Challenges for Brain Emulation:
Why is Building a Brain so Difficult?

Rick Cattell

SynapticLink.org

rick@cattell.net

Alice Parker

Department of Electrical Engineering

University of Southern California

parker@eve.usc.edu

February 5, 2012

Abstract – In recent years, half a dozen major research groups have simulated or constructed sizeable networks of artificial neurons, with the ultimate goal to emulate the entire human brain. At this point, these projects are a long way from that goal: they typically simulate thousands of mammalian neurons, versus tens of billions in the human cortex, with less dense connectivity as well as less-complex neurons. While the outputs of the simulations demonstrate some features of biological neural networks, it is not clear how exact the artificial neurons and networks need to be to invoke system behavior identical to biological networks and it is not even clear how to prove that artificial neural network behavior is identical in any way to biological behavior. However, enough progress has been made to draw some conclusions and make comparisons between the leading projects. Some approaches are more scalable, some are more practical with current technologies, and some are more accurate in their emulation of biological neurons. In this paper, we examine the pros and cons of each approach and make some predictions about the future of artificial neural networks and the prospects for whole brain emulation.

Keywords – biomimetic, neuromorphic, electronic, artificial brain, neuron, intelligence

1. Introduction

Reverse-engineering the brain is one of the Grand Challenges posed by the United States National Academy of Engineering [1]. In this paper, we assess current status in approaching this difficult goal of brain emulation. We contrast competing approaches, and examine the major obstacles.

Artificial neurons and neural networks were proposed as far back as 1943, when Warren McColluch and Walter Pitts [2] proposed a "Threshold Logic Unit" with multiple weighted binary inputs combined to produce a binary output based on a threshold value. More sophisticated neural models were subsequently developed, including Rosenblatt's popular "perceptron" model [3] and others we examine in this article. In 1952, Hodgkin and Huxley [4] published a model of ionic currents that provided the first basis for mathematical modeling and simulation of biological neurons and their action potentials, with the help of Wilfred Rall's [5] theory of spatiotemporal integration, non-linear summation, and conductance of synaptic signals. These models have likewise been enhanced over the years by researchers examining synaptic transmission, integration, and plasticity.

Over the past 50 years, advances in technology have successively and phenomenally increased our ability to emulate neural networks with speed and accuracy.[1] At the same time, and particularly over the past 20 years, our understanding of neurons in the brain has increased substantially, with imaging and microprobes contributing significantly to our understanding of neural physiology.

These advances in both technology and neuroscience make possible the projects we discuss in this paper, aimed at modeling large numbers of interconnected neurons. Today it is feasible to emulate small but non-trivial portions of the brain, for example thousands of neurons in the visual cortex. Each approach has advantages and shortcomings when meeting the challenges posed by an artificial brain. We will examine the leading approaches and technologies, along with their pros and cons. We will conclude with a discussion of technological and architectural challenges for an artificial brain, and some debate on future research directions.

1.1 Motivation for Brain Emulation

Three motivations are frequently cited for brain emulation:

1. Researchers hope to gain a better understanding of how the brain works (and malfunctions) by creating simulations. A model can provide insight at all levels, from the biochemistry and neurochemical behavior of individual cells to the behavior of networks of neurons in the cortex and other parts of the brain.

2. Some researchers feel progress in artificial intelligence over the past 50 years has been insufficient to lead to intelligent behavior. Ideas from simulations of neural networks may yield new ideas to develop intelligent behavior in computers, for example through massive parallelism. Neural networks are already being used for applications such as computer vision and speech understanding, and many algorithmic approaches are bio-inspired, but their biological basis is, for the most part, simplified from the more-detailed models used by neuroscientists. Autonomous vehicles and other robotic applications are likely targets for such brain-like systems.

3. For the most part, computers still use the same basic architecture envisioned by John von Neumann in 1945. Hardware architectures based on the massive parallelism and adaptability of the brain may yield new computer architectures and micro-architectures that can be applied to problems currently intractable with conventional computing and networking architectures.

The projects described in this paper generally cite all three of these reasons for their work. However, there are differences in emphasis. Projects focused on understanding the brain require a more-detailed and more computationally-expensive model of neuron behavior, while projects aimed at the second or third goal may use simpler models of neurons and their connections that may not behave exactly as biological neural networks

---

[1] Some authors refer to "simulating" neurons in software and "emulating" neurons in hardware, but for simplicity in this paper we use the term "emulation" to refer to hardware, software, and hybrid implementations.

behave. An additional advantage of attempts at whole brain emulation is to further understanding of prosthetic device construction. While the research in that general area has focused on the difficult task of providing connectivity between electronics and biological neurons (e.g. Berger [6]), more complex emulated neural networks might one day provide prosthetic devices that adapt to an individual's brain, providing functions missing due to surgery, accidents or congenital defects.

1.2 Challenges to Brain Emulation

In spite of the progress in many brain emulation efforts, there are major challenges that must still be addressed:

- Neural complexity: In cortical neurons, synapses themselves vary widely, with ligand-gated and voltage-gated channels, receptive to a variety of transmitters [7]. Action potentials arriving at the synapses create post-synaptic potentials on the dendritic arbor that combine in a number of ways. Complex dendritic computations affect the probability and frequency of neural firing. These computations include linear, sublinear, and superlinear additions along with generation of dendritic spikes, and inhibitory computations that shunt internal cell voltage to resting potentials or decrease the potential, essentially subtracting voltage. Furthermore, some neuroscientists show evidence that the location of each synapse in the dendritic arbor is an important component of the dendritic computation [8], essential to their neural behavior, and there is growing consensus among neuroscientists that aspects of dendritic computation contribute significantly to cortical functioning. Further, some propagation of potentials and other signaling is in the reverse direction, affecting first-order neural behavior (for example, see the reset mechanism affecting dendritic spiking plasticity) [9, 10]. The extent of the detailed modeling of dendritic computations and spiking necessary for brain emulation is an open question.

- Scale: A massive system is required to emulate the brain: none of the projects we discuss have come close to this scale at present. The largest supercomputers and computer clusters today have thousands of processors, while the human cortex has tens of billions of neurons and a quadrillion synapses. We are a long way from cortex scale, even if one computer processor could emulate thousands of neurons, and, as we will see, it is unclear whether that emulation would be sufficiently accurate.

- Interconnectivity: Emulation of the cortex in hardware represents a massive "wiring" problem. Each synapse represents a distinct input to a neuron, and each postsynaptic neuron shares synapses with an average of 10,000 (and as many as 100,000) other presynaptic neurons. Similarly, the axon emerging from each neuronal cell body fans out to an average of 10,000 destinations. Thus each neuron has, on average, 10,000 inputs and 10,000 outputs. If the connections were mostly local, the wiring would not be so complicated; however, recent research by Bassett et al [11] derives a Rent exponent for the biological brain that could be used to compute the quantity of connections emerging from a volume of brain tissue. Early indications are that this Rent exponent is sufficiently large (many

distal connections) so as to cause connectivity problems with conventional CMOS electronics.

- Plasticity: It is generally accepted that an emulated brain with static neural connections and neural behavior would not produce intelligence. Synapses must be "plastic": the strength of the excitatory or inhibitory connection must change with learning, and neurons must also be able to create new synapses and hence new connections during the learning process. Research on the mechanisms by which neurons learn, make and break connections, and possess memory is ongoing, with hypotheses and supporting data appearing frequently. These studies have led to a basic understanding of synaptic and structural plasticity. In the last decade, attention has been given to the role of glial cells in neural behavior, glial cells being much more numerous in the brain than neurons. The role of astrocytes, a type of glial cell, in learning and memory is being actively investigated [12] and neuromorphic circuits constructed [13].

- Power consumption: A final, indirect problem is the power consumed by a brain emulation with 50 billion neurons and 500 trillion connections, and the dissipation of the associated heat generated. The human brain evolved to use very little power, an estimated 25 watts. We do not have computing technology anywhere near this power efficiency, although nanotechnology and ultra-low power design offer promise.

We will examine how each major project addresses these challenges. Although the brain emulation field is in its infancy, progress has been made in a very short time.

1.3 Other Surveys

Other surveys of brain emulation are worth reference here. They provide a different perspective than ours.

Sandberg and Bostrom [14] prove an excellent survey of the overall issues in brain emulation, although they do little discussion of actual brain emulation projects. They cover different levels of emulation, different neural models, computational requirements of emulation, and brain mapping technologies,

De Garis, Shuo, Goertzel, and Ruiting [15] provide the most similar survey to ours, covering half of the projects mentioned here. This is a good reference for another perspective on these projects. It is part one of a two-part survey. The second part, written by Goertzel, Lian, Arel, de Garis, and Chen [16], surveys higher-level brain models aimed at producing intelligent behavior, inspired by human intelligence but not based on emulation of neural networks; this work is closer to classical AI.

2. Background

There are three key components to any artificial neural network:

1. Neurons: the models used to emulate the computation and firing behavior of biological neurons, and the technology used for the emulation,

2. Connectivity: the models and technologies used for the synaptic connections between neurons, and

3. Plasticity: the models and technologies to create changes in the behavior of neurons and their synapses.

In this section, we provide some background on these models and technologies. This background will provide a basis for understanding brain emulation projects in the remainder of the paper.

