AutoSSD: CXL-Enhanced Autonomous SSDs for Low Tail Latency

Mingyao Shen[†]**,** Heewoo Kim[‡]**,** Suyash Mahar[†], Joseph Izraelevitz[‡], Steven Swanson[†]

[†]UC San Diego, [‡]University of Colorado, Boulder

HPDC 2025
2025/07/21

Outline

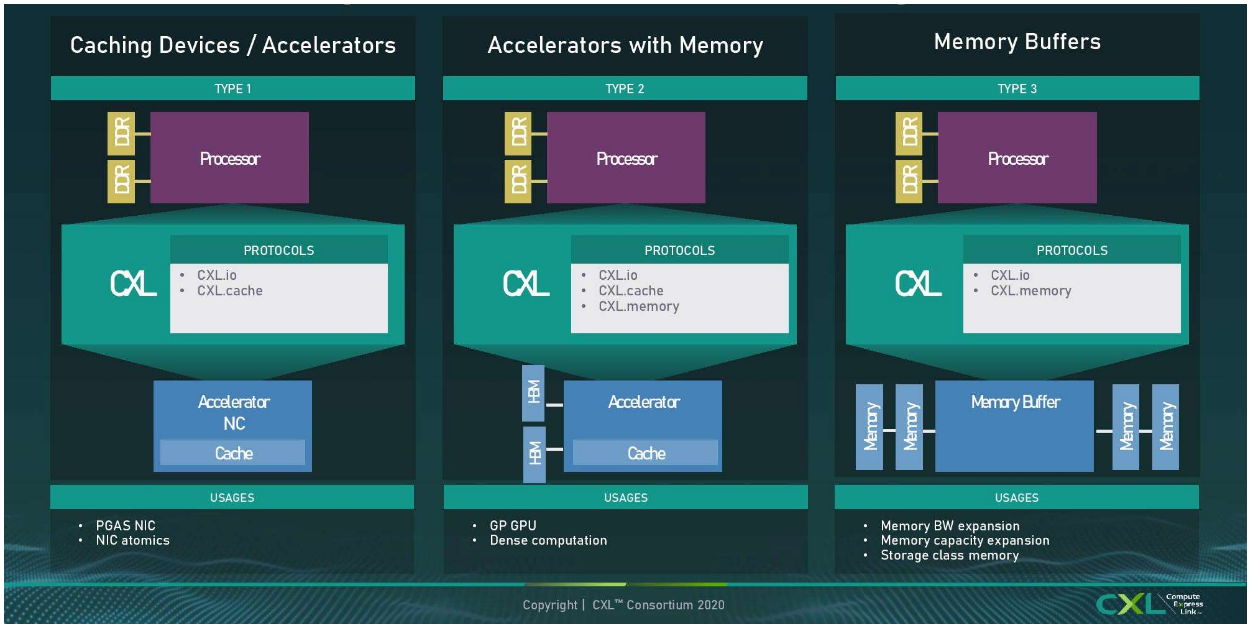
- Background and Motivation
- Problems
- Solutions
- Evaluation & Conclusion

Outline

- Background and Motivation
- Problems
- Solutions
- Evaluation & Conclusion

Compute Express Link (CXL) is Here

- Standard for high-speed communication
- Three sub-protocols
  - CXL.io
  - CXL.mem
  - CXL.cache

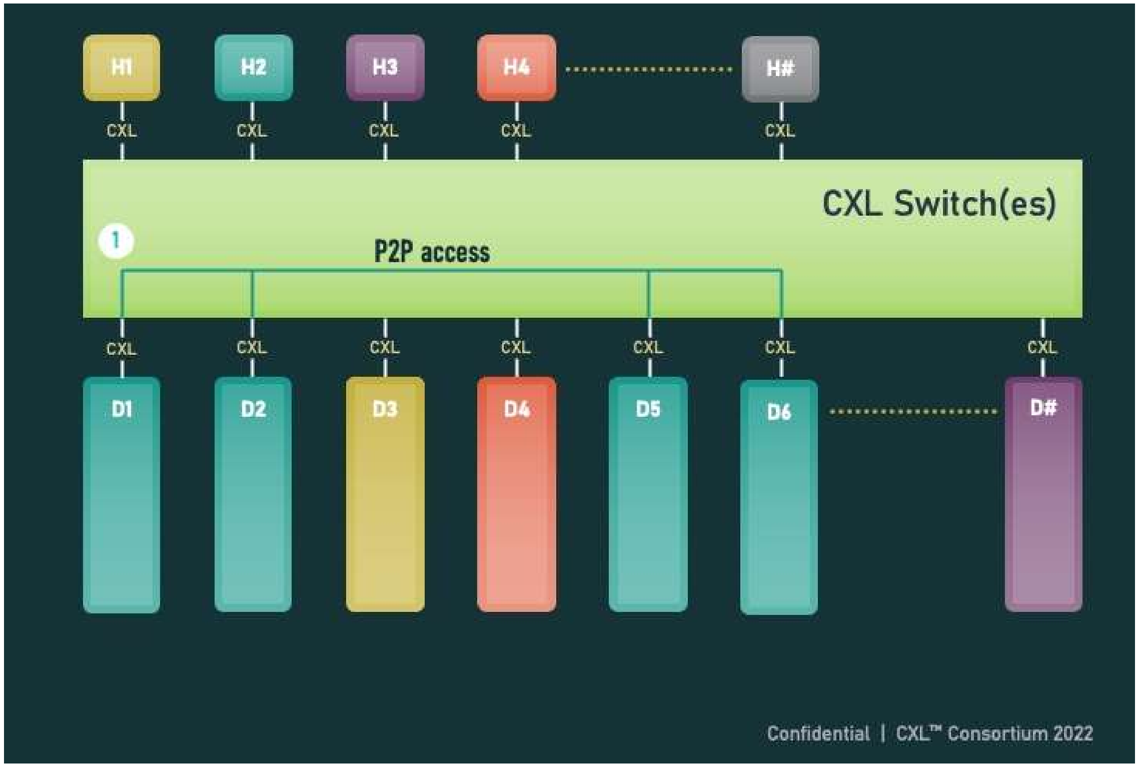


Representative CXL Usage*

*https://www.anandtech.com/show/16227/compute-express-link-cxl-2-0-switching-pmem-security

Compute Express Link (CXL) is Here

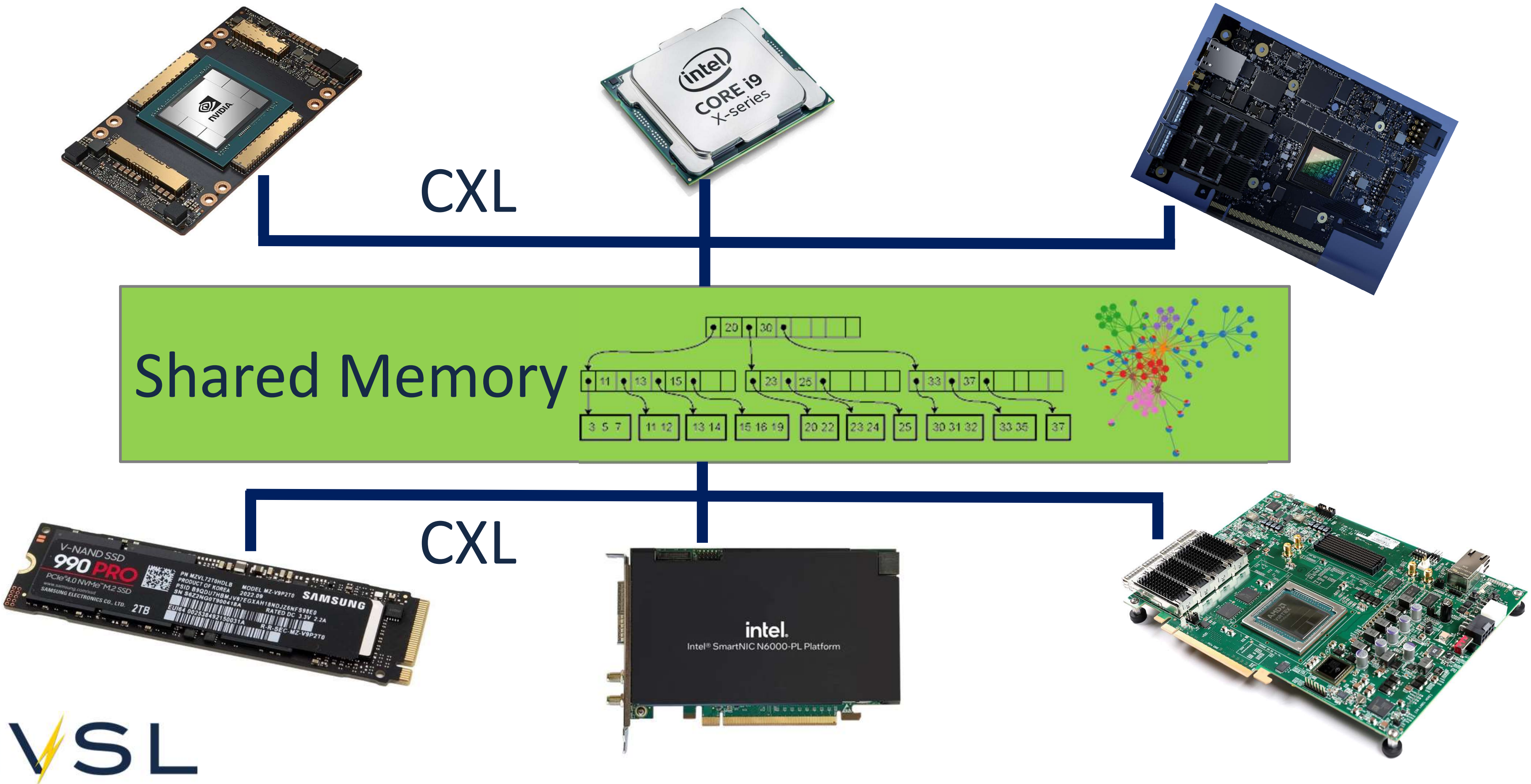
- Standard for high-speed communication
- Three sub-protocols
- Provides cache-coherent memory access
  - – CPU to device
  - – Device to CPU
  - – Device to device



CXL P2P Memory Access*

*https://www.servethehome.com/compute-express-link-cxl-3-0-is-the-exciting-building-block-for-disaggregation/

