Computation and Data Committee
Final Report to the CASCA Long Range Plan Committee May, 2019

Computing and Data Committee 2010-2020

The Computing and Data Committee (CDC) evolved out of the Theory and Computation Committee. Over time the committee and its members tended to be involved with simulations, data and computation and did not have contact with traditional theory areas (such as relativity or particle physics). In addition, the committee has had close involvement with CADC and CANFAR (via having members representing CADC/NRC) for well over 10 years. Thus the focus had become Computation and Data. It was felt that the mandate was too broad and the committee should focus on Computation and Data. A decision was made around 2010 to make this explicit and call it "The Computation and Data Committee". Unfortunately, this did not propagate throughout CASCA properly. For example, the website changed the name but not the associated descriptive text which still referred to theory as of Sept 2019. For these reasons we restate the committee mandate as it has been practiced for the last 10 years here:

> "The Computation and Data Committee has the important mandate of providing expert opinion on computation and large database astronomy."
> - Keeping track of Canadian computing infrastructure: hardware, funding opportunities and people who support computing and data.
> - Maintain dialogue with CADC and CANFAR
> - Engagement with Compute Canada (since ~ 2006), CFI and related groups (e.g. consortia, Provincial groups)

In the lead up to LRP 2020 it has become clear that there is no particular committee for Theory. At some level CITA and its members cover theory in Canada. CITA was formed out of recommendations from previous community planning exercises. The theory white paper (W40) focuses on the roles of theory and CITA in Canada. CASCA may wish to study the need for better or more explicit representation for theory within CASCA rather than getting it indirectly via CITA.

Current CDC membership:

James Wadsley (McMaster) (Chair)
Pauline Barmby
Catherine Lovekin
J. J. Kavelaars (HIA/NRC/CADC):
Erik Rosolowsky

James Wadsley has served as chair from around 2010 and has thus authored the reports from 2010-2019. JJ Kavelaars has also been a member over this period and Erik Rosolowsky for a large portion of it. Prior members over the past 10 years have included Jonathan Dursi, Jason Fiege and

Hugo Martel.

To address its current mandate there are important requirements on members. The committee should include one or more members who are:

- Engaged with National and provincial level computing providers
- Engaged with CADC/CANFAR
- Involved with data/computation: observations and simulations
- Knowledgeable regarding growing areas: big data, pipelines, cloud and middle-ware

Computing and Data in Canadian Astronomy 2010-2020

The past 10 years have had many notable developments regarding Computation and Data in Canada. Committee reports stretching back over the past years are archived on the CASCA website Compute and Data Committee website: **https://casca.ca/?page_id=273** . We summarize these developments below.

Computing Infrastructure in Canada

Canadian High Performance Computing (HPC) has evolved considerably over the past three decades. A key milestone was the creation of CFI in 1997, offering an effective mechanism to fund computing hardware and storage. HPC is often defined as computing significantly bigger than a desktop. The CFI era began with systems 10's or 100's of times bigger than office workstations all across the country and gave a big boost to parallel computing and simulations in Canada.

Over time there has been a focus on larger scale collaboration and consolidation. This led to the creation of multi-university consortia such as MACI (Alberta, 1997), HPCVL (Ontario. 1998), SHARCNET (Ontario. 2001) and CLUMEQ (Quebec, 2001). This was followed by regional consortia. e.g. Westgrid (covering BC, Alberta 2001) and ACENET (covering the Atlantic provinces in 2003). Ultimately it led to a new national organization, Compute Canada, in 2006. Compute Canada has worked closely with the regional consortia who manage the systems and provide front-line services to researchers. Under Compute Canada, we moved to a common national model for access to systems, account management and to apply for large CPU time and storage allocations.

Over the entire period from 1997 until 2021, all major academic HPC systems have been or will be funded by CFI with matches from provincial governments. The overall result has been that Canada has developed a national network of powerful, world-class systems available to all its university researchers. These systems also enable services provided by CADC and CANFAR. Accompanying this has been the development of a network of experienced staff, funded by CFI and provincial grants. These staff are typically located at universities, supporting users and providing essential training to researchers and their students. We note that there have been other large research systems such as those maintained by Environment Canada, NRC and occasionally, provincial initiatives. Access to these systems has typically been more restrictive.

CFI created special competitions for HPC infrastructure but they were not run regularly, leading to a boom and bust cycle whereby Canada's academic HPC capacity would temporarily become comparable with other G20 nations and then sink down, relatively, for several years before the next refresh. There were also difficulties associated with provincial matches (not always forthcoming) and funding for operating the systems and technical staff. Some of these difficulties were related to tension between Compute Canada, whose management showed a preference to centralize, and regional organizations. These tensions were directly related to delays in new funding and caused disruption for researchers. CASCA and other research groups periodically played important roles in defusing this tension and refocusing the compute providers on serving their users. A key example was a researcher petition circulated in 2013 and sent to Compute Canada, CFI and government. Key concerns included a lack of transparency by Compute Canada, inadequate consultation with the research community and poor coordination with key stakeholders (including provinces and regional organizations). The committee directs the interested reader to a previous CDC position paper, submitted to CASCA in 2012. Such tensions have re-surfaced repeatedly over the last 13 years and ultimately led to a rethink of federal support for large scale computing with Compute Canada's involvement to be phased out over 2-3 years from 2019.

