K Pre-Post Cloud

Tutorial for the use of GPGPU instances

RIKEN R-CCS

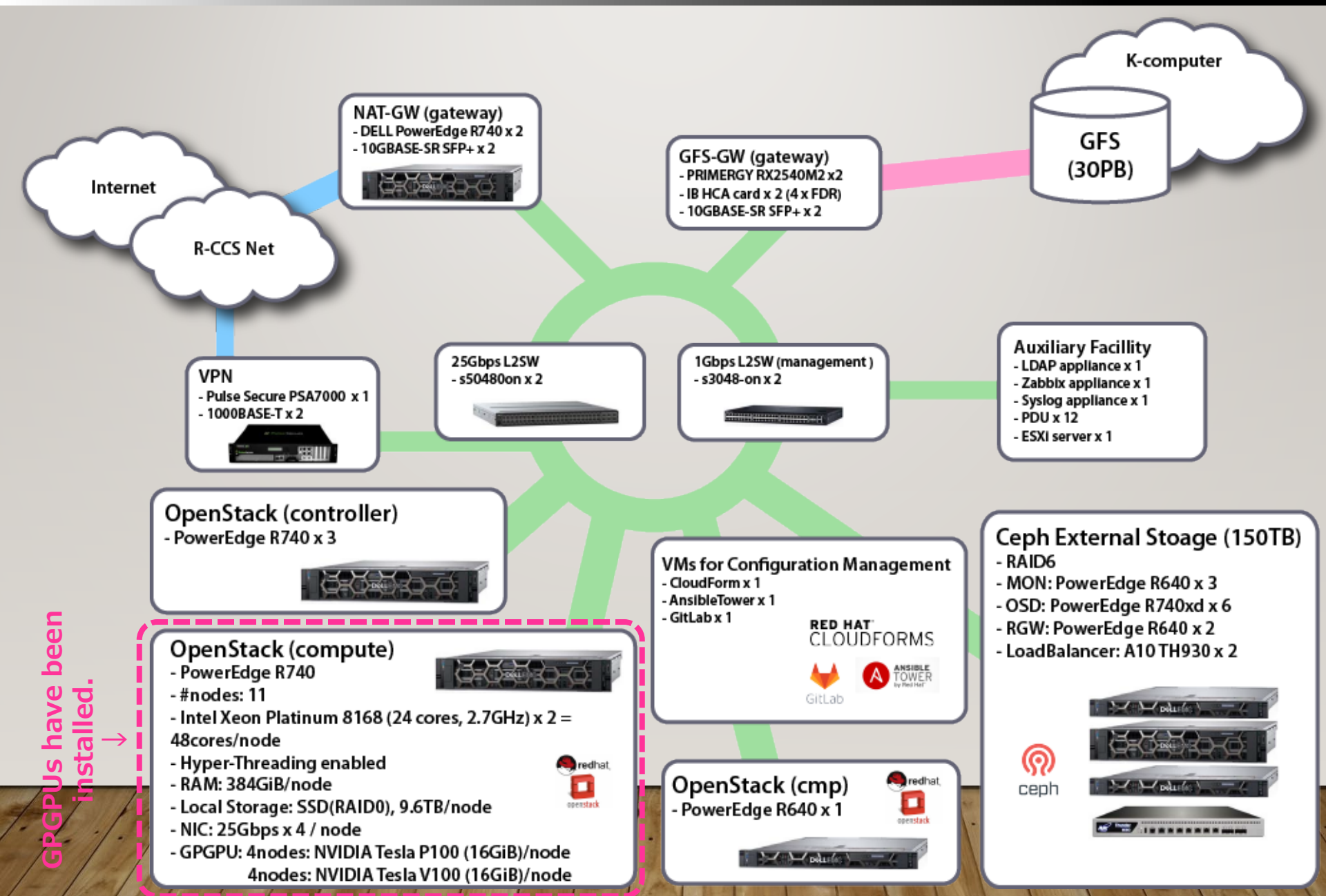
MARCH 29, 2019

About this Slides

This material provides additional information regarding the use of GPGPU instance. (GPGPUs are installed in March 2019.) and is based on the previously released tutorial named “Tutorial for basic usage.”

If you’ve never seen the tutorial, we recommend referring the tutorial before getting started.

