Computing Infrastructure Planning Working Group Report

17th September 2018

Contents

Forward

This document is the final report of the AAL Computing Infrastructure Planning Working Group. The Working Group was commissioned by AAL in October 2015 to report on appropriate investments in astronomy computing infrastructure over the 5-year period beginning 1st July 2016. The report was finally tabled in October 2016. This version of the report has been amended in September 2018 for use in the community consultation process as AAL develops a four year Astronomy Data and Computing Investment Plan.

1 Executive Summary

Astronomers are increasingly challenged by the size, dimension and complexity of their data, the need to develop and run sophisticated data processing and analysis pipelines, and the need for compute-intensive theoretical simulations in order to compare with and interpret the observations. Over the next five years, it will become essential for astronomers to have skills and resources across a wide range of areas including:

- **Storage** – *data storage is no longer a matter of using a single hard drive, nor in-house parallel disk arrays. Rather storage will have to be housed in federated buffers and wide bandwidth Internet connections will be needed to transfer part of those data for processing.*
- **Curation** – *databases and storage media need to be properly designed with appropriate hardware and software and security. It is all too easy for data to get lost, erased or for retrieval time to be too long.*
- **Interoperability –** *To the maximum extent possible, astronomical data sets should work together seamlessly and this requires a common architecture framework. Wherever possible, databases and data access services should be Virtual Observatory compliant.*
- **Processing** – *single-threaded codes on conventional processors and consumer hard drives are inadequate to complete data reduction or run advanced simulations. Astronomers will increasingly require high data input/output rates and custom disk solutions on supercomputers with massively parallel processors to process data on reasonable timescales.*
- **Analysis** – *human inspection/classification techniques developed when data was restricted in dimension, no longer scale. Advanced analysis, statistics, machine learning and artificial intelligence are now vital tools in a scientist's armoury.*
- **Presentation** – *science increasingly demands the opportunity to reprocess and reproduce results derived from sub-sections or, at times, entire datasets. If no appropriate portal exists, this opportunity is denied, raising questions about the authenticity of published results.*

The Square Kilometre Array (SKA) and its pathfinders, gravitational wave data processing, and next-generation spectrometers will only magnify these issues in coming years and finding ways to address these issues with limited funding will be an ongoing challenge. The amount AAL has invested in eResearch hardware and software historically has been a very small fraction of the overall astronomy infrastructure budget. Therefore, our recommendations seek to increase the level of targeted AAL investment to support the community's eResearch requirements, and to maximise the impact from any investment, by augmenting and complementing existing infrastructure to leverage the greatest benefit.

To fully exploit Australian investment in Big Data-generating telescopes, it is necessary to plan, cost and adequately resource the data and computing infrastructure well in advance of the commissioning of the telescopes. It is also clear that existing computing, data and software infrastructure could be better exploited by the astronomy community if there was a dedicated team of experts to help coordinate and facilitate access to required computing infrastructure and provide specialist training and expertise.

AAL recognised these challenges and established the Computing Infrastructure Planning Working Group in late 2015 to advise AAL on investments over the next 5-years in

computing infrastructure to augment and better exploit existing eResearch facilities and services. This document presents the working group's recommendations regarding the eResearch infrastructure, services, and skills that the Australian astronomy community needs if it is to address the science priorities in the Decadal Plan 2016-2025. This document also presents rough estimated costings to give AAL an indication of the scale of the associated investment. However, the working group recommends ongoing user consultation in order to provide more detailed recommendations around the requirements and specifications for future hardware, as well as further investigation and engagement with service providers to refine the costings.

1.1 Key recommendations

Key recommendation 1: AAL should invest in the provision of Astronomy Data And Computing Services (ADACS) to provide discipline-specific training, support and expertise to allow astronomers to maximise the scientific return from data and computing infrastructure. ADACS resources would be used to aid our astronomers to generate, process, store, curate, analyse and visualise data. Australian astronomers could draw on ADACS resources and expertise to help them generate/process and interpret theoretical and observational data across the electromagnetic spectrum and from multi-messenger and gravitational wave detectors. It should engage with the astronomy community through a user committee and its strategic direction should be determined by a steering/advisory committee (e.g. AAL's Astronomy eResearch Advisory Committee) and AAL.

ADACS would provide discipline-specific interface to, and coordination with, National eResearch programs/facilities and data fabric. ADACS would enable astronomers to better exploit the existing eResearch resources (including National computing centres, NeCTAR Research Cloud, ANDS and RDS services, virtual observatory services) by helping users connect to the facilities and to the data, and through software and tool development. ADACS should also advise National supercomputing centres on the design, purchase and operation of hardware to support astronomical research.

ADACS services should be delivered by a small team growing over several years to reach approximately 12 FTE (~$2M/year for personnel and operating costs) comprising software engineers, developers, and scientists, with a range of skills from data acquisition, to parallel processing, machine learning, advanced visualisation and hardware support. Some fraction of ADACS resources should be available for allocation to research projects by a resource assignment committee based on merit. A critical mass of ADACS staff could be based in 1 or 2 key hubs, augmented by satellite staff distributed around the country at major astronomical groups and/or computing and data hubs, where appropriate. ADACS should be delivered by a host organisation or consortium, selected through a tender process.

ADACS has the potential to drive research and development in the "big data/data science" across the whole electromagnetic spectrum. Its team of experts will have capacity to develop innovative tools and data analysis techniques to enhance existing astronomy projects, and pioneer new methodologies for future large-scale projects. Many of these tools may be transferable to other research disciplines.

ADACS would create a pool of experts that have long-term careers, thereby ensuring that projects are sustainable and adequately supported over their full life-time. This pool of

experts can also provide critically needed training for our PhD students and postdocs, and be competitive in bids for SKA construction contracts, many of which will be in the data/supercomputing domain.

ADACS should also create opportunities for academia-industry engagement through incubator workshops and industry placements, and open up alternative career paths for astronomy graduates whose highly transferable computational skills are sought after in other sectors. Expanding the capabilities of Australian industry in this rapidly growing "big data/data science" sector will be of national benefit.

Key recommendation 2: We recommend that an investment of approximately $7-15M every five years for astronomy-dedicated data storage and computing resources (serving both theory and observation) would be needed - in addition to the available National eResearch infrastructure - if the astronomy community's growing eResearch infrastructure needs are to be met. This infrastructure could be designed by, or in consultation with, ADACS experts in order to maximise astronomy data throughput, analysis and system uptime. The community's unmet storage, compute and performance requirements could be met through the provision of hardware housed at existing National supercomputing facilities and/or through cloud computing services. Note: AAL and ADACS should aim to maximise the extent to which the identified infrastructure gaps can be filled with external resources/funding by working with national and international eResearch facilities and programs etc.

1.2 Investment principles

Any AAL investments into computing infrastructure, training and support should adhere to the following key principles:

1. New investments should leverage and augment, rather than duplicate, existing resources. Many existing resources are already available to Australian-based astronomers, including compute and storage at National computing facilities, virtual observatory-compliant data archives, and a wide range of training options and HPC support. ADACS should therefore: 1) fill gaps in expertise and services that are not already available to the broad astronomy community, and 2) interface with and (where appropriate) coordinate the existing resources to ensure maximum uptake and value to the astronomy community.
2. Storage and compute resources are expensive and energy-intensive and therefore should be used as efficiently as possible. The astronomy community should be required and supported to: 1) write efficient and optimised code, and 2) develop and implement appropriate data compression and management strategies, particularly for data-intensive experiments.

2 Five-year plan

2.1 Investment approach

There are many computational bottlenecks for astronomers that are a barrier to scientific discovery, including access to sufficient long- and short-term storage, adequate computational cycles to process data or perform simulations, and the skills to construct optimal codes and pipelines for data processing. Once data are generated and/or processed, their curation and presentation to the outside world requires suitable databases, media and portals.

A two-pronged approach to computing infrastructure is required to solve these problems for the astronomy community:

1. **Investing in people.** To ensure efficient and effective use of astronomy data and the available computing hardware, it is crucial that AAL invest in a team of experts to provide astronomy-specific training, support, and software development, as part of the proposed Astronomy Data and Computing Services. These experts will help the astronomy community increase: the quality and optimization of software/code; the development and adoption of common data architectures and protocols to enable interoperability; and the use of modern statistical, informatics and data mining techniques to properly interpret data. ADACS will also play an important role in supporting career paths for data scientists who can provide the interface needed between scientific use of data by astronomers and the high quality software engineering needed to produce effective code.
2. **Investing in hardware.** "Hard" infrastructure (including network, storage, HPC, and cloud computing) is currently available to the astronomy community through a combination of National eResearch resources and facilities, small-to-mid-scale institutional facilities, international programs and partnerships, and commercial services. Our investigations have identified a growing gap between the hardware available to astronomers and the community's requirements to address the science drivers in the 2016-2025 Decadal Plan. We recommend augmenting the available resources with additional storage and compute capacity that could be housed and managed at existing National computational facilities and/or provided through commercial services, as appropriate.

2.2 Personnel: Astronomy Data and Computing Services

AAL has the opportunity to invest in critical IT personnel to provide expert training and support in software development, data management, and advanced analysis, in order to ensure that the investment in hardware is put to best use. These personnel should be based at host institution(s) and have on-going positions to provide support for the Australian astronomical community.

The personnel should be built up to critical mass over several years, beginning in 2016/17. A core team growing from ~4 to ~12 FTE over 3 years should eventually cover the following areas of expertise: astronomy IT training, software engineering (CPU, GPU, MPI, Astropy, Python, C++, etc), databases, virtual observatory standards and protocols, web-programming, astro-statistics, machine learning and other analytical methods, visualisation,

and IT support. The team should provide astronomy-focused training and support that complements, leverages and builds on skills development services provided by relevant national and international eResearch groups (e.g. NCI, Pawsey, NeCTAR, ANDS, and other eResearch groups).

We understand that developing an implementation plan for ADACS is not part of the working group's terms of reference, and this task would be undertaken by AAL should it decide to invest in this concept. However, we make the following broad recommendations about governance and organisational structure, for AAL's consideration:

- An AAL-appointed steering/advisory committee (e.g. AAL's Astronomy eResearch Advisory Committee) should advise on strategic direction and high-level resource allocation across ADACS's broad areas of responsibility.
- A users committee should be established that comprises representative users of eResearch infrastructure across the areas of radio, optical, theoretical/computational, and multi-messenger astronomy. The users committee should provide information and recommendations to AAL, ADACS and other National eResearch facilities/programs, as appropriate, on eResearch operational matters and astronomy users' detailed eResearch infrastructure and skills requirements.
- A director/coordinator should oversee operations and engage with AAL and the ADACS steering committee on the strategic direction of ADACS.
- Once most of the ADACS personnel is in place, some fraction of ADACS effort should be made available to astronomers/projects via a merit-based resource allocation process to be seconded onto projects e.g., for major code optimization, for VO-integration of datasets etc.
- The fields of software engineering, programming, HPC use etc often lack diversity relative to the general population and even relative to the overall Australian astronomy population. In establishing ADACS governance, policies and implementation, AAL should be aware of these challenges and ensure that there is a strong focus on encouraging diversity and inclusiveness in any activity ADACS undertakes or supports.
- To ensure sustainability, AAL should look to leverage co-investment from the host institution(s), including underwriting short-term employment contracts, if necessary, and co-funding positions. We suggest that most of the core ADACS positions are fully-dedicated support roles, rather than split science/support positions.

We also propose a strawman timeline for ADACS establishment and growth over its 3-year ramp-up period:

2016/17

- Tender process to select ADACS host organisation(s)/consortium.
- Establish steering committee (or make use of an existing AAL committee such as the Astronomy eResearch Advisory Committee) and user committee, and recruit first tranche of core positions.
- ADACS to work with AAL to develop a detailed implementation plan.
- Develop a training plan in consultation with the astronomy community and eResearch facilities and programs.
- ADACS to work with the advisory/steering committee to establish a process for merit-based resource allocation to make services such as code optimization and software

development available to the community. The process may not be implemented until year 2 of ADACS after some critical mass has been achieved.
- ADACS to begin running training program, including running (or supporting) workshops, hackathons, and webinars to train the community on advanced data analysis, use of data portals and HPC, programming, code optimization etc. The level of effort on this activity will ramp up over the first 3 years of ADACS.
- ADACS to begin project management of the All Sky Virtual Observatory and interfacing with the Australian and International VO communities. Consult with advisory/steering committee and users to develop priorities and plan for the management and growth of ASVO.

2017/18
- Recruitment of additional staff
- Establish a coordinated HPC support helpdesk to facilitate better coordination between the users and facilities.
- Begin running a new flagship program offering 10-20M CPU hours/year.
- Begin tool/software development program to support users to develop and share their code and bring their code to the data by directly interfacing with Australian VO data archives.
- Implement merit-based ADACS resource allocation scheme.
- Develop hardware upgrade and procurement plan, including consulting with National computing facilities to understand options for supporting astronomy requirements through their future hardware upgrades.
- Begin providing advanced data analysis support services, in the areas such as astro-statistics, astro-informatics, machine learning etc.
- Increase industry engagement through industry placements, industry-academia workshops/hackathons etc
- Begin liaising with international counterparts and HPC facilities to try to boost the available resources through collaborations and partnerships.

2018/19
- Recruitment of additional staff to reach ~12 FTE
- Develop and promote astronomy-specific guidelines for data management and software development best practices, in consultation with National eResearch facilities.
- Begin providing visualization support services.

Cost: Salary + on-costs will come to $130-200K per annum/person with an average around $150K and $1.8M/year for 12 people. ADACS should also include an operations budget for travel, workshops and industry engagement ($100K/year). We recommend that AAL leverage co-investment from ADACS's host organization(s) to cover additional operational costs, and work with industry to co-fund training events, workshops and industry internships. AAL and ADACS should also pro-actively seek additional government and industry funding to allow ADACS to undertake or oversee additional eResearch projects or activities.

Investment timeline: 2016/17 $700K, 2017/18 $1.3M, 2018/19+ $1.9M/year.

2.3 Hardware

We have identified a core/minimum level of hardware that the community will need if it is to handle the increasing computing and data demands over the next five years. We have also identified a preferred level of hardware capacity that would allow Australia to be an international leader in key scientific areas. In keeping with the above principle of leveraging existing resources, **we recommend that AAL investments in hardware above the core/minimum level should be prioritized towards those that leverage significant co-investment from other organisations or industry**.

Note: for the purposes of producing the costings below, we have generally assumed that the additional storage and compute requirements will be met through the procurement of hardware to be housed and managed at National computational facilities (Pawsey, NCI and Swinburne). However, computing technology evolves very rapidly, and a growing number of cloud computing options are becoming available and viable. Therefore, the first step in the procurement process will be to engage with the various National and commercial service providers to determine the most cost effective and appropriate solution, which may involve a hybrid mix of technologies and services.

2.3.1 Storage

Disk Storage

Many projects require many passes through Petabyte (PB) scale datasets. In the next five years Australian astronomers will generate over 50 PB of data. It is desirable for at least 20% of these data to be accessible to major computational facilities at any one time. High-quality, redundant and fast (>1 GB/s) IO 1.5 PB disks cost around $700K each. Several of these could be purchased for use by astronomers at the major computational facilities (Pawsey, NCI and Swinburne) to greatly accelerate our science. Where high-volume Lustre (or equivalent) systems exist and are stable it makes more sense to either purchase additional disks and servers. However, standalone disks are also desirable because they guarantee high IO and availability. An investment of ~$4M over the next five years in this area is desirable. As disks become cheaper with time (the cost halves about every three years), purchasing capacity prior to required use is unwise. Immediate (2016/17) purchase of circa 1-1.5PBs of additional storage at Swinburne, Pawsey and NCI for the community is recommended ($2M for 5PB), with additional disk space purchased in 2019/2020 ($2M for ~10PB).

Investment timeline: 2016/17 $2M (minimum $1M), 2019/2020 $2M (minimum $1M).
Total Cost over 5 years in disk: $4M AUD for 15 PB (minimum $2M for ~5-7 PB).

