Record Simulation of the Full-Density Spiking Potjans-Diesmann-Microcircuit Model on the IBM Neural Supercomputer INC 3000

Arne Heittmann[1], Georgia Psychou[1], Guido Trensch[2], Markus Diesmann[4,5,6], Winfried Wilcke[3], Charles Cox[3] and Tobias G. Noll[1]

Motivation: This contribution reports on the results of a study conducted in the framework of the "*Advanced Computing Architectures (ACA) - towards multi-scale natural-density Neuromorphic Computing*" project at Forschungszentrum (FZJ) Juelich, Germany. The ACA project is a joint cooperation of several institutes at FZJ, RWTH Aachen University, Germany, The University of Manchester, Great Britain, and Heidelberg University, Germany. It is targeting the Neuroscience simulation application area as a pilot project preparing a long-term Neuromorphic Computing research initiative. Its main goal is the specification of a future Neuromorphic Computing architecture, including the definition of requirements and target performances, the development of workflows for a systematic validation and benchmarking of neuromorphic architectures, and the development of efficient Neuromorphic Computing concepts.

Currently, full-scale, biologically plausible, spiking microcircuit models, containing about 100,000 neurons with roughly 10,000 synapses each (i.e., with natural density), can be simulated at approximately biological real time (BRT) [Knight, 18, Rhodes, Kurth]. Very-large-scale, so-called scaled multi-area models [Schmidt], consisting of 32 microcircuit equivalents, can be simulated approximately 30-times slower than BRT [v.Albada, 20, Knight, 21]. This is sufficiently fast to study the interplay between local and global dynamics [v. Albada, 20]. But these are static networks, only, and adding plasticity enabling system-level learning – crucial for understanding learning - , will slow down the simulation of future more advanced models even further. This is relevant as learning unfolds over long stretches of biological time [Stapmanns]. Consequently, the goal of this study is to identify the simulation-speed bottlenecks of state-of-the art systems and architectures, to find remedies to it and to search for ways how to bring new architectures optimized for brain simulation far ahead of general purpose supercomputer technologies.

Materials: The motivation for the creation of the IBM Neural Supercomputer (INC) family originated as part of the IBM Artificial General Intelligence (AGI) project at IBM Research in Almaden, California. Two INC-3000 systems have been built so far, one for IBM Almaden and the other for Forschungszentrum Juelich, Germany. The IBM INC-3000 system consists of sixteen 14" x 22" INC cards, each hosting 27 software-programmable and field-reconfigurable Xilinx-SoC nodes ZYNQ XC7Z045 with 1-GB DDR SDRAM. Each SoC consists of a programmable logic (*PL*, 218,600 reconfigurable look-up tables and 437,200 flip-flops) and a processing system (*PS*, 2 ARM Cortex-A9 cores) as well as 16 serial transceivers with up to 12.5 Gb/s data rate in both directions and state-of-the-art node-to-node latency (1st-bit-to-1st-bit) of about 1 μs. Using these transceivers, on one INC card the 27 nodes are interconnected by a 3x3x3 mesh network topology and for the whole INC-3000 system the 432 nodes are interconnected by a 12x12x3 mesh network topology via a thick backplane.

The cortical microcircuit model [Potjans] has become a benchmark network for the comparison of spiking neural network simulations [e.g., v. Albada, 18, Knight,18, Rhodes, Kurth] and represents approximately a 1-mm^2 patch of early sensory cortex. It consists of a balanced four-layer network of 8 populations with about 80,000 spiking leaky integrate-and-fire (LIF) neurons (80% excitatory and 20% inhibitory) in total. The four layers correspond to L2/3, L4, L5, and L6 in the biological neo-cortex. This is the smallest network which combines a realistic number of about 10,000 static synapses per neuron with a connectivity of 10%. The connectivity is cell-type specific but laterally uniformly random.

