The Computation and Data Needs of Canadian Astronomy

The Computation and Data Committee
on behalf of
The Canadian Astronomical Society

Summary

In this SPARC2 white paper, we review the role of computing in astronomy and astrophysics from a current and forward-looking perspective. This whitepaper largely repeats the content presented in the original SPARC consultation; however, a brief update reflecting the 2016 landscape is given at the end of the document.

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 CASCA members 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. Astronomical users 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. Preserving our rich astronomical data legacy under this transition will be essential.

Introduction

The field of Astronomy is defined by the classes of objects under study rather than by a specific set of methods. Until recently, astronomy has been indistinguishable from astrophysics, namely the application of physics to astronomical systems. However, with modern observational advances discovering a rich chemistry at play in space and the discovery of exoplanets with the possibility of hosting life, the fields of astrochemistry and astrobiology are quickly developing. Because most objects of interest are inaccessible to experiment and rarely available for direct *in situ* study, the field is driven forward by the ongoing dialogue between observations and theory. Both sides of this scientific dialogue are driving rapidly expanding requirements in computation, data storage, and network connectivity.

In this white paper, we summarize our changing needs and make our best forecast for the immediate future. This paper represents the Canadian Astronomical Society / Société Canadienne d'Astronomie (CASCA), which is the national society representing Canadian astronomy and astrophysics researchers. Since 2000, CASCA has played a leading science policy and planning role for the discipline, issuing a decadal Long Range Plan (LRP) after a year-long process supported by NSERC, NRC, CSA, and CFI and involving nationwide

consultation, proposals, town halls, and written submissions. In concert with this process, external reviewers identified Canadian astronomy as one of the highest-impact disciplines within the entire Canadian research endeavour. Ranked in terms of a normalized literature impact factor, Canadian Space Science (including Astronomy) was top ranked compared to other nations (France, UK, US, Germany, Italy, Japan). This distinction was shared with Clinical Medicine.[1] Part of Canadian Astronomy's disproportionate impact in the international field can be attributed to careful stewardship of our national data and computational resources.[2]

The Astronomical Use-Cases

In addition to standard desktop computing, astronomy establishes four domains in which national commitments to computing are required. These are

- High-Performance Computing
- Astronomical Software Development
- Data Archiving and Access
- Data-Intensive Computing

Each of these domains projects separate requirements onto processing capacity, software, storage, and network requirements. We have tried to capture the requirements by requesting usage statistics from Compute Canada and the Canadian Astronomical Data Centre. We have also polled our 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 resources. In astronomy, HPC requirements are largely driven by the theoretical side of the field, where a physical model of a system (a star, a galaxy or the Universe itself) is numerically simulated. Computation is one of the only means astronomers have of accessing the time domain since the natural timescales on which many astronomical phenomena play out dwarf human lifetimes. Such simulations are a natural fit for HPC resources since different physical regions can be distributed to separate cores, providing an obvious channel for parallelizing the simulation. More recently, observational astronomy has taken advantage has taken advantage of HPC resources for jobs that can be divided trivially into separate tasks (i.e., the processing thousands of independent images using the same algorithm).

Since most HPC resources in Canada are managed through Compute Canada (CC), the statistics on of CC usage are representative of how astronomers use high-performance computing resources. Figure 1 below shows the distribution of usage as a function of the number of cores. Most of astronomical computing on CC resources uses 1 core or a moderate

[1] For the years 2005-2009. From "The Global Citation Race", *Inside Higher ED*. June 10, 2010.
[2] "Astronomy in Canada," a policy report by HAL Innovation Policy Economics, 2011

number (64 to 512) cores, though astronomy uses nearly half of all time in the 8192-16383 core bin.

