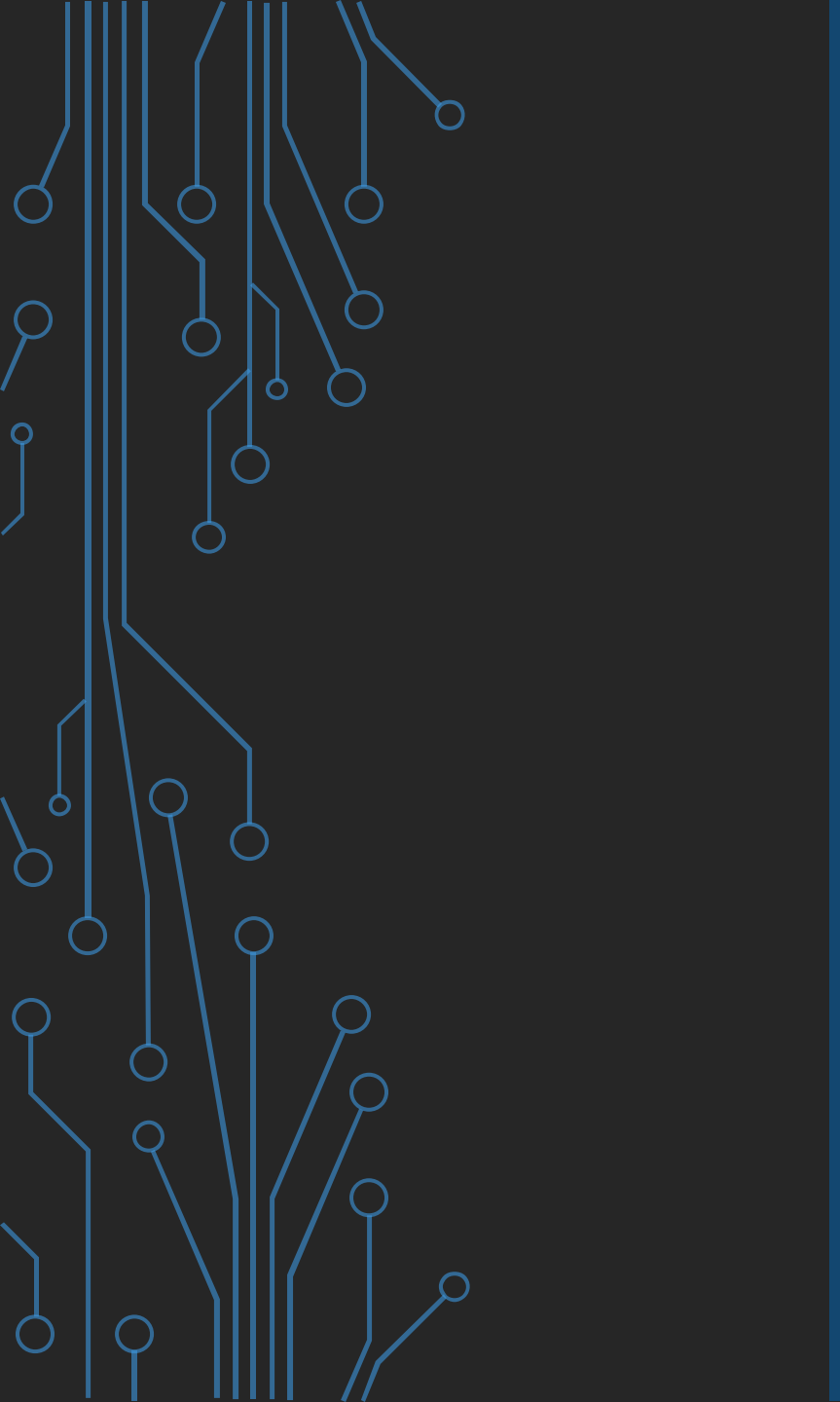
SURVEY ON LATEST DEVELOPMENTS IN DESIGN OF ARTIFICIAL INTELLIGENCE SYSTEM ON CHIP (SOC)

PAPER INFORMATION

- K. J. Lee, J. Lee, S. Choi and H. -J. Yoo, "The Development of Silicon for AI: Different Design Approaches," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 12, pp. 4719-4732, Dec. 2020, doi: 10.1109/TCSI.2020.2996625.
- **URL**: http://ieeexplore.ieee.org.ssl.sa.skku.edu:8080/stamp/stamp.jsp?tp=&arnumber=9104667&isnumber=9275400
- **Published in**: IEEE Transactions on Circuits and Systems I: Regular Papers (Volume: 67, Issue: 12, Dec. 2020)
- **Date of Publication**: 01 June 2020

TABLE OF CONTENTS

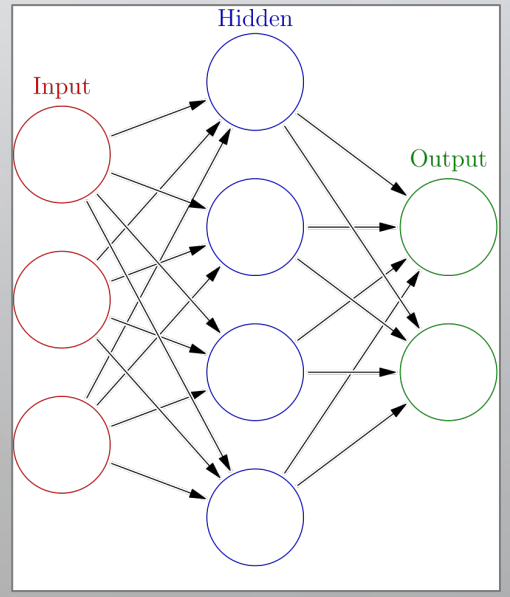
- **Section I:** Introduction
- **Section II:** Related Works about NN processor design
  - Previous NN processor designs
  - Today's DNN and Neuromorphic Processors
- **Section III:** Review Analog, Digital, and Mixed-Mode hardware design implementations
- **Section IV:** Review of today's fully digital DNN Processors
- **Section V:** Review Neuromorphic Processors
- **Section VI:** Insights and perspectives on future research directions
- **Section VII:** Conclusion



SECTION 1
INTRODUCTION

NEURAL NETWORK (NN): DEFINITION

- Subset of machine learning and are the heart of deep learning algorithms.
- Reflect the behavior of the human brain, allowing computer programs to recognize patterns and solve common problems in the fields of AI, machine learning, and deep learning
- Comprised on node layers (input layer, one or more hidden layers, and an output layer)

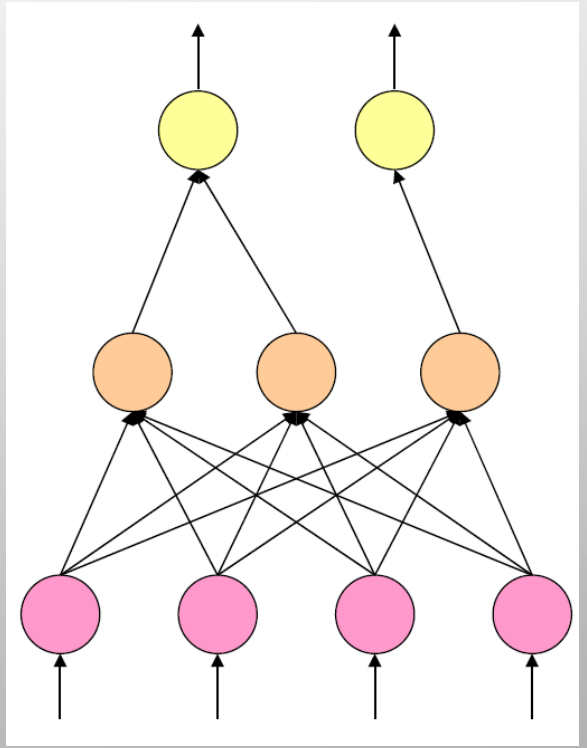


Three Different type NNs

- **Multi-layer perceptrons (MLPs):** comprised on input layer, a hidden layer or layers, and an output layer. Comprised of Simoid neurons. Computer vision, natural language processing
- **Convolutional neural network (CNNs):** utilized for image recognition, pattern recognition, and/or computer vision. Harness principles of linear algebra, particularly matrix multiplications to identify patters within an image
- **Recurrent neural networks (RNNs):** identified by their feedback loops. Use time-series data to make predictions about future outcomes, such as stock market predictions or sales forecasting

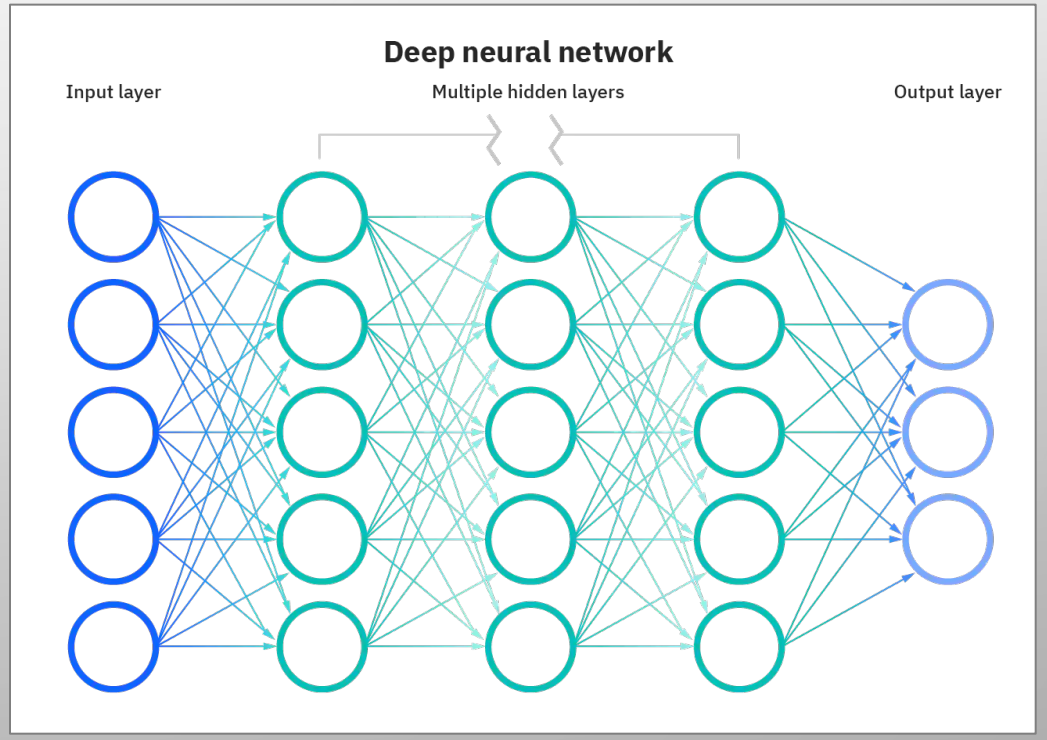
NEURAL FUZZY (NF): DEFINITION

- Combination of Neural Networks and Fuzzy Logic
- Fuzzy logic is based on the observation that people make decision based on imprecise and non-numerical information
- Fuzzy models or sets are mathematical means of representing vagueness and imprecise information
- Main strength of Neuro-Fuzzy systems is that they are universal approximators with the ability to solicit interpretable IF-THEN rules



DEEP NEURAL NETWORK (DNN): DEFINITION

- “Deep” in deep learning refers to the depth of layers in a Neural Network
- Neural Network with more than 3 layers can be considered a deep learning algorithm



SYSTEM ON CHIP (SOC): DEFINITION

- SoC is an integrated circuit that integrates all or most components of a computer or other electronic system
- These components almost always include a CPU, memory, input/output ports and secondary storage, often alongside other components such as radio modems and graphics processing unit (GPU) all on a single substrate (silicon) or microchip
- Use less power and take up less space than multichip counterparts
- Becoming increasingly popular with the growth of IoT, Edge, and mobile computing

NEUROMORPHIC CHIPS: DEFINITION

- Neuromorphic Computing is concerned with **emulating the neural structure and operation of the human brain,** as well as probalistic computing, which creates algorithmic approaches to dealing wit the uncertainty, ambiguity, and contradiction in the natural world
- Key challenges in Neuromorphic Research are **matching a human's flexibility, and ability to learn** from unstructured stimuli with the energy efficiency of the human brain
- The computational building blocks within neuromorphic computing systems are **analogous to neurons, Spiking Neural Networks (SNNs)** are a novel model for arranging those elements to emulate neural networks that exist in biological brains
- **Neuromorphic Chips** are based on specialized architecture that is optimized for SNN algorithms, supports the operation of SNNs that do not need to be trained in the conventional manner, and become smarter over time

HISTORY OF DEEP NEURAL NETWORK (DNN)

- There were various types of machine intelligence algorithms
  - Multilayer perceptron (MLP)
  - Fuzzy inference system (FIS)
  - Neuro-fuzzy system (NF)
- It tool several decades for them to become today's Deep Learning used for wide range of computer vision applications such as image classification, object detection, and autonomous vehicles
- Variant of DNN used for different applications
  - Convolutional Neural Network (CNN) : image processing
  - Recurrent Neural Network (RNN) : natural language processing

