Summary - Infrastructure Renewal Plan

In November 2014, CFI released a draft call for proposals for their Cyberinfrastructure Initiative. As part of this initiative, “Challenge 2” presents Compute Canada with an opportunity to modernize its infrastructure to better support Canadian researchers. Challenge 2 is divided into two stages. Stage 1 involves proposal submission in April 2015 and a final decision by the CFI board in June 2015. The focus of this summary is on the Compute Canada infrastructure plan for stage-1 infrastructure renewal.

Compute Canada maintains an extensive network of Advanced Research Computing (ARC) facilities across the country. Infrastructure renewal will bring some consolidation of these facilities and systems. This document also summarizes this infrastructure consolidation plan. Please note that Compute Canada is committed to maintaining a network of ARC experts at sites across Canada, whether or not those sites host physical infrastructure.

This document is the basis for further consultations on the Compute Canada submission to the CFI cyberinfrastructure initiative. While comments are welcome any time via sparc@computecanada.ca, those received by February 10, 2015 will be given full consideration in our April 2015 submission to CFI. CFI and Compute Canada will be running consultations in 6 Canadian cities between January 20-22, as well as an online consultation session on January 26. Please see the Compute Canada website for dates, times and locations.

The Context - CFI Challenge 2, Stage-1

Quotes from CFI Draft Call For Proposals - Challenge 2, Stage-1
Stage 1: Up to $15 million will be provided for the upgrading and modernization of the computational and data storage capacity of the pan-Canadian advanced research computing platform. As the managers of this platform, Compute Canada will be invited to submit a proposal on behalf of the advanced research computing community; Stage 1: The renewal of the pan-Canadian advanced research computing platform will be conducted in two stages. For Stage 1, the CFI invites Compute Canada, on behalf of its member institutions, to propose three distinct options for the capabilities and services that will enable leading-edge research and address the most pressing immediate needs. This proposal will focus on the upgrading and modernizing of the computational and data storage capabilities managed by Compute Canada.

With a total project cost of up to $37.5M, stage-1 will not be large enough to meet all of the ARC needs of Canadian researchers. The vast majority of current Compute Canada systems are already at or beyond their nominal 5-year lifespan. Aging systems lead to increased operational costs due to less energy efficiency, increased likelihood of system failure and increased warranty costs on key components. These systems are also poorly adapted to modern use-cases and are becoming uncompetitive on the world stage. As such, the focus of the stage-1 renewal is on replacement of existing capacity and to “address the most pressing immediate needs”.

While the focus of stage-1 funding is necessarily on replacement of existing capacity, it is neither possible, nor desirable to replace existing systems one-for-one with new systems of similar design. The design of the new systems must be adapted to modern workloads based on the current and future needs of Canadian researchers.

System Consolidation

Compute Canada currently operates 50 distinct systems in 27 data centres across the country. Most of these are shared systems, though some are also dedicated researcher-contributed systems. It would not be efficient to deploy systems in 27 data centres from stage-1 funding. Further, Compute Canada is moving towards a model in which fewer, larger systems are maintained. This consolidation of systems and data centres will allow for more focused investments, scaleable workloads, and efficient use of resources.

In stage-1, Compute Canada proposes to invest in 4 new ARC systems. In autumn 2014, Compute Canada published a call to its member universities, seeking hosts for this next generation of ARC infrastructure. Responses from 9 potential host institutions, from across the country, are currently being evaluated by an international panel of experts. Once this evaluation is complete, hosting scenarios (combinations of up to 4 sites) will be created and evaluated, considering both technical and financial factors which could influence the delivery of service to Canadian researchers. The proposal submitted to CFI in April 2015 will include placement of each system described in this document.

Compute Canada has performed a cost-benefit analysis of existing systems and has shown that systems commissioned in 2010 or earlier are the most expensive to operate per unit capacity. Compute Canada has chosen a set of systems to be de-funded in the 2016-2017 fiscal year. These systems will not be de-funded until the stage-1 systems are commissioned. In an appendix of this document we provide a table of systems which will be in operation at the end of stage-1.

The Evolving Needs of the Research Community

To inform this plan, Compute Canada has undertaken broad research community consultation. This includes a detailed survey of user needs in the fall of 2013, a series of in-person and online consultations in December 2013 - February 2014 while writing the organization’s strategic plan, and the SPARC call for white papers in the summer of 2014. We have also used internal data from our usage tracking and annual resource allocation process to inform the hardware plan. We recognize that this is still not sufficient to capture the needs of the entire Canadian research community and are committed to continuing the consultation process. This document is being shared for comment as part of the SPARC process in the hope that we can continue the conversation with the researchers we serve.

