compute canada | calcul canada

Compute Canada Technology Briefing

November 2017

1. Overview

This technology briefing is intended for Compute Canada stakeholders and suppliers. It provides a snapshot of status for the technology refresh program resulting from CFI's cyberinfrastructure initiative, being implementation from 2015-2019. It also describes plans for future growth.

This document builds on prior versions (2015 and 2016[1]). It is intended to be useful and accurate at the time of writing, but does not guarantee any particular actions or outcomes. All activity within Compute Canada is subject to oversight by the Canada Foundation for Innovation (CFI), as well as member institutions, regional partner organizations, and other parties.

2. Compute Canada Overview

Compute Canada (CC) exists to enable excellence in research and innovation for the benefit of Canada, through effective, efficient, and sustainable deployment of a state-of-the-art advanced research computing (ARC) infrastructure, supported by world-class expertise across multiple disciplines. Compute Canada's focus is on the seamless integration of technology and services, working to make them readily accessible and useful to researchers across Canada. CC is an essential component of the multi-layered digital research infrastructure (DRI) in Canada.

Compute Canada focuses its activities in four strategic areas[2]:

- Providing excellent research support services;
- Building research capacity;
- Creating a positive effect on Canadians through the research it supports; and,
- Aspiring for organizational excellence and sustainability.

[1] Compute Canada Technology Briefing 2016

[2] Compute Canada Strategic Plan 2014-2019, July 2014

Several core elements are required in order to achieve these strategic objectives. These include: state-of-the-art advanced research computing (ARC) infrastructure; highly qualified, client-focused personnel; a secure environment; and a commitment to professional and organizational excellence.

Compute Canada supports projects at scales ranging from a single faculty member, up to the largest “Big Science” projects in the country. The CC “Facility” refers to the collection of ARC infrastructure and services managed by the federation, distributed in various geographic locations in Canada. The Facility enables internationally competitive research in many disciplines. The Facility also supports researchers from large and small research institutions from coast-to-coast, across many sectors. The diverse community needs are addressed through CC’s delivery of services, which include traditional tightly coupled High Performance Computing (HPC), High Throughput Computing (HTC), big data and data analytics, serial computing, cloud computing and storage, visualization, and other technology solutions.

The Compute Canada federation operates as a member institution. A not-for-profit corporation serves as project manager, and the membership includes the majority of Canada’s research universities, along with a number of other research institutions. As of late 2017, there are 37 member institutions, spanning four geographic regions.

3. Technology Refresh

The Canada Foundation for Innovation (CFI) is the major funding agency for the Compute Canada project. Funding is matched by provinces, member institutions, and other sources. CFI is funded through Canada’s Ministry for Innovation, Science and Economic Development (ISED). Compute Canada is one of 17 projects under CFI’s Major Science Initiatives (MSI) program. CC began its second 5-year round of funding in 2017. The MSI provides major support for operations and maintenance of the Facility.

A major technology refresh program is underway, through CFI’s Cyberinfrastructure Challenge 2 program. Challenge 2 Stage 1 is nearing completion, and Stage 2 is getting underway. Together, they represent investment valued at $125M across approximately three years, which will replace nearly all cyberinfrastructure within the Facility.

As part of the Challenge 2 effort, CFI has guided consolidation of the number of data centres. Prior to Challenge 2, there were 27 member institutions that operated cyberinfrastructure (with 58 systems across Canada). The new cyberinfrastructure purchased through Challenge 2 Stages 1 and 2 is being installed at 5 new national hosting sites.

Through a formal competition among Compute Canada member institutions, four sites were selected to host the Stage 1 systems and associated services: the University of Victoria (UVic), Simon Fraser University (SFU), the University of Waterloo (Waterloo), and the University of Toronto (UofT). Table 1 provides a summary of Stage 1 technology investments.

Table 1: Stage 1 Investment Summary
(includes cash expense; in-kind contributions add a further 20%). Total value of $75M.

UVic	Compute (GP1)	$4,390
	Storage & infra	1,610
SFU	Compute (GP2)	11,550
	Storage & infra	5,150
UW	Compute (GP3)	10,615
	Storage & infra	4,985
UT	Compute (LP)	16,850
	Storage & infra	2,850
Software and Services	All 4 sites	2,000
	Grand total	**$60,000**

For Stage 2, another competition was held to identify one or more additional national hosting sites. Compute Canada submitted the Stage 2 proposal[3] in May 2016, and was informed of results in September 2016. CFI has not yet announced outcomes of the Stage 2 competition; therefore, this document does not name the selected sites. However, planning is underway for the Stage 2 technology refresh to be substantially complete by September 2018.

