Streamline Data Exchanges & In-Storage Processing
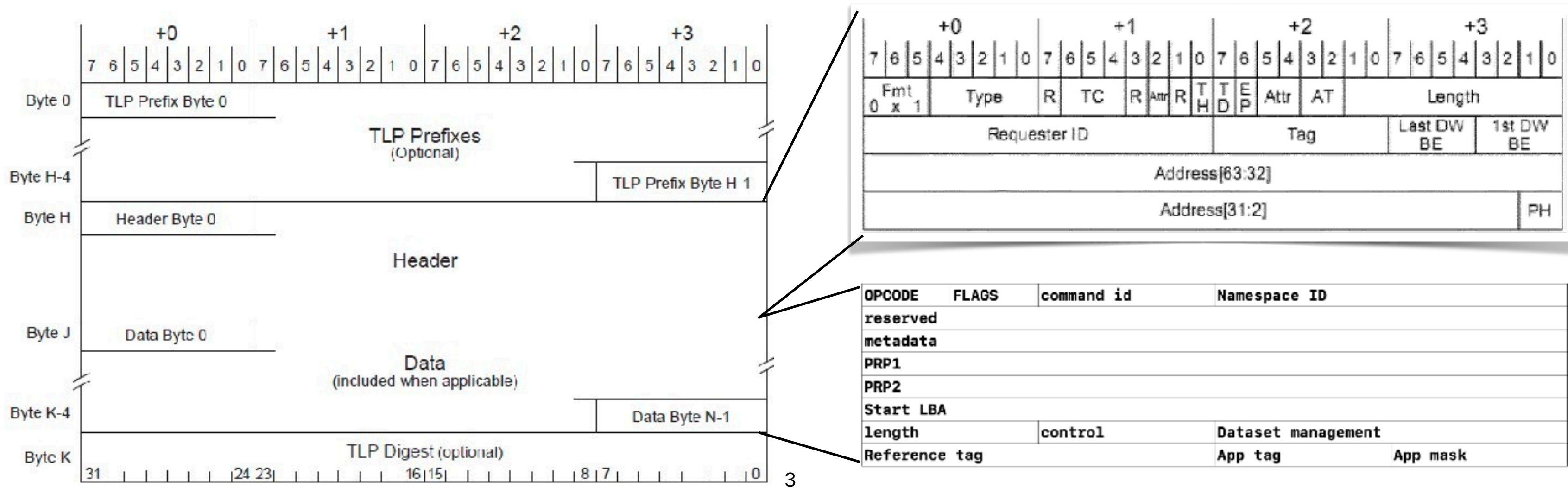
Hung-Wei Tseng

Recap: Garbage Collection in FTL

CPU

Write 0x3241

SSD Controller

M Buffer

Erase

LBA	Flash Block	Flash Page
0x3241	**0**	**2**
0x3242	0	63
0x3243	1	3
0x3244	2	4
0x3245	3	6
0x3246	2	7

