Performance Benchmarks for Neuromorphic Systems at Scale

Johanna Senk[1], **Sacha J. van Albada**[1,2], **Markus Diesmann**[1,3,4]

1. Institute of Neuroscience and Medicine (INM-6), Institute for Advanced Simulation (IAS-6), JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany

2. Institute of Zoology, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany

3. Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany

4. Department of Psychiatry, Psychotherapy and Psychosomatics, School of Medicine, RWTH Aachen University, Aachen, Germany

Summary. A neuroscientifically relevant model of a local cortical circuit has become a benchmark challenging different conventional and neuromorphic simulation approaches in terms of accuracy and efficiency. The cross-disciplinary performance comparison calls for discussions on 1) future benchmark models representing the multi-level organization of the brain at scale, and 2) demands on the design of neuromorphic hardware to enable corresponding network simulations.

In the development of neuromorphic systems, the brain not only provides design criteria but also sets the ultimate bar for the performance of these systems. While neuroscientists are still in the process of uncovering basic principles of brain function, it is their responsibility to formulate plausible constraints for software and hardware emulation [1, 2]. To test whether neuromorphic hardware systems fulfill these constraints, performance benchmarks with neuroscientifically relevant network models need to be conceived, implemented, and executed. A network model suited for this purpose should account for brain structures and dynamics at realistic spatial and temporal scales. The biological mechanisms described by the model should be sufficiently understood by the field and recognized as both fundamental and generic. The model should provide potential for extension and further development such as upscaling to larger networks and inclusion of more complex features. It is often useful for the hardware and software requirements to be moderate to enable routine simulations with off-the-shelf computing systems.

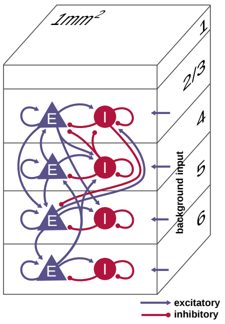

Figure 1: Cortical microcircuit model; taken from [3].

A universal building block for brain-like computing is a local cortical microcircuit: Its network architecture below 1 mm^2 of cortical surface is similar across cortical areas and for different mammalian species, from mouse to human. At natural density, this network has on the order of 10^4–10^5 neurons, each of which has 10^3–10^4 connections. Its connection probability of about 0.1 is an upper bound; larger cortical networks are less densely connected. A prototype computational model for this circuit [4] represents each cortical layer by an excitatory and an inhibitory population of leaky integrate-and-fire neuron models with cell-type-specific recurrent connectivity derived from experimental data (Fig. 1). The model has been employed in a number of studies focusing on neuroscientific questions, and there exist multiple implementations for different simulators. Recently, it has evolved into a standard model for comparing the performance of different simulation technologies.

Fully deterministic simulation engines, which combine conventional high-performance computing systems with dedicated simulation software, serve as reference technology. Neuroscience can here benefit from the continuous advancement of flexible, general-purpose hardware driven by other domains. Reference simulations establish a baseline in terms of accuracy and efficiency to be used for verification and validation. The simulation results of the microcircuit model can only be compared on a statistical level due to inherent differences between simulators regarding algorithms, numerical resolutions, or random number generators. Moreover, the network dynamics is chaotic and minimal deviations may amplify. We analyzed the spiking activity of the same model simulated with NEST on CPUs for reference and with the neuromorphic hardware system SpiNNaker [3]. A good match between the simulation results was obtained once SpiNNaker had overcome certain challenges, as the design specifications of the neuromorphic hardware differed from the model demands (e.g., a small integration time step and a high number of connections per neuron). We further compared the time-to-solution and energy-to-solution as real-time or even accelerated simulations at low power consumption are imperative for future simulation-based research. At that time, neither technology enabled simulating the microcircuit in real time, and the required power exceeded the demands of the natural brain by orders of magnitude. Our study was soon picked up by the community, and other simulation approaches for neuronal network models were compared against our results: The microcircuit model was evaluated with similar metrics using the GPU-based simulators GeNN [5, 6] and NEST GPU [7], and an FPGA-based computing system [8]. Further simulator development has led to improved efficiency of SpiNNaker [9] and NEST [10].

