

Unblinding the OS to Optimize User-Perceived Flash SSD Latency

Woong Shin^{*}, Jaehyun Park^{**}, Heon Y. Yeom^{*}

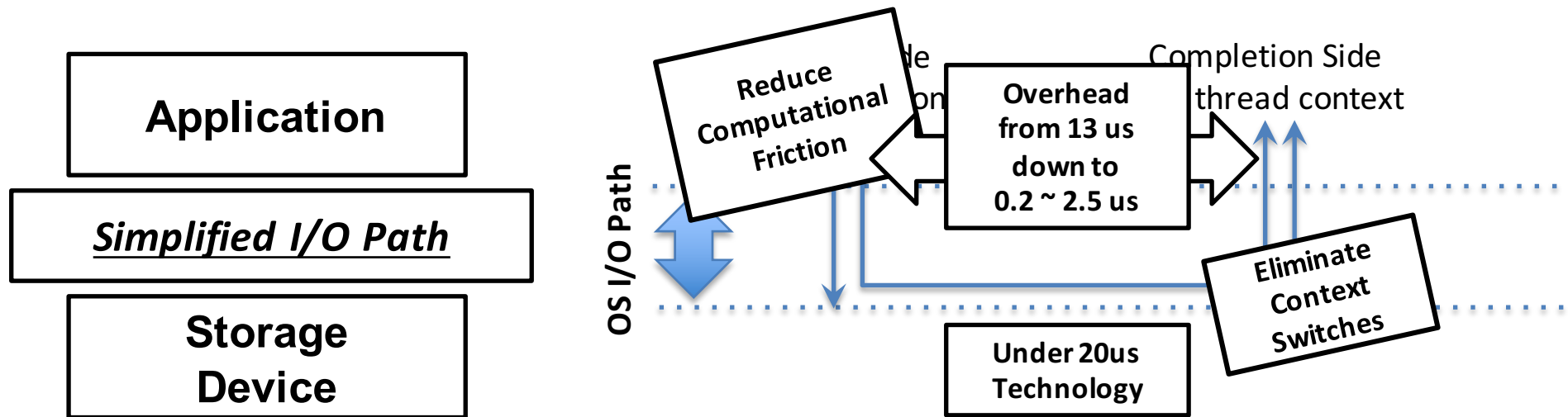
^{*}Seoul National University

^{**}Arizona State University

USENIX HotStorage 2016

Jun. 21, 2016

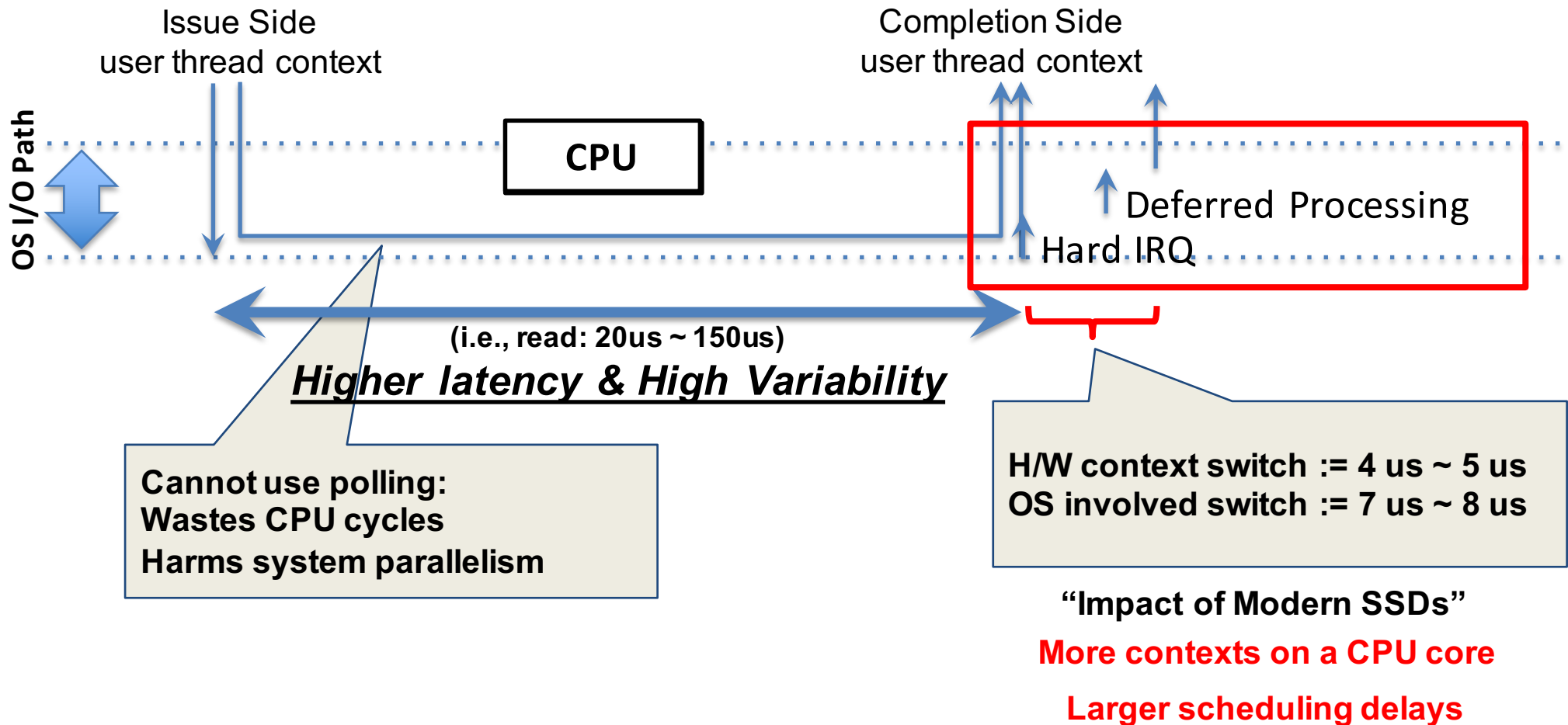
OS I/O Path Optimizations: Reducing S/W Overheads



Memory technology with ultra low (nanoseconds), predictable latencies.

“Block CPU (poll) while waiting for the I/O to complete”

To Block the CPU (sync) or to Yield the CPU (async)



Impact of Modern SSDs

700,000 IOPS NVM-e SSD

Bandwidth:

$$\underline{4\text{kB} \times 700 \text{ kIOPS} = 2.8 \text{ GB/s}}$$

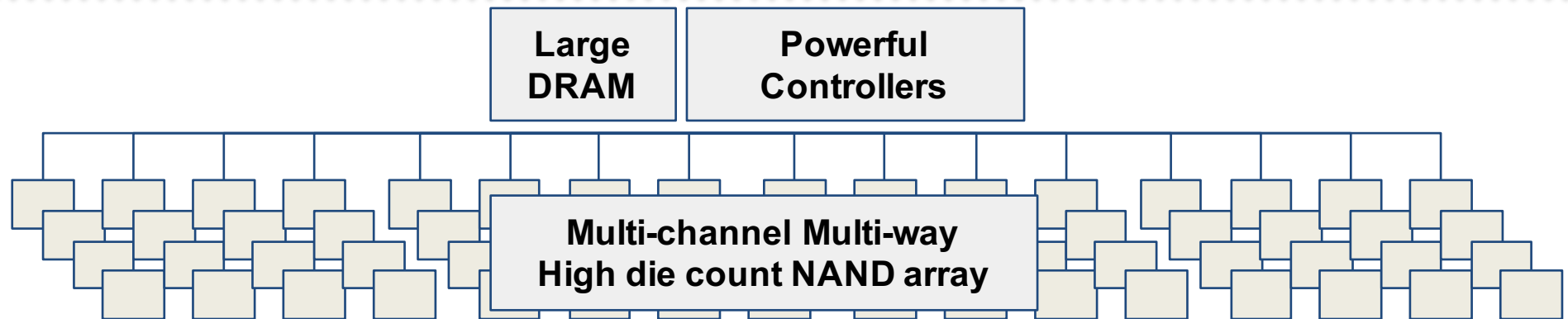
Latency:

~~$$1 \text{ sec} / 700 \text{ kIOPS} = 1.42 \mu\text{s}?!?$$~~

Impact of Modern SSDs

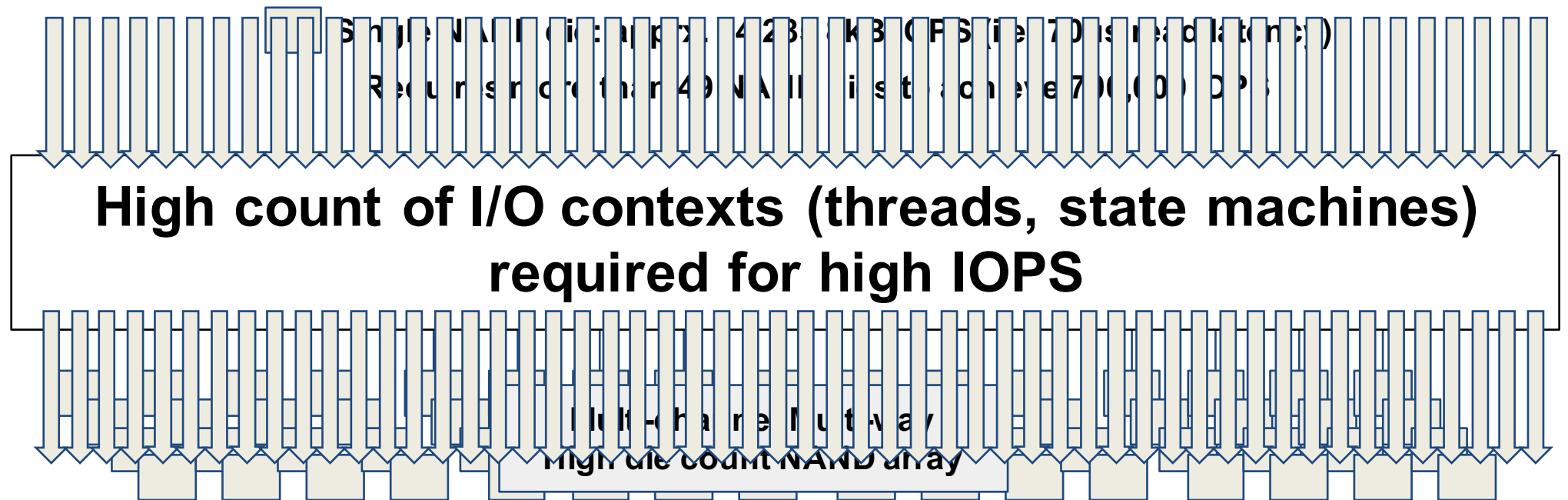
700,000 IOPS NVM-e SSD

Single NAND die: apprx. 14,285 8kB IOPS (i.e, 70us read latency)
Requires more than 49 NAND dies to achieve 700,000 IOPS

