Use of Neural Models for Cognitive Processing
2004 Sandia-UNM Cognitive Systems Workshop

James C. Brakefield © 2004

Abstract:

Results of a feasibility study show that system level neuron and synapse models can be implemented in affordable scalable real-time systems. Current limitations in scientific knowledge make it impossible to accurately model complete mammalian nervous systems. However, cognitive systems can be designed and tested without biological details, and therefore allow an avenue of research using neurological based cognitive processing. Example pulse mode neurological models and implementation architectures will be presented.

Contents:

Use of Neural Models for Cognitive Processing (cont'd)

Biases and Prejudices

Quantum effects don't matter:

Anthony J. Bell: ***Levels and loops: the future of artificial intelligence and neuroscience***, Phil. Trans. R. Soc. Lond. B (1999) 354, 2013-2020.

Consciousness irrelevant:

Look to Gazziniga's interpreter to understand consciousness:
M.S. Gazzaniga: The Social Brain: Discovering the Networks of the Mind, 1985, Basic Books. http://www.dartmouth.edu/~cogneuro/Gazzaniga.html

Intelligence a function of scale and architecture and chemistry

Scale:	Typical neurons have 6000 inputs & outputs
Human brain:	10^{10}-10^{12} neurons
Architecture:	More than just a uniform neo-cortex surface
Chemistry:	Dopamine effects?

Use of Neural Models for Cognitive Processing (cont'd)

Biases and Prejudices

Scale matters

Lesson from Mandlebrot: new tools lead to new science:
Computer graphic workstation allowed visual investigation of fractals:
Benoit B. Mandelbrot: The fractal geometry of nature, 1977, W.H. Freeman.

Lack perspective for devices with thousands of inputs and outputs.
Sparse Distributed Memory (SDM) and others typically use 500+ inputs.
http://www.rni.org/kanerva/homepg.html

Usual methodology for high In/Out counts: Operate in time domain

Usual methodology for high unit numbers: Analyze in continuous domain

Use of Neural Models for Cognitive Processing (cont'd)
Biases and Prejudices

Timing matters

A general purpose simulator must handle axon and dendrite transmission delays

Izhikevich E.M., Gally J.A., and Edelman, G.M.: ***Spike-Timing Dynamics of Neuronal Groups*** (2004) Cerebral Cortex, in press.

Spiking models now proven necessary to match biological data:

Simon Thorpe: ***Ultra-rapid scene categorization with a wave of spikes***.
(150 ms to do image processing, or ~10 ms per neural stage)

Jacques Sougne & Robert French: ***Synfire Chains and Catastrophic Interference***,
Proceedings of 23rd Annual Conference of the Cognitive Science Society.
(millisecond accuracy of neuron firing after a delay of 300ms)

Use of Neural Models for Cognitive Processing (cont'd)

Biases and Prejudices

System level modeling

Must do science on the whole to completely understand the whole:

Some ~50 brain regions with extensive interconnections, so each region in isolation presents an incomplete picture.

Cortex exhibits same phenomena in even greater complexity.

The brain is an engineered machine, not a general-purpose computer.

Attempts to use neural models to engineer cognition somewhat like early history of airplanes.

Use of Neural Models for Cognitive Processing (cont'd)

Brain Simulation Issues

General purpose models:

Some three to four thousand distinct neuron types of operation and morphology in mouse brain, most of which have not been studied as yet.

A parameterized model which encompasses most neural types of operation:

Eugene M. Izhikevich: *Which Model to Use for Cortical Spiking Networks*, IEEE. Trans. On Neural Networks (2004, submitted). Also: *Simple Model of Spiking Network*, ibid. (has Matlab & C++ source code).

Missing biological details: connectivity and long-term synapse behavior !!!

Connections have higher information content than synapse state.
Log base 2 of number of potential synapse inputs is ~34 bits
Log base 2 of number of synapse configurations < 8 bits
Log base 2 of synaptic delay ~6 bits (40ms at 0.5ms resolution)
33 >> 8 + 6
Also holds true for digital chips

Use of Neural Models for Cognitive Processing (cont'd)

Brain Simulation Issues

Hardware models of learning have characteristics not present in software models
Hardware is always active, e.g., computing
No penalty for parallelism
Must consider physical implementation of connectivity

Learning can take place via circuit design:
Valiant, LG (1984) ***A theory of the learnable***, Communications of the ACM 27(11): 1134--1142.

