AI On Chip

Farah Fahim
Fermilab, Northwestern University

Content

Co-design Strategy

Preliminary Demonstrator for HL LHC : Data Concentrator ASIC for HG Cal

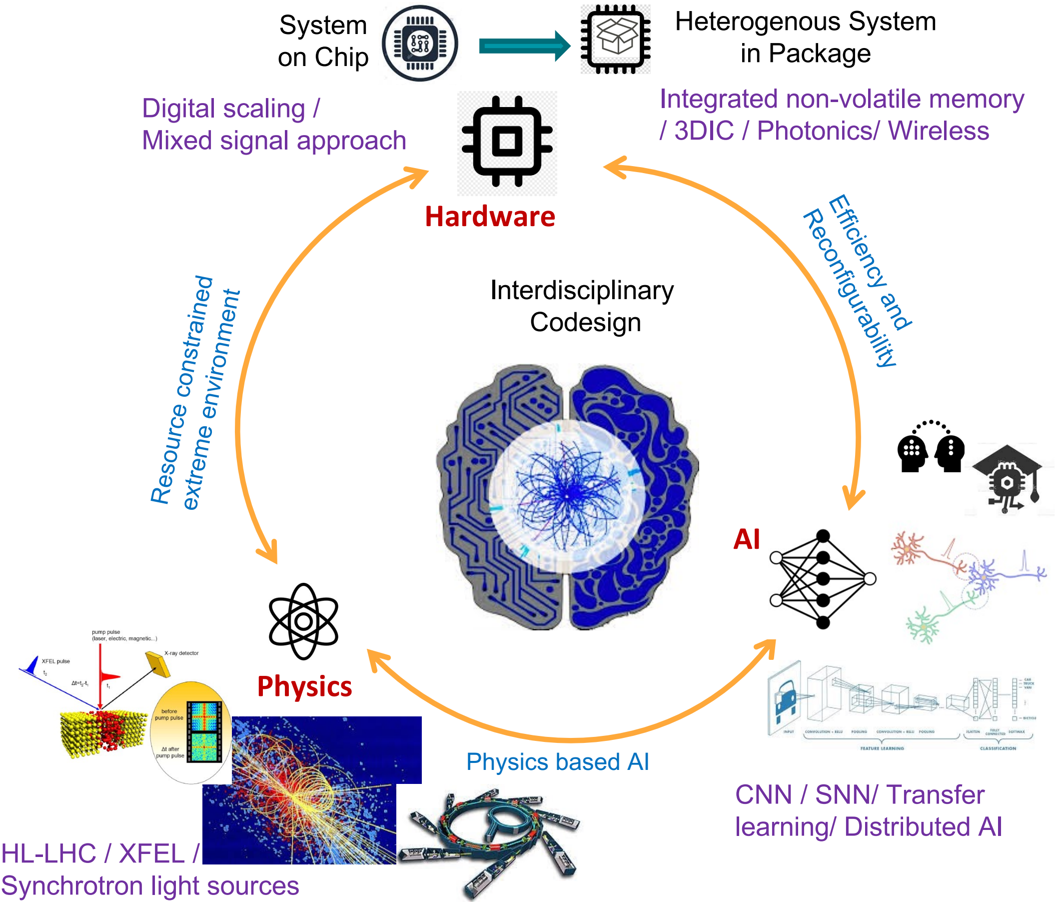
- Establish feasibility
- Establish methodology

Moving Intelligence to the data source (Ideas to be demonstrated)

- Neuromorphic Digital implementation
- Analog implementation using floating gates
- Analog implementation using memristive cross-bar arrays

