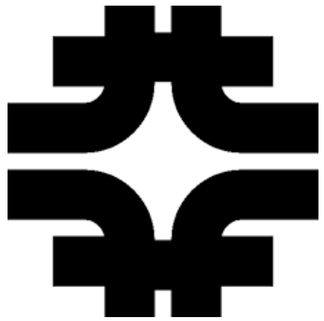
SONIC

(+bonus AI hardware show&tell)

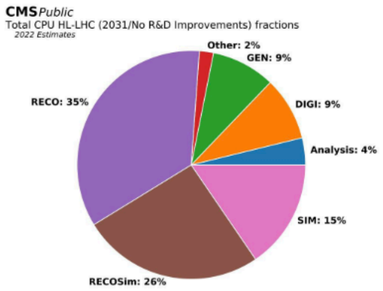
Javier Campos, Yongbin Feng, Nhan Tran
Fermilab
May 22, 2024

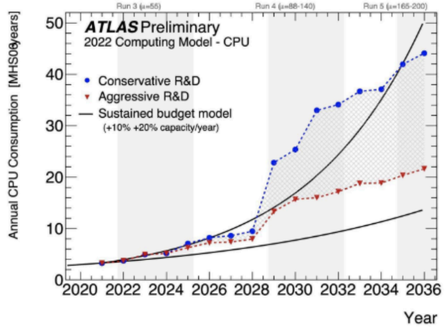


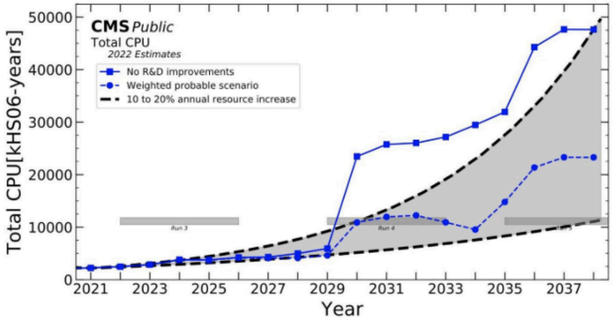
What you've heard already

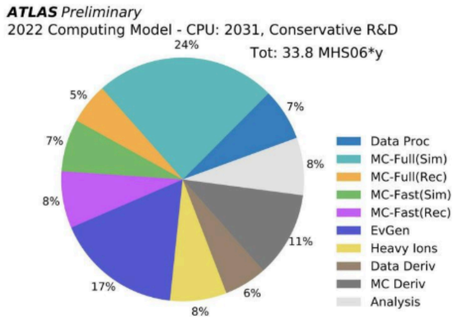
Computing Challenge in High Energy Physics

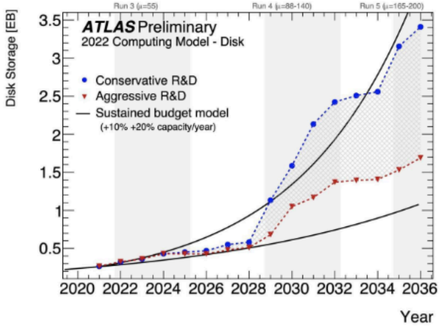
- Large-scale Monte Carlo (MC) Simulations
- Data Analysis from Particle Detectors
- Complex Computational Models (e.g., Lattice QCD)

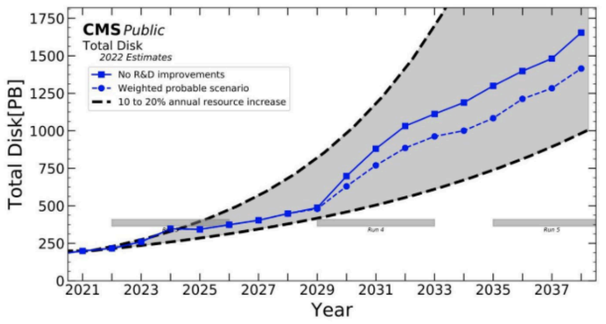






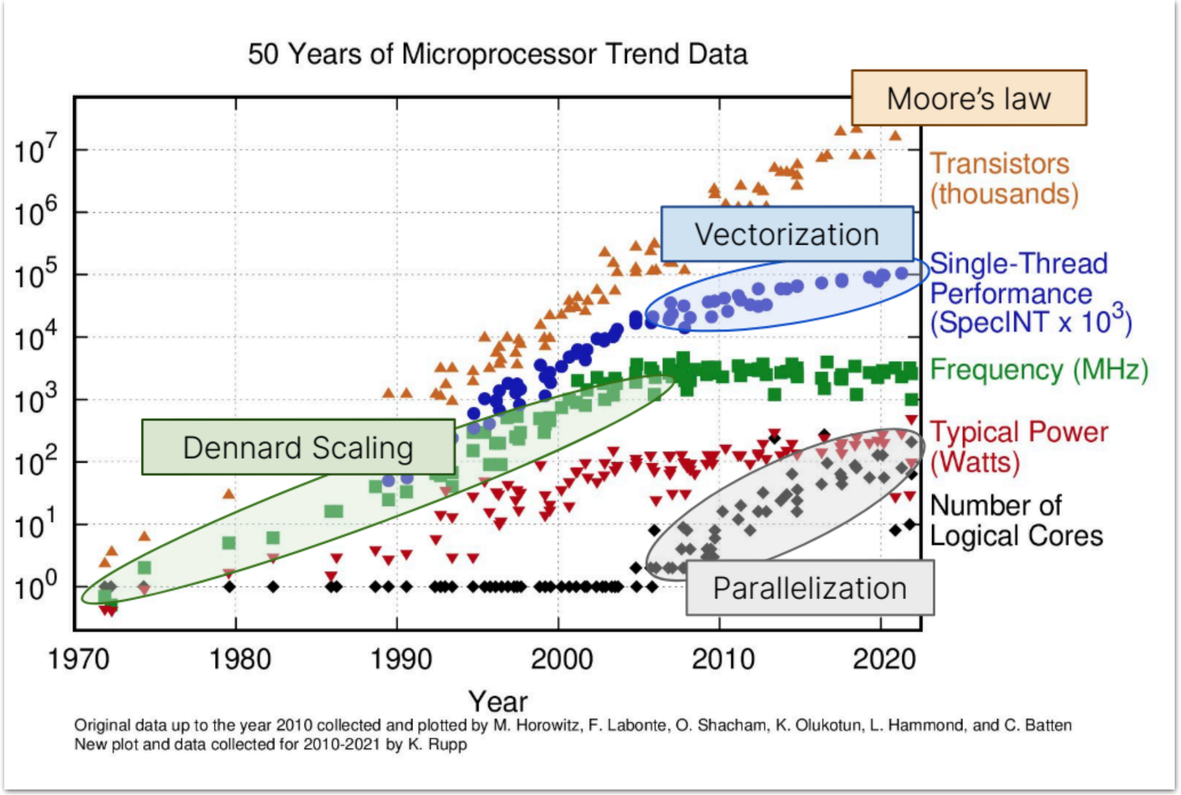






CMS Offline and Computing Public Results & ATLAS Software and Computing HL-LHC Roadmap

What you've heard already



CPU vs GPU: Chip Structure

CPU

- Small number of powerful cores (~10)
  - Branch prediction, out-of-order execution, etc.
- Large caches
- Single instruction on multiple data (**SIMD**)

GPU

- Many number of cores (≳1000)
  - *A lot* simpler
- Small caches
- Single instruction on multiple threads (**SIMT**)

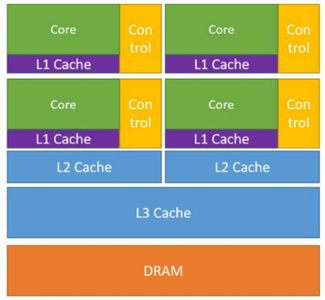
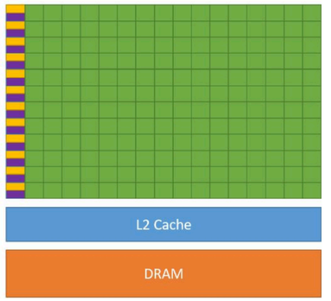
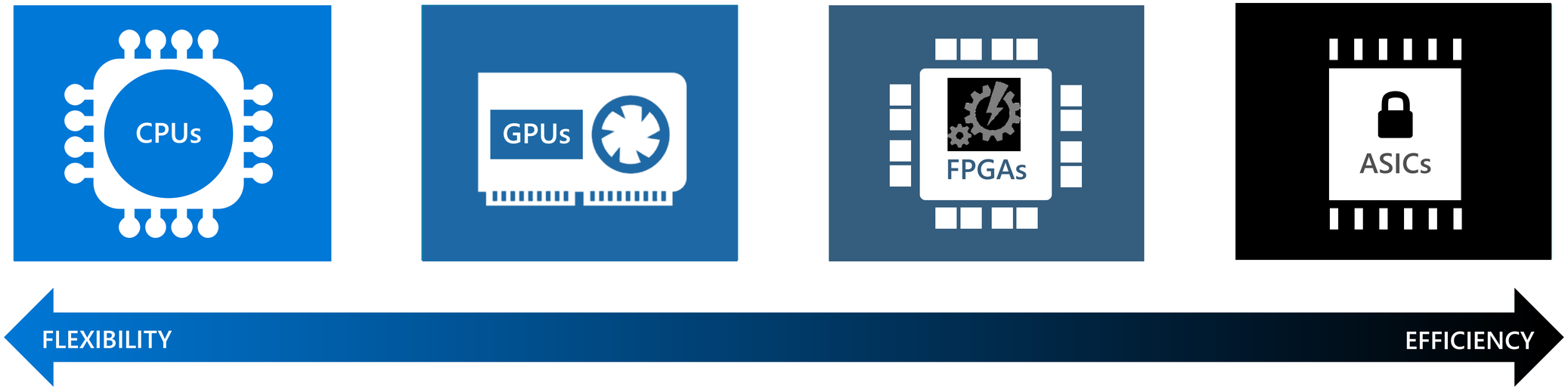


Image credit



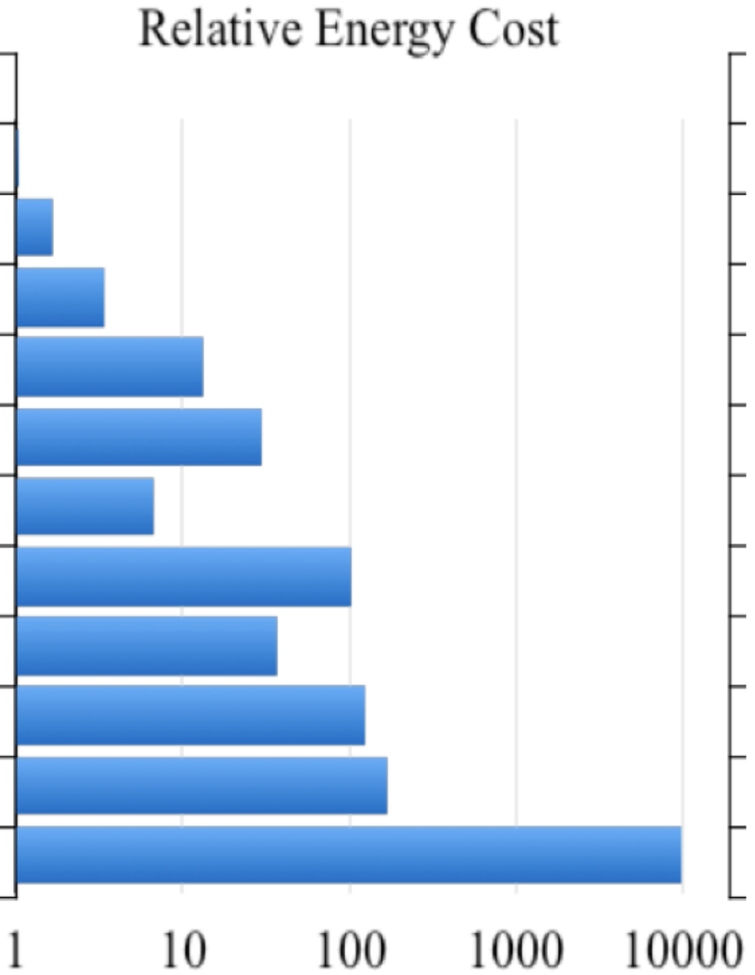
Practically GPU can outperform CPU with parallelizable and relatively simple algorithms

Computing technology



Move data expensive, compute cheap

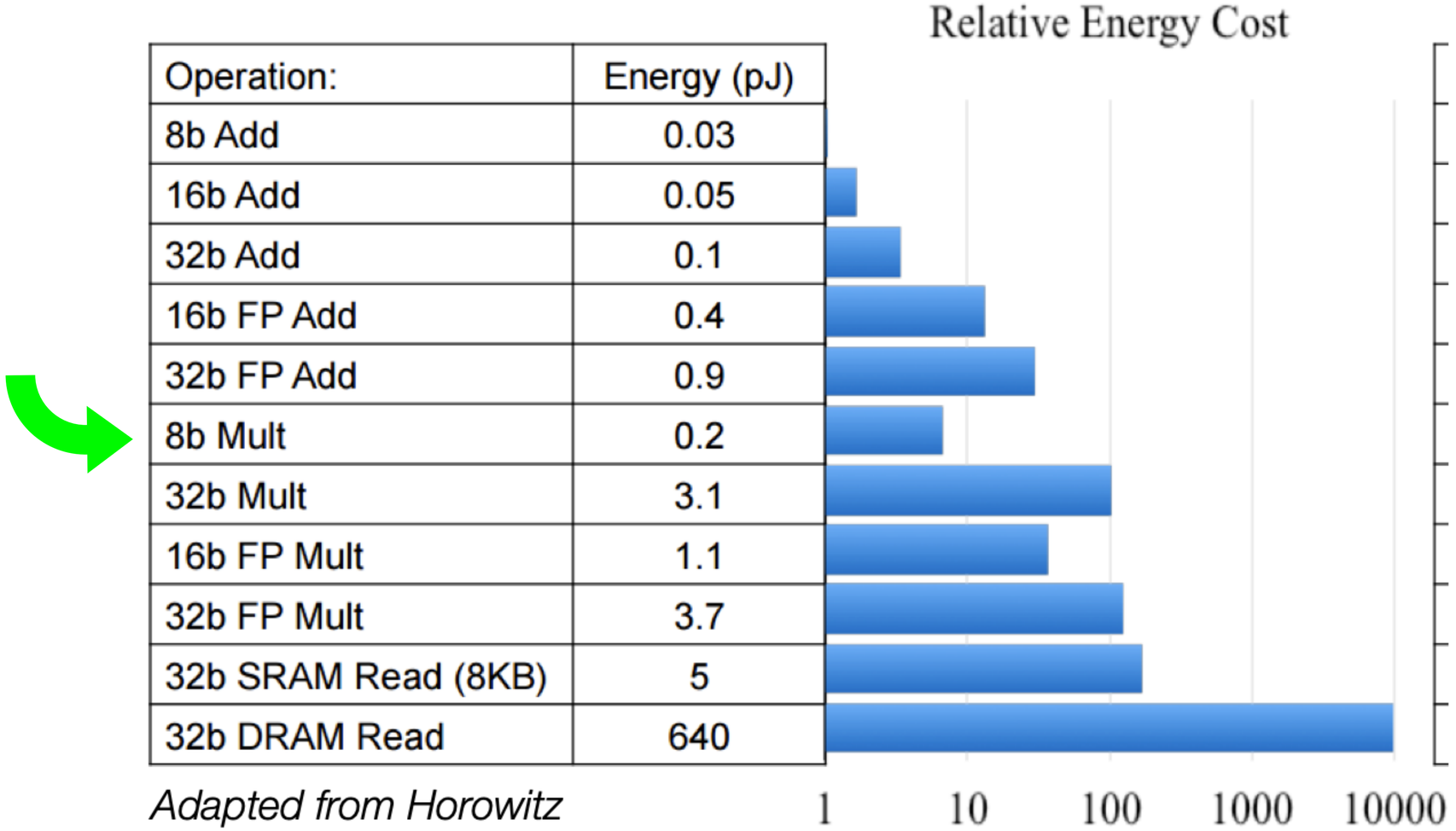
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



Adapted from Horowitz

Move data expensive, compute cheap

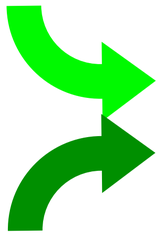
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

