Use Cases in Database-SSD Co-Design

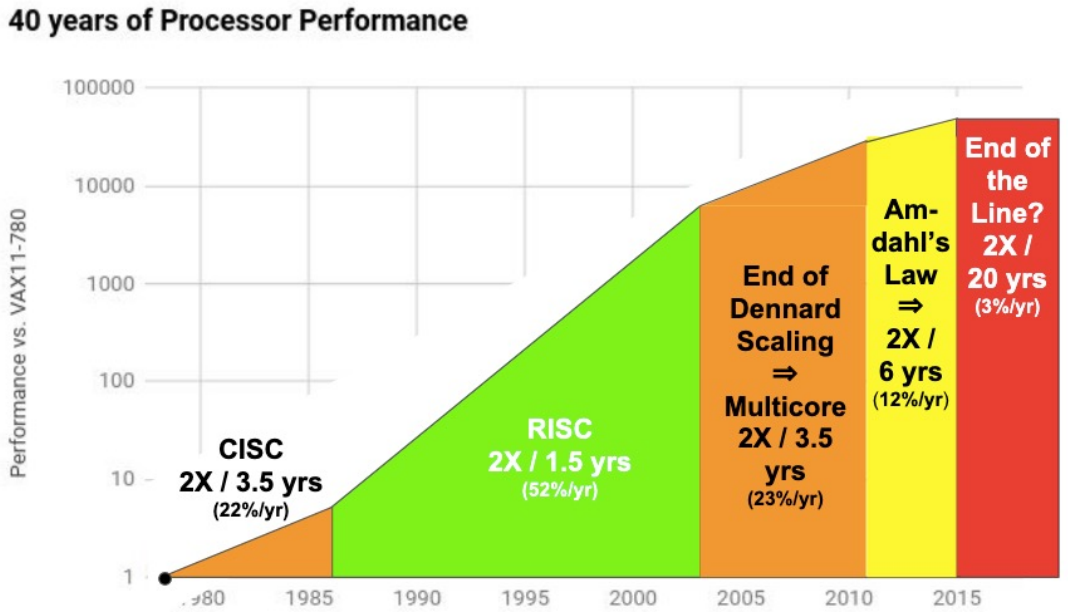
Alberto Lerner – eXascale Infolab
University of Fribourg – Switzerland

Joint work with ENC Lab at Hanyang University and Samsung

FG DB Spring Symposium 2022

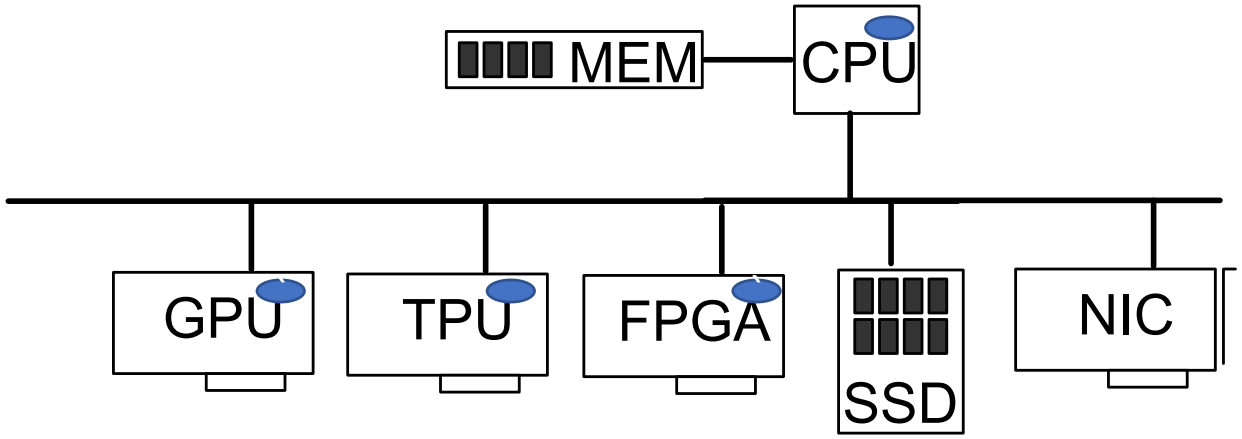
Motivation

- End of growth of single program speed (Patterson and Hennessy Turing Award lecture @ ISCA'18)
- Specialization is the answer!



