FERMILAB-SLIDES-24-0096-CSAID

Fermilab | U.S. DEPARTMENT OF ENERGY | Office of Science

Accelerating Scientific Discoveries with Real-Time Intelligent Sensing

Nhan Tran, Fermi National Accelerator Laboratory

SCSP AI Expo

7 May 2024

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

ai expo 24 for national competitiveness

Outline

- Why Fast ML for Science?
- The intelligent edge of tomorrow
- Outlook

Outline

- Why Fast ML for Science?
- The intelligent edge of tomorrow
- Towards ultra-fast automated experimentation

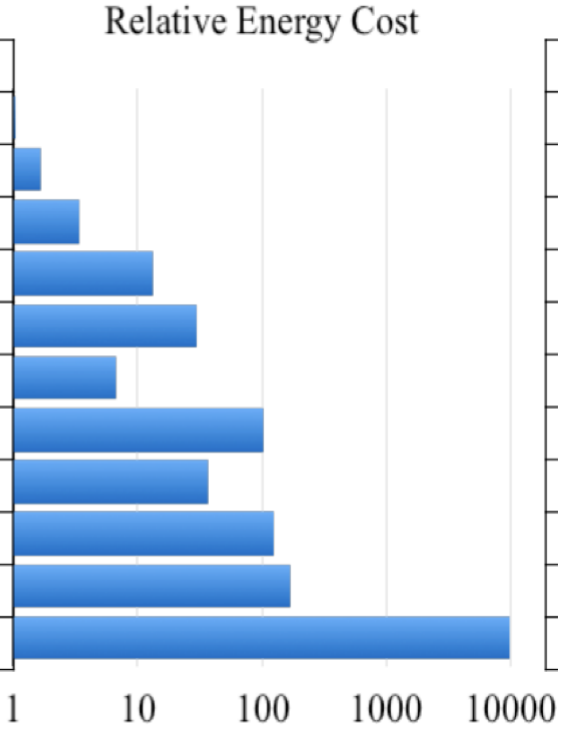
"Scientific discoveries come from groundbreaking ideas and the capability to validate those ideas by testing nature at new scales - finer and more precise temporal and spatial resolution. This is leading to an explosion of data that must be interpreted, and ML is proving a powerful approach. The more efficiently we can test our hypotheses, the faster we can achieve discovery. To fully unleash the power of ML and accelerate discoveries, it is necessary to embed it into our scientific process, into our instruments and detectors."

