compute
canada

Part B: CFI Cyberinfrastructure Initiative Challenge 2

Stage 2 | Proposal

Contents

1. Need for the Infrastructure

As demonstrated in Part A, Compute Canada (CC) supports a vibrant program of research spanning all disciplines and regions of Canada. This support is delivered by providing Canadian researchers access to world class infrastructure and expert personnel.

The advanced research computing (ARC) needs of the Canadian research community are growing. Growth comes from new scientific instruments and experiments, from use in a broadening list of disciplines, from generation and access to new datasets and the innovative analysis and mining of those datasets, and from the mutual reinforcement of technological and scientific advances that inspire researchers to construct ever more precise models of the world around us. Canada's ARC infrastructure needs constant update, to keep pace with the needs of its researchers.

In order to promote effective and efficient use of new infrastructure, CC will continue to adapt service offerings to modern workloads. This will include ease of access to well managed systems, services and support, as well as leading edge software environments and data management tools on a national network of facilities.

On behalf of the Canadian ARC community, CC has constructed three options for infrastructure refresh for CFI's "Challenge 2 stage 2" Cyberinfrastructure Initiative. These options include a preferred option, which provides a balanced approach to meeting the most pressing needs of Canadian ARC community. In addition, two alternatives are presented, which each favour a certain technological direction which would benefit specific science and application use cases.

1.1. Existing Usage Information

CC has studied usage information collected over the last 5 years. For example, the chart below shows CPU usage from 2010 through the end of 2015. CPU usage and allocations of computational resources are measured in core years, representing a single CPU core's utilization for one calendar year. The different colours show the usage broken down by discipline. The decrease in 2015 was expected, due to a decrease in available compute resources as CC needed to decommission older systems which have exceeded their normal life span (the largest single system contributing to this supply is now 7 years old). This chart illustrates that a significant number of different disciplinary areas share the CC facility, each bringing their own resource needs.

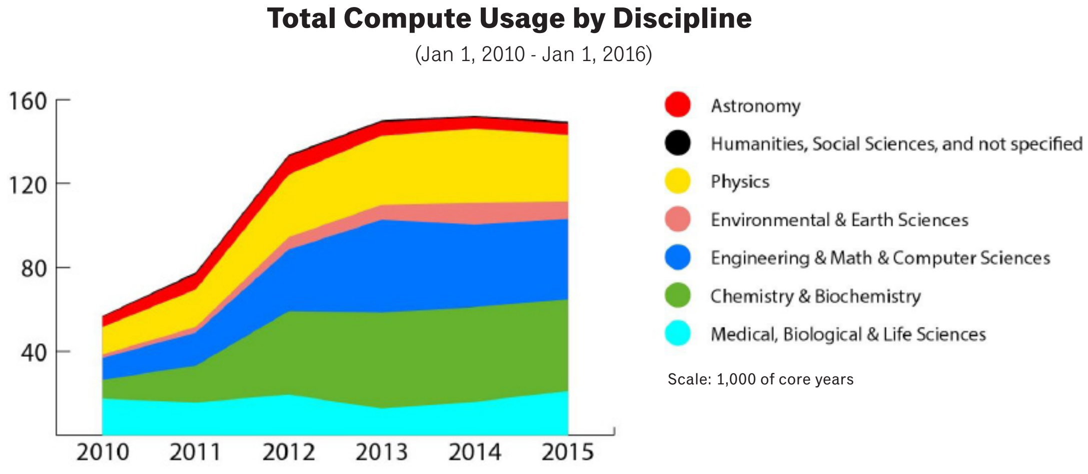
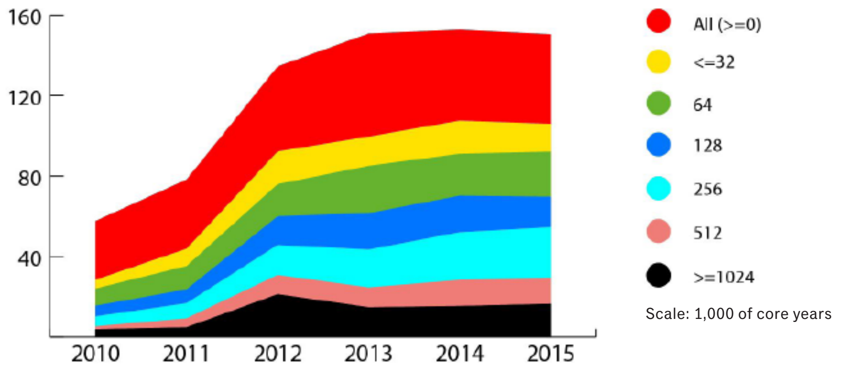


Figure 1.1: CPU usage by discipline as a function of time

CC supports a wide range of computational needs on its shared infrastructure. One way to examine this is through the number of cores used in a single batch job, which is the dominant method for use of these resources, and the types of jobs for which resources allocations are granted (contributed systems, cloud systems, platforms & portals, and other modalities are mentioned below). The chart below shows the number of core years used in CC per year. The colours illustrate the fractions of those core years in bins of cores-per-job. It shows, for example, that the largest single category in 2015 is serial or low-parallel computation (fewer than 32 cores), which represents about 30% of the total. Meanwhile, nearly 50% of CPU consumption in 2015 was by jobs using at least 128 cores.

Core Years Used, Binned by Core Used per Job



Scale: 1,000 of core years

Figure 1.2: CPU usage binned by number of cores used per job as a function of time.

It should be noted that the size and configuration of CC's current systems limits the ability of Canadian researchers to submit jobs at the largest scales, and this has limited the growth of the highly parallel bins. Even for the larger resources, queue wait times (via the "fairshare" workload management policy in effect for most systems) would create challenges for completing large multi-job computational campaigns.

As noted in Part A, the overall capacity within Compute Canada is currently inadequate to meet the growing need of the Canadian community. After technical validation, for 2016 CC was only able to allocate 54% of the requested computational resources (down from 85% in 2012) and 20% of GPU requests. With respect to storage, 93% of requests were granted in 2016, although this was enabled by deferred allocation of storage to as-yet-uninstalled Stage 1 resources. Without Stage 1 storage, the allocation rate for storage would have been 65%.

1.2. The Future Needs of the Canadian ARC Community

Extensive community consultation was undertaken to ensure that this proposal is anchored in the anticipated future needs of the Canadian ARC community. Consultation, described in Part A, included in-person community meetings, online surveys, collection of community white papers, and user interviews.