System Overview



Overview of GPGPU installation

In FY2018, we installed 8 GPGPUs to the 8 compute nodes, respectively.

The 8 compute nodes consists of

- 4 compute nodes that have 4 GPGPUs (NVIDIA Tesla P100 (16GiB) x 1), resp, and
- 4 compute nodes that have 4 GPGPUs (NVIDIA Tesla V100 (16GiB) x 1), resp.

Each GPGPU is exclusively assigned to a single instance.

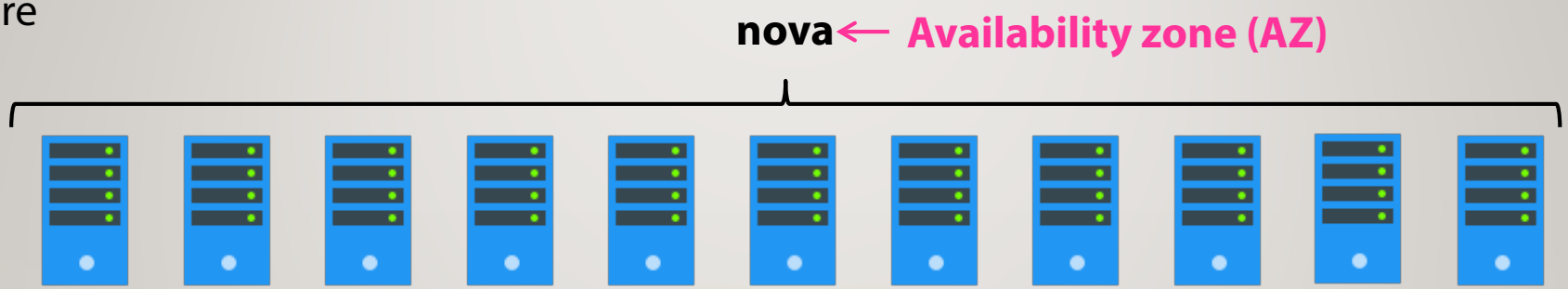
- 8 GPGPU instances can be used simultaneously in the system.
- If 8 instances have been already used, a user's request to create an additional GPGPU instance will be failed.
- Also, the service does not support to share a GPGPU by several instances (e.g., VDI).

Changes:

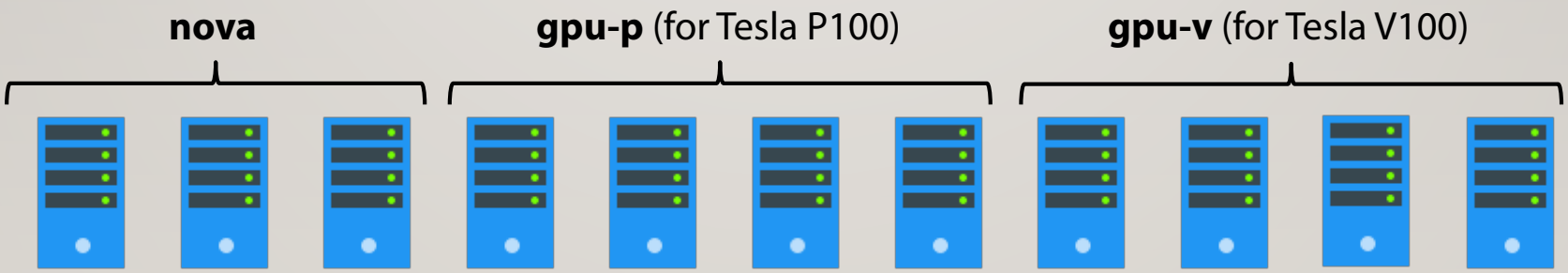
- New Availability Zones (gpu-p/v) are added.
- New Flavors (A8.huge.gpu-p/v) are added.

