The Challenges of Building AI Infra on Virtualization

Xin He & Hao Hong

System Technologies and Engineering (STE) team, ByteDance

ByteDance

Agenda

- Background
- Design & Implementation
- Future Work

Background

ByteDance

Background

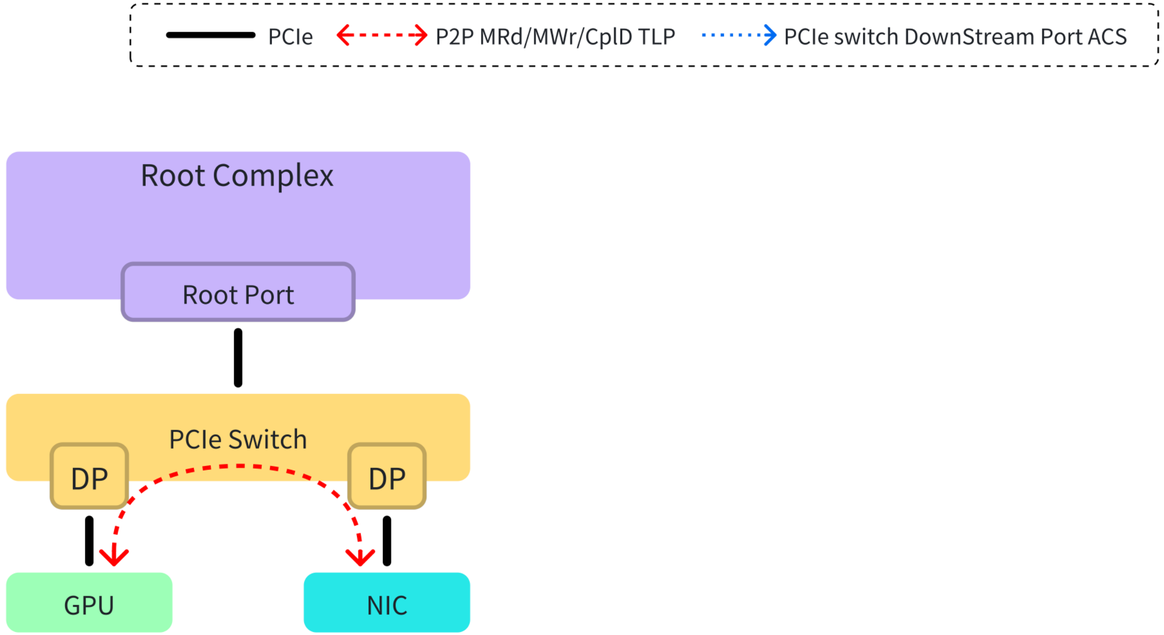
Main issue:

- In virtualization scenarios, there will be a serious performance degradation for PCIe peer-to-peer (p2p) communication due to the enablement of IOMMU
- Various high-precision (millisecond-level) monitoring agents can result in a high number of VMEXITs due to frequent PIO and RDPMC operations

Design & Implementation — Direct P2P

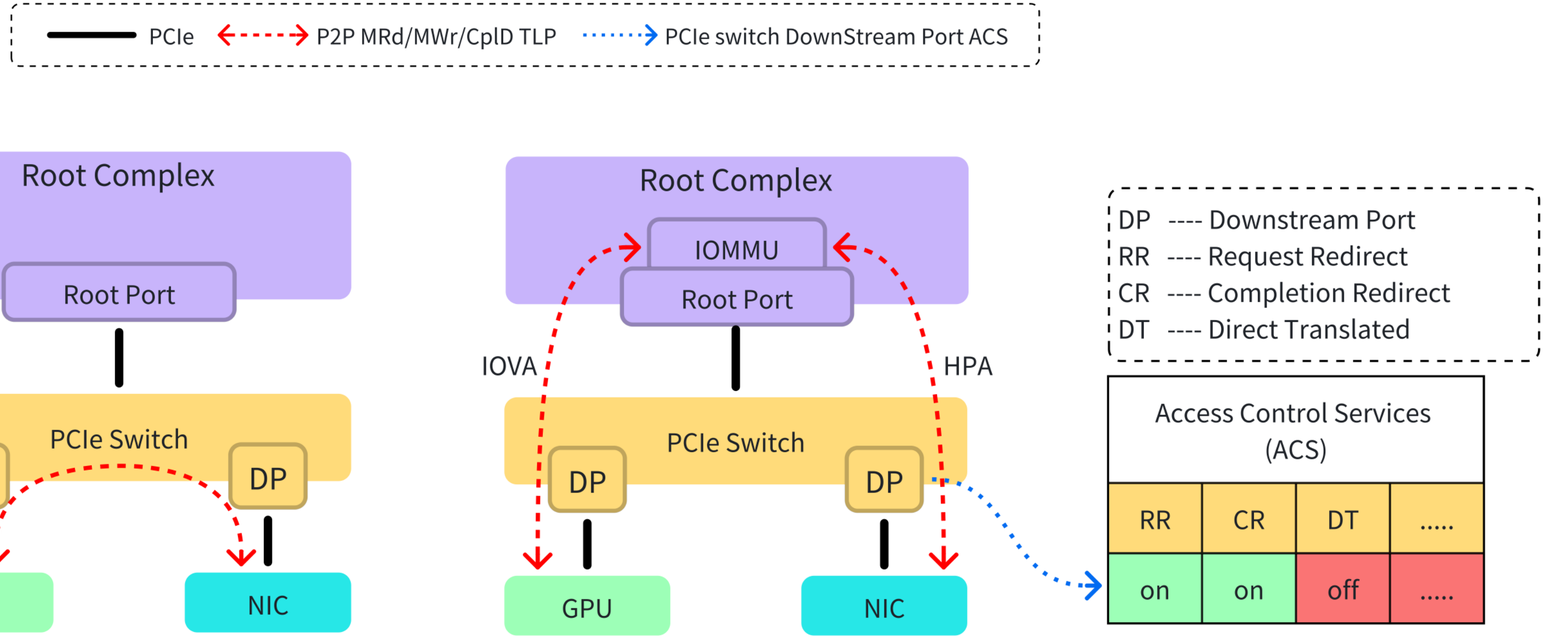
ByteDance

Direct P2P



Host(Disable IOMMU)

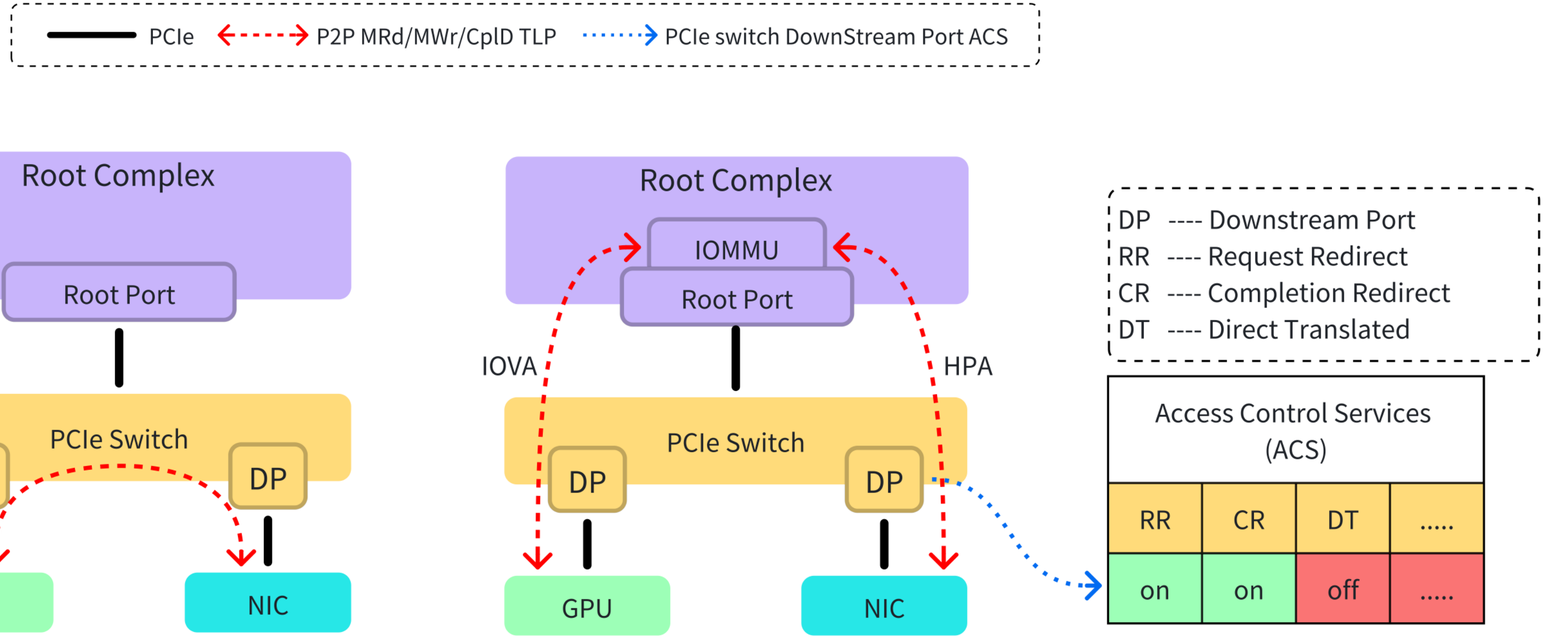
Direct P2P



Host(Disable IOMMU)

