The impact of

Gaussian and exponential lateral connectivity on distributed spiking neural network simulation

E. Pastorelli, P.S. Paolucci, F. Simula, A. Biagioni, F. Capuani, P. Cretaro, G. De Bonis, F. Lo Cicero, A. Lonardo, M. Martinelli, L. Pontisso, P. Vicini, R. Ammendola

On their behalf

Pier Stanislao Paolucci,
the APE parallel/distributed computing lab - INFN Roma

Specific motivation for this study (1/2)

Recent experimentally defined model of probability of intra – areal connection among excitatory neurons:

Shnepel et al (2015) Cerebral Cortex Horizontal Connections…

slow exponential decay with distance

$P_{conn} = A \exp(-r/\lambda_{s,t})$, r:=distance between neuron

$150\ \mu m < \lambda_{s,t} < 350\ \mu m$ (depends on the source and target layers and neuron kind)

inside a cortical area ($r < 1$ cm), at least 75% synaptic connections have a long-range origin (i.e. from outside the local neural column).

Stepanyants et al. (2009) PNAS The fraction of short- and long- range connections …

…not counting inter-areal connectivity arriving through white-matter fibers ($r > 1$ cm)

… total number of synapses per neuron approaching 10 K

Specific motivation for this study (2/2)

- Previous studies estimated at about 75% local interconnectivity (and simulations accordingly performed)

  previous faster, shorter range gaussian decay of connection probability

  $P_{conn} = B \exp(-r^2/2\sigma^2)$,

  r:=distance between neuron

  typical $\sigma \sim 100\ \mu m$

Potjans, Diesmann (2014)
Cerebral Cortex
The cell-type specific…

… and a lower total number of synapses per neuron

-> longer range (and higher number of) connections to be supported in simulations -> **impact on distributed simulations**

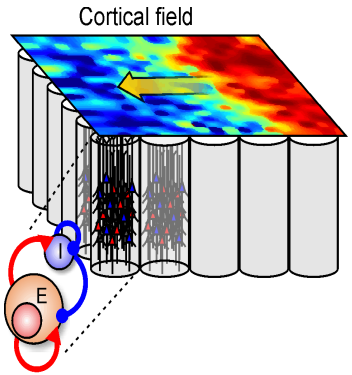
Cortical Slow Wave Activity (SWA)

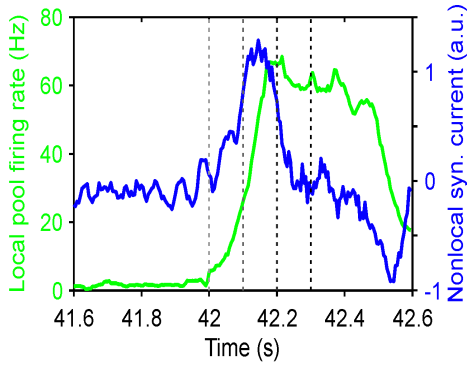
The transition from and toward expression of SWA, a phenomenon of great theoretical and applied interest.

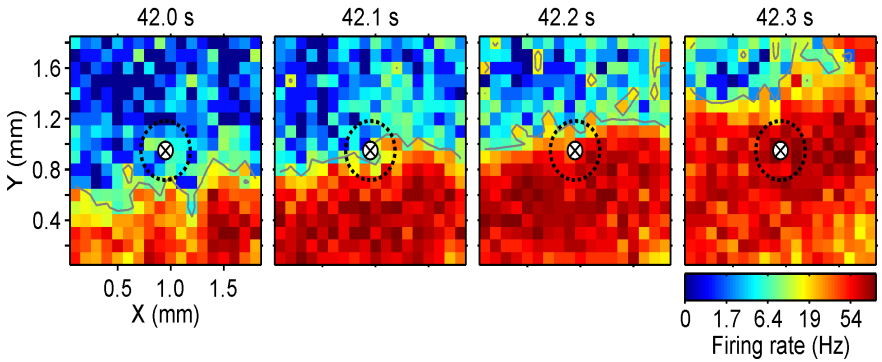
When Slow Waves appear, consciousness fades out, but they are an essential, fundamental mode of activity:

- ❑ SWA during dreamless deep sleep (every night the first phase of a good sleep), deep anaesthesia (unconscious kinds of), default mode of activity of isolated cortical modules
- ❑ SWA ubiquitous across animal species
- ❑ SWA more frequent, and essential, in juveniles
- ❑ Probable effects of physiologic SWA include improvement of coding of memories acquired during wakefulness and restoration of optimal cortical working point

Example of simulation requiring realistic intra-areal connections: cortical slow wave activity, single area simulated at high resolution (1/2)







Capone, Rebollo et al. (2017) Cerebral Cortex. Slow Waves…

Cortical area, described using a two-dimensional grid of cortical column.

Thousands of spiking neurons per column (excitatory and inhibitory).

Thousands of synapses per neuron

Simulation of a large field of cortical columns (pixels of the bottom snapshots),

Top-right panel: firing rate of the central column (green) of the cortical field and the net synaptic input it receives from neighboring columns (blue): local vs global contribution.

Parameters of the theoretical model defined by ISS (M. Mattia, P. Del Giudice, C. Capone)

Targeting the simulation of 1 cm^2 of cortex at biological resolution …

Species and brain area dependent requirements.

For the rat neocortex (V1)

Neural density 54 K neurons / mm^2, 5 K synapses / neuron

->

1 cm^2, 5.4 M neurons, 27 G synapses

… and measuring the impact of shorter and longer range interconnects

In this study, three problem sizes are mapped from 1 to 1024 hardware cores / MPI processes to evaluate the impact of connectivity on strong and weak scaling and memory occupation

GRID	COLUMNS	NEURONS	Number of SYNAPSES		MPI Processes / hardware cores	
			Gaussian shorter range interconnect	Exponential longer range interconnect	MIN	MAX
24x24	576	0.7 M	1.2 G	1.8 G	**1**	64
48x48	2304	2.9 M	3.5 G	5.9 G	4	256
96x96	9216	11.4 M	**14.2 G**	**23.4 G**	64	**1024**

… when executed on our own DPSNN (Distributed Plastic Spiking Neural Network) simulation engine, which is grounded on a data distribution strategy oriented to memory locality

Mean number of synapses (in thousands) projected according to Gaussian and exponential laws