Applications and Techniques for Fast Machine Learning in Science

https://doi.org/10.3389/fdata.2022.787421

Moving 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

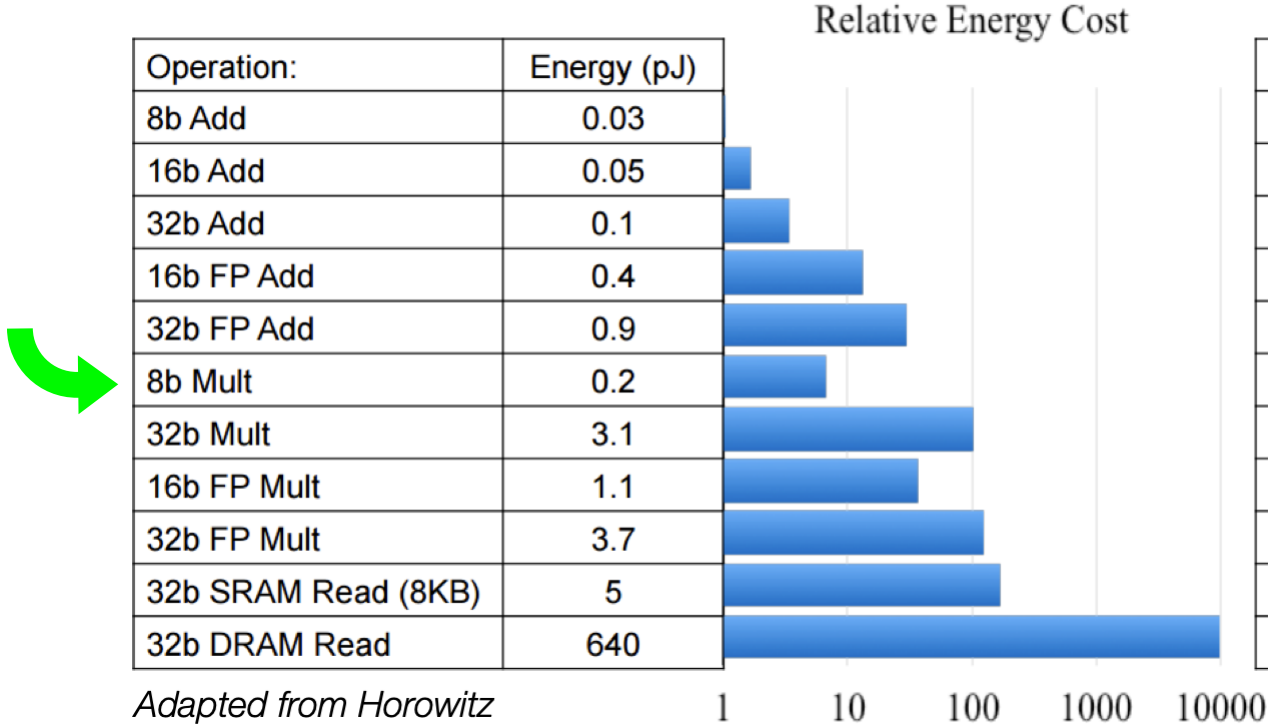


Adapted from Horowitz

Fermilab

Moving 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

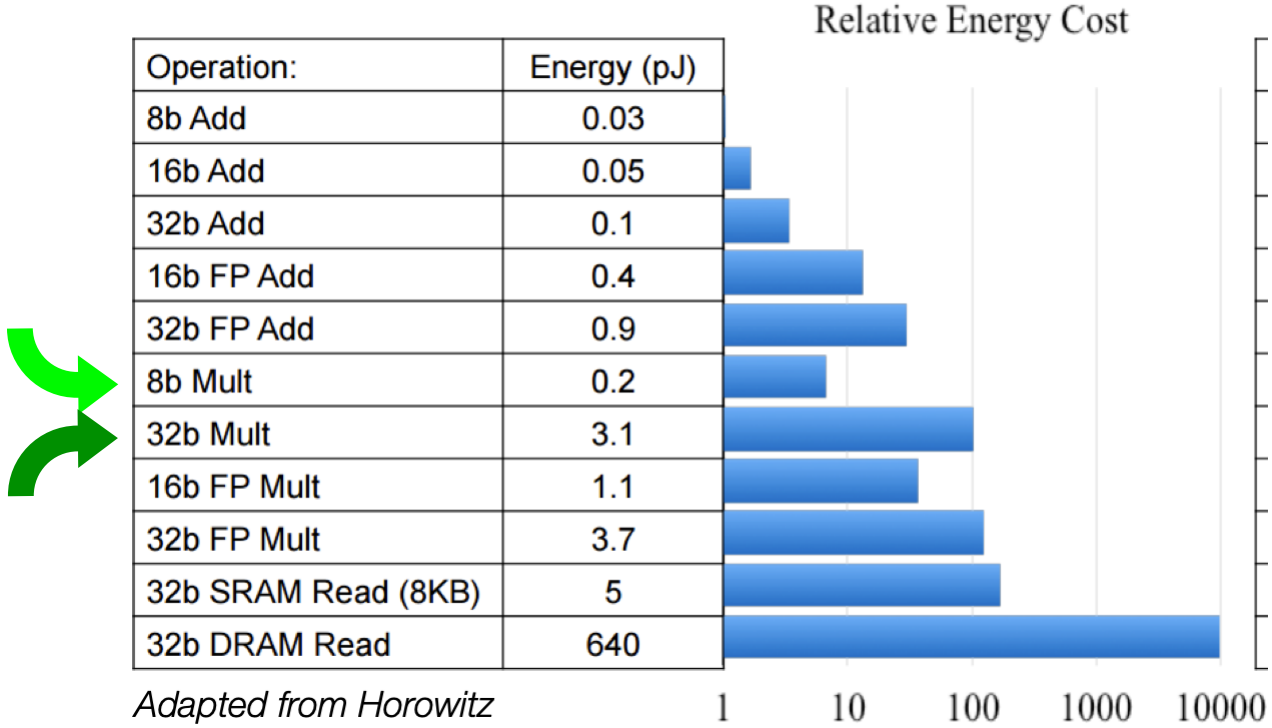


Adapted from Horowitz

Fermilab

Moving 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

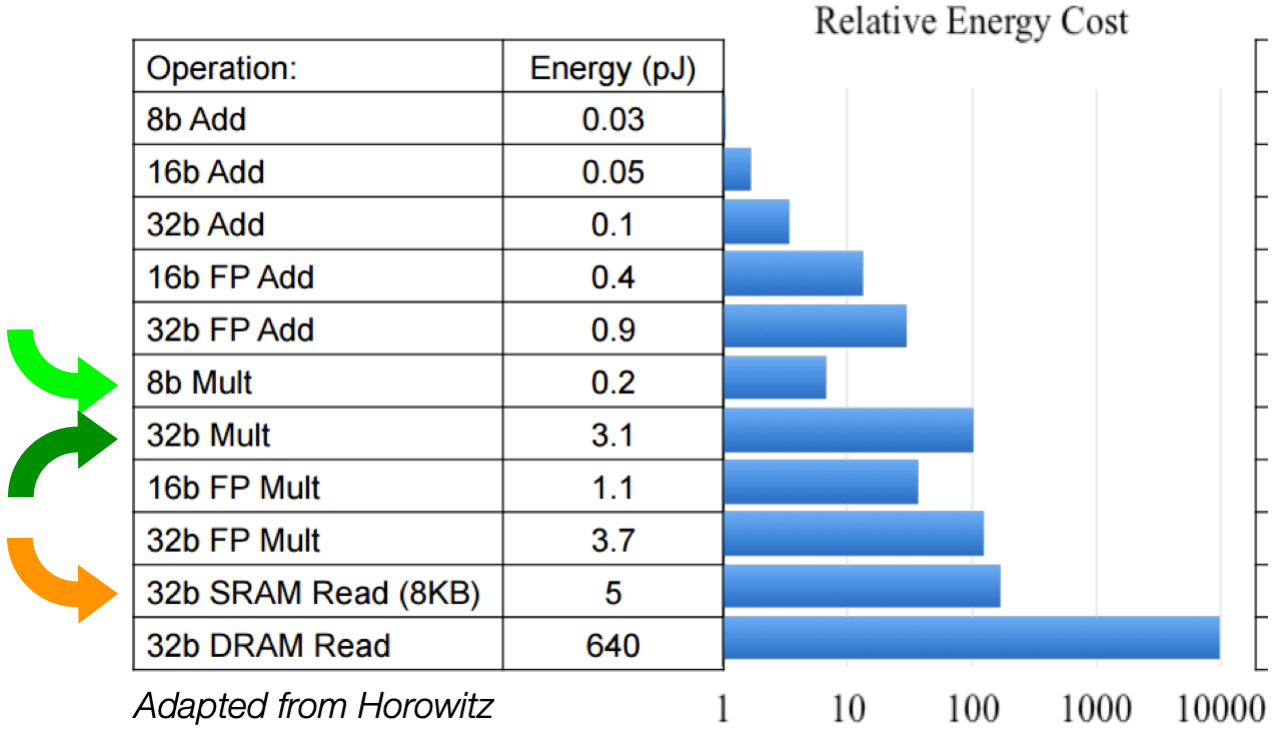


Adapted from Horowitz

Fermilab

Moving 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

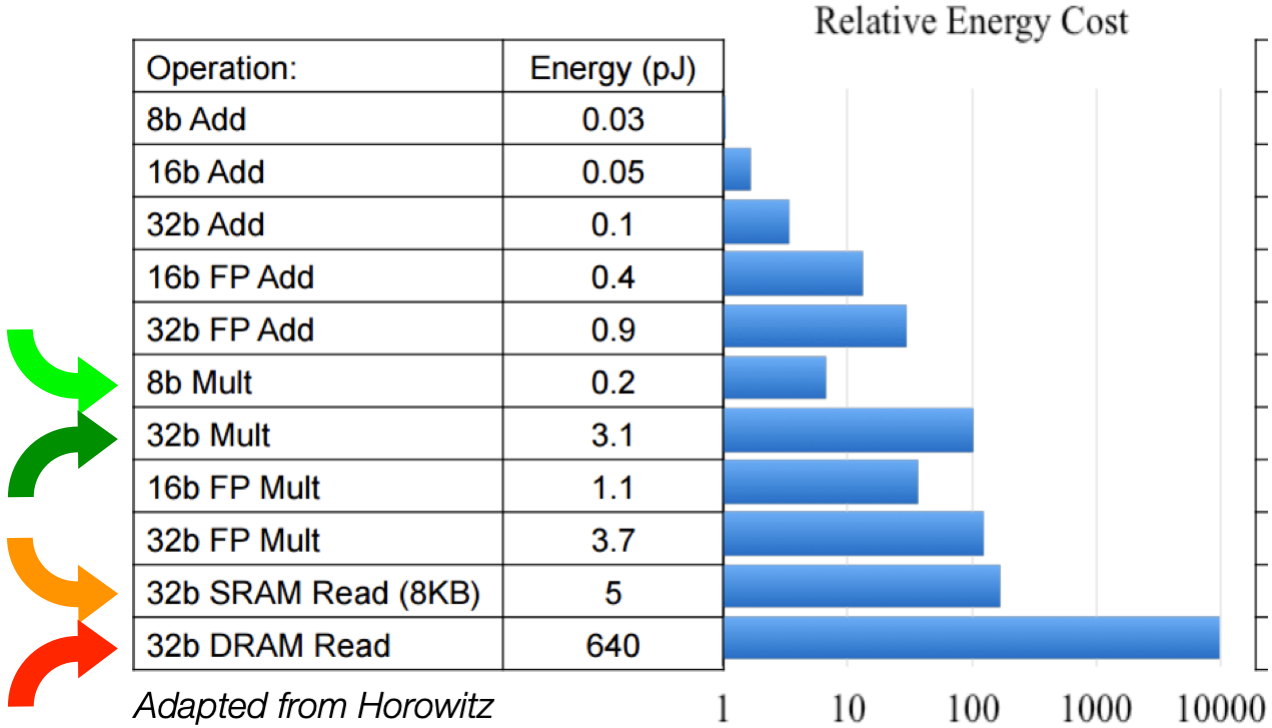


Adapted from Horowitz

Fermilab

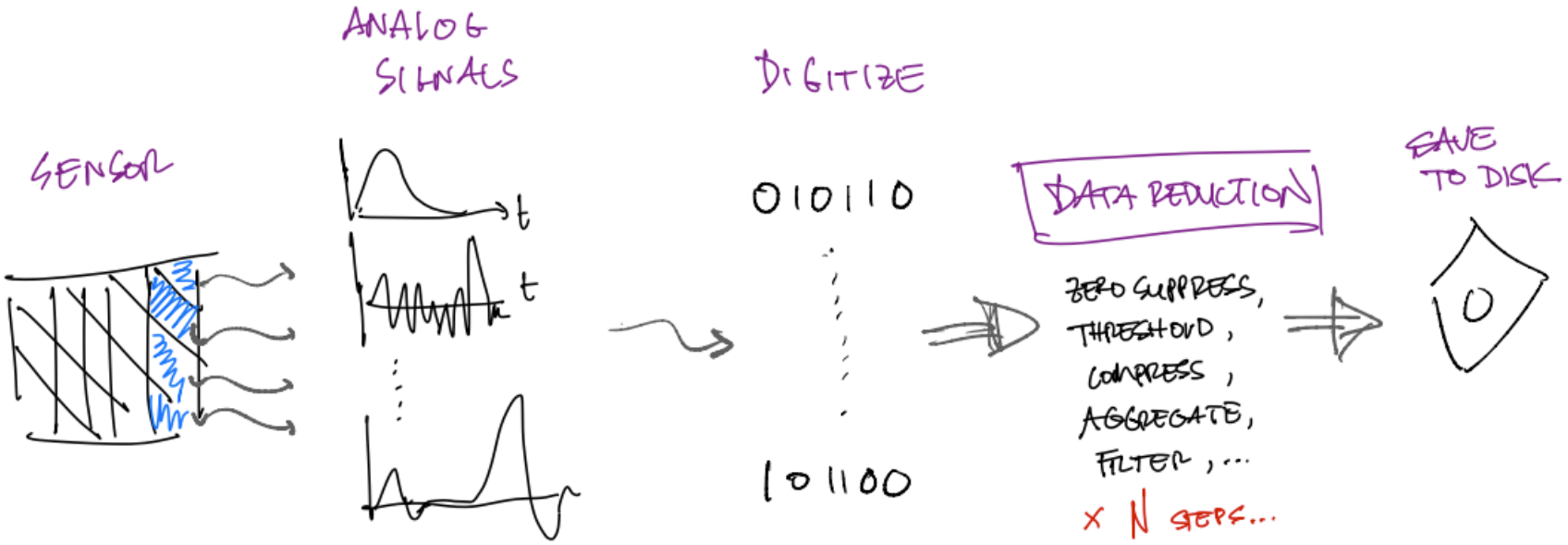
Moving data expensive, compute cheap

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

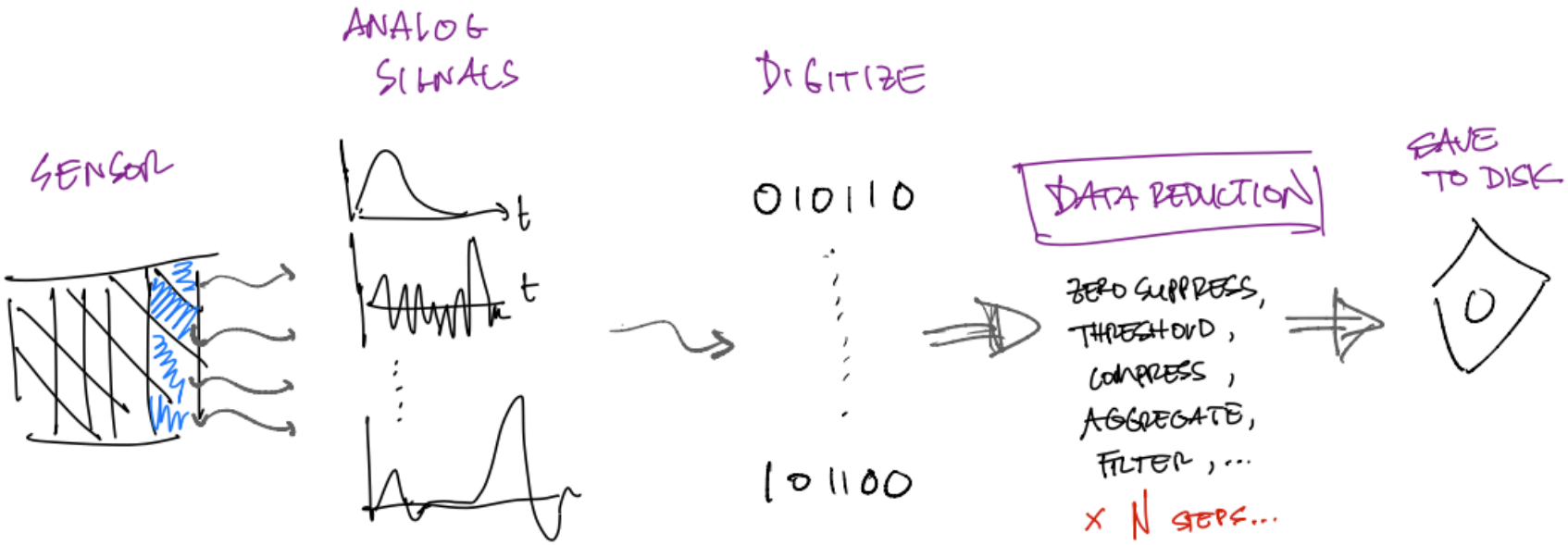


Adapted from Horowitz

Fermilab

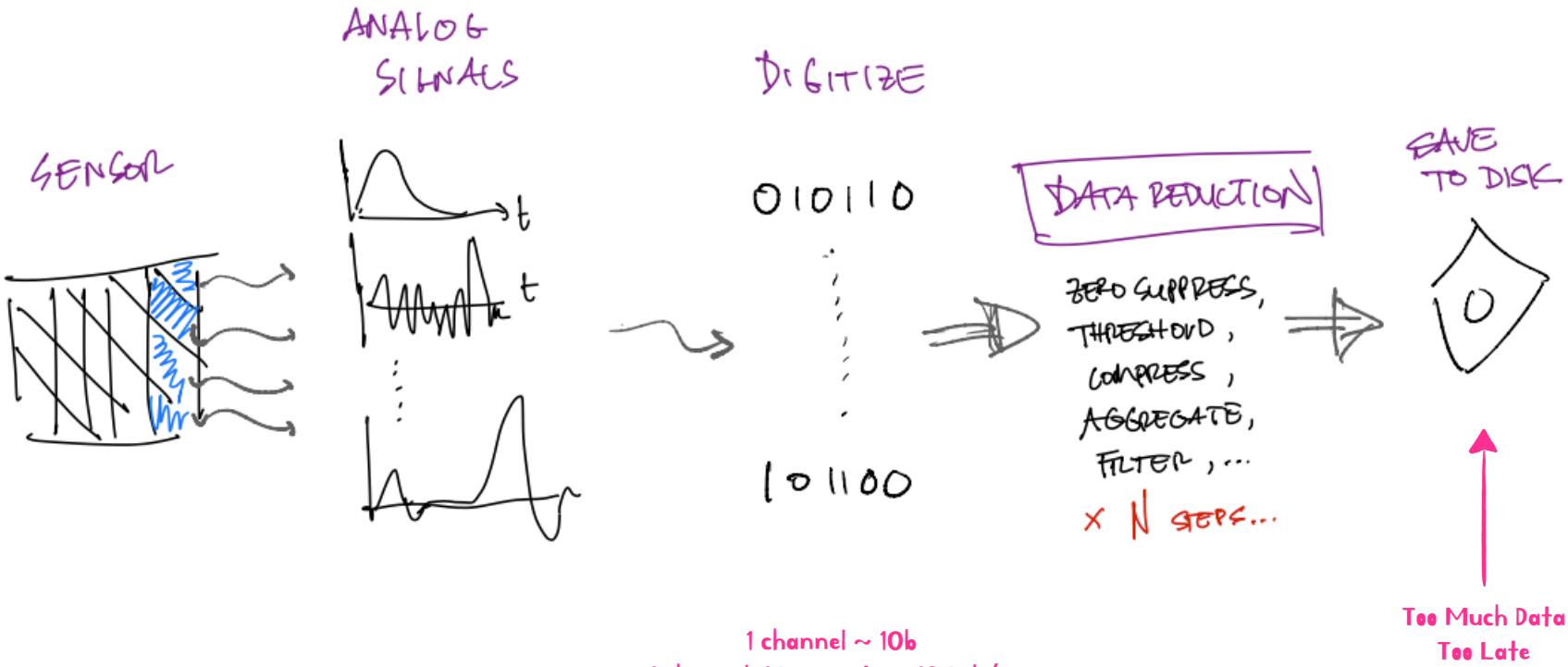

SENSOR
ANALOG SIGNALS
t
t
DIGITIZE
010110
101100
DATA REDUCTION
ZERO SUPPRESS,
THRESHOLD,
COMPRESS,
AGGREGATE,
FILTER, ...
x N STEPS...
SAVE TO DISK

Fermilab

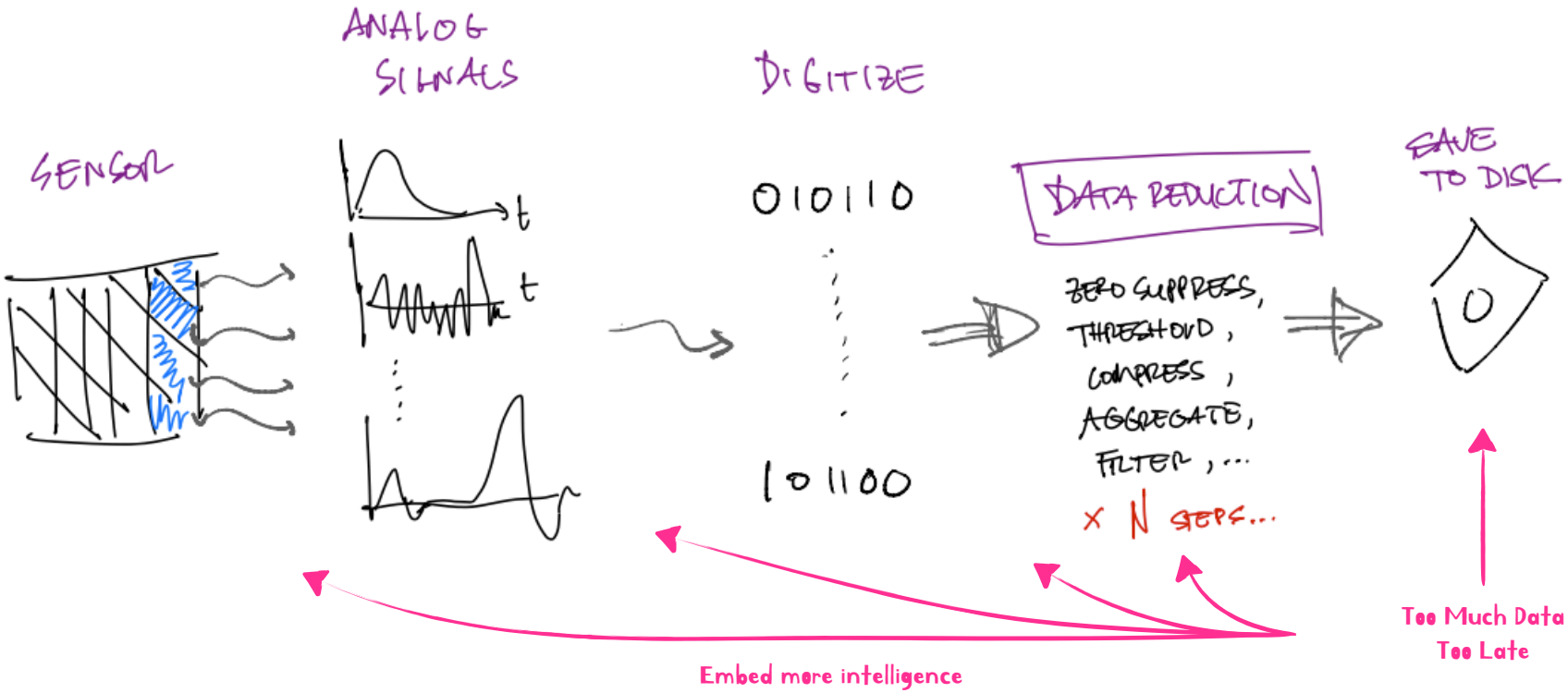


1 channel ~ 10b
1 channel, 1 MHz rate ~ 10 Mb/s
100k channels, 1 MHz rate ~ 1 Tb/s

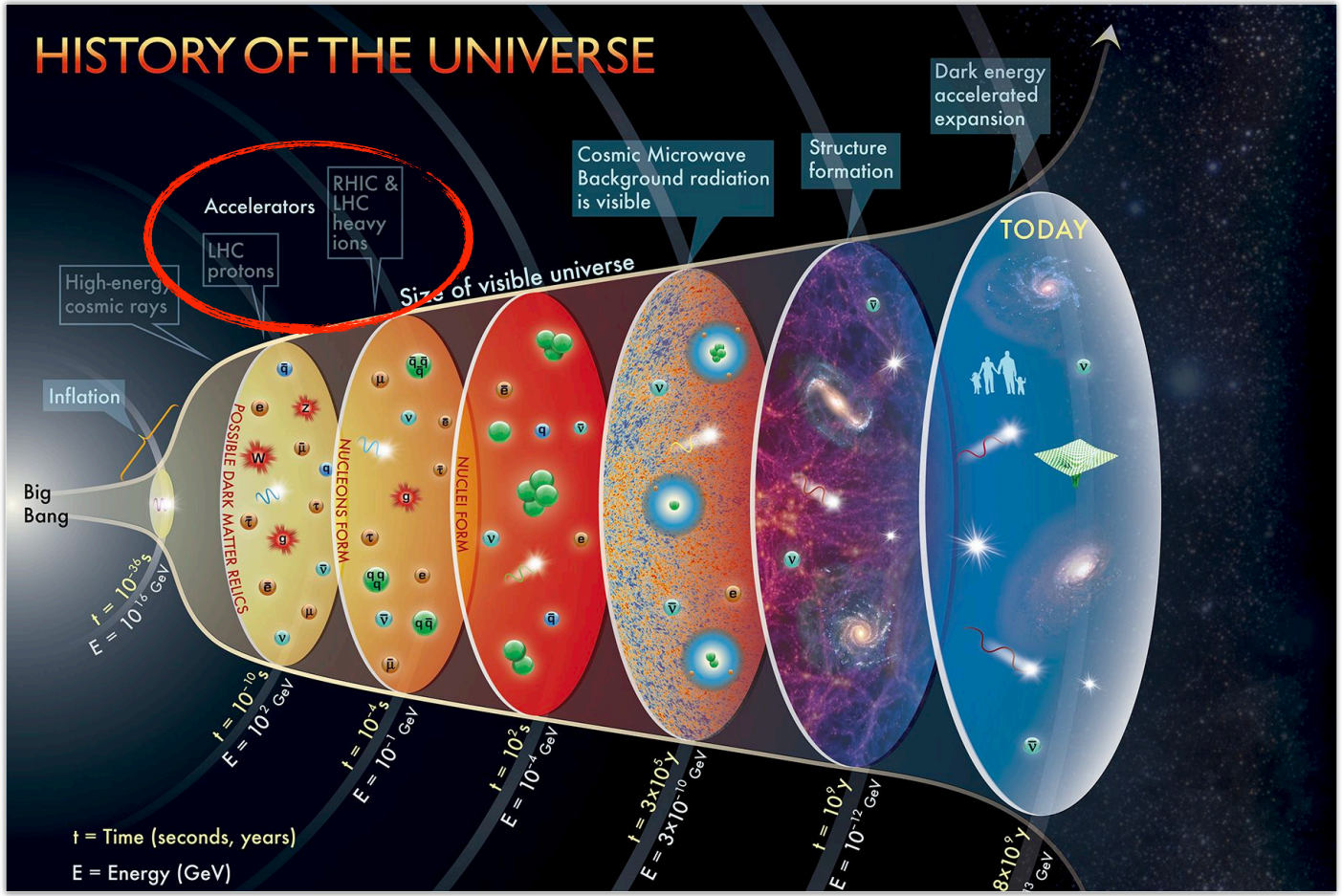
Fermilab


ANALOG SIGNALS
SENSOR
DIGITIZE
010110
101100
DATA REDUCTION
ZERO SUPPRESS,
THRESHOLD,
COMPRESS,
AGGREGATE,
FILTER, ...
X N STEPS...
SAVE TO DISK
Too Much Data
Too Late
1 channel ~ 10b
1 channel, 1 MHz rate ~ 10 Mb/s
100k channels, 1 MHz rate ~ 1 Tb/s

Fermilab

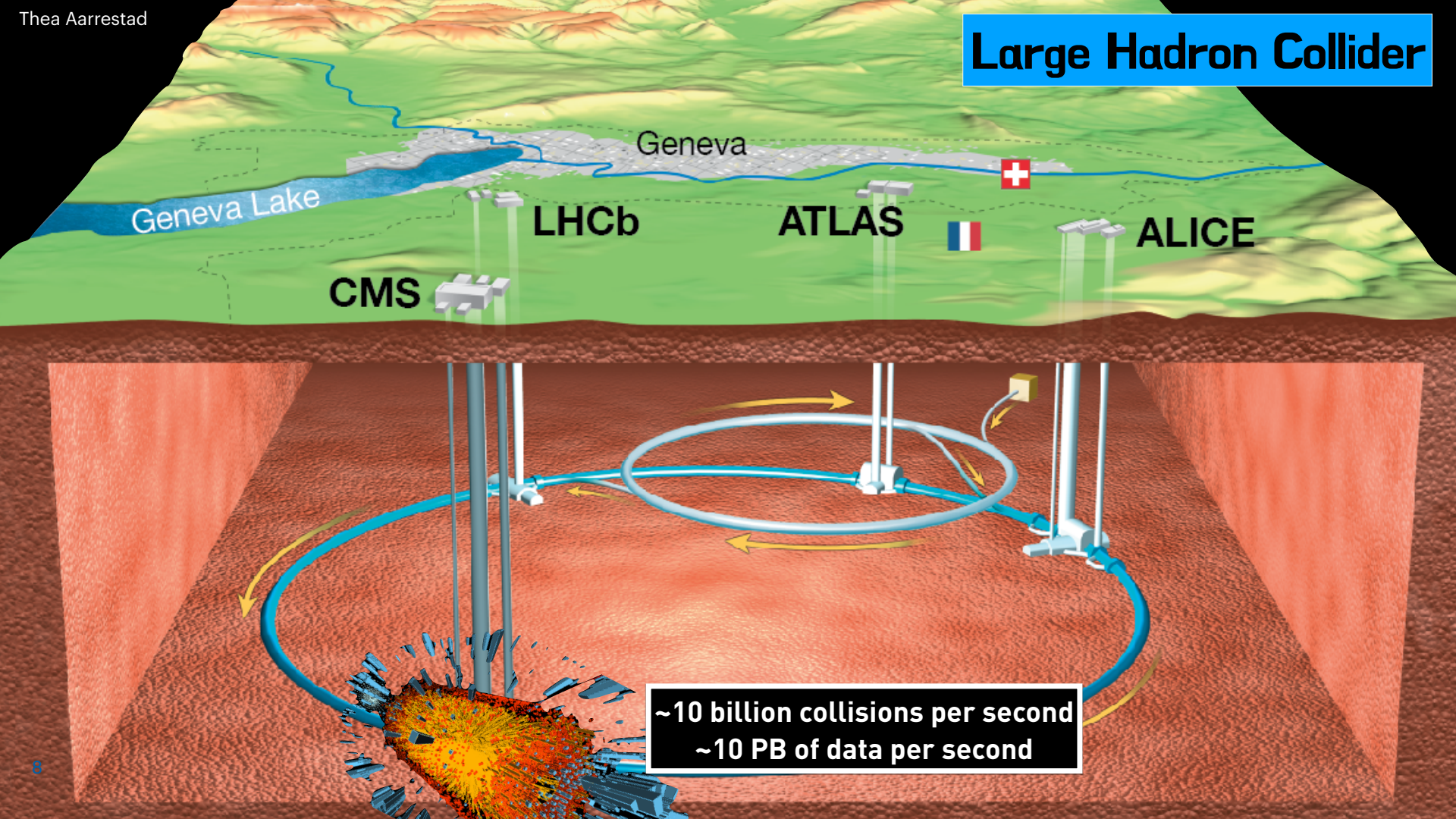

ANALOG SIGNALS
SENSOR
DIGITIZE
010110
101100
DATA REDUCTION
ZERO SUPPRESS,
THRESHOLD,
COMPRESS,
AGGREGATE,
FILTER, ...
X N STEPS...
SAVE TO DISK
Too Much Data
Too Late
Embed more intelligence

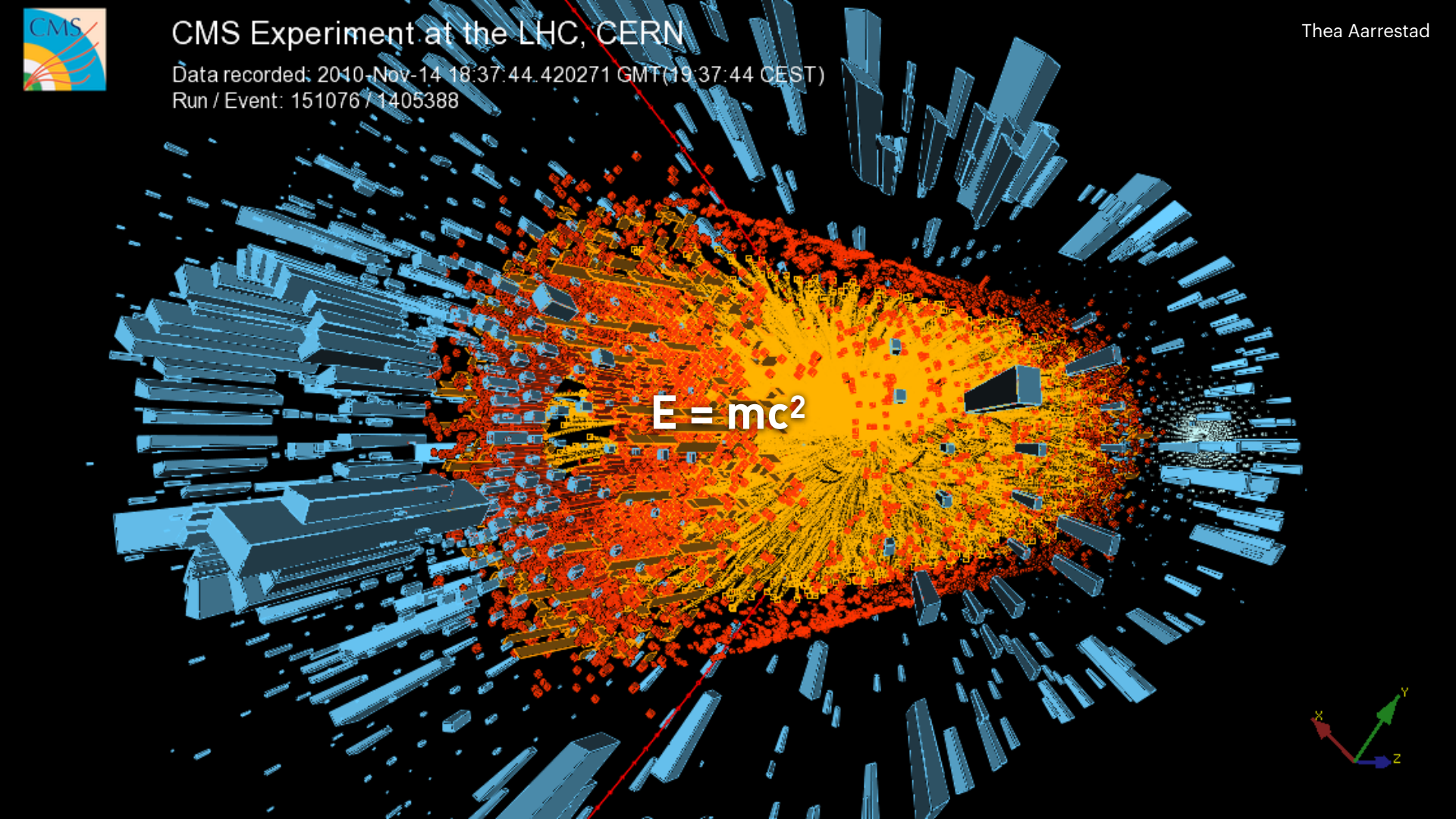
Fermilab


HISTORY OF THE UNIVERSE
Dark energy accelerated expansion
Cosmic Microwave Background radiation is visible
Structure formation
TODAY
Accelerators
RHIC & LHC heavy ions
LHC protons
High-energy cosmic rays
Size of visible universe
Inflation
Big Bang
POSSIBLE DARK MATTER RELICS
NUCLEONS FORM
NUCLEI FORM
t = 10^-36 s
E = 10^16 GeV
t = 10^-10 s
E = 10^2 GeV
t = 10^-4 s
E = 10^-1 GeV
t = 10^2 s
E = 10^-4 GeV
t = 3x10^5 y
E = 3x10^-10 GeV
t = 10^9 y
E = 10^-12 GeV
t = Time (seconds, years)
E = Energy (GeV)

