HRL LABORATORIES

A Compiler for Scalable Placement and Routing of Brain-like Architectures

Narayan Srinivasa

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

International Symposium on Physical Design 2013
March 26, 2013
Lake Tahoe, CA

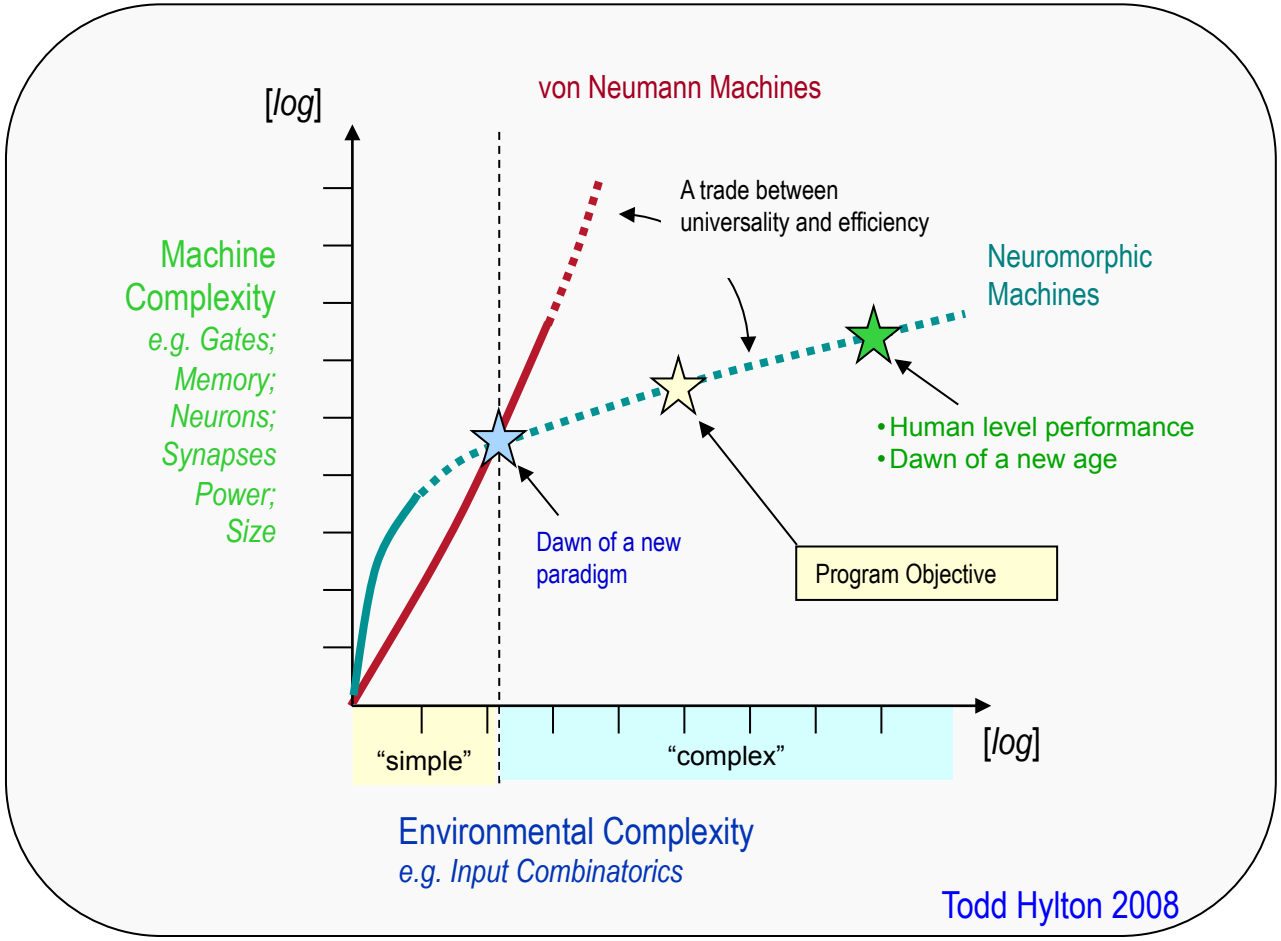
Computers vs. Mammalian Brains

Parallel distributed architecture		Serial architecture
Spontaneously active		No activity unless instructed
Composed of noisy components and operates at low speeds (< 10 Hz)		Precision in components and operates at very high speeds (GHz)
Low power (30W), small footprint (1 liter)		High power (100MW), Large footprint (40M liters)
Asynchronous (no global clock)		Synchronous (global clock)
Analog computing, Digital communication		Digital computing and communication
Integrated memory and Computation		Memory and Computation are clearly separated
Intelligence via Learning thru BBE interactions		Intelligence via programmed algorithms/rules

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 seeks to break the programmable machine paradigm by developing neuromorphic machine technology that scales to biological levels

Program Structure

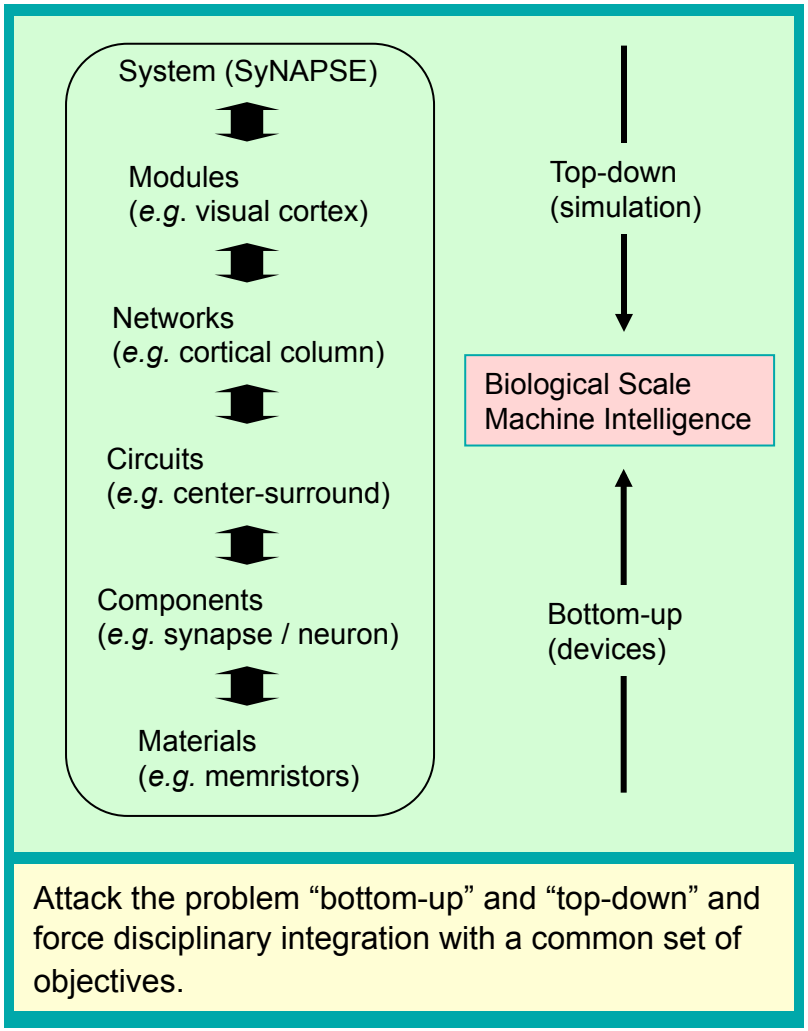
Structure	Period of Performance
Baseline/Phase 0	October 7, 2008 - September 6, 2009
Option 1/Phase 1	September 7, 2009 - March 28, 2011
Option 2/Phase 2	March 29, 2011 - January 27, 2013

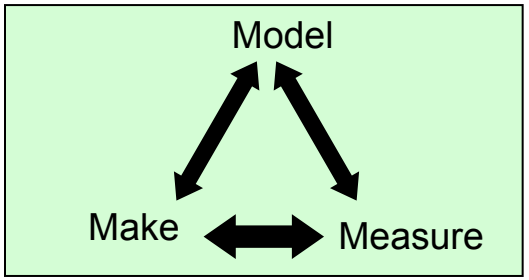
Performers

- HRL (prime)
- Subcontractors
  - University of Michigan
  - Stanford University
  - Neurosciences Institute
  - Boston University
  - University of California, Irvine
  - George Mason University
  - Portland State University
  - SET Corporation