invalid page

valid page

free page

Block #0

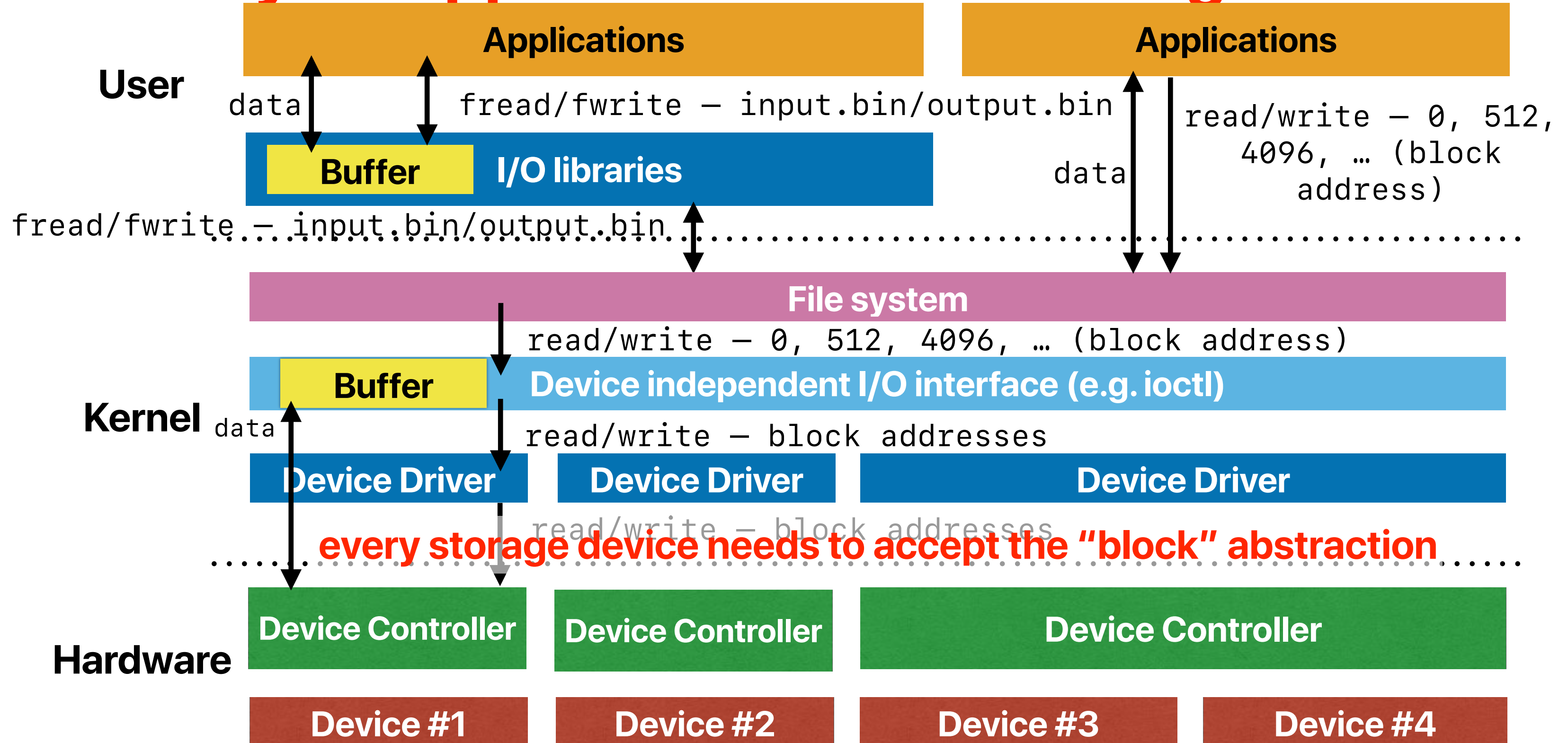Block #1

Block #2

Block #3

Block #4

2
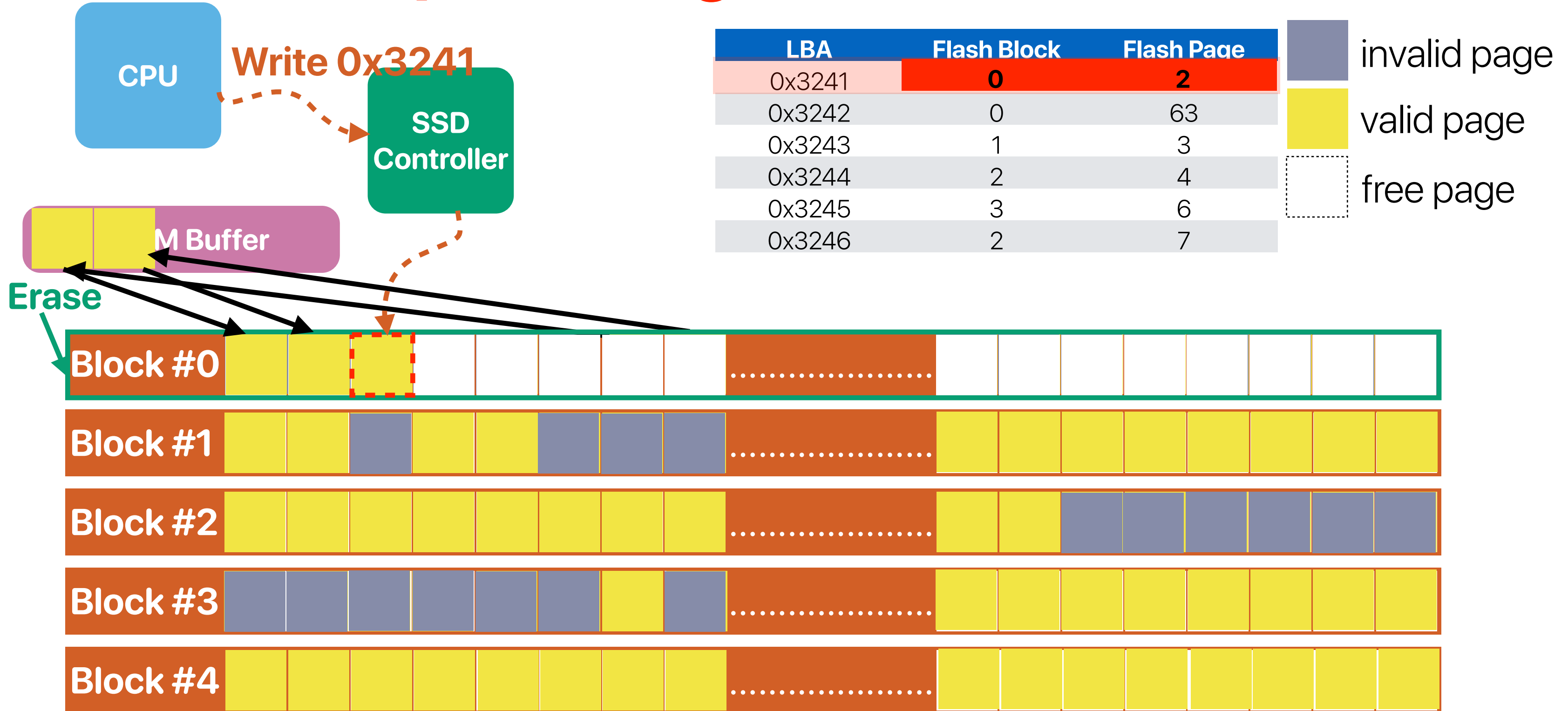
Recap: NVMe is the payload/data of PCIe

- Very similar to computer networks
- Use "memory addresses" as the identifier for routing

