Computing like the Brain – From Dream to Reality?

Karlheinz Meier
Ruprecht-Karls-Universität Heidelberg

Brain-Inspired Computing Workshop at ISCA 40

June 24, 2013
Tel Aviv, Israel
Situation
Realizations
Future
What is Neuromorphic Computing?

A Physical Model rather than a Mathematical Model

\[ C \frac{dV(t)}{dt} = -g \left( V(t) - U \right) \]
You heard the Arguments many times before

Neuromorphic Systems are ...

Low Power solving the energy problem
Fault Tolerant solving the reliability problem
Self Organized solving the software problem
Fast solving the simulation time problem
Compact solving the volume problem
And there is more to this..

A radical departure from what is probably the most successful “mathematics-to science-to-technology roadmap” of humankind:

Boole (Theory) – Shannon (Circuits) – Turing (Programmability)
   – von Neumann (Architecture) – Kilby (Integration)

The brain does NOT:

• use Boolean logic (Boole)
• use electronic realisations of logic gates (Shannon)
• use programmed code (Turing)
• use separate memory and compute units (von Neumann)
• use quasi two-dimensional wiring (Kilby)
So, why has neuromorphic computing not taken over?

Reason 1
So, why has neuromorphic computing not taken over?

Reason 2
So, why has neuromorphic computing not taken over?

Reason 3
So, why has neuromorphic computing not taken over?

$L_G = 30\text{nm}$

Reason 4
How much does a Neural Computation „cost“ ?
A rough (and incomplete) estimate of 2 contributions

Approx. $10^9$ ATP molecules to be hydrolyzed for action potential
Approx. $10^5$ ATP molecules to be hydrolyzed for synaptic transmission

D. Attwell and S. B. Laughlin

Obtain $10^{-19}$ Joule per ATP molecule


$10^{-10}$ Joule (100,000 fJ = 0.1 nJ) per action potential
$10^{-14}$ Joule (10 fJ) per synaptic transmission

100 Billion neurons firing at 1 Hz : $10$ J / s = 10 W
$10^{15}$ Synapses transmitting at 1 Hz : $10$ J / s = 10 W

SMALL (even realistic..) numbers
Electronics vs. Biology on the device level - Not a big difference!

Metal (300 nm times 300 nm)
Insulator (Oxide) < 100 Atom Layers
Semiconductor

„Switching“ of a MOS element: approximately 1 fJ (much less today)

Synaptic Transmission: approximately 10 fJ

„10 low-tech CMOS Transistors“
The communication problem

(almost) twodimensional system of „connecting wires“

Spend typically 1000 times more energy in wires compared to transistors (as long as leakage currents are still small)

Energy Problem

Use this network to transport Boolean information

Architecture Problem
So, 2 remaining reasons

Lack of neuroscience knowledge: Calling for Flexibility

No serious implementations: Calling for large scale, usable systems

... sounds pretty much like supercomputers .....
Santiago Ramon y Cajal (1852-1934)

Individual cells in the brain are spatially separate objects

“interaction over a distance”
and

“spatial and temporal integration”
"Spike“ Communication across dynamic links (synapses)

Charge integration across cell membrane (neurons)

\[ C \frac{dV(t)}{dt} = -g \ (V(t) - U) \]

Stereotypic action Potential „Spike“
Continuous time scale
Biology vs. Numerical Simulation
A naive and not quite correct scaling exercise

Cortical Column
0.000003 Watt
100.000 Watt

Mouse Brain
0.03 Watt
1.000.000.000 Watt

Human Brain
30 Watt
1.000.000.000.000 Watt

Installed in D (2010) : 170.000.000.000 Watt
AND : typically 100 times slower than biology
The Key **Arguments** for Neuromorphic Computing

- Low Power (energy per fundamental operation)
- Fault Tolerance
- Plasticity / Learning / Development („no algorithm“)
- Speed
- Scalability

The Key **Challenges** for Neuromorphic Computing

- Neuroscience Knowledge, *Flexibility*
- Configurability, *Technologies for distributed memory*
- Integration Density, *nano-components, 3D Integration*
- Circuit re-Use, *CAE Tools*
- User Access, *Unified software toolset*
Developing and Tuning Neuron Models – From Biology to Mathematics

Gerstner, Naud, Science 326 (2009) 379 + references therein
Adaptive-Exponential Integrate-and-Fire Neuron Model (I)

\[ C \frac{dV}{dt} = -g_L(V - E_L) + g_L \Delta_T \exp \left( \frac{V - V_T}{\Delta_T} \right) + I - w, \]  

