

CERN IT GPU Update

Ricardo Rocha - IT-CM-RPS

Compute Accelerator Forum, Oct 13th 2021

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

Reminder

<https://clouddocs.web.cern.ch/gpu/README.html>

GPU availability on virtual machines, batch, kubernetes clusters, ...

And higher level services: *lxplus*, *gitlab*, *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)

* Number of GPUs required

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

Usage

Requests for GPU access coming at a steady pace

Spread over different departments and projects

Many requests forwarded to higher level services (batch, notebooks, ...)

Overall we're always full regarding assignment of available resources

But keep flexibility so we can accommodate big requests (tutorials, ...)

Dedicated assignments are made with limited duration for non shared services

What's New

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

<https://indico.cern.ch/event/975007/>

GPUs on lxplus: done - lxplus-gpu.cern.ch

GPUs on Gitlab CI (as shared runners): done

Other items

GPUs on SWAN (done)

GPUs on baremetal (done, possible on request)

Ongoing Work

Preparing support for Multi-Instance GPU (MIG)

Kubernetes (**done**) and OpenStack (**ongoing**)

Adding support for GPU profiling in vGPU nodes

vGPU is not physical partitioning but time sharing up to 4x != MIG

Lack of support for profiling has been a stopper for general usage of vGPUs

Profiling possible with the new 13.x Nvidia drivers (already validated)

Upgrade default drivers in our recipes and automated installations

Adding support for CUDA 11.1/11.2, popular request

Questions?