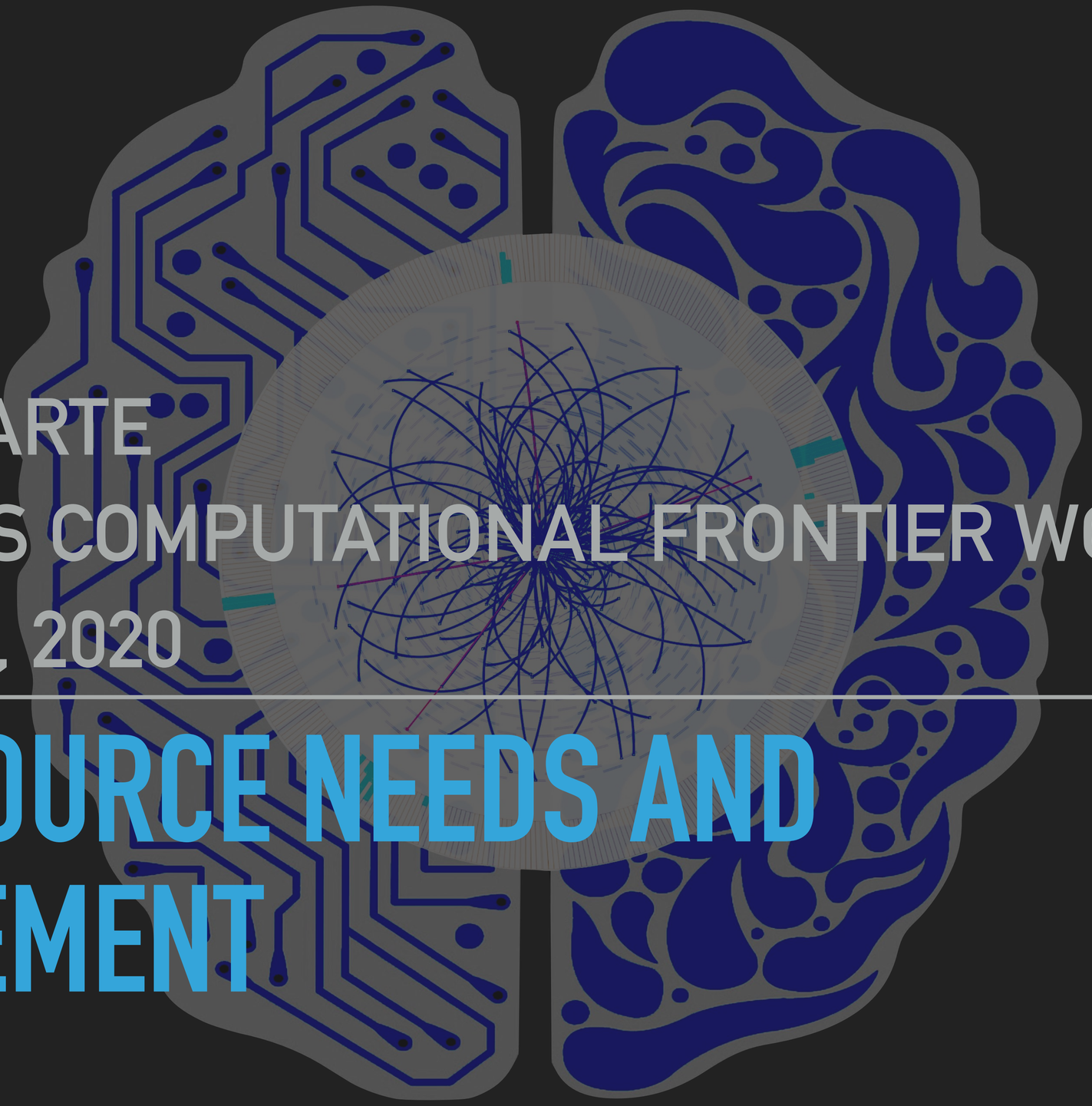
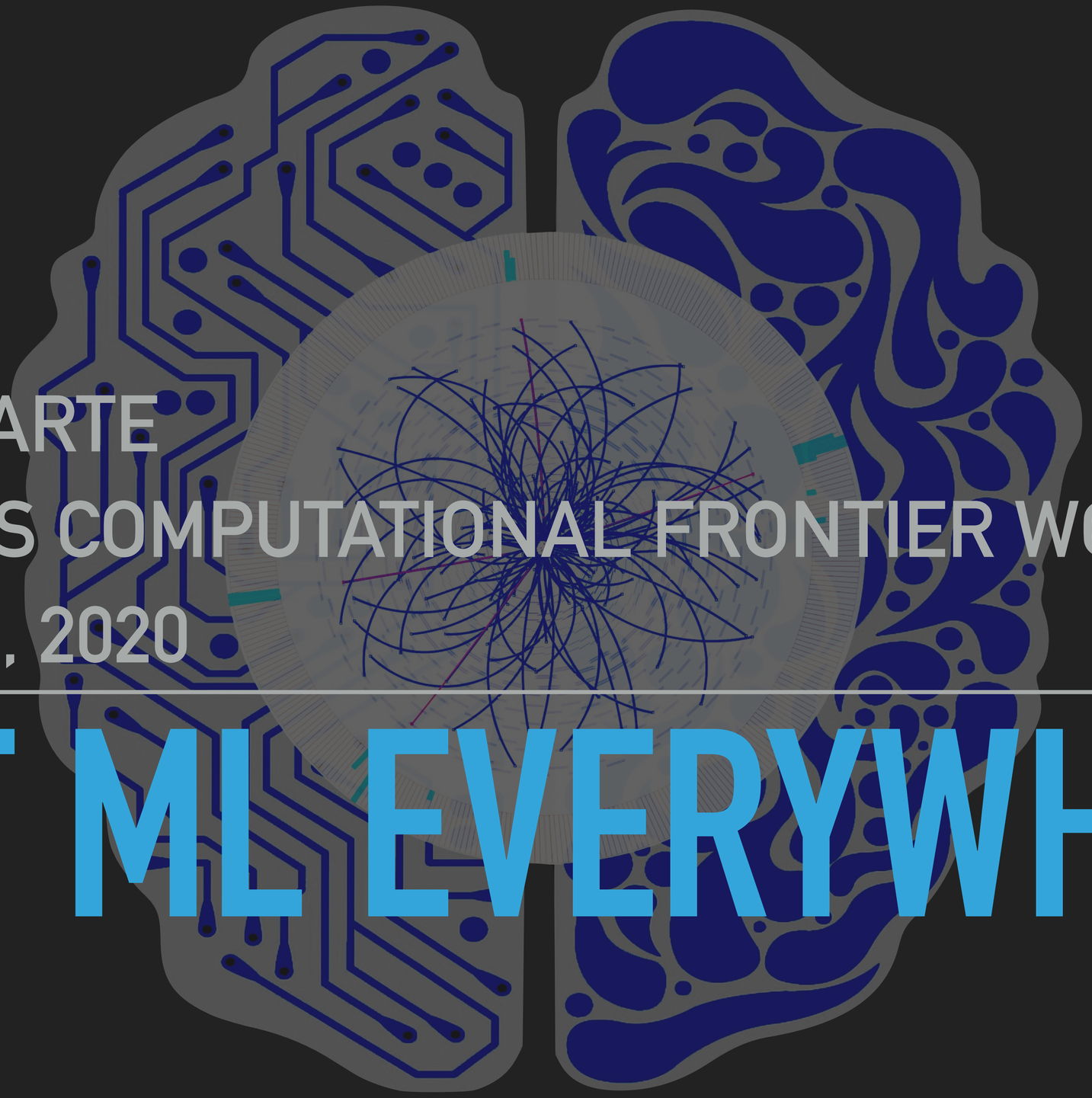


JAVIER DUARTE

SNOWMASS COMPUTATIONAL FRONTIER WORKSHOP

AUGUST 11, 2020

ML RESOURCE NEEDS AND MANAGEMENT



JAVIER DUARTE

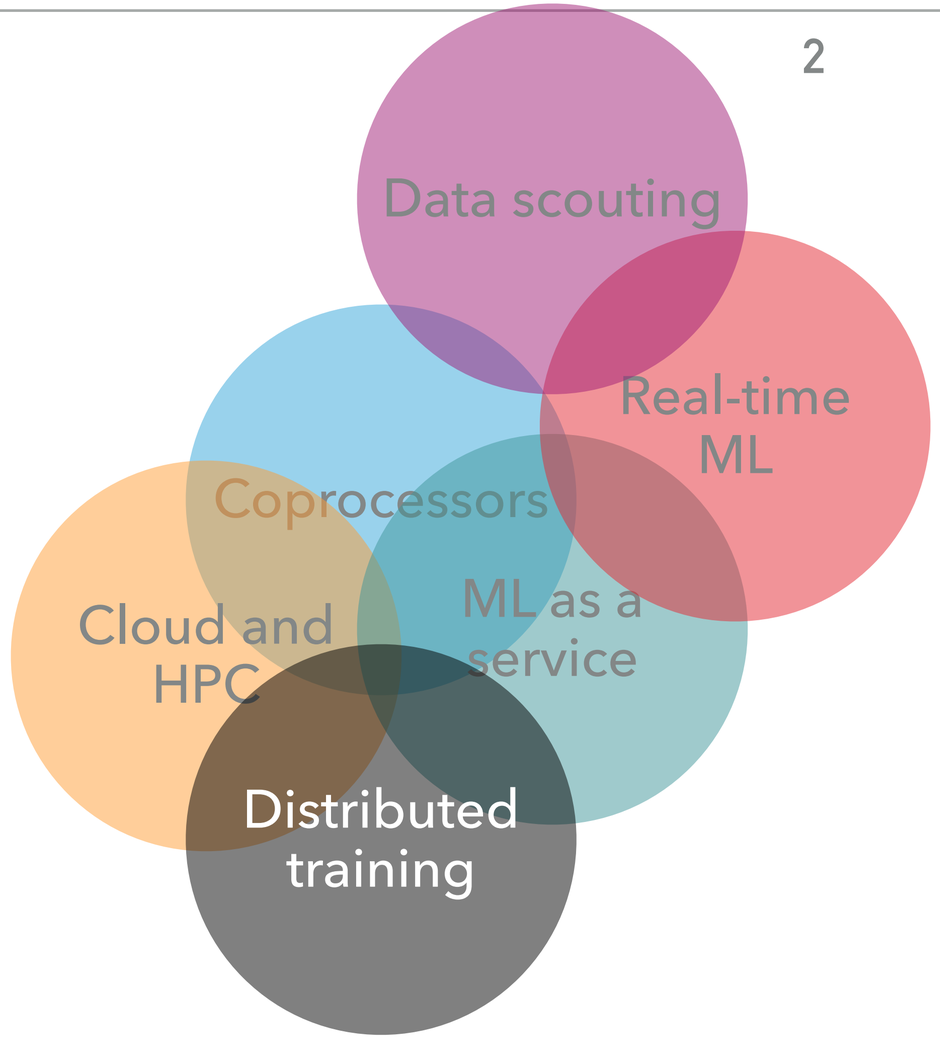
SNOWMASS COMPUTATIONAL FRONTIER WORKSHOP