Device Collaboration Enabled

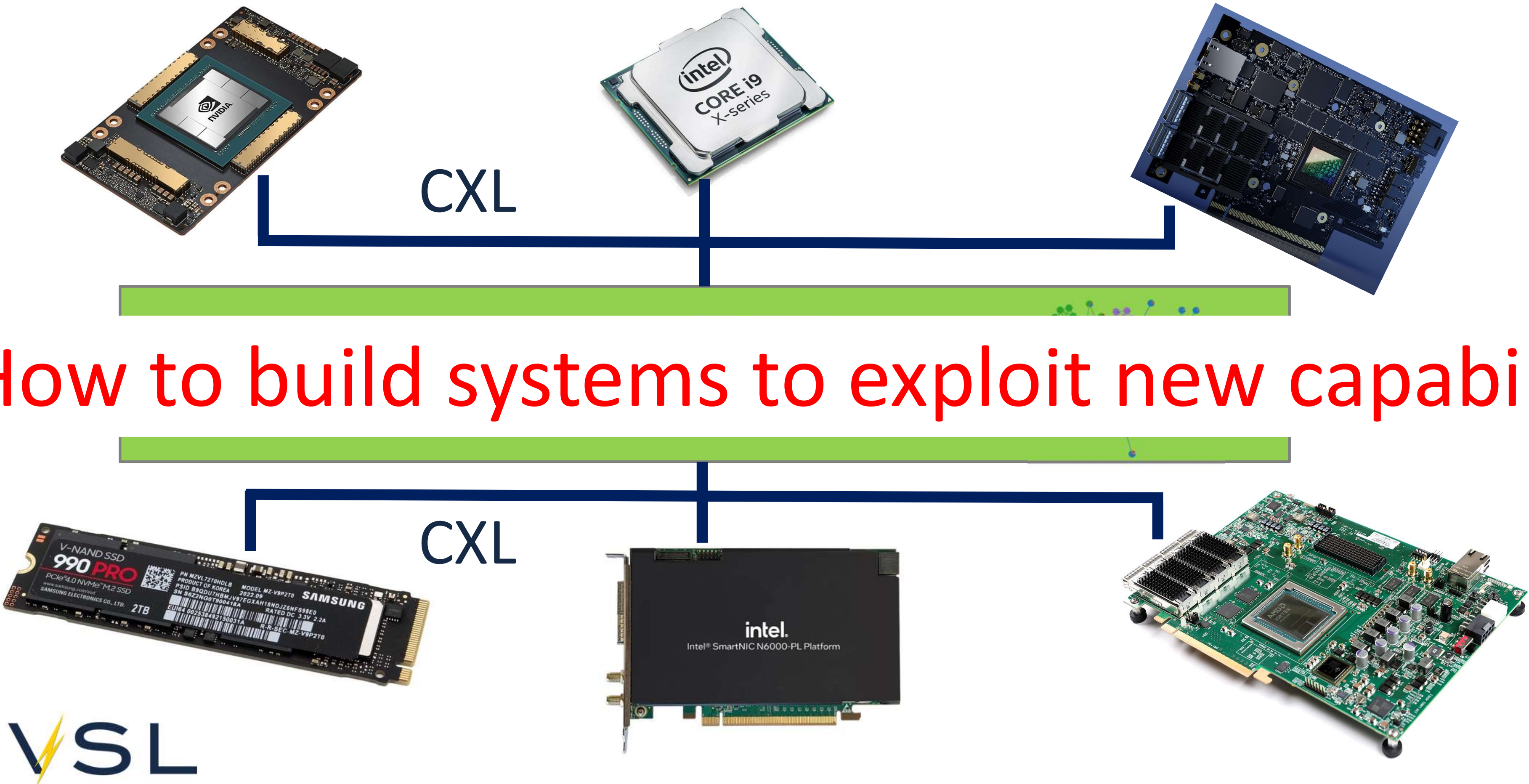


CXL

Shared Memory

CXL

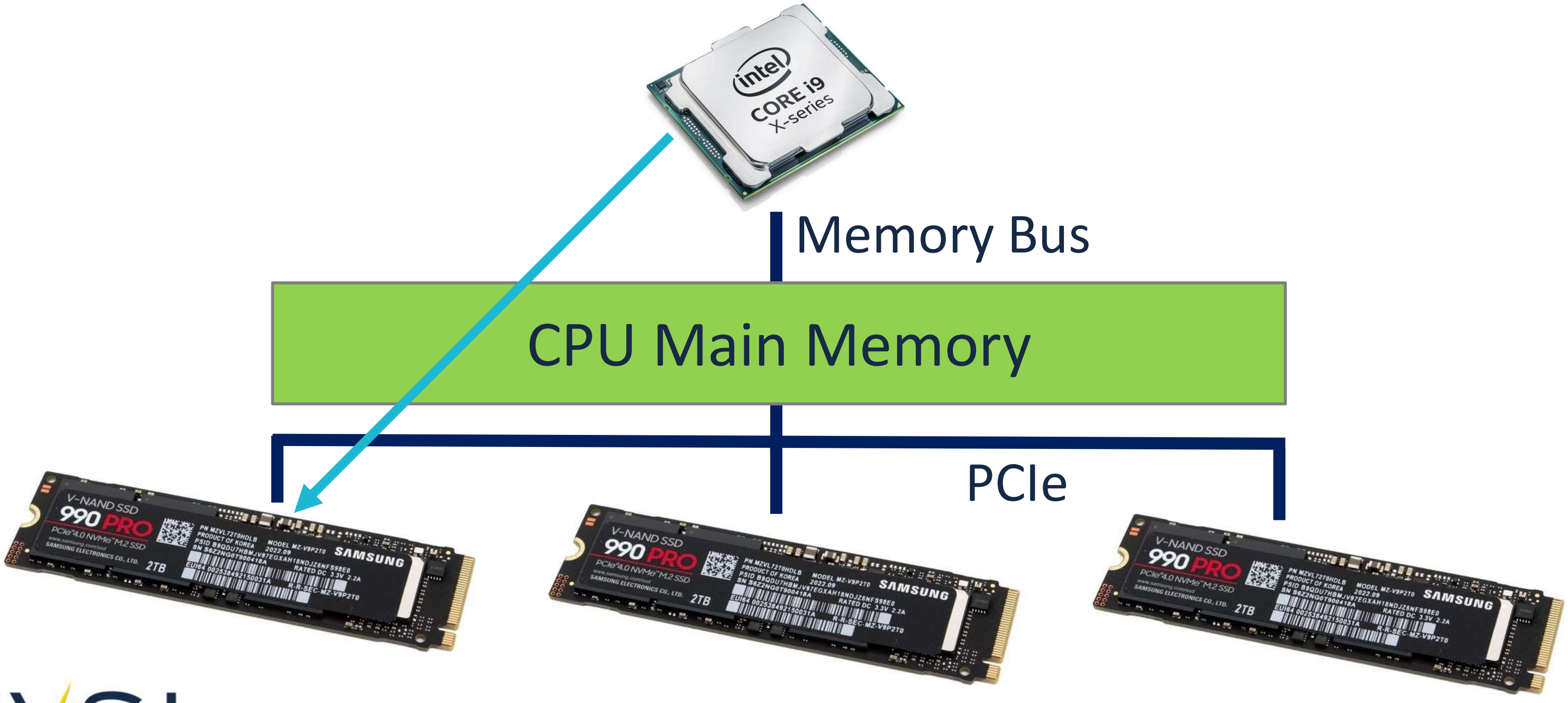
But why?



How to build systems to exploit new capability?

