GPUs in CERN IT

Ricardo Rocha
CERN IT-CM-RPS

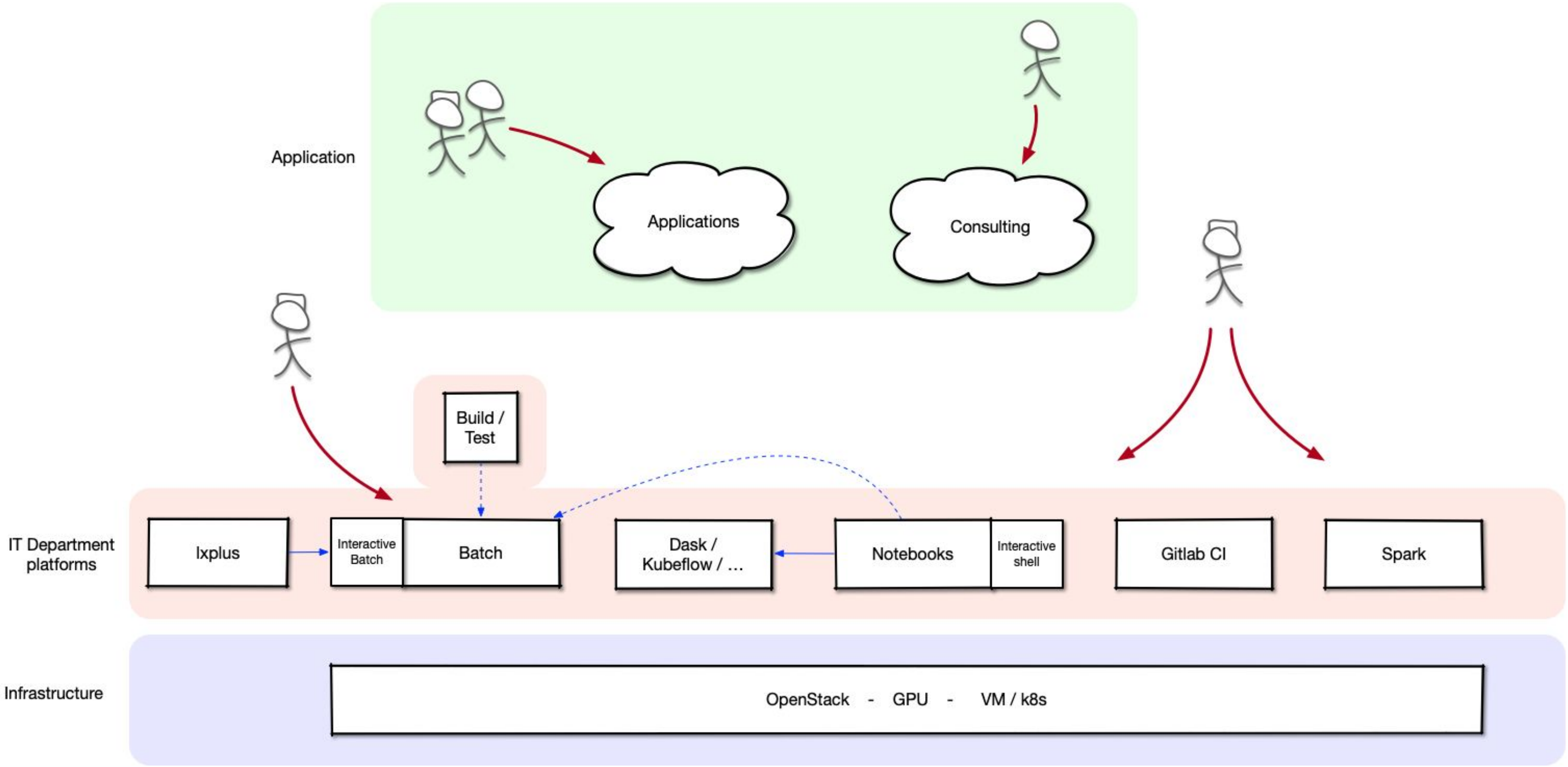
Compute Accelerator Forum - Oct 1st 2020