The aggregated community need for computational resources has been projected as described in Part A. Survey analysis predicts 12x growth in computational need over 5 years, while the white paper analysis predicts a 7x increase over the same period, with different annual increase rates among submissions. The chart below shows these need projections assuming an exponential growth profile, in units of allocatable Haswell-equivalent core years. The shaded band covers the range between the 7x and 12x 5-year projections. Three supply curves are shown: 1) (red) assuming only the Stage 1 award, 2) (light blue) assuming the Stage 1 award and success of this Stage 2 proposal, and 3) (dark blue) projecting a $50M Stage 3 award in 2018.

While Stage 3 is not a confirmed competition at this time, it is a planning assumption used throughout this document based on information from CFI in February 2016. This chart shows that Stage 1 alone leads to a short-term increase in core count (by about 50%), followed by a marked decrease based on the decommissioning of older pre-existing systems over the next 2 years. Stage 2 funding will lead to an approximate doubling of allocatable cores by 2019 with respect to the baseline. Stage 3 funding could allow the supply to approach the need curve in 2019.

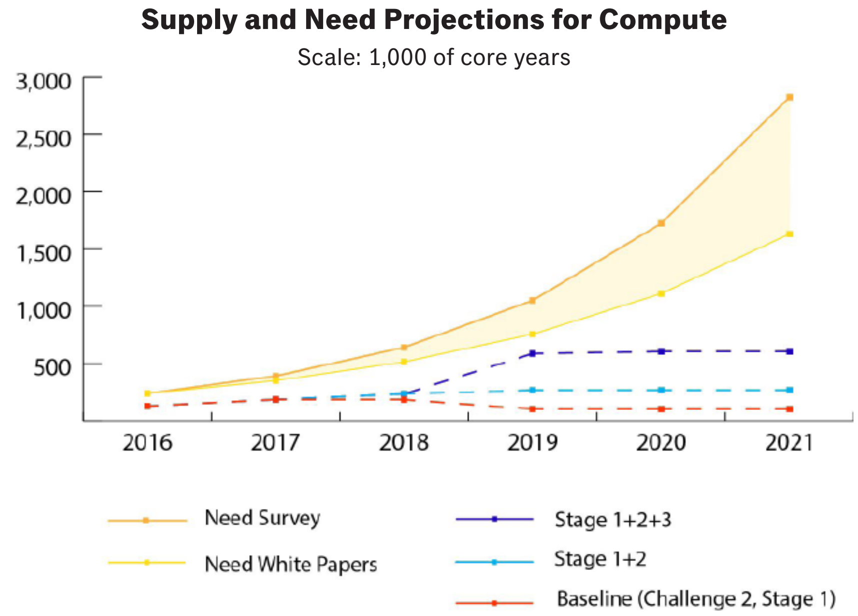


Figure 1.3: Supply and need Projections for Compute (core years)

The aggregated need for storage resources has also been projected, as described in Part A. The survey analysis predicts 19x growth in storage need over 5 years, while the white paper analysis predicts a 15x increase over the same period. The storage projection chart below shows the 15x-19x range for the three stages of investment described above. The storage supply and need both represent allocatable disk storage. However, replication factors, potential future object storage adoption rates and usage of tape (for nearline storage) to alleviate disk need are difficult to predict prior to Stage 1 storage deployment and evaluation of the rate of adoption of object storage. As a result, in the chart below we scale raw storage supply downward by a factor of 1.4 (i.e. we assume approximately 70% disk usage efficiency).

In addition to aggregated need information, the user survey responses revealed specific requests for additional features, new architectures and special node types. They include requests for:

- Overall increased compute capacity,
- Better support for Big Data use-cases,
- Encrypted cloud storage and other steps to enable research on sensitive datasets,
- Increased access to large memory nodes,
- Specialized resources to support bioinformatics,
- Greater accelerator (e.g., GPU) capacity,
- Better support for interactive and visualization-focused use cases,
- Better support for long-term data storage and enterprise-class data backup,
- Platforms to support new hardware development (IT and computer engineering-related research),
- Increased training and improved documentation.

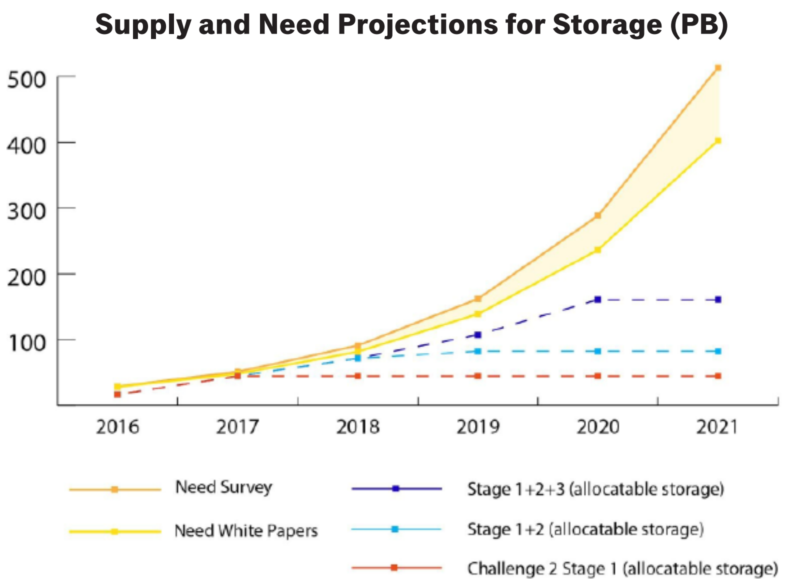


Figure 1.4: Supply and Need projections for compute (in core-years/year) and storage (in PBs).

White paper submissions also revealed a number of emerging trends that help to drive the technology choices laid out in this proposal. These include need for:

- Large data storage driven by improved instrumentation in genomics, neuroimaging, astronomy, light microscopy and subatomic physics,
- Large memory nodes (at least 512GB) from astronomy, theoretical subatomic physics, quantum chemistry, some use-cases in bioinformatics, humanities, some use-cases in AMO physics, and institutional responses,
- Expanded accelerator capacity (primarily GPUs) from subatomic physics, chemistry, artificial intelligence,
- Robust, secure storage options from the digital humanities,
- Expanded cloud services from digital humanities and astronomy,
- Expanded capacity for tightly coupled processing, including jobs that exceed 1,024 cores.

There is evidence of need for systems with far larger homogeneous partitions than reflected in Stage 1 planning for the LP system. This includes researchers who have offshored or outsourced their computation away from CC resources.

