中国移动 China Mobile

移动云

SPDK Based High Performance Distributed Storage

China Mobile Cloud Center | Xie Changlong

2021 Dec. 15th

中国移动 China Mobile

移动云

01 Background

02 Improving EBS Performance

03 Next Steps

Background

中国移动 China Mobile

5G 移动云

SCI-TECH Development

5G

AR/VR

IOT

Autopilot

Blockchain

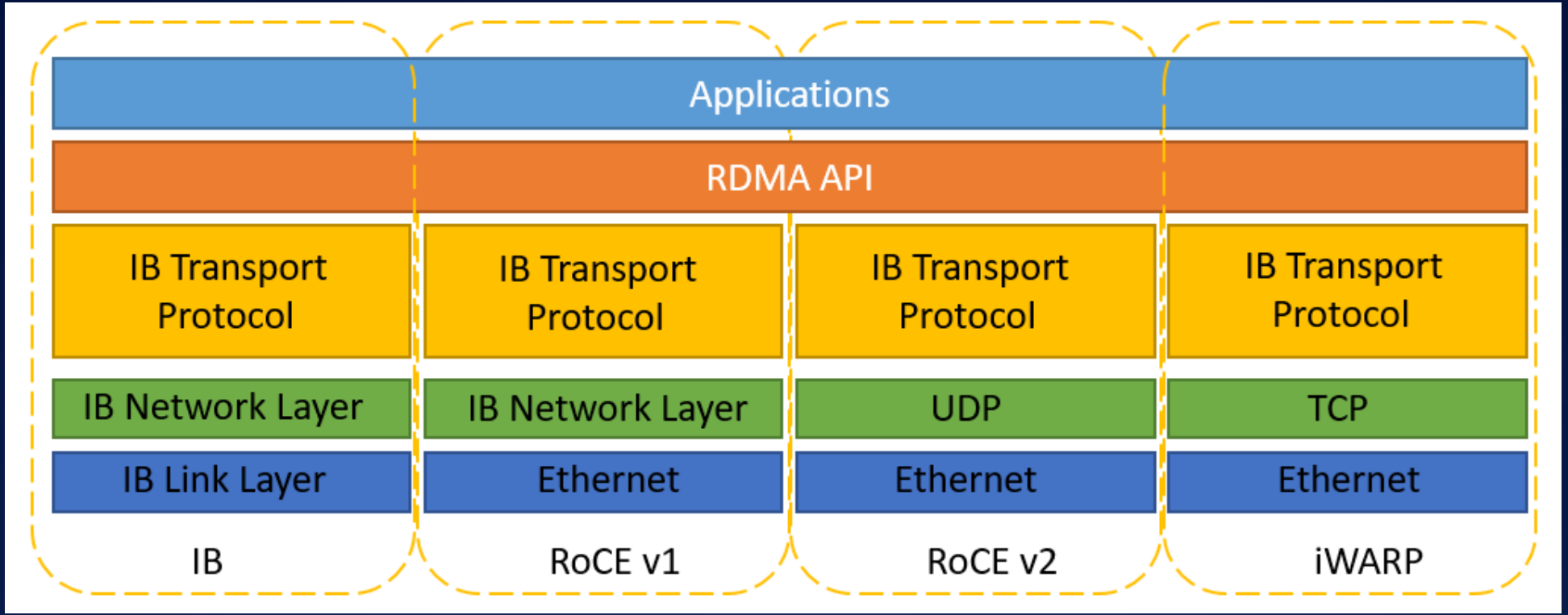
Cloud Computing

Big Data

AI

Capacity **Performance** **TCO**

Network Challenges



Traditional TCP/IP VS RDMA (RemoteDirect Memory Access)

