



Astronomy Australia Ltd.

Response to NeCTAR consultation paper

November 2010

Written in consultation with:

Matthew Bailes; Lindsay Botten; Brian Boyle; Tim Cornwell; Darren Croton; Andrew Hopkins; Jarrod Hurley; Peter Quinn; Andrew Rohl; Brian Schmidt; Steven Tingay.

PO Box 2100, Hawthorn, VIC 3122
Telephone: +61 3 9214 8758; Fax: +61 3 9214 4396
amanda.beasley@astronomyaustralia.org.au
www.astronomyaustralia.org.au

Executive Summary

Astronomy Australia Ltd's response to NeCTAR's consultation paper is in the form of a vision for an astronomical project leveraging off new facilities due to generate a flood of data within the next few years.

Astronomers in Australia today use a variety of facilities from optical and radio telescopes to supercomputers to conduct their research. Their data paradigm is 1-10 Gb/s data rates or 50 TB simulations.

New facilities for all-sky survey astronomy that have recently been funded through a more than \$200 million input into astronomy include SkyMapper (an optical telescope in NSW expected to generate 500 TB of data over 5 years), HERMES (the next major instrument for the Anglo-Australian Telescope (AAT) run by the Australian Astronomical Observatory (AAO)), ASKAP (a radio telescope in WA expected to retain 5-10 PB per year of data), MWA (a radio telescope in WA that will achieve at least 6 PB per year of data for 3-5 years) and gSTAR (a supercomputing facility at Swinburne University that will have a peak theoretical performance in excess of 100 Tflops). In addition, two Data Centres that support astronomy among other sciences, the Pawsey Supercomputing Centre for SKA Science (run by iVEC), and the National Computational Infrastructure (NCI), have received significant funding to enable their creation (for the Pawsey Centre) or their upgrade (for NCI).

The Australian Decadal Plan for Astronomy seeks to position Australian astronomy as a significant, competitive and productive player in the global-scale programs and projects of 21st century astronomy. To maximize scientific return from the new investments, the astronomy community has advised Astronomy Australia Ltd that a system for federation of astronomy data is needed. Data federation would involve creating the hardware, tools and services to bring together data from ASKAP, MWA, SkyMapper, the AAO and gSTAR, i.e. radio, optical and simulation data from all parts of the southern sky, under a Virtual Observatory. The federation of astronomy data would be an effective mechanism to enable astronomers to participate in global, multi-wavelength survey science and to share the science projects based around Australian survey facilities.

iVEC, NCI and Swinburne University, as existing astronomy data hubs, are the natural hosts for the various hardware and software required to achieve data federation. The work to enable federation will involve creation of data ingestion software systems, VO-compliant services, high performance database systems, VO-compliant data access layers, and VO systems and interface tools specific to MWA, ASKAP, SkyMapper, the AAO and gSTAR, including service directories and desktop functionality. Some of the expertise generated in the creation of technology and software for astronomy could be transferrable to other sciences served by these data centres.

This project will be dominated by manpower and will require of order 30-35 person-years of effort over a 4 year period with a cost of order \$5 million. There will be additional hardware, software and support costs in the range of \$2-4 million depending on further subsystem analysis of Pawsey and NCI planning. Hence the total 4 year project cost will be in the \$7-9 million range.

Contents

Introduction	4
Astronomy today	4
Planning for the future.....	5
Astronomy tomorrow	5
New Facilities	5
Maximizing scientific return.....	7
Achieving Federation	8
Conclusion.....	12
Appendix	13
eResearch Tools	13
Virtual Laboratories	14
Research Cloud	15
National Servers.....	17
ARCS tools and services	17

Introduction

This response is in the form of a description of a project for astronomy that comes under the banner of NeCTAR and RDSI investments. Answers by astronomers to specific questions asked in the NeCTAR consultation paper can be found in the appendix.

Astronomy today

Astronomers in Australia today use a variety of telescopes and facilities. National facilities of note include the 3.9m optical Anglo-Australian Telescope (AAT), operated by the Australian Astronomical Observatory (AAO) at Siding Spring Observatory in NSW, the CSIRO operated 64 m radio telescope at Parkes, NSW ("The Dish"), and the six-dish Australia Telescope Compact Array (ATCA) also operated by CSIRO at Narrabri, NSW.