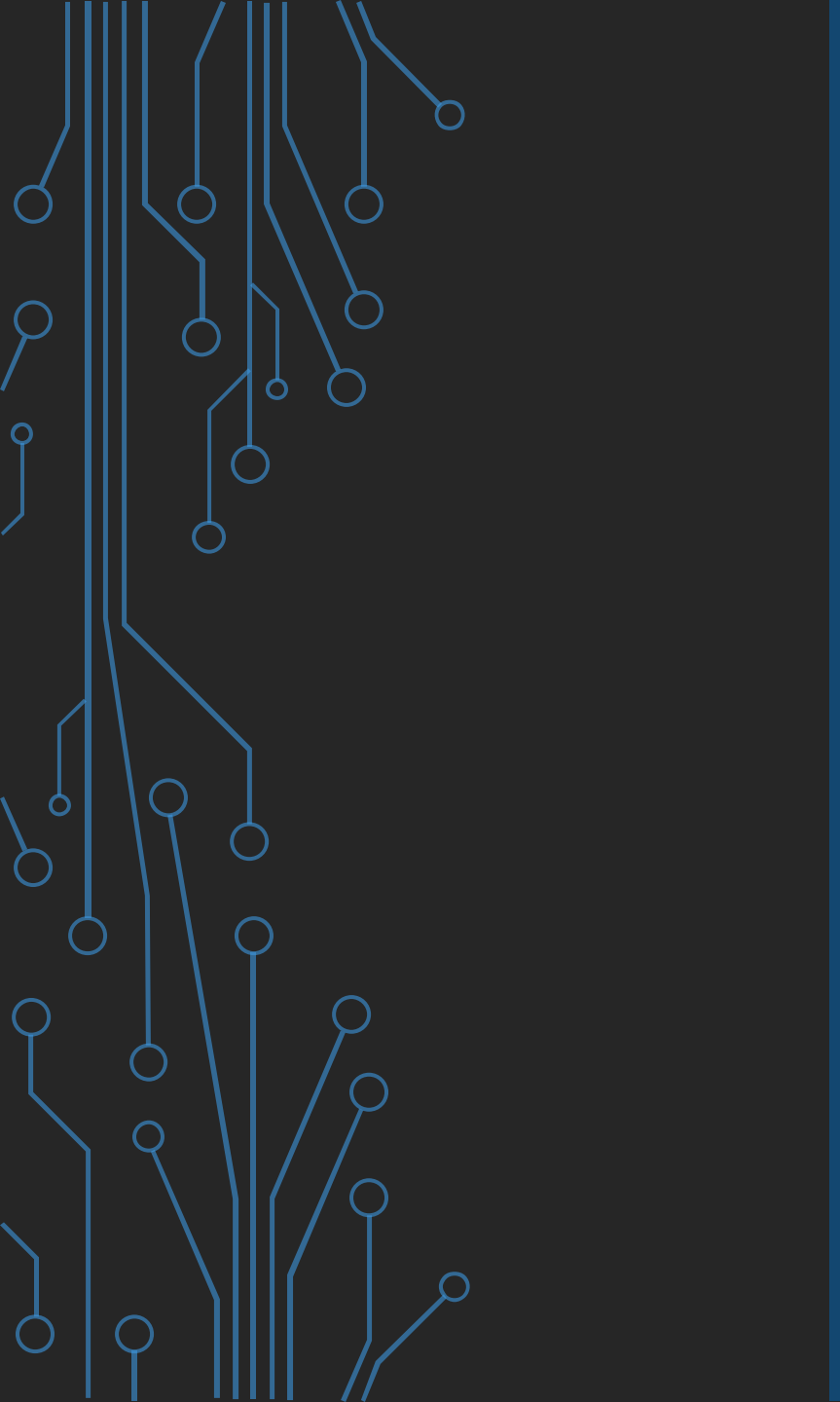
DEEP NEURAL NETWORK (DNN)

- Consist of hundreds of layers requiring huge number of computations and memory footprints
  - Most applications target bulky systems with high performance GPUs or servers rather than mobile applications
  - Needs of mobile DNN are increasing nowadays
  - Mobile DNN applications lack processing speed requirements, limited hardware resources and limited power budget
  - Hardware accelerator design is gaining more attention worldwide and they are actively investigated to efficiently run DNNs in real time with low power

PROCESSOR HARDWARE DESIGN APPROACHES

There have been many efforts to develop dedicated processors from the early age of neural networks with various hardware design approaches and it is important to analyze the previous approaches to develop advance System on Chips (SoCs).

- **Analog** : VLSIs enable low-cost parallelism with low-power computation, but their inaccurate circuit parameters induced by noise and low precision degrade accuracy
- **Digital :** Can achieve high accuracy, flexibility, and programmability, but they consume huge power and area due to the large amount of data transaction and fast operation speed
- **Mixed-Mode :** Have advantages of both analog and digital implementations obtaining low-power consumption within small area, but it suffers from domain conversion overhead cost (speed, area, power)



SECTION II
RELATED WORKS

NEUROMORPHIC DESIGNS

2014: Kim
proposed sparse coding ASIC to enable training of sparse representation of images for feature detection and recognition using spiking neural network

2015: Kim
developed a simple object recognition system composed of spiking neural network inference module

Jan 2015: Lu
implemented clustering algorithm in analog domain with floating gate non-volatile memory

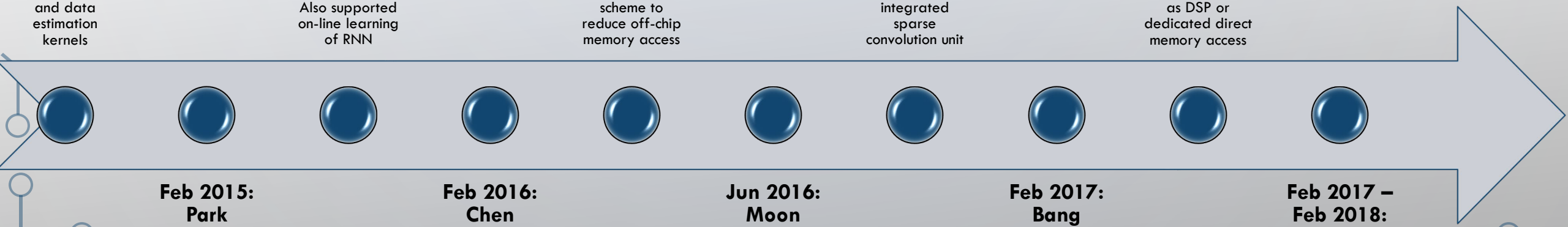
Feb 2015: Zhang
implemented matrix-multiplying ADC that enables multiplications with input samples

Feb 2016: Lee
designed energy-efficient matrix multiplier with switched capacitor scheme for classification applications

Jun 2016: Ambrogio used resistive switching memory RRAM to emulate spiking neural network

Jun 2016: Zhang
Proposed in-memory computation scheme using standard 6-T SRAM array

DNN PROCESSOR DESIGN



June 2014:
Tsai proposed digital DL processor for big-data applications such as data filtering and data estimation kernels

Feb 2015:
Park
attempted to implement DNN training on silicon with massive parallelism. Develop user experience using embedded DL SoC. It is capable of both simple NN training and inferencing

Feb 2016:
Lee developed SoC for advanced driver assistance system using RNN and FIS accelerators. Also supported on-line learning of RNN

Feb 2016:
Chen
proposed a CNN accelerator utilizing data reuse pattern and its dedicated hardware architecture

Feb 2016:
Sim designed digital CNN processor with multi-range MAC unit and kernel compression scheme to reduce off-chip memory access

Jun 2016:
Moon
designed a CNN processor with 2-D MAC array and scalable bit precision, and enhanced the energy-efficiency with zero guarding and wide-range voltage-frequency scaling

Jun 2016:
Kang
proposed a convolution restricted Boltzmann machine processor to infer support vector machine classifier with integrated sparse convolution unit

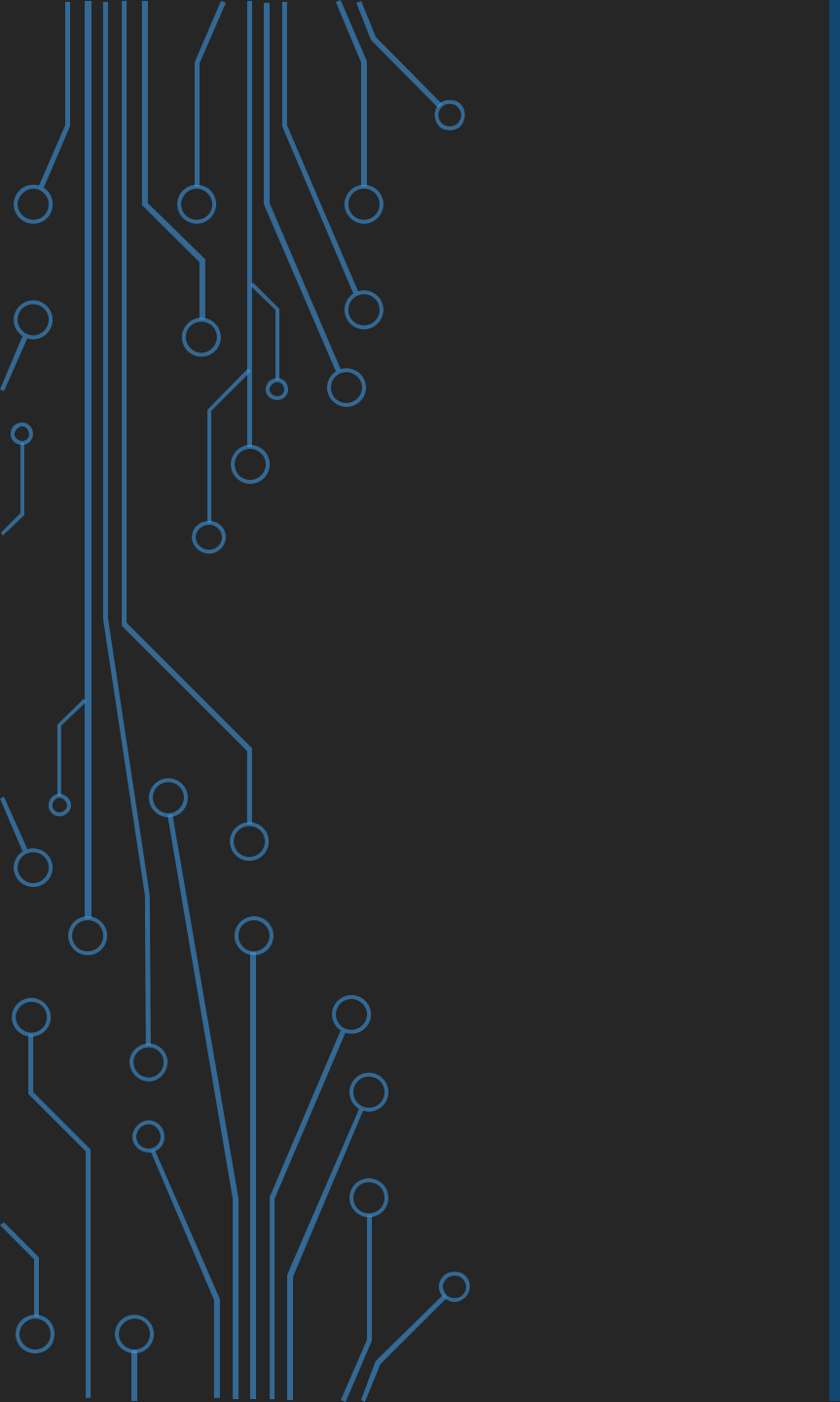
Feb 2017:
Bang
introduced efficient accelerators for fully-connected DNN with their hardware architecture optimized for matrix multiplication

Feb 2017:
Desoli
integrated low power DNN SoC with other hardware components such as DSP or dedicated direct memory access

Feb 2017 – Feb 2018:
More works were exposed to integrate digital circuits for different type of DNNs from FC-DNNs to CNN and RNN

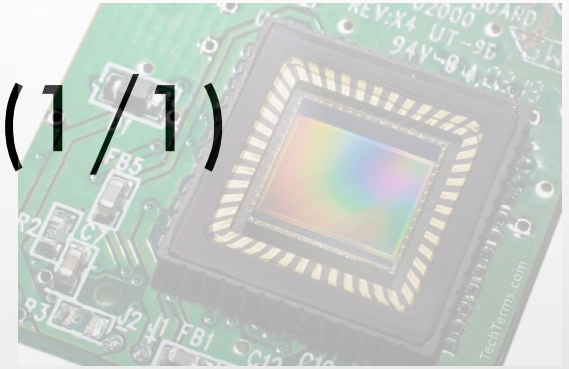
DNN & NEUROMORPHIC PROCESSOR DESIGN

The trend of recent design techniques show high-performance DNN SoCs are implemented in digital while ultra-low-power SoCs are mixed-mode implementation

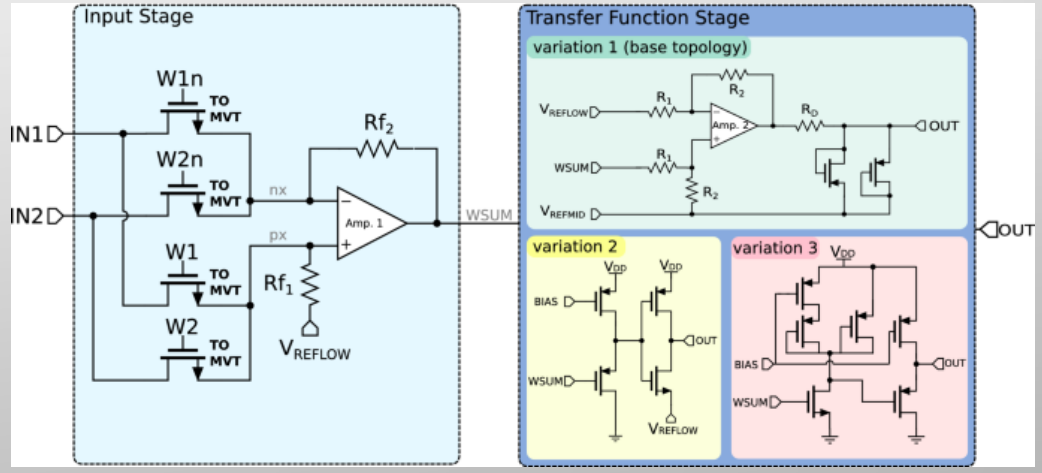


SECTION III
REVIEW ANALOG, DIGITAL, AND MIXED-MODE HARDWARE DESIGN IMPLEMENTATIONS

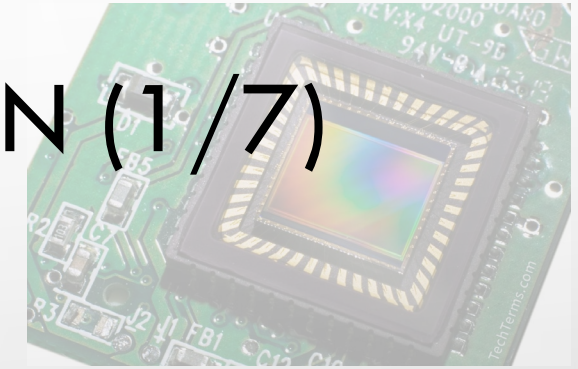
ANALOG IMPLEMENTATION (1/1)



