Data and Computing White Paper

Australian Astronomy Decadal Plan Mid-term Review

Authors: Simon O’Toole (AAO-Macquarie), Bart Pindor (U Melbourne), Chris Power (ICRAR/UWA)

Overview

Data management and analysis tools, and high performance computing (HPC), are essential ingredients in any modern astronomical research program. Without this infrastructure, our investments in telescope access, instrument development, and scientific expertise will fall short of meeting their full potential. This white paper addresses the current status of astronomical data and computing in Australia, as well as the requirement for future research success.

The All-Sky Virtual Observatory (ASVO) has led the way for data management and access to Australia-held data sets, including optical and infrared, radio and theory. The next stage is to expand and consolidate this work into the three key data centres for the future: the Optical Data Centre (ODC), the Gravitational Wave Data Centre (GWDC), and the Australian SKA Regional Centre (AusSRC). The technologies developed and implemented by the five ASVO nodes will be essential in creating an Australian Astronomical Data Centre that federates the ODC, GWDC and AusSRC.

High-Performance Computing

Overview

The Australian Astronomy Decadal Plan (2016-25) identifies as one of its five equally weighted top-level science infrastructure priorities, “world-class high performance computing (HPC) and software capability for large theoretical simulations, and resources to enable processing and delivery of large data sets” from cutting-edge optical/infrared and radio telescopes.

Community demographics presented in the Decadal Plan (2016-25) highlighted that **theory and computation** comprises a significant proportion of the research activity undertaken now by Australian astronomers (approximately 40%); a defining feature of these research activities is the central role played by HPC. Taken together with **radio astronomy** (approximately 20%), which relies on supercomputing to process the data coming off the Murchison Widefield Array (MWA), the Australian SKA Pathfinder (ASKAP), and ultimately the SKA, this highlights the **fundamental, ongoing, importance of access to world-class HPC for Australian astronomical community.**

The Australian astronomical community accesses national HPC resources via the **National Computational Infrastructure (NCI)** in Canberra and the **Pawsey Supercomputing Centre** in Perth, which are directly funded as part of the National Research Infrastructure, and **OzSTAR** at Swinburne University of Technology in Melbourne, which receives funding from Astronomy Australia Limited (AAL). The main routes to access HPC on NCI and Pawsey are via the **National Computation Merit Allocation Scheme** (NCMAS), which has an annual call open to all computationally intensive research disciplines (e.g. climate, genomics, geosciences, materials, etc…) and makes available of order **300M core hours per annum**, although the size of the **largest individual allocations are of order 5-10M core hours per annum**; and via **partner schemes** with universities, such as The Australian National University (with NCI) and Curtin University and The University of Western Australia (with Pawsey), and with Australian Research Council (ARC) Centres of Excellence, such as CAASTRO or ASTRO 3D. Access to OzSTAR is managed independently. In many cases, astronomers are also utilising their **institutional and state HPC facilities**, as well as leveraging collaborations to access **overseas facilities.**

Current Status

Two recent snapshots of the Australian astronomical community underline one of the key outcomes of the Decadal Plan (2016-25) - that **ongoing access to world-class HPC continues to be critical to carry out the full breadth of our community's research programme**. They also highlight a **key area of community need** - access to multiple **large (>10-15M CPU hours), individual, allocations** of HPC time.

The first snapshot, from May 2019, is based on a national survey carried out by the AAL Science Advisory Committee's (SAC) HPC working group, and covers the entire astronomical HPC user community. Based on a sample of 65 respondents, it shows that,

- the astronomical HPC user community is predominantly concentrated in **computational/theoretical astrophysics** and **radio astronomy** - approximately 75% - but there is also activity in **optical/IR/UV**, **high energy**, and **gravitational wave astronomy**; and
- approximately 80% of this community consider HPC to be either **important** or **critical** to carrying out their research programmes.

It shows also that,

- currently, approximately **80%** of users are running research programmes that require annual allocations of **less than 5M core hours**, while **10%** are running programmes that require annual allocations of **more than 10M core hours**;
- in the coming 5 years, the number of users requiring annual allocations of **more than 10M core hours** will increase - from **10% to 25%** - while those requiring **less than 5M core hours** will decrease - from **80% to 65%;**
- typical jobs are memory intensive and require fast I/O, with CPU being preferred, but with GPU and cloud computing becoming increasingly important;
- current **total reported usage** lies in the range of **150-200M core hours per year**, coming mostly from NCI, Pawsey, and OzSTAR, as well as from institutional and state HPC facilities. Approximately **25% of this usage** was reported to come from **overseas** - Europe and North America - accessed via collaborators.

