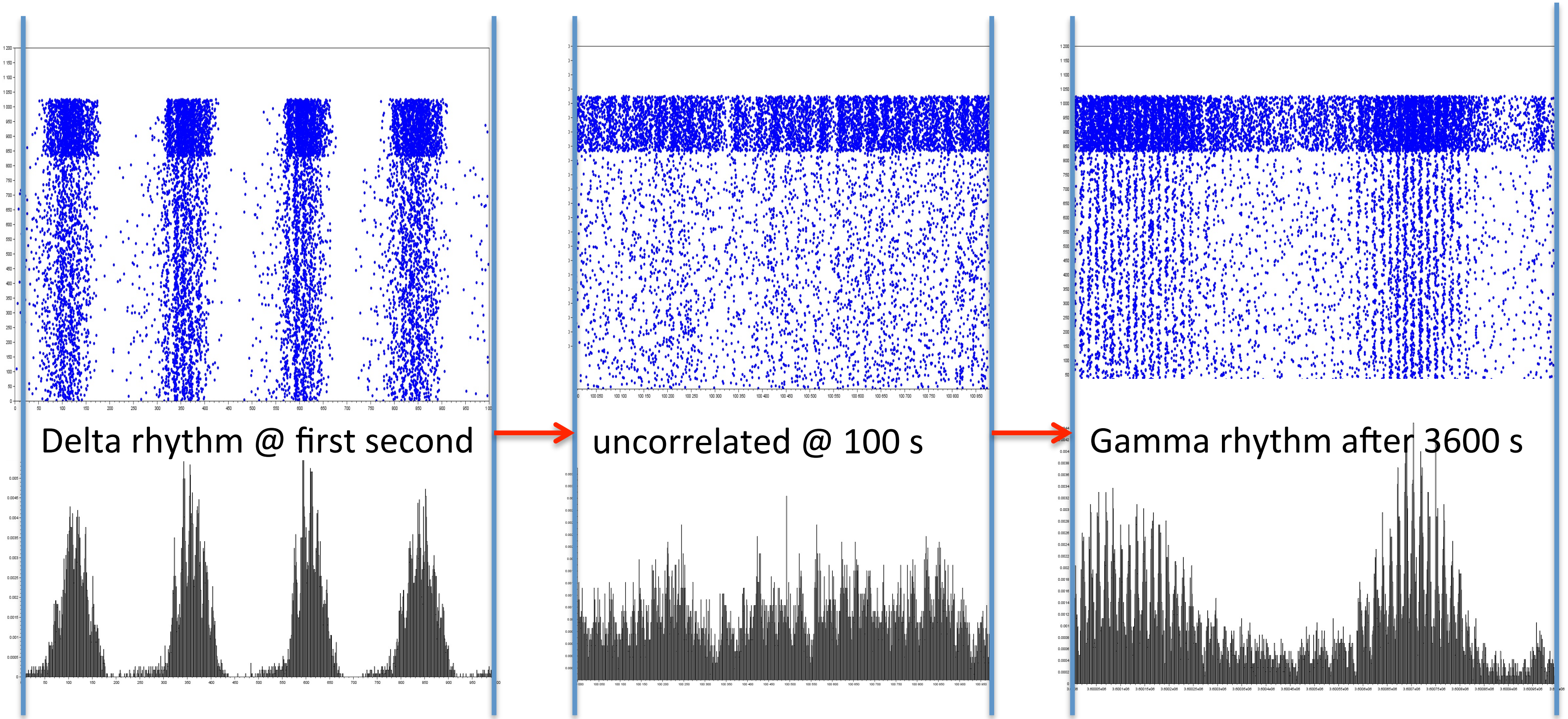
Large scale modeling of neuro-synaptic activity ad plasticity

Pier Stanislao Paolucci
for INFN Roma APE LAB.