Also consider decomposition theory:
B. Wurth, K. Eckl, and K. Antreich**, *Functional multiple-output decomposition: Theory and an implicit algorithm***, in Proc. ACM/IEEE Design Automation Conf., June 1995, pp. 54--59.
Michael Burns, Marek Perkowski, and Lech Jozwiak, ***An Efficient Approach to Decomposition of Multi-Oputput Boolean Functions with Large Set of Bound Variables,*** in Proc. 1998 Euromicro, pp. 16-23, Vasteras, Sweden, August 25-27, 1998.
Christoph Scholl, ***Multi-output Functional Decomposition with Exploitation of Don't Cares***, May 1997.

For practical work with learning via circuit design, take a VHDL or Verilog compiler (from Exemplar, Synplicity, Synopsys) and compile incompletely specified truth tables into ASIC gates, FPGA lookup tables or CPLD macro-cells.

Use of Neural Models for Cognitive Processing (cont'd)

Hardware Simulation Alternatives

FPGA implementation superior to microprocessor implementation

Greater range of implementation choices

- Data width adjustable
- Pipelined processing

High speed self clocking serial interfaces readily available (800 Mhz and higher)

Parallelism readily available

Greater memory bandwidth (500 signal pins typical)

Same chip does store & forward function as well as processing

For digital circuits the information content is in the connections

ASICs: 2 and three input gates, thousands to millions of possible inputs

- Three bits to identify gate type, 20 bits per input to identify each input
- E.g., 40 bits for inputs, 3 bits for logic

FPGAs: 3, 4, 5, 6 input lookup tables (LUTs), thousands of possible inputs, input routing takes many more bits than LUT contents:

16 bits to configure 4-input LUT, ~16 bits * 4 to identify all inputs, in practice it takes ~50 bits per input to route each signal to the LUT.

- E.g., ~200 bits for inputs, 16 bits for logic

Use of Neural Models for Cognitive Processing (cont'd)

Hardware Simulation Alternatives

Implementation choices

(all using 3D torus interconnect of identical processing units)

1) Microprocessor farm (e.g. Red Storm).

2) Micro-controller + DRAM (~1GB for synapse data) + FPGA (for store & forward network). (e.g. ARM microprocessor & memory on small circuit board)

3) FPGA (with on chip micro-processor) + DRAM (~1GB).

4) Synchronous simulation: FPGA + DRAM (~1GB).

5) Asynchronous simulation: FPGA + DRAM (~1GB) + dual-port SRAM (event queues).

In all cases small FPGA or portion of FPGA or an ASIC needed for 3D torus interconnect.

Use of Neural Models for Cognitive Processing (cont'd)
Hardware Simulation Alternatives

Asynchronous simulation

~ 1ms time step

Computational effort is lower than with synchronous simulation

Requires dual-port RAM for event queues

- One port to post events, driven from incoming and outgoing spikes
- One port to process events
- Higher cost & circuit board area due to dual-port RAM chip(s)

DRAM Memory used mostly in random access mode, so high memory bandwidth not possible

Architectural parallelism:

1) Store & forward network
2) Incoming spike processor
   - DRAM operated in pipeline mode (for synapse examination)
3) Event processors
   - Four way parallel processing possible
   - DRAM operated in random access mode

Use of Neural Models for Cognitive Processing (cont'd)
Hardware Simulation Alternatives

Synchronous simulation

Minimum update rate of 40Hz, 100Hz affordable

Memory bandwidth requirement scales with update rate

To increase update rate

Increase number of FPGAs

Increase DRAM data rate

Total DRAM stays the same

Memory accessed in pipeline (streaming) mode

High parallelism is pin limited

Conversion of asynchronous models to synchronous implementation:

Spike activity in previous cycle encoded into ~8-bit representation

Architectural parallelism:

1) Store & forward network

2) Synapse processors (~20 operating at 200 Mhz)

DRAM interface (480 bits wide operating at 600 Mhz?)

Each incoming axon has an on-chip RAM word for "spike" representation

Use of Neural Models for Cognitive Processing (cont'd)
Hardware Simulation Alternatives

Experiment Process Flow

Given the connectivity and model parameterization:

1) Derive axon, synapse, spine & dendrite-tree delays
2) Neurons allocated to FPGA chip/modules topographically
3) Parallelism requires allocation of synapses and/or neurons to various pipes
4) Memory locations allocated
5) Axon destination ranges (voxels on the 3D torus) allocated
6) FPGA chips configured
7) DRAM initialized
8) Run made
9) DRAM contents saved for analysis

Use of small scale Processor Farm (e.g. Red Storm) for support purposes:

- Disk drives to load/save DRAM
- Monitor runs
- One processor per ~1K FPGA modules?

