Towards a Modular Supercomputer Architecture: The DEEP-EST Project

Viktor Khristenko (CERN), Maria Girone (CERN)

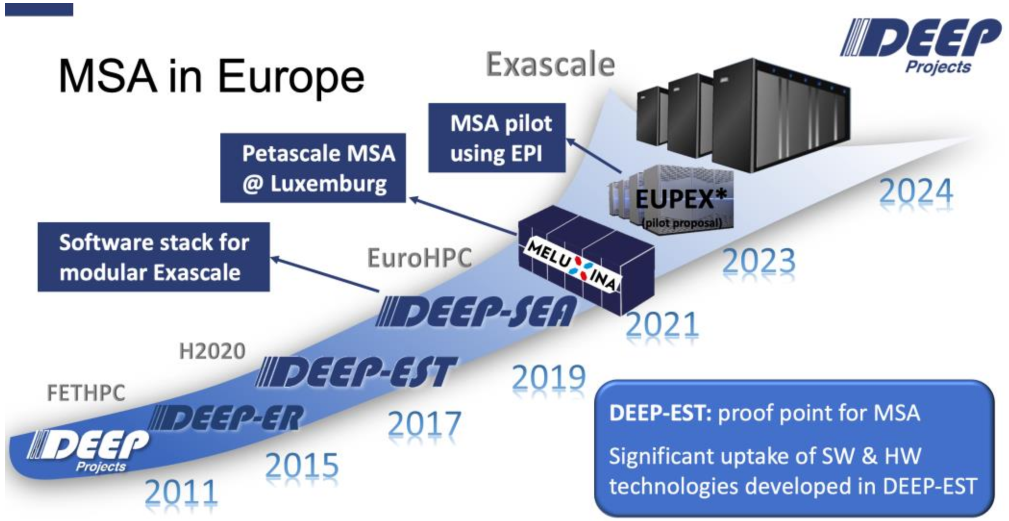
DEEP-EST Modular Supercomputer

- Prototype for the Modular Heterogeneous HPC system

- Convergence of HPC and HPDA worlds

- Variety of hardware to enable wide range of applications

- Software Hardware co-design driven by 6 applications (reconstruction in CMS included)



MSA in Europe

Exascale

MSA pilot using EPI

Petascale MSA @ Luxemburg

Software stack for modular Exascale

EUPEX* (pilot proposal)

