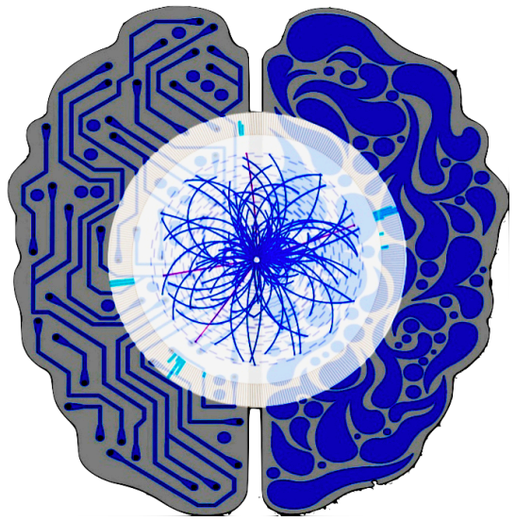


FPGAs-as-a-Service Toolkit (FaaST)

Dylan Rankin, Jeffrey Krupa, Philip Harris
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Maria Acosta Flechas, Burt Holzman, Thomas Klijnsma, Kevin Pedro, Nhan Tran
Fermi National Accelerator Laboratory
Batavia, IL 60510, USA

Scott Hauck, Shih-Chieh Hsu, Matthew Trahms, Kelvin Lin, Yu Lou
University of Washington
Seattle, WA 98195, USA

Ta-Wei Ho
National Tsing Hua University
Hsinchu, Taiwan 300044, R.O.C.

Javier Duarte
University of California San Diego
La Jolla, CA 92093, USA

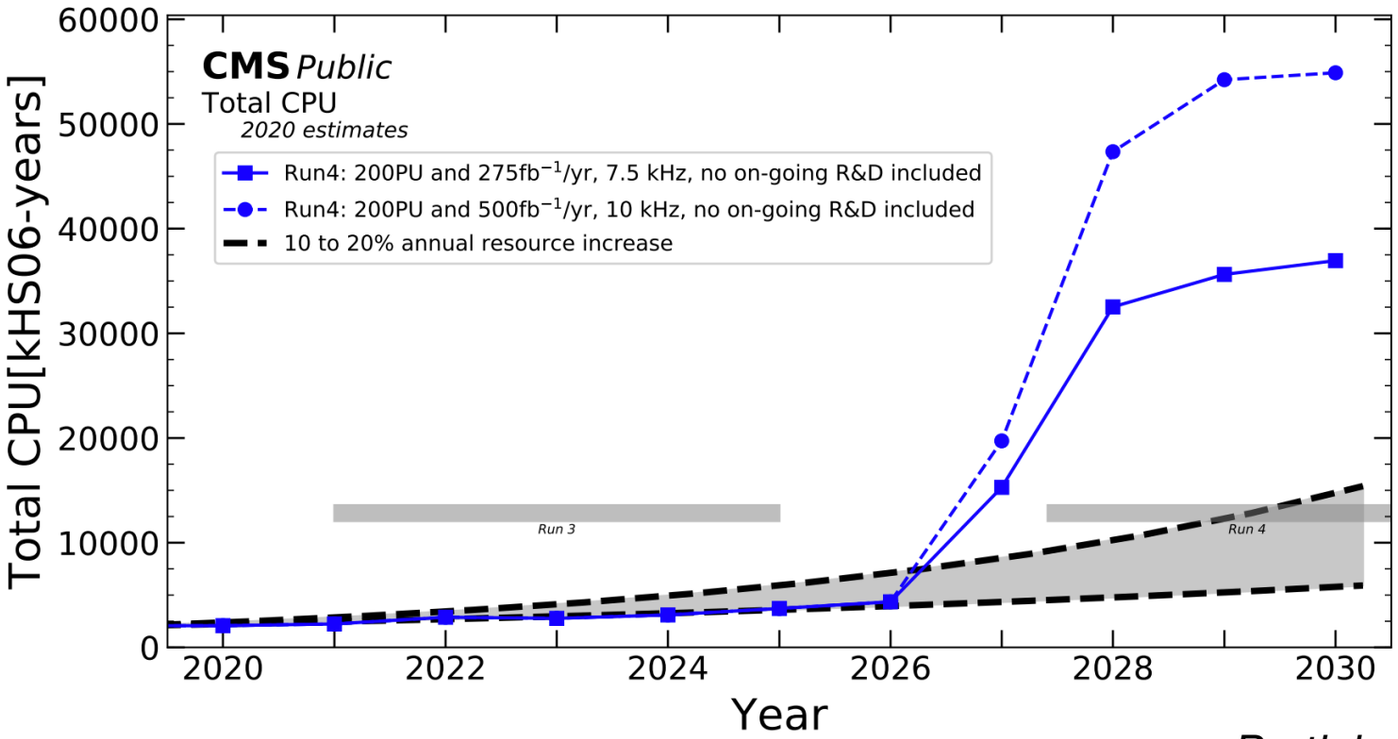
Mia Liu
Purdue University
West Lafayette, IN 47907, USA

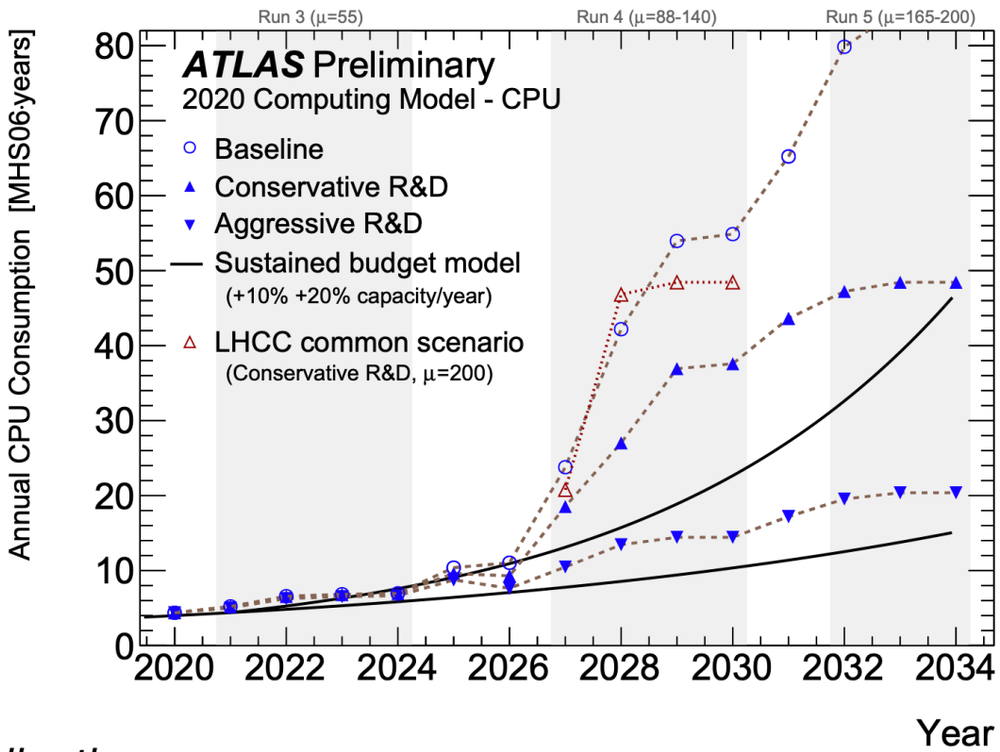
November 13, 2020

Sixth International Workshop on Heterogeneous High-performance Reconfigurable Computing (H²RC'20)

Motivation

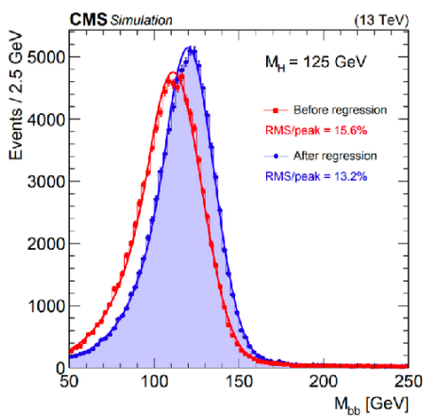
- Computing projections for high energy physics (HEP) greatly outpace CPU growth, interest in ML rapidly increasing



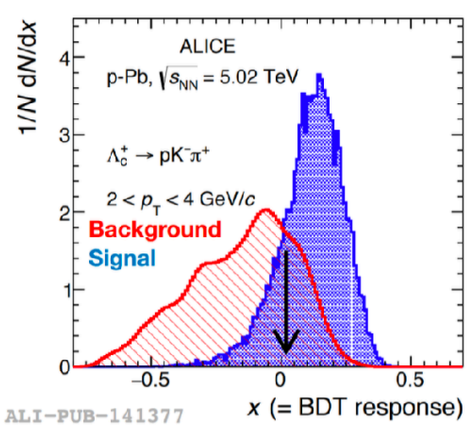


- **We see FPGAs as possible solution**
- How can we best use FPGAs for ML computing tasks in HEP?
  - → *As-a-service computing*

