ML in FPGA
(& ASICs, etc)
(for embedded systems)
(for science, particularly the EIC)

Nhan Tran, Fermilab
October 12, 2022

Outline

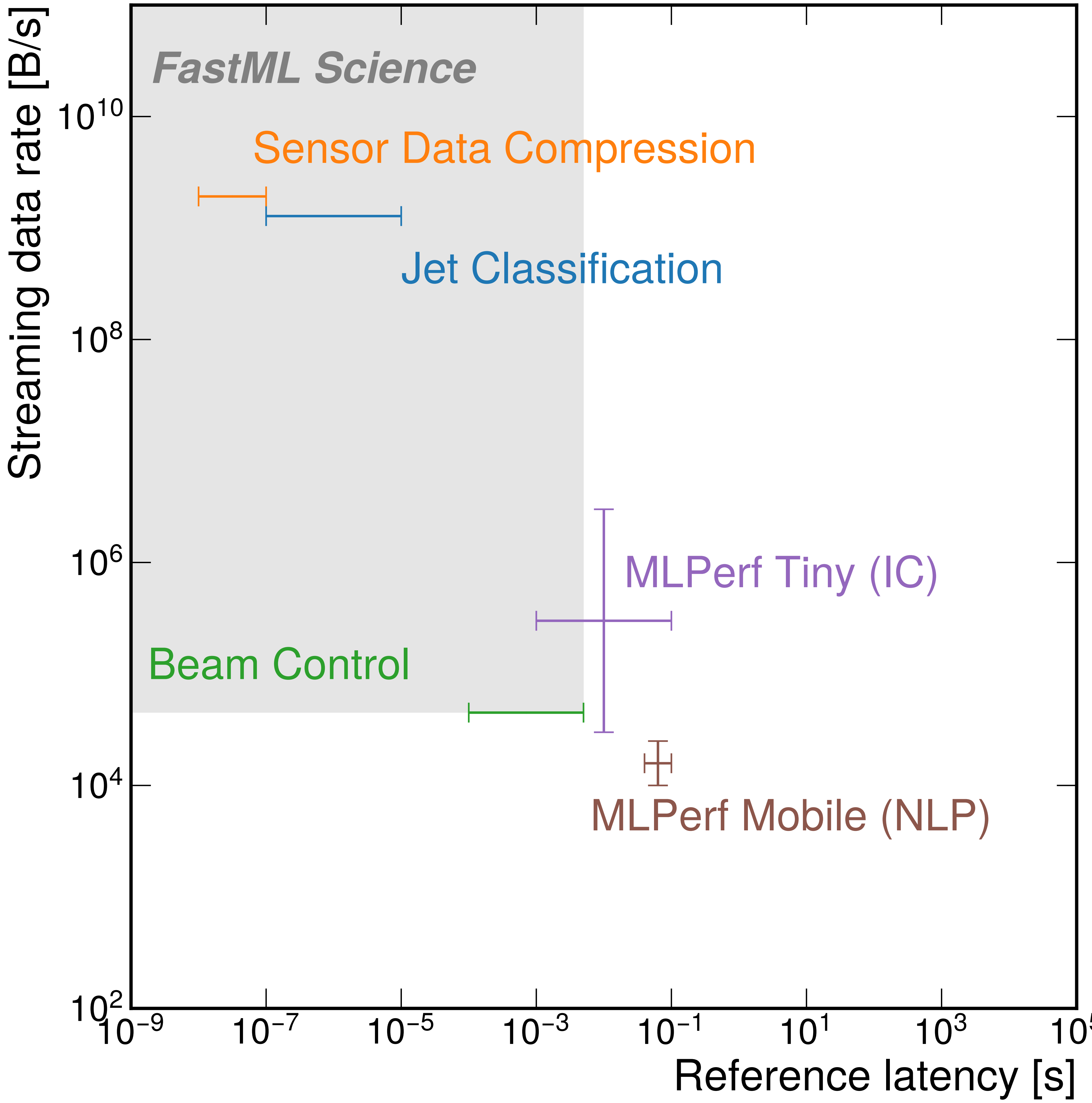
- Motivation
- State-of-the-art workflow for FPGA/ASIC
  - Towards a sustainable and robust ecosystem
  - Emerging technologies

This is a big area!
My approach — present key important topics and provide a lot of references.
Follow references if you are interested in learning more;
reach out if you are even more interested after following the references.

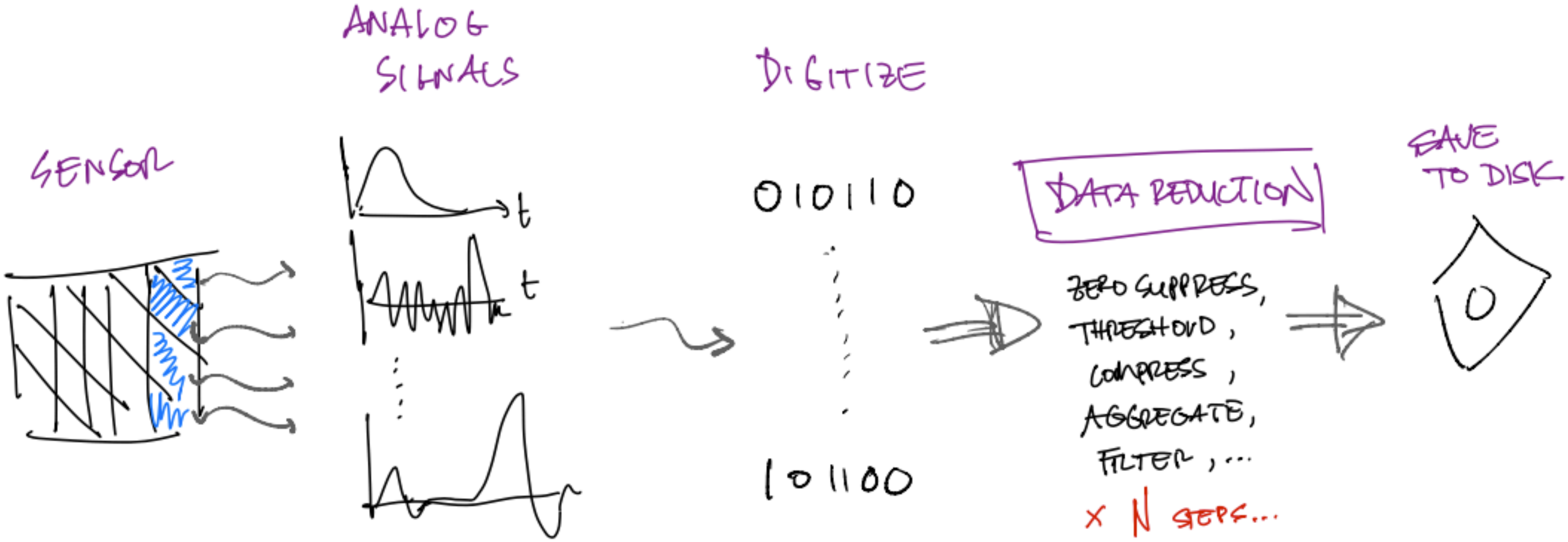
Motivation

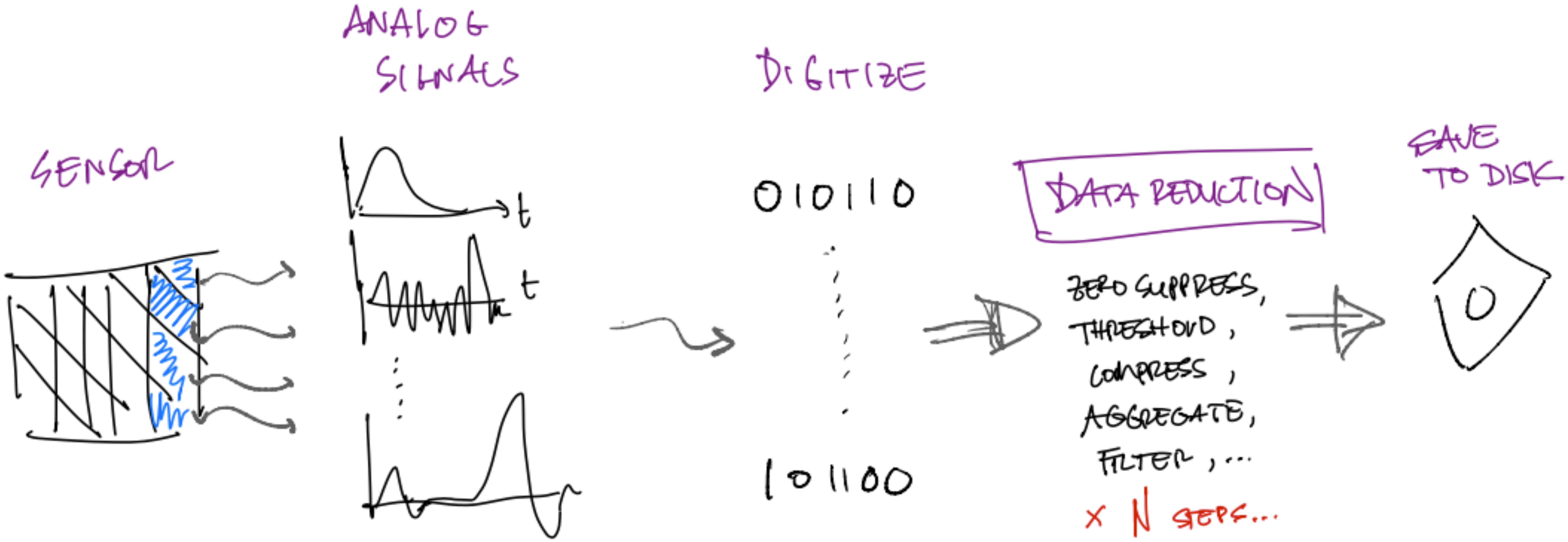
High intensity collider experiments explore nature at the **finest temporal and spatial scales** Leads to data rates far surpassing industry — requires developing **innovative techniques**

- ML in specialized embedded architectures require in ***real-time*** to reduce and filter data
- Optimal data selection enables **more efficient operation, saves lost data, and accelerates time-to-discovery**