Neuro-synaptic activity and Plasticity

- Key areas of present INFN activity on large scale neural modeling
  - Coding of scalable Parallel/Distributed simulator
    - INFN developed the DPSNN-STDP simulator in the EURETILE FET Project. Proven simulation up to 6.6 G synapses, 128 cores.
    - See arXiv:1310.8478 (Apr 2014)
  - Comparison with experimental neuro-biological data and calibration of the INFN simulator
    - Will be performed in the CORTICONIC FET project (starting from Oct 2014, end Dec 2015) (cooperation with ISS, TUM, IDIBAPS)
  - Interface with experimental systems / inclusion of the simulator into robotic platforms
    - Will be investigated by the INFN "COSA" (iniziativa di gruppo 5), start Jan 2015
  - Co-design of simulation code and execution platform
    - The plan is to start from "COSA" and "CORTICONIC" to prepare the participation to a future European project on this topic

Emergent Biological Behaviour: Spontaneous Evolution of Rythmic Activity due to Polychronism and Synaptic Plasticity

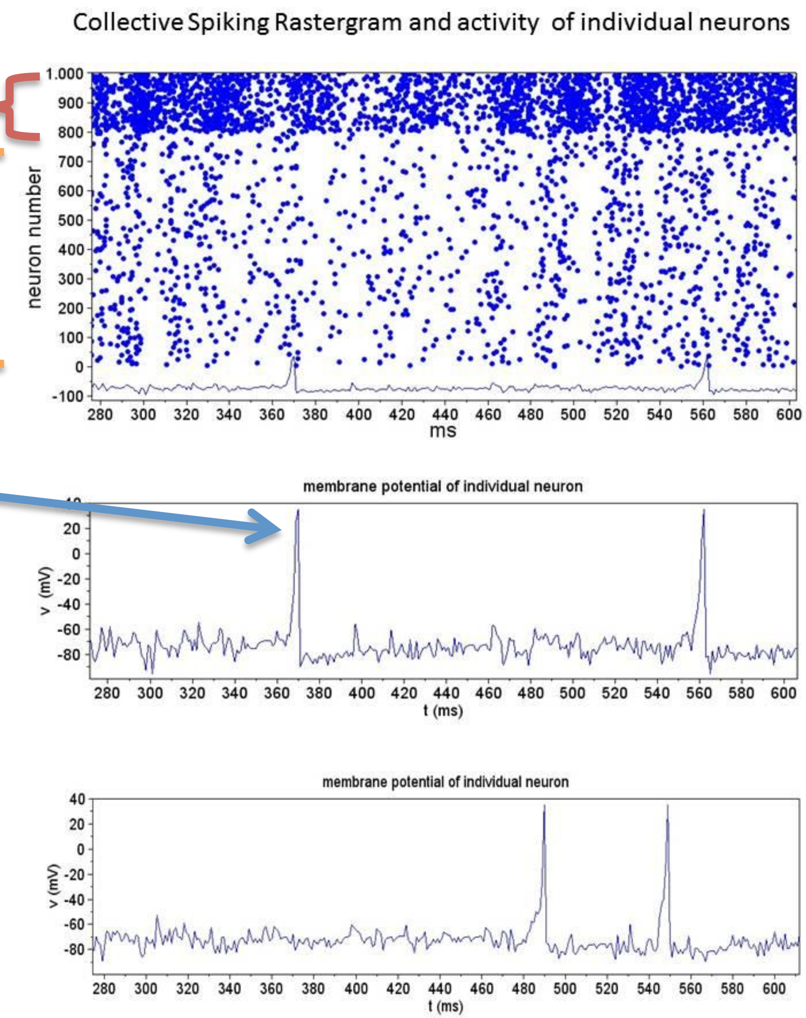


- As synaptic weights evolve according to STDP (synaptic spike-timing dependent plasticity, initial **delta** frequency oscillations (2-4 Hz @ first second activity) dissolves for a while into **uncorrelated** Poissonian activity (activity @ 100 seconds) and then **gamma** frequency activity emerges (30-100 HZ @ 3600 seconds)

INFN Istituto Nazionale di Fisica Nucleare

DPSNN-STDP simulates Spiking Activity and Synaptic Plasticity (already proved from 100 K up to 6.6 Giga synapses, from 1 to 128 software processes)

- The picture represents the evolution of a neural network computed by the DPSNN-STDP code
- This picture:
  - 200 inhibitory neurons
  - 800 excitatory neurons
  - total 100 000 synapses
  - Time resolution: 1ms (horizontal axis)
  - Each dot in the raster gram represents an individual spike
  - The evolution of the membrane potential of each neuron is simulated
  - The evolution of individual synaptic strength is computed (not shown in the picture)
  - Polychronism: individual synaptic delays are taken into account
  - Individual connections and neural types can be programmed



