Summary of Fast ML for Science Workshop

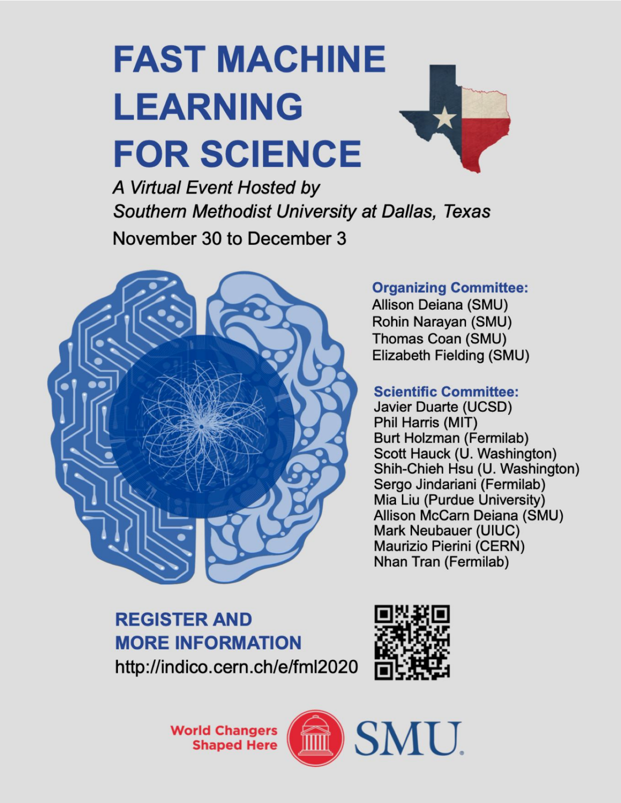
https://indico.cern.ch/event/924283/

Shih-Chieh Hsu
University of Washington

Dec 16 2020

HSF Trigger Reco Working Group

Fast Machine Learning For Science 2020



FAST MACHINE LEARNING FOR SCIENCE

A Virtual Event Hosted by
Southern Methodist University at Dallas, Texas
November 30 to December 3

Organizing Committee:
Allison Deiana (SMU)
Rohin Narayan (SMU)
Thomas Coan (SMU)
Elizabeth Fielding (SMU)

Scientific Committee:
Javier Duarte (UCSD)
Phil Harris (MIT)
Burt Holzman (Fermilab)
Scott Hauck (U. Washington)
Shih-Chieh Hsu (U. Washington)
Sergo Jindariani (Fermilab)
Mia Liu (Purdue University)
Allison McCarn Deiana (SMU)
Mark Neubauer (UIUC)
Maurizio Pierini (CERN)
Nhan Tran (Fermilab)

REGISTER AND MORE INFORMATION
http://indico.cern.ch/e/fml2020

World Changers Shaped Here SMU

History

- ML for LHC (Apr 2018) at MIT [URL]
- UltraFast DNN (Feb 2019) Zurich [URL]
- 1st Edition (Sep 2019) at FNAL [URL]
- 2nd Edition (30 Nov - 3 Dec, 2020) Virtual (this workshop)
  - Allison Deiana (LOC chair)
  - 3 days of workshops (581 registrations)
    - 47 plenary talks
  - 1 day of FastML tutorials (60 participants)

Mini-workshop on Portable Inference (Dec 4 2020)

- An IRIS-HEP Blueprint Workshop (32 participants)
- Mark Neubauer (Host)
- https://indico.cern.ch/event/972791

Nhan Tran

Why we’re here...

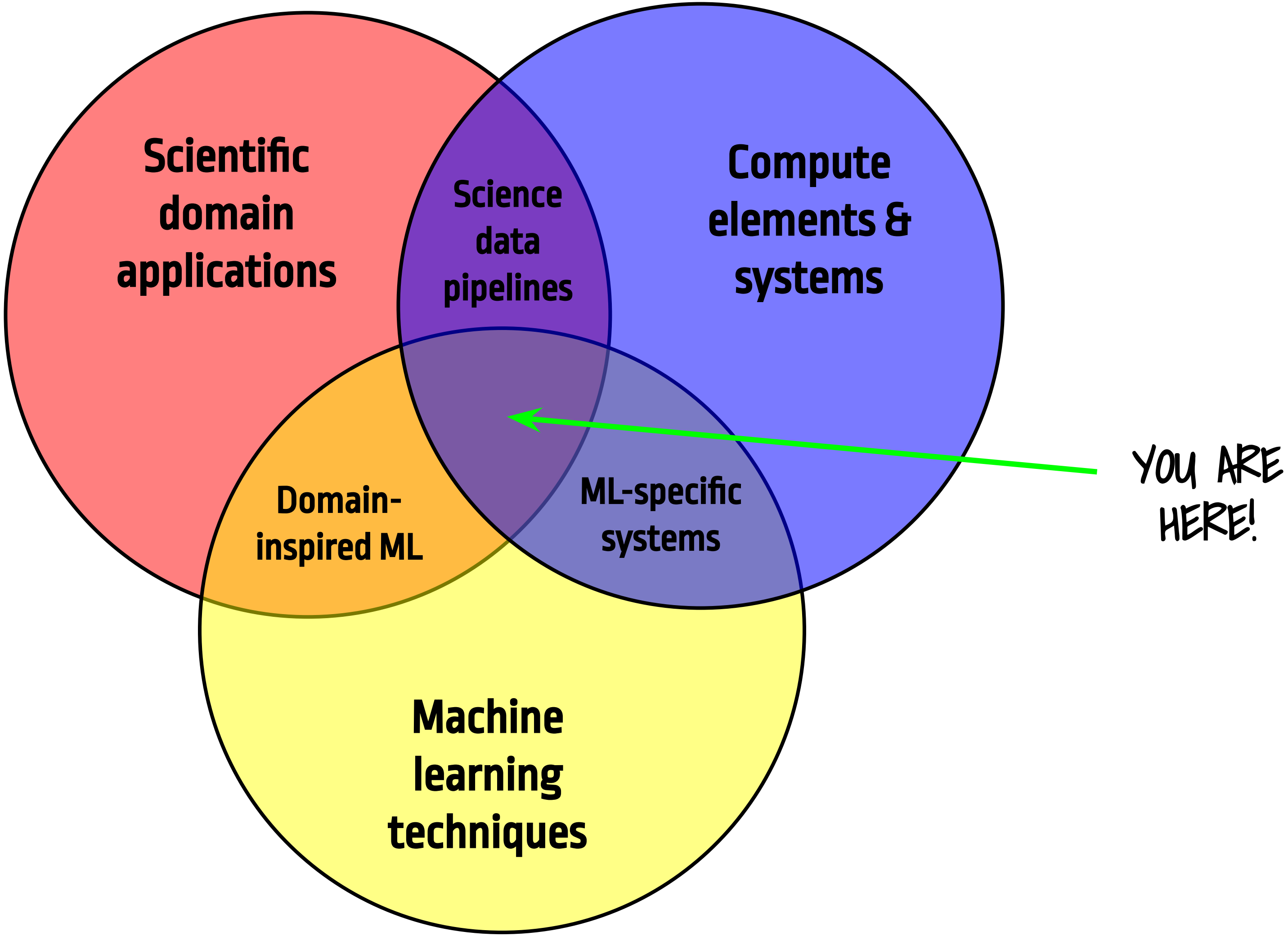
Cutting edge science requires:

Faster, more precise, bigger, more granular,...

Real-time, accelerated machine learning can greatly accelerate ***time to science***, allowing us to:

- test hypotheses significantly faster
- enhance and automate performance of detectors/accelerators
- save and maximize potentially lost data

A Multi-Disciplinary Workshop



Scientific domain applications

Compute elements & systems

Machine learning techniques

Science data pipelines

Domain-inspired ML

ML-specific systems

YOU ARE HERE!