HRL SyNAPSE Team

Overall Approach

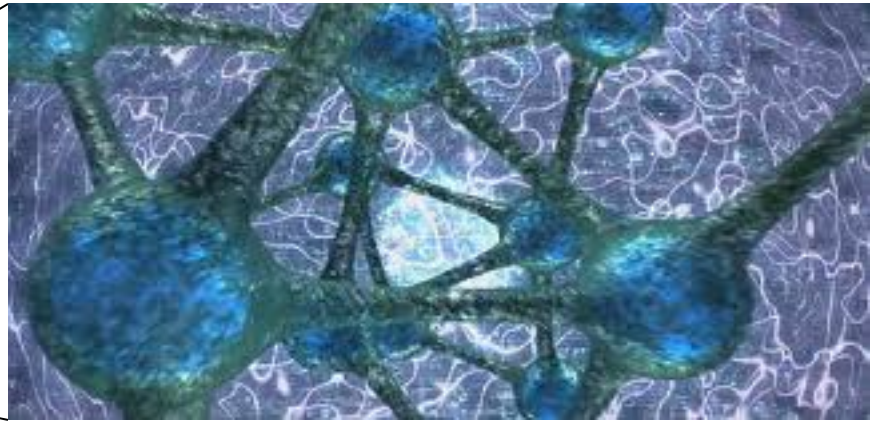




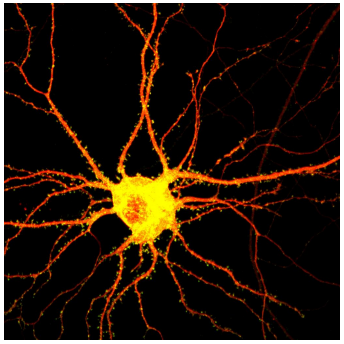
Todd Hylton 2008

Brain Architecture

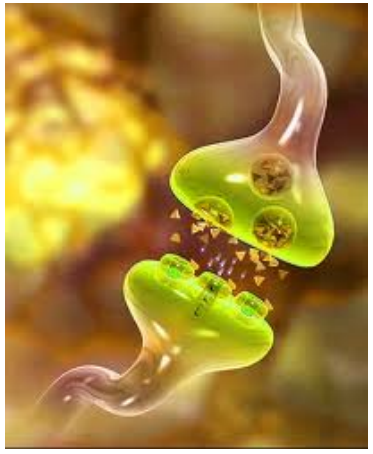
Dense Network



Neurons

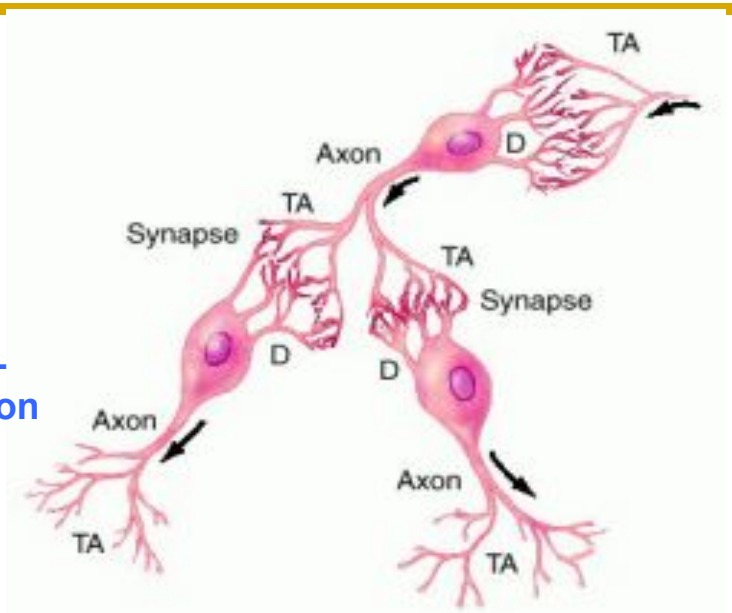


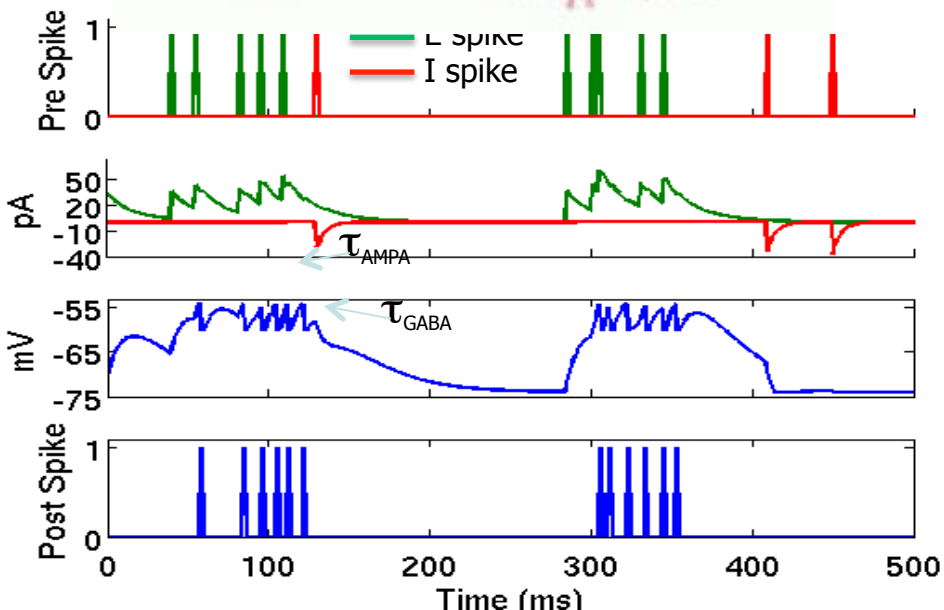
Synapses



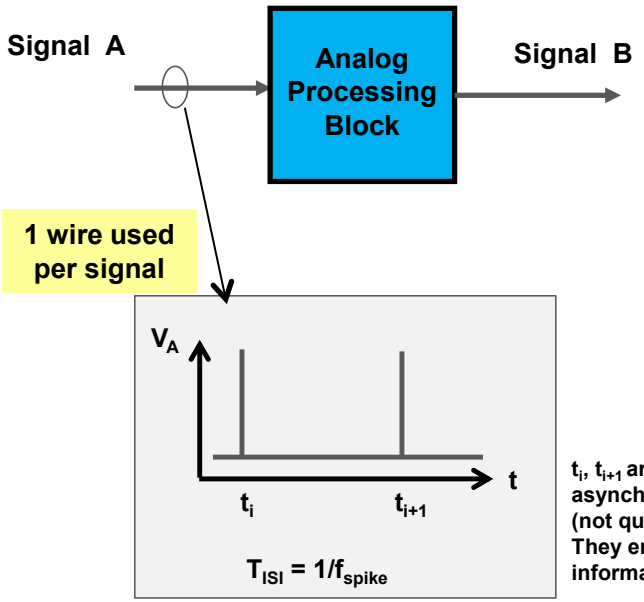
Brain is composed of 10^{11} neural cells with 10^{15} synapses:
Very High Density (10^{10} synapses/cm^2) and Connectivity (1:10^4)

Architecture Dynamics: Leaky Integrate and Fire Neuron



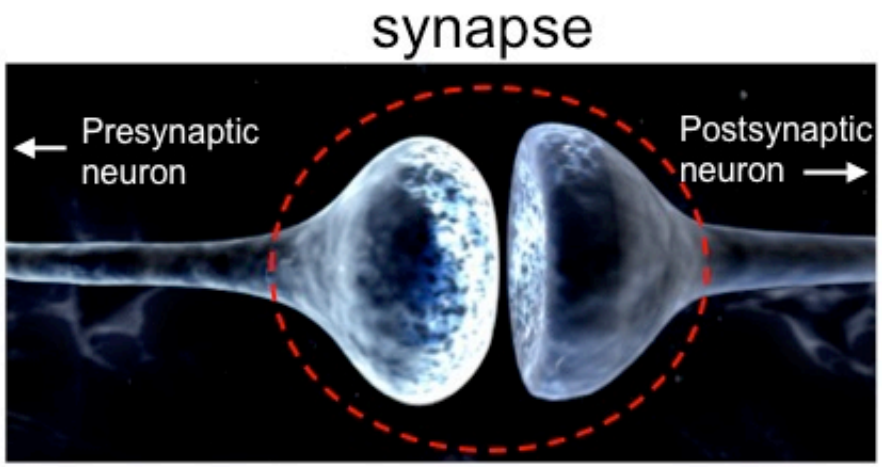


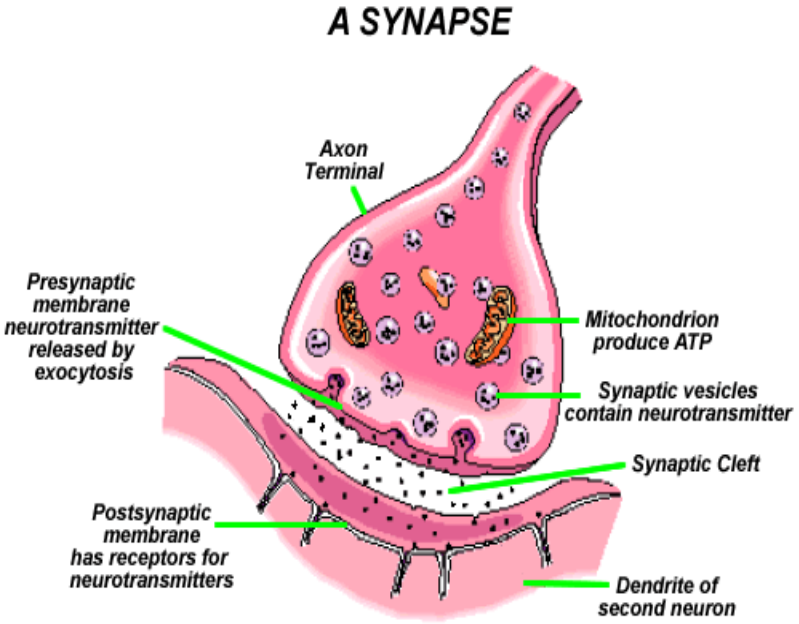
Analog Spiking (Mixed Signal)

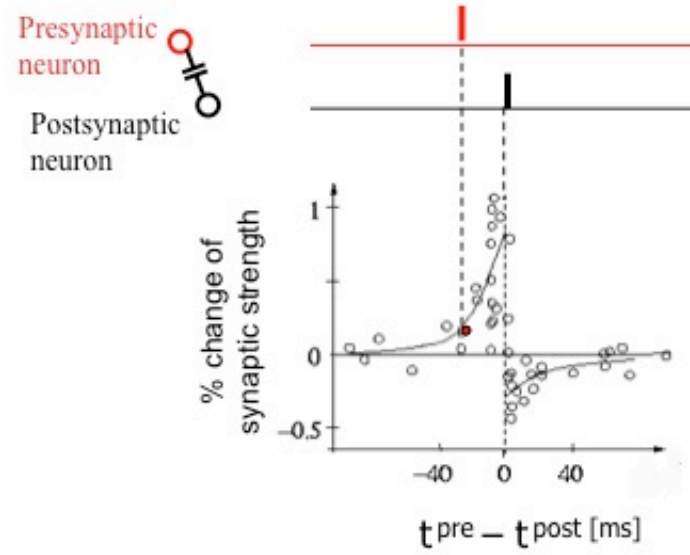


- Single wire used to represent spike signals which encode analog information
- Dissipate power only during spike events
- Spiking system **less prone to noise and variations (only needs to maintain timing information)**
- Cascaded spiking analog processing blocks is **less prone to noise accumulation due to spikes combined** with learning and adaptation

