P 1300 113 112 E info@Hyperscalers.com

HyperScalers Redhat Gluster appliance with RHEV-Manager

HyperScalers Pty Ltd.
Conducted at HyperScalers Proof of Concept (PoC) Lab 11th Jan 2018

HyperScalers is an Australian registered company.
ABN - 83 600 687 223
ACN - 600 687 223

Table of Contents

1. Executive Summary

The objective of this proof of concept is to create a Redhat Gluster software-defined-storage appliance. RedHat Gluster Storage is designed to handle the requirements of traditional file storage—high-capacity tasks like backup and archival as well as high-performance tasks of analytics and virtualization. Unlike traditional storage systems, Red Hat Gluster Storage isn't rigid and expensive. It easily scales across bare metal, virtual, container, and cloud deployments.

The appliance runs on virtual machines launched on Redhat's RHEV-M IaaS. RHEV-Manager is a self-hosted IaaS which executes on the same hosts as managed by that RHEV-Manager. The virtual machine is created as part of the host configuration, and the Manager is installed and configured in parallel to the host configuration process.

The appliance creates a storage volume using two or four peer nodes. It verifies the design on QCT OCP rack with high speed storage JBOD and 40Gbps data path.

2. Introduction

The hardware infrastructure for this PoC consists of Rackgox X300 rack. It is a compute-intensive solution designed for the most computing-intensive applications. With a total rack power cap of 25K watts, it can install up to 16 units of F06A servers that are designed for optimized performance and space. QCT's X300 features up to 64 independent 2-socket half width servers that are capable of running complex workloads using highly scalable memory, I/O capacity and fibre network options.

The rack is populated with compute, storage and network components as mentioned in the table shown above. It uses eight compute nodes, two JBRs and one apiece leaf and spine switches.

2.1 Rackgox F06A compute node

The rackgox F06A is designed for the highest compute density with four nodes in a 2 OU space. Each node can install up to two SATADOMs for the operating system and up to four extra hot-swappable SSD/HDDs for cache or data storage. Its RAID-ready configuration preserves data integrity and avoids data corruption.

2.2 2U JBR storage box

The JBR is based on hidden-shelf chassis design to fit 28x 3.5 inch hard disks in a 2 OU space.

2.3 T5032 -LY6 network switch

The QuantaMesh T5032-LY6 is a high performance and low latency layer 2/3/4 Ethernet switch with 32 40GbE QSFP+ ports in a 1U form factor.

2.4 T1048-LY4 network switch

The QuantaMesh T1048-LY4 family is the new generation of layer 2 and layer 4 Ethernet standalone switches that provide 48x10/100/1000Base-T downlink plus 2 1/10GBase-X SFP+ uplink ports.

2.5 Data path design over network switches

The network switches are specially configured to support 40G data path for all Rackgox nodes.

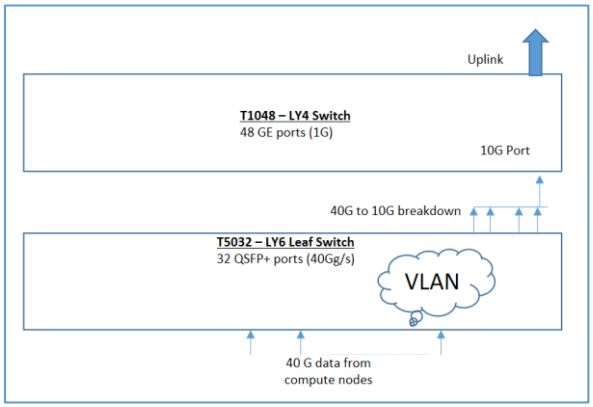


Figure 1: Network configuration for 40G data path

The nodes are equipped with 40G OCP mezzanine ports, which act as the data path. The ports are connected to one port on the LY6 switch as depicted in the diagram. The 40G ports in LY6 switch are configured with common VLAN ID, so that all packets in RedHat configured nodes are accessible to each other. The uplink port from LY6 switch is in trunk mode with the 10G port of LY4 switch. The breakout cable "Octopus" splits 40G ports in 4 10G ports. The 10G port in LY4 switch has DHCP configured; which feed dynamic IP address to all nodes connected to leaf switch. The uplink port of LY4 switch goes to external router for secured internet connectivity. The management ports of all nodes are directly connected to the GE port of LY4 switch. The data and management ports are segregated with dedicated subnets and VLAN IDs.

