BM-Store: A Transparent and High-performance Local Storage Architecture for Bare-metal Clouds Enabling Large-scale Deployment

Yiquan Chen, Jiexiong Xu, Chengkun Wei, Yijing Wang, Xin Yuan, Yangming Zhang, Xulin Yu, Yi Chen, Zeke Wang, Shuibing He, Wenzhi Chen

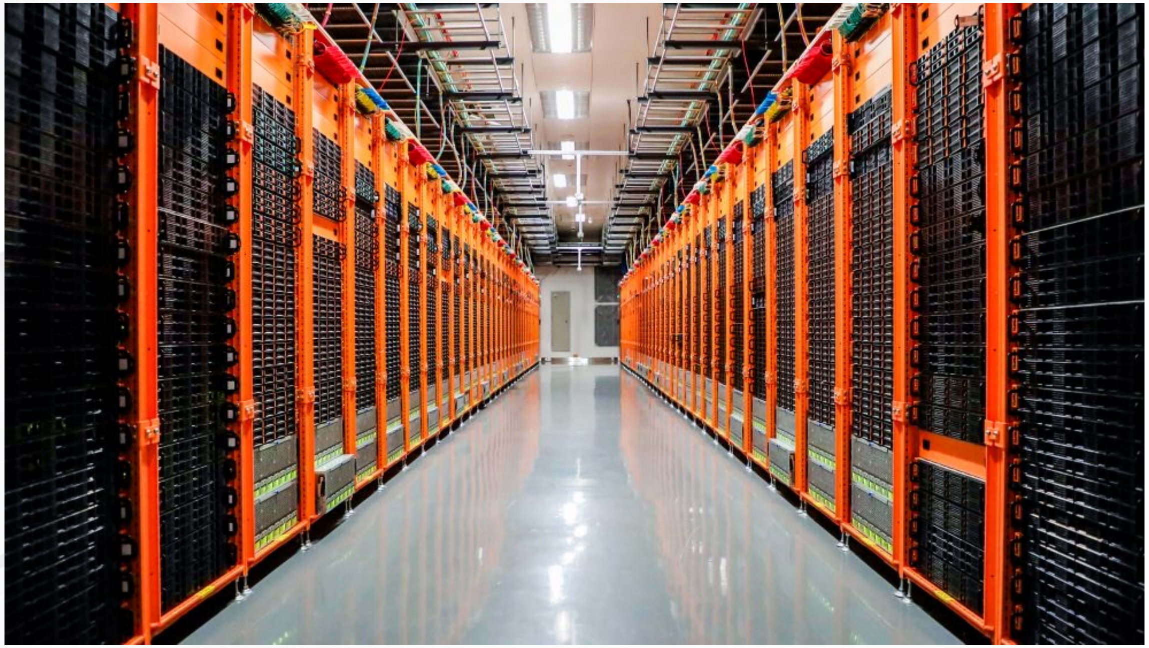
Presenter: Yuejian Xie

Alibaba Group

浙江大學
ZHEJIANG UNIVERSITY

- **Background**
- Motivation
  - Software-based / Hardware-assisted approaches
  - Management and availability challenges
- BM-Store: Storage Virtualization Architecture for Bare-metal Clouds
- Evaluation
- Conclusion

Bare-Metal instance is essential in cloud computing



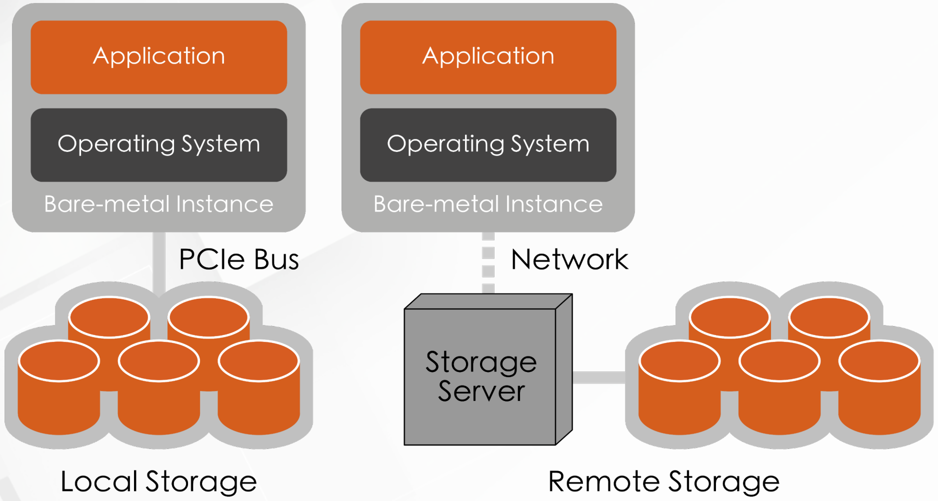
Benefits of bare-metal instances:

- Ultimate Performance
- Delivery in Minutes
- Secure Physical Isolation
- Tenants exclusively own hardware resources and host operating systems

