Introduction to the K Pre-Post Cloud Service

RIKEN R-CCS
Aug. 23, 2018
The Goal of this Material

• The goal of this material is to make you:
  • Be familiar with technical terms in OpenStack;
  • Understand the service contents of K Pre-Post Cloud;
  • Know how to get started the service.
  • Understand to create an instance through a demonstration.

• It is assumed that you are familiar with a Linux distribution
  and its configuration (not need to be an expert).

• Also, it is desirable that you have already experienced to
  use a cloud service because OpenStack provides schemes
  and APIs that resemble other cloud services.
Who the Service is for?

• The following cases we suppose are a part of examples for use of the K Pre-Post Cloud.

• **Data processing**
  • Generate a mesh file
  • Compress result files to archive
  • Transfer files to other supercomputer centers and data centers with rsync
  • Use many cores (up to 96cores per a VM) or memory usage (up to 320GiB per a VM)
  • Use high-throughput disk I/O with SSD

• **Visualization**
  • Use a remote visualization

• **Others**
  • Use open-source software
  • Use the latest Linux distribution
  • Use Windows OS (We don’t provide the OS image and its license.)
  • Use ISV software (We don’t cover any cost for paid-software.)
  • Control VMs with CLI/REST API
  • Run tasks immediately without a queuing process
  • Run extra simulations with a small number of nodes when their assigned resource has exhausted (probably, at the end of the fiscal year).
Outline

• Features (summary)
• Hardware Overview
• Service Guide
  • Software-defined resource
  • Flavor
  • Storage
  • Network
  • Quotas
  • Getting Started
• Demonstration
Background

• Issues regarding the pre-post environment
  • A lack of compute resource for pre-post processing in K
    • In the K computer environment, there are four pre-post servers installed. However, the servers are quite small-scale than the compute nodes of K. Part of users requires beefing up the facility.
  • Isolated ecosystems for open-source software
    • Even though there are myriad open-source software available on the Internet, only a part of them is available on K because most of them are developed and optimized for x86-based architecture.
  • Unsupported architecture for paid-software
    • Most of the paid-software does not support for the K computer or cannot be installed due to a software environment reasons (e.g., root privilege, incompatible shared library). At least, IA servers (x86-based pre-post servers) are suitable for the case. This kind of demands was requested by industrial users.

• In FY2017, in the K-computer environment, we added a private cloud (IaaS) as a new experimental platform to address the above issues.
# Features of the K Pre-Post Cloud

<table>
<thead>
<tr>
<th><strong>x86-based</strong></th>
<th><strong>Virtualization</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>This private cloud employs the Intel x86-based architecture to quickly use abundant software in the ecosystems, without formidable porting process. Eventually, we expect that you can reduce time-to-result.</td>
<td>This private cloud was built by the OpenStack framework to achieve virtualization. Virtualization provides huge benefits to you and operators. As an obvious benefit, the private cloud allows you to run a command as root user.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Operating System</strong></th>
<th><strong>Internet</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Various types of guest operating system (e.g., CentOS, Ubuntu) are available in the private cloud. Also, Windows Server and other third-party operating systems are bootable on a VM if you have a license and an image.</td>
<td>Every virtual machine (VM) can access the Internet. This feature helps you to easily install/update open-source software and push/pull any contents from the Internet. Also, you can configure own ingress/egress communication policy for each VM.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Storage</strong></th>
<th><strong>CLI/REST API</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A VM can use high-throughput disk I/O with SSDs for installation space of a guest OS and your processing data. There is external storage to back up VMs in the private cloud. Also, VMs can access the GFS on K. This feature allows you to use large working space in pre-post processing.</td>
<td>OpenStack framework provides well-organized Python-based command line interface (CLI) and REST API. To remotely control your compute resources in the private cloud, you can develop your application injected with code snippets using the CLI/API.</td>
</tr>
</tbody>
</table>
The vendors who played the role of building the private cloud.
- Digital Technologies Cooperation
- Red Hat K.K.
- Dell Inc.
- Fujitsu Limited (GFS-GW)
### Old and New Pre-Post Facilities

