Exploiting GPU Resources at VEGA for CMS Software Validation

Daniele Spiga[1], Adriano Di Florio[2], Andrea Bocci[3], Antonio Perez-Calero Yzquierdo[4,5], Christoph Wissing[5], Jose Hernandez[4]

[1]INFN, [2]IN2P3, [3]CERN, [4]CIEMAT, [5]PIC, [6]DESY

This Talk

HPC resources integration at CMS: brief intro

CMS Grant at VEGA EuroHPC in Slovenia

- Motivation
- Strategy
- Results

Summary and Lessons

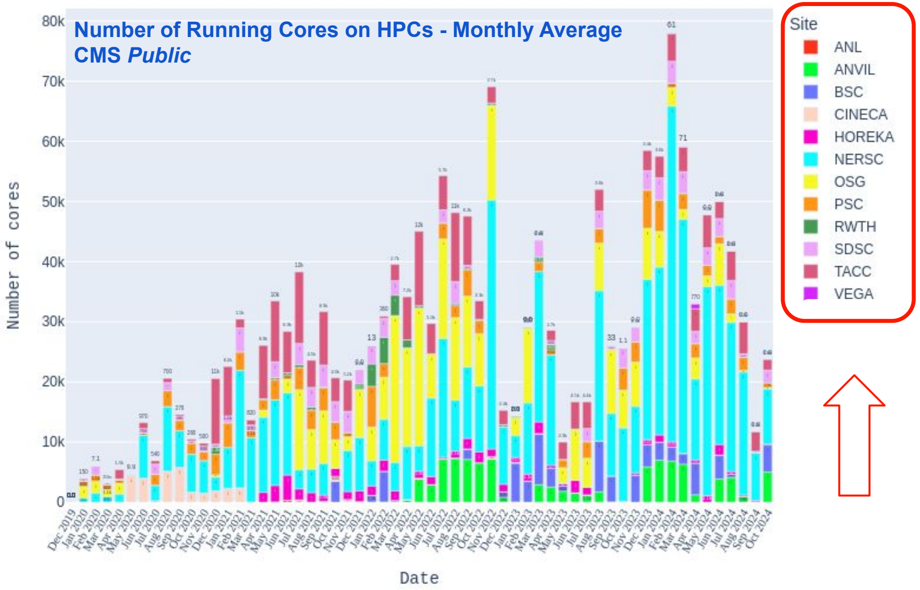
HPC integration in the CMS Computing

HPC integration one of the key assets of the CMS Computing: a number of HPC machines has been integrated and continuously been used in production mode throughout the year 2020.

Distinct strategies developed and deployed:

1. **Overlay batch model**
2. **Site Extension**
3. **HTCondor split starter mechanisms for filesystem based communication**

CMS wants to further **increase the HPC exploitation, particularly in the EU Zone**



spiga@pg.infn.it

The CMS Grant at VEGA HPC in Slovenia

VEGA is the first of eight peta and pre-exa-scale EuroHPC Joint Undertaking.
Hosted at IZUM in Maribor

- 960 CPU nodes
  (overall 1920 CPUs AMD Epyc 7H12 – 122000 cores)
- 60 GPU nodes (overall 240 GPUs NVidia A100)
- Sustained performance of HPC Vega is 6,9 PFLOPS (peak performance is 10.1 PFLOPS).

EuroHPC Application

ID: EHPC-BEN-2023B12-002

Project Application - Jose Hernandez

The Project

Project details

Project title: Running CMS simulation workflows on a high-performance computing platform

Project summary (abstract):

The Compact Muon Solenoid (CMS) is one of the two general-purpose particle physics detectors
European Laboratory for Particle Physics (CERN). The LHC experiments are facing unprecedente
substantial surge in data volume and complexity. Significant additional compute and storage resou
process and analyze the large volume of recorded data and to produce the required simulations o

High-Performance Computing Centers (HPC) stand out among the largest processing resources
Unlocking the full potential of these world-class HPC facilities holds the key to significantly boostin
processing endeavors, addressing the critical need for expanded computing capacity. Ongoing Re
are actively underway to harness the computational capacity offered by these facilities and propel
model, seamlessly integrating their capabilities. This project aims to benchmark the execution of C
showcasing the optimal utilization of HPC resources for High Energy Physics computing.