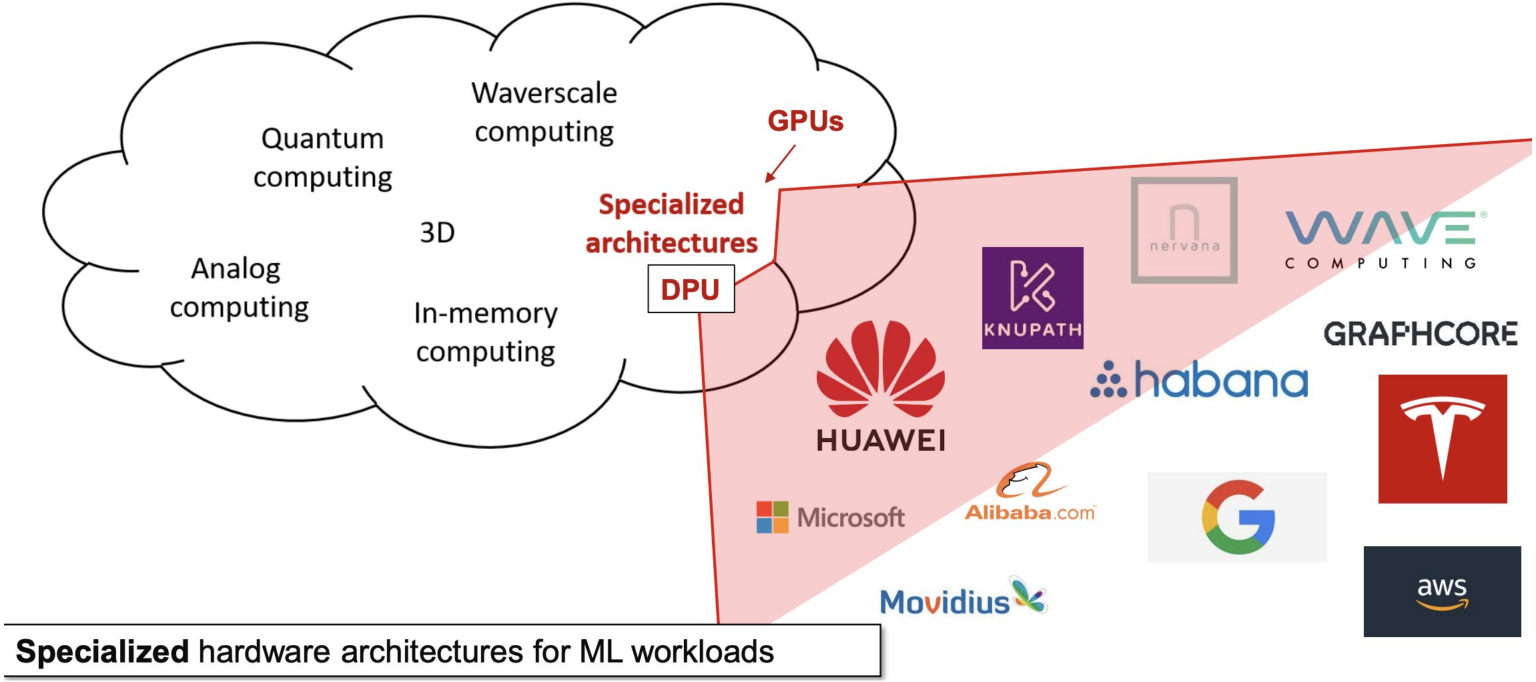
Outline

- The FastML workshop has a good mix between coprocessor and low latency
- This talk will highlight coprocessor specific contents
    - Advances in heterogeneous computing
        - Compute trends and tools
    - Applications for accelerated ML
        - Challenges in LHC physics
        - Challenges in adjacent science domains

Trends in computing architectures

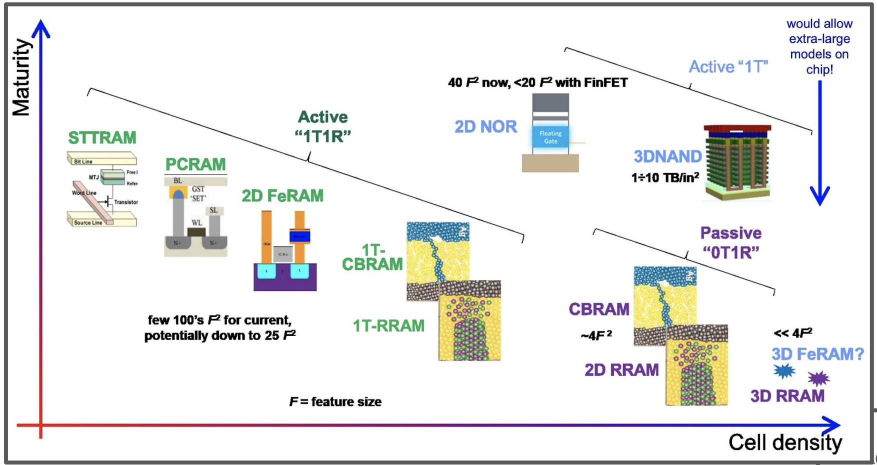
Michaela Blott

Limits of semiconductor chip development Moore's law, Denard scaling, High manufacturing cost of 3 nm/ Trends driven by big data explosion & computationally expensive ML methods.

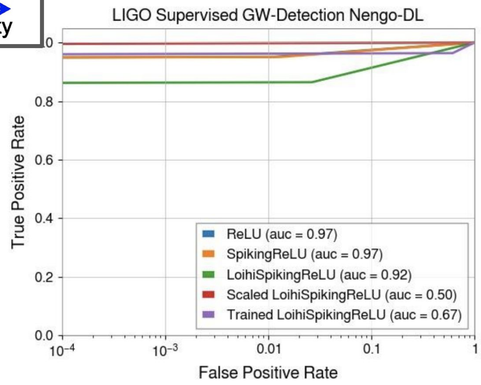


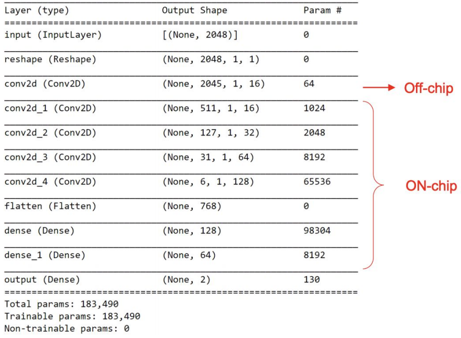
Specialized hardware architectures for ML workloads

More types of hardware



Beyond CMOS tech maturity, Dmitri Strukov





```
Layer (type)                 Output Shape              Param #
=================================================================
input (InputLayer)           [(None, 2048)]            0
_________________________________________________________________
reshape (Reshape)            (None, 2048, 1, 1)        0
_________________________________________________________________
conv2d (Conv2D)              (None, 2045, 1, 16)       64
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 511, 1, 16)        1024
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 127, 1, 32)        2048
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 31, 1, 64)         8192
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 6, 1, 128)         65536
_________________________________________________________________
flatten (Flatten)            (None, 768)               0
_________________________________________________________________
dense (Dense)                (None, 128)               98304
_________________________________________________________________
dense_1 (Dense)              (None, 64)                8192
_________________________________________________________________
output (Dense)               (None, 2)                 130
=================================================================
Total params: 183,490
Trainable params: 183,490
Non-trainable params: 0
```

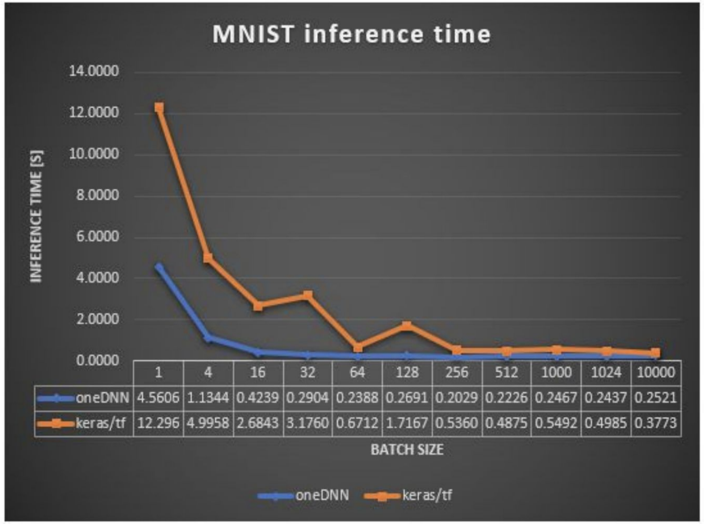
First exploration on neuromorphic chips, Bartlomiej Borzyszkowski, Eric Moreno

Executed on Loihi chip

Tools for rapid processing of ML algorithms

Intel OneAPI leads to fast CPU inference

Vladimir Loncar



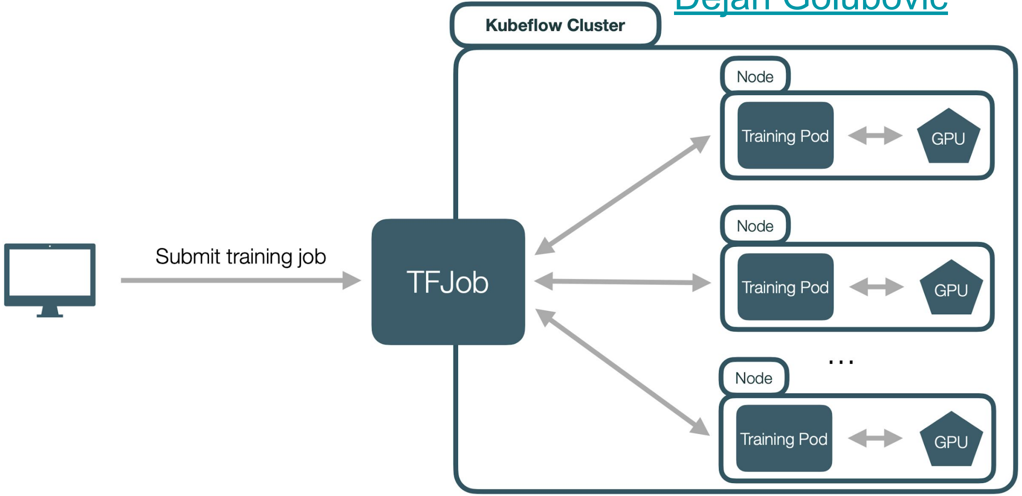
	1	4	16	32	64	128	256	512	1000	1024	10000
oneDNN	4.5606	1.1344	0.4239	0.2904	0.2388	0.2691	0.2029	0.2226	0.2467	0.2437	0.2521
keras/tf	12.296	4.9958	2.6843	3.1760	0.6712	1.7167	0.5360	0.4875	0.5492	0.4985	0.3773

