Distributed, Intelligent, and Scalable COmputing DISC Lab

Hardware-Software Co-Design for On-Chip Learning in AI Systems

M. L. Varshika, A. K. Mishra, N. Kandasamy and A. Das

Associate Professor, Drexel University, Philadelphia

AMBITION CAN'T WAIT

Executive Summary

- Neuromorphic hardware can reduce the energy consumption of machine learning
  - An attractive solution for embedded/edge devices where power is limited
- On-chip learning or online learning is a step-forward in the development of neuromorphic hardware
  - Enables a system to learn from a constant stream of data
- ECHELON: A tile-based neuromorphic hardware with on-chip learning capabilities
  - Each tile consists of
    - Neural Processing Units (NPUs)
    - On-chip Learning Units (OLUs)
    - Special Function Units (SFUs)
  - Tiles interconnected using Network-on-Chip (NoC)
- Co-Design: Develop hardware and software architecture for a given machine learning model
- Evaluation: FPGA based implementation and co-design evaluation using 8 machine learning workloads

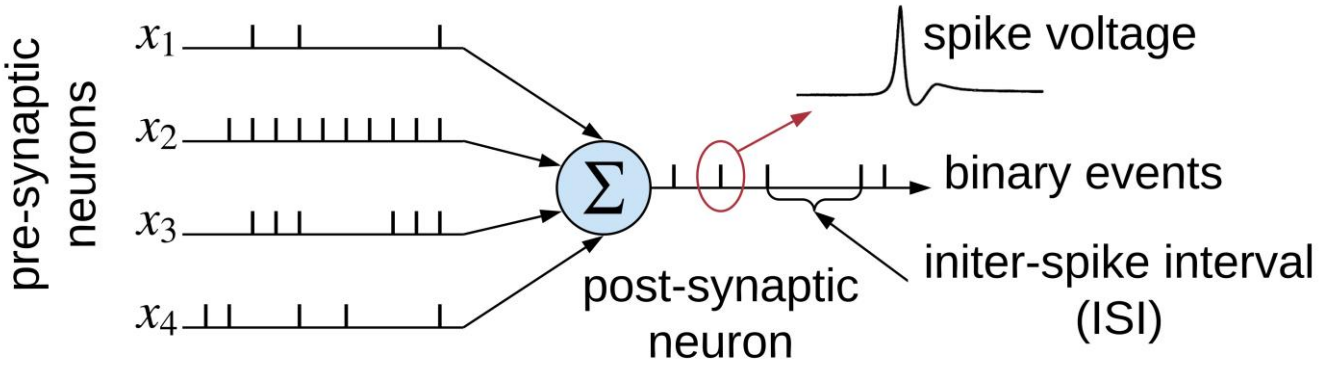
Outline

- Introduction
- ECHELON
  - Tile Architecture
  - Interconnect Architecture
- System Software
- Co-Design
- Evaluation
- Conclusion

Introduction (I): Neuromorphic Computing

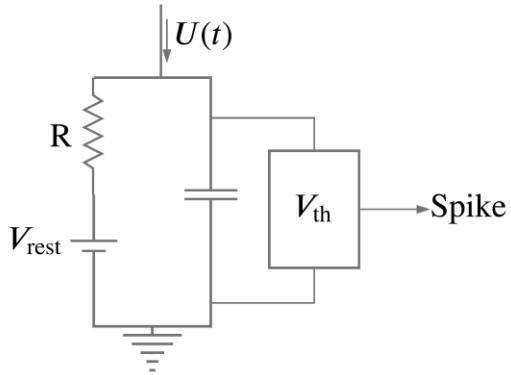
- Powerful computation capability compared to ANN
  - Spatiotemporal information encoding

- Dynamics of an SNN is complicated compared to ANNs

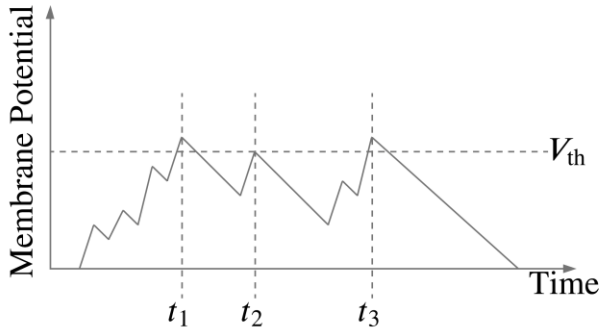


Introduction (I): Neuromorphic Computing

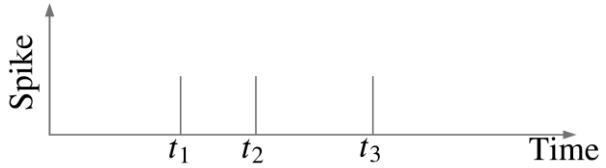
- Spiking Neurons



Leaky Integrate-and-Fire (LIF) neuron



Membrane Potential of the neuron



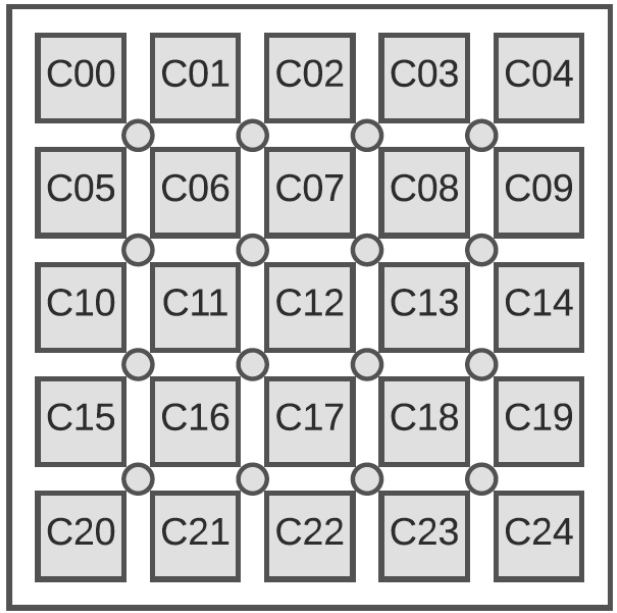
Spike output of the neuron

Introduction (I): Neuromorphic Computing

- SNN vs. ANN for Convolutional Neural Networks (CNNs)
  - Communication between layer facilitated using spikes (compared to tensors in ANNs)
  - Computational units are (leaky) integrate and fire (LIF) neurons (compared to sigmoid activations in ANNs)
- Many of recent works focuses on SNN inference
  - Train SNN
  - Deploy trained synaptic weights to a neuromorphic hardware such as DYNAPs, Tianji, TrueNorth, etc.