EuroHPC

MELUXINA

DEEP-SEA

H2020

DEEP-EST

FETHPC

DEEP-ER

DEEP Projects

2011

2015

2017

2019

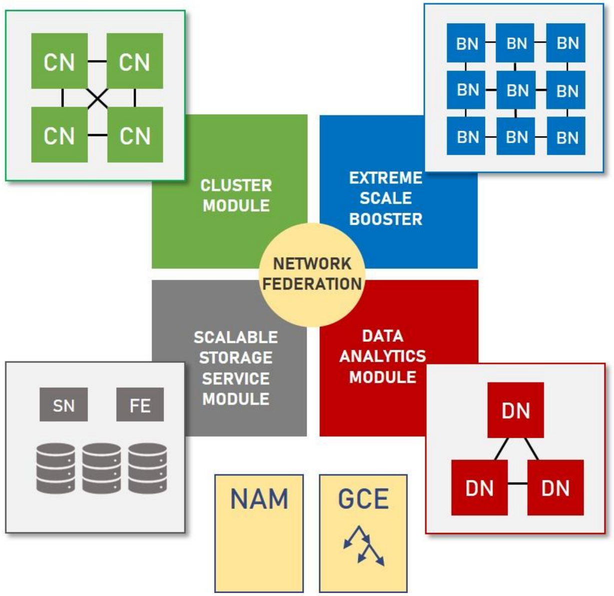
2021

2023

2024

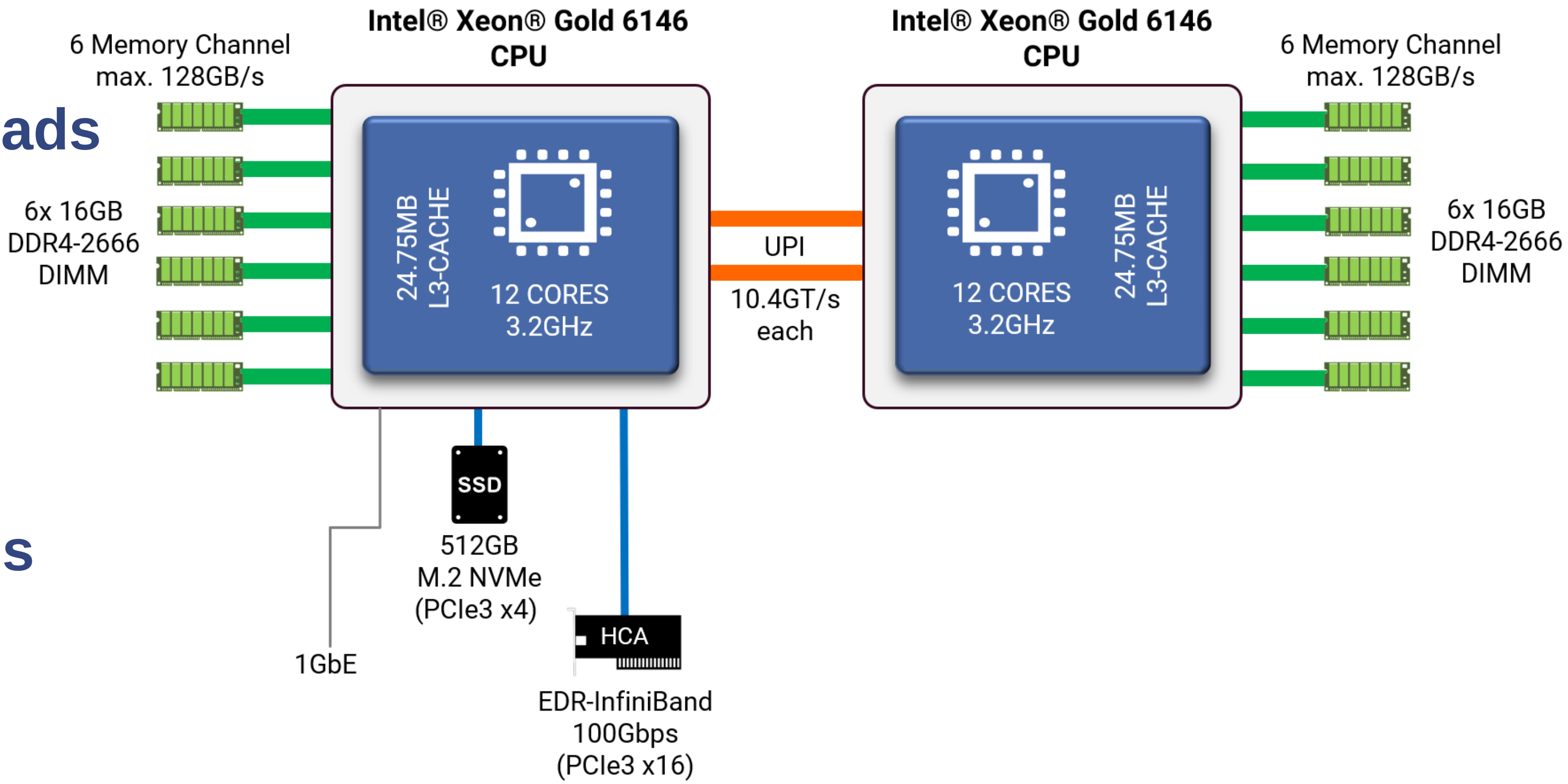
DEEP-EST: proof point for MSA

Significant uptake of SW & HW technologies developed in DEEP-EST



Cluster Module

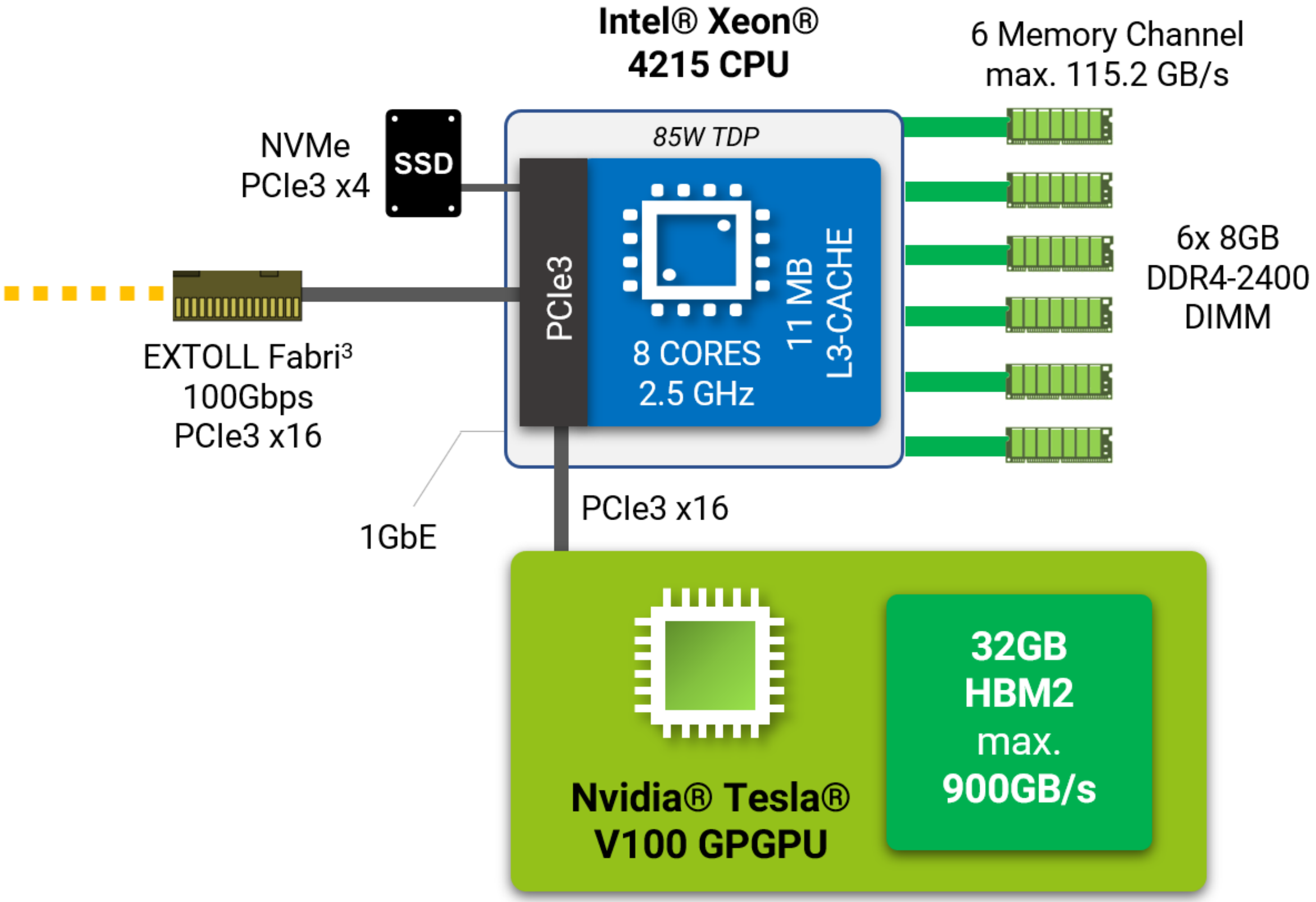
- **Overall 50 nodes**
- **Aimed at CPU-bound workloads**
- **To/from ESB**
  - **Infiniband/Extoll Bridge**
- **To/from DAM**
  - **Inifiband/Ethernet 40Gbps Bridge**



Intel® Xeon® Gold 6146 CPU
Intel® Xeon® Gold 6146 CPU
6 Memory Channel max. 128GB/s
6 Memory Channel max. 128GB/s
6x 16GB DDR4-2666 DIMM
6x 16GB DDR4-2666 DIMM
24.75MB L3-CACHE
12 CORES 3.2GHz
24.75MB L3-CACHE
12 CORES 3.2GHz
UPI
10.4GT/s each
SSD
512GB M.2 NVMe (PCIe3 x4)
HCA
EDR-InfiniBand 100Gbps (PCIe3 x16)
1GbE

Extreme Scale Booster

- Overall 75 nodes
- GPU-based, Nvidia V100
- Extoll / Infiniband Network Fabric
  - ~ Quarter of nodes uses Extoll