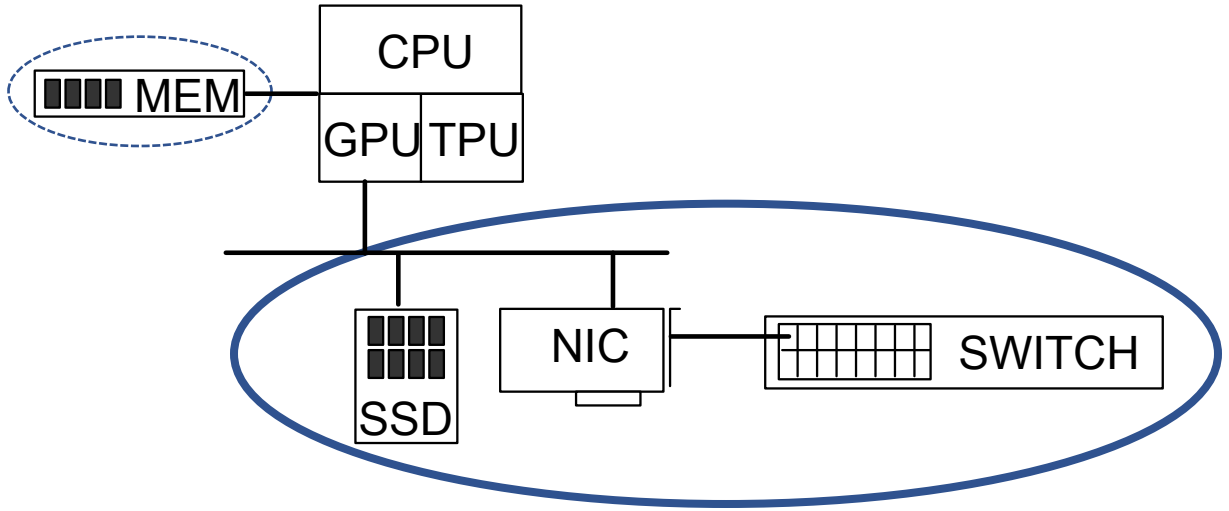
Specialization I

- Different computing units offer different functionalities



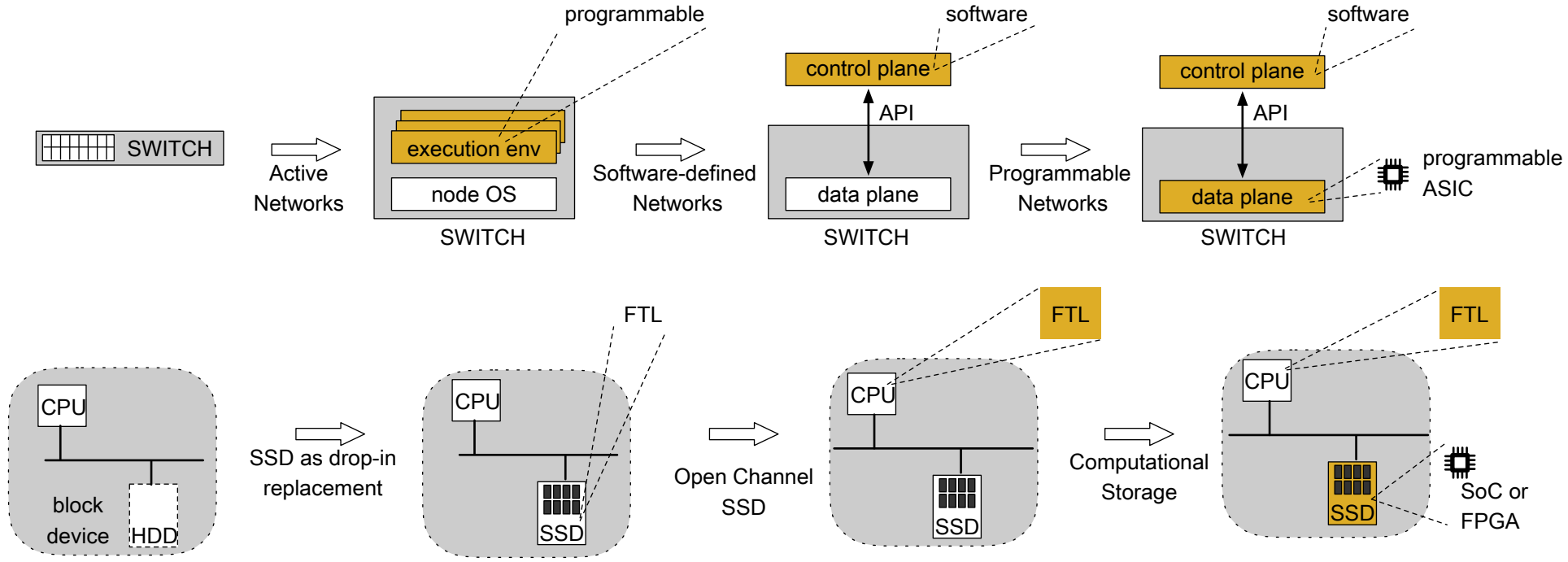
Specialization II

- Different computing units offer different functionalities
- A recent example: the M1 chip from Apple
- Push functionality to units that were “passive” so far
  - No I/O should go untapped!



