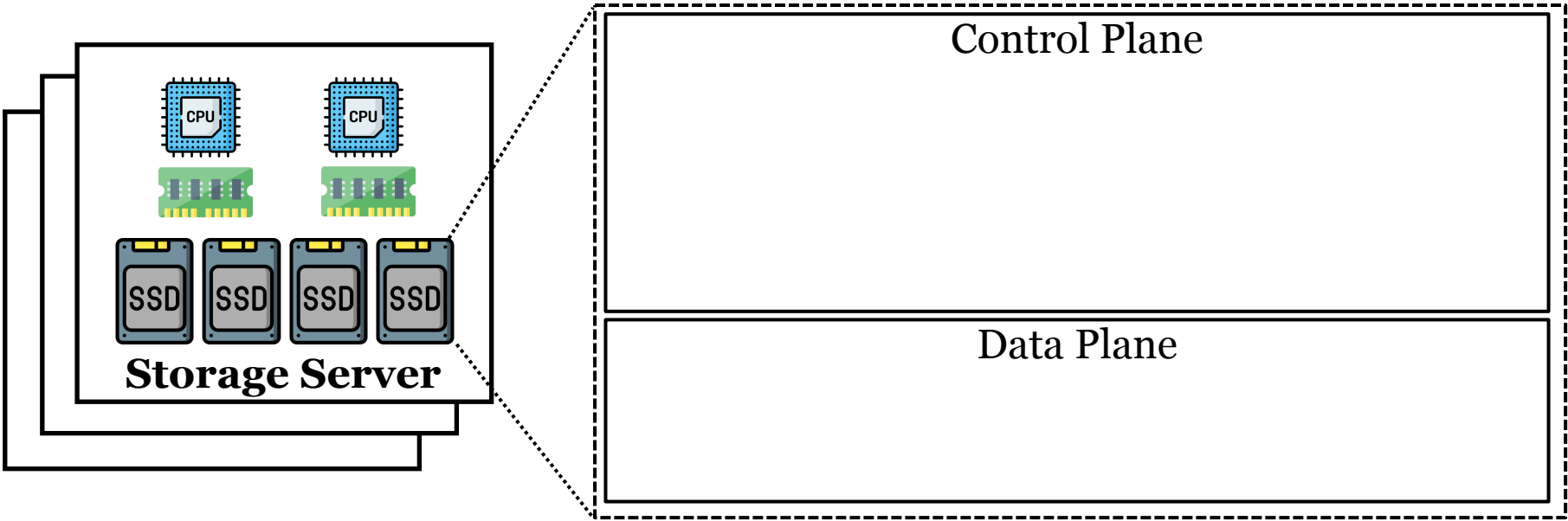
Allow software to **manage** storage chips



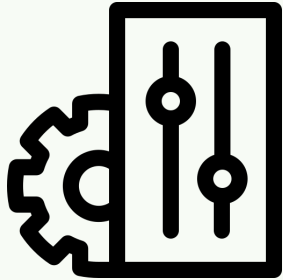
Increasing performance



Decreasing Cost



Background on Solid-State Drives



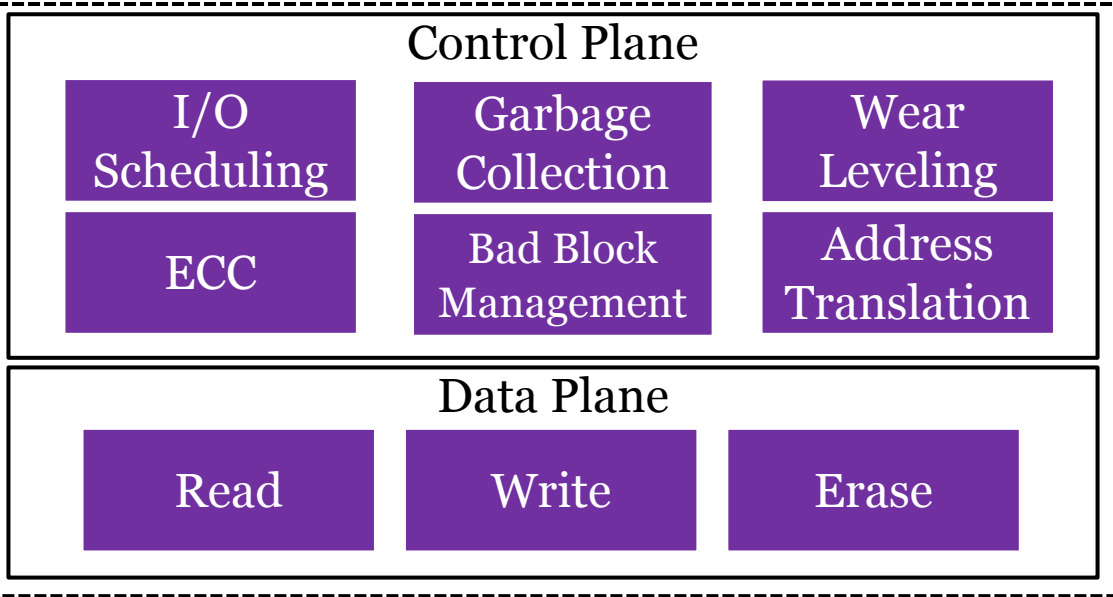
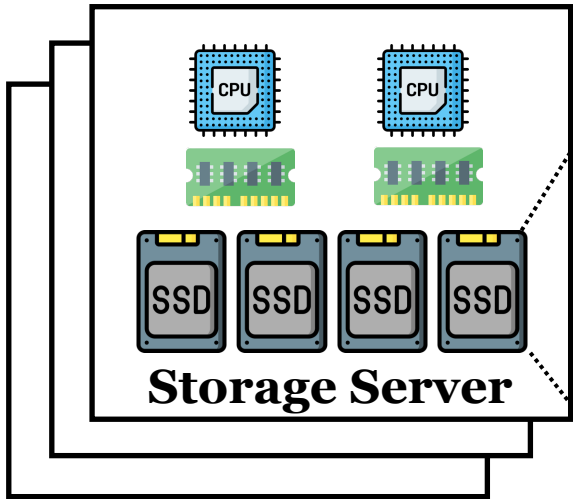
Allow software to **manage** storage chips



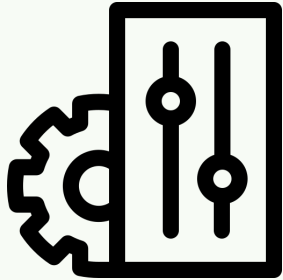
Increasing performance



Decreasing Cost



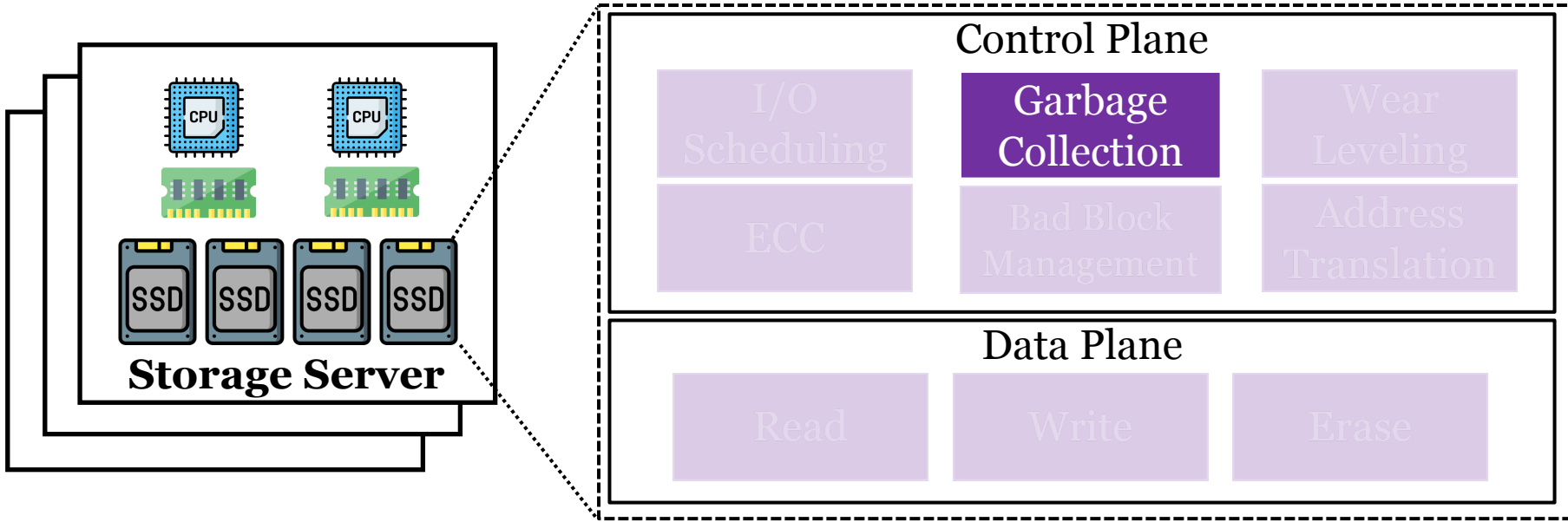
Background on Solid-State Drives



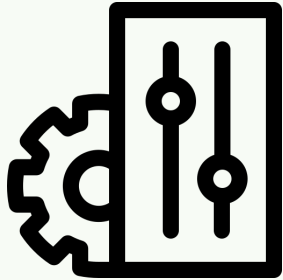
Allow software to **manage** storage chips



Garbage collection for out-of-place updates



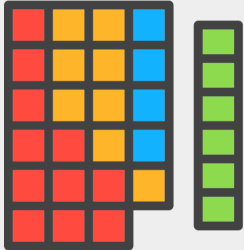
Background on Solid-State Drives



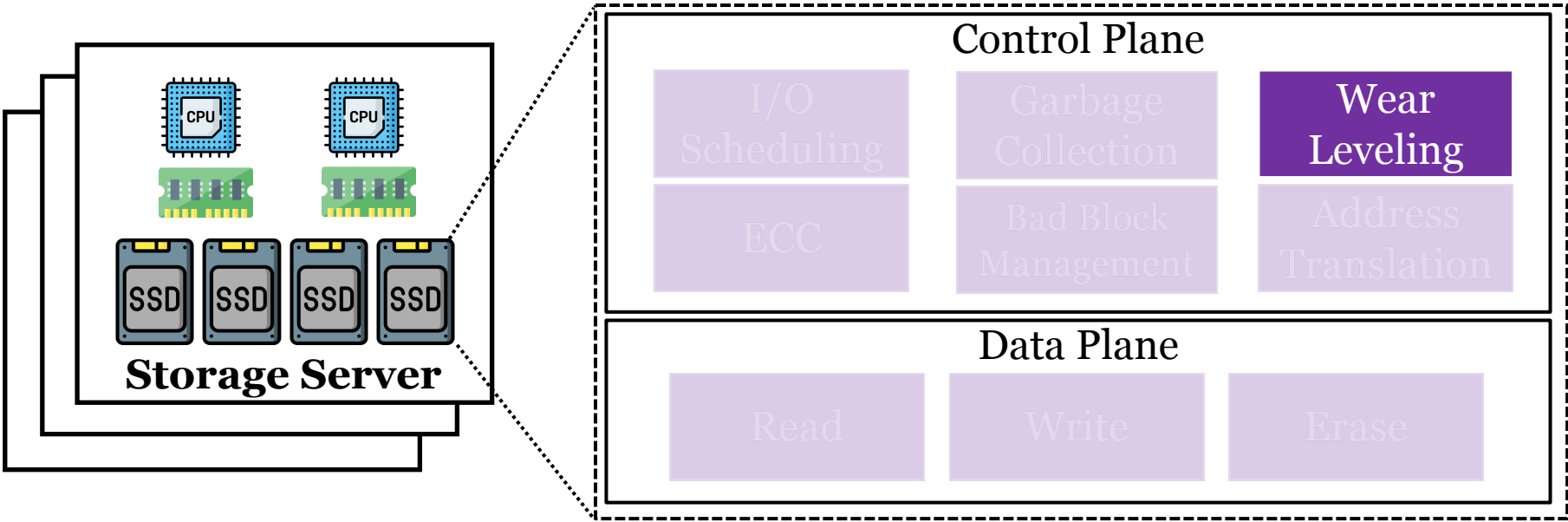
Allow software to **manage** storage chips



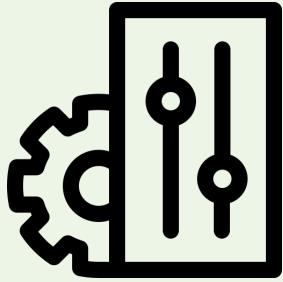
Garbage collection for out-of-place updates



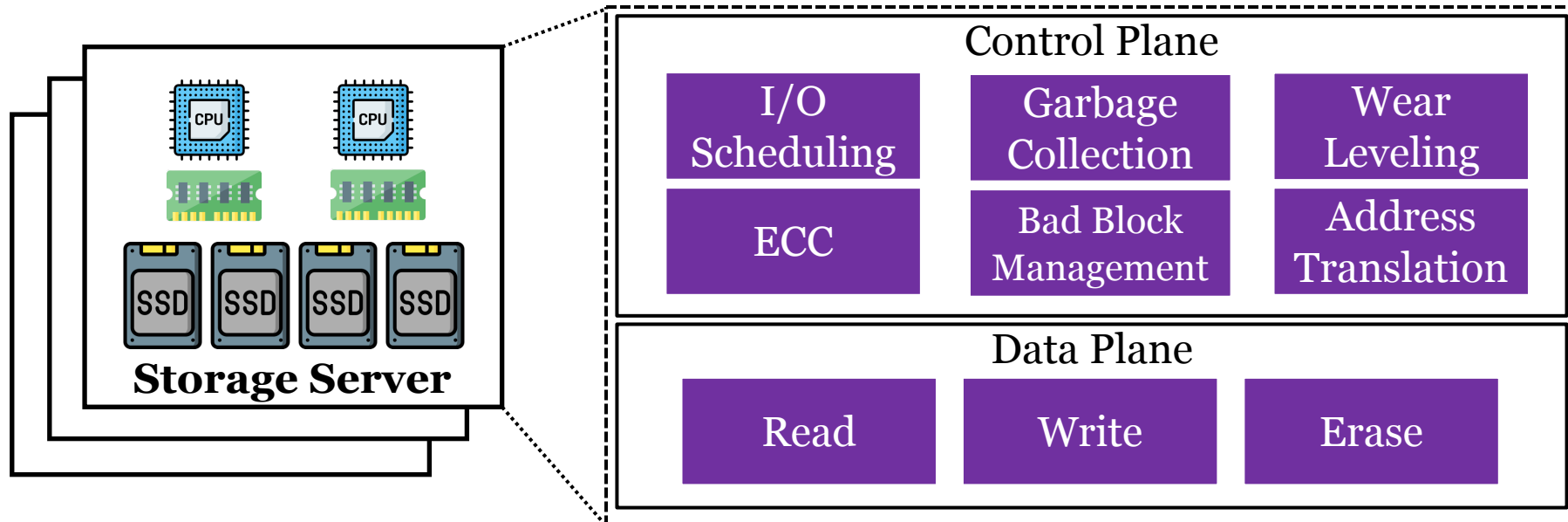
Wear leveling distributes writes across flash blocks



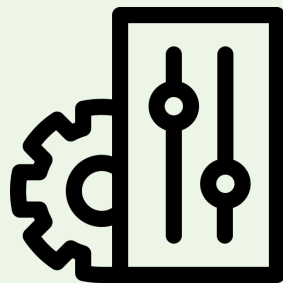
What is Software-Defined Flash?



Allow software to **manage** storage chips



What is Software-Defined Flash?



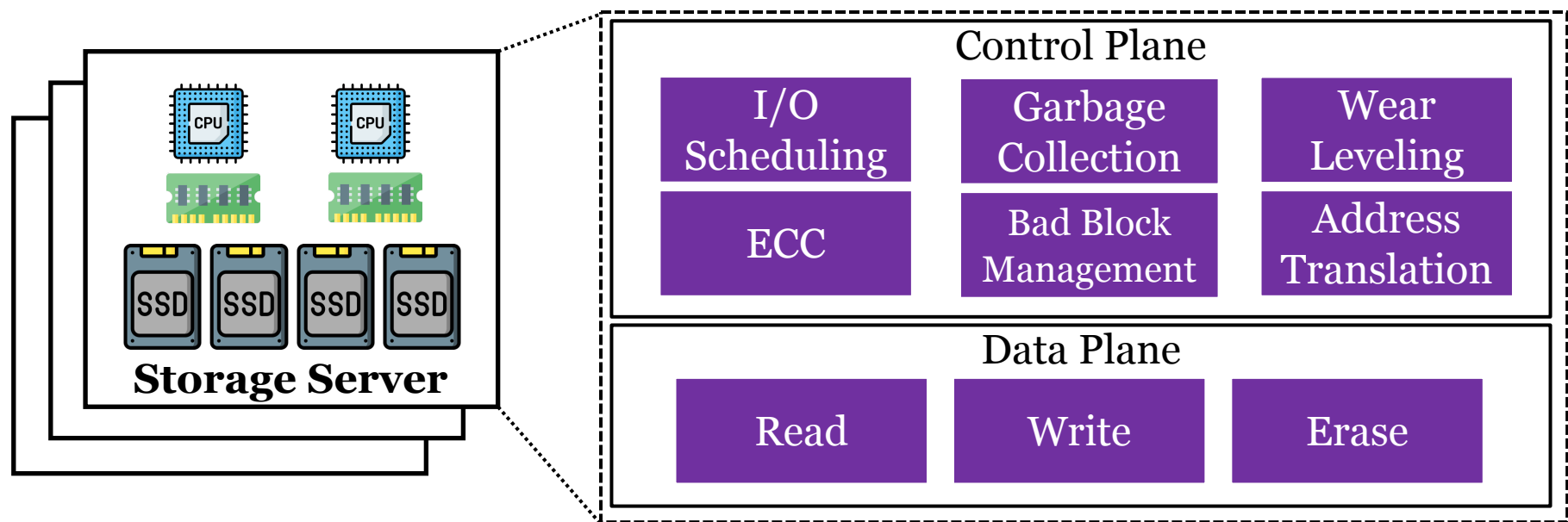
Allow software to **manage** storage chips



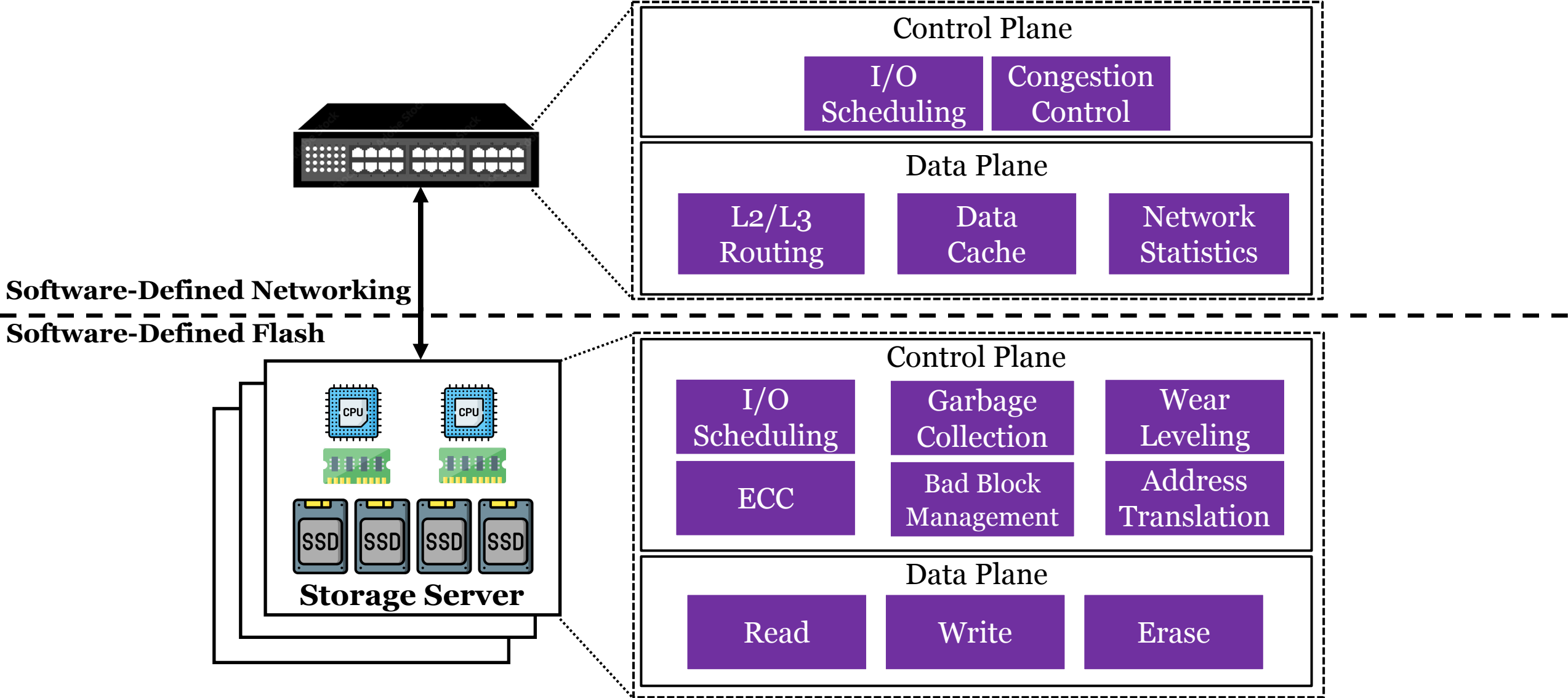
Improved resource utilization



Predictable storage performance

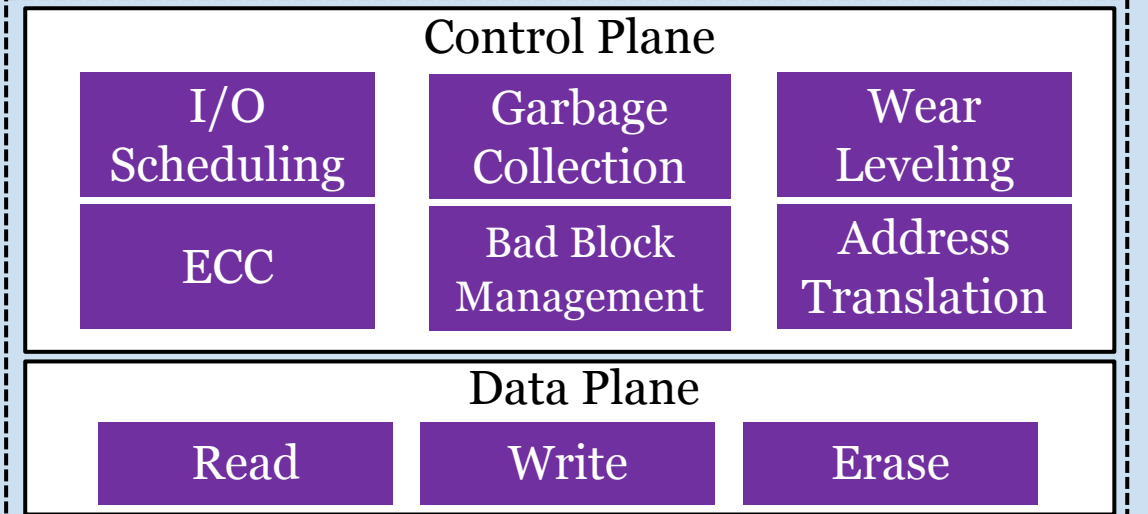
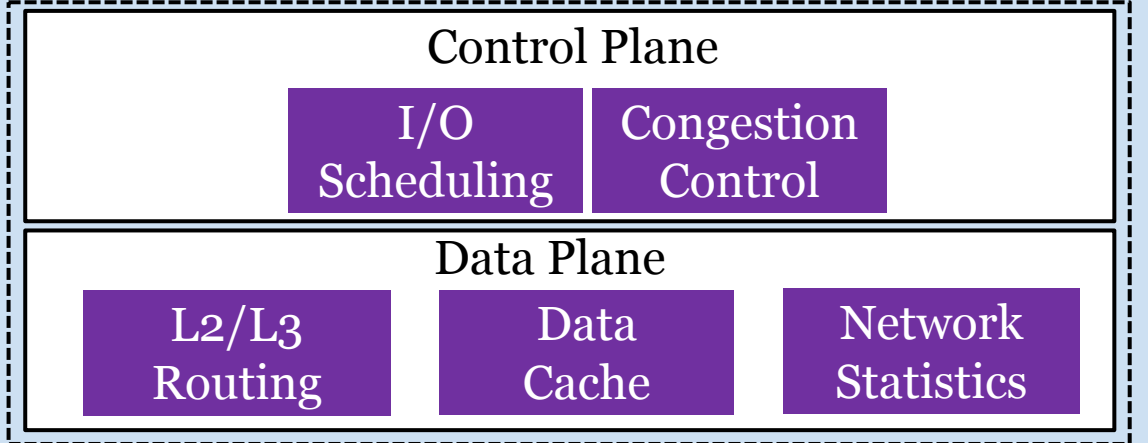


SDN/SDF are Managed Separately Today



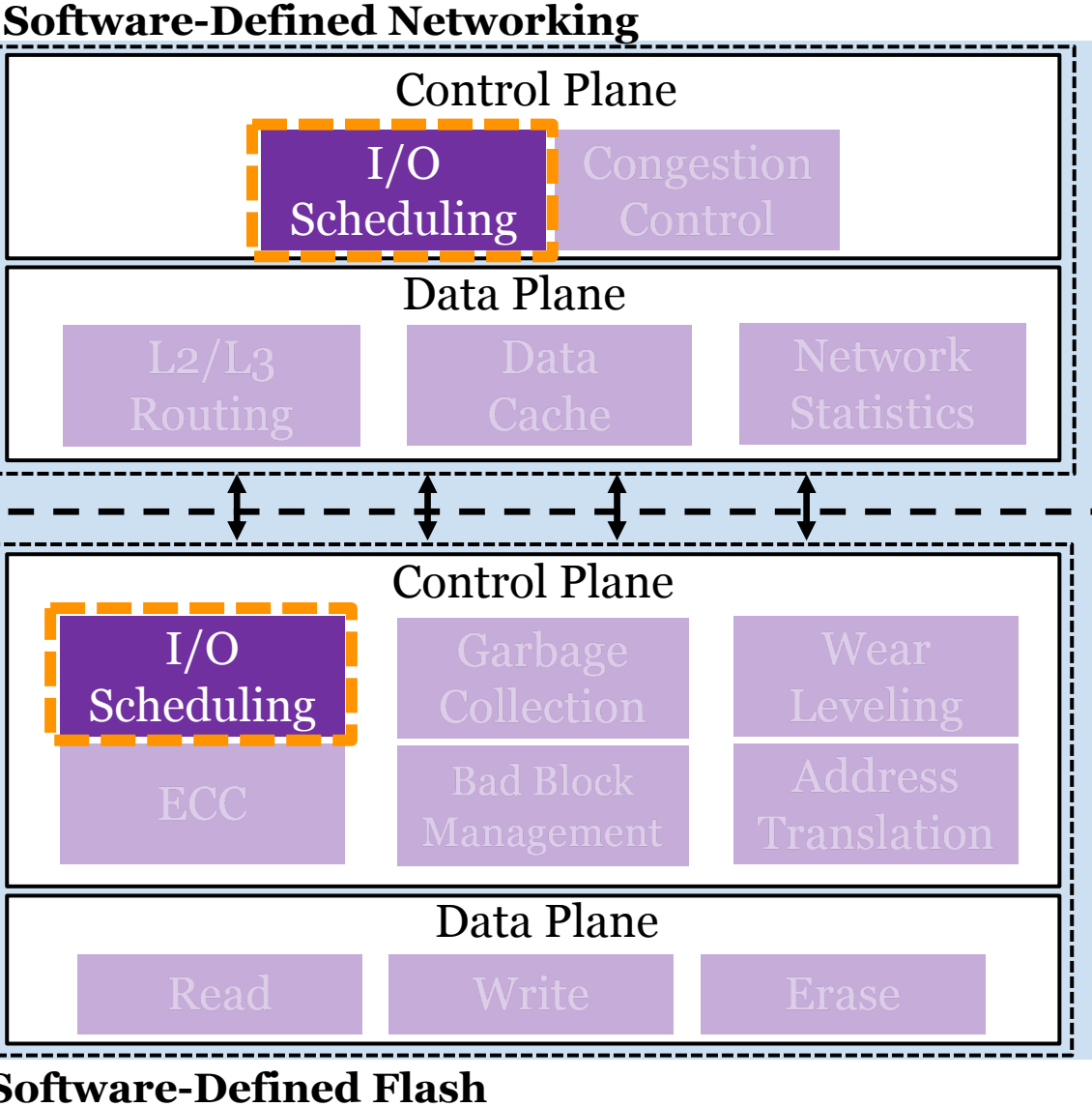
Research Problem: Lack of Coordination between SDN/SDF

Software-Defined Networking

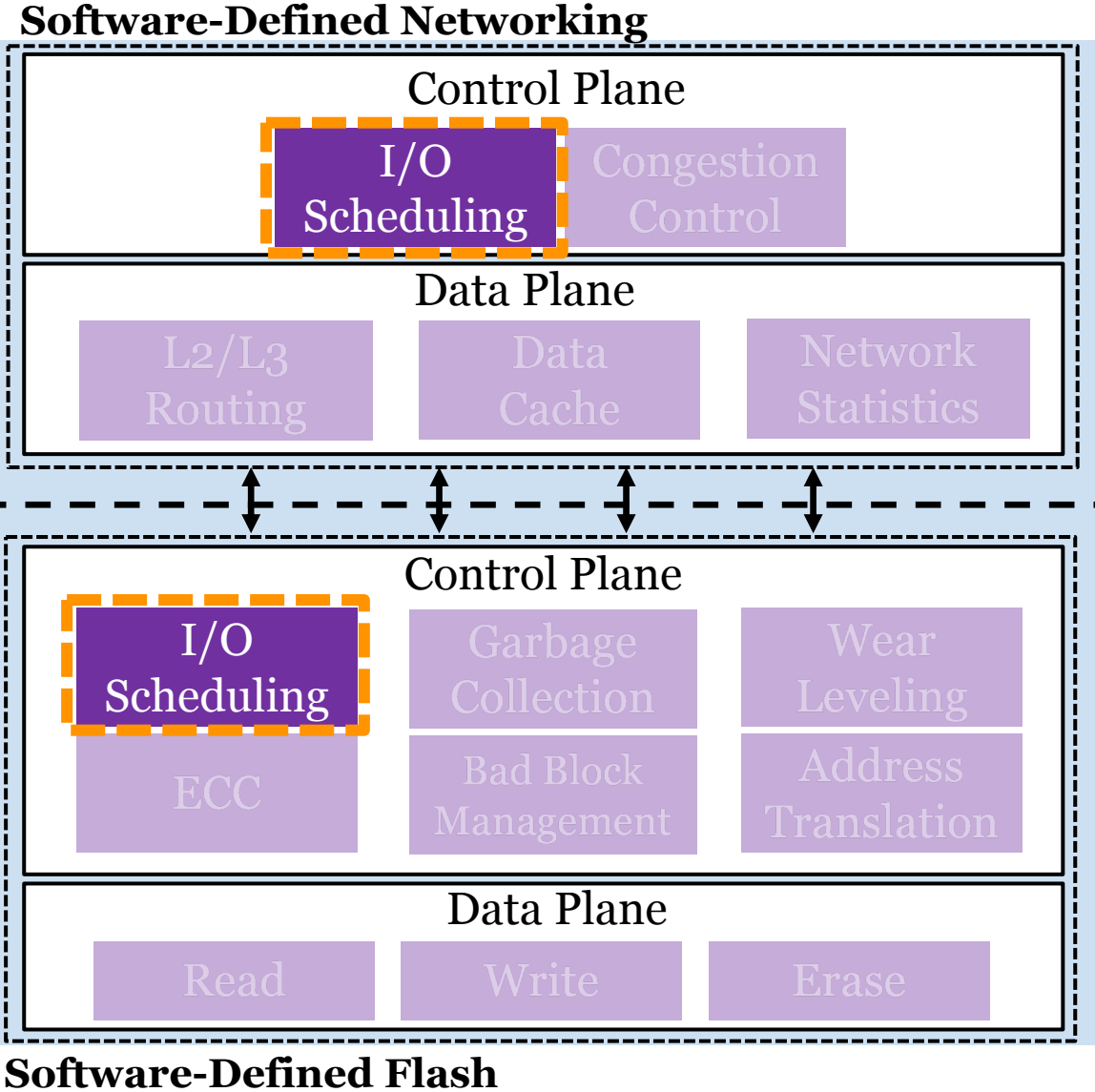


Software-Defined Flash

Research Problem: Lack of Coordination between SDN/SDF



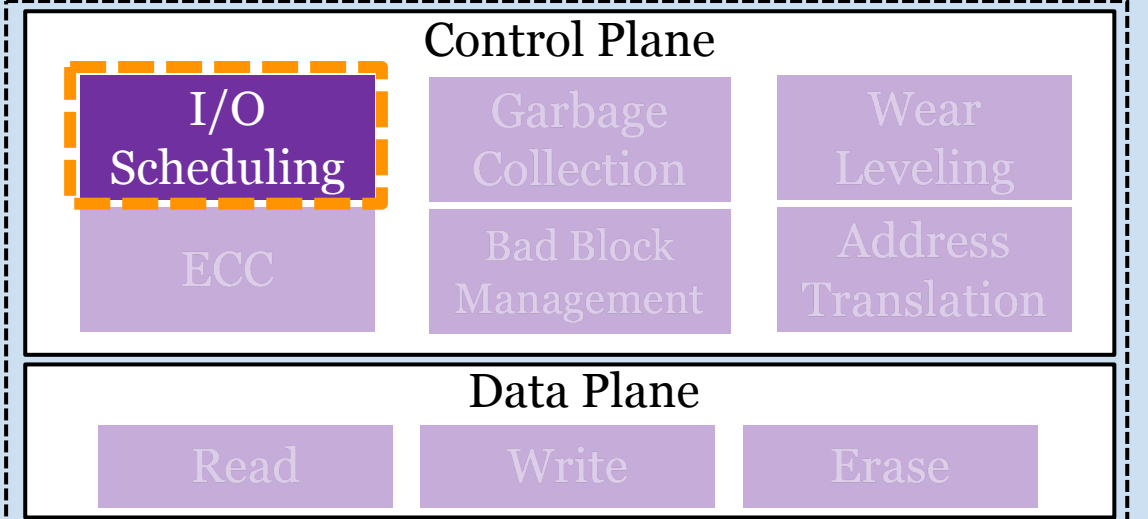
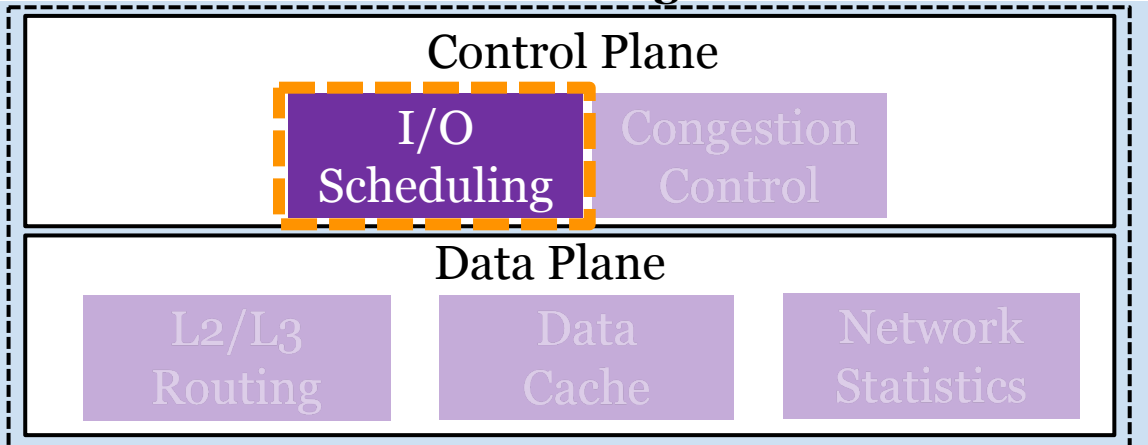
Research Problem: Lack of Coordination between SDN/SDF



Conflicting policies!