Bare-metal instances in cloud computing have become an essential part that leases dedicated physical servers to tenants

Local storage vs remote storage

Remote Storage ← **Storage Options of Bare-metal Instances** → **Local Storage**



Bare-metal instances access

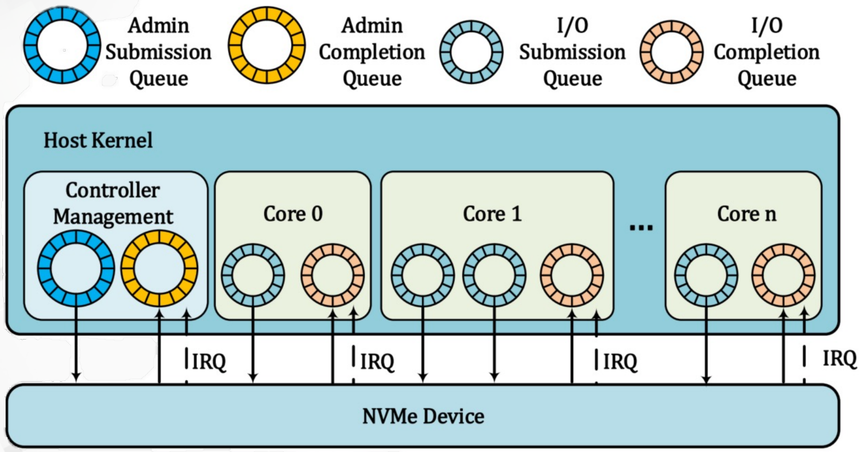
- local storage through **PCIe bus**
- remote storage through the **network**

Bare-metal tenants prefer to choose local storage *(NVMe SSD, SATA HDD & SSD)* **for low cost, high throughput, and low latency**

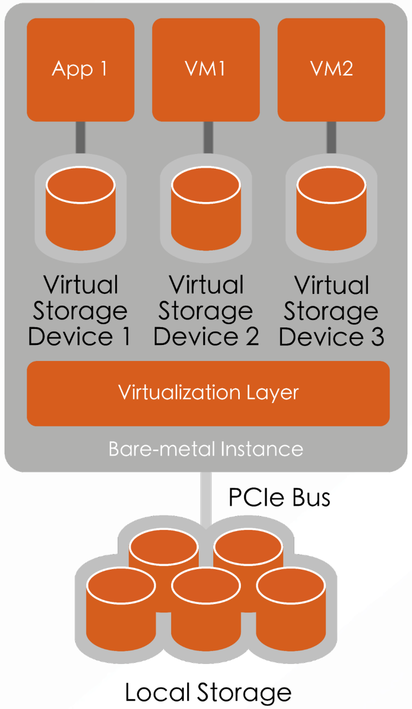
Local storage virtualization

NVMe & Storage Virtualization

- Provide high I/O performance
  - ➢ Set multiple I/O submission/completion queue (CQ/SQ) pairs.
  - ➢ Enable highly parallel I/O processing on multiple CPU cores



NVMe SSD is widely used in cloud computing.



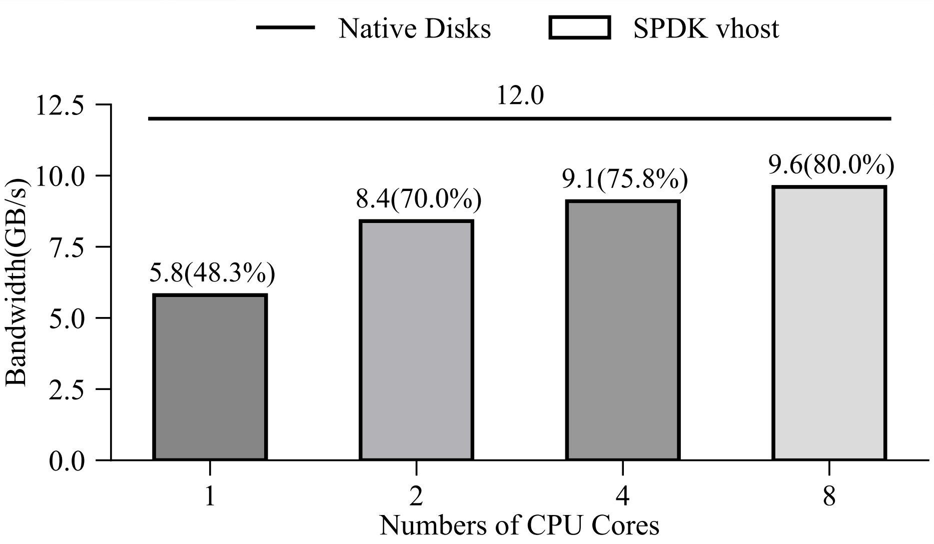
- Virtualization on Bare-Metal
  - ➢ Tenants usually run virtual machines on bare metal instances and must virtualize the local disk.
  - ➢ Bare-metal tenants want to benefit from virtualization.