Kubeflow tool for fast distributed training

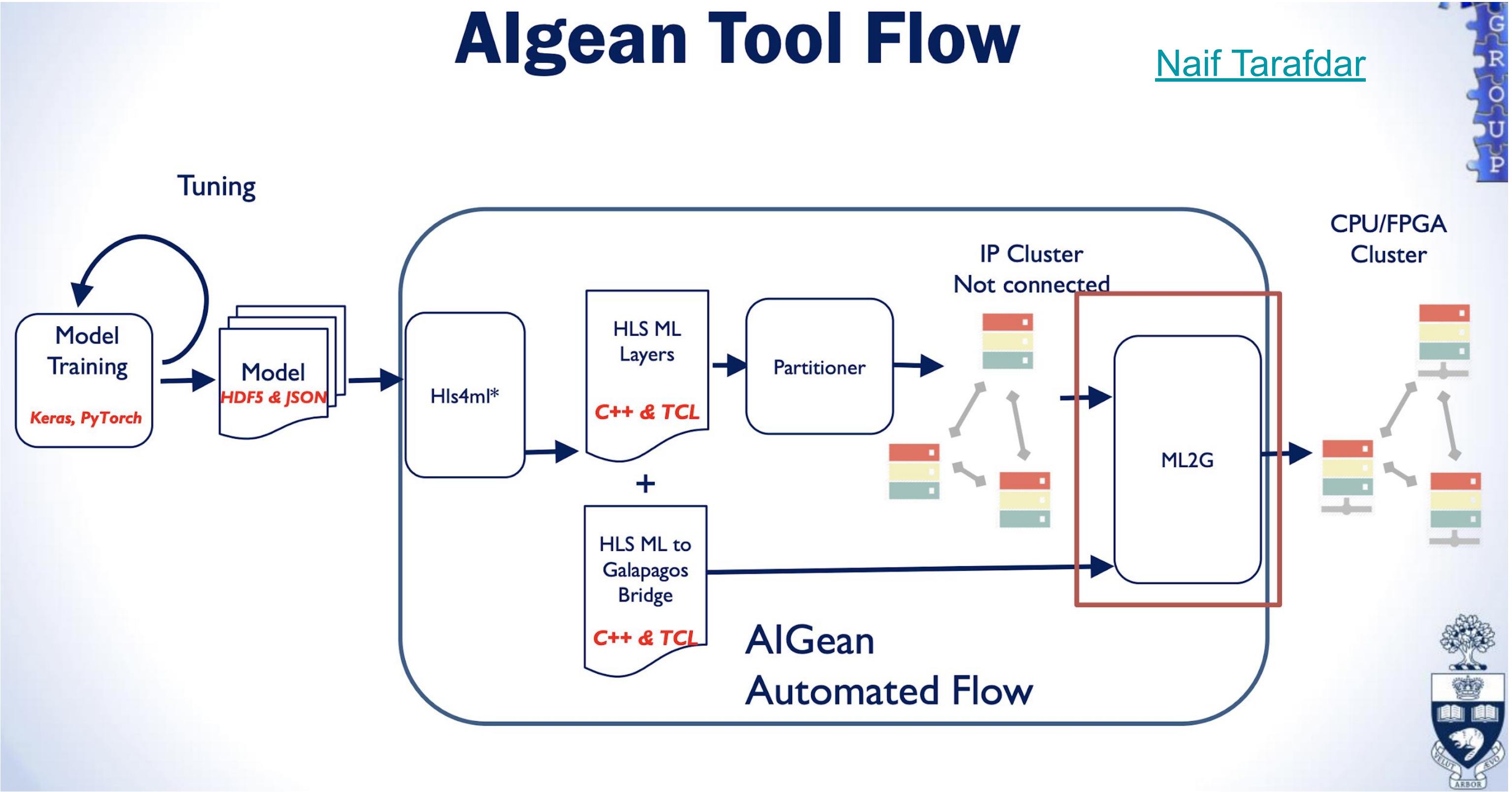
Large scale control of the system

deployed and available at CERN

Dejan Golubovic



Tools for large distribution of networks on FPGAs



Algean Tool Flow

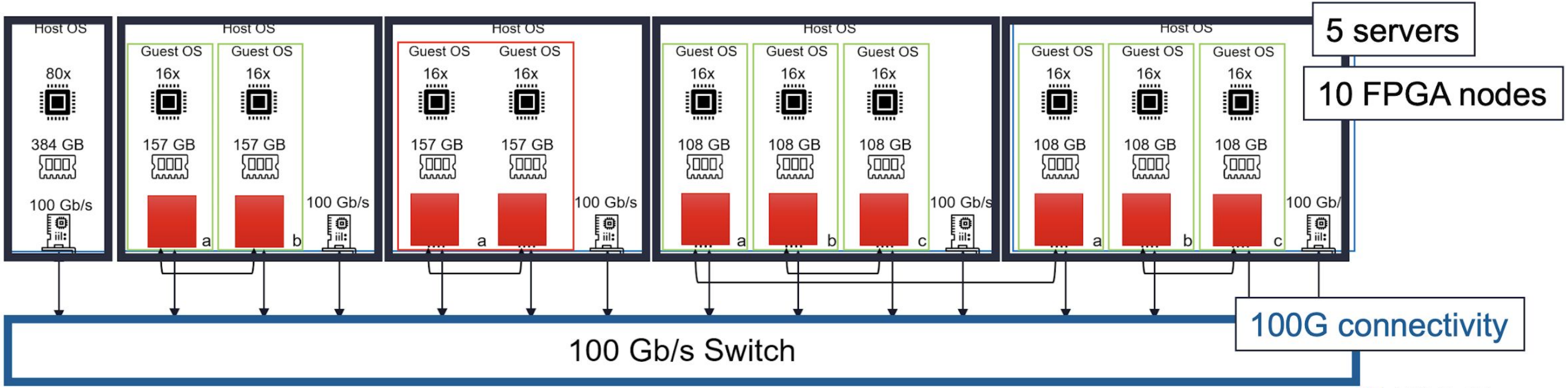
Naif Tarafdar

Tools for large distribution of networks on FPGAs

A large FPGA cluster to study algorithms across many FPGAS

FPGAs are linked with high speed connection

Leads to a very different use of HPC cluster



5 servers

10 FPGA nodes

100G connectivity

35

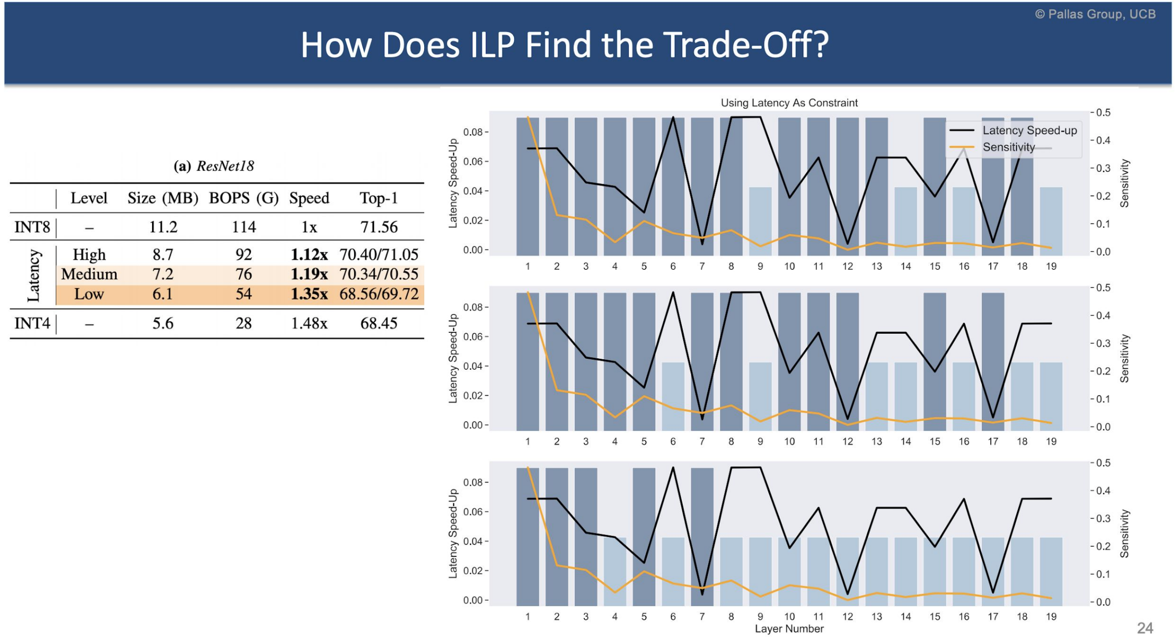
© Copyright 2020 Xilinx

XILINX

Michaela Blott, Xilinx

Efficient Design at the Edge

- What optimizations in precision and design can lead to speed-ups and power reductions?



© Pallas Group, UCB

How Does ILP Find the Trade-Off?