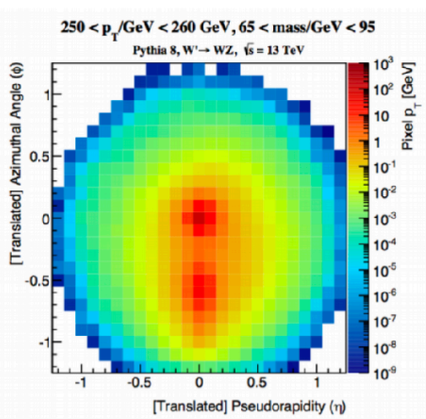
Particle collection energy regression



Signal/background classification

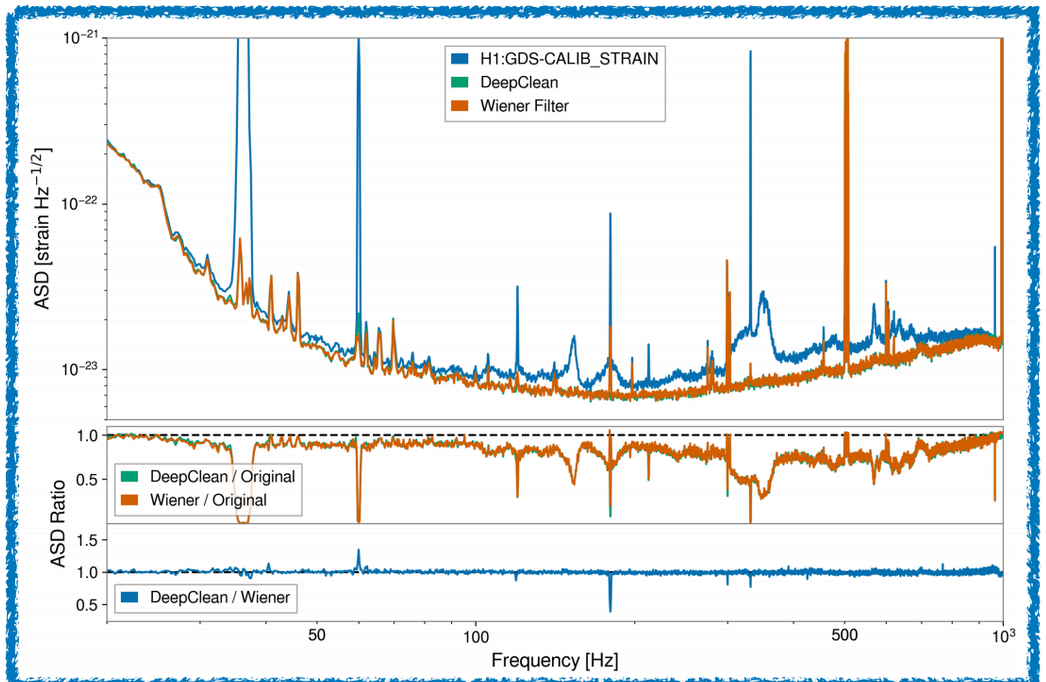


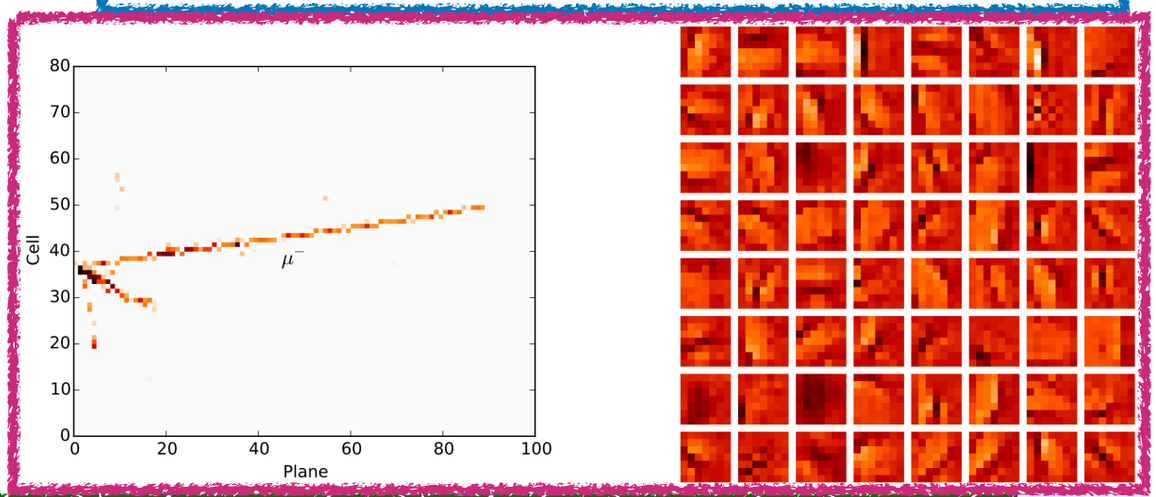
Particle classification

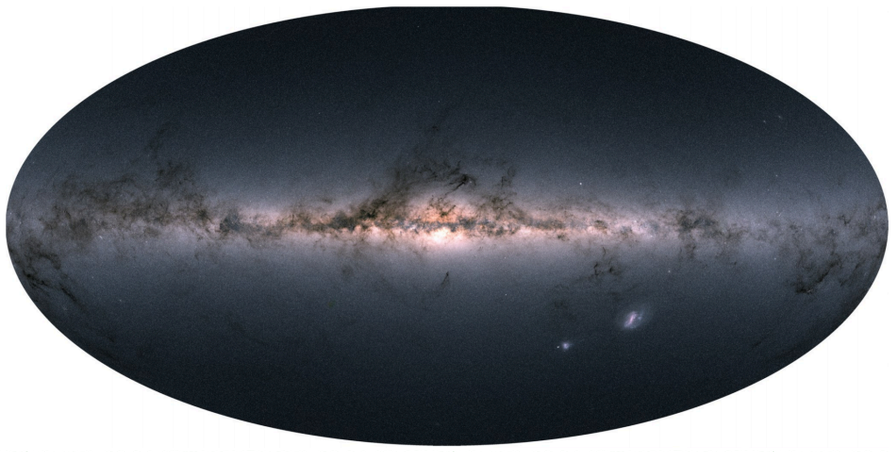


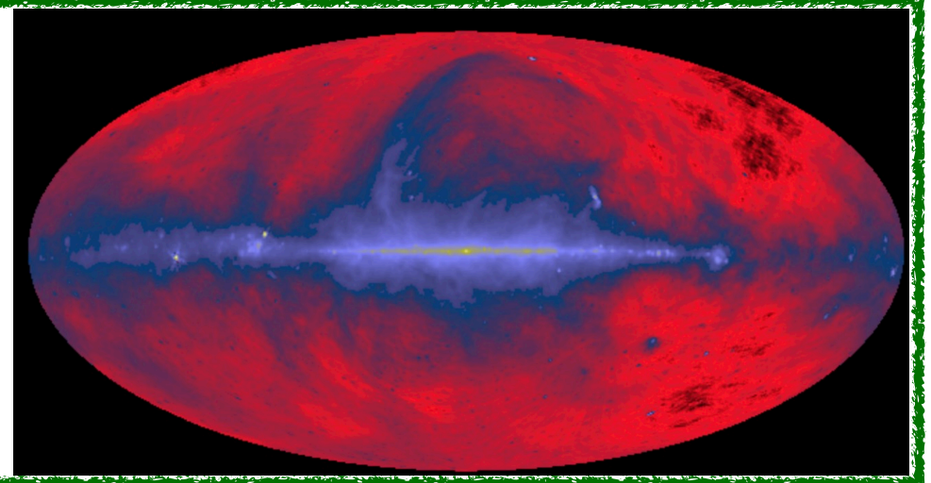
Applications

- FPGA compute as-a-service not only beneficial for our particular experiments
- Gravitational waves
- Neutrinos
- Multi-messenger astronomy



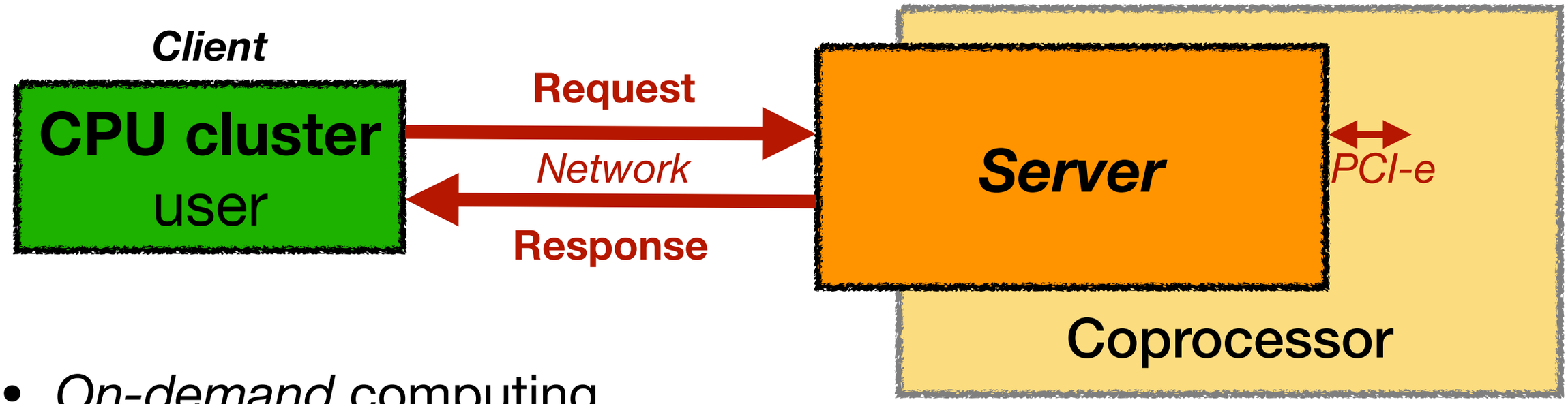






As-a-service Computing

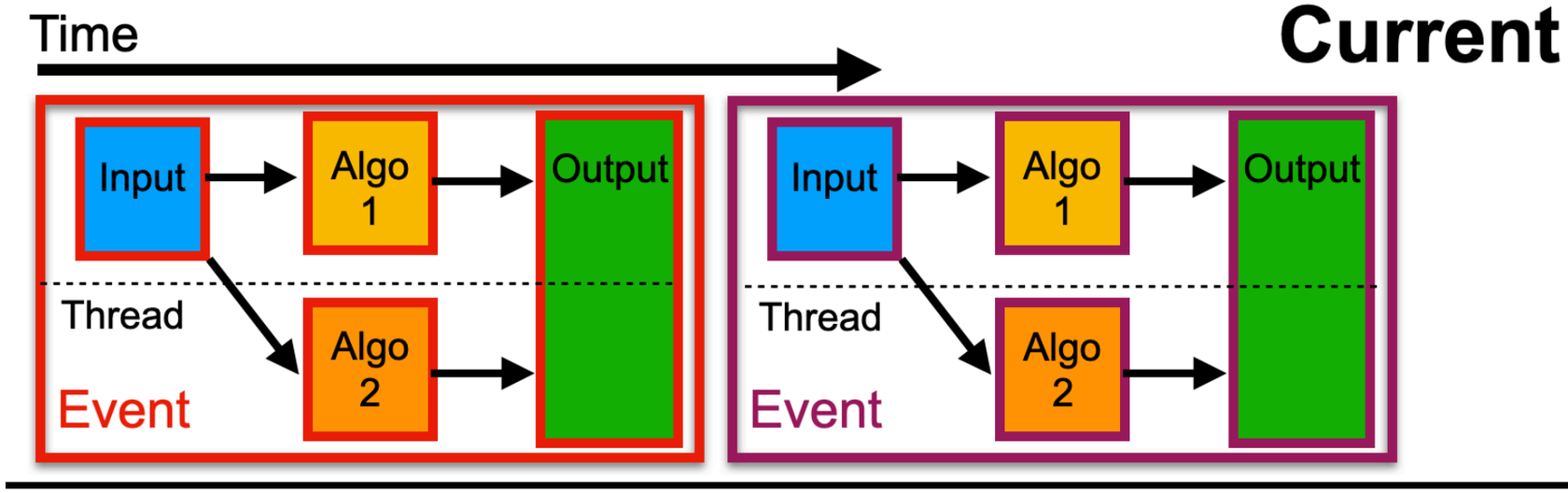
- As a user, I just want my workflow to run quickly



