OPTIMIZATION AND PERFORMANCE STUDY OF LARGE-SCALE BIOLOGICAL NETWORKS FOR RECONFIGURABLE COMPUTING

Mohammad A. Bhuiyan
Ananth Nallamuthu
Melissa C. Smith
Vivek K. Pallipuram

This work supported in part by NSF Grant No. CCF-0916387
Motivation

• Goal: Enable large-scale cortical simulations
  – $10^{11}$ neurons, $10^{14}$ synapses

• Cortical model simulation requires
  – Highly-parallel processing network
  – High-speed processing power
  – Extremely large storage

• Can Reconfigurable Architectures be exploited for cortical simulations?
Outline

• SRC 7 H MAP
  – Hardware
  – Software
• Spiking Neural Network
  – Biological Models
  – Network design
• Mapping of the network to the SRC system
• Optimization strategies used
• Effect of optimizations on the performance
• Energy efficiency
FPGA System

- Processor Module or Hi-Bar Switch Port
- OBCM 1 GB
- Controller
- OBCM 1 GB
- FPGA 1
- FPGA 2
- OBM 64 MB
- GPIOX

- Intel Xeon dual core, 3.0 GHz
- 7.2 GB/s
- 150 MHz Altera Stratix™v
- 128-bit data channel
- 16 OBM banks BW = 19.2 MB/s

Fig: http://www.srccomp.com
Programmer exploits parallelism

- Two FPGAs
- CPU
- Memory hierarchy

MAP compiler

- Extract parallelism from the code
- Generate pipelined hardware logic

Fig: http://www.srccomp.com
Spiking neuron Models
- Integrate & Fire (1907)
- Hodgkin-Huxley (1952)
- Wilson (1999)
- Izhikevich (2003)
- More ........

Spiking Neural Networks (SNN) are widely employed for mammalian brain simulations

E. Izhikevich on Integrate & Fire Model:
“The model cannot exhibit even the most fundamental properties of cortical spiking neurons, and for this reason it should be avoided by all means.”

More accurate Models:
- Hodgkin-Huxley
- Izhikevich
- Wilson model
The ratio of total flops required for all neuron updates to the total memory required for these updates.

<table>
<thead>
<tr>
<th>SNN Model</th>
<th>Differential Equations</th>
<th>Flops per Neuron Update</th>
<th>Flop/Byte Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Izhikevich</td>
<td>2</td>
<td>13</td>
<td>0.65</td>
</tr>
<tr>
<td>Wilson</td>
<td>4</td>
<td>38</td>
<td>0.86</td>
</tr>
<tr>
<td>Hodgkin-Huxley</td>
<td>4</td>
<td>246</td>
<td>6.02</td>
</tr>
</tbody>
</table>
Network Design

- Two layer network
- Adapted from Gupta and Long paper
- Binary input fed to neurons in Level 1

![Network Diagram]

- Least compute intensive (<5% of total computation). Requires transfer of large weight matrix. Computed by the Host.
- Most compute intensive layer (90% of total computation). Requires transfer of dynamic variables. Computed by the FPGA.
Optimization Strategies

- **Opt. 1: Most Naïve implementation**
  - Invoke the MAP function each simulation cycle
  - Non-overlapping computation of both network levels
  - Block RAM (1.1 MB) used; smaller network (10,000 neurons)

- **Opt. 2: POSIX Threads**
  - MAP function called once for entire simulation
  - Overlapping computation of both network levels

- **Opt. 3: OBM banks instead of Block RAM**
  - Network now scales up to 2 million neurons

- **Opt. 4: Streaming technique for data communication**
  - Firing information communicated with streaming technique instead of OBM banks

- **Opt. 5: Loop unrolling and multiple pipelines**
### Optimization Strategies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Izhikevich</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wilson</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hodgkin-Huxley</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNN Model</th>
<th>Logic (%)</th>
<th>OBM (Total=16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Izhikevich</td>
<td>66</td>
<td>16</td>
</tr>
<tr>
<td>Wilson</td>
<td>49</td>
<td>12</td>
</tr>
<tr>
<td>Hodgkin-Huxley</td>
<td>95</td>
<td>12</td>
</tr>
</tbody>
</table>
Performance

Izhikevich Model

Wilson Model

Hodgkin-Huxley Model
Speed-Up: Izhikevich

![Graph showing speedup versus number of neurons (thousands)]
Speed-Up: Wilson

![Graph showing speedup vs. neurons (thousands)]
Speed-Up: Hodgkin-Huxley
Effect of Optimizations

Izhikevich Model

Wilson Model

Hodgkin-Huxley Model
Optimizations: Izhikevich

- Serial comp
- Parallel Comp
- Streaming comm
- Two loop unrolling
- Three pipelines

Speedup
Optimizations: Wilson

![Graph showing speedup for Serial comp, Parallel Comp, and Streaming comm in the Wilson Model. The graph indicates significant improvement in speedup for Parallel Comp and Streaming comm compared to Serial comp.]
Optimizations: Hodgkin-Huxley

Hodgkin-Huxley Model

Speedup

Serial comp  Parallel Comp  Streaming comm
Energy Efficiency

- Only using one of the two FPGAs
- Compared with Intel dual core Xeon (80 watt)
- 2 million Izhikevich neurons
  - 5 fold power savings
- 0.5 million Wilson neurons
  - 2.9 fold power savings
- 0.5 million HH neurons
  - 19 fold power savings
Conclusions

• SRC system’s HW/SW approach can simulate
  – Large biological neural network
  – 2 million Izhikevich neurons
  – 0.5 million Wilson and HH neurons

• Acceleration with SRC H MAP
  – 10x - Izhikevich
  – 6x - Wilson
  – 38x - HH

• Optimization strategies and their effects studied
• Memory is the current bottleneck
• Varying energy efficiencies
Future Work

• Improve performance with both FPGAs
  – May be limited by memory
• Investigate other optimizations that can
  – Avoid performance saturation
  – Simulate larger networks
• Multi-node SRC system
  – To enable biological scale networks
Thank you

(mbhuuiya, anallam, smithmc,kpallip)@clemson.edu