Research Problem: Lack of Coordination between SDN/SDF

Software-Defined Networking



Software-Defined Flash



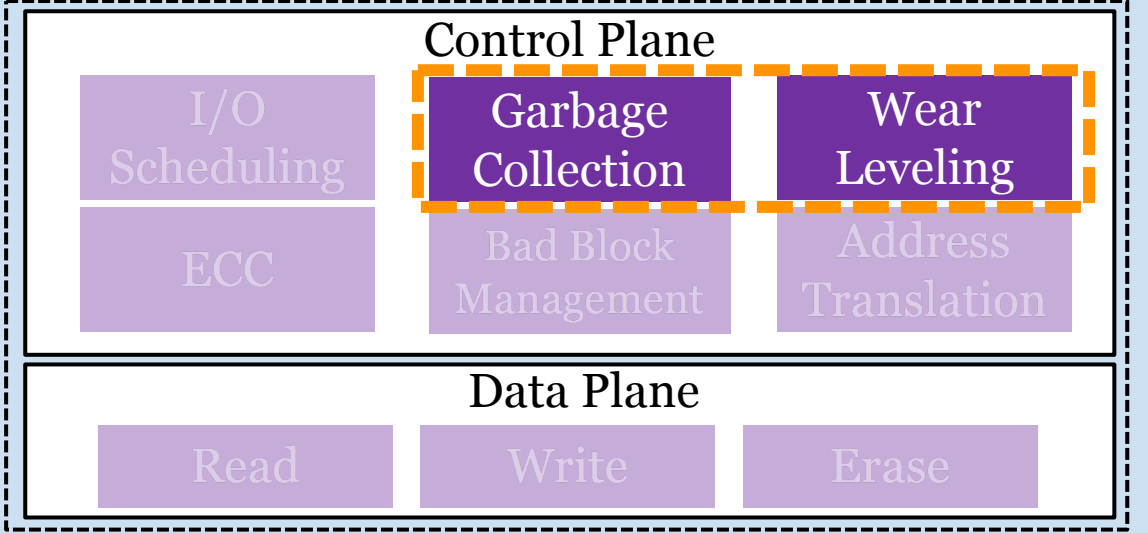
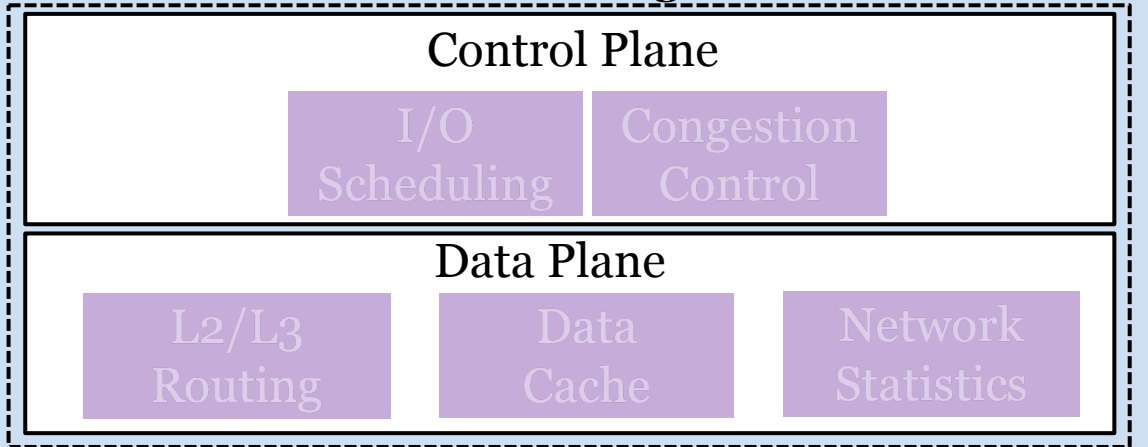
Conflicting policies!



Redundant effort!

Research Problem: Lack of Coordination between SDN/SDF

Software-Defined Networking



Software-Defined Flash



Conflicting policies!

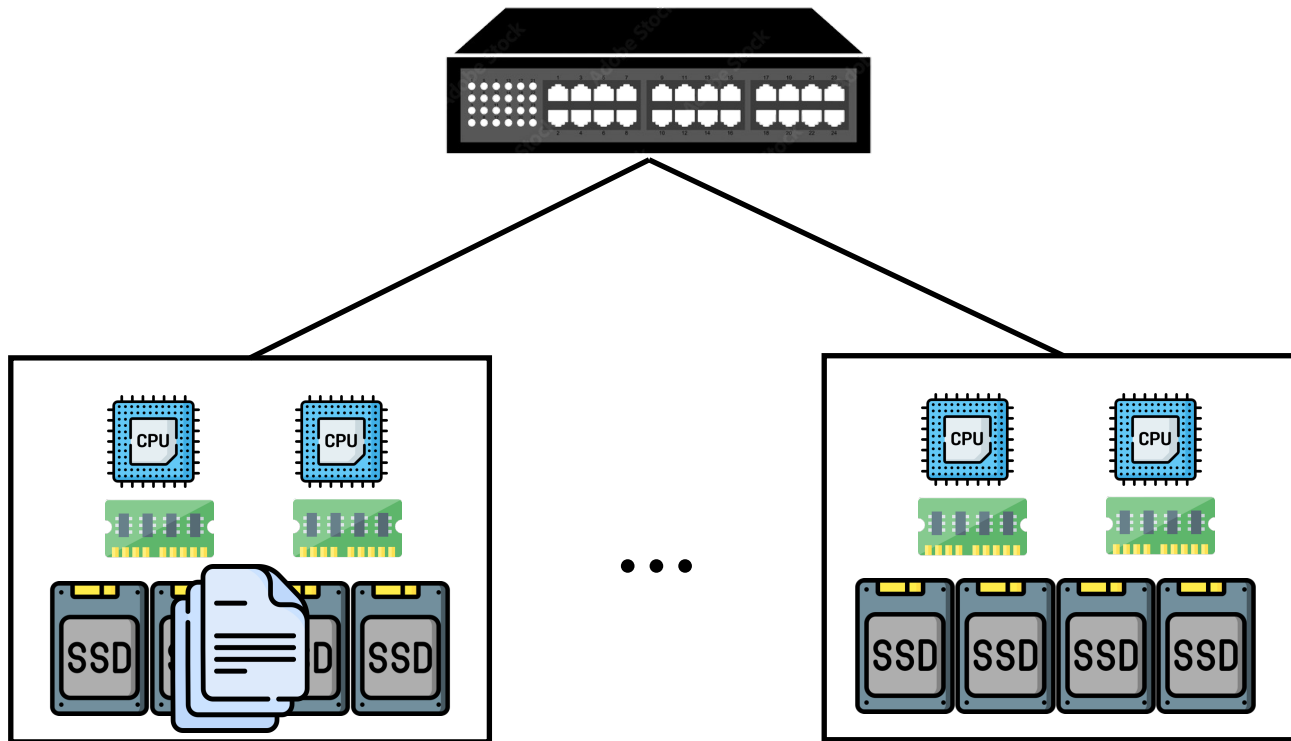


Redundant effort!

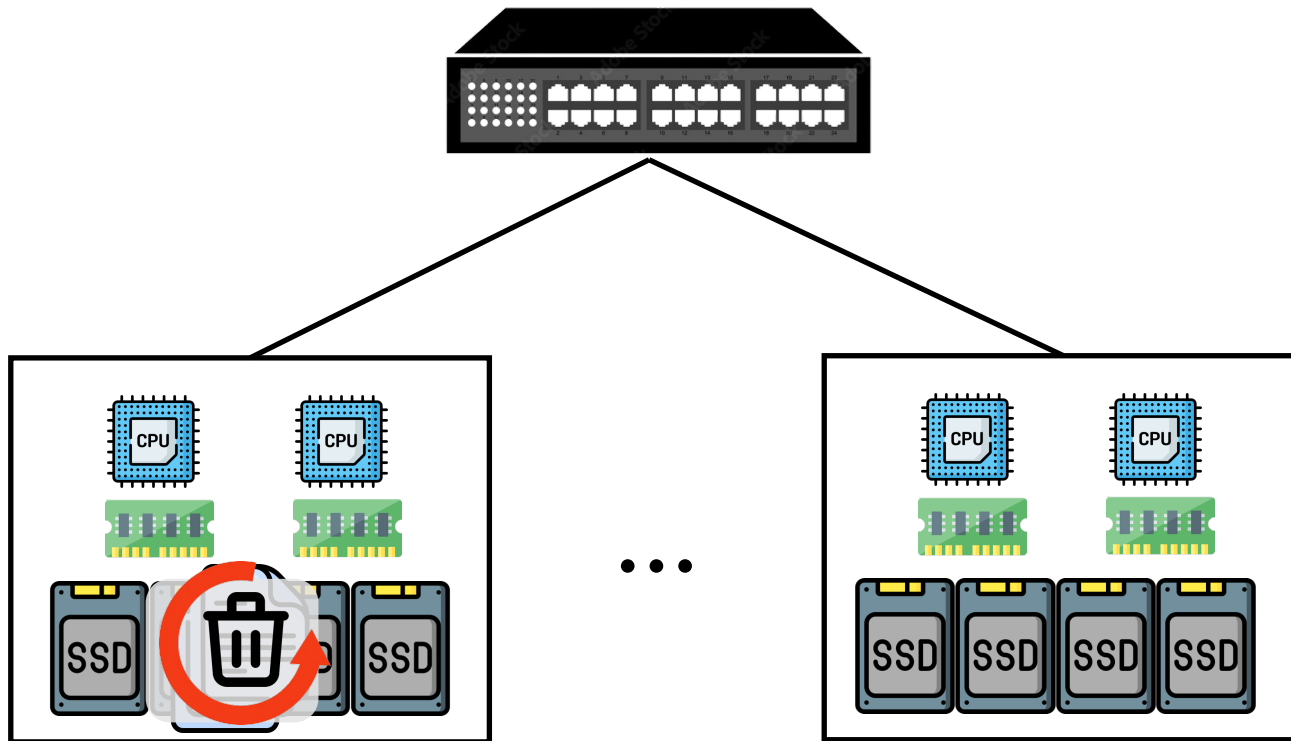


Missed opportunities for rack-scale optimization!

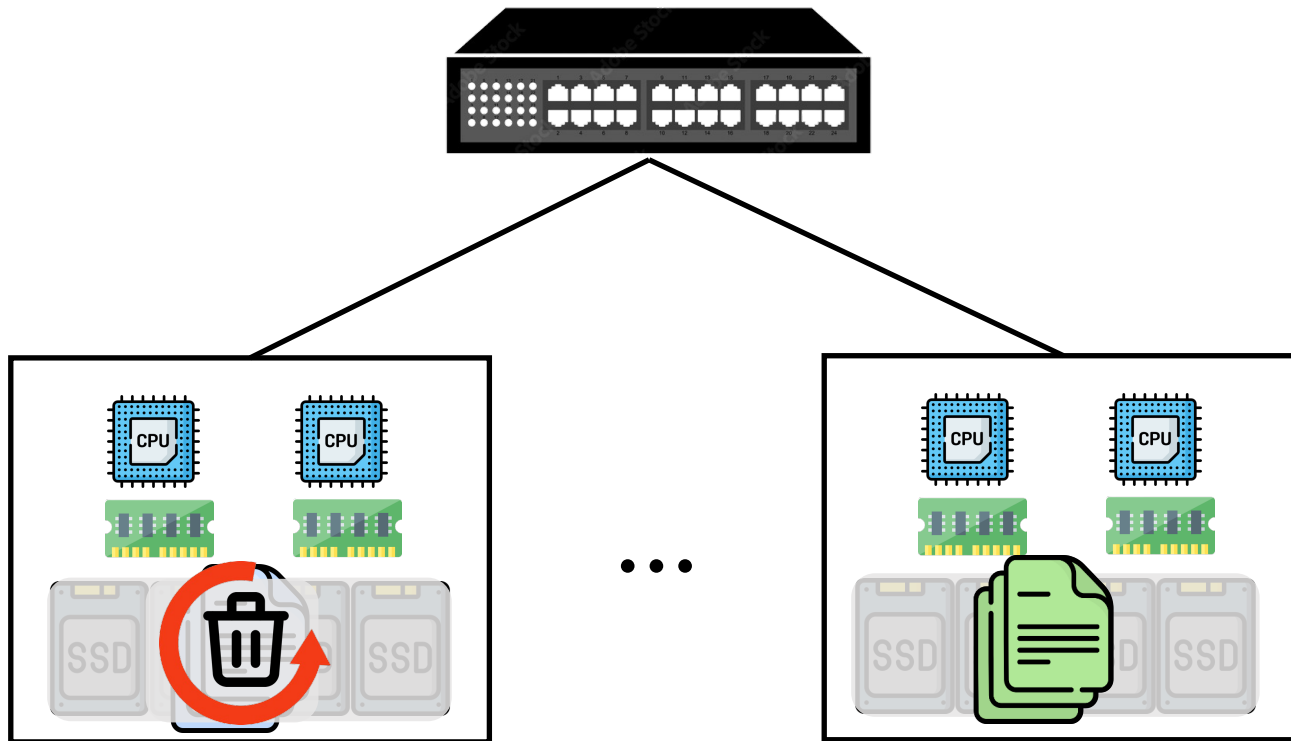
Research Problem: Lack of Coordination between SDN/SDF



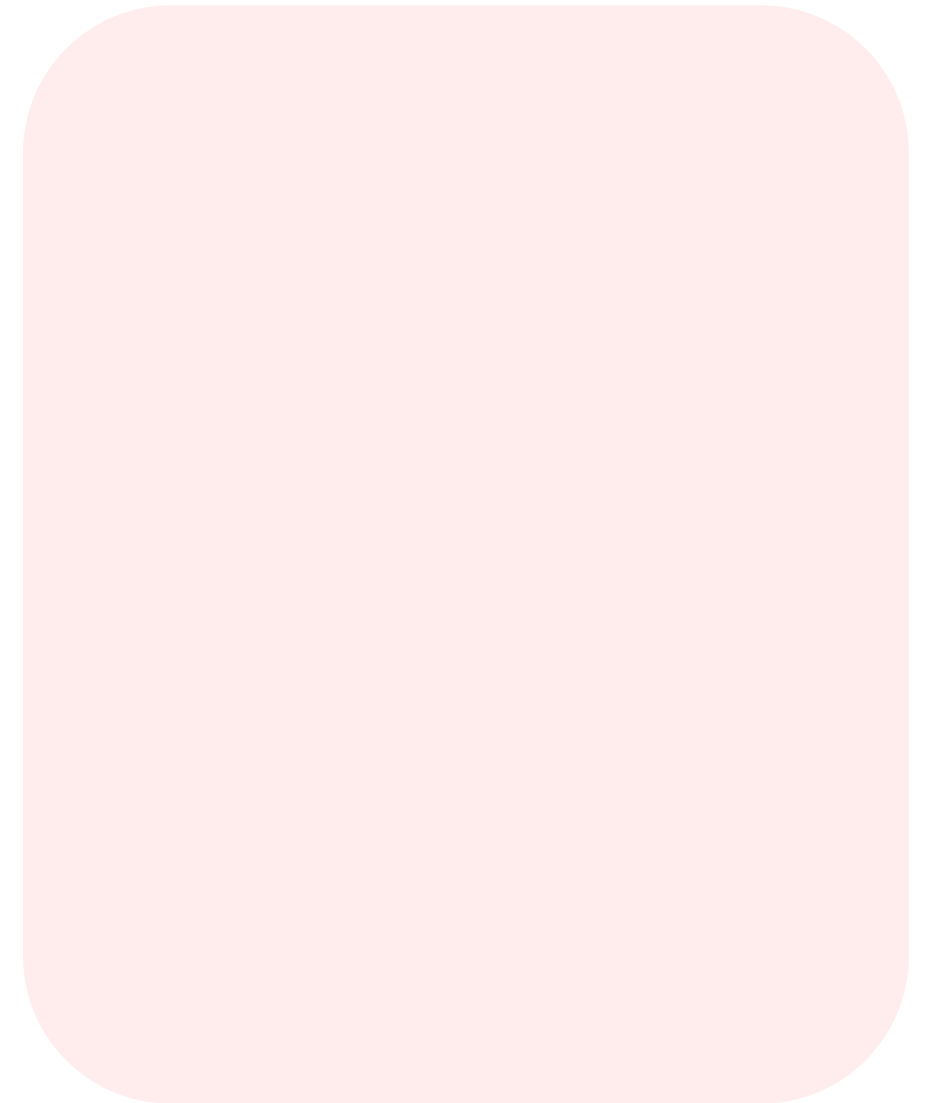
Research Problem: Lack of Coordination between SDN/SDF



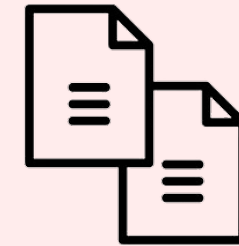
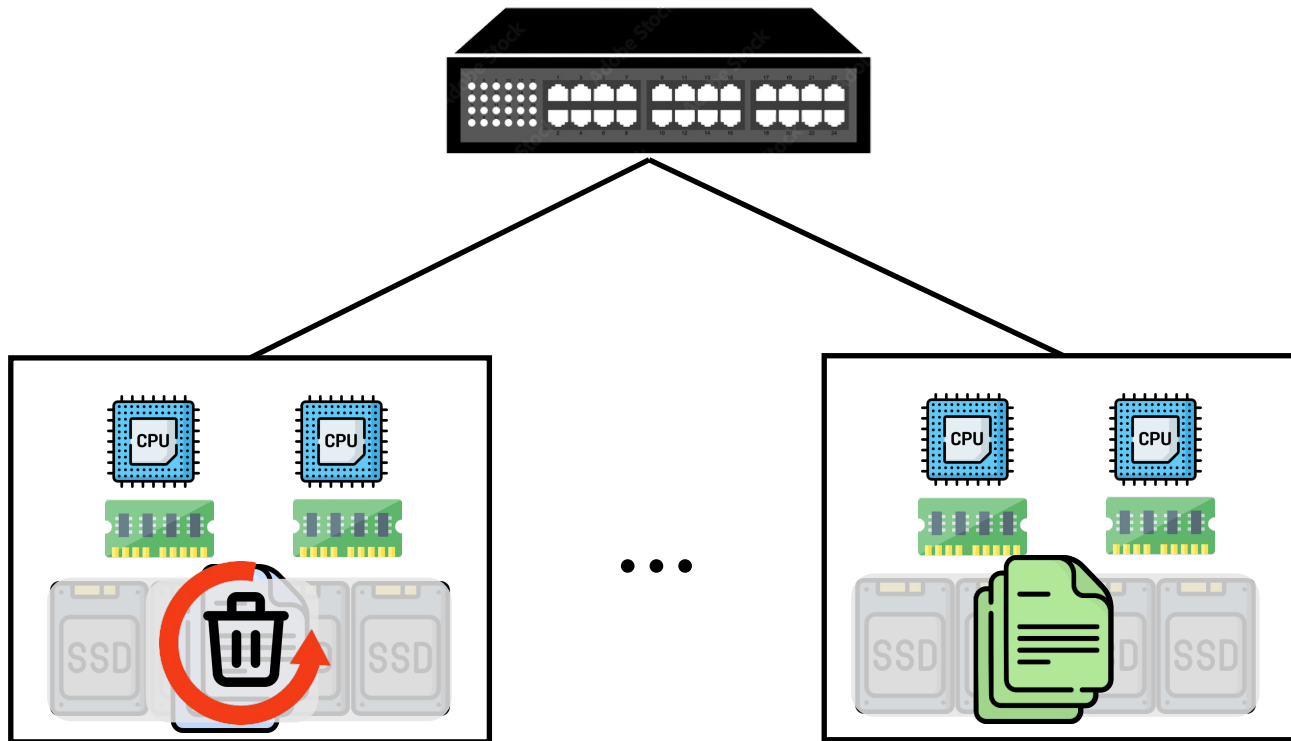
Research Problem: Lack of Coordination between SDN/SDF



Data is **replicated** by default!

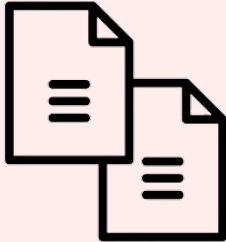
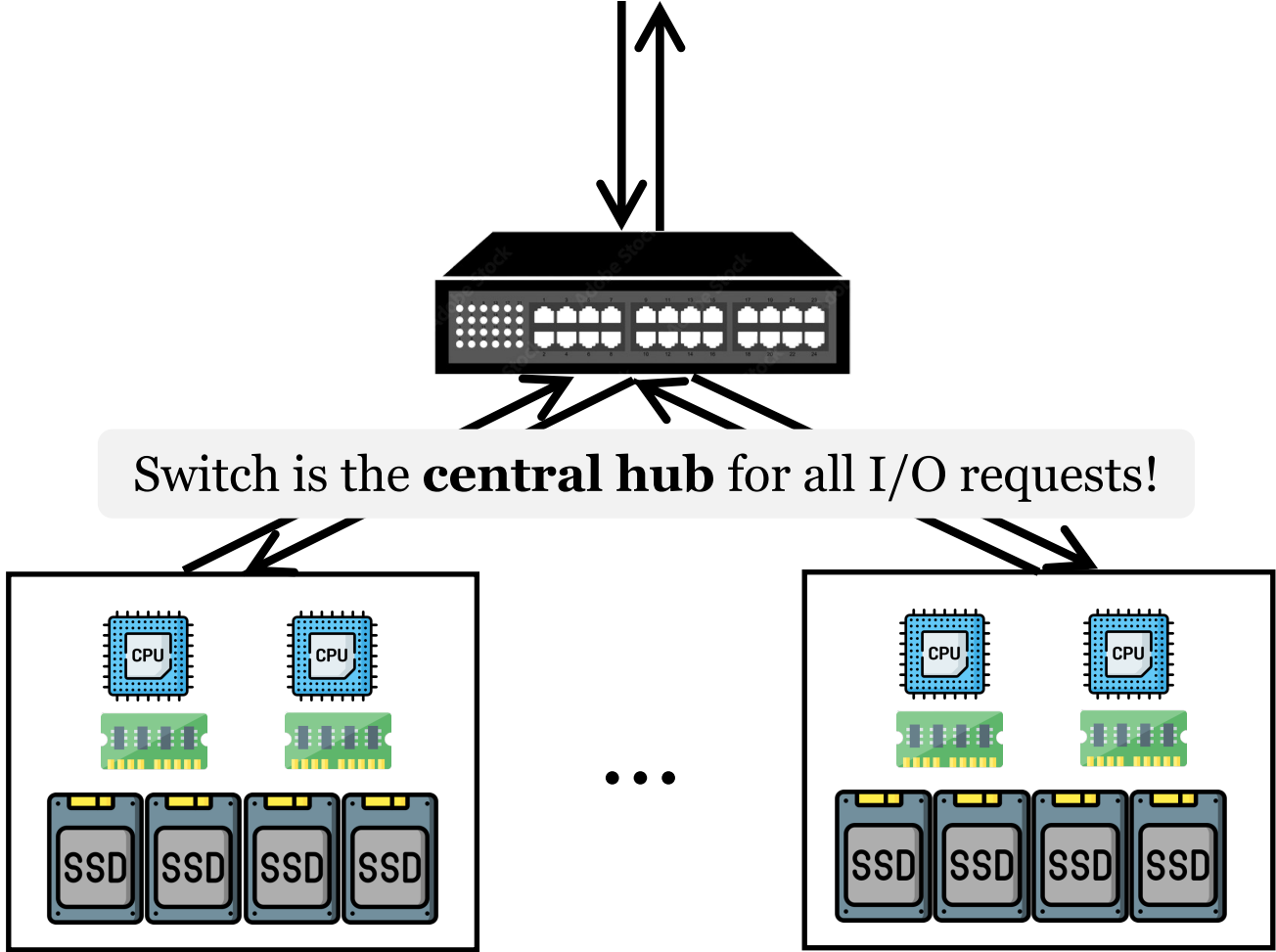


Research Problem: Lack of Coordination between SDN/SDF



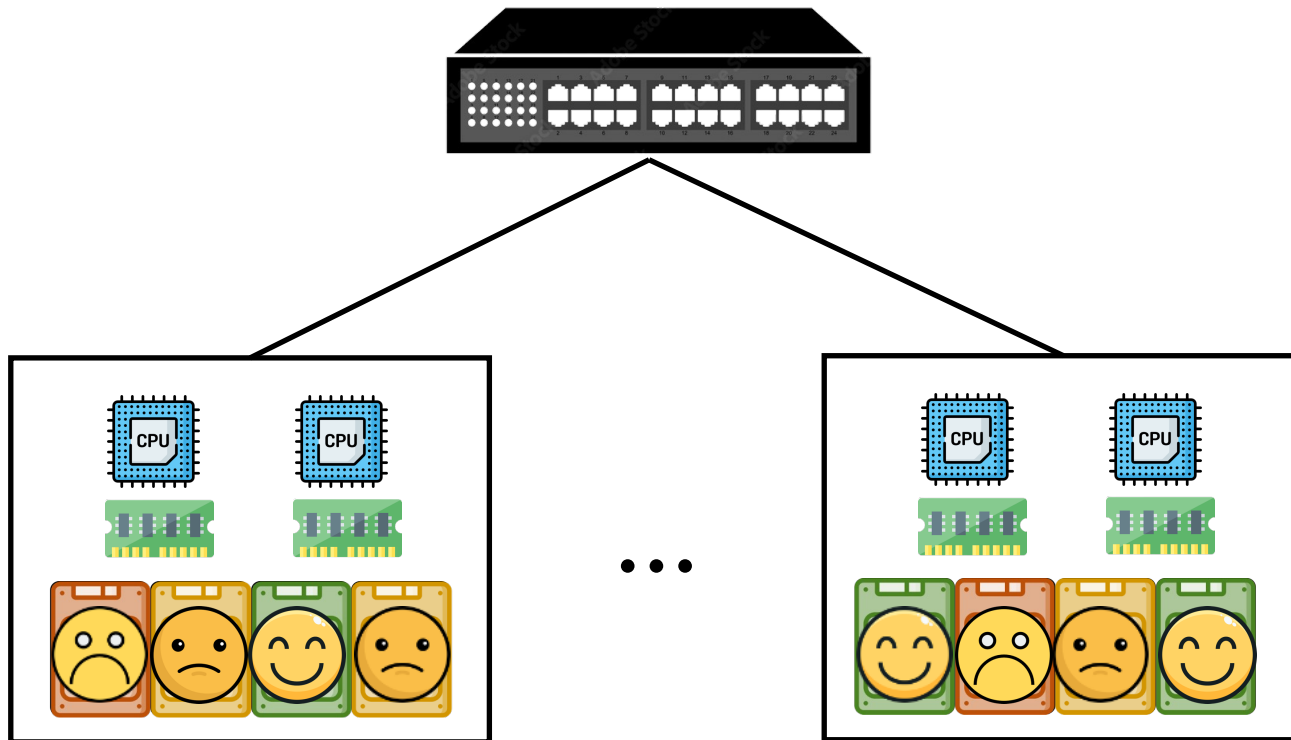
**Does not leverage
existing data replicas
during GC!**

Research Problem: Lack of Coordination between SDN/SDF

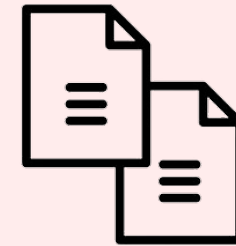


Does not leverage existing data replicas during GC!

