Co-Design Methodologies for Edge Computing ASICs in Scientific Research Environment

Sandeep Miryala, Gabriella Carini, Grzegorz Deptuch
ASIC Group, Instrumentation Division
smiryala@bnl.gov

Date11/26/2021

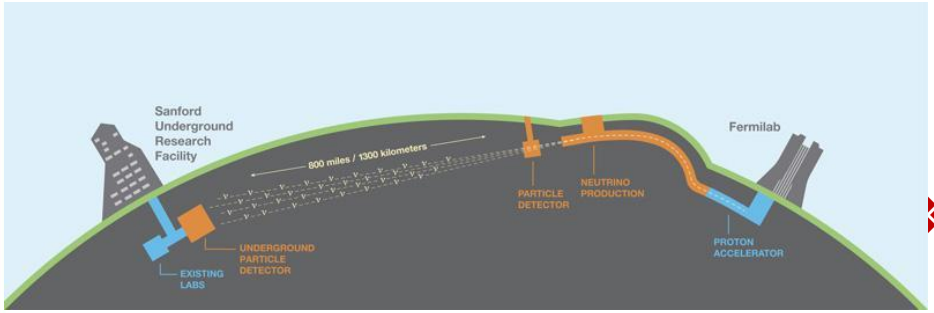
@BrookhavenLab

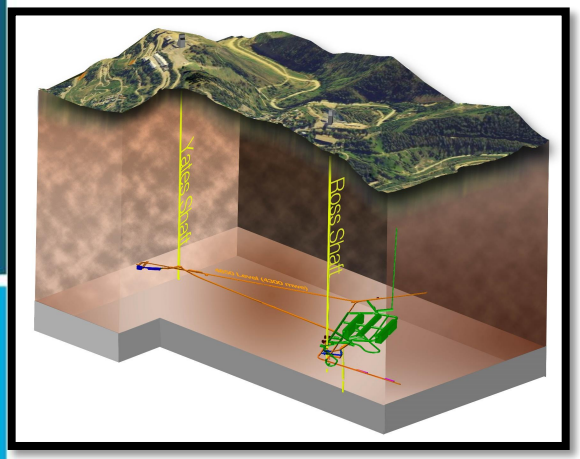
Outline

- Introduction
  - Detector overview and Read Out Integrated Circuits (ROICs)
  - Neural processor on ROICs
- Edge Computing
  - Waveform processing using neural networks
  - Quantization and Compression of a neural network
- Design Methodologies
  - Role of High-level Synthesis tools
- Novel devices: Memristors
  - Device characteristics and modeling
  - Cross Bar Arrays (CBAs)
- Non Von-Neumann Architectures
  - Processing in Memory (PIM)

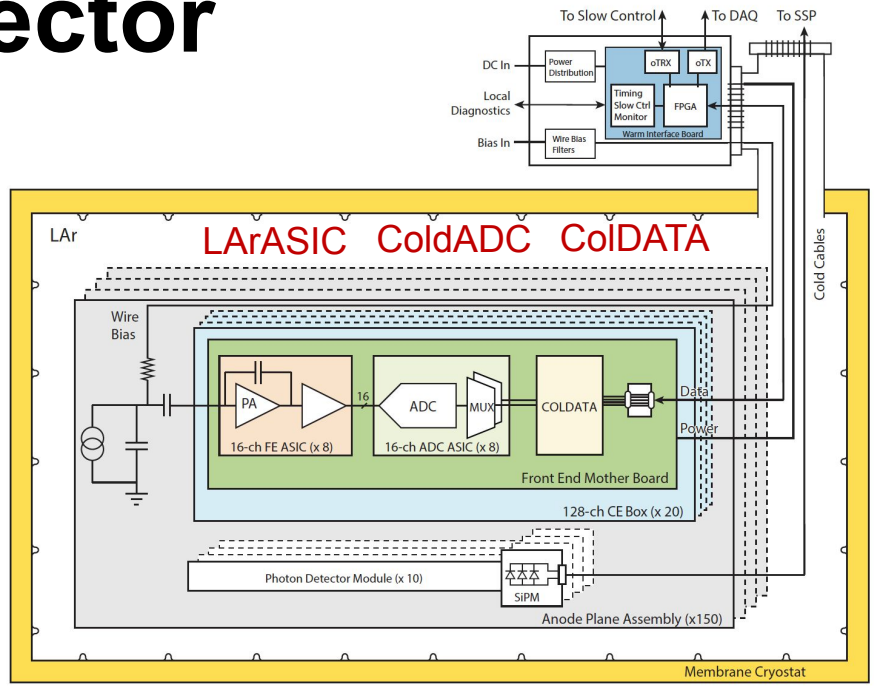
DUNE Far-End Detector

- Precisely measure neutrino properties using a beam from Fermilab
- Detect neutrinos from galactic-core supernovae
- Search for nucleon decay





- Located almost 1 mile underground to reduce backgrounds from cosmic rays

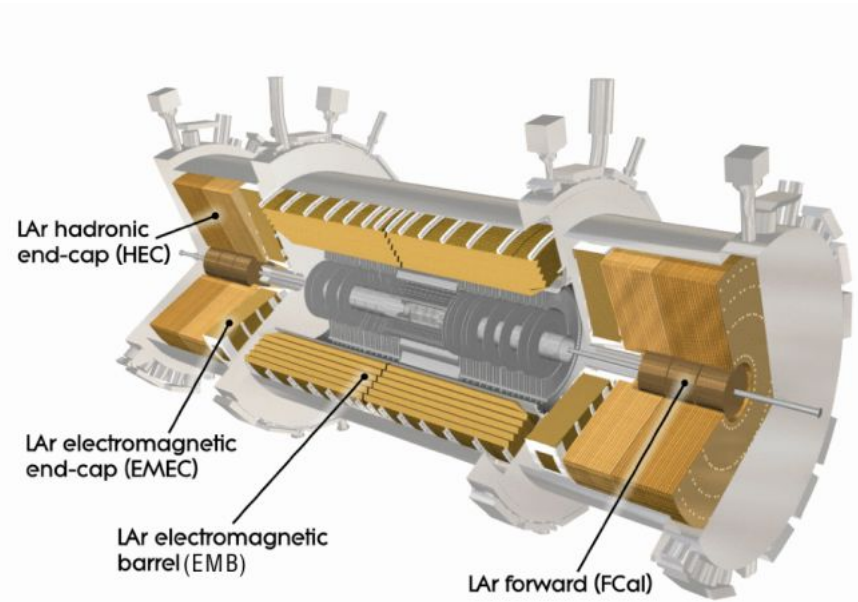


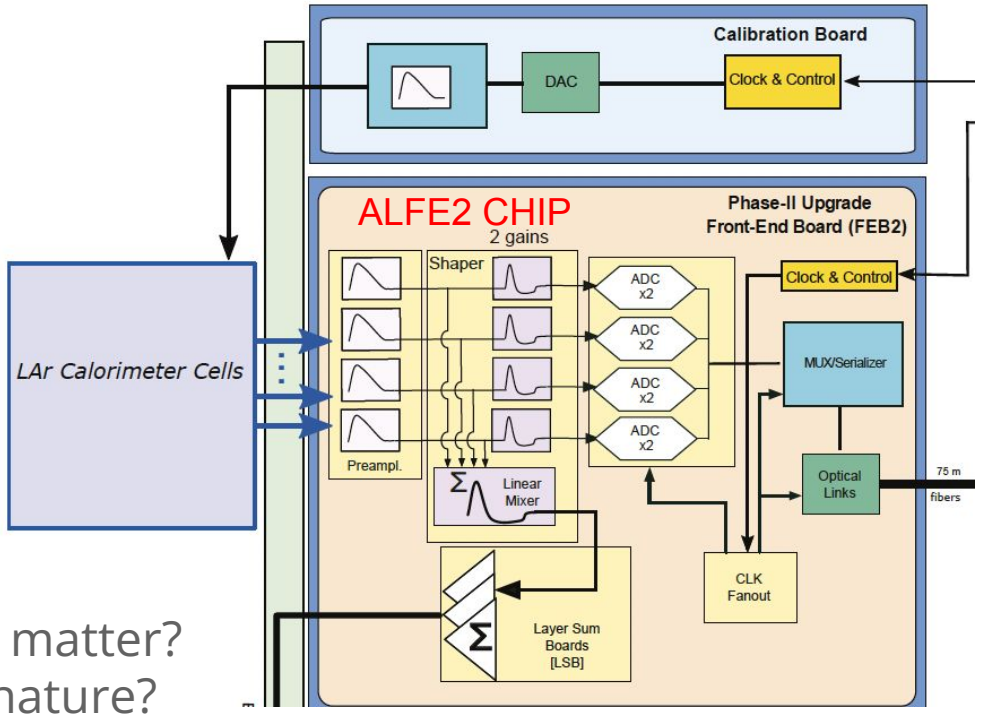
- One 10 kTon detector has
  - 3000 128-channel Front End Mother Boards
  - 24000 FE ASICs, 24000 ADC ASICs, 6000 COLDATA ASICs
  - 12000 1.28 Gbps links (9.2 Tbps of waveform data)

HUGE DATA

S. Miryala et al, "CDP1: A Data Concentrator Prototype for the DUNE" , IEEE TNS 20

CERN ATLAS Experiment





What are the basic building blocks of matter?
What are the fundamental forces of nature?
What is dark matter made of?