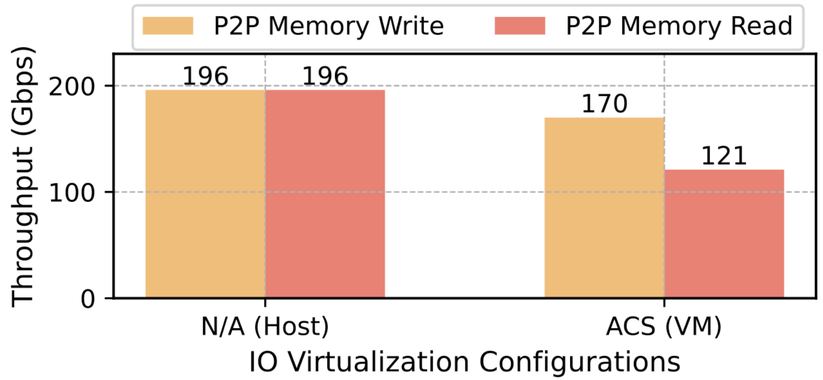
VM(Enable IOMMU & ACS)

Direct P2P



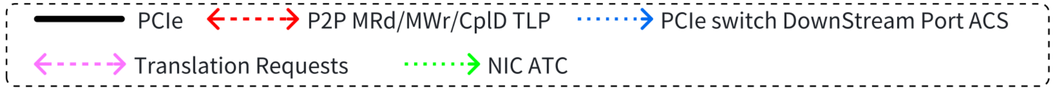
Host(Disable IOMMU)

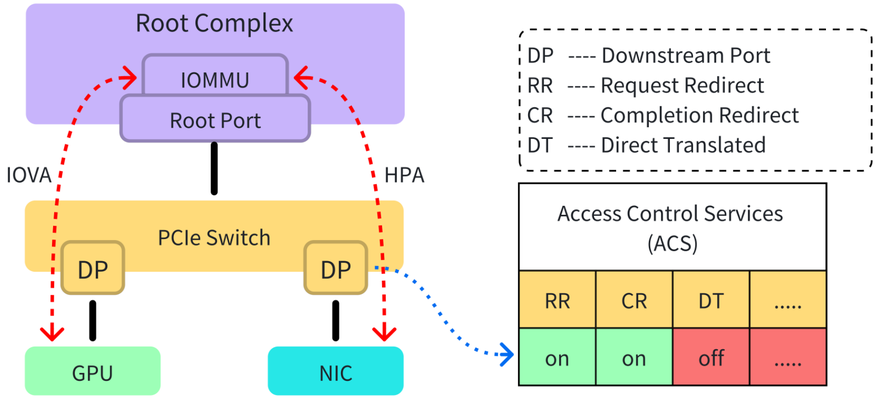
VM(Enable IOMMU & ACS)



CPU: Intel iceLake
GPU: A800
NIC: CX6

Direct P2P



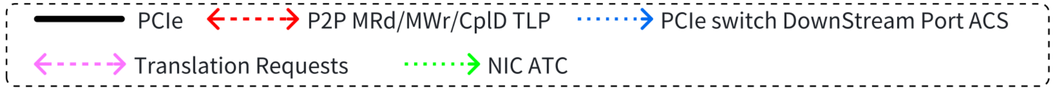


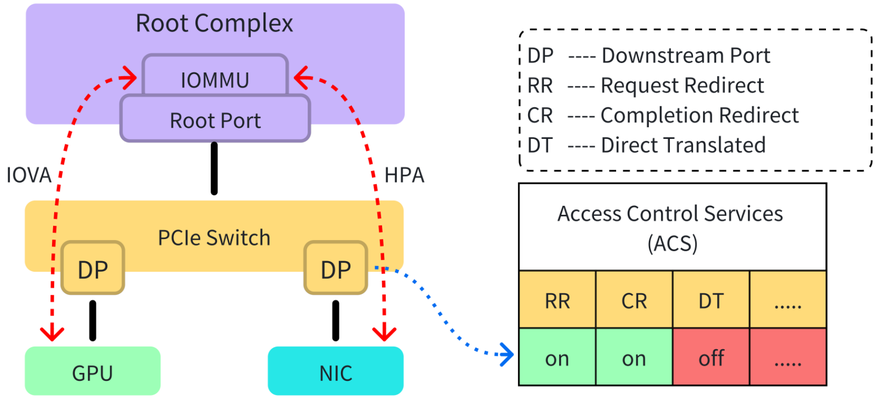
DP ---- Downstream Port
RR ---- Request Redirect
CR ---- Completion Redirect
DT ---- Direct Translated

Access Control Services (ACS)			
RR	CR	DT	.....
on	on	off	.....

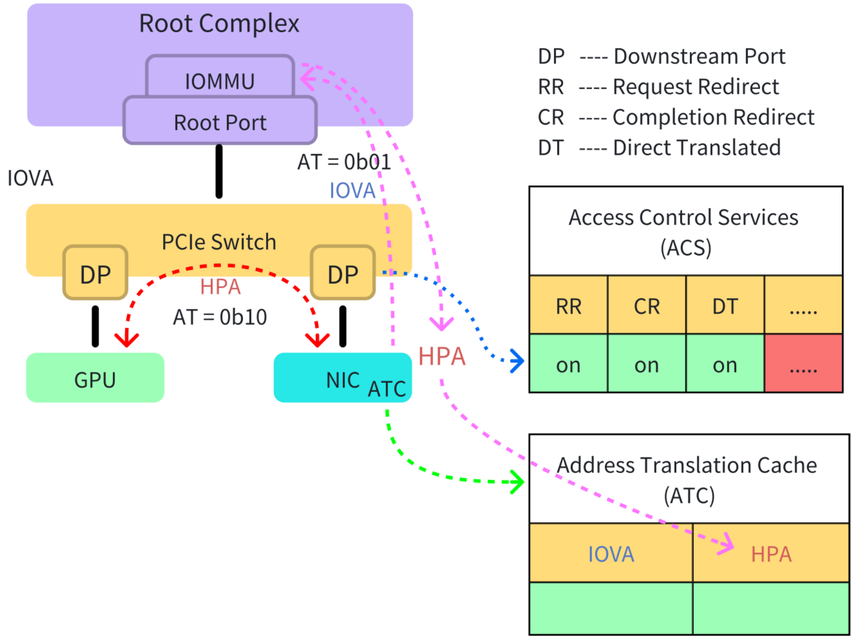
VM(Enable IOMMU & ACS)

Direct P2P



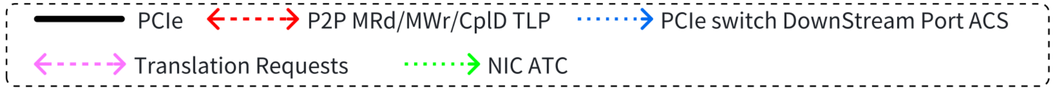


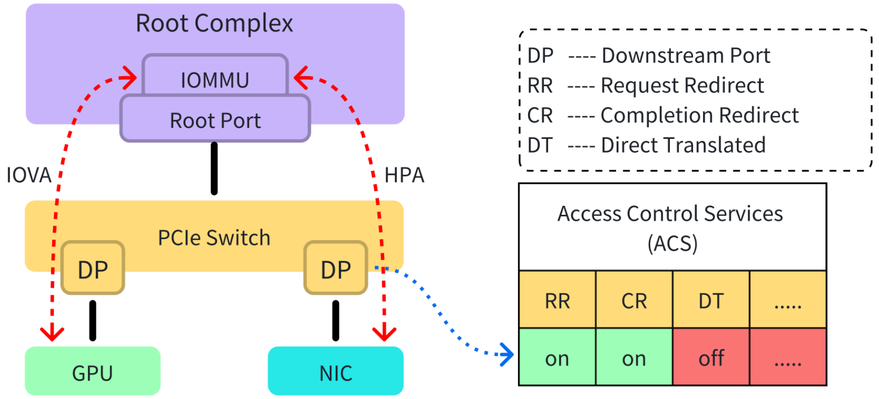
VM(Enable IOMMU & ACS)



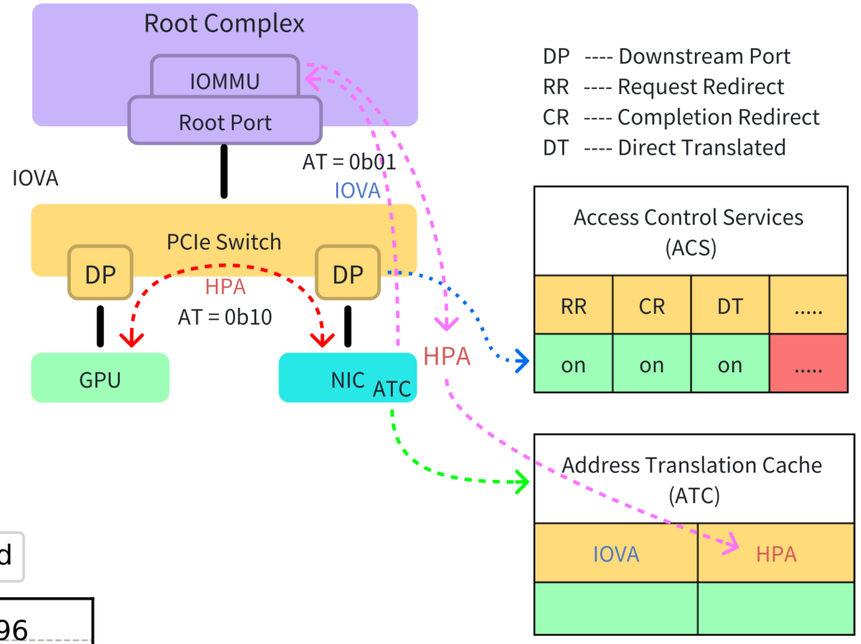
VM(Enable IOMMU & ACS & ATS)

