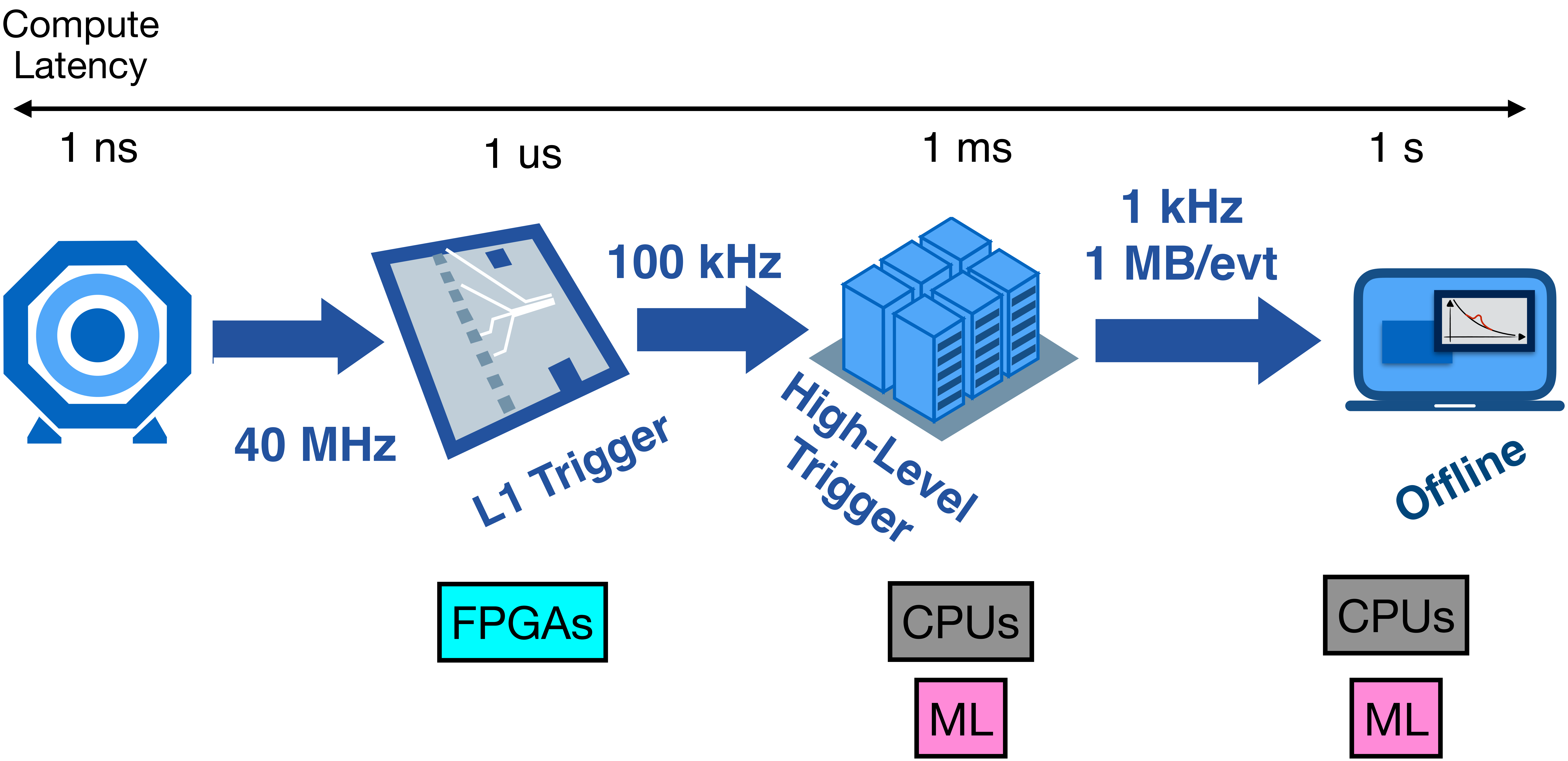
Real-time on-detector AI

Nhan Tran

+ Javier Duarte, Lindsey Gray, Sergo Jindariani, Kevin Pedro, Bill Pellico, Gabe Perdue, Ryan Rivera, Brian Schupbach, Kiyomi Seiya, Jason St. John, Mike Wang,…