Availability Zone

Before



After (from March 11, 2019)

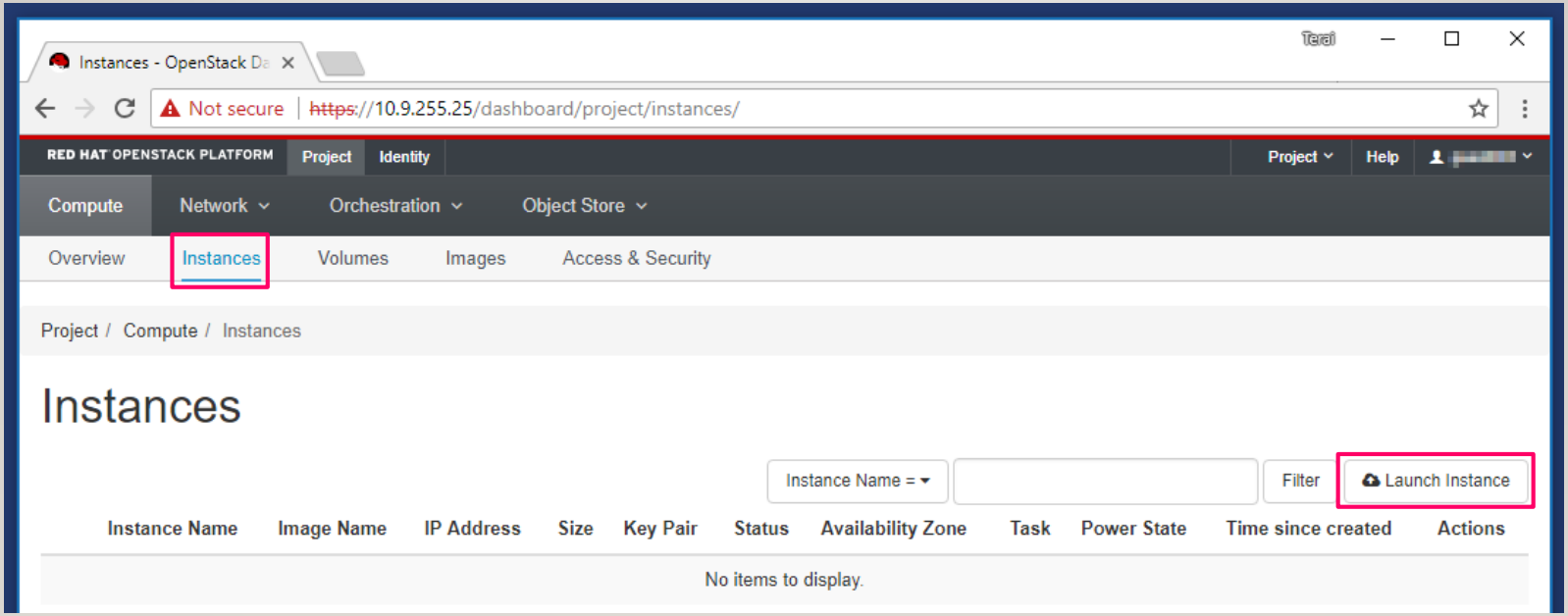


In nova, users can choose several flavors ranging from 1vCPU to 96vCPU as before.

In this slide, we ignore "cmp" availability zone because it's used internally in the system.