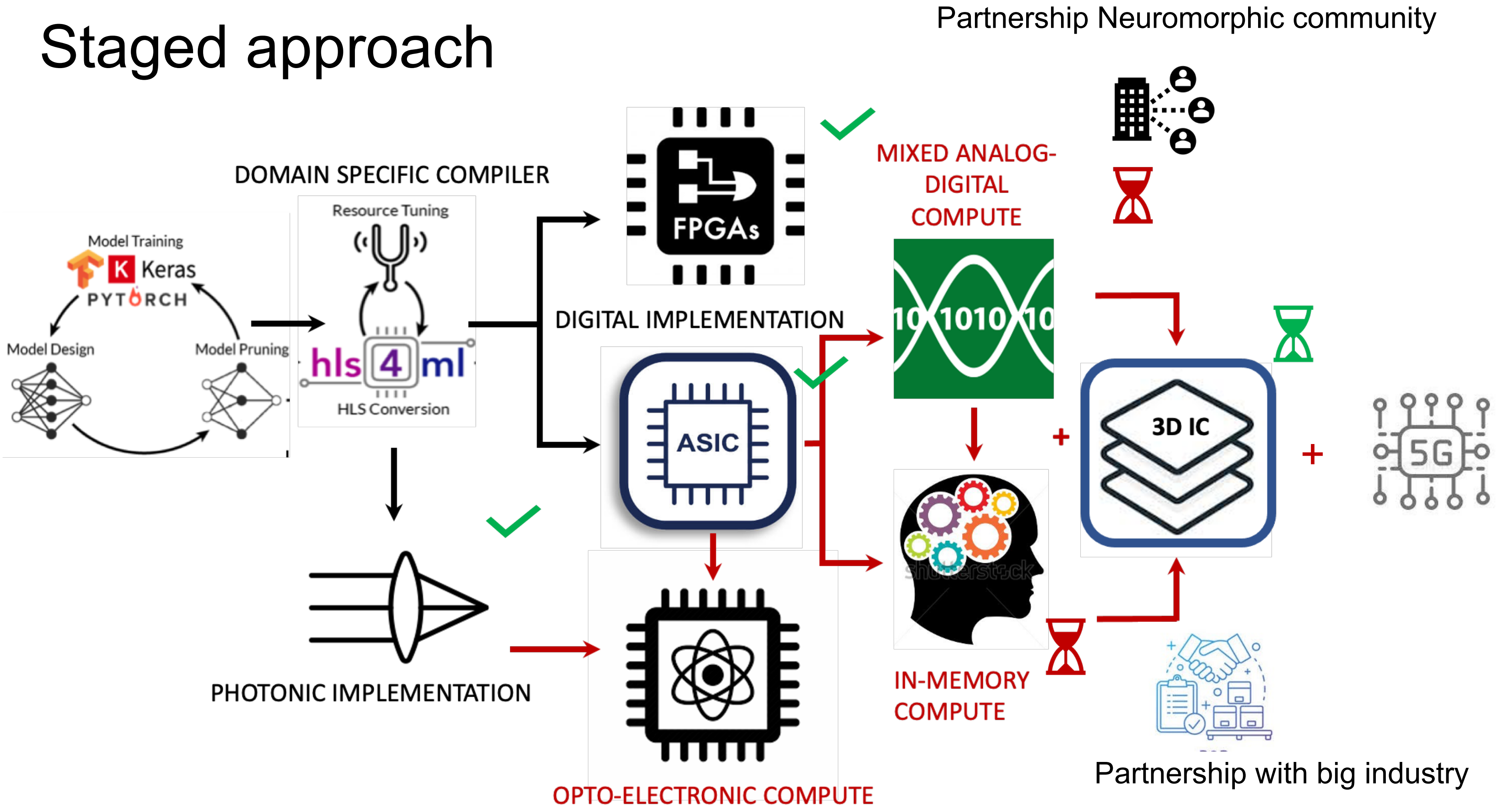
Creating a distributed AI model

On-detector intelligence using on-chip Machine Learning

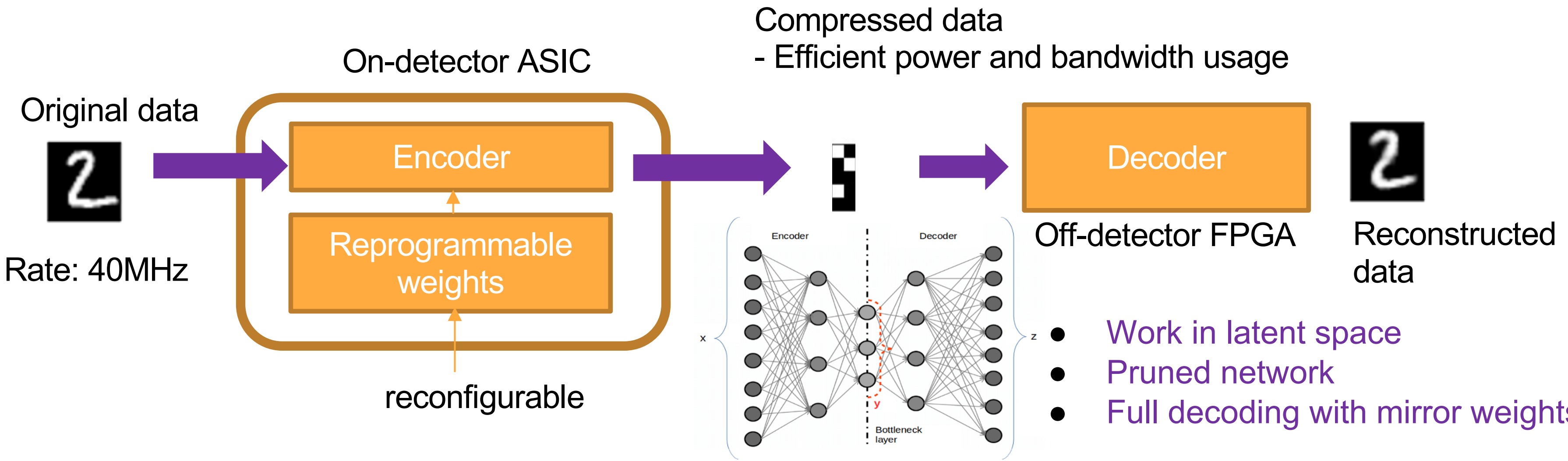


- Resource constrained environment
  - High Radiation
  - Limited Power/Material budget
  - Where should this intelligence be added
- Efficiency and reconfigurability
  - Ultra-low energy per inference at extremely high rates (10's ns)
  - Reprogram both network and parameters
  - On-chip learning / inference
- Physics based Algorithms
  - Independent events
  - Depth vs. classification

