Neuromorphic Electronics for Energy Efficient Processing

Narayan Srinivasa

Center for Neural and Emergent Systems
Information and Systems Sciences Lab
HRL Laboratories LLC
Malibu, CA

ITRS ERD Presentation, Stanford University
02/26/15

Mammalian Brain vs. Computers: Architectural Aspects

Brain	Computer
Operates at low speeds (< 10 Hz)	Operates at very high speeds (> GHz)
Spike Encoding	Digital Encoding
Power and energy efficient	Power and energy hungry
Asynchronous (no global clock) – *clock free*	Synchronous (global clock) – *clock aligned*
Analog computing and communication	Digital computing and communication
Memory and computation are fully integrated	Memory and computation are separated
Small world network organization with *multi-scale interactions*	Access to a single pre-determined scale with local interactions only

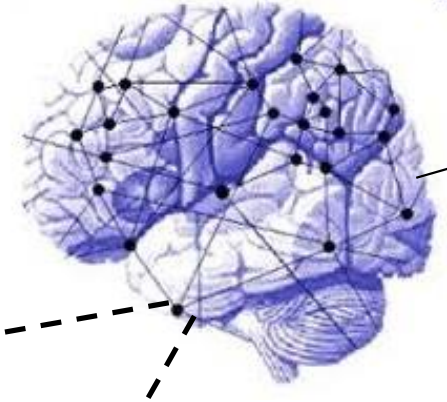
N. Srinivasa, Design considerations for a computational architecture of human cognition, *Emerging Nanoelectronic Devices*, First Edition, Ann Chen, James Hutchby, Victor Zhirnov and George Bourianoff, John Wiley and Sons, Ltd, January, 2015.

Brain Architecture

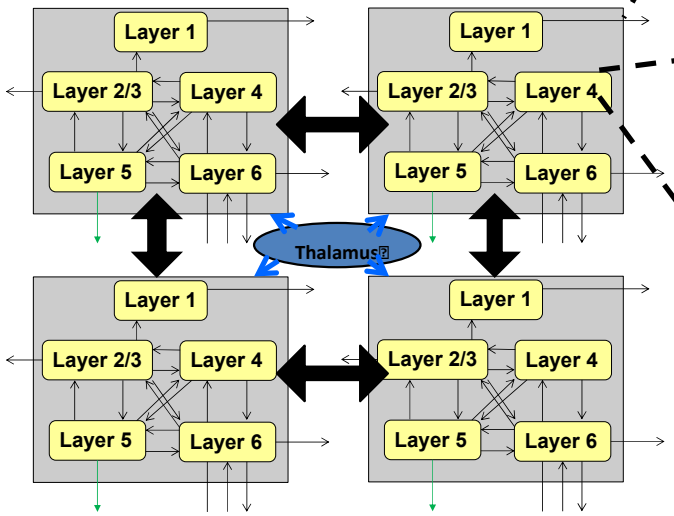
Evolutionary Objective: Keeping synaptic path length fairly constant with brain size

Organization:
Degree of Clustering
Degree of Separation

Both are needed to achieve large scale traffic with minimum wiring – Efficient communication

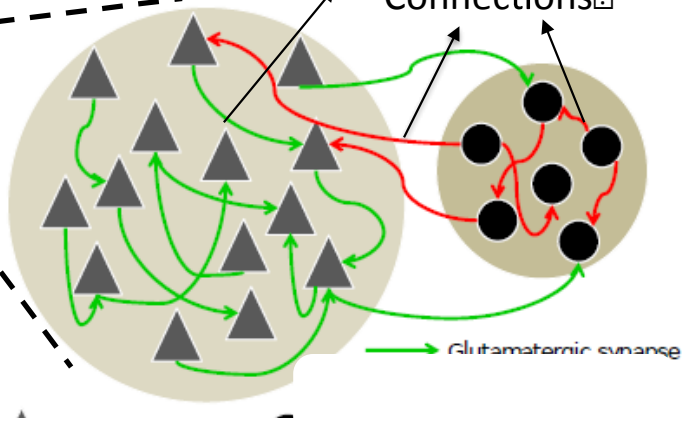


Sparse Long Range Connections



Thalamocortical Circuits

Dense Short Range Connections



E-I Network

N. Srinivasa, Design considerations for a computational architecture of human cognition, *Emerging Nanoelectronic Devices*, First Edition, Ann Chen, James Hutchby, Victor Zhirnov and George Bourianoff, John Wiley and Sons, Ltd, January 2015.

