CERN IT GPU Update

Ricardo Rocha, CERN IT

Compute Accelerator Forum, June 8th 2022
https://indico.cern.ch/event/1073643/

Reminder

https://clouddocs.web.cern.ch/gpu/index.html

Currently T4 and V100 GPUs

GPU availability on:

- Virtual Machines, Batch, Kubernetes clusters
- Higher level services: *lxplus-gpu, GitLab runners, SWAN*

Request for GPUs: *GPU Platform Consultancy Functional Element*

- https://cern.service-now.com/service-portal?id=functional_element&name=gpu-platform

#GPU channel on IT-dep mattermost

Access to GPU resources

Server Provisioning Service (Ticket created in FE = GPU Platform Consultancy)

Fill this form for requesting **access to GPU resources**.
If you don't need access but you have another kind of request for the GPU Platform Consultancy, please use **this form** instead.

N.B. it will create a ticket directly into "GPU Platform Consultancy" 2nd level.

Usage pattern expected (spiky if <30% overall usage, full if >80%)

(•) Spiky ( ) Medium ( ) Full

Specific performance requirements for floating point precision

(•) Double ( ) Single ( ) None

Type of interface desired

(•) Notebook ( ) Batch ( ) Kubernetes ( ) VM ( ) Other

Openstack project name (required for Kubernetes and VM)

CUDA drivers and versions required (custom if you need specific drivers)

(•) Custom ( ) Any

ML framework being used (for machine learning workloads only)

(•) Tensorflow ( ) PyTorch ( ) scikit-learn ( ) Other

Distributed training possible or desired (for machine learning workloads only)

No

* Number of GPUs required

* Project Description (overview, purpose, software, specific requirements)

Usage

Still getting GPU access requests at a steady pace

- Many requests forwarded to higher level services (batch, notebooks, …)
- Improved resource sharing, preferable vs directly allocation

Overall we're always full regarding assignment of available resources

- But keep flexibility so we can accommodate big requests (tutorials, …)
- Example with recent CERN School of Computing workshop
- Please release resources when no longer needed

What's New

Action items from the last update in this forum (Oct 2021)

https://indico.cern.ch/event/975015/

Upgrade Nvidia drivers (470.82.x) for CUDA 11.4 (done)

Support for GPU profiling in vGPU nodes (done)

Reminder: vGPU is not physical partitioning but time sharing up to 4x

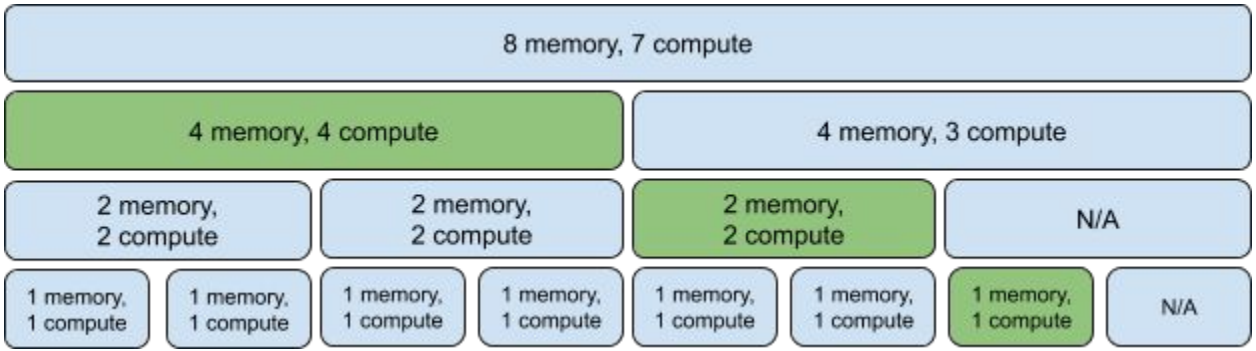
Profiling possible with vGPUs - with new drivers (done)

General availability of vGPU setup (instabilities detected)

Ongoing Work

Focus on getting the **new A100 GPUs** available

- ~70 new cards, physical split possible up to 7x
- Very significant delays due to supply chain issues - expected end of July 2022
- Support for Multi-Instance GPU (MIG) - done for Kubernetes, OpenStack unclear



Ongoing Work

Initiative approved in the ATS-IT collaboration in the GPU area

- Starting now: https://indico.cern.ch/event/1142396/
- Consolidate procurement and improve overall resource usage
- Catalog short, medium, long term requirements for GPUs and accelerators
- Adapt service offering to these needs
- Prototype hybrid deployments where appropriate

Questions?