Is Bare-metal I/O Performance with User-defined Storage Drives Inside VMs Possible?

Benchmarking libvfio-user vs. Common Storage Virtualization Configurations

Sebastian Rolon & Oana Balmau

McGill UNIVERSITY

CHEOPS ‘23
May 8th, 2023
Rome, Italy

NUTANIX

Virtualization is everywhere

- Datasets keep growing
- We want storage to be efficient

Source: Intel [1]. An Intel Optane PCIe NVMe SSD.

Source: Google [2]. Inside a Google Datacenter.

aws

Azure

Google Cloud Platform

What is missing from storage virtualization?

1. Good performance with loose coupling
2. Rapid device prototyping
3. Live migration
4. Userspace drivers

McGill UNIVERSITY

What is missing from storage virtualization?

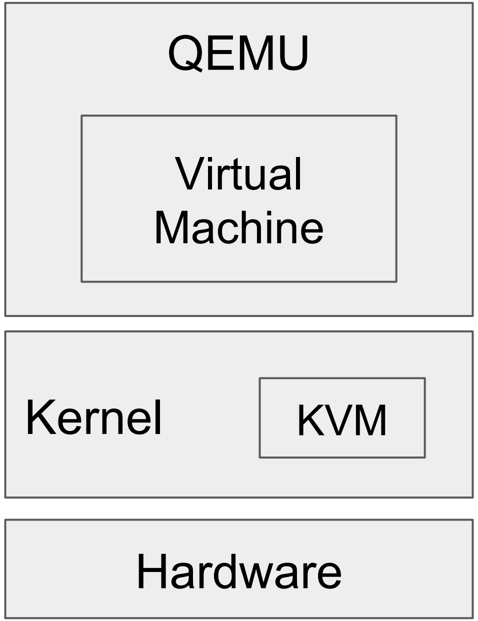
1. Good performance with loose coupling
2. Rapid device prototyping
3. Live migration
4. Userspace drivers

NUTANIX

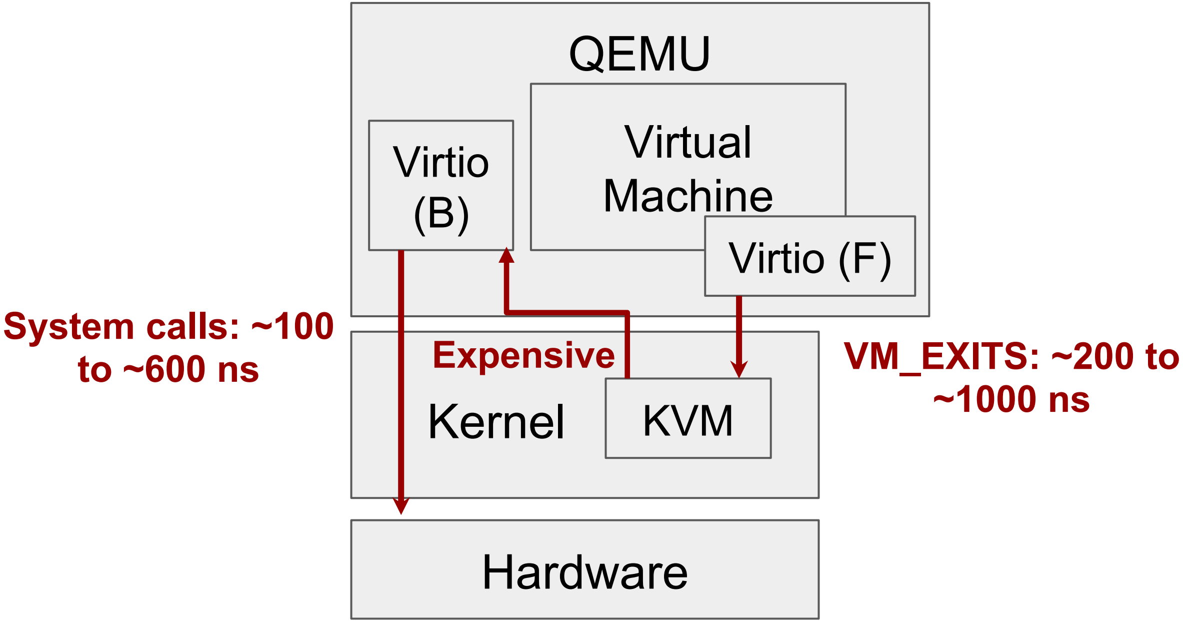
vfio-user

Can **vfio-user** be used as an alternative to current VM storage?

Virtualization and QEMU/KVM

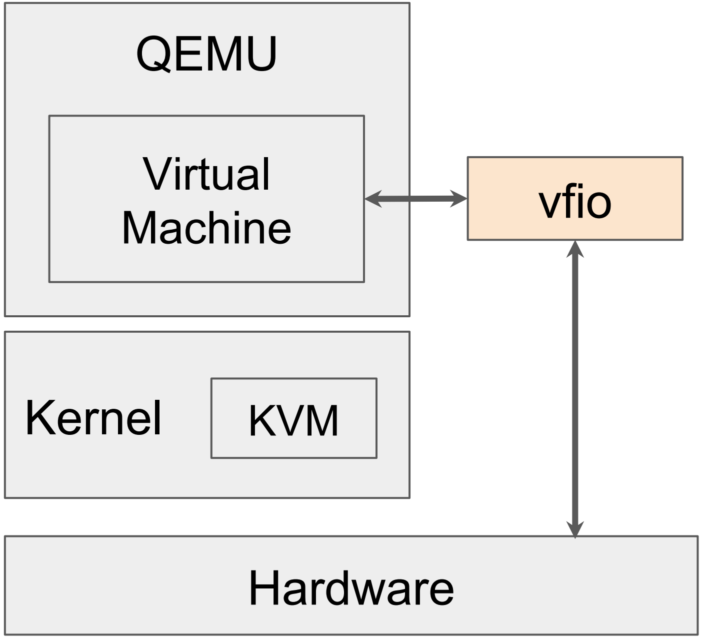


Context switches are expensive

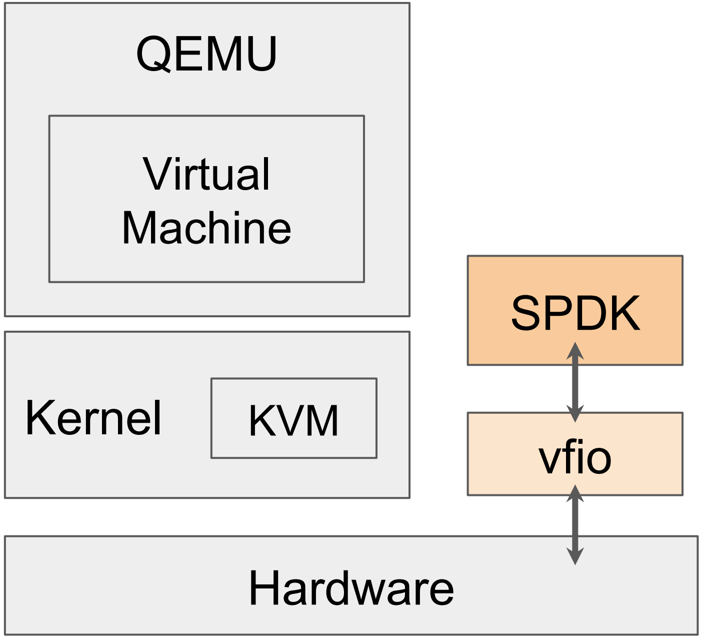


McGill UNIVERSITY

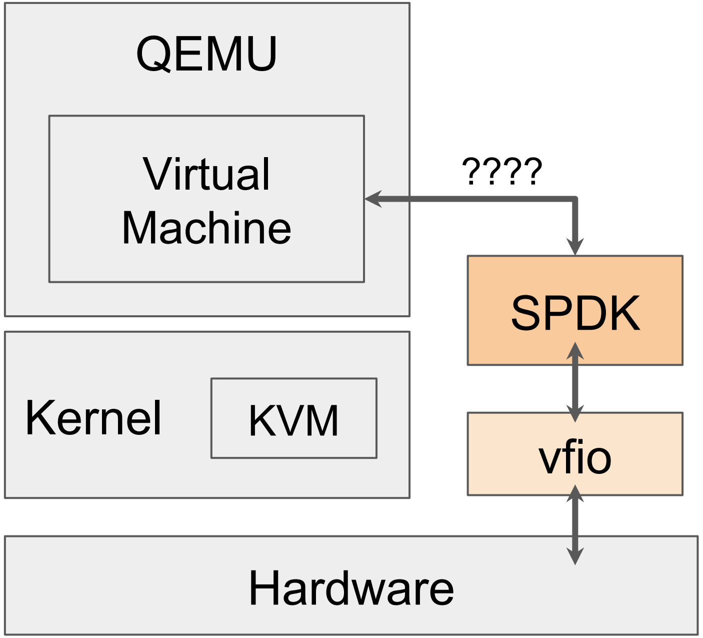
Userspace hardware access



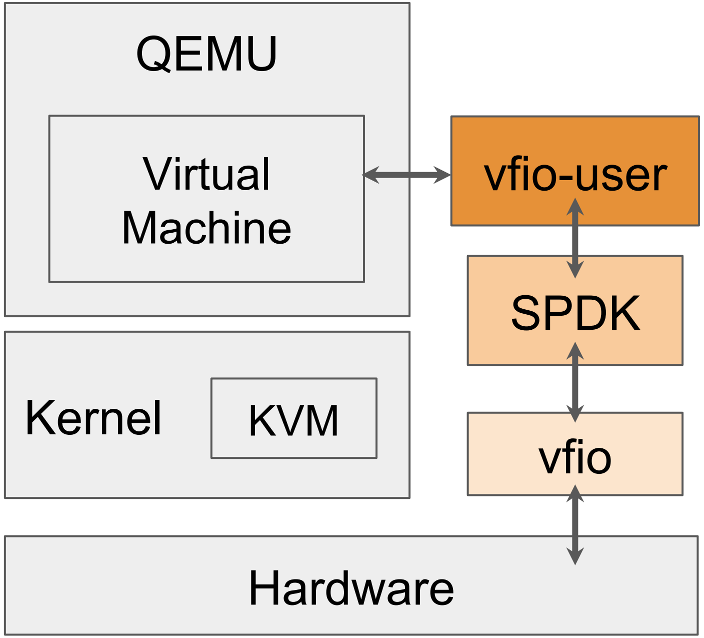
Abstracting hardware access



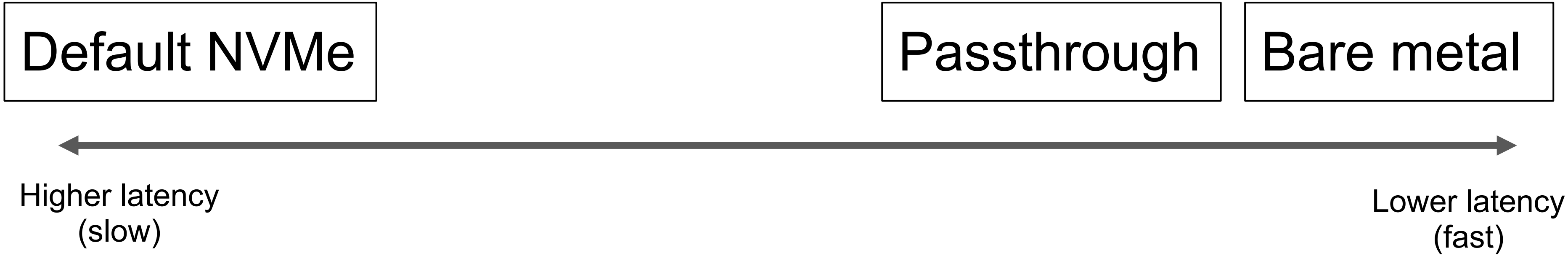
How do we connect after abstracting?



