Canada's Future Digital Research Infrastructure (DRI) Ecosystem: A perspective from the Bioinformatics National Team (BNT)

Current, future vision and needs of the BNT in the context of NDRIO

Jose Sergio Hleap, PhD
HPC technical consultant (SHARCNET)
BNT interim lead
jshleap@sharcnet.ca

François Lefebvre
Bioinformatics Manager (C3G)
BNT's Canadian Bioinformatics Helpdesk coordinator

Gemma Hoad
Bioinformatics Application Support (SFU)
BNT Bioinformatic Software Specialist

Jean-François Lucier
HPC technical consultant (Calcul Quebec)
BNT Bioinformatic Software Specialist

Table of Contents

Overview

Responding to the call for input from the New Digital Research Infrastructure Organization (NDRIO) leadership team, the Bioinformatics National Team (BNT) wanted to put forward our vision for the team within the NDRIO organization, with emphasis on our current status and issues, as well as aspiration for the future of the BNT with specific suggestions and expectations.

What We'll Cover in This White Paper

- Current Issues: In this section we will provide the status of the service model of the BNT within Compute Canada (CC), and will list strengths and weaknesses of the current model
- Vision: Here we will cover the expectations and ideas of a better functioning DRI from the perspective of the BNT to better serve the needs of our users
- How to Bridge the Gap: Threading from the previous point, a set of suggestions will be made

Current Issues

Status of the BNT within Compute Canada

The BNT started in September 2016 reporting to the Scientific Leadership Committee (SLC) of Compute Canada (CC) and it is composed of 15 members across Canada. Since then the BNT has supported the CC community in various ways:

- The Canadian Bioinformatics Helpdesk: Launched in 2018, has been offering online support for the Canadian bioinformatics community, as well as promotion of bioinformatic related activities.
- The BNT has provided bioinformatics repositories to CC systems in partnership with C3G, a Genome Canada Genomics Technology Platform.
- The BNT has provided internal support to Compute Canada analysts via a dedicated Slack channel and the user support ticketing system.
- Training and documentation: The BNT has produced approximately 10 documentation pages in the wiki, and has participated in over 15 training events (mainly in conjunction with SHARCNET and WESTGRID). Additionally, the team has answered more than 200 questions in the Canadian Bioinformatics Helpdesk.

Challenges and opportunities

Lack of experts across Compute Canada partners

As a subject specific national team, we are always looking for bioinformatic experts in the CC regional partners (ACENET, Calcul Quebec, SHARCNET, SCINET, and WESTGRID) to serve as a link to the systems and the member institutions. This is not always possible since the regions need to prioritize their budget in ways that might not warrant hiring bioinformaticians. Hence the BNT is sometimes relying only on the

expertise provided by the C3G and a few staff members. This impacts the reach of the team across regional partners, preventing a more concerted effort for the benefit of our users. It also impacts the reach of initiatives that the BNT might put forward as well as the gathering of information on regions' sites (i.e local users needs, uptake by regional users of initiatives, impact if training, etc).

Data redundancy and availability

Most bioinformatics tools rely on sizable reference datasets (e.g. Human genomes, BLAST "databases", etc) that are queried multiple times in a single run. The cost of this is three-fold: First, there is data redundancy in all clusters (i.e. just in Graham project filesystem, there more than 20 copies of BLAST's NR dataset) causing overcrowding of the systems and affecting researchers' user experience; secondly, not all datasets can comfortably fit in the project filesystem, and the users are forced to download reference datasets on the scratch filesystem, which also overcrowds the system and is volatile. The BNT and C3G have been working on a biological datasets repository solution, however, maturation to a proper Compute Canada resource never happened because of insufficient resources available to support this initiative; and thirdly, most bioinformatic applications are IO heavy, affecting all filesystems when reading or writing the same files repeatedly in a short amount of time. This does not only affect bioinformaticians. Artificial Intelligence (AI) for example, also suffers from these shortcomings of data management.

Expert visibility and user engagement

Given the current structure and funding of the BNT, engaging users by having visibility as a "go-to" person for bioinformatics guidance is challenging. For one, the BNT is not only composed of CC staff, which hinders the possibility of more engagement with CC users. Second, the lack of funding to send both CC staff (although offset by the regions) and non-CC staff to conferences, talks, etc, limits the visibility of the team across Canada. This is extra challenging while working from home, while conferences are not happening and while universities are limiting campus access.

Funding model of national teams

Within Compute Canada there is no established funding model for the national teams, hindering our ability to perform outreach, promotion, user and staff training. The BNT initiative included a discussion forum as a service model, however, the uptake of this initiative was lacking due to lack of promotion and budget.

Vision

BNT within NDRIO

The BNT will develop strategies aimed at optimizing the use of computational resources through development of new tools, refactoring of existing tools, training users and staff in best practices with current software, as well as adjusting the usage

of tools and data to NDRIO systems. The BNT will also act as a conduit for the NDRIO bioinformatics community to be heard on a national level, and to develop strategies that benefit this community.

Challenges

From the bioinformatics community there are 3 main challenges to overcome in the new DRIO:

1. Efficient access to secure data repositories: Health data has important safety considerations. To better serve the Canadian bioinformatics community we need to provide a suitable solution beyond site-specific solutions.
2. Bioinformatics staff in all regions: This allows for the BNT initiatives to reach all communities across Canada.
3. Training and outreach for CC staff and non-CC staff: Having the capability to send BNT representatives to conferences, talks, and local training events widens the visibility of the team, and the impact on our users.

How to Bridge the Gap

Better data management

The BNT has put a lot of effort into shared repositories. However, without the hardware infrastructure to accompany it (such as high efficiency storage), the users pay the price in time and effort wasted. To increase data throughput, the BNT proposes adding a data staging infrastructure in each of the systems. This comprises a new solid state file system connected to the infiniband network, managed through the Slurm Burst Buffer. The proposed storage system design integrates a tier of solid-state burst buffers in the storage system to absorb application I/O requests for data and IOOps heavy applications (Liu et al., 2012) such as bioinformatics and artificial intelligence. This is the rationale for Niagara having a solid state burst buffer, but this strategy should be extended to the rest of the systems. High IOOps reference datasets can be placed as persistent data on this filesystem, and provide semi-persistent space to users on request.

Funding model

The BNT proposes that NDRIO include in their funding model, additional funds (not including the regional partners' allocations) for the subject specific national teams. Allowing subject specific national teams to "buy" full-time or part of the time of hires to better populate the NDRIO with subject specific experts. Also, a modest budget to support training of users and staff in the subject specific matters is important to increase the outreach, expertise, and capabilities of the team. This model will allow the national teams to better fulfill its mandate, as well as provide a better quality of bioinformatics services by guaranteeing time dedicated to researcher support. We

believe that this could be applied to all subject-specific national teams, and will position NDRIO as a paramount resource for research in Canada.

Conclusion

We consider that the subject-specific national teams can add an important source of collective expertise to NDRIO and the partners regions. However, there are important challenges that need to be addressed so that the subject-specific national teams can be more integrated with the regional sites, and that regional users can get the most of such integration. The main challenge is the funding model for these teams, and we consider that the subject-specific team can become a better asset to the regions, having a separate source that will allow us to support the regional partners.

References

Liu, N., Cope, J., Carns, P., Carothers, C., Ross, R., Grider, G., Crume, A., & Maltzahn, C. (2012). On the role of burst buffers in leadership-class storage systems. *2012 IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, *2012*(0), pp. 1-11. 10.1109/MSST.2012.6232369.