SPDK+: Low Latency or High Power Efficiency? We Take Both

Endian Li, Shushu Yi, Li Peng, Qiao Li, Diyu Zhou, Zhenlin Wang, Xiaolin Wang,Bo Mao, Yingwei Luo, Ke Zhou, Jie Zhang

CHASELab Computer Hardware And System Evolution Laboratory

PEKING UNIVERSITY

Michigan Tech

Background: SSDs Become Dominant Storage Media

SSDs → Superb **performance** & **reliability**!

DELL EMC VMAX

PureStorage FlashArray

Datacenters

Supercomputers

FUJITSU ETERNUS

NetApp AFF

SSDs are widely adopted in diverse domains.

[1] Lin et al. " Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management. " 2024

Background: Evolvement of SSD

Continual advancement in SSD performance

16x

SATA SSD: 100 KIOPS

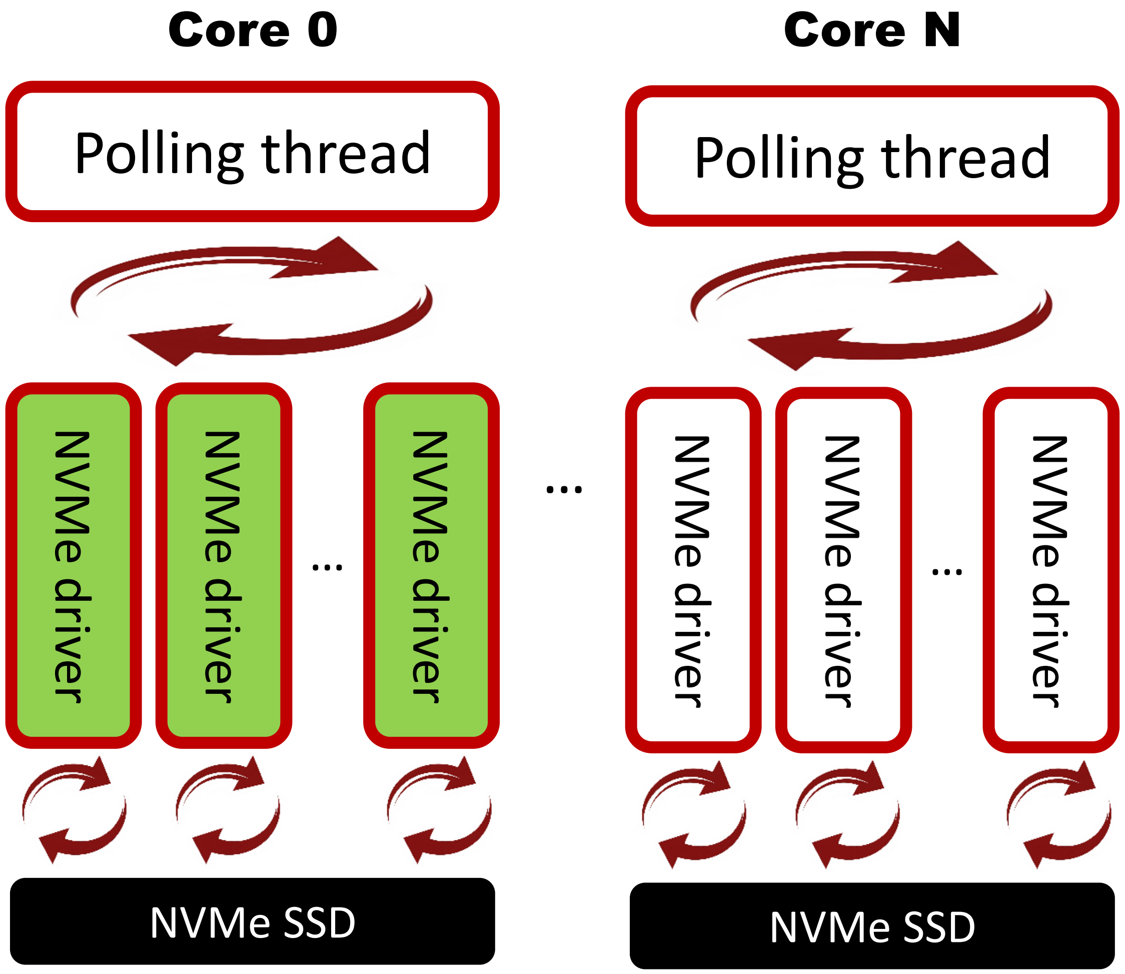
PCIe5.0 SSD: 1600 KIOPS

How can I/O storage stack fully exploit high-performance SSDs?

[1] Lin et al. " Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management. " 2024

Background: SPDK

- The Storage Performance Development Kit (SPDK)
  - User space I/O engine
  - concurrent multi-thread accesses based on a lock-free principle
  - Run-to-complete thread
  - Polling method
  - Advantage: Low latency



Internal details of SPDK

[1] Shehabi, et al. "United states data center energy usage report." 2016.
[2] IEA. "Global data centre energy demand by end use." 2019.
[3] Ganesh, E. N. "Analysis of Low Power Data Server in Distributed Environments." 2022.

Background: Challenge in Polling methods

Polling vs sync I/O: 23.1% latency ↓ & 10x CPU usage ↑

CPU usage ↑ CPU power consumption ↑

CPU Usage

$$\boxed{P} = c + k \cdot f^3 \cdot \boxed{U_{cpu}}$$ [1]

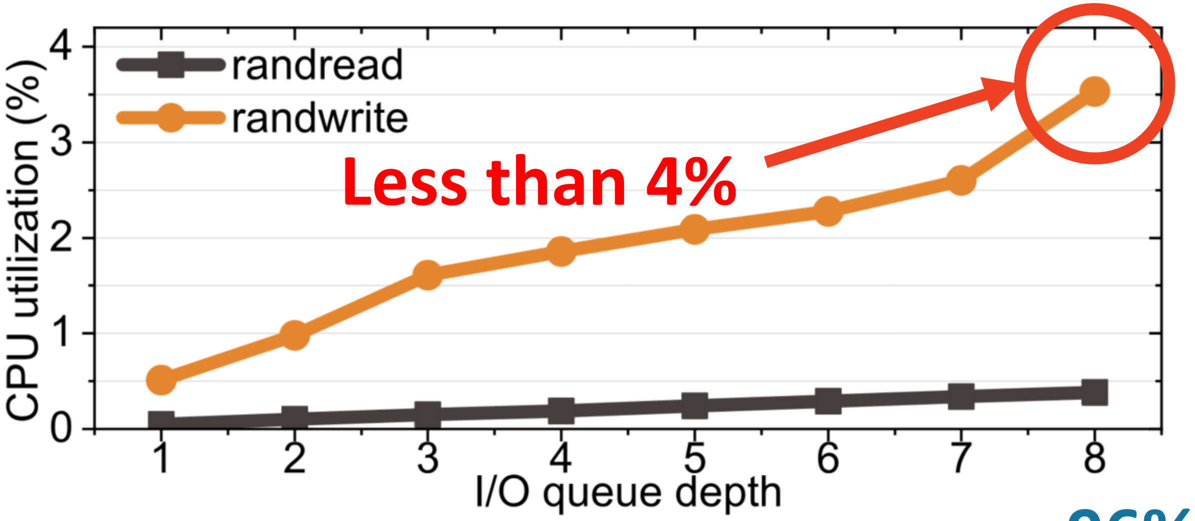
CPU Power

The CPU power consumption in polling methods is high.