- 1 FEB2 mother board streams out ~225 Gbps of data
- A total of 1524 FEB2 boards for entire (FCal) system, ~345 Tbps

Future proton colliders will exceed these data rates

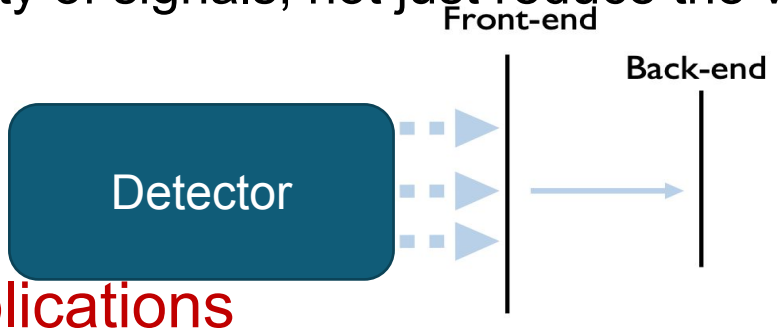
HUGE DATA

Brookhaven National Laboratory

https://cds.cern.ch/record/2285582/files/ATLAS-TDR-027.pdf

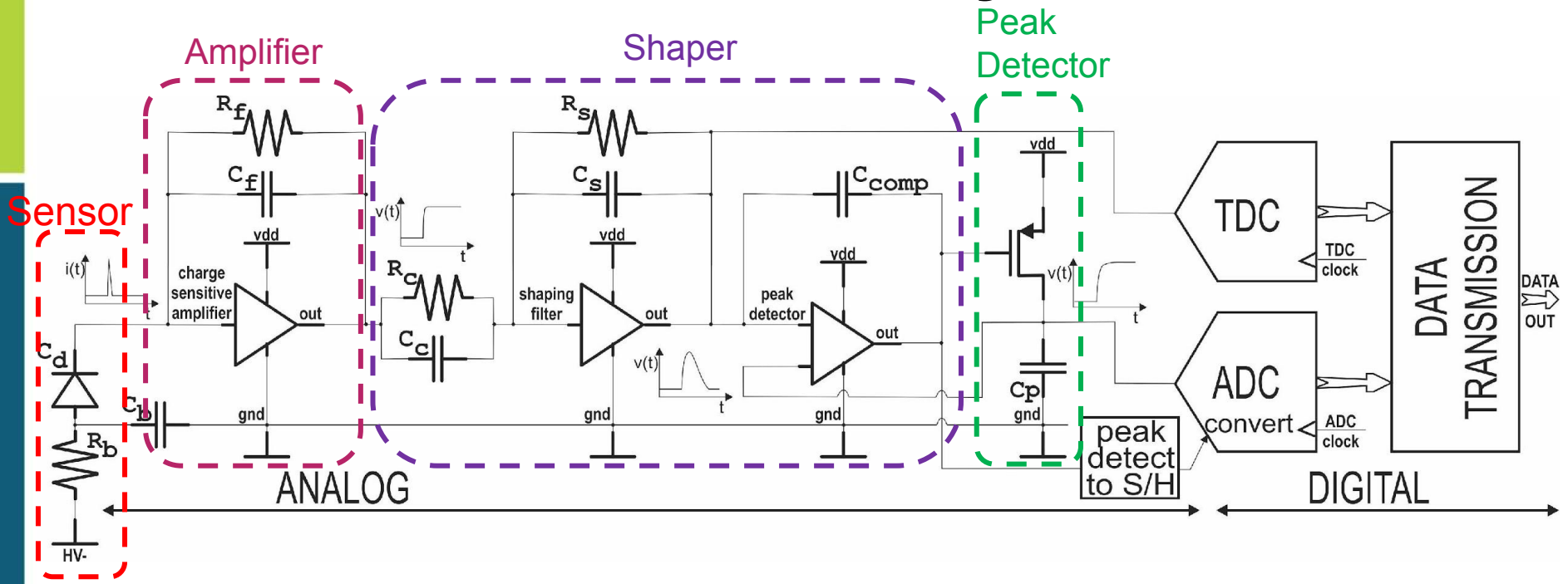
Towards Edge Computing

- Edge computing
  - Traditionally, data processing is mostly done outside of front-end ASICs using commercial FPGA/GPU/DSP/Neural_Chips
  - Introduce smartness by bringing processing inside with artificial neural networks **could be the future for front-end ASICs**
  - Include more logic on the Front-End Electronics (FEE)
  - Improve quality of signals, not just reduce the volume



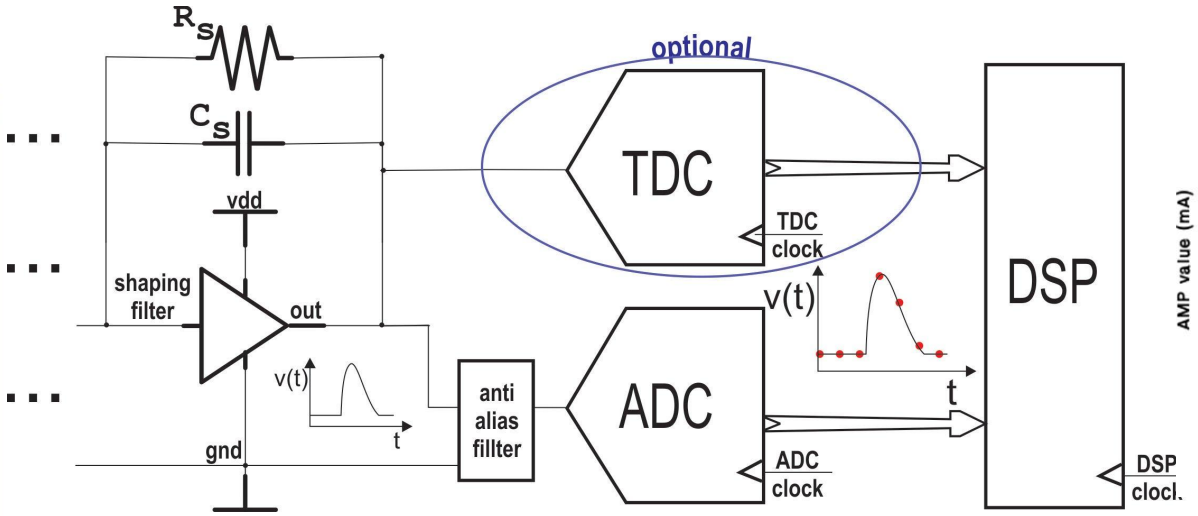
- Immediate applications
  - Waveform: Denoising, digital interpolating filters for processing of sampled waveform, e.g. improve energy resolution through digital peak finding in readout circuits
  - Spatial Distribution: Enhancing of 2D or 3D spatial resolution and data reduction filtering, e.g. solving charge or light sharing problems in pixel detectors or PET scanners
  - Data Concentrator: Event reconstruction

Conventional Readout System

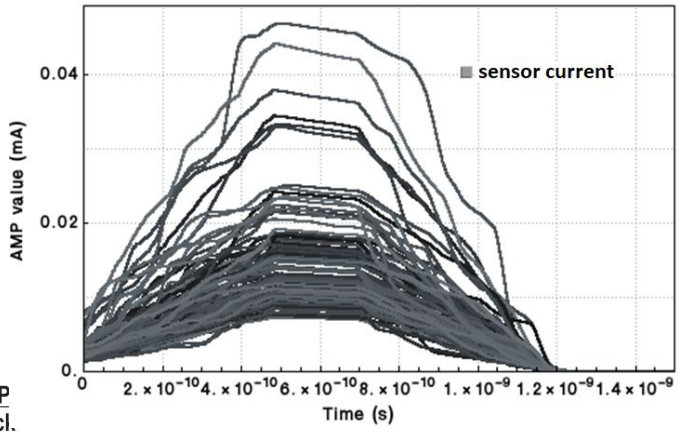


- analog front end: charge sensitive preamplifier, shaping filter, peak detector, front end of ADC and/or TDC
- digital processing chain: ADC and/or TDC, digital signal processing (zero-suppression, encoding, transmission)
- analog circuitry requires precise design and digital assistance needed for trimming its parameters only one sample (at peak) is processed to estimate energy deposited in sensor
- OnDemand A-to-D conversion, whereas high resolution ADC perform better if ran continuously
- no possibility of increasing accuracy using averaging
- Industrial trend: to reduce analog strictly to front-end up to antialiasing filter of ADC