(1)

\[ \tau_w \frac{dw}{dt} = a(V - E_L) - w, \]  

(2)

R. Naud et al.  
The Adaptive-Exponential IF Neuron Model (II)

Brette, Gerstner, Adaptive Exponential Integrate-and-Fire Model as an Effective Description of Neuronal Activity, J Neurophysiol 94: 3637-3642, 2005
Physical Models of Spike based Computation – The Communication Aspect
Physical Models of Spike based Computation - Approaches

Discrete Time, Discrete Signal

Continuous Time, Discrete Signal

Continuous Time, Continuous Signal
The UK SpiNNaker Project

- 18 ARM 968 Cores per chip
- Integer Operations
- 200 MHz Processor Clock
- Shared system RAM on die
- 128 Mbyte DRAM stacked on die
- Each Chip 6 bi-directional links
- 6 million spikes per second per link
From Mathematics to Electronic Physical Models

layout drawing of two neurons: 150x20 μm²

Combining multiple membrane circuits with 256 synapses each allows neurons with up to 16k pre-synaptic inputs.
Neural Processing Unit,
200,000 Neurons, 50,000,000 plastic Synapses, 16,000 syn. I/P per neuron
Separation of Neural Circuits and Monitoring/Readout/Control

Control and Communication FPGAs
Control and Communication Board with digital communication ASICs
Neural Network Wafer (8 inch)

*Wafer-scale integration of analog neural networks*
J. Schemmel, J. Fieres and K. Meier
BrainScaleS HMF in Heidelberg Lab
2 Wafer System in Commissioning Stage
6 Wafer System to be delivered in 2014
Multi-Scale Circuit Structure on 8 inch CMOS Mixed-Signal Wafer (180nm)

Plastic Synapses, 50.000.000 Million Instances on Wafer, Length Scale 10 μm
4-bit SRAM Weights, STP, STD, STDP

High Input Count Network Chips, 400 Instances on Wafer, Length Scale 1 cm network routing

AdEx Neurons, 200.000 Instances on Wafer, Length Scale 300 μm, Analog Floating Gate Parameter Storage Poisson Noise Generators
Communication Architecture Overview

Hierarchical 2 Layer Communication Setup

- **Layer 1**: On-Wafer continuous, asynchronous, fixed delay spike transmission
- **Layer 2**: Off-Wafer Packet based digital communication for medium and long-range spike transmission, network set-up, read-out and control.
Aggregate Bandwidth from off-Wafer Connections exceeds 1 Terabit/s
On-wafer Information Flow >> 10 Terabit/s

- Printed Circuit Board
- 10,000 Vertical Links Deliver approx. 500 A to 320 cm² of active silicon Send and Receive 2 Gbit/s/Link
- Wafer-to-Wafer Routing Circuits
Energy efficiency?
Current Status – Not optimized for low power

Accelerated BrainScaleS ANN wafer at conservative (maximum) rate: 3 Watts/cm²

Approximately $10^{-10}$ J per synaptic transmission

Biology: $10^{-14}$ j per synaptic transmission

X 10,000 gap downwards to biology but another large upwards gap to HPC simulations

F Pollack
**Energy Scales**

Energy used for a synaptic transmission

14 orders of magnitude difference for „the same thing“

Physical models (*Neuromorphic*)

- Typically 10.000.000 times more energy efficient than state-of-the art HPC (comparable model)
- 10.000 less efficient than biology

From: HBP project report
Continuous Time Integrating Neural Cell Membrane Model - *Neuromorphic*

\[ C_m \frac{dV}{dt} = g_{\text{leak}} (E_{\text{leak}} - V) \]

- **R** = \( \frac{1}{g_{\text{leak}}} \)
- **V(t)**

<table>
<thead>
<tr>
<th>Biology(*)</th>
<th>( \Delta V ) [V]</th>
<th>( g_{\text{leak}} ) [S]</th>
<th>( C_m ) [F]</th>
<th>( (g\Delta V)/C ) [V/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology(*)</td>
<td>( 10^{-2} )</td>
<td>( 10^{-8} )</td>
<td>( 10^{-10} )</td>
<td>( 10^{0} )</td>
</tr>
<tr>
<td>VLSI</td>
<td>( 10^{-1} )</td>
<td>( 10^{-6} )</td>
<td>( 10^{-13} )</td>
<td>( 10^{6} )</td>
</tr>
</tbody>
</table>

