P 1300 113 112 E info@Hyperscalers.com

HyperScalers-RedHat Cloud Appliance

HyperScalers Pty Ltd.
Conducted at HyperScalers Proof of Concept (PoC) Lab
3rd Apr 2017

Table of Contents

1. Executive Summary

The objective of this proof of concept is to execute RedHat IaaS, PaaS and SDS (Software-defined Storage) solutions on QCT open rack platforms. The experiment utilizes OCP complaint hardware solutions from QCT called as rackgox to design an appliance running RedHat software solutions like OpenStack, OpenShift and CEPH.

Rackgox is Quanta's rack system inspired by the OCP standards. Equipped with abundant innovative features, it is a perfect solution not only for the hyper-scale datacentres, but also for the usage in the public & private cloud environments. The rackgox series includes hardware offerings like rack systems, compute nodes, storage nodes, network switches and mother boards. RedHat is pioneer in providing open software solutions for cloud computing based appliances. Red Hat provides storage, operating system platforms, middleware, applications, management products, and support, training, and consulting services. The PoC is done in HyperScalers research laboratory to design a cloud appliance with:

- Open hardware and software architecture, fully compliance with OCP initiatives
- Integrating RedHat cloud solutions on QCT rackgox platform and benchmarking the performances
- Defining process to design an open IaaS, PaaS and SDS appliance; in a private cloud environment

2. Infrastructure Design

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.

10 Gb/s Spine / Management
40 Gb/s Leaf
Compute | Compute
Compute | Compute
JBR
JBR
Compute | Compute
Compute | Compute

Compute	**RackgoX – F06A** 4 Nodes / chassis 2xIntel 2698 v3 CPU 2 x 32G Dimm 2x240G SATA ssd LSI 3008 ISCSI Mellanox OCP 40G Network Adapter
Storage	**2U JBR (JBOD)** Lock-in Mini-SAS Module Front Load Screw-less HDD Trays Support up to 28 x 3.5 SAS HGST drives
Network	**T5032 – LY6 Leaf Switch** 32 QSFP+ ports (40Gg/s) **T1048 – LY4 Switch** 48 GE ports (1G) 2 1/10G SFP+ ports

Figure 1 : Rackgox hardware components for RedHat cloud appliance

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.

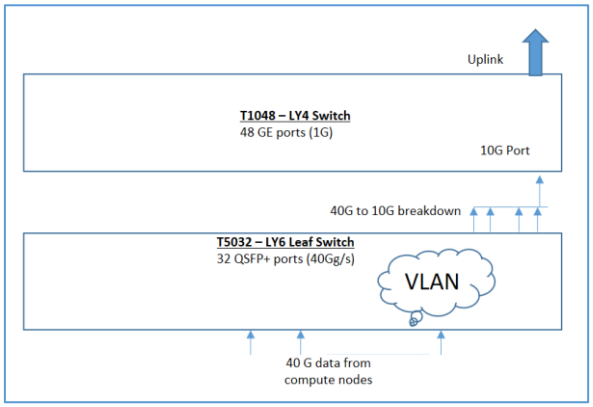

Figure 2: Network configuration for 40G data path

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. RedHat Appliances

This section describes three appliances designed using RedHat cloud software OpenStack, CEPH and OpenShift.

3.1 RedHat OpenStack IaaS

The installation uses RedHat packstack on one dedicated node of Rackgox. It utilizes the internal SSD as storage and RHEL7.3 as the base operating system. RedHat packstack is a utility to deploy various parts of OpenStack on multiple RHEL pre-installed servers over SSH automatically or manually. The PoC installs OpenStack 10 using packstack in a non-interactive method. The packstack command is provided with configuration options via a text file, referred to as an answer file, instead of via standard input.

- Use packstack to generate a default answer file.
- Edit the answer file inserting desired configuration values.
- Execute the packstack command providing the completed answer file as a command line argument.
- Packstack will then attempt to complete the deployment using the configuration options provided in the answer file.

The edited answer file shown below is used as the configuration for installations in PoC.

```
[general]
CONFIG_SSH_KEY=/root/.ssh/id_rsa.pub
# Default password to be used everywhere (overridden by passwords set
# for individual services or users).
CONFIG_DEFAULT_PASSWORD=*****

# The amount of service workers/threads to use for each service.
# Useful to tweak when you have memory constraints. Defaults to the
# amount of cores on the system.
CONFIG_SERVICE_WORKERS=%{::processorcount}

# Specify 'y' to install MariaDB. ['y', 'n']
CONFIG_MARIADB_INSTALL=y

# Specify 'y' to install OpenStack Image Service (glance). ['y', 'n']
CONFIG_GLANCE_INSTALL=y

# Specify 'y' to install OpenStack Block Storage (cinder). ['y', 'n']
CONFIG_CINDER_INSTALL=y

# Specify 'y' to install OpenStack Shared File System (manila). ['y',
# 'n']
```

After the installations, the host provides OpenStack webui and interface to be launched.

