Digital Research Infrastructure in Astronomy

Dr. JJ Kavelaars
Head, Canadian Astronomy Data Centre
Herzbeg Astronomy and Astrophysics Research Centre
National Research Council of Canada

December 15, 2020

1 Summary

In this White Paper we use the term Digital Research Infrastructure (DRI) to refer to the hardware (e.g. compute, storage, networks) and software (e.g. OS, middle ware, science platform) layers and to the support of users on those systems.

The Canadian astronomy community requires reliable and stable access to DRI that is significantly resourced and presented via interfaces designed to serve the science user. The astronomy research community is acutely dependent on DRI. Major telescope infrastructure projects which Canadian astronomers are partners in, such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST), CHIME and the Square Kilometer Array (SKA), require DRI at scales that are not currently available to this community. The National Research Council's Canadian Astronomy Data Centre (formed in partnership with the Canadian Space Agency) has a long history of partnering with Compute Canada and Canarie to ensure that Canada's astronomy community has the DRI they need to be successful. In this white paper we propose that NDRIO continue this partnership work with CADC, and the university led Canadian Advanced Network for Astronomy Research (CANFAR), to ensure astronomy continues to have access to DRI.

Established in 1982, the Canadian Astronomy Data Centre provides a science data portal that provides Canadian and international astronomers with a broad range of observational astronomy data. The CADC is funded by the National Research Council of Canada and the Canadian Space Agency. The CADC is housed within the Herzberg Astronomy and Astrophysics Research Centre with storage and compute systems provided by Shared Service Canada (SSC). The continuous innovation needed to ensure the utility of the CADC has been enabled by the partnership between the professional astronomers, software developers and operations personnel who make up the CADC staff. Today, the CADC provides high availability search and retrieval interfaces to over 300 million files occupying 1.7 PBytes of storage. In 2019 the CADC delivered 600 million data objects comprising over 1 PByte of storage to over 10,000 distinct IP address.

On behalf of the university research community, the CADC also operates a the CANFAR science portal on Compute Canada's Arbutus cloud. CANFAR consists of discipline specific VM computing systems and astronomy specific interfaces to large storage volumes, providing specialized science processing to over 200 science teams. CANFAR is the base layer on which Canada's next generation astronomy science portals for CHIME, SKA and LSST will be built. The CANFAR partnership between NRC/CADC and the university research community has allowed Canadian astronomers to gain access to high-value astronomy facilities (such as the European Space Agency's Euclid Space Observatory) by leveraging our astronomy domain data expertise. Continued support and development of CANFAR storage and compute capacity is needed for Canadian astronomy to continue to be successful.

Canadian astronomy needs to increase access to DRI and we must develop new community-wide structures to manage this infrastructure. The CADC, backed by continuous NRC funding, is a central element of this infrastructure due to its expertise in building and maintaining stable, world-leading DRI that is has developed over the last

30 years. Currently, funding of the hardware and software platform for research is being maintained via project focused-CFI grants (such as CIRADA, CLASP, CHIME, etc.) within a funding model that excludes direct support for the CADC. Separately, NRC has continued to support CADC's efforts to build the CANFAR science platform capacity so that this can act as unifying system for the various individual projects. Maintaining the CADC's world-leading expertise in the current system of funding via grants whose funding distribution excludes CADC involvement is highly challenging. Particularly challenging, from a data volume and computing needs view, will be supporting SKA, CHIME and possibly LSST data within the increasingly complex data landscape. Ongoing and increased funding to CADC is needed to deal with the challenges of creating an astronomy cyber platform that integrates Canada's astronomy data sets.

Where university-based groups are able to create and manage some components of the necessary science infrastructure, they should be encouraged to do so. The principal challenges for university groups are 1) to create infrastructure that serves a broad science community rather than a focused research project, 2) to integrate the infrastructure they create with existing national and international astronomy infrastructure and, most importantly, 3) to ensure the continued operation and ongoing development of their infrastructure reasonably far (more than a decade) into the future.

2 Introduction

As can be seen from even a cursory examination of the Canadian Long Range Plan for Astronomy 2020 (LRP), Canadian astronomy is a data-rich research endeavour. Astronomical research has evolved to become a digital science, dependent on methods of analysis, digital infrastructure, and the collection of large survey data sets. The data-related theme that runs through the the LRP report is that Canada has benefited strategically from the creation of a discipline specific archive and science data centre (the CADC) but the community's ambitions reach well beyond the capacity of the current facility.

To tackle fundamental scientific questions, astronomers are turning to ever larger data sets, made from surveying the sky with a variety of observational facilities operating across the energy spectrum. We are building up a picture of the universe over a broad range of resolutions, timings and messengers. At the same time, computing and mathematical methods are evolving towards techniques that are driven by AstroInformatics, Statistical Learning or Machine Learning (such as stellar classification and redshift determination) and even evolving towards systems that use artificial intelligence in their analysis. Astronomy is picking up the pace in deploying these new technologies and in training the workforce. These new modes of operating require significant and sustained investment in DRI.