The Compute Canada user community has grown to include more than 2500 faculty-led research groups, spanning all computational and data intensive disciplines. Recent years have seen significant growth in the needs of all traditional HPC communities, explosive growth in some already data-intense areas and significant adoption of compute and data-intensive methodologies in “new” ARC disciplines. As a result, Compute Canada’s new infrastructure must be capable of hosting a wide array of services for a variety of communities, while still providing the raw compute power required by traditional communities. This plan attempts to strike an appropriate balance.

Compute

Compute Canada currently operates nearly 200,000 computational cores on behalf of the Canadian research community. The majority of these cores are provided in relatively large clusters (thousands of cores), many with some flavour of high-speed interconnect. While some of our current systems were competitive systems at time of commissioning, we do not currently operate any single system in the top 200 worldwide. We cannot properly serve Canadian researchers if our capacity to serve computationally intense research continues to decline.

When we surveyed our users in fall 2013, researchers ranked computational resources as their number 1 current and future need from Compute Canada. The SPARC white papers demonstrated a broad need for increased computational resources over the next 5 years as shown in the table below.

White Paper	Predicted Increase from Current to 2020
Numerical Relativity	3x
Subatomic Physics	3x
Materials Research	5x
Canadian Genome Centres	8x
Canadian Astronomical Society	10x
Theoretical Chemistry	12x

In absolute numbers, by 2020, the theoretical chemistry need alone represents four times the entire current Compute Canada core capacity.

It is clear that Canadian researchers require an increase in computational capacity to be competitive. Nearly 60% of core-years provided in 2014 were used by parallel jobs spanning multiple nodes. We also have a number of excellent researchers who now request allocations of more than 5,000 core-years annually, and who need to have their allocations split over several systems. It is most efficient for both Compute Canada and the researchers we serve to build and operate individual systems capable of supporting users with large needs.

Storage

Compute Canada currently hosts approximately 20 petabytes (PB) of disk storage in addition to significant tape-based resources. While computational cores can be reallocated in a dynamic fashion, using a priority system, data storage allocations are very difficult to re-assign. Researchers accumulate datasets which grow in time and it is not reasonable to simply reallocate storage resources in the midst of a multi-year project. As such, when storage resources come under pressure, as they are now, there is little flexibility in the system.

The Canadian subatomic physics community has some of the largest storage allocations on Compute Canada resources today. This discipline represents "traditional big data". The long timelines of the associated experiments and relative maturity of the field means that the storage growth rate is predictable and controlled. This provides us with an example of a large base experiencing only modest growth. The subatomic physics community in Canada submitted a white paper to the SPARC process showing the storage evolution for the field. A summary table from that submission is included below (note that current numbers include storage provided at non-Compute Canada facilities such as TRIUMF). This shows a modest factor of 3-4 growth in storage needs over the next 5 years.

Subatomic Physics Storage Requirements (IPP+CINP White Paper)								
	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Disk (PB)	12.9	14.9	19.4	22.6	26.5	30.4	37.0	43.9
Tape (PB)	5.5	7.2	10.4	13.7	16.0	23.4	30.9	40.7
Total (PB)	**18.4**	**22.1**	**29.8**	**36.3**	**42.5**	**53.8**	**67.9**	**84.6**

While subatomic physics represents a large current usage undergoing slow growth compared to some other disciplines, the Canadian genomics community is experiencing an exponential growth in sequencing capacity and in concomitant storage needs. The four largest Canadian Genome Centres submitted a white paper to the SPARC process which predicts a factor of 30 increase in required disk and tape storage capacity by 2020. **The disk storage need alone by 2020 is projected to exceed 450PB (tape storage needs are similar)**. Genomics represents an already large base experiencing very rapid growth. While much of this base has been provided outside Compute Canada in the past, we are seeing a significant migration of bioinformatics users to Compute Canada managed resources.

Subatomic physics and bioinformatics are only two examples of disciplines which require large quantities of fast disk storage and are chosen to illustrate a range of growth rates (3-30x) over the next 5 years. The cumulative need is growing for each researcher, while the number of researchers we support has more than doubled in the last 5 years and is continuing to grow. In addition, there is a growing expectation that Compute Canada should act as the primary data repository for many researchers who now have no other place to store and archive their growing datasets. As such, they request tape copies, or multiple live disk copies on Compute Canada systems. While the needs of an individual researcher are small relative to the numbers shown above, the sum is significant and growing.