The federal government, through the Ministry of Innovation, Science and Economic Development (ISED), has decided to create a new non-profit corporation responsible for Digital Research Infrastructure with a greater focus on data. This new organization will take on roles expanded beyond computing to include Data Management and Research Software which are currently associated with Research Data Canada and Research Libraries. ISED hopes the new model will resolve issues associated with Compute Canada's lack of consultation, CFI's intermittent funding and the unwillingness of some provinces to match funding. National systems will be largely fully funded by ISED. However, ISED expects that provinces and universities will take over the majority of support for technical support staff. CANARIE will retain responsibility for networks and related infrastructure and development. Another driver for the reorganization was the recognition of the importance of big data, analysis and its computational requirements and recommendations arising from Canada's Fundamental Science Review (2017), led by David Naylor.

The sole applicant to form this new corporation is a consortium created by the U15 (Canadian research intensive universities). On their behalf and in consultation with stakeholders that included CASCA (via the CDC committee), a working group submitted a proposed process to develop the new organization to ISED. This proposal now has ministerial approval (Summer 2019, see: https://www.ic.gc.ca/eic/site/136.nsf/eng/home). A total of $375M funds covering roughly 6 years are expected to start flowing in 2019 to establish the new non-for-profit corporation. The transition to this new organization is expected to run until 2021, during which time existing organizations (including Compute Canada) will continue in their current roles. The proposed governance structure has a membership consisting of Canada's universities and colleges with an extended set of stakeholders including research librarians (and related organization affiliated with data curation), information officers (CIO and university IT staff), researcher groups (such as CASCA) and industry groups. The new organization will have a new user committee with stronger connections to the board (including contributing up to 2 user board members).

Ongoing consultation with researchers and stakeholder groups is planned and there are many unresolved issues. While roles for specific stakeholders such as universities and researcher representatives are described, others are left vague. Notably absent from the consultation committee are the domain specific data centres, such as CADC and the Institute of Particle Physics, both major curators of science data in their disciplines. The future role of regional consortia that currently manage most aspects of HPC in Canada are unclear. These consortia also speak for stakeholders (primarily researchers) in their regions in a fairly comprehensive and representative way which must continue in some form. Another concern is how to fund staff currently performing training, user and software support. A final key concern is that the new focus on data may inadvertently undercut simulation work and software development related to research. It will be critical that CASCA pay close attention over the next 2-3 years and help guide the development of this new organization so that it is transparent and responsive to the needs of the astronomical community.

Candian Computing Infrastructure Status, June 2019

Canada's major CFI-funded academic research systems are listed below in Table 1. Computing power is measured in Teraflops (Tflops) achieved on a standard linear algebra problem in parallel using all the cores on the system in parallel (both normal CPU and graphics GPU). A modern single serial core might achieve 0.1 TFlops. A comparison to top systems in other countries shows that our largest systems are around a factor of 5 smaller than flagship systems elsewhere. Canadian astronomy groups have successfully applied for large blocks of computing time to remain fairly competitive. For example, 3700 core years for 3D hydrodynamics simulations of stellar convection with nuclear burning (or 6% of the Niagara system for a year for the research group led by Falk Herwig, UVic, 2019). Successfully getting such large allocations may not be sustainable as needs grow. For comparison 15000 core years were required for the Illustris-TNG cosmological galaxy simulations (or 5% of the German S-MUC system, 2018) which would be ¼ of Niagara, our largest system today.

Top 500 Rank	Cluster	Cores (CPU+GPU)	TFlops	Notes:
69	Niagara (Toronto)	60000	3000	#53 in 2018
106	Béluga(McGill)	70000	2300	GPU (New)
256	Cedar-2 (SFU)	55000	1600	

312	Cedar (SFU)	35000	1300	GPU
350	Graham (Waterloo)	50000	1200	GPU
1	Summit (US)	2400000	150000	GPU
3	TiahuLight (China)	11000000	93000	
6	Piz Daint (Swiss)	390000	21000	GPU
8	ABCI (Japan)	390000	20000	GPU
9	S-MUC (Germany)	300000	19500	
11	Pangea3 (France)	290000	18000	GPU

Table 1. Current top National Systems in Canada, located at the indicated universities and accessible to all Canadian university researchers via Compute Canada. Included below are top systems for 6 other countries for comparison. Rankings provided are from the June 2019 Top 500 list (top500.org).

As noted in Figure 1, the aggregate computing power in Canada is also somewhat low. Canada has gone through repeated boom and bust cycles for computing infrastructure. We are currently coming off a recent wave of CFI investment but we are still low: in terms of TFlops per GDP average, Canada has half the average value of G8 countries. During the bust part of the cycle around 2015 we slipped to around 20% of the G8 average. The new Digital Research Infrastructure approach, funded directly by the Federal Budget (ISED) should provide more sustained investment.

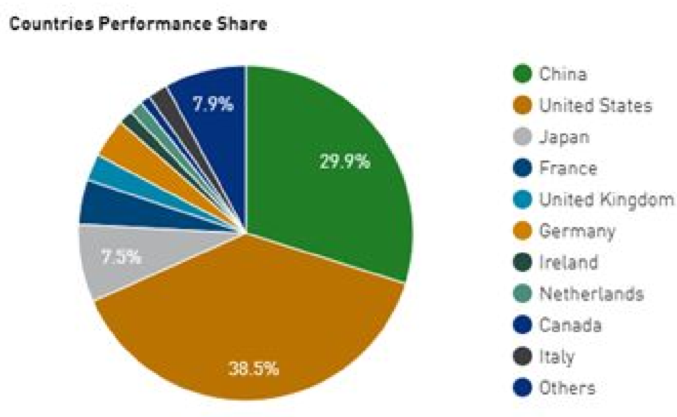


Figure 1. Selected Countries Performance Share (Source top500.org June 2019 Rankings). Canada is 15th for research systems. Out total computing power (15000 TFlops) is about 2% that of the US or half of what is typical for a G8 country based on GDP (the smallest wedge in the figure).

The current total storage associated with Compute Canada systems is of order 25 Petabytes in globally visible short (disk) and long term storage (disk and tape).