Fermilab

Large Hadron Collider

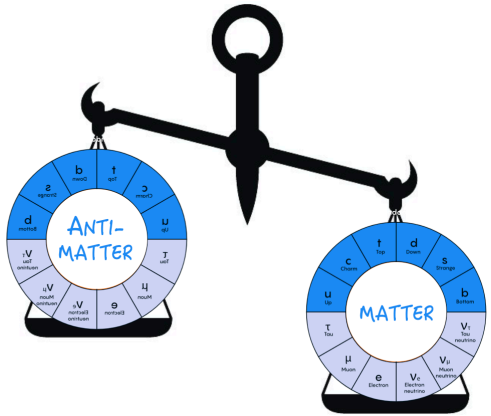



CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388
Thea Aarrestad
E = mc²
Y
X
Z

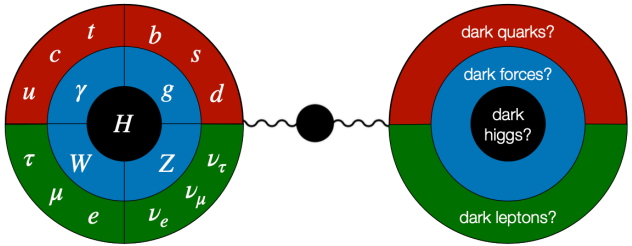
The Standard Model describes the **stuff (matter)** we are made of and how that stuff **interacts (forces)**

Open fundamental questions

Why are we here?
(Why more matter than anti-matter)

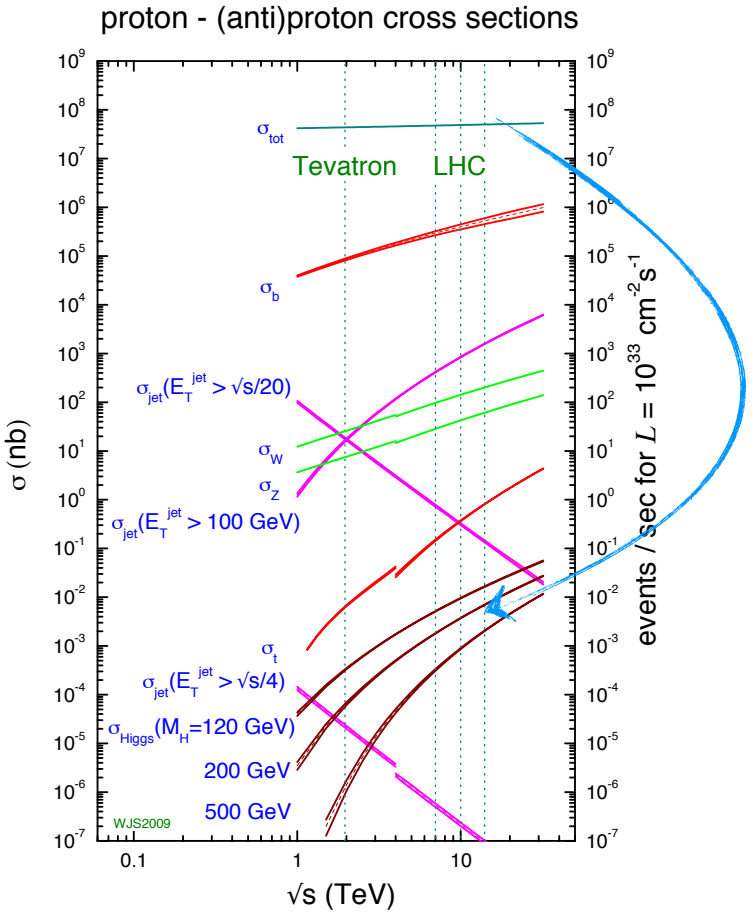


What is dark matter?



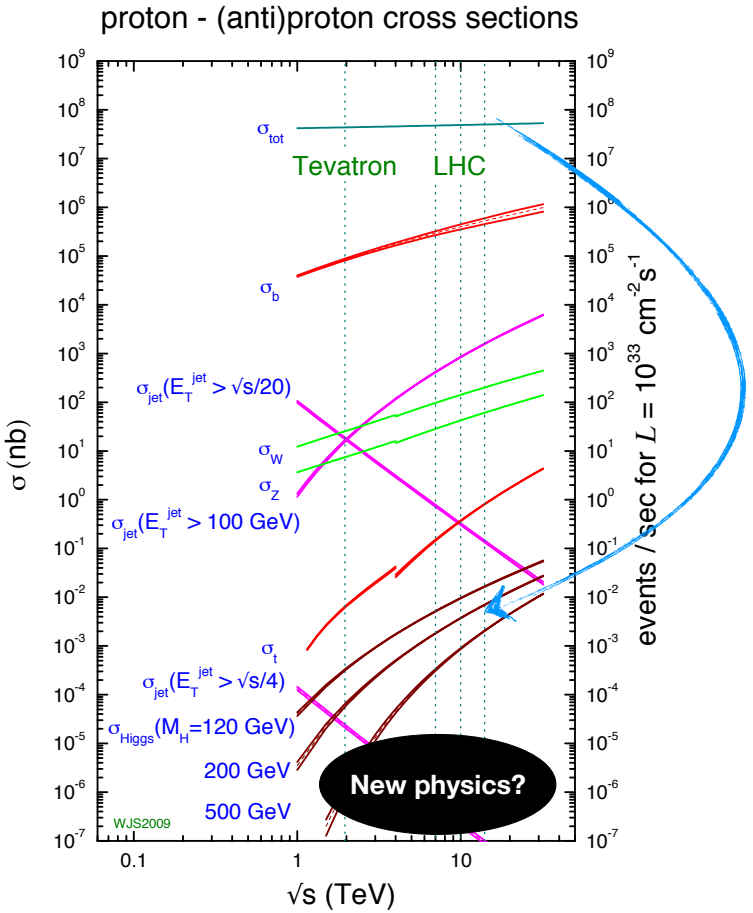
Why is gravity so weak?
What is the fate of the universe (is it stable)?
etc.

LOTS of collisions

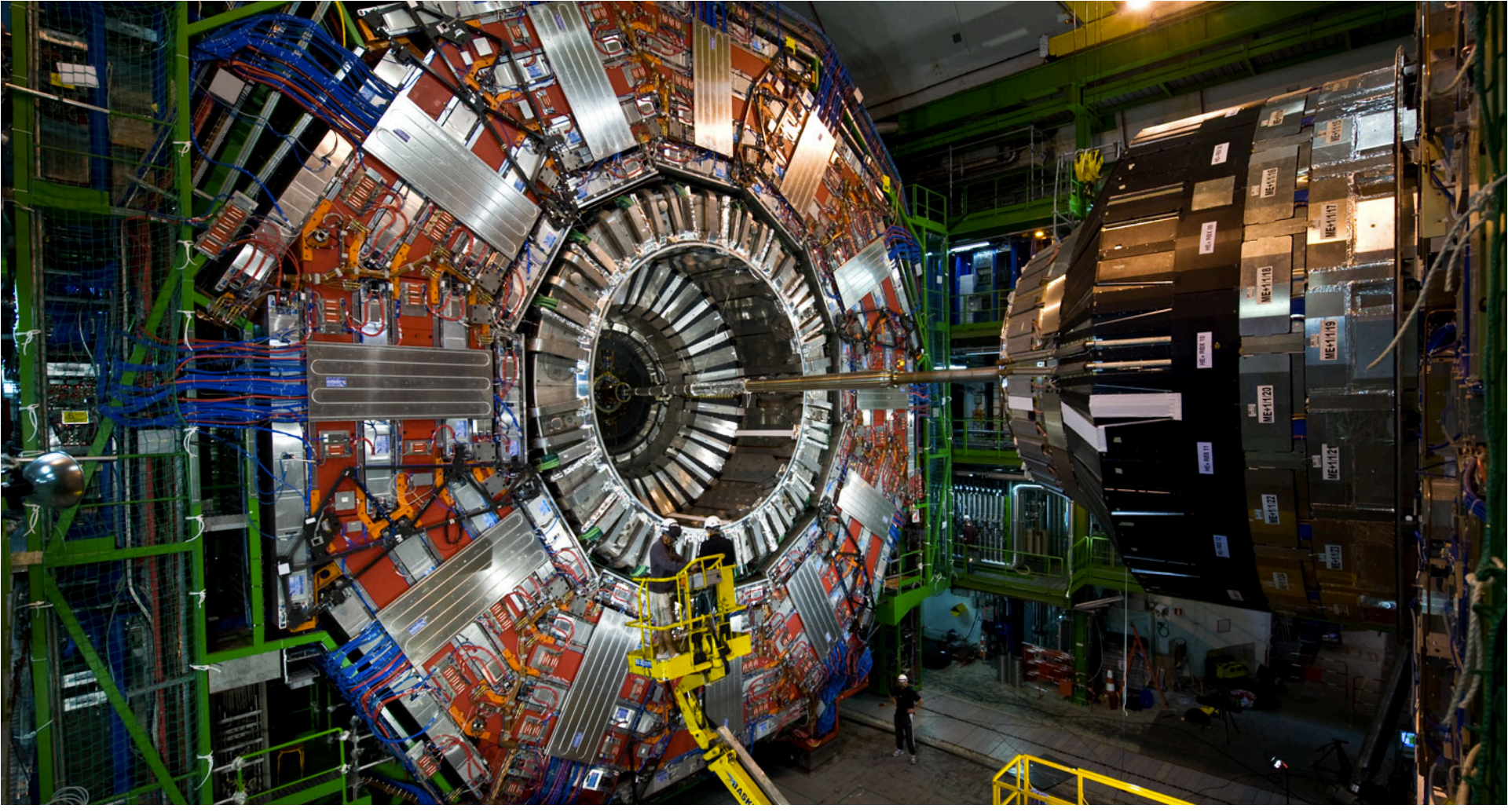


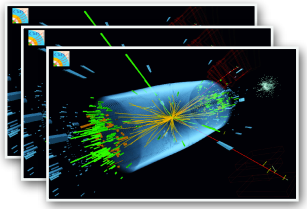
Higgs produced once every billion collisions

LOTS of collisions



Higgs produced once every billion collisions

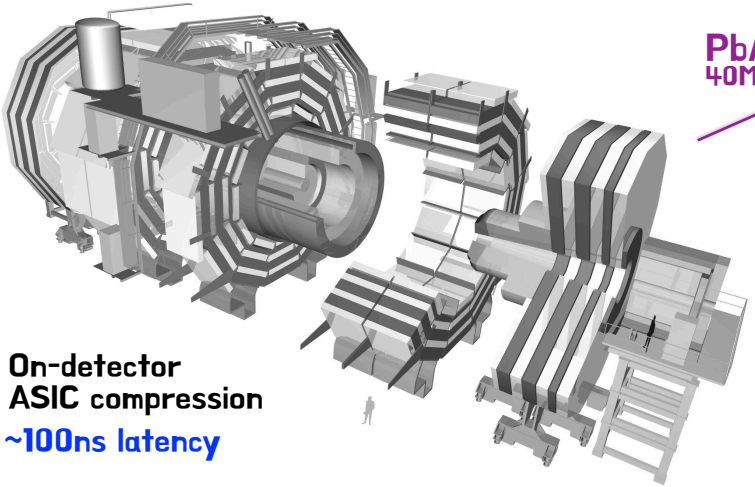

ME+1/1/8
ME+1/1/7
ME+1/1/6
ME+1/1/5
ME+1/1/4
ME+1/1/3
ME+1/1/2
ME+1/1/1



CMS Experiment

40MHz collision rate

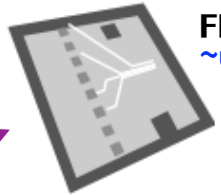
~1B detector channels



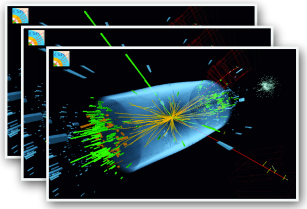
On-detector
ASIC compression

~100ns latency

Pb/s
40MHz



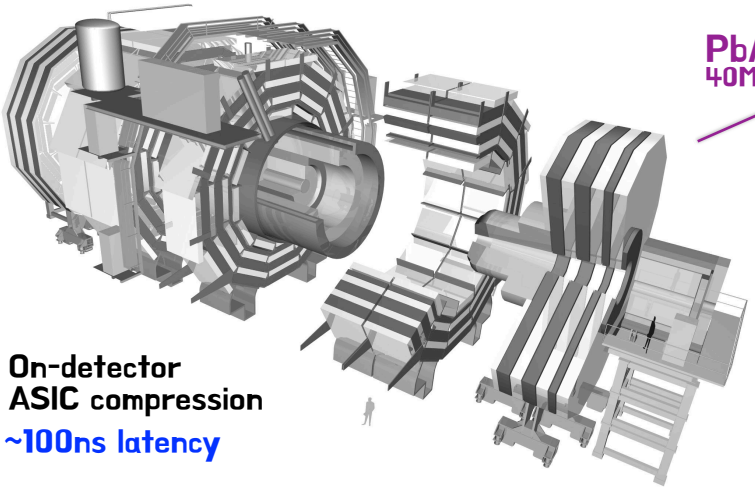
FPGA filter stack
~us latency



CMS Experiment

40MHz collision rate

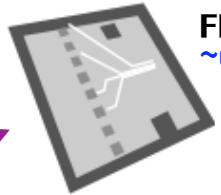
~1B detector channels



On-detector
ASIC compression

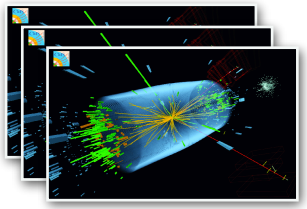
~100ns latency

Pb/s
40MHz



FPGA filter stack
~us latency

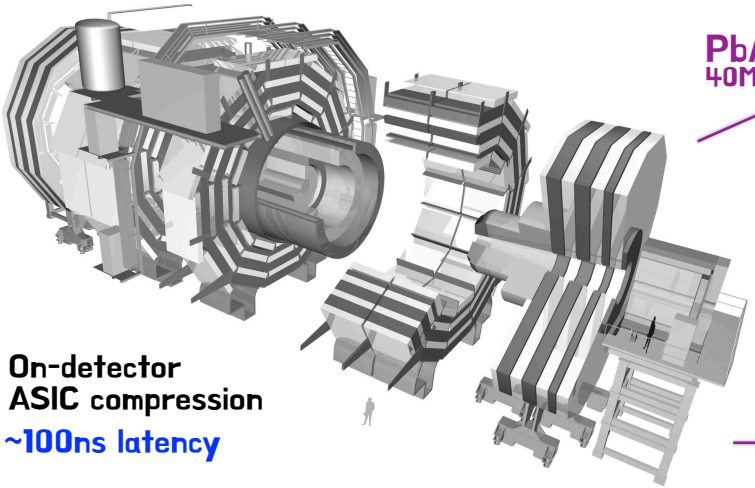
COMPARABLE TO GLOBAL INTERNET TRAFFIC BANDWIDTH



CMS Experiment

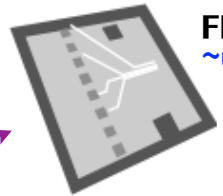
40MHz collision rate

~1B detector channels



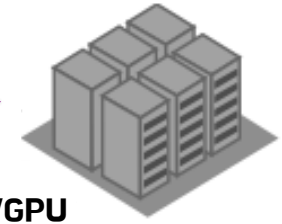
On-detector
ASIC compression
~100ns latency

Pb/s
40MHz



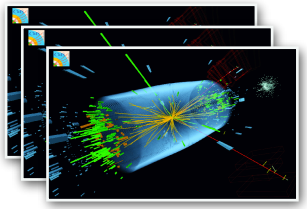
FPGA filter stack
~us latency

10s Tb/s
100s kHz



On-prem CPU/GPU
filter farm
~100 ms latency

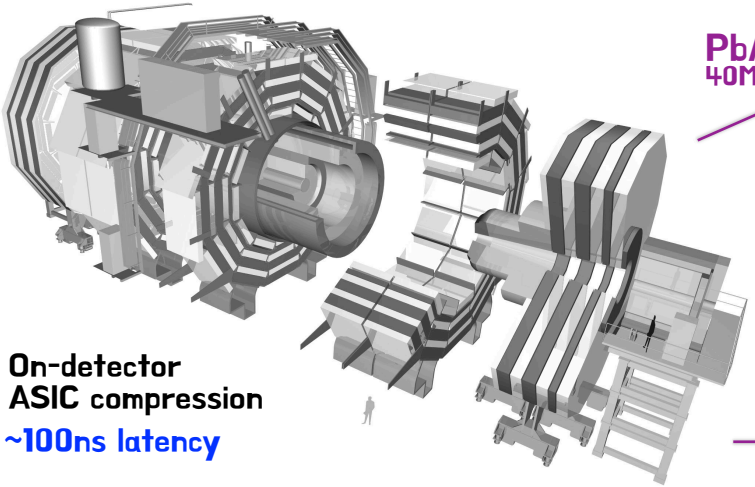
COMPARABLE TO GLOBAL INTERNET TRAFFIC BANDWIDTH