Vfio-user virtualizes hardware over a channel



I want my process to see an NVMe



Default NVMe

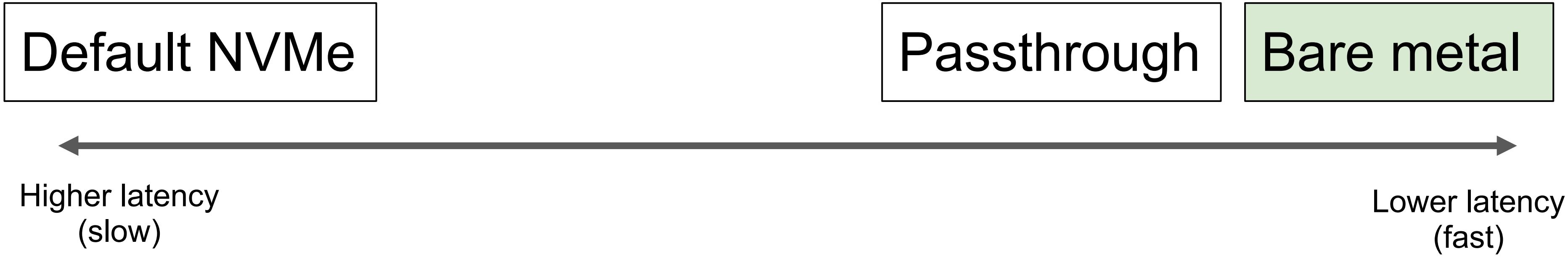
Passthrough

Bare metal

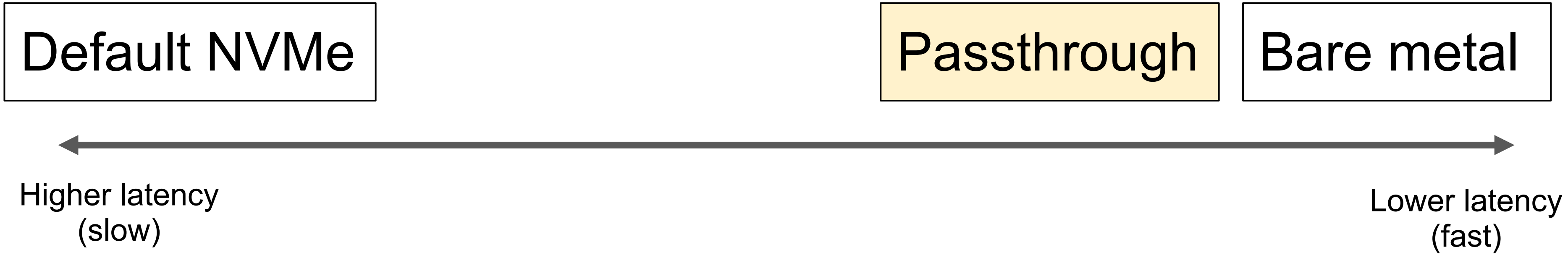
Higher latency
(slow)

Lower latency
(fast)

Fast and simple



Fast, involves some configuration



Default NVMe

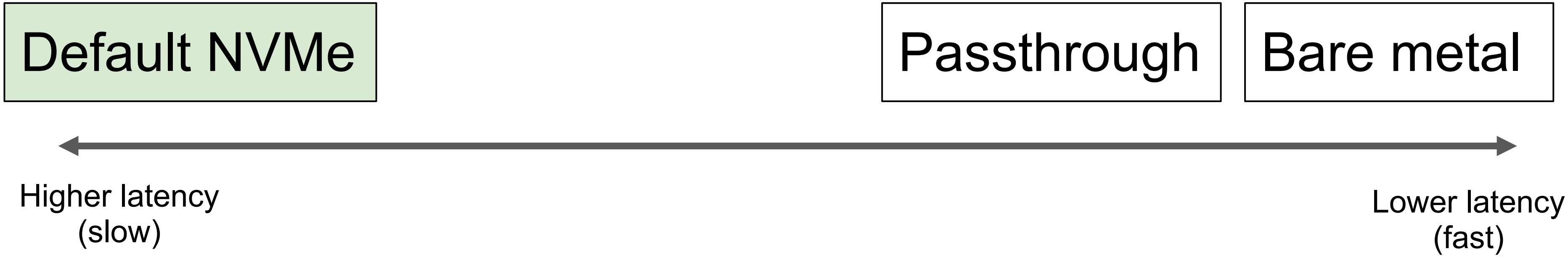
Passthrough

Bare metal

Higher latency
(slow)

Lower latency
(fast)

Slow, requires one command-line flag



Default NVMe

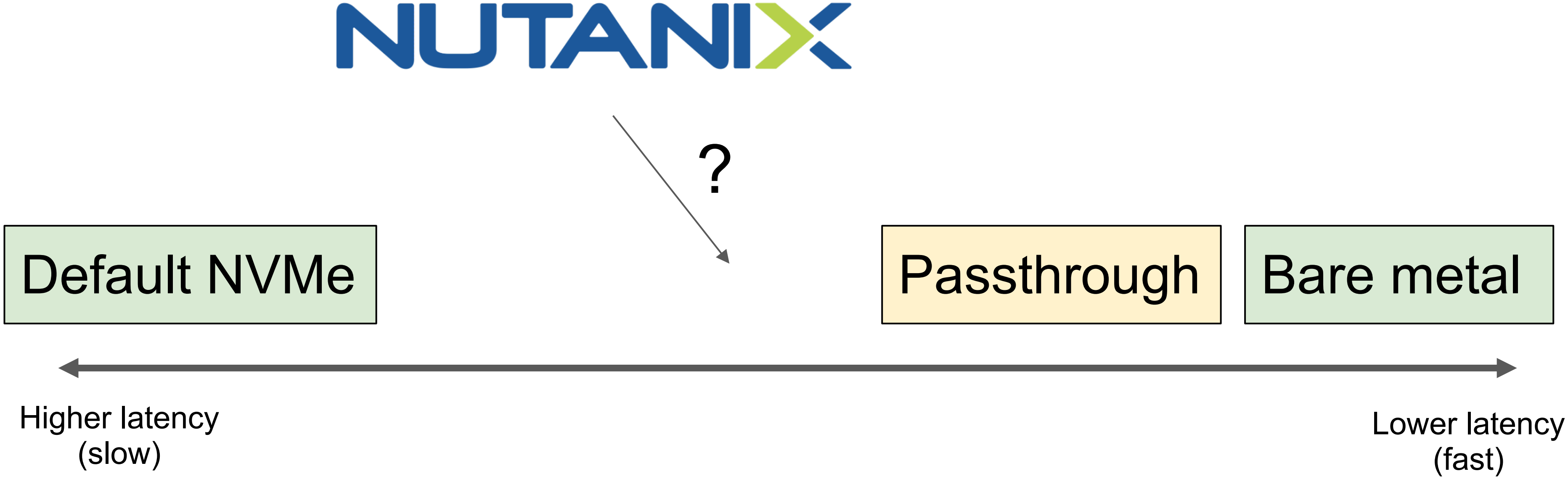
Passthrough

Bare metal

Higher latency
(slow)

Lower latency
(fast)

Where does vfio-user fit?



NUTANIX

?

Default NVMe

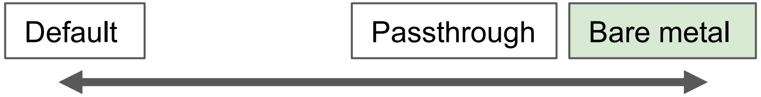
Passthrough

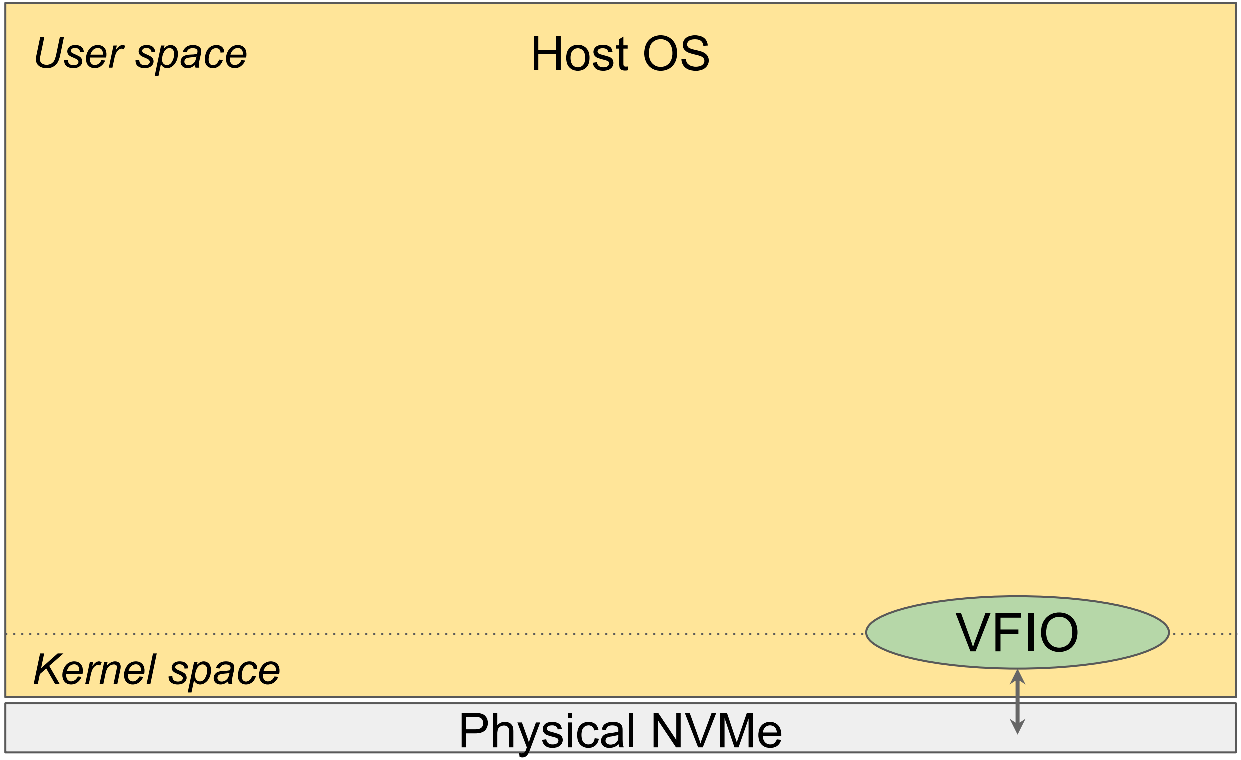
Bare metal

Higher latency
(slow)

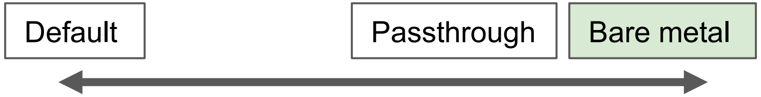
Lower latency
(fast)