Supercomputing facilities available to astronomers nationally include the National Computational Infrastructure (NCI) National Facility in Canberra (a Sun Constellation with a peak performance of 140 Tflops) and Swinburne University's Green Machine in Melbourne (10 TFlop supercomputer (theoretical peak) with 1,280 cores (8 cores/node; Clovertown), 16 Gb memory/node, a 250 Tb file system and Infiniband interconnect). These computing facilities, when in use by astronomers, are mostly used to create astronomical simulations to answer specific questions about the universe, but also to host data from telescopes requiring powerful processing.

Astronomers work in a paradigm of data rates of order 1-10 Gb/s (for example for VLBI streaming from Parkes). This data can sometimes be "reduced" on site (initial processing), and then transported to the astronomer's home institution via DVD, or over the internet. Alternatively, multi-Gb/s of data will be transported raw to the home institution for processing – in some specialized cases, data will be transferred to a tape or disk drive (for example, VLBI requires a 10-20 TB raid disk set). The data are often processed by the astronomer using their own institution's facilities.

Supercomputer generated datasets, on the other hand, are necessarily very large. For example, the largest dataset generated on Green to date is a 50 TB cosmological simulation of the large-scale structure of the universe under the influence of dark matter and dark energy.

The data astronomers collect are generally for themselves or their collaboration. If an astronomer outside the collaboration wishes to access it (for example, after reading a journal article about results from the dataset), they must wait a proprietary period (often 18 months), after which they can download the dataset from the telescope facility. Alternatively astronomers can request early access to data through personal contact with a collaboration member. Due to their size, supercomputer simulations the datasets must generally remain at the host facility with access granted via a login.

This system of data access can scientifically limit the work of astronomers. For example, an optical astronomer investigating galaxy rotation may wish to view the one area of sky at both optical and radio wavelengths, or perhaps compare their observations with a simulation. Currently this is a laborious task

for the astronomer, and may not even be possible in some cases. The existing method of data access, processing and transport is only feasible when the datasets involved are relatively small.

In the future this limitation will become more apparent with new telescopes and supercomputers generating very large amounts of data.

Planning for the future

Astronomy Australia Ltd's (AAL) High Performance Computing Working Group (HPCWG) spent 2009/10 investigating the HPC facilities currently available to astronomers, and provided advice to the AAL Board in the form of a report "The Future of eScience and High Performance Computing for all Australian Astronomers" (available from http://astronomyaustralia.org.au/publications/HPCWG_Final_Report.pdf). The purpose of these investigations was to discover the preferred methods astronomers wanted to use to tackle the next generation of data coming online in the future.

In addition to the work of the HPCWG, AAL held two workshops during 2010. The first workshop, held in May, was for the astronomy community to discuss their needs and plan for the future. The second workshop, held in September was for astronomers and eResearch providers to come together to determine a way forward. Attendees at this meeting included DIISR, NCI, AARnet, Intersect, ANDS, ARCS and AeRIC, as well as senior astronomers.

Astronomy tomorrow

In the past three years, Australian astronomy has been very fortunate in receiving a significant investment in research facilities funding from Federal and State Governments. In particular astronomers have received more than \$200 million for the funding of ASKAP (Australian Square Kilometre Pathfinder), the MWA (Murchison Widefield Array), and the Pawsey HPC Centre for SKA Science in Perth. The AAO has received funding for its next major instrument for the AAT: HERMES. In addition NCI received \$50M under the Super Science Initiative to upgrade to a next-generation facility. Similarly, Swinburne University will receive \$1M from AAL's EIF grant to build gSTAR (GPU Supercomputer for Theoretical Astrophysics), which will be part of a larger supercomputing facility run by the University.

New Facilities

SkyMapper

SkyMapper is a 1.35 m optical telescope located at Siding Spring Observatory in NSW, and run by the ANU. SkyMapper's mission is to robotically create the first comprehensive digital survey of the entire southern sky. The survey will be a massively detailed record of over a billion stars and galaxies, to a depth that is one million times fainter than the human eye can see. SkyMapper will photograph each part of the Southern Sky 36 times and create approximately 500 TB of data over five years. SkyMapper has recently seen "first light" and will begin operations in 2011.