Storage Challenges

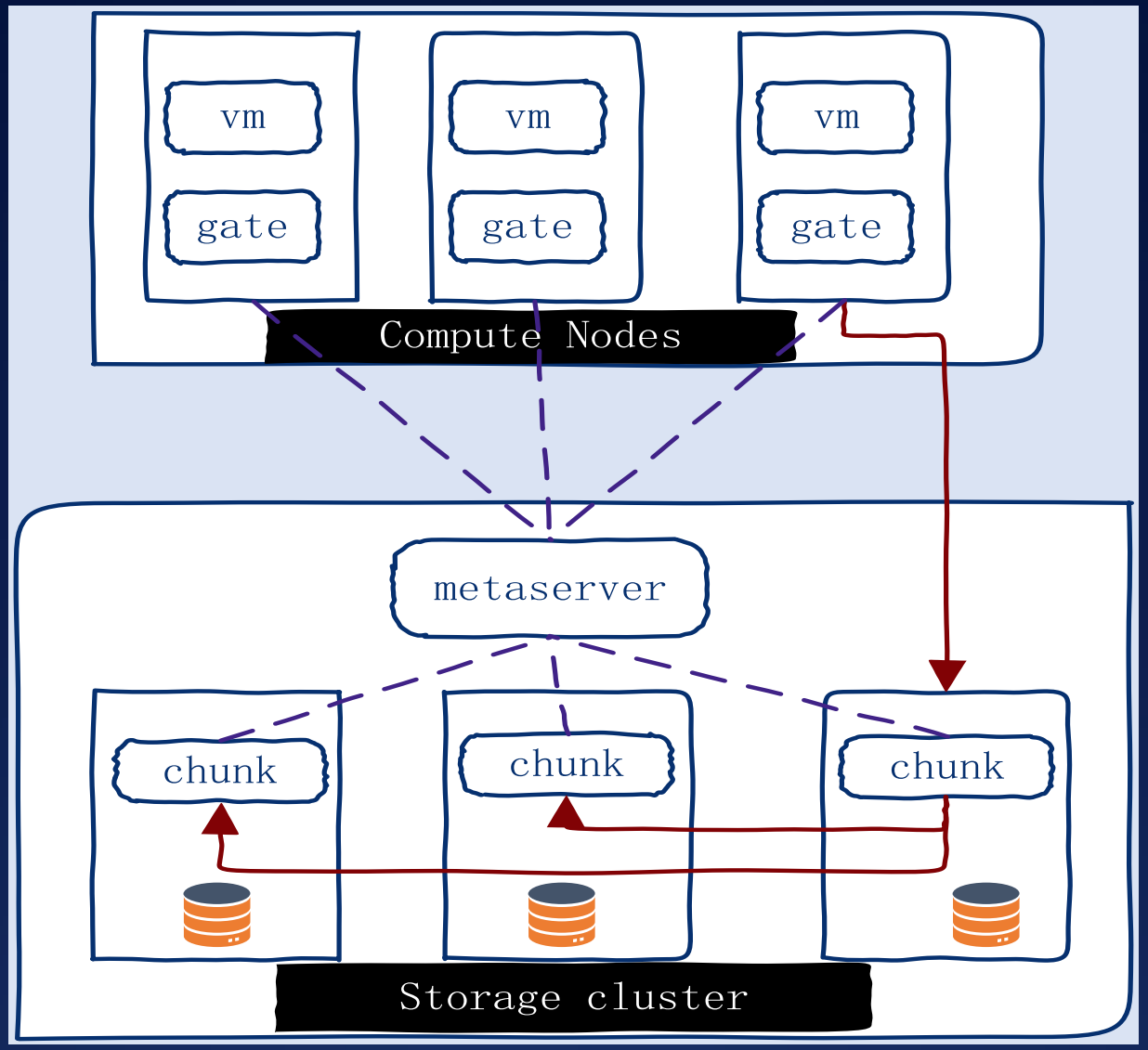
Device	Type	IOPS	Form Factor & Interface
7,200 rpm SATA drive	HDD	~75-100	SATA 3Gbit/s
15,000 rpm SAS drive	HDD	~175-210	SAS
Intel X25-M G2 (MLC)	SSD	~8,600	SATA 3Gbit/s
Intel D3-S4620 (144L TLC 3D NAND)	SSD	~60,000-90,000	SATA 3.0 6Gb/S
Intel D7-P5510 (144L TLC 3D NAND)	SSD	read 930,000 write 190,000	U.2 PCIe 4.0 x4, NVMe
Intel Optane P5800X (PCIe x4, 3D XPoint)	SSD	read 1,500,000 write 1,150,000	PCIe 4.0 x4, NVMe