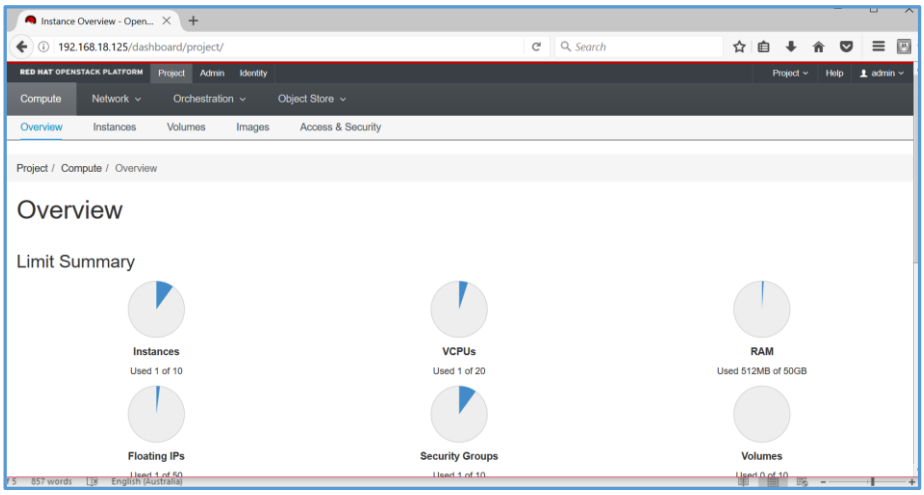

Figure 3: RedHat Openstack portal

The webui can be used to launch the virtual machine, containers and other IaaS services.

horizon	Web browser-based dashboard that you use to manage OpenStack services.
keystone	Centralized service for authentication and authorization of OpenStack services and for managing users, projects, and roles.
neutron	Provides connectivity between the interfaces of OpenStack services.
cinder	Manages persistent block storage volumes for virtual machines.
Nova	Manages and provisions virtual machines running on hypervisor nodes.
swift	Allows users to store and retrieve files and arbitrary data.

The Rackgox node has 40G OCP port connected to the T5032 switch for data path. The internal SATA SSD is used as the storage for launching services and once ceph is installed in other nodes, it would be integrated with the openstack setup for complete storage solutions.

3.2 RedHat CEPH storage

RedHat Ceph Storage is a scalable, open, software-defined storage platform that combines the most stable version of the Ceph storage system with a Ceph management platform, deployment utilities, and support services. Red Hat Ceph Storage is designed for cloud infrastructure and web-scale object storage. Ceph is designed to run on commodity hardware, which makes building and maintaining petabyteto- exabyte scale data clusters economically feasible.

The PoC installs Ceph 1.3; it requires minimum 5 nodes and 2 JBRs for complete installations. All the nodes involved in ceph installation uses RHEL 7.3 OS.

3.2.1 Ceph Calamari

Ceph calamari is a management and monitoring system for Ceph storage cluster. It provides a dashboard user interface that makes Ceph cluster monitoring simple and handy. Calamari was initially a part of Inktank's Ceph Enterprise product offering and it has been open sourced by Red Hat. One node of rackgox would be dedicated for calamari and would utilize the internal SSD as storage.

3.2.2 Ceph Monitor

The Ceph monitor is a data store for the health of the entire cluster, and contains the cluster log. RedHat strongly recommends using at least three monitors for a cluster quorum in production; though for the PoC purpose one rackgox node is used. Monitor nodes typically have fairly modest CPU and memory requirements. Because logs are stored on local disk(s) on the monitor node, it is important to make sure that sufficient disk space is provisioned. The node uses internal SSD for storing data.

3.2.3 Ceph OSD

ceph-osd is the object storage daemon for the Ceph distributed file system. It is responsible for storing objects on a local file system and providing access to them over the network. The PoC dedicates 3 rackgox nodes for OSD and connect 2 JBRs as storage pool. Each JBR consists of 2 sets of 14 HDDs; hence altogether 3 x14 drives are dedicated for the OSD storage purposes. The setup ensures that network interface, controllers and drive throughput don't leave any bottlenecks— e.g., fast drives, but networks too slow to accommodate them. The datapath used is 40G and JBR drives are connected through high speed LSI iSCSI interface.