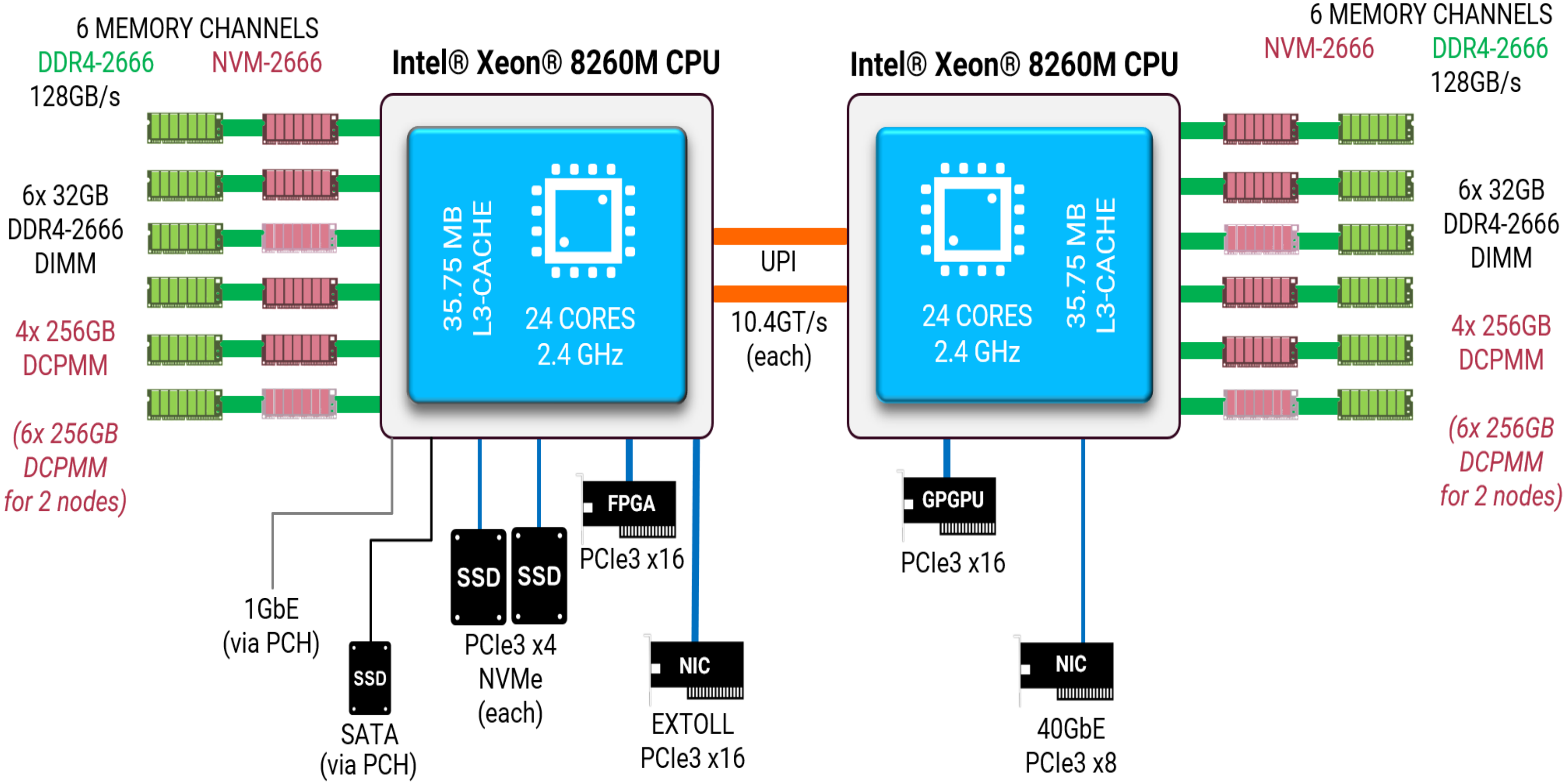


Data Analytics Module

- 16 nodes

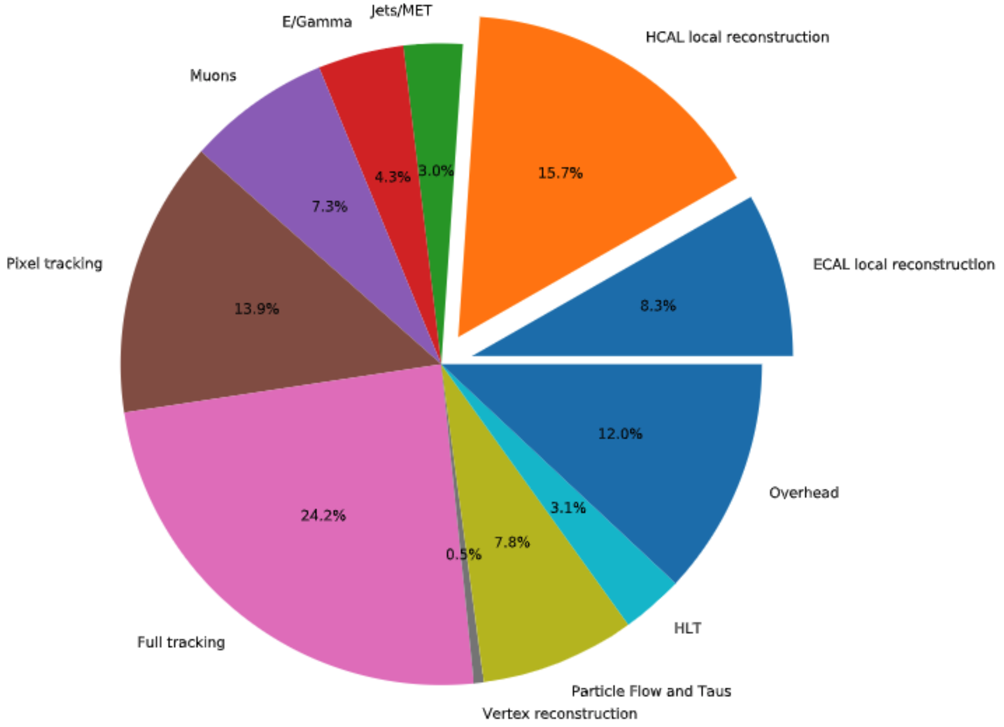
- 2 accelerators per node
  - 1 GPU - Nvidia V100
  - 1 FPGA - Intel Stratix 10

- Memory
  - 2-3TBs Intel Optane Memory
  - 384GB DDR4



DEEP-EST: Heterogenous data processing

- Heterogenous Execution for CMSSW within Patatrack Effort
- Porting of HCAL / ECAL Local Energy Reconstruction
- Current Calorimeters consume ~15-20% of the total HLT time
  - Both (Ecal/Hcal) utilize the same algorithm (the core part) for energy regression