Use of Neural Models for Cognitive Processing (cont'd)
System Simulation "Rational"
Real-Time Digital Simulation of the Human Nervous System

The objective is to study the feasibility of a real-time system level simulation of the human brain. The result is that it is considered feasible using FPGA and DRAM technology at a chip cost of $100+ Million US dollars.

Why even consider such an ambitious project?

A) The study itself adds insight into current knowledge of the brain.

B) It provides a framework from which to organize existing knowledge.

C) It forces component level models of neurobiology to be considered from the system or whole.

D) Silicon costs are expected to continue to decline. In a decade, the issue will not be why, but when and by whom. (see www.ad.com/tech.html, www.cybernetics.demon.co.uk/Plan.html, www.artificialbrains.com)

E) A configuration reduced by one thousand times, is suitable for the system level simulation of small mammals. Silicon cost also reduces by one thousand, leading to a tool affordable by many research institutions.

F) The reduced configuration is compact and therefore may have military use.

G) The system level perspective puts new light on what is overlooked by researchers not interested in system level issues.

H) A facility and the reduced scale facility are useful for many other kinds of complex system modeling.

Use of Neural Models for Cognitive Processing (cont'd)
Micro-Electronics Background

AMD "Hammer" Opteron PC Chip: 2-3Ghz, 64-bit address space,
"Red Storm": http://www.cray.com/news/0210/sandia_redstorm.html, 10K chips,
3D mesh inter-connect, for Sandia National Laboratories

Samsung "Halla" ARM Chip: ~600Mhz, 32-bit address space, integer only, also Intel PXA27x
http://www.convergencepromotions.com/pdf/Samsung_article.pdf

Intrinsity ***Fast***MATH signal processor: 2Ghz 32-bit processors, 17 per chip
http://www.intrinsity.com/products/products_fastmath_set.htm

FPGAs with 200-300 Mhz clock rates: Xilinx (www.xilinx.com)
Spartan-3 XCS2000: 565 I/Os, **$25** each in high volume, Has:
DDR (double data rate) I/Os, (40) 18K dual port RAMs, (40) 18x18 multipliers

Other FPGAs: Altera(www.altera.com) & Lattice Semiconductor(www.latticesemi.com)
Altera: Stratix & Stratix II: similar to Spartan-3 and/or Virtex II
Xilinx: Virtex II Pro has embedded PowerPC processors

RLDRAM: Commodity DDR DRAM: ~$0.10 per megabyte
Dual-Port RAM: Cypress and IDT: up to 18 Mbits, up to 72-bit I/O.

Use of Neural Models for Cognitive Processing (cont'd)
Human Brain 101

Basic Facts (http://faculty.washington.edu/chudler/facts.html):

1,300-1,400 grams:	$1.3*10^6$ mm^3	
60% gray matter:	$0.8*10^6$ mm^3	(gray matter is neurons, white matter is wiring)
cortex surface area:	$0.20\text{-}0.25*10^6$ mm^2	(cortex is largest part of human brain)
cortex thickness:	average is 2.0 mm,	range of 1.5 to 4.5 mm
neurons in cortex:	$10\text{-}20*10^9$,	100K-150K per mm^2
synapses in cortex:	$60\text{-}240*10^{12}$,	10^8-10^9 per mm^3
Koch:	page 87: average of $6*10^8$ per mm^3,	up to 10^9 per mm^3
Shepherd:	page 471 (Beaulieu & Colonnier):	$2.78*10^8$ per mm^3
Sejnowski:	page 51:	10^9 per mm^3, recently $2*10^9$ per mm^3