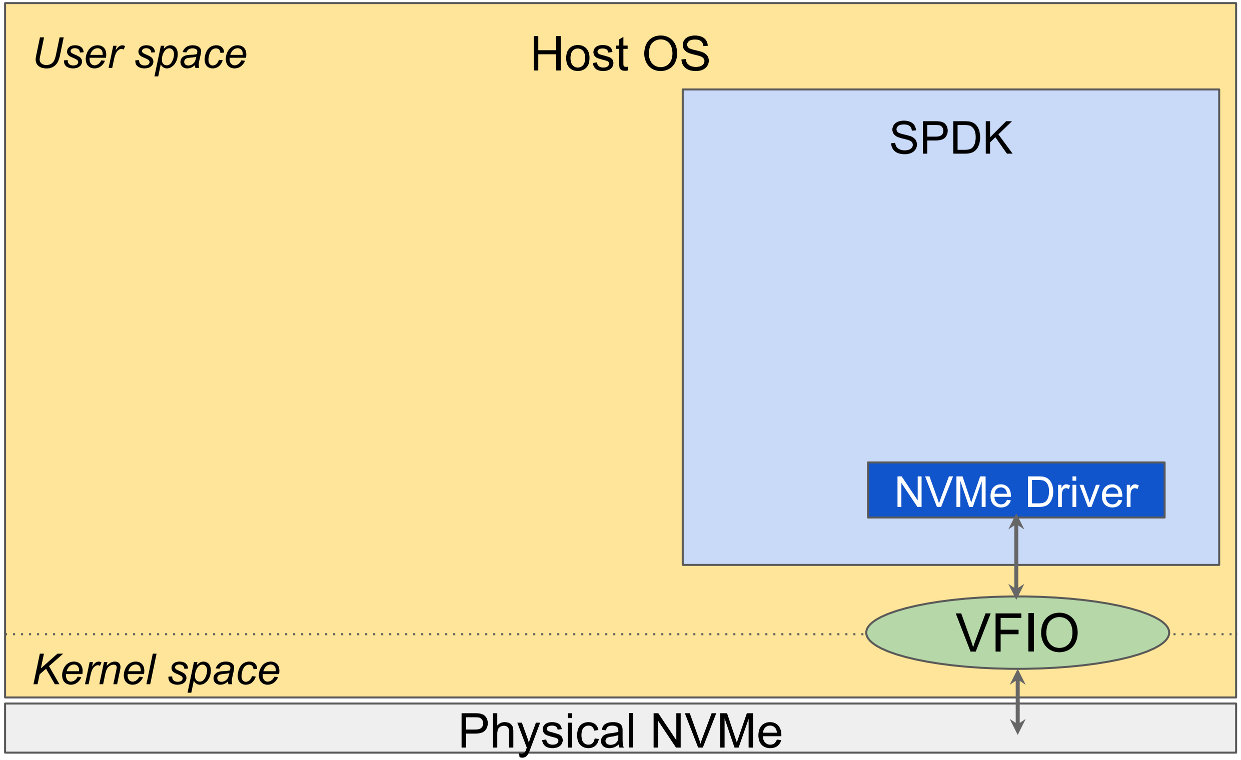
Bare metal configuration



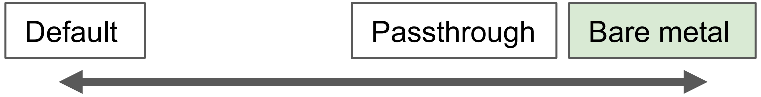


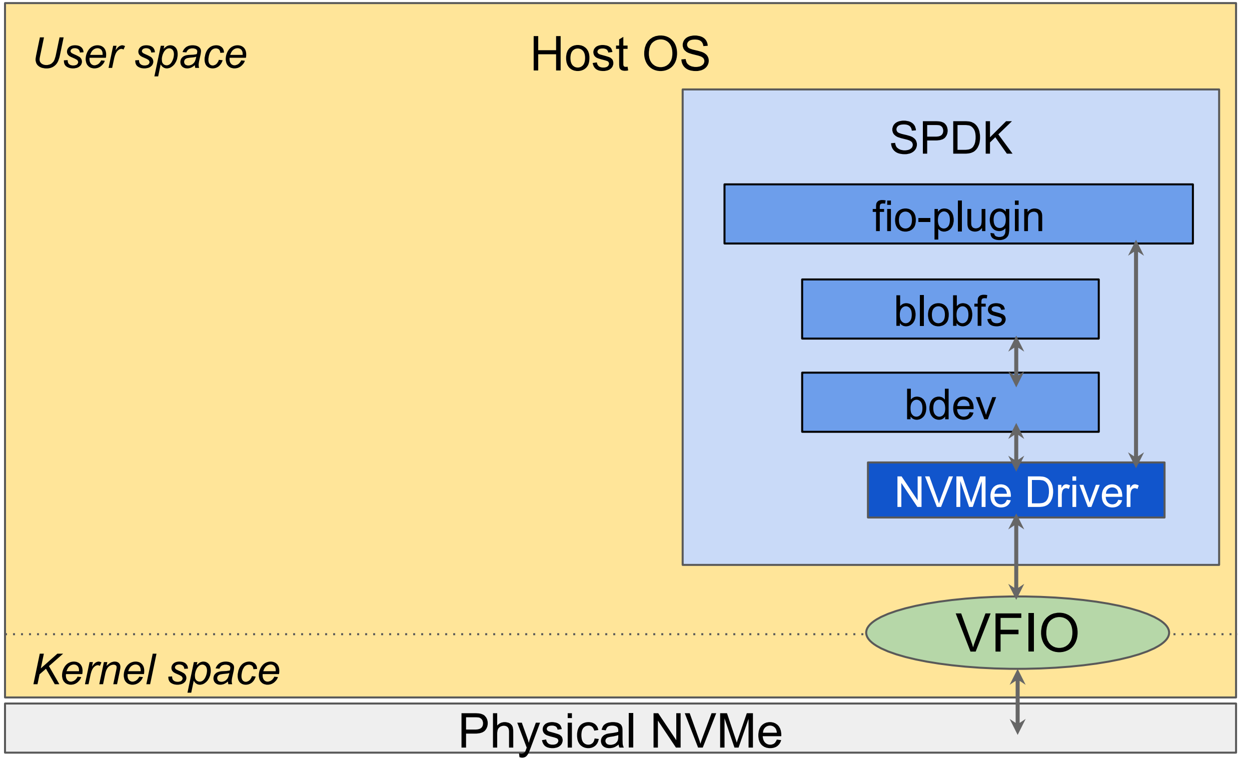
Bare metal configuration



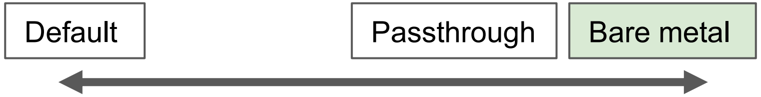


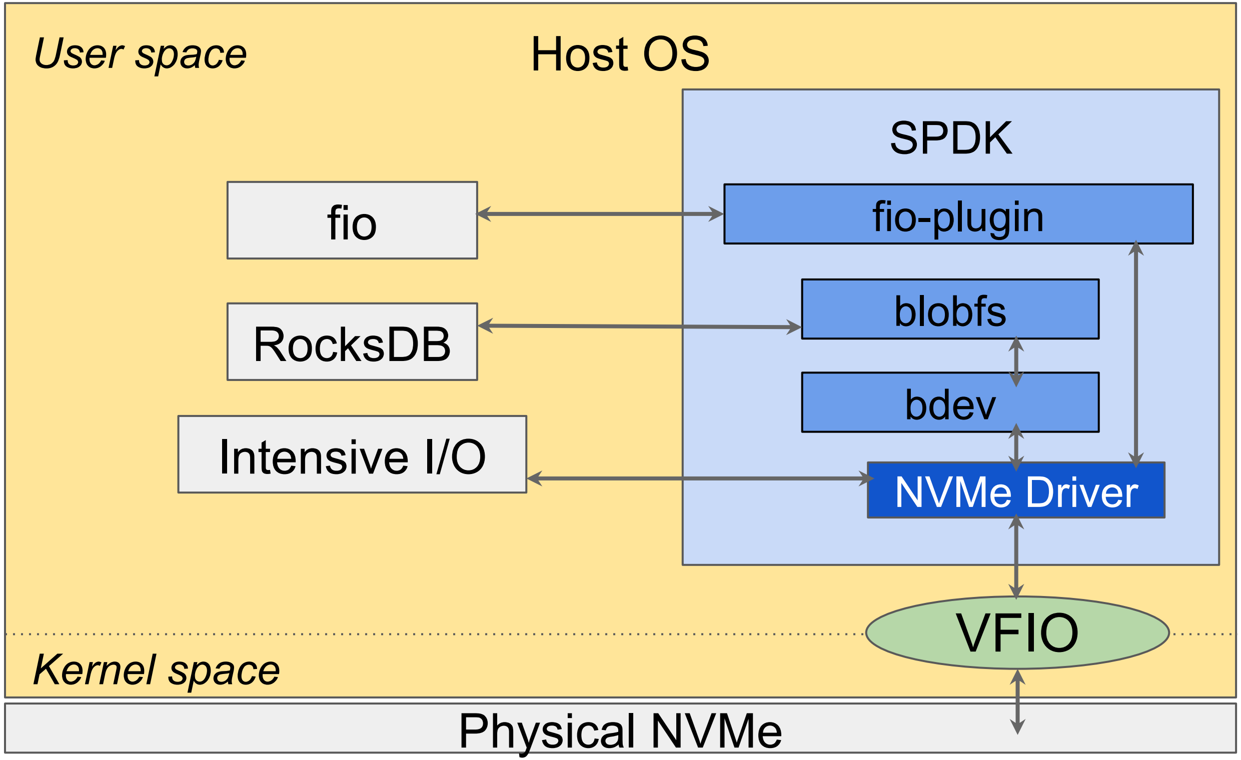
Bare metal configuration



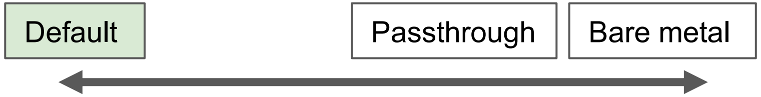


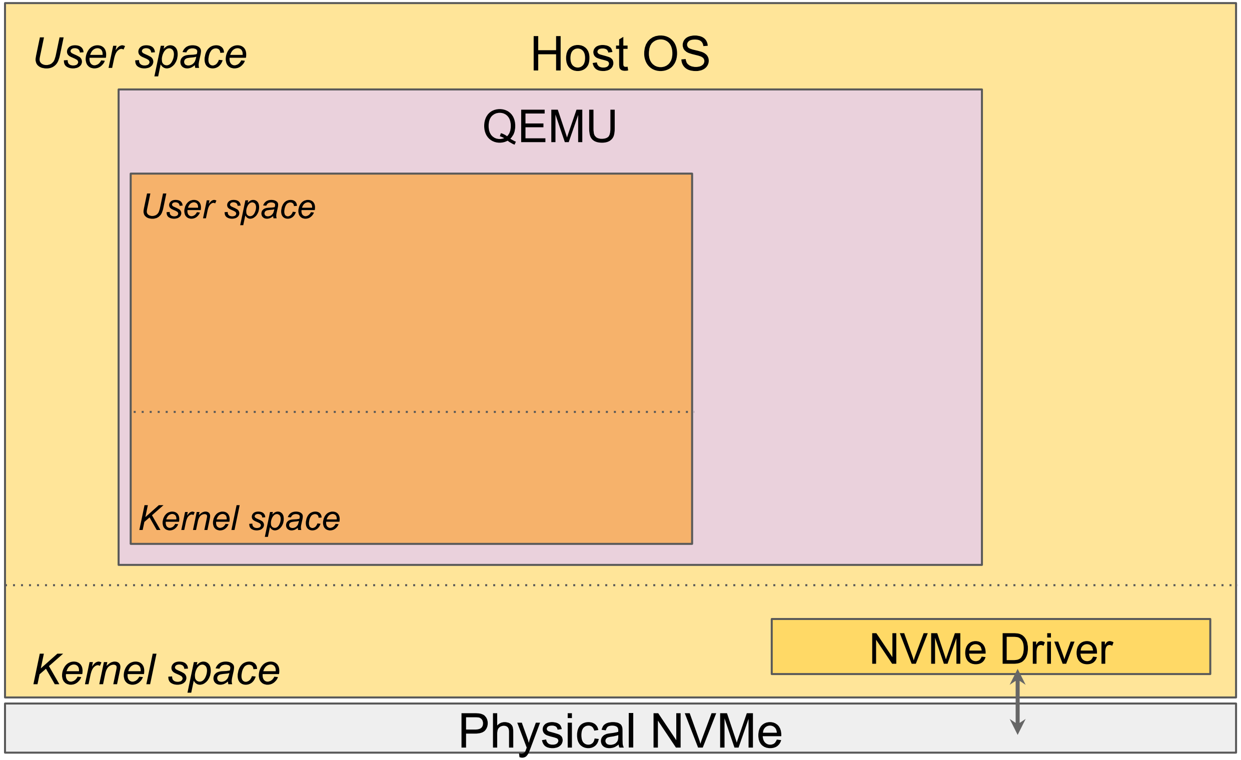
Bare metal configuration



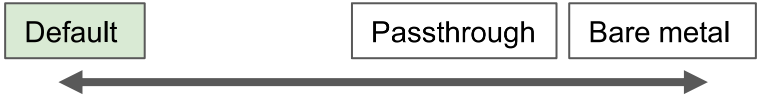


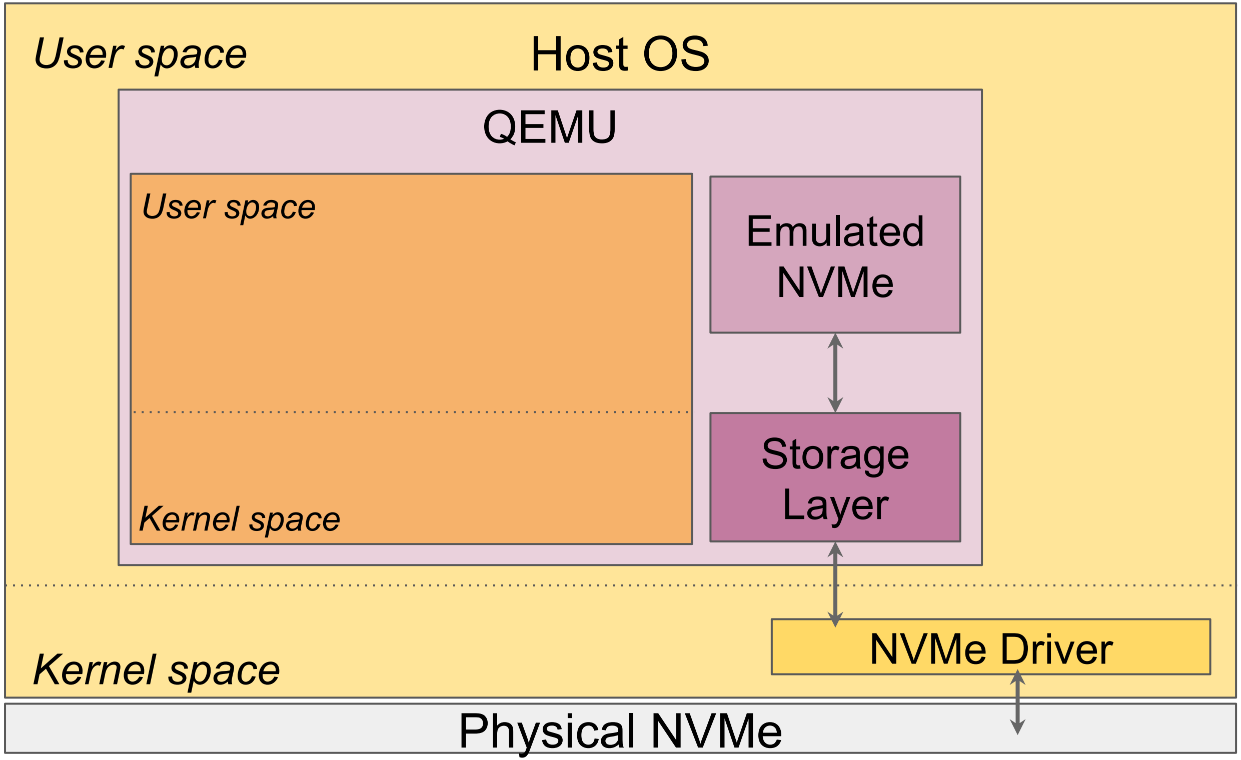