New CFI investment of $90 M (total including CFI and matching funds) is planned for the next 2 years to expand 4 of these current systems. This will be the last CFI investment within the old Compute Canada framework. The 2018 Federal Budget earmarked $350 M over 6 years for the new Digital Research Infrastructure framework ramping up from late 2019. The new framework is funded directly by ISED and will not require CFI-type matching but provinces are expected to pick up operating costs.

Canada's research Networks are managed by Canarie who will maintain an independent role alongside the new DRI framework with approximately $140 M invested over the next 4 years. Canarie's networks currently offer 100 Gigabits/s or 80 seconds to transfer 1 TB of data.

CANFAR, CADC and Supporting National Facilities

On the data research front the past decade has been marked by a transformation in the ability to analyze large datasets using 'cloud computing' resources (such as AWS, Google Cloud and the Compute Canada cloud). This transformation is occurring at the same time that astronomers are increasing the rate at which data acquisition occurs. Many telescope projects (CTIO's DECam and CFHT's MegaPrime for example) are now easily managed by such infrastructure. Indeed, the ease of cloud computing and storage was a primary enabling factor in the Gemini telescopes decision to move to an 'in house' archive solution, rather than continuing to contract to the CADC for this service.

Within Canada, DRI is provided via Compute Canada (CC) who have, over the last decade, evolved to support an increasing capacity of cloud computing. US and European researchers have been enabled to take advantage of commercial cloud. Although more fiscally expensive, commercial cloud has evolved their services offerings (beyond raw infrastructure) at a much more rapid rate than has been possible for Compute Canada. In 2008, the CADC helped establish the Canadian Advanced Network For Astronomical Research, CANFAR, to serve the data-intensive storage, access, and processing needs of university groups and centers engaged in astronomy research. CANFAR's primary goal has been to build on top of the bare CC infrastructure to provide the service offerings (eg. customized virtual machines, managed storage, batch processing using virtual machines, group access rights) for astronomy research. CADC has also worked with CC to develop these services as generic research infrastructure. The challenge in continuing to evolve CANFAR has been to find the correct mix of funding through NRC and CFI to enable the required evolution. In addition, CC's support staff have been unable to sustain the deployment load that CANFAR evolution creates and NRC has needed to supplement support funding within the CC cloud system to ensure that CANFAR services can evolve on a reasonable time scale.

Projects like UNIONs, CIRADA and DragonFly are now making use of the specialized CANFAR layer. The number of projects using the CANFAR project has grown substantially over the last decade, starting with 3 or 4 projects in 2011 when CANFAR became operational to currently supporting 186 compute users across 53 different research projects and delivering over a PB of data to 1000s of users around the globe.

Some new facilities however, like the Jansky VLA and CHIME, are creating orders of magnitude more data and observations from such facilities are not well matched to the cloud infrastructure currently available. These projects are not making use of community infrastructure as their internal requirements are significantly more demanding than such facilities can deliver. In addition, the network capacity between NRC's DRAO where CHIME is located and the SFU computing centre is significantly under-powered compared to the data rate. The lessons that facilities like CHIME are providing should inform the next phase of development of CANFAR services as we move into the LSST, SKA and MSE era.

In addition to the growth of datasets has come the growth of data processing requirements and in particular the very recent growth in Machine Learning computing applied to large astronomy observation sets (as opposed to catalogs). This growth is stimulated by advancements in machine learning techniques and the exposure of those techniques with easy-to-learn packages, such as scikit-learn in Python and 'TensorFlow'. These specialized software packages are optimally executed on GPU enabled computers. At this time astronomy research demands for GPU cycles exceeds Compute Canada's capacity to deliver and GPUs have become a strictly rationed resources.

CADC continues to work with the CANFAR user community to expand the capacity of that platform to meet the future needs of Canadian astronomy. For example, beginning in 2020 CANFAR hopes to integrate GPUs into the platform and CANFAR/CADC are currently exploring new ways of delivering virtual desktops for interactive computing under an ALMA funded development project. Such efforts will, however, need to be enhanced as Astronomy really is now in its own 'big data' era.

Survey of CASCA Membership and Facility Projections

From June until September 2019, the CASCA-CDC conducted a survey of CASCA membership to project the community's usage of computing and data resources over the next decade. Of the 572 members identified in the CASCA Secretary's report to the AGM, we received 87 responses (15%). Compared to CASCA numbers, 55 reported their career status as Faculty or Staff members, and 16 reported being postdocs or temporary staff, which can be compared to the 324 Ordinary CASCA members. Sixteen respondents reported as students compared to 142 student members. The population of graduate students involved in astronomy is significantly larger than the CASCA membership number and we might anticipate that some of the student respondents are not CASCA members (i.e., the completeness for students is difficult to gauge and quite low). We also asked whether the respondents were affiliated with a Canadian institutions and 82/87 respondents indicated they were. In addition to surveying the membership, we reviewed the computational requirements of different facilities and initiatives considered for the Long Range Planning Process. We want to identify those facilities that have computing requirements that necessitate dedicated planning.

We summarize the results of the survey below with the raw data available here.

Facility Usage: The survey asked whether respondents were likely to be users of given facilities over the next five years on a qualitative scale including responses "Unlikely to Use", "Potential User", "Likely User" or "Certain User". We aggregate these responses into a fractional community usage to identify which facilities the survey respondents would see the largest community usage of the 2020-2030 decade. We assign a weight of 0 to "Unlikely to Use", 0.2 for "Potential User", 0.5 to "Likely User" and 1.0 to "Certain User", resulting in the following usage weights for respondents. This fractional usage for these facilities is given in the Table below listed in order of aggregate expected use.