Research Problem: Lack of Coordination between SDN/SDF

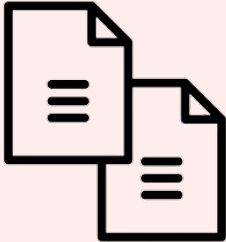
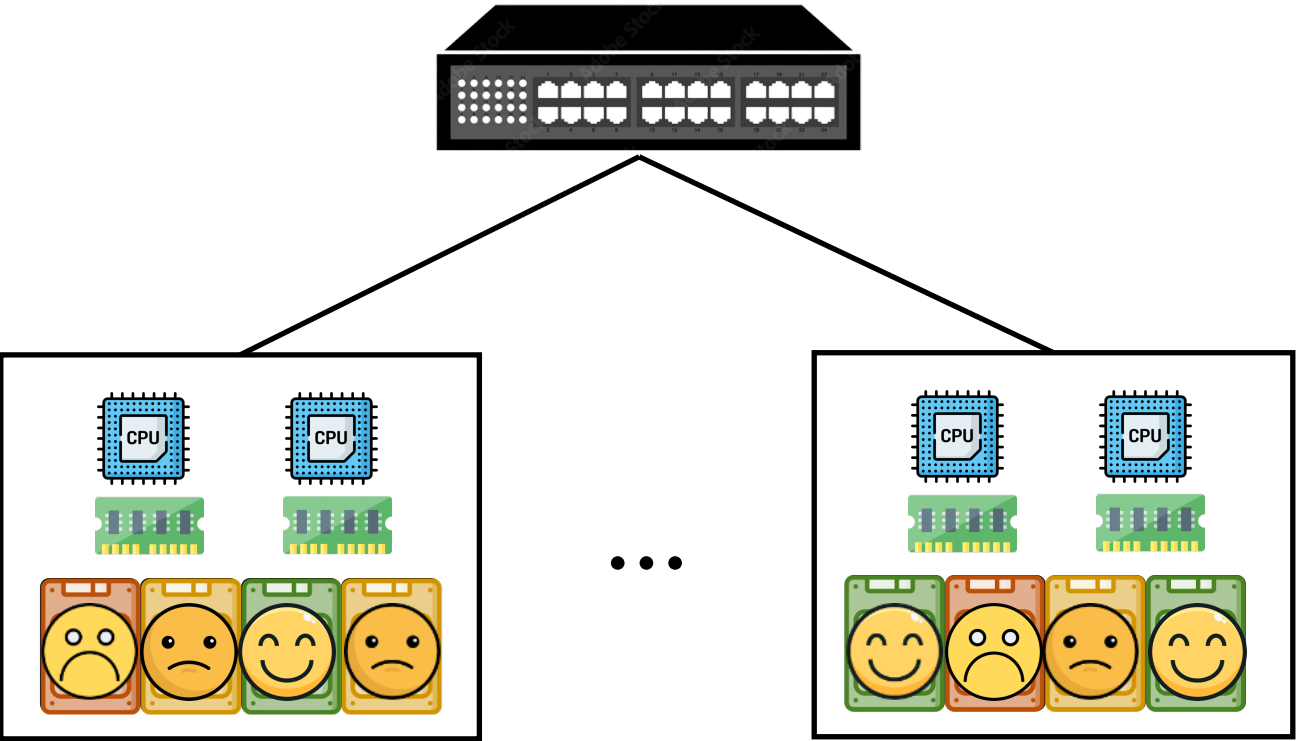


Different SSDs may have **different write traffic!**

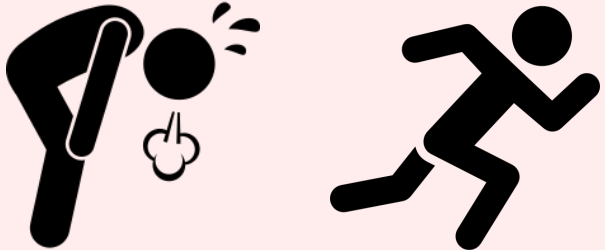


**Does not leverage
existing data replicas
during GC!**

Research Problem: Lack of Coordination between SDN/SDF

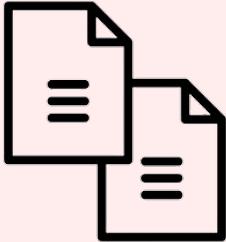
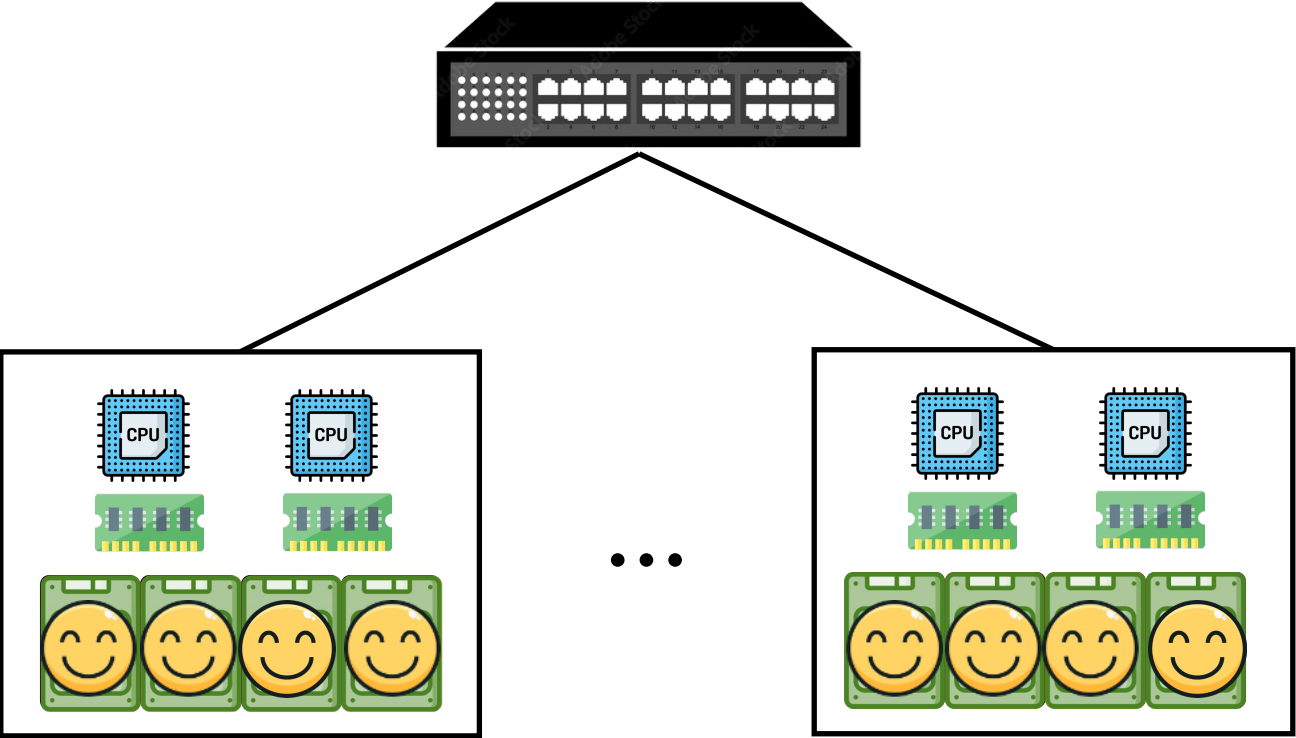


Does not leverage existing data replicas during GC!

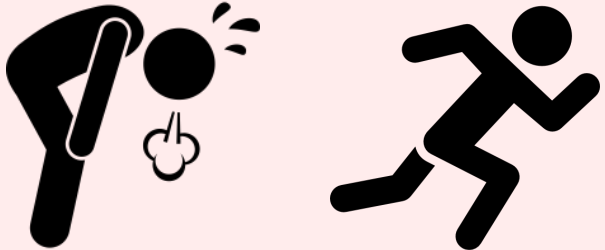


SSDs wear out an uneven rate!

Research Problem: Lack of Coordination between SDN/SDF

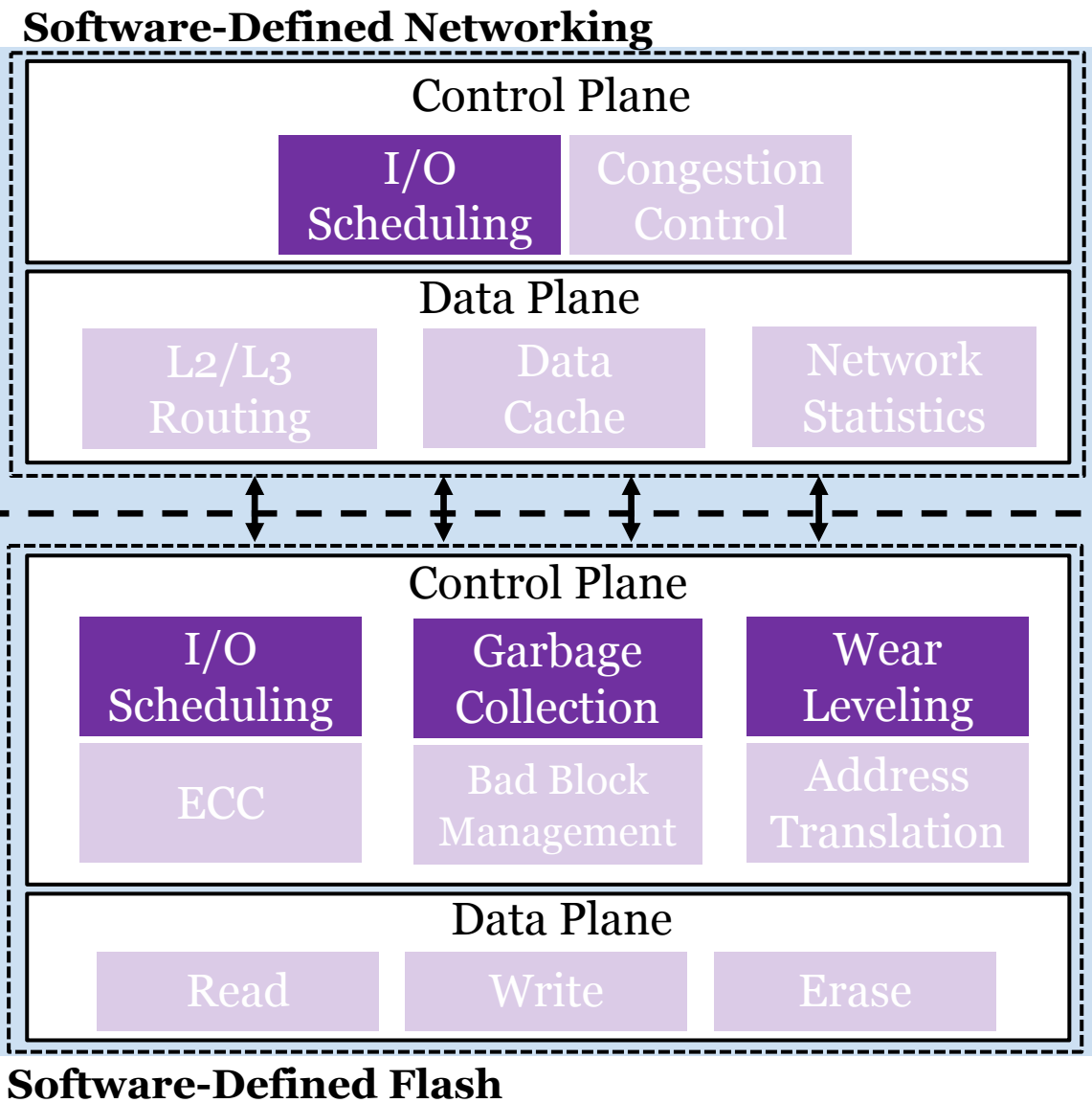


Does not leverage existing data replicas during GC!



SSDs wear out an uneven rate!

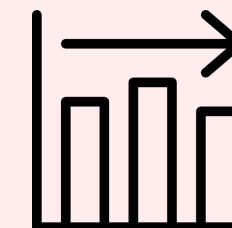
Research Problem: Lack of Coordination between SDN/SDF



Conflicting policies!



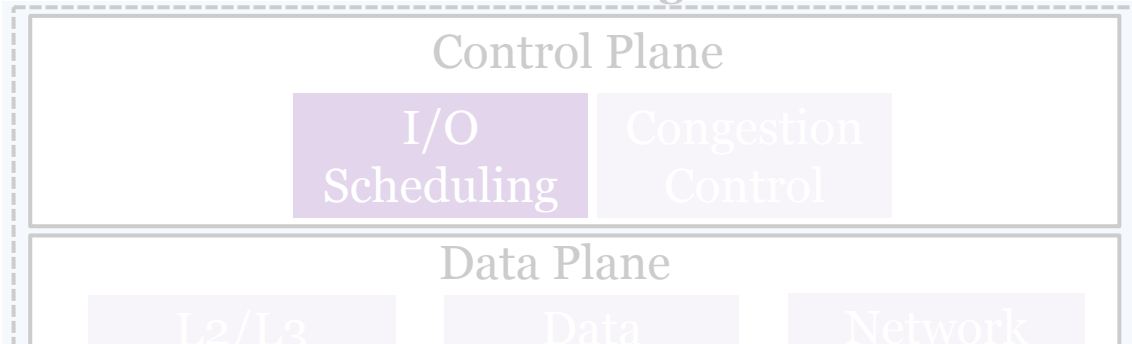
Redundant effort!



Missed opportunities for rack-scale optimization!

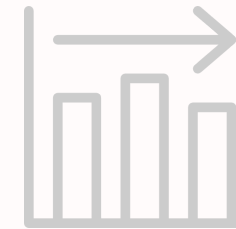
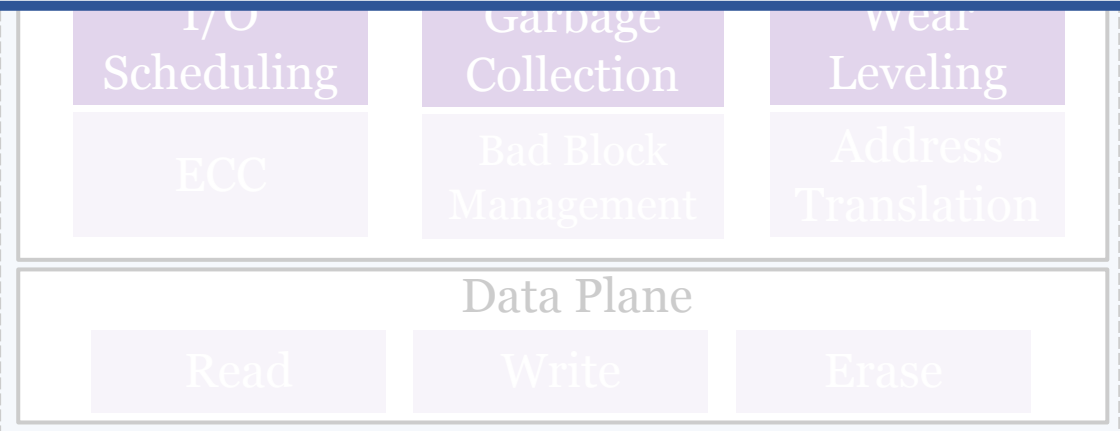
Research Problem: Lack of Coordination between SDN/SDF

Software-Defined Networking



Conflicting policies!

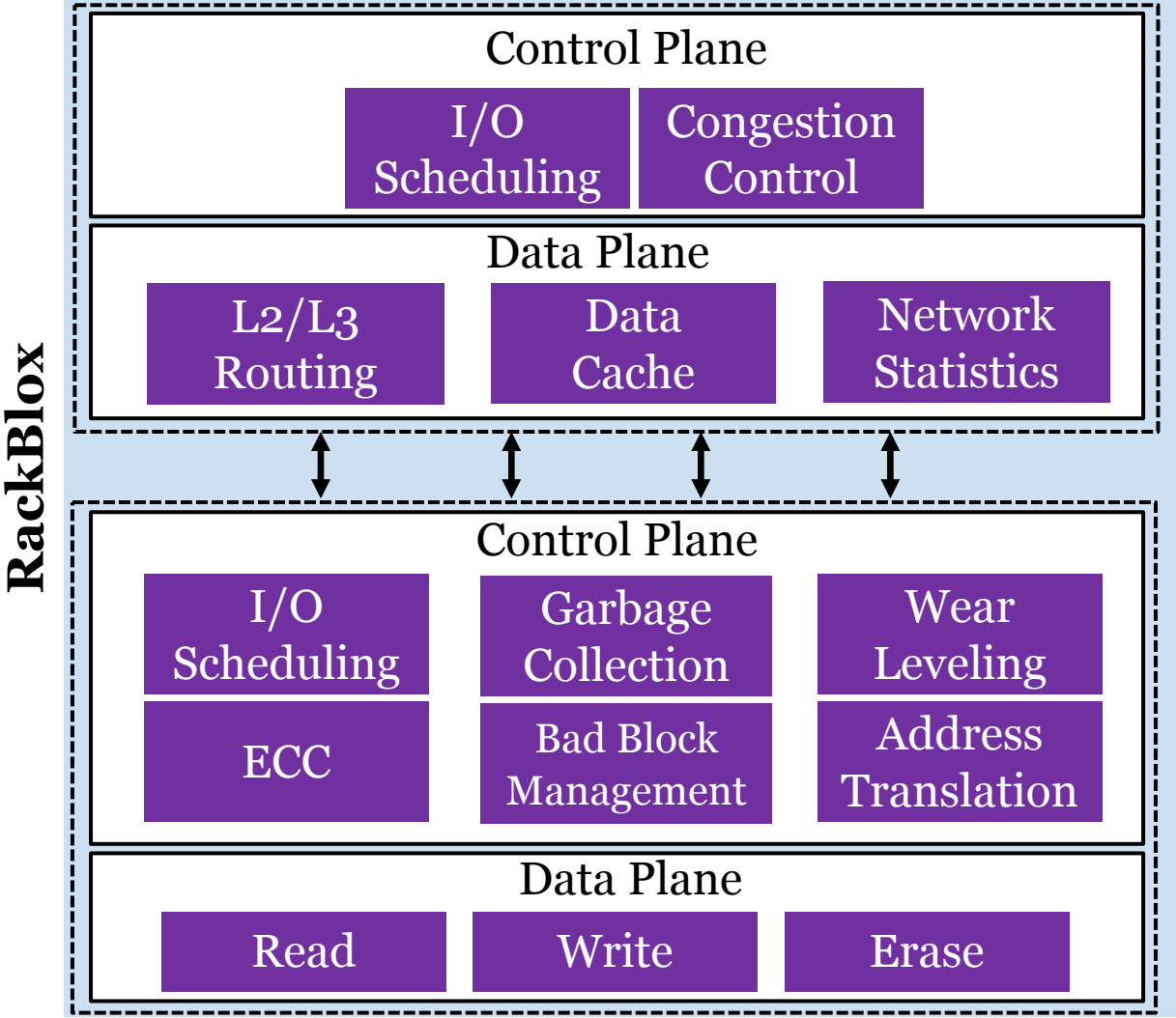
SDN and SDF should coordinate!



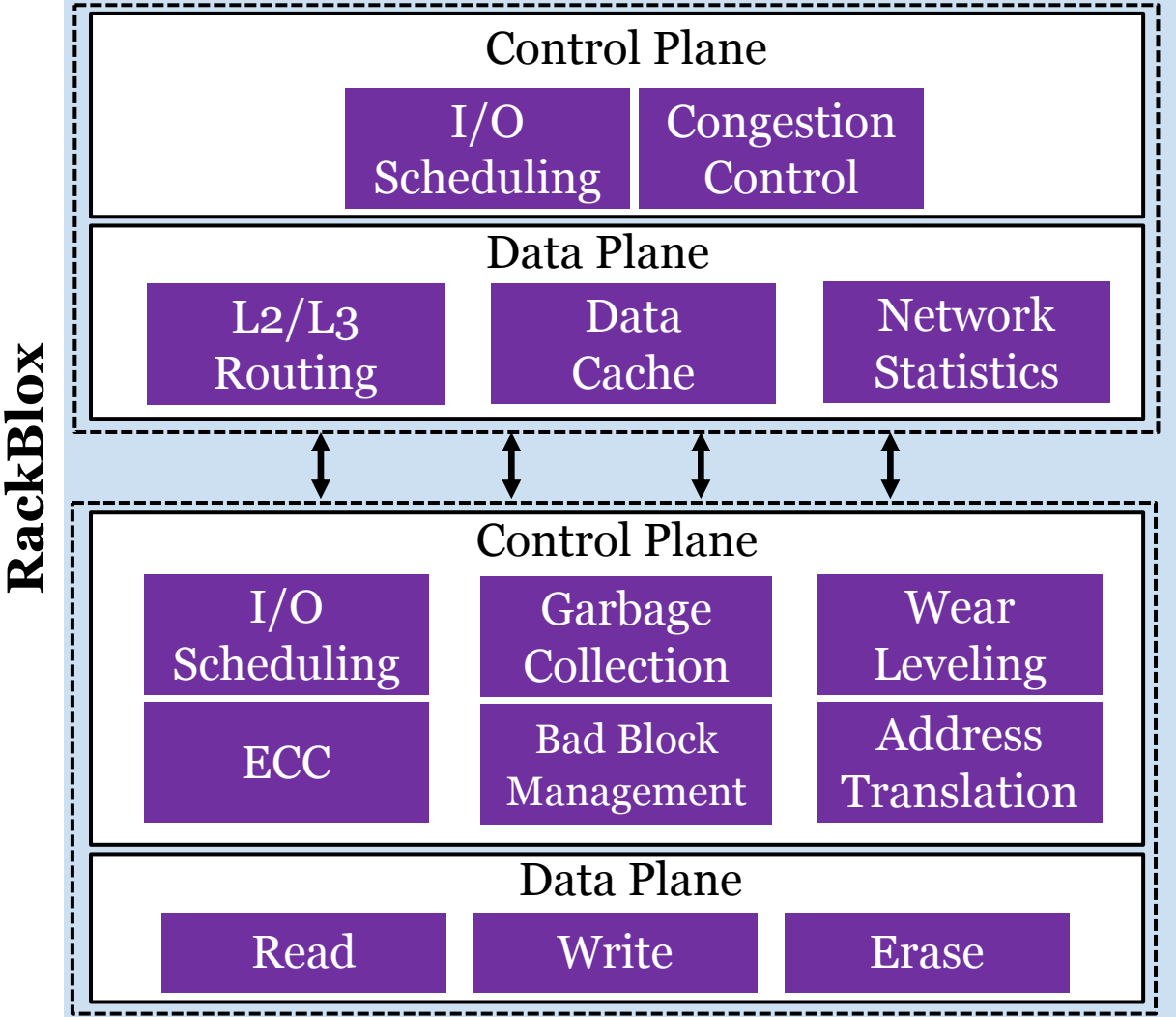
Missed opportunities for rack-scale optimization!