Mammalian Brain vs. Computers: Other Features and Dynamics

Brain	Computer
Composed of faulty, noisy components that are short-lived but can regenerate – *fault tolerant*	Made of precise components which have longer lives but cannot regenerate – fault sensitive
Fully distributed processing that can integrate multiple sources of information and process multiple goals simultaneously - *grid-free*	Serial/Parallel processing that can integrate multiple sources of information towards a common goal
Very limited encoding into symbols – *symbol-free*	Encoding of concepts using symbols
Spontaneously active	No activity unless instructed
Intelligence via Learning thru BBE interactions	Intelligence via programmed algorithms

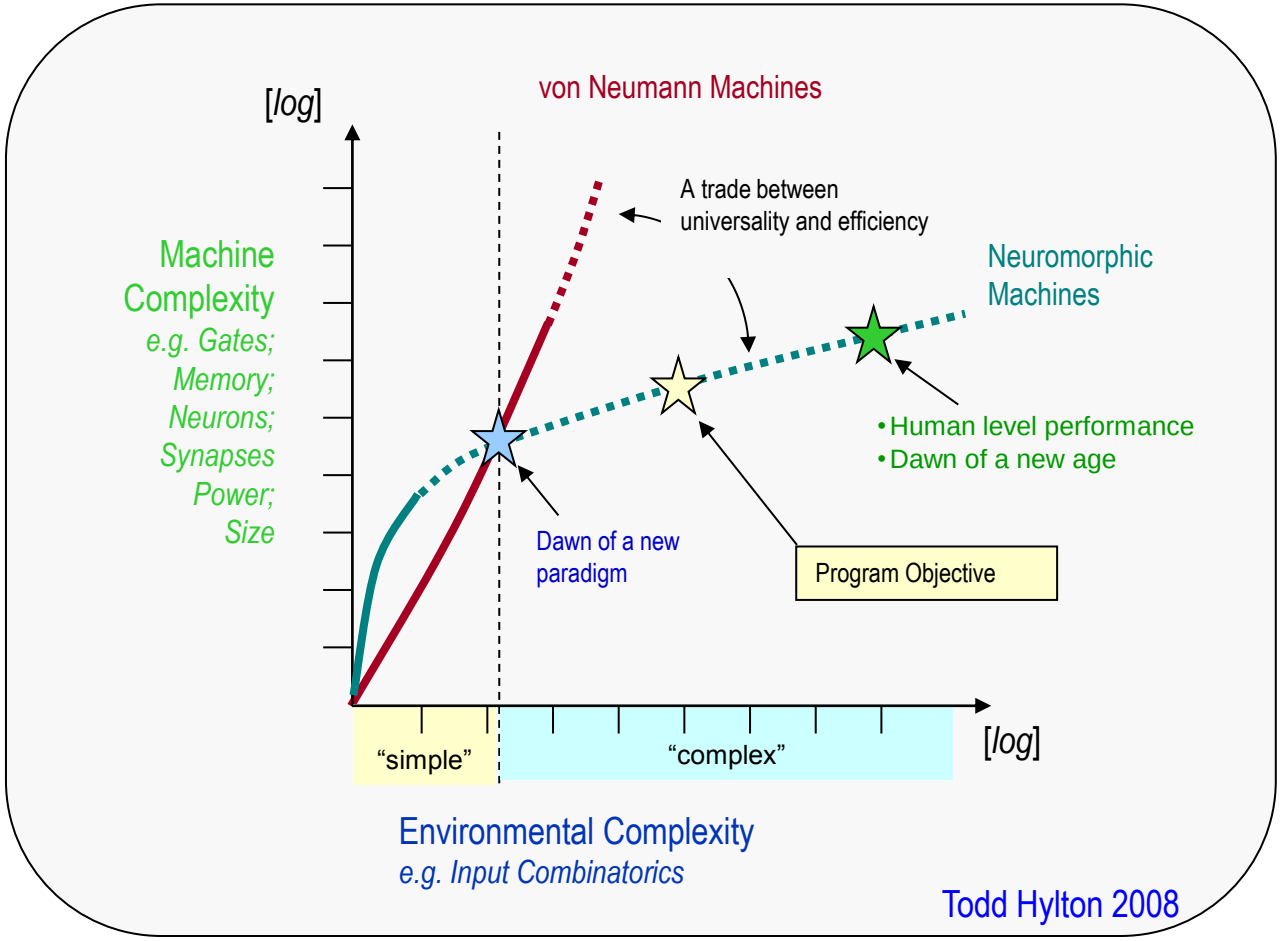
N. Srinivasa, Design considerations for a computational architecture of human cognition, *Emerging Nanoelectronic Devices*, First Edition, Ann Chen, James Hutchby, Victor Zhirnov and George Bourianoff, John Wiley and Sons, Ltd, January, 2015.

DARPA SyNAPSE: Motivation and Objective

Problem

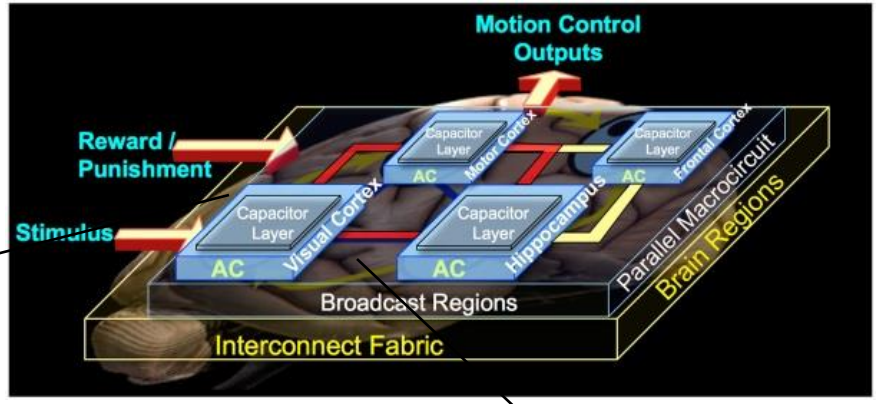
- As compared to biological systems, today's intelligent machines are less efficient by a factor of a million to a billion in complex environments.