HERMES at the AAT

HERMES is the next major instrument for the 3.9 m Anglo-Australian Telescope (AAT), to be operational from 2012, currently being constructed by the Australian Astronomical Observatory (AAO). The HERMES system is built upon the AAT's existing two-degree field (2dF) optical fibre positioner, which can collect the light from 400 stars at a time. The positioner, which currently provides the highest multiplex for obtaining optical spectroscopy in the world, feeds a powerful new spectrograph which covers four optical bands simultaneously at a spectral resolution of 28000. The key HERMES science project is the "Galactic Archaeology" Survey, which aims to reconstruct the history of our Galaxy's formation from precise multi-element abundances of one million stars derived from HERMES spectra. The AAT continues to be the most efficient telescope in the world for rapidly obtaining large numbers of optical spectra, and to date the AAO has measured about 40% of every galaxy redshift ever measured.

ASKAP (Australian Square Kilometre Array Pathfinder)

ASKAP will be a 36-antenna array radio telescope located at the Murchison Radio Observatory (MRO) in WA, owned and operated by CSIRO as a National Facility. Each dish is 12 m in diameter, and the full telescope will be completed in 2012. The first dish is already operational and recently participated in an Australasia-wide connection of telescopes to create a detailed image of the heart of the Centaurus A galaxy. During ASKAP's first five years of operation at least 75% of its time will be used for large Survey Science Projects. ASKAP's data rate is expected to be about 80 PB/year and the project intends to keep 5-10 PB/year of this.

MWA (Murchison Widefield Array)

The MWA will be a 512-tile array radio telescope, also located at the MRO and is currently being built by an international consortium. MWA has recently achieved a 32-tile prototype and the full array is expected to be operational in 2011. The science MWA is designed for includes investigating Galactic and extra-Galactic phenomena including pulsars, the interstellar medium and early universe astrophysics. MWA could expect to generate in the order of 6 PB/year for 3-5 years depending on observational parameters. The total volume of data over the life of the project could be significantly higher.

The Pawsey Supercomputing Centre for SKA Science

As part of the Super Science Initiative, the Commonwealth Government allocated \$80 million to Western Australia's supercomputing hub iVEC to establish a petascale supercomputing facility. The aim is to enhance Australia's position in the international supercomputing field and to boost its bid for the Square Kilometre Array Project. The Pawsey Centre will provide a top 20 supercomputing facility to support the needs of the Australian radio astronomy research community, as well as researchers in other areas of computational and data-intensive science, such as nanotechnology, biotechnology and geoinformatics. The investment will also provide additional support for the computational and data processing capabilities required to fully implement the Australian Square Kilometre Array Pathfinder (ASKAP) and Murchison Widefield Array (MWA) radio telescopes.

gSTAR (GPU Supercomputer for Theoretical Astrophysics)

gSTAR is a specialised computing cluster to be based at the Centre for Astrophysics and Supercomputing at Swinburne University starting mid-2011, and is part of a \$3 million upgrade of the Green Machine. The gSTAR supercomputer will provide the national astrophysical community with a CPU/GPU facility for performing world-class simulations and to enable rapid processing of telescope data (e.g. Parkes). gSTAR is expected to have a theoretical peak performance of in excess of 100 Tflops of computing power and will be connected to a peta-scale storage system.

NCI (National Computational Infrastructure)

NCI is the national, high-end, research computing service, established under an agreement between DIISR and ANU in 2007. It builds on the successful National Facility program of APAC (2000–06), and the ANU Supercomputing Facility, established in 1987, and delivers a world-class service supporting climate and environmental science, astronomy and astrophysics, computational biology and chemistry, engineering and fluid mechanics, medicine, physics and photonics. Under the Super Science initiatives announced in the 2009 Commonwealth Budget, \$50 million was allocated to NCI focused on providing a substantial enhancement in the modelling capability for climate science, earth system science and national water management. The Project Plan signed by ANU and DIISR in May 2010 is carrying this forward through an extension of the role of NCI, and provides for the establishment of a comprehensive “digital laboratory” to serve the national priority in climate change and other areas of high impact science (which include astronomy and astrophysics).

Maximizing scientific return

The facilities described above are the result of the research community’s appreciation that some of the most fundamental questions that can be asked of the Universe will only be answered through the statistical strength, the all-sky integrated signals and the discovery of rare systems and events, that result from an all-sky approach to gathering data. This universal “buy-in” is strongly reflected in the European, US and Australian decadal planning processes and priorities. In Australia, a large fraction of the astronomical community (at Sydney University, Curtin University, the University of Western Australia, the University of Melbourne, Swinburne University of Technology, the Australian Astronomical Observatory, and the Australian National University) is currently collaborating on survey science through the recently funded \$28M ARC CoE in All-Sky Astrophysics (CAASTRO) which enables an all-sky approach to research from panoramic multi-wavelengths facilities.

