Optimizing communication in brain-scale multi-area model simulations

Melissa Lober[1,2*], Markus Diesmann[1,3], Susanne Kunkel[4]

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

[2] RWTH Aachen University, Aachen, Germany

[3] Department of Physics, Faculty 1, & Department of Psychiatry, Psychotherapy, and Psychosomatics, Medical School, RWTH Aachen University, Aachen, Germany

[4] Neuromorphic Software Ecosystems (PGI-15), Jülich Research Centre, Jülich, Germany

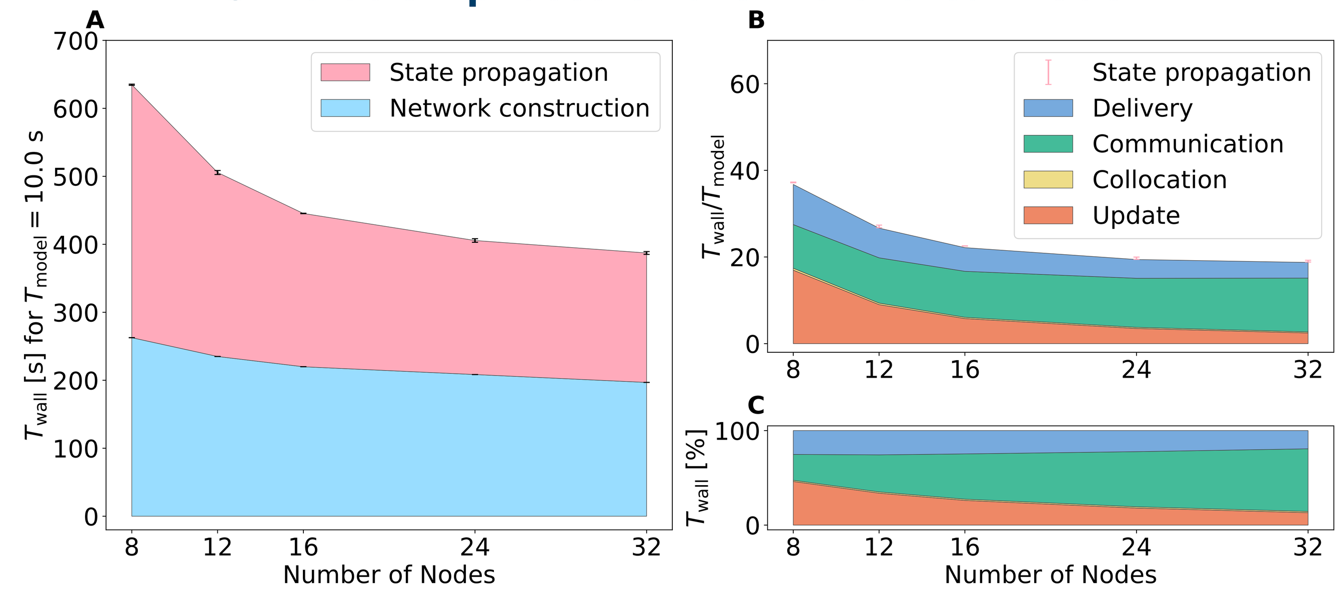
* m.lober@fz-juelich.de

Motivation

- continuous improvement of CPU-based simulation techniques create challenging benchmarking targets for neuromorphic platforms
  - neuronal simulations on conventional hardware still maintain higher flexibility at potentially lower cost compared to novel dedicated hardware [1]
- spike communication is the bottleneck in simulations of brain-scale networks [2]
- e.g. the multi-area model of macaque visual cortex [3]
  - 32 interconnected areas modelled as microcircuits [4]
  - realistic connectivity
  - single neuron resolution

⇒ structure-aware neuron distribution scheme combined with optimized spike-communication framework to speed up neuronal simulations