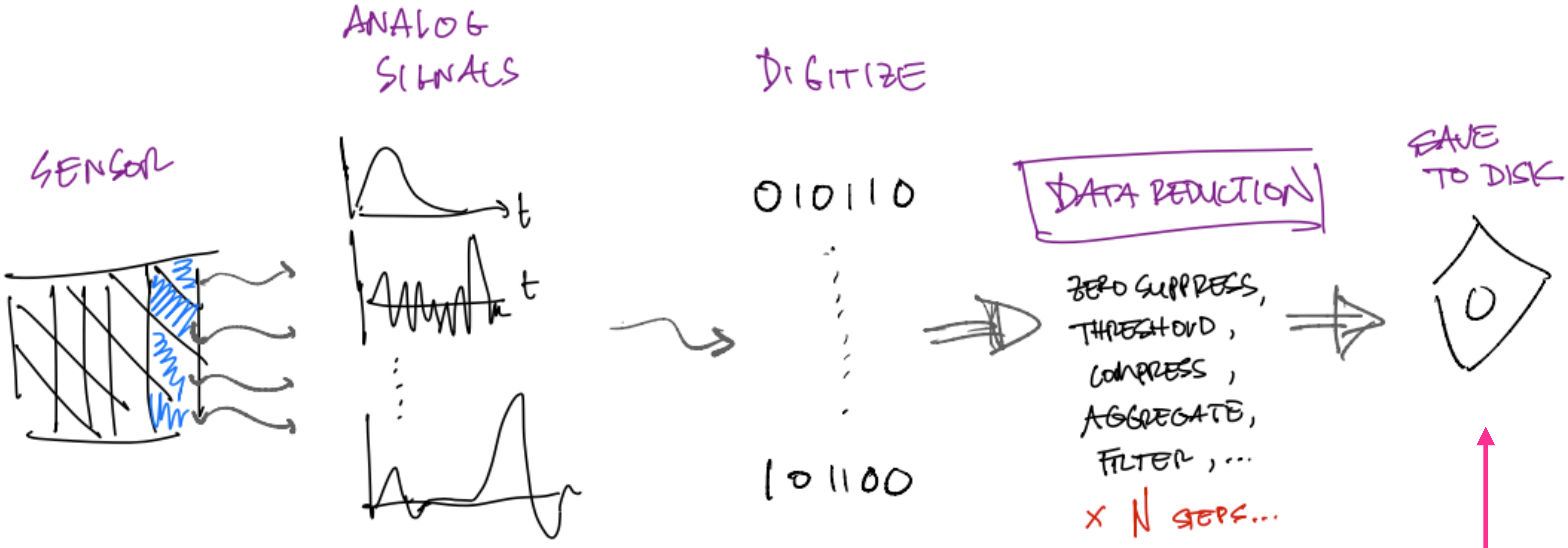


https://arxiv.org/abs/2207.07958


ANALOG SIGNALS
DIGITIZE
SENSOR
SAVE TO DISK
DATA REDUCTION
t
t
010110
101100
ZERO SUPPRESS,
THRESHOLD,
COMPRESS,
AGGREGATE,
FILTER, ...
x N STEPS...

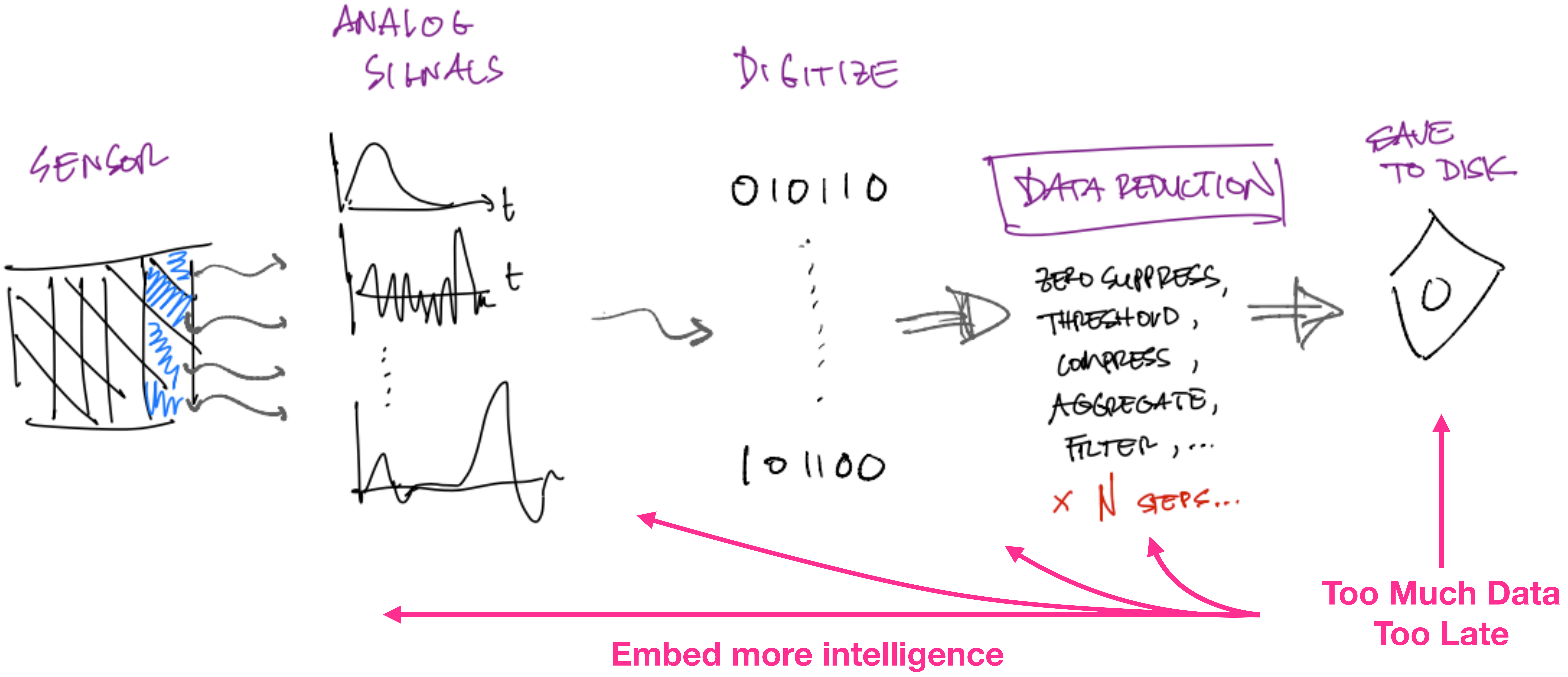


1 channel ~ 10b
1 channel, 1 MHz rate ~ 10 Mb/s
100k channels, 1 MHz rate ~ 1 Tb/s



Too Much Data
Too Late

1 channel ~ 10b
1 channel, 1 MHz rate ~ 10 Mb/s
100k channels, 1 MHz rate ~ 1 Tb/s


ANALOG
SIGNALS
DIGITIZE
SENSOR
t
t
010110
101100
DATA REDUCTION
SAVE
TO DISK
ZERO SUPPRESS,
THRESHOLD,
COMPRESS,
AGGREGATE,
FILTER, ...
x N STEPS...
Too Much Data
Too Late
Embed more intelligence

Applications in nuclear physics and beyond

https://indico.cern.ch/e/fml2022

A workshop dedicated to real-time applications of ML across the sciences

See also:

Applications and Technique in Fast Machine Learning for Science

https://www.frontiersin.org/articles/10.3389/fdata.2022.787421/full

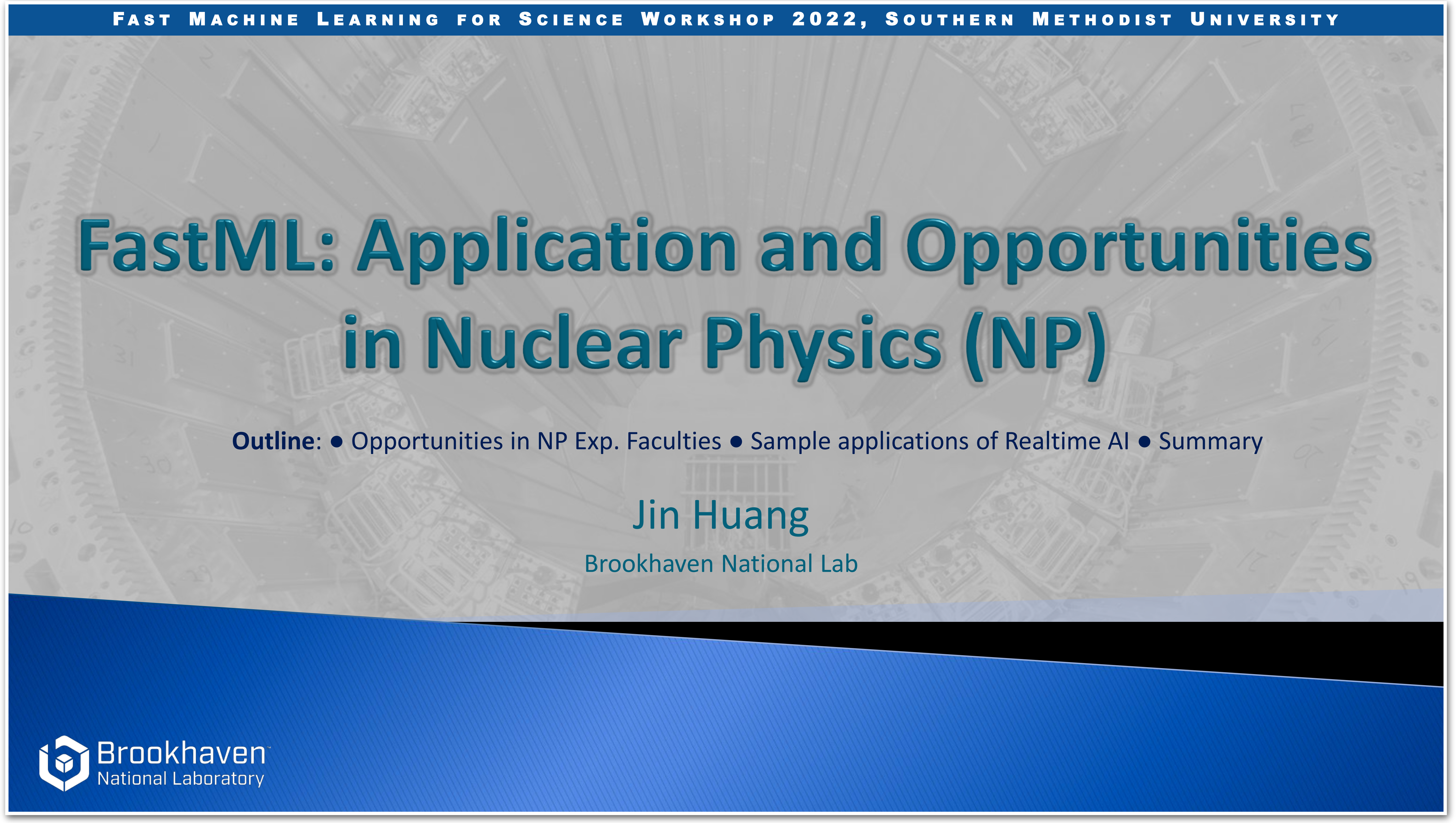
Applications in nuclear physics and beyond