Software-Defined Flash

RackBlox: A Software-Defined Rack-Scale Storage System

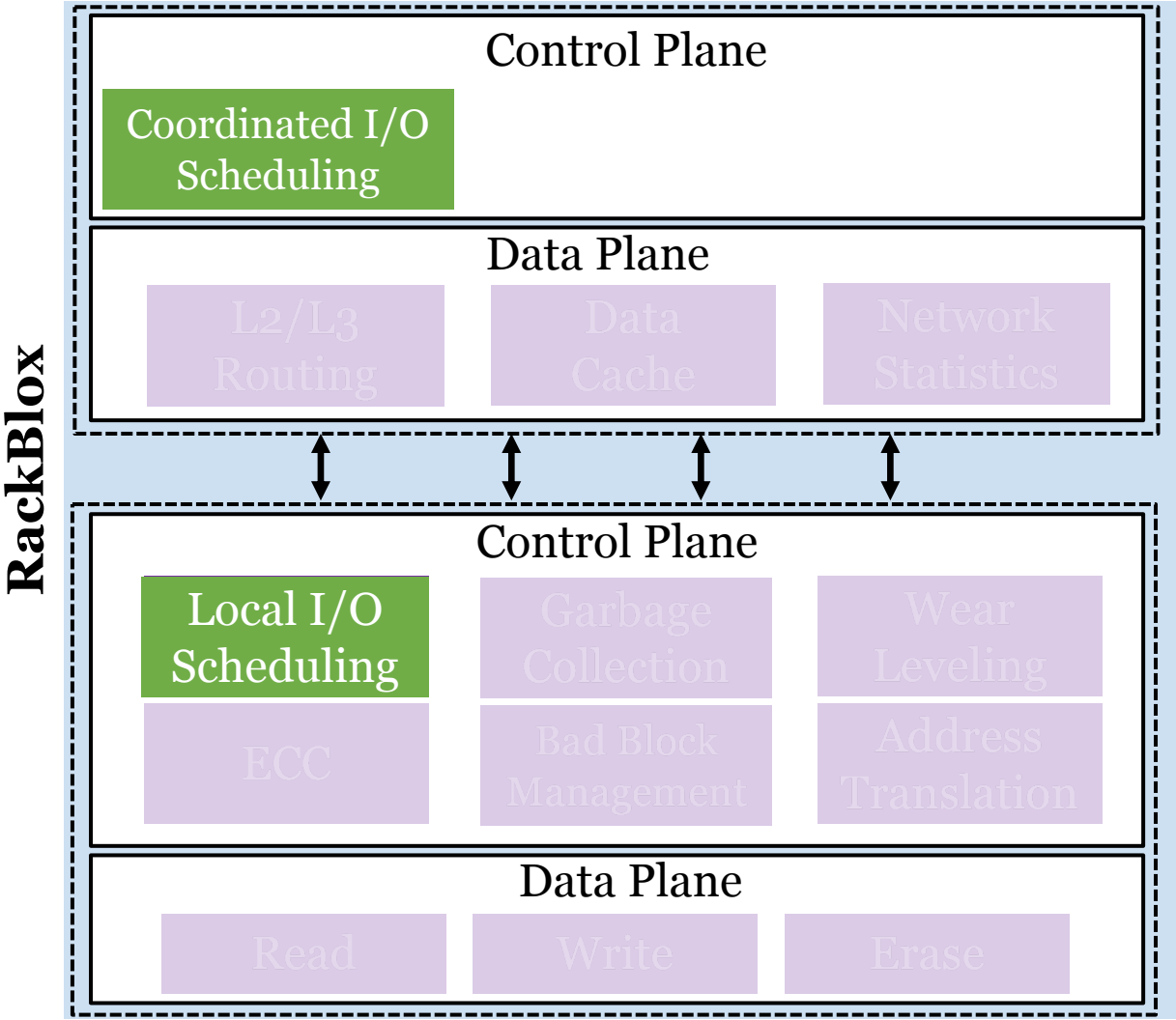


RackBlox: A Software-Defined Rack-Scale Storage System




Decouple storage management

RackBlox: A Software-Defined Rack-Scale Storage System

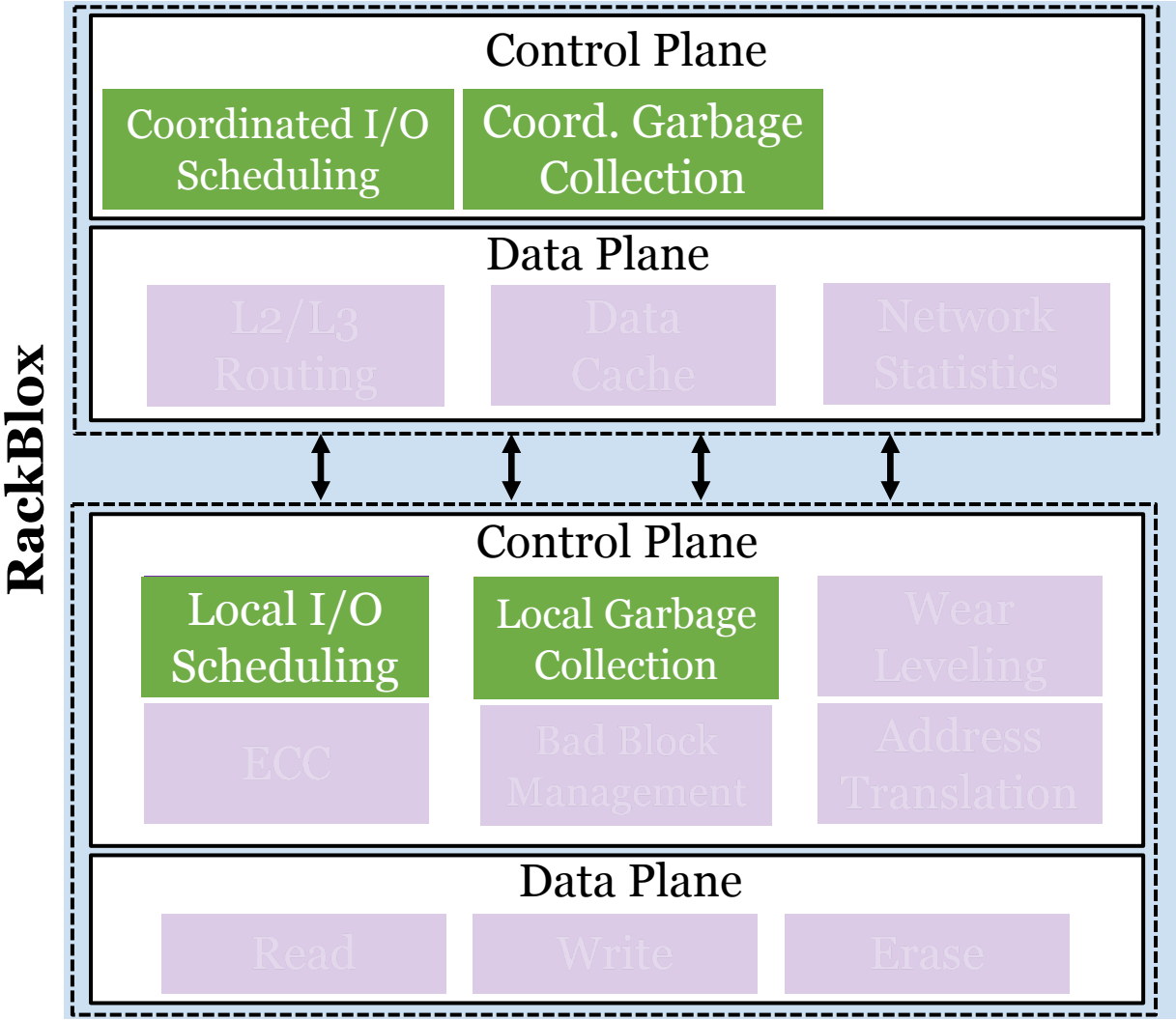


Decouple storage management




Enable coordinated I/O scheduling

RackBlox: A Software-Defined Rack-Scale Storage System



Decouple storage management

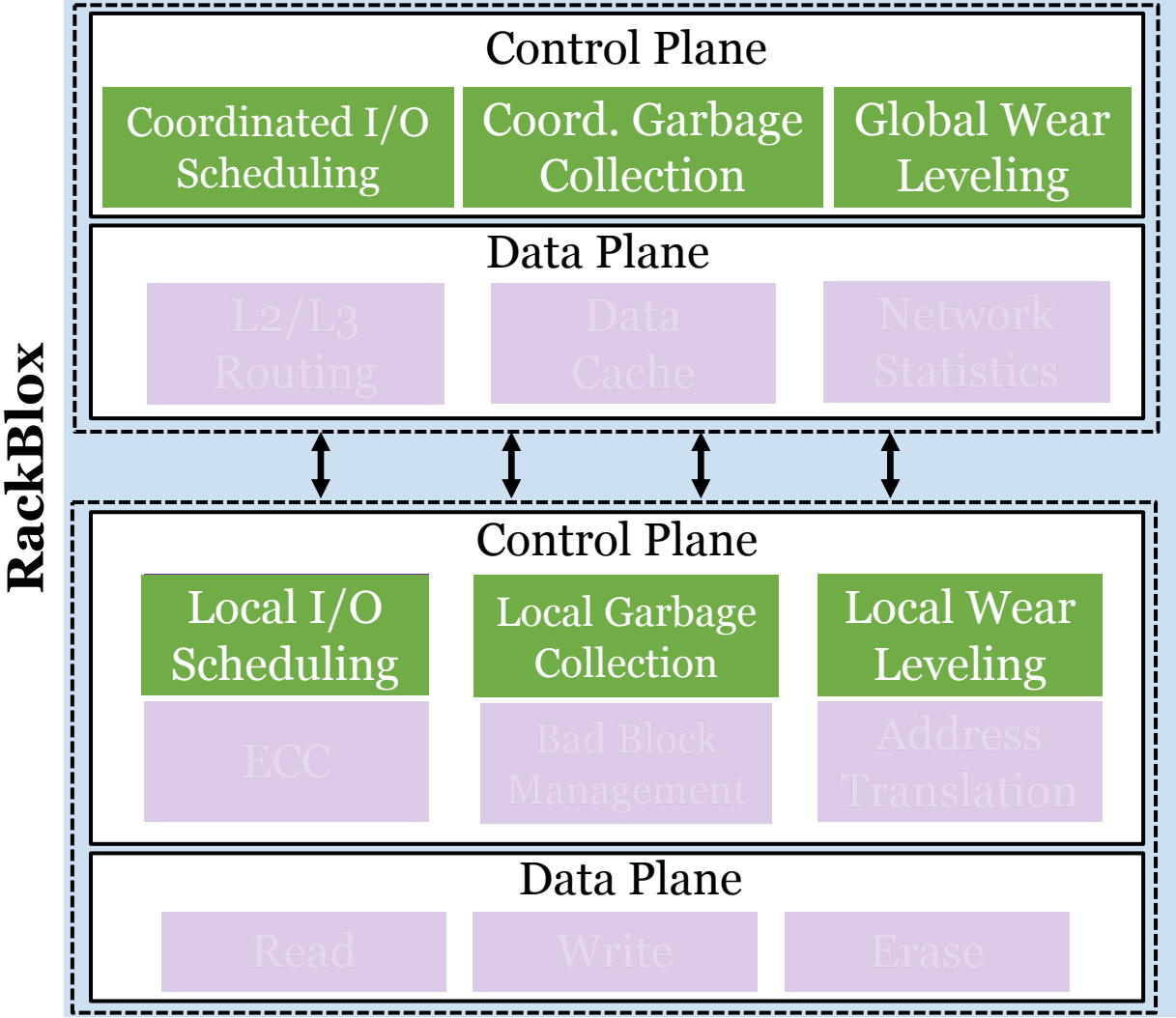


Enable coordinated I/O scheduling




Enable coordinated garbage collection

RackBlox: A Software-Defined Rack-Scale Storage System



Decouple storage management



Enable coordinated I/O scheduling

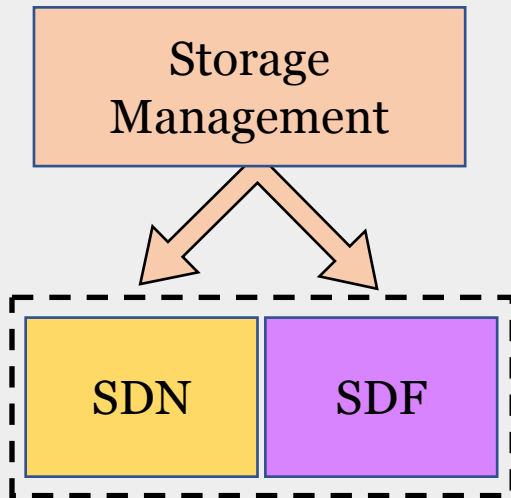


Enable coordinated garbage collection



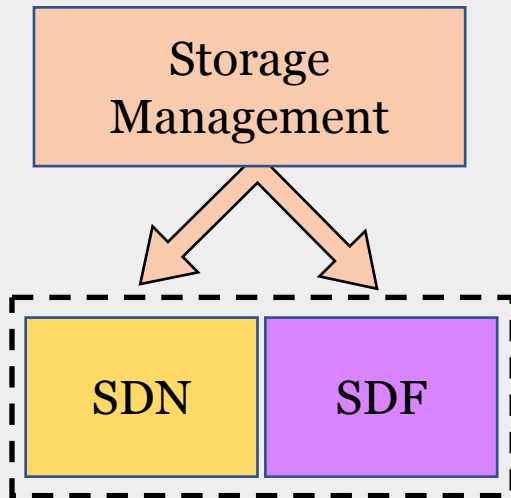
Enable rack-scale wear leveling

Enabling SDN/SDF Codesign is Challenging



How to **decouple** the storage management

Enabling SDN/SDF Codesign is Challenging

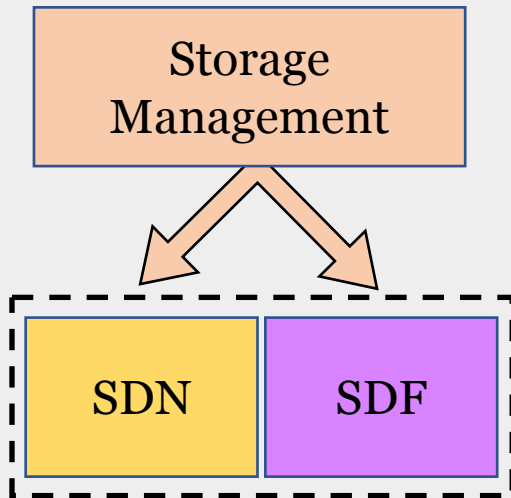


How to **decouple** the storage management



Limited hardware resources in the switch

Enabling SDN/SDF Codesign is Challenging



How to **decouple** the storage management

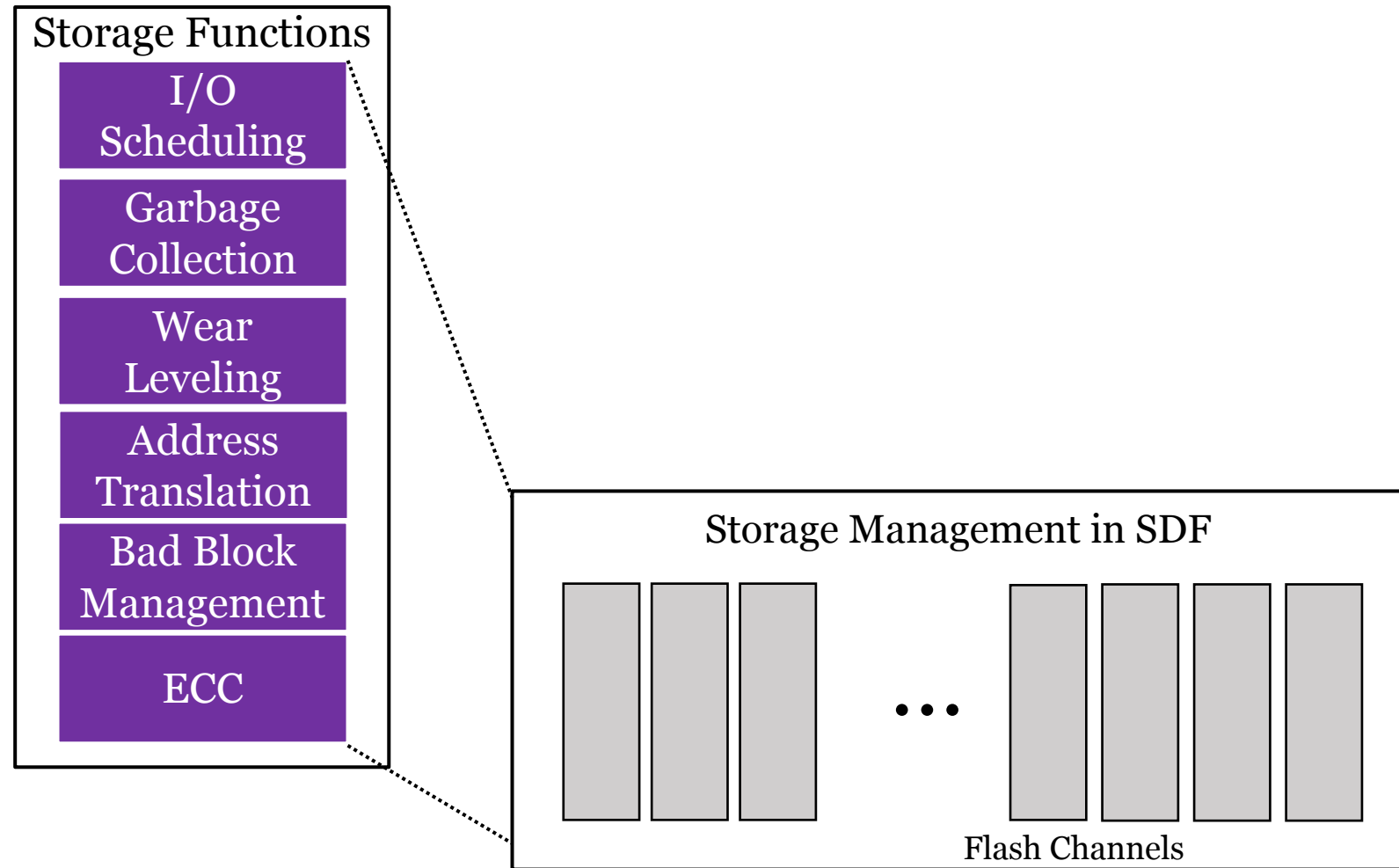


Limited hardware resources in the switch

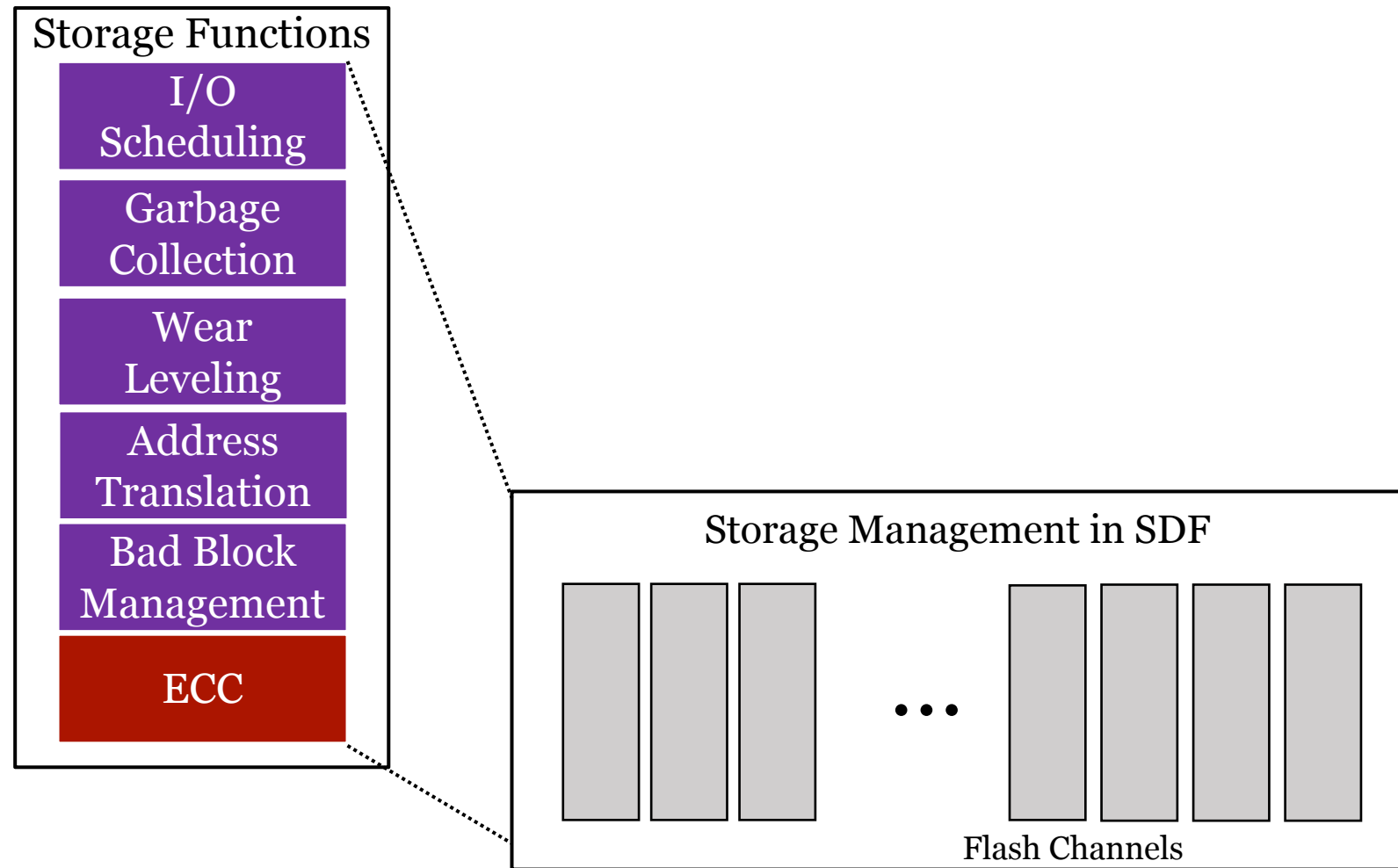


Ensure **flexibility** and **ease-of-use**

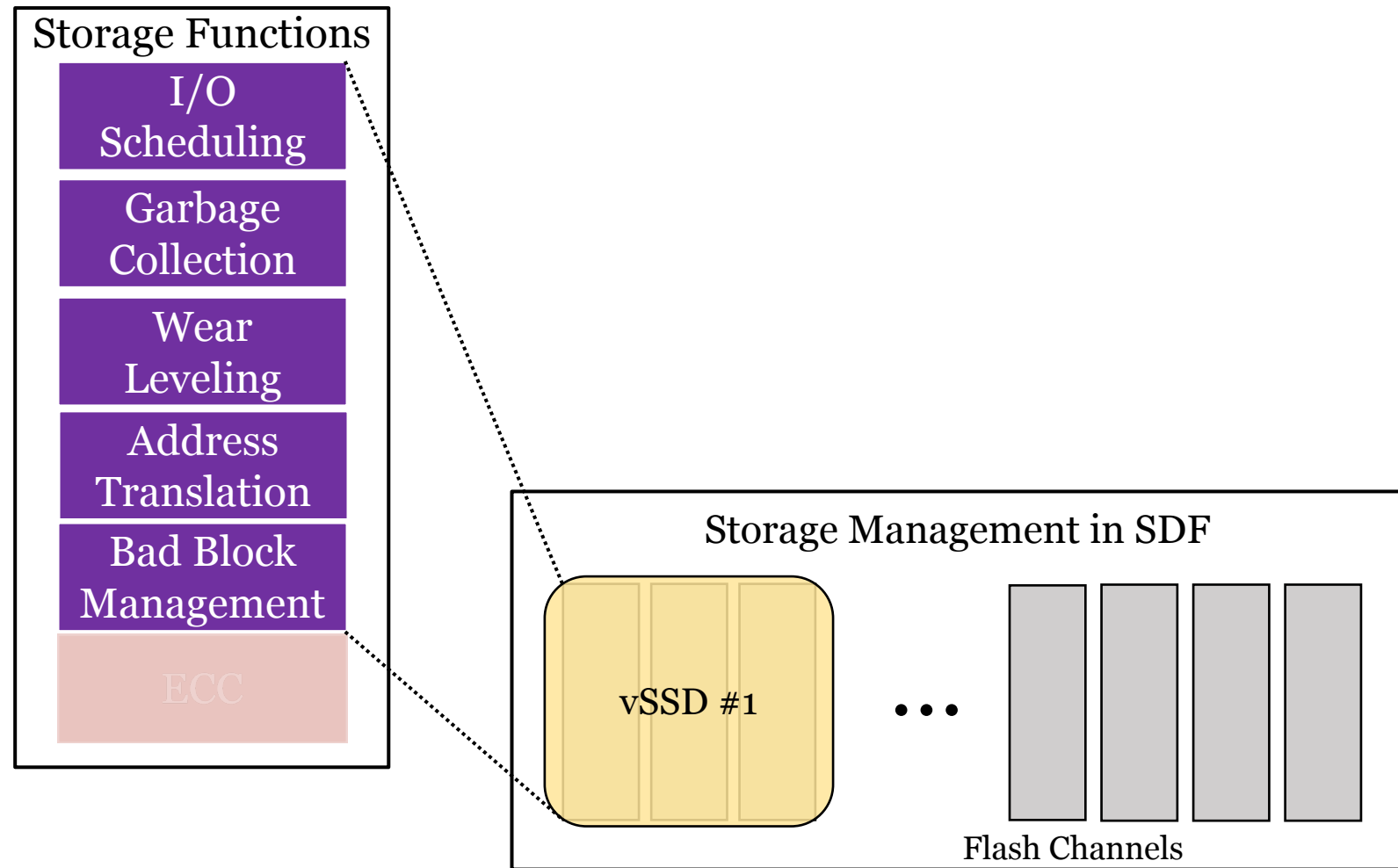
Decoupling the Storage Management Across SDN/SDF



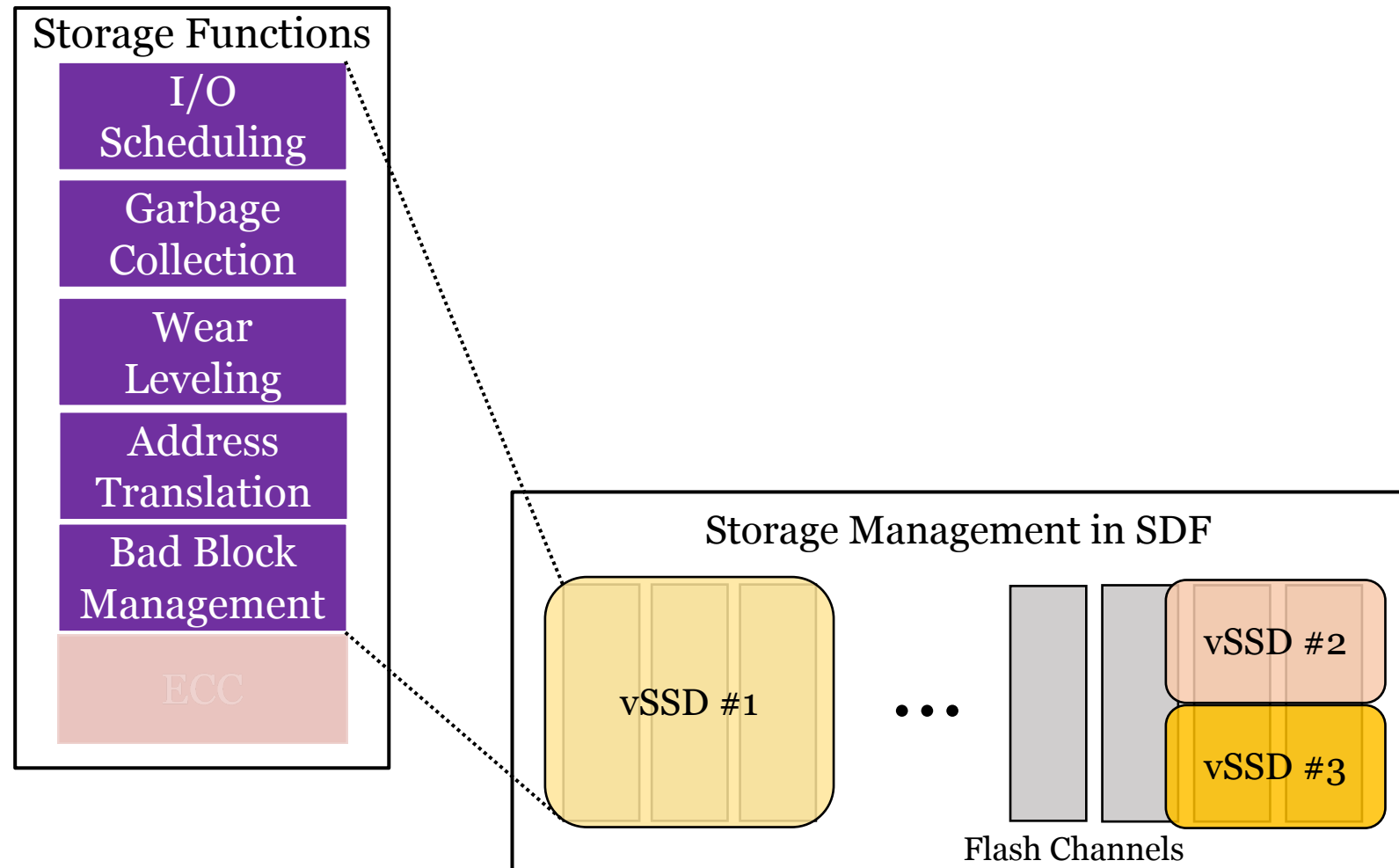
Decoupling the Storage Management Across SDN/SDF



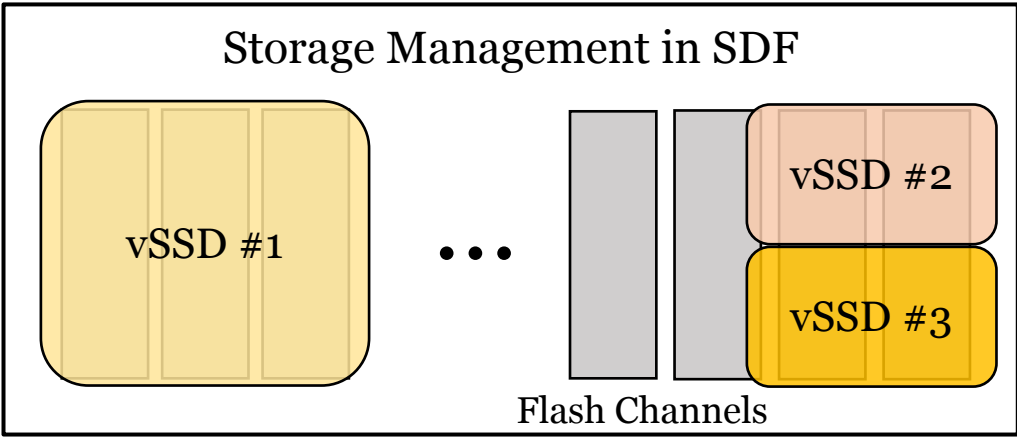
Decoupling the Storage Management Across SDN/SDF



Decoupling the Storage Management Across SDN/SDF




Decoupling the Storage Management Across SDN/SDF



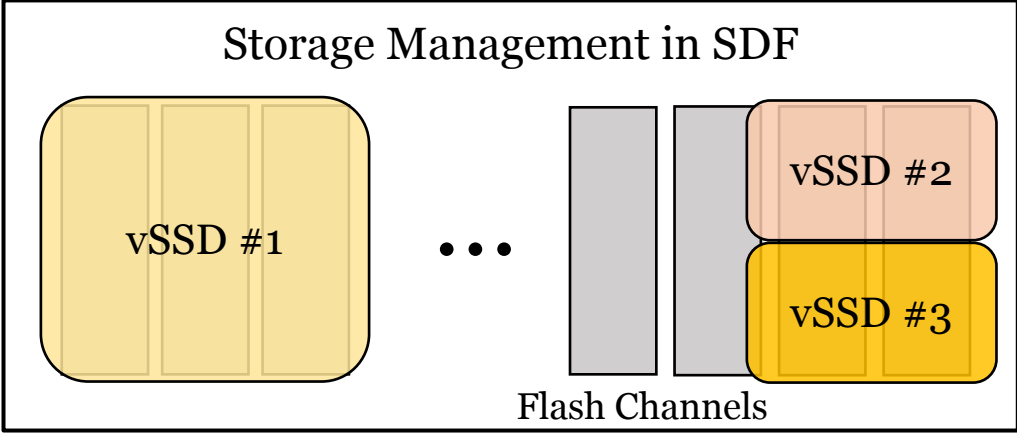
Decoupling the Storage Management Across SDN/SDF



Benefit from coordination




State retained in switch?



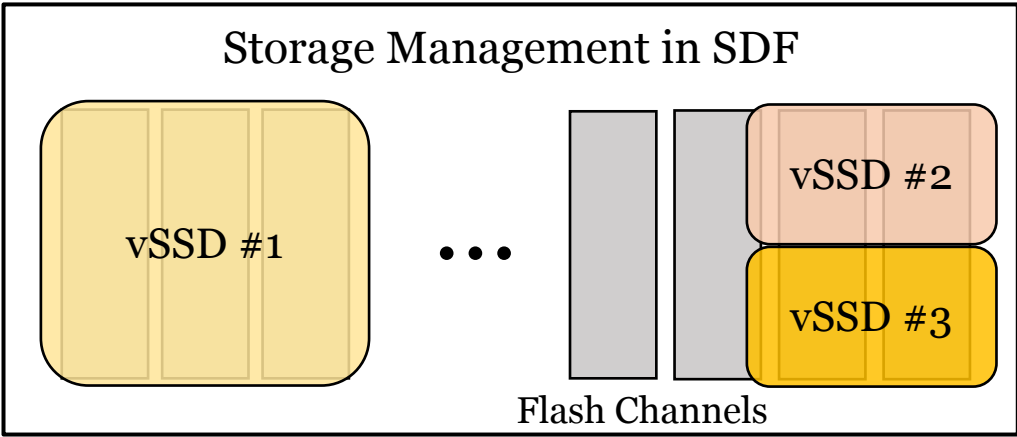
Decoupling the Storage Management Across SDN/SDF



Benefit from coordination




State retained in switch?



Decoupling the Storage Management Across SDN/SDF



Benefit from coordination



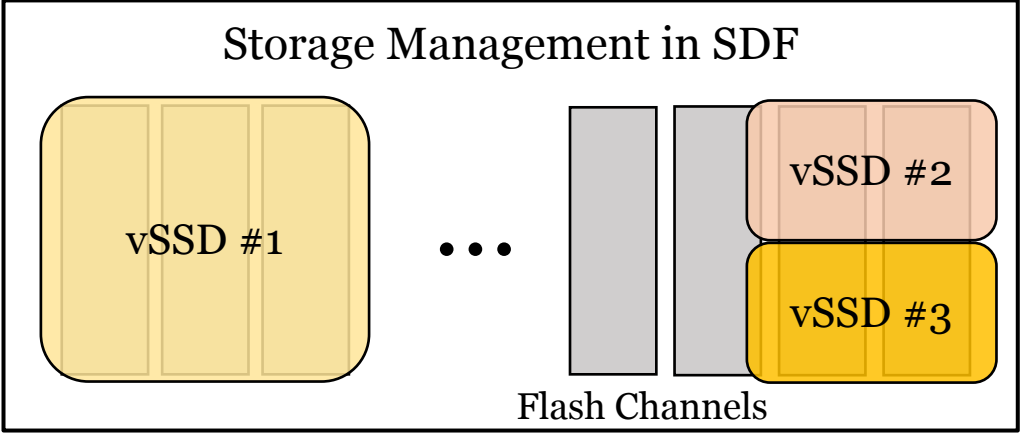
State retained in switch?



Storage Management in SDN

vSSD_ID	GC Status	Replica vSSD_ID
vSSD1	1	vSSD12
vSSD2	0	vSSD20

vSSD_ID	GC Status	Server IP
vSSD1	1	10.0.0.16
vSSD2	0	10.0.0.20



State Communication Between SDN/SDF

Customized packet for RackBlox

ETH	IP	TCP/UDP	OP	vSSD_ID	Latency	Payload
------------	-----------	----------------	-----------	----------------	----------------	----------------

State Communication Between SDN/SDF

Requires **no** hardware changes!

Customized packet for RackBlox

ETH	IP	TCP/UDP	OP	vSSD_ID	Latency	Payload
------------	-----------	----------------	-----------	----------------	----------------	----------------

State Communication Between SDN/SDF

Requires **no** hardware changes!

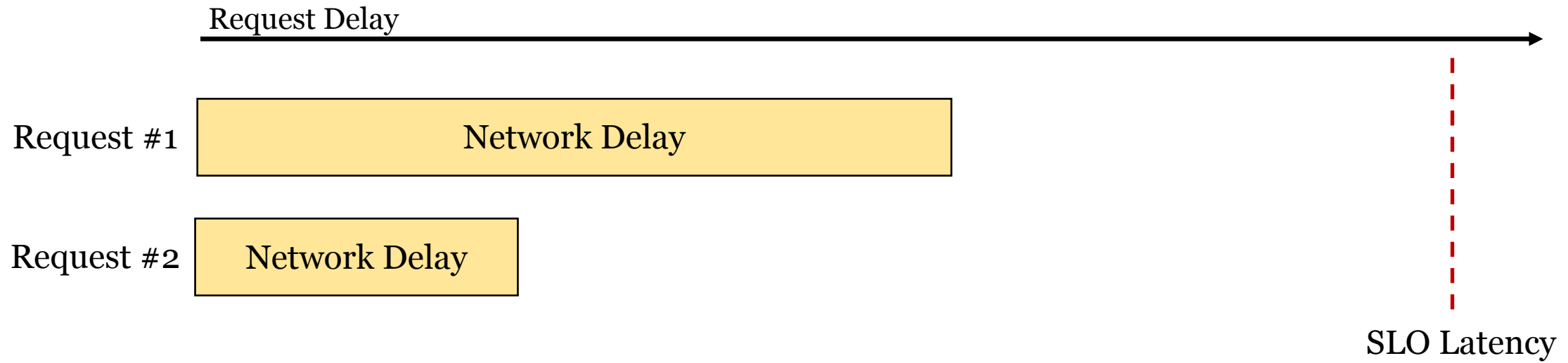
Customized packet for RackBlox

ETH	IP	TCP/UDP	OP	vSSD_ID	Latency	Payload
-----	----	---------	----	---------	---------	---------

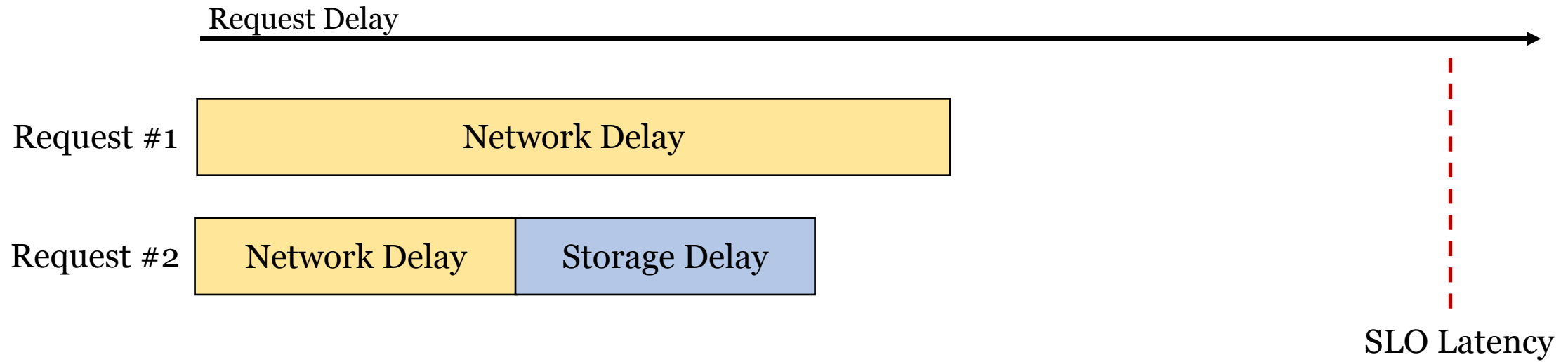


OP code	Operation Name	Description
000	create_vssd	Register new vSSD in switch
001	del_vssd	Remove vSSD from switch
010	write	Client write
011	read	Client read
100	gc_op	Packet to update GC in switch