Tape Libraries

Pawsey and NCI are well-served by high-capacity (>10PB) mass storage facilities. Swinburne currently has no publicly-available tape system, and the existing system is limited to 60 TB in capacity. 10PB tape libraries with media cost $1M, and 25 PB tape storage facilities cost $2M (there is a slight improvement in cost/PB with scale. Additional storage of 50 PB will require an estimated $2.9M total over five years, or a minimum of $1M for an extra 10 PB.

Investment timeline: 2016/17 $400K (minimum $0), 2017/18 $1M (minimum $0), 2019/2020 $1.5M (minimum $1M)
Total cost over five years: $2.9M for 50 PB of tape storage (minimum $1M).

Data Portals

Impact of the products of data reduction and computation can be maximized by their presentation to the community through appropriate portals. This requires disk space and servers plus appropriate IT support personnel. Unfortunately security patches and other updates often require "maintenance" to continue the presentation of data. This work should be performed by on-going teams of IT professionals, and may be supported by AAL/ADACS through the provision of expertise and/or resources, as appropriate.

2.3.2 Compute

CPU Clusters

We recommend investing in astronomy-dedicated CPU hours, to augment the resources already available through merit-based and other access schemes. The Australian astronomy community currently uses approximately 50 M CPU hours per year *in total* on National computational facilities, yet international competitors are running world-leading simulations that require 100+ M CPU hours *per simulation*. Many Australian astronomers don't even attempt >M CPU core hour simulations because of the limited available resources. Clearly, simply maintaining the current level of HPC access will see Australia fall further behind the international astronomy community, and will jeopardise our ability to interpret and fully exploit the data coming from National facilities like ASKAP, MWA, SkyMapper, and the future SKA.

AAL should secure additional astronomy-dedicated time at a minimum level of 10-20M CPU hours/year. This demand could be met through the acquisition by AAL of shares in the peak National computational facilities, sufficient to support flagship astronomy projects. Assuming a rate of approximately $0.035 per core hour, 20M CPU hours/year would cost $700K/year on a peak facility. AAL could also choose to invest directly in hardware to augment and be managed by National facilities. The major CPU clusters in Australia for astronomers are housed at the NCI and Pawsey, and the design and procurement of any future astronomy-dedicated cores should be timed with the major hardware investments at these facilities. The gravitational wave community would like access to dedicated (circa 1000) CPU cores to act as a Tier-2 LIGO data centre as soon as possible, but only if sufficient operations funds to run the facility can be secured from other sources (10 M CPU core hours/year costs approximately $700K to purchase the hardware).

Investment timeline: Gravitational wave cluster in 2016/17: $700K (minimum $0K if funding is not available from other sources to run the facility) and renewal in 2020/21: $700K (minimum $0K). Astronomy clusters/time at NCI and/or Pawsey: $700KM per year ($350K/year minimum).
Total cost of over five years: $4.9M ($1.75M minimum)

GPU (or other specialized) Clusters

Highly-parallel codes that don't require double-precision operations are most efficiently performed by GPU clusters. The development time required to make efficient use of GPUs is long, so GPUs are only efficient when a problem exceeds some critical dimension. Currently reasonably large GPU clusters exist at Swinburne, CSIRO, Pawsey and NCI. GPUs are more efficient in terms of watts/flop than traditional CPUs. GPUs are only available from a few vendors, and it makes most sense to purchase GPU clusters shortly after the release dates of major upgrades to the technology, that typically double the performance of the cards.

Swinburne is currently planning to expand its GPU cluster in 2017 timed with the release of the next generation of Nvidia gaming cards. It may be most appropriate for double-precision "workstation-class" GPUs to be housed at NCI or Pawsey, and that Swinburne uses gaming cards. However, before any procurement decisions are made, AAL and ADACS should work with the facilities to ensure that the cost/benefit of other platforms (e.g. Intel XEON Phi) have been fully explored.

Investment timeline: 128 GPU servers with 4 x gaming cards at Swinburne in 2017/18 ($1M AAL contribution, with Swinburne to co-invest). 32x4 GPU servers with double precision cards for Pawsey and NCI in 2018/19: $1.3M ($700K minimum).
Total cost of over five years: $2.3M ($1.7M minimum)

High-performance workstations and/or cloud computing

Although HPC clusters support the largest-scale processing needs, many astronomers still require smaller-scale compute facilities beyond the capacity of their departments. It makes sense to have a large number of mid-scale compute facilities and/or cloud computing access for "interactive data reduction" and day-day processing that don't fit well into the national facility processing models (that are most efficient by demanding batch queue operations). Internet links through AARNET are now good enough to make the use of workstations at the national computing centres practical. We recommend the immediate purchase of hardware (or the equivalent capacity through cloud services) to deliver performance equivalent to 16 servers, each with at least 128 GB RAM, 1 GPU, 20 CPU cores and 4x2TB SSD drives to facilitate efficient data reduction and interactive data reduction. These should be replenished every four years and housed at data centres.

Investment timeline: $224K in 2016/17, $224K in 2020/2021 ($0 minimum, if sufficient resources can be secured through other funds or facilities)

2.4 Five-year Investment Timeline

The investment recommendations from Section 2.2 and Section 2.3 are summarised in Table 1. The total recommended investment by area is divided into base and recommended further investments.

Year	Base Investment ($M)			Recommended Further Investment ($M)		TOTAL ($M)
	ADACS	Storage	Compute	Storage	Compute	
2016/17	0.70	1.00	0.35	1.4	1.27	4.72
2017/18	1.30	0.00	1.35	1.0	0.35	4.00
2018/19	1.90	0.00	1.05	0.0	0.95	3.90
2019/20	1.90	1.00	0.35	0.5	0.35	4.10
2020/21	1.90	1.00	0.35	1.0	1.27	5.52
TOTAL	**7.70**	**3.00**	**3.45**	**3.9**	**4.19**	**22.24**

Table 1 Data and computing investment schedule

3 Background

Modern astronomy datasets can be characterised as being abundantly rich in information yet challenging in size and complexity. Such data enable science that is simply not possible with small datasets, such as data stacking and data mining from telescope data. In addition, they provide the opportunity to interpret the data and gain new insight into physics and cosmic evolution with sophisticated and computationally-demanding theoretical simulations.

How modern datasets (both theoretical and observational) will be exploited is a question that will shape the next decade of astronomy investment. It is usually true that the more costly a dataset was to obtain the richer it is and the more science can be derived from it. As the richness of modern astronomy data continues to increase, the way to maximise its value is to make the data as widely available as possible, and to provide services and tools that supports its use.

Allowing broad access to astronomy data yields many benefits for individuals, institutions and national/international science teams. Increasingly science is being carried out by large teams with participation from many countries. To facilitate such collaborations, data need to be publicly and easily available. For major facilities such as ALMA and the SKA this will be largely facilitated through the construction of Science Data Centres. Such Centres may also act as a hub to provide the data services needed for smaller and/or less well-resourced facilities.

There are significant challenges when building effective data support facilities. Such facilities require robust long-term data storage together with a high level of data integrity and validation, fast access to high performance computing, user-friendly interfaces. Resources are also needed for data scientists who can provide expert support across HPC, data processing, data mining, astro-statistics etc.

The immensity of big data demands new approaches and techniques to store, analyse, interpret and explore data. In order to effectively harness large volumes and complex datasets, today's astronomer needs to be proficient in scientific computing and programming languages, be able to exploit the plethora tools available for handling large datasets and running sophisticated simulations (e.g. HPC, cloud computing, VO services and data portals), and have good understanding of the data-mining methodologies appropriate for the next-generation astronomy surveys.

3.1 Working Group Terms of Reference

AAL has recognised these challenges and established the Computing Infrastructure Planning Working Group in October 2015 to advise AAL on appropriate investments in computing infrastructure over the 5-year period beginning 1/7/2016, in order to achieve the science goals in the Australian Astronomy Decadal Plan 2016-2025. Taking into account the broader context of "eResearch" funding and facilities in Australia, the working group was asked to:

- Assess the Australian astronomy community's computing and software infrastructure requirements over the next 5 years, including:

  - Access to astronomy datasets. The working group should define the: types of datasets that should be supported, modes of access and storage, options for federation of datasets, and software requirements;
  - Access to computational infrastructure. The working group should define the technical specifications, and minimum level of access, for computational infrastructure that would meet the community's needs for the next 5 years;
- Identify expected gaps in astronomy's computing and software infrastructure over the next 5 years;
- Recommend investments in computing and software infrastructure for astronomy over the next 5 years, taking into account expected levels of funding through AAL and the broader context of eResearch funding and facilities in Australia;
- Recommend strategies and investments to provide support and training to the astronomy community to maximise the effective use of computing resources;
- Identify and report on international examples of successful implementations of collaborative computing and software infrastructure for astronomers, and/or successful local/international examples in other disciplines.

3.2 Working Group Members

	Name	Institution
1	Matthew Bailes (**WG Chair**)	Swinburne
2	Simon O'Toole	AAO
3	Jessica Chapman	CSIRO
4	Allan Williams (delegate Lindsay Botten)	NCI
5	Neil Stringfellow (delegate Jenni Harrison)	Pawsey
6	Jarrod Hurley	Swinburne
7	Alex Heger	Monash
8	Orsola De Marco	Macquarie
9	Andreas Wicenec	UWA
10	Arna Karick	Swinburne
11	Greg Poole	UoM
12	Chris Power	UWA

3.3 Consultation process

A draft report from the working group dated 26th April 2016 was provided to the following groups for feedback:

- AAL's optical telescopes advisory committee (OTAC)
- AAL's radio telescopes advisory committee (RTAC)
- AAL's multi-messenger astronomy working group (MMAWG)
- The Australian National Institute for Theoretical Astrophysics (ANITA) Chapter of the Astronomical Society of Australia (ASA).

The key feedback from these groups is summarised in Appendix C. In addition, AAL undertook a community-wide user survey to which 120 individuals responded. The survey results (minus free-text responses) are included as Appendix D, with the key results summarised in Appendix C.

The Computing Infrastructure Planning Working Group has endeavoured to reflect most of the feedback from the user survey, AAL's committees/working groups, and ANITA in this version of its report. However, the working group notes that different sub-fields within astronomy have different priorities for data and computing infrastructure, services, and skills development. For instance, additional CPU cycles is the top priority for theoretical and computational astronomers, as represented by ANITA. AAL's optical telescopes advisory committee prioritised training/skills development as well as sharing data through virtual observatory portals. AAL's radio telescopes advisory committee prioritised expert support to develop pipelines, algorithms, and other software to deal with Big Data.

Despite different stated priorities from these groups, it is worth noting that the user survey revealed some commonalities in the expected future limitations across sub-fields of astronomy (see Appendix C – Table 3). For instance, lack of long-term data storage, insufficient network speed or bandwidth, difficulty sharing data with other researchers, and lack of training/expertise in advanced statistical and informatics techniques, were expected to be key limitations across all sub-fields.

The working group recommends that AAL invest in a balanced and cost-effective way to ensure that the highest priority areas for each main sub-field are addressed over the next 3 years.

3.4 Structure of this report

The working group has identified the community's computing and software infrastructure challenges, requirements, gaps and solutions under the following broad and inter-related themes:

- Storage, curation and interoperability
- Processing, analysis, and presentation
- Training and support
- International examples of data infrastructure models

Detailed reports and recommendations from these four themes are in the remaining sections 4-7, with supplementary materials in the Appendices.

4 Storage, curation and interoperability

4.1 Storage demands for Australian Astronomy Observational Facilities

4.1.1 Data Storage Requirements

Radio and optical astronomy facilities generate a high volume of data. These data are stored and made available for science use to the global community of astronomers. Considerable operational resources are needed for the construction and ongoing operations of data archives and to facilitate the best science use of their data. (See Appendix A – Table 3 and Table 4 for a summary of astronomy data archives in Australia[1].)

For the five years to June 2021 data storage requirements will be strongly dominated by radio astronomy, in particular for the Murchison Widefield Array (MWA), Australian SKA Pathfinder (ASKAP), and Parkes telescopes. In addition, by June 2021 some early commissioning data for SKA-Low may be archived.

As at June 2016, the total volume of astronomy data stored in managed Australian archives will be approximately 13 Petabytes (PB). This figure corresponds to the data volume for a single or 'primary' copy of the data. Including data backups, the total volume doubles to ~26 PB.

For the facilities listed in Table 3 we estimate that there is a 'gap' in the provision of data storage of approximately 2 x 15 PB (corresponding to the primary data sets + plus one backup copy) that will be needed to handle the archives until June 2021. This is in addition to storage that is already available or planned.

As discussed in section 4.4 – it may be possible to reduce data storage requirements if data compression and/or other means of reducing incoming data volumes can be effectively used for some types of data products and/or unprocessed data.

4.1.2 Post-processing archive requirements

Table 3 does *not* include the storage requirements for science data products that are produced by science teams using data extracted from Australian or international archives. As a scenario to illustrate this, a science team might wish to extract 1000 astronomy images from an archive and download these to another location. The team then uses advanced image stacking techniques to produce a final set of 10 images that have extremely low noise levels. These 10 images form the substance of a research paper and the image data are published in an archive together with digital object identifiers. Providing access to final data sets in the type of way is now becoming de rigeur.

It is difficult to estimate the total storage for such 'post processing' across many different science areas but this may add perhaps 5% to archival storage requirements. This post-

[1] The information in Table 3 and Table 4 has been compiled in consultation with T Ambaum, T Booler, J Dempsey, N Hurley-Walker, O De Marco, A Jameson, A Kosmynin, C Onken, S O'Toole, C Trott, R Wayth, A Wicenec and M Whiting.

processing could add several PB to the long-term archival storage requirements over the five years.

4.1.3 Data backups and mirroring

For data collections of national significance, it is standard to archive at least one backup copy that is held at a different location to the primary data set. This is a high level requirement for the SKA. For major international facilities such as ALMA and LOFAR, the primary data archives are mirrored to several locations to provide additional data security and easier global access. Currently for ASKAP and the MWA two data copies are stored using robotic tape systems. This introduces a significant risk as a major failure (fire/flood etc) in the Pawsey Supercomputing Centre could result in near-total loss of ASKAP and MWA data. Future planning should take into account the need for offsite backups where this is deemed to be an essential requirement.

4.1.4 Data types and VO services

Column five of Table 3 indicates the types of data products stored in the archives. Additional information on the use of Virtual Observatory (VO) services is included in Table 4.

There is an increasing trend for archives to provide science-ready data products where raw data from the telescopes are processed using automated pipelines to produce data products such as images and catalogues. Science data processing may be carried out either in quasi real time (as for LSST, ASKAP and SKA), or over a longer timescale (as for MWA).

There is a high degree of commonality between data products produced for optical and radio surveys; for most (though not all) radio and optical surveys, processed data products include images/image cubes and source catalogues as key data products.

The use of Virtual Observatory standards is well adopted in the optical community with increasing use and development to provide VO services in the radio community. For further discussion on the use of VO see section 4.3.

A modified concept of Virtual Observatory has been adopted by the theoretical community working on extragalactic astronomy, where large simulations conducted in large collaborations are subsequently mined for years by many groups. The Theoretical Virtual Observatory (TAO, part of the All-Sky Virtual Observatory). We recommend that additional theoretical database work alongside TAO.

4.1.5 Operational costs

In Table 4 (Appendix A), columns six to eight indicate whether the level of operational funding available at present is likely to be sufficient for 2016 to 2021.

Once an archive has been constructed and data are available to science users, significant resources are needed for the ongoing operational costs of data archives and their associated data management systems. Resources are needed in particular for:

- Infrastructure: Ongoing maintenance and upgrades of the physical infrastructure – tapes, disks, networks, supercomputers etc.
- Cloud computing: Government sponsored and/or commercial cloud services will feature more prominently in the portfolio of the computational infrastructures used by

astronomers, who will need training and/or resources to effectively exploit these services.
- Software systems: Ongoing maintenance and upgrades for the software systems that interface between the infrastructure and provide the data access interfaces to users.
- System and User Support: Database and system administration and monitoring, communications, documentation, support for users who wish to access and download data and use this for their subsequent scientific work.

For many telescope facilities, the ongoing operational costs are estimated as 5 – 10% of the construction costs. This ballpark value is likely to also apply to large-scale data facilities.

