STOCKHOLM BRAIN INSTITUTE


KTH Campus

Very large-scale brain simulations using supercomputers

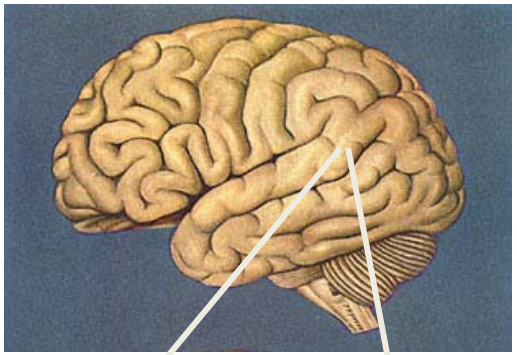
Anders Lansner

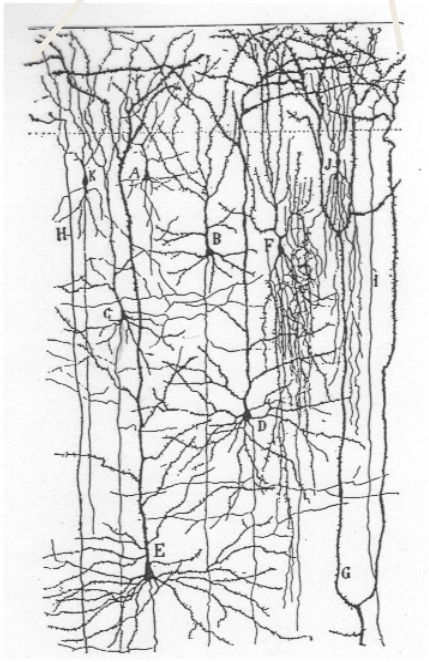
Dept of Computational Biology
KTH and Stockholm University

Our brain = ”alien technology”

STOCKHOLM BRAIN INSTITUTE

- Volyme: 1.5 liters
- Cortex (80%)
  - Dimension: 0,3 cm × 2400 cm^2
  - N:o nerve cells: 20 billion
  - N:o connections (synapses): ~ 2 10^{14}
- Power: 25 W, ~PC-processor
- 1 rack BG/L: 35 kW
- Fiber length: 10 million km
  - > 20 × Earth – Moon distance
  - > 6000km/cm^2
  - 90 nm chip: 10 km/cm^2
- N:o messages/s: 2 10^{13}
- ”Alien technology”
  - Scientific challenge to understand!
  - … and to model/simulate/emulate