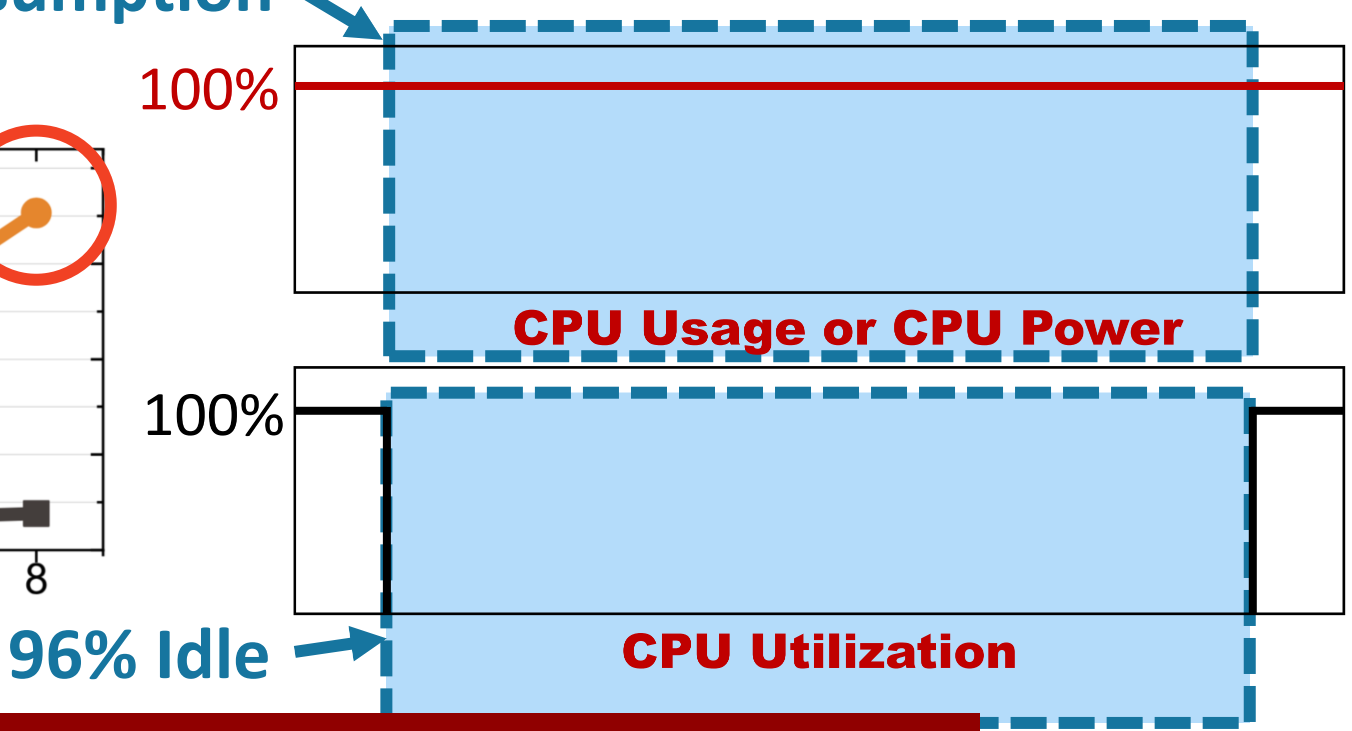
[1] Liu, Zhen, Xiang, Yongchao, Qu, Xiaoya, Workload-Aware and CPU Frequency Scaling for Optimal Energy Consumption in VM Allocation, Mathematical Problems in Engineering, 2014, 906098, 12 pages, 2014. https://doi.org/10.1155/2014/906098

Background: Challenge in Polling Methods

Is the power consumption of these CPUs worth it?



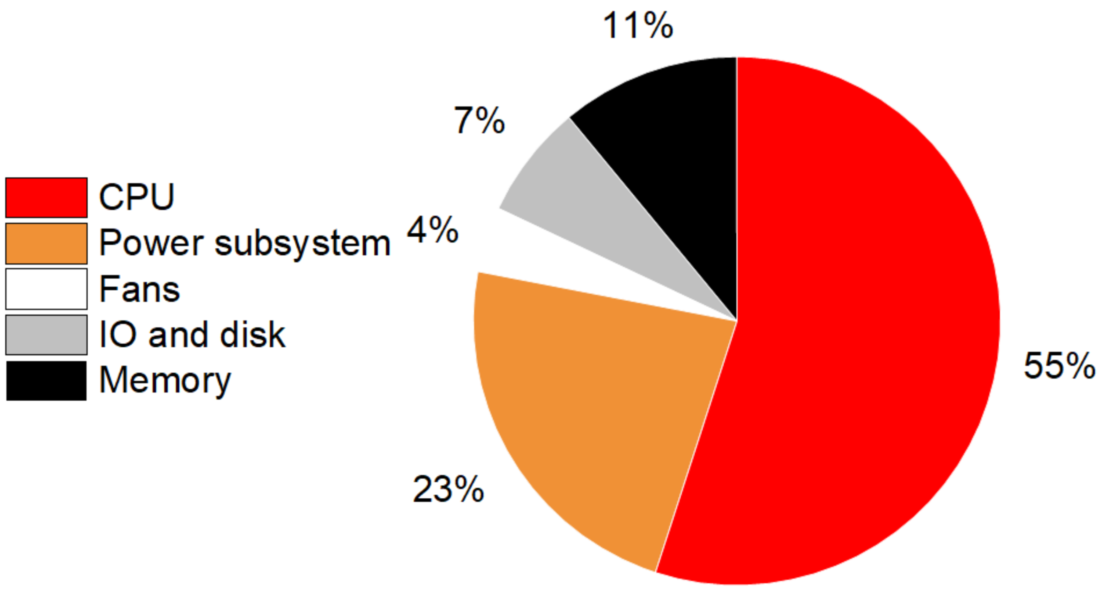
Less than 4%

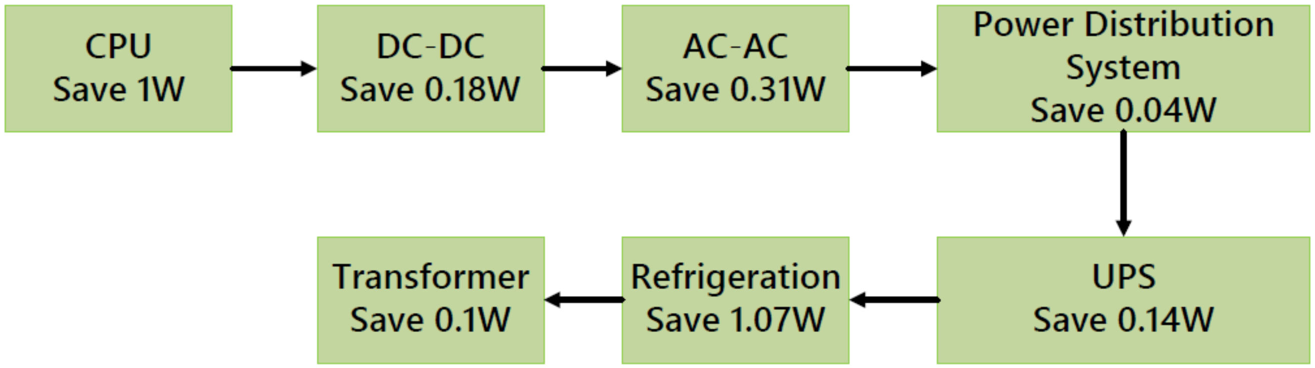


A large amount of CPU power consumption is wasted

[1]童琳."服务器各部件节能技术分析[J].通信与信息技术."2015.