Staged approach



Deep Neural Network: Autoencoder for data-compression



- Work in latent space
- Pruned network
- Full decoding with mirror weights

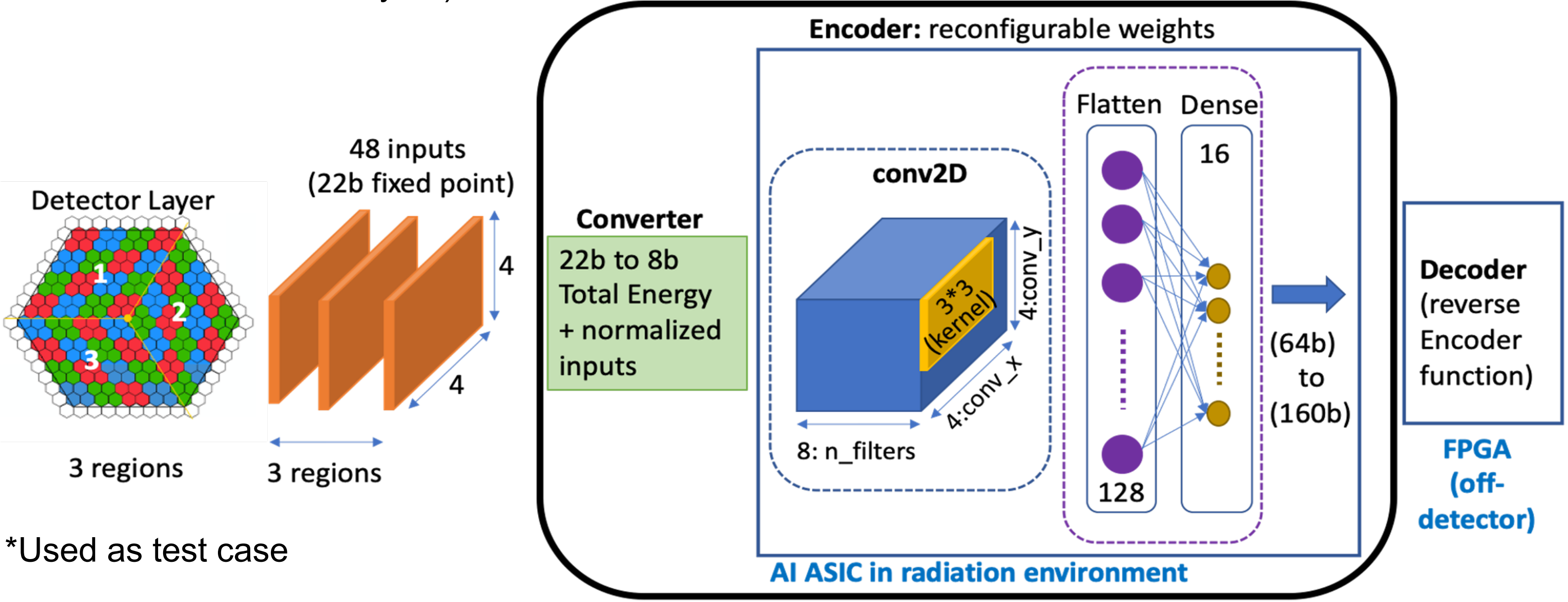
- Enable edge compute : Data compression for efficient usage of power and bandwidth
- Programmable and Reconfigurable: ability to reprogram weights to adjust for detector conditions and eventually lead to self-learning intelligent detectors
- Hardware – Software codesign : Algorithm driven architectural approach
- Optimized : Low power and Low latency
- Operating in extreme radiation environment: 200 M rad

- Autoencoder for data compression is the first use case towards a DNN based on-chip learning and inference

HL LHC High Granularity Calorimeter*: Data flow

CNN: Encodes information by correlating spatial features

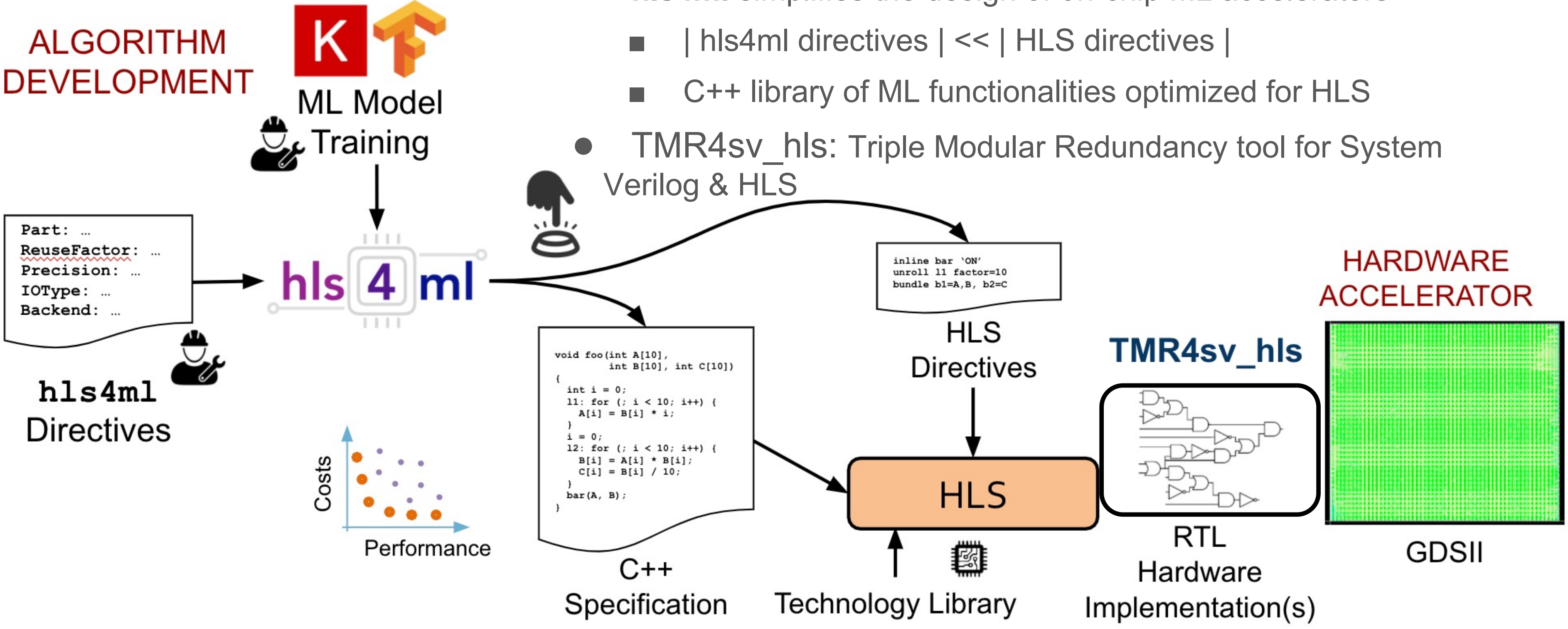
- **conv2D layer** – extract spatially corelated geometric features
- **Flatten layer** – Vectorizes the 2D image from the conv2D layer [8 x 4 x 4 = 128 x 1]
- **Dense layer** – aggregates the various features to provide higher order information
- **ReLU** – an activation function which introduces non-linearity by applying thresholds (part of both the conv2D and dense layers)