Results: CMS Hcal/Ecal only

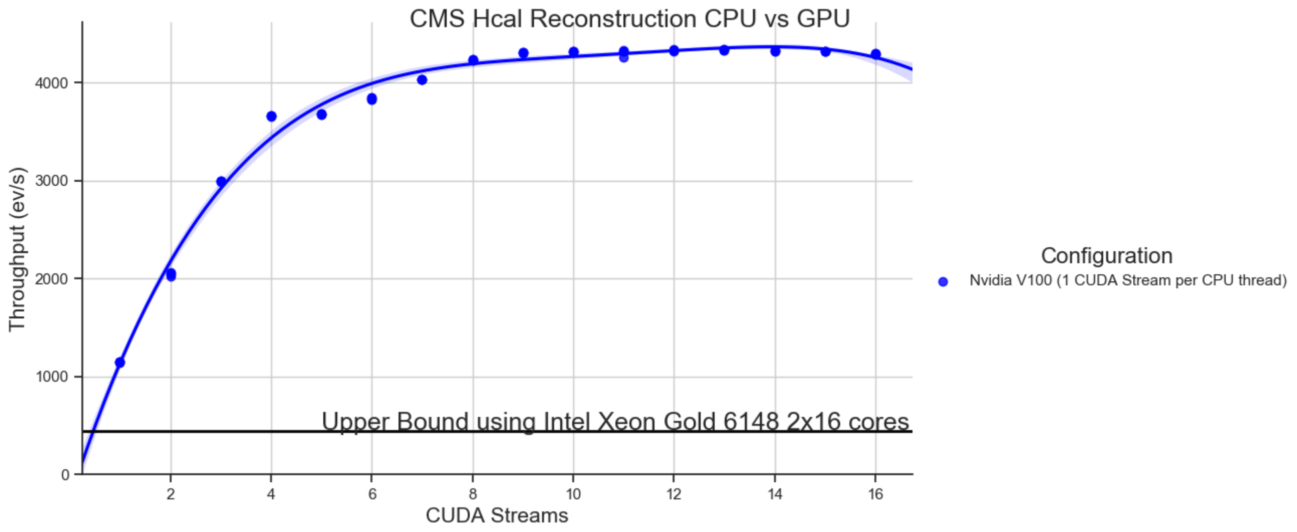
- http://opendata.cern.ch/record/12303

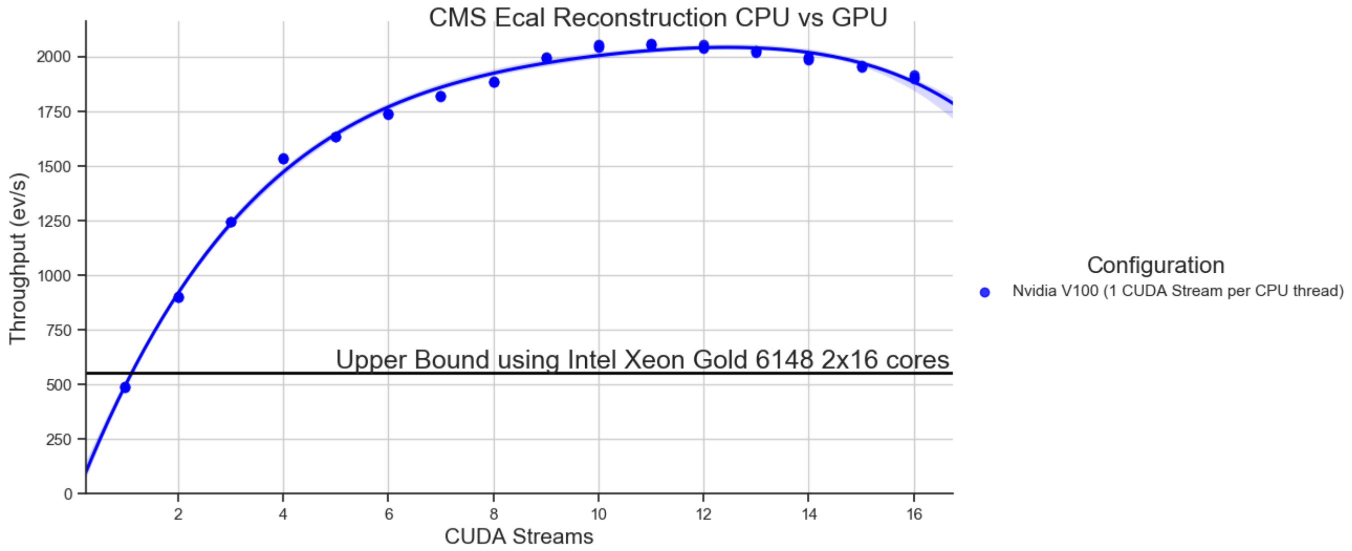
- 20K events. Replicate twice

- Hcal -> speed of 7-8x
  - Using Nvidia V100 GPU

- Ecal -> speed of 3-4x
  - Using Nvidia V100 GPU

Intel Xeon Gold 6148 used for the baseline



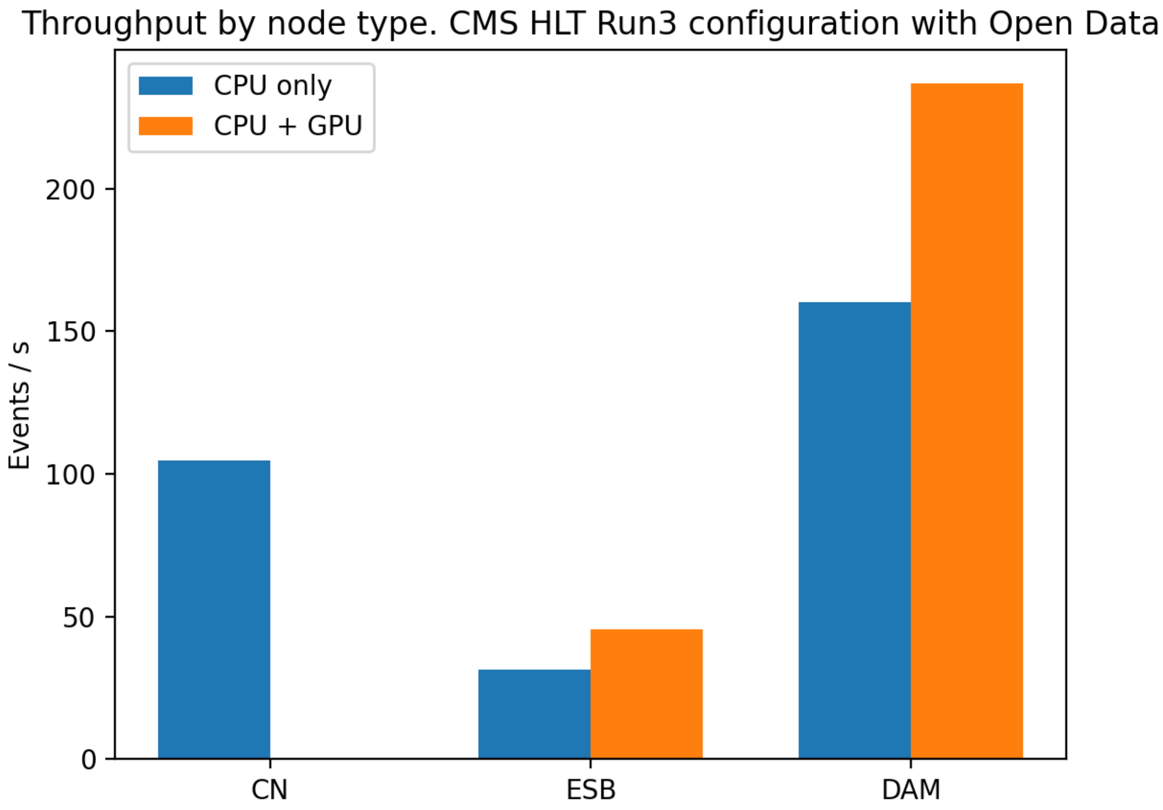


Results: CMS HLT-like Run3

- http://opendata.cern.ch/record/12303 - Open Data used

- CMS HLT-like Run3 configuration
  - Includes Patatrack GPU developments

- **50% more out of nodes with Nvidia GPUs (V100 here)**



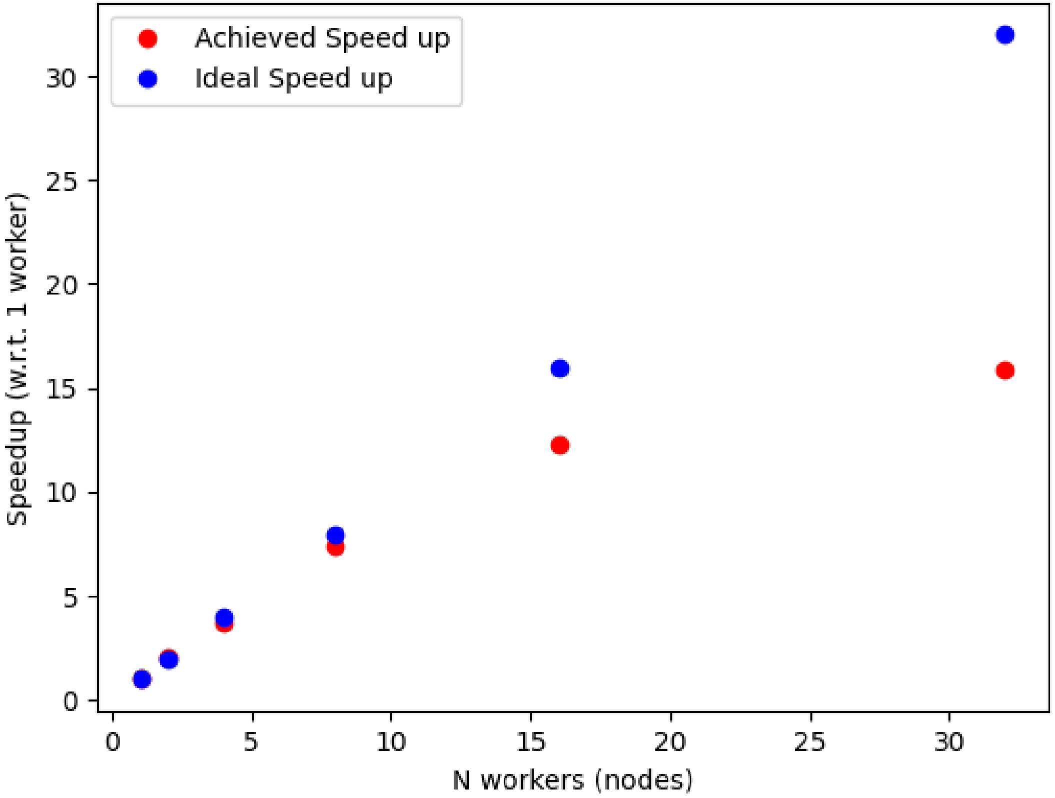
Results: Testing AI Workflows