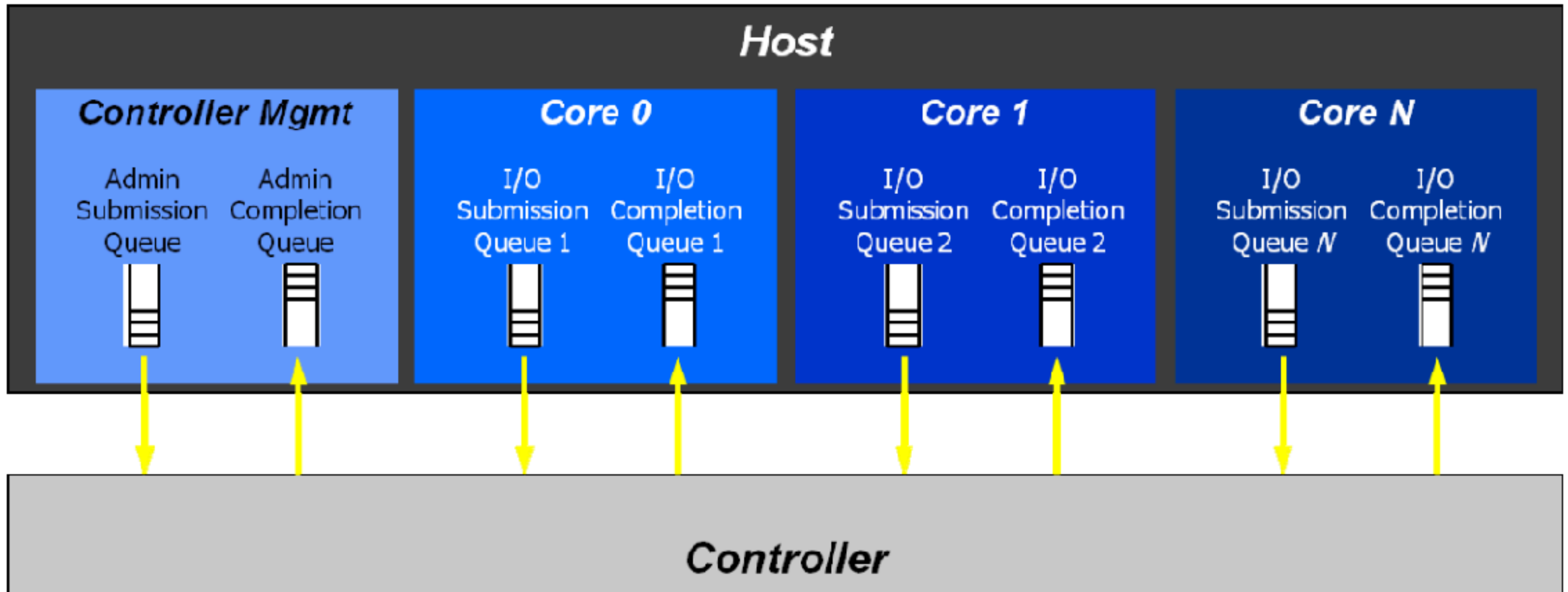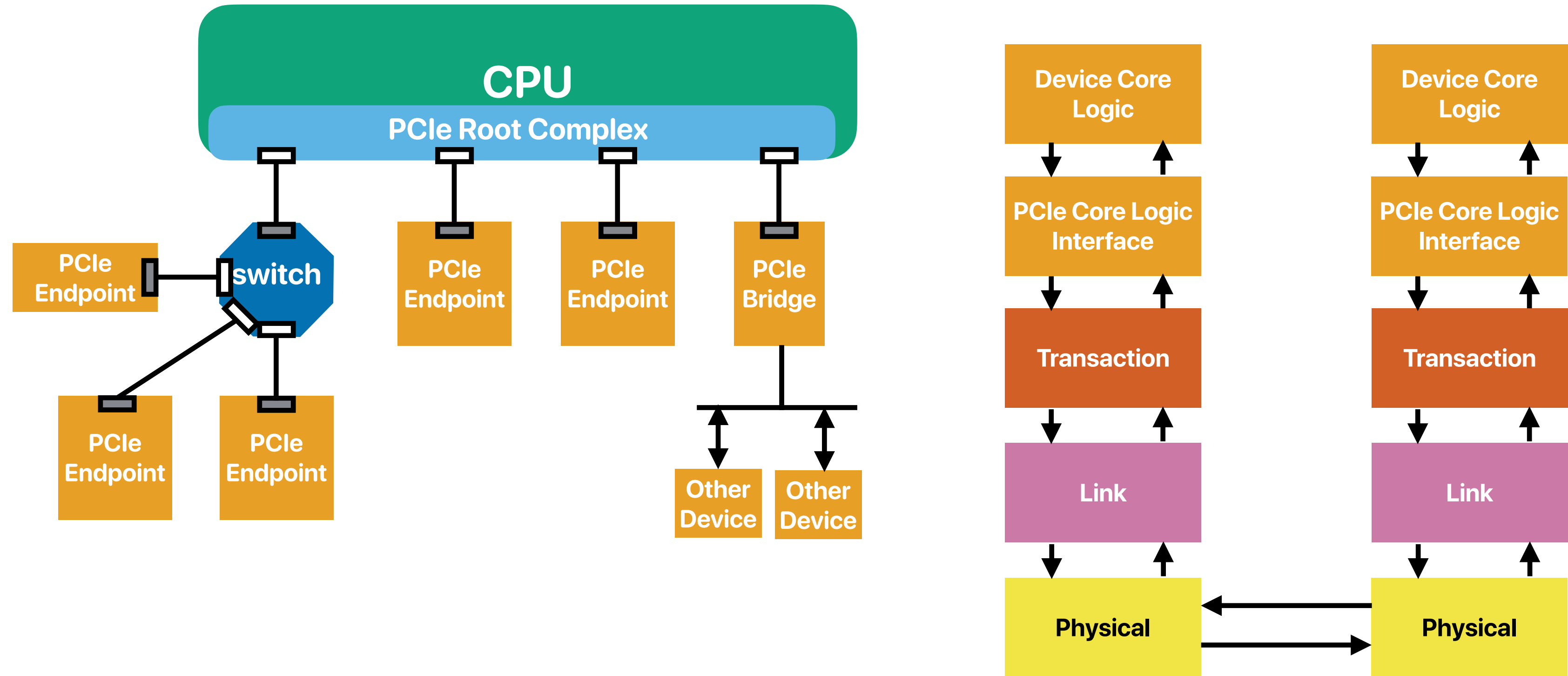# How your application reaches storage devices

User

Applications

Applications

data

fread/fwrite — input.bin/output.bin

read/write — 0, 512, 4096, … (block address)

data

Buffer I/O libraries

fread/fwrite — input.bin/output.bin

File system

read/write — 0, 512, 4096, … (block address)

Buffer Device independent I/O interface (e.g. ioctl)