Default NVMe configuration



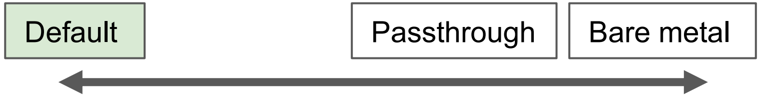


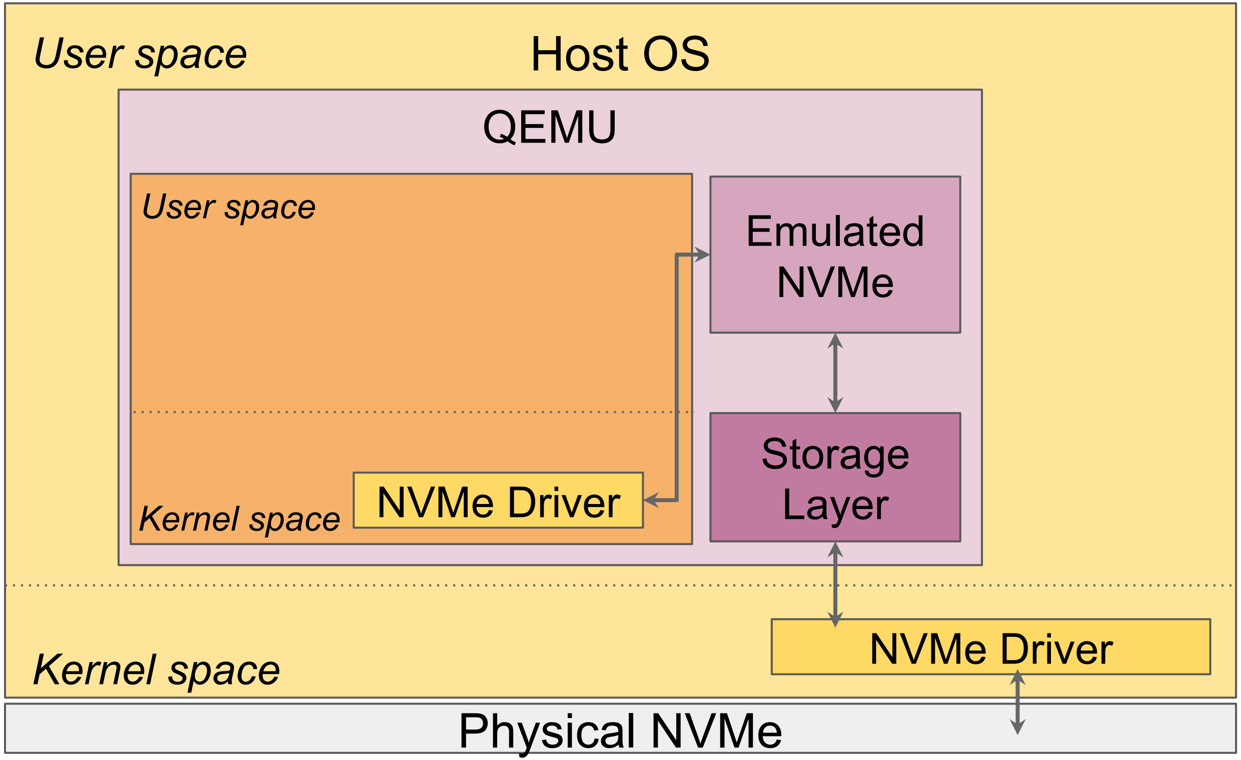
Default NVMe configuration



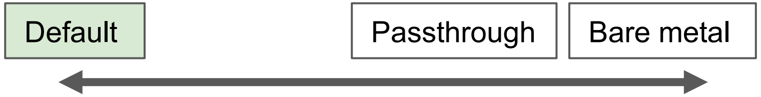


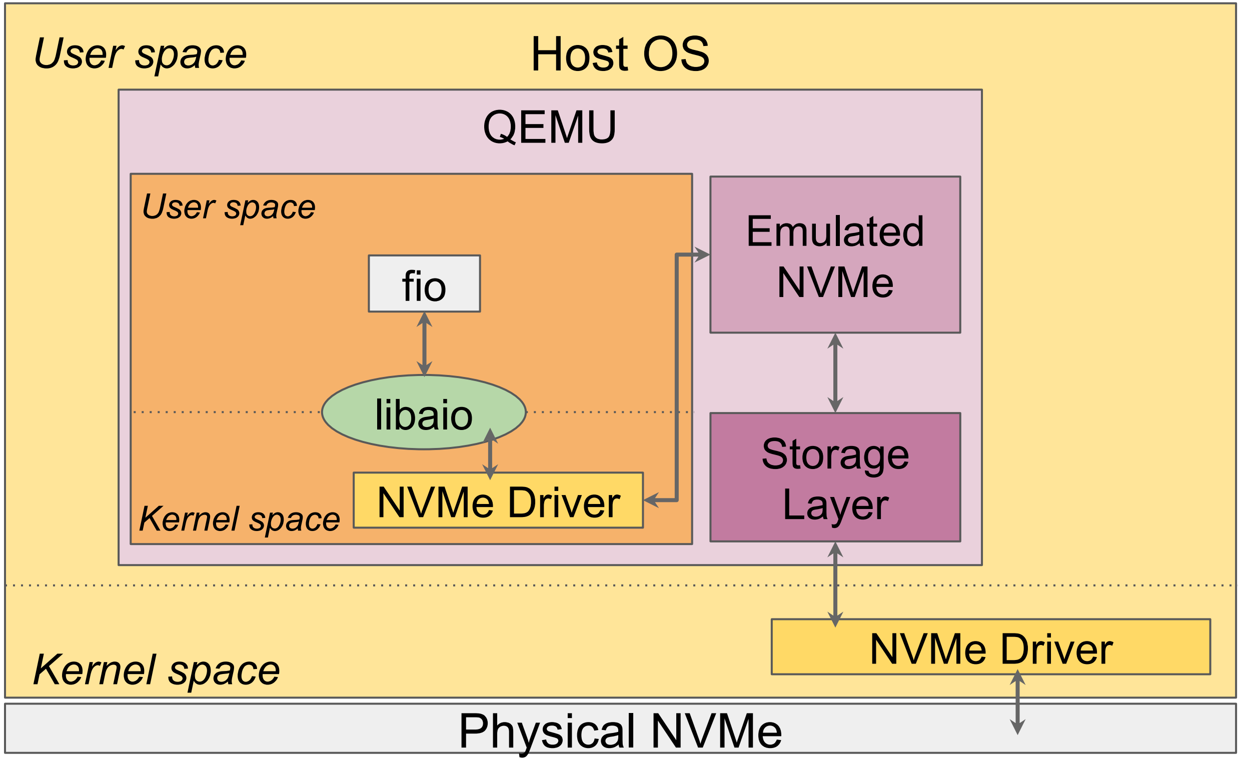
Default NVMe configuration



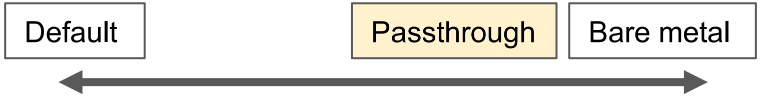


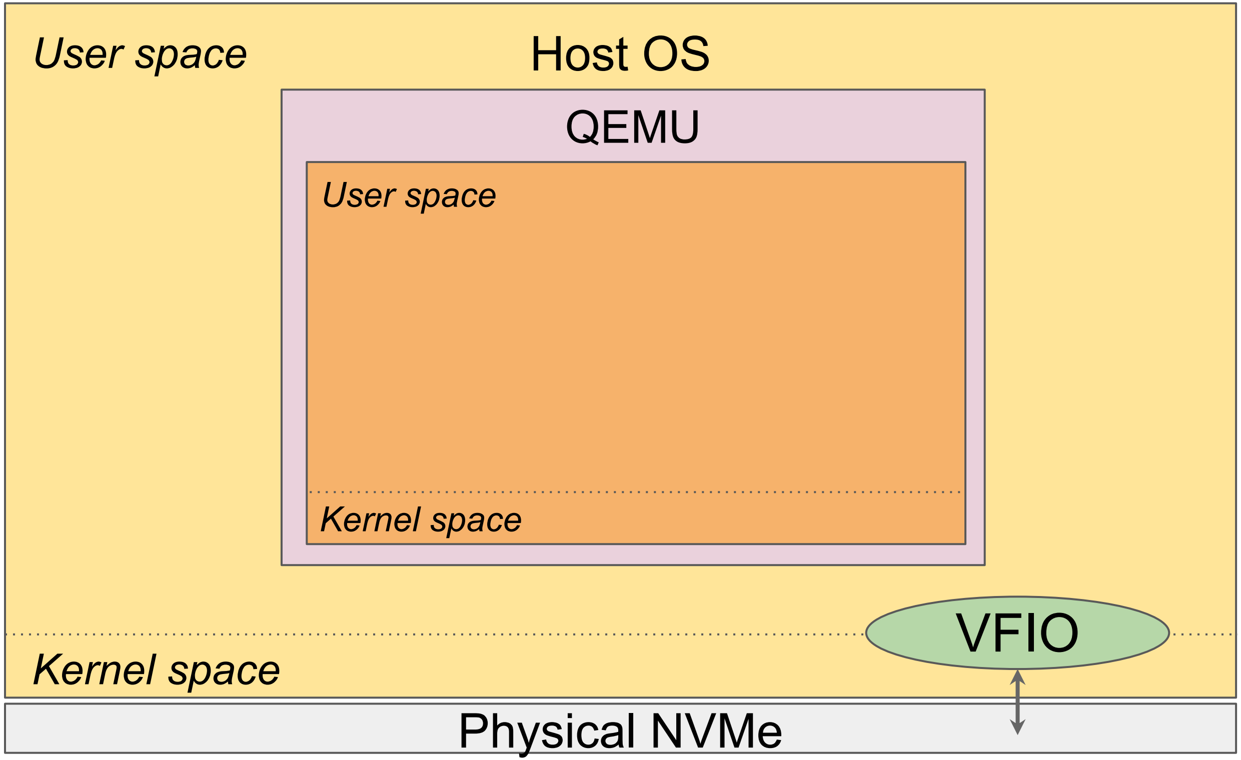
Default NVMe configuration



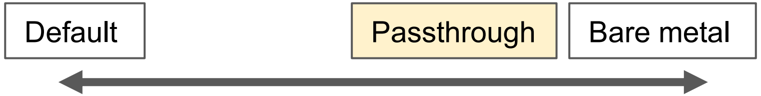


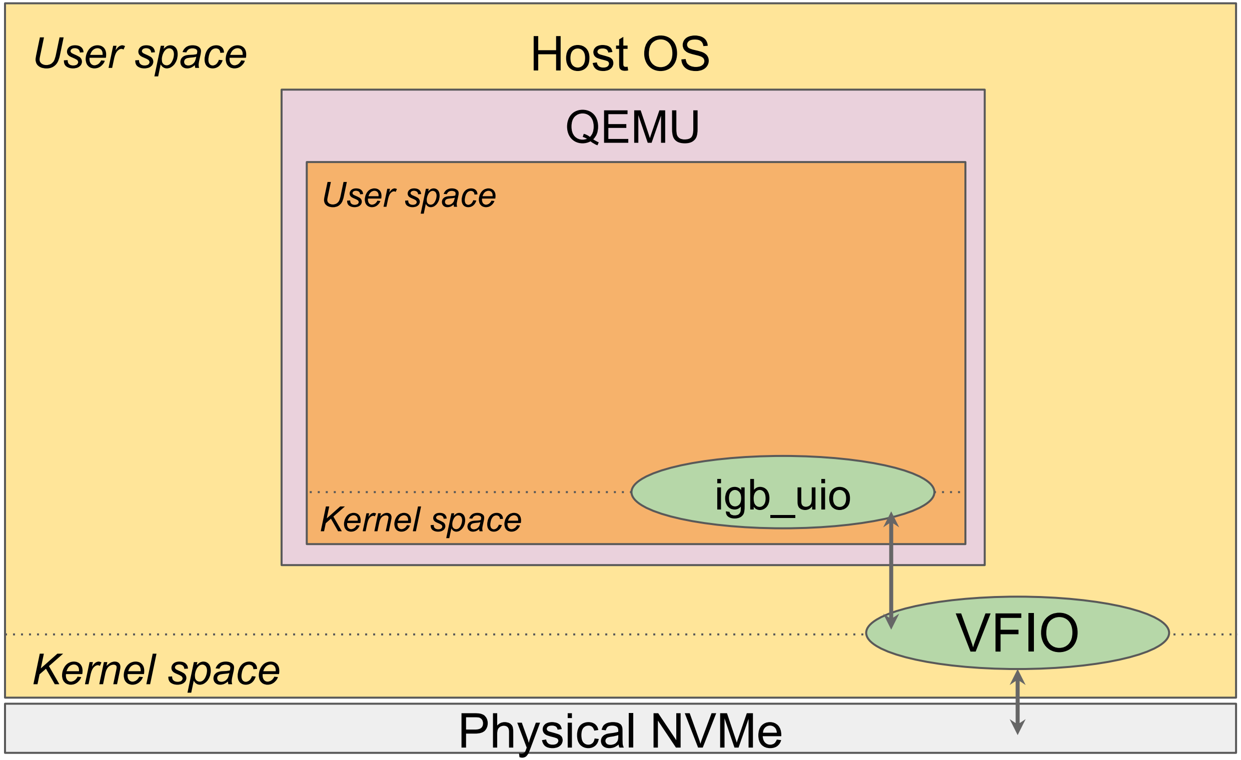
Passthrough configuration