Bare-metal tenants require virtualization for local storage to enable better **isolation** and higher **resource utilization**.

- Background
- **Motivation**
  - **Software-based / Hardware-assisted approaches**
  - **Management and availability challenges**
- BM-Store: Storage Virtualization Architecture for Bare-metal Clouds
- Evaluation
- Conclusion

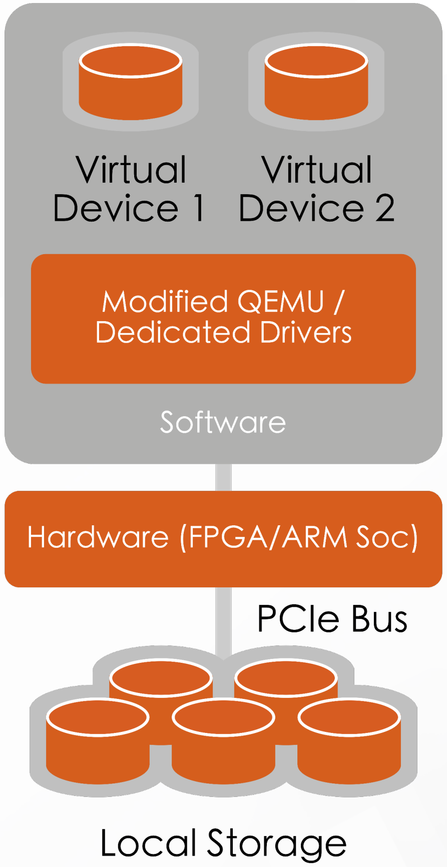
Software-based virtualization for local storage

- **SPDK vhost:** Polling-based
  - ➢Software-based high-performance approach
  - ➢Need dedicate CPU cores to emulate virtual devices
  - ➢Consumes too many valuable CPU cores for virtualization



Traditional hardware-assisted virtualization

- **Offload virtualization functions to FPGA or ARM SoC**
  - ➢Save the host CPU cores required for storage virtualization
  - ➢Need modified host/guest OS, QEMU, and inevitable customized drivers for initialization and configuration
  - ➢Difficult to be deployed in bare-metal instances because cloud vendors can’t access host OS.



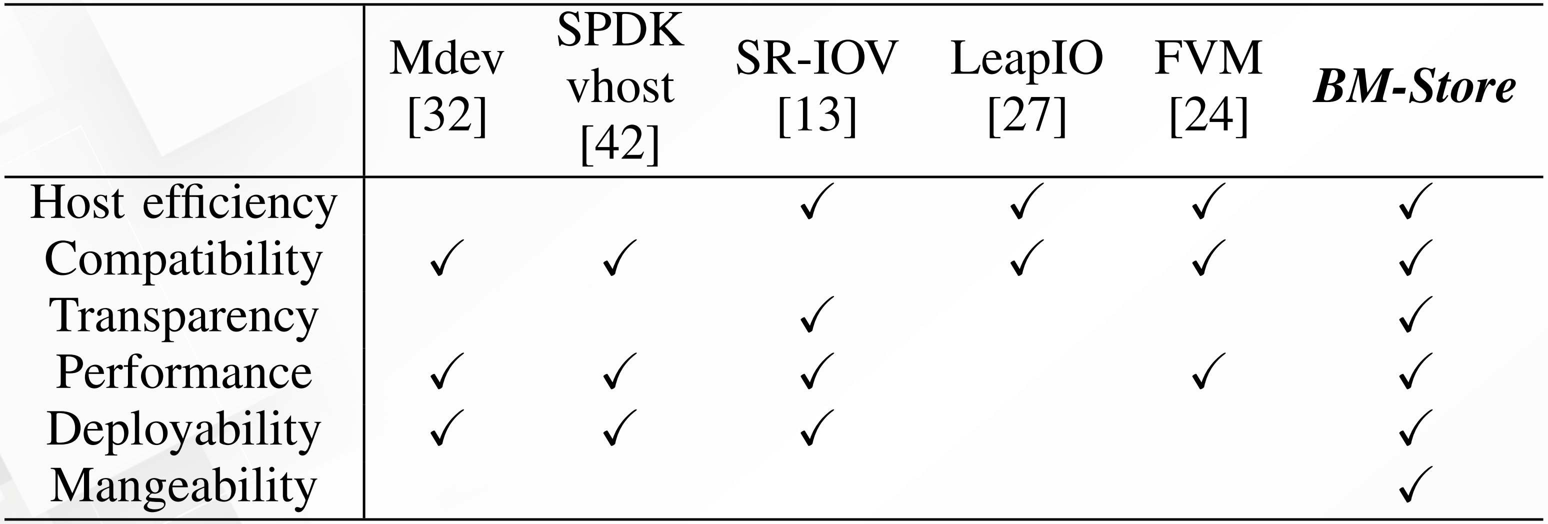
Management and availability challenges of Bare-metal

- Cloud vendors must provide
  - Hardware configuration and management
  - Hot upgrade and monitor for availability

- Existing approaches
  - Do not focus on the manageability and availability challenges of local storage services in the production environment.
  - Difficult to monitor or manage storage devices, such as health info, performance status, etc.