Table 2: Stage 2 Investment Summary
(includes cash expense; in-kind contributions add a further 20%). Total value of $50M.

GPx expansion	Compute	$9,283
	Storage & infra	3,468
GPx expansion	Compute	0
	Storage & infra	1,000
GP4	Compute	16,788
	Storage & infra	7,463
Experimental systems	All 3 sites	1,500
Software and Services	All 3 sites	500
	Grand total	**$40,000**

[3] Cyberinfrastructure Challenge 2 Full Report

Stage 1 yielded four major new computational systems and associated storage. Two sites, UW and SFU, were identified as the “deep storage” sites, to receive backups and other long-term data. All four sites worked together to foster development of research data management software, middleware, and related services to improve Compute Canada’s capability to meet known needs of users. Stage 1 also included emphasis on storage, with 20-24% of all expenditures devoted towards persistent (disk or tape-based) storage. Various infrastructure services, such as a national LDAP, and centralized monitoring systems, were also deployed as part of Stage 1.

For Stage 2, a fifth national hosting site will be added. This will host another large “general purpose” (GP) system, with a variety of node types, and is intended to meet a wide range of computational needs, with job sizes up to approximately 1024 CPU cores. All GP-type systems include cloud partitions, and GP1 (at UVic) is entirely focused on OpenStack cloud services.

Stage 2 will also see a large expansion of one of the Stage 1 GP-type systems. Storage expansion, including capacity for backups and nearline use, will be included. Software and middleware for research data management and other purposes will again be included. A new category of investment is included for experimental systems, which will allow relatively small investments in technologies that are of interest to one or more current users, and which will help to inform future technology choices.

Table 3: Stage 1-2: Major Computational Systems

System	**Description**	**Production Date (planned)**
(GP1) Arbutus - University of Victoria	OpenStack cloud environment. 7000+ CPU cores. Persistent & on-demand VMs.	September 8, 2016
(GP2) Cedar - Simon Fraser University	General purpose cluster. 27K+ CPU cores, 580+ GPU devices. Many node types.	July 25, 2017
(GP3) Graham - University of Waterloo	General purpose cluster. 33K+ CPU cores, 360 GPU devices. Many node types.	July 28, 2017
(LP1) Niagara - University of Toronto	Large parallel system. 60K cores, Skylake.	(Early 2018)
GP4 - Site 5	General purpose cluster.	(Mid 2018)

4. Major Components of the National Platform

The following sections provide some detail on the major components of the Compute Canada technology refresh, which comprise Canada's updated national platform for advanced research computing. This Briefing does not include the full background of user need and predecessor infrastructure choices that led to these plans. That background is included in the MSI and cyberinfrastructure proposals, which are available at www.computecanada.ca/publications.

4.1 General Purpose Clusters

Prior to the Stage 1 submission, Compute Canada engaged in nation-wide consultations to understand the needs of current and future users. These consultations were intended to identify pain points or shortfalls in the national platform as it existed then, and also to forecast future needs for infrastructure. One of the major findings was that substantial need exists for computational systems designed for diverse workloads, from single cores to around 1000 cores. Different node types were needed, including GPU nodes, large memory nodes, and nodes with local (on-node) storage. Systems of this type were labeled "general purpose" clusters, and Stage 1 resulted in two large GP systems, and one smaller one focused mainly on running an OpenStack cloud.

Arbutus (University of Victoria): Arbutus (previously known as "GP1") is an OpenStack cloud, with emphasis on hosting virtual machines and other cloud workloads. The system has 6,944 CPU cores across 248 nodes, each with on-node storage and 10Gb networking. It accesses 1.6PB of persistent storage, primarily via Ceph in a triple-redundant configuration. The system became operational in September 2016, as an expansion to the Compute Canada "Cloud West" system. In 2018, expansion by approximately ⅓ of the Arbutus compute and storage capacity will take place.