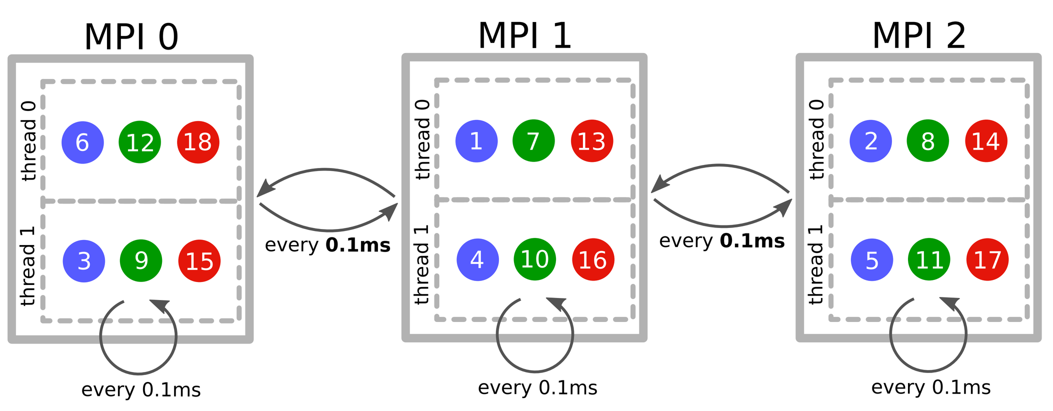
Simulation phases of the multi-area model



Strong-scaling benchmark of macaque multi-area model performed with NEST v3.6 on Jülich Supercomputer JURECA.

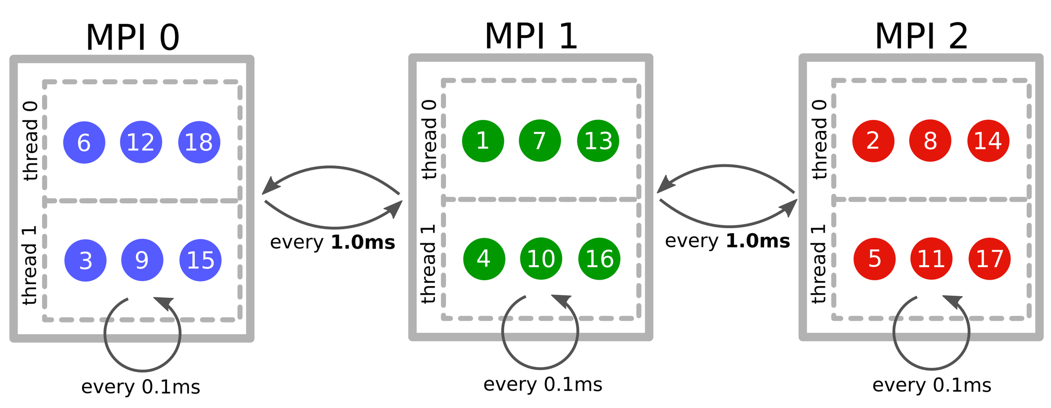
Algorithm

Conventional neuron distribution scheme



- uniform occupation of compute nodes: "round-robin"
- neurons of the same area are spread out on the hardware
- communication between compute nodes every smallest delay of e.g. 0.1 ms

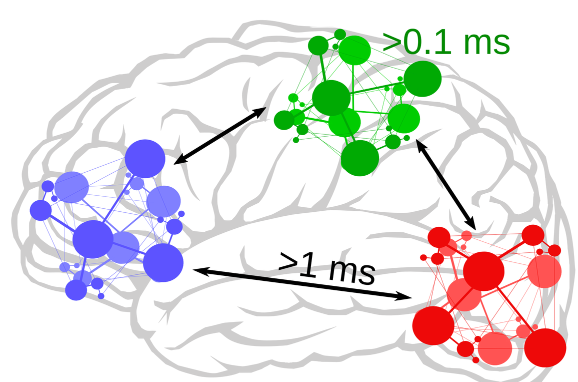
Structure-aware neuron distribution scheme



- one/few compute nodes per area
- two communication pathways
  - within an area: short delays (e.g. 0.1 ms)
  - between areas: long delays (e.g. 1.0 ms)