(a) *ResNet18*

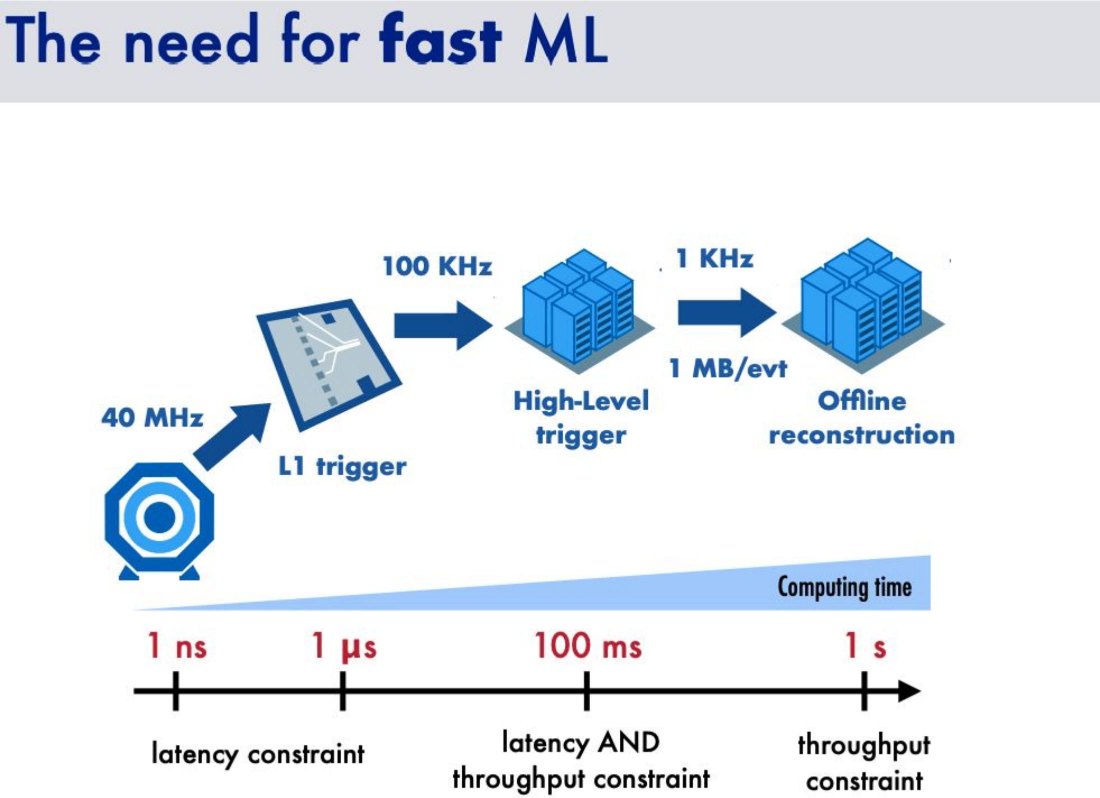
	Level	Size (MB)	BOPS (G)	Speed	Top-1
INT8	–	11.2	114	1x	71.56
Latency	High	8.7	92	**1.12x**	70.40/71.05
	Medium	7.2	76	**1.19x**	70.34/70.55
	Low	6.1	54	**1.35x**	68.56/69.72
INT4	–	5.6	28	1.48x	68.45

24

Challenges in LHC Physics

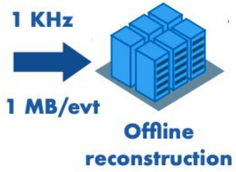
Offline Reconstruction (1 KHz, 1 s latency)

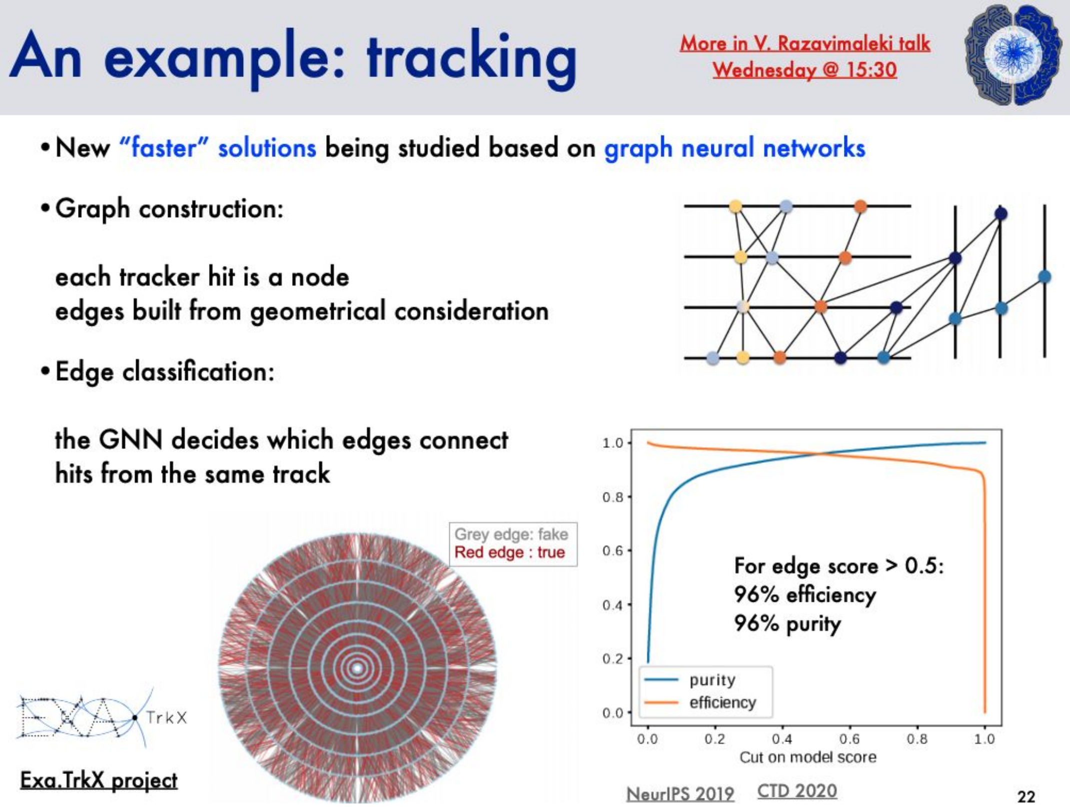
Jennifer Ngadiuba



The need for **fast** ML

Offline Reconstruction (1 KHz, 1 s latency)





An example: tracking

More in V. Razavimaleki talk
Wednesday @ 15:30

- New "faster" solutions being studied based on graph neural networks
- Graph construction:

  each tracker hit is a node
  edges built from geometrical consideration
- Edge classification:

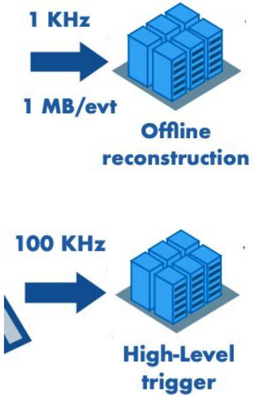
  the GNN decides which edges connect
  hits from the same track

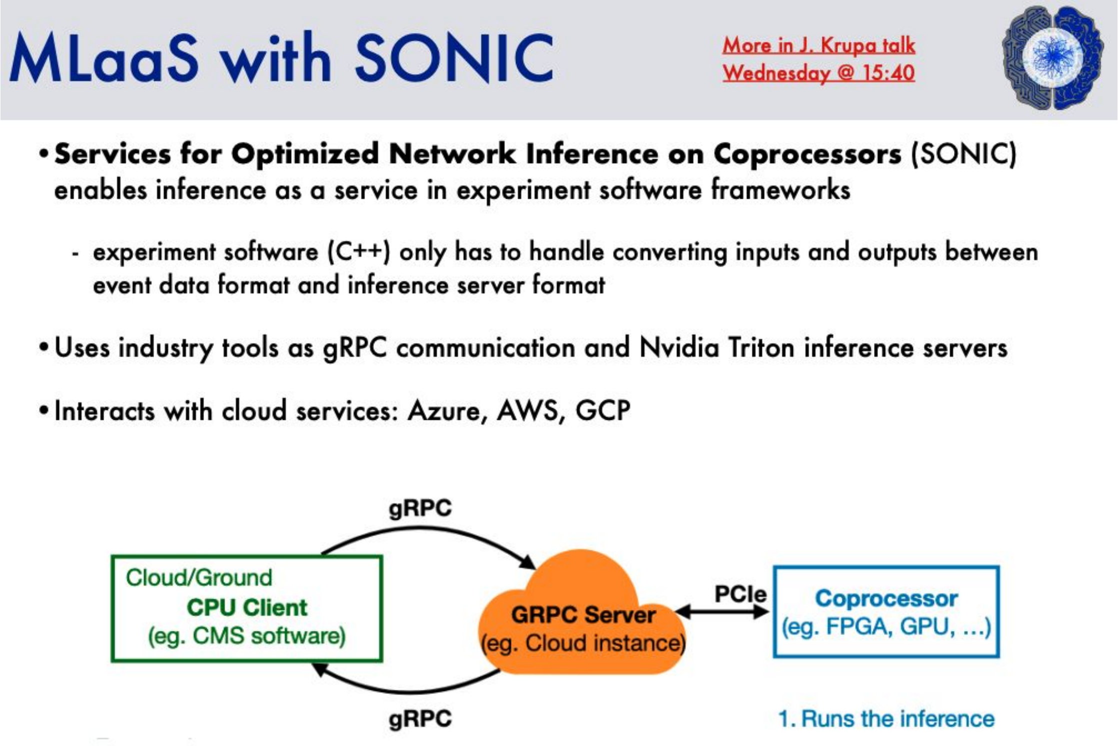
For edge score > 0.5:
96% efficiency
96% purity