Cedar (Simon Fraser University): The Cedar system (previously known as "GP2") is a heterogeneous cluster, suitable for a variety of workloads. The system is liquid cooled and is installed in the newly renovated SFU Water Tower data centre. It has over 27,000 CPU cores, and 576 GPU devices. Node types include "base" and "large" compute nodes with 128GB and 256GB of memory, as well as bigmem nodes with 512GB, 1.5TB and 3TB of memory. More information: https://docs.computecanada.ca/wiki/Cedar

Graham (University of Waterloo): The Graham system (previously known as "GP3") has a similar design to Cedar, as well as a shared software stack and other configuration consistencies to provide workload portability for and ease of use. Like Cedar, Graham hosts a small OpenStack partition and has various diverse node types. Graham provides over 33,000 CPU cores and 320 GPU devices. Islands in the topology are 1280 cores, and one island has 4096 cores. More information: https://docs.computecanada.ca/wiki/Graham

GP4 (Site not yet announced): The GP4 system will have a similar profile to Cedar and Graham and will have a very consistent look and feel for users. It will include various node types, including nodes with accelerators. It will provide a small OpenStack partition, federated with the other GP-type systems. Full production is targeted for the 2nd half of 2018.

4.2 Large Parallel Computing

Many users run jobs that are larger than the target designs of Cedar, Graham or GP4. The "large parallel" system profile is designed to support these larger jobs, as well as being suitable for smaller jobs, "big data" workloads, and other purposes. Stage 1 funded the Niagara system, which is the only system of this class which CFI has supported so far within Challenge 2.

Niagara (University of Toronto): Niagara (previously known as "LP") will be deployed in late 2017, for production by April 2018. It will have 60,000 CPU cores, a burst buffer storage layer, and a high performance interconnect with adaptive routing. More information: https://docs.computecanada.ca/wiki/Niagara

4.3 National Data Cyberinfrastructure

The national data cyberinfrastructure (NDC) provides user-facing capacity for online storage of various types. It also provides networking among the sites to enable data transport for a wide range of needs. Software/middleware, such as Globus, is utilized to facilitate data access and movement.

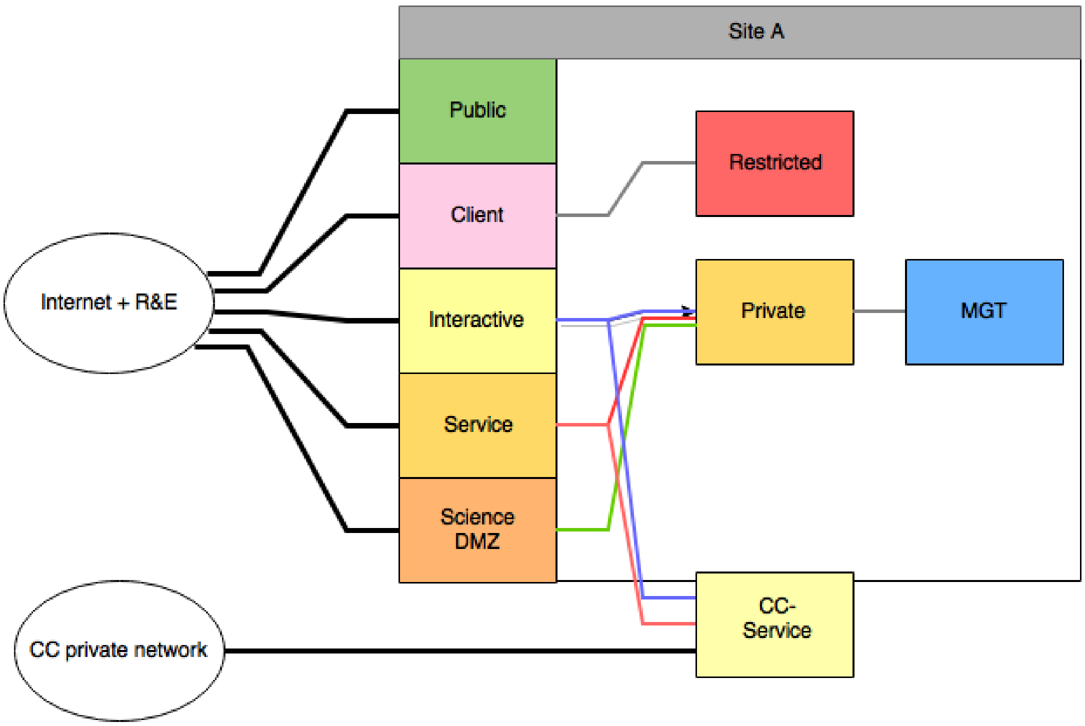
All five national hosting sites deploy elements of the NDC, with specific configurations and capacities dependent on each site's workload. Total online storage capacity from Stage 1 reached approximately 62PB, and Stage 2 will see growth to approximately 100PB. Reliance upon commodity storage building blocks (SBBs) and software-defined storage enables efficiency of management, cost-effective capacity, and flexibility for different storage types and configurations.