The second snapshot, from August 2019, is based on a survey carried out by Prof Mark Krumholz at The Australian National University, at the request of the AAL SAC HPC working group. The focus here was on research team usage as reported by team leaders; there were 28 responses from most of the major astronomical research institutions in Australia, covering the requirements of approximately 100 researchers, predominantly working in theory and computation. This shows that,

- of order **130M core hours per annum** is used on **Australian facilities** - approximately 60% at NCI, 15-20% at Pawsey and OzSTAR, and the remainder coming from institutional and state HPC facilities;
- of order **150M core hours per annum is used on overseas facilities**, in Asia, Europe, and North America. Most (>95%) of this usage is concentrated in **large, individual allocations of >15M core hours**, and is used by the groups that also have the largest allocations on Australian facilities.

There are a number of points worthy of note here;

- It is surprisingly hard to pin down the exact HPC usage per annum by the Australian astronomical community, but based on the surveys, a **reasonable estimate** of the total current annual usage is **of order 300M core hours**.
- However, this number is an **underestimate** - the current HPC requirements of radio astronomy data processing are of order 50-100M core hours per annum, and will rise with full ASKAP, etc... This suggests a current total astronomical HPC requirement of **400M core hours per annum**, likely rising to **500M core hours per annum in the next 5 years**.
- At present, the current **NCMAS framework does not provide a route to accessing large, individual, allocations of time** (>15M core hours), which is possible via overseas schemes such as PRACE. This places a **hard limit on the kind of science that can be done on Australian HPC facilities**, which has implications not only for the kinds of communities that Australian astronomy can support, but also the effectiveness of existing communities.

The Theory Perspective

While the Australian theory community is notable for the breadth and diversity of its research programme and the techniques it employs, **HPC is the fundamental component of the community's research activities**, whose success links intimately to ongoing investment in world-class HPC facilities and related services.

A strong, consistent, message from the theory community is the need for **regular, ongoing, investment in world-class astronomy-dedicated HPC access – approximately 300M core hours per annum,** with access to **multiple, large, individual allocations of HPC time (>15M core hours).** This can be supported via **regular** - and periodically major - **upgrades of, and co-investment into, our national HPC facilities**, as well as opening up access to overseas Tier-0 facilities, such as Partnership for Advanced Computing in Europe (PRACE). Investment should also go into securing continuous access cloud computing (~5k VCPUs), and ring-fencing OzSTAR as Australian astronomy's Tier-1 facility with regular 3-5 year upgrades.

Equally strong and consistent is the message that a **well-resourced and sustainable Astronomical Data and Computing Service** (ADACS) is essential to provide **crucial specialist expertise** to develop new algorithms and software, and to maximise the scientific return on investment in HPC.

The Radio Astronomy Perspective

Processing of radio astronomical, particularly interferometric data, is an extremely challenging and intensive computational task, which can involve multiple algorithmic and human iterations over enormous data volumes before producing science-ready outputs. Consequently, **access to HPC platforms and support are every bit as crucial to the success of projects such as the MWA and ASKAP as the instruments themselves.**

The Pawsey Centre for Supercomputing provides dedicated computing resources for ingest processing of ASKAP data through YANDAsoft (formerly ASKAPsoft) on Galaxy, a ~200 Tflop machine dedicated to radio astronomy; this corresponds to ~40M CPU hrs/yr. Before ASKAP became fully operational, Galaxy also served as an important processing resource for MWA science teams. A replacement for Galaxy is a priority within the Pawsey capital refresh programme.

MWA and ASKAP science teams also use Pawsey, NCI, and OzSTAR resources for further science processing. In 2019, radio astronomers were awarded 8.5M CPU hours on Pawsey, the majority of this (~7M hrs) going to MWA processing. **As planned ASKAP surveys begin taking data**, it should be anticipated that the **processing requirements of ASKAP science teams will increase significantly**, probably comparable to the dedicated Galaxy resources. Likewise, i**f the MWA completes an upgrade to a 256T system**, processing of those data would require a **substantial increase in compute resources**.

By 2025, it is plausible that **effective utilization of MWA/ASKAP data will require >100M core hours per annum**. If the capital refresh of Pawsey delivers ~10 times the compute infrastructure of the existing hardware, and if ~25% of this infrastructure continues to be devoted to radio astronomy, then these resources are achievable. However, this will require strategic direction and increased partner investment (~$1M+/yr).