Exa.TrkX project

NeurIPS 2019 CTD 2020

High-Level Trigger (100 KHz, 100 ms latency)



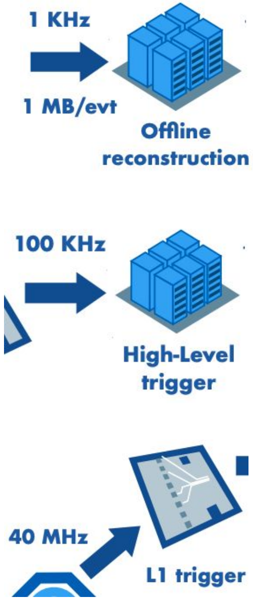


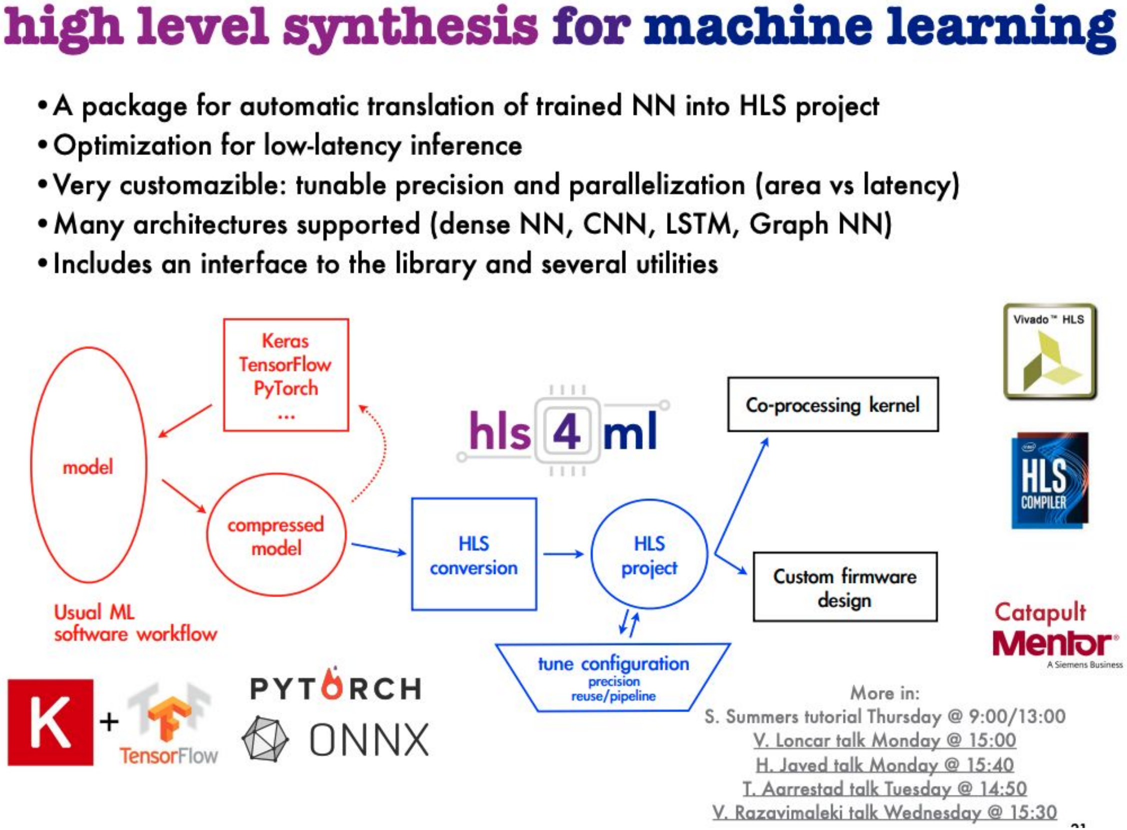
MLaaS with SONIC

More in J. Krupa talk Wednesday @ 15:40

- **Services for Optimized Network Inference on Coprocessors** (SONIC) enables inference as a service in experiment software frameworks
  - experiment software (C++) only has to handle converting inputs and outputs between event data format and inference server format
- Uses industry tools as gRPC communication and Nvidia Triton inference servers
- Interacts with cloud services: Azure, AWS, GCP

Level 1 Trigger (40 MHz, 1 μs latency)



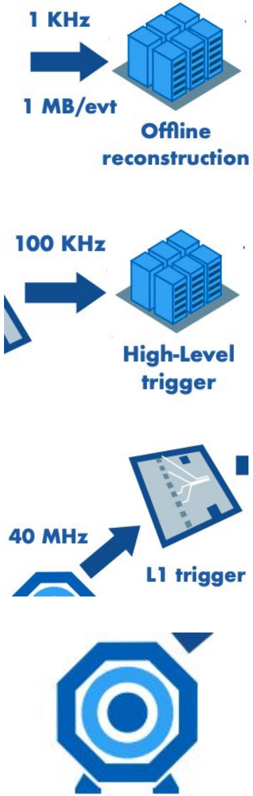


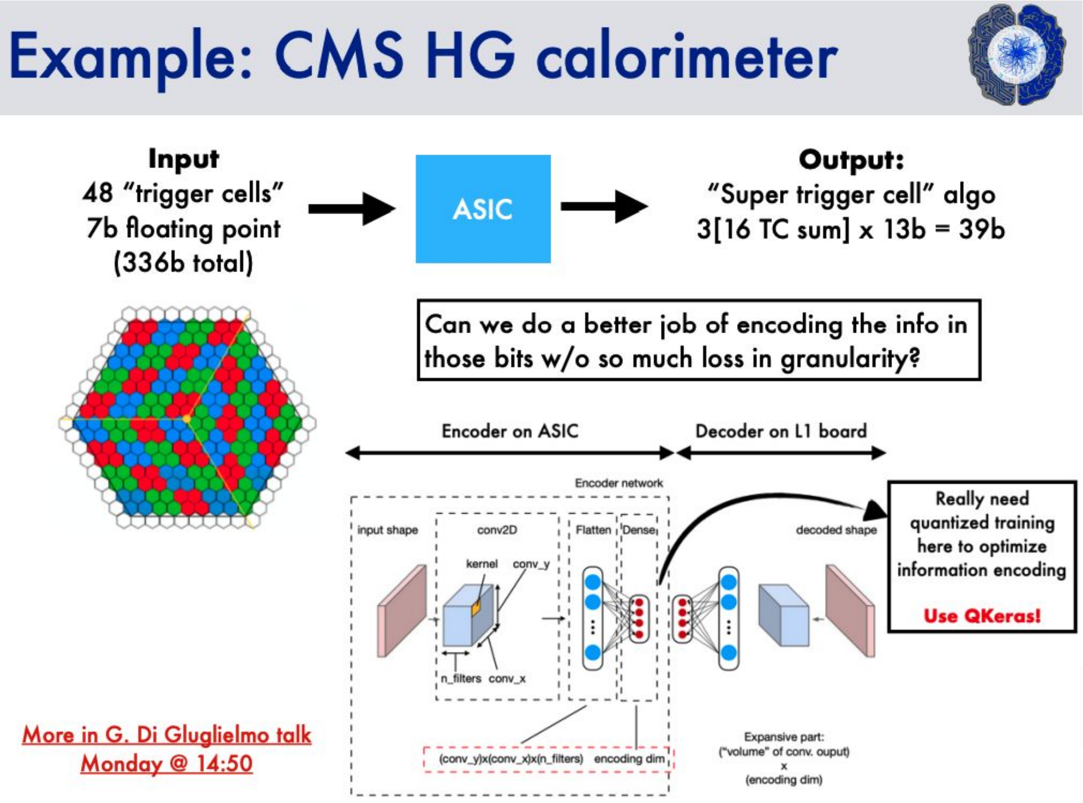
high level synthesis for machine learning

- A package for automatic translation of trained NN into HLS project
- Optimization for low-latency inference
- Very customazible: tunable precision and parallelization (area vs latency)
- Many architectures supported (dense NN, CNN, LSTM, Graph NN)
- Includes an interface to the library and several utilities

More in:
S. Summers tutorial Thursday @ 9:00/13:00
V. Loncar talk Monday @ 15:00
H. Javed talk Monday @ 15:40
T. Aarrestad talk Tuesday @ 14:50
V. Razavimaleki talk Wednesday @ 15:30

Ultra Fast (1ns)





Example: CMS HG calorimeter

Input
48 "trigger cells"
7b floating point
(336b total)

ASIC

Output:
"Super trigger cell" algo
3[16 TC sum] x 13b = 39b

Can we do a better job of encoding the info in those bits w/o so much loss in granularity?

Encoder on ASIC

Decoder on L1 board

Really need quantized training here to optimize information encoding

Use QKeras!