Methods: Initially, the INC concept was designed to yield high performance for rate-coded artificial neural networks. But it turns out, that these machines are also an ideal playground for the elaboration of concepts towards future dedicated accelerators for biologically plausible neural networks in Computational Neuroscience. The cortical microcircuit was implemented on the INC-3000 system for simulations on a 0.1-ms time grid as follows:

The sole goal of these design experiments was to explore the ultimate performance limits of the INC-3000 machine, regardless of any efficiency aspects. Consequently, only the programmable logic parts of the SoC nodes are used with maximum parallelism. Instead of the 1 GB SDRAM, only on-SoC low-latency block RAMs (BRAMs) are used as storage for the synapse data. All arithmetic is implemented with single-floating point precision and high-level synthesis with a target clock frequency of 150 MHz was applied in order to allow for quick design-space-exploration experiments.

AER (Adress Event Representation) spike packets are communicated by deterministic routing via a central master node and broadcasting. The processes on the nodes are synchronized by barrier messages, also using this communication infrastructure. In addition to the LIF neurons and exponential-decay shaped CUBA-synapse models of the original Potjans-Diesmann model, also Izhikevich and AdEx models and alpha- and beta-shaped COBA-synapses were implemented. For advancing the state of the neuron models on the time grid, Forward-Euler-, Exact-Exponential-, Runge-Kutta-, and Parker-Sochacki-ODE-solver methods, are applied. As BRAMs are quite limited resources on the SoCs, for the sake of a minimal memory footprint and in order to shorten simulation-set-up time as much as possible, procedural network generation, as already proposed in [Roth], is used. Thereby, synaptic connections are drawn in a *deterministic* fashion using pseudo-random number generators again and again during simulation. Variants of *Walker's alias methods* [Walker] are applied as highly-optimized PRNGs.

Correctness of the simulation results was successfully verified as described in [v.Albada, 18], i.e. by comparing the probability distributions of average firing rates, coefficients of variation, and Pearson's correlation coefficients with that derived from reference simulations using NEST [Gewaltig].

Results: Under these constraints, it is optimal to have each INC-3000 node to host about 256 neuron models, resulting in the use of 305 nodes, equivalent to 70.6 % of the INC-3000 machine. The maximum Manhattan distance in the 12x12x3 mesh network is 24 hops and placing the master node in the center of the 305-node cluster results in a round-trip latency of about 18 μs. So, for a 0.1-ms time grid, the maximum speed-up over BRT is 100 μs / 18 μs = 5.56 X. The measured simulation time for 1 second of BRT is 246.3 ms for the original Potjans-Diesmann microcircuit, i.e. a speed-up of 4.06 X, is to the best of our knowledge, the fastest simulation of the microcircuit reported so far. From this, it is evident that in this machine the performance of the communication system establishes a brick wall to any further simulation-time improvement. This is also underlined by the fact that applying more sophisticated neuron and synapse models as well as ODE-solver methods increases the simulation time only marginally.

Discussion: For future very-large Neuroscience networks with sophisticated plasticity models a significant improvement of simulation times will be required. At a first glance, there are three mayor challenges in this: First at all, to "bring the spikes to the FLOPS" and to synchronize the massively parallel processes, i.e., ultra-low latency communications. Secondly, to access the synapse states (including plasticity information) in huge data structures, i.e., ultra-low latency memory accesses. And, last but not least, to allow for quick network generation in the preparation of the simulation itself, avoiding hours-long set-up times. Neither today's HPC systems nor dedicated neuromorphic computing systems are well suited to these requirements and new, network centric system concepts and architectures are needed to overcome the current limitations. At least the simulation of cortical intra-area spike communication via electrical interconnects sets stringent limits to the spatial distance between the compute nodes. Multi-area models will enforce the use of improved hierarchical communication and synchronization infrastructures. Very-large on-chip or at least on-package memories will become a must. The simulation of ultra-large unscaled cortical areas may require new concepts and hardware support for spatial mapping. Moreover, sophisticated plasticity models with compartmental dendrite models may also challenge arithmetic performance and make the use of efficient accelerators a must. Although, initially area and energy efficiency do not appear as a mayor goal for such systems, like in biology, the stringent communication-latency constraints do enforce a high integration density at least at the level of cortical areas and consequently the need for low power consumption. It appears not to be reasonable to expect that future development of general-purpose HPC systems will follow these directions.