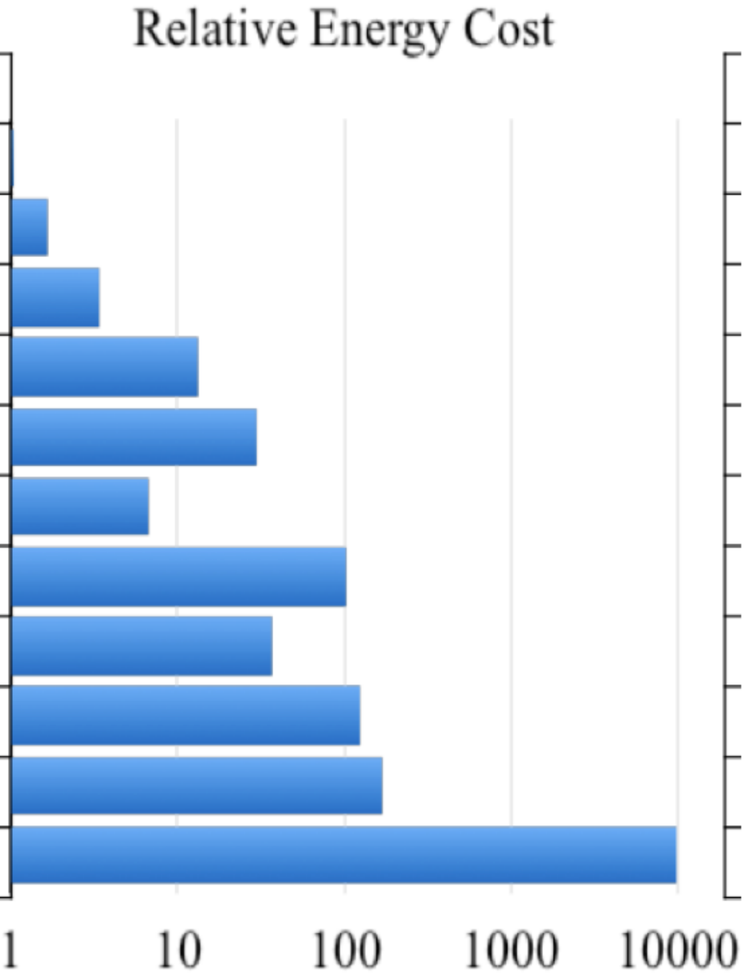


Adapted from Horowitz

Move data expensive, compute cheap

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

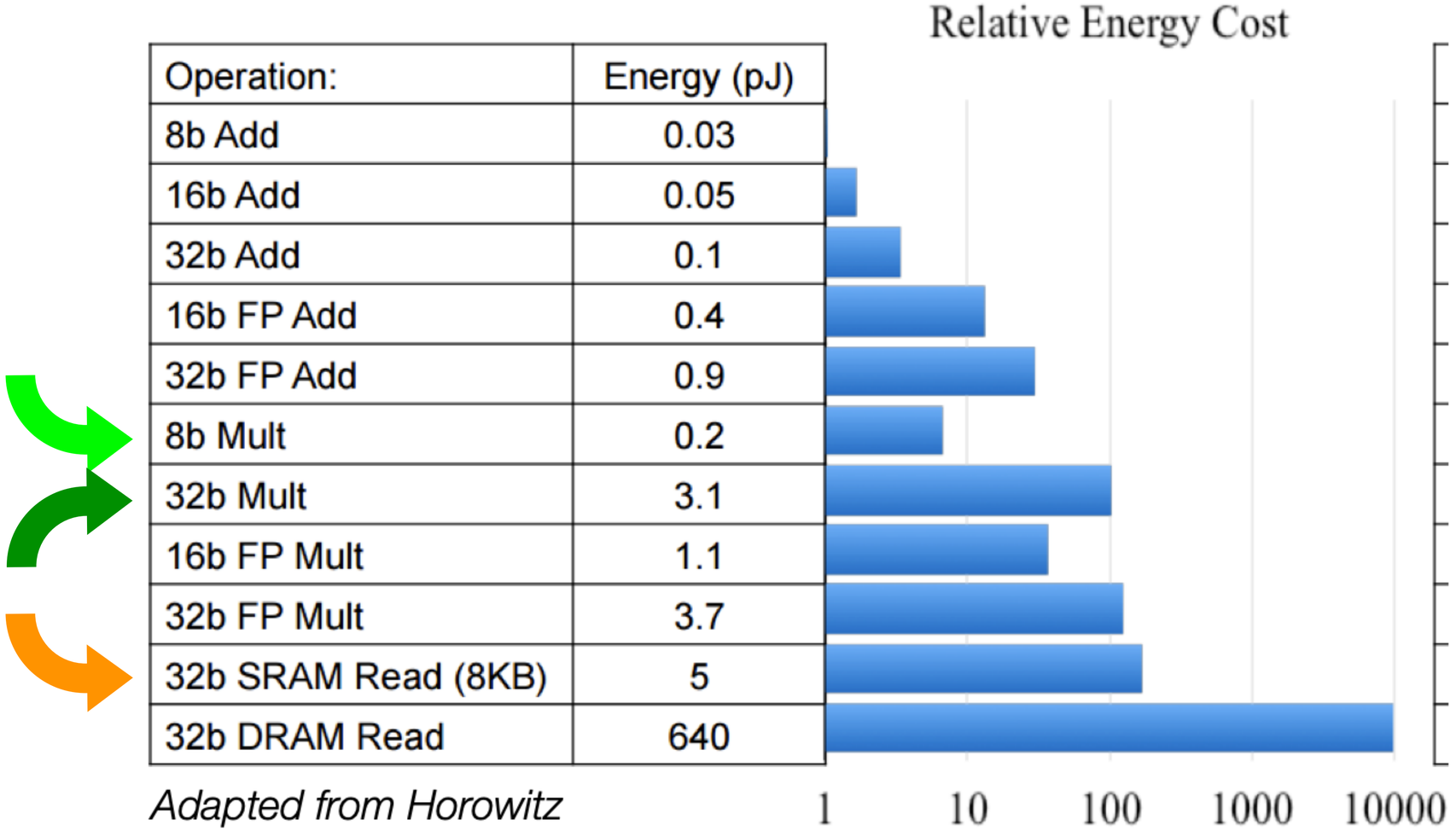




Adapted from Horowitz

Move data expensive, compute cheap

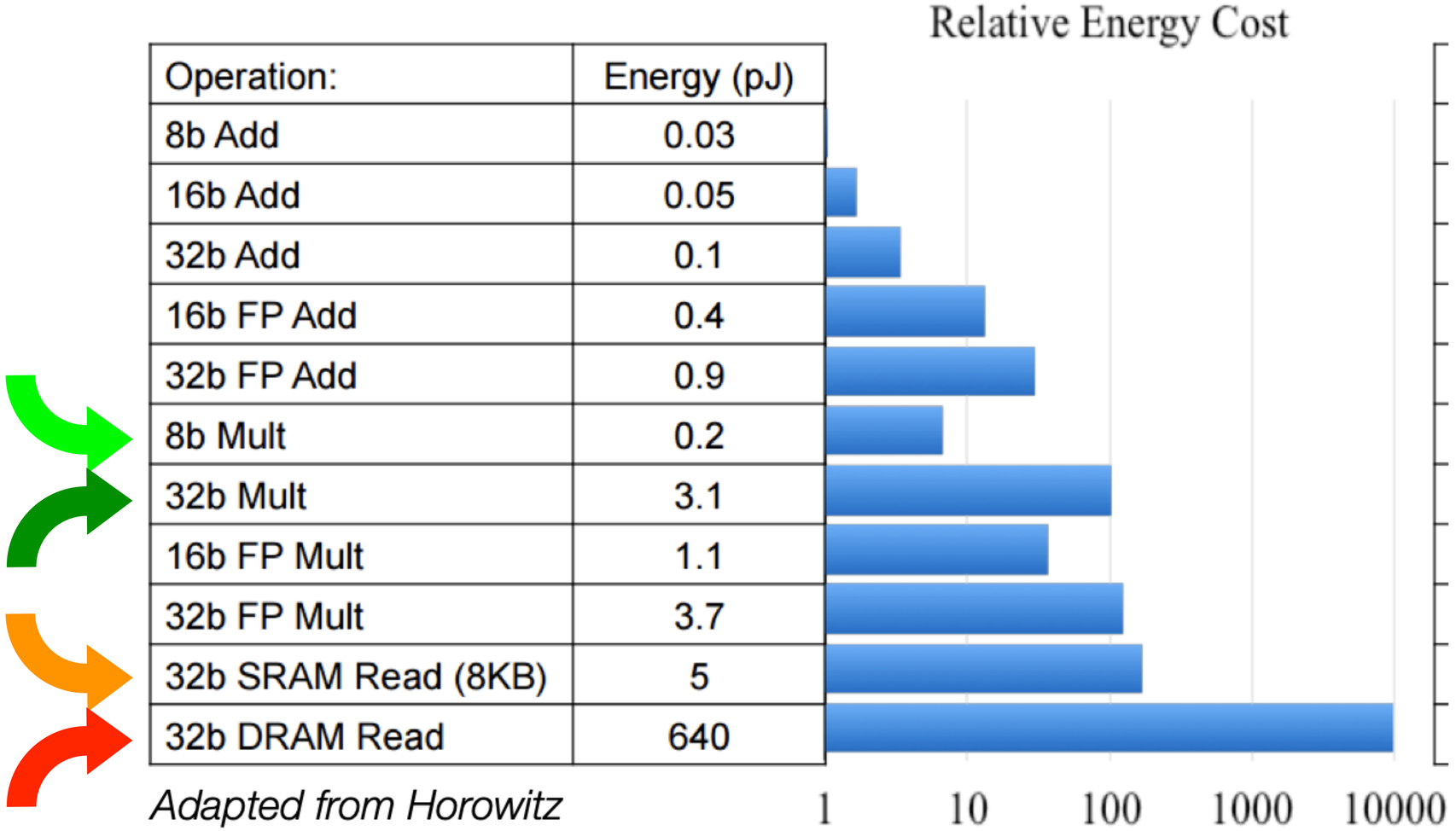
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



Adapted from Horowitz

Move data expensive, compute cheap

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



Adapted from Horowitz

Accelerated compute

Embedded Systems

Embedded in our experiments; often (hard) real-time latency constraints, custom architectures

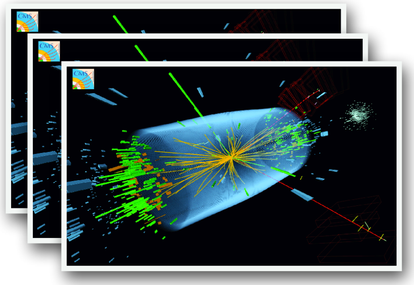
Coprocessors

Traditional datacenter-scale compute; throughput-driven; general purpose architectures

Embedded hardware demo

Embedded Systems

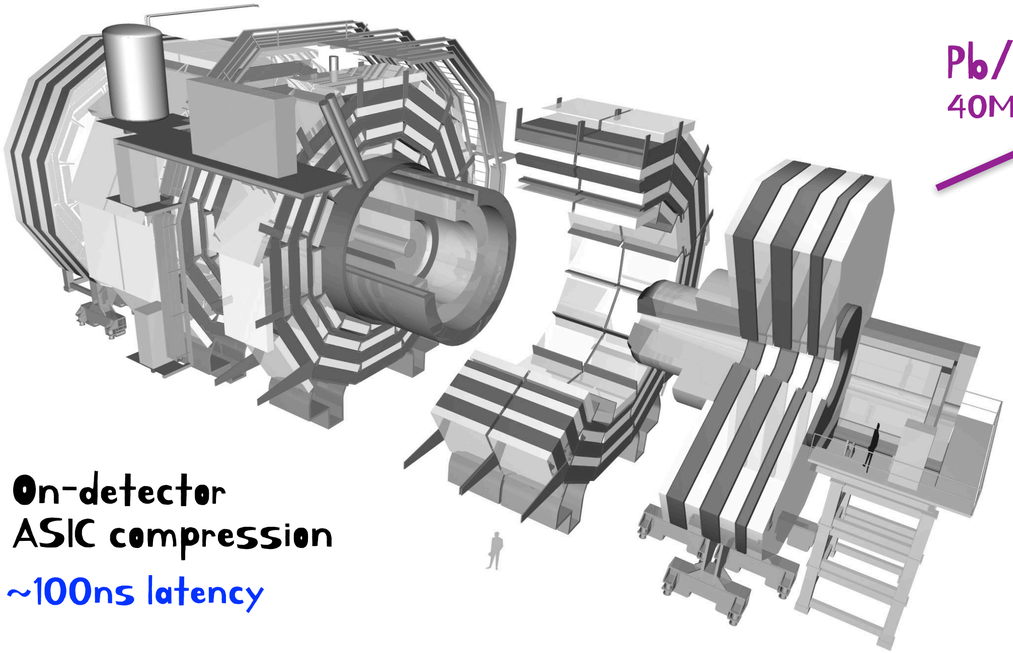
Embedded in our experiments; often (hard) real-time latency constraints, custom architectures



CMS Experiment

40MHz collision rate

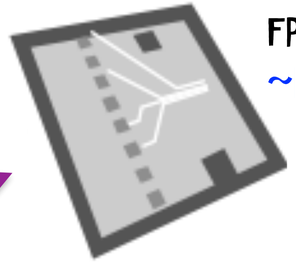
~1B detector channels



On-detector
ASIC compression

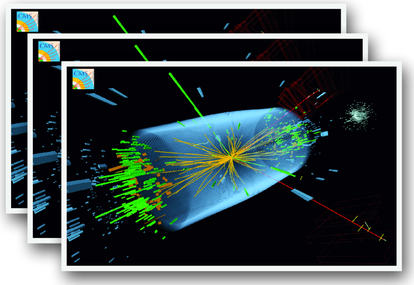
~100ns latency

Pb/s
40MHz



FPGA filter stack

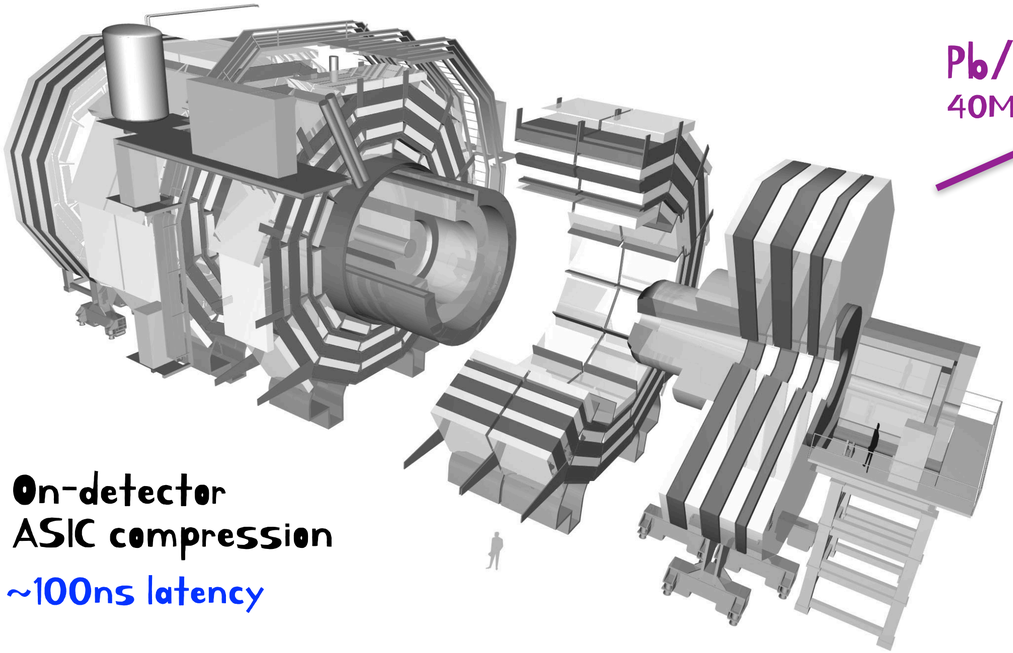
~µs latency



CMS Experiment

40MHz collision rate

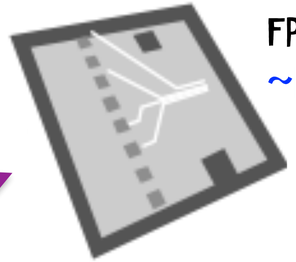
~1B detector channels



On-detector
ASIC compression

~100ns latency

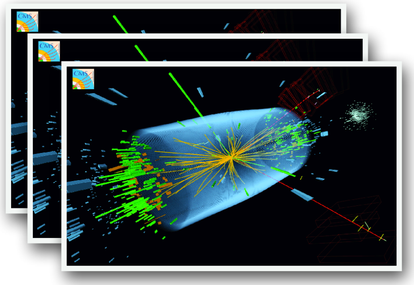
Pb/s
40MHz



FPGA filter stack

~µs latency

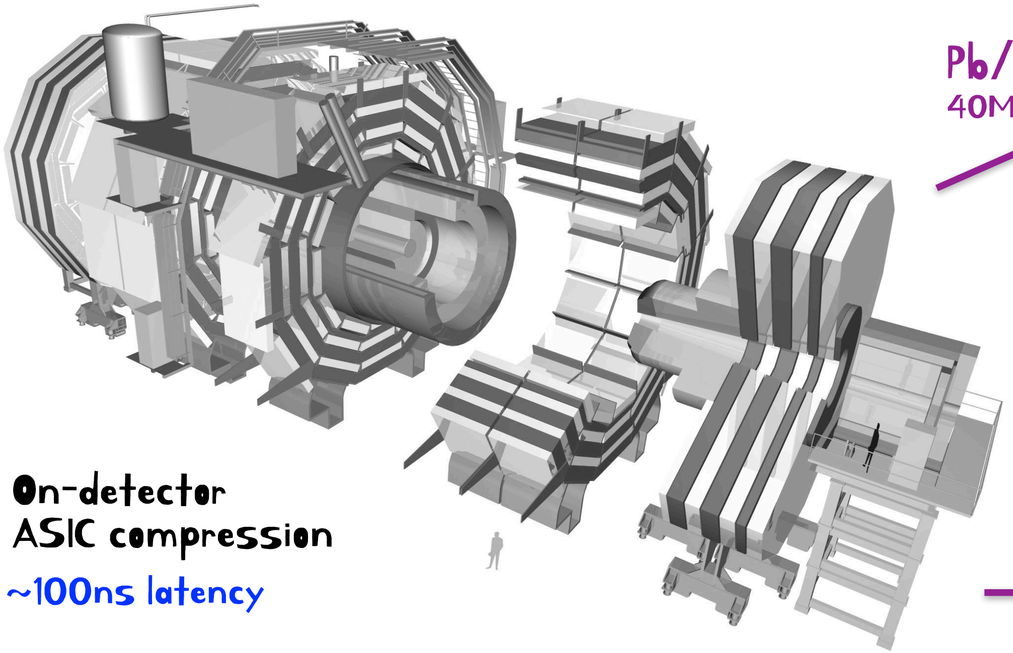
10x AVERAGE INTERNET TRAFFIC IN NORTH AMERICA (2021)



