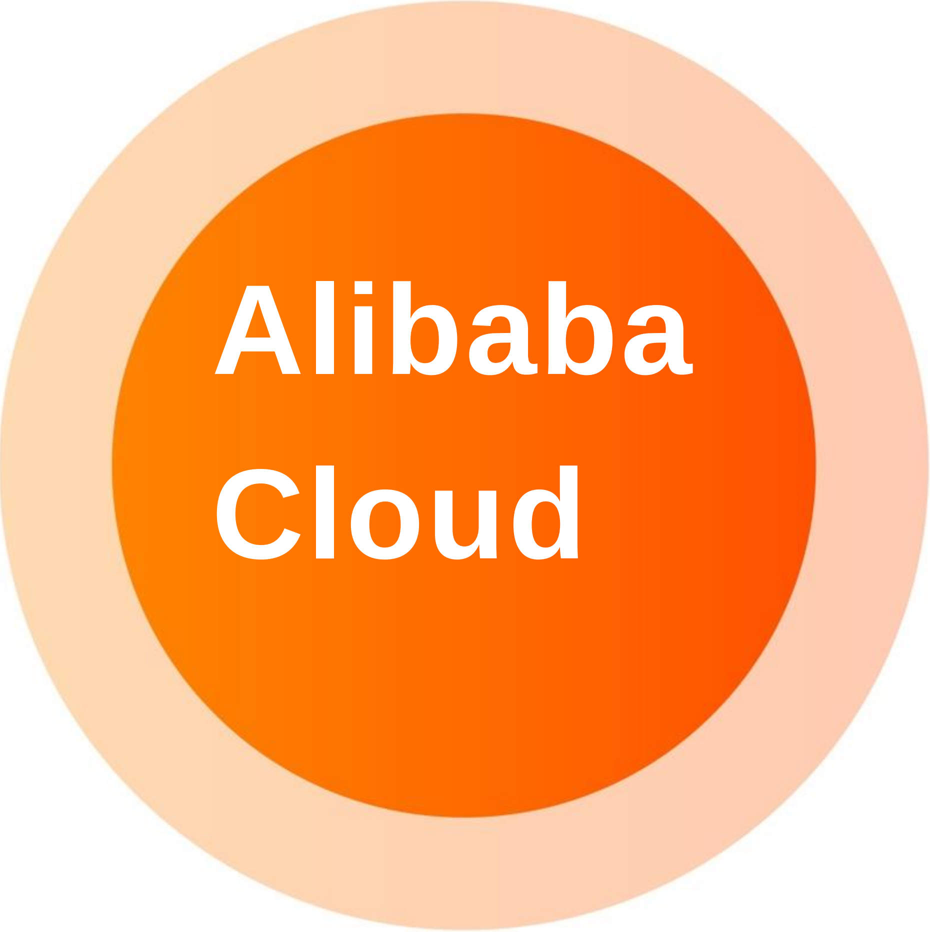
Alibaba Cloud | Worldwide Cloud Services Partner

HotChips

High-density Multi-tenant Bare-metal Cloud with Memory Expansion SoC and Power Management

Authors:
xiantao.zxt@alibaba-inc.com
zhengxiao.zx@Alibaba-inc.com
justin.song@alibaba-inc.com

Why Baremetal Cloud and What is X-Dragon?



1 For security and isolation

2 For multi-tenancy and cost efficiency

3 For single-thread performance

4 For interoperability and manageability

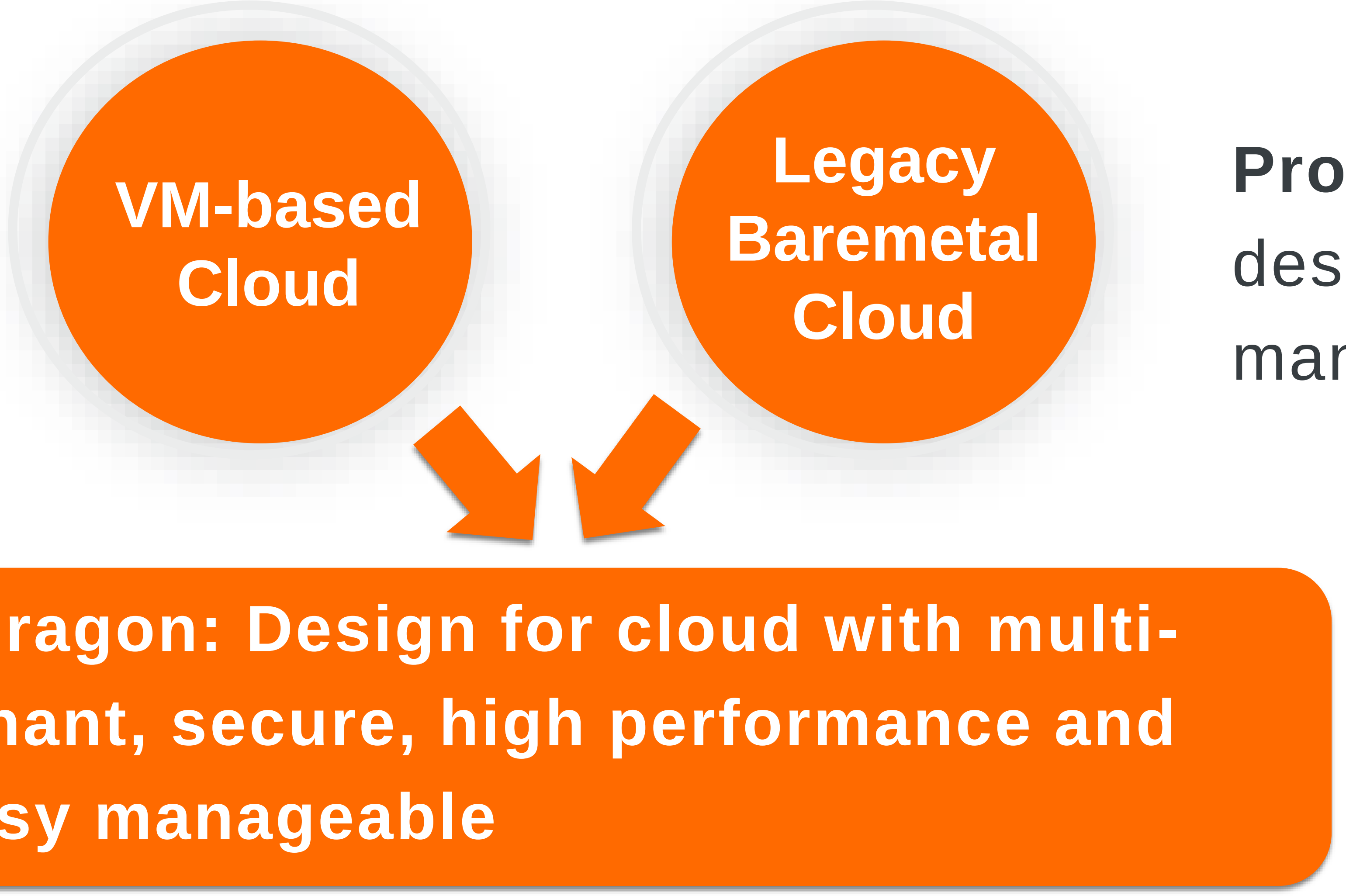
X-Dragon: multi-tenant BM-Guests in same Server

Problems

There are VM-based cloud, single-tenant bare-metal cloud and BM-Hive(Multi-tenants bare-metal cloud) in Datacenter

Problem1: VM-based Cloud has non-ignorable virtualization overhead, isolation/security concern and limited single thread performance, but good manageability

Problem2: Existing bare-metal cloud design for single tenant, lack of manageability and also costly



VM-based Cloud

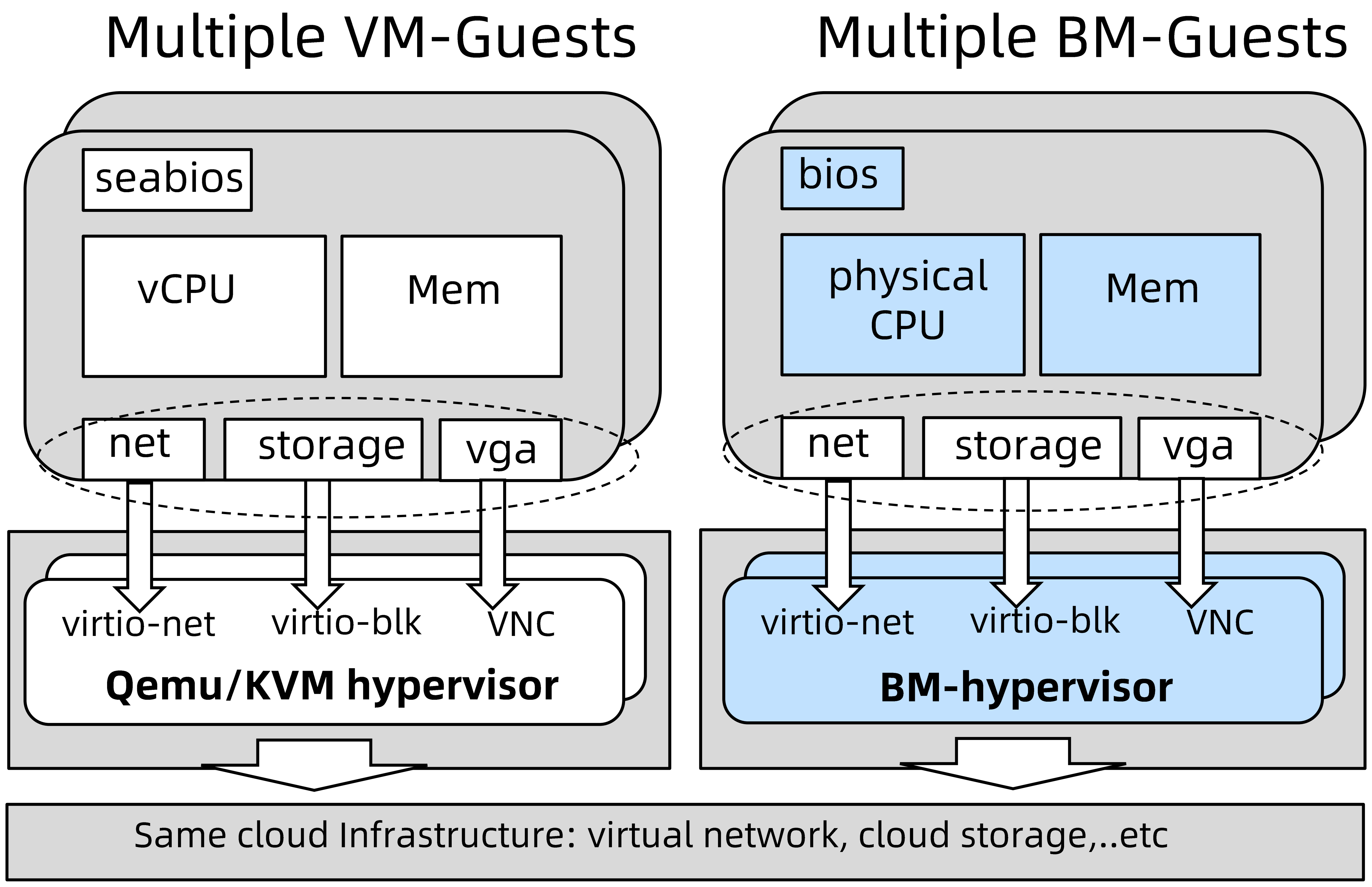
Legacy Baremetal Cloud

Xdragon: Design for cloud with multi-tenant, secure, high performance and easy manageable

Service	Security	Isolation	Performance	Density
VM-based cloud	Side-channel and Dos attacks because of resource sharing	Weak isolation because of resource sharing	CPU, Memory, and I/O overhead caused by virtualization	Very high density through server over-provisioning
Single-tenant bare-metal cloud	N/A	Strong isolation due to exclusive access to system	Native performance	Very low density, one user per server, leading to high cost
X-Dragon	No side-channel or Dos attacks due to hardware-based isolation; Protected hardware resources, particularly the firmware	Strong hardware-based isolation	Native CPU and memory performance; para-virtualized I/O with minor overhead	High, 16 BM-Guests per server at most