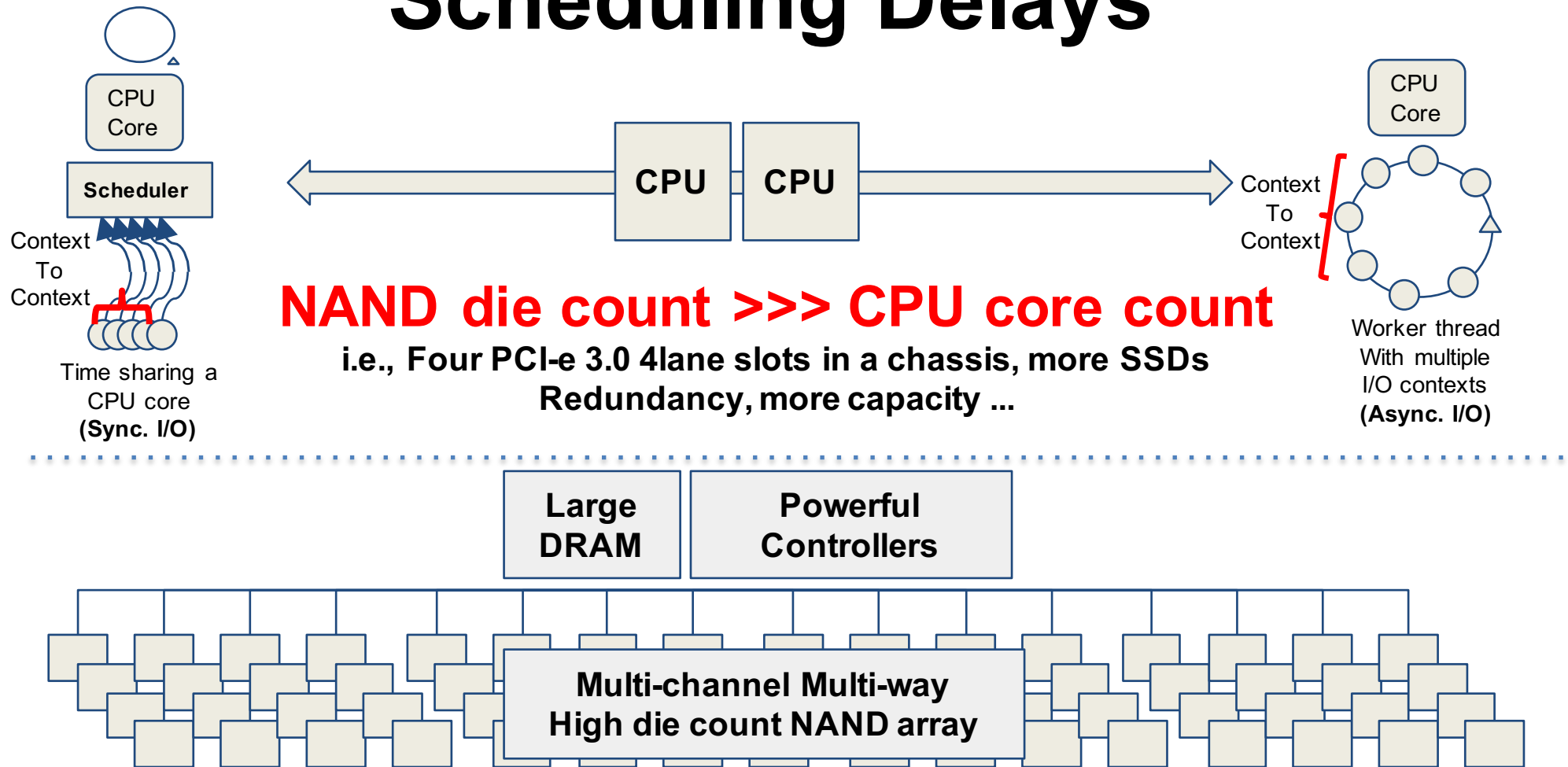


Impact of Modern SSDs

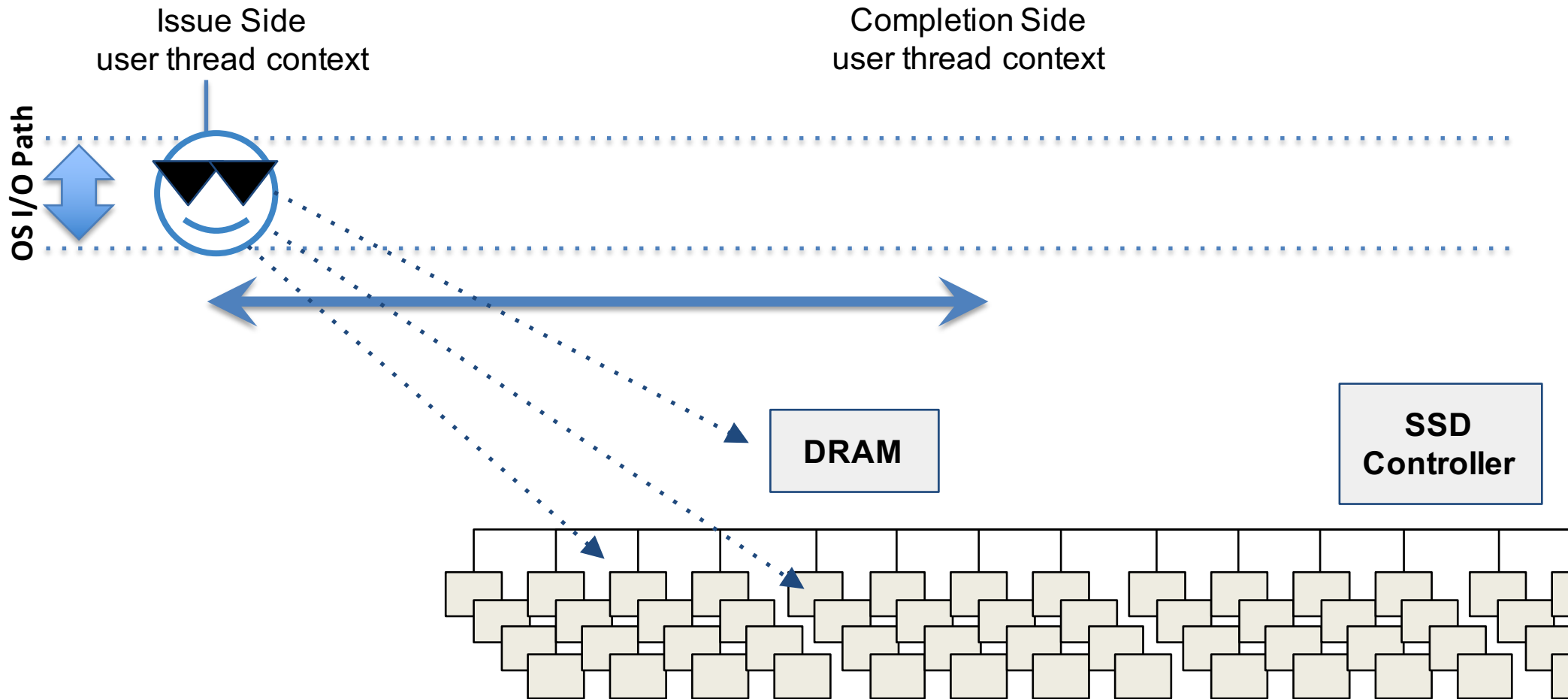
700,000 IOPS NVM-e SSD



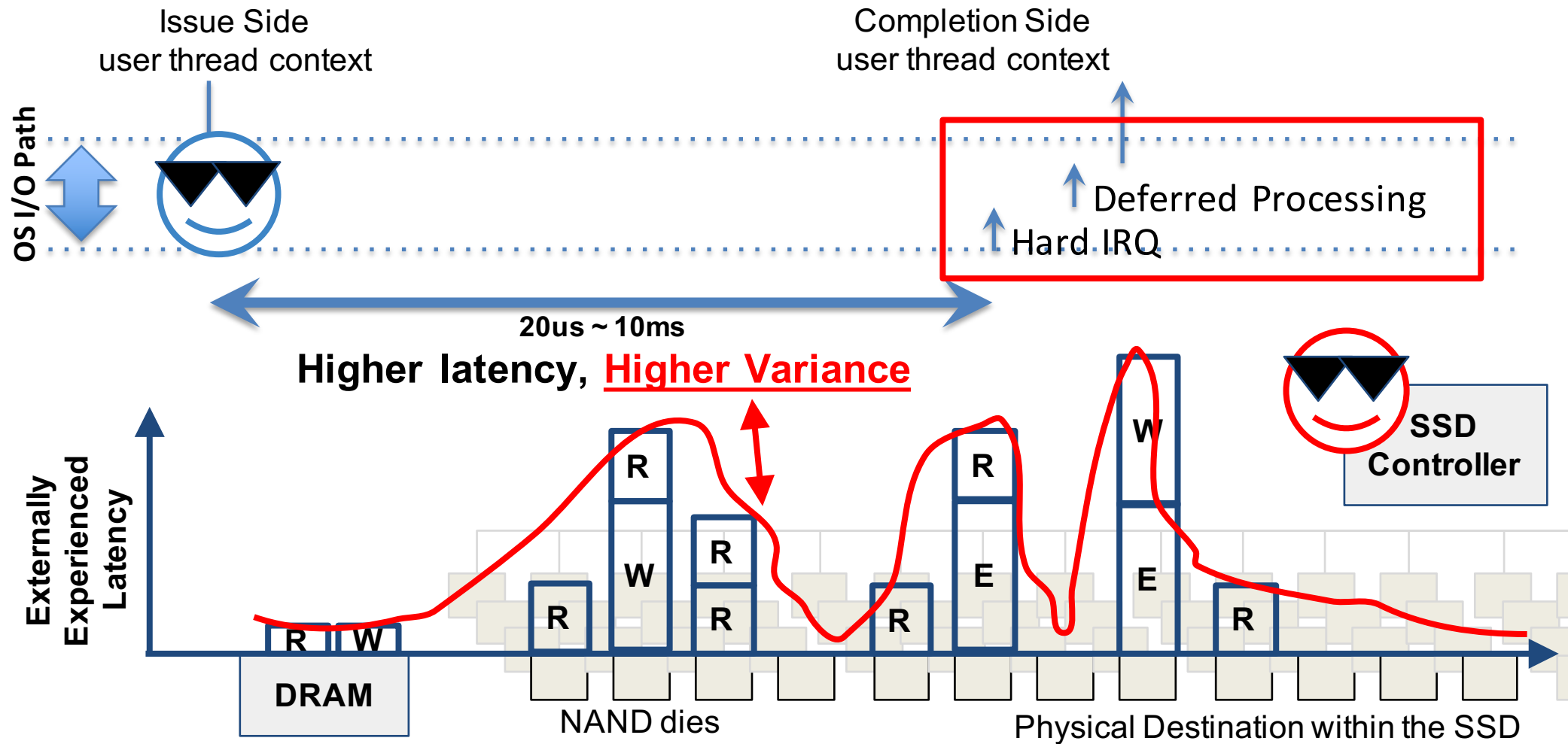
Higher Context Multiplexing Cost “Scheduling Delays”



