DOI:10.1145/1978542.1978559

Unite neuroscience, supercomputing, and nanotechnology to discover, demonstrate, and deliver the brain's core algorithms.

BY DHARMENDRA S. MODHA, RAJAGOPAL ANANTHANARAYANAN, STEVEN K. ESSER, ANTHONY NDIRANGO, ANTHONY J. SHERBONDY, AND RAGHAVENDRA SINGH

Cognitive Computing

WHAT IS THE mind? Neither scientists nor philosophers agree on a universal definition or specification. Colloquially, we understand the mind as a collection of processes of sensation, perception, action, emotion, and cognition. The mind can integrate ambiguous information from sight, hearing, touch, taste, and smell; it can form spatiotemporal associations and abstract concepts; it can make decisions and initiate sophisticated coordinated actions.

Cognitive computing aims to develop a coherent, unified, universal mechanism inspired by the mind's capabilities. Rather than assemble a collection of piecemeal solutions, whereby different cognitive processes are each constructed via independent solutions, we seek to implement a unified computational theory of the mind. AI pioneer Allen Newell described it as "a single set of mechanisms for all of cognitive behavior. Our ultimate goal is a unified theory of human cognition."

Historically, many disparate fields have taken radically different approaches to exploring mind-like computation, some of which we cover here. On the one hand, strong artificial general intelligence, or AI,[10] a branch of cognitive science, takes a system-level approach to synthesizing mind-like computers. Since the mind arises from the wetware of the brain, neuroscience[18] takes a component-level approach to understanding how it gives rise to the mind. Proceeding top-down in a reductionist fashion, cognitive neuroscience[9] seeks to integrate theoretical cognitive science with experimental psychology and organism-level neuroscience. In contrast, proceeding bottom-up in a constructive fashion, systems neuroscience[18] seeks to combine experimental data at multiple spatial and temporal scales. The diversity of thought implicit in this plurality of approaches is essential, given the profound technological importance and scientific difficulty of the mind-brain problem. Science thrives on multiple groups taking different, complementary, parallel perspectives while working at different levels of abstractions.

Against this backdrop, our novel, promising approach is to operationalize vast collections of neuroscience data by leveraging large-scale computer simulations. Today, a thoughtful selection from the riches of neurophysiology and neuroanatomy can be combined to produce near real-time

» key insights

- **Cognitive computing will lead to novel learning systems, non-von Neumann computing architectures, programming paradigms, and applications that integrate, analyze, and act on vast amounts of data from many sources at once.**
- **A new white-matter long-distance network spanning the entire Macaque monkey brain consisting of 383 regions and 6,602 connections opens fresh ways to analyze, understand, and eventually, imitate the network architecture of the brain.**
- **Path-breaking developments in cortical simulation algorithms enable cat-scale cortical simulations on Lawrence Livermore National Laboratory's Dawn Blue Gene/P supercomputer with 147,456 CPUs and 144TB of main memory.**

Measuring White-Matter Pathways in the Human Brain

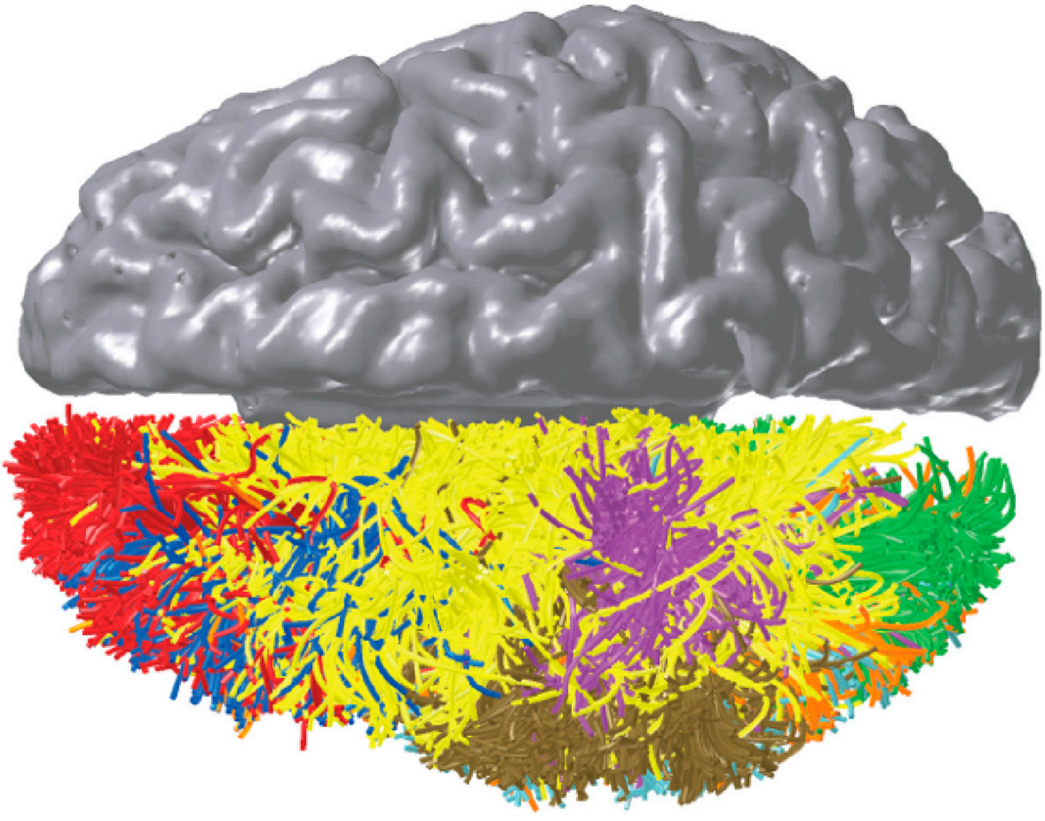
The white matter of the human brain comprises more than 150 kilometers of long-range projections. Understanding the architecture of these projections (the "projectome") is important for understanding brain function and has led to fundamental discoveries in normal and pathological brains (see Figure 1). Despite these findings, the bulk of the human projectome, or complete map of axonal projections in the brain, remains unexplored, with many exciting questions to answer.

Recent advances in diffusion-weighted magnetic resonance imaging (DW-MRI) have allowed noninvasive measurement of the human white-matter network across the entire brain. DW-MRI acquires an aggregate description of the microscopic diffusion of water molecules, along many directions within millimeter-size chunks of brain tissue. The dense packing of axon bundles within the white matter imposes oriented obstacles faced by water molecules. By measuring diffusion patterns produced by these obstructions, DW-MRI can help determine the location and orientation of axon bundles.

Unfortunately, the aggregation of the microscopic diffusion measurements at the millimeter-scale spatial resolution introduces ambiguity in the inference of the underlying projectome. Resolving it demands evaluation of an enormous number of potential pathways in order to estimate the full projectome from DW-MRI data. As a consequence, DW-MRI axonal-tracing techniques often estimate only one projection at a time, attempting to trace a single fiber through the white matter using only local measurements, with no regard for the paths of other fibers. These local, greedy optimization methods are not well suited for estimating the entire projectome, as they ignore critical global criteria (such as data prediction, where the model predicts the diffusion data that matches the measurements, and physical-volume constraints, where white-matter volume is finite).

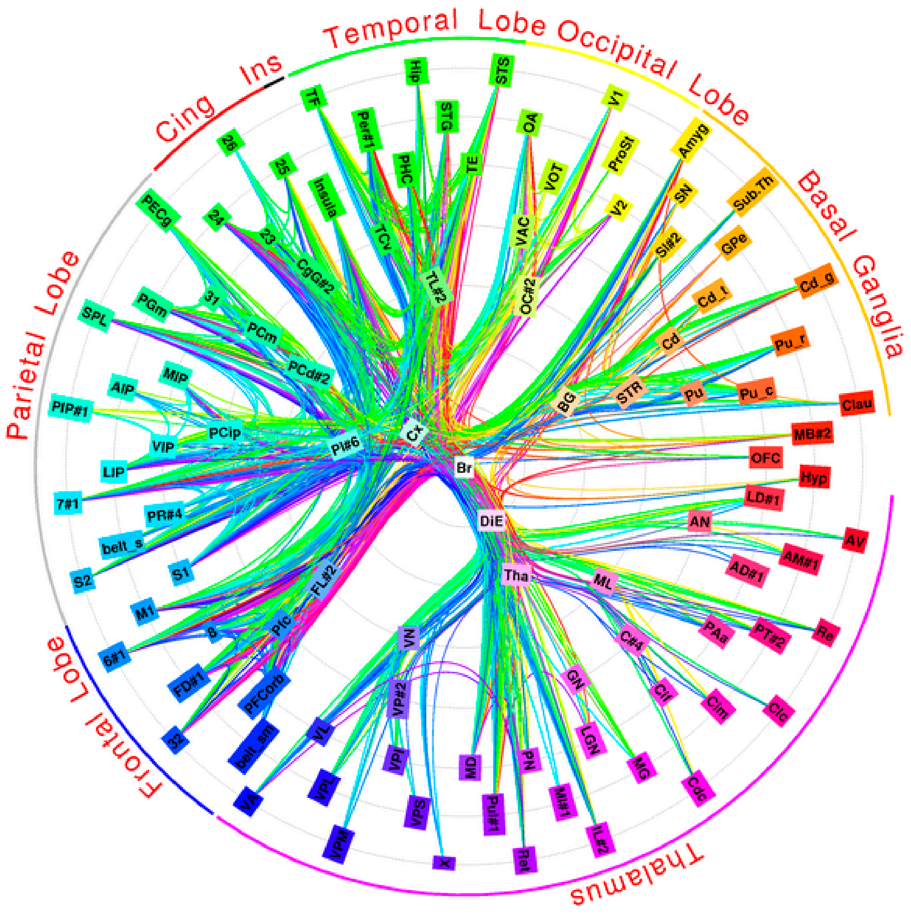
To address these shortcomings we have developed a parallel algorithm for global projectome evaluation that uniquely accounts for global prediction error and volume conservation.34 Leveraging the IBM Blue Gene/L supercomputing architecture, the algorithm first creates a massive database of 180 billion candidate pathways using multiple local tracing algorithms, then employs a global-optimization algorithm to select a subset of these candidates as the projectome. The estimated projectome (in the figure) accounts for 72 million projections per square centimeter of cortex and is therefore the highest-resolution, volume-conserved, collaborative projectome of the human brain. This surpasses previously achieved projectome resolutions by a factor of at least a thousand.

Figure 1. Top view of a subset of the human-brain projectome created using our algorithm. Shown are major long-range connections of the brain, estimated through diffusion-weight MRI, bottom, with several important groups of pathways assigned distinct colors; the cortex (gray area) is for reference in the opposite hemisphere.



Analyzing White-Matter Pathways in the Macaque Monkey Brain

Figure 2. A white-matter graph in the macaque monkey brain consisting of 102 hierarchically organized super-areas and 1,138 connections. Each node is represented by a colored rectangle, labeled with an abbreviation from CoCoMac nomenclature. Each edge is represented by a colored curve such that the color transitions smoothly from the color of its target node to the color of its source node. Bundles of white-matter projections criss-cross the expanse of the brain.



Anatomical tracing in experimental animals has historically been the pervasive technique for mapping white-matter pathways. In these experiments, a dye is injected in one brain area and its percolation studied to discover white-matter projections to other brain areas. Thousands of such measurements, collected over decades, have generated a vast, but sometimes inconsistent, database of projections. We undertook the challenge of constructing, visualizing, and analyzing a unified, consistent white-matter graph spanning the macaque brain.[28]

We derived a novel white-matter graph incorporating 410 published anatomical tracing studies of the macaque brain from the neuroinformatic database CoCoMac.[19] Our graph consists of 383 hierarchically organized areas spanning cortex, thalamus, and basal ganglia; it also has 6,602 directed edges and captures well-known cortico-cortical, cortico-subcortical, and intra-subcortical white-matter pathways. This graph is three times larger than the largest previous white-matter network of the macaque brain and is eight times larger than one of the most commonly analyzed white-matter networks of the macaque brain.

We have unearthed several critical insights by leveraging the unprecedented scale of our graph and state-of-the-art tools from network theory, which has also proved invaluable to understanding the hidden structure of graphs (such as the Web, metabolic pathways, and social networks). The degree distribution of the graph is consistent with an exponential distribution and is not scale-free, thus settling a much-debated, foundational open question. The graph has six degrees of separation, is a small-world network, and is characterized by the principle of organized complexity. Additionally, the graph revealed that the prefrontal cortex, the seat of executive function, contains the lion's share of topologically central areas. Finally, the graph embodies a tightly integrated core circuit that corresponds extremely well with a network believed to be the substrate for higher cognition and consciousness. It is quite remarkable and reassuring that the graph recapitulates critical known fiber pathways in the visual system, the dorsal-ventral pathways, thalamocortical relays, and numerous corticocortical, corticosubcortical, and subcorticocortical fiber systems implicated in specific cognitive functions. Simulation of circuits incorporating these fiber systems may yield specific insights into these cognitive functions.

To compactly visualize the large, high-resolution graph and make it amenable to simulation, we aggregate neighboring areas into a smaller number of super-areas, thus sacrificing resolution. A connection between two brain areas in the original graph results in a connection between corresponding super-areas. The smaller, low-resolution graph (in the figure) contains 102 super-areas and, after eliminating duplicates and self-loops, 1,138 connections.

simulations at the scale of small mammalian brains. Though we have only humble achievements to report, our aspirations are lofty. We seek nothing less than to discover, demonstrate, and deliver the core algorithms of the brain and gain a deep scientific understanding of how the mind perceives, thinks, and acts. This will lead to novel cognitive systems, computing architectures, programming paradigms, practical applications, and intelligent business machines.

Rationale

Our rationale was aptly and eloquently captured by Churchland and Sejnowski, writing, "It would be convenient if we could understand the nature of cognition without understanding the nature of the brain itself. Unfortunately, it is difficult if not impossible to theorize effectively on these matters in the absence of neurobiological constraints. The primary reason is that computational space is consummately vast, and there are many conceivable solutions to the problem of how a cognitive operation could be accomplished. Neurobiological data provide essential constraints on computational theories, and they consequently provide an efficient means for narrowing the search space. Equally important, the data are richly suggestive in hints concerning what might really be going on and what computational strategies evolution might have chanced upon."

Neuroscience today is rich in detailed biological observations, as reflected in the sheer size—1,414 pages—of *Principles of Neural Science*, a modern introductory textbook by Kandel et al.[18] As neuroscientists, we view these observations as a web of clues to the biological mechanisms of cognition. As engineers, we view them as something else entirely. The brain is an example solution to the problem of cognitive computing, and the observations of neuroscience are a partial set of constraints on the form of that solution. The trick to leveraging neuroscience in the name of cognitive computing is to separate the wheat from the chaff.

Here, we explore the fundamental neuroscientific constraints on building a functional simulation of the brain, first describing structural constraints learned from the wiring diagram of the brain. The central message is the brain's neuronal network is a sparse, directed graph organized at multiple scales. In particular, local, short-range connections can be described through statistical variations on a repeating canonical subcircuit, whereas global, long-range connections can be described through a specific, low-complexity blueprint. We highlight what neurophysiology has taught us about the dynamics of computation and communication within this network. Our thesis is that the computational building blocks of the brain (neurons and synapses) can be described by relatively compact, functional, phenomenological mathematical models, and that their communication can be summarized in binary, asynchronous messages (spikes).

The overarching motivation of our approach is the fact that the behavior of the brain apparently emerges via non-random, correlated interactions between individual functional units, a key characteristic of organized complexity. Such complex systems are often more amenable to computer modeling and simulation than to closed-form analysis and often resist piecemeal decomposition. Thus, empowered by strides in supercomputing-based simulation, the rationale for our approach rests in our conviction that large-scale brain simulations, at the appropriate level of abstraction, amount to a critical scientific instrument, offering opportunities to test neuroscientific theories of computation and to discover the underlying mechanisms of cognition.

A critical judgment must be made as to the appropriate level of abstraction for simulation. This conundrum must be faced when modeling any physical system. If we would choose too high a level of abstraction, the black boxes within the model will themselves be hopelessly complicated and likely map poorly onto reality as our understanding grows. If we abstract away too little and work at too high a resolution, we will squander computational resources and obscure our own understanding with irrelevant detail. Unfortunately, no oracle exists to instruct us as to the correct balance between abstraction and resolution at the outset. The only solution is to experiment and explore as a community. It is a virtue that different schools of thought have emerged,[13,14,16,25,38] each with an argument for its own chosen level of abstraction in conceptualizing and modeling the brain. The most established traditions are at relatively high levels of abstraction and include efforts in AI, cognitive science, visual information processing, connectionism, computational learning theory, and Bayesian belief networks.[4,5,9,21–24,27,30,35,37] Meanwhile, other efforts have sought to pin down the opposite end of the spectrum, striving for ever-higher levels of reductionist biological detail in simulating brain tissue.[20,36] Here, we strike a balance between these extremes[11] and advocate a middle path,[12,15,17] one more faithful to the neuroscience than to an abstract connectionist model, yet less detailed than an exhaustive, biophysically accurate simulation. Depending on context, a telescope, a microscope, and binoculars each has a place in a scientist's repertoire.

Neuroanatomy

A central tenet of neuroscience, sometimes called the "neuron doctrine," posits that specialized cells in the brain, the neurons, are the biological substrate of brain computation. The function of individual neurons is covered later in the section on neurophysiology, but for now, neuronal function can be abstracted to receiving, integrating, and sending binary messages. These messages are communicated at points of contact, dubbed synapses by Sir Charles Sherrington in England in 1897. Through messaging, neurons collaborate to form networks that engender powerful capabilities, vastly more sophisticated than the processing capacity of individual neurons. To understand brain function, it is crucial to understand the organization of neural circuitry.

Connectivity in the brain is sparse. Adult humans have about 100 trillion synapses, six orders of magnitude less than would be required to completely and directly connect the tens of billions of neurons that make up the brain. Moreover, there is strong evidence that biology has a relatively compact algorithm for assembling this sparse network; animals with much larger brains and higher cognitive abilities (such as apes) do not have proportionately larger swaths of their genome devoted to neural function than animals with more modest brains and abilities (such as rodents). Perhaps it is unsurprising that neuroanatomists have not found a hopelessly tangled, arbitrarily connected network, completely idiosyncratic to the brain of each individual, but instead a great deal of repeating structure within an individual brain and a great deal of homology across species.

At the surface of the brains of all mammals is a sheet of tissue a few millimeters thick called the cerebral cortex, or simply cortex, thought to be the seat of higher cognition. It is organized at multiple scales, including layers running horizontally, vertically integrated columns spanning its depth, large functionally defined areas consisting of millions of columns, and ultimately networks of multiple areas that are tightly linked. Within this system, neurons are connected locally through gray-matter connections, as well as through long-range white-matter connections that leave the cortex to travel to distant cortical regions or sub-cortical targets (see Figures 1 and 2).

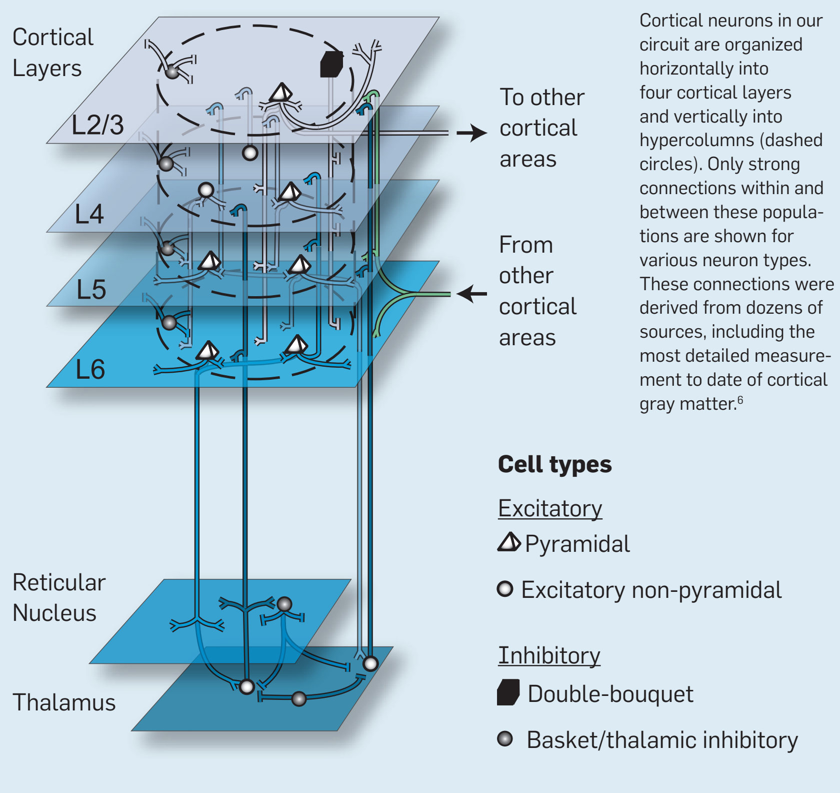
One of the earliest discoveries suggesting structure within cortex was the six distinct horizontal layers spanning the thickness of the cortical sheet. A specific network of connections between and within these cortical layers has been identified and studied,[6] giving rise to characteristic patterns of interlaminar activity propagation. We adapted this canonical laminar cortico-thalamic architecture into an archetypical gray-matter network amenable to simulation (see Figure 3).

The connections between layers are principally vertical, with limited lateral spread, leading to a columnar structure tens or hundreds of microns in diameter, referred to as a "cortical column." In many cortical areas, it has been demonstrated that neurons within the same column share related functional characteristics, suggesting that columns are functional, as well as structural, entities.[6,29] The information collected by measurements at the columnar scale has been instrumental in creating our large-scale brain models, as in Figure 3.

Cortical columns organize into cortical areas that are often several millimeters across and appear to be responsible for specific functions, including motor control, vision, and planning. Suggesting the possibility of a specific cortical circuit for each function, the famous Brodmann atlas, *Localization in the Cerebral Cortex*, offers a segmentation of the brain into cortical areas based on cellular density variations within the six cortical layers.[18] For example, Brodmann area 17 has been definitively linked to core visual-processing functionality. Decades of work by hundreds of scientists have focused on understanding the role each cortical area plays in brain function and how anatomy and connectivity of the area serve that function.

While overwhelming evidence in the 20th century supports the functional specialization of cortical areas, the brain also demonstrates a remarkable degree of structural plasticity. For example, it has been demonstrated that an area normally specialized for audition can function as one specialized for vision, and vice versa, by rewiring the visual pathways in the white matter to auditory cortex and the auditory pathways to visual cortex in the developing ferret brain. This astonishing natural reconfigurability gives hope that the core algorithms of neurocomputation are independent of the specific sensory or motor modalities and that much of the observed variation in cortical structure across areas represents a refinement of a canonical circuit; it is indeed this canonical circuit we wish to reverse engineer. The existence of such a canonical microcircuit is a prominent hypothesis,[29] and while a great deal about the local cortical wiring has been measured,[6] the exact form of this microcircuit remains unknown and its role in neurocomputation undemonstrated. Even if a base canonical circuit can be found, to unlock its potential we must also identify and implement the accompanying plasticity mechanisms responsible for tailoring, refining, and elaborating the canonical circuit to its specific function during development and in adult

Figure 3. A circuit diagram of the thalamocortical system simulated on the C2 cortical simulator.



Cortical neurons in our circuit are organized horizontally into four cortical layers and vertically into hypercolumns (dashed circles). Only strong connections within and between these populations are shown for various neuron types. These connections were derived from dozens of sources, including the most detailed measurement to date of cortical gray matter.[6]

learning. We later revisit the topic of plasticity, specifically its possible local, synaptic mechanisms.

At the coarsest scale of neuronal system organization, multiple cortical areas form networks to address complex functionality. For example, when reading, the brain executes a deft series of intricate eye movements that scan and fixate within words to extract a series of lines and edge combinations (letters) forming intricate spatiotemporal patterns. These patterns serve as keys to unlock a tome of linguistic knowledge, bathing the brain in the sights, sounds, smells, and physicality of the words' meaning. It is astounding that this complex functionality is mediated by a small network of tightly connected, but spatially distant, brain areas. This gives hope that distinct brain functions may be supported by signature subnetworks throughout the brain that facilitate information flow, integration, and cooperation across functionally differentiated, distributed centers. In 2009 and 2010, our group at IBM Research-Almaden achieved two breakthroughs[28,32] in measuring and analyzing the white-matter architecture of macaque and human brains as a means of furthering our understanding of networks of brain areas (see the sidebars "Analyzing White-Matter Pathways in the Macaque Monkey Brain" and "Measuring White-Matter Pathways in the Human Brain").

Neurophysiology

The adaptation of a biological cell into a structure capable of receiving and integrating input, making a decision based on that input, and signaling other cells depending on the outcome of that decision is a truly remarkable feat of evolution. Such cells, the neurons, were proposed to underlie information processing and cognitive function within the brain by Nobel laureate Santiago Ramón y Cajal in 1891. Neuronal function arises from a neuron's three main structural components: dendrites, tree-like structures that receive and integrate inputs; a soma, where decisions based on these inputs are made; and an axon, a long narrow structure that transmits signals to other neurons near and far. It is quite impressive that while cells are typically microscopic, axons can grow to more than a meter in length.

The adaptation of a biological cell into a structure capable of receiving and integrating input, making a decision based on that input, and signaling other cells depending on the outcome of that decision is a remarkable feat of evolution.

At the root of signal integration and transmission within a neuron are fluctuations in the neuron's membrane potential, the voltage difference across the membrane that separates the interior and exterior of a cell. These fluctuations occur when ions cross the neuron's membrane through channels that can be opened and closed selectively. If the membrane potential crosses a critical threshold, the neuron generates a spike (its determination that it has received noteworthy input), which is a reliable, stereotyped electrochemical signal sent along its axon. Spikes are the essential information couriers of the brain, used in the sensory signals the retina sends down the optic nerve in response to light, in the control signals the motor cortex sends down the spinal cord to actuate muscles, and in virtually every step in between.

When a spike arrives at the end of its axon, the nature of the signal changes. Synapses are tiny structures that bridge the axon of one neuron to the dendrite of the next, transducing the electrical signal of a spike into a chemical signal and back to electrical. The spiking neuron, called the "presynaptic neuron" in this arrangement, releases chemicals called neurotransmitters at the synapse that rapidly travel to the other neuron, called the "postsynaptic neuron." The neurotransmitters trigger ion-channel openings on the surface of the postsynaptic cell, subsequently modifying the membrane potential of the receiving dendrite. These changes can be either excitatory, meaning they make target neurons more likely to fire, or inhibitory, making their targets less likely to fire. Both the input spike pattern received and the neuron type determine the final spiking pattern of the receiving neuron. Through this process, the essentially digital electrical signal of the spike sent down one neuron is converted first into a chemical signal that can travel between neurons, then into an analog electrical signal that can be integrated by the receiving neuron.

The magnitude of this analog postsynaptic activation, called "synaptic strength," is not fixed over an organism's lifetime. Thus, the influence one neuron has on another can change, altering the functional relationships within a network of neurons. Canadian psychologist Donald O. Hebb's famous conjecture for synaptic plasticity is "neurons that fire together, wire together," or that if neuron A and B commonly fire spikes at around the same time, they will increase the synaptic strength between them. One modern refinement of Hebb's idea is that synaptic strengths may change depending on the relative timing of pre- and post-synaptic spikes through a mechanism called "spike-timing dependent plasticity," or STDP,[33] so neuron A strengthens its connection to neuron B if A tends to fire just before B fires, while connection strength is weakened if the firing order is reversed. There are also ongoing research efforts to link neuromodulatory chemicals, like dopamine, to more complex mechanisms for synaptic plasticity that resemble update rules from reinforcement learning.[31,34] The mechanisms of synaptic plasticity are a focus of active research, but no one can say for certain which mechanisms are most prevalent or most significant. However, it is widely believed among brain researchers that changes in synaptic strength underlie learning and memory, and hence that understanding synaptic plasticity could provide crucial insight into cognitive function.

While this provides a rough outline of neuron behavior, neuroscientists have uncovered a much more detailed picture of neuron function, including a host of different ion channels that produce oscillatory changes in membrane potential and regulate firing patterns, different synapse types that operate over a range of time courses, neuromodulators that produce changes in neuron behavior, and many other features that influence function.[18] Many different types of neurons can be distinguished based on these features, which have been captured in a number of models.

It should be noted that though it is widely agreed that spikes are the brain's primary information couriers, considerable debate concerns how spikes encode information. The dominant view has been that cortical neurons encode information in terms of their instantaneous firing rates, and the relative timing between spikes is essentially irrelevant. Studies have shown there is additional value in the precise timing of spikes, though the lion's share of the information is available in the spike rate. Further, recent evidence suggests the brain is able to detect and exploit artificially induced precise spiking timing.

We embrace the proposition that, because spikes are the universal currency of neuronal communication, a simulated network that reproduces the brain's temporal pattern of spiking must necessarily constitute a sufficient simulation of neural computation. In an idealized thought experiment, this simulation would predict the exact temporal pattern of spikes across the entire brain (including neuron bursting, correlations between neurons, and temporal synchrony) in response to arbitrary stimuli and contextual placement. It is uncontroversial to say that such an achievement, were it possible, would implicitly recapitulate biological neurocomputation. However, many researchers dispute that this implies that spikes are the correct level of abstraction at which to study and simulate the nervous system. Those who believe that working at a higher level of abstraction is preferable argue that the details of such a spiking simulation are irrelevant to the fundamental principles of cognition and actually obscure the key algorithms of brain-based computation.[27] On the other hand, those who believe that working at a finer resolution is required assert

Cortical Simulator Design and Implementation

Since 2007, we have been developing the C2 near-real-time mammalian-scale cortical simulator[2,3] to harness the distributed memory multiprocessor architecture of IBM Blue Gene systems (see Figure 4). Here, we discuss the core architecture of the simulator and highlight key innovations along the dimensions of memory, computation, and communication.

The cortical simulator includes a clock-driven component with discrete time steps, as well as an event-driven component. In the former, the state of the neurons is updated once every time step, typically either one millisecond or one-tenth of one millisecond of simulated time. In the latter, when a neuron fires, it creates a spike event that is then delivered to the synapse of a target neuron after a tunable axonal delay. The spike event has two essential functions: change the membrane potential of the target neuron and possibly trigger a change to the strength of the synapses on the axon and dendrites of the spiking neuron.

The entire state of the simulation (consisting of neurons, synapses, and transient spike messages) is evenly distributed among the local memories of the multiprocessor system. Each processor maintains the state of a group of neurons and all synapses providing input to these neurons. A notable C2 innovation is the memory-efficient representation of synaptic state, facilitating significantly increased model scales.

C2 harnesses a large number of processors while fully exploiting the computational capacity of each processor to achieve near-real-time simulation. Its design ensures that the number of computational operations at every time step is proportional to the number of spikes, rather than to the vastly larger number (typically a thousandfold) of synapses.

Most notably, C2 employs a novel synchronization technique requiring only two communication steps, in sharp contrast to previous algorithms that used communication steps in proportion to number of processors. When simulating with more than a hundred thousand processors, such communication optimizations are indispensable.

Neurons and synapses in representative mammals.

	Mouse	Rat	Cat	Monkey	Human
Billions of Neurons	0.016	0.055	0.763	2	20
Trillions of Synapses	0.128	0.442	6.10	16	200

that a precise reproduction of spike trains, though sufficient, is unattainable without including in the simulation the detailed dynamics of dendritic compartments, ion concentrations, and protein conformations. However, few brain researchers dispute that neurons are the fundamental cellular units of computation and that spikes are the messages passed between them. Because spikes therefore constitute a preferred level of description of neural communication, we focus on simulations dealing directly with spikes.[2,3,15]

Supercomputing Simulations

Neuroanatomy and neurophysiology, together, have produced a rich set of constraints on the structure and the dynamics of the brain. In our own work, we have aimed to integrate and operationalize a judiciously chosen subset of these constraints into a single computational platform, a mammalian-scale brain simulator.

The essential ingredients of the simulator include phenomenological model neurons exhibiting a vast behavioral repertoire, spiking communication, dynamic synaptic channels, plastic synapses, structural plasticity, and multi-scale network architecture, including layers, minicolumns, hypercolumns, cortical areas, and multi-area networks, as in Figure 3. Each of these elements is modular and individually configurable, so we can flexibly test a multitude of biologically motivated hypotheses of brain structure and dynamics. The ensuing enormous space of possible combinations requires that simulations run at speeds that permit rapid, user-driven exploration.

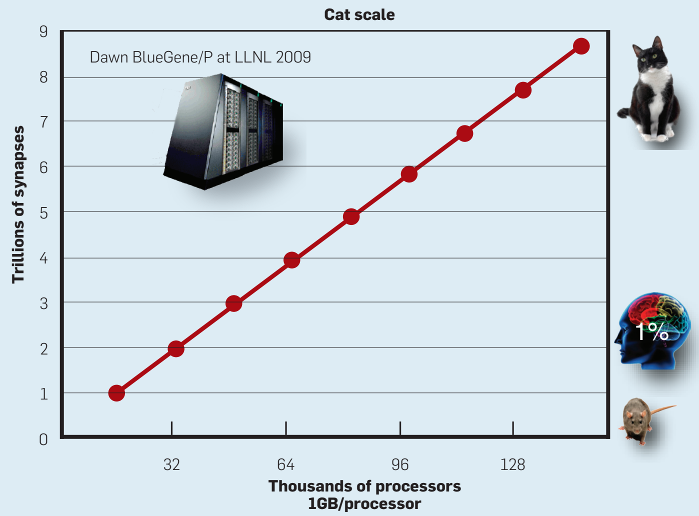
Neural simulations have a rich history dating to the 1950s. Since then, research in cortical simulations (see Figure 3) has progressed along two paths: detail and scale. Several publicly available simulators, including NEURON and GENESIS, allow for detailed simulation of a small number of neurons.[7] Unfortunately, incorporating such fine biophysical detail renders the task of near-real-time simulations at mammalian scale computationally impractical. On the other hand, using compact phenomenological neurons, other studies have demonstrated simulations of millions of neurons and billions of synapses.[3,15] Our objective is to push the boundaries of the state of the art along the dimensions of model scale and neuroanatomical detail while achieving near-real-time simulation speed.

Simultaneously achieving scale, speed, and detail in one simulation platform presents a formidable challenge with respect to the three primary resources of computing systems: memory, computation, and communication. For example, the cat cerebral cortex has almost a billion neurons and more than six trillion synapses (see the table opposite). Since synapses outnumber neurons by a factor of 10,000, memory requirements representing the state of the simulation scale in direct proportion to the number of synapses. Consequently, even if we could represent the state of a synapse in 1B, a cat-scale simulation would require at least 6TB of main memory; efficient synaptic data structures require about 16B of memory per synapse. Further, assuming that each neuron is updated once every millisecond, the dynamical difference equations governing neuronal state evolution must be computed one trillion times per second. With a biologically plausible average neuron firing rate of once per second, most synapses would receive a spike once a second, so six trillion spike messages must be communicated across the network. To meet this demand, we leverage magnificent strides in supercomputing, coupled with key innovations in algorithms and software architecture.

Along the hardware dimension, the Blue Gene supercomputer system offers large numbers of computational processors, vast amounts of main memory in a distributed architecture, and low-latency, high-bandwidth communication subsystems. Along the software dimension, we have developed a cortical simulator we call C2 that exploits the distributed-memory multiprocessor architectures. We have performed simulations of increasing scale and incorporated progressively richer neurophysiological and neuroanatomical constraints in our simulations (see the sidebar "Cortical Simulator Design and Implementation").

Since 2007, Our simulations have grown steadily in scale, beginning with early work at a scale of mouse and rat[3] cortices. We obtained our most recent result (see Figure 4) in May 2009 in collaboration with Lawrence Berkeley National Laboratory using the Dawn Blue Gene/P system, achieving

Figure 4. Scaling cortical simulations with C2.



Simulations on a Blue Gene/P supercomputer, including a 1% human-scale cortical model with 1.97 trillion synapses employing 32,768 processors and 32TB of main memory, culminating in a cat-scale cortical model with 8.61 trillion synapses using 147,456 processors and 144TB of main memory.

the newsworthy milestone of cat-scale cortical simulations, roughly equivalent to 4.5% of human scale, fully utilizing the memory capacity of the system.[2] The networks demonstrated self-organization of neurons into reproducible, time-locked, though not synchronous, groups.[3] The simulations also reproduced oscillations in activity levels often seen across large areas of the mammalian cortex at alpha (8Hz–12Hz) and gamma (> 30Hz) frequencies. In a visual stimulation-like paradigm, the simulated network exhibited population-specific response latencies matching those observed in mammalian cortex.[2] A critical advantage of the simulator is that it allows us to analyze hundreds of thousands of neural groups, while animal recordings are limited to simultaneous recordings of a few tens of neural populations. Taking advantage of this capability, we were able to construct a detailed picture of the propagation of stimulus-evoked activity through the network; Figure 5 outlines this activity, traveling from the thalamus to cortical layers four and six, then to layers two, three, and five, while simultaneously traveling laterally within each layer.

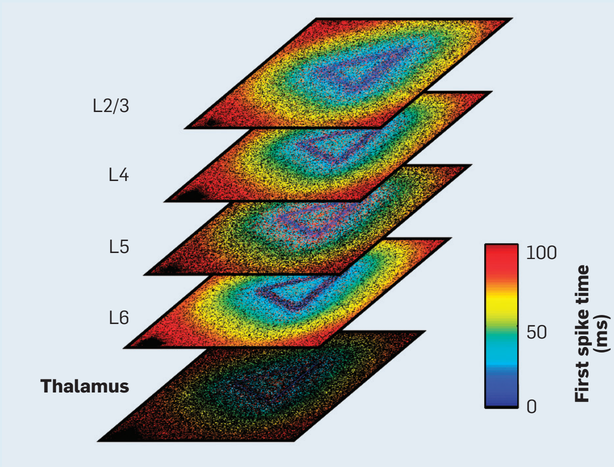
The C2 simulator provides a key integrative workbench for discovering algorithms of the brain. While our simulations thus far include many key features of neural architecture and dynamics, they only scratch the surface of available neuroscientific data; for example, we are now incorporating the long-distance white-matter projections (see the first two sidebars and Figures 1 and 2), other important sub-cortical structures (such as the basal ganglia), and mechanisms for structural plasticity. We remain open to new measurements of detailed cortical circuitry offered by emerging technologies.

The realistic expectation is not that cognitive function will spontaneously emerge from these neurobiologically inspired simulations. Rather, the simulator supplies a substrate, consistent with the brain, within which we can formulate and articulate theories of neural computation. By studying the behavior of the simulations, we hope to reveal clues to an overarching mathematical theory of how the mind arises from the brain that can be used in building intelligent business machines. In this regard, the simulation architecture we built is not the answer but the tool of discovery, like a linear accelerator, laying the groundwork for future insight into brain computation and innovations in neuromorphic engineering.

Prospective

The quest for intelligent machines ultimately requires new breakthroughs in philosophy, neuroanatomy, neurophysiology, computational neuroscience, supercomputing, and computer architecture orchestrated in a coherent, unified assault on a challenge of unprecedented magnitude. The state of today's effort in cognitive computing was best captured by Winston Churchill: "Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."

Figure 5. Simulated response of thalamocortical circuitry to a triangle-shape stimulus.



Included are 1.6 billion neurons and 8.61 trillion synapses. The plot is the propagation of the evoked response through the circuit, depicting the latency of the first spike at every topographic location in each cortical layer and the thalamus following stimulus delivery.

Included are 1.6 billion neurons and 8.61 trillion synapses. The plot is the propagation of the evoked response through the circuit, depicting the latency of the first spike at every topographic location in each cortical layer and the thalamus following stimulus delivery.

On the heels of the unprecedented simulation scale and the trends in development of supercomputer technology, the good news is that human-scale cortical simulations are not only within reach but appear inevitable within a decade.

The bad news is that the power and space requirements of such simulations may be many orders of magnitude greater than those of the biological brain. This disparity owes its genesis to the salient differences between the von Neumann architecture and the brain itself.[39] Modern computing posits a stored program model, traditionally implemented in digital, synchronous, serial, centralized, fast, hardwired, general-purpose, brittle circuits, with explicit memory-addressing imposing a dichotomy between computation and data. In stark contrast, the brain uses replicated computational units, neurons and synapses, implemented in mixed-mode analog-digital, asynchronous, parallel, distributed, slow, reconfigurable, specialized, fault-tolerant biological substrates, with implicit memory addressing blurring the boundary between computation and data.

The elegance and efficiency of biology entices us to explore entirely new computing architectures, system designs, and programming paradigms. Under the umbrella of the U.S. Defense Advanced Research Projects Agency (DARPA) Systems of Neuromorphic Adaptive Plastic Scalable Electronics initiative, beginning in 2008, we have embarked on an ambitious program to engender a revolutionarily compact, low-power neuromorphic chip comprising one million neurons and 10 billion synapses per square centimeter by exploiting breakthroughs in nanotechnology and neuromorphic very large-scale integration.[26]

Finally, the ugly news is that the core set of algorithms implemented within the brain are as yet undiscovered, making our task as replete with uncertainty as it is rich with opportunity. Confronting this challenge requires the sustained, coherent effort of a dedicated interdisciplinary community of researchers endowed with substantial resources.[1] At the moment, this grand endeavor proceeds in parallel at multiple scales of investigation: abstract cognitive primitives and artificial neural networks; extremely detailed biological models; and fundamental language of spiking communication favored by us and others. We hope future discoveries will demonstrate these approaches to be complementary, each with its own virtues and each contributing to a unified solution to the challenge of cognitive computing. We share the inspired enthusiasm of U.S. Secretary of Energy and Nobel laureate Steven Chu: "I do not underestimate the difficulty of meeting these challenges, but I remain optimistic that we can meet them. I believe in the vitality of our country and our economy, and as a scientist, I am ever optimistic at our ability to extend the boundaries of what is possible."

Acknowledgments

The research reported here was sponsored by DARPA's Defense Sciences Office's Program: Systems of Neuromorphic Adaptive Plastic Scalable Electronics (SyNAPSE) under Contract HR0011-09-C-0002. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or of the U.S. Government.

We would like to thank Horst Simon for his collaboration and the U.S. Department of Energy National Nuclear Security Administration Advanced Simulation and Computing Program for time on the Dawn Blue Gene/P supercomputer at Lawrence Livermore National Laboratory operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration, under Contract DE-AC52-07NA27344. We have also benefited tremendously from supercomputing resources at IBM Research-Almaden, IBM Research-Thomas J. Watson Research Center, and KAUST/IBM Center for Deep Computing Research.

We are grateful to T. Binzegger, R.J. Douglas, and K.A.C. Martin for sharing their thalamocortical connectivity data and indebted to Gregory Corrado for numerous insightful discussions and significantly clarifying our exposition here.

References

1. Albus, J.S., Bekey, G.A., Holland, J.H., Kanwisher, N.G., Krichmar, J.L., Mishkin, M., Modha, D.S., Raichle, M.E., Shepherd, G.M., and Tononi, G. A proposal for a decade of the Mind Initiative [Letter]. *Science 317*, 5843 (Sept. 7, 2007), 1321.
2. Ananthanarayanan, R., Esser S.K., Simon H.D., and Modha, D.S. The cat is out of the bag: Cortical simulations with 10^9 neurons and 10^{13} synapses. Gordon Bell Prize Winner. In *Proceedings of the ACM/IEEE Conference on Supercomputing* (Portland, OR, Nov. 14–20). ACM, New York, NY, 2009, 1–12.
3. Ananthanarayanan, R. and Modha, D.S. Anatomy of a cortical simulator. In *Proceedings of the ACM/IEEE Conference on Supercomputing* (Reno, NV, Nov. 10–16). ACM, New York, NY, 2007, 3–14.
4. Arbib, M., Ed. *The Handbook of Brain Theory and Neural Networks.* MIT Press, Cambridge, MA, 2002.
5. Baumgartner, P. and Payr, S. Eds. *Speaking Minds: Interviews with 20 Eminent Cognitive Scientists.* Princeton University Press, Princeton, NJ, 1995.
6. Binzegger, T., Douglas, R.J., and Martin, K.A.A. quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience 24*, 39 (Sept. 2004), 8441–8453.
7. Brette, R. et al. Simulation of networks of spiking neurons: A review of tools and strategies. *Journal of Computational Neuroscience 23*, 3 (July, 2007), 349–398.
8. Gazzaniga, M.S. *The Cognitive Neurosciences, Fourth Edition.* MIT Press, Cambridge, MA, 2009.
9. George, D. and Hawkins, J. Towards a mathematical theory of cortical microcircuits. *PLoS Computational Biology 5*, 10 (Oct. 2009), e1000532.
10. Goertzel, B. and Pennachin, C. *Artificial General Intelligence.* Springer, Berlin, Heidelberg, 2009.
11. Herz, A.V.M., Gollisch, T., Machens, C.K., and Jaeger, D. Modeling single-neuron dynamics and computations: A balance of detail and abstraction. *Science 314*, 5796 (Oct. 2006), 80.
12. Hill, S. and Tononi, G. Modeling sleep and wakefulness in the thalamocortical system. *Journal of Neurophysiology 93*, 3 (Nov. 2005), 1671–1698.
13. Hodgkin, A.L. and Huxley, A.F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology 117*, 4 (Aug. 1952), 500–544.
14. IEEE Symposium on Information Theory. MIT, Cambridge, MA, 1956.
15. Izhikevich, E.M. and Edelman, G.M. Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Sciences of the USA 105*, 9 (Mar. 2008), 3593–3598.
16. Jeffress, L.A. *Cerebral Mechanisms in Behavior: The Hixon Symposium.* John Wiley & Sons, Inc., New York, 1955.
17. Johansson, C. and Lansner, A. Towards cortex-sized artificial neural systems. *Neural Networks 20*, 1 (Jan. 2007), 48–61.
18. Kandel, E.R., Schwartz, J.H., and Jessell, T.M. *Principles of Neural Science, Fourth Edition.* McGraw-Hill Medical, New York, 2000.
19. Kötter, R. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics 2*, 2 (June 2004), 127–144.
20. Markram, H. The Blue Brain Project. *National Review of Neuroscience 7*, 2 (Feb. 2006), 153–160.
21. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman, Cambridge, MA, 1983.
22. McCarthy, J. Generality in artificial intelligence. In *Formalizing Common Sense*, V. Lifschitz, Ed. Ablex, Bristol, U.K., 1990, 226–236.
23. McClelland, J.L. and Rumelhart, D.E. Parallel Distributed Processing. In *Psychological and Biological Models, Volume 2.* MIT Press, Cambridge, MA, 1987.
24. McCulloch, W.S. and Pitts, W. How we know universals: The perception of auditory and visual forms. *Bulletin of Mathematical Biology 9*, 3 (Sept. 1947), 127–147.
25. McCulloch, W.S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5* (1943), 115–133.
26. Mead, C. *Analog VLSI and Neural Systems.* Addison Wesley Publishing Co., Boston, 1989.
27. Minsky, M. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* Simon & Schuster, New York, 2006.
28. Modha, D.S. and Singh, R. 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.
29. Mountcastle, V.B. *Perceptual Neuroscience: The Cerebral Cortex.* Harvard University Press, Cambridge, MA, 1998.
30. Rosenblatt, F. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 6 (Nov. 1958), 386–408.
31. Schultz, W., Dayan, P., and Montague, P.R. A neural substrate of prediction and reward. *Science 275*, 5306 (Mar. 1997), 1593–1599.
32. Sherbondy, A.J., Dougherty, RF., Ananthanaraynan, R., Modha, D.S., and Wandell, B.A. Think global, act local: Projectome estimation with BlueMatter. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention Society, Lecture Notes in Computer Science* (London, Sept. 20–24). Springer, Berlin, 2009, 861–868.
33. Song, S., Miller, K.D., and Abbott, L.F. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience 3*, 9 (Sept. 2000), 919–926.
34. Sutton, R.S. and Barto, A.G. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review 88*, 2 (Mar. 1981), 135–140.
35. Teddington Symposium on the Mechanization of Thought Processes (National Physical Laboratory, Nov. 24–27). Her Majesty's Stationary Office, London, 1958.
36. Traub, R.D., Contreras, D., Cunningham, M., Murray, H., LeBeau, F.E.N., Roopun, A., Bibbig, A., Wilent, W.B., Higley, M., and Whittington, M.A. Single-column thalamocortical network model exhibiting gamma oscillations, spindles, and epileptogenic bursts. *Journal of Neurophysiology 93*, 4 (Nov. 2005), 2194–2232.
37. Valiant, L.G. *Circuits of the Mind.* Oxford University Press, New York, 2000.
38. von Hayek, F.A. *The Sensory Order: An Inquiry Into the Foundations of Theoretical Psychology.* University of Chicago Press, Chicago, 1952.
39. von Neumann, J. *The Computer and the Brain, Second Edition.* Yale University Press, New Haven, CT, 2000.

For recommended additional reading and sources, see the online appendix in the ACM Digital Library.

Dharmendra S. Modha (dmodha@almaden.ibm.com) is the manager of Cognitive Computing at IBM Research-Almaden, San Jose, CA.

Rajagopal Ananthanarayanan (ananthr@google.com) is a member of the technical staff of Google, Inc., Mountain View, CA, and was at IBM Research-Almaden, San Jose, CA, when he participated in this work.

Steven K. Esser (sesser@us.ibm.com) is a research staff member of IBM Research-Almaden, San Jose, CA.

Anthony Ndirango (andirang@gmail.com) was at IBM Research-Almaden, San Jose, CA, when he participated in this work.

Anthony J. Sherbondy (tony@addepar.com) is an engineer in Addepar, Mountain View, CA, and was at IBM Research-Almaden, San Jose, CA, when he participated in this work.

Raghavendra Singh (raghavsi@in.ibm.com) is a research staff member of IBM Research-India, New Delhi.