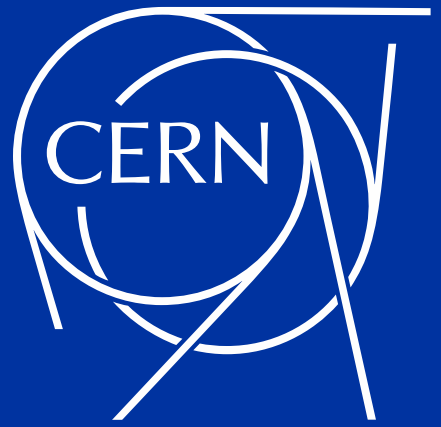


Proposal of implementation of an Heterogeneous and non-x86 Architectures Platform

A project collaboration with IT-GOV-INN, IT-GOV-ENG, IT-CD, IT-FA

M. Girone, L. Atzori, A. Wiebalck

CERN IT, Innovation Review Board, 26 April 2023

A bit of history

- Innovation: First Project Charter on *Heterogeneous Architectures Testbed* presented at the IRB #1
  - M. Girone et al. – June 2022 https://indico.cern.ch/event/1170829/
- ARB: Public Cloud Usage Framework v. 2.0 includes the *Testbed for new architectures* as an “Initial Use Case”
  - D. Van der Ster et al. - Oct 2022 + several iterations in the following months, final version approved by the DHO in January 2023 (https://it-arb.web.cern.ch/frameworks/IT-Cloud-Framework-v2.1.pdf)
- Engagement: *Access and Support for non-x86 Architectures* presented at the RCS-ICT #2
  - A. Wiebalck – February 2023 https://indico.cern.ch/event/1233298/
- Engagement: PSO presented at the RCS-ICT Steering Committee
  - A.Wiebalck, M. Girone, J. Blomer, D. Piparo – March 2023

https://docs.google.com/document/d/14L_2zKwb1OeEJYGgGzpEOapVnn5pcEYIs5fbZcggtG4/

Stakeholders: Innovation, ARB, Engagement, CERN IT, CERN-EP, LHC experiments, CERN-TH, ATS

Purpose of the proposal

- Heterogeneous and non-x86 architectures have become increasingly important in high energy physics
  - GPUs provide the massive parallelism needed for modern data processing and ML algorithms
  - Processors with non-x86 architectures such as ARM, RISC-V come with a better event/Watt output and are on the roadmaps for the LHC experiments' online
  - Programmable hardware such as FPGAs offer high performance and low latency for the specific tasks of pre-processing, triggering and filtering
- Integrating, building and testing data processing algorithms on multiple platforms is instrumental to guarantee high code quality standards
- Maximise the chances to get advantage of non-WLCG resources such as High-Performance Computing centres
- Provide access to rapidly evolving, cutting-edge technologies to demonstrate capabilities before making capital commitments

Overall Goals

- Since the original “testbed” proposal in June 2022, many more discussions have happened resulting recently in the Engagement PSO bringing additional requirements beyond the initial purpose.

- CERN IT is asked to provide the user communities with flexible access and support to on-prem and remote infrastructures for heterogeneous and non-x86 resources, with the aim to cover (from the PSO):
  - **evaluation,** granting preview access for early assessment of the usefulness of a given technology, including a basic software ecosystem; where applicable, have direct contact with the providing hardware companies or technology experts; *(this is the original scope of the testbed)*
  - **development and porting of applications:** adequate provisioning of identified heterogeneous and non-x86 resources as part of IT infrastructure services to ensure a low entry threshold and a stable platform; to be used for all development phases: building (main and nightlies), testing, debugging (e.g. via interactive access), and release into official repositories;
  - **time-limited production needs:** ensure quick technology turn-around and address short-notice requirement bursts.

How would it work?