In the 2015 resource allocation competition, 15.5PB of disk storage was allocated to research projects. Given that a) it is not advisable to allocate 100% of a filesystem and b) some of the disk space must be set aside for "home space" and "scratch space" (temporary storage for active jobs), the Compute Canada disk resources are essentially fully allocated for 2015. Given the historical year-over-year growth in storage requests we project a disk storage crisis in the 2016 competition. It is clear that the newly purchased systems must include a significant increase in total disk storage capacity and in tape capacity.

Accelerators (GPUs)

Graphics Processor Units (GPUs) have become a standard component of ARC due to the very powerful computational capabilities required for modern graphics and games systems. Major GPU vendors now provide special versions of these processors which include only the computational components, without the graphics head. Jobs offload compute-intensive, vector-oriented computations to the GPU co-processor. Compute Canada has a few systems with GPUs attached to a set of nodes, and there has been a continuous increase in uptake as researchers learn how to effectively exploit this technology. The 2015 RAC competition saw saturation of Compute Canada's GPU resources. Based on user data, we predict continued strong growth in adoption of accelerators, such as GPUs, in computationally intense research.

Security and Data Privacy

Compute Canada's ability to effectively serve researchers in medicine, social sciences and industry has traditionally been limited by our capacity to host and manage private data on a large scale. We are currently developing a comprehensive security framework to address policy issues related to hosting private data. However, our future facilities must also incorporate the necessary physical and network security arrangements required to implement such a comprehensive security framework. While not every system and data centre needs to be certified for highly sensitive data, there is an advantage in designing this requirement into multiple regionally distributed physical locations, to facilitate compliance with local regulations.

Memory

Compute Canada's current clusters are, for the most part, comprised of nodes with less than 4GB per core of system memory (RAM). The corresponding memory per node is therefore generally less than 100GB. We have seen a growing number of requests to provide single nodes with more than 256GB of RAM. This has been especially prevalent in image processing and bioinformatics applications, two rapidly growing areas.

How the Resource is Served - Virtualization, Cloud, Gateways and Portals

Traditionally, Compute Canada made resources available to the research community through traditional batch-oriented environments accessed through a command line interface. The nature of ARC has changed to include a growing need to provide virtual machines, cloud interfaces, scientific gateways, database services, web portals and data repositories. Compute Canada has attempted to address this growing need by offering virtualization on some existing hardware resources, and through the launch of a national Research Platforms and Portals competition. However, in some cases, this has meant meeting the researcher's need on non-optimal hardware resources. The new hardware systems need to be designed to better support these new resource delivery paradigms.

Stage-1 System Architectures

Stage-1 is intended to replace most of the computing capacity required as equipment installed in 2010 and before is defunded. This capacity will be replaced by a mix of systems designed to serve the diverse needs of the Canadian research community, as described above.

The final list of hardware to be purchased will only be known after a competitive procurement (RFP) process. Therefore, this document should not be interpreted as a precise list of hardware to be deployed by Compute Canada after stage-1 funding. Further, CFI has requested that Compute Canada provide 3 distinct options in its final proposal. Those options are not detailed herein and will only be created after the community consultation initiated by this document.

Types of Systems

In this document we describe two system types, distinguished in part by hardware and in part by the way they are configured and allocated. These types are:

1. **Large Parallel (LP):** a system optimized for running large message passing (e.g., MPI) jobs, focused on serving applications using 512 cores and more in any single parallel job. This type of system will have a high-speed interconnect and a relatively homogeneous set of nodes with relatively low requirements on memory/node.
2. **General Purpose (GP)**: a system optimized for running a wide range of applications including serial computation and parallel applications spanning a relatively small number of nodes. This type of system will be comprised of a heterogenous set of nodes (eg. some with large memory, some with GPUs) and will be well-suited to data-intensive applications.

The current proposal is to purchase 1 LP and 3 GPs in stage-1. The rationale is provided below.

Of the capabilities that will need to be replaced from the defunded systems, there is a need for at least one LP system. This system will have approximately 4 GB of RAM per processor core and have a homogeneous configuration. High performance storage suitable for the I/O requirements of the large parallel jobs will be purchased with this system as well as a small amount of mid tier storage for cost effective handling of data that can be staged to fast storage when needed. Buying a single system of this type ensures that it is as large as possible given the available budget to allow the largest scale jobs to be run. The scale of this system has been chosen such that it will have approximately the same number of cores as the largest of the current CC systems. The aim is to have this system mainly run large parallel jobs with smaller jobs run on other CC systems. This split in functionality allows for more efficient scheduling of jobs requiring larger numbers of cores than what we provide today. Users requiring parallel jobs at this scale on Compute Canada resources today include those performing computational fluid dynamics calculations (eg. aircraft design, plasma physics, stellar evolution), ocean and atmospheric modelling and some materials science calculations.