In 2016, the SCINET consortium (Ontario) contacted 58 Canadian faculty members who run large parallel jobs. Respondents were primarily users who had submitted at least one job requiring at least 1,024 cores, either on CC resources, the 66,000 core Blue Gene/Q at SOSCIP, or international facilities. Of these, 26 were interviewed to discuss their usage patterns and future needs. If resources were available today, in total they would use approximately 250,000 cores per year on a homogeneous, tightly-coupled, large parallel machine - with much larger jobs, and requiring many more cores, than the LP system planned via Stage 1 for mid- to late-2017 (below). One individual within the group had already run a 330,000 core job on the Tianhe-2 machine (China), and expressed need to scale to 1M cores in the future. Due to lack of current availability in CC, it can be reasonably assumed that the LP demand is significantly underestimated and researchers are tailoring their areas of investigation and ARC usage to maximize their productivity. It is envisioned by Compute Canada that, over time, larger systems with larger homogeneous partitions will be provided, thereby enhancing the ability for users to pursue larger-scale investigations.

1.3. User Community Growth and Diversification

Goal number 1 of the Compute Canada Strategic Plan is to, “Provide services which enable excellent Canadian research across a broad range of disciplines.” Social Sciences, Humanities and Health Sciences were explicitly named in the Plan as areas in which CC would need to expand support. In the two years since the strategic plan was written, the adoption of ARC methods by those communities has accelerated, reinforcing the need for new services to support them.

An area of need that results from some of these disciplines is for partitioning of resources to better support isolation of computation and data within shared systems. This reflects the needs of research in human health, human behavior and cultural traditions, criminology, and others with need to work under stringent privacy and security regimes. Expansion of cloud computing capabilities for Stage 2 include building on the success of sites already engaged with such researchers, to deploy shared secure cloud computing as a component of all Stage 1 and 2 cloud partitions. Over the last 3 years, several (at least 5) projects funded through CFI’s John Evans Leadership Fund were not integrated into the CC platform specifically due to privacy and security concerns which could be addressed by such a model; in the future, the needs of those and other researchers will be able to effectively utilize CC.

Adopting an enhanced security posture (both in policy and technology) is vital to supporting social science researchers. For example, over the last year, CC has engaged in detailed discussions with the Canadian Research Data Centre Networks (CRDCN) and Statistics Canada concerning access by researchers to Statistics Canada survey and administrative data. CC is assisting with the design of the refresh of CRDCN’s platform, specifically its transition to a thin client environment. The ability to provision needed computational resources to the social science community will open new avenues of study and constitute a major new service to thousands of researchers across the country.

Beyond the issues of security and privacy, the broadening of the user base brings the need for enhanced middleware services. In Stage 1, CC set aside $2M for the development of services in support of services infrastructure, via efforts in Research Data Management (RDM), user authentication, authorization and ID management services, file transfer services, resource monitoring, and publication services. These services are currently being built and deployed during the course of Stage 1. Further development of services is envisioned for Stage 2 in support of the broadening community supported by CC, as described below.

1.4. Services Infrastructure Development in Support of CFI Challenge 1 and Broader Community Needs

CC has a special role to play in support of CFI Cyberinfrastructure Challenge 1 projects. These projects will implement data platforms on CC resources. The first 7 successful projects have been identified but not yet publicly announced at the time of writing. CC consulted with all 18 projects invited to submit proposals in 2015, and is contacting the successful projects to discuss any evolution of their hardware, middleware and software needs. The seven projects represent relatively modest compute needs, but 11-13 PB of aggregate storage need. These projects will benefit from services such as a common authentication, authorization and ID service, data movement services, standard and secure cloud services, monitoring, and resource publishing services. The projects require a diverse set of node types and are well-suited for general purpose cluster architectures.

Challenge 1 needs are reflective of broader community needs. As planning and implementation for Stage 1 has advanced, many user communities are seeking benefits for ease of use, better time-to-solution, workload portability, and software flexibility. Many of those needs are met by access to cloud resources, which let experienced users deploy and manage their own virtual machines. Other uses require significant developer resources to design, implement and sustain the needed research software and computational environment.

The CC National Teams (described further in the accompanying Management Plan) will provide some of the needed new services infrastructure as part of their normal operations. Some of the most relevant teams for these efforts include Cloud Operations, Storage National Team, Monitoring National Team, and Platforms & Services National Team. For Stage 1, services infrastructure investment includes procurement of software, as well as software developer personnel. Similar needs are anticipated during Stage 2.

2. Capital Investment Options: Current Status

This Part B proposal is the capital infrastructure request for the MSI described in Part A. Capital investment is requested to add and extend advanced research computing systems, associated storage, and other systems and services to the portfolio overseen by Compute Canada in the MSI. This $20M Stage 2 proposal builds on the $30M Stage 1 cyberinfrastructure investment announced in July, 2015. Additional details concerning Stage 1 were published in a 2015 Technology Briefing (www.computecanada.ca/featured/compute-canada-technology-briefing/), with additional updates online under www.computecanada.ca/renewing-canadas-advanced-research-computing-platform/. A brief synopsis follows.

GP1, hosted by the University of Victoria: An OpenStack cloud system with approximately 6,000 CPU cores and 5 petabytes (PB) of disk for persistent storage. Both system and storage are being purchased, as of late May 2016, and will be deployed in early summer 2016. An expansion to compute and disk resources of ⅓ of the total expenditure is planned for mid-2017. Total CFI funding: $3M.

Persistent storage, for all Stage 1 and 2 systems, is not inclusive of temporary/scratch disk, which will be purchased with the system. All core counts in this proposal are based on Haswell; actual core count will vary proportionally.

GP2, to be hosted by Simon Fraser University: A large heterogeneous cluster with approximately 25,000 CPU cores, 1,536 GPU devices, and 15PB of disk for persistent storage. “GP” for GP2 and GP3 refers to general purpose clusters, with various node types intended for diverse workloads. SFU is a destination for backups and deep storage (see below). Storage is being purchased, and the system RFP is nearly final, with installation and commissioning expected Fall 2016. Total CFI funding: $8.35M.

GP3, to be hosted by the University of Waterloo: A large heterogeneous cluster with approximately 19,000 CPU cores, 192 GPU devices, and 15PB of disk. UWaterloo is a destination for backups and hierarchical storage management. Storage is being purchased, and the system is scheduled for purchase in October 2016. An expansion to compute and disk resources of ⅓ of the total expenditure is planned for mid-2017. Total CFI funding: $7.8M

LP, to be hosted by the University of Toronto: A large parallel supercomputer with a balanced high performance interconnect, anticipating large homogeneous partitions with one or two node types, possibly including accelerator/manycore. Approximately 66,000 CPU cores will be purchased in mid/late-2017. 5PB of persistent storage is being purchased, and a further 10PB will be purchased with the system itself. Total CFI funding: $9.85M.