The Importance of CPU Power Conservation



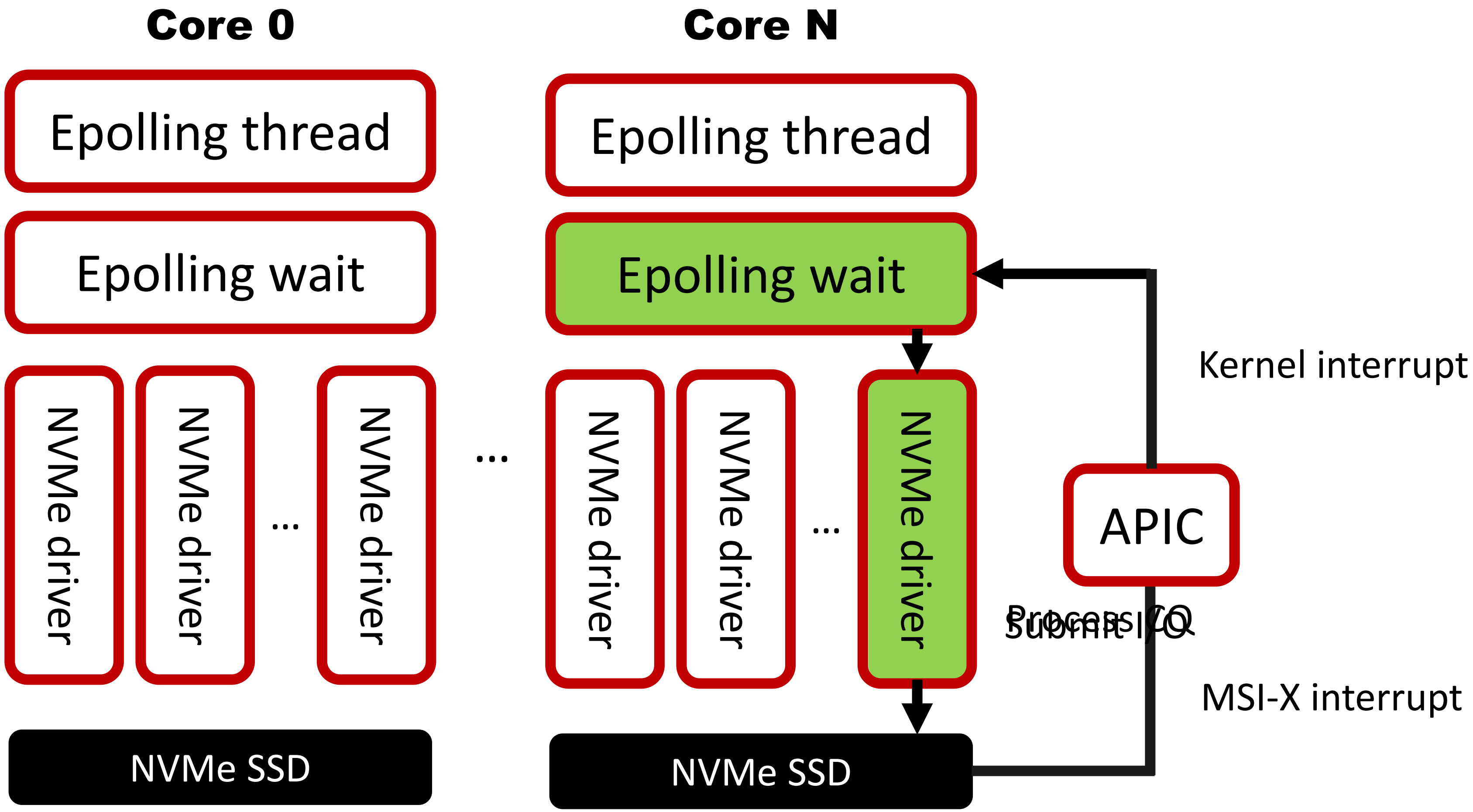


It is essential to improve the power efficiency of the CPU

[1]童琳.”服务器各部件节能技术分析[J].通信与信息技术.”2015.
[2] Ganesh, E. N. "Analysis of Low Power Data Server in Distributed Environments.“ 2022.

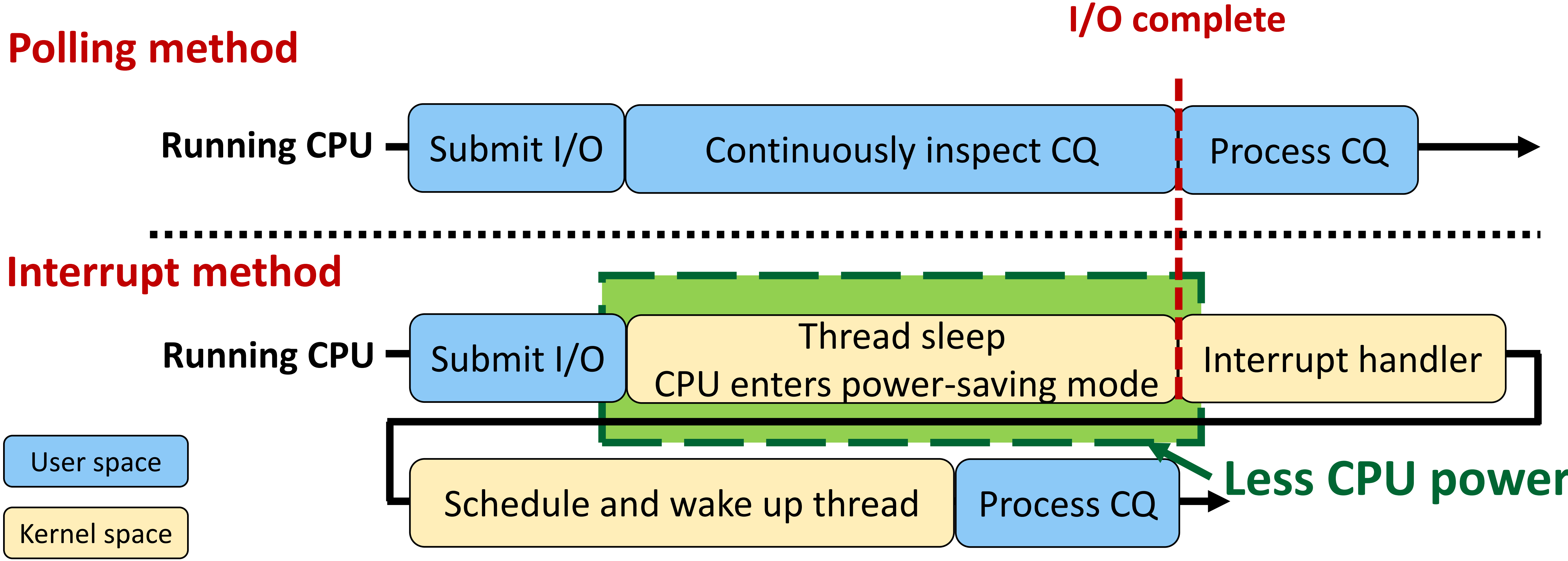
Background: Alternative Approach – Interrupt Mode

- SPDK-IM: SPDK added support for interrupt method.