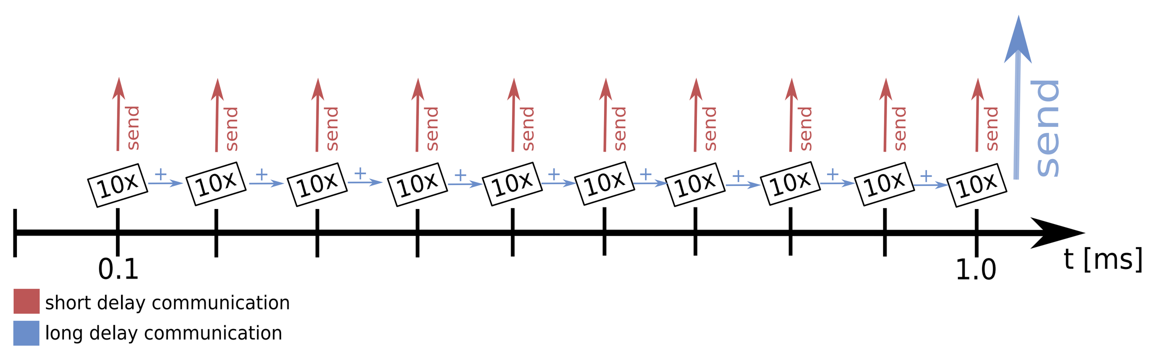
⇒ faster communication within areas

⇒ fewer communication between areas



Example: structure-aware approach

- 10.000 neurons per area; average firing rate 10 spikes/s

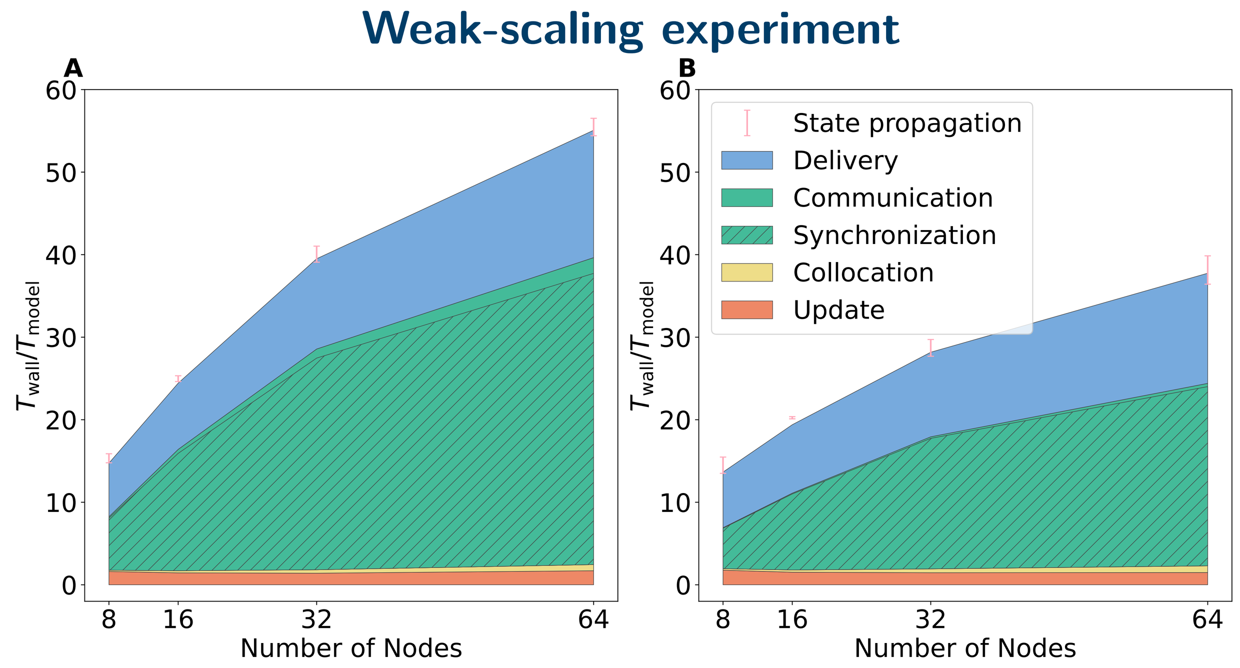


Results

Setup

- neuronal simulator tool NEST [5,6]
- benchmarking model
  - similar to macaque multi-area model in connectivity and work load
  - easily scalable while retaining constant activity levels
  - ≈ 130.000 neurons per area
  - ≈ 3000 inter- and intra-area connections per neuron, respectively
  - average spike rate of 2.5 spike/s
- Jülich Supercomputer JURECA
  - 2 areas per compute node
  - 2 MPI process per node; 64 threads per MPI Process
- communication phase
  - synchronization between all compute nodes (only long-range communication)
  - spike data exchange (both short-range and long-range communication)