Goal

Make GPUs available to a variety of services

Ensure a common infrastructure allowing flexible (re)assignment of resources

Provide consultancy on GPU usage to our users


Application
Applications
Consulting
Build / Test
IT Department platforms
lxplus
Interactive Batch
Batch
Dask / Kubeflow / …
Notebooks
Interactive shell
Gitlab CI
Spark
Infrastructure
OpenStack - GPU - VM / k8s

Hardware

Focusing on Nvidia enterprise cards: P100, V100(S), T4

Older 1080(ti)s cards have been decommissioned

Current numbers

- V100 32GB: 16x split in 4 nodes
- V100S 32GB: 24x split in 6 nodes
- T4 16GB: 64x in 64 nodes

All with PCIe Gen 3 interconnects - no NVLink

GPU Showdown

T4

SPECIFICATIONS

GPU Architecture	NVIDIA Turing
NVIDIA Turing Tensor Cores	320
NVIDIA CUDA® Cores	2,560
Single-Precision	8.1 TFLOPS
Mixed-Precision (FP16/FP32)	65 TFLOPS
INT8	130 TOPS
INT4	260 TOPS
GPU Memory	16 GB GDDR6 300 GB/sec
ECC	Yes
Interconnect Bandwidth	32 GB/sec
System Interface	x16 PCIe Gen3
Form Factor	Low-Profile PCIe
Thermal Solution	Passive
Compute APIs	CUDA, NVIDIA TensorRT™, ONNX

V100

SPECIFICATIONS

	V100 PCIe	V100 SXM2	V100S PCIe
GPU Architecture	NVIDIA Volta		
NVIDIA Tensor Cores	640		
NVIDIA CUDA® Cores	5,120		
Double-Precision Performance	7 TFLOPS	7.8 TFLOPS	8.2 TFLOPS
Single-Precision Performance	14 TFLOPS	15.7 TFLOPS	16.4 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS	130 TFLOPS
GPU Memory	32 GB /16 GB HBM2		32 GB HBM2
Memory Bandwidth	900 GB/sec		1134 GB/sec
ECC	Yes		
Interconnect Bandwidth	32 GB/sec	300 GB/sec	32 GB/sec
System Interface	PCIe Gen3	NVIDIA NVLink™	PCIe Gen3
Form Factor	PCIe Full Height/Length	SXM2	PCIe Full Height/Length
Max Power Comsumption	250 W	300 W	250 W
Thermal Solution	Passive		
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC®		