CMS Experiment

40MHz collision rate

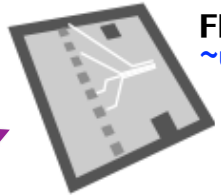
~1B detector channels



On-detector
ASIC compression

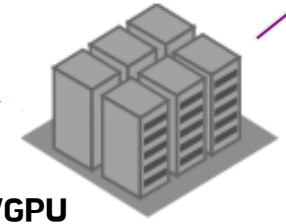
~100ns latency

Pb/s
40MHz



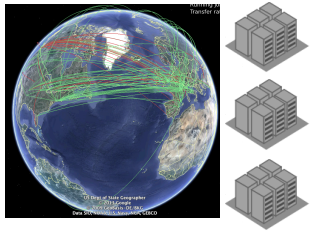
FPGA filter stack
~us latency

10s Tb/s
100s kHz



On-prem CPU/GPU
filter farm
~100 ms latency

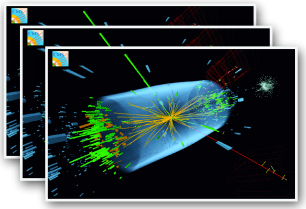
10s Gb/s
~5 kHz



Worldwide
computing grid

Exabyte-scale
datasets

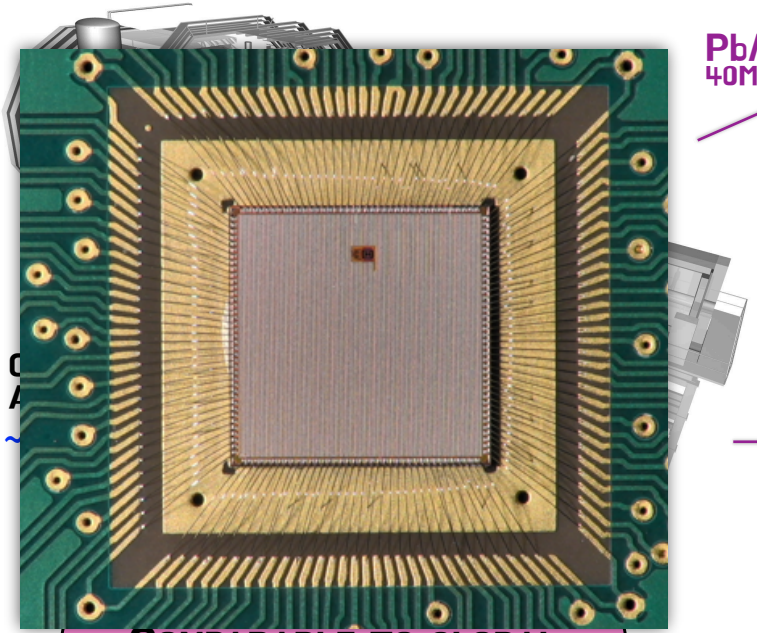
COMPARABLE TO GLOBAL
INTERNET TRAFFIC BANDWIDTH



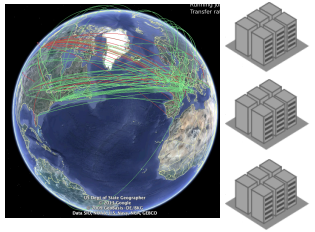
CMS Experiment

40MHz collision rate

~1B detector channels



Pb/s
40MHz



Worldwide computing grid

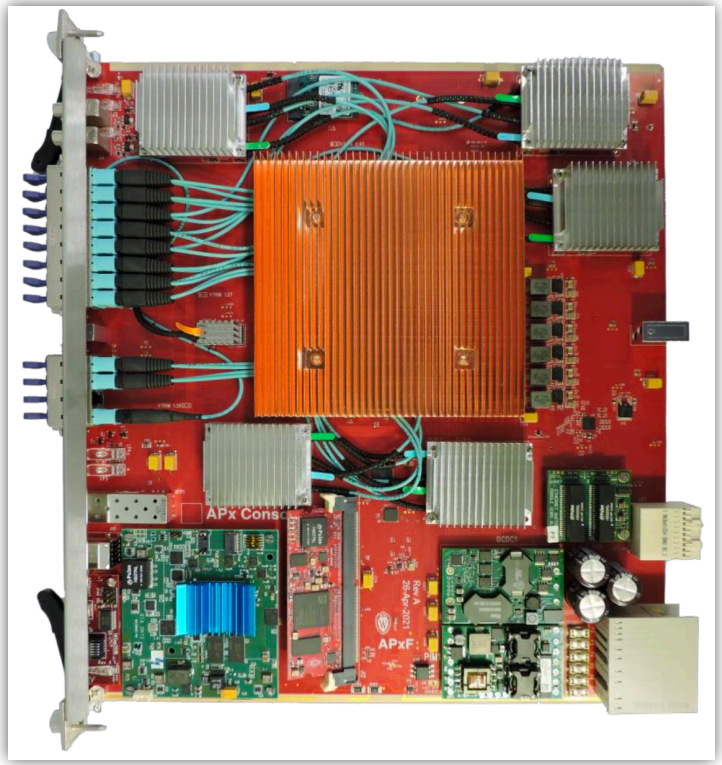
Exabyte-scale datasets

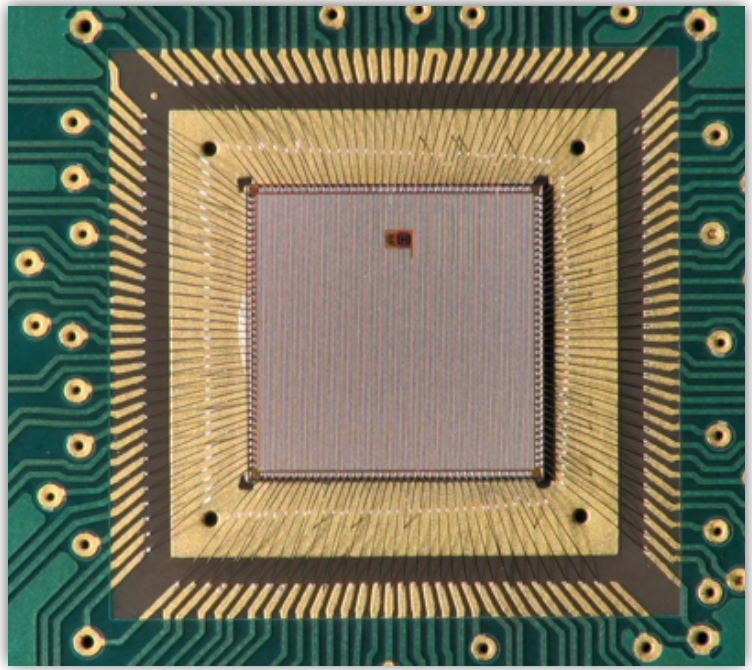
On-prem CPU/GPU filter farm

~100 ms latency

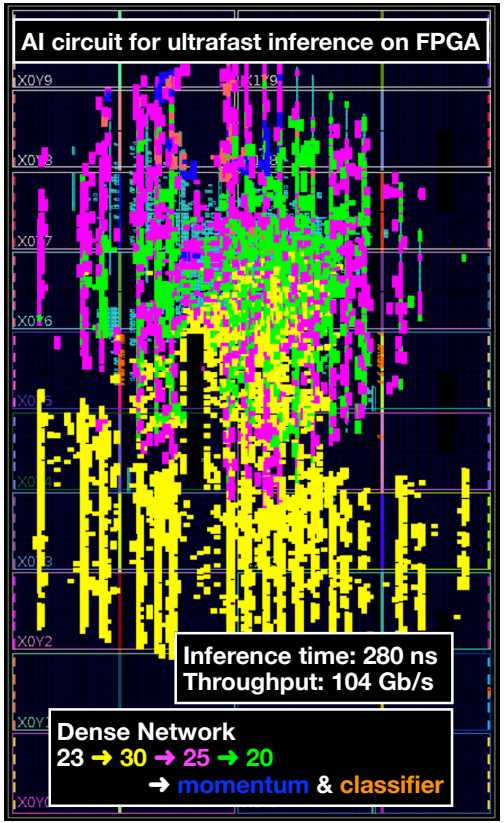
COMPARABLE TO GLOBAL INTERNET TRAFFIC BANDWIDTH

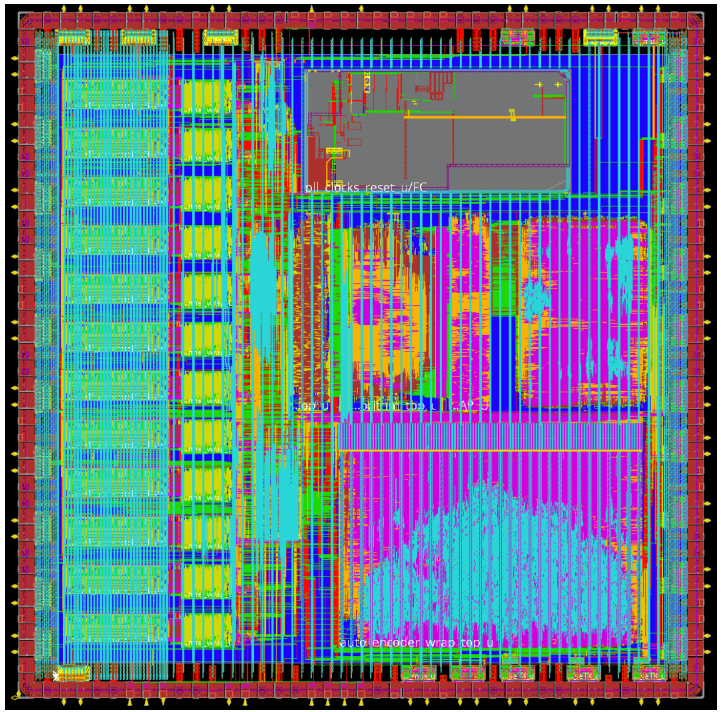
AI-on-chip





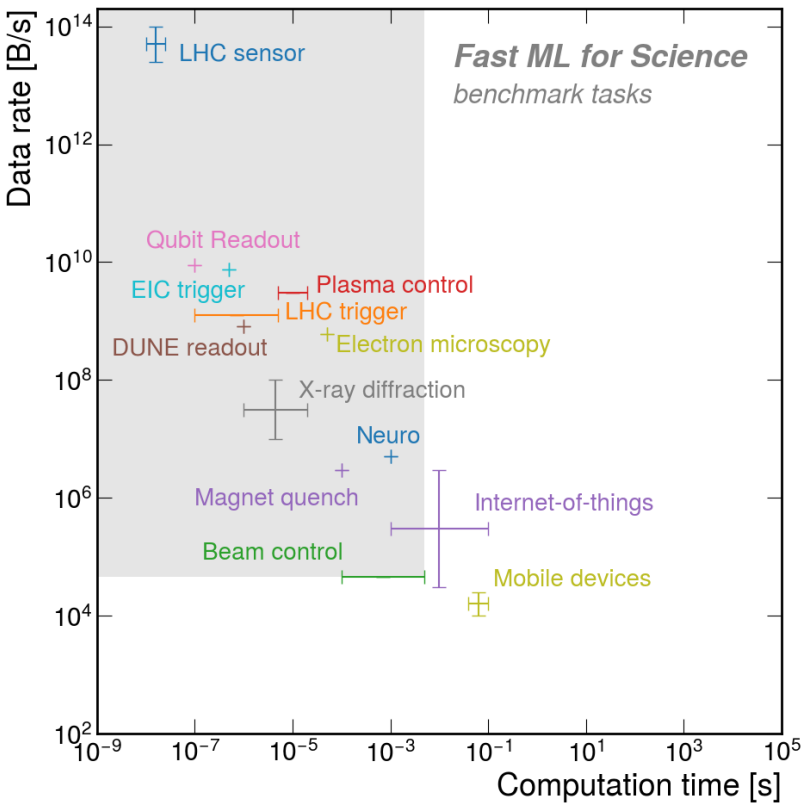
AI-on-chip





Fermilab

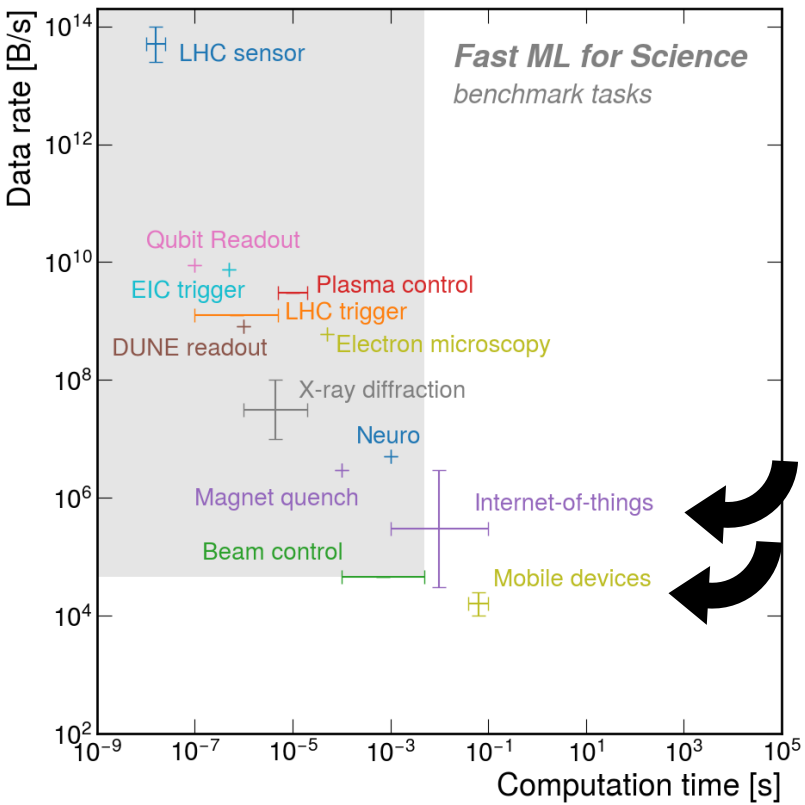
Fast ML for Science



Benchmarks bring innovation,
Grand challenges spark imaginations!

Fermilab

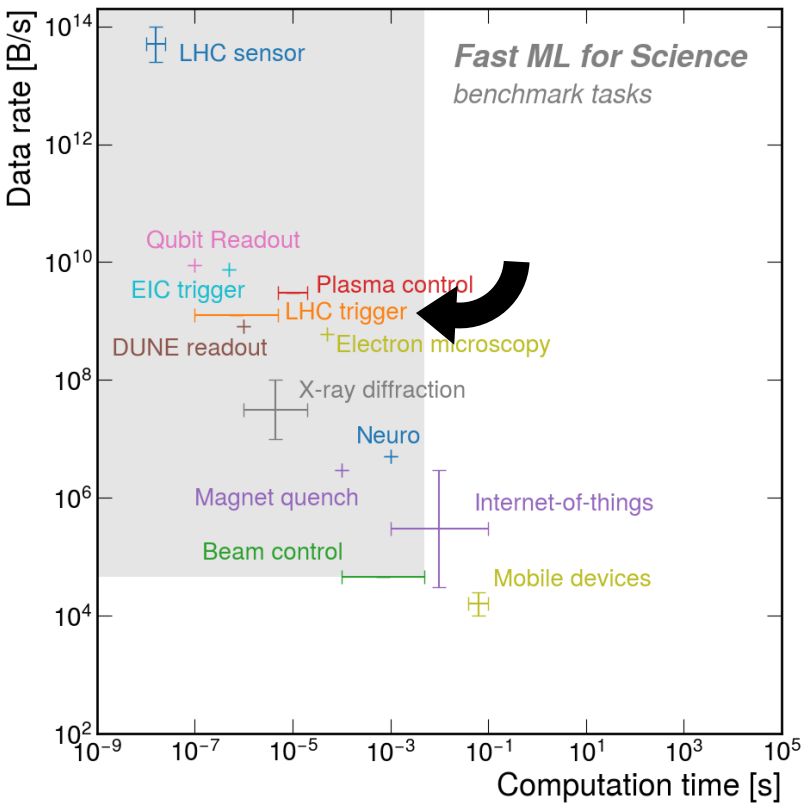
Fast ML for Science



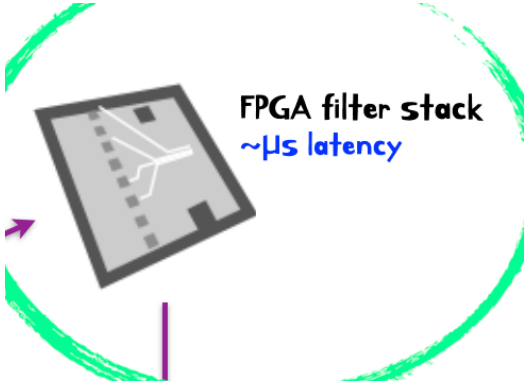
Benchmarks bring innovation,
Grand challenges spark imaginations!

Fermilab

Fast ML for Science

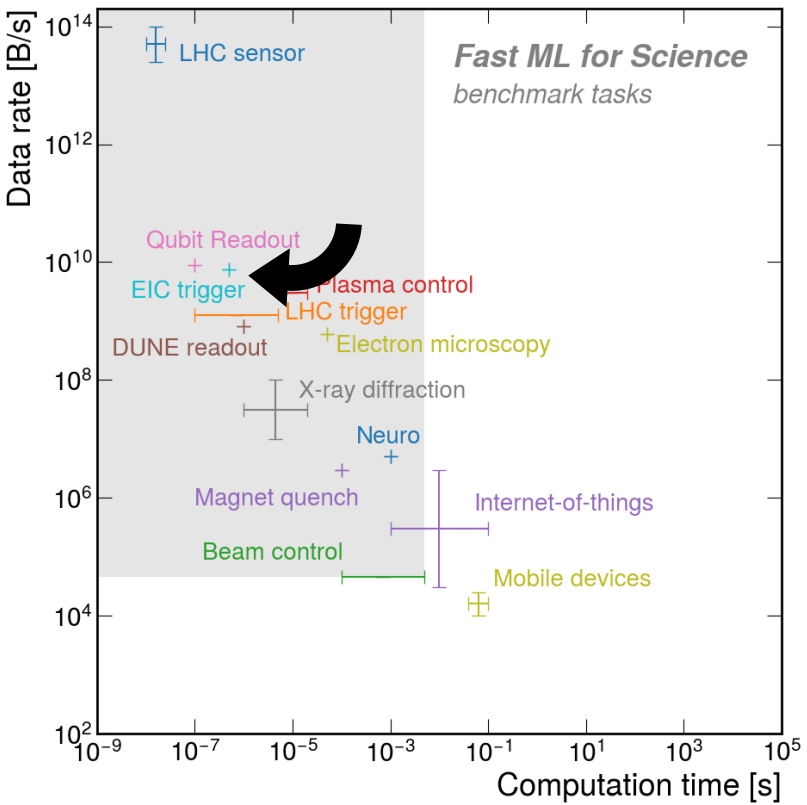


Benchmarks bring innovation,
Grand challenges spark imaginations!