- Common “access platform” providing support for different activities to the users, possibly through a common portal to gather requests and list availability of devices/services
- Resources provided via different mechanisms

1. **Industry collaborations, CERN openlab partnerships** (example: E4/NVIDIA project: access to next generation processors)
2. **HPC supercomputer testbeds**
3. **Emerging** IT offerings, such as the **IT Public Cloud initiative** (example: GPUs/DPUs)
4. **Established IT services**, such as OpenStack, GitLab, or LxPlus (example: ARM VMs or runners)

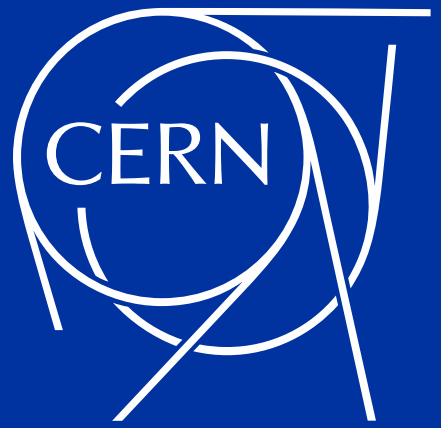
More specifically for the development “testbed”...

1. **Industry collaborations, CERN openlab partnerships**
2. **HPC supercomputer testbeds**
3. **Testbed for New Architectures (IT Cloud Framework v 2.1)**

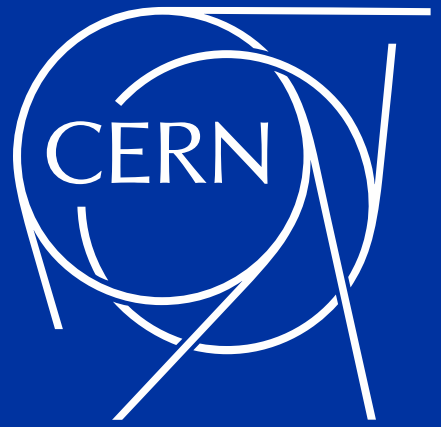
- **on-prem/remote via CERN openlab (e.g. CERN openlab E4/Nvidia, Intel, IBM projects)**
  - NVIDIA CPUs and GPUS, including GRACE and HOPPER
  - ARM Marvell and Ampere CPUs
  - RISC-V CPUs
  - Intel CPUs (Sapphire Rapids, Ice Lake, ...) and GPUs
  - Power
  - FPGAs
  - AI-specialised architectures
  - Quantum simulators
- **remote, via HPC supercomputers testbeds**
  - Leveraging on the strong connection with PRACE and EuroHPC
- **remote, via commercial cloud-hosted systems**

Personnel Resources Needs

- Currently under discussion, as it goes beyond the testbed discussion
- Key: ensure mid-term/long-term continuity of the activity at an engineering level
- Thoughts to support at least initially the testbed (input from IT-FA)
  - experienced Computing Engineer benefiting from a 5-year LD contract
  - (junior) fellow



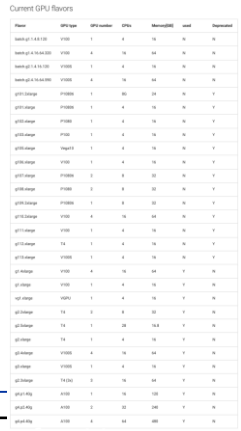
Discussion



Backup Slides

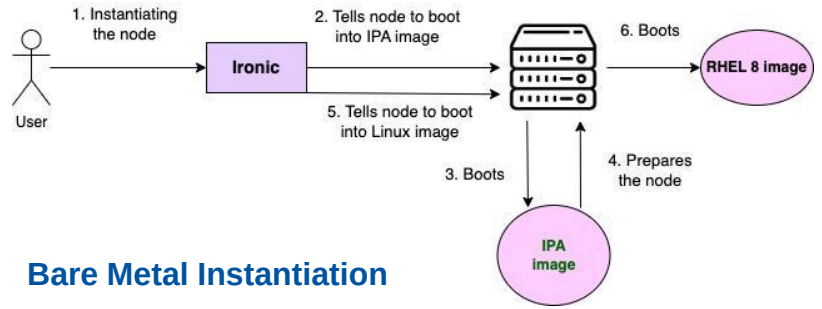
(input collected by A. Wiebalck for the WLCG workshop, CHEP2023)