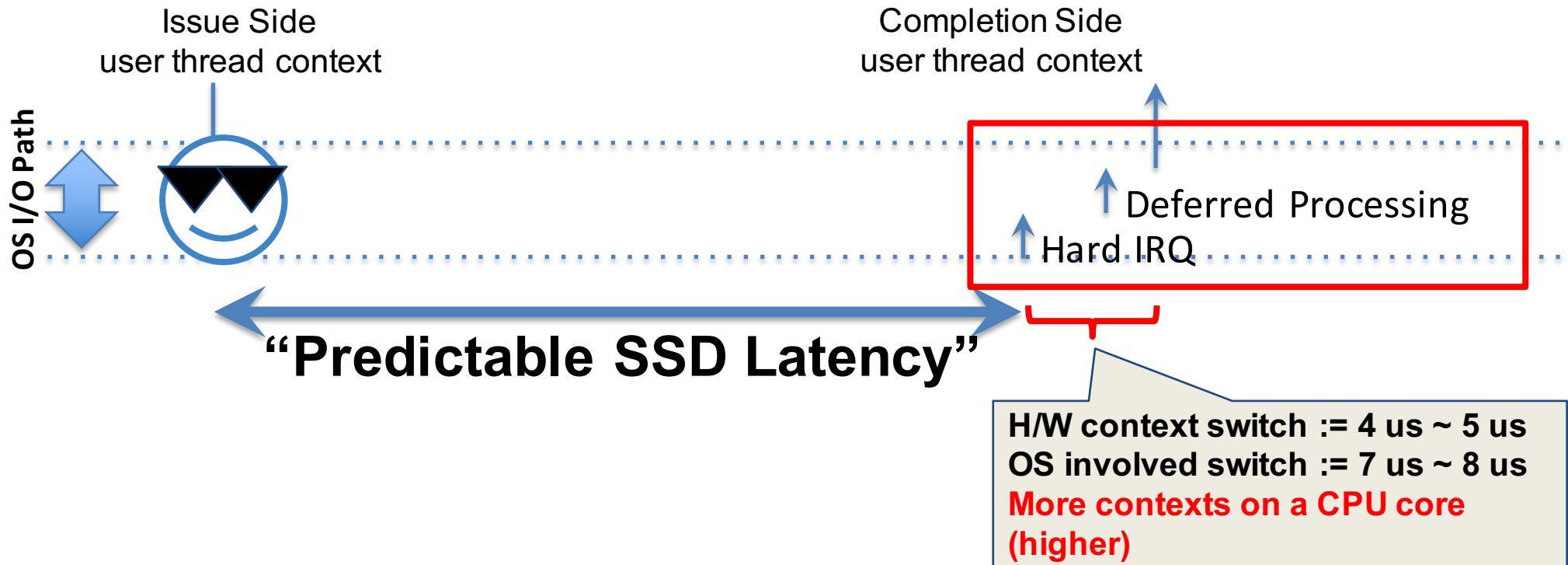
The OS is Blind: Conservative Strategies



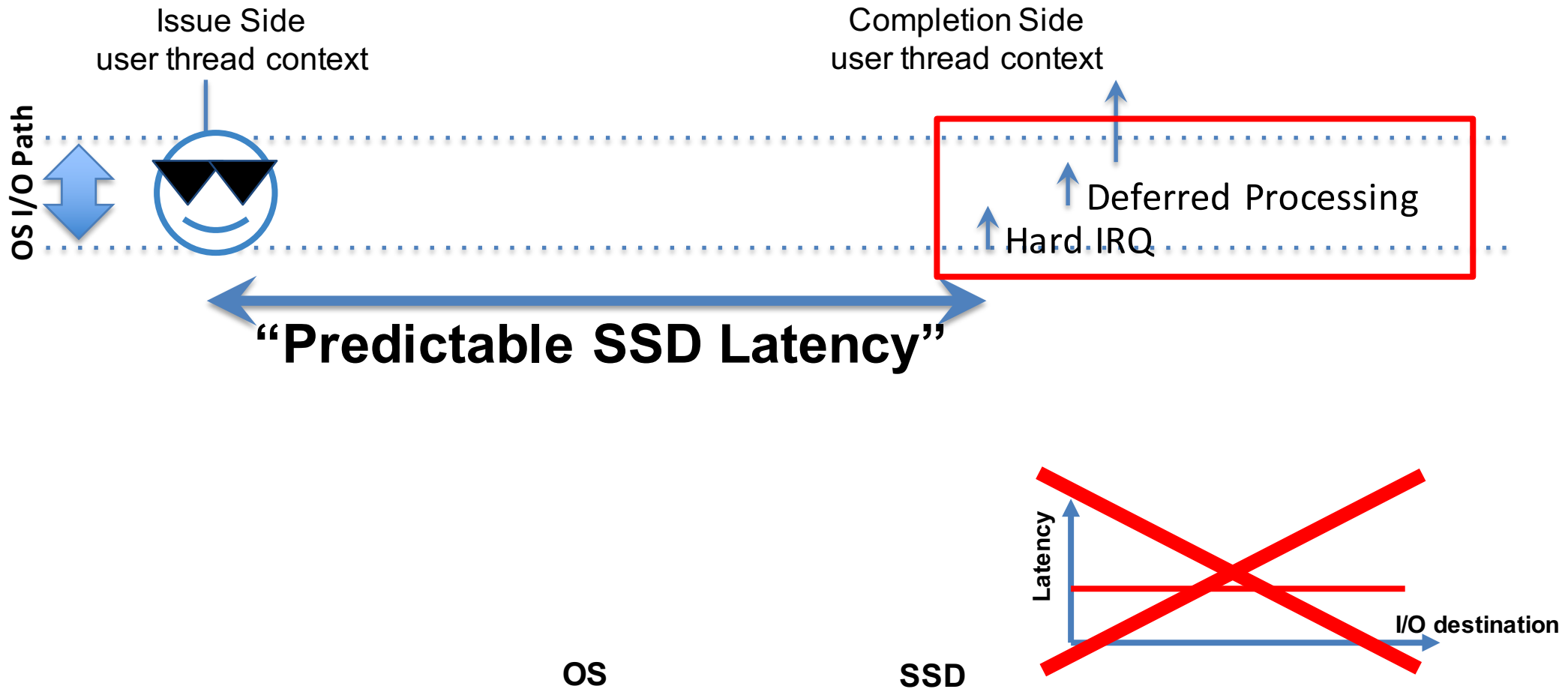
The OS is Blind: Conservative Strategies