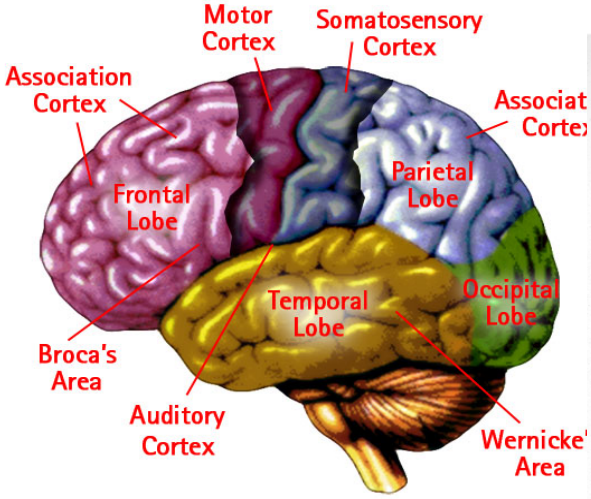


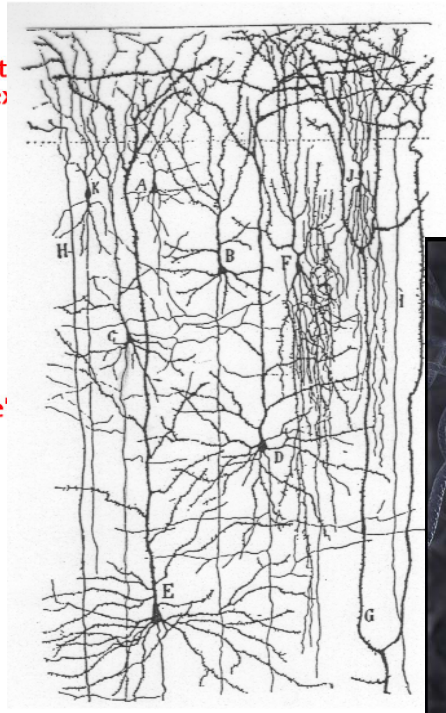


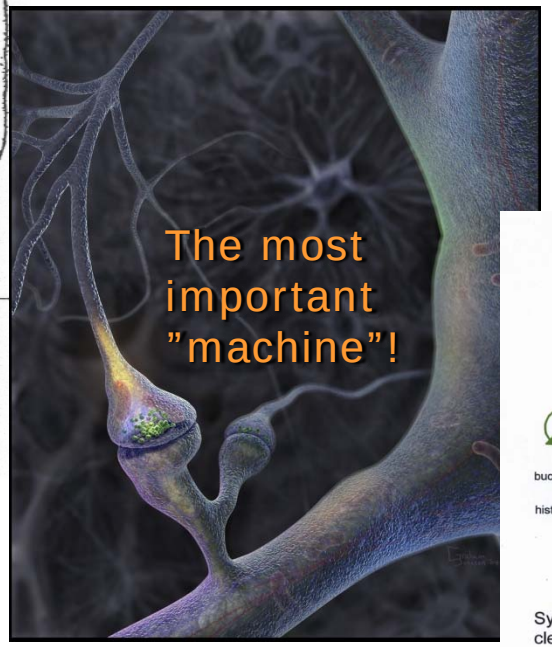
Brain simulation

A multi-scale problem

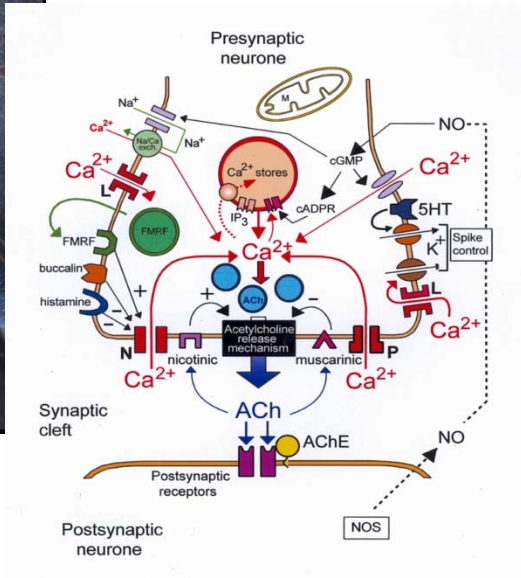
STOCKHOLM BRAIN INSTITUTE







Graham Johnson Medical Media, Boulder, Colorado



STOCKHOLM **BRAIN** INSTITUTE

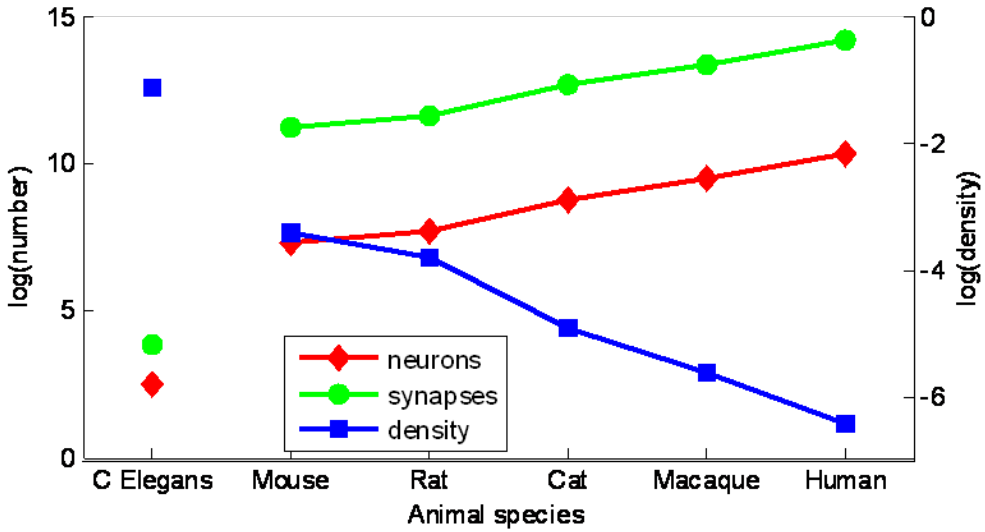
Why simulate the brain?

- To undercover the mystery
- To understand function and dysfunction
- To be able to interface with the brain
- To design brain-inspired technology
  - ”Cognitive computing”

STOCKHOLM BRAIN INSTITUTE

Why use HPC?

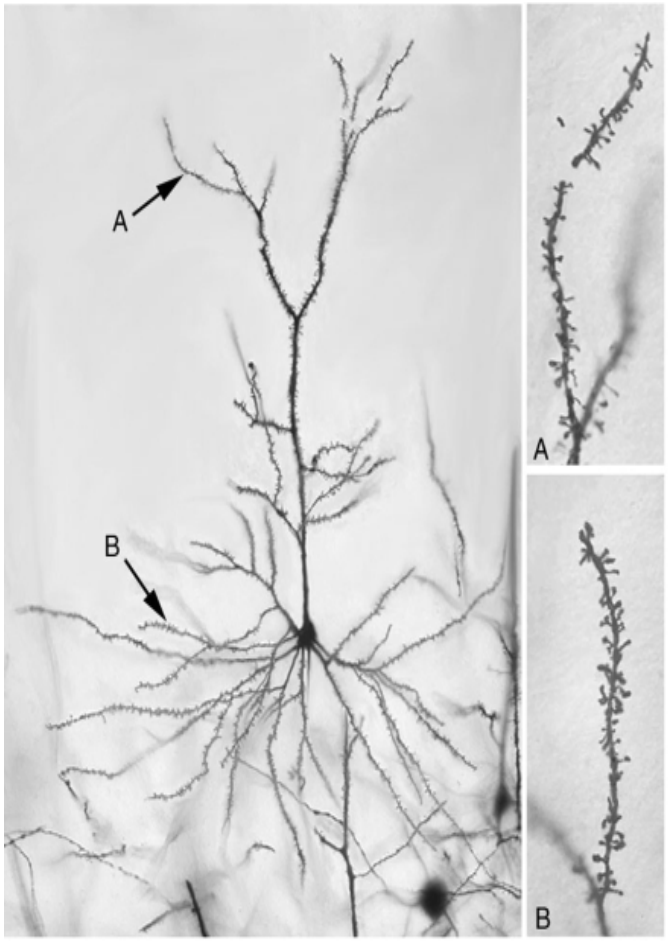
- Many neurons and synapses in reality
- Network is key
- Simulating small networks distorts results
- Distributed representation and processing for technical applications

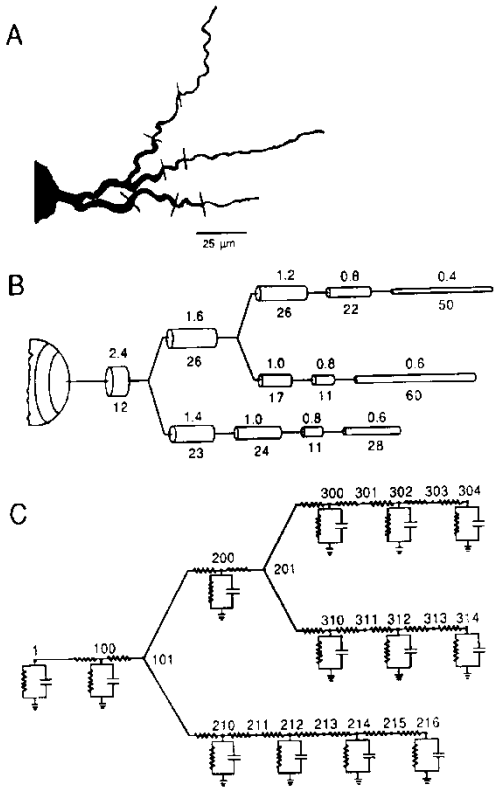


- 1000 neurons on PC slow

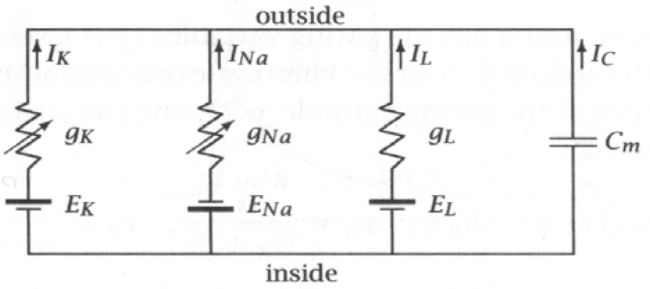
STOCKHOLM BRAIN INSTITUTE

Single cell modeling





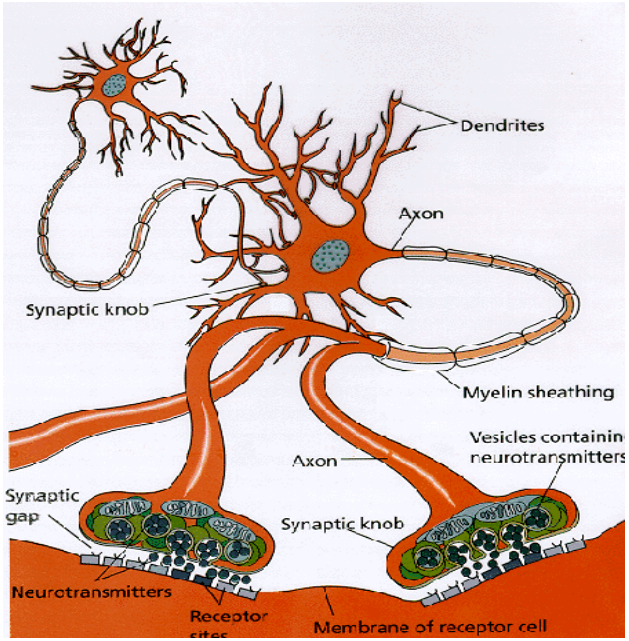
- Connectionist/Mesoscopic
- Integrate-and-fire
- Hodgkin-Huxley
  - Single/multiple compartment
- Equivalent electrical circuit