2.1 Modeling Neurons

A variety of neural models are used in the projects we describe.

Most neural modeling involves the ion channels responsible for spike generation at the axon hillock, or the synapse, where spikes are transformed into post-synaptic potentials. The Hodgkin-Huxley [4] biological neural model discussed earlier, with Ca++, Na+, and K+ currents through ion channels, can require relatively expensive computations. Simulation is further complicated when one takes into account the 3-dimensional layout of axons and dendrites, requiring spatiotemporal integration. Cable theory and compartmental models have been used to account for the latter. Various improvements have been proposed to simplify computation while maintaining some level of faithfulness to biological neurons. A survey of this work is beyond the scope of this paper; the interested reader is referred to [17].

Some of the projects we discuss use a very simple model of neuronal behavior. The simplest model is an integrate-and-fire "point neuron," summing weighted input from synapses and comparing the resulting sum to a threshold, arriving at a binary decision whether and when to generate an output spike. This model is commonly extended to include a decaying charge, as a "leaky integrate and fire" neuron. The model can also be enhanced in other ways: non-linear summation, time-dependent thresholds, programmable delays in the delivery of spikes, and other variations. The point neuron models require only modest computation and hardware, in contrast to biological ion-channel models with spatiotemporal integration.

Izhikevich [18] provides a good recent survey of hybrid spiking neural models, comparing their computational costs and their ability to handle a range of 20 different spiking behaviors observed in neurons in vivo. Each model is represented by a set of ordinary differential equations that define the change in neuron membrane voltage over time, and the computational cost is measured by the number of floating point operations required in each time-step in the simulation. Izhikevich assumes 1 millisecond time steps in his paper. The Hodgkin-Huxley model is the most expensive one he considers, requiring about 1200 floating-point operations per millisecond.

Izhikevich concludes by advocating an enhanced leaky-integrate-and-fire model for neurons that is a compromise between computational cost and computational power, able to exhibit all 20 of the spiking waveforms he surveys. The differential equations for his model are

$$v' = .04\ v^2 + 5v + 140 - u + I \ (1)$$

$$u' = a\ (bv - u) \ (2)$$

if v > 30 then reset v ← c and u ← u+d (3)

Where *a*, *b*, *c*, *d,* and *I* are parameters that define the neuron's behavior, *v* is a variable representing the membrane potential in millivolts, and *u* is a variable representing membrane recovery. The parameter *I* represents the synaptic current resulting from the combination of post-synaptic potentials. Each millisecond of simulation requires only 13 floating-point operations in this model, about 100 times fewer floating point operations than Hodgkin-Huxley, yet the model still retains the capability to exhibit all of the same spiking behaviors as Hodgkin-Huxley, given appropriate values of the parameters.

More sophisticated neuron models, in contrast to the "point" models surveyed by Izhikevich, emulate components of the neuron separately. For example, synapses may be modeled separately from signal integration in the remainder of the neuron, followed by a spike-generator modeling the axon hillock, or a neuron may be modeled as dozens of small compartments, applying ion-migration equations to each compartment separately.

The fundamental feature of synapses is the voltage response over time of the neuron cell membrane to rapid input spikes that cause post-synaptic potentials to sum temporally and spatially, and that decay over time with time courses that vary depending on each individual synapse. The nonlinear sum of the excitatory post-synaptic potentials (EPSPs) might be offset by hyperpolarizing inhibitory post-synaptic potentials (IPSPs) that essentially subtract potential, or might be entirely negated by shunting inhibitory synapses that return the cell membrane to resting potential, with location of each synapse playing a role in the computation. The Blue Brain project we discuss models these dendritic computations in more detail than the other major projects.

As we shall see, the actual implementation of neuron models can be in software or in hardware, or a combination of the two. The purely-hardware implementations we discuss use neuromorphic analog circuits, as do the hardware portions of the hybrid implementations. We will discuss the pros and cons of these technology choices.

2.2 Modeling Connections

Modeling connections between neurons may seem trivial, given a hardware or software model of the neurons. However, one of the biggest challenges to brain emulation is the immense problem "wiring" the connections: the synapses, dendrites, and axons.

The connection-wiring problem differs depending how neurons are modeled and implemented. As we will see in the next section, three different approaches have been used to implement neurons:

1. *Supercomputers*, used to model neurons and their connections in software,

2. *Neuromorphic analog integrated circuits*, with an array of special-purpose neural-modeling circuits on each chip, and

3. *Special-purpose digital integrated circuits*, emulating neurons in software using many small CPUs networked together.

Corresponding to these neuron emulation technologies, there are several different approaches to implementing synaptic connectivity between neurons. In the supercomputer case, synaptic activity can be communicated through simple procedure

calls or inter-process calls. In the case of neuromorphic analog circuits, direct wiring between artificial neurons has been used locally. However, since neurons contain many distinct synapses with differing effects on neural behavior, there is high connectivity fan-in for off-chip signals. As a result of the high connectivity fan-in and fan-out, with current technologies, direct wiring has only been practical for connections between "nearby" analog neurons. For longer connections in the analog case, and for all connections in the digital case, a networking approach has been required.

The approaches used for this networking in the major projects examined here are almost all based on Mahowald's pioneering *Address Event Representation* (AER) architecture [19]. Networking and AER are based on a simplifying assumption that continuous connectivity between neurons is not necessary for an accurate emulation. Instead, they assume communication is necessary only when a neuron fires, generating an action potential. The emulated neurons are networked together, generally with a topology of many nested networks, as on the Internet, to allow scaling. When a neuron fires, network packets are sent out to all of the neurons that synapse upon it, notifying them of the spike.

As on the Internet, each network node (a neuron in this case) is assigned a network-wide unique address, and some form of routing tables are required for the system to know what nodes and subnetworks a packet must go through to reach its destination. However, in typical network communication on the Internet, each network packet contains a source address, a destination address, and the data to be communicated. In contrast, the AER approach includes only the source address (the "address event" of the neuron that spiked) in the packet. A destination address is not used because it is not practical: every neuron would need to generate many thousands of packets each time it spiked.

Instead, in the AER approach, all the synaptic connectivity information is stored in tables used by network routers. Other information may be stored there as well, for example, the strength of the synaptic connection, and the desired delivery delay for the spike.

There may or may not be data associated with each packet, as we will see. No data is necessary with a model that simply conveys a spike. However, a more sophisticated model may deliver a spike rate or a waveform for spikes, through A/D conversion of the output of neuromorphic analog circuits, or could even send continuous waveforms, delivering packets whenever significant changes in voltage occurred.

We will discuss the trade-offs in these connectivity approaches, as well as trade-offs in the neuron modeling, after describing the projects in more detail. There are important differences in scalability, emulation speed, power consumption, and biological accuracy between the connectivity approaches.

2.3 Modeling Plasticity

A static model of the neurons fulfills only some of the requirements for an artificial brain. The other key component is a model of plasticity: how neurons "learn" over time through changes in synaptic sensitivity and through generation of new synaptic connections. Synaptic strength varies in several ways. Presynaptic strength (neurotransmitter availability) is up- or down-regulated (the synapses are facilitated or depressed) through a retrograde process that is not completely understood. Postsynaptic strength is up- or down-regulated through potentiation or depression, by the availability of receptors on the

post-synaptic side of the synapse that receive neurotransmitters released on the presynaptic side of the synapse. Post-synaptic strength is modulated by several mechanisms including spike-timing-dependent plasticity (STDP), that increases receptor concentration (synaptic strength) when a positive post-synaptic potential is followed by a spike generated in the axon hillock, and decreases synaptic strength when the increase in post-synaptic potential is either late with respect to the spiking activity or does not occur at all. Parallel synapses can form at locations where existing synapses are highly active, and synapses can dissolve when activity is absent for some time. Silent synapses that do not respond to presynaptic activity can be awakened via messenger proteins expressed by the neuron's RNA, and new synapses can form over time, possibly due to other protein expression. While the post-synaptic synapse formation is believed usually to occur initially with spine growth as a precursor, followed by the presynaptic growth, there is some evidence that pre-synaptic formation can occur at the same time, or earlier.

The projects we describe assume a limited form of learning, long-term potentiation, and STDP in the brain. They generally implement at the least some form of basic Hebbian learning [20], i.e., when an axon synapsing on a post-synaptic neuron repeatedly takes part in firing the neuron, the synapses on that axon are strengthened. More-complex and more-specific models of plasticity (e.g. STDP) are implemented in some cases. Various more-sophisticated forms of synaptic plasticity have been proposed and studied in neuropsychology. For example, Gordon Allport [21] posits that repeated patterns of activity become an auto-associated engram, exciting neurons that are part of the pattern, and inhibiting those that are not. And finally, in addition to strengthening and weakening of synapses, there is evidence in biological neurons, even in mature brains, for the growth of entirely new dendritic spines, dendrites and synapses (e.g., [22, 23]).