Human Brain Project

Example:
24 x 24 neural columns,
Grid step, 100µm

Green: Gaussian Shorter range decay of connection probability, σ=100µm

Orange: exponential longer range decay, λ=290µm

Locally projected: 992 K synapses in both cases

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1				1.2																				
2		1.2	5.0	8.7	5.0	1.2																		
3		5.0	22.3	37.2	22.3	5.0																		
4	1.2	8.7	37.2	992	37.2	8.7	1.2	0.7	0.8	0.9	1.0	1.1	1.2	1.2	1.2	1.1	1.0	0.9	0.8	0.7				
5		5.0	22.3	37.2	22.3	5.0	0.7	0.9	1.1	1.2	1.4	1.5	1.6	1.7	1.6	1.5	1.4	1.2	1.1	0.9	0.7			
6		1.2	5.0	8.7	5.0	1.2 0.8	1.0	1.2	1.4	1.7	2.0	2.2	2.3	2.4	2.3	2.2	2.0	1.7	1.4	1.2	1.0	0.8		
7				1.2	0.7	1.0	1.2	1.5	1.9	2.3	2.7	3.0	3.2	3.3	3.2	3.0	2.7	2.3	1.9	1.5	1.2	1.0	0.7	
8				0.7	0.9	1.2	1.5	2.0	2.5	3.1	3.7	4.2	4.6	4.7	4.6	4.2	3.7	3.1	2.5	2.0	1.5	1.2	0.9	0.7
9				0.8	1.1	1.4	1.9	2.5	3.2	4.1	5.0	5.8	6.4	6.6	6.4	5.8	5.0	4.1	3.2	2.5	1.9	1.4	1.1	0.8
10				0.9	1.2	1.7	2.3	3.1	4.1	5.3	6.6	8.0	9.0	9.4	9.0	8.0	6.6	5.3	4.1	3.1	2.3	1.7	1.2	0.9
11				1.0	1.4	2.0	2.7	3.7	5.0	6.6	8.6	10.7	12.5	13.2	12.5	10.7	8.6	6.6	5.0	3.7	2.7	2.0	1.4	1.0
12				1.1	1.5	2.2	3.0	4.2	5.8	8.0	10.7	14.0	17.2	18.7	17.2	14.0	10.7	8.0	5.8	4.2	3.0	2.2	1.5	1.1
13				1.2	1.6	2.3	3.2	4.6	6.4	9.0	12.5	17.2	22.8	26.4	22.8	17.2	12.5	9.0	6.4	4.6	3.2	2.3	1.6	1.2
14				1.2	1.7	2.4	3.3	4.7	6.6	9.4	13.2	18.7	26.4	992	26.4	18.7	13.2	9.4	6.6	4.7	3.3	2.4	1.7	1.2
15				1.2	1.6	2.3	3.2	4.6	6.4	9.0	12.5	17.2	22.8	26.4	22.8	17.2	12.5	9.0	6.4	4.6	3.2	2.3	1.6	1.2
16				1.1	1.5	2.2	3.0	4.2	5.8	8.0	10.7	14.0	17.2	18.7	17.2	14.0	10.7	8.0	5.8	4.2	3.0	2.2	1.5	1.1
17				1.0	1.4	2.0	2.7	3.7	5.0	6.6	8.6	10.7	12.5	13.2	12.5	10.7	8.6	6.6	5.0	3.7	2.7	2.0	1.4	1.0
18				0.9	1.2	1.7	2.3	3.1	4.1	5.3	6.6	8.0	9.0	9.4	9.0	8.0	6.6	5.3	4.1	3.1	2.3	1.7	1.2	0.9
19				0.8	1.1	1.4	1.9	2.5	3.2	4.1	5.0	5.8	6.4	6.6	6.4	5.8	5.0	4.1	3.2	2.5	1.9	1.4	1.1	0.8
20				0.7	0.9	1.2	1.5	2.0	2.5	3.1	3.7	4.2	4.6	4.7	4.6	4.2	3.7	3.1	2.5	2.0	1.5	1.2	0.9	0.7
21					0.7	1.0	1.2	1.5	1.9	2.3	2.7	3.0	3.2	3.3	3.2	3.0	2.7	2.3	1.9	1.5	1.2	1.0	0.7	
22						0.8	1.0	1.2	1.4	1.7	2.0	2.2	2.3	2.4	2.3	2.2	2.0	1.7	1.4	1.2	1.0	0.8		
23							0.7	0.9	1.1	1.2	1.4	1.5	1.6	1.7	1.6	1.5	1.4	1.2	1.1	0.9	0.7			
24								0.7	0.8	0.9	1.0	1.1	1.2	1.2	1.2	1.1	1.0	0.9	0.8	0.7				

Comparing the speed of simulations (1/2)

Simple and effective measure to compare the speed of neural simulations on different problem sizes and activity regimes of spiking neurons with instantaneous synaptic current injection:

N := total # neurons,

M := mean # synapses / neuron,

N := mean firing rate of neurons (and synapses),

T_S := simulated time, T_E := elapsed execution time

Total number of synaptic events to be simulated: $N \times M \times \nu \times T_S$

Execution time per simulated event = $T_E / (N\ M\ \nu\ T_S)$

Or the reciprocal, a simulation speed:

 # simulated synaptic events / second

Comparing the speed of simulations (2/2)

MOTIVATION: There are N neurons, but NM synapses

- Simulation of individual spiking neuron: integration of a low dimentional differential equation. For example, LIF-SFA, Leaky Integrate and Fire neuron with rate dependent Spike Frequency Adaptation. Two dynamic variables, first order differential equations, plus forcing by synaptic current injection. If a threshold is exceeded, a spike is emitted. After spike, reset to post-spike values
- For each spike of an individual neuron, all its projected synapses are activated and have to inject a current in the target neuron
- NOTE AGAIN. For each neuron, there are thousands of synapses
- Current injection, an event driven approach can be adopted. Instantaneous synapses are not active when not spiking, and a time driven approach would be highly inefficient.