May 10, 2019

CMS EVENT PROCESSING



Compute Latency

1 ns

1 us

1 ms

1 s

40 MHz

L1 Trigger

100 kHz

High-Level Trigger

1 kHz

1 MB/evt

Offline

FPGAs

CPUs

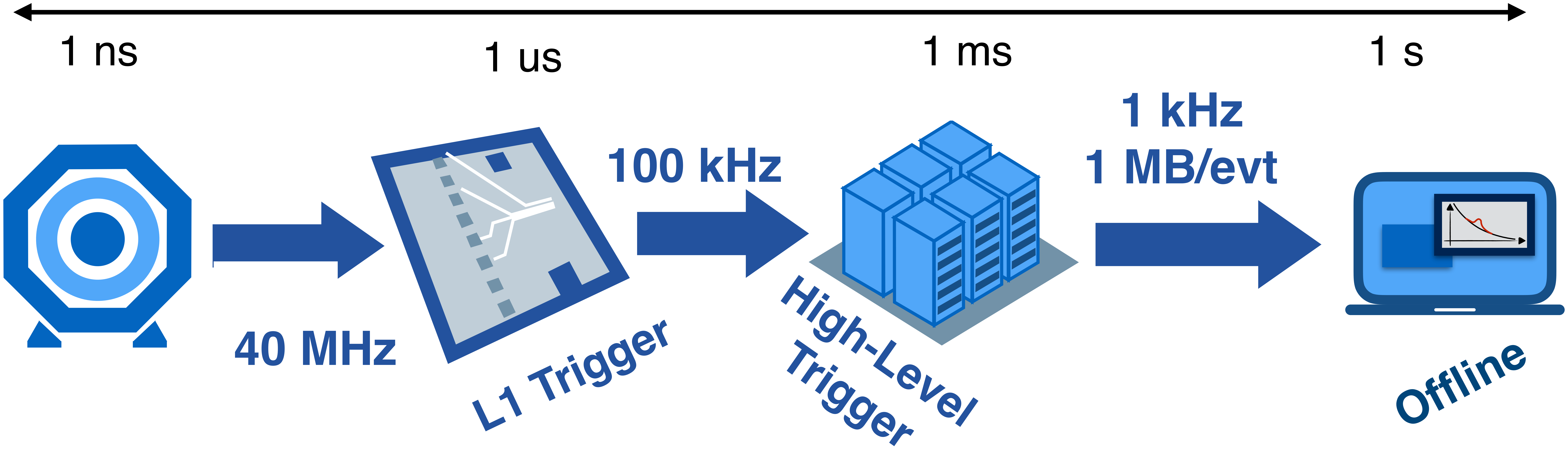
ML

CPUs

ML

CMS EVENT PROCESSING

Compute Latency



1 ns

1 us

1 ms

1 s

40 MHz

L1 Trigger

100 kHz

High-Level Trigger

1 kHz
1 MB/evt

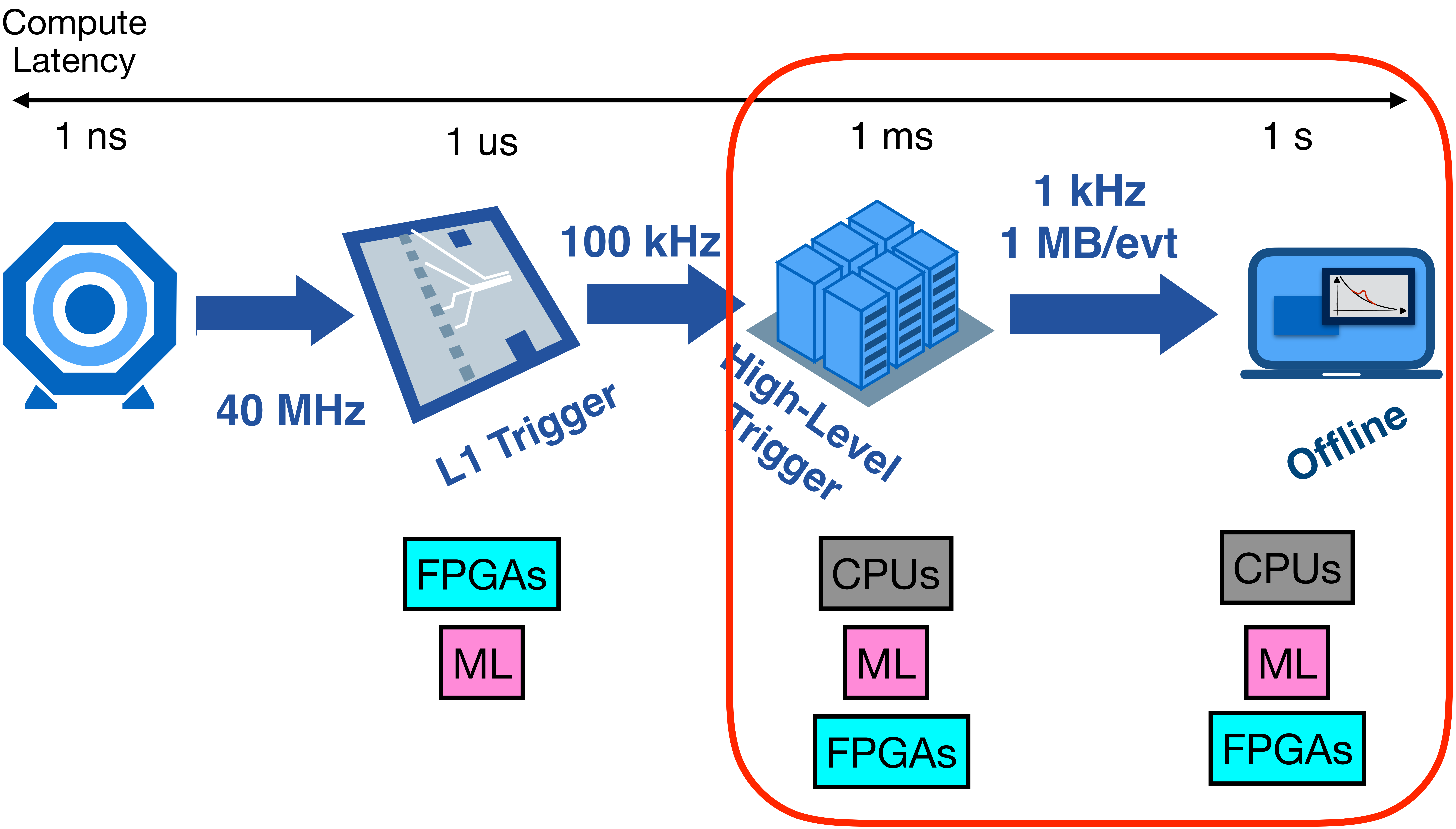
Offline

L1 Trigger: FPGAs, ML

High-Level Trigger: CPUs, ML, FPGAs

Offline: CPUs, ML, FPGAs

CMS EVENT PROCESSING



Compute Latency

1 ns

1 us

1 ms

1 s

40 MHz

L1 Trigger

100 kHz

High-Level Trigger

1 kHz
1 MB/evt

Offline

FPGAs

ML

CPUs

ML

FPGAs

CPUs

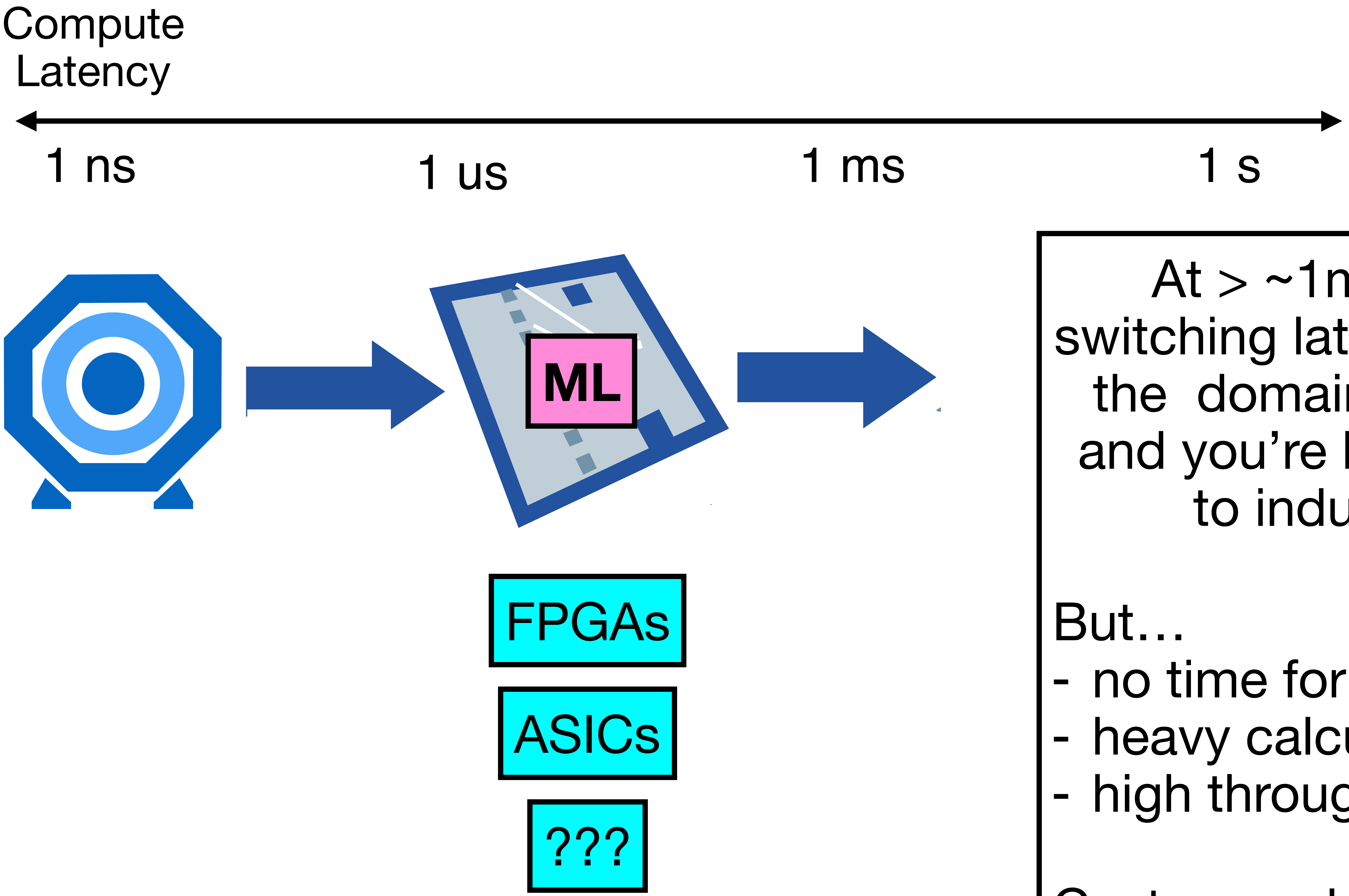
ML