Providing resources for the long-term operations of data archives can be highly challenging, especially for smaller institutions who may need to support a number of data archives. It is unfortunate that whilst it may be possible to find funding to cover construction costs, longer term operation costs are rarely considered adequately in funding proposals. Ongoing support can rely heavily on key individuals causing single-point failures. There is a significant risk that data management services set up by a small number of individuals will become unusable once they move on.

The approach taken by the ASVO and other groups to provide federated systems that can support many different archives, together with the expert support to set up archives and provide ongoing user support could greatly improve this situation.

4.1.6 Recommendations

1) We identify a storage 'gap' (including one back-up copy) of approximately 2 x 15 PB for 2016-2021.

2) ADACS should establish national guidelines for astronomy data management services. Such guidelines might include for example:
   - Use of data compression (and perhaps also other means of reducing data rates)
   - Use of VO standards
   - Data access restrictions
   - Data backups and mirroring
   - Duration of archival storage
   - Storage options and associated costs

3) Additional resources will be needed to support the construction and ongoing operational costs of small-medium archives, such as the All Sky Virtual Observatory and the gSTAR Data Portal that host both theoretical and observational data. AAL should consider funding such activities where funding gaps are evident.

4) Wherever possible, knowledge and software systems should be shared between the optical and radio communities.

4.2 Storage demands for simulations

Large simulations take a considerable amount of CPU time and their outputs should be made openly accessible to get the most science out of the data. Smaller simulation datasets that may not have large user bases should also be stored publicly, too, for similar reasons and because competing groups will be able to check each other's results and this will

encourage best practice. Open access simulations will guarantee the integrity and repeatability of published results (see new PASA policy). The storage requirements for simulations tend to be much smaller than for large observational surveys, hence storing and backing them up presents a relatively small incremental cost.

The Millennium and Illustris simulations, which take up ~100TB and 265TB of data respectively, are highly successful, high-impact international enterprises. The idea of storing and sharing other simulation data is best taken at the national or community level. The Australian Astronomical community, under the AAL umbrella, has now an opportunity to lead the way into the storage and publishing (intended as "making public") of simulation data.

4.2.1 Recommendations

- AAL should fund infrastructure to store and serve simulation databases calculated on NCI, Swinburne or other computational infrastructure. For instance, the Theoretical Astrophysical Observatory (TAO) of the All Sky Virtual Observatory (ASVO) and the gSTAR Data Platform (under development) could both be expanded to host a broad variety of theory data types.
- Funding should be devoted primarily to personnel, who can facilitate the transfer of data into suitable storage and help the researchers to document it and link their datasets to publications and other locations. Swinburne has plans to cater for some of these needs through the gSTAR Data Portal; therefore any additional AAL plans should be integrated with those.

4.3 Virtual Observatory Services

The value of astronomy data sets is greatly increased through their connection, or federation. For example, both observations and simulations can play a complementary role in facilitating the science objectives of each individual community, by providing theorists with the latest observations with which they can refine their models, and by providing observers with the latest simulations with which they can interpret their results. Similarly, observational datasets stored in individual data centres taken across many wavelengths (say by the optical, infrared and radio communities) provides a multi-dimensional picture of the universe that cannot be obtained from any single data set alone.

The Virtual Observatory is, in effect, a concept that *astronomical datasets should work together seamlessly* (www.ivoa.net). The International Virtual Observatory Alliance was established in 2002 to work towards this vision by establishing technical data standards that facilitate access and sharing of data across a wide range of facilities. It acts as an *alliance of worldwide VO projects that develops the required standards and coordinates global aspects of the infrastructure* (Mark Allen, CDS). The VO standards provide a means of combining and/or comparing astronomical data, for example between different observational facilities or between simulations and observations. Many data services such as Aladin or TOPCAT make extensive use of the VO protocols whilst the SKA is expected to be fully VO compliant. Using VO protocols together with interfaces provides very powerful tools for working with image and catalogue data.

In Australia, the last few years has seen significant progress towards implementing VO protocols and building national expertise in this area. In particular, VO implementations are

a major component of the CSIRO ASKAP Science Data Archive (CASDA) that has been successfully deployed at Pawsey with V1 fully operational with V2 in development, as well as the AAL-funded All Sky Virtual Observatory (ASVO) that comprises the following "Nodes" in various stages of maturity:

- ASVO-SkyMapper (V2 operational, V3 in development at ANU/NCI) provides VO-compliant access services to optical imaging and catalogue metadata from the SkyMapper Southern Sky Survey. Test data is currently available with the first fully calibrated data release scheduled for 2016B.
- ASVO-Theoretical Astrophysical Observatory (TAO) (V2 operational, V3 in development at Swinburne) provides access to N-body cosmological simulations and semi-analytic galaxy formation models, which can be further processed via science modules to produce mock light cones, spectral energy distributions, images and predicted observations from major telescopes, instruments and surveys. V3 will expand TAO to include hydrodynamic simulations and is an important component of the CoE bid for CAASTRO 3D that is currently under assessment by the ARC. The IVOA standards for theory data are much less well defined than for optical data, and therefore only key tabular TAO datasets are available through the VO. While a very powerful tool, TAO's science applications are relatively narrow, and there has been demand from the community to broaden the scope of TAO to include other types of model and simulation data.
- ASVO-AAT (V1 in development at AAO) is being developed to provide highly flexible VO-compliant infrastructure to cater for past and future AAT datasets, which will include complex multi-dimensional data from modern instruments that can collect spatially resolved spectral data.
- ASVO-MWA (V1 concept design completed) would be the first radio interferometry dataset supported within ASVO.

Whist good progress is being made both in Australia and internationally, there have been some issues with the adoption of VO standards. The implementation of the protocols can be quite challenging for developers, whilst some improvements are needed in the underlying data models. The VO standards available to date are better matched to optical astronomy than to radio astronomy and some additional standards may be needed, for example to support the full range of astronomy data types.

Ongoing international effort through the IVOA will be needed to further develop VO standards and to make these usable by developers with user friendly services for astronomers. It is important for Australia has adequate representation on the IVOA.

4.3.1 Recommendations

- Wherever feasible, data archives should support (but not necessarily be limited to) user access through VO services (such as TOPCAT and Aladin) using VO standards.
- To avoid duplicating effort and to share knowledge, new infrastructure should build, as far as possible on existing infrastructure. In particular, software that has already been developed for the implementation of VO protocols should be openly shared and reused.
- Resources should be made available for building user expertise with using VO through demonstrations, training sessions and documentation.

- Stronger Australian engagement in the IVOA with consultation to establish community needs would be of benefit.

4.4 Compression and appropriate bit-depth of data

The allocation of 8 bits per byte is a historical accident, as is the 32-bit floating point and 64-bit double precision depth of data. All too often astronomers default to 8 or 32 bits to store their data, and sometimes this is quite unnecessary. Similarly many astronomers like the convenience of storing their data in ascii files, even though this takes up more room than binary data.

There are two good reasons to compress data. Firstly, it takes up less disk space, and secondly the cost of transporting it (both in time and money) are reduced. Here transporting it can mean from a tape backup to a hard drive, or from one site to another, or from a disk to a computer's memory for processing.

In essence there are two types of data compression, lossless and lossy. Lossless compression can be perfectly undone with no information loss. In computer graphics, *tiff* files are an example of lossless compression - they remove redundant and pointless information, *gif* files use signal processing "tricks" to remove subtle details the eye can't really detect but that astronomers might care about.

In many experiments the cost of exploring whether data compression can aid in reducing storage and transportation costs is hardly worth it. At any time in Australian astronomy data storage is usually dominated by fields where information generation is cheap but pointless without vast quantities of data, and one of the experiment's main costs is actually the data storage.

The dominant users of storage at the present time are the radio projects being conducted at the Parkes and Australian SKA sites. The High Time Resolution Universe survey for pulsars and fast radio bursts generates 52 MB/s and was responsible for over 1PB of disk space. Its follow-up survey (SUPERB) generates data at the same rate (~5TB/day). Two copies of the data are typically recorded but the experiment only uses 2 bits per sample resulting in the loss of all polarisation information and a reduction from the original 8-bit data. Ideally the experiment would record four polarisations and 8-bit data at a rate of 80 TB/day.

The Murchison Widefield Array (MWA) uses 128 tiles that each record two orthogonal linear polarisations. From these tiles 128x127/2 = 8128 baselines are recorded for each of four stokes parameters in thousands of frequency channels over 30 MHz. To date ~10 PB of data has been recorded. The MWA currently uses 32 bits to store its visibilities, but it is unlikely that this is necessary. They have been conducting experiments to see the results of reducing the bit depth significantly. The gain of an individual tile is < 0.01K/Jy, and the receiver and sky temperature so hot that the SNR in a ten second integration is usually less than unity. It can be shown that in such a regime, 4-bit data is usually more than adequate to produce near optimal science.

When ASKAP begins routine operations in 2017 with at least 30 PAFs, the data quantities in spectral line mode will be very significant - maybe 5-10 PB/year. There are very good reasons to compress these data. Mass storage is currently very expensive. A server-class 8 TB disk is around $1K/disk and to house it in a data centre with appropriate redundancy,

servers and chassis often close to 250-300K/PB of spinning disk. It is not hard to imagine requiring 10 PB/disk/year to have such data online. Within 5 years the data, and data products could well amount to 100 PB.

If the data is on a tape storage system and can only be retrieved at 250 MB/s, then to "replay" a PB dataset at 100% efficiency takes 46 days! If appropriate compression can reduce the amount of data stored by a factor of ~4, then this reduces to about 4-5 days.

4.4.1 Recommendations

- All experiments seeking to use large quantities of "national facility" storage should exhaustively test data compression and bit-depth reduction strategies prior to the majority of their data entering the facility.
- Forecasting of the likely re-use of data and its likely impact on the storage and computational facilities that house or process it will be vital to avoid "bottlenecks". If experiments are being devised that are likely to be reprocessed many times, there will need to be proper planning for the IO (input/output) infrastructure to cope.

5 Processing, analysis, and presentation

Here we look at the data processing needs of the astronomical community, considering existing usage and future growth in computational demand from simulations and telescopes. The focus is on compute capacity (creation and processing of data, as well as interactive data analysis) and associated software development/optimization requirements. There is an assumption that the associated infrastructure to transport data to the compute hardware is in place.

5.1 Existing usage of the National supercomputing facilities

As described below, Australian astronomers' current level of access equates to ~10-15% of a top100 supercomputer like Raijin/Magnus. The Decadal Plan 2016-2025 recommends that Australian astronomers have access to the equivalent of 30% of a top100 supercomputer.

Peak facilities (NCI & Pawsey): The National Supercomputing facilities at Pawsey and NCI include, Magnus and Raijin, respectively, which are both currently in the top100 worldwide. Raijin is a ~1.2 petaflop machine that has approximately 500M CPU hours available per annum, while Magnus is a ~1 petaflop machine with approximately 300M CPU hours available annum[2].

Through various merit allocation and partner share schemes, Australian astronomers access 35M-45M CPU hours/year at NCI and Pawsey, and an additional ~10M CPU hours per year on the Swinburne supercomputer (described below). The NCI and Pawsey National merit-based schemes (NCMAS) are typically oversubscribed by a factor of 2-3. For comparison, to reproduce the current international benchmark for cosmological simulation, Illustris, would require 19M CPU hours, i.e. approximately one third of Australia's annual astronomy HPC usage.

Specialised mid-scale facility usage (Swinburne g2/gSTAR): The Swinburne facility is a hybrid CPU/GPU machine. It aims to provide astronomers nationally with dedicated HPC resources regardless of whether they use GPUs or CPUs for their computation and regardless of their field/type of research (e.g. usage includes telescope data processing as well as simulations). It can be classed as a mid-level facility, with ~17M CPU-hours and ~2M GPU hours available annually, linked to a 3 PB Lustre system. The resources are accessible via a job queue and through a merit-based Time Allocation Committee (TAC) process. In 2014/15, astronomers used ~10M CPU hours on the Swinburne system, of which ~2M hours were through the merit-based TAC and ~8M hours through the queue.

In 2015 the astronomy usage was split by institution as 25% (Monash), 12% (Macquarie), 3% (Melbourne), 3% (Sydney, UWA, Curtin, ANU, UQ: combined), 2% (international) and the remainder taken up by Swinburne. There are 280 astronomy account holders: 50% are PhD students, 30% are female, 45% are from Swinburne, 30% from other Australian institutions and 25% international. Across any quarter, on average, 70-80 account holders will be actively submitting jobs through the job queue and ~150 accounts will be utilized in some way.

[2] Note: Pawsey also hosts the Galaxy supercomputer, a ~250 teraflop CPU + GPU hybrid machine dedicated to ASKAP real time processing and other radioastronomy needs.

Note: Without further funding from AAL, g2/gSTAR is only funded to operate as a national facility until the end of 2016.

5.2 Usability of National supercomputing facilities

(1) Access - the current mode of access to guaranteed time, via ASTAC, NCMAS (including ANU NCMAS), partner schemes, and CAASTRO allocations, provides plenty of opportunities to apply for supercomputing time allocations, and it's possible for organised groups of researchers to receive of order 5-10 MCPU hrs in a calendar year at present. This is a very useful amount and it's possible to carry out very respectable research programmes with this. However, this time is usually obtained piecemeal and, because of the nature and size of allocations, it caps the more ambitious projects that could be done. Comparison with schemes in the USA and Europe (e.g. PRACE) suggests that we could:

- identify a fixed total allocation for astronomy projects (e.g 50-100 MCPU hrs per year), and
- restructure the time allocation process between starter, small, medium, and large projects (similar to what is done already in NCMAS), where large projects would be >10 MCPU hrs.

Detailed technical cases with proper benchmarking would be a requirement (see 3), to demonstrate that codes are properly optimised for the facilities we have access to, and clear community benefit would have to be demonstrated for the larger allocations.

(2) Efficient Handling and Storage - arguably this is the main bottleneck in how our science is being done at the moment, and it will continue to be a problem if we don't start to make changes. Partly it's an I/O problem — there will be as many approaches as there are codes used within the community (e.g. binaries, ASCII, HDF5; parallel I/O vs serial; etc) and I/O rates are increasingly become a bottleneck as I/O volumes and frequency increase — and partly it's a choice of architecture problem — you write to scratch, do your analysis, and either retain the data into deeper storage or copy it to a remote location, or delete. There are the related issues of visualisation and handling of reduced data products (see below), which will need to be done remotely as data volumes increase.

Here support (3) becomes crucial — collaboration with software and hardware experts is essential to ensure that codes are doing their I/O in the most sensible manner, tailored to the architecture and problem. This would involve a move away from e.g. binaries to e.g. HDF5 and e.g. parallel I/O, and inlining of analysis so that data volumes that need to be written can be reduced in size and frequency. A more active and intelligent form of storage system, in which a pipeline is in place to move data into tiered storage, some of which should include cloud storage, will be essential, preferably virtual environments that allow analysis software to run on data, agnostic about its underlying structure (a uniform interface as seen by the user overlaying a complex backend to interface with the data). All of this will avoid costly transfer of data.

(3) Support — efficient intelligent codes that fully exploit the systems we have currently, and will have in the future, requires expert support. Support for specialised software packages for some users is desirable, including providing documentation, sample scripts, special MPI settings for large jobs, help with core pinning, etc. Current support can be very

good, but we need a core team of software engineers and computer scientists that can work in close collaboration with the research teams.

5.3 Computational demands of next-generation radio telescopes

ASKAP and MWA currently have exclusive access to the Galaxy computer at the Pawsey centre. Galaxy has 472 nodes with a total of 9440 cores plus another 64 nodes with 512 cores and 64 Nvidia K20X GPU. The 472 CPU-only nodes are designed to meet the baseline compute requirements for ASKAP real-time processing. The Galaxy computer has been designed, scoped, and procured to meet the needs of ASKAP and is an integral part of ASKAP as an instrument. It has the capacity to meet the compute requirements of ASKAP science data processing to produce data products over the coming five years. The recent addition of a GPU based capability to Galaxy enables it to also meet the operational processing requirements of the MWA. As ASKAP is not yet at full capacity, the CPU nodes are not at full capacity, and are shared with MWA, although the usage is expected to go to close to 100% as the Early Science program begins later in 2016. At this point in time MWA is using significant amount of time on Galaxy as well to process its voltage and EOR data. In addition the GLEAM survey is also using Galaxy to a lesser degree.