Facility	Use	Data Impact	Processing Impact
CADC	0.66	High	High
HPC Facilities	0.61	High	High
CANFAR	0.49	High	High
CFHT/MSE	0.37	Low (50 TB/year)	Moderate (50 cores)
JWST	0.35	Low (20 TB/year)	Low
ALMA	0.31	Moderate (500 TB/year)	Moderate (50 cores)
HST	0.29	Low (1 TB/year)	Low
LSST	0.29	High (2 PB/year)	Moderate (400 cores)

Gemini Telescopes	0.27	Low	Low
SKA & Precursors	0.27	High (7 PB/year)	High (10000 cores)
TMT	0.26	Low	Low
CHIME/CHORD	0.16	Moderate (700 TB/year)	High (1200 cores)
JCMT	0.16	Low (50 TB/year)	Low (50 cores)
Subaru	0.15	Moderate (100 TB/year)	Low
CCAT-prime	0.11	Low	Low
DRAO/ST	0.1	Low	Low

Users also identified interest in using WFIRST (2), Euclid (4), SDSS-V (2), GBT (3), OMM (1), AstroSat (1), DAO (1), XRISM (1). Within the table, we have also profiled the data and processing requirements from different facilities. We directly requested this information from whitepaper authors when there was a relevant whitepaper submitted. For those facilities that provided data (ALMA, SKA) we included their numbers. For those facilities with publicly available data and processing plans, we also included those values (for LSST, we assumed a 10% share of data and processing hosted in Canada). In all other cases, we estimated the relative impact of the facilities based on the typical requirements of comparable facilities. We organized those facilities into tiers of impact:

- Low: Less than 100 TB per year of data growth, processing data requires less than 100 dedicated computer cores. These can be supported with existing cyberinfrastructure, although in aggregate they may represent a substantial requirement that is met via pooled resources applications within CADC, HPC and CANFAR (see below).
- Moderate: Between 100 TB and 1 PB per year of data growth, processing requires between 100 and 1000 dedicated computer cores. These would require planning and growth within existing cyberinfrastructure.
- High: Requires more than 1 PB per year of data growth; processing requires more than 1000 dedicated cores. These would require dedicated new cyberinfrastructure, such as dedicated computational facilities or network capacity.

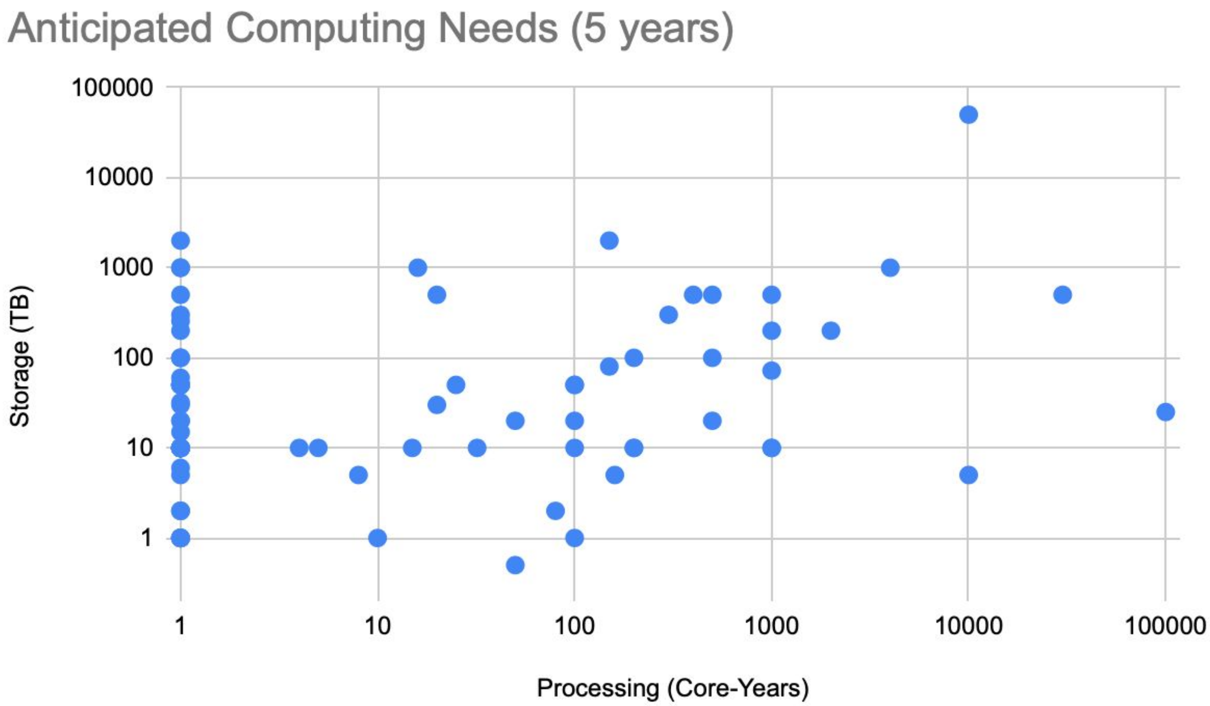
The three facilities with the highest expected use (CADC, HPC Facilities, CANFAR) do not have specific data and compute requirements and would support many of the facilities identified below. We flag these as having High computing and processing requirements given their umbrella support for many other facilities.

<u>Projected Computing Needs:</u> We also asked about the typical computing facilities that the respondents required to complete their work and how frequently they used those resources. The fractions of users in these different categories is indicated below.

