Real-time cortical simulations: energy and interconnect scaling on distributed systems (ARM and INTEL cores)

F. Simula, E. Pastorelli, P.S. Paolucci, M. Martinelli, A. Lonardo, A. Biagioni, C. Capone, F. Capuani, P. Cretaro, G. De Bonis, F. Lo Cicero, L. Pontisso, P. Vicini, R. Ammendola

On their behalf

Pier Stanislao Paolucci,
the APE parallel/distributed computing lab - INFN Roma

Highlighting the importance of low latency interconnects for energy efficient application of bio-intelligence (spiking networks) in real-time embedded scenarios, after you took care of the usual suspect (memory locality)
Summary

- Addressing real-time spiking simulations highlights the importance of low-latency interconnects for distributed computing platforms (small-packets), after you took care of the usual suspect (memory locality) and implemented axonal like MPI interconnect structure.

- Comparison of speed and energy consumption of Intel and Arm platforms on real-time cortical simulations.
- Strong and Weak scalings on Intel and Arm platforms.

- Metrics:
  - joule per simulated synaptic event
  - real-time ratio

- Perspective: application of thalamo-cortical spiking bio-intelligence to embedded real-time scenarios: e.g.
  - Sleep-like slow oscillations improve image classification during wakefulness. arXiv:1810.10498
Example of motivation for (real-time) simulations of thalamo-cortical plastic networks:
model sleep effects and transition to wakefulness

- Sleep is present in all animal species
- Young humans pass the majority of time sleeping, exactly when learning in faster
- Sleep deprivation is detrimental for cognition
- On a small-scale model, we have evidence of improvement of post-sleep classification of images (MNIST handwritten characters), see our:
Starting point, presented at PDP2018

Pastorelli et al (2018) PDP 2018
Gaussian and exponential lateral connectivity on distributed spiking neural network simulations

On our DPSNN simulation engine, we measured and presented @PDP2018 the (manageable) impact of biologically plausible dominant long-range connectivity (75% remote, 25% local) on the simulation of large scale cortical spiking models (1cm^2), at realistic synapses per neuron (thousands syn/neu) and realistic neural densities (45K neurons / mm^2), sleep-like and wake-like states

<table>
<thead>
<tr>
<th>GRID</th>
<th>COLUMNS</th>
<th>NEURONS</th>
<th>Number of SYNAPSES</th>
<th>hardware cores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gaussian shorter range interconnect</td>
<td>Exponential longer range interconnect</td>
</tr>
<tr>
<td>24x24</td>
<td>576</td>
<td>0.7 M</td>
<td>1.2 G</td>
<td>1.8 G</td>
</tr>
<tr>
<td>48x48</td>
<td>2304</td>
<td>2.9 M</td>
<td>3.5 G</td>
<td>5.9 G</td>
</tr>
<tr>
<td>96x96</td>
<td>9216</td>
<td>11.4 M</td>
<td><strong>14.2 G</strong></td>
<td><strong>23.4 G</strong></td>
</tr>
</tbody>
</table>

Shnepel et al (2015) Cerebral Cortex Horizontal Connections...

motivations for previous PDP2018 study

Stepanyants et al. (2009)PNAS
The fraction of short- and long-range connections ...
DPSNN engine: mixed event and time driven execution flow.

Exploits memory locality (neighbouring neurons and incoming synapses on the same process) + axonal arborization inspired 3 step MPI strategy
Simulation requiring realistic intra-areal connections: cortical slow wave (deep sleep) activity, single area simulated at high resolution.

Cortical area, described using a two-dimensional grid of cortical column.

Thousands of spiking neurons per column (excitatory and inhibitory). Leaky integrate and fire with spike frequency adaptation.

Thousands of synapses per neuron.

Simulation of a large field of cortical columns (pixels of the bottom snapshots).

Top-right panel: firing rate of the central column (green) of the cortical field and the net synaptic input it receives from neighboring columns (blue): local vs global contribution.

Parameters of the theoretical model defined in cooperation with ISS (M. Mattia, P. Del Giudice).

Capone, Rebollo et al. (2017) Cerebral Cortex. Slow Waves...
Strong scaling, deep sleep waves, DPSNN simulation engine (@PDP2018), exploiting memory locality (+ clever axon like communication strategy) is enough at large scale

For large scale networks good scaling, but were still at least 12-30 times slower than real time. Asynchronous spiking regime @3Hz, single compartment neuron: Leaky Integrate and Fire with Spike Frequency adaptation. Instantaneous current synapses.

Execution Platform:
GALILEO @ CINECA
Up to 64 IBM Nodes, 1024 cores.

Two Intel Xeon Haswell 8-core E5-2630 V3 processor per node @ 2.4 GHz.

Infiniband network, 4x QDR switches. Hyper-threading off.
Impact of longer-range, more biologically plausible, exponential connectivity, still manageable on DPSNN engine (@PDP2018), for large scale simulations (not aiming to real time).

(75% remote synapses, 25% local synapses)

Long range
Intra-areal later connectivity adds a factor 2 cost to simulations

Clear hint for the cost of transporting spiking messages
PDP2019:
For smaller networks, can we reach real-time?

Yes, but see what happens for smaller networks already at small number of processes

**Red-line: soft real-time threshold**

Average firing rate: 3 Hz
Asynchronous regime

Leaky-Integrate and Fire neuron with Spike-Frequency Adaptation
80% exc, 20% inh

Small net -> target synapses distributed among all processes
Communication of spikes is the main cost

Red line: real time threshold
AER representation of spiking messages, 12B/spike:
(Spiking Neuron Id, Spiking Time)

Synchro step (messages) every 1 simulated ms
Small inter-processor packets size.
Estimate: bytes/packet @ 32 processes

\[
(20480\text{neu} \times 3\text{Hz} \times 12\text{B/spike}) / (32\text{processes} \times 1000\text{step/s}) = 23\text{ B/packet}
\]

KEY LESSON: DESIGN OF LOW LATENCY INTERCONNECTS FOR SMALL PACKETS ESSENTIAL
And for ARM based platform (prototype of ExaNeSt systems)?