Scientific advantage and new opportunities for discovery from these large all-sky surveys can be gained through astronomers being able to easily access data and results from different types of facilities. Creating the fabric to allow the new generation facility data to be “federated” was a key recommendation from the HPCWG’s report. Data federation will allow astronomers across the country to engage with these new facilities coming online, and undertake new areas of investigation.

The new generation of telescopes being built have been designed to be flexible in data rate output. Data output can be increased or decreased depending in the computational and storage facilities available. For example, the output data rate of MWA will be able to be varied depending on the level of averaging

in time and/or frequency. The greatest scientific return is naturally gained when the data output is highest. Additional investments in HPC/storage/network infrastructure and tools will allow astronomers to extend the capabilities of new instruments, and maximize the scientific return from them.

Achieving Federation

To bring into existence the overall vision of federation of astronomy data from ASKAP, MWA, SkyMapper, AAT/HERMES and gSTAR, the “connective tissue” is required.

The data path from telescope and supercomputer to researcher is slightly different for every facility, but the core requirements are similar. Between the researchers and the large hardware resources, there will need to be systems that enable:

- discovery of data products,
- queries on those products across a range of telescopes and simulation,
- delivery and sharing of the results of those queries.

This middle layer consists of database technologies, data ingestion and access layers and data interoperability standard systems which are tightly coupled to the HPC below and above to specially designed interfaces on researchers desktops. Many of the standards and systems needed to enable data discovery and multi-wavelength interoperability have been developed by the International Virtual Observatory Alliance (17 international VO projects) in the past eight years.

To create an efficient federated system, all component datasets need to comply with these standards. SkyMapper data are already planned to be IVOA compliant, as are simulation datasets from gSTAR. Formatting data to IVOA standards allows the data to become persistent over time, leading to the possibility of long term archiving.

Taking full advantage of already-funded Data Centres, astronomy data federation can be achieved through the co-operation of three natural hosts – iVEC, ANU and Swinburne University. These organisations hold the computing facilities that would accommodate data storage, processing and database hosting for the SkyMapper, MWA, ASKAP, the AAO and gSTAR.

These hosts, each already with a strong astronomy component of their operations, would increase the likelihood that the data federation model could continue past the lifetime of the projects and onto the future generation of facilities, such as the SKA.

Work to enable federation

The work to enable federation across these diverse facilities will involve the following:

1. design, build and installation of data ingestion software systems

2. design, build and installation of VO-compliant services
3. design, build and installation high performance database systems
4. design, build and installation of VO-compliant data access layers
5. design, build and delivery of AUS-VO systems and interface tools specific to MWA, ASKAP, SkyMapper, the AAO and gSTAR needs including service directories and desktop functionality

The ingestion systems and high performance database systems would be particular to each host, but the VO-compliant services, data access layers and AUS-VO systems would be largely a shared effort.

Scale

This project will be dominated by manpower and will require of order 30-35 person-years of effort over a 4 year period with a cost of order \$5 million. There will be additional hardware, software and support costs in the range of \$2-4 million depending on further subsystem analysis of Pawsey and NCI planning. Hence the total 4 year project cost will be in the \$7-9 million range. Realising the theory arm of data federation would require 1-2 PB of storage as an extension to gSTAR and 1-2 support personnel for two years to design and maintain the virtual laboratory and tools. This would also require the appropriate hardware (e.g. servers) and database software (custom or stock) that would form the backbone of the system.

The grand view is a federated data network that enables new science, theoretical and observational, underpinned by the hardware of NCI, Pawsey and gSTAR. Thus maintaining a high-speed network (10 Gb/s minimum) between the Data Centres is essential.

The Hosts

1. *iVEC in Perth, WA, runs the recently funded Pawsey Centre. Connected to the MRO, the Pawsey Centre is the planned host for initial MWA and ASKAP data, as well as data generated by geophysicists.*

iVEC was established in 2000 to foster and promote scientific and technological innovation, through the provision of supercomputing and eResearch services to the research community, commercial organisations and government agencies. It is an unincorporated joint venture of CSIRO and the four public universities in Western Australia and has been strongly supported by the WA Government, receiving nearly \$13 million to date. iVEC is a distributed organization and provides access to supercomputing, large scale data storage and visualization facilities across the Perth metropolitan area. The radio astronomy community has extensively used the data storage facilities, with iVEC currently holding more than 400 terabytes of astronomy data. Through being tasked with establishing and operating the \$80 million Pawsey Centre, iVEC is set to become one of the most significant supercomputing facilities in the world.