Kernel data

read/write — block addresses

Device Driver Device Driver Device Driver

read/write — block addresses

every storage device needs to accept the "block" abstraction

Device Controller Device Controller Device Controller

Hardware

Device #1 Device #2 Device #3 Device #4

4

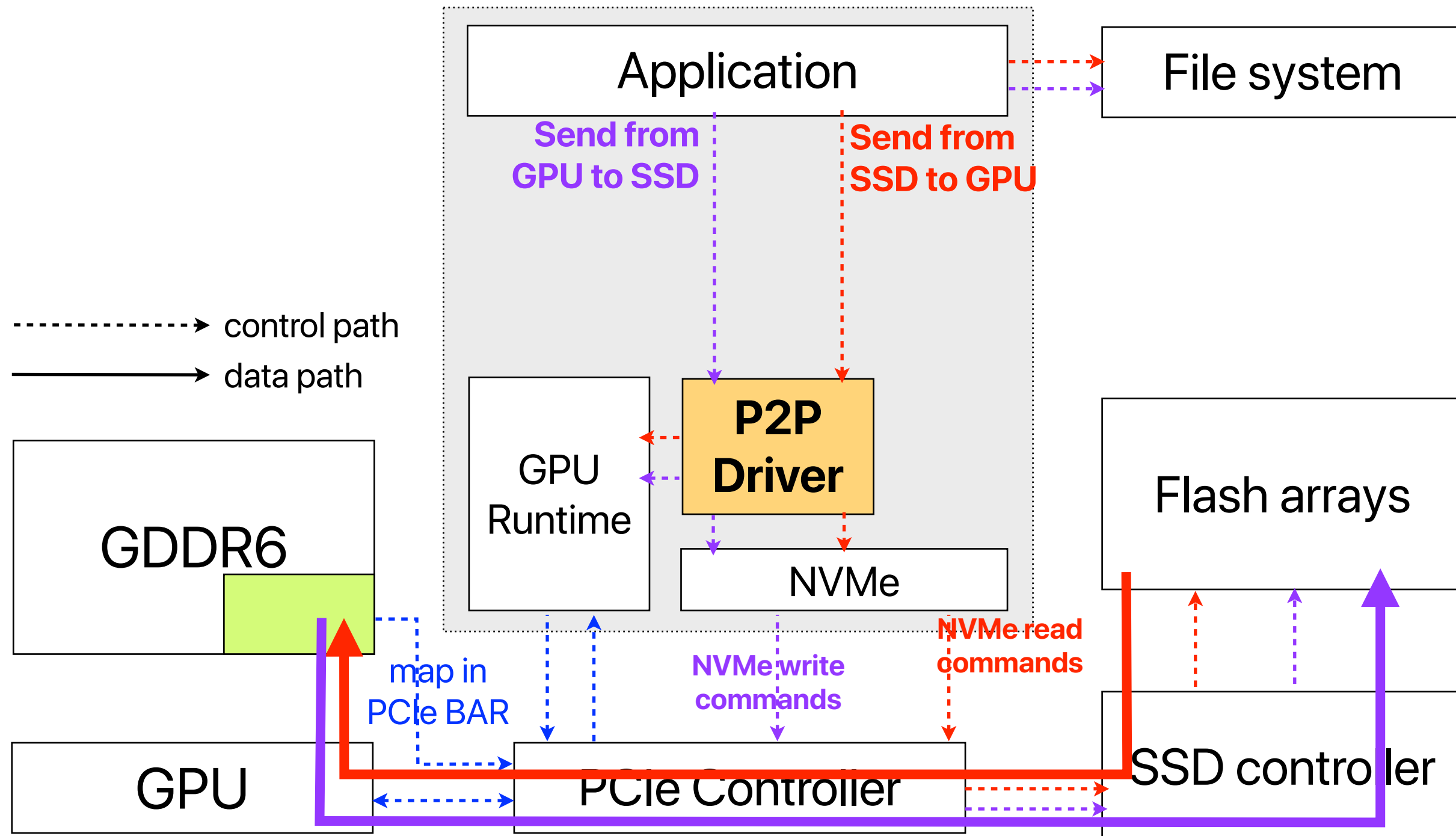# Recap: NVMe Model

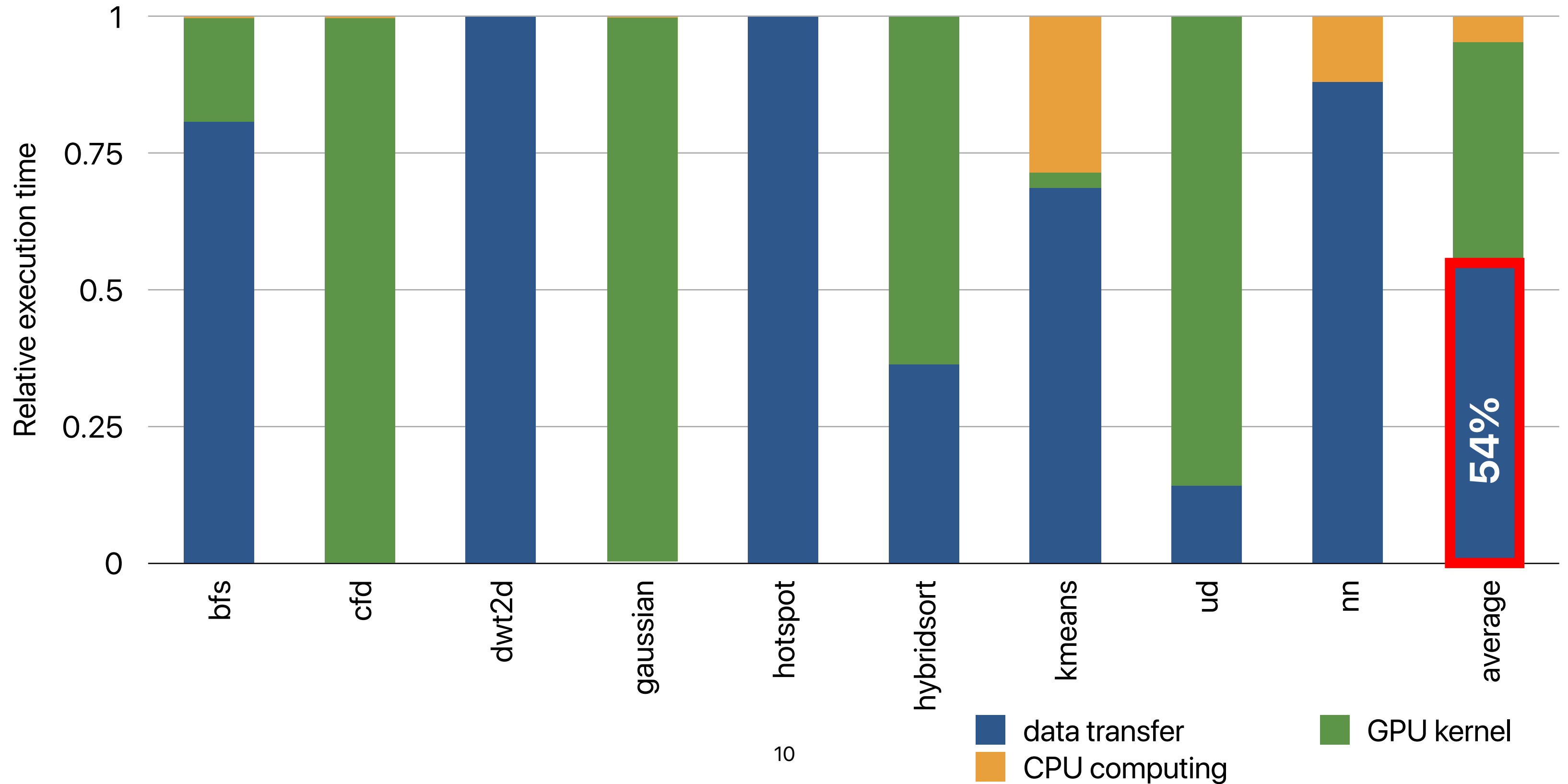Figure 3: Queue Pair Example, 1:1 Mapping