	Daily	Weekly	Monthly	Annually	Never
Desktop/Laptop	0.83	0.11	0.02	0.00	0.04
Small Server (10-50 cores)	0.20	0.30	0.22	0.09	0.19
Small Cluster (50-200 cores)	0.07	0.17	0.29	0.15	0.32
Medium Cluster (200-1000)	0.05	0.06	0.17	0.22	0.50
Large Cluster (1000-10000 cores)	0.01	0.03	0.08	0.10	0.78
XLarge Cluster (>10000 cores)	0.00	0.00	0.01	0.03	0.96

We asked respondents how much computing they would need to support their research programs over the next 5 years in terms of core-years of processing and storage, receiving the following responses. For cases where respondents did not report a value, we have filled in a value of 1. We present the results as a bivariate graph to show how users view the balance between these two different resources. We also asked if users were regular users of GPU processing and 40% of the respondents indicated that they were. Finally, we asked users to anticipate the growth of their storage and processing requirements over the next 5 years. Nearly all respondents (90%) indicated a three- to ten-fold increase over their current usage in the next five years.

We asked users to rate how four different types of limitations would impact their research program, specifically, limitations in access to process, limitations in access to storage and data, limits to specific types of resources (e.g., GPUs), limitations in the ability to transport large data sets, and access to improved software. Respondents reported on a qualitative scale to which we assign weights: No Impact (0), Minimal Impact (0.2), Moderate Impact (0.5), and Heavy Impact (1.0). We then aggregate over responses to determine which factors are viewed as the primary risks that users perceive. We then use these scores to rank these perceived risks. The respondents identify limitations in storing large data sets as having the largest impact on their work (score of 0.70) followed closely by access to sufficient processing (0.68). The ability to transport large data sets was deemed to have a slightly lower impact (0.58). Access to specialized computing resources or improved software was viewed has having less impact (both with scores of 0.44).



Dedicated Compute Priorities: We also asked respondents regarding their priorities for new investment that would support community efforts. They were asked to rank the priorities of the following facilities, for which we report the mean rank and standard deviation of ranks from first (1) to last (4) priority:

- Astronomy-focused Computing Centre (2.14, 1.09)
- New Data Intensive Computing Facility (2.26, 0.94)
- New High Performance Computing Facility (2.74, 1.24)
- Next Generation Software (2.82, 1.01)

Of note, the responses for a *New High Performance Computing Facility* were bimodal leading to the high standard deviation. Respondents tended to rank this priority either first or last.

Compute Canada: Finally, we asked respondents if they directly used Compute Canada facilities and 45/87 respondents said yes. Of these 45 respondents, we asked them about their satisfaction with different providers of research computing. We used a 5 point Likert Scale with options Very Unsatisfied (1), Unsatisfied (2), Neutral (3), Satisfied (4), Very Satisfied (5) and No Basis to Judge (Reported separately) The resulting mean scores were:

- Compute Canada: 3.8, with 1 "No Basis"
- Regional Consortia: 4.3 with 9 "No Basis"
- University IT: 3.5 with 5 "No Basis"

Comments cited some concerns about limited access to the Compute Canada facilities either through cluster queue wait times or the clusters being unstable.

Commentary: The results of the survey are limited by the number of respondents, with <20% of CASCA membership responding and only a limited number of students responding. However, with 50% of the respondents indicating they use Compute Canada resources, the survey results may represent those faculty and permanent staff researchers who rely most heavily on computing in their work. We conjecture that we are underestimating the number of small and moderate users of computational resources.

Nearly every respondent to our survey anticipated significant growth in their computational requirements over the next five years. This growth is reflected in their projected requirements, requiring 10s to 100s cores of processing and 100s of TB of storage.

The survey showed broad support for CADC and CANFAR and the regional computational consortia. While not addressed broadly, several users mentioned the need for better software development in astronomy, notably using expert software development rather than astronomical software. High quality is frequently overlooked in terms of the efficient use of the available resources.

Community Whitepaper Submissions:

We reviewed whitepaper submissions to LRP2020 from the community and highlight several reports that identify facilities and research programs that require significant computational resources. These needs are also partly accounted for in the survey of the community, but this study presents a complementary view on community needs. In the Appendix to this report, we summarize the computational and data needs of individual whitepapers. We did not identify any additional areas where the community faces risks from lack of computational resources; however, many of the community driven initiatives are subject to high risk if there is not expanded investment in computational and data resources. The facilities will require specific computation and data investments to be successful. These risks tend to cluster in three distinct areas.

- *Simulations:* There are several science areas where order-of-magnitude expansions in the high performance computing are required to carry out simulations essential to progress. In particular the addition of the flagship niagara facility was noted as essential for recent Canadian progress and regular investments of this scale are required for continued progress. There is also a need for a data-intensive computing centre for post-processing the simulation data for richer insight.
- *Radio Astronomy*: Radio interferometers and high time-resolution observational facilities require extensive computation and data support. This need has been appreciated for the SKA, but all long-wavelength facilities (CHORD, CHIME, GBT, ngVLA) will present data challenges.
- *LSST:* With the community's growing enthusiasm for LSST, there are opportunities to build an LSST-focused data center. Given the volume of LSST data and the benefits for being in the collaboration, a successful Canadian participation in LSST will require dedicated computational investment.

Several whitepapers cite specific need for improved GPU resources. Many whitepapers cite the success of CADC and Canadian cyberinfrastructure projects like CANFAR and CIRADA as field-leading examples that provide a competitive advantage in the field. There is also broad enthusiasm for community-wide access to astronomical data centres that expand the scope of CADC, notably toward LSST and SKA. These facilities are not necessarily tied to CADC, but the whitepapers broadly cite the need to use the CADC expertise to participate in these observational projects.

LRP Criteria Discussion Points

1: How does the proposed initiative result in fundamental or transformational advances in our understanding of the Universe?