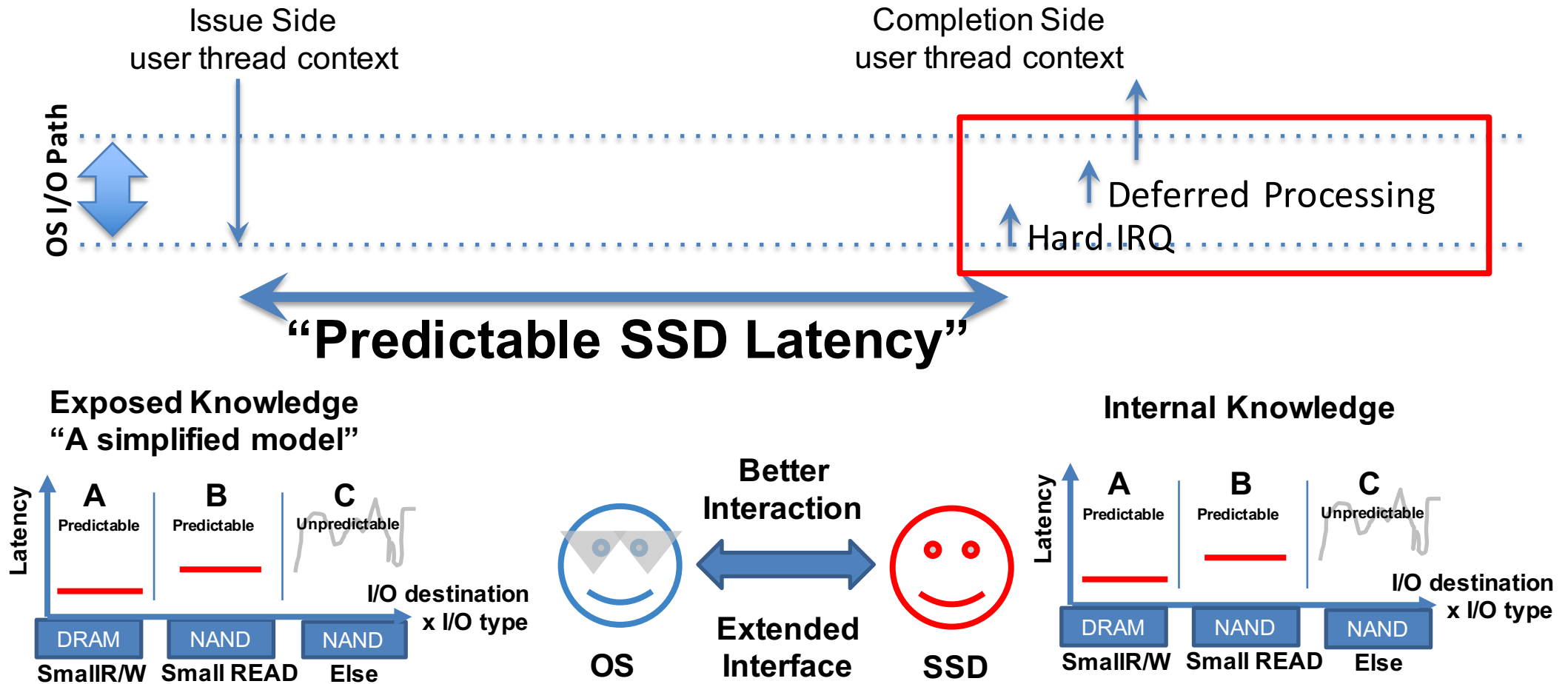
This Work: Unblinding the OS



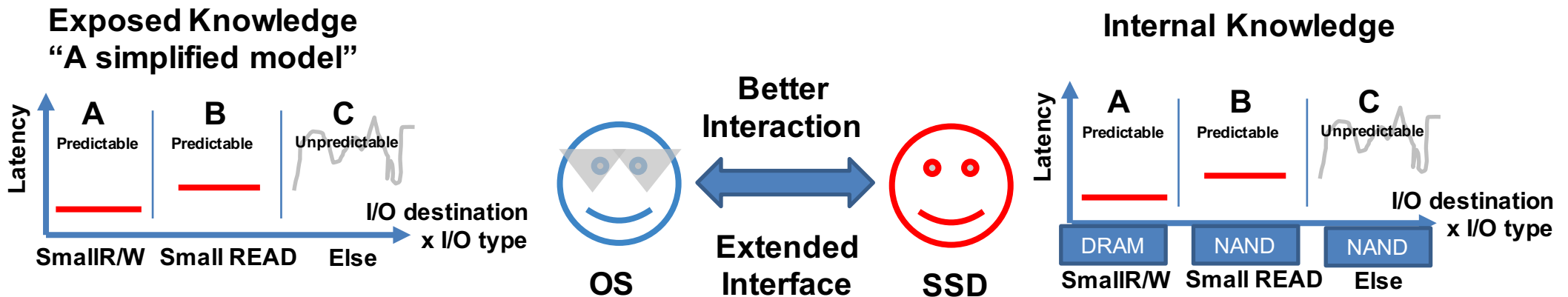
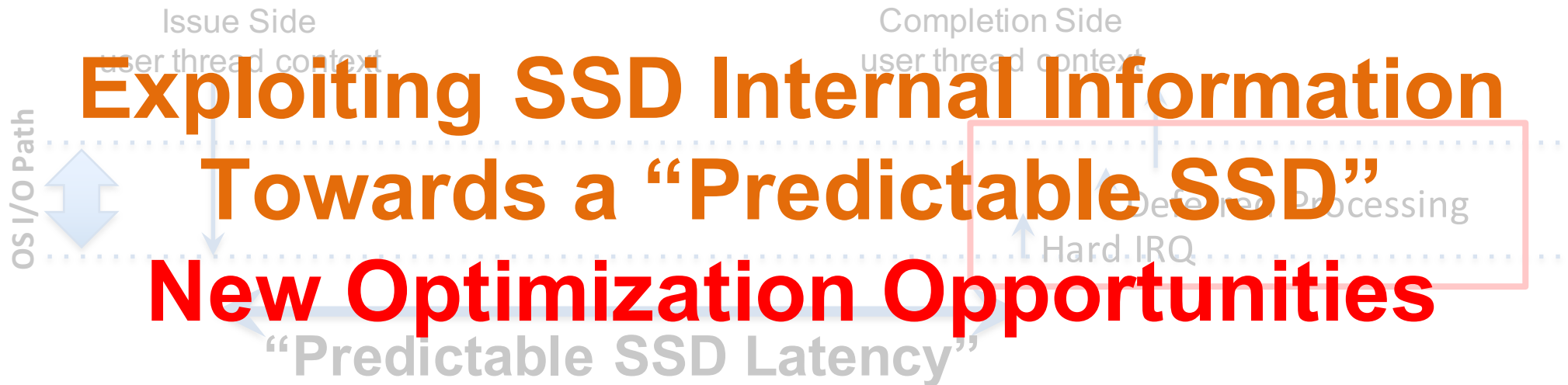
This Work: Unblinding the OS



This Work: Unblinding the OS



This Work: Unblinding the OS



Exploiting SSD Internal Information



OS



SSD

**The Unpredictable
SSD**

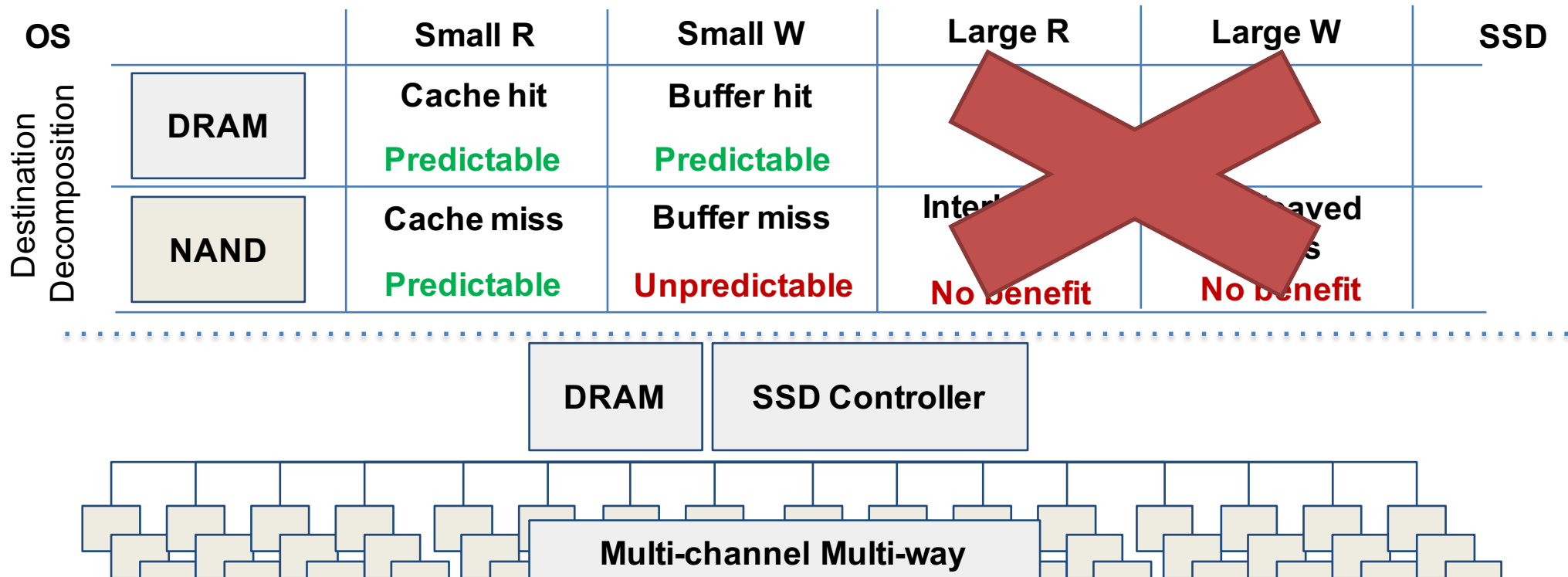
Multi-channel Multi-way

Exploiting SSD Internal Information

“Decomposition & Classification”



I/O request Classification

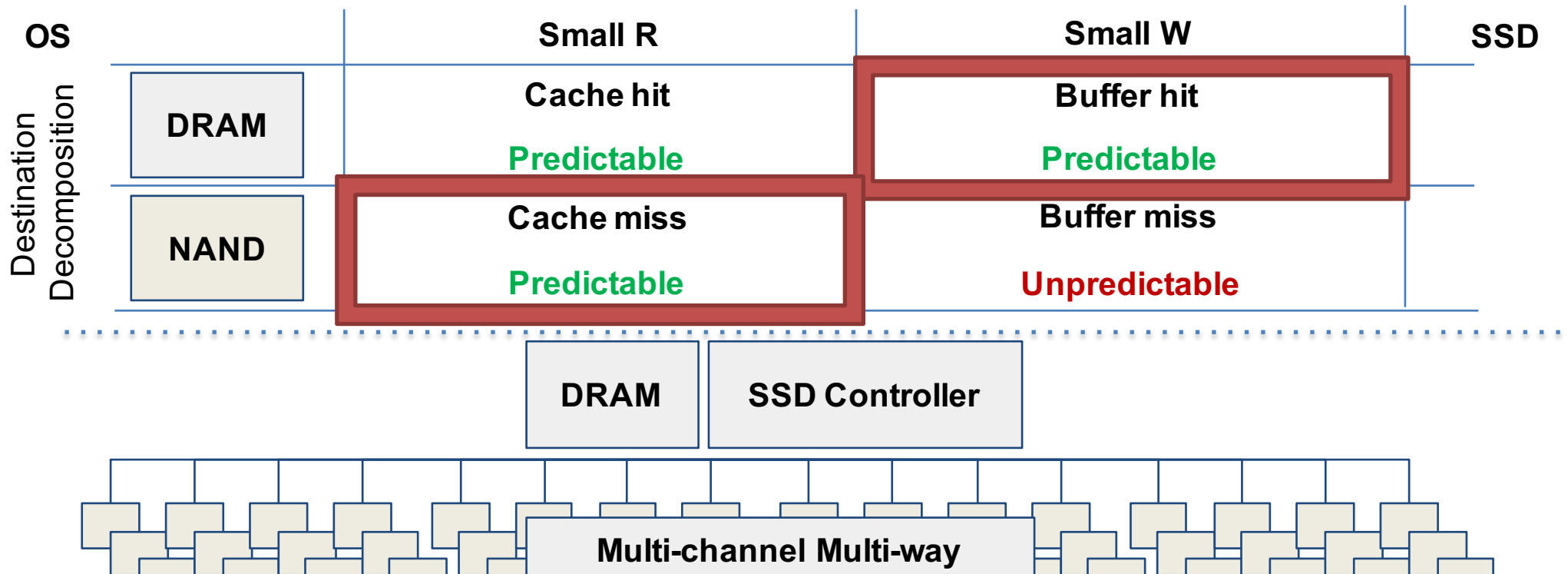


Exploiting SSD Internal Information

“Decomposition & Classification”