2.1 The diversity of DRI

When considering digital research infrastructure solutions in astronomy one must keep in mind the diverse scope of the discipline of Astronomy and Astrophysics. The community studies physical processes that span scales from the sub-atomic to, literally, the size of the universe. The various fields study processes that are driven by Newtonian physics, General Relativity, Fluid dynamics, weak and strong forces, and chemical and mechanical actions. This broad array of physical scales and mathematical approaches requires support from a diverse cyber-infrastructure. While some problems are well addressed by massively parallel **supercomputer** systems using high-speed interconnected processors, other problems are better addressed through the use of dedicated GPU processing on a set of **specialized hardware**. Still other problems are better considered as **high throughput computing** where massive amounts of data are processed in parallel via machines with some limited interconnected capacity and access to high-speed storage systems. Other problems are better suited to systems that straddle interactive and high-throughput computing making use of *virtualization and cloud infrastructure*. When considering the solutions for cyber-infrastructure in astronomy it is important to keep in mind that this broad range of scales and capacities must all be satisfied for our diverse field to flourish and maintain its world-leading impact.

In the petabyte era the lines between software, technology and science are blurred - the likelihood to doing science with petabytes-scale datasets without major infrastructure designed and operated to meet the need of the science user will be very low. In a modern astronomical research community DRI should present to the community

a domain-specific view of the resources, providing access to software systems and tools that are designed to meet the research goals of that community.

3 Science Platforms

Canadian astronomy requires digital research infrastructure that can bring together these various digital pieces: computing, storage, networks, databases and software. **Given the expressed science need for data collections to cut across sub-domains (such as X-ray, optical, optical, infrared and radio astronomy) a single astronomy domain aware science portal that enables use of the full spectrum of this data is needed.** Such a system must be built in an agile way and respond on short timescales to the changing requirements of emerging science activities and evolving technologies. In addition to CANFAR, a few science platforms already exist internationally (notably JHU's SciServer, NOAO DataLab and ChinaVO) and many more are being planned or built, (e.g. Project Escape and LSST Science Portal). Through the International Virtual Observatory Alliance (IVOA), the astronomy community is working towards standardization of these science platforms. Canadian astronomy must include a comprehensive science platform as part of our DRI.

As the primary host of the Canadian Virtual Observatory, the CADC has been participating in the IVOA efforts from their outset to ensure that data centres provide standardized interfaces that enable interoperability. The success of this effort can be seen, for example, in the rich universe of IVOA/TAP services that allow astronomers to probe an incredibly diverse sets of astronomically relevant catalogs of information. These efforts lead to what might be considered 'open data' in astronomical research. As we move into the Science Platform realm this concept of standardized service layers is even more critical as they provide the opportunity for platform interoperability, permitting scientists to have their research investigations to span between centres and commercial computing clouds. As we work towards the concepts of 'bringing the code to the data,' which is a fundamental driver of the science platform concept, we must keep in mind the interoperability of these science platforms or we will lose the achievement of open data by closing off the resources needed to access those data.

At this time, CANFAR has more computational resources than most of the existing astronomy science platforms (all except for ChinaVO which, as of Fall 2019, has over 2000 active research users 10s of PB of storage and 10,000s of cores). However, the user interface to CANFAR requires substantial development to allow the same ease of use that is being achieved (or planned) elsewhere. These science portals are executing a vision of bringing astronomy research computing into a cloud-based environment that crosses boundaries across wavelength and research domains. The *SciServe portal* reaches beyond the astronomy community, providing similar capabilities across the sciences and humanities research groups at Johns Hopkins University. The concept of a science portal as the gateway to computing has become a ubiquitous modality.

To be useful, these portals require stable long-term funding on top of the base infrastructure funding. Without this stability, researchers will become reliant on systems that perpetually shift in their behaviour, and this will substantially impact productivity.

Long-term stable funding is key for successful DRI.

4 Needed infrastructure

The astronomy research community, along with other disciplines, has a growing and strong need for increased storage and compute capacity accessed via domain specific science platforms. Over the next decade storage capacity must grow from the current scale of a few petabytes of online storage (enabling the storage of all Canadian astronomy data on live disks) to the capacity of 30-100s of petabytes of online storage and many times more in near-line capacity (i.e., tape). This capacity is at the scale of one of the entirety of one of CC's current general compute centres (Cedar, Arbutus, Graham, Béluga and Niagra). The CHIME project, for example, already generates petabytes per year of operation while LSST will generate tens of PB and SKA will generate hundreds of PB per year. However, the data from each of these projects must be linked together to give astronomers a complete view of astrophysical systems. And astronomy is not alone: high energy physics, genomics and climate (for example)

are also demanding increased capacity for large data. **The existing capacity within the CC's centres will be exhausted on very short timescales.**

Along with the predicted increase in storage data volumes will come a need to increase our computing capacity. Within CANFAR, for example, the majority of computing is to reprocess observational datasets in novel ways to extract new information. Thus, if we increase the stored data by a factor of 10–30$\times$ we must also anticipate a need for increasing our compute capacity by a similar factor. The need for increased computational capacity is driven higher still by the development of new approaches to data analysis, in particular convolutional neural network (CNN)-based machine learning. The CNN approach invariably requires access to GPU-based computing to make possible the numerical computation needed to train deep networks. During the most recent round of CC computing allocations, the demand for GPUs significantly out-stripped the available capacity by an oversubscription rate that is even larger than for standard CPU processing. With growing data volumes and increased efficacy of techniques, the pressure on classical and GPU computing will only continue to grow. Here too, the data-intensive research communities must organize themselves so that the appropriate models for delivering computing infrastructure can be developed.

5 Timeline

The major data missions will require storage and processing capacity to rapidly grow, beginning in the early 2020s. By the end of the next decade, Canadian astronomy will require a fully operational storage and processing capacity capable of handling ≥ 50 PB of astronomical data. This is about 25 times the capacity that is currently (Fall 2019) being deployed for astronomy research needs. We must act quickly to be able to meet the needs of the community and remain world-leading researchers in astronomy.