Future Outlook

By 2025, the Australian astronomical HPC user community, which cuts across predominantly radio astronomy and theory, but also includes the gravitational wave, high energy, and optical/IR/UV astronomy communities, will require of order **500M core hours per annum** to carry out its various research programmes.

Based on current usage and planned investments, this figure should be achievable. Arguably the most significant stumbling block will be gaining access to **multiple, large, individual allocations of HPC time (>15M core hours).** However, we recommend **co-investing** as a **partner in our national HPC facilities** and supporting their **regular, and periodically major, upgrades.** At the same time, we recommend building **relationships with overseas partners** to give our community access to Tier-0 facilities, such as Partnership for Advanced Computing in Europe (PRACE). Finally, we recommend making **smarter use** of our existing **HPC** facilities to ensure that they are used as efficiently and effectively as possible (e.g. migrating workflows onto the cloud, redesigning codes to exploit the architectures, etc...), which utilises the **specialist expertise** in the **Astronomical Data and Computing Service** (ADACS).

Data Management and Storage

Overview

Astronomical data management and access is one of the fundamental pillars of research. Each of the fundamental questions posed in the Decadal Plan (2016-2025) are currently supported by the ASVO nodes and will continue to be supported by the three data centres in development. The data being stored by the ASVO also supports ARC Centres of Excellence, in particular, ASTRO 3D.

It is essential that users feel engaged with the data and tools being offered by the ASVO and the future data centres. The ASVO has carried out a User Preferences survey that found that the two main areas to be addressed are **interoperability of data centres/ASVO nodes** and **long-term sustainability**. Implementing the set of protocols developed by the International Virtual Observatory Alliance (IVOA) goes some way to addressing the interoperability question, and many of the ASVO nodes are part way through doing this. Addressing long-term sustainability is one of the goals of this white paper.

Current Status

Radio

Australia has a long history at the forefront of radio astronomy through instruments such as Parkes and the ATCA. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) provides access to data from these instruments through the Australia Telescope Online Archive and the Parkes Pulsar Data Archive. As identified in the decadal plan, a top-level priority for Australian astronomy is the continued development and operations of the Murchison Widefield Array (MWA) and the Australian Square Kilometer Array Pathfinder (ASKAP). Both of these instruments produce enormous data rates which present challenges of storage, processing, and human capacity.

ASKAP is, or will be soon, fully operational, producing raw data rates of ~75PB/yr. These data are processed through YANDAsoft (formerly ASKAPsoft), stored at Pawsey, and managed by the CSIRO ASKAP Science Data Archive (CASDA). CASDA serves these science data products to both the ASKAP dedicated surveys and community users through an ASVO portal.

The MWA archive has an allocation of 40PB at the Pawsey centre which is at present about 75% full (~5 PB/year since 2013), mostly in the form of uncalibrated visibilities. Should the MWA expand its signal chain to allow correlation of the 256 existing tiles, there would be an increase in the data rate of at least ~4X. MWA data are available to the community though the MWA-ASVO portal which to date has served over 1.5PB. MWA and CASDA ASVO are supported by AAL/NCRIS ($0.7M / 2yr).

MWA and ASKAP science users also have access to ~2.7PB of disk storage at Pawsey which is used for science processing on Pawsey platforms. Science teams also make use of smaller merit/partner storage allocations at NCI and OzSTAR (<1 PB total). It should be noted that the capacity and availability of this storage has often been one of the limiting factors in the effective exploitation of MWA/ASKAP data.

Optical & Infrared

Australia has a strong history of maintaining access to raw observational data through the Anglo-Australian Telescope (AAT) archive, however archives for optical science data products have had some success (e.g. the Galaxy And Mass Assembly survey) and some challenges (e.g. the WiggleZ survey). Since the Decadal Plan (2016-2025) was published, the AAO Data Central project and the SkyMapper archive have come online and largely resolved this problem, by providing long-term stable access to optical and infrared data sets of national significance. What is needed now is a way to allow astronomers to carry out research through a data portal rather than downloading all the data.

The Optical Data Centre (ODC) is building on this success by providing not just access to optical and infrared data, but the tools for survey science more generally. The system is being built as an extension to AAO Data Central ASVO node and incorporates the SkyMapper ASVO node. It is currently funded at the level of $400K/yr for the next two years, with an additional $345K/yr to support integration into NCI. The ODC currently provides access to data from the Anglo-Australian Telescope (AAT), the UK Schmidt Telescope (UKST), the Siding Spring Observatory 2.3m telescope (SSO2.3m) and SkyMapper. It's total data holdings are around 1.2PB, with the bulk of that made up of SkyMapper imaging data.