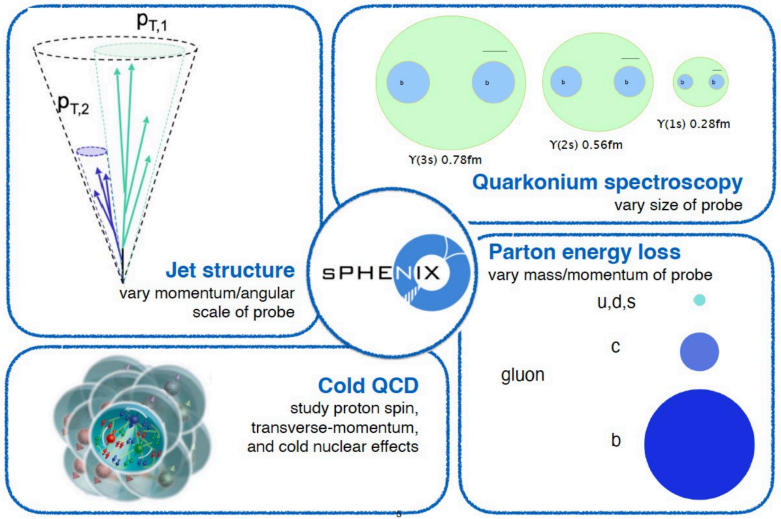


Fermilab

Fast ML for Science

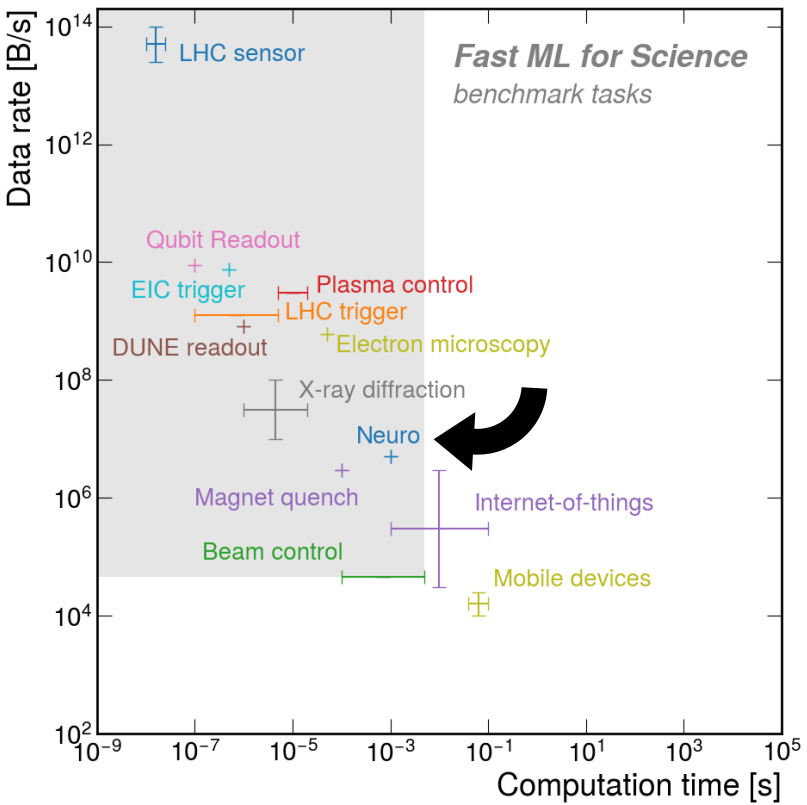


Benchmarks bring innovation,
Grand challenges spark imaginations!

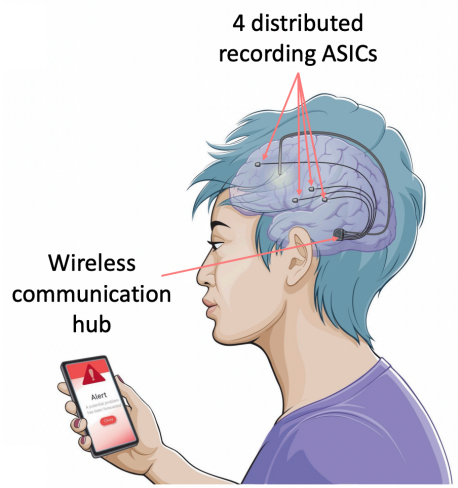


Fermilab

Fast ML for Science



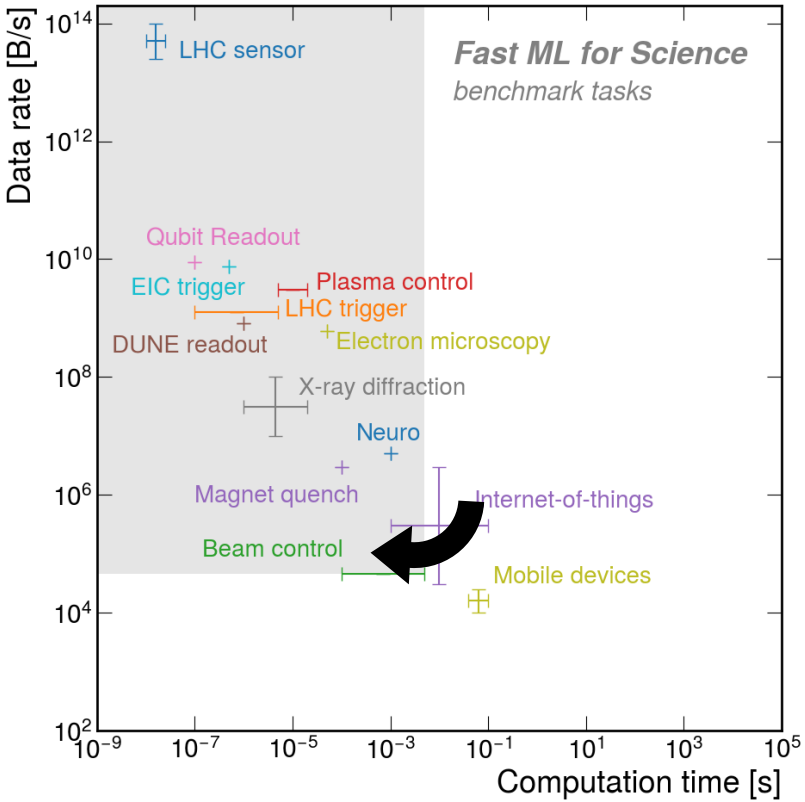
Benchmarks bring innovation,
Grand challenges spark imaginations!



Real-time seizure detection

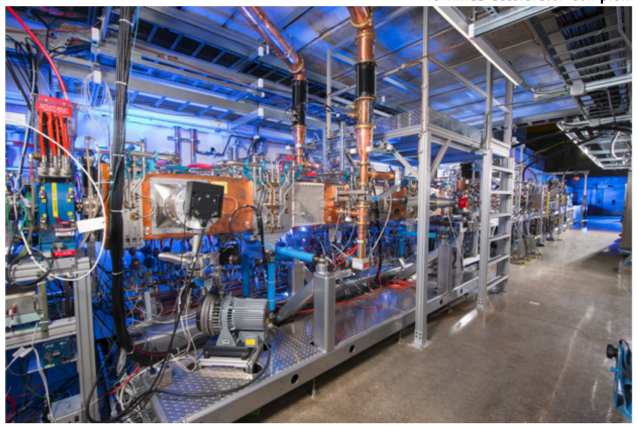
Fermilab

Fast ML for Science



Benchmarks bring innovation,
Grand challenges spark imaginations!

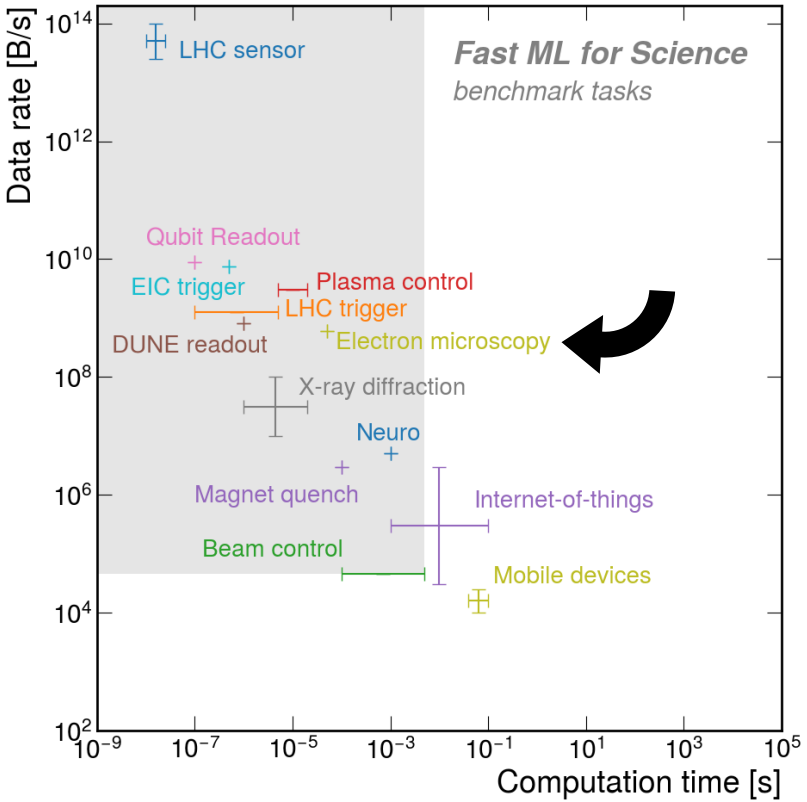
Fermilab accelerator complex



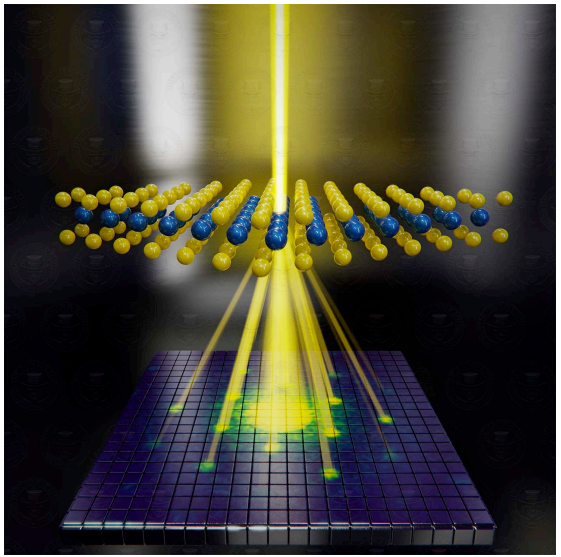
Particle accelerator controls

Fermilab

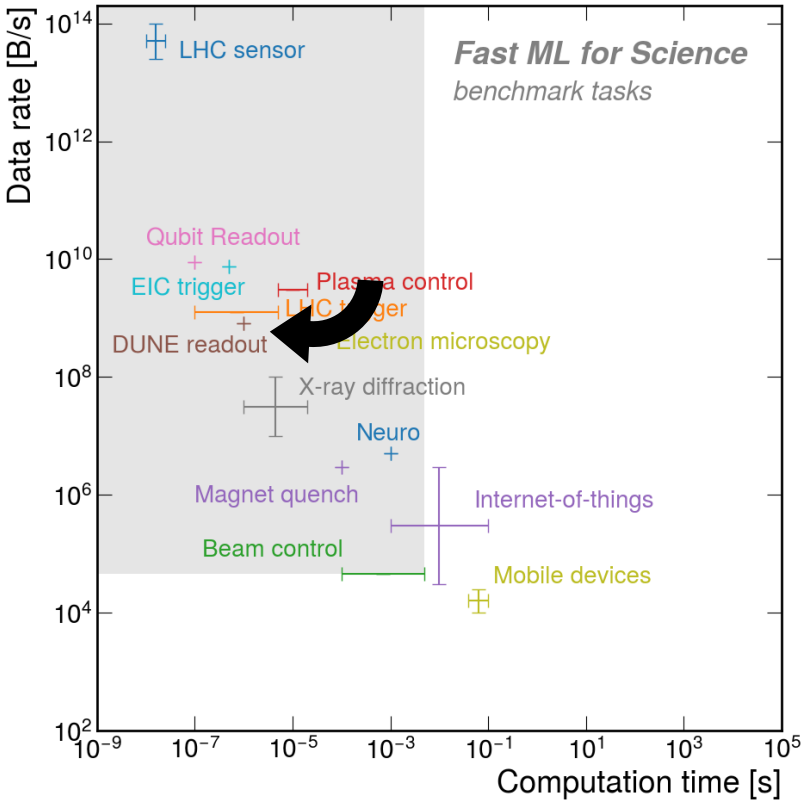
Fast ML for Science



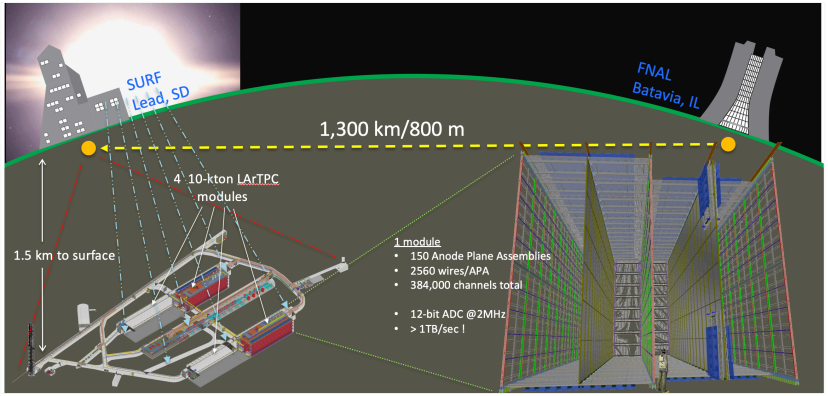
Benchmarks bring innovation,
Grand challenges spark imaginations!



Fast ML for Science



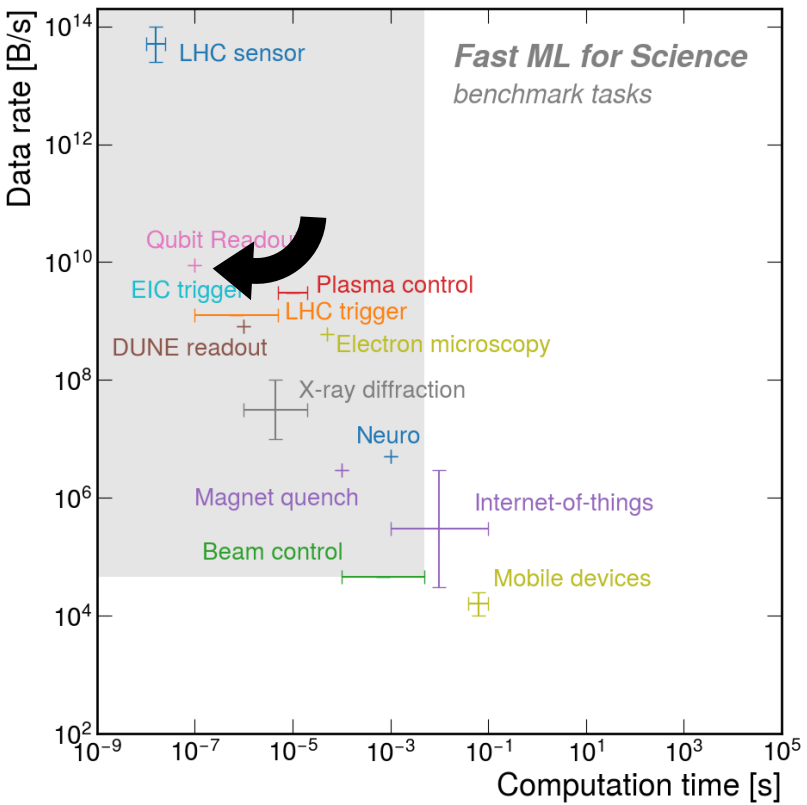
Benchmarks bring innovation,
Grand challenges spark imaginations!



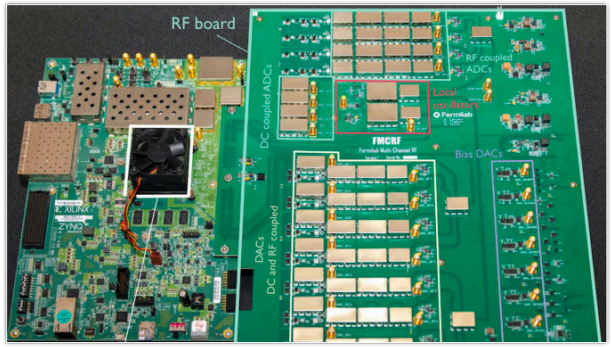
Supernova detection and multi-messenger astronomy

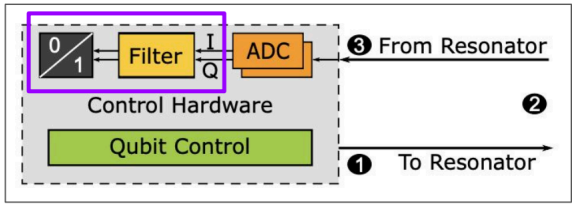
Fermilab

Fast ML for Science



Benchmarks bring innovation,
Grand challenges spark imaginations!

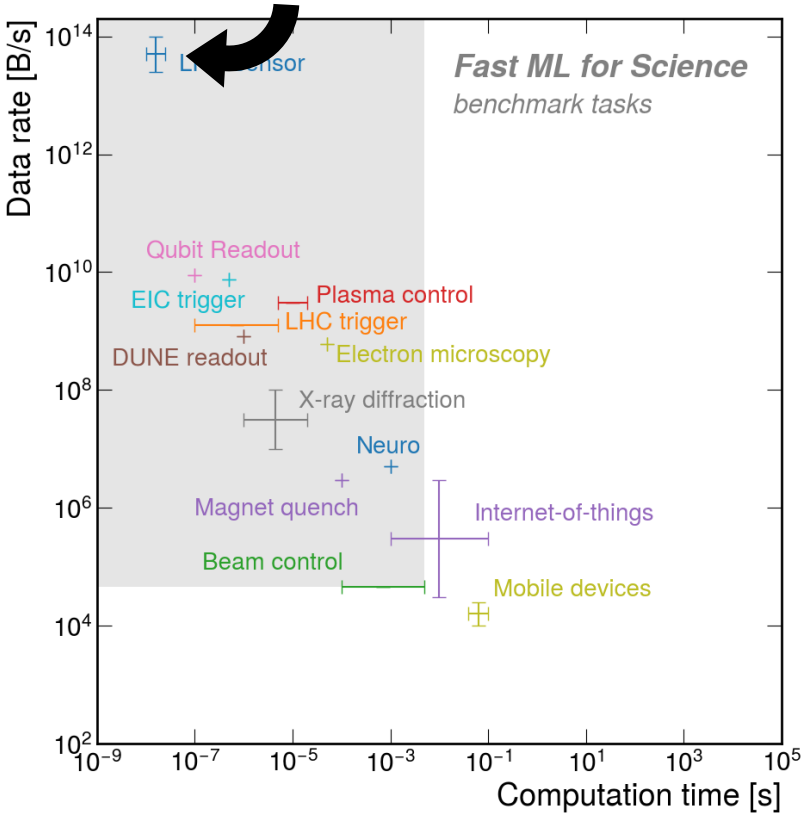




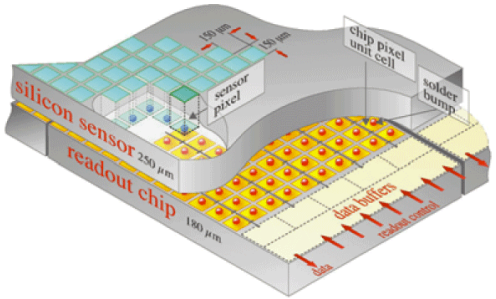
Qubit readout and control

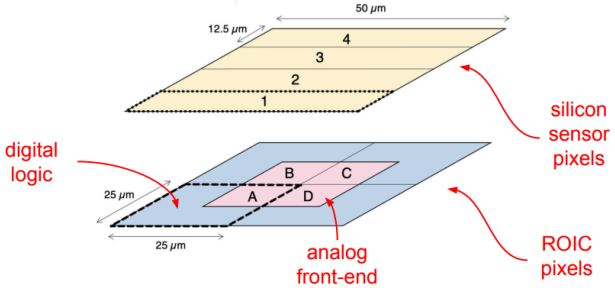
Fermilab

Fast ML for Science



Benchmarks bring innovation,
Grand challenges spark imaginations!



