

Neuro-Inspired Computation : Beyond Stored Program Architecture and Moore's Law Limits

Murat Okandan

February 25, 2013

Beyond von Neumann/Turing and Moore's Law

Inspiration



- Drivers :
- End of physical device scaling
 - Power limitations
 - Applications with large, incomplete, noisy ("natural") data sets – big data

Current approaches

Probabilistic -Bayesian Computation
Novel Architectures



Probability processing



Probabilistic computing



Intelligent computing



Asynchronous high speed PLD



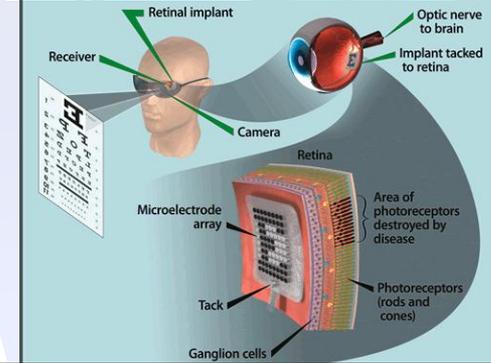
CMOS integrated optical comm.



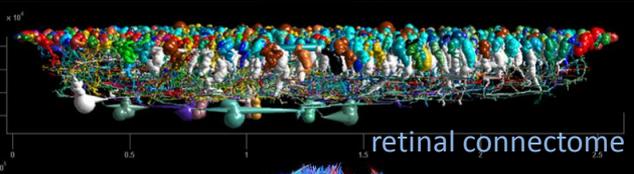
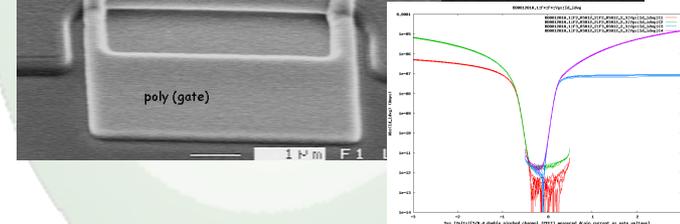
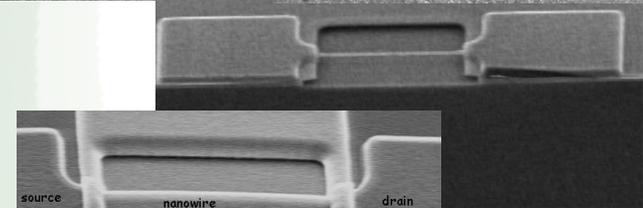
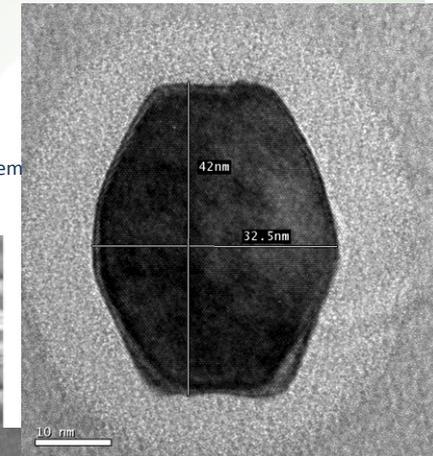
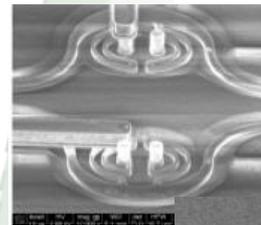
parallel processing

Future

Emulating pattern recognition, abstraction and prediction model of neural systems – from device physics up to system level



computation not with "1s-and-0s" but 1s-0s in and around the system to provide coupling to conventional systems



retinal connectome



human connectome

Overview

Problem

*Conventional computational approaches (von Neumann architecture) have reached physical limits (end of Moore's Law, power dissipation limits, programmability) –
and cannot address the highest impact problems.*

Challenge

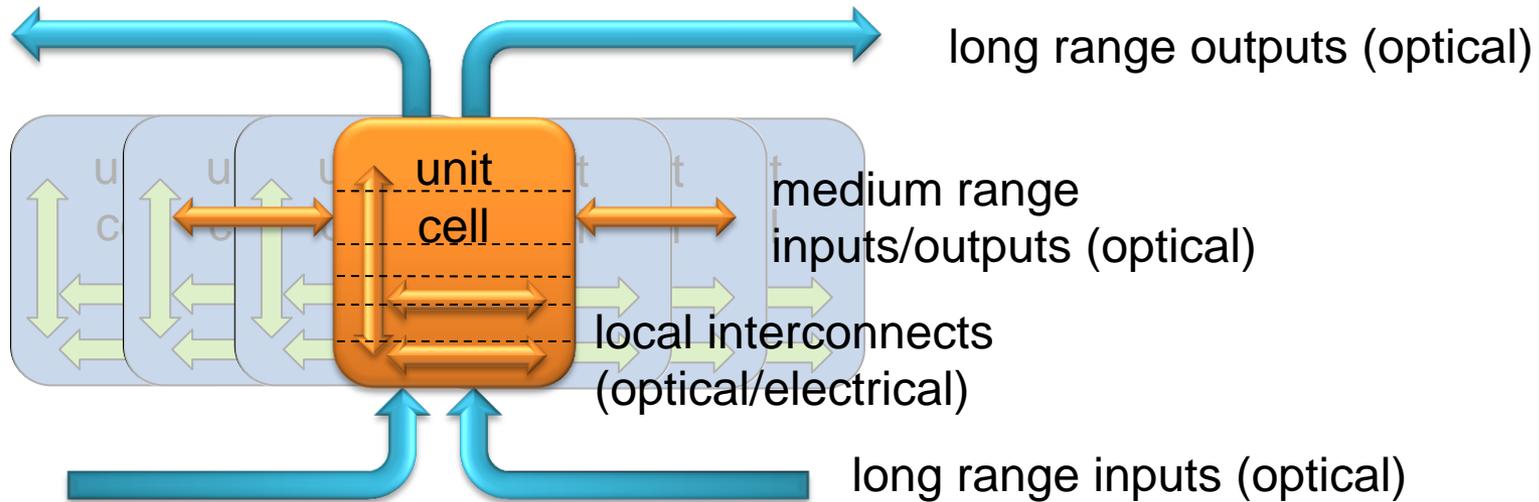
How do we figure out what is next?