Key Differences

Memory, CUDA and Tensor Cores

- x2 in V100

V100 Double Precision

Power Consumption

- x3.5 in V100

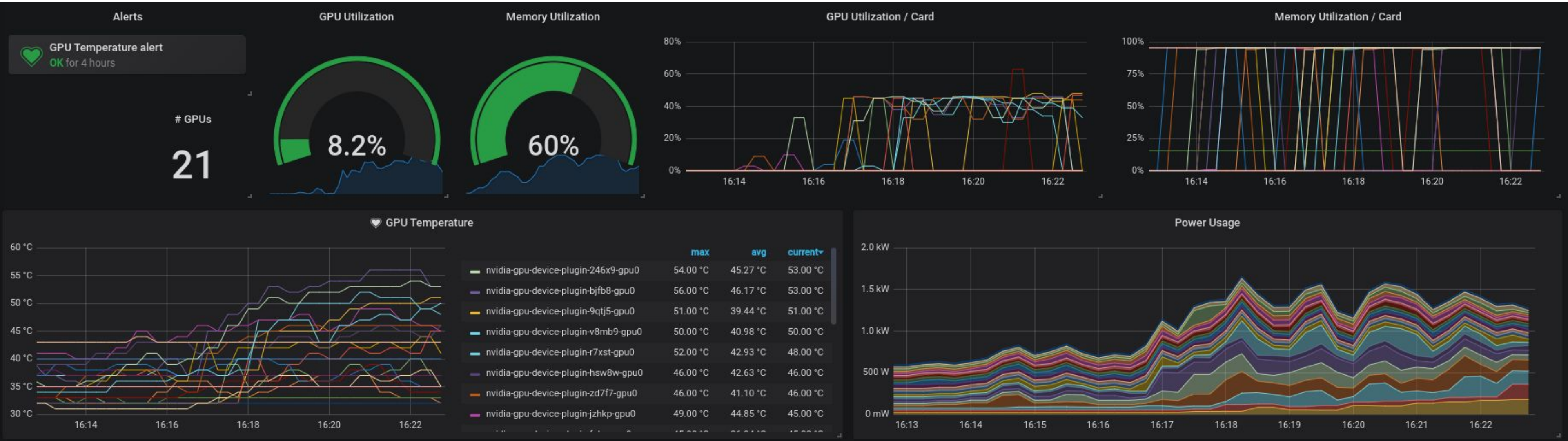
APIs Supported

Alerts
GPU Temperature alert
OK for 2 hours
GPUs
V100S
21
GPU Utilization
71%
Memory Utilization
79%
GPU Utilization / Card
Memory Utilization / Card
GPU Temperature
max avg current
nvidia-gpu-device-plugin-jzhkp-gpu0 76 °C 54 °C 76 °C
nvidia-gpu-device-plugin-hsw8w-gpu0 74 °C 51 °C 74 °C
nvidia-gpu-device-plugin-v8mb9-gpu0 72 °C 50 °C 72 °C
nvidia-gpu-device-plugin-bjfb8-gpu0 72 °C 43 °C 72 °C
nvidia-gpu-device-plugin-r7xst-gpu0 72 °C 50 °C 71 °C
nvidia-gpu-device-plugin-fxwg8-gpu0 69 °C 47 °C 69 °C
nvidia-gpu-device-plugin-9qtj5-gpu0 68 °C 47 °C 68 °C
nvidia-gpu-device-plugin-zbp6p-gpu0 69 °C 47 °C 68 °C
Power Usage
Alerts
No alerts
GPUs
T4
34
GPU Utilization
100%
Memory Utilization
90%
GPU Utilization / Card
Memory Utilization / Card
GPU Temperature
max avg current
nvidia-gpu-device-plugin-2gkqp-gpu0 84 °C 82 °C 84 °C
nvidia-gpu-device-plugin-8x54n-gpu0 85 °C 62 °C 84 °C
nvidia-gpu-device-plugin-xrwnv-gpu0 84 °C 62 °C 84 °C
nvidia-gpu-device-plugin-2xg4x-gpu0 84 °C 64 °C 84 °C
nvidia-gpu-device-plugin-vf9zh-gpu0 77 °C 63 °C 76 °C
nvidia-gpu-device-plugin-6nj66-gpu0 77 °C 63 °C 76 °C
nvidia-gpu-device-plugin-2rn2b-gpu0 76 °C 61 °C 76 °C
nvidia-gpu-device-plugin-dz2n5-gpu0 77 °C 63 °C 76 °C
Power Usage