*Used as test case

Physics Driven Hardware Co-design

- Algorithm development based on Physics data
- **hls4ml** simplifies the design of on-chip ML accelerators
  - | hls4ml directives | << | HLS directives |
  - C++ library of ML functionalities optimized for HLS
- TMR4sv_hls: Triple Modular Redundancy tool for System Verilog & HLS

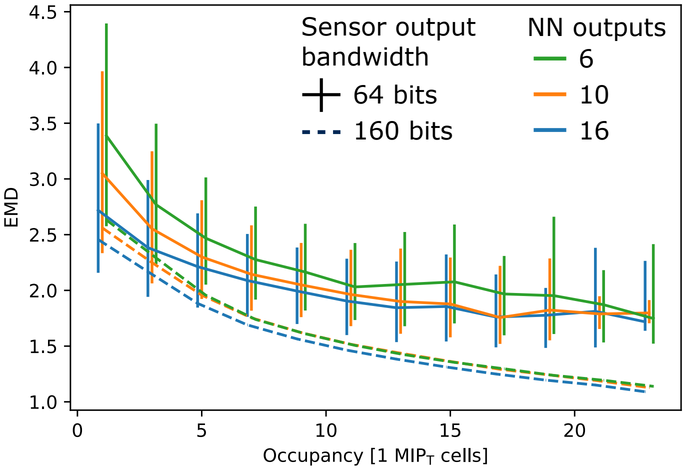


Rapid design prototyping
Neural Network architecture optimization

Lower EMD is better

	Network Architecture					Relative Power & Area		Relative Performance	
Test feature	Geometry	# filter	kernel	stride	pooling	# params	# operations	EMD Mean	EMD RMS
Reference	4x4x3	8	3x3	1	none	1.00	1.00	1.00	1.00
4x4x3 -> 8x8	8x8	8	3x3	1	none	2.73	1.76*	0.64	0.41
max pooling	8x8	8	3x3	1	2x2	0.71	0.97*	0.59	0.33
3x3 -> 5x5 kernel	8x8	8	5x5	1	2x2	0.99	2.76	0.64	0.35
pooling -> stride=2	8x8	8	3x3	2	none	0.94	0.59	0.76	0.46
8 -> 10 filters	8x8	10	3x3	2	none	1.17	0.73	0.73	0.43
8 -> 6 filters	8x8	6	3x3	2	none	0.70	0.44	0.85	0.57

* zero operations removed



Step	Type	Run Time	Iterations	Size
Model generation	D	1s	50–100	1.1k lines of C++
C Simulation	V	1s		
HLS	D	30 min	3–100	40k lines of verilog
RTL simulation	V	1 min		
Logic synthesis	D	6 hrs	6	750k gates
Gate–level sim	V	30 min		
Place and route	D	50 hrs		780k gates
Post–layout sim	V	1 hrs		
Post–layout parasitic sim	V	2 hrs		
SEE simulation	V	4 hrs		
Layout	D	20 min	1	7.6M transistors
LVS and DRC	V	1 hr		