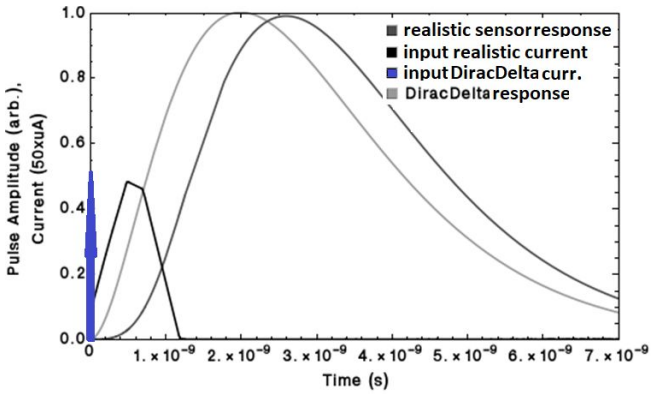
Streaming ADC Based Readout



- real sensor current



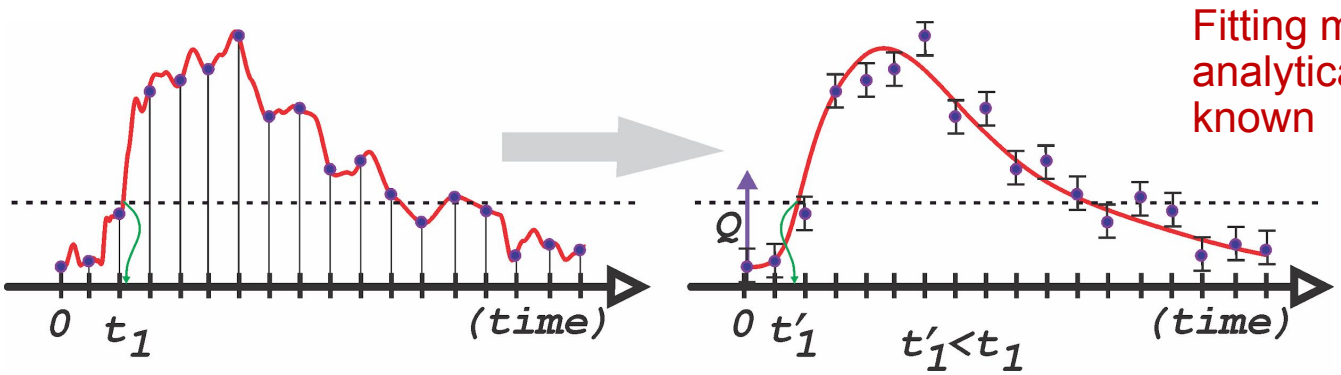
- ideal and convolved response



- continuous A-to-D conversion (waveform sampling)
- Digital Signal Processor (DSP) can perform fitting or interpolation leading to"
  - peak finding
  - time of arrival estimation
- Assuming ideal forms of pulses fitting can be very precise
- analytical fitting difficult on chip => FIR or Neural Network

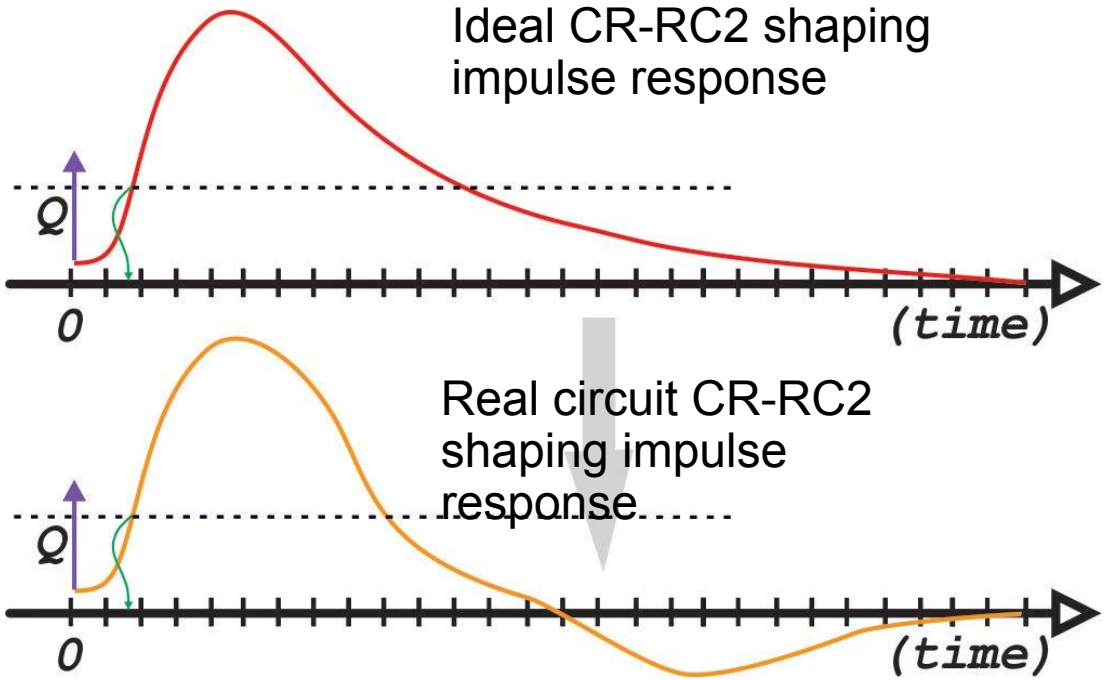
Brookhaven National Laboratory

Nonidealities in signal response



Where is the actual peak, or threshold crossing?

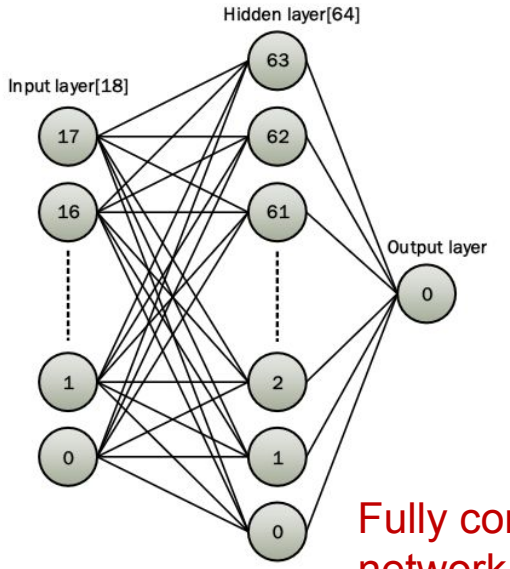
- Can Neural Network "learn each channel signal shape" and do deconvolution in general case?



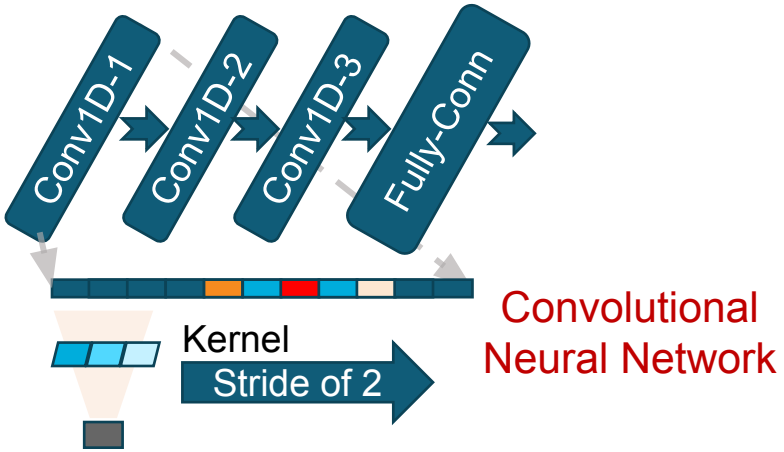
Brookhaven National Laboratory

Machine Learning Algorithms

- Investigated two types of neural network for peak amplitude prediction
  - Multi Layer Perceptron (MLP)
  - Convolutional Neural Network (CNN)



Fully connected network or MLP



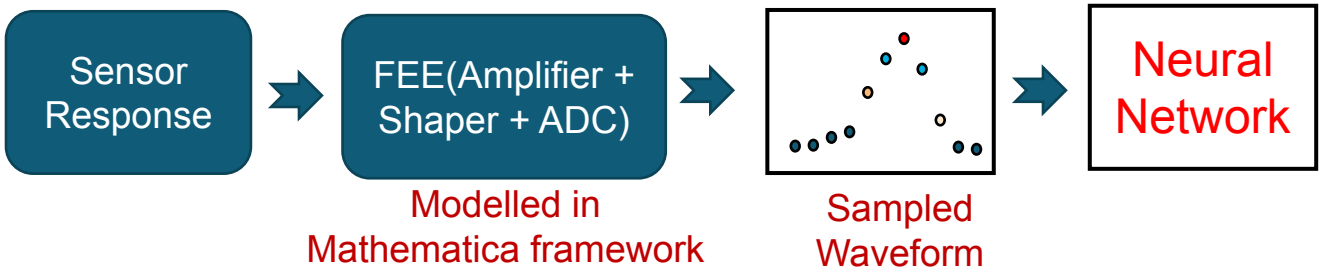
Convolutional Neural Network

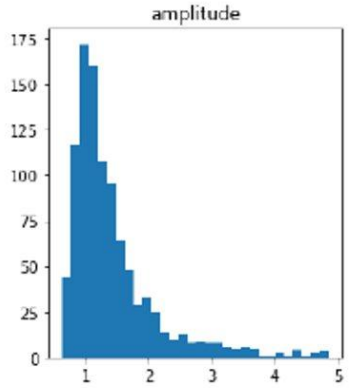
- How many layers?
- How many neurons on each layer?
- Compression of neural networks?
- Quantization of data variables?