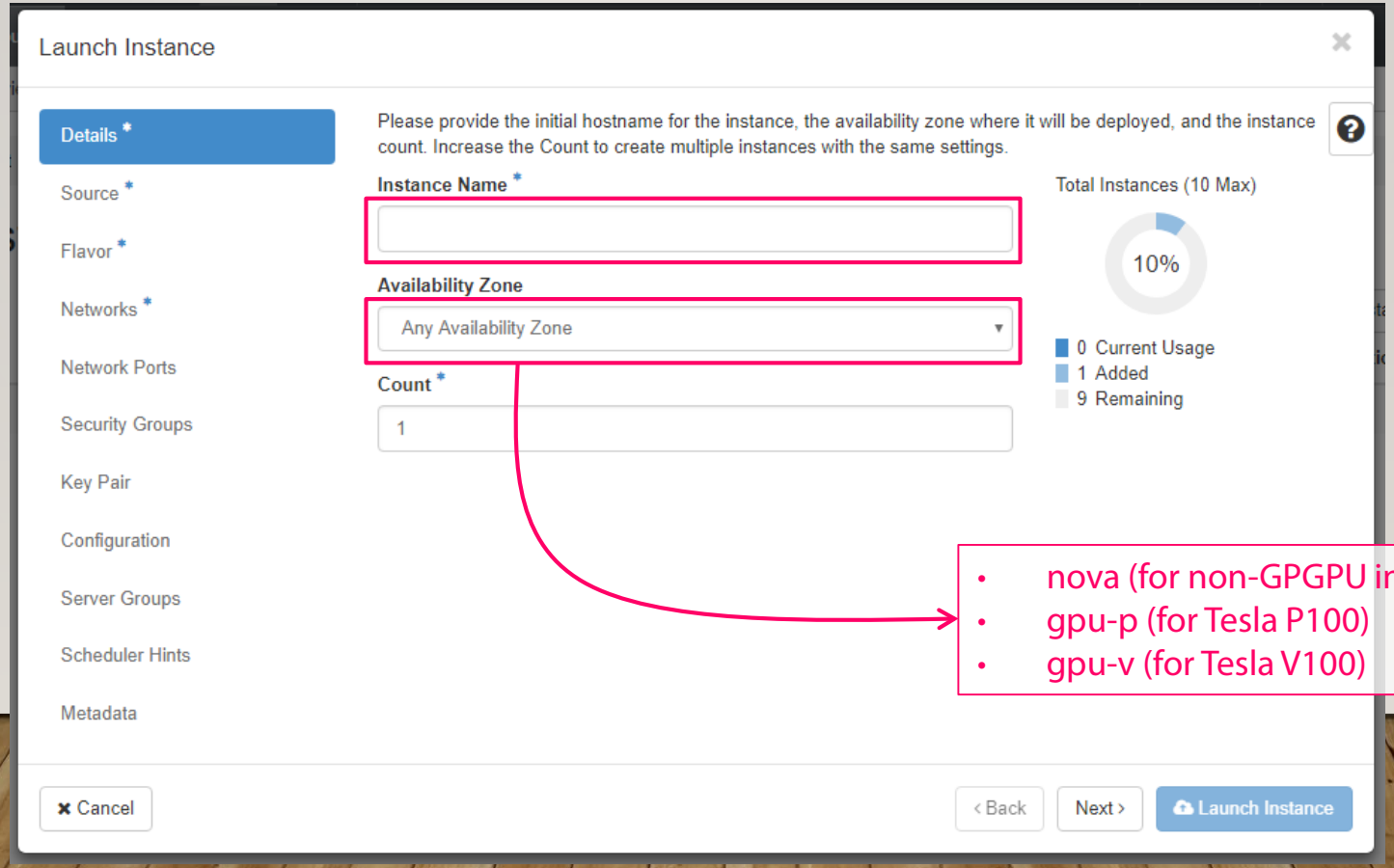
Create a GPGPU Instance (1/6)

- The first step is the completely same as the normal usage without GPGPU node.
- Click [Project] -> [Compute] -> [Instances] on the navigation bar.
- Click the [Launch Instance] button.