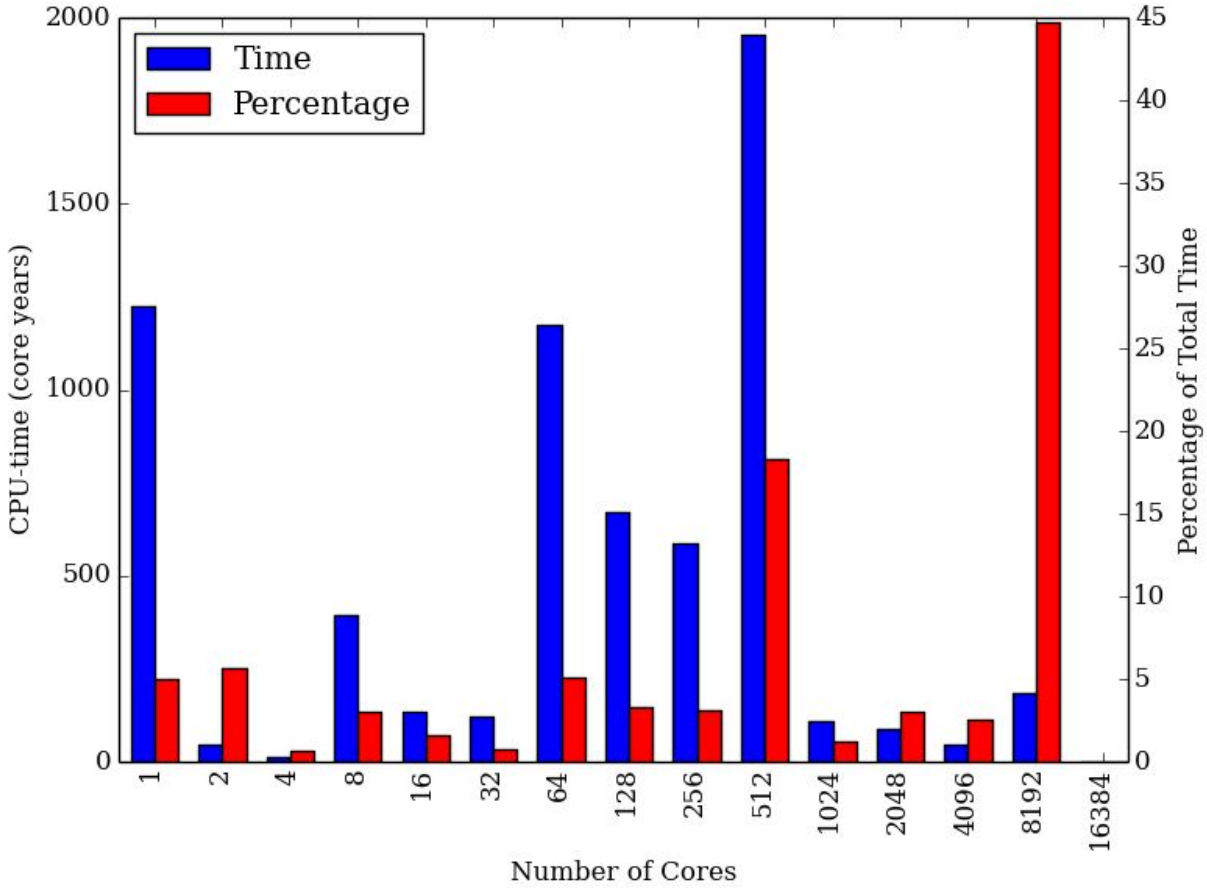

Figure 1: Distribution of CPU-time used on Compute Canada resources for Astronomy/ Astrophysics research. The left axis (blue bars) shows total compute time used by astronomical resources as a function of the number of cores used in each job. Bins include all jobs sizes up to the bottom of the next higher bin. The right axis (red bars) shows the fraction of time used for jobs of that size across all disciplines. Data represents usage over June 2012 to June 2014.

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. This usage fraction may be an underestimate as some astronomers report their discipline as “Physics” for Compute Canada self-identification purposes.

Some HPC users can make use of hardware acceleration components such as GPUs, though this requires the development of specific software to use these accelerators. Some problems in astrophysical analysis can be recast into a form amenable to GPU use but a large fraction of the domain specific software cannot use this acceleration. Smaller groups do not always have the

resources to adapt their software as coding overheads are much greater than for CPU based development. In theoretical astrophysics, GPU applications to date are typically for individual algorithms that represent a small part of the physics required for modern simulations. Internationally there are no complex simulation codes that are fully ported to GPUs. In this sense GPUs are an unproven technology for general HPC use. In our poll of the CASCA membership, 34% of the respondents indicated that they have made some use of GPU acceleration for research purposes. Of this fraction, 60% made use of Compute Canada resources directly. While GPUs are essential for a small number of researchers, many Canadian astronomers rely exclusively on CPU processing.