CMS Experiment

40MHz collision rate

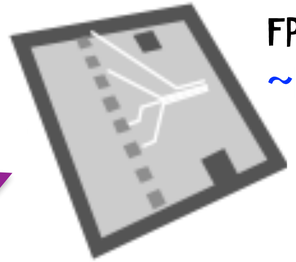
~1B detector channels



On-detector
ASIC compression

~100ns latency

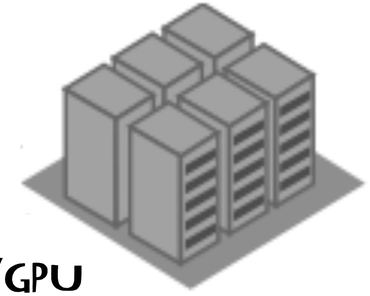
Pb/s
40MHz



FPGA filter stack

~µs latency

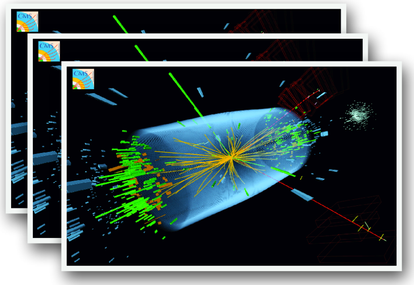
10s Tb/s
100s kHz



On-prem CPU/GPU
filter farm

~100 ms latency

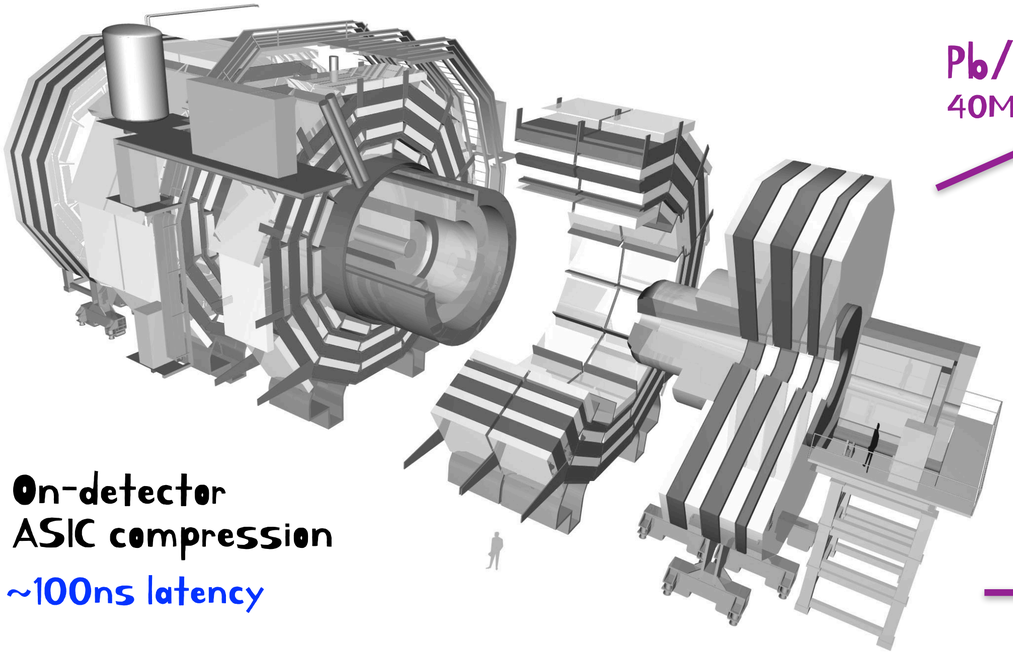
10x AVERAGE INTERNET TRAFFIC IN NORTH AMERICA (2021)



CMS Experiment

40MHz collision rate

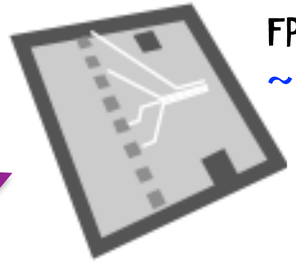
~1B detector channels



On-detector
ASIC compression

~100ns latency

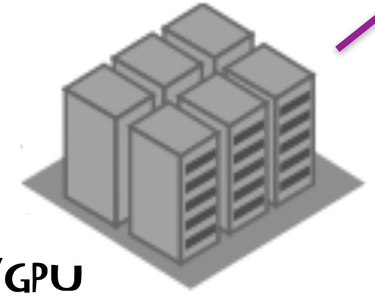
Pb/s
40MHz



FPGA filter stack

~µs latency

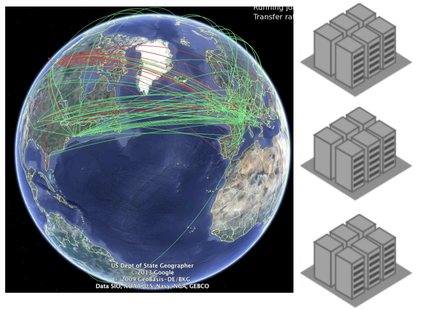
10s Tb/s
100s kHz



On-prem CPU/GPU
filter farm

~100 ms latency

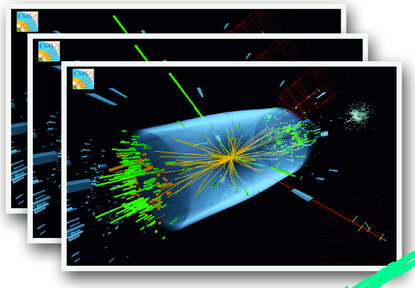
10s Gb/s
~5 kHz



Worldwide
computing grid

Exabyte-scale
datasets

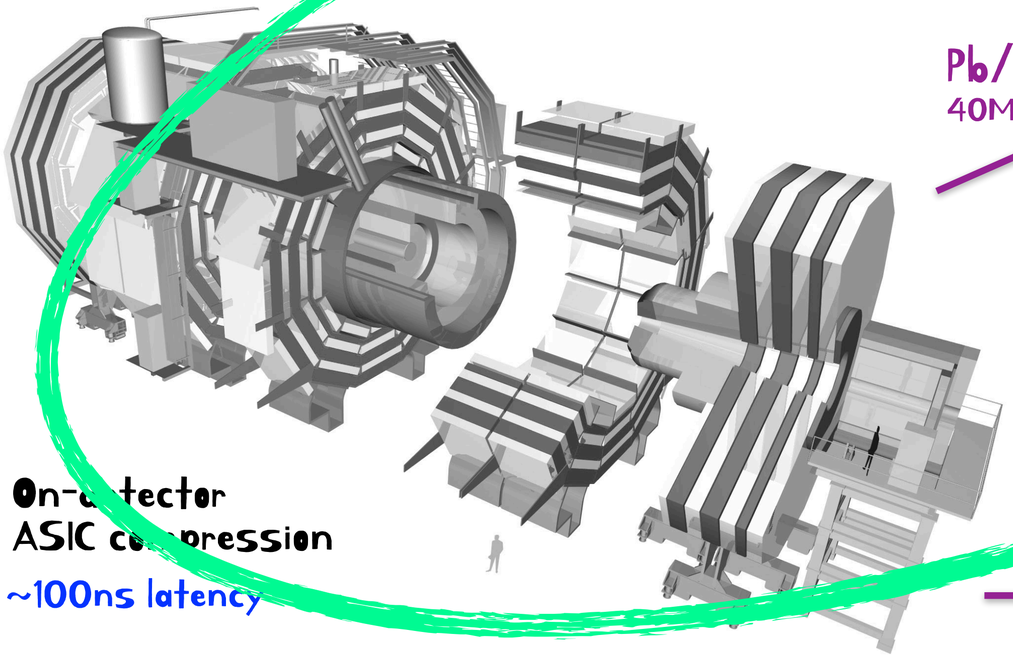
10x AVERAGE INTERNET TRAFFIC IN NORTH AMERICA (2021)



CMS Experiment

40MHz collision rate

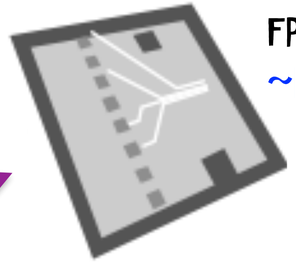
~1B detector channels



On-detector
ASIC compression

~100ns latency

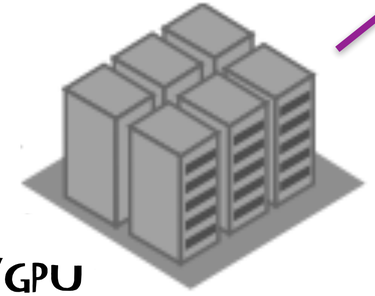
Pb/s
40MHz



FPGA filter stack

~µs latency

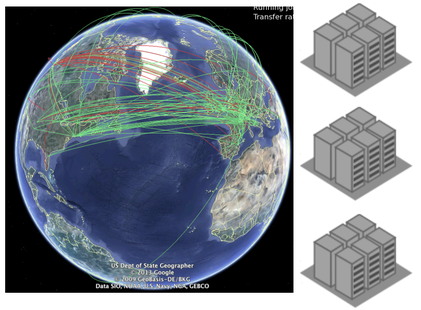
10s Tb/s
100s kHz



On-prem CPU/GPU
filter farm

~100 ms latency

10s Gb/s
~5 kHz

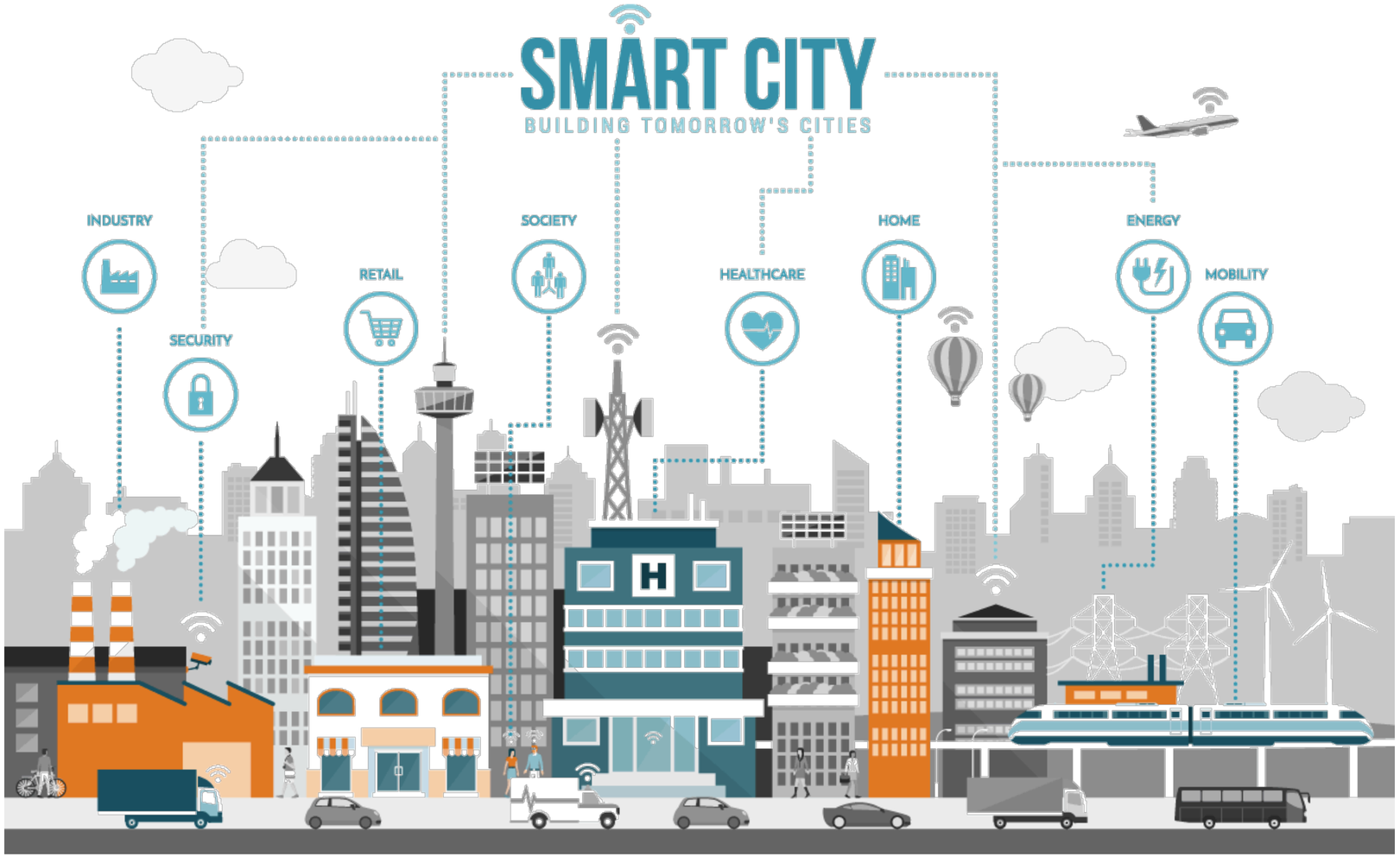


Worldwide
computing grid

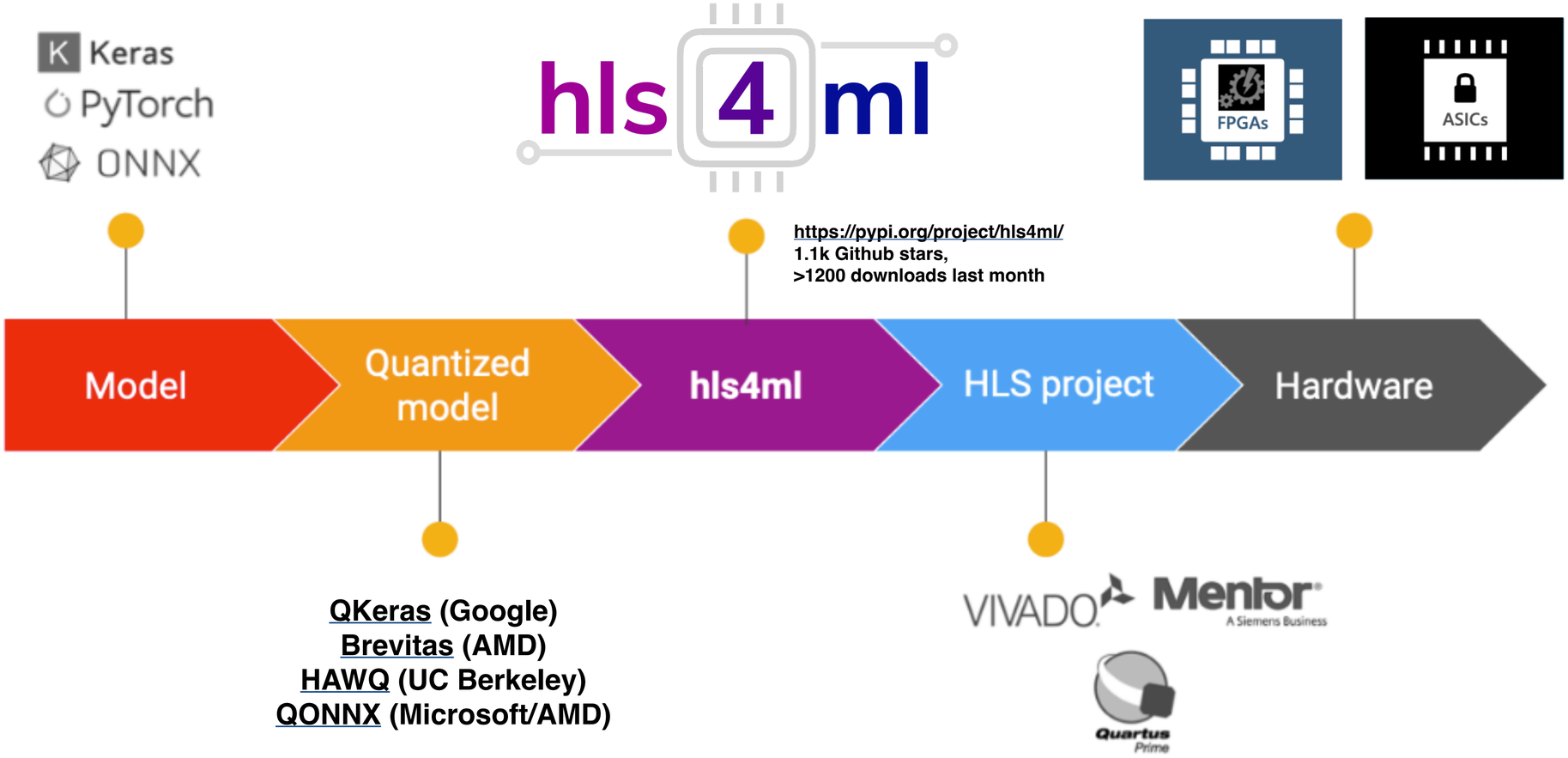
Exabyte-scale
datasets

10x AVERAGE INTERNET TRAFFIC IN NORTH AMERICA (2021)

IoT


SMART CITY
BUILDING TOMORROW'S CITIES
INDUSTRY
SECURITY
RETAIL
SOCIETY
HEALTHCARE
HOME
ENERGY
MOBILITY
H

Efficient codesign tools for



adapted from Vladimir Loncar

MLCommons launches machine learning benchmark for devices like smartwatches and voice assistants

by Ben Wodecki 6/16/2021

With experts from Qualcomm, Fermilab, and Google aiding in its development

MLCommons, the open engineering consortium behind the MLPerf benchmark test, has launched a new measurement suite aimed at 'tiny' devices like smartwatches and voice assistants.

MLPerf Tiny Inference is designed to compare performance of embedded devices and models with a footprint of 100kB or less, by measuring

Fermilab, UCSD, Columbia, teamed up with AMD/Xilinx for IoT submissions for MLCommons benchmarks

yahoo!finance Search for news, symbols or companies

My Portfolio News Markets Sectors Screeners Personal Finance Videos

Siemens simplifies development of AI accelerators for advanced system-on-chip designs with Catapult AI NN

PR Newswire
Tue, May 21, 2024, 8:00 AM CDT • 5 min read

In This Article:

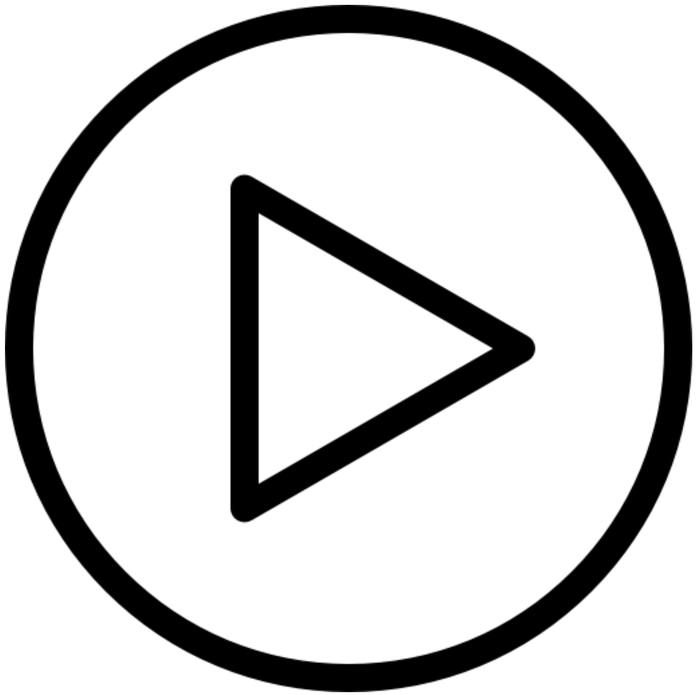
SIEGY -0.78% SMAWF +0.35%

- **Catapult AI NN offers software engineers a comprehensive solution to synthesize AI Neural Nets**
- **Enables software development teams to seamlessly translate AI models designed in Python into silicon-based implementations, facilitating faster and more power-efficient execution compared to standard processors**

PLANO, Texas, May 21, 2024 /PRNewswire/ -- Siemens Digital Industries Software today announced Catapult™ AI NN software for High-Level Synthesis (HLS) of neural network accelerators on Application-Specific Integrated Circuits (ASICs) and System-on-a-chip (SoCs). Catapult AI NN is a complete solution that starts with a neural network description from an AI framework, converts it into C++ and synthesizes it into an RTL accelerator in Verilog or VHDL for implementation in silicon.

