A NETWORK SIMULATOR FOR THE ESTIMATION OF BANDWIDTH LOAD AND LATENCY CREATED BY HETEROGENEOUS SPIKING NEURAL NETWORKS ON NEUROMORPHIC COMPUTING COMMUNICATION NETWORKS

ROBERT KLEIJNEN¹, MARKUS ROBENS¹, MICHAEL SCHIEK¹, STEFAN VAN WAASEN¹,²

¹CENTRAL INSTITUTE OF ENGINEERING, ELECTRONICS AND ANALYTICS - ELECTRONIC SYSTEMS (ZEA-2), FORSCHUNGSZENTRUM JÜLICH GMBH, JÜLICH, GERMANY
²FACULTY OF ENGINEERING, COMMUNICATION SYSTEMS, UNIVERSITY OF DUISBURG-ESSEN, GERMANY

MCSoC-2021 SINGAPORE
SESSION 7-A: EMBEDDED NEUROMORPHIC COMPUTING SYSTEMS

Email: r.kleijnen@fz-juelich.de
OVERVIEW

- Introduction
  - Aims of the “Advanced Computing Architectures” (ACA) – project
  - General aspects of neuromorphic computing systems
- The Python based network simulator
- Estimated network load for homogeneous neural networks
- Heterogeneous neural networks
  - An ACA-project test case: Cortical Microcircuit Model
  - Impact of neuron mapping
- Summary and Outlook
INTRODUCTION

Aims of the “Advanced Computing Architectures” (ACA) – Project [1]

Development of a novel modular (digital) NC simulation platform for computational neuroscience

- Biological connectivity level – on average 10,000 synapses per neuron
- Large scale – aiming towards human/mammal brain size \[\approx 10^{11}\] neurons
- Simulated biological time step 0.1 ms
- 100x faster than biological real time simulation
- Perform and explore online learning by simulating different learning rules

[1] https://www.fz-juelich.de/aca/EN/Home/home_node.html
INTRODUCTION

General aspects of Neuromorphic Computing Systems

Two main components of a NC system:

• A large number of processing units, each simulating a (small) number of neurons* and their input synapses

• A communication network to connect all nodes

Spike events are communicated as spike packets

⇒ small information packets containing a neuron address and potentially a number of header and debug bits

Resulting communication traffic is a very large number of small information packets.

We focus on the communication task at hand

⇒ Computation/simulation of neuron and synapse behavior not considered in our simulator

*In this work, the number of neurons per node, $N_pN$, was set to 100
THE PYTHON BASED NETWORK SIMULATOR

How to simulate the network load

1. Defined NN, either fixed as a netlist or as connectivity statistics

   Netlist:
   1: 2
   2: 4 6
   3: 2 5
   4: 1 2
   5: 4
   6: 3 5
   7: 5 6 7

2. Create a graph to represent the communication network

   Different graphs can be created to represent different network topologies

3. Specify the location of the neurons in the graph, i.e. mapping of the neurons

   Necessary step for heterogeneous NNs

4. Mimic a spike event for every neuron and simulate the movement of the resulting spike packets through the network

   Spike movement depends heavily on the used casting and routing scheme
NETWORK LOAD OF HOMOGENEOUS CONNECTIVITY MODELS

Comparison to Analytical Model

\[ BW_{avg} = \frac{n \cdot N_{targets} \cdot D_{avg}}{L} \cdot FR_{avg} \] \[2\]

Comparison done with a randomly connected neural network (RNDC NN), with an uniform connection probability \( \epsilon = 0.048 \), \( FR_{avg} = 1 \) and \( NpN = 100 \)

Calculated mean corresponds well to the simulated mean value

NETWORK LOAD OF HOMOGENEOUS CONNECTIVITY MODELS

Advantages of the simulator over the Analytical Model

• The analytical model is not able to consider the variations due to boundary effects, or irregular topologies.
  o Maximum network load is a critical metric as it will form the bottleneck of the system
• $N_{targets}$ and $D_{avg}$ are difficult to define for certain connectivity models and casting types.
HETEROGENEOUS NEURAL NETWORK

ACA-project test case: The Cortical Microcircuit Model [3]

NN consists of:

- Four layers with an excitatory and inhibitory population each
- One separate TC population

Each population includes a different number of neurons and has its own specific connection probability to other populations.