Design goals

- Host-efficient
- Transparent and high compatibility
- Virtualization and high performance
- Manageability and high availability



	Mdev [32]	SPDK vhost [42]	SR-IOV [13]	LeapIO [27]	FVM [24]	***BM-Store***
Host efficiency			✓	✓	✓	✓
Compatibility	✓	✓		✓	✓	✓
Transparency			✓			✓
Performance	✓	✓	✓		✓	✓
Deployability	✓	✓	✓			✓
Mangeability						✓

- Background
- Motivation
  - Software-based / Hardware-assisted approaches
  - Management and availability challenges
- **BM-Store: Storage Virtualization Architecture for Bare-metal Cloud**
- Evaluation
- Conclusion

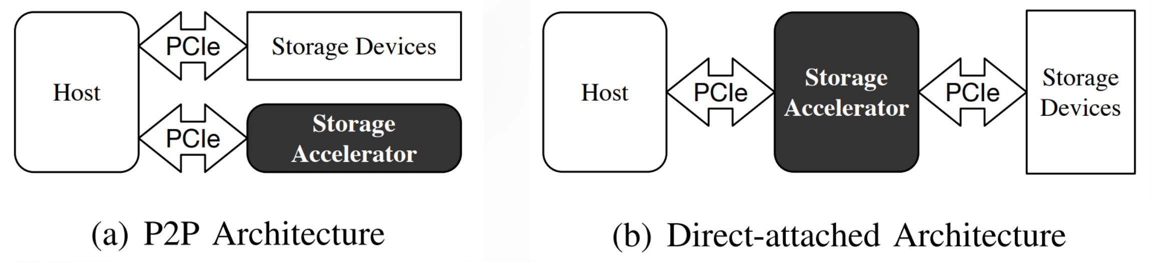
Key ideas and benefits

- Hardware-based
  - **Save the host resources** required for storage virtualization
- Transparent architecture
  - **Do not need modified QEMU / customized drivers** to deploy on bare-metal instances
- High-performance
  - Achieve **near-native** performance
- Manageability and high availability
  - Enable cloud vendors to manage local storage even if they cannot access the operating system in bare-metal instances.

Transparent architecture

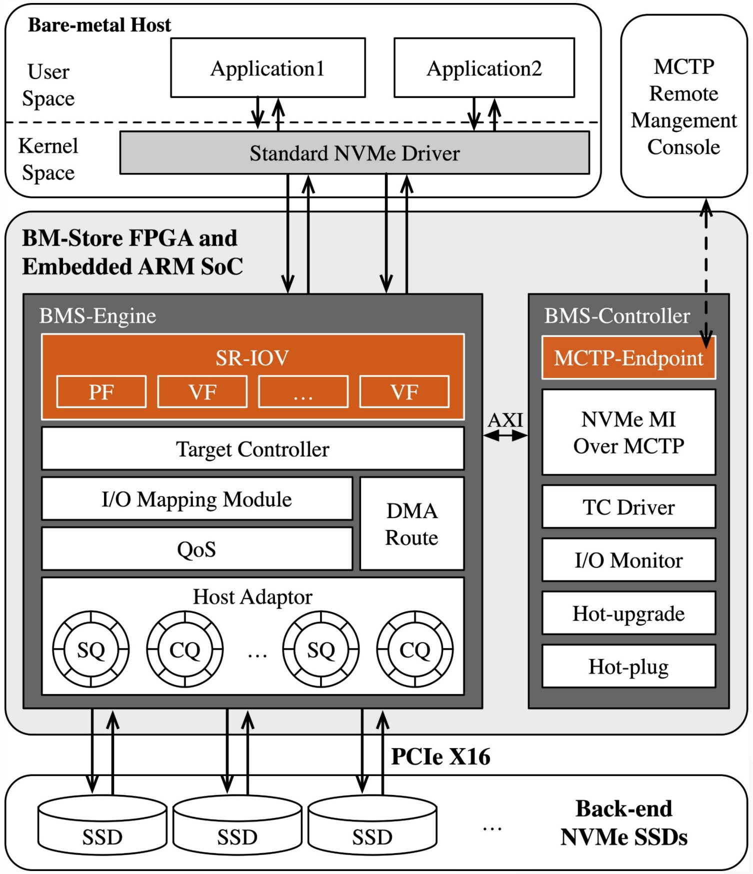
Alibaba Group 浙江大学 ZHEJIANG UNIVERSITY

- Direct-attached
- Standard SR-IOV layer
- Supports HDDs and SATA SSDs
- MCTP out of band management
  - No customized drivers
  - No modification to host OS/QEMU