INFN Istituto Nazionale di Fisica Nucleare

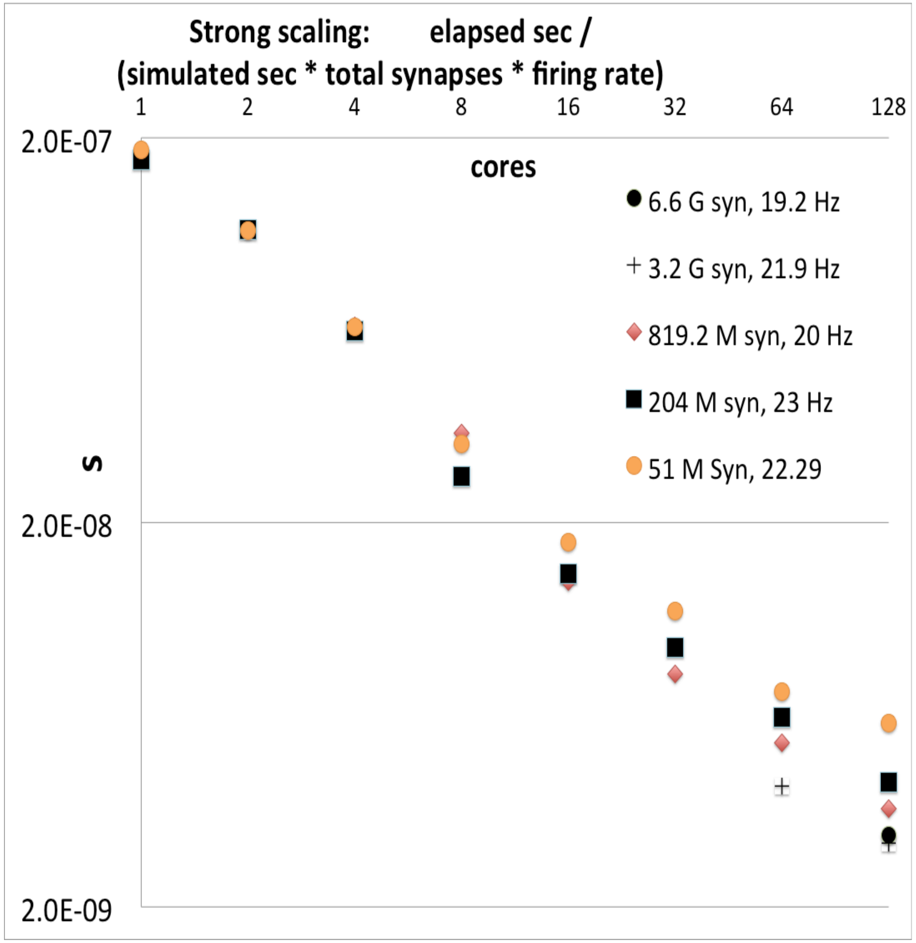
Measurements - MPI version of DPSNN-STDP

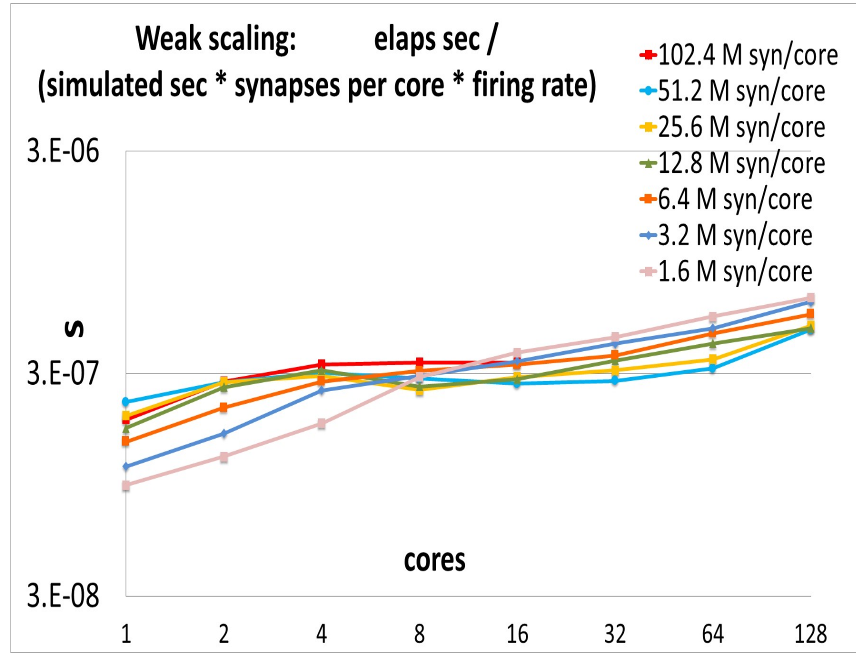
- From 200 K to 6.6 Giga synapses, 1 K to 32.8 Million neurons
- From 1 to 32 K cortical columns, max bi-dim grid 256x128
- From 1 to 128 software processes mapped onto 2.4 GHz cores

