CERN

Providing ARM and GPU resources in the CERN Private Cloud Infrastructure

Maryna Savchenko

HEPiX Spring
28 March 2023

Outline

- Overview
- Bootstrapping ARM
- Ironic changes
- Adaptations to the PXE boot service
- Offering ARM VMs
- GPU
  - *PCI passthrough*
  - *Virtual GPU*
  - *Multi-instance GPU*
- Summary

On-premises: OpenStack

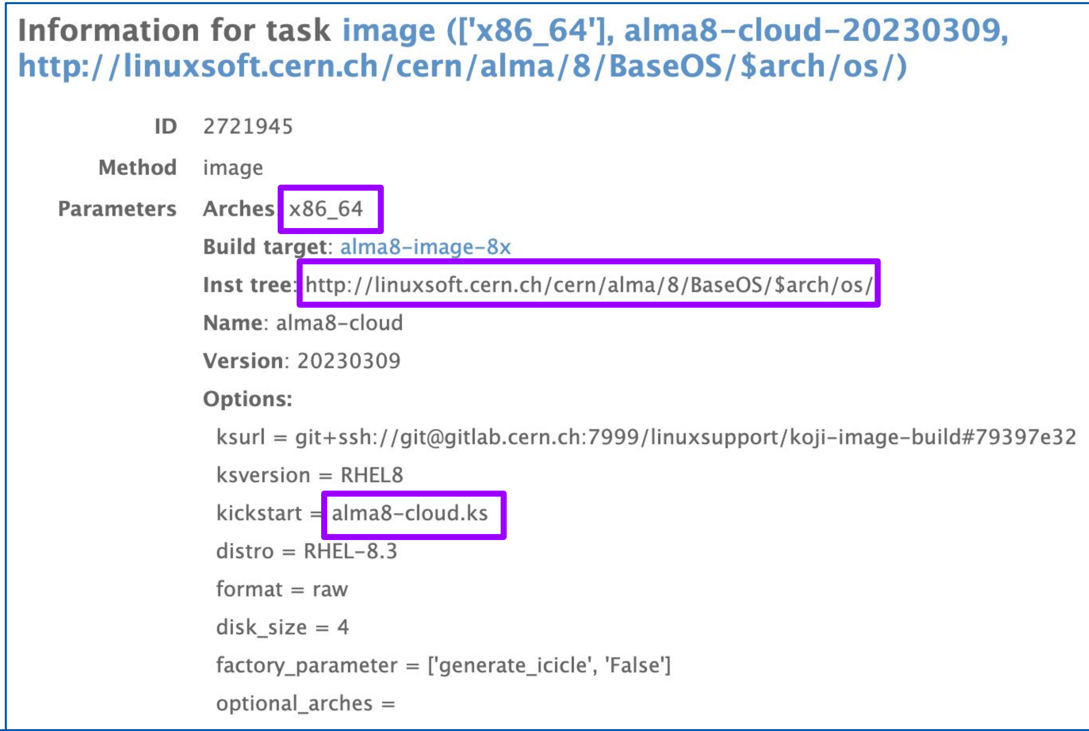
openstack

<table>
<tr><th>Resource</th><th>Spec</th><th>Use cases</th></tr>
<tr><td><b>ARM</b> Altra
“Mt. Snow”
AMPERE</td><td>5 nodes:
v8.2 Neoverse-N1 2.8GHz,
256GB</td><td>all LHC experiments
HEP benchmarking
IT: Linux, gitlab, lxplus, Ceph</td></tr>
<tr><td><b>GPU</b>
NVIDIA</td><td><table>
<tr><th>Model</th><th>Nodes</th><th>Cards</th></tr>
<tr><td><i>V100</i></td><td>5</td><td>17</td></tr>
<tr><td><i>V100S</i></td><td>6</td><td>24</td></tr>
<tr><td><i>T4</i></td><td>73</td><td>76</td></tr>
<tr><td><i>A100</i></td><td>18</td><td>72</td></tr>
</table></td><td>V100(S): batch
T4: batch, SWAN, ML …
A100: batch, ML</td></tr>
</table>

Building Linux image

```
koji image-build ...
```

1. QEMU emulator
2. Installing VM using kickstart file
3. Snapshot of VM is an image



Information for task image (['x86_64'], alma8-cloud-20230309, http://linuxsoft.cern.ch/cern/alma/8/BaseOS/$arch/os/)