Catapult AI NN brings together hls4ml, an open-source package for machine learning hardware acceleration, and Siemens' Catapult™ HLS software for High-Level Synthesis. Developed in close collaboration with Fermilab, a U.S. Department of Energy Laboratory, and other leading contributors to hls4ml, Catapult AI NN addresses the unique requirements of machine learning accelerator design for power, performance, and area on custom silicon.

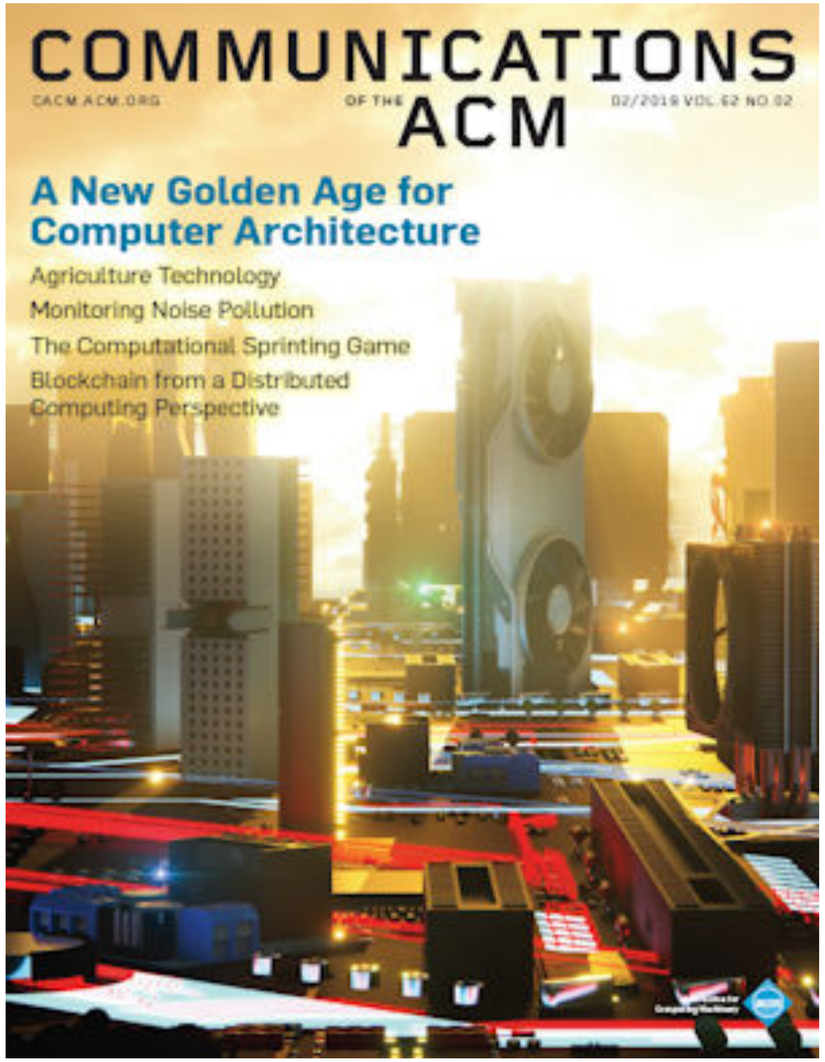
Accelerated compute



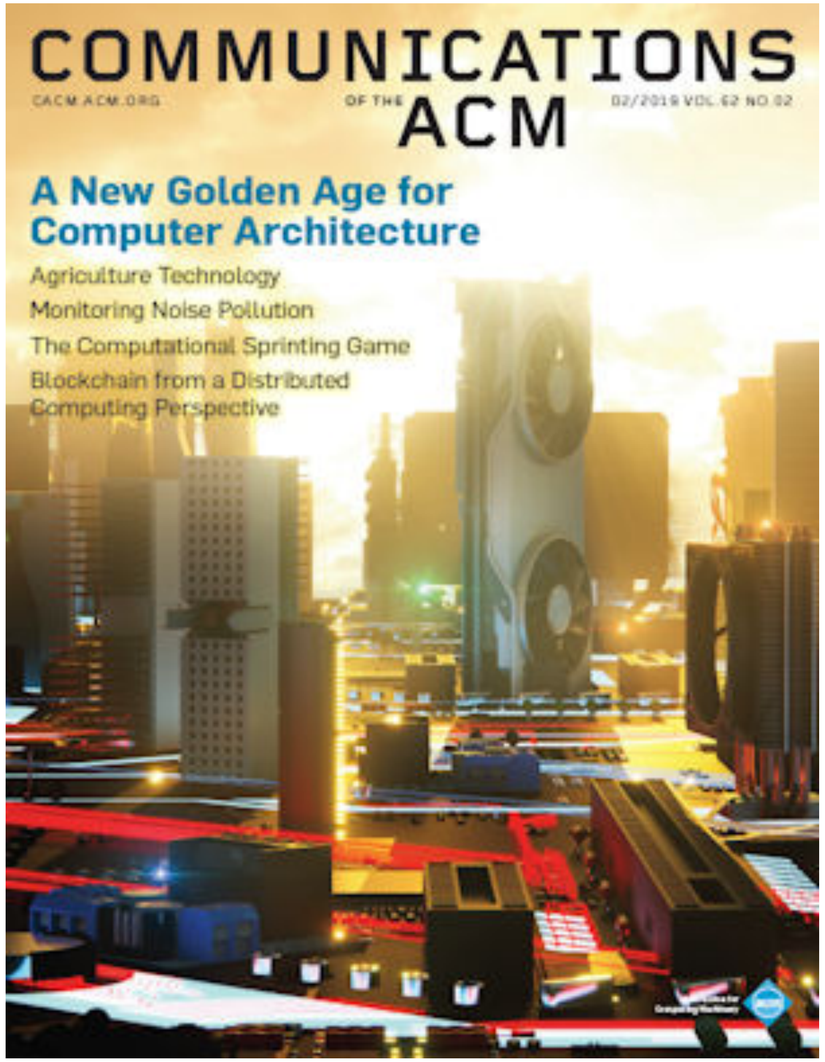
Coprocessors

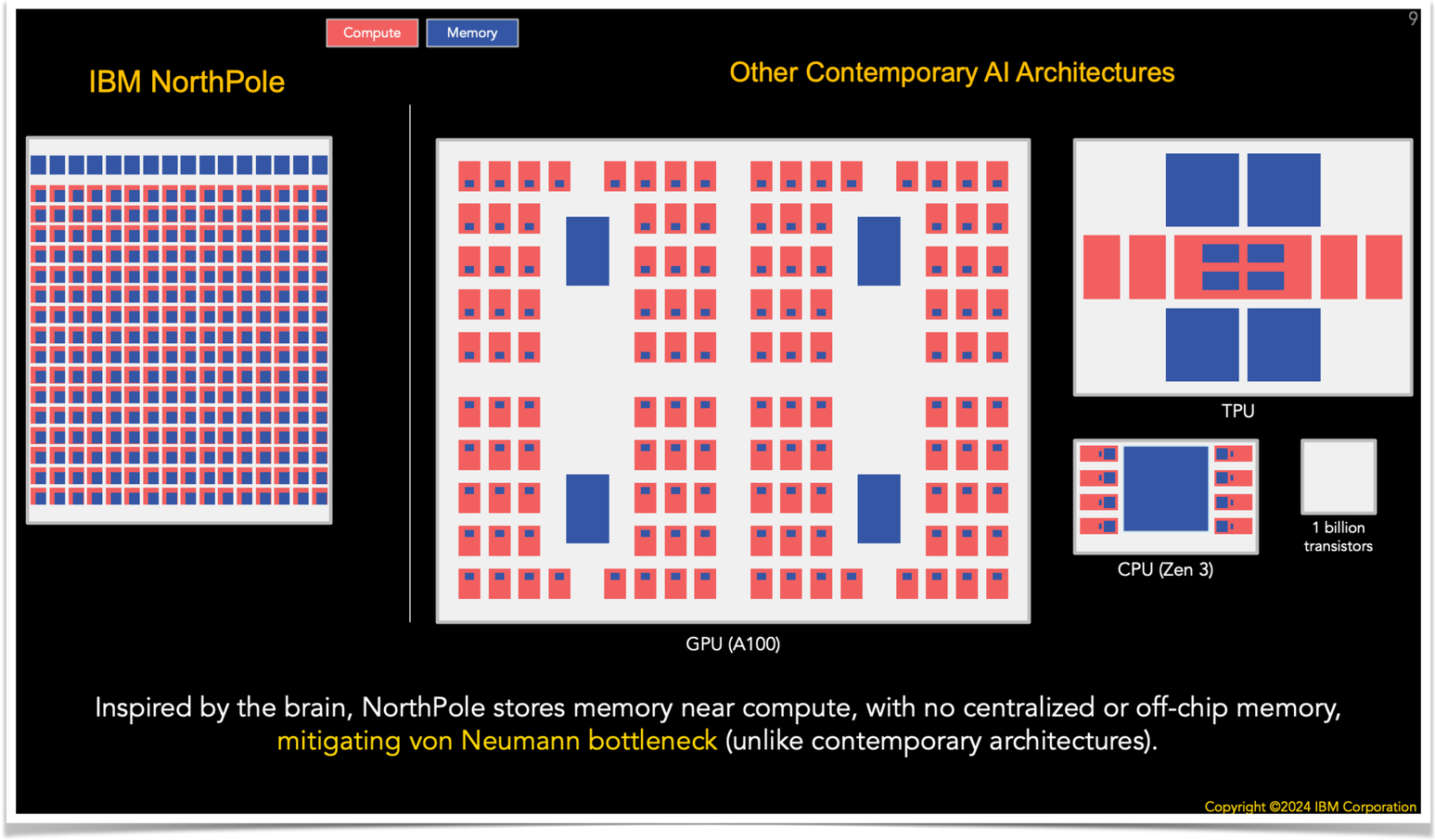
Traditional datacenter-scale compute; throughput-driven; general purpose architectures