Service Infrastructure Development: Research Data Management (RDM) and other software infrastructure investments in support of all systems and services, including the needs of CFI’s Challenge 1. Personnel and software expenses to be shared across the four Stage 1 sites. Total CFI funding: $1M.

Infrastructure systems: High-availability servers for critical infrastructure services have been deployed at SFU and uWaterloo. All four sites are receiving new network routers to support a multi-site Science DMZ (see fasterdata.es.net), software-defined networking, and 100Gb connectivity. Funding included in site totals above.

3. Capital Investment Options Introduction and Common Elements

Concepts in this Stage 2 proposal were derived from the Stage 1 program, adjusted with a deeper understanding of user needs as described above and in Part A. In preparation for this proposal, Compute Canada ran an open RFP for Stage 2 hosting site institutions, in which hosting proponents presented their analysis and ideas, often augmenting earlier understanding and introducing new ideas for Stage 2.

3.1. Stage 2 Elements: Systems, Storage and Software

GP1x: OpenStack cloud system. Building on the successes of Cloud East (Sherbrooke) and Cloud West (UVic), the goal is to have a federated cloud across Compute Canada. The cloud primarily provides Infrastructure as a Service (IaaS) and Platform as a Service (PaaS), to a rapidly growing constituency. Software as a Service (SaaS) is also available, and expected to grow. Cloud federation will benefit users with workload and storage portability and resiliency, single sign-on and namespace, and a common software stack. The Stage 2 hosting RFP solicited a GP1b system, to complement the UVic system (now denoted GP1a).

GPx: Heterogeneous cluster with elastic OpenStack partitions. Clusters with a variety of node types, including nodes suitable for OpenStack, big memory nodes, and nodes with GPUs; most nodes will have local storage. The high-performance interconnect might not be fully balanced for all nodes, but will have some partitions suitable for multiple jobs of at least 1024 cores. The systems will grow over time as funding allows, including via contributed systems. The Stage 2 site selection RFP solicited GP4/GP5 hosts, as well as possible expansion of GP2/GP3 at SFU and uWaterloo.

LPx: Large parallel supercomputer. An expansion of the LP to be deployed via Stage 1. Only UofT was permitted to propose expansion/update of LP.

Experimental systems: In service to Compute Canada's drive to provide access to new resource types, it is envisioned that a number of relatively small experimental systems will be deployed. These may be purchased, loaned, or developed, over different durations. Some experimental systems may become production resources, or guide future procurements of larger systems. Stage 2 site selection proponents could select this as an additional option to the main three system types. Compute Canada has been in communication with numerous vendors who may participate in an experimental system program (including D-WAVE, Huawei, IBM, Intel, Obsidian, and PHEMI).

A component of experimental systems is commercial cloud hosting. CC is often asked about outsourcing to commercial hosting services. To explore this area, CC can run an open RFP to select one or more in-Canada cloud providers, and then work to develop easy mechanisms for users to span their workload among CC resources and commercial clouds. Microsoft, Amazon, HP and Softlayer (IBM) all have cloud offerings based entirely in Canada, and each has expressed interest in working with CC on this initiative. The cost/node for individual purchases of cloud computing is high (at least 4x greater than CC's in-house systems for retail pricing, although this may be mitigated via an RFP for bulk purchase and partnership), so this resource must be deployed carefully. At the same time this will add capabilities of interest from commercial clouds, which tend to be more feature-rich than our OpenStack environment. Emphasis would be on providing ease of use for constituents who wish to move between CC's cloud and a commercial cloud, or in the other direction. This will include situations where users pay for the cloud capacity themselves, but CC enables workload and feature portability.

Deep storage and persistent storage: At least one new site was sought to augment SFU and uWaterloo in hosting backups and other nearline storage. These will consist mainly of tape libraries and associated software and infrastructure. Proponents could select this as an additional option to the main three system types. In all options, persistent storage for each site will be added as part of the total system cost. Overall, plans include 20-25% of capital to go towards storage (discussed further below). This will reach over 100PB of online persistent (not temporary) storage across all Stage 1 and 2 sites, along with needed tape capacity for geographically separate backups.

Local/regional data caches: Relatively small resources to provide local/regional access to the national data infrastructure. These would be distributed roughly in proportion to storage need (as readers or as writers), and would expand the efficiencies of large-scale procurement and operations from the national data infrastructure to smaller sites. Proponents could not bid explicitly for this resource type.

Services infrastructure: Further investments in the service infrastructure development efforts funded through Stage 1 are envisioned. Stage 2 hosting proponents were not permitted to bid explicitly for this resource type. As described in Sections 1.3 and 2. Investment to date has focused on personnel to develop or adapt services, including those needed for Challenge 1 as well as other common services. The philosophy is that if multiple users/groups express a need for a service, such as via user surveys or white papers, then CC should consider making that a national offering. Investments to date have included a software partnership with Globus, to develop Globus data publication services entirely based in Canada (to avoid offshoring of metadata).

3.1.1. Continuum of System Types and Configurations

Compute Canada has engaged in planning since the Stage 1 site selection process, in which labels for system types were first introduced. GP systems have bifurcated into GPx (diverse node types, including for cloud) and GP1x (cloud only), while LP represents the largest homogenous partitions. In fact, these are continua. LP-type systems could have more node types, while GP-type systems could include larger homogeneous low-latency partitions, and GP1x-type systems could be used for parallel computations. Moreover, the coordinated storage infrastructure (below) and increased emphasis on common schedulers and other aspects of workload portability mean that users will be more easily able to utilize multiple resource types, when and as needed.

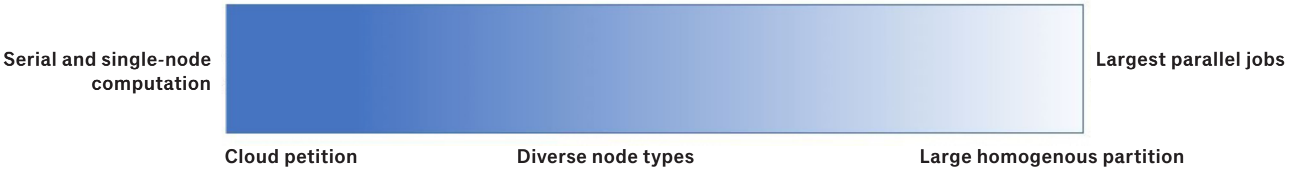

Serial and single-node computation
Largest parallel jobs
Cloud petition
Diverse node types
Large homogenous partition

Figure 1.5: Continuum of System and Node types

The three system types described here (GP1x, GPx, LP) are associated with points on the continuum which are anticipated to fit best with the profile of users that currently exists. The experimental systems, and the ability of the GPx systems (especially) to accommodate a mixture of node types and workloads, will help provide additional flexibility and capabilities in the future. We can imagine future need for very large memory systems, for machine learning, for different processor and accelerator types, etc. It is in the interest of our users that Compute Canada is able to continually evolve system and service offerings, and to help researchers to utilize the best technologies available to achieve their goals.