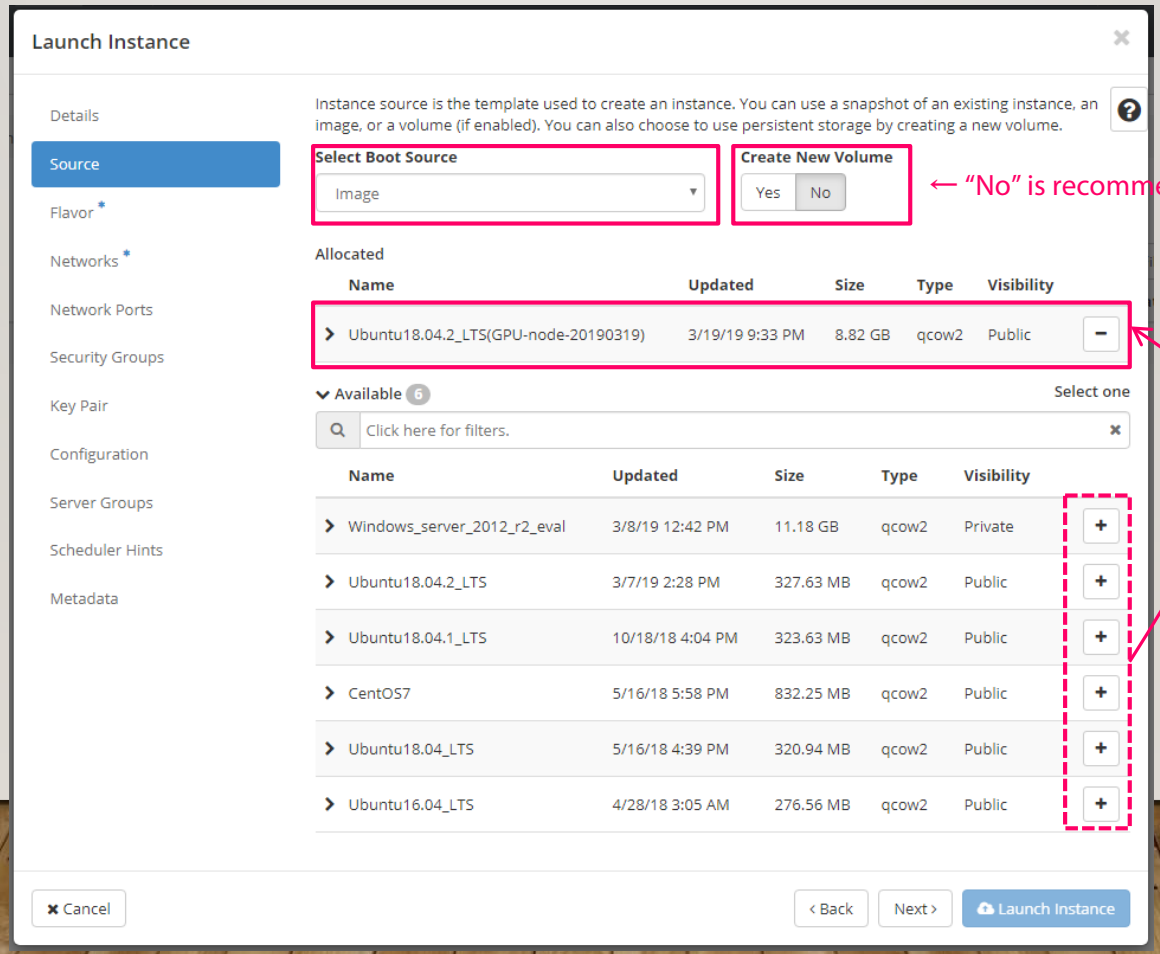
Create a GPGPU Instance (2/6)

- You can see the wizard dialog to create an instance.
- In the [Details] step of the wizard, input your instance name to the [Instance Name] field.
- Select an availability zone from the [Availability Zone] list that includes
  - **nova** (default): for instance(s) without GPGPU(s),
  - **gpu-p**: for Tesla P100, and
  - **gpu-v**: for Tesla V100.