Internal details of SPDK interrupt methods

Background: SPDK-IM

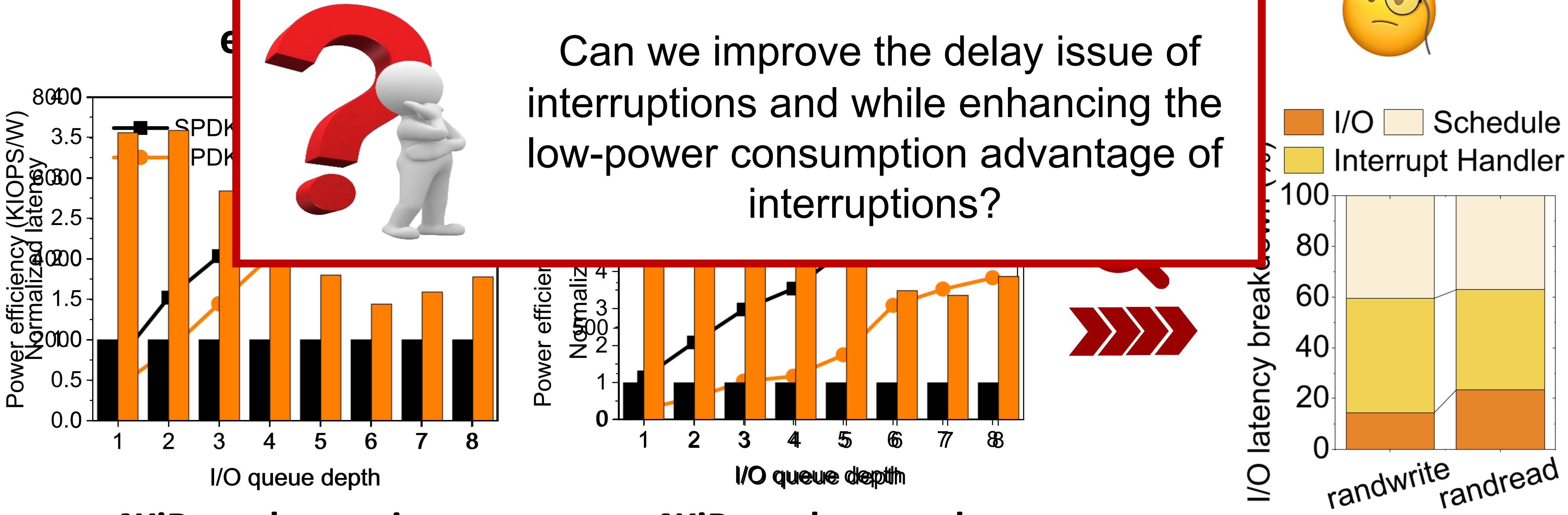


Interrupt method consumes less CPU power

Background: Challenge in Interrupt Methods

In small I/O shallow queue: SPDK-IM < SPDK (Power efficiency)

Because high software latency in SPDK-IM



Can we improve the delay issue of interruptions and while enhancing the low-power consumption advantage of interruptions?

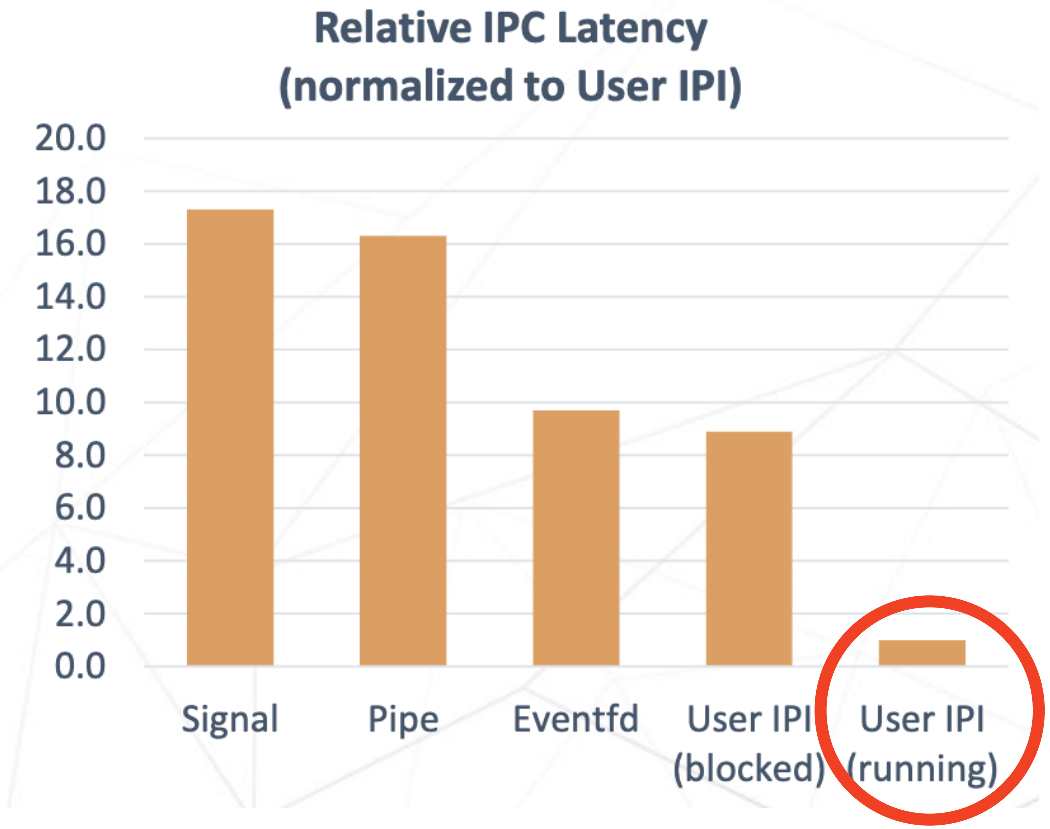
4KiB random write

4KiB random read

Key Insight: The User Interrupt Feature Provides Low Latency

User interrupt: a new feature in Intel CPU, aim to reduce inter-process communication latency

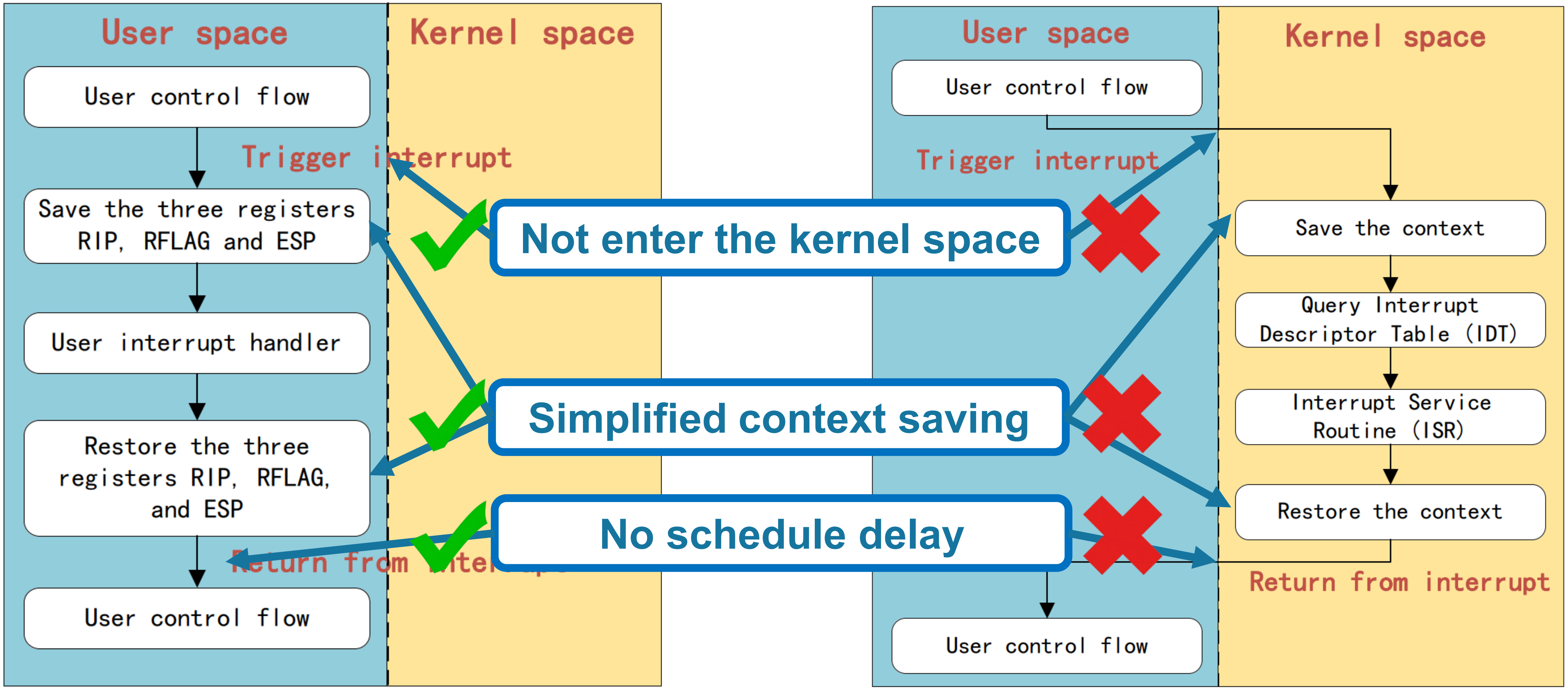
User IPI: Inter-Processor Interrupt supported by user interrupt



The user interrupt delay can be reduced to 1 μs

[1] Mehta, Sohil. "User Interrupts – A faster way to signal." Linux Plumbers Conference 2021, Contribution 985, Attachment 756. 2021. PDF file.