Explain the scientific case of the project for which you intend to use the code(s):

Particle physics is the study of the most elementary components of matter and their interactions. T
particle accelerators, which bring particles such as protons into high-energy collisions. The Large
world's highest-energy accelerator. This very high collision energy allows for the possibility of the p
seen in the laboratory before, giving the LHC great
opportunities for discovery. The Compact Muon Solenoid (CMS) is one of the two general-purpose
CMS Collaboration co-discovered the Higgs boson in 2012, has provided constraints on many mo
precise measurements of the properties of known particles.

The LHC community has strategically outlined a roadmap to proactively address impending shifts
storage infrastructures, software, and technologies. This forward-looking plan is crucial for navigati
challenges that the LHC experiments will encounter over the next decade. In the current data-takir
demand for computational resources necessitates primarily evolutionary changes. This is due to th
collisions delivered compared to Run 2 (2015-2018). However, the prospect of the high luminosity
a formidable challenge, with CMS set to collect approximately 20 times more data volume than the

EuroHPC Proposal Approval Information for HPC Vega

Access Call Information:

Call Information:	https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-ju-benchmark-an development-access-calls/
Type/Cut-off Date:	Benchmark, 01/12/2023
System, Hosting Entity:	HPC Vega, Institute of Information Science, Maribor, Slovenia
Name of HE coordinator:	Dejan Valh, dejan.valh@izum.si

Proposal Information:

Proposal ID:	2023B12-002
Project name:	Running cms simulation workflows on a high-performance computing platform
Application type/domain:	Fundamental Constituents of Matter
Duration of the project:	3 month(s)
Partitions requested:	Vega GPU
Total resources requested:	1TB
Primary Investigator (PI) info:	Jose Hernandez, male, jose.hernandez@ciemat.es
PI organisation, country:	Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain

Approval Information:

Evaluation comments:	Use of cluster is technically feasible.
Evaluation result:	ACCEPTED

CMS successfully got a grant of type "benchmark"

Why a grant at VEGA: main motivations

Although Slovenia is not member of CMS Collaboration

An excellent opportunity to prove the flexibility of the CMS computing system by several dimensions:

- Integration of a **world class HPC center logical partition of a existing WLCG site**
  - Demonstrate the feasibility of a transnational site extension
- Exploiting **large fraction of opportunistic accelerated nodes**
  - To dynamically extend the pool of GPUs already available to CMS Offline Computing
- Proving the capability **to execute "any type of workflow" at HPC sites**
  - To keep achieving experience with operational model for a long term sustainability

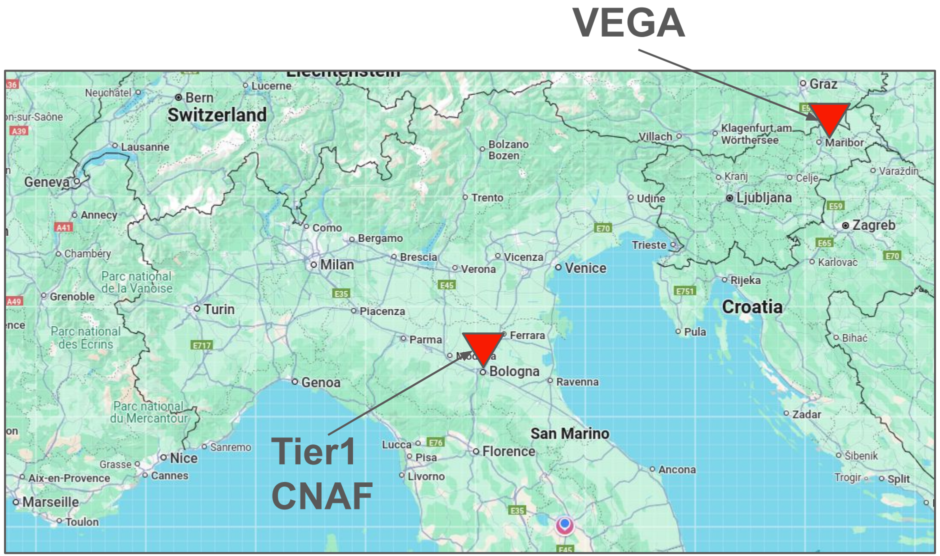
Moreover:

- **A valuable "playground" to contribute to the GPU initiatives of the CMS Software**
  - Software - Computing interplay

Transnational Site Extension

VEGA transparently integrated as a sub-site extension to the Italian Tier-1 site at CNAF.

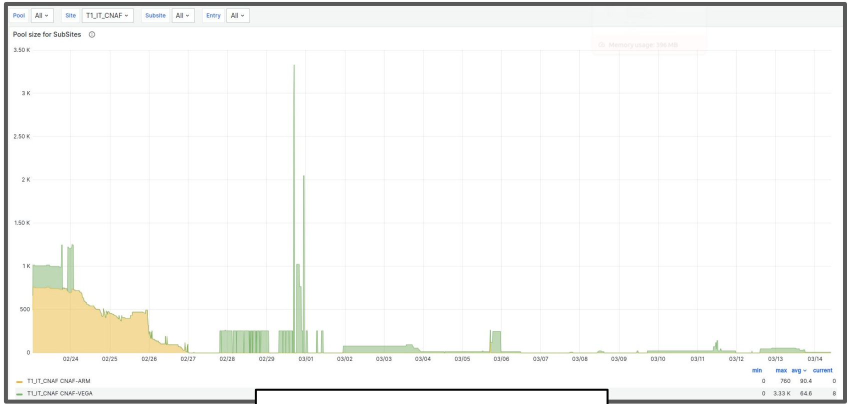
- **Storage-less** site relying on
  - **Remote data access** via xrootd federation (AAA)
  - **Stage out to CNAF storage**
- Regular **CMS Pilot but started "manually"** via slurm at VEGA
  - Pressure both manually and via interLink (**See ID: 511**)
- Squid, Apptainer, CVMFS available at site
- **Local scratch are on NVMe**



First time we successfully integrated an HPC resources located in a different country from the Grid site

