NDRIO: Call for White Papers on Canada's Future DRI Ecosystem

Title: **The Future of Cloud as DRI**

Submission from Deborah Stacey (Technical Committee Chair) on behalf of the Linked Infrastructure for Networked Cultural Scholarship project - LINCS (lincsproject.ca)

Current State of Cloud

The Compute Canada Cloud (CCC) currently provides Cloud at the level of Infrastructure as a Service (IaaS). Its main services are the provision of virtual machines (VMs), Cloud storage, and HPC tools over SSH. Beyond the VMs and storage, there are few other services that are available currently. As an IaaS platform it is missing the following features:

- Load balancing to allow the scaling up and scaling down of resources
- Many flavours of backups – file-based, volume-based, etc.
    - This is improving (as are all aspects of CCC) and more storage services (such as the S3 storage recently added at Arbutus) are being added

It is obvious from our experience at the LINCS project (lincsproject.ca), that there are a number of Cloud services that would have made our setup simpler:

- Container orchestration services such as Kubernetes provided as a CCC service (currently this has to be set up by the user)
- Microservice capabilities, e.g. serverless functions
- Backups as a service – various backup services provided by CCC and subscribed to by the user/project

Another major concern with using the CCC is the need to interact with the CCC personnel to set up the most basic of infrastructure needs such as storage and backups. This is not a criticism of the CCC personnel, who are more than competent and eager to engage with research project needs. It goes to the level of maturity of the Cloud system. Most users of public/commercial[1] Cloud platforms rarely need to engage with the platform to set up basic infrastructure and are rarely impacted by the need to upgrade or maintain the infrastructure.

[1] In the literature on Cloud computing, services like AWS, Azure, and GCP are referred to as public Clouds since anyone can purchase services while the CCC would be called private because usage is determined by membership, *i.e.* the public does not have access. This comes from the analogous notion of public utilities like water and electricity - everyone has access but not necessarily free access. The public versus private term does not come from the source of the funding for the infrastructure but from the body of people who can have access (but not necessarily free access). CCC is only public in the sense that public money paid for it but it is not public in terms of access. But to make it clear as to the nature of the Cloud provider, the term commercial Cloud will be used in the remainder of this document to refer to Cloud providers such as AWS, Azure, GCP, *etc*.

Future State of Cloud

The CCC team is very skillful and none of this critique is meant to suggest that they are not doing an amazing job. It is very impressive that they can provide as much as they can with the number of people and resources that they have. That being said there is a distinct and vast difference between using the CCC and using commercial Cloud providers such as AWS, Azure, GCP, IBM, *etc*. These platforms provide the following that it is not reasonable to assume that CCC can deliver:

- Platform as a Service
- Guaranteed Quality of Service
- 24/7 Help and Services
- Software as a Service
- Bleeding-edge hardware

Platform as a Service (PaaS)

PaaS refers to a collection of hardware and software tools that are made available by the Cloud platform. There are a number of tools that should be offered by CCC with container management at the top of the list. A **container** is a package of **software** that includes application code and all its dependencies so the **application** runs quickly and reliably from one computing environment to another. This has become a standard way for deploying software in the Cloud and elsewhere. There are numerous container management systems with the most powerful and popular one being Kubernetes. As CCC evolves, it will be necessary to provide container management for its users. There are numerous 3rd-party software packages that can also help with managing containers/Kubernetes, such as *Rancher*, that are either free to use or have a small cost. CCC has to move beyond just allowing users to instantiate VMs; if it is to serve the needs of the Canadian research community it must provide best-of-breed software to manage Cloud infrastructure and software applications.