中国移动
China Mobile

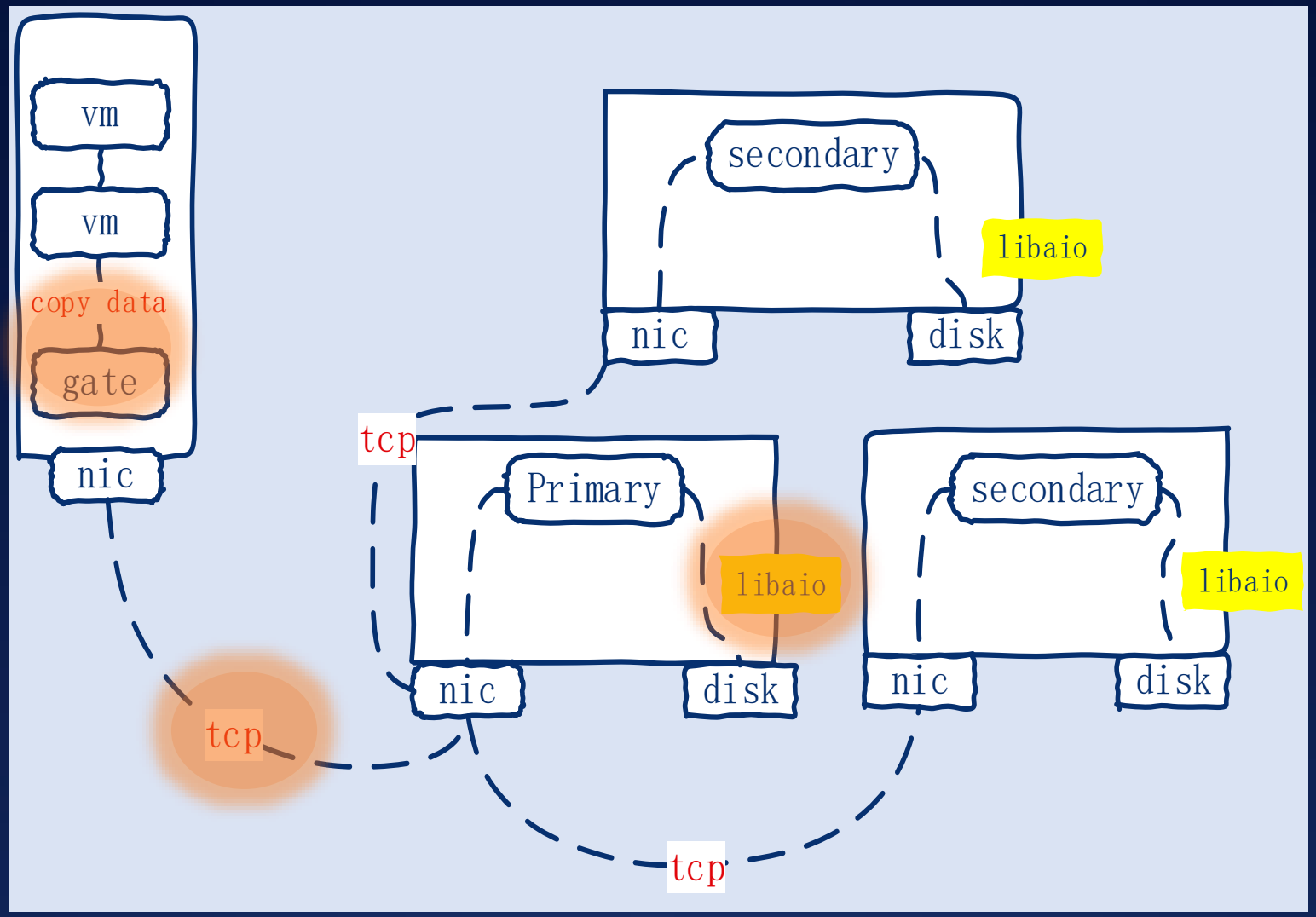
移动云

Improving EBS Performance

EBS Architecture



EBS IO Path


vm
vm
copy data
gate
nic
secondary
libaio
nic
disk
tcp
Primary
libaio
nic
disk
secondary
libaio
nic
disk
tcp
tcp