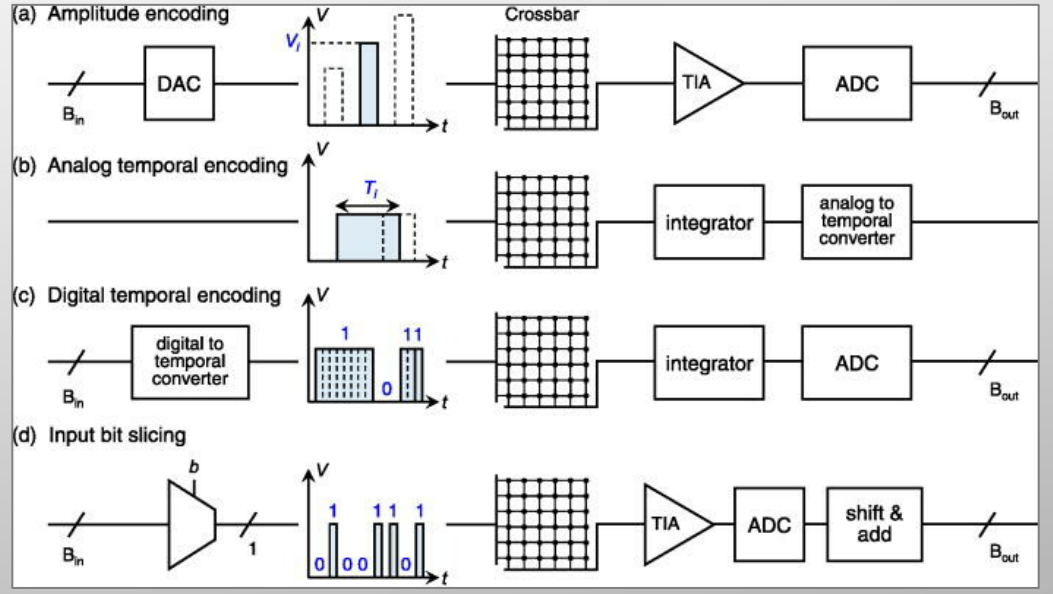
- Proposed for functional blocks used in NN/NF such as synapse or sigmoid activation generation
- Analog circuit lacks programmability while NN/NF operations requires training and setting many parameters by nature
- Most of the analog-based neural networks were assisted by digital circuits for flexible control of the analog arrays
- Full integration of NN/NF could be classified as mixed-mode and digital design



MIXED-MODE IMPLEMENTATION (1/7)



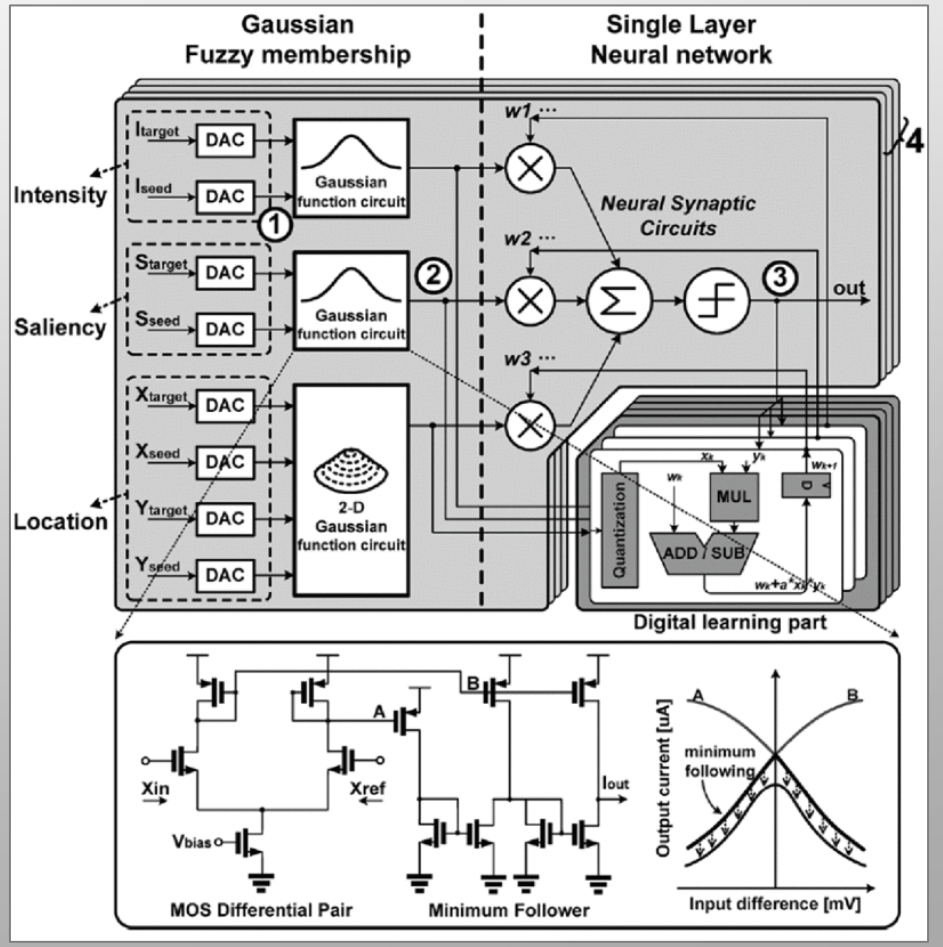
- **Analog Circuits**: used for its low power consumption and save area (current mode, voltage mode)
- **Digital Circuits:** used for training and controlling the analog parameters for its high programmability and accurate calculations
- Although analog/mixed-mode design facilitates low-power and energy efficient designs, it is often complicate due to PVT variation as well as the domain conversion overhead in terms of conversion speed, area, and power
- Analog and digital domains must be carefully divided and balanced



MIXED-MODE IMPLEMENTATION (2/7)

- Designed as a functional block to achieve high performance and low power
- Detects contour of target objects around the ambiguous object boundaries in the input image
- Gaussian fuzzy membership function and single-layer perceptron used to measure the similarity of neighboring pixels and to clarify boundaries
- Consists of a current-mode analog datapath for feedforward neuro-fuzzy operation and a digital processor for training the synaptic weights
- For weight multiplication, a binary weighted current mirror is used as current multiplier
- Fabricated in 0.13 µm CMOS technology and reduced area and power by 59% and 44% compared with the fully-digital implementation in the same process technology

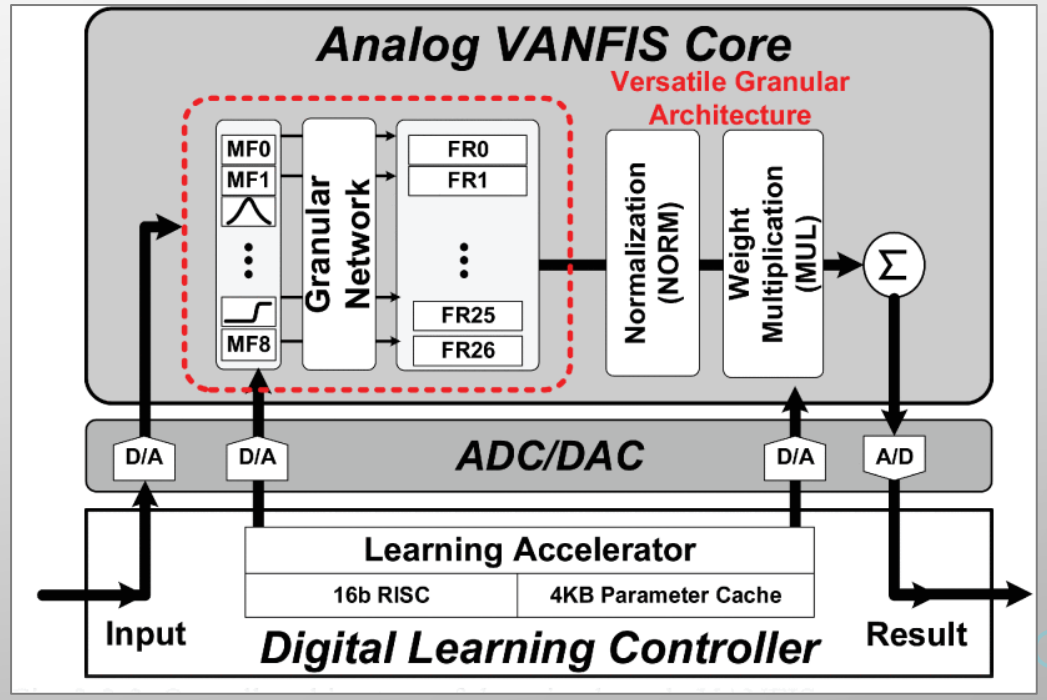
NF Mixed Mode Circuit for Object Detection Hardware Architecture

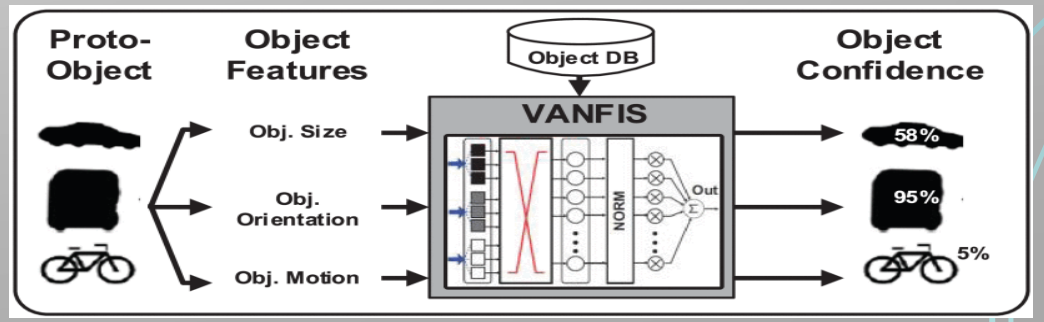


MIXED-MODE IMPLEMENTATION (3/7)

- Used for object classification and dynamic workload prediction to increase energy efficiency of the object recognition SoC
- Measures the similarity between proto-objects and fuzzy rules of target objects to estimate the confidence level of input object
- Also used for efficient hardware control by comparing current status of the SoC with pre-trained workload history to predict the future workload to make SoC achieve the highest energy efficiency
- Implemented with current-mode circuits, fabricated in 0.13 µm CMOS technology, VANFIS processor saved area and power by 56% and 85%, respectively, compared with the equivalent digital implementation

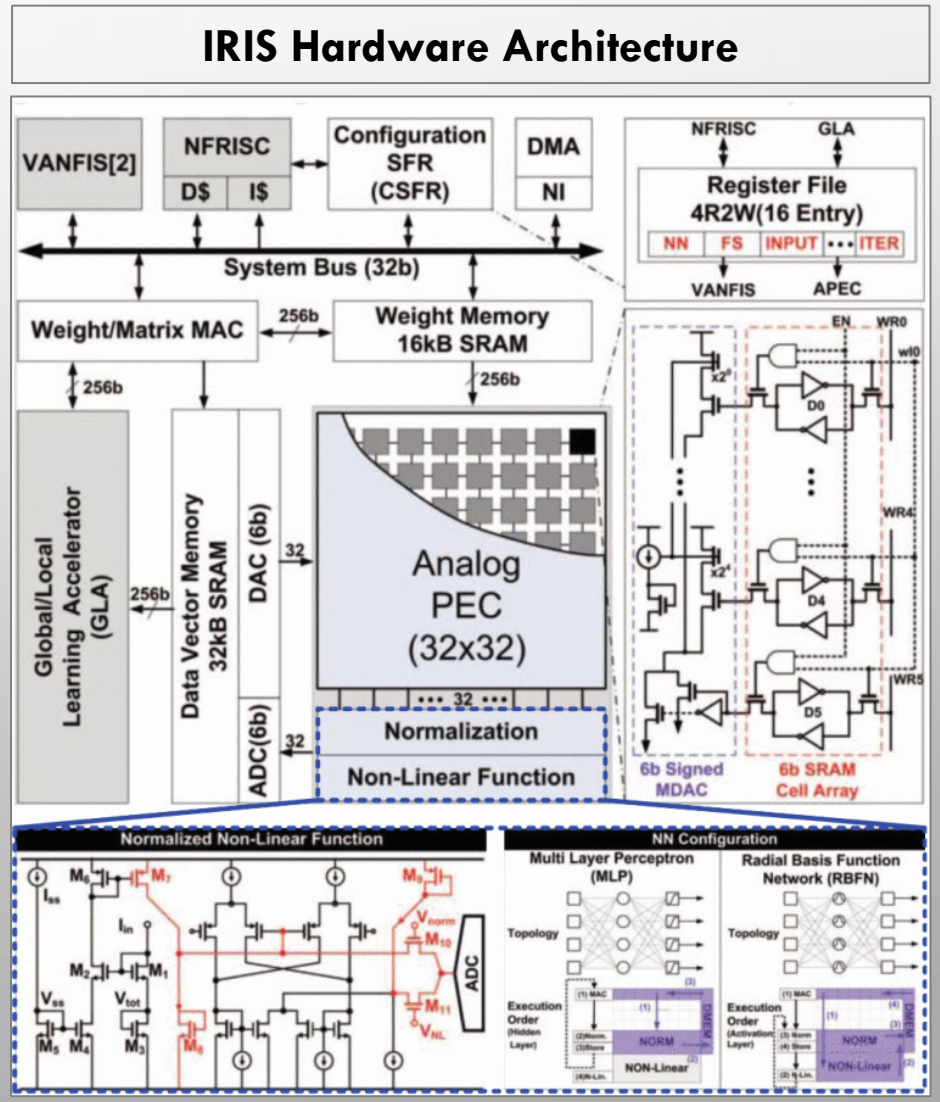
VANFIS Hardware Architecture





MIXED-MODE IMPLEMENTATION (4/7)