Relatively little is written about the plasticity and learning processes used in the projects we cover. However, the learning mechanism is generally encoded in software that can easily be changed, so the projects do offer an opportunity to experiment with various models of learning.

3. Project Summaries

Many projects around the world have aimed at emulating neural networks.[2] In this paper we have attempted to limit our scope to the most advanced and pragmatic approaches to large-scale neural emulation. In particular, we only consider projects intended to scale to millions of neurons, and projects that have fabricated and tested their designs, at least on a small scale, with currently available technologies. Given this scope, although there are innovative, successful projects with more limited scope, due to space and time limitations, we elected to focus on six projects in this paper that have the most ambitious scope and the most demonstrable results:

1. The SpiNNaker [24] project at Manchester University in the U.K.,

2. The Blue Brain [25] project at École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland,

---

[2] This field is rapidly evolving, so our descriptions reflect a single point in time for each project represented. The reader is cautioned to consult the latest publications for the most accurate information.

3. The C2S2 SyNAPSE [26, 27] project at IBM Research in California,

4. The FACETS [28] project at Heidelberg University in Germany,

5. The Neurogrid [29] project at Stanford University in California, and

6. The IFAT [30, 31] and NeuroDyn [32] projects at the University of California at San Diego.

In the following subsections we look at each of these projects in more detail. In the last subsection, we discuss a few related projects, with a focus on emerging technologies.

3.1 SpiNNaker

The SpiNNaker project at Manchester University is based on fabricating many small CPUs on a chip, the cores communicating through a network on-chip and through a network between chips. The principal investigator, Steve Furber, was a co-designer of the ARM 32-bit RISC microprocessor, and a simplified ARM 968 processor is used for the CPUs on the SpiNNaker chips. Each CPU is designed to simulate about 1,000 neurons, communicating spike events to other CPUs through packets on the network.

The SpiNNaker chip is designed to include

- 18 low-power ARM CPUs, each with about 100KB of local RAM used to store its programming and data,

- 128MB of RAM shared by all 18 CPUs through a DMA controller, used to store synaptic weights and other information, and

- An on-chip network and packet router that connects the 18 CPUs and also connects to 6 adjacent SpiNNaker chips, to reach other CPUs.

The routing of packets in SpiNNaker is carefully designed to balance complexity and bandwidth. AER packets are used, as with most of the other projects described here. Routing tables stored in a content-addressable memory tell the router which packets must be routed to which CPUs, whether off-chip or on-chip. The SpiNNaker chips are connected to adjacent SpiNNaker chips in a 2-dimensional toroid mesh network; each chip has 6 network ports, connected to adjacent chips. The router need not know the eventual destination(s) of a packet, it only needs to know which port(s) to send it to. Routing tables are built and maintained by a separate (programmable) background process responsible for connectivity, plasticity, and learning [33].

SpiNNaker is initially using a simple algorithm for neurons based on Eugene Izhikevich's point neuron model [34]. For the purposes of this paper, we analyze SpiNNaker based on that model, although their software-based architecture could support a variety of more sophisticated neural models.

Their point neuron algorithm is programmed into the local memory of each of the SpiNNaker CPUs. Post-synaptic weights for synapses are stored in the SpiNNaker chip's shared memory; the algorithm fetches the corresponding weight into local CPU memory whenever a spike arrives at one of its "synapses," and recomputes neuron action potentials at 1ms simulation intervals, based on Izhikevich's equations. 16-bit fixed-point

arithmetic is used for most of the computation, to avoid the need for a floating-point unit and to reduce computation and space costs.

Because spike delivery time in SpiNNaker is designed to be faster than a biological brain (assuming the network and routing delays are adequately controlled), SpiNNaker allows a delay of up to 15ms to be inserted in delivery of AER packets, in order to simulate longer axons. The goal is to allow the globally asynchronous, locally synchronous design to operate similarly to a biological brain possessing the same neural network.

The maximum number of SpiNNaker chips supported by the packet-address structure is 2^{16} (65,000 chips). About a billion neurons could be simulated in this configuration, if the physical chip placement, network, and other constraints do not limit scalability. The group has done some limited simulations to determine when the network and CPUs become saturated [35]. We will further discuss scalability in the last section.

Work continues on the SpiNNaker project. It is expected that a full 65,000-chip configuration with 18-CPU chips will be built some time in 2012.

3.2 Blue Brain

The Blue Brain Project at EPFL in Switzerland uses an IBM Blue Gene supercomputer with 8,000 CPUs to simulate neurons and STDP in software. Henry Markram at EPFL's Brain Mind Institute is the principal investigator.

The Blue Brain group constructed a 10,000 neuron model of a neocortical column from the somatosensory cortex of a 2-week-old rat, and simulated it on the Blue Gene supercomputer. The simulation ran about ten times slower than biological neurons.

The modeled cortical column is about .5mm in diameter and about 2.5mm in height. The model is not a map of real connections in any particular rat; the connections are randomly derived based on the percentage connectivity of neurons of different types in different layers of rat cortical columns. However, the model does attempt to account for the 3D morphology of the neurons and cortical column, using about 1 billion triangular compartments for the mesh of 10,000 neurons. A multi-processor adaptation of the NEURON simulation software [36] was run at this fine grain using Hodgkin-Huxley equations, resulting in gigabytes of data for each compartment, and presumably a high level of bio-realism. Timing, e.g. propagation delays along the simulated compartments of an axon, are incorporated into the simulation. Synaptic learning algorithms are also introduced, to provide plasticity. A visual representation of parts of the cortical column can be displayed for the simulation, allowing researchers to focus on particular parts or phases of the simulation in more detail.

The Blue Brain project is unusual in its goal to simulate the ion channels and processes of neurons at this fine-grain compartmental level. Had the project simply used a "point neuron" model integrating incoming spikes, the simulation could have delivered orders of magnitude higher performance, but Markram opted for a higher level of bio-realism.

Of course, software emulation of neurons on large computers, including the bio-realistic fine-grain compartmentalized emulation used in Blue Brain, has been used widely in computational neuroscience laboratories; we mention some other projects at the end of

10

this section. However, we chose to include the Blue Brain project in this paper as the best example of this approach, because of its combination of large scale and bio-realism.

Work on the Blue Brain project is now progressing to a second phase of work. The team cites two new directions: incorporating molecular level processes, and simulating more of the brain through additional parallelism. No publications are yet available on this work, to our knowledge.

3.3 C2S2

Dharmendra Modha's Cognitive Computing Group at IBM Alamaden Research Lab received funding in 2008 from DARPA's SyNAPSE initiative with their proposal "Cognitive Computing via Synaptronics and Supercomputing (C2S2)." Modha has in turn funded professors from 5 universities (Cornell, Columbia, Stanford, Wisconsin Madison, and UC Merced) as part of their project, bringing in expertise in neuroscience, psychology, VLSI, and nanotechnology. We will refer to Modha's project as "C2S2".

Modha's team studied data on biological brains to work toward a "connectome" database of neural connectivity [37], using experimental data from diffusion tensor imaging (DTI) and other techniques. They created a massively parallel cortical simulator called C2, which was initially used at the scale of a rat cortex, and more recently at the scale of a cat cortex, running on IBM's Dawn Blue Gene/P supercomputer, with 147,456 CPUs and 144TB of main memory. In the latter case C2 simulated 1.6B cortical neurons and 9 trillion synapses, using experimentally measured thalamo-cortical connectivity. The simulations incorporated STDP and controlled axon delays.

The C2 simulation used a much simpler model of neurons than the Blue Brain, with single-compartment spiking Iszhikevich-type neurons. As with the Blue Brain, the connectome used did not match the actual connectome of any particular biological brain: it is an approximation based on the tools currently available. However, Modha points out that much can be learned even with these approximations. He reported oscillations in neural firing patterns seen over large areas of the simulated cortex at the alpha and gamma frequencies seen in mammal brains, and groups of neurons exhibited population-specific response latencies matching those in the human cortex.

More recently, Modha has published papers on new "cognitive computing chips" [28], suggesting that IBM research will now turn to hardware for brain emulation. The prototype chip emulates 256 neurons, using a crossbar connecting 1024 input axons to the 256 neurons with weighted synapses at the junctions. Variations of the chip have been built with 1-bit and 4-bit synapse weights stored in SRAM. Another was built with low leakage to reduce power consumption.

Cross-chip spikes are conveyed asynchronously via AER networking, while the chips themselves operate synchronously. Synapses are simulated using the Izhikevich leaky integrate-and-fire model. The results are identical to the same equations simulated in software, but all 256 neurons on the chip update their membrane voltage in parallel, at 1ms intervals. The details of the AER networking are not specified, so it is not possible to speculate on how that will scale at this time.

3.4 FACETS and BrainscaleS

The FACETS project (Fast Analog Computing with Emergent Transient States) is a consortium of 15 groups in 7 European countries, led by professors Johannes Schemmel and Karlheinz Meier of the Electronic Visions lab at the University of Heidelberg.

In their early work, the "Spikey" neuromorphic ASIC chip was developed. A Spikey chip hosts a total of 128K synapses; it could simulate, for example, 8 neurons with 16K inputs, or 512 neurons with 256 inputs. The goal was to simulate analog neuron waveforms analogous to biological neurons on the same input.