The data federation arm for data generated by radio telescopes will be closely linked to the Data Intensive Research Pathfinder (DIRP) project within the International Centre for Radio Astronomy Research (ICRAR) who will be implementing necessary foundation layers for federation within the Pawsey Centre.

There is a great deal of similarity between the data cubes produced by radio astronomy and the data cubes resulting from geoscience surveys of the Earth's crust. The data sizes and processing (algorithms, services) complexity are similar. Searching within geoscience cubes for mineral resources is comparable to finding transients and/or galaxies within radio astronomy data. ICRAR will be pursuing this synergy within the DIRP project. This can be extended to the developments contained within the data federation project and effectively connects the geoscience and space science super science priorities.

2. **ANU in Canberra, ACT, runs the National Computational Infrastructure (NCI).** *NCI will host SkyMapper data and also runs large scale astronomical simulations. NCI also has a strong focus on supporting the activities of climate scientists.*

The work of NCI is supported by the Commonwealth through funding provided through NCRIS and the Super Science program, and through the substantial co-investment of a number of partner organisations (including ANU and CSIRO) whose combined annual co-investment is approximately \$7.5 million in 2010.

In addition to the services provided to researchers through partner shares, NCI supports leading Australian science through its Flagship Merit Allocation Scheme funded by NCRIS investments. This presently supports three ARC Centres of Excellence—one of which is CAASTRO that has been allocated 4M hours per annum. NCI also invests in the provision of support for two national research communities— climate science and astronomy. In the case of the astronomy, NCI funds two full-time positions for two years at ANU and Swinburne, respectively focusing on optical astronomy data and high scaling astrophysical simulation, and radio astronomy and GPU-based computation.

The infrastructure funding from the Commonwealth provides for a new data centre (approx. 900 sq. m.), and a new petascale HPC facility, while the recurrent costs (which exceed the annualised value of the infrastructure by around 40%) must be provided through the investment of co-investing partners—the three largest of which will be ANU, CSIRO and the Bureau of Meteorology, and which between them will be providing for at least three quarters of the recurrent expenses.

The present peak facility, commissioned in April 2010, is a Sun Constellation which provides an internationally significant peak performance of 140 Tflops, and which is a well-balanced facility that sustains high throughput, and is thus ideal for highly scaling astrophysical simulation. Also at the NCI National Facility at ANU is a data cloud and storage (that supports the needs of climate science and astronomy, including SkyMapper, major surveys with the AAT, and other optical instruments), and a strong support team that is acknowledged nationally and internationally for its expertise.

The next generation peak system will be commissioned in mid-2012. It will provide full service operation and will comprise a high performance computing system (petaflop performance), large scale data storage for nationally and internationally significant collections, a high throughput cloud to provide for

crucial data-intensive services, a comprehensive layer of computational / data services, and a strong support / development team possessing the expertise needed to deliver these services, to address international research challenges, and to add value to the infrastructure investment.

Presently, there are 16 (systems and applications) staff associated with the operations of the NCI NF, and a further six staff associated with specialist support in astronomy, climate science, and the development of cloud computing services funded by NCRIS resources. The NCI Business Plan anticipates around eight additional staff, funded by partner co-investment in order to boost the impact of high-end computing services in the designated national and partner priority areas, and further staff through the NeCTAR Tools and Virtual Laboratories programs to deliver the data-intensive services required in climate science, impact and adaptation, and astronomy and astrophysics.

NCI further expects to host a major national node of the RDSI initiative serving climate science, the environment, astronomy, and other research communities.

3. **Swinburne University** in Melbourne, VIC, hosts the *Centre for Astrophysics and Supercomputing (CAS)*. One of the areas of expertise in CAS is theoretical astronomy and runs large astronomical simulations. gSTAR will also generate data that would be part of federation.

Theoretical astrophysics represents a major research effort in Australia, with a growing reliance on High Performance Computing to solve some of the most complex problems in astrophysics. Unfortunately, making sense of the observed Universe is a messy business. This is due in large part to the fact that many of the processes that lead to complexity and diversity in the Universe are still rather poorly understood. Supercomputer simulations of the co-evolution of dark matter, galaxies, gas and black holes are essential in this respect - they are the virtual "laboratory" in which competing ideas about galactic and cosmic evolution can be tested in a controlled environment. For many of the key problems in astronomy, there is no other way.