Introduction (I): Neuromorphic Computing

- Neuromorphic platform
  - Many-core hardware
  - Cores interconnected using a shared interconnect
- Each neuromorphic core
  - Neuron circuitry
  - Synapse circuitry
  - Interface to communicate spikes on the shared interconnect



Introduction (II): On-chip Learning

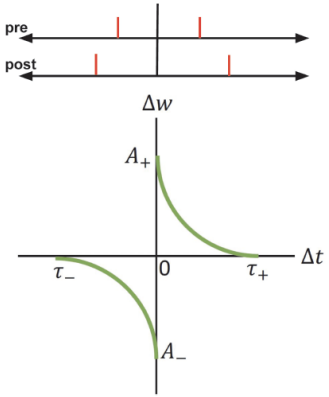
- STDP: On-chip Learning
  - Spike timing dependent plasticity (STDP)
  - A form of Hebbian Learning
- STDP Operation
  - Long-term Potentiation (LTP)
    - Increase of synaptic weight when a pre-synaptic spike arrives before a post-synaptic spike
  - Long-term Depression (LTD)
    - Decrease of synaptic weight when a pre-synaptic spike arrives after a post-synaptic spike
  - Use a spike window to evaluate LTP or LTD

Introduction (II): On-chip Learning

- STDP Dynamics: Exponential STDP Model
  - Calculate time delta between pre and post spike

$$\Delta t = t_{post} - t_{pre}$$

  - Evaluate weight increment and decrement



$$\Delta W = \begin{cases} A_+ \exp(-\Delta t/\tau_+) & \text{if } \Delta t > 0 \\ A_- \exp(-\Delta t/\tau_-) & \text{if } \Delta t \leq 0 \end{cases}$$

  - Apply the weight change

$$W_{new} = W_{old} + \beta + \alpha(\Delta w_+ + \Delta w_-)$$

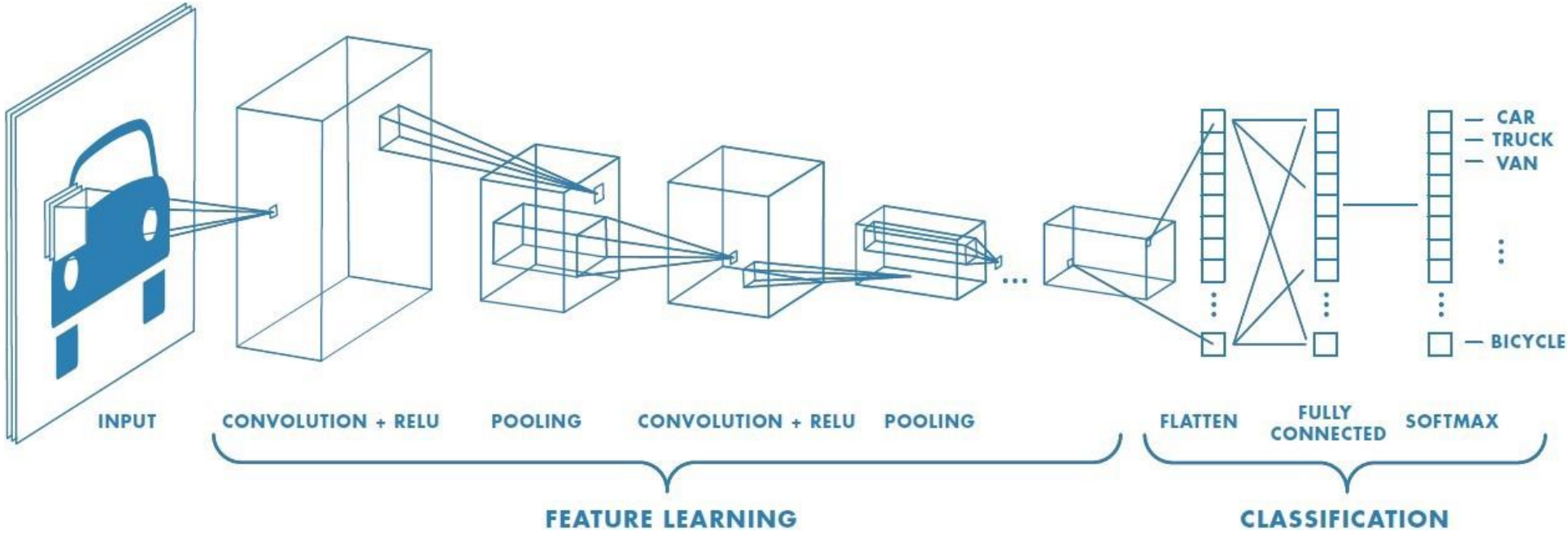
Introduction (II): On-chip Learning

- On-chip Learning hardware
  - ROLLS
    - Short-term plasticity (STP)
    - Long-term plasticity (LTP)
  - Loihi
    - STDP

  - ROLLS cannot support large models while Loihi cannot support arbitrary model architectures
  - Our Objective: Design a many-core neuromorphic hardware supporting STDP-enabled convolutional neural networks (CNNs)

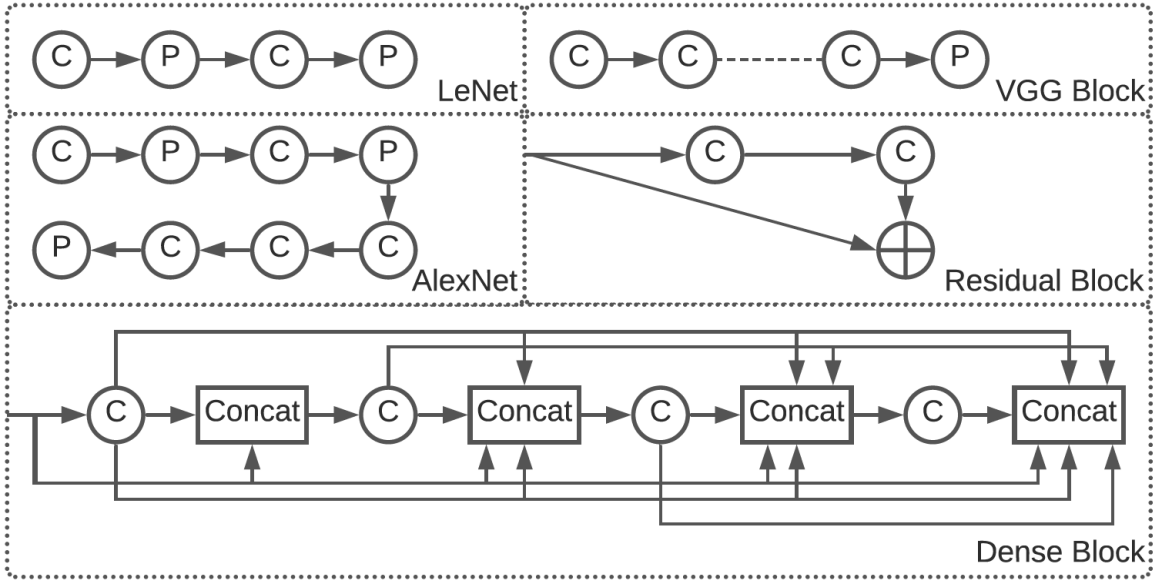
Introduction (III): STDP-Enabled CNNs

- Convolutional Neural Network (CNN)



