R·I·T

Energy Efficient and Scalable Neuromemristive Computing Substrates

Dhireesha Kudithipudi†, Cory Merkel‡, James Mnatzaganian†, Nicholas Soures†, Qutaiba Saleh†

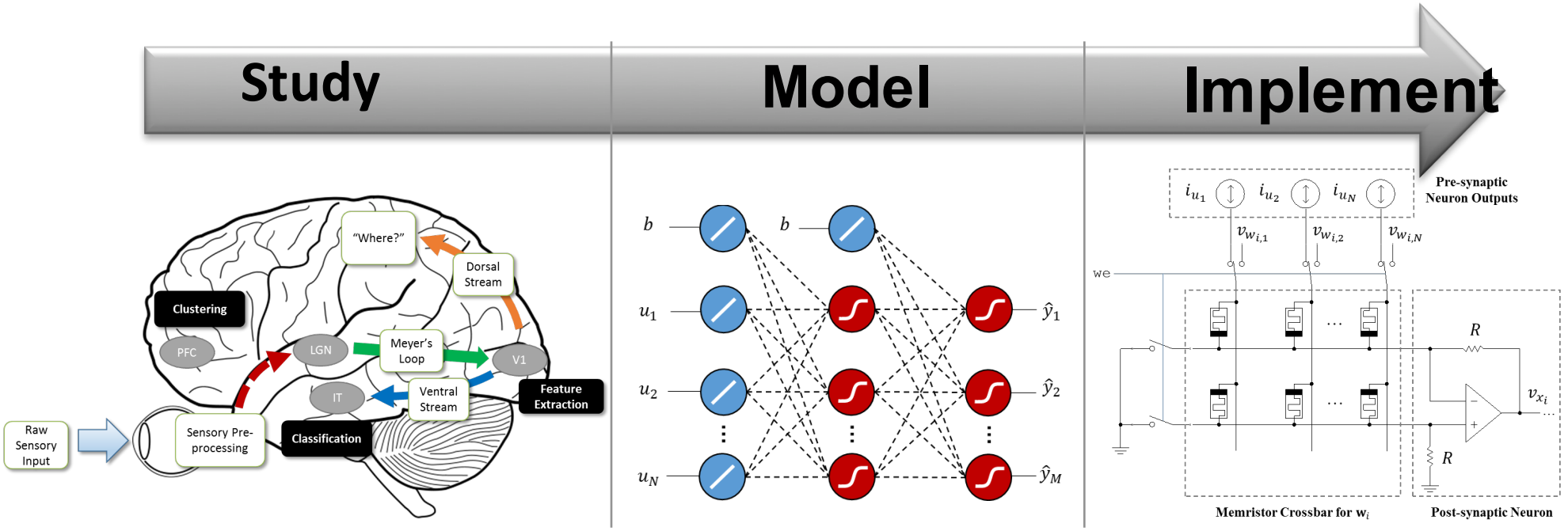
†NanoComputing Research Lab

Rochester Institute of Technology

‡Information Directorate

Air Force Research Laboratory

Neuromemristive Systems

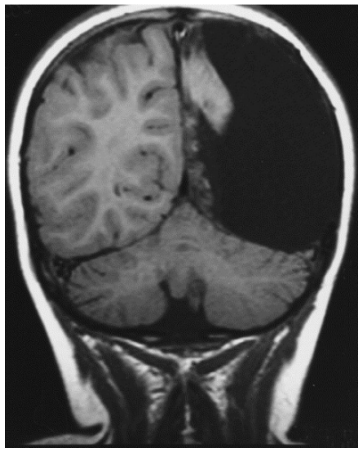


- **Brain-inspired adaptive computing platforms based on nanoscale resistive memory (memristors)**
- Memristor characteristics facilitate efficient computation and learning
- Improve the efficiency (over conventional computers) of *natural* processing tasks

Why the Brain?

- Incredibly versatile
  - Can learn anything!
- Energy efficient
  - ~10^{16} ops/sec @ a few watts!
- Robust/Resilient
  - Functions with noise!
  - Unreliable and damaged components!

Why *not* the brain?



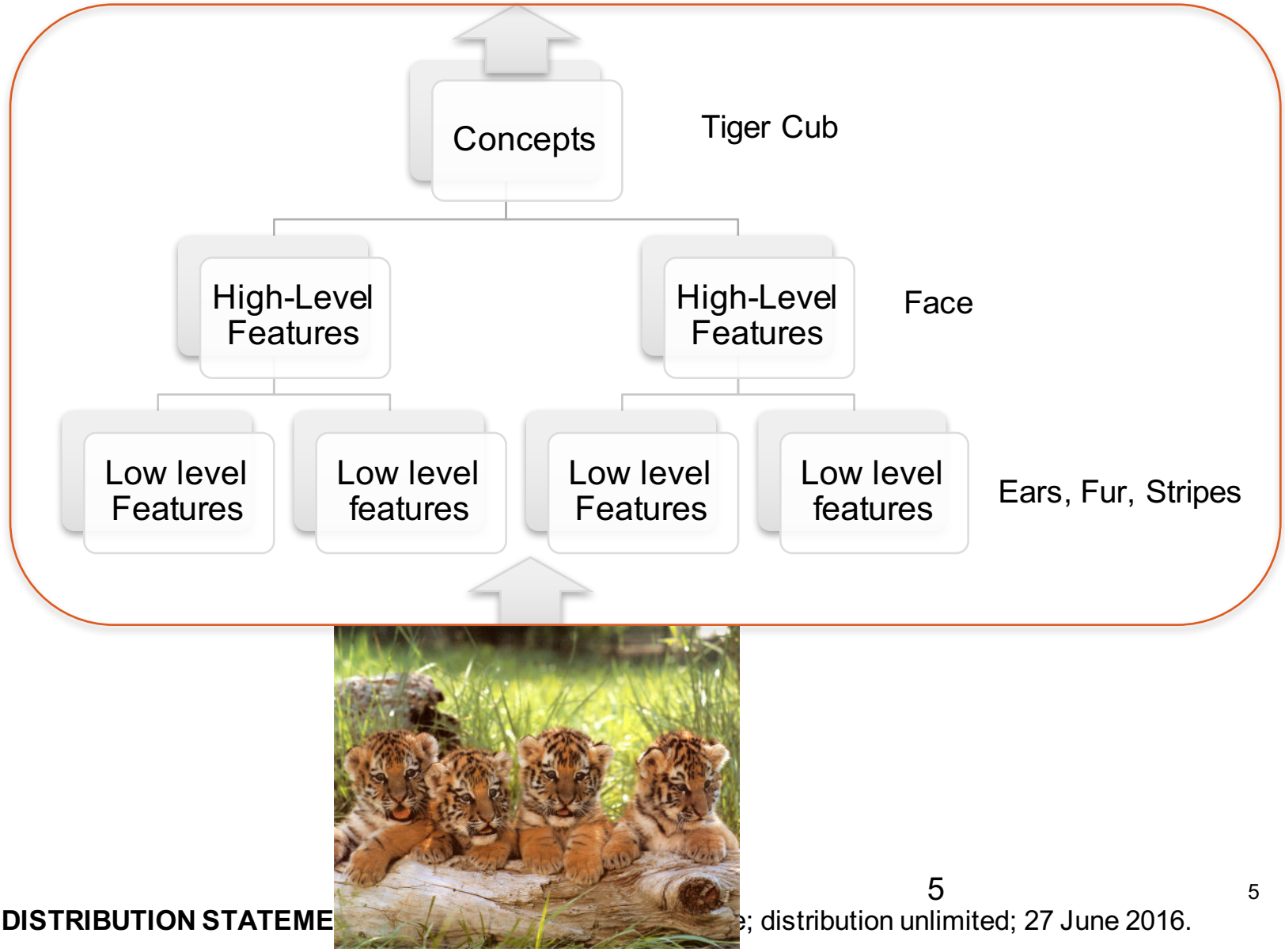
[1] *Scientific American*
[2] www.transhumanist.com
[3] http://sites.psu.edu/cigerber02141993/2014/04/14/are-two-halves-better-the-one-whole/