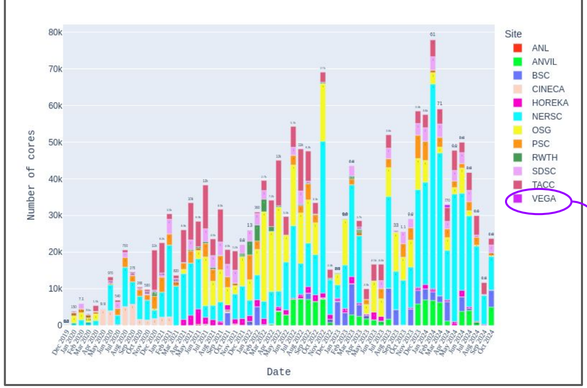
VEGA in the CMS Stack

CMS Job Monitoring

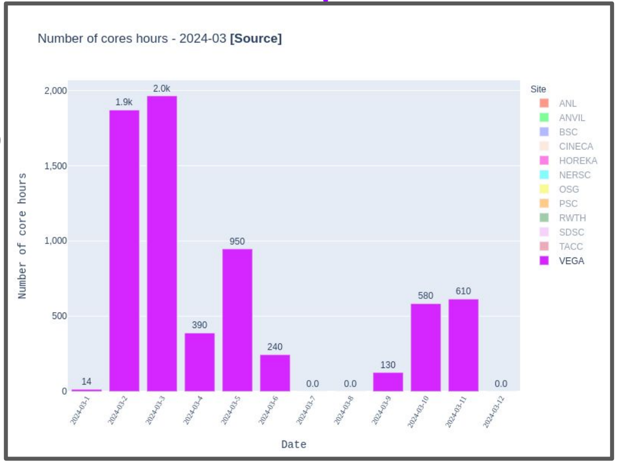


Commissioning phase

HPC Usage @CMS since 2020



CMS HPC Monitoring



Vega seamlessly joined the CMS production stack

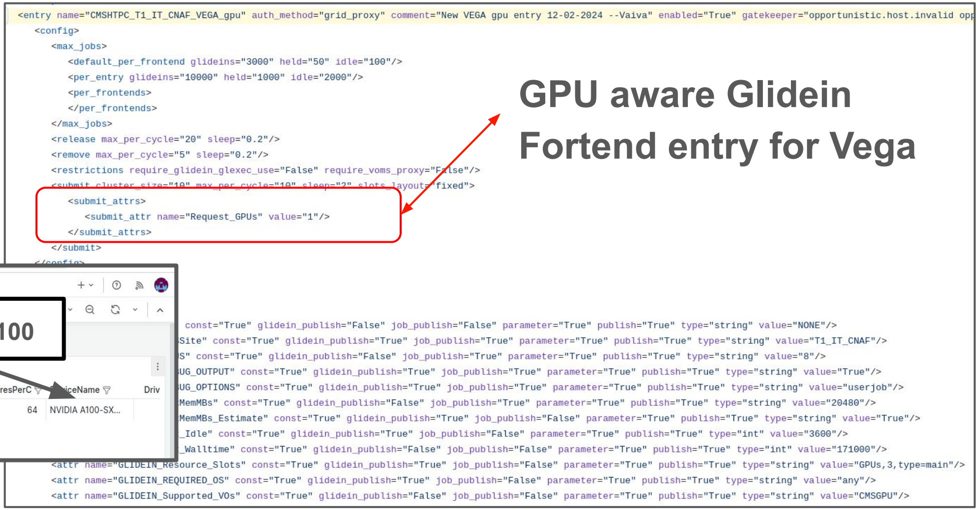
CMS operations team does not “see” Vega, it only targets Tier1 CNAF