3.2.4 Ceph installation architecture

RedHat ceph uses five nodes and 3 channels for JBR for installations. The calamari and monitor uses 1 node each and OSD uses 3 nodes for installations.

<table>
<tr><td colspan="2">10 Gb/s Spine / Management</td></tr>
<tr><td colspan="2">40 Gb/s Leaf</td></tr>
<tr><td colspan="2">Calamari Node</td></tr>
<tr><td></td><td>JBR 1 (14 Drives)</td></tr>
<tr><td>JBR 2 (14 Drives)</td><td>JBR 3 (14 Drives)</td></tr>
<tr><td>Monitor Node</td><td>OSD1 Node</td></tr>
<tr><td>OSD2 Node</td><td>OSD3 Node</td></tr>
</table>

Figure 4: RedHat CEPH hardware components

Ceph relies on packages in the Red Hat Enterprise Linux 7 Base content set. Each Ceph node must be able to access the full Red Hat Enterprise Linux 7 Base content. To do so, ceph nodes are connected to the Internet to the Red Hat Content Delivery Network (CDN) and registered with the redhat

customer portal. The HyperScalers laboratory partners with RedHat repository portal and registered all nodes with relevant ceph subscriptions.

After complete installations, the calamari provides an interface to configure the storage utilizing ceph drives.

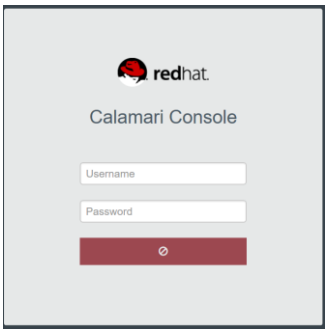

Figure 5: Ceph monitoring interface

The ceph CLI provides commands to create cluster using SAS drives. As part of PoC 3 OSD servers are installed, each utilizing 5 HDD from three channels of JBR (1,2,3). The architecture followed is

- Ocprack is rackname on which jbrs are placed
- Hyperlab is lab in which Ocprack is placed
- Hyperceph is the datacentre in which Ocprack is configured

The dump from ceph tree command shows the crush hierarchy as designed above. There are 15 drives showing up state as configured at three OSDs.

```
[root@hsmonitor ~]# ceph osd tree
ID  WEIGHT   TYPE NAME            UP/DOWN REWEIGHT PRIMARY-AFFINITY
 -1 76.07983 root default
 -2 22.77994     host hsosd1
 -5 15.00000     datacenter hyperceph
 -6 15.00000         room hyperlab
 -7        0             row ocprack
 -8        0                 rack jbr1
 -9        0                 rack jbr2
-10        0                 rack jbr3
 29  1.00000                     osd.29         up  1.00000          1.00000
 30  1.00000                     osd.30         up  1.00000          1.00000
 31  1.00000                     osd.31         up  1.00000          1.00000
 32  1.00000                     osd.32         up  1.00000          1.00000
 33  1.00000                     osd.33         up  1.00000          1.00000
 19  1.00000                     osd.19         up  1.00000          1.00000
 20  1.00000                     osd.20         up  1.00000          1.00000
 21  1.00000                     osd.21         up  1.00000          1.00000
 22  1.00000                     osd.22         up  1.00000          1.00000
 23  1.00000                     osd.23         up  1.00000          1.00000
 24  1.00000                     osd.24         up  1.00000          1.00000
 25  1.00000                     osd.25         up  1.00000          1.00000
 26  1.00000                     osd.26         up  1.00000          1.00000
 27  1.00000                     osd.27         up  1.00000          1.00000
 28  1.00000                     osd.28         up  1.00000          1.00000
```

The CLI can be used to check the health and warning details for the OSD drives.

```
[root@hsmonitor ~]# ceph -s
    cluster 812a337e-59e6-47f1-8750-5bfa62db06c5
     health HEALTH_WARN
            712 pgs stuck inactive
            712 pgs stuck unclean
     monmap e1: 1 mons at {hsmonitor=192.168.18.124:6789/0}
            election epoch 2, quorum 0 hsmonitor
     osdmap e137: 34 osds: 15 up, 15 in
      pgmap v465: 712 pgs, 4 pools, 0 bytes data, 0 objects
```

```
539 MB used, 55716 GB / 55716 GB avail
  712 creating
```

As part of enhancement on this PoC; ceph would be used to create block storage pool and integrated with the openstack installation. Also it would upgrade to ceph2; which uses Ansible scripts for configurations and storage maintenance.