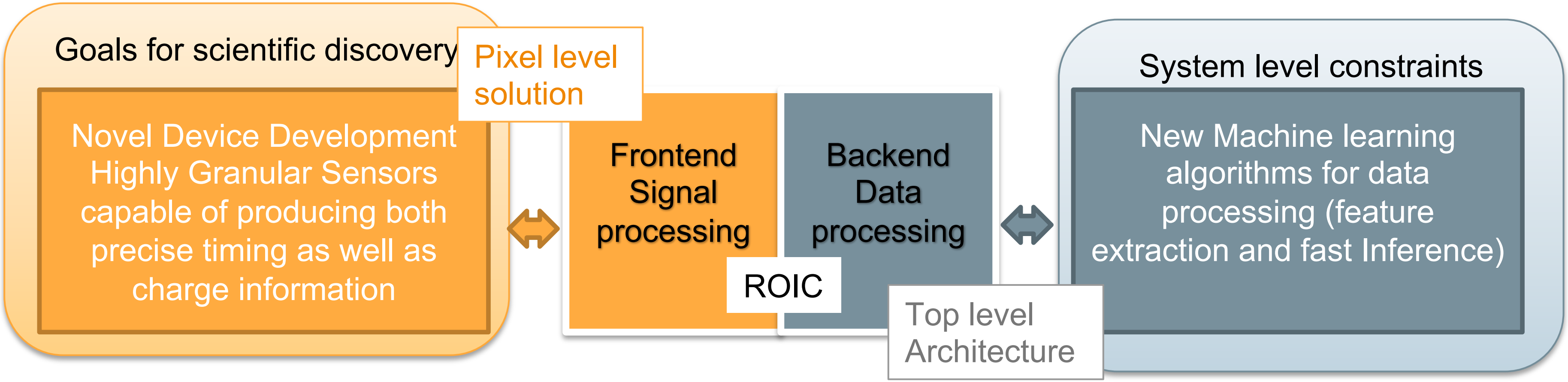
Network optimization

Design optimization

Increasing time and complexity

Metric	Simulation	Target
Power	48 mW	<100 mW
Energy / inference	1.2 nJ	N/A
Area	2.88 mm^2	<4 mm^2
Gates	780k	N/A
Latency	50 ns	<100 ns

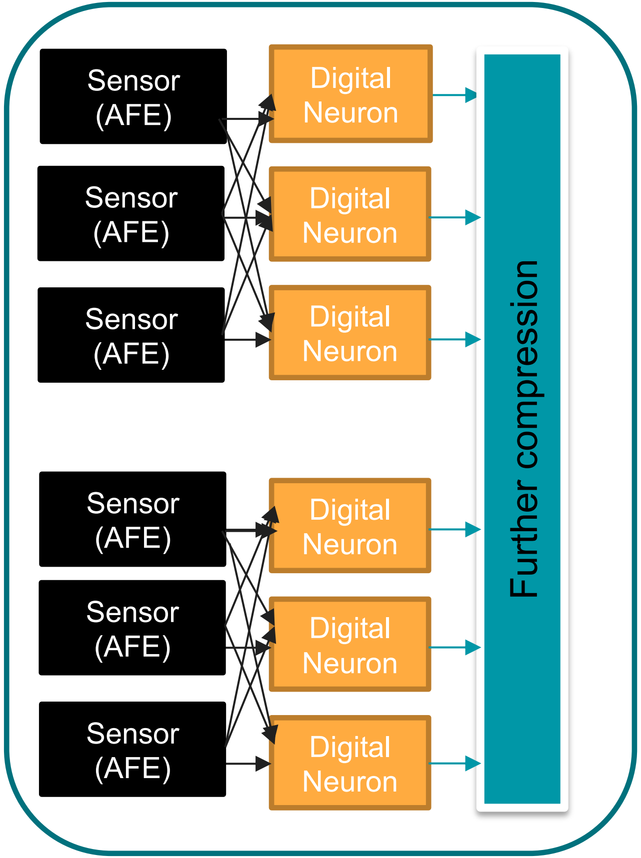
Co-design with algorithm



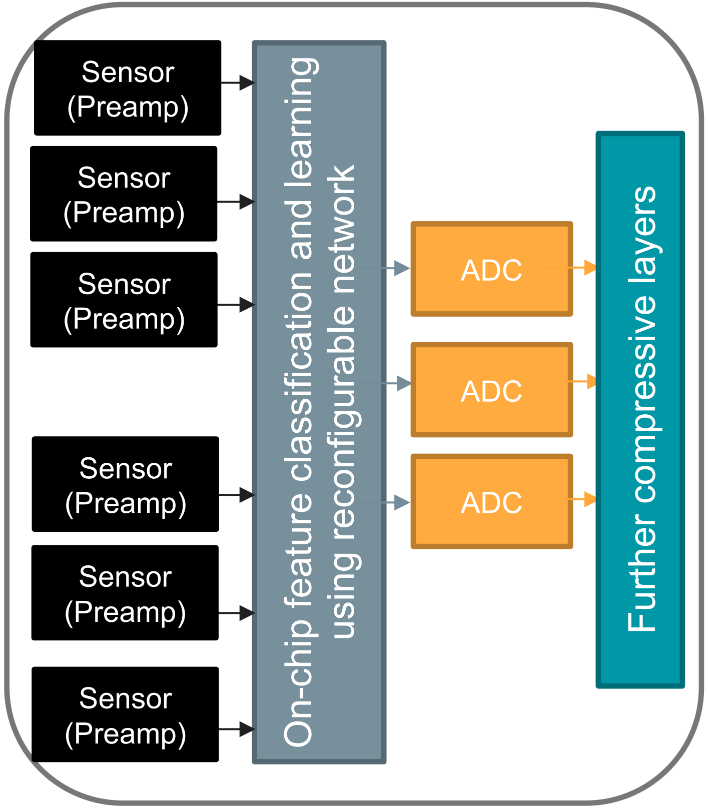
- Convert raw data to physics information
- Reconfigurable pixel clusters for classification dependent on detector geometries
- Create hierarchical network and enable parallel computation.