Architecture Dynamics: Synaptic Plasticity







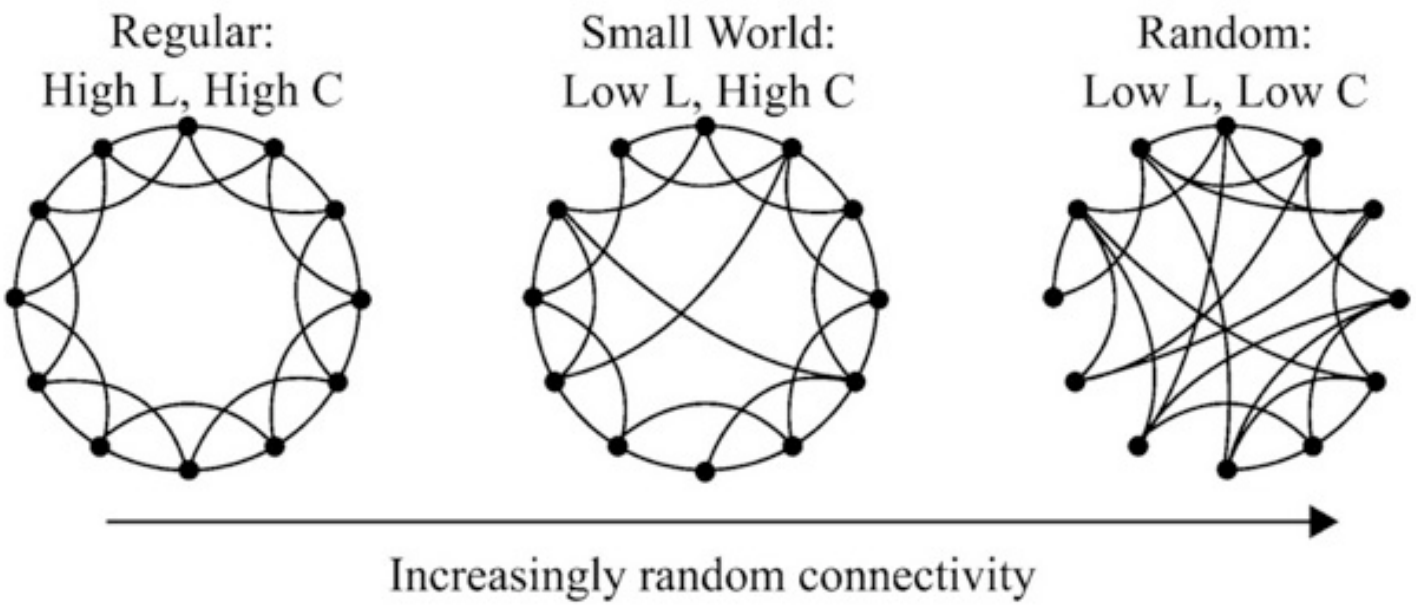
Electrical ➔ Chemical ➔ Electrical

Speed, Specificity, Timing

Spike Timing Dependent Plasticity (STDP)

(Markram et. al 1997; Bi and Poo, 1998)

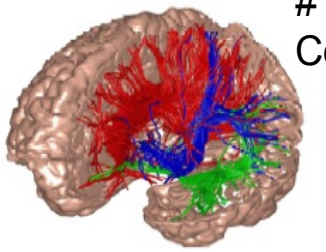
Architecture Design: Small World Connectivity



Strogatz 2000; Sporns, 2004)

- Cortex (> 85% of the brain) is organized as a small world network of neurons
- Dense local connections and sparse long range connections
- The typical distance or synaptic path length *L* between two randomly chosen neurons grows as $L \propto N$ where *N* is the number of neurons in network
- Efficient communication despite network complexity – needed for survival

Large Scale System (Analog Core)



Brain Architecture

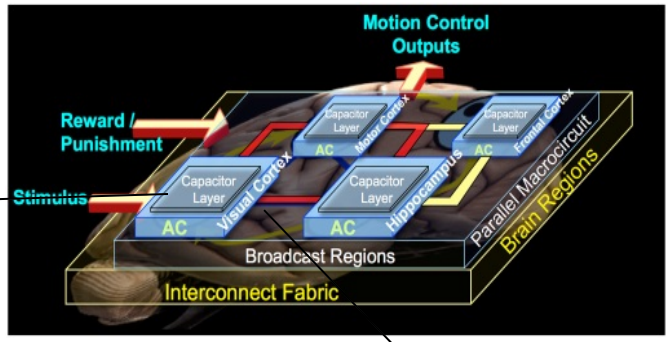
Neurons,
Synapses,
Connectivity

Neuromorphic Compiler

Routing, Neuron Placement

Digital Memory

Programmable Front-End



Set switch states

Acquire Switch states

Analog Core with Cortical Fabric (neurons, synapses)

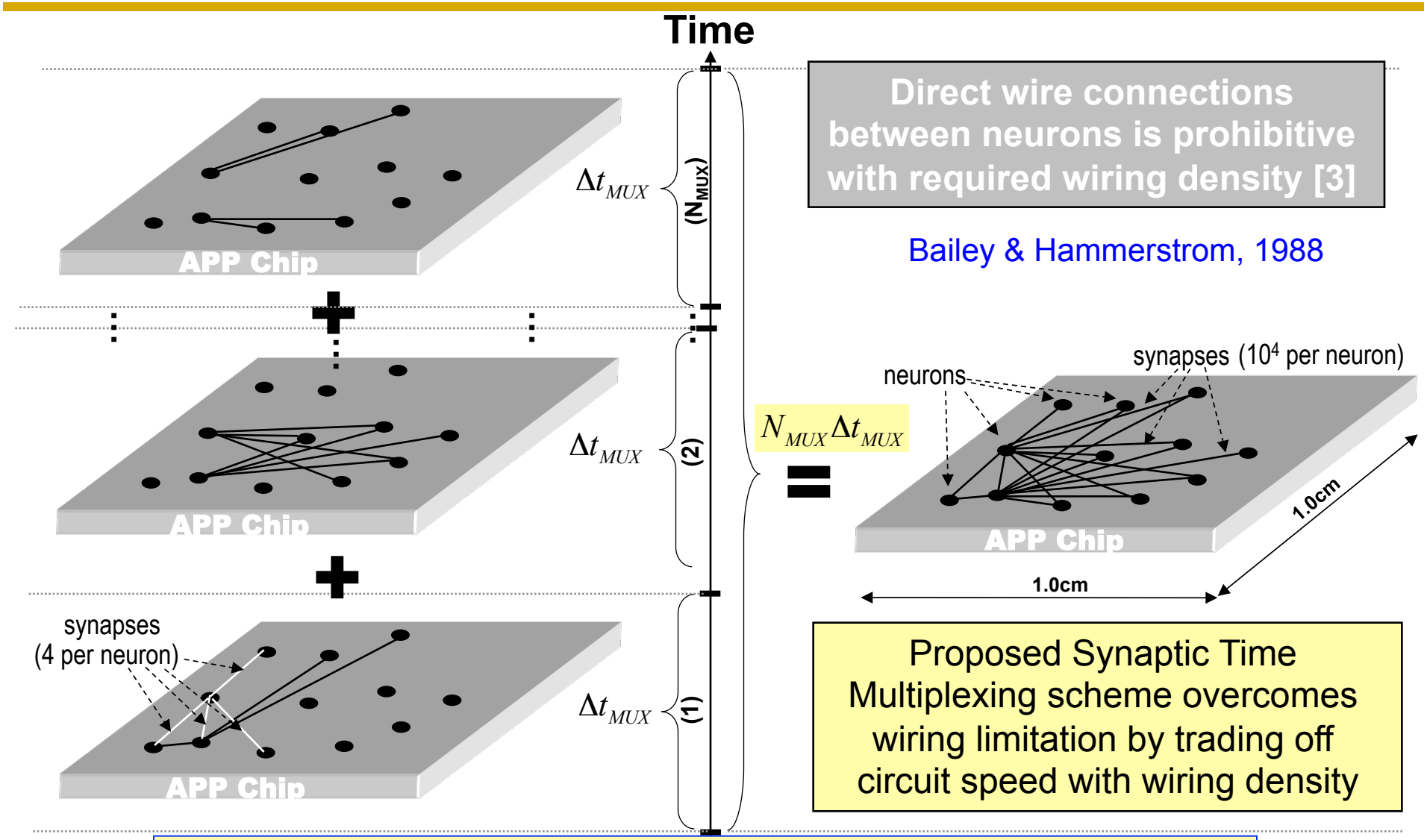
Store

Retrieve

Analog Memory (store synaptic conductances)

Overall Design Goal: 10^6 neurons and 10^{10} synapses in cm^2 consuming 1 W of power

Synaptic Time Multiplexing (STM)



Direct wire connections between neurons is prohibitive with required wiring density [3]