ID 2721945

Method image

Parameters **Arches** x86_64

Build target: alma8-image-8x

Inst tree: http://linuxsoft.cern.ch/cern/alma/8/BaseOS/$arch/os/

Name: alma8-cloud

Version: 20230309

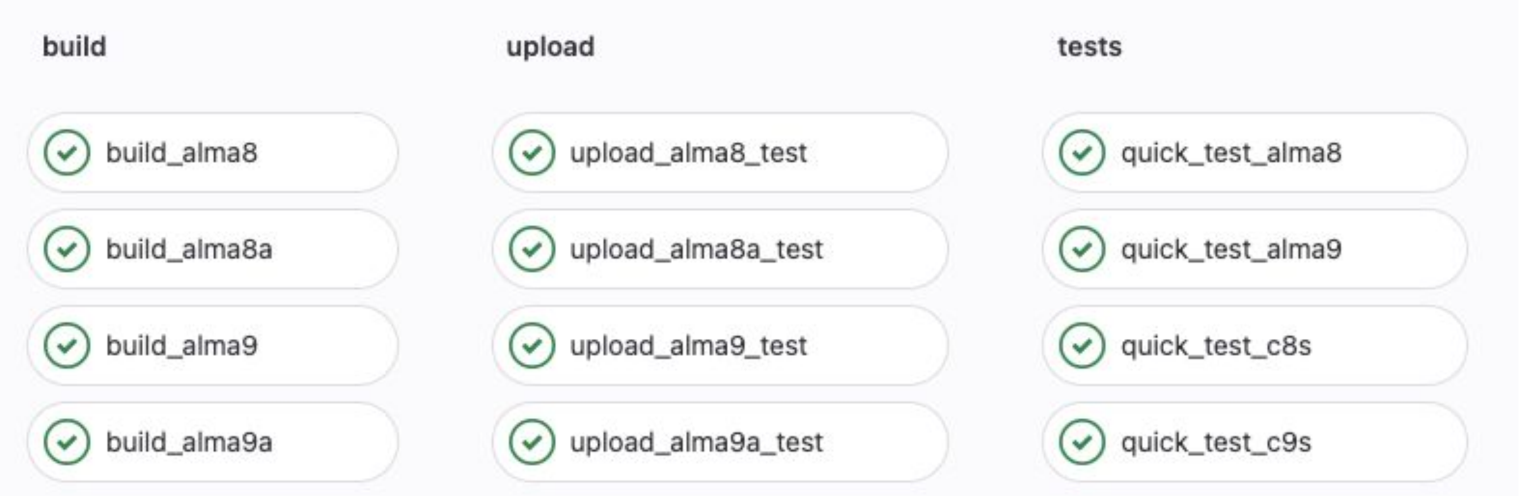
Options:

```
ksurl = git+ssh://git@gitlab.cern.ch:7999/linuxsupport/koji-image-build#79397e32
ksversion = RHEL8
kickstart = alma8-cloud.ks
distro = RHEL-8.3
format = raw
disk_size = 4
factory_parameter = ['generate_icicle', 'False']
optional_arches =
```

Bootstrapping ARM

1. No AArch64 physical node
   a. *Own version QEMU emulator AArch64*
   b. *RHEL 8 kernel on koji builder to run AArch binaries*
2. Koji builds packages for AArch64
3. Installing VM using kickstart file
4. Snapshot of VM is an image