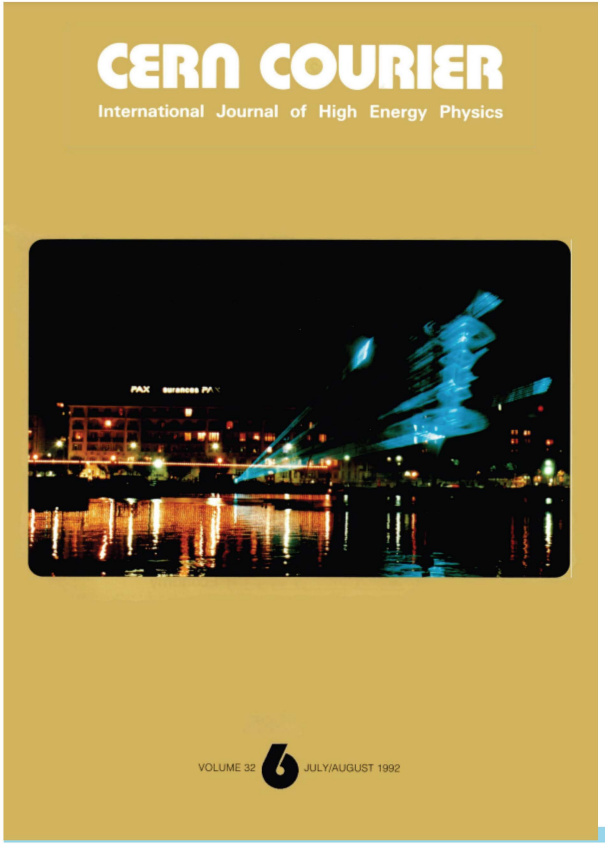


Full 40 MHz readout with smart pixel detectors

Fermilab

Outline

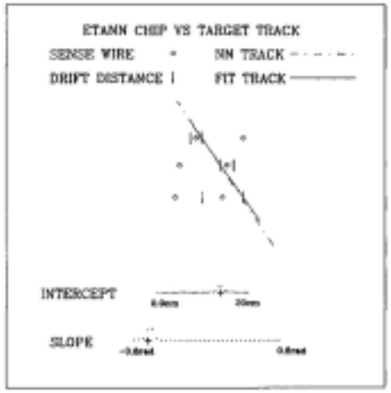
- Why Fast ML for Science?
- The intelligent edge of tomorrow
- Outlook



ware of a high energy physics experiment.

The first such application comes from a recent Fermilab test beam experiment, where a VLSI neural network chip was interfaced to the data acquisition system of a prototype drift chamber. Drift time information from the sense wires, encoded as voltages, was passed to the neural network, which calculated the slope and intercept of the track traversing the chamber and sent this information back to the mother readout board to be read out with the rest of the event, without any dead time.

Neural network hardware is also finding its way into other trigger systems. The CDF experiment has



three neural network triggers in place for its 1992 run: an isolated endplug electron trigger, an isolated central photon trigger, and a semileptonic B particle trigger.

Also at Fermilab's Tevatron collider, a group in the D0 experiment is studying the use of neural networks in the muon trigger for the D0 Muon Upgrade. A neural network trigger for H1 at DESY has been under development for some time and will be tested in the current run. Several R&D projects at CERN are looking at the feasibility of neural networks for LHC experiment trigger systems.

Another application of neural networks under study is in adaptive control systems for accelerators. A group at SLAC recently simulated how a neural network control system could be trained both to emulate and control a section of beamline.

These new artificial intelligence techniques could go on to play an important role in the acquisition and analysis of experimental data for the coming generation of proton colliders.

From Bruce Denby and Clark Lindsey (Fermilab) and Louis Lyons (Oxford)

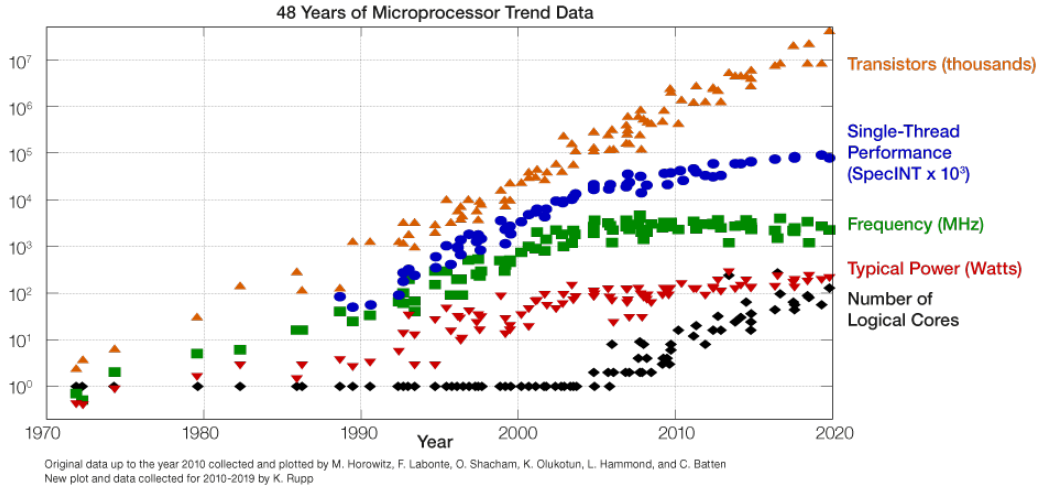
Fermilab

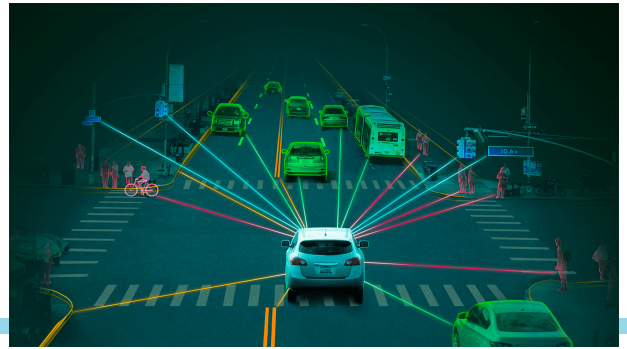
Edge of Tomorrow

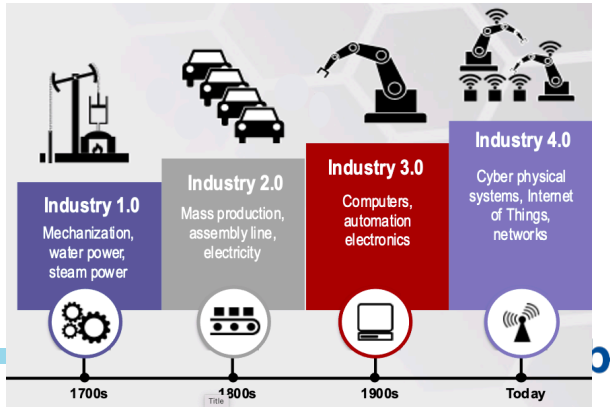
- Necessity
- Hardware
- ML Research
- Tools

Edge of Tomorrow

- Necessity
- Hardware
- ML Research
- Tools

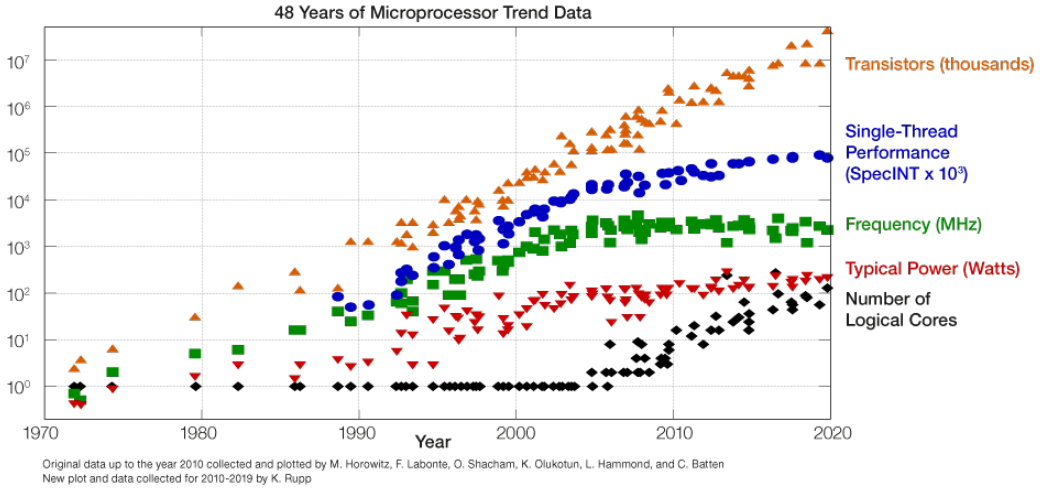


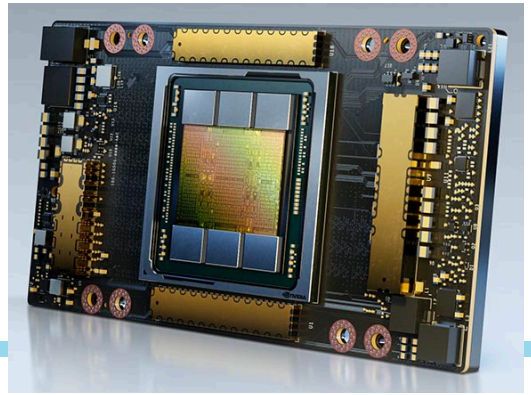




Edge of Tomorrow

- Necessity
- Hardware
- ML Research
- Tools





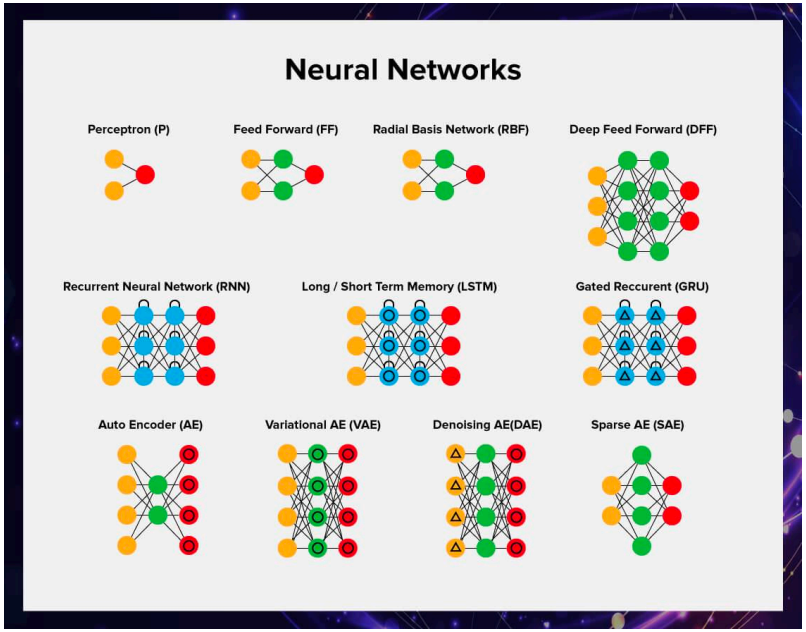
Technology size	Year	Technology size	Year
10 um	1971	130 nm	2001
6 um	1974	90 nm	2004
3 um	1977	65 nm	2006
1.5 um	1982	45 nm	2008
1 um	1985	32 nm	2010
800 nm	1989	22 nm	2012
600 nm	1994	14 nm	2014
350 nm	1995	10 nm	2017
250 nm	1997	7 nm	2018
180 nm	1999	5 nm	2020

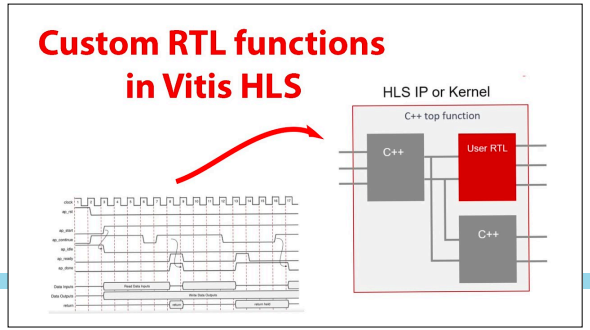
Edge of Tomorrow

- Necessity
- Hardware
- ML Research
- Tools

TensorFlow

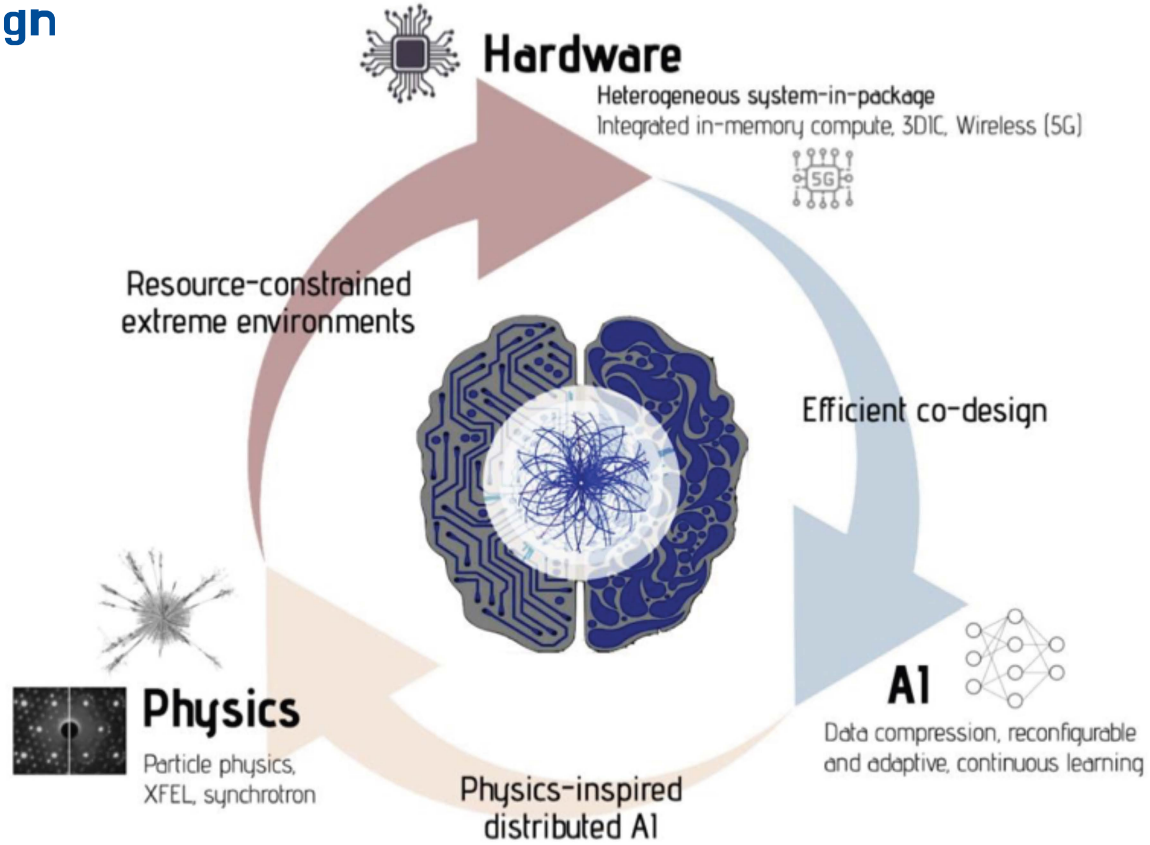
PyTorch





NVIDIA CUDA

Efficient codesign



Fermilab

Efficient codesign

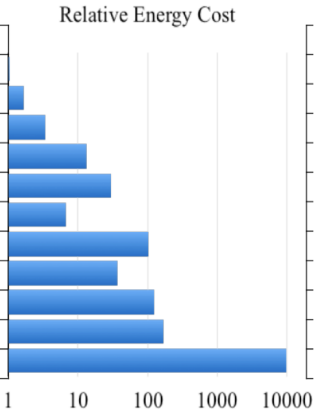
Efficient codesign

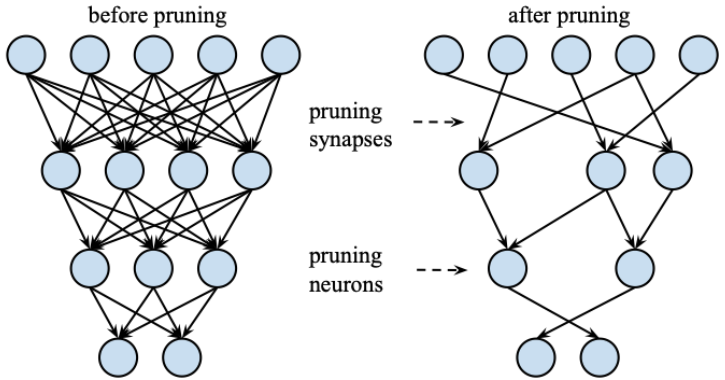
algorithm

A lot of literature on sparsity and quantization as very generalizable techniques

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





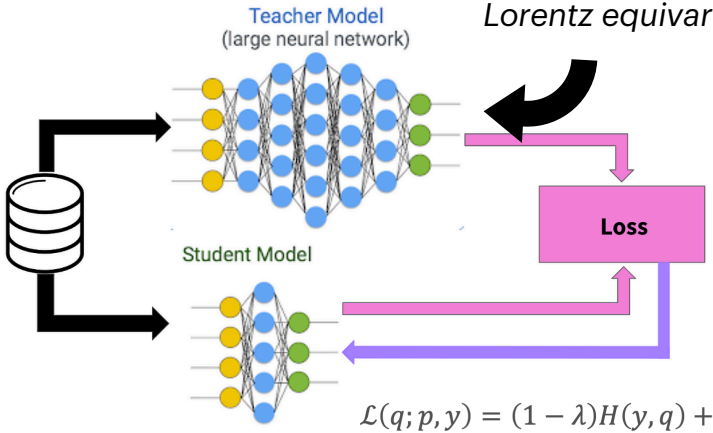
Fermilab

Efficient codesign

algorithm

More interesting directions — distillation and inductive bias

Teacher model knows about Lorentz equivariance



$$\mathcal{L}(q;p,y) = (1-\lambda)H(y,q) + \lambda D_{KL}(\tilde{p}||\tilde{q})$$

q: student's prediction, p: teacher's prediction, y: ground truth, $\tilde{p}$: temperature soften distribution.