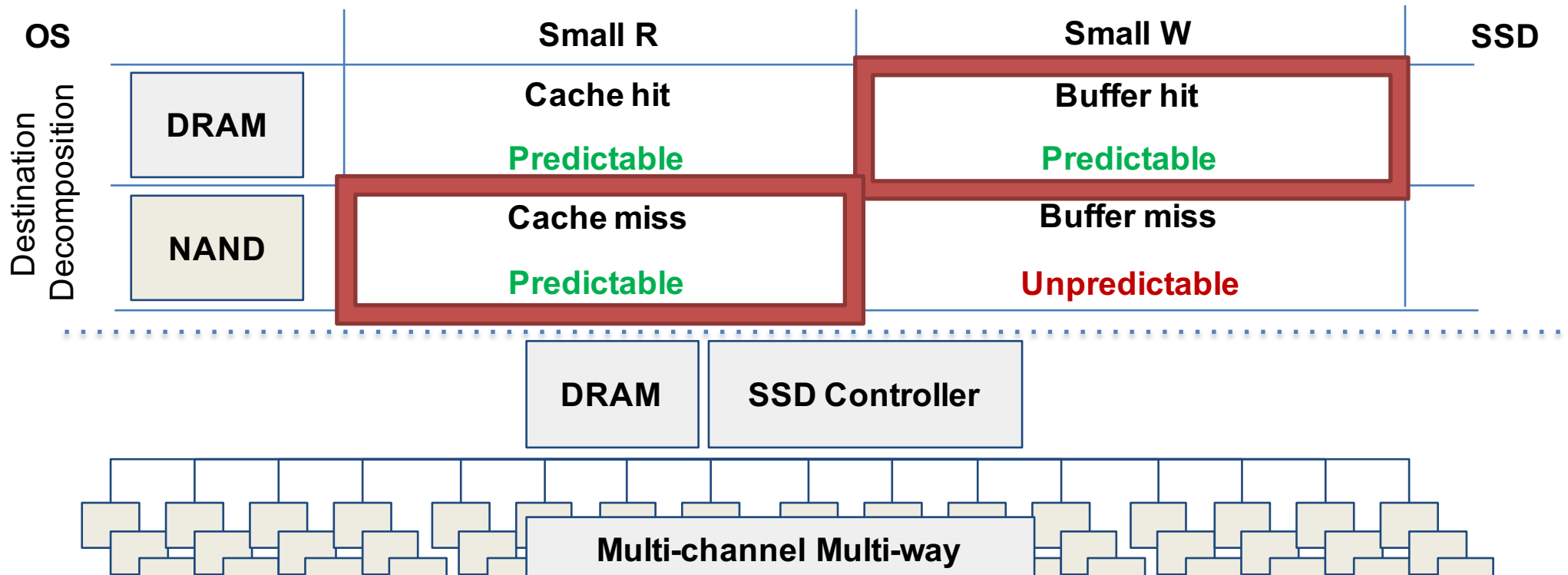
I/O request Classification



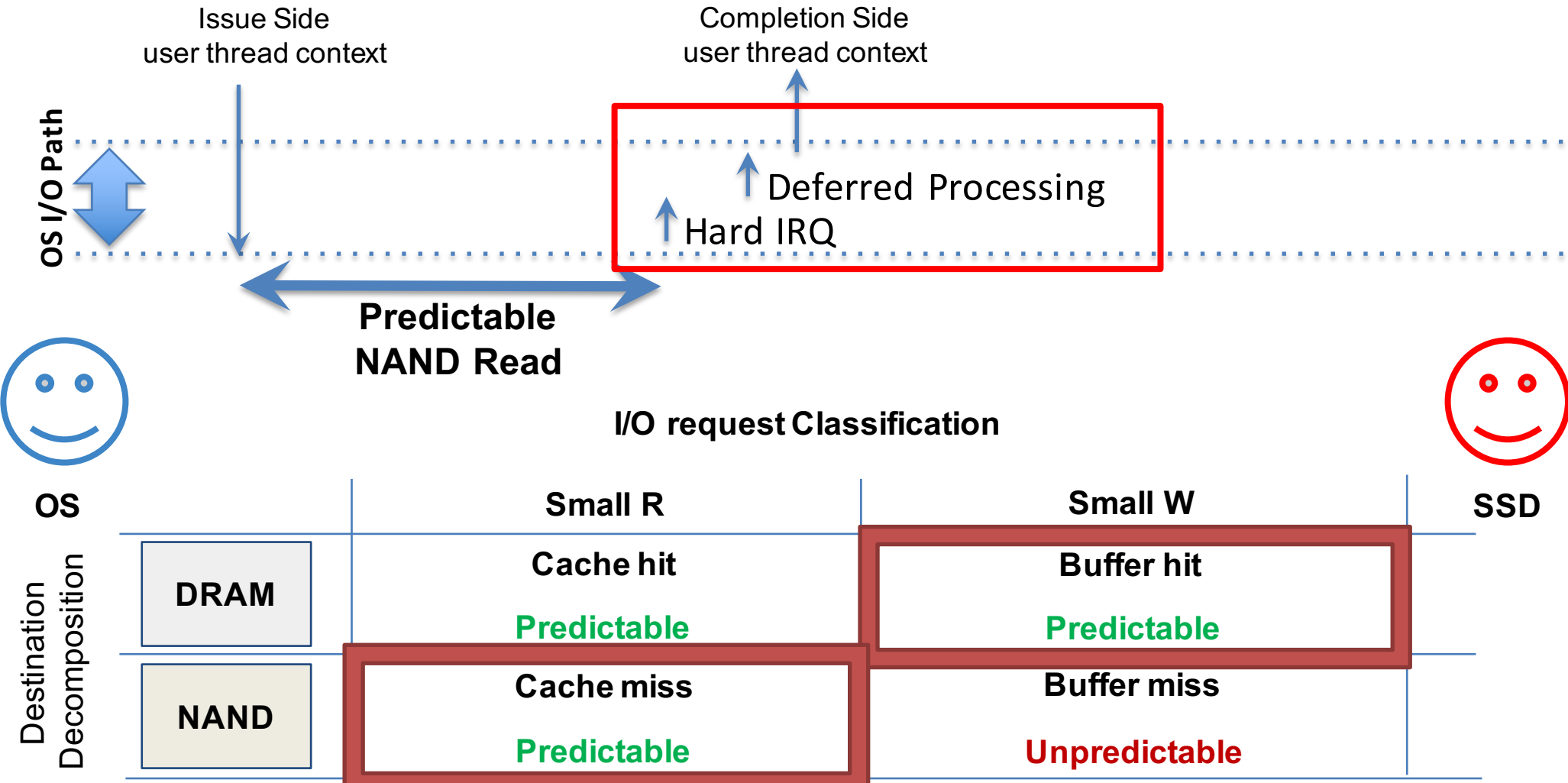
Exploiting SSD Internal Information “Decomposition & Classification”