More in G. Di Gluglielmo talk
Monday @ 14:50

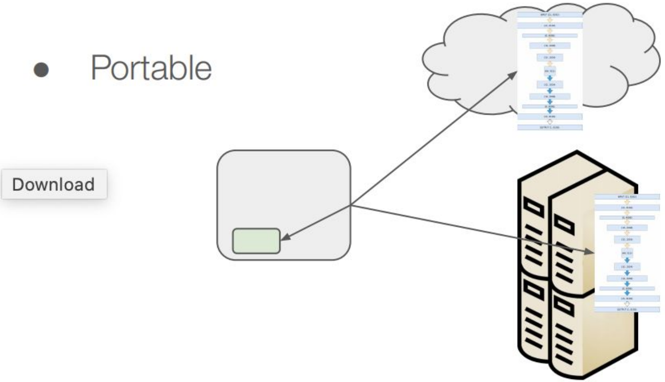
Challenges in adjacent science domains

Gravitational Waves Challenges

Inference-as-a-Service

LIGO deep clean algorithm, Alec Gunny

- Portable



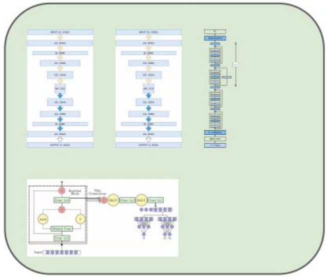
- Framework and architecture agnostic
- Critical for applications like DeepClean that need frequent retraining

Caffe2

TensorRT

TensorFlow

ONNX

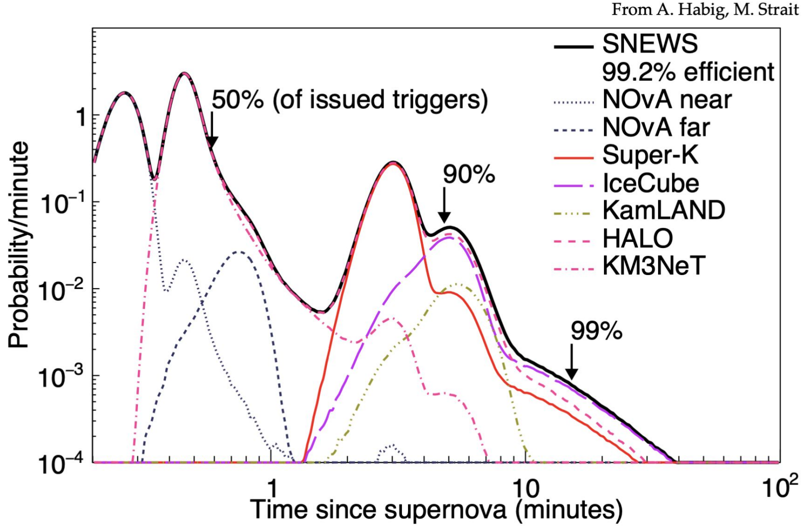


- Co-locate downstream models for better resource allocation/autoscaling
- Manage and accelerate end-to-end latency of DeepClean + downstream algorithms to meet requirements

Neutrino Astrophysics challenge

Kate Scholberg

SNEWS Alert Latency



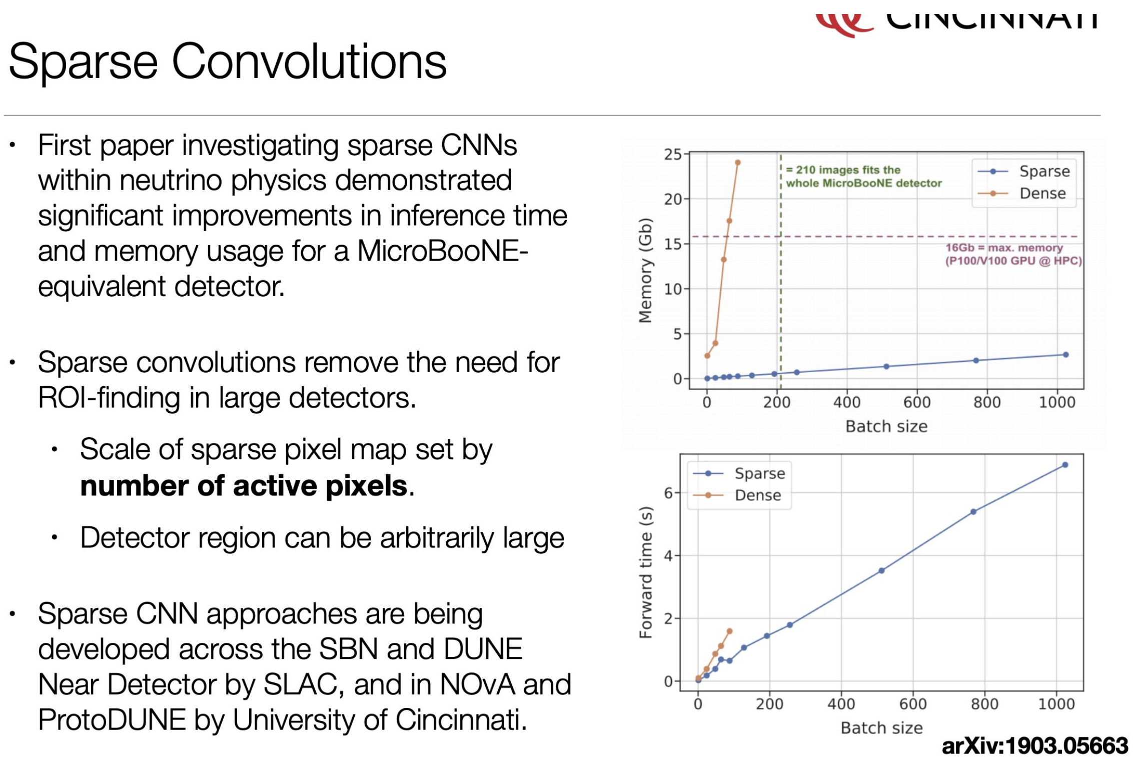
Now, the example where FastML may help!

Methods for Pointing Using Neutrinos

- ❑ **Anisotropic neutrino interactions**
  combined with detector technology that can exploit it, using the burst neutrino signal
- ❑ **Triangulation**
  using inter-detector timing
- ❑ **Oscillation pattern pointing**
  in high-energy resolution detectors
- ❑ **High-energy (~GeV) neutrino follow-on pointing**
  in directional detectors, using later neutrinos
- ❑ **All of the above!**

Accelerator Neutrinos reading out and processing

Jeremy Hewes



Sparse Convolutions

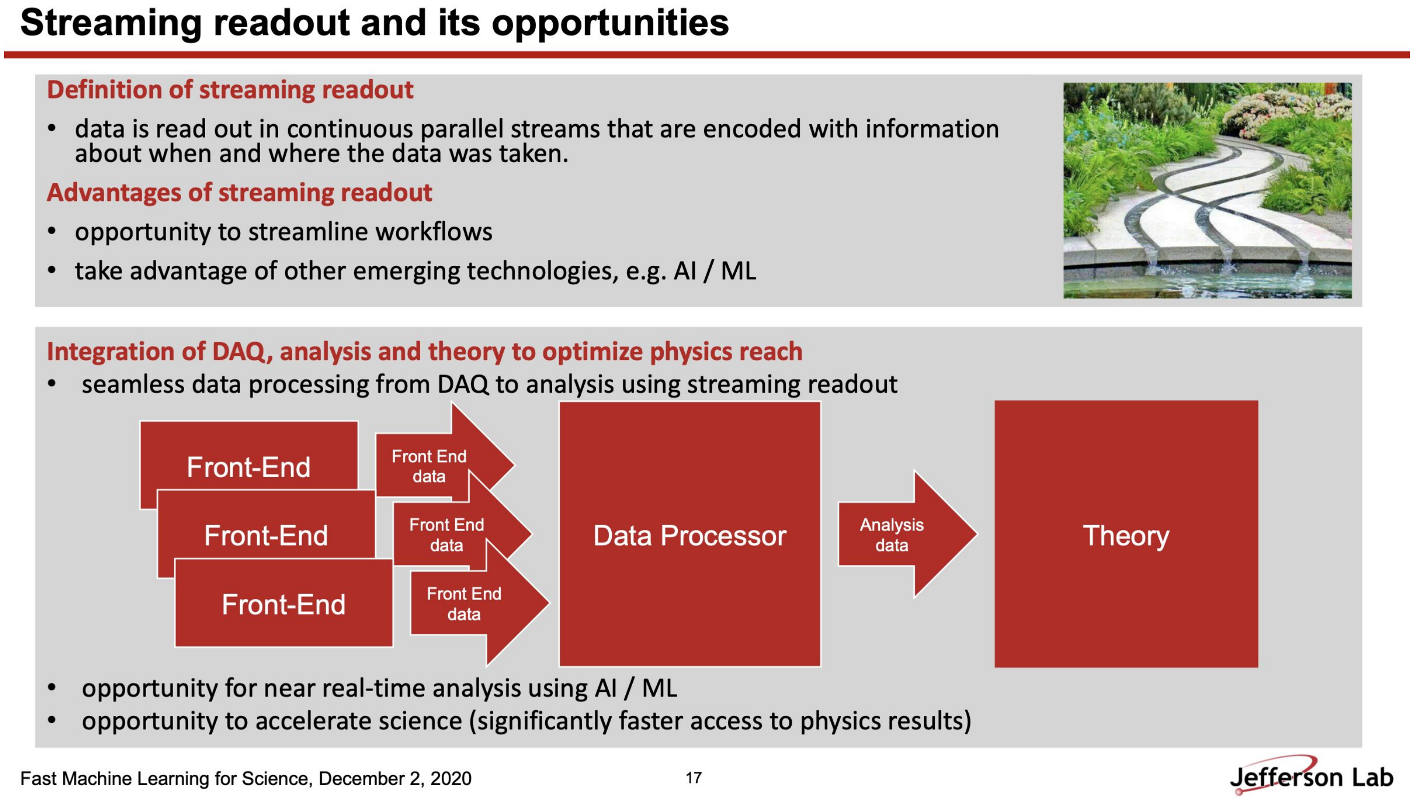
- First paper investigating sparse CNNs within neutrino physics demonstrated significant improvements in inference time and memory usage for a MicroBooNE-equivalent detector.
- Sparse convolutions remove the need for ROI-finding in large detectors.
    - Scale of sparse pixel map set by **number of active pixels**.
    - Detector region can be arbitrarily large
- Sparse CNN approaches are being developed across the SBN and DUNE Near Detector by SLAC, and in NOvA and ProtoDUNE by University of Cincinnati.

arXiv:1903.05663

Computational challenge in Accelerator Neutrinos
Are quickly becoming large

Challenges at Electron Ion Collider (EIC)

Markus Diefenthaler



Streaming readout and its opportunities

Definition of streaming readout

- data is read out in continuous parallel streams that are encoded with information about when and where the data was taken.