Key Insight: The User Interrupt Feature Provides Low Latency



User space

Kernel space

User control flow

Trigger interrupt

Save the three registers RIP, RFLAG and ESP

User interrupt handler

Restore the three registers RIP, RFLAG, and ESP

Return from interrupt

User control flow

User space

Kernel space

User control flow

Trigger interrupt

Save the context

Query Interrupt Descriptor Table (IDT)

Interrupt Service Routine (ISR)

Restore the context

Return from interrupt

User control flow

Not enter the kernel space

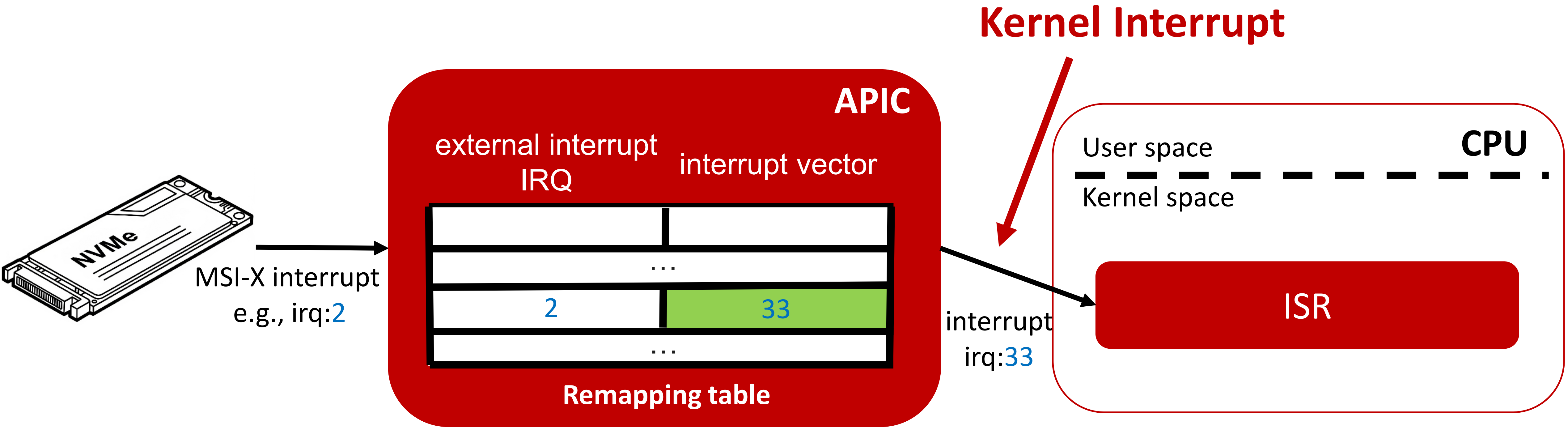
Simplified context saving

No schedule delay

User interrupt process | **Kernel interrupt process**

Design: Use User Interrupt to process MSI-X Interrupt

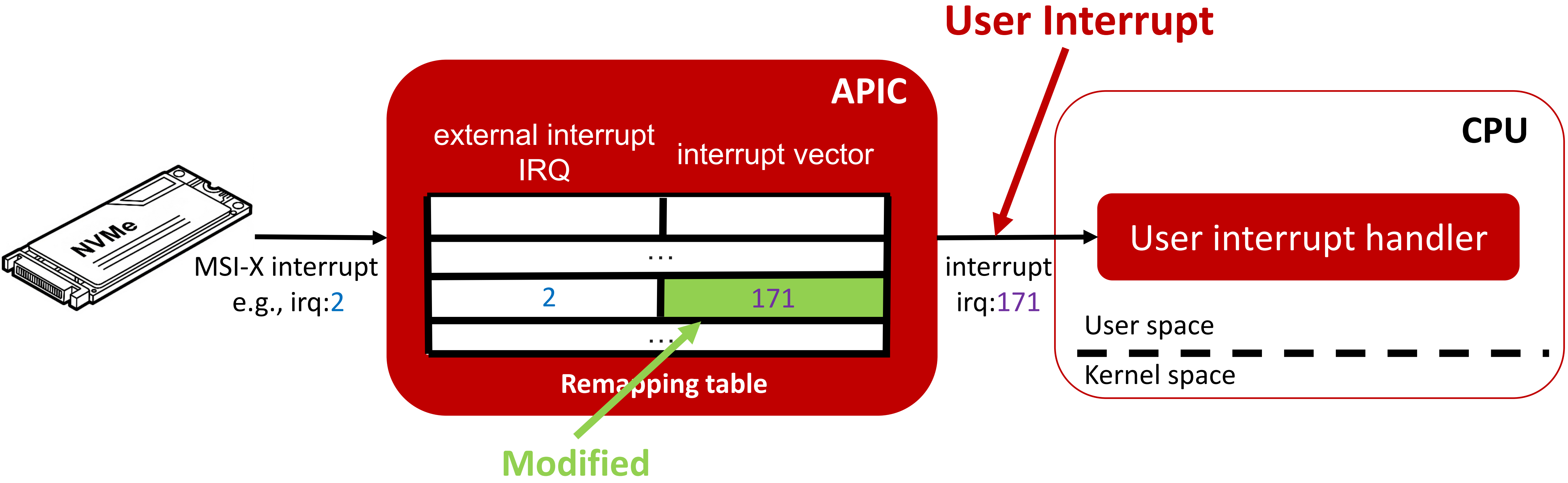
Assumption: user interrupt vector: 171



external interrupt IRQ	interrupt vector
…	
2	33
…	

Design: Use User Interrupt to process MSI-X Interrupt

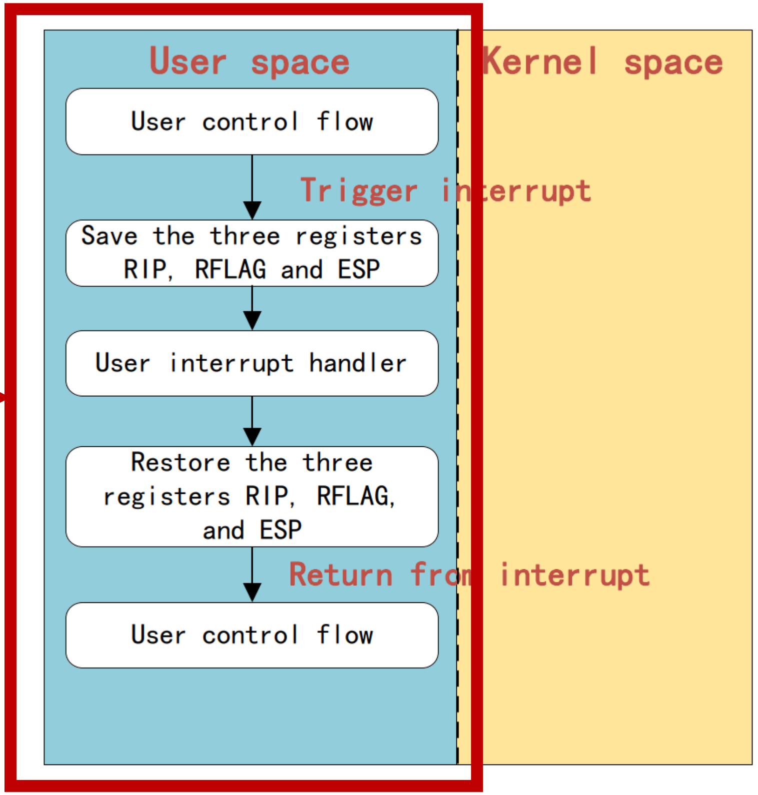
Assumption: user interrupt vector: 171



Key Insight: User wait instructions reduce power consumption

What is the cost of low latency?