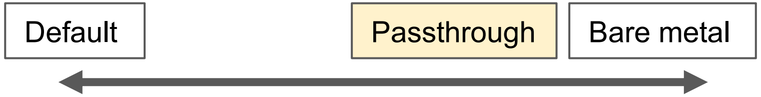


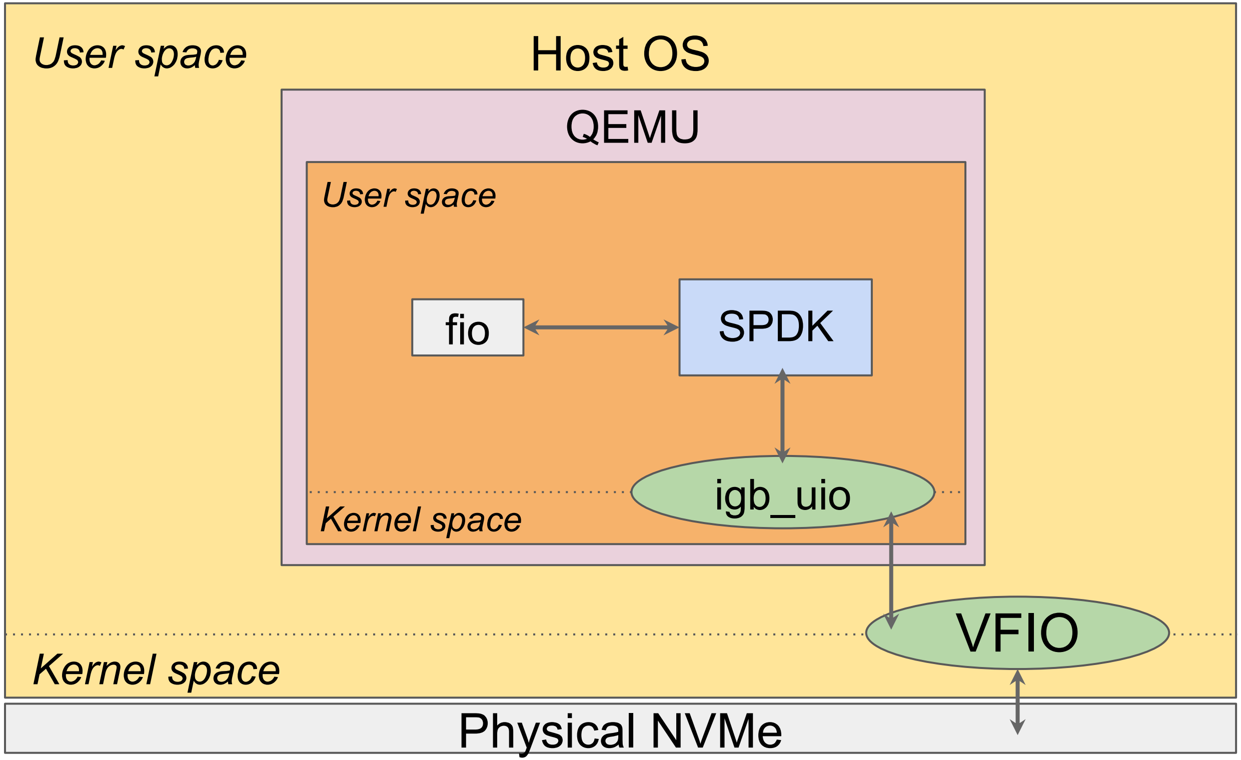
Passthrough configuration



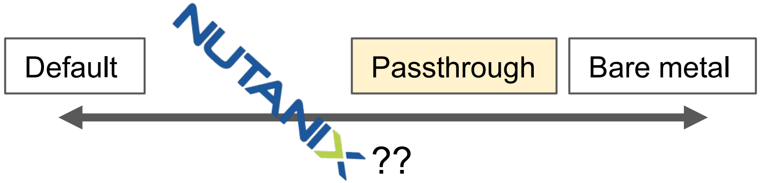


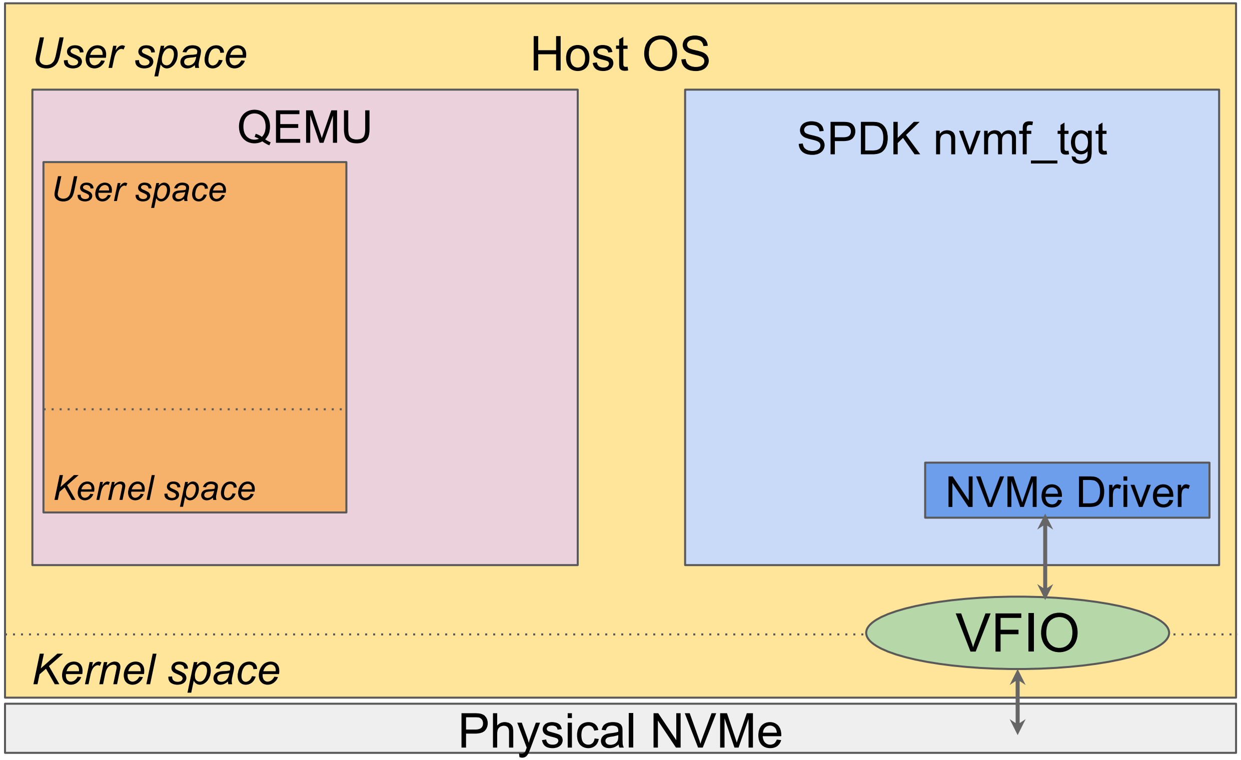
Passthrough configuration



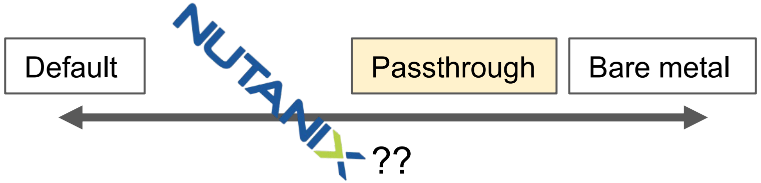


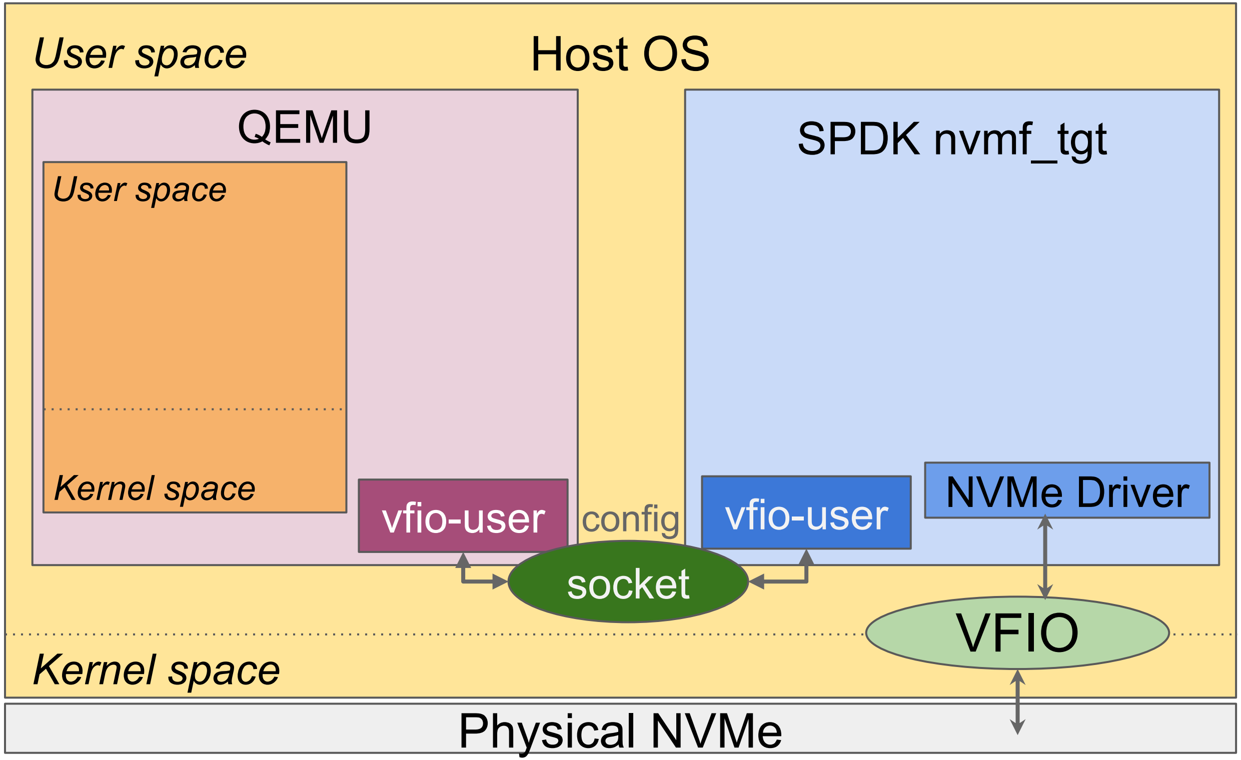
Vfio-user configuration



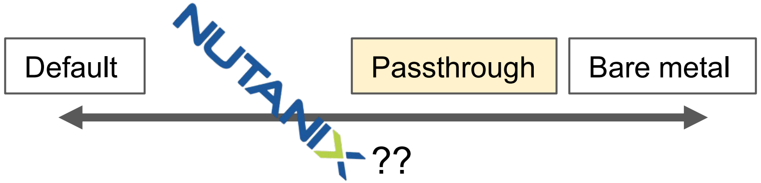


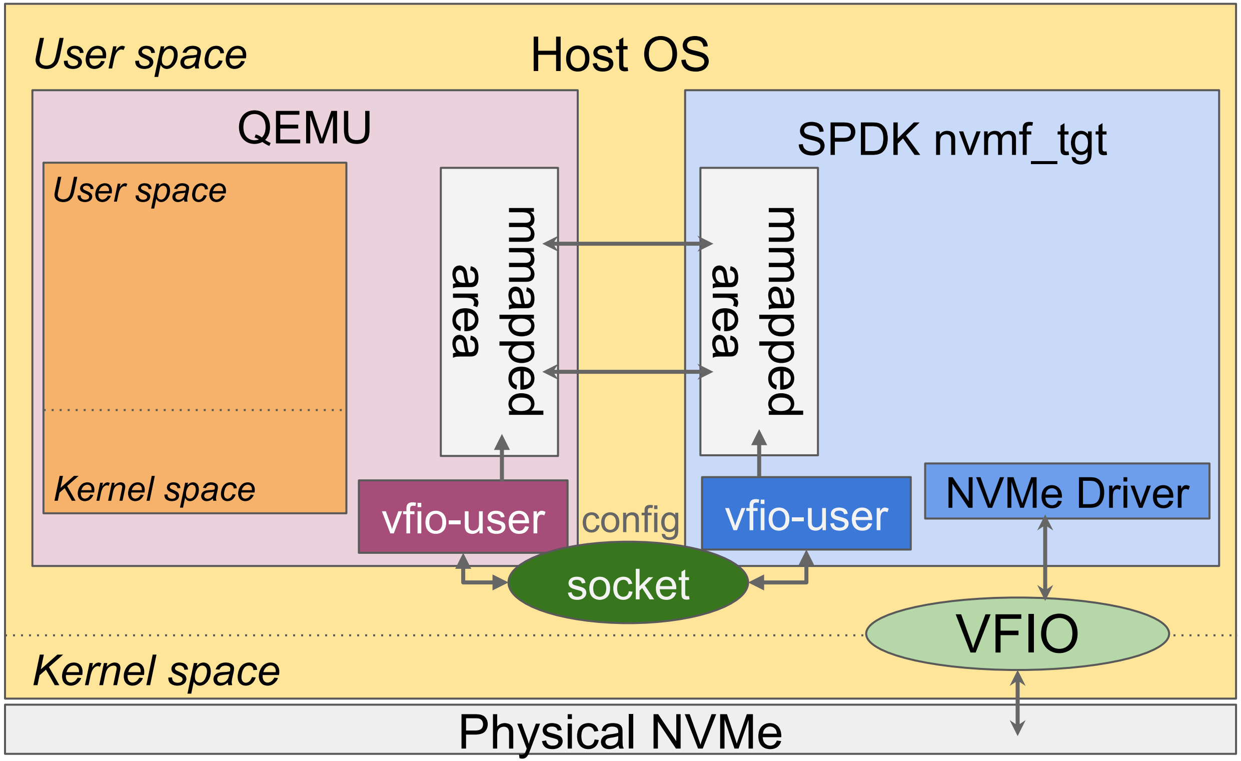
Vfio-user configuration