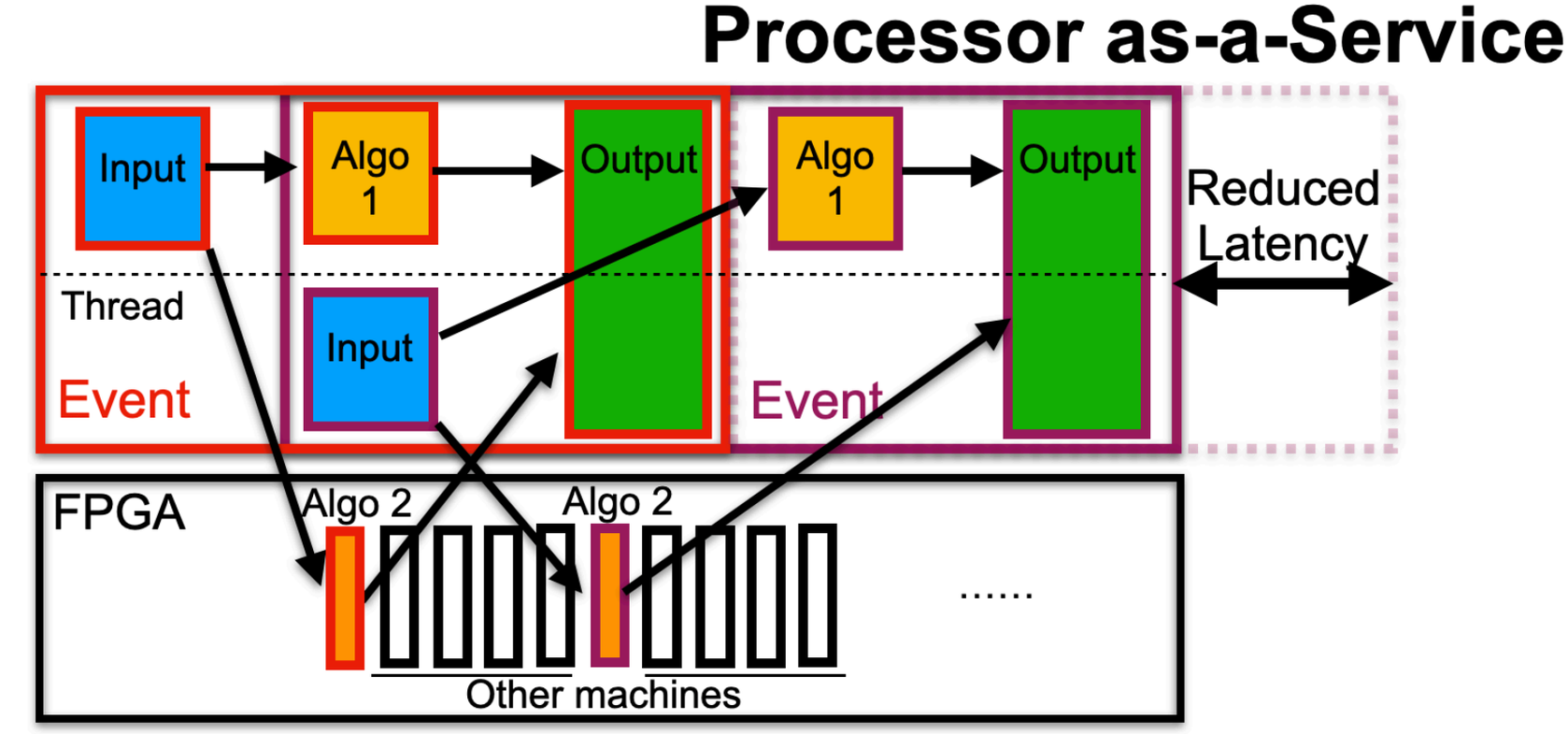
- *On-demand* computing
    - Client communicates with server CPU, server CPU communicates with coprocessor
- Many existing tools from industry, cloud

As-a-service Computing

- Can provide large speed up w.r.t traditional computing model
  - Scheduling important to improvement

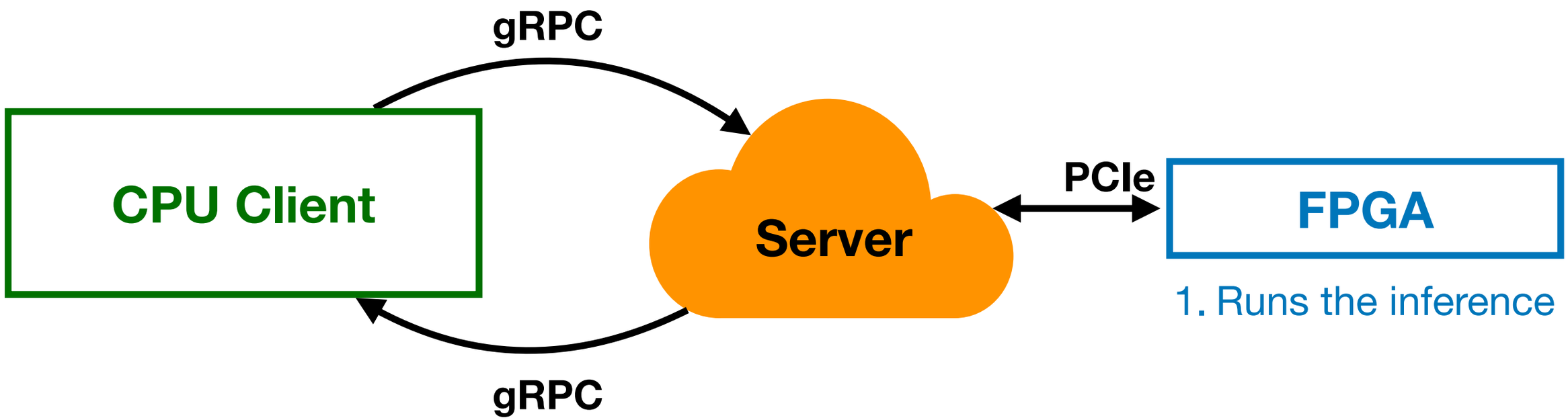


- Machine learning is particularly well-suited for as-a-service
  - Small number of inputs relative to large number of operations
  - Large speedups w.r.t CPU



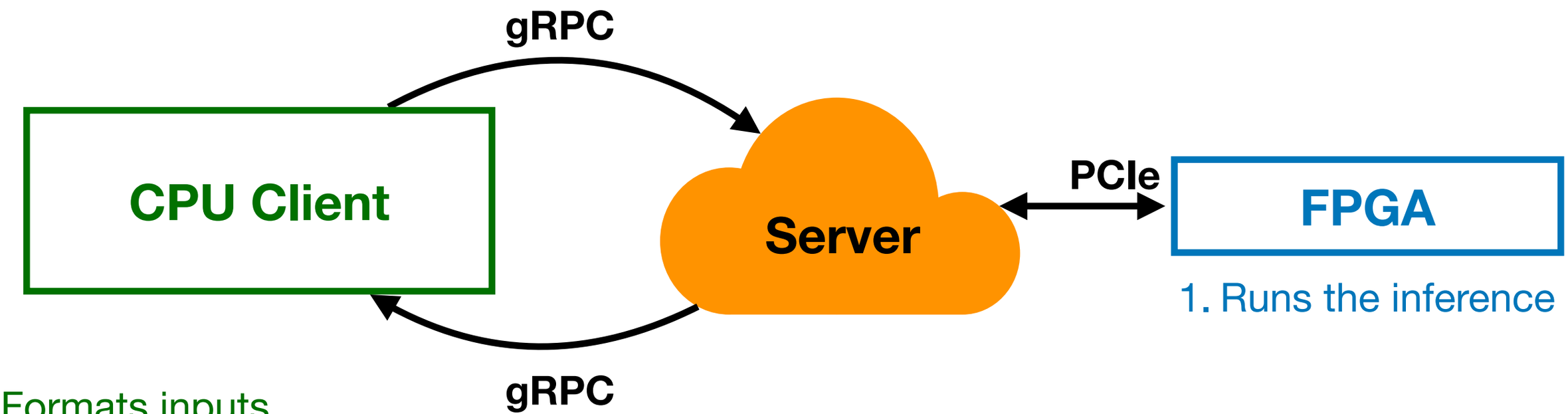
FPGAs-as-a-Service Toolkit

- Have developed cohesive set of implementations for range of hardware/ML models - refer to as ***FPGAs-as-a-Service Toolkit*** **(FaaST)**
- For fast inference we focus on gRPC protocol
  - Open source remote procedure call (RPC) system developed by Google



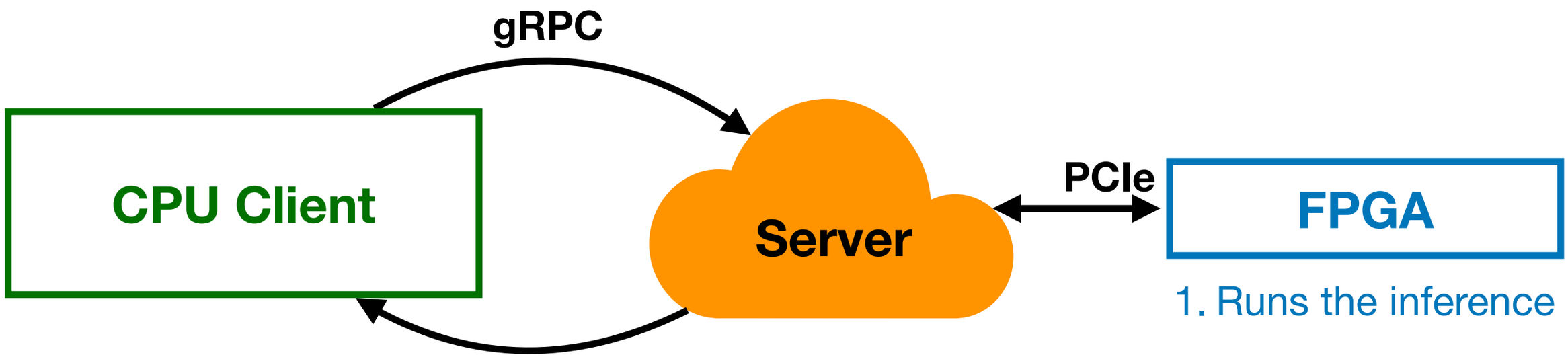
FPGAs-as-a-Service Toolkit

- Have developed cohesive set of implementations for range of hardware/ML models - refer to as ***FPGAs-as-a-Service Toolkit*** **(FaaST)**
- For fast inference we focus on gRPC protocol
  - Open source remote procedure call (RPC) system developed by Google



FPGAs-as-a-Service Toolkit

- Have developed cohesive set of implementations for range of hardware/ML models - refer to as ***FPGAs-as-a-Service Toolkit*** **(FaaST)**
- For fast inference we focus on gRPC protocol
  - Open source remote procedure call (RPC) system developed by Google

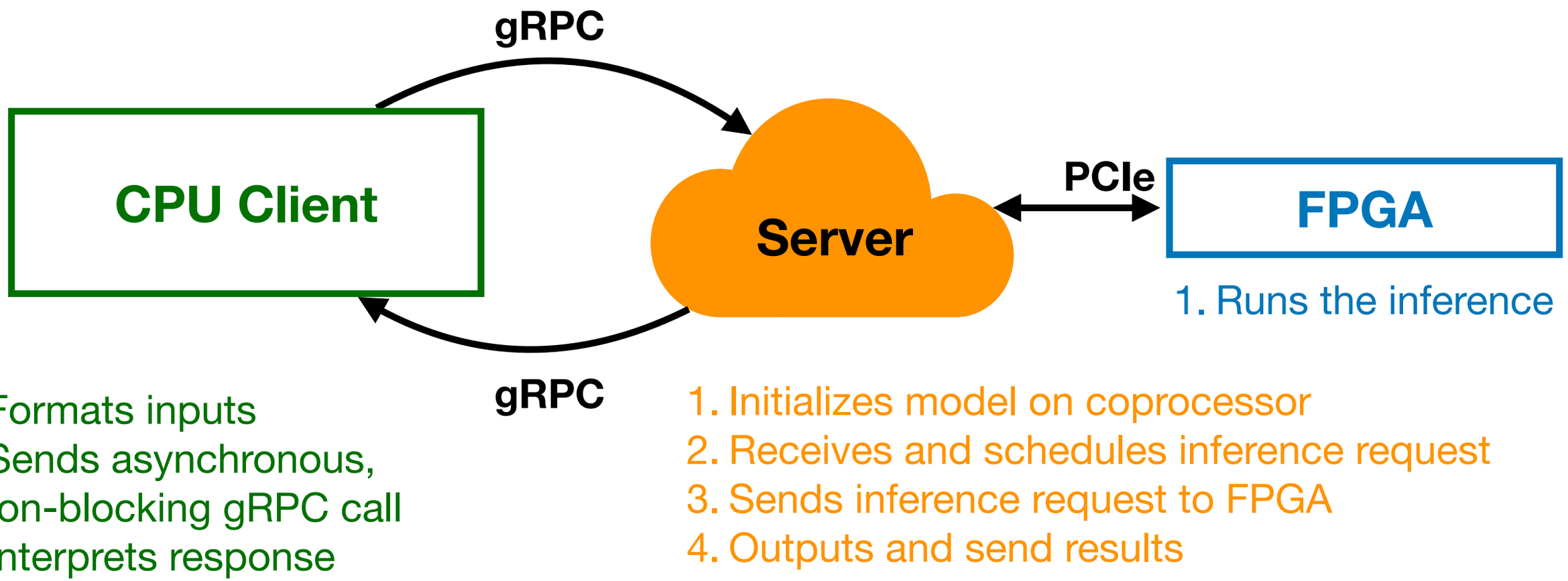


1. Formats inputs
2. Sends asynchronous, non-blocking gRPC call
3. Interprets response

1. Initializes model on coprocessor
2. Receives and schedules inference request
3. Sends inference request to FPGA
4. Outputs and send results