Direct P2P

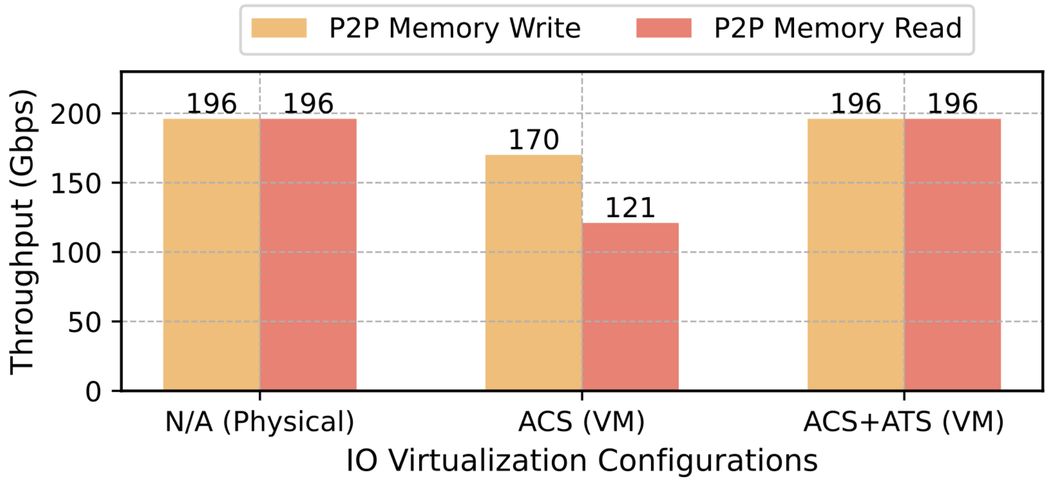




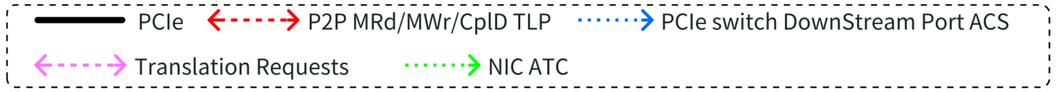
VM(Enable IOMMU & ACS)

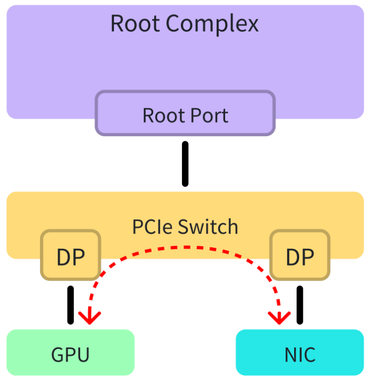


VM(Enable IOMMU & ACS & ATS)

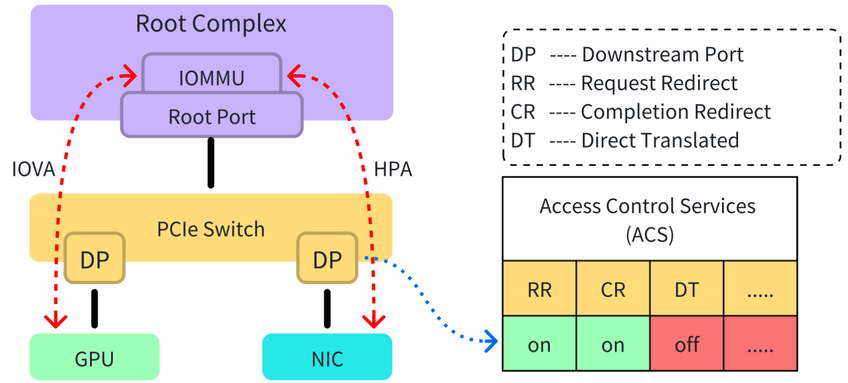


Direct P2P

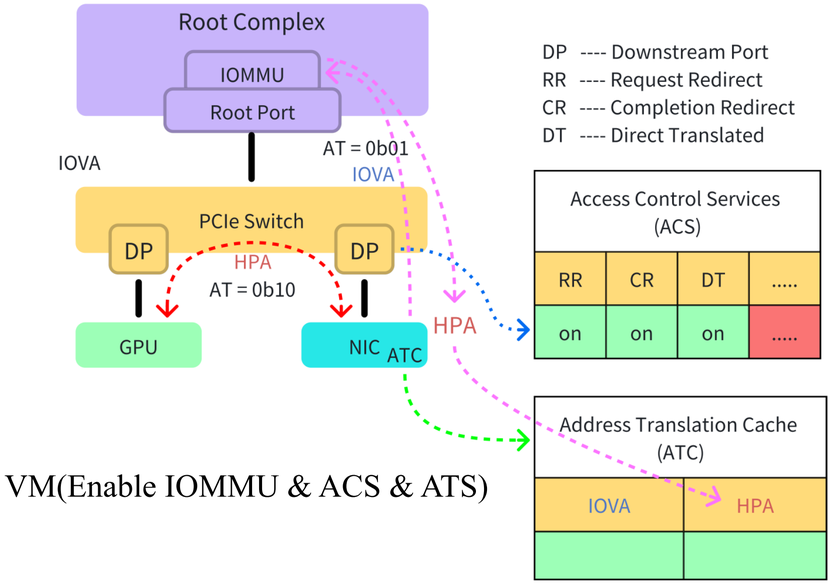




Host(Disable IOMMU)

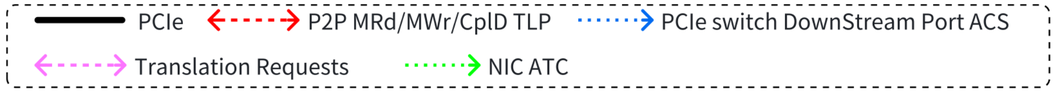


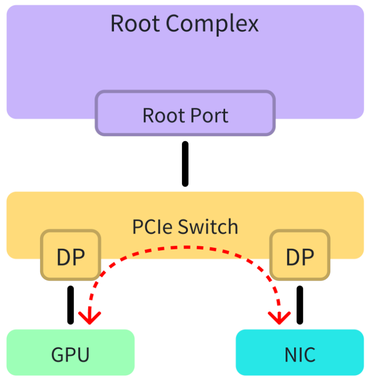
VM(Enable IOMMU & ACS)



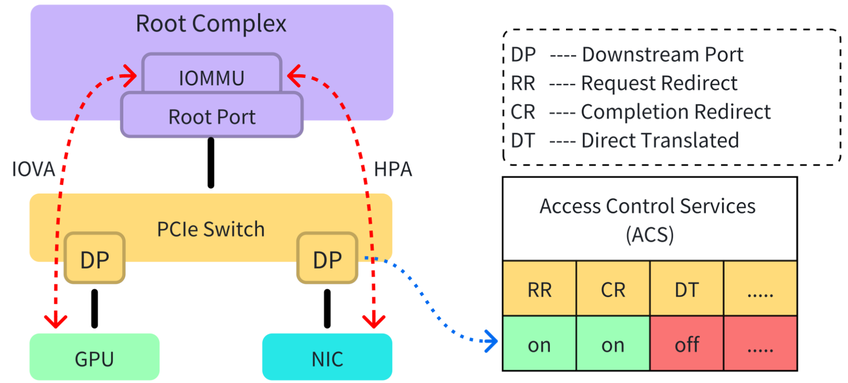
VM(Enable IOMMU & ACS & ATS)

Direct P2P

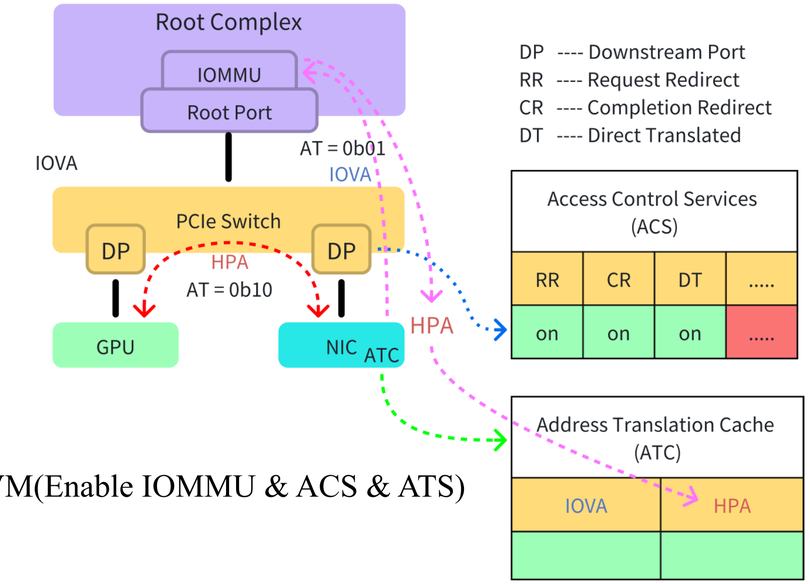




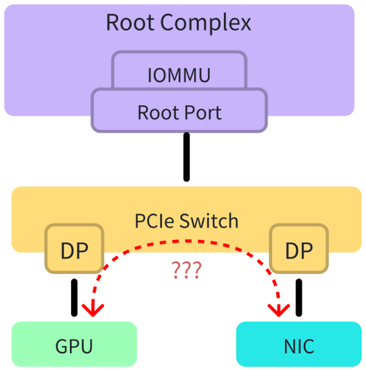
Host(Disable IOMMU)



VM(Enable IOMMU & ACS)

