Title

International Neuroscience Initiatives Through the Lens of High-Performance Computing

Journal

Computer, 51(4)

ISSN

0018-9162

Authors

Bouchard, Kristofer E
Aimone, James B
Chun, Miyoung
et al.

Publication Date

2018-04-01

DOI

10.1109/mc.2018.2141039

Peer reviewed

PERSPECTIVES

International Neuroscience Initiatives through the Lens of High-Performance Computing

Kristofer E. Bouchard, Lawrence Berkeley National Laboratory and UC Berkeley
James B. Aimone, Sandia National Laboratories
Miyoung Chun, Kavli Foundation
Thomas Dean, Google and Stanford University
Michael Denker and Markus Diesmann, Jülich Research Center
David D. Donofrio, Lawrence Berkeley National Laboratory
Loren M. Frank, UC San Francisco and Howard Hughes Medical Institute
Narayanan Kasthuri, Argonne National Labs and University of Chicago
Christof Koch, Allen Institute for Brain Science
Oliver Rübel and Horst D. Simon, Lawrence Berkeley National Laboratory
F.T. Sommer, UC Berkeley
Prabhat, Lawrence Berkeley National Laboratory

Neuroscience initiatives aim to develop new technologies and tools to measure and manipulate neuronal circuits. To deal with the massive amounts of data generated by these tools, the authors envision the co-location of open data repositories in standardized formats together with high-performance computing hardware utilizing open source optimized analysis codes.

Many international neuroscience initiatives are in different stages of progression—although they have different goals, they will all produce large amounts of data. Much attention has been focused on the technological challenges of measuring and manipulating neural activity from large numbers of sites for long periods, but much less attention has been paid to the computing challenges associated with the vast amounts of data these technologies will generate. As a result, potential advances offered by neurotechnologies are threatened by a lack of computing tools. The neuroscience community is not alone in this challenge, as other science fields are being transformed by advanced analytics being applied to an ever-increasing volume of experimental data. Co-location of massive datasets hosted in open repositories with high-performance computing (HPC) will allow for community-driven exploratory analysis and integration with simulations. This is required to extract universal design principles of biological computation, which might

0018-9162/18/$33.00 © 2018 IEEE

provide insight into and inspiration for new models of in silico computation.

GRAND CHALLENGE PROBLEMS IN NEUROSCIENCE

To understand the brain is to know how its structure (wiring diagram) and function (activation dynamics) give rise to specific computations and behaviors. Following the goals of the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative and the EU Human Brain Project (HBP), we propose four grand challenge problems in neuroscience for which HPC will likely play an important role (see Figure 1): neuroanatomy and structural connectomics; neural population dynamics and functional connectomics; linking sensations, brains, and behaviors; and synthesis through simulations. While each challenge would provide useful information by itself, integrating them will result in synergistic understanding. Taken together, the results of these challenges will deepen our understanding of network mechanics that generate complex behaviors and the transformation of sensory inputs into neural representations of sensations. Furthermore, the results will provide insight into how brains achieve near-optimal computing capabilities and link structure to function across many spatiotemporal scales.

Neuroanatomy and structural connectomics

Across species, brains consist of hundreds to billions of individual neurons that are connected by thousands to trillions of synapses (see Figure 2). These anatomical features are the structural backbone from which all neuronal function is generated. In its most microscopic form, *structural connectomics* refers to the reconstruction of tissue at sufficient resolution to trace the finest neuronal processes and identify synaptic connections. Currently, the only approaches that offer the required resolution over large volumes are based on automated serial electron microscopy.

Advances in microscope design have improved resolution and acquisition time so that segmentation and annotation is the rate-limiting step.[1] The result of such anatomical reconstruction could be a full 3D representation of each neuron or a graph of the resulting structural connectivity matrix with some measure of synaptic strength (the matrix C_s: N neurons x N neurons). A structural connectome would provide a compact summary sufficient for some analyses, and is required to link structure to function in the nervous system.

Neural population dynamics and functional connectomics

Although neuroanatomy defines the possible interactions among neurons, it is the dynamically modulated spatiotemporal patterns of activations across neurons that give rise to sensations, actions, cognition, and consciousness. New technologies enable increasingly large numbers of brain signals to be recorded simultaneously, and these signals can be derived from diverse recording modalities. Furthermore, the duration of recordings is concurrently increasing, and it will soon be possible to record continuously for weeks to months. Thus, it is becoming increasingly important to develop data-analysis methods capable of revealing structure from heterogeneous, nonstationary time-series measurements at scale.