The notion of a continuum of system types is relevant for CC's anticipation of Stage 3 funding. Stage 3, if offered, will support further growth of storage, and will need to address needs for cloud and the mixture of GPx-type node configurations to accommodate the workload. A highlight of Stage 3 discussions to date have focused on a much larger LP-type system. A $75M cash investment, for example, could yield a true leadership-class system for Canada. Such a system would create challenges for single-province matching mechanisms, as would the larger operational costs for power than a single host institution would be expected to shoulder. CC sees this as an opportunity, and a way of growing the size of questions that may be addressed computationally. LP systems are natural big data systems, able to handle the workload of many of CC's users. The Option to expand the Stage 1 LP, discussed below, would impact planning for Stage 3, as well as the needed future balance of GPx/GP1x/LP-type systems.

3.2. National Data Infrastructure

For Stage 2, as for Stage 1, significant effort and investment is focused on meeting storage needs. The four Stage 1 sites formed a purchasing consortium that successfully and economically built a base of products and services that will underpin storage across all Stage 1 and Stage 2 sites. User need forecasts received by Compute Canada indicate that 20-25% of total capital must be spent on storage through at least 2020, in order to handle the volume of growth. This takes into account likely technology changes over time, using Kryder's Law and market research to forecast increases in storage density.

The new national data infrastructure is based on the concepts of storage building blocks (SBBs) and software-defined storage. These concepts are intended to give scalable mechanisms for adding commodity storage over time and in multiple locations, thereby enabling expansion when needed, with the appropriate capacity and performance profile, at market prices. At time of writing, the first set of Stage 1 storage purchases are underway, for deployment at all four Stage 1 sites.

In addition to filesystems (including multiple petabyte parallel filesystems, as well as traditional Linux filesystem types), the Stage 1 Storage purchase includes object storage software. Object storage provides a mechanism for cost-effective and highly resilient data replication, which is required by some of Compute Canada's largest storage users. Object storage also provides access control and other mechanisms directed at multi-tenant systems, which is required for health information and other sensitive data types.

Facilities for deep storage are critical components of the infrastructure. Nearline capacity management will allow lesser-used data to be moved to tape and then retrieved when needed. Backups will provide a high degree of resiliency and durability, with two copies of most data (configurable by policy): one each at SFU and uWaterloo, with a third site planned as part of Stage 2. Sites will rely heavily on CANARIE and the regional networks, as they ramp up to 100Gb speeds, in order to facilitate data movement. Starting in July 2016, approximately 15PB of data from legacy systems will be migrated into the new systems.

Overall storage capacity projections were included in the Storage RFP, covering filesystems, backups, and object storage. SBB types suitable for metadata, local Ceph/object storage, databases, etc. were specified. Overall online storage capacity across the four sites is projected to reach 40PB by the end of 2016, and 62PB by the end of 2017. Additional sites via Stage 2 will add to this capacity for 2017 and onwards. The Storage RFP included access to the same discount levels and products for Stage 2.

3.3. Elastic Secure Cloud Services

For this Stage 2 proposal, notions of federated cloud sites and local/regional data caches have been expanded to incorporate elastic secure cloud services. Stage 2 site selection RFP responses indicated strong need, as well as existing capabilities, for secure cloud (from six out of the ten proponents). The main current use case for these services is hosting of health information, including personally identifiable information (PII). PII is reflected in one of the largest data growth areas (genetic sequences and brain imaging, which are also major elements of Challenge 1). Emergent use cases are in the social sciences, where controlled access to datasets is the norm. Researchers in criminology, labour statistics, and other areas have similar needs.

Our capital investment options propose to build a secure multi-tenant environment based on the concepts of OpenStack cloud and local/regional data caches. The intention is that the same OpenStack cloud environment as other Compute Canada cloud resources, with the same storage environment, will implement logical partitioning such that the needed levels of isolation for data and compute are enforced. This design is informed by the highly successful HPC4Health implementation by Compute Canada member institutions in Ontario (www.hpcforhealth.ca). The model will be expanded and enhanced to meet the needs of other provinces. It is proposed that secure cloud capabilities will be part of all OpenStack systems or partitions on Stage 1 and Stage 2 sites (i.e., all GP1x and GPx systems).

The “elastic secure cloud services” label is chosen to convey several qualities. First, any of the cloud partitions on GPx systems are intended to be resized as needed in response to user demand, with allocation of appropriate computational/storage resources. As mentioned above, all cloud systems will be able to provide a secure environment, via logical partitioning of compute and storage resources. Such logical partitioning is used by HPC4Health and some other current implementations by CC members, and is in most cases adequate (i.e., physical partitioning and air gaps are not necessary, but separate filesystem mount points and VLANs are). The secure partitions within a cloud will, generally, be assigned to a particular tenant (such as a hospital department, or a Challenge 1 data analysis platform). The tenant would have the needed control over authentication, authorization, logging, etc. Those secure partitions would also be elastic as needed over time, so that they can expand, shrink, or gain access to a different resource mix.

4. Capital Investment Options

The “need for the infrastructure” and the desired investment components, both described above, exceed the capital available for the contemplated Stage 2 investment. Moreover, as shown by the table below, the potential Stage 2 hosting proponents have a willingness to match funding and fund operations in Stage 2 that exceeds available funding by more than 2.5x. Given this excess of demand and of needs, CC has worked with its community of member institutions to identify the type and scale of investment that will best respond to their own needs and preferences. Since different institutions have different priorities, this process has identified multiple options for consideration by the CFI, and each is described below.

Consistent with CFI’s Stage 2 guidelines, hosting institutions are not identified in connection with the preferred options; although we provide general indications about whether a given investment would represent expansion of Stage 1 system(s) or system(s) to be located at a data centre other than one of the Stage 1 sites. Over the coming months CC will work to finalize site selection for forthcoming investment, under each of the options described below.