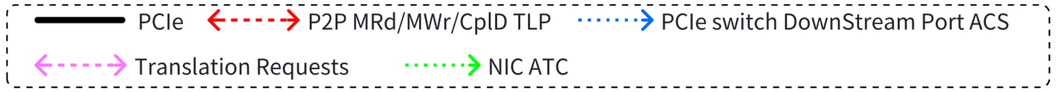


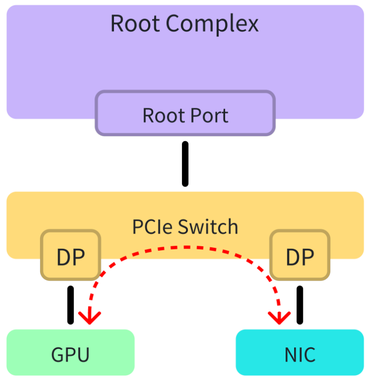
VM(Enable IOMMU & ACS & ATS)



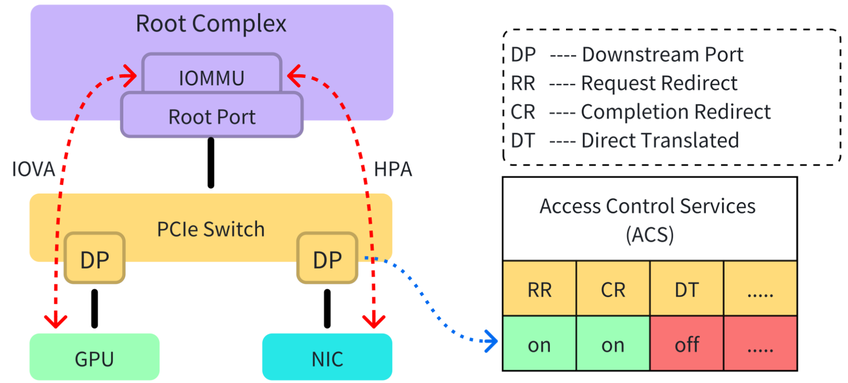
VM(Enable IOMMU & ACS & NO_ATS)

Direct P2P

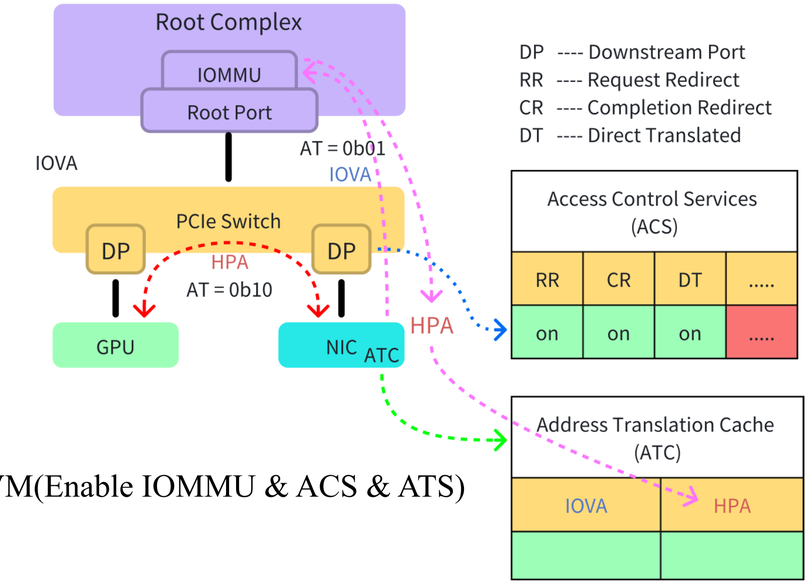




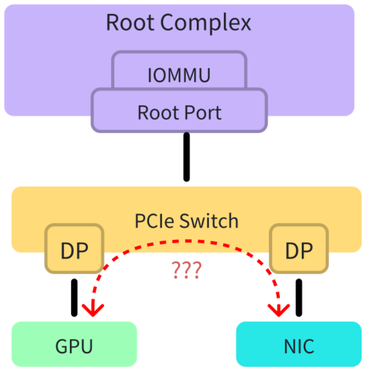
Host(Disable IOMMU)



VM(Enable IOMMU & ACS)



VM(Enable IOMMU & ACS & ATS)

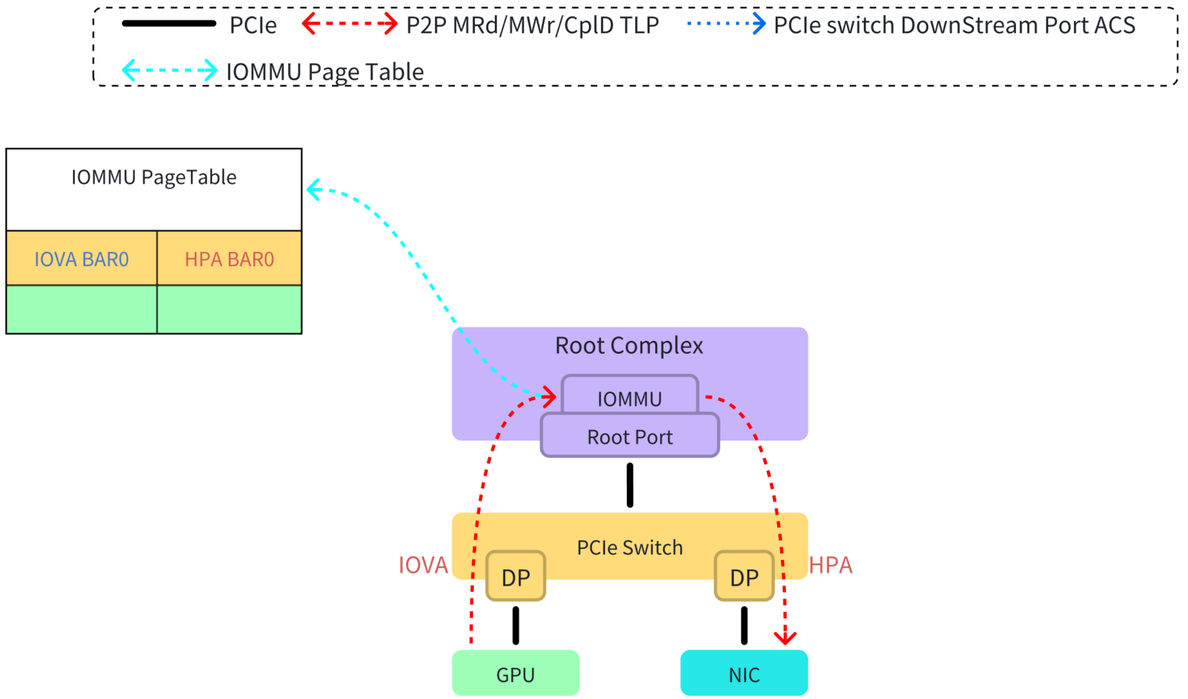


VM(Enable IOMMU & ACS & NO_ATS)

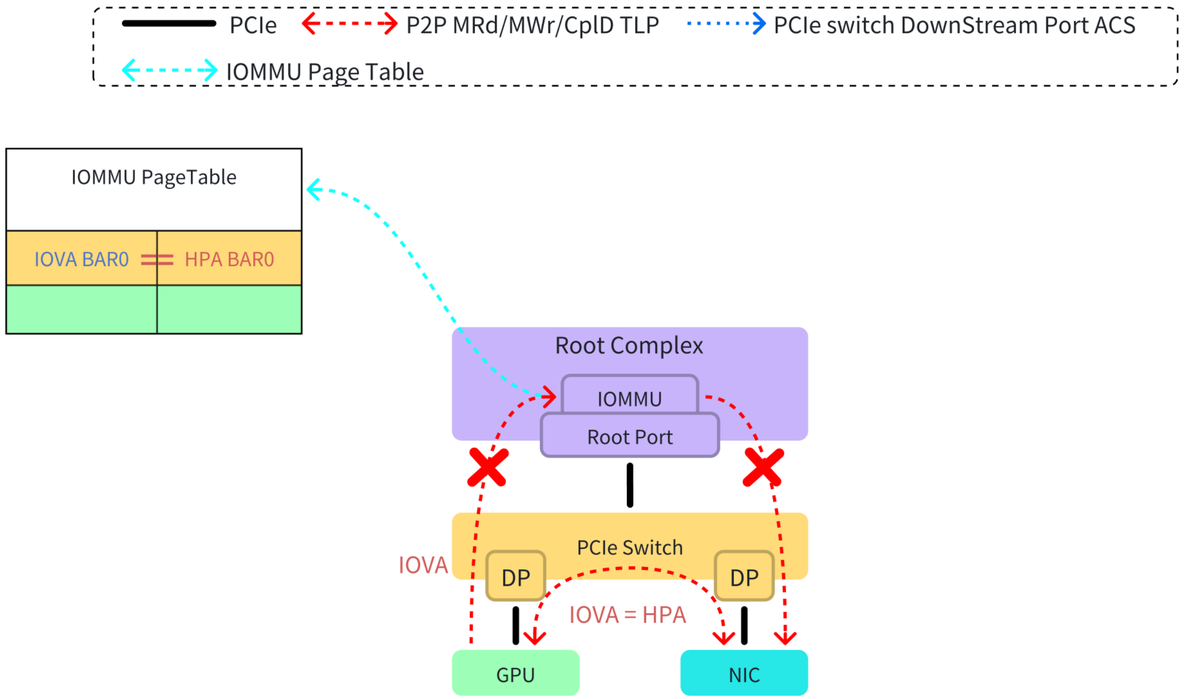
Q1: P2P without IOMMU?