Coprocessors

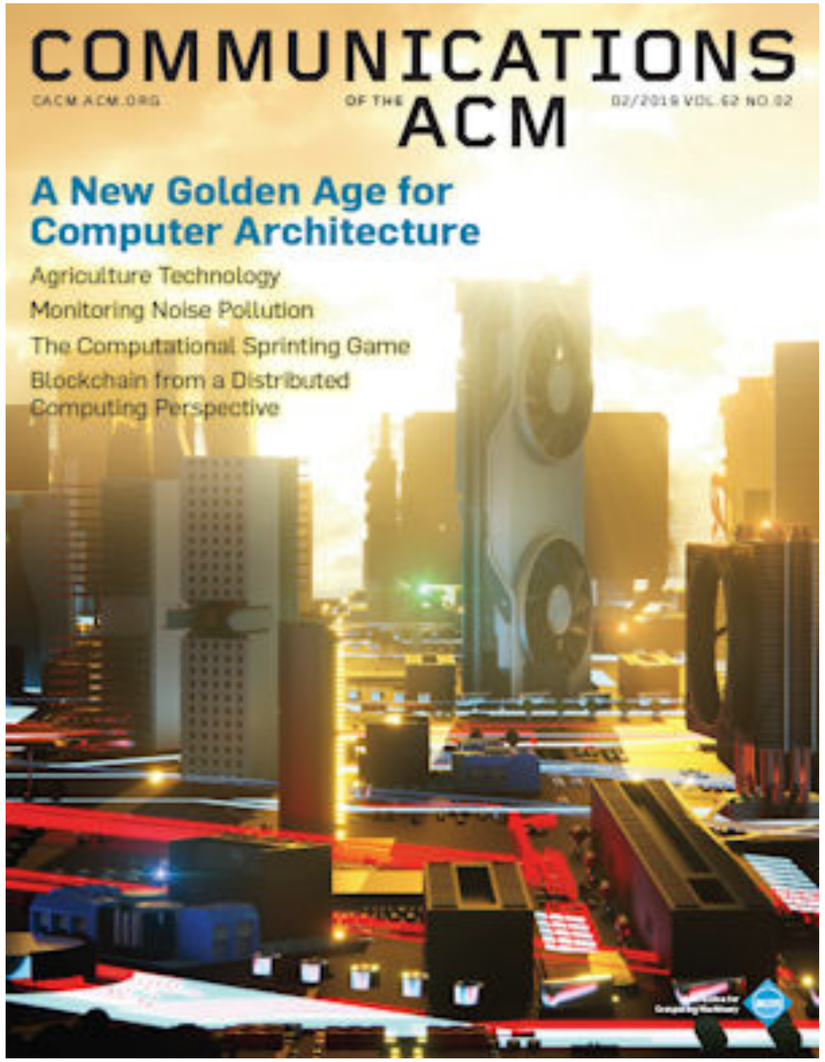


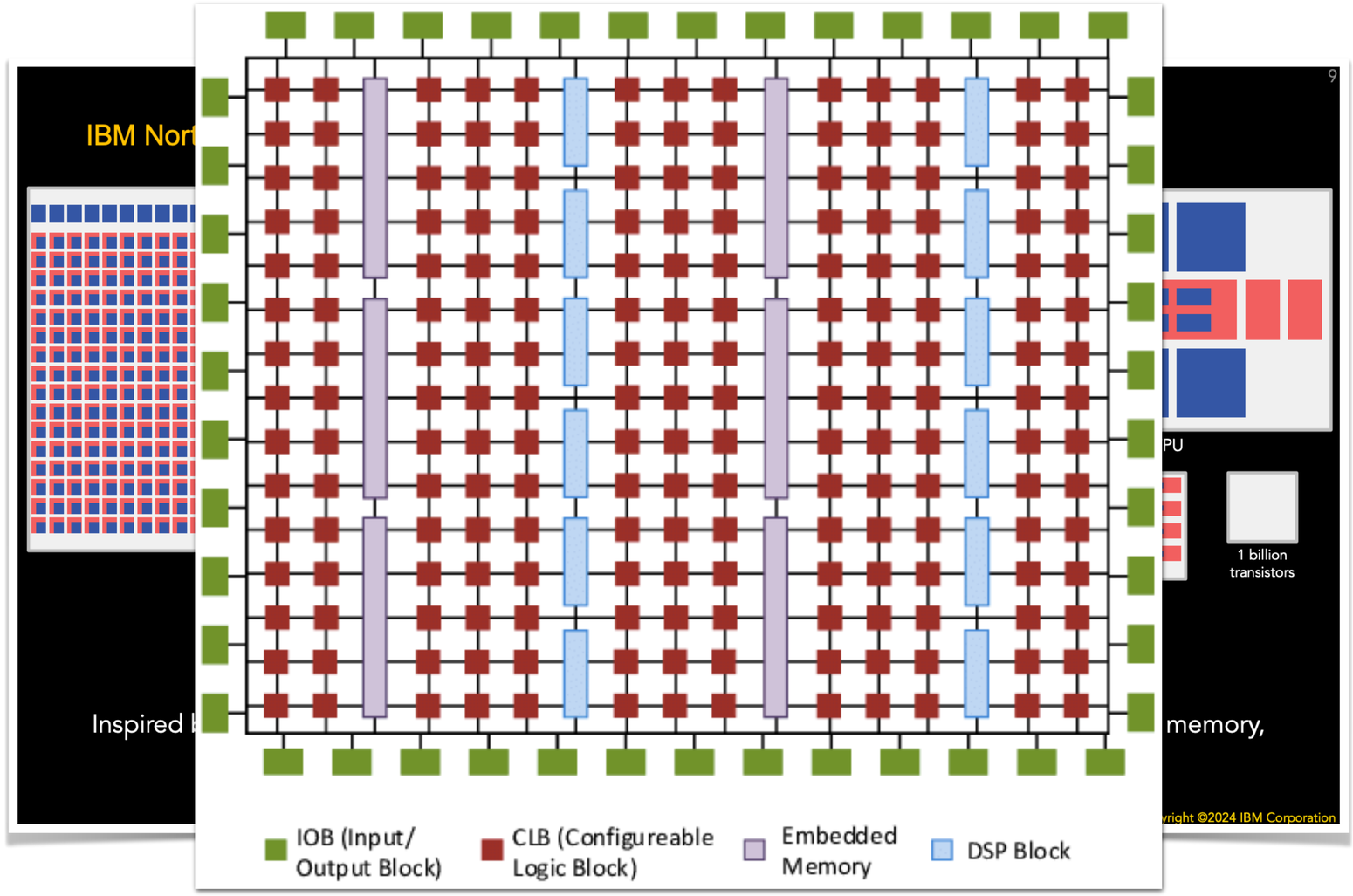
Coprocessors





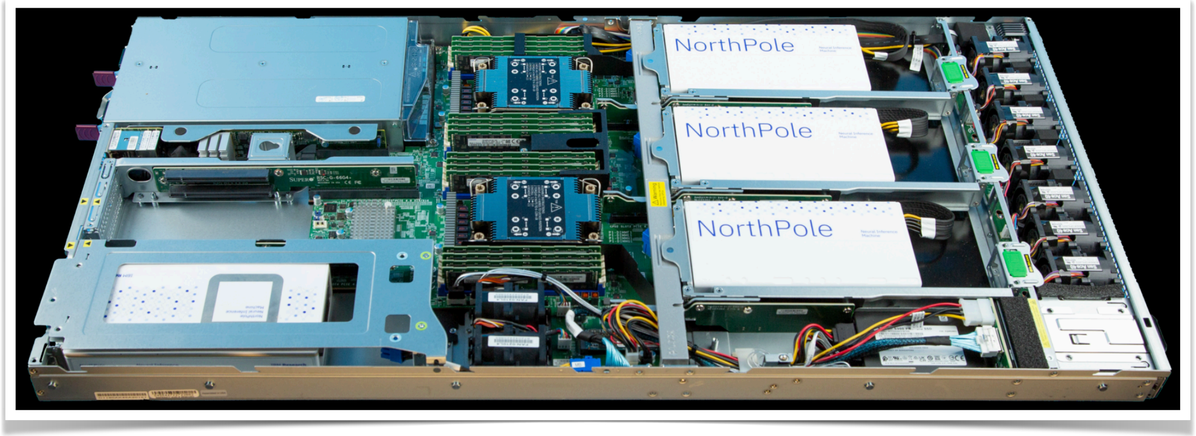
Coprocessors



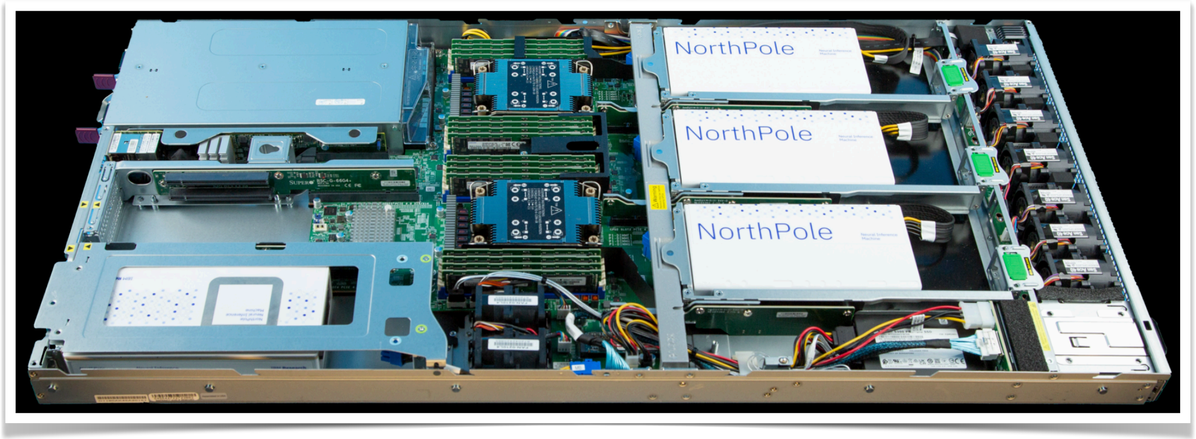


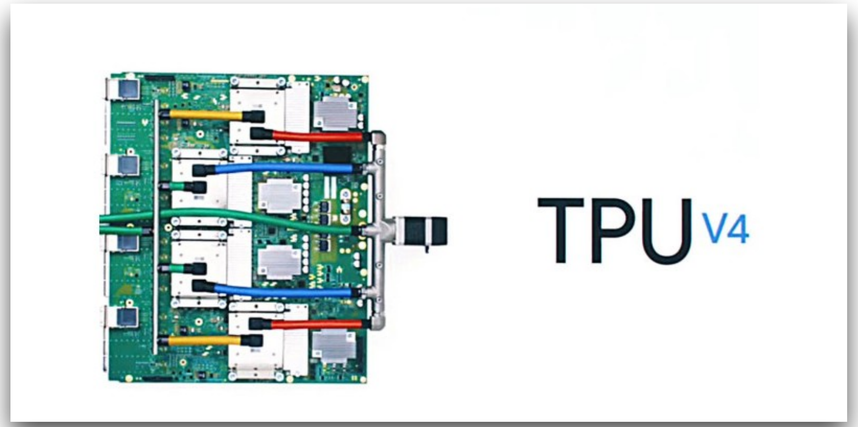
Coprocessors

Coprocessors

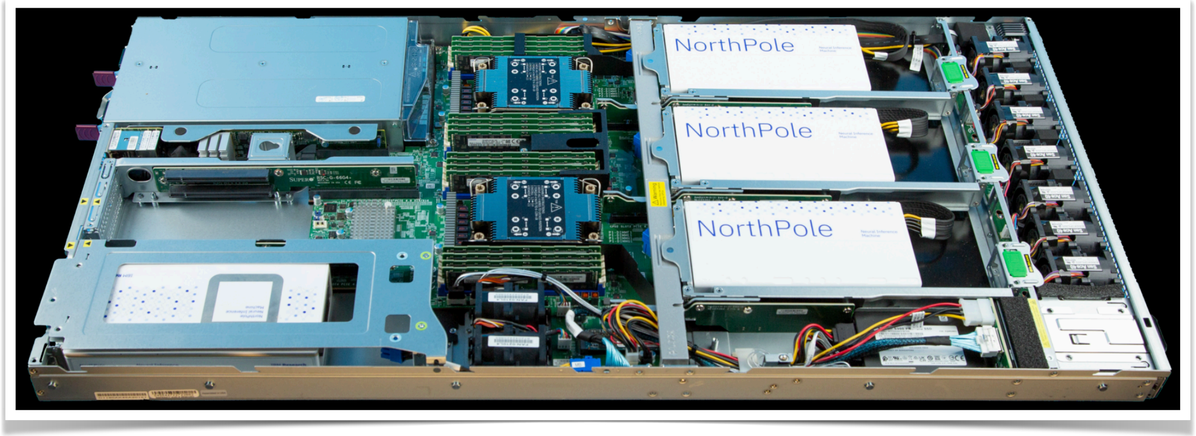


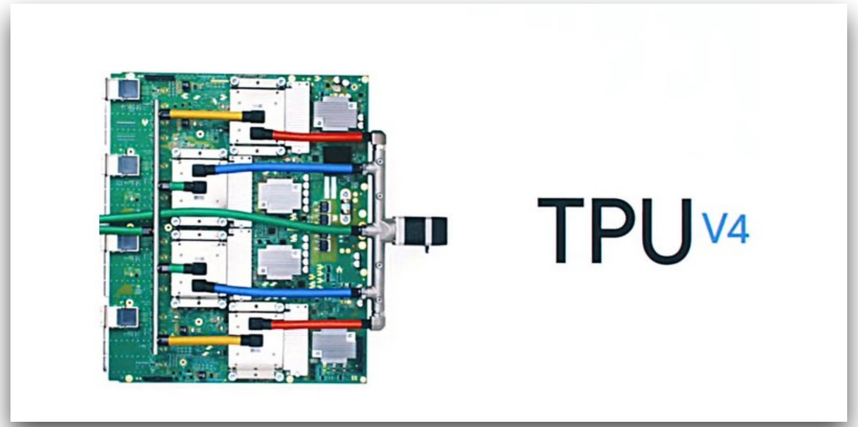
Coprocessors

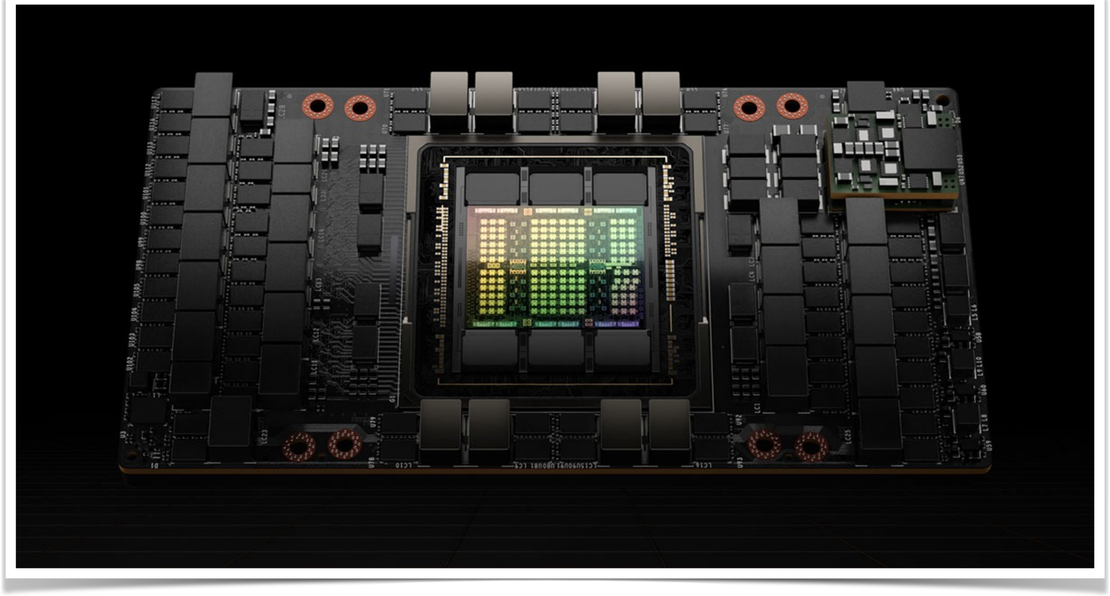




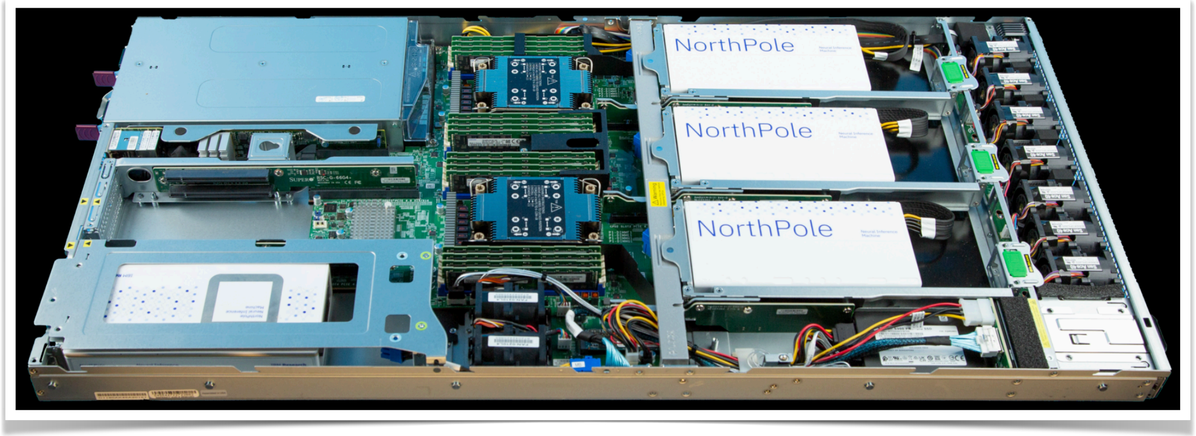
Coprocessors

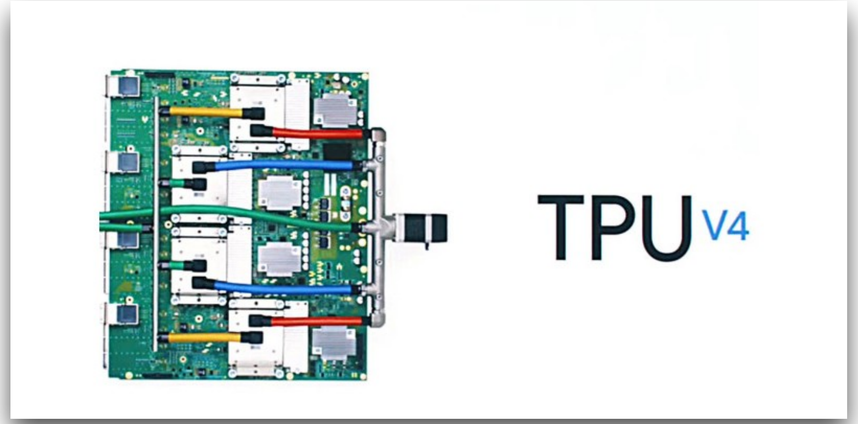


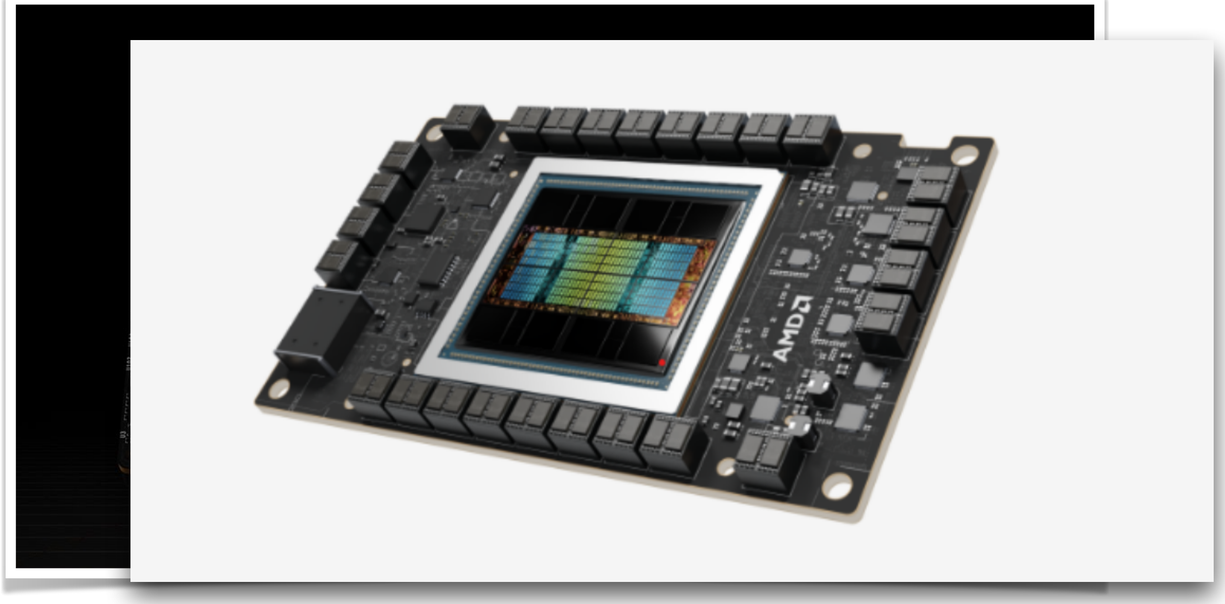




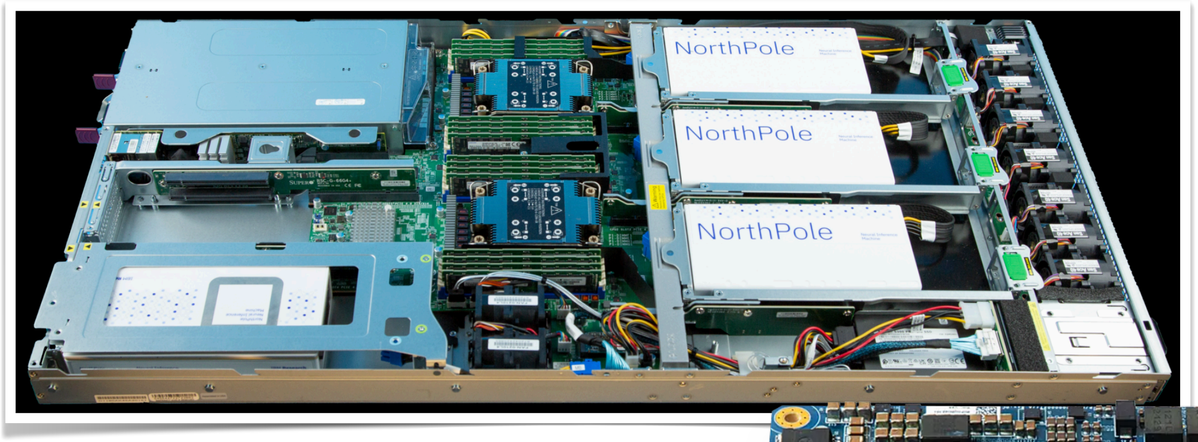
Coprocessors

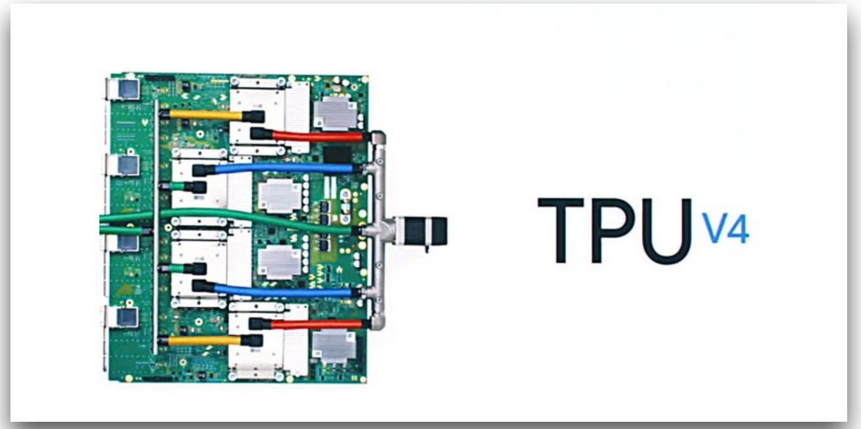


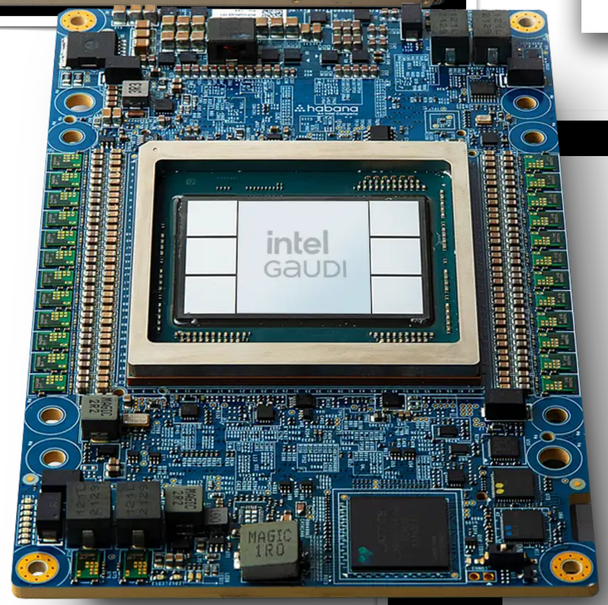


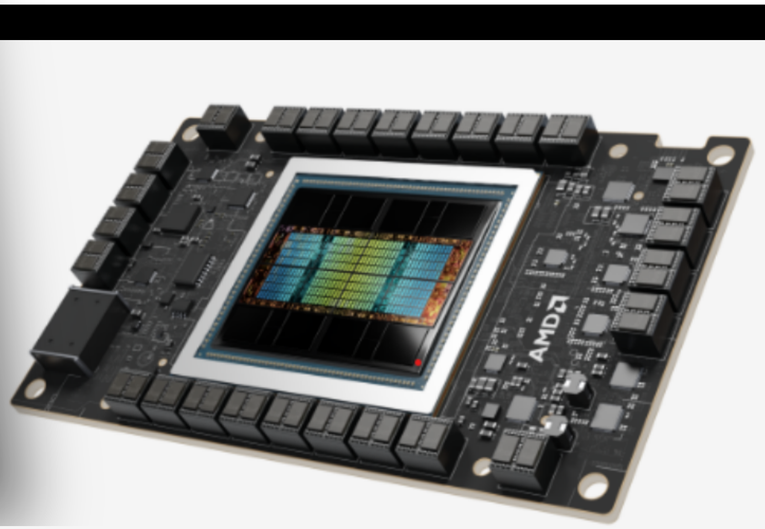


Coprocessors

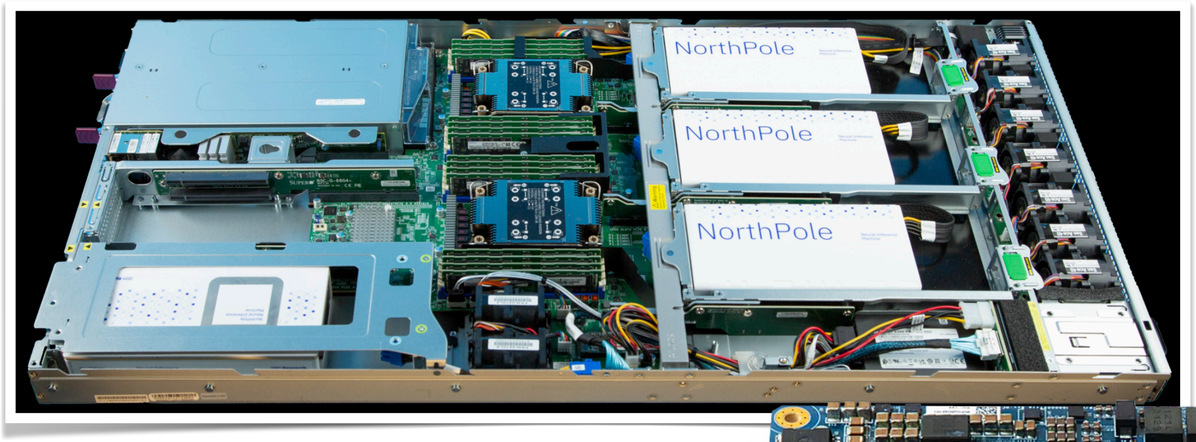


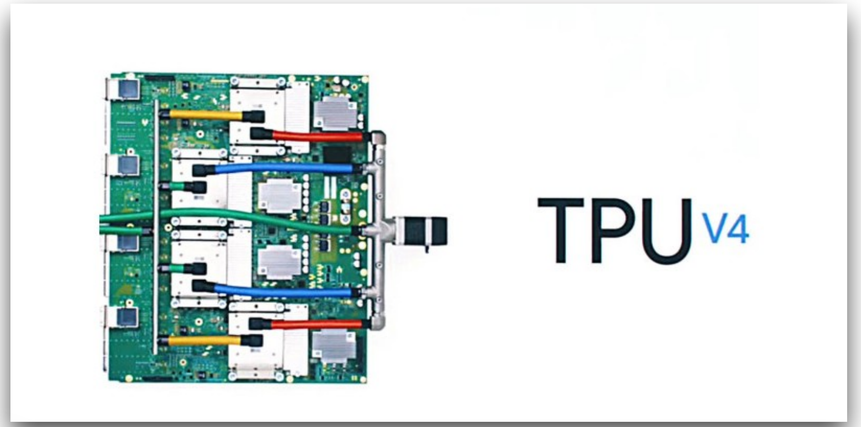


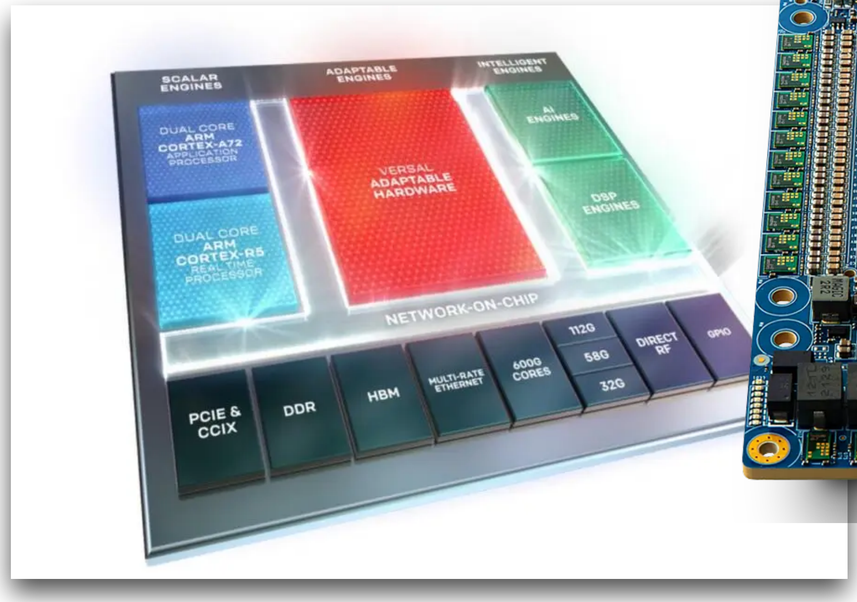


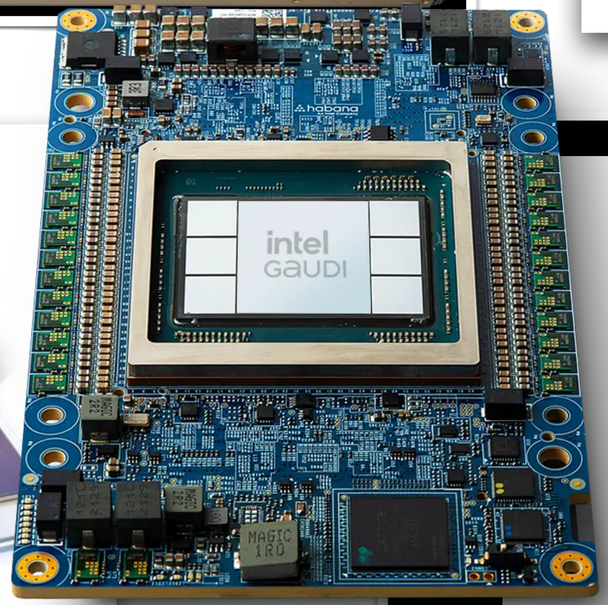