Swinburne is already a national focal point for small, medium and large-scale "grand challenge" computer simulation projects. Swinburne's Centre for Astrophysics and Supercomputing has staff skilled in running computer simulations ranging from globular clusters to galaxies to the large-scale structure of the Universe. In addition, Swinburne University has the only astronomy centre in the country which is contained within a Faculty of Information and Communication Technology, with a large part of its focus on HPC.

The new facility, gSTAR, will have around 600 TB of disk but this will only service the expected approximately 150 user accounts with no capacity for long-term storage of data products. However, the gSTAR storage facility will be scalable, and by adding 1-2 PB of disk can greatly enhance the service provided to the astronomical community so that it can act as a theory arm of the proposed astronomy data federation model.

An example of the estimated storage needs required by the theory arm of the federation model can be gauged by assuming the theory community will run three "grand scale" simulation projects on gSTAR across a projected five year lifetime of the facility. As data products tend to double in size every

approximately 1.5 years, the first can be expected to require 75 TB of storage (the current size of the largest cosmological simulation run in the USA), the second 150 TB, and the third 300 TB, totalling a combined 525 TB of data. Space will be needed for an additional number of medium and small-scale simulations, Parkes radio data (for GPU processing), copies of critical components from the Pawsey and NCI virtual laboratories (MWA, ASKAP, the AAO, SkyMapper), as well as space for data processing and analysis, and for the general operation of the middleware and VO layers.

An example of the scientific impact of such a cloud-based simulation database is the German Astrophysical Virtual Observatory (GAVO; <http://www.g-vo.org/Millennium>). This VO interface offers public SQL access to the Millennium Simulation, a large simulation of the evolution of cosmic structure in the Universe. Through GAVO, the user base of the simulation was vastly increased, and hence also the scientific return on the initial simulation effort. To date (2006-2010) over 330 refereed papers have been published in A or A* journals by the international community using data drawn from the GAVO online tool, garnering over 6500 citations. The Swinburne theory arm of the data federation model proposed here would encompass the functionality of the GAVO online database while vastly expanding the number of simulations available to the community (from one to as many as can be housed on the petabyte data store), adding new VO tools (a light cone factory and telescope simulators), and connecting directly with next generation observational datasets from ASKAP, SkyMapper, the AAO and MWA.

Conclusion

The Australian Decadal Plan for Astronomy seeks to position Australian astronomy as a significant, competitive and productive player in the global-scale programs and projects of 21st century astronomy. The community's ability to join efforts like the SKA will depend on the scientific success and attraction of Australian facilities and the ability to work collaboratively with the international community. The federation of astronomy data will be an effective mechanism to enable astronomers to participate in global, multi-wavelength survey science and to share the science projects based around Australian survey facilities.

Appendix

In addition to the vision described above, the following specific responses to questions in the NeCTAR Consultation Paper have been sent to AAL from members of the astronomy community. They are reproduced below.

eResearch Tools

RT1: What principles could underpin the prioritisation of Research Tool development?

- Concentration on national research priorities, with an emphasis on large scale projects that have requirements across different forms of research infrastructure (networks, storage and compute) are mostly likely to generate tools of generic interest to a wide range of disciplines over time. An underlying principle in the development of tools should be the coordination of access and utilisation of these different forms of infrastructure.
- Alignment with established national research priorities, community readiness, and cost-benefit.

RT2: While we are not soliciting proposals for eResearch Tool creation at this stage, can you identify the priority areas for research tool improvement in your disciplinary area?

- The curation of large datasets (100s of TB to PB) from single instruments and the federation of these multi-wavelength datasets in single locations will require tools to support the transfer of massive datasets between several locations in Australia.
- Astronomy visualisation. Virtual Observatory (VO) queries and analysis. Pulsar analysis.

RT3: Are you aware of any existing tools or services that would be ready, at an early stage, to engage with NeCTAR as an 'exemplar' project and demonstrate value to the sector?

- Current work on a national frequency standard distribution network over standard network infrastructure has wide applicability and involves a number of communities. The work involves AARNet.
- The CASS/ATNF Parkes Pulsar Data Archive.
- The Theoretical Astrophysical Observatory (TAO) currently being developed at Swinburne.