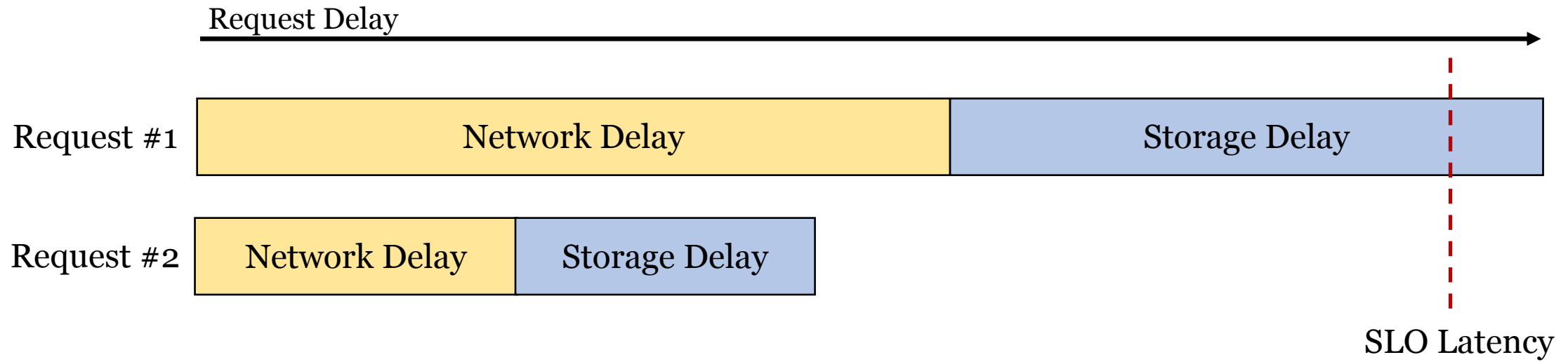
Enabling Coordinated I/O Scheduling



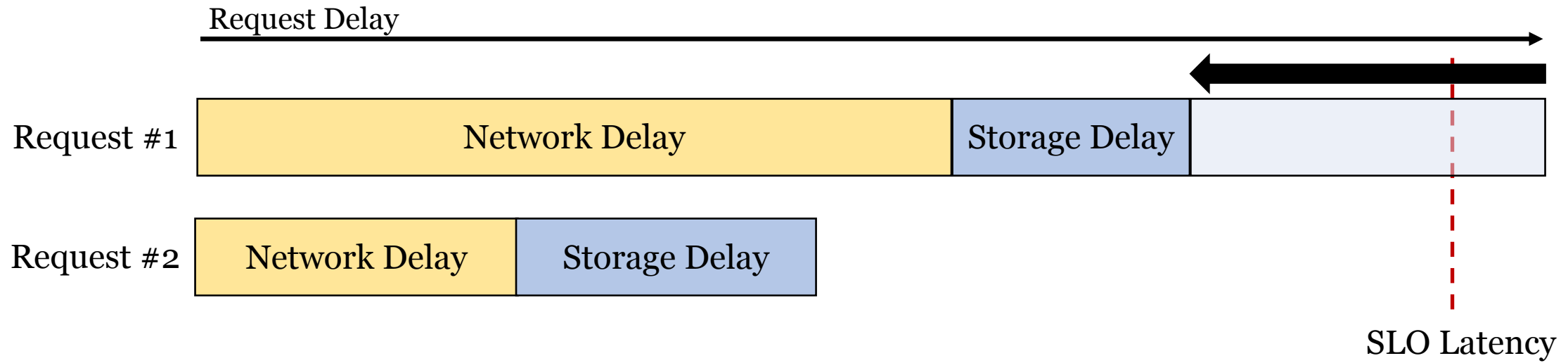
Enabling Coordinated I/O Scheduling



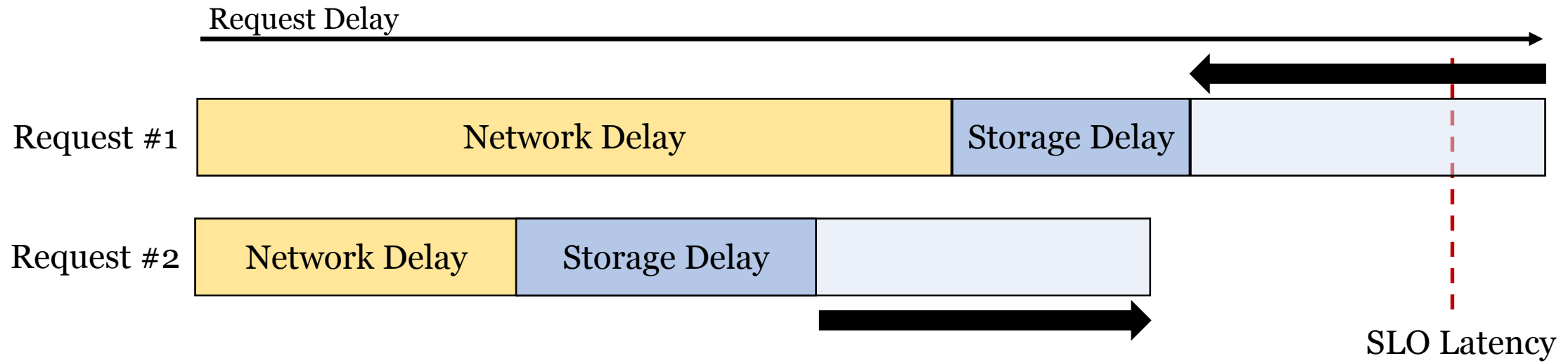
Enabling Coordinated I/O Scheduling



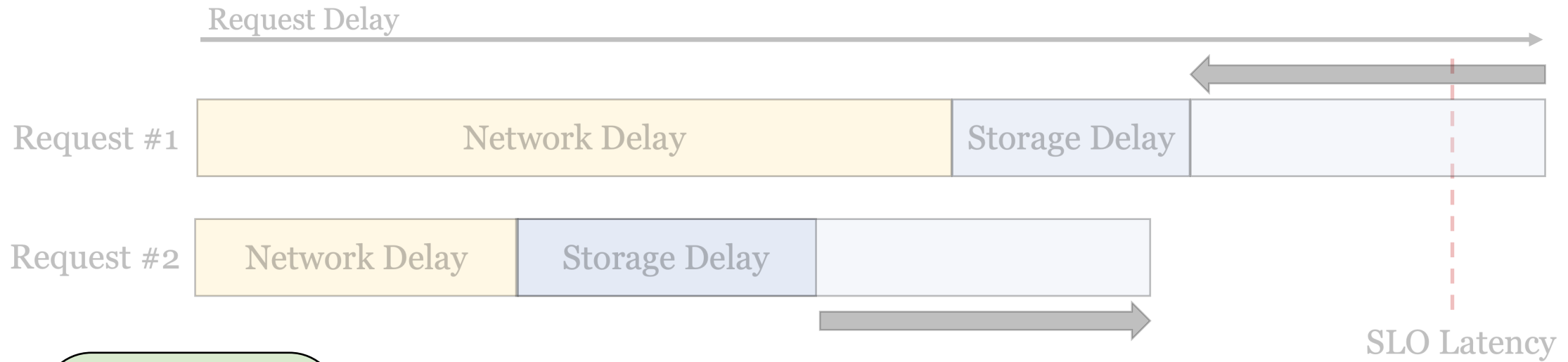
Enabling Coordinated I/O Scheduling



Enabling Coordinated I/O Scheduling

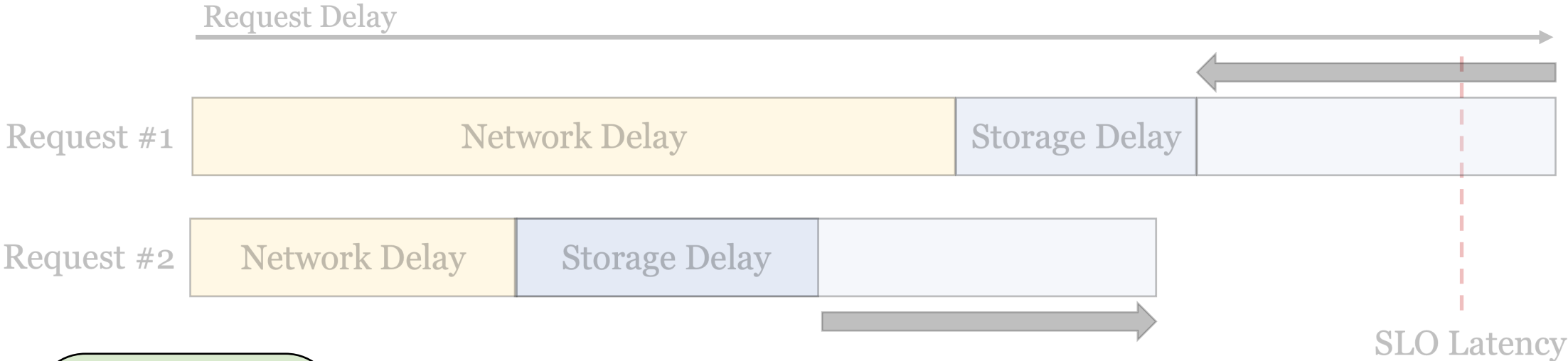


Enabling Coordinated I/O Scheduling



$$\begin{aligned} Prio_{sched} &= \\ &Net_{time} \\ &+ \\ &Storage_{time} \\ &+ \\ &Predict_{time} \end{aligned}$$

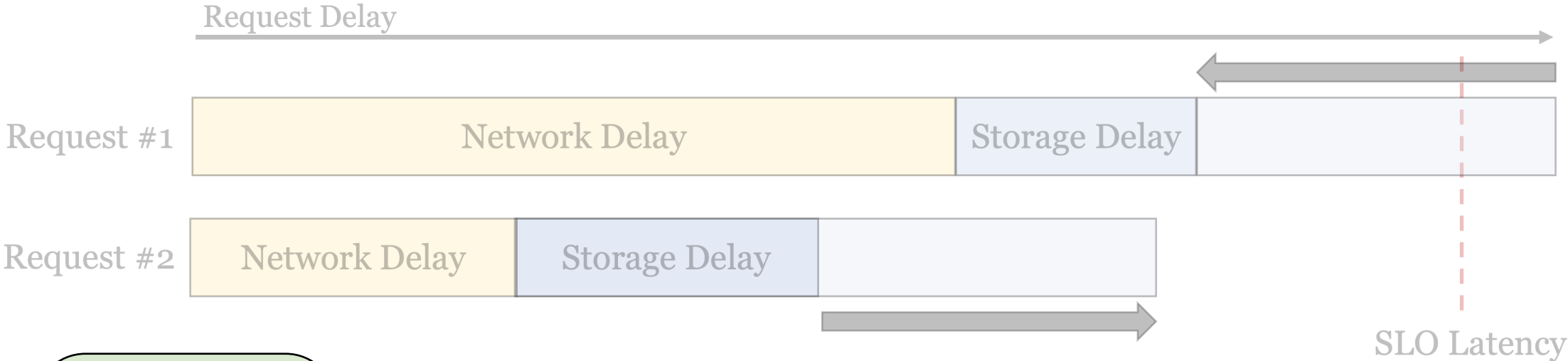
Enabling Coordinated I/O Scheduling



$$\begin{aligned} Prio_{sched} &= \\ &Net_{time} \\ &+ \\ &Storage_{time} \\ &+ \\ &Predict_{time} \end{aligned}$$

In-Network Telemetry 

Enabling Coordinated I/O Scheduling

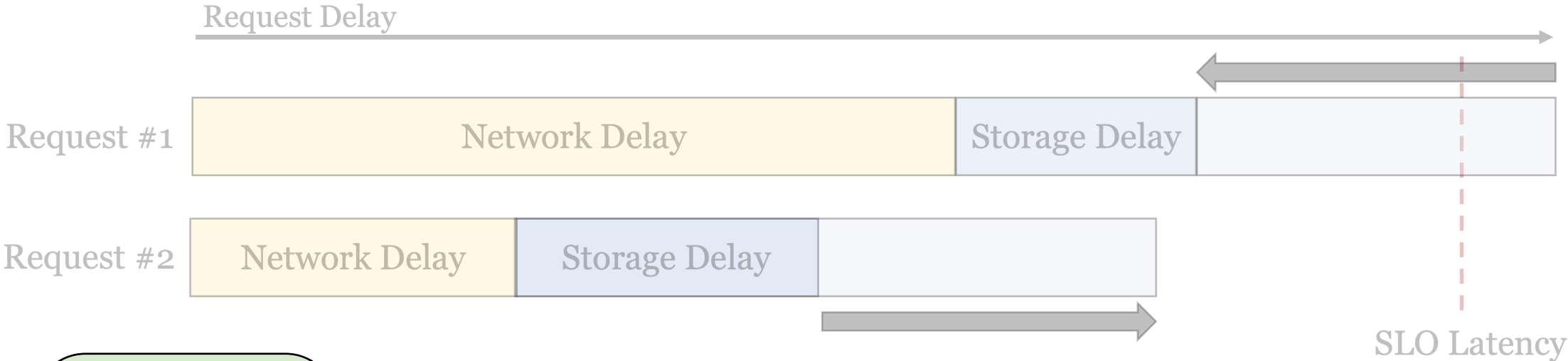


$$\begin{aligned} Prio_{sched} &= \\ &Net_{time} \\ &+ \\ &Storage_{time} \\ &+ \\ &Predict_{time} \end{aligned}$$

In-Network Telemetry 

Storage Queuing Delay 

Enabling Coordinated I/O Scheduling



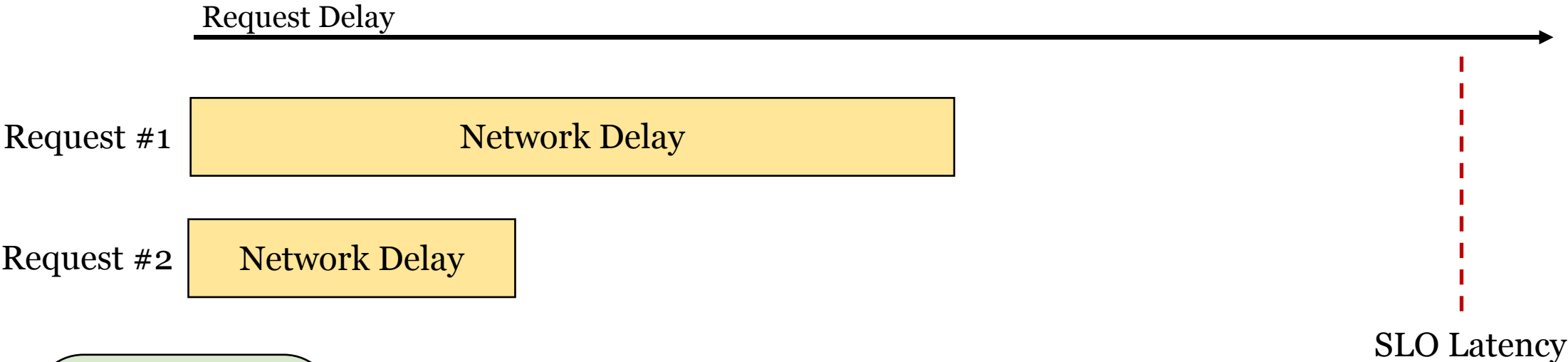
$$\begin{aligned} Prio_{sched} &= \\ &Net_{time} \\ &+ \\ &Storage_{time} \\ &+ \\ &Predict_{time} \end{aligned}$$

In-Network Telemetry 




Storage Queuing Delay 

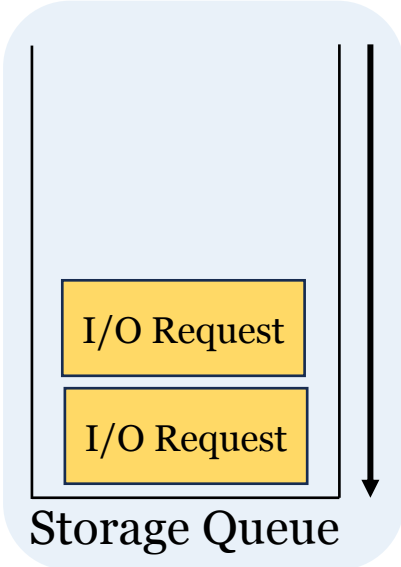
Sliding Window Predictor 

Enabling Coordinated I/O Scheduling

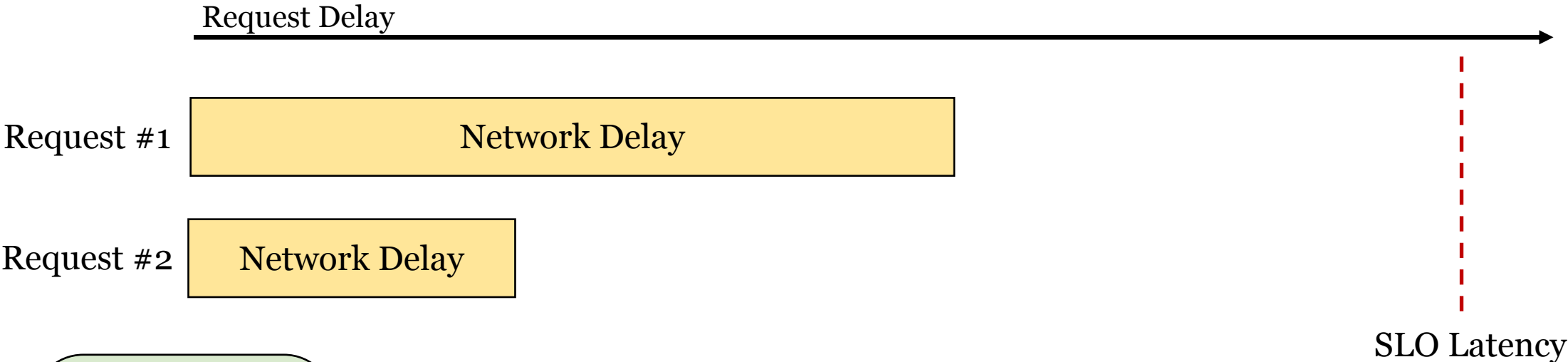


$$\begin{aligned} &Prio_{sched} \\ &= \\ &Net_{time} \\ &+ \\ &Storage_{time} \\ &+ \\ &Predict_{time} \end{aligned}$$




- In-Network Telemetry 
- Storage Queuing Delay 
- Sliding Window Predictor 

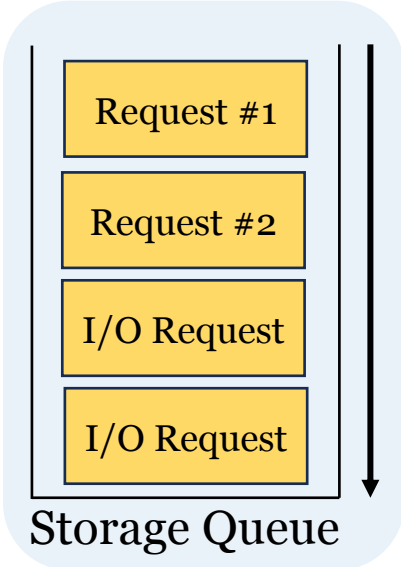


Enabling Coordinated I/O Scheduling

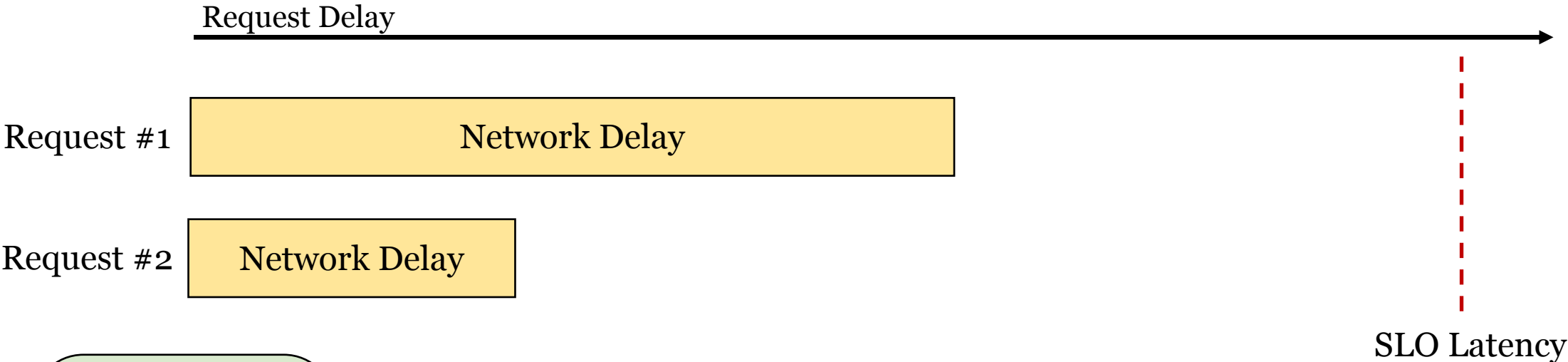


$$\begin{aligned} &Prio_{sched} \\ &= \\ &Net_{time} \\ &+ \\ &Storage_{time} \\ &+ \\ &Predict_{time} \end{aligned}$$




- In-Network Telemetry 
- Storage Queuing Delay 
- Sliding Window Predictor 

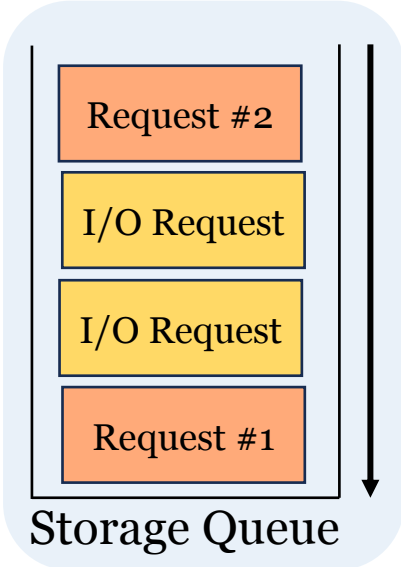


Enabling Coordinated I/O Scheduling

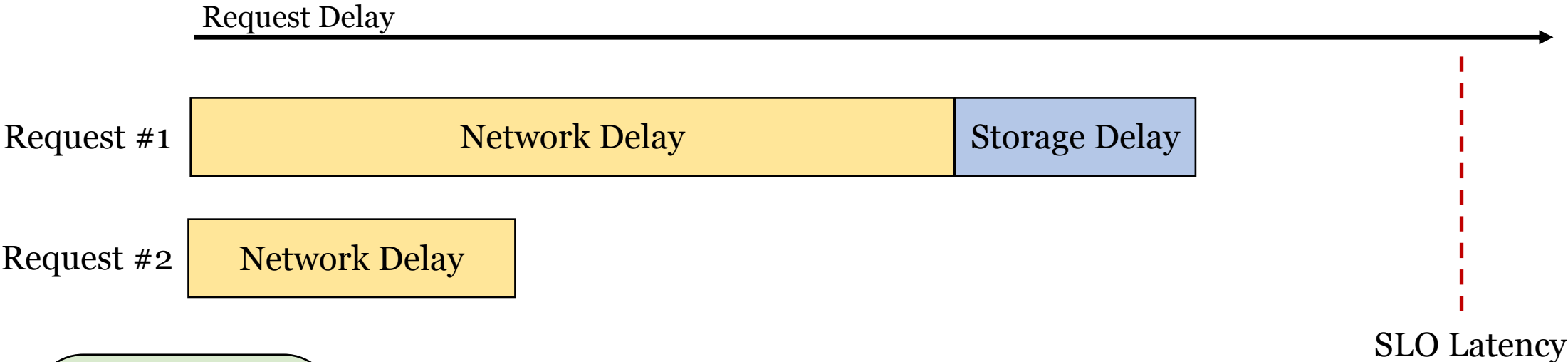


$$\begin{aligned} Prio_{sched} &= \\ &+ Net_{time} \\ &+ Storage_{time} \\ &+ Predict_{time} \end{aligned}$$




- In-Network Telemetry 
- Storage Queuing Delay 
- Sliding Window Predictor 

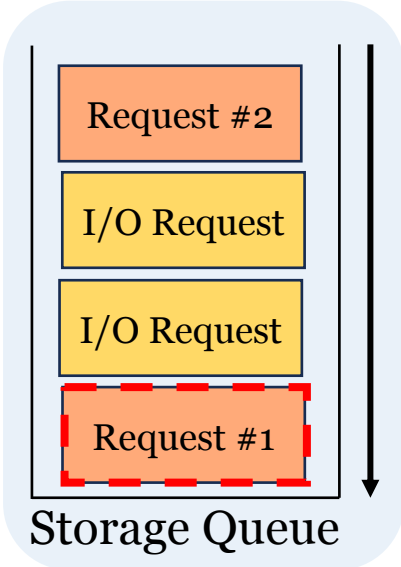


Enabling Coordinated I/O Scheduling

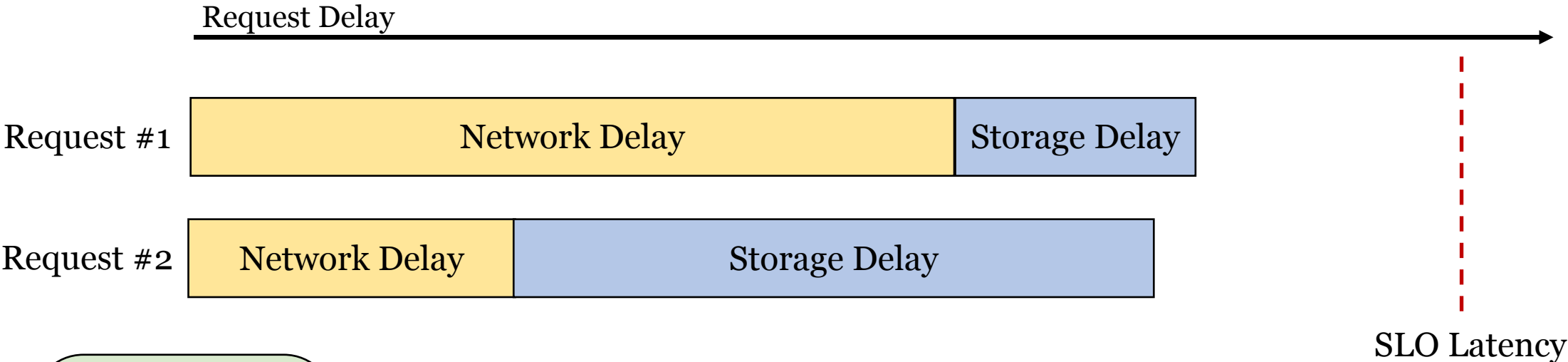


$$\begin{aligned}
 &Prio_{sched} \\
 &= \\
 &Net_{time} \\
 &+ \\
 &Storage_{time} \\
 &+ \\
 &Predict_{time}
 \end{aligned}$$




- In-Network Telemetry 
- Storage Queuing Delay 
- Sliding Window Predictor 

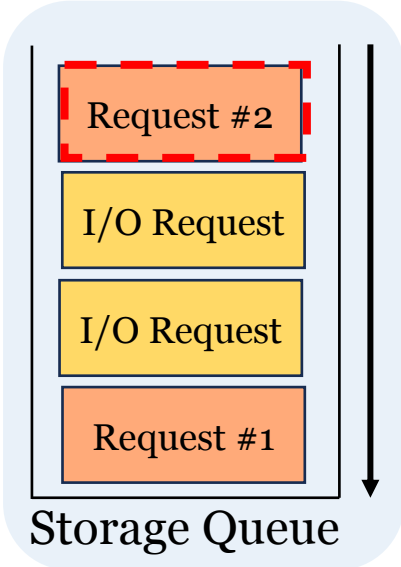


Enabling Coordinated I/O Scheduling

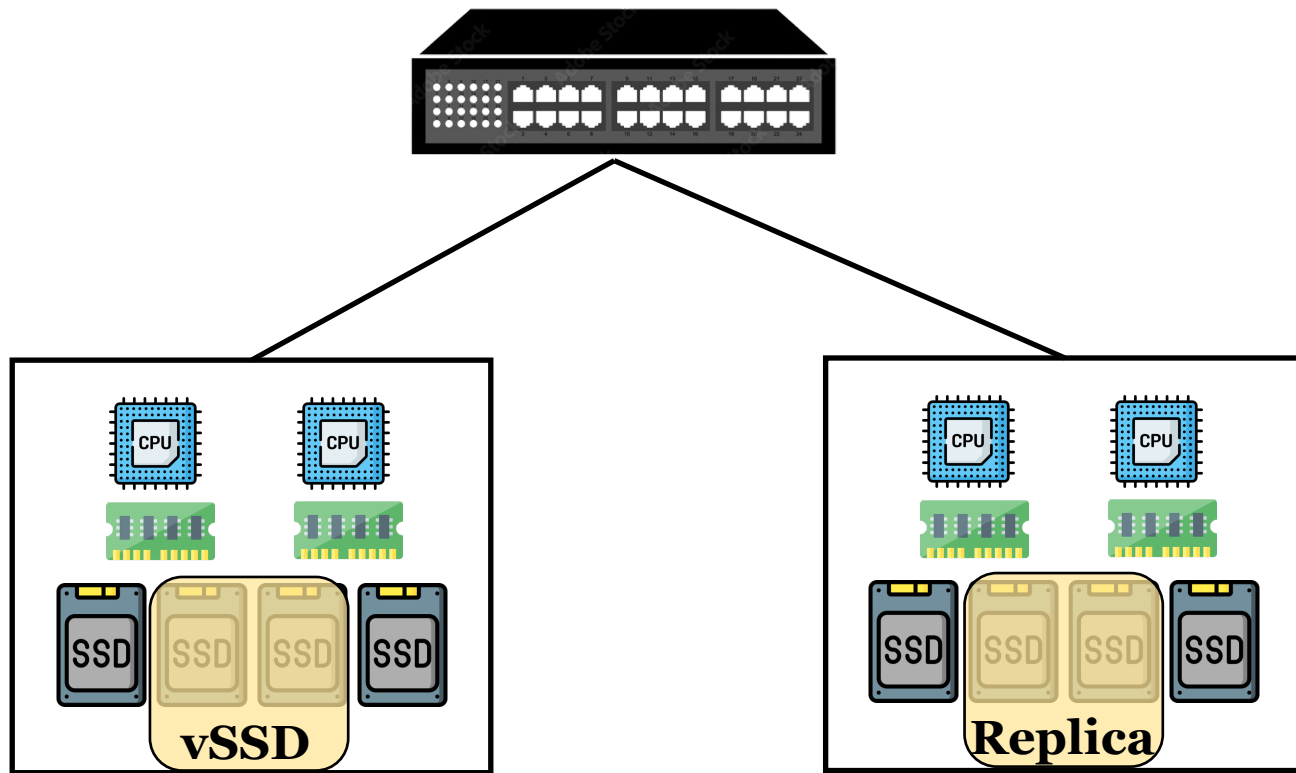


$$\begin{aligned}
 &Prio_{sched} \\
 &= \\
 &Net_{time} \\
 &+ \\
 &Storage_{time} \\
 &+ \\
 &Predict_{time}
 \end{aligned}$$

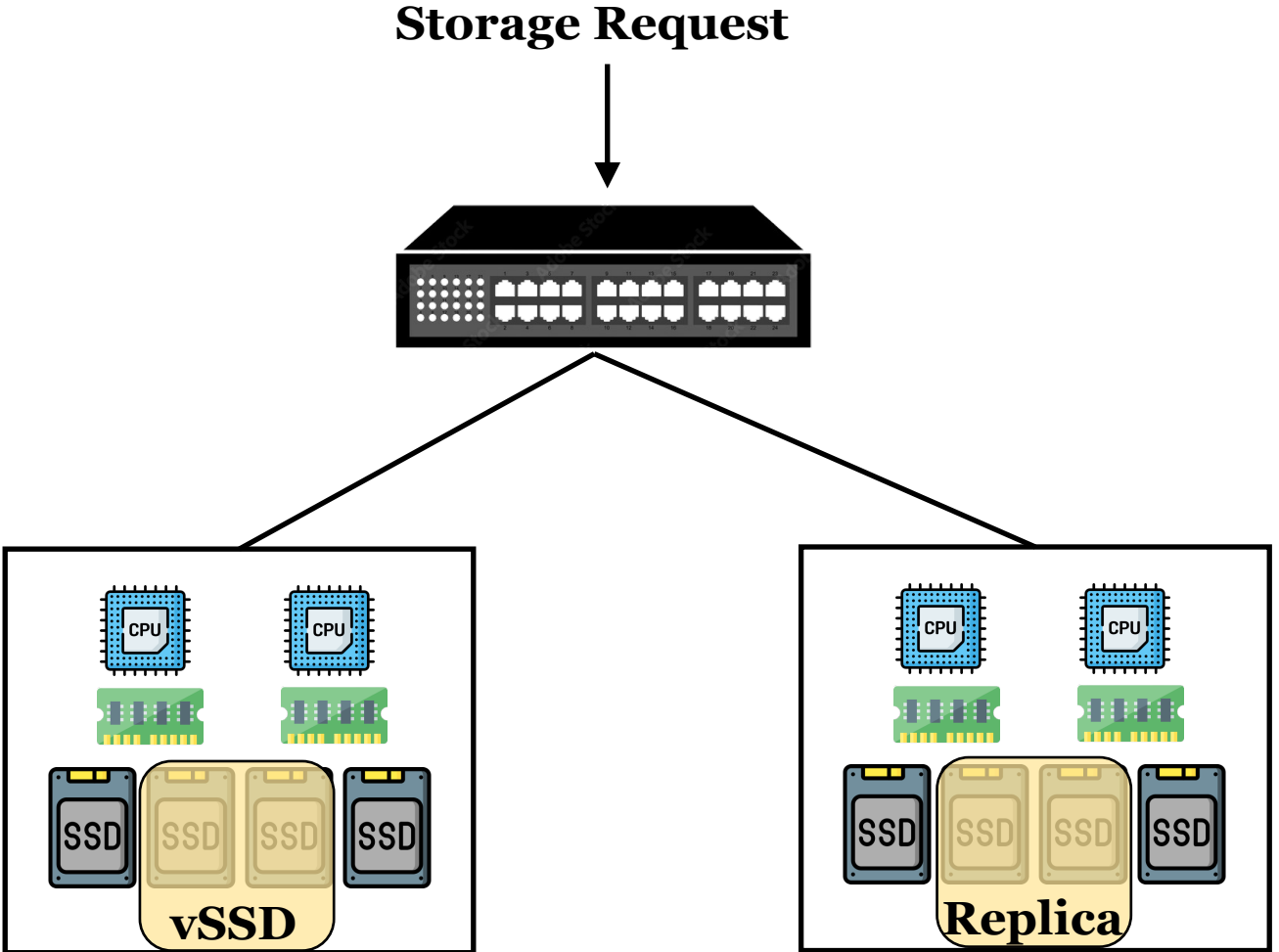
- In-Network Telemetry 
- Storage Queuing Delay 
- Sliding Window Predictor 