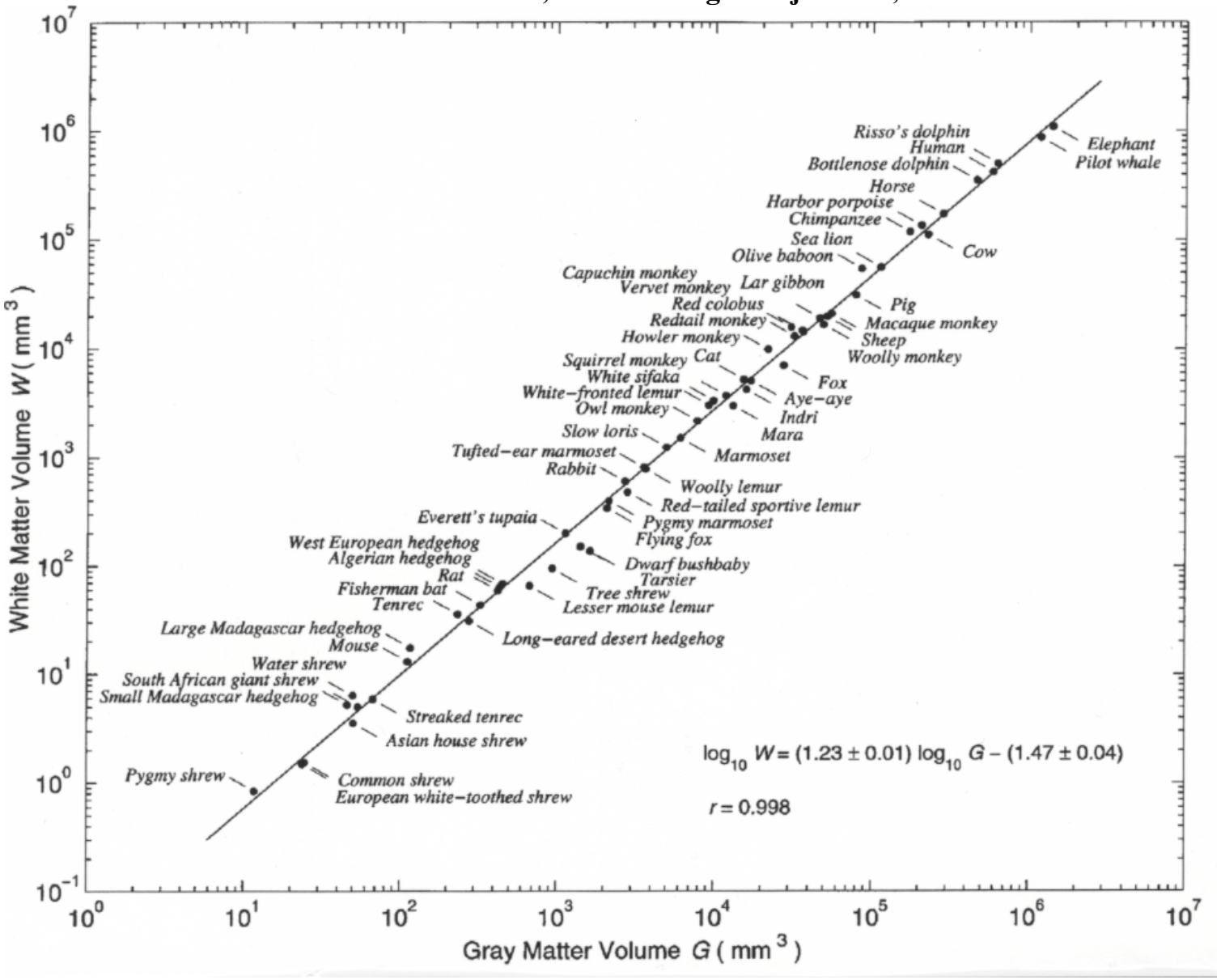
(Cortex dominates brain, cerebellum has ~$30*10^9$ neurons and only ~$6*10^{12}$ synapses)

neuron firing rate: 1-200 Hertz, resting: 1-10 Hertz, visual cortex: ~40 Hertz

Shepherd: The Synaptic Organization of the Brain, 4th ed., 1998, Oxford University Press
Koch: Biophysics of Computation: Information Processing in Single Neurons, 1999, Oxford Press
Churchland & Sejnowski: The Computational Brain, 1992, MIT Press

Use of Neural Models for Cognitive Processing (cont'd)

Mammal Brain Sizes, From Zhang & Sejnowski, PNAS 97:10:5622



Use of Neural Models for Cognitive Processing (cont'd)
Neuronal Modeling

System Level Neuron Models:

Izhikevich: http://www.nsi.edu/users/izhikevich/publications/
Mathematician, Matlab and C++, Single parameterized model

Maass: http://www.lsm.tugraz.at
Computer Scientist, Matlab and C++, Pulse mode, adaptive
Maass & Bishop eds: Pulsed Neural Networks, 1998, MIT Press