What impact does this development have on our mission space?

Critical impact:

- 1) First/earliest users: game changing technology for national security applications
- 2) Industrial paradigm shift – microelectronics: \$ 300B/yr., >>\$1T on top
- 3) Game theory – theory of mind, how individuals and societies interact

Devices and Systems : Future



“cortical column” - hierarchical, temporal memory
(On Intelligence, Jeff Hawkins)

3D hybrid integration – opto-electronics, TSV, novel devices, ...

key characteristics:

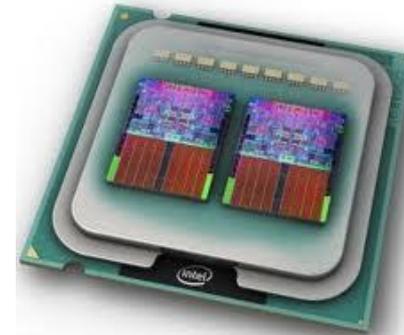
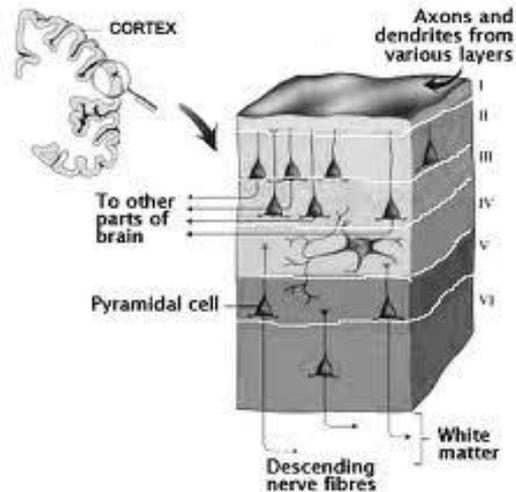
- Plasticity/adaptability at native (device) level functionality
- Massive interconnect/fanout at system level

Neural Computation as Inspiration for Next Computational Architecture

- not a von Neumann architecture system
- Pattern recognition – abstraction – prediction – adjustment
 - dynamic
 - multi-dimensional
 - difference engine + accumulator with goals: *
acquire sustenance, avoid predators, reproduce
- How is the data represented, stored and processed?
 - still an open question
 - multiple mechanisms and time scales

* not necessarily the optimum solution, it worked – and it is conserved

Neo-Cortex – CPU/GPU Comparison



~75% of human brain volume

2500 cm² x 0.4 cm thick

(large dinner napkin, 1/6 inch thick)

10 billion neurons (10^{10})

massively interconnected, (10^4 synapses per neuron)

10-100 trillion synapses (10^{13} - 10^{14})

dynamic, fault tolerant, low power (~ 25W for system)

~10ms per unit cell, ~100ms for perception

(photons in → object recognition)

3-7 cm² x 0.2 cm thick

(~10um active thickness, transistors + metal stack)

1-4 billion transistors (10^9)