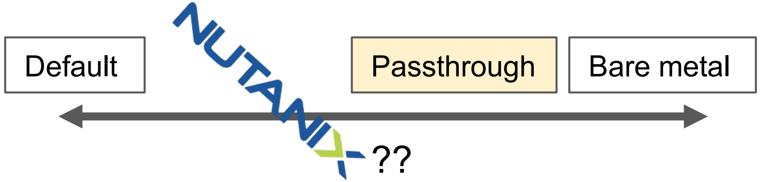


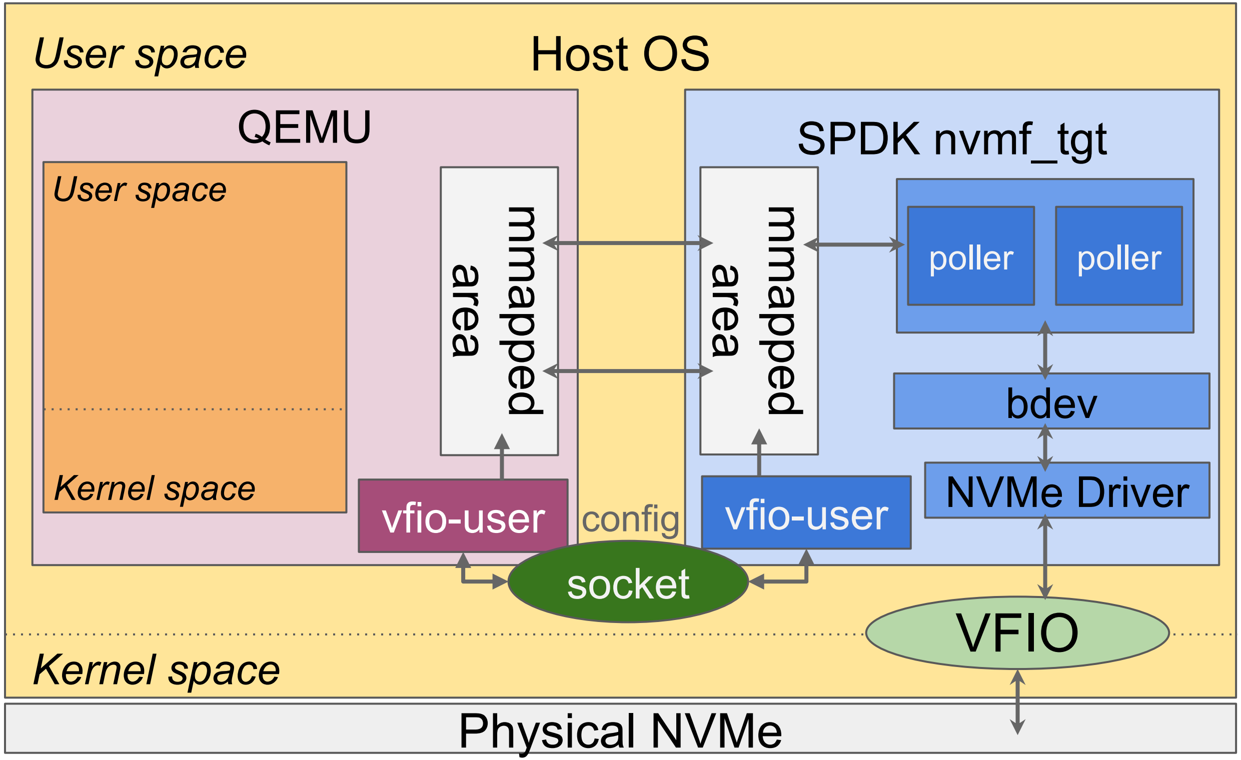
Vfio-user configuration



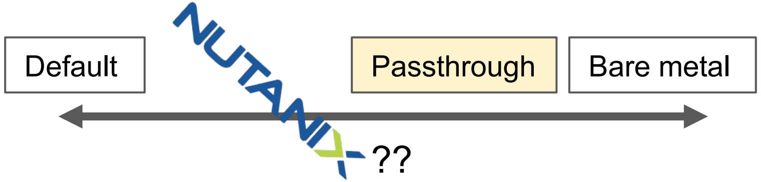


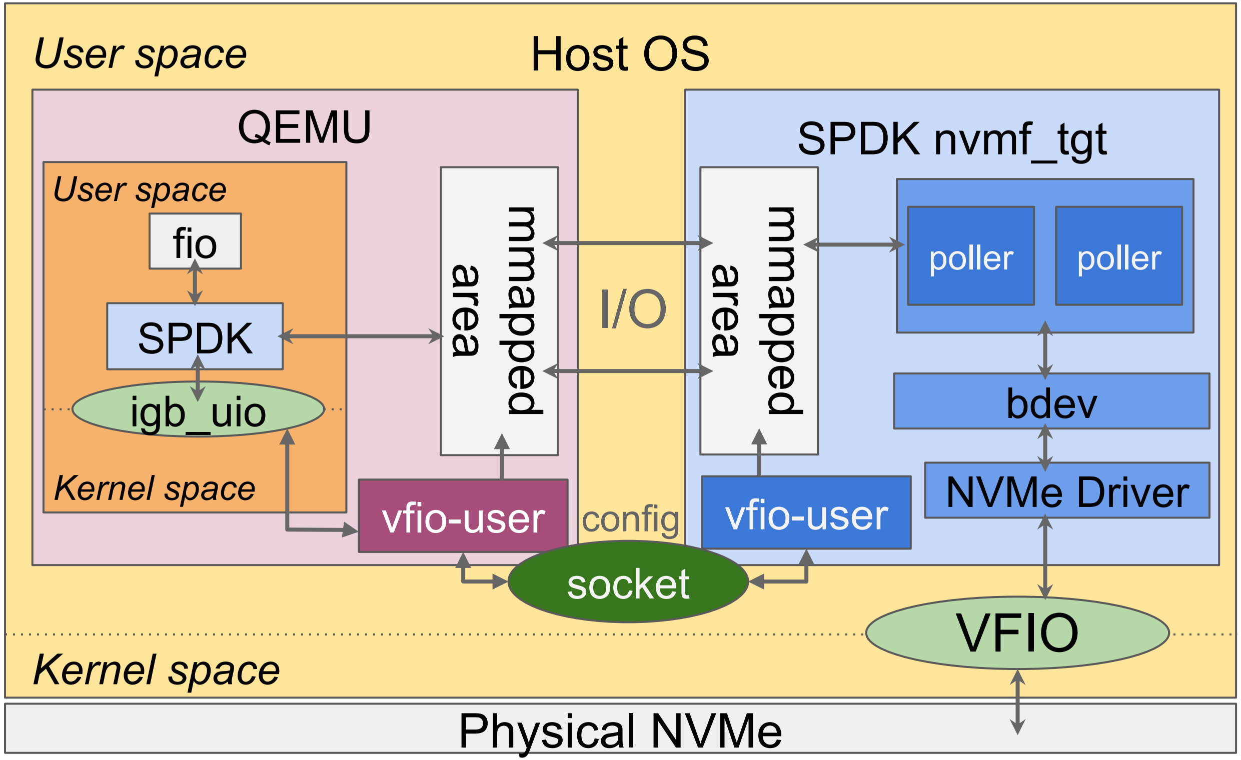
Vfio-user configuration





Vfio-user configuration

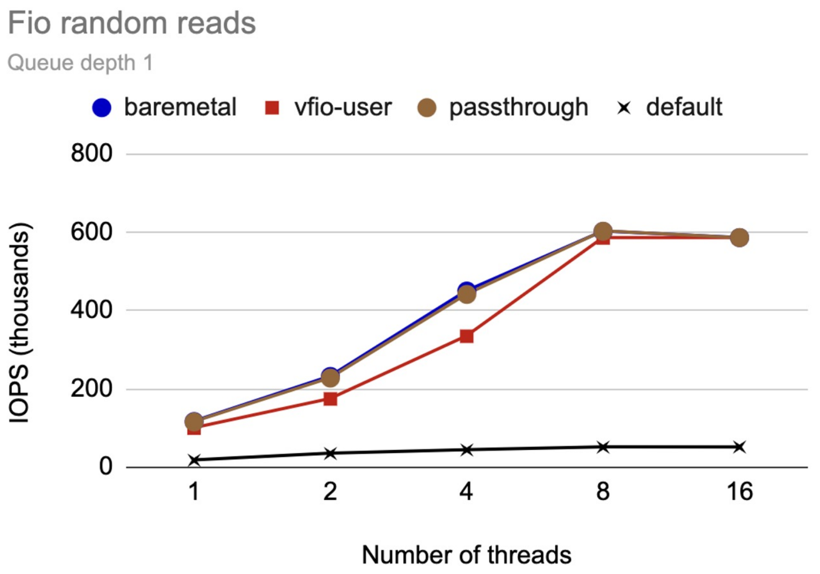


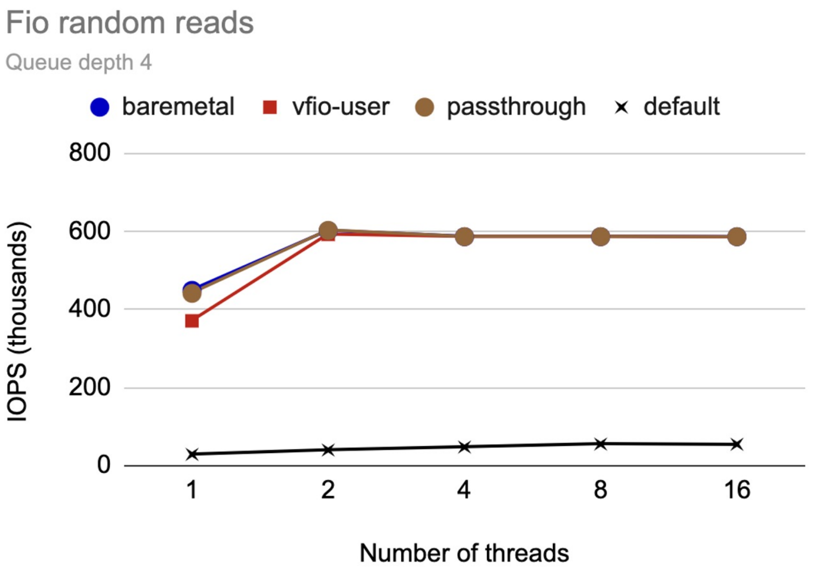


Experimental evaluation

Hardware Specifications	
CPU	36 Core Xeon Gold 6240L @ 2.40 GHz
Memory	768 GiB 3200MHz DDR4 DIMM
SSD	375GiB Dell Express Flash NVMe P4800X