Brain vs. Conventional Computing

	Brain	Computer
Signals	Mixed signal	Digital
Precision	Low	High
Parallelism	Very high	Low
Information Density	High	Low
Processing-Memory	Closely-coupled	Separated
Mutability	Plastic	Constant

Brain-like computing is better for massively parallel applications with noisy data and relaxed precision requirements

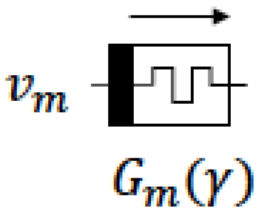
Neuromemristive Systems



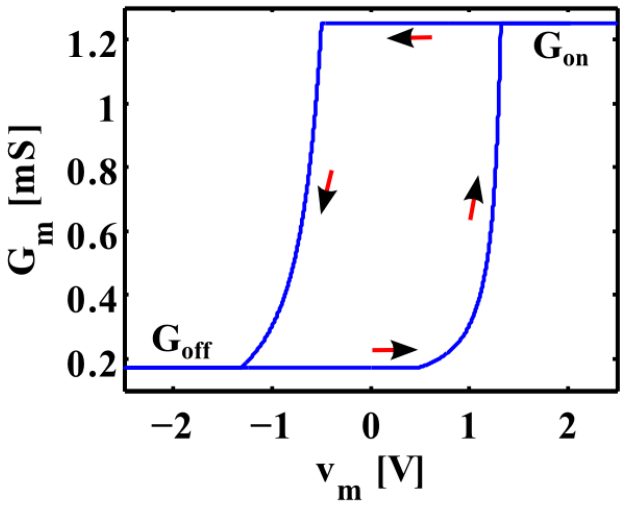
5

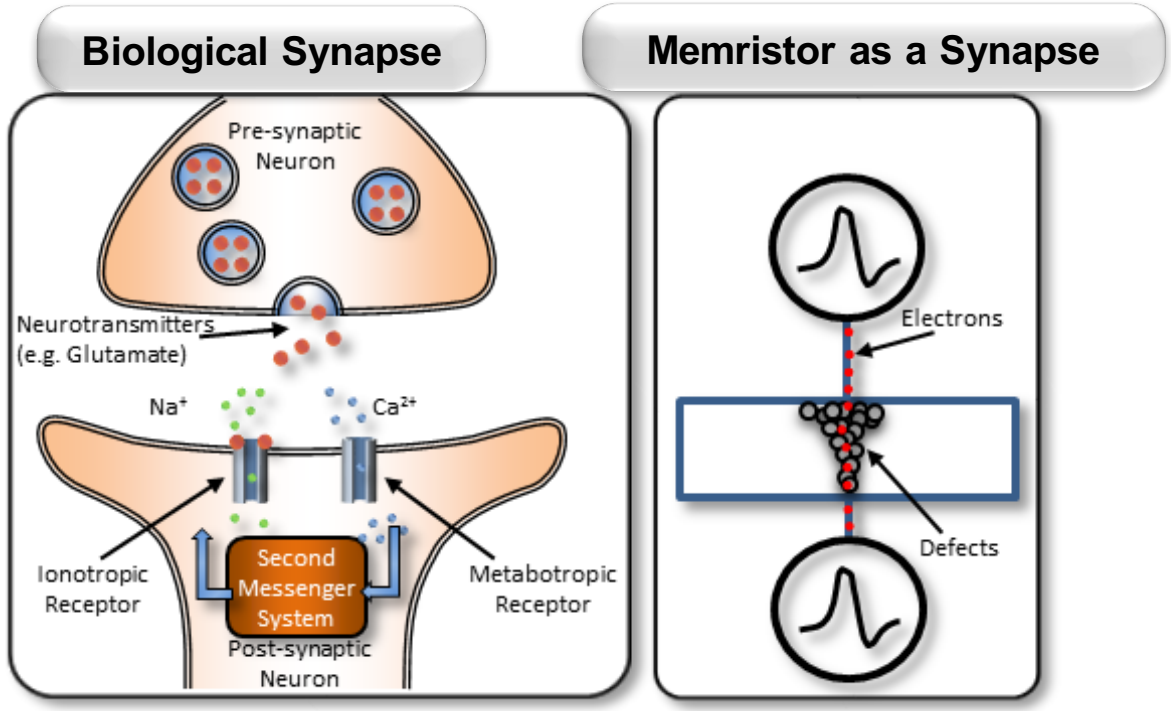
Memristors for Plasticity

$$i_m = G_m(\gamma)v_m$$



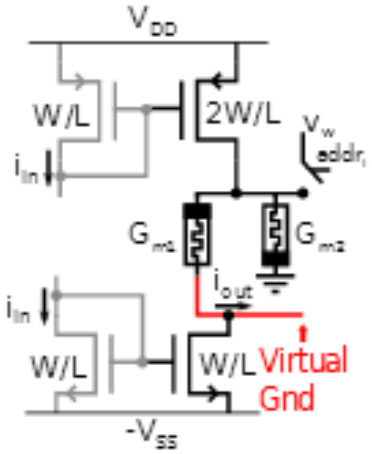
2-terminal device with state-dependent Ohm's Law



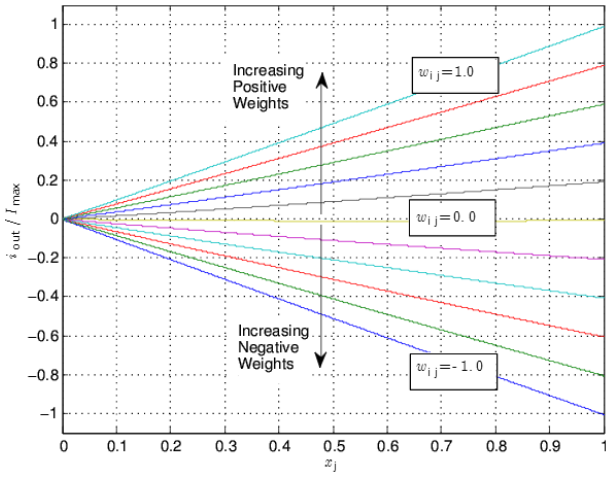


- Compatibility with CMOS
- Memristor characteristics facilitate efficient computation and learning

Reconfigurable Synapses



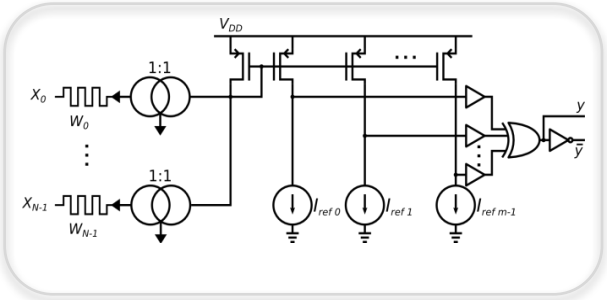
(a)



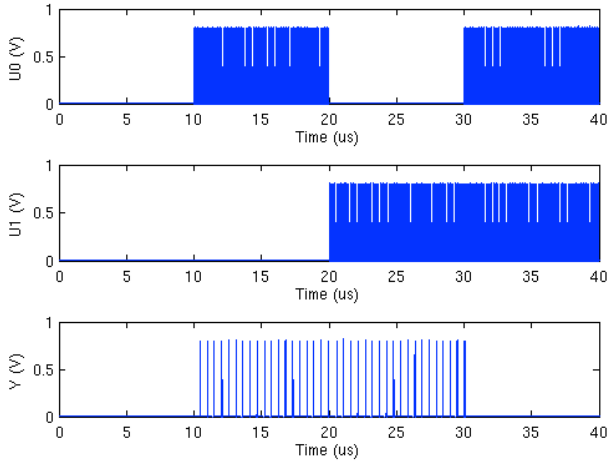
$$s_{ij} = \left(2\frac{G_{m1j}}{G_{m1j}+G_{m2j}} - 1\right)x_j = w_{ij}x_j$$

