Programmable and Configurable, Neurmorphismically Inspired, Ultra-Low Power Computation

Professor Jennifer Hasler
ph67@gatech.edu
Georgia Tech
Why Analog (Physical Based) Processing?

Digital Hitting Limits of Power Efficiency

- Power Efficiency wall (Production Ics)
  - 1 MAC in 100pJ (10MMAC/mW )

Source → $V_T$ Mismatch (some I-V as well)

Mead Hypothesis: Analog x1000 efficiency improvement

- Analog (VMM):
  - 100 fJ / MAC (10MMAC/µW)

- Other Analog SP similar:
  - Freq Decomp / Analog FT
  - VMM, GMM
  - Classifiers
  - Adaptive Filters

Performance Advantage for Emerging Architectures

- Power Efficiency Scaling
  - 10MMAC(/s)/W → 1st DSPs (1978 - 1981)
  - 100MMAC(/s)/W
  - 1MMAC(/s)/mW → Energy Efficiency Wall (32bit inputs)
  - 10MMAC(/s)/mW
  - 100MMAC(/s)/mW
  - 1MMAC(/s)/uW
  - 10MMAC(/s)/uW → Typical Analog VMM
  - 100MMAC(/s)/uW → FPAA compiled Dendritic Classifier (350nm IC)
  - 1MMAC(/s)/nW
  - 10MMAC(/s)/nW
  - 100MMAC(/s)/nW
  - 1MMAC(/s)/pW → (<) Biological Neuron

Analogs SP
Large-Scale FPAAAs → Practical Analog SP

Integrated FPAA + FPGA (2011)

15 different working FPAA ICs since 2004

FPAA computing through routing fabric
One FPAA, Four Examples

- Mixed-Signal FIR
- Image Convolution
- Arbitrary Waveform
- DAC in Routing
- Constrained Optimization, Path-Planning (R grids)
- VMM, Classifiers, Baseband Comm
FPAA Tool Infrastructure
Neuromorphic Algorithms ➔ Improved Apps

Brain is highly power efficient
- highly constrained by power available,
  key to its design (and for Si)

Building applications (i.e. robotics) makes
power constraints real

Leverage Analog SP ICs for robust system
development

- Neuromorphic processing = event based processing
uses power only when useful signals are present
(“always on” in sensors or further processing)
Blocks for Large-Scale Neuromorphic Systems

Transistor Channel Models

Single Transistor Learning Synapses

- Single Transistor Learning Synapses
  [Hasler, et. al, NIPS 1994, BMES 1994]

Utilizing the physics of physical medium (Si) to efficiently implement computation

Si CMOS approach can achieve densities while avoiding issues with device integration with Si
Comparing Physical to Digital Computations

FG enables analog precision
- not imprecise or overly noisy components
- Power constrains digital to similar resolutions, worse ODE dynamics
Dense Synapse + Neuron IC

- Synapse Array with configurable neuron blocks, STDP and programmable synapses
- Address Event I/O available
- Compiled from standard cells
- Mismatch Programming Essential

WTA network

Synapse Array
(30k synapses ~ 3mm² space
10k synapses/mm²)
Larger Neuron Experiments

EventIn

Synfire chain (0-19)

Synapse Matrix

Neurons (29-44)

EventOut

N Neurons in Sequence

“Synfire chain”

• Neuromorphic Path Planning
Why Dendritic Computation?

- Largest supercomputer (~ 3000TMAC) is $10^4$ factor smaller than required for neural computation (~ 10^7 TMAC)
Dendrite-Model Wordspotting Classifier

[George and Hasler, 2011]

Sensitivity to particular delay window → coincidence detection

Decreasing event delay on resulting soma signal

[Ramakrishnan, et. al, 2012]
Scaling of Neuromorphic Computation

- Neurobiology: constrained by energy cost of communication (local comm is critical)

- Silicon systems also constrained by energy cost and complexity cost of communication.
Neural Classifier Approaches

VMM-WTA block

Neurons are configurable processors (synapse → dendrites → soma)

[Ramakrishnan, et. al, 2012]
Summary of Learning & Adaptation

Neuron Implementations: FG based learning
   Synapse types: excitatory, inhibitory, NMDA
   Experimentally demonstrate