General Purpose (GP) systems will serve researchers with a wide range of needs, including those with very large data requirements. These researchers either run serial jobs, jobs that run on single computing nodes or jobs that use a small number of nodes for message passing applications. An increasing number of these jobs also require access to large amounts of data and have high I/O demands. These data centric jobs have placed considerable demands on the current systems and replacement systems must have suitable I/O performance to address this issue.

Many users have a mix of types of jobs that make up the overall workflow to produce their science output, leading to a preference for systems with a mix of capabilities at a single site. GP systems will be comprised of nodes of different memory sizes. Also, in order to address the changing needs of the CC user base, these systems will also include accelerators (e.g., GPUs) and the capability to support virtualization and containers. Currently the majority of jobs that utilize GPUs run on individual nodes, which is why the current recommendation is that the GPUs be placed in the GP systems. GP systems will also be designed to run virtualized environments and will host some shared block storage.

The GP systems will also be architected in such a way that at least two security zones are available on each system. These zones will permit isolation of users (and data) with stringent data privacy needs from general purpose usage. This will allow for some limited support of data with higher security requirements than is required for the majority of CC users. At least one of the GP systems may be designated for even higher security datasets.

While the LP system is designed to put the maximum number of cores at a single site, this consideration is less important for a GP system. In fact, several large current GP users require at least two geographically separate sites to ensure that large data stores are always accessible when a single site is offline. This means that the absolute minimum number of GP systems which can be purchased in stage-1 is 2, or else Compute Canada cannot transition these users from existing systems to the new systems. While two GP systems is the minimal option, it is inefficient to split the GP component of this purchase into many small pieces, as datasets would be unnecessarily split, it is more work to manage more systems, and it is more problematic to effectively schedule small systems. For this reason, we propose to buy 3 GP systems in stage-1.

Reference System Configurations

While the best technical solutions for delivering LP and GP systems within the available budget will only be known after a competitive bidding process, it is important to construct some reference configurations in advance. This allows budgets to be built and power and cooling requirements to be estimated.

For these reference systems, core counts and pricing are based on the Haswell Intel processor family and the NVidia K20 GPU accelerators. With Haswell processors the assumptions are:

- 12 cores per socket and 2 sockets per node (24 cores per node)
- FDR Infiniband (IB), 4 to 1 blocking (for GP systems)

Using 24 cores per node should replace close to the number of cores that will be defunded in 2016/2017 in the most cost effective way. The exact number of cores per socket and speed of the processors will be determined based on the best value and the technology solutions available at the time of purchase. High capacity interconnects are needed for message passing jobs and to support the large I/O requirements of nodes with large numbers of processor cores. A more performant interconnect than this will be needed for the LP system.

We also plan to have two visualization access nodes at each site. These will enable remote visualization of data created at that site without the need for users to move data to other systems. A small number of GPUs will be available at every site to accelerate these visualization sessions.

Example configurations for the systems is given in the table below. Such a configuration would have a total project cost of roughly $37.5M Canadian dollars (i.e. a $15M CFI contribution). This pricing is based on estimates blending recent purchases of large Intel Xeon “Haswell” based systems in the US, and using node pricing from previous Canadian based quotes, assuming that Xeon nodes will be priced similarly at the time of the stage-1 purchase. This should not be interpreted as a commitment to systems of this type - these systems are provided as examples to establish reasonable pricing and power numbers.

In addition to these systems, the budget includes upgrading the sites holding GP1 and GP2 to have 100Gb/s wide-area networking. Also, the budget includes the cost of two new tape libraries that should be located at sites with 100Gb/s networking so that it can be used to backup key data for all CC systems. Note that GP3 includes some accelerators with an alternative architecture than is used in the other systems to provide some diversity.

The total number of cores will be less than will be defunded in 2016/2017, but due to the increase of performance in CPUs and the addition of over 1000 GPUs, it will result in an overall increase in available floating point operations per second (FLOPS) across the CC platform. We believe that all of these systems should be installed as soon as possible in order for CC researchers to be able to take advantage of contemporary systems that are not currently available to them.