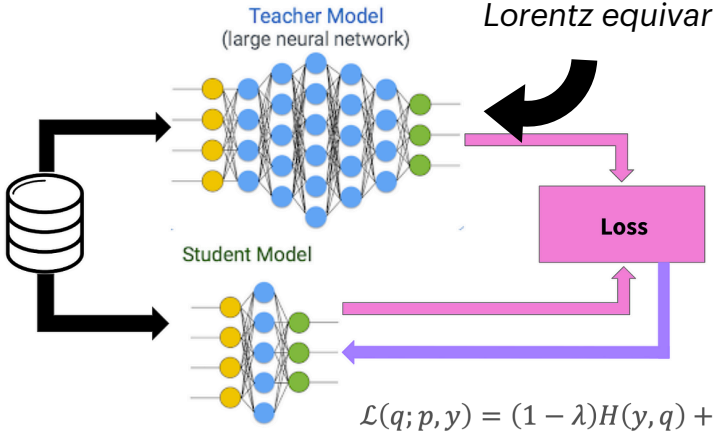
Fermilab

Efficient codesign

algorithm

More interesting directions — distillation and inductive bias

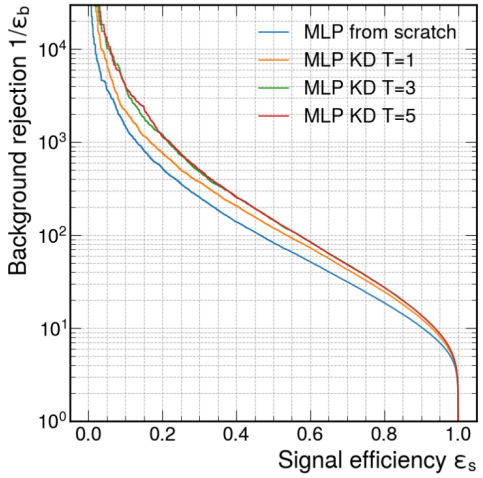
Teacher model knows about Lorentz equivariance



$$\mathcal{L}(q; p, y) = (1 - \lambda)H(y, q) + \lambda D_{KL}(\tilde{p}||\tilde{q})$$

q: student's prediction, p: teacher's prediction, y: ground truth, $\tilde{p}$: temperature soften distribution.

Model performance improves with distillation of expert knowledge, and more robust (see talk)

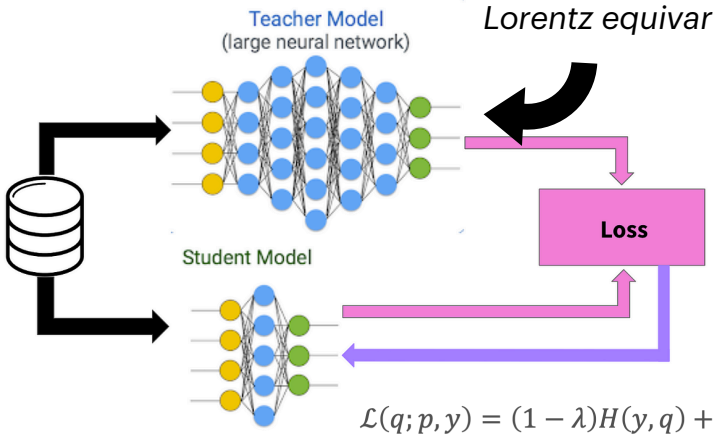


Fermilab

Efficient codesign

algorithm

More interesting directions — distillation and inductive bias



Teacher model knows about Lorentz equivariance

$$\mathcal{L}(q;p,y) = (1-\lambda)H(y,q) + \lambda D_{KL}(\tilde{p}||\tilde{q})$$

q: student's prediction, p: teacher's prediction, y: ground truth, $\tilde{p}$: temperature soften distribution.

19 Parameters Is All You Need: Tiny Neural Networks for Particle Physics

Alexander Bogatskiy
Center for Computational Mathematics
Flatiron Institute, New York, NY, U.S.A.
abogatskiy@flatironinstitute.org

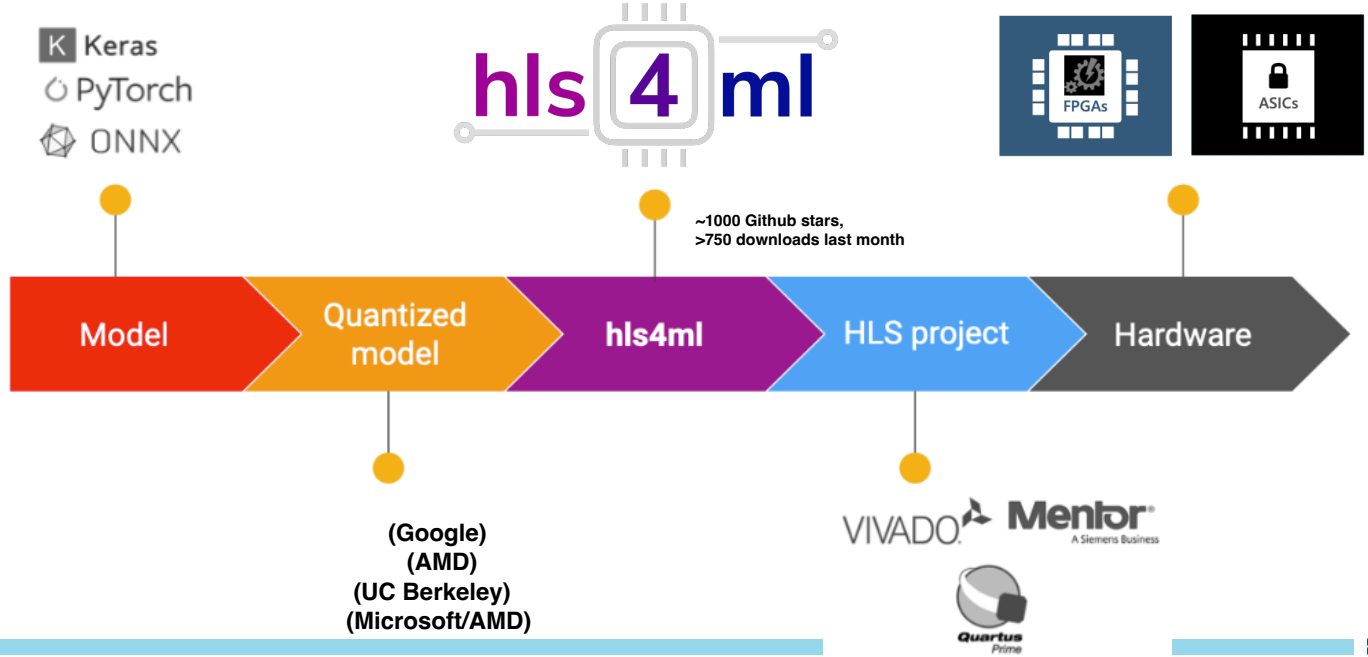
Timothy Hoffman
Department of Physics, University of Chicago
Chicago, IL, U.S.A.
hoffmant@uchicago.edu

Jan T. Offermann
Department of Physics, University of Chicago
Enrico Fermi Institute
Chicago, IL, U.S.A.
jano@uchicago.edu

n.b. not necessarily computationally light, TBD

Fermilab

Efficient codesign tools for

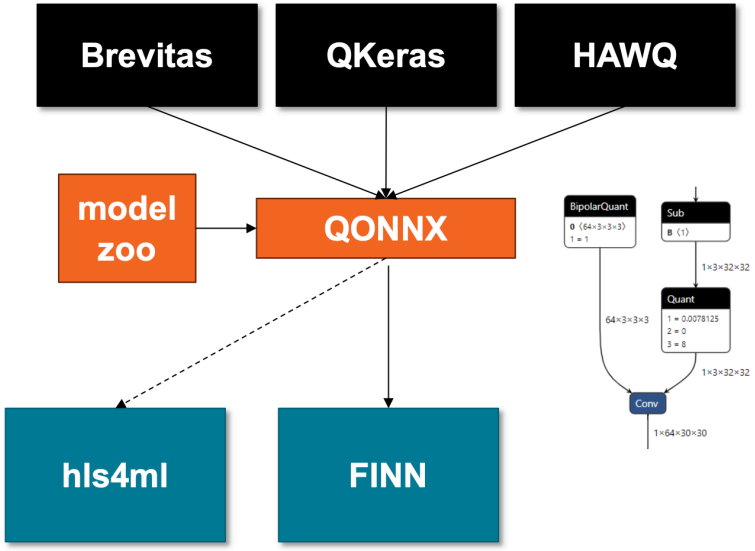


Fermilab

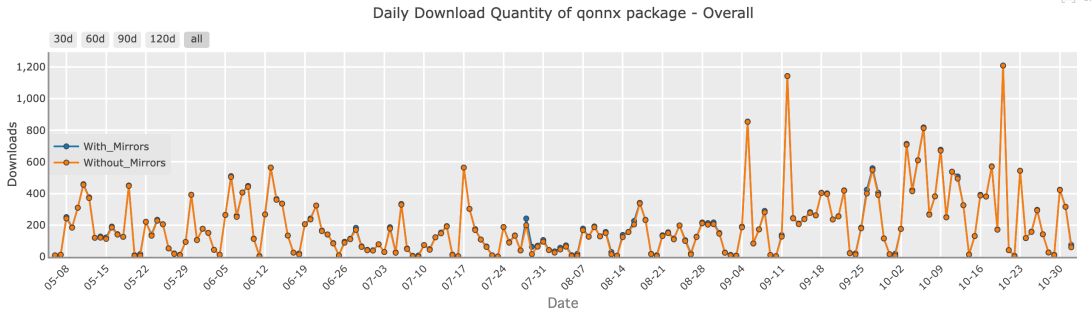
adapted from Vladimir Loncar

Efficient codesign tools for

Credit: Yaman Umuroglu



	Arbitrary precision	Rounding variants	Below 8-bits precision	Weights-only quantization	Avoid op. duplication	High-precision output
QONNX (this work)	✓	✓	✓	✓	✓	✓
QCDQ (this work)	×	×	✓	✓	✓	✓
Quantized op. with clipping (this work)	×	×	✓	×	×	×
QDQ [12]	×	×	×	✓	✓	✓
Integer op. [12]	×	×	×	×	×	✓
Quantized op. [12]	×	×	×	×	×	×



Fermilab

Efficient codesign workflows and collabs for

- Open-source
- Community-supported
- User-driven
- Accessible and usable

N

Use hls4ml to convert a neural network from Keras to HLS

`hls4ml` is a package developed to convert machine learning models (like those from Keras) into high-level synthesis (HLS) code. This allows the deployment of such models onto FPGAs. The following steps will guide you through converting a Keras neural network model to HLS using `hls4ml`:

1. **Installation**:
   If you haven't installed `hls4ml` yet, you can do so using `pip`:

```
pip install hls4ml
```

2. **Prepare Your Keras Model**:
   Let's assume you've trained a simple model in Keras or you have a Keras model ready:

```python
from keras.models import Sequential
from keras.layers import Dense

# Create a simple model
model = Sequential()
model.add(Dense(64, input_shape=(16,), activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accur
```

   Make sure you've saved your trained model:

```python
model.save('keras_model.h5')
```

3. **Convert with hls4ml**:

```python
import hls4ml

# Configure the converter
config = hls4ml.utils.config_from_keras_model(model, granularity='model')
print(config)

# Convert the model
hls_model = hls4ml.converters.convert_from_keras_model(model,
                                                       hls_config=config,
                                                       output_dir='hls4ml_p
                                                       fpga_part='xcvu9p-fl

# Print the model configuration to check
hls4ml.utils.plot_model(hls_model, show_shapes=True, show_precision=True, t
```

   In the `hls4ml.converters.convert_from_keras_model` function:
   - `hls_config`: is the configuration for the conversion. We generated it using `config_from_keras_model` for simplicity, but you can customize this as needed.
   - `output_dir`: is the directory where the HLS project will be created.
   - `fpga_part`: specifies the FPGA part. Adjust this based on your target FPGA.

4. **Build the HLS Project**:
   Once the model is converted, you can compile it into HLS:

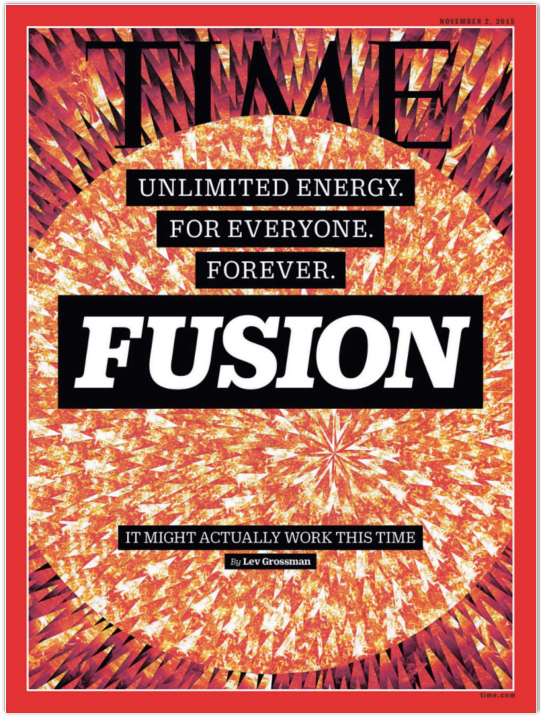
```python
hls_model.compile()
```

5. **Run the HLS Simulation (Optional)**:
   Before synthesizing for FPGA, you can run a C-simulation to check if the model works correctly in HLS:

```python
hls_model.build(csim=True)
```

After this, you'll have an HLS project in the specified `output_dir` that you can use with FPGA development tools to generate bitstreams for FPGA deployment.

A seemingly different topic - rolling it all up!



Google DeepMind

Accelerating fusion science through learned plasma control

16 FEBRUARY 2022

Pulsar Team, Swiss Plasma Center* (* External authors)

Share

REUTERS® World Business Markets Sustainability Legal Breakingviews Te

Energy | Grid & Infrastructure | Nuclear

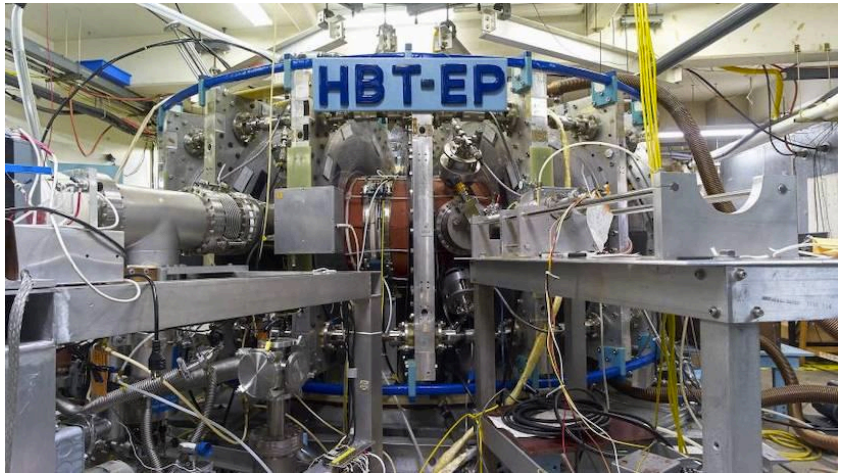
US scientists repeat fusion ignition breakthrough for 2nd time

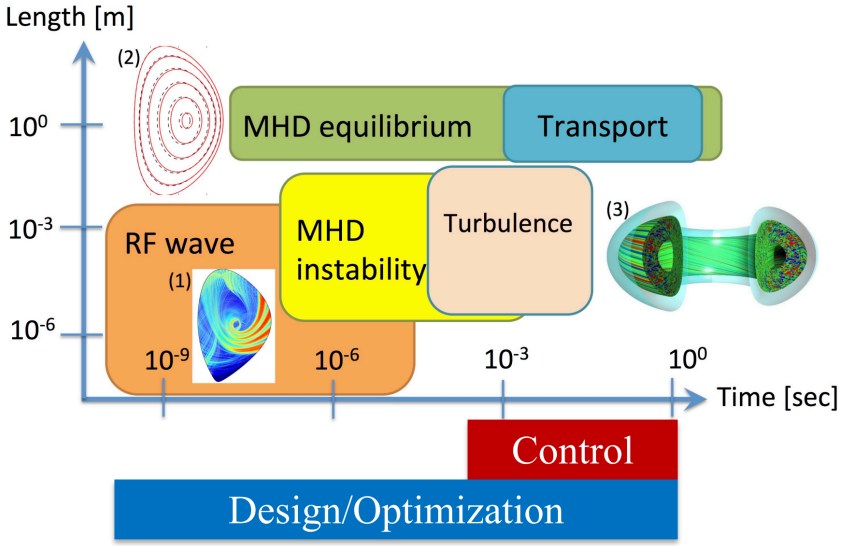
Reuters

August 7, 2023 2:07 AM CDT · Updated 3 months ago

Fermilab

Plasma control

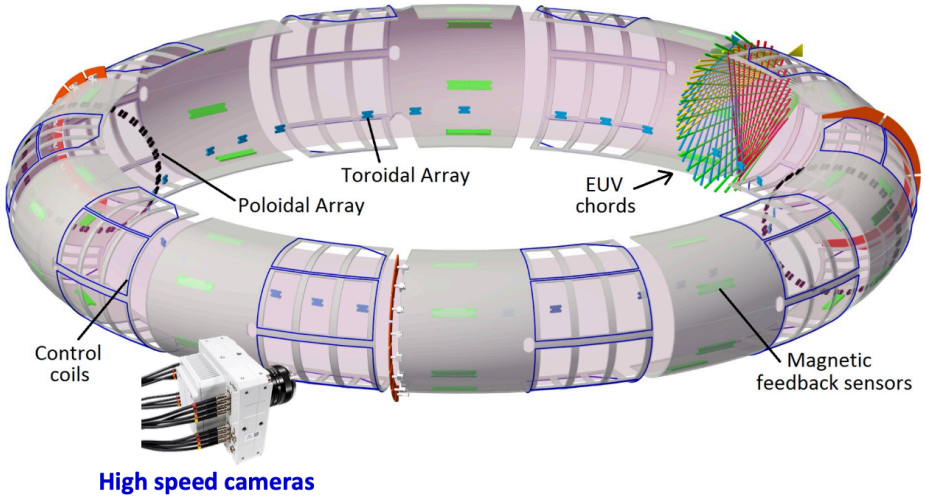


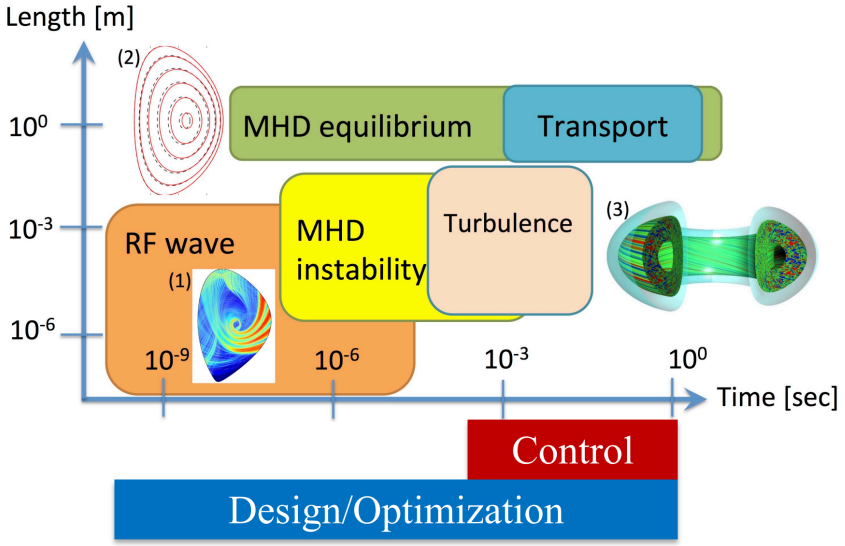


Credit: Chris Hansen

- Magneto-hydrodynamic (MHD) instabilities form when magnetic field lines become distorted, become critical on the order of microseconds
  - Leads to confinement loss on contact with vacuum chamber wall and damage to the reactor
  - One of the major roadblocks preventing **lasting** sustained fusion reactions