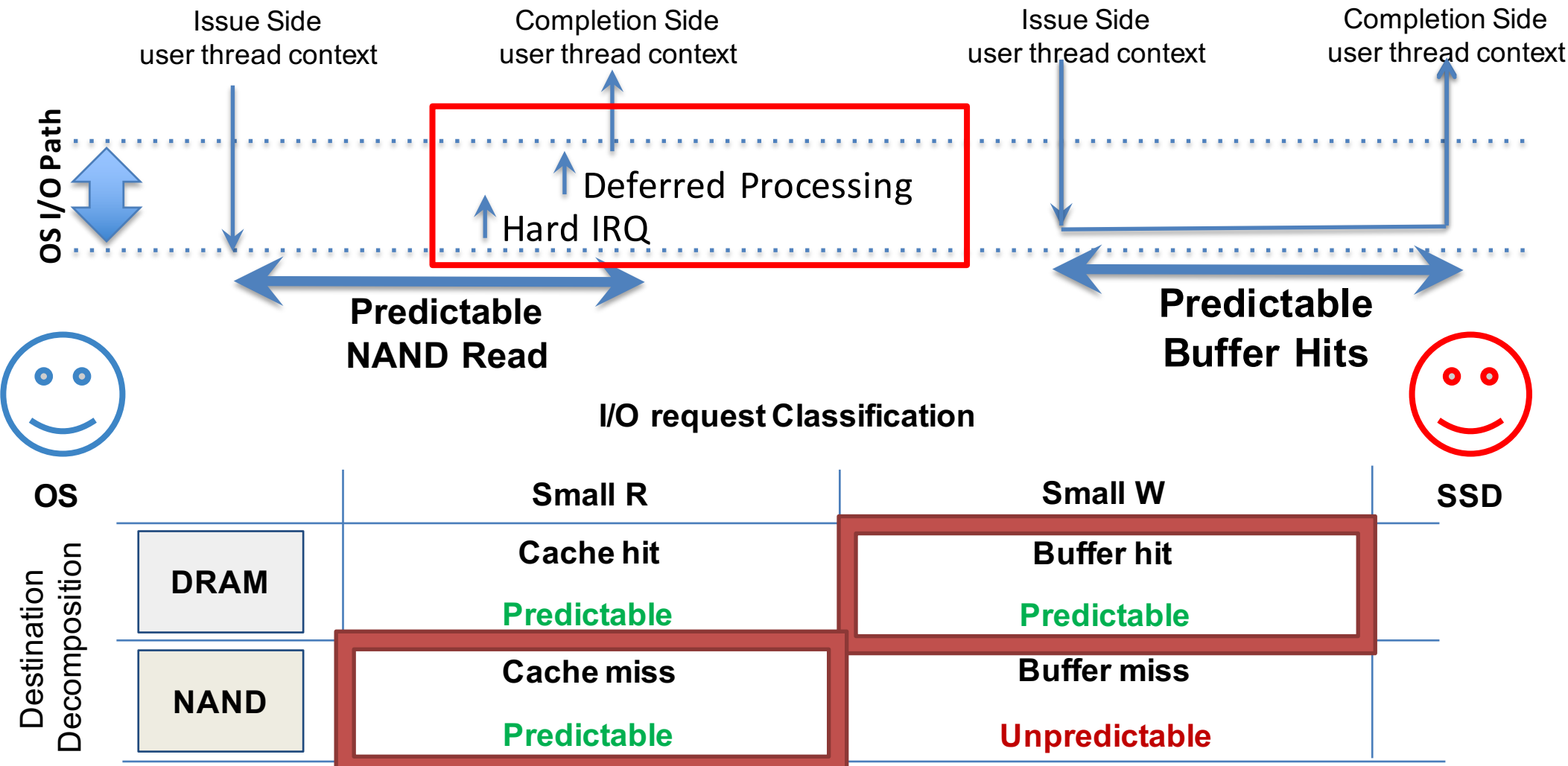
I/O request Classification



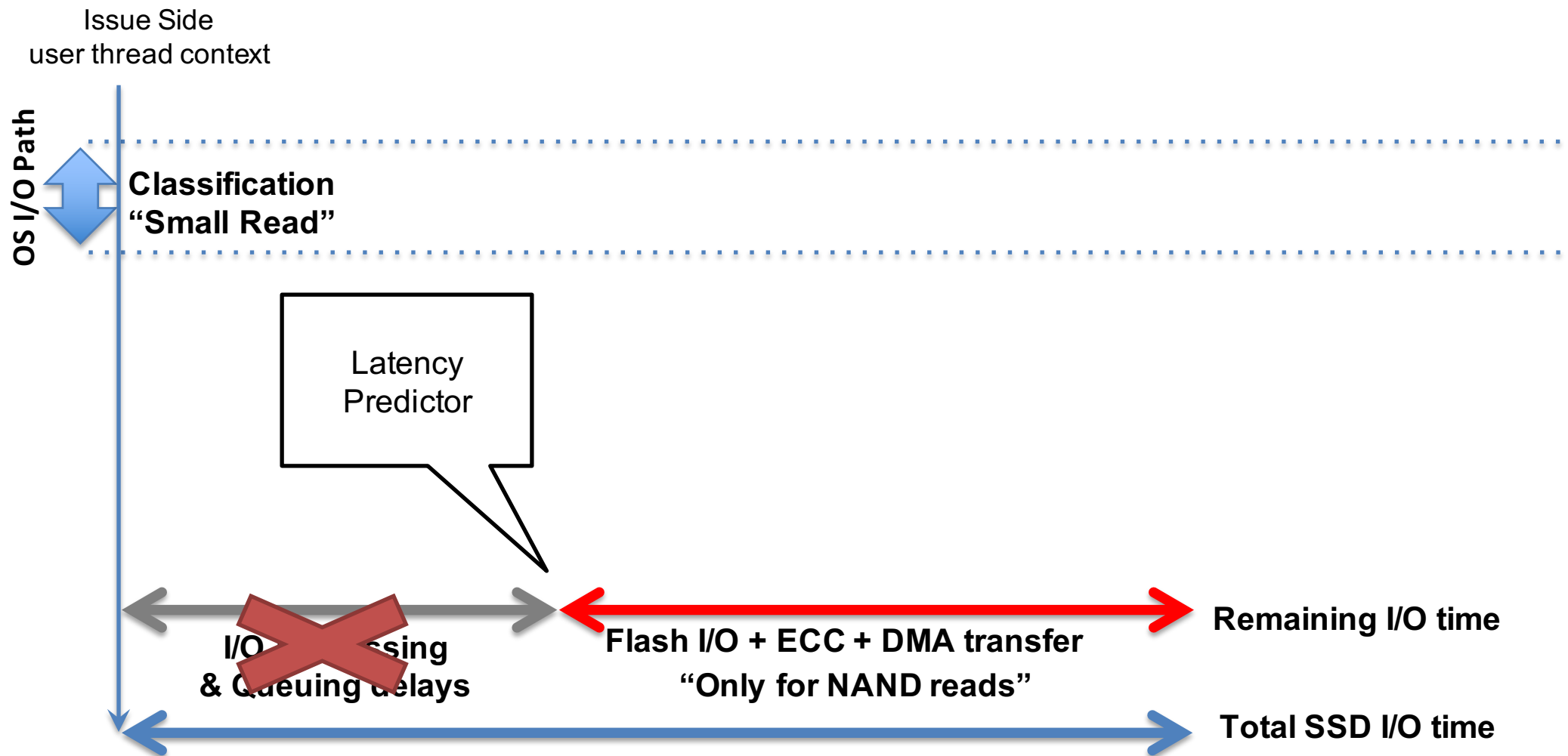
Mitigating the Impact of Scheduling Delays



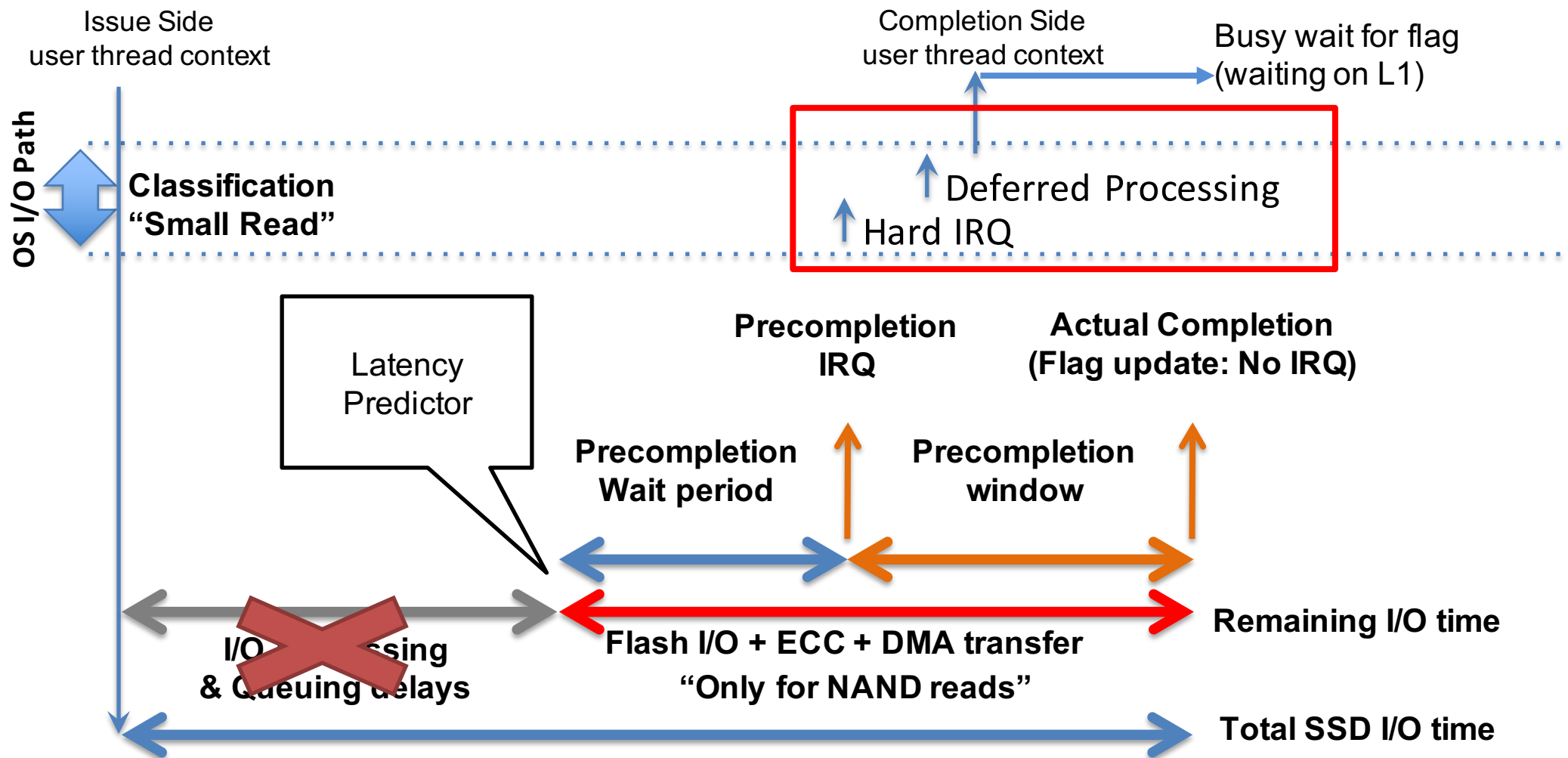
Mitigating the Impact of Scheduling Delays



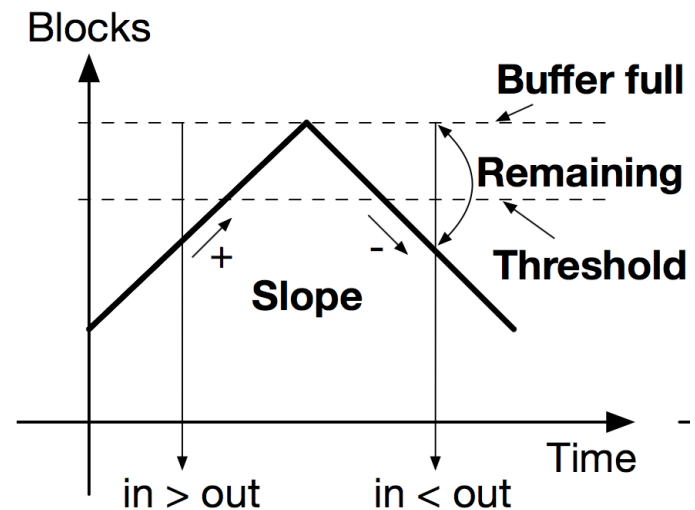
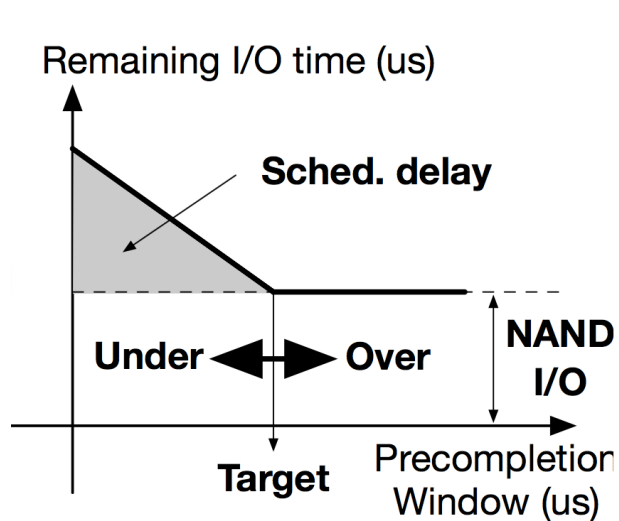
Accurate Latency Prediction: Remaining I/O Time