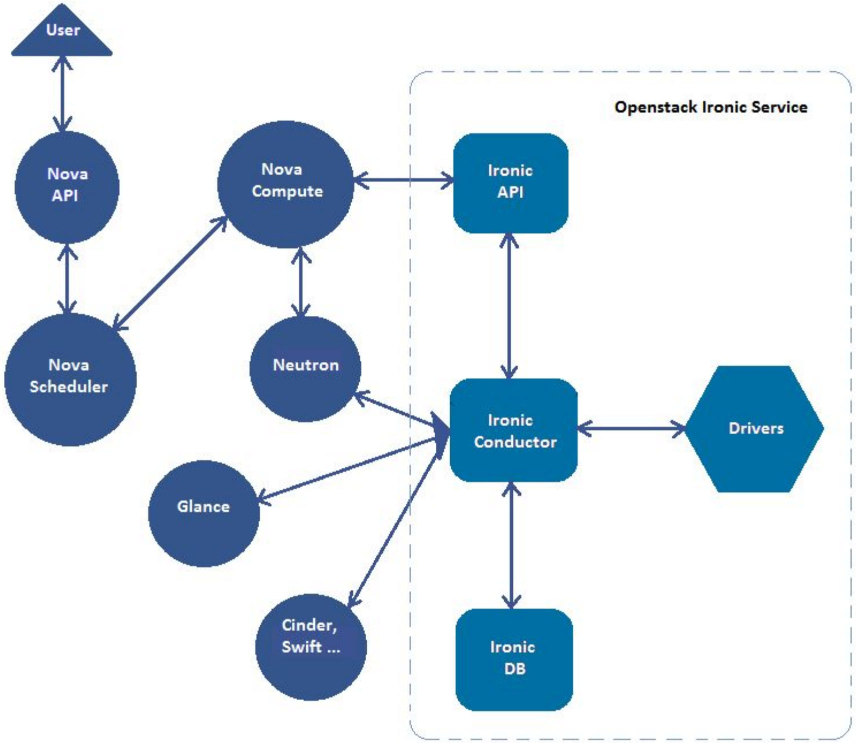
Openstack ARM image



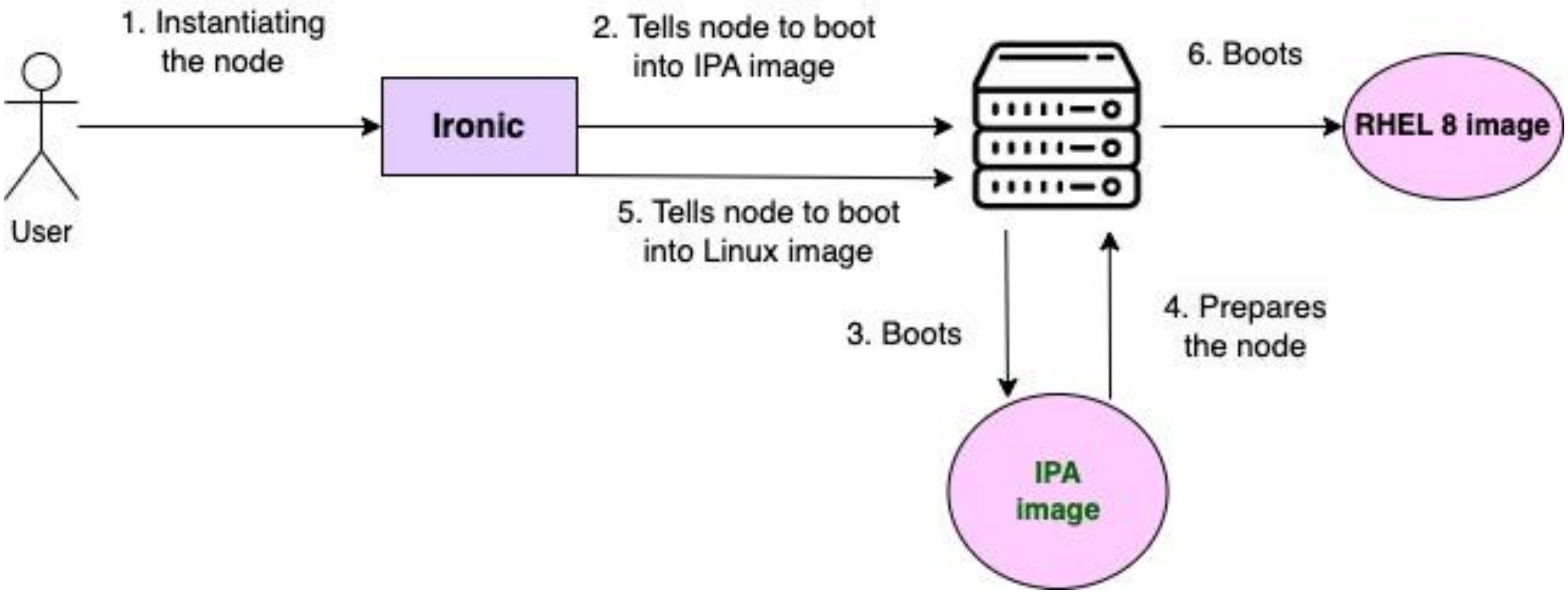
build	upload	tests
✓ build_alma8	✓ upload_alma8_test	✓ quick_test_alma8
✓ build_alma8a	✓ upload_alma8a_test	✓ quick_test_alma9
✓ build_alma9	✓ upload_alma9_test	✓ quick_test_c8s
✓ build_alma9a	✓ upload_alma9a_test	✓ quick_test_c9s