Early (unpublished) data on receptive field
development, different event encodings

Application sets IC data flow speed
   (minimizing memory blocks)

Learning investigations: faster than real time
   - same structure, direct scaling of timescales

Why Learning?  How about loading cortex…

<table>
<thead>
<tr>
<th>Load time</th>
<th>15 minutes</th>
<th>1 day</th>
<th>10 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication Rate</td>
<td>11.3Tbit/s</td>
<td>116Gbit/s</td>
<td>11Gbit/s</td>
</tr>
<tr>
<td>Power</td>
<td>10.4kW</td>
<td>109W</td>
<td>11W</td>
</tr>
</tbody>
</table>

Memristors: Multi-timescale Adaptation

Synapses: Not competitive with
   1T EEPROM cell for density
   (32nm cell ~ 50nm x 50nm device)
   Memristor arrays hard (Liu: 40 x 40)

Neurobiology: multi-timescale devices
   - modulation timescales from 1s to hours
   - adaptive FG allows approach
   - memristors enable capability

\[ I = W R V, \quad \varepsilon \tau \frac{dW}{dt} = f(V(t)) \]

Why Learning? How about loading cortex…

FPPA + Memristor (A)
Building Silicon Cortex

Neuromorphic \(\Rightarrow\) Higher Power Efficiency

Power Efficiency Scaling

- 10MMAC(s)/W 1st DSPs (1978 - 1981)
- 100MMAC(s)/W
- 1MMAC(s)/mW
- 10MMAC(s)/mW Energy Efficiency Wall (32bit inputs)
- 100MMAC(s)/mW
- 1MMAC(s)/uW
- 10MMAC(s)/uW Typical Analog VMM
- 100MMAC(s)/uW FPAA compiled Dendritic Classifier (350nm IC)
- 1MMAC(s)/nW
- 10MMAC(s)/nW Dendritic Classifier (350nm IC)
- 100MMAC(s)/nW
- 1MMAC(s)/pW (<) Biological N……n
- 1000’s of inputs, 1000’s of channel populations, one output

10nm: \(~4M\) Pyramidal cell neurons \(\Rightarrow\) $20M$ IC cost for Cortex (digital: 1000 in parallel) (~100k chips)
Neural Classifier IC Applications

For commercial 10nm potential

<table>
<thead>
<tr>
<th>Sensor Inputs</th>
<th>Sensor Signal Conditioning</th>
<th>Sensor Signal Processing</th>
<th>Signal to Symbol Conversion</th>
<th>First Layer Cortical Classification</th>
<th>Second Layer Cortical Classification</th>
<th>Refined Symbols</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Speech Recognition</th>
<th>Microphone Interface / filtering</th>
<th>Cepstrum</th>
<th>Basic Auditory Features (VQ, GMM)</th>
<th>Phoneme Classification</th>
<th>Low SNR Wordspotting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Processing</td>
<td>Image acquisition, color calculations</td>
<td>Retina (edge enhancement)</td>
<td>Edge / Corner Detection</td>
<td>Movement Sequence Classification</td>
<td>Gesture Recognition, etc.</td>
</tr>
<tr>
<td>Baseband Communications</td>
<td>Demodulation of desired band</td>
<td>Frequency Decomposition</td>
<td>Fundamental Comm Symbol Detection</td>
<td>Frequency Hopping Recognition</td>
<td>Complex Signal Detection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Die Size</th>
<th>Consumer IC</th>
<th>Processor IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1mm²</td>
<td>S2</td>
<td>40mm²</td>
</tr>
<tr>
<td>Chip Cost</td>
<td>$100</td>
<td>$5000</td>
</tr>
<tr>
<td>Neurons</td>
<td>60,000</td>
<td>3,000,000</td>
</tr>
<tr>
<td>MAC</td>
<td>10 TMAC</td>
<td>500 TMAC</td>
</tr>
<tr>
<td>Comp Power</td>
<td>1mW</td>
<td>50mW</td>
</tr>
<tr>
<td>Out Events</td>
<td>1,000 / s</td>
<td>10,000 / s</td>
</tr>
<tr>
<td>Comm Power</td>
<td>70 nW</td>
<td>8μW</td>
</tr>
</tbody>
</table>