https://indico.cern.ch/e/fml2022

A workshop dedicated to real-time applications of ML across the sciences

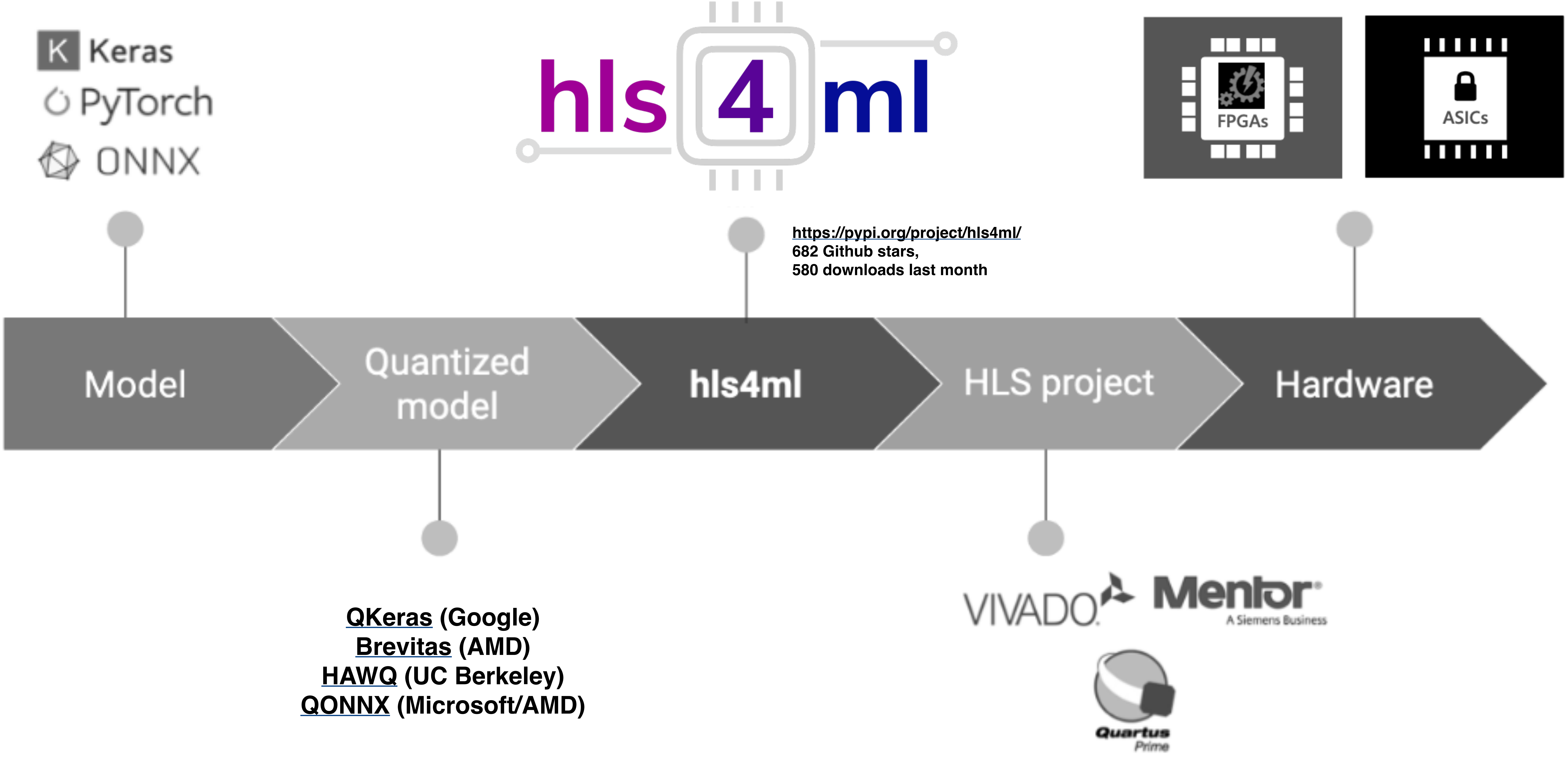
See also:
Applications and Technique in Fast Machine Learning for Science
https://www.frontiersin.org/articles/10.3389/fdata.2022.787421/full

See Jin Huang's talk for a great overview of exciting real-time applications



I will be referencing other talks from the Fast ML workshop to point to other examples of state-of-the-art studies

State-of-the-art



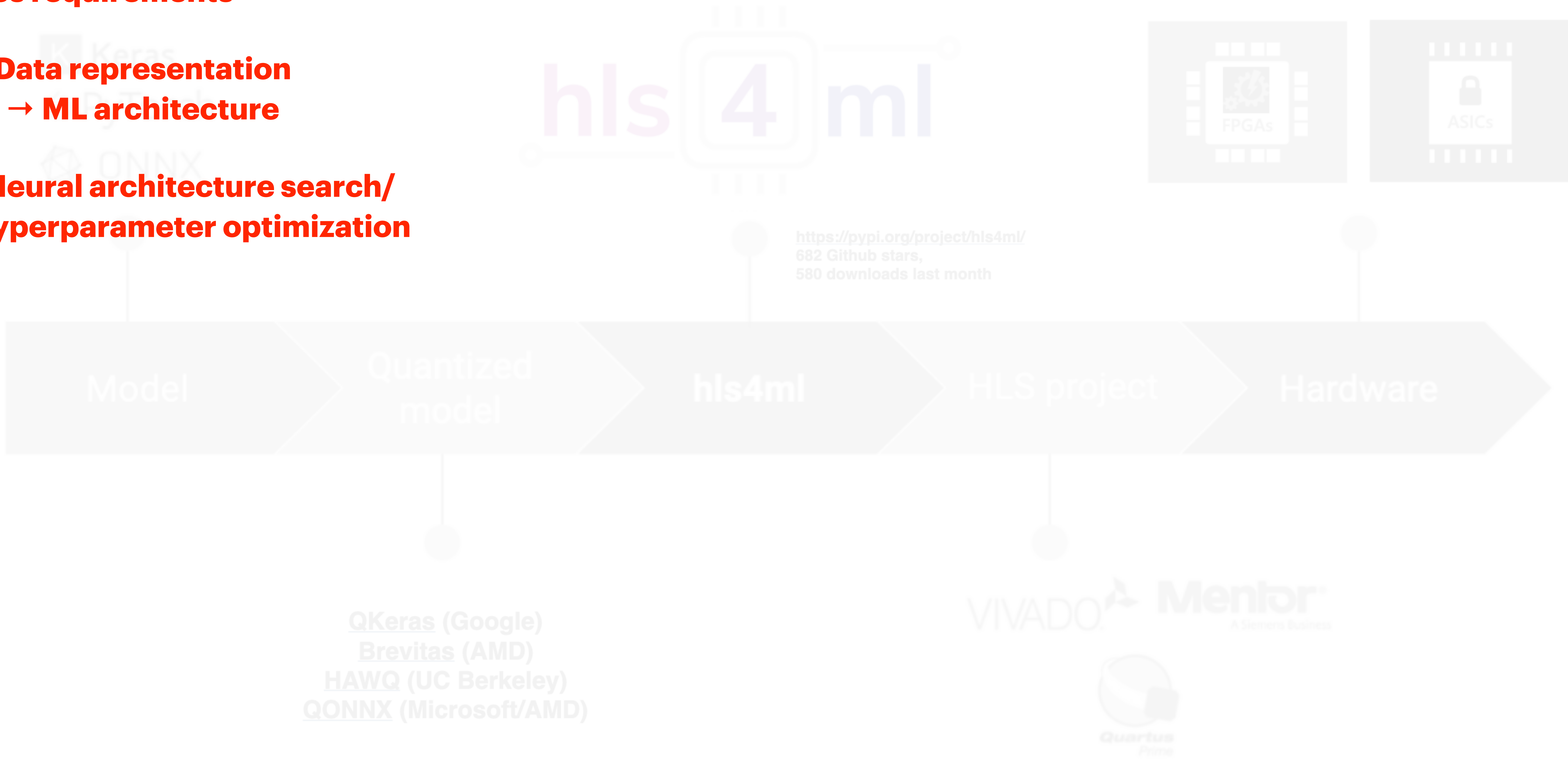
State-of-the-art

State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization



State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

What kind of platform?

Latency?
Pipeline Interval?

How many resources?

Area/power?
Radiation?
Cryo?