X-Dragon High Level View in Cloud

Alibaba Cloud
Worldwide Cloud Services Partner



KVM vs X-Dragon

Same cloud infrastructure

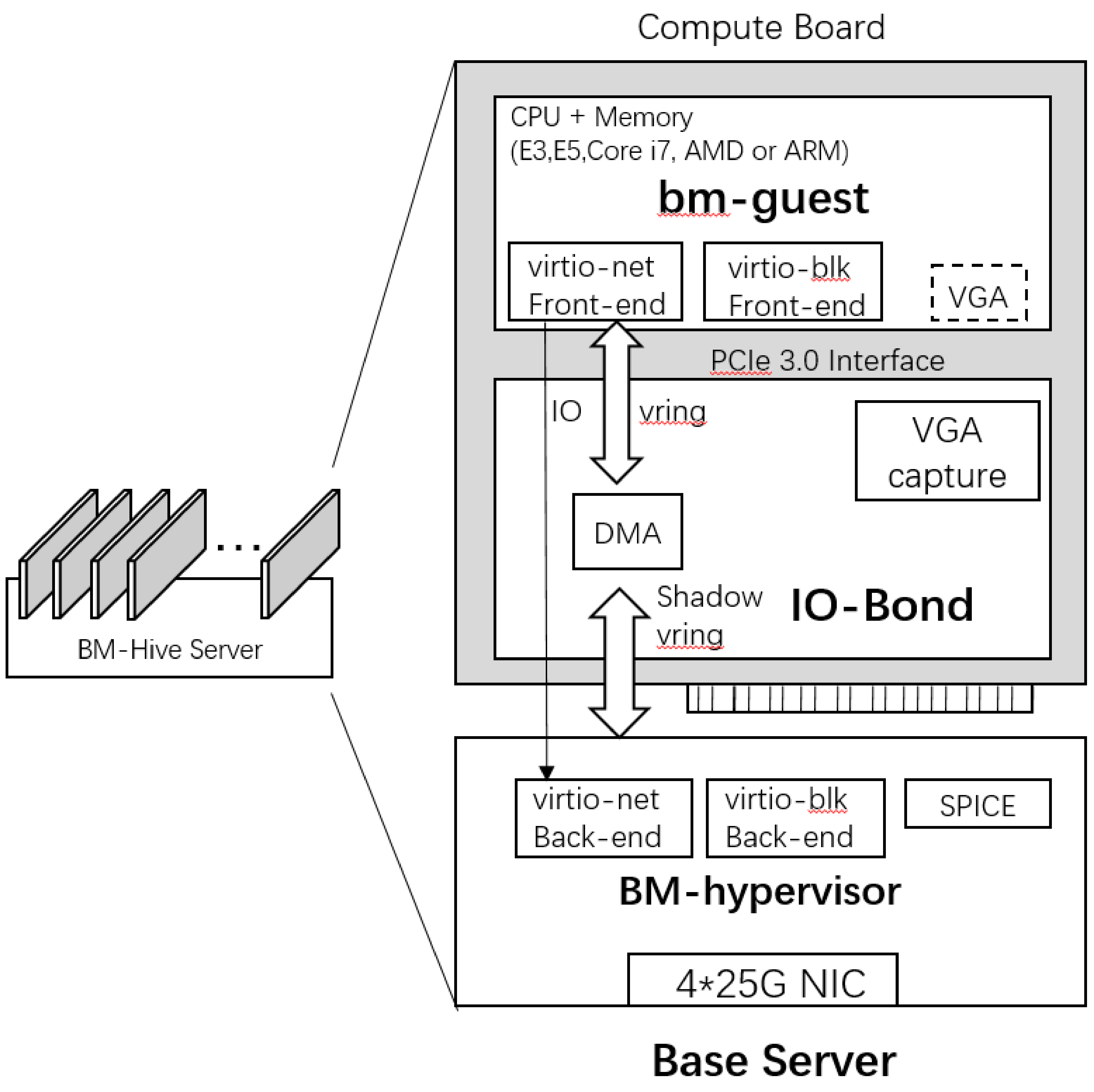
Same tools to manage

Both Multi-tenants

More secure and selectable bare-metal performance

X-Dragon System Architecture

Alibaba Cloud
Worldwide Cloud Services Partner

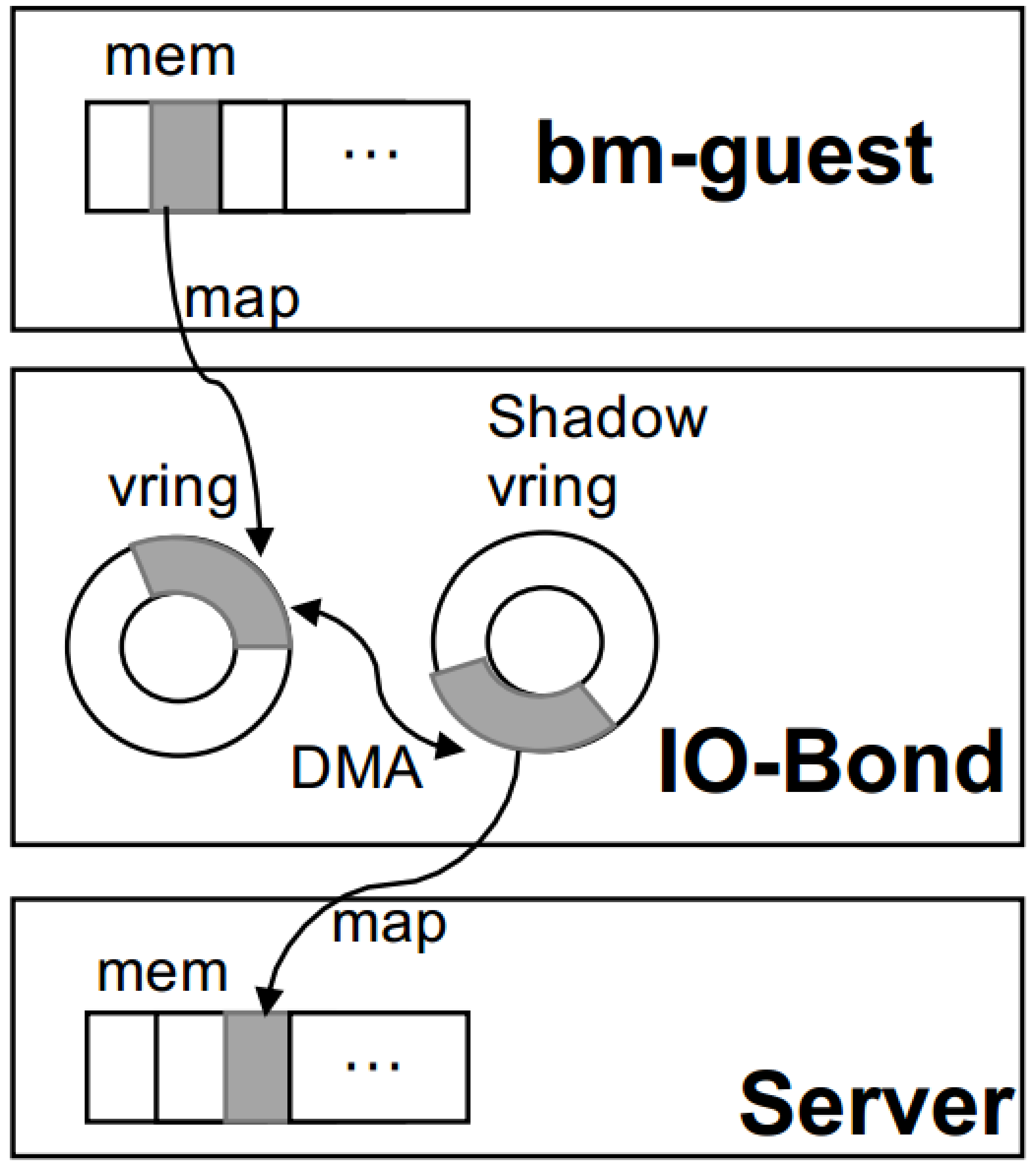


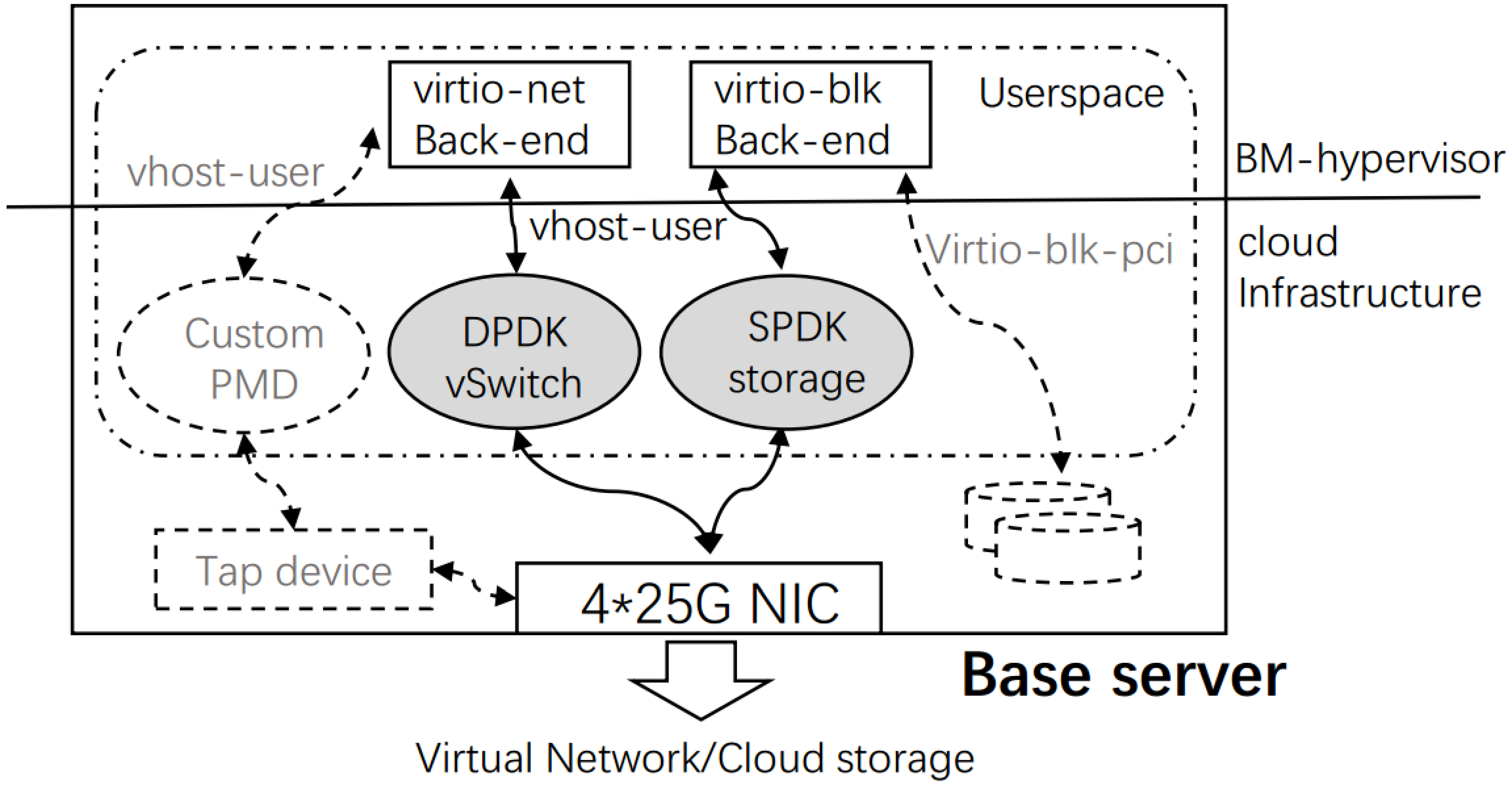
1 Compute Boards + Base Server

2 Hardware implementation of virtio devices

3 Custom backend: BM-Hypervisor

X-Dragon: IO Bond and Backend





Shadow Ring buffer design
Transfer data between computing board and backend base server

BM-Hypervisor design
Emulate virtio-devices, and connect into existing cloud infrastructure