Proponent	Site	Province	GP1x	GPx	Deep store	Exper	CFI Contribution (matched)
ACENET	Dalhousie	Nova Scotia		x	x	x	$5,000,000
Calcul Québec	ETS, Montréal	Québec	x	x	x	x	$14,500,000
McMaster	McMaster	Ontario	x		x		$1,000,000
SFU	Burnaby Campus	British Columbia	x	+	already	x	no maximum
U. Alberta	Edmonton N.	Alberta	x	x			$10,500,000
UBC	UBC Point Grey	British Columbia	x			x	$11,500,000
U. Calgary	Shaw Data Center	Alberta	x	x		x	$590,000
U. Manitoba	Ft. Garry (Winnipeg)	Manitoba	x	x	x	x	$4,000,000
U. Toronto	Vaughan	Ontario		LP	x		$6,000,000
U. Waterloo	Waterloo	Ontario		+	already		$2,000,000

4.1. Capital Investment Option 1

This is the preferred option proposed by Compute Canada. The proposed configuration:

System/service type	CFI capital	Notes
Deep storage	$2,500,000	One additional deep storage site, plus additional capacity for the current two sites.
Experimental systems	$750,000	Small experimental systems at some Stage 2 sites; modest investment in commercial cloud.
Services infrastructure	$250,000	1 FTE for 2 years, plus small purchases of existing software and/or services.
Elastic secure cloud (ESC)	$750,000	One standalone ESC site.
Expand LP	-	No expansion of LP.
GPx[1]	$15,750,000	Expansion of one or more GPx systems, and addition of one or more new GPx systems. All GPx systems will have ESC partitions.
TOTAL	**$20,000,000**	

Option 1 prioritizes most of the possible investment components described above, except for expansion of LP. The GPx investment is high, allowing for expansion and addition of sites. GPx systems will provide a mixture of node types (bigmem, GPU, tightly coupled compute partitions), and an elastic secure cloud (ESC) partition. Mid-level investments in experimental systems and a single standalone elastic secure cloud site will help to yield a rounded portfolio. A modest investment in services infrastructure (versus $1M for Stage 1) will augment the efforts of National Teams. The Stage 1 LP investment, to be made in 2017, will stand as-is at $8,425,000, and will not be increased with Stage 2. Further description of estimated yielded core capacity is in the Budget Justification section.

Option 1 Estimated capacity	Stage 1	Stage 2	Total
Ncores (Elastic Secure Cloud)	(GP1) 8,500.	5,486.	13,986.
Ncores (LP expansion)	(LP) 66,000.	0.	66,000.
Ncores (GPx)	(GP2+3) 52,000.	89,250.	141,250.
Total cores	126,500.	94,736.	221,236
New persistent storage (PB online)	62.	38.	100.

In this option, investment is focused on general purpose computing (i.e., GPx-type systems, with multiple nodes types), with the addition of a standalone ESC site. The strength of this option is that GPx systems will address the needs of the majority of users/projects, adding needed capacity. The node configurations of new GPx systems and expansion of Stage 1 GxP systems will be adjusted to reflect early experiences with the

[1] In the event additional funds are available, we would add one additional standalone ESC site and increase our investment in GPx systems with the balance.

Stage 1 systems - for example, it may be desirable to have larger partitions for tightly-coupled workloads, or to have larger bigmem nodes, or different configurations for local storage, or variations on the cloud partition sizes or node configurations. Such adjustments, as funding allows, may mitigate the fact that the Stage 1 LP system would stay at the currently planned size, without augmentation. Progress towards an envisioned larger LP-type system for Stage 3 would be based on the base LP size, which is already planned in Stage 1 to have approximately 2x the core count of any current CC system. The strength of the ESC addition is to develop a new model for local/ provincial/ regionally-focused systems, at a relatively low cost but with very high value. ESC systems highlight capabilities of CC's systems and staff, give needed features to stakeholders, and provide on-ramps to larger computational and storage resources. A weakness of ESC is that it reduces capital for other purposes, and dilutes the effort towards consolidation of fewer, larger sites and resources.

4.2. Capital Investment Option 2

This is the second choice option requested by Compute Canada. The proposed configuration:

System/service type	CFI capital	Notes
Deep storage	$2,500,000	One additional deep storage site, plus additional capacity for the current two sites.
Experimental systems	$250,000	Small experimental systems at some Stage 2 sites; modest investment in commercial cloud.
Services infrastructure	$125,000	.5 FTE for 2 years, plus small purchases of existing software and/or services.
Elastic secure cloud (ESC)	$1,500,000	Two standalone ESC sites.
Expand LP	$6,000,000	This would add to the Stage 1 capital for LP, at the University of Toronto.
GPx[2]	$9,625,000	Expansion or addition of two or more GPx systems. All GPx systems will have ESC partitions.
TOTAL	**$20,000,000**	

[2]

Option 2 includes a focused investment to expand LP from $8,425,000 allocated in Stage 1, to a combined total of $14,425,000. This would enable a top-tier system in Canada, likely exceeding 100K cores and facilitating utilization by more large-scale users. It would also accommodate a variety of smaller workloads, including data-intensive workloads, for users that do not require the specific node types or software environments provided by GPx systems. By reaching a total facility power draw peaking over 3MW, SCINET anticipates a reduced power rate versus with Stage 1 alone.

GPx expansion or addition would be at a lower level in this option. Investments in experimental systems would be at ⅓ of Option 1, and services infrastructure investments at ½. Estimated yielded core capacity is based on the same parameters as the other options.

[2] In the event additional funds are available, we would increase our investment in GPx systems accordingly

Option 2 Estimated capacity	Stage 1	Stage 2	Total
Ncores (Elastic Secure Cloud)	(GP1) 8,500.	10,971.	19,471.
Ncores (LP expansion)	(LP) 66,000.	47,003.	113,003.
Ncores (GPx)	(GP2+3) 52,000.	54,542.	106,542.
Total cores	**126,500.**	**112,516.**	**239,016.**
New persistent storage (PB online)	62.	38.	100.

This option yields the strength of a larger LP system. As described earlier, there is clear demand for this capacity, and an LP-type system can be used for some of the GP-type workload (without provisions for cloud partitions). Because LP nodes are, in current planning, less expensive than other node types, more cores will be yielded per dollar (see the Budget Justification section below for tradeoffs in memory size, local storage, and other characteristics versus GPx systems). One of the chief anticipated benefits is supporting growth of CC's largest users, and potentially fostering on-shoring of some currently expatriated users - thereby creating opportunities for discovery at scales not previously possible with CC resources. This will assist in steps towards even larger systems, such as are contemplated for Stage 3. One weakness of LP investment at this stage is that it comes at the expense of GP-type investments, whose specific node types are needed by many users. Another weakness is that operational costs for the Stage 1 LP system (power & personnel) are higher than in other provinces, thereby resulting in a higher operational cost/core than similar systems elsewhere would have. In this option, the smaller investment in GPx will result in fewer/smaller cloud partitions - with the benefits & weaknesses discussed for Option 1.