AUGUST 11, 2020

FAST ML EVERYWHERE

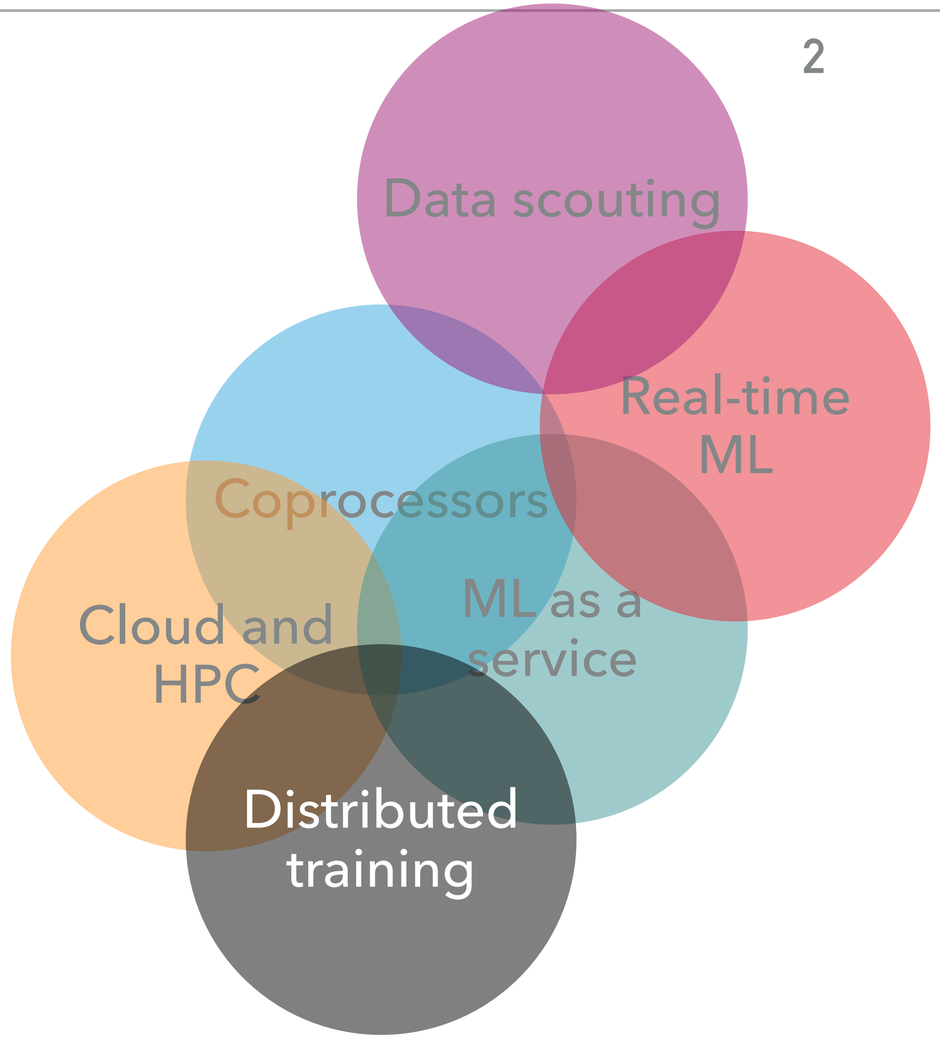
OUTLINE

- Introduction
- "Offline"
  - Coprocessors (FPGAs, GPUs, ASICs)
  - ML as a service
  - Cloud and HPC resources
  - Distributed training
- "Online"
  - Real-time ML
  - Data scouting
- Summary and outlook



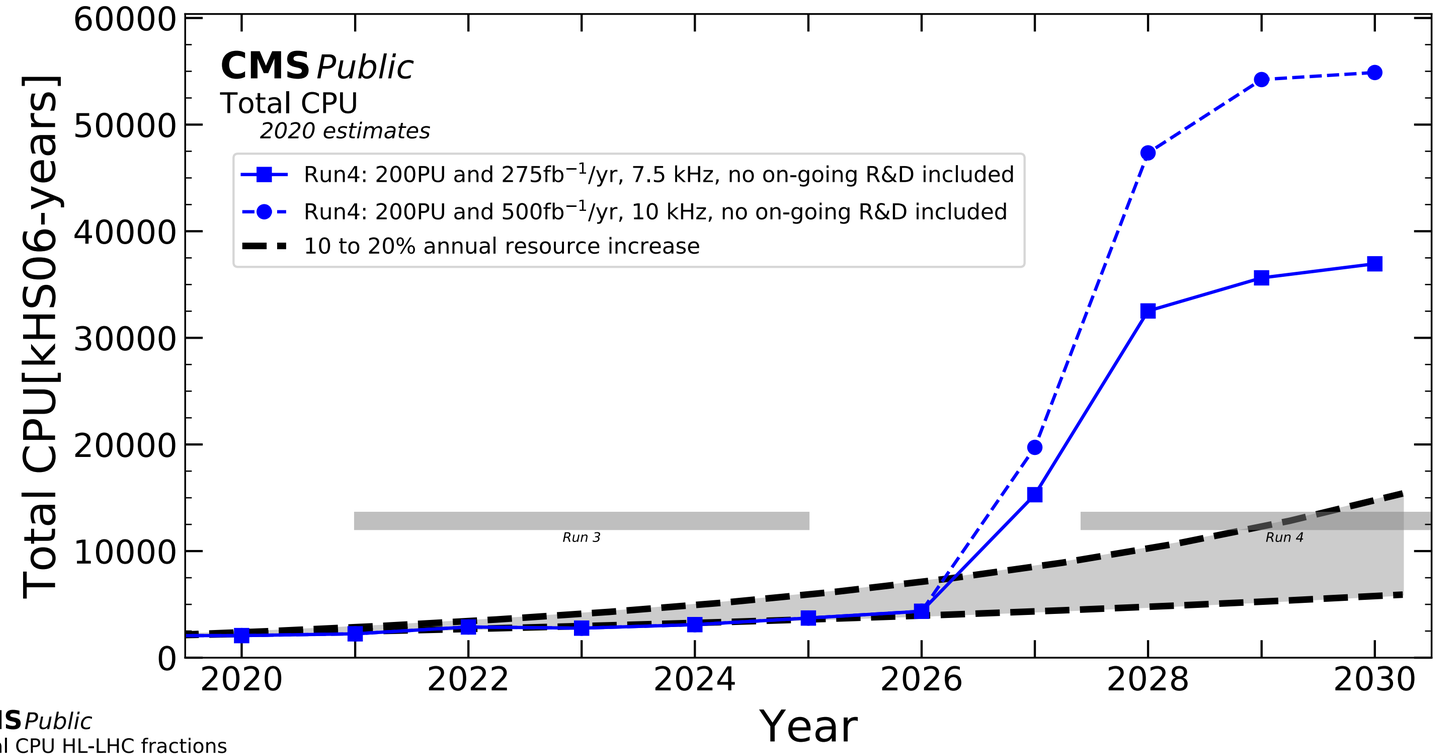
OUTLINE

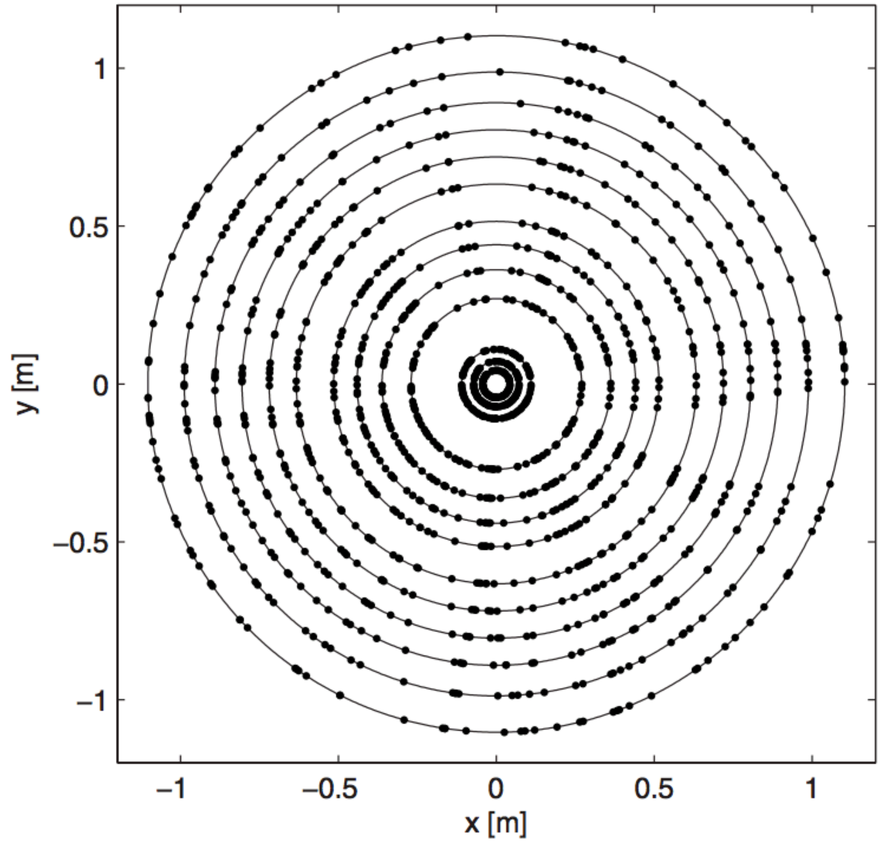
- Introduction
- "Offline"
  - Coprocessors (FPGAs, GPUs, ASICs)
  - ML as a service
  - Cloud and HPC resources
  - Distributed training
- "Online"
  - Real-time ML
  - Data scouting
- Summary and outlook

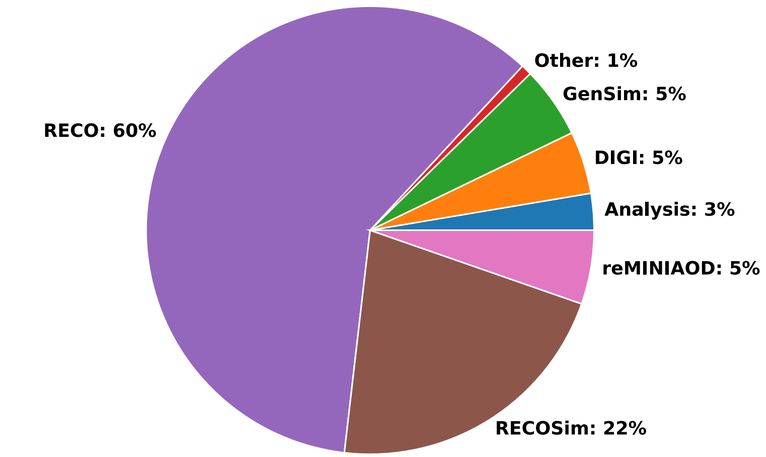


- Note: (biased) sampling of some relevant topics!
  - For more info, see links!