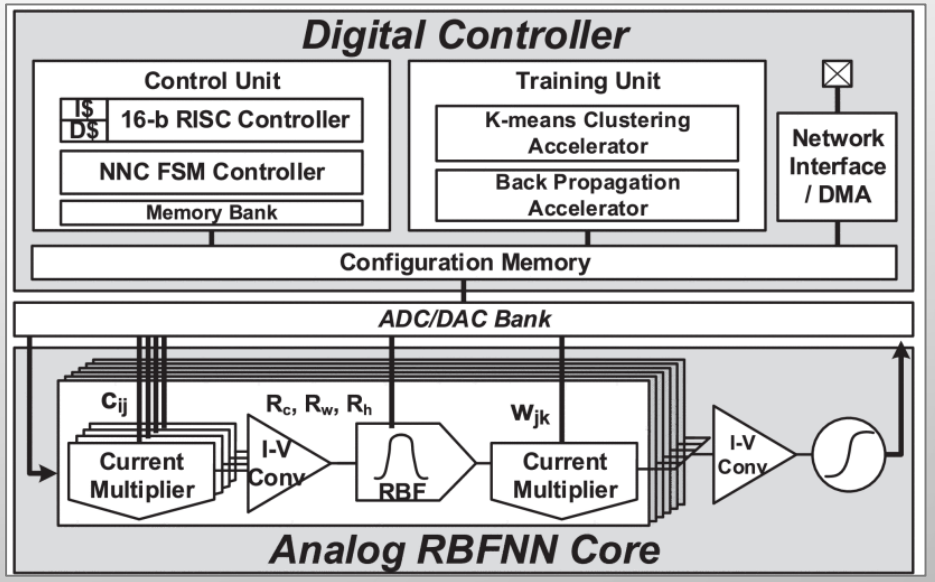
- Introduced for multi-purpose application of NN and FIS
- Composed of a reconfigurable analog PE cluster that is capable of computing various NN and FIS topologies, a global/local learning accelerator, a MAC unit, memory banks, a RISC controller, and data converters
- Analog core contains 32 x 32 Pes for parallel MAC operation of NN
- Capable of accelerating MLP, radial basis function neural network (RBFNN) and RNN by changing signal paths of the analog PE cluster
- Reduced power and area by 71.2% and 54% compared with equivalent digital design
- Fabricated in 0.13 µm CMOS technology, achieves 1 mJ/frame energy efficiency and consumes 57 mW on average for object recognition



MIXED-MODE IMPLEMENTATION (5/7)

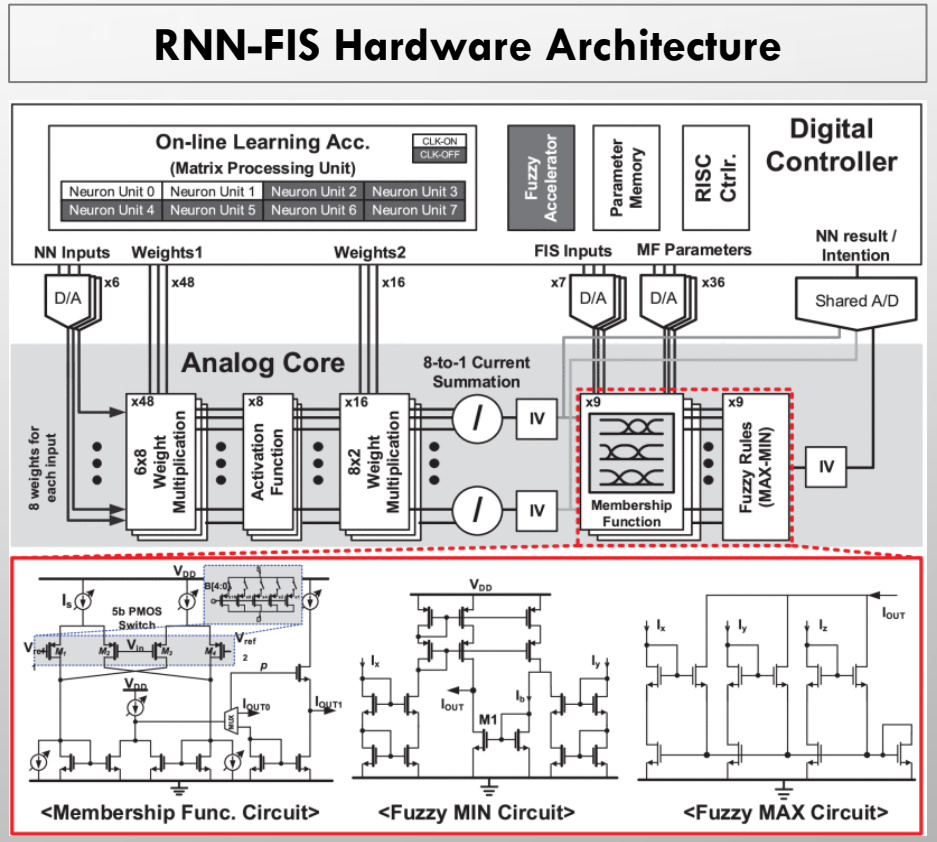
- Widely used as a classifier for its high accuracy
- RBFNN classifier in current-mode circuit for low power yet highly accurate scene classification
- Contains temperature and supply voltage variation compensation circuits, which outputs stable current despite variations
- Classification is performed in the analog core, while digital controller is in charge of controlling configuration of analog connections and training of RBFNN parameters
- Classifies global scene by taking HMAX features as input
- Fabricated in 0.13 μm CMOS technology, saves area and power by 84% and 82% respectively, compared with fully digital implementation

RBFNN Hardware Architecture



MIXED-MODE IMPLEMENTATION (6/7)

- NF processor proposed for automotive black box
- Algorithm alerts drivers to the risky objects that are about to be collided while it triggers surveillance recording when object is getting closer
- Hardware consists of analog core for feedforward RNN-FIS acceleration and a digital controller for online training and control of the analog parameters
- Fabricated in 65 nm CMOS technology and achieves high performance (502 GOPS) in driving mode and low power (0.984 mW) in parked mode
- Total area and power consumption are reduced by 64% and 39% respectively, compared with the fully digital implementation

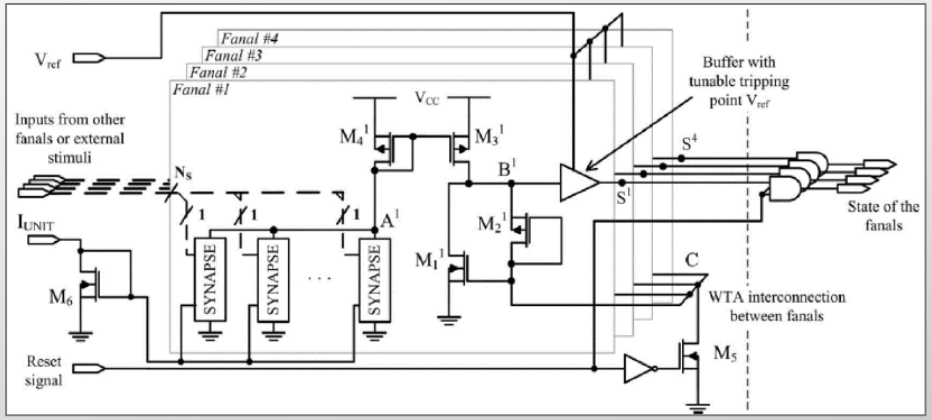


MIXED-MODE IMPLEMENTATION (7/7)

Digitally programmable Multipliers

- Mixed-mode neural network by proposing digitally programmable multipliers for linearization of NTC thermistor
- Gain of analog multiplier is controlled by digital switches, but need adder circuit since they used voltage mode circuits
- Designed with 0.18 µm CMOS, consumes 0.538 mW with 100 MHz
- Focused on the neuronal circuit design rather than complicate NN applications, containing only 2 hidden neurons which is very primitive

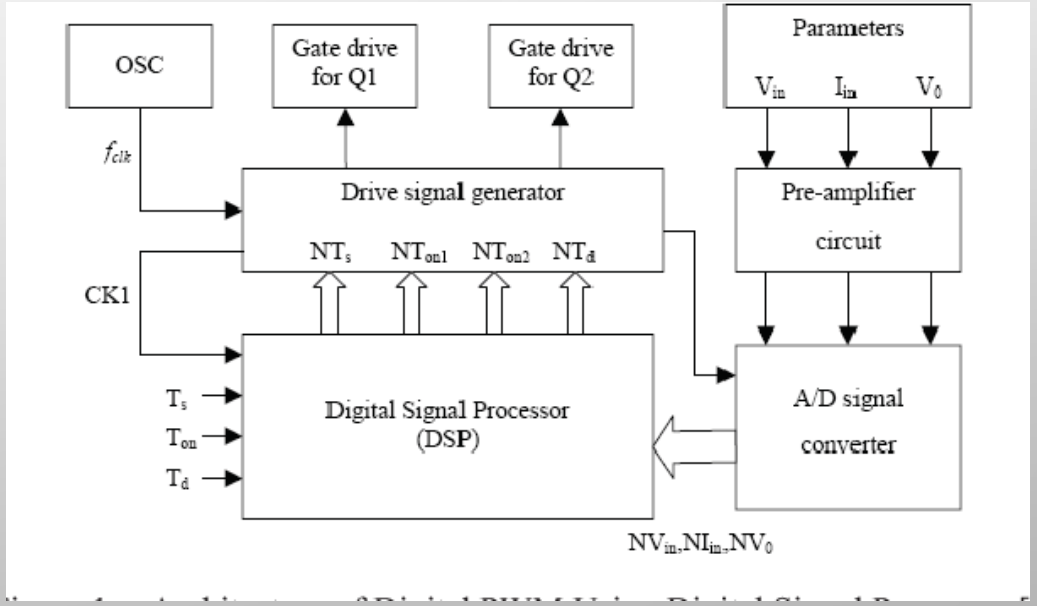
Encoded NN Hardware Architecture



- Consists of analog neuron computing nodes for low power and cost design with digital network for communication among the nodes
- Fabricated in 64 nm CMOS technology and achieves 68 fJ ultra low energy operation
- Too simple compared to other processors

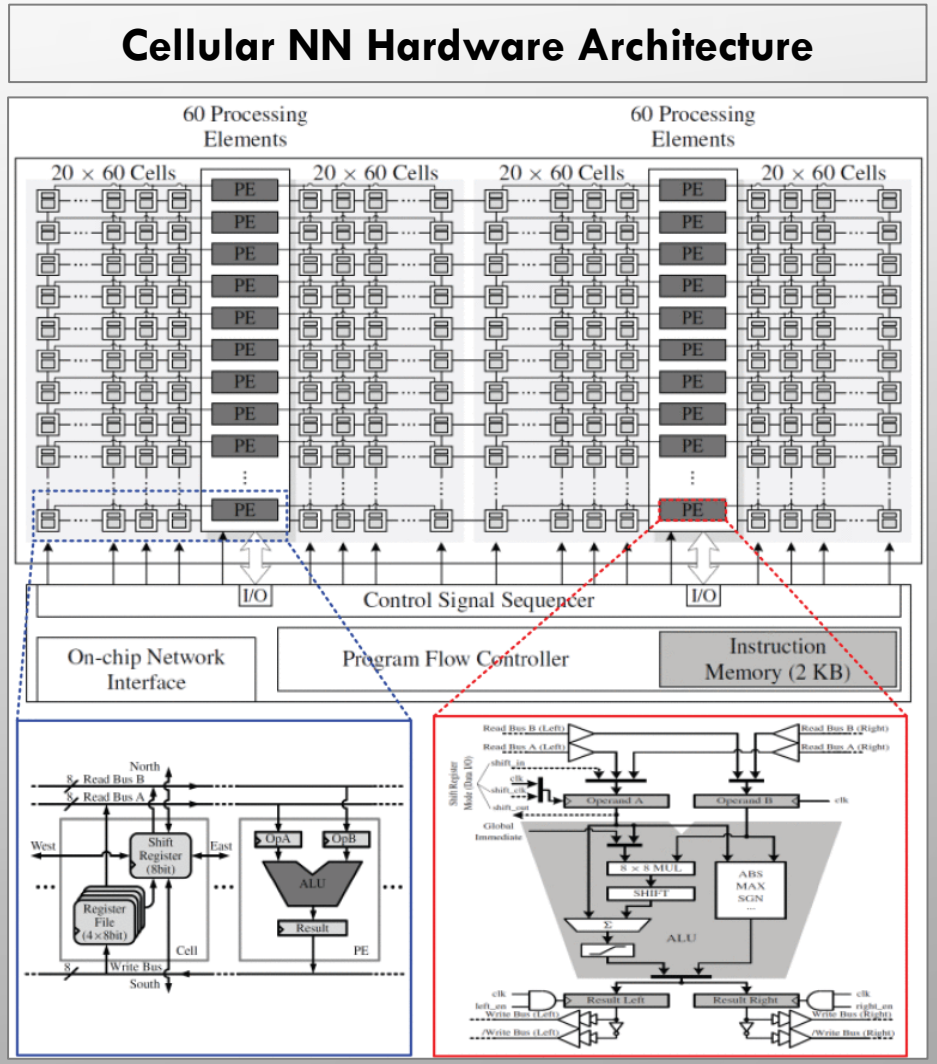
DIGITAL IMPLEMENTATION (1/4)

Although mixed-mode design has advantages in low power and small area implementation, digital processor has advantages in its speed and high precision as the technology gets smaller in addition to the removal of data conversion overhead in image processing



DIGITAL IMPLEMENTATION (2/4)