Measures on DPSNN engine

DPSNN (Distributed Plastic Spiking Neural Network simulation engine)

Developed by INFN APE Parallel/Distributed Computing Lab. Objectives: maximum speed on selected problems. Benchmarking tool for application specific computing/ interconnect architectures

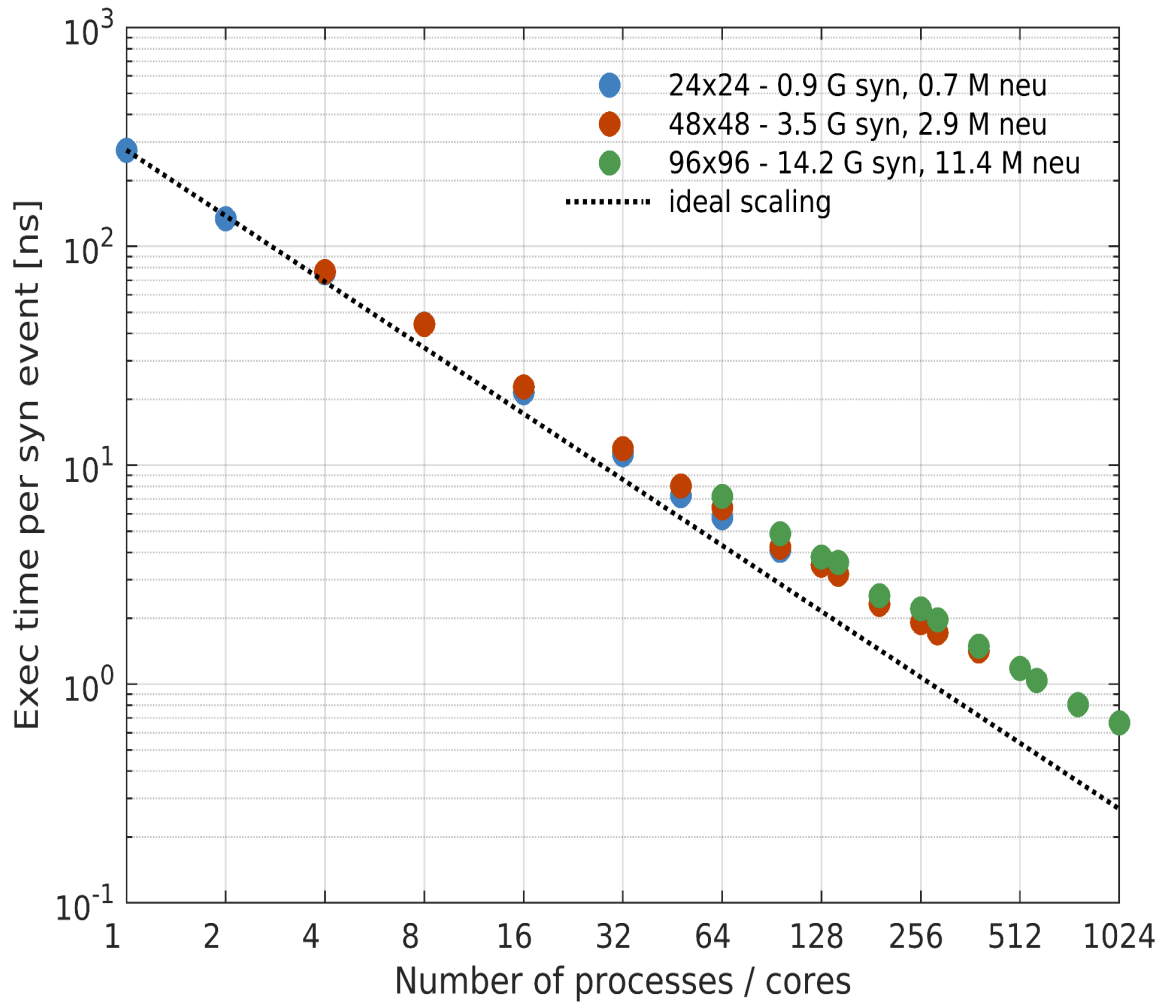
Natively distributed, exploits memory and temporal locality.