FPGAs

A whole other talk,
mostly for computing group

https://arxiv.org/abs/1904.08986

CMS EVENT PROCESSING



At > ~1ms (network switching latencies), this hits the domain of CPU/GPU and you're better off going to industry tools.

But…

- no time for CPU
- heavy calculation
- high throughput

Custom real-time detector AI applications are for you!

HIGH RATE AND INTELLIGENT EDGE

O(50-100) optical transceivers running at ~O(15) Gbs

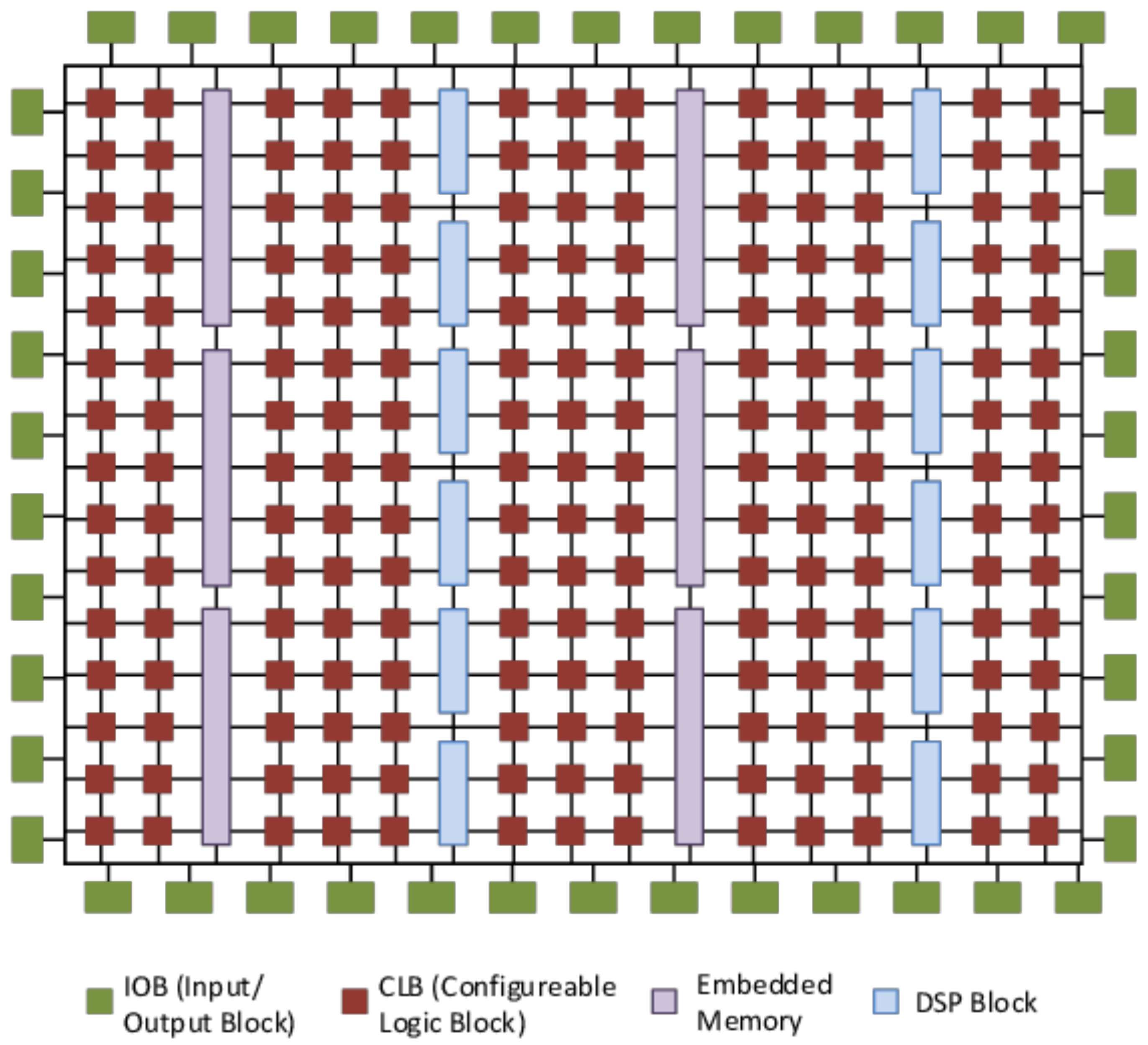
Traditionally, FPGAs programmed with low-level languages like Verilog and VHDL

High level synthesis (HLS)

New languages C-level programming with specialized preprocessor directives which synthesizes optimized firmware;