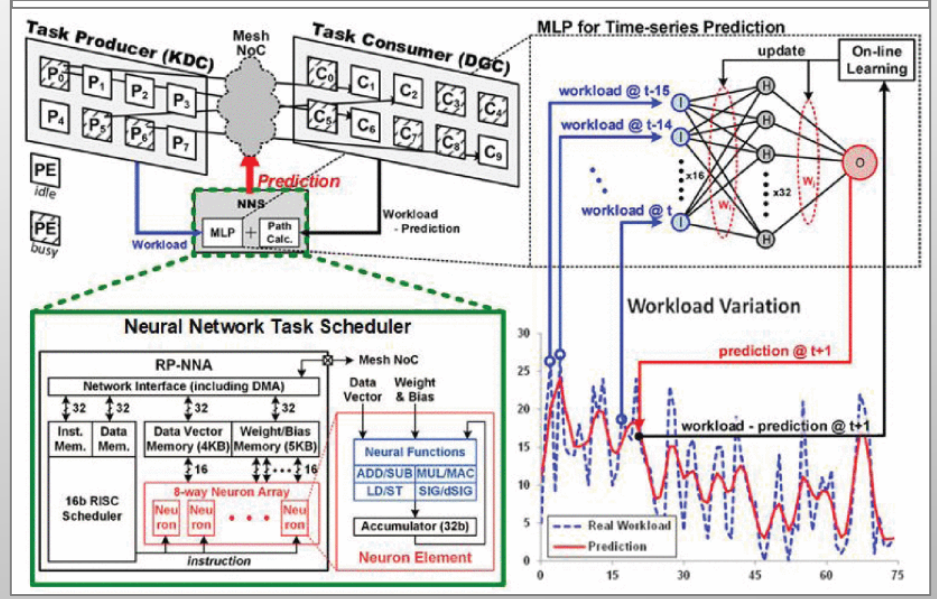
- Combines the flexibility of digital approach with high performance of fully parallel cell topology of analog approach
- Implemented in 0.13 µm CMOS technology and occupies 4.5 mm² area
- Consumes 84 mW running at 200 MHz
- Incorrect local features are drastically reduced to increase frame rate by 83% and reduce energy/frame by 45% without degradation in recognition accuracy



DIGITAL IMPLEMENTATION (3/4)

- NN task scheduler is designed for workload prediction to enhance energy efficiency by reducing time overhead on core-to-core allocation in a multi-core vision processor
- Estimates workloads in future frames with MLP and allocates producer-consumer pairs in advance to reduce network congestion
- Fabricated in 65 nm CMOS technology, consumes 4.9 mW while achieving 12.7 mJ/frame energy efficiency by reducing 24.4% of network latency on average

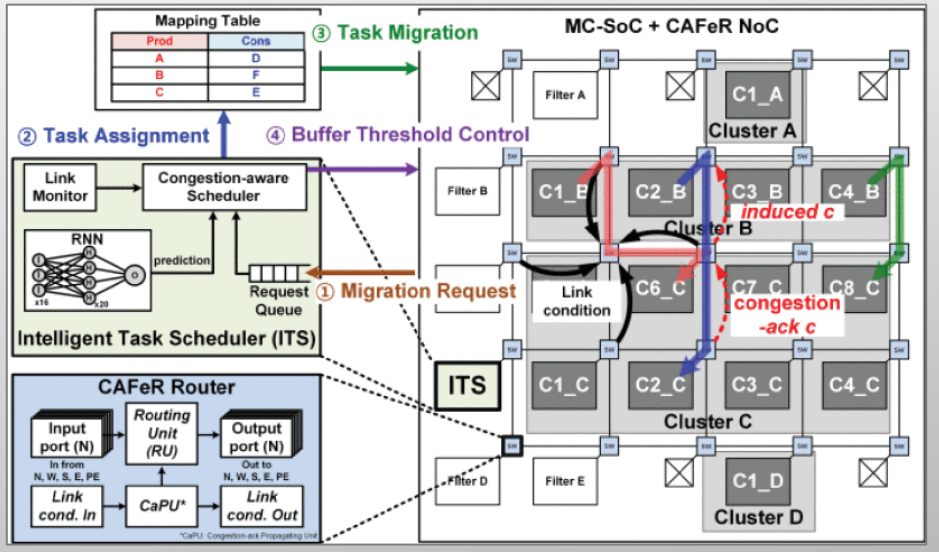
NN Task Scheduler Hardware Architecture



DIGITAL IMPLEMENTATION (4/4)

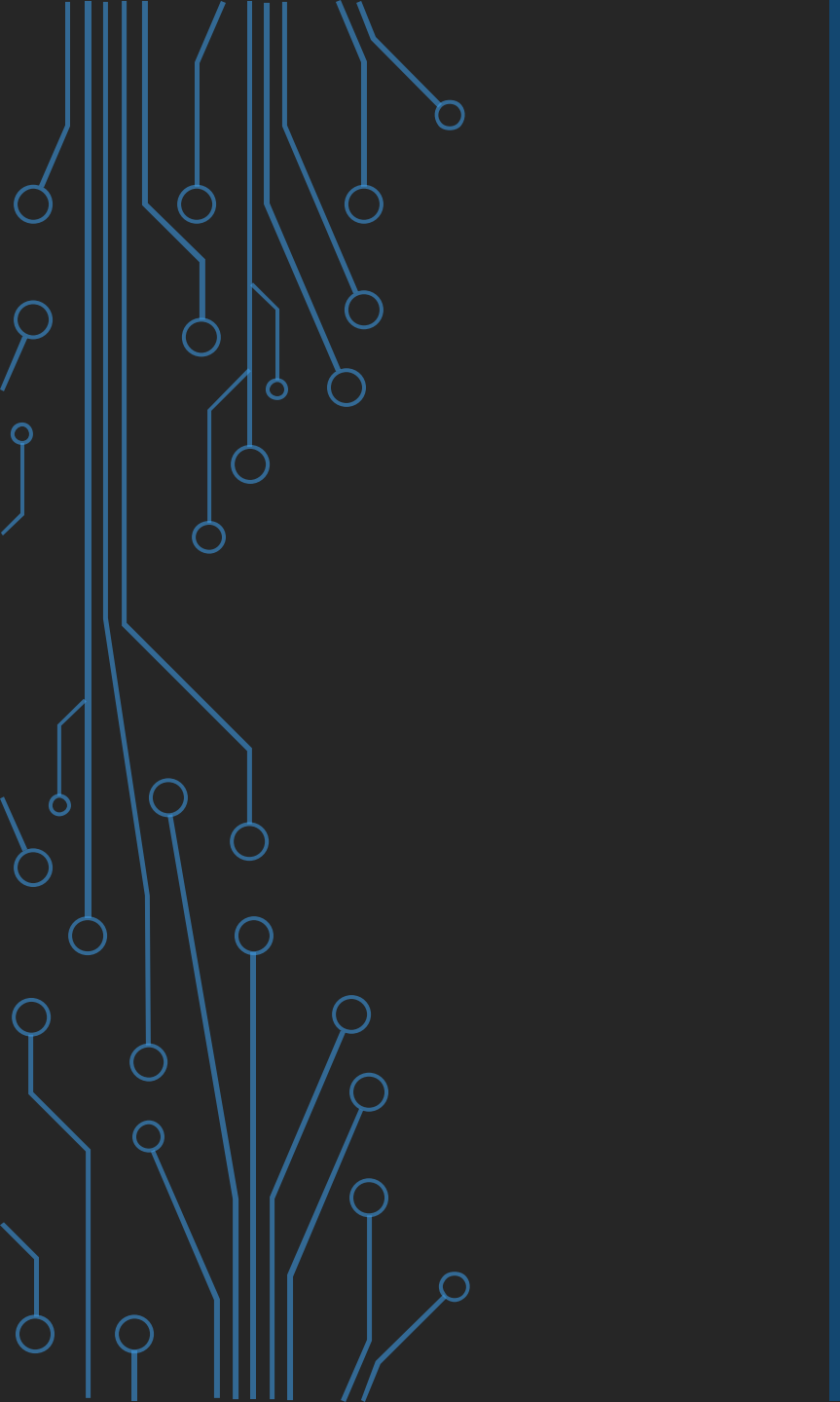
- To enhance the performance, an advanced version of NN scheduler using RNN with new NoC architecture dedicated to SIFT-based object recognition is proposed
- Also designed as a part of the object recognition SoC
- Improved the workload prediction accuracy to 91.4% and system throughput by 50.2%

RNN Task Scheduler Hardware Architecture



SUMMARY: HARDWARE IMPLEMENTATION

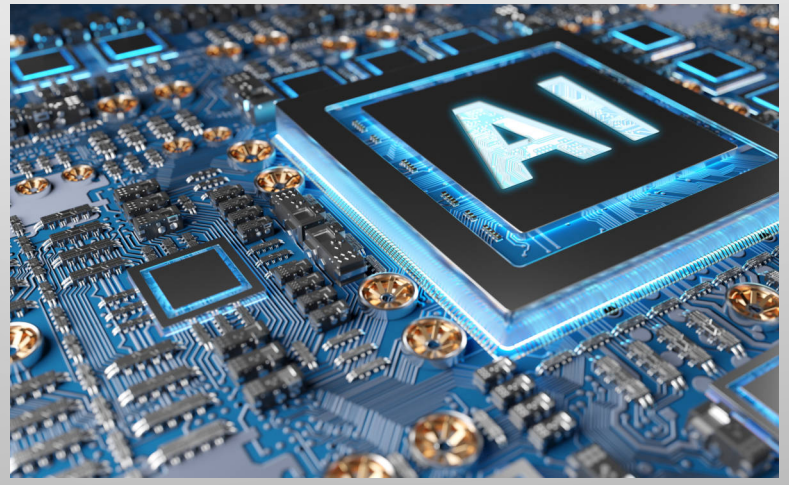
- Utilizing NN/NF algorithms as a functional block of SoC brings great improvements if they are used in right place with dedicated hardware architecture
- Although analog/mixed-mode design facilitates low-power and energy efficient designs, it is often complicate due to PVT variation as well as the domain conversion overhead in terms of conversion speed, area, and power
- Analog and digital domains must be carefully divided and balanced
- Although mixed-mode design has advantages in low power and small area implementation, digital processor has advantages in its speed and high precision as the technology gets smaller in addition to the removal of data conversion overhead in image processing



SECTION IV
DEEP LEARNING / DEEP NEURAL NETWORK SYSTEM ON CHIPS

ACCELERATOR: DEFINITION

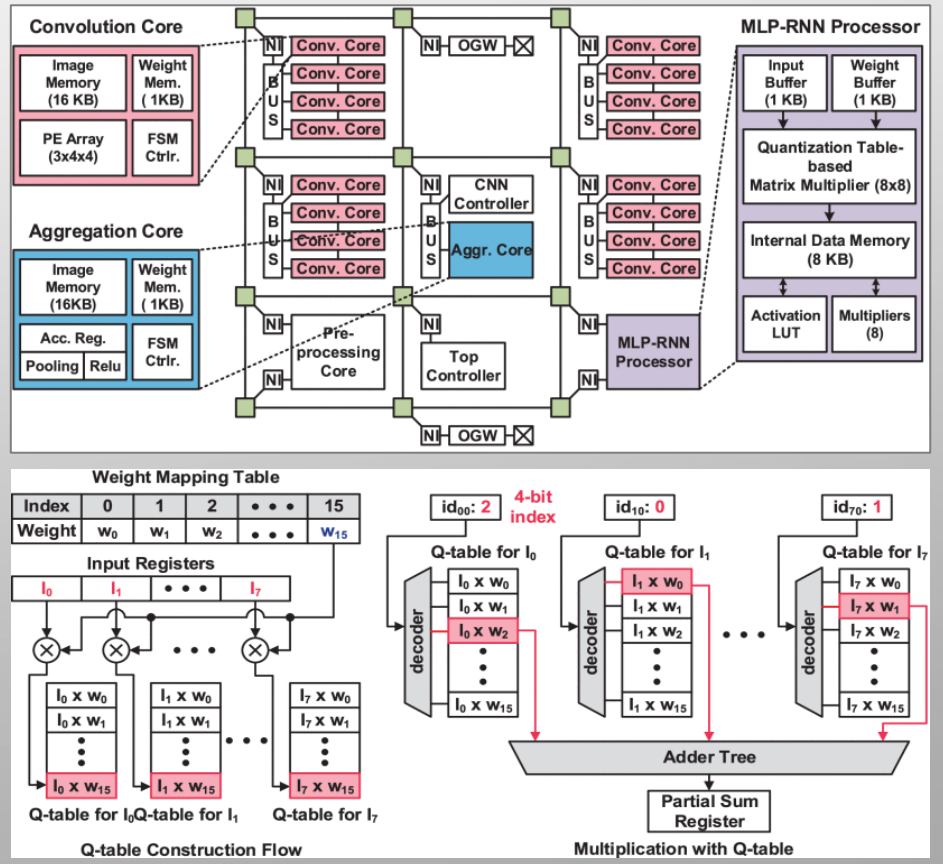
- A group of processors optimized for system performance by offloading and accelerating specific classes of software algorithms from the primary processor



ACCELERATING DL/DNN ALGORITHM IN SOC (1/4)