Pixel Detector: Proposed ML implementation

Digital neuromorphic implementation



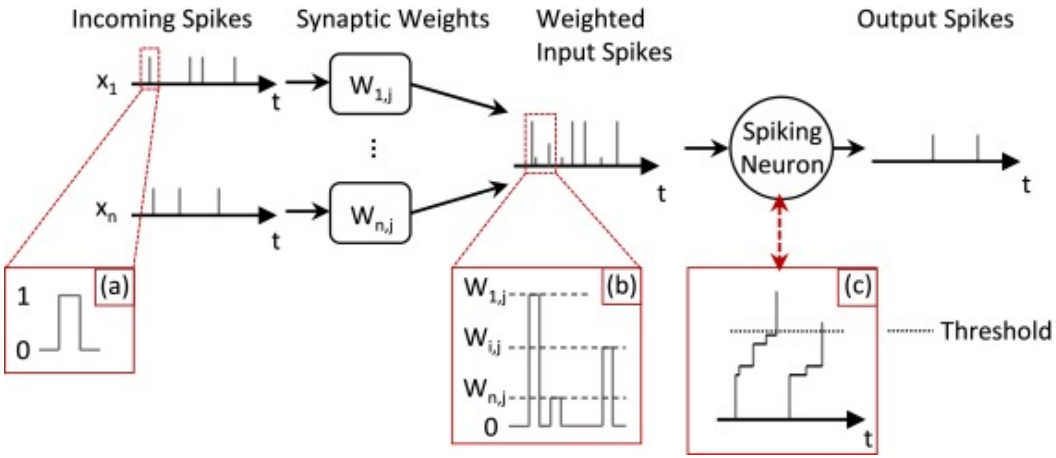
Analog – Mixed Signal implementation using floating gates or memristive cross-bar arrays



- Ability to work in the latent space (downstream resources)
- Reconfigurability vs. pruning?
- On-chip inference vs. on-chip training?
- Light weight models?
- Can lead to self calibrating detectors?

Use digital spiking neural network

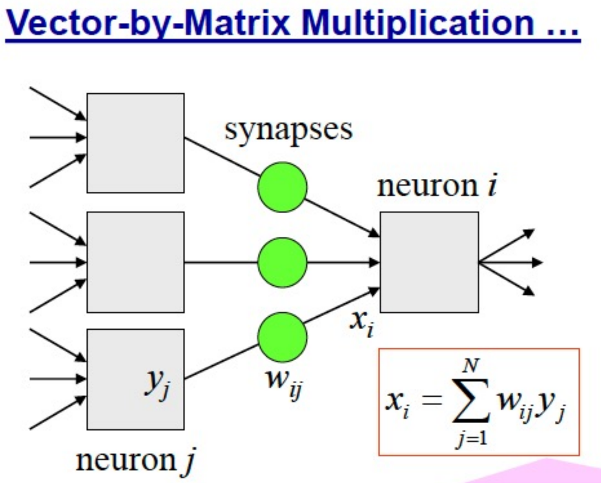
- Local neighborhood for 1st stage of classification
    - Compute total energy and track angle by spatio-temporal correlation
- Mature systems are based on these SNNs e.g. Loihi, True-North, Spinnakar (in adv geometry nodes)
- Low–power since it runs without a clock
- HL- LHC case- since events are uncorrelated between bunches we don't need a complex network requiring historic information