Why put logic on IO devices?

- They are powerful computations devices
- They have been naturally evolving toward programmability [DEBull'20]



Myths about Programmable Devices

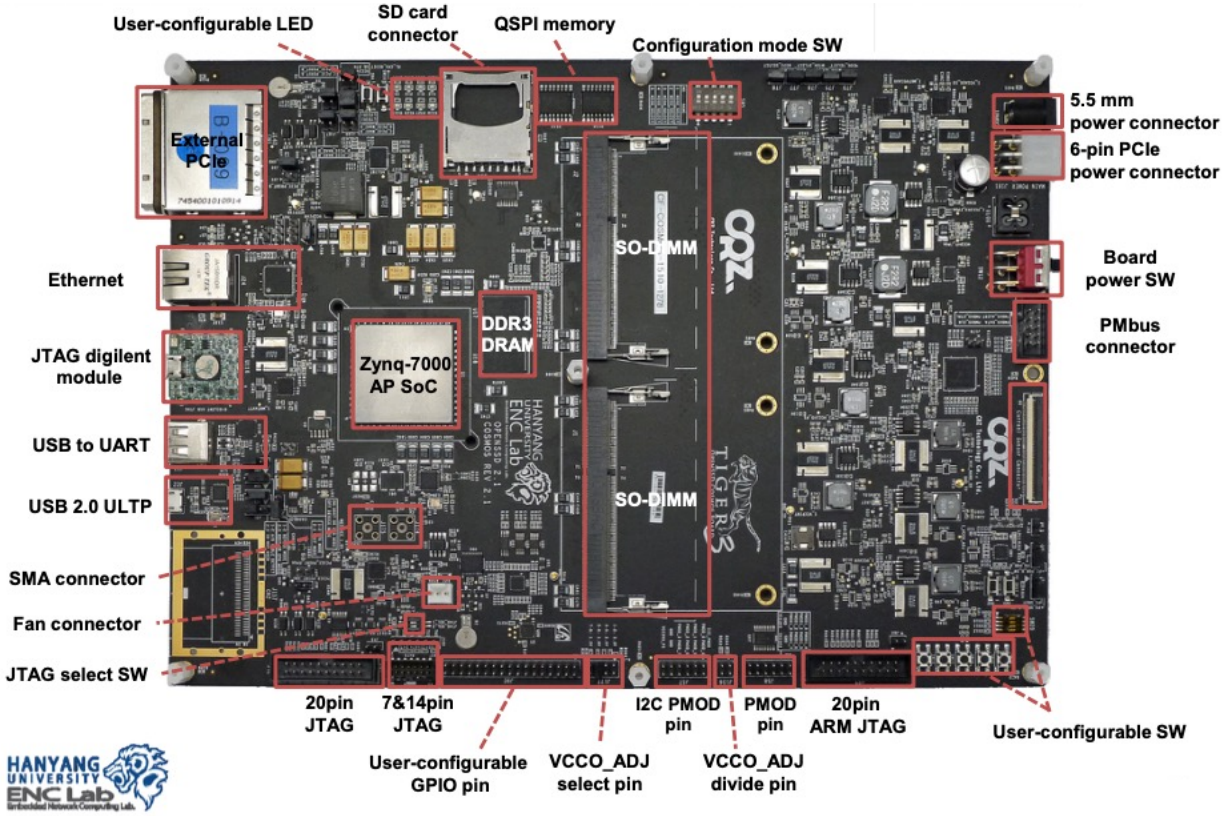
- They are unobtainable
  - Arista and Cisco use programmable switching ASIC
  - We have colleagues that have ZNS SSDs already
- They require hardware engineering (EE vs CS)
  - Many NIC and Switches are fully software programmable
  - Cosmos+ SSD's main logic is in the firmware
- They are not standard
  - P4 is a public language for NICs and switches
  - SNIA computational devices initiative
- Improvements are not portable