Guaranteed Quality of Service (QoS)

One of the distinctions between the commercial Cloud platforms and CCC is in the area of geographic distribution and replication. The commercial Cloud providers maintain a large collection of geographically distributed data centres that are able to provide replication of data, replication of computational capacity, and provide Cloud resources close to all of the users of a service/research application even if they are distributed across Canada/world. It is obvious that NDRIO/CCC cannot compete with this infrastructure, since it would be well beyond the financial and logistic capabilities of the organization. So obviously we cannot expect CCC to provide replication, guaranteed uptime, disaster recovery, and edge computing capabilities. If this level of service is required for a research project, then it might be prudent to look at supplementing CCC with selected services from commercial Cloud platforms. These services/infrastructure might only be used as a failover for CCC services/infrastructure or they might be used when guaranteed QoS is required. An example of this is the secure storage of keys, passwords,

certificates, *etc*. for protecting secrets and other sensitive data. One software application that provides this service is Vault (vaultproject.io). While it is possible to set this up in the CCC, because there can be no guarantees of stability (uptime, replication, *etc*.), it would be better to set this up on a commercial Cloud provider (a Key Management Service is available through the AWS Marketplace). The cost for this service is reasonable and can be estimated accurately.

Another example of the prudent deployment of commercial Cloud services is load balancing. Many commercial Cloud platforms allow their load balancers to include non-platform infrastructure to create a hybrid system where commercial Cloud resources might only be used when there is a need that cannot be serviced by the CCC.

24/7 Help and Services

It seems unlikely that NDRIO will ever be able to fully resource 24/7 help and maintenance services for the CCC. If there is an aspect of a research project that needs to be available continually (*e.g.* reading and processing sensor data for the control of actuators, or the provision of Linked Open Data resources essential to a research community) then a minimally viable version of this service could be deployed on a public Cloud platform to supplement the normal CCC service. It could be a fail-over option. Since many commercial Cloud services are pay-as-you-go, *i.e.* you are only charged when you use the service, this fail-over option in the commercial Cloud could be very cost effective.

Software as a Service (SaaS)

This last category is one that can be slowly achieved over time and in consultation with the major users of the CCC. Here we are talking about providing services such as Cloud Delivery Networks (CDN) for content delivery, fast database services, messaging applications such as Slack, *etc*. There is also the opportunity to take advantage of mature AI/ML (artificial intelligence/machine learning) products such as AWS Textract that are useful for laborious text, image, and speech tasks and are unlikely to be available through CCC.

For CCC to provide aspects of SaaS, they will need to study the major projects on the current CCC, study proposals for CCC usage, and survey past, current, and potential users of CCC.

Bleeding Edge Hardware

Hardware development has always proceeded at a blistering pace – often at a pace that government funding agencies cannot match. There may be a need for some projects to use some types of hardware that are not in the CCC inventory. This again is not a criticism of CCC, since it is not reasonable to assume that they will have the funding to constantly refresh their hardware at the pace that is present in the industry. The major commercial Cloud platforms can and do make available *bleeding edge* hardware in a timely manner. There may be select opportunities for hardware innovation, however. For instance, a Cloud platform for “real” Quantum Computing is available from the Canadian company D-Wave (www.dwavesys.com). A

partnership between NDRIO and D-Wave would be a major victory for Canadian research computing.

Bridging the Gap

The main proposal of this white paper is that NDRIO adopt a hybrid approach to Cloud Computing. That entails the combination of the CCC and services and infrastructure from the commercial Cloud providers. Currently the granting councils and CFI do not allow researchers to pay for commercial Cloud services. It is reasonable to direct researchers to the CCC, but it is not reasonable to restrict them to this platform if it is not able to satisfy their needs. A hybrid approach would be the most effective mechanism for maximizing the benefit of NDRIO's investment in the CCC with the sophisticated Cloud needs of Canadian researchers. Researchers have always had to justify where and how they spend grant funds and this would just be another category of appropriate expenditures. To maximize the effectiveness of a policy that allows researchers to use aspects of the commercial Cloud platforms, NDRIO should explore partnership agreements with the major Cloud platforms.

Cloud computing is having an even greater impact on all aspects of computing (research, services, teaching, *etc*.) now that we have discovered that working virtually as teams is more than just doable. Many more Cloud services will appear over the next few years and NDRIO and CCC must pivot to focus Cloud offerings to fit in this future world and also to take advantage of services essential to Canadian research that are provided by others.