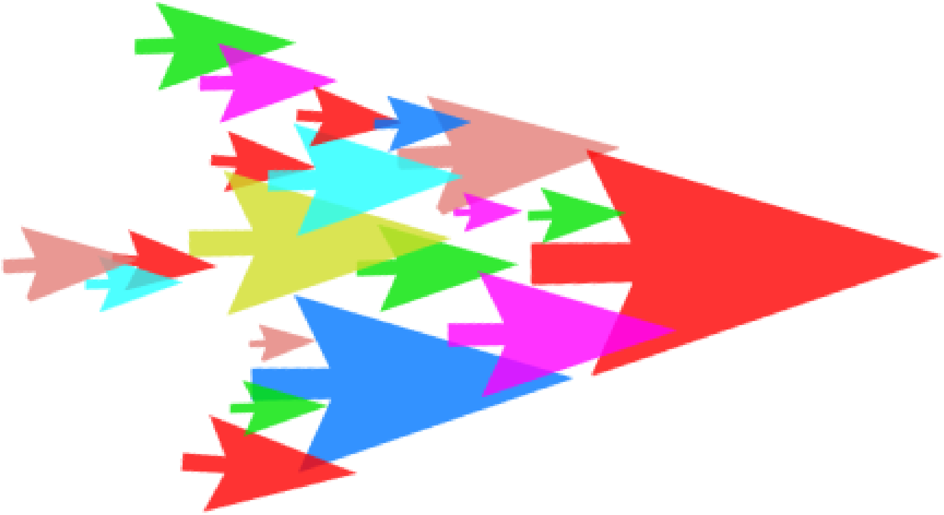
Rebooting Compute Canada

Building a **successful** and **federal** computational research enterprise, together



A Draft Discussion Paper For
The Compute Canada Community
April 4, 2017

Executive Summary

A New Federalism Over time, the national Compute Canada project has gone from being a loose association of occasionally-cooperating independent sites to a highly centralized expensive central office making extremely specific technical and personnel prescriptions for the entire country. Neither approach is sustainable.

- **Standardization**
- **Facilitation**

Researcher Focus foo

Research Focus bar

Background

Compute Canada has failed. We need the next one to succeed.

Compute Canada — the national project — was assembled to drastically improve the capabilities of Canadian research. Investigators were too often limited in the scholarship they could perform because of lack of local research computing expertise, or availability of storage or computational resources. The solution to this problem was clear then, and is still clear now. The most scarce and valuable of our resources, the expertise of research computing staff, actually grows, rather than is diminished, by applying it to many diverse problems; and economies of scale apply to computational and storage resources, which can be efficiently shared. Thus, the best way to support Canadian research was to pool resources nationally, rather than building many small silos. The federal government agreed, providing a steady stream of federal operating funding, far above what similar organizations have available elsewhere, freeing our effort from relying solely on institutions or even provinces.

But at this point it is clear that Compute Canada — the actual existing organization — has failed. Rather than building on the strengths of the regional organizations and the national perspective of a central group, the existing structure has radically worsened the natural tensions and distrust between them, escalating into a spiralling cycle of control and dismissive centralization on one side, and increasing disengagement and resistance on the other.

We thus find ourselves in a situation where Compute Canada has a central office costing millions of dollars a year — the equivalent of approximately 40 average-sized NSERC Discovery grants at a time when funding council budgets are frozen — with little direct benefit to researchers. The central office drives all technical decision making with a one-size-fits-all approach, such as the sorry and continuing saga of the multi-year "emergency" procurement of storage. But even with this level of centralization, there still is very little meaningful cooperation or knowledge-sharing between institutions; and Ph.D.-level computational scientists, rather than being active research partners in projects suited to their expertise regardless of their location across the country, largely spend their time answering help-desk-level questions about compiler errors and file system issues from users who happen to use the same computer that they sit beside. Lacking both the efficiency that comes from centralization and the flexibility that comes from decentralization, the community currently has the worst of both extremes.