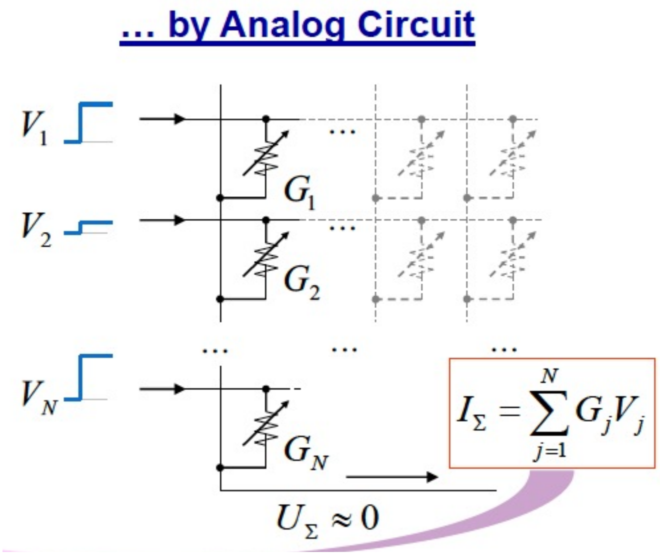


Vector Matrix Multiplication

- Basic building block of neural network

- Analog implementation
- Small footprint, programmable, large resistors

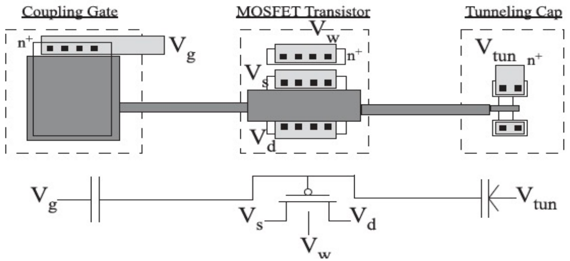


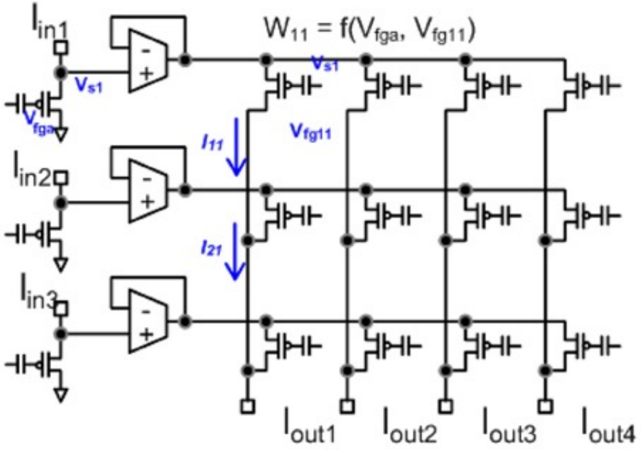


J. Hasler, Georgia Tech

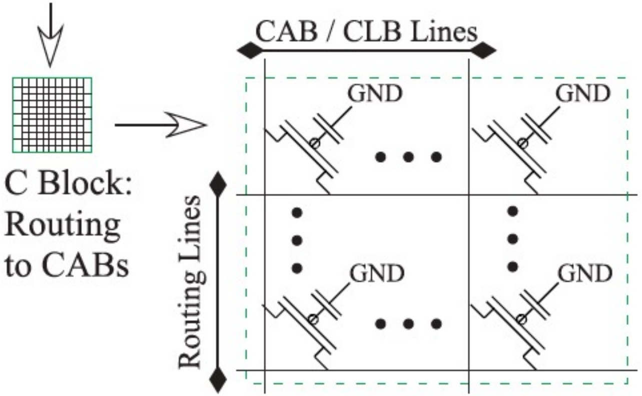
Integrate a Field programmable Analog Array (FPAA)

- Programmable Floating gate transistors for weights and switches
- Structures are available in standard CMOS process (have been demonstrated in 350 nm to 40 nm nodes)
- Radiation performance – unknown
- Uses operation transconductance amplifier (OTA) with floating gates for Vector Matrix multiplication
- Reconfigurable architecture by using a switch matrix and Manhattan routing to define interconnections





$$\begin{bmatrix} I_{out1} \\ I_{out2} \\ I_{out3} \\ I_{out4} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \end{bmatrix} \begin{bmatrix} I_{in1} \\ I_{in2} \\ I_{in3} \end{bmatrix}$$



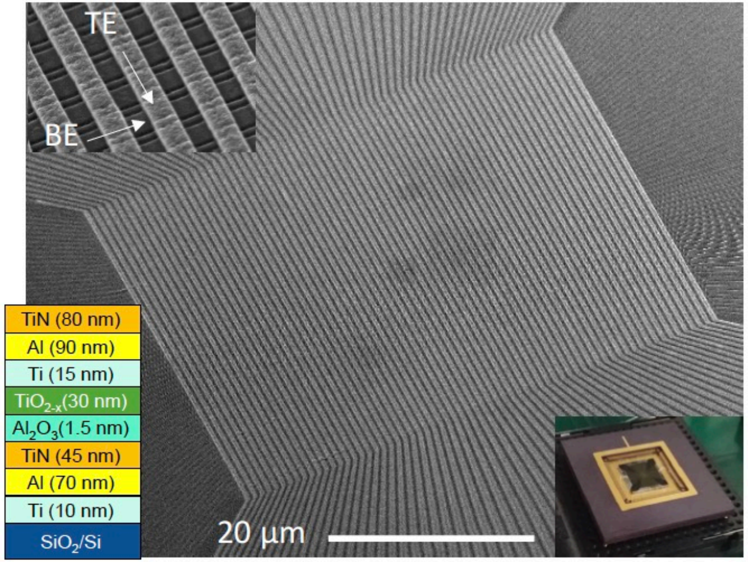
Memristive cross-bar arrays

D. Strukov, UCSB

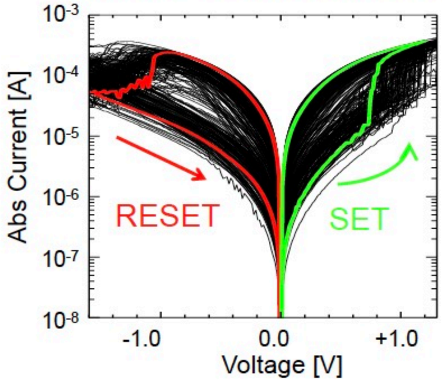
- Use of programmable resistors (1 – 10 GΩ
- Small footprints (< 1μm²)

UC Santa Barbara's Metal-Oxide Memristors

- **64 × 64 passive crossbar circuit**



- **Typical I-V characteristics**



Details:
- Al_2O_3/TiO_{2-x} active bilayer by reactive sputtering
- CMOS-compatible CMP/dry etching process and TiN/Al electrodes for higher conductance
- ~250 nm wide lines
- The largest functional analog-grade passive memristor crossbar circuit supported by proper statistics