Introduction (III): STDP-Enabled CNNs

- Irregular CNN Architectures



Introduction (III): STDP-Enabled CNNs

- STDP-Enabled CNNs
  - Learning in the convolution layers
  - When a new image is presented, neurons of a convolutional layer compete and one which fire earlier triggers STDP and learn the input pattern

Kheradpisheh et al., *STDP-based spiking deep convolutional neural networks for object recognition*, Neural Networks 2017

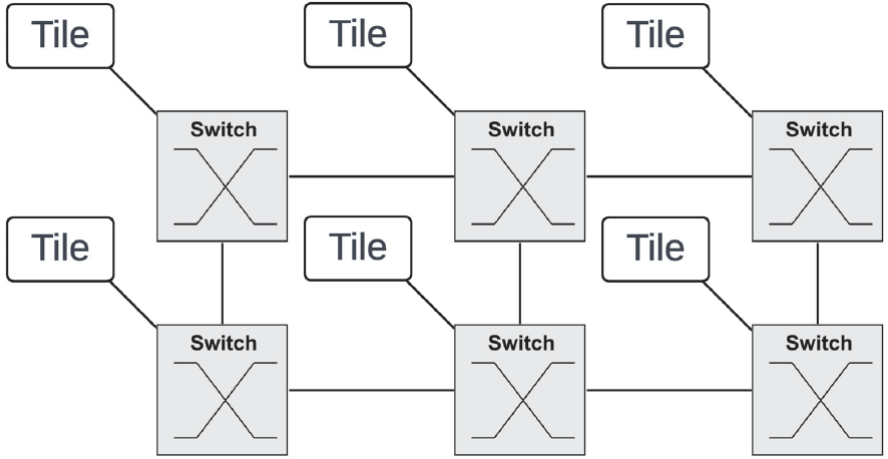
Outline

- Introduction
- ECHELON
  - Tile Architecture
  - Interconnect Architecture
- System Software
- Co-Design
- Evaluation
- Conclusion

ECHELON: Tiled Hardware Supporting STDP-Enabled CNN

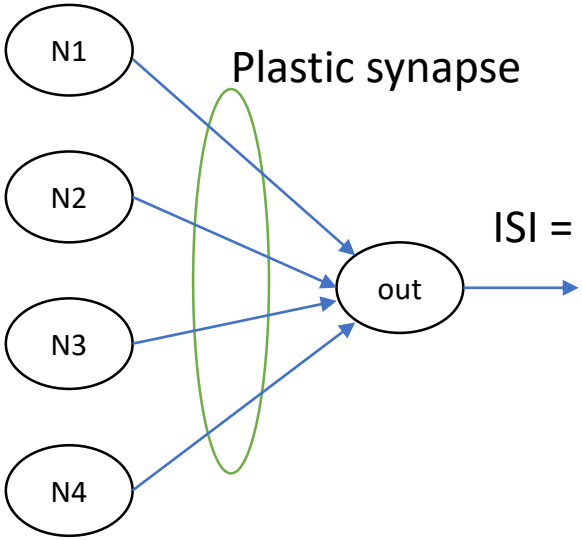
- ECHELON
  - Tiled neuromorphic hardware
  - Tiles interconnected using Network-on-Chip (NoC)

- ECHELON Tile
  - Neural Processing Unit (NPU)
    - Convolution and dense layers
  - Special Function Unit (SFU)
    - Pooling, concatenation, batch normalization
  - On-chip Learning Unit (OLU)
    - STDP learning

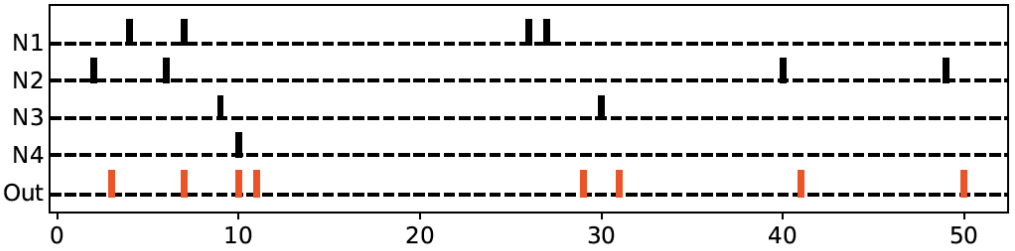


ECHELON Tile Architecture

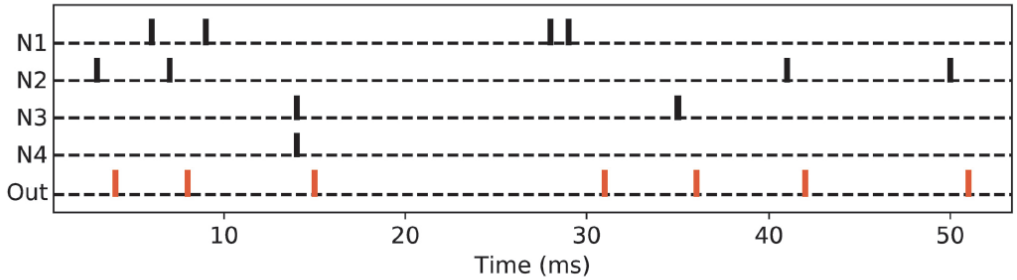
- STDP Learning is sensitive to spike timing



ISI = Inter-Spike Interval



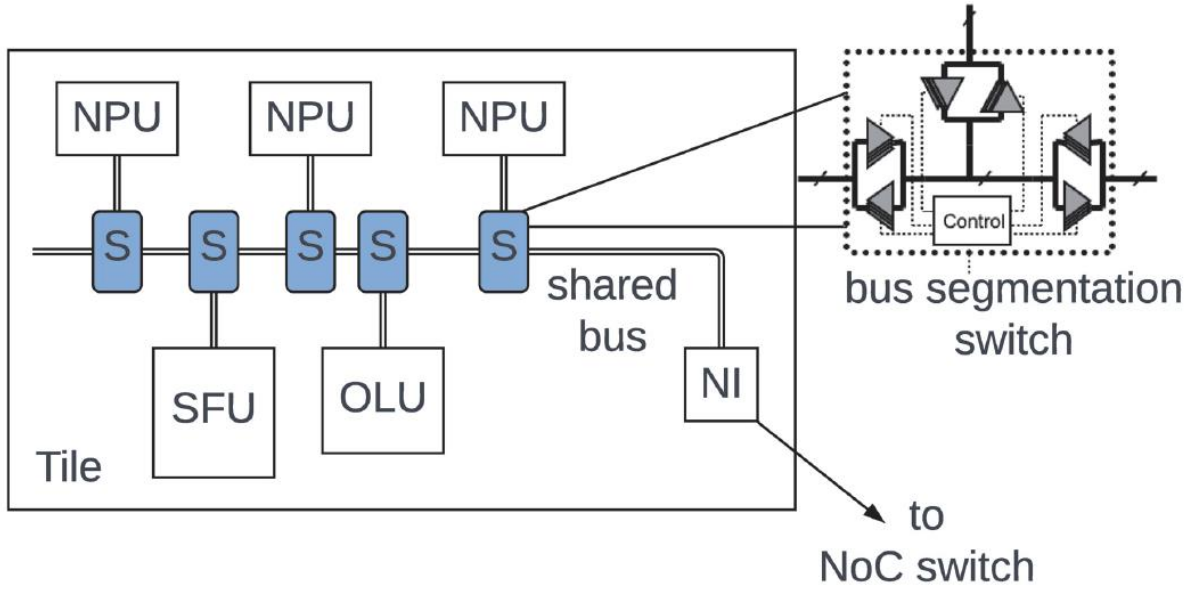
(a) No spike delay (ISI = 6.71).