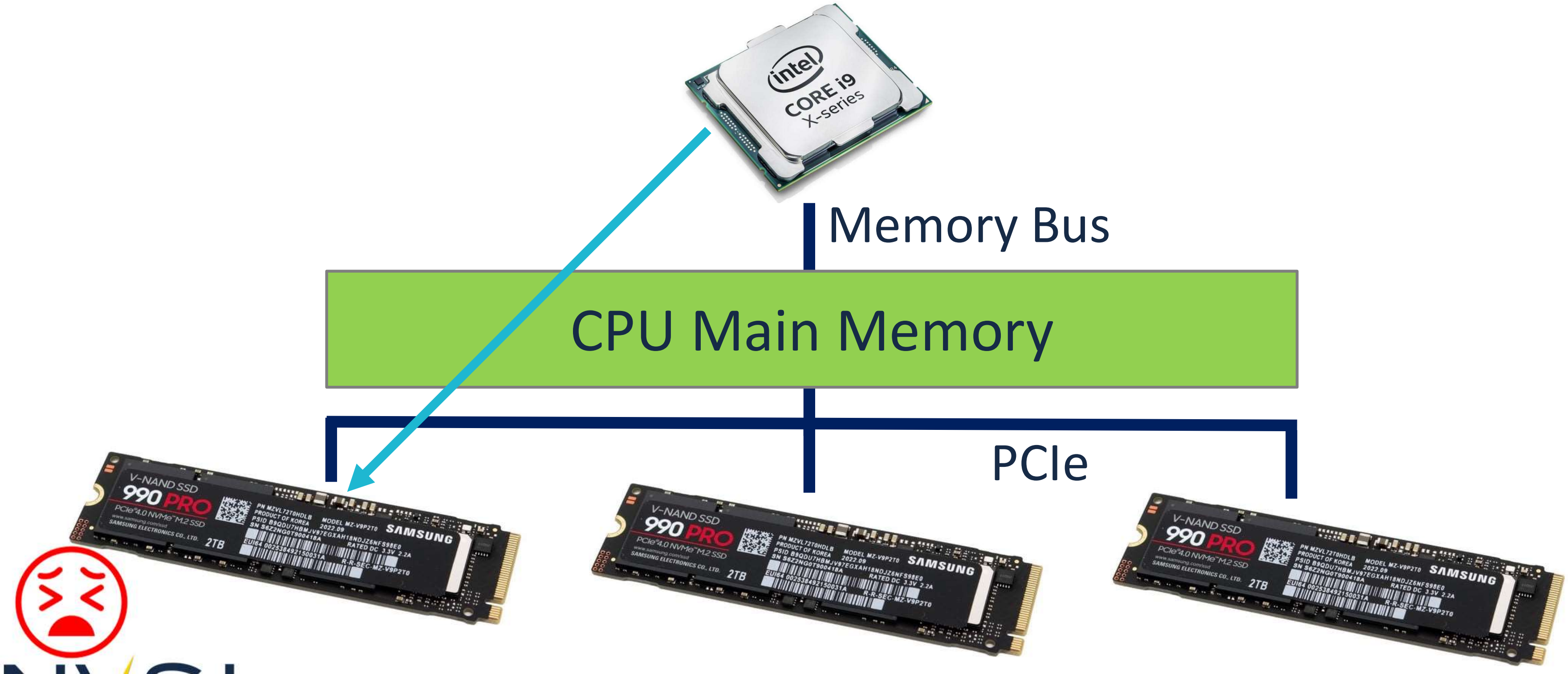
OS on CPU in Charge Now

- Peripherals are managed as black boxes



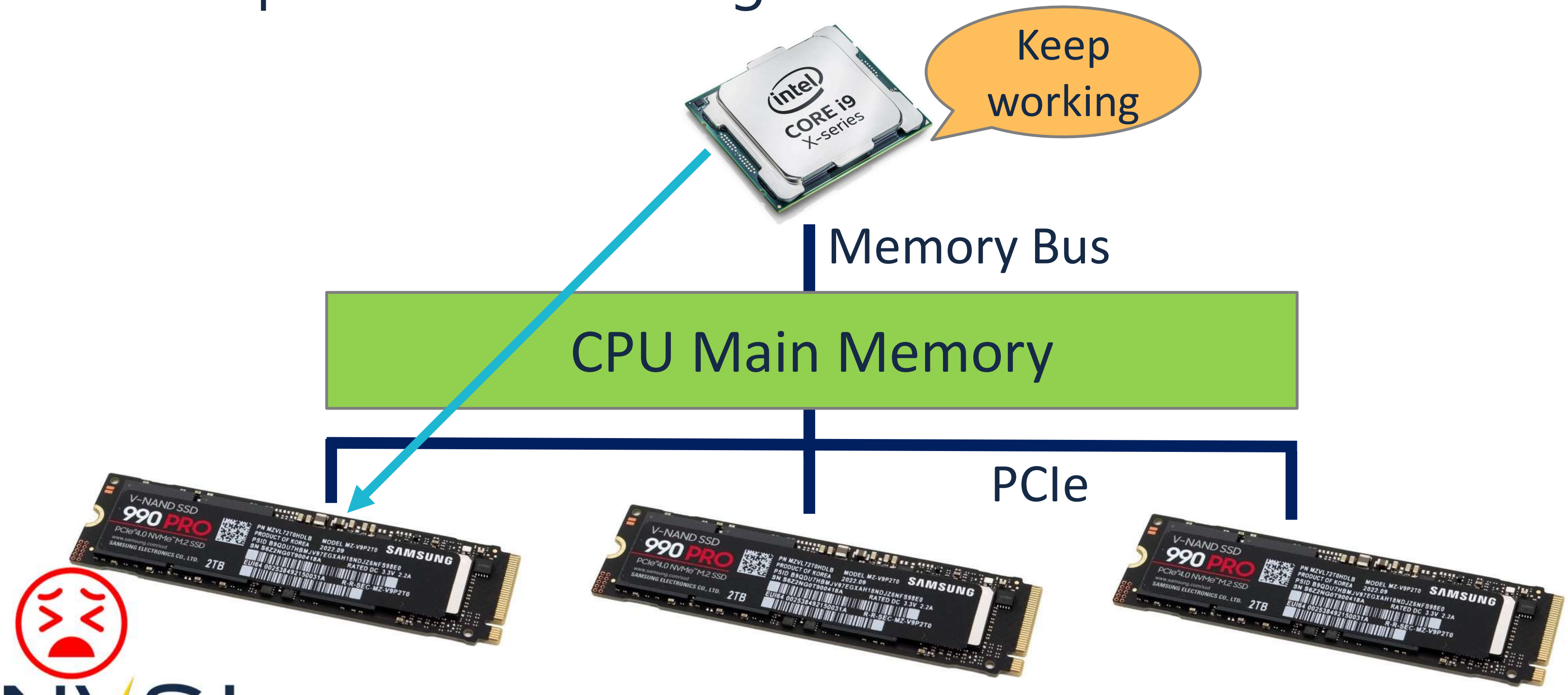
OS on CPU in Charge Now

- Peripherals are managed as black boxes



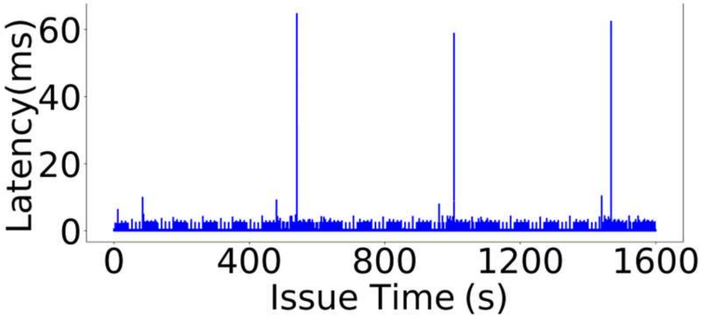
OS on CPU in Charge Now

- Peripherals are managed as black boxes



SSD Tail Latency Problem

- SSD’s performance jitters
  - Garbage collection
  - Wear leveling



Sample spike behavior in write tests on an SSD (write request size at 64KB).
Source: Figure 4(a) from FusionRAID

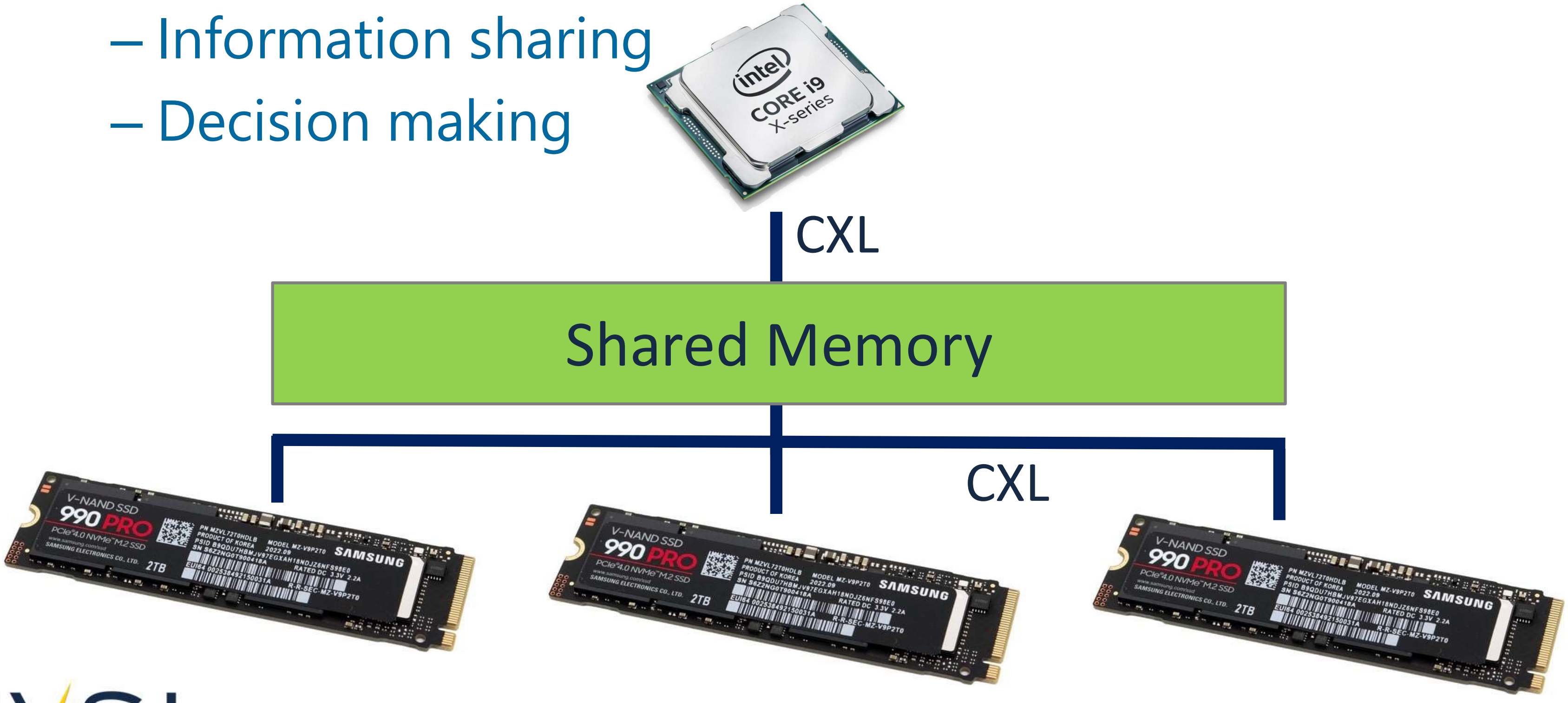
- SSD RAID’s tail latency is worse
  - One operation touches multiple SSDs
  - Any SSD’s spike latency affects the whole operation

SSDs	Median latency (ms)	Avg. latency (ms)	P99 latency (ms)	P999 latency (ms)
SSD 0	0.049	0.68	0.42	24.4
SSD 1	0.049	1.26	0.46	702.24
SSD 2	0.05	0.63	0.39	30.16
SSD 3	0.049	1.64	0.53	895.02
SSD 4	0.05	1.71	0.64	827.91

Application Exchange latency, individual aged SSDs within RAID.
Source: Table 2 from FusionRAID