(* Brette/Gerstner, J. Neurophysiology, 2005

Inherent speed gap: \( 10^{6} \) Volt/second

\[ \rightarrow \text{accelerated neuron model} \]
## Time Scales

<table>
<thead>
<tr>
<th></th>
<th>Nature</th>
<th>Simulation</th>
<th>Accelerated physical model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection of causality</td>
<td>$10^{-4}$ s</td>
<td>0.1 s</td>
<td>$10^{-8}$ s</td>
</tr>
<tr>
<td>Plasticity</td>
<td>1 s</td>
<td>100 s</td>
<td>$10^{-4}$ s</td>
</tr>
<tr>
<td>Learning</td>
<td>day</td>
<td>100 days</td>
<td>10 s</td>
</tr>
<tr>
<td>Development</td>
<td>year</td>
<td>100 years</td>
<td>3000 s</td>
</tr>
</tbody>
</table>

### 12 Orders of magnitude

| Evolution | > millenia | > 100 millenia | > month |

### > 15 Orders of magnitude

Temporal dynamics is key to understanding (and using) the computational paradigms of the brain.
The Neural FPGA Concept for rapid Prototyping CONFIGURABILITY!

Biological Databases → Model Building Cells, Network, Plasticity → Theory

Formal Network Description → Technology Mapping

Technology Routing → Hardware Platform Constraints

Simulation and Verification → Neuromorphic Substrate
Implement Custom Digital Circuits: A Success Story for Configurable Hardware

- Code
- Map and Route
- Synthesize
- Verify
- Load
- Verify
- Use
Typical Configuration Space for a Neuromorphic System approx. 40 MB for a full Wafer

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>Type Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuron circuits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synapse line drivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synapses</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4: Sector diagram of the parameter space to configure one HICANN chip. For a full wafer, the configuration data volume is 44 MB large.
Experiments and Applications under Study
In *BrainScaleS* and *HBP* - 4 Categories

I. Fundamental dynamical properties of isolated circuits
   *Synchronisation, coincidence detection, stability, order-chaos*

II. Implement and test fundamental, generic concepts and theories
    *Liquid computing, probabilistic inference, neural sampling*

III. Biologically realistic, reverse engineered circuits in closed loops
    *Cortical structures, cortical columns, functional units*

IV. Generic neuromorphic computing outside neuroscience
    *Neuromorphic controllers, spatio-temporal pattern detection in data streams, causal relations in big data, approximate computing*
Synfire Chain with Feed-Forward Inhibition, Kremkov et al. 2010

Balanced random network (Brunel 2000)
 asynchronous, irregular firing
Prerequisite for information processing using stochastic inference

Hopfield Network

- Deterministic mapping of input and output patterns
- Network configuration through cost (energy) minimization in local minimum
- Network in fixed state

Boltzmann Machine

- Stochastic Neurons
- Probability for states given by Boltzmann distribution in thermal equilibrium
- Learn probability distributions of network states
Liquid State Machine (Maass et al., 2002) with binary tempotron classifier readout (Gütig&Sompolinsky, 2006)

Distinguish spike train segments in a continuous data stream composed of two templates (X, Y) with identical rates

Cortical Layer 2/3
attractor memory network
Lundqvist et al. 2006, 2010

Insect antennal lobe inspired classification of multidimensional (odor) data
(Schmuker&Schneider 2007)
Decorrelating inputs from sensory neurons - association to choices

The appeal of neuromorphic architectures lies in

i) their potential to achieve (human-like) intelligence based on unreliable devices typically found in neuronal tissue

ii) their strategies to deal with anomalies, emphasizing not only tolerance to noise and faults, but also the active exploitation of noise to increase the effectiveness of operations

iii) their potential for low-power operation.

Traditional von Neumann machines are less suitable with regard to item i), since for this type of tasks they require a machine complexity (the number of gates and computational power), that tends to increase exponentially with the complexity of the environment (the size of the input). Neuromorphic systems, on the other hand, exhibit a more gradual increase of their machine complexity with respect to the environmental complexity.