NEXT-GEN COMPUTING DEMANDS

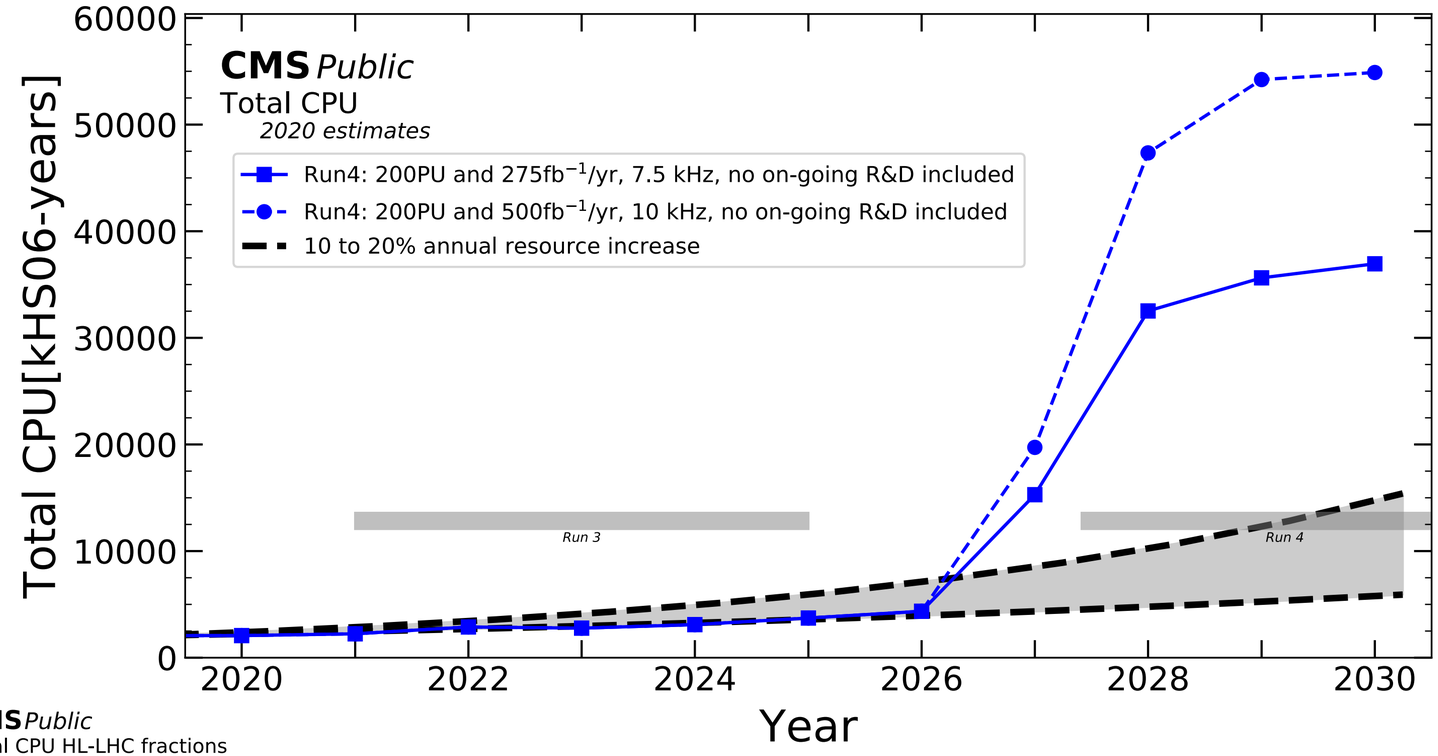


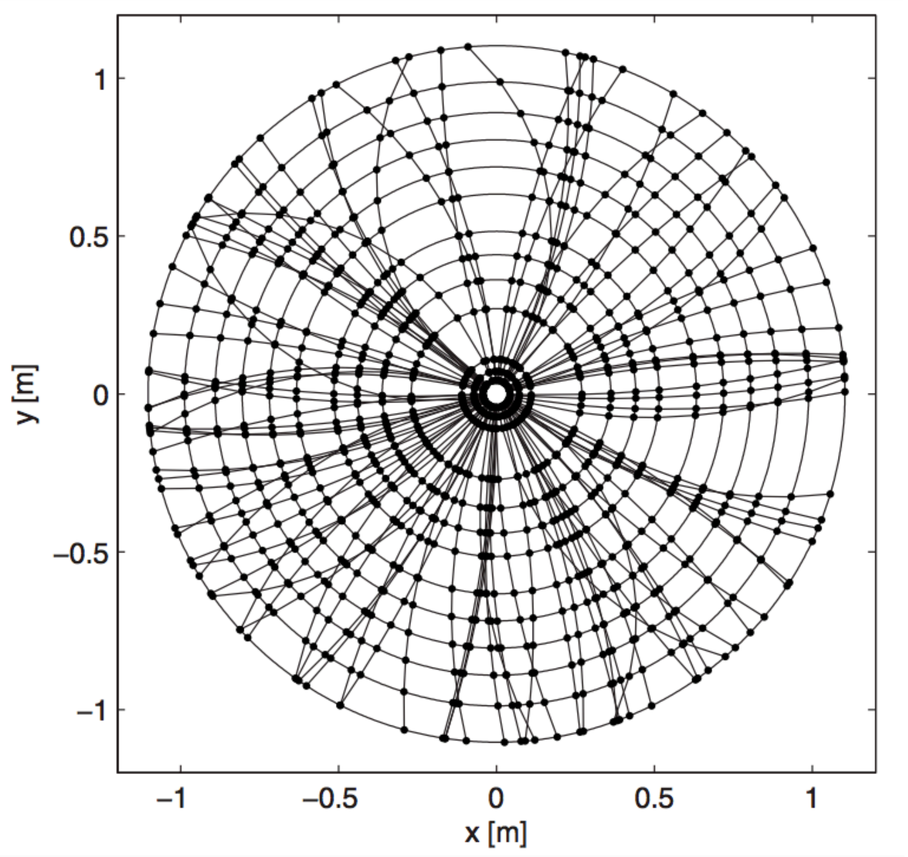


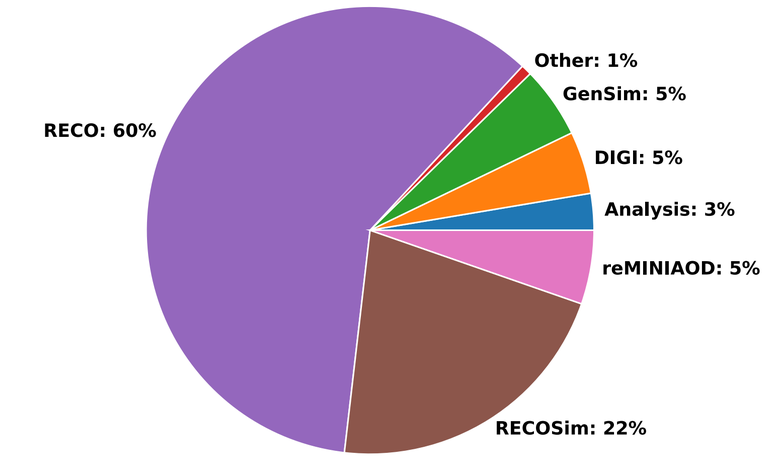


- Need more (or alternative) processing power to keep up with data demands in next-gen experiments (especially for reconstruction)
- Example: tracking at HL-LHC

NEXT-GEN COMPUTING DEMANDS







- Need more (or alternative) processing power to keep up with data demands in next-gen experiments (especially for reconstruction)
- Example: tracking at HL-LHC

ML FOR X AND CHALLENGES

- ML is being used for more tasks in HEP: ML for X
  - where X = trigger, calorimetry, tracking, particle-flow, reconstruction, simulation, …

ML FOR X AND CHALLENGES

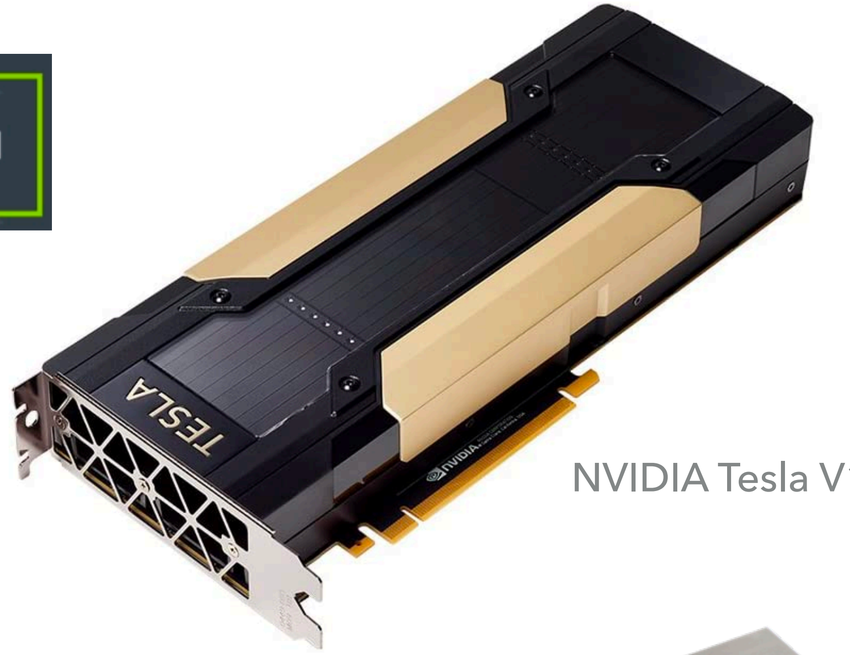
- ML is being used for more tasks in HEP: ML for X
  - where X = trigger, calorimetry, tracking, particle-flow, reconstruction, simulation, …
- Challenges for using ML in HEP:
  - Large algorithms with many ops for inference and training
  - Integration into exp. workflows
  - Utilization of heterogenous resources (GPUs, FPGAs, etc.)
  - Real-time latency constraints

ML FOR X AND CHALLENGES

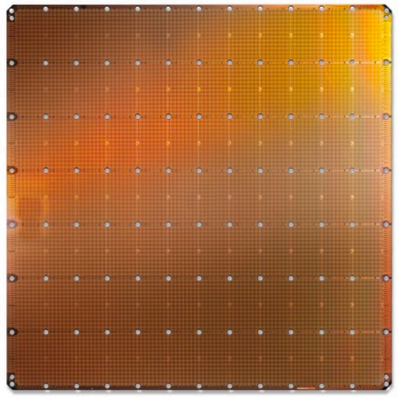
- ML is being used for more tasks in HEP: ML for X
  - where X = trigger, calorimetry, tracking, particle-flow, reconstruction, simulation, …
- Challenges for using ML in HEP:
  - Large algorithms with many ops for inference and training
  - Integration into exp. workflows
  - Utilization of heterogenous resources (GPUs, FPGAs, etc.)
  - Real-time latency constraints
- How can we best utilize [HEP, cloud, heterogeneous] resources for [distributed, accelerated, real-time] [training, inference, hyperparameter optimization] of [large, compressed] ML algorithms

COPROCESSORS / ACCELERATORS

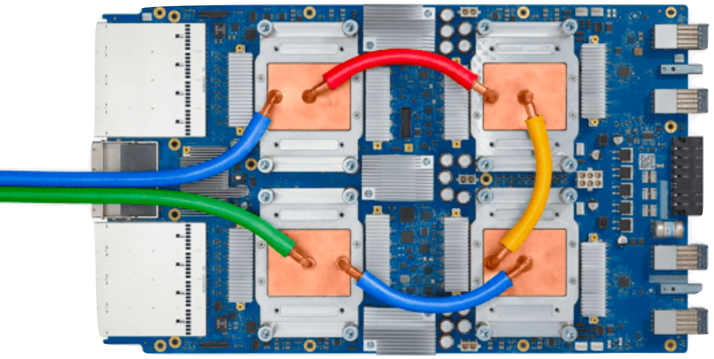
- Heterogeneous computing resources used in tandem with CPUs
  - Can be connected directly or remotely (using client-server internet communication protocols)