A better future is possible, but the necessary changes will not just happen on their own. It is not in CFI's power or ability to step in and impose a new vision of what a Canadian computational research support enterprise will be. Nor will the ministry insert itself into the operations of a scientific support organization; if it did, dozens of similar organizations across the country would rightly be up in arms.

So if the situation is to be fixed, it is up to us — the members, the regions, and the researchers — to do so. The members and regions together do have the power, and the duty to researchers, to reorganize Compute Canada and rebuild a more responsive national platform for computational research support.

But we can't begin to fix our shared national computational research support enterprise without first knowing what "fixed" looks like. An inwards-focused Operational Plan exercise was promised after the externally-focused strategic planning process in 2014; this never happened. As a result, difficult but critically important internal questions were not debated and never received consensus answers — What services do researchers need? How do we best provide them? What is the role of a central organization? What is the role of the regions? How do decisions get made?

This document is one proposed set of answers to these questions; a sketch of one possible alternate future for providing computational research support in Canada. Its intention is to spur debate and inspire other different sets of answers, starting a discussion that should have happened three years ago. When a consensus emerges within our community, we will have both the knowledge of where we want to go, and the power to get there. Canadian research needs computing and data, and Canadian research deserves better than the current situation for providing them.

Principles

Any structure for a Compute Canada, present or future, must be judged against a set of principles we have for the running of such a research support organization.

We propose six such principles, listed below. In this section we describe them at length, and their current status.

Principle	Description
Researcher-Centred	The driver for every decision is researcher needs, with technology a means to an end.
Service Oriented	The purpose of the organization is enabling research, not just providing cycles.
Modern	Full advantage of current best practices are taken where they improve researcher experience.
National	Resource decisions are not made based on location - neither to researchers nor internally.
Interoperable, not Identical	All parts of the natural platform must interoperate seamlessly, but they should not be identical.
Equal Federal Partners	As equal funding partners, provinces and national operations share different but equally important responsibilities.

Researcher-Centred

In any long-running organization, there is a tendency to lose the perspective of the clients and instead to make decisions based on what is easiest or best internally. Groups that solve problems using technology are doubly prone to this, as the technology begins to seem important for its own sake, rather than simply being a way to achieve success for a client.

Modern agile software development addresses this problem by having desired “User stories” – a new task a user would want to perform – drive software development, with a product owner in charge of prioritizing the user stories so that they genuinely reflect the needs of the clients. While this is more meaningful for software development than service delivery, the basic method shows what one approach to keeping the clients needs firmly centred in decision making looks like, and how seriously many large companies take it.

In earlier incarnations of Compute Canada, a goal of maintaining a system in the top twenty of the Top 500 list of large systems was occasionally suggested. This correctly was never made into a formal priority, because it is simply not a legitimate goal of a research support organization. Specific technical benchmarks are end goals of technical organizations, not research support organizations.

The end goals of the latter can only be to effectively support particular projects and programmes of research. Some of those efforts may indeed end up requiring such a system, or access to such a system, but that would simply be a means to achieving a true goal of the organization.

The difference between an organization that is focussed on its clients and one whose focus is internal is reflected quite starkly their behaviours, in particular where time and money is spent. A researcher-focussed technical organization casts decision making in terms of researcher needs and successes rather than technical implementation details, deferring such details until the last possible minute, and pushing such decisions as close to the researchers as possible. A researcher-centered technical organization would never, as an example, begin the process of issuing RFPs for compute systems by drawing up prescriptive, detailed discussions of interconnects, processors, and core counts, but instead the metrics would be described in terms of use cases, job mixes, waiting times, and other researcher-facing metrics.

Similarly, a researcher-centred technical organization does not take urgent researcher needs such as storage, and pre-impose specific technical architectures upon the storage types before issuing RFPs to vendors; researcher-facing metrics are used, and any feasible solution with sufficiently good metrics and costs are quickly and efficiently procured.