Total synapses	**200 K**	**800 K**	**3.2 M**	**12.8 M**	**51.2 M**	**204.8 M**	**819.2 M**	**3.2 G**	**6.6 G**
Total neurons	1 K	4 K	16 K	64 K	256 K	1024 K	4.096 M	16.4 M	32.8 M
Grid of neural columns	1 x 1	2 x 2	4 x 4	8 x 8	16 x16	32 x 32	64 x 64	128x128	256x128
Mean firing rate (Hz)	27	24	26	23	22	23	20	22	19
Used cores1 **(min-max)**	1-8	1-32	1-128	1-128	1-128	1-128	4-128	64-128	64-128
MPI processes	1-8	1-32	1-128	1–256	1-256	1-256	4-256	64-256	128
Execution time2 **(execution sec / simulated sec)**	0.15	0.4	1.80	3.05	6.85	20.0	59	211	386
Normalized execution time3**: execution time / (firing rate × total syn × simulated second)**	2.73×10^{-8}	5.36×10^{-9}	2.41×10^{-8}	4.22×10^{-9}	6.0×10^{-9}	4.22×10^{-9}	3.61×10^{-9}	2.94×10^{-9}	3.07×10^{-9}

1 Each cores @2.4 Ghz part of a quad-core Intel(R) Xeon(R) E5620.
2 Using the "max" number of cores reported in this table
3 See the "Strong scaling" section for a discussion about the unit of measure for the normalized execution speed.