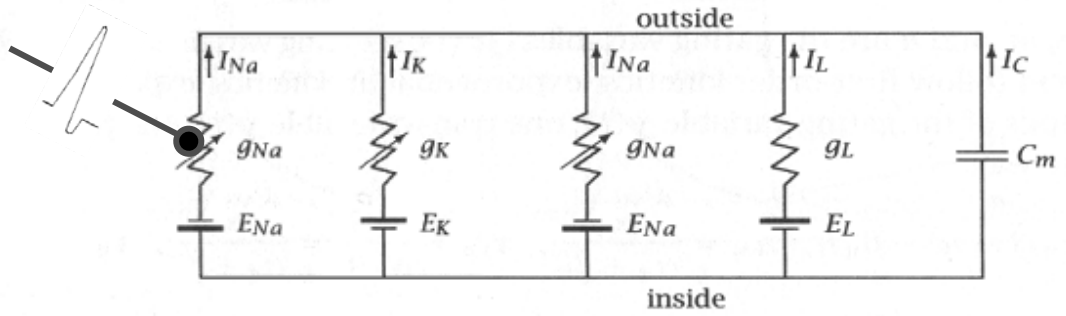
- Hodgkin-Huxley formalism
  - Na, K, K_{Ca}, Ca-channels
  - Ca_{AP} and Ca_{NMDA} pools
  - …
- Quantitative fit possible
- IBM-EPFL Blue Brain project

STOCKHOLM BRAIN INSTITUTE

Synaptic transmission and plasticity



- Conduction delay
- Synaptic conductance, kinetics, PSP-shape
- Reversal potential
  - Glutamate (AMPA&NMDA)
  - $GABA_A$



- Quantitative fit possible
- Synaptic plasticity
  - Fast: Facilitation, Depression
  - LTP/LTD: STDP
    - Biochemistry

APPENDIX
MODEL HODGKIN–HUXLEY NEURON EQUATIONS

This Hodgkin–Huxley implementation was based on Ekeberg et al. (1991). With the membrane potential V and the Nernst potential E_i for $i \in \{Na, K, Ca, K_{Ca}\}$ and given Ohm's law:

$I_i = g_i(V - E_i)$ combined with Kirchoff's laws, yields:

$$I_m = C_m \frac{dV}{dt} + g_{Na}(V,t)(V - E_{Na}) + g_K(V,t)(V - E_K) + g_{Ca}(V,t)(V - E_{Ca}) + g_K(V,t)(V - E_{Kca}) + g_{Leak}(V - E_{Leak})$$

Where t is time, g_{Leak} is the constant leak conductance and E_{Leak} is the leak equilibrium potential. Parameter values are given in **Table A1**. The dynamic conductance $g_i(V,t)$ can be expressed with a gating model for individual ion channels. Hodgkin and Huxley (1952) discovered these dynamics for Na^+ and K^+ ion channels (but not Ca^{2+}) after analyzing empirical data on action potentials in squid neurons. This resulted in the Hodgkin–Huxley equation:

$$I_m = C\frac{dV}{dt} + \bar{g}_{Na} m^3 h (V - E_{Na}) + \bar{g}_K n^4 (V - E_K) + g_{Leak}(V - E_{Leak})$$

Where $\bar{g}$ is the maximal conductance when a channel is open. Gating variable m is Na^+ channel activation, h is K^+ channel activation and h is Na^+ channel inactivation. The gating variables can be expressed as the following differential equations:

$$\frac{dm}{dt} = \alpha_m(1-m) - \beta_m m \quad \text{with} \quad \alpha_m = \frac{A(V-B)}{1 - e^{(B-V)/C}} \quad \beta_m = \frac{A(B-V)}{1 - e^{(V-B)/C}}$$

$$\frac{dh}{dt} = \alpha_h(1-h) - \beta_h h \quad \text{with} \quad \alpha_h = \frac{A(B-V)}{1 - e^{(V-B)/C}} \quad \beta_h = \frac{A}{1 + e^{(B-V)/C}}$$

$$\frac{dn}{dt} = \alpha_n(1-n) - \beta_n n \quad \text{with} \quad \alpha_n = \frac{A(V-B)}{1 - e^{(B-V)/C}} \quad \beta_n = \frac{A(B-V)}{1 - e^{(V-B)/C}}$$

Where constants A, B and C are independently defined for α and β of each channel. Ca^{2+} is treated differently, because Ca^{2+} pools are assumed to be inside the cell near the cell membrane and can activate K_{Ca} channels to achieve hyperpolarization. Using q to represent Ca^{2+} activation, a relation similar to the Na^+ channel activation (m) holds:

$$\frac{dq}{dt} = \alpha_q(1-q) - \beta_q q \quad \text{with} \quad \alpha_q = \frac{A(V-B)}{1 - e^{(B-V)/C}} \quad \beta_q = \frac{A(B-V)}{1 - e^{(V-B)/C}}$$

change in concentration $[Ca_{AP}]$ is equivalent to the rate of ions entering the pool less the ions leaving the pool:

$$\frac{d[Ca_{AP}]}{dt} = (E_{Ca} - V) Q_{AP} q^5 - \delta_{AP}[Ca_{AP}]$$

Where Q_{AP} is the rate of $[Ca^{2+}]$ influx and δ_{AP} is the rate of decay. The concentration $[Ca_{AP}]$ occurs in the soma of neurons and will activate K_{Ca} channels inside the cell membrane with the following current:

$$I_{K(Ca)AP} = \bar{g}_{K(Ca)AP}(E_k - V)[Ca_{AP}]$$

After sustained firing activity, calcium buildup in the neuron will cause hyperpolarization and a reduction in the firing rate (adaptation).

SYNAPTIC EQUATIONS

Like voltage-gated ion channels, the neurotransmitter-gated synaptic channels have ionic current produced by a voltage driving force, maximum conductance $\bar{g}_{syn}$ and synaptic activation level s with 1 being the most active. Given this, the synaptic AMPA and $GABA_A$ inward currents are:

$$I_{syn} = \bar{g}_{syn}(E_{syn} - V)s \quad 0 \le s \le 1$$

All synapses are saturating as defined by Lytton (1996) and depressing as defined by Varela et al. (1997). The combination of saturation and depression implemented in the same synapse (using NEURON) was not found in previous publications and appears to be new. Each pre-synaptic spike results in neurotransmitter release into post-synaptic receptors with a binding rate α, unbinding rate β, and synaptic conductance r. During neurotransmitter release C_{dur}, $r = r_{on}$ and conductance increases to $R_\infty = \alpha/(\alpha + \beta)$ with a time constant $\tau_{on} = 1/(\alpha + \beta)$. After C_{dur}, $r = r_{off}$ and conductance decays to zero with a time constant $\tau_{off} = 1/\beta$. With Δt being time since a spike, the dynamics are as follows:

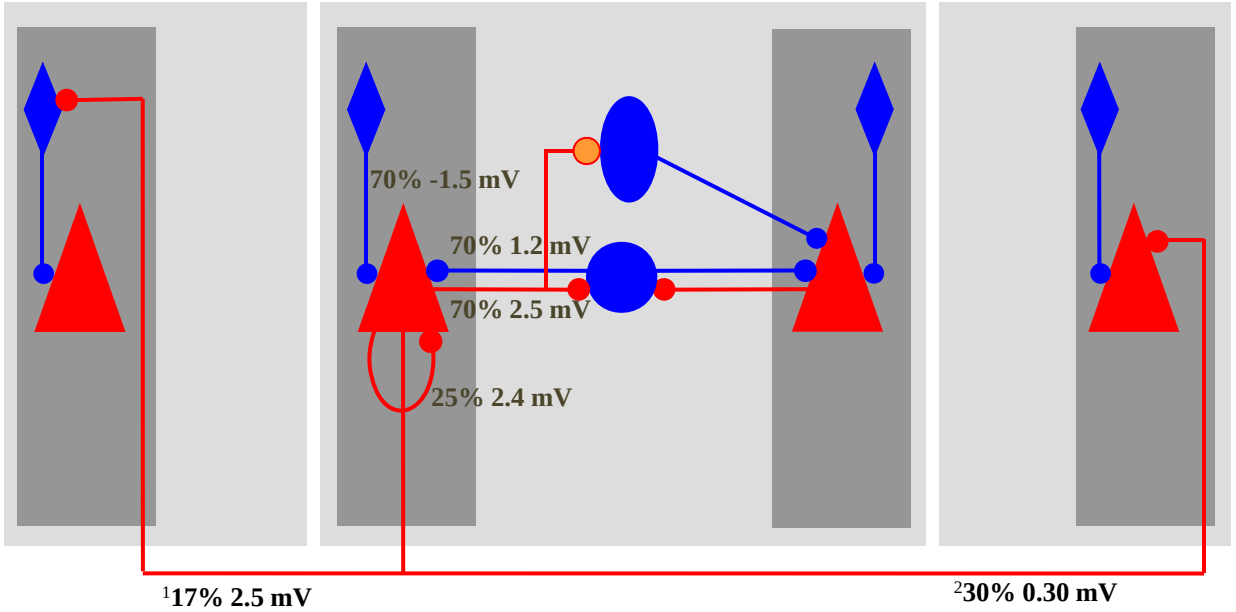
$$\frac{dr_{on}}{dt} = (\alpha + \beta)(R_\infty - r_{on}) \quad \Delta t \le C_{dur}$$

$$\frac{dr_{off}}{dt} = -\beta r_{off} \quad \Delta t > C_{dur}$$

$$s = \frac{r_{on} + r_{off}}{R_\infty}.$$

A layer 2/3 cortex model
Microcircuit layout “Icecube - Potts” like

STOCKHOLM BRAIN INSTITUTE

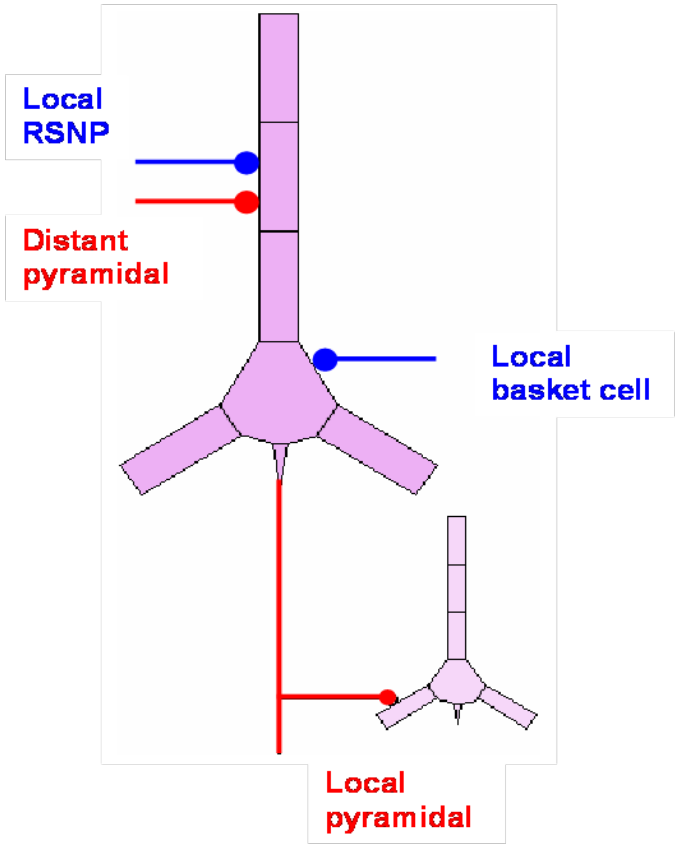


- Minicolumns/local sub-networks with
  - 30 pyramidal cells, connected 25%
  - 2 dendritic targeting, vertically projecting inhibitory interneurons
    - RSNP, e.g. Double bouquet
- Hypercolumns (soft WTA modules) with
  - Pool of Basket cells
  - Martinotti cells, with facilitating synapses from pyramidal cells
  - Large models: 100 minicolumns, 200 basket + Martinotti cells per hypercolumn
- Currently rudimentary layers 4 and 5

The layer 2/3 cortex model

Synaptic properties and connectivity

STOCKHOLM BRAIN INSTITUTE



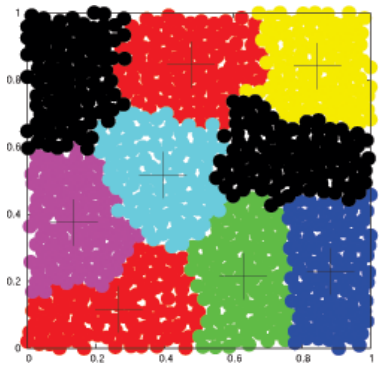
- **Synaptic transmission**
  - **Glutamate (AMPA & voltage dependent NMDA)**
  - **Depressing synapses**
  - **$GABA_A$**
- **Synaptic targeting of soma and dendrites**
- **3D geometry ⇒ delays**
  - **0.1 - 1m/s conduction speed**
- **Realistic amplitude of PSP:s in larger network models**

Pyramidal-pyramidal fast synaptic depression
[Tsodyks, Uziel, Markram 2000]