Once ASKAP is fully operational it is expected that there will be a shortage in available core and GPU hours to cover both ASKAP and MWA. With the current upgrade of MWA and the potential further upgrade the situation will be very tense and alternate solutions need to be sought. The biggest current limitation for ASKAP is the amount of memory/core. On the Galaxy GPU nodes this ratio is slightly better, due to the fewer number of cores. However, the amount of GDDR5 memory on the GPUs is fairly low as well and thus requires very aggressive data movement schemes for data intensive applications, which will affect the overall efficiency of the code. On both GALAXY and MAGNUS the I/O contention on the scratch file space is very noticeable and thus a reliable prediction of execution times is not possible. In particular for ASKAP and the SKA1-LOW and their near real-time processing requirements this is not satisfactory, since it will affect the overall efficiency of the observatory. The MWA is currently processing the data off-line. In particular after the second upgrade this situation is likely to change, since the data volume will then require significant data reduction in order to be able to store the results long term.

The amount of post-processing critically depends on the amount of real-time processing. For ASKAP and the SKA1-LOW most of the data stored longer term will be processed to more advanced, or higher level data products. The MWA is storing output from the correlator directly and also output from the voltage capturing system. That means that the post-processing of MWA data first has to reach the same level as ASKAP and SKA1-LOW. After the MWA upgrades the processing required will be of the same order of magnitude as ASKAP and thus lies in the range of 10% of the SKA1. Precise numbers are almost impossible to state, since everything is highly dependent on the actual science cases. The science teams will then pick-up the data from the ASKAP and MWA archive(s) and in a similar way later from the SKA1-LOW archive and apply further processing. Again the amount of processing depends very much on the science case, but in general the maximum required is assumed to be of the same order of magnitude as the real-time processing. Partially this is also due to the fact that very often this kind of processing is more experimental in nature, has to be re-done several times, or using a number of different algorithms.

The data products for ASKAP are released as soon as they have been validated with the data validation processes carried out by the science teams. ASKAP has no proprietary period but science teams have access to the data products prior to validation. The five-year span of this report will include a large part of ASKAP Survey Science operations, where the processing is expected to be largely steady-state, with well-defined requirements. Unknowns from ASKAP's point of view over this period include the impact from MWA operations on the I/O performance of the scratch filesystem (the ASKAP correlator writes directly to this filesystem, and so it is in effect part of the instrument), and any replacement platform that may be acquired by Pawsey once Galaxy comes out of warranty in 2017. The ASKAP processing performed on Galaxy produces science-ready data products that are made available through the archive. Post-processing requirements that ASKAP science teams may have are still somewhat unclear. For the ASKAP Early Science program, an allocation of Magnus time has been awarded, and this will be used, in part, to refine the post-processing requirements.

For MWA the products produced by the real-time processing will be public after some limited proprietary period, which typically is 12-18 months. The products produced by science teams in general will be published along with the papers, which typically require another 12-18 months, or longer if the processing is very complex. Very often in survey projects data is released in multiple stages, in yearly or more frequent releases. Within the MWA proprietary periods, data will have to be shared with the whole survey team, which usually is globally distributed. Data sharing includes also sharing the access to processing capabilities, since moving or copying large amounts of data is very time consuming and in-efficient. Data volumes of science projects can span several orders of magnitude, with the biggest ones being about a factor of 10 less than the real-time processing products. As an example, the EOR experiment will produce the largest amount of data within the SKA for a single science project. The final result can essentially be represented as a single two-dimensional plot. However, the intermediate data produced on the way to this plot might be extremely valuable to all kinds of other science projects. Storage of those products does not need to be inside the current centres, but that for sure is one viable option. In any case the critical part is not the storage, but the globally enabled access, which requires dedicated highly available services. While the IVOA provides a number of protocols to deal with a whole range of standard products, very often new science projects, by the nature of science, produce novel and often quite complex products, requiring special services. Astronomical data has a very long life-span, photographic plates and observations, which are already centuries old, can be very valuable. With the current exponential growth of data volumes, we will need to be both critical and very careful about which data to retire and what to keep. Very often the lower level data products will be prone to become unreadable if not actively maintained. Maintenance includes both the data, the storage media and devices as well as the reading software.

Future: The SKA1-LOW will require a machine with processing power of around 300 PFLOPs. Given the discussion above, the post-processing in the worst case will require up to the same amount, more realistically probably between 1-10% in Australia. Part of this demand can be provided by facilities of various flavours and ownerships, including the large-scale facilities, the smaller scale institutional facilities as well as public and private clouds. In addition we will need very capable networks in order to share the load and also the data between centres in Australia and across the globe. In order to facilitate this we should try to

secure multiple 1 Terabit per second links by the time the SKA is operational. The data volumes produced by the SKA will be of the order of several 100 PB/year, with the potential to go up to 700 PB/year if we would store the so-called discovery space of the telescope. This would maximise the potential 'archival' and also serendipity science, but obviously also maximises costs. A better approach might be to invest more into trying to extract as much information as possible and use dedicated progressive lossy compression techniques to reduce the total amount of data kept.

For both the processing and the intelligent archiving and data maintenance we will need highly skilled people, data scientists, astronomers, computer scientists, software engineers and HPC and storage system specialists. We will also need an established culture of working collaboratively together, across these fields, in an efficient and solution oriented way.

5.4 Computational demands of next generation simulations

As documented in the 2016-2025 Decadal Plan, theoretical and computational astrophysics has grown from ~12% to ~32% of the astronomy community over the past ten years, partly fuelled by a growth in computational facilities at the institutional and national levels. Continued investment in such facilities will be necessary to enable the simulation sector to grow further into the future. Looking ahead, we need to ensure that researchers have access to Exa-scale facilities as they come online, so as to enable world-leading grand-challenge simulations (e.g. high-resolution cosmological simulations, 3-D stellar evolution and Supernova models, magneto-hydrodynamical radiative transfer simulations for reionisation modelling) to be led from Australia.

In developing this report, the Working Group canvassed major HPC users[3] within the Australian astronomy community to understand their computing requirements over the next five years and to identify likely barriers to achieving the science drivers in the astronomy Decadal Plan. The key outcomes from this consultation are:

- The preferred level of HPC access to do world-leading simulations varies across the sub-domains. At the upper end, Alex Heger's 3D supernova simulations at preferred level of resolution would require 1000 M CPU hours in 2016 (a factor of 100 higher than his actual usage of 10M CPU hours) growing to 1 T CPU hours in 2021. For most other large HPC users the current level is around 1M-10M CPU hours in 2016 with the preferred level typically 1-5 times higher, and forecast growth over the next five years to be in line with expected science and technology advances (i.e growing by a factor of 5-10).
- Some of the largest HPC users rely heavily on international HPC access because the National facilities cannot meet their requirements.
- The ability to store, back-up and curate the generated data products was identified as a key limitation by many HPC users consulted, with the severity of the problem expected to become more serious over the next five years. Most users consulted are each generating about 10-100 TBs of data in 2016, expected to rise to 100-1000+ TBs in 2021.
- Most HPC users share most of their simulation data with collaborators, but not publicly.

[3] The following investigators were asked in April 2016 to provide information on current and future usage and unmet demands on behalf of their research groups: Martin Asplund (ANU, Christoph Fedderath (ANU), Alex Heger (Monash), Chris Power (UWA), Chris Blake (Swinburne), Orsola De Marco (MQ), Luz Penaloza, Pascal Elahi (Uni of Sydney), David Parkinson (UQ), John Webb (UNSW). Together, these research teams represent more than half of the astronomy HPC usage within Australia.

- In addition to CPU and storage capacity, other limitations to compute-intensive research are: remote visualisation, RAM, I/O and network speeds, and HPC architecture causing bottlenecks.

USE CASE: **Next-generation Australian telescopes and simulations unlocking the secrets of the early Universe**

A major focus of astronomy over the coming decade will be on understanding how the Epoch of Reionization (EoR) has influenced the formation and evolution of galaxies. National facilities such as ASKAP, MWA, and SkyMapper, and ultimately the SKA, will allow us to connect the conditions under which galaxies formed during the EoR, when the Universe was barely a billion years old, through their evolution over the subsequent 90% of cosmic time to the properties of galaxies that we observe in the nearby Universe.

Cosmological simulations provide a powerful tool for astronomers to interpret these observations, and a goal of Australian computational astrophysicists is to create a suite of state-of-the-art simulations of the Milky Way and the nearby Universe to study how the properties of galaxies, ranging from our own Galaxy and M31 to their satellite galaxies to the Virgo galaxy cluster, are shaped by conditions in the early Universe, and what these local galaxies can tell us about the EoR. These state-of-the-art simulations will incorporate cutting edge physical prescriptions to model star formation, the growth of black holes, the build-up of the chemical elements, and the propagation of radiation, requiring a resolution capable of tracking individual star forming clouds and the accretion of gas onto black holes in the innermost parts of galaxies.

Extrapolating from the current generation of simulations, we expect that individual simulations will require 5-10 million CPU hrs, generating a few hundred terabytes of data per simulation, and we will need to run multiple realisations over a 2-3 year baseline. The ability to run such simulations would capitalise on Australia's leading expertise in this field and is vital to interpreting the data from next-generation Australian astronomical facilities and Australian-led science teams. When compared to simulations such as the Millennium Simulation (Springel et al. 2005), which has been widely used by the astronomical community and has produced nearly 500 papers with tens of thousands of citations, we expect this simulation to provide a similarly exploited dataset, drawing on, for example, the next generation of the Theoretical Astrophysics Observatory for delivery.

Priority areas for future computing infrastructure are:

- Access to sufficient CPU-time to enable massively parallel simulations of a grand challenge nature to be attempted. Often these represent risky science and may require more time than is feasible for a competitive allocation from a general pool. However, the payoff can be large in terms of impact factor for Australian science. Buying dedicated time that can be made available to 1-2 projects per year could be considered (and added on top of the pooled idea above). Ideally this would be on the next generation of Australian-based national facilities but looking overseas is also a viable option. The level of dedicated astronomy HPC time that should be ring-fenced for flagship projects needs to be a minimum of 10-20 million CPU hours in 2016 (noting that the current benchmark for world-leading simulations is 20-100 million CPU hours *per simulation* in many fields

of astronomy), growing year-by-year to keep pace with technological and scientific advances.

- Increased storage capacity for, and curation of, simulated HPC data products. This includes more short-term local storage as well as long-term archival storage, back-ups and services to support curation and sharing of data.

- Technical support and training for researchers to ensure that they are best prepared to submit competitive proposals in merit-based schemes for large allocations of CPU-time on national facilities (overlaps with software items below and Training);

- Access to next generation accelerator technology, ideally via an astronomy dedicated facility and most likely GPU-based. Many specialised data processing pipelines are now GPU-dependent in addition to simulation code which requires accelerators to be computationally viable (e.g. cosmological and star cluster N-body, microlensing maps). gSTAR has a combined speed of ~500 Tflops which will be slow in comparison to future GPU clusters.

5.5 Efficiency of software development in the Big Data era

Large astronomical projects such as the LSST and SKA will produce volumes of publicly available data that have not been seen before. This presents the community with significant challenges, raising questions such as:

- Where is the data stored?
- How do astronomers access the data?
- How do astronomers process and analyse the data?

While the first of these questions is currently being addressed through various implementations of the Virtual Observatory, questions around data access and analysis remain largely unresolved. Public data releases circumvent access issues to some degree, as data is simply stored and served to users. But what can a user do with a 50GB+ catalogue table? What tools are available to analyse such a dataset? How does one write such code, if required? If the user wants more than just tables, but images, spectra or data-cubes, how do they analyse many terabytes of such data? These are instances where downloading data and running software locally is no longer feasible.

Bring the compute to the data

One answer that is being raised more in the astronomical community is the “Bring the compute to the data” model. In this model, users log on to some high-performance computing system, either through a web front-end or at the terminal, and run some code on the data. Whether the code is self-written or provided by the data repository is unclear, but since flexibility is always preferable, some combination of these is desirable.

This model removes the need to download the data, but presents software challenges, as well as data access issues. These are challenges and issues that astronomers alone should not be required to solve.

Who writes the software?

This is the biggest question. Manipulating extremely large datasets often involves using libraries and tools that are not familiar to most astronomers.

In large part, astronomers are familiar with scripting languages such as python and/or compiled languages such as Fortran and C/C++. Accessing databases (of whatever flavour) is something that most astronomers are not familiar with; most astronomers do not want to have to write SQL queries, for example. What are required are tools that astronomers are comfortable with and/or are straightforward to learn.

To develop these tools, a combination of software developers and software-minded astronomers is required. The demands to publish and the general lack of credit they receive for writing software means that research astronomers should not be spending most of their time writing and debugging complex code.

The most efficient way to develop software for astronomy is therefore to hire trained software and database engineers who engage with astronomers to develop the tools required for Big Data management and analysis. Software engineers could either develop software in-house at data repositories or be contracted out to survey teams as required.

5.6 Optimisation of software

Highest priorities should go to tasks that have the largest impact on a wide cross-section of the community and on short timescales. In particular, this comprises basic and simple tasks for an expert, like benchmarking and identification of hotspots, performance measurements and identification of bottlenecks. It would be useful to teach people to use the requisite tools effectively. Special support that could be offered on a routine basis would include benchmarks and help with profiling at a subroutine level. Researchers who know their code and algorithms may be able to fix codes quickly themselves or with minimal help. Benchmarking support and obtaining of key performance parameters when porting code to new machines, and including for proposal development on such new machines for merit allocation schemes, would be very helpful.

Only if significant resources for code development can be found, large-scale parallelization support including parallelisation of entire codes should be provided. One could argue, however, that providing such a service would be an invaluable investment when building multi-million dollar machines. Even then, this could likely be done only for a select number of codes and should not compromise the support for a broader community. For such projects it is likely more appropriate for a researcher to find independent professional support as part of a separate grant.

Lowest priority should be given to optimisation support for specialised hardware. These may be transient and change on a rather modest timescale. It is hard to predict where the evolution will go, and hence it may also be a risky investment for the long-term and limit portability. If at all, this is a service that should be provided by a local computing centre and be included in the operating cost for machines using highly specialised hardware.

We acknowledge that it can be difficult to draw the line between specialized and general hardware. What portion of the community needs to adopt ‘specialised’ hardware before it can be considered essential/general? GPUs present such as a case, with widespread use but not across all sections of the astronomy community.

Investment in code quality vs machine dimension.

On the question of investment in code quality compared to machine dimension it should be noted that current machines are already highly oversubscribed, machine dimension should be a large priority as long as it is not at the expense of machine quality such as serial CPU performance and memory.

Porting existing codes to new architectures.

We have experience of this already with GPUs. Those early adopters who ported code to GPU usage have reaped the benefits with marked increase in performance, to the extent that these codes are now highly dependent on GPUs. Such codes include pure simulation (e.g. N-body) and data processing (Parkes, Molonglo, MWA post-processing). However, many astronomers have not found the time to invest in porting code to GPUs (even in some part), reflecting the traditional pattern of sticking with what works at the time. Moderate success has been achieved by having a software engineer available to assist but community uptake has not been overwhelming. The story may be similar if it became necessary to port to other architectures, e.g. non-x86 CPUs, although in that case the majority of HPC users will be forced to act (unless they look elsewhere). Non-astronomers are generally better placed to move to different architectures, reflecting a greater reliance on software packages that release architecture-specific versions (e.g. Gaussian).

5.7 Interactive Data Analysis and Visualisation

Here we concentrate on the use of scientific visualization to achieve scientific results rather than the more standard information-based visualization use cases which can be addressed with existing platforms. The architecture of the current HPC facilities and related services are tuned towards batch processing applications with minor attention to interactive / GUI-based / Web-based data analysis. There is a need to give similar attention to the interactive processing uses-cases in order to facilitate the intermediate steps required to build big-data analysis and visualization capability.

This can be split into three sub-areas based on data-size.

Analysis and visualization of small datasets (up to 4 GB):

The problem for these datasets is the lack of appropriate modern tools that are built to utilize the latest improvements in hardware and software infrastructure to provide astronomers with an integrated data analysis and visualization virtual workspace. To make such a service available to astronomers requires appropriate hardware resources.