Q2: ACS Configuration

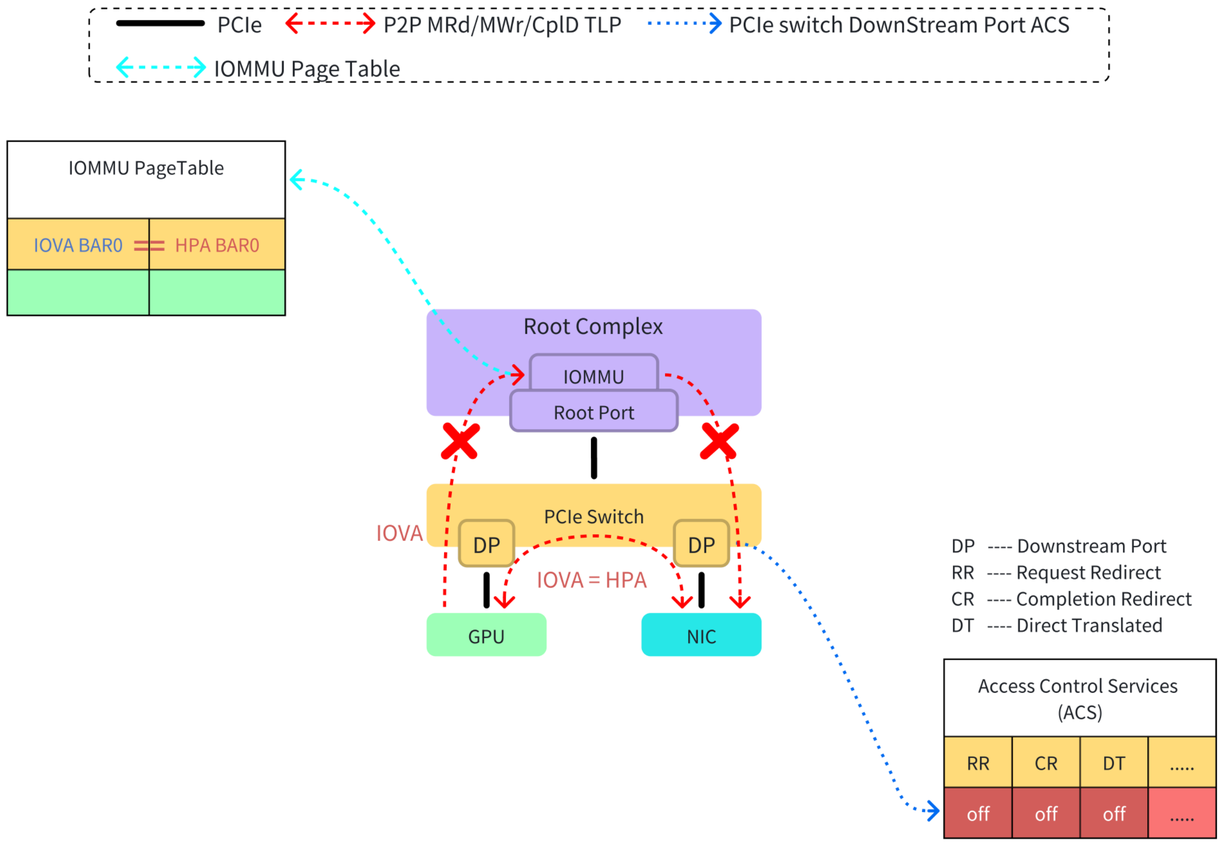
Direct P2P



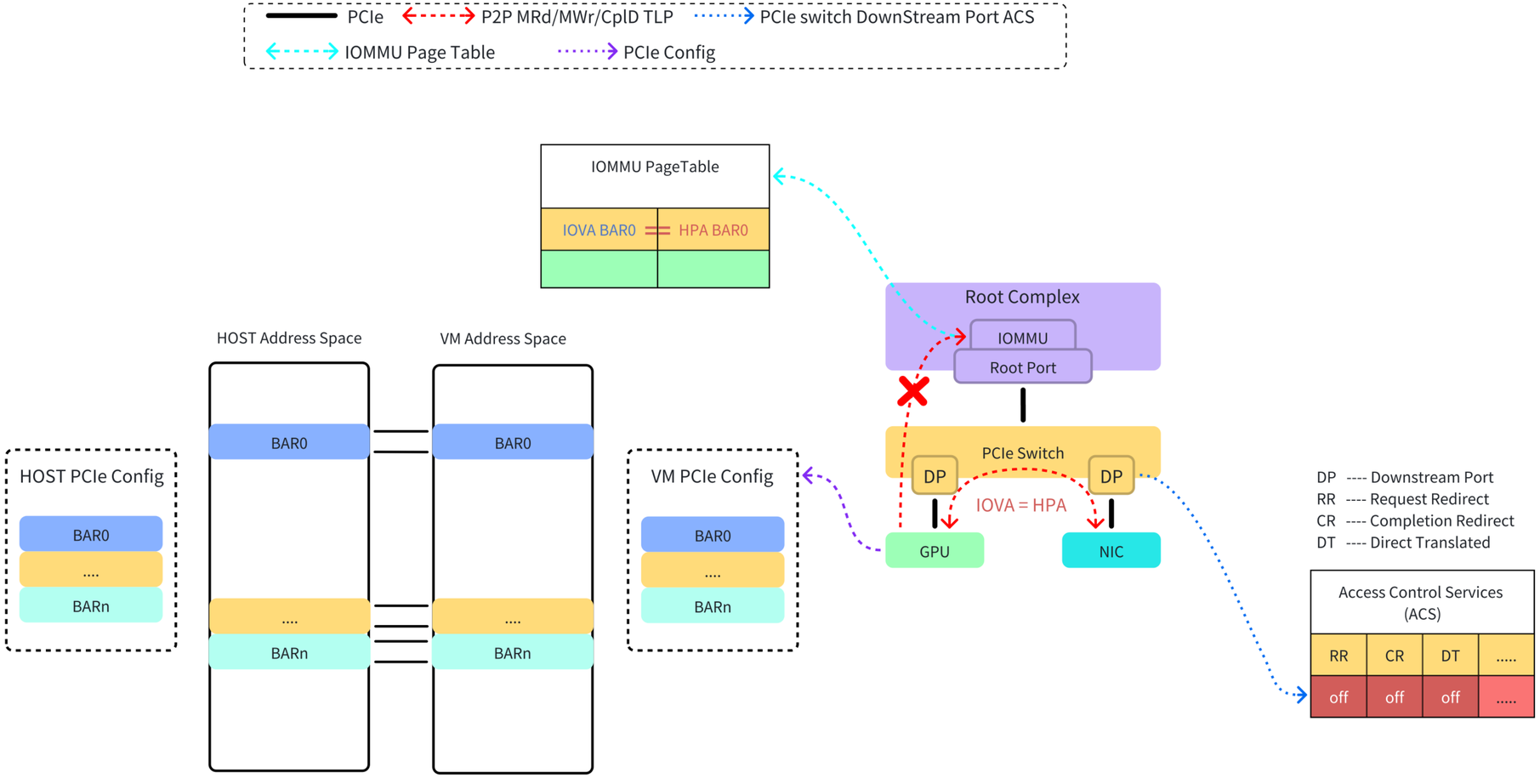
Direct P2P



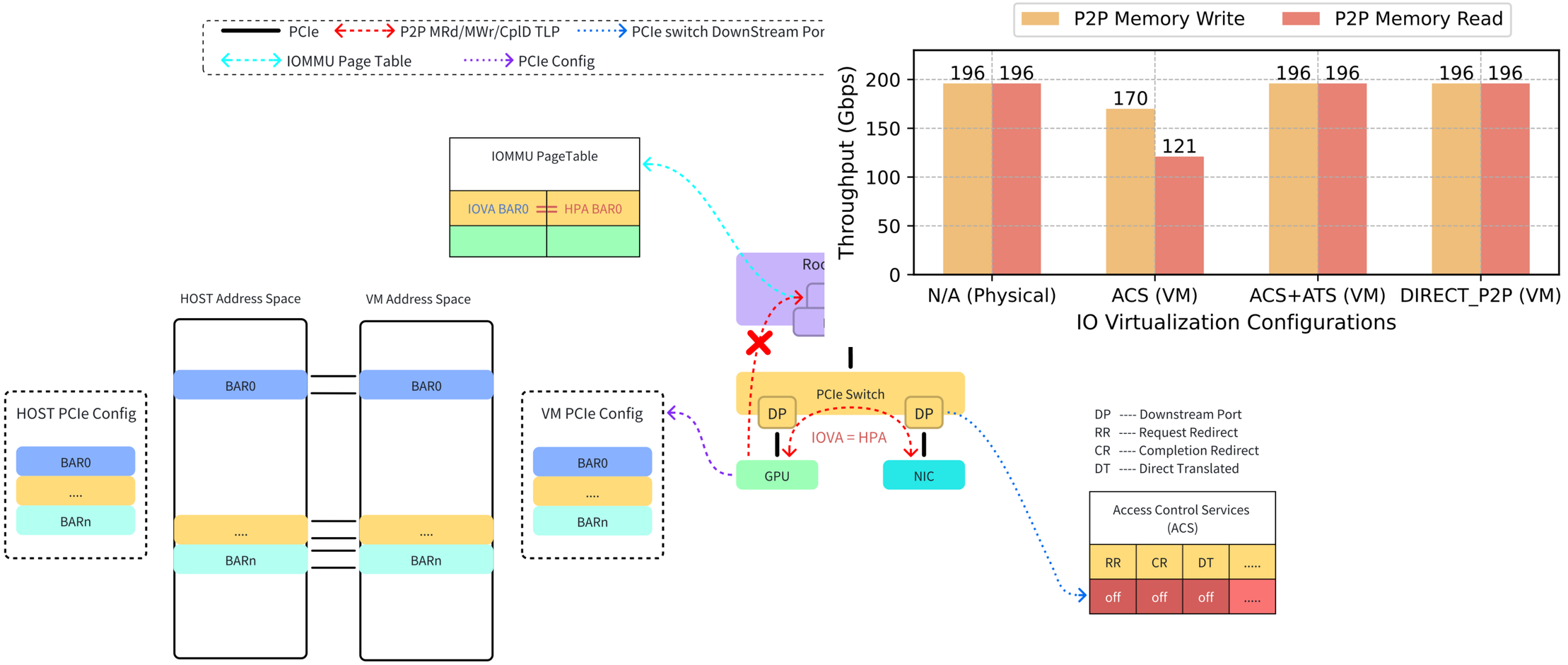
Direct P2P



Direct P2P



Direct P2P

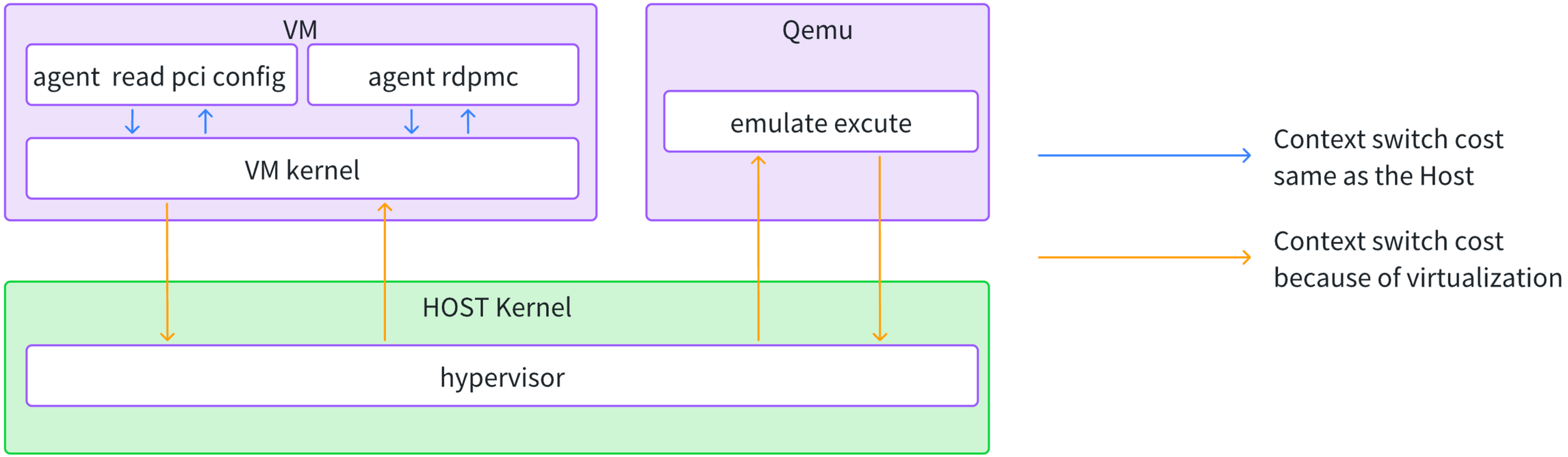


High-precision montoring agents

ByteDance

High-precision monitoring agent

PCI config read/rdpmc Path