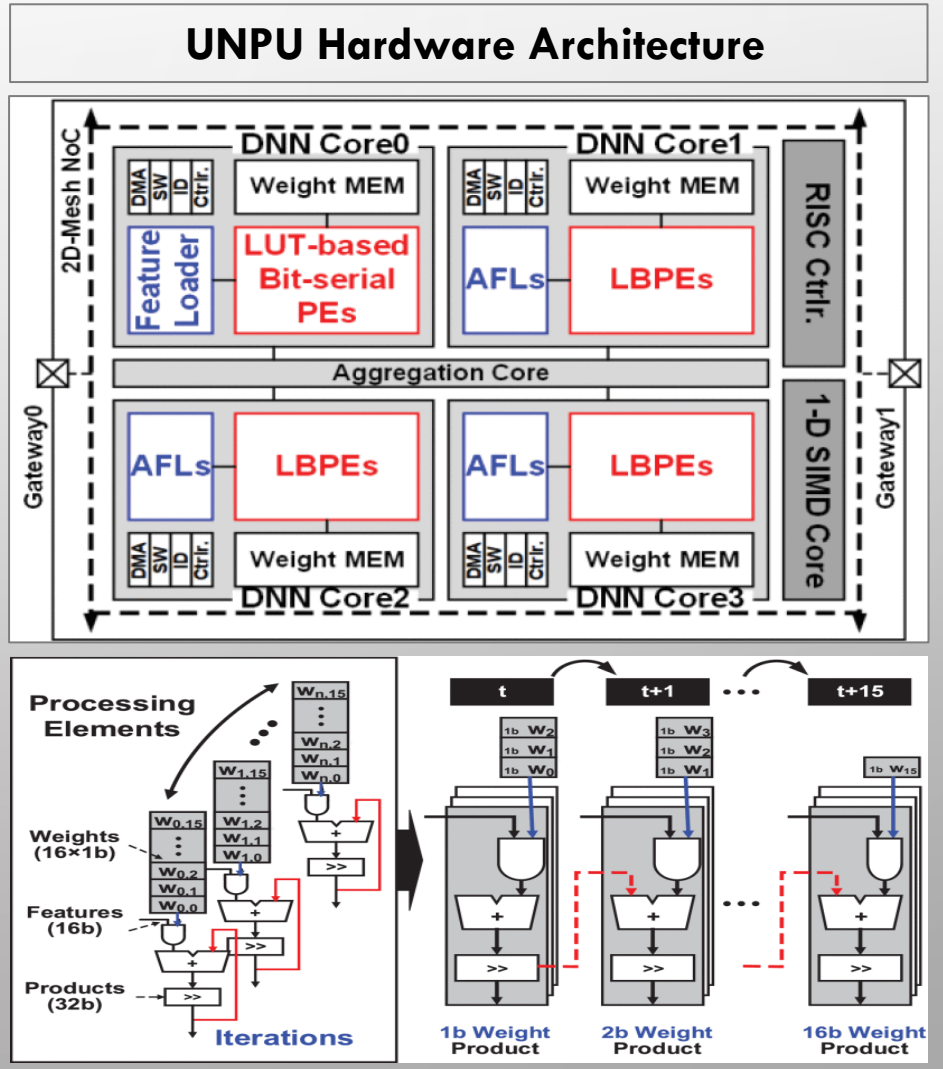
- DNN processing unit DNPU was introduced to accelerate both CNN and RNN in mobile environment
- Features to optimize CNN and RNN
  - Heterogeneous multi-core architecture
  - Mixed channel division scheme for CNN computation
  - Quantization table (Q-table) for efficient matrix product
- Matrix multiplication is performed by accessing table values without using multipliers, reducing the number of multiplication by 99%
- Fabricated in 65 nm CMOS technology and achieves 8.1 TOPS/W at 50 MHz, 0.77 V where peak power is 279 mW

UNPU Hardware Architecture



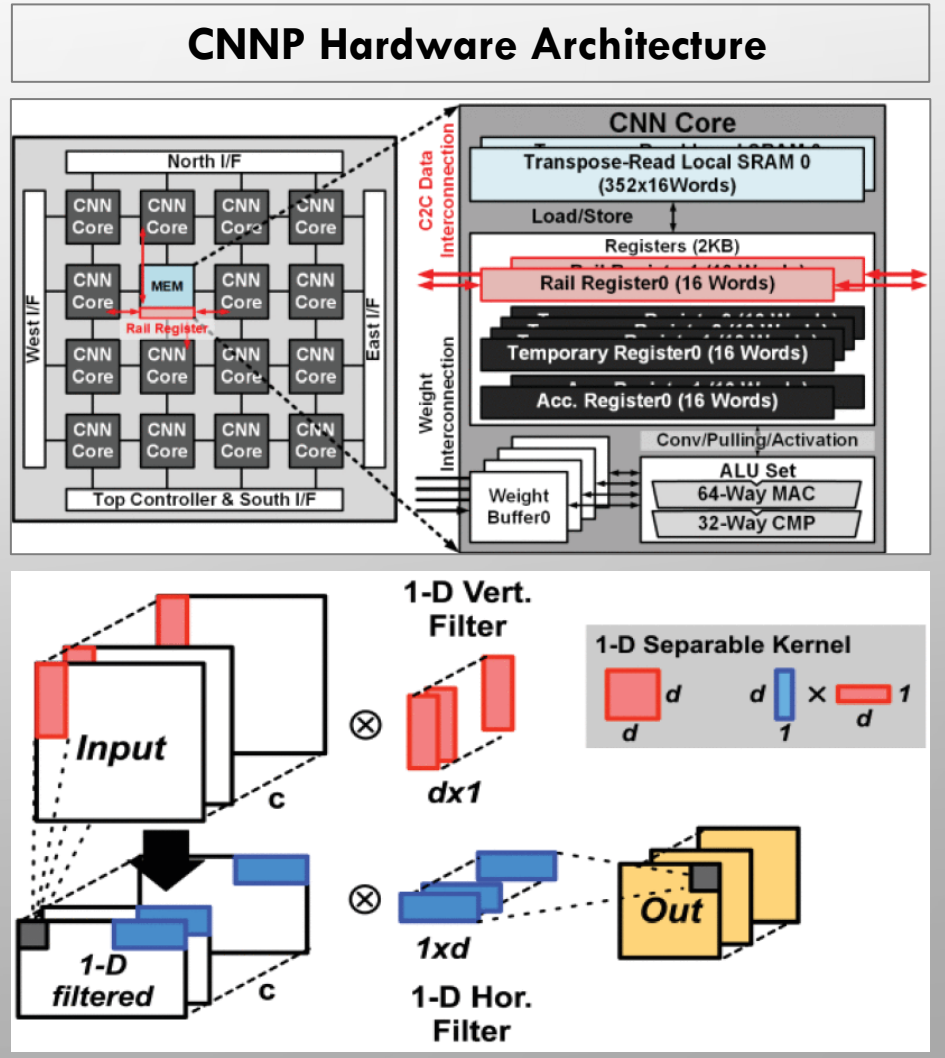
ACCELERATING DL/DNN ALGORITHM IN SOC (2/4)

- DNPU limited hardware utilization, if a DL application requires only one of CNN or RNN, the unused part of the DNPU wastes hardware resources
- Unified neural processing unit (UNPU) was presented to resolve limitations of DNPU
- Features:
  - Unified DNN core architecture
  - Fully variable weight bit precision
- Reduced energy consumption of MAC operation by 23.1% (16 bit), 27.2% (8 bit), 41.0% (4 bit) and 53.6% (1bit)



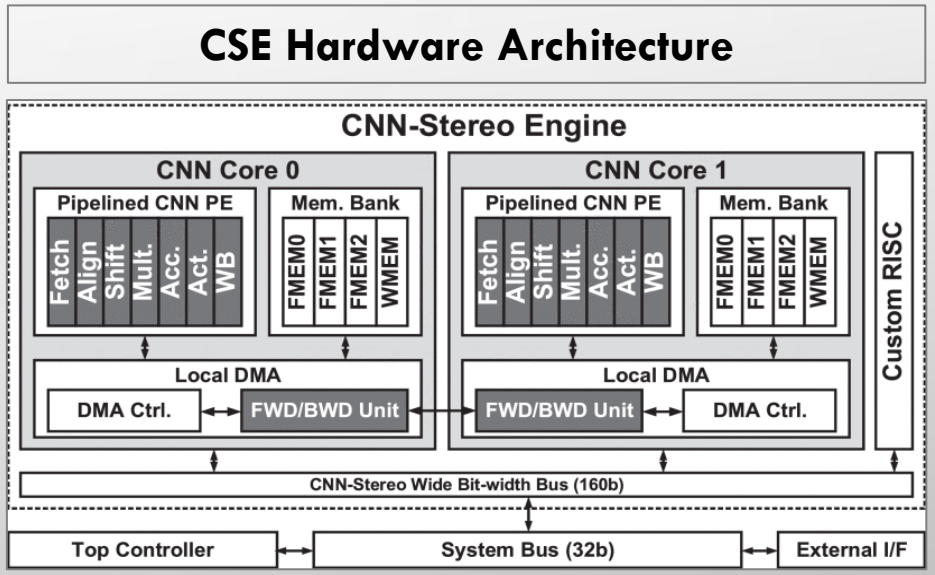
ACCELERATING DL/DNN ALGORITHM IN SOC (3/4)

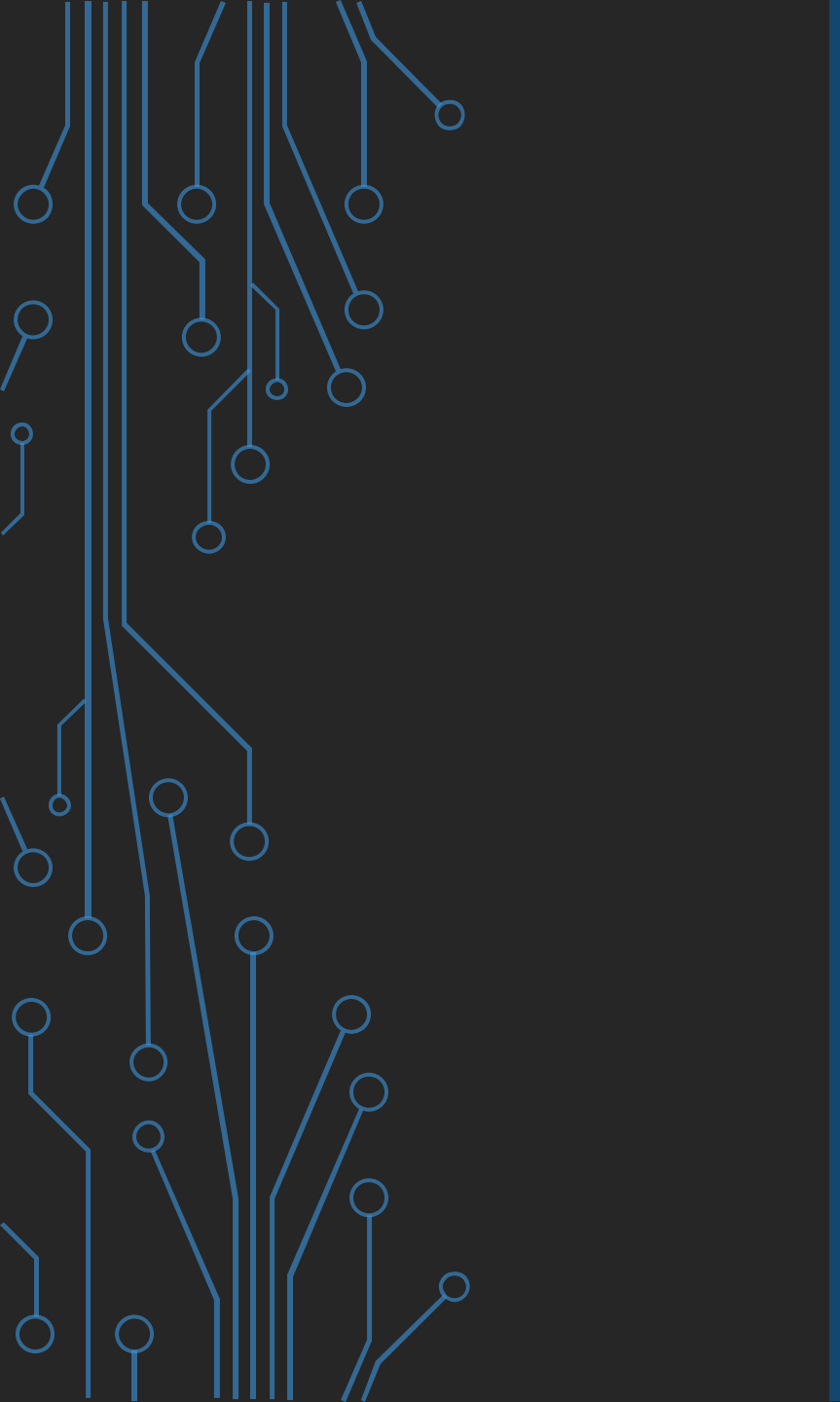
- CNN processor (CNNP) targeted user identification application on always-on sensor node, requiring highly accurate face recognition with low power consumption
- Main features for low power operation
  - Distributed memory CNN architecture
  - Separable filter approximation
- Utilized distributed memory architecture without separate global routing and memory to lower routing complexity, and bring external memory into multiple local on-chip memories for lower power CNN operation
- Reduced 78% overall energy, fabricated in 65 nm CMOS technology and achieved 5.3 mW



ACCELERATING DL/DNN ALGORITHM IN SOC (4/4)

- CNN-Stereo Engine (CSE) introduced to maximize the hardware utilization with two main features
  - Channel-wise parallel 1-D MAC operation
  - Core-to-core data balancing
- Increased parallelism in the channel direction rather than spatial direction
- Designed to handled 1D operations to remove the redundancy caused by the 2D workload method of CNNP
- Achieved 20% higher energy efficiency than CNNP
- Improved overall performance by 23.9% compared to CNNP