RT4: What computing platform and interoperation standards in your field need to be taken into consideration in planning for this investment?

- Standards based on the output data formats of the national astronomical instrument suite and Virtual Observatory (VO) standards should be taken into account.
- VO standards for data formats, and remote server access.

RT6: What support do you see as necessary to promote the success of NeCTAR and the provision of sustainable eResearch tool capabilities in your field of research?

- Development of tools via development staff embedded within large priority projects (data sources) and also closely associated with major national HPC facilities.
- A broadly disseminated and demonstrated commitment to ongoing support and development.

Virtual Laboratories

VL1: While we are not soliciting proposals for Virtual Laboratory development at this stage, what research areas should be prioritised for involvement in the Virtual Laboratory component of the project?

- The suite of national astronomy instruments, including new large scale instruments with massive data volume outputs, connected with HPC centres such as Pawsey, NCI and Swinburne University of Technology, all connected via high speed networks, will form a natural Virtual Laboratory for the Australian and international astronomy communities.
- The national astronomy instruments, including those of the CASS (CSIRO Astronomy and Space Science Division)/Australia Telescope National Facility, and upcoming large data-generation facilities such as ASKAP, SkyMapper, and MWA. There is a clear national commitment to all sky astronomy (as also demonstrated in the establishment of CAASTRO), and so all sky astronomy should be prioritised.

VL2: Are there existing Virtual Laboratory or similar capabilities that would be ready to engage early with NeCTAR as an exemplar and demonstrate value to the sector?

- MWA, via ICRAR, and connected to Pawsey, would be interested in doing this.
- ASKAP is very interested and able to engage early with NeCTAR as one partner in a VL. Such an engagement would be especially timely given the current planning and development of the ASKAP Data Archive, which will be a key underlying national resource. The outcome could be envisioned as a Virtual Laboratory for All Sky Astronomy (VLASS), thus complementing CAASTRO. The value of a Virtual Laboratory would clearly be reliant on the inclusion of SkyMapper and MWA.
- The Theoretical Astrophysical Observatory (TAO) currently being developed at Swinburne.

VL3: What additional support might be required to bring these exemplars to fruition? It should be noted that this might take the shape of existing institutional or other resources that might assist with implementation.

- Existing infrastructure in high speed networks (eResearch tools to manage data transfer), large scale computing, and human resources to provide the VL framework connection to the instrument suite.
- Support would be required in the form of a well-established and productive professional quality software development team, working in conjunction with astronomical domain experts. CSIRO has such a software development team now working on the implementation of the Parkes Pulsar Data Archive (as funded by ANDS). This team serves as an exemplar and could be able to assist with the implementation given the necessary funding (similar to the afore-mentioned ANDS funding).

VL4: Early expressions of interest are sought to gain a sense of the size of the budget allocation and co-investment required.

- The level of budget allocation for a VLASS is difficult to specify with accuracy at the moment, but would be in the range \$5M to \$10M. The co-investment for CASS/ASKAP would consist of the resources (people only) being allocated to the ASKAP Data Archive, approximately \$1.5M. We note that the existing ANDS investment in the Parkes Pulsar Data Archive (\$486K) is being built upon in the ASKAP development.

Research Cloud

RC1: Does your research area have a preferred model for delivery of this component?

- The European e-VLBI community is adopting a cloud approach for the next generation of e-VLBI collaboration, in the EU-funded program NEXPRES, connecting global assets. Australian organisations working in e-VLBI are partners in NEXPRES. It would be preferable if any model for cloud delivery in Australia consider the delivery model for similar projects elsewhere in the world.
- A cloud model could be suitable given the appropriate performance characteristics. We also note that the Virtual Observatory has an evolving server-side execution model that could serve as an exemplar.

RC2: While we are not soliciting proposals for Research Cloud development at this stage, what existing applications should be prioritised for migration to a research cloud infrastructure?

- (1) e-VLBI could be a clear user of cloud.
- (2) Existing radio astronomy data analysis packages such as MIRIAD, AIPS, and CASA, and some VO analysis tools. Running these in a cloud environment would significantly ease use by astronomers, removing the need for maintenance of the software packages on individual computers or servers.

- (3) Access to theoretical data, both raw and mocked to mimic observational data, will be cloud based.

What resources might be required to carry out this migration?

