The Computation and Data Needs of Canadian Astronomy

The Computation and Data Committee

Summary

In this white paper, we review the role of computing in astronomy and astrophysics and present the Computation and Data Committee's assessment of future priorities for the community. Astronomers make disproportionately heavy use of the national computing resources provided primarily through Compute Canada (CC) and the Canadian Astronomy Data Centre (CADC). Based on a survey of research astronomers and usage statistics from CC and CADC, we estimate that meeting research needs over the next five years will require at least an order of magnitude increase in processing power and a factor of 30 increase in storage. CASCA members ranked the need for a high-performance computing refresh at the top of their research priorities followed closely by a new national facility dedicated to data-intensive computing. In addition to these research needs, astronomers will have to move their data archives from National Research Council facilities to an alternative facility.

The Astronomical Use-Cases

In addition to standard desktop computing, we review four domains in which national commitments to computing are required: High-Performance Computing, Astronomical Software Development, Data Archiving and Access, and Data-Intensive Computing. We include user requirements estimated through usage statistics from Compute Canada and the Canadian Astronomical Data Centre. We also directly polled the CASCA membership in July 2014 to get direct input for current use and forecasts.

High-Performance Computing

High-Performance Computing (HPC) represents a classic use case for Compute Canada (CC) resources.[1] Numerical simulations of astrophysical phenomena are a natural fit for HPC resources. More recently, observational astronomy has taken advantage of HPC resources for processing and calibration tasks.

Since most HPC resources in Canada are managed through CC, the statistics on of CC usage are representative of how astronomers use high-performance computing resources. Astronomers make significant use of multi-core systems, with the majority of compute time being spent on jobs with 64 to 512 computer cores, though some jobs require in excess of 10^4 cores. Over the past two years, astronomers ranked 6th across disciplines in terms of overall CC processing usage, accounting for 4.7% of total usage though the ~350 Canadian astronomers comprise only 0.5% of Canadian scientific research community.

Data Archiving and Access

[1] Compute Canada (CC) is an umbrella organization that coordinates and manages regional high performance computing (HPC) consortia such as WestGrid (BC, AB, SK, MB), Compute Ontario (ON), etc.

Astronomical research drives three requirements from our data archives: (1) long term maintenance of our observational legacy, (2) flexible access to data, most notably the ability to extract subsets of archives according to specific search criteria, and (3) the network capacity to transport data between hosting and processing resources.

The Canadian Astronomical Data Centre (CADC), part of the National Research Council Canada (NRC), has long been an international leader in the storage and curation of astronomical data. In polling CASCA members, 45% of the respondents indicated they have used CADC archival resources in the past 12 months suggesting around 150 Canadian users annually. This number is dwarfed by the number of international users. Based on download statistics, CADC estimates there are nearly 6000 international users of the CADC archives, representing ~60% of the entire global community. CADC hosts 700 TB of astronomical data on NRC-supported resources and currently maintains a large fraction of the archive as a copy on Westgrid resources.

Simulation data for astrophysical systems also have significant data impact. Theoretical studies must store their outputs as code save-points and post-processing products for subsequent analysis. On Westgrid alone, there are currently 1.2 PB of astronomical data including storage for users creating simulations, the CADC holdings, and the CADC's VOSpace system (a cloud storage platform tailored for astronomical use). Given Westgrid usage statistics, we estimate roughly 30% of this fraction is for simulations. The fraction will be substantially higher for other Compute Canada consortia since observational use of CC is concentrated on Westgrid.

Astronomical Software and Platform Development

Large astronomical software analysis falls into three major categories, namely simulation (numerical codes for theoretical analysis), calibration (processing/reducing observational data) and analysis (including data mining, modelling, and informatics). Code development represents a significant fraction of actual research productivity in astronomy, though this fraction varies from researcher to researcher and from project to project.

Much of the research code base in astronomy is closed source, being limited to individual researchers or their research groups. In LRP2010, the community identified the need to develop robust, open-source "community codes." Robust community codes would be better validated than closed-source codes, enable reproducibility, and minimize the inefficiency of having several groups develop similar tools. Since LRP2010, collaborative development tools (notably, github and bitbucket) have enabled several open-source community-driven software projects (ENZO, astropy). However, these success stories have, at their core, dedicated personnel whose role it is to shepherd the project. To ensure Canadian research needs are met in the future, the community requires dedicated effort to support the development of these kinds of tools.

Canadian astronomers are also leading large software development projects. For example, the Canadian Advanced Network For Astronomical Research (CANFAR) is deploying a platform for managing virtual machines and cloud storage in a form that is tailored to the

astronomical use case. With an eye to the future, developers are creating the CyberSKA platform, a platform for interacting with the immense data sets that will be created by the Square Kilometre Array without transporting those data to the user.

Data-Intensive Computing

Historically, Compute Canada infrastructure has been used for theoretical work, but there is growing usage for data-intensive computing. With large data sets becoming increasingly common, the ability to support their transport and storage by individual researchers is shrinking. This includes data generated by simulations. Frequently, analysis must occur on dedicated hardware. For example, image processing from the SCUBA2 instrument on the James Clerk Maxwell Telescope requires large amounts of memory and is currently limited by access to 512 GB memory nodes available through CANFAR and Compute Canada resources. Astronomers are also making use of advanced algorithmic methods for analyzing large data sets. The growing domain of astroinformatics and astrostatistics uses Big Data processing methods for astronomy. Applications of Machine Learning techniques to astronomical data sets remain promising but require large amounts of processing when applied to large data volumes.