Bailey & Hammerstrom, 1988

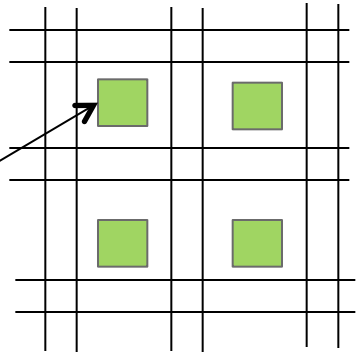
Proposed Synaptic Time Multiplexing scheme overcomes wiring limitation by trading off circuit speed with wiring density

Scalable solution to enable CMOS based neuromorphic chip design

Reconfigurable Fabric vs. Crossbar

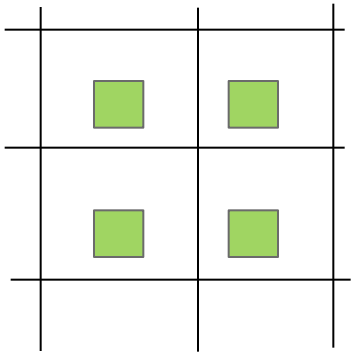
Reconfigurable Fabrics

Time multiplexed Fabric (HRL)



Advantages
- Flexible topology
- High effective density (Wires reused for different axons)

Broadcasting (HRL)

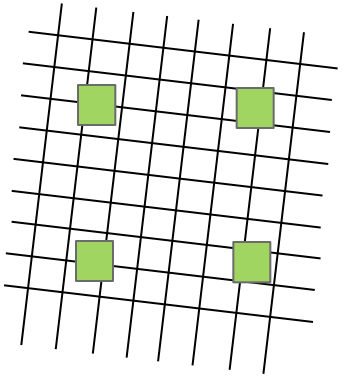


Advantages
- Flexible topology
- High effective density (Wires reused for different axons)

Limitations
- High multiplexing ratio needed for large networks

Fixed Fabrics

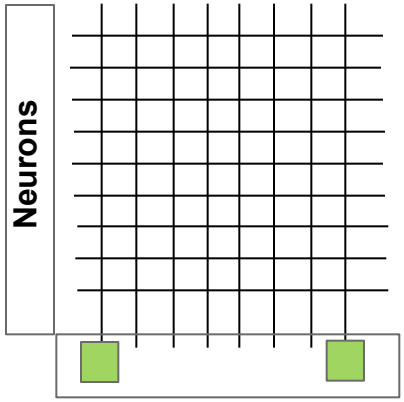
Crossbar (SUNY)



Advantages
- No multiplexing simplifies synapse design

Limitations
- Fixed topology
- Synapse density limited by wiring (axons not multiplexed)

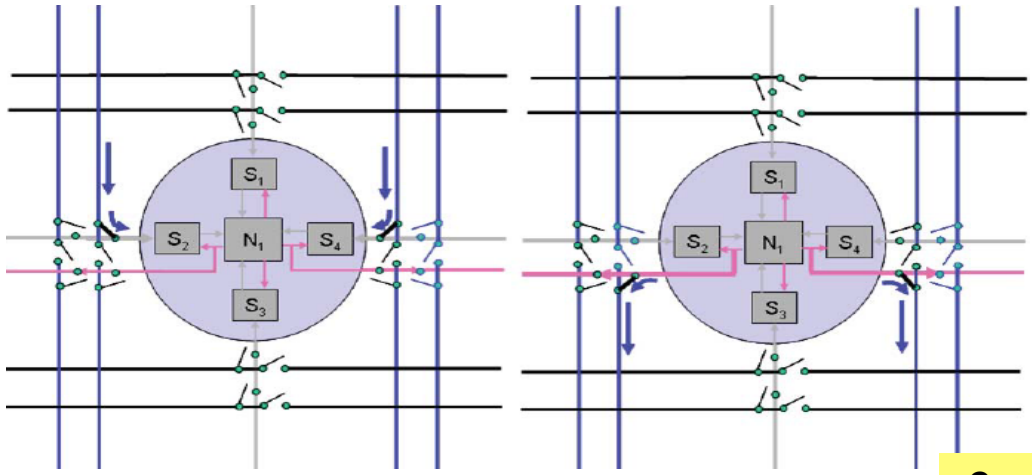
Synapse in 2D array. Neurons in 1D arrays (HP, IBM)



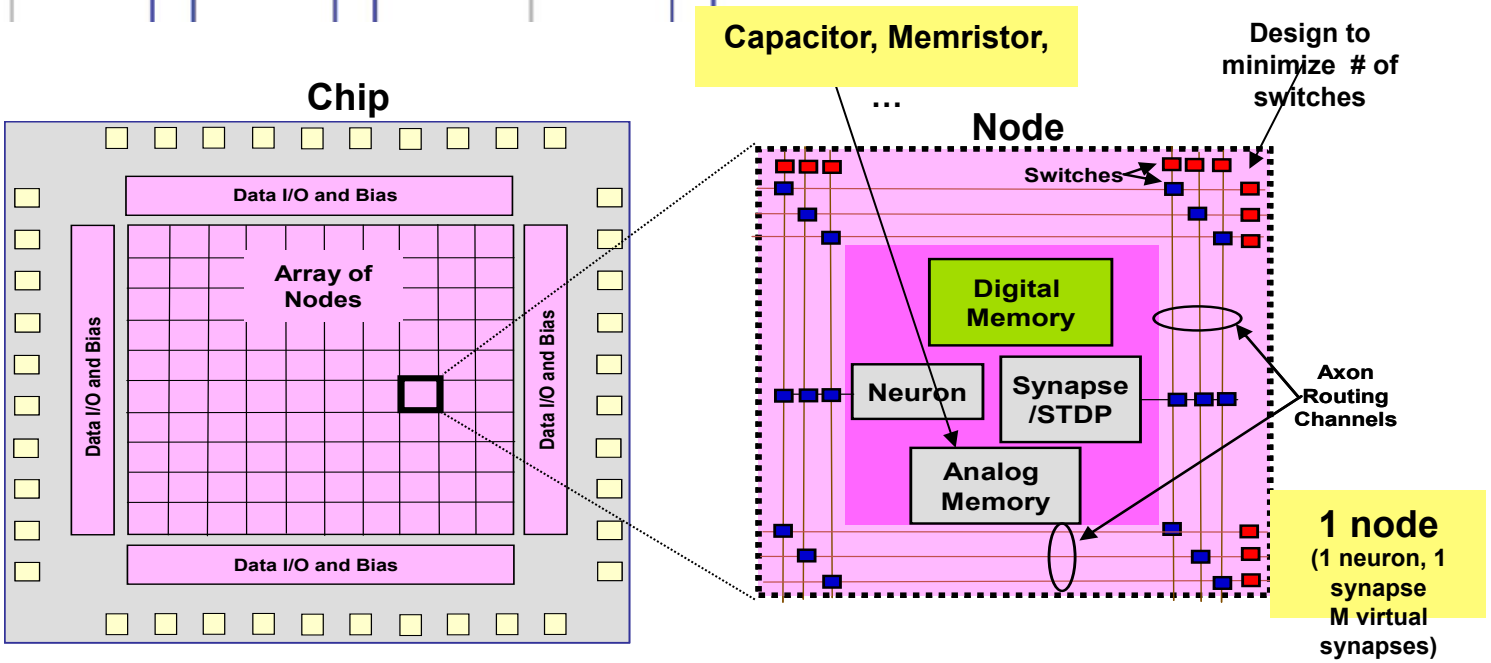
Advantages
- No multiplexing simplifies synapse design

Limitations
- Fixed topology
- Number of neurons scale less than linearly with chip area
- Synapse density limited by wiring

STM Fabric & Analog Core Chip Architecture



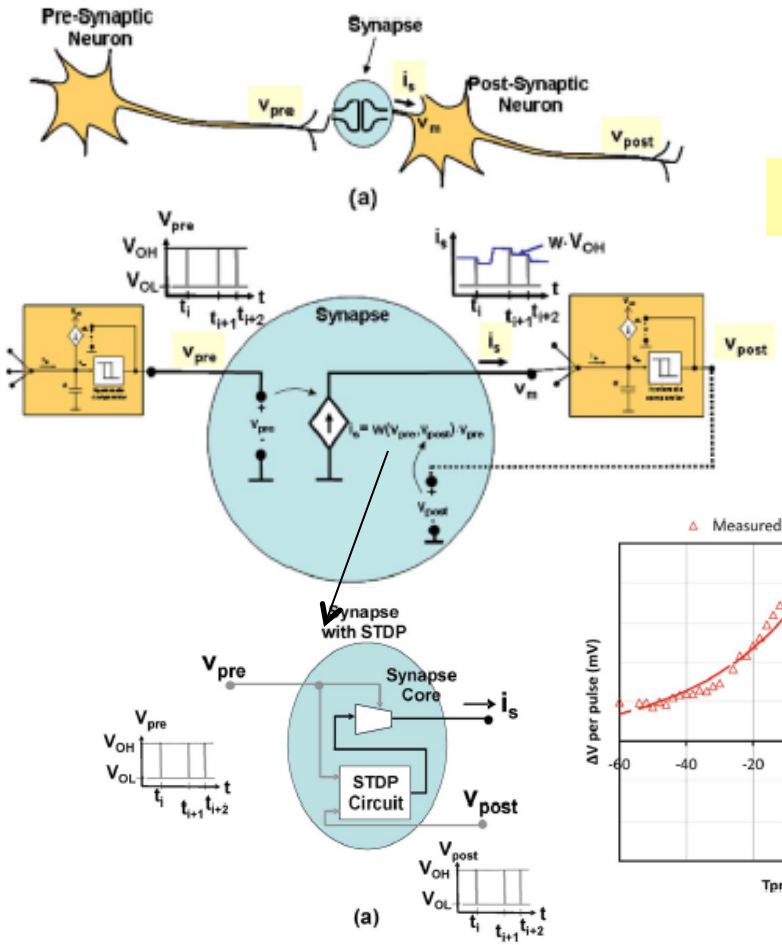
K. Minkovich, N. Srinivasa, J. M. Cruz-Albrecht, Y. K. Cho and A. Nogin, "Programming Time-Multiplexed Reconfigurable Hardware Using a Scalable Neuromorphic Compiler," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 889-901, June 2012.