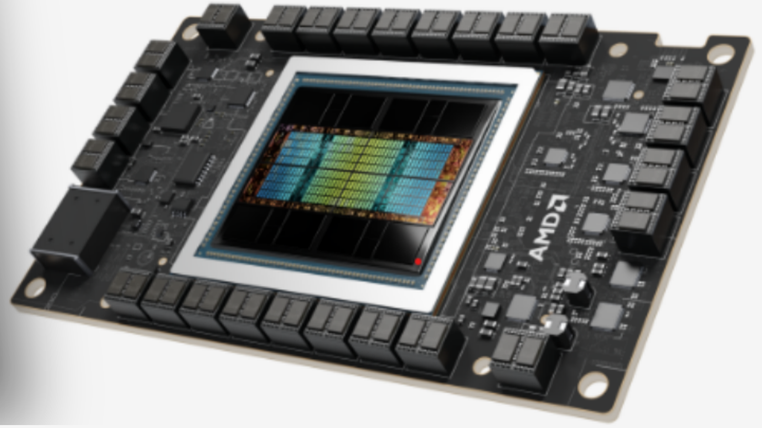
Coprocessors



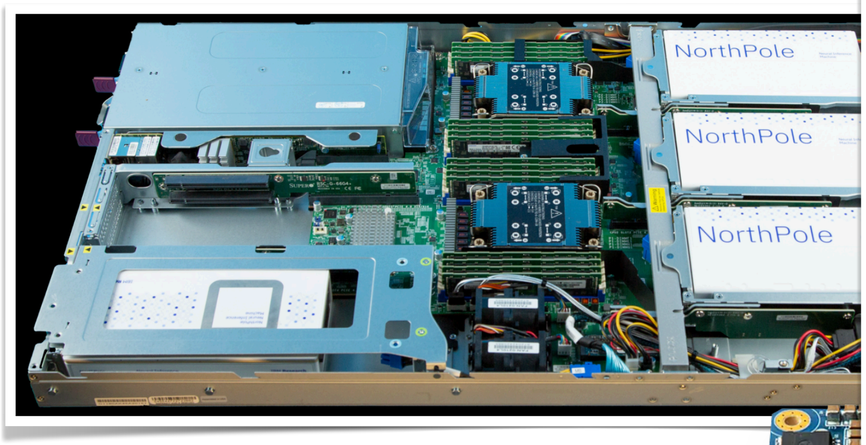


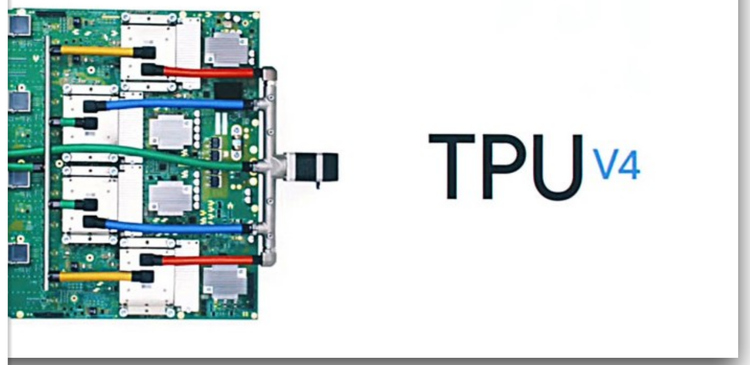


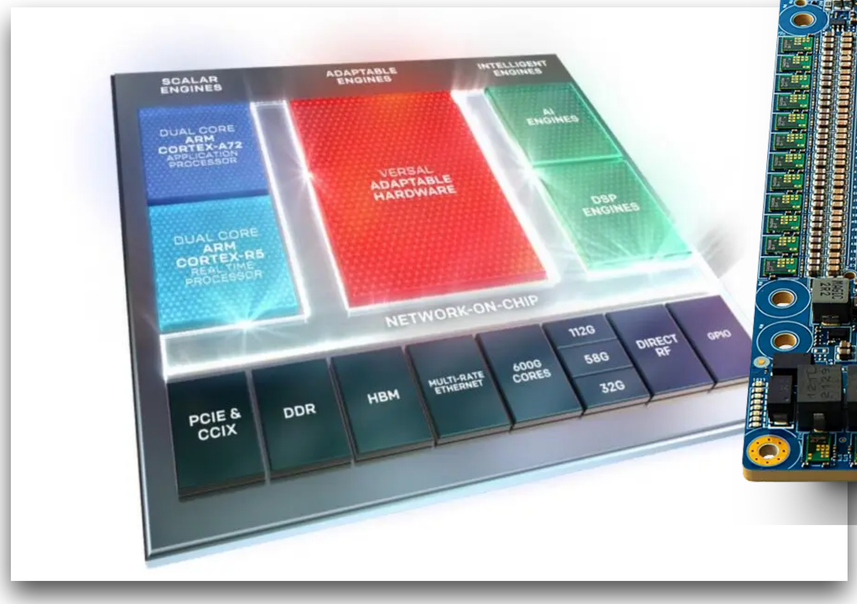


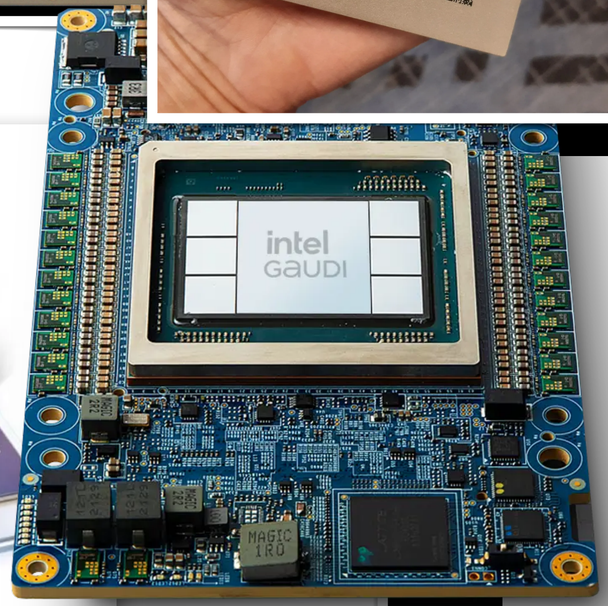


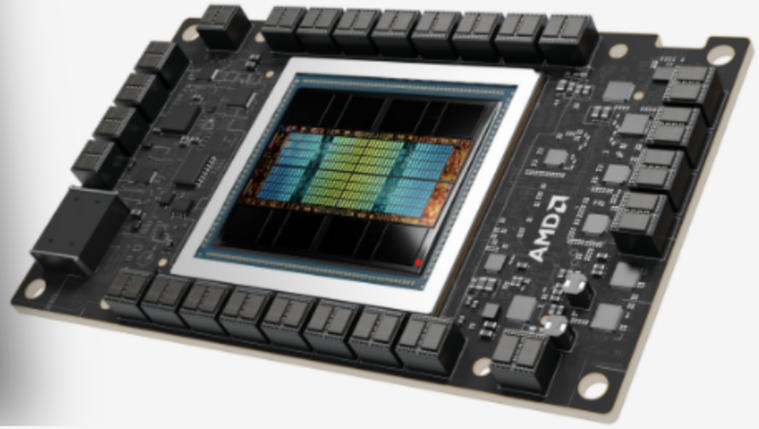
Coprocessors



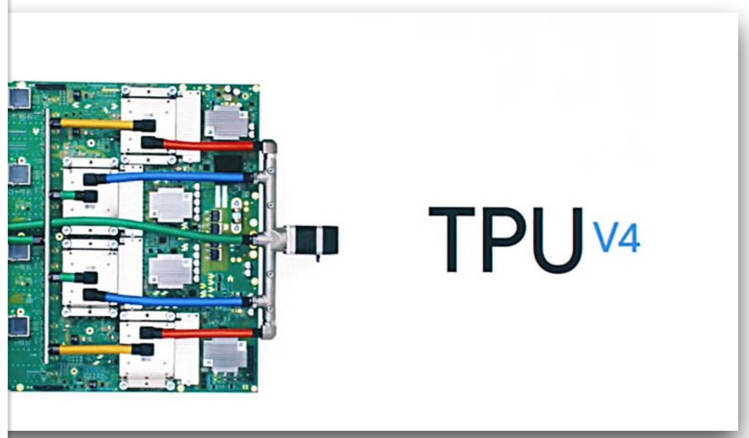


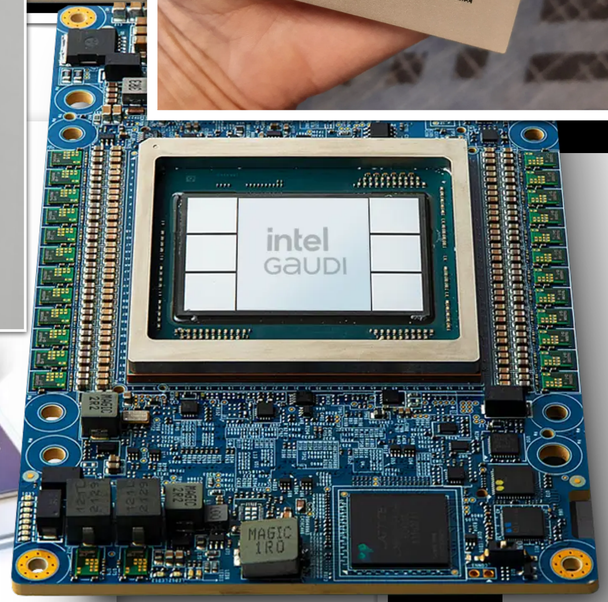


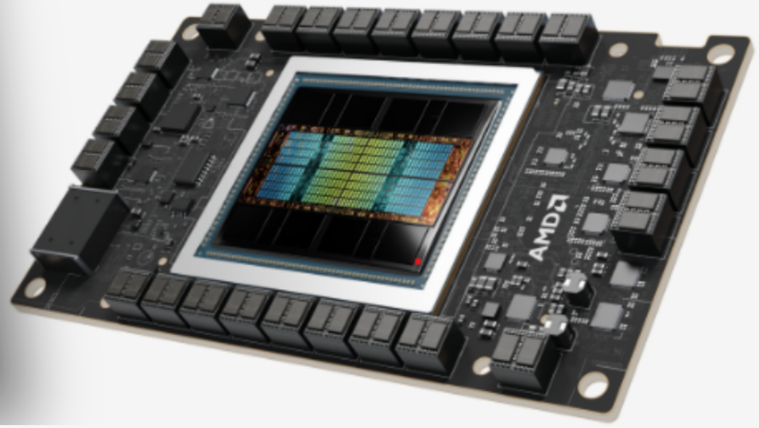




Coprocessors

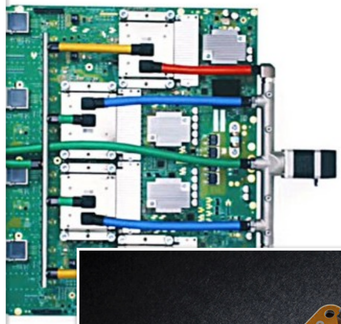


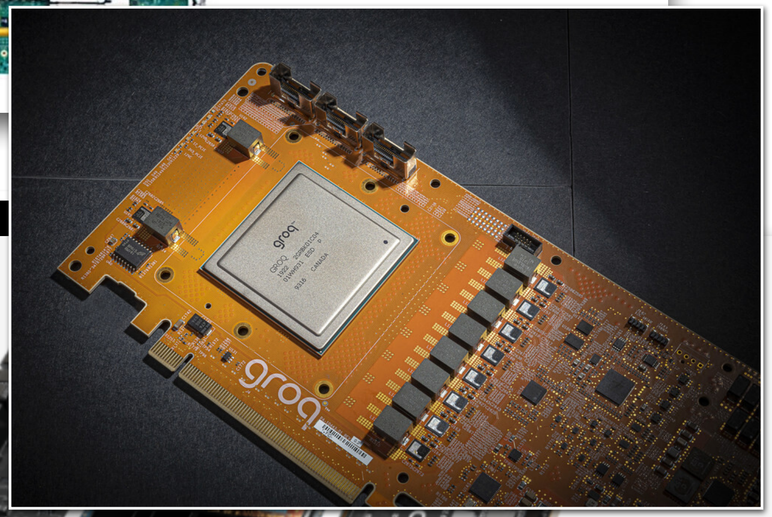


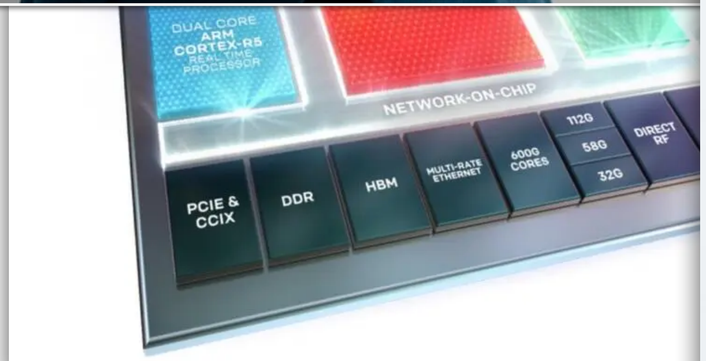


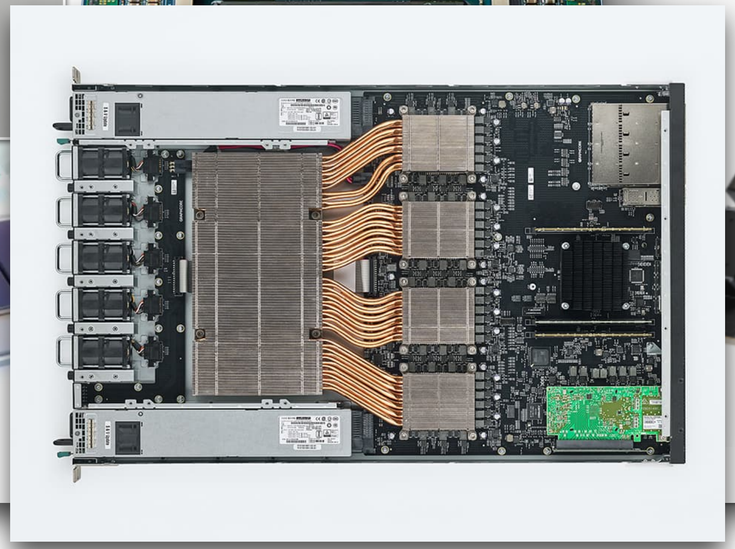
Coprocessors

Coprocessors









Coprocessors for Science

- We are not the primary customers of these chips
- How can we leverage advancements in industry?
  - Be **flexible** - map the right architecture to right application
  - Build **benchmarks** for physics workloads
  - **Programmability** important for accessibility - why GPU is leading the pack