FPGAs-as-a-Service Toolkit

- Have developed cohesive set of implementations for range of hardware/ML models - refer to as ***FPGAs-as-a-Service Toolkit*** **(FaaST)**
- For fast inference we focus on gRPC protocol
  - Open source remote procedure call (RPC) system developed by Google



Tools:

hls4ml

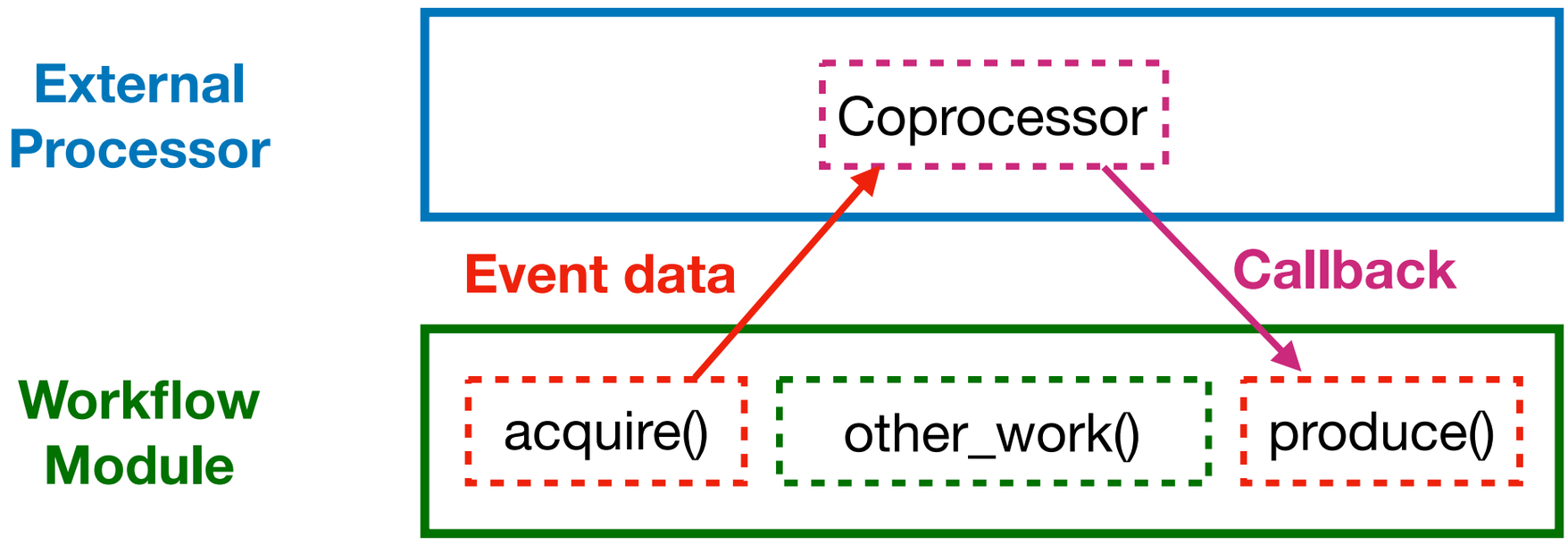
XILINX VITIS

XILINX Xilinx ML Suite

Azure Machine Learning Studio

SONIC

- FaaST compatible with Services for Optimized Network Inference on Coprocessors (SONIC) framework
- Integration of as-a-service requests into HEP workflows
  - Works with any accelerator
  - Requests are asynchronous, non-blocking



FaaST Server

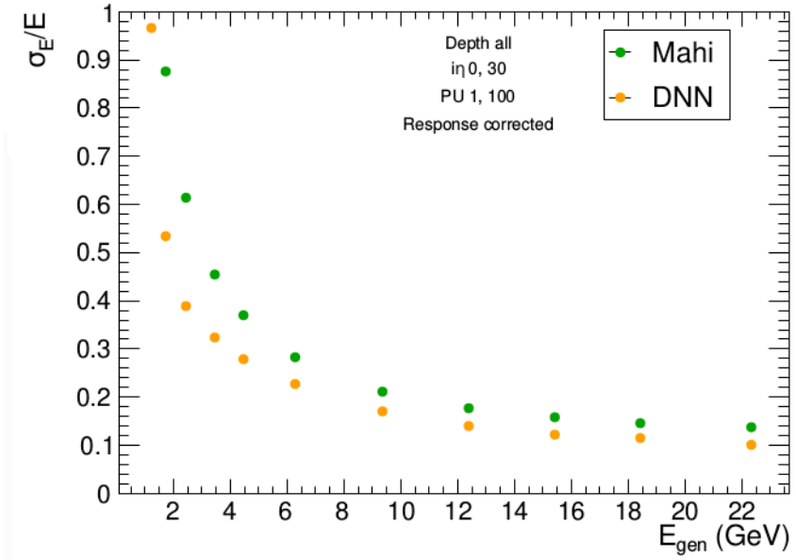
- Triton inference server developed by Nvidia for as-a-service inference on GPUs
  - Supports gRPC protocol
- FaaST designed to use same message protocol as Triton

- Server designed using various tools for different benchmarks
  - **FACILE:** XILINX VITIS™ + hls4ml (Alveo U250 & AWS f1)
  - **ResNet-50:** XILINX (AWS f1)
  - **ResNet-50:** Xilinx ML Suite, Azure Machine Learning Studio (Azure Stack Edge)

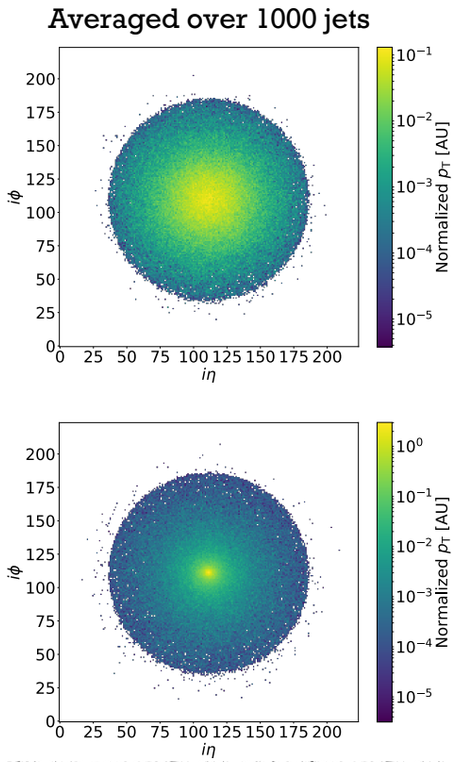
Benchmarks

FACILE

calorimeter energy regression
3-layer MLP

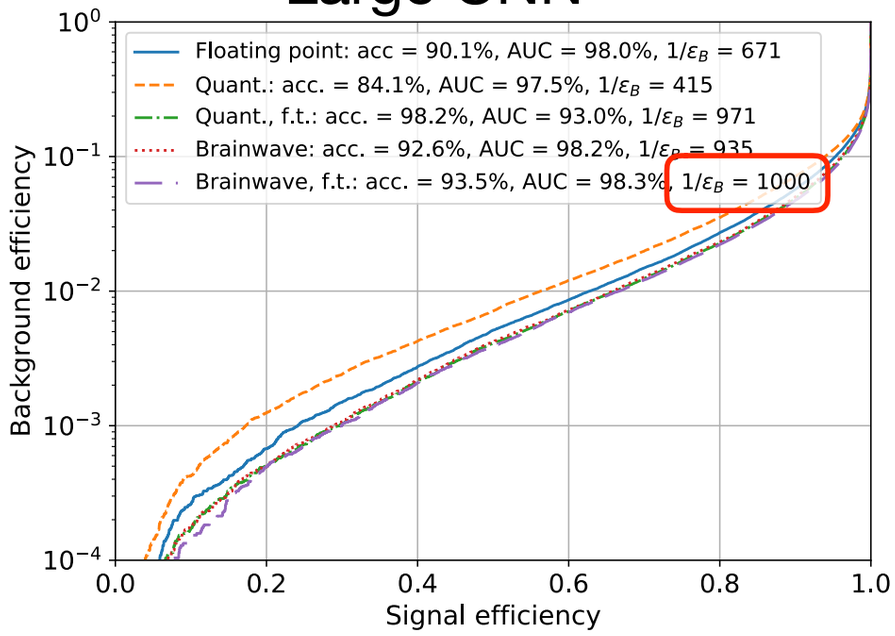


batch 16000



ResNet-50

top quark image classification
Large CNN

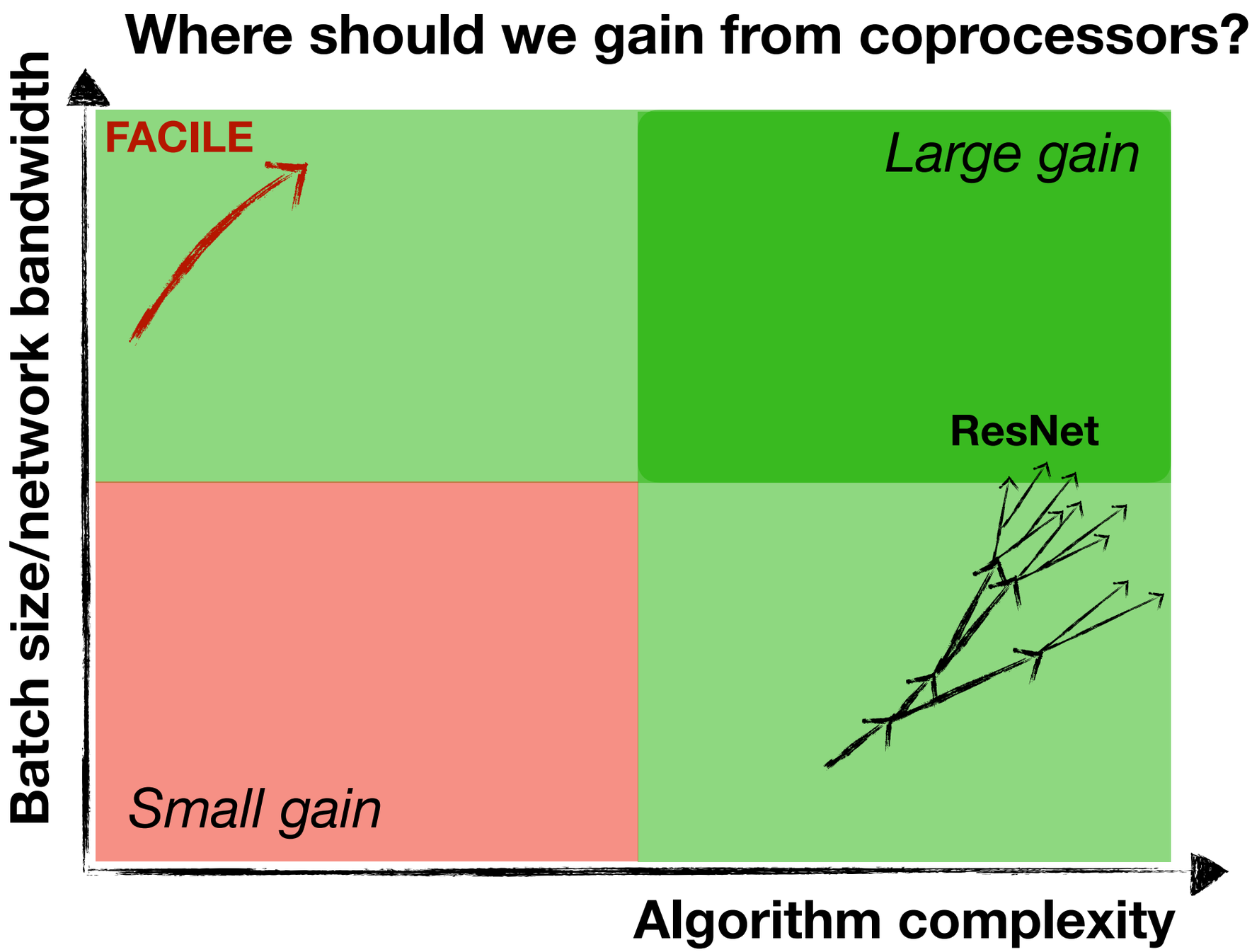


batch 10/batch 1

- Standard HEP data processing proceeds event-by-event
  - Batch sizes limited by event characteristics → smaller batches