- A key to the sustainable operations

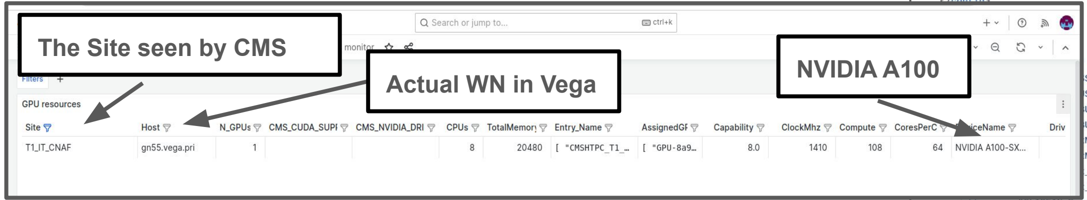
Vega and the CMS Pool of GPUs

Successfully extended the GPU pool

- Vega GPUs are seen as CNAF

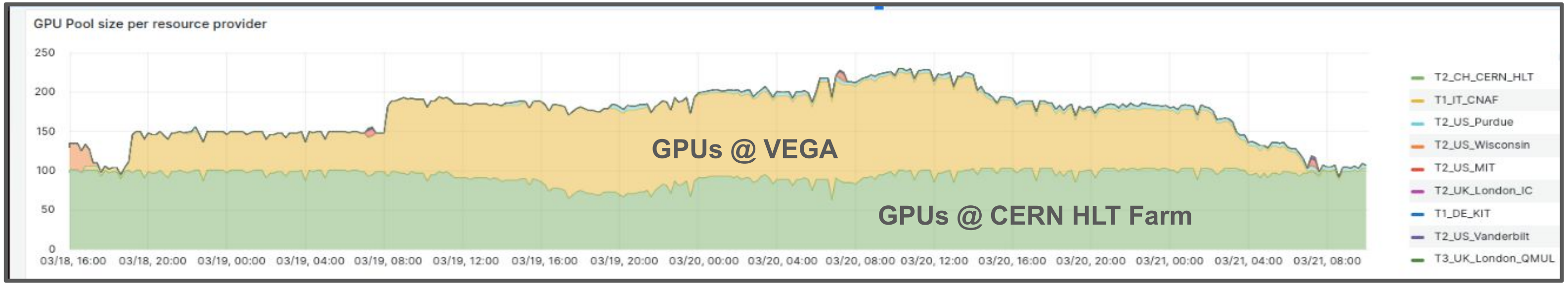


```
<entry name="CMSHTPC_T1_IT_CNAF_VEGA_gpu" auth_method="grid_proxy" comment="New VEGA gpu entry 12-02-2024 --Vaiva" enabled="True" gatekeeper="opportunistic.host.invalid opp
  <config>
    <max_jobs>
      <default_per_frontend glideins="3000" held="50" idle="100"/>
      <per_entry glideins="10000" held="1000" idle="2000"/>
      <per_frontends>
      </per_frontends>
    </max_jobs>
    <release max_per_cycle="20" sleep="0.2"/>
    <remove max_per_cycle="5" sleep="0.2"/>
    <restrictions require_glidein_glexec_use="False" require_voms_proxy="False"/>
    <submit cluster_size="10" max_per_cycle="10" sleep="2" slots_layout="fixed">
      <submit_attrs>
        <submit_attr name="Request_GPUs" value="1"/>
      </submit_attrs>
    </submit>
  </config>
 const="True" glidein_publish="False" job_publish="False" parameter="True" publish="True" type="string" value="NONE"/>
Site" const="True" glidein_publish="True" job_publish="True" parameter="True" publish="True" type="string" value="T1_IT_CNAF"/>
S" const="True" glidein_publish="False" job_publish="True" parameter="True" publish="True" type="string" value="8"/>
UG_OUTPUT" const="True" glidein_publish="True" job_publish="True" parameter="True" publish="True" type="string" value="True"/>
UG_OPTIONS" const="True" glidein_publish="True" job_publish="True" parameter="True" publish="True" type="string" value="userjob"/>
MemMBs" const="True" glidein_publish="False" job_publish="True" parameter="True" publish="True" type="string" value="20480"/>
MemMBs_Estimate" const="True" glidein_publish="True" job_publish="False" parameter="True" publish="True" type="string" value="True"/>
_Idle" const="True" glidein_publish="True" job_publish="False" parameter="True" publish="True" type="int" value="3600"/>
_Walltime" const="True" glidein_publish="False" job_publish="False" parameter="True" publish="True" type="int" value="171000"/>
<attr name="GLIDEIN_Resource_Slots" const="True" glidein_publish="True" job_publish="False" parameter="True" publish="True" type="string" value="GPUs,3,type=main"/>
<attr name="GLIDEIN_REQUIRED_OS" const="True" glidein_publish="True" job_publish="False" parameter="True" publish="True" type="string" value="any"/>
<attr name="GLIDEIN_Supported_VOs" const="True" glidein_publish="False" job_publish="False" parameter="True" publish="True" type="string" value="CMSGPU"/>
```

Site	Host	N_GPUs	CMS_CUDA_SUPI	CMS_NVIDIA_DRI	CPUs	TotalMemory	Entry_Name	AssignedGF	Capability	ClockMhz	Compute	CoresPerC	…iceName	Driv
T1_IT_CNAF	gn55.vega.pri	1			8	20480	["CMSHTPC_T1_…	["GPU-8a9…	8.0	1410	108	64	NVIDIA A100-SX…	

CMS Submission Infrastructure Monitoring



Workflows selection: the rationale

Commissioning phase: Clones of CMS regular workflow (CPU only)

- Standard MC workflow with remote read of the pre-mixed pileup; Data reprocessing with remote read of primary input

Production phase: Release Validation workflows of Alpaka-based version of the HLT.