(a) P2P Architecture (b) Direct-attached Architecture

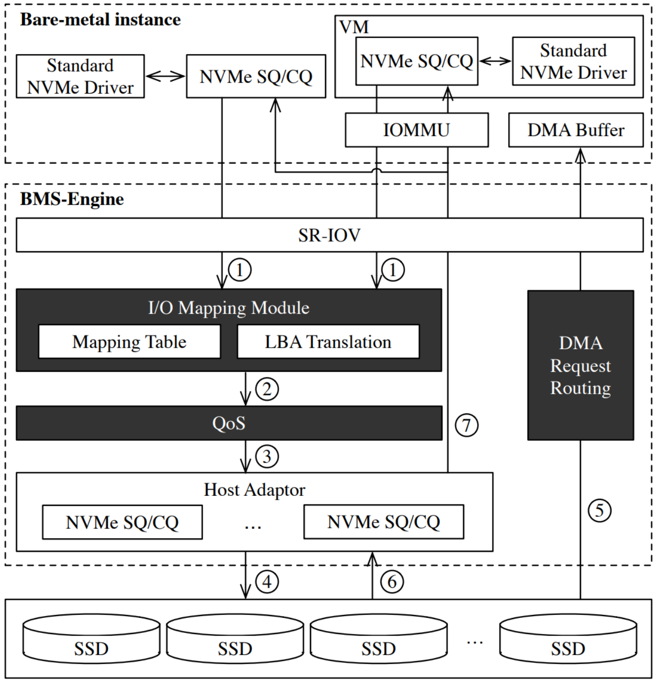
Tenants can access virtualized storage resources through standard NVMe drivers. BM-Store is transparent to host to enable deployment in bare-metal instances



Hardware accelerated I/O path

Alibaba Group · ZHEJIANG UNIVERSITY

- FPGA-accelerated virtualization layer
- Adopting **DMA request routing** to enable Zero-Copy
  - ➢ Originally, the data must be transferred to the FPGA memory and then copied to the host memory
  - ➢ DMA request routing can **eliminate duplicate data copies** and achieving near-native performance



BM-Store achieves extreme performance close to native disks through **FPGA-accelerated I/O path** and **zero-copy mechanism**

Out-of-Band management

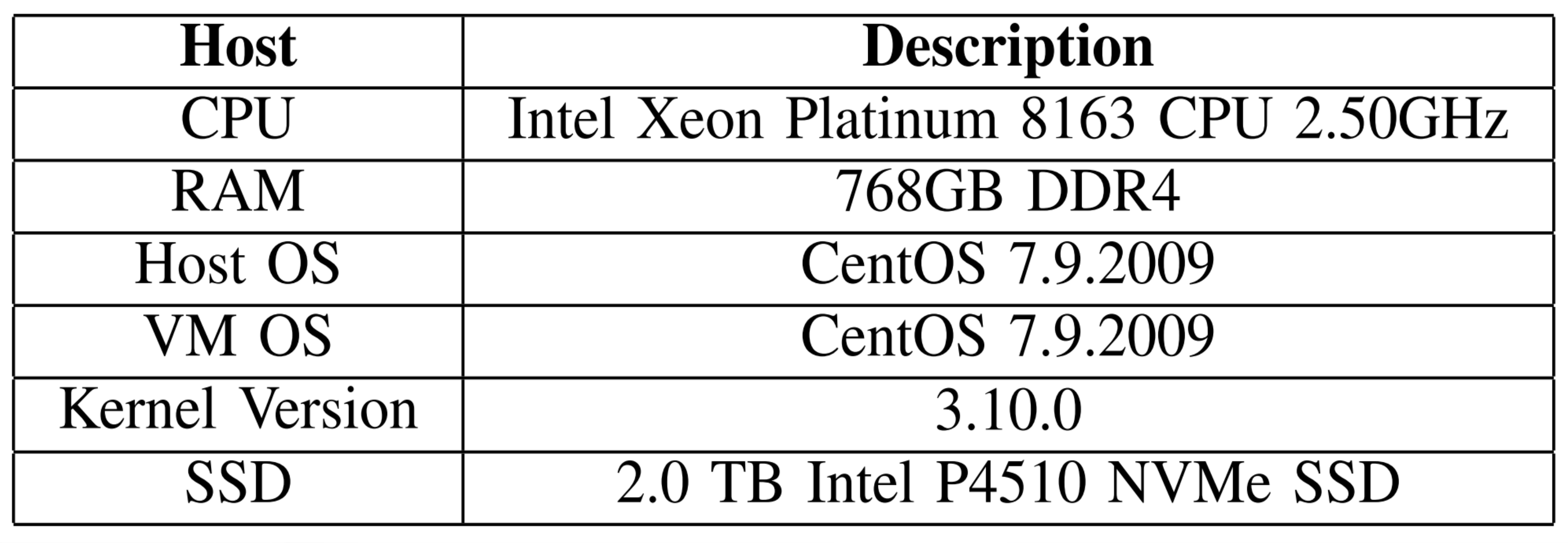
- MCTP out-of-band management
- Hot-upgrade and hot plug to enhance local storage service availability

Tenants can access virtualized storage resources through only standard NVMe drivers. BM-Store is transparent to the host to enable large-scale deployment in bare-metal instances