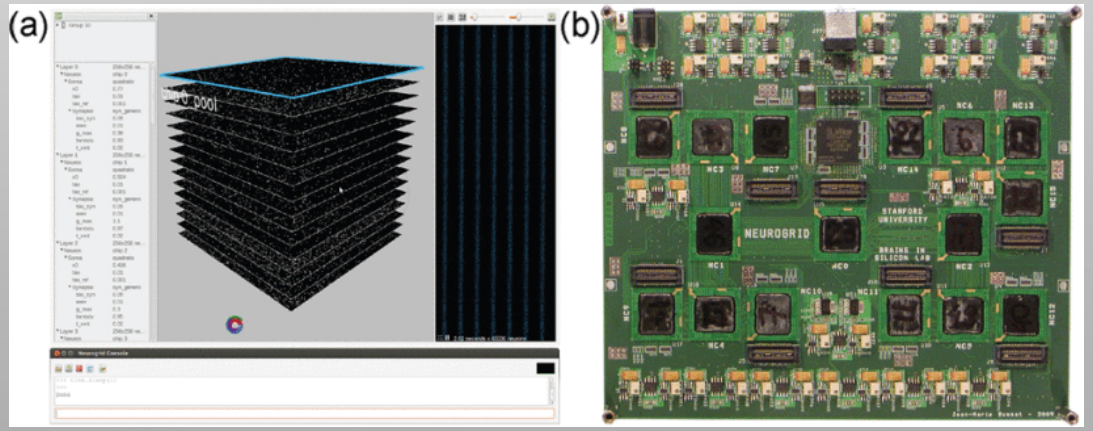


SECTION V
NEUROMORPHIC PROCESSORS

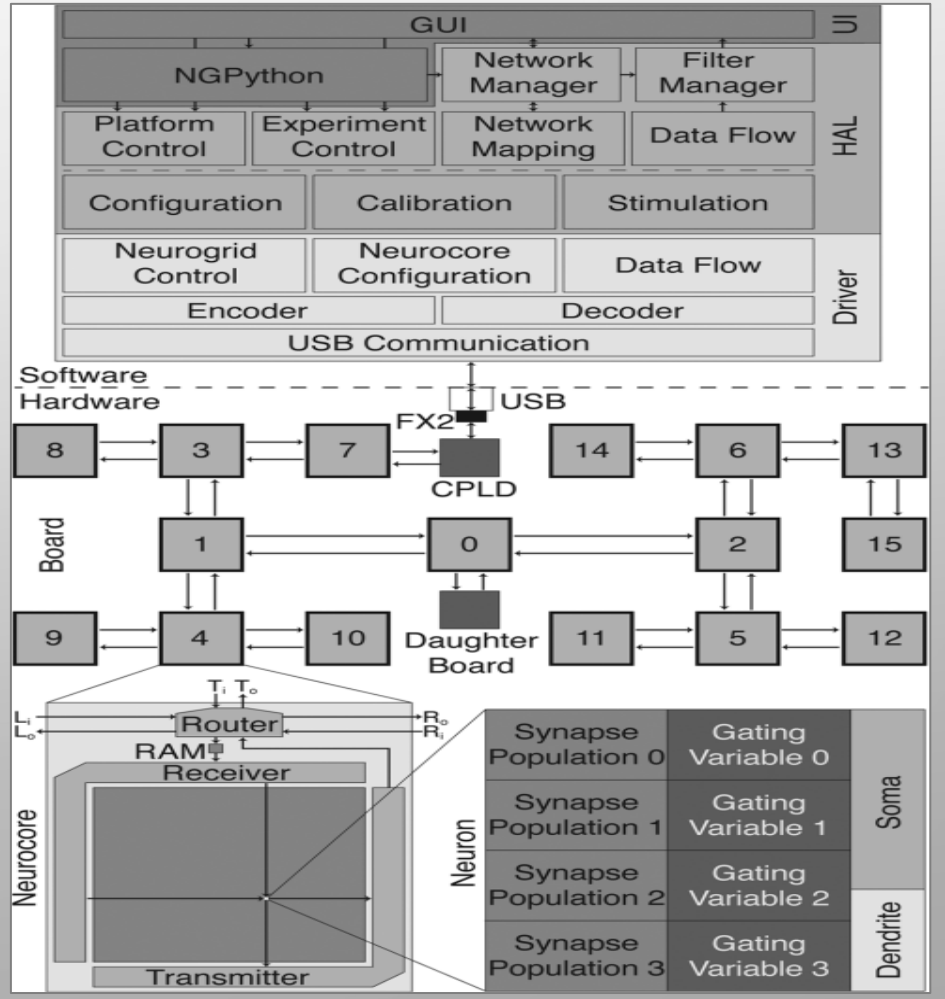
NEUROMORPHIC PROCESSORS (1/5)

- Neurogrid is a mixed-mode multichip system of which neuron array is designed in analog while spike TRX and memory are designed in digital with 0.18 µm CMOS
- Simulates 1 million neurons and 8 billion synapses in real time

Neurogrid. (a) GUI: Enables a user to change his or her model parameters (left), view spike activity in the model's various layers (middle), plot spike rasters from a selected neural layer (right), and enter commands (bottom). (b) Board



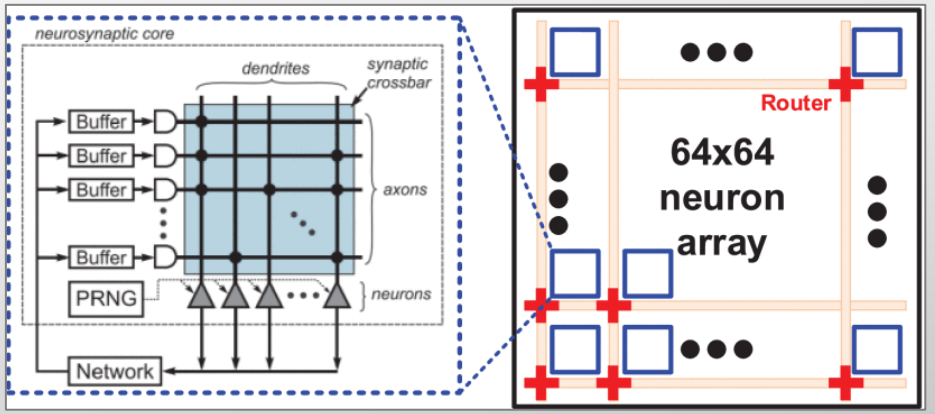
Neurogrid System Architecture



NEUROMORPHIC PROCESSORS (2/5)

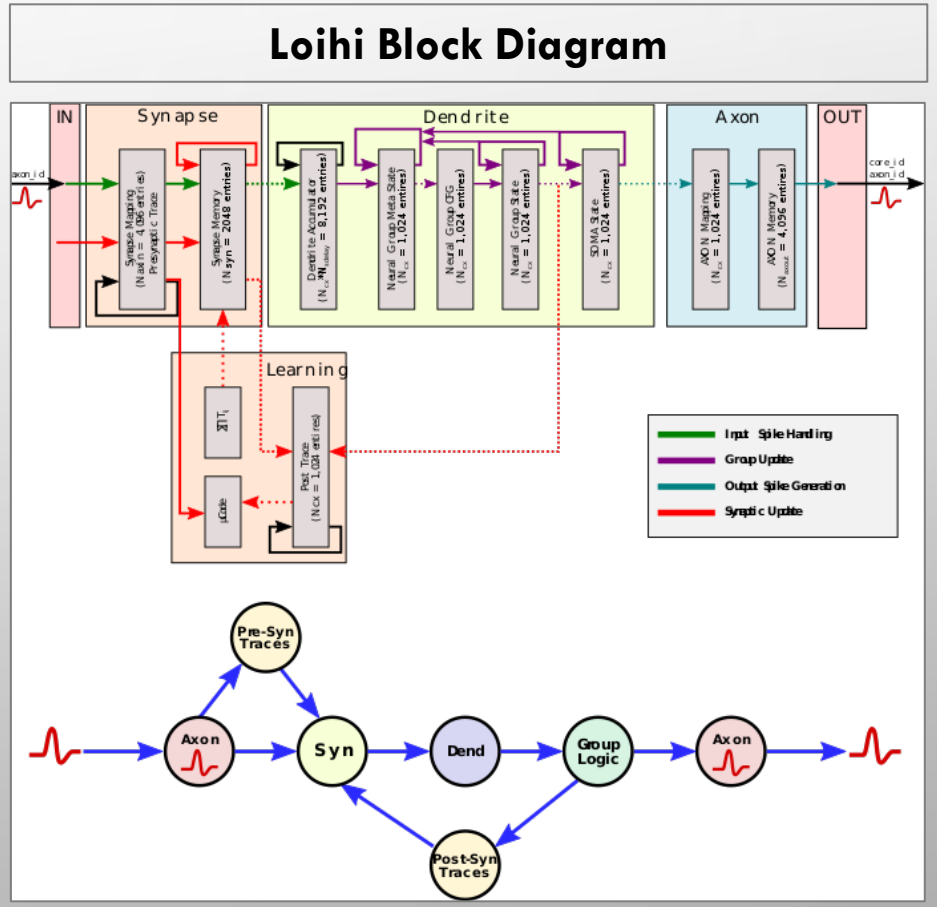
- TrueNorth implemented 4,096 neurosynaptic cores of SNN that are connected via scalable routing network
- Simulates 16 million neurons and 4 billion synapses in real time
- Application: Visual Recognition, Speech Recognition
- Each core consists of 256 digital axons and neurons which are connected by 256 x 256 crossbar for dendrite design
- The large-scale chip (4.3 cm^2) fabricated in 28 nm CMOS was applied to multiobject detection with 63 mW with 400 x 240.30 fps video

TrueNorth Hardware Architecture



NEUROMORPHIC PROCESSORS (3/5)

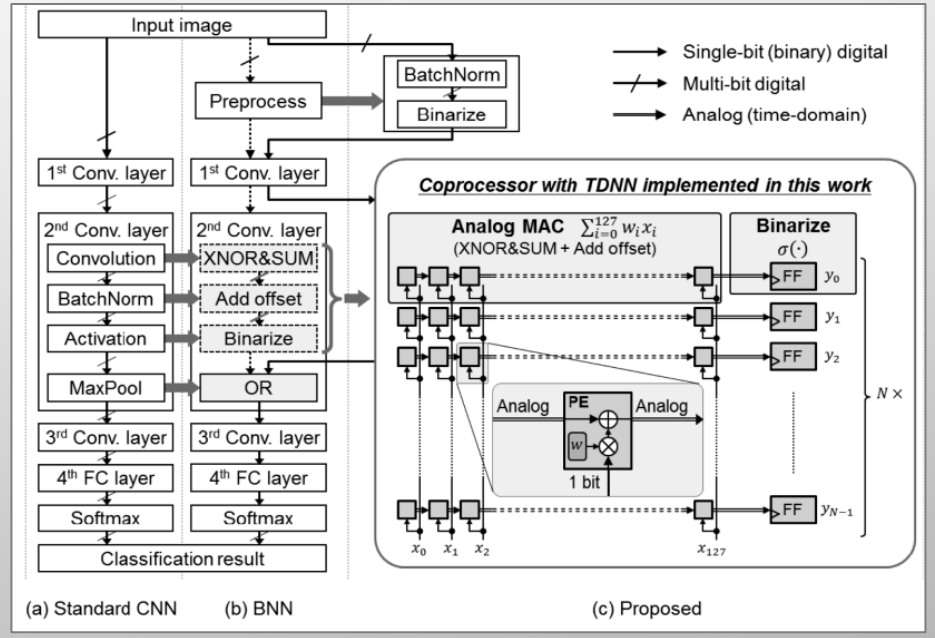
- Loihi, also is fully digital architecture implementing 128 cores of 1,024 SNN units and supports sparsity as its performance is evaluated with LASSO optimization
- The chip has a high level of completion that is also contained three x86 cores for message control for the neuromorphic cores
- The shole SoC is implemented in 14 nm FinFET technology and occupies 60 mm^2
- Although their die area were very large, both TrueNorth and Loihi are energy efficient consuming only ~10mW



NEUROMORPHIC PROCESSORS (4/5)

- Take advantage of analog computing arrays
- TDNN: time domain neural network processor in part of Binarized NN with analog MAC circuits in CMOS technology
- Employ differential signaling to prevent precision degradation from PVT variation
- The chip fabricated in 65 nm CMOS technology achieved high energy efficiency of 12.9 fJ per synaptic operation

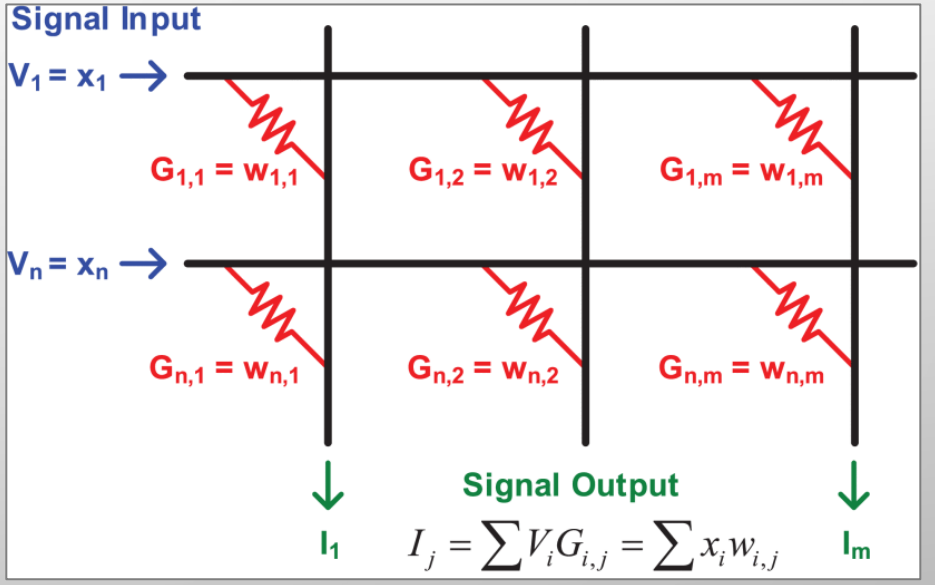
TDNN Hardware Architecture



NEUROMORPHIC PROCESSORS (5/5)

- Some neuromorphic processors make use of nonvolatile memory (NVM) such as ReRAM and MRAM
- The analog accelerators show significant performance improvement than digital processors
- It is important to ensure that implementation details and material properties of the NVM must be aligned with the requirements of NN algorithms
- Have problems of their limited functionality
- Most of the works were capable of only simple classification task while DNN/DL SoCs provide higher-level intelligence with much complex dataset

Conceptual diagram of analog vector MAC array



SECTION VI
DISCUSSION & FUTURE DIRECTION

MIXED MODE NN/NF PROCESSOR (1/2)