- (1) Support to interface assets to the cloud and to coordinate with global partners.
- (2) 2-3 FTE-years experienced cloud software developer, plus of course suitable servers and data storage connectivity.
- (3) The theory community is just beginning this process and still looking for the optimal implementation.

What external requirements, if any, must be met for such a migration?

- (1) See above. Cloud developments funded in Europe by the EU.
- (2) Minimum level of agreement for developmental support from package developers (or local substitutes).
- (3) Our simulation data products are large so we need large storage systems and metadata tagging.

What is the expected scale of demand for these resources in your research field?

- (1) Within Australia, CSIRO plus two universities as part of radio astronomy national facilities that service user demand from the Australian astronomy community (many universities) and global partners in Europe, US, Japan, China and India.
- (2) CASS/ATNF observers - potentially hundreds globally, but substantially fewer concurrently.
- (3) Internationally demand is high for such simulation VOs (e.g. the German Astrophysical Virtual Observatory). Almost every observing proposal submitted in the modern age uses simulated data to justify their science goals/claims.

What special security issues, if any, might apply to such a migration in your research field?

- (1) None.
- (2) Authentication and authorisation via AAF as well as internationally accepted equivalent.
- (3) None.

RC3: Are there any other types of infrastructure (non-cloud) that would better meet your needs?

No.

National Servers

NS1: What special server requirements does your research require, including the need for 24x7 support, uptime requirements, network connectivity and redundancy?

- MWA data archive will require office hours support, 95% uptime, 10 Gbps connectivity and at least dual site redundancy.
- ASKAP would expect a non-critical level of support, such as 99% uptime and support with guaranteed QOS (e.g. first response within 12 hours). Multi-site redundancy is required.
- gSTAR will be supported by VPAC and internally.

NS2: What existing special servers do you currently maintain?

- None. MWA will have need on timescale of 18 months.
- CASS maintains servers for the Australia Telescope Online Archive (ATOA).

NS3: What are their server infrastructure requirements?

- ATOA currently requires a single server with less than 1PB storage.

NS4: What sort of criteria should be devised for admitting services for virtual server support and resource allocation?

- Projects with greatest data volume and broadest user demand should be supported first.
- Alignment with established national research priorities, community readiness, and cost-benefit.

ARCS tools and services

Responses to the ARCS questions provided by ICRAR in consultation with ARCS.

ARCS1: What functionality delivered through ARCS tools and services is important for your research?

The Long Baseline Array project started to make heavy use of the **ARCS Data Transfer Service** in March 2009. Since then, we have transferred approximately 285 TB of data. The project requires large amounts of data to be transferred from six telescopes in the eastern states to Perth for processing. Traditionally, this would be facilitated by shipping up to 180 disks holding the observation data from the telescope sites to Perth and back.

The use of the **ARCS Data Transfer Service** has significantly simplified and streamlined the process by allowing us to transfer all experimental data over the network from all telescope sites with network connectivity.

The benefits for us are:

- significantly reduced time between observation and result
- the permanent availability of disks at the stations allows for rapid response target of opportunity observations that were previously not possible.
- an increase in the effective disk pool allowing more observations (disks are no longer removed from the observation pool while being transported and processed).
- reduced cost
 - no charges for shipping of disks (approximate cost: \$11,000 per year)
 - significant savings in man hours not spent handling disks at the telescopes and in Perth (man hours saved: approximately: 300 per year)
- crucially increased data security due to removed risk of disks getting lost or damaged in transit
- increased workflow flexibility
 - the processing workflow can be completely operated remotely
 - more data is readily available for access at any given time

Other ARCS Services that we deem to be potentially useful for us in the future are the **ARCS Data Fabric** and **EVO**.

We plan to investigate the use of the **ARCS Data Fabric** as part of our data distribution system in order to simplify the sharing of our experimental results with our colleagues around Australia and overseas. In particular, developments such as QuickShare, the establishment of a project space for us on the **ARCS Data Fabric** and the possibility to integrate the **ARCS Data Fabric** into our output workflows are very promising.

Further, we plan to explore the possibilities **EVO** offers to our distributed teams. We are excited about the availability of a phone bridge as well as the option to hold meeting in plenary mode.

ARCS2: What changes or improvements to these services would you like to see in order to better deliver this functionality?

We can see two areas in which improvements to the **ARCS Data Transfer Service** could benefit our workflow.

#1 a simplified setup of data transfer endpoints

#2 an easy to use management interface individual researchers can operate without special knowledge.