- Goal: Understand scalability of distributed DL training on HPC
  - specifically for HEP workloads

- Used
  - JEDI-net (arxiv: 1908.05318) model
  - NNLO package for distributed MPI
    - *ParastationMPI on DEEP-EST prototype*
    - *With GPUDirect support*
  - Data from zenodo

- Using up to 32 DEEP-EST ESB nodes
  - 1 Nvidia V100 GPU per node
  - 100Gbps Infiniband interconnect

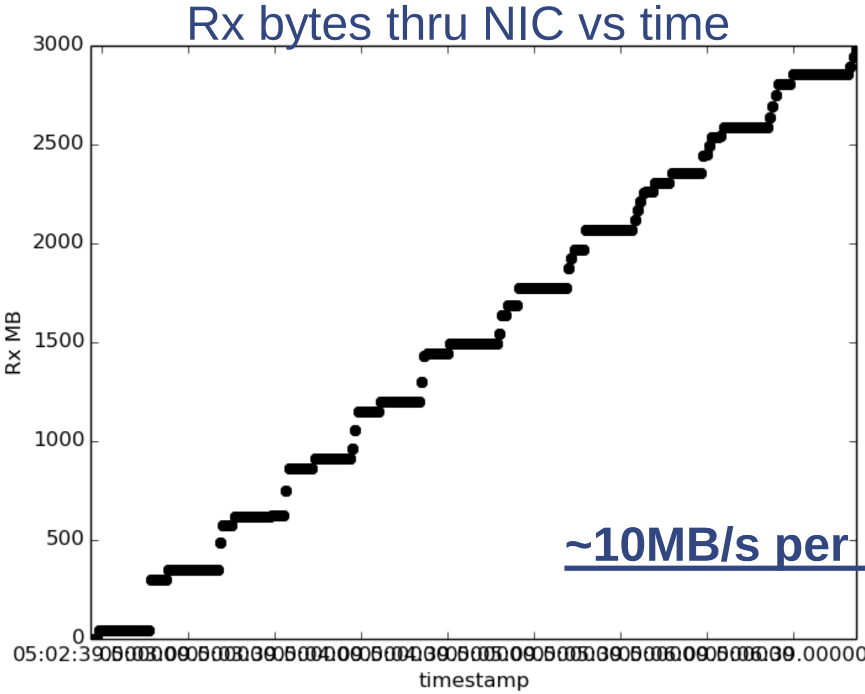
- **Good scalability observed**



Data Access / HPC:

RAISE
Center of Excellence

- How will HEP process Exabytes of data?!
- **Would HPC Storage Systems be enough**
  - Or Exascale HEP workflows will require extra
- CMS MINIAOD2NANO workflow
  - Running up to 64 nodes
    - *~640MB/s aggregate from Ceph (SDSC Popeye)*
    - *This is on average (HEP i/o is bursty)*
- Hypothetically, Exascale HPC O(1M) cores total -> O(10K) nodes -> O(100GB/s) aggregate
  - Although we might never use from a single site



~10MB/s per node

Conclusions

- The DEEP-EST Project concludes end of March 2021

- The DEEP-EST Project proved to be an invaluable platform for
  - Collaboration with HPC experts from other sciences/centers
  - HEP Tests and Developments towards the usage of Exascale HPCs

- Contribution to the Patatrack effort has been integrated into CMS Experiment's framework and will be used in Run 3.