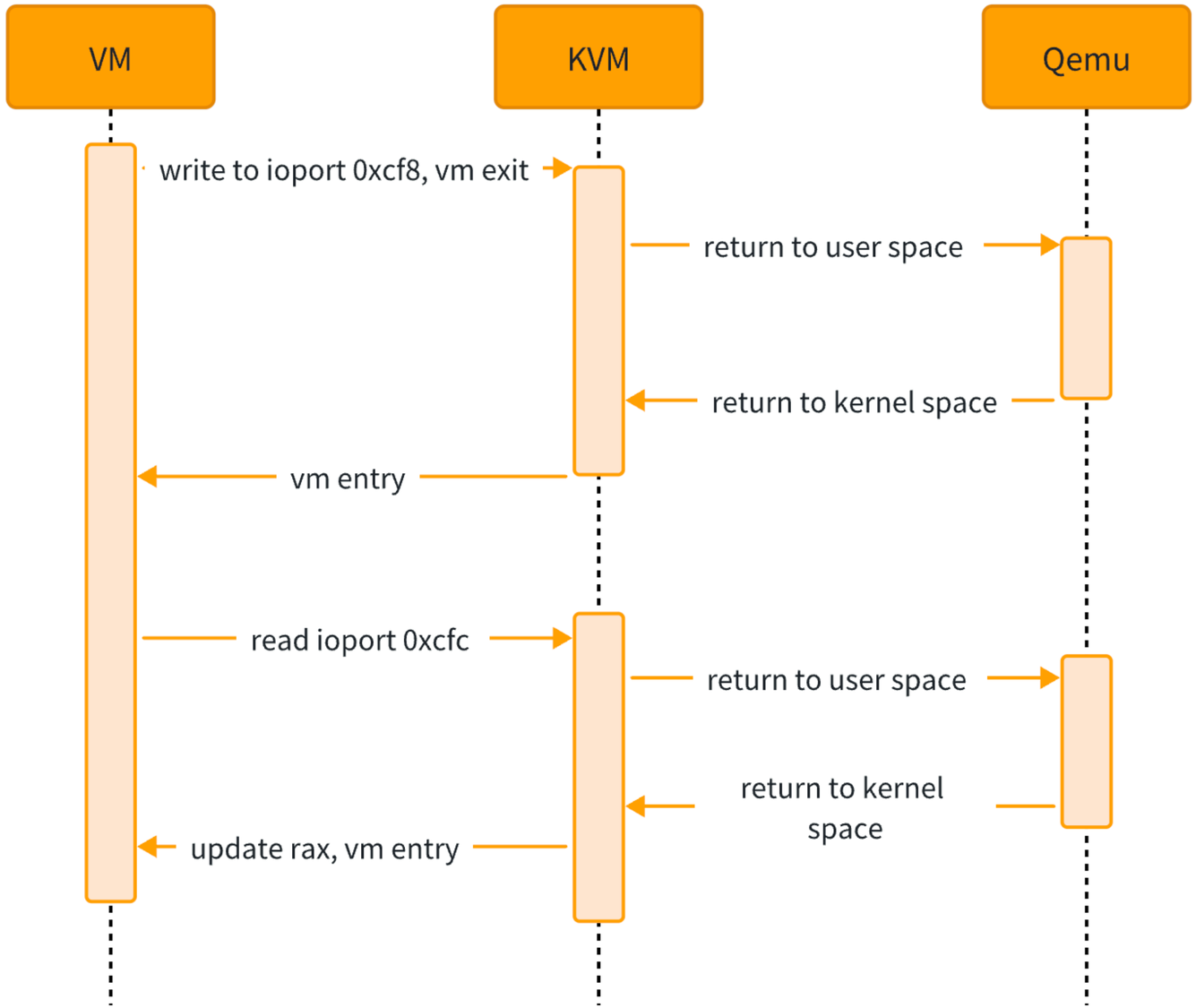


Design & Implementation — PCI PortIO

ByteDance

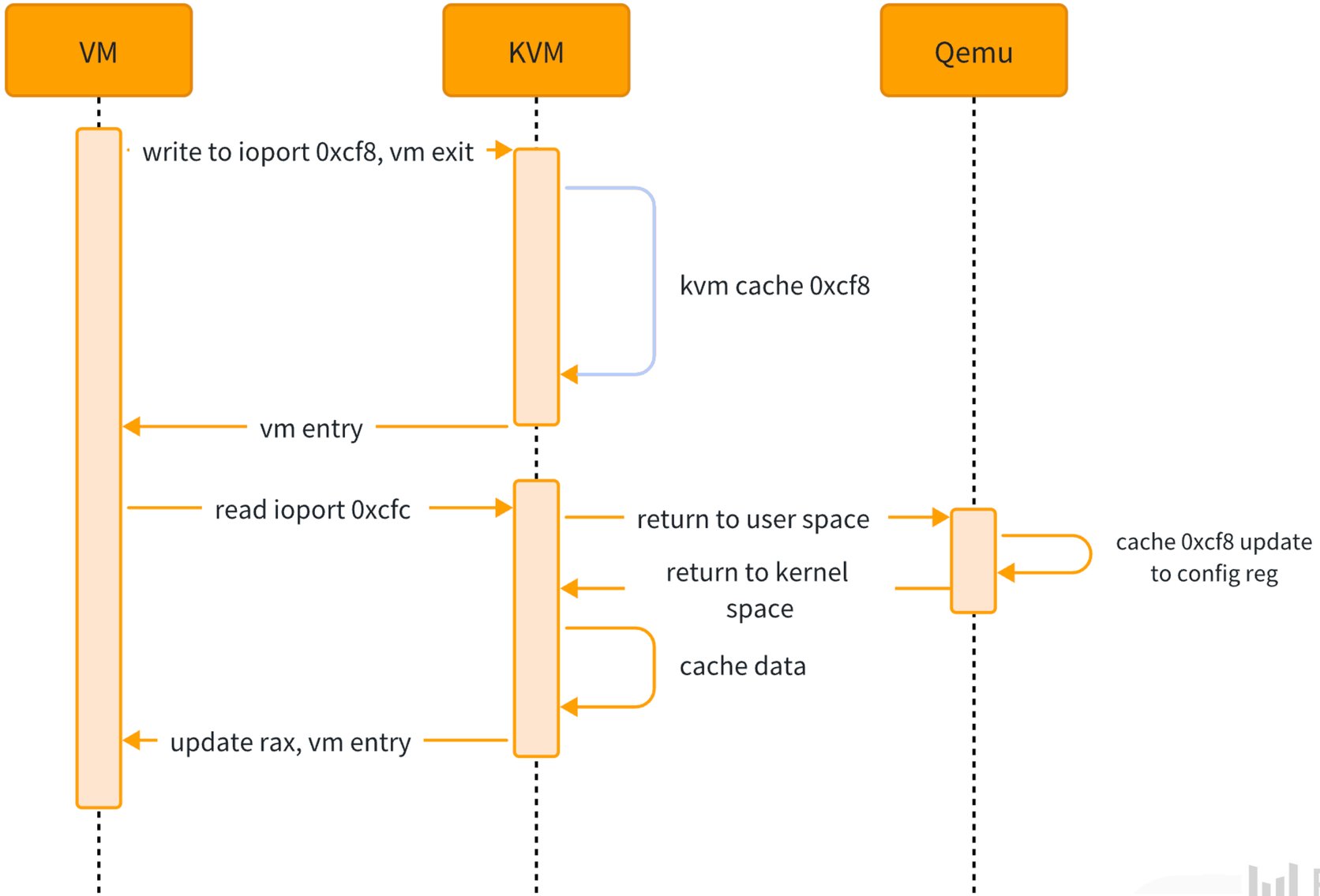
PCI PortIO

Original read path



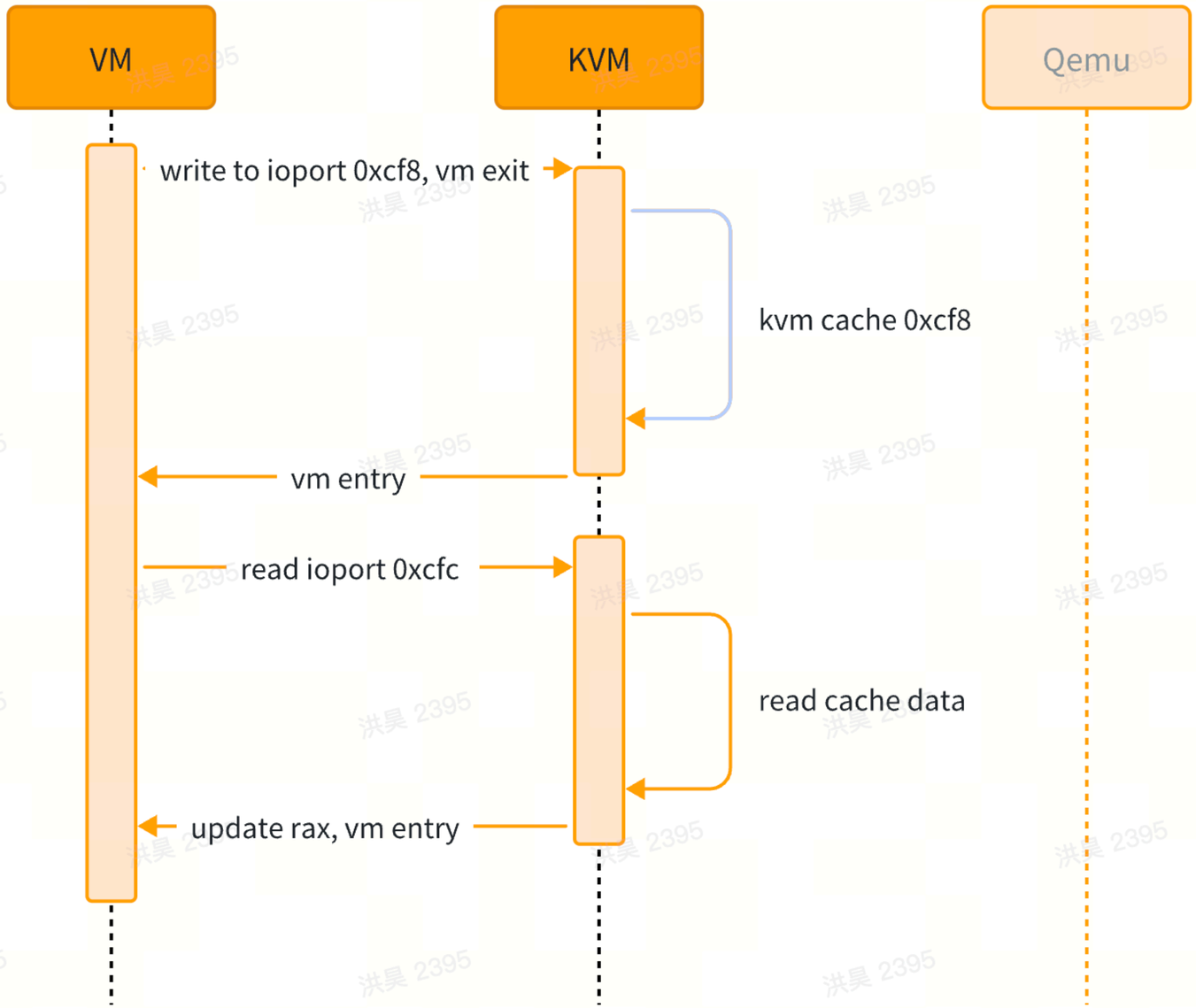
PCI PortIO

First read path

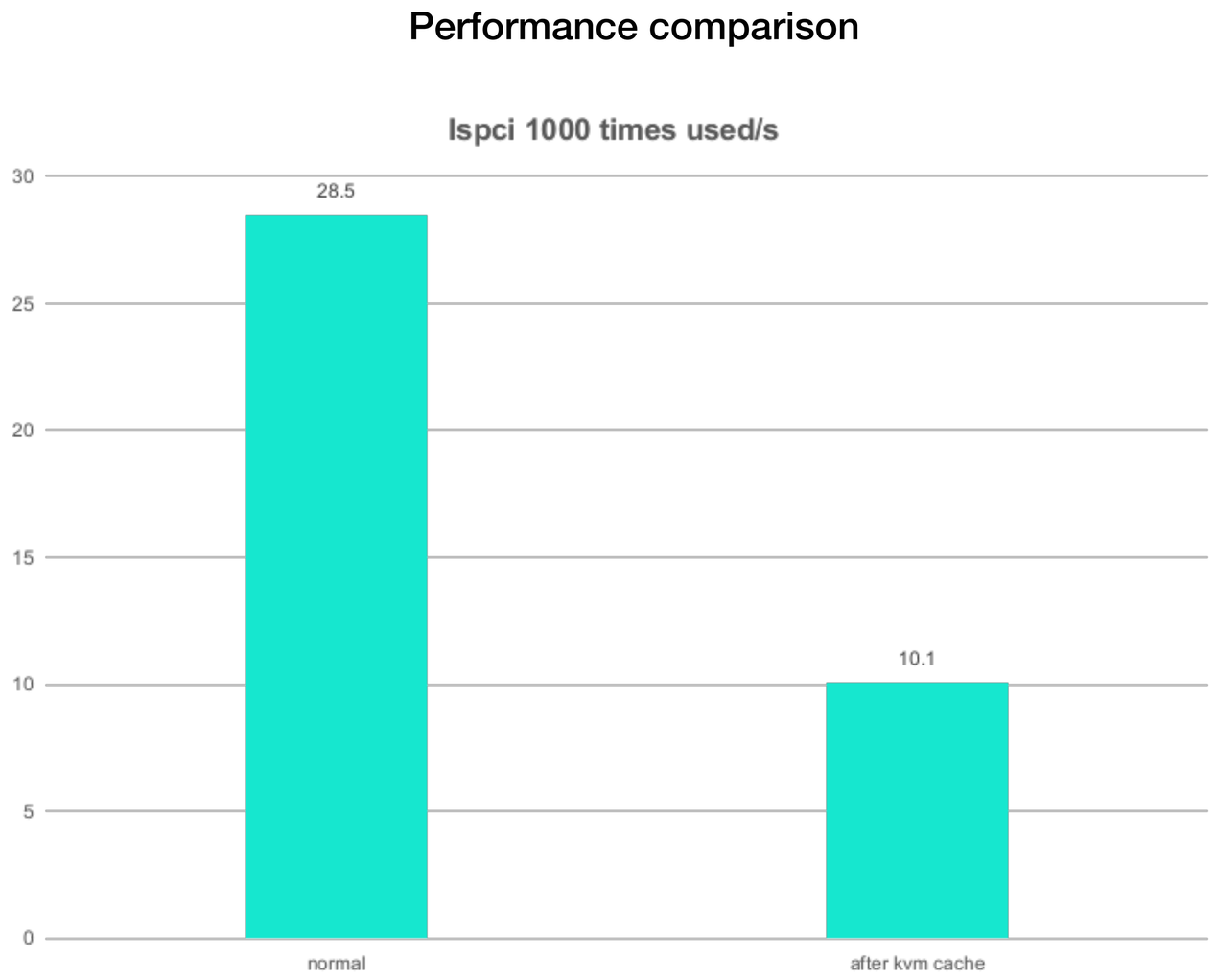


PCI PortIO

Read cache path



PCI PortIO



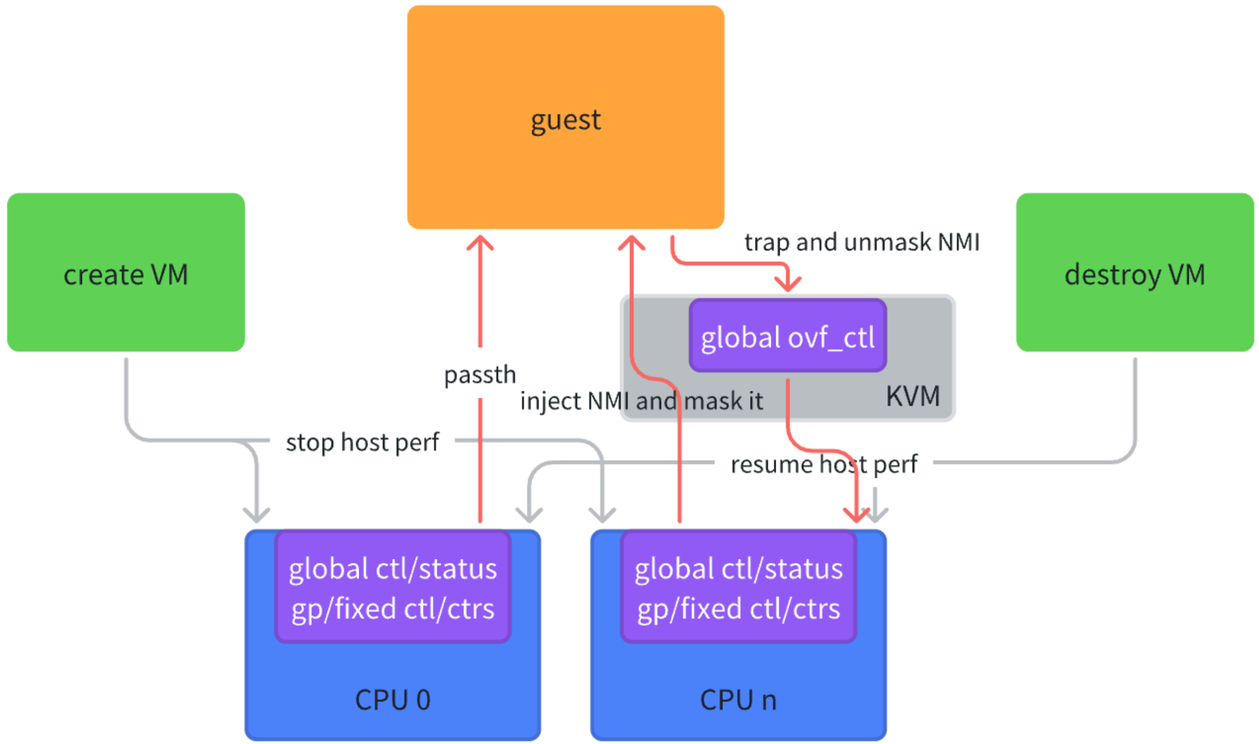