Our choice to achieve the best out of the Grant

- We use the Vega GPU-equipped nodes to execute online workflow
- **We contributed to Alpaka** (**GPU**) validation campaign needed for the HLT in view of the 2024 data taking
  - With the start of Run3, CMS has successfully offloaded part of the online reconstruction on NVIDIA GPUs (with CUDA). This includes the reconstruction of HLT pixel tracks and vertices.
  - From 2024 data taking, CMS has chosen Alpaka as performance portability library in order to target different CPUs and GPUs with a single code base
  - Alpaka is an header-only library that provides portability of code for various backends by adding an abstraction layer for the backends parallelism. Also, the performance w.r.t. native CUDA have stayed basically untouched.

Validation strategy: two setups

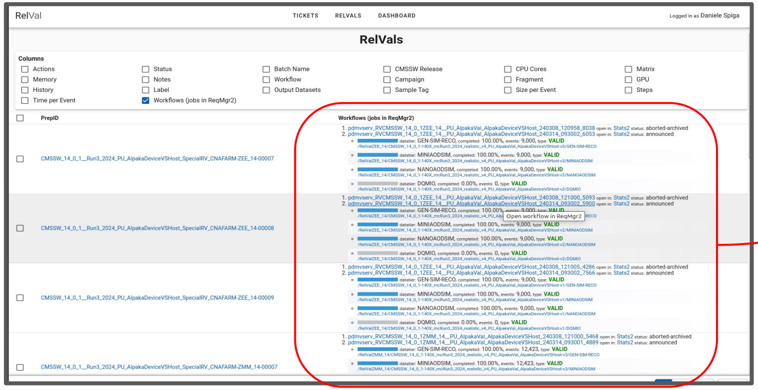
Running the HLT reconstruction on the same input events (RAW) with four setups running the HLT pixel tracks:

- pre-Alpaka migration on CPU,
- pre-Alpaka migration on GPU with native CUDA
- on CPU with Alpaka;
- on GPU with Alpaka;

- Comparing the four outputs.

Running the Alpaka (CUDA) GPU and CPU reconstruction in the same job and for each event.

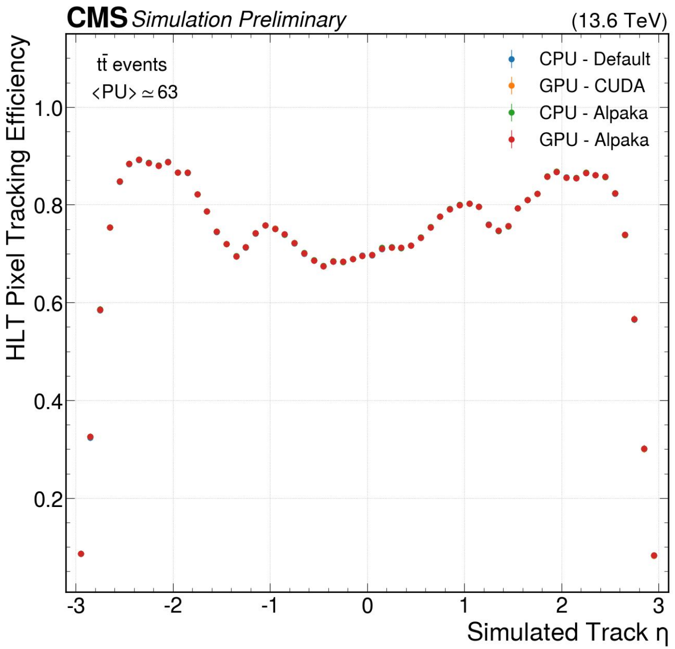
- Comparing GPU and CPU quantities event by event.