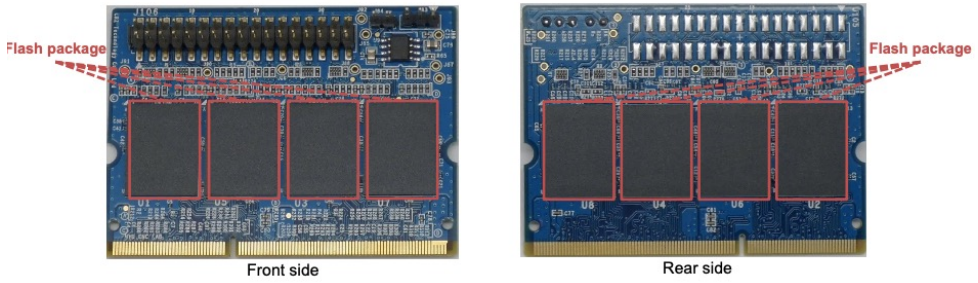
Case I – Performance Counters

Cosmos+ OpenSSD

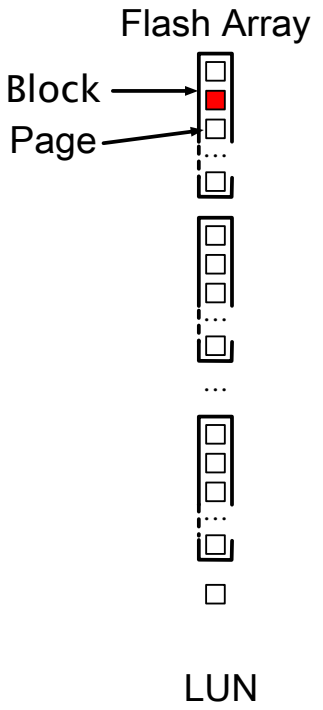
- SSD rapid prototyping platform
- Fully NVMe compatible
- Open-source firmware (C code)
- Next generation is available now