Challenges	
Structural connectomics	
Scaling:	neurons^3
Dataset:	serial electron microscopy of brains
Product:	circuit wiring diagram
Functional connectomics	
Scaling:	neurons^2 × time
Dataset:	in vivo activation levels of neurons
Product:	circuit dynamics
Sensations, brains, and behaviors	
Scaling:	modalities × time
Dataset:	sensory and/or behavioral measurements
Product:	computations performed by circuits
Biophysically detailed simulations	
Scaling:	synapses × time
Dataset:	connectivity and biophysics of neurons
Product:	bridge across spatiotemporal scales

FIGURE 1. Grand challenge problems in neuroscience. We pose four grand challenge problems in neuroscience, which, at scale, will require high-performance computing (HPC). This figure summarizes how problems approximately scale with key features and provides example inputs (data types) and outputs (insights gained) associated with each problem. Note that each of these problems scales approximately as the product of at least two key features of the dataset (for example, neurons^2 x time). (Source: Christian Swinehart, Samizdat Drafting Co.)

Two complementary approaches to this problem are dimensionality reduction and functional connectomics. Dimensionality reduction methods aim to find low-dimensional spaces that concisely summarize high-dimensional spatiotemporal patterns of activity, and can be used to gain insight into network dynamics. Functional connectomics aims to determine time-resolved causal influences among spatially distributed neural recordings (for example, the data array C_f: N neurons x N neurons x time for cellular-level data). Together, these complementary methods will provide insight into dynamic interactions among individual neurons and neural populations that can be linked to underlying structural connectomics and behavior.

Linking sensations, brains, and behaviors

Brains have evolved to produce behaviors in response to sensory events that

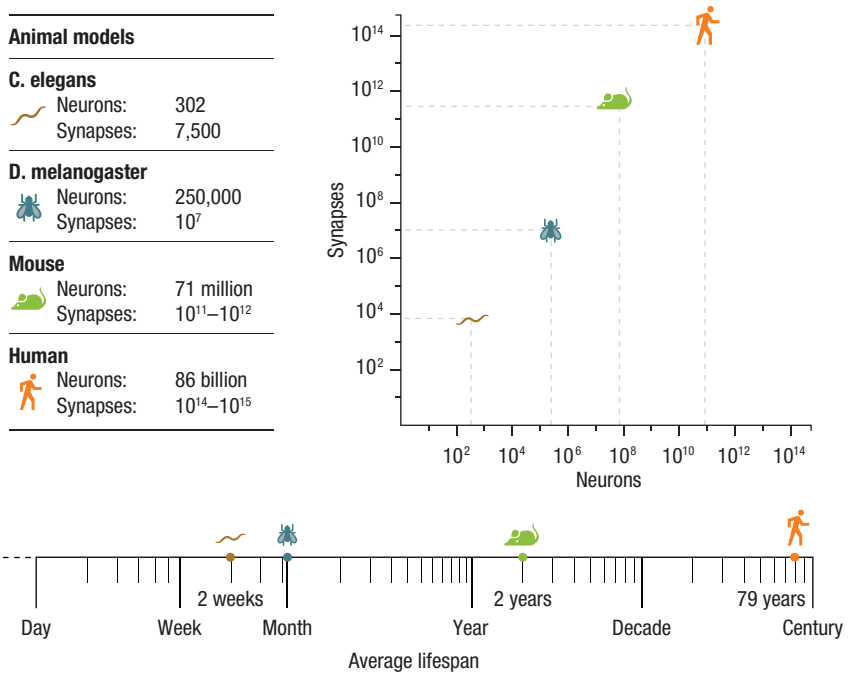

FIGURE 2. Neuroscience datasets will scale exponentially across species. Neuroscientists study brains across several species of varying complexity. This figure depicts the exponential grown in brain size (number of neurons and synapses) and lifespan across c.elegans (worms), d. melanogaster (flies), mice, and humans. (Source: Christian Swinehart, Samizdat Drafting Co.)

increase the probability of organismal survival and reproductive fitness. As such, our ability to infer brain function depends on linking brain measurements to events and objects in the external world. In many cases, the stimuli and behaviors used in neuroscience studies have been relatively simple (low-dimensional), which eases data analyses but elicits neural activity far removed from activity underlying naturalistic sensations and behavior. Indeed, it has recently been noted that simple tasks result in simple neural activity patterns.[2] Thus, it is insufficient to record from more neurons without simultaneously monitoring behavior during increasingly complex sensory, motor, and cognitive tasks. This brings with it challenges of acquiring, analyzing, and integrating such multimodal data (for example, visual, audio, haptic, and movement) with the brain data.

A causal understanding of the brain requires combining measurements with closed-loop manipulations of neural circuits triggered by neuronal and behavioral observations during complex tasks. However, this is very challenging because the activity of many neurons is rapidly modulated (tens of milliseconds) by sensations, cognitive tasks, and behavior. Given this complexity, manipulations need to be targeted to specific neurons when they are engaged in specific types of information processing, requiring real-time analysis of neural signals.