Weak-scaling experiment

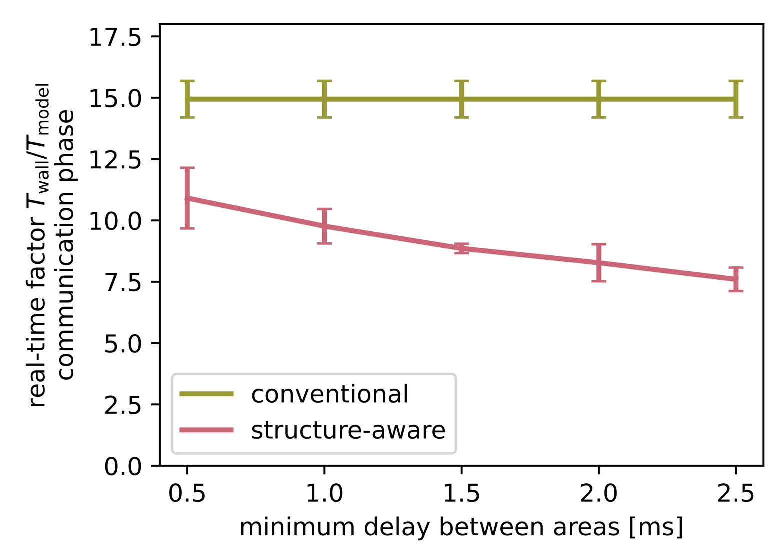


A Conventional round-robin neuron distribution with single communication pathway.

B Structure-aware neuron distribution with separate communication pathways for short- and long-range connections.

- significant speed up of spike communication
- speed of other simulation phases is maintained
- benefit lies in reduced time spent on compute node synchronization
- promising scaling behavior for large number of areas and compute nodes

Delay dependence



- delay distr. within an area: $\mathcal{N}(1.25, 0.625)$
- delay distr. between areas: $\mathcal{N}(5.00, 2.50)$
- lower cutoff of inter-area delay distribution defines inter-node communication frequency in structure-aware approach

⇒ benefit of implementation increases with decreasing amount of inter-area communication

References

[1] Kurth AC, Senk J, Terhorst D, Finnerty J, Diesmann M (2022) Sub-realtime simulation of a neuronal network of natural density. Neuromorphic Computing and Engineering, 2:021001. doi: 10.1088/2634-4386/ac55fc

[2] Albers J, Pronold J, Kurth AC, Vennemo SB, Haghighi Mood K, Patronis A, Terhorst D, Jordan J, Kunkel S, Tetzlaff T, Diesmann M and Senk J (2022) A Modular Workflow for Performance Benchmarking of Neuronal Network Simulations. Front. Neuroinform. 16:837549. doi: 10.3389/fninf.2022.837549

[3] Schmidt M, Bakker R, Hilgetag CC, Diesmann M and van Albada SJ (2018) Multi-scale account of the network structure of macaque visual cortex. Brain Structure and Function, 223: 1409 https://doi.org/10.1007/s00429-017-1554-4

[4] Potjans TC, Diesmann M (2014), The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. Cerebral Cortex, 24 3 785–806 https://doi.org/10.1093/cercor/bhs358

[5] Gewaltig M-O and Diesmann M (2007) NEST (Neural Simulation Tool) Scholarpedia 2(4):1430

[6] https://nest-simulator.readthedocs.io/en/latest

Outlook

- benchmarking of networks with inhomogeneous activity or size
- benchmarking state of the art models (e.g. multi-area model of macaque visual cortex)

Acknowledgments

This research was supported by the Joint Lab "Supercomputing and Modeling for the Human Brain", as well as "Neurosys" as part of the initiative "Clusters4Future" funded by the Federal Ministry of Education and Research BMBF (03ZU1106CB).