OCP Use Case:
Using a Software Defined Storage on OCP Lightning (NVMe JBOF)

Yusuke Sato, Akira Kamio
Yahoo Japan Corporation
Yusuke Sato
Private Cloud,
Compute/Storage Architect

Akira Kamio
OpenStack,
Hypervisor(KVM/QEMU)
Introduction of Yahoo! JAPAN

- Yahoo! JAPAN is one of the largest internet company in Japan
- Our 100+ services earn 68billion PV per month
Datacenter

• 6+ DCs (1DC in the US)
• 75,000+ Baremetal Servers
• 130,000+ Virtual Machines
• 60PB+ Storages
Private Cloud

openstack™
kubernetes
OpenStack Environments

• 100+ OpenStack Clusters
  130,000+ instances
• Operated by less than 20 engineers
• For reducing operating costs
  • Automated provisioning
  • ChatOps
  • Centralized logging
  • Monitoring & Visualizing status
Storage Complexity

openstack

Cinder
Manila
Swift

openstack

Cinder
Manila
Swift

openstack

Cinder
Manila
Swift

kubernetes

Persistent Storage

Open. Together.
Software Defined Storage (SDS)

- Short Lead Time
- Flexible
- Scalable
Software Defined Storage (SDS)

SDS on commodity hardware

- Short Lead Time
- Flexible
- Scalable

Storage Appliance
Why We Choose OCP Lightning

• Flexible NVMe JBOF
• High-Density
  • Up to 30 NVMe SSDs in 2U
• High-Performance & Easy to Operate
  • Host can access NVMeSSDs directly via Mini SAS HD Cable
  • All SSDs are Hot-swappable

Why We Choose Quobyte

Object

Cinder

Manila

Quobyte

kubernetes

Unified Storage System
Challenges of SDS

- SDS is harder to operate than Storage Appliance
- SDS is a type of Distributed System
- Network is equivalent to a backplane
Network of Distributed System

• Network is one of the most important piece for SDS
• System Configuration
  • High-bandwidth network for East-West Traffic
• Monitoring
  • Error counter of NICs
  • All nodes ping each other

Pingmesh: A Large-Scale System for Data Center Network Latency Measurement and Analysis
Server Configurations

Storage Node
• CPU: Xeon E5-2683v4 x2
• MEM: 512GB
• Disk:
  • 240GB SATA SSD
  • JBOF (3.8TB NVMe x15)

Compute Node
• CPU: Xeon E5-2683v4 x2
• MEM: 512GB
• Disk: 240GB SATA SSD
Storage Server Configurations

Clos Network

PCIe Switch

NVMe x15

Host (Storage Node)

PCIe Gen3 x16: 31GBps (248Gbps) @bi-direction

25Gbps Ether x2: 50Gbps

JBOF (Lightning)

Open. Together.
Storage Server Configurations

Clos Network

Host (Storage Node)

PCIe Gen3 x16: 31GBps (248Gbps) @bi-direction

25Gbps Ether x2: 50Gbps

JBOF (Lightning)

PCIe Switch

NVMe x15
Storage Server Configurations

Clos Network

PCIe Switch

NVMe x15

JBOF (Lightning)

Host (Storage Node)

PCIe Gen3 x16: 31GBps (248Gbps) @bi-direction

25Gbps Ether x2: 50Gbps

Performance?
Performance Test: Environment

Storage Node
• CPU: Xeon E5-2683v4 x2
• MEM: 512GB
• Disk:
  • 240GB SATA SSD
  • JBOF (3.8TB NVMe x15)

Compute Node
• CPU: Xeon E5-2683v4 x2
• MEM: 512GB
• Disk: 240GB SATA SSD

Num of servers:
• 18 Storage Nodes
• 70 Compute Nodes

Benchmark Tool: fio
Benchmark Patterns:
• Sequential Read/Write
• Random Read/Write (4k/8k/16k/32k/64k/128k)
Performance: SDS Cluster

Peak Read IOPS: 1000k IOPS+
Peak Read BW: 40GB/s+
Performance: SDS Cluster

Peak Write IOPS: 300k IOPS+
Peak Write BW: 12GB/s+
Conclusions

• Yahoo! JAPAN is using Lightning as SDS backend
• Lightning is best for configuring SDS with OCP
• Better performance
Open. Together.

OCP Global Summit | March 14–15, 2019