Drastically reduces development times for firmware

FPGA

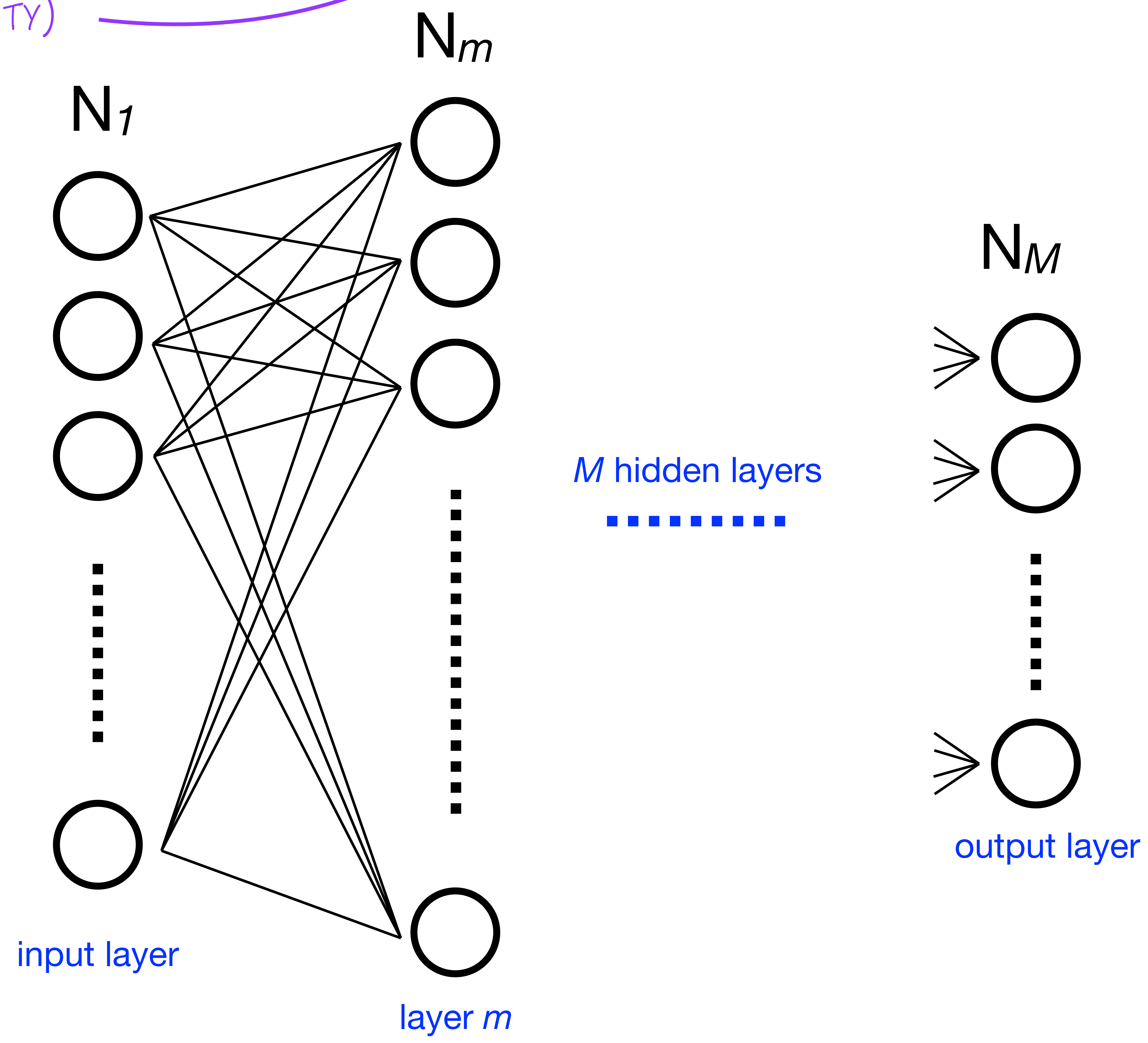


DSPs (multiply-accumulate, etc.)
Flip Flops (registers/distributed memory)
LUTs (logic)
Block RAMs (memories)

Neural networks and linear algebra

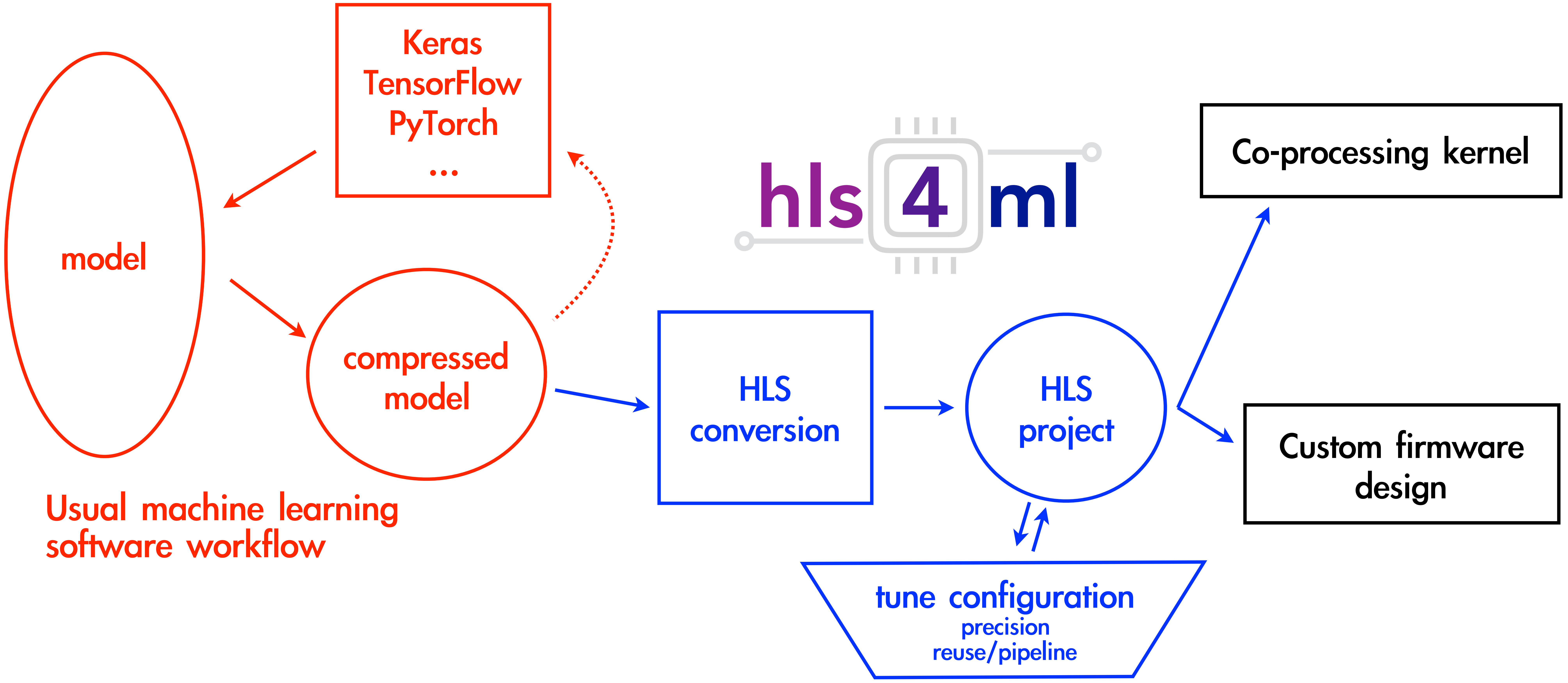
$$\vec{O}_j = \phi(\vec{I}_i \times \overleftrightarrow{W}_{ij} + \vec{b}_j)$$

ϕ = ACTIVATION FUNCTION (NON-LINEARITY)