The implementation of `lspci` is consistent with the high-precision monitoring agents. They are both read pci config space.

Design & Implementation — Passthrough PMU

ByteDance

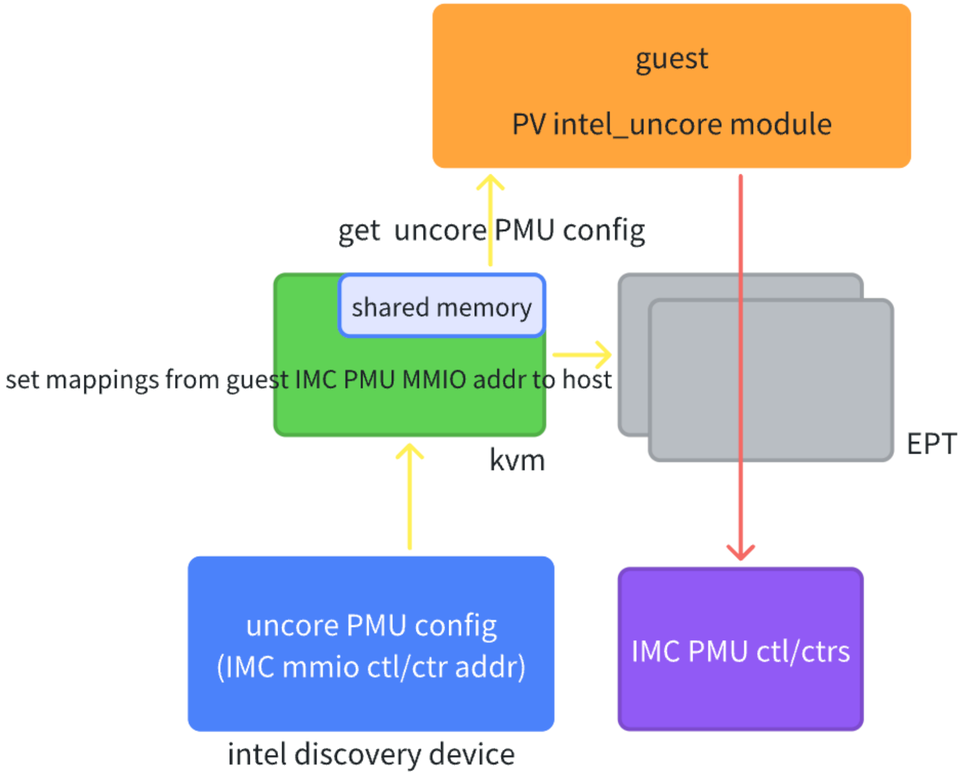
Passthrough Core PMU

eliminate core PMU MSR VM-Exits overhead



Uncore PMU Virtualization

need to deploy high-precision monitoring agents in guest, take IMC PMU for example