(b) With spike delay (ISI = 7.83).

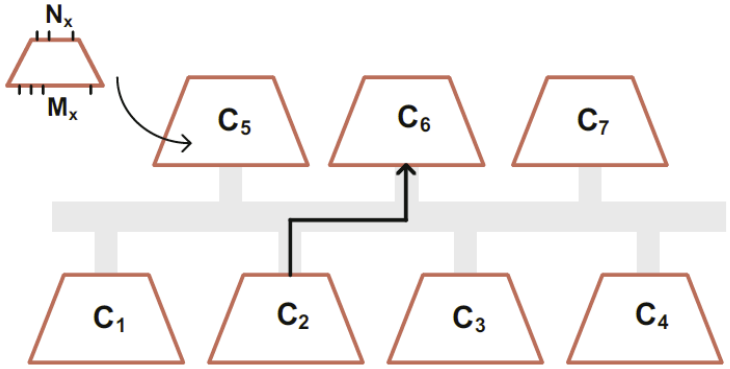
ECHELON Tile Architecture

- Use segmented bus for interconnecting tile components

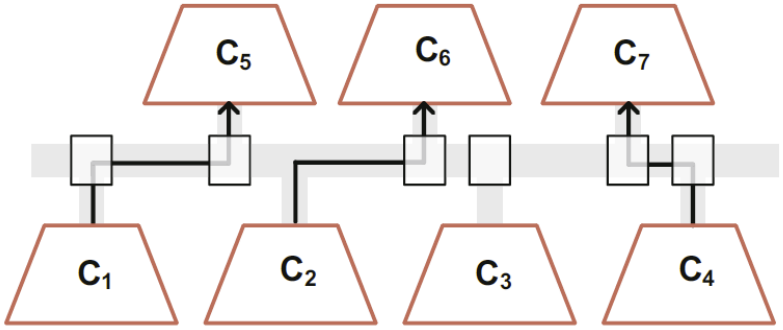


Intra-Tile Interconnect

- **Segmented Bus**
  - A bus lane is partitioned to allow concurrent connections