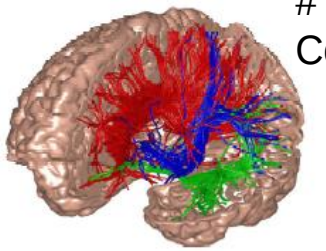
- For intelligent machines to be useful, they must compete with biological systems.



The SyNAPSE program sought to break the programmable machine paradigm by developing neuromorphic machine technology that scales to biological levels

Large Scale Neuromorphic System





Brain Architecture

Neurons,
Synapses,
Connectivity

Neuromorphic Compiler

Programmable Front-End

10^6 neurons
10^6 synapses
10^{10} virtual synapses

Routing, Neuron Placement

Digital Memory

Set switch states

Acquire Switch states

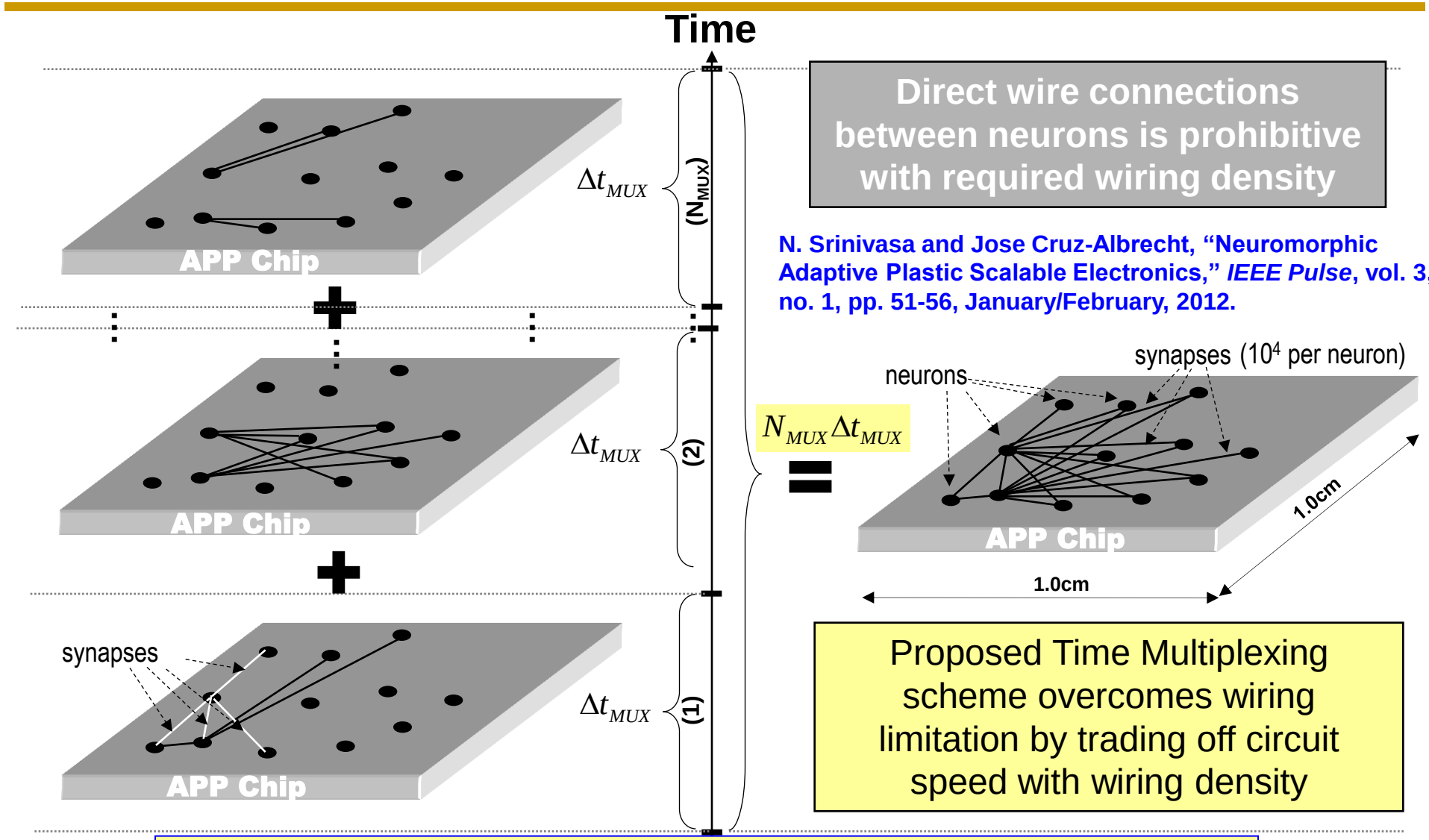
Analog Core with Cortical Fabric
(neurons, synapses)