Computing and Data are now central to the practice of astronomy and astrophysics. Data volumes require not just numerical analysis but high performance computing systems to do so. These large and complex datasets have been transformative for our understanding of the universe.

2: What are the main scientific risks and how will they be mitigated?

We defer discussion of specific science to associated white papers. However, some general trends appear.

Firstly, the amount of data being created continues to increase. This applies to observations and simulations in all sub-areas. This requires triage of datasets at the point of creation and selective storage, often in a hierarchical fashion. Set against this are tri-council requirements for storing raw data and making it available. Creative strategies may be required, such as making initial conditions and codes available for simulations where it is not practical to store outputs. In addition, just managing large datasets requires increasingly specialized expertise and software solutions, often called middleware or cyberinfrastructure. Resources to develop, deploy and maintain such software can be hard to acquire. There is also a need to coordinate university level, CADC, consortium/provincial and national roles. It is currently unclear how future funding, formerly channeled through Compute Canada, will operate under the new DRI framework. A particular concern is how expert staff currently in consortia will fit into the new framework. As noted above, they fulfill critical user-support roles currently and are difficult to replace.

Secondly, the associated processing needs often scale non-linearly with data set sizes. Keeping up may require specialized hardware (GPUs), parallel analysis and developing middleware and specialized extensible toolkits. It can be difficult for individual researchers to get the tools they need and/or train their group to use tools developed by others. Canada has a lot of expertise but it is scattered among universities, NRC (CADC) and the current computing organizations at the provincial (consortia) and national levels (Compute Canada, Canarie and others). Various attempts have been made to fund new development under CFI short-term programs (e.g. cyberinfrastructure) but we need ways to develop and retain expertise long term. CADC is currently in a good state but somewhat under-resourced if it were to take on a larger role.

3: Is there the expectation of and capacity for Canadian scientific, technical or strategic leadership?

Canada has access to world class facilities for observing and simulations. We have data storage, management and analysis solutions in place and plans to move forward. We have many talented individuals who function close to the cutting edge with connections to leading international groups. However, we operate on a relatively low level of funding and personnel that means we tend to be slightly behind the curve. This means we can take advantage of solutions that external groups produce but it limits our ability to be right at the forefront in many areas related to data, analysis and software development.

Strategically, the CADC has received long-term support for the mission of astronomy data management. This long term vision has enabled the CADC to provide leadership towards data-layer standards in astronomy. The use of standardized access modes and formats substantially reduces costs and can enable more rapid science results from merging of datasets.

4: Is there support from, involvement from, and coordination within the relevant Canadian community and more broadly?

The Canadian astronomical community is highly integrated into the international astronomy community at the level of individual instruments and science collaborations. We are well integrated into existing Canadian support groups for Computing and Data. However, as noted above, we typically function with a relatively low level of support (e.g. to fund staff and specialists).

5: Will this program position Canadian astronomy for future opportunities and returns in 2020-2030 or beyond 2030?

See above.

6: In what ways is the cost-benefit ratio, including existing investments and future operating costs, favourable?

All astronomers must live within the same constraints on datasets and analysis. In many cases Canada's astronomers operate within international collaborations that collectively implement solutions to these constraints. Canada is somewhat under-resourced for Canadian groups to undertake projects in either observations or simulations that would push us to the cutting edge in some areas.

We do not make cost estimates in this document but the projected order-of-magnitude growth in computation and data needs will require investment comparable to participation in an observational facility (CAD 30M to 100M over the decade). Computing needs and resources are both growing and users tend to consume all the resources provided. In addition, the budget envelopes are somewhat fixed for now and the main national support organization is in a state of flux. Much of computing is a service in support of other things. As the DRI organization takes shape, we should be able to identify gaps in terms of infrastructure and support in advance. Beyond the committee itself (CDC) there are several groups focussing on individual facilities. CASCA should monitor developments and seek to head off any serious mismatch between resources and needs. In particular, we should continuously advise the new DRI the organization to help it direct appropriate resources to different components that affect astronomy in general and individual facilities in particular.

7: What are the main programmatic risks and how will they be mitigated?
We interpret programmatic risks to refer to those associated with the underlying infrastructure for computing and data. We have discussed above how this infrastructure (in terms of facilities and people) connects to scientific outcomes. The big picture concerns for infrastructure are two-fold. Firstly, we must ensure that investment is sustained. Historically, with CFI as the primary provider, this has been somewhat problematic but with DRI we are fairly optimistic that it will be sustained. However, with larger funds available and more players there is a risk of that competing priorities and turf-wars may leave important gaps. This leads to the second concern: that the investment is balanced among compute hardware (including CPUs and GPUs and how it is made available via cloud, resource competitions and so on), storage (both in the successor to Compute Canada, CADC and in telescope associated facilities, short and long term) and personnel. We have experts in the form of tenured academics but we need full time support for using the infrastructure (including training), developing and maintaining software and managing the datasets we produce. CASCA must pay close attention to how the new, well-funded DRI organization develops to ensure we get this support and retain expertise though viable career paths for non-tenured specialists. Key groups include postdocs, local staff (universities, consortia), NRC (CADC, CANFAR) and national staff (currently Compute Canada). A pressing concern is that the new national DRI framework has no funding for local staff with a currently unmet expectation that provinces or others will pick up the tab.

The most direct way CASCA can mitigate risks is to play an active role engaging with the development of the new DRI organization (and associated groups include CANARIE, NRC and governments). For the past decade, sporadic consultation with selected individual researchers has been the primary approach of larger compute organizations, such as Compute Canada. However, provincial groups, most notably consortia, have filled the void fairly well. They tended to know their users (local researchers) quite well and thus had a fairly balanced view which they carried forward to the national level. Historically, roles for organizations like CASCA have been suppressed partly because other comparable professional groups are not as engaged as we are. We would benefit if more groups were organized and presented more of a consensus view of their disciplines. We could then jointly seek a greater role in decision making. If the role of consortia diminishes, as expected, this may be the only effective way to ensure astronomers needs are met.