System	LP	GP2	GP3	GP1
Cores	30k+	16k+	16k+	10k+
Fast storage	3PB+	4PB+	4PB+	2PB+
Mid tier storage	2PB+	4PB+	4PB+	2PB+
GPUs	4 - for vis.	768	256	4 - for vis 32- alt. arch
Minimum Expected Power Draw (base kW, before overhead of cooling and other auxiliary systems)	540 kW	430 kW	360 kW	200 kW
Estimated Cash Purchase Price (net of vendor in-kind) ($million	9.27	8.28	7.13	4.97
Estimated CFI Contribution ($million)	4.64	4.14	3.56	2.49

Appendix: Systems Operational after Stage-1

Stage-1 will involve the purchase of 4 new systems and the de-funding of several current Compute Canada systems. The table below lists the existing shared systems which we expect to remain in service with some support from Compute Canada (in addition to the 4 new purchases) at the end of stage-1. Of course, system failures, and changing funding conditions (positive or negative) can change this list. Thus, at the end of stage-1, we expect more than 100K currently existing computational cores will remain in service. Concerns about the envisioned remaining systems should be raised as soon as possible.

System	Location	Commission year
glooscap	Dalhousie University	2012
mp2	Université de Sherbrooke	2011
cottos	Université de Montréal	2009
psi	Concordia University	2011
briarée	Université de Montréal	2011
guillimin	McGill University	2011/2013
helios	Université Laval	2014
East-cloud	Université de Sherbrooke	2014
sw	Queen's University	2008-2013
monk	Waterloo	2012
orca	Waterloo	2011
grex	University of Manitoba	2010
parallel	University of Calgary	2012
hungabee	University of Alberta	2012
jasper	University of Alberta	2009/2012
bugaboo	Simon Fraser University	2009/2011
orcinus	University of British Columbia	2009/2011
hermes	University of Victoria	2009/2012
West-cloud	University of Victoria	2014

Please note, that most systems in this table will be past their service contract dates (5-years from commissioning) on time scales relevant for stage-2 cyberinfrastructure funding. Therefore, the nominal expectation is to decommission those systems in 2018.

Appendix: Feedback to this Plan

Compute Canada welcomes written feedback to this plan through the sparc@computecanada.ca email address or during our January 2015 consultation sessions. All comments are welcome.

This appendix provides some examples of structured feedback that could be provided by researchers to influence the planning.

Science Competitiveness and Opportunities:

1. Does this plan provide you with the resources you need to remain competitive in your field over the next few years?
2. Are there near-term science opportunities that this refresh should be tuned to enable? In your opinion, would the existing plan enable researchers to take advantage of those opportunities?

General Hardware Questions:

1. Based on this draft plan, what (if any) hardware you require will be completely missing from the Compute Canada platform in 2017?
2. One of the hardest optimization questions for Compute Canada is the balance between CPU, GPU, storage, parallel computing, serial computing, etc. Should the balance of funding in the current draft plan be shifted? (eg. for your research you will require more GPUs).

General Questions About Your Future Needs:

1. What is your most pressing hardware need for 2017? Compute, storage, GPUs, high memory nodes, etc.?
2. What type of services are you looking for? VMs, interactive systems, secure data, archiving, etc.?

Storage Questionnaire:

Researchers can request at least four different types of storage:

- Fast Storage - Disk-based, suitable for processing large datasets as quickly as possible. High-throughput systems which feed CPUs in a highly efficient way.
- Normal Storage - Bulk storage. Can be disk or tape-based or a hybrid system which automatically migrates infrequently-accessed files to tape and pulls them back quickly as needed.
- Archival Storage - Usually tape-based. Generally either a second-copy with primary data stored elsewhere or the researcher is requesting storage of multiple redundant copies. May include geographic redundancy in some cases.
- Database Storage - Some researchers host large databases for a community of researchers. These databases require special servers and fast disk to optimize performance.

We would like to determine the breakdown of your storage needs as a function of time.

Please take the time to fill the table below as best you can.

	2014	2015	2016	2017	2018	2019	2020	2021
Fast								
Normal								
Archival								
DB								

We are also interested in your answers to the following questions related to data and storage:

- What are your data transfer needs? Do you need to transfer data over the network from, for example, a single global laboratory or observatory into CC systems? Do you need to transfer data to/from CC storage in order to process and analyze the data at a different physical location, and what are the locations?
- Do you need to share your data with others? If so, is it shared within the Canadian community (ie. everyone can have a CC account) or is it international? Do you allow anonymous access to your data? What is the size of the data and how does it need to be shared (technology)?
- Is any of your data subject to special privacy restrictions due to embedded personal information? What fraction of the data is subject to these restrictions?