Data Archiving and Access

Telescopes from across the world return their data to Canadian resources for storage, processing and long-term archiving. These telescopes survey the Universe in light with a wide range of wavelengths, provide complementary views on astronomical objects. Astronomers rely on the ability to integrate the data from across the electromagnetic spectrum. Ready access to archival data is a necessity for the field, particularly through flexible access to those resources.

The Canadian Astronomical Data Centre (CADC), part of NRC-Herzberg, has long been an international leader in the storage and curation of astronomical data. The CADC archives and data from Canadian national telescopes like the Canada-France-Hawaii Telescope as well as providing archive support and supplementary data processing for other international telescopes like the Hubble Space Telescope and the Atacama Large Millimetre/submillimetre array. 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 Compute Canada resources (silo on Westgrid).

Simulation data for astrophysical systems also have significant data impact. Theoretical studies must store their outputs as as code save-points and post-processing products for subsequent analysis. On Westgrid resources 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 the inputs, we estimate roughly 30% of this fraction is for simulation output. The fraction will be substantially higher on other consortia since CADC's activities are concentrated on Westgrid resources.

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. In particular, the flexible access requires good curation of the data and establishing rigorous metadata standards. CADC has

been an international leader in establishing these standards and is one of the leaders in furnishing flexible access through their own access methods as well as through the International Virtual Observatory standards.

Astronomical Software and Platform Development

Both simulations and observations require specialized software to cope with the specific needs of the discipline. Numerical simulations require domain-specific codes that incorporate the physics applied to the evolution of a given system. Observational astrophysics frequently makes use of computationally intensive software pipelines to provide optimally calibrated data. The scope of both of these data types dwarfs the ability for human analysis, and astronomers must employ analysis codes to reduce the data volume into a summary form. These three classes of codes (simulation, calibration and analysis) all require significant development effort from astronomers. Code development represents a significant fraction of actual research productivity in astronomy, though this fraction will vary 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 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 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. Commercial cloud providers are cost-prohibitive when meeting large data or large memory jobs. Such providers are unable to provide support in creating an environment suitable for astronomical development. Thus, these tools are being deployed on CC resources. Community developers also contribute to telescope projects, contributing code to large international projects such as the Atacama Large Millimetre/submillimetre Array. 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.

As astronomical research becomes ever closer 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. For example, compilation of complex numerical codes under different architectures and shared library versions was listed as a deterrent in our survey. Similarly, the adaptation of existing codes to hardware accelerators requires substantial effort and was identified as an area for the improvement of CC services.

Data-Intensive Computing

While most processing on national computing infrastructure has been used for theoretical work, 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. Frequently, analysis must occur on dedicated hardware. As an example, users of the Jansky Very Large Array commonly produce data sets with sizes of several 10s of TB. The observatory furnishes a high-performance storage array connected to a data processing cluster for the calibration of telescope data but no similar facility exists for the analysis of those data. 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 set. 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 big data volumes.

Astronomy and Compute Canada

As discussed above, astronomers make use of Compute Canada resources for both high-performance computing for simulations and for the storage and processing of observational data. While the usage statistics are summarized previously, we also solicited satisfaction with different providers of computer support of a 5-point Likert Scale (“Very Dissatisfied”=1 to “Netural” =3 to “Very Satisfied”=5). The respondents to our survey were most positive about their interactions with their regional computing consortia (e.g., WestGrid, SHARCNET, etc.; mean of 4.25/5), positive about their interactions with Compute Canada (mean = 4/5) and less positive about their interactions with their University or Department IT (mean = 3.5/5).

We also solicited feedback about the satisfaction Compute Canada resource allocation process and whether it meets the research needs of groups. Overall there was a neutral reaction (3.1 / 5) but some concern when asked whether the allocation process would be able to meet future needs (2.8 / 5).

In narrative feedback, repeated user comments emerged in three areas: (1) the comparatively low performance compared to international HPC resources, (2) 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 deploying new computer codes.

A significant interface between the astronomical community and CC is through the CANFAR project. While the project has seen many successes enabled by this collaboration, CANFAR management noted that process of development was limited by downtime of CC resources, slow administrator responses, and slow deployment of approved allocations.