Advantages of streaming readout

- opportunity to streamline workflows
- take advantage of other emerging technologies, e.g. AI / ML

Integration of DAQ, analysis and theory to optimize physics reach

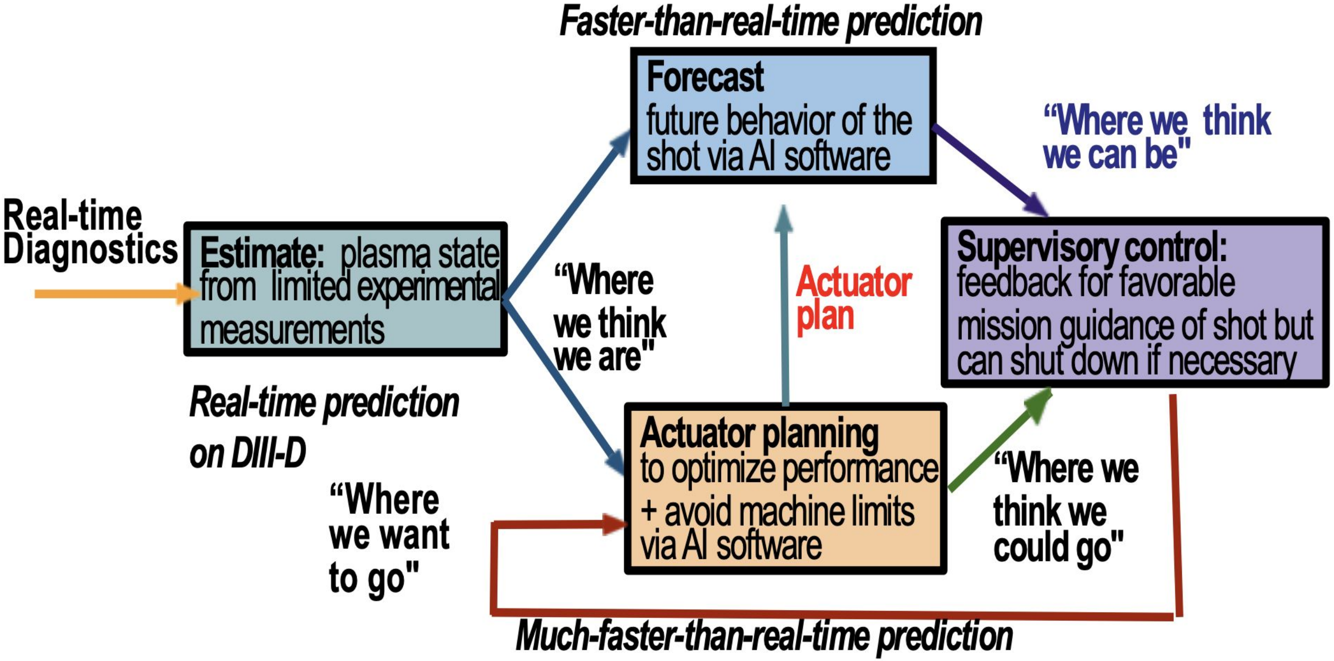
- seamless data processing from DAQ to analysis using streaming readout

- opportunity for near real-time analysis using AI / ML
- opportunity to accelerate science (significantly faster access to physics results)

Plasma Physics: From Prediction to Control

Bill Tang

- How do we control a plasma?



Faster-than-real-time prediction

Forecast future behavior of the shot via AI software

"Where we think we can be"

Real-time Diagnostics

Estimate: plasma state from limited experimental measurements

"Where we think we are"

Actuator plan

Supervisory control: feedback for favorable mission guidance of shot but can shut down if necessary

Real-time prediction on DIII-D

Actuator planning to optimize performance + avoid machine limits via AI software

"Where we think we could go"

"Where we want to go"

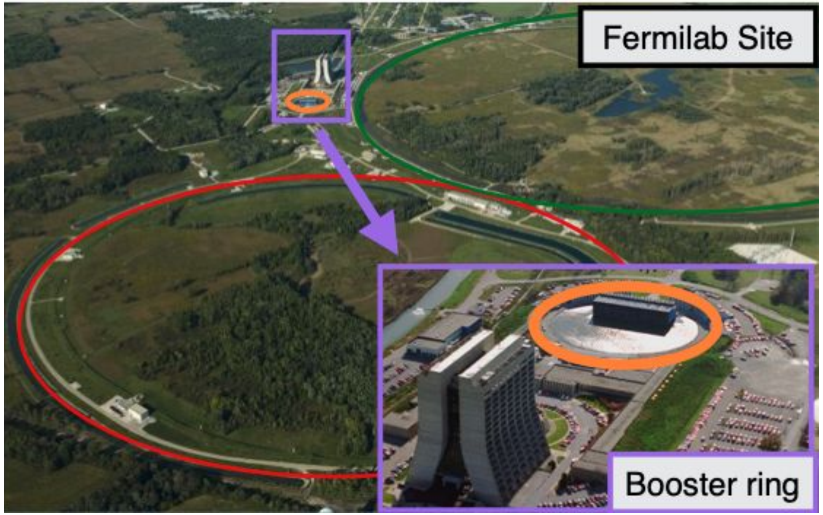
Much-faster-than-real-time prediction

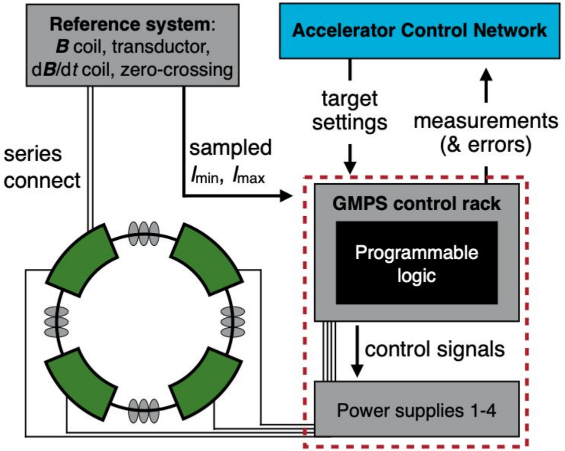
Accelerator Control

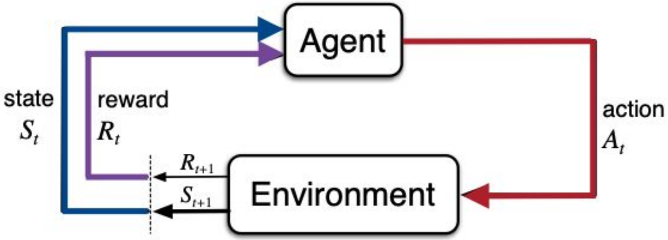
- How do we control proton accelerator?