8: Does the proposed initiative offer specific tangible benefits to Canadians, including but not limited to interdisciplinary research, industry opportunities, HQP training, EDI, outreach or education?

Astronomy can sometimes find it challenging to sell itself to commercially minded government and the wider community. However, Computing and Data are huge areas of growth, both in academia and beyond, making HQP a key strength. Our students and postdocs get the opportunity to build related skills that enable them to do very well outside of academia, particularly in growth areas such as Big Data (as data scientists) and machine learning.

In addition, Computation and Data activities could contribute to education and outreach. Astronomy produces huge amounts of data from both simulations and observations. While processing these data sets involves HQP, we can also take advantage of these in terms of education and outreach. Many of these data sets result in beautiful images, and these could be made accessible to the public in the form of websites or a youtube channel (e.g. movies of simulations). Digital delivery is an increasingly effective pathway. As a community, we produce world-class research, and we need to present those discoveries to the public and prospective students.

Appendix: Review of Whitepaper Submissions

Here we present our review of those whitepapers with significant computational or data requirements or those that would have bearing on how our computational needs are addressed.

- W004 - Machine Learning Advantages in Canadian Astrophysics - This report advocates for the computational resources to apply Convolutional Neural Network machine learning methods to several upcoming survey facilities including the Euclid mission, LSST, CHIME/CHORD, and spectroscopic facilities like MSE. While the whitepaper cites a lack of facility access as a risk, no quantitative estimates of requirements are provided but the processing and training of neural networks can have significant processing requirements. Limited access to high performance computing poses a moderate risk to this program that could be mitigated by preprocessing and sample selection.
- W005 - Signposts of planet formation in protoplanetary disks - This report cites Canadian computational facilities and management (e.g., through CANFAR and Compute Canada) as a strength for the program. Continued access to HPC is required for both the data reduction and analysis as well as numerical simulations. Lack of access to improved HPC poses a moderate risk to the data reduction aspect of the program and a high risk to the simulations mentioned in the report.
- W009 - Low-redshift 21cm Cosmology in Canada - This report advocates for a science case using low-frequency radio telescope arrays (CHIME, CHORD, HIRAX in this report), which have extremely large data rates. While not explicitly mentioned as a risk, these facilities have large computation and data requirements that are considered in their construction and in the above facility models. Lack of access to improving HPC, both for computation and large data archiving facilities, is a high risk to these facilities.
- W010 - Astrostatistics in Canada - This report does not describe computational and data requirements explicitly but many of the research directions advocated in the report require Big Data and the computational facilities capable of processing those data. Lack of access to HPC represents a range of risks to science programs described here as astrostatistical approaches range from requiring little data and computation to being very computationally intensive.
- W012 - High-redshift 21cm Cosmology in Canada - This presents similar computational risks and requirements as W009.
- W015 - Canadian Participation in the LSST - The submission presents a direct estimate of data requirements to support Canadian researchers citing 2 to 15 PB range for potentially interesting data products over the 2020-2030. The whitepaper does not describe the computational requirements for analyzing this large amount of data, but they do cite the CANFAR resource as the means by which this initiative will support Canadian processing. This Canadian initiative is infeasible without the data resources outlined in the whitepaper. The lack of access to HPC resources, i.e., through CANFAR presents a high risk.
- W016 - Pulsar Timing Arrays - While this whitepaper does not specifically describe data and computation requirements, pulsar data processing remains one of the most data and

computationally intensive branches of observational astronomy. Processing relies heavily on GPU resources and generates large data volumes. Within Canada, these needs have been met by Compute Canada and other CFI-funded resources. However, continuing to grow this area of research will require further investment in storage and computational facilities, notably GPUs.

- W017 - Star Formation in the Galactic Ecosystem - This whitepaper cites the need for HPC to support large simulations of star formation and galaxy evolution. A lack of improved computational facilities presents a moderate risk to the science program outlined here.
- W018 - CASTOR: A Flagship Canadian Space Telescope - This whitepaper notes a raw data rate of 200 Gbytes / day placing its facility impact moderately data intensive (<100 TB/year) with computational requirements that are likely moderate. These requirements could be met with existing infrastructure, thus this initiative has low computational risk.
- W019 - Development Plans for the Atacama Large Millimeter/submillimeter Array (ALMA) - The author team for this whitepaper prepared data and processing estimates as outlined in the Facilities section. These present moderate computational and data challenges that could likely be met within existing facilities. This initiative also relies heavily on advanced processing of data through the ARCADE system which relies on CANFAR and Compute Canada infrastructure. Failure to maintain that infrastructure would present a moderate risk to the initiative described here.
- W020 - The Euclid Mission - This whitepaper advocates for participation in the Euclid mission, using the data available through the Canada-France Imaging Survey (CFIS). The data challenge presented by Euclid is large requiring 100 PB of storage and 20,000 cores for the entire mission (arXiv: 1701.08158), but it is unclear what the computational requirements for Canadian involvement would be. However, making significant computational contributions would require investments comparable to a Canadian LSST effort.
- W021 - Astronomy in a Low-Carbon Future - This whitepaper points out that the power generation required to support intensive computing has significant carbon impact. This whitepapers argues that shifting to data centres and cloud computing providers that mitigate the carbon footprint of astronomy. This vision can be harmonized with many of the initiatives proposed here if the research community preferentially selects computing providers with low-carbon power generation.
- W023 - Fundamental Physics with Pulsars - This whitepaper initiative has similar considerations as W016 (Pulsar Timing).
- W024 - Star Clusters Near and Far - This whitepaper outlines observational and theoretical initiatives for scientific progress in the physics of Star Clusters. While the observational requirements are outlined in the requirements for different facilities, the simulation aspect is identified. In particular, this science path requires investment in expanded computational capabilities, particularly GPU resources for efficient gravitational+hydrodynamic simulations. The whitepaper also cites the need for continued software development to leverage new resources.
- W025 - The next decade of optical wide field astronomy in Canada - This whitepaper specifically identifies expansion of the role of CADC/CANFAR to meet the science needs of

this initiative. In particular, this effort identifies LSST Science as a major new direction as outlined above.