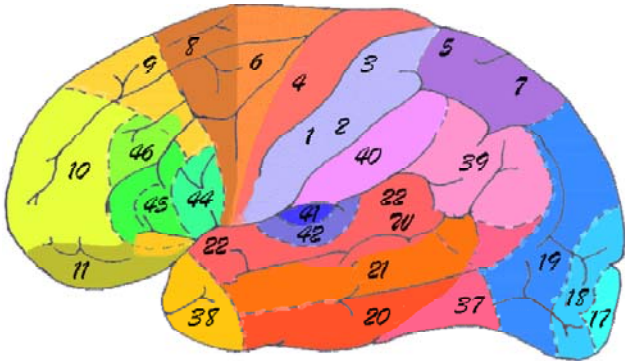
Conceptual model of Neocortex

STOCKHOLM BRAIN INSTITUTE

Hypercolumns



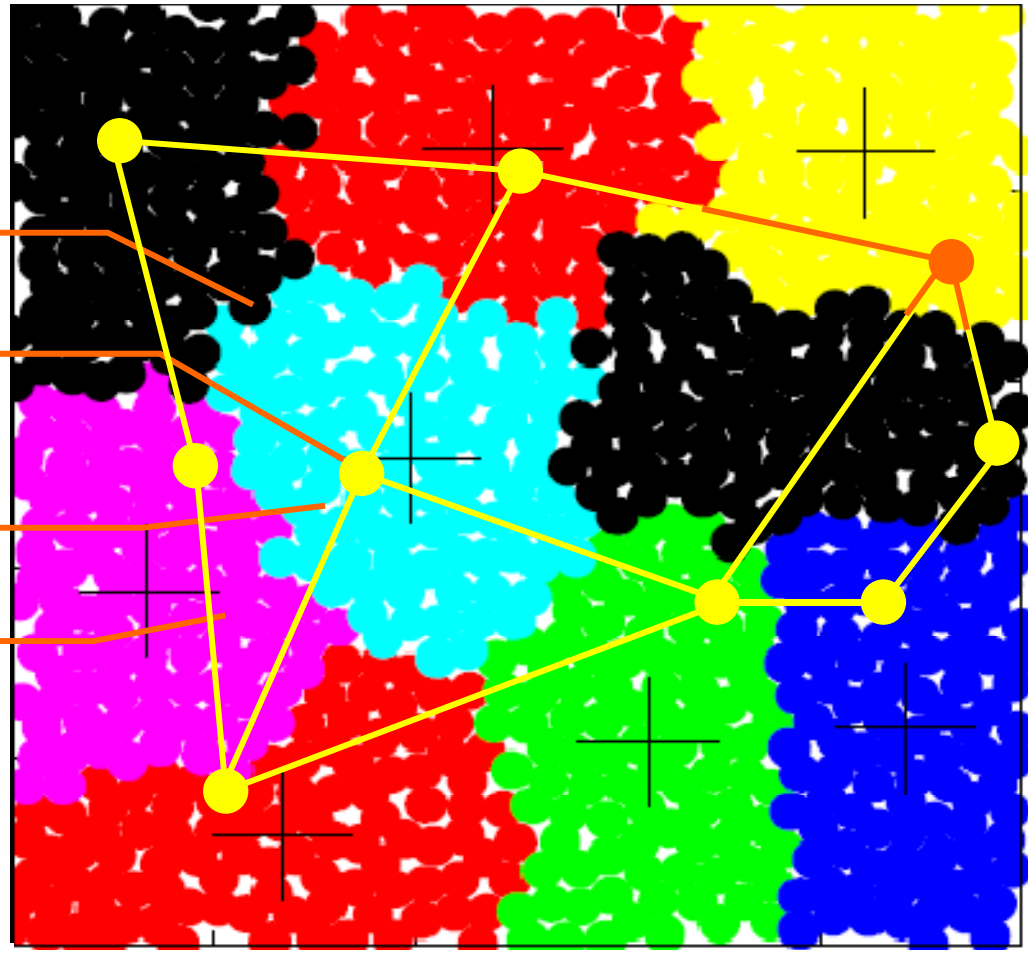
Cortical areas



- Hypercolumns are grouped into cortical areas of various sizes
  - Human V1 has ~40000 hypercolumns
- Human neocortex has about 110 cortical areas (Kaas, 1987)

Simulation model network layout

STOCKHOLM BRAIN INSTITUTE



- 1x1 mm patch
- 9 hypercolumns
- Each hypercolumn
  - 100 minicolumns
  - 100 basket cells
- 29700 neurons
- 15 million synapses
- 100 patterns stored

- W trained offline
- (A-)symmetric

STOCKHOLM BRAIN INSTITUTE

9 hypercolumns

Spontaneous activity

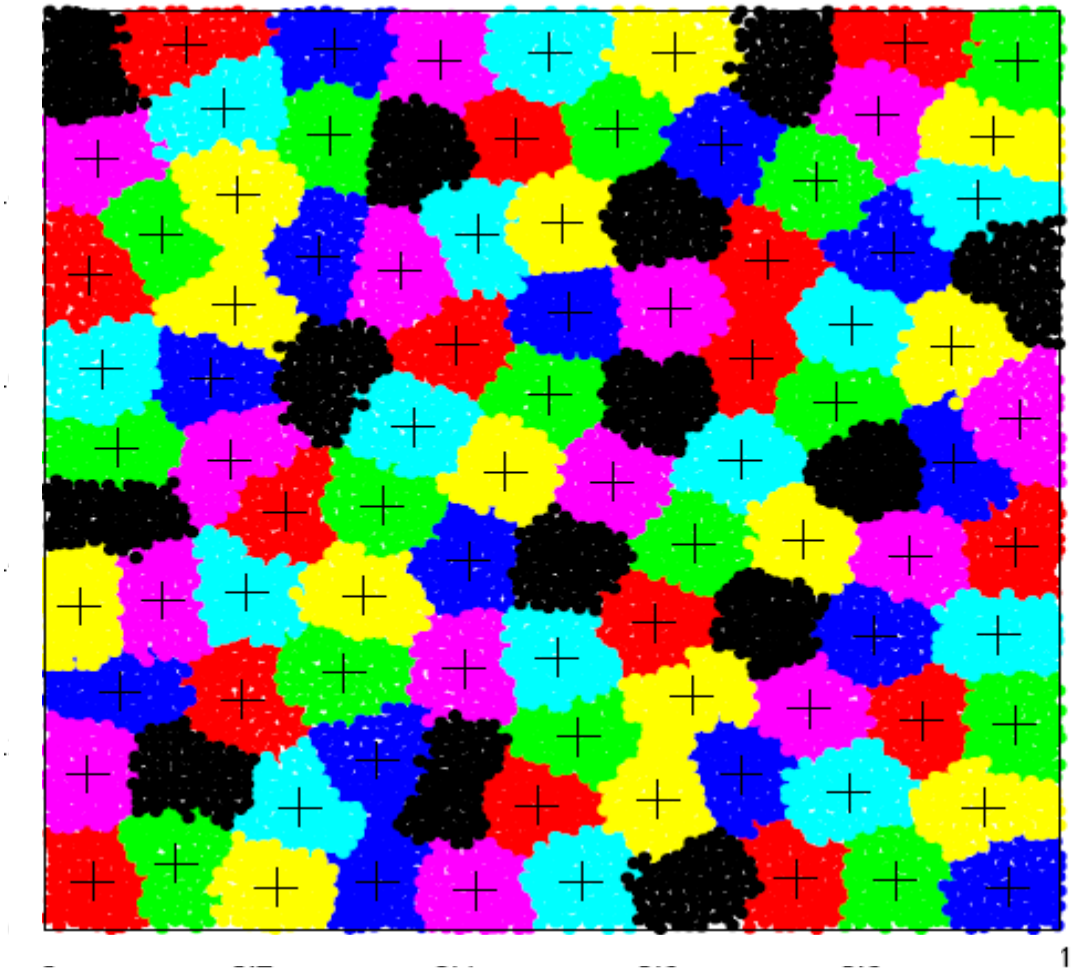
- 1x1 mm patch
- 9 hypercolumns
- Each hypercolumn
  - 100 minicolumns
  - 100 basket cells
- 29700 neurons
- 15 million synapses
- 100 patterns stored