Inhibitory and Excitatory Synapses

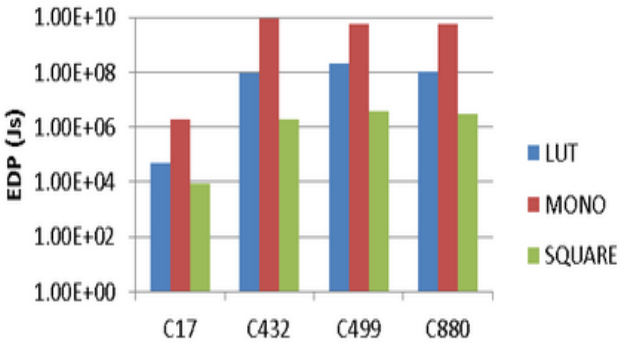
Reconfigurable Neurons



Non-Monotonic Neuron



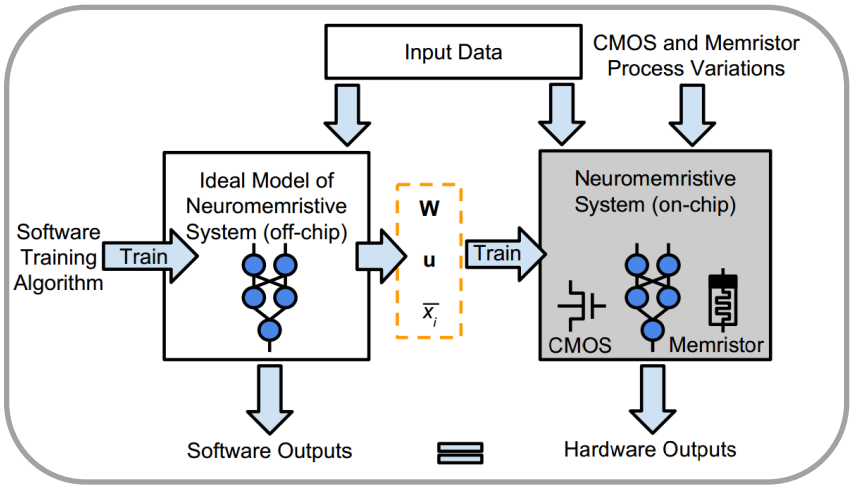
Edge Detection



Energy-Delay Product

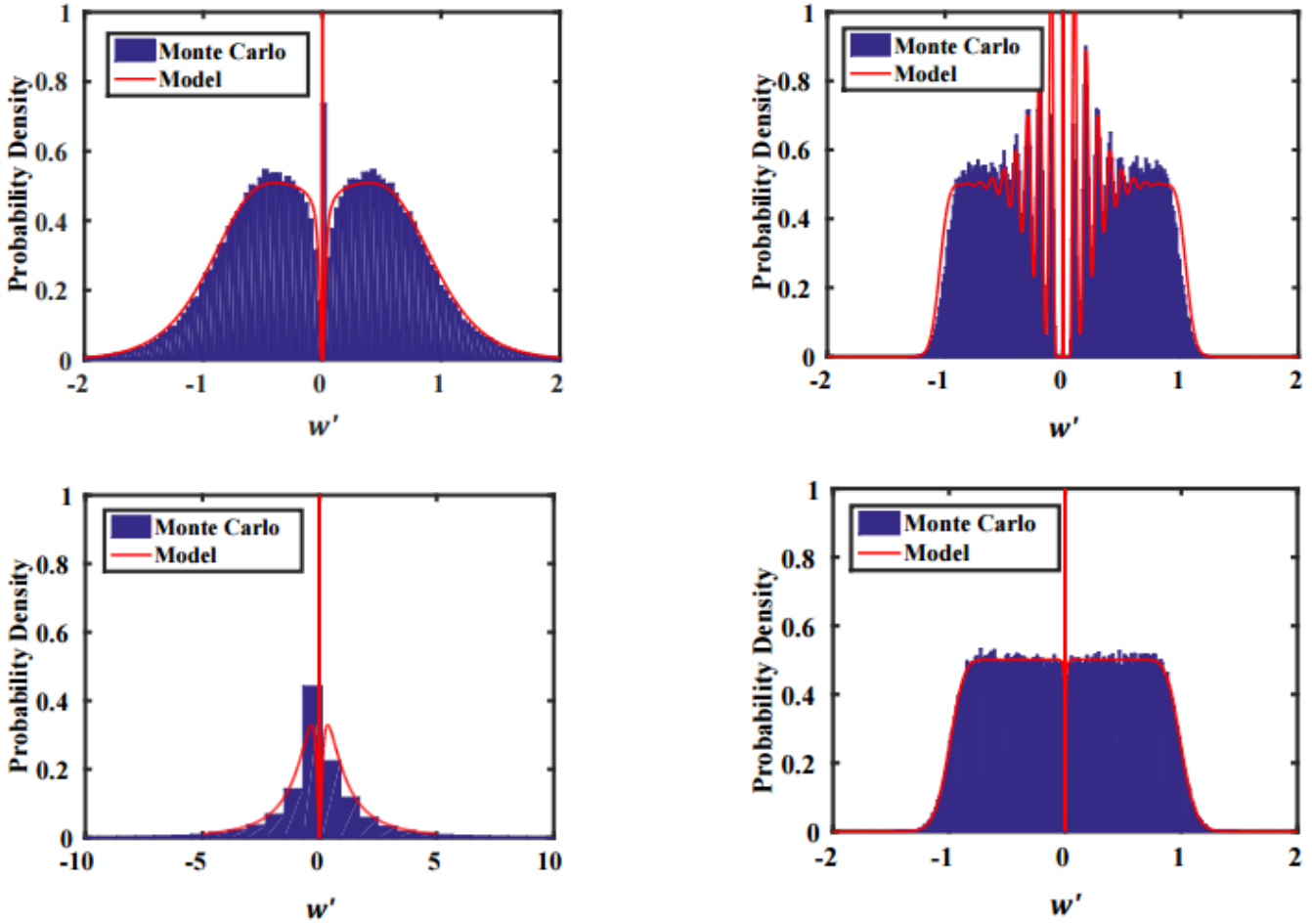
On-Chip Training

- Variation is exploited in the training process



Random Weight Synapses

- Exploit random mismatch in current mirrors
- Control distribution with sizing



On-Chip Training

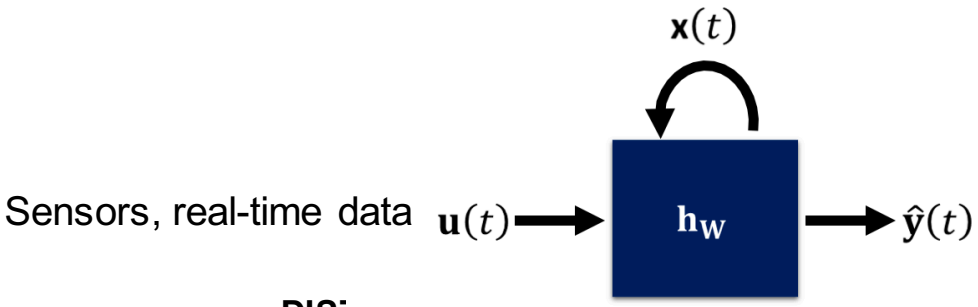
- Least-mean squares (LMS) training algorithm:

$$\Delta w_{ij}^{(p)} = \alpha u_j^{(p)} \left(y_i^{(p)} - \hat{y}_i^{(p)} \right)$$