NVIDIA Tesla V100 GPU



Cerebras CS-1



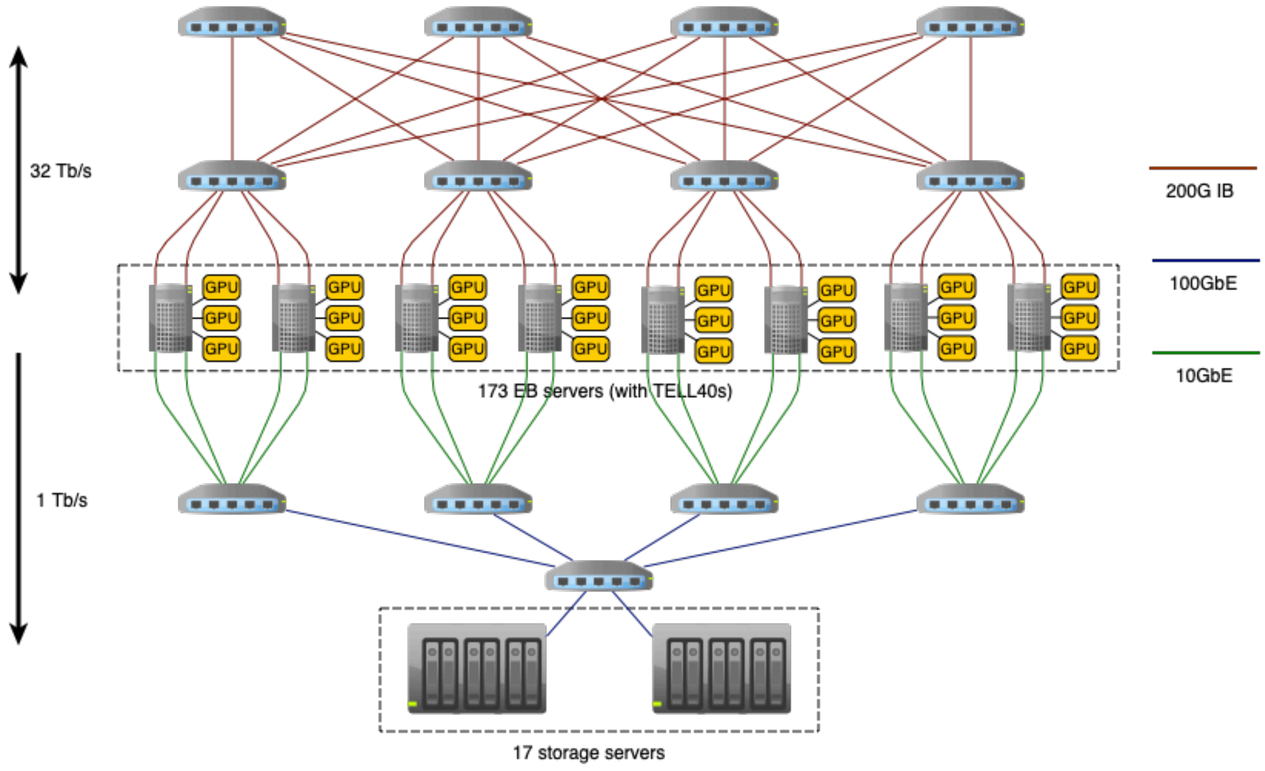
Google Cloud TPU v3

Xilinx Alveo U250 FPGA

COPROCESSOR APPLICATIONS

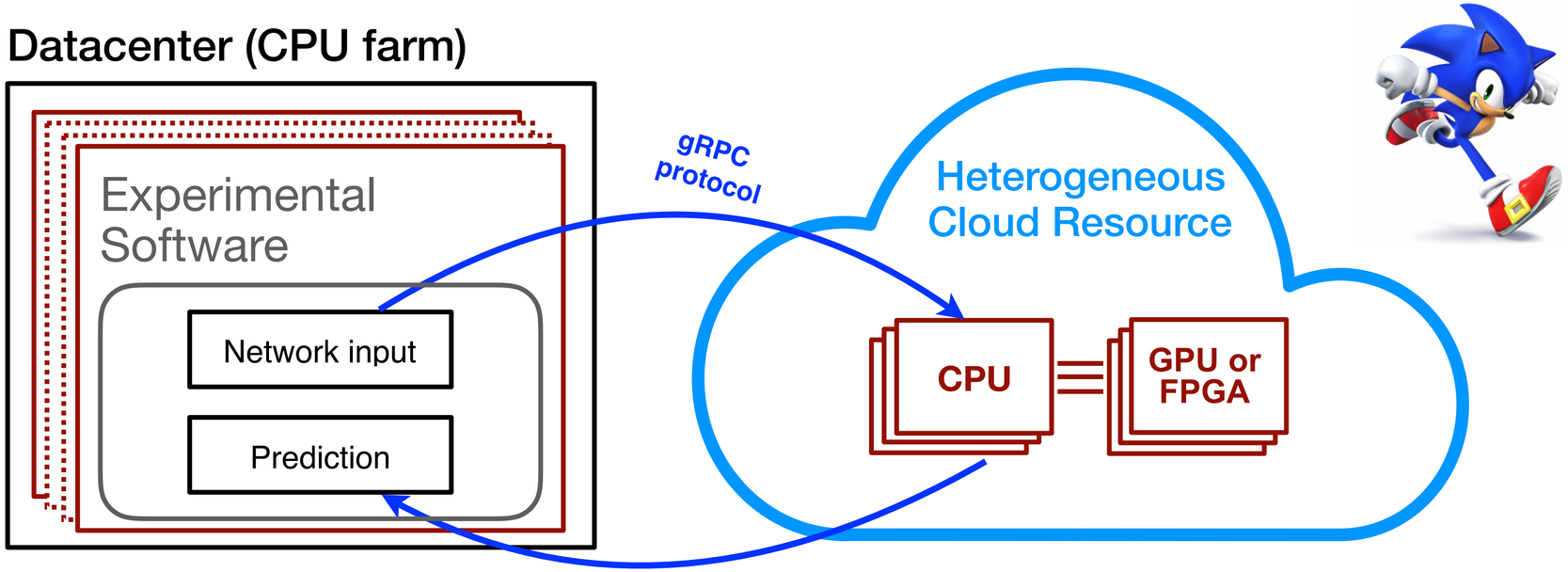
- Accelerate our (domain-knowledge and ML) algorithms with coprocessors
  - Allen: fully GPU HLT1 in LHCb [arXiv:1912.09161]
  - GPU-enhanced HLT in CMS
    - ExternalWork [arXiv:2004.04334]; Patatrack [CMS-DP-2018-059]: pixel local reco., pixel-only track reco., and vertex reco.; SONIC [arXiv:2007.10359]
  - CNNs for event interpretation in multiple neutrino physics experiments

Allen (LHCb)



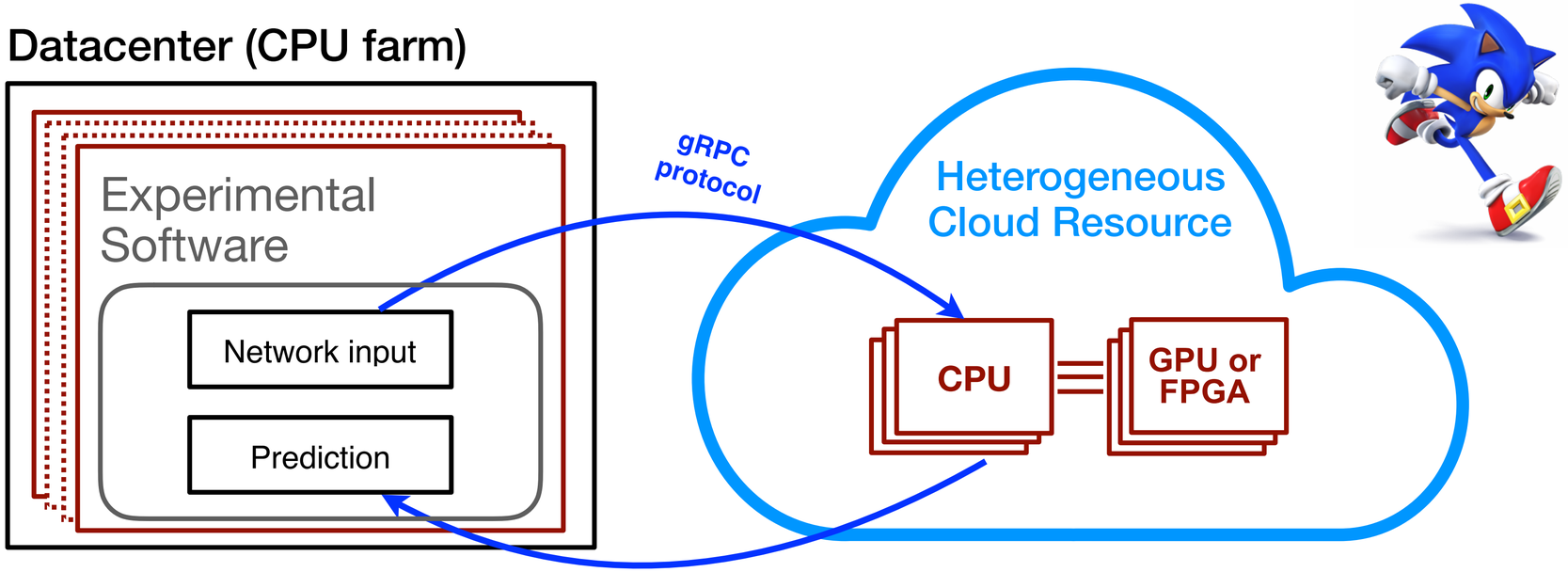
Allen can be implemented in ~500 GPUs @ full collision rate of 30 MHz processing 40 Tb/s

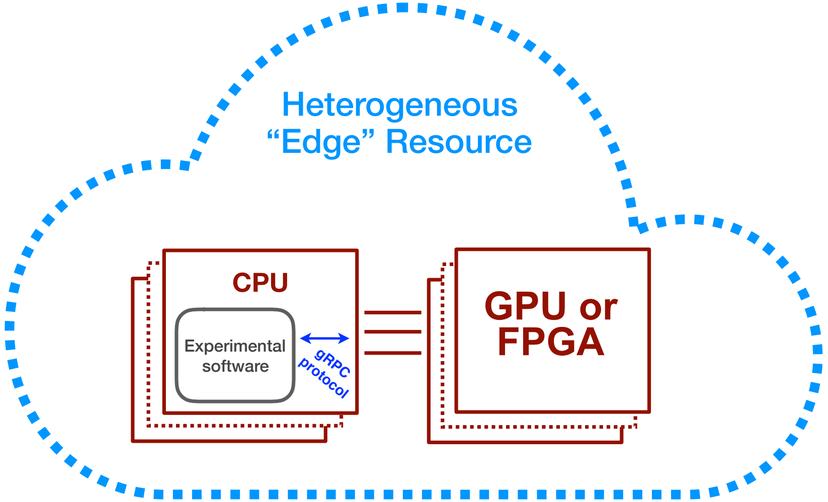
SONIC "AS-A-SERVICE" APPROACH



- Services for Optimized Network Inference on Coprocessors (SONIC)
  - Send data (e.g. jet images) from exp. CPUs to GPUs or FPGAs

