High-Performance Computing in Neuroscience for Data-Driven Discovery, Integration, and Dissemination

Kristofer E. Bouchard, 1,2,3,* James B. Aimone, 4 Miyoung Chun, 5 Thomas Dean, 6 Michael Denker, 7 Markus Diesmann, 7,8,9 David D. Donofrio, 10 Loren M. Frank, 2,11,12 Narayanan Kasthuri, 13,14 Chirstof Koch, 15 Oliver Ruebel, 10 Horst D. Simon, 16 Friedrich T. Sommer, 17 and Prabhat 18,*

http://dx.doi.org/10.1016/j.neuron.2016.10.035

Opportunities offered by new neuro-technologies are threatened by lack of coherent plans to analyze, manage, and understand the data. High-performance computing will allow exploratory analysis of massive datasets stored in standardized formats, hosted in open repositories, and integrated with simulations.

Introduction

The diverse national and private neuroscience initiatives across the globe aim to accelerate our understanding of brains in a variety of ways (Grillner et al., 2016). Central to many initiatives is developing tools to investigate neuronal circuits at unprecedented resolution and scales. Much attention has been focused on the technological challenges of measuring and manipulating neural activity from large numbers of sites (e.g., neurons) for long periods. However, much less attention has been focused on the computing challenges associated with the vast amounts of data that these technologies generate. Specific and coherent plans to manage, share, analyze, and understand these data remain in a nascent stage. Thus, the opportunities offered by new neuro-technologies are threatened by the deluge of data generated.

High-performance computing (HPC) has revolutionized many scientific fields,

in particular those studying systems of many heterogeneous elements governed by complex interaction across many spatiotemporal scales: brains fit this mold well. Scientific HPC has traditionally focused on large-scale simulations, and this is also true for neuroscience (e.g., Markram et al., 2015). However, there is a paradigm shift occurring within the HPC community toward harnessing this massive computing power to processing and analyzing experimental data (Bethel et al., 2016). The neuroscience community is not alone in the challenges of utilizing HPC-other scientific fields are being rapidly transformed through the application of HPC to process and analyze ever-increasing volumes of experimental data.

The computational power, memory, and storage capabilities offered by HPC resources, while not yet suitable for many neuroscience labs, can enable the *neuroscience community* to address

several "Grand Challenge Problems." Here we focus on four topics: processing and analysis of structural connectomics data, extreme-scale statistical data analytics on massive physiology datasets (i.e., electrical and optical recording), repositories for hosting and sharing multi-modal data stored in standardized formats, and data synthesis through neural circuit simulations.

Processing and Analysis of Structural Connectomics Data

Across species, brains consist of hundreds-to-billions of individual neurons connected by thousands-to-trillions of synapses. These circuits form the structural backbone from which brain function is generated. Structural connectomics refers to neuropil reconstruction at sufficient resolution to trace neuronal processes and unambiguously identify synaptic connections. The current fastest approach for structural connectomics



¹Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

²Kavli Institute for Fundamental Neuroscience, UC San Francisco, San Francisco, CA 94158, USA

³Helen Wills Neuroscience Institute, UC Berkeley, Berkeley, CA 94720, USA

⁴Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185, USA

⁵The Kavli Foundation, Oxnard, CA 93030, USA

⁶Google Research, Mountain View, CA 94043, USA

⁷Institute of Neuroscience and Medicine (INM-6), Institute for Advanced Simulation (IAS-6), JARA BRAIN Institute, Jülich Research Centre, 52425 Jülich, Germany

⁸Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52062 Aachen, Germany

⁹Department of Physics, RWTH Aachen University, 52062 Aachen, Germany

¹⁰Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

¹¹Howard Hughes Medical Institute, UC San Francisco, San Francisco, CA 94158, USA

¹²Department of Physiology, UC San Francisco, San Francisco, CA 94158, USA

¹³Nanoscience Division, Argonne National Laboratory, Lemont, IL 60439, USA

¹⁴Department of Neurobiology, University of Chicago, Chicago, IL 60637, USA

¹⁵Allen Institute for Brain Science, Seattle, WA 98109, USA

¹⁶Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

¹⁷Redwood Center for Theoretical Neuroscience, UC Berkeley, Berkeley, CA 94720, USA

¹⁸NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

^{*}Correspondence: kebouchard@lbl.gov (K.E.B.), prabhat@lbl.gov (P.)

Neuron **NeuroView**

involves scanning multiple electron beams over a brain sample in parallel and is already producing ~50 TB/day. HPC systems are well positioned to store, host, and process these data volumes.

The major computational bottleneck in structural connectomics is automating segmentation, alignment, and annotation of the raw data (Lichtman et al., 2014). The best current algorithms are relatively slow when one considers the total amount of computation required to render any reasonably sized connectome (e.g., a cubic millimeter of tissue). The segmentation challenge is compounded by two facts: the segmentation task requires the tracing of all (or most of) the objects in the field of view; and segmentation accuracies have to be very good, as even small mistakes could result in thousands of downstream synapses being incorrectly assigned. Large-scale image/volumetric processing has been routinely conducted on HPC systems for three-dimensional (3D) and four-dimensional (4D) data generated by physics simulations (e.g., combustion, cosmology, etc.). Thus, collaboration between the communities should be able to effectively leverage existing computational tools and resources.

The result of processing structural connectomics data could be either a 3D geometric representation of each cell and its attendant processes and/or a "graph" of the resulting structural connectivity matrix with identified synapses and measures of synaptic strength. A weighted, directed graph characterizing the connection strength between neurons, along with a 3D embedding of the synapses and neurons for visualization purposes, would provide a compact summary sufficient for analyses - and is a critical step in linking "structure-to-function" in the brain. Performing exploratory analysis (e.g., "clique counting," Milo et al., 2002) on graphs consisting of 100 billion nodes (number of neurons in human brain) and one trillion edges (10% average connectivity) requires optimized methods that can exploit the structure of the graph and are accelerated by HPC resources.

Extreme-Scale Data Analytics on Massive Neurophysiology Datasets

The extraction of information from largescale neurophysiology experiments requires the application of sophisticated data analysis methods. There are computing challenges associated with data preprocessing (e.g., image processing, spike sorting) and application of statistical data analytics (SDA) to the processed data. In particular, neuroscience researchers often implicitly or explicitly interpret the output of their data analysis tools as reflecting the true state of nature. Therefore, neuroscience requires SDA algorithms that are simultaneously interpretable and predictive. By interpretable, we mean that inferred models yield insight into the (bio)-physical processes that generated the data; by predictive, we mean that the model predicts the data with high accuracy. There is often a tension between the goals of interpretability and predictability, and SDA methods that satisfy both criteria for massive datasets often require large-scale computing systems (Jordan, 2013). As there is a cost associated with all computing resources, and HPC in particular, optimized analysis codes developed by experts within the open-source community are required for efficient HPC utilization. More broadly, building and using workflows and frameworks for distributed computing, and embedding them in a collaborative setting, requires expertise well outside that of typical neuroscientists, and interactions with HPC centers could be transformative.

HPC will eventually be required to understand the dynamically modulated. spatiotemporal patterns of neural activations occurring across multiple brain areas, which ultimately give rise to sensation, action, consciousness, and behavior. New technologies enable increasingly large numbers of "brain signals" to be recorded simultaneously from diverse recording modalities (e.g., Ca2+ imaging, multi-unit silicon arrays, electrocorticography [ECoG], fMRI). Concomitantly, the duration of recordings from the same neuronal population has been steadily increasing, and it will soon be possible to record continuously for weeks at a time (generating 100 s of TBs of data). Some data processing and analyses in neurophysiology can be performed independently on individual "channels" (e.g., pixels, neurons), exhibiting what is known as "natural" or "trivial" parallelism. For example, determination of encoding models, which describe how, e.g., a single

neuron's activity is modulated by sensory inputs/motor outputs, can be computed independently of the other neurons that are recorded. In contrast, extracting structure from the concerted action of neural populations necessarily requires that calculations be performed jointly on all (or nearly all) the data at the same time, making the parallelization challenging. Therefore, it is important to develop scalable SDA methods capable of revealing structure from the large-scale neurophysiology datasets.

Two complementary approaches to understanding neural population dynamics are "dimensionality reduction" and "functional connectomics." Dimensionalityreduction methods aim to find low-dimensional spaces that concisely summarize the high-dimensional spatiotemporal patterns of neural activity; these lower-dimensional spaces can be used to gain insight into higher-order network dynamics (Cunningham and Yu, 2014). Owing primarily to its (relative) computational efficiency, the most common dimensionality-reduction method in neuroscience (and many other fields) is principal components analysis (PCA). Even for PCA, application to very large-scale datasets is non-trivial, but it has recently been performed on TB-sized matrices using HPC systems (Gittens et al., 2016).

The second complementary approach to neural-population dynamics is functional connectomics, which involves using temporally directed analysis methods to infer "causal" influences among spatially distributed neural recordings (Brovelli et al., 2004). The resulting functional connectomes explicitly represent the time-varying, causal interactions between individual physical recordings and are thus potentially more interpretable than results from methods like PCA. Furthermore, functional connectomes are naturally represented as graphs, and so may be mapped to structural connectomes to enhance understanding of "structure-function" relationships. However, even using linear techniques (e.g., vector autoregressive models), the extraction of functional connectomes from large-scale neurophysiology datasets will require HPC systems running highly optimized implementations that estimate sparse connectivity in an unbiased fashion.

These examples illustrate that performing even standard analysis on large-scale neurophysiology datasets will benefit from HPC systems. Continued investigations into interpretable statistical data analysis coupled with implementations on high-performance computing systems (e.g., to distribute optimization calculations) are required to address these challenges and to ensure that extraction of useful information from neuroscience data is unhindered by computing bottlenecks.

Repositories for Hosting and Sharing Multi-modal Data Stored in **Standardized Formats**

Neuroscience is a diverse field producing a wide variety of datasets. It is essential share data and code through standardized and extensible data models and management solutions (e.g., Garcia et al., 2014; http://nwb.org). The data model requirements in neuroscience are primarily driven by the data generators (e.g., experimenters) and the data consumers (e.g., data analysts). At a high level, data-generator requirements relevant for HPC include: fast data-write capabilities and efficient storage of large data volumes: common data standards to enhance portability and usability for efficient storage; and capacity for collection of metadata with the raw data for reproducibility and interpretation. Dataconsumer requirements include: the need for fast data read for efficient data analysis; semantic data access through data query, annotations, and relationships; and integration of distributed, multi-modal data sources. This last point-integration of multi-modal datais particularly salient for understanding neurophysiology recordings because, as recently noted, it is insufficient to record increasing numbers of neurons without simultaneously monitoring behavior during increasingly complex sensory, motor, and cognitive tasks (Gao and Ganguli, 2015). The increased complexity of naturalistic, real-world stimuli and behaviors brings with it challenges of integrating such multi-modal data and metadata (e.g., visual, audio, haptic, and movement) with the brain data.

The single greatest impediment to fully extracting the return on investment into neuroscience data collection is the lack of community alignment and coordination around standards for experiments, data, and metadata. Sharing and reuse of all data will enable validation of experiments and enhance reliable scientific interpretation. Ultimately, data analysis results in commodities that become shared and reused. The path to understanding often involves a complex range of processing steps requiring different areas of expertise. Similar to the role of metadata for raw measurements, data provenance (including the denotation of methods, parameters, etc.) is required for reliable interpretation of analysis. To maximize the utility of HPC for neuroscience data. ensuring "machine readability" of metadata is an important consideration that requires knowledge and expertise from experimentalists, data model designers, and data analysts.

The close integration of computing resources with data repositories (e.g., http://crcns.org) will enable effective data-driven discovery. This includes integration of hardware resources for efficient processing, reduction of costs for large data transfers, and management/analysis software stacks to enable large-scale analysis. In addition to the challenges associated with storing large datasets from a single individual (i.e., a single brain or single experimental subject), persistent access to a large number of datasets across individuals drastically increases data volumes. Critically, combining across individuals will reveal universal design features of a species' brain and will constrain models. Perhaps as important, comparing across individuals could expose "quirks" that make each individual unique and thus provide insight into resilience/susceptibility to failure modes (e.g., Alzheimer's disease, environmental challenges, etc.). Meeting all these needs requires advanced, high-performance data and computing infrastructure that HPC centers are ideally positioned to provide.

Data Synthesis through Neural Circuit Simulations

Ultimately, the goal of neuroscience is to achieve a deeper, broader understanding of brains in health and disease that extends across vast spatial and temporal scales. Many domains utilize simulation where extensive experimentation

is intractable due to either cost or feasibility of data acquisition. As other fields have demonstrated, it can be extremely productive to have computational/theoretical frameworks that permeate research directions and guide the scaling-up of data acquisition and analysis. By following a similar pattern, neuroscience can reduce the challenge of integrating information from very different perspectives of the brain. While a "theory of brain" seems a distant goal, many research domains focused on highly complex, multi-scale problems (e.g., climate, high-energy physics, cosmology) have effectively leveraged HPC capabilities to integrate data and analysis into theoretical frameworks through the use of simulations (Bauer et al., 2015).

Full-scale simulations of neurons with all their synapses on supercomputers have recently enabled several research groups to construct anatomically detailed models of local cortical networks (e.g., Markram et al., 2015). Modeling approaches can 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, and more effort in this direction is required.

The challenge still is to scale up networks to the size of the full brain (Kunkel et al., 2014), and further research may allow use of the upcoming exascale supercomputers for this purpose. Another challenge facing large-scale neural simulations is understanding how to evaluate the quality of results from models constrained by relatively limited amounts of experimental data, though some efforts in this domain exist (Hawrylycz et al., 2016). Conversely, it should be possible to examine the accuracy of data analysis algorithms (e.g., spike sorting, functional connectivity estimation) on data from simulations, in which the ground truth is known. Techniques for understanding the uncertainty of model outputs and overall sensitivity of models relative to input parameterization is an area that needs to be further explored and will greatly increase computational demand.

Neuron **NeuroView**

Box 1. Summary and Recommendations

- The computational power, memory, and storage capabilities offered by HPC resources will enable the neuroscience community to address several "Grand Challenge Problems." HPC needs to be more thoroughly integrated into neuroscience research strategies.
- Persistent storage of large-scale neuroscience datasets from multiple brains in massive repositories is required to extract universal design principles and identify unique differences across individuals. HPC systems are uniquely positioned to enable this capability and should be utilized.
- Leveraging HPC resources for neuroscience "Grand Challenge Problems" requires significant investment in standardization of data and metadata as well as development and optimization of data preprocessing and analysis codes, as well as enabling data analysis frameworks (e.g., Spark) on HPC systems.
- Activities focused on enabling "Cloud" computing for individual labs (e.g., "International Brain Station") are important endeavors but do not address the challenges for which HPC will be required, and both should be pursued in parallel.
- International research and funding agreements, such as occurs in high-energy physics, should ensure coordination across different national and private neuroscience initiatives.

Community reliance on simulation codes requires the treatment of these codes as complex scientific infrastructure that need to be maintained. HPC will allow continued scaling (both in number and accuracy) of network simulations, as well as tighter integration with experimental data and analysis, which is required to generate specific predictions and hypotheses across spatiotemporal scales.

Conclusion

While the computing challenges for different areas of neuroscience require somewhat non-overlapping sets of expertise, it is not enough to embark upon these enterprises independently (Box 1). To fully achieve the goals of the many neuroscience initiatives, data standards and analyses need to be coordinated, and the derived products from data analysis and simulations need to be synthesized into coherent, self-consistent descriptions of the brain. This endeavor has its own set of computing requirements: allowing exploratory analysis across the (international) neuroscience community will require major improvements to reproducibility of analysis workflows; massive data "meta-repositories" to persistently host the many repositories collecting physiological data from many experiments and individuals; that data will need to be stored in common formats and collocated with HPC hardware; and it must be analyzed using open-source, optimizedanalysis tools guided by, and integrated with, brain-scale simulations. The insights generated from this endeavor will have high-payoff outcomes: they will greatly support neuroscience efforts to reveal both universal design features of a species' brain, as well as understand what makes each individual unique—a central concept of precision medicine. Activities focused on enabling "Cloud" computing for individual labs (e.g., "International Brain Station") are important endeavors but do not address the challenges for which HPC will be required, and both should be pursued in parallel. Harnessing the power of HPC resources will require neuroscientists to work closely with computer scientists and will take time, so we recommend rapid and sustained investment in this important endeavor now.

ABOUT THE AUTHORS

This NeuroView article was written by an international, interdisciplinary group with expertise spanning applied mathematics, computer science, physics, and experimental and computational neuroscience. The group is comprised of members and leaders of both Department of Energy National Labs and U.S./European academic institutes, as well as private research and funding organizations.

REFERENCES

Bauer, P., Thorpe, A., and Brunet, G. (2015). Nature 525, 47-55.

Bethel, E.W., Greenwald, M., van Dam, K.K., Parashar, M., Wild, S.M., and Wiley, H.S. (2016). DoE ASCR Workshop Report. http://www.mcs.anl. gov/papers/P6017-0616.pdf.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S.L. (2004). Proc. Natl. Acad. Sci. USA 101, 9849-9854

Cunningham, J.P., and Yu, B.M. (2014). Nat. Neurosci. 17. 1500-1509.

Gao, P., and Ganguli, S. (2015). Curr. Opin. Neurobiol. 32, 148-155.

Garcia, S., Guarino, D., Jaillet, F., Jennings, T., Pröpper, R., Rautenberg, P.L., Rodgers, C.C., Sobolev, A., Wachtler, T., Yger, P., and Davison, A.P. (2014). Front. Neuroinform. 8, 10.

Gittens, A., Devarakonda, A., Racah, E., Ringenburg, M., Gerhardt, L., Kottalam, J., Liu, J., Maschhoff, K., Canon, S., Chhugani, J., et al. (2016). arXiv, arXiv:1607.01335, https://arxiv.org/ abs/1607.01335.

Grillner, S., Ip, N., Koch, C., Koroshetz, W., Okano, H., Polachek, M., Poo, M.-M., and Sejnowski, T.J. (2016). Nat. Neurosci. 19, 1118-1122.

Hawrylycz, M., Anastassiou, C., Arkhipov, A., Berg, J., Buice, M., Cain, N., Gouwens, N.W., Gratiy, S., Iyer, R., Lee, J.H., et al.; MindScope (2016). Proc. Natl. Acad. Sci. USA 113, 7337-7344.

Jordan, M.I. (2013). Frontiers in Massive Data Analysis (National Academies Press).

Kunkel, S., Schmidt, M., Eppler, J.M., Plesser, H.E., Masumoto, G., Igarashi, J., Ishii, S., Fukai, T., Morrison, A., Diesmann, M., and Helias, M. (2014). Front. Neuroinform. 8, 78.

Lichtman, J.W., Pfister, H., and Shavit, N. (2014). Nat. Neurosci. 17, 1448-1454.

Markram, H., Muller, E., Ramaswamy, S., Reimann, M.W., Abdellah, M., Sanchez, C.A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., et al. (2015). Cell 163, 456-492.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Science 298,