For Arbutus, the largest capacity is for Ceph storage, in support of OpenStack. Graham, Cedar and GP4 have (or will have) project space (i.e., Lustre "/project") of 20-30PB, and various special-purpose storage including database servers, Ceph, dCache, etc.

4.3.1 Network Architecture and Science DMZ

As a component of the NDC, sites purchased new edge routers to connect at 100Gb to the regional network and from there to CANARIE, the national research and education network. A network architecture is being implemented to provide needed service isolation, performance and routing. The Science DMZ concept is applied to keep traffic from some services within a network zone that is routed over the wide-area network, but not externally.

Figure 1: Network Architecture Zones



4.3.2 Deep Storage

Deep storage refers to storage optimized for costs, rather than performance. It includes nearline storage (automated or semi-automated disk-to-tape) and backups.

Three deep storage sites are included in Challenge 2: SFU, Waterloo, and the new GP4 sites. Each will have 80+ PB of tape capacity for backups, archival copies and similar purposes. Generally, backups will include an off-site second copy at another deep storage site.

4.3.3 Capacity Management

All Stage 2 sites will participate in a new consortium RFP (or similar selection process) for capacity management solutions. Capacity management solutions were an intended part of the 2016 Storage RFP (as HSM, hierarchical storage management), but were not purchased at that time. Now that other elements of the NDC and deep storage are known and form the basis of storage capacity for Stage 1 and Stage 2 sites, specifications for capacity management solutions will be more attuned to the actual technology mix.

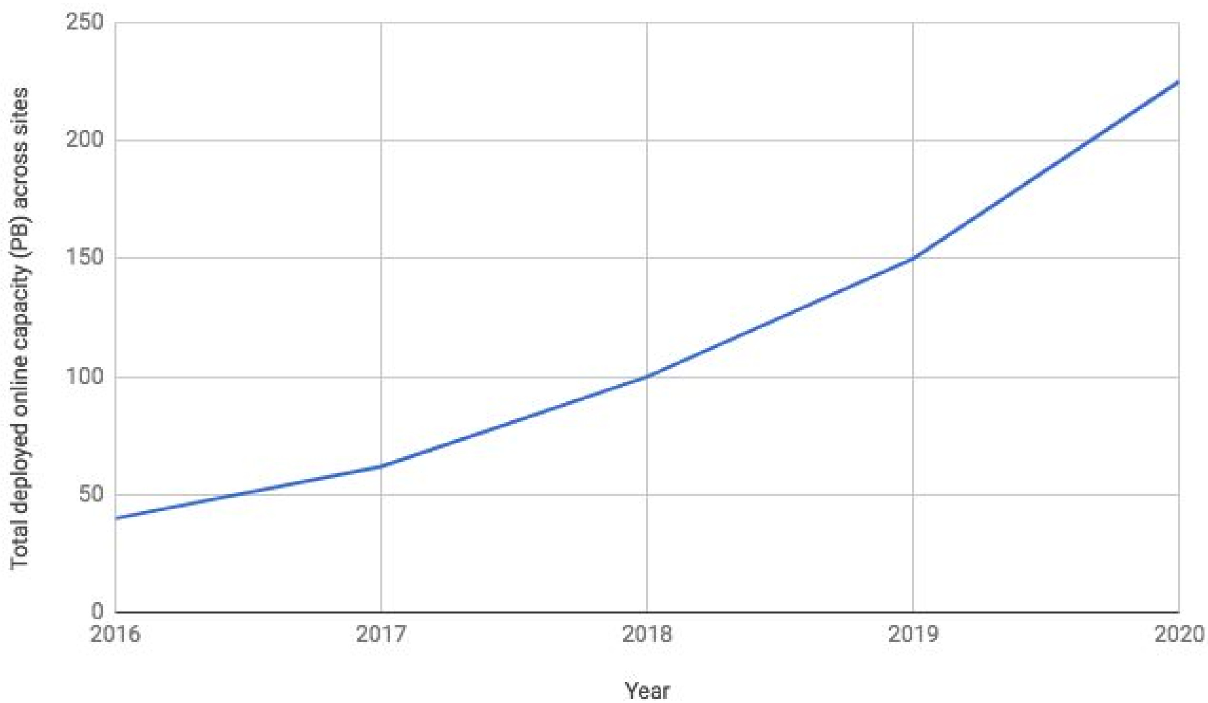
The idea of capacity management is to offload capacity from more expensive online modalities, to less expensive nearline modalities. In practice, this involves automated relocation of data from disk to tape, but there are numerous approaches to implementation. Multiple solutions may be required, to address the full span of need.