Profiling on Trenz - 20480 neu, 23M syn

TEBF0808 Trenz Boards equipped with Zync Ultrascale+ MPSoC.

Each Zync includes quad-core ARM Cortex-A53

Similar need for low latency, interconnect optimized for small packets
Energy to solution: Intel

Sim. of 10s of activity of 20480 neuron, 23 Mega Syn, asynchr. regime, 3Hz spiking rate

<table>
<thead>
<tr>
<th>x86 cores</th>
<th>Time (s)</th>
<th>Power (W)</th>
<th>Energy to solution (J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>150.9</td>
<td>48</td>
<td>7243.2</td>
</tr>
<tr>
<td>2 HT</td>
<td>121.8</td>
<td>53</td>
<td>6455.4</td>
</tr>
<tr>
<td>2</td>
<td>80.7</td>
<td>62</td>
<td>5003.4</td>
</tr>
<tr>
<td>4</td>
<td>37.4</td>
<td>92</td>
<td>3440.8</td>
</tr>
<tr>
<td>8</td>
<td>25.3</td>
<td>124</td>
<td>3137.2</td>
</tr>
<tr>
<td>16</td>
<td>26.1</td>
<td>166</td>
<td>4332.6</td>
</tr>
<tr>
<td>32 plus ETH</td>
<td>30.0</td>
<td>342</td>
<td>10260.0</td>
</tr>
<tr>
<td>32 plus IB</td>
<td>19.7</td>
<td>318</td>
<td>6264.6</td>
</tr>
<tr>
<td>64 plus ETH</td>
<td>69.3</td>
<td>531</td>
<td>36798.3</td>
</tr>
<tr>
<td>64 plus IB</td>
<td>32.1</td>
<td>501</td>
<td>16082.1</td>
</tr>
</tbody>
</table>
Energy to solution: ARM

Sim. of 10s of activity of 20480 neuron, 23 Mega Syn, asynchr. regime, 3Hz spiking rate

**TABLE III**

<table>
<thead>
<tr>
<th>ARM cores</th>
<th>Time (s)</th>
<th>Power (W)</th>
<th>Energy to solution (J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>636.8</td>
<td>2.2</td>
<td>1273.6</td>
</tr>
<tr>
<td>2</td>
<td>334.1</td>
<td>3.4</td>
<td>1135.9</td>
</tr>
<tr>
<td>4</td>
<td>185.0</td>
<td>6.0</td>
<td>1110.0</td>
</tr>
<tr>
<td>8</td>
<td>133.8</td>
<td>10</td>
<td>1338.0</td>
</tr>
</tbody>
</table>
INTEL vs ARM on DPSNN neural simulation

TABLE IV
Comparison of energetic efficiencies.

<table>
<thead>
<tr>
<th>DPSNN simulator</th>
<th>Compass/TrueNorth sim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARM</td>
<td>Intel</td>
</tr>
<tr>
<td>1.1 (μJ / syn event)</td>
<td>3.4 (μJ / syn event)</td>
</tr>
<tr>
<td>Intel</td>
<td>5.7 (μJ / syn event)</td>
</tr>
</tbody>
</table>

- 4 ARM cores about 3x greener than 4 Intel cores ...
- ... but 4 ARM cores about 5x times slower than 4 Intel
Conclusions / Acknowledgements

- Low Latency interconnects for small packets essential to reach real-time simulation of spiking neural networks in asynchronous (and sleeping) regimes (after you took care of the usual suspect, memory locality, and exploited axonal arborization like MPI structure)

- 4 ARM cores about 3x greener than 4 Intel cores ...

- ... but 4 ARM cores about 5x times slower than 4 Intel ARM

Acknowledgements:

- The Human Brain Project, EU grant No. 785907 (HBP SGA2)
- The ExaNeSt Project, EU grant No. 671553
- INFN-CINECA Computational Theoretical Physics Collaboration
Sleep Memories interaction

Cristiano Capone
Pier Stanislao Paolucci

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
Introduction on Sleep Questions

- 1) Sleep role in *biological intelligence*
  - Homeostatic processes
  - Optimization of performances
  - Optimization of energy consumption

- 2) Optimization mechanisms to use in *artificial intelligence*:
  - Mechanisms similar to the ones already used in AI algorithms
    - Regularizations
    - Batch normalization
  - New mechanisms ????
Additional material on sleep-memory interaction
A minimal thalamo-cortical spiking model is considered. Population are composed of AdEx neurons. This model is exposed to static images (e.g. MNIST dataset). It can be trained to store and recall them.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
Sleep in a minimal thalamo-cortical model

During “deep sleep” the contextual signal and visual signal are absent.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
Sleep induced Synaptic association

- Synapses between neurons coding for examples in the same class are **potentiated**.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
- Synapses between neurons coding for examples in the same instance are *depressed*.
- Synapses between neurons coding for examples in the same class are *potentiated*.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
Classification task

- This architecture can be trained to classify images (MNIST dataset).
- The learning is incremental.
- The accuracy is 93% after 30 training examples per class.
- And already 85% after 10.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
- Sleep progressively improves performances in a visual recognition task.
- The effect is stronger in presence of thalamic feedback.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
- The minimal model reproduces dynamical properties observed in experiments.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]
- Synapses between neurons coding for examples in the same class are **potentiated**.

[Capone, Pastorelli, Golosio, Paolucci, arXiv. 1810.10498v4]