→ Impacts Hardware (Power, Performance, Area)

Brookhaven National Laboratory

Ground Truth Data





Objective

Estimate the peak amplitude

Data sets

Sensor response (Practical)

Sampled Waveform Set

3400 points on each waveform

10000 waveforms

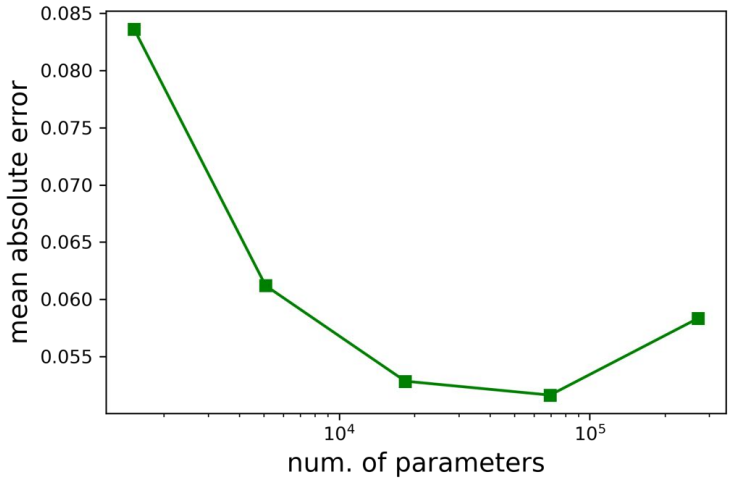
Data set split into 80% train, 10% validation, 10% test

Brookhaven National Laboratory

Inferencing Accuracy (MAE)

MLP Setting	# of parameters	MAE (Mean Absolute Error)
16-32-16	1521	0.0836098
32-64-32	5089	0.0611916
64-128-64	18369	0.0528343
128-256-128	69505	0.0516194
256-512-256	270081	0.0583198

- Varied number of neurons on hidden layers
- Analysis is performed for sensor response
- Acceptable inferencing accuracy

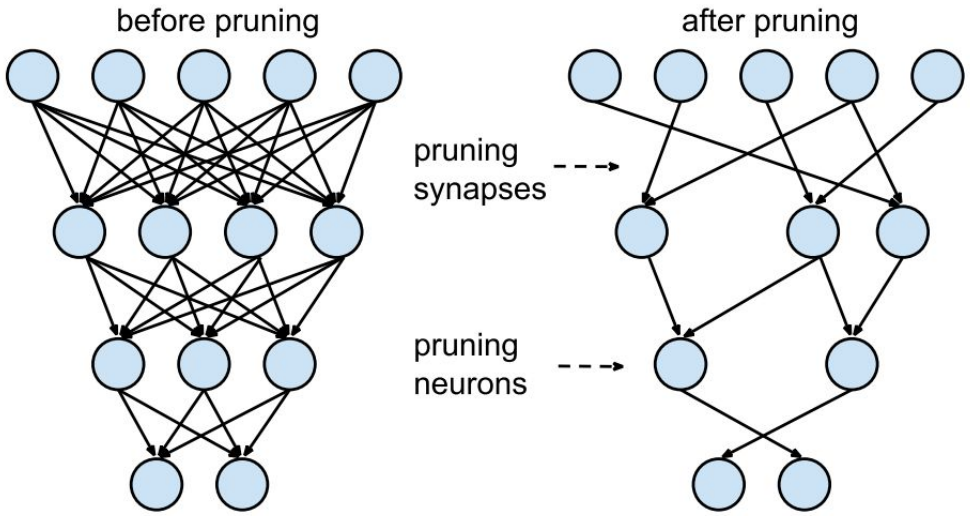


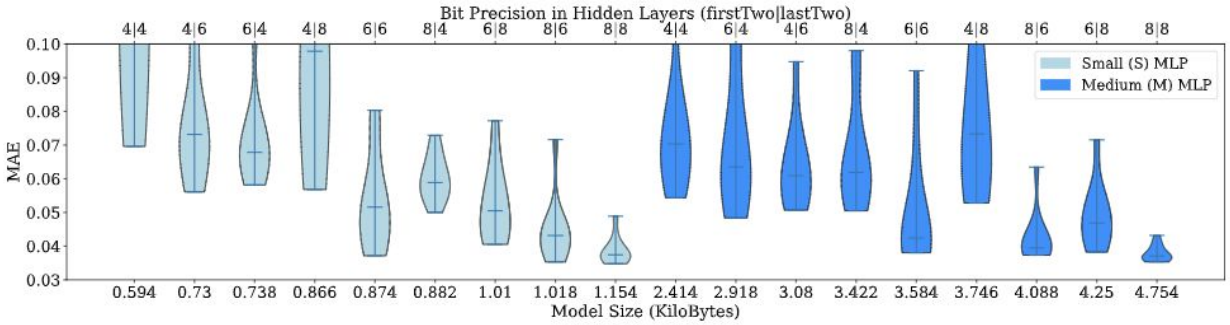
Brookhaven National Laboratory

S. Miryala et al, "Waveform Analysis Using Machine Learning Algorithms on the Front-End Electronics" , iWorid, 2021

Quantization and Pruning

- Quantization of weights as 4-bit, 6-bit and 8-bit fixed point instead of 32-bit floating bit numbers
- Pruning reduces the network
- Pruning with Quantitation Aware Training (PQAT) for efficient inferencing accuracy



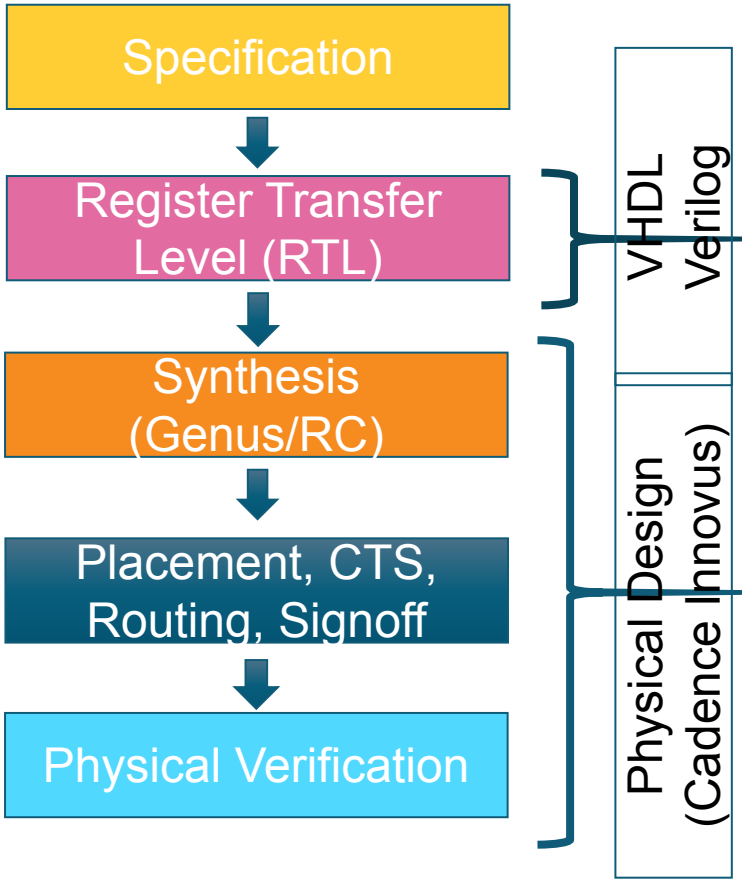


Evaluation of Quantization Aware Training

Brookhaven National Laboratory

ASIC Design Methodology

Conventional Digital Design Flow



- Design of Neural Network
  - First, building and optimizing Neural Network model
    - tools available: Tensor Flow, PyTorch or Caffe2 frameworks
  - Second, training of NN model to estimate kernel weights, in GPUs
  - New tools and methodologies to bridge the conventional flow with python-based frameworks
  - Quickly adapt to the changes in neural network
  - Verification automation at various stages