Precompletions: Overlapping I/O & Scheduling Delay



OS & SSD Interaction: Simple Behavioral Models

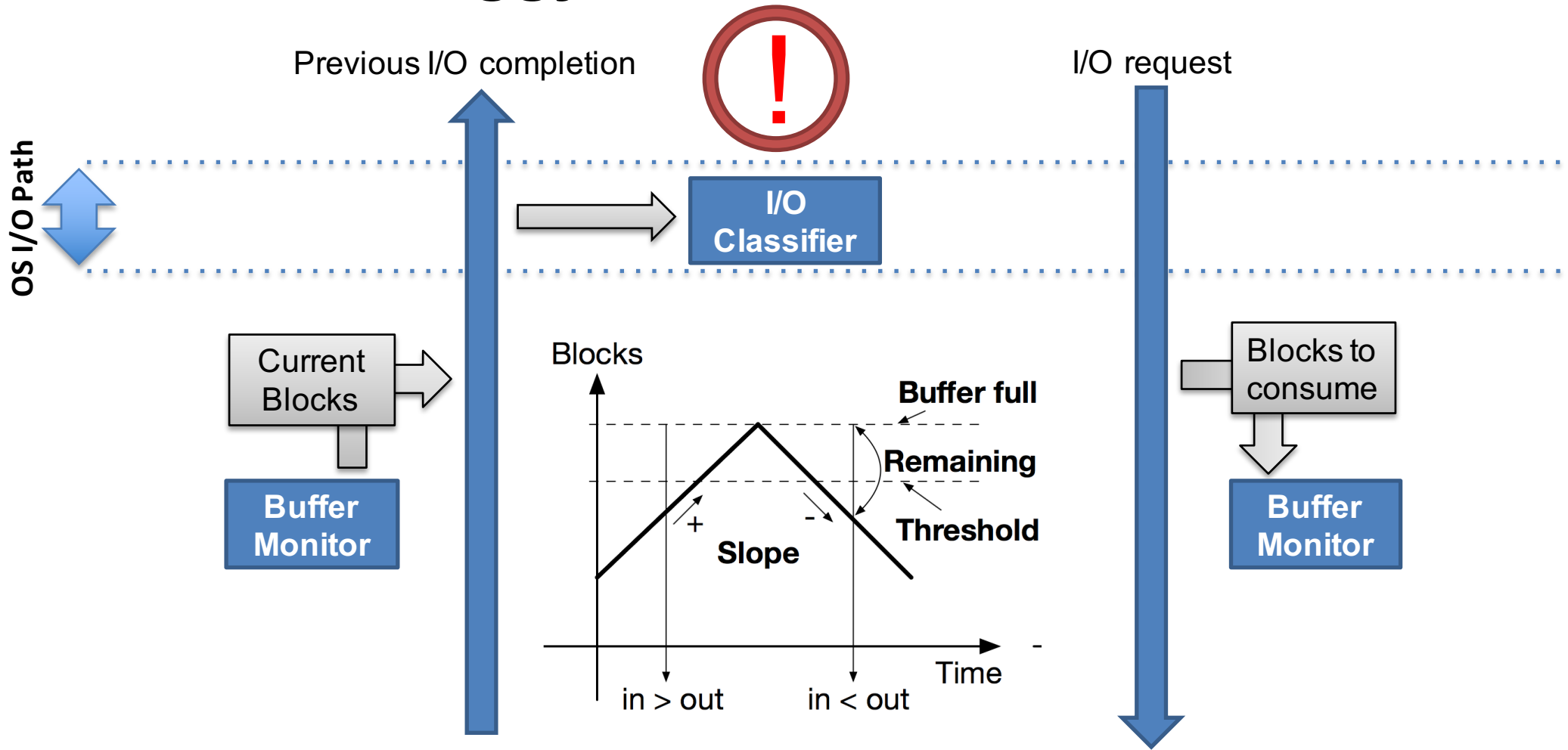


I/O request Classification

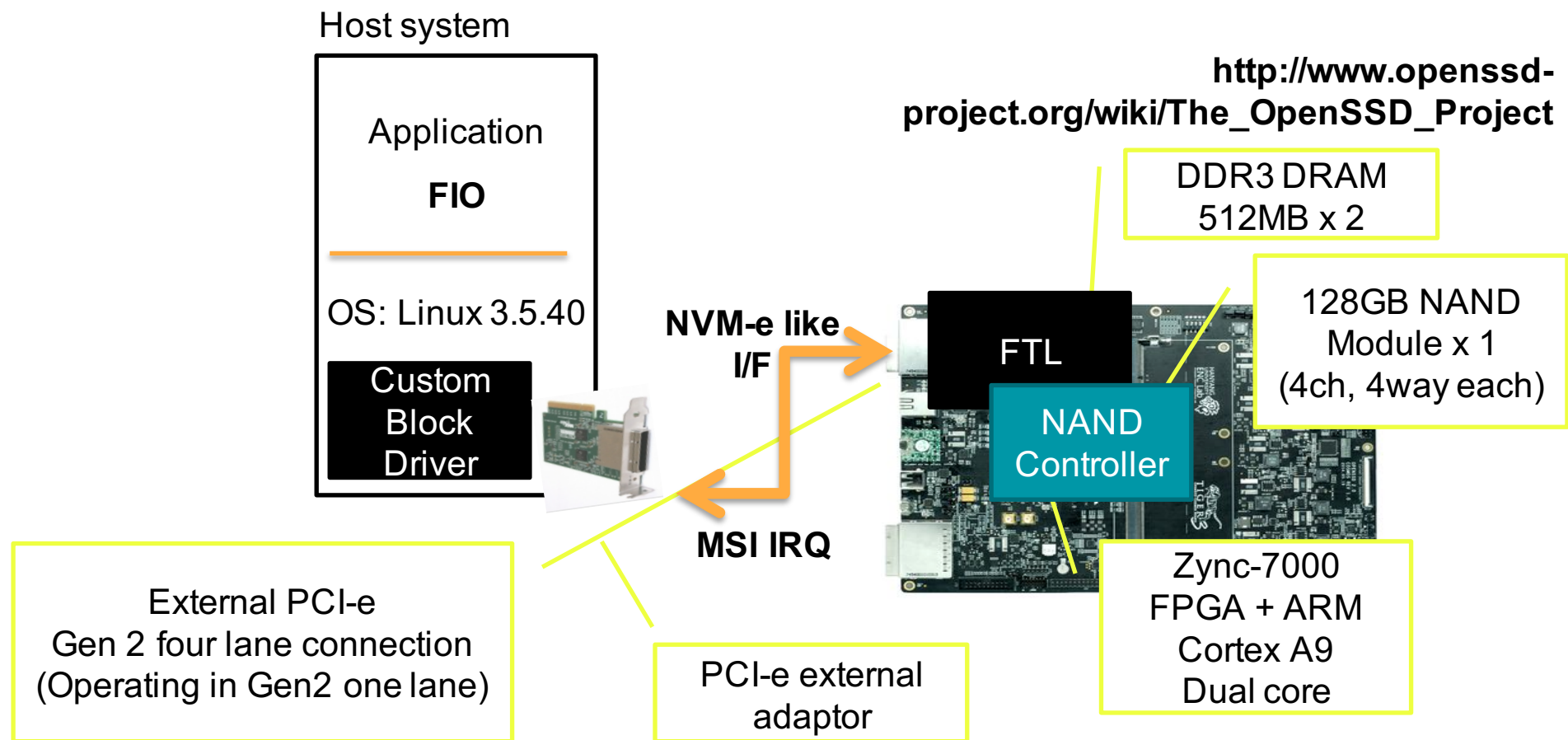
OS		Small R	Small W	SSD
Destination Decomposition	DRAM	Cache hit Predictable	Buffer hit Predictable	
	NAND	Cache miss Predictable	Buffer miss Unpredictable	

In-band Communication Channel

“Piggybacked Information”



Implementation & Evaluation



Implementation & Evaluation

Limitation

- Single I/O Depth
- Unoptimized FPGA NAND Controller (Higher Latencies)
- Fixed latency
- Slow DMA transfers (low freq. bus)
- PCI-e Gen2 one lane

http://www.openssd-project.org/wiki/The_OpenSSD_Project

DDR3 DRAM
512MB x 2

Latency Prediction

**Impact of
precompletion**

External PCI-e
Gen 2 four lane connection
(Operating in Gen2 one lane)

PCI-e
adaptor

Dual core

Predicting SSD Latency (Small NAND Read)

Flash: NAND I/O + ECC

Prediction: three value moving average

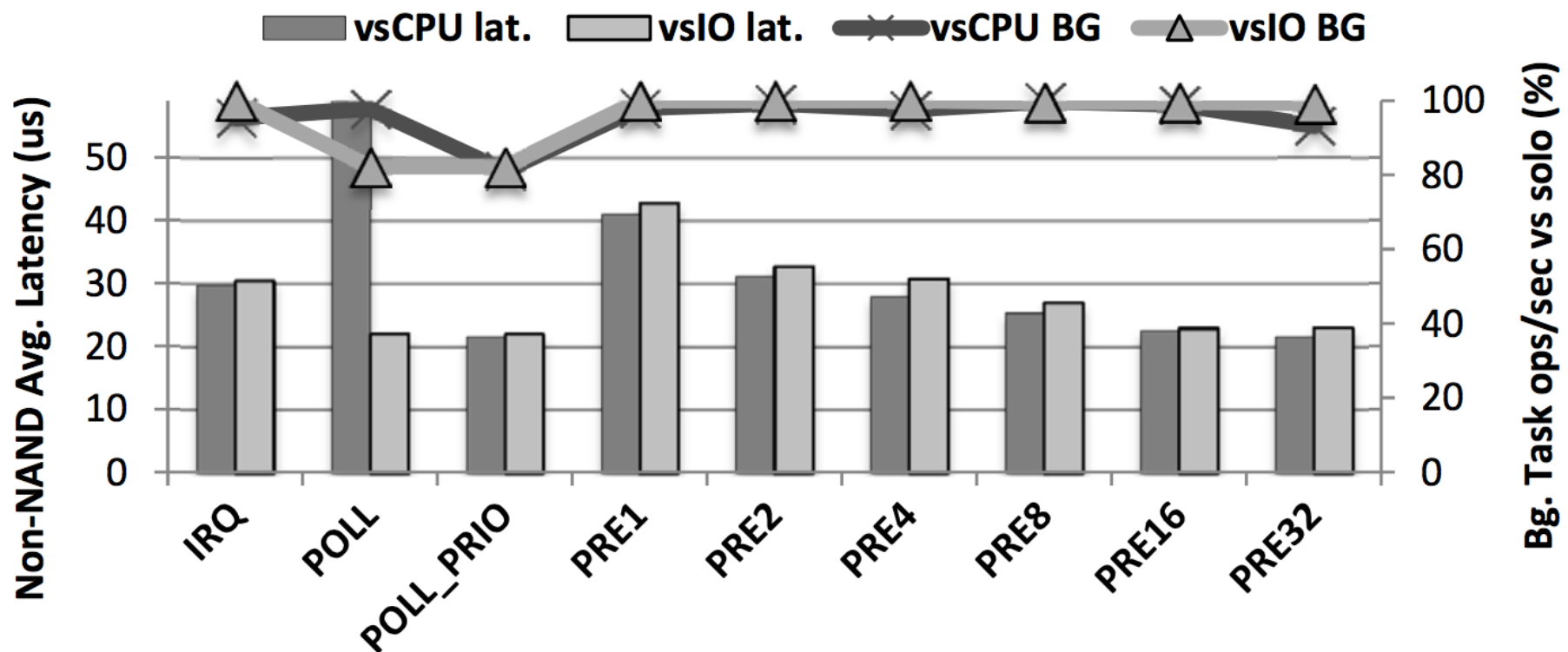
H/W	Measured	Std. dev	Predicted	Error
Flash	352 μ s	0.66 μ s	352 μ s	0.94 μ s
DMA	9 μ s	0.26 μ s	9 μ s	0.56 μ s

DMA: device to host transfer (4kB)

Low variance
Low Error
“Predictable”