4.3. Capital Investment Option 3

This is the third choice option requested by Compute Canada. The proposed configuration:

System/service type	CFI capital	Notes
Deep storage	$2,500,000	One additional deep storage site, plus additional capacity for the current two sites.
Experimental systems	$750,000	Small experimental systems at some Stage 2 sites; modest investment in commercial cloud.
Services infrastructure	$250,000	1 FTE for 2 years, plus small purchases of existing software and/or services.
Elastic secure cloud (ESC)	-	No standalone ESC sites.
Expand LP	-	No expansion of LP.
GPx[3]	$16,500,000	Expansion of one or more GPx systems, and addition of two or more GPx systems.
TOTAL	**$20,000,000**	

[3] In the event additional funds are available, we would increase our investment in GPx systems accordingly.

Option 3 is similar to Option 1, except with an even larger investment in GPx. This increase over Option 1 comes at the expense of dropping standalone ESC sites. However, this may be offset by an additional GPx site (with ESC partition) beyond that of Option 1. Investment in services infrastructure and experimental systems are at the same levels as Option 1, as is the lack of increase to LP. Estimated yielded capacity is based on the same parameters as the other options.

Option 3 Estimated capacity	Stage 1	Stage 2	Total
Ncores (Elastic Secure Cloud)	(GP1) 8,500.	0.	8,500.
Ncores (LP expansion)	(LP) 66,000.	0.	66,000.
Ncores (GPx)	(GP2+3) 52,000.	93,500.	145,500.
Total cores	126,500.	93,500.	220,000.
New persistent storage (PB online)	62.	38.	100.

This option has essentially the same strengths and weaknesses as Option 1, but allocates the ESC investment of that option towards GPx in Option 3. This would allow for larger GPx systems, and possibly an additional GPx site. Because GPx systems will include ESC partitions, the use cases associated with ESC will still be pursued. GPx systems also include uniform low-latency partitions suitable for HPC workloads, and Option 3 may accommodate larger partitions. Stage 3, if offered, will allow renewed focus on larger-scale LP systems.

5. Sustainability

An operations and maintenance (O&M) plan was submitted to CFI as part of the Stage 1 finalization. The plan will be updated during the lifetime of Stage 1 and the MSI, and Stage 2 sites will join the plan.

Operating cost estimates for the three options presented are dependent on assumptions of core counts acquirable by the funding available, as well as the location at which equipment is operated. The following chart shows the range of 5-year operating cost estimates for the three options based on various ways of distributing capital among the candidate hosting sites listed above.

	Incremental 5-year Operating Cost Estimates*		
Range	Baseline (Option 1)	Option 2 Difference from Baseline	Option 3 Difference from Baseline
High Estimate	$4,969,000	+ $1,575,000	$202,000
Low Estimate	$4,216,000	+ $917,000	$202,000

** Incremental Operating costs are electricity only. The systems administration staff required for Stage 2 are already part of the Compute Canada team; once Stage 2 site selection has been decided it is expected that some system operation positions will be redirected towards other CC objectives.*

Advanced Research Computing (ARC) in Canada is a national resource that is being managed collaboratively at a national level by Compute Canada. CC and the host sites' ability to operate and manage ARC resources optimally is dependent on the MSI funding, per Part A. The goal of the O&M plan is to assure all parties that any cyberinfrastructure deployed will be maintained and made available for national use throughout its useful lifespan.

For Stage 2, as with Stage 1, infrastructure items to be purchased will have a nominal lifespan of 5 years. Generally, equipment purchase price will include lifetime extended warranty support. Prior to every major purchase, the hosting institutions - which will purchase and own the equipment or other infrastructure - will confirm to CFI that they will continue to operate the equipment over its lifespan.

Key elements of the O&M plan for the Stage 2 cyberinfrastructure:

- Costs of acquiring the cyberinfrastructure are included in this Part B proposal,
- Operations and maintenance will be as described in the 5-year MSI renewal and accompanying Budget and Budget Justification sections (Part A),
- An O&M plan with the points in this section was provided to CFI as part of Stage 1, and will be updated as part of Stage 2,
- All Stage 2 sites, along with Compute Canada, will be party to the O&M plan submitted to CFI and updated as needed. This is a requirement of being a hosting institution and receiving CFI funding for Challenge 2,
- Capital match for acquisitions was identified as a component of becoming a hosting institution, and,
- Prior to every major acquisition, hosting institutions will confirm to CFI the intention to operate cyberinfrastructure for its useful lifespan.

6. Benefits to Canadians

"We need Infrastructure that supports change"

– PM Trudeau, Davos, January 20, 2016

Benefits to Canadians for the national platform provided by Compute Canada are described in Part A. In this section, the focus is on the benefits of specific aspects of the proposed Challenge 2 cyberinfrastructure investment:

Deep storage: Although CC does not provide archival services, a very high level of data protection is the norm. As part of the national data infrastructure described above, Stage 2 will extend from 2 to 3 sites with robust, enterprise-grade facilities for long-term retention of data, mostly on tape. Tape provides capacity at under ⅓ the cost of online storage, and does not use electricity when data are at rest - providing a more "green" alternative to online storage. The typical configuration for any item stored on tape will be to have two copies, at two different locations. This will apply to all data, whether as a backup, a replica of an online object, or a nearline extent of an online file (i.e., via hierarchical storage management or similar techniques).

For Canada's research community, this aspect of the national data infrastructure means that data will be kept safe. CC's users generate data primarily through simulation, or through observation. If through simulation, the ability to regenerate lost output data may exist, or may not (depending on whether researchers can fully reproduce the software environment and input data that generated the output). When such output required many core years to produce, it may be impractical to regenerate.

For observational data, there are data sets for which CC is the main repository, the only repository, or the repository of last resort. Whether observations come from genetic sequences, radio astronomy observations, light source experiments, particle physics experiments, clinical observations, microscopy, or other sources, it may be difficult or impossible to recreate the observational data. CC provides a key value for Canada's researchers, large and small, in having robust mechanisms and support infrastructure for ensuring data are safeguarded.

- Benefits to Canadians are the growth and stewardship of a safe and reliable destination for storing the valuable data generated by users of CC resources.