average fan-out of

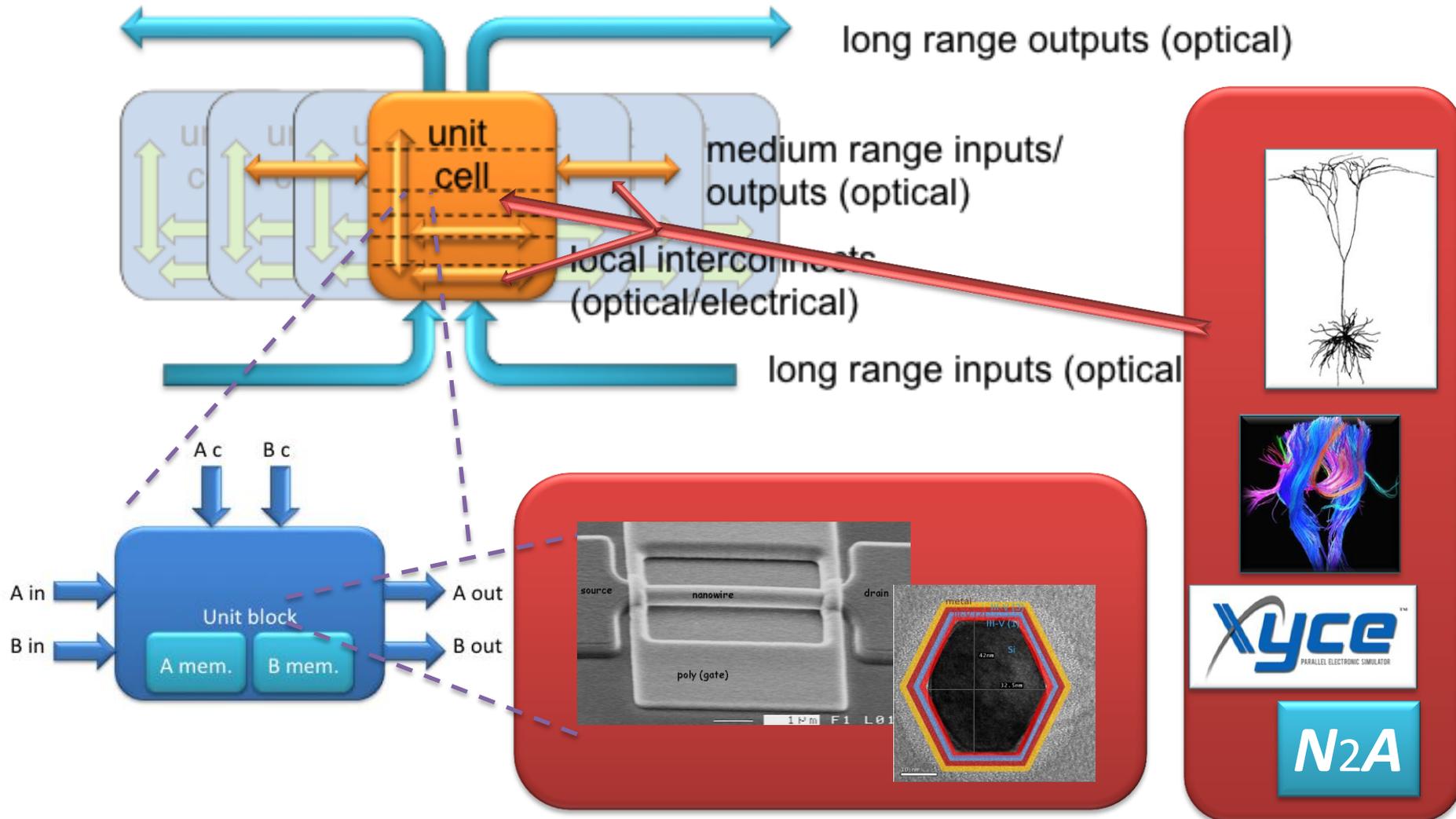
2-4 per transistor

40-100W (per chip)

2-4 GHz

What?

How?



Devices and Systems : Now

Current Technology: Probabilistic/Bayesian Computation

- Benefits :
- 1/10th power
 - 1/10th to 1/30th chip size

The screenshot shows a web browser window displaying the Lyric Semiconductor website. The page title is "Lyric Semiconductor | Technology: Gates" and the URL is "http://www.lyricsemiconductor.com/technology-gates.htm". The navigation menu includes "Background", "Technology", "Company", "Contact", and "News". The main content area is titled "Gates: The fundamental building blocks". It features a comparison between a "Traditional LDPC primitive in TSMC 65 nm" and "Lyric's LDPC primitive in TSMC 65 nm". The traditional primitive is shown as a large, complex circuit diagram with a "scale: 20um" dimension. Lyric's primitive is shown as a much smaller, simpler circuit diagram. A text box states: "Smaller and lower power than a traditional implementation using 15nm CMOS". Below the comparison, there is a paragraph explaining that at the most fundamental level, computers are an assembly of gates used to perform basic operations. For problems in the probability domain, these gates must determine the probability that a bit is a 1 or 0. The final paragraph states that Lyric's gates are designed to model relationships between probabilities natively in the device physics, allowing for mathematical operations in the probability domain with just a handful of transistors, creating power and area savings of more than 10X over traditional implementations.

Applications: error correction, communication systems, data compression, ...

Devices and Systems : Now

The Future of Parallel Processing!

CogniBlox is a **stackable module** allowing versatile, fully parallel, pattern recognition architectures for data mining, high-performance cognitive computing, sensor fusion, high-speed video analytics, hyper spectral images, genomics and more. The common denominator of any configuration of CogniBlox is that for a given chain of cognitive memory processors (of any size), the recognition of a vector submitted to this chain will take only 10 microseconds.



**If you need to match 1 pattern (up to 256 bytes) to 4K, 40K, or more...
CogniBlox will do it in 10µS per vector consuming 250mW per 1 thousand patterns**

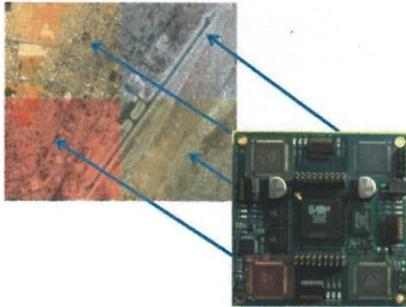
Thanks to the natively hardwired parallel architecture of the CM1K cognitive memory processors, the CogniBlox boards can be stacked vertically via a "spine" connector, but also connected horizontally to build massively parallel networks.

CogniBlox for Data Mining

Recognize and classify vectors against large datasets or knowledge bases.

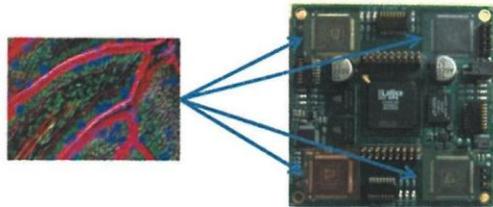
CogniBlox for Video Analytics

Process images N times faster by distributing the recognition workload to multiple CM1K chips.