2k parameters ⟷ **10M parameters**

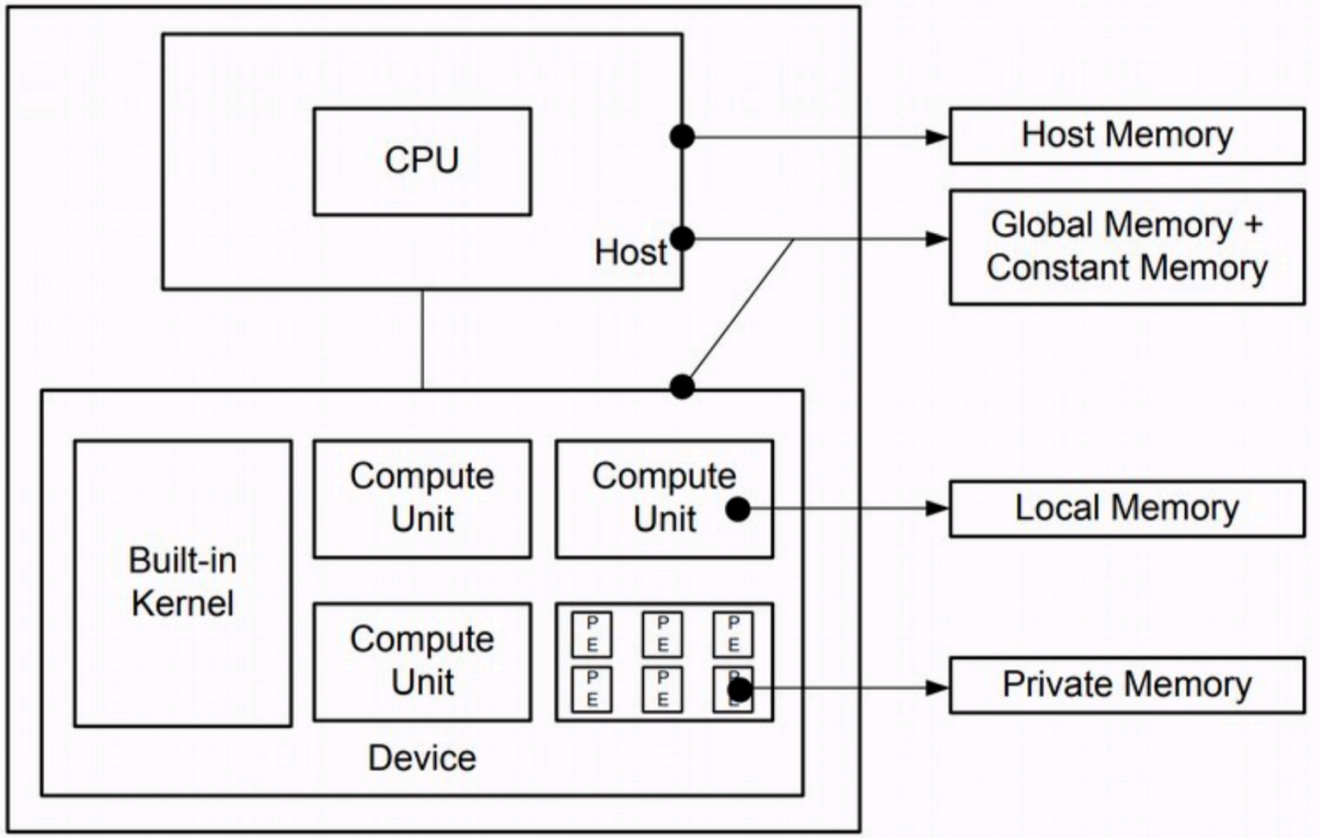
Gains


Where should we gain from coprocessors?
Batch size/network bandwidth
FACILE
Large gain
ResNet
Small gain
Algorithm complexity

- hls4ml is a software package for creating implementations of neural networks for FPGAs and ASICs
  - https://fastmachinelearning.org/hls4ml/
  - arXiv:1804.06913
- Supports common layer architectures and model software, options for quantization/pruning
  - Output is a fully ready high level synthesis (HLS) project
- Customizable output
  - Tunable precision, latency, resources

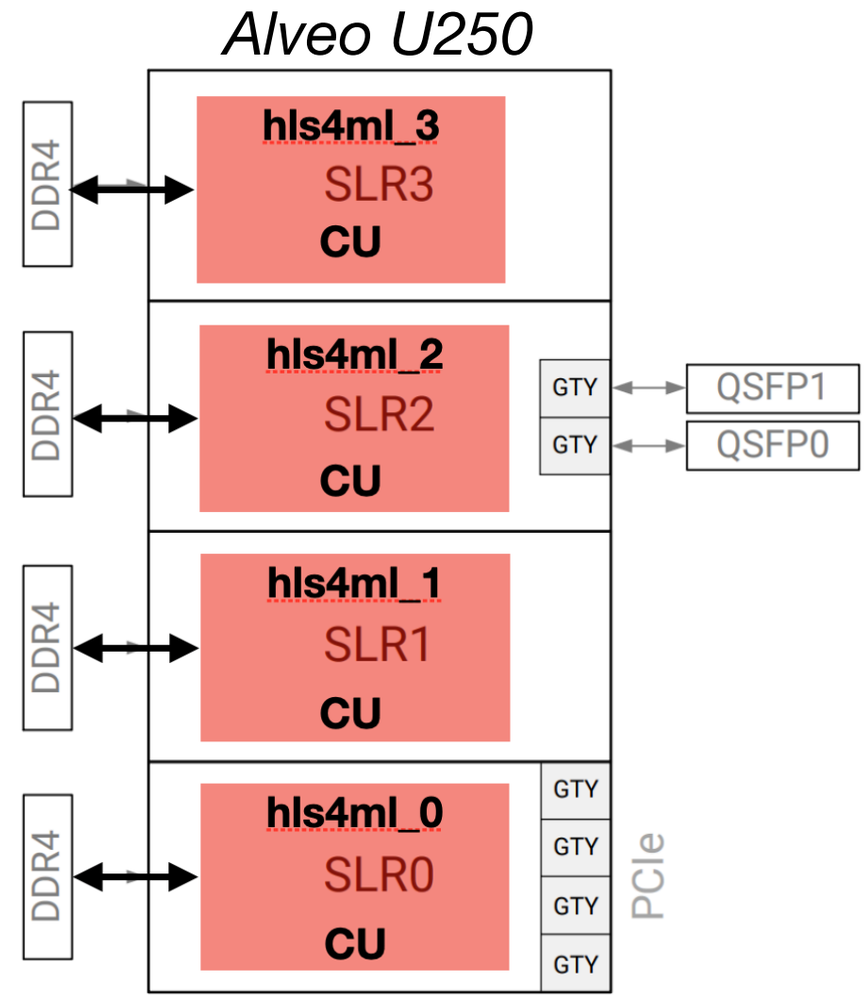
FACILE Server (XILINX VITIS + hls4ml)

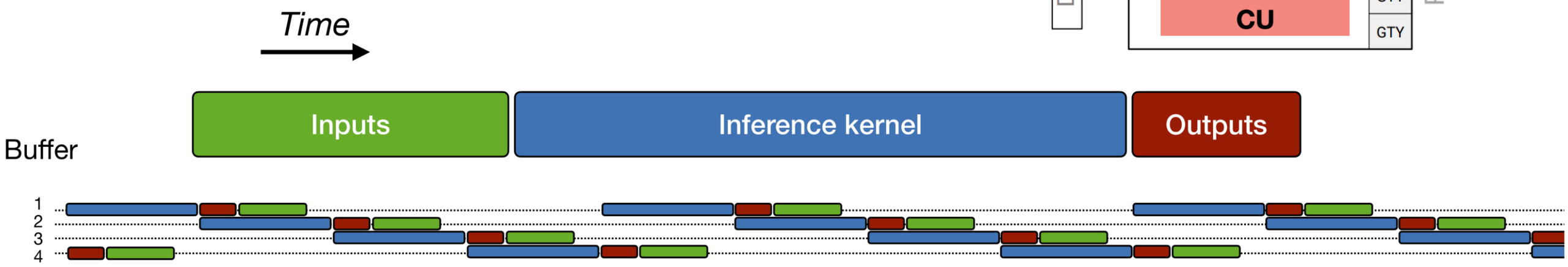
- Use Vitis Accel to manage data transfers, kernel execution
- Basic scheduling:
  - Copy batch 16000 inputs from host to FPGA DDR
  - Run hls4ml kernel
    - Tuned for low latency, pipelined, ~104 ns/inference
  - Copy 16000 batch outputs from FPGA DDR to host
- Server responsible for transferring input to dedicated buffers in host memory
- Set up for Alveo U250, AWS f1



FACILE Server (XILINX VITIS + hls4ml)