State-of-the-art

Physics requirements

Data representation
→ ML architecture

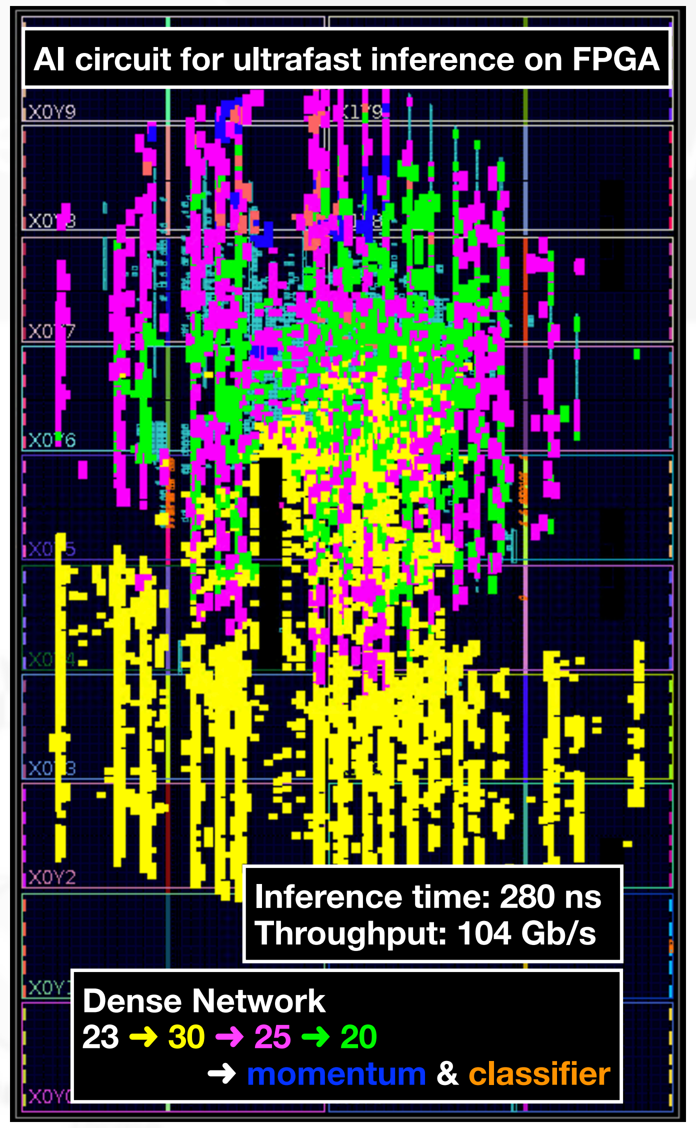
Neural architecture search/
Hyperparameter optimization

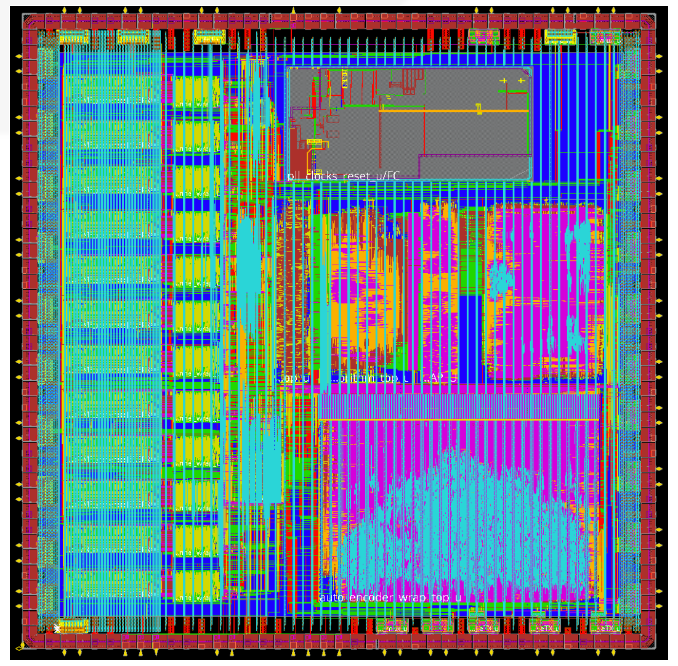
What kind of platform?

Latency?
Pipeline Interval?

How many resources?

Area/power?
Radiation?
Cryo?





State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

Quantize network

What kind of platform?

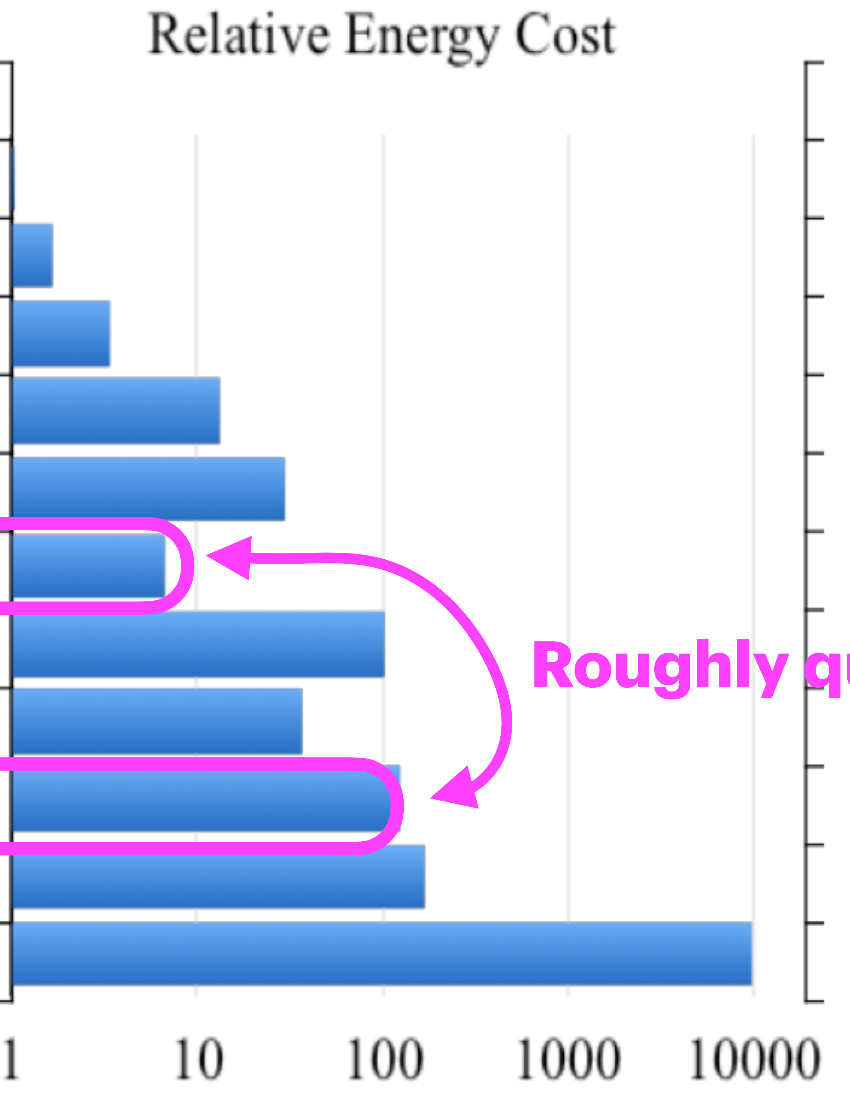
Latency?
Pipeline Interval?

How many resources?

Area/power?
Radiation?
Cryo?

See tools like:
QKeras
HAWQ
Brevitas

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



Roughly quadratic

Adapted from Horowitz

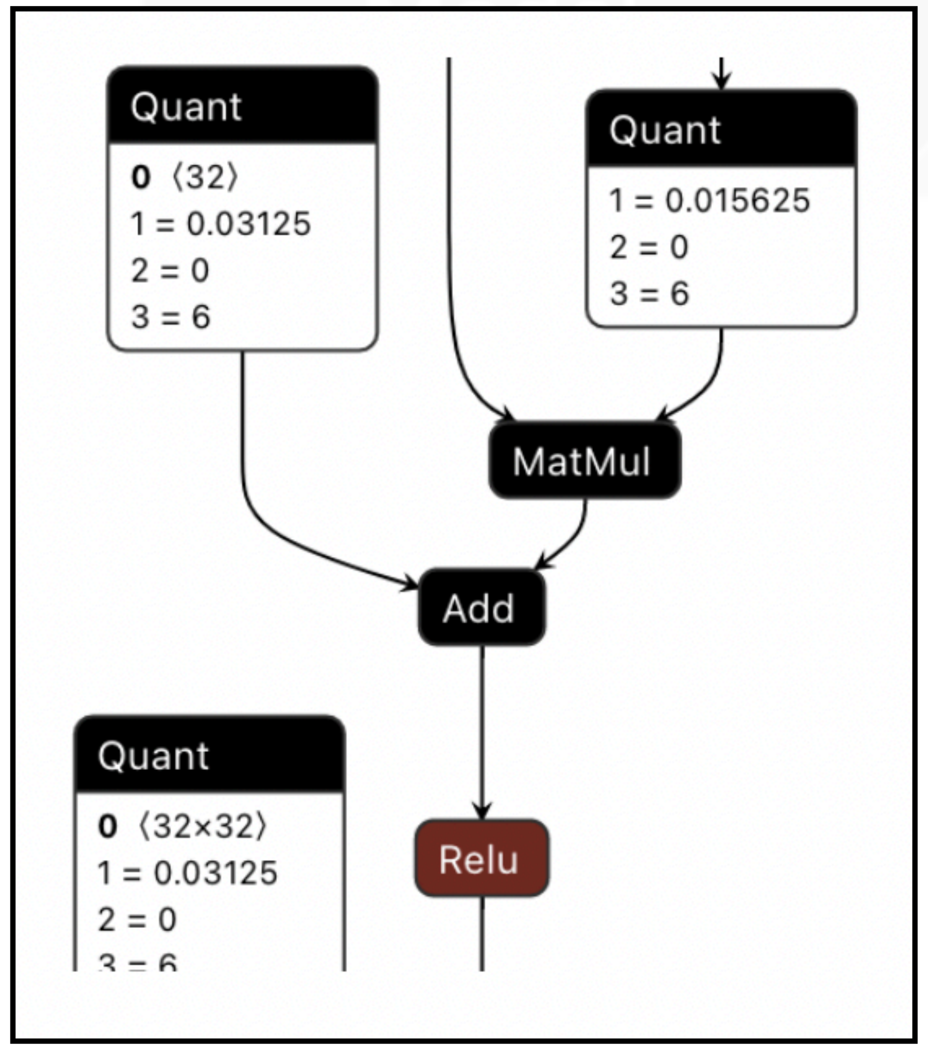
State-of-the-art

Physics requirements

Data representation → ML architecture

Neural architecture search/ Hyperparameter optimization

Quantize network



Intermediate (quantized) representations

See proposal for QONNX

What kind of platform?

Latency? Pipeline Interval?

How many resources?

Area/power? Radiation? Cryo?

State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

What kind of platform?

Latency?
Pipeline Interval?