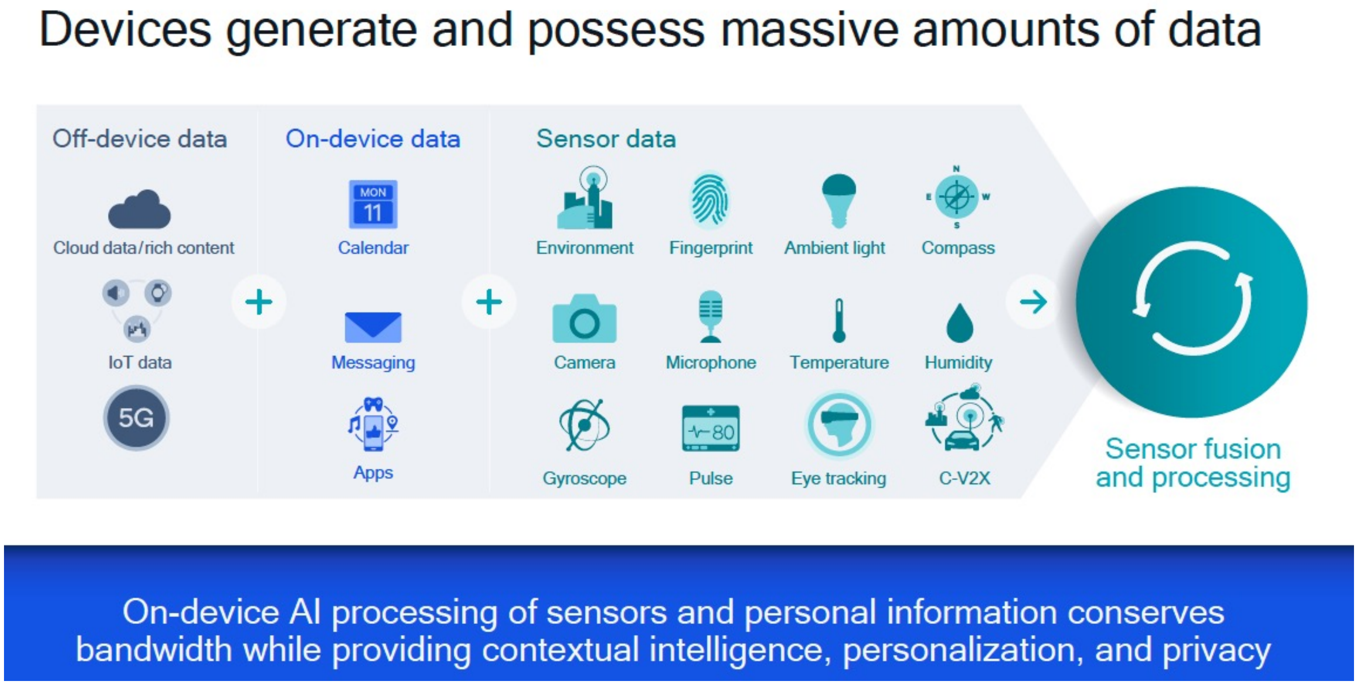
H. Kim et al. arXiv 2019

Background work: *M. Prezioso et al., Nature 521, 61 2015, M. Prezioso et al. IEDM'15 p. 17.4.1, 2015, F. Merrikh Bayat et al. Nature Comm., 2018*

Data collection and processing split at 3 levels

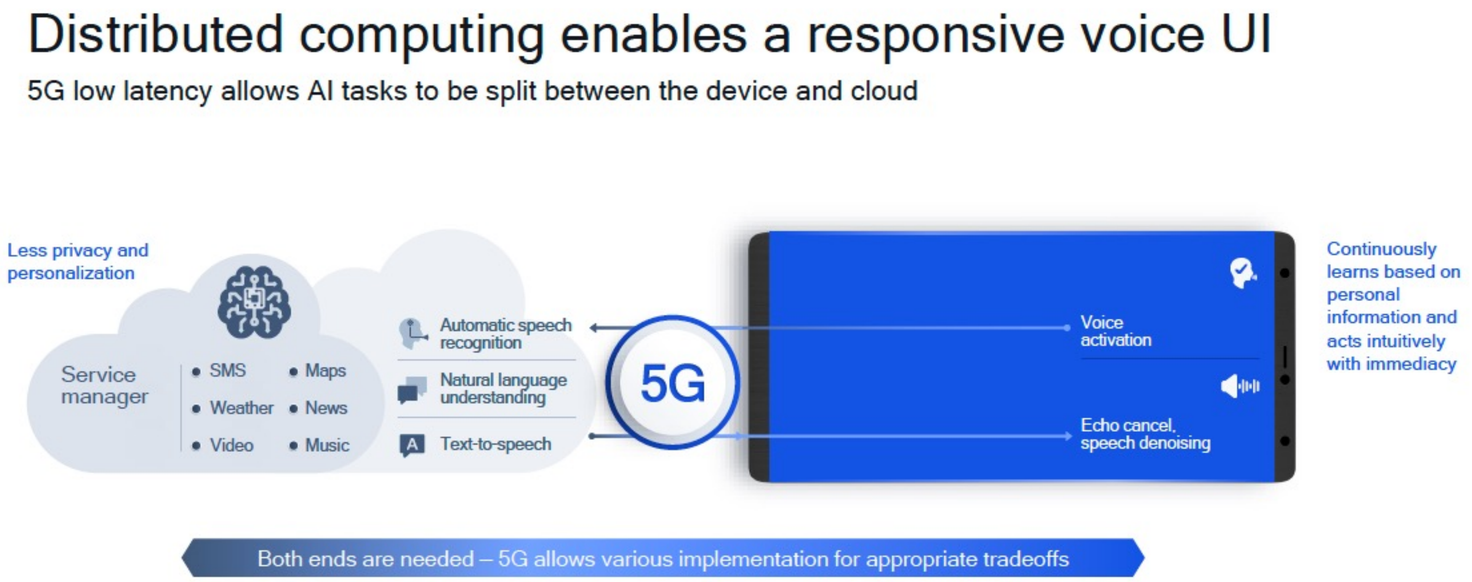
- Sensors: Single ASICs
- Devices: Detectors / Data concentrators
- Off Devices: Processing farms

Commercial - Qualcomm



Enhance AI performance using 5G/wireless

Commercial - Qualcomm



- Communicate between layers for more efficient data processing
- Correlations between layers can provide the best compression
- Local edge cloud can allow for low latency partial processing offload
- Use as continuous learning, additional capacity and maybe increase precision

Conclusion

- Algorithms, Design Tools and Hardware must be co-designed for efficient implementation
- Analog-Mixed Signal & Beyond CMOS techniques can provide energy efficient computation, and their reliability in extreme environments need to be explored
- 5G/6G/Wireless can enable distributed AI for responsive autonomous systems
- A staged approach could lead to robust solutions