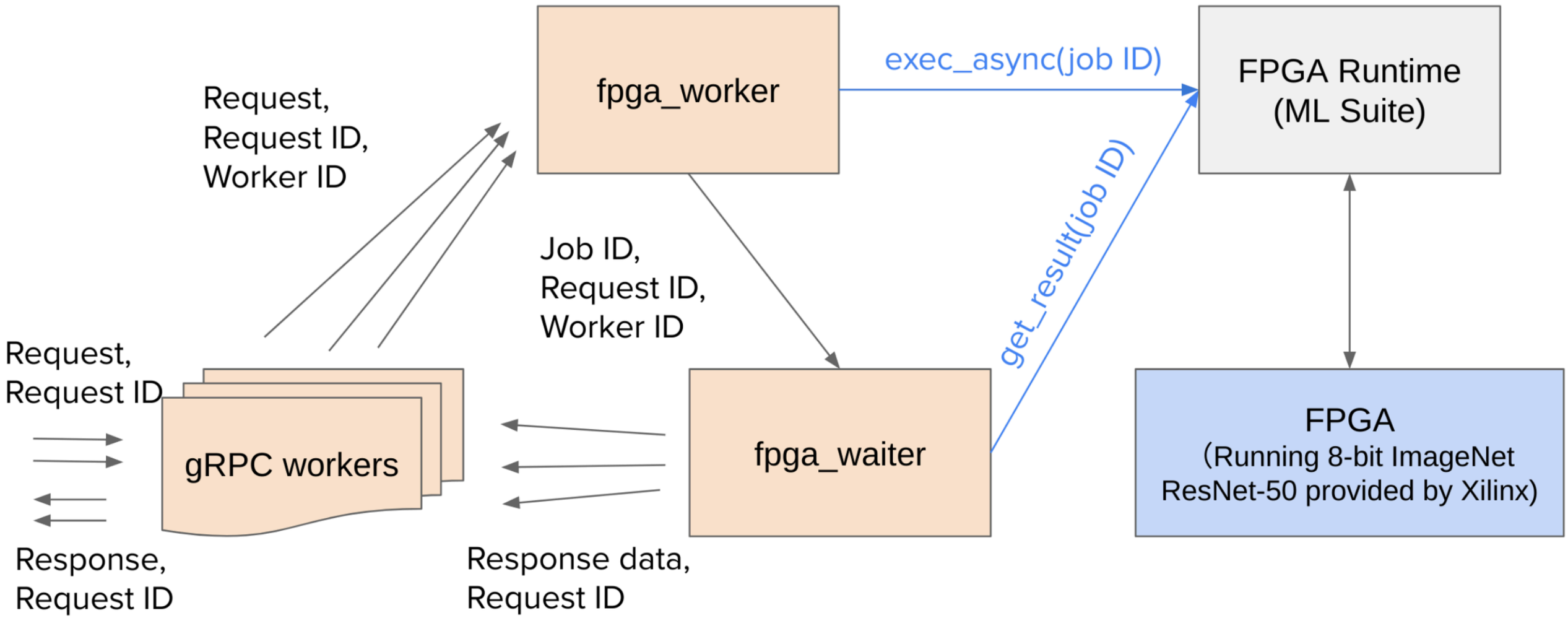
- Large amount of server optimization
- Can create multiple copies of hls4ml inference kernel on separate SLRs
- Can create buffer in DDR for multiple inputs, cycle through buffers





ResNet Server (XILINX Xilinx ML Suite)

- Similar server interface designed for ResNet / Xilinx ML Suite
- Set up for AWS f1

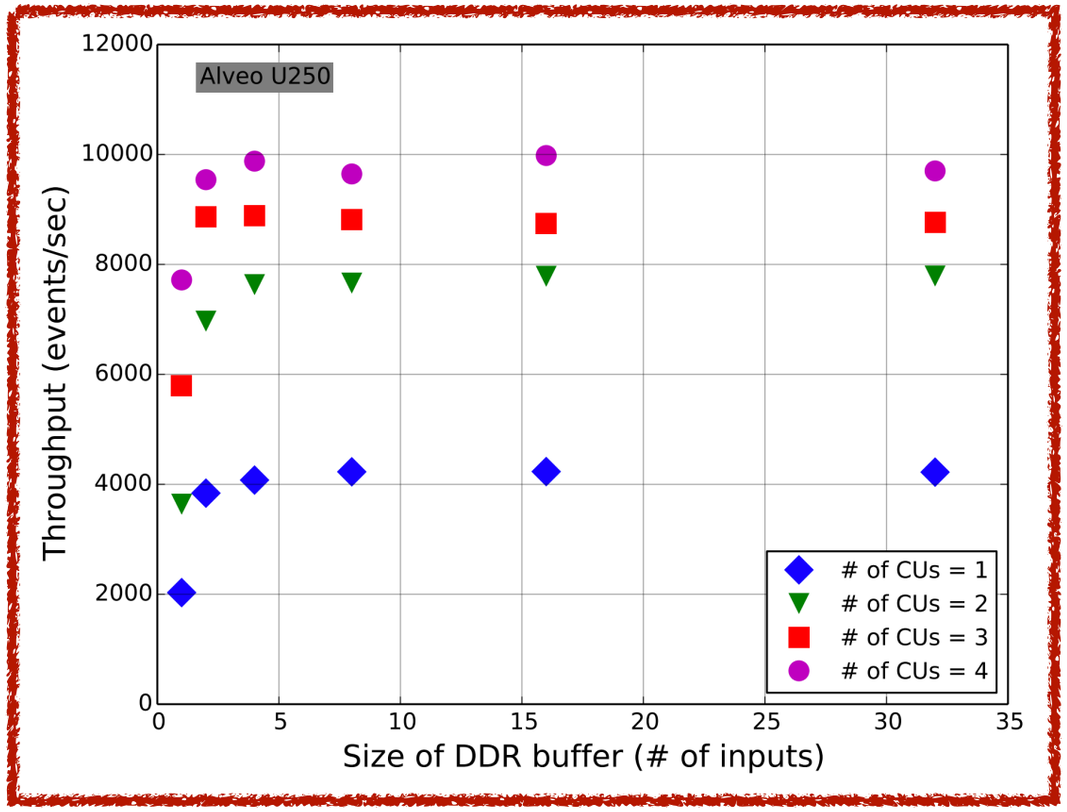


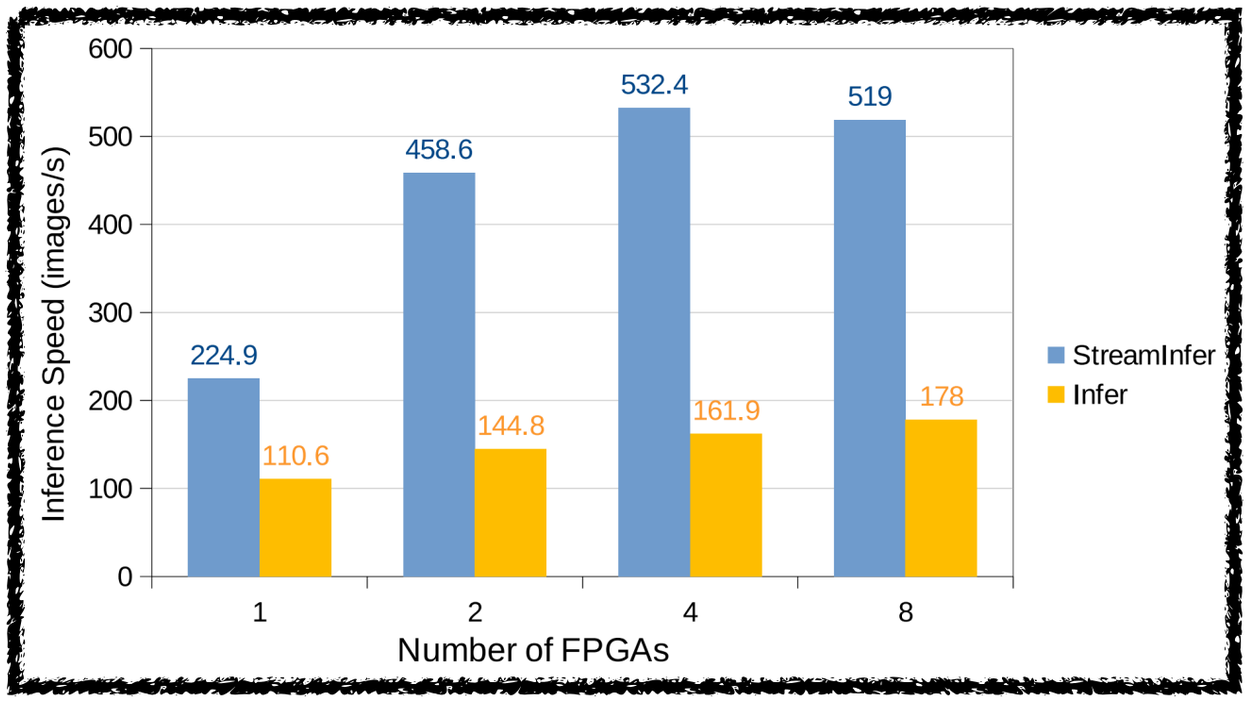
ResNet Server ( )

- Microsoft Azure Machine Learning Studio works with Azure Stack Edge server
  - Intel Arria 10 FPGA
  - Predefined list of ML models (including ResNet-50)
- Out-of-the-box solution accepts gRPC calls
- Installed locally at Fermilab

Server Optimization

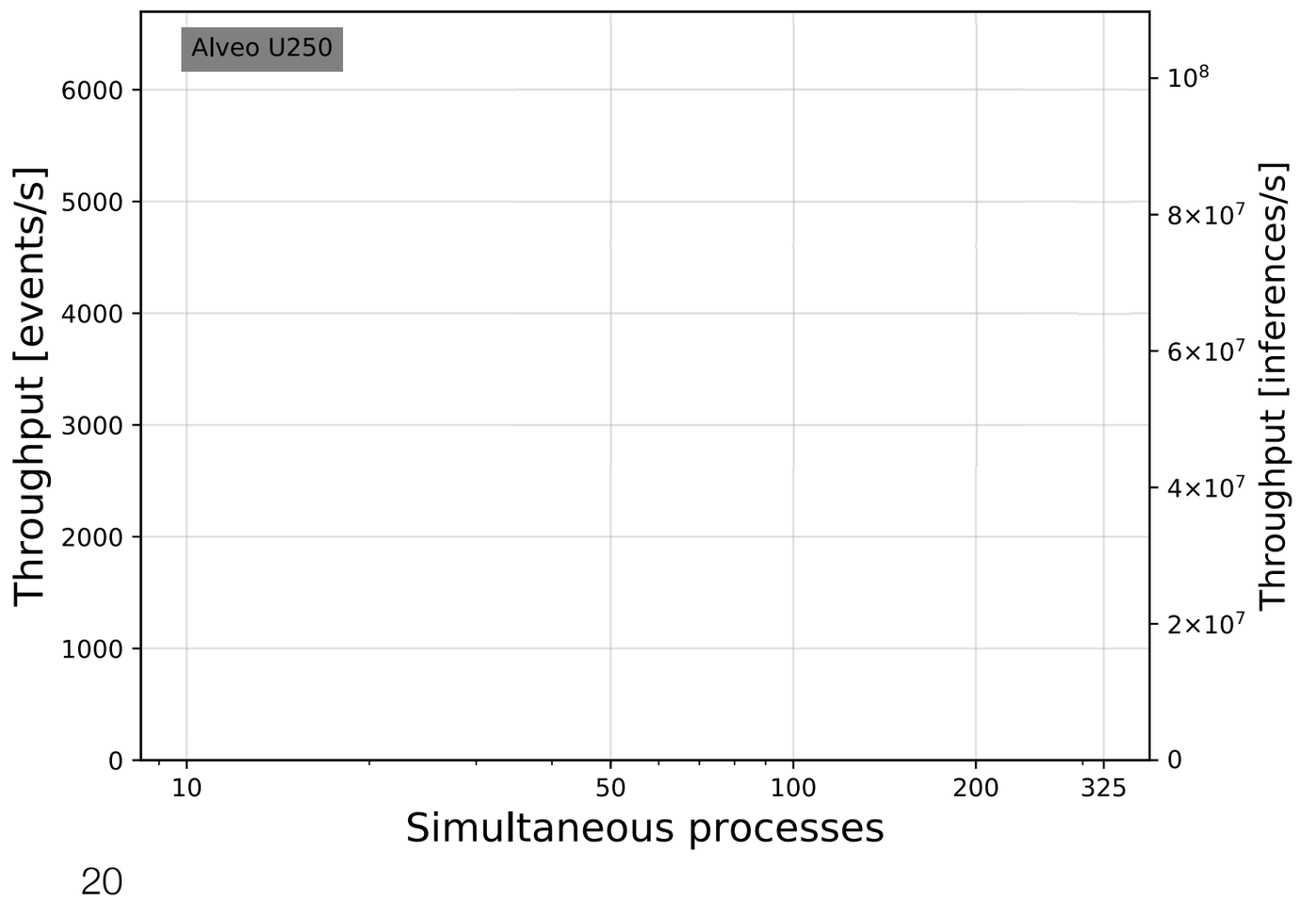
- Many settings to tune
- **FACILE**: scan of CU duplication and DDR buffer size
- **ResNet**: streaming gRPC inference calls found to greatly increase throughput
- Both: proxies to manage requests, distribute to multiple gRPC server endpoints