SONIC "AS-A-SERVICE" APPROACH

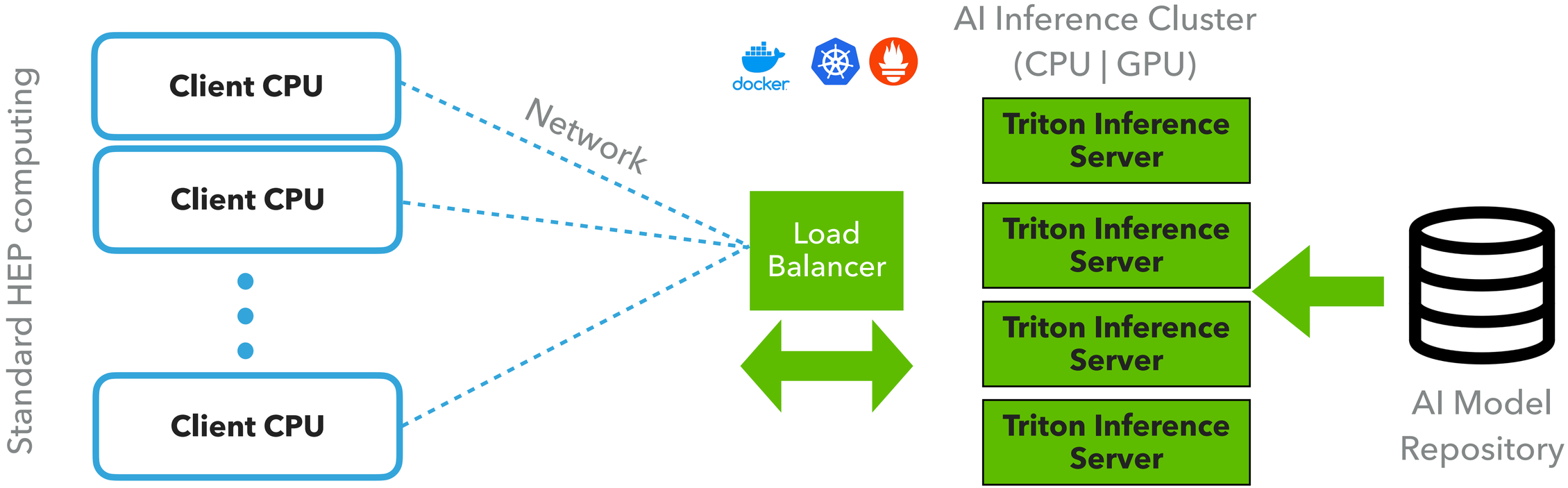




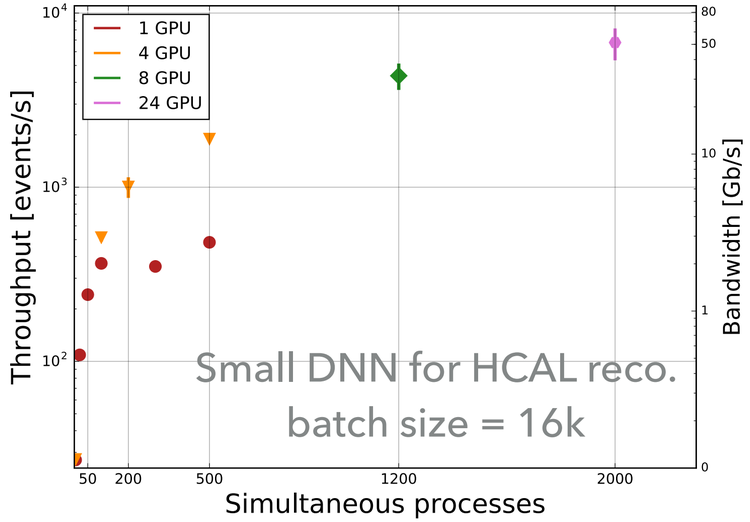
- Services for Optimized Network Inference on Coprocessors (SONIC)
  - Send data (e.g. jet images) from exp. CPUs to GPUs or FPGAs
- Two modes: cloud service and on premises

ML AS A SERVICE WITH SONIC + TRITON + GPUS

arXiv:2007.10359



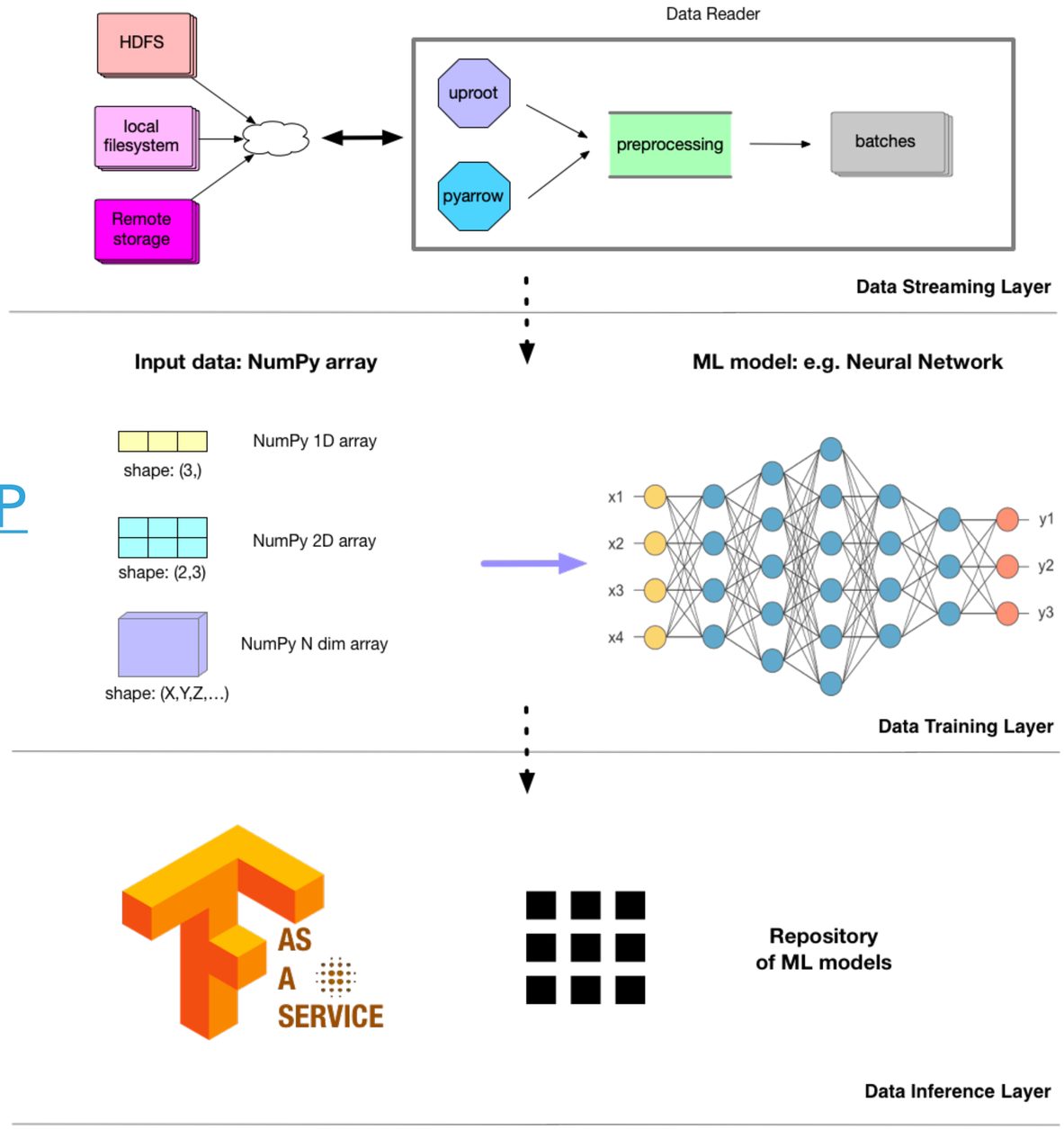
- SONIC + Triton Inference Server demonstrates ML as a service on GPUs for a variety of collider [arXiv:2007.10359] and neutrino physics tasks [https://indico.slac.stanford.edu/event/371/contributions/1210/]



ML TRAINING AS A SERVICE

arXiv:1811.04492

- Utilize commercial tools and interface them with HEP data to train directly on them (e.g. ROOT files):
  - https://github.com/vkuznet/MLaaS4HEP
- Other training repos:
  - https://github.com/hqucms/weaver
  - https://github.com/DL4Jets/DeepJetCore



CLOUD AND HPC COMPUTING

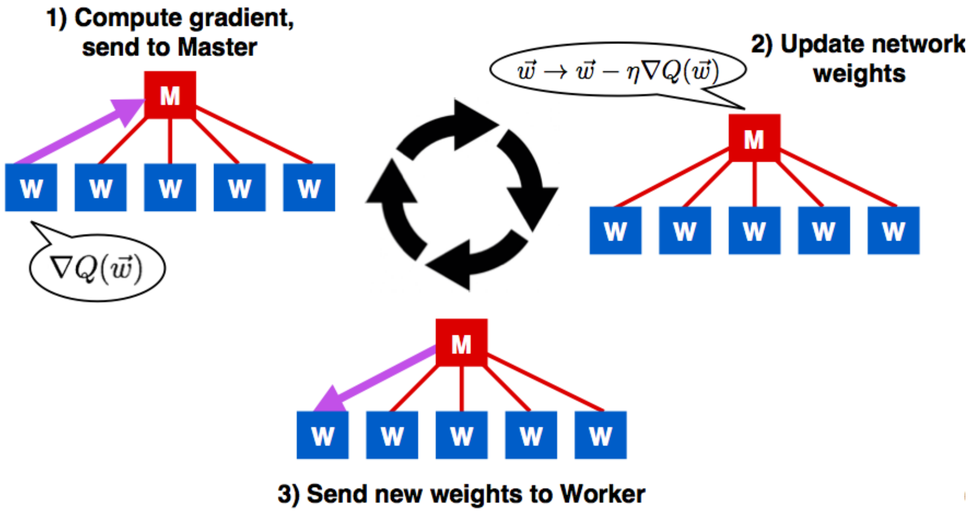
- HEPCloud: integrating commercial cloud (AWS, GCP), allocation-based (HPC), with traditional experiment resources [arXiv:1710.00100, doi:10.1088/1742-6596/898/5/052019, arXiv:1806.03224, doi:10.1051/epjconf/201921403031, arXiv:1904.08988]
  - ML for optimization of computing workflow management

- Some (growing) experience of using HPCs in HEP experiment workflows
  - Especially distributed training