中国移动 China Mobile

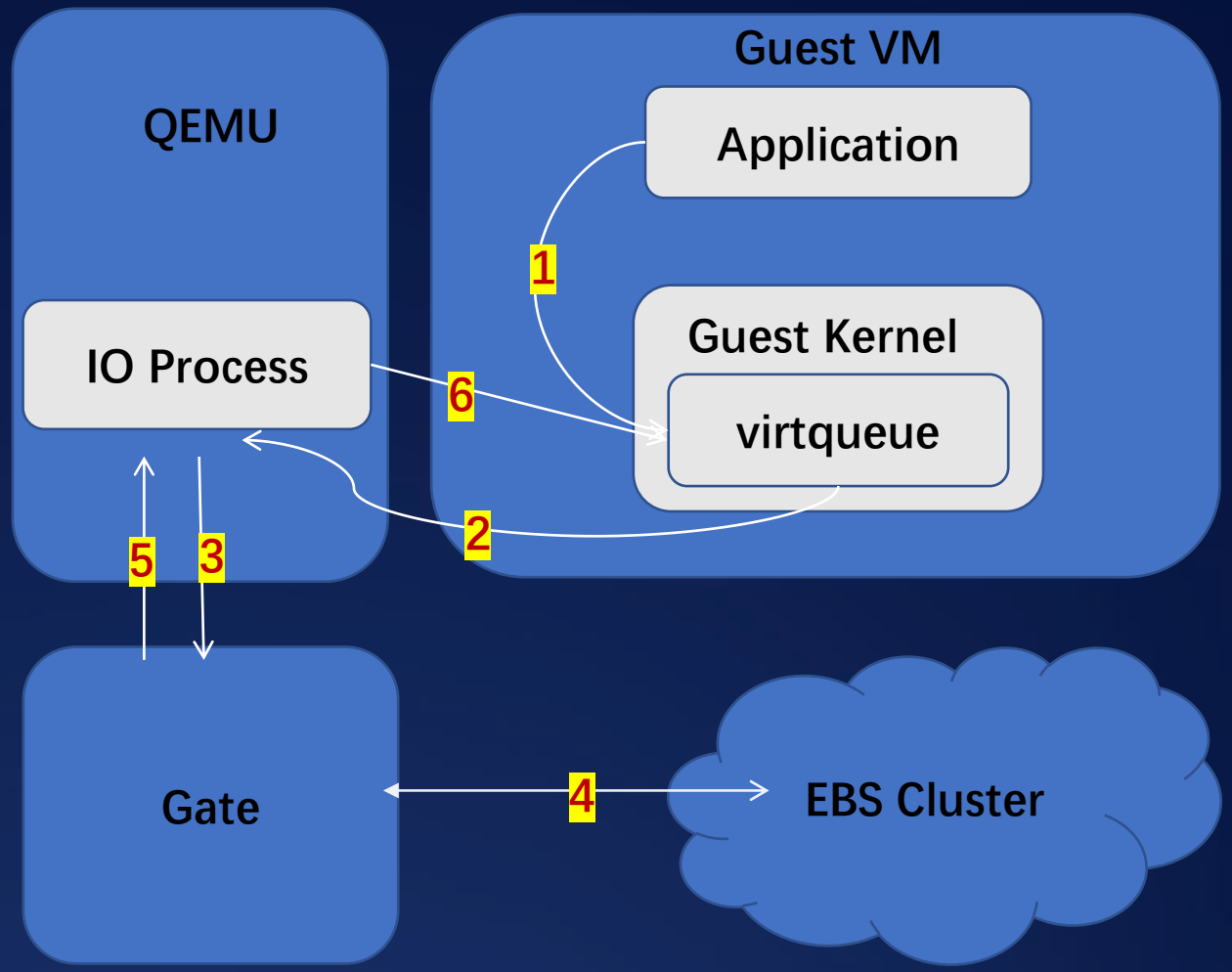
移动云

EBS Evolution

Based on Megrez，introduce new EBS architecture

QEMU Virtio

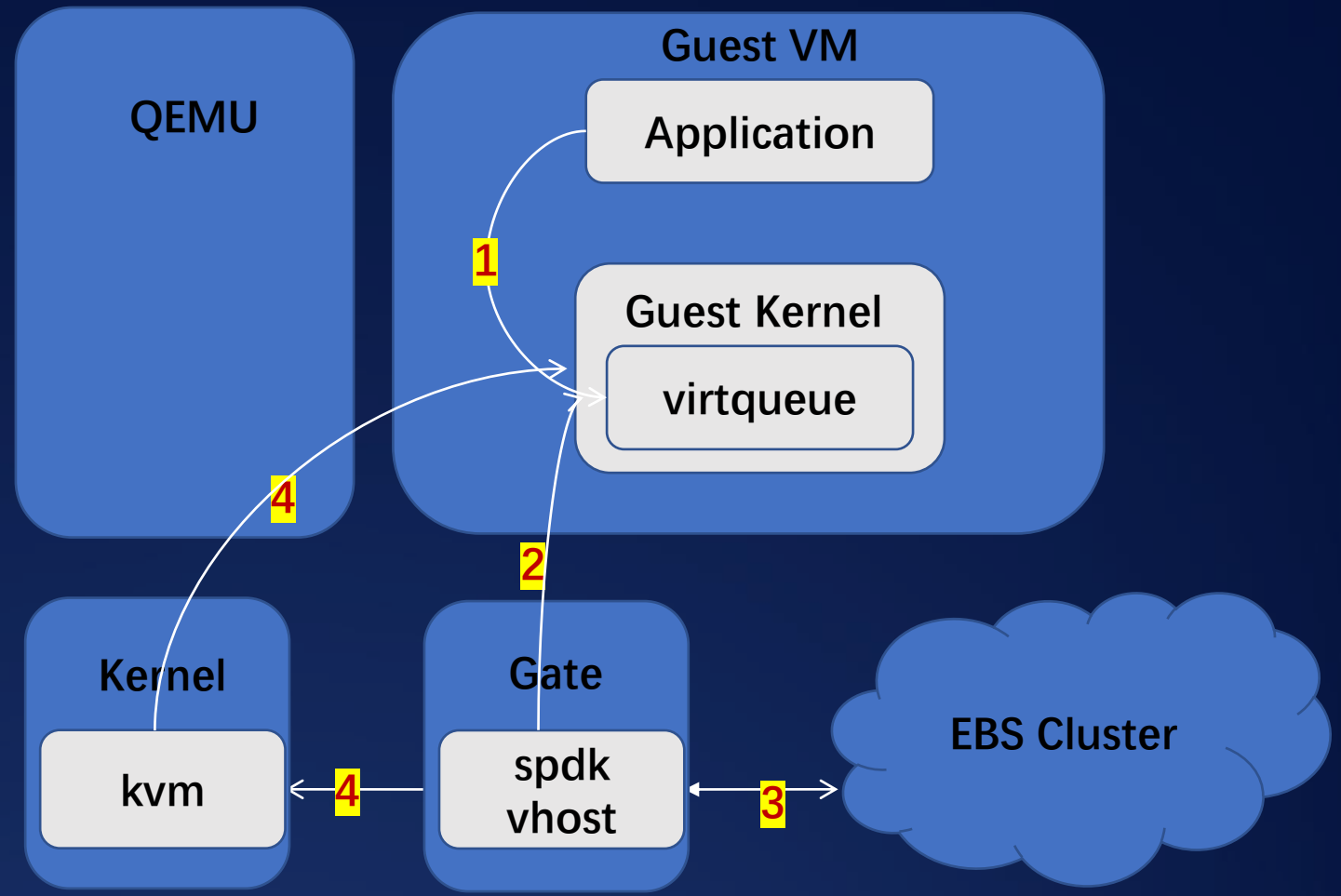
1. Submit io to VQ
2. IO event notify qemu
3. Qemu transfer IO to Gate
4. EBS processing IO
5. Gate forward IO
6. Irqfd notify guest，io complete