High Level Synthesis (HLS) tools for a neural processor design

Open-source framework in the community

Intermediate format, Python to C++

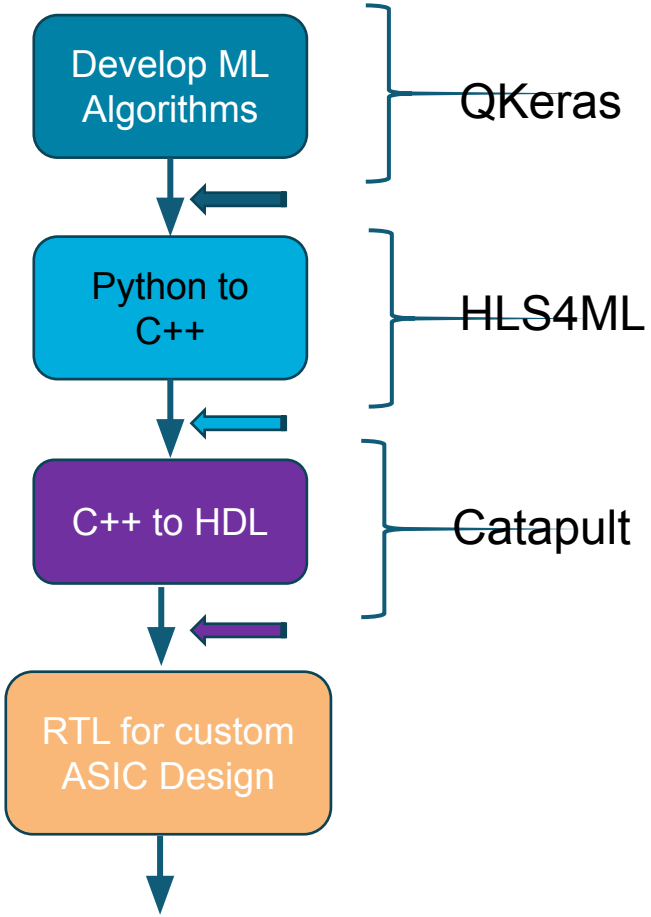
Commercial HLS tool from Siemens C/C++ to HDL

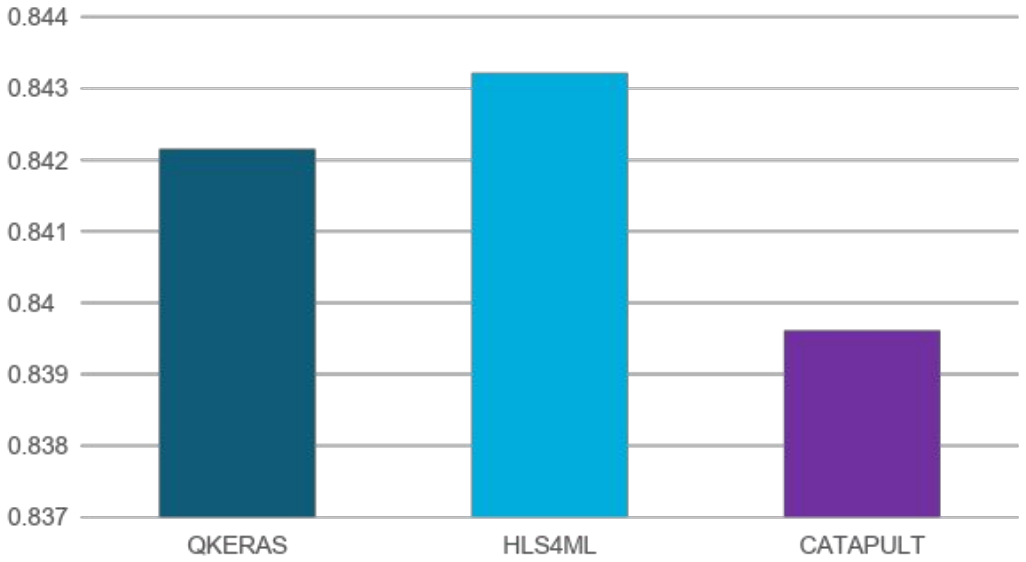
RTL Generation

Digital Design Flow

- **HLS4ML framework**
  - Generates C++ code for the neural network
  - Reads in weights and bias from Qkeras
  - Supports only MLP and CNN
- **Catapult**
  - Maps C++ code to RTL (Verilog / VHDL)
  - Also offers verification framework

Preliminary RTL Results and Verification

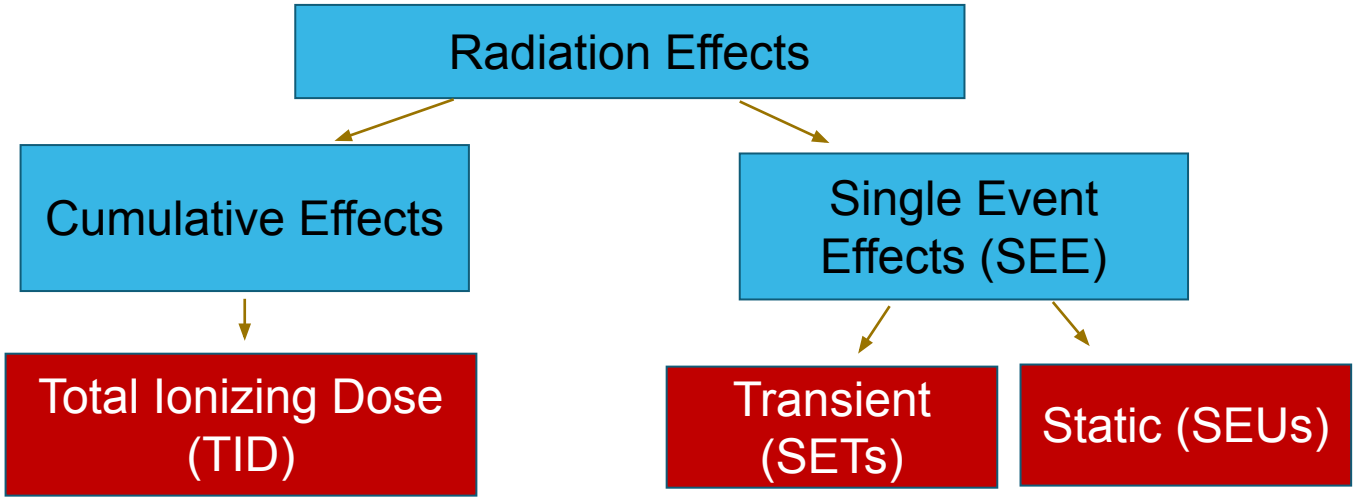




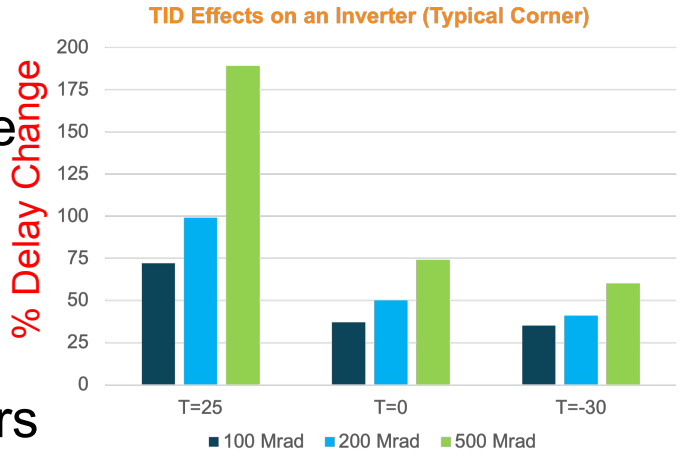
- MAE Verification is performed at various level
- Good match (< 1%)

Brookhaven National Laboratory

AI ASICs for Radiation Environment



- **TID tolerant digital design**
  - Performance degradation of devices over time
  - Neural processor must be rad-hard
  - New process corners are introduced
- **SEE tolerant digital design**
  - Triple Modular Redundancy (TMR) on registers
  - Protect all the configuration registers

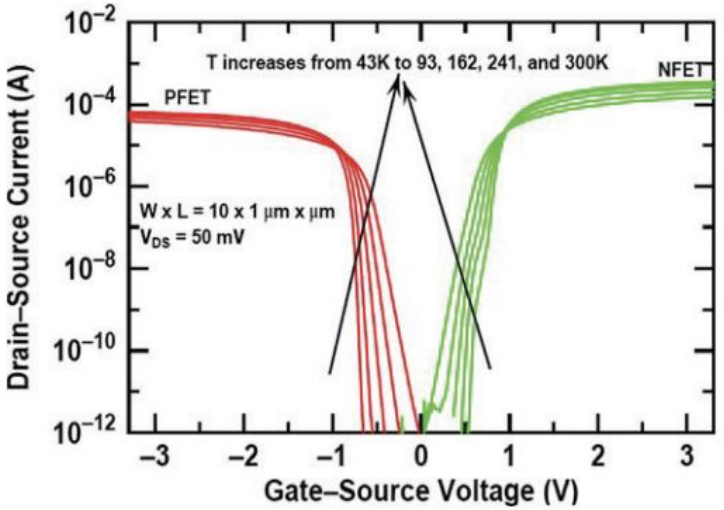


Brookhaven National Laboratory

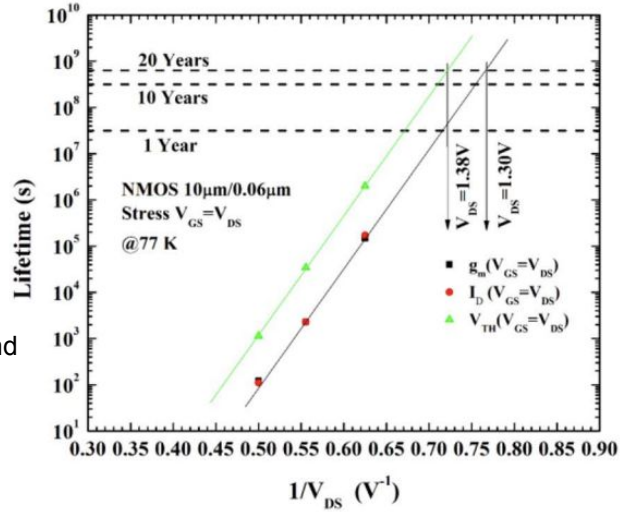
S. Miryala et al, "Single Event Upset (SEU) Cross-Section Measurements of Sequential Elements with TMR in a 65nm process," IEEE NSS, 2020

AI ASICs for Cryogenic Detectors

- Commercial process design kits for ASIC design support from -40ºC to 125ºC
- Cryogenic temp range from -185ºC to -269ºC are of interest for scientific applications
- CMOS device reliability is an issue due to Hot Carrier Injection (HCI)
- Custom SPICE model developments supporting cryogenic range
- Custom standard cell libraries and timing libraries
- Neural processors for cryogenic detectors must adapt



T. Chen et al., "CMOS reliability issues for emerging cryogenic Lunar electronics applications," Solid State Electron., vol. 50, pp. 959–963, Jun. 2006.