- Background
- Motivation
  - Software-based / Hardware-assisted approaches
  - Management and availability challenges
- BM-Store: Storage Virtualization Architecture for Bare-metal Clouds
- **Evaluation**
- Conclusion

Experimental configuration

- We evaluate BM-Store on servers in Alibaba Cloud.
- The configurations are as follows:

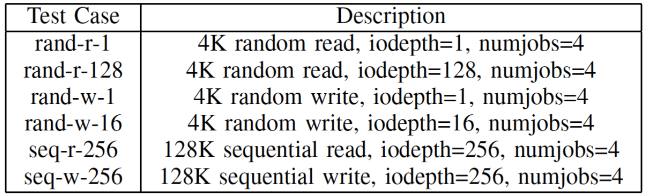


Host	Description
CPU	Intel Xeon Platinum 8163 CPU 2.50GHz
RAM	768GB DDR4
Host OS	CentOS 7.9.2009
VM OS	CentOS 7.9.2009
Kernel Version	3.10.0
SSD	2.0 TB Intel P4510 NVMe SSD

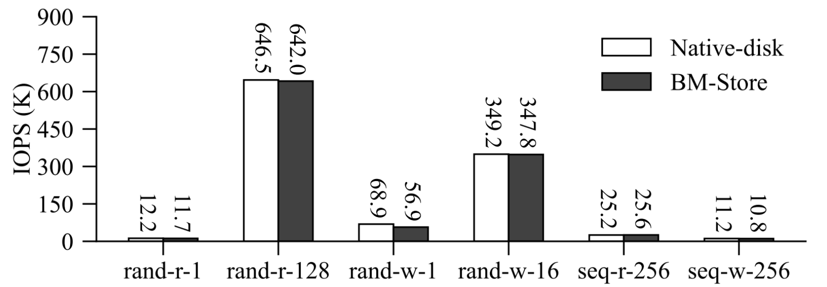
Bare-Metal I/O performance

- **BM-Store vs Native disk in bare-metal machine**

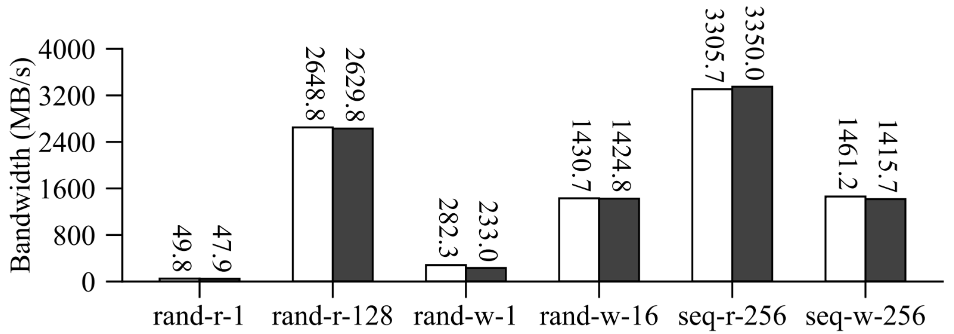
BM-Store adds only 3us extra latency compared to the native disk.



Test Case	Description
rand-r-1	4K random read, iodepth=1, numjobs=4
rand-r-128	4K random read, iodepth=128, numjobs=4
rand-w-1	4K random write, iodepth=1, numjobs=4
rand-w-16	4K random write, iodepth=16, numjobs=4
seq-r-256	128K sequential read, iodepth=256, numjobs=4
seq-w-256	128K sequential write, iodepth=256, numjobs=4



(a) IOPS

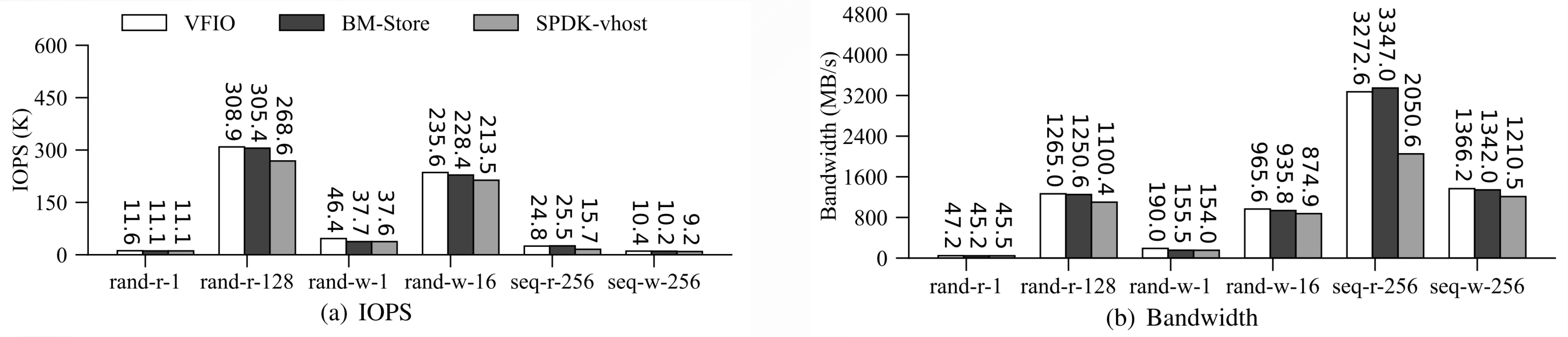


(b) Bandwidth

BM-Store can achieve near-native performance from 96.2% to 101.4% .