Experimental systems were of interest for Stage 1, but not proposed to CFI due to the desire to devote maximum resources to putting cores and PB in the hands of researchers. With Stage 2, CC and members desire to allocate a few percent of the capital towards experimental systems. This is a relatively small investment with potentially large payoffs. The first payoff is to the CC research community, by providing early access to technologies that may, in the future, become mainstream. A second payoff is in assessing the utility of new technologies, in order to decide whether or not to pursue them. The third payoff is in giving an opportunity to Canada's high tech industry to work with CC and members, to gain insights into product utility, functionality and promise.

- Benefits to Canadians are in fostering development and adoption of next-generation technologies and services.

Services infrastructure to be developed during Stage 1 and 2 is explicitly intended to be easy to deploy within CC resources and elsewhere. Many outcomes will be cloud-based, and others will be built in partnership with other platforms, such as where single sign-on will leverage the Canadian Access Federation (CAF) project managed by CANARIE. CC's efforts for services infrastructure are directed at the broad community of users, including Challenge 1 users. Efforts are directed, as much as possible, at general-purpose solutions which will be freely available to others in Canada. One current example is Globus Data Publication Service, in which the Stage 1 investment is resulting an open platform for all users. Another example, related to CAF, is identity management. CC is working within Canada and internationally to make it easier for Canada's researchers to have authenticated access to resources globally.

- Benefits to Canadians are from services infrastructure outcomes being available to the broad research community in Canada, not just on CC systems.

Elastic secure cloud (ESC): Currently, cloud services within CC are primarily for experts who have the need for a custom software configuration. These configurations are typically implemented in a Linux virtual machine, and then deployed and managed within CC's OpenStack environment. During Stage 1, Stage 2 and beyond, CC is working to make its cloud offerings more flexible, easier to use, and with increased capabilities for secure cloud features. In the long term, CC forecasts decreased need for differentiating cloud-type resources from HPC-type resources, and has already taken steps towards the future by designing cloud partitions for GPx-type systems. In future environments, HPC (i.e., batch parallel computation with a balanced high-performance interconnect) will be just one service offered in the CC environment. We will allow essentially any combination of needed software stacks, isolation of compute and data, use of in-house authentication and authorization layers, particular mixes of node types, persistent services, performance and capacity tiers for storage. ESC is a key leverage point for moving towards this envisioned future.

- Benefits to Canadians will come from expert support for the flexibility and configurability of cloud services, to adapt service offerings for specific requirements of ARC users.

Expansion of LP: Whether or not the Stage 1 LP investment is recommended for expansion now, there is a clear need within Canada for ever-larger resources of this type. Canada lags its G8 peers in access to high-end HPC resources. Large partitions of tightly coupled, uniform node types are needed to address the largest problems in numerous scientific disciplines, including climate and environmental study, physics and material science, astronomy, and chemistry. As the SCINET survey demonstrates, there are scientists in Canada, today, who cannot get their work done on CC's resources. The Stage 1 LP investment will mitigate the situation, but not solve it. Meanwhile, there are many users who have reported self-limiting the size of their jobs, in the hopes the jobs will launch sooner. CC desires to provide ever-larger resources, and also to support greater use of HPC resources when such use will increase the accuracy or verisimilitude of findings.

- Benefits to Canadians will come from new discovery or understanding that could not occur without the use of large-scale computations, and the attraction and retention of scientists (and their intellectual property) who rely on availability of a world-class system.

GPx systems and expansion: With over 3,100 active projects and 11,000 active users, CC serves a very diverse clientele. GPx-type systems are intended to grow over time (through expansion or contributed systems), to best meet the needs of users. While GP2 is targeted at jobs of 1,024 or fewer cores, future GPx systems could elect to have a far larger uniform partitions, even up to LP scale. For most users, it will be the specific node types - larger memory, with local disk; big memory (512GB-3TB); GPU nodes; cloud partitions - which let them best match resources to their jobs.

- Benefits to Canadians will come from the discoveries, innovations, and other outcomes of CC's users who are able to leverage these resources for their computationally-based research.

7. Budget Table

In the presentation of options, budget reflects capital expenditures. Every expenditure will consist of the usual CFI 40/40/20 split. For example, $400K of CFI capital would be matched by $400K (typically via a provincial agency), plus $200K in vendor in-kind. Operational expenses are not included in this capital budget, as will be met by the MSI described in Part A.

System/service type	CFI capital	Total
Deep storage	$2,500,000	$6,250,000
Experimental systems	$750,000	$1,875,000
Services infrastructure	$250,000	$625,000
Elastic secure cloud (ESC)	$750,000	$1,875,000
Expand LP	-	-
GPx	$15,750,000	$39,375,000
TOTAL	**$20,000,000**	**$50,000,000**

Option 1 Capital Budget. Total reflects 40% CFI capital, 40% match, and 20% vendor in-kind

8. Budget Justification

Estimated yielded capacity as shown in the options were based on assumptions of core density per node, for purchase in mid- to late-2017. These projections are consistent with projections used in Stage 1, which will be adjusted as technology evolves, the market changes, and other factors (such as $CAD fluctuation) occur. Cores/$ is expected to increase over time, and therefore the time of purchase - ranging from early 2017 to mid/late-2018 for Stage 2 - will impact the capacity yield per $. A per-system per-site technology spending plan for Stage 2 will be developed, with higher fidelity forecasts, once the chosen option is identified.

For mid-late 2017, forecasts are based on an average of 40 cores/node. This average is intended to partially account for CPU core density increases, and also account for the possibility of manycore nodes. It does not account specifically for GPU-type cores. It assumes 5 years NBD support will be purchased for systems and subsystems. Anticipated yielded core counts are based on these estimates:

- GP1x: 3,657 cores per $1M. Larger-memory HPC base nodes with local storage. 10GbE interconnect. Appropriately sized high performance temporary storage. Purchase price also includes a proportion of persistent storage, external to the cluster.
- LP: 3,917 cores per $1M. At most two computational node types, including base HPC nodes and possibly nodes with accelerator/manycore. Balanced high-performance interconnect. Appropriately sized high performance temporary storage. Purchase price also includes a proportion of persistent storage, external to the cluster.
- GPx: 2,833 cores per $1M. A mixture of node types, including base HPC nodes, larger-memory HPC base nodes, bigmem, and GPU. All nodes with local storage. High-performance interconnect, not necessary fully balanced. Appropriately sized high performance temporary storage. Budgeted amount also includes a proportion of persistent storage, external to the cluster. GPx-type systems are expandable, and expansion paths are part of the RFP (for different groups of node types: base, large, GPU, bigmem; also temporary storage). This gives a pathway to accommodating contributed systems, and also a mechanism for larger or differently balanced configurations.