Synapses mapped in the same process of target neurons

Clusters of neurons and incoming synapses in a process

Really fast and scalable… INFN APE Lab. Since 1984 developed several generations of application specific parallel/distributed platforms.

Strong scaling for Gaussian Connectivity, DPSNN simulation engine



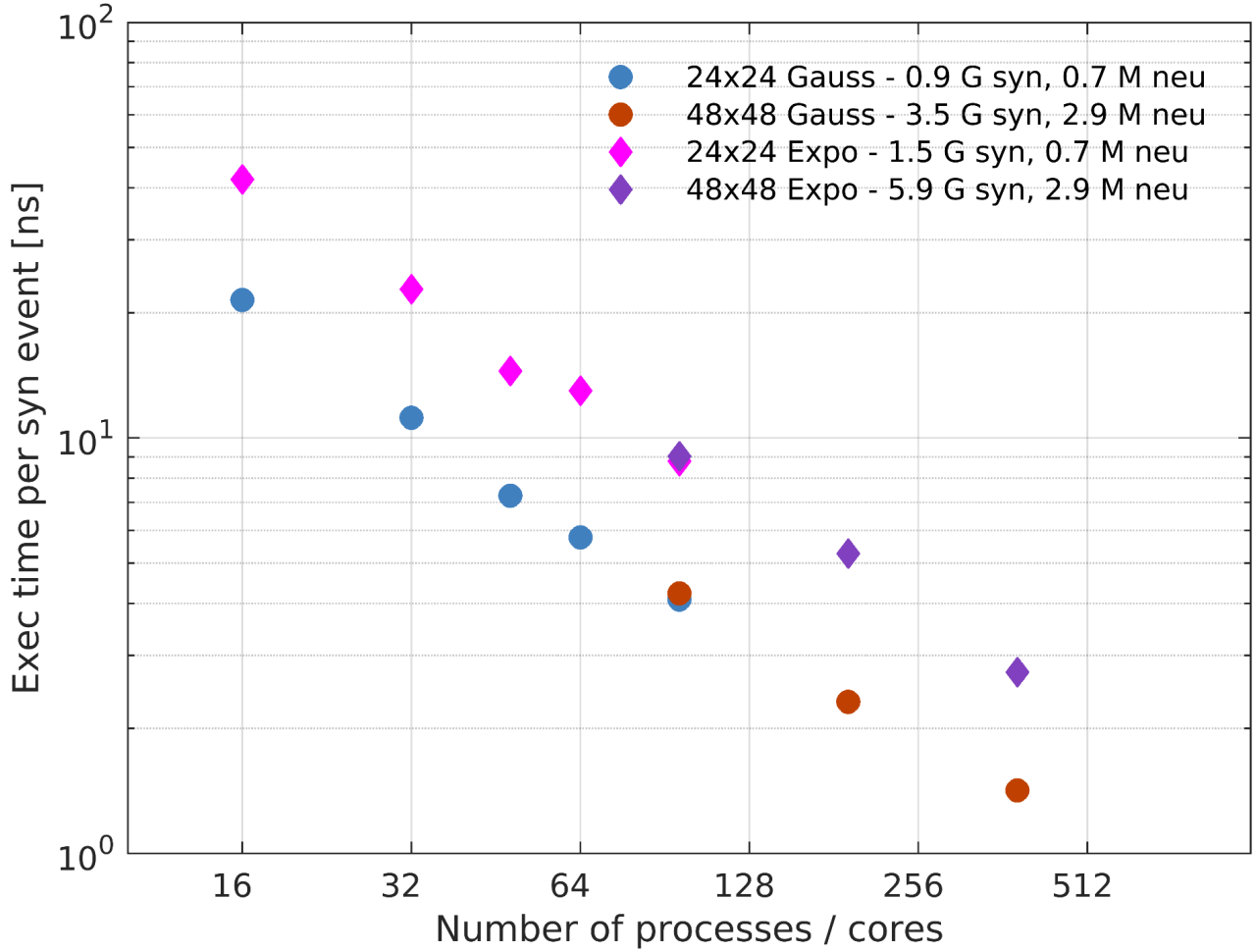
Execution Platform:
GALILEO @ CINECA

Up to 64
IBM Nodes, 1024 cores.