☹️ The user program is always staying in the foreground



User space

Kernel space

User control flow

Trigger interrupt

Save the three registers RIP, RFLAG and ESP

User interrupt handler

Restore the three registers RIP, RFLAG, and ESP

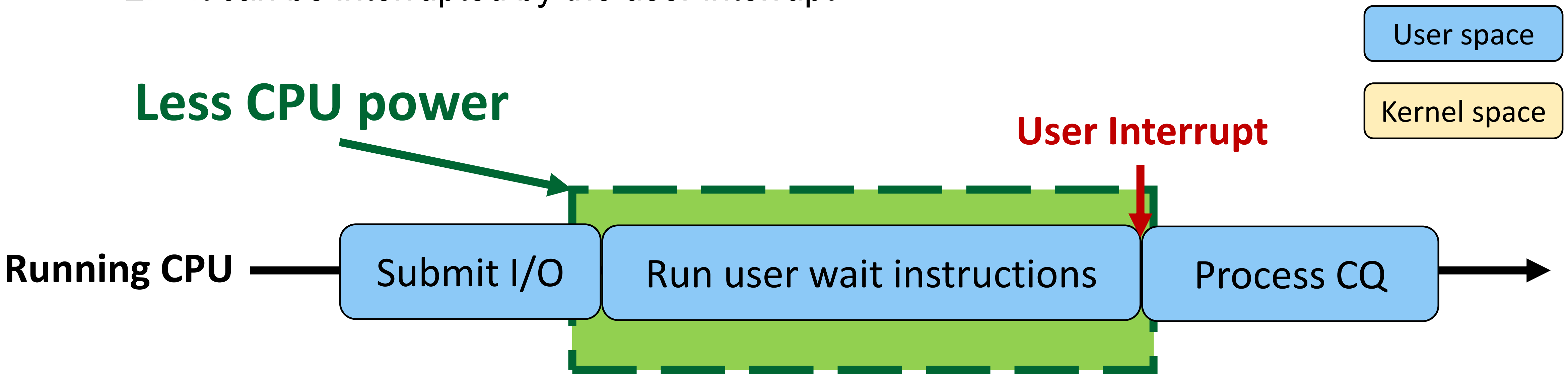
Return from interrupt

User control flow

Key Insight: User wait instructions reduce power consumption

What should the CPU do while waiting for I/O to complete? 🤔

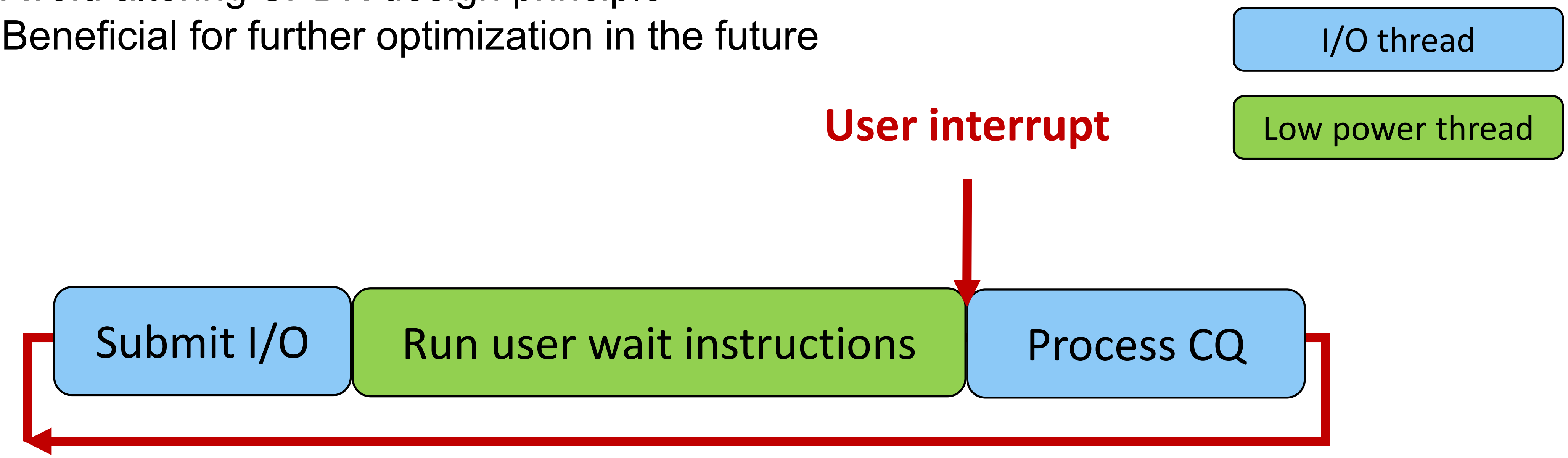
- **key insight:** user wait instructions
  1. Allow the CPU to directly enter a low-power state in user mode
  2. It can be interrupted by the user interrupt



Design: User-mode Scheduling Framework

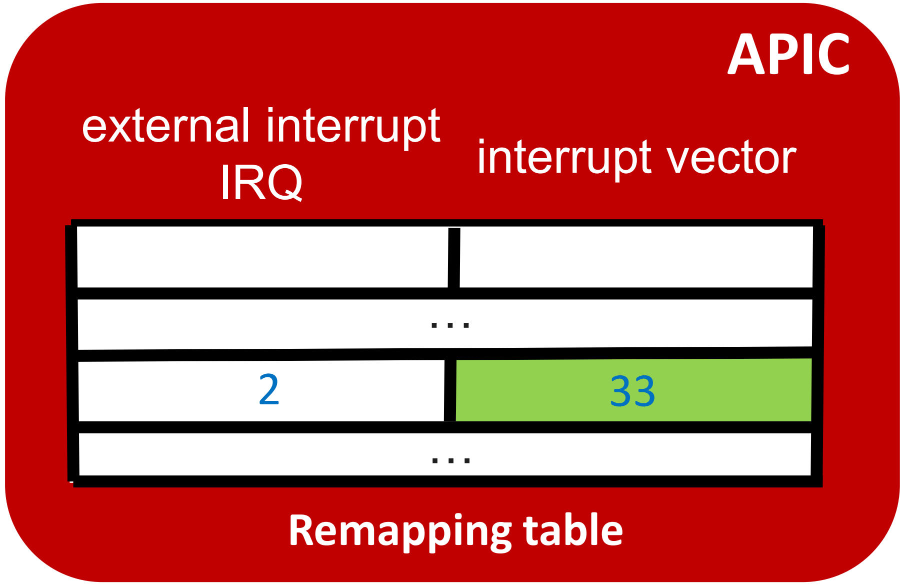
How to get user interrupt and user wait instruction together?