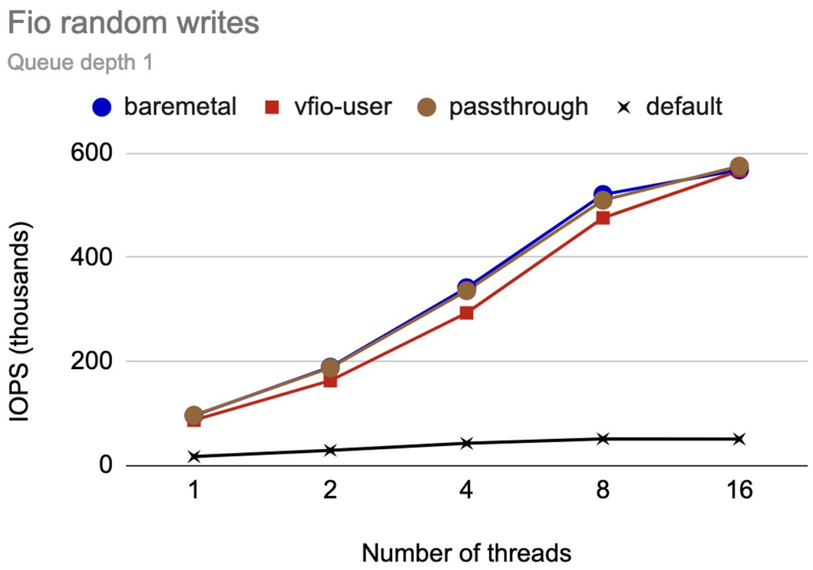
Results: fio random reads

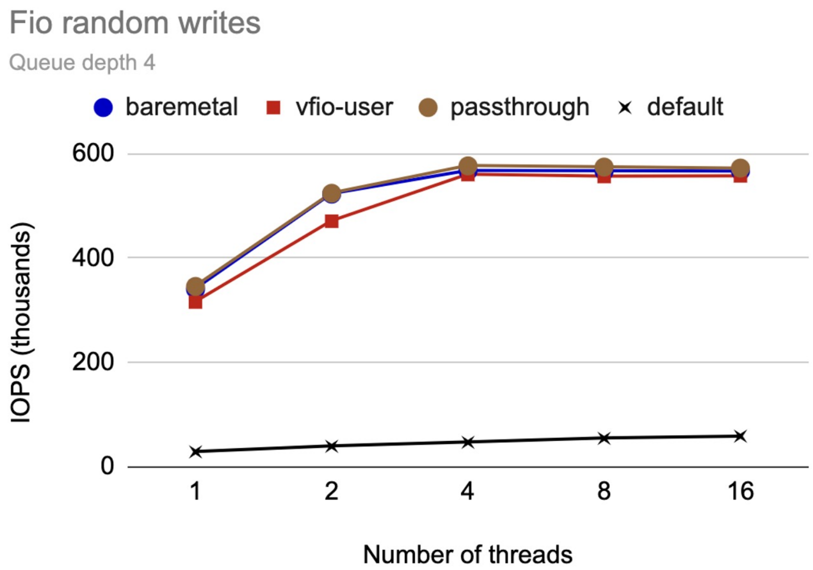




McGill
UNIVERSITY

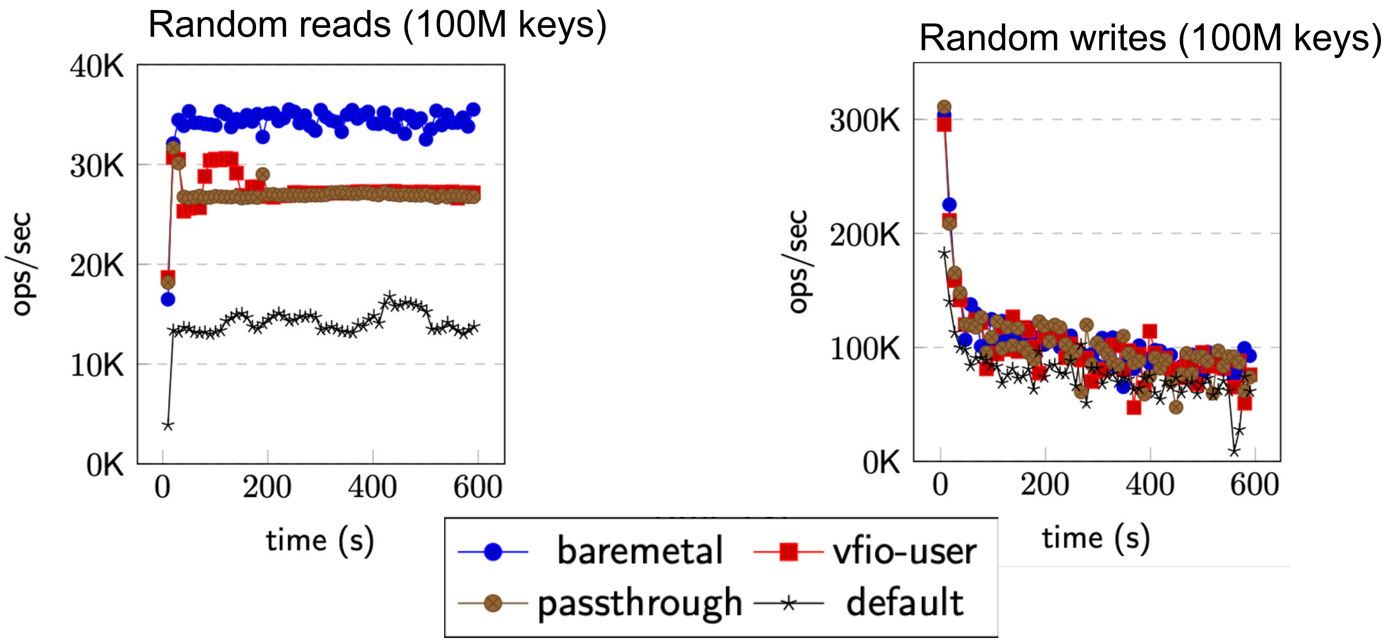
Results: fio random writes





McGill UNIVERSITY

Results: RocksDB benchmarks

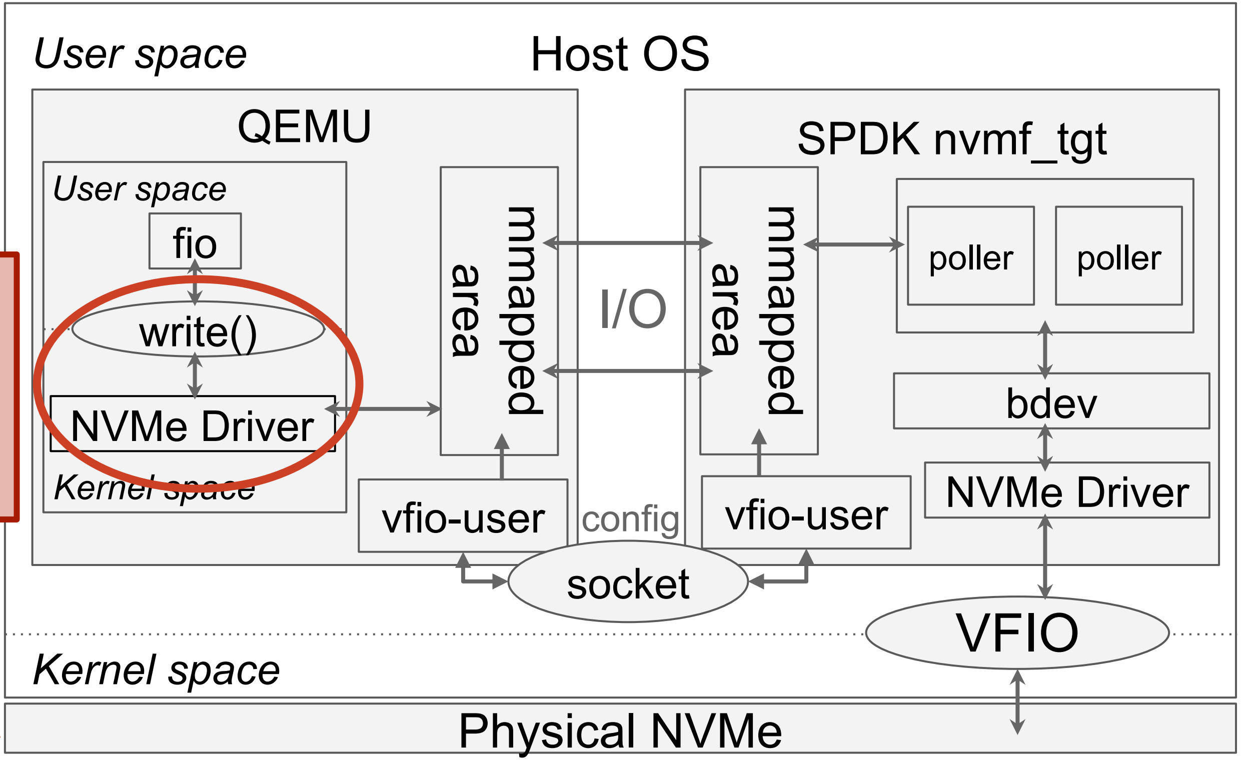


Random reads (100M keys)

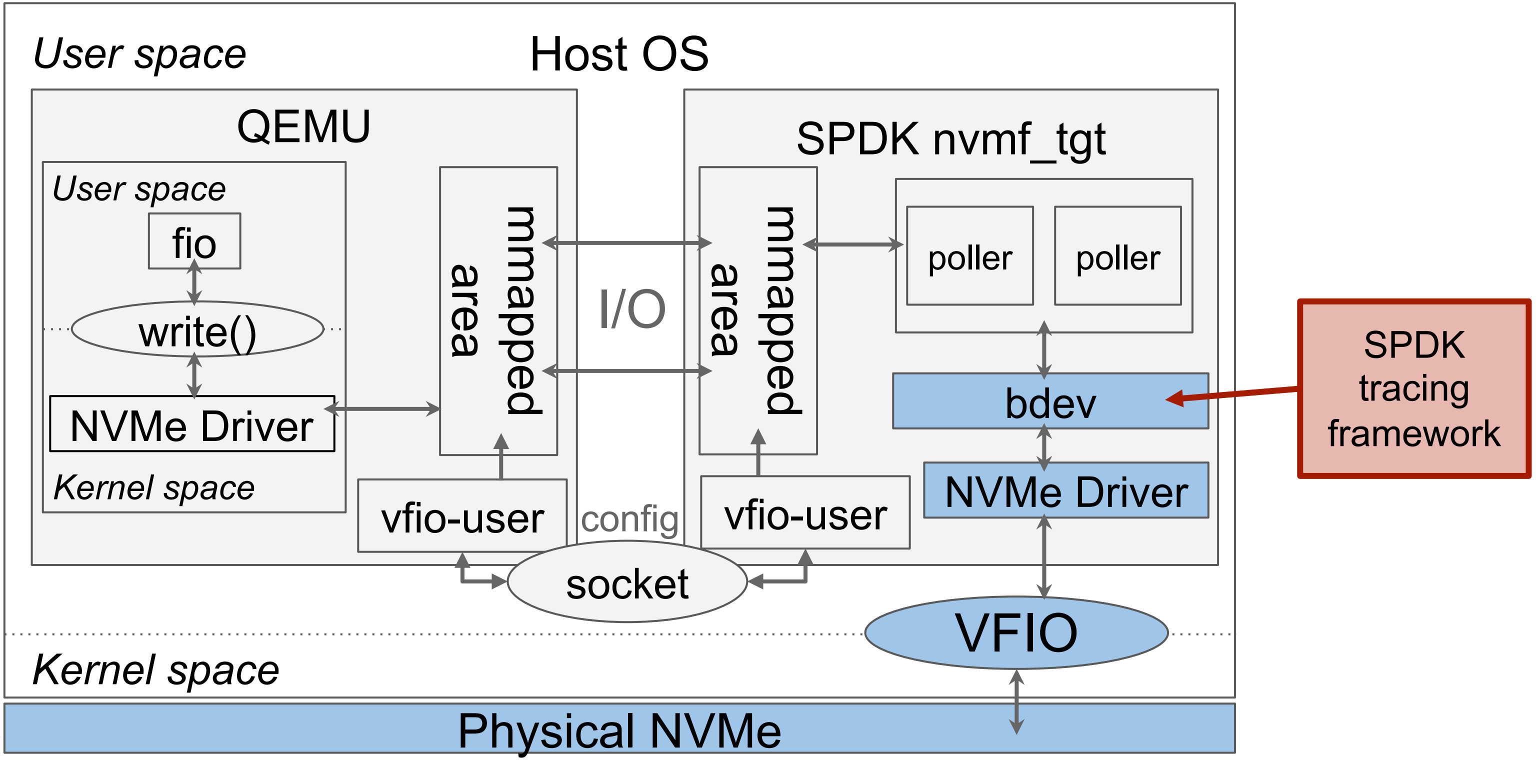
Random writes (100M keys)