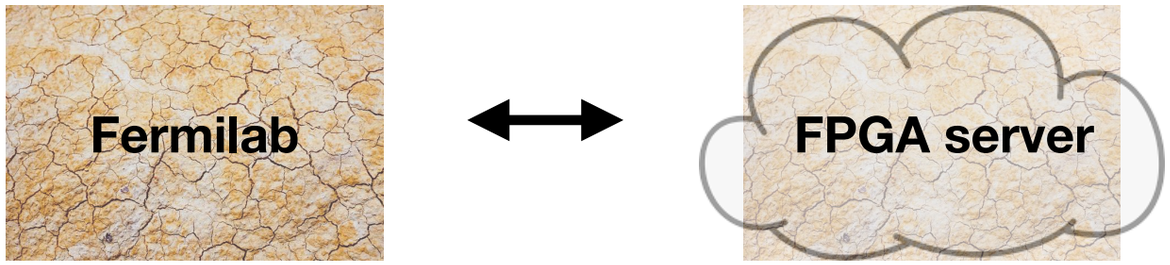
Throughput Tests

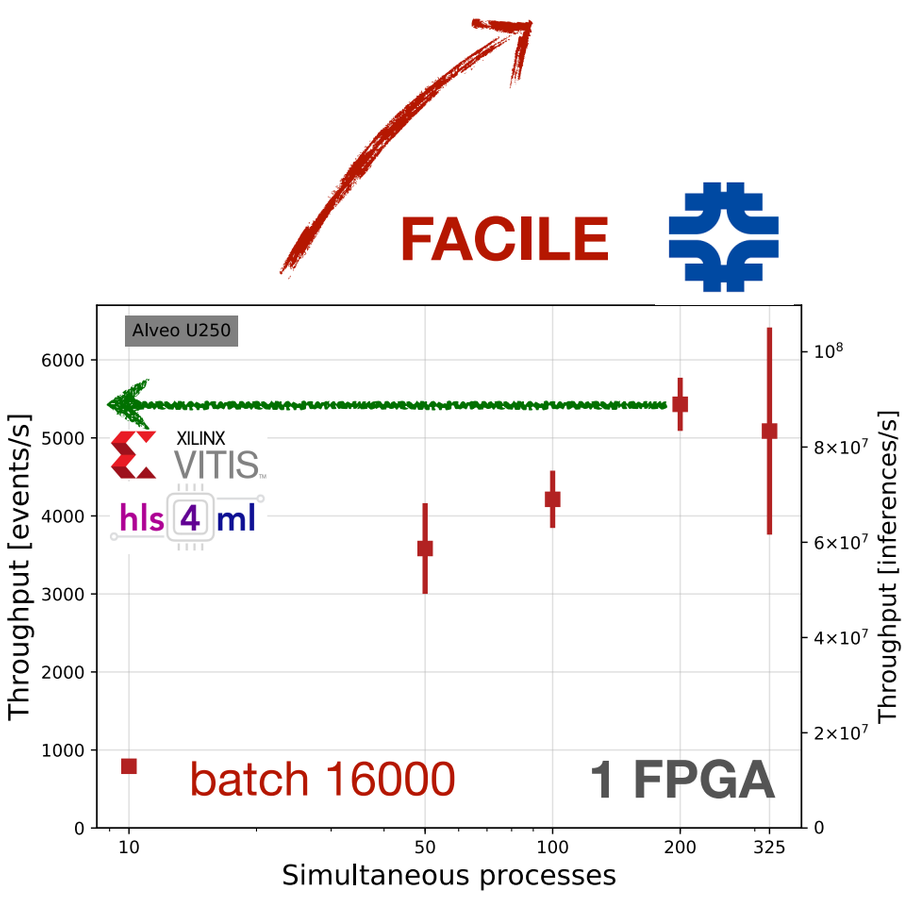
- *What is the maximum throughput of the server?*
- Start server (local/cloud), create N client processes at Fermilab computing cluster
    - Workflow contains only accelerated processing module
- All processes begin running at the same time
    - Fixed number of events
- Measure time/throughput for each process

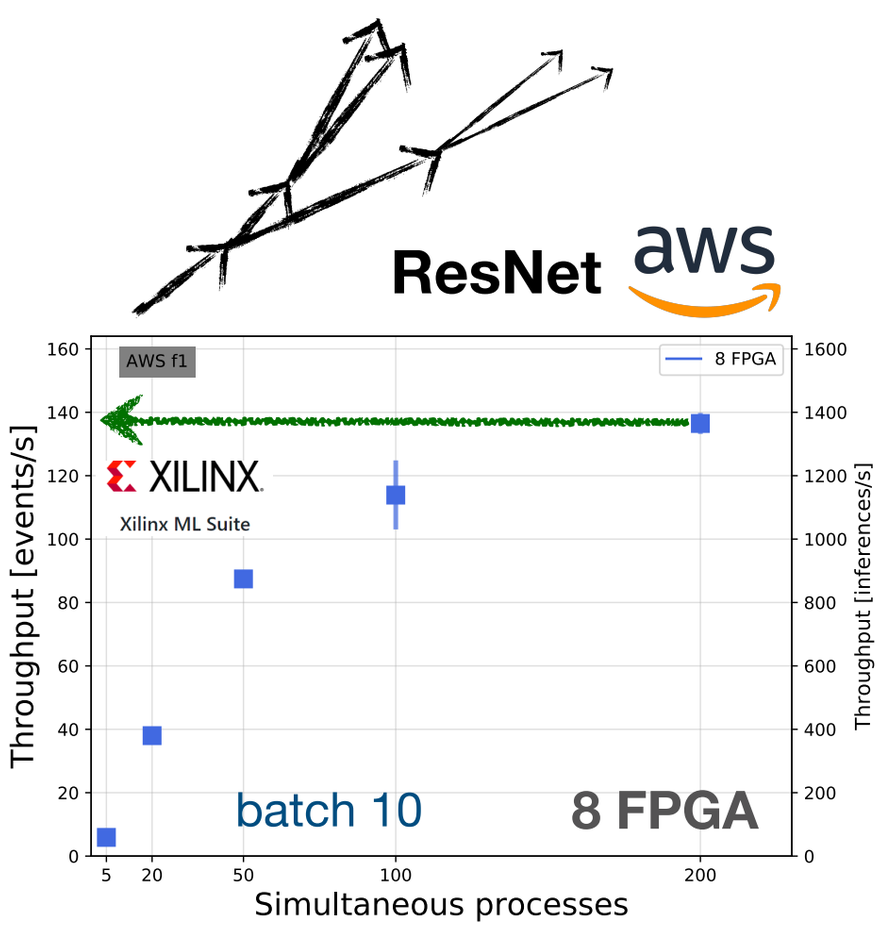


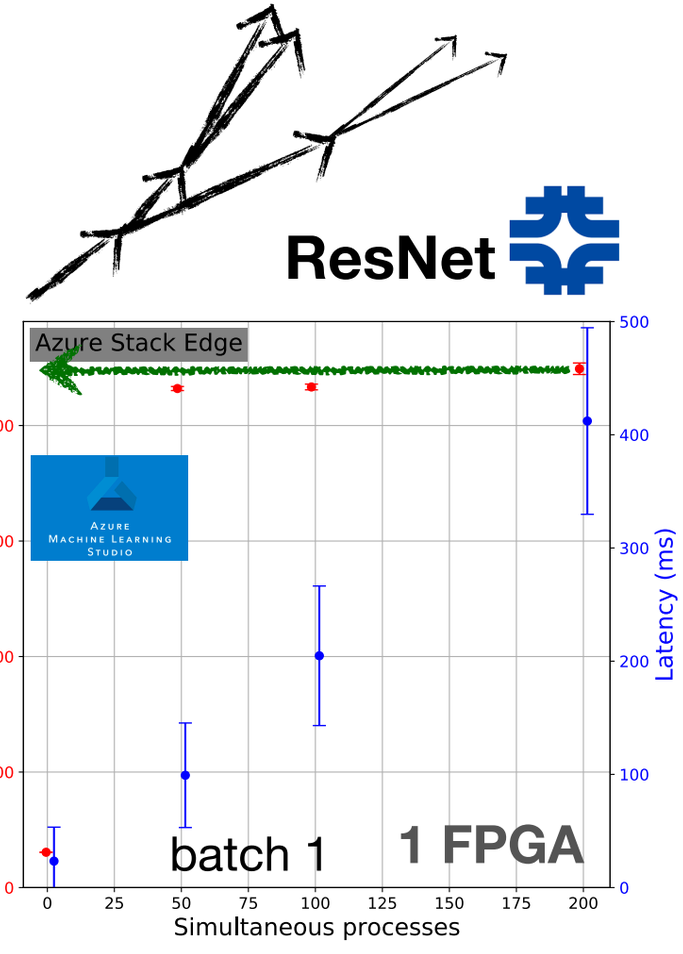
Throughput Tests

- With small **FACILE** network, server able to process over 5000 events/s
  - Limitation from CPU
- **ResNet** performance depends on hardware/specs