- **Design:** user-mode scheduling framework
  1. Low switch delay
  2. Avoid altering SPDK design principle
  3. Beneficial for further optimization in the future



SPDK+： the whole process

Assumption: user interrupt vector: 171

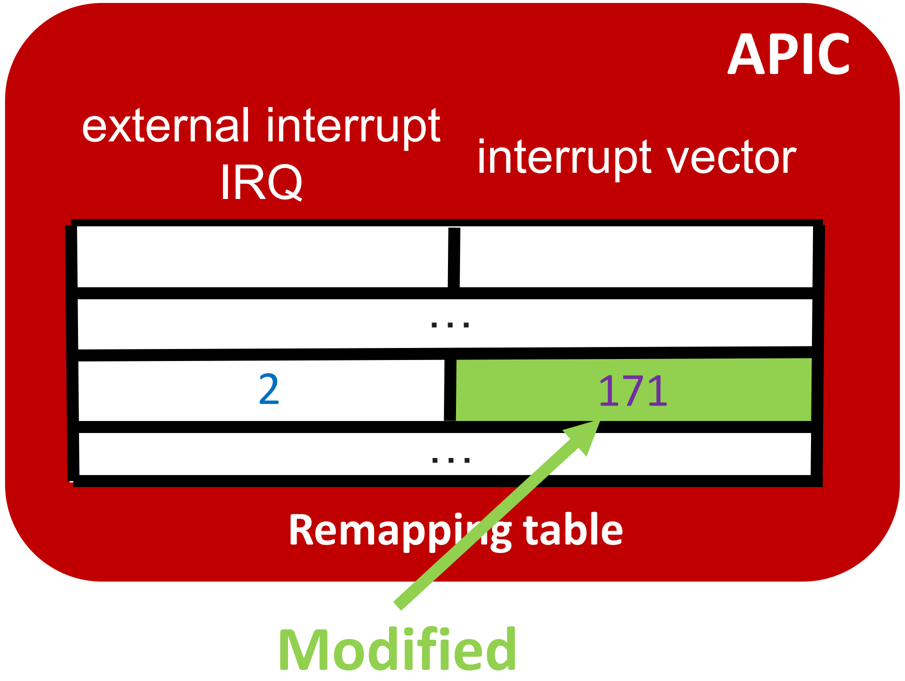


external interrupt IRQ	interrupt vector
…	
2	33
…	

Remapping table

SPDK+： the whole process

Assumption: user interrupt vector: 171



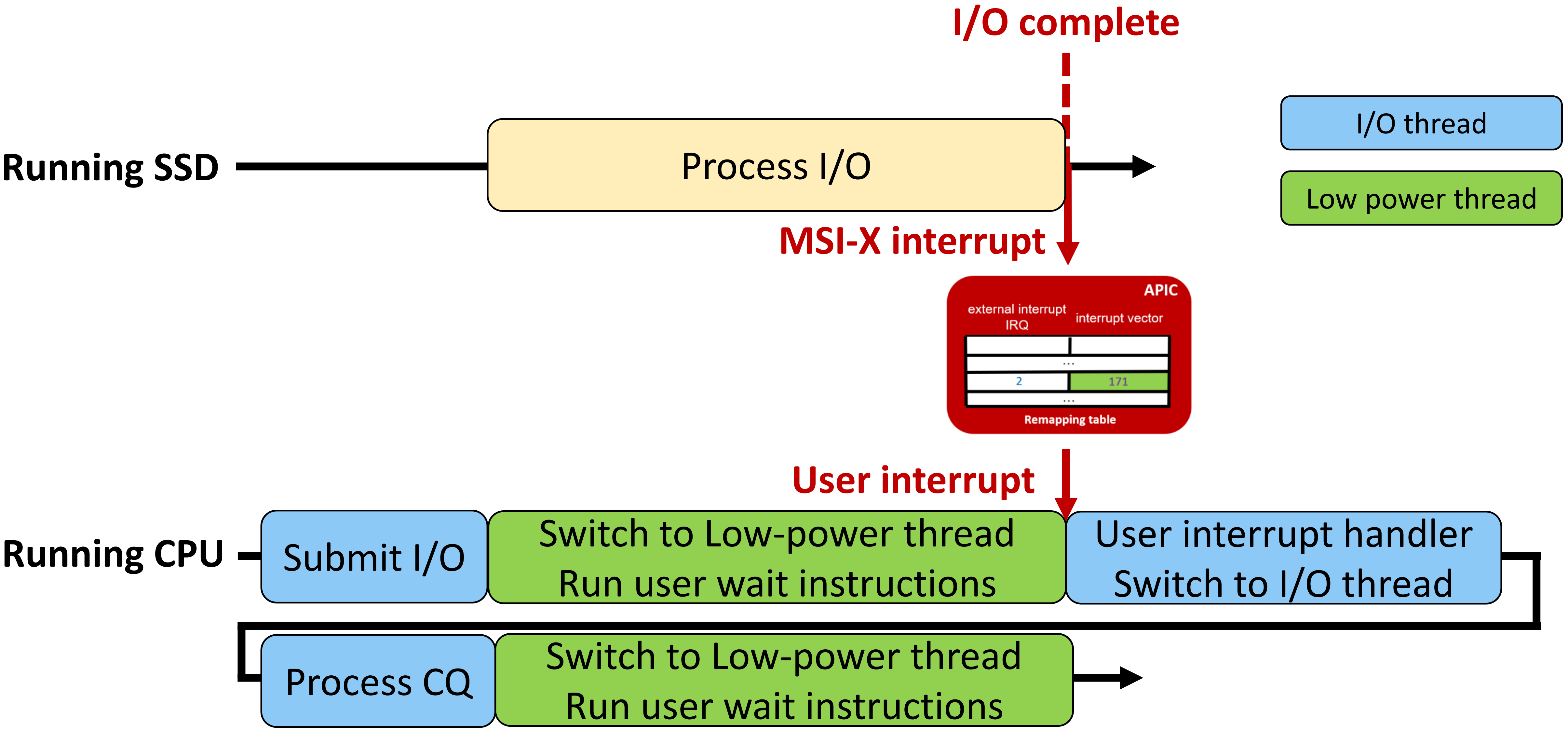
APIC

external interrupt IRQ	interrupt vector
…	
2	171
…	

Remapping table

Modified

SPDK+： the whole process



I/O complete

Running SSD

Process I/O

MSI-X interrupt

APIC

external interrupt IRQ	interrupt vector
…	
2	171
…	

Remapping table

User interrupt

Running CPU

Submit I/O

Switch to Low-power thread
Run user wait instructions

User interrupt handler
Switch to I/O thread

Process CQ

Switch to Low-power thread
Run user wait instructions

🟦	I/O thread
🟩	Low power thread

Prototype and Testbed Setup

Testbed Configuration

Component	Configuration
CPU	Intel Xeon PLATINUM 8558, 48 cores 2.1 GHz without hyper-threading
NVMe SSD	Up to 7 × TiPro9000 1TB Rand read/Rand Write : 2000K IOPS/1800K IOPS
OS	Ubuntu 22.04 LTS, modified Linux v6.8.10
Frequency scaling governor	Ondemand
Software	Spdk_nvme_perf v24.09
Power measure	Model Specific Register 64DH