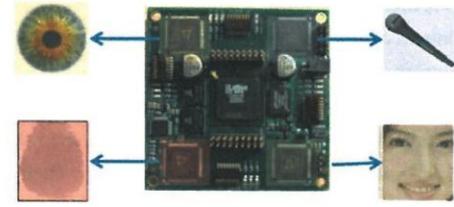
CogniBlox for Complex Recognition

Build robust diagnostics using multiple recognition engine(s) and hypothesis generation.



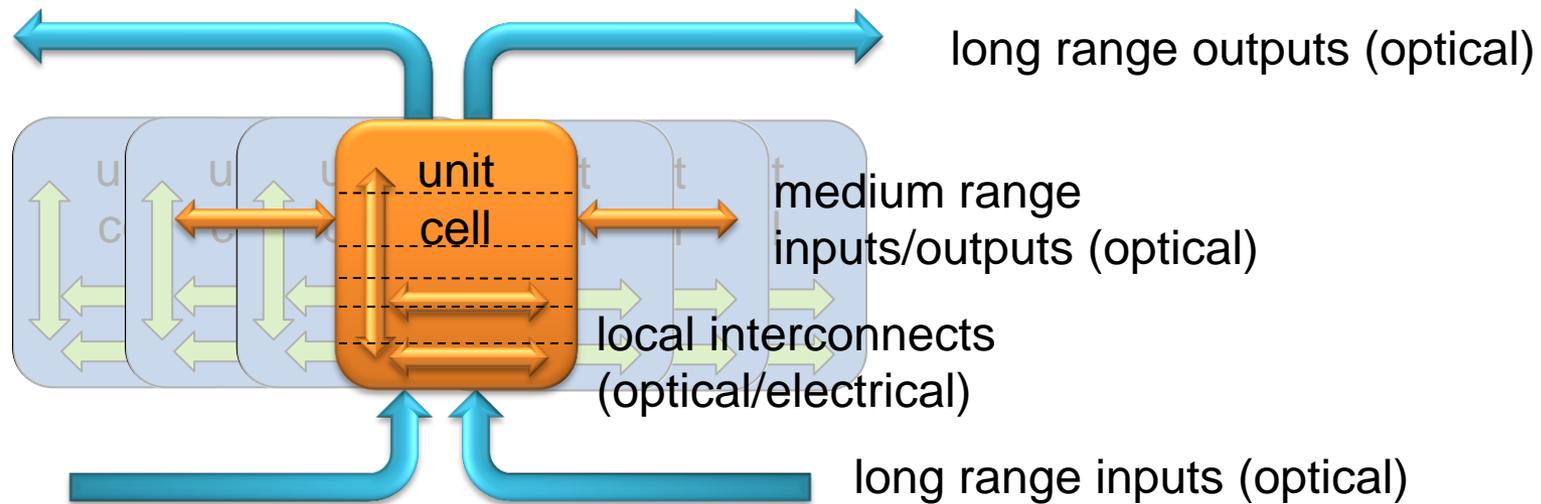
CogniBlox for Sensor Fusion

Multiple sensor inputs (video, sound, accelerometer, etc.) for composite recognition.



COGNITEM™
Technologies, Inc.