DISTRIBUTED TRAINING

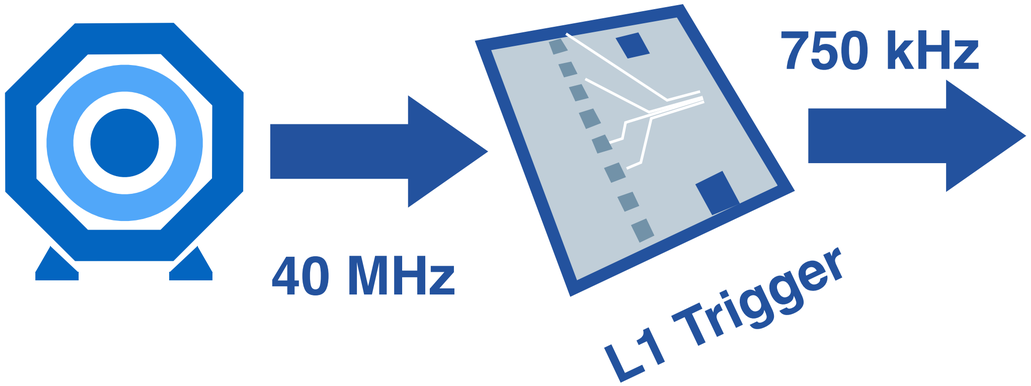
See: https://indico.cern.ch/event/822126/contributions/3500187/
https://indico.slac.stanford.edu/event/371/contributions/1231/

- Flavors of distributed training:
  - Data distribution, gradient distribution, model distribution
  - Hyperparameter optimization: bayesian optimization, evolutionary algorithms, …
- Exa.TrkX group working to demonstrate distributed training and hyperparameter optimization on HPCs for particle tracking
- Neutrino physics applications as well [https://indico.slac.stanford.edu/event/371/contributions/1231/]



REAL-TIME ML, E.G. CMS LEVEL-1 TRIGGER

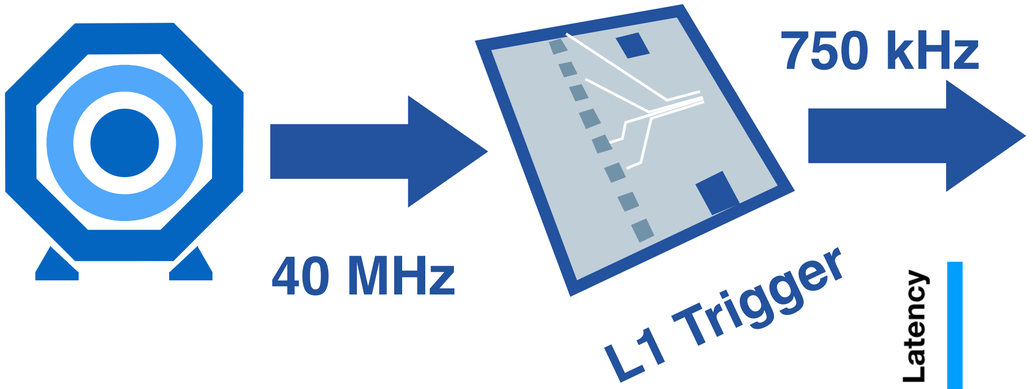
CMS-TDR-021



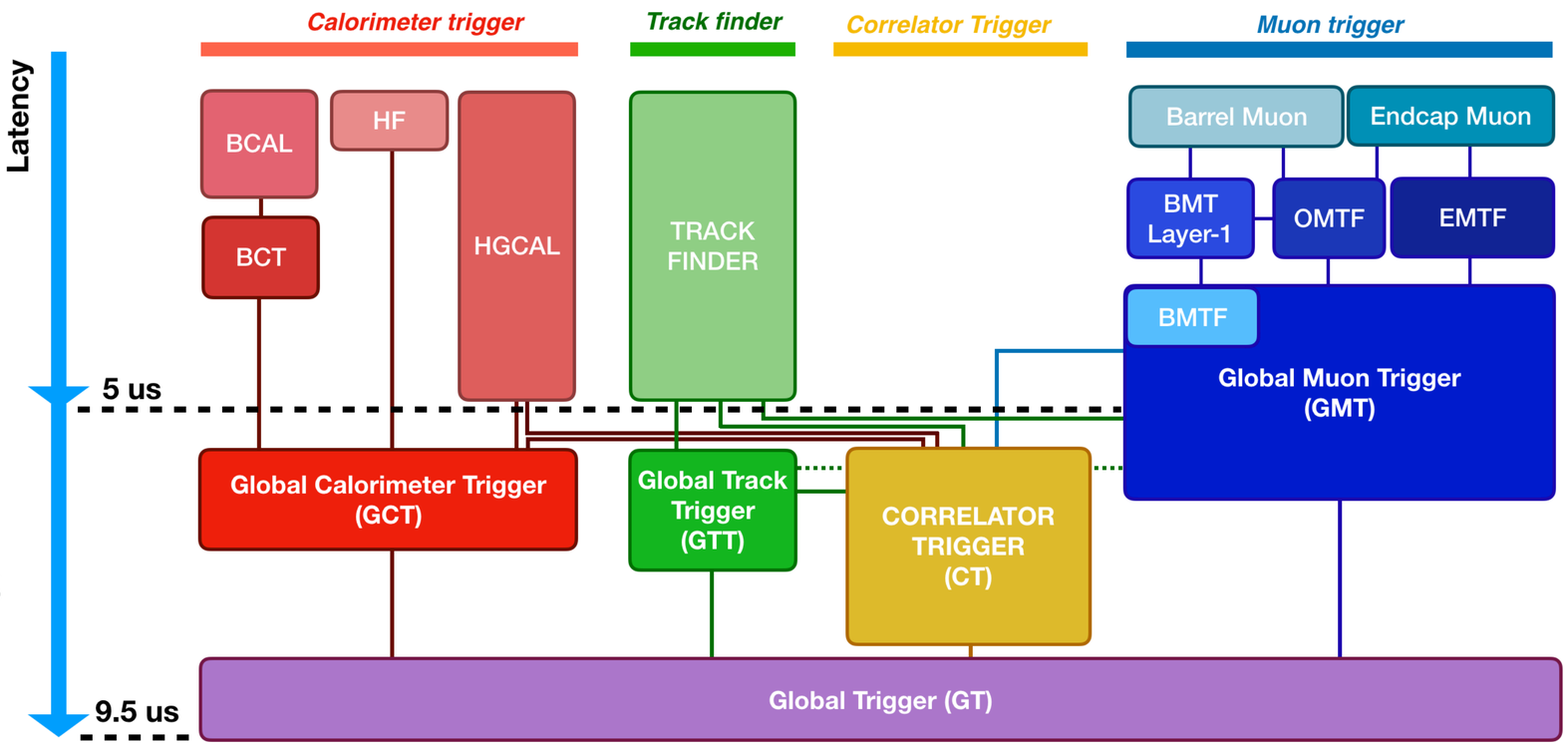
- Upgraded Level-1 Trigger:
  40 MHz → 750 kHz

REAL-TIME ML, E.G. CMS LEVEL-1 TRIGGER

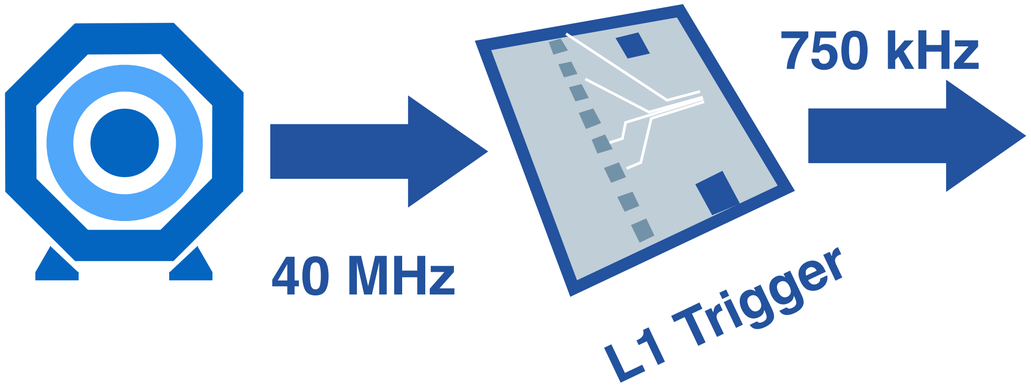
CMS-TDR-021



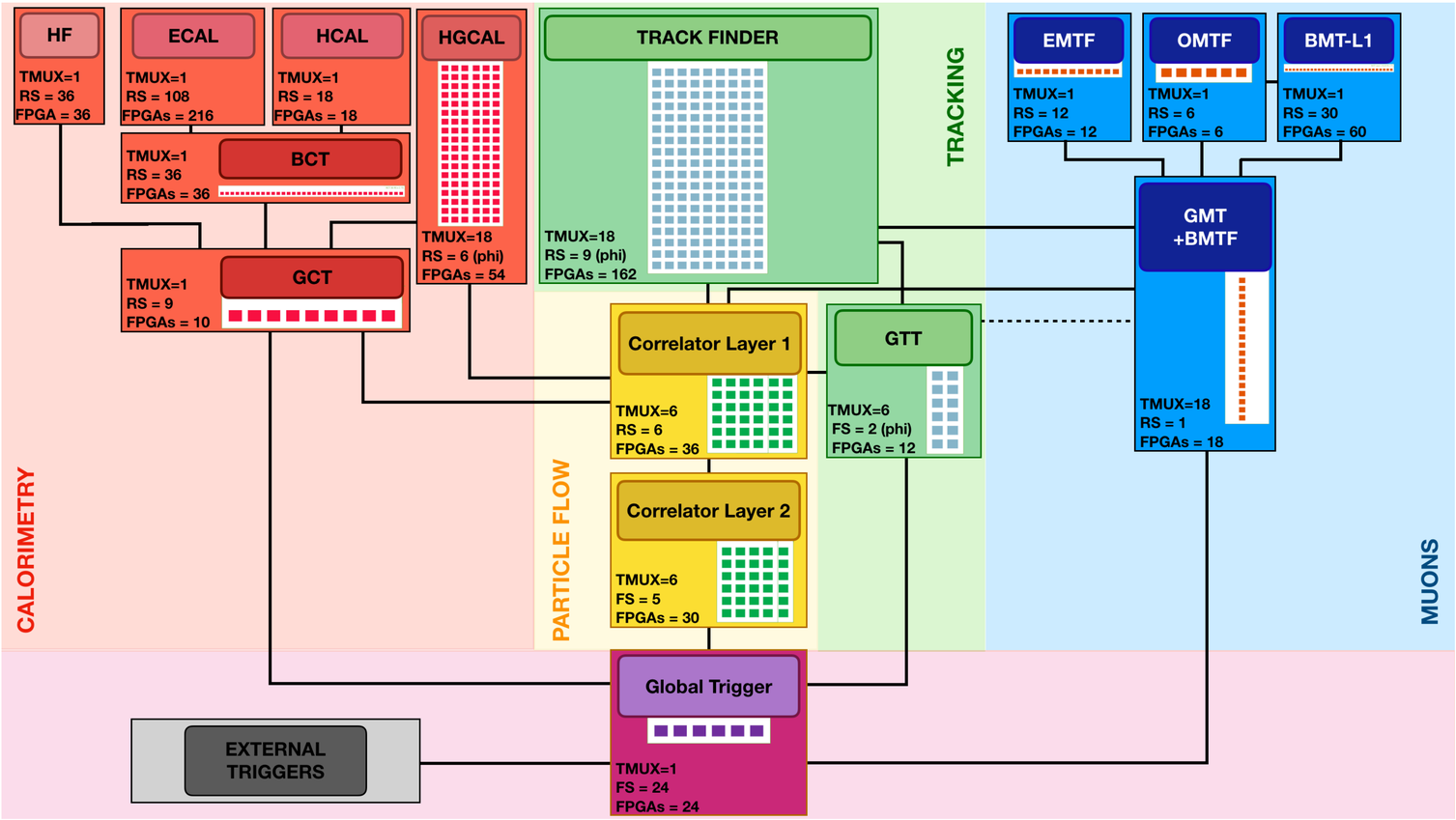
- Upgraded Level-1 Trigger: 40 MHz → 750 kHz
- Reconstruct all events and reject 98% of them in ~12.5 μs