PROJECT OVERVIEW

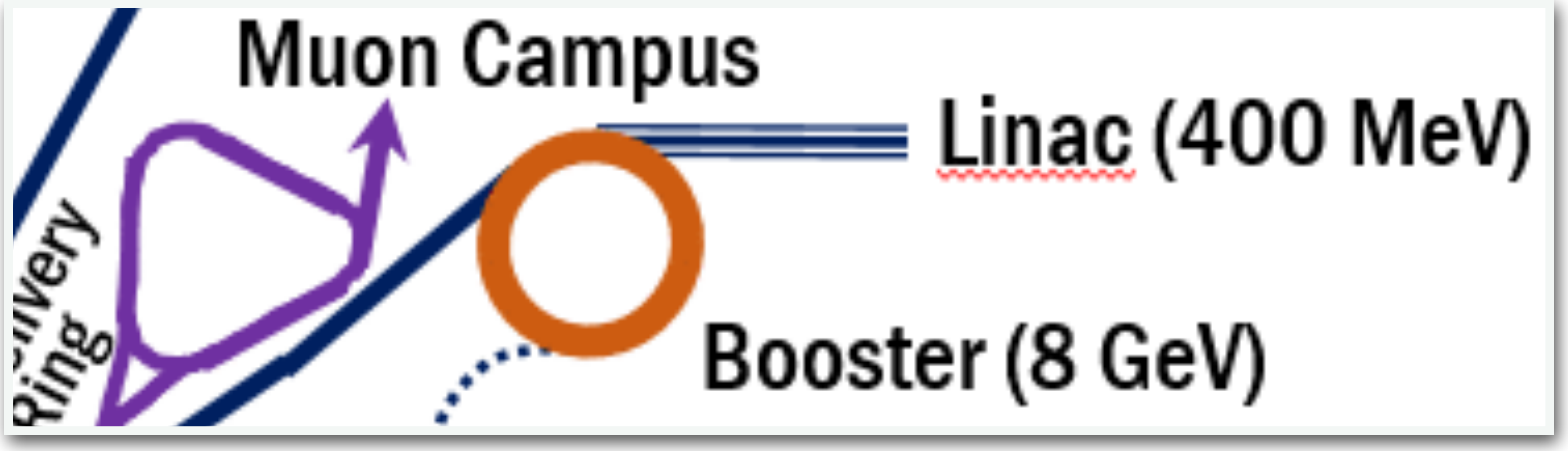
Quantization, Compression, Parallelization made easy with hls4ml!



Results and outlook:

4000 parameter network inferred in < 100 ns with 30% of FPGA resources!
Muon pT reconstruction with NN reduces rate by 80%
Larger networks and different architectures actively developed (CNN, RNN, Graph)

TECH TRANSFER



LDRD:

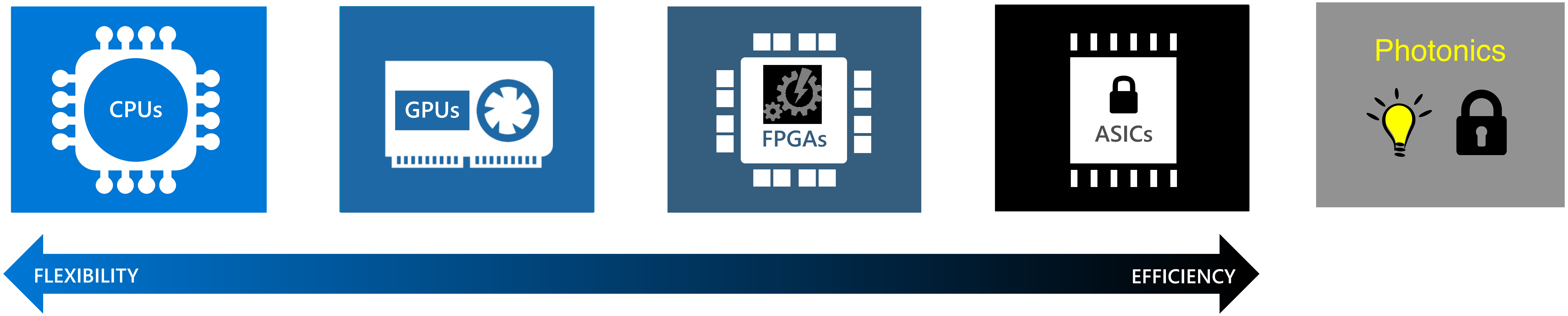
Add “reinforcement learning” to improve accelerator operations

Tuning the Gradient Magnet Power Supply (GMPS) system for the Booster

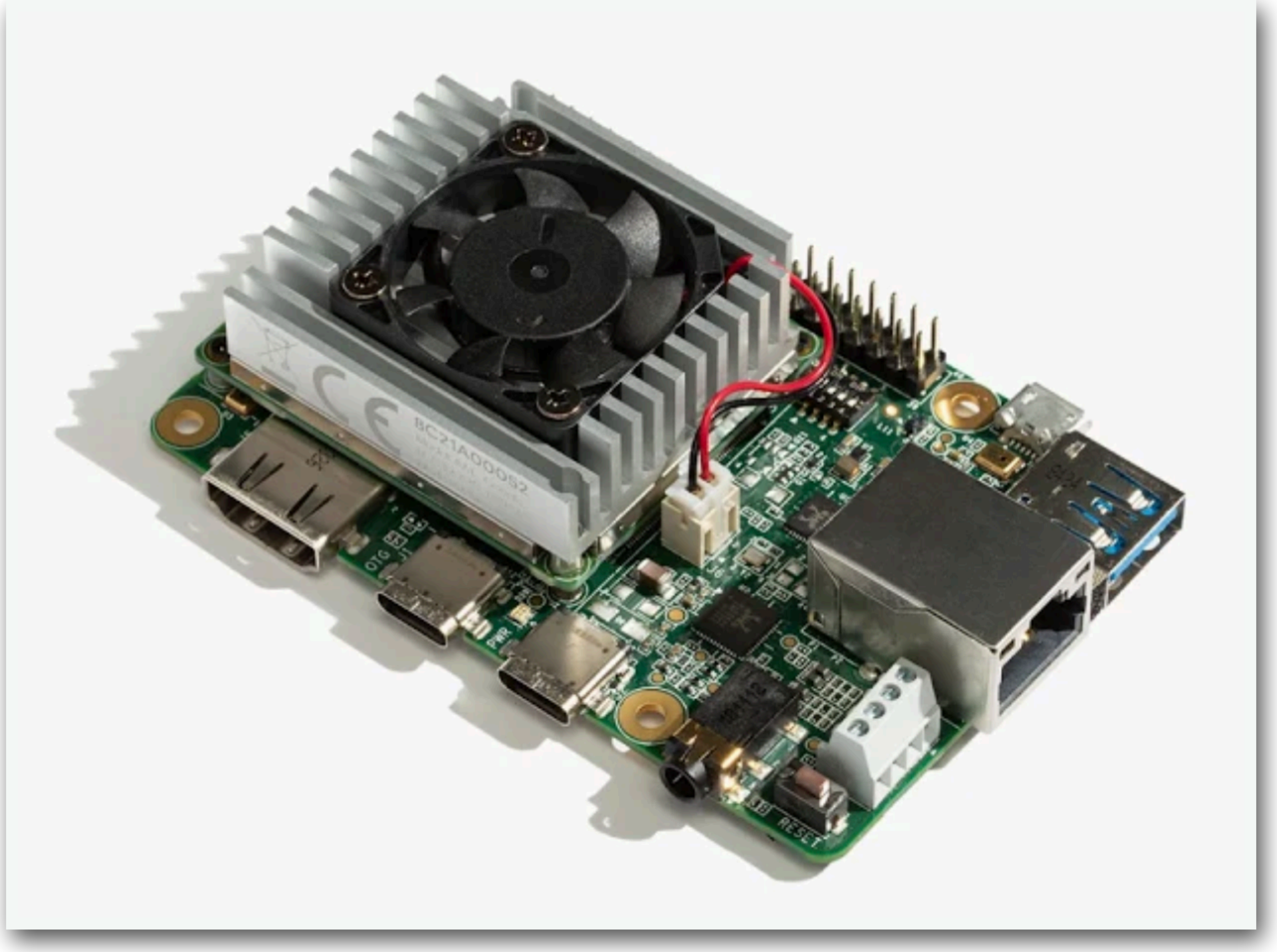
will be a first for accelerators and critical for future machines