Evaluation: CPU/Mem/IO performance

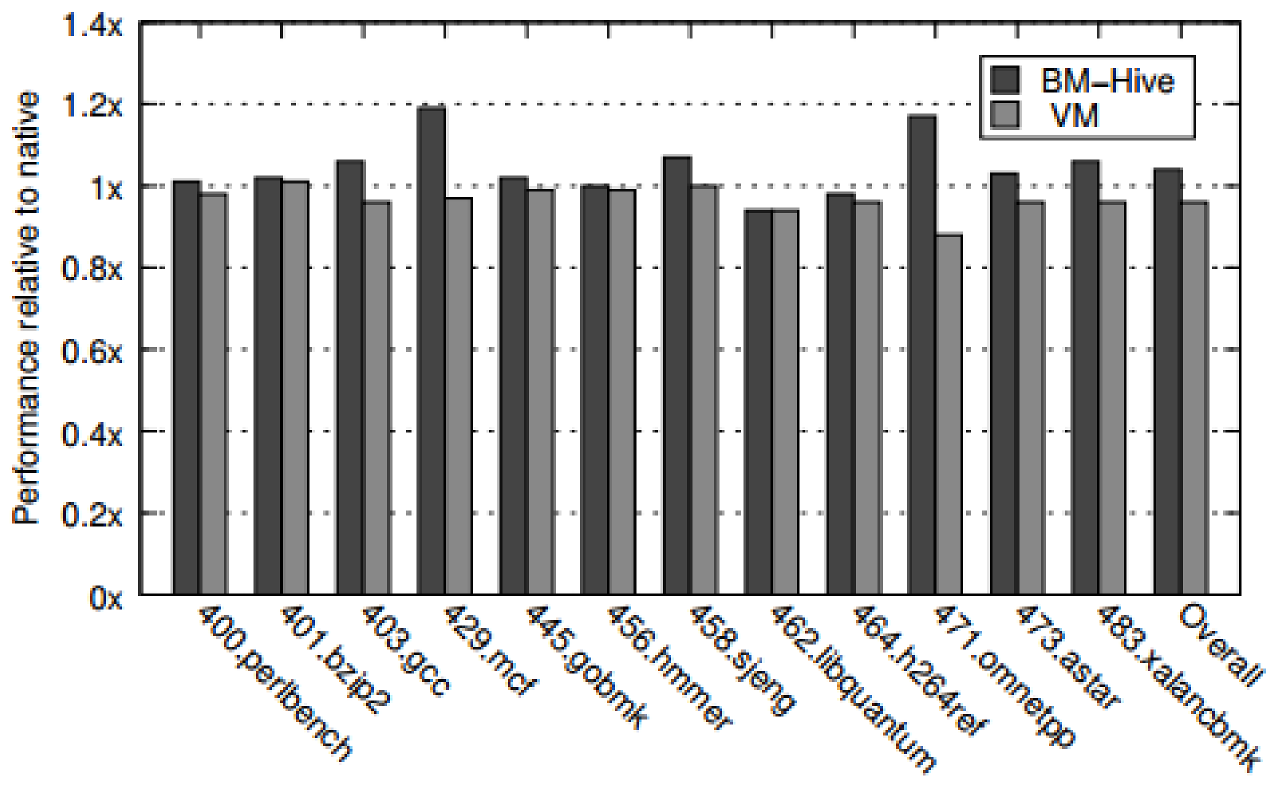


Figure 7. CPU performance by SPEC CPU2006

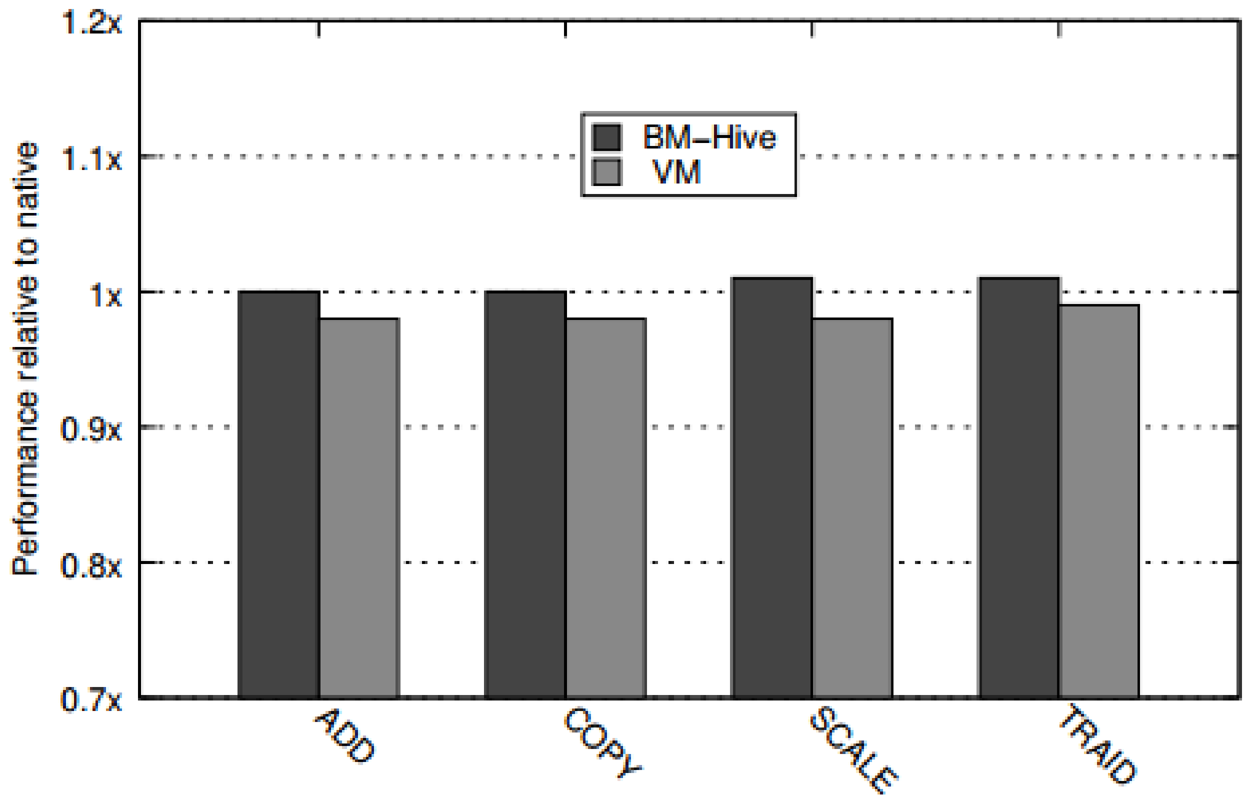


Figure 8. Memory bandwidth by STREAM multi-thread

- X-Dragon BM-Guest vs Native vs VM: BM-Guests are slightly better performance than VM
- Memory bandwidth： BM-Guests are same as Native. VM 98% of BM-Guests under load
- Network PPS: Same PPS rate, however more implied volatility.
- Latency: Same in application level, longer path then DPDK bypass-kernel testing
- Storage: substantially better than VM from latency and long tail.