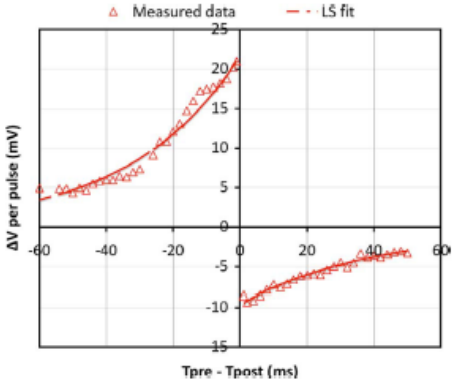
Time-multiplexing ensures scalability of hardware using conventional CMOS technology

HRL SyNAPSE Fabricated Phase 0 Hardware Base Components

Synapse with STDP

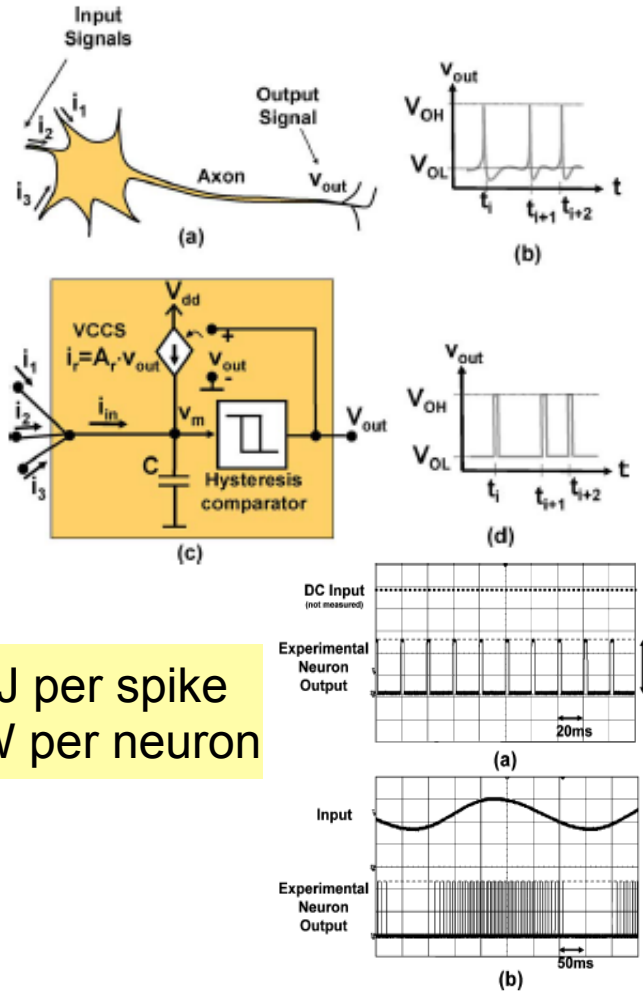


90nm CMOS



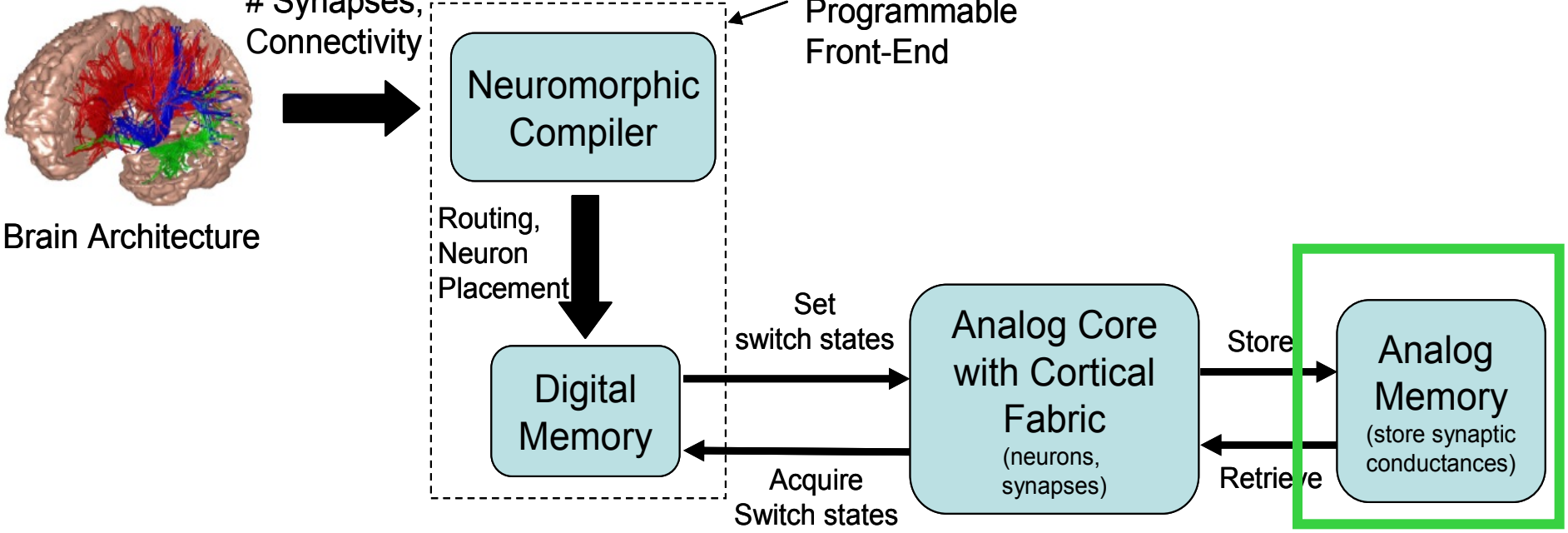
0.4pJ per spike
10nW per neuron

Integrate & Fire Neuron



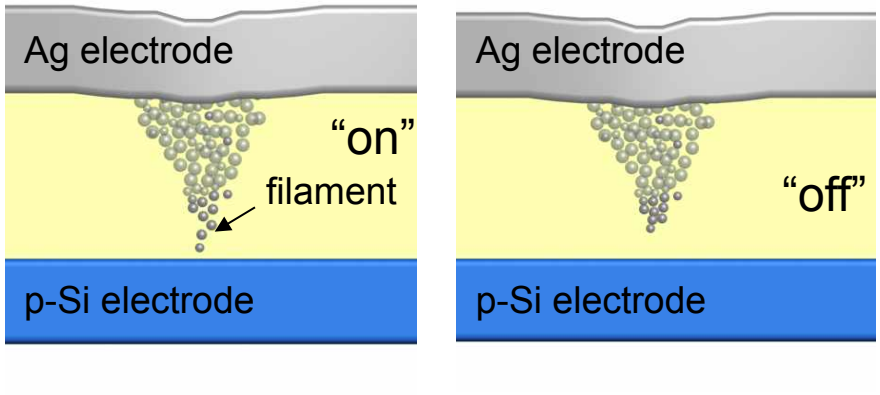
Jose Cruz-Albrecht, Michael Yung, Narayan Srinivasa, "Energy-Efficient, Neuron, Synapse and STDP Integrated Circuits, " *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6. No. 3, pp. 246-256, June, 2012.

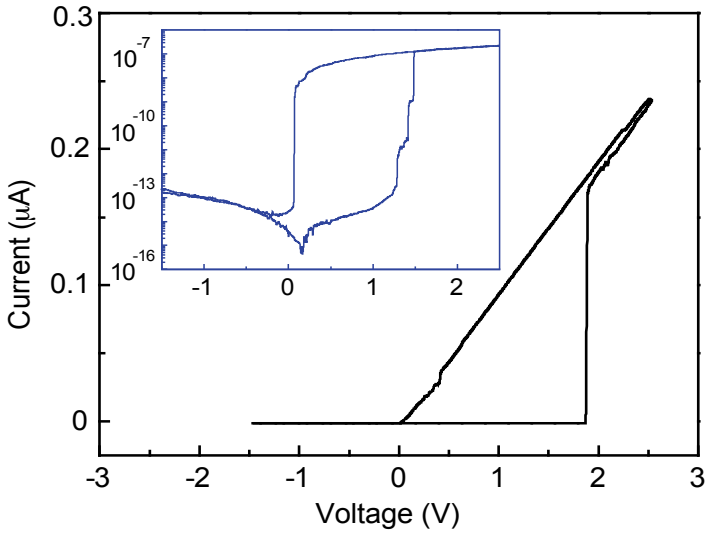
Large Scale System (Analog Memory)



Absolute vs. Incremental Memristors

Abrupt Resistance Switching



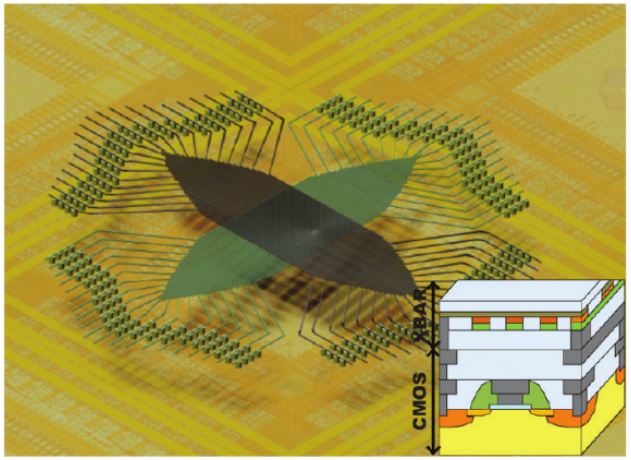


- Two terminal resistance switching device
- Nanoscale a-Si switching area
- Small cell size, < 50 nm x 50 nm (density > $10^{10}/cm^2$)
- 3.5 bits or 10 levels of storage per device
- Endurance $3*10^8$ cycles and retention is for months
- CMOS compatible materials and processes