28.03 at 11.45 - *Fully automated: Updates on the Continuous Integration for supported Linux distributions at CERN*

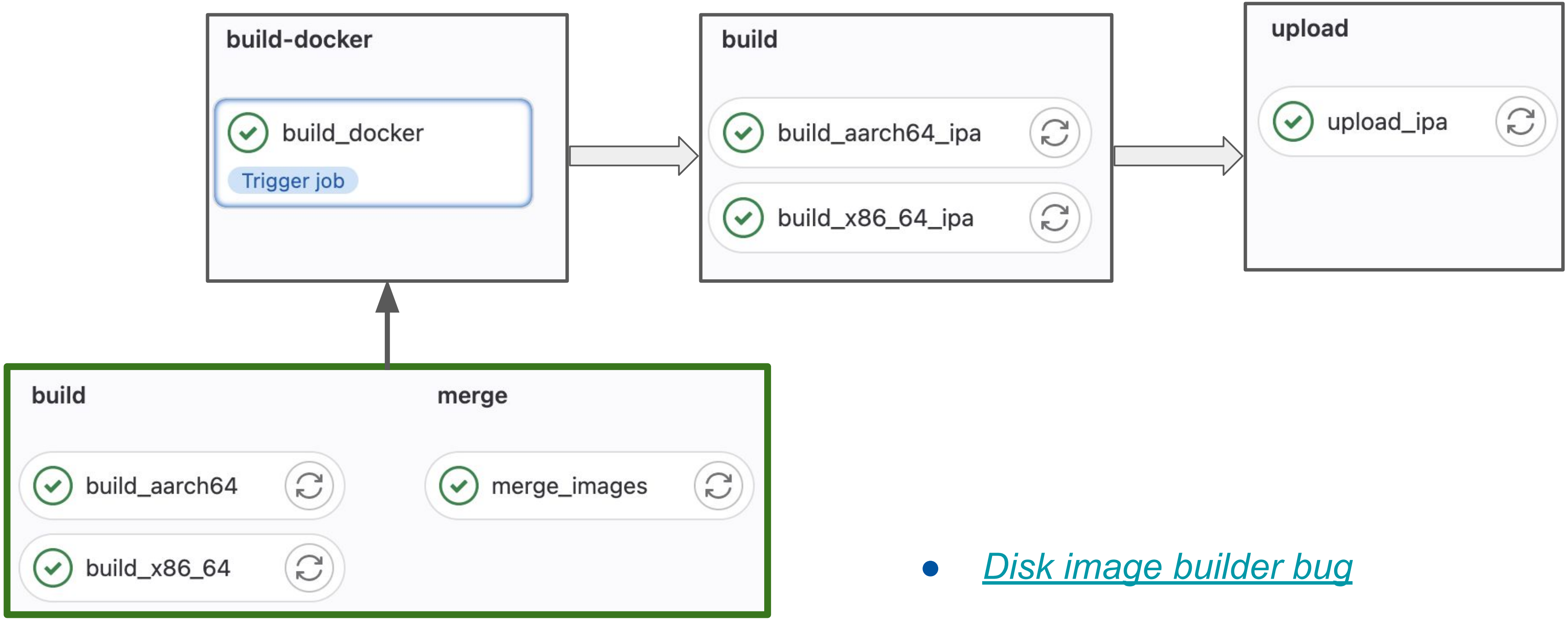
Providing physical resources with Ironic



