NEST GPU simulations scale up to networks of billions of spiking neurons and trillions of synapses

José Villamar[1,2], Gianmarco Tiddia[3], Luca Sergi[3,4], Pooja Babu[1,5], Luca Pontisso[6], Francesco Simula[6], Alessandro Lonardo[6], Elena Pastorelli[6], Pier Stanislao Paolucci[6], Bruno Golosio[3,4], Johanna Senk[1,7]

1. Institute for Advanced Simulation (IAS-6), Jülich Research Centre, Jülich, Germany
2. RWTH Aachen University, Aachen, Germany
3. Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Cagliari, Monserrato, Italy
4. Department of Physics, University of Cagliari, Monserrato, Italy
5. Simulation and Data Laboratory Neuroscience, Jülich Supercomputing Centre, Jülich Research Centre, Jülich, Germany
6. Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Roma, Italy
7. Sussex AI, School of Engineering and Informatics, University of Sussex, Brighton, United Kingdom

Efficient simulation of large-scale spiking neuronal networks is important for neuroscientific research, and both the simulation speed and the time it takes to instantiate the network in computer memory are key factors. NEST GPU is a GPU-based simulator under the NEST Initiative written in CUDA-C++ that demonstrates high simulation speeds with models of various network sizes on single-GPU and multi-GPU systems [1–3]. On the path toward models of the whole brain, neuroscientists show an increasing interest in studying networks that are larger by several orders of magnitude compared to local circuits. Here, we show the performance of our simulation technology with a scalable network model across multiple network sizes approaching human cortex magnitudes. For this, we propose a novel method to efficiently instantiate large networks on multiple GPUs in parallel. Our approach relies on the deterministic initial state of pseudo-random number generators (PRNGs). While requiring synchronization of network construction directives between MPI processes and a small memory overhead, this approach enables dynamical neuron creation and connection at runtime. The method is evaluated through a two-population recurrently connected network model designed for benchmarking an arbitrary number of GPUs while maintaining first-order network statistics across scales. The benchmarking model was tested during an exclusive reservation of the LEONARDO Booster cluster. While keeping constant the number of neurons and incoming synapses to each neuron per GPU, we performed several simulation runs exploiting in parallel from 400 to 12,000 (full system) GPUs. Each GPU device contained approximately 281 thousand neurons and 3.1 billion synapses. Our results show network construction times of less than a second using the full system and stable dynamics across scales. At full system scale, the network model was composed of approximately 3.37 billion neurons and 37.96 trillion synapses (∼25% human cortex). To conclude, our novel approach enabled network model instantiation of magnitudes nearing human cortex scale while keeping fast construction times, on average of 0.5 s across trials. Note, however, that the model used here only establishes local connections instantiated between neighboring groups of GPUs while long-range or inter-area connections are more costly to create and require significantly longer construction times. The stability of dynamics and performance across scales obtained in our model is a proof of feasibility paving the way for biologically more plausible and detailed brain scale models.

Acknowledgements

ISCRA for awarding access to the LEONARDO supercomputer (EuroHPC Joint Undertaking) via the BRAINSTAIN - INFN Scientific Committee 5 project, hosted by CINECA (Italy); HiRSE_PS, Helmholtz Platform for Research Software Engineering - Preparatory Study (2022-01-01 - 2023-12-31), the Horizon Europe Grant 101147319, Joint lab SMHB; FAIR CUPI53C22001400006 Italian PNRR grant.

References

[1] B. Golosio et al. “Fast Simulations of Highly-Connected Spiking Cortical Models Using GPUs”. In: *Frontiers in Computational Neuroscience* 15 (2021). ISSN: 1662-5188.

[2] B. Golosio et al. “Runtime Construction of Large-Scale Spiking Neuronal Network Models on GPU Devices”. In: *Applied Sciences* 13.17 (2023), p. 9598.

[3] 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). ISSN: 1662-5196.