DPSNN-STDP: MPI version - Strong and Weak Scaling



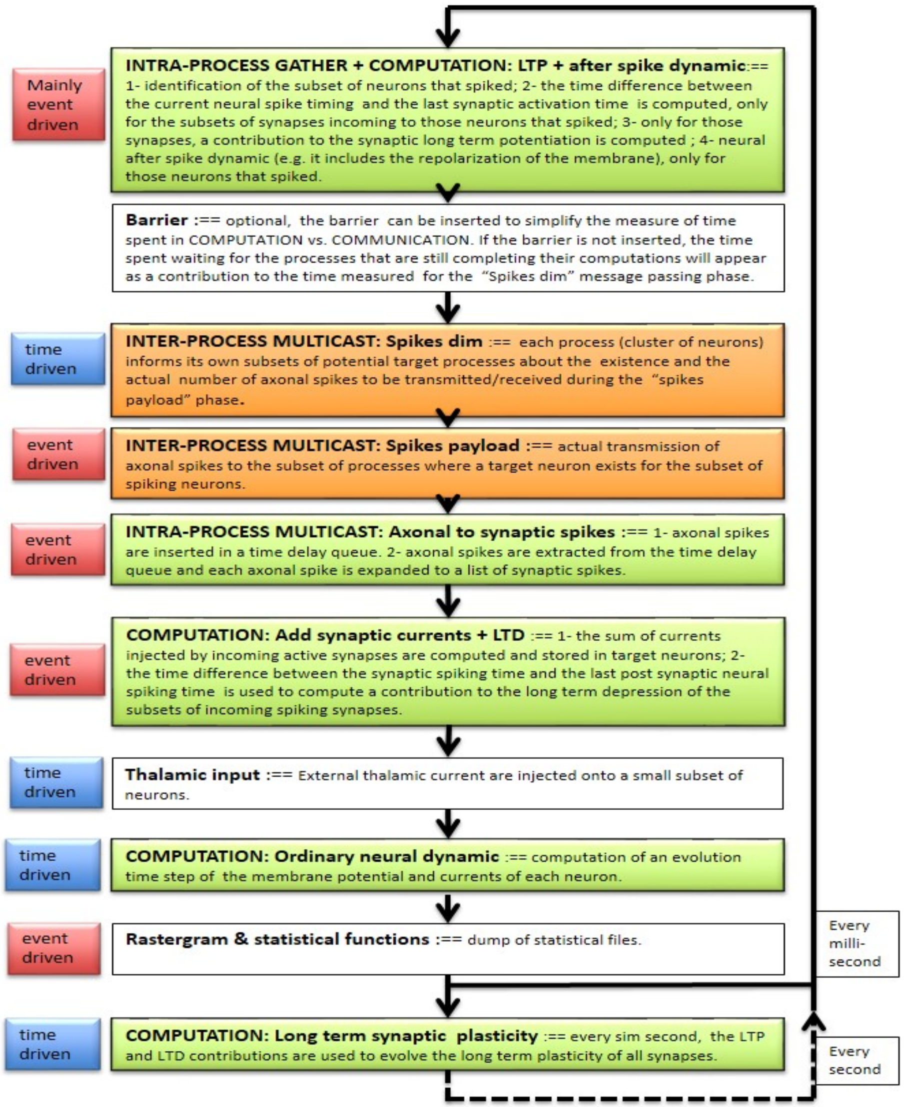


- Weak scaling for various local network sizes. Exec time normalized to synapse count.

- Strong scaling. From 1 to 128 cores @ 2.4 GHz simulate various total network sizes (from 51 Mega syn to 6.6 Giga synapses). Exec time normalized to synapse count.