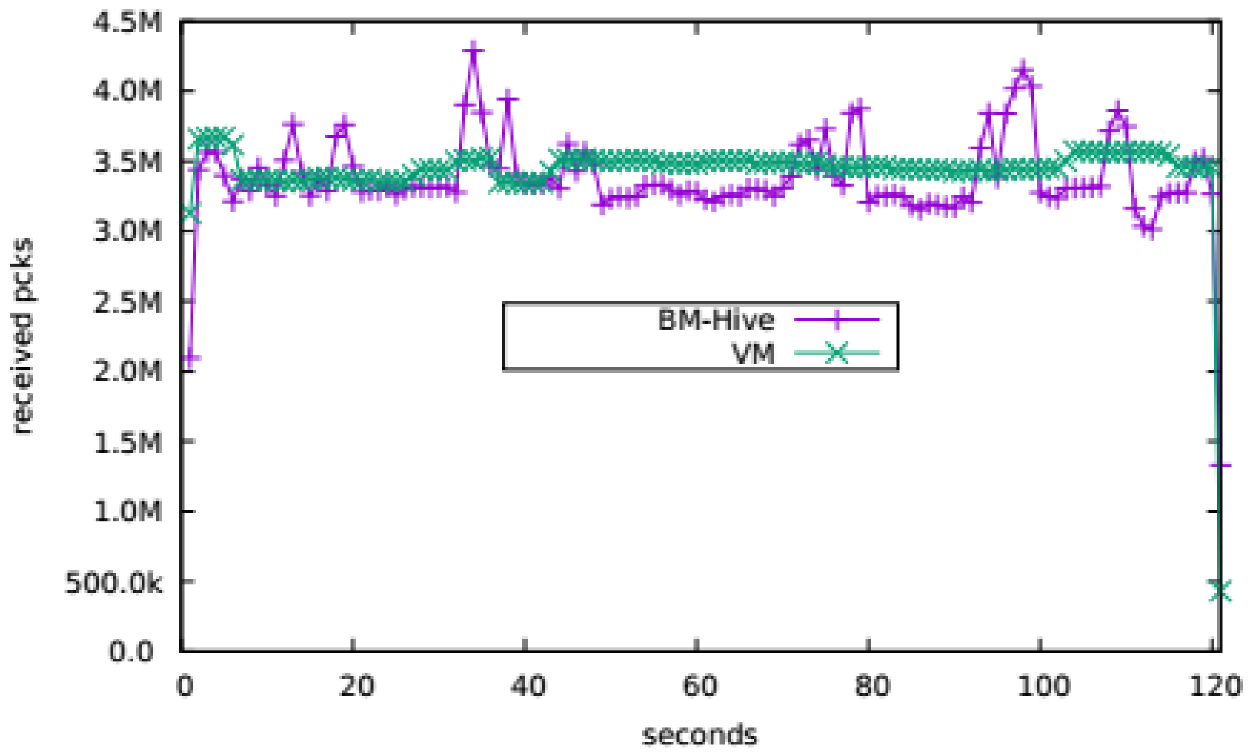


Figure 9. UDP packet receive rate

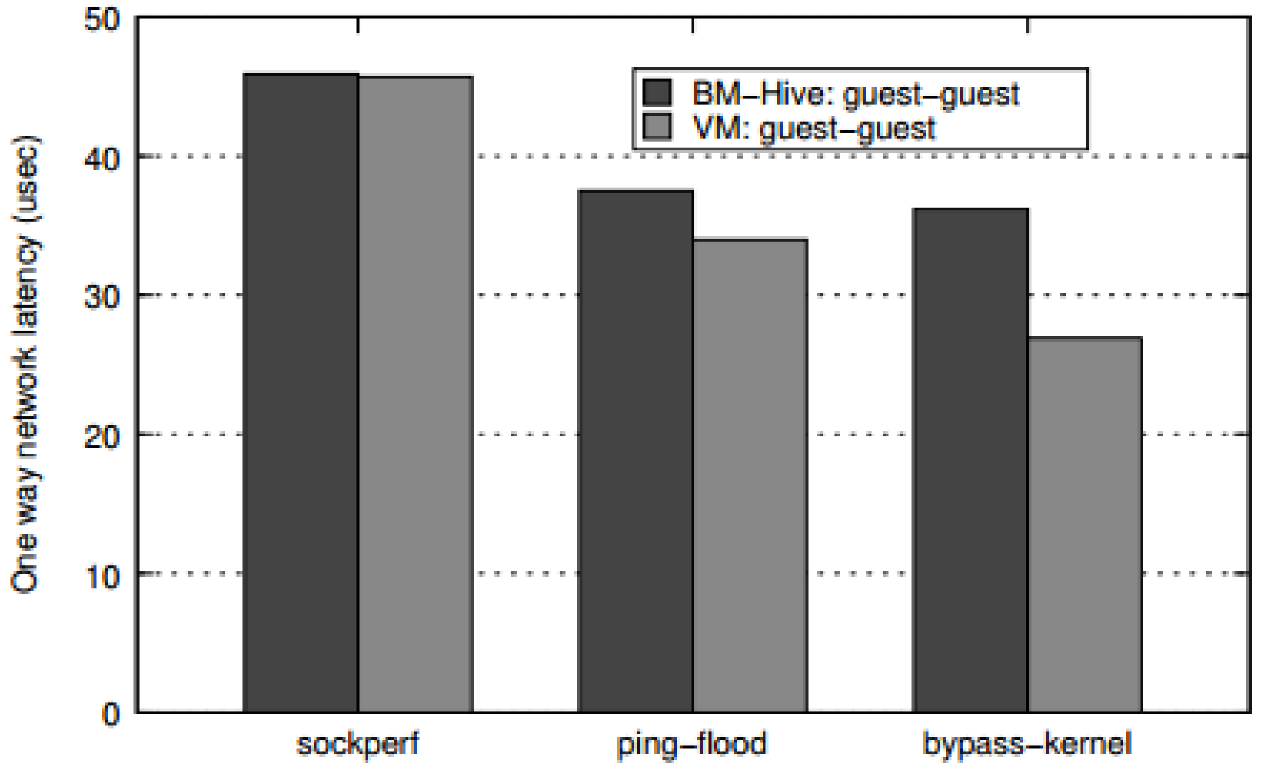


Figure 10. UDP and ping latency

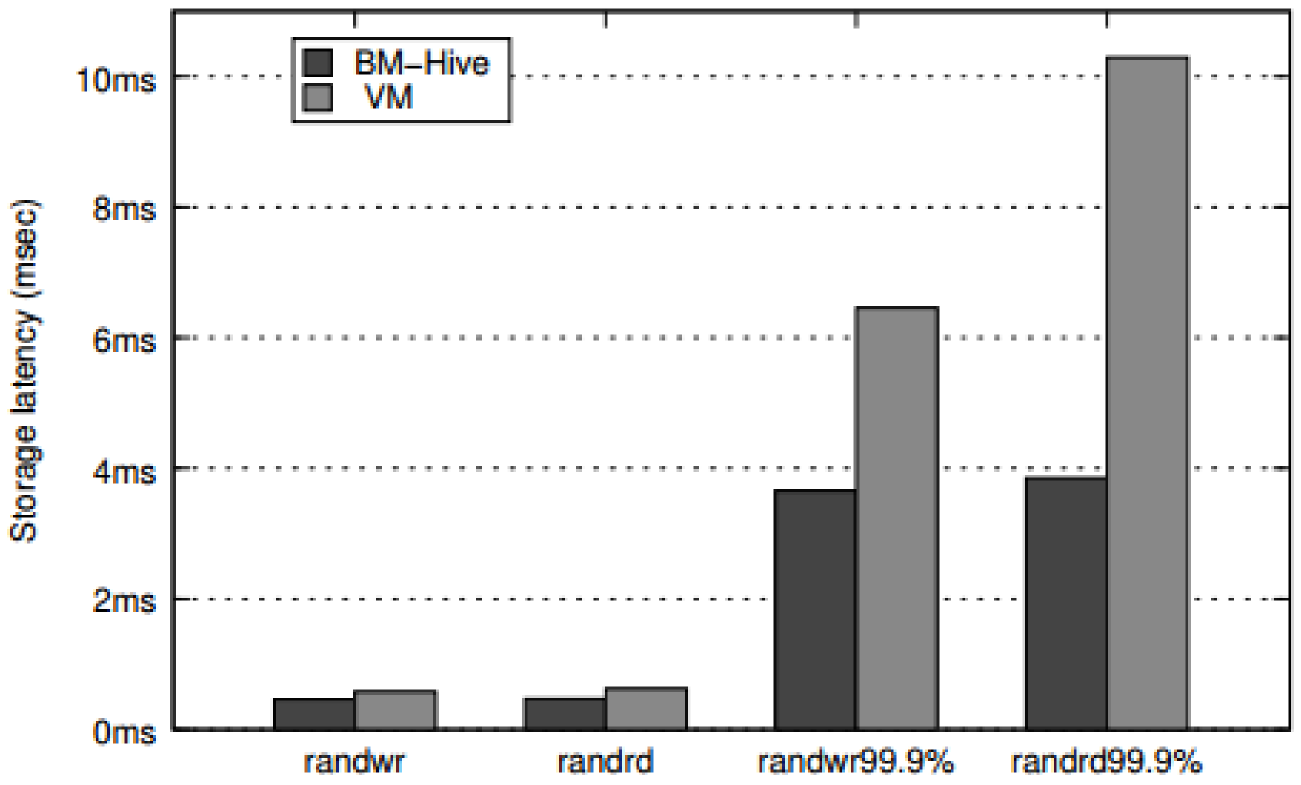


Figure 11. Storage I/O latency

Evaluation: Real business

Alibaba Cloud | Worldwide Cloud Services Partner

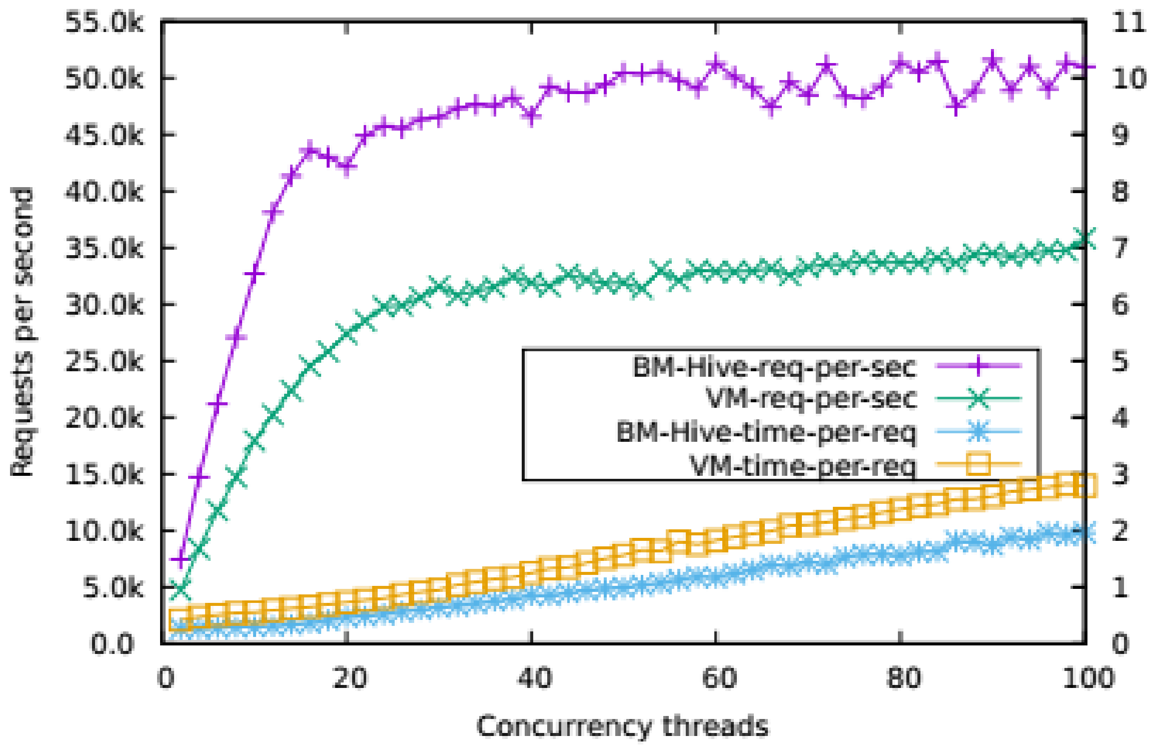


Figure 12. NGINX

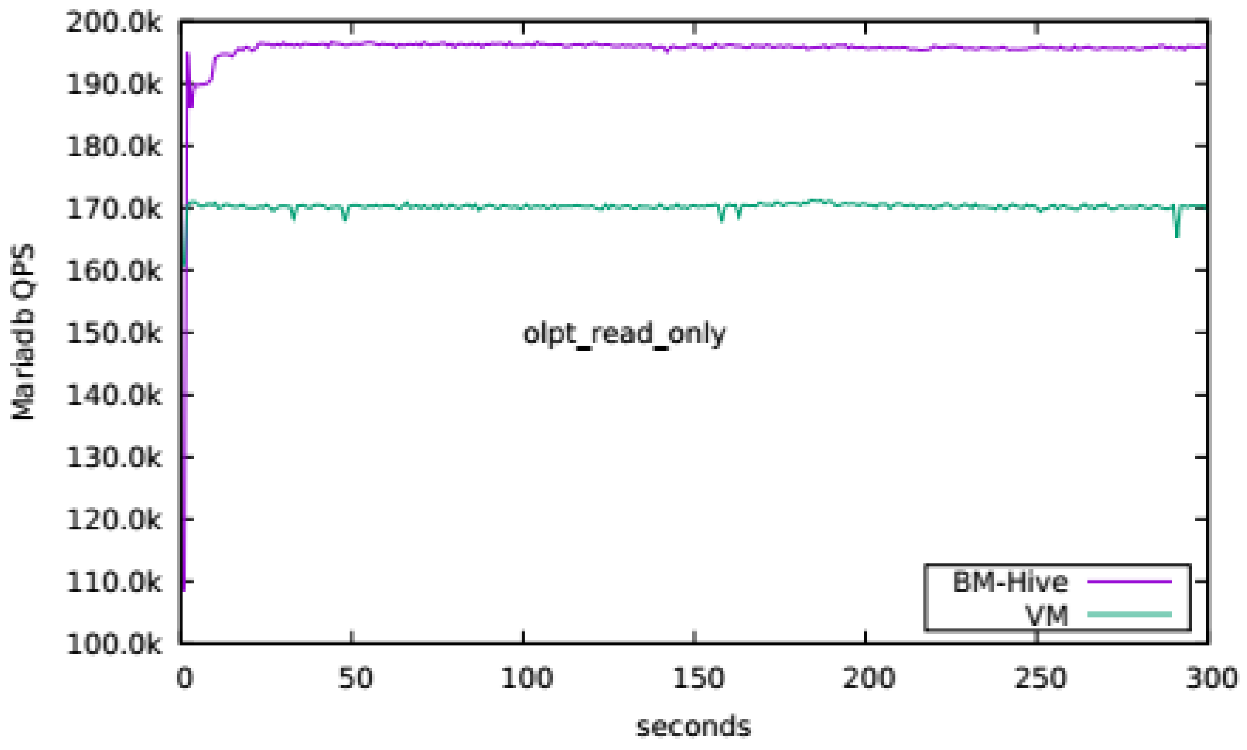


Figure 13. MariaDB ready-only