Recap: PCIe "Interconnect"

How can you let SSDs to communicate with GPUs directly & why does SSD-GPU P2P make sense?

How to implement GPU-SSD P2P?

P2P between GPU and SSD

Cost of moving data

Relative execution time

Categories (x-axis): bfs, cfd, dwt2d, gaussian, hotspot, hybridsort, kmeans, ud, nn, average

54%

10

Legend:
- data transfer
- CPU computing
- GPU kernel

GPU-accelerated architecture

GPU

CPU

DRAM

Desired transfer

SSD
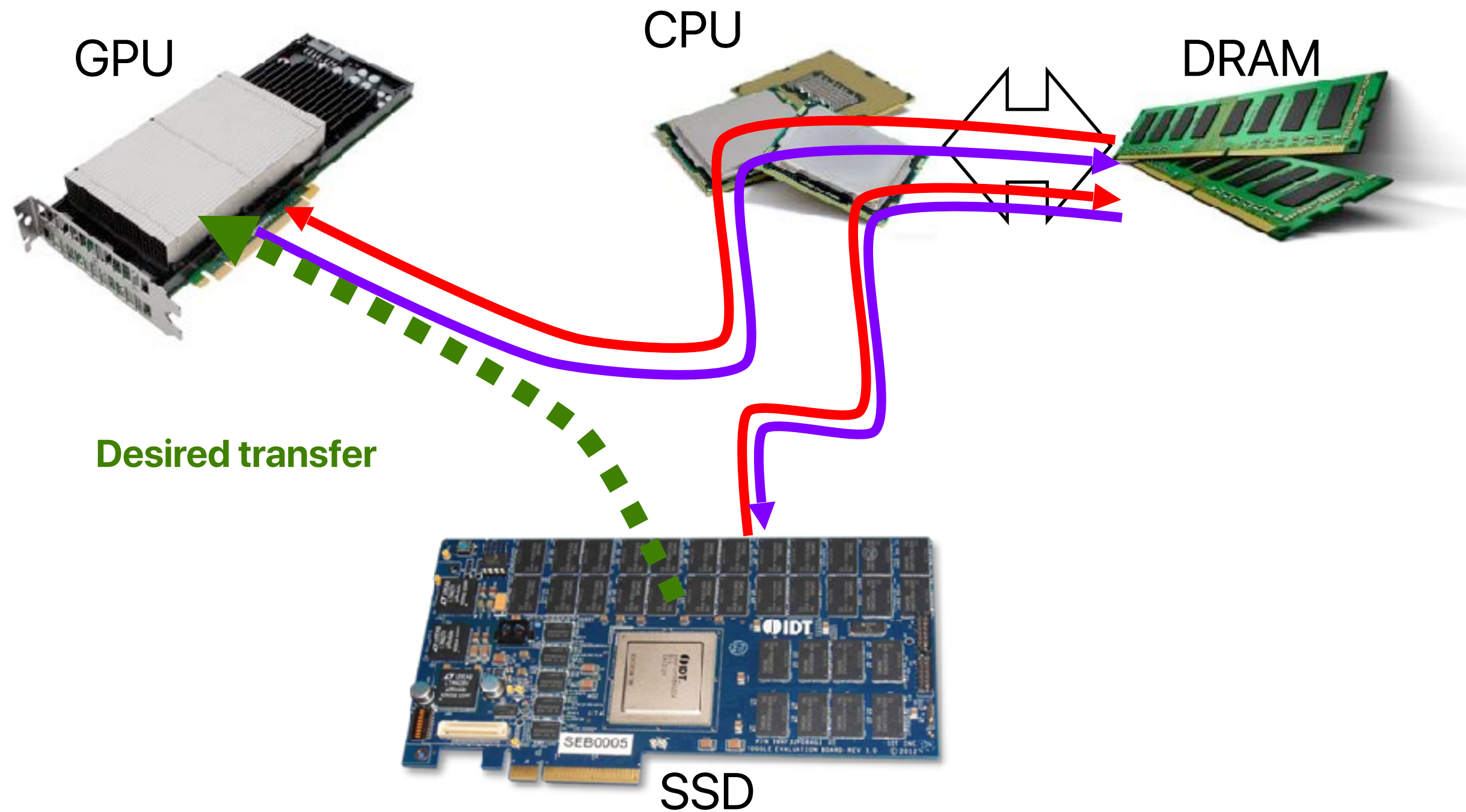
Recap: Multi-threaded I/O helps

Bar chart of Bandwidth (MB/sec) vs. configuration.