<table>
<thead>
<tr>
<th></th>
<th>(old) Pre-Post Server</th>
<th>K Pre-Post Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>Intel Xeon X7560 (Nehalem-EX) (8cores/2.26GHz/24MB) x 8 (/node)</td>
<td>Intel Xeon Platinum 8168 (Skylake) (24cores/2.7Ghz/33MB) x 2 (/node)</td>
</tr>
<tr>
<td><strong>#nodes</strong></td>
<td>2 (front nodes) + 2 (batch nodes)</td>
<td>11 (compute nodes)</td>
</tr>
<tr>
<td><strong>Total #cores</strong></td>
<td>128 cores (batch nodes)</td>
<td>528 cores (1056 vCPUs, Hyper-Threading enabled)</td>
</tr>
<tr>
<td><strong>RAM</strong></td>
<td>0.5TiB/node or 1TiB/node (The batch nodes have memory devices in different sizes.)</td>
<td>384GiB/node</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>GFS(30PB)</td>
<td>SSD(9.6TB/node)+Ceph(150TB)+GFS(30PB)</td>
</tr>
</tbody>
</table>
| **OS**           | RHEL 6.5                                                                              | HostOS: RHEL 7.4
GuestOS: CentOS, Ubuntu, etc
(A user can choose a guest OS.) |
|                  | A batch job management system (SLURM) are installed. A user can submit his/her job to the batch servers via the batch manager. | A service portal provides an interface (Web/CLI/REST API) to control his/her VM. Through the interface, a user can get his/her VM on demand. |
Features of Cloud Computing

Target resources for virtualization
- CPU (vCPU)
- RAM
- Storage
- Network

1. Server virtualization
   This technology can divide a physical server into multiple isolated virtualized environment to share resources with users. In the virtualized environment, each virtual machine can be installed a different operating system.

2. Multitenancy
   OpenStack can provide complete separation between VMs.

3. On demand
   Users can require resources by themselves as needed.
OpenStack

• A framework to build an IaaS cloud computing service
  • IaaS = Infrastructure as a Service

• OpenStack is open-source software.

• The OpenStack community is working to produce open source training materials available on the Internet.

• Please refer the following URL if you want to know OpenStack in more detail.
  • https://www.openstack.org/

• A community version of the OpenStack will be updated twice a year.
  • https://releases.openstack.org/

• Red Hat offers enterprise OpenStack solutions and support.

• There are numerous configurations depending on the system design and versions of the service components. That is, the OpenStack configuration is not unique.

• Red Hat’s solutions alleviate the complexity of open-source software.
OpenStack Architecture

- OpenStack employs loosely coupled design and consists of several service components. Except for mandatory core components, administrators can choose components based on their system design.
- These services that control compute, storage, and networking resources.
- Each service has APIs to control the service itself.

The cloud can be managed with a web-based dashboard (Horizon) or command-line clients, which allow administrators/users to control, provision, and automate OpenStack resources.

Please refer the following URL if you want to know in more detail.
- https://www.openstack.org/software/project-navigator/openstack-components#main-services
User/Group/Project

Naming rules
• “User name” is based on K-user ID and is added a postfix character ‘c’.
  • e.g., a15003 → a15003c
• “Group name” is the same with K-group ID.
• “Project name” is the same with K-group ID.
Software-defined Resource

• Your virtual machine can divide into several software-defined parts (vCPU, RAM, SSD, Ceph, and Network).

• We provide templates of resource configuration called “Flavor.”

• A user can choose the flavor that defines the size of a virtual machine that can be launched within the approved quotas.

• Ceph is external storage in the private cloud and is designed for storing VM images.

• At this time, we provide a router and an internal network. Any customizable network as a service is unavailable.
- At the moment, we provide resources based on the following flavors.