Store

Retrieve

Analog Memory
(store synaptic conductances)

Time Multiplexing Concept to Address Connectivity Challenge



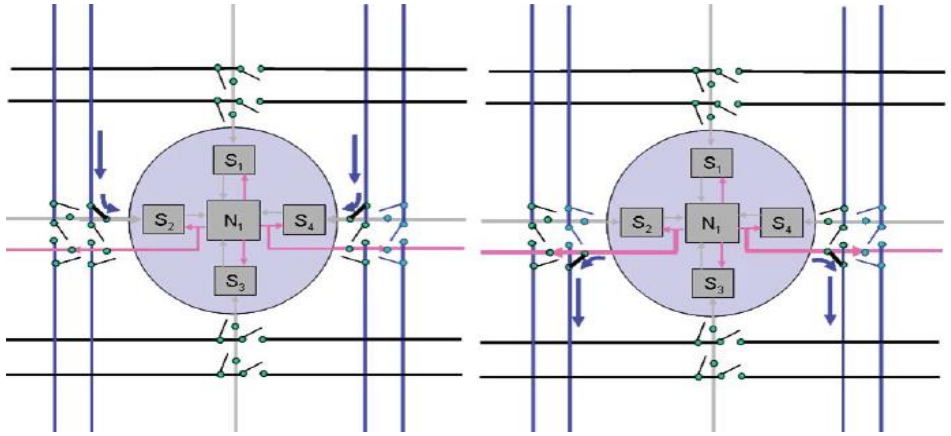
Direct wire connections between neurons is prohibitive with required wiring density

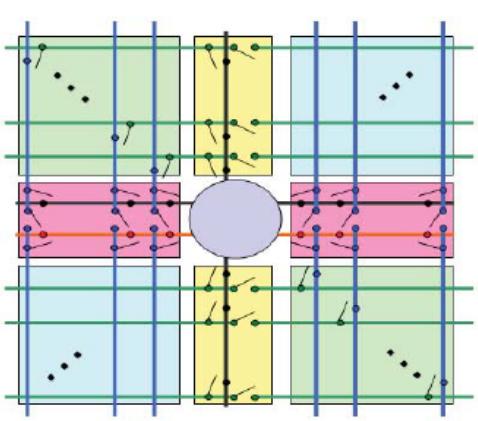
N. Srinivasa and Jose Cruz-Albrecht, "Neuromorphic Adaptive Plastic Scalable Electronics," *IEEE Pulse*, vol. 3, no. 1, pp. 51-56, January/February, 2012.

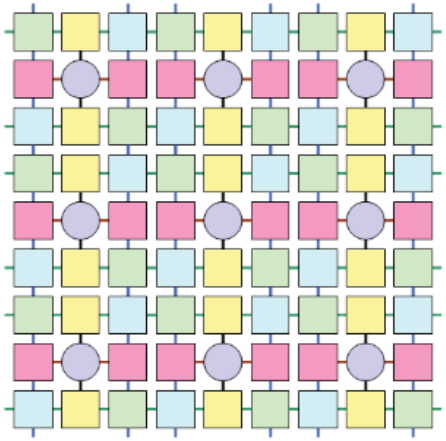
Proposed Time Multiplexing scheme overcomes wiring limitation by trading off circuit speed with wiring density

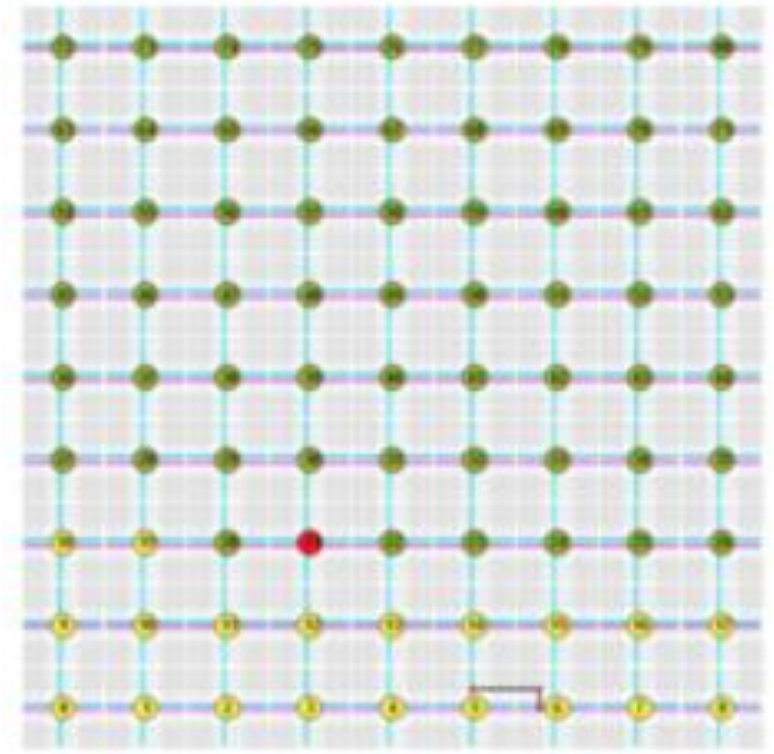
Scalable solution to enable CMOS based neuromorphic chip design