Future Work

ByteDance

Future Work

- upstream
- solve H2D & D2H virtualization degradation by DMA devirt

Thank You

ByteDance

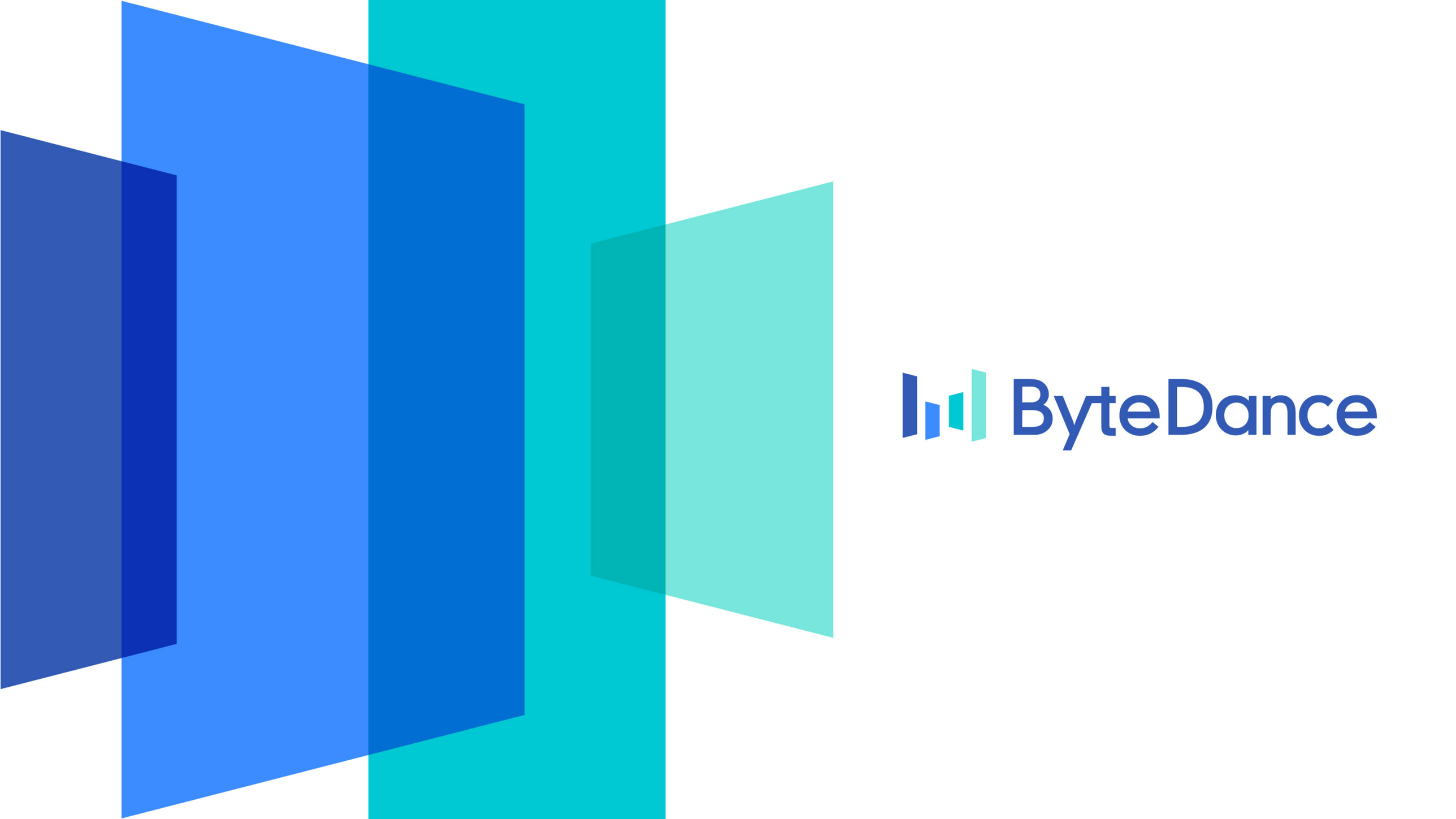
Q & A

Contact Info：hexin.op@bytedance.com
honghao.dante@bytedance.com

ByteDance

Can’t overcommit

- Core PMU: Host cannot use perf when VM running
- Uncore PMU: Can’t used in multiple VM


ByteDance