Devices and Systems : Future



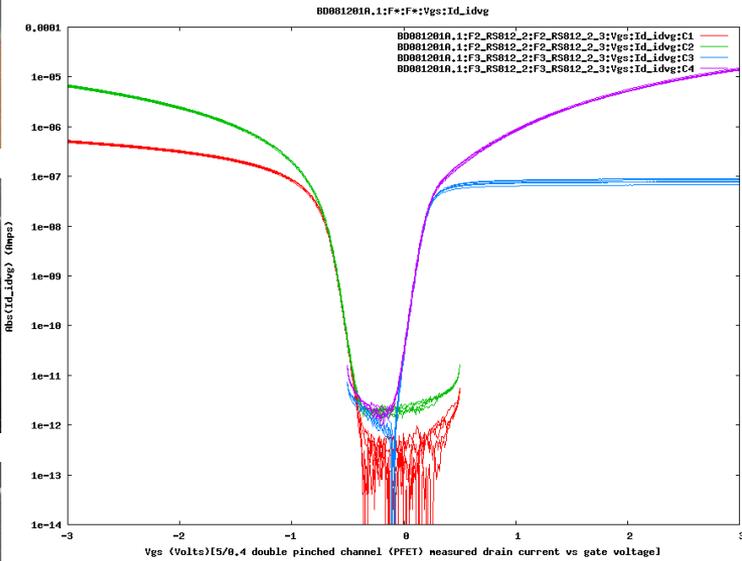
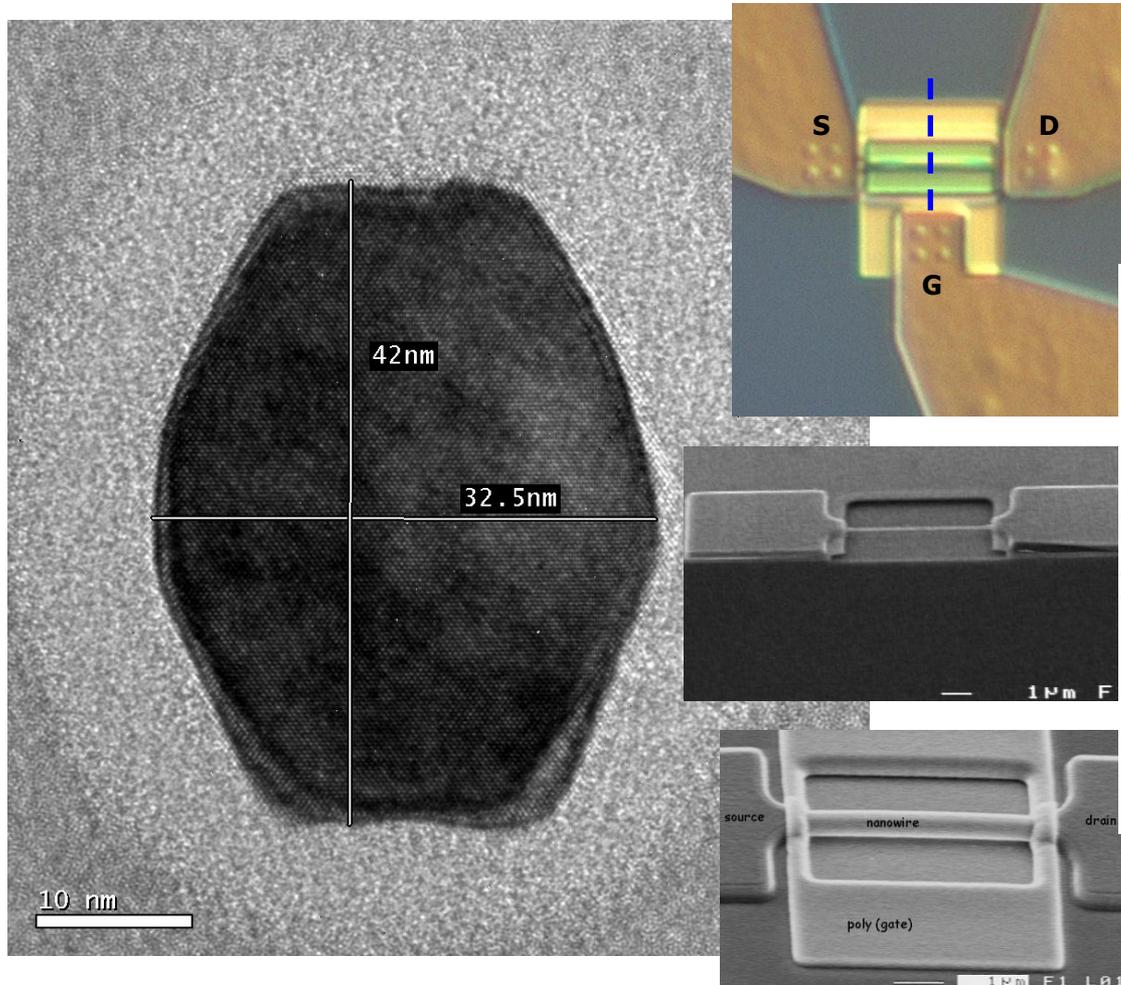
“cortical column” - hierarchical, temporal memory
(On Intelligence, Jeff Hawkins)

3D hybrid integration – opto-electronics, TSV, novel devices, ...

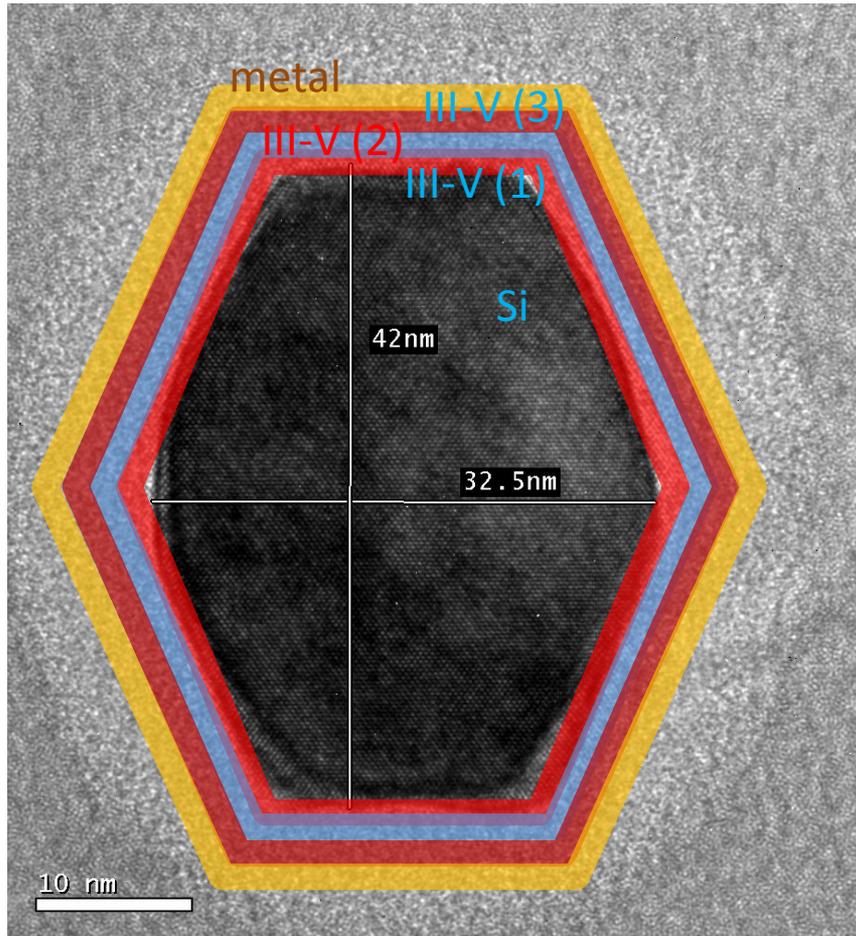
key characteristics:

- Plasticity/adaptability at native (device) level functionality
- Massive interconnect/fanout at system level

CMOS embedded Si Nanowires (MESA)



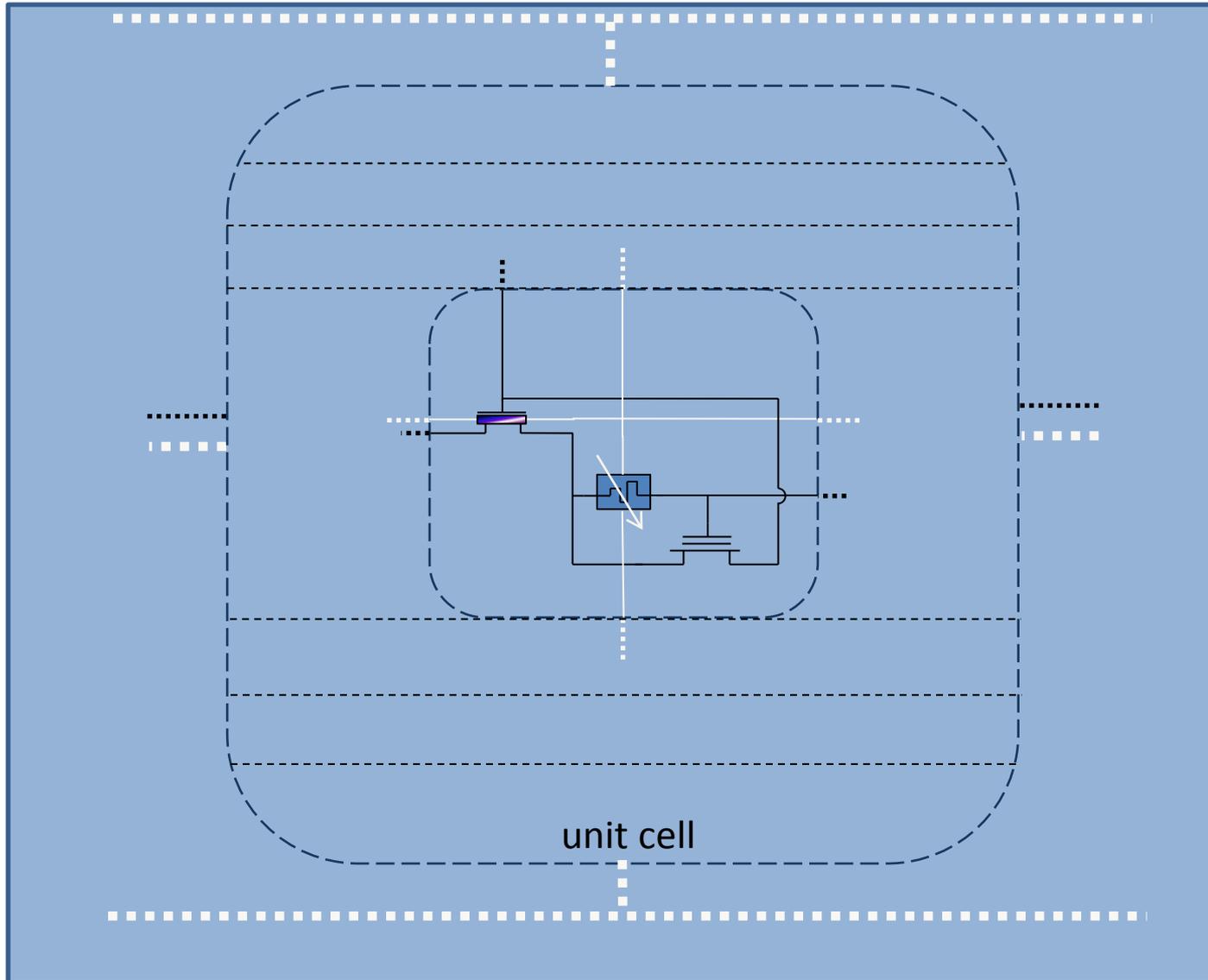
Nanoscale Optically Active Devices in CMOS – Basic Building Block