Chance for Collaboration

- Devices can collaborate
  - Information sharing
  - Decision making



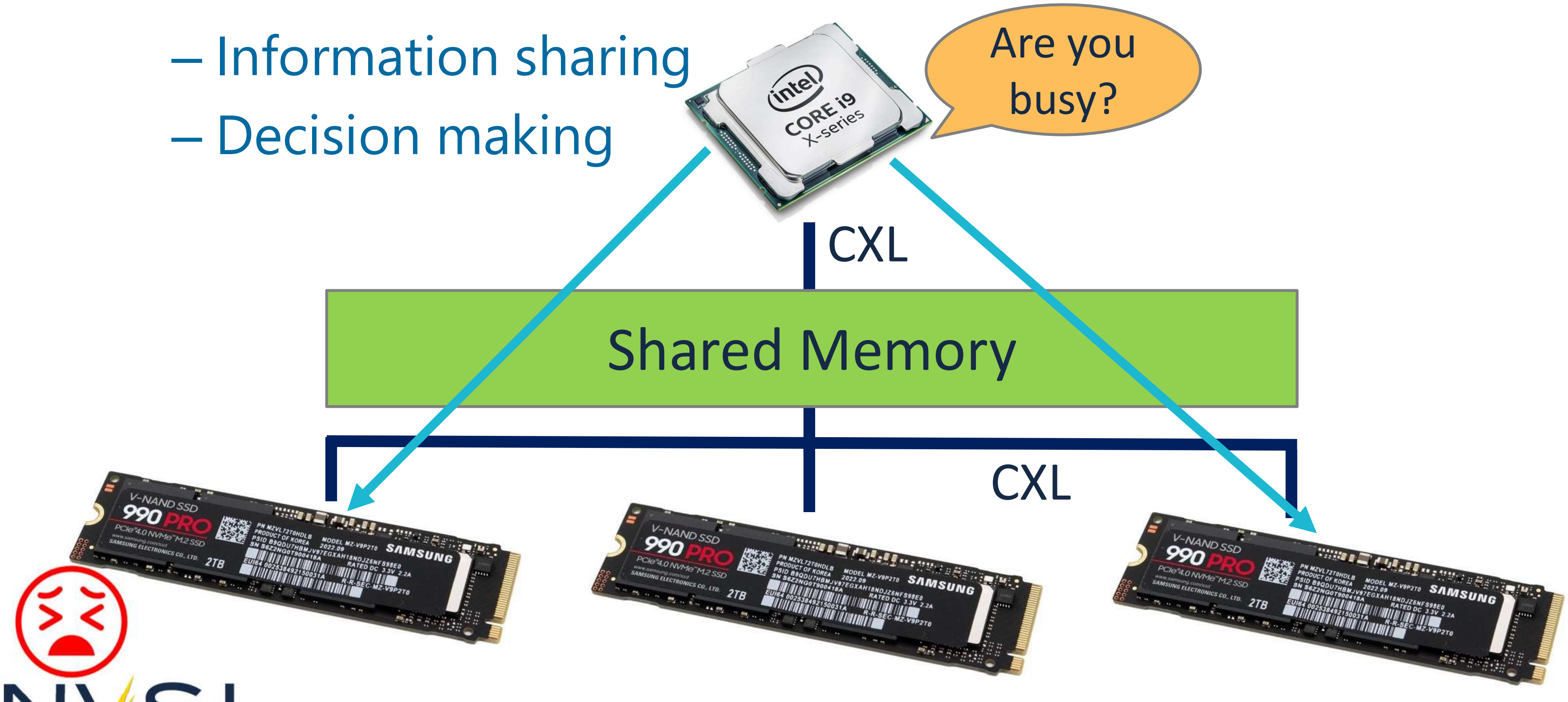
CXL

Shared Memory

CXL

Chance for Collaboration

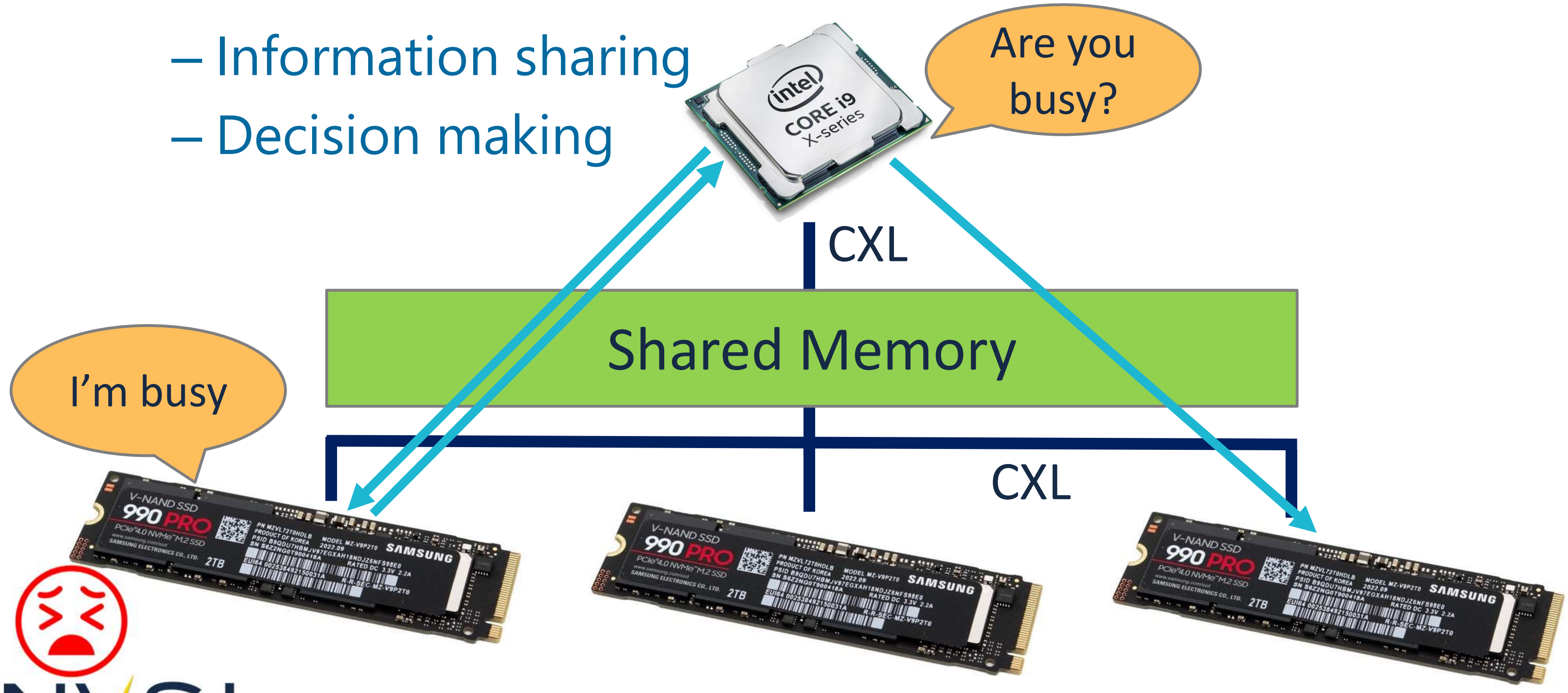
- Devices can collaborate
  - Information sharing
  - Decision making



NVSL

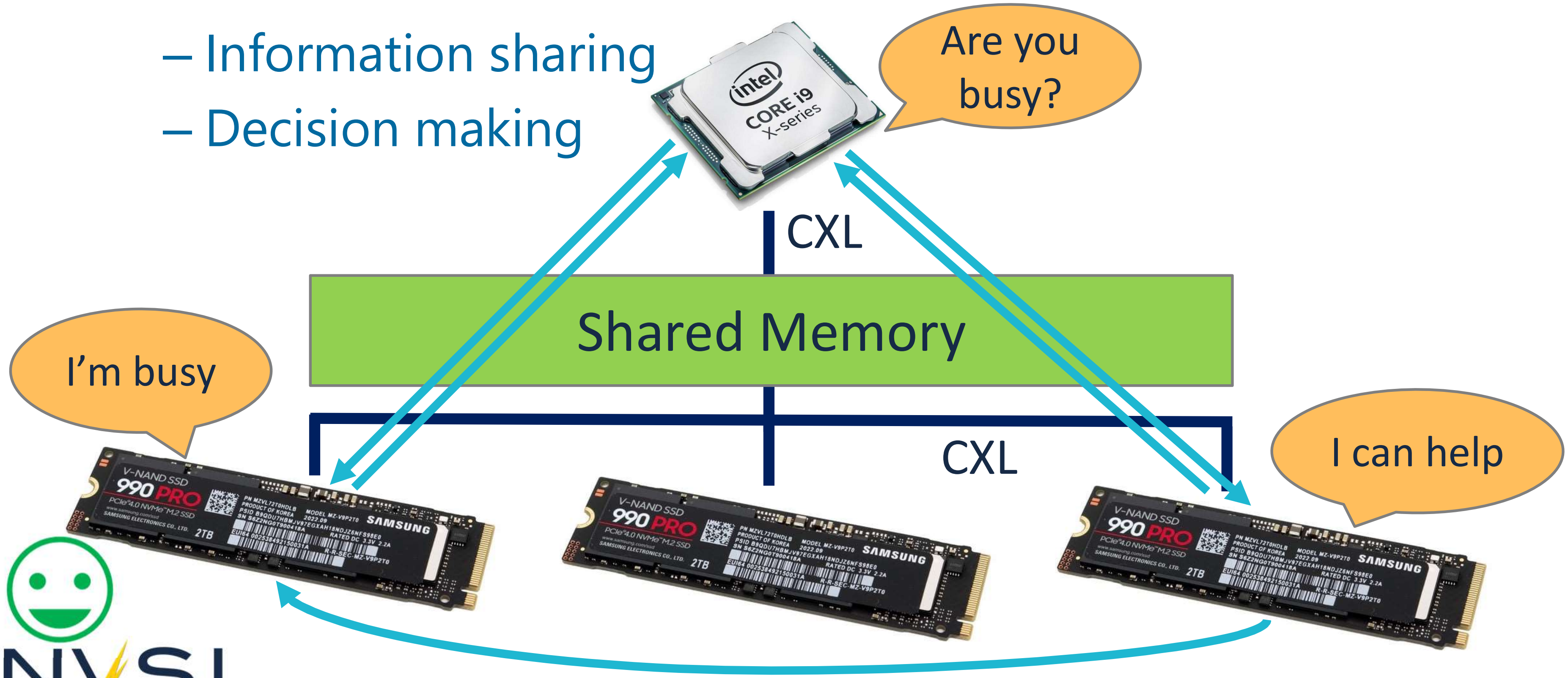
Chance for Collaboration

- Devices can collaborate
  - Information sharing
  - Decision making



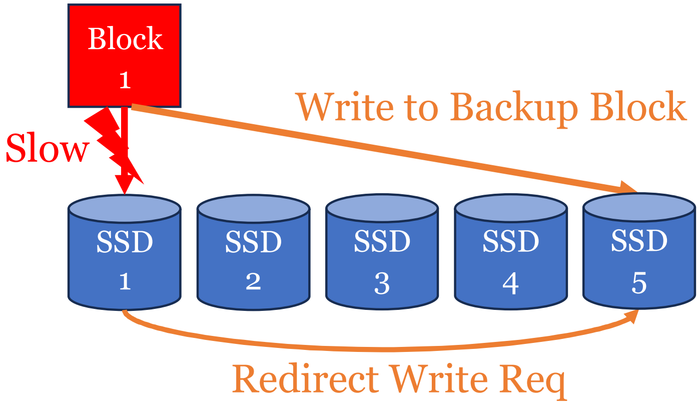
Chance for Collaboration

- Devices can collaborate
  - Information sharing
  - Decision making

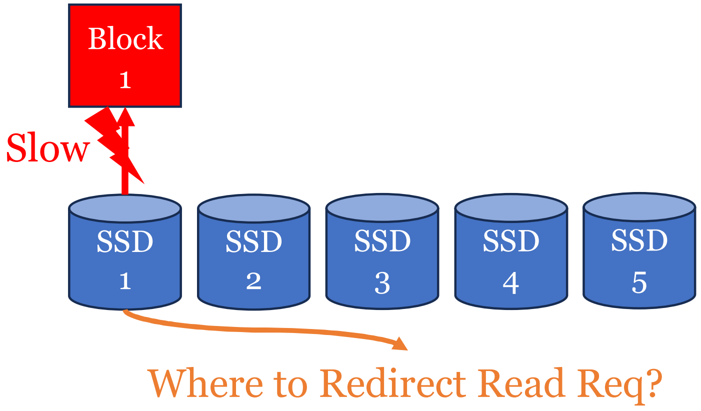


Straightforward Solution

Write operation: redirect



Read operation: redirect



Outline

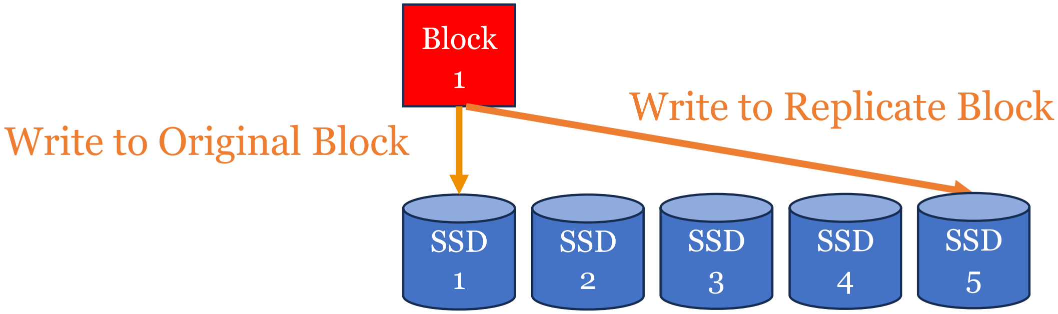
- Background and Motivation
- Problems
- Solutions
- Evaluation & Conclusion