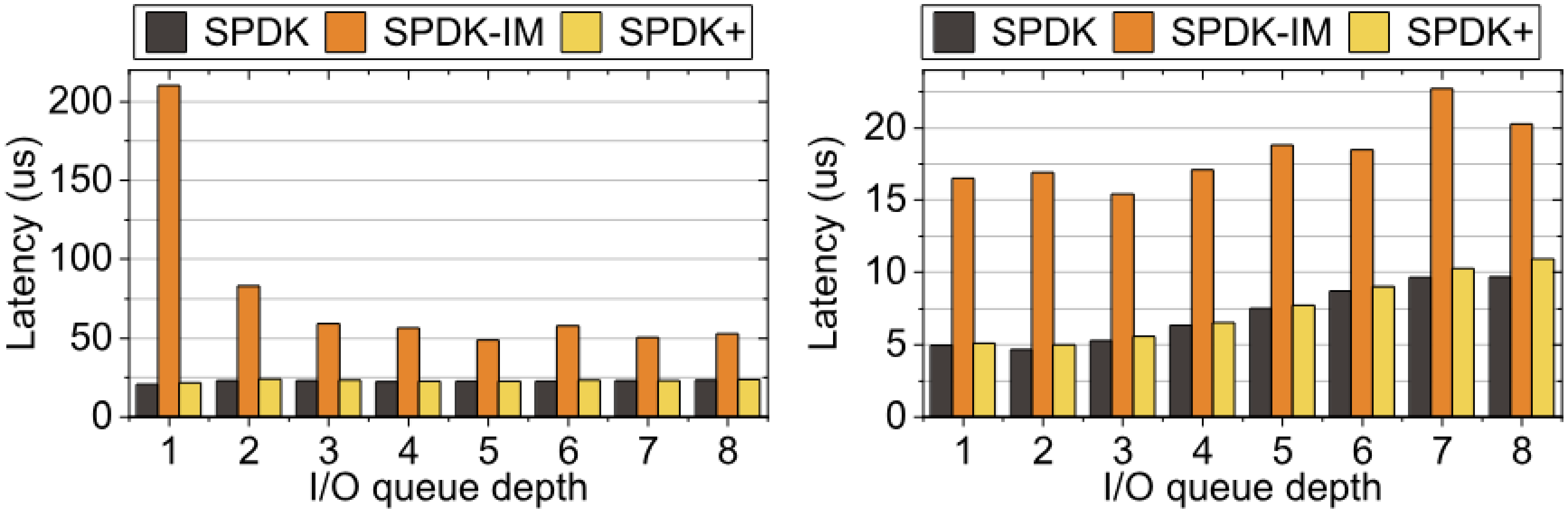
Definition of CPU Power Efficiency

$$CPU\ Power\ Efficiency = \frac{Avg\ IOPS}{Avg\ Power}$$

Subjects

Abbreviation	Description
SPDK	Using polling method
SPDK-IM	Using interrupt method
SPDK+	Our work using user interrupt method

Latency

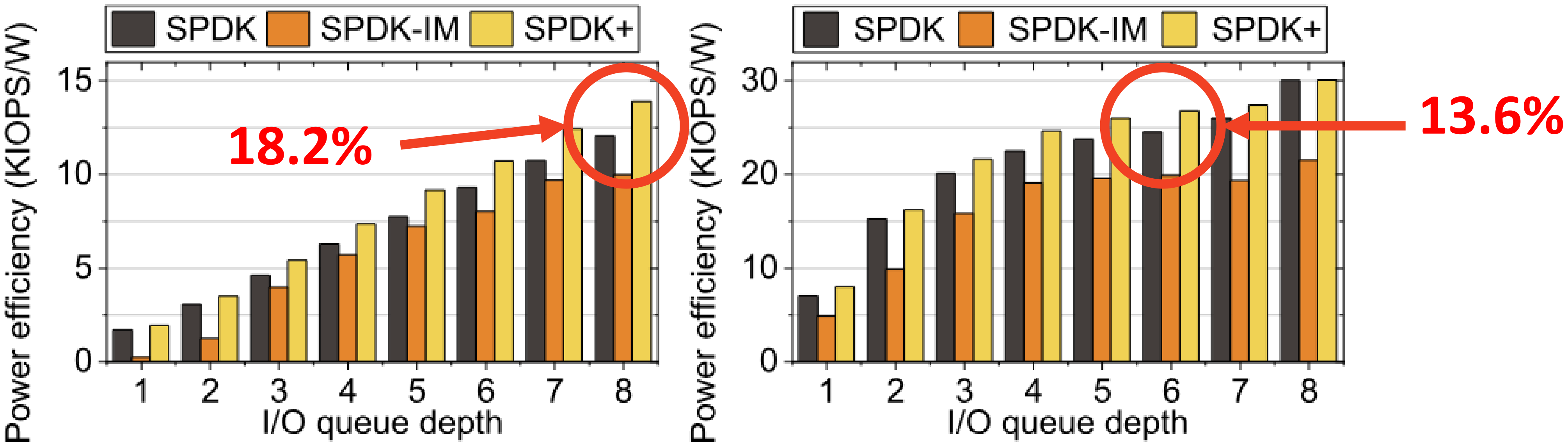


(a) 4KiB random read @ 7 cores.

(b) 4KiB random write @ 7 cores.

The latency of SPDK+ is almost the same as that of SPDK

Power Efficiency



18.2%

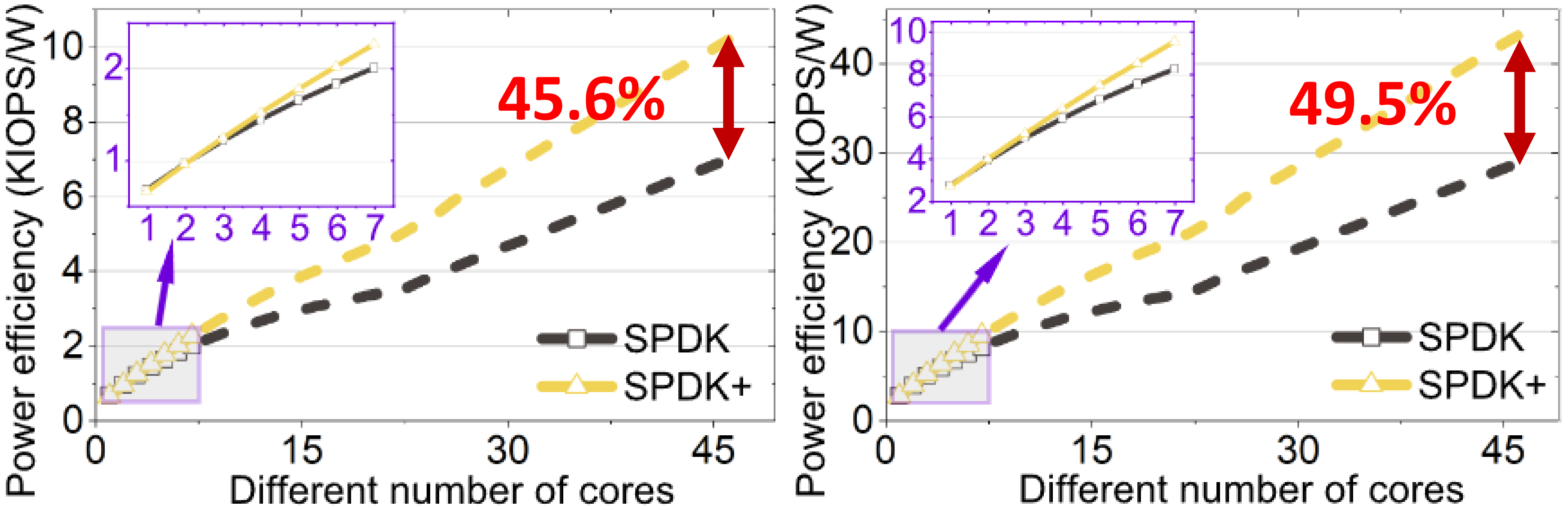
13.6%

(a) 4KiB random read @ 7 cores.

(b) 4KiB random write @ 7 cores.

SPDK+ improves power efficiency by up to 18.2%

Scalability



(a) 4KiB random read @ 1-46 cores. (b) 4KiB random write @ 1-46 cores.

As the number of cores increases, the CPU power efficiency of SPDK+ is further enhanced

Summary

- **Background:** The current I/O software stack does not achieve the best power efficiency in the case of **small I/O and shallow queues**
- **SPDK+:** Optimizing CPU power efficiency in the small IO shallow queue
- **Insights**:
  - The utilization rate of the polling mechanism is very low
  - The poor efficiency of interruption is due to the high interruption delay
- **Designs**:
  - **User interrupt** reduces interruption delay
  - **User wait instructions** reduce IO power consumption
  - The user-mode scheduling framework is used to connect the above two designs
- Significantly improve power efficiency while keeping the delay unchanged

Thanks & QA

Email: littledictionaryled@gmail.com

Endian Li, Shushu Yi, Li Peng, Qiao Li, Diyu Zhou, Zhenlin Wang, Xiaolin Wang,Bo Mao, Yingwei Luo, Ke Zhou, Jie Zhang

CHASELab
Computer Hardware And System Evolution Laboratory

PEKING UNIVERSITY

Michigan Tech