Within Compute Canada, needs for capacity management include:

- For large filesystems (mainly Lustre, also some GPFS), to automatically migrate data to and from tape, based on policy. Policy will be implemented based on characteristics such as time since last use, size, and user or group quotas. This is the traditional domain of hierarchical storage management (HSM) systems and may be similarly applied to other filesystem types.
- For object storage (S3 or similar mechanisms), to provide backup mechanisms, or mechanisms for data replicas to include an offline/nearline copy.
- For "cold storage" of data objects, such as when a user desires to check-in datasets for long-term archival purposes.
- Performance tiering, by which data objects move from high performance tiers, to mid-performance, to nearline.
- Data replication and load balancing, by which multiple copies (online and nearline) may be maintained efficiently across multiple sites.
- Data availability, stage-in, and stage-out, by which data objects will become available locally with high performance for computation, and then relocated afterwards.

Because of the large capacities involved across sites, and large number of users, Compute Canada prefers bulk licensing models. CC is also investigating free or low-cost alternatives, even when some level of staffing effort is required. Considerations involve modeling the total cost of ownership (TCO) for different ranges of capacity.

Table 4: Online Storage Capacity Projections



4.4 Experimental Systems

All three Stage 2 sites will host experimental systems. A lightweight proposal process will be utilized at each site to balance immediate utility against expense and the needed level of Compute Canada effort. The desire is to identify, rapidly deploy, and assess promising technologies, as well as to be responsive to requests from users for new technologies that were not previously identified as part of Stage 2.

Technologies of interest identified in the Stage 2 proposal include:

- commercial cloud integration and cloud bursting;
- next-generation computing technologies including quantum computing;
- ultra-fast storage tiers;
- accelerators and reconfigurable devices for computation, storage and networking;
- Cybersecurity; and,
- software and solutions for data analytics, deep learning and disciplinary research.

Experimental systems can include purchases of equipment, software or services. Equipment may be leased or located off-site, if appropriate. Expenses can also include personnel, training, or other necessary elements to enhance utility of the experimental system.

4.5 Services Infrastructure

Further investments in the service infrastructure development efforts funded through Stage 1 are planned. In Stage 1, this was labelled “Research Data Management.” Both stages reflect the need for software, middleware and services to meet a variety of needs. These activities are mainly focused on software development within Compute Canada, often leveraging existing software.

The philosophy is that if multiple users/groups express a need for a service, such as via user surveys or white papers, then Compute Canada should consider making that a national offering. Activities will include: implementing single sign-on technologies for authentication, authorization, and identity management; deploying monitoring and resource publication services; reporting; data transfer services; batch and cloud scheduler integration; metascheduling; workflow management for compute and data; and, research data management.

5. Vision 2020

Compute Canada, as a leading provider of digital research infrastructure (DRI), is taking an integrated approach to data and computational infrastructure in order to benefit all sectors of society. As a result of the technology refresh and modernization supported by CFI’s Challenge 2 Stages 1 and 2, world-class Canadian science will benefit from modern and capable resources for computationally-based and data-focused research.

Compute Canada is cooperating with government funding agencies and with other digital research infrastructure (DRI) providers to provide the world’s most advanced, integrated and capable systems, services and support for research. Future researchers will have seamless access to DRI resources, integrated together for maximum efficiency and performance, without needing to be concerned with artificial boundaries based on different geographical locations or providers.

By 2020, Compute Canada will offer a comprehensive catalog of resources to support the full data research cycle, allowing researchers and their industrial and international partners to compete at a global scale. In cooperation with Canada’s other DRI providers, Compute Canada’s systems and services will facilitate workflows that easily span different resources: from the lab or campus, to national computational resources, analytical facilities, publication archives, and with collaborators. Local support and engagement will remain a hallmark of delivering excellent service to all users. The pathway to this future has begun, with the modernization of Compute Canada’s national data cyberinfrastructure through the CFI Challenge 2 investments.