Integrate SPDK Vhost

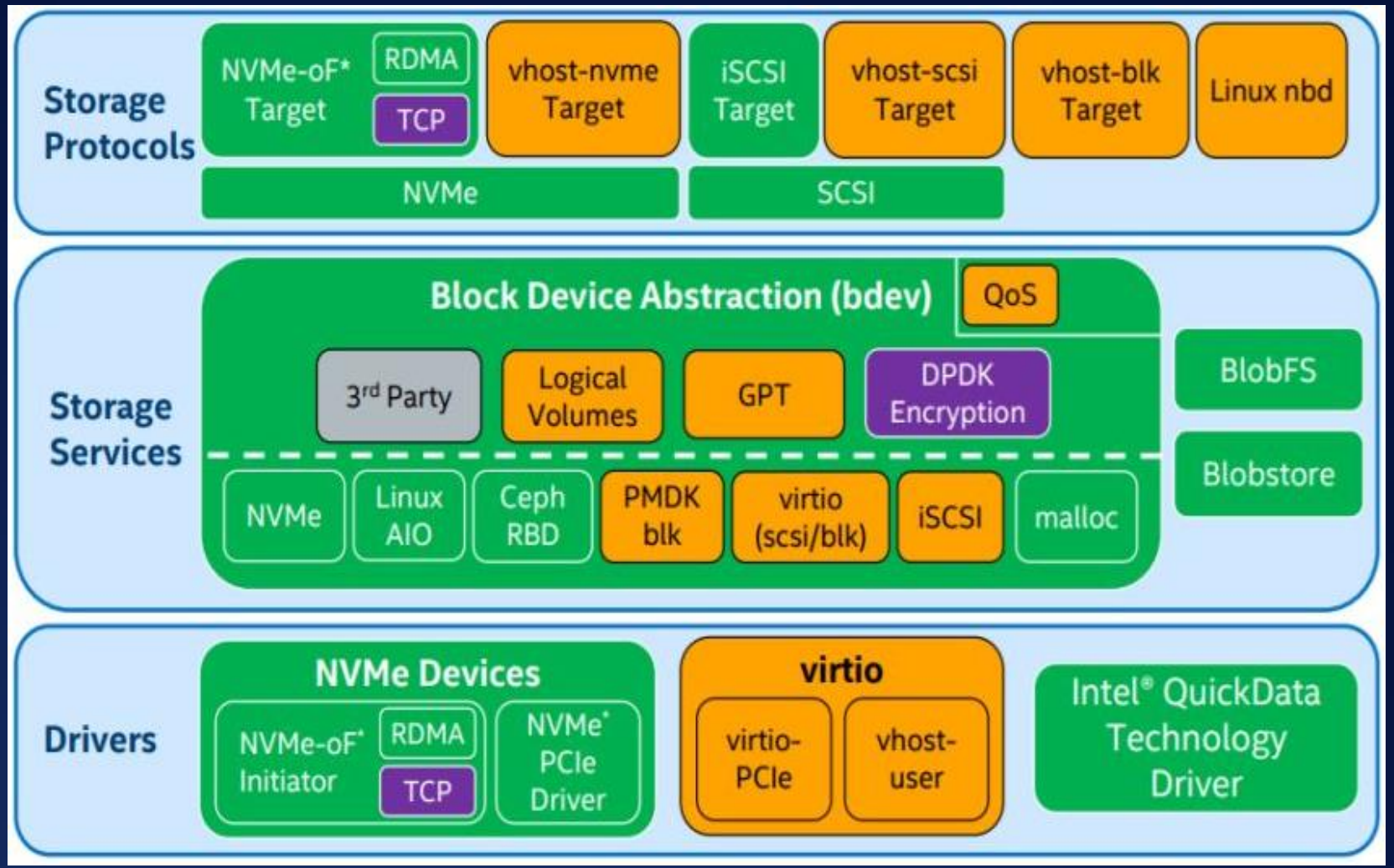
Vhost and Guest share HugePage

1. Submit io to VQ
2. Polling VQ
3. EBS processing IO
4. Irqfd notify guest, io complete



Integrate SPDK on Node

- NVMe-oF tgt(tcp/rdma), ISCSI tgt,Vhost tgt