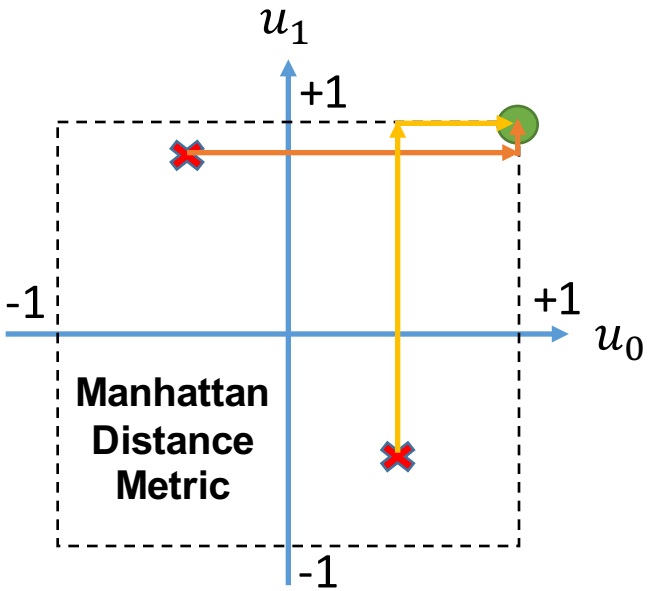
Expected output

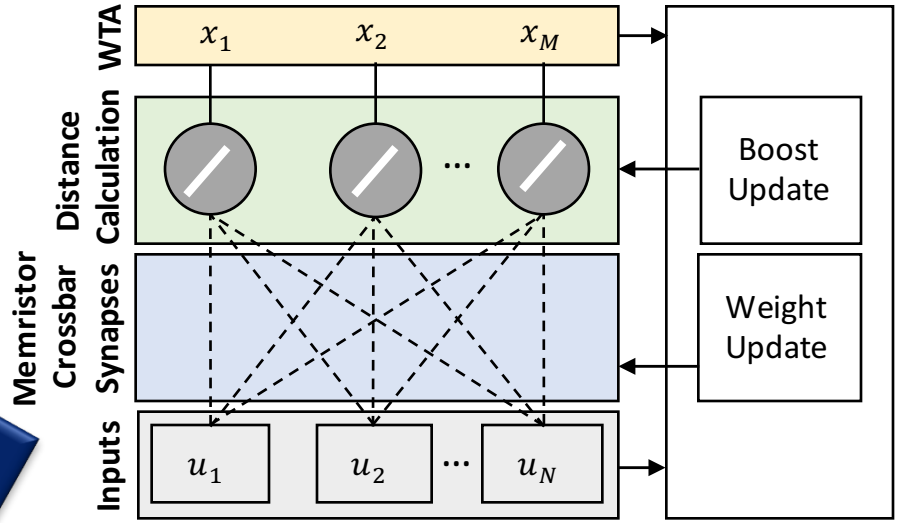
- Converted to stochastic LMS (SLMS) using proposed method:

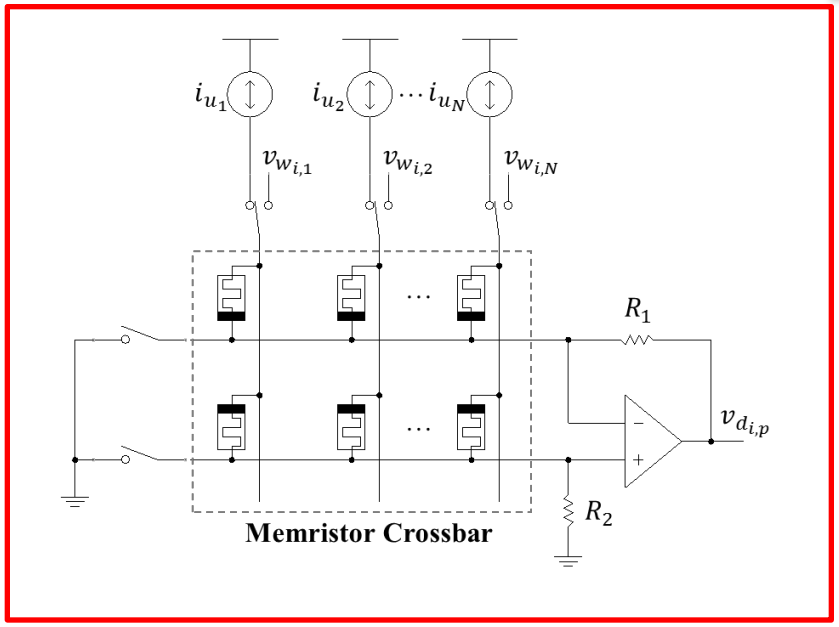
$$\Delta w_{ij}^{(p)} = \alpha \mathrm{sgn}\left(y_i^{(p)} - \hat{y}_i^{(p)} \right) X_j^{(p)} \left| Y_i^{(p)} - \hat{Y}_i^{(p)} \right|$$

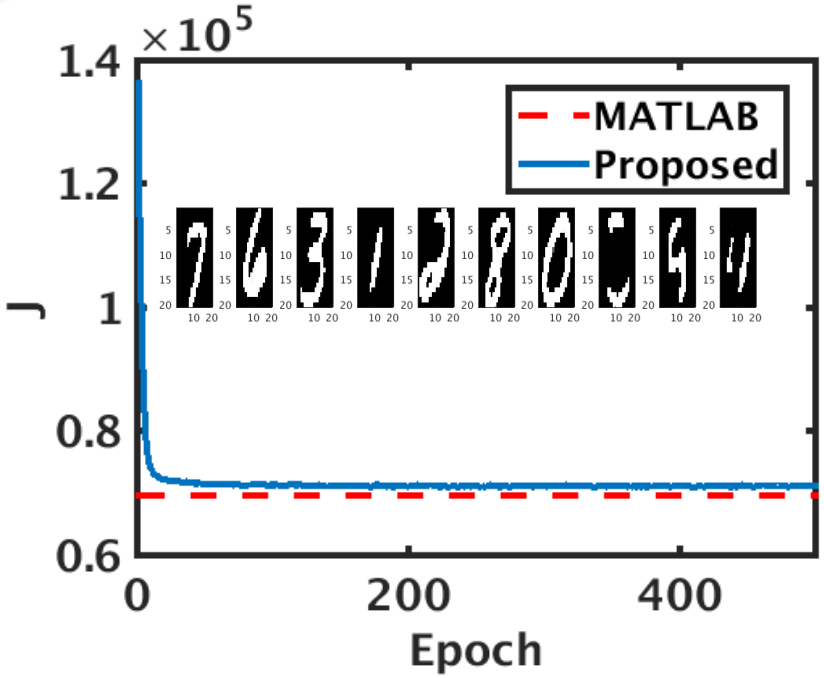


Unsupervised Clustering





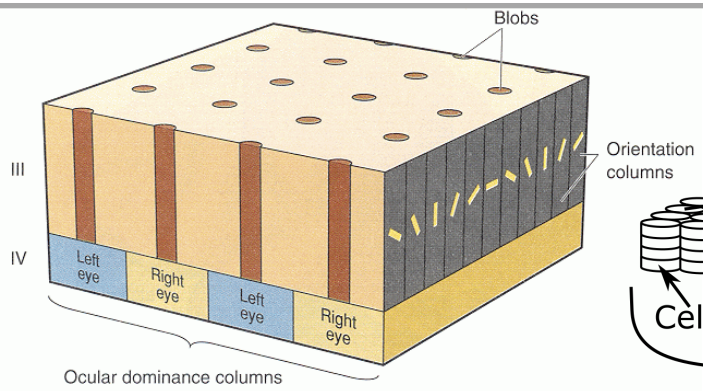


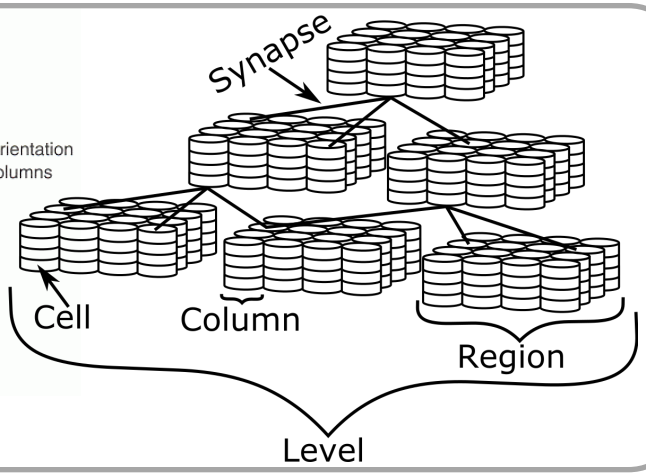


Hierarchical Temporal Memory

Inspiration/Motivation

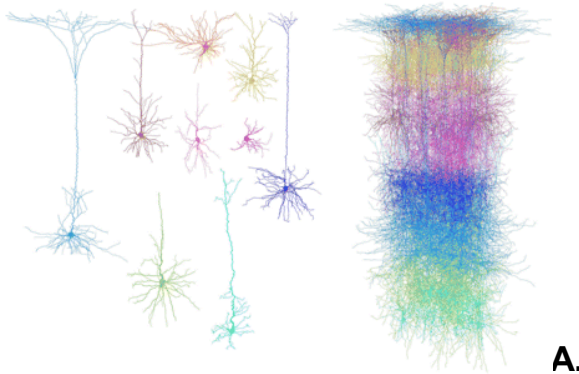
- Inspired by the neocortex
- Highly parallelizable
- Suitable for hardware design