Future Needs and Forecasting

As part of its activities, CASCA develops a Long Range Plan (LRP) for astronomy, outlining the consensus priorities for telescopes and other facilities for the national community. This planning exercise takes place every 10 years with a Mid-Term Review occurring halfway through. The current LRP covers the period 2010-2020 and, as of this writing, we are preparing for the Mid-Term Review. 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 retired in the next 5 years.

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. The SKA will become operational in the beginning of the 2020s. While the CADC will not be responsible for archiving the full data stream, Canadian astronomical users will want to process and analyze these data sets. In our poll to our membership, we asked the how likely they were to be users of the various LRP-identified priorities facilities. We scaled the different categories to a fractional number of users ("Unlikely" = 0, "Potential User" = 0.25, "Likely User" = 0.5, "Certain User" = 1) and scaled the number of respondents to the size of the full CASCA membership. We also categorized the facilities into Low, Medium and High data impact based qualitatively 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 Number of 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 significant larger growth in anticipated needs and we estimate the overall growth in processing requirements for the community as a factor of ~10.

When asked to anticipate their storage needs required to support their research (without redundancy), we received the distribution of answers indicated shown in Figure 2 below. 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.

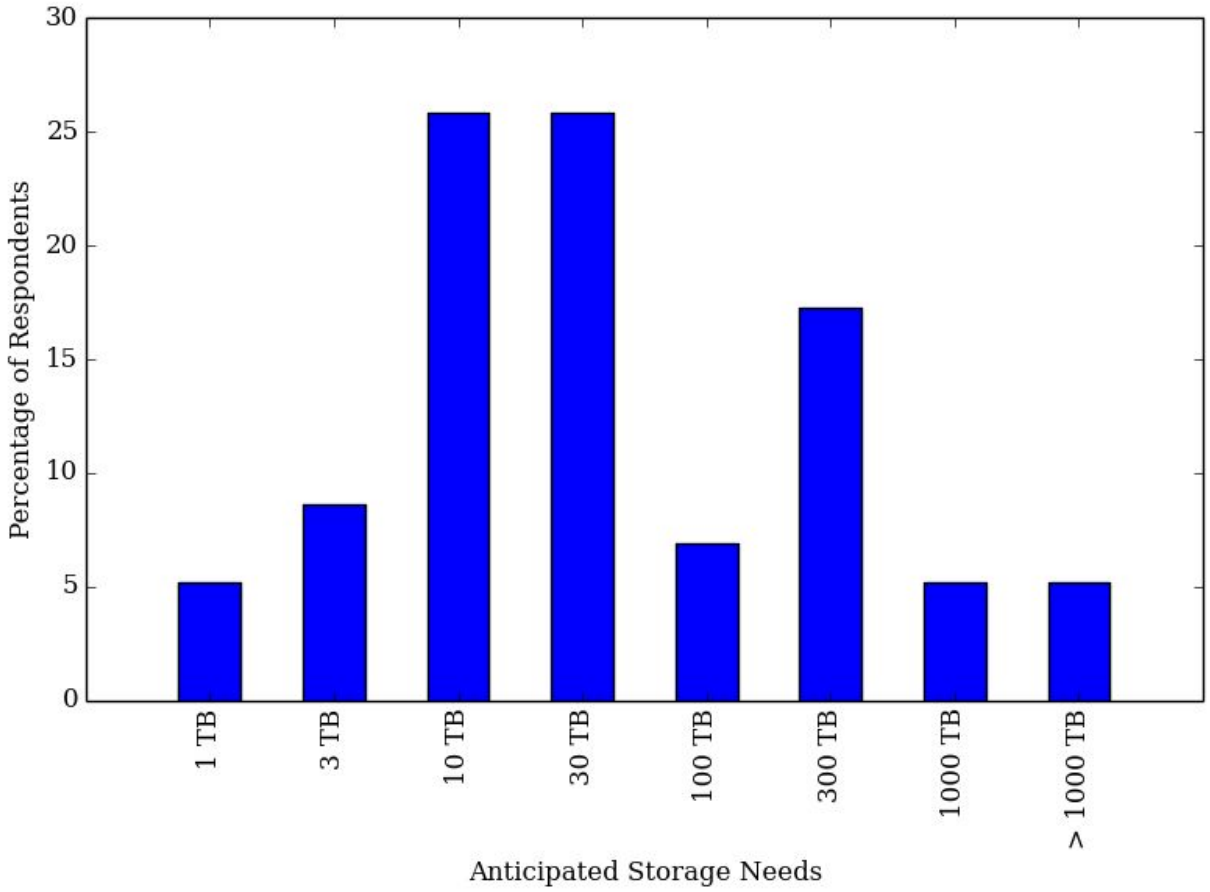

Figure 2: Anticipated Storage Needs for Individual Canadian Astronomers. Data are derived from a poll of CASCA membership and asked respondents to estimate the minimum amount of storage they would need to support their research program, with no margin for redundancy or backup in the figure. Most astronomers anticipate growing but modest data impact but the highest end users will require new solutions to their data needs.