- bdev qos, io split



Network

TCP

CPU overhead

Data copy

UserSpace to Kernel

High latency

RDMA

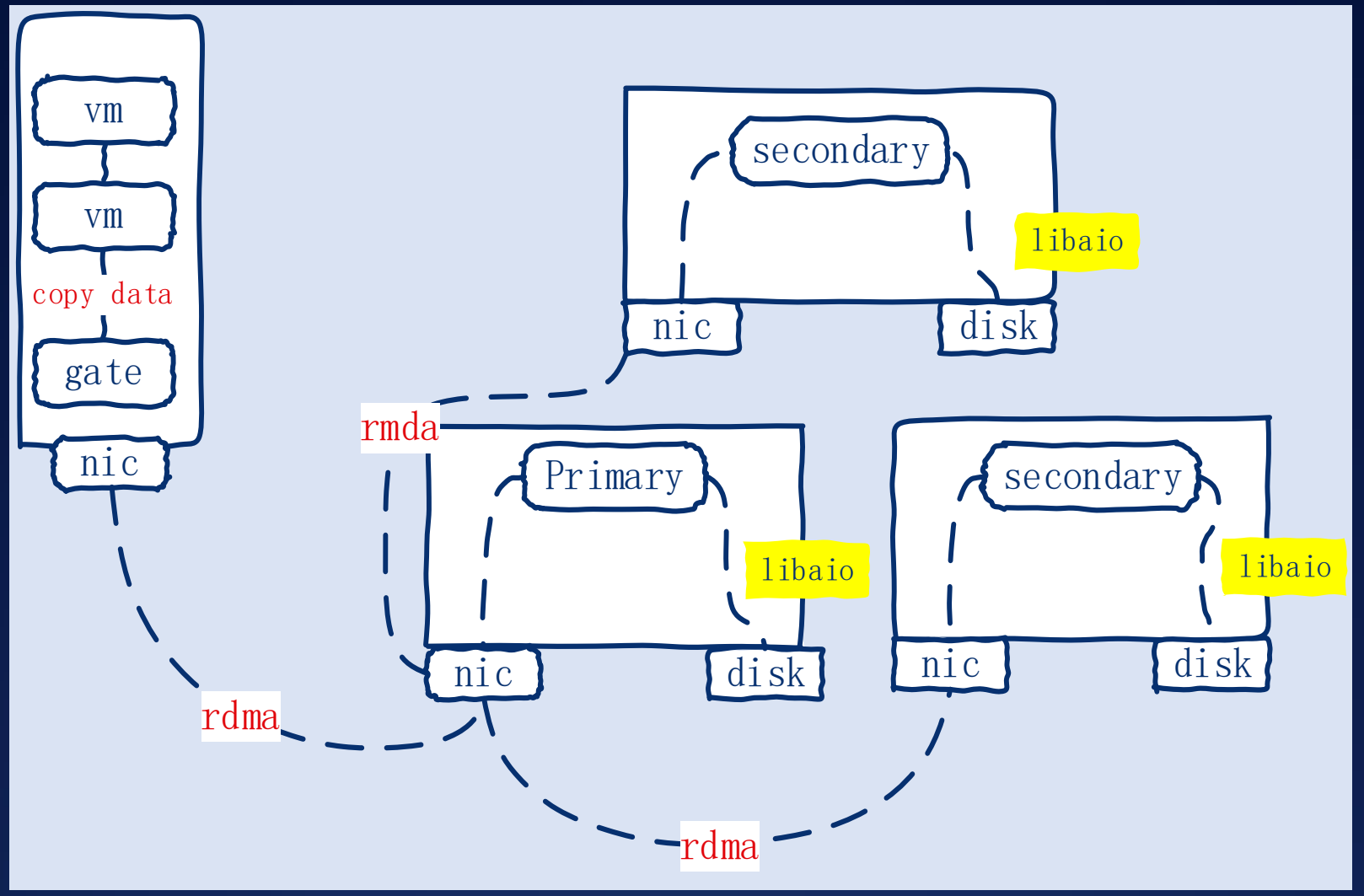
CPU offload to NIC

Zero copy

No context switch

Low latency

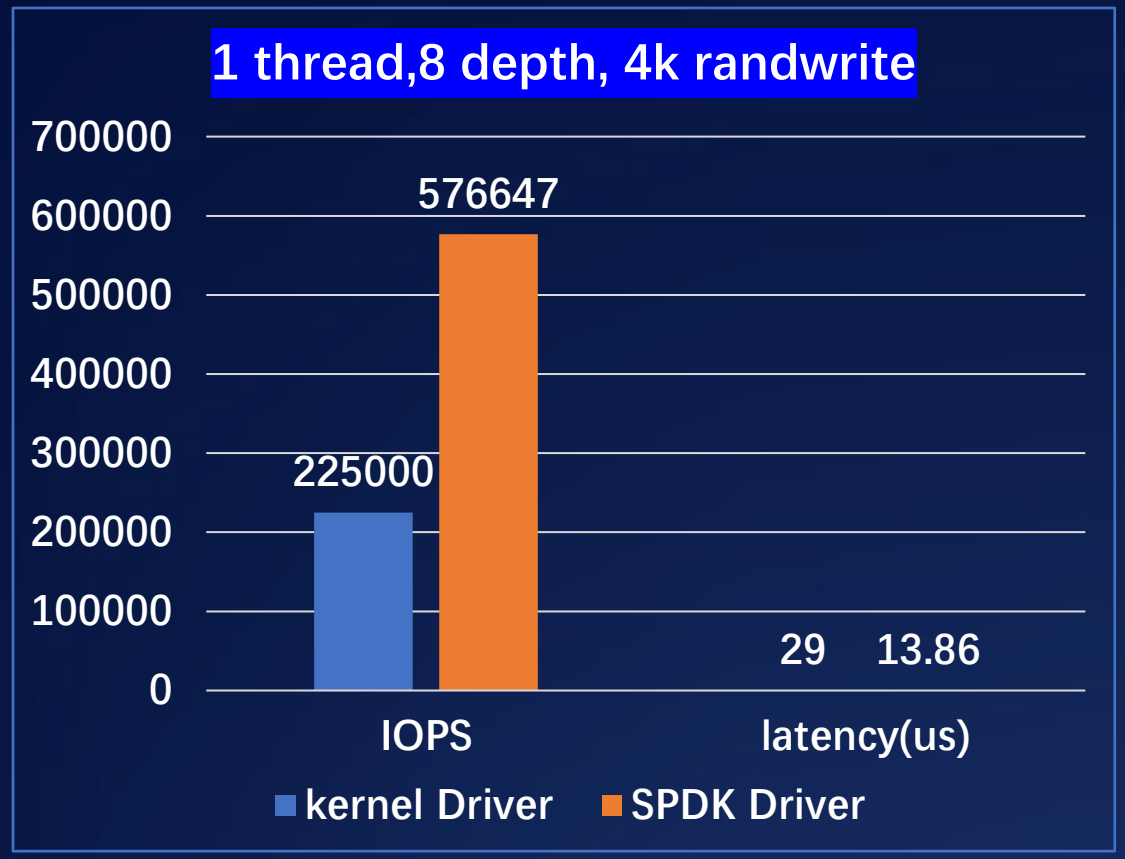
Integrate RDMA