Problems

- Replication Overhead
- Block Tracking
- CPU Centric
- Redirection Performance

Problem 1: Replication Overhead

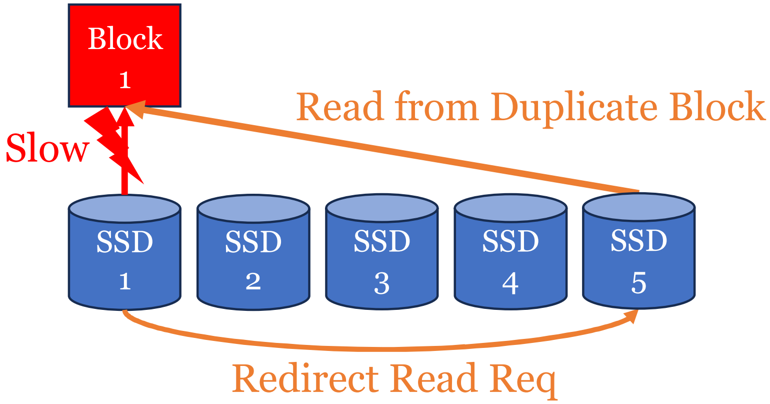
Normal Write



Problem 1: Replication Overhead

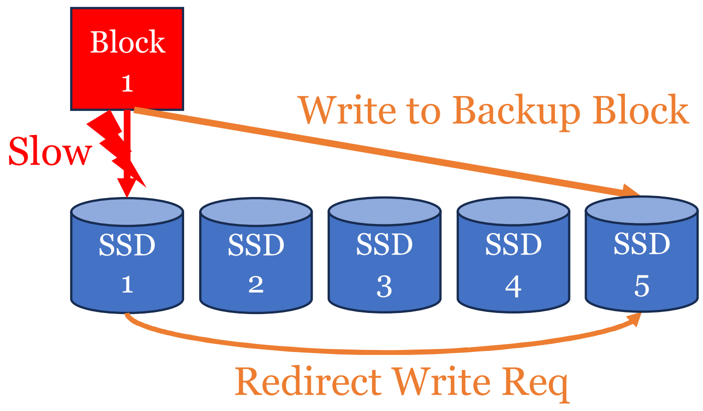
- Hurt normal write performance
- Duplication wastes space

Read operation: redirect



Problem 2: Block Tracking

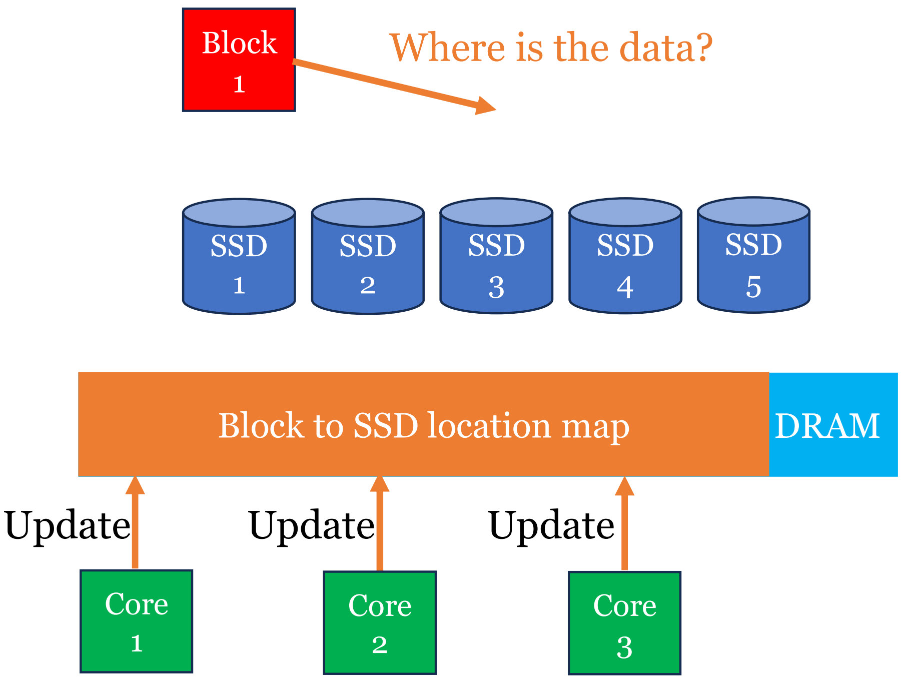
Write operation: redirect



Problem 2: Block Tracking

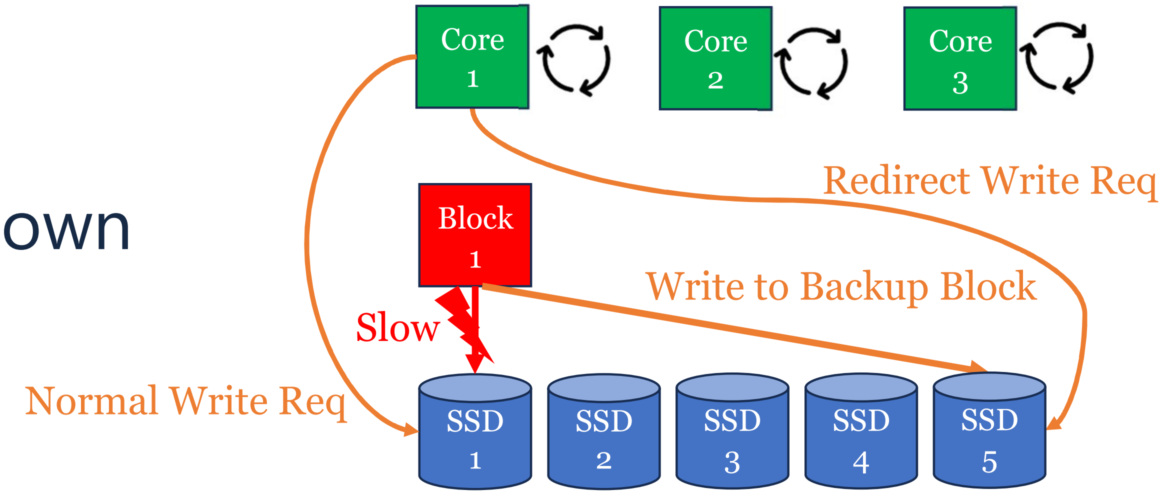
- Memory footprint
- Concurrent map update

Normal Read



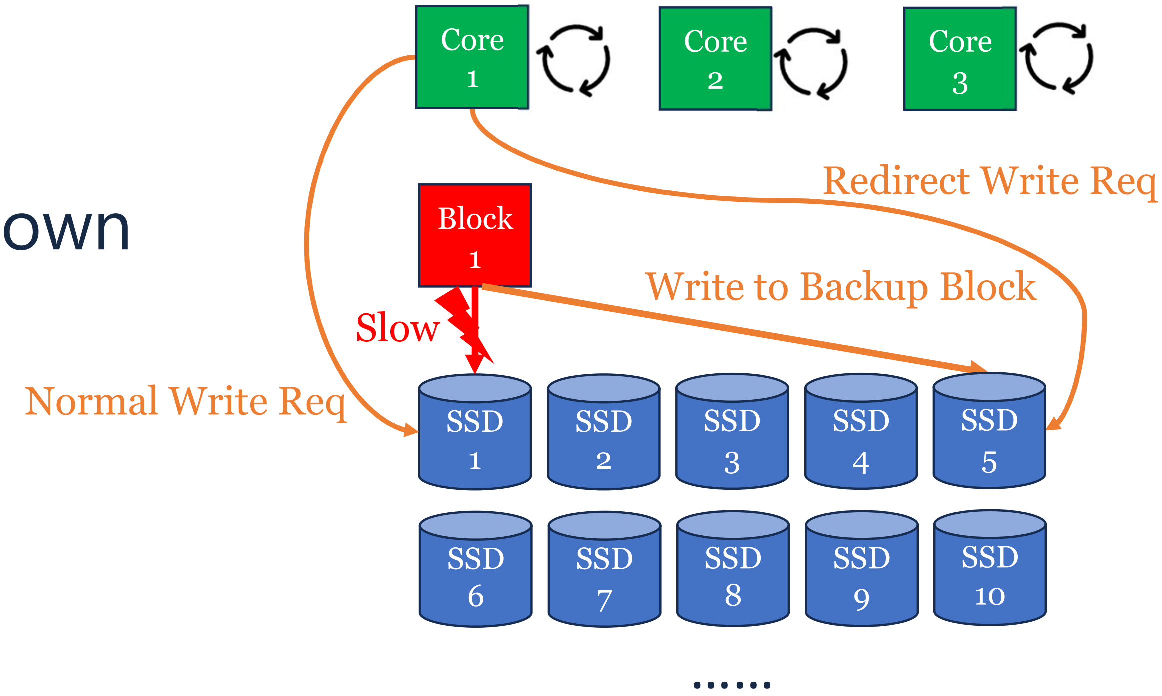
Problem 3: CPU Centric

- Interference with normal work
  - Monitoring
  - Redirection
- SSD's internal state unknown
  - Late redirecting
  - Late resuming