Time-Multiplexed Fabric & Neuromorphic Compiler to address programmable connectivity challenge







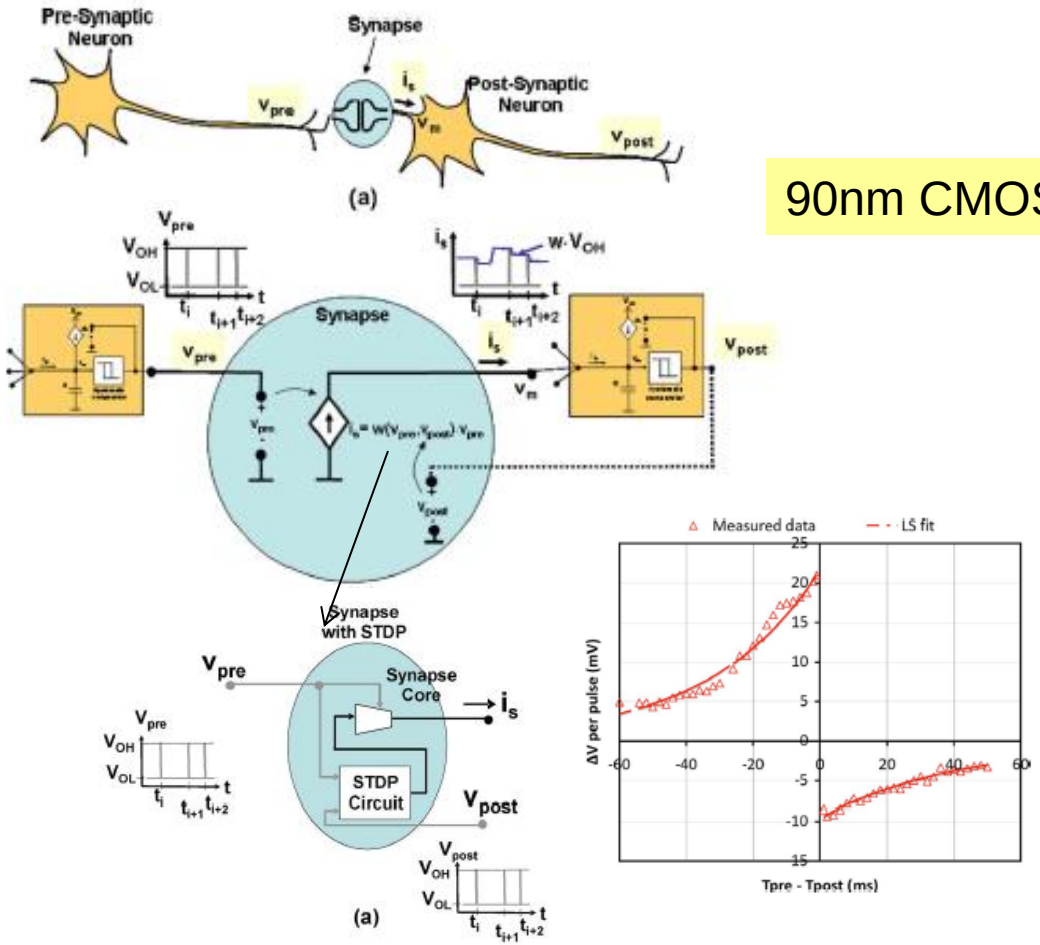


K. Minkovich, N. Srinivasa, J. M. Cruz-Albrecht, Y. K. Cho and A. Nogin, "Programming Time-Multiplexed Reconfigurable Hardware Using a Scalable Neuromorphic Compiler," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 889-901, June 2012.

Time-multiplexing fabric with novel compiler ensures flexible connectivity in CMOS