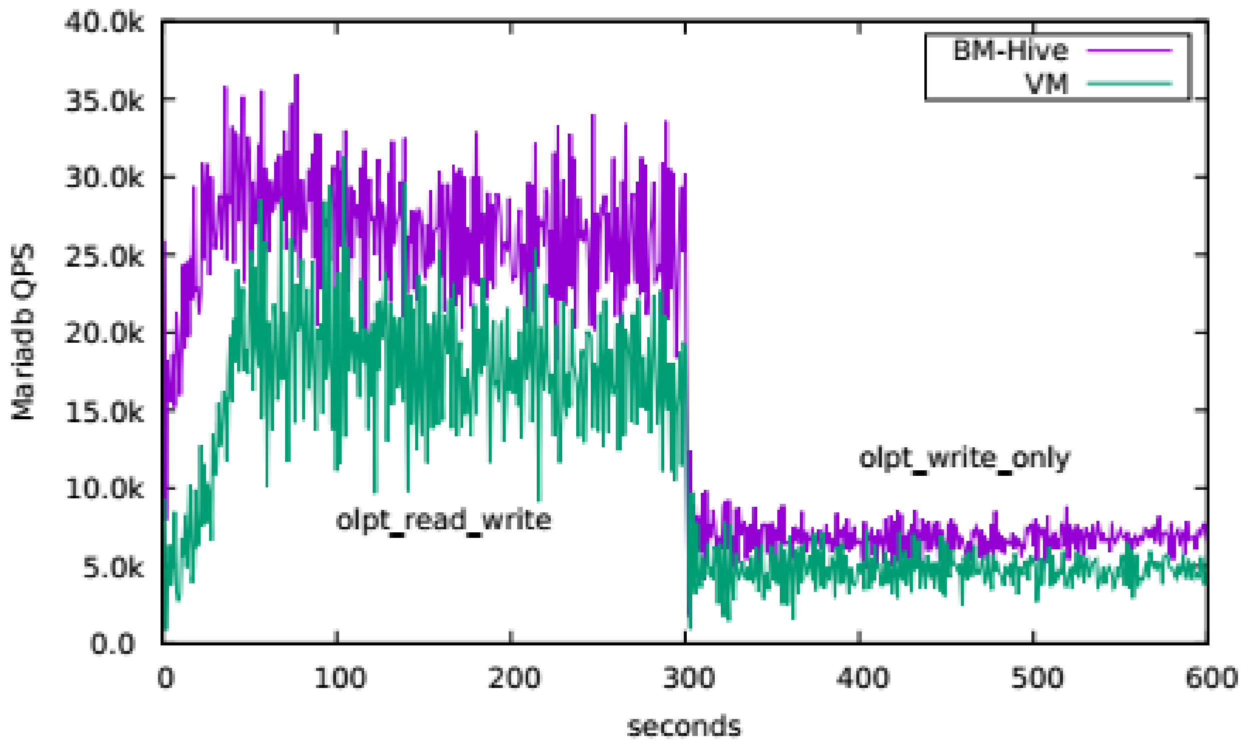


Figure 14. MariaDB rd/wr and wr-only

- Nginx
- MariaDB
- Redis

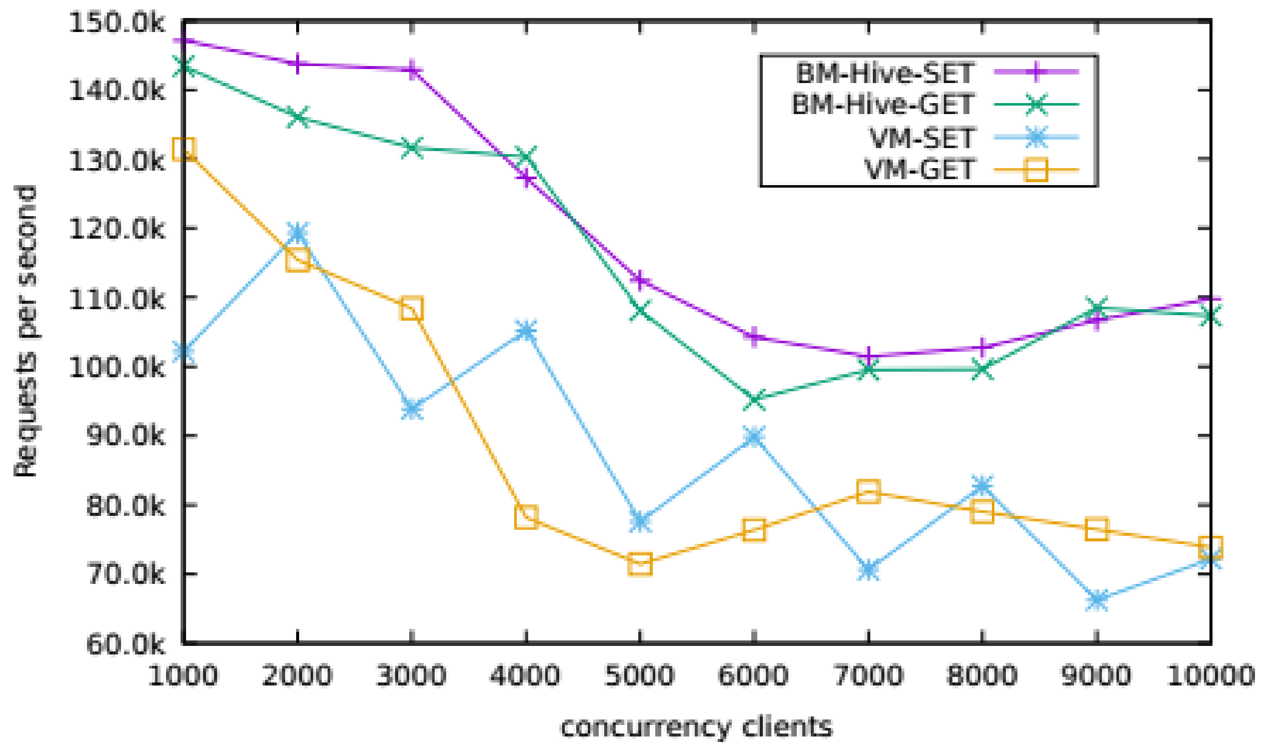


Figure 15. Redis with varying clients

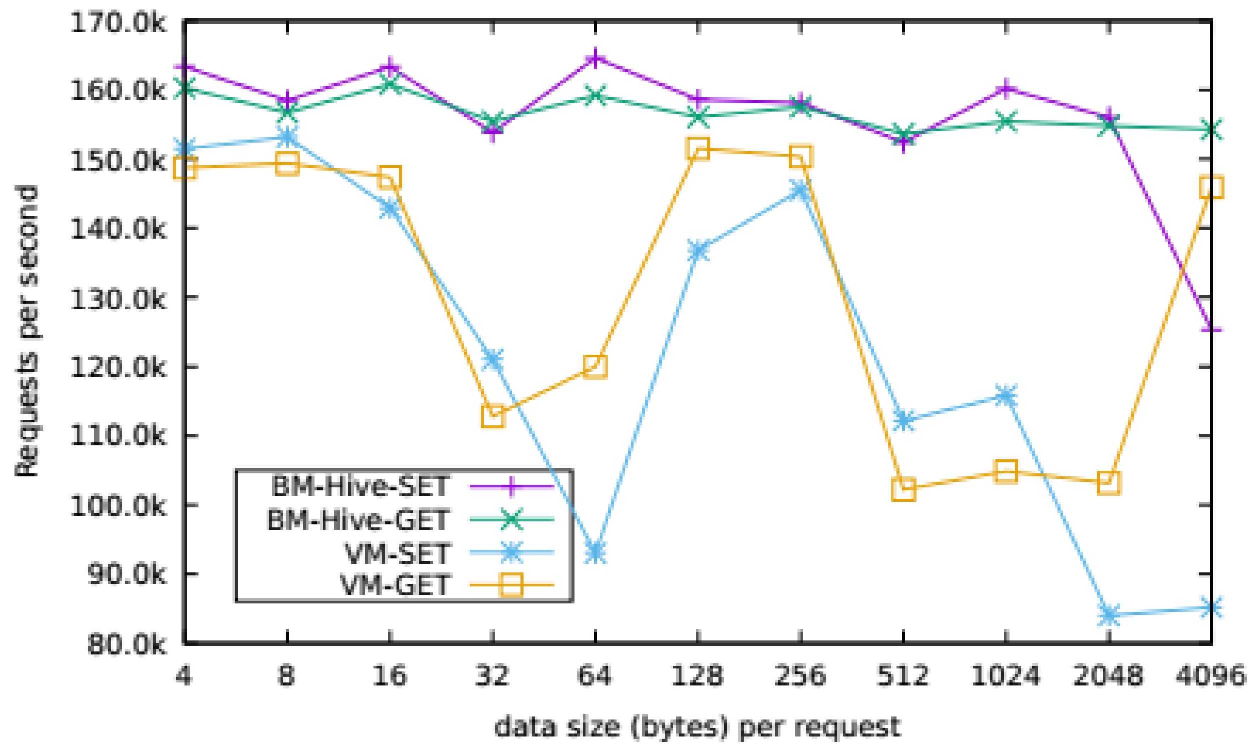


Figure 16. Redis with varying data size

X-Dragon BM guest performs substantially better than the virtualization-based cloud service for the popular applications used in the cloud