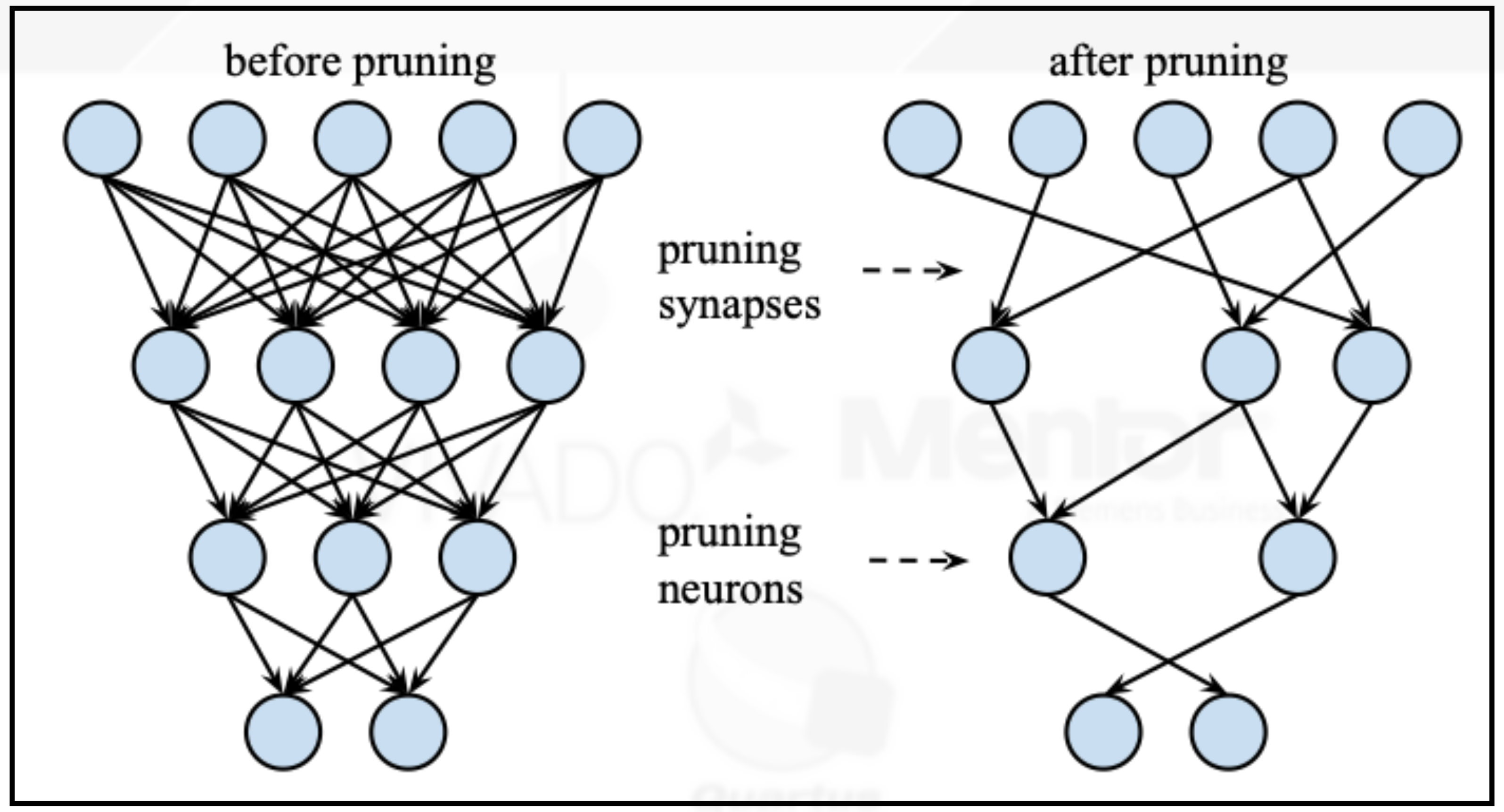
How many resources?

Area/power?
Radiation?
Cryo?

Quantize network

Intermediate (quantized) representations

Pruning/sparsity?



State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

Quantize network

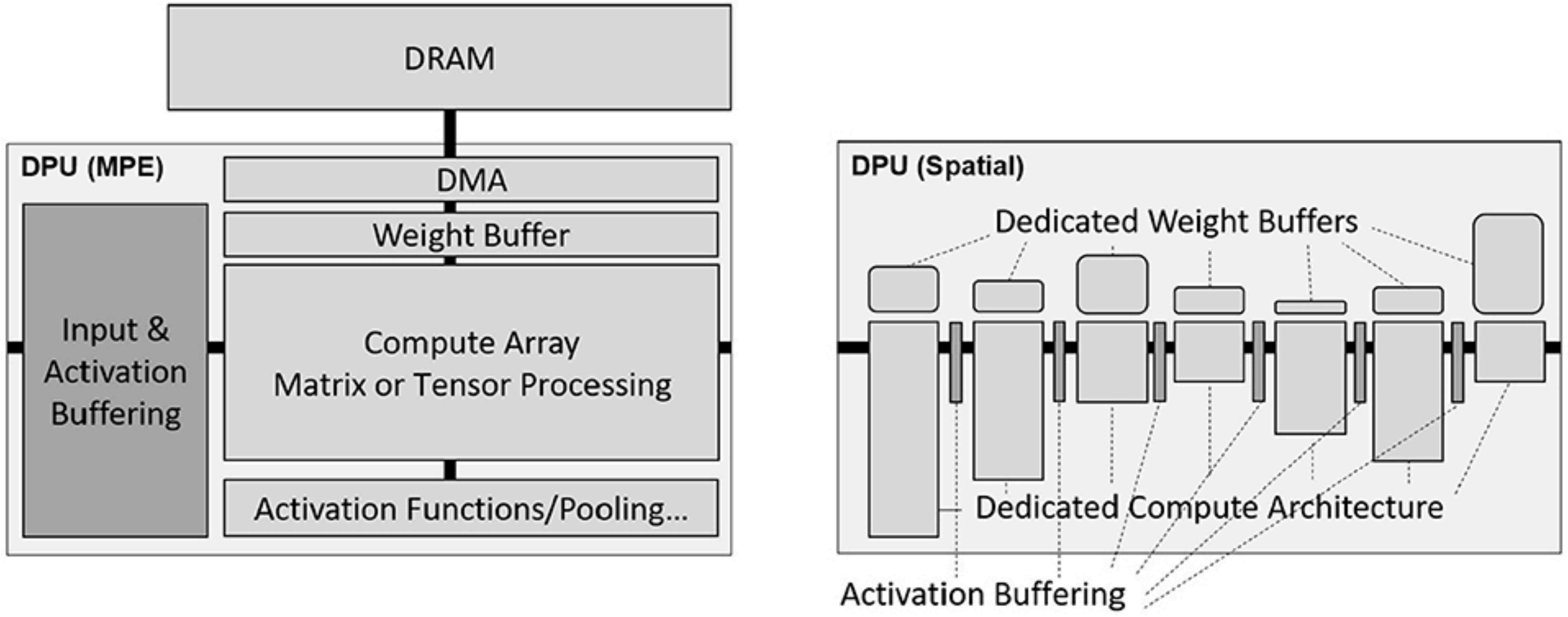
Intermediate (qu
representat

Pr

What kind of platform?

Latency?
Pipeline Interval?

Microarchitecture



Matrix of processing elements (Systolic Array)

Spatial Dataflow

State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

Quantize network

Intermediate (quantized) representations

Pruning/sparsity?

Microarchitecture

Parallelization

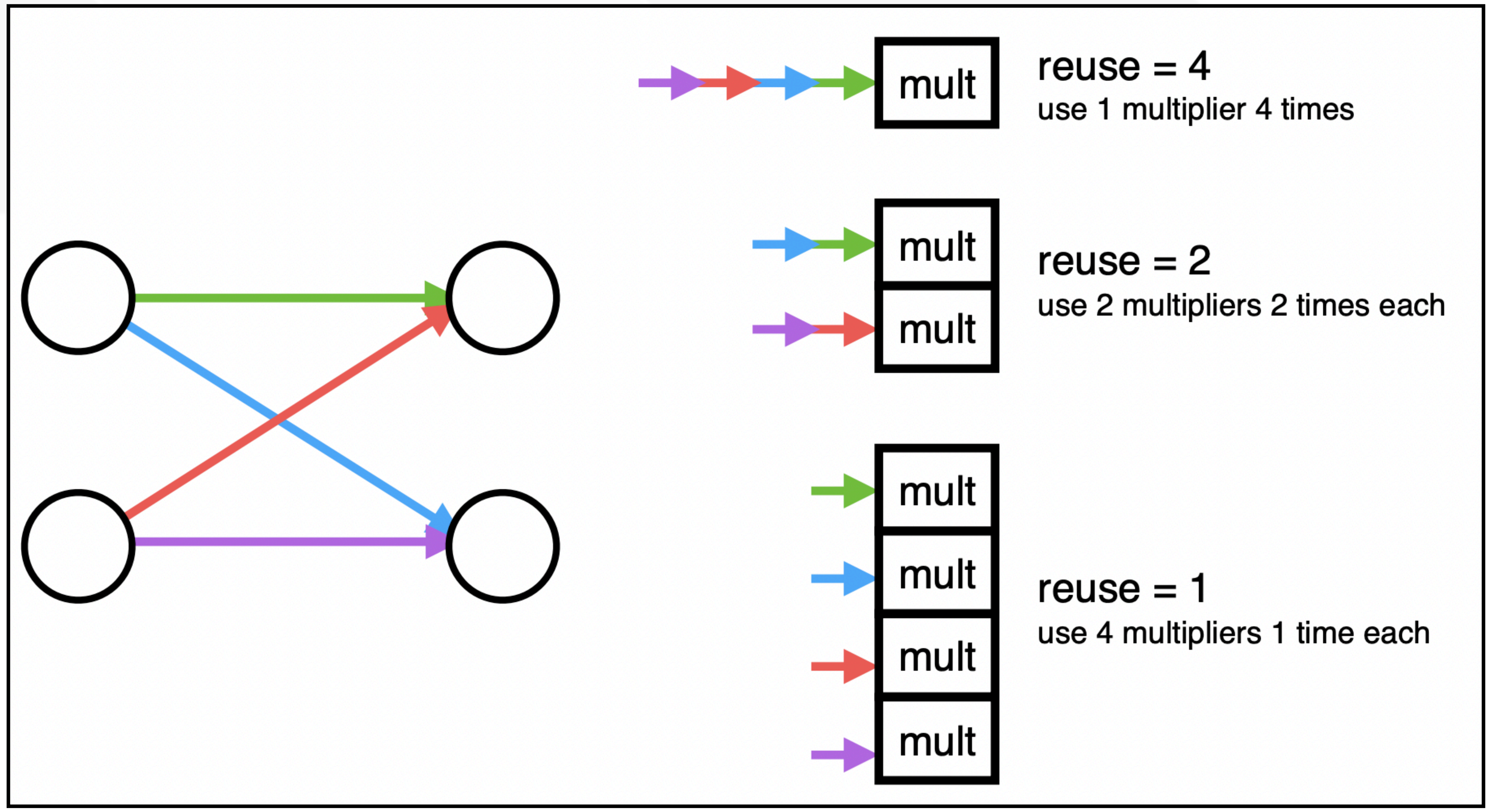
What kind of platform?