Mixed Signal CMOS neurons and synapses to address energy efficiency challenge

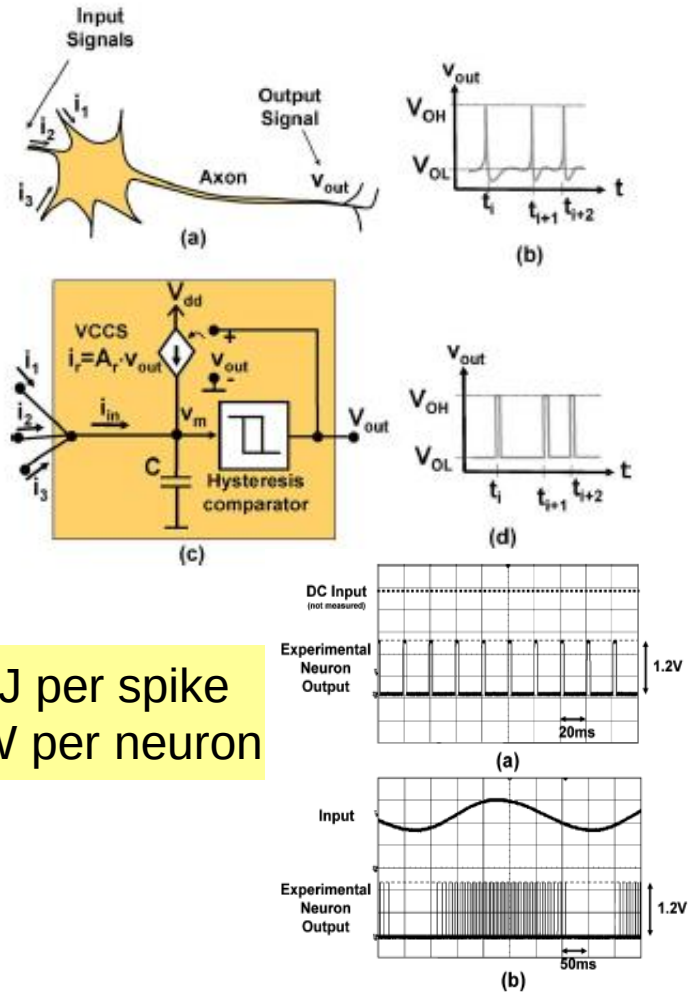
Synapse with STDP



90nm CMOS

0.4pJ per spike
10nW per neuron

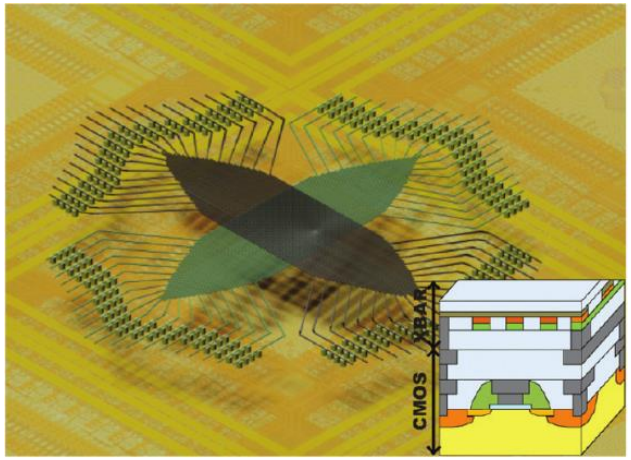
I & F Neuron



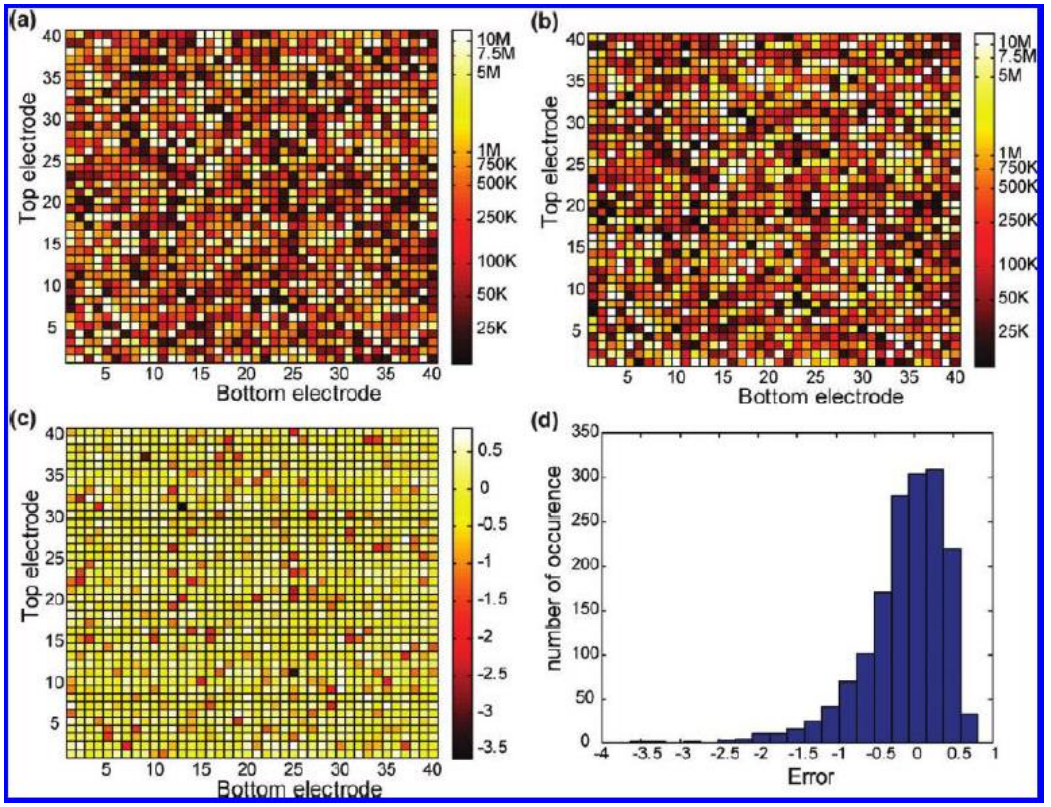
Jose Cruz-Albrecht, Michael Yung, Narayan Srinivasa, "Energy-Efficient, Neuron, Synapse and STDP Integrated Circuits, " ***IEEE Transactions on Biomedical Circuits and Systems*****, vol. 6. No. 3, pp. 246-256, June, 2012.**