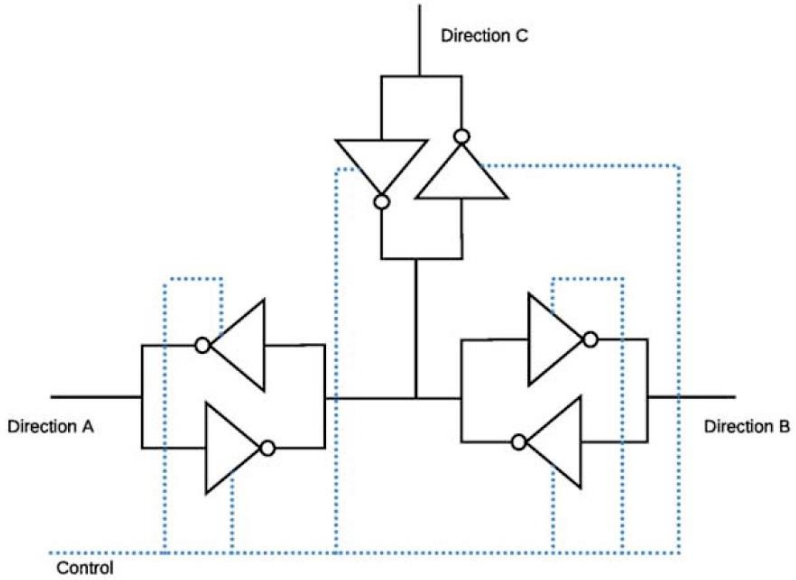
Conventional Bus



Segmented Bus

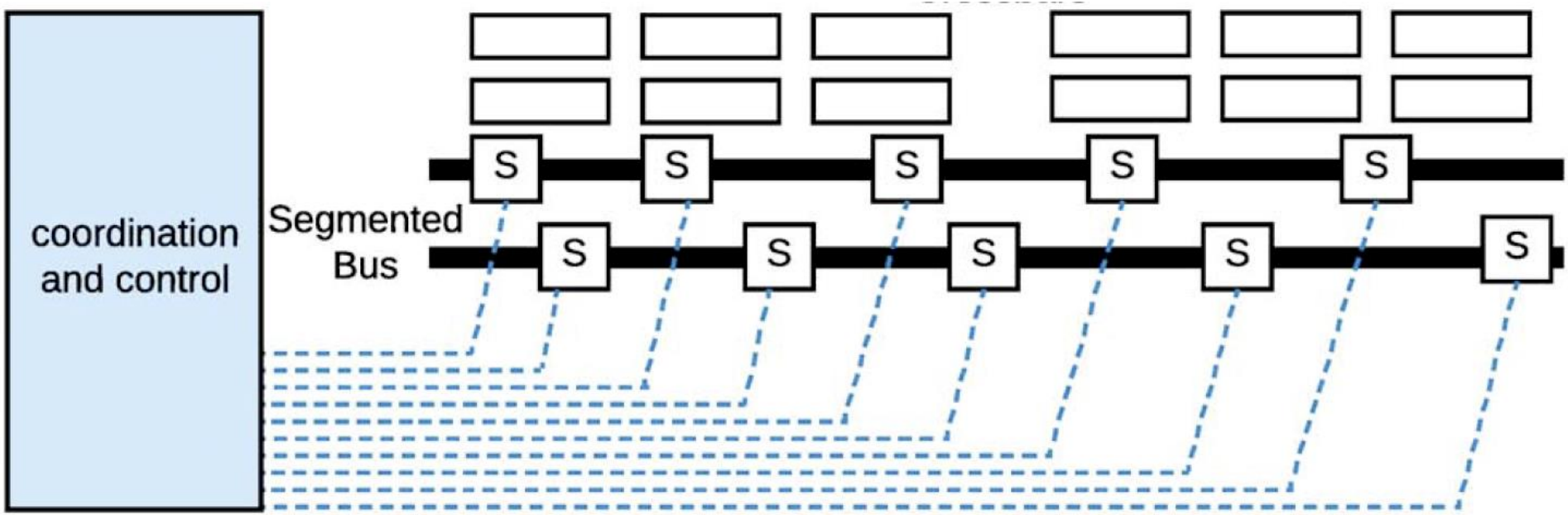
Intra-Tile Interconnect

- **Segmentation Switches**
  - Three-way switch



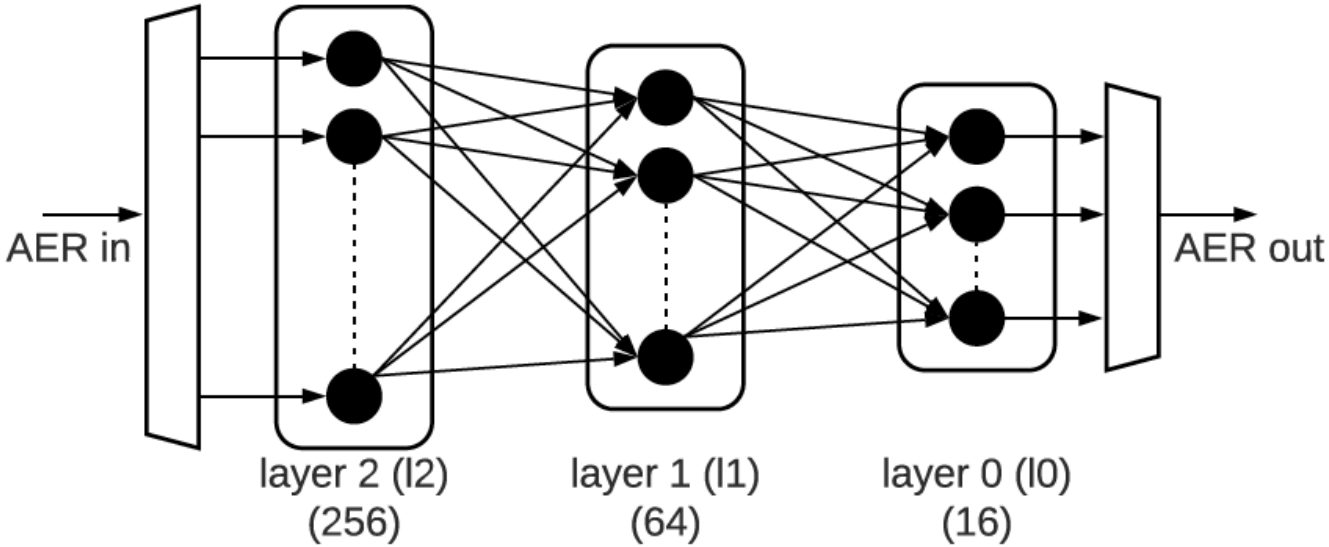
Intra-Tile Interconnect

- **Parallel Segmented Bus**



ECHELON Tile Component: NPU

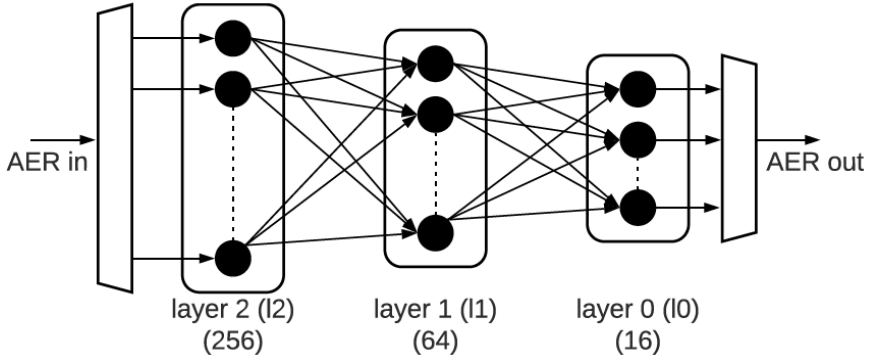
- μBrains design



*Varshika et al. "*Design of Many-Core Big Little μBrains for Energy-Efficient Embedded Neuromorphic Computing*", DATE 2022*

ECHELON Tile Component: NPU

- μBrains design
  - 3 layers of neurons with full connectivity between layers and recurrent connections within layer
  - Layer 1: 256 neurons
  - Layer 2: 64 neurons
  - Layer 3: 16 neurons



- Design specification
  - 336 neurons
  - 38K synaptic connections (feedforward and recurrent)
  - 40nm CMOS technology (2.82mm2 including pads)

ECHELON Tile Component: NPU

- **Handwritten Digit Recognition**

J. Stujit et al. "μBrain: An event-driven and fully synthesizable architecture for spiking neural networks", Frontiers in Neuroscience 2021

	μBrain	Frenkel et al. (2018)	Park et al. (2019)	Cho et al. (2019)	Chen et al. (2018)	Moradi et al. (2017)	Davies et al. (2018)
MNIST accuracy (%)	91.7 (16 × 16)	91.4 (16 × 16)	97.83	91.6 (16 × 16)	97.9	–	96.4
Neuron/Synapses used for MNIST	74/17k	10/2.5k	410/199k	2048/149k	1546/666k	–	10/7840
VDD (V)	1.1	0.55–1.1	0.8	0.7	0.525–0.9	1.3–1.8	0.5–1.25
Energy/Prediction (nJ)	308	15 @ 75 MHz, 54 @ 1.3 MHz	236.5	–	1700	–	85,52*
Technology (nm)	40	28 FDSOI	65	40	10 FinFET	180	14 FinFET
Physical neurons cores/total neurons	336/336	1/256	410/410	2048/2048	4096/4096	1024/1024	128/131072
Power	73 μW	35–447 μW	23.6 mW	46.6 mW (2.3 uW * 4096 neurons)	94 mW	400 μW @ 10 Hz average firing rate	110 mW
Area (mm^2)	2.68 (1.42 core only)	0.086**	10.08	2.56	1.7	43.79	60
Synaptic resolution # bits	4	4	>10	2/3	7	2 (analog)	1–9
Clock frequency	Event-driven	75 MHz	20 MHz	Global Async. Locally sync 110 MHz (neurons)	105 MHz	Event-driven	Event-driven
Fully synthesizable	Yes	Yes	Yes	Yes	Yes	No (Analog Mixed Signal design)	Yes
Supported algorithm	SNN feed-forward, recurrent	SNN online learning, feed-forward	SNN on-line learning	SNN feed-forward, recurrent	SNN/BNN online-learning, feed forward, recurrent	SNN feed-forward, recurrent	SNN, online-learning, feed-forward, recurrent