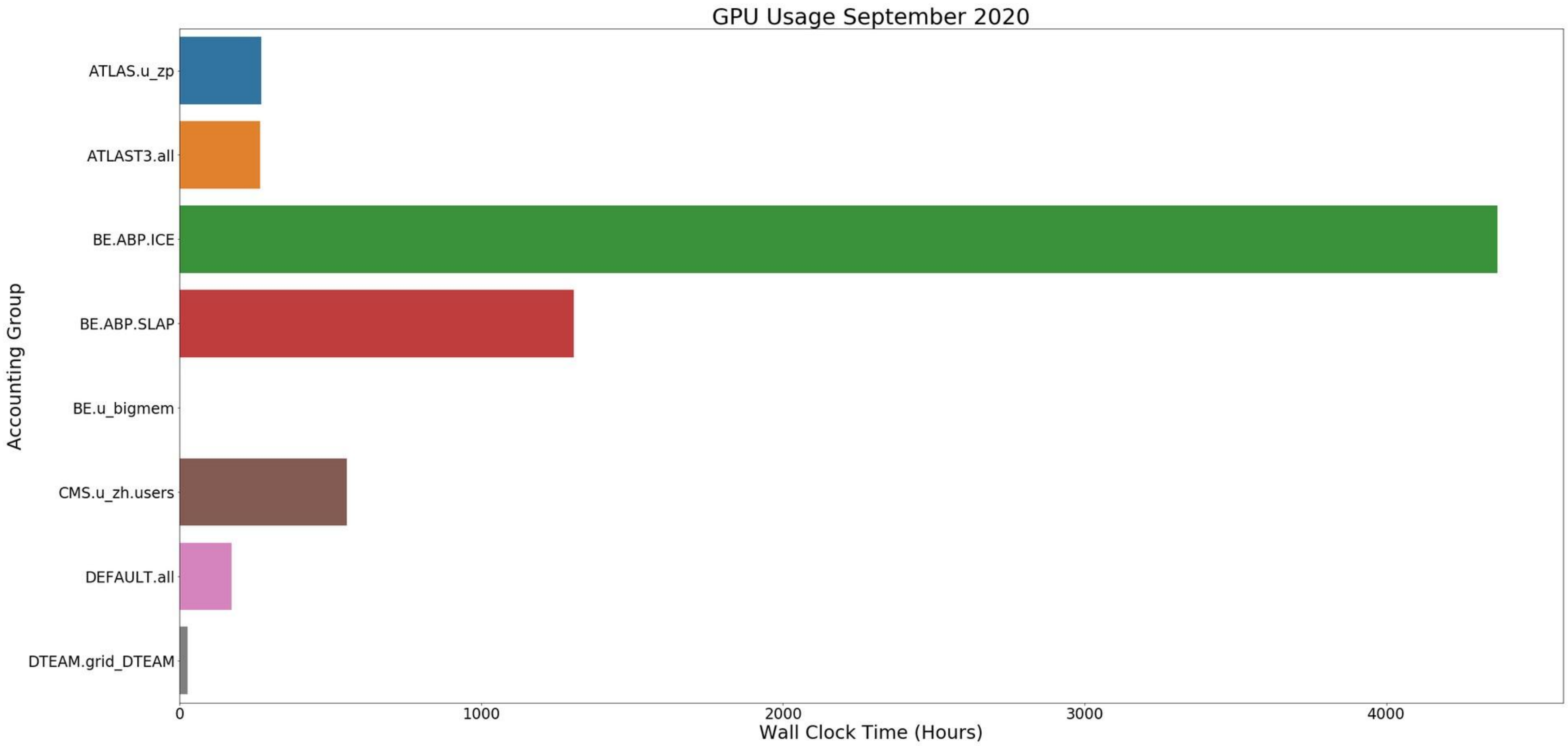
A more realistic workload...

Distributed TensorFlow Job

Example below using 8 out of 21 available GPUs in a cluster



Batch Usage Overview



Access to GPUs

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

Virtual Machines: when full flexibility is required

Custom kernels, modules and drivers

We can help with a base setup - puppet based?

Kubernetes: automated GPU setup and built-in monitoring

Batch: possible to also get an interactive session

GitLab, Notebooks/SWAN, etc: ongoing work

Access to GPUs

Step 1: Request to the *GPU Platform Consultancy Functional Element*

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

Allows us to redirect users to the best solution

And build a knowledge base on the different use cases

Access to GPUs

Step 1: Request to the *GPU Platform Consultancy Functional Element*

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

“ I would like access to GPUs for project X,

we will use them to build our software daily ”

⟶Gitlab?

Access to GPUs

Step 1: Request to the *GPU Platform Consultancy Functional Element*

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

“ We would like access to GPU resources to validate our software

including testing against multiple Nvidia drivers ”

⟶Virtual Machines?

Access to GPUs

Step 1: Request to the *GPU Platform Consultancy Functional Element*

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

“ We would like to use GPUs to train our ML models using

TensorFlow/Pytorch. We could benefit from distributed training. ”

⟶Batch?