Latency?
Pipeline Interval?

How many resources?

Area/power?
Radiation?
Cryo?

QKeras (Google)
Brevitas (AMD)
HAWQ (UC Berkeley)
QONNX (Microsoft/AMD)



State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

Quantize network

Intermediate (quantized)
representations

Pruning/sparsity?

Microarchitecture

Parallelization

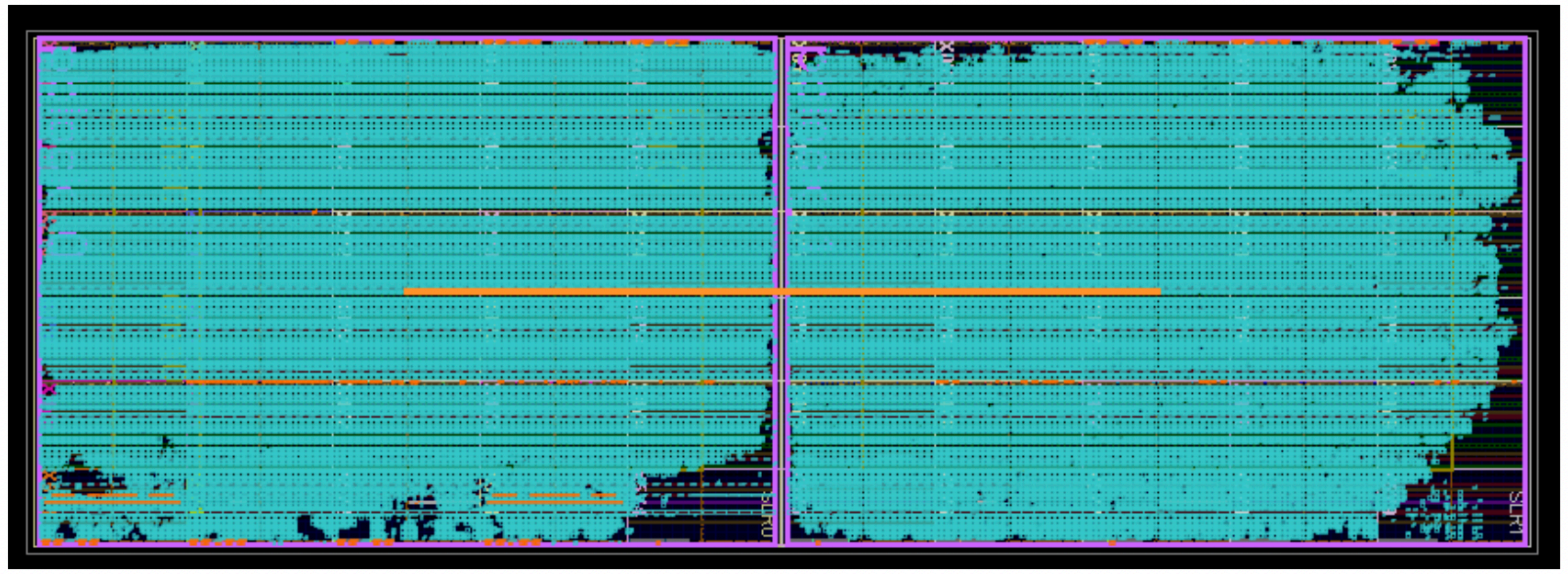
What kind of platform?

Latency?
Pipeline Interval?

How many
resources?

Area/power?
Radiation?
Cryo?

Synthesize, validate design,
satisfy design rules/timing



BNL711 FELIX Firmware Floorplanning

State-of-the-art

Physics requirements

Data representation
→ ML architecture

Neural architecture search/
Hyperparameter optimization

Quantize network

Intermediate (quantized)
representations

Pruning/sparsity?

Microarchitecture

Parallelization

What kind of platform?

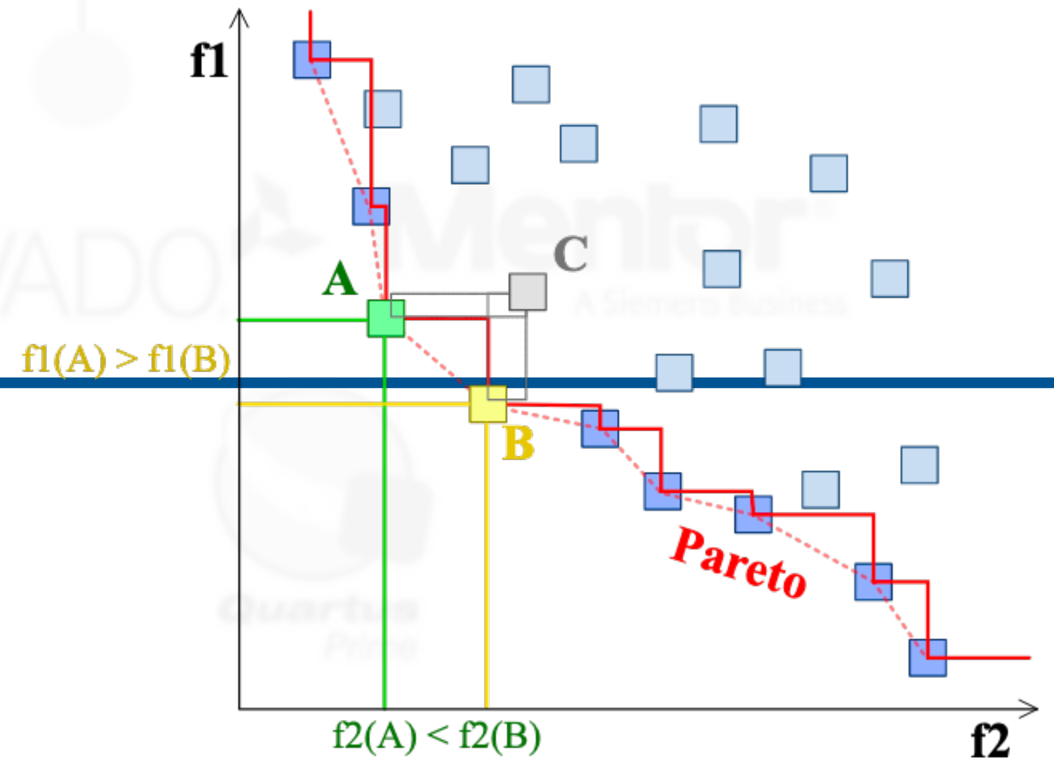
Latency?
Pipeline Interval?

How many
resources?

Area/power?
Radiation?
Cryo?

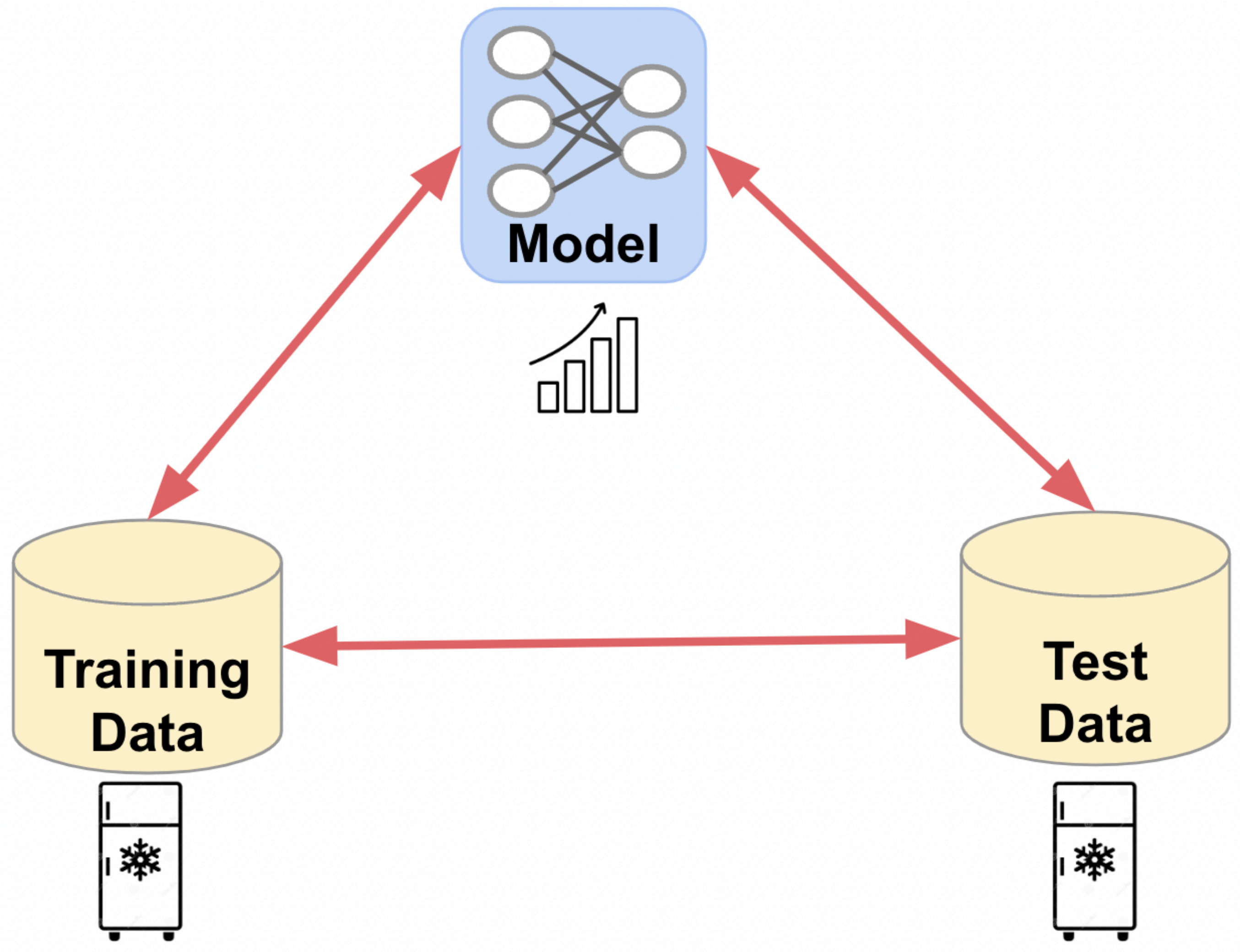
Synthesize, validate design,
satisfy design rules/timing