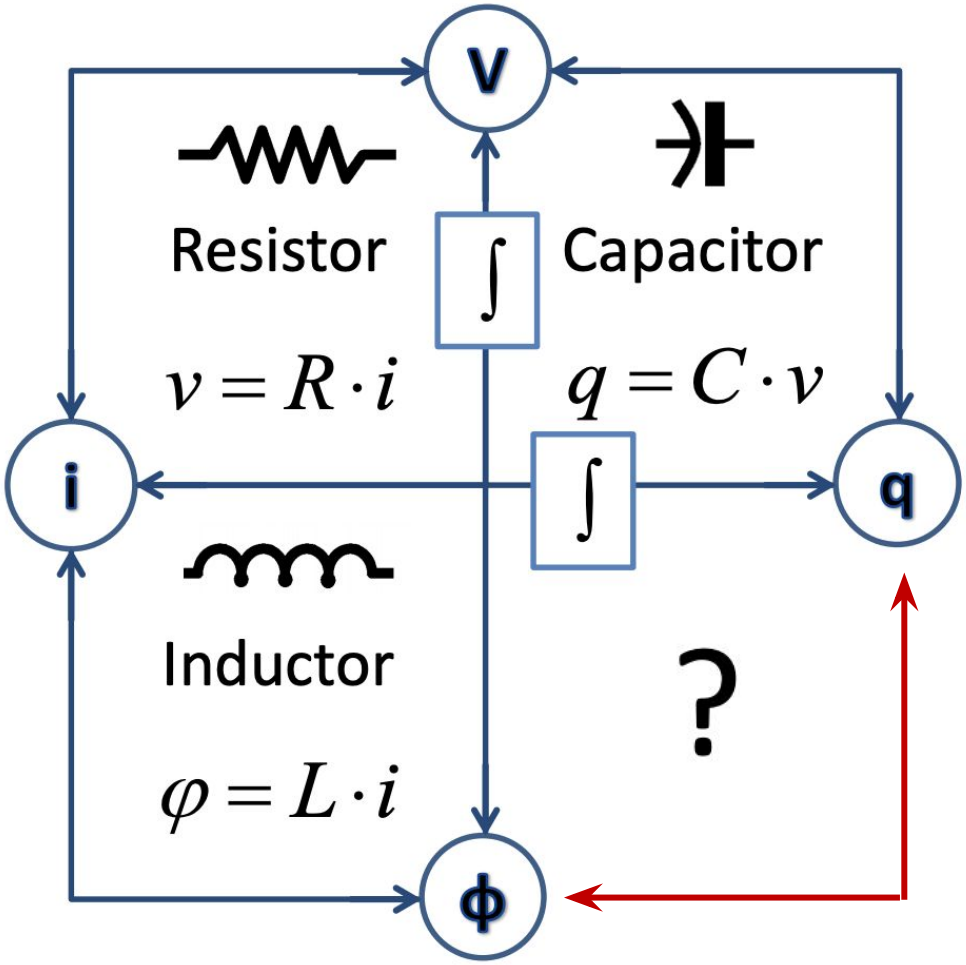


J. Hoff et.al., "Cryogenic Lifetime Studies of 130 nm and 65 nm nMOS Transistors for High-Energy Physics Experiments", TNS, 2015

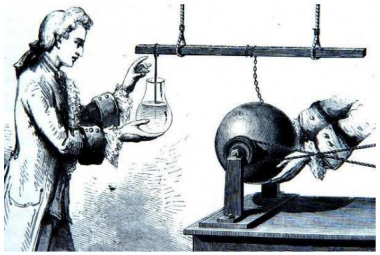
Brookhaven National Laboratory

Novel Devices - Memristors

Fundamental circuit elements



Missing fourth element ?



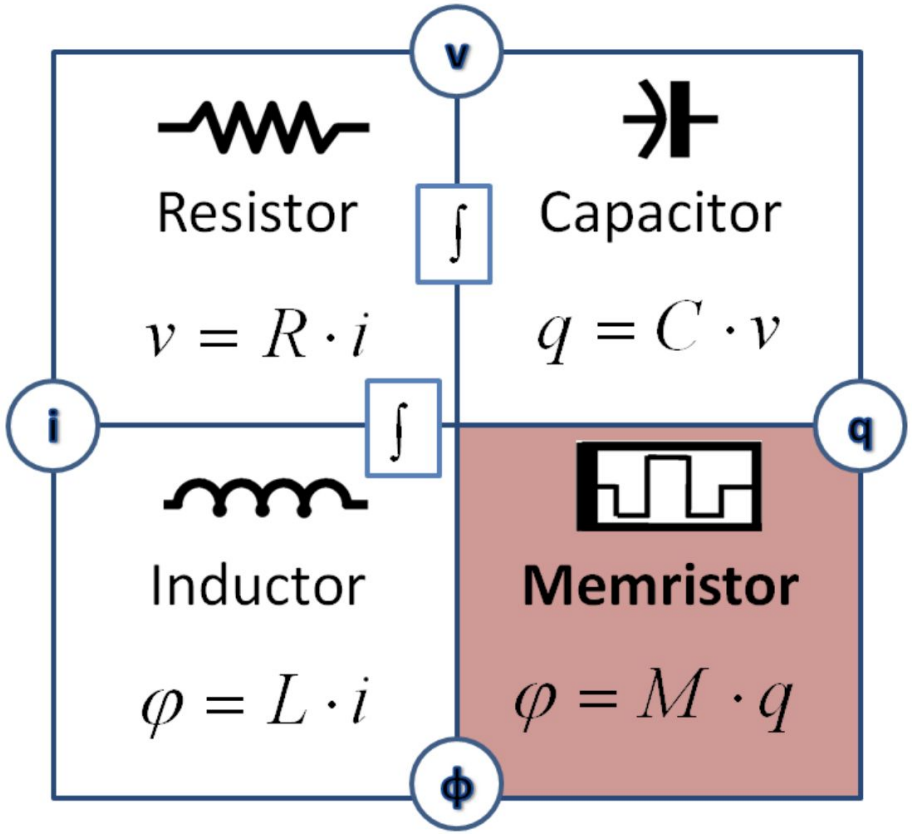
Capacitor
von Kleist 1745

Resistor
Georg Ohm 1827

Inductor
Michael Faraday 1831

Brookhaven National Laboratory

Novel Devices - Memristors



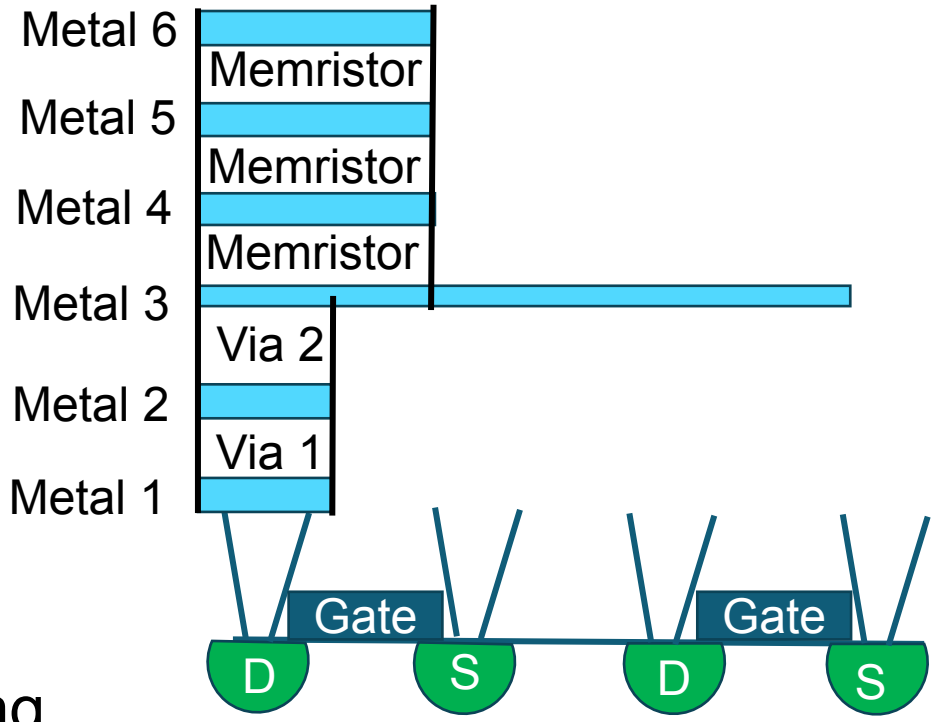
IEEE TRANSACTIONS ON CIRCUIT THEORY, VOL. CT-18, NO. 5, SEPTEMBER 1971 507