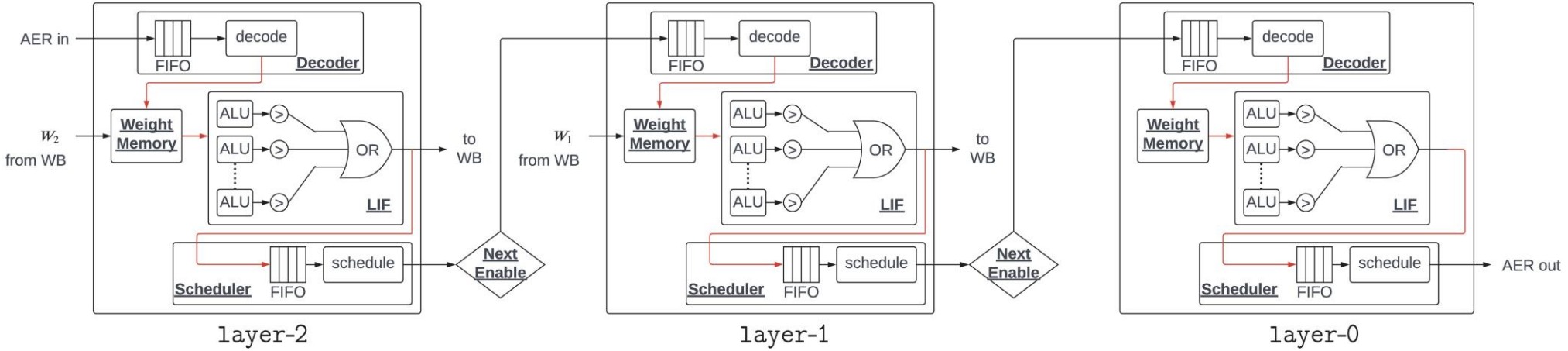
ECHELON Tile Component: NPU

- μBrain Scaling

μBrain Configuration	Normalized	
	Static Power	Area
256-64-16	1X	1X
1024-256-16	9.9X	7.8x
4096-1024-16	280.2X	222.1X
16384-4096-16	4943.0X	3920.1X

ECHELON Tile Component: NPU

- Digital design

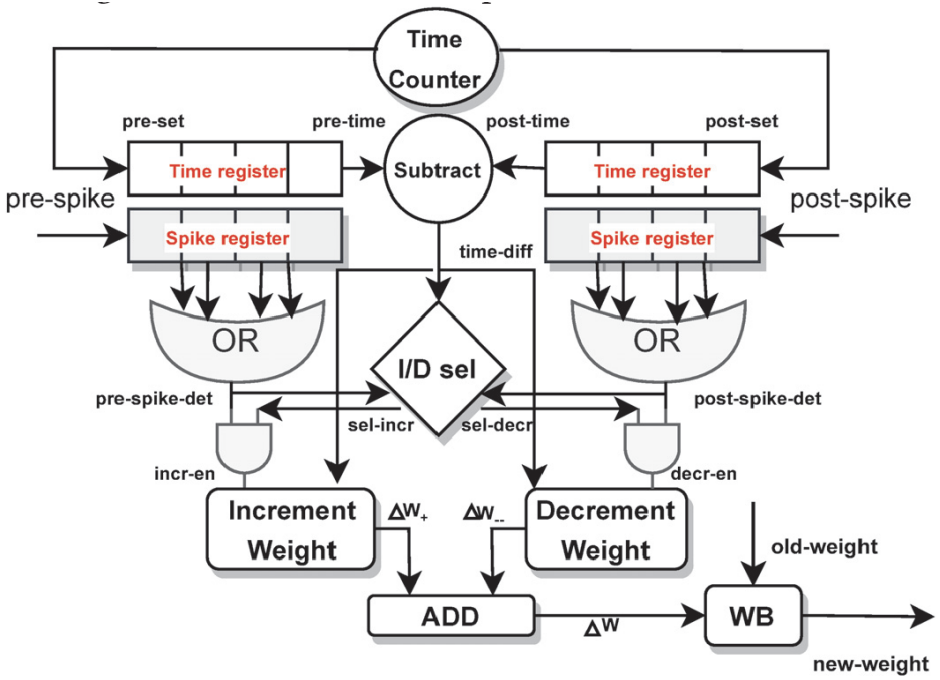


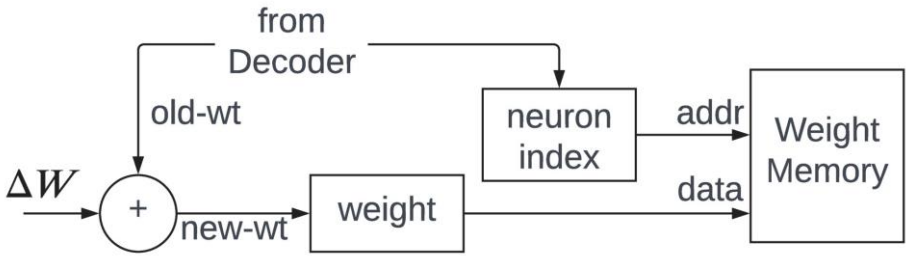
ECHELON Tile Component: SFU

- Carefully configuring the layers and synaptic connections, μBrain can be used for implementing pooling, concatenation, and other custom operations

ECHELON Tile Component: OLU

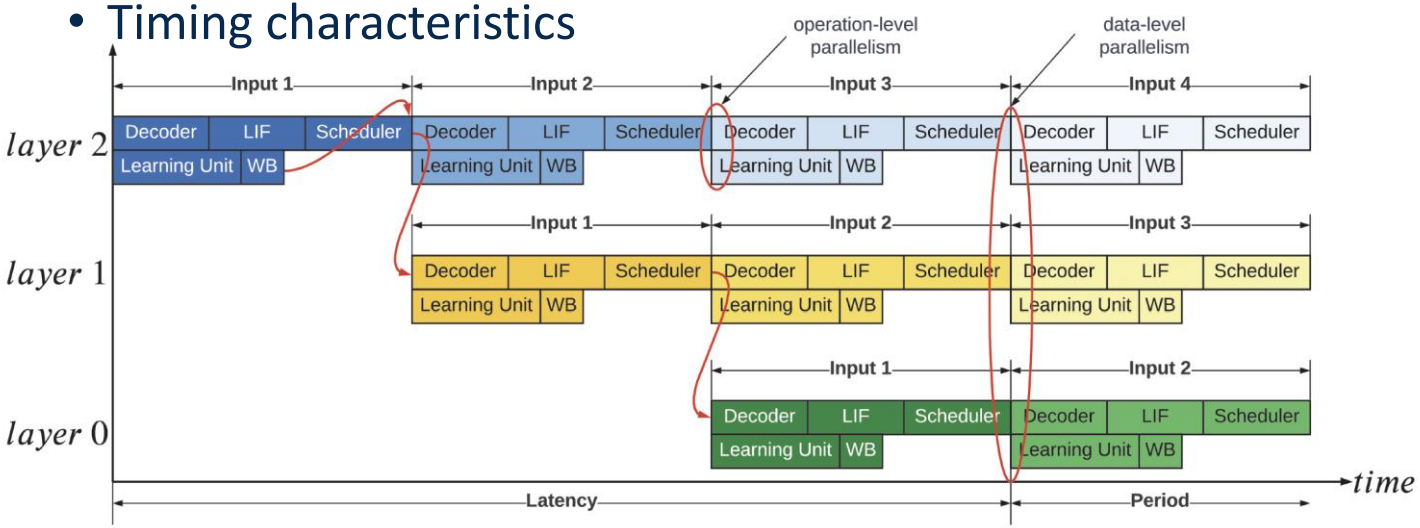
- STDP Unit and interfacing with weight memory