- ***Idea: let's instrument an actual device!*** **[CIDR'20]**

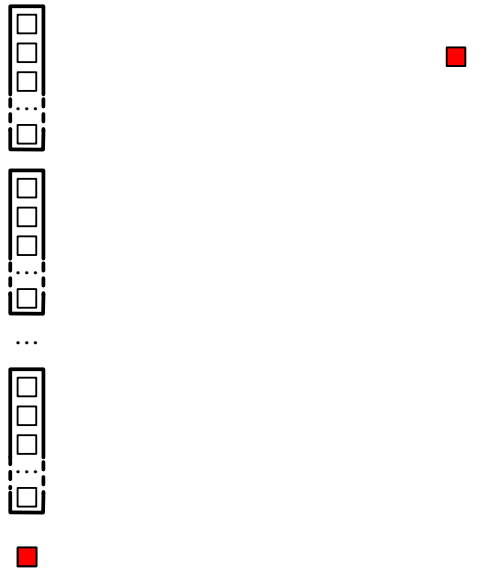




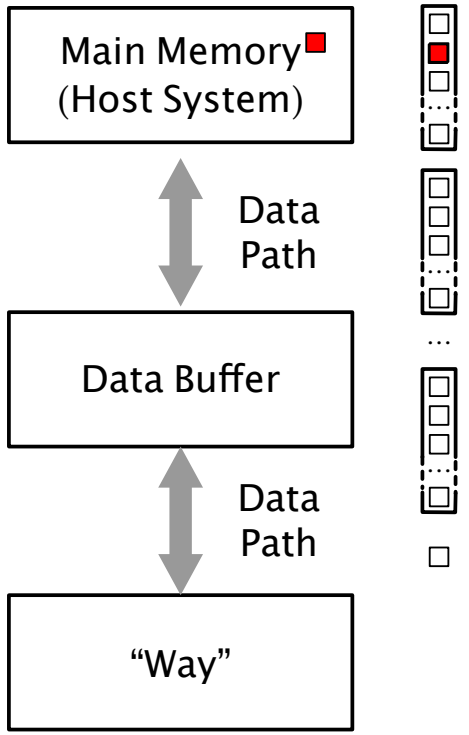
Lifetime of a Write



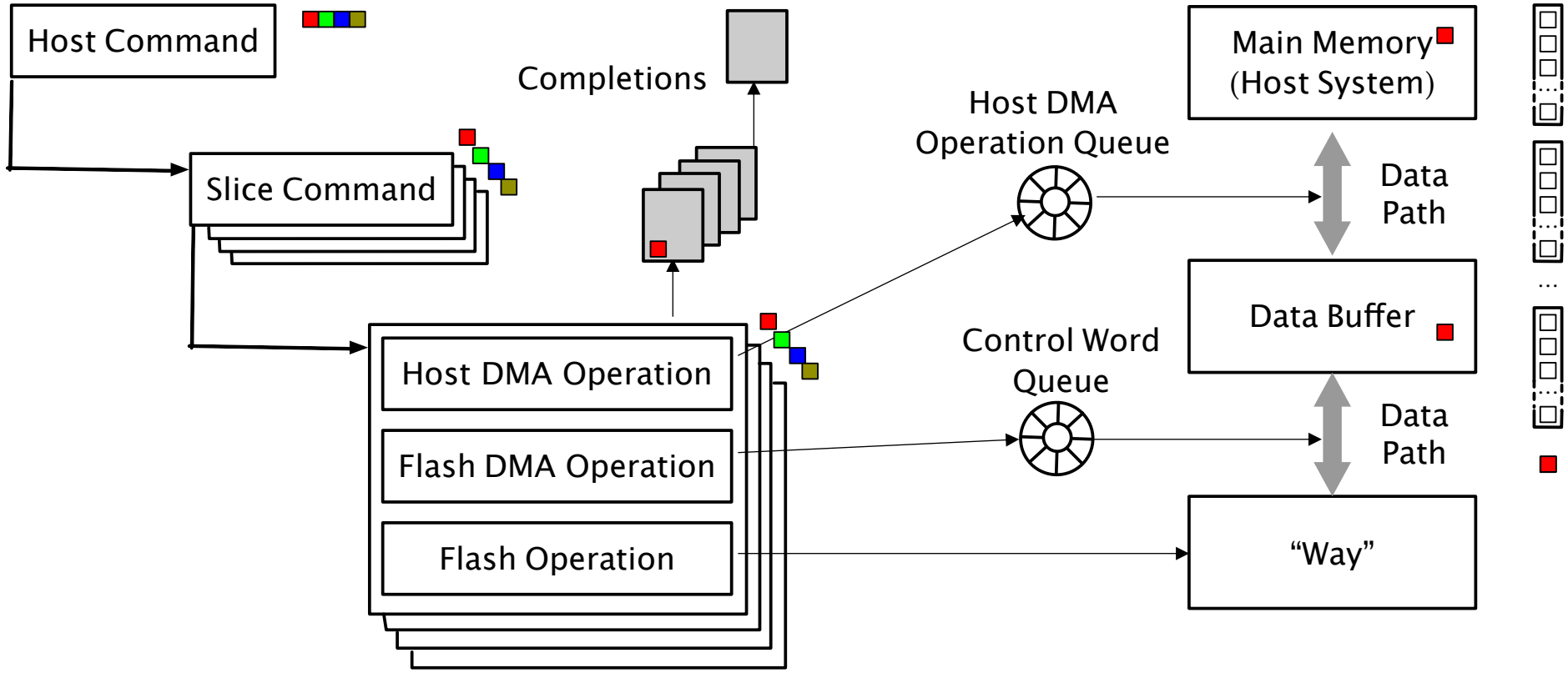
Lifetime of a Write



Lifetime of a Write

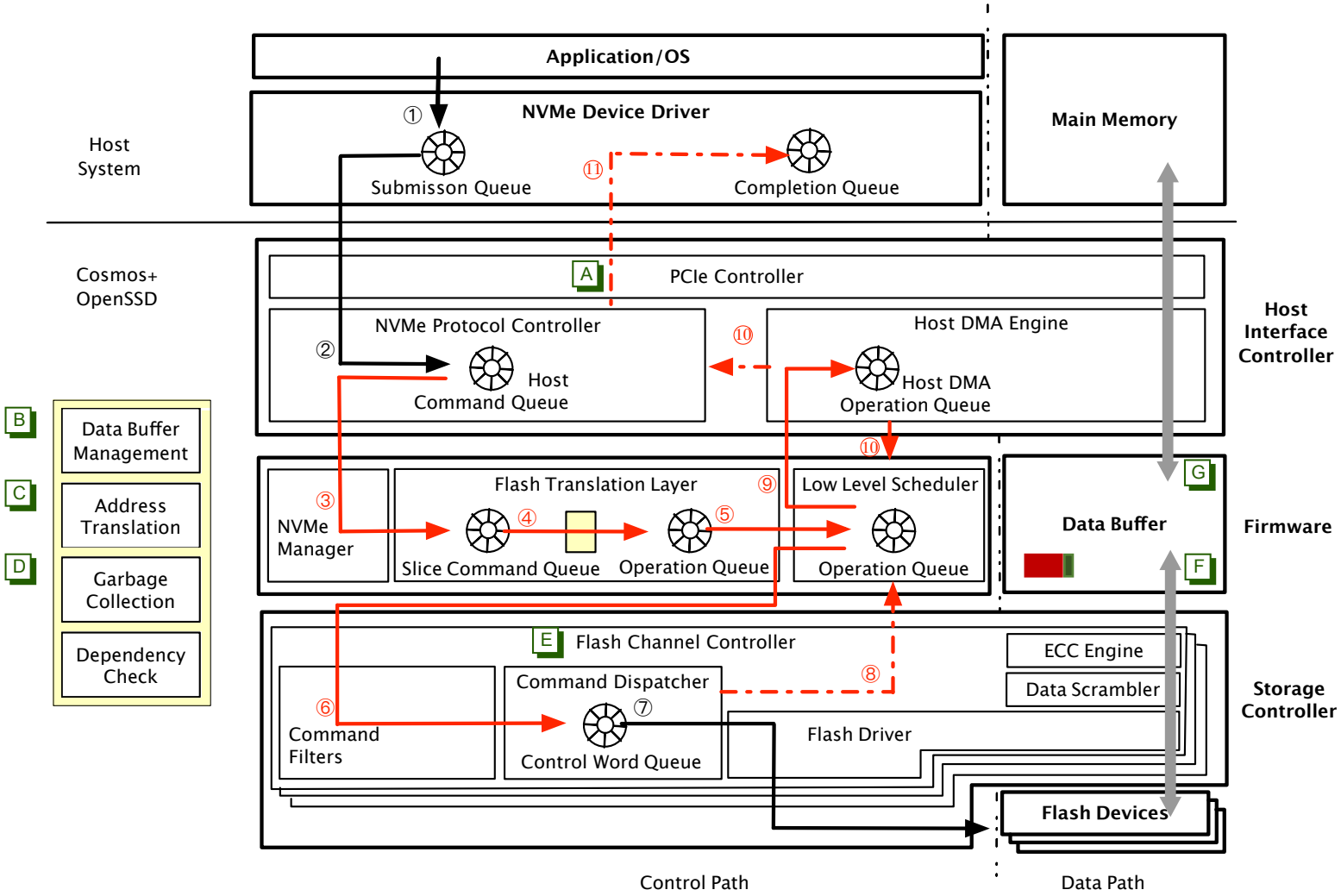


Lifetime of a Write



Instrumentation

- Timestamping
- Counters
- Data extraction commands
- Page map
- Trigger