The ODC has preliminary interoperability with four of the ASVO nodes (Data Central, SkyMapper, CASDA and the MWA ASVO node) via a common cone search tool, with more advanced features to come. All ASVO nodes except MWA can currently be accessed via Single-Sign-On against Data Central accounts.

Theory

The Australian theory community is noteworthy for its breadth and diversity, with active communities working on topics that include **cosmology**; **planet, star, and galaxy formation**; **dynamical modelling of star clusters**; **stellar astrophysics**; and **astro-particle physics.** A range of techniques are employed to perform the calculations involved, and, as noted already in the section on high performance computing, they are computationally intensive and can produce enormous volumes of data. For example, large cosmological N-body simulations that underpin the latest generation of galaxy formation models produce >100 TB of raw particle data and take >1M core hours per run, while the latest state-of-the-art calculations of turbulence using adaptive mesh refinement hydrodynamical simulations produce similar data volumes and can take >10M core hours per run. Typically, such rich datasets are mined for years after the calculation, most often in the form of derived data products that can be utilised by the wider community. The sophistication with which this dissemination is done will depend on the precise nature of the data and how it is to be used.

Currently, the theory ASVO node is the **Theoretical Astrophysical Observatory (TAO)**, which provides a portal for the delivery of mock galaxy catalogues built from theoretical (semi-analytical) galaxy formation models that have been run on the outputs from a suite of large N-body simulations, including the Millennium Simulation. These mock catalogues can be constructed using the same selection criteria of an ASKAP HI survey, say, using the predicted properties of galaxies in the models. TAO can deliver catalogues that are either static and pre-made, or dynamic and tailored to user criteria, and can be utilised by teams of observational astronomers to plan their surveys and interpret their results, while also providing computational astronomers with the means to share their data products.

Future Outlook

Optical Data Centre

The future data requirements of the ODC are **a minimum of 2PB**; the kind of secure storage required is only available via a national facility such as NCI. The ODC will host and provide access SkyMapper data during and after the Southern Sky Survey, and the data will become more integrated into the system. Data from optical and infrared surveys of national significance will continue to be hosted.

The ODC will provide access to, and possibly host, data from the LSST and ESO telescopes. The cost of hosting the LSST archive is very high however, given the size of the final dataset (500PB after 10 years), so **the return on investment for hosting LSST is low**. While the cost of hosting a mirror of the ESO archive is much more manageable (current data volume is ~1PB), there is no science case currently for hosting the data locally. Combining virtual observatory tools and the innovative technologies used by the ODC, it is possible to provide access to these systems for Australian users "under-the hood".

One of the main goals of the ODC is to support the full data life cycle for all data it collects. This includes data management and processing during a survey, access to survey teams to their data before publishing a data release, publishing the data, and long-term hosting of raw and science data products. As part of this work, the following components are planned or already underway:

- To support the growing number of transient science projects, the ODC must move to **real-time processing of optical data** collected by the AAT and the SSO2.3m. Over the next two or so years, this system will be implemented first at Data Central and then at NCI.
- **Authenticated access** to private survey team data, coupled with **an authorisation scheme**, is a critical part of a federated national data centre.

Australian SKA Science Regional Centre

Towards the end of the decadal plan period the provisioning of the Australian SKA Science Regional Center (AusSRC) will have a major impact on data and HPC processing for radio astronomy in Australia. Ultimately, the AusSRC is expected to be a national scale facility which will provide archiving and dissemination of SKA data, employ SKA processing experts who will provide software, workflows and user support to facilitate the utilisation of SKA data, and provide physical computing infrastructure to allow Australian researchers to compute at the data. Before SKA data become available, the AusSRC is expected to take a leading role in the data processing and archiving needs of ASKAP and the MWA as SKA precursors. However, the interaction and degree of overlap between these resources and the forthcoming capital refresh of the Pawsey centre, as well as the roles of domain specialists with the AusSRC and national resources such as an expanded ADACS, remain open areas of strategic planning.

Gravitational Wave Data Centre & Fast Radio Bursts

Gravitational waves and Fast Radio Bursts (FRBs) are two areas of astronomical research which have achieved international prominence and the growing importance to the Australian community in a manner not fully anticipated within the original decadal plan.

The importance of Australia's involvement in gravitational wave (GW) research has been recognized through the NCRIS funding ($2.8M over two years) of the Gravitational Wave Data Centre (GWDC) which will provide dedicated infrastructure and personnel to support the hosting, real-time processing, and more extensive analysis of GW data from existing and pending GW observatories as well as pulsar timing experiments. The continued support of this facility should be considered a high strategic priority.