Astronomy and Compute Canada

This management body provides the primary channel for interaction with CFI regarding the purchase and maintenance of HPC resources nationally. CC has recently undergone significant management reorganization which the Computation and Data Committee has monitored closely. Despite some concerns in the process, the new management has emerged with a significant role for scientific input and appears to be acting as a good steward for computing in Canada.

As astronomical research becomes ever more closely tied to computing, our users are taking advantage of national computing resources. In our survey of CASCA members, 60% of users responded that they currently use CC resources and 65% reported that they are “Likely” or “Certain” users of CC facilities over the next 5 years. However, several respondents indicated that a major impediment to leveraging national resources has been the adaptation of codes to specific resources. We also solicited opinions about different computing providers and respondents had strong positive opinions about their regional computing consortia, significantly better than their opinions of university-based computing providers. In narrative feedback, repeated user comments emerged in three areas: (1) the comparatively low performance compared to international HPC resources, (2) a need for better management of existing resources, in particular tailoring job submission to machine architecture and enabling multi-year allocations for storage, and (3) better support for developing new computer codes.

Future Needs and Forecasting

In forecasting for this whitepaper, we assessed (1) the data impact of future telescope facilities, (2) community priorities that bear directly on new computing infrastructure, and (3) clear risks to the field that need to be addressed in the next 5 years.

Data Impact: The data impact of telescopes varies widely. The Thirty Metre Telescope (TMT) is the top priority for ground-based astronomy identified in LRP2010. However, the archival data impact for the TMT is relatively modest for many use-cases (<1 TB/day). In contrast, the archive data rate of the Square Kilometre Array Phase I, the second ranked priority, is 8 PB / day. In our poll to our membership, we asked how likely they were to be users of the various LRP-identified priorities facilities. We extrapolated expected use of the different respondents to the scale of the full CASCA membership. We also categorized the facilities into Low, Medium and High data impact based on how disruptive the data management strategy would be, given current facilities. The response shows that a significant number of users are anticipated for the highest impact data facilities but the Low impact facilities will see the largest number of users.

Data Impact	Example Facilities	Expected Users
High	Square Kilometre Array, Jansky Very Large Array, Atacama Large Millimetre/submillimetre Array	75
Medium	Canada France Hawaii Telescope, James Webb Space Telescope, Dark Energy Satellite Mission	80
Low	Thirty Metre Telescope, Gemini	150

We also asked users to anticipate the factor by which their processing, storage and memory requirements would change over the next 5 years. The geometric mean across responses suggests that the typical astronomical user anticipates a factor of ~3 increase in these domains. However, the upper end of HPC users as indicated by their past HPC allocations indicated significantly larger growth in anticipated needs and we estimate the overall growth in processing requirements for the community as a factor of ~10. We asked users to anticipate their storage needs required to support their research. Taking these numbers as representative of the needs of the community as a whole, we estimate that the field will require 30-100 PB of storage for processed data, a figure which neglects the volume of raw data to be reduced or any margin for backup and redundancy.

New Facilities: We solicited community for feedback about how to prioritize facilities that support computing in astronomy in the future. We asked users to prioritize three different classes of facilities for national computing infrastructure. Ranked in order of the number of respondents who selected these facilities as a top priority:

i. (54%) *New High Performance Computing Facility* -- A new national supercomputer capable of performing heavily parallelized jobs requiring many nodes. Such a facility would rank highly on, e.g., the Top500 Supercomputer List and be used for large simulations and large-scale analysis (e.g. CMB).
ii. (33%) *New Data Intensive Computing Facility* -- A national computing resource involving cluster computing connected to a large volume of high performance disk storage. Such a facility would be used for data processing and mining.

iii. (13%) *Next Generation Software* -- Funding allocated to developing data and processing management across existing computational resources, including options like (a) enabling GPU processing of astronomical data, (b) creation of flexible reduction pipelines, (c) developing better simulation codes and code frameworks.

Identified Risks: Summarizing the above forecasting and collecting data suggests there are specific risks that national computing infrastructure will need to minimize.

i. *Long-term Support for CADC Archiving and Processing Facilities* -- NRC has indicated that the Canadian astronomical data archives should move off NRC facilities onto national or university computer infrastructure in the near term. The current Compute Canada model cannot support the community's long term archiving needs.

ii. *Need for an HPC Refresh* -- Canadian HPC resources remain poor compared to other nations, even when normalized by size of research community. While many users are able to take advantage of the relatively small resources, the upper envelope of HPC users is scientifically limited by the available HPC machines.

iii. *Need to Analyze a Growing Data Volume* -- A significant fraction of the Canadian astronomical community will require access to both large data archives and the ability to host those large data sets with fast connections to processing.

Of particular note with respect to these risks, CFI has recently announced a new Cyberinfrastructure Fund with Compute Canada putting forward a Sustainable Planning for Advanced Research Computing (SPARC) process to meet the call. This call may directly address all three of these risks and the Mid-term Review should facilitate a community-wide response to this call.

Acknowledgements

The report made extensive use of material presented in the white paper "Astronomy and Astrophysics Research Computing Needs: Present and Future" by the CDC submitted to Compute Canada in 2013. We are also grateful to Rob Simmonds of Compute Canada and to David Schade of the CADC for providing information regarding usage of their respective organization's resources. Finally, we are grateful to the CASCA membership for taking the time to furnish detailed responses to our survey.