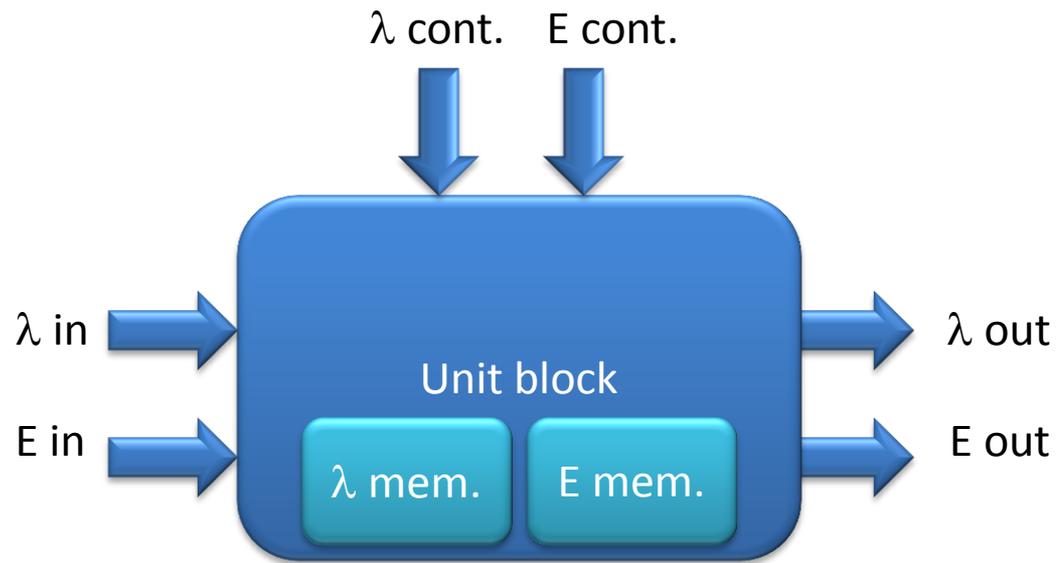
key characteristics:

- Plasticity/adaptability at native (device) level functionality
- Massive interconnect/fanout at system level

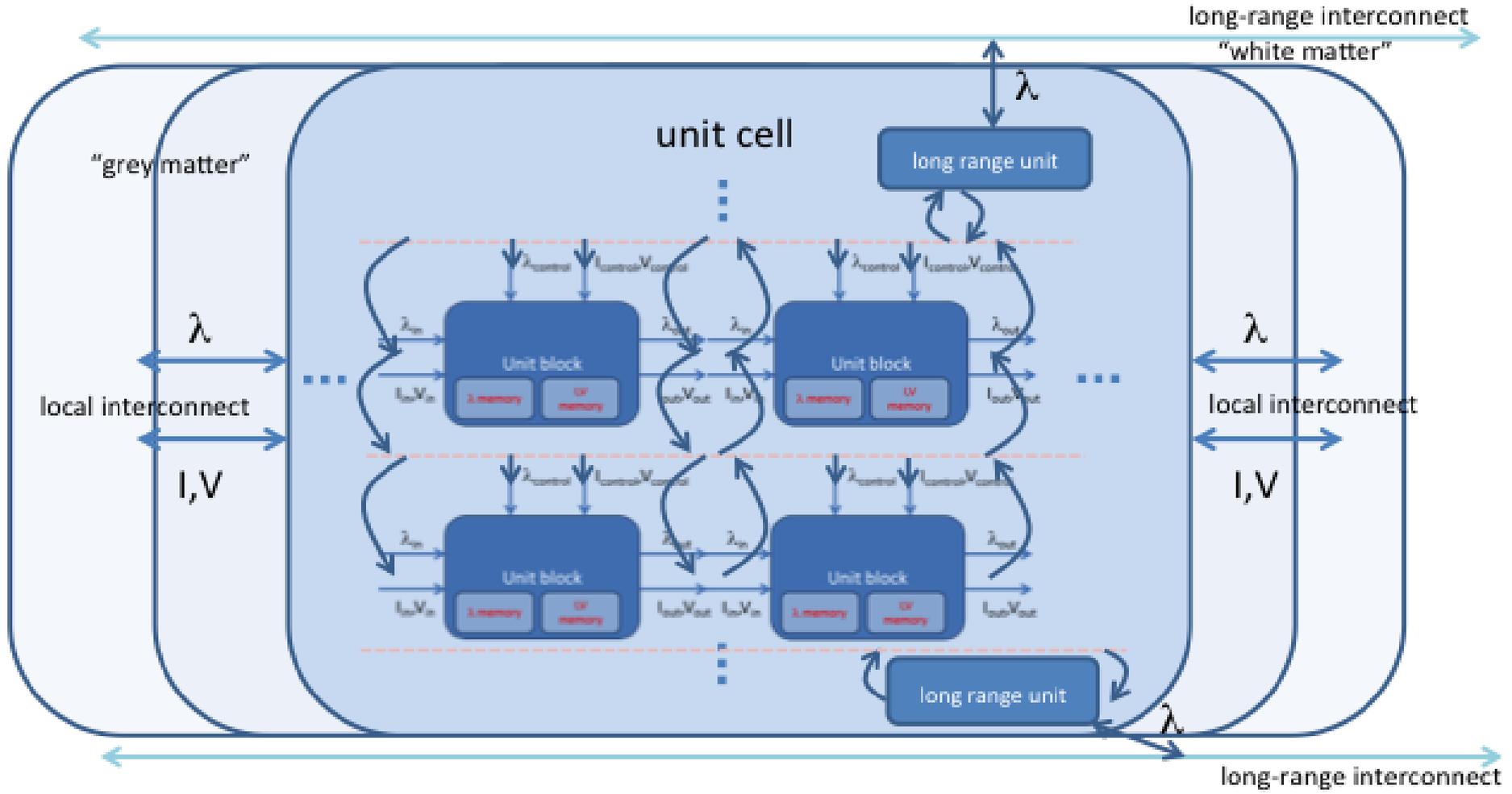
Notional Block Diagram for Unit Cell



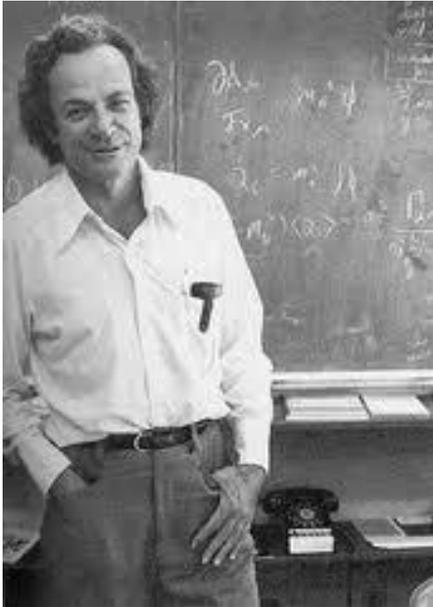
Unit Block – 6 port element



Unit Cell



What we are going to do with it...