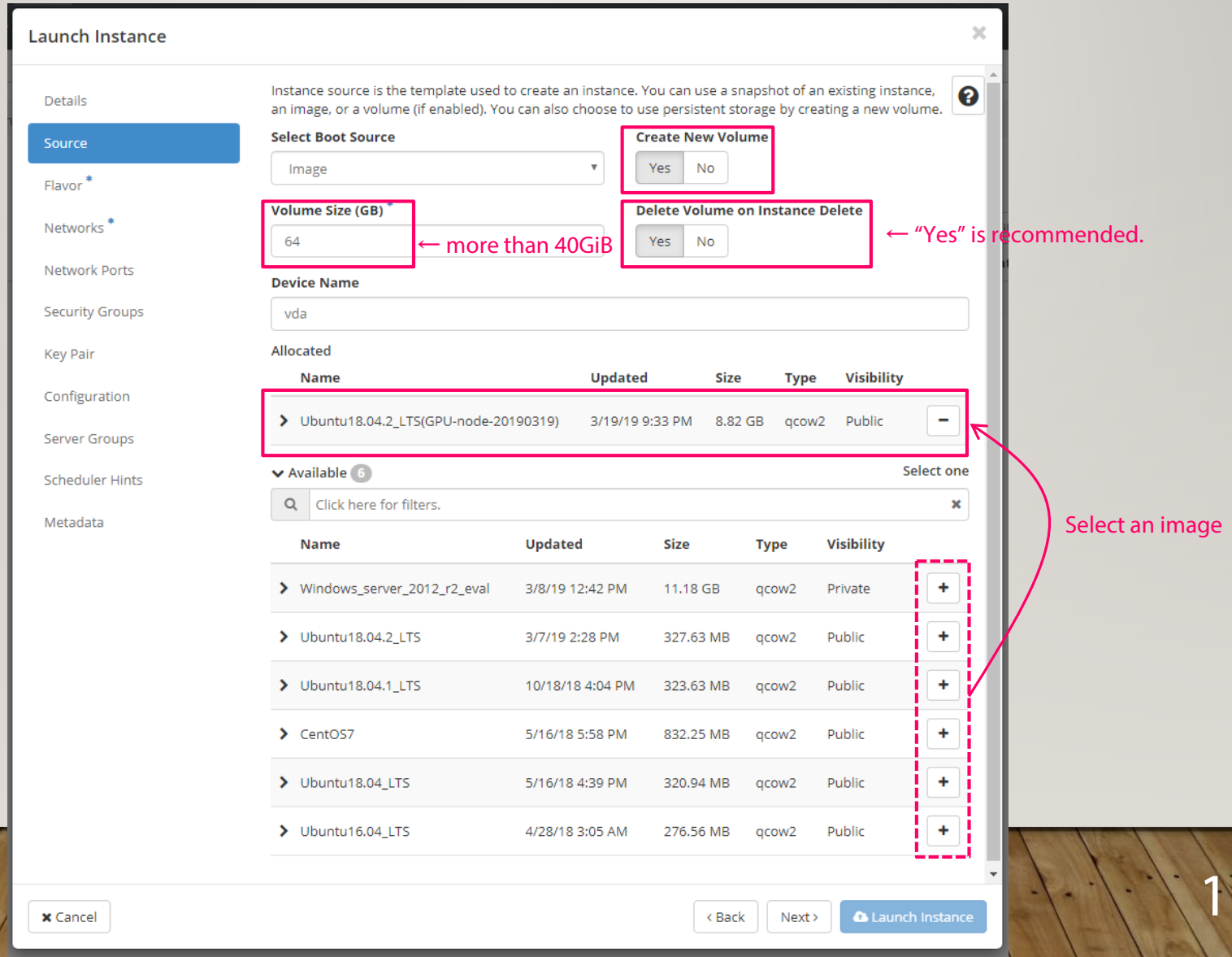
Create a GPGPU Instance (3/6)

- Select the [Image] item from the [Select Boot Source] pull-down menu.
- Select the [No] button in the [Create New Volume] switch.
- Add an OS image in the [Available] list.
  - At the end of FY2018, Ubuntu18.04.2_LTS(GPU-node-20190319) is available to create a GPGPU instance.