- Non-symmetric W

STOCKHOLM BRAIN INSTITUTE

100 hypercolumns

Spontaneous activity



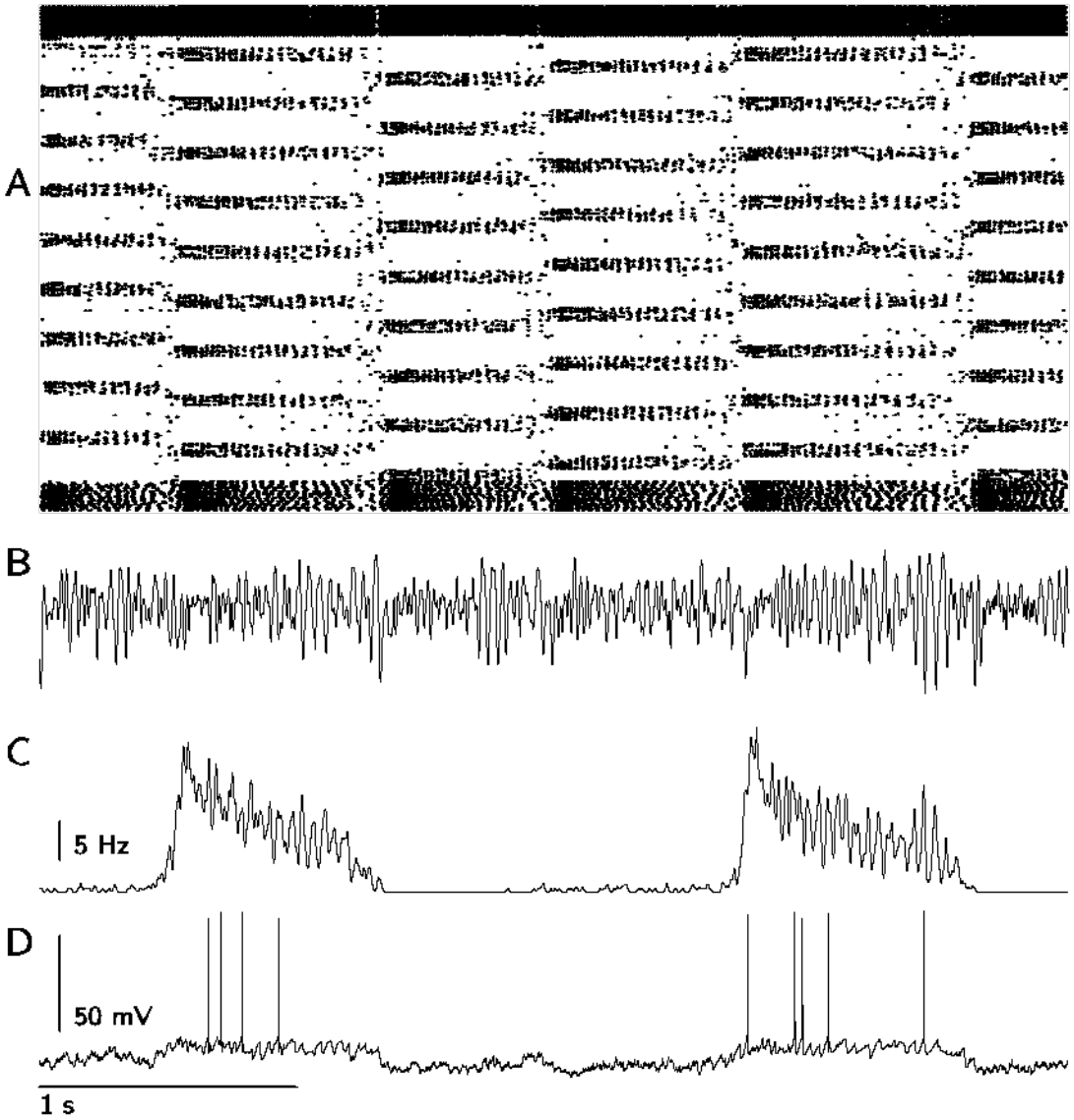
- 330000 neurons
- 161 million synapses

≈ 4x4 mm

STOCKHOLM **BRAIN** INSTITUTE

- **2000+ neurons**
- **250000+ synapses**

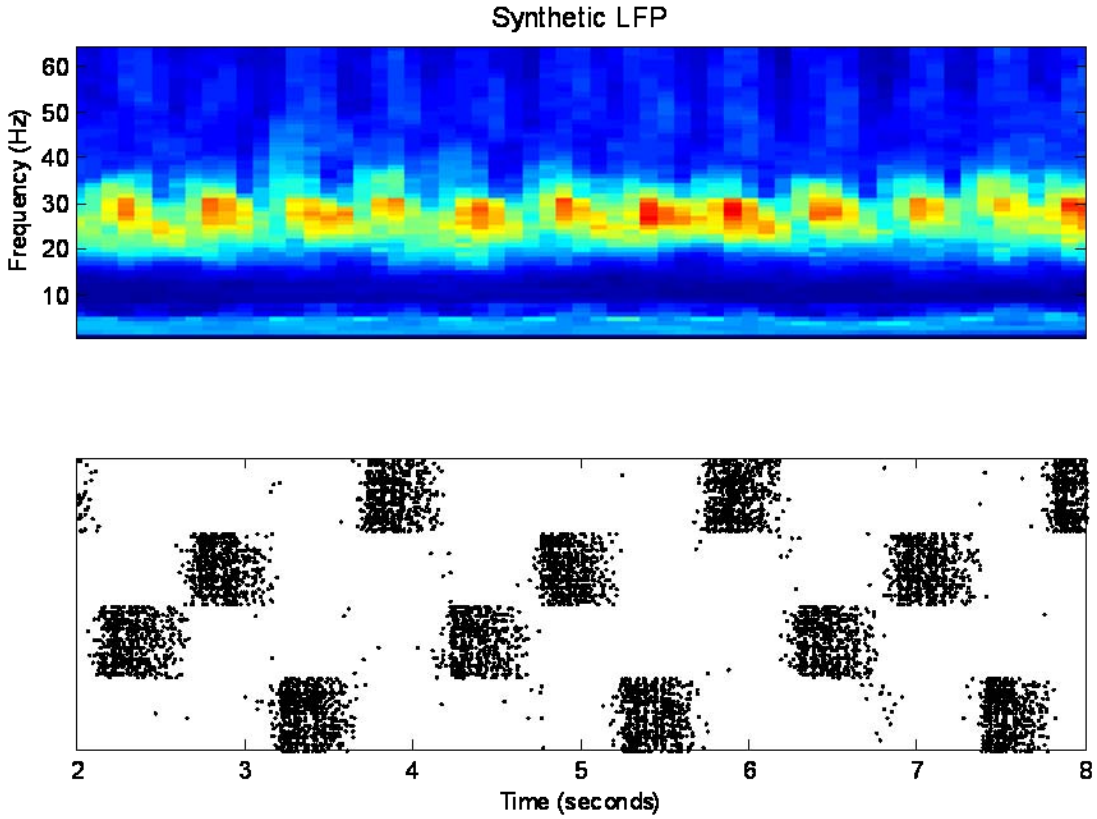
Spontaneous "resting activity" Tsodyks, Grinvald et al. 1999