- W026 - Probing Diverse Phenomena through Data-Intensive Astronomy - This initiative describes the need to restructure large computing resources to be more useful for data processing initiatives. In particular, it advocates for Canadian computing centres to move toward a high throughput computing model where data processing facilities are a community resource and data are collocated with computational resources. These shared computing resources should also allow for prototyping and development.
- W028 - The Canadian Hydrogen Observatory and Radio-transient Detector (CHORD) - The data and computational requirements for CHORD are substantial but these requirements are incorporated into the instrument design. The facility plans to rely on Compute Canada for its computational requirements and failure to secure these computing resources would make the project infeasible.
- W032 - The Next Generation Very Large Array - This facility would only have a computation and data impact in the post-2030 timeframe. However, like most radio interferometers, the facility will likely require focused attention on computation and data analysis. However, with fewer elements in the array compared to, e.g., SKA-Low, the requirements will not be as demanding as for SKA under most use cases.
- W034 - Revealing the Origin and Cosmic Evolution of Supermassive Black Holes - This whitepaper mentions requirements for extensive computational efforts. This includes both in the forward modelling of observations and on carrying out the simulations of material accreting near the event horizon of black holes. These requirements are not clearly outlined, but it is likely that a lack of computational resources would pose a moderate risk to the science goals.
- W037 - DRAO Synthesis Telescope - This proposal outlines a suite of upgrades for the DRAO Synthesis telescope. With only 7 elements, the data and computation requirements for the interferometer are modest compared to larger interferometers. The proposed science cases should not have large data rates, but if the facility were to move into fast time-resolved data, the rate would dramatically increase. At scoped, a lack of investment computational facilities poses low risk to this initiative.
- W038 - Astrophysics and Cosmology with Line Intensity Mapping - The science presented in this whitepaper cites numerical simulation as one of their requirements and the need for sufficient resources for the science case to succeed. However, the requirements are not described. It is likely that a lack of computational resources would pose a moderate risk to the science goals.
- W040 - Theoretical Astrophysics in Canada - Computation is a central tool for theoretical astrophysics and this whitepaper outlines several concerns with the current computational resources in Canada. First, the Compute Canada model is poorly suited to executing large simulations requiring a large fraction of a computing resource (e.g., all of niagra for a day). The data needs of the theoretical community are also large and simulation results are minimally shared because of the volume. This porblem is more acute in Canada than elsewhere due to lack of resources and people to support it. This whitepaper advocates broadening the

community storage resources dedicated to simulations and providing close connection to further computational power for post processing and analysis. Given the central role computation is playing for theoretical work, lack of investment in computational and data infrastructure poses a high risk for this initiative.

- W041 - The cosmic origin and evolution of the elements - This whitepaper notes the central role for numerical simulation in carrying out the scientific development. The authors specifically require successor machines to the niagara computational facility, with tripling the computational capacity of a single resource by 2023 and a nine-fold increase by 2028. Lack of this investment in computational and data infrastructure poses a high risk for this initiative.
- W044 - Canadian Investigations of the Interstellar Medium - This scientific whitepaper cites the need for expanded computational facilities to manage radio data and large numerical simulations of ISM physics. No specific requirements are provided, but it is likely that a lack of computational resources would pose a moderate risk to the science goals.
- W046 - Canada and the SKA from 2020 - 2030 - This whitepaper makes the scientific case for Canadian participation in the SKA. The computation and data requirements for the facility are well studied and participation in the SKA regional centre network is estimated to cost 45 million CAD over 2021-2030. Lack of access to the computational and data support for facility would prevent effective Canadian participation in the project and represents a high risk to the initiative.
- W051 - Science with the Large Synoptic Survey Telescope - This whitepaper outlines science opportunities with LSST, emphasizing the benefits of establishing the Canadian LSST Advanced Science Platform, a cyberinfrastructure system built to facilitate Canadian science use of the LSST data. This would require a heavy investment in data and computation but given the LSST collaboration may consider this contribution as “in-kind” buy in to the project, there is an opportunity to take advantage of Canadian strength in this area. Lack of access to this cyberinfrastructure poses a high risk to Canadian participation in LSST.
- W052 - Planetary Astronomy-Understanding the Origin of the Solar System - This whitepaper specifically highlights the role of data and computing in making advances in planetary astronomy, primarily by leveraging LSST data. Lack of access to LSST data and the computational facility poses a moderate risk to this science case.
- W055 - Cosmic Magnetism - This whitepaper outlines new opportunities in unravelling the role of the magnetic field in astrophysics. Many approaches are relatively light, but the data impact and computational requirements for using polarized radio-continuum data from current and next generation interferometers represents a substantial requirement. Lack of access to sufficient computational resources poses a moderate risk to this science case.
- W058 - Radio Transients - This whitepaper does not specify the computation and data needs to pursue this science goal but cites access to data and computation as a key ingredient in recent scientific progress. Given the data intensive radio astronomical facilities required for progress, this initiative faces a moderate risk from lack of investment in computational resources.