X-Dragon based Infrastructure Enhancement

Alibaba Cloud
Worldwide Cloud Services Partner

Alibaba Cloud

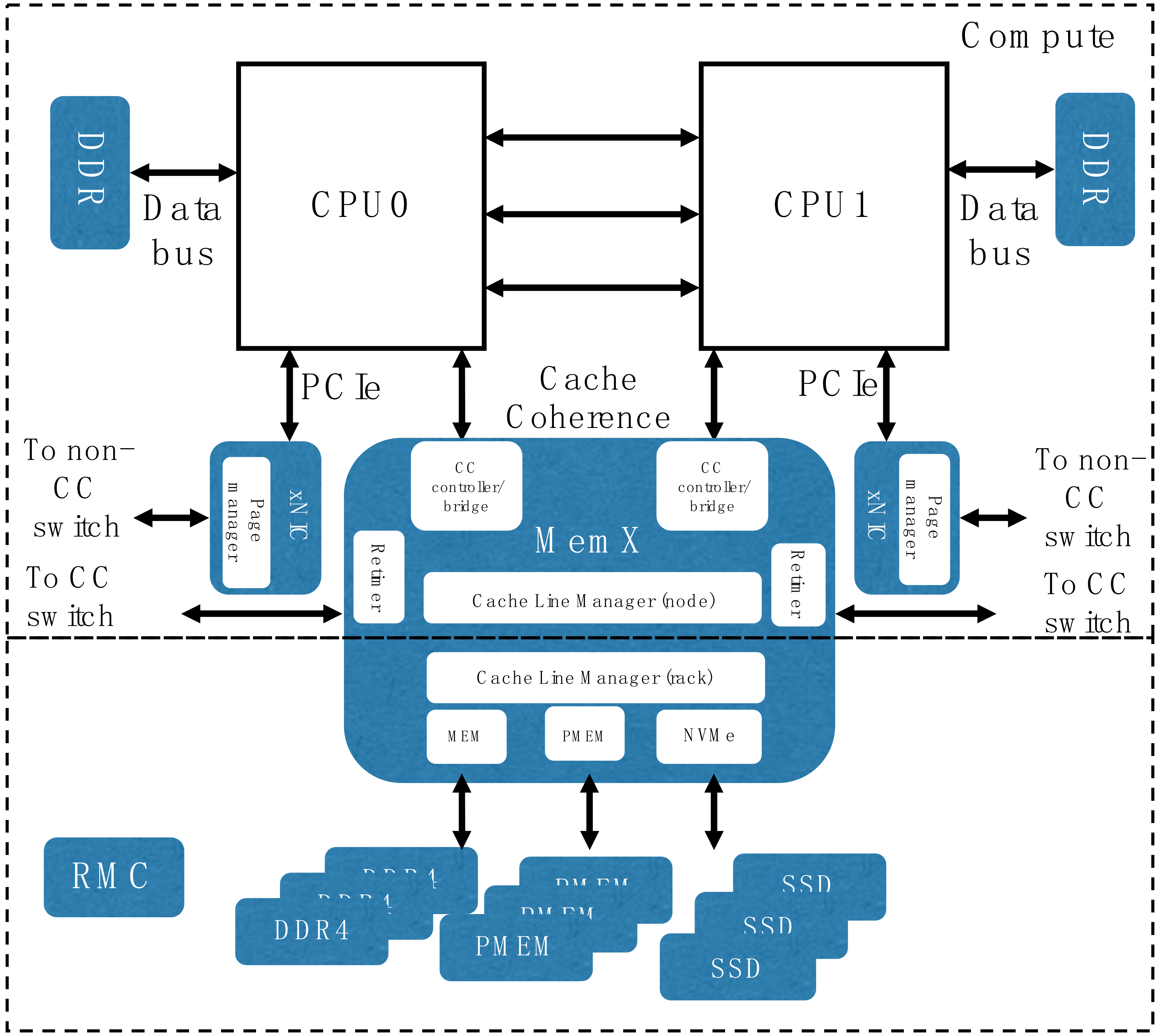
1 Memory Pool

2 Cloud App Aware Power Management

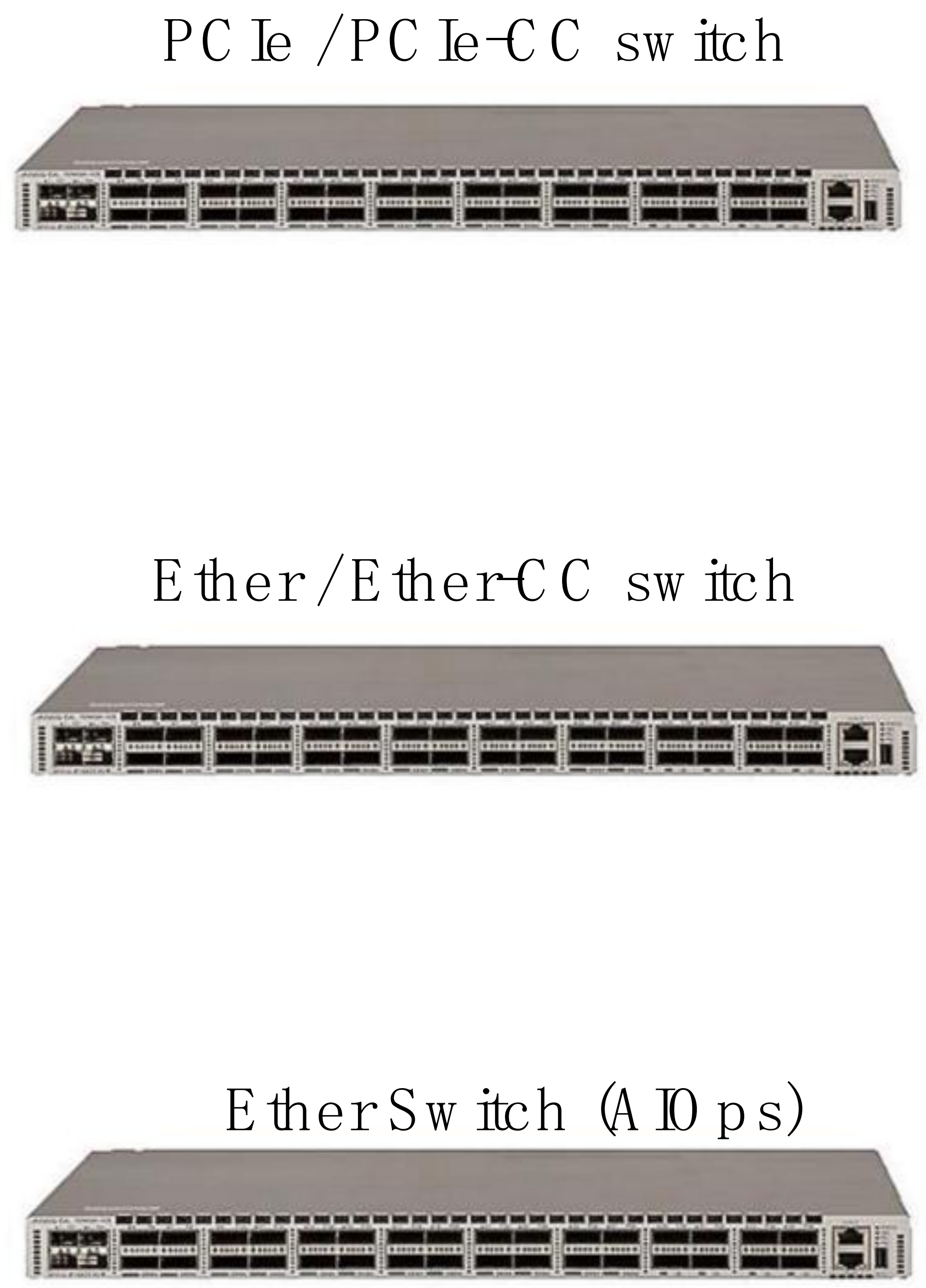
IaaS TCO optimization and new usage models enabling

Memory Pool

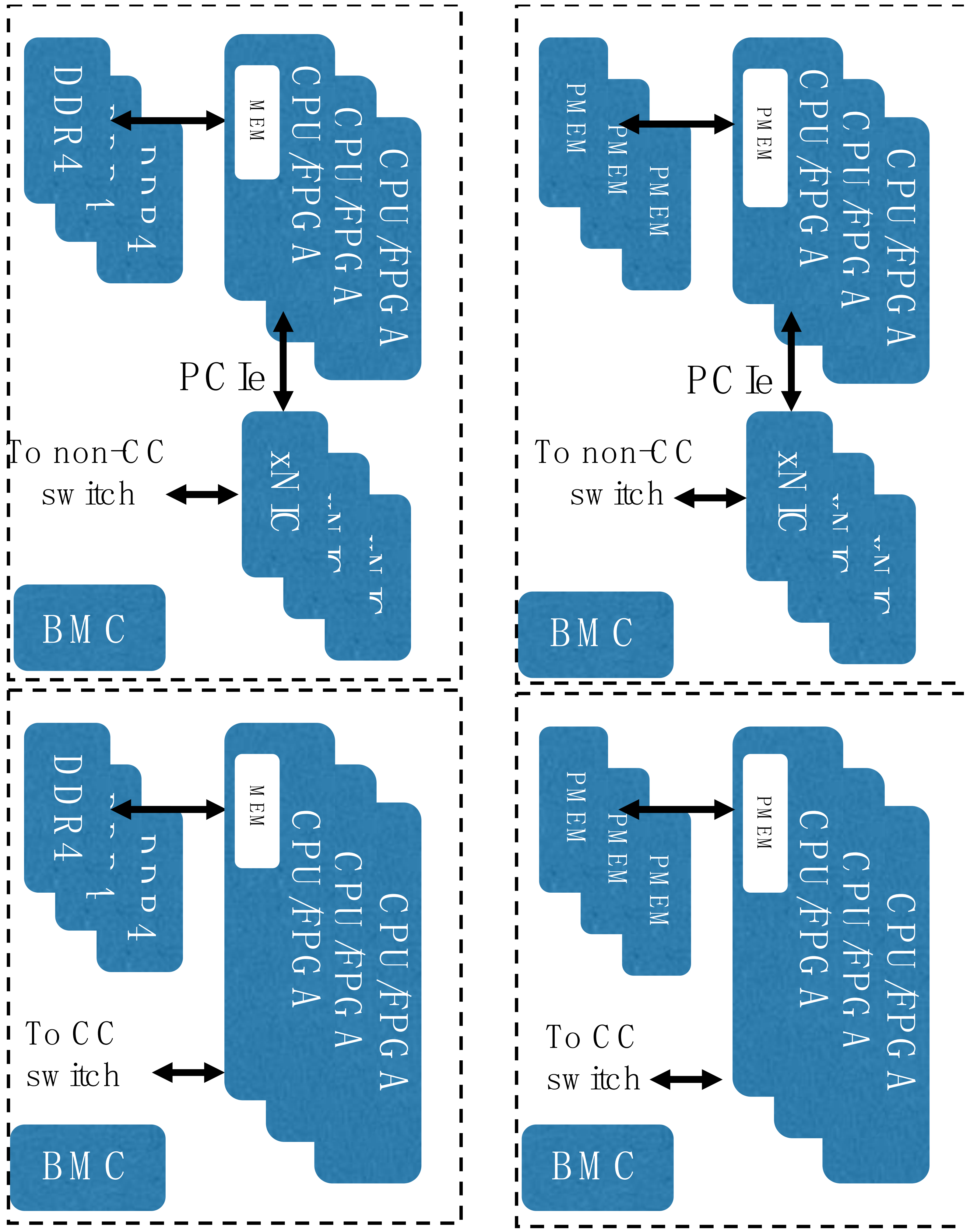
Alibaba Cloud
Worldwide Cloud Services Partner



Rack / Local Pool



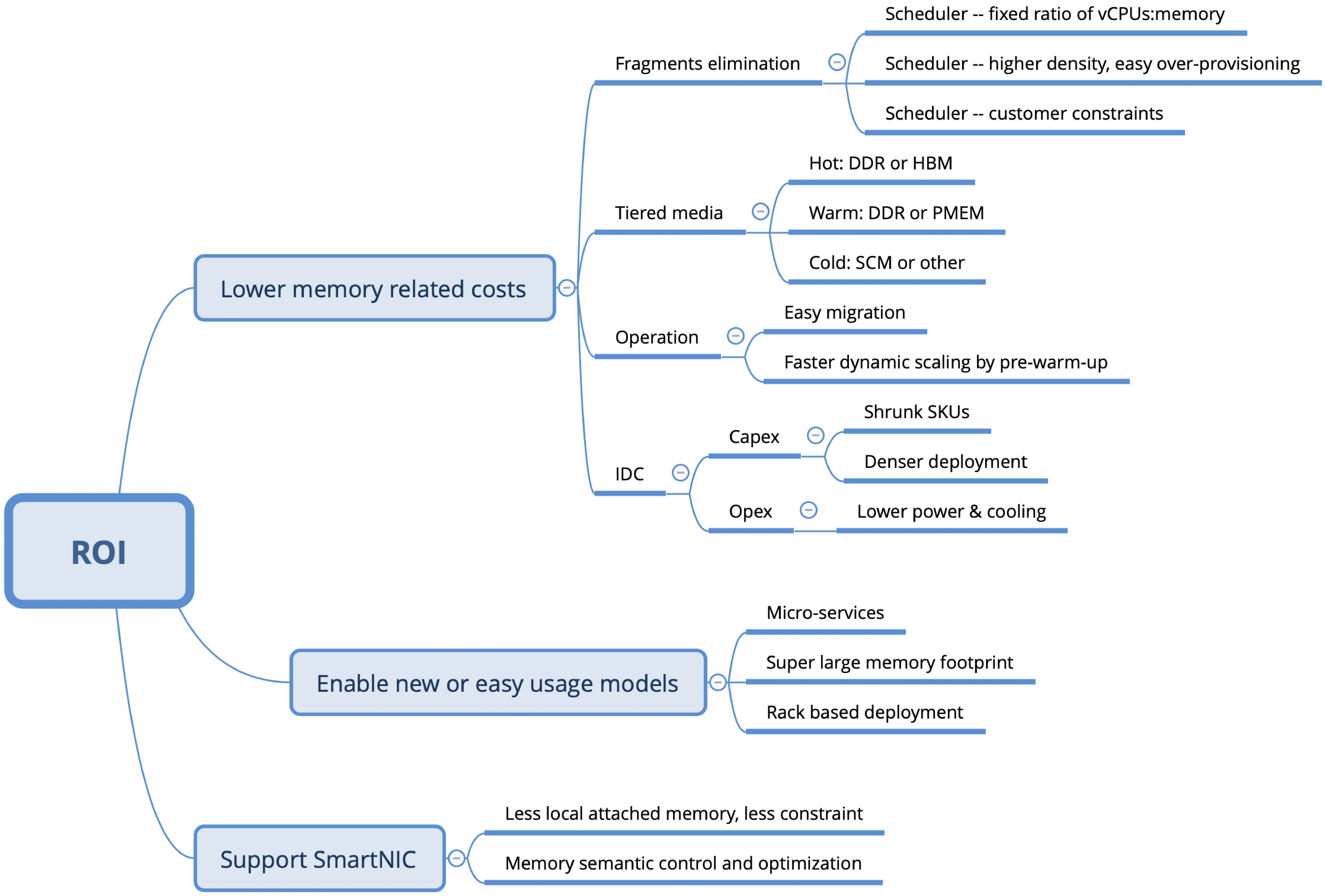
Switch Fabric



Remote Pool

ROI Analysis

Alibaba Cloud
Worldwide Cloud Services Partner



ROI

Lower memory related costs

- Fragments elimination
  - Scheduler -- fixed ratio of vCPUs:memory
  - Scheduler -- higher density, easy over-provisioning
  - Scheduler -- customer constraints
- Tiered media
  - Hot: DDR or HBM
  - Warm: DDR or PMEM
  - Cold: SCM or other
- Operation
  - Easy migration
  - Faster dynamic scaling by pre-warm-up
- IDC
  - Capex
    - Shrunk SKUs
    - Denser deployment
  - Opex
    - Lower power & cooling

Enable new or easy usage models

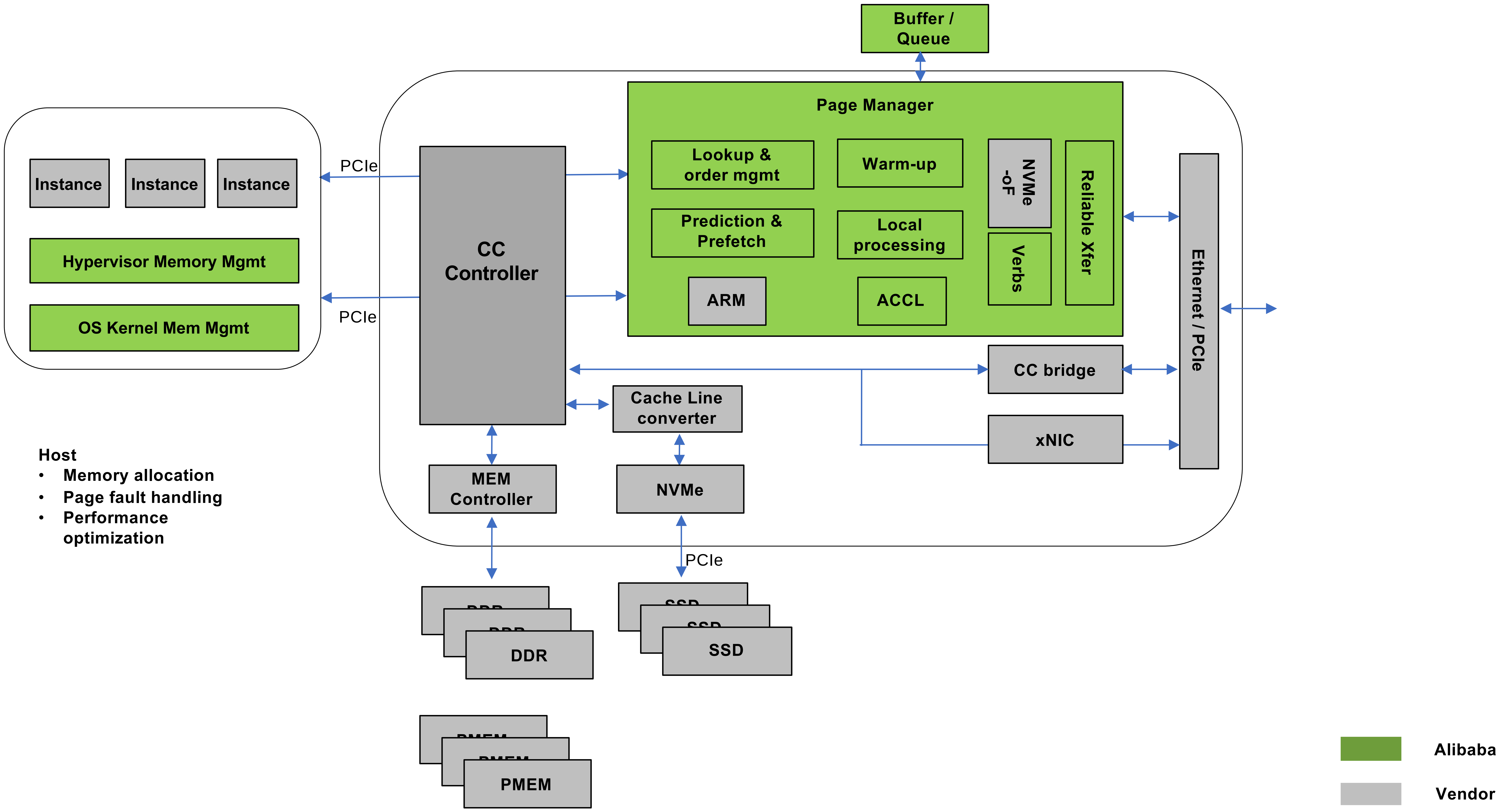
- Micro-services
- Super large memory footprint
- Rack based deployment