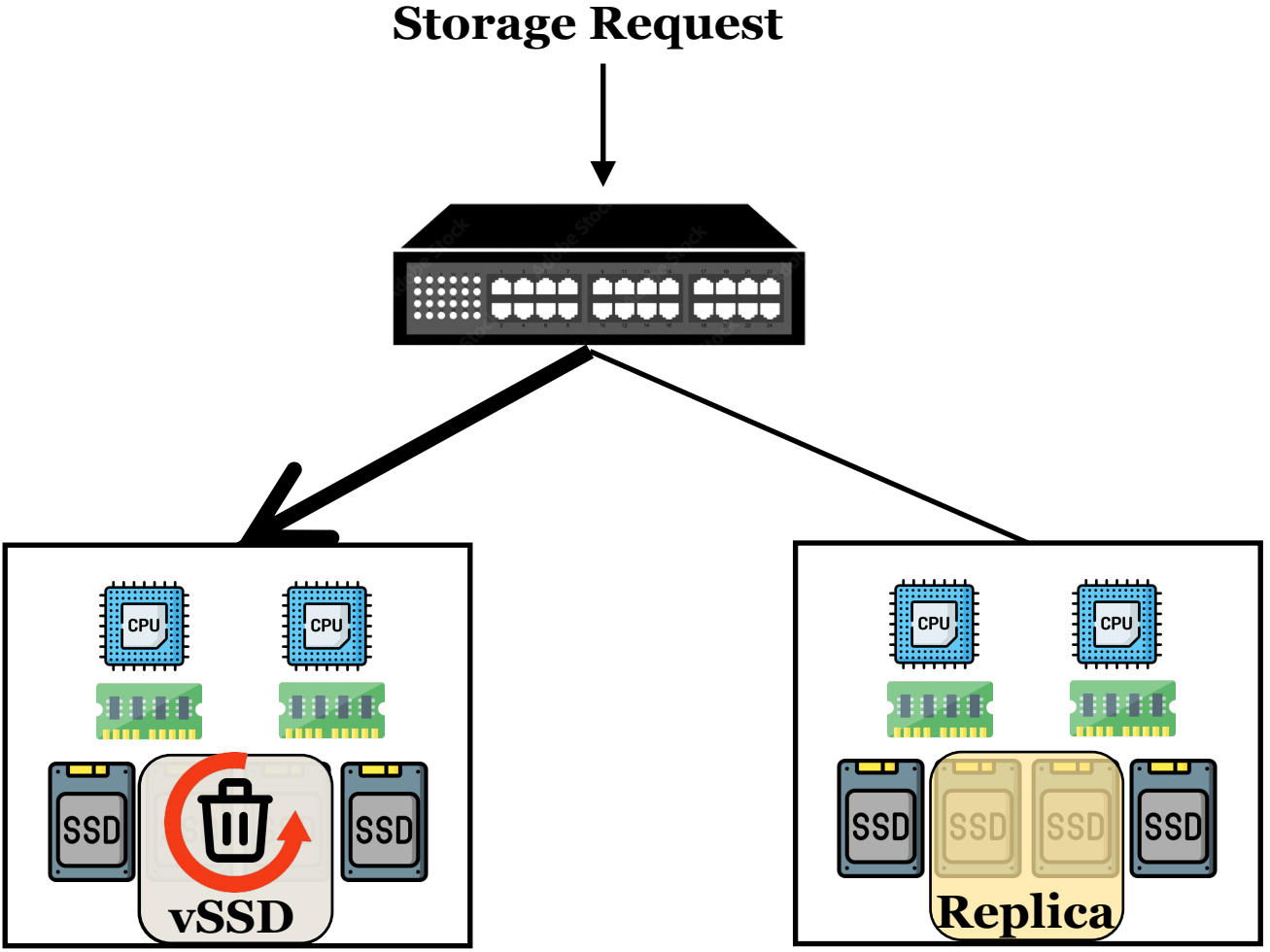
Enabling Coordinated Garbage Collection



Enabling Coordinated Garbage Collection

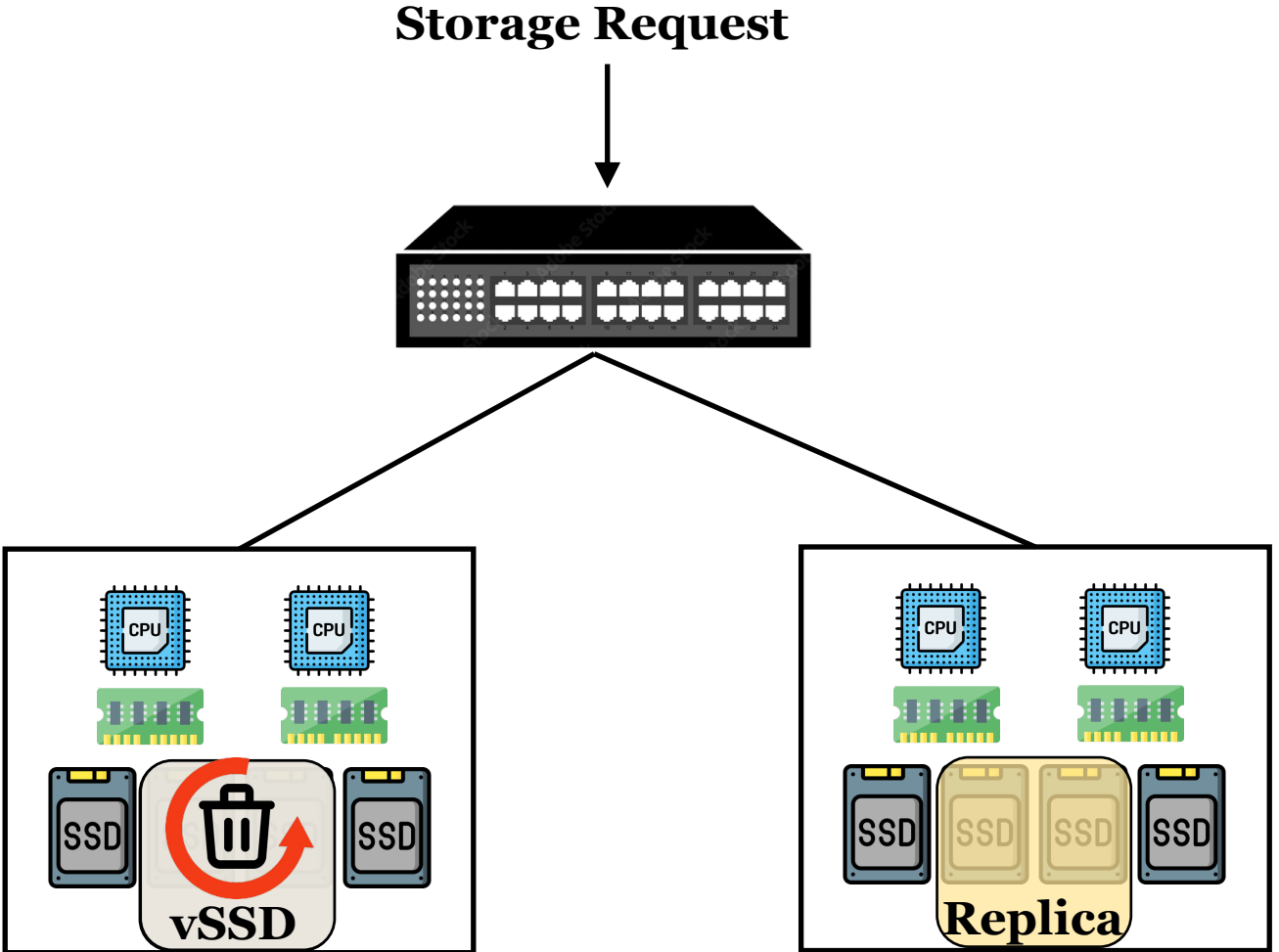


Enabling Coordinated Garbage Collection

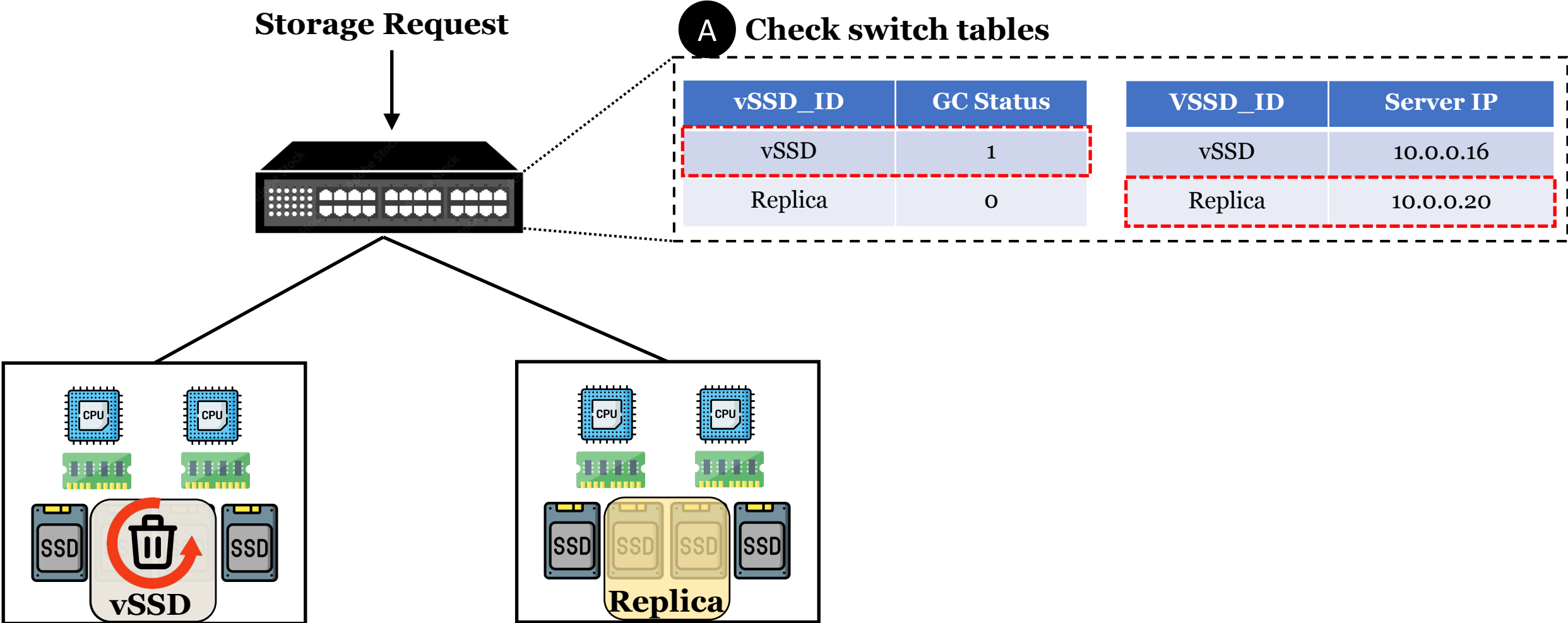



GC can add **significant storage delay!**

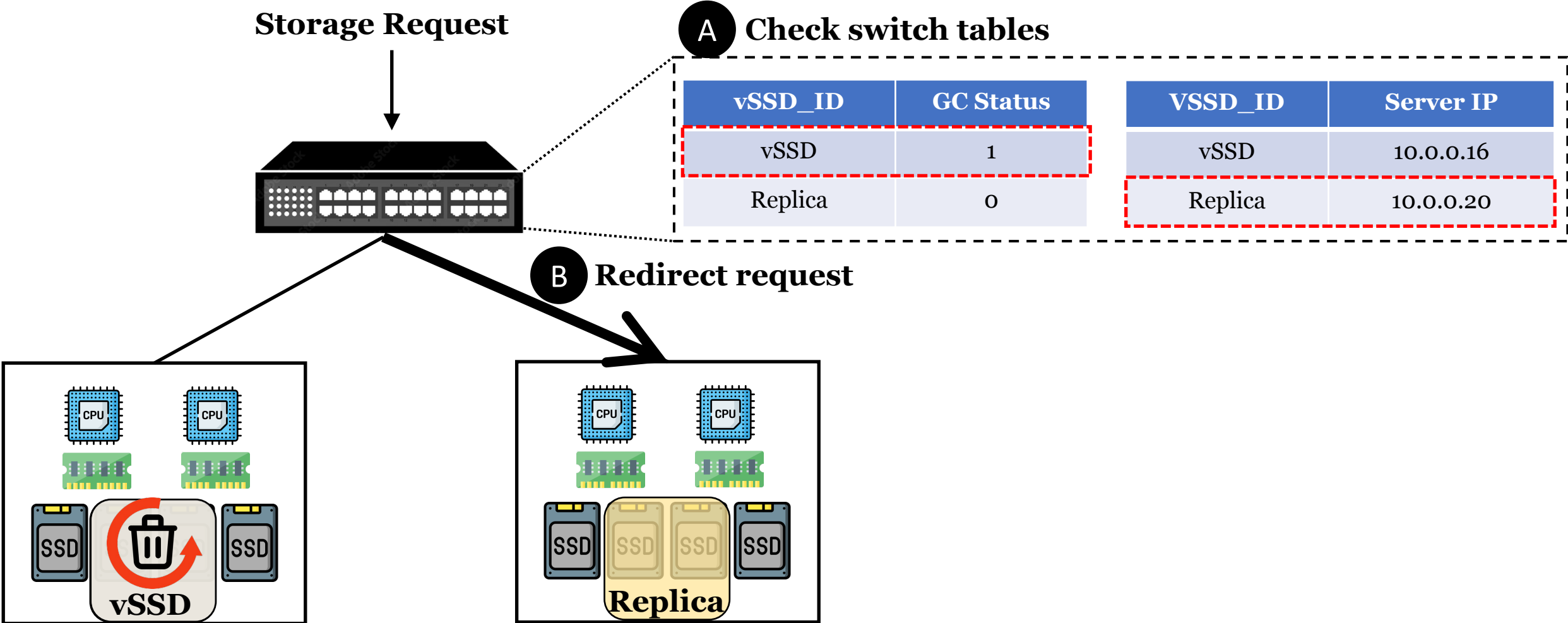
Enabling Coordinated Garbage Collection



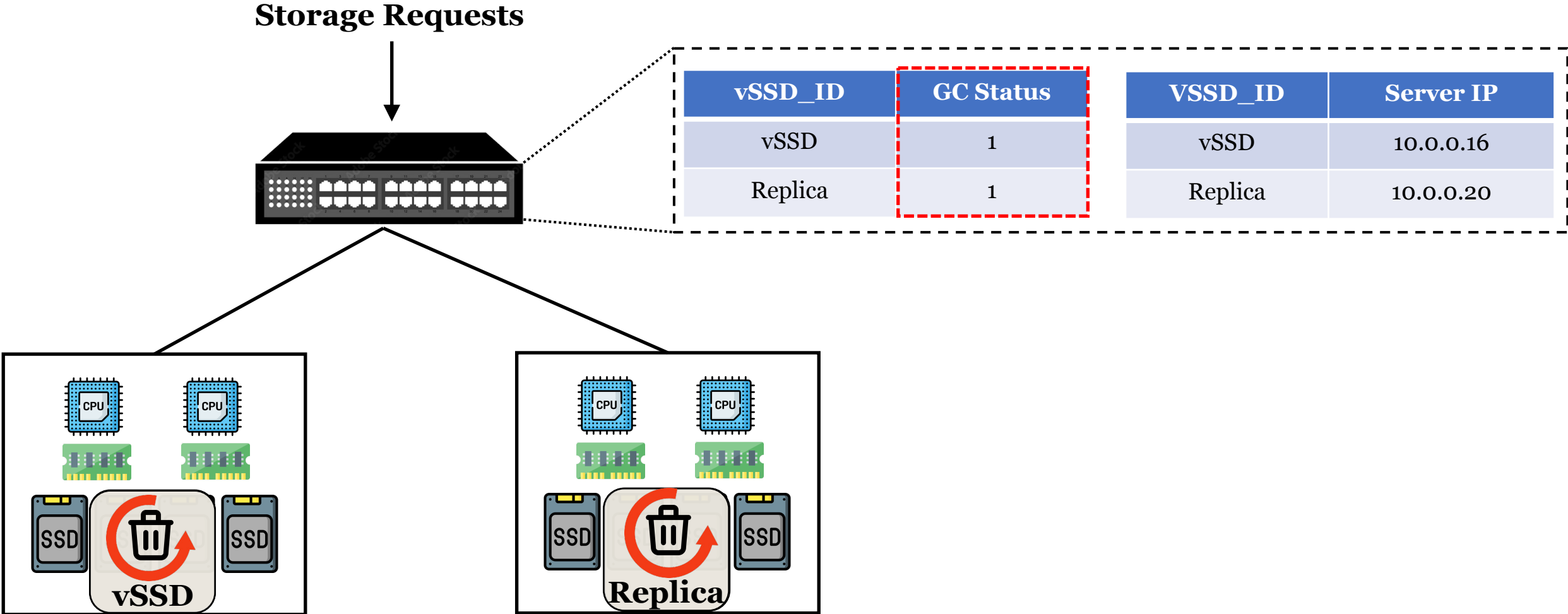
Enabling Coordinated Garbage Collection



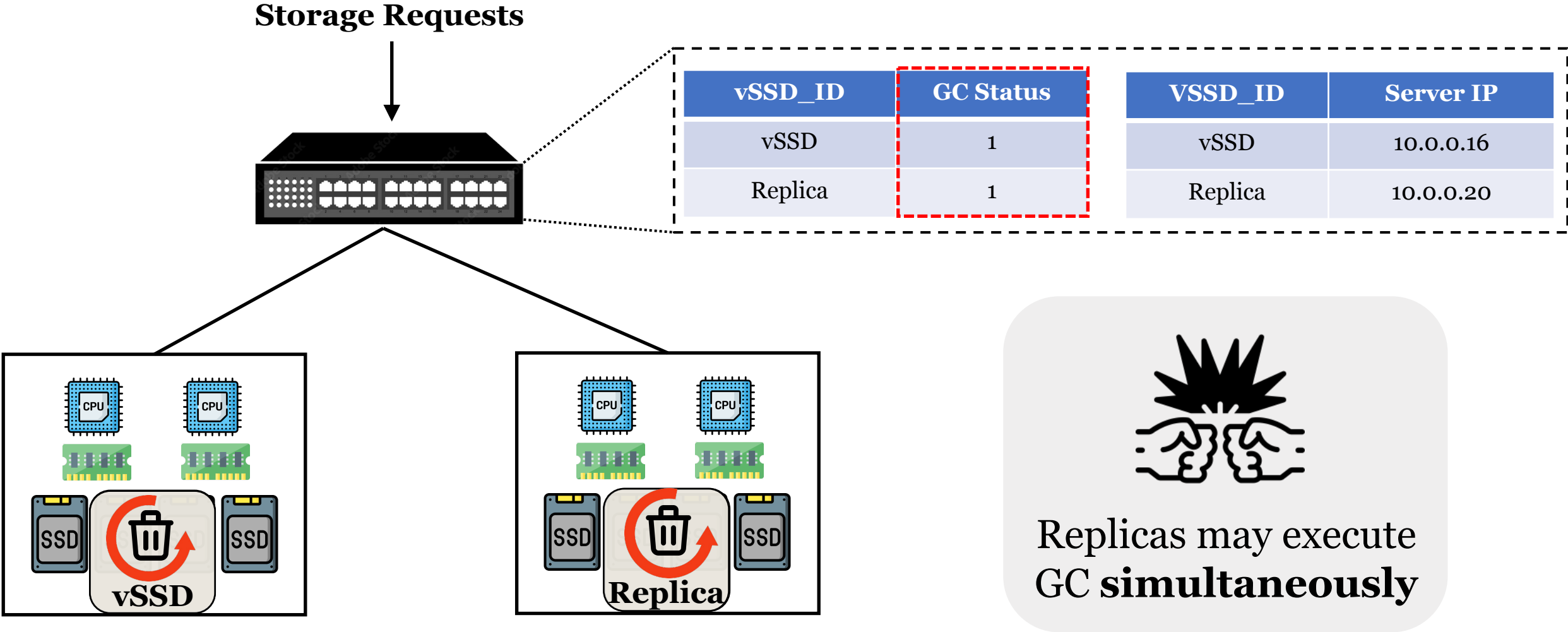
Enabling Coordinated Garbage Collection



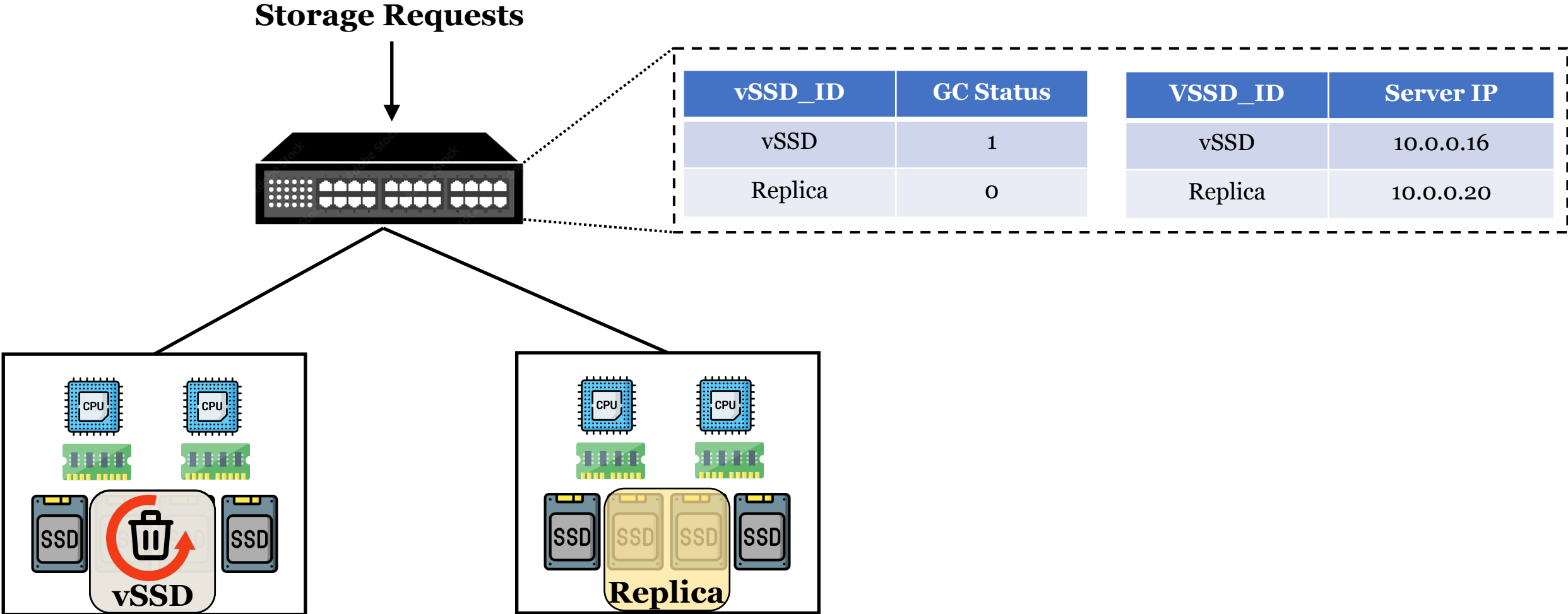
Enabling Coordinated Garbage Collection



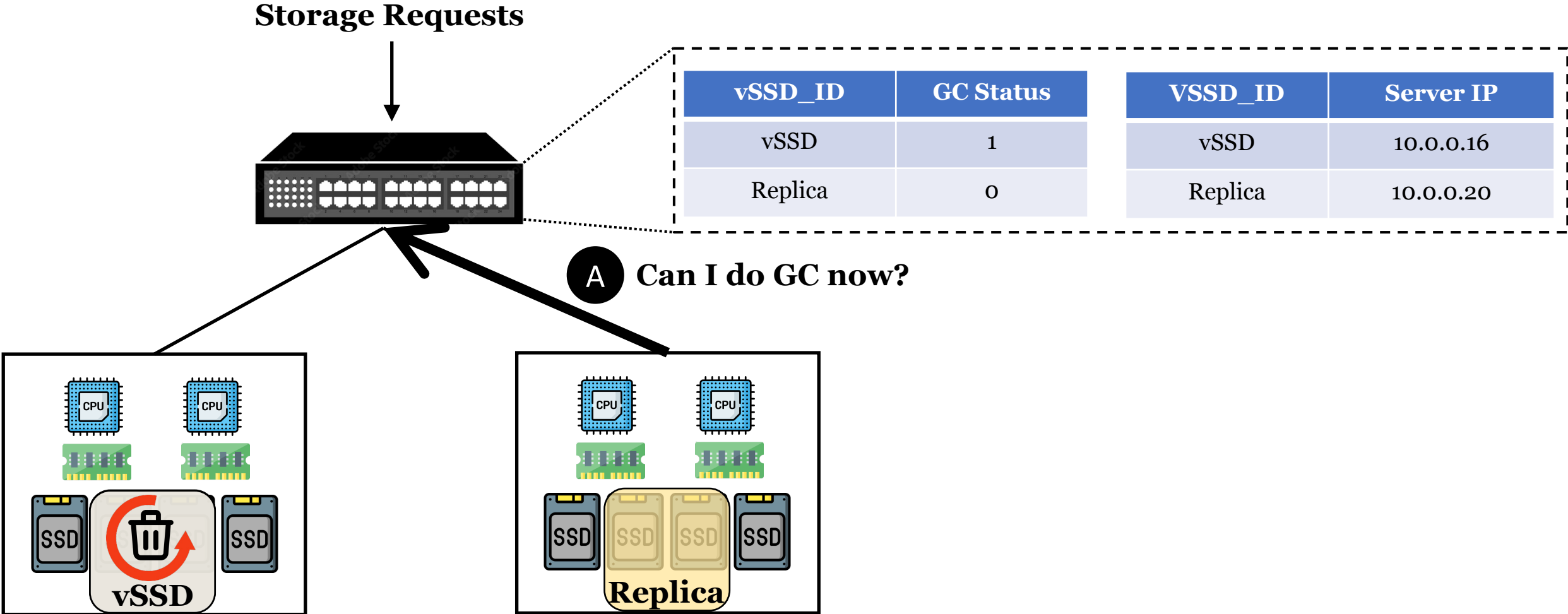
Enabling Coordinated Garbage Collection



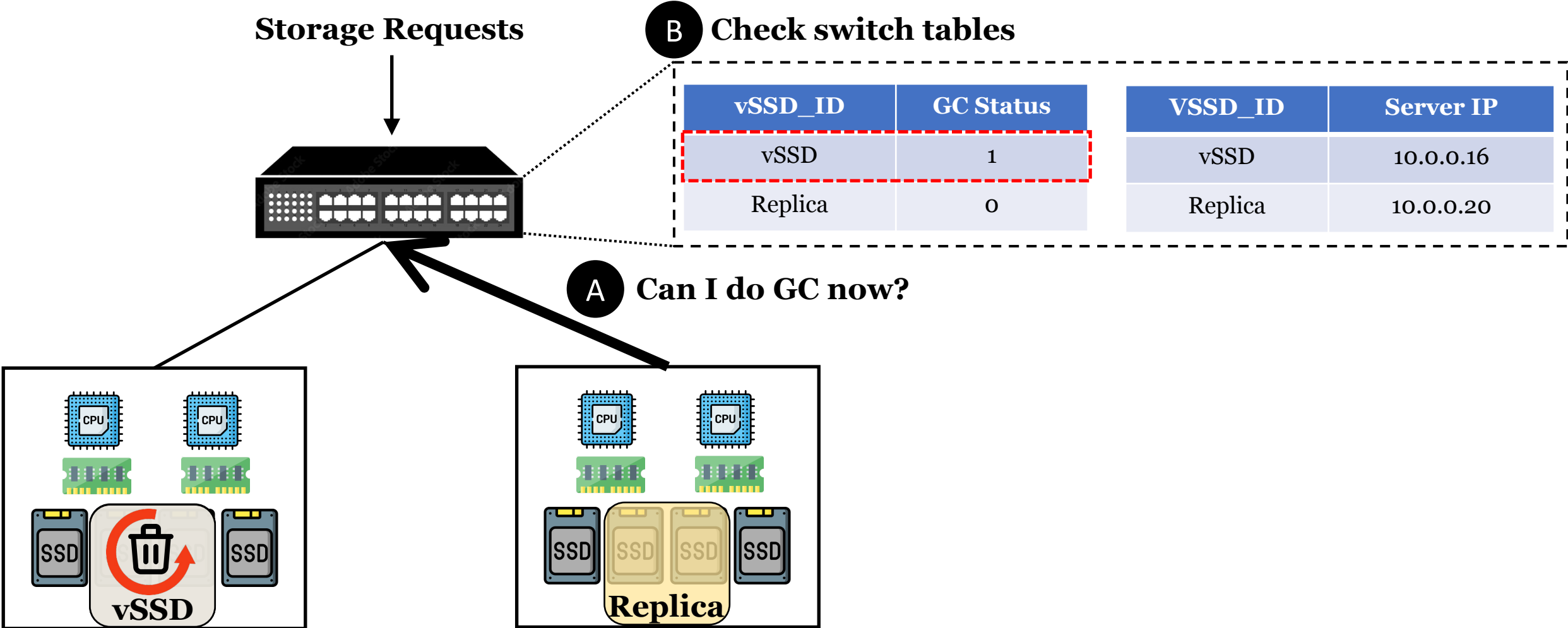
Enabling Coordinated Garbage Collection



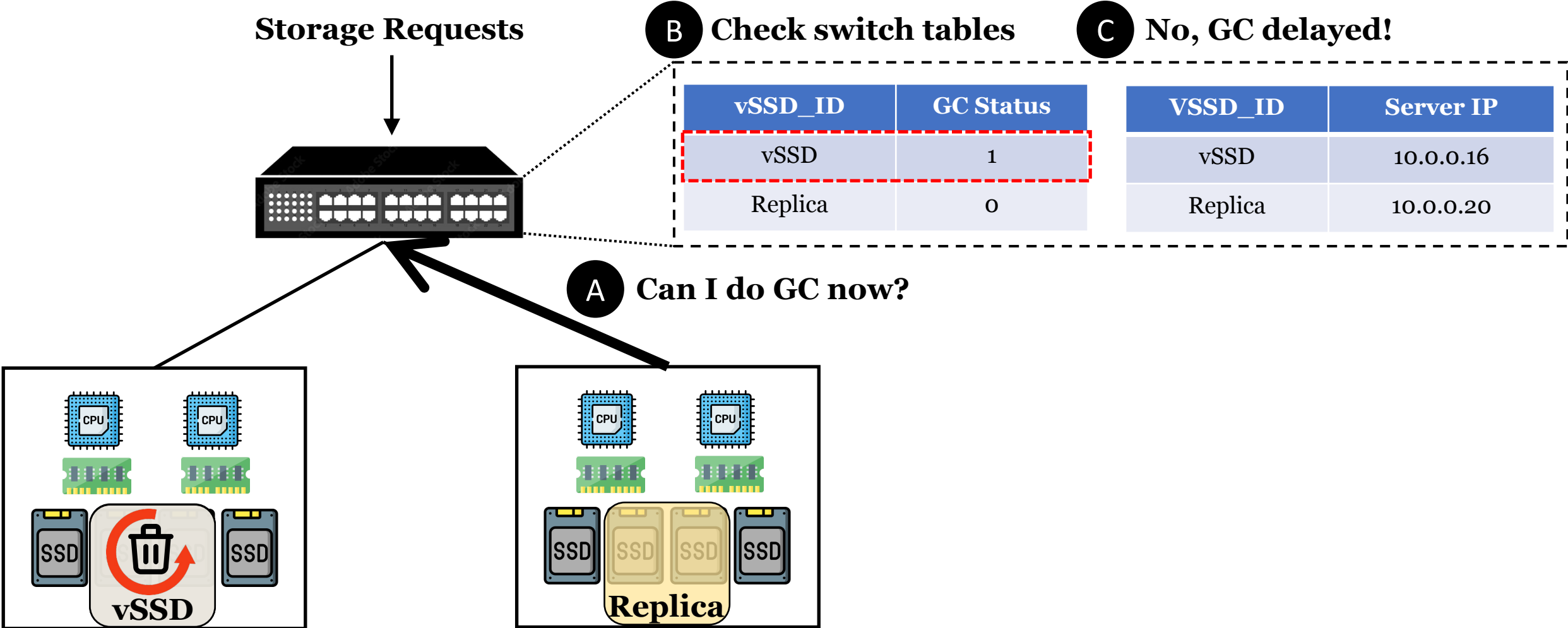
Enabling Coordinated Garbage Collection



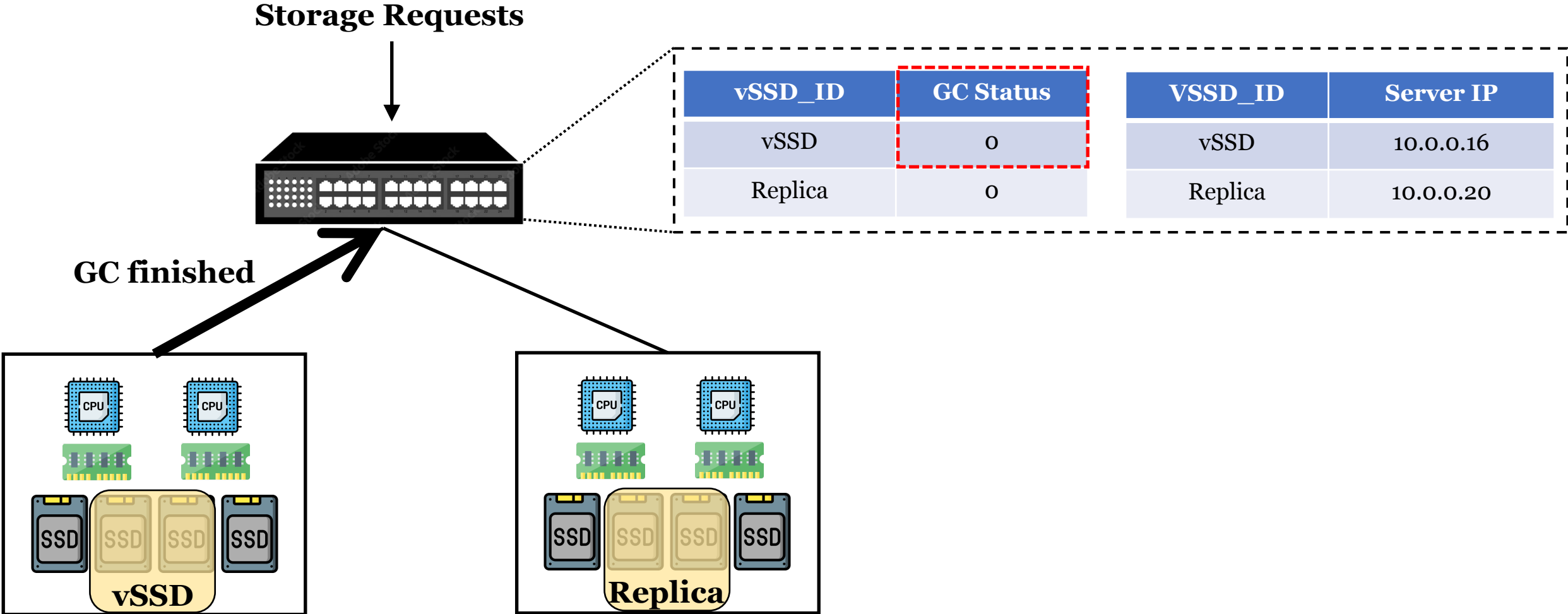
Enabling Coordinated Garbage Collection



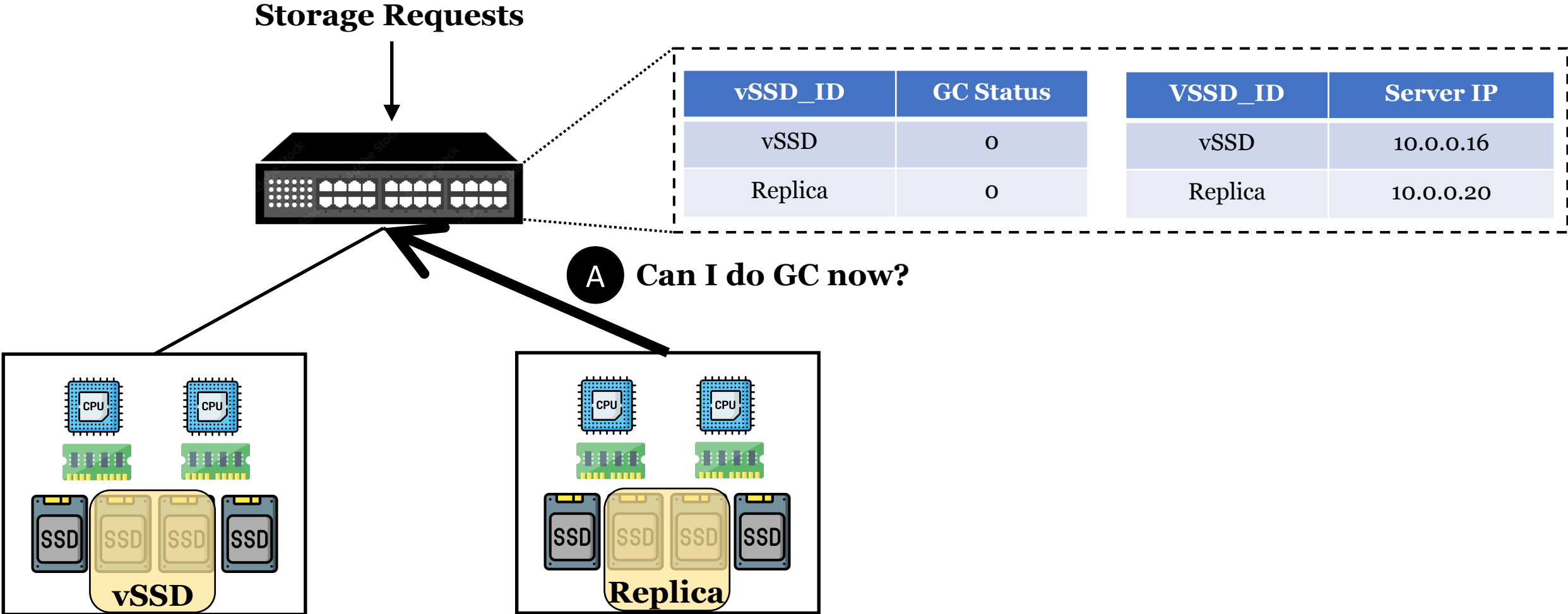
Enabling Coordinated Garbage Collection



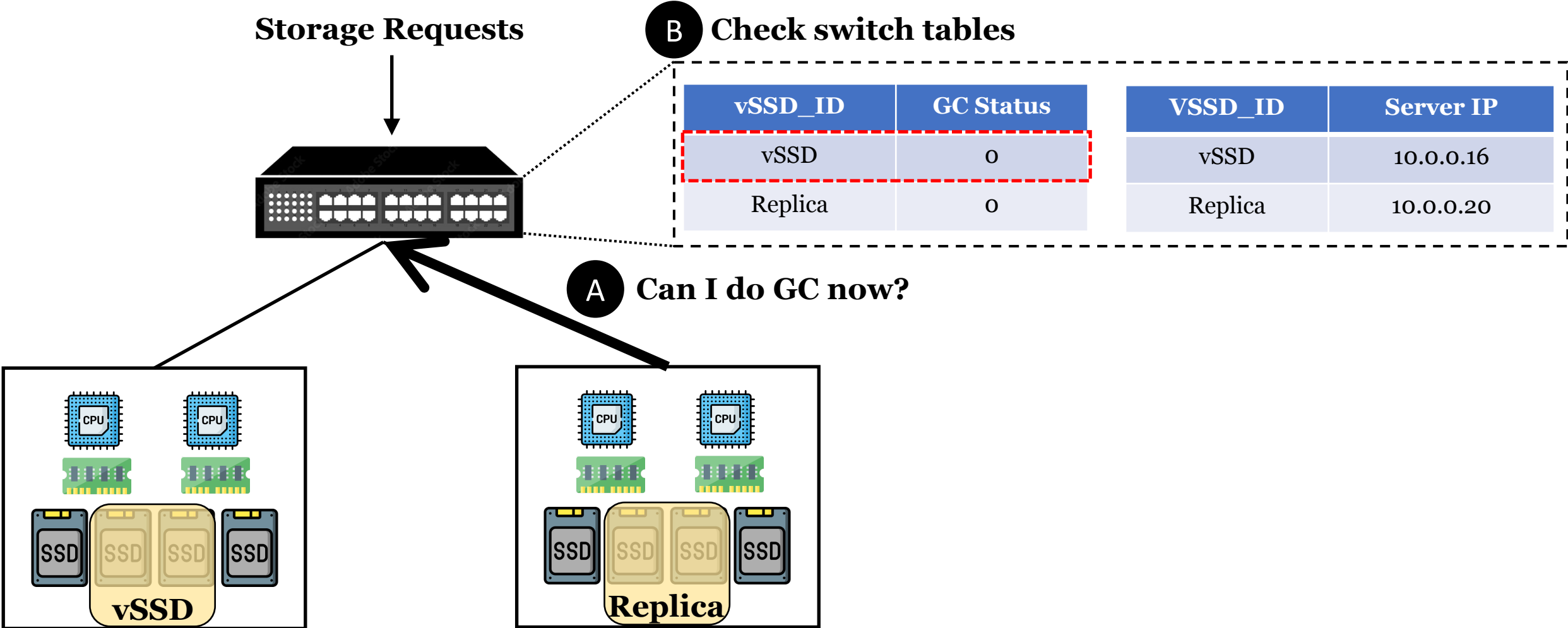
Enabling Coordinated Garbage Collection



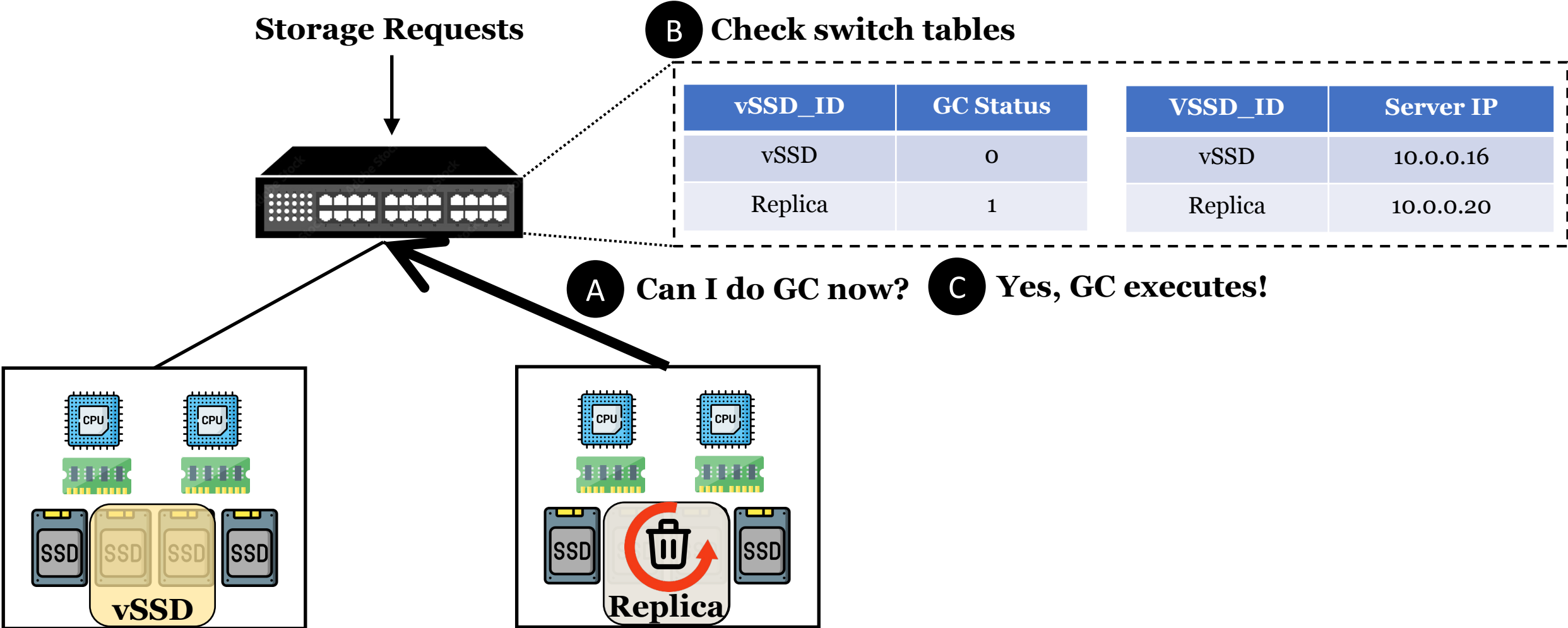
Enabling Coordinated Garbage Collection



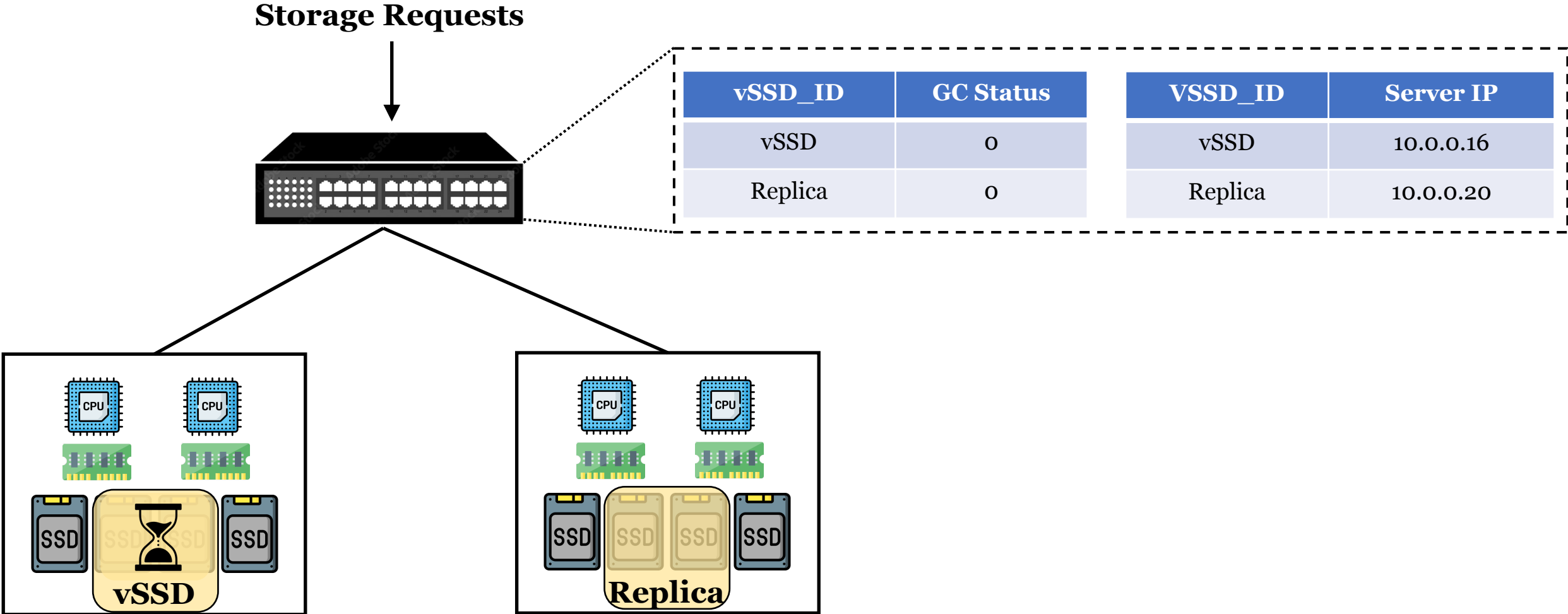
Enabling Coordinated Garbage Collection