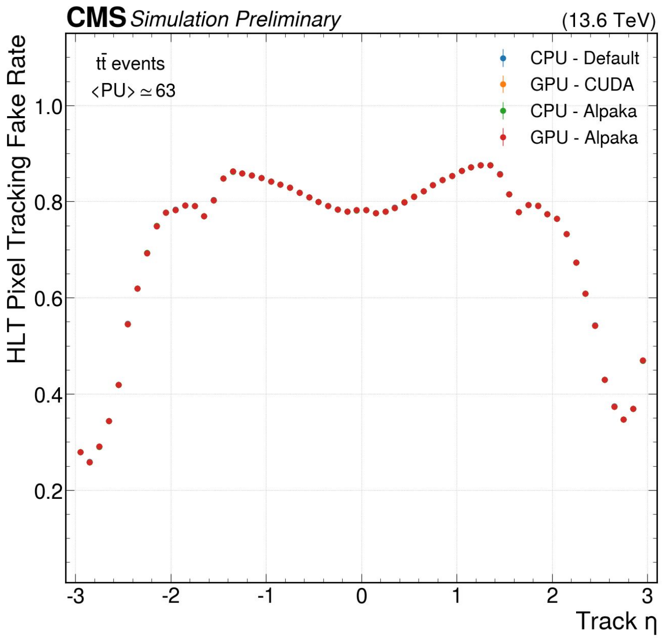


Employing the established production pipeline (RelMon)

Results - Validation I

Perfect match between the four setups (CPU,native CUDA, Alpaka CPU and Alpaka GPU).

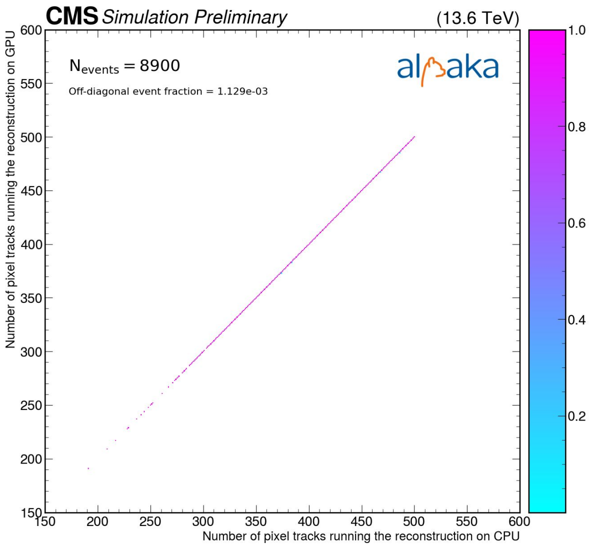


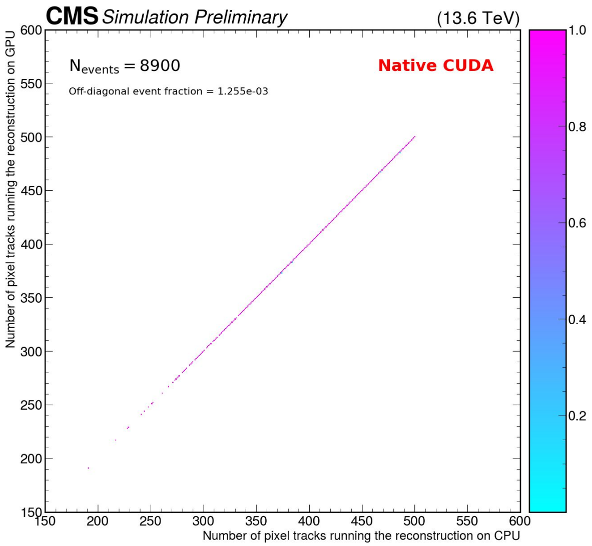


Results - Validation II

GPU vs CPU comparison with Alpaka compatible with native CUDA with the same (minor) fluctuations.

- In the plots comparison of the number of pixel tracks event by event.
  - Each bin is normalised by the number of entries of the column to which it belongs.
- The x-axis (y-axis) represents the number of pixel tracks running the reconstruction on CPU (GPU) with Alpaka/CUDA.
- **Very good agreement!**





Summary

CMS applied for a development grant at VEGA Slovenian EuroHPC

- Not huge amount of resources but a key system to challenge CMS
- Extremely valuable to fine tune dynamic resource integration process as well as to contribute to the GPU mission at CMS

CMS computing system confirms to be high flexible and agile in accommodating unconventional resources

- As well as heterogeneous architectures

The success of this first Vega integration represents a track record for a higher scale exploitation

Acknowledgments

This work was possible also thanks to the support made available by VEGA admins and in particular by A. Filipcic and T. Prica