Analysis and visualization of medium datasets (up to 100 GB):

In addition to the lack of software tools to address these datasets, astronomers don't have the appropriate hardware resources to analyse / visualize these datasets. Interactive virtual desktops can be a great approach to facilitate dealing with these datasets, e.g. the NeCTAR funded Monash Characterization Virtual Laboratory (https://www.massive.org.au/cvl). This will solve the problem from the hardware side and make it more accessible to astronomers. Deploying this solution should be considered as part of future HPC facilities. Allocation of resources can be done using a regular queue-like system or via committees such as ASTAC.

Analysis and visualization of large datasets (beyond the workstation/server memory limit):

Software solutions are under development for this purpose but are not yet complete or ready for community release. Fortunately, they are not yet needed by a large number of astronomers and will be limited to radio astronomy in the near future. For this problem, we don't have consistent access to HPC visualization clusters, high-resolution displays (e.g. Opti-portal or CAVE-like), or software tools and solutions. A lack of consistent access here means that all existing arrangements are very limited and not accessible to the wider community. This area is not an immediate priority given the amount of investment required and the lack of use-cases. This decision should be considered after ASKAP is fully operational in order to more readily assess the need for such tools and infrastructure.

5.8 Recommendations

- AAL should make funding available to buy or access additional HPC time for performing grand-scale simulations that align with the Decadal Plan. Applications would be awarded time by an expert committee based on merit, with 1-2 proposals funded per year. The level of dedicated astronomy HPC time that AAL should secure for flagship projects needs to be a minimum of 10-20 million CPU hours in 2016, growing year-by-year to keep pace with technology and scientific developments.
- AAL should invest in increased storage for, and curation of, simulation data generated on HPC facilities. This includes more short-term local storage as well as long-term storage for archives and back-ups and services to support curation and sharing of data.
- AAL should continue to fund national access to readily-accessible mid-level facilities.
- ADACS should employ staff to engage with national facilities on behalf of the astronomy community and represent interests in design/procurement/access decisions.
- ADACS should investigate pooling the astronomy time available through allocation schemes and restructuring the allocation process so that it is centrally managed.
- ADACS experts should provide assistance to HPC users on software development and optimisation.
- Ensure that interactive data analysis, remote visualisation, and data ingestion needs are prioritised in facility procurement plans. Investment here is modest, requiring only a small number of large memory nodes (preferably with a high-end GPU, e.g. Titan).

6 Training and support

6.1 Training

The rise of “big data” and “data science” in the technology industry and its prevalence in academic research is creating a new generation of savvy astronomy that are adopting more general, industry standard tools and practises. This has led many astronomy research groups and individuals to shift their efforts towards dealing with issues around data-intensive research, and to create new tools to address next-generation datasets. The majority of tools are developed in collaboration by early and mid-career researchers in dual software development/data science roles and research roles, and are widely used by the community.

In some cases universities have established successful Data Science Institutes that support dual astronomy research/technical roles and offer comprehensive training in data science and scientific computing. Notable institutes include the Berkeley Institute for Data Science, eScience Institute (UW), Centre for Data Science (NYU), and Centre for Data-Driven Discovery (Caltech). The majority of these were initiated by astronomers involved in large survey projects (e.g. LSST), or leaders in astroinformatics. Training in scientific computing and new methodologies is a core part of Institute activity. Promoting best practises in scientific computing, exploring new machine-learning techniques, initiating collaborative coding projects (e.g. AstroPy), and developing community resources (e.g. IPython Notebook, Jupyter, AstroML, scikit-image, GlueViz) has become increasingly important and reflects the broader cultural shift in the way data-intensive research is conducted. It also reflects a growing trend for astronomers to seek alternative careers outside of astronomy.

Programs that aim to increase uptake of tools and promote best practise range from one-off workshops (e.g. SciCoder, Software Carpentry, AstroPy, SciPy), to community connected and regularly scheduled tutorials and meet-ups (e.g. Hacker Within). The result is that researchers are no longer limited to specific astronomy data analysis tools and traditional IT infrastructure. Satellite conferences and hackathons (e.g. SciPy, DotAstronomy, Astro Hack Week), connect researchers, foster collaboration between research institutes, and provide a forum for engaging with industry.

The research/computing training needs for Australian astronomers can be divided into a number of categories: training critical for specific Australian-led survey projects, training specific to big data and high performance computing, and more general scientific computing/tools training that provides researchers at all levels with new skills, enabling them to be more effective researchers and/or to prepare them for alternative career paths within astronomy or the tech industry.

The ANITA Chapter of the ASA is a well-established community of theoretical astronomers, and among other things provides scientific computing training specifically targeted to theoretical astronomers (particularly those at the early– and mid– career level). ANITA hosts an annual summer school and runs a series of somewhat *ad hoc* online workshops. Past workshop have addressed specific issues around big data and data mining and include *databases and SQL – applied, computational Bayesian theory, R Statistics for astronomers.* In recent years *astroinformatics* has also fallen under their domain. As a solid training platform ANITA has the potential to expand its suite of workshops and lectures to include more rigorous *scientific computing* and *data science methodologies.*

Swinburne's gSTAR supercomputing team offers community-wide HPC training for astronomers nationally and Swinburne users from other disciplines. The gSTAR webinar series aims to support existing users and introduces prospective users to GPU programming, code testing and optimisation. Given the nature of online learning it's unclear how effective these are, although those that do attend certainly benefit. Many gSTAR users who don't attend, run inefficient code so understanding why they don't participate is important. It may be that a more effective medium is face-to-face training and/or building a properly supported online e-Learning platform (similar to Coursera) that includes comprehensive lesson plans, videos, exercises and a discussion forum to promote peer-driven user support. Infrastructure to support remote workshops is also desirable, although it may be that adequate telepresence facilities already exist at Swinburne.

The Pawsey Supercomputing Centre offers an extensive on-site and off-site HPC training and short courses (https://portal.pawsey.org.au/docs/Training/Courses). Some of the course documentation is available online. The astronomy community would benefit from increased coordination and engagement between all HPC facilities (e.g gSTAR, NCI and Pawsey), in terms of training programs and resourcing, and sharing of expertise. A first step could be a single online Help Desk for Astronomy HPC that support astronomers nationally and better connects facilities.

Currently, there is a lack of comprehensive training in astrostatistics, data mining, and machine learning in astronomy, necessary for analysing large optical survey datasets (e.g. LSST, SkyMapper). Developed by LSST and SDSS astronomers, "*Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data" (Ivezic, Z, Connolly, A.J, VanderPlas, J. and Gray, A.)* is now regarded as one of the foremost texts in astroinformatics. For all applications described in the book, Python code and example data sets are provided, making this a perfect resource for any research/computing training program.

Increased research training around Virtual Observatory projects, data access portals, data and software citation, version control, database languages (e.g. SQL, XML) and architectures, advanced visualisation (3D cubes, interactive plotting), linked-data analysis and representation (e.g. Python+GlueViz), statistical and graphical computing (e.g. R Statistics), cloud computing (AWS, Digital Ocean, NeCTAR Cloud), would benefit the whole astronomy community and also provide early– and mid– career astronomers with the skills they need to transition into alternative careers; either the tech industry (e.g. data science), or within astronomy (e.g. UI/web developer, software developer or astronomy data science research and support roles).

6.2 Data Institutes

There is a trend for foundations and governments to invest in data science institutes. The motivation for this is obvious with the rise of companies like Google and the exploitation of the large-scale harvesting and processing of data.

Data Institutes seek to bring together experts in the many aspects of data science that are required to minimise the time to scientific interpretation by using advanced software engineering, statistics and appropriate computational infrastructure. Many senior astronomers commenced their careers at a time when software engineering and computer

programming were not part of their undergraduate experience. Others may have learnt about programming, but ascended the management tree and have failed to keep pace with modern developments.

In the same way that students have the opportunity to visit national facilities and go on training workshops at places like Narrabri and the AAO, it would be tremendously helpful if Australian PhD students could visit a national astronomy data science facility to aid them in preparing tools for the storage, processing and analysis of their data from a range of experts in programming, processing and statistics and access data "helpdesks" to teach them the skills they will need in either future scientific or industrial careers.

6.3 Recommendations

1. A Research Training Strategy should be developed in consultation with the wider astronomy community, that addresses the specific needs of researchers (particularly PhD, early– and mid– career level) engaged in data-intensive research and/or large survey projects. Consideration should be given to an expanded ANITA training program (or similar) that serves both observational and theoretical astronomers; improved coordination of NCI, Pawsey, gSTAR user support and HPC training; more comprehensive structured courses (similar to those developed for Coursera) in scientific computing (at all levels); astroinformatics, and new data science methodologies. This requires a mix of experts: astronomers, computer scientists and professional software developers at observatories and/or partner organisations willing to volunteer their time.
2. The creation of an astronomy data science institute - the Australian National Data Observatory - that brings together expertise within the astronomy community and facilitates engagement with peak facilities, industry and expertise from other disciplines. Astronomers should have the ability to second ADACS staff to work on their projects. A possible model is QCIF where Australian academics can employ QCIF staff on an hourly rate to work on projects, but that the staff have ongoing positions. ADACS should have the resources to run workshops both on-site and nationally, modern tele/video-conferencing options to aid in helpdesk operations, and enough staff to cover the domains of software engineering, statistics/informatics, and high performance computing. ADACS could comprise a hub with distributed affiliates located across the country in universities, observatories and partner organisations. A small leadership team would be responsible for overall management, coordination, with the support of steering committee. Nodes could be established where there is a critical mass of experts. As a platform, it has a number of clear benefits;
    - Provides the Australian astronomy community with a directory of experts for consultation.
    - May help to increase collaboration between astronomers and software developers
    - Provides the critical mass of expertise required to develop strategies and long term plans (~5+ years) and effective coordination of research/computing training.
    - Can build relationships with other data science institutes, foster cross-disciplinary and industry collaboration, and support exchanges.

- Could provide alternative career paths for software/tool building astronomers (retention of expertise), and greater stability for software developers/engineers working on long-term projects (e.g. SKA).
- Provides resources/expertise for workshops and training programs.
- Can be used to build relationships with industry: Microsoft Research/GitHub/Mozilla Science Lab already sponsor workshops and hackathons, and specific software projects associated with SKA, LSST and SDSS.

7 International examples of data infrastructure models

In looking at successful models from around the world, in many cases Australia has already adopted world class practices for the delivery of merit based allocations of HPC, cloud and large data storage at a number of its facilities. There is of course always room for improvement and efforts to evolve these models should not necessarily just focus on dedicated overseas astronomy facilities but potentially seek to adopt and adapt models from other disciples that are facing similar challenges.

Our review of researchers at many of overseas facilities found that they are dealing with many of the same issues that are faced by Australian astronomy users.

The following recurrent themes were expressed at a number of institutions:

- Insufficient Computational Resources
- Increasing Data Volumes
- Legacy code
- Lack of IT skills
- Rapid changes in technology

All of these are interrelated and different counties/institutions/facilities have attempted to use different models to address these recurrent themes.

Of the numerous different overseas models that are currently in use however all of them are generally shaped by the prevailing funding models that are used within that country or by the facility/institution proving the compute and/or storage resources. As a consequence, there is no one clear model that could be identified as “best practice” however there were a number of aspects to their various approaches to solve these problems that might be applied within the Australian context or could be used to influence the future directions within the Australian research landscape.

The issue of insufficient resources was ubiquitous as there was never enough compute /storage resources (or the funding of these) to meet research needs. It is no surprise then that the only solution to this appears to have been the prioritisation of research based on strategic directions or meritorious need. The way this was approached differed slightly from country to country but generally involved some level of national funding and an application scheme through which researchers could apply.

Within the UK, the Research Councils of the UK (RCUK) allows researchers to apply for access to large-scale compute and storage facilities as part of their grant application. These facilities are largely funded by the council and therefore the council is aware of the resourcing that is available to be granted to researchers and can therefore balance both the supply and the demand sides of the equation.

This is in contrast to the Australian ARC that may fund researchers who then need to secure compute or storage resources separately. This potentially risks the research outcome delivery if the researcher is unable to secure sufficient resources to meet the original research goals.

While it is beyond the direct control of the AAL, it would be possible to advocate for changes in government research funding to encourage the direct funding of compute and storage at

facilities proportional to the allocations requested by a researcher to be funded. The researcher would then receive a cash component and a fully funded compute/storage allocation at a designated facility.

Another approach to address the resourcing issue that has been taken by the UK is the creation of dedicated clusters to service a specific line of scientific enquiry. This allows the aggregation of similar requirements at a national level to create a shared resource. This is in effect this is what AAL has already done with the investment into the gSTAR cluster. This approach has been extremely useful in meeting demand where there is a specialised need or a requirement for testing new technology. The downside to this type of approach can be the proliferation of mini clusters too small to meet research needs and their additional flag fall system administration overhead costs that need to be found. Similarly a proliferation of purpose built clusters can lead to orphaned technology that becomes unusable with little or no uptake if too specialist or the research field moves on.

It is therefore recommended that AAL seek to limit investment into new facilities to avoid diluting investments through unnecessary overheads but seek instead to augment and leverage existing facilities.

Within Germany, the German Climate Computing Centre (DKRZ) has a different approach with access again granted by merit but overseen at the facility level not at the national level. The DKRZ has a scientific advisory group that administers/allocates resources funded by Federal Ministry of Education and Research (BMBF) and advises facility shareholders/partners and the Ministry on scientific direction.

Under a co-ordinated approach with government and the facilities, the AAL could propose a model where by the allocation of time for astronomy research could be managed on Australian facilities and provide regular input into strategic directions for these facilities.

Another nice feature of the DKRZ was the existence of a user group that provides a more direct feedback mechanism to the facility on service operations and directions. AAL could consider working with the facilities to ensure coordinated representation on these facility user groups or to assist with their creation.

In an overview of Italian HPC support for Astronomy (Memorie della Società Astronomica Italiana Supplement, v.1. , p.223 (2003)) the author raised the issue of a lack of funding and advocated strong support for key projects where Italian researchers could make the most impact. HPC resources would be allocated via a competitive process to ensure that resources are used to address the highest-impact challenges.

Along similar lines, the Leibniz Supercomputing Centre (LRZ), runs a competitive process to select a few key projects to run on the entire machine enabling researchers to run large scaling jobs without contention. This is done in conjunction with their major annual maintenance. The LRZ take the system offline for up to two weeks and allow 2 -3 key projects to have use of the entire machine. They also provide facility staff to help the researchers to profile and optimise their code with the aim of having code that is capable of running across the entire machine or at least significantly increased scale to when the project started.

In order to have the greatest impact, the AAL might consider a competitive process to select a few key projects and dedicate a significant resource to optimise particular codes or work with facilities to provide large computational blocks of time to address a specific grand challenge.

For large data intensive problems where the bulk of data is kept offline or on tape for reasons of cost, a similar process could be envisaged but instead of a using a large amount of compute, a coordinated effort is undertaken to bring all the data online at once to enable the rapid processing/reprocessing of a large data set for a period of time before returning it back to tape.

In recognition of the limited resources in compute and storage, many of the overseas facilities offer services to assist researches with using the facilities by providing varying degrees of support. Typically they all offer basic access support but many also offer training and optimisation tools or software support to assist with porting of legacy codes. Although discussed elsewhere, the software engineering effort in porting of legacy code to newer technologies can lead to significant improvements in scaling and speed. A single investment on improving a widely used code can lead to reduced computational time allowing more researchers access to a finite resources.

The National Science Foundation (NSF) *has a vision of a Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) identifies advancing new computational infrastructure as a priority for driving innovation in science and engineering. Innovation occurs through advances in computing facilities, scientific instruments, software environments, advanced networks, data storage capabilities, and the critically important human capital and expertise. Software is thus an integral enabler of computation, experiment and theory and a central component of the new computational infrastructure. Scientific discovery and innovation are advancing along fundamentally new pathways opened by the development of increasingly sophisticated software. Software is also directly responsible for increased scientific productivity and significant enhancement of researchers' capabilities.*

The AAL could adopt the role of the NSF in the diagram below to provide a pool of software developers/engineers to coordinate activities across facilities and research groups to improve commonly used codes or through a priority allocation mechanism, select key projects or codes to provided targeted focused effort.