ECHELON Tile Component: OLU

- Timing characteristics



	Decoder	LIF	Scheduler	Learning Unit	WB
delay	3	5	3	5	2

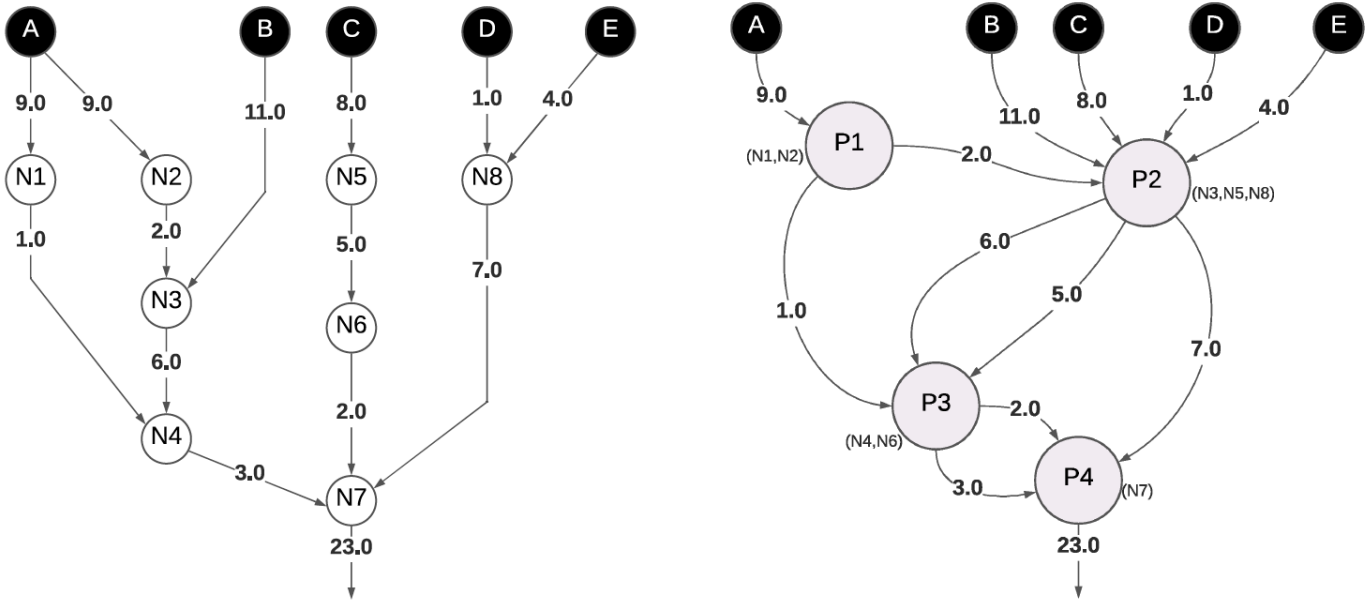
latency	33 clock cycles
long-term execution period	11 clock cycles

Outline

- Introduction
- ECHELON
  - Tile Architecture
  - Interconnect Architecture
- System Software
- Co-Design
- Evaluation
- Conclusion

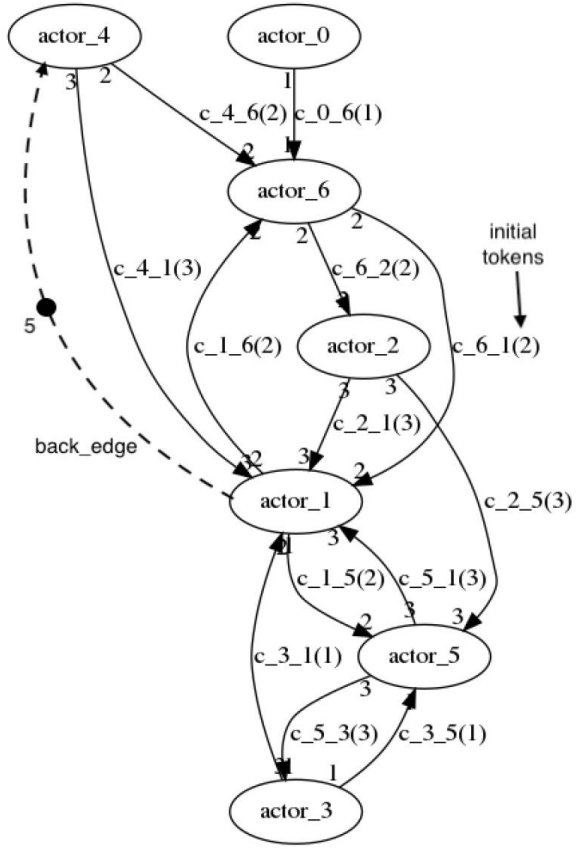
ECHELON System Software

- SNN partitioning to sub-networks



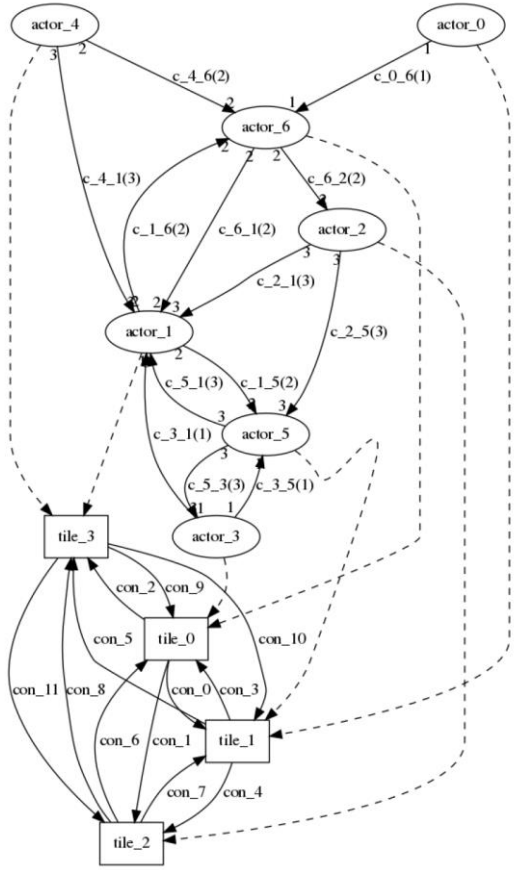
ECHELON System Software

- Clustered SNN graph
  - Nodes (called actors) are clusters
  - A single cluster maps on to a crossbar of the hardware in its entirety
  - Edges represent inter-cluster communications
  - Number (called token) on an edge represent spike rate between clusters
  - Initial tokens represent spikes from previous iterations of the application