INFN Istituto Nazionale di Fisica Nucleare

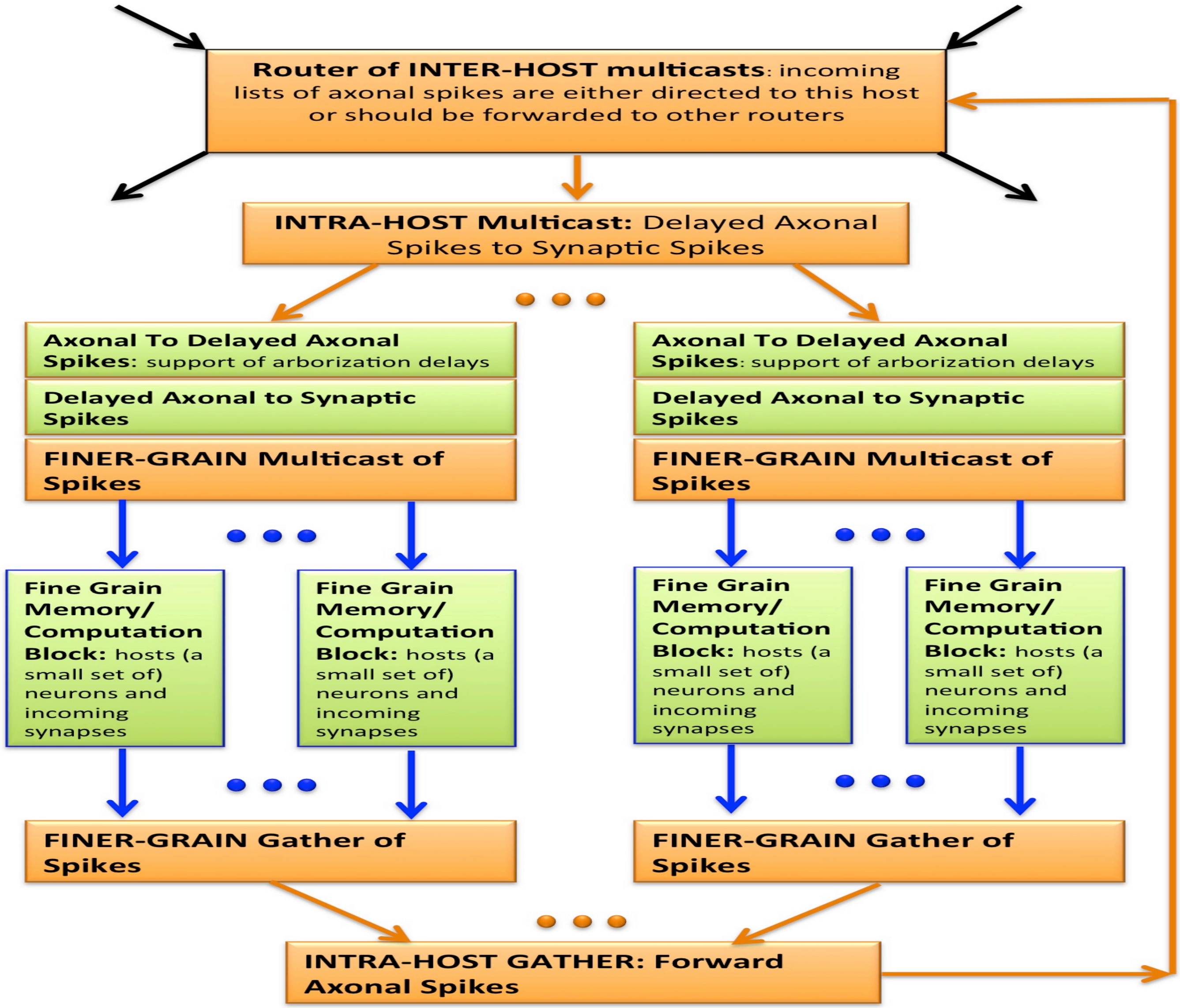
From Program Flow and Profiling …



Function of the block	Relative execution time	Note
Long term potentiation + after spike dynamic	(9.7 ±0.7)%	Gather**1** + computation
Barrier (optional)**2**	(29.9±6.1)%	Workload fluctuations
Communication: inter-process multicast: Spikes dim	(0.77±0.10)%	Message passing
Communication: inter-process multicast: Spikes payload	(0.82±0.20)%	Message passing
Axonal to synaptic spikes: intra-process multicast	(16.8±2.3)%	Dereferencing**3**
Add synaptic currents + long term depression	(19.2 ±2.7)%	Computation
Thalamic input**4**	0.01%	Simplified model
Ordinary neural dynamic	(11.8±1.4)%	Computation
Rastergram & other statistical functions	(1.9±0.1)%	Computation
Long term synaptic plasticity	(9.2±1.8)%	Computation

1 The benchmarked software implementation is based on sparse accesses from the target neuron to the global list of incoming synapses. In a hardware implementation, based on several independent memory banks, if all synapses incoming to the same neuron were stored in contiguity, this task could be easily accelerated.
2 If the barrier is not inserted, the time spent waiting for the processes that are still completing their computations appear as a contribution to the time measured for the "spikes dim" communication phase. We verified that the deviations from ideal strong scaling, can be entirely attributed to the cost of measured fluctuations in workload execution (represented by the Barrier block) and to (a very small) cost of communications.
3 The task to be performed is an "intra-process" multicast, from axons to specific lists of synapses. Instead, the benchmarked software implementation is based on two levels of dereferencing.
4 In this simulation the thalamic input is computed using a simple statistical model. Actually, this is one of the interface between the neural network and the "external" world, so its weight would greatly increase and add to that of other interfaces to be added.