Christian Herwig

Booster: 400 MeV—> 8 GeV



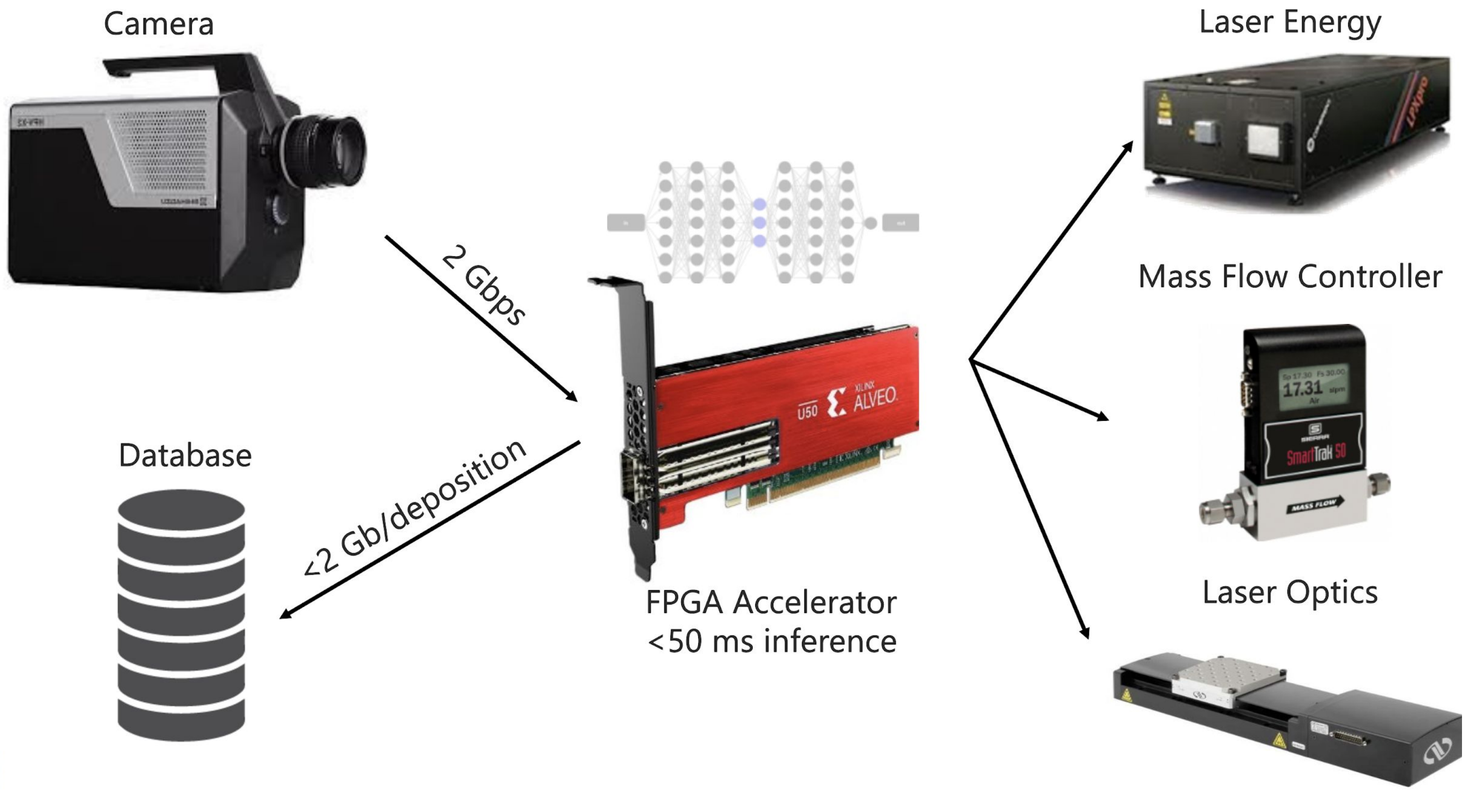




Material Science

Materials Science, Josh Agar

Processing Large amounts of data in near real-time



Camera

2 Gbps

Database

<2 Gb/deposition

FPGA Accelerator
<50 ms inference

Laser Energy

Mass Flow Controller

Laser Optics

Conclusions

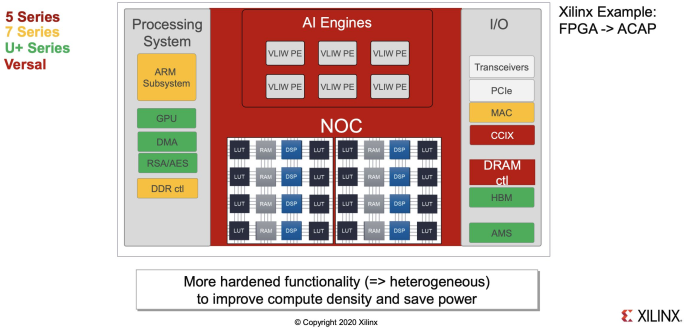
New developments in computing allow for larger use of computing

- New tools are emerging to allow for more advanced applications
  - Tools for large scale use of GPUs, FPGAs
  - Tools to improve algorithm design to reduce latency and resources
- New opportunities for use of fast ML in design
  - Real-time readout and processing in LHC collider physics
  - Real-time readout in neutrino physics, astrophysics and gravitational wave
  - Real-time control and operations in accelerator, plasma and materials science
- Supplement workshop with forward-looking white paper to
  - Establish key applications and identify overlaps between domains scientific applications for resource-constrained ML (tentative timeline Feb 15 2021)

Backup

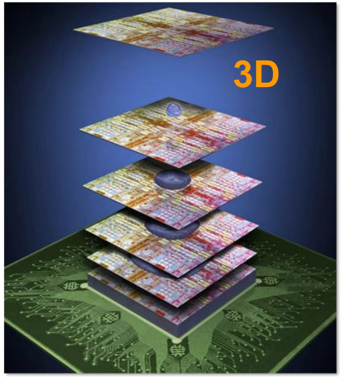
Trends in Computer Architecture

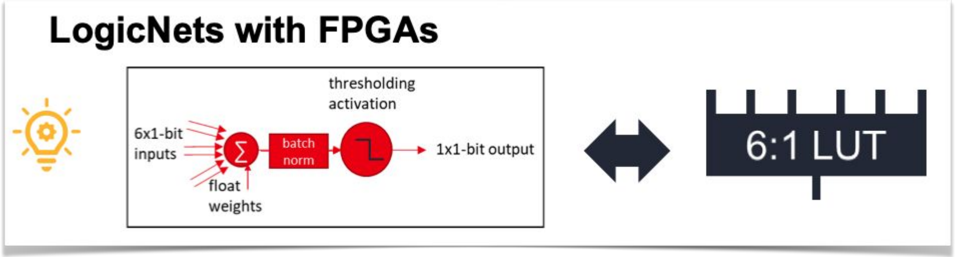
Specialization = > Increasingly Heterogeneous

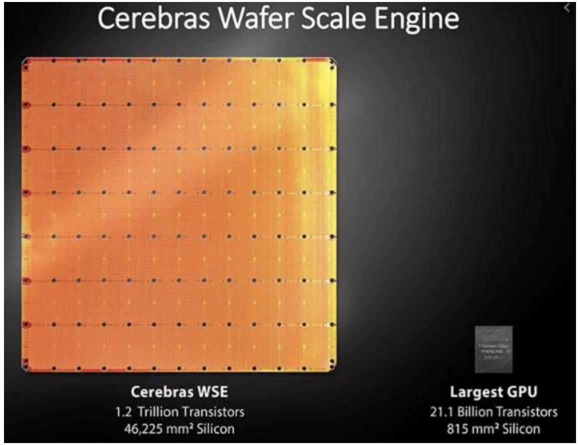


Xilinx Example:
FPGA -> ACAP

More hardened functionality (=> heterogeneous)
to improve compute density and save power

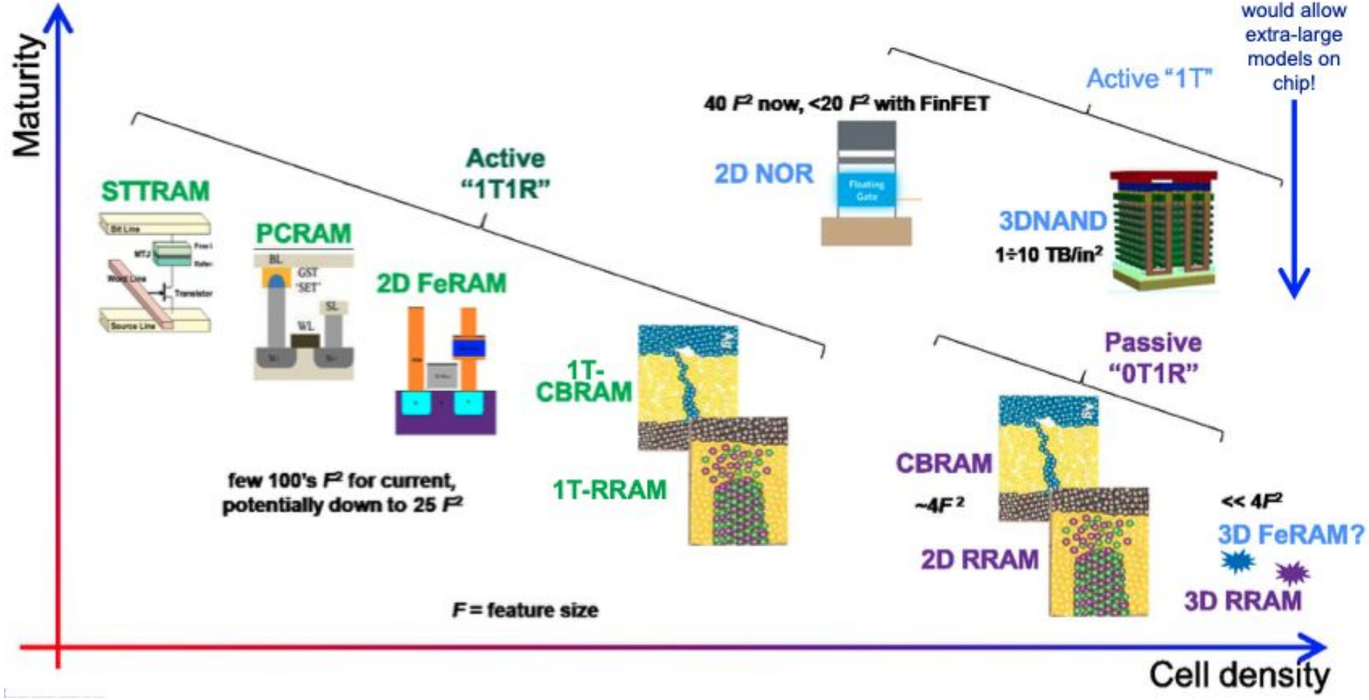






Source: HotChips2019

Beyond CMOS: Neuromophic Computer



Dmitri Strukov (UCSB)

Analog Circuit: natural fit for vector*matrix multiplications