REAL-TIME ML, E.G. CMS LEVEL-1 TRIGGER



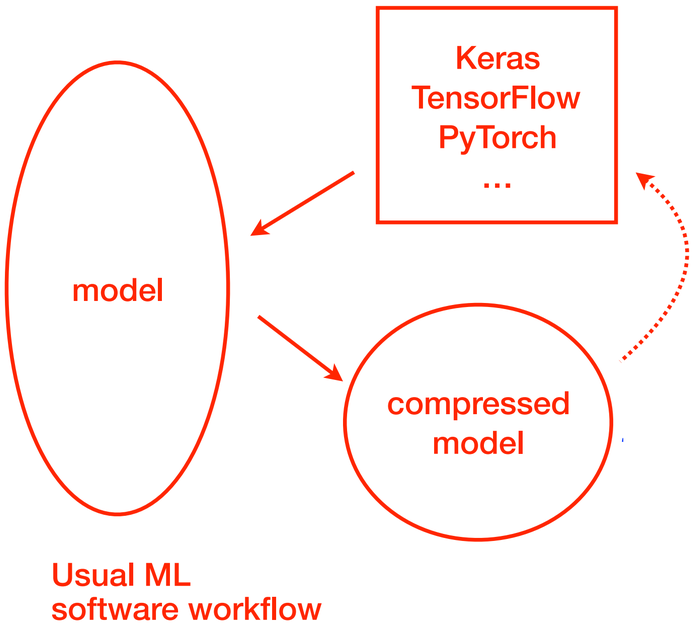
- Upgraded Level-1 Trigger: 40 MHz → 750 kHz
- Reconstruct all events and reject 98% of them in ~12.5 μs
- Latency necessitates all **FPGA** design (729 FPGAs!)



REAL-TIME ML WITH HLS4ML

J. Instrum. 13, P07027 (2018)

- **hls4ml** for physicists or ML experts to translate **ML algorithms** into **FPGA firmware**

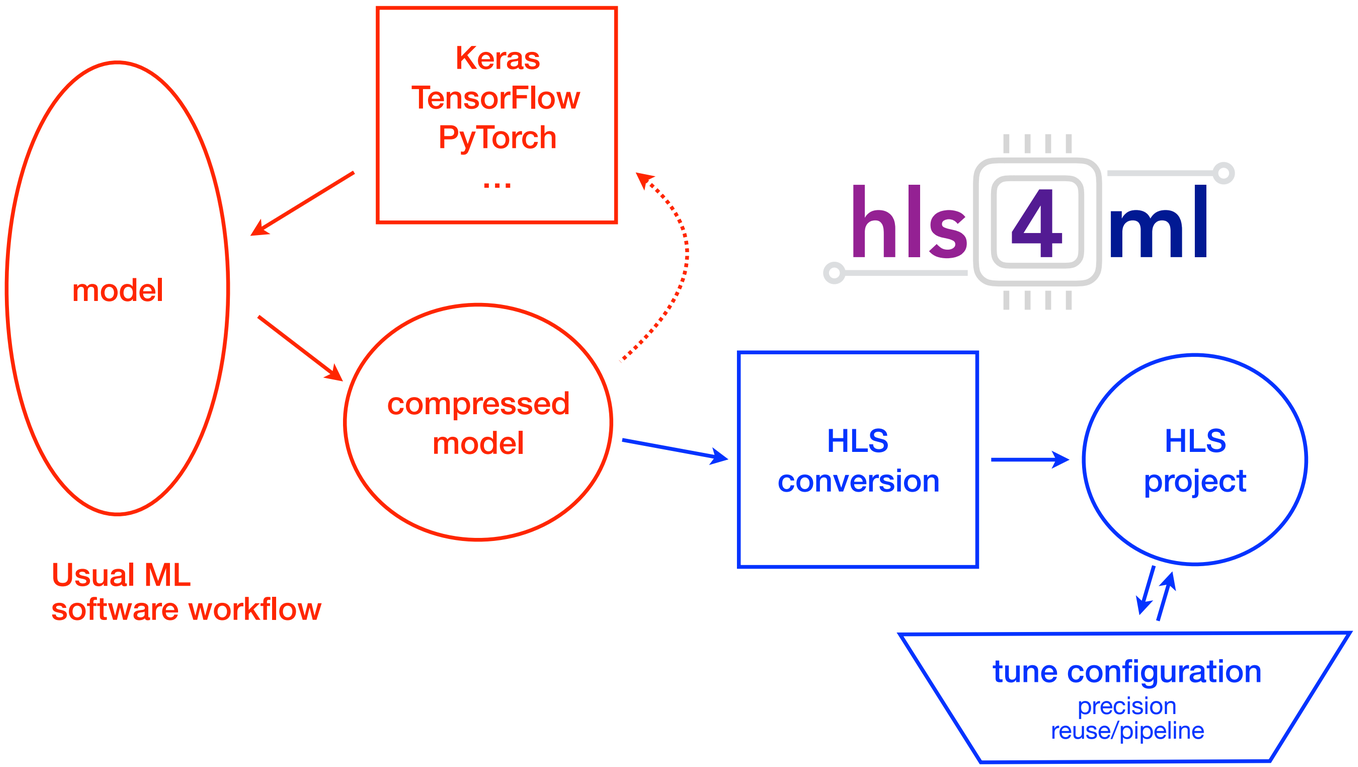


hls 4 ml

REAL-TIME ML WITH HLS4ML

J. Instrum. 13, P07027 (2018)

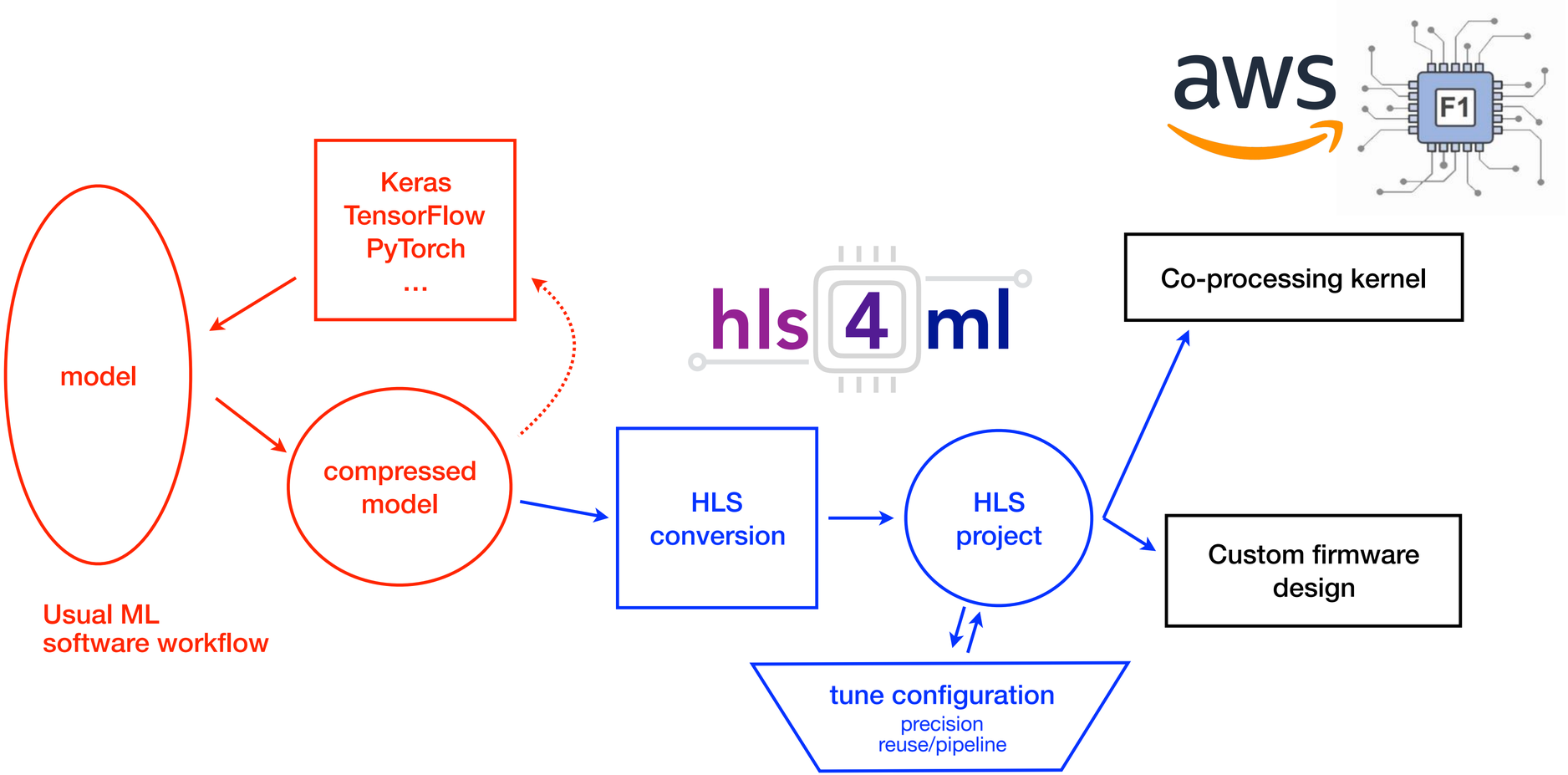
- **hls4ml** for physicists or ML experts to translate **ML algorithms** into **FPGA firmware**



REAL-TIME ML WITH HLS4ML

J. Instrum. 13, P07027 (2018)

- **hls4ml** for physicists or ML experts to translate **ML algorithms** into **FPGA firmware**



MANY INTERESTING APPLICATIONS

arXiv:2006.10159

arXiv:2002.02534

Physics data representations are unique and varied — we want the right tool for the job!