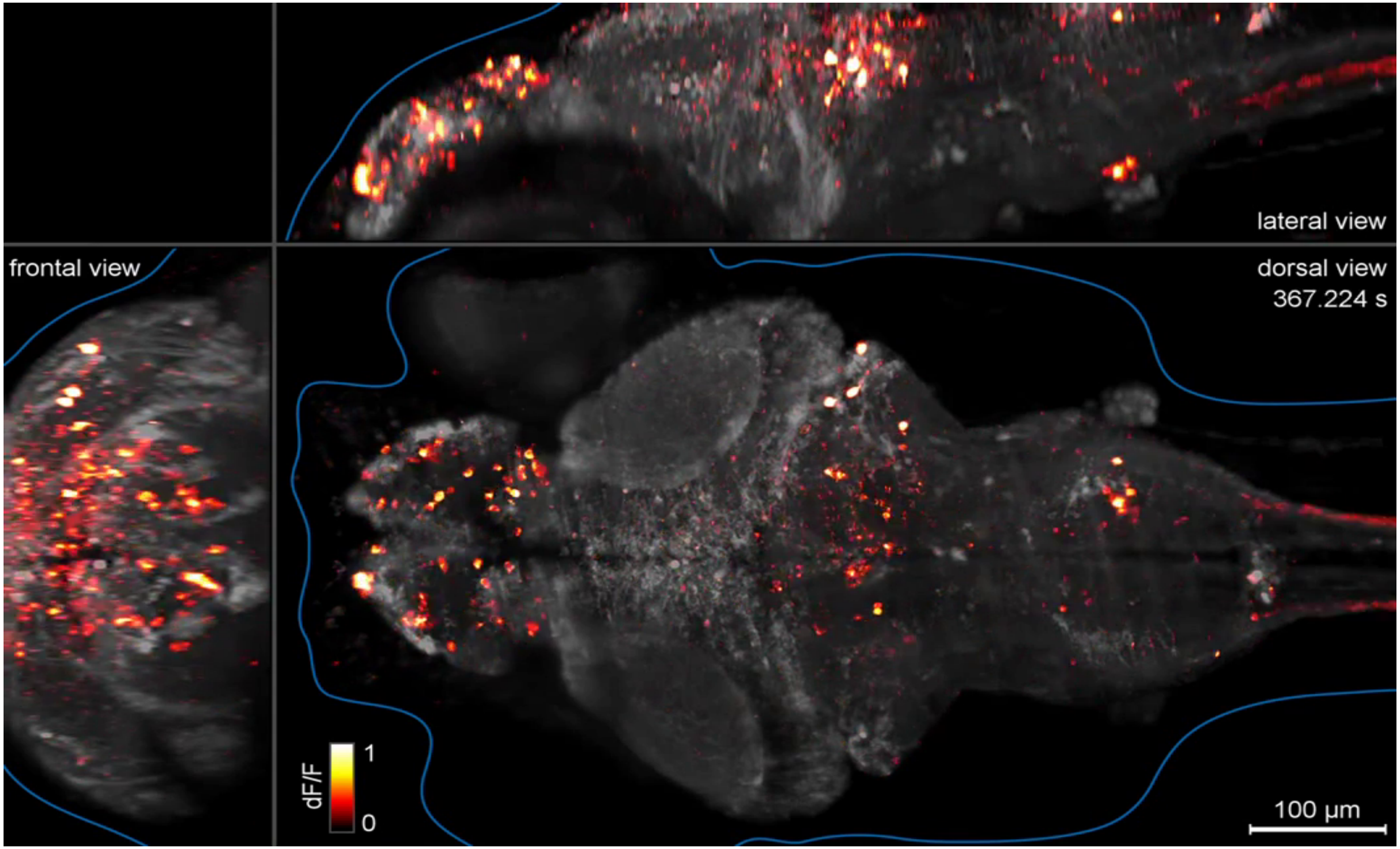
… toward Hardware Acceleration



CORTICONIC FET project

- Experimental techniques used to stimulate and measure cortical activities on animals and humans: opto-genetic, electrode arrays, trans-cranial magnetic stimulation, electro encephalographic arrays, drug perfusion…
- Large scale simulations
- DPSNN will be improved to simulate biological networks:
    - Cooperation between INFN and ISS
- Comparison with in-vivo/in-vitro experimental results
    - IDIBAPS (Barcellona), TUM

Spiking activity of individual neurons observed in real-time (e.g. in a Zebra Fish Larva)



Misha B Ahrens, Philipp J Keller, «Whole-brain functional imaging at cellular resolution using light-sheet microscopy», Nature Methods, 18 March 2013, DOI:10.1038/NMETH.2434

Howard Hughes Medical Institute, 3D recording of temporal spiking activity of ~100 000 neurons.

Note: the effective time resolution is still only ~1 s.

Hardware-Software Co-design opportunity

- Huge potential for architectural improvements driven by this benchmark:
  - the brain performs with 50 W computations that would require more than 50 MW on present generation HPC architectures
- Strategic research area…

Conclusion

- present INFN activity on cortical/brain modeling
  - Coding of scalable Parallel/Distributed simulator
  - Comparison with experimental neuro-biological data and calibration of the INFN simulator
  - Interface with experimental systems / inclusion of the simulator into robotic platforms
  - Co-design of simulation code and execution platform

- A strategic Large Scale Computing research theme…