The processing of FRB data from ASKAP and UTMOST is comparable to other planned ASKAP surveys but has a distinct nature; particularly the coherent processing of real-time data would be best served by dedicated computational resources not typically available through normal HPC merit allocations.

Theory

Theoretical datasets are no different from observational datasets insofar as the data needs to be **managed**, it is usually **made public** after some proprietary period, and it needs to be **hosted long-term**. The key question is how these data get used - by small collaborations that can interact with the raw data on the HPC facility where it is generated, or by a larger community that queries a database of derived data products. As highlighted already, the theory community is broad and diverse; the data are heterogeneous; and the mode in which the data is used is field-specific. However, there are some general observations worth making.

In general, **processing of raw data should take place at the HPC facility where it is generated and stored**. How much raw data is retained or placed into deep storage will likely depend on the expense of generating it at some later time, or how much information is necessary to reconstruct it. In some instances, e.g. large volume cosmological N-body simulations, processing (e.g. halo catalogues, merger trees, production of lightcones) will happen increasingly inline, reducing the data volumes that need to be stored. In other instances, the richness of the dataset means that preserving the full N-dimensional information is essential. The creation of realistic mock observables that allow for more accurate comparisons of model predictions and observational data will continue to grow in importance, which will require additional (at times, computationally expensive) post-processing and appreciable data volumes.

These considerations suggest that, at least in the case of data that is generated at the primary Australian HPC facilities used by Australian theorists and simulators, these **data will be co-located at the site of existing data centres** (e.g. the Optical Data Centre at NCI), and the public-facing data can utilise the same infrastructure - although the details will differ, the methods for querying and delivering the data should be the same. Datasets that are generated at other Australian institutional or state facilities, or overseas, should be ingested to one of the data centres that could act as long-term host.

Ensuring that the framework for theoretical datasets follows the FAIR guidelines (findable, accessible, interoperable, and re-usable) is crucial - boosting the impact of both the theoretical datasets and the observational teams that can use them in their analyses, as well as facilitating collaborations between theorists (e.g. gravitational wave astronomers, dense stellar cluster simulators, and theoretical stellar astrophysicists).

Towards a Federated National Data Centre

One of the key outcomes for the three data centres discussed in this white paper **must** be that they are **seamlessly interoperable**. It is important that the ODC, GWDC and AusSRC build upon the work done by the ASVO if there is to be a truly federated national data centre. A coordinated approach is required to avoid “re-inventing the wheel” and it is **important to engage with the international community** who are also working in this space. It is also essential that each of the data centres has the Subject Matter Experts it needs in-house.

ADACS – Astronomy Data And Computing Services

As ever larger datasets and computational tasks become the norm in astronomy, there is a broad consensus that there is an opportunity to train and retain a new class of specialist researchers. These roles fall between more traditional researchers, who primarily strive to produce publishable scientific results, and HPC platform specialists, who administer, optimize and maintain the physical computing infrastructure. They typically write software to acquire and process data from increasingly complex and powerful instruments, as well as creating software tools and workflows, which allow other researchers to effectively access and extract scientific results from huge datasets.

Despite the recognition of their importance, there are a number of issues and risks surrounding these roles:

- Their work often does not lend itself to publication in astronomical journals; hence they are not well-positioned to advance into tenured academic positions.
- Their work often requires a high degree of astronomical domain knowledge which makes them not entirely suited for positions with general HPC facilities such as NCI/Pawsey.
- Their employment is often funded on fairly short funding cycles which can lead to difficulties in attracting and retaining qualified individuals, particularly in more senior roles.
- Access to their skills and expertise is not always effectively aligned with the needs of the community

One existing effort within the Australian community to address these issues is the Astronomical Data And Computing Services (ADACS). ADACS provides training and workshops for software skills, as well as merit-based software engineering support to researchers or projects. Presently, ADACS is funded through AAL over a two year cycle for ~$1M/year. This model offers a possible path to increase the astronomical software workforce; a software support position(s) which is embedded in a specific project, but remains closely tied to a nationally federated cadre of e-researchers which can more flexibly serve the community’s needs as they arise and which can serve to smooth out cyclical funding of a particular project. ADACS builds upon the already existing capabilities of the astronomical software teams at ANU and AAO Macquarie, both of whom have been serving the community via instrumentation and data pipeline projects. **Each of these capabilities should be maintained**.