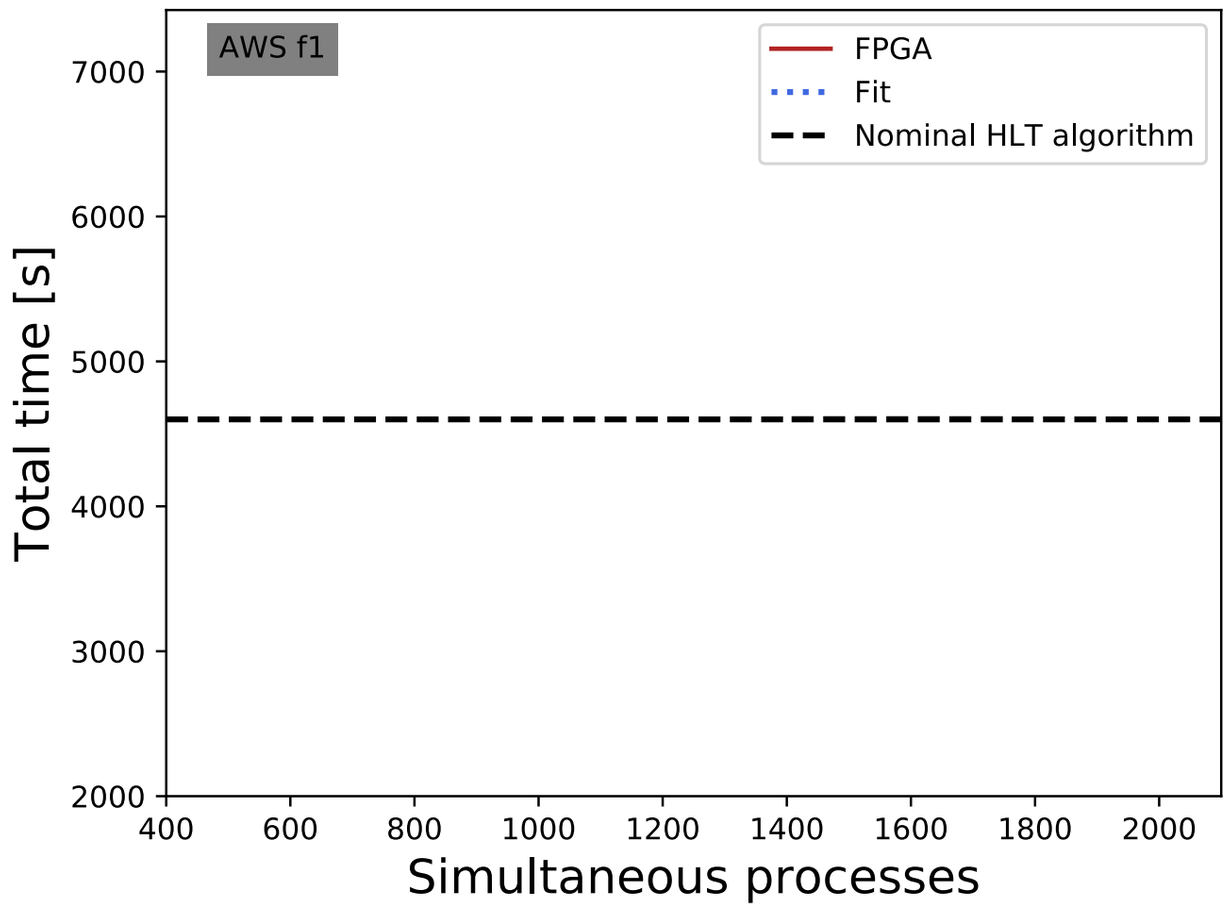






Scalability Test

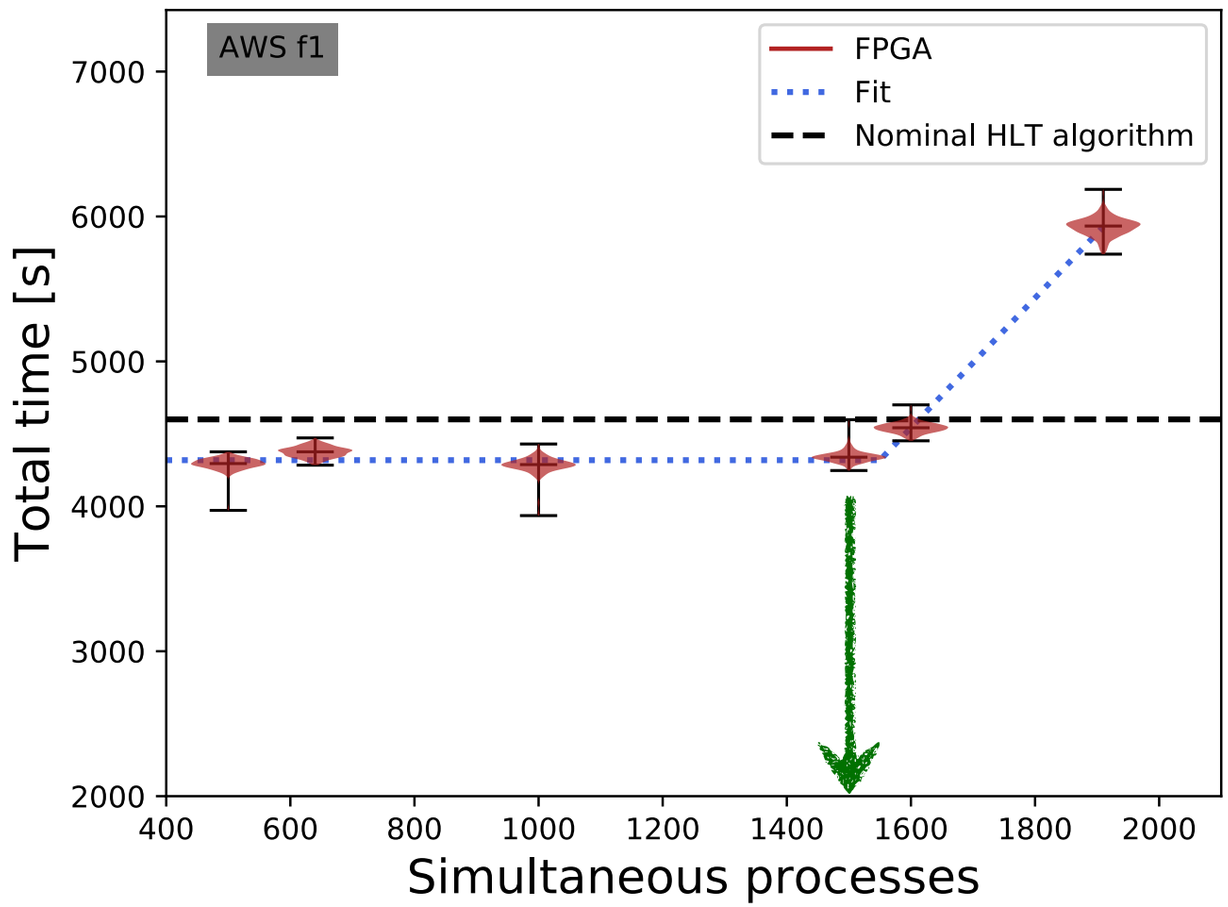
- *How many processes can a single server realistically serve?*
- Start server, create N client processes
  - Running realistic HEP high level trigger (HLT) workflow
  - HLT is fast reconstruction during data-taking traditionally performed using large CPU farm
- Compare **standard HLT** to **HLT with calorimeter reconstruction replaced by FaaST server running FACILE**
- Use HEPCloud to manage clients



Scalability Test

- **10% reduction in computing time operating as-a-service**
  - Consistent with fraction of time spent on calorimeter reconstruction w.r.t total HLT time
  - → Maximal achievable reduction
    for this single algorithm

- No increase in latency until **1500 clients**
  - Single FPGA can service
    1500 HLT instances
- Limited by AWS bandwidth (25 Gbps)
  - On Alveo U250, without network limit, estimate saturation at ~3300 clients



Summary

- Comparison of results to GPUaaS results (arXiv:2007.10359)
- **FaaST** greatly outperfoms **GPUaaS** for **FACILE**
  - Small network, large batch is ideally suited for FPGA
- Comparable performance between **FaaST** and **GPUaaS** for **ResNet**

Algorithm	Platform	Number of Devices	Batch Size	Inf./s [Hz]	Bandwidth [Gbps]
FACILE	AWS EC2 F1	1	16,000	36 M	23
FACILE	Alveo U250	1	16,000	86 M	55
FACILE	T4 GPU	1	16,000	8 M	5.1
ResNet-50	AWS EC2 F1	8	10	1400	6.7
ResNet-50	V100 GPU	8	10	1,700	8.1
ResNet-50	ASE	1	1	460	2.2
ResNet-50	T4 GPU	1	10	250	1.2

Conclusions

- FPGAs have been used in HEP for decades
- As-a-service paradigm, recent developments in ML inference, provide opportunity to leverage FPGA compute for many additional applications
- ***FPGAs-as-a-Service Toolkit (FaaST)*** can help facilitate integration of FPGA compute into existing workflows
  - Our results focus on HEP (and LHC particularly)
  - Applicable many other fields
    - Astronomy, neutrinos, gravitational waves
- Look forward to the growth of heterogeneous computing for science

Thanks!

Google Cloud Platform

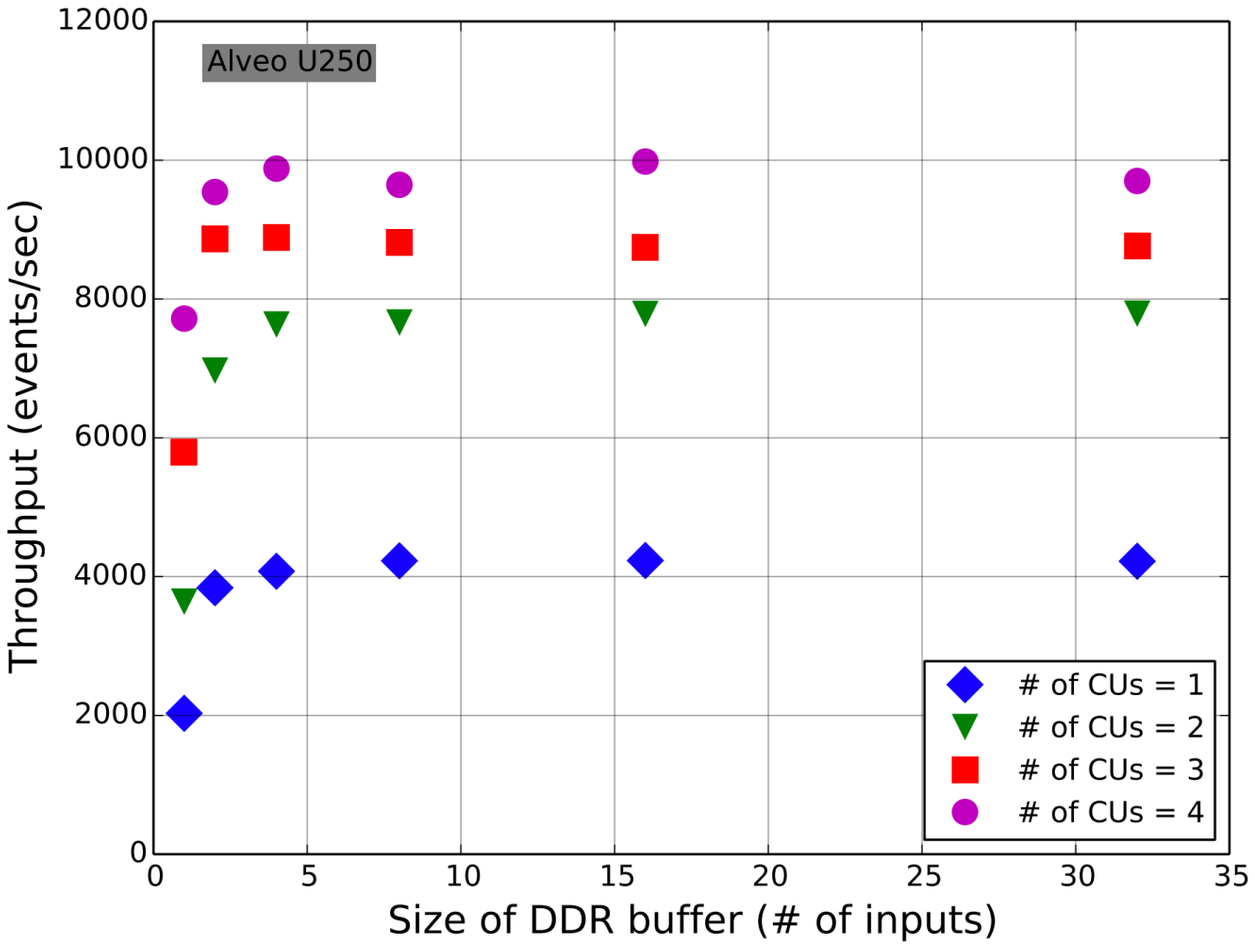
Institue for AI and Fundamental Interactions

Fast Machine Learning Lab

BACKUP

FACILE Optimization

Alveo U250



AWS f1