Configuration	Bandwidth (MB/sec)
baseline	~580
baseline-2T	~815
baseline-3T	~905
baseline-4_threads	~1045
NVMeD	~815
NVMeD+2threads	~995
NVMeD+3threads	~1050
NVMeD+4_threads	~1125
NVMeD	~885
NVMeD-2T	~1180
NVMeD-3T	~1210
NVMeD-4T	~1230
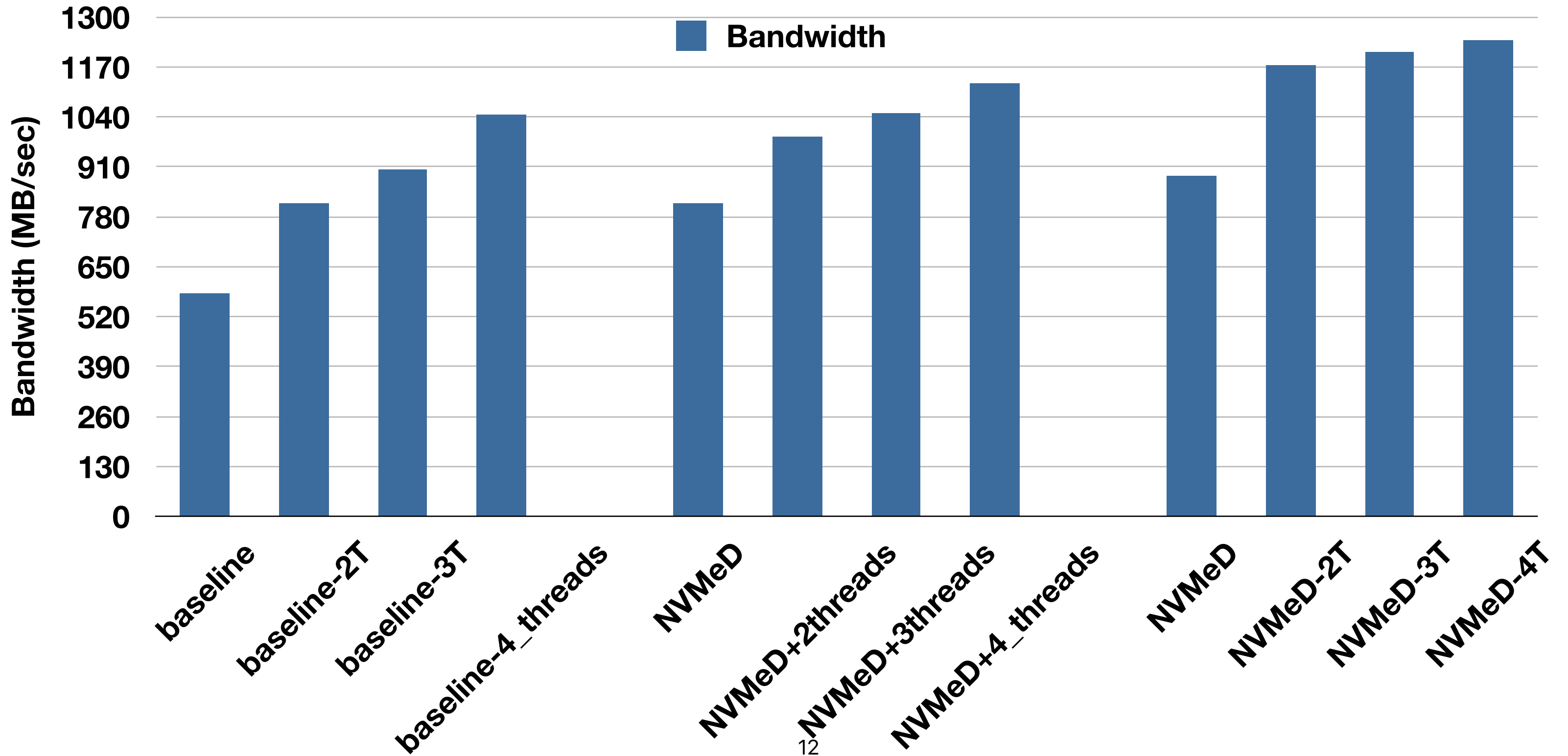
Limited Interconnect Bandwidth

The PCIe Root Complex on a processor has only limited (e.g., 48) PCIe 3.0 lanes

Bandwidth LIMIT 4

Limited Bandwidth

SSD

SSD

SSD

PCI EXPRESS

PCIe Root Complex

intel Xeon processor

CPU

x16 (i.e., 16 GB/sec)

GPU

GPU

x16

x4

System Interconnect (e.g., PCIe)

System Interconnect (e.g., PCIe)