Two Intel Xeon
Haswell 8-core E5-2630
V3 processor per node
@ 2.4 GHz.

Infiniband network,
4x QDR switches.
Hyper-threading off.

Impact of longer-range exponential connectivity, manageable on DPSNN engine



Memory cost: normalized measure

Ideally, the memory cost should be proportionally to the number of represented synapses. A normalized measure to compare for different problem sizes and mappings

Normalized memory cost := divide the peak memory occupation by the number of total represented synapses

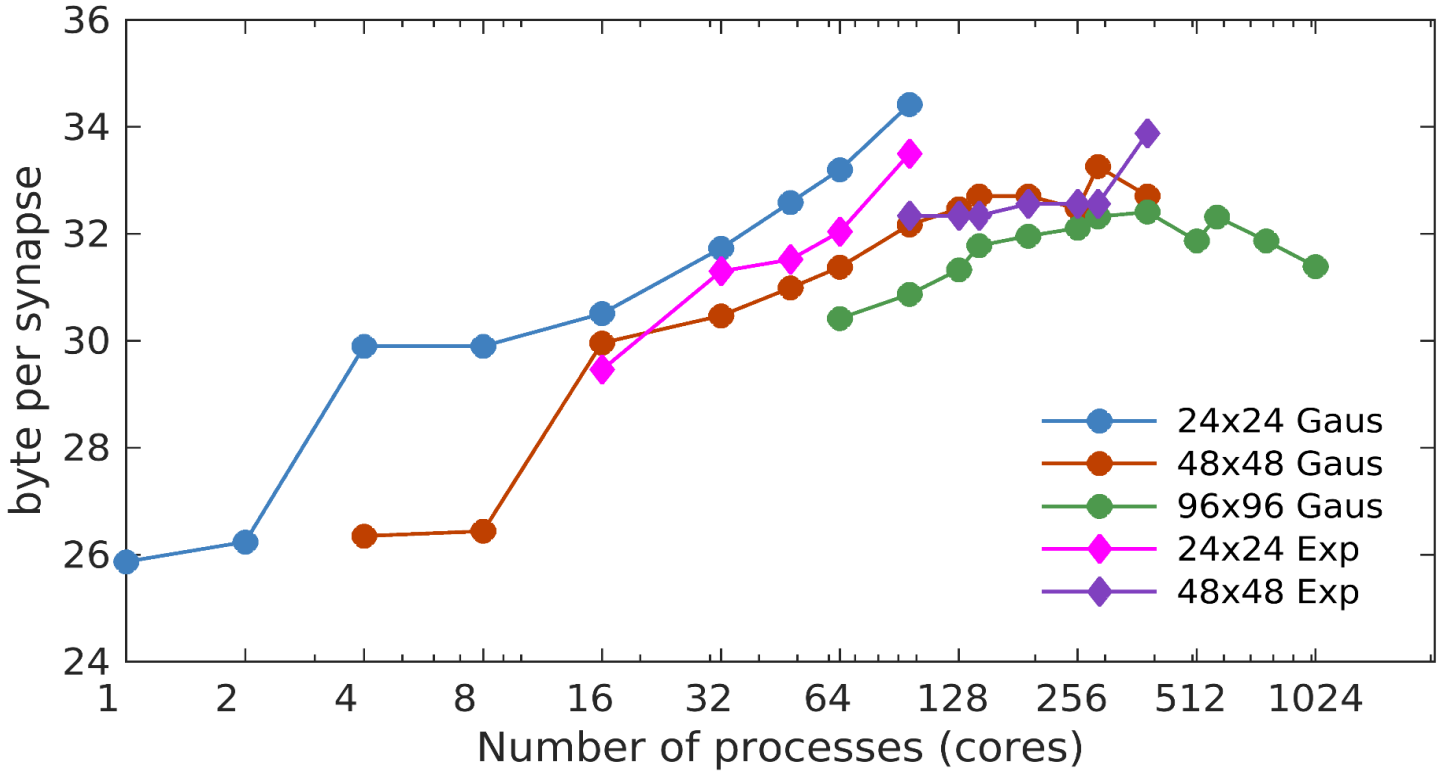
On DPSNN, when plasticity is switched-off (static synapse), a synapse is represented during execution by 12 Bytes on the process storing the target neuron:

4 B source neuron id, 4 B target neu id, 2 B weight, 2 B syn kind

During initialization, synapses are represented on both source and target neuron -> minimum expected peak occupation:

24 B / represented synapse

Memory cost on DPSNN for shorter and longer range interconnects



the overhead is mainly due to: 1) MPI buffers 2) Internal DPSNN structures to demultiplex from axonal spike event messages to synaptic events

DPSNN simulation engine: objectives of INFN APE parallel/ distributed computing lab

1) Quantitative assessment of requirements / benchmarking during development of embedded and HPC systems specialized for neural simulations, focusing on either:

- ❑ Specialized interconnects, for ARM and Intel based platforms
- ❑ Power efficiency, e.g. on ARM processors
- ❑ Acceleration of kernels (e.g. on FPGAs)
- ❑ Invention and test of improved distributed coding techniques on standard message passing software infrastructures to be ported on general purpose neural simulation platforms

2) Acceleration of specific scientific simulations: e.g. INFN coordinator of the WaveScalES experiment in the Human Brain Project (specific objectives in WaveScalES: simulation of cortical Slow Wave Activity (SWA) and matching with experimental results, understand interaction between sleep (SWA) and memories, transition from unconscious/anaesthesia states (characterized by SWA) to conscious states (asynchronous activity, gamma rythms…)