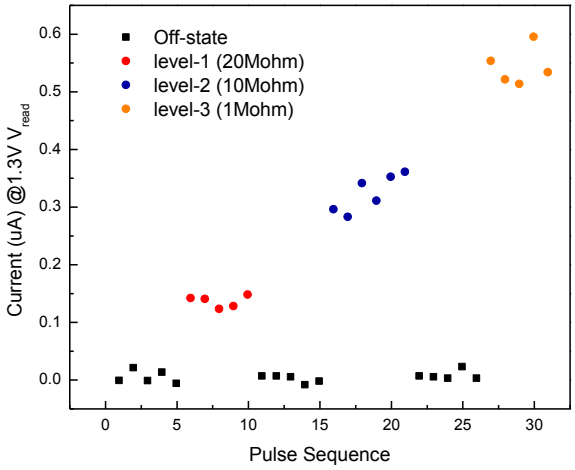
Developed CMOS compatible memristors to enable memristor array fabrication

Functional Memristor Array with CMOS Integration

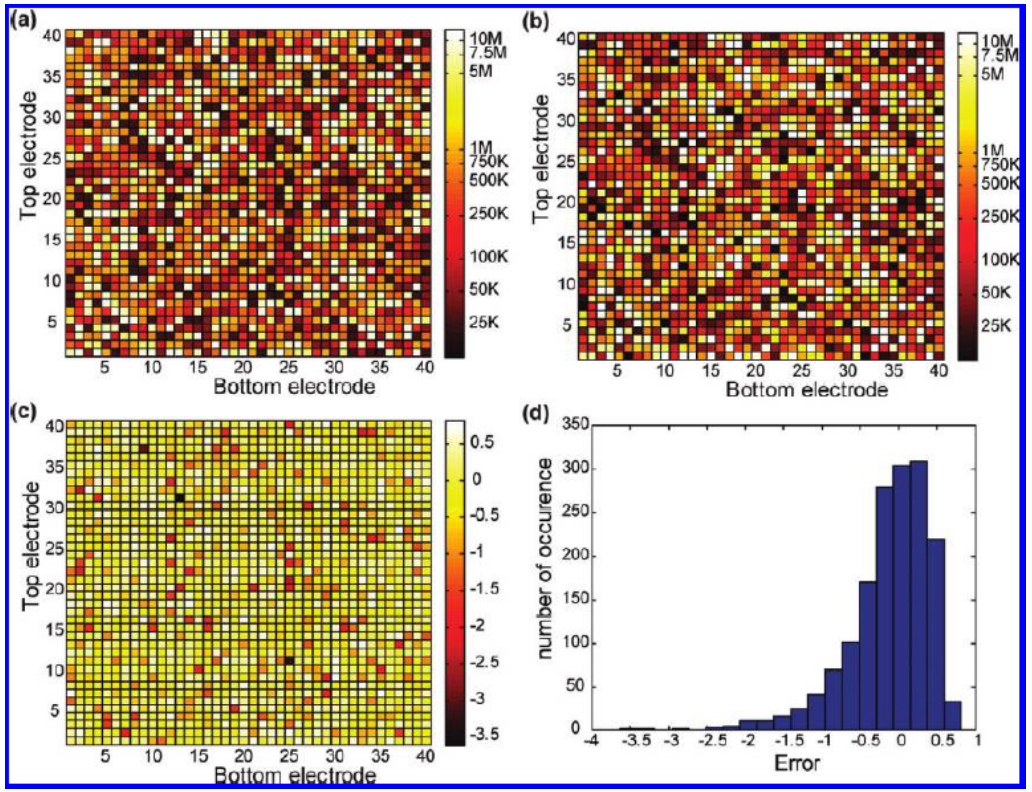
CMOS circuit with memristor



Multibit values written on memristor device within integrated chip

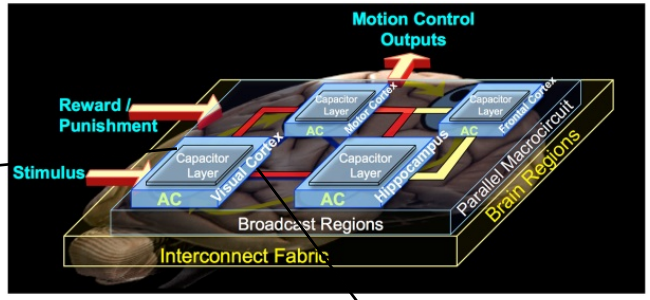


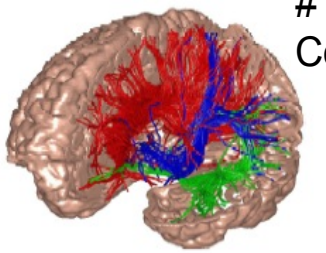
Data written on memristor array (40x40)



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.

Large Scale System (Neuromorphic Compiler)





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

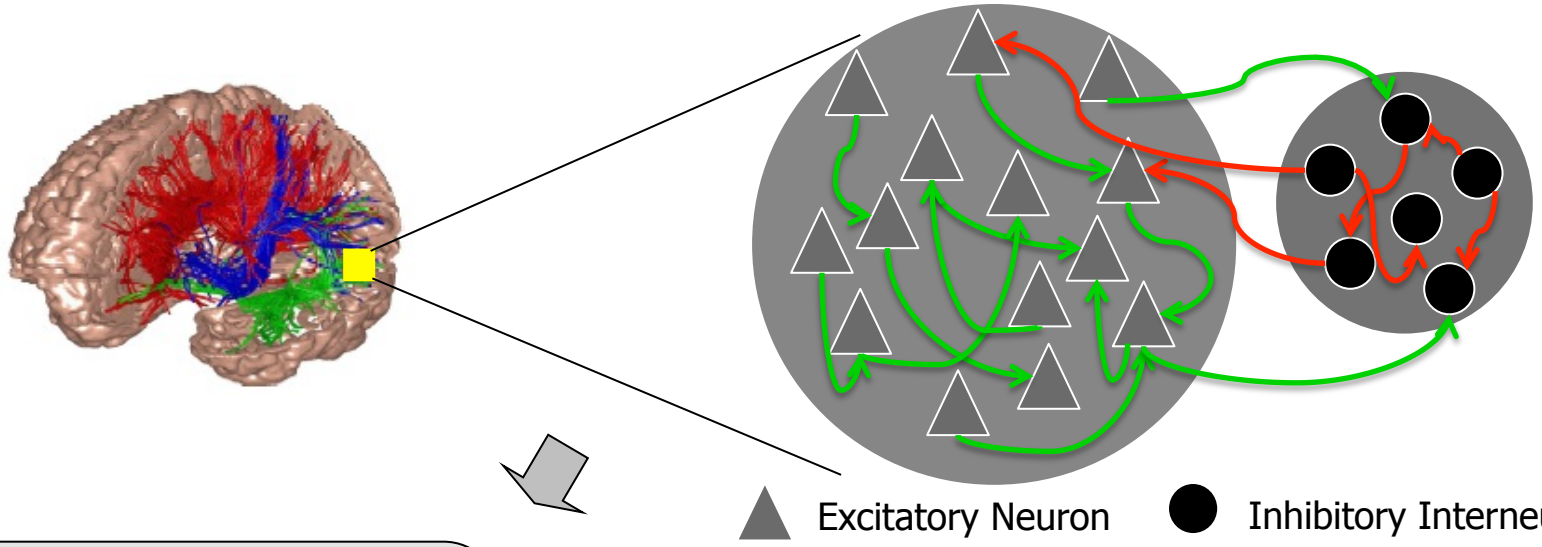
Analog Core with Cortical Fabric
(neurons, synapses)

Store

Retrieve

Analog Memory
(store synaptic conductances)

Scalable Neuromorphic Compiler



$$C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \end{bmatrix}$$

Connectivity Matrix (Neuron A connects to B, D, F etc)

Placement Algorithm → Routing Algorithm → Switch states for TMF across allotted time-multiplexing steps