Finally, 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:

1. (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.
2. (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.
3. (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.

1. 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 computer infrastructure in the near term. To meet this directive under a Compute Canada managed model, the community would need:
    a. Long-term storage allocations.
    b. Reallocation of storage requirements to accommodate changing needs on short (< 1 year) timescales.
    c. Web servers on CC infrastructure.
    d. Database servers on CC infrastructure.
    e. Elastic allocation of cloud processing.
2. 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. The needs of the community are diverse and GPU resources are required for some applications but pure CPU resources are most urgently needed. It is likely that at least two computing facilities would be needed to meet this goal. However, a major ongoing problem is the lack of predictability in the provision of systems. Users need to know that competitive resources will continue to be available on 10 year timescales or longer to be able to commit to compute intensive research.
3. 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. While currently rare, members of the community anticipate needed access to PB scale data sets for research in the next five years. A national facility with fast local storage and large amounts of memory per node would be needed to retire this risk. A common site to host large survey data with attached processing capabilities would likely reduce the overall data impact of astronomy since fewer redundant copies would be required.

Update for SPARC2 Consultation

The forecasts presented in 2014 remain largely accurate and the identified risks remain relevant, even with the proposed refresh. Since the submission of the original SPARC whitepaper, Canadian astronomy has undergone a discipline-wide review conducted through the Long Range Planning initiative. While not finalized, the review panel has recently presented their draft report[3] to the community. Of note, the report affirms support for participation in all of the current and upcoming facilities outlined in this whitepaper. Notably, this support includes data intensive observational projects, including the Square Kilometre Array which is the top priority for Canadian ground based astronomy over the next five years (see separate whitepaper).

Recent commitments to funding for computation and cyberinfrastructure through Compute Canada and CFI have reduced some of the concerns about the ability for CC to meet the upcoming needs of the community. However, concerns remain about the ability for the resource plan to meet the upper envelope of research needs for the largest HPC requirements. Astronomy users running the largest jobs (>10^3 cores simultaneously) or requiring accelerators like GPUs have found that the CC model does not accommodate their research effectively, citing long queue times, frequent queue "stalls," and significant downtime. Prioritizing large, parallel jobs in hardware build out and queue structure would support these research agendas. CC can also foster a more responsive staff that can support individual research groups when, e.g., libraries are upgraded. Even with the proposed refresh, CC resources will not be competitive globally, so a focus on efficiency must help minimize the gap.

Data-intensive astronomy is becoming the norm and more observational astronomers are turning to Compute Canada to meet their research needs. Notably, the CASCA Mid-term review made the following (draft) recommendation:

> *Recommendation: CC provide services such as authentication/authorization, and efficient distributed storage platforms that encompass both archiving and user spaces in a scalable way. Services like this will enable the development of increasingly sophisticated analysis platforms.*

Furthermore, the CANFAR platform has developed further and become tightly integrated with the entire process of Canadian astronomy. Recent attempts to secure funding for ongoing development of the platform to meet upcoming challenges have been rejected, owing to the CFI

[3] http://casca.ca/wp-content/uploads/2016/02/MTR_draft_v1.12.pdf

funding model being unable to accommodate partial support for NRC activities. Given the vital role played by the CADC in the entire scientific endeavour, the MTR made the following (draft) recommendation.

> *Recommendation: Given the increasingly blurred relationship between data and analysis products, the MTRP recommends that the Agency Committee for Canadian Astronomy (ACCA) meet to discuss possible strategies to avoid policy traps that preclude the funding of the development of critical and innovative software infrastructure for CANFAR.*

In summary, access to adequate research computing remains a risk to scientific agendas in the most recent assessment of the computing landscape in Canada.

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.