Main + and - points of our DPSNN simulation engine

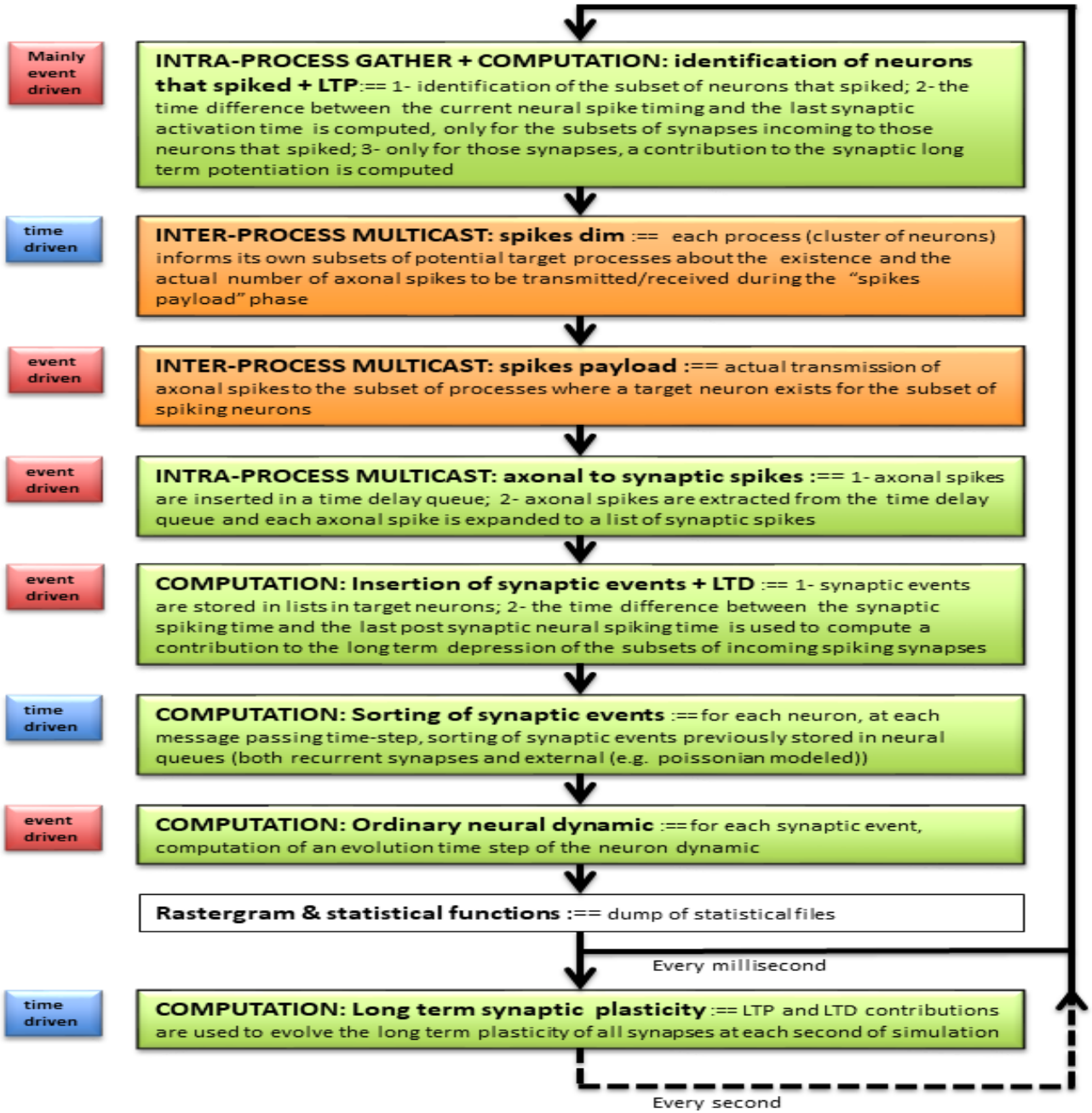
- Fast distributed network initialization
- Mixed time-driven (axonal messages between software processes) and event-driven (synaptic dynamic) scheme -> high temporal resolution on individual synaptic event AND good scalability on high number of MPI processes
- Highly application specific – dirty down to the essential – no bells and whistles -> speed / scalability potential
- Easy, essential benchmark kernel for hardware architecture
- Limited flexibility / configurability of models of neurons, synapses, connectivity
- In house maintenance and reconfiguration for different problems needed
- Not a platform for the general neuroscientist. Requires the support of the developer

DPSNN engine: internals (1/3)

- Mixed time and event driven simulation engine
  - Event driven: synaptic events and integration of neural dynamics
  - Time driven: exchange of spiking messages among processes
- Data distribution strategy:
  - Synapses are localized in memory near the target neurons
  - Neurons and synapses contiguous in space are stored in the same process
- DPSNN processes are agnostic of the specific message passing library agnostic:
  - in this study DPSNN uses MPI,
  - in other studies DPSNN processes have been e.g. nodes of a Kahn network
- Natively parallel initialization:
  - each DPSNN process creates in total autonomy its own set of neurons and manages the creation of the connections of its own set of projected synapses and incoming synapses

DPSNN engine: internals (2/3)

- Axonal arborization of spiking messages to a set of target synapses deferred to the processes that host target neurons
- Suppression of superfluous interprocess communications using a pruning strategy performed in several steps: (two steps during initialization, two steps at each iteration of the simulation)
- supported e.g. by a sequence of MPI_alltoallv() calls addressing subsets of targets of decreasing size, ….



DPSNN engine: internals (3/3) mixed event and time driven execution flow

Conclusions / Acknowledgements

- Long-range intra-areal synaptic connections projected according to distance dependent decay laws of probability compatible with experimental evidence have a measurable but manageable impact on scalability of spiking neural network simulations distributed on up to 1K cores and performed at biological resolution using simulators that exploit memory and time locality, like our DPSNN engine.

- Acknowledgements:
  - The Human Brain Project, EU grant No. 720270 (HBP SGA1)
  - The ExaNeSt Project, EU grant No. 671553
  - INFN-CINECA Computational Theoretical Physics Collaboration