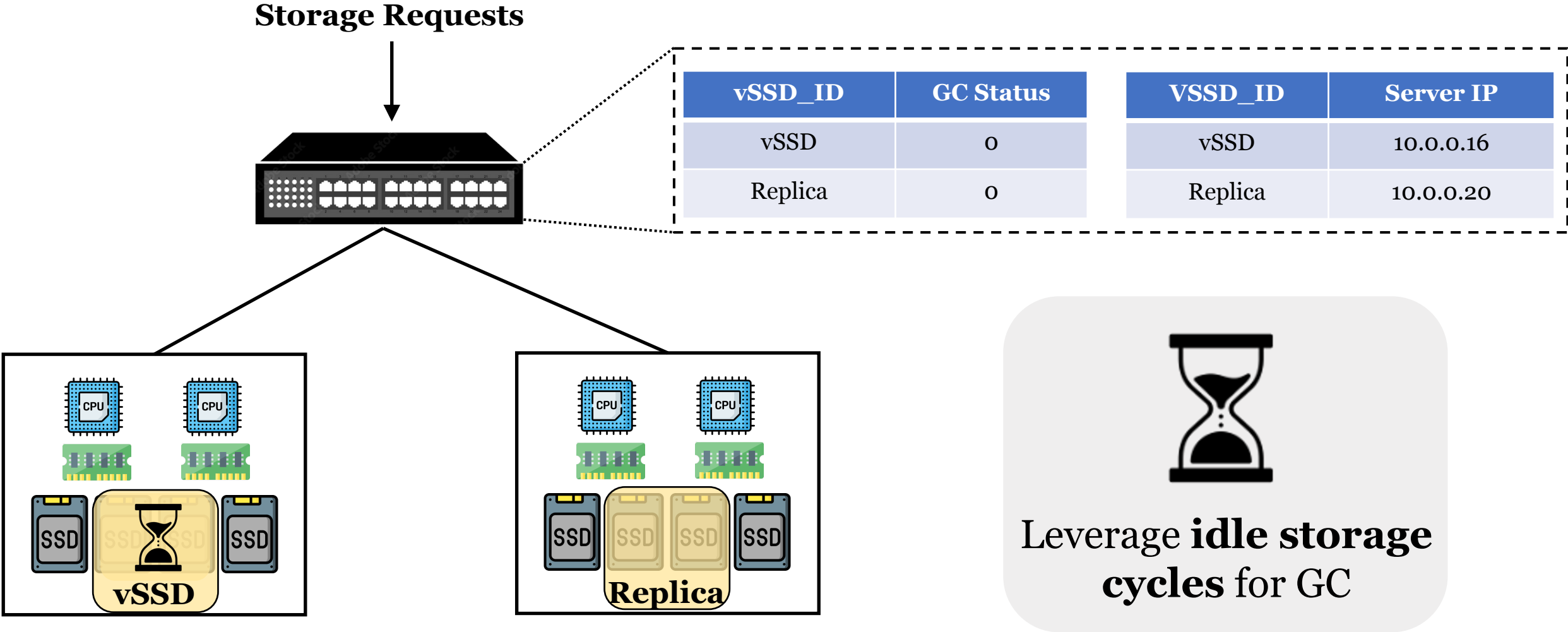
Enabling Coordinated Garbage Collection



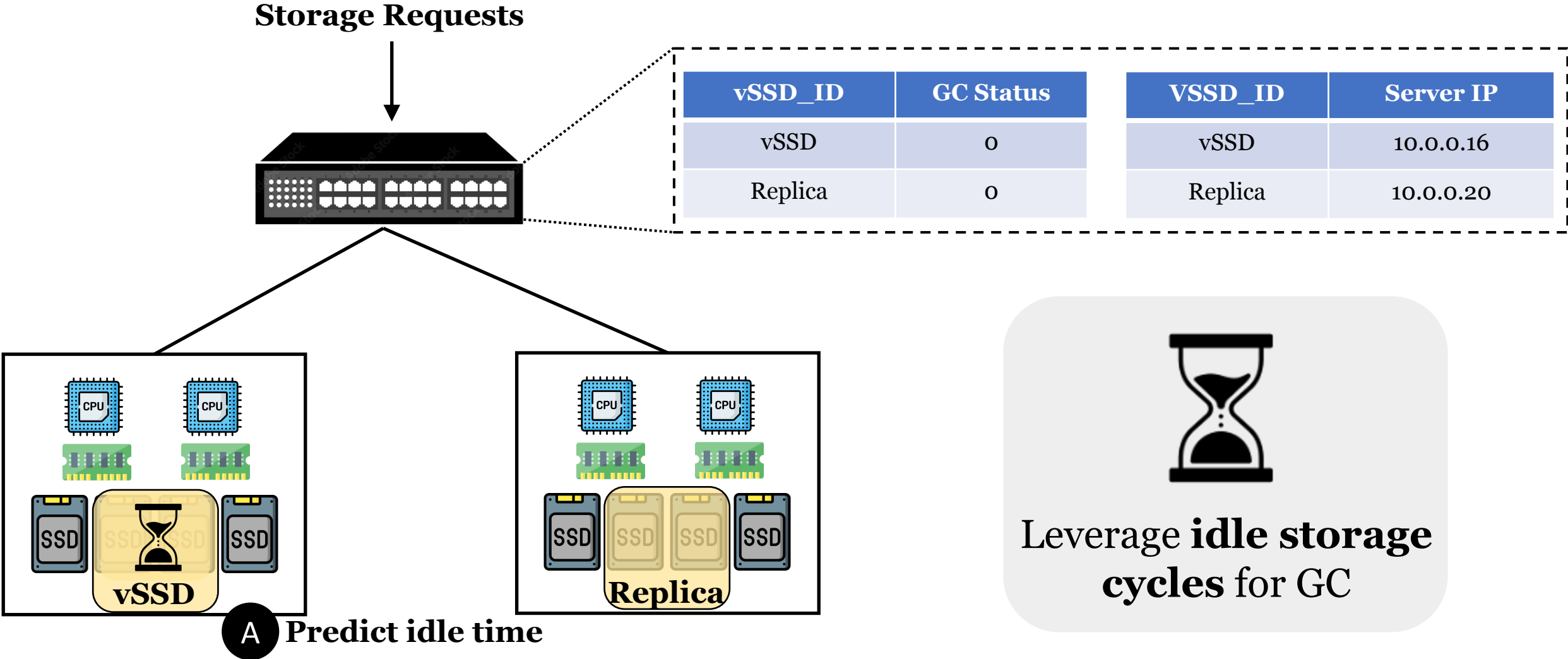
Enabling Coordinated Garbage Collection



Enabling Coordinated Garbage Collection

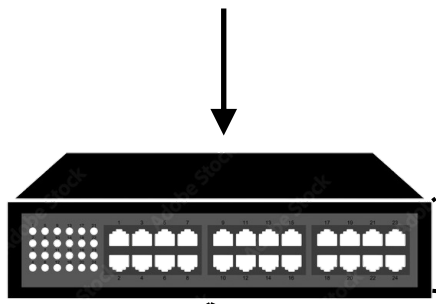


Enabling Coordinated Garbage Collection



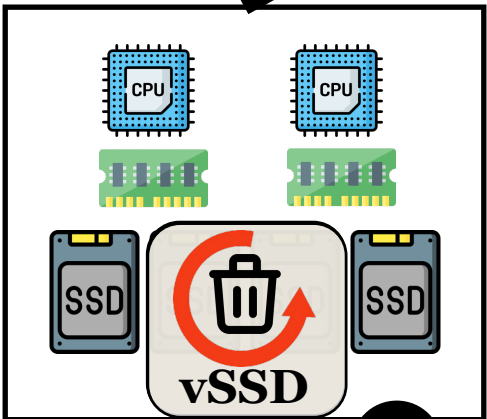
Enabling Coordinated Garbage Collection

Storage Requests


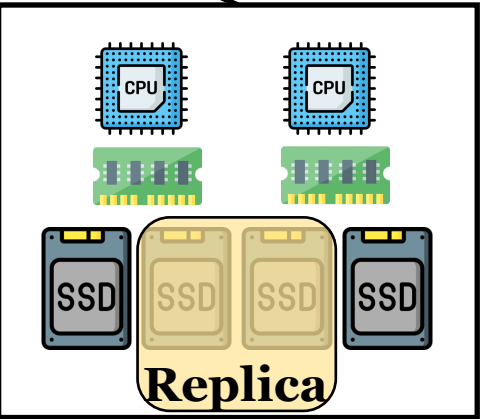


vSSD_ID	GC Status	VSSD_ID	Server IP
vSSD	0	vSSD	10.0.0.16
Replica	0	Replica	10.0.0.20

B Notify Switch



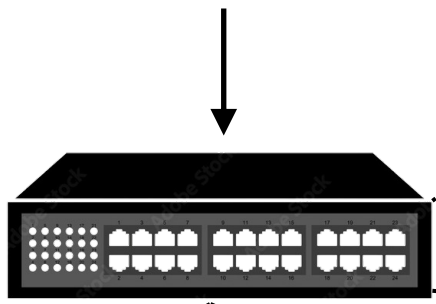
A Predict idle time



Leverage **idle storage cycles** for GC

Enabling Coordinated Garbage Collection

Storage Requests

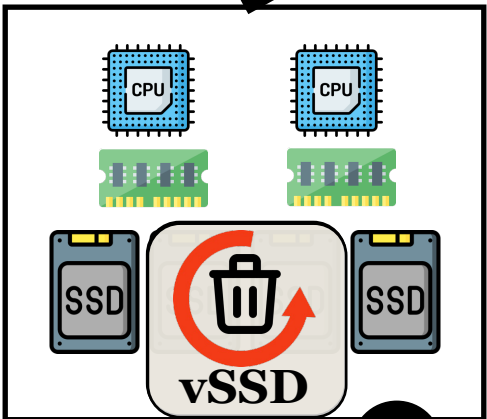


C Update switch tables


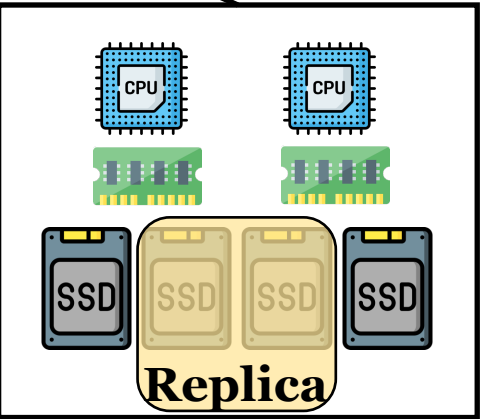
vSSD_ID	GC Status
vSSD	1
Replica	0

VSSD_ID	Server IP
vSSD	10.0.0.16
Replica	10.0.0.20

B Notify Switch



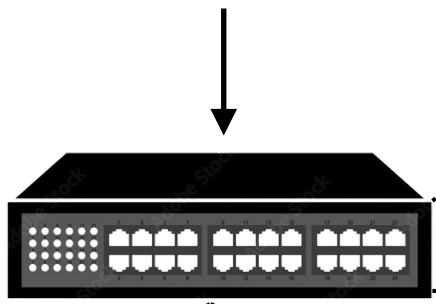
A Predict idle time



Leverage **idle storage cycles** for GC

Enabling Coordinated Garbage Collection

Storage Requests

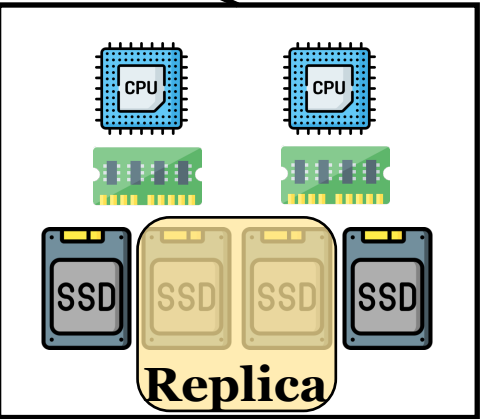
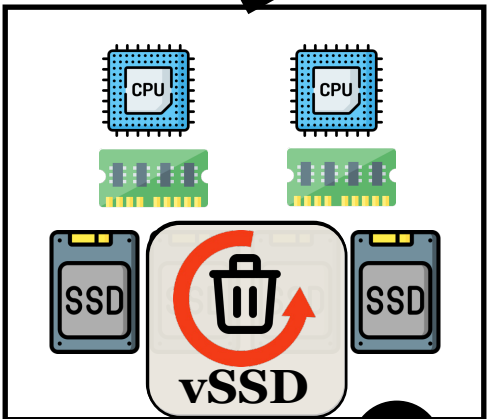


C Update switch tables

vSSD_ID	GC Status
vSSD	1
Replica	0

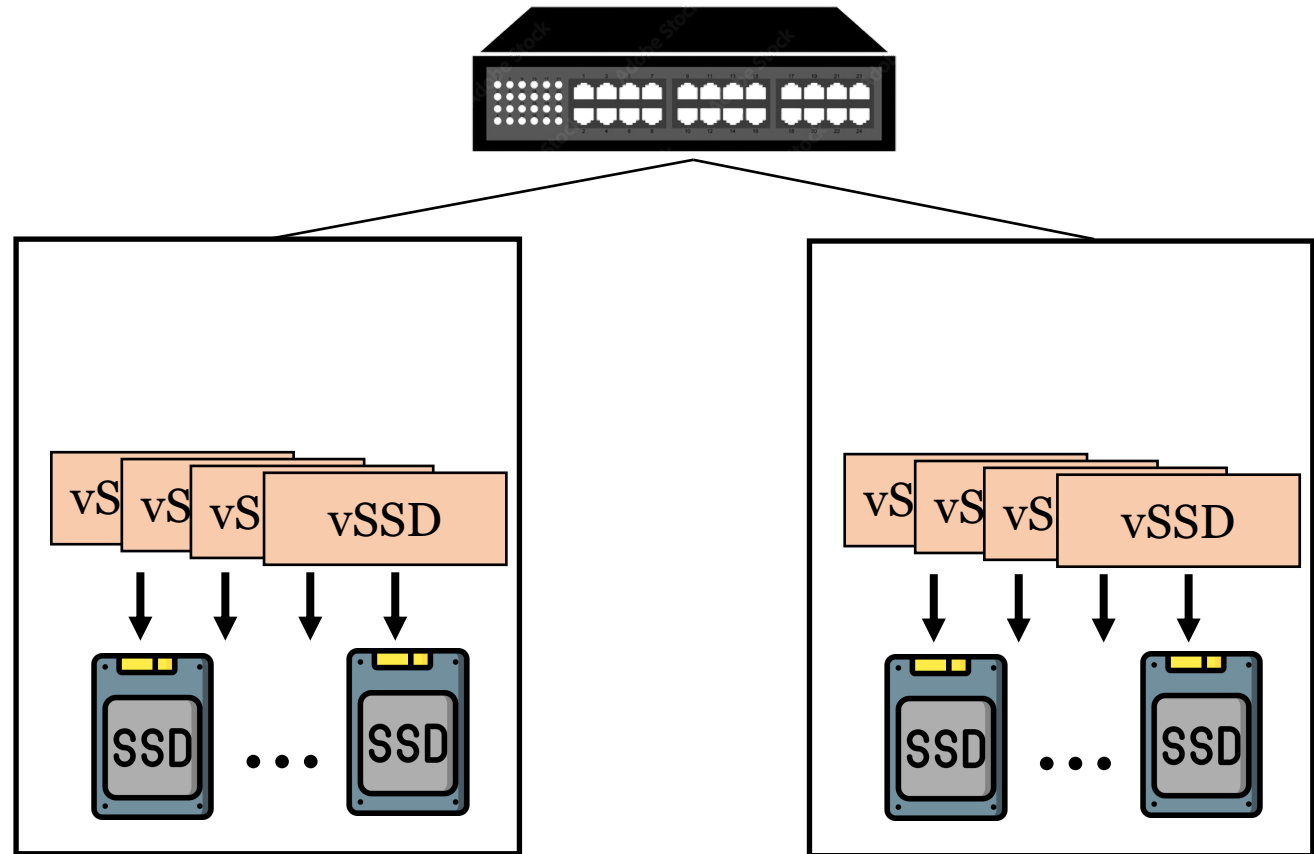
VSSD_ID	Server IP
vSSD	10.0.0.16
Replica	10.0.0.20