Providing physical resources with Ironic

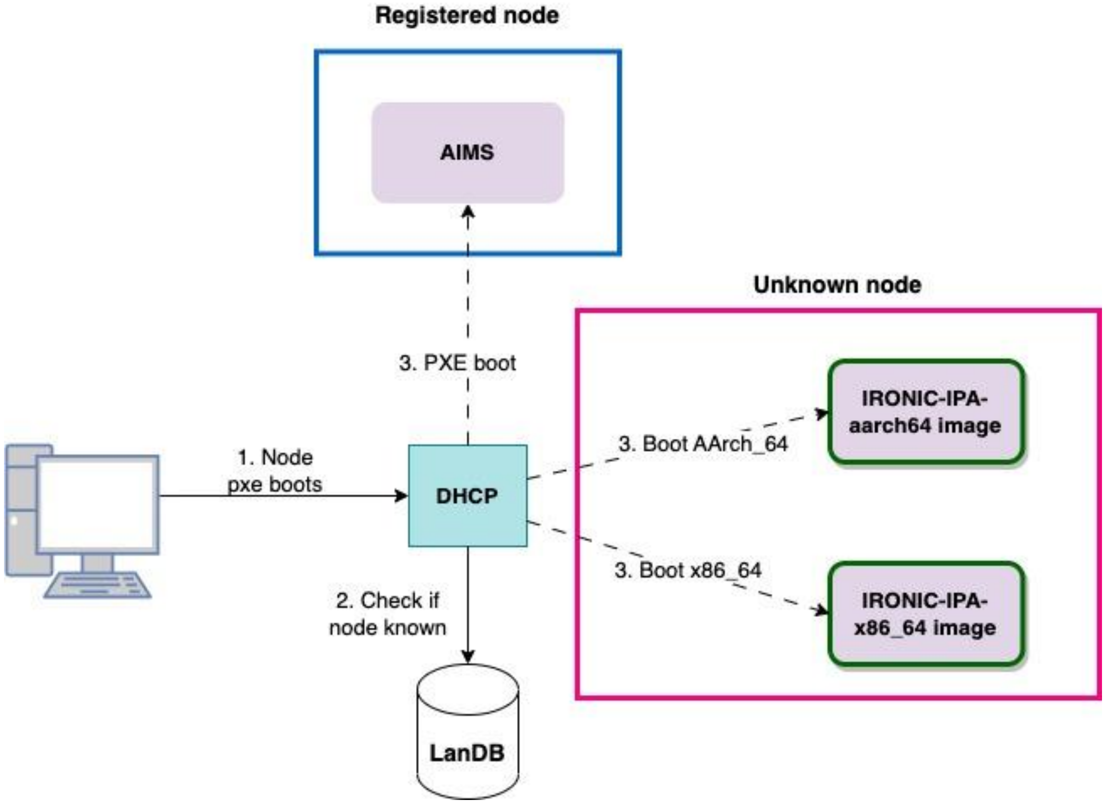

1. Instantiating the node
User
Ironic
2. Tells node to boot into IPA image
5. Tells node to boot into Linux image
6. Boots
RHEL 8 image
3. Boots
4. Prepares the node
IPA image

Ironic Python Agent Image Building



- *Disk image builder bug*

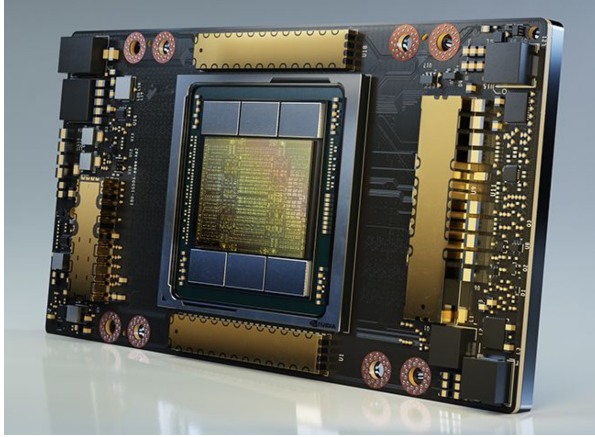
Adaptations to the PXE boot service



Provisioning VMs on ARM

- Required EL8
- Adapt configuration
- Libvirt bug
  - *Unknown processor*
- Image filtering
- Flavor capabilities host filtering

GPU Overview



- *Models: T4, A100, V100s, V100*
- *Provided as VMs*

Method of providing	**Type of access**	**Model**
PCI-passthrough	Full access	T4, A100, V100s, V100
vGPU	Time sharing	T4, A100, V100s, V100
Multi-instance GPU	Partition sharing	A100

PCI-passthrough

⊕ Direct access to the graphics card from the guest

⊖ No monitoring of the GPU usage on the hypervisor

⊖ One device per GPU - no sharing

- EL7 with newer kernel on hypervisor
- Out of the box for EL7 guests
- Additional kernel boot options for EL8 and EL9 guests

Virtual GPU

- ⊕ Hypervisor drivers give access to GPU usage information
- ⊕ Physical card shared between multiple virtual machines