SONIC

Services for Optimized Network Inference on Coprocessors

SONIC

Services for Optimized Network Inference on Coprocessors

- SONIC: tools for inference-as-a-service within experimental SW frameworks
- Abstract software AND hardware with modern containerization tools
- Scalable, Flexible, Adaptable, Non-disruptive

Options

Domain Algo

ML

as a Service (aaS)

direct connect

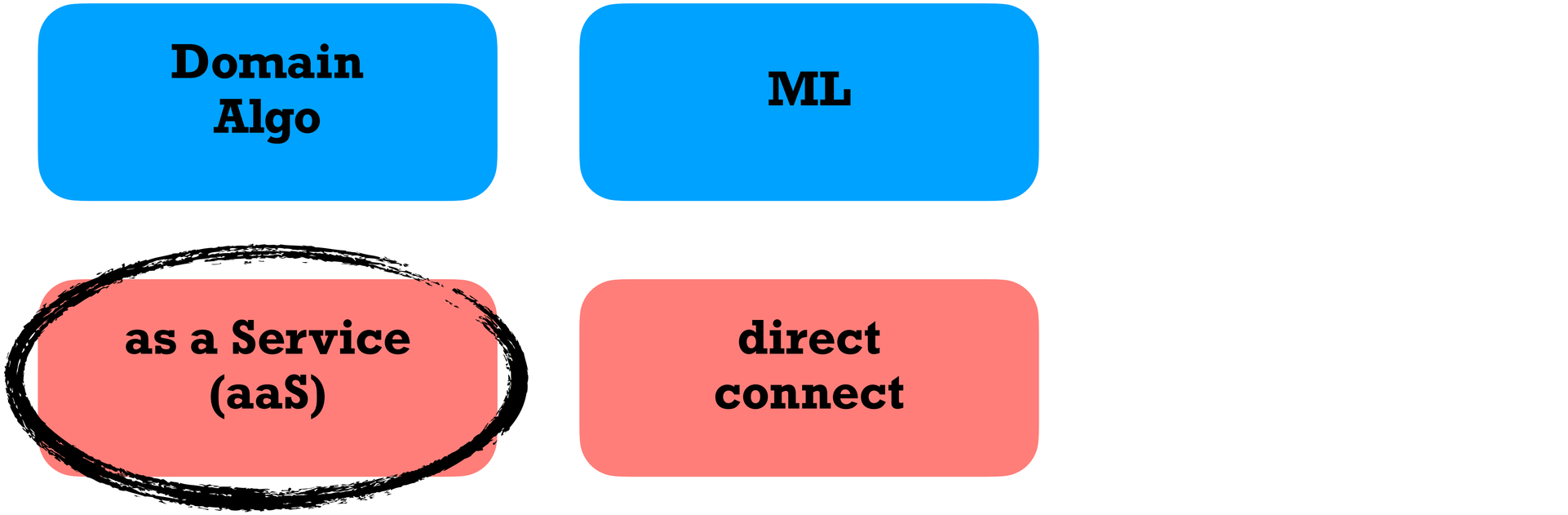
GPU

FPGA

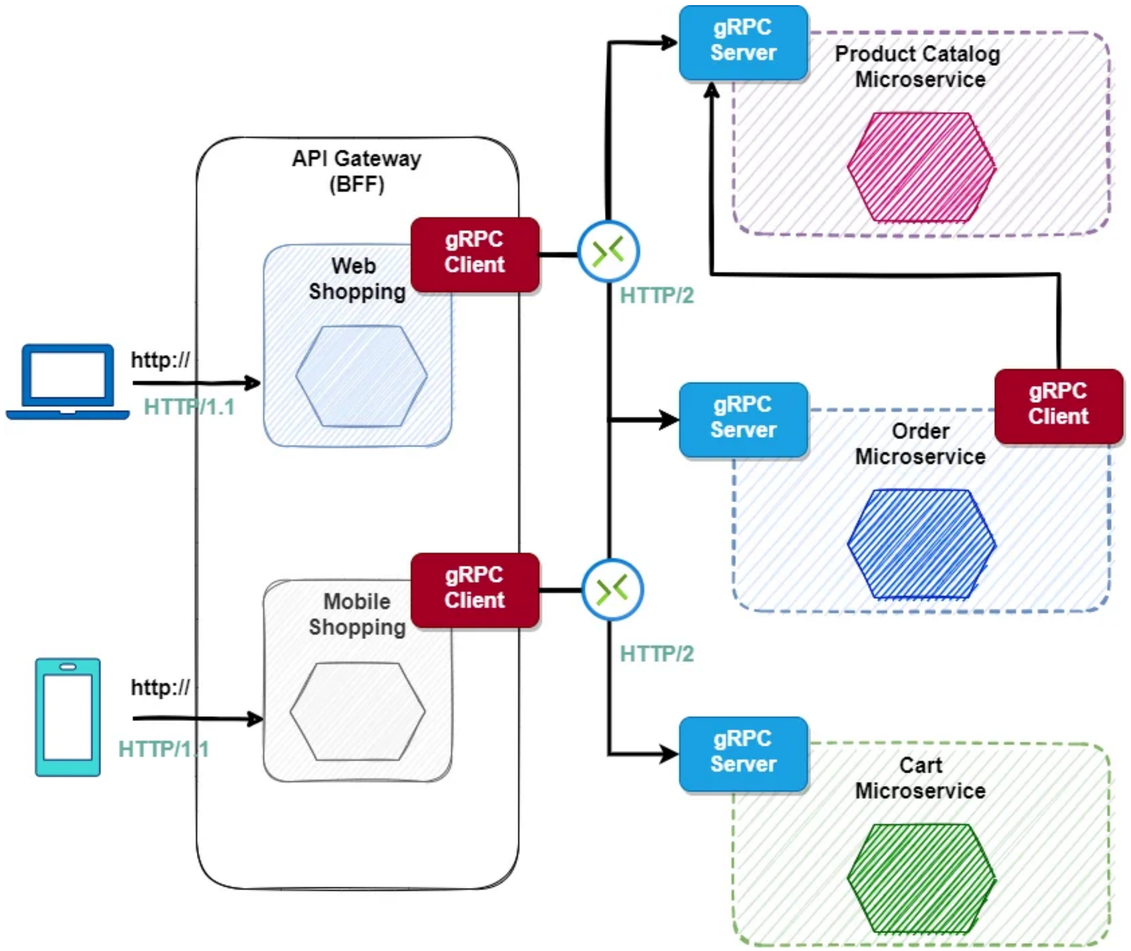
ASIC

...

Options



aaS

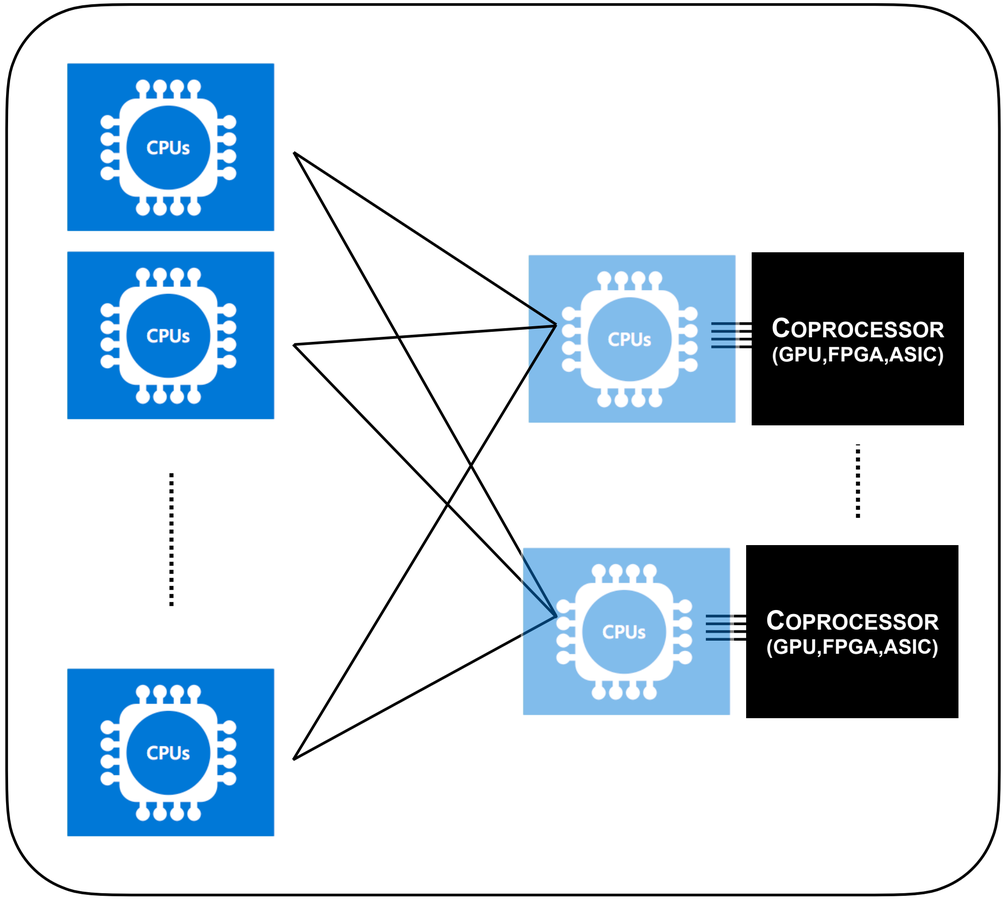


gRPC is a cross-platform open source high performance remote procedure call framework...

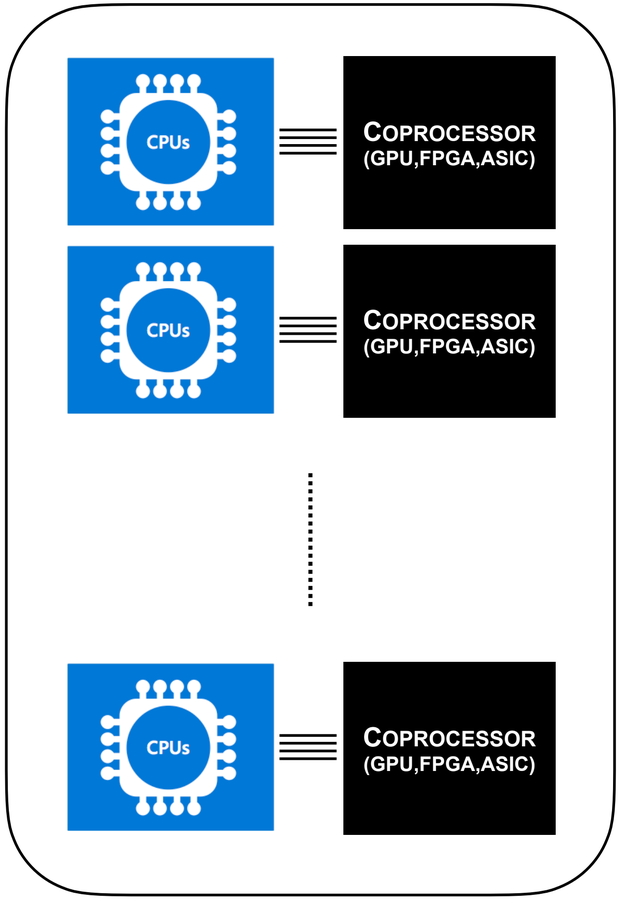
It uses HTTP/2 for transport, Protocol Buffers as the interface description language, and provides features such as authentication, bidirectional streaming and flow control, blocking or nonblocking bindings, and cancellation and timeouts. It generates cross-platform client and server bindings for many languages.

https://en.wikipedia.org/wiki/GRPC

aaS vs direct connect

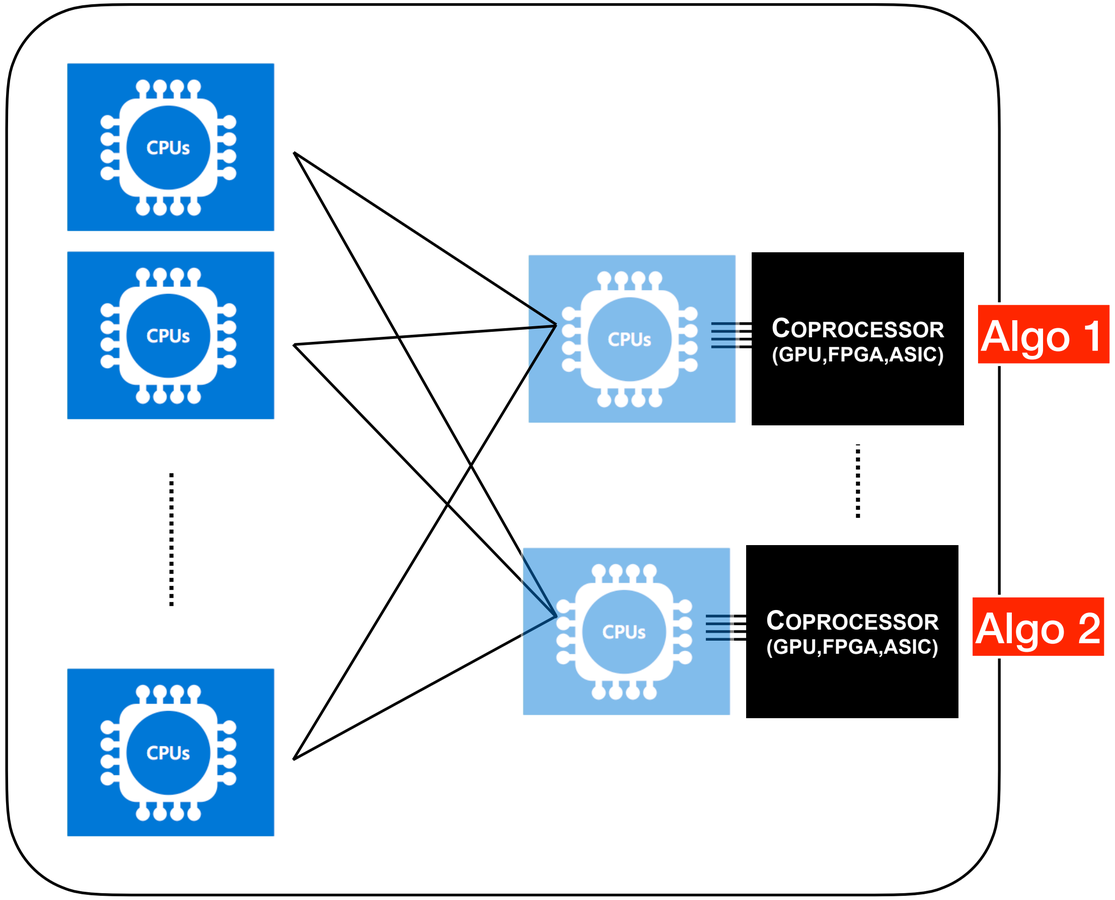


Pros:
scalable algorithms
scalable to the grid/cloud
heterogeneity (mixed hardwares)

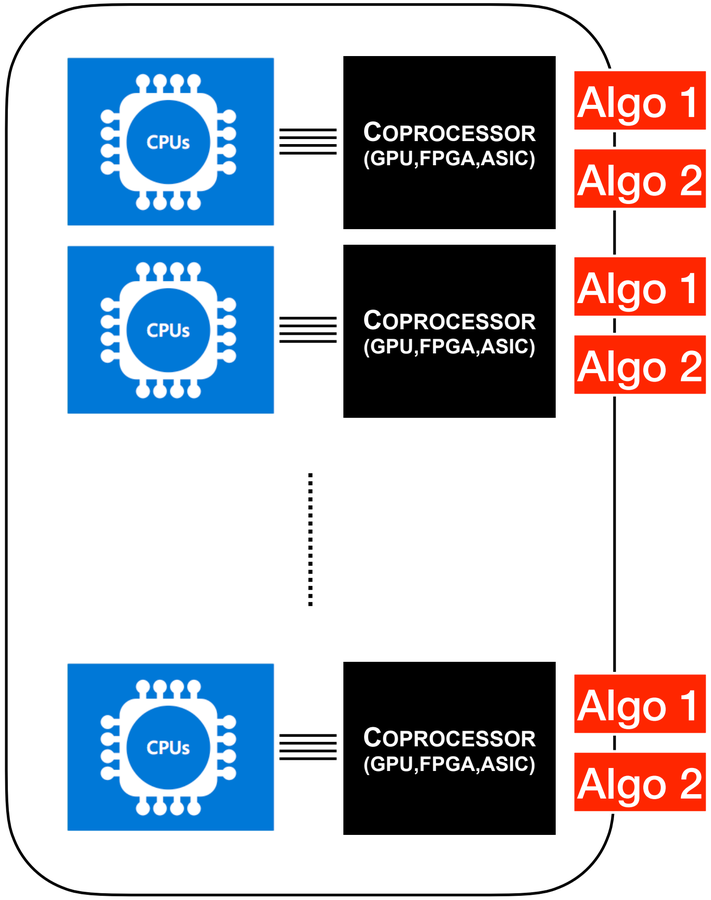


Pros:
less system complexity
no network latency

aaS vs direct connect



Pros:
scalable algorithms
scalable to the grid/cloud
heterogeneity (mixed hardwares)