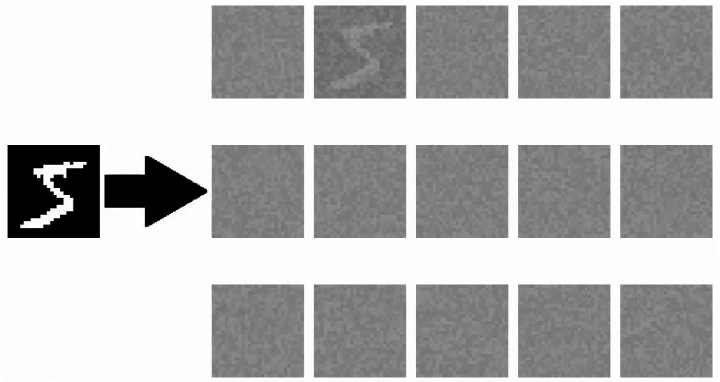


Critical Aspects

- Spatiotemporal data
- Online, unsupervised learning
- Classification & prediction
- Distinct learning components
- Customizable architecture



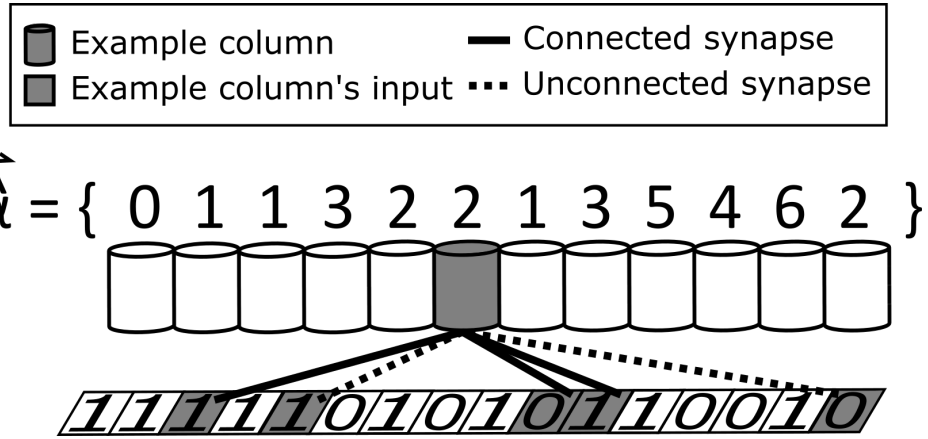
Applications/Results



```
Given     : (A)  11111-.....-.....
Predicted : (?)  .....-.....-.....
```

A. Approved for public release; distribution unlimited; 27 June 2016.

Mathematical Formalization of the Spatial Pooler



http://arxiv.org/abs/1601.06116

https://github.com/tehtechguy/mHTM

$$\vec{\alpha} \equiv \begin{cases} \vec{\hat{\alpha}}_i \vec{b}_i & \vec{\hat{\alpha}}_i \geq \rho_d, \\ 0 & \text{otherwise} \end{cases} \quad \forall i \;\; \vec{\hat{\alpha}}_i \equiv \boldsymbol{X}_i \bullet \boldsymbol{Y}_i$$

Overlap

$$\vec{\hat{c}} \equiv \mathrm{I}(\vec{\alpha}_i \geq \vec{\gamma}_i) \;\forall i$$

$$\vec{\gamma} \equiv \max(\mathrm{kmax}(\boldsymbol{H}_i \odot \vec{\alpha}, \rho_c), 1) \;\forall i$$

inhibition

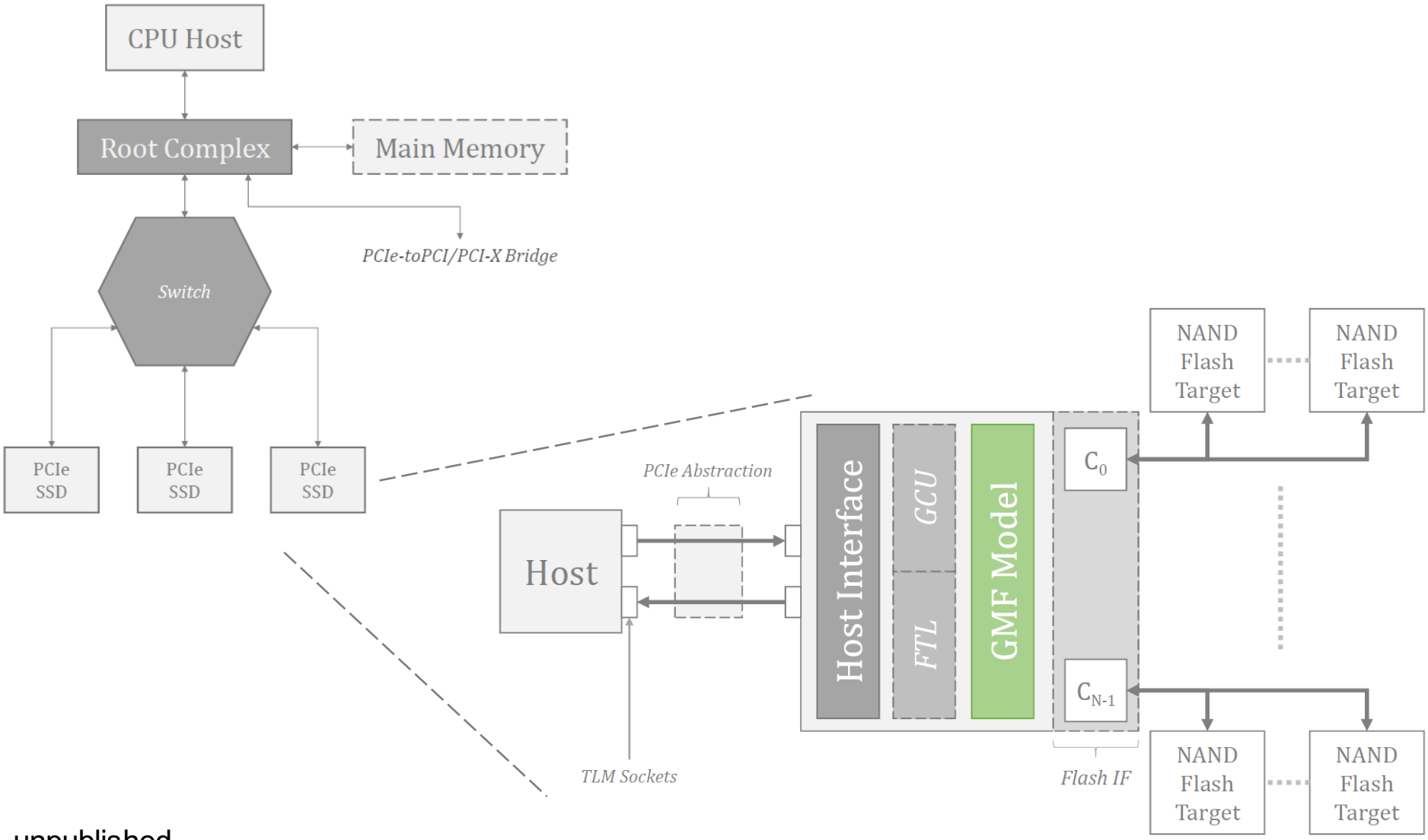
$$\boldsymbol{\delta\Phi} \equiv \vec{\hat{c}}^{T} \odot (\phi_+ \boldsymbol{X} - (\phi_- \neg \boldsymbol{X}))$$

$$\boldsymbol{\Phi} \equiv \mathrm{clip}\,(\boldsymbol{\Phi} \oplus \boldsymbol{\delta\Phi}, 0, 1)$$