- ⊖ Timesharing
- ⊖ Licenses for virtualisation drivers

- Puppet configuration:
  - CUDA
  - Drivers

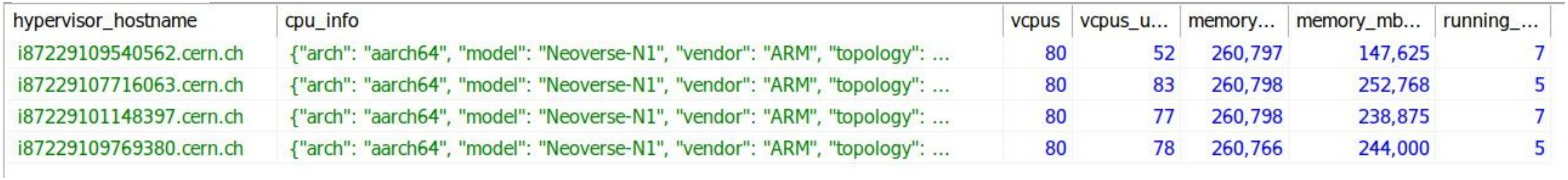
Multi-instance GPU

- ⊕ Physical card shared between multiple virtual machines
- ⊕ Physical chunk, not timeshared
- ⊕ Thermal and power consumption per card only

- ⊖ All cards in a single HV have to be partitioned the same way
- ⊖ Only 1 device per VM
- ⊖ Licenses for virtualisation drivers

- Required a backport for UUID treatment for Nova

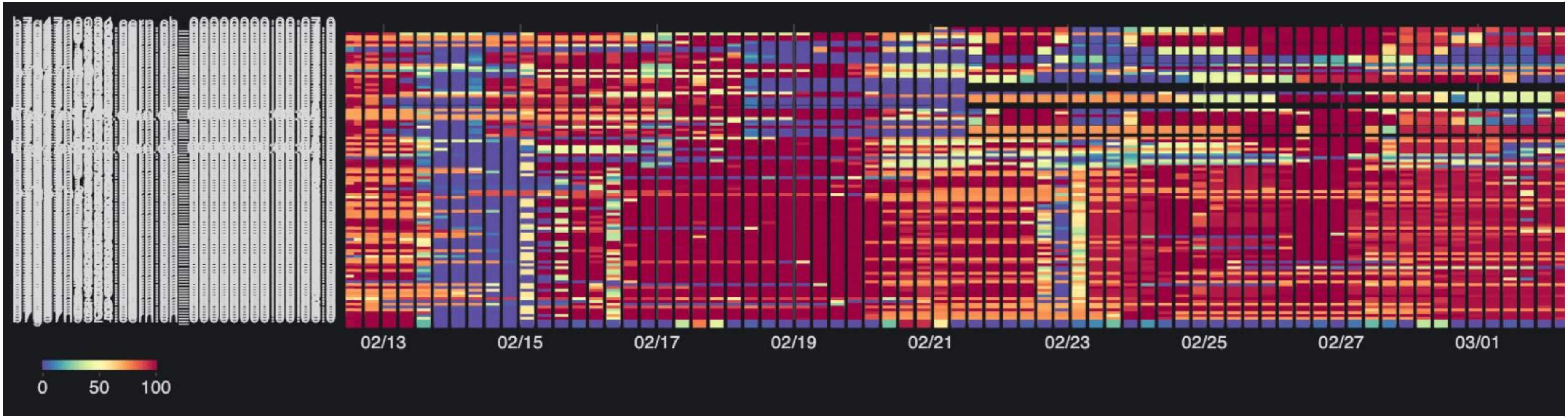
Summary - ARM utilization



hypervisor_hostname	cpu_info	vcpus	vcpus_u...	memory...	memory_mb...	running_...
i87229109540562.cern.ch	{"arch": "aarch64", "model": "Neoverse-N1", "vendor": "ARM", "topology": ...	80	52	260,797	147,625	7
i87229107716063.cern.ch	{"arch": "aarch64", "model": "Neoverse-N1", "vendor": "ARM", "topology": ...	80	83	260,798	252,768	5
i87229101148397.cern.ch	{"arch": "aarch64", "model": "Neoverse-N1", "vendor": "ARM", "topology": ...	80	77	260,798	238,875	7
i87229109769380.cern.ch	{"arch": "aarch64", "model": "Neoverse-N1", "vendor": "ARM", "topology": ...	80	78	260,766	244,000	5

Summary - GPU utilization

- *PCI passthrough over vGPU*



Plans

- High demand for Non-x86 resources
- GPU:
  - *Multi-instance GPU*
  - *Ironic burn-in of GPU*
  - *GPU benchmarking*

Thank you!

All our **open source** code is available on https://gitlab.cern.ch/cloud-infrastructure

My email: *maryna.savchenko@cern.ch*

CERN

home.cern