14
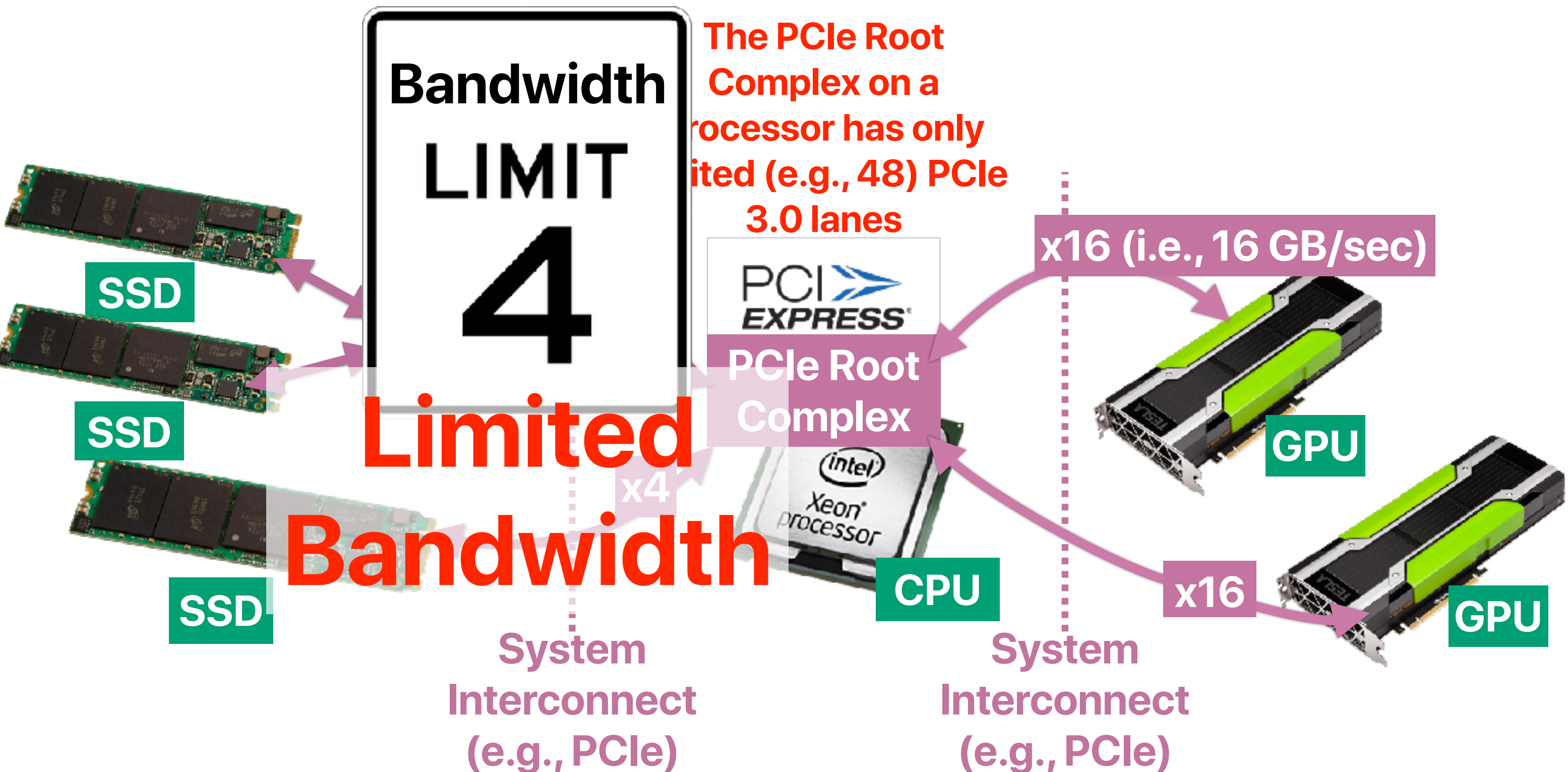
Software overhead

Label	Description	Baseline latency (μs) Write	Read
OS/User	OS and userspace overhead	1.98	1.95
OS/User	Linux block queue and no-op scheduler	2.51	3.74
Schedule	Get request from queue and assign tag	0.44	0.51
Copy	Data copy into DMA buffer	0.24/KB	-
Issue	PIO command writes to Moneta	1.18	1.15
DMA	DMA from host to Moneta buffer	0.93/KB	-
Ring	Data from Moneta buffer to mem ctrl	0.28/KB	-
PCM	4 KB PCM memory access	4.39	5.18
Ring	Data from mem ctrl to Moneta buffer	-	0.43/KB
DMA	DMA from Moneta buffer to host	-	0.65/KB
Wait	Thread sleep during hw	11.8	12.3
Interrupt	Driver interrupt handler	1.10	1.08
Copy	Data copy from DMA buffer	-	0.27/KB
OS/User	OS return and userspace overhead	1.98	1.95
Hardware total for 4 KB (accounting for overlap)		8.2	8.0
Software total for 4 KB (accounting for overlap)		13.3	12.2
File system additional overhead		5.8	4.2
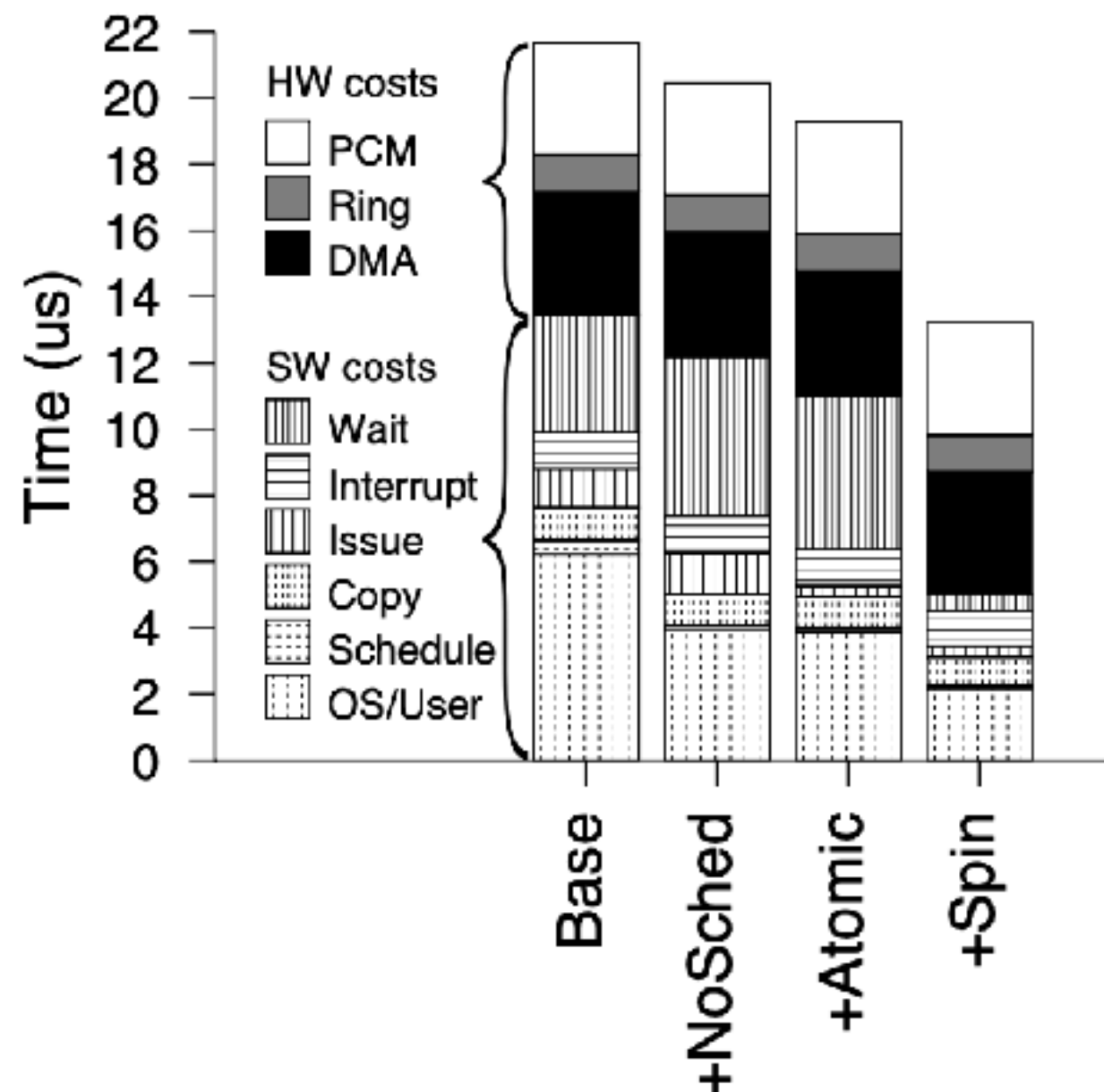
A. M. Caulfield, A. De, J. Coburn, T. I. Mollow, R. K. Gupta and S. Swanson, "Moneta: A High-Performance Storage Array Architecture for Next-Generation, Non-volatile Memories," 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture, 2010, pp. 385-395, doi: 10.1109/MICRO.2010.33.