Problem 3: CPU Centric

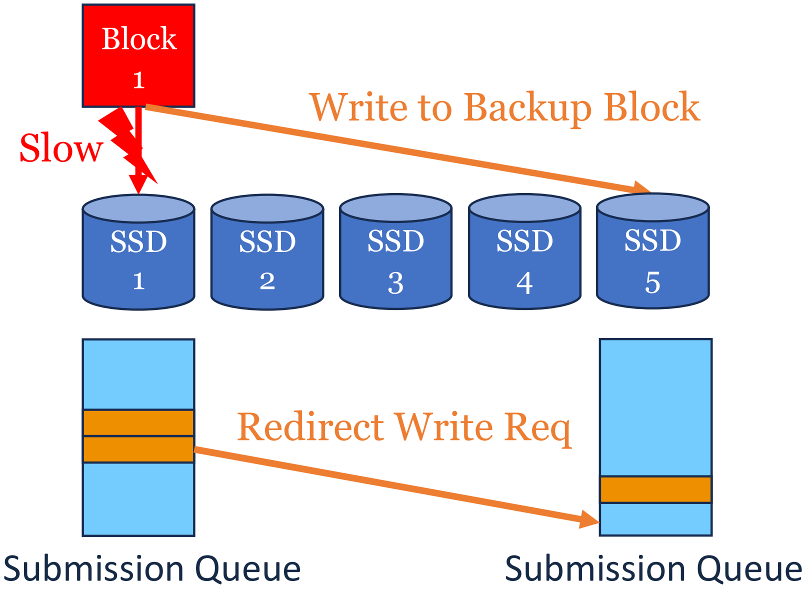
- Interference with normal work
  - Monitoring
  - Redirection
- SSD's internal state unknown
  - Late redirecting
  - Late resuming
- Not scalable



NVSL

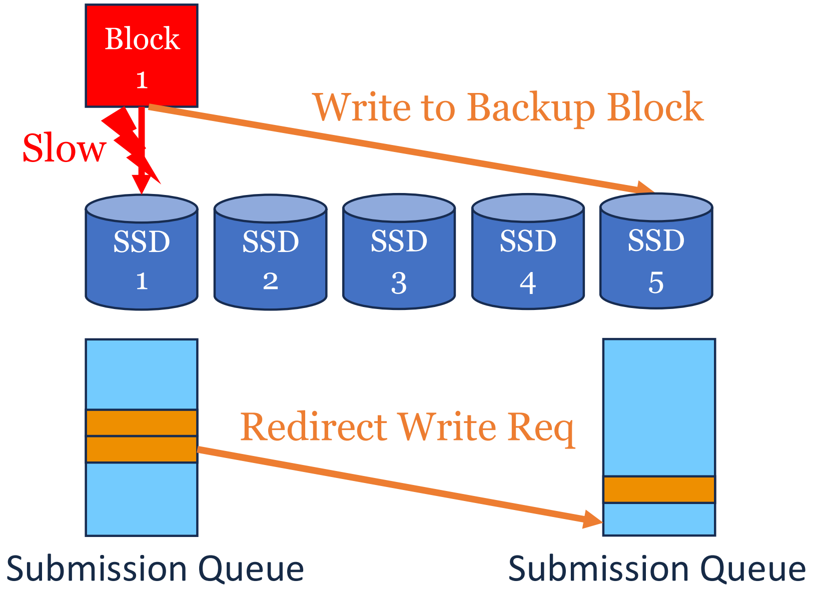
Problem 4: Redirection Performance

- CPU: duplicate request
  - Shouldn’t remove requests
  - Waste bandwidth



Problem 4: Redirection Performance

- CPU: duplicate request
  - – Shouldn’t remove requests
  - – Waste bandwidth
- SSD: DMA not suitable
  - – Initialization and setup overhead
  - – Interrupt notification

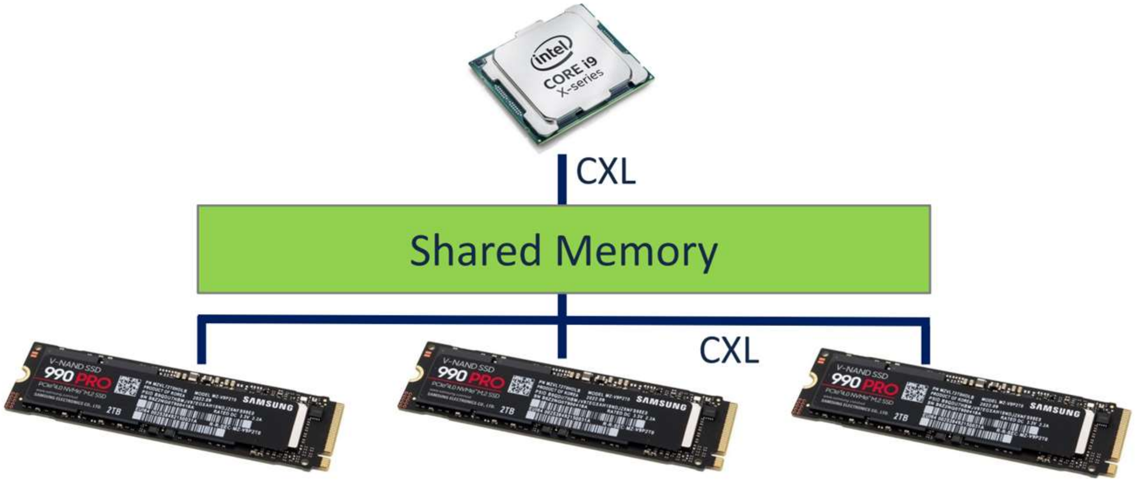


Outline

- Background and Motivation
- Problems
- Solutions
- Evaluation & Conclusion

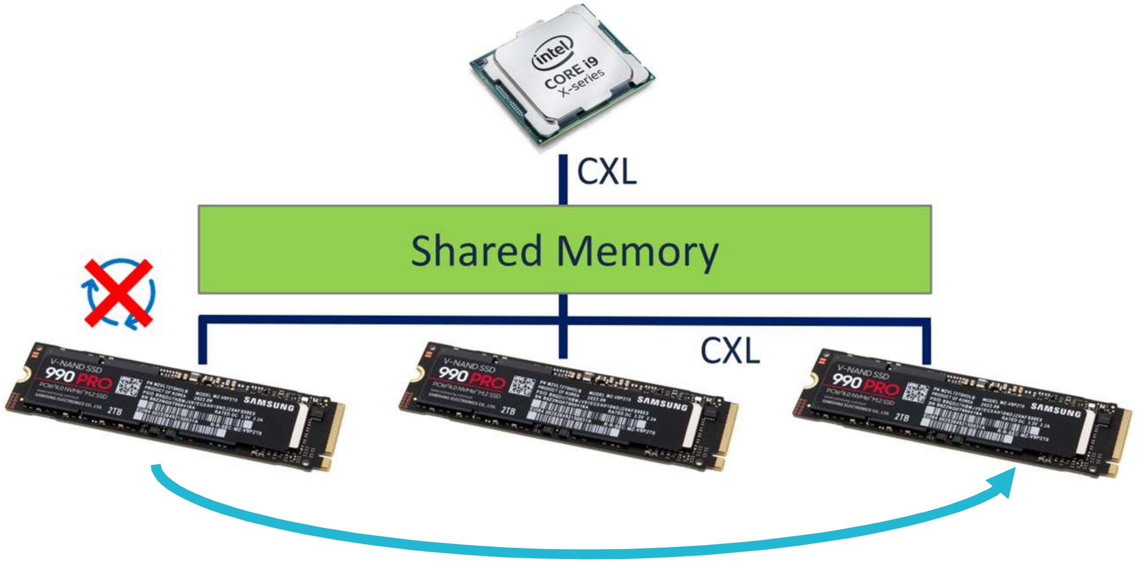
CXL-based SSD Autonomic and Collaborative Scheduling

- CXL-based
  - High-performance communication



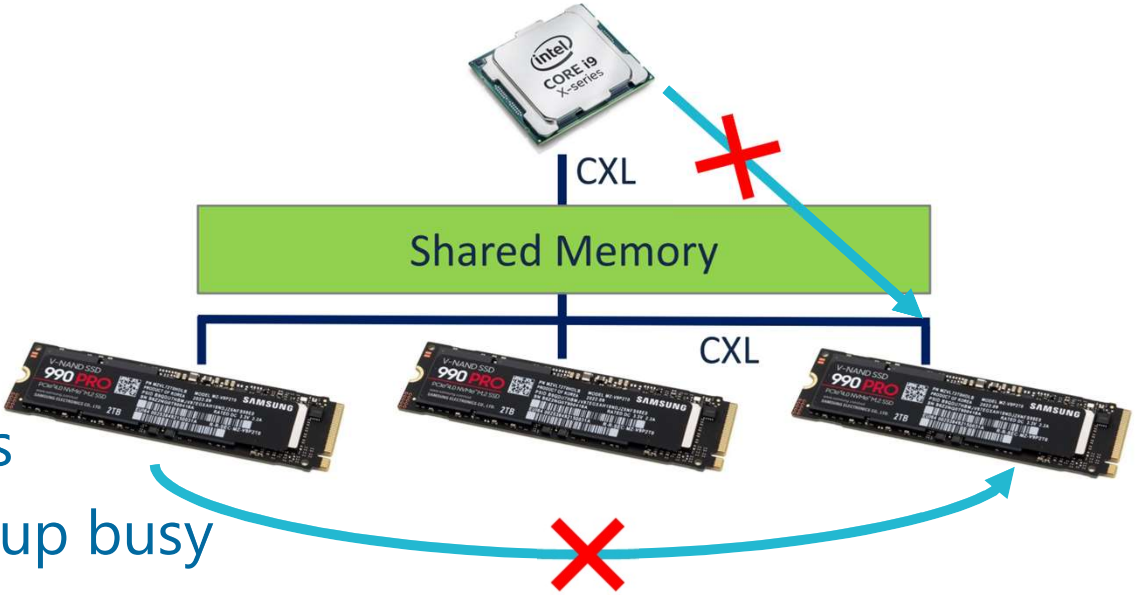
CXL-based SSD Autonomic and Collaborative Scheduling

- CXL-based
  - High-performance communication
- SSD autonomic
  - Stop polling when busy
  - Redirect / rebuild



CXL-based SSD Autonomic and Collaborative Scheduling

- CXL-based
  - High-performance communication
- SSD autonomic
  - Stop polling when busy
  - Redirect / rebuild
- Share status
  - Flag busy to stop requests
  - No redirection when backup busy