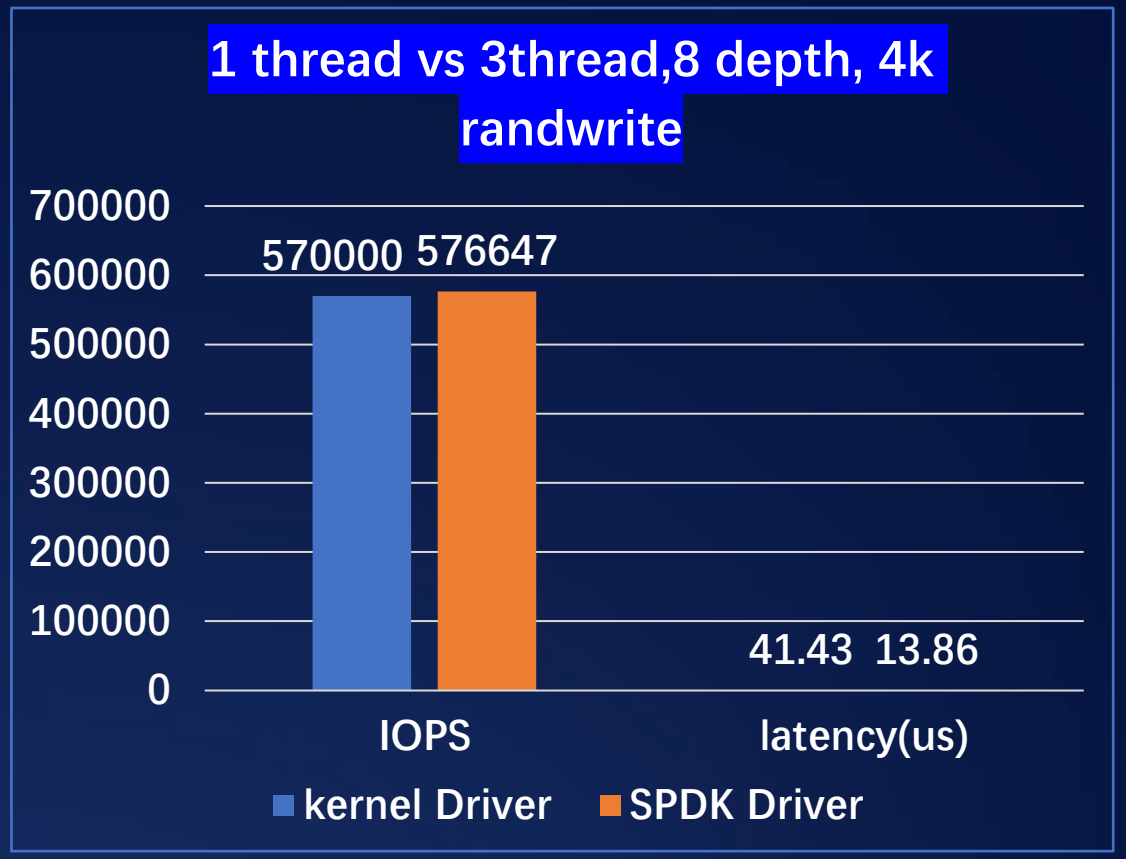


Kernel Driver VS SPDK Driver

kernel Driver	SPDK Driver
Interrupts	Asynchronous Polled Mode
Synchronization	Lockless
System Calls	Userspace Hardware Access
DMA Mapping	Hugepages
Generic Block Layer	Specific for Flash