- From reviewed mixed-mode processors, they employ current-mode circuits for area and power reduction compared with fully digital implementation
- However, amount of power reduction becomes less significant as the technology node goes down to nm
  - The average power reduction using 0.13 µm technology was 70%, while that of 65 nm technology was only%. This is because the analog circuit cannot be scaled down with the same ratio of technology scaling
- Has the advantage of having smaller but flexible digital controller as well as utilizing natural parallelism of circuit operations
- Analog cores can be reconfigurable with some design techniques to provide versatile functions even though it can be less accurate than digital computations
- Analog circuits can also be programmable assisted by digital controller
  - By having shared datapath to ADC with the digital learning controller, the RNN-FIS processor can handle four different NN/NF topologies

MIXED MODE NN/NF PROCESSOR (2/2)

- However, analog and digital domains must be carefully divided in mixed-mode design to reduce the domain conversion overhead cost in terms of conversion speed, area, and power
- Overhead can be resolved when the processing comes into sensors and computes with analog data before A-D conversion
  - In [103], pupil edge detection and glint corner detection circuits were deployed into CMOS image sensor, each of which circuit configures pixel array to extract edge in analog voltage and compares charge difference, respectively
  - In [104], matrix multiplication was integrated into ADC that multiplies analog input with digital signal to reduce additional energy required for A-D conversion
  - Kim [105] designed an image sensor dedicated to stereo matching. Its reconfigurable pixel array computes sparse image rectification to align two different inputs and census transformation is executed in analog domain with simple switch network and comparators
  - Since those values are computed with analog signal before ADC the digital processor integrated in the sensor excluded additional blocks for rectification and census transformation
  - All the works achieved high energy efficiency by putting computations into the sensors, and showed new design paradigm

DIGITAL NN/NF PROCESSOR (1/1)

- Reviewed digital NN/NF processors as a functional block of the machine vision systems
- They reduced the amount of overall computation with visual attention and improved system throughput with workload prediction
  - It is noticeable that taking NN/NF functions for specific purpose reduces energy efficiency of the whole machine vision SoC
- This implies that processor does not always have to be dedicated for accelerating end-to-end DNN algorithm but applying small NN/NF system to right places with dedicated hardware architecture for that specific purpose brings performance enhancement

DEEP NEURAL NETWORK SOC: DESIGN CHALLENGES (1/4)

POWER CONSUMPTION

- Although the majority of computing engine for general purpose DNN acceleration are traditional CPU and GPU, they consume too much power
- In contrast, ASIC designs reduce power consumption to the order of mW, which makes ASICs suitable for mobile and embedded system
- Table III shows the summary of the DNN SoC

Ref.	Type	Application	Process	Peak Power [mW]	Area [mm^2]	Supply Voltage [V]	Frequency [MHz]
[70]	RNN/CNN	General Purpose	65 nm	279	16.0	0.77 ~ 1.0	50 ~ 200
[72]	RNN/CNN	General Purpose	65 nm	297	16.0	0.63 ~ 1.1	5 ~ 200
[78]	CNN	Always-on Face Recognition	65 nm	211	16.0	0.46 ~ 0.8	5 ~ 100
[80]	CNN-Stereo	Depth Estimation	65 nm	21.3	~ 8	0.7 ~ 1.2	10 ~ 100
[35]	CNN	Image Classification	65 nm	278 (conv. only)	16.0	0.82 ~ 1.17	100 ~ 250
[40]	FCL	Always-on DNN	40 nm	N/A	7.10	0.63 ~ 0.9	1.9 ~ 19.3
[48]	CNN/FCL/RNN	General Purpose	40 nm	2,083	122	0.77 ~ 1.1	75 ~ 330

AlexNet running on ASIC design consumes only 290mW while traditional CPU/GPU consumed 2.7W

DEEP NEURAL NETWORK SOC: DESIGN CHALLENGES (2/4)

MEMORY BANDWITH

- Although they are designed for low power consumption, memory bandwidth becomes crucial problem
- The processors [38], [43], [47], [79] tried to deploy data locality for maximizing on-chip data reuse in order to achieve lower power yet high throughput
  - This approach requires a large memory bandwidth and complex global routing to high parallelism, which is not suitable for low-power consumption
- So [78] utilized distributed memory architecture without global routing and memory to lower routing complexity and resolve memory bandwidth bottleneck
  - Such distributed memory architecture extends to Processing-in-Memory (PIM) architecture
- External Memory Bandwidth is another bottleneck as on-chip data reuse technique advances
  - 3D stacking techniques with DNN accelerators as ported in [48] will bring immense improvement in external bandwidth overhead

DEEP NEURAL NETWORK SOC: DESIGN CHALLENGES (3/4)

DEPENDENCY ON DNN ARCHITECTURE AND ACCURACY

- Unlike CPU and GPU, the DNN SoCs introduced so far are dedicated to specific system and application which induces dependency on DNN architecture and accuracy
- GPU and TPU are preferable than the ASICs introduced in this paper if programmability of DNN algorithm is versatile, or the target application requires precise computation for high accuracy such as autonomous driving

TECHNOLOGY SCALING

- Thanks to technology scaling, the future researchers on hardware accelerators for high performance DNN applications will be digital to maintain high throughput and precise computation
- It will bring integration of higher-level intelligence with complicate DNN architectures with energy-efficient digital implementation

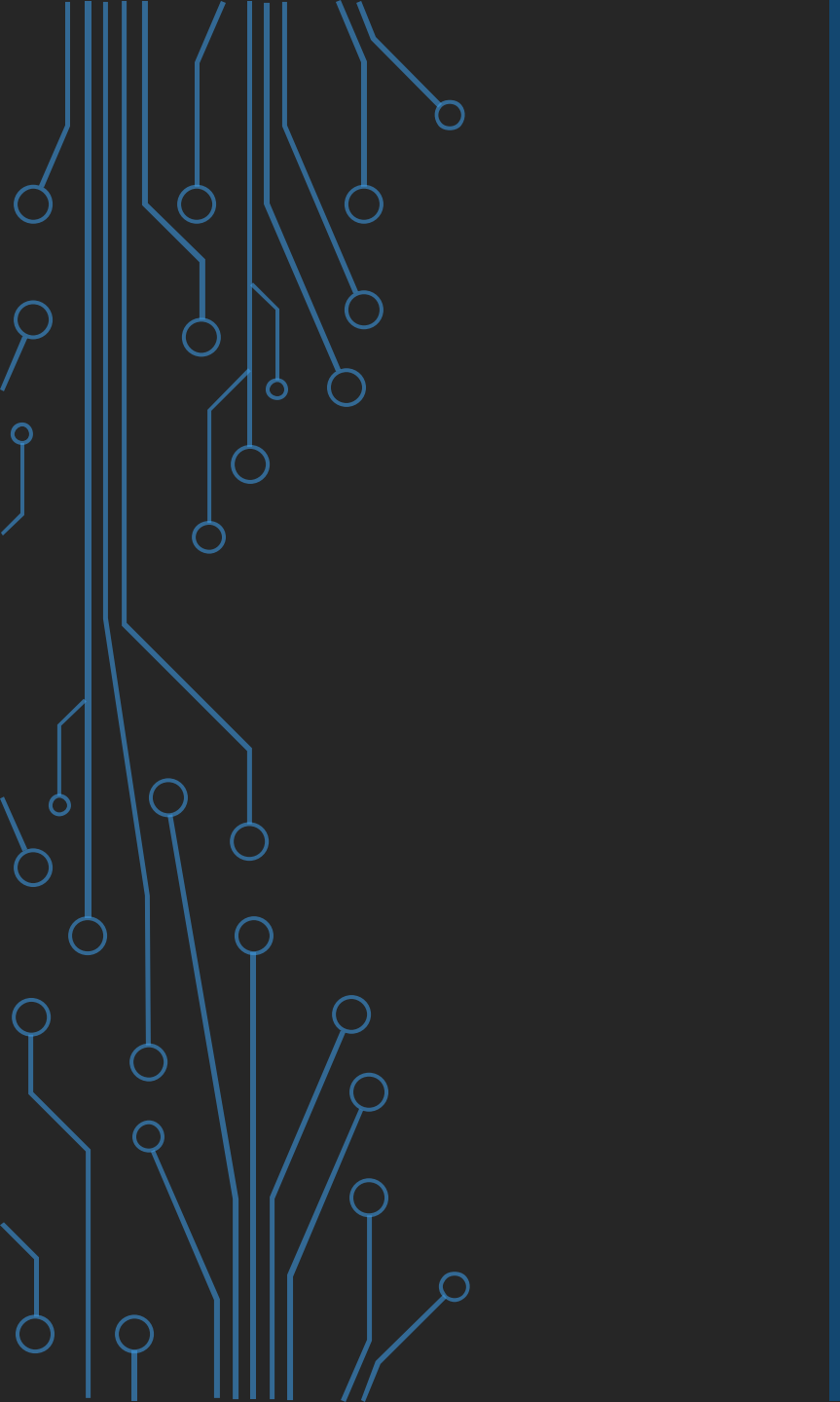
DEEP NEURAL NETWORK SOC: DESIGN CHALLENGES (4/4)

ULTRA-LOW POWER APPLICATIONS

- Neuromorphic processors will be investigated for ultra-low power applications with simple intelligence by utilizing energy-efficient analog computations

RELIANCE ON DEVELOPMENT OF NEW TECHNOLOGIES

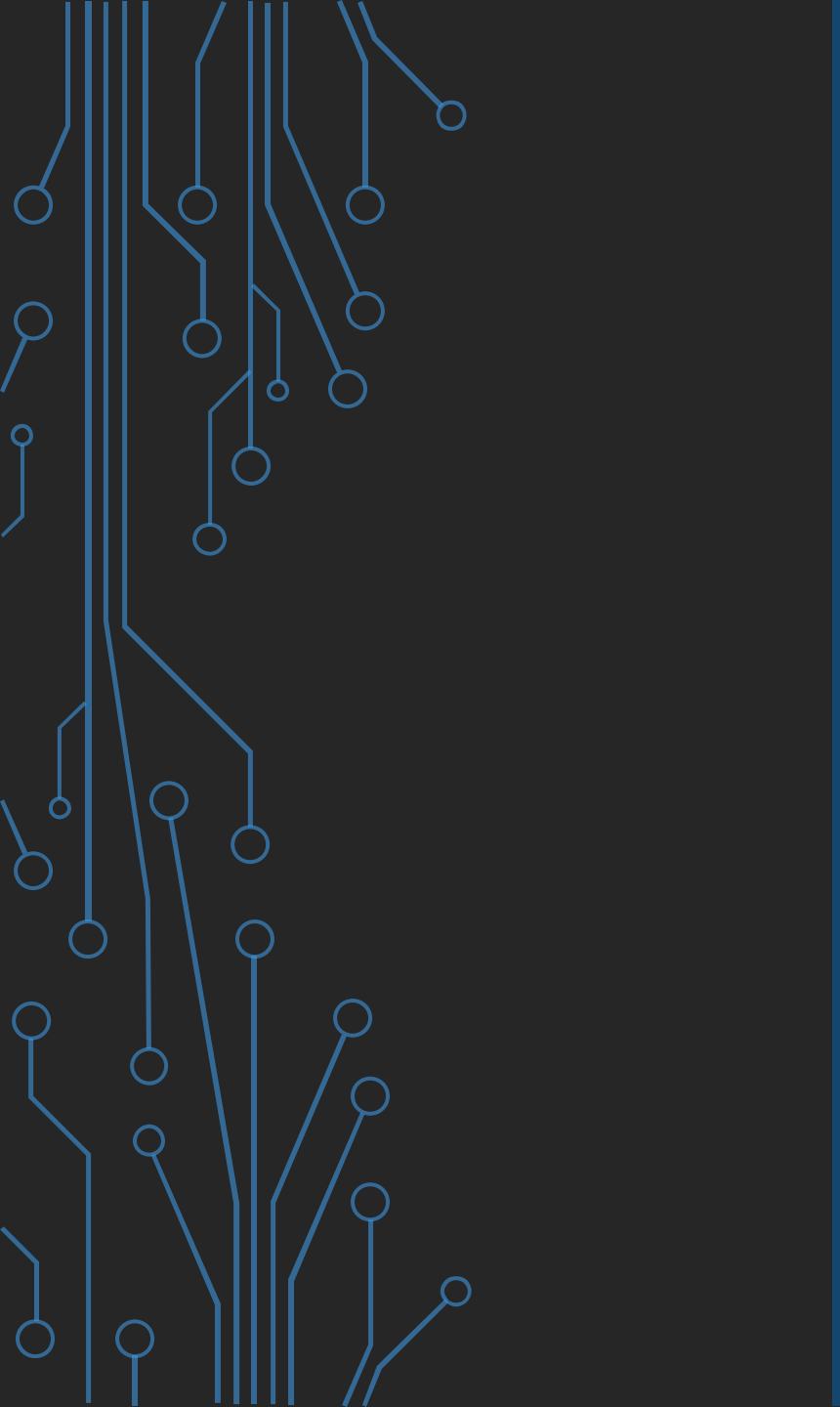
- Many neuromorphic processors highly reply on the development of new technologies such as ReRAM or MRAM
- These technologies are still being researched but its time-to-market until mass production is unpredictable
- In the meantime, it is believed that mixed-mode implementation and PIM architectures of neural network processor that uses current CMOS technology is promising for ultra-low-power DNN applications until the neuromorphic devices become in mass production



SECTION VII
CONCLUSION

CONCLUSION

- This paper provides a review of dedicated processors for neural network, neuro-fuzzy system, deep learning and deep neural network processors in various design approaches of analog, digital, and mixed-mode implementation over decades
- Provide both mixed-mode and fully-digital implementations of neural network / neuro-fuzzy processors that are deployed as a functional building block of machine vision systems
- Provide the comparisons upon the different design approaches
- The most recent designs of deep learning processors were introduced
- Neuromorphic chips are highly dependent on the development of new technologies, but no one can assure how long it will take for the technologies to be in mass production
- Before then, it is believed that mixed-mode implementation of neural network processor is promising for ultra-low-power applications while digital implementation will fit for high-performance applications



THANK YOU