Funding: This work is partially supported by Helmholtz Association Initiative and Networking Fund under project number SO-092 (Advanced Computing Architectures, ACA) and European Union's Horizon 2020 Framework Programme for Research and Innovation under Specific Grant Agreement No. 785907 (Human Brain Project SGA2) and No. 945539 (Human Brain Project SGA3),

References

[Knight, 18] Knight, J.C., et al. (2018). GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model. *Front. Neurosci.12,* doi: 10.3389/fnins.2018.00941

[Rhodes] Rhodes, O., Peres, L., Rowley, A.G.D., Gait, A., Plana, L.A., Brenninkmeijer, C., Furber, S.B. (2020). Real-time cortical simulation on neuromorphic hardware. *Philosophical Transactions of the Royal Society A*: Mathematical, Physical and Engineering Sciences, Vol.378, No.2164, doi: 10.1098/rsta.2019.0160

[Kurth] Kurth, A., Finnerty, J., Terhorst, D., Pronold, J., Senk, J., Diesmann, M. (2020). Sub realtime simulation of a full density cortical microcircuit model on a single compute node. *Bernstein* Conference 2020. doi: 10.12751/nncn.bc2020.0221

[Schmidt] Schmidt, M. et al. (2018). A Multi-Scale Layer-Resolved Spiking Network Model of Resting-State Dynamics in Macaque Visual Cortical Areas. Ed.Lyle J. Graham. PLOS Computational Biology 14.10

[v. Albada, 20] van Albada, S.J., Pronold, J., van Meegen, A., Diesmann, M. (2020). Usage and Scaling of an Open-Source Spiking Multi-Area Model of Monkey Cortex, arXiv:2011.11335v1

[Knight, 21] Knight, J.C. , Nowotny, T. (2021). Larger GPU-accelerated brain simulations with procedural connectivity. *Nat. Comput. Sci. 1*, 136–142. doi: 10.1038/s43588-020-00022-7

[Stapmanns] Stapmanns, J., Hahne, J., Helias, M., Bolten, M., Diesmann, M., Dahmen, D. (2021). Event-Based Update of Synapses in Voltage-Based Learning Rules. Front. Neuroinform., 10 June 2021, doi: 10.3389/fninf.2021.609147

[Potjans] Potjans, T.C., Diesmann, M. (2014). The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. *Cereb. Cortex*, 24, 785-806. doi: 10.1093/cercor/bhs358

[v. Albada, 18] van Albada, S.J., Rowley, A.G., Senk, J., Hopkins, M., Schmidt, M., Stokes, A.B., David R. Lester, D.R., Diesmann, M., Furber, S.B. (2018). Performance Comparison of the Digital Neuromorphic Hardware SpiNNaker and the Neural Network Simulation Software NEST for a Full-Scale Cortical Microcircuit Model. Front. Neurosci., 23 May 2018. doi: 10.3389/fnins.2018.00291

[Roth] Roth, U., Eckardt, F., Jahnke, A., Klar, H. (1997). Efficient on-line computation of connectivity: Architecture of the connection unit of NESPINN. *Proc. MicroNeuro '97*, Dresden, 1997, pp. 31-39

[Walker] Walker, A.J. (1974). New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronics Letters.* **10** (8): 127, doi:10.1049/el:19740097

[Gewaltig] Gewaltig, M.-O., and Diesmann, M. (2007). NEST (NEural Simulation Tool). Scholarpedia 2:1430. doi: 10.4249/scholarpedia.1430

Author's Affiliations

[1]JARA-Institute Green IT (PGI-10), Jülich Research Centre, D-52425 Jülich, Germany

[2]Simulation & Data Lab Neuroscience, Jülich Supercomputing Centre Jülich Research Centre, D-52425 Jülich, Germany

[3]IBM Research Division, Almaden Research Center, San Jose, CA 9512

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

[5]Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany

[6]Department of Psychiatry, Psychotherapy and Psychosomatics, School of Medicine, RWTH Aachen University, Aachen, Germany