<table>
<thead>
<tr>
<th>vCPUs</th>
<th>RAM [GiB]</th>
<th>Root (ephemeral) Disk Size (SSD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>tiny 16GiB</td>
</tr>
<tr>
<td>2</td>
<td>A2</td>
<td>small 128GiB</td>
</tr>
<tr>
<td>6</td>
<td>A3</td>
<td>medium 512GiB</td>
</tr>
<tr>
<td>12</td>
<td>A4</td>
<td>large 2TiB</td>
</tr>
<tr>
<td>24</td>
<td>C1 A5 B1 B2 B4</td>
<td>huge 8TiB</td>
</tr>
<tr>
<td>48</td>
<td>C2 C4 A6 B3 B5</td>
<td>24vCPUs 64GiB + 512GiB (SSD)</td>
</tr>
<tr>
<td>96</td>
<td>C3 C5 C6 A7 A8</td>
<td>Example: A5.medium</td>
</tr>
</tbody>
</table>

Instance (VM) Type
A1-8: standard
B1-5: memory-oriented
C1-6: compute-oriented
VM Duration (Important)

- To give more users an opportunity to use the private cloud, we introduce a simple mechanism that automatically terminates old VMs in a given period of time depending on the flavors.

- This policy is based on that a bigger resource consumer tends to be imposed short duration. Meanwhile, the policy allows smaller VMs to live longer.

<table>
<thead>
<tr>
<th>vCPUs</th>
<th>VM Maximum Duration (tiny, small, medium)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAM [GiB]</td>
</tr>
<tr>
<td>1</td>
<td>4 8 16 32 64 128 256 320</td>
</tr>
<tr>
<td>2</td>
<td>inf</td>
</tr>
<tr>
<td>6</td>
<td>inf</td>
</tr>
<tr>
<td>12</td>
<td>inf</td>
</tr>
<tr>
<td>24</td>
<td>4w 4w 2w 2w 1w</td>
</tr>
<tr>
<td>48</td>
<td>4w 2w 2w 1w 1w</td>
</tr>
<tr>
<td>96</td>
<td>2w 2w 1w 1w 1w</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>vCPUs</th>
<th>VM Maximum Duration (large, huge)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAM [GiB]</td>
</tr>
<tr>
<td>1</td>
<td>4 8 16 32 64 128 256 320</td>
</tr>
<tr>
<td>2</td>
<td>1w</td>
</tr>
<tr>
<td>6</td>
<td>1w</td>
</tr>
<tr>
<td>12</td>
<td>1w</td>
</tr>
<tr>
<td>24</td>
<td>1w 1w 1w 1w 1w 1w</td>
</tr>
<tr>
<td>48</td>
<td>1w 1w 1w 1w 1w 1w</td>
</tr>
<tr>
<td>96</td>
<td>1w 1w 1w 1w 1w 1w</td>
</tr>
</tbody>
</table>

**TIPS**
- We provide a backup space (Ceph storage) to store VM snapshots.
- A backup (snapshot) file size depends on the root disk size of your VM.
- The Ceph storage space is not enough to save all the user data. Thus, we recommend using tiny, small, or medium root disk size to save the storage resource.
- Anyway, to prevent losing your VM, we recommend to back up your VM by the snapshot feature as needed.

1w: 1week
2w: 2weeks
4w: 4weeks
inf: the end of the fiscal year or the expiration date
**Storage**

- We provide several types of storage you can choose.

  - **SSD (ephemeral, root disk)**
    - The storage space uses RAID0-based disk arrays installed in compute nodes.
    - You can see a block storage on your VM.
    - In default, this storage space is used for a guest OS installation and storing user data.
    - We call it “root disk.” Carefully, it’s not called “volume” in OpenStack.
    - **By the termination of a VM, the root disk is deleted. (This is not persistent storage.)**

  - **Ceph (volume)**
    - It’s external storage space to store VM images.
    - This storage automatically replicates data with three-redundant copies and makes it fault-tolerant using cluster nodes.
    - You can use this space instead of the SSD volume to install a guest OS (not recommend).
    - At the moment, the storage space is not enough to store the bulk of input/output data of your simulation.

  - **Global File Storage (GFS) on K**
    - The private cloud allows your VM to access the GFS space using SFTP or SSHFS.

  - **Other Storage via the Internet**
    - Your VM can access any resources on the Internet.
Storage (Important)

- This figure shows the dialog window for creating a VM in Horizon.
- **If you want to use the SSD device (we recommend), choose “No” in the “Create New Volume” switch.**
- If you choose “YES” in the switch, your VM can be attached to arbitrary size space from Ceph storage. (At the time, the disk size defined in the flavor you choose is ignored.)
- Also, if you attach Ceph storage in your VM, the snapshot feature does not work appropriately. (Snapshot size will be zero bytes.)