What have we found?

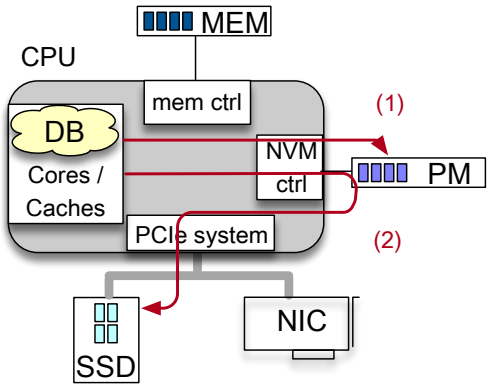
- Running a transaction log along with a snapshot (for checkpointing)
- Mainly three sources of problem
  - Interference
  - Interference
  - Interference

- Read access to a LUN that was writing
- Frame in data buffer “stolen” by heavier workload
- …

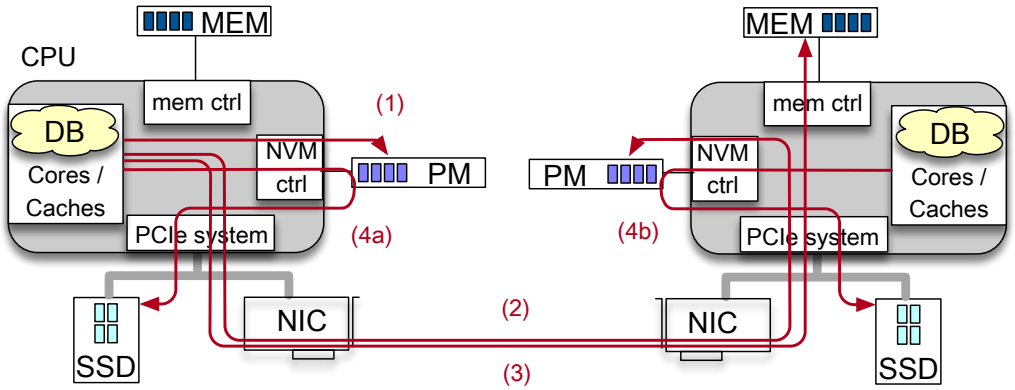
- In general: heavier workload overpowering the lighter one