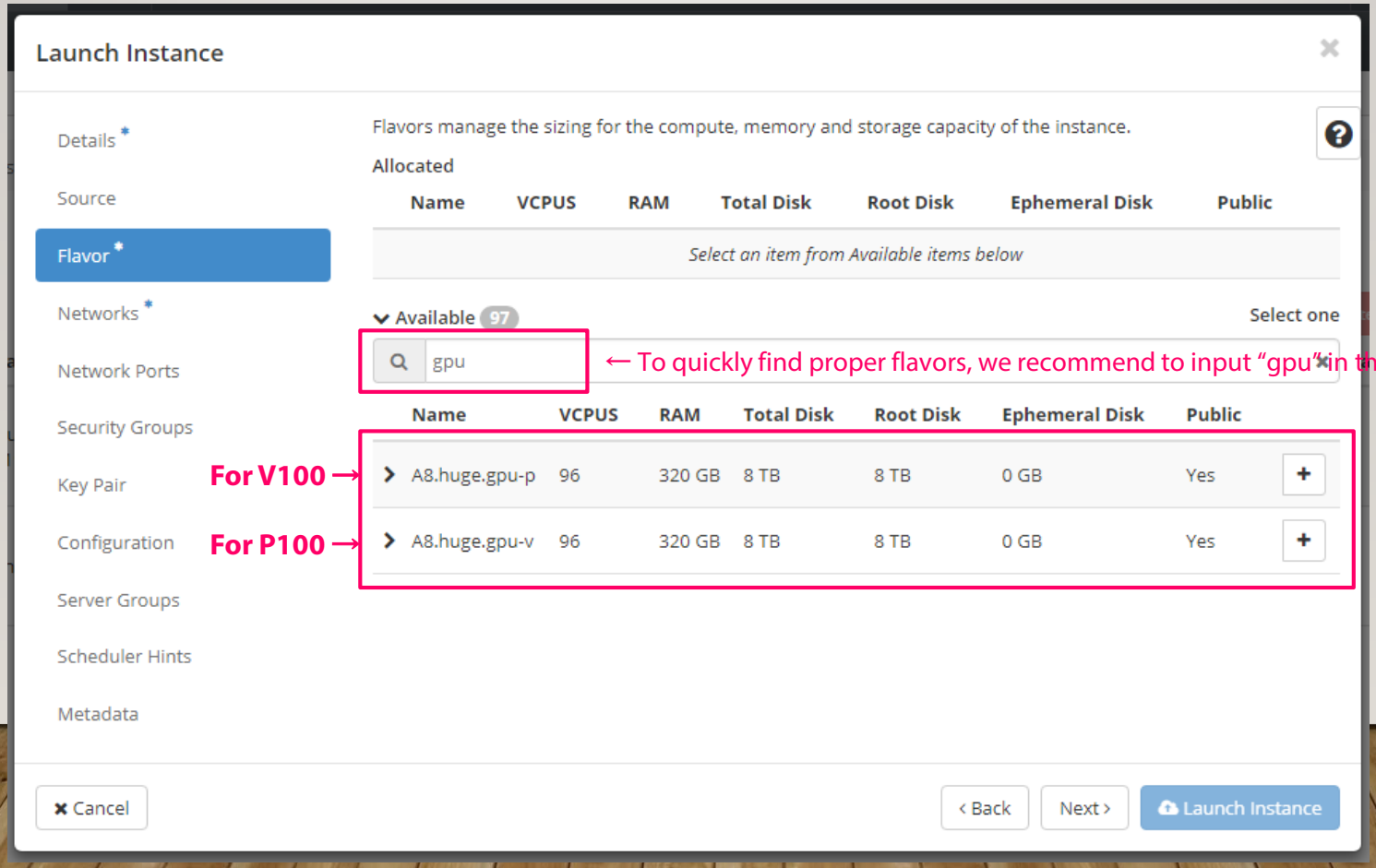
Create a GPGPU Instance (4/6)

- If you choose the [Yes] button in the [Create New Volume] switch, you must specify more than 40GiB in the volume size.



Create a GPGPU Instance (5/6)

- Add a flavor from the [Available] list.
- We newly provide GPGPU flavors (A8.huge.gpu-p/v) that consumes the whole of a compute node as well as A8.huge.
- For P100/V100, select A8.huge.gpu-p/v, resp.
- If you select other flavors for instance(s) without GPGPU(s), the request will be failed.



Create a GPGPU Instance (6/6)

- The rest of the steps are the same with the usage for instance(s) without GPGPU(s).
  - Add an internal network.
  - Add security group(s).
  - Add key pair(s).
  - Click the [Launch Instance] button.
  - After about 3 minutes, the instance using root disk will be launched.
  - Assign the Floating IP address to the instance.
  - The instance is ready to access using SSH.

Image for GPGPU instance

- Currently (in March 2019), we provide a single image (Ubuntu 18.04.2LTS based) for GPGPU instance.
- The file name depends on the updated date.
- As of March 28, 2019, this image includes
  - NVIDIA Driver version 410.48,
  - CUDA Toolkit release 10.0,
  - Docker (Engine 18.09.3, Client 18.09.3), and
  - NVIDIA-Docker 2.0.3.

TIPS

- If you find an error when the system spawns a new GPGPU instance, please check the following points.
  - Check the combination of the availability zone and the flavor you chose.
    - gpu-p + A8.huge.gpu-p
    - gpu-v + A8.huge.gpu-v
  - Check the quota of your project and unallocated resources that has sufficient space to launch your GPGPU instance.
    - In default settings, a single project can create a few GPGPU instances.
    - If you need to expand the quota, please contact us.
  - There is no available resource to launch GPGPU instance in the system.
    - If the system has already launched 8 GPGPU instances (including reserved/error instances), your request to create an additional GPGPU instance will be failed.
    - In this situation, it's difficult to sort out the problem by a user, please contact us.
    - Also, an error instance remains to be reserved a GPGPU node. Therefore, please release the instance with the error.