ECHELON System Software

- Tile-allocation: A greedy strategy to allocate actors to tiles
  - Load-balancing to distribute actors evenly on tiles
- Actor-ordering: Time-division multiple access (TDMA) scheme to allocate time slices to actors mapped to the same tile
  - Apply Max-Plus Algebra formulation on resource-aware SDFG
- Actor execution: Self-timed execution to execute actors
  - Exact firing times of actors from design-time analysis are discarded retaining only the order
    - Static-order schedule
  - At run-time, ready actors are fired using the static-order schedule

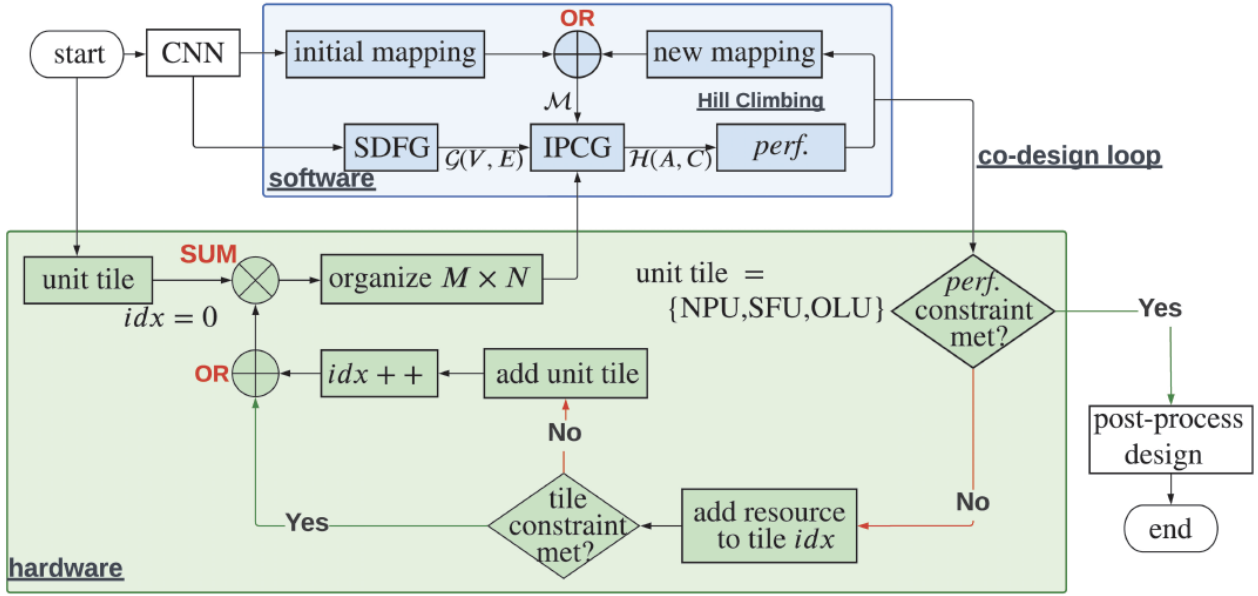


Outline

- Introduction
- ECHELON
  - Tile Architecture
  - Interconnect Architecture
- System Software
- Co-Design
- Evaluation
- Conclusion

Co-Design Using ECHELON

- Start with initial configuration
- Add tile components
  - Explore cost-tradeoff
- Add more tiles



Outline

- Introduction
- ECHELON
  - Tile Architecture
  - Interconnect Architecture
- System Software
- Co-Design
- Evaluation
- Conclusion

Evaluation: Area and Power

- Xilinx Virtex-7 FPGA
- 8-bit weight precision
- 10MHz frequency

Area

	128 × 128				256 × 64 × 16			
	neurons	synapses	learn. units	LUTs	neurons	synapses	learn. units	LUTs
NPU/SFU	256	16,384	–	17,750	336	17,408	–	23987
OLU	–	–	128	14,075	–	–	320	18,970

Power

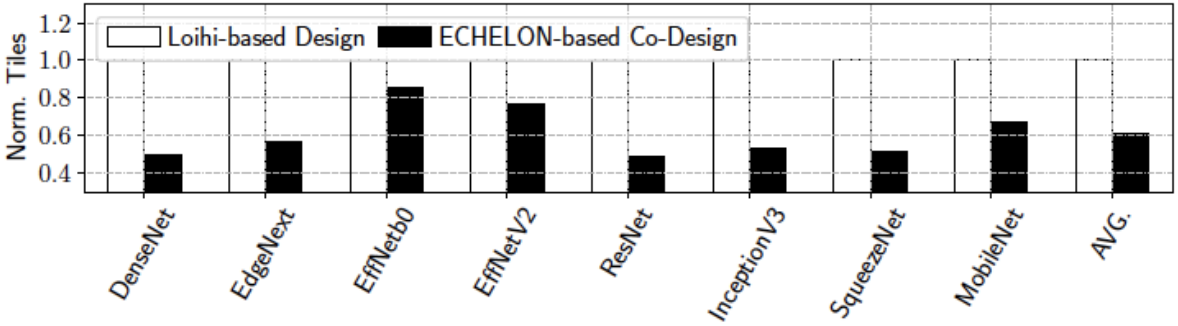
	2 × 1		4 × 2 × 2	
	Static	Total	Static	Total
NPU/SFU	0.242W	0.253W	0.242W	0.252W
NPU/SFU + OLU	0.242W	0.253W	0.242W	0.351W

Evaluation: Co-Design

- Evaluated Models

Models	Params./Acc.	Models	Params./Acc.	Models	Params./Acc.	Models	Params./Acc.
DenseNet	6.9M, 75.0%	EdgeNext	1.3M, 79.4%	EffNetb0	3.9M, 77.1%	EffNetV2	20.1M, 78.7%
ResNet	23.5M, 76.0%	Inception	21.8M, 77.9%	SqueezeNet	0.7M, 57.5%	MobileNet	3.2M, 70.4%

- Co-design Results



Outline

- Introduction
- ECHELON
  - Tile Architecture
  - Interconnect Architecture
- System Software
- Co-Design
- Evaluation
- Conclusion

Conclusions

- Neuromorphic hardware can reduce the energy consumption of machine learning
  - An attractive solution for embedded/edge devices where power is limited
- On-chip learning or online learning is a step-forward in the development of neuromorphic hardware
  - Enables a system to learn from a constant stream of data
- ECHELON: A tile-based neuromorphic hardware with on-chip learning capabilities
  - Each tile consists of
    - Neural Processing Units (NPUs)
    - On-chip Learning Units (OLUs)
    - Special Function Units (SFUs)
  - Tiles interconnected using Network-on-Chip (NoC)
- Co-Design: Develop hardware and software architecture for a given machine learning model
- Evaluation: FPGA based implementation and co-design evaluation using 8 machine learning workloads

Distributed, Intelligent, and Scalable COmputing DISC Lab

Hardware-Software Co-Design for On-Chip Learning in AI Systems

M. L. Varshika, A. K. Mishra, N. Kandasamy and A. Das

Associate Professor, Drexel University, Philadelphia

Web: www.anupkdas.com

Email: anup.das@Drexel.edu

AMBITION CAN'T WAIT