All steps in this process are shown in the tutorial material below. (This introduction omits the details.)
Network (overview)

- The internal network among VMs and the Ceph storage space uses a 25GbE network.
- The private cloud system provides private IP addresses (10.9.0.0/16) to VMs.
- Through VPN connection, you can access the private cloud network and VMs.
- Your VM can access the Internet via NAT-GWs (gateway) with 10GbE.
- Also, your VM can access the global file storage space on K via GFS-GWs.
- VM is not allowed to access from the outside of the private cloud without a VPN session to comply with the RIKEN security policy.

In the next slide, the inside of the dotted frame is depicted in more detail.
• In default, each project has one router that works as SNAT.
• The external network is shared with all projects.

**TIPS**
• The private cloud provides a firewall (packet filter) called security group.
• By the feature, you can configure to permit (or not to permit) ingress/egress TCP/UDP ports and ICMP.

private IP: 192.168.0.0/16
public IP (external): 10.9.0.0/16
Network (Important)

• This figure shows the dialog window for creating a VM in Horizon.
• To appropriately connect the Internet from your VM, the VM needs to attach a given internal network.
• If your project name is “guest,” “guest-internal” is the correct internal network as shown in the figure.

Naming rules
• If your project name is “project1”, the given internal network is “project1-internal.”

All steps in this process are shown in the tutorial material below. (This introduction omits the details.)
Quotas

- These are part of quotas per a project. (Quotas can be defined for each project by administrators.)
- If you consumed resources exceeding one of the quotas, your VM creation process would be failed/rejected.
- We can change the quotas based on your request.

<table>
<thead>
<tr>
<th>Type of Quota</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute</td>
<td></td>
</tr>
<tr>
<td>#vCPUs</td>
<td>192</td>
</tr>
<tr>
<td>#Instances</td>
<td>20</td>
</tr>
<tr>
<td>RAM [MB]</td>
<td>327680</td>
</tr>
<tr>
<td>Volume/Snapshot</td>
<td></td>
</tr>
<tr>
<td>#Volumes</td>
<td>10</td>
</tr>
<tr>
<td>Total size of Volumes and Snapshots in Ceph [GiB]</td>
<td>8192</td>
</tr>
<tr>
<td>Network</td>
<td></td>
</tr>
<tr>
<td>#Security Groups</td>
<td>20</td>
</tr>
<tr>
<td>#Security Group Rules</td>
<td>50</td>
</tr>
<tr>
<td>#Floating IPs</td>
<td>10</td>
</tr>
</tbody>
</table>
Getting Started

• The Service can be utilized free of charge.
• Every K user is eligible to use this service.

• Application method
  • To get started the service, you need to apply via the website below.
    • http://www.r-ccs.riken.jp/ungi/prpstcloud/

  • In the website, we also provide useful information/slides to use the service.

• Contact
  • If you have questions about general issues, please send an e-mail to the following address. Your feedbacks help improving our service.
    • r-ccs-k-desk@riken.jp (K support desk)
Tentative Schedule

- Apr. to Jul. in 2018
  - Preliminary phase by the unit members and part of the users.

- Aug. in 2018
  - Announcement to the users in R-CCS and a meeting will be held to explain the service.
  - Experimental phase

- Oct. in 2018
  - Announcement to all the K users and a meeting will be held to explain the service.
  - Experimental phase

- After shutting down the K computer, this service will continue.

- Also, in this fiscal year, we will add GPUs into the private cloud.
Disclaimer

• “K Pre-Post Cloud” is a private cloud service as an experimental platform in the K computer environment in order to enhance pre-post data processing features.

• Since we aim to provide the Service as an experimental service to obtain technical knowledge and know-how of operation for pre-post servers installed in the supercomputer environment, therefore, this Service may be inferior in certain ways including its contents and procedures.

• You can see all the terms of service on the website below.
  • http://www.r-ccs.riken.jp/ungi/prpstcloud/

• The terms strictly define the rules, but we wish to help you as long as we can.
Demonstration