Layer-by-layer latency measurement

Replaced internal SPDK with synchronous I/O and guest NVMe driver

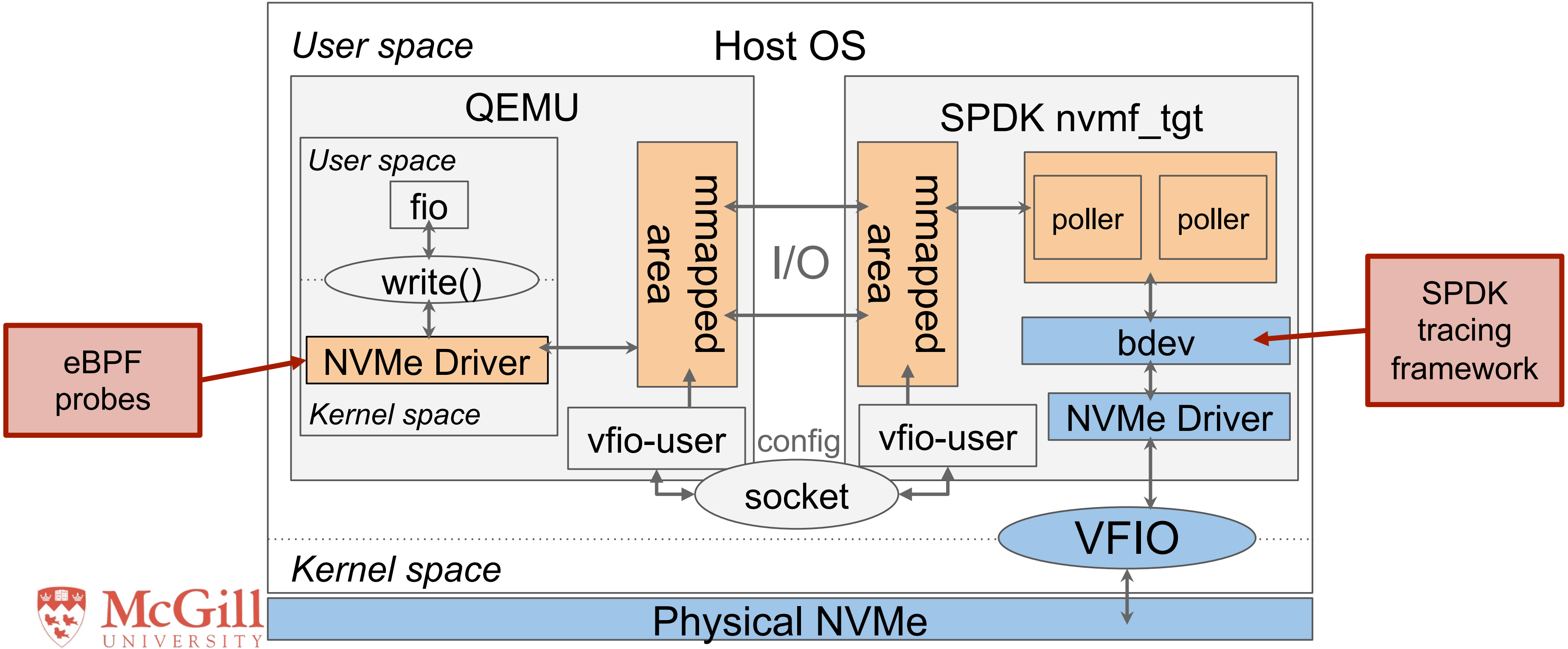


Layer-by-layer latency measurement



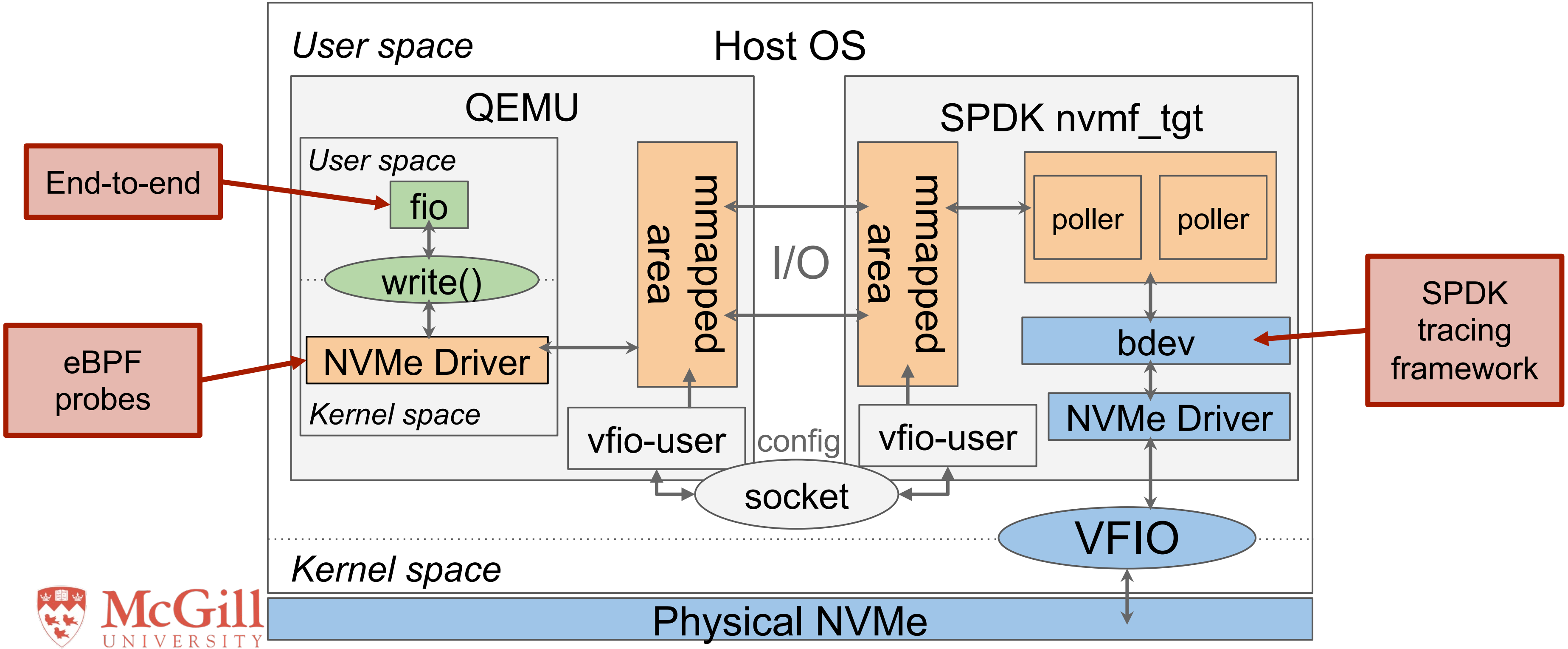
McGill UNIVERSITY

Layer-by-layer latency measurement

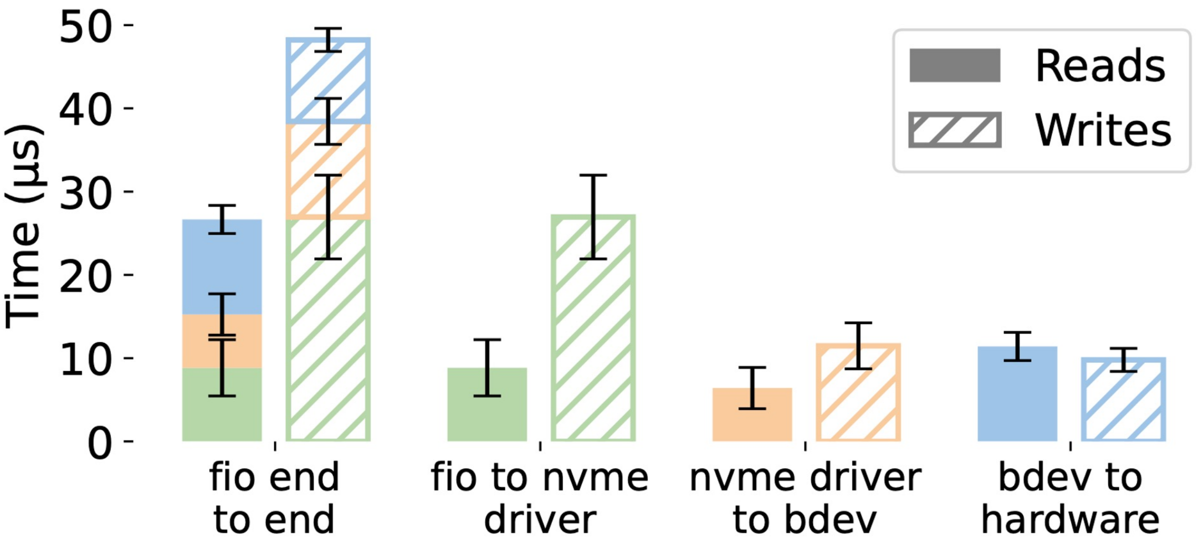


McGill UNIVERSITY

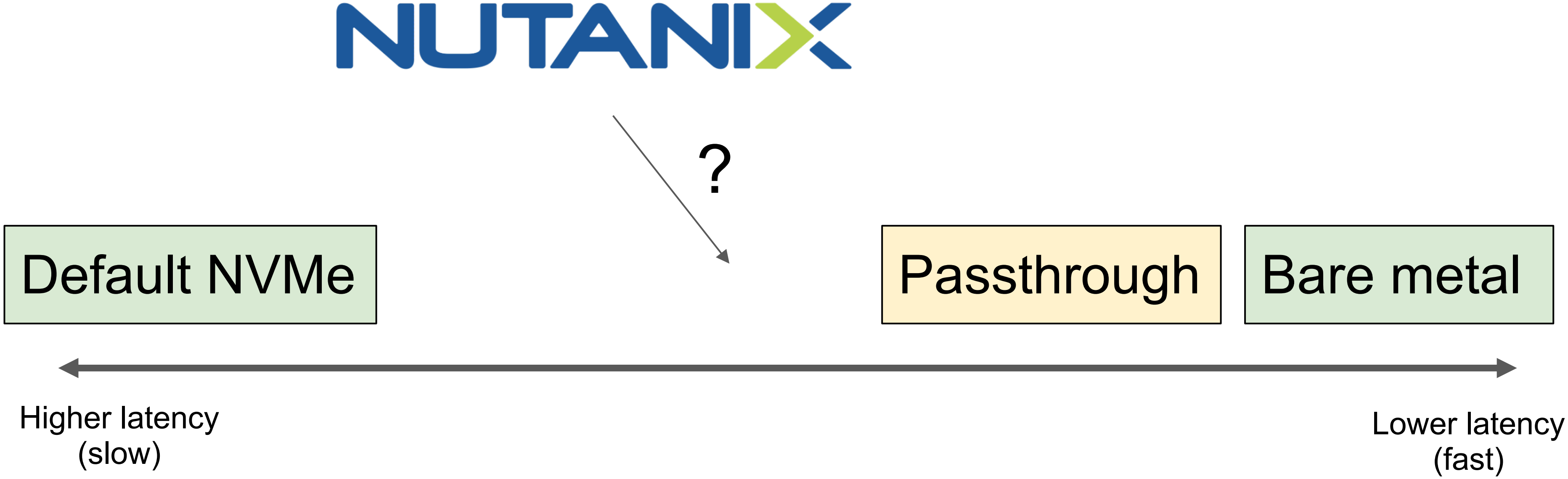
Layer-by-layer latency measurement



Layer-by-layer latency measurements



Where does vfio-user fit again?



NUTANIX

?

Default NVMe

Passthrough

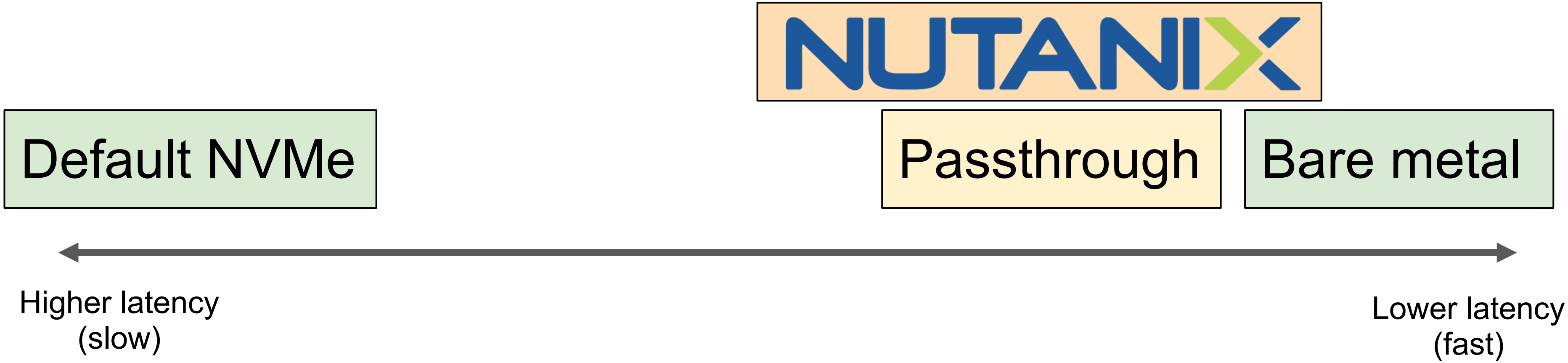
Bare metal

Higher latency
(slow)

Lower latency
(fast)

McGill
UNIVERSITY

Right about next to Passthrough



NUTANIX

Default NVMe

Passthrough

Bare metal

Higher latency
(slow)

Lower latency
(fast)

Conclusion

- **Vfio-user** appears to have comparable performance to passthrough
- Could be viable for VM storage
- See more benchmarks and analysis in our paper!

Visit the DISCS Lab:
https://discslab.cs.mcgill.ca

On the job market!
js@rolon.co

McGill
UNIVERSITY