Synthesis through simulations

Ultimately, the goal of neuroscience is to achieve a deeper, broader understanding of the brain that extends across spatial and temporal scales. However, simply acquiring data without simultaneously developing guiding (theoretical) principles will impede extracting understanding from the data, and runs the risk of misguided investment into costly experiments. As other fields have shown, common computational and theoretical frameworks should permeate research directions while scaling up data acquisition and analysis to reduce the challenge of integrating information from very different levels of system granularity.

Many research domains focused on highly complex, multiscale problems (for example, climate, high-energy physics, and cosmology) have effectively leveraged HPC capabilities to integrate data and analysis into evolving theoretical frameworks through the use of simulations. This has been useful in precisely those conditions where extensive experimentation is intractable due to either cost or feasibility of data acquisition. While a general "theory of brain" seems a distant goal, using HPC for large-scale simulation of neural circuits and networks has a long history. Continued scaling (both in number and accuracy) of neural circuit simulations, as well as tighter integration with experimental data, is critical to connect spatiotemporal scales.

DIVERSE COMPUTING PLATFORMS FOR A DIVERSE COMMUNITY

Modern computing solutions are as diverse as the needs of the neuroscience community, and it is unlikely that there will be a one-size-fits-all solution to all needs. Indeed, there are several important tradeoffs in performance, cost, and accessibility associated with different computing platforms. Neuroscientists should be aware of these when deciding on current solutions and planning long-term investments.

While many neuroscientists are familiar with computing capabilities contained within a single laptop or by a shared cluster, there is less familiarity with the resources available or what problems may be tackled by cloud computing or supercomputing facilities.

DEFINITIONS OF TERMS

Brain signals: There are many diverse signals that can be used as measures of brain function, including intracellular and extracellular electrical recordings, optical imaging of neuronal voltage/calcium, electrocorticography (ECoG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). The spatial and temporal resolution of different signals is generally proportional to the invasiveness of the methods used to sense the signal.

Closed loop: Used here to describe an experiment that uses the signals generated by a system (such as the brain) to trigger a perturbation of that system.

Electrophysiological recordings: The recording of electrical activity generated by the biophysical processes involved in neuronal signaling. Electrophysiology provides the highest temporal resolution measurements of brain signals, capable of resolving the precise timing of individual action potentials ("the speed of thought").

Neuronal skeletons: The external structure of single neurons, including the cell body, dendrite, axon, and associated synapses.

Performance: A metric typically correlated with obtaining a quick turnaround time to solution. A well-balanced computational system will optimize for storage, network, memory, and compute performance.

Productivity: A metric typically associated with effort spent by developers and programmers in developing code and utilizing computational infrastructure.

Spike sorting: The problem of assigning recordings of action potentials ("spikes") to individual neurons from extracellular electrophysiological data, which are typically composed of a superposition of electrical signals from multiple neurons.

Cloud computing infrastructure fundamentally relies on commodity hardware with somewhat better network interconnects (10 Gbytes) than typical university clusters. The cloud is popular because of the ease with which resources can be provisioned and software services can be utilized, without exposing users to fault-prone hardware. HPC provides well-balanced CPU and memory subsystems, tightly coupled with high-performance interconnects, and data is typically read over massively parallel file systems capable of terabyte-per-second read/write performance. In aggregate, state-of-the-art HPC centers are capable of holding hundreds of terabytes of data in memory, and can calculate at the rate of petaflops. In contrast to cloud computing, where productivity is the norm, the software stack on HPC resources is tuned for performance, and it does take some degree of expertise and familiarity to fully utilize such systems. As the rate of improvement in computing processors decreases from a slowing of Moore's law, commodity computers (and clusters of them) will not be able to address the ever-increasing volumes of neuroscientific data.

In neuroscience, there also exists a need for experimenters to rapidly interrogate their data to receive timely feedback on results of experiments (for example, to evaluate the position of a recording device). There are other experiments (such as closed-loop perturbations) that require real-time analysis (<5–10 ms). These latencies can be achieved with carefully designed PC-based systems or in a more cost-effective fashion through specialized hardware utilizing field programmable gate arrays (FPGAs). In this context, the power of FPGAs comes from their flexible and high bandwidth (terabytes per second) connections and their ability to manipulate data from multiple sources. However, programming FPGAs can be challenging, and while familiar to many in the computing world, it is a skill rarely found in the neuroscience community. Finally, for on-sensor processing of massive datastreams from large-scale experimental equipment for which the processing algorithm has been established and agreed upon, application-specific integrated circuits (ASICs) can provide a critical filter, dramatically reducing the amount of data that needs to be moved and written to disk.