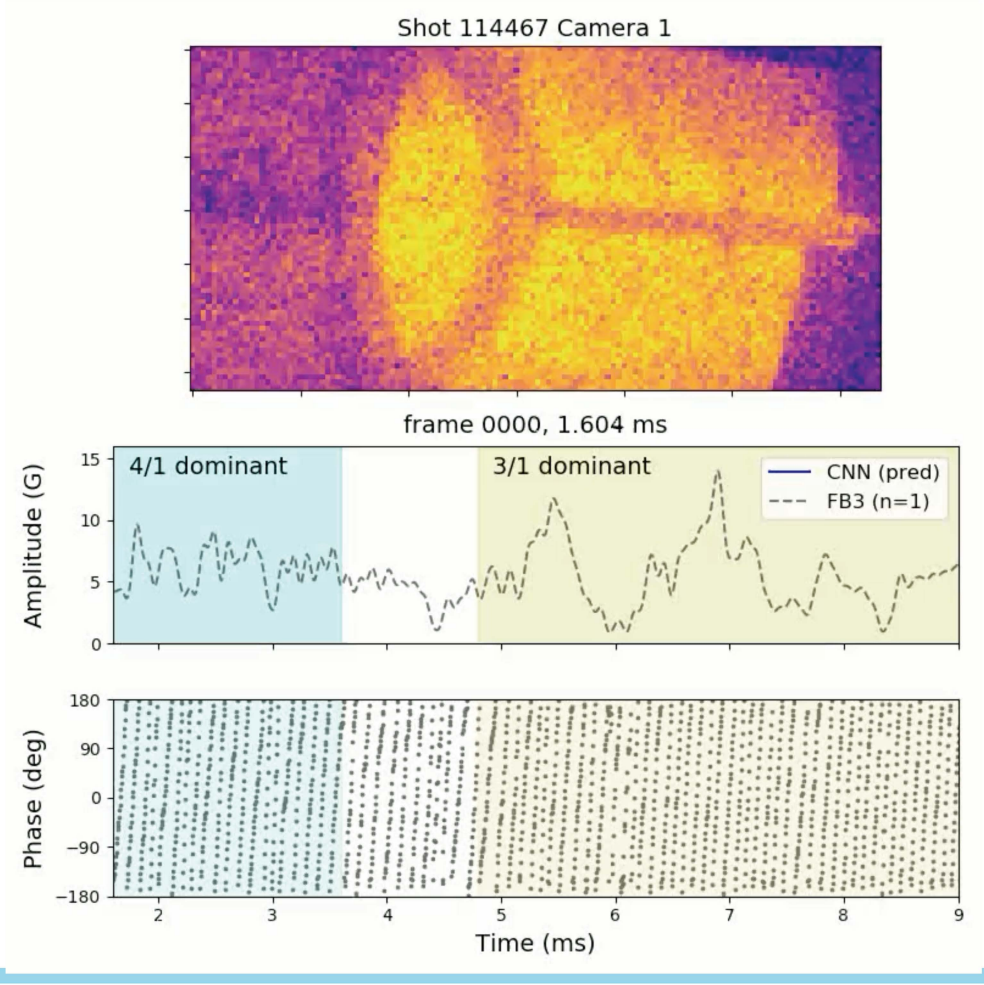
Plasma control





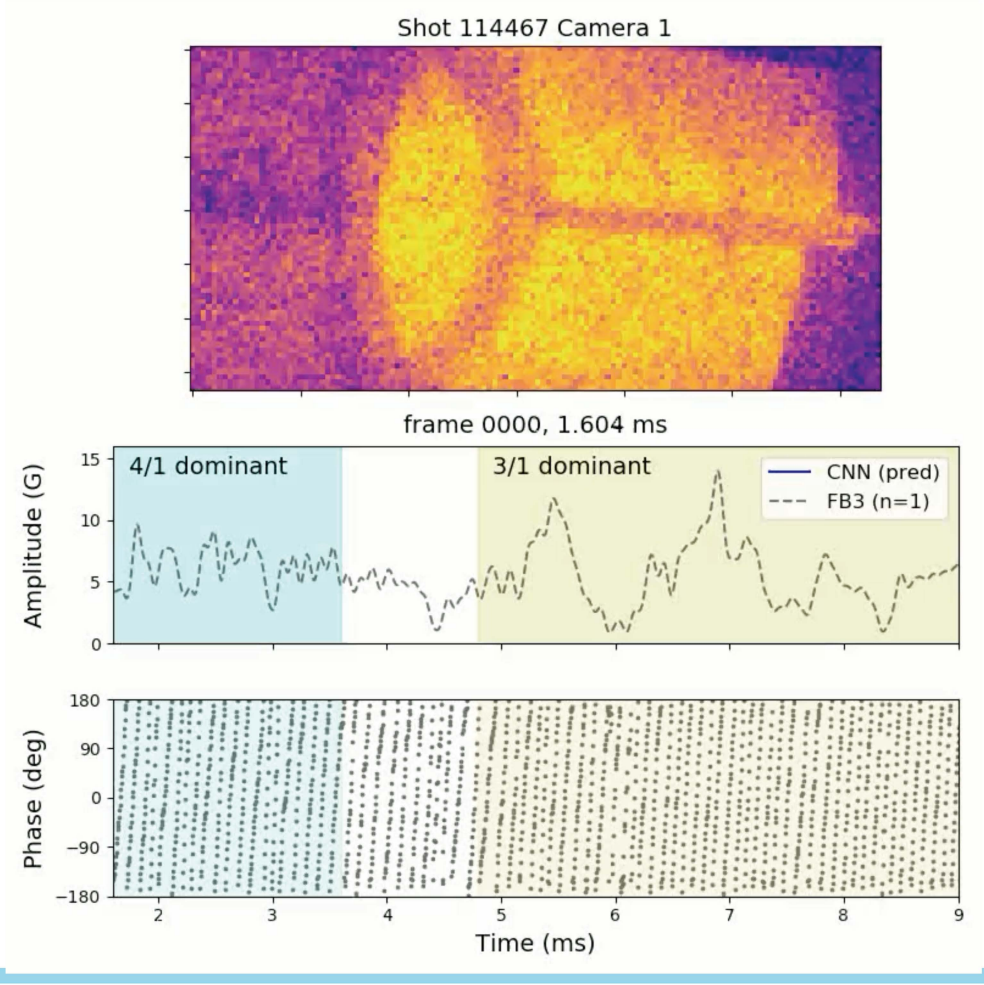
Credit: Chris Hansen

- Magneto-hydrodynamic (MHD) instabilities form when magnetic field lines become distorted, become critical on the order of microseconds
  - Leads to confinement loss on contact with vacuum chamber wall and damage to the reactor
  - One of the major roadblocks preventing **lasting** sustained fusion reactions



Credit: Yumou Wei

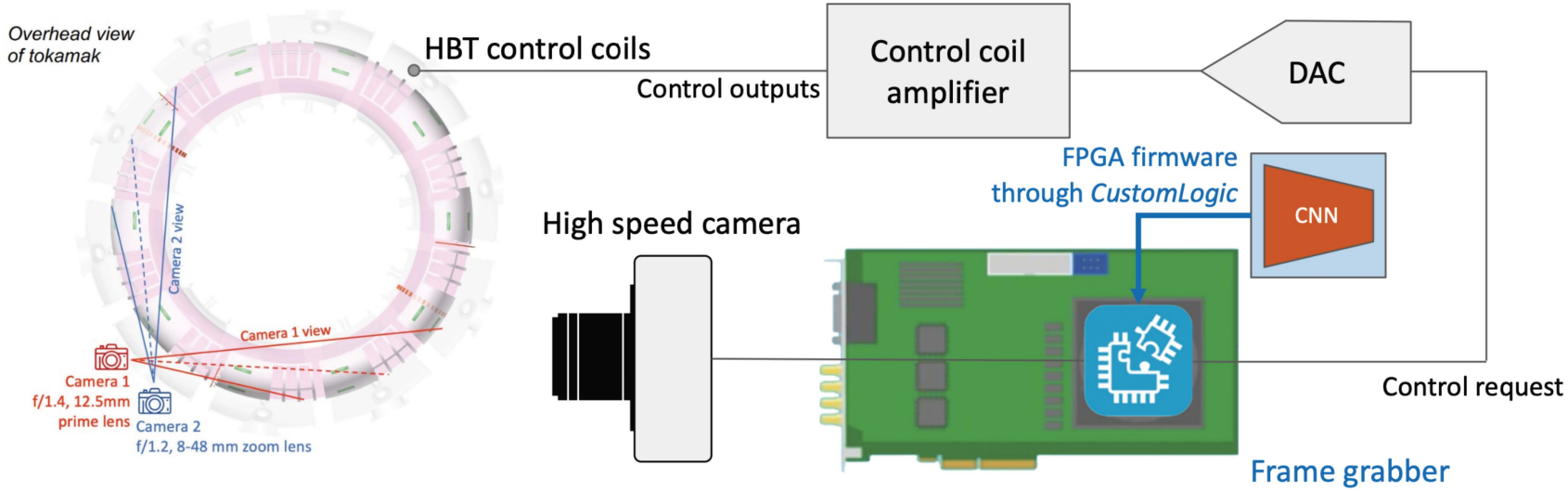
Fermilab



Credit: Yumou Wei

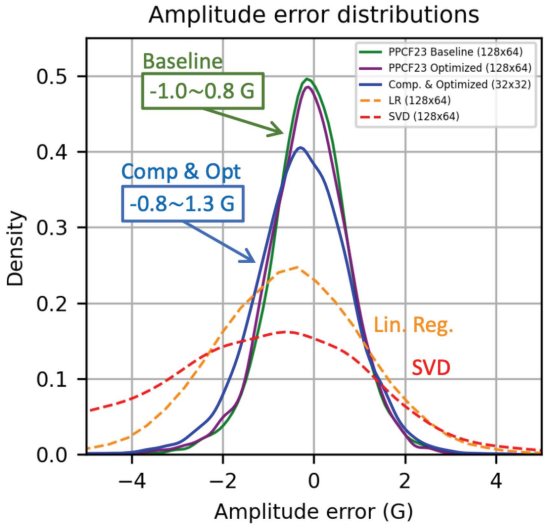
Fermilab

OTS, open-source high speed camera system



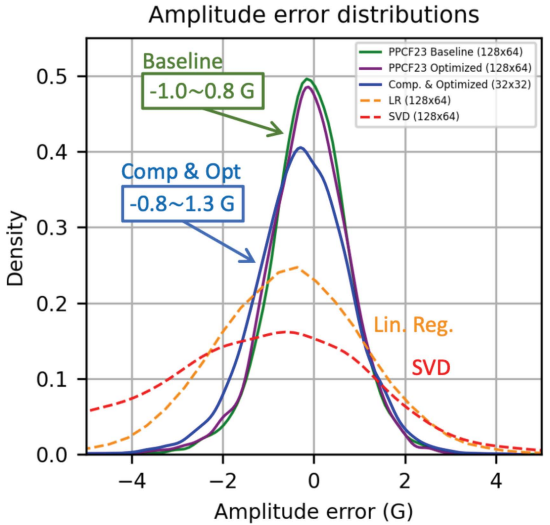
Fermilab

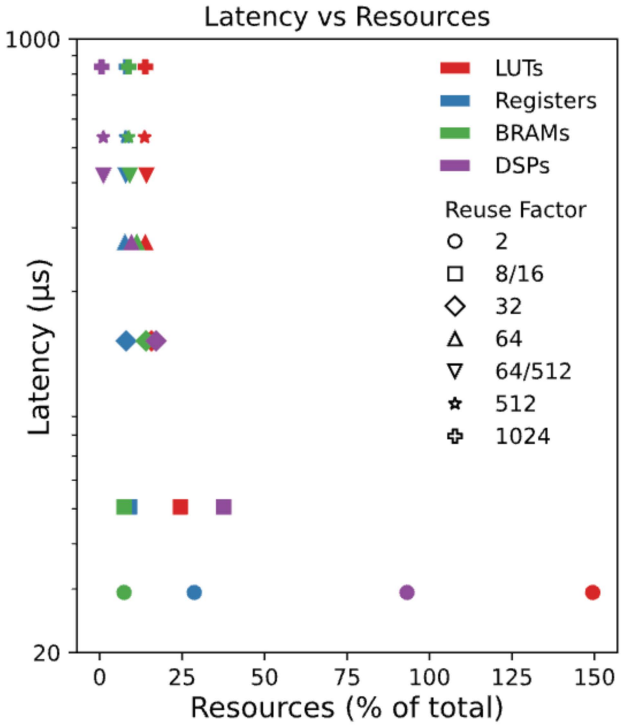
Performance and optimization



Model Name	PPCF23 Baseline	QAT+Pruning	Optimized
Image Resolution	128×64	128×64	32×32
Conv layer filters	{8,8,16}	{8,8,16}	{16,16,24}
Dense layer widths	{256,64}	{256,64}	{42,64}
Total parameters	362,730	362,730	12,910
Parameter precision	PTQ, 18 bits	QAT, 8 bits	QAT, 7 bits
Sparsity	none	80%	50%
Bit Operations	6.74e13	x	4.52e11

Performance and optimization

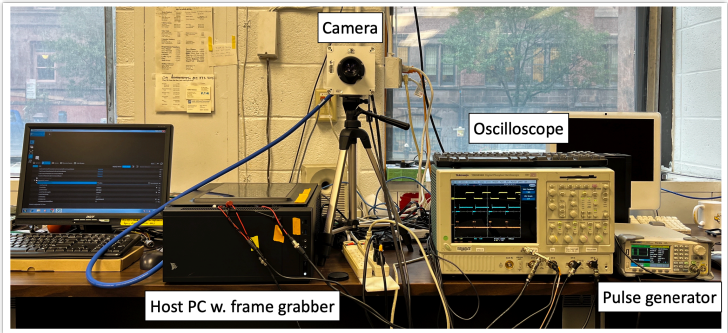




Model Name	PPCF23 Baseline	QAT+Pruning	Optimized
Image Resolution	128×64	128×64	32×32
Conv layer filters	{8,8,16}	{8,8,16}	{16,16,24}
Dense layer widths	{256,64}	{256,64}	{42,64}
Total parameters	362,730	362,730	12,910
Parameter precision	PTQ, 18 bits	QAT, 8 bits	QAT, 7 bits
Sparsity	none	80%	50%
Bit Operations	6.74e13	x	4.52e11

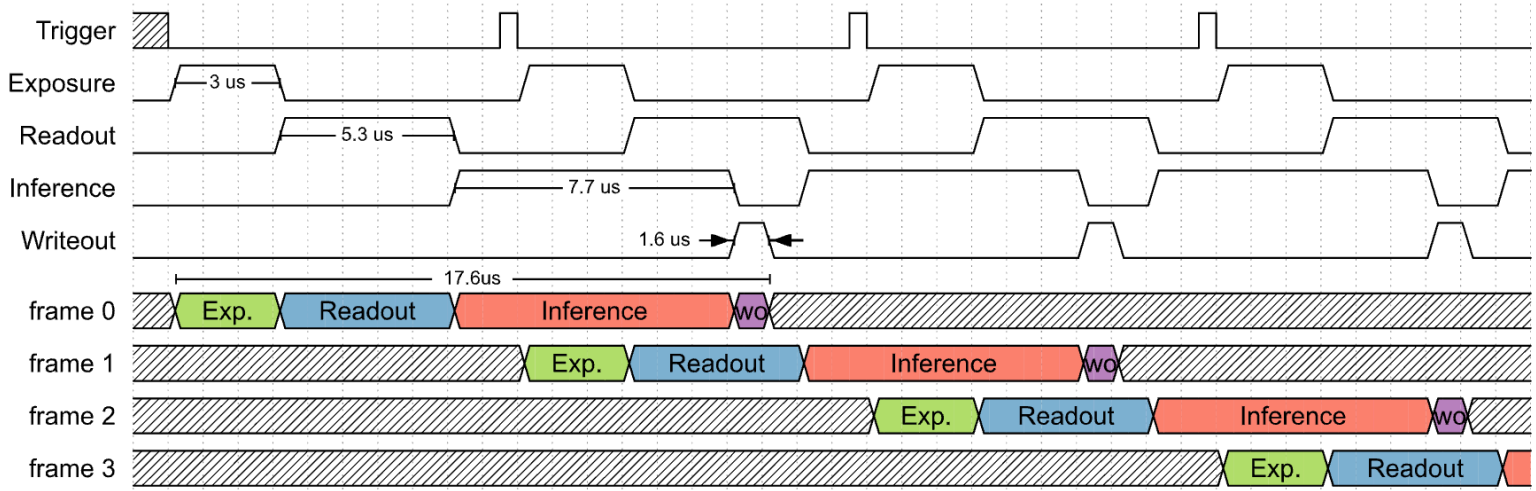
Fermilab

Prototype system



120 kfps throughput
17.6 µs latency

Enabling new capabilities for fusion experiments!



rmilab

Outline

- Why Fast ML for Science?
- The intelligent edge of tomorrow
- Towards ultra-fast automated experimentation

Outlook