Pros:
less system complexity
no network latency

SONIC

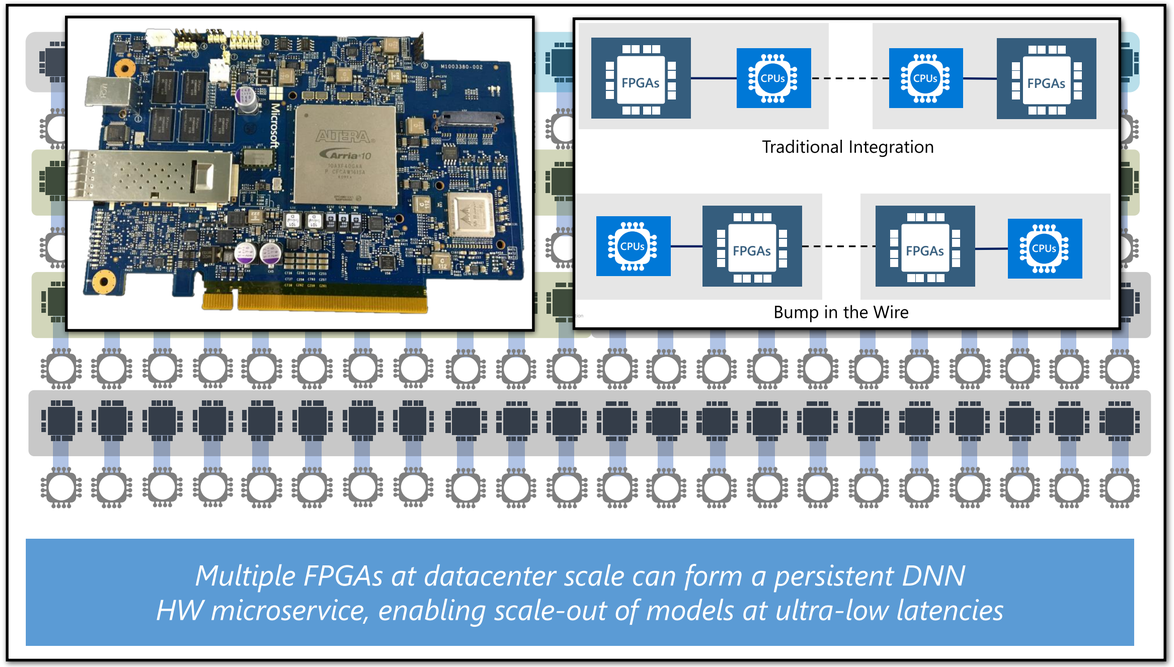
Services for Optimized Network Inference on Coprocessors

- **Scalable**: Not bound to coprocessors directly connected to CPUs at a given node
- **Flexible**: Can be used on any hardware and can "right-size" the number of coprocessors needed based on the task
- **Adaptive**: Abstracts server software stack including compiling latest TF or Torch, custom libraries or languages
- **Non-disruptive**: builds off current experimental infrastructure, can offload as needed without changing current paradigm

A brief history

https://arxiv.org/abs/1904.08986
https://arxiv.org/abs/2009.04509
https://en.wikipedia.org/wiki/Amdahl%27s_law

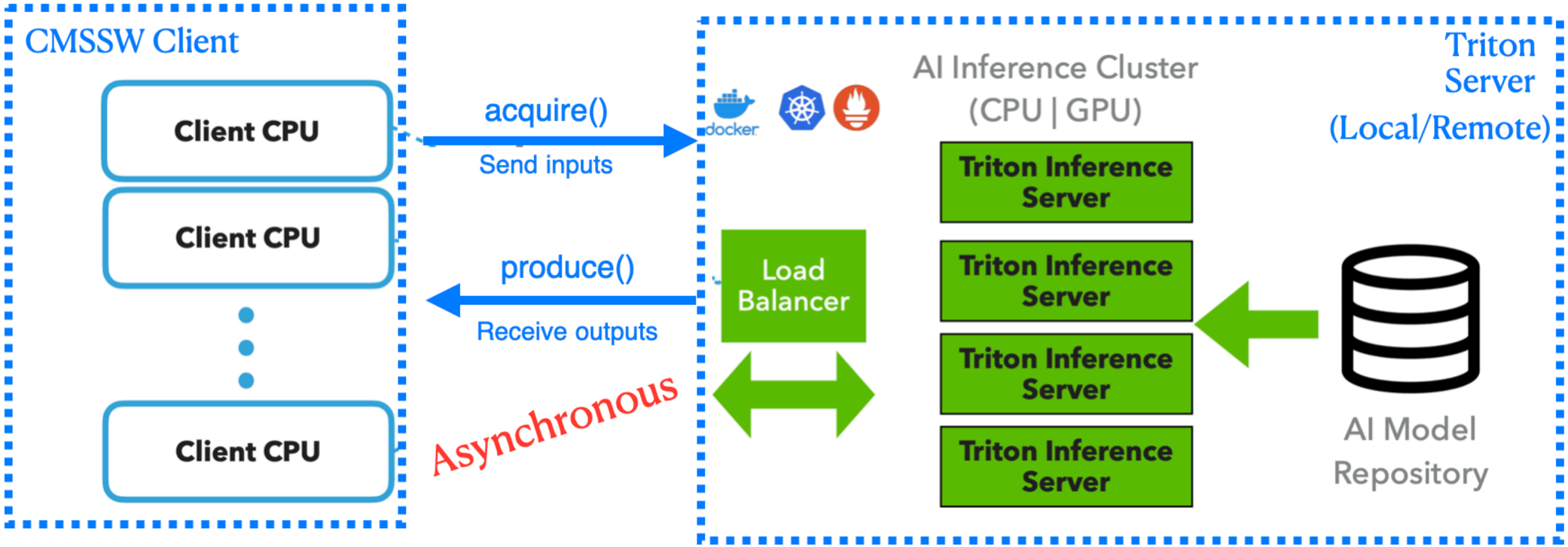
- First deployed on FPGAs for CMS in collaboration with Microsoft Brainwave
- Then deployed on GPUs for LHC and neutrino applications
- Demonstrated for GW experiments and now explored in many other areas
  - See recent SONIC mini-workshop: https://indico.cern.ch/event/1372201



	Wall time (s)		
	ML module	non-ML modules	Total
CPU only	220	110	330
CPU + GPUaaS	13	110	123

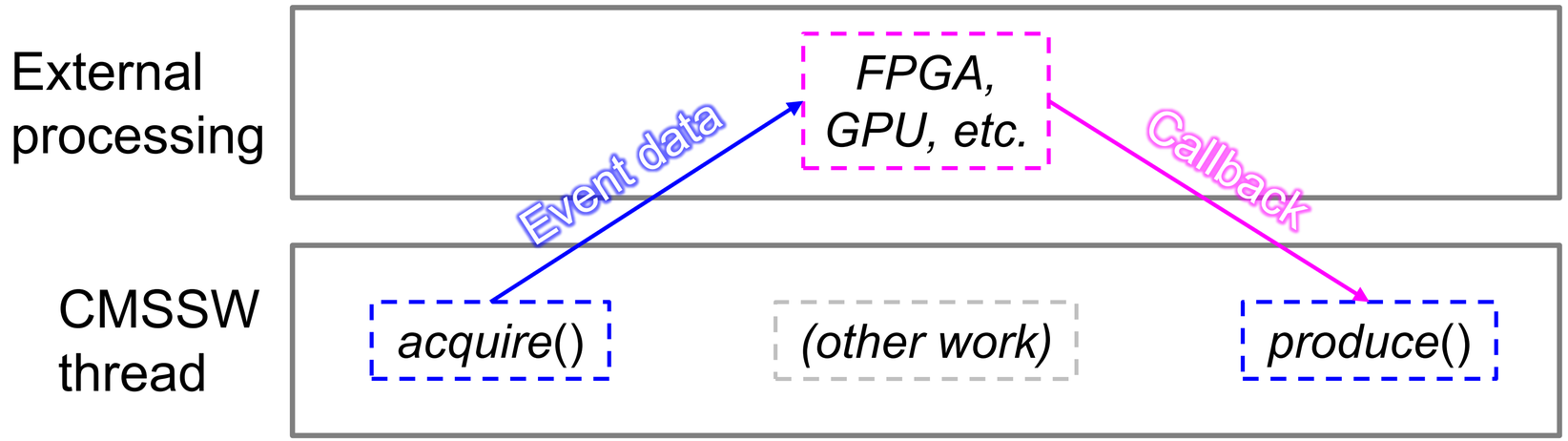
State-of-the-art

- Most popular/flexible workflows currently use GPUs and leverage Nvidia Triton Server



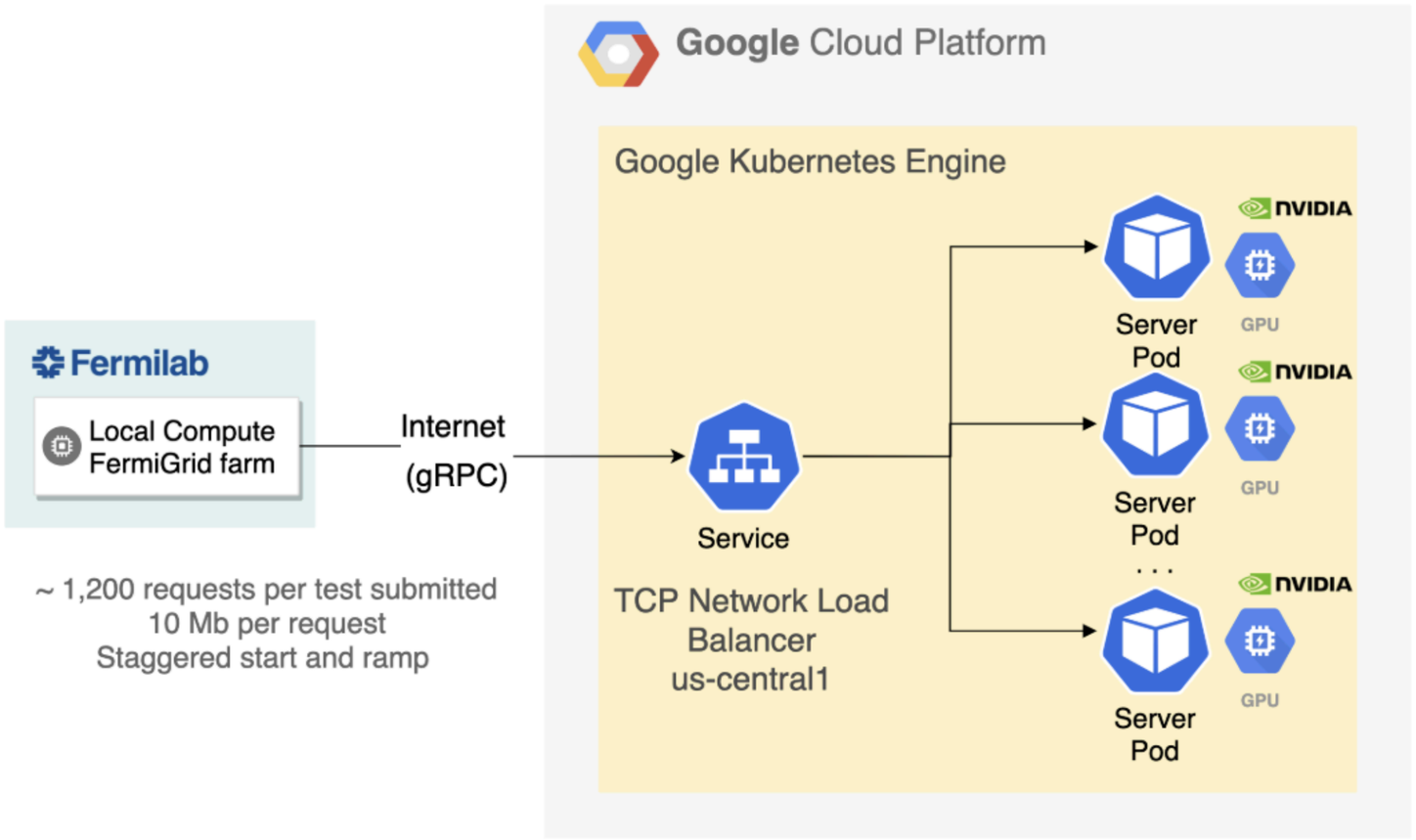
State-of-the-art

- Most popular/flexible workflows currently use GPUs and leverage Nvidia Triton Server

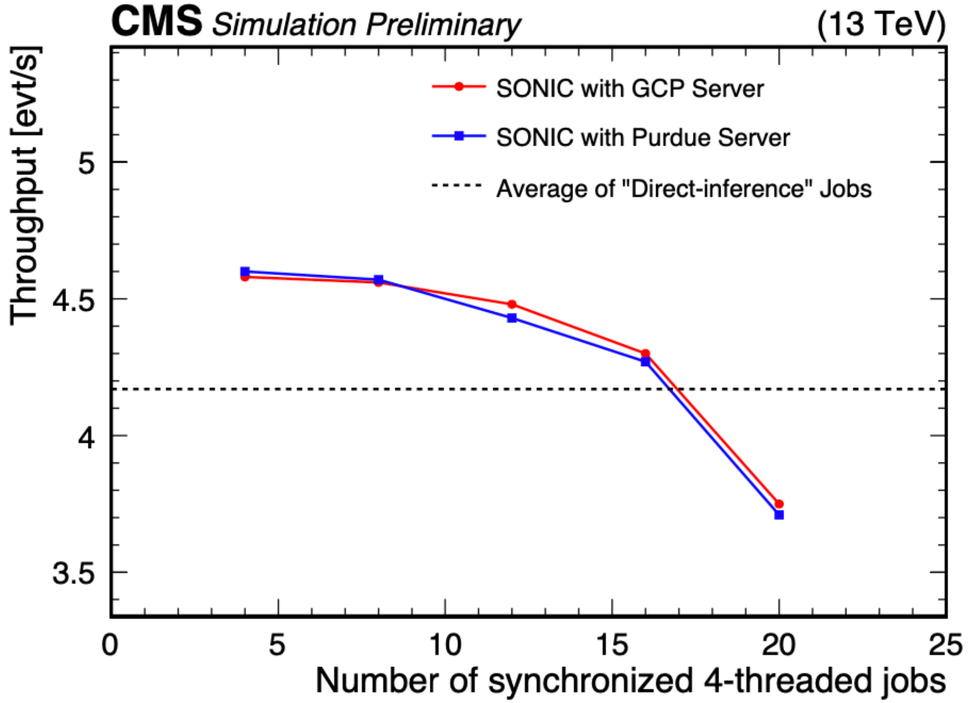


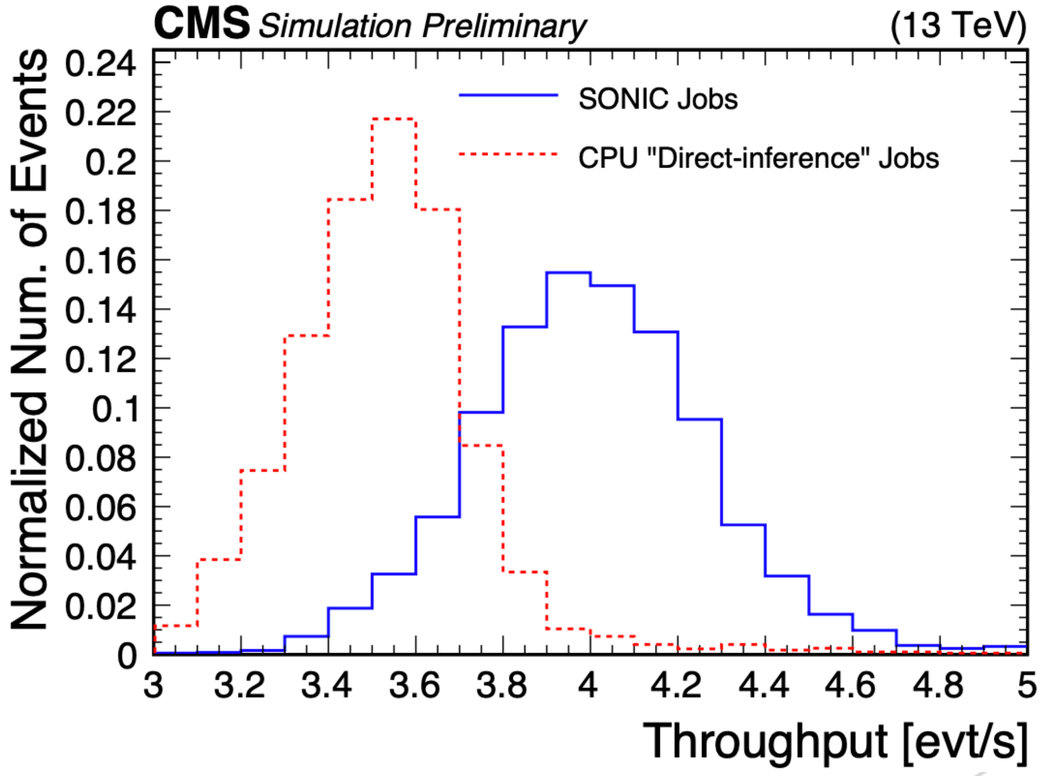
State-of-the-art

- Most popular/flexible workflows currently use GPUs and leverage Nvidia Triton Server



Highlight: first CMS paper





- Similar performance for servers running at different sites. Network overload impacts is small!
- Scales up well with large number of CPUs and GPUs
- CMS-PAS-MLG-23-001: First CMS paper systematically studying the computing performance on GPUs, and also first study of the inference as-a-service approach

Outlook

- SONIC is one of the most promising approaches to using coprocessors for accelerating computing workloads in science
- A lot of interesting areas to work on: benchmarking & tech survey, scale-out, resource orchestration, etc.

Tutorial time

- Now you have a chance to try SONIC/Triton tools out yourselves!