DATA MANAGEMENT CHALLENGES FOR NEUROSCIENCE

Advanced data management and sharing is required to accommodate the increasing complexity, data rates and acquisition times, and multimodality of neuroscience data. Similar problems occur with neuronal circuit simulations, where standardization of formats and data is an ongoing endeavor. The data-management requirements correspond to specific needs of experimentalists and analysts. Experimentalist requirements include fast write and efficient storage of large data volumes, resilience to corruption to minimize loss, extensible data standards, and collection of metadata with the raw data for reproducibility. Analyst requirements include: common standards to enhance portability and usability for storage, sharing and access; fast read for efficient analysis; integration of distributed, multimodal data sources; and provenance for interpretation and reuse. Standardization of data formats and data sharing are critical to maximize the return on investment into acquisition of experimental data, and will require close interactions among neuroscientists, data model designers, and data analysts.

Data standardization requires convergence of file storage formats and organization, as well as metadata standards and ontologies, and is an unresolved challenge in the field. File format and metadata standards have to be fully supported and well-integrated with acquisition, ideally through automation. To enable efficient analysis, standards also need to be well-integrated with the analysis pipelines, requiring advanced APIs. The need to interpret data in place, combined with efficient discoverability across hardware resources, means metadata should be centrally accessible and machine readable. Integration of multiple modalities requires effective modeling of complex semantic and structural relationships among data. Together, these capabilities will enable the neuroscience community to effectively store, analyze, and share data, accelerating discovery and enhancing reproducibility.

Sharing and reuse of data is essential to enable validation of neuroscience results, which will enhance reliable and unbiased scientific interpretation. The close integration of computing resources with data will enable effective data-driven discovery. This includes co-location of hardware resources to enable efficient processing while reducing costs for large transfers, as well as integrated management and analysis software stacks for analysis at scale. To enable the utilization of shared resources, centralized science gateways/portals that collect data and make it searchable and accessible are needed. Ultimately, data analyses result in commodities that become shared and reused. Similar to the role of metadata for raw measurements, data provenance (including the denotation of methods, parameters, and so on) is required for reliable interpretation of analyses. Meeting all these needs requires advanced, high-performance infrastructure that computer science and HPC centers are ideally positioned to provide.

COMPUTING CHALLENGES FOR NEUROSCIENCE

Repositories hosting data in standardized formats co-located with HPC resources will present the opportunity for creation of automated analysis pipelines at scale. The extraction of information from most modern neuroscience experiments and simulations requires application of sophisticated data-analysis methods. Indeed, each grand challenge problem has data processing (such as segmentation and spike sorting) and analytics challenges associated with it. As there is a cost associated with all computing resources, optimized analysis codes developed by experts in an open source community are required for efficient resource utilization. Building workflows and frameworks for distributed computing, and embedding them in a collaborative setting, requires expertise that is well outside that of typical neuroscientists. Here, although we describe specific analysis issues in the context of a single grand challenge problem, many of the issues apply to all of the problems. Figure 3 schematizes the computing resource requirements, using different species as anchor points, and emphasizes the need for advanced computing solutions at scale.

Neuroanatomy and structural connectomics

The current fastest approach for acquiring structural connectomics data involves scanning multiple electron beams over a brain sample in parallel and is already producing approximately 50 terabytes per day. The major time-limiting step in the analysis pipeline is the stitching, alignment, segmentation, and annotation required to obtain an accurate reconstruction of the brain.[1] The segmentation challenge is compounded by two facts: the task requires tracing most

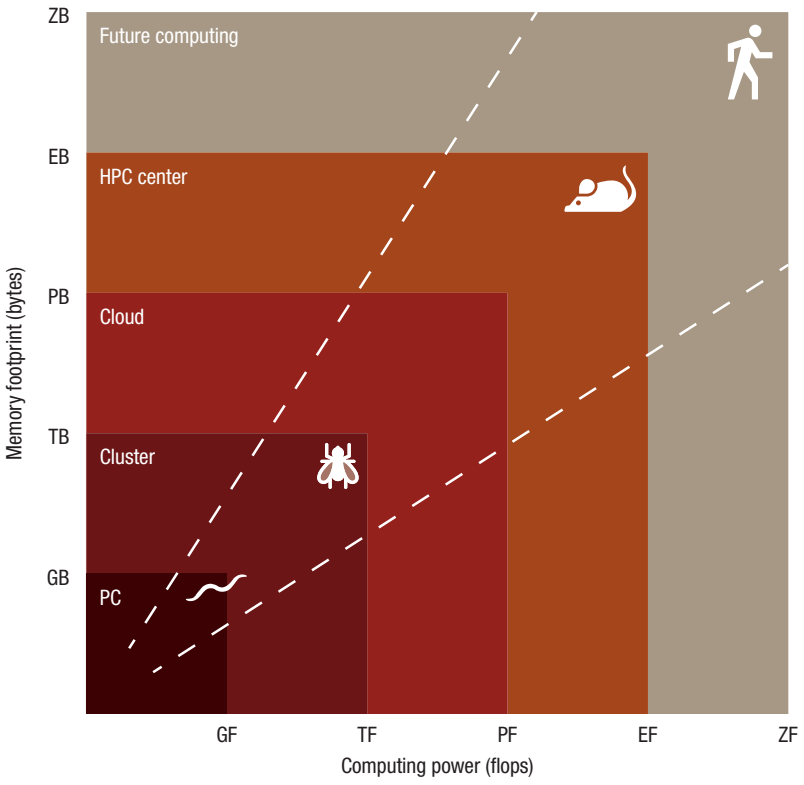

FIGURE 3. Grand challenge problems in neuroscience will push the boundaries of computing. Schematic of the computational demands (computing power [flops] and memory footprint [bytes]) of the grand challenge problems associated with four species. We project that these problems will scale approximately within the boundaries outlined by the dashed line. GB: gigabytes, TB: terabytes, PB: petabytes, EB: exabytes, ZB: zettabytes, GF: gigaflops, TF: teraflops, PF: petaflops, EF: exaflops, ZF: zettaflops. (Source: Christian Swinehart, Samizdat Drafting Co.)

objects in the field of view, and accuracies must be good because even small mistakes tracing an axon could result in thousands of synapses being incorrectly assigned.

The best algorithms for segmentation use recurrent 3D convolutional neural networks to automatically segment raw data.[3] This method performs 60 Mflops per voxel, which for a 100 μm^3 cube of cortex requires 200,000 GPU hours. This is relatively slow when one considers the total amount of computation required to render any reasonably sized circuit. However, once the necessary precision has been achieved, it is reasonable to assume that code optimization will allow processing of cubic millimeter scale samples.

While a cubic millimeter generates about a petabyte (PB) of data, an entire mouse brain will require close to an exabyte (EB) of data. This implies the need to avoid storing raw data long term, and instead processing data as it is generated by co-locating the imaging and computing hardware. This will require special-purpose hardware (such as ASICs and FPGAs) optimized for structural connectomic pipelines. Much of the computational challenge of structural connectomics is image processing, for which HPC systems have long been used. In addition to the 3D reconstruction of all neuropil in the sample, morphological analysis of neuronal skeletons is another important and computationally demanding task. Finally, once raw data has been processed and the resultant connectivity matrix has been extracted, analyses will need to be performed. Rigorous analytics on graphs consisting of 100 billion nodes (number of neurons in the human brain) and 10 billion edges (10 percent average connectivity) requires numerical linear algebra methods that can both exploit the structure of the graph (for example, sparse vs. dense and local vs. global connectivity) and are tailored to the available computing resources.

Neural population dynamics and functional connectomics

In 5 to 10 years, technologies will allow electrophysiological recordings in individual animals from (approximately) 10^6 neurons from brain networks of intermediate size, continuously (temporal resolution of 1 kHz) for many days (for example, 10 days: 8.64 x 10^8 ms), corresponding to approximately 3.5 PB of time-series data for a single animal (multiply by 25 for the high sampling rate raw data). Much of the analysis for large-scale neural population recording is focused on data-driven discovery, where our knowledge of the ground truth is vague at best. The accuracy of many machine learning/statistical data-analysis methods will be greatly enhanced by increased recording durations (and hence increased numbers of samples) afforded by next-generation recording technologies. Application of state-of-the-art analysis methods to large-scale functional datasets will benefit from efficient implementations in HPC systems.

Neural population activity and dynamics have been examined with dimensionality reduction methods for more than 15 years, and this approach has recently experienced a resurgence. Owing primarily to its computational ease, the most prevalent method in neuroscience is principal components analysis (PCA), the core calculation of

which is the singular-value decomposition (SVD). Performing very-large-scale SVDs is nontrivial, but has recently been performed on terabyte-sized matrices using HPC systems with techniques from randomized linear algebra.[4] Moving forward, convolutional-based methods (for example, convolutional non-negative matrix factorizations [NMF], which can result in interpretable "parts-based" decomposition) allow for the simultaneous extraction of spatiotemporal basis. Convolution-based methods can require significant computational resources and so implementations on GPUs and HPC architectures could greatly reduce compute time.

A complementary approach to dimensionality reduction for understanding spatiotemporal structure of neural populations is to use temporally directed analysis methods to infer causal influences among neurons. Such "functional connectomes" explicitly represent the influences between individual physical recordings and are thus potentially more interpretable than PCA. Open challenges in the application of such methods (such as generalized vector autoregressive models) to large-scale neurophysiology datasets include improving scalability of implementations, increasing the sparsity of estimated connectivity while retaining temporal contiguity, scaling of sparse identification of nonlinear dynamical systems to large dimensions (neurons), and accounting for the nonstationarity of data obtained in complex behavioral experiments. Other scalable methods to identify recurring spatiotemporal activity patterns in neuronal data (such as recurrent neural networks) should also be investigated.

Analysis of time-series data, while ubiquitous across the physical sciences and central to neural recordings, has not been performed in many biological fields. For example, as the cost of genetic sequencing and proteomics continues to fall, it is likely that biologists (and clinicians) will not only collect more samples from larger populations, but also more samples from the same individual over time. Thus, continued investigation in data analytics for time-series data, coupled with implementations in HPC systems (for example, to distribute the optimization calculation underlying many of these methods) are likely to become important for interpreting biomedical data in the near future.

Linking sensations, brains, and behaviors

In natural environments, animals process many high-dimensional sensory signals that are co-modulated over diverse time scales to produce complex sequences of behaviors toward achieving goals. Neuroscientists traditionally use hand-engineered features to characterize sensory and behavioral events, but the inherent arbitrariness of this selection impedes insight into neural coding principles. Sparse coding methods extract features from naturalistic sensory datasets (primarily visual and auditory) that can provide a principled account of response properties in primary sensory areas.[5] Characterizing the joint high-order statistics of natural sensory signals is challenging, as it requires the analysis of large amounts of multimodal data collected over long periods.

Neuronal activity is modulated on rapid time scales by many factors, including attention, reward, variation in arousal, and so on. Disentangling the contributions of different factors requires causal perturbations of neural activity during task engagement. Methods for rapid, cell-type-specific manipulations of neural activity in awake, behaving animals provide the capability to perform high-resolution closed-loop experiments. To be effective, this requires incoming brain signals to be processed in real time and analyzed to identify specific patterns, and then to trigger the manipulation in milliseconds (<10 ms), necessitating sufficient computational power and fast interconnects. Modern approaches to distributed computation in local computer clusters can likely solve current problems, but scaling to future data volumes might be challenging. Alternatively, FPGAs are ideally suited for real-time processing of massive datastreams, but programming such hardware is outside the purview of most neuroscientists.

Synthesis through simulations

While simulations have long been a tool of neuroscience, until recently most have been downscaled in size. Downscaled networks can preserve first-order statistics of neuronal activity (means) but higher-order statistics (cross-correlations) are generally not preserved. Mesoscale measures of activity, which are key tools for understanding the human brain, are driven by the fluctuations of neuronal populations, which are dominated by correlations. Thus, if a simulation is inaccurate in the second-order statistics of the microscopic activity, predictions of the mesoscopic activity obtained by forward modeling might be misleading.

HPC systems have recently enabled researchers to construct anatomically detailed models of local cortical circuits and perform full-scale simulations of neurons with all their synapses.[6] While work on cortical microcircuits is progressing, current implementations have fundamental shortcomings. First, many brain functions are distributed over several areas, and thus cannot be understood by studying an isolated microcircuit. Second, each neuron receives about half of its excitatory inputs from distant sources; thus, isolated models of cortical microcircuits are severely underconstrained. These challenges might be addressed by increasing the scale of neural circuit simulations—a task well-suited to HPC.

While detailed biophysical simulations are important, other approaches have targeted more abstract simulations with less biological fidelity.[7] These modeling approaches have different goals: "bottom-up" models aim to emulate high-level phenomena by constraining low-level parameters, while "top-down" models aim to replicate specific computations in a region. Linking the biophysical reality of bottom-up models with the well-defined computations of top-down models could reveal the biophysical mechanisms of neural computations. More effort in this direction is required.

A central challenge facing very-large-scale neural circuit simulations is understanding how best to evaluate the quality of results of models underconstrained by relatively limited amounts of experimental data. Conversely, it should be possible to examine the accuracy of a data-analysis algorithm (such as spike sorting and functional connectivity estimation) on data from simulations for which the ground truth is known. Techniques for understanding the uncertainty of model outputs (uncertainty quantification) and overall sensitivity of models relative to input parameterization are areas that need to be further explored.

Now that major obstacles in memory usage have been removed, reduction of the simulation time becomes the relevant target. Next-generation supercomputers (the so-called exascale systems) will be able to represent major parts of the human brain at microscopic resolution, and could be used to make predictions of the effects of pharmaceuticals testable in humans with mesoscale measurements.

THERE IS A GROWING REQUIREMENT FOR HPC ARCHITECTURES TO BE LESS SIMULATION FOCUSED AND MORE DATA INTENSIVE.

FROM BIOLOGICAL BRAINS TO DIGITAL BRAINS AND BACK

Exponential increases in computational capabilities have fueled the establishment of simulation as the third leg of modern science (complementing classic theory and experiments). Importantly, because Moore's law is slowing down, HPC has found itself at a crossroads, as illustrated by the National Strategic Computing Initiative (NSCI). The challenge of making HPC systems bigger and faster is joined by a change in their desired use.

Specifically, simulations have been the driving application behind HPC's growth, and development of these systems has been focused on their requirements. However, there is a growing requirement for HPC architectures to be less simulation focused (higher flops) and more data intensive (higher-performance I/O bandwidth to memory and between nodes).[8] We expect that neuroscience, with its special demands on the balance between memory access and compute power, will further drive this shift in supercomputing.

Many neuroscientists are envisioning *interactive supercomputing*—using a supercomputer like a superworkstation for exploring large datasets. This requires a supercomputer to be managed more like a telescope: individual research groups would be assigned time, as opposed to a scheme that executes a job when it optimally fits on the system. The price HPC centers will pay for this is a decline in the overall system utilization metric, but the benefit will be a much broader scientific user base.

As Moore's law draws to end, alternatives to von Neumann's model of

computing are being explored. Neuromorphic hardware is being developed, which might enable large-scale brain models and potentially advance more brain-like computing systems. Hardware implementations must make engineering tradeoffs to achieve specific computational advantages.

For instance, IBM's TrueNorth is a specialized chip achieving large-scale computation at low power, but with relatively restricted connectivity, low-precision synapses, and no on-chip learning. In contrast, the Spinnaker system modifies commercially available ARM cores that are directly programmable, and thus permits flexible implementations of neural circuits, albeit without energy savings. Finally, the BrainScales system uses analog approaches for accelerating simulations of neural circuits to study long-time scale processes like learning and development, but these approaches are nontrivial to implement and challenging to program.

Innovations in neuromorphic hardware might inspire new technology for classical systems. Perhaps design concepts that enable low-power computation in the brain might be used in next-generation supercomputing facilities to satiate their voracious power consumption. Additionally, computer scientists might find continued inspiration for stable, adaptive computing systems from the brain's synaptic and cellular learning processes.

Collaboration between brain scientists and computer scientists has a long history. Shortly before his death, von Neumann started writing about the similarities and differences between computers and brains.[9] Today, many burgeoning collaborations benefit both fields. The droves of data produced by the world's neuroscience initiatives could be an application area that ushers in a new age for HPC focused on experimental and observational data. Drawing inspiration from brain circuitry has enabled the development of low-power computing chips. Modern computer systems and algorithms are able to leverage massive datasets to train deep neural networks to effectively solve many problems for which progress had essentially plateaued. Together, these collaborative ventures have revived interest in artificial intelligence—the true nexus of the two fields.

There are many opportunities for neuroscience to benefit from HPC and computer science. Co-location of open neuroscience data repositories with HPC hardware will greatly support neuroscience efforts to reveal universal design features of species' brains and to understand what makes each individual unique—a central concept of precision medicine. Relating the structural and functional connectomes is necessary to deepen biophysical understanding of neural computations by mapping anatomical wiring diagrams to functional properties. Quantitative methods for understanding structure–function relationships is a ubiquitous problem in many fields (structural biology and material science, for example). In this context, developing a theoretical framework for understanding learning in deep neural networks (where we have precise knowledge of the structural connectivity, the activation of every unit, the objective function, input statistics, and learning dynamics) is a prerequisite to a normative "theory of brain" that links structure and function (this observation has been made by Stanford University physicist Surya Ganguli, among others). However, we speculate that a "theory of brain" would require much more than deep learning and would likely build on concepts from nonequilibrium statistical mechanics, information theory, optimal control and decision theory, (deep) learning theory, sparse coding, Bayesian inference, and sensor fusion.

We believe that addressing the challenges described here would create infrastructure to enable the neuroscience community to utilize HPC systems, and would provide a prototype for long-term strategic engagements between HPC centers and other scientific communities in the age of data-driven discovery. This would impact scientific return from major federal investments across multiple initiatives—immediate and sustained investment is required.

REFERENCES

1. J.W. Lichtman, H. Pfister, and N. Shavit, "The Big Data Challenges of Connectomics," *Nature Neuroscience*, vol. 17, no. 11, 2014, pp. 1448–1454.
2. P. Gao and S. Ganguli, "On Simplicity and Complexity in the Brave New World of Large-Scale Neuroscience," *Current Opinion in Neurobiology*, vol. 32, 2015, pp. 148–155.
3. M. Januszewski et al., "Flood-Filling Networks," arXiv:1611.00421, 2016; https://arxiv.org/abs/1611.00421.
4. A. Gittens et al., "Matrix Factorization at Scale: A Comparison of Scientific Data Analytics in Spark and C+MPI Using Three Case Studies," arXiv:1607.01335, 2016; https://arxiv

ABOUT THE AUTHORS

KRISTOFER E. BOUCHARD is principal investigator of the Neural Systems and Engineering Lab at Lawrence Berkeley National Laboratory (LBNL) and the Helen Wills Neuroscience Institute at UC Berkeley. He received a PhD in systems neuroscience from UC San Francisco. Contact him at kebouchard@lbl.gov.

JAMES B. AIMONE is a principal member of technical staff in the Center for Computing Research at Sandia National Laboratories. He received a PhD in neurosciences from UC San Diego. Contact him at jbaimon@sandia.gov.

MIYOUNG CHUN is executive vice president of science programs at the Kavli Foundation. She received a PhD in molecular genetics from the Ohio State University. Contact her at chun@kavlifoundation.org.

THOMAS DEAN is a research scientist at Google and teaches computational neuroscience at Stanford University. He received a PhD in computer science from Yale University. Contact him at tld@google.com.

MICHAEL DENKER is a research scientist at the Institute of Neuroscience and Medicine (INM-6), Computational and Systems Neuroscience at the Jülich Research Center. He received a PhD in biology from the Free University of Berlin. Contact him at m.denker@fz-juelich.de.

MARKUS DIESMANN is director of the Institute of Neuroscience and Medicine (INM-6), Computational and Systems Neuroscience, and director of the Institute for Advanced Simulation (IAS-6), Theoretical Neuroscience at the Jülich Research Center. He is also a professor of computational neuroscience at RWTH Aachen University. He received a PhD in physics from Ruhr University Bochum. Contact him at diesmann@fz-juelich.de.

DAVID D. DONOFRIO leads the Computer Architecture Group at LBNL. He received a bachelor's degree in computer engineering from Virginia Tech. Contact him at ddonofrio@lbl.gov.

LOREN M. FRANK is a professor in the Department of Physiology at UC San Francisco and an investigator at the Howard Hughes Medical Institute. He received a PhD in systems neuroscience from MIT. Contact him at loren@phy.ucsf.edu.

NARAYANAN ("BOBBY") KASTHURI is a neuroscience researcher at Argonne National Labs and an assistant professor in the Department of Neurobiology at the University of Chicago. He received a PhD in neurophysiology from Oxford University (Rhodes Scholar). Contact him at bobbykasthuri@anl.gov.

CHRISTOF KOCH is the president and chief scientific officer at the Allen Institute for Brain Science. He received a PhD in physics from the Max Planck Institute for Biological Cybernetics. Contact him at christofk@alleninstitute.org.

OLIVER RÜBEL is a research scientist in the Computational Research Division at LBNL. He received a PhD in computer science from the University of Kaiserslautern. Contact him at oruebel@lbl.gov.

HORST D. SIMON is deputy director and chief research officer at LBNL. He received a PhD in mathematics from UC Berkeley. Contact him at hdsimon@lbl.gov.

FRIEDRICH T. SOMMER is an adjunct professor at the Redwood Center for Theoretical Neuroscience and the Helen Wills Neuroscience Institute at UC Berkeley. He received a PhD in physics from the universities of Düsseldorf and Tübingen. Contact him at fsommer@berkeley.edu.

PRABHAT leads the data and analytics services group at the National Energy Research Scientific Computing Center at LBNL. He received an MS in computer science from Brown University. Contact him at prabhat@lbl.gov.

.org/abs/1607.01335.

5. B.A. Olshausen and D.J. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, vol. 381, no. 6583, 1996, pp. 607–609.
6. S. Kunkel et al., "Spiking Network Simulation Code for Petascale Computers," *Front Neuroinform*, vol. 8, 2014, p. 78.
7. J.J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proc. Nat'l Academy of Sciences USA*, vol. 79, no. 8, 1982, pp. 2554–2558.
8. E. Bethel et al., "Management, Analysis, and Visualization of Experimental and Observational Data—The Convergence of Data and Computing," *Proc. IEEE 12th Int'l Conf. e-Science*, 2016; doi: 10.1109/eScience.2016.7870902.
9. J. von Neumann, *The Computer and the Brain*, Yale Univ. Press, 1958.