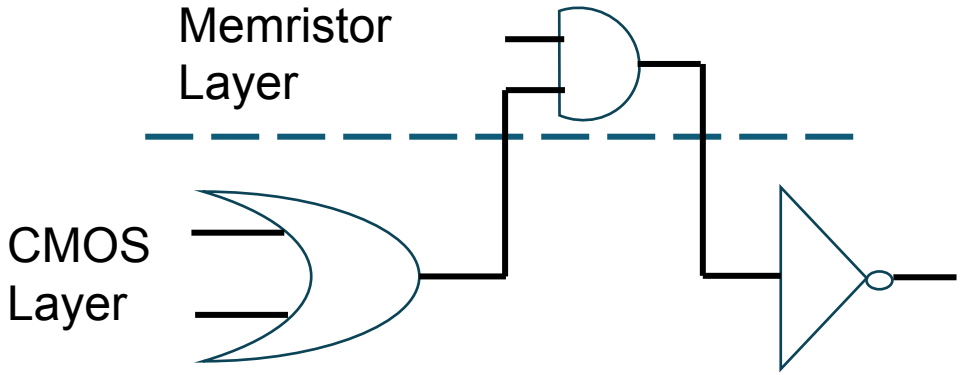
Memristor—The Missing Circuit Element

LEON O. CHUA, SENIOR MEMBER, IEEE

Advantages

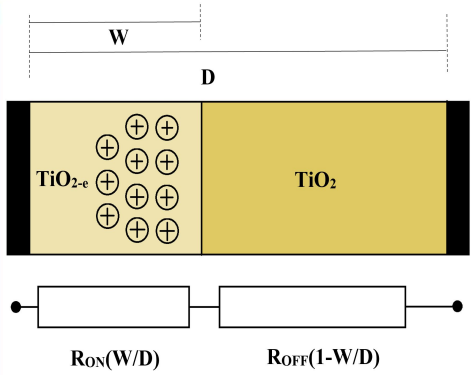
- CMOS compatible
- Memory benefits
  - Dense
  - Nonvolatile
  - Fast
  - Low power
  - High endurance
- Rad-Hard
- Beyond Moore: Integrating memristors with standard logic

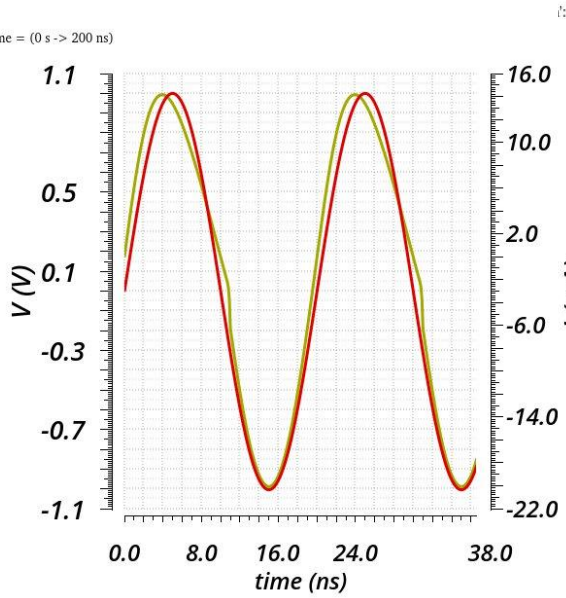


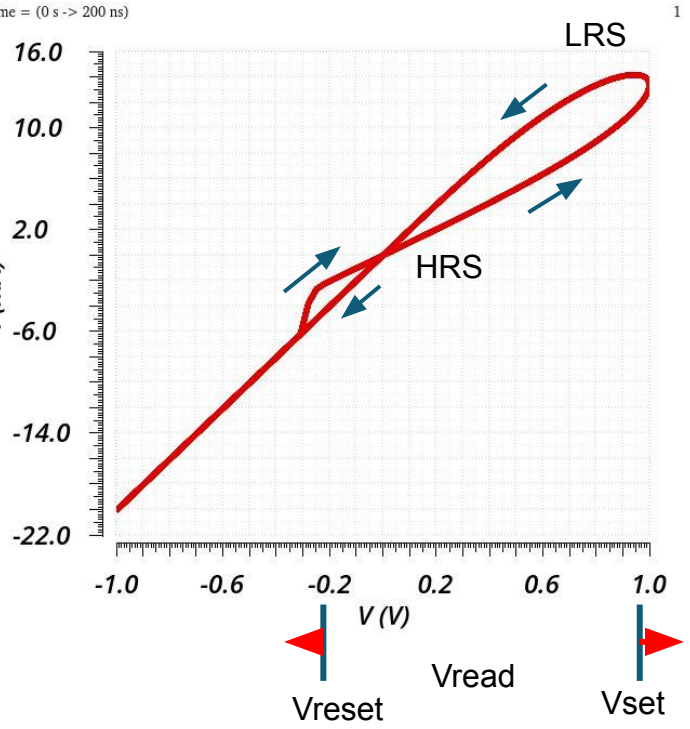


Brookhaven National Laboratory

I-V Characteristics and device modeling





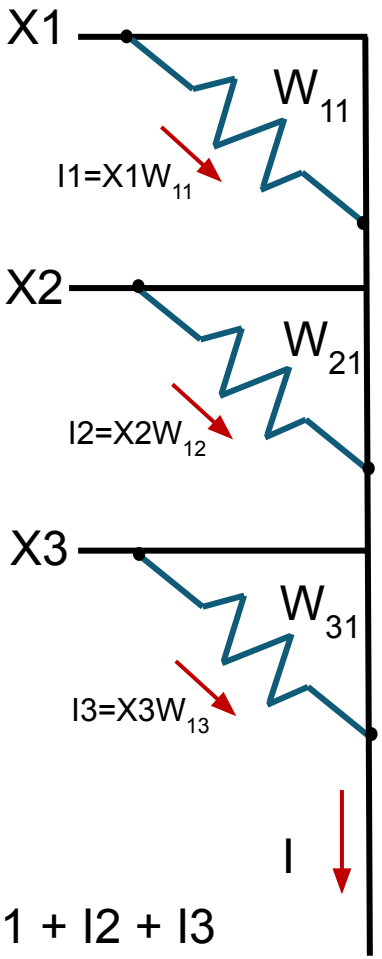


Response to a sinusoidal excitation

- Resistor with varying resistance, Low Resistive State (LRS), High Resistive State (HRS)
  - Neuromorphic Computing, Memristor as a Synapse, Memristor as a Neuron
  - Device models are essential for circuit simulations
  - Identify commercial fab houses that can fabricate at wafer level