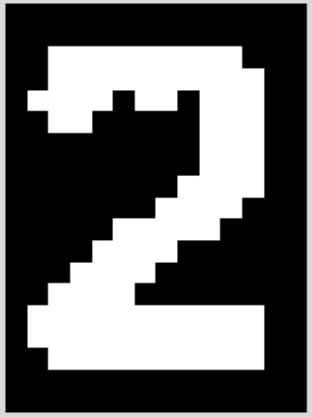
Learning

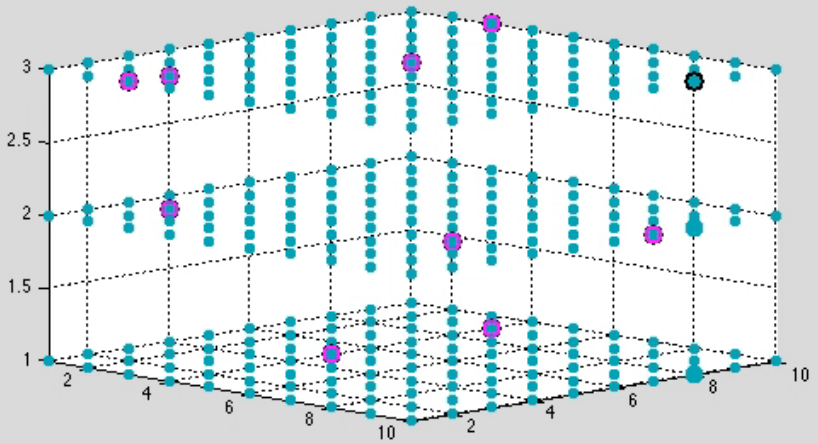
Reconfigurable HTM Architecture

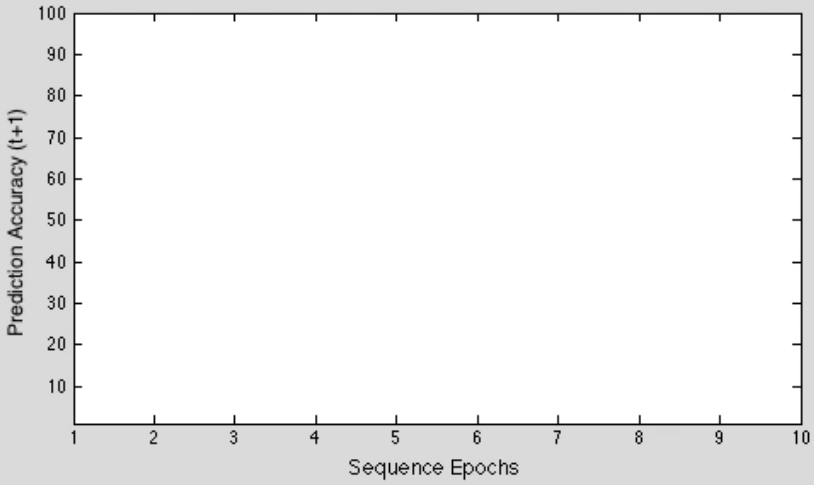
Storage processor units may leverage PCIe SSD technology