A researcher-centred organization doesn't shift internal bookkeeping burdens onto the researchers, such as having multi-page sign-up forms requiring third-party authorization and several day waiting periods before access (compare for instance XSEDE or Amazon).

A researcher-centred organization allocates funding based on clear and concrete current or near-term researcher needs, and avoids spending large amounts of money on nebulous goals with no immediate driving need, such as untested "Research Data Management" solutions.

Not Researcher Centred	Researcher Centred
Users must fill out many elaborate forms	Easy sign-up, renewals
Technology drives decision-making	Researcher goals drive decision-making
RFPs specify architecture, interconnects, feeds and speeds	RFPs specify job mixes, researcher-facing metrics
Projects and collaborations are launched for their own sake	Projects and collaborations undertaken to meet specific, concrete, researcher needs.
Allocation of funding driven by various priorities	Allocation of funding addresses current and near-future researcher needs.

In a researcher-centred organization, significant decisions can always be justified in terms of making it easier for specific researchers to tackle concrete current or proposed projects of theirs, and the amount of resources allocated to that decision are proportional to those goals.

Service Oriented

Keeping the researcher central to decision-making will not automatically ensure that one is offering the most valuable services possible; researchers will not necessarily know to ask for services that have not been routinely provided in the past. To ensure one is offering a full range of valuable research-enabling services, one must constantly try new offerings, but in a disciplined and researcher-centred way.

New services can be routinely and inexpensively trialled with pilot projects, particularly when the services revolve around expertise rather than hardware (and it is these expertise services which will generally provide the greatest value-add for most researchers). This approach can only work, however, when it is paired with a commitment to ruthlessly prune services that provide little value before incurring too much cost. An area where this approach is taken successfully is training and education efforts, led by the regions and with little central involvement, where enrollment provides immediate feedback as to interest.

In a technology-focussed research computing organization, the main research computing service offered tends to be helpdesk-style questions about logging in, compiler errors, or queuing jobs — literally the lowest-level, least-value-added services that could be meaningfully offered beyond having the systems running. Compute Canada currently has approximately 60 Ph.D.-level staff who spend much of their work time performing this level of support.

Other organizations elsewhere offer a much richer set of services. Both XSEDE with their extended collaborative support services[1] and the growing Research Software Engineering[2] role in the UK embed staff inside research groups for extended periods of time to provide a variety of expertise, which can be particularly valuable for groups new to computational research or trying new-to-them approaches. Such staff participate deeply in the research, often to the level of authorship, and manifestly enable research that would have happened more slowly or not at all otherwise. In the 2013 staff survey, this level of participation was mentioned often as a stated wish of the regions' trained and ambitious technical staff.

SHARCNET has long offered dedicated programmer time, one type of such services, and it has been quite successful and indeed very popular with both researchers and staff. Such efforts have not yet been trialled nationally.

While Compute Canada is staffing on programmer staff currently, significantly, such staff are being hired centrally, report solely to senior executives, and skill sets have been selected and staff hired without any discussion with the researchers that they are proposed to enable. The intention appears to be for middleware development, so that such staff have little research computing background to speak of, but it is only now after most of the hiring has been done that assessment of what middleware is needed by the RPP and CyberInfrastructure projects they aim to support is being done.

Not Service Oriented	Service Oriented
New services are chosen centrally and rolled out full-scale nationally.	New services are piloted, tested, and scaled-up or phased out.
Services tend to be low-level and low-value-add.	Services range from hardware-provision to research partnership.
Services are either devised centrally, or done "the way things have always been done"	Best practices and new services used successfully elsewhere are routinely trialled.

A service-oriented research support organization ensures that services are offered to enable research at all stages of a project and at all levels of involvement, taking full advantage of expertise and resources available to the organization. New services are rigorously trialled with pilots before rolling out nationally, and service offerings are pruned if unnecessary.