Connection Probability $C_{X,Y}$

<table>
<thead>
<tr>
<th>Source Population ($X$)</th>
<th>Population Size # neurons</th>
<th>Target population ($Y$)</th>
<th>L2/3E</th>
<th>L2/3I</th>
<th>L4E</th>
<th>L4I</th>
<th>L5E</th>
<th>L5I</th>
<th>L6E</th>
<th>L6I</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2/3E</td>
<td>20683</td>
<td>L2/3E</td>
<td>0.1009</td>
<td>0.1346</td>
<td>0.0077</td>
<td>0.0691</td>
<td>0.1004</td>
<td>0.0548</td>
<td>0.0156</td>
<td>0.0364</td>
</tr>
<tr>
<td>L2/3I</td>
<td>5834</td>
<td>L2/3I</td>
<td>0.1689</td>
<td>0.1371</td>
<td>0.0059</td>
<td>0.0029</td>
<td>0.0622</td>
<td>0.0269</td>
<td>0.0066</td>
<td>0.0010</td>
</tr>
<tr>
<td>L4E</td>
<td>21915</td>
<td>L4E</td>
<td>0.0437</td>
<td>0.0316</td>
<td>0.0497</td>
<td>0.0794</td>
<td>0.0505</td>
<td>0.0257</td>
<td>0.0211</td>
<td>0.0034</td>
</tr>
<tr>
<td>L4I</td>
<td>5479</td>
<td>L4I</td>
<td>0.0818</td>
<td>0.0515</td>
<td>0.1350</td>
<td>0.1597</td>
<td>0.0057</td>
<td>0.0022</td>
<td>0.0166</td>
<td>0.0005</td>
</tr>
<tr>
<td>L5E</td>
<td>4850</td>
<td>L5E</td>
<td>0.0323</td>
<td>0.0755</td>
<td>0.0067</td>
<td>0.0033</td>
<td>0.0831</td>
<td>0.0600</td>
<td>0.0572</td>
<td>0.0277</td>
</tr>
<tr>
<td>L5I</td>
<td>1065</td>
<td>L5I</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0003</td>
<td>0.0</td>
<td>0.3726</td>
<td>0.3158</td>
<td>0.0197</td>
<td>0.0080</td>
</tr>
<tr>
<td>L6E</td>
<td>14395</td>
<td>L6E</td>
<td>0.0076</td>
<td>0.0042</td>
<td>0.0453</td>
<td>0.1057</td>
<td>0.0204</td>
<td>0.0086</td>
<td>0.0396</td>
<td>0.0658</td>
</tr>
<tr>
<td>L6I</td>
<td>2948</td>
<td>L6I</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2252</td>
<td>0.1443</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>902</td>
<td>TC</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0983</td>
<td>0.0619</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0512</td>
<td>0.0196</td>
</tr>
</tbody>
</table>

Randomly mapping the cortical microcircuit model yields a similar result as a RNDC NN with an equivalent average connectivity.

Even a rudimentary mapping, such as sequential, has a significant impact on the network load.
HETEROGENEOUS NEURAL NETWORK

Impact of Neuron Mapping

Network load distribution for different mapping schemes

Randomly mapped “Cortical Microcircuit” model

Sequentially mapped “Cortical Microcircuit” model

Sequential mapping of the “Cortical Microcircuit” model
SUMMARY & OUTLOOK

Introduced a network simulator, specifically designed to simulate NC communication networks:

- Simulates network bandwidth
- Estimates spike packet latency

Unique features:

- Analyse network load distribution caused by non-uniformities in the communication network
- Simulation of biological realistic heterogeneous NNs (derived from scientific publications)
- Potential to evaluate different neuron mapping algorithms

Future steps

- Use the simulator in combination with the Multi area model [4], a significant larger test case, to evaluate different topologies and routing- & casting types
- Develop communication concepts, able to facilitate the accelerated simulation of large scale neural networks