LESSONS FROM BRAIN CONNECTIVITY

For NextGen 3D NanoICs

Jan M. Rabaey
University of California @ Berkeley

Biomimetic Nanotechnology
Nov. 17, 2015
Computing at the Crossroads

The waning days of Moore’s Law

Speed, energy and efficiency (and economics) plateauing

And Variance/Uncertainty
Computing at the Crossroads

The Memory Wall

“What’s needed today is not so much the ability to process each piece of data a great deal, it’s the ability to swiftly sort through a huge amount of data”, Mark Dean, VP IBM

The Interconnect Woes

“On-chip interconnect is dominating device power consumption, performance and cost”, Wu & Kumar, Applied Materials
One possible path forward:

Get inspiration from biology to empower nanotechnology
The Brain Promise

- **2-3 orders more efficient** than today’s silicon equivalent
  (>10^{16} FLOPS with ~20 W)

- Robustness in presence of component failure and variations
  - Neural response is highly variable ($\sigma/\mu \approx 1$) [Faisal]

- **Amazing** performance with mediocre components
  - E.g. sensory pathways—auditory, olfactory, vision, …
Some Basics about the Human Brain

- 2% of total body weight
- 20W avg power dissipation (20%)
- Gray matter – the home of computation and memory
  - 85 billion neurons
  - 150 trillion synapses
- White matter – the communication infrastructure
- Ratio of white/gray: 1.3-1.5

Brain facts and figures: https://faculty.washington.edu/chudler/facts.html
Ratio of White versus Gray Matter

Ratio constant over 5-6 orders of magnitude

Lower than ICs where wires take up to 90% of volume

\[ W \approx G^{4/3} \]

[From: “Communication in Neuronal Networks,” S. Laughlin and T. Sejnowski, 2003]
Neural Communication 101

- **Dendrite**
- **Axon**
- **Cell body (Soma)**
- **Synapse**

**Neuron fan-in:** 100s - 10,000
**Axon length ranges from 100 µm to 1m meter (sciatic nerve)**

**Configurable interconnection point**
(electro-chemical)
Neural Communication 101

Propagation speed between nodes:
- up to 120 m/sec
- 100 times faster than without myelin!
- VERY slow compared to copper/silicon (≈ 150,000 km/sec)
  - Delay over longest axon: ~ 10 msec

Source: wikipedia
Neural Communication 101

Spike (Pulse) Based Physical Layer

- Accumulation of charge as results of input events leads to neuron firing

![Recorded neural signal (rat cortex)](image)

- Leads to different sources of noise and sensitivity to variation
- Favors event-driven (asynchronous) computational paradigms
Neural Communication 101

Information encoded in many different ways – rate, interval, phase, ...

More than single wires: Spatial or pattern-based data representations

Different types of spikes

[E. Izhikevich, IEEE Transactions on Neural Networks (2004)]

Example: Analyte identification in olfactory system
Brain network topology

No media access layer

- Point-to-point (arguably)
- Many short and some long wires
- Semi-stationary yet reconfigurable
- Random yet structured

- 3D !!!
Interconnect Structure

Random?

Seems so at first glance ...
Combined with massive redundancy

A powerful but expensive proposition
Interconnect Structure

Random?

Seems so at first glance ...
Combined with massive redundancy

A powerful but expensive proposition

But ...

Mixture of short and long, thick and thin connections ...

Similar to FPGAs

[Perge et al, J. of Neuroscience, Jan 2012]
Interconnect Structure

Random?

But ... Cerebellar Cortex
An associative memory!
(3 layer structure)

Seems so at first glance ... Combined with massive redundancy
A powerful but expensive proposition
Interconnect Structure

Random?

Seems so at first glance ...
Combined with massive redundancy

A powerful but expensive proposition

But ... Cerebral Cortex
Cortical Columns
(6 layer structure)

Inter-column

Intra-column

To inner regions
The Grand Scheme of Things

- Functions mapped directly to local regions, connected by global interconnect and some global signaling
- Reactive model of computation

See movie ...
Begs the question:

Is any of this relevant for future nano-technology ICs, their interconnect, architecture, organization and operation?
The 3D Nano Promise

From:
Essentially connections between points in a 2D plane

To:
Huge numbers of devices distributed over multiple stacked layers (3D)
Intertwining memory and computation

Samsung 3d Nand Flash
The 3D Nano Challenge

- Huge variability
- Large error and failure rates
- Low SNR operation

Can inspiration from our “biological computer” come to the rescue?
A Case Study

The cerebellar cortex, associative memory and hyper-dimensional computing

The cerebellum:
Plays important role in motor control.
may also be involved in cognitive functions such as attention and language,
and in regulating fear and pleasure responses
A Nano-Technology 3D Equivalent

Random Indexer:
- **Ultra-dense** 3D integration of RRAM+CNFET delay cells
- **Exploit** High RRAM + CNFET Variability

Associative memory array:
True 3D stacked silicon/CNT/RRAM technology
[Shulacker, IEDM 2014]
Structured Associative Memory
Approximate distributed in-memory computation

Sparse
< 1 % elements switching at any time

Random-Variation Resistant
Law of the numbers

Leakage Resistant
Pulse (event) driven operation
Enables Novel Computational Models

**Hyperdimensional computing**
Representations with dimension “much” (> 10,000) larger than needed to cover space
- Extremely robust against **most failure mechanisms and noise**
- Purely statistical, thrives on randomness
- Information distributed in space
- **Supports full algebra**

**Superb properties:**
- Allow for computation **under very low SNR and high variability conditions**

Extremely inefficient or even impossible in current 2D platforms

[Ref: P. Kanerva, An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors, 2009]
Random Indexing/Associative Memory

Example: Identifying Languages

21 languages
1000 sentences/language
Letters only

"Je kunt een scherm verwerken in een bril en een toetsenbord in een polsband, die dan met elkaar communiceren."

Random Indexer

Input Sentence

Random Indexer

10,000 bit random vector

Associative Memory

"dutch"

Associative Memory

21,000 10,000D vectors
Stored in Associative Memory

Identified Language

Correct language chosen 97.8% of time
Equally applicable to speech, music, ...

[Courtesy: P. Kanerva, UCB]
The Big Picture: In-Memory Computing

Consistent with evolution in computing architectures
- Increasing parallelism
- Processing moving to memory
- From hot to dark processors
Extend to More Complex Cortical Topologies?

**Deep Learning**
- Very successful in pattern recognition, adaptive control, etc
- Based on neural-net architectures
- Mostly implemented on traditional architectures (GPU)

Recent progress in building large neural nets in 2D standard silicon
- E.g. IBM TrueNorth: 70 mW, 4096 processors, 46 billion synapses/sec
- But – extremely hard to program, still far from brain efficiency

[D. Modha, IBM]
Lessons for Future Nano-ICs

• 3D!!!
• Locality!
• Massive concurrency with specialization
  – Distributed processing
  – In-memory computing
• Event-driven execution
• Robust (random) data representations
• Optimization of cell complexity and size versus interconnect density and speed
High Order Bits

• Waning days of Moore’s law requires fresh look at how computational engines are built
• The paths between physical and biological information processing may be converging
• Interconnect architecture is key in both!
• Need true understanding of information-theoretic and thermodynamic trade-off’s