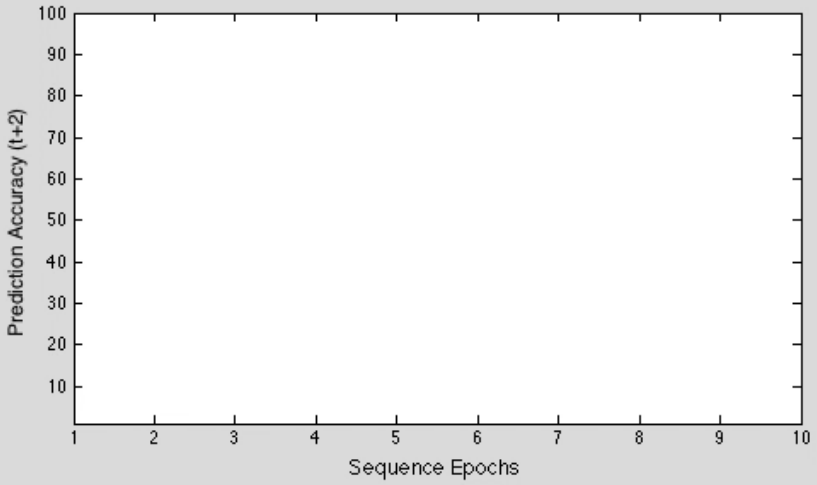


Reconfigurable HTM Architecture

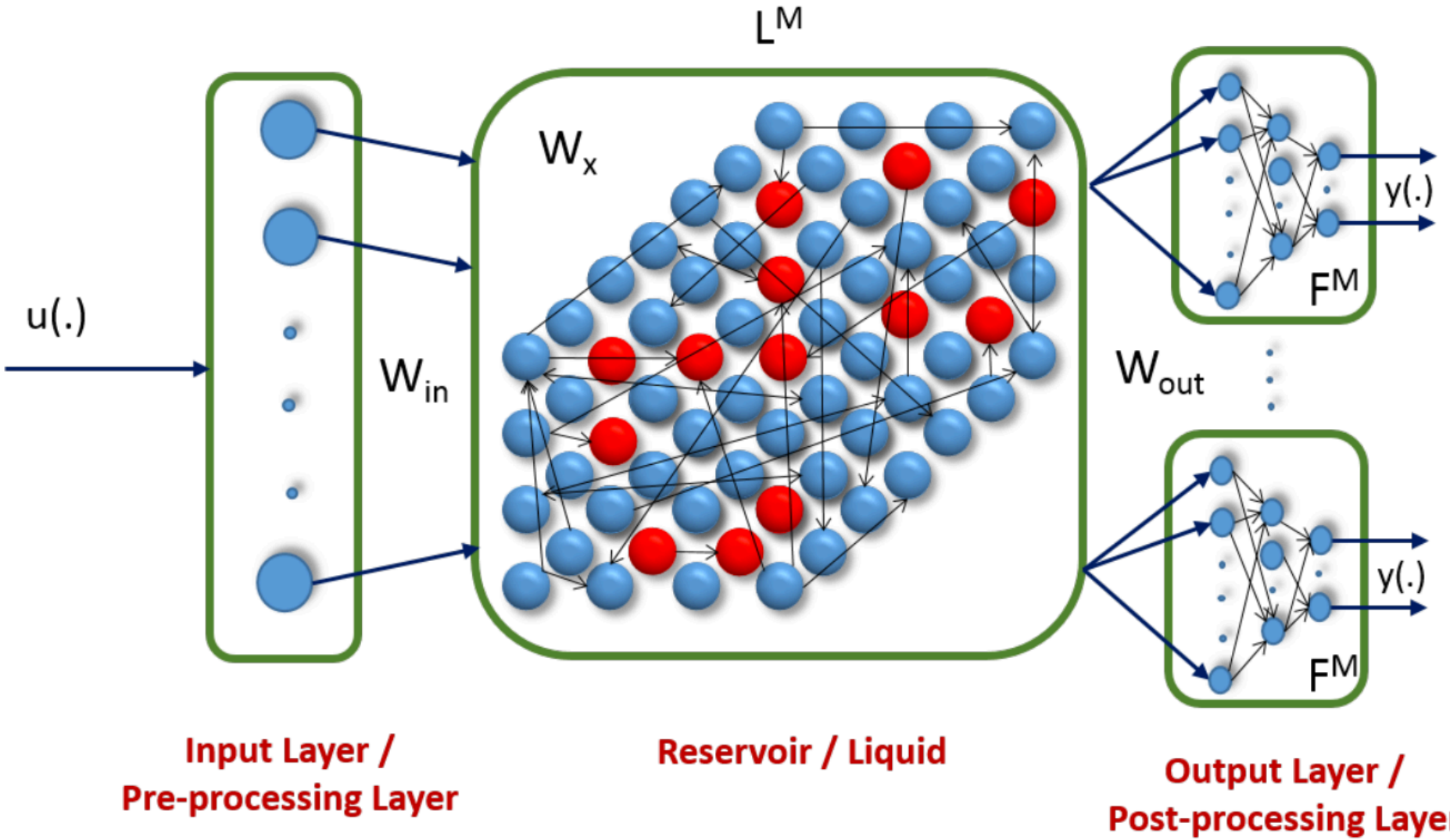




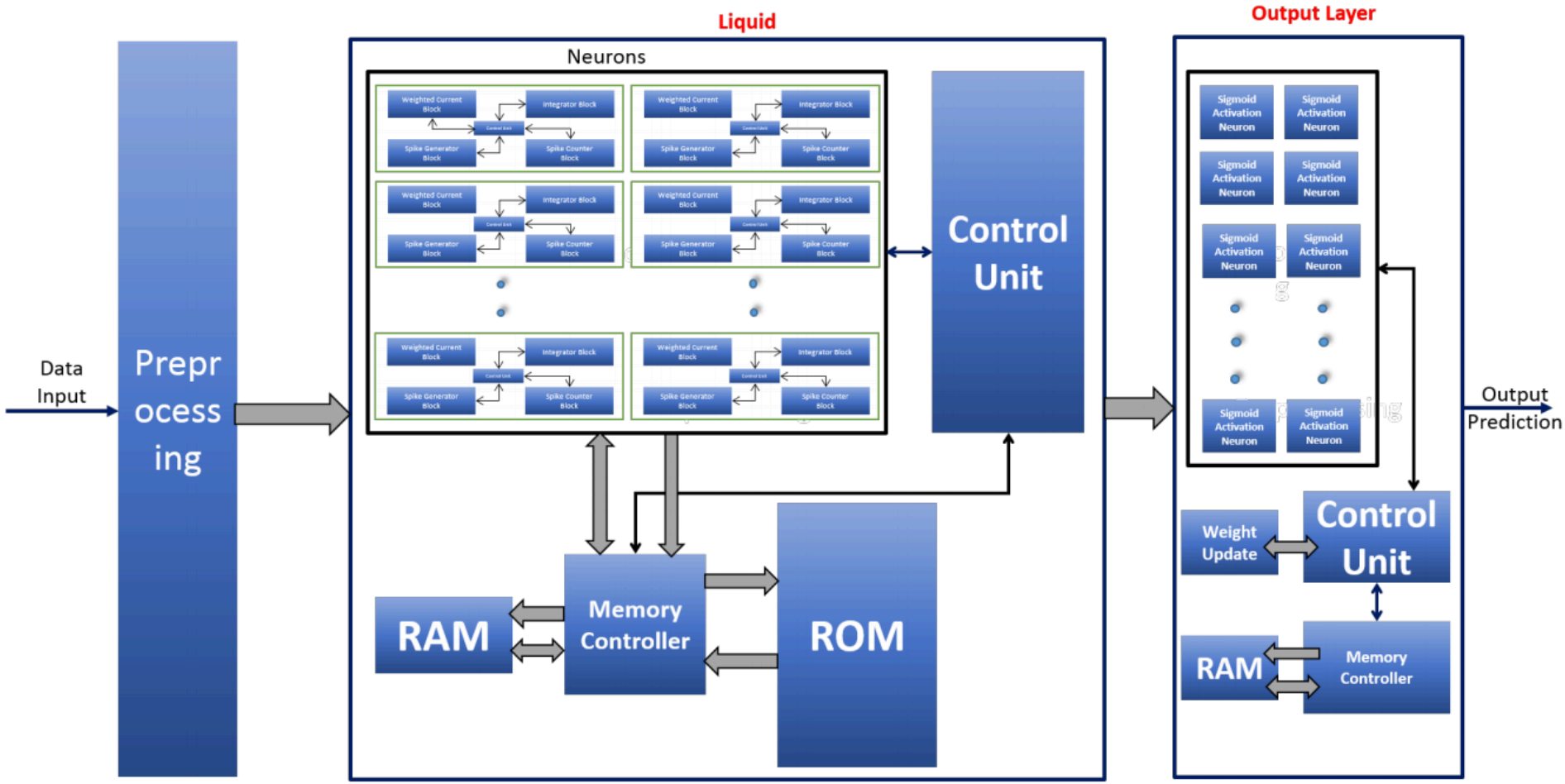




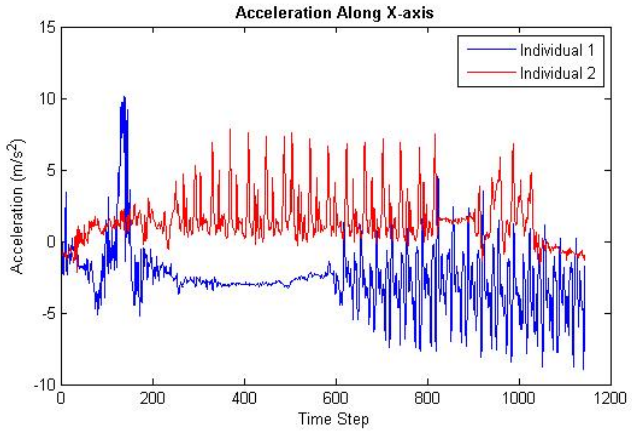
Generalizable Intelligence Engine

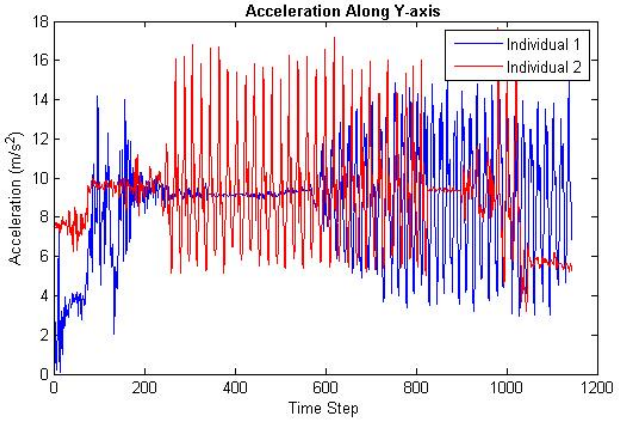


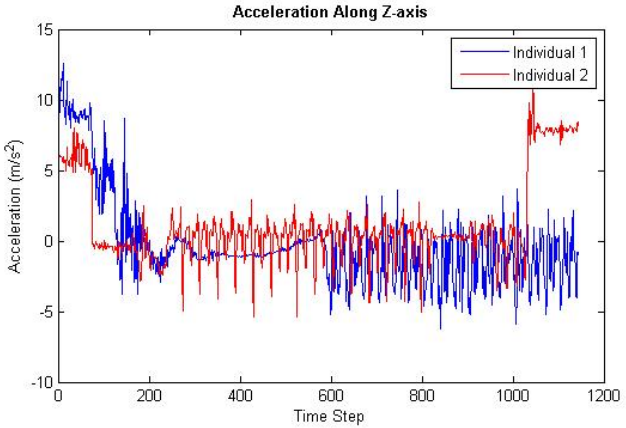
Reconfigurable Reservoir Architecture

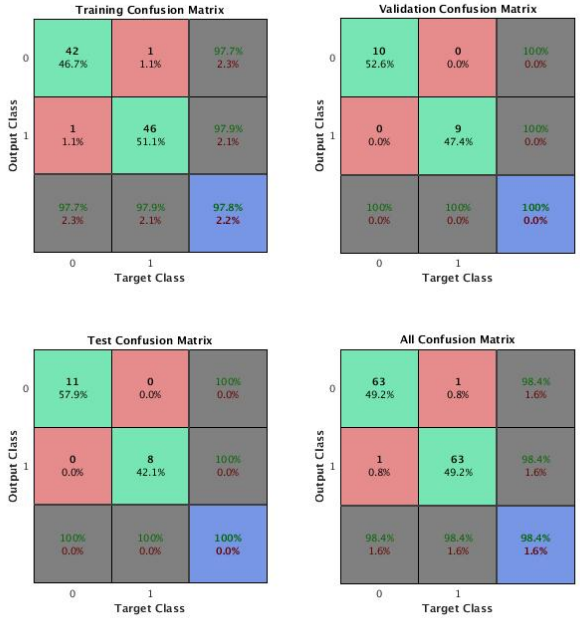


Reconfigurable Reservoir Architecture





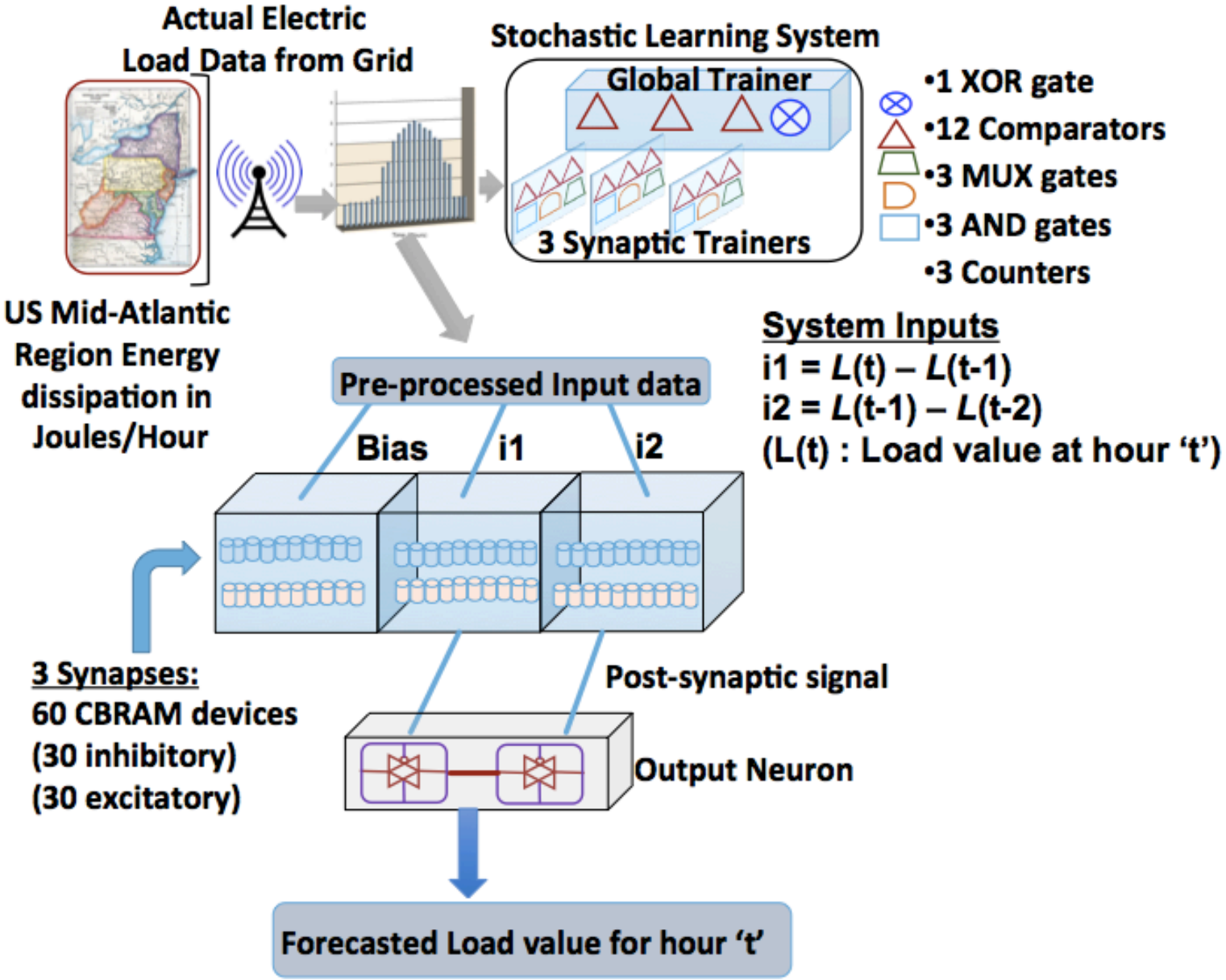