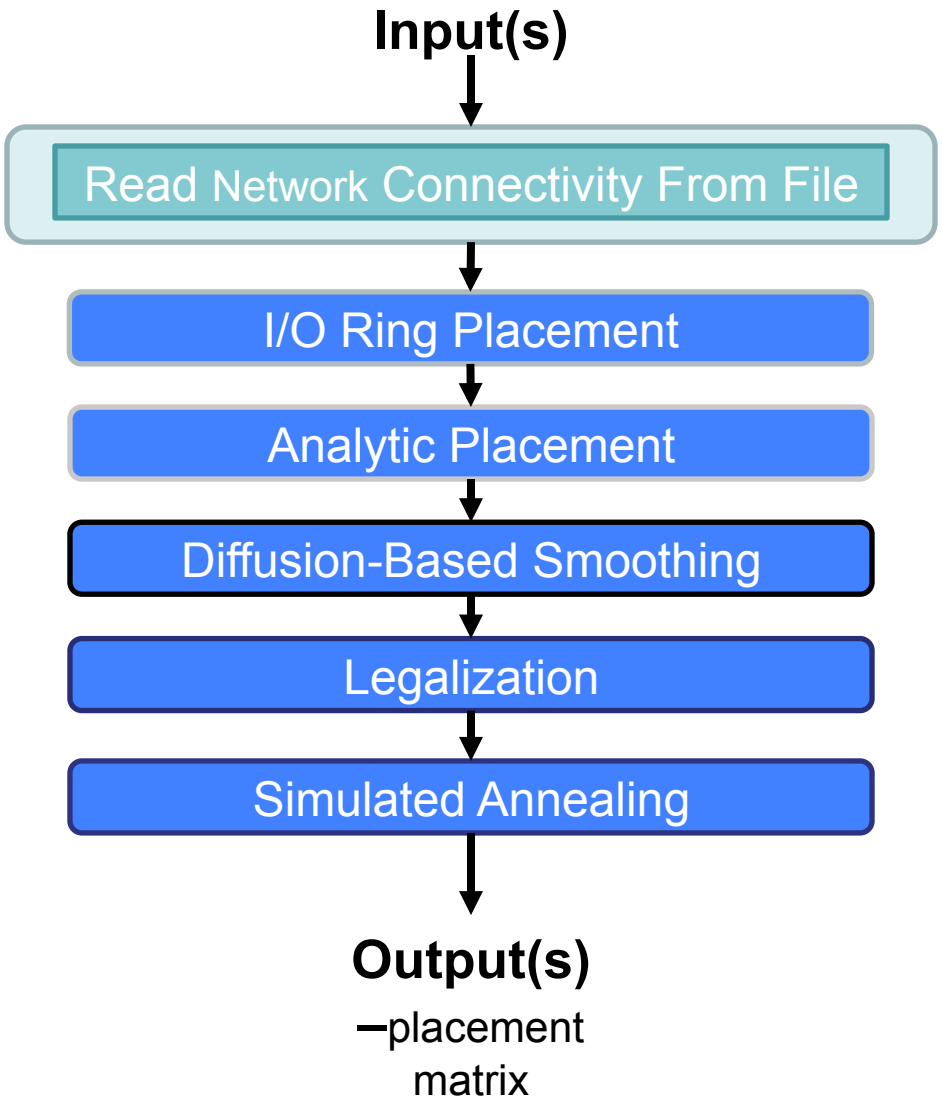
K. Minkovich, N. Srinivasa, J. M. Cruz-Albrecht, Y. K. Cho and A. Nogin, "Programming Time-Multiplexed Reconfigurable Hardware Using a Scalable Neuromorphic Compiler," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 889-901, June 2012.

Enables rapid and efficient translation of microcircuits into time-multiplexed hardware

Placement: Overview

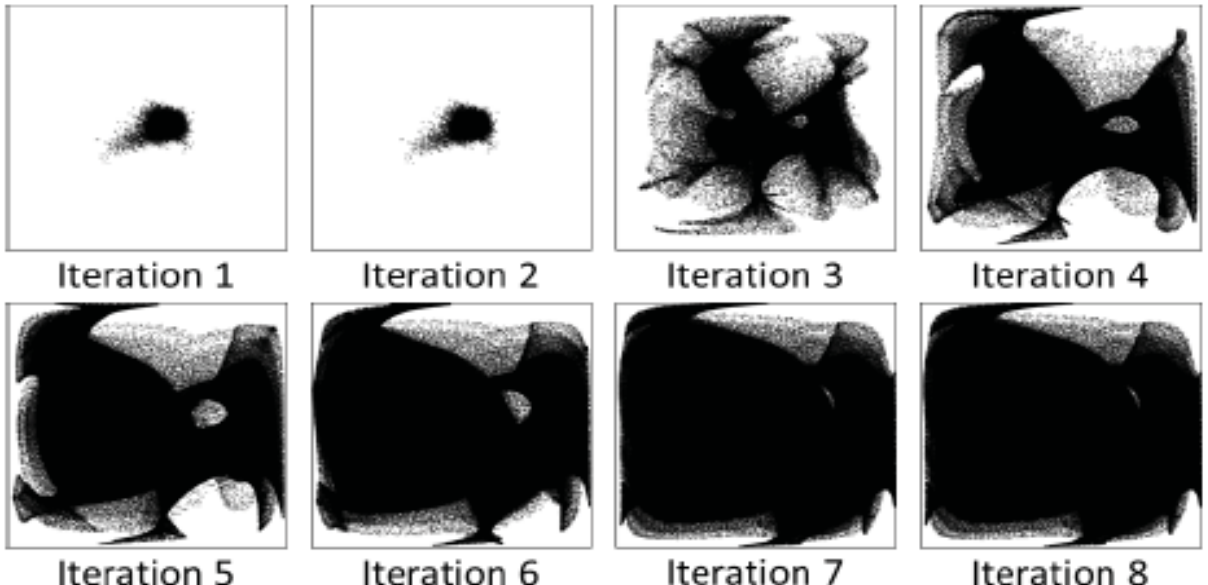
Purpose: Assign network neurons to physical hardware nodes

Goal: Minimize congestion and allow for evenly distributed synaptic communication



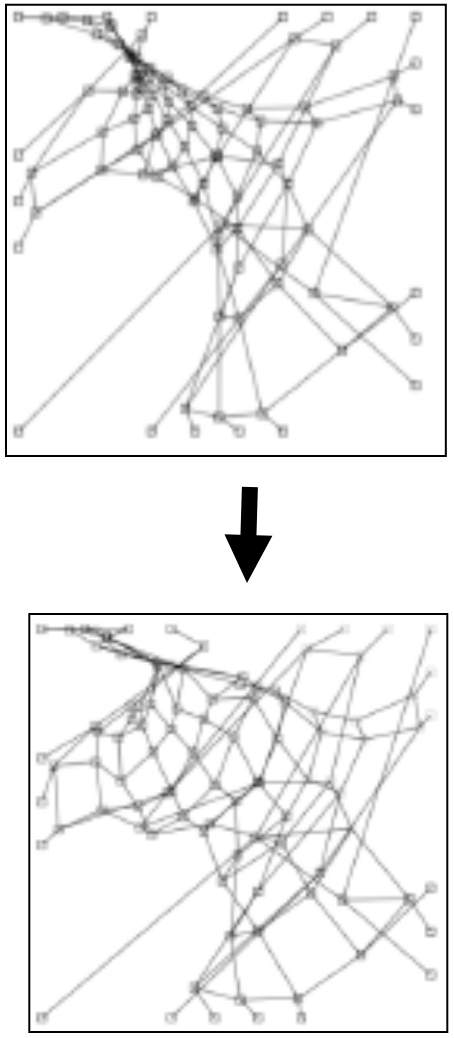
Analytic Placement

- Generates initial placement solution iteratively
- Quadratic wire-length minimization problem
  - Synaptic pathways ➔ springs
  - Neurons ➔ connection points
  - Minimizes total potential energy of springs (quadratic function of length)
- Converts one-to-many synaptic pathways into pair-wise springs based on neural star model
- Average synaptic path length sees 3X reduction – directly correlates to reduction in required STM timeslots



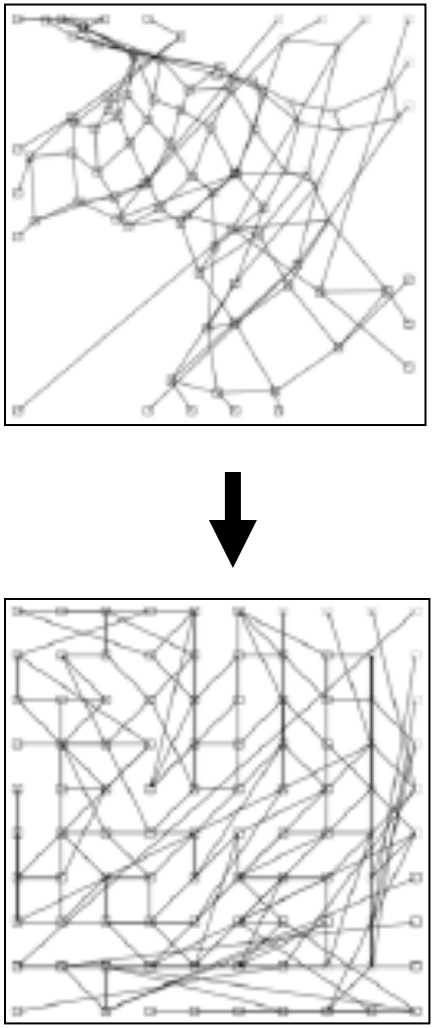
Diffusion-Based Smoothing

- Aims to smooth out densely-connected clusters of initial placement solution
- Adds forces based on density of layout and iteratively spreads out placement
- Neurons "migrate" to final equilibrium positions using velocity functions based on local density gradient



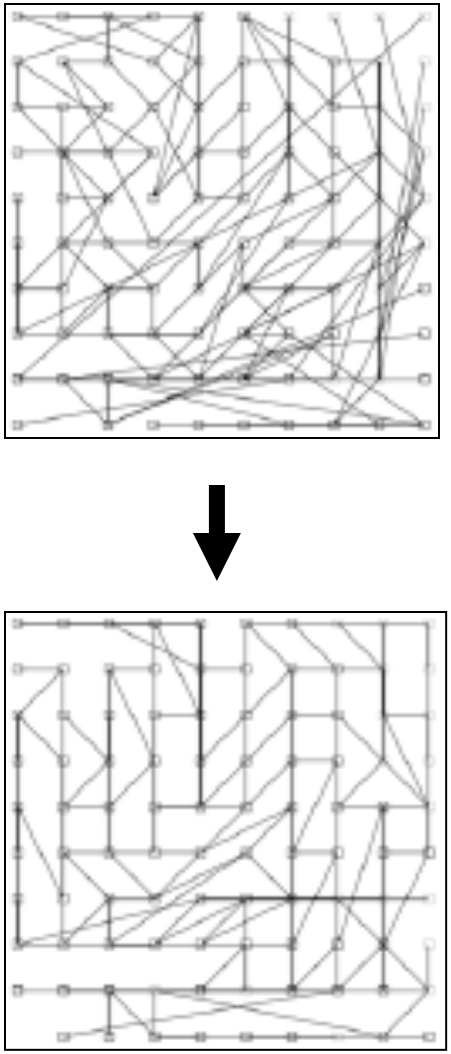
Legalization

- Assigns neurons to actual grid-based locations
- Ensures all neurons are placed and no node contains more than 1 neuron
- Sorts nodes by connectivity and pushes neurons outward in spiral pattern onto unoccupied nodes