[1] https://www.xsede.org/ecss

[2] http://rse.ac.uk

Modern

A research service organization which uses technology to address researcher needs must stay on top of new tools so that they can meet those needs as effectively as possible. Researcher needs must always be the driver, but solutions change quickly.

Those tools can certainly be new hardware — NVMe, FPGAs, and server-class ARM CPUs are all technologies which could have significant impact on research computing in the quite-near future — but they can also be new technologies for robustly and efficiently providing technical services.

As more and more companies rely on computer infrastructure, the past decade and a half have led to improved approaches to ensuring the services they provide are reliable and effective. For instance, servers and network connections fail; rather than being blindsided by the predictable, Netflix took the approach of routinely and automatically testing of failure to ensure that individual failures did not adversely effect users. Similarly, Google pioneered a now widely-adopted Systems Reliability Engineer (SRE) approach[3] which emphasizes extensive automation, minimizing human intervention on routine operations (even failures), allowing staff to focus on providing better kinds of services.

An organization which adopts modern tools ensures there is paid staff time and training for learning about new hardware and new approaches to deploying them. It continually provides small experimental systems to the staff (and interested researchers) to explore the suitability of new hardware and new provision techniques for suitability of research systems. It tests, modifies, and deploys new approaches to systems management. It also takes seriously the possibility of using commercial cloud providers as service provision options for some use cases.

Does Not Use Modern Tools	Uses Modern Tools
No availability of experimental systems	Invests in new technology for staff to explore for suitability for researcher use
Little paid staff training	Provides staff with time and training in new methods and techniques
Runs computer systems as they were run in late '90s	Modern SRE approaches are explored, customized, and used, such as heavy automation, routine failure testing
Interaction with users same as in late '90s	Interaction with researchers followed using tools like CRMs, so new staff anywhere in country can quickly come up to speed
Commercial cloud providers are the competition	Commercial cloud providers are one of many provision options

An research support organization which adopts modern tools also takes advantage of tools used elsewhere to provide better services, such as following researcher interactions and project progress using tools like Customer Relationship Management (CRM) packages, so that staff anywhere in Canada who might be able to bring their expertise to bear to assist the researcher can quickly be brought up to speed.

[3] https://landing.google.com/sre/book.html

National

Any conversation about Compute Canada has to have as a starting point that the reason for the effort is that pooling resources nationally is the best way to support Canadian researchers, and that their location in the country cannot matter for the type and level of services receive.

Truly national provision of resources to researchers, particularly resources as diverse and important as expertise, is something which takes active effort on the part of the research support organization; it can't be neglected as something which is allowed in principle but left to the researcher to pursue on their own. Presenting researchers with a list of national staff and bullets list of their expertise, and leaving the researcher to try contacting staff members in turn to recruit them to collaborate in their project, is a woefully inadequate approach to enabling computational research projects.

A truly national organization must make sure that Canadian researchers in all fields and institution types are adequately supported. Researchers in biological and life sciences (particularly human health), social sciences, and scholars in the digital humanities remain poorly served by Compute Canada; applied research work in colleges and polytechnics (over $200M/yr of external funding, approximately 40% of which comes from the private sector) is essentially completely ignored.

A truly national research support organization can't revert to using funding formulas that divvy funding up by the number of users in geographical catchment areas, but must fund services and providers to support researchers nationally.

Not Truly National	Truly National
Researchers are given a list of national resources available for them to investigate themselves	National teams of resources are actively assembled for a project
Researchers in some fields or institution types are overlooked	Researchers are supported equally across the country, across all institution types
Providers and services are funded based on the number of users near their location	Providers and services are funded based on the value they provide to the national community

Interoperable, not Identical

The internet is arguably the most important computational tool for enabling faster and better research made in modern times, and yet the central internet body, the IETF, does not specify brands of computer and browser, nor enforce a list of services that every website must provide. Instead, strict interoperability requirements, coupled with the freedom to innovate within those standards, have combined to make the internet such a powerful research tool.