Solutions

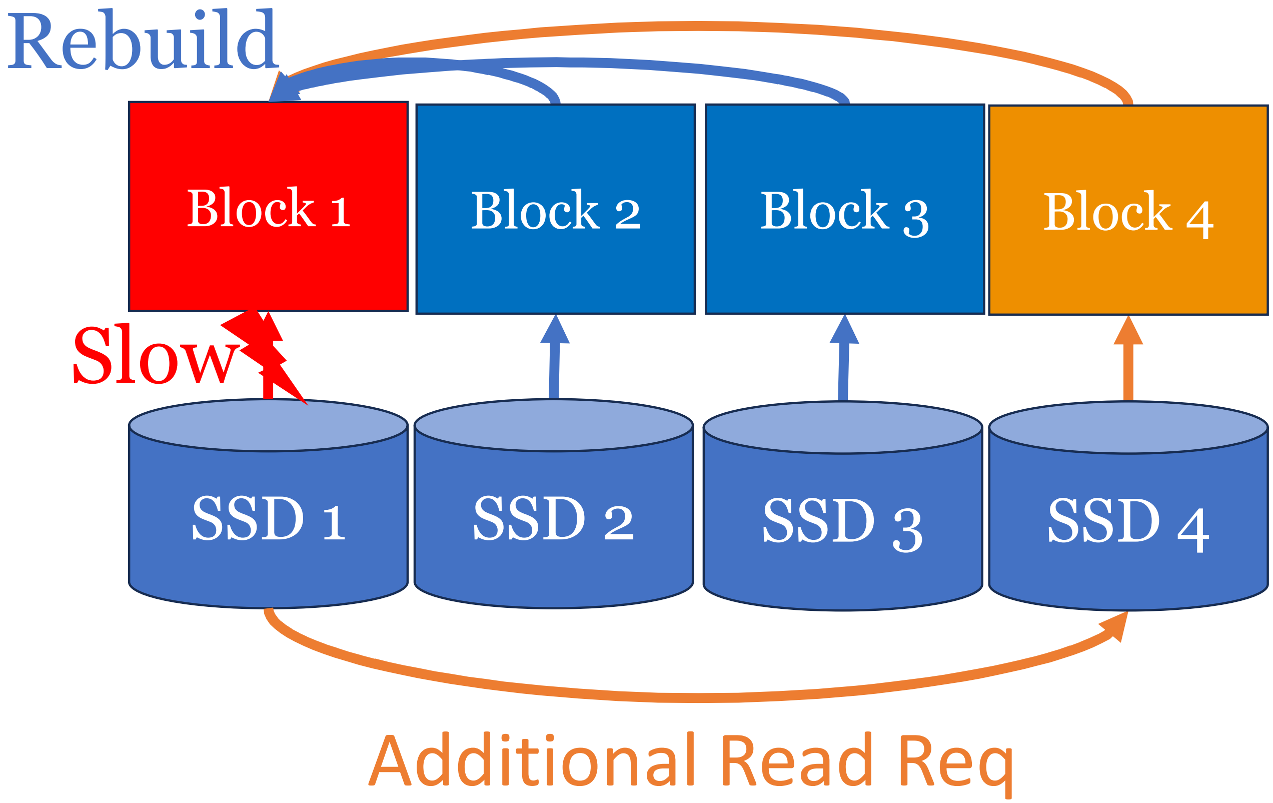
- Replication Overhead -> RAID Rebuild
- Block Tracking -> Stack + Dynamic Block Mapping
- CPU Centric -> CPU + SSD Collaboration
- Redirection Performance -> CXL Based Data Sharing

Solutions

- Replication Overhead -> RAID Rebuild
- Block Tracking -> Stack + Dynamic Block Mapping
- CPU Centric -> CPU + SSD Collaboration
- Redirection Performance -> CXL Based Data Sharing

Using RAID Rebuild for Read

Read operation: rebuild



Solutions

- Replication Overhead -> RAID Rebuild
- Block Tracking -> Stack + Dynamic Block Mapping
- CPU Centric -> CPU + SSD Collaboration
- Redirection Performance -> CXL Based Data Sharing

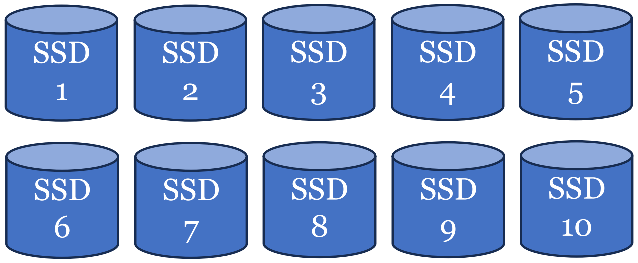
Static + Dynamic Block Mapping

- Deterministic block mapping calculation
- N (N <= stripe size) backup blocks for one stripe

Strips

Data		Parity	Backup
SSD 1 Blcok 0	SSD 2 Blcok 0	SSD 3 Blcok 0	SSD 4 Blcok 0
SSD 2 Blcok 1	SSD 3 Blcok 1	SSD 4 Blcok 1	SSD 5 Blcok 0
SSD 3 Blcok 2	SSD 4 Blcok 2	SSD 5 Blcok 1	SSD 6 Blcok 0
SSD 4 Blcok 3	SSD 5 Blcok 2	SSD 6 Blcok 1	SSD 7 Blcok 0

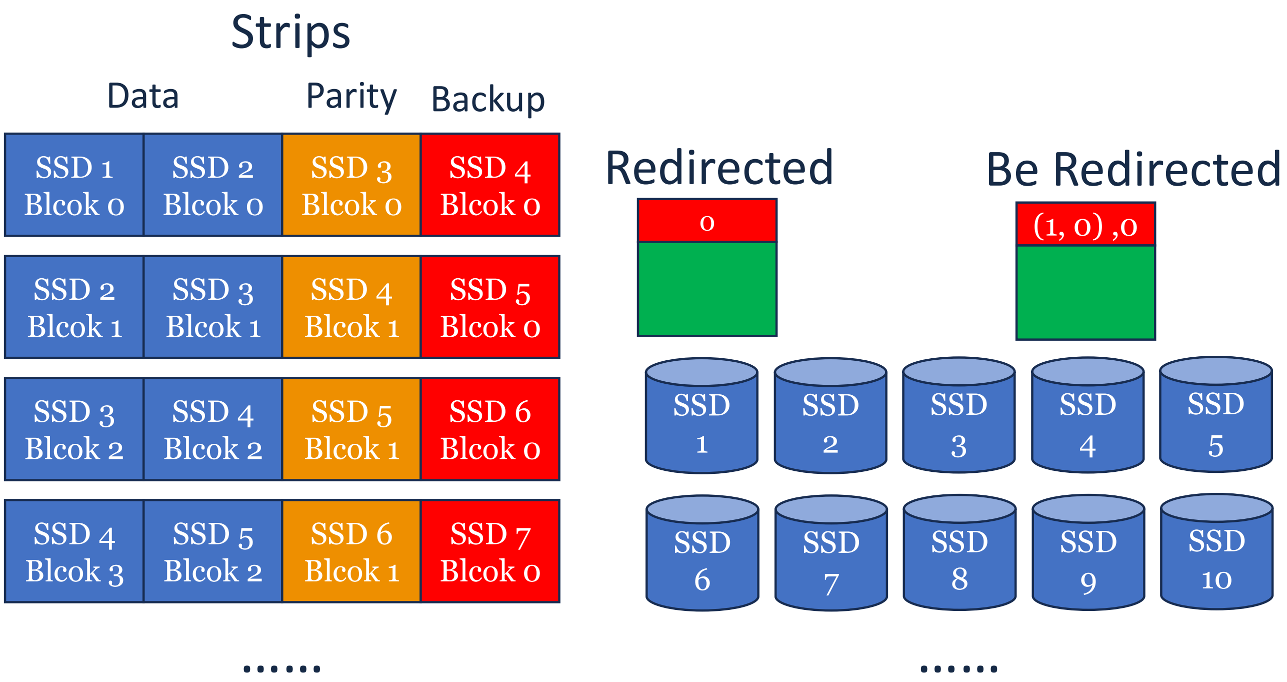
......



......

Static + Dynamic Block Mapping

- Deterministic block mapping calculation
- N (N <= stripe size) backup blocks for one stripe
- Dynamically tracking redirected data



Strips

Data		Parity	Backup
SSD 1 Blcok 0	SSD 2 Blcok 0	SSD 3 Blcok 0	SSD 4 Blcok 0
SSD 2 Blcok 1	SSD 3 Blcok 1	SSD 4 Blcok 1	SSD 5 Blcok 0
SSD 3 Blcok 2	SSD 4 Blcok 2	SSD 5 Blcok 1	SSD 6 Blcok 0
SSD 4 Blcok 3	SSD 5 Blcok 2	SSD 6 Blcok 1	SSD 7 Blcok 0

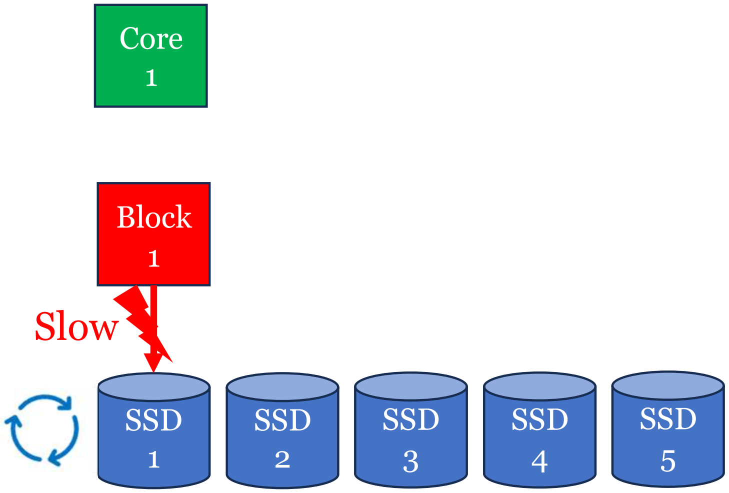
……

Solutions

- Replication Overhead -> RAID Rebuild
- Block Tracking -> Stack + Dynamic Block Mapping
- CPU Centric -> CPU + SSD Collaboration
- Redirection Performance -> CXL Based Data Sharing