Feynman's Corollary on new technology

“Like everything else new in our civilization,
it will be used for entertainment.”

Feynman's second nanotechnology talk, 1983

Potential Applications

Native probabilistic computation

- low power, dynamic
- robotics, communication systems, on-board decision support...
- abstract reasoning
- unhackable, non-reverse engineerable systems

Large datasets, datastreams

- Fraud prevention, anomaly detection, ...
- without needing >MW power levels

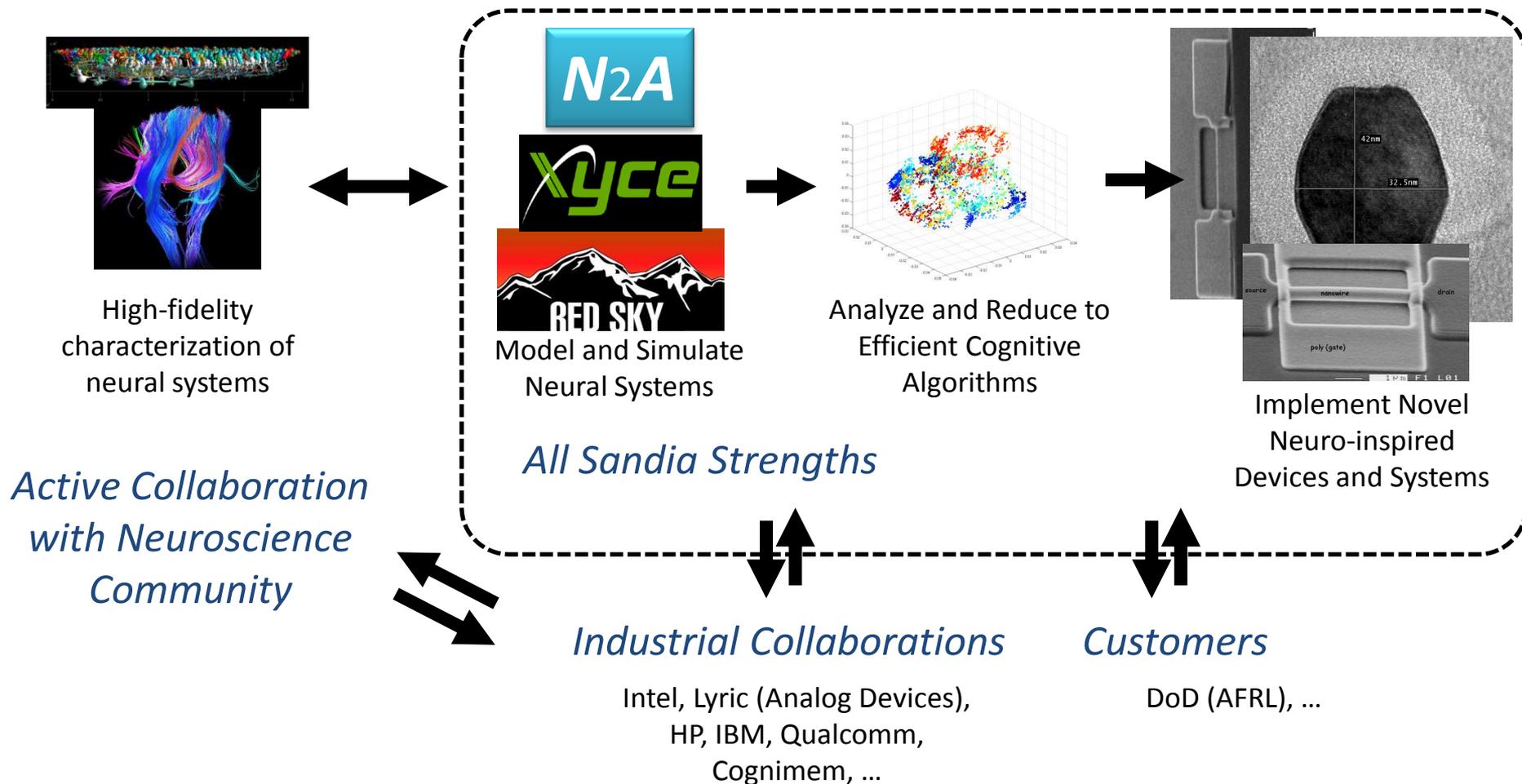
Modeling large, complex, probabilistic systems

- Physics, anthropology, economics, markets, (history?), ...
- Uncovering patterns not readily observable.

Link between electronic and biological (neuro) systems

- Neural prosthesis
- Augmentation

Path Forward



1st Neuro-inspired Computational Elements Workshop (25-27 February 2013)