Case II – Support for Transaction Logging and Replication

Logging to PM



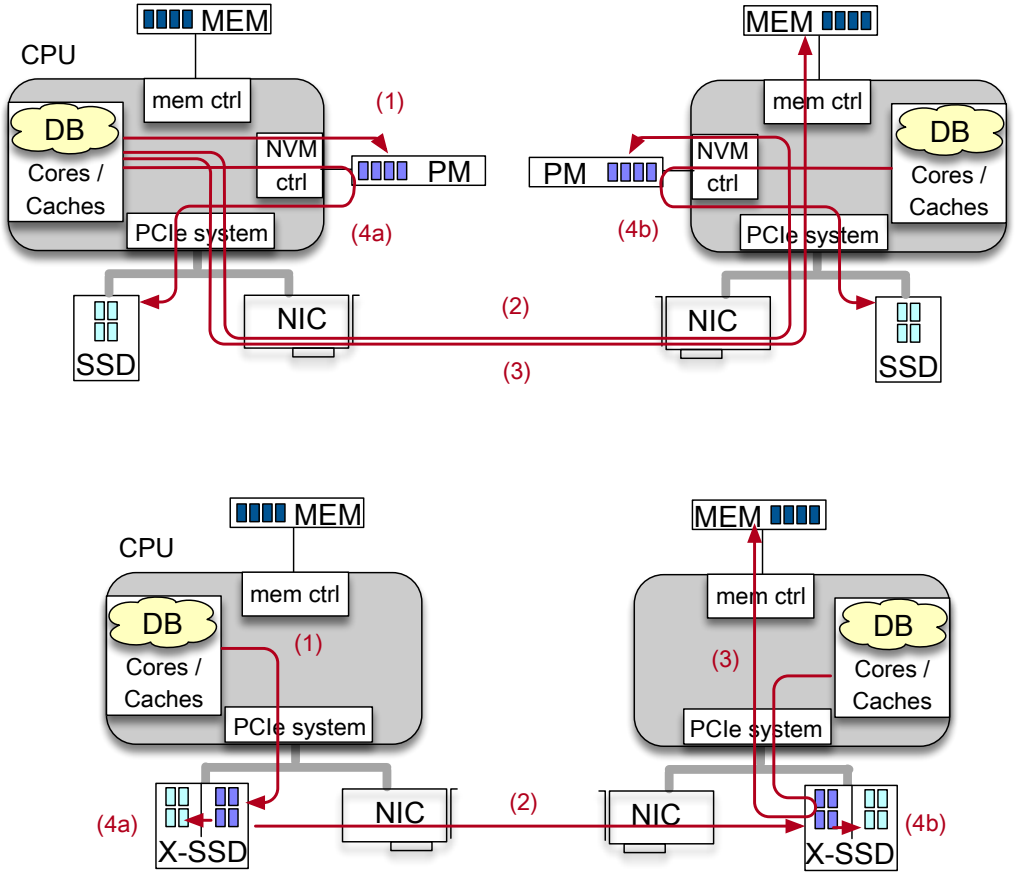
Remote Logging to PM



Issues:

- Portability
- Correctness
- Programming Difficulty

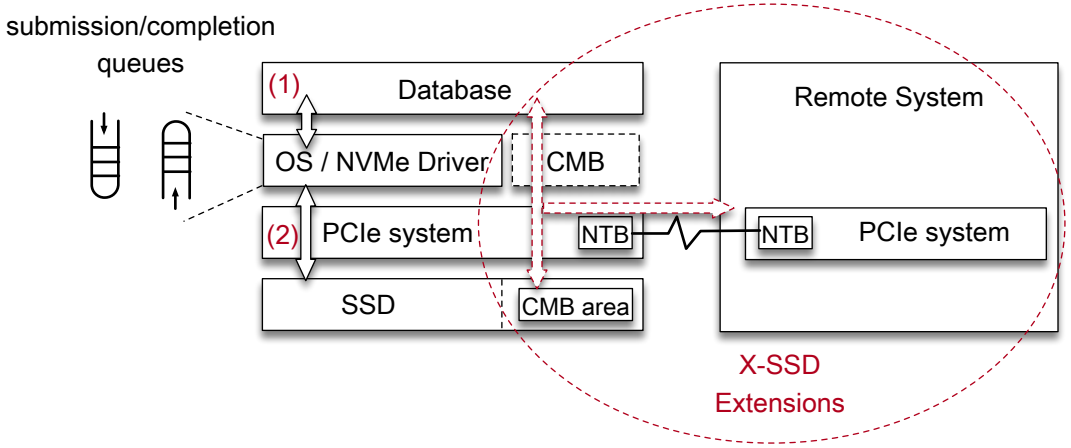
Can we do any better?