Lundqvist M, Rehn M, Djurfeldt M and Lansner A (2006). Attractor dynamics in a modular network model of the neocortex. *Network: Computation in Neural Systems*: 17, 253-276

STOCKHOLM **BRAIN** INSTITUTE

Spontaneous attractor ”hopping”



Can be compared to experimental data

- **Memory replay at theta**
  - **Fuentemilla et al. Curr Biol 2010**

Computational demands
Mouse brain ≈ 1/1000 Human brain

STOCKHOLM BRAIN INSTITUTE

- Number of state variables in model neurons (6 compartments) ≈ 30
- Flops per neuron update: ≈ 100

- Number of state variable in model synapses: ≈ 3
- Flops per synapse update: ≈ 20

- Number of neurons: 30 M
- Number of neurons: 100 G
- Memory 30*4B*30M + 3*4B*100G ≈ 1TB

- Simulation time step: 50 μs
- Flops/simulated second: $(100*30M + 20*100G)/(50*10^{-6}) \approx 4*10^{16}$
- Simulation time per simulated second (1 Pflop) = 40 sec

STOCKHOLM BRAIN INSTITUTE

Parallelization of brain simulations

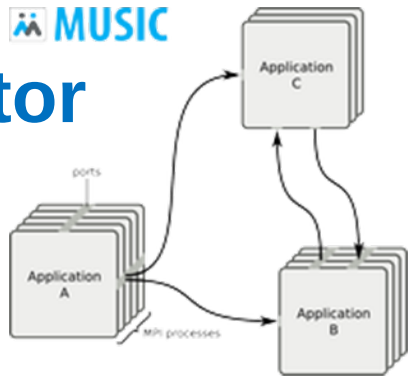
- Distributed processing, local memory (synapses mostly)
- Very sparse connectivity (human cortex 10^{-6})
- Spiking (event based) communication
  - Some graded interactions

- $\Rightarrow$Readily parallelized

Available HPC simulators

STOCKHOLM BRAIN INSTITUTE

- SPLIT (Hammarlund, Ekeberg 1998)
- PGENESIS, Parallel Genesis (Bower, ...)
- PNEURON, Parallel Neuron (Hines, ...)
- NEST (NEST Initiatiative)
- MOOSE (Bhalla, ..., β-version)
  - Integrated subcellular dynamics
- MUSIC – MUltiSImulation Coordinator
- GPU simulators



STOCKHOLM BRAIN INSTITUTE

8 rack BG/L simulation

October 2006

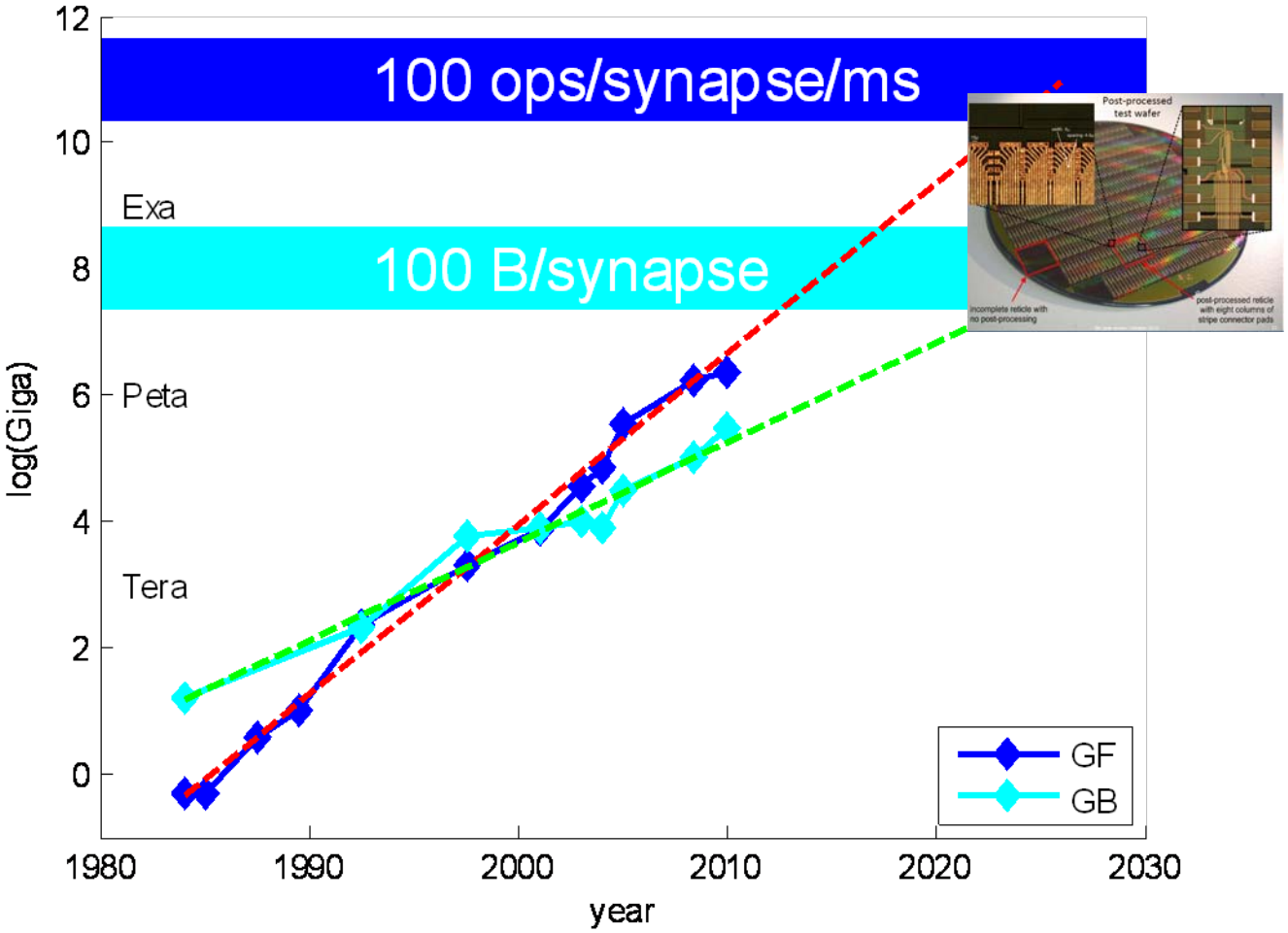
- 22x22 mm cortical patch
  - 22 million cells, 11 billion synapses
- SPLIT simulator by KTH
- 8K nodes, co-processor mode
  - used 360 MB memory/node
- Setup time = 6927 s
- Simulation time = 1 s in 5942 s
- Massive amounts of output data
  - Network operation similar as smaller versions of the model

- 77 % estimated speedup
  - Point-point communication slows (?)
  - Linear speedup to 4K nodes

Djurfeldt M, Lundqvist M, Johansson C, Rehn M, Ekeberg Ö, and Lansner A (2008): Brain-scale simulation of the neocortex on the Blue Gene/L supercomputer. IBM J R&D 52:31-41

STOCKHOLM BRAIN INSTITUTE

When can we simulate the human brain in real time?