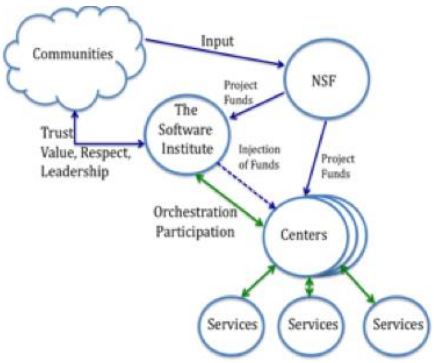


Table 2 shows some early work on the types of compute codes that might be suitable for use on a GPU cluster.

Field	High Efficiency	Moderate Efficiency
Simulation	-Direct N-body -Fixed-resolution mesh simulations -Semi-analytic modelling -Gravitational lensing ray-shooting -Other Monte-Carlo methods	-Tree-code N-body and SPH -Halo finding -Adaptive mesh refinement
Data reduction	-Radio-telescope signal correlation -General image processing -Flat-fielding etc. -Source extraction -Convolution and deconvolution	-Pulsar signal processing -Stacking/mosaicking -CLEAN algorithm -gridding visibilities and single-dish data
Data analysis	-Machine learning -fitting/optimisation -numerical integration Volume rendering	-Selection via criteria-matching

Table 2 Initial assessment of suitability of codes for GPU clusters

Clearly taking time to baseline the efficiency of various codes from across the field can assist researchers selecting the right technology on which to run their codes. With the rapid development of GPU from NVidia and now the Knights Landing Xeon Phi from Intel, the ongoing need to baseline codes and port them to the newer technologies is an ongoing need that will require expanded training (as previously discussed) and potentially re-training of researchers in order to take advantage of these advances.

It is also important for the future planning of many of the facilities supporting researchers to have clear understanding of the technology that is best able to meet their needs.

The question therefore of what to invest in comes down to where the investment will have the greatest impact in achieving the strategic goals for the AAL.

7.1 Recommendations

- Invest in training to enable graduates to assist with the porting and development of new code bases on the latest technology
- Prioritise key projects to focus investment of software engineering staff to enable the greatest impact
- Invest in existing facilities to leverage skilled staff and compute/storage capabilities.
- AAL to consider taking on a more active role in the allocation of astronomy project resources to ensure the success of priority projects aligned with strategic aims through influencing government funding agencies, facility operators or investing directly in compute or storage resources and then allocating these potentially as major blocks of time or through key projects.
- AAL/ADACS should work with international counterparts and HPC facilities to boost the available resources through collaborations and partnerships.

Appendix A – Detailed Storage Requirements

Ref	Telescope Facilities	Archive name	Primary data storage Location	Data Types	Data storage, excluding backups June 2016 (PB)	Total data storage, excluding backups June 2021 (PB)	Volume allocated and/or likely to be approved for future use	Storage 'GAP', excluding backups June 2021 (PB)	Number of additional backup copies
1	BETA ASKAP	ASKAP Commissioning Archive	Pawsey Supercomputing Centre (Pawsey)	Engineering and science commissioning data Unprocessed correlated visibilities (note 1)	0.36	1.2	0.6	0.6	0
2	ASKAP	CSIRO ASKAP Science Data Archive (CASDA)	Pawsey	Data products for: a) images and image cubes b) calibrated visibilities c) source catalogues d) transients e) pulsar timing/searches d) and e) not yet available	0.01 Pawsey disk and tapes	21	10	**11**	1 Pawsey tapes
3	Parkes	CSIRO Data Access Portal	CSIRO Data Centre (CDC), Canberra	Unprocessed pulsar data	0.4 CDC	1.4	1.4 (as needed)	**0**	1+

		(DAP): Pulsar data Parkes pulsar data managed by CSIRO (note 2)		from Parkes pulsar search and timing observations					Other CSIRO data centres
4	Parkes	gSTAR project P002 Parkes pulsar data managed by Swinburne	Swinburne	Unprocessed and Processed data from Parkes pulsar searches and timing observations	1.8 Tape and disk	2.3	2.3 (as needed)	**0**	0 (tbc)
5	ATCA Mopra Parkes LBA: Long Baseline Array	Australia Telescope Online Archive (ATOA)	CSIRO, Marsfield	Unprocessed data for the ATCA, Mopra and Parkes (non-pulsar) Correlated VLBI data Mopra image cube data products	0.2 disks	0.3	0.3 (as needed)	**0**	1 Backup copy on tapes in Marsfield
6	Murchison Widefield Array (MWA)	MWA Archive (unprocessed data)	Pawsey	MWA visibility data in proprietary format. Voltage data in proprietary format.	10 Pawsey tapes	22	20	**2**	1 Pawsey tapes
7a	MWA (see note 3)	MWA Survey Archive	currently ad hoc at UWA	Calibrated visibilities Images and catalogues	0.25	1.0	small	**1**	1

7b	MWA	MWA EoR	Pawsey	Gridded UV data in proprietary format, data cubes power spectra, calibration log files	0.20	1.7	1.7	0	1 (in part)
8	AAT and UK Schmidt	ASVO-AAT	AAO, Sydney	1. AAT & UK Schmidt raw data 2. Data products for: a) AAT surveys (legacy, current and future) b) UKST surveys (inc Taipan and FunnelWeb) c) Other AAT projects	0.03	0.23	0.07	0.16	To be established
9	Sky Mapper Telescope	SkyMapper	NCI, Canberra	Raw images processed images source catalogues	0.3	2	2	0	1+ Backups of all public data releases (details TBD)
10	SKA-Low	SKA Data Centre	TBD	Processed data products for: a) images and image cubes	0	early commissioning data	as needed	0	1

				b) calibrated visibilities c) source catalogues d) pulsar timing and search e) transient events					offsite copy is a SKA L1 requirement
				Storage Totals (PB)	**13.1**	**50.4**	**36.7**	**14.8**	

Table 3 Data storage for observational astronomy data archives located in Australia

Notes

1: ASKAP unprocessed data is only stored for commissioning observations. Commissioning data may not be held indefinitely.

2: The CSIRO DAP supports data management across CSIRO, including many different science areas. The DAP pulsar data archive and the DAP services for CASDA are specialised parts of the DAP.

3. Some MWA science data products may later be managed through the ASVO. An initial design study has been carried out. However, funding for construction is not yet secured. Information in Table 3 and Table 4 are for the major science surveys.

Ref	**Archive name**	**Access restrictions**	**User access notes**	**VO services provided**	**Current level of operations funding adequate for:**		
					Ongoing user support?	**Maintenance and upgrades of systems and software?**	**Maintenance and upgrades of physical infrastructure?**
1	ASKAP Commissioning Archive	Access restricted to ASKAP commissioning team.	Archive provided through *Live Arc.*	None	Yes	Yes	Yes
2	CASDA	Open access following validation of data products	a) DAP web interfaces b) VO services	TAP SIAP (v2) supports 3-d image cut-outs cone-searches ADQL	Yes	Yes	No

3	CSIRO DAP: Parkes Pulsar Data Archive	Open access from 18 months after date of observations	a) DAP web interfaces b) VO services	TAP Cone Search ADQL (DAP VO services to be replaced with CASDA implementations)	Yes	Yes	Yes
4	gSTAR project P002 Parkes pulsar data managed by Swinburne	Restricted to Swinburne Pulsar Staff and Collaborators	terminal access (ssh)	None	?	?	?
5	ATOA		Web interface VO services (under development)	TAP Cone search ADQL (not yet available)	Yes	Yes	Yes
6	MWA Archive	Access to released data upon request to MWA director	Web interface python scripts	TAP Cone Search ADQL	No	Yes	No
7a	MWA Survey Archive	Public access 18 months after last observation	python web VO	SAMP TAP postage stamps	Y	N	N
7b	MWA EoR	Access restricted to MWA EoR Members	*	*	Yes	Yes	?
8	ASVO-AAT	Open access Survey team data requires authentication	Web interface RESTful services	TAP Cone Search SIAP, SSAP ADQL To be implemented	No	Yes	No

9	Sky Mapper Telescope	Access within Australia for one-year proprietary period, then open access	VO services, web interface	TAP Cone Search SIAP ADQL	No	No	No
10	*LSST*						
11	SKA Data Centre	Open after a proprietary period (that is still to be determined)	APIs, web interface	TBD VO services, including VOEvents	Still in early planning stages. SKA construction is not funded at this point in time, but commitments had been made for 80% of the required money.		

Table 4 Data access and operations for Australian astronomy data archives

Notes

ASVO-AAT: Will be managed through AAL for first two years. This is expected to host the AAT archive, including all legacy AAT surveys.

Appendix B – Further Detail on Training

Summary of recommendations (for discussion)

Scientific computing/data science in astronomy:

- Data Science tools/e-Research Infrastructure: databases, cloud computing, APIs, data scraping, data visualisation, data communication.
- Best practise in scientific computing: version control, collaborative coding, testing, software and data citation, reproducible science.
- Astro-statistics and Machine Learning for large survey data: many tools have already been developed for astronomers, they just need to be taught.
- Research training is more effective when combined with small projects, building tools, experimenting with early release datasets, large infrastructure projects.

HPC training for astronomers:

- Best practise in HPC is a high priority: code testing, optimization
- Webinar success is difficult to evaluate, but it seems that more interactive courses and face-to-face teaching may be more effective.
- Better coordination of gSTAR, NCI and Pawsey HPC support for astronomers (what is the demand?)

A strategy for research training:

- National guidelines for data and software citation and best practises in research – data management, reproducible science.
- Workshops and training programs require significant resources (materials, people, sometimes IT infrastructure), and running many different one-off workshops can be hard to organise. It would be more effective to invest in a number of regular workshops to address specific needs.
- A strategy that identifies key research training needs, consolidates existing efforts would be useful for costing workshops/training over the next 5 years and the required resources
- Universities/Research Centres are responsible for research training programs but the quality varies across Institutions. AAL shouldn't be responsible for general computer science/programming etc.

Related things:

- Coding for STEM? Debugging the gender gap.
- Research training for alternative careers

Notes for reference

Astronomers' ability to harvest the sky for information has rapidly escalated since the dawn of the digital astronomy era. Recent advances in radio and optical astronomy instrumentation are leading to Petabyte datasets that are challenging to record, archive and process for science. Increasingly the big data bottleneck is an impediment to scientific progress. Many of our senior astronomers are well schooled in the planning and interpretation of observations, but were not trained in the IT skills to curate and process these petabyte datasets. In short, astronomers are facing a fire-hose of data and are ill-

prepared for it. The immensity of the data demands new approaches and techniques to used to store, analyse, interpret and explore data. New approaches are required now, including harnessing novel technologies such as graphics processing units (GPUs) for massively parallel computation, in order to prepare for this large/complex data paradigm shift.

The rise of "big data", "data-science" and new methodologies

The rise of "big data" and "data science" in the technology industry and its prevalence in academic research is creating a new generation of savvy researchers that are adopting more general, industry standard tools and practises. The rise of data science institutes, particularly those initiated by leading astronomers; e.g. the Berkeley Institute for Data Science (University of California), eScience Institute (UW), Centre for Data Science (NYU), and others, have lead many research groups and individuals to shift their efforts on dealing with issues and creating new tools to better handle data-intensive research. This includes promoting best practises in scientific computing, exploring new machine-learning techniques, initiating collaborative coding projects (e.g. AstroPy), and developing community resources (e.g. IPython Notebook, Jupyter, AstroML, scikit-image, GlueViz). The majority of tools are developed in collaboration by early and mid-career researchers in dual software development/data science and/or pure academic research roles, and are widely used by the community. Training programs that aim to increase uptake of tools and promote best practise range from one-off workshops (SciCoder, Software Carpentry), conferences and hackathons (e.g. SciPy, DotAstronomy, ANITA summer schools), to community connected and regularly scheduled meet-ups (Hacker Within). The result is that researchers are no longer limited to specific astronomy data analysis tools and traditional IT infrastructure. For better or worse the line between astronomer and software-developer is blurring. One positive aspect is that astronomers can more easily transition into more lucrative data science positions in the tech industry. The astronomy community also recognises the value of retaining expertise, with STScI and the Institute of Cosmology & Gravitation (Portsmouth) now offering Data Science Fellowships.

***Recommendation*: An Astronomy Data Science Centre (or Network) could provide the necessary platform for addressing the various issues around data-intensive research:**

- Provides the whole Australian astronomy community with a directory of experts for consultation
- Provides a critical mass of expertise that can help drive strategies and develop longer term plans (~5+ years) for training and infrastructure etc.
- Better enable the community to develop clear strategy for astronomy research & computing training directory/network of experts
- Can be used to connect to existing data science institutes around the world and foster new collaborations
- If properly supported, would provide alternative career paths for astronomers and more stability for current instrument scientists/software developers. With the rise of Data Science Fellowships we are likely to see an exodus of astronomers to the tech industry.
- Can be used to coordinate community workshops and training programs.
- Can be used to build relationships with tech companies: Microsoft Research/GitHub/Mozilla Science Lab already sponsor workshops and hackathons, and at some level individual research projects.

Current community-wide research/computing training workshops

(excluding observational techniques workshops)

ANITA/ASA Lectures (http://anita.edu.au/lectures/): ANITA supports PhDs, and postdocs by hosting a series of online workshops throughout the year. A number of past lectures have addressed specific issues around big data and data mining and include topics such as *databases and SQL – applied, computational Bayesian theory, R Statistics for astronomers.* Although ANITA is a well supported community the /lectures appear to sporadic and organised somewhat *ad hoc* depending on current needs/trends.

ANITA/ASA Workshop/summer school: Every two years (roughly) the ANITA summer schools alternate between concepts/research training specific to theoretical astronomy (simulations/supercomputing) and more general astroinformatics. Although still biased towards theoretical astronomers, the astroinformatics workshops are designed to serve as an introduction to a wide range of IT tools that are essential for the modern astronomer, especially PhD students. These skills include programming (Python, version control), Unix scripting, database construction (SQL) and use, internet technologies (VO tools, data formats) and data mining.

Dot Astronomy – Day Zero (Sydney - Nov 2015): The 2015.Astronomy7 conference, included a "Day Zero" hack training day: http://dotastronomy.com/events/seven supported by a **Web Development & Research Tools for Astronomers** guide (soon to become its own website/community resource). This was the first time we held a pre-hack training day for Dot Astronomy and I think it made a huge difference, especially for PhDs and researchers new to version control, collaborative coding and hacking. There were a number of reasons for incorporating Day Zero into the conference, but it was mainly in response to a cry from the community. You can read about it here: http://dotastronomy.com/blog/2015/09/countdown-to-day-zero/

Recommendation: *The ANITA community appears to be well-supported if not always well resourced. More regular workshops, particular astro-informatics/data-science that target both observational astronomers and theorists. We should think about how to grow this.*

Specific scientific computing recommendations from Amr Hassan:

- Build a clear sequence of courses, that move the researchers from beginner to intermediate in a pre-defined path should be the goal.
    - Introduction to Linux and Shell Scripting
    - Level 1 and 2 of Python with some software engineering and testing concepts
    - SQL Level 1 and 2 with Astronomy related examples and some DB design concepts.
    - Basics of Structure programming and Object oriented programming
    - Matlab and/or R beginner to intermediate.
    - Light introduction to Cloud Computing
    - C/C++ Level 1 and 2 (Optional)
- Building courses for Coursera will be a great community service, but it needs more budget. It is a secondary objective of course.

Support and training for high performance computing

gSTAR HPC/GPU webinars: The Swinburne Supercomputing Group also runs a number of webinars to support existing users and introduce prospective users to code testing and GPU programming. It's unclear how effective these are, although those that do attend certainly benefit. Many gSTAR users who don't attend, run inefficient code so understanding why they don't participate is important. It may be that are more effective medium is face-to-face training. Optimisation and code testing are a high priority. Infrastructure to support remote workshops is desirable, although it may be that facilities already exist at Swinburne

Pawsey: The Pawsey Supercomputing Centre offers an extensive on-site HPC training and short courses (https://portal.pawsey.org.au/docs/Training/Courses) but this requires participants to have access to adequate travel funding. Some of the course documentation is available online. The astronomy community would benefit from increased coordination and engagement between all HPC facilities (e.g gSTAR, NCI and Pawsey), in terms of training programs and resourcing, and sharing of expertise. A first step could be an single online Help Desk for Astronomy HPC that support astronomers nationally and better connects facilities.