Multi-objective
design space optimization

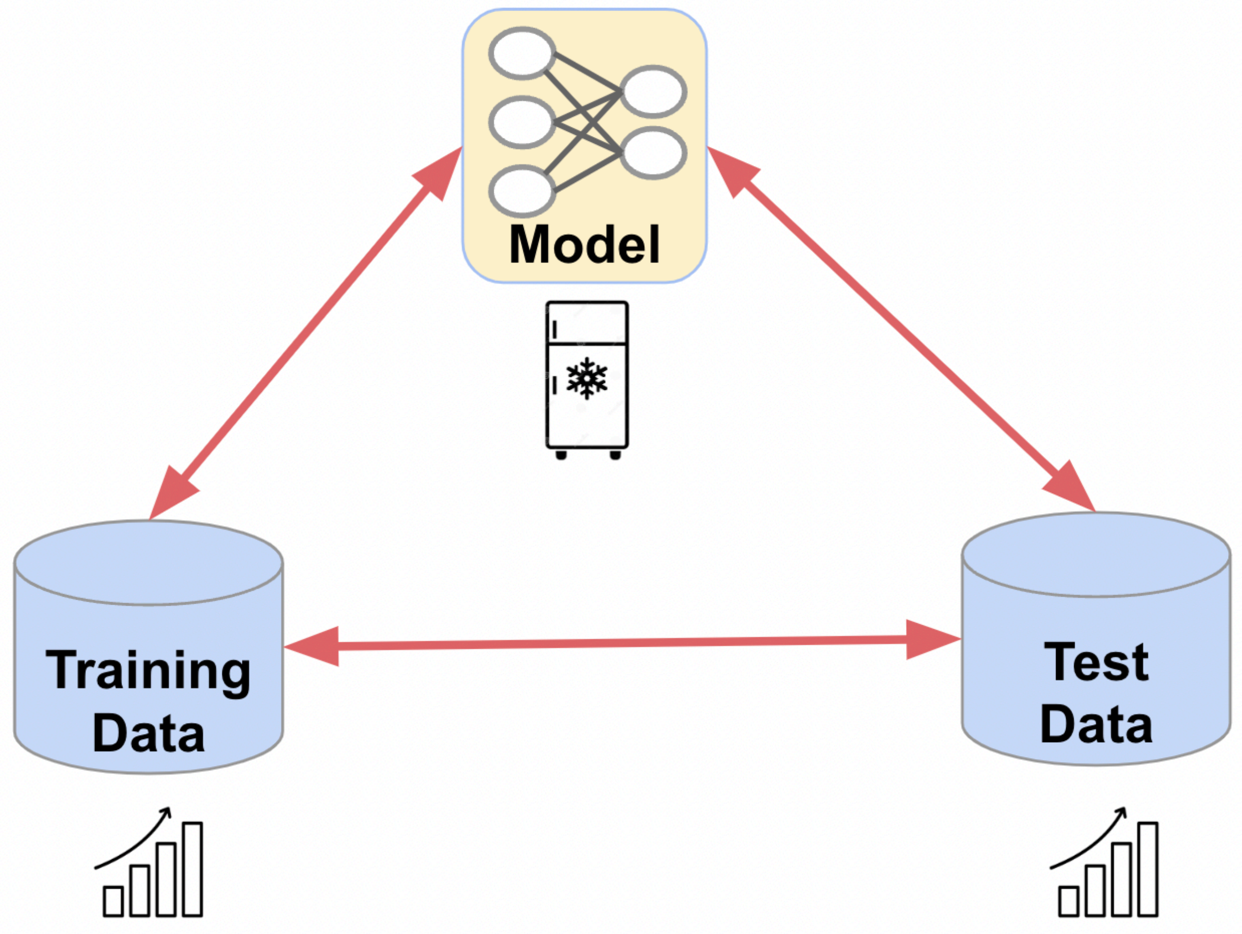


Towards sustainability and robustness

- Robustness and fault tolerance
  - DataPerf (V. Reddi)
  - FKeras (O. Weng)
  - Continual learning? (B. Radburn-Smith)
- Implementation within FW infrastructure, synthesize effectively
  - Issues, tricks, and tips (M. Rigatti, D. Hoang)
  - Emulating NNs in experimental SW?



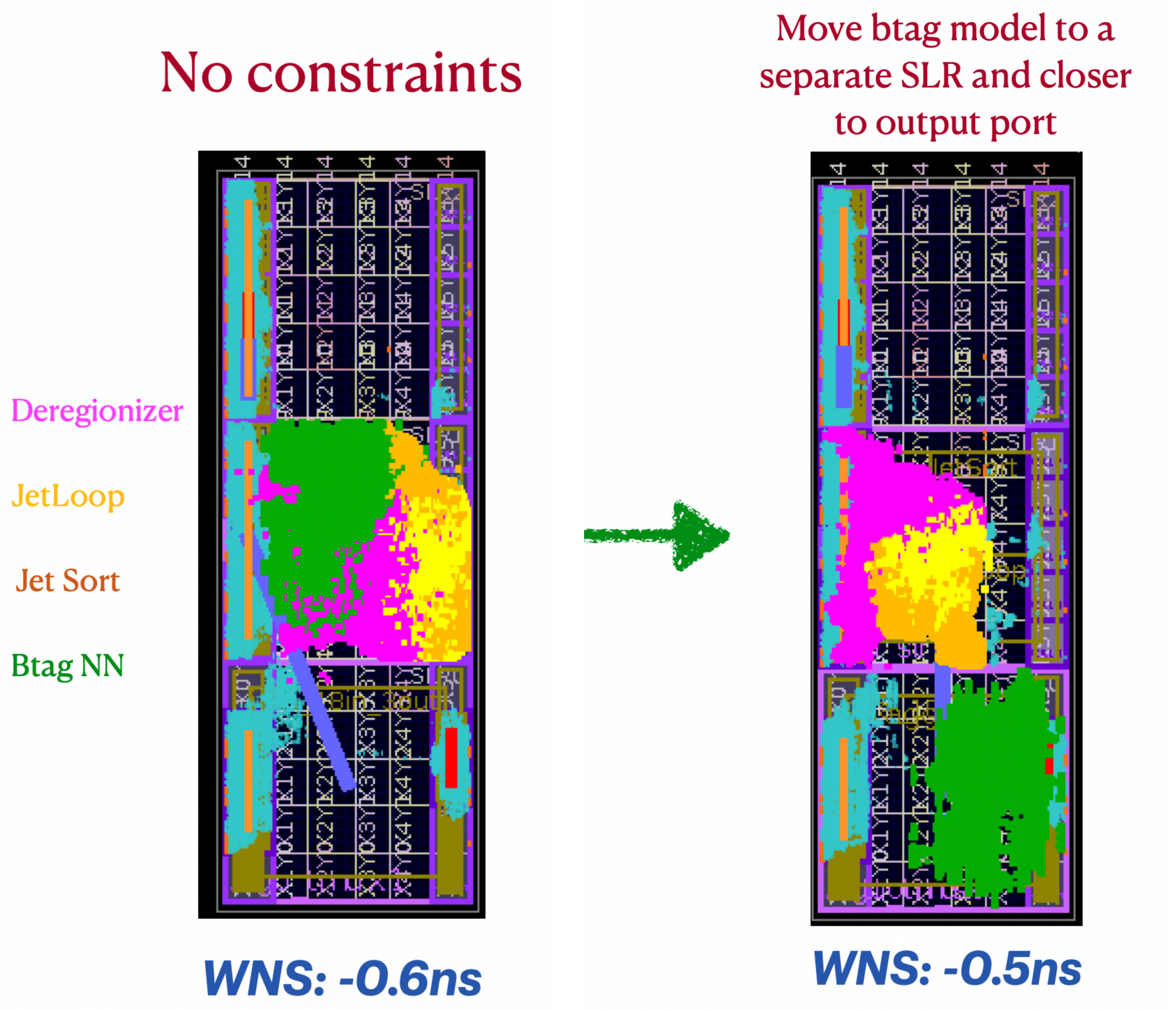
Model-centric paradigm



Data-centric paradigm

Towards sustainability and robustness

- Robustness and fault tolerance
  - DataPerf (V. Reddi)
  - FKeras (O. Weng)
  - Continual learning? (B. Radburn-Smith)
- Implementation within FW infrastructure, synthesize effectively
  - Issues, tricks, and tips (M. Rigatti, D. Hoang)
  - Emulating NNs in experimental SW?