The Impact of Precompletion

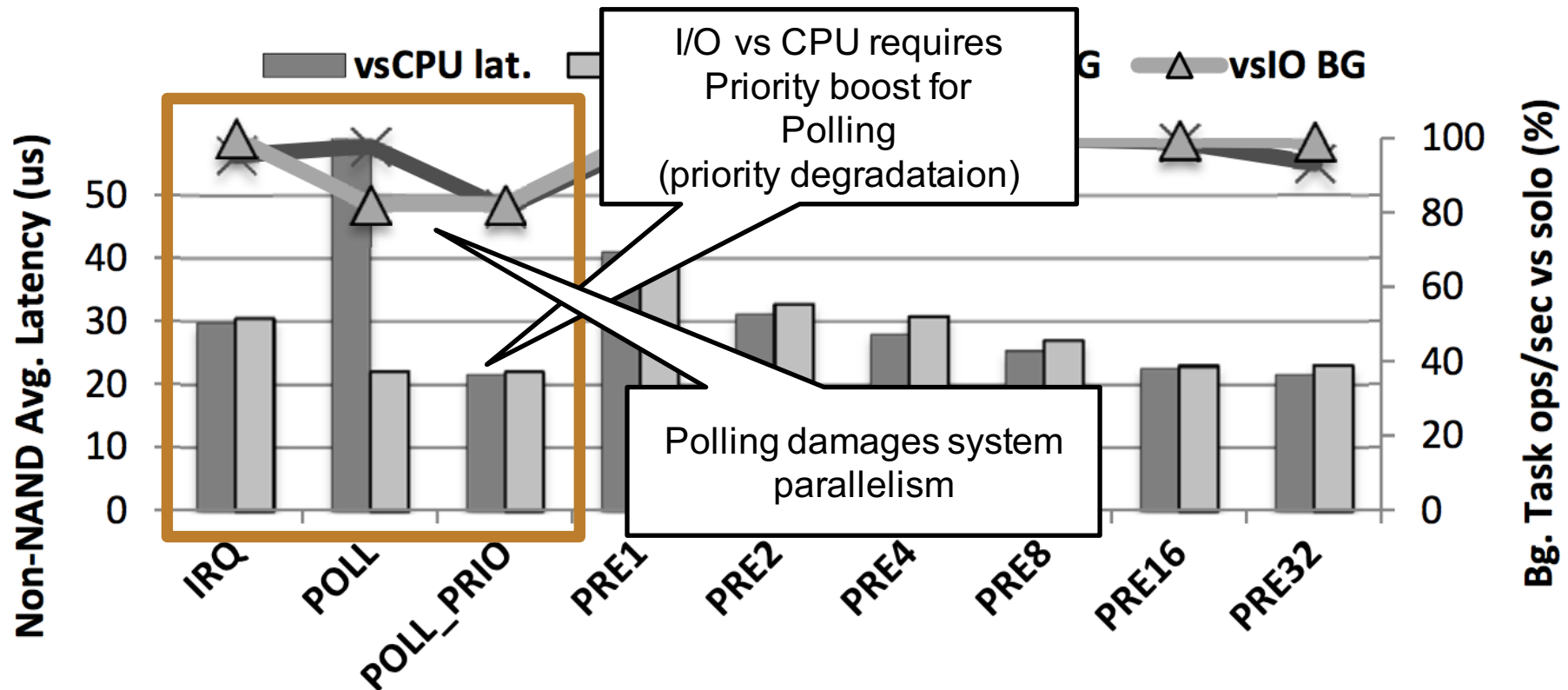
fio I/O thread vs background threads



Non-NAND Avg. latency: Measured AVG latency – 352us (NAND latency)

The Impact of Precompletion

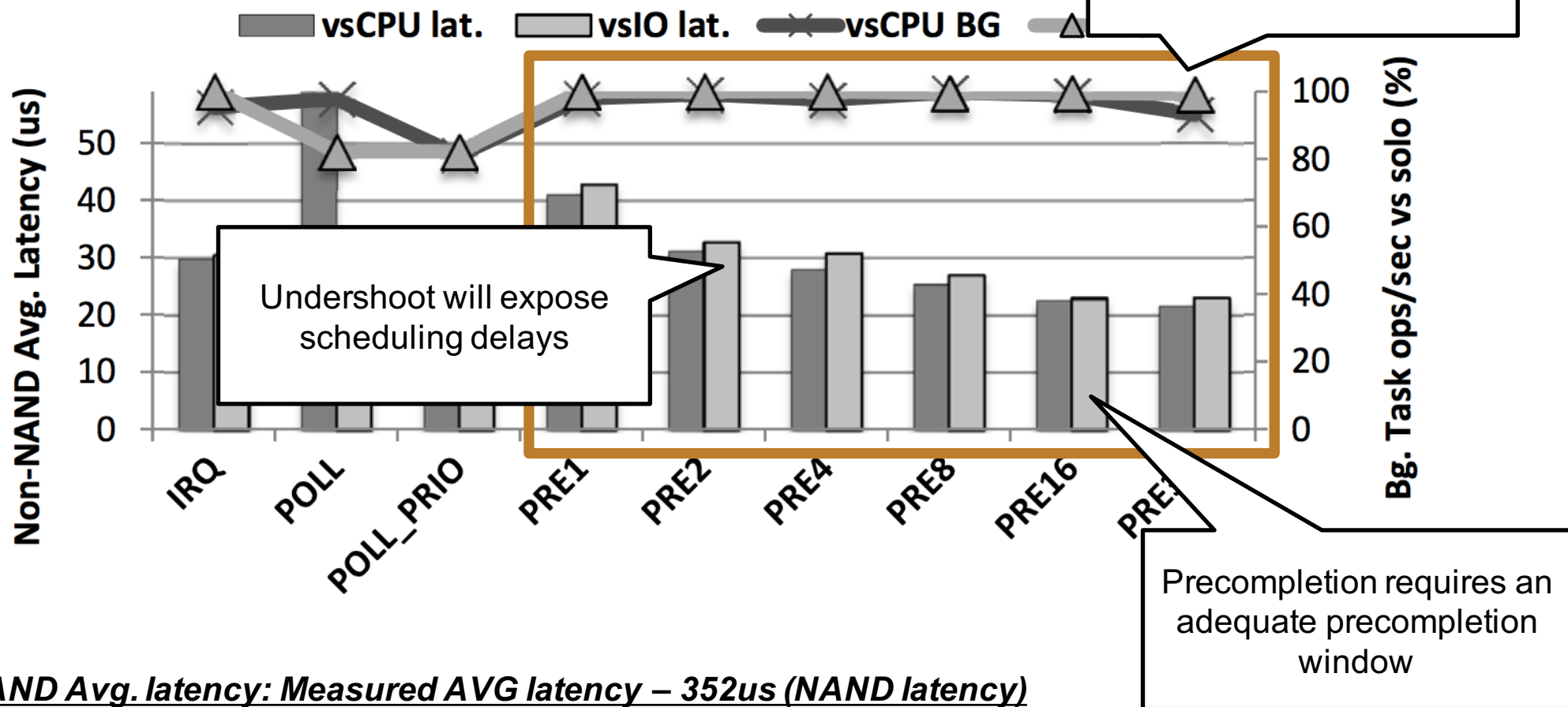
fio I/O thread vs background threads



Non-NAND Avg. latency: Measured AVG latency – 352us (NAND latency)

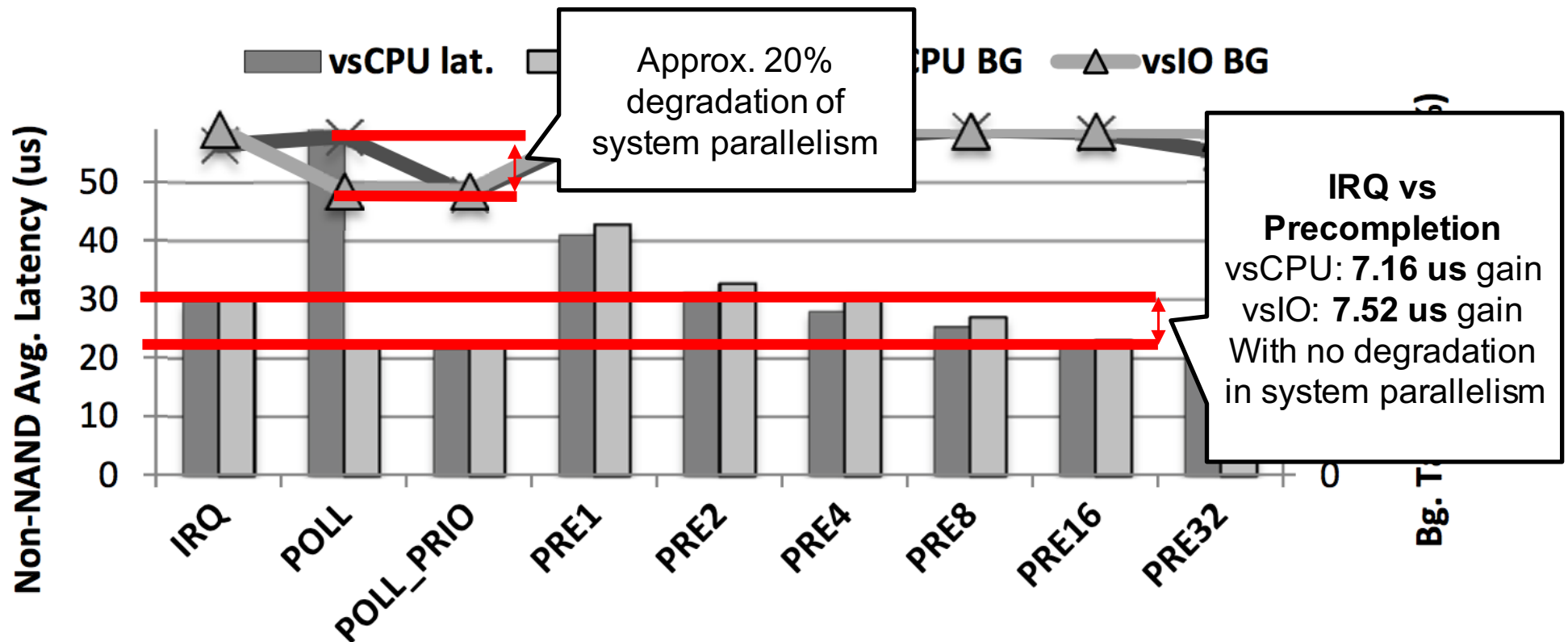
The Impact of Precompletion

fio I/O thread vs background threads



The Impact of Precompletion

fio I/O thread vs background threads



Non-NAND Avg. latency: Measured AVG latency – 352us (NAND latency)

Summary

- **Unblinding the OS - Cross layer optimization**
 - Achieved a partially predictable SSD “*decomposition / classification*”
 - Exploit SSD internal information - “*Remaining I/O time*”
 - Protecting SSD proprietary internals – “*Abstracted behavioral models*”
- **Mitigating scheduling delays**
 - Exploiting predictability of certain I/O requests
 - Pre-completion - Projection (1 I/O depth vs other threads)

	IRQ	Polling	This work
Latency	Bad	Good	Good
Parallelism	Good	Bad	Good

Future Work

- **Future Implementation & Evaluation**
 - Full blown SSD
 - **Projection (1 I/O depth – this work)**
 - Simulation (varying tech latency & etc)
 - Real implementation
- **Cross layer optimization**
 - More models
 - More use cases
 - More backend technologies rather than flash