A first proof-of-concept, could apply across the accelerator complex

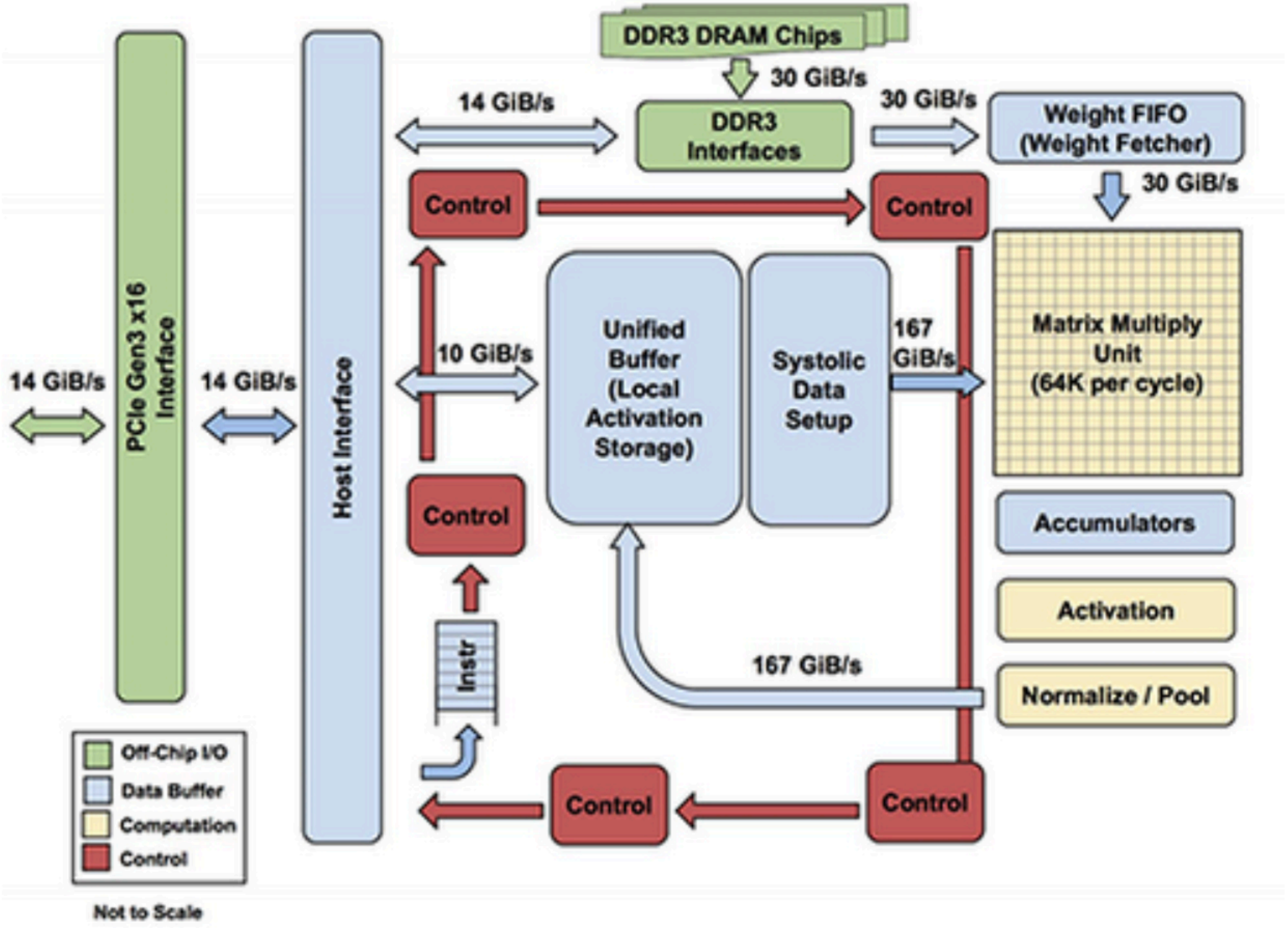
Futurisitic ideas



ASICs

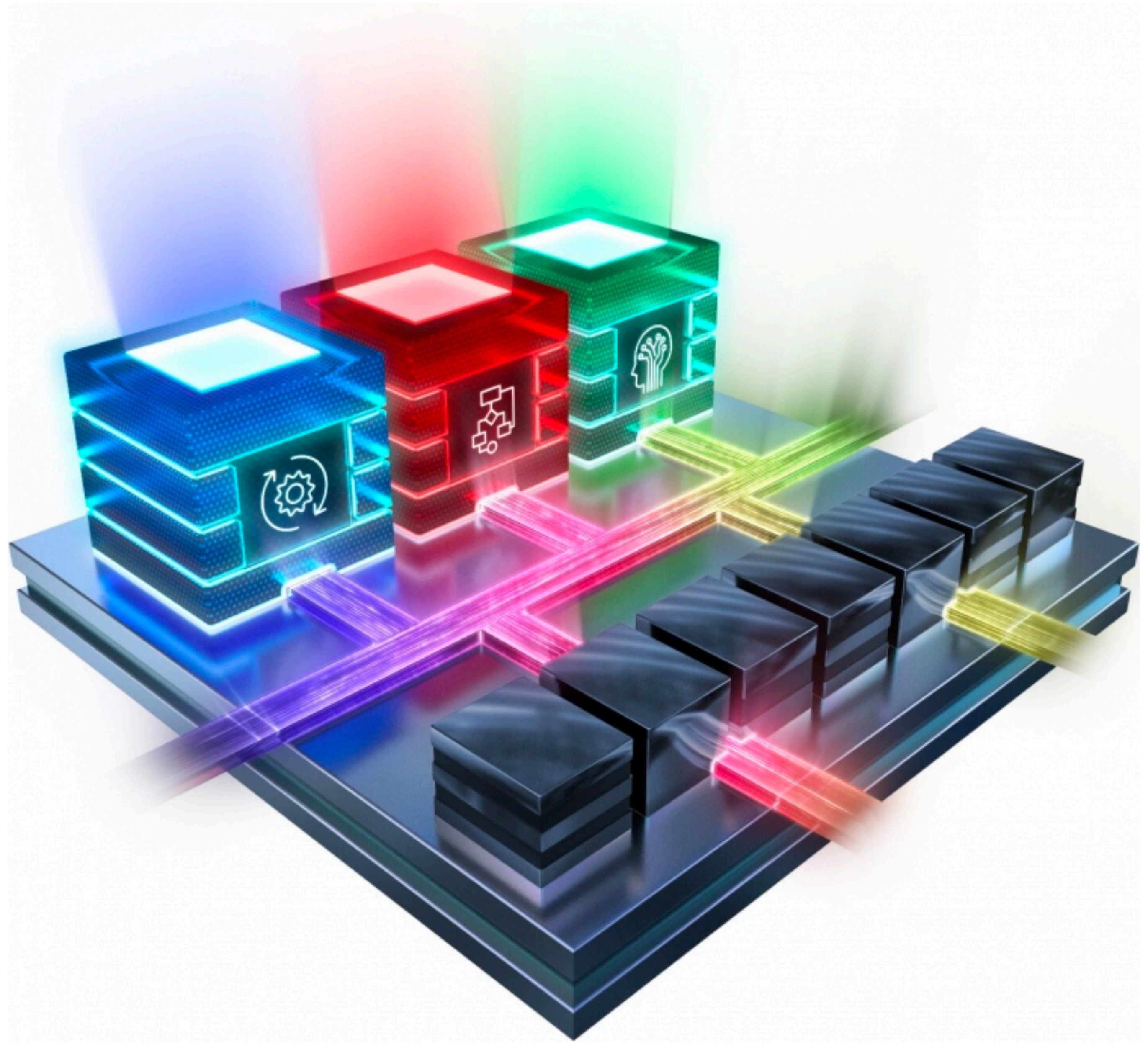


Edge TPU

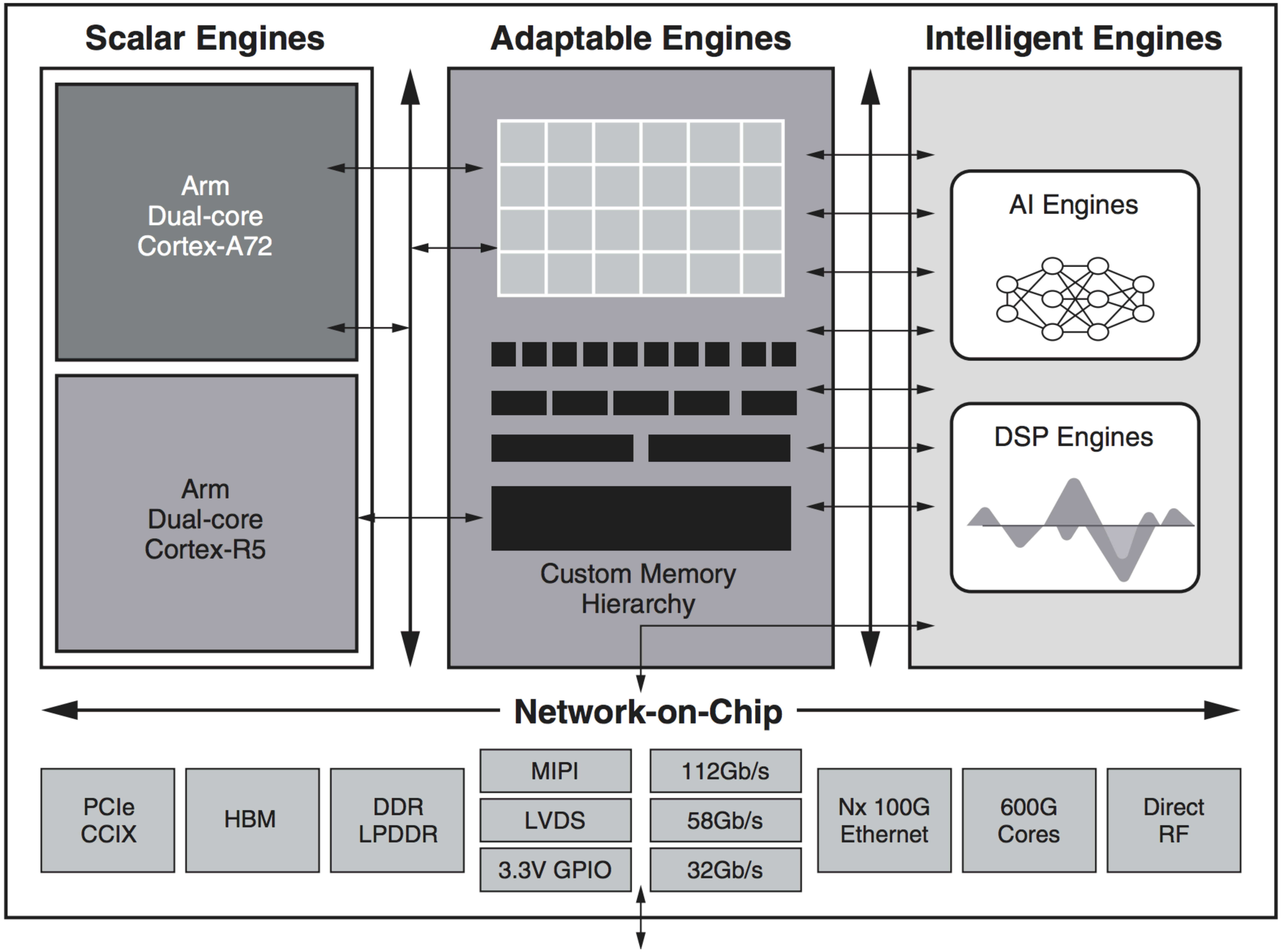


TPU Block Diagram

FRANKENSTEINS



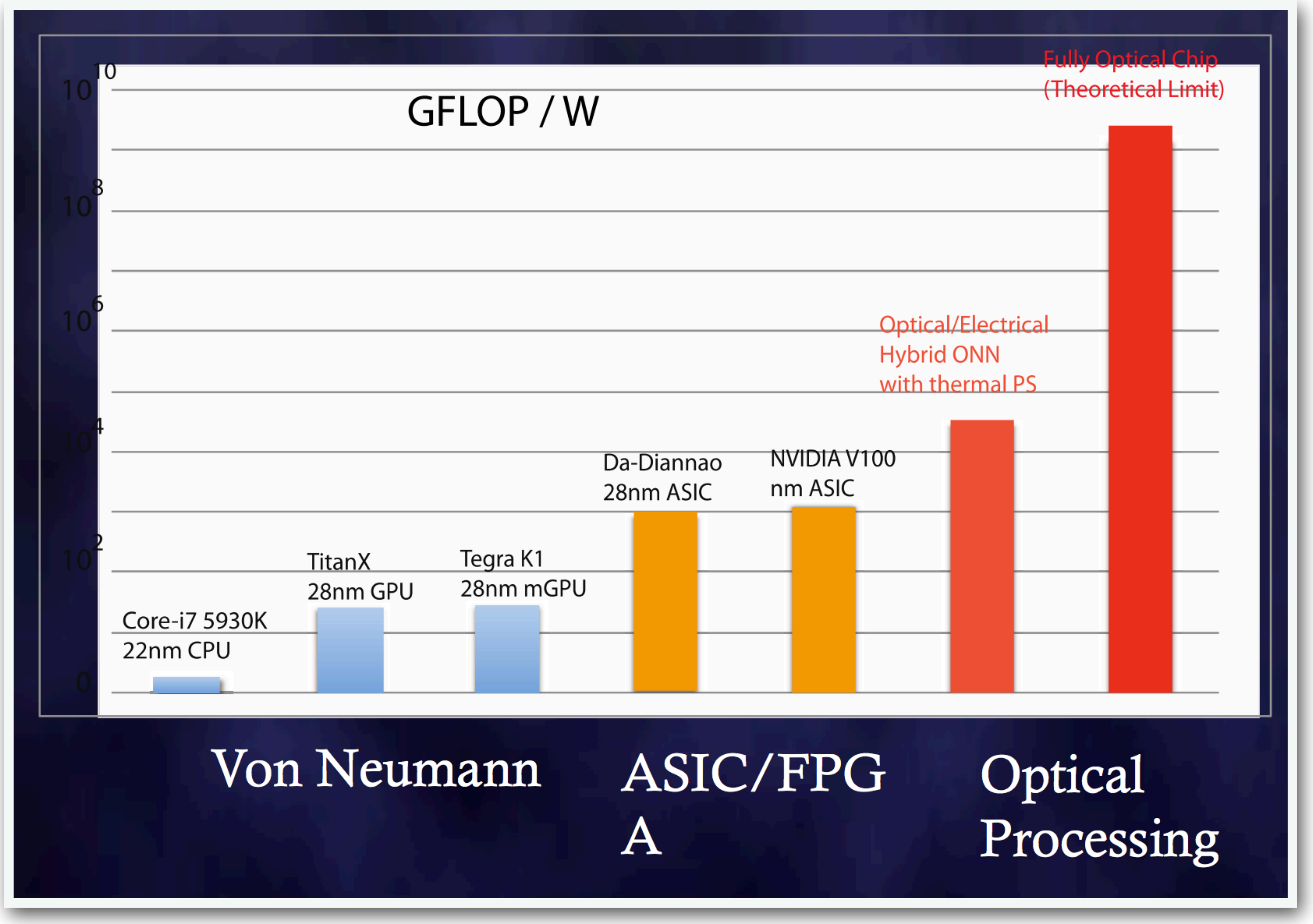
Xilinx Versal

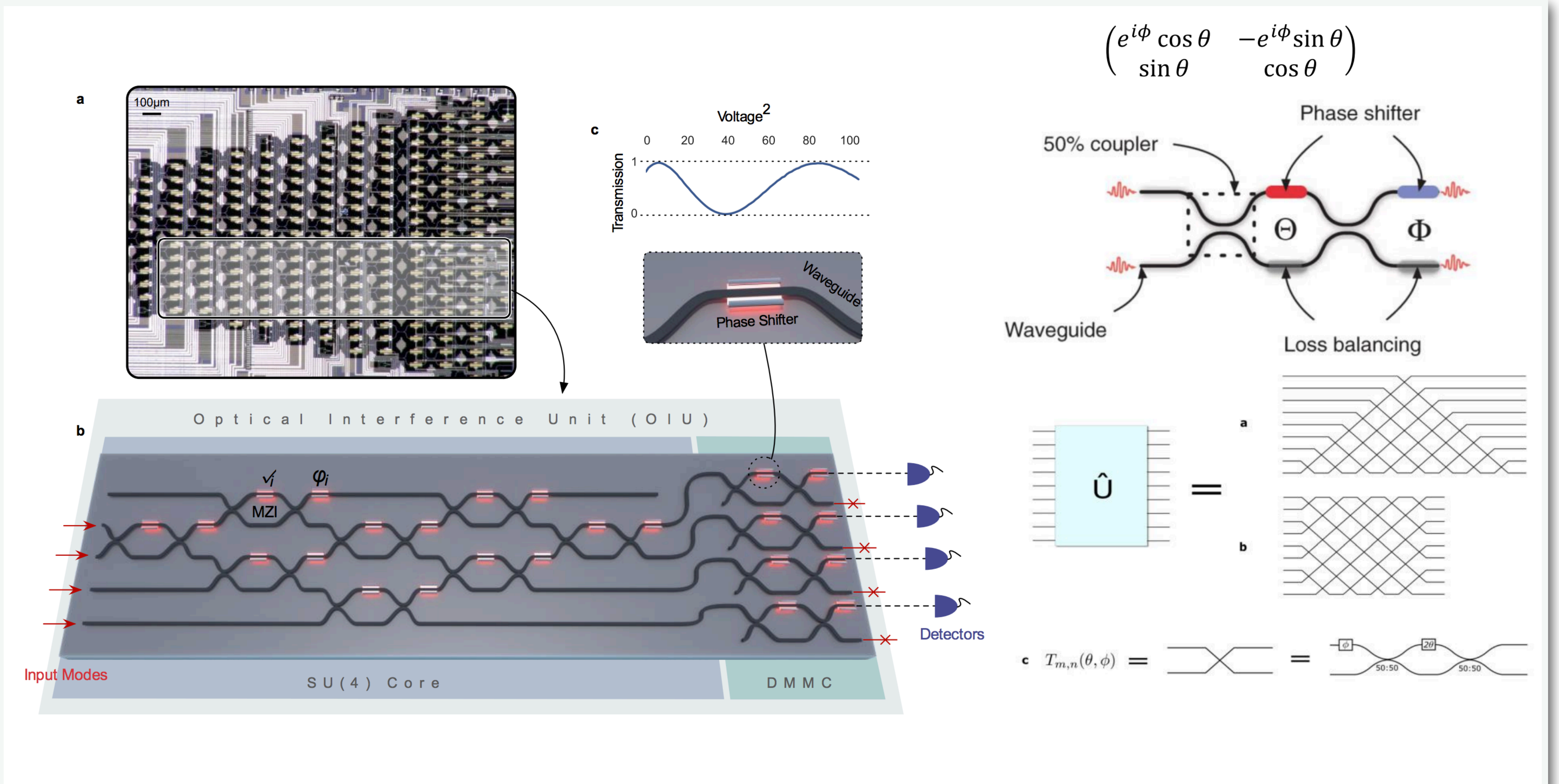


PHOTONICS

Even faster — a neural network photonics “ASIC”

Recently fabrication processes have become more reliable





Y.Shen and N. Harris et al. *“Deep Learning with Coherent Nanophotonic Circuit” Nature Photonics* **11**, 441–446 (2017)

In contact with 2 groups (MIT, Princeton) on possible photonics prototypes

SUMMARY

- Real-time AI brings processing power on-detector
  - Improves losses in efficiency/performance for triggers - gains back physics
    - Other physics scenarios? A lot of efficiency loss from high bandwidth systems…
  - Want to demonstrate helps with automation and efficiency of system operation

- Futuristic technologies could bring even more front end processing power
  - Hardened vector DSPs, electronics and photonics