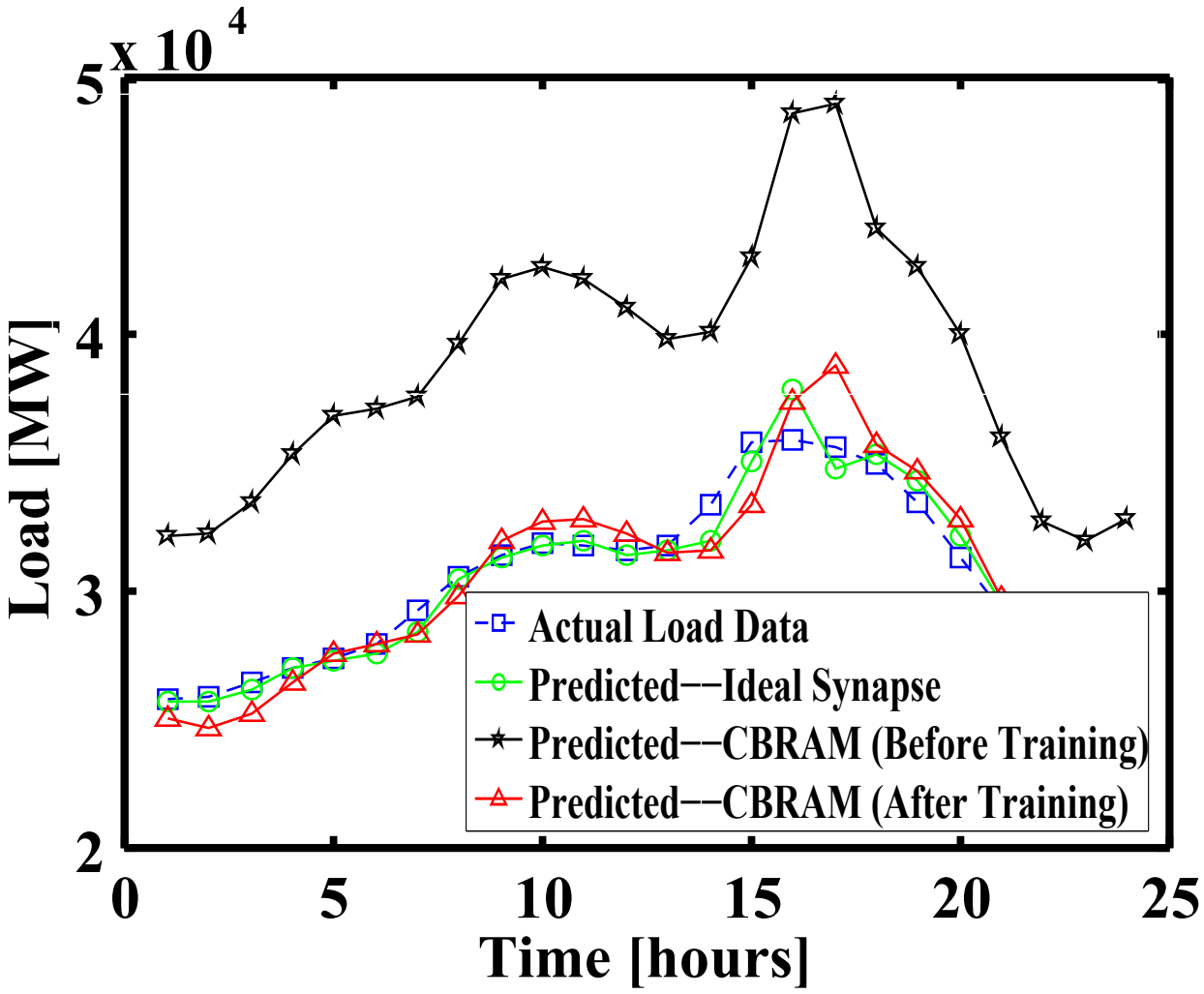
User Authentication based on Gait Patterns

Rebooting'16

Smart Grid Load Forecasting



Smart Grid Load Forecasting



Summary

- Reconfigurability is integral to the nature of computation
  - –Precomputation is occurring in communication channels
  - –No standardized metrics/benchmarks to evaluate
  - – Designing technology agnostic vs. technology aware systems
- Looking forward
  - – one shot learning ….

Team & Collaborators