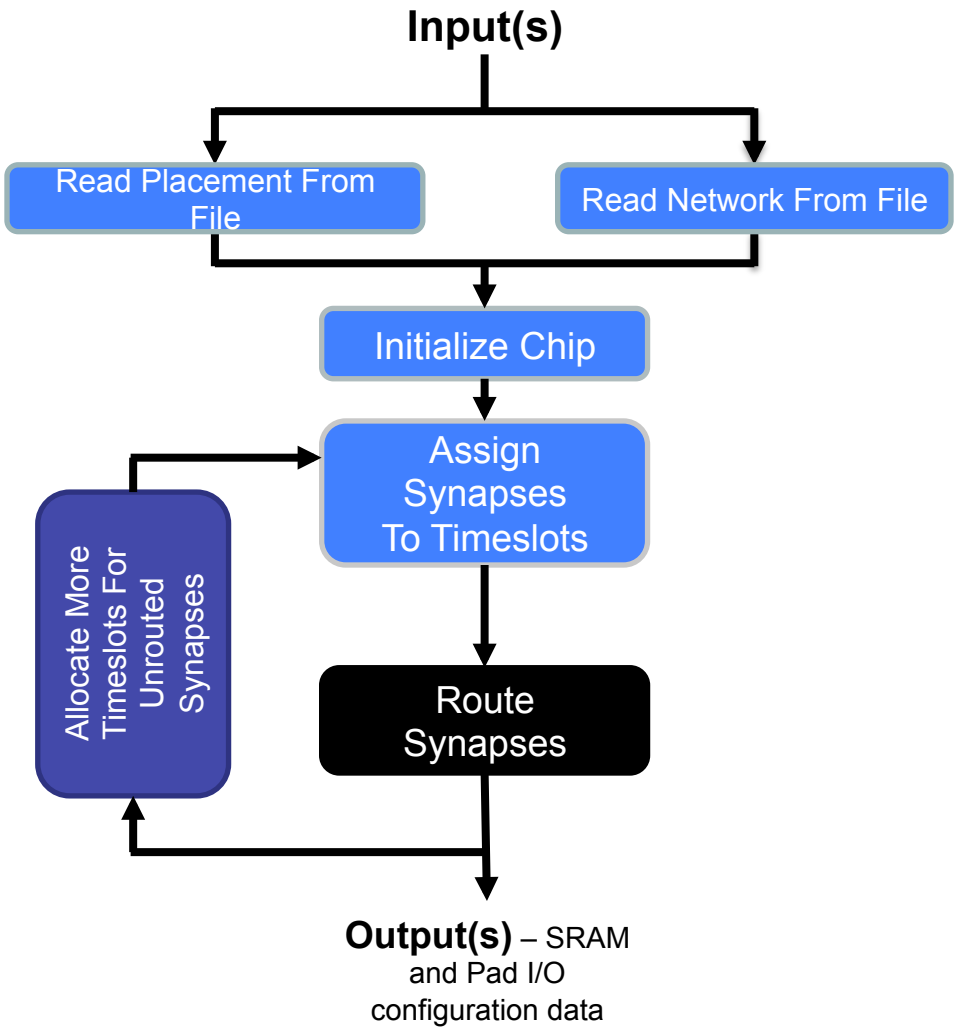


Simulated Annealing

- Aims to further reduce grid wire-length after legalization
- Attempts to move neurons to their "ideal" locations via chain of relocations
- When chain intersects itself, series of relocations is guaranteed to reduce grid wire-length



Routing: Overview



Cost of path:

Manhattan Distance

Number of switches required

Timeslot Assignment

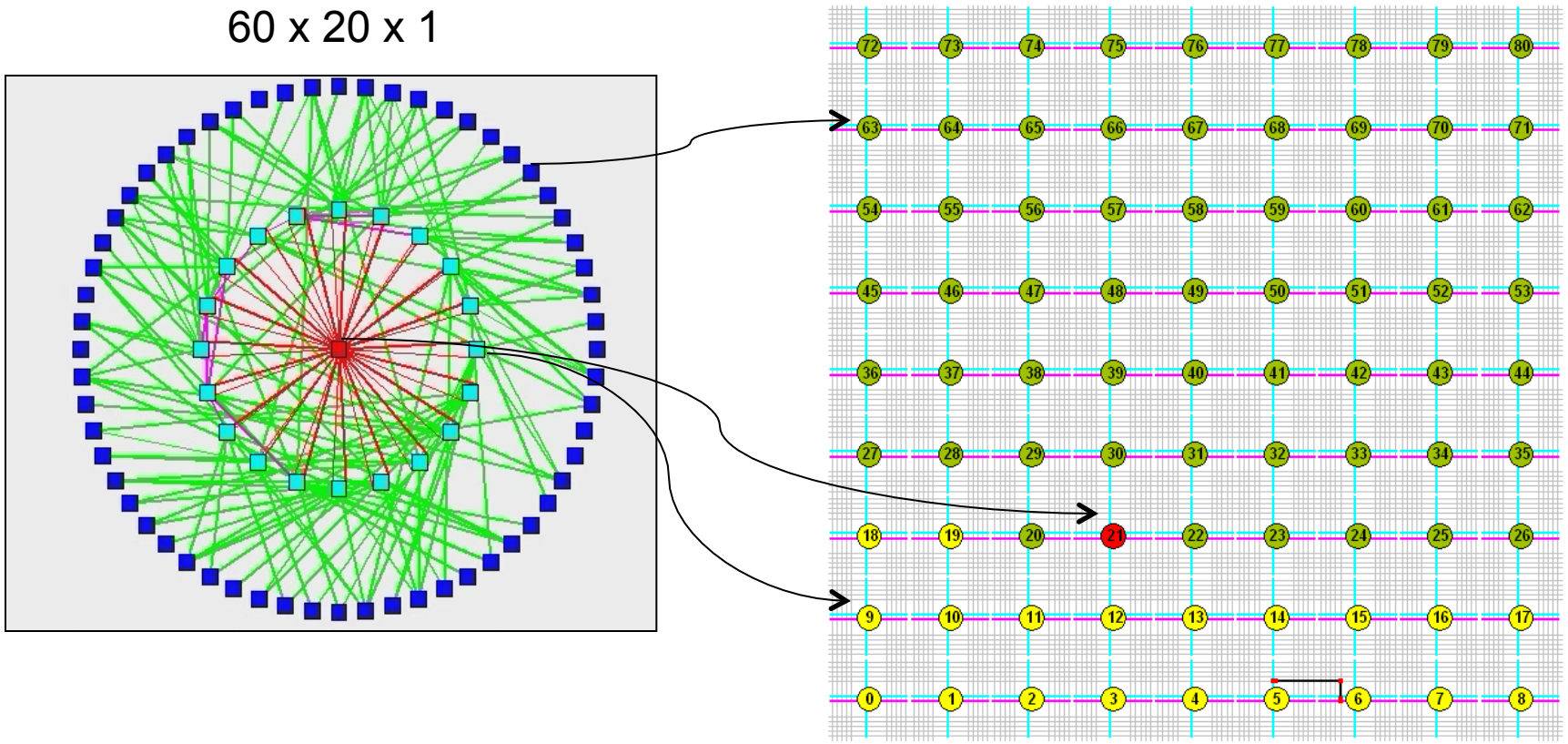
- Determine minimum number of timeslots required based on fan-in/fan-out restrictions
- Sort synapses in increasing order by Manhattan distance, pre-synaptic neuron, and post-synaptic neuron
- Assign synapses in round-robin fashion
- When synapse is assigned to given timeslot, assign other synapses with same pre-synaptic neuron and within range of same Manhattan Distance within same timeslot

Synaptic Routing

- For each timeslot:
  - Group assigned synapses by pre-synaptic neuron
  - Loop over all available gridlines
  - For each gridline, try routing as many unrouted synapses as possible
- To route a given synapse:
  - Use A-star based search
  - Minimize cost of path

Example of Compilation

60 x 20 x 1



Capable of compiling 1M neurons and 10B synapses in about 5 minutes

Summary

- Hybrid Mixed Signal Circuit architecture design (discrete signal and continuous time)
  - Analog for neural and synaptic computation
  - Digital for spike transmission
  - Low power, small footprint (1 M neurons and 10 B synapses in cm^2 using 1 W)

- Flexible Connectivity
  - Programmable STM fabric with compiler enables scalable arbitrary connectivity

- Scalable Design
  - Modular arrangement of nodes enable rapid scaling with CMOS technology

- Currently porting several spiking models on to chip for verifying functional performance

Challenges

- Absence of analog tools for rapid chip design, verification and debugging makes it impossible to scale rapidly
- Multichip implementation is necessary to scale to mammalian levels – however current interconnect methods such as AER are error prone and power hungry – maybe 3D CMOS architectures plus other interconnect designs will help here
- So far we have only considered plasticity in the form of *reweighting* the synapses
  - *reconnection*, *rewiring* and *regeneration* – currently no solution available
- Showing emergent behavior via learning and w/o programming is key for useful applications – slowly making inroads here but still will be limited due to above