NCI: In terms of cloud computing with HPC astronomers can access NCI's Tenjin (private cloud). Unlike the NeCTAR Research Cloud, and open source/commercial services like Digital Ocean, Tenjin is housed entirely within NCI, providing access to NCI's 20PB global high-speed parallel filesystems.

Recommendation: *There seems to be a lack of coordination between NCI, Pawsey, and gSTAR, in terms of HPC training and astronomy user support services. A more coordinated approach that enables sharing of expertise, in addition to an online support system (that connects support at each facility) would benefit the community.*

Improved communication/sharing of resources across the community.

There still seems to be a lack of consistency across the community with respect to available data storage options, best practices around collaborative coding and version control, software and data citation. Many researchers still don't understand what DOIs, Zenodo, ASCL, FigShare, and GitHub are for and where they should go to generate DOIs for data and software citation. Should software be deposited in GitHub? ASCL? Or SourceForge? Should researchers be making data collections available on Research Data Australia? Attribution statements can also vary across journals.

The AAS now provides a set of guidelines for citing data repositories in AAS Journals (AJ/ApJ):

- Citing data/software repositories:

https://github.com/AASJournals/Tutorials/blob/master/Repositories/CitingRepositories.md

- Using data/software repositories:

https://github.com/AASJournals/Tutorials/blob/master/Repositories/UsingRepositories.md

Improving community-wide communication is one reason we are turning the .Astronomy7 **Web Development & Research Tools for Astronomers** guide into a web resource.

Tools for Astronomical big data: training for next five years

Linked-data, linked-views, and 3D interactivity will make it easier to deal with large, complex datasets. The ability to build community tools is much easier and quicker than previously, purely because collaborative coding repositories like GitHub lead to 100s of contributors. eg. AstroPy.

- Dataverse and Authorea now link to WWT
- Glue and GlueViz is now funded by NASA
- On the horizon: CARTA/NOAO (visualiser for big data cubes)
- Nightlight app will be launched in 2016: http://nightlightapp.io
- Data cubes will be ubiquitous e.g. http://nanocubes.net
- Visualisation of large imaging data: e.g. http://www.noao.edu/meetings/bigdata/files/moolekamp_2015-2-17_toyz.pdf

Astroinformatics training:

Statistics, Data Mining, and Machine Learning in Astronomy presents a wealth of practical analysis problems, evaluates techniques for solving them, and explains how to use various approaches for different types and sizes of data sets. For all applications described in the book, Python code (e.g. AstroML, Scikit) and example data sets are provided. The supporting data sets have been carefully selected from contemporary astronomical surveys (for example, the Sloan Digital Sky Survey) and are easy to download and use. The accompanying Python code is publicly available, well documented, and follows uniform coding standards. Together, the data sets and code enable readers to reproduce all the figures and examples, evaluate the methods, and adapt them to their own fields of interest.

- Describes the most useful statistical and data-mining methods for extracting knowledge from huge and complex astronomical data sets
- Features real-world data sets from contemporary astronomical surveys
- Uses a freely available Python codebase throughout
- Ideal for students and working astronomers

Recommendation: A research/computing training program should be developed around this book.

Other training/workshops:

The current workshops outlines previously appear to serve a good fraction of the community. However, they don't cover all aspects of computing training and they rely heavily on the availability of astronomers with the necessary skills to volunteer their research time. They also focus on mastering individual topics that tend to be either introductory or advanced, rather than combining tools that build on another.

Increased research training around ASVO projects, data access portals, data–science in general (Inc. data mining methods and data visualisation), and data storage and visualisation of Petascale astronomy and best practises in data and software citation would benefit the

whole astronomy, and provide early– and mid– career astronomers with the skills they need to transition into alternative careers in either the tech industry (e.g. data science careers), or within astronomy (e.g. UI / web developer, software developer or astronomy data science roles).

Ideally all astronomers should have the skills/confidence to[4]:

- Query existing databases (e.g. SDSS) more effectively. SQL is a hurdle for many PhD and early– and mid– career researchers.
- Setup and manipulate databases, and export/import queried data into new databases
- Use databases to build multi-wavelength datasets for their everyday research (farewell spreadsheets, ascii files, printed tables of data)
- Setup and use personal VMs for everyday research and for hosting projects e.g. pre-configured cloud instances such as Digital Ocean, NeCTAR Cloud, DIT4c, and DockerHYPERLINK "https://www.docker.com/what-docker"
- Build simple web frontends (e.g. a responsive personal website or project website)
- Implement version control in their everyday research (e.g. Regularly use Git with GitHub or BitBucket)
- Document their own software/code properly (best practise in reproducible science)
- Develop web projects to enhance existing research outputs, for outreach or purely for skills development (e.g. Interactive websites with javascript)
- Use (or build)) tools for the community that make research easier.
- Make their data VO compliant from the outset (if possible)
- Be able to transition between coding languages (or at least be familiar with different languages)
- Be able to make informed decisions/assess the different options for storing and sharing data.
- Be able to make informed decisions about data/software citation and licensing.
- Participate in hackathons or other workshops that rely on collaborative teamwork
- Know where or who to go for to ask help on any of the above topics.

Aside from research, this is important for another reason. Permanent jobs in astronomy are becoming more competitive. Building up a set of skills that enable researchers to better transition to jobs outside of academia. It also enables astronomers to propose or even build demo modules (MVPs) or data-visualisation add-ons for ASVO projects (or similar) or for existing Python/Astropy routines. These types of projects are invaluable for those seeking alternative career paths in data science. While astronomers are in high-demand, data science fellowships are becoming more and more competitive. It's not enough to have a PhD in astronomy.

Examples of community connected, informal, peer-driven workshops

The Hacker Within (THW) began as a student organisation at the University of Wisconsin-Madison, and is now reborn as a collection of such chapters around the world. Current

[4] ***Note: The following list does not replace the need for properly curated data archives such as the CSIRO Data Access Portal, the gSTAR/PASA Data Sharing cluster, and various ASVO nodes. Nor does it replace the need for designing data access portals to be user friendly (e.g. minimal SQL queries), or the need for dedicated software developers/engineers. We don't want to promote bad computer science practises***

chapters include the University of California (Berkeley), Illinois, and the University of Wisconsin-Madison. [NOTE: the links are to THW websites] Nascent chapters include Yale University, Swinburne University of Technology, the University of Melbourne, and Michigan State University.

Each of the chapters convenes a community of researchers, at all levels of their education and training, to share their knowledge and best practices in using scientific computing to accomplish their work. Typically they run as 2.5 hour weekly meet-ups that focus on specific topics relevant to scientific computing. The goal is to introduce researchers to the plethora of open-source tools that can be exploited, to increase productivity and enhance existing projects, and to encourage the development of off-shoot projects and contribution to community-developed research tools. There are two reasons why this is important:

- The tenets of scientific research (e.g., data control, reproducibility, and peer review) suffer in projects that fail to make use of current development tools such as version control, testing, and comprehensive/automatic documentation. To avoid these pitfalls, the numerous Hacker Within Chapters exist for the purpose of sharing skills and best practices for computational scientific applications.
- To prepare researchers for alternative careers in the technology industry, for example via the Insight Data Fellows and Science to Data Science fellowships, or for Data Science Fellowships with in academia.

Connected GitHub repositories allow each chapter to share materials, and help researchers at other universities/groups join the program. The **Berkeley Institute for Data Science** now recognises THW as an important part of research and runs this as one of their core training programs.

Longer workshops /boot camps that focus on developing expertise in specific topics.

e.g. ANITA summer school, Astro Hack Week (a Moore–Sloan Data Science initiative). Example topics:

- Exploratory Data Analysis and Visualisation
- Working with databases and understanding workflows (e.g. SQL, MySQL). Introduction of basic, concepts, main advantages, programming language/syntax, real-world astronomy examples e.g. working with/exploring early release datasets.
- Advanced workshops: Hadoop (Hadoop Data File System – HDFS Apache spark?)
- Best practice in software citation and version control (DOIs/Zenodo/ASCL/GitHub)
- Software and Web Development tools
- Data Mining, Machine Learning Algorithms, Statistics, Bayesian Inference
- Sampling methods: Monte Carlo, Approximate Bayesian and beyond.
- Statistics with R
- Interactive visualisation, e.g. D3js, GlueViz.
- Relational databases and SQL

Hack Days:

Successful Hack Days have been running at the AAS and RAS-NAM for the past two years, initiated by the members of the ever-growing Dot Astronomy and Astro Hack Week

communities. Hack Days are traditional events in software development circles, where people with skills, ideas, and the willingness to dedicate a day of their lives get together to make interesting projects happen. The day usually begins with pitches for hack ideas ranging from novel data analysis techniques to new public outreach sites.

AAS Hack Day 2016: http://www.astrobetter.com/wiki/AASHackDayHYPERLINK "https://github.com/AASJournals/Tutorials/blob/master/Repositories/UsingRepositories.md"

AAS Hack Day 2014: http://www.astrobetter.com/blog/2014/01/22/aas-hack-day-2014/HYPERLINK "http://www.astrobetter.com/blog/2014/01/22/aas-hack-day-2014/"

RAS NAM Hack Day 2014: http://www.nam2014.org/hackdaywiki/index.php5/Main_PageHYPERLINK "http://www.nam2014.org/hackdaywiki/index.php5/Main_Page"

Example hack projects:

- A working demo of a new or modified algorithm — for example: optimize.js
- Outline of a larger project, with some key features scoped – for example: The journal of brief ideas
- A new visualisation of an old dataset
- A modified algorithm applied to a new dataset
- A demo module (pitch) for an ASVO project or other community data hub
- Setting up your own virtual machine and perhaps creating your first database for your data
- A mash up of pre-existing pieces of code to perform some new function – for example: lens modelling

Appendix C – Summary of feedback on draft report

Summary of key feedback from AAL's committees/working groups and ANITA:

- AAL's optical, radio, eResearch and multi-messenger advisory committees/working groups broadly agreed that there are a range of gaps in the astronomy community's skills, expertise, and ability to maximise the scientific return from eResearch infrastructure and data. While there was general support for some level of increased AAL investment in people to address these gaps, there were a range of views about the scale and the focus of such investment.
- Of the recommendations in the draft report, OTAC prioritised investment in staff to build and operate data archives for optical/IR facilities, and in training (as described under section 6.1 of the report) to upskill the community. OTAC noted that future needs for large data storage and HPC computing resources are driven largely by the requirements of the new radio telescopes and theoretical simulations, and therefore, OTAC did not make any formal recommendations about the working group's recommendations regarding storage and HPC.
- RTAC recognised that the astronomy community is facing growing challenges in developing pipelines, algorithms, and other software, particularly to deal with radio Big Data. RTAC broadly supported modest levels of investment in experts to help address those challenges.
- The MMAWG generally supported the working group's recommendations. Priorities for MMAWG included computing pipelines and personnel for gravitational wave astronomy, and infrastructure to enable integration and mining of data from multi-wavelength and multi-messenger facilities.
- ANITA supported AAL investment into both hardware and people/skills development, provided it benefited the broad Australian astronomy community. Highest priority for ANITA was securing additional astronomy-dedicated time at a minimum level of 10-20M CPU hours/year. Regarding investment in people, ANITA preferred that staff are distributed around the country rather than concentrated in a single institution. Furthermore, ANITA recommended that personnel be funded via a national "computational fellows" scheme, where institutions could make a co-contribution in order to host a `fellow'.
- Significant investment by AAL in "hard" infrastructure such as data storage and high performance or cloud computing was generally not supported by OTAC and RTAC. Instead, these groups preferred that AAL use the proposed pool of experts to interface with the existing National computing facilities in order to maximise astronomer's access to the National computing resources, and the extent to which those facilities are designed to meet astronomers' needs.
- There was general consensus that a new organisation should not be established to deliver the eResearch services. Instead, they should be provided by existing organisation(s), building on established teams of experts with strong track records in astronomy eResearch projects, and leveraging support from the host in order to provide employment security for the staff and ensure longevity of the services.
- It was also agreed that any AAL investment in eResearch infrastructure and services should not duplicate existing services and resources. Instead, it should leverage National eResearch programs and resources and fill in gaps not covered by other groups or schemes.

- It was generally agreed that the proposed services should be built up gradually over time - with assessment points to measure performance and return on investment - and administrative and governance overheads should be minimised.

Summary of key results from community-wide user survey:

Users were asked to what extent they expect a range of factors will limit their research over the next year. Table 5 lists these factors along with the percentage of users who expect that their research will be moderately or severely limited by these factors over the next year. The overall responses are presented, along with responses broken down by the sub-fields[5] that users indicated were one of their main areas of astronomy. *Italics* indicate limiting factors that the proposed Astronomy Data and Computing Services could begin to address, while the remaining factors largely rely on "hard" infrastructure.

	% indicating their research will be severely or moderately limited by this factor			
Limiting factor in the next year	**Overall (N=100)[6]**	**Radio/submm/mm (N=50)**	**Optical/IR (N=40)**	**Theoretical/ Computational (N=52)**
Not enough long-term data storage	60%	60%	50%	63%
Insufficient network speed or bandwidth	51%	54%	48%	46%
Difficulty sharing your data with other researchers	49%	54%	49%	43%
Lack of training/expertise in advanced statistical and informatics techniques	49%	45%	49%	51%
Not enough short-term data storage (e.g. temporary storage for data generated on HPC facilities)	46%	52%	31%	44%
Difficulty performing advanced visualisation of data/simulations	46%	52%	28%	55%
Not enough CPU computing/ processing time	46%	34%	45%	56%
Lack of training/expertise in programming or software development	43%	38%	48%	44%
Lack of training/expertise in machine learning/artificial intelligence techniques	42%	42%	40%	42%
Insufficient I/O when using high-performance computers	36%	39%	25%	37%
Insufficient memory when using high-performance computers	35%	43%	15%	39%
Difficulty combining, comparing or cross-matching multiple external datasets	34%	35%	40%	29%
Not knowing what storage and computing resources are available, or how to access them	32%	32%	30%	41%
Lack of training/expertise in software version control and	32%	24%	35%	37%

[5] Multi-messenger astronomy (MMA) sub-field is not shown separately in this table, as only 8 respondents indicated that MMA was one of their main fields of research. However, their responses are included in the overall numbers.
[6] While the survey had 120 respondents, only 100 completed this question.

best-practises in software citation				
Lack of training/expertise in database management	32%	34%	30%	37%
Difficulty accessing external datasets	30%	36%	26%	24%
Lack of training/expertise in creating virtual observatory-compliant databases and services	22%	28%	20%	25%
Not enough GPU computing/processing time	19%	24%	10%	20%
Difficulty sharing your software/code with other researchers, or developing code collaboratively	16%	12%	10%	18%

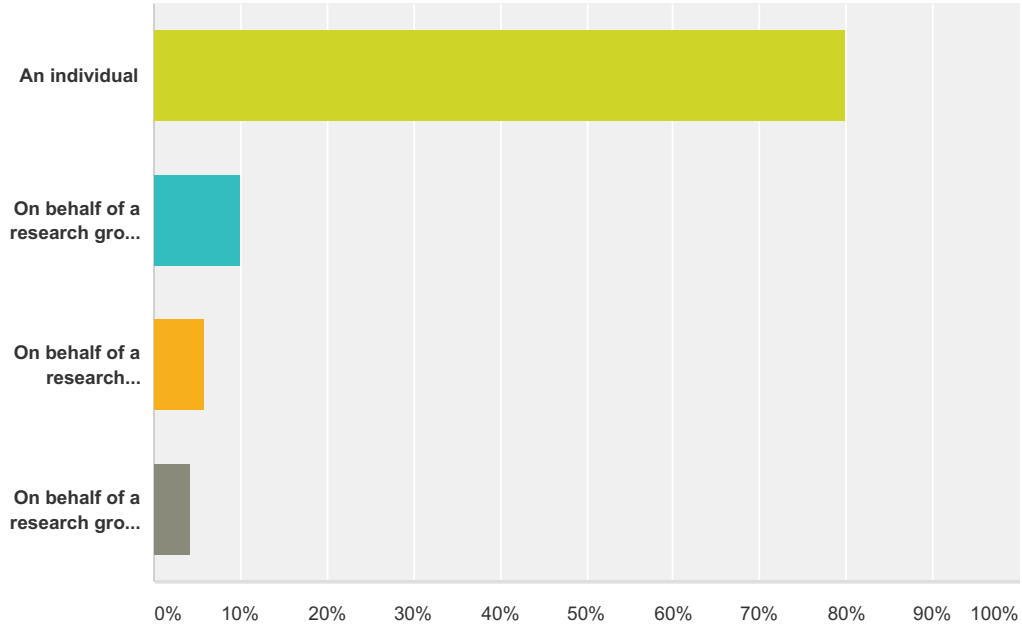
Table 5 Expected factors limiting research in the next year

Survey respondents were also asked what tools and languages they would be interested in learning. Of the 58 users who answered this free text question, 25 listed Python or Python-related languages/tools (e.g. Advanced Python, PythonGPU, PythonSQL), making this the most common response. Other tools/languages mentioned by multiple respondents were: PostgreSQL/SQL, machine learning tools, statistical analysis tools, VO-related tools, C, C++, CUDA, R, IDL, MPI, Julia, GPU programming.