Simulation output pre-post processing

STOCKHOLM BRAIN INSTITUTE

- Preprocessing
  - Setting up simulation model
    - Predefined networks of neurons and synapses
    - W files may be large!
  - Building/Growing network
    - E.g. By self-organization
    - From Large datasets
- Postprocessing
  - TB output files ...
  - Sub-sampling of output, online postprocessing
  - Synthesis of measurements
    - VSD, LFP, EEG, MEG, BOLD
    - Spike analysis
  - 3D visualization

STOCKHOLM BRAIN INSTITUTE

HPC results overview

- Full scale lamprey swimming simulations
- First (?) parallelizing Hodgkin-Huxley brain simulator
- Largest ever Hodgkin-Huxley cortex network models
  - with David Silverstein
  - Brain simulation, Cortex modelling
  - Currently scaling up (JUGENE) using NEURON
- Most parallel ever spiking and non-spiking neural network simulations
  - with Simon Benjaminsson
  - February 2011 on JUGENE
    - Largest multi-core installation (294912 cores)

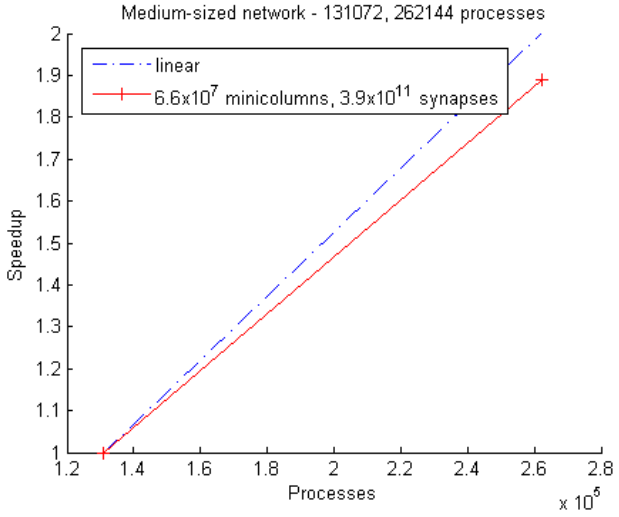
- Brain imaging data analysis

STOCKHOLM BRAIN INSTITUTE

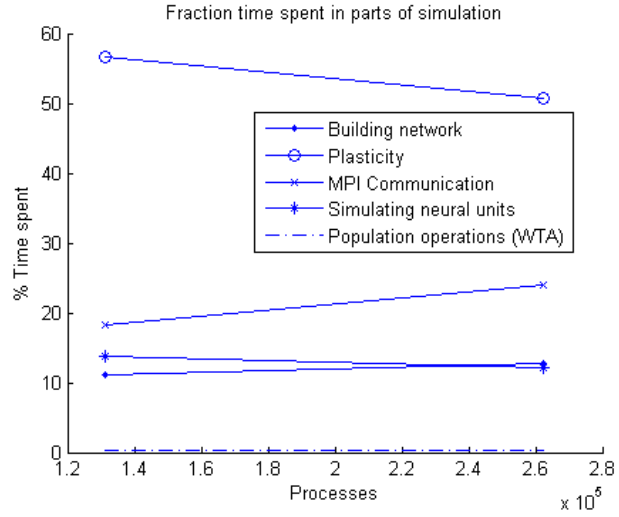
Recent scaling experiments: ANSCore program

TRITA-CSC-CB 2011:01

- **ANSCore on full machine**
  - Non-spiking ANS code (Simon Benjaminsson)
  - Collective communication
  - ANSCore good strong scaling
    - N = 66M (non-spiking units, “minicolumns”)
    - C = 390G
    - Inner loop time = 2200 ms (w/o plasticity)



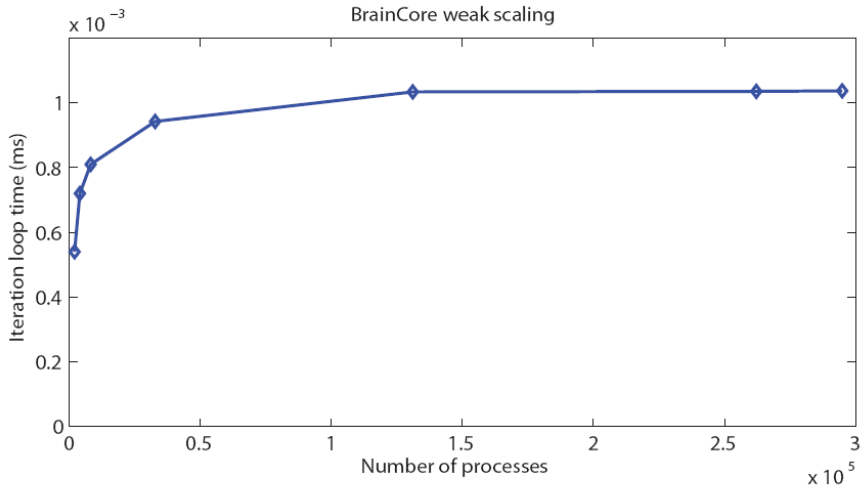
- **Recent tests on Cray XE6 (ANSCore)**
  - 6000 cores, 4-5 times faster than JUGENE, 3 times ← faster clock ...?



Recent scaling experiments. BrainCore program

STOCKHOLM BRAIN INSTITUTE

- BRAINCore on full JUGENE
  - >29M spiking units, event based communication
    - MPI point-point communication
    - Largest network 0.034 % processor-processor communication
      - ... as in the real brain
  - >295G trainable connections
  - Perfect weak scaling
    - Inner loop time < 1 ms (real time learning and recall)
    - Size ≈ mouse brain
    - Complexity << mouse brain
  - Remember:
    - 25x25x2 m machine
    - 1 MW



STOCKHOLM BRAIN INSTITUTE

What difficulties we encountered?

- MPI programming
  - Mapping problem to parallel architecture
  - Debugging, profiling
- New methodology/Technology for optimal use of resources
  - Scripting, temporary saves, ...
  - Model fitting possibilities
- Queing Policy – different user groups needs
  - We use MANY cores but currently run short jobs ...
- Hard to build large networks
  - Public simulators not programmable enough
  - Need input from file (W) – large!
- Potentially huge simulations output
  - Subsampling, postprocessing necessary

STOCKHOLM BRAIN INSTITUTE

Conclusions and Outlook

- Next generation simulators for brain simulation
  - Rapidly growing HPC activities
  - Multi-scale
  - Complex multi-network brain-like architecture

- What would have been 'good to get' from HPC in future?
  - Exascale computers will allow full scale human brain simulation with considerable detail
    - Need one on desktop ...
  - Expert support on
    - MPI programming
    - Data postprocessing
    - 3D visualization

STOCKHOLM **BRAIN** INSTITUTE

Thanks for listening!