Multi-scale levels of investigation in analysis of the central nervous system (adapted from Churchland and Sejnowski 1992) and corresponding neuromorphic synthesis of highly efficient silicon cognitive microsystems. Boltzmann statistics of ionic and electronic channel transport provide isomorphic physical foundations.

Scaling of Task and Machine Complexity

Achieving (or surpassing) human-level machine intelligence requires a convergence between:

- Advances in computing resources approaching connectivity and energy efficiency levels of computing and communication in the brain;

- Advances in deep learning methods, and supporting data, to adaptively reduce algorithmic complexity.


Adiabatic CID-DRAM SVM (*Kerneltron*)
R. Karakiewicz et al, 2013

Synaptic Sampling Machine (SSM)
E. Neftci et al, 2016
Scaling and Complexity Challenges

- Scaling artificial neural systems to performance and efficiency approaching that of the human brain will require:
  - Scalable advances in silicon integration and architecture
    - Scalable, locally dense and globally sparse interconnectivity
      - Hierarchical address-event routing
    - High density \(10^{12}\) neurons, \(10^{15}\) synapses within 5L volume
      - Silicon nanotechnology and 3-D integration
    - High energy efficiency \(10^{15}\) synOPS/s at 15W power
      - Adiabatic switching in event routing and synaptic drivers
  - Scalable models of neural computation and synaptic plasticity
    - Convergence between cognitive and neuroscience modeling
    - Modular, neuromorphic design methodology
    - Data-rich, environment driven evolution of machine complexity
# Large-Scale Reconfigurable Neuromorphic Computing

## Technology and Performance Metrics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Die Size (mm²)</td>
<td>102</td>
<td>430</td>
<td>50</td>
<td>168</td>
<td>16</td>
</tr>
<tr>
<td>Neuron Type</td>
<td>Digital</td>
<td>Digital</td>
<td>Analog</td>
<td>Analog</td>
<td>Analog</td>
</tr>
<tr>
<td>Neuron Type</td>
<td>Arbitrary</td>
<td>Accumulate &amp; Fire</td>
<td>Conductance Integrate &amp; Fire</td>
<td>Shared-Dendrite Conductance I&amp;F</td>
<td>2-Compartment Conductance I&amp;F</td>
</tr>
<tr>
<td># Neurons</td>
<td>5216 ¹</td>
<td>1M ²</td>
<td>512</td>
<td>65k</td>
<td>65k</td>
</tr>
<tr>
<td>Neuron Area (µm²)</td>
<td>N/A ¹</td>
<td>3325 (14) ²</td>
<td>1500</td>
<td>1800</td>
<td>140</td>
</tr>
<tr>
<td>Peak Throughput (Events/s)</td>
<td>5M</td>
<td>1G</td>
<td>65M</td>
<td>91M</td>
<td>73M</td>
</tr>
<tr>
<td>Energy Efficiency (J/SynEvent)</td>
<td>8n</td>
<td>26p</td>
<td>N/A</td>
<td>31p</td>
<td>22p</td>
</tr>
</tbody>
</table>

¹ Software-instantiated neuron model
² Time-multiplexed neuron (256x)


Comparison of synaptic connection topologies for several recent large-scale event-driven neuromorphic systems and the proposed hierarchical address-event routing (HiAER), represented diagrammatically in two characteristic dimensions of connectivity: expandability (or extent of global reach), and flexibility (or degrees of freedom in configurability). Expandability, measured as distance traveled across the network for a given number of hops $N$, varies from linear and polynomial in $N$ for linear and mesh grid topologies to exponential in $N$ for hierarchical tree-based topologies. Flexibility, measured as the number of target destinations reachable from any source in the network, ranges from unity for point-to-point (P2P) connectivity and constant for convolutional kernel (Conv.) connectivity to the entire network for arbitrary (Arb.) connectivity. MMAER: Multicasting Mesh AER; WS: Wafer-Scale.
Hierarchical Address-Event Routing (HiAER) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing [Park et al, 2012; Yu et al, 2012]. (a) Dynamic reconfigurable synaptic connectivity across IFAT arrays of addressable neurons is implemented by routing neural spike events through DRAM synaptic routing tables (SRT). (b) The IFAT neural array multiplexes and integrates (top traces) incoming spike synaptic events to produce outgoing spike neural events (bottom traces). (c) Full-size HiAER-IFAT network with 4 boards, each with 4 IFAT modules, serving 1M neurons and 1G synapses, and spanning 4 levels in connection hierarchy. (d) Each IFAT chip module comprises a 65k-neuron Tezzaron 130nm CMOS IFAT microchip, Xilinx Spartan-6 FPGA (Level 1 HiAER), and two 2Gb DDR3 SDRAM SRTs serving 65M synapses. (e) Each neural cell models conductance based membrane dynamics in proximal and distal compartments for synaptic input with programmable axonal delay, conductance, and reversal potential. IFAT chip measured energy consumption is 48 pJ per spike event, several orders of magnitude more efficient than emulation on CPU/GPU platforms.

Yu et al, BioCAS 2012; Park et al, TNNLS 2016
Large-Scale Reconfigurable Neuromorphic Computing

- Integrate-and-fire array transceiver (IFAT) as digitally programmable analog neural supercomputer
- Biophysical detail in neural and synaptic continuous-time dynamics
- Record high density: 65k two-compartment neurons with 65M reconfigurable conductance-based synapses
- Record low energy: 22 pJ per synaptic event
- Real-time at 73M spikes per second

Hybridization and nanoscale integration of CMOS neural arrays with phase change memory (PCM) synapse crossbar arrays. (a) Nanoelectronic PCM synapse with spike-timing dependent plasticity (STDP) [Kuzum et al., 2011]. Each PCM element implements a synapse with conductance modulated through phase transition as controlled by timing of voltage pulses. (b) CMOS IFAT array vertically interfacing with nanoscale PCM synapse crossbar array by interleaving via contacts to crossbar rows. The integration of IFAT neural and PCM synapse arrays externally interfacing with HiAER neural event communication combines the advantages of highly flexible and reconfigurable HiAER-IFAT neural computation and long-range connectivity with highly efficient (fJ/synOP range energy cost) local synaptic transmission.
Spiking Synaptic Sampling Machine ($S^3M$)

*Biophysical Synaptic Stochasticity in Inference and Learning*

- Stochastic synapses for spike-based Monte Carlo sampling
  - Models biophysical origins of noise in neural systems
  - Activity dependent noise: multiplicative synaptic sampling rather than additive neural sampling
  - Sparsity in neural activity and in synaptic connectivity

- Online unsupervised learning with STDP
  - Biophysical model of spike-based learning
  - Event-driven contrastive divergence

---

Time-varying Bernoulli random masking of weights

Synaptic stochasticity as biophysical model of continuous *DropConnect*

The $S^3M$ requires fewer synaptic operations (SynOps) than the equivalent Restricted Boltzmann Machine (RBM) requires multiply-accumulate (MAC) operations at the same accuracy.

---

1.2 TMACS / mW

- adiabatic resonant clocking conserves charge energy
- energy efficiency on par with human brain (10^{15} SynOP/S at 15W)
Sub-Micropower Analog VLSI Adaptive Sequence Decoding
Chakrabartty and Cauwenberghs, 2004

Silicon support vector machine (SVM) and forward decoding kernel machine (FDKM)

Forward decoding MAP sequence estimation

840 nW power

Biometric verification

Subthreshold translinear MOS circuits
Programmable with floating-gate non-volatile analog storage