Virtual machine I/O performance

- **BM-Store** *vs* **VFIO** and **SPDK** vhost in virtual machine
  - SPDK vhost has to consume extra 1 core for 1 SSD
  - BM-Store and VFIO do not need host CPU resources



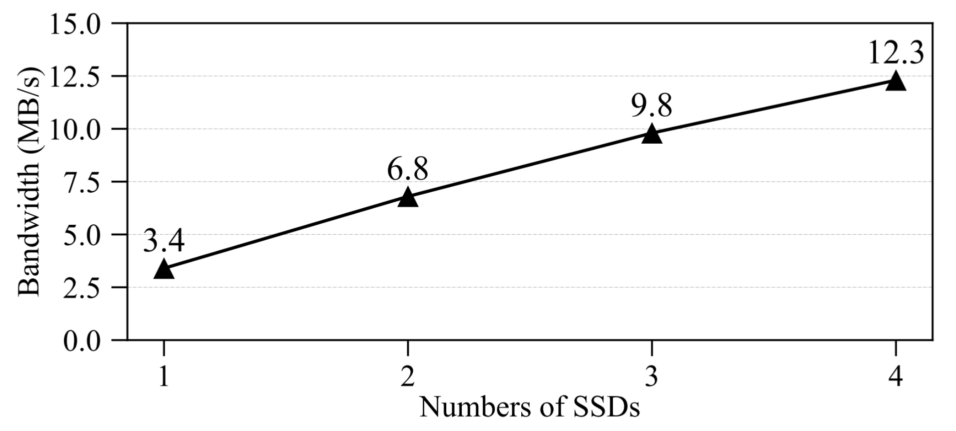
(a) IOPS

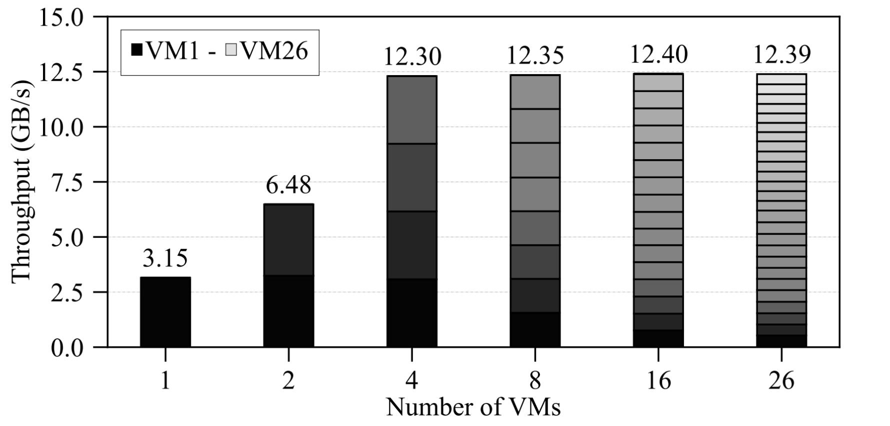
(b) Bandwidth

BM-Store can achieve virtualization performance close to VFIO and outperform SPDK vhost

Scalability and fairness

- **BM-Store with different number of SSDs in virtual machines**
  - Evaluate the bandwidth with 1 – 4 NVMe SSDs
  - Evaluate the bandwidth with 1 NVMe SSDs in 1 – 26 VMs

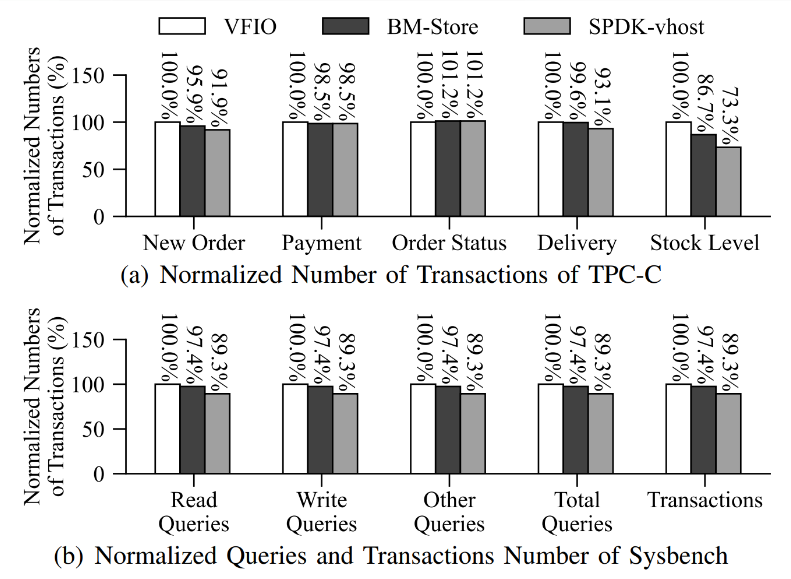




BM-Store can ensure promising scalability and maintain the fairness of each virtual machine as well as the overall performance of I/O