Support SmartNIC

- Less local attached memory, less constraint
- Memory semantic control and optimization

On Compute & Rack

Alibaba Cloud
Worldwide Cloud Services Partner



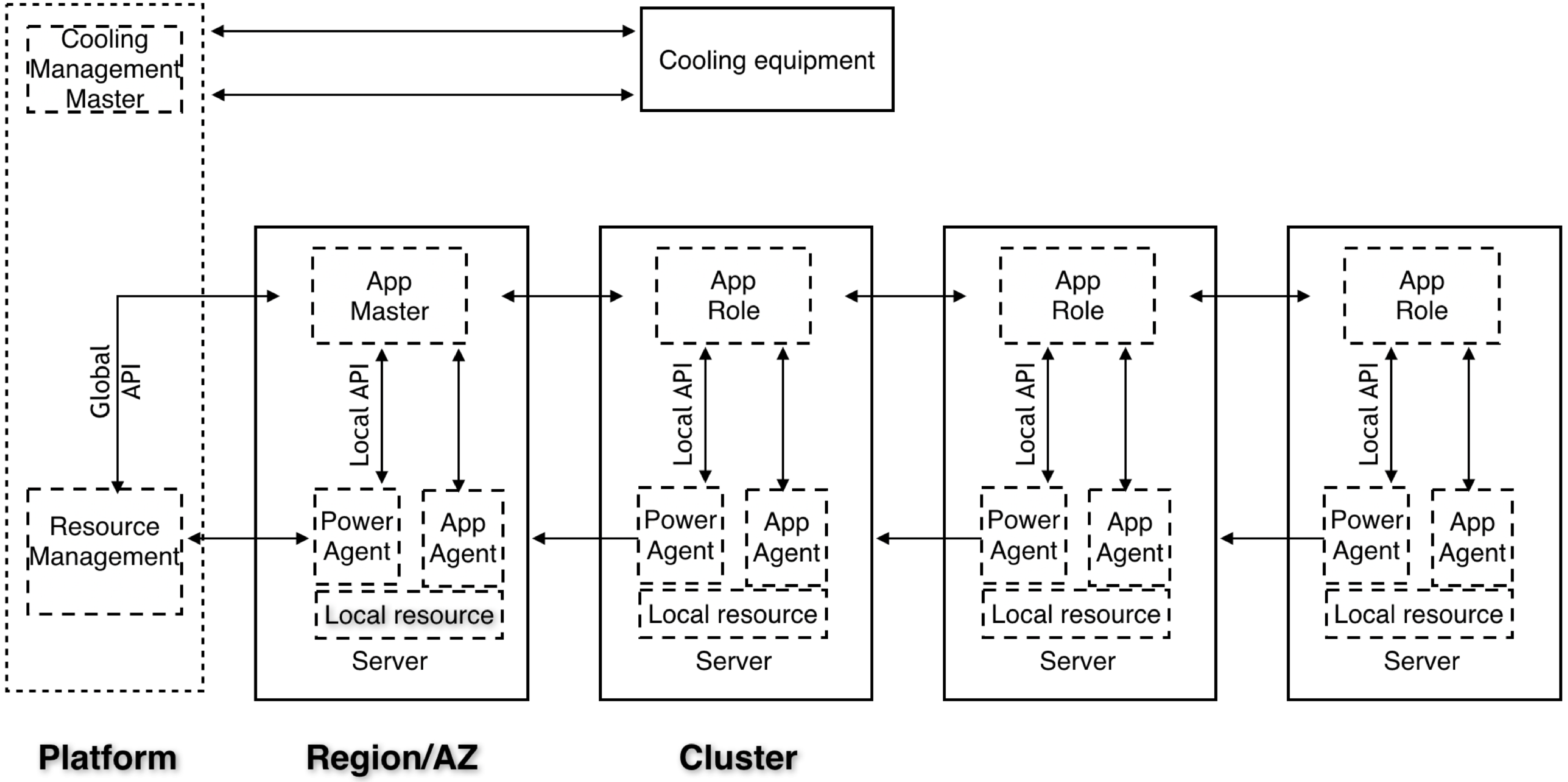
Host
- Memory allocation
- Page fault handling
- Performance optimization

Workloads & Potential Benefits

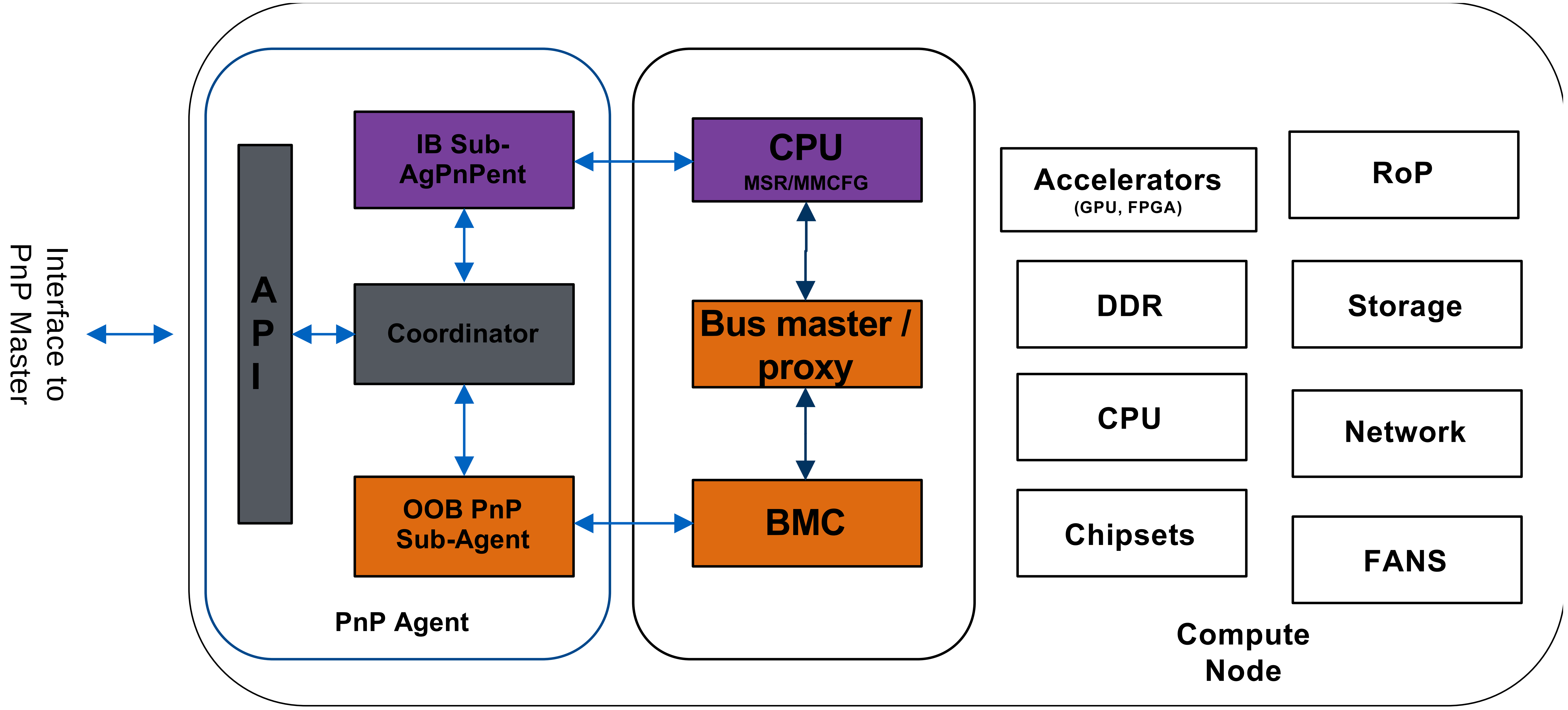
Type	Test	Potential Benefits
Traditional Compute	Mid to high utilization	Lower performance, higher density
Middleware	E-Commerce	Lower performance, higher density
Micro Services	E-Commerce	Lower performance, higher density
AI	Ali Native training & inference	Unacceptable for training
Encyption & Compression	Standard payload pre-/post-processing	Easier to scale out
Placement & Migration	Large instances	Faster; saving network b/w
Checkpointing & Mirroring	Cloud based HPC	High performance checkpointing enabled
NFV	Host gateway	Depends; easier to provision
Database	In-memory DB	Cost down significantly
Graph	Large social apps	Cost down significantly; minor programming model change
Upgrade & Deployment	Patching & initialization	Faster upgrade & composing