- Now starting to work on RAISE HPC project

New CoE RAISE: Research on AI- and Simulation-Based Engineering at Exascale

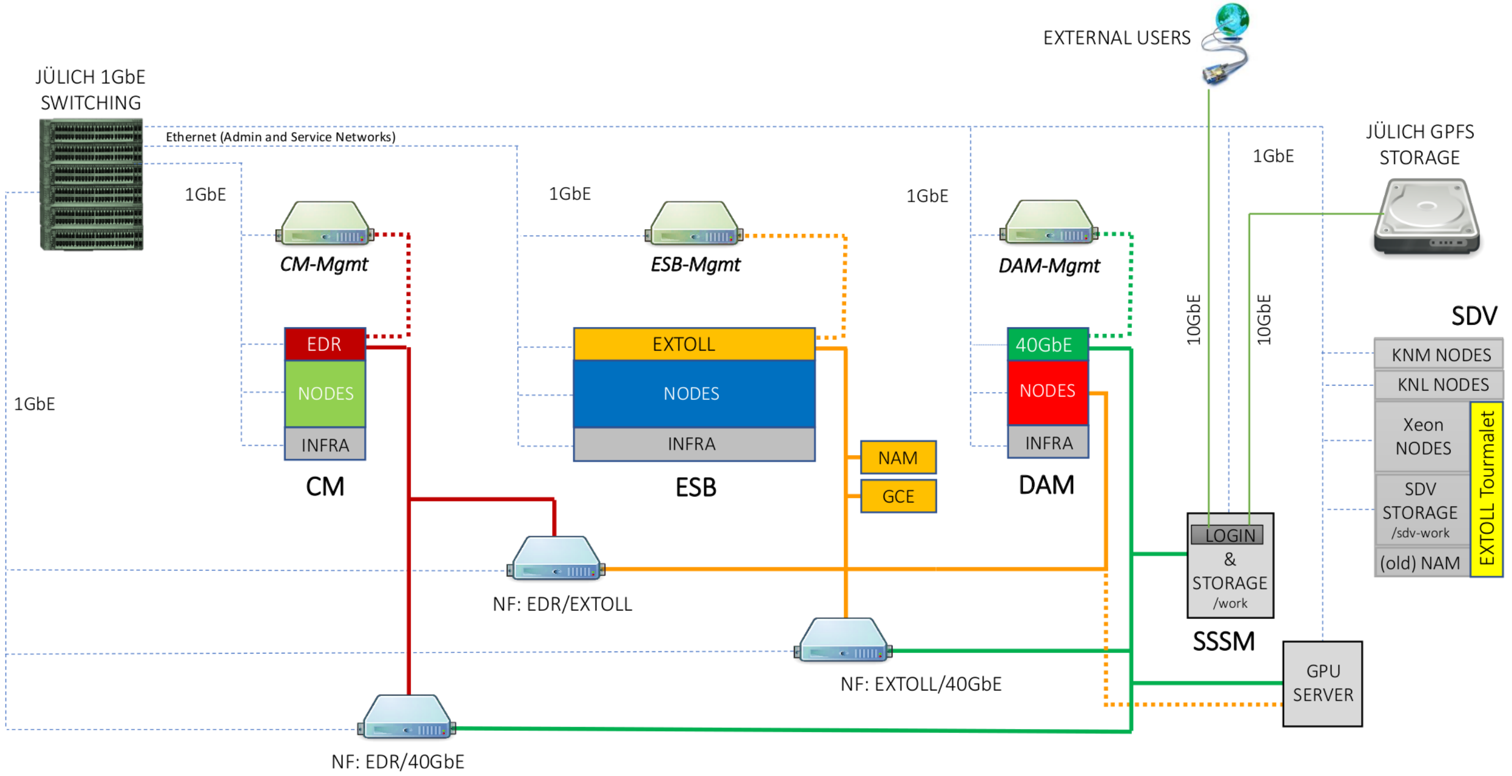
DEEP
Projects

The DEEP projects have received funding from the European Union’s Seventh Framework Programme (FP7) for research, technological development and demonstration and the Horion2020 (H2020) funding framework under grant agreement no. FP7-ICT-287530 (DEEP), FP7-ICT-610476 (DEEP-ER) and H2020-FETHPC-754304 (DEEP-EST).

Network Federation + Auxiliary

- Multiple fabrics
  - 100Gbps Infiniband
  - 100Gbps Extoll
  - 40 Gbps Ethernet
  - Bridges
- Network Attached Memory NAM
  - Extoll’s FPGA based solution
  - 128GBs DDR4
  - TB(s) SSDs
- Global Collective Engine GCE
  - Extoll’s FPGA based solution
  - Accelerate MPI-collective operations

DEEP-EST Prototype – Schematic Network Overview



Results: Testing Intel OneAPI

- **Ported standalone CMS Ecal Reconstruction algorithm to use Intel OneAPI**

- Results between regular c++ and Intel OneAPI implementations match 100%
  - Tested on a CERN’s VM with CPUs only

- Employed Intel OneAPI Compatibility Tool to convert from CUDA-based implementation
  - Almost 0 modifications after the conversion
  - Had to adapt slightly Eigen to be used within OneAPI kernels

- Github repo: https://github.com/vkhristenko/cmsregr-oneapi