This systematic comparison brings out advantages and disadvantages of different simulation technologies with respect to the benchmark model. Creative algorithmic strategies have been developed to make best use of the respective hardware and likewise to cope with limitations. These efforts fuel discussions on topics such as number representations and resolutions [11], or the possibility of offloading certain computations to specialized hardware. In addition, the comparison has led to a performance gain for the community: In only a few years, the milestone of real-time simulation has been reached and surpassed for the microcircuit model.

We consider the benchmarking endeavors around the microcircuit model as a starting point for a cross-disciplinary co-development of simulation technologies and neuroscientific models. The complexity of the brain calls for complementary benchmark models with different biological detail and computational demands such that large-scale neuromorphic hardware systems will not be optimized for only a single model type. One next challenge is to scale up from local to brain-size networks: A multi-area model of the visual system of macaque monkey uses an adapted version of the microcircuit model for each of the 32 areas represented [12]. This model can already be simulated with NEST, NEST GPU [13], and GeNN [14] but not yet in real time. Slow processes like development and long-term learning still cannot be studied in large-scale models even with state-of-the-art simulators. Another challenge therefore consists in devising and maturing additional representative benchmark models (in particular functional models), which require different metrics for evaluating the performance as well as unified benchmarking tools for running simulations and comparing results [2].

Computational neuroscientists and simulation system developers symbiotically benefit from diversity in both the types of neuronal network models studied and the simulation technologies developed. On the road towards understanding the brain and at the same time making use of the gathered knowledge to advance technology, we emphasize the importance of points of convergence for different disciplines to come together and learn from each other through rigorously defined and executed performance benchmarks.

References

[1] J. Senk et al. "Connectivity concepts in neuronal network modeling". In: *PLOS Computational Biology* 18.9 (2022), e1010086.

[2] J. Albers et al. "A Modular Workflow for Performance Benchmarking of Neuronal Network Simulations". In: *Frontiers in Neuroinformatics* 16 (2022).

[3] S. J. van Albada et al. "Performance Comparison of the Digital Neuromorphic Hardware SpiNNaker and the Neural Network Simulation Software NEST for a Full-Scale Cortical Microcircuit Model". In: *Frontiers in Neuroscience* 12 (2018).

[4] T. C. Potjans and M. Diesmann. "The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model". In: *Cerebral Cortex* 24.3 (2014), pp. 785–806.

[5] J. C. Knight and T. Nowotny. "GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model". In: *Frontiers in Neuroscience* 12 (2018).

[6] J. C. Knight, A. Komissarov, and T. Nowotny. "PyGeNN: A Python Library for GPU-Enhanced Neural Networks". In: *Frontiers in Neuroinformatics* 15 (2021).

[7] B. Golosio et al. "Fast Simulations of Highly-Connected Spiking Cortical Models Using GPUs". In: *Frontiers in Computational Neuroscience* 15 (2021).

[8] A. Heittmann et al. "Simulating the Cortical Microcircuit Significantly Faster Than Real Time on the IBM INC-3000 Neural Supercomputer". In: *Frontiers in Neuroscience* 15 (2022).

[9] O. Rhodes et al. "Real-time cortical simulation on neuromorphic hardware". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378.2164 (2019), p. 20190160.

[10] A. C. Kurth et al. "Sub-realtime simulation of a neuronal network of natural density". In: *Neuromorphic Computing and Engineering* 2.2 (2022), p. 021001.

[11] S. Dasbach et al. "Dynamical Characteristics of Recurrent Neuronal Networks Are Robust Against Low Synaptic Weight Resolution". In: *Frontiers in Neuroscience* 15 (2021).

[12] M. Schmidt et al. "A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas". In: *PLOS Computational Biology* 14.10 (2018), e1006359.

[13] G. Tiddia et al. "Fast Simulation of a Multi-Area Spiking Network Model of Macaque Cortex on an MPI-GPU Cluster". In: *Frontiers in Neuroinformatics* 16 (2022).

[14] J. C. Knight and T. Nowotny. "Larger GPU-accelerated brain simulations with procedural connectivity". In: *Nature Computational Science* 1.2 (2021), pp. 136–142.