Power Management Platform

Alibaba Cloud
Worldwide Cloud Services Partner



Highly Available Management



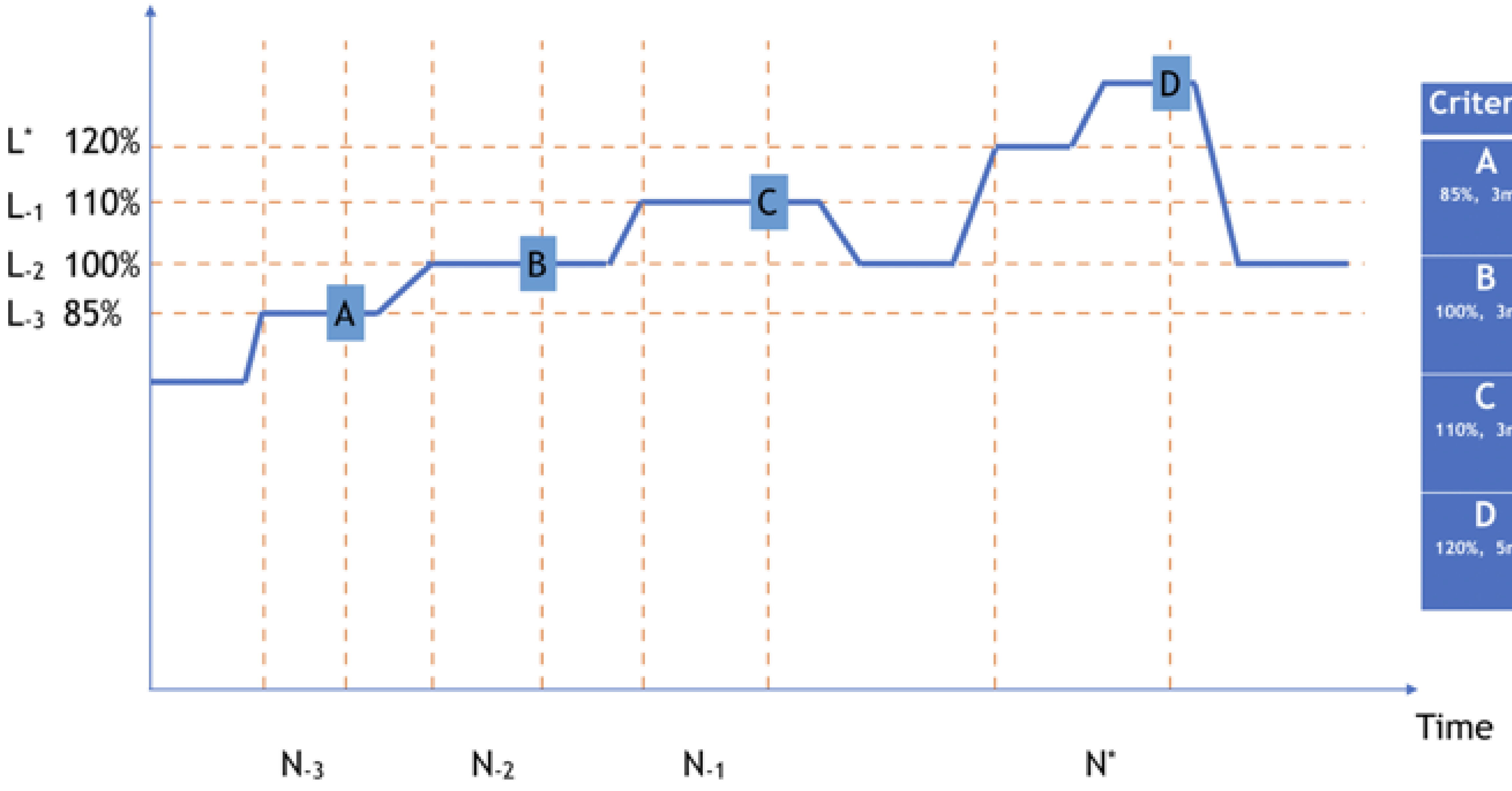
Alibaba Power Agent

- In-Band Power Management
- Out-of-Band Power Management

Server Platform

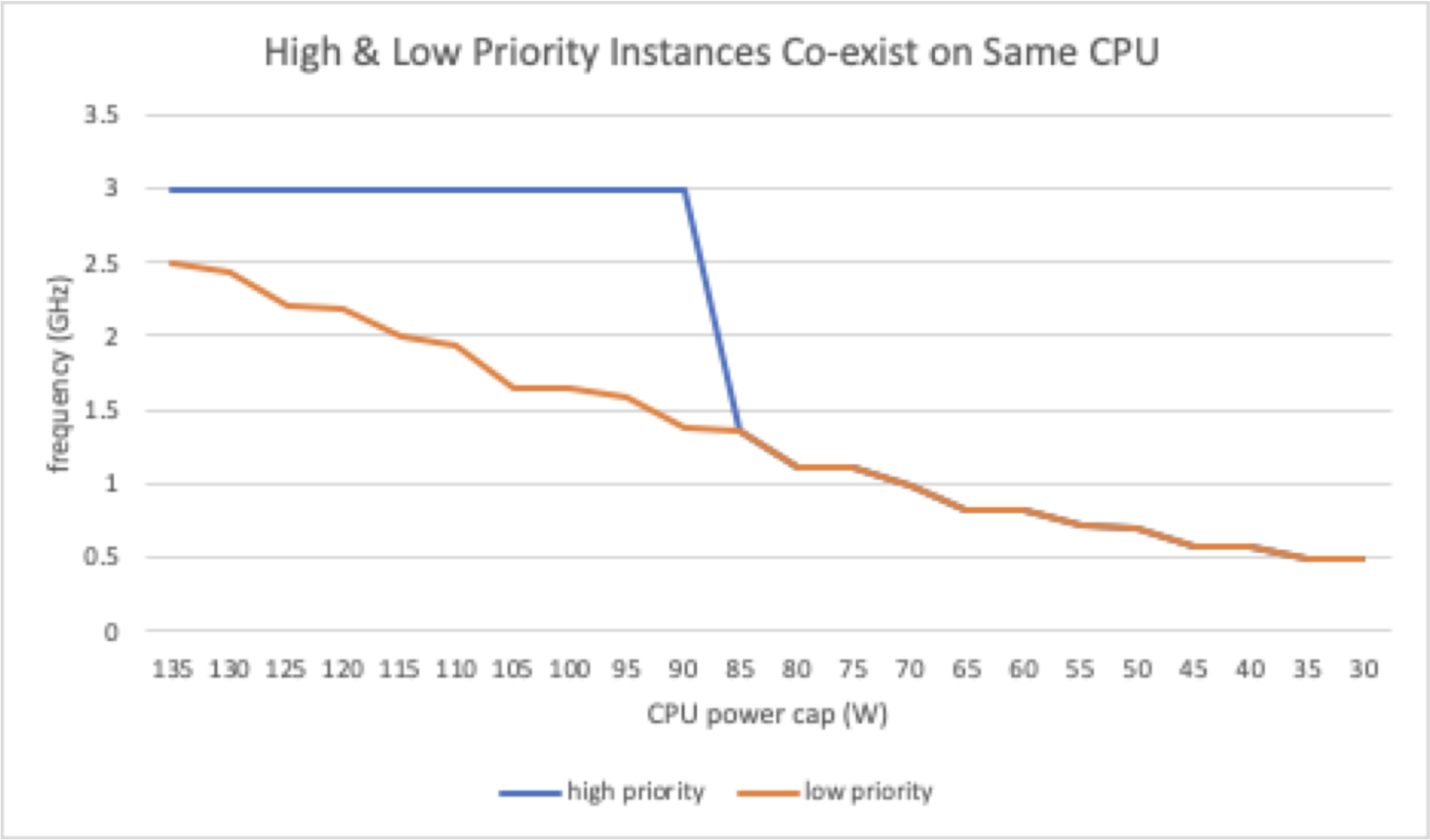
- Fine granularity power and performance telemetry & control knobs
- In-Band and Out-of-Band Control Channels

Capping & Budgeting



Criteria	Operation
A 85%, 3min	Rebalance
B 100%, 3min	Stop adding instances
C 110%, 3min	Super apps flow control
D 120%, 5min	Power capping with performance SLA

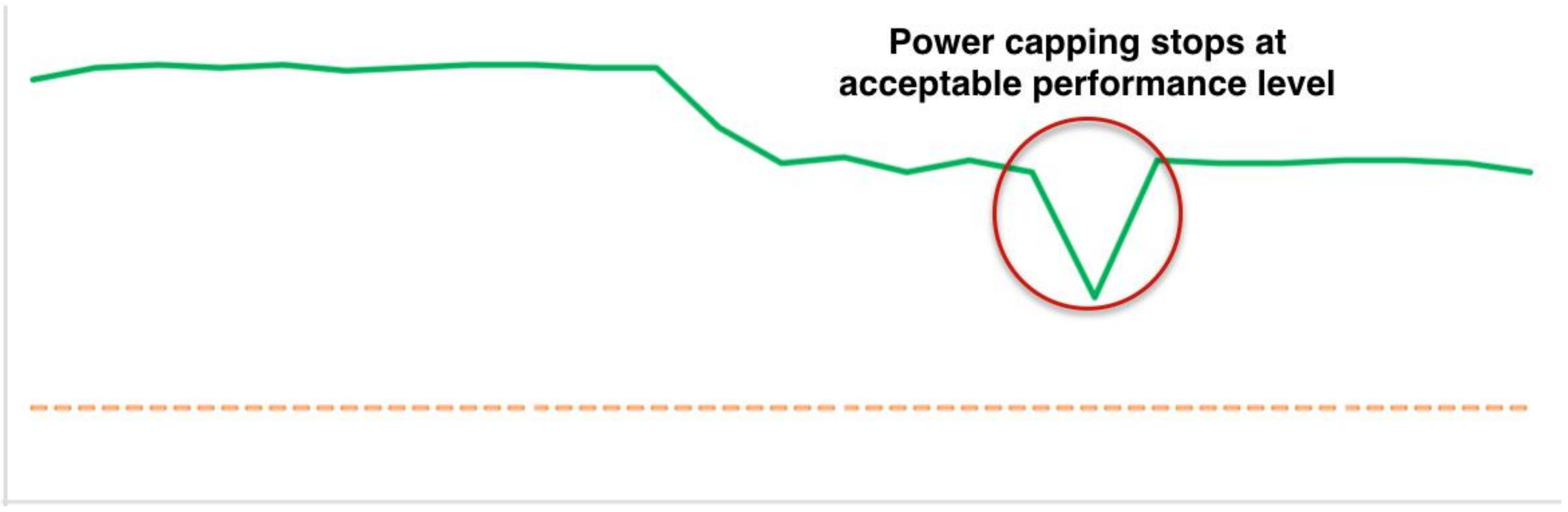
Rack/Node Power Capping



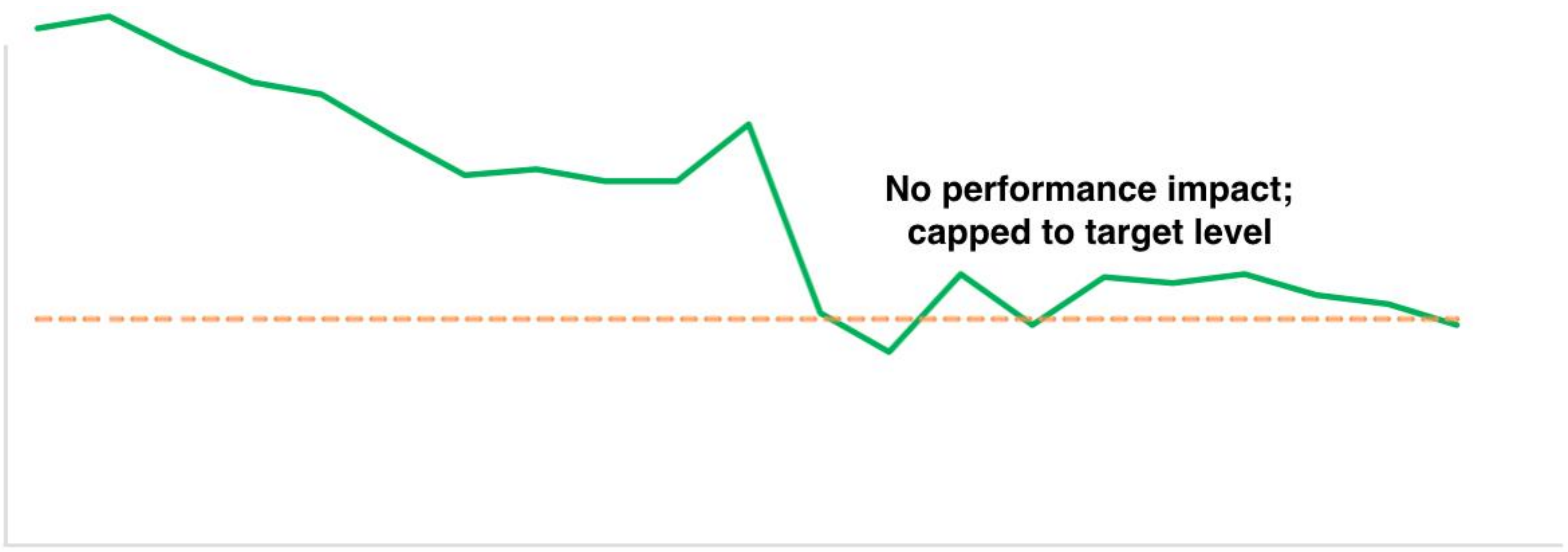
App Driven Power Budgeting

Performance Awareness

Type	Target	Note
Availability		Align w/ apps
Service delay	1s	Local
	30s	Global
Models coverage	Based on spec & test results	
Racks coverage	Based on spec & test results	
Power watermarks	Defined by apps & platform	
Capping accuracy	5%	
Priority	Defined by apps	Low priority nodes first capped
Fmin	Defined by apps Lifted by AI	Anytime higher than Fmin
Granularity	By core (CPU), rank (mem), link (IO) and device (storage)	
Capping - DVFS	Minimal performance impact	Defined by apps
Capping - CCx	**Minimal latency impact**	Defined by apps
In-Band	supported	
Out-of-Band	Partially supported	
Thermal watermarks	Defined by apps and platform	
Failover	Unconditional capping, autonomous capping, or S5	



Power capping stops at acceptable performance level



No performance impact; capped to target level

Identify IDC, server and app control knobs with least performance impact

Alibaba Cloud

Worldwide Cloud Services Partner

WWW.ALIBABA CLOUD.COM