Ongoing Work: vGPUs

Currently only possible to offer access to full GPUs - via PCI passthrough

In many cases this is not required and a fraction of a GPU would be enough

- Builds, code validation, notebooks, ...

Working on virtualizing GPU resources to improve overall resource usage

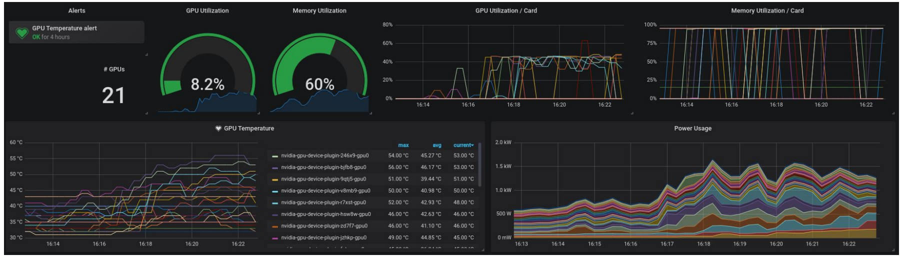
Ongoing Work: Monitoring

Very useful to better understand efficiency

Built-in for Kubernetes clusters, we could offer a solution for VMs as well

Working on how to relate workloads with the monitoring data

Similar feedback to end users for managed services would also be good



Ongoing Work: Vendors

Monitoring the GPU / Accelerator landscape

Today this is mostly a Nvidia (enterprise) world but things might change

Ongoing Wor

Monitoring the GPU /

Today this is mostly a ... ht change

Community supported AMD ROCm build for TensorFlow

TensorFlow Follow
Sep 17, 2019 · 2 min read

A guest post by Mayank Daga, Director, Deep Learning Software, AMD

Deep Learning has burgeoned into one of the most important technological breakthroughs of the 21st century. GPUs have played a critical role in the advancement of deep learning. The massively parallel computational power of GPUs has been influential in reducing the training time of complex deep learning models, hence accelerating the time to discovery. The availability of open source frameworks like TensorFlow is another cornerstone for the fast-paced innovation in deep learning. At AMD, we strongly believe in the open source philosophy and have purposely developed Radeon Open eCosystem (ROCm), an open-source software foundation for GPU computing on Linux®.

Other Possibilities

We have (limited) access to resources on public clouds

Thanks to multiple OpenLab collaborations

What is possible today

Access to A100 GPUs and/or TPUs in the Google Cloud Platform

Access to IPUs in Azure

Contact us if you have a good use case for any of these

Other Possibilities

We have (limited) access to resources on public clouds

- Thanks to multiple OpenLab collaborations

What is possible today

- Access to A100 GPUs and/or TPUs in the Google Clouc
- Access to IPUs in Azure

Contact us if you have a good use case for any of these

SYSTEM SPECIFICATIONS (PEAK PERFORMANCE)

	NVIDIA A100 for NVIDIA HGX™	NVIDIA A100 for PCIe
GPU Architecture	NVIDIA Ampere	
Double-Precision Performance	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOPS	
Single-Precision Performance	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS | 312 TFLOPS*	
Half-Precision Performance	312 TFLOPS | 624 TFLOPS*	
Bfloat16	312 TFLOPS | 624 TFLOPS*	
Integer Performance	INT8: 624 TOPS | 1,248 TOPS* INT4: 1,248 TOPS | 2,496 TOPS*	
GPU Memory	40 GB HBM2	
Memory Bandwidth	1.6 TB/sec	
Error-Correcting Code	Yes	
Interconnect Interface	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**
Form Factor	4/8 SXM GPUs in NVIDIA HGX™ A100	PCIe
Multi-Instance GPU (MIG)	Up to 7 GPU instances	
Max Power Consumption	400 W	250 W
Delivered Performance for Top Apps	100%	90%
Thermal Solution	Passive	
Compute APIs	CUDA®, DirectCompute, OpenCL™, OpenACC®	

* Structural sparsity enabled

** SXM GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to 2 GPUs

Questions ?