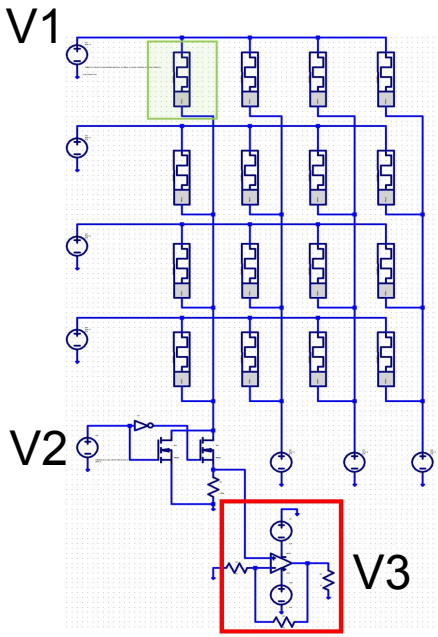
Brookhaven National Laboratory

Cross Bar Arrays and Simulations

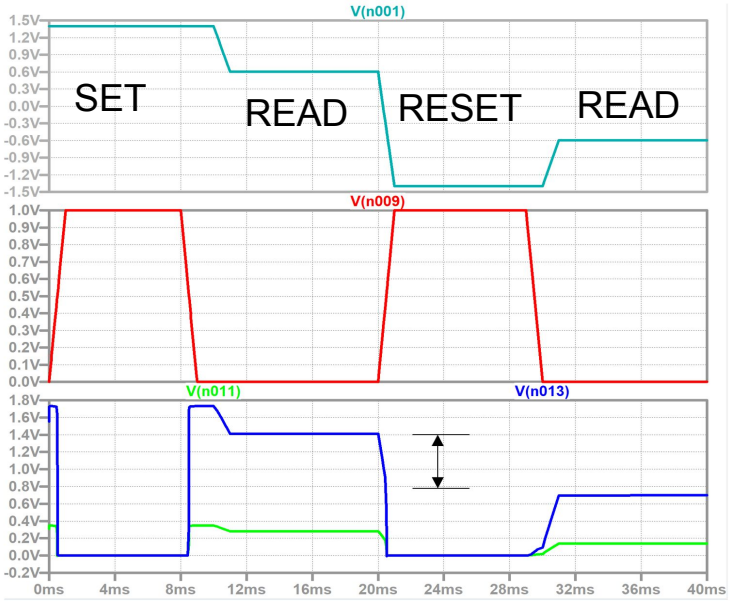


$I = I1 + I2 + I3$

$I = X1W_{11} + X2W_{21} + X3W_{31}$



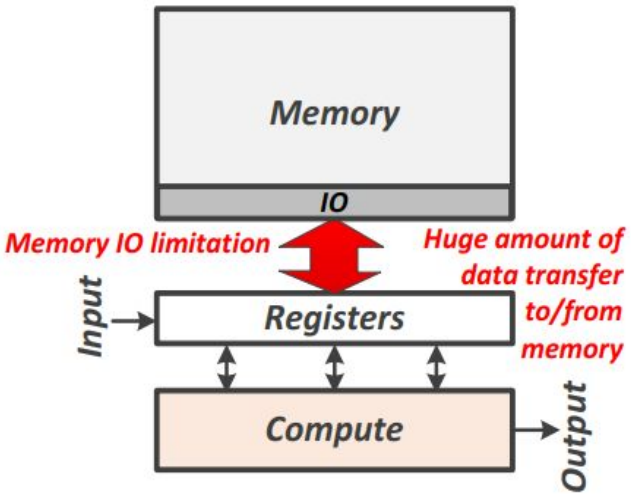
Level Shifter



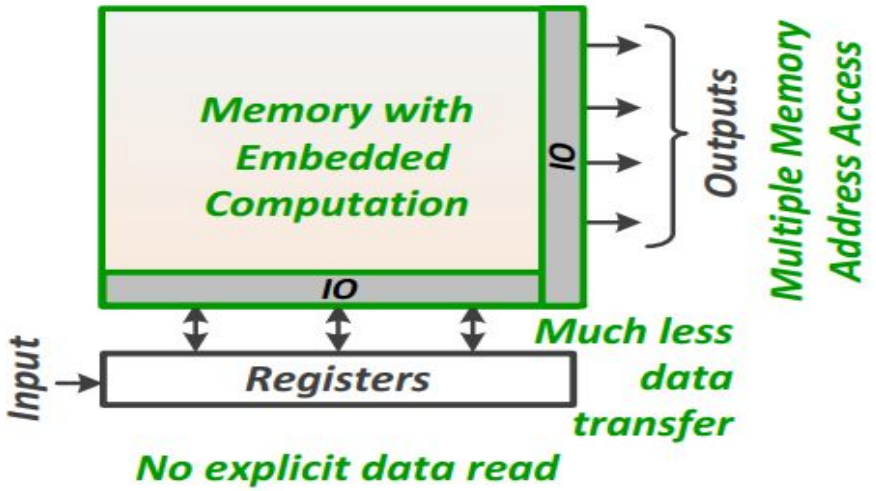
- Reading and writing to 1st row 1st column

W_{11}, W_{12}, W_{13} are the conductance

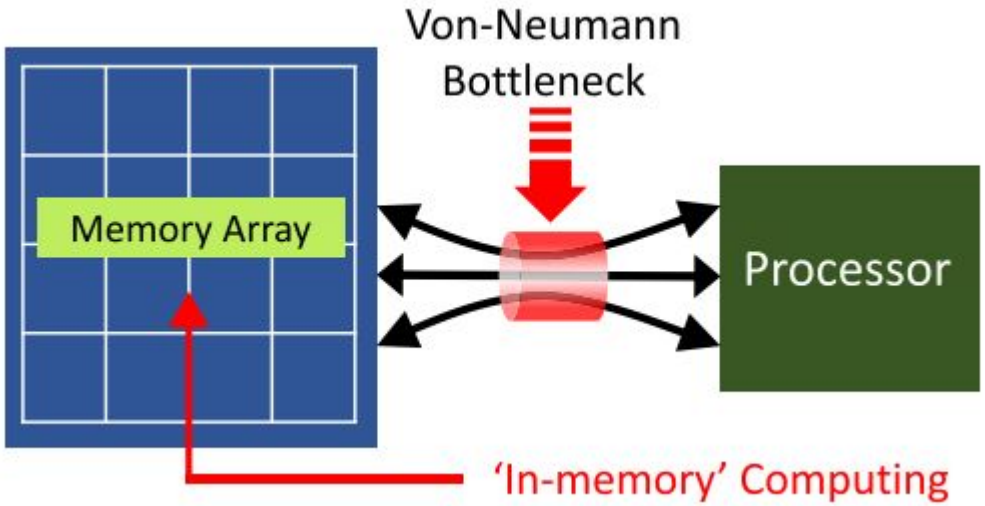
Von-Neumann Architecture: Issues



Conventional Architecture

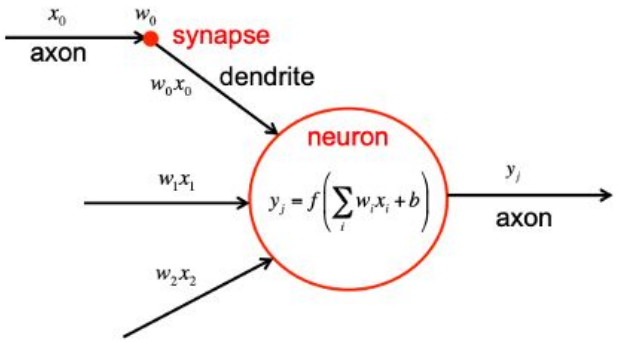


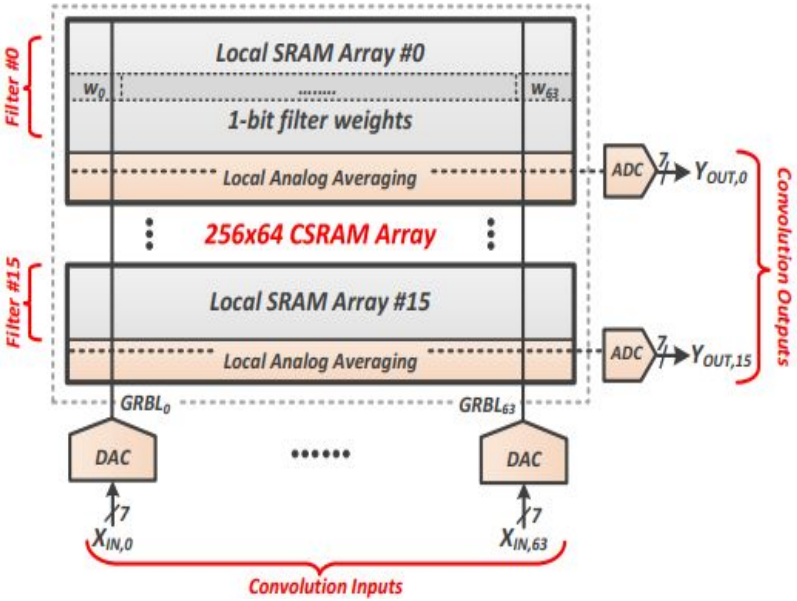
In-memory Architecture



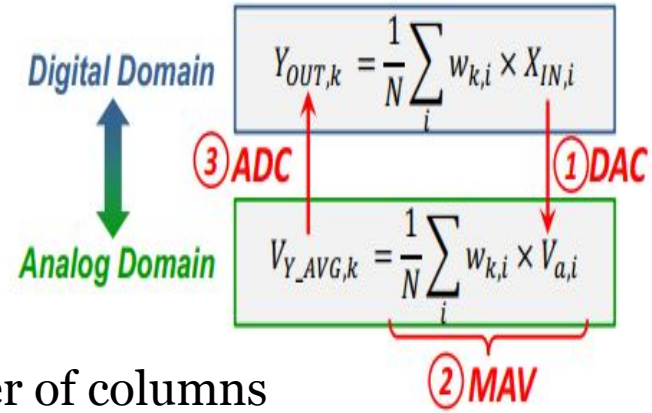
[A. Biswas et al, JSSC'19]

In-Memory Computing





- Multiply And Accumulate (MAC) implementation could be carried out either in Digital or Analog
- Improves the energy efficiency of the system by reducing the data transfer between the processor and memory unit.
- Improve the memory bandwidth because of multiple memory access for parallel processing.



N- Number of columns
W- filter weights
V_a - analog voltage

Acknowledgements

- Towards Edge Computing LDRD
  - Collaboration between Instrumentation Division (IO) and Computational Science Initiative (CSI)
  - https://www.bnl.gov/ldrd/
- Instrumentation Division
  - Adnan Zaman, Sioan Zohar, Jack Fried
- Computational Science Initiative
  - Yihui Ren, Sandeep Mittal, Shinjae Yoo
- Physics Department
  - Jin Huang
- Deep Underground Neutrino Experiment (DUNE)
- CERN ATLAS Experiment