Idea:

- Move PM to SSD
- Offer byte addressable interface
- Data propagation services

X-SSD: An NVMe Extension



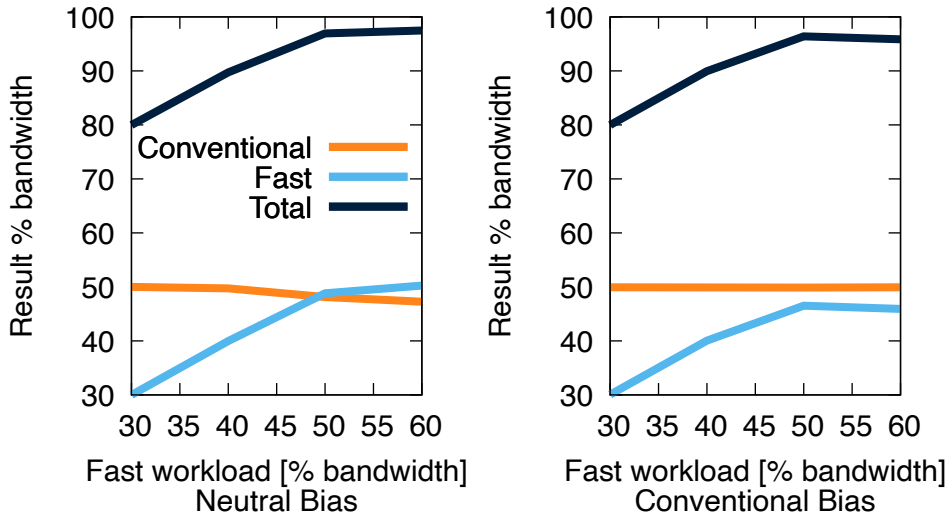
A new type of SSD [SIGMOD'22]

- CMB for user data
- A Transport module
- A Destage module

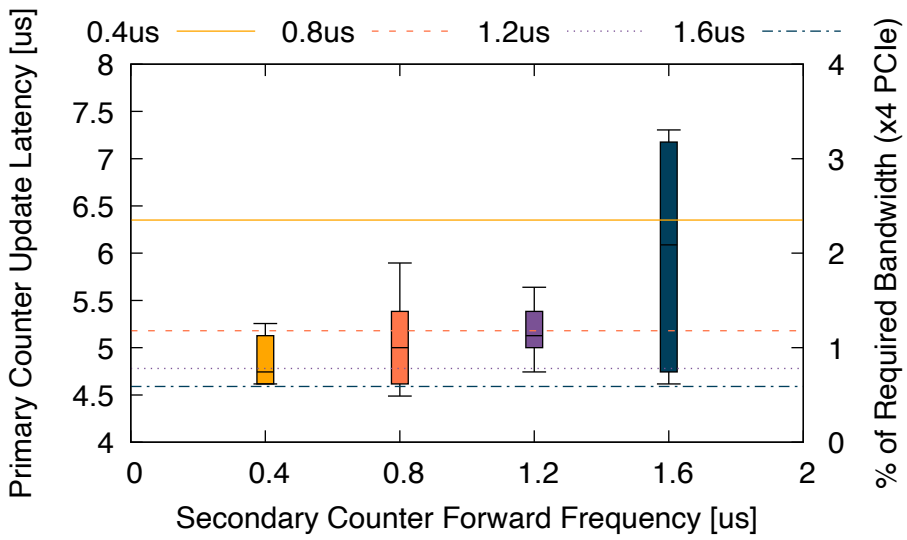
Future

- All module to be programmable

Benefits



Opportunistic Destaging



Low Latency

Conclusion

- Co-designing can be a rich source of performance and “ergonomics” improvements

- Programmable devices are here to stay and are evolving
  - We should present our design requests

References

- Alberto Lerner, Rana Hussein, André Ryser, Sangjin Lee, and Philippe Cudré-Mauroux. “**Networking and Storage: The Next Computing Elements in Exascale Systems?**.” *IEEE Data Engineering Bulletin* 43, no. 1 (March 2020): 60–71.
- Alberto Lerner, Jaewook Kwak, Sangjin Lee, Kibin Park, Yong Ho Song, and Philippe Cudré-Mauroux. “**It Takes Two: Instrumenting the Interaction between In-Memory Databases and Solid-State Drives**.” In *CIDR 2020, 10th Conference on Innovative Data Systems Research*, 2020.
- Sangjin Lee, Alberto Lerner, André Ryser, Kibin Park, Chanyoung Jeon, Jinsub Park, Yong Ho Song, Philippe Cudré-Mauroux, “**X-SSD: A Storage System with Native Support for Database Logging and Replication**.” SIGMOD 2022 (To Appear).