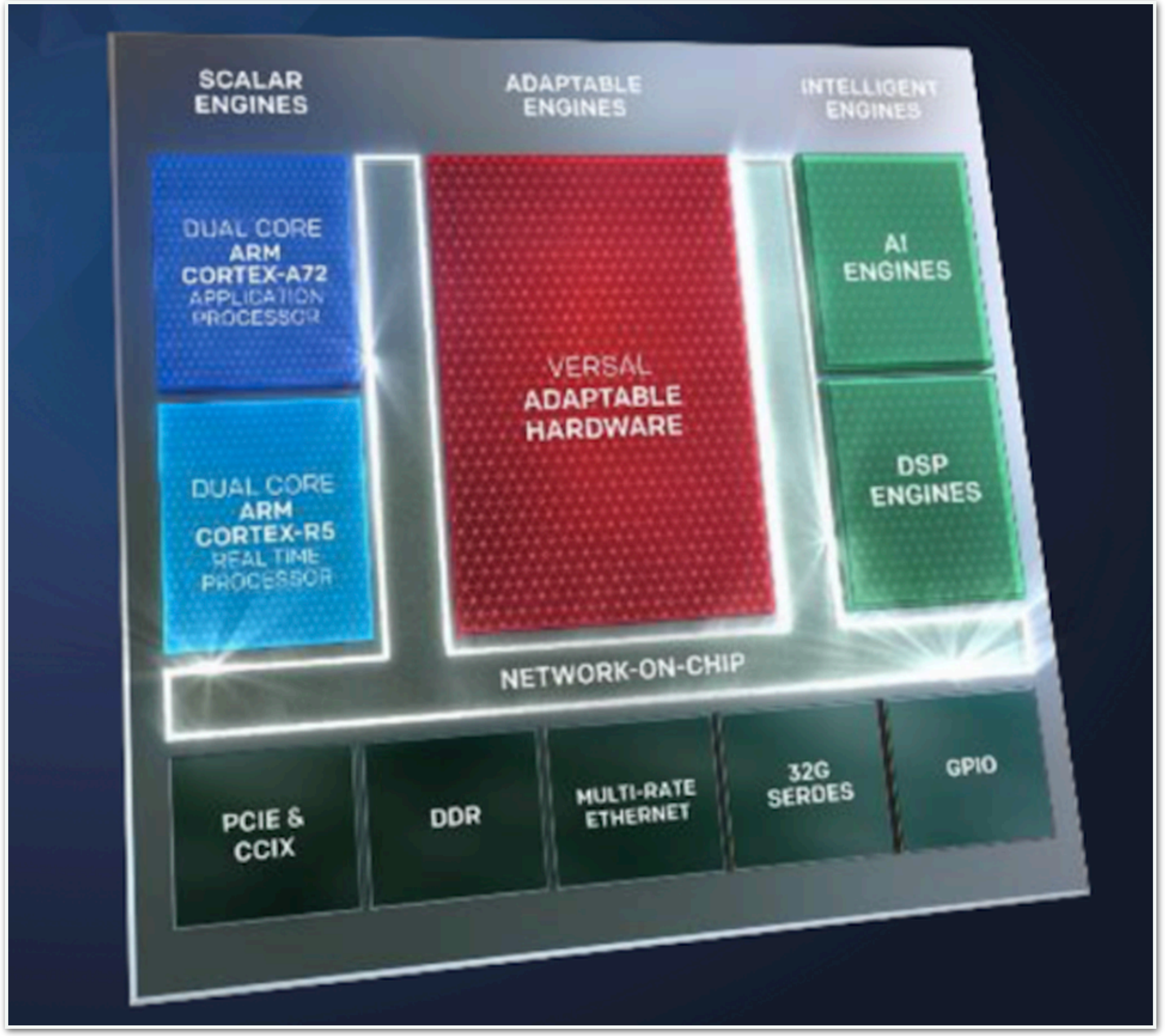


Emerging technologies

- Emerging computing architectures
- Emerging neural architectures
  - Spiking, inductive bias, physics-inspired,…
- Emerging microelectronics technologies

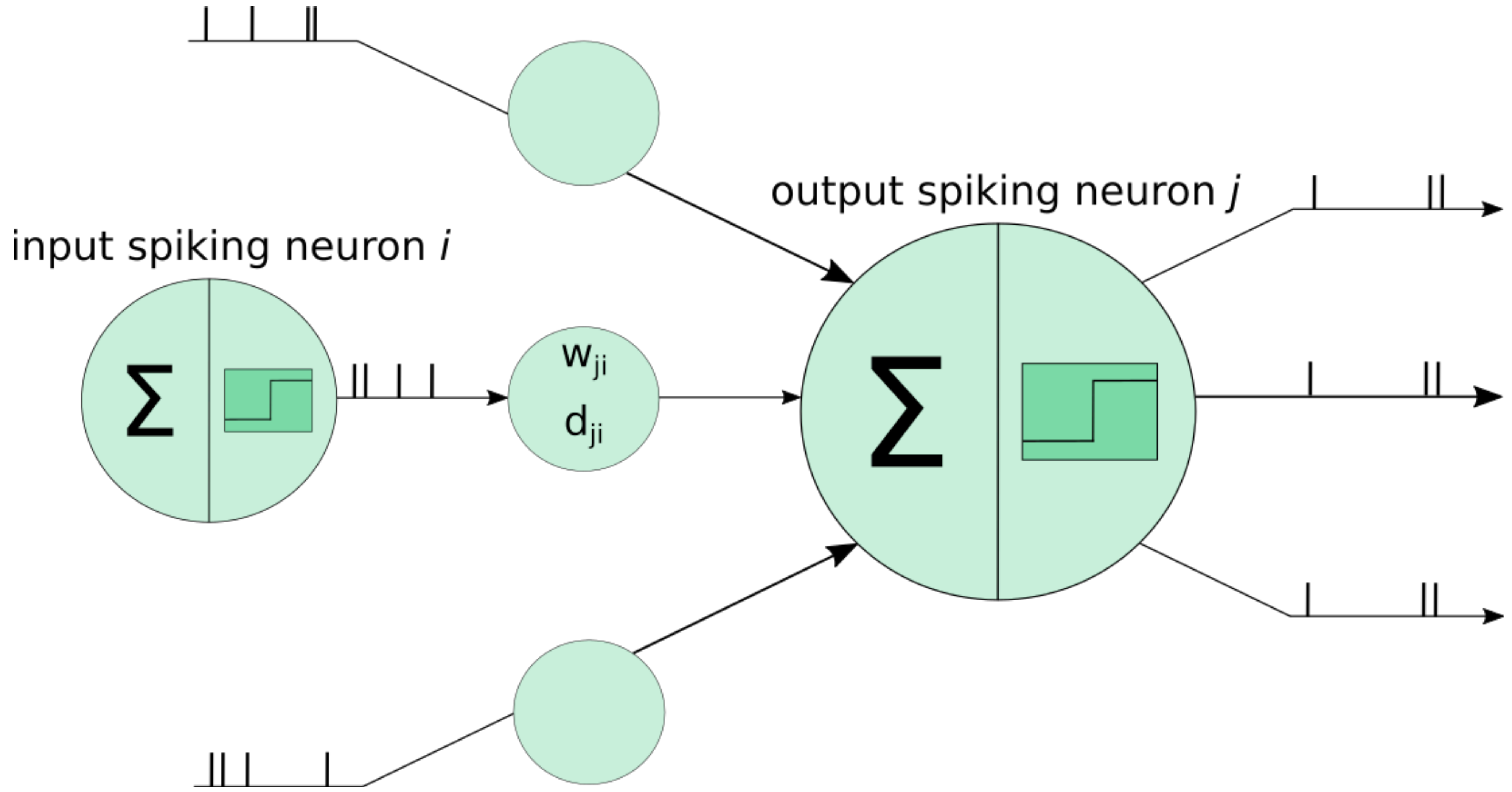
Emerging technologies

- Emerging computing architectures
- Emerging neural architectures
  - Spiking, inductive bias, physics-inspired,...
- Emerging microelectronics technologies



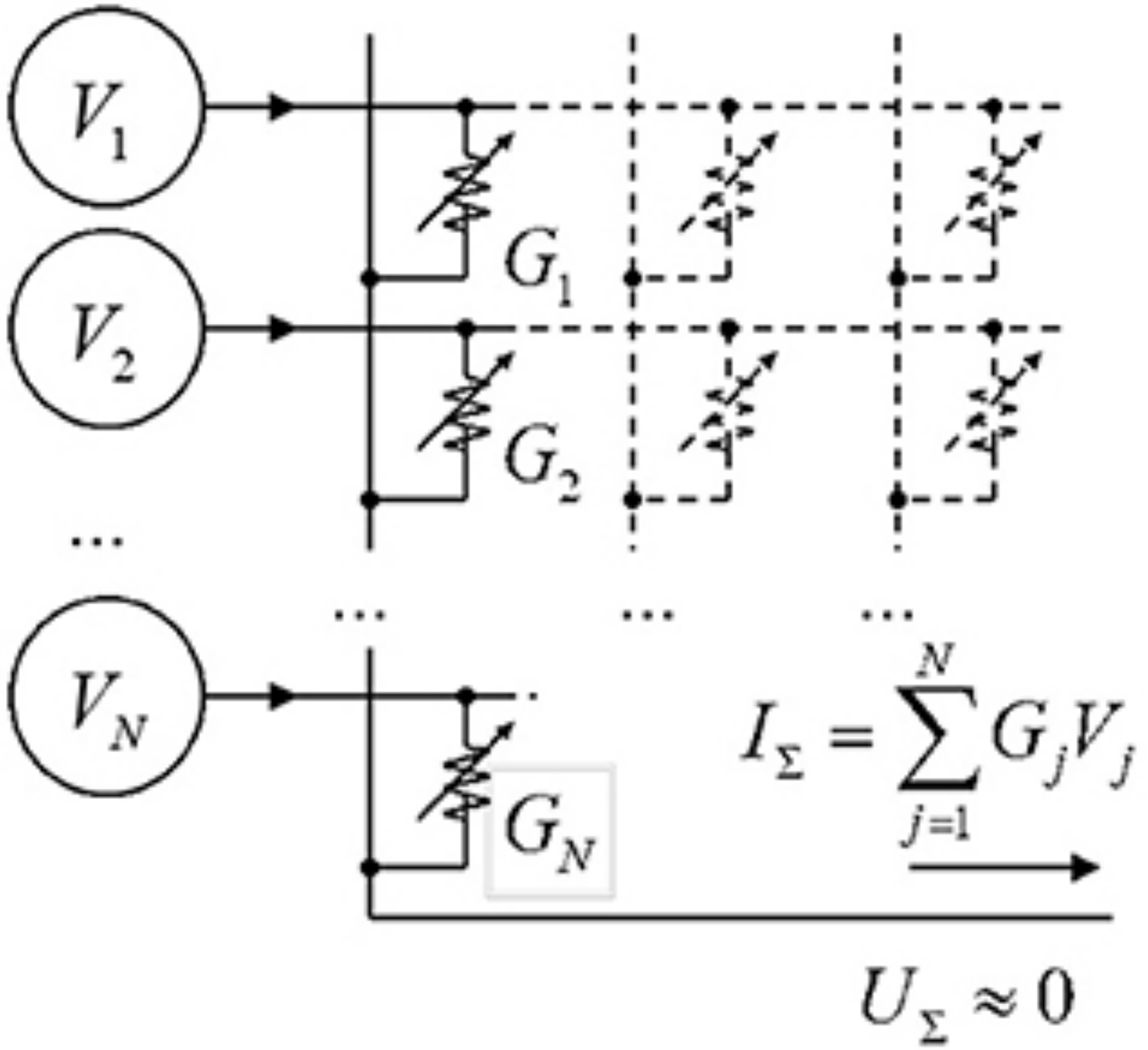
Emerging technologies

- Emerging computing architectures
- Emerging neural architectures
  - Spiking, inductive bias, physics-inspired,…
- Emerging microelectronics technologies



Emerging technologies

- Emerging computing architectures
- Emerging neural architectures
  - Spiking, inductive bias, physics-inspired,...
- Emerging microelectronics technologies



Summary

- A whirlwind tour through elements of developing embedded real-time ML!
- With hls4ml we try to make cutting edge techniques accessible to non-experts; open-source tools for scientific applications
  - https://github.com/fastmachinelearning/hls4ml-tutorial
- Powerful techniques exist
  - But there is still plenty of exciting research to do — ML techniques, computing architectures, microelectronics technologies

Backup