Analog memristor array with CMOS integration to address high density memory challenge ($3*10^{10}$ bits/cm^2)

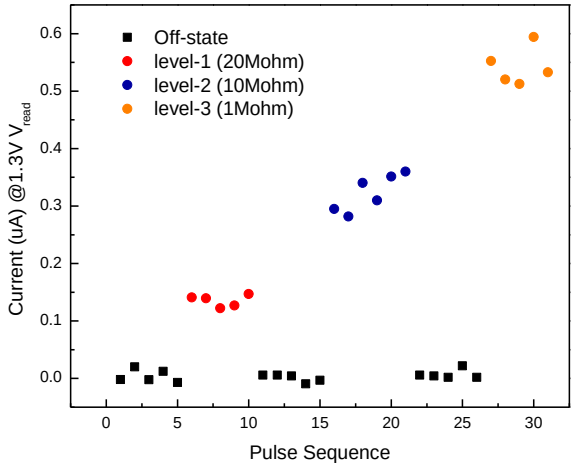
CMOS circuit with memristor



Data written on memristor array (40x40)



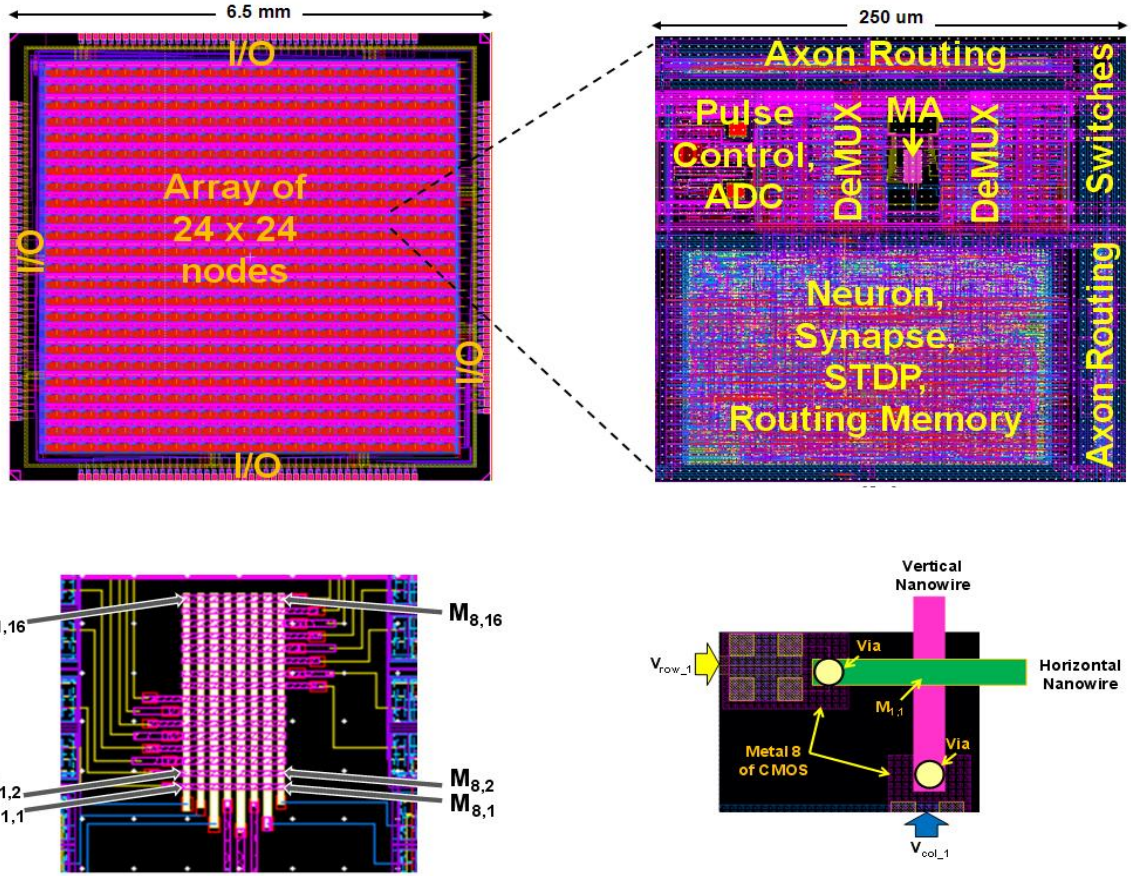
Multibit values written on memristor device within integrated chip



K. H. Kim, S. Gaba, D. Wheeler, J. Cruz-Albrecht, T. Hussain, N. Srinivasa and W. Lu, "A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications" *Nano Letters*, vol.12, no. 1, pp. 389–395, February/March 2012.