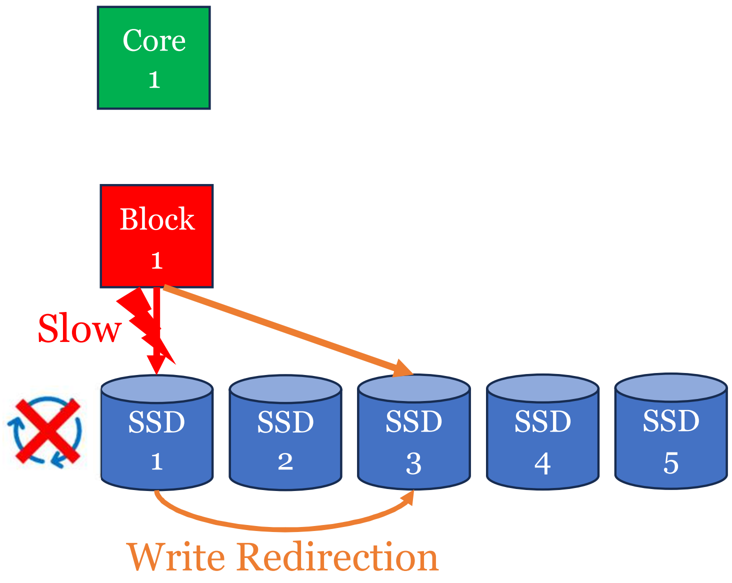
CPU + SSD Collaboration

- SSD autonomic



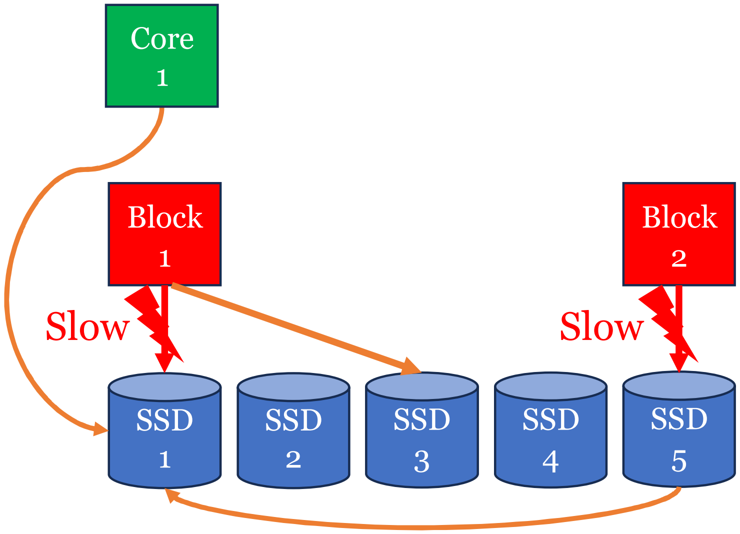
CPU + SSD Collaboration

- SSD autonomic
  - – Stop polling when busy
  - – Redirect / rebuild



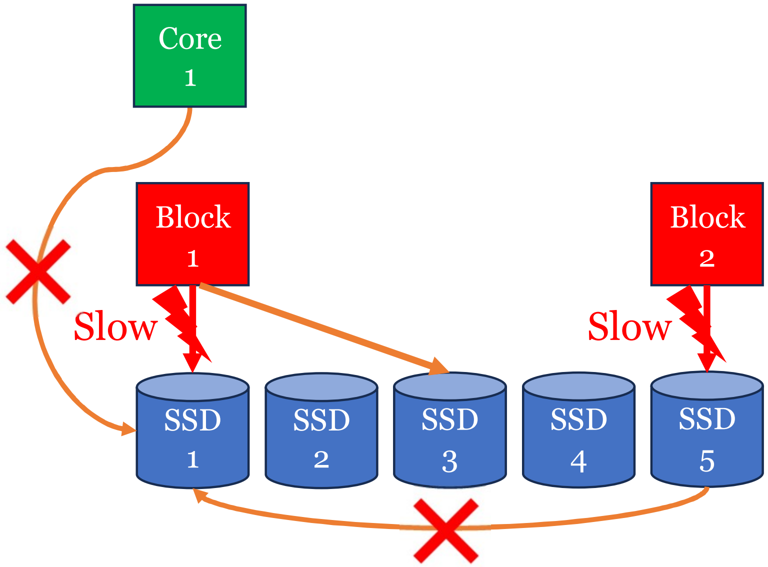
CPU + SSD Collaboration

- SSD autonomic
  - Stop polling when busy
  - Redirect / rebuild
- Share status



CPU + SSD Collaboration

- SSD autonomic
  - Stop polling when busy
  - Redirect / rebuild
- Share status
  - Flag busy to stop requests
  - No redirection when backup busy

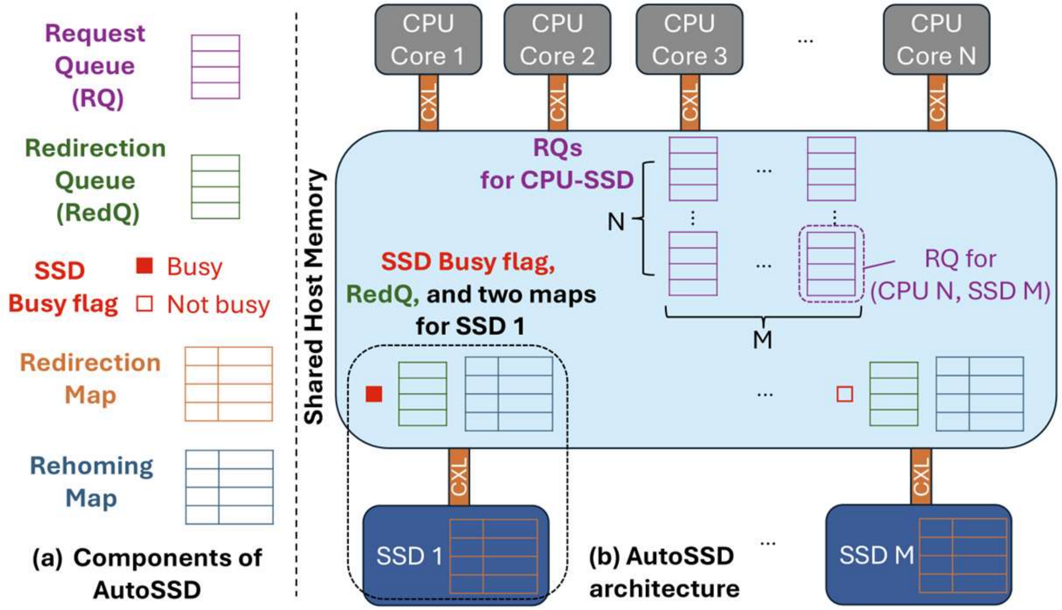


Solutions

- Replication Overhead -> RAID Rebuild
- Block Tracking -> Stack + Dynamic Block Mapping
- CPU Centric -> CPU + SSD Collaboration
- Redirection Performance -> CXL Based Data Sharing

CXL Based Data Sharing

- Communication through high-performance shared memory
- Optimized data structures



(a) Components of AutoSSD

(b) AutoSSD architecture

Outline

- Background and Motivation
- Problems
- Solutions
- Evaluation & Conclusion

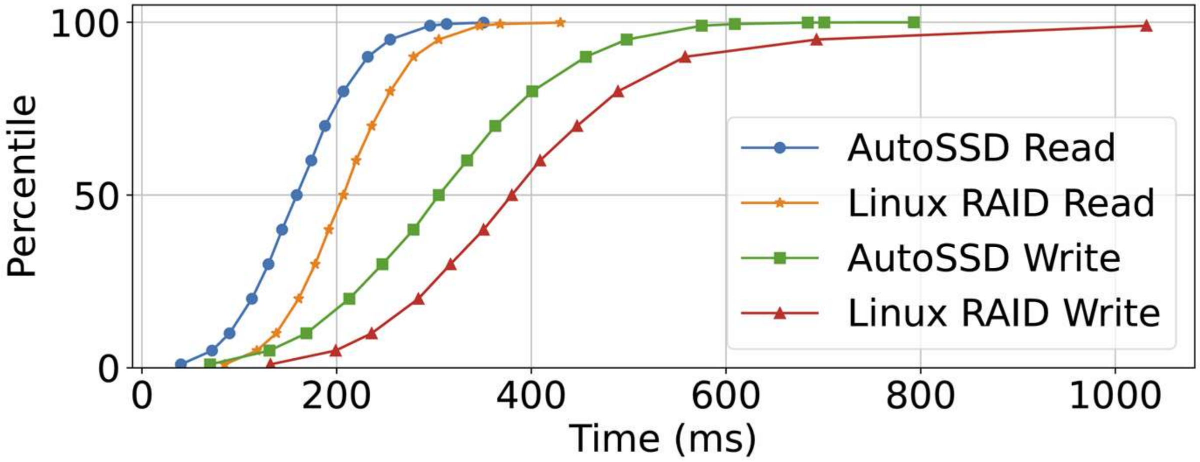
Evaluation

- Dual socket machine
- Remote socket as CXL memory

Evaluation

- Dual socket machine
- Remote socket as CXL memory
- ~15% and ~45% decreased P99 latency for random reads and writes

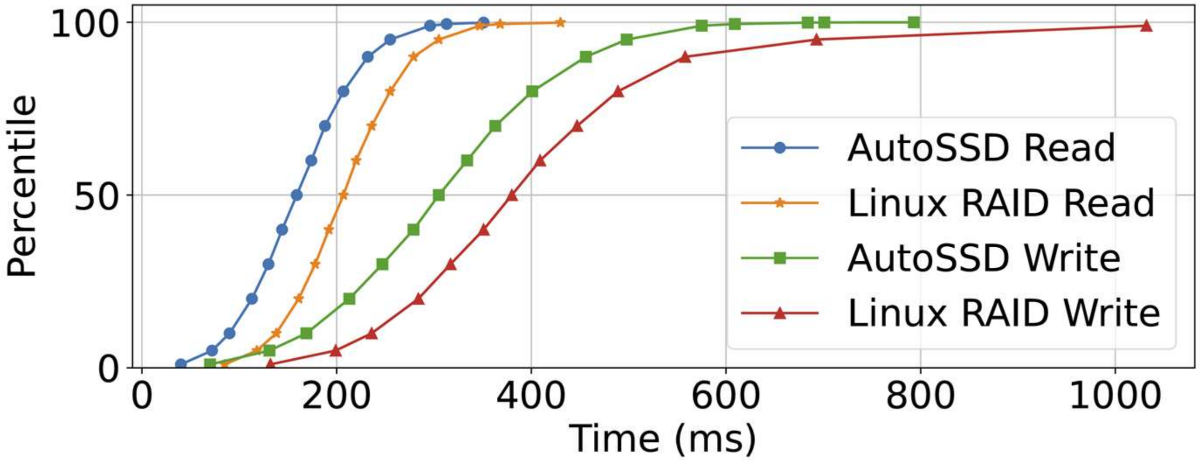
Random Read/Write



Evaluation

- Dual socket machine
- Remote socket as CXL memory
- ~15% and ~45% decreased P99 latency for random reads and writes

Random Read/Write



Conclusion

- CXL shared memory provides chances for devices' collaboration

Conclusion

- CXL shared memory provides chances for devices' collaboration
- SSDs can utilize it to improve their tail latency
  - Internal status sharing
  - Autonomic decision making
  - Collaboration

Conclusion

- CXL shared memory provides chances for devices' collaboration
- SSDs can utilize it to improve their tail latency
  - Internal status sharing
  - Autonomic decision making
  - Collaboration
- Other peripherals could also exploit it

Questions

Q&A