15

Recap: Think different

Near-Data Processing — having processing power near data

DDR4
26 GB/sec

Processor

CPU-Memory BUS

DDR4

Memory Controller

DRAM

CPU

intel Xeon processor

PCI EXPRESS

1 GB/sec

PCIe Root Complex

x4
4GB/sec

Processor

SSD

22 GB/sec

GPU

105 GB/sec

TPU

ALVEO

FPGA

Movidius MA2485 Myriad X

XPU

Processor

NIC

16
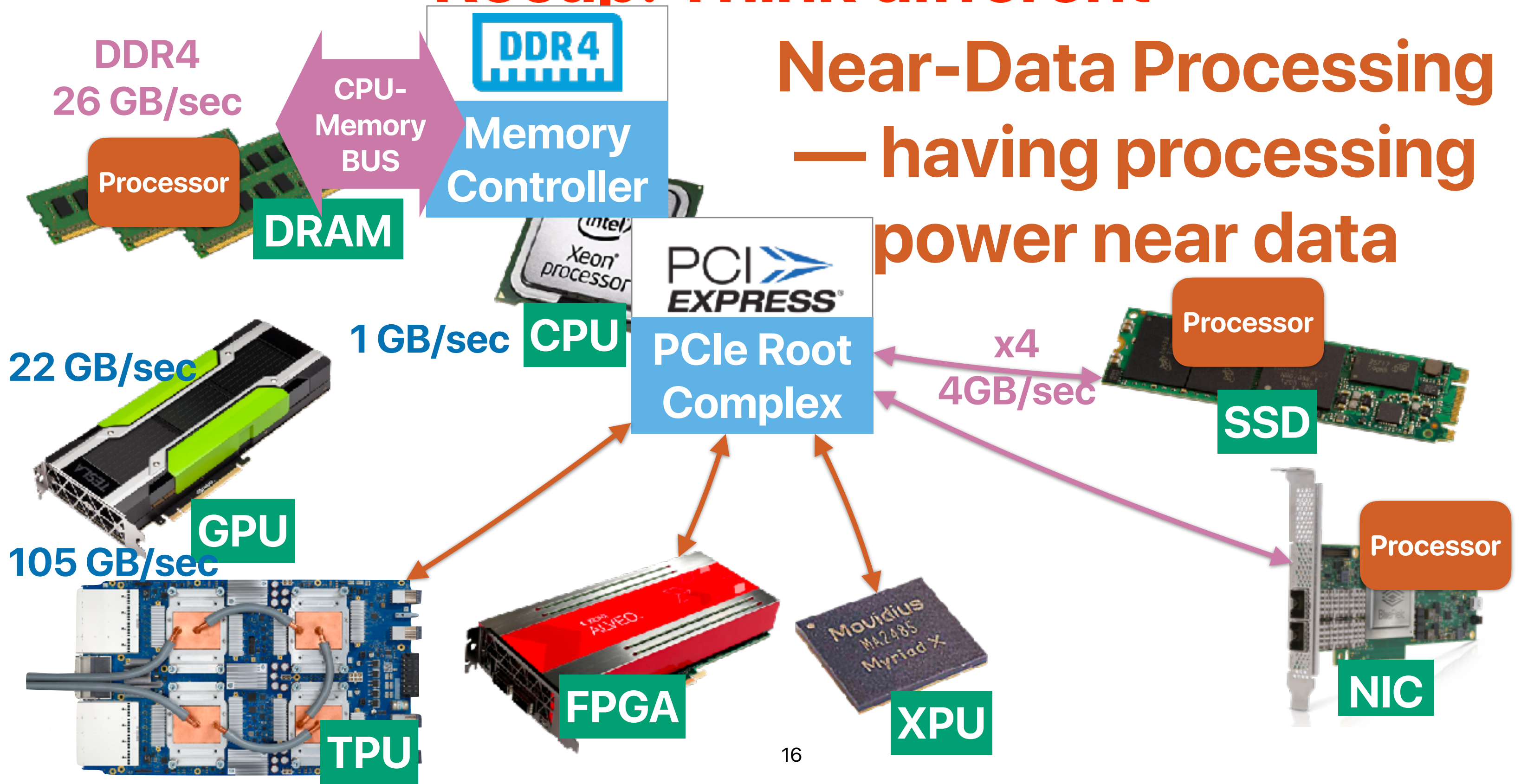
Where do you see the opportunity of in-storage processing in address the bandwidth limitation and software overhead?

Announcement

- What to expect in the final presentation
  - Brief recap of the why — 3-5 minutes
  - What has been tried — 5-7 minutes
  - What concrete experimental results have been measured — 5-7 minutes
    - You need to have convincing experimental setup
    - You need to have several result figures (i.e., bar charts)
  - What insights have you learned that would guide future research — 3-5 minutes

Electrical

Computer Science

Engineering

つづく