Koch: http://klab.caltech.edu/index.shtml
Neurobiologist
Koch & Segev eds: Methods of Neuronal Modeling, 2nd ed, 1998, MIT Press

Hopfield: http://neuron.princeton.edu/~moment/Organism/index.html
Physicist, Matlab

Van Essen & Anderson: http://stp.wustl.edu/
Neurobiologist, NESim written in Matlab, Java user interface

Thorpe, Delorme & VanRullen: SpikeNET: http://www.spikenet-technology.com
French/USA reseachers, SpikeNET does fast image recognition

Use of Neural Models for Cognitive Processing (cont'd)
Neural Biology Web sites

Boston University & Harvard & others: http://synapses.mcg.edu/
Reconstruction from electron microscope sections
3D animations(from reconstructions), electron-microscope pictures, really great web site !!

Instituto Cajal: http://www.cajal.csic.es/valverde/
Web site has three Gogli method slices of postnatal mouse, for more slices:
Facundo Valverde: Gogli Atlas of the Postnatal Mouse Brain, 1998, Springer-Verlag
See also: George Paxinos & Keith Franklin: The Mouse Brain in Stereotaxic Coordinates, Deluxe Second Edition (with CD-ROM), 2000, Academic Press.

Cajal Neuroscience Research Center: http://marlin.life.utsa.edu/
Has axon arbors as well as dendrite trees
Bennett & Wilson: "Synaptology and Physiology of Neostriatal Neurons" in: Brain Dynamics and the Striatal Complex, Miller & Wicken eds, in press, Harwood Academic

Dissection: http://www.vh.org/adult/provider/anatomy/BrainAnatomy/BrainAnatomy.html

Mammal brain collections: http://brainmuseum.org/

Mammalian cortex sizes: Zhang & Sejnowski, "A universal scaling law between gray matter and white matter of cerebral cortex", PNAS, May 9, 2000, vol.97, no.10, pgs 5621-5626 (on the web).

Use of Neural Models for Cognitive Processing (cont'd)
Engineering Considerations

Synapse state: 48 bits per synapse:

- Connectivity: if 10^{10} neurons, then 33 bits per synapse for source identification
  - Because axon arbors are compact,
  - Half of connectivity handled by routing, and
  - Other half by RAM address (~16 bits)
- Aggregate delay: 8 bits
- Strength and type of synapse: 8 bits
- State: 8 bits, double buffered ("current" and "update" buffers)

FPGA chip + DRAM

- ~2 inch square circuit board
- FPGA on one side, ~16 DRAM chips on other
- 400-600 Mhz DRAM data rate on 480 FPGA pins
- Processing at 200-300 Mhz rate on each of 20 parallel processing streams
- ~25,000 neurons per circuit board at a 40 Hertz update rate
- May need a heat sink (~10 watts per circuit board)

Interconnections

- Flattened 3D torus with 200 Mhz serial links between nodes
- Part of each FPGA chip devoted to store & forward mechanism

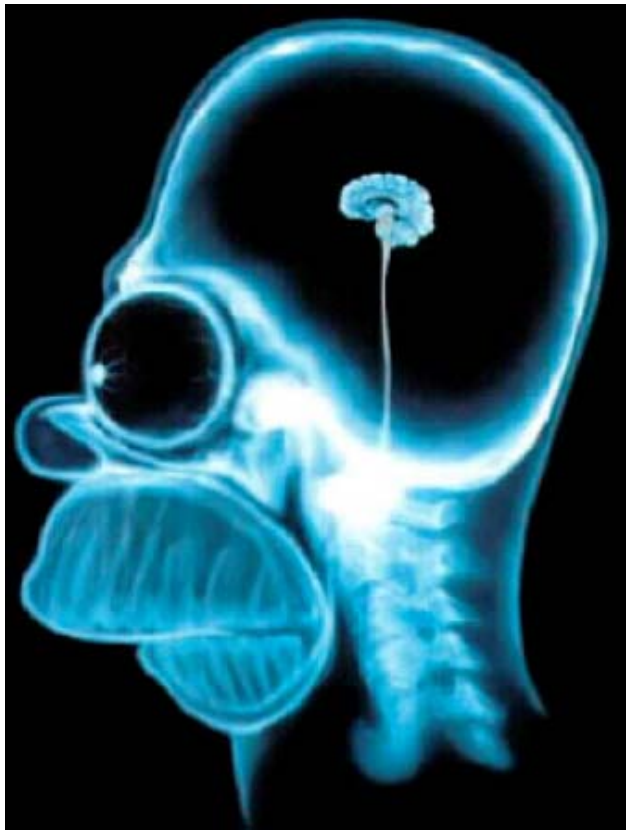
For 10^{14} synapses (a low-end estimate) & 40 Hertz synchronous update rate:

- 10^6 modules:
  - 10^6 FPGAs: $25 M
  - $6*10^{14}$ bytes DRAM: $60 M

Use of Neural Models for Cognitive Processing (cont'd)

Sample Compute Module: www.simtec.co.uk/products/EB110MOD/

Real-Time Digital Simulation of the Human Nervous System (cont.)
Reduced Configurations (~1000X reduction in size)



Simplified Human

Small Mammal

Military Applications

Real-Time Digital Simulation of the Human Nervous System (cont.)
The Biological Connectivity Problem

Two automated approaches to tracing neurons:

Knife Edge Scanning Microscopy:
Wonryull Koh & Bruce McCormick: http://research.cs.tamu.edu/bnl
Specifications for Volume Data Acquisition in Three-Dimensional Light Microscopy
500nm x 200nm x 200nm voxels, 200 Mhz pixel rate.

Two photon laser scanning microscopy:
"All-Optical Histology Using Ultrashort Laser Pulses", Neuron, vol. 39, pp. 27-41, 2003,
Tsai, Friedman, Ifarraguerri, Thompson, Lev-Ram, Schaffer, Xiong, Tsien, Squier & Kleinfeld.
http://www.tsienlav.ucsd.edu/Publications

Neither has 100nm resolution (dendrite spines are about 100nm in diameter).

Effort to determine neuron type from biochemistry:

Paul Allen's Brain Atlas: http://www.brainatlas.org/

Real-Time Digital Simulation of the Human Nervous System (cont.)
Engineering Risk Assessment

Uncertainty in synapse count:

Prefer Shepherd's numbers, $ 85 M DRAM chip cost
Koch's numbers OK, $170 M DRAM chip cost
Sejnowski's numbers considered high due to:
Perforated synapses: http://synapses.mcg.edu/anatomy/radiatum/synapses.stm
And multiple synapses per axon/dendrite pair

Biological unknowns:

Axon arbor sizes: 400 um axon & dendrite arbor diameter implies 50K axons within reach of each dendrite (assumes 100K incoming axons per mm^2 of cortex).
Adequate models of inputs and outputs (vision, hearing, musculature, etc)

Reliable operation:

600 Mhz DRAM data rate per pin
200 Mhz FPGA pipeline rate: faster chips coming (using 90nm fabs)
Reliable chips: neutron flux from cosmic rays may require shielding
http://www.actel.com/products/rescenter/ser/docs/SERWP.pdf

Routing and configuration:

Interconnect bandwidth & routing: 3D torus configuration has excess interconnect capacity
Placement: Distributing axon data across 20 parallel "processors" per FPGA

Real-Time Digital Simulation of the Human Nervous System (cont.)
Additional References

Many new books in Computational Neurobiology:

J. Feng ed, Computational Neuroscience: A Comprehensive Approach; 2002, CRC-Press.
Kotter ed, Neuroscience Databases: A Practical Guide; 2003, Kluwer Academic.
M.A. Arbib ed, The Handbook of Brain Theory and Neural Networks, 2nd ed; 2002, MIT Press.
P. Dayan & L.F. Abbott, Theoretical Neuroscience: Comptational and Mathematical Modeling of Neural Systems; 2001, MIT Press.
F. Rieke, D. Warland, R.R. vanSteveninck & W. Bialek, Spikes: Exploring the Neural Code, 1999, MIT Press.

Ontogenesis:

D.H. Sanes, T.A. Reh & W.A. Harris, Development of the Nervous System, 2000, Academic Press.

Ultrastructure of the nervous system:

Josef Spacek: Atlas of Ultrastructural Neurocytology; http://synapses.mcg.edu/index.asp

Beginner's introduction to FPGAs & CPLDs, has history & glossary:

K. Parnell & N. Mehta: Programmable Logic Device Quick Start Handbook; 2002, PDF version available from Xilinx web site.

Large Scale Computing:

T. Sterling, P. Messina & P.H. Smith, Enabling Technologies for Petaflops Computing; 1995, MIT Press (the 10^6 FPGA system of this paper does ~30 ops * 40 Hertz * 10^{14} synapses = 120 Peta-ops).