The Spikey neurons communicate with each other digitally, although the neuron circuit is analog. Digital action potentials are routed to synapse drivers, that convert them to voltage pulses that, in turn, control synaptic conductance. The synapse drivers also implement STDP; synaptic weight storage is implemented as static RAM. Synaptic conductance is modulated by an exponential onset and decay.

Whenever an analog neuron circuit reaches an action potential, digital monitoring logic generates a spike event with the event time and the address of the spiking neuron. This event is transmitted on a network to multiple destination neurons that need not be on the same Spikey chip. About 1/3 of the Spikey chip is digital control logic that implements the digital communication between neurons. 16 Spikey chips can be operated on a custom backplane that implements high-speed digital communication between the chips with a fixed and guaranteed latency.

The Spikey chip outputs, inputs, circuit parameters, and neuron interconnections can be monitored and controlled from software running on a host computer. Selected neurons can then be stimulated with experimental spikes, and neuron outputs can be recorded.

More recently, the FACETS researchers developed the HICANN (High Input Count Analog Neural Network) chip and "wafer scale integration" to achieve higher connectivity between simulated neurons. HICANN bears some resemblance to Spikey in that neural emulation is analog, with digital circuits for communication and STDP. However, there are a number of differences. Instead of placing each HICANN chip in a separate package as with Spikey, the entire multi-chip wafer is enclosed in a single sealed package with horizontal and vertical "Layer 1" channels that connect the HICANN chips within and between reticles on a wafer. A total of 352 HICANN chips can be interconnected on the multi-chip wafer, producing 180,000 neurons with a total of 40 million synapses.

Synapses are implemented with groups of DenMem (Dendrite Membrane) circuits. A hybrid analog/digital solution is used for the synapses, and a hybrid of address-encoding and separate signal lines is used for communication. Each DenMem can receive as many as 224 pre-synaptic inputs based on a 6-bit address sent via a Layer 1 channel. The synaptic weight is represented in a 4-bit SRAM with a 4-bit DAC. The post-synaptic signal is encoded as a current pulse proportional to the synapse weight, and can be excitatory or inhibitory. Neuron circuits integrate the DenMem signals. A digital control circuit implements STDP based on temporal correlation between pre- and post- synaptic signals, updating the synaptic weight.

A packet-based "Layer 2" routing protocol is used to communicate between wafers, using pads on the HICANN chips that connect them to the PCB. Layer 2 channels provide 176GB/sec from the wafer to PCB, allowing 44 billion events/second to be communicated between wafers. The Layer 2 wafer-to-wafer channels are handled by FPGAs and OTS switches on the PCB with 1-10 Gbit Ethernet links.

The HICANN chips implement an adaptive exponential integrate and fire (AdExp) model of neurons. This model is somewhat more sophisticated than the standard integrate and fire model used in SpiNNaker, but less sophisticated (and less computational expensive) than Blue Brain's multi-compartmental Hodgkins-Huxley-based model. The FACETS group is now investigating more sophisticated models.

The FACETS neural networks are described in PyNN, a simulator-independent language maintained by neuralensemble.org. PyNN is Python-based and includes operations to create populations of neurons, set their parameter values, inject current, and record spike times. PyNN can be run on a simulator such as NEURON, or can be used on the FACETS host computer to initialize and control the chips. In addition, a neuralensemble.org framework called NeuroTools has been developed to assist in the execution of experiments, and the storage and analysis of results. In recent work [38], software has been developed to automatically translate a PyNN design into a hardware implementation in several stages, optimizing the physical placement of the neural components and connections on HICANN chips.

A follow-on to the FACETS project, BrainscaleS [39], was started in 2011. To date, only high-level directions have been published on BrainscaleS. Two key goals of BrainscaleS are in-vivo recording of biological neural networks and the construction of synthesized cortical networks with similar behavior. The focus is on perceptual systems. The BrainScaleS project is establishing close links with the Blue Brain project and with Brain-i-Nets [40], a consortium producing a set of learning rules based on synaptic plasticity and network reorganization.

3.5 Neurogrid

The Neurogrid project at Kwabena Boahen's "Brains in Silicon" lab at Stanford University uses programmable analog "neurocore" chips. Each 12x14 mm^2 CMOS chip can emulate over 65,000 neurons, and 16 chips are assembled on a circuit board to emulate over a million neurons. The system is built and functional.

Neurogrid uses a two-level simulation model for neurons, in contrast to the point neuron model used in SpiNNaker, and in contrast to the thousands of compartments used in Blue Brain's simulation. Neurogrid uses this approach as a compromise to provide reasonable accuracy without excessive complexity. A quadratic integrate-and-fire model is used for the somatic compartment. Dendritic compartments are modeled with up to four Hodgkin-Huxley channels. Back-propagation of spikes from somatic to dendritic compartments are supported.

Neurogrid uses local analog wiring to minimize the need for digitization for on-chip communication. Spikes rather than voltage levels are propagated to destination synapses. To simplify circuitry, a single synapse circuit models a neuron's entire synapse population of a particular type, and each of these circuits must be one of four different

types. The synapse circuit computes the net postsynaptic conductance for that entire population from the input spikes received. Although this approach limits the ability to model varying synaptic strength, and it does not model synaptic plasticity, it greatly reduces circuit complexity and size.

Like SpiNNaker, Neurogrid uses an AER packet network to communicate between-chip spikes. Unlike SpiNNaker's grid organization, Neurogrid's chips are interconnected in a binary tree with links supporting about 80M spikes/second (this is a change from earlier work [41] in which Boahen used a grid network). Routing information is stored in RAM in each router. This AER-based networking is referred to as "softwire" connections.

To reduce communication overhead, a single inter-chip spike can target multiple neurons on the destination chip. The postsynaptic input triggered in a target neuron can be propagated to neighboring neurons with a programmable space-constant decay. This requires only nearest-neighbor connections: the synaptic potentials superimpose on a single resistive network to produce the net input delivered to each neuron. A single cross-chip spike can thus reach a hundred neurons. This is analogous to cortical axons that travel for some distance and then connect to a number of neurons in local patch arbors in another cortical column.

Unlike FACETS, which is designed to run orders of magnitude faster than biological neurons, the Neurogrid neuron array is designed to run in real-time. This means that a single AER link can easily service all of the cross-chip spikes for 65,000 neurons. Furthermore, the on-chip analog connections can easily service their bandwidth, and it seems likely that the binary routing tree connecting the 16 Neurogrid chips on a circuit board can easily support a million neurons. Thus, the only potential bottleneck for Neurogrid might be in routing between multiple boards in the future.

Like FACETS, the neurocore chips are programmable. Each neurocore models the ion-channel behavior and synaptic connectivity of a particular neuron cell type or cortical layer. The Neurogrid neuron circuit consists of about 300 transistors modeling the components of the cell, with a total of 61 graded and 18 binary programmable parameters. Synapses can be excitatory, inhibitory, or shunting. The Neurogrid group has demonstrated that their neurons can emulate a wide range of behaviors.

The Neurogrid team has encouraged others to build on their work, teaching courses training students to build neural networks on their framework, and making their silicon compiler available to allow others to design neuromorphic systems for fabrication. The descriptions are written in Python.

3.6 IFAT and NeuroDyn

Like the Neurogrid and FACETS projects, Gert Cauwenberghs and colleagues at the Institute for Neural Computation (INC) at the University of California at San Diego chose to use analog neuromorphic circuit chips to model neurons. They have produced two different chips, IFAT and NeuroDyn, with different goals.

The initial IFAT (Integrate and Fire Array Transceiver) chip, prototyped in 2004, could emulate 2400 simple neurons. A separate microcontroller on the same circuit board used analog-digital converters and an AER lookup table to deliver spikes to the IFAT chips

based on a global "clock cycle." The INC group applied the IFAT chips to various applications, including Laplacian filters to isolate vertical edges on images, and spatiotemporal filters to process a spike train from an artificial retina, constructing velocity-selective cells similar to those found in the medial-temporal cortex in the human brain, demonstrating brain processing.

The latest version of the IFAT chip emulates 65,000 neurons. The new system, called HiAER-IFAT (Hierarchical AER IFAT), uses a tree of routers for delivery of AER events [42]. The tree is built using Xilinx Spartan-6 FPGAs connecting to the IFAT chips at the leaves. HiAER-IFAT has been demonstrated with 250,000 neurons. Like SpiNNaker, all of the connectivity information is held in RAM in the routing tables of the intermediate nodes, in this case the non-leaf nodes of a hierarchy. Unlike SpiNNaker, the maximum number of routing "hops" is logarithmic in the number of neurons. However, it is possible that the HiAER-IFAT routers in the highest level of the hierarchy could become overloaded if there is insufficient locality of reference.

The INC group has also designed a "NeuroDyn" chip, which is the most sophisticated of all of the neuromorphic chips discussed in this paper, in terms of bio-realism and neuron emulation. Their neuron emulation supports 384 parameters in 24 channel variables for a complex Hodgkin-Huxley model. This level of emulation is important, for example, in examining the effects of neuromodulators, neurotoxins, and neurodegenerative diseases on ion channel kinetics. However, NeuroDyn is not designed for large-scale brain emulation: each chip emulates only 4 neurons and 12 synapses.

In contrast to IFAT and all the other projects that generate discrete spike events to be delivered by AER or other means, NeuroDyn emulates neural and synaptic dynamics on a continuous basis. Matlab software on a workstation can monitor and control each neuron's membrane potential and channel variables, and can adjust the 384 NeuroDyn emulation parameters to tune to any desired neuron behavior. The parameters are stored on chip in digital registers. Experiments analogous to patch-clamping biological neurons can be performed on NeuroDyn neurons through the software.

3.7 Other projects

Some other projects are worth mention because they address the challenges of an artificial brain in novel ways, although they have not yet progressed enough to include in our comparison at this time. Additional projects are also surveyed in papers by de Garis et al [43], although a number of those projects are aimed at higher-level models of the brain, not the direct emulations surveyed here.

The BioRC [44] project at the University of Southern California, led by the second co-author of this paper, is worth mention because of its radically different technology approach: artificial neurons and connections are proposed to be built from carbon nanotubes and other nanodevices like nanowires or graphene transistors. The long-term goal of the BioRC research project is the development of a technology and demonstration of electronic circuits that can lead to a synthetic cortex or to prosthetic devices. However, this project is still at an experimental stage, designing individual neurons and small neural networks, so we did not include it in our comparison.

The BioRC project aims to meet all of the challenges discussed earlier in this paper, and the choice of emerging nanotechnologies is posited to be required in order to achieve all the challenges posed. While experiments to date have involved carbon nanotube FET transistors, other nanotechnologies are under investigation. Carbon nanotubes have some distinct advantages, not provoking an immune system reaction or corroding in contact with living tissue, as well as the obvious advantages of being extremely small (a few nm in diameter) and low power. Finally nanotechnologies like carbon nanotubes offer a possible future advantage if they can be arranged and rearranged into 3-D structures of transistors and circuits to support the connectivity and structural plasticity challenges faced when building an artificial brain.

The BioRC neural circuits can be arranged to implement neurons with many variations in structure and behavior. The neural circuits are also designed with inputs that act as "control knobs" to vary neural behavior. The control knobs can be used to create neurons with differing characteristics (e.g. spiking vs. bursting), or can be used as control inputs representing external influence on neural behavior (e.g. neurohormones). A variety of synapse circuits, axon hillocks, and dendritic arbors have been designed to illustrate temporal and spatial summation, STDP, dendritic computations, dendritic spiking, dendritic plasticity, and spiking timing variations. A CMOS chip containing many of the circuits has been fabricated. Finally, a single synapse with a carbon nanotube transistor has been constructed and tested in collaboration with Chongwu Zhou [45].

BioRC's neural models, especially the synaptic models, are more complex than most of the major projects, with the exception of Markram's Blue Brain. Interconnections in existing CMOS technology are believed to be the primary challenge to whole brain emulation for this project, although newer flip-chip technologies can ameliorate connectivity problems significantly. Self assembly with nanodevice transistors, like that performed by Patwardhan *et al.* [46] shows promise for future whole brain emulation with analog circuits.

Memristors are another nanotechnology being implemented in neural circuits, with the pioneering work at HP, where the first fabricated memristors were invented [47].

In addition, various other research groups have made progress towards more advanced neural simulations:

- Eugene Izhikevich, CEO of the Brain Corporation, together with Nobel prize winner Gerald Edelman, simulated a million spiking neurons and 500 million synapses tuned to approximate recorded in-vitro rat cortex neural responses [49]. Their neural model was slightly more sophisticated than the one used in Modha's simulations, separating the soma from multiple dendritic compartments. Like Modha, they found that waves and rhythms emerged. They also found their simulation highly sensitive to the addition or removal of a single neural spike.

- Giacomo Indiveri's Neuromorphic Cognitive Systems Lab at the Institute of Neuroinformatics at the University of Zurich have built biomimetic hybrid analog / digital CMOS VLSI chips for specific functions such as real-time sound recognition and optic flow sensors, using quite detailed neuronal models [50].

- The Computational Neurobiology Lab at Salk Institute as well as the INC lab at UCSD perform detailed software neuron simulation and advanced recording of biological brains, for example to model learning [51]. The MCell project simulates detailed diffusion machinery and other biomolecular processes at synapses.

- Farquhar and Hasler at Georgia Tech describe a programmable neural array [52],with analog neural circuits.

4. Analysis and Comparisons

Each of the projects we discuss address some challenges to artificial brain construction directly. However, none of the projects masters all of them. In this section of our paper, we examine four challenges:

1. Bio-realism of the neural computation model, i.e., the project's ability to emulate the behavior of biological neurons,

2. Bio-realism in neural connectivity, including fan-in and fan-out,

3. Bio-realism in synaptic and structural plasticity, i.e., whether an artificial brain will learn and adapt like a biological brain, and

4. Scalability of all the above, including power and space requirements, for tens of billions of neurons and hundreds of trillions of connections.

4.1 Neuron Emulation

The projects we focused on in this paper use three different technological approaches to the emulation of neurons. SpiNNaker uses a "neuroprocessor" approach, emulating neurons in software on loosely-coupled (networked) CPUs. Blue Brain and the original C2S2 work use a "neurosimulation" approach, emulating neurons in software on tightly-coupled (shared memory) CPUs in a supercomputer. FACETs, Neurogrid, and NeuroDyn use a "neuromorphic" approach, with analog circuitry for neural computations.

Independent of the technological approach, the projects differ substantially in the level of bio-realism and computational sophistication in their emulation of neurons and synapses:

1. The simplest approach is the *point neuron* model, as recommended by Izhikevich, in which a neuron's synaptic inputs enter into differential equations to compute the output of the neuron over discrete time intervals. SpiNNaker and the C2S2 work have used such a model.

2. A point neuron model implemented in analog circuitry is potentially more sophisticated, depending on the complexity of the circuit, since the circuit can perform continuous real-time integration of signals in contrast to the discrete-=time intervals used in software emulations. The NeuroDyn chip implements a particularly sophisticated point neuron, with hundreds of parameters.

3. A *two-level* analog model such as Neurogrid's two compartments, or the FACETs HICANN chip's separate dendritic membrane circuits, allows more sophisticated neural emulations, depending on the complexity of the compartment emulations.

17

4. The most bio-realistic approach among the projects is Blue Brain's *fully compartmentalized* model of the neuron, representing a biological neuron as hundreds of independent compartments, each producing an output based on adjacent ion channels and regions. These result in an integrated neural output at the axon hillock compartment, but also allow for local dendritic spikes and back-propagation of action potentials to dendrites. Blue Brain uses computationally expensive Hodgkin-Huxley equations to compute the potential bio-realistically in each compartment.

The neuromorphic approach avoids the substantial computational overhead of software simulation, and may produce a more biologically-accurate result in less time than point neuron software simulations using Izhikevich's equations. On the other hand, while neuromorphic analog circuits can produce results many orders of magnitude faster than real neurons or a software simulation like Blue Brain, there is still a remaining question about whether their fixed neuronal structure adequately captures biological neuronal behavior.

Because the connectome used in the Blue Brain simulations is not identical to any biological brain, it is difficult to observe identifiable functional behavior from the cortical column they simulate, except in very abstract ways. Since none of the systems can be directly compared to biological brains, it remains an open question what neural complexity is required to demonstrate biological behavior.

Keep in mind that biological neurons are slow in comparison to current electronics, differing by at least a factor of 10^6 in speed, if we compare the speed of a simple logic gate with the speed of a neuron. However, it takes many machine cycles and tens of thousands of gates executing software instructions in order to simulate a neuron. There is also significant overhead due to communication between processors, further slowing execution. At one point, the Blue Brain neurons were about ten times slower than biological neurons in simulations, and used about 8000 processors to simulate 10,000 neurons in a cortical column of a rat. This highlights the need for massive parallelism, and the performance degradation when simulation is performed on serial processors.

Note that the artificial brain projects can be grouped into two overall categories: simulating neurons with digital hardware (Blue Brain, C2S2, and SpiNNaker), or simulating neurons with in analog hardware (FACETs, NeuroDyn, and Neurogrid). Most projects seem to rest at the extremes of processing variations: massive multiprocessor software simulations or analog neuromorphic circuit emulations. One could speculate that special-purpose digital hardware built with FPGAs or as ASICs would explode in size and complexity due to the non-linear additions and multiplications occurring in the dendritic arbor, forcing digital implementations that implement the dendritic arbor for each neuron to be significantly simplified over software implementations. Because of the relative simplicity of analog computations compared to digital computations, most hardware approaches have exploited the ability to manipulate currents and voltages by means of analog electronics, inspired by the seminal work of Misha Mahowald [15] and Carver Mead [53]. While the analog computations are inexact and subject to fabrication and environmental variations, biological neural networks exhibit variability in behavior as well, and still perform well under differing circumstances.

The troubling thought is that there are no definitive results to indicate how detailed the model of the brain and neurons must be in order to demonstrate intelligence. Non-linear dendritic computations and dendritic spiking are shown to occur in the biological brain (e.g., by Polsky [7]), but perhaps such biological structures could be supplanted with more intricate connectivity between simpler neuronal structures in an artificial brain, much as the analogy between Turing machines and modern supercomputers, with elaborate programming schemes in a Turing machine replacing software running on a more-complicated execution engine. Thus, while some attempts are less bio-realistic in their models of neuronal computations, they might be able to demonstrate equivalent intelligence with added sophistication of connectivity over other models.

4.2 Synaptic Connectivity

As with the neuron models, there are a number of technical approaches to modeling synaptic connectivity. Blue Brain uses software calls. The others generally use a digital networking approach, but Neurogrid, FACETS, NeuroDyn, and IFAT use direct wiring for short distances.

As with neuron emulation, and independent of technical approach, the important differences in connectivity approaches are the resulting bio-realism and capabilities:

1. *Delivered content:* the synaptic connectivity model may simply deliver a *spike* as an event to another neuron, the entire *spike voltage waveform* may be delivered, or there may be *continuous connection,* at least at some small time granularity, so that sub-threshold voltages can affect a synapse. There are key questions for neuroscience to answer, here, before we can judge what must be delivered. There is some evidence that implementing the spikes as events alone is adequate, and this would vastly simplify the circuits and technology to emulate synaptic connectivity, but there is dissent concerning this assumption.

2. *Connection distance:* the projects differ in their ability to deliver output to distant *vs.* nearby neuron synapses, or in the properties of their short *vs.* long connections, e.g. with direct connections versus AER packets.

3. *Connection delays:* biological axon/synapse connections differ in their time delays, particularly for axons and dendrites that synapse over longer distances, or through the thalamus. A model that treats all connections the same cannot model this. However, it may be possible to insert delays to simulate longer connections in all of the projects, and even with direct wiring, delay circuits could be inserted. AER packet delays can be unpredictable, but a variable delay can be inserted to achieve any desired delivery time, if AER delays can be adequately bounded. Of course, tracking and correcting delivery times add complexity to those systems.

4. *Fan-in/fan-out:* There are limitations in the fan-in and fan-out allowed by all the technologies, and there will be bigger delays with larger fan-in and fan-out, with either direct connections or with AER packet delivery. We will examine connectivity scalability in Section 4.4.

5. *Timing and other issues:* A final challenge related to connectivity is the neural sensitivity to spike-arrival timing at the synapse. Late spikes result in synaptic

depression in biological synapses. Arrival of spikes in a predictable manner supports rate coding, believed to be a mechanism that conveys more information than a more-binary interpretation of spikes with spike/no-spike processing. Thus, connecting the brain physically is a major challenge, but predictable spike-arrival timing further complicates the connectivity problem enormously. In addition, communication between proximal neurons occurs via astrocytes as well and is postulated to occur via electromagnetic waves and other signals, further complicating the wiring.

An architecture with *synchronous* delivery of spikes introduces a timing issue. For example, connectivity in the original IFAT chip was based on delivering spikes on global clock cycle intervals, with neurons computing their output state on each cycle, while HiAER-IFAT provides asynchronous operation.

Continuous analog connectivity works well for short distances, as demonstrated by a number of the projects, but direct wiring to all neurons at the scale of an artificial brain requires massive connectivity not yet possible in modern digital circuits. The brain does seem to follow a Rent's rule [11] just as digital systems do, in that there is a relationship between the number of connections emerging from a volume of brain tissue compared to the size of the brain tissue enclosed. However, all modern digital systems exploit some form of multiplexing for communications at any distance. An artificial brain that did not multiplex connections, sharing wires between distant parts of the brain, would likely be unable to support 10,000 connections per neuron using current technologies.

Thus, all of the projects have a challenge with the brain's dense synapse fan-in and axon fan-out. To date, the solution of choice is AER packet networking. In the case of neuromorphic analog circuits (i.e., for all but SpiNNaker and Blue Brain), the use of AER requires some form of A/D and D/A conversion, or at least detection and delivery of a spike threshold voltage or spike event. This is problematic, because the per-neuron circuitry required for conversion and for the routing and decoding logic for the very large address space (37 bits in the worst case) is much larger than the neuron emulation circuit itself, perhaps orders of magnitude larger. Recent evidence points to significant connectivity outside each individual cortical column, implying many non-local connections [11].

In SpiNNaker's "neuroprocessor" approach, there is no A/D conversion and the cost of the network routing logic is amortized over 1,000 emulated neurons per CPU. However, both SpiNNaker and the neuromorphic analog circuits face another problem with AER networking: routing packets in real-time from tens of billions of neurons is a major challenge. We will examine networking scalability in Section 4.4.

Another issue with AER networking is the timing of spikes. Neurons adapt to early and late signals over time, and some signal timing tuning is performed by the oligodendrocytes that form a myelin sheath on the neurons' axons. In an emulated brain, with packet communication over a network, timing of emulated spikes originating from the same neuron is uncertain from second to second. However, proper synchronization can be achieved by inserting delays or reserving bandwidth [32, 54].

4.3 Plasticity and Learning

Plasticity and learning present one of the biggest challenges to artificial brain projects, partly because we don't fully understand how these work in the brain, partly because such biological mechanisms are quite complex (e.g. Kandel's seminal research on memory [55]), and partly because our technologies are far less plastic than neural tissue. We have experimental evidence for some basic synaptic plasticity mechanisms, but knowledge about plasticity, learning, and memory is far from complete. Basic Hebbian learning could likely be the "tip of the iceberg". There is already evidence that Hebbian learning applies not just at the level of individual synapses, but to groups of interacting neurons, and many forms of learning may have evolved in the brain, or in different parts of the brain [22]. There is also evidence for neurons growing new dendrites and synapses to create new connections as well as changing the "weight" of existing synapses by increasing or decreasing the number of neurotransmitter vesicles or receptors for the neurotransmitters. Finally, learning can also occur through neurons dying or being created.

The projects have generally not specified the mechanisms for learning in an artificial brain. However, they generally explain how learning would be programmed, once a learning algorithm is specified. There are three approaches:

1. In an all-software solution like Blue Brain, the learning algorithm is of course implemented in software, which makes it easy to change. Connections can added and removed, their synaptic strengths can be changed, new neurons can be created, and existing ones can be removed.

2. A separate software module can be used for learning with a solution that implements synapses through AER packets, and which uses tables to store both the connection information and the "weight" of each connection. Connections can be created and destroyed in the table as well. This is the case for SpiNNaker's tables, the tree of connections in FACETs and Neurogrid, and the separate tables used in IFAT. It should also be possible to create and remove neurons via the separate process, given an appropriate interface to the neuron implementation.

3. In a solution with direct wiring, at least some of the learning mechanism must be implemented in the artificial neuron itself. FACETS includes hardware-based STDP in the HICANN chips. However, a separate mechanism for the creation and removal of neurons would be needed in this case, and there is currently no practical solution for the growth of new synaptic connections as direct "wires," particularly over longer distances.

Although neuroscience's understanding of learning and plasticity is currently limited, and artificial brain projects have offered incomplete solutions to date, these projects will likely prove important in our understanding going forward. It is fortunate that most of the brain emulation projects have left the learning mechanism open, to be programmed in software. It is very difficult to understand synaptic changes and learning in a heavily-interconnected biological brain: experiments *in vivo* are very difficult, and experiments *in vitro* do not deal with sufficiently-complex neuron networks. In an artificial brain, in contrast, it is possible to experiment with many different plasticity mechanisms,

monitoring all the neuron connections over time, and observing the learning behavior. These experiments could prove to be the most useful contribution of artificial brain projects over the coming decade.

In addition to the learning problem, an unsolved problem is the initial configuration of an artificial brain prior to developmental change and learning. Embryologically, neurons are of many different types and have at least some genetically-determined connections or connection patterns that are almost certainly essential to intelligent behavior, and without which learning would probably not occur. It is known that connections are pruned as the brain matures and knowledge specializes the neural circuits. Even at the individual neuron level, the initial configuration of synaptic strengths, threshold voltages, and integration behavior in an artificial brain, along with models of learning mechanisms to be used, will certainly determine whether learning occurs later. In Ishikevich's favored learning model, for example, no initial values for the parameters a, b, c, and d are specified for his equations. An open question is what should these be.

4.4 Overall Scalability

Armed with some understanding of the technologies and mechanisms proposed for artificial brains, we can now examine practical issues of overall scalability.

All-software solutions such as the Blue Brain project do not directly address scalability. Instead, the problem is reduced to finding a sufficiently large supercomputer to run the software at the desired scale. Since that project is already experiencing limitations running on one of the largest supercomputers available to date, and they are emulating less than .0000001% of the neurons in the human cortex, there are obviously scaling issues here, and there are power and scaling constraints on the largest supercomputers that can be built. Another alternative would be a software approach decomposed into processes that could run on many distributed computers, e.g. using computing power on many sites or cloud computing. We are not aware of a solution using this approach to date, but the communication latency and bandwidth could prove to be a problem with this approach, even if enough processing power could be secured.

In all the other projects, a hardware solution is proposed. The scaling challenges in these projects fall into four categories:

- *Physical size and packaging:* Each project proposes integrated circuit chips that emulate some number of neurons and connections. For example, the NeuroDyn chip can perform Hodgkin-Huxley simulations of 4 neurons with limited fan-in, the Spikey chips can simulate 8 neurons with fan-in comparable to cortical neurons, and the current SpiNNaker chip is projected to perform about 18,000 much simpler point-neuron simulations. To scale to the billions of neurons in the mammalian cortex, millions of chips would be required using present-day CMOS technology, even with the simple point-neuron model.

- *Connectivity:* A separate problem involves the fan-in and fan-out of connections between the emulated neurons within and between the chips. No current technology allows for reasonably-sized emulated neuron circuits with an average of 10,000 inputs and outputs to other neuron circuits on a chip: the combinatorial explosion of input wires to each neuron would overwhelm any integrated circuit

layout with more than a few dozen neurons. Thus, almost all of the projects we detailed have turned to digital networking. However, there are bandwidth and circuitry limitations to networking, as we will discuss shortly.

- *Power and heat issues:* The human brain is incredibly power-efficient. It is a 3-dimensional computer with liquid coolant/power delivery. Even with the most energy-efficient integrated circuit technologies we have, heat dissipation and total power requirements will be a problem in scaling to the size of biological brains. The neuromorphic solutions (FACETS, Neurogrid, NeuroDyn, IFAT) are most promising in terms of lowest power consumption per emulated neural time unit. For example, the HICANN wafer uses only 1 nJ of energy per synaptic transmission, less than a single instruction on an ARM processor. At the same time, the speed of neural emulation and communication in neuromorphic solutions can run 10,000 times faster than biological equivalents.

- *Ancillary issues:* The artificial brain projects propose ancillary mechanisms that must also scale along with the artificial brain. For example, if a software process separate from the actual neuron implementations is responsible for learning, then it must scale alongside. A question arises as to whether the learning process be decomposed into many cooperating learning processes on multiple processors distributed throughout a network.

How well does AER networking scale? If there are about 40 billion neurons in the human cortex and thalamus, with an average axon fan-out to 10,000 synapses, firing at an average of 10 times per second, then AER networking would need to deliver about 4×10^{14} packets per second, with at least 4×10^{10} originating packets per second. To put this in perspective, it is recently estimated that the total U.S. user Internet traffic averages about 8×10^8 packets per second. Admittedly, the AER packets are fewer bytes, and over short distances, but the routing overhead is comparable, and the routing tables are much bigger, given 40 billion destinations. Even if the firing rate is significantly lower than our estimate, the total traffic is staggering when taken as a whole.

Luckily, the actual number of inter-chip packets might be much smaller. There is evidence that interconnectivity is much higher within a cortical column and minicolumn. With an estimated 10,000 neurons in a cortical column, cortical columns could fit entirely on planned second-generation chips for projects such as SpiNNaker and FACETS. If 99% of connectivity is within a column, this reduces the inter-column and inter-chip bandwidth 100 times, and earlier-mentioned research on a "Rent exponent" by Bassett *et al.* [11] suggests that locality of connectivity may extend beyond cortical columns.

Even without locality of reference assumptions, the SpiNNaker group provides some evidence that their AER network can scale to 1 billion neurons with a fanout of 1,000 synapses per neuron [56]. While falling short of the scale of the human cortex, this is a promising result. Their torus of 65,000 interconnected chips, each with its own router, and each connected to 6 neighbors, allows more even distribution of load than hierarchical networks and subnetworks. Thus with the right networking topology, some help from locality of reference, and a significant portion of computing power dedicated to routing, it may be possible to support AER networking on the scale of the human brain, but this remains to be demonstrated.

Turning our attention from neuron connectivity to neuron emulation, there are scalability issues there as well.

As just discussed, Blue Brain emulates a small, fractional percent of brain neurons in much less than real time.

The analog neuromorphic circuit approach requires less hardware due to the special-purpose nature of the circuits, and the economy of analog "computations," absent any interconnection hardware. Because the computations are inexact, direct comparison is not possible. However, neurons with complexity somewhere between Blue Brain and SpiNNaker could be realized with synaptic plasticity and dendritic computations with less than a million transistors per neuron [37]. Given the potential for over a billion transistors per chip with current technology, each neuromorphic chip could hold over 1,000 neurons. However, note that many millions of chips would still be required for an artificial brain, and the connectivity and structural plasticity problems with neuromorphic analog circuits remain. Major breakthroughs in nanotechnology that allow 3-dimensional construction and real-time modification of electronic circuits would be required to achieve an analog whole brain.

The highest scale is achieved by SpiNNaker, with its simpler neuron model. However, even using SpiNNaker chips with 18 CPUs, over a million chips would be required for the human cortex. And for bio-realism of the complexity of Hodgkin-Huxley with two or more levels per neuron, and synapses and dendritic arbors with commensurate complexity, a much smaller number of neurons could be emulated by each CPU.

In summary, SpiNNaker's neuroprocessor approach gives the highest scalability but with limited bio-realism, the neurosimulation approach gives the highest bio-realism with scalability limited by the largest supercomputer available, and the neuromorphic approaches are in between in bio-realism, being limited in scalability until the "wiring" and circuit density problems are solved.

4.5 Conclusions

We are a long way from a working artificial brain. Given our limited of understanding of biological neurons and learning processes, the connectivity scalability issues, and the substantial computing power required to emulate neuronal function, readers may understandably be skeptical that a group of interconnected artificial neurons would possibly behave in a fashion similar to the simplest animal brains, let alone display intelligence.

With the drawbacks of all three approaches we discussed, it seems that there is not one good approach to all of the problems. In the near term, software is easier to experiment with than hardware, and connectivity and structural plasticity are practical to attain in partial brain emulations. Experiments with software may produce more progress on requirements for neural modeling. At the same time, experiments with analog hardware might demonstrate economies of scale and use of nanotechnologies for future whole brain emulation.

In our opinion, the most promising approaches depend on the goals and timeframe:

1. In the short term, scalability to the size of a mammalian brain is not practical, but software simulation seems most promising for emulation and understanding of small networks of neurons, e.g. to experiment with learning algorithms, since software is so easily modified.

2. An approach like SpiNNaker's, with loosely-coupled processors and AER networking, seems most likely to yield neural emulation on the scale of the entire brain in the medium term. The result might or might not behave like a biological brain, given the simplifying assumptions, e.g. using a point neuron model and delivering spikes rather than continuous synaptic connectivity.

3. In the long term, if a solution such as DNA-guided self-assembly of nanoelectronics is developed to allow self-guided wiring and re-wiring of dendrites and axons, a neuromorphic analog solution seems the only approach that can give bio-realism on a large scale.

Neuroscientists are uncovering additional mechanisms and aspects of neural behavior daily, and neural models for artificial brains may become significantly more complex with future discoveries. However, an artificial brain, limited and simplified, does provide a test bed for research into learning and memory, and we expect that substantial progress will be made through new technologies and ideas over the coming decades. In addition, research on smaller-scale artificial neural networks still provides considerable value in applications such as signal processing and pattern recognition, and experiments with neural networks may give us insights into learning and other aspects of the brain.

Independent of any value to neuroscience, efforts on artificial brain emulation also provide value to computer science. Projects such as SpiNNaker yield new computer architectures that have other valuable applications, and emulation of the brain may yield new approaches to artificial intelligence. If the goal is artificial intelligence or new computer architectures rather than bio-realistic brain emulation, then it is also possible that simpler neuron emulations would be adequate. Thus, we see value in continued research despite our pessimism about the timeframe and current technologies for brain emulation.

Acknowledgements

References

[1] National Academy of Engineering (nae.edu), *Grand Challenges for Engineering,* www.engineeringchallenges.org, 2010.

[2] W. McColloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity", in *Bulletin of Mathematical Biophysics* Vol 5, pp 115–133, 1943.

[3] F. Rosenblatt, "A Probabilistic Model for Information Storage and Organization in the Brain," Cornell Aeronautical Laboratory, *Psychological Review,* v65, No. 6, pp. 386–408, 1958.

[4] A.L. Hodgkin, A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerves," *Journal of Physiology* 117, pp 500–544, London, 1952.

[5] W. Rall, "Branching dendritic trees and motoneuron membrane resistivity," *Experimental Neurology* 1, pp 491–527, 1959.

[6] T. Berger, et al, "A cortical neural prosthesis for restoring and enhancing memory," *Journal of Neural Engineering 8*, 4, August 2011.

[7] Shepherd, G. "Introduction to Synaptic Circuits," in *The Synaptic Organization of the Brain*, edited by Gordon Shepherd, 5th edition, Oxford University Press, 2004.

[8] A. Polsky, B. W. Mel, J. Schiller, "Computational subunits in thin dendrites of pyramidal cells, *Nature Neuroscience,* www.nature.com/natureneuroscience, 2004.

[9] A. Losonczy, J. K. Makara, and J. C. Magee, "Compartmentalized dendritic plasticity and input feature storage in neurons," Nature, vol. 452, pp. 436–441, March 2008.

[10] S. Remy, J. Csicsvari, and H. Beck, "Activity-dependent control of neuronal output by local and global dendritic spike attenuation.," Neuron, vol. 61, pp. 906–916, March 2009.

[11] S. Bassett, D. L. Greenfield, A. Meyer-Landenberg, D. R. Weingerge, S. W. More, E. T. Bullmore, Efficient Physical Embedding of Topologically Complex Information Processing Networks in Brains and Computer Circuits, PLoS Computational Biology 6, 4, 2010.

[12] R. D. Fields, *The Other Brain: From Dementia to Schizophrenia, How New Discoveries about the Brain Are Revolutionizing Medicine and Science*. Simon & *Schuster, 1 ed., December 2009.*

[13] J. Joshi, A. Parker, K Tseng, "An *in-silico* glial microdomain to invoke excitability in cortical neural networks," IEEE International Symposium on Circuits and Systems, Rio de Janeiro, Brazil, May 2011.

[14] A. Sandberg, N. Bostrom, *Whole Brain Emulation: A Roadmap,* Technical Reprot 2008-3, Future of Humanity Institute, Oxford University, 2008.

[15] H. de Garis, C. Chuo, B. Goertzel, L. Ruiting, "A world survey of artificial brain projets, Part I: Large-scale brain simulations," Neurocomputing 74, Issues 1-3, pp. 3-29, December 2010.

[16] B. Goertzel, R. Lian, I. Arel, H. de Garis, S. Chen, "A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures," Neurocomputing 74, Issues 1-3, pp 30-49, September 2010.

[17] C. Koch, I. Segev, *Methods in Neuronal Modeling*, 2nd Edition, Massachusetts Institute of Technology, Cambridge, MA, 1998.

[18] Izhikevich, E. M., "Which model to use for cortical spiking neurons?," *IEEE Trans. Neural Networks* 15, 2004, pp 1063-1070.

[19] M. Mahowald, *Computation and Neural Systems,* PhD thesis, California Institute of Technology, 1992.

[20] D. Hebb, *The Organization of Behavior,* Wiley & Sons, New York, 1949.

[21] D. Allport, "Distributed memory, modular systems, and dysphasia," in S. Newman and R. Epstein, *Current Perspectives in Dysphasia,* Churchill Livingstone, Edinburgh, 1985.

[22] N. Ziv, "Principles of glutamatergic synapse formation: seeing the forest for the trees," Current Opinion in Neurobiology, vol. 11, pp. 536–543, October 2001.

[23] C. Verderio, S. Coco, E. Pravettoni, A. Bacci, and M. Matteoli, "Synaptogenesis in hippocampal cultures," Cellular and Molecular Life Sciences, vol. 55, pp. 1448– 1462, August 1999.

[24] S. Furber and S. Temple, "Neural Systems Engineering*", in *Studies in Computational Intelligence,* Springer Verlag, 2008.

[25] H. Markram, "The Blue Brain Project", in *Nature Reviews,* Volume 7, February 2006, pp 153-160.

[26] D. Modha, et al, "Cognitive Computing: Unite neuroscience, supercomputing, and nanotechnology to discover, demonstrate, and deliver the brain's core algorithms," Communications of the ACM 54, 8, pp 62-71, August 2011.

[27] P. Merolla, J. Arthur, F. Akopyan, I. Nabil, R. Manohar, D. Modha: "A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45pJ per Spike in 45nm", IEEE Custom Integrated Circuits Conference (CICC), San Jose, 2011.

[28] J. Schemmel et al, "A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neuron Modeling", in *Proceedings of the 2010 IEEE International Symposium on Circuits and Systems,* 2010.

[29] R. Silver, K. Boahen, S. Grillner, N. Kopell, K. Olsen, "Neurotech for Neuroscience: Unifying Concepts, Organizing Principles, and Emerging Tools", *Journal of Neuroscience,* pp 11807-11819, October 2007.

[30] R. Jacob Vogelstein et al, "A Multichip Neuromorphic System for Spike-Based Visual Information Processing", in *Neural computation,* Volume 19, 2007, pp 2281-2300.

[31] J. Park, T. Yu, C. Maier, S. Joshi, G. Cauwenberghs, "Hierarchical Address-Event Routing Architecture for Reconfigurable Large Scale Neuromorphic Systems," *IEEE International Symposium on Circuits and Systems*, Rio de Janiero, May 2011.

[32] T. Yu and G. Cauwenberghs, "Analog VLSI Biophysical Neurons and Synapses with Programmable Membrane Channel Kinetics", *IEEE Transactions on Biomedical circuits and Systems,* 4, 3, p 139-148, June 2010.

[33] X. Jin, A. Rast, F. Galluppi, S. Davies, S. Furber, "Implementing Spike-Timing-Dependent Plasticity on SpiNNaker Neuromorphic Hardware," IEEE World Congress on Computational Intelligence, Barcelona, Spain, July 2010.

[34] E.M. Izhikevich, "Simple Model of Spiking Neurons," *IEEE Trans. Neural Networks*, vol. 14, no. 6, 2003, pp. 1569–1572.

[35] A. Rast, S. Furber, et al, "Scalable Event-Driven Native Parallel Processing: The SpiNNaker Neuromimetic System," *CF '10,* Bertinoro, Italy, May 2010.

[36] NEURON simulation software, http://www.neuron.yale.edu/neuron, 2005.

[37] D. Modha and R. Singh, "Network architecture of the long-distance pathways in the macaque brain," Proceedings of the National Academy of Sciences of the USA 107, 30 (June 2010), 13485–13490.

[38] M. Ehrlich, et al, "A software framework for mapping neural networks to a wafer-scale neuromorphic hardware system," Proceedings Artificial Neural Networks and Intelligent Information Processing Conference, 2010.

[39] BrainScaleS Project Overview, http://brainscales.kip.uni-heidelberg.de.

[40] Brain-i-Nets, consortium overview, http://brain-i-nets.kip.uni-heidelberg.de/.

[41] J Lin, P Merolla, J Arthur and K Boahen, "Programmable Connections in Neuromorphic Grids," *49th IEEE Midwest Symposium on Circuits and Systems*, pp 80-84, IEEE Press, 2006.

[42] S. Joshi, S. Deiss, M. Arnold, J. Park, T. Yu, G. Cauwenberghs, "Scalable Event Routing in Hierachical Neural Array Architecture with Global Synaptic Connectivity," *Proceedings 12th International Workshop on cellular Nanoscale Networks and their Applications (CNNA),* 2010.

[43] H. de Garis, C. Shuo, B. Goertzel, L. Rulting, "A World survey of artificial brain projects, Part I: Large-scale brain simulations", Neurocomputing 74, Issues 1-3, December 2010, pp 3-29.

[44] A. Parker et al, "Towards a Nanoscale Artificial Cortex", *2006 International Conference on Comuting in Nanotechnology,* June 26, 2006, Las Vegas, USA.

[45] J. Joshi, J. Zhang, C. Wang, C. Hsu, A. Parker, "A Biomimetic Fabricated Carbon Nanotube Synapse with Variable Synaptic Strength," IEEE/NIH 5th Life Science and Applications Workshop, 2011.

[46] J. Patwardhan, C. Dwuyer, A. Lebeck, D. Sorin, "Circuit and System Architecture for DNA-Guided Self-Assembly of Nanoelectronics," Prcoeedings of Foundations of Nanoscience, ScienceTechnica, 2004.

[47] G. Snider, "From Synapses to Circuitry: Using Memristive Memory to Explore the Electronic Brain," IEEE Computer 44, 20, p21-28, 2001.

[49] E. Izhikevich and G. Edelman, *Large-scale Model of Mammalian Thalmocortical systems,* The Neurosciences Institute, San Diego, CA, 2007.

[50] G. Indiveri et al, "Neuromorphic Silicon Neuron Circuits," *Frontiers in Neuroscience 5,* 73, May 2011.

[51] A. Meltzoff, P. Kuhl, J. Movellan, T. Sejnowski, "Foundations for a new science of learning," Science 325: 284-288 (2009).

[52] E. Farquhar, C. Gordon: "A field programmable neural array," IEEE International Symposium on Circuits and Systems, May 2006.

[53] C. Mead, *VLSI and Neural Systems,* Addison-Wesley, 1989.

[54] S. Philipp, A. Grubl, K. Meier, J. Schemmel, "Interconnecting VLSI Spiking Neural Networks using Isochronous Connections," in Proc 99th International Work-Conference on Artificial Neural Networks, Springer LNCS 4507 pp 471-478, June 2007.

[55] E. Kandel, *In Search of Memory*, W.W. Norton, 2006

[56] J. Navaridas, M. Lujan, J. Miguel-Alonso, L. Plana, S. Furber, "Analysis of the Multicast Traffic Induced by Spiking Neural Networks on the SpiNNaker Neuromimetic System," Proceedings 23rd ACM Symposium on Parallelism in Algorithms and Architectures, San Jose, June, 2011.