3. Test Environment

The test environment consists of following hardware and software components:

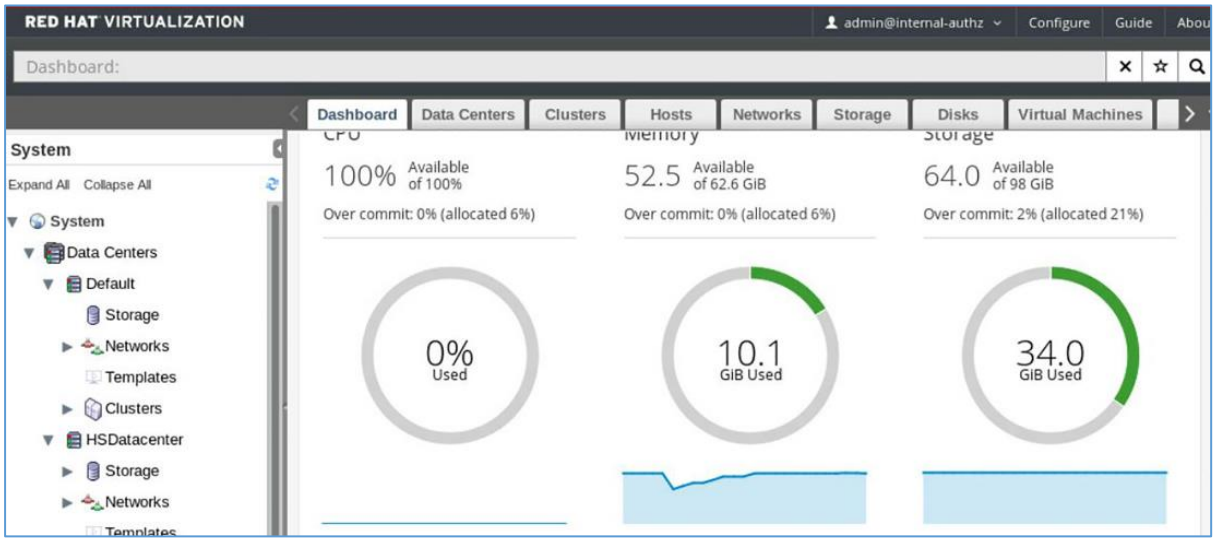
Hardware	Quanta QuantaPlex T21P-4U converged server • E5-2603 CPU • 128 (32x4) GB DDR4 RAM • 2x120GB SATA SSD • 20x8TB 3.5" HDD • LSI SAS 3008 HBA
Software	• Redhat Virtualization Manager 4.1 • Redhat Gluster 3.3

4. Software Appliances

Following sections describe software layers for appliance installation.

4.1 RHEV Manager

A self-hosted engine is a virtualized environment in which the Red Hat Virtualization Manager, or engine, runs on a virtual machine on the hosts managed by that Manager. The virtual machine is created as part of the host configuration, and the Manager is installed and configured in parallel to the host configuration process.



The RHEV-M provides a dashboard of resources, on which appliance can create virtual machines to act as peer nodes of Gluster storage. The RHEL installation is done on the local SATA SSD while the storage pool is connected through JBOD. A pool of 14 hard disks each 4TB are used for the appliance. It creates virtual network bridge running on 40Gbps uplink data path though QCT OCP PCIe card.

4.2 Redhat Gluster

The RHEV-M hosts multiple virtual machines as the peers to create a GlusterFS pool.

```
Number of Peers: 1

Hostname: server2
Uuid: ab9107ec-d81c-44c7-9ea4-b82d0c07dd01
State: Peer in Cluster (Connected)
```

```
Number of Peers: 1

Hostname: server1
Uuid: c612b555-01c4-42a9-bb70-f0d0f307f7d0
State: Peer in Cluster (Connected)
```

Once the pools are created and peers start communicating; the client machines is connected to the pool and acts as the NFS loaded GlusterFS file system. In the appliance, there are bricks created on all virtual machines and they are configured as a replicated pool. That enabled the GlusterFS t sync the data in all bricks every-time any of the brick is edited.

```
[root@server1 gv0]# ls -l
total 1064176
-rw-------. 2 root root     157989 Jan  9 23:37 copy-test-001
-rw-------. 2 root root     157989 Jan  9 23:37 copy-test-002
-rw-------. 2 root root     157989 Jan  9 23:37 copy-test-003
```

```
[root@server2 gv0]# ls -l
total 15600
-rw-------. 2 root root 157989 Jan  9 23:37 copy-test-001
-rw-------. 2 root root 157989 Jan  9 23:37 copy-test-002
-rw-------. 2 root root 157989 Jan  9 23:37 copy-test-003
```

In the screen dump, there are bricks of two virtual machines highlighted. The client load NFS and edits it mounted file space; the edited changes are replicated in all bricks. It's observed that there are same data in both bricks; and they are replicated.

```
[root@server1 gv0]#  dd if=/dev/zero of=test1 bs=1G count=1 oflag=dsync          1+0 records in
1+0 records out
1073741824 bytes (1.1 GB) copied, 17.4624 s, 161.5 MB/s
[root@server1 gv0]#
```

The disk performance tests are executed on the client system which shows around 162MB/s. The performance expected is around 250MB/s; so as a next step, the compute and storage path would be changed with extra cores and disk speed.

5. Accessibility

The appliance can be accessible to the customers using WAP DDNS "http://hyperscalers.asuscomm.com/". Depending on the customer requirements; the administrator can open a port accessible via DDNS VPN.

6. Conclusion

The appliance shows that Redhat GlusterFS solution can be efficiently designed as appliance on QCT open racks. The performance on a disaggregated architecture is benchmarked would be improved with better CPU cores and storage drives. The RHEV-M is utilized as IaaS for the appliance efficiently and it hosts a self-managing web interface; it's a convenient interface to create a virtualized compute, network and storage environment.