HRL Neuromorphic Chip – Gen 1

- 576 point neurons with receptor kinetics
- > 9K synapses with STDP
- Synaptic time multiplexing
- Axonal delays
- Fully programmable
- 90 nm CMOS
- Integrated memristors
- SRAM
- ~ 36 mm^2
- ~ 30 mW



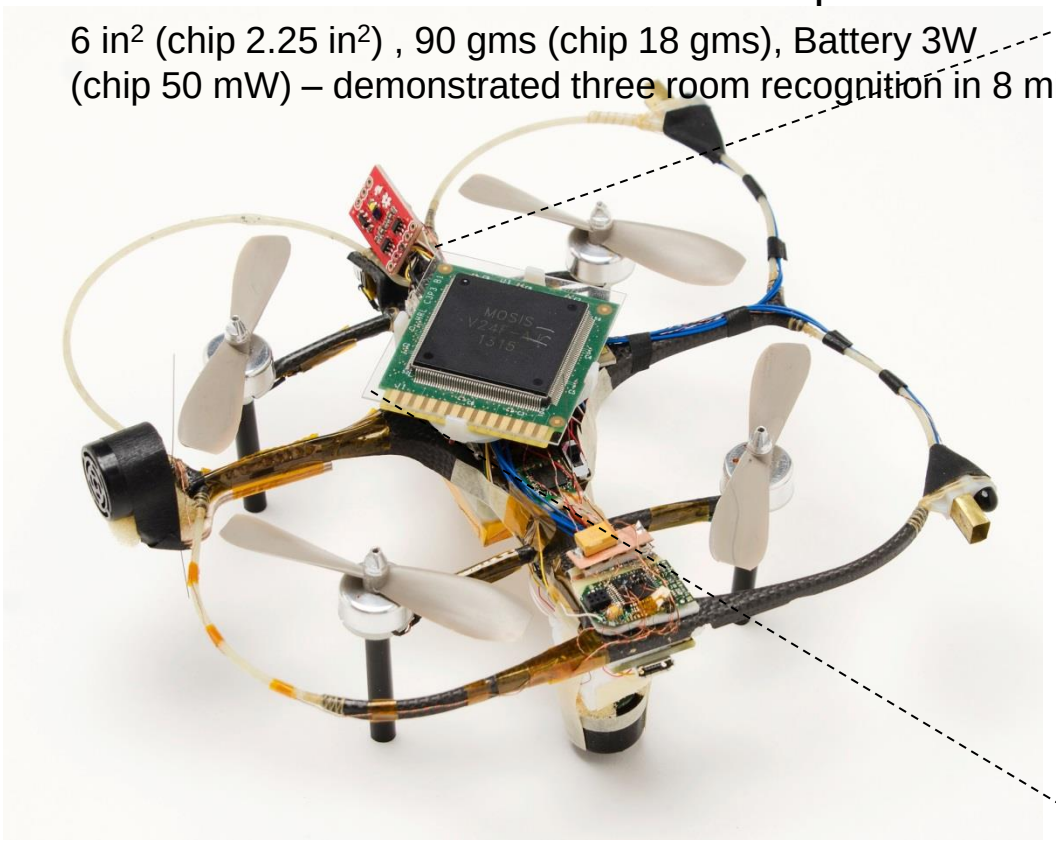
J. Cruz-Albrecht, T. J. Derosier and N. Srinivasa, "Scalable neural chip with synaptic electronics using CMOS integrated memristors", *Nanotechnology*, Special Issue on synaptic electronics, September 2013. doi:10.1088/0957-4484/24/38/384011.

HRL neuromorphic chip provides a key starting point for the development of the real world applications with SWAP constraints

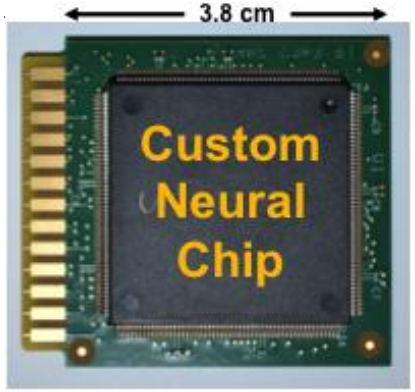
NAV Experimental Integration for Room Recognition

Nano Air Vehicle with Chip

6 in^2 (chip 2.25 in^2) , 90 gms (chip 18 gms), Battery 3W (chip 50 mW) – demonstrated three room recognition in 8 min



(a)



(b)

(c)

Connectors for SPI (Serial Peripheral Interface) with Neural Inputs and Outputs

Connectors for initial configuration of Neural Chip

Featured in MIT Tech Review http://www.technologyreview.com/news/532176/a-brain-inspired-chip-takes-to-the-sky/

Summary

- HRL neuromorphic technology enables
  - spiking enables energy efficient processing
  - can meet low SWAP requirements
  - learning enables autonomous applications
  - currently pursuing a variety of autonomous applications including some for our owners