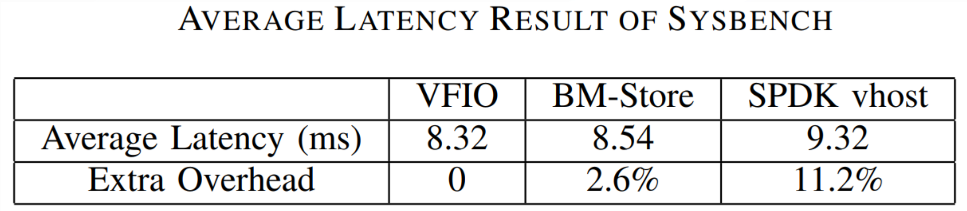
RocksDB and MySQL performance

- BM-Store vs VFIO and SPDK vhost in virtual machine



(a) Normalized Number of Transactions of TPC-C

(b) Normalized Queries and Transactions Number of Sysbench



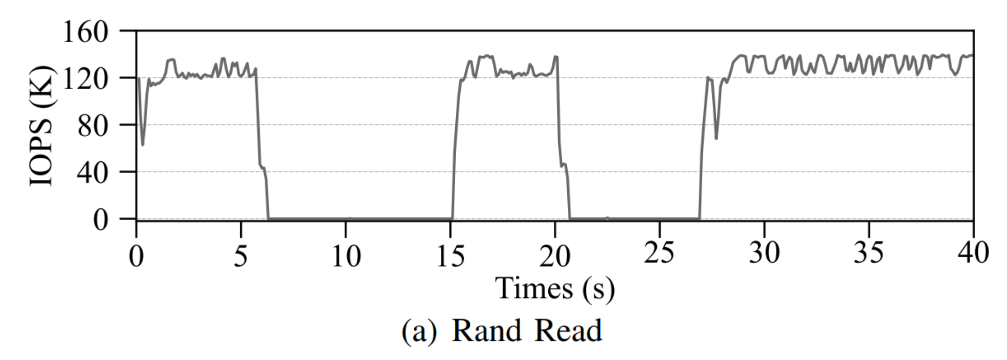
AVERAGE LATENCY RESULT OF SYSBENCH

	VFIO	BM-Store	SPDK vhost
Average Latency (ms)	8.32	8.54	9.32
Extra Overhead	0	2.6%	11.2%

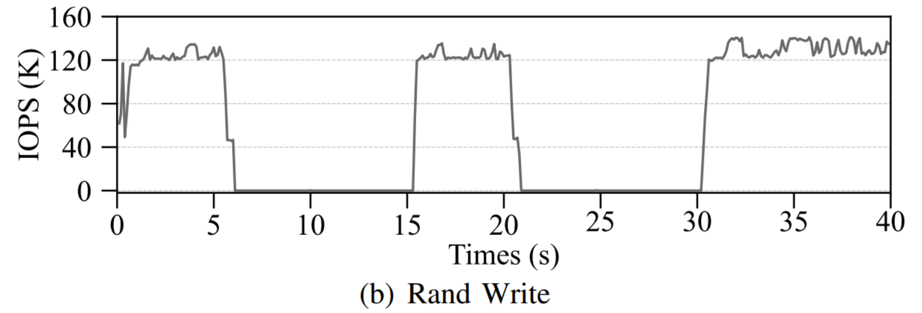
BM-Store architecture provides closed-to-native disk performance for real-world applications.

Hot-upgrade for availability

- **Evaluate the hot-upgrade time of BM-Store**
  - ➢ Performing hot-upgrade of BM-Store when doing random read/write



(a) Rand Read



(b) Rand Write

BM-Store can provide high availability for local storage services in the production environment.

Compatibility and TCO analysis

•Compatibility of BM-Store Architecture

- ➢ Use standard NVMe driver and no additional software modification
- ➢ Can further easily support various storage devices such as SATA HDDs and ZNS SSDs.

•TCO Analysis

- ➢ SPDK vhost consumes 16 HT CPUs for 16 SSDs on each server and causes resource fragments (128 GB memory/2 SSDs).
- ➢ BM-Store can release 16 HT CPUs to sell two more instances (8 HT CPU/64 GB Memory/1 SSD) and get about 11.3% TCO benefit.

THANK YOU!

http://arc.zju.edu.cn/

ARClab

浙江大学
计算机系统结构实验室
Computer Architecture Laboratory
of Zhejiang University

Alibaba Group

浙江大學
ZHEJIANG UNIVERSITY

yuejian.xie@Alibaba-inc.com