Similarly, the Global Alliance for Genomics and Health (GA4GH[4]) is an international effort to build computational research tools to make full use of the increasing volume of genomics data to improve human health. A recurring mantra of the effort is to "Co-operate on interfaces, compete

[4] http://genomicsandhealth.org

on implementations". By building interoperability standards, allowing specialization for implementations, and working iteratively, the project is enabling new efforts like the Beacon Network[5] for data discovery and the Matchmaker Exchange[6] for better understanding rare diseases.

The Canadian research environment can be strengthened by ensuring that each project has the potential to access the complete national portfolio of computational science resources. But specifying exact model numbers of hard drives to use nationally, or that each region provide copies of the same services to the national research community, profoundly misunderstands the point of working together and pooling resources.

Onboarding

Focused on Identical	Focused on Interoperable
Infrastructure is specified in terms of technical specifications	Infrastructure requirements specified in terms of SLAs and interfaces to other infrastructure
Services are replicated across the country	Sites and regions can specialize in service provision with clear interoperability specifications
New sites cannot fully join the platform without wholesale replacement of infrastructure, procedures	New sites can easily fully join the platform by exposing services, infrastructure via interfaces

Equal Federal Partners

Canada has one of the most fiscally decentralized federal governments in in the G20, particularly when it comes to funding of research. This flexibility has real benefits, but it introduces complexities that are just as real. It is for this reason that we can not simply copy successful organizational models from the UK, or from XSEDE (from the US, where states generally play very little role in funding research). Even the EU is of little help to us here; their pan-Europe effort, PRACE, focuses exclusively on one type ("Tier 0") of research computing, with all other aspects of research computing support expected to be supplied to researchers by their member states or institutions.

Rebuilding To Our Principles

Having laid out a set of guiding principles, one can begin to design a new national organizational structure that adheres to them, aligning better to the expectations of the community.

The main building blocks of our national project include a national office; the provincial or regional organizations (henceforth, the regions), which we take to include the sites and the institutions that host them.[7] The two have different relative strengths.

[5] https://beacon-network.org/

[6] http://www.matchmakerexchange.org

[7] The relationships between the institutions and the regional organizations are important and complex, but they will quite rightly vary from province to province; since it's not meaningful to have a national consensus on the nature of those relationships, they aren't discussed here.

Role of a National Office

Role of the Provinces and Regions

Decision Making

Budgeting

Capital

Operations

What's Next

Our shared national project of enabling Canadian research with a country-wide portfolio of resources and expertise is too important to do poorly; and it is too important for us to retreat back into silos and limit researchers to those experts and computers that happen to be nearby.

After having read this document, you likely have something to say. Hopefully there were parts of this proposal that you strongly agree with; even better, there are probably parts you disagree with, think are missing, or think should go missing.

The purpose of this document is not to advocate in particular for the proposals contained within (although the points made here are genuinely-held, not just straw-man arguments). The purpose is to start in earnest a conversation that should have been launched officially three years ago, allowing the community and stakeholders to come to a consensus about what the internal organizational structure of Compute Canada should be, how it should make decisions, and how it should offer services to Canadian researchers and scholars.

The most important next step, then, is for you to have this discussion with colleagues locally and across the country, disagreeing vehemently initially on some points, and coming to agreement on others. We have put together one forum to have such discussions at `https://www.rebootcompute.ca`, where we would also be delighted to host competing proposals, but the location of the discussions doesn't matter; that they take place does.

The members and regions can rebuild a Compute Canada that works, and works the way the community wants it to, but they cannot do so without knowing what destination the community feels they should aim for. As the title of this document suggests, getting there from here will require completely turning off Compute Canada before starting it up again, with a completely new board and staff that are completely committed to the model and priorities that the community have chosen. But this process can happen in months, not years, and the result will be a Canadian research community served by a successful, truly federal, national computational research support organization. The Canadian research and research computing communities can do great things together. Let's get started.