Kernel Driver VS SPDK Driver

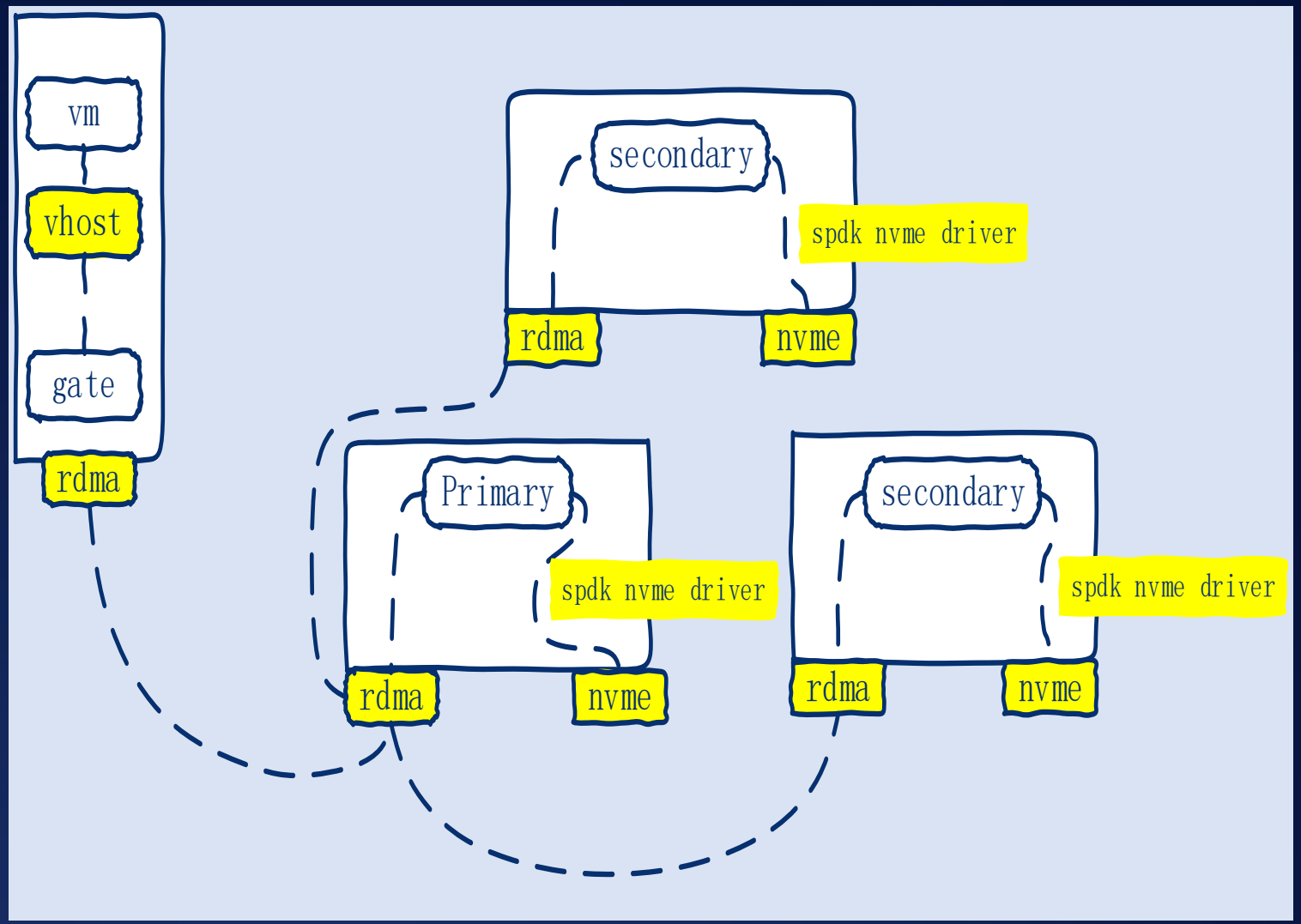




Intel® Optane™ SSD DC D4800X Series (750GB, 2.5in PCIe 2x2, 3D XPoint™)

EBS After Optimization

- 25Gb NIC
- High-density SSD disk
- From qemu virtio to SPDK Vhost
- From TCP to RDMA
- Upgrade Libaio with kernel Driver to SPDK NVMe Driver
- IO path lockless
- Polling on bonding core



中国移动 China Mobile

移动云

Summary

- SPDK Vhost accelerates Host IO path
- SPDK NVMe Driver speeds up backend storage IO path
- RDMA reduces CPU usage

Next Steps

Future Work

- Compatible with HDD
- Replace Brpc to reduce io latency and process switch
- Message passing VS Lock-Based concurrency schemes

中国移动
China Mobile

移动云