Therefore, at the level of human-like computing tasks, neuromorphic machines have the potential to be superior to von Neumann machines.
Noise, Anomalies and Faults

- **Temporal** noise (at frequencies small compared to relevant system time constants)
  - Incoherent
    - Not user controllable
      - **thermal** noise
      - **shot noise** (discrete charge carrier statistics)
    - **User controllable**
      - on-purpose integrated pseudo or quantum noise sources (spikes)
      - Irregular network network activity („spikes, sea of noise“)
  - Coherent
    - Somewhat user controllable
      - Cross-talk from digital control signals to analogue membranes

- **Spatial** (fixed-pattern) noise
  - **Static** device mismatch
    - To some extend user controllable (hard)
      - Neuron parameters (voltages and timing), „Calibration“
      - Synapse parameters (voltages and timing), „Calibration“
    - Not user controllable
      - Trial-to-trial variations from analogue parameter setting
      - Dead or faulty components
Full chip $E_L$ calibration

Target: $E_L = 655$ mV

![Graph showing membrane potential distribution](image)

Mean = 658.9 mV | Std = 28.5 mV
Full chip $E_L$ calibration

Target: $E_L = 655$ mV

Mean = 654.3 mV | Std = 5.1 mV
Pyloric rhythm of the crustacean stomatogastric ganglion

20,000,000 model networks with 17 random cell parameters, fixed connectivity (Neuron)
400,000 networks found with de-generate timing behaviour in measured biological range

Sensitivity of single parameters within „de-generate“ solutions

Marder, Taylor Nature Neuroscience 14, Nr 2, 2011
Generic Loop

Phenomenological Model ($P_i$)

- modify

Computational performance

- measure

Systematic Workflow for Model and Theory Development / Verification / Falsification

HBP Neuromorphic Computation Platform

- modify

resolution scales: $\Delta x, \Delta t$

output: $O(\Delta x, \Delta t, t)$

config. space: $P_i$

runtime: $\Delta T$

Biological Loop

Experimentation Biology

- in-vitro, in-vivo

Theory

- verify / falsify

- build

- compare
Hybrid (Neuromorphic-HPC) Multiscale Modelling BrainScales HMF

Virtual environment

Action Perception Rewards

Microscopic – macroscopic Milliseconds – years

Rapid cycling of experiments

Bridging Scales \( (x,t) \)

\[ c_m \frac{dV}{dt} = -g_{\text{leak}} (V - E_1) \]

\[ \Delta t = t_{\text{post}} - t_{\text{pre}} \]

Input of microscopic and macroscopic theory
A large-scale coordinated effort over 10 years
Budget 1.1 B€

Director: Henry Markram (EPFL)
Co-Directors: Richard Frackowiak (CHUV)
Karlheinz Meier (Heidelberg)

Public Report available:
www.humanbrainproject.org

Approval of 30 months ramp-up phase received in January 2013

Expected start: October 2013
Future Computing Platforms
what will they provide? – a few selected items

High Performance Computing
- Interactive, visual. Exascale supercomputing
- Massive distributed volumes of heterogeneous data
- Convergence with neuromorphic technology

Neuromorphic Computing
- First large-scale neuromorphic systems superior to HPC
- Non-von Neumann (Multicore) + non-Turing (Neuromorph)
- Technology integration (3D, non-CMOS backends)

Neurorobotics
- Virtual robots with two-way, closed loop interfaces
- Link to brain models and neuromorphic systems
- Physical prototypes and applications
Manchester Small-Scale Experimental Machine (SSEM)
World's first stored-program computer
Frederic C. Williams, Tom Kilburn and Geoff Tootill
First program on 21 June 1948
NM-MC-1
The SpiNNaker Many-Core Machine

103 machine: 864 cores, 1 PCB, 75W

104 machine: 10,368 cores, 1 rack, 900W
(NB 12 PCBs for operation without aircon)

105 machine: 103,680 cores, 1 cabinet, 9kW

106 machine: 1M cores, 10 cabinets, 90kW
NM-PM-1
The Physical Model Machine

Neuromorphic Physical Model (NM-PM-1)

20 Neuronal Network Wafer
20 Peta-Connections/s

1000 Kintex FPGAs

Conventional Computing Cluster
4 T op/s

1 Tbit/s

4 Million Neurons
1 Billion Synapses

20 Intel Haswell
BRAIN-DERIVED COMPUTING

- Consistent concept for a novel, brain derived, non-von Neumann, non-Turing computer architecture
- Accessible to available technologies (CMOS) and attractive application for future component technologies (nanoelectronics)
- Key features: Universality, scalability, fault tolerance, power efficiency, speed, learning
- Accelerated operation: Only known approach to bridge all timescales relevant for circuit dynamics
- Important next step: Give up simulation as a reference, exploit device mismatch and noise
THANKS!

brainscales.eu