3.3 RedHat OpenShift PaaS

Redhat openshift container platform is RedHat's on-premise private platform as a service product, built around a core of application containers powered by Docker, with orchestration and management provided by Kubernetes, on a foundation of RHEL 7.3. Openshift adds developer and operational centric tools to enable rapid application development, easy deployment and scaling, and long-term lifecycle maintenance for small and large teams and applications.

The PoC installs openshift 3.4 on a single dedicated rackgox node. The setup needs three dedicated physical nodes or virtual machines. These nodes must be in same network and configured in same vlan.

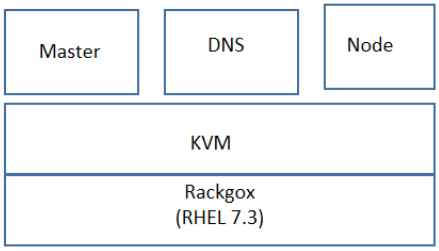

Figure 6: Openshift software installation stacks

The dedicated rackgox node is installed with RHEL7.3 and using KVM; it creates 3 VMs for master, node and DNS servers. Openshift platform requires a fully functional DNS server in the environment. This is ideally a separate host running DNS software and can provide name resolution to hosts and containers running on the platform. The installation requires all VMs to be registered with Redhat partner portal and download all requisite RPMs.

The PoC uses quick installation method with an interactive CLI utility, the **atomic-openshift-installer** command, to install openshift container platform across a set of hosts. This installer can deploy openshift components on targeted hosts by either installing RPMs or running containerized services. This installation method is provided to make the installation experience easier by interactively gathering the data needed to run on each host. The installer is a self-contained wrapper intended for usage on a Red Hat Enterprise Linux (RHEL) 7 system.

After installations, the master node should give list of all configured nodes though CLI:

```
[root@osmaster ~]# oc get nodes
NAME              STATUS    AGE
ose3-master.lab.com   Ready     20d
ose3-node.lab.com     Ready     20d
```

The master node as well provides a web interface to manage the pods:

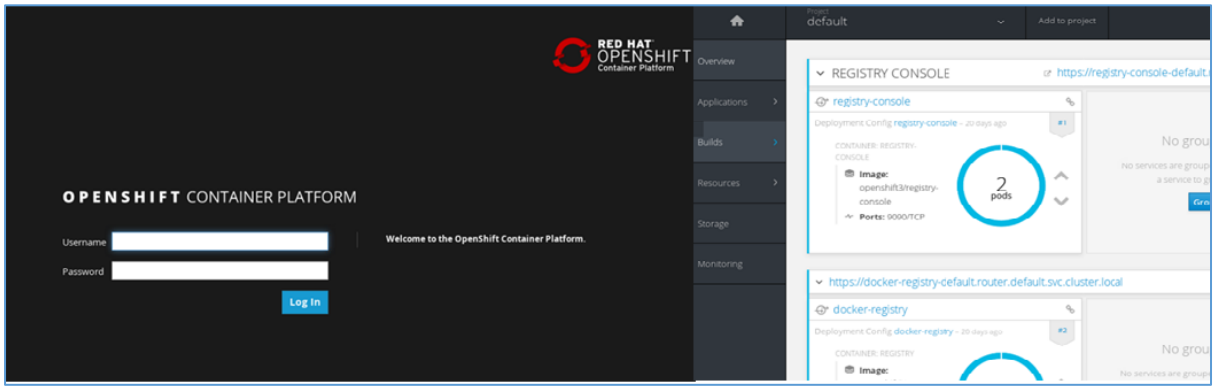

Figure 7: Openshift web management console

With OpenShift, developers deploy resources and code using either the web console or the command-line interface. The command options are high level. For example, a single command can deploy an application stack, including database, application server and Web server. Applications are designated as scalable or non-scalable, and can be deployed with an HAProxy gear for load balancing.

4. Performance and Accessibility

4.1 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.

5. Conclusion

The objective of this PoC was to showcase a design with which RedHat IaaS, PaaS and storage solutions can be installed on QCT OCP gear hardware. The configuration steps described in the document demonstrated that Openstack, Openshift and CEPH can be efficiently installed and executed with that design. As next enhancements on these appliances; the applications would be developed and they would be benchmarked with their system performance.