B Notify Switch

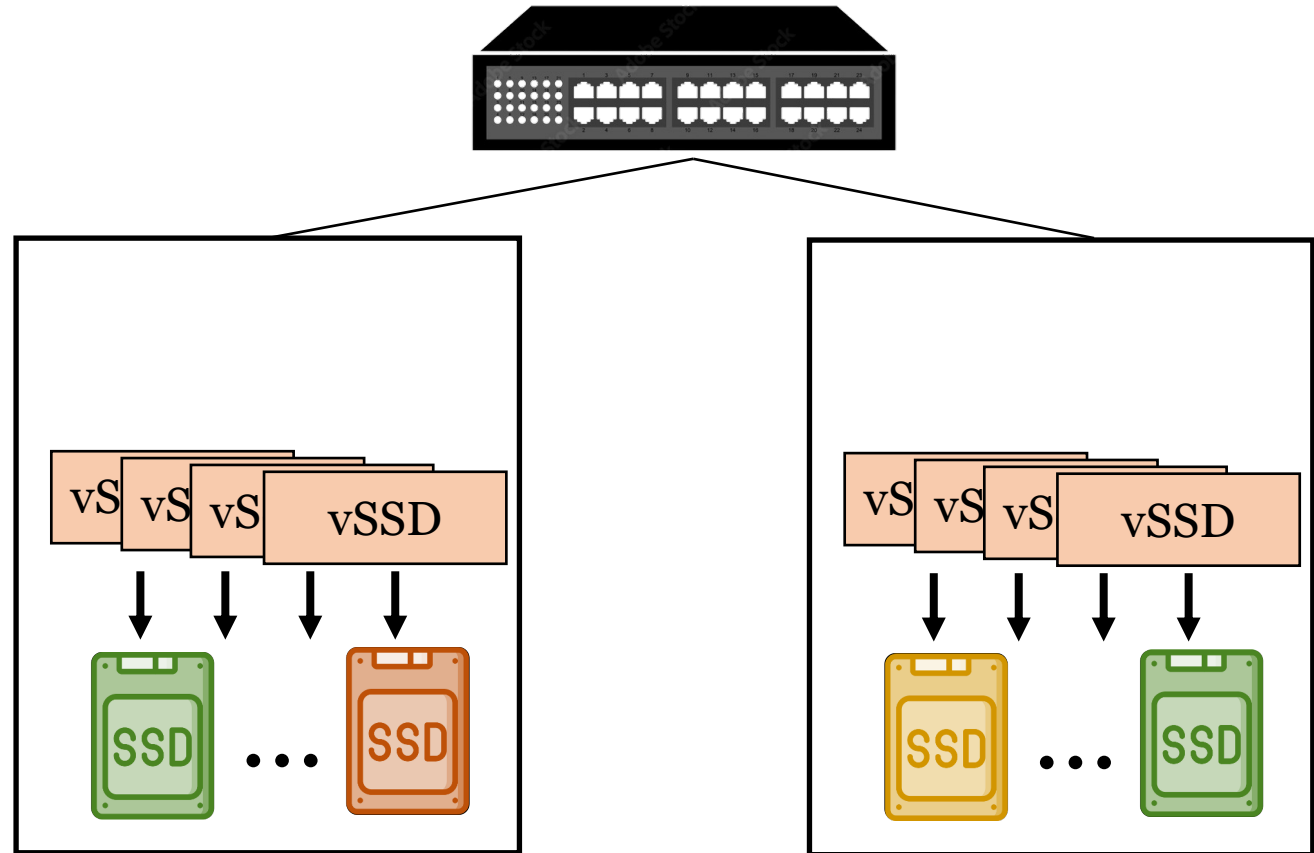


A Predict idle time

Enabling Rack-Scale Wear Leveling



Enabling Rack-Scale Wear Leveling

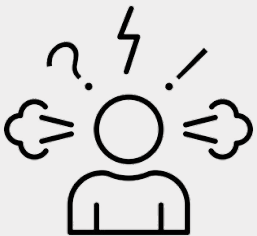


Workload instances have **diverse write intensity!**

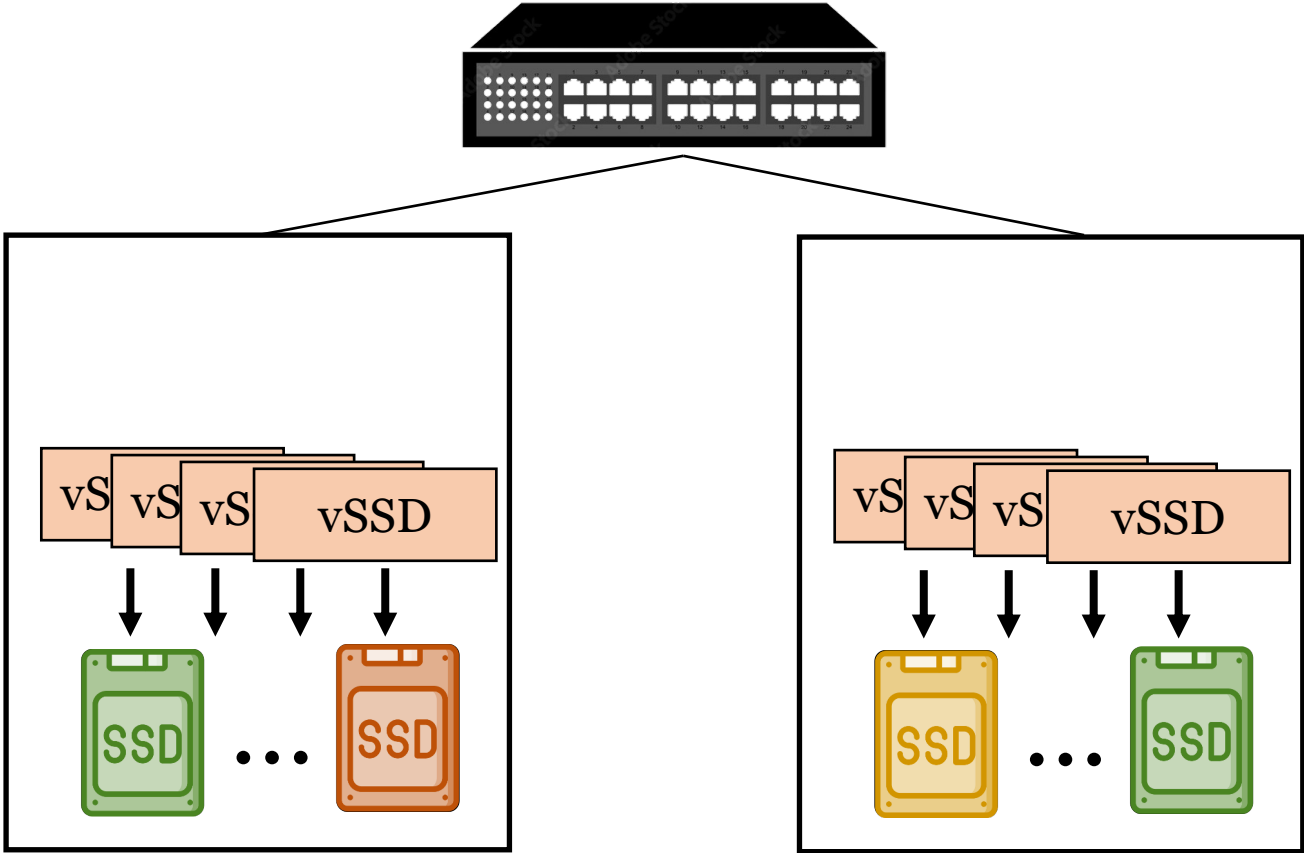
Enabling Rack-Scale Wear Leveling



Increased cost

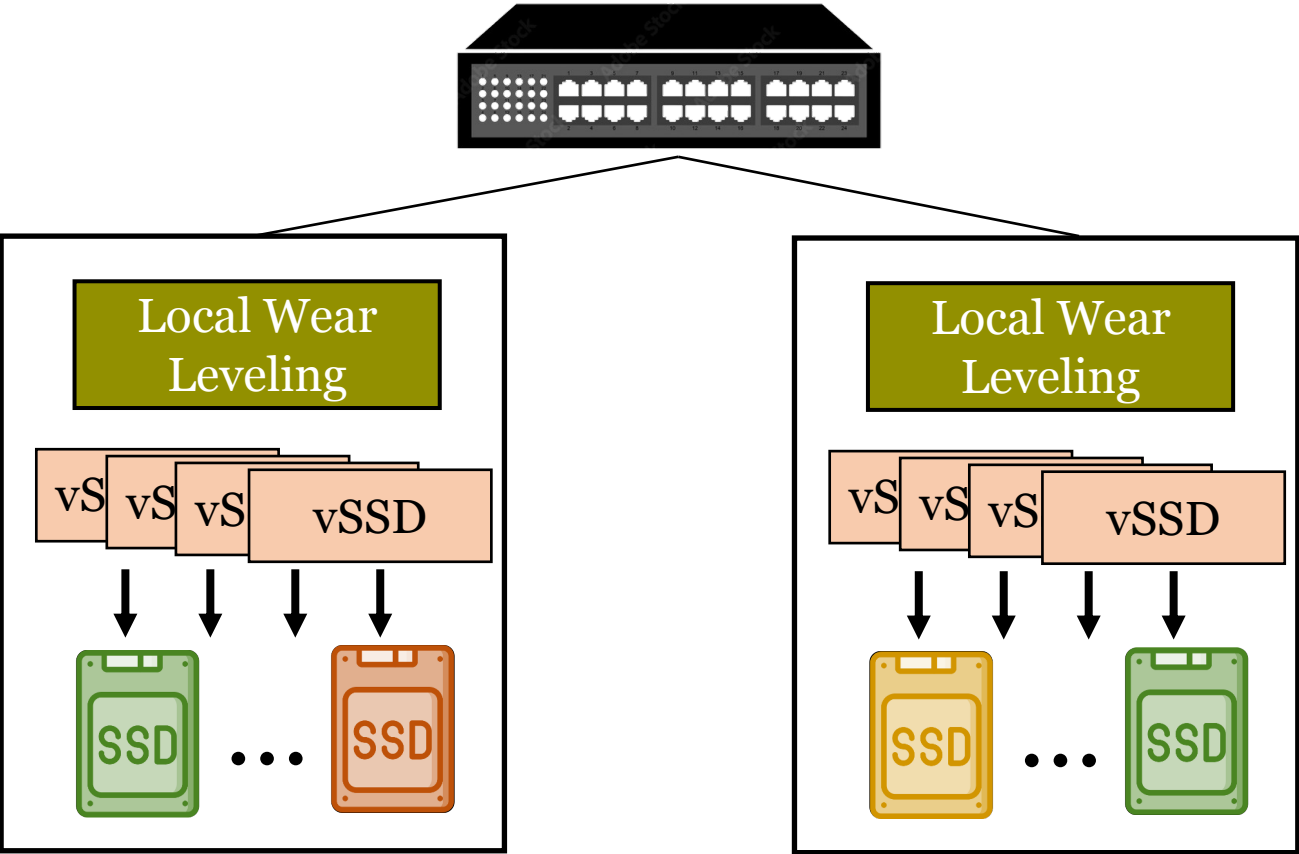


Increased management complexity

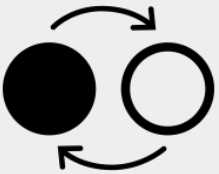


Workload instances have **diverse write intensity!**

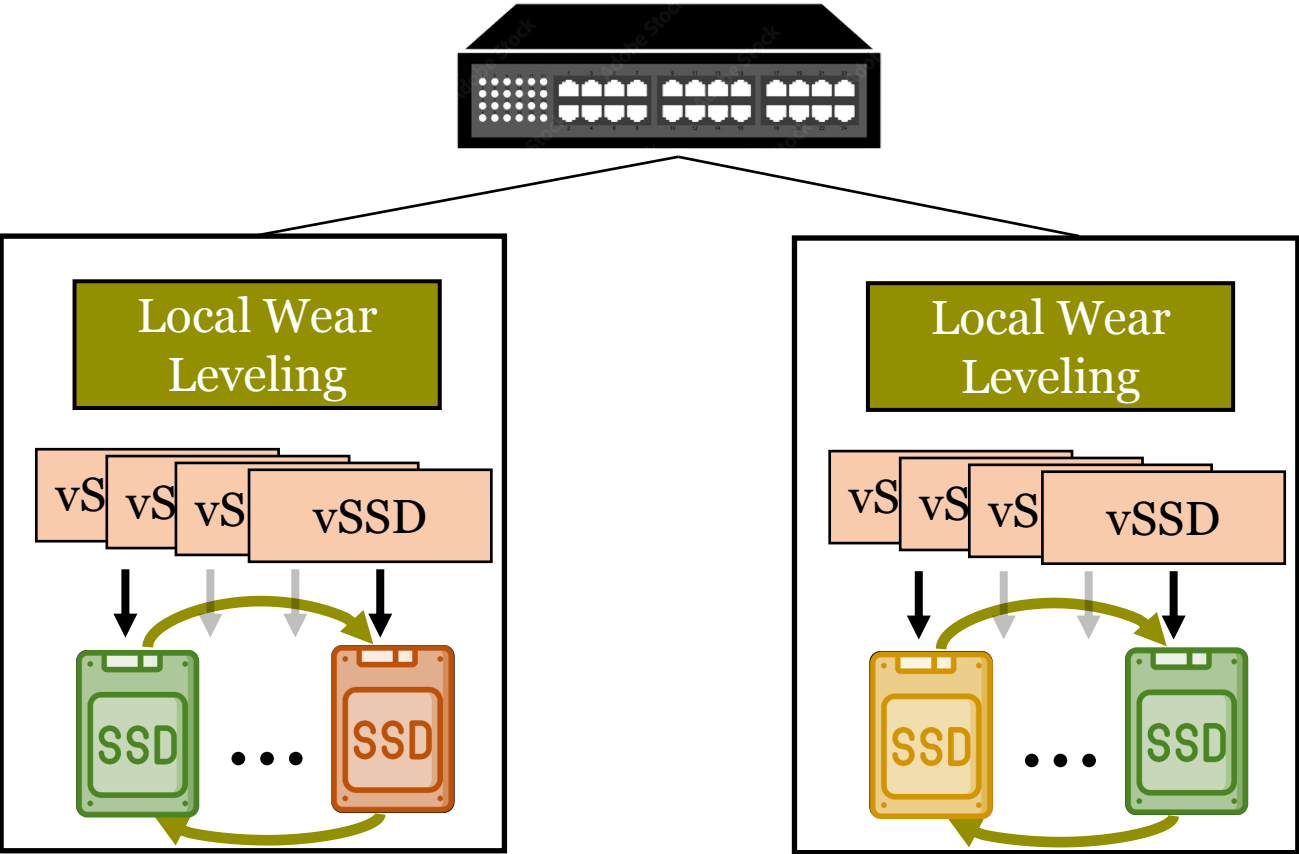
Enabling Rack-Scale Wear Leveling



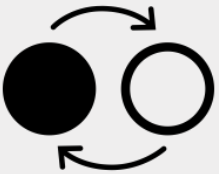
Enabling Rack-Scale Wear Leveling



Swap SSDs **within** servers **every 16 days**

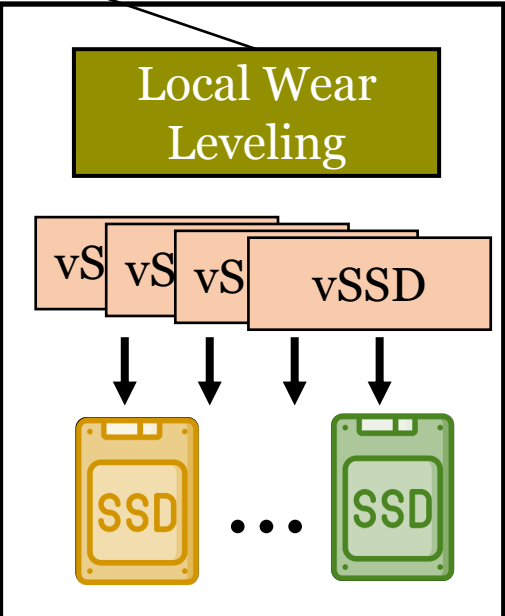
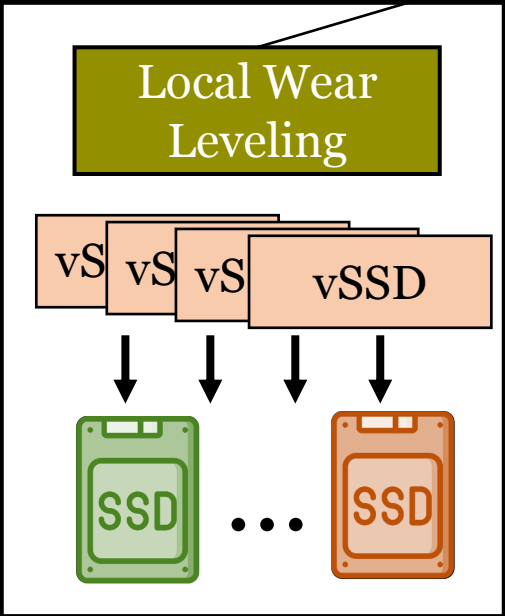


Enabling Rack-Scale Wear Leveling

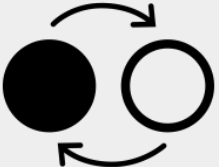


Swap SSDs **within** servers **every 16 days**

Global Wear Leveling



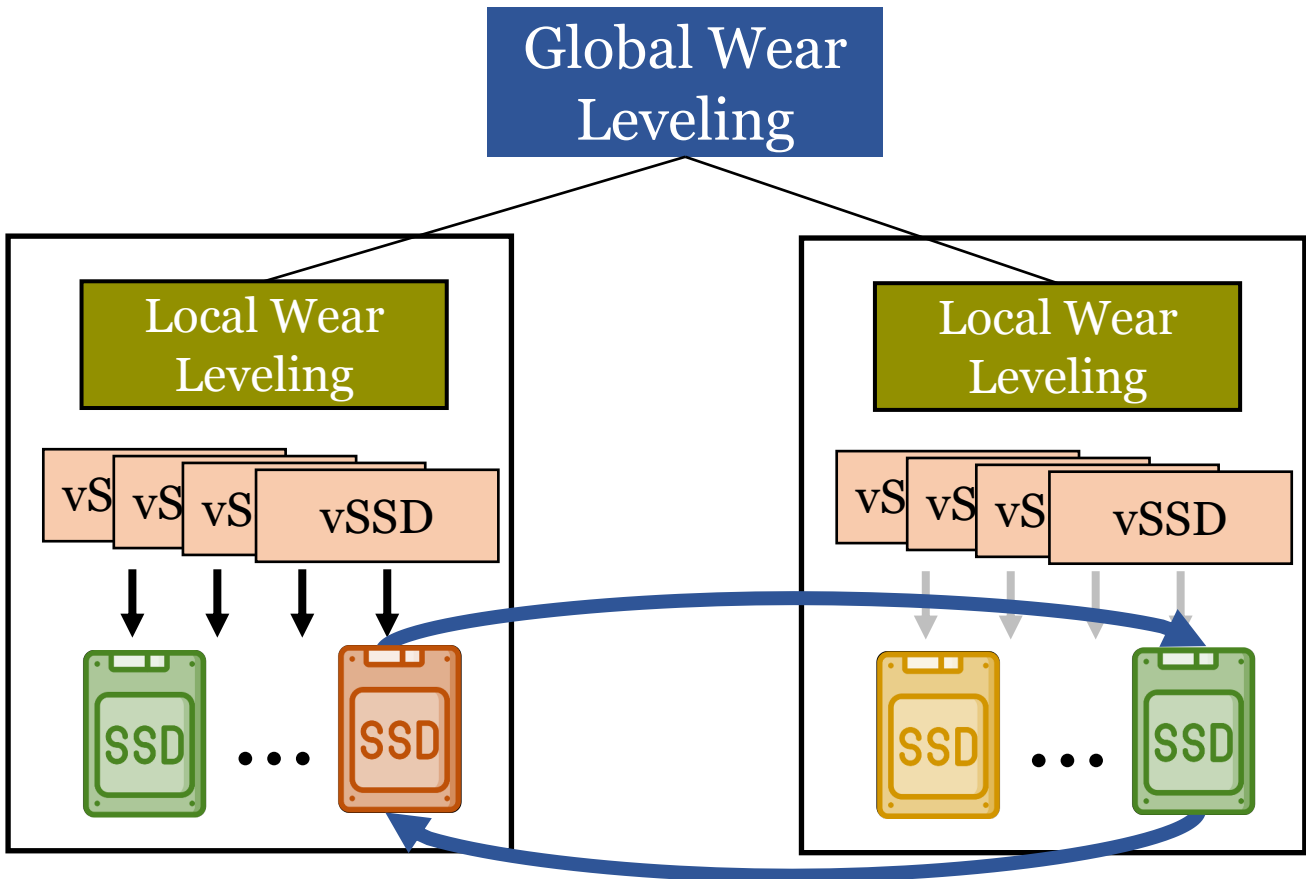
Enabling Rack-Scale Wear Leveling



Swap SSDs **within** servers **every 16 days**



Swap SSDs **across** servers **every 8 weeks**



RackBlox Implementation

Programmable SSDs

1 TB

16 Channels

16 KB/page

70 MB/s per channel

RackBlox Implementation

Programmable SSDs

1 TB
16 Channels
16 KB/page
70 MB/s per channel

Programmable Switch

Intel Tofino Switch

RackBlox Implementation

Programmable SSDs

1 TB
16 Channels
16 KB/page
70 MB/s per channel

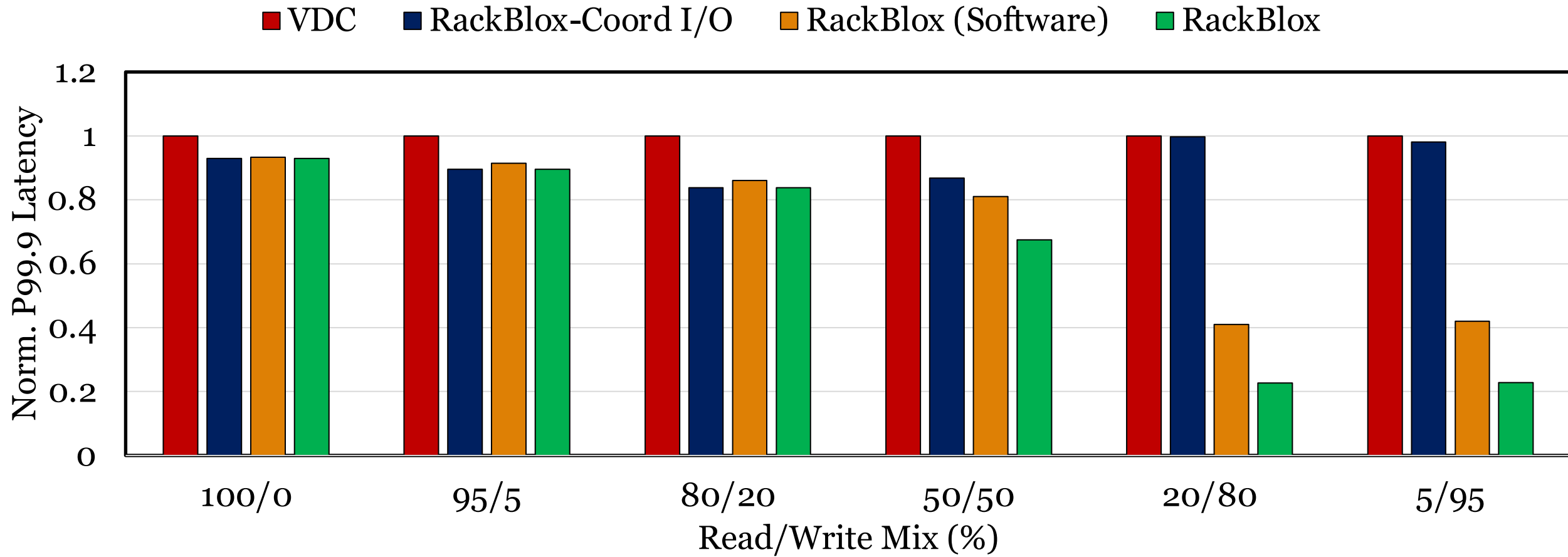
Programmable Switch

Intel Tofino Switch

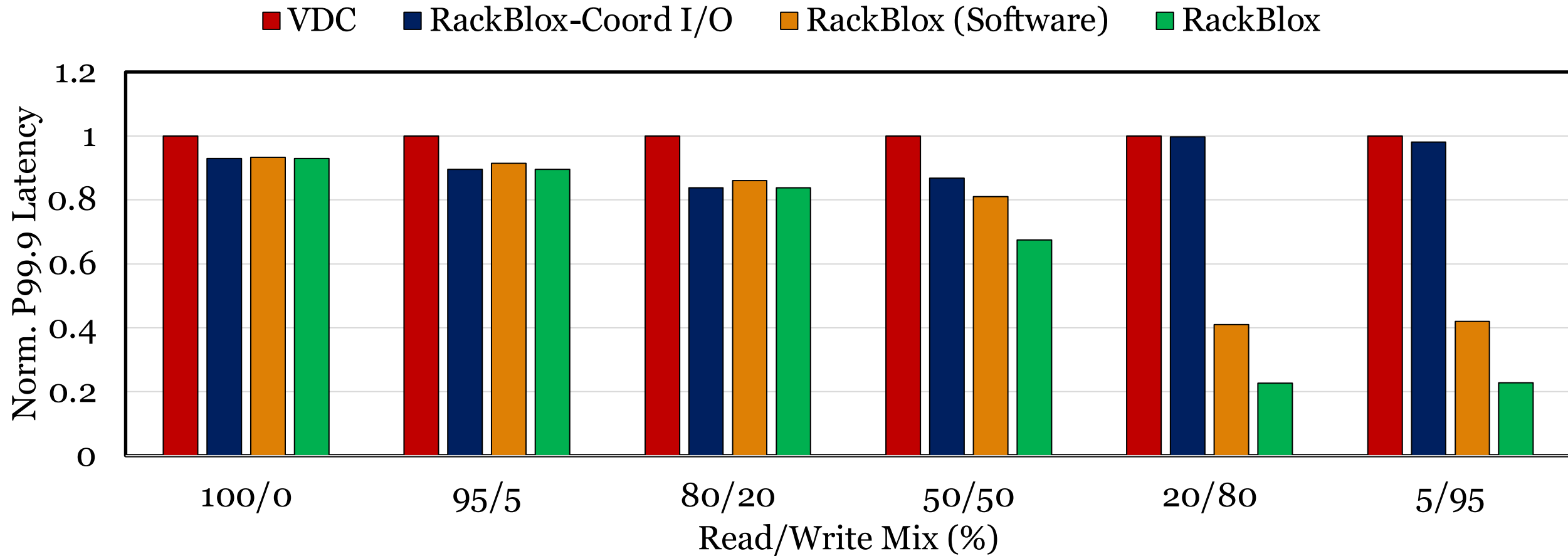
Workloads

YCSB
TPC-H
TPC-C
AuctionMark

Improving End-To-End I/O Performance

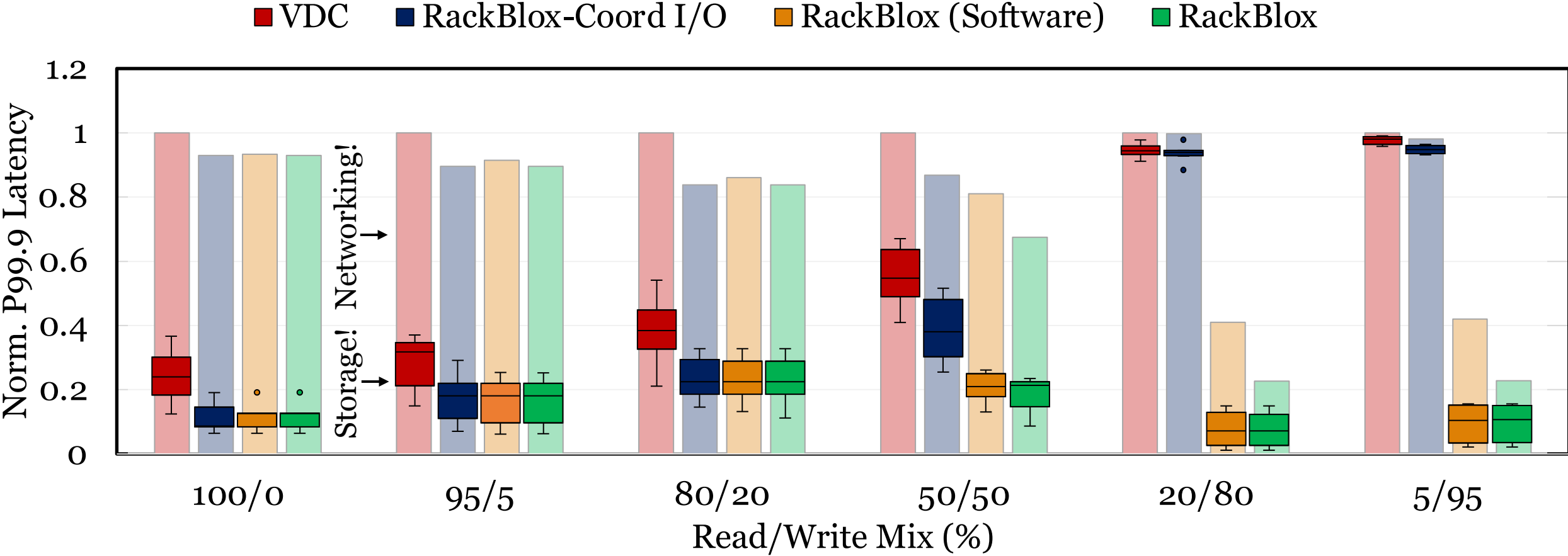


Improving End-To-End I/O Performance



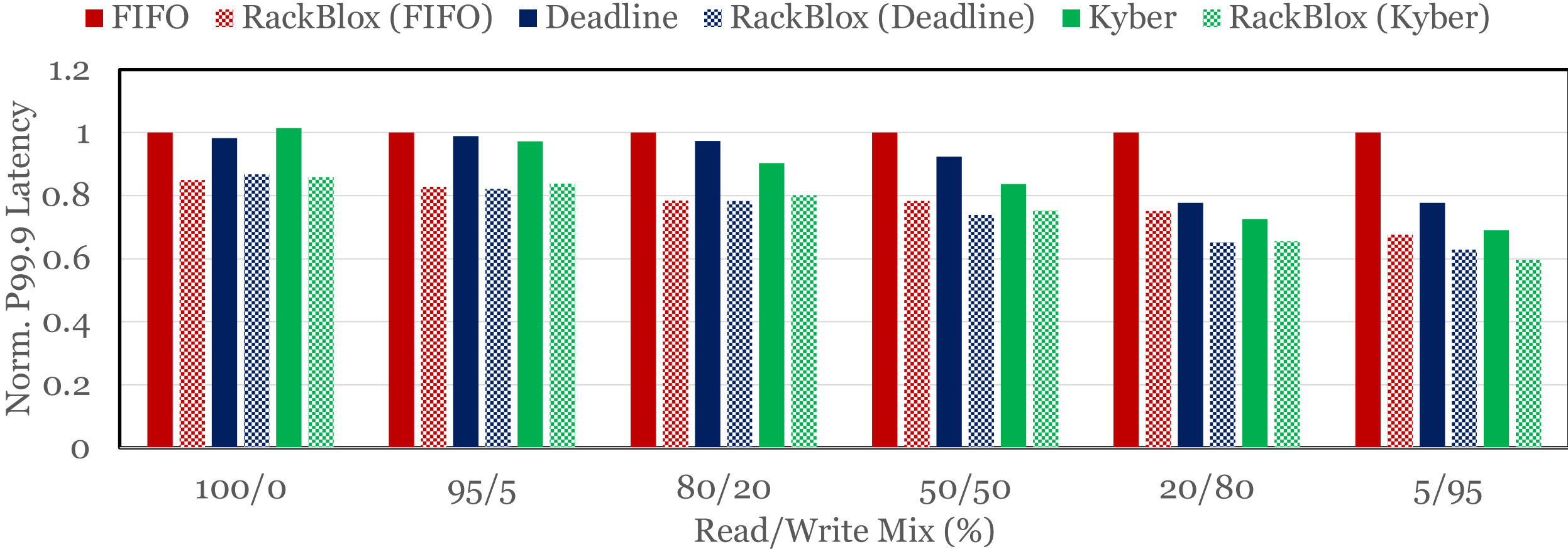
RackBlox reduces tail latency by up to 4.8x!

Improving End-To-End I/O Performance



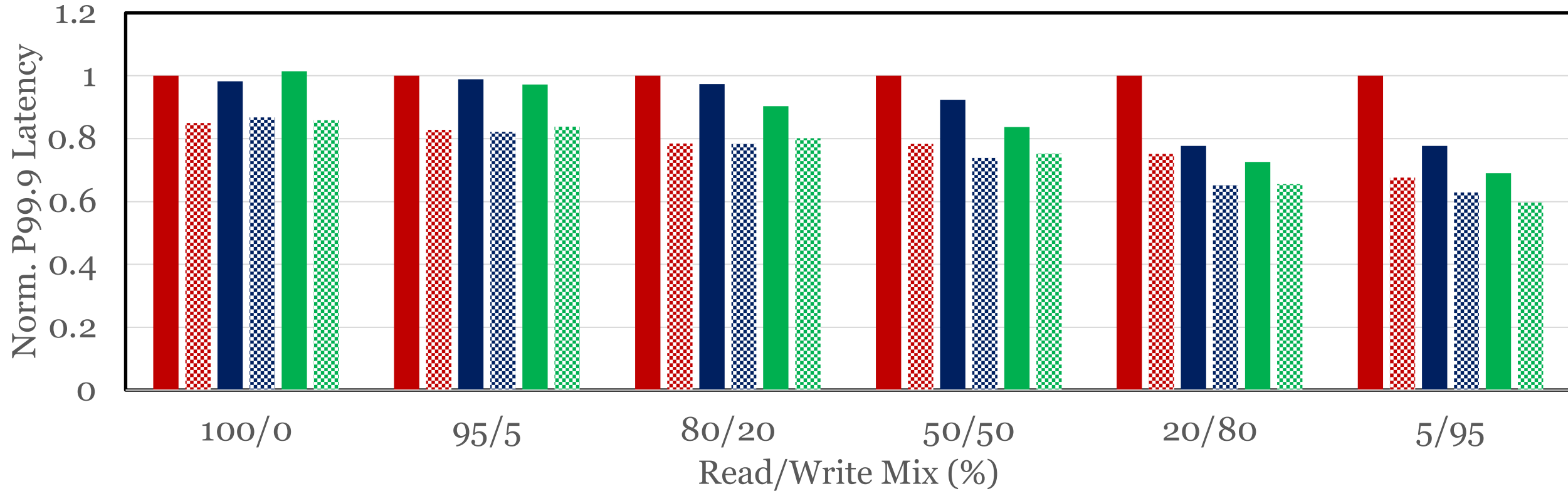
RackBlox reduces tail latency by up to 4.8x!

RackBlox Supports Different Storage Schedulers



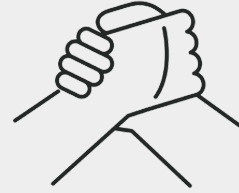
RackBlox Supports Different Storage Schedulers

■ FIFO ■ RackBlox (FIFO) ■ Deadline ■ RackBlox (Deadline) ■ Kyber ■ RackBlox (Kyber)



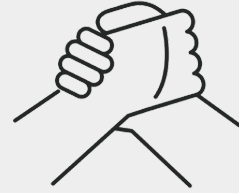
Coordinated I/O Scheduling always outperforms incoordination!

RackBlox Summary



Network/Storage Codesign

RackBlox Summary

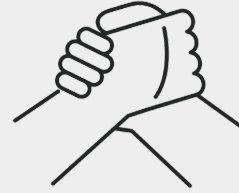


Network/Storage Codesign



**Improves End-to-End I/O
Performance**

RackBlox Summary



Network/Storage Codesign



**Improves End-to-End I/O
Performance**



**Ensures Rack-Scale Wear-
Balance**

Thank You!

Benjamin Reidys

Yuqi Xue Daixuan Li Bharat Sukhwani

Wen-mei Hwu Deming Chen Sameh Asaad Jian Huang

Systems Platform Research Group



UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN