McGill submission to Canada's Call for White Papers for Sustainable Planning for Advanced Research Computing Phase II (SPARC2)

Authors

This submission was prepared with input from members of the McGill community representing different areas and whose research relies heavily on the Compute Canada network. This document is approved by the Vice-Principal, Research and International Relations.

- Nikolas Provatas Professor, Canada Research Chair Department of Physics and Director of McGill Compute Canada HPC site
- Alan Evans, Professor of Neurology and Neurosurgery, Medical Physics, and Biomedical Engineering, Director of Ludmer Centre for Neuroinformatics and Mental Health
- Guillaume Bourque, Associate Professor, Department of Human Genetics and Director of McGill University Genome Quebec Innovation Centre Bioinformatics Platform (C3G)
- Kristina Ohrvall, Director Strategic initiatives, Office of the Vice-Principal, Research and International Relations

Context

McGill has a longstanding history of supporting computationally based science and engineering in a wide range of research fields, including traditional areas such as physics, materials science, aerospace engineering, and computer science, and more recently in emerging fields such as digital humanities and medical sciences, the latter of which is quickly expanding to be among Calcul Quebec's largest users of computing capacity.

As a leading user of High Performance Computing (HPC) in Canada and founding member of Calcul Quebec, McGill has a significant vested interest in the future of the Compute Canada platform. In order to meet the growing needs of our research community for advanced computing and the growing need for 'big data' approaches, McGill wholeheartedly supports the development of a sustainable and efficient pan-Canadian CC network. As such, McGill continues to play a leading role in the management of the Compute Canada platform and contributes financially to its operations. Furthermore, a proposal to house the next generation ARC system for investment under the CFI Challenge 2 Stage 2 Cyberinfrastructure Initiative is in preparation and the proposed single site solution for Quebec is the existing McGill HPC site housed at ETS. If selected, this site will be co-managed on behalf of Compute Canada by all member institutions of Calcul Quebec. Through McGill's involvement in shaping the future of CQ, we will ensure capacity to continue support to researchers for advanced computing across all disciplines.

Science Description: Current Use of Advanced Research Computing at McGill

McGill researchers in physics, advanced materials, engineering, bioinformatics, and neuro-informatics are already large-scale users of the Compute Canada platform (around 10% of the national usage) and we foresee continued growth across the fields as described below:

Physics and Engineering: McGill has been and continues to be a world-leader in physics. As such, it has the distinction of being one of a select number of international universities involved in the world-renowned high-energy physics ATLAS and Belle 2 grid projects. The McGill HPC Centre is one of four sites in Canada that receives and manages data from the ATLAS project and for researchers across the country. Notably, the McGill grid site hosts disk space for the ATLAS Higgs analysis group. This is a clear success story for our ability to support "Data Intensive Science". It is expected that a new Calcul Quebec site would continue to support ATLAS activities. Other HPC-intensive projects that are ramping up include a new proposed Materials Informatics training platform, partnership with the US-lead Materials Genome Initiative (MGI), as well as collaborative research between École Polytechnique, McGill and Bombardier. Taking a leadership role in the maintenance of a dedicated Compute Canada facility and maintaining technical HPC expertise is *crucial* to the support of McGill participation in these prestigious global scientific activities.

Neuroinformatics: Linked to the greatest advances and most important opportunities in neuroscience today, neuroinformatics unites neuroscientists, mathematicians, computer experts, and physicists who deploy enormous processing power to build new knowledge and discern patterns about the brain. Their raw materials are newly emerging tsunamis of data derived from mental health animal models, brain imaging, genetic, epigenetic, demographic, social and cultural, clinical, environmental, and other sources. New technologies for data storage allow high rate data capture for neuroimaging, genetic, epigenetic, various 'omics, and behavioural studies that are transforming neuroscience at a rapid pace. A single high-resolution dataset for a full brain now requires over 200TB of disk space for raw data only, whereas five years ago the most state of the art dataset was 1TB. Despite the staggering volume and complexity, these data are essential to advancing knowledge of how the brain operates. As an early adopter of neuroinformatics, McGill has amassed global expertise and infrastructure. McGill is home to CBRAIN (Canadian Brain Research and Informatics Network), and LORIS (Longitudinal Online Research and Imaging System), two advanced platforms for high-capacity computing and web-based data management which together support more than 400 users around the world.

Bioinformatics for Genomics: With projects such as the Genetics and Genomics Analysis Platform (GenAP) and C3G – a Genome Canada funded national platform – genomics researchers are major users of the Compute Canada platform. Over the next decade, almost every biomedical investigation in basic and clinical research will be enabled through characterization of an accompanying genome sequence. With next-generation sequencing technologies revolutionizing the life sciences data processing and interpretation, rather than data production, has become the

major rate-limiting step toward discovery of new therapies. The 2015 CFI award to acquire this advanced sequencing technology at McGill and two other Genome Centres, brings an immediate urgency to the need for advanced research computing resources for the genomics community. In addition, collaboration between Quebec and the Ontario HPC4Health initiative to better integrate and share genomics data further necessitates the creation of a major ARC system to serve the Quebec-Ontario corridor in the area of HPC for health sciences. The joint submission of a white paper from the four largest Genome Centers in Canada further elaborates the need for advanced computing resources for the genomics research community.

Overview of future usage profile

Research area	**Future needs** (2017-2022)
Materials physics and Engineering	The rise of integrated computational materials design (ICME) will impose stringent requirements on processing speed and data storage. Simulations to model the microstructure of single IC circuit interconnect sample (microelectronics), or one alloy powder sample (additive manufacturing) requires no less than 512-1024 tightly coupled cores running continuously for 3-5 days. Data generated from each sample run in 3D is no less than 5-10TB. One typical project requires running 10 runs per sample (for statistics) and ~10 X10 different sample (process parameters). This brings the total to about 5-10 PB per project in the domain of materials design. We can expect similar storage and cycles resource needs for aerospace simulation work.
Neuroinformatics	CBRAIN will to continue to serve the neuroscience community in Canada. A "Big Data" approach is required for further development and to expand the user base to other fields of research. Cloud access is desirable. CBRAIN will benefit from many of the common services including data transfer, authentication and ID management, resource monitoring, etc. A major increase of storage capacity is expected: • 2016: 1,200 CPU core years and 1.6PB disk storage • 2020: 15,000 CPU core years and 25PB disk storage
Bioinformatics for Genomics	Computing resources will have to double every 2 years; online disk and tape backup resources will have to triple. • Current (2016): 2,604 cores; 3 PB of high-performance online storage and 6 PB of tape storage. • Forecasted needs (2022): 20,000 cores; 81 PB of high-performance online storage and 162 PB of tape storage.