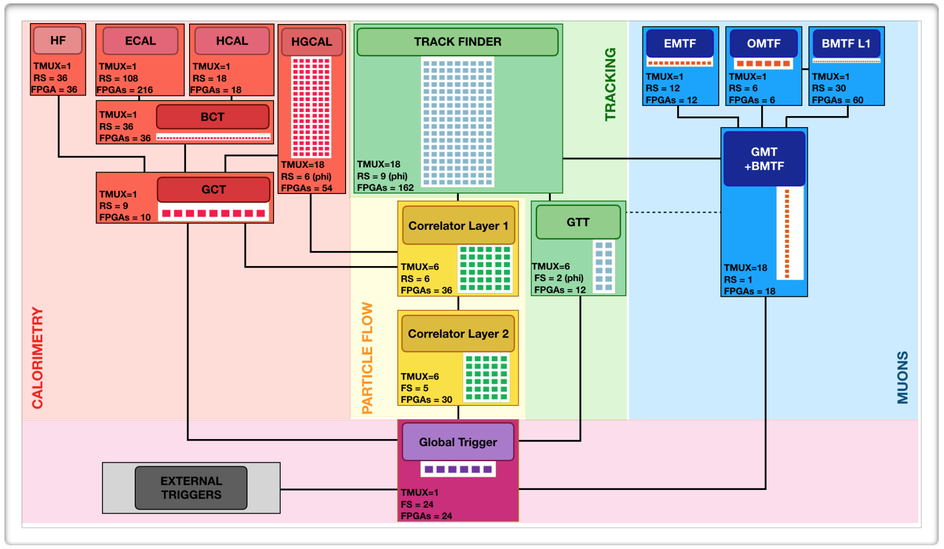
Quantization-aware training

Physics-inspired architectures

BDT

MLP

arXiv:1804.06913



LSTM

Conv2D

BNN/TNN

arXiv:2003.06308

Autoencoder

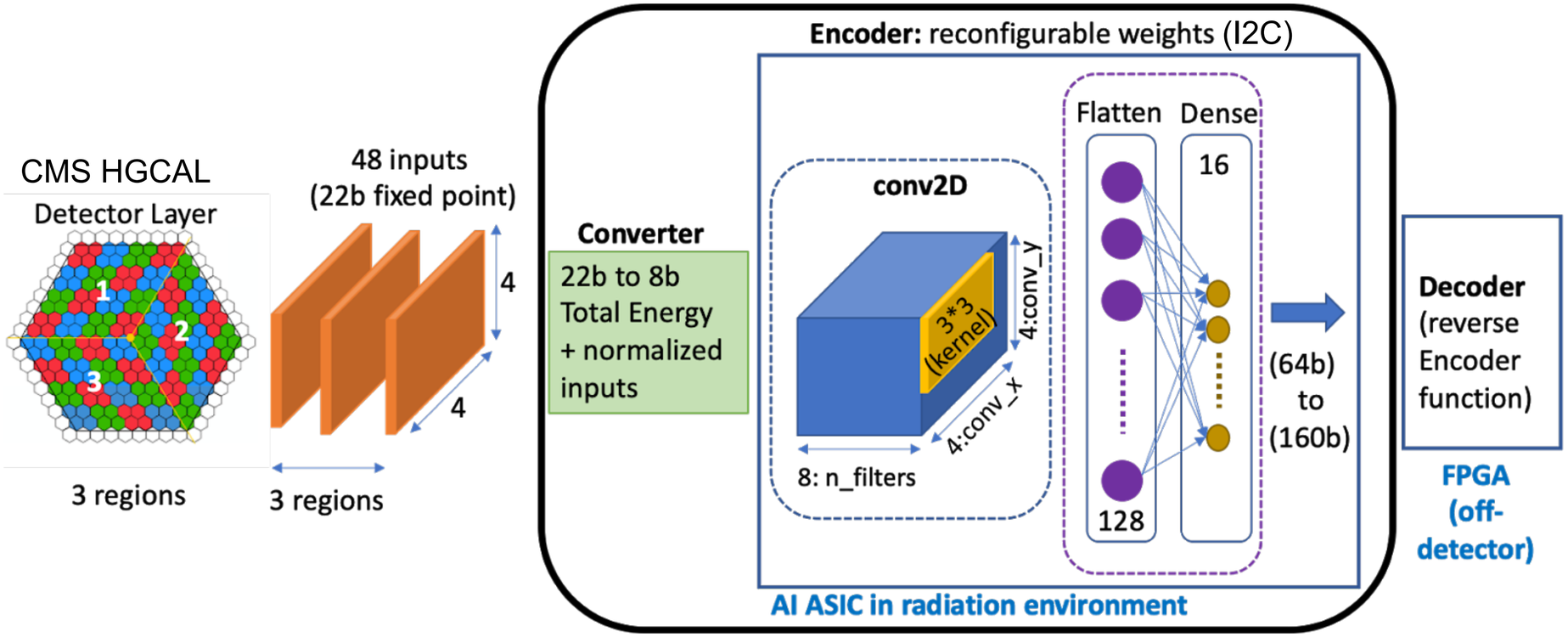
Graph NN

arXiv:2008.03601

ML "ON DETECTOR"

See: https://indico.fnal.gov/event/44698/contributions/192923/

- ML can be extended to detector frontend ASICs: bringing processing / data reduction as close to detector close as possible
- Example: autoencoder for data compression in frontend CMS HGCal ASIC

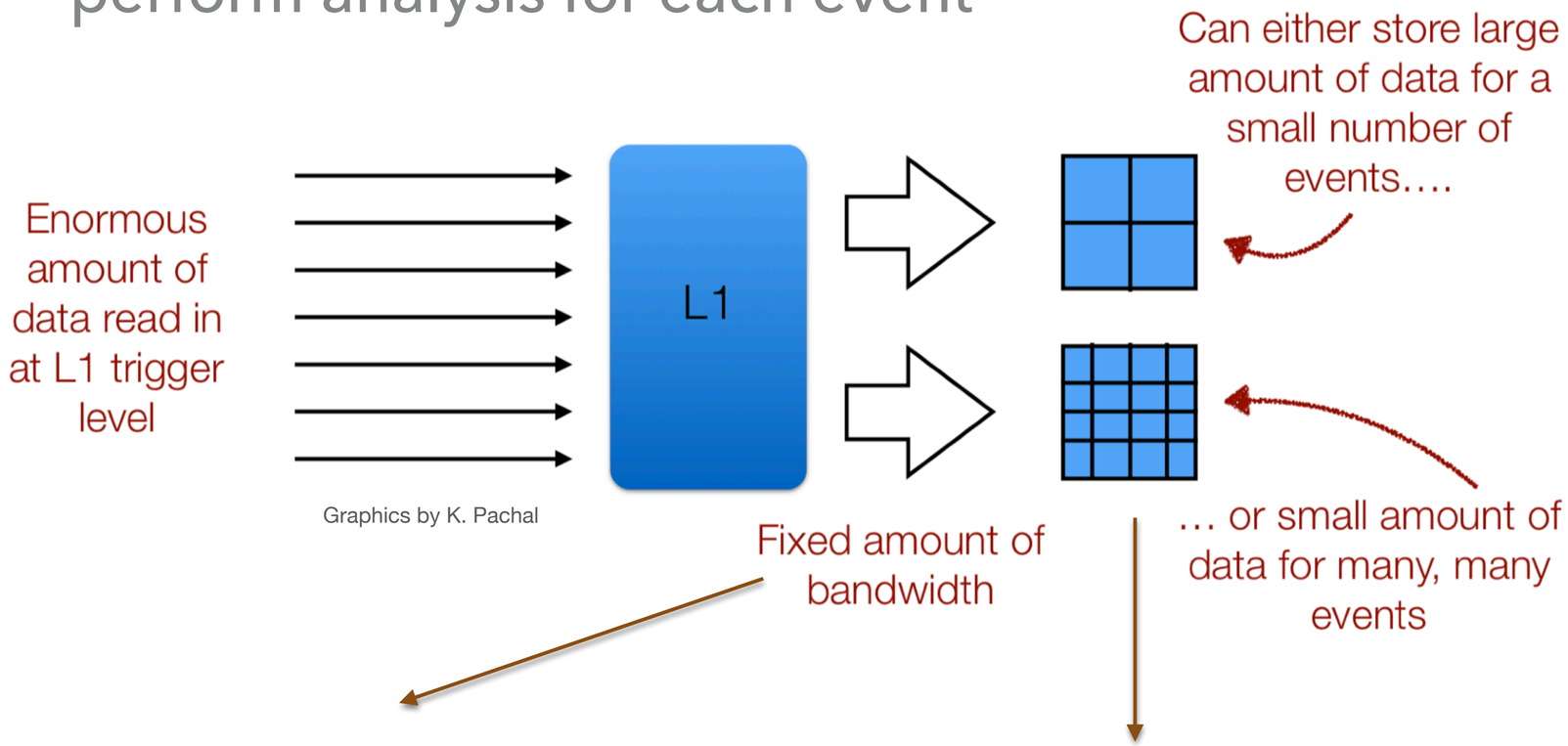


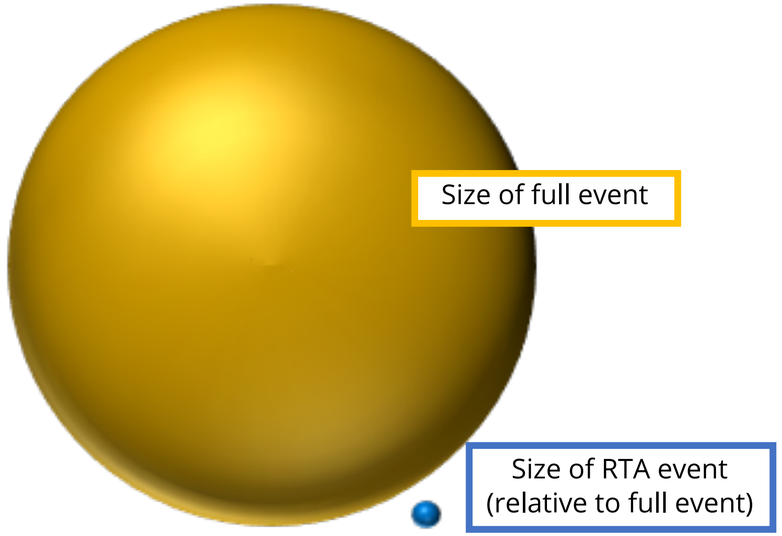
"DATA SCOUTING" OR SELECTIVE PERSISTENCE OF EVENTS

- Related: real-time analysis & selective persistence
  - Record only essential information to perform analysis for each event

See: https://indico.fnal.gov/event/44402/contributions/191632

Data scouting (CMS)
Trigger-level Analysis (ATLAS)
Turbo stream (LHCb)
Mu3e





Perform as much "analysis" as possible @ HLT

Reduced data formats

SUMMARY AND OUTLOOK

- ML is here to stay in HEP
- How to leverage existing and future computing resources/platforms for ML-based workflows is an active area of research
  - Both “offline” and “online” applications abound
- Interested groups include
  - Fast ML: https://fastmachinelearning.org/
  - RealTime ASG: https://realtime.blogg.lu.se/
  - SMARTHEP: http://www.smarthep.org/
  - IRIS-HEP: https://iris-hep.org/
  - HSF: https://hepsoftwarefoundation.org/
  - Exa.TrkX: https://exatrkx.github.io/
- Connections to many other Snowmass groups:
  - IF04 Trigger and DAQ, IF07: Electronics/ASICs, CompF01: Experimental Algorithm Parallelization, CompF04: Storage and processing resource access (Facility and Infrastructure R&D)

SnowMass2030

- Snowmass Frontiers

MACHINE LEARNING FRONTIER

JAVIER DUARTE

SNOWMASS COMPUTATIONAL FRONTIER WORKSHOP

AUGUST 11, 2020

BACKUP