On-premises IaaS on OpenStack

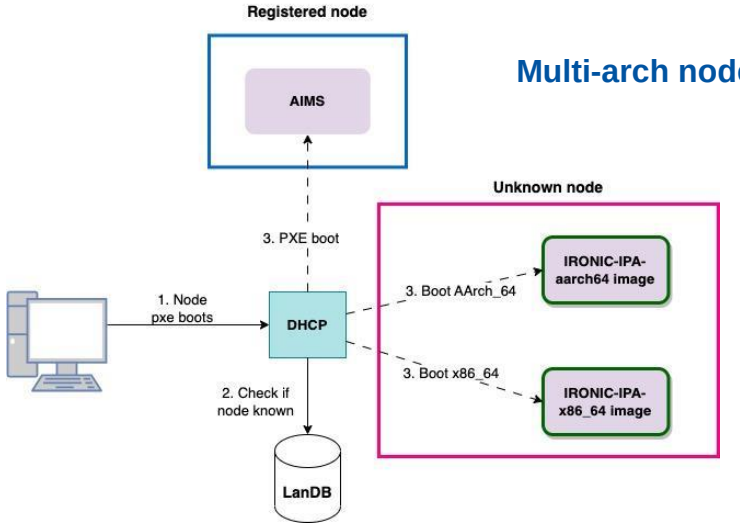
Resource	Spec	Available (Total)	Access	Used by
ARM Altra "Mt. Snow"	v8.2 Neoverse-N1 2.8GHz, 256GB	4 (5)	Via VMs and services	all LHC experiments (CI) HEP benchmarking, SIS EP: CernVM-FS, SFT IT: Linux, gitlab, lxplus, Ceph
GPU	Total cards: Model / nodes / cards P100 / 1 / 1 V100 / 5 / 17 V100S / 6 / 24 T4 / 73 / 76 A100 / 18 / 72		Via Services (next slide)	V100(S): batch T4: batch, SWAN, ML, TTaaS, … A100: batch, ML

On-premises IaaS Integration: ARM

- **Virtual Machines … "easy"**
  - Required EL8
  - Image and capabilities filtering
  - Libvirt bug
- **Physical servers … "not so easy"**
  - Image boot-strapping with AArch64 QEMU
  - Ironic as CERN IT's fleet manager
  - Multi-arch docker images & PXE
  - Detailed talks at HEPiX & CHEP



Bare Metal Instantiation



Multi-arch node discovery

On-premises IaaS Integration: GPUs

- ➔ **Various hardware models**
- ➔ **Various access modes**
- ➔ **Used by higher level services**

Provisioning	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 GPUs

- ⊕ 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 GPUs

- ⊕ 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 GPUs

- ⊕ 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

On-premises: GPU access

- https://clouddocs.web.cern.ch/gpu/index.html
- GPUs are available …
  - as virtual machines
  - via the batch service
  - on kubernetes clusters
  - via Lxplus
  - via Gitlab
  - via Swan
  - ml.cern.ch
- Request for GPUs: GPU Platform Consultancy FE
- Mattermost for GPUs: ~GPU

Public Cloud: Oracle Cloud

<table>
<tr><th>Resource</th><th>Spec</th><th>Amount</th><th>Access</th><th>Used by</th></tr>
<tr><td>Ampere Altra
AMPERE</td><td>BM.Standard.A1.160
160 cores, 1TB RAM (3.0GHz)</td><td>~10 VMs 1 physical</td><td rowspan="2">Via keys on pre-created instances</td><td>{ALICE, ATLAS} builder
WLCG benchmarking
IT: lxplus & lxbatch, GitLab, Linux, Monit</td></tr>
<tr><td colspan="3">- VM.GPU2.1 12 72. (Nvidia Tesla P100 16GB, 1 per node)
- BM.GPU.A10.4 64 1024 (Nvidia A10 Tensor Core, 4 per node)
- BM.GPU2.2 28 256 (Nvidia Tesla P100 16GB, 2 per node)</td><td>(not used at the moment)</td></tr>
</table>

Collaborations: openlab (1)

Resource	Spec	Amount	Access	Used by
GPU NVIDIA	V100 T4	20 8	Direct on (shared) bare metal nodes	openlab projects (ML, QTI, Medical)
GPU intel	ATS-P	1		
DPU* NVIDIA	Bluefield-2	2		

*DPU: Data Processing Unit combining CPUs, network links, and programmable accelerators.

Collaborations: openlab (2)

Resource	Spec	Amount	Access	Used by
ARM 	ThunderX	2	Direct on (shared) bare metal nodes	Build nodes for SFT & experiments
	ThunderX2	3		Linux team
Power 	Power8	4		CMS & LHCb build nodes
	Power9	2		-

➔ **Power8 Minsky server acquired.**

➔ **Future of Power9 nodes to be discussed**