Appendix D – User survey results

Q1 You may respond to this survey as an individual or on behalf ofa research group. If the latter, you are encouraged toconsult with your research group in order to complete this survey. Please indicatewhether you are responding to this survey as:

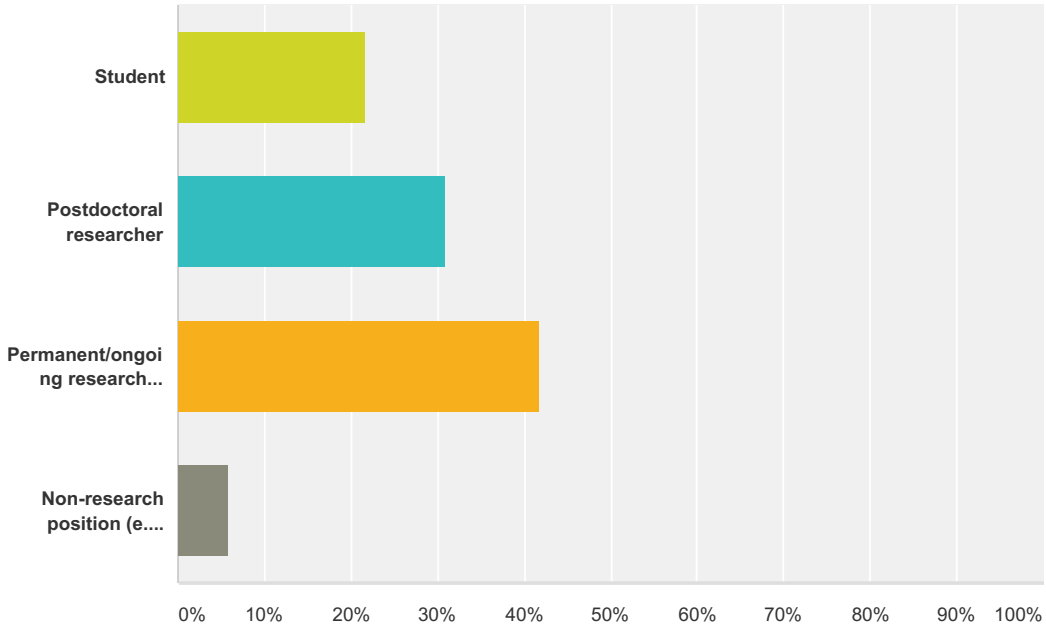
Answered: 120 Skipped: 0



Answer Choices	Responses	
An individual	**80.00%**	96
On behalf of a research group comprising2-5 people	**10.00%**	12
On behalf of a research groupcomprising6-10 people	**5.83%**	7
On behalf of a research group comprising more than 10 people	**4.17%**	5
Total		**120**

Q2 Your employment status is best described as:

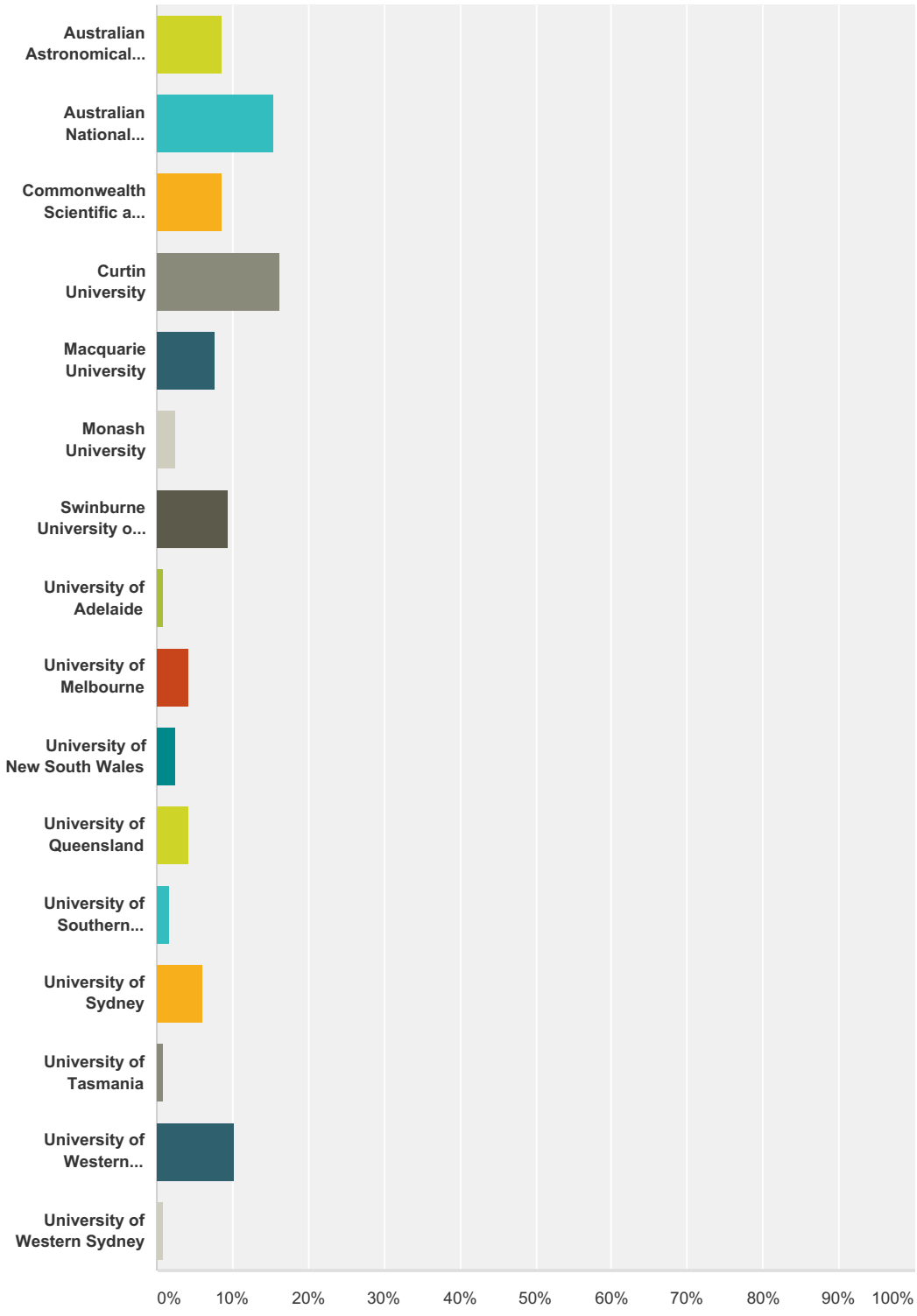
Answered: 120 Skipped: 0



Answer Choices	Responses	
Student	**21.67%**	26
Postdoctoral researcher	**30.83%**	37
Permanent/ongoing research position	**41.67%**	50
Non-research position (e.g. non-research IT personnel)	**5.83%**	7
Total		**120**

Q3 What institution are you from (choose from dropdown menu or else indicate in text box)?

Answered: 117 Skipped: 3

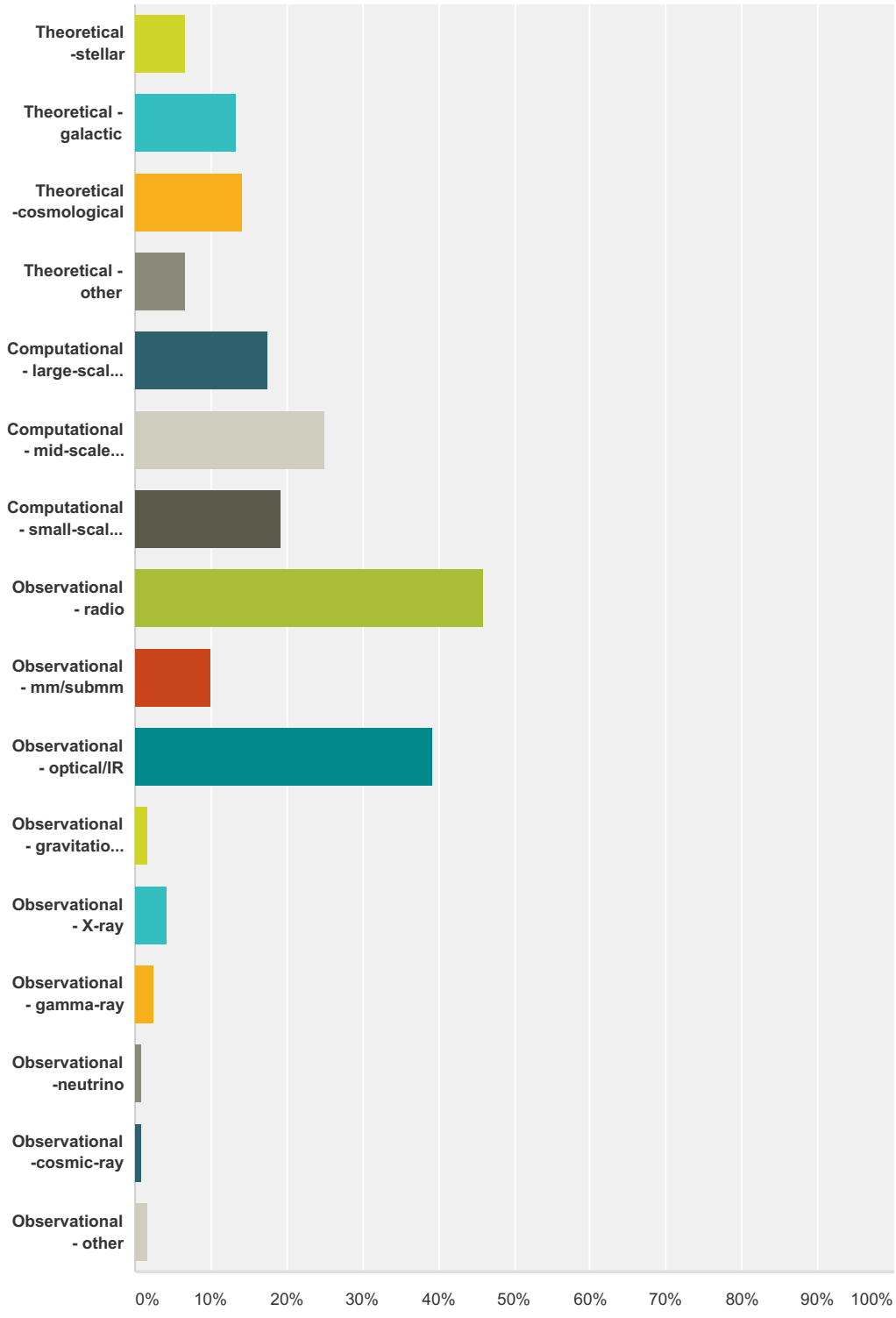


Answer Choices	Responses	
Australian Astronomical Observatory	8.55%	10
Australian National University	15.38%	18

Commonwealth Scientific and Industrial Research Organisation	**8.55%**	10
Curtin University	**16.24%**	19
Macquarie University	**7.69%**	9
Monash University	**2.56%**	3
Swinburne University of Technology	**9.40%**	11
University of Adelaide	**0.85%**	1
University of Melbourne	**4.27%**	5
University of New South Wales	**2.56%**	3
University of Queensland	**4.27%**	5
University of Southern Queensland	**1.71%**	2
University of Sydney	**5.98%**	7
University of Tasmania	**0.85%**	1
University of Western Australia	**10.26%**	12
University of Western Sydney	**0.85%**	1
Total		**117**

Q4 What are your main fields of research (you may select multiple fields):

Answered: 120 Skipped: 0

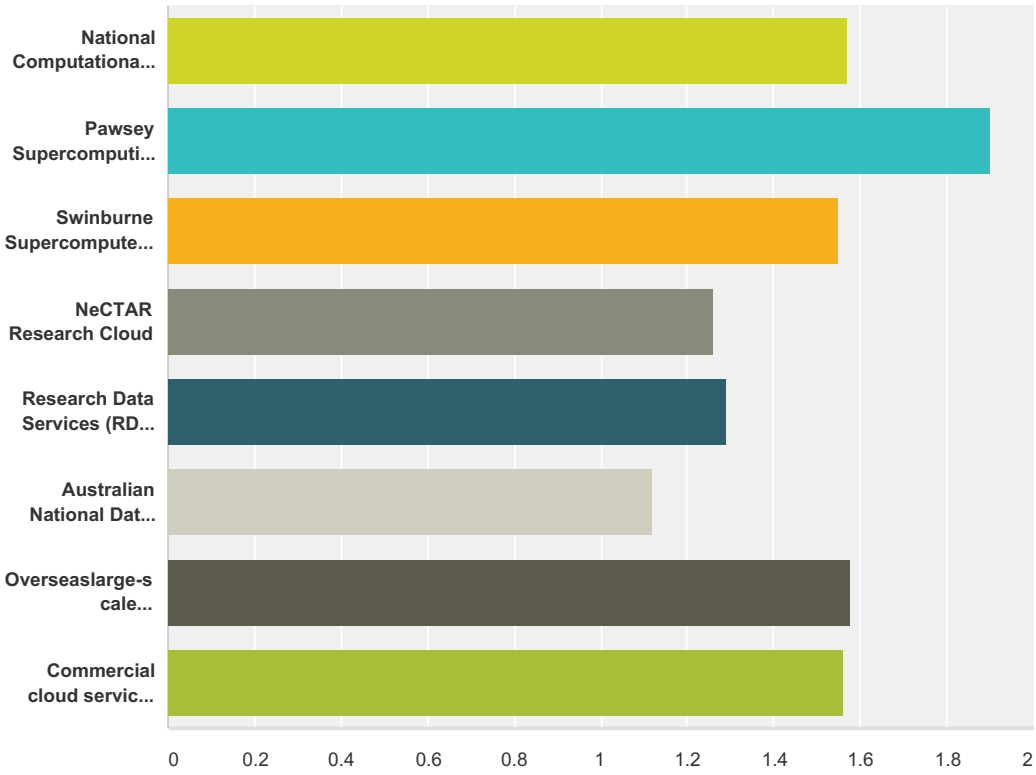


Answer Choices	Responses	
Theoretical -stellar	**6.67%**	8
Theoretical - galactic	**13.33%**	16
Theoretical -cosmological	**14.17%**	17

Theoretical - other	**6.67%**	8
Computational - large-scale (e.g. requires significant time on large-scale HPC facilities)	**17.50%**	21
Computational - mid-scale (e.g. canbe performed on university-scale or mid-scale HPC facilities)	**25.00%**	30
Computational - small-scale (e.g. canbe performed on local workstations)	**19.17%**	23
Observational - radio	**45.83%**	55
Observational - mm/submm	**10.00%**	12
Observational - optical/IR	**39.17%**	47
Observational - gravitational wave astronomy	**1.67%**	2
Observational - X-ray	**4.17%**	5
Observational - gamma-ray	**2.50%**	3
Observational -neutrino	**0.83%**	1
Observational -cosmic-ray	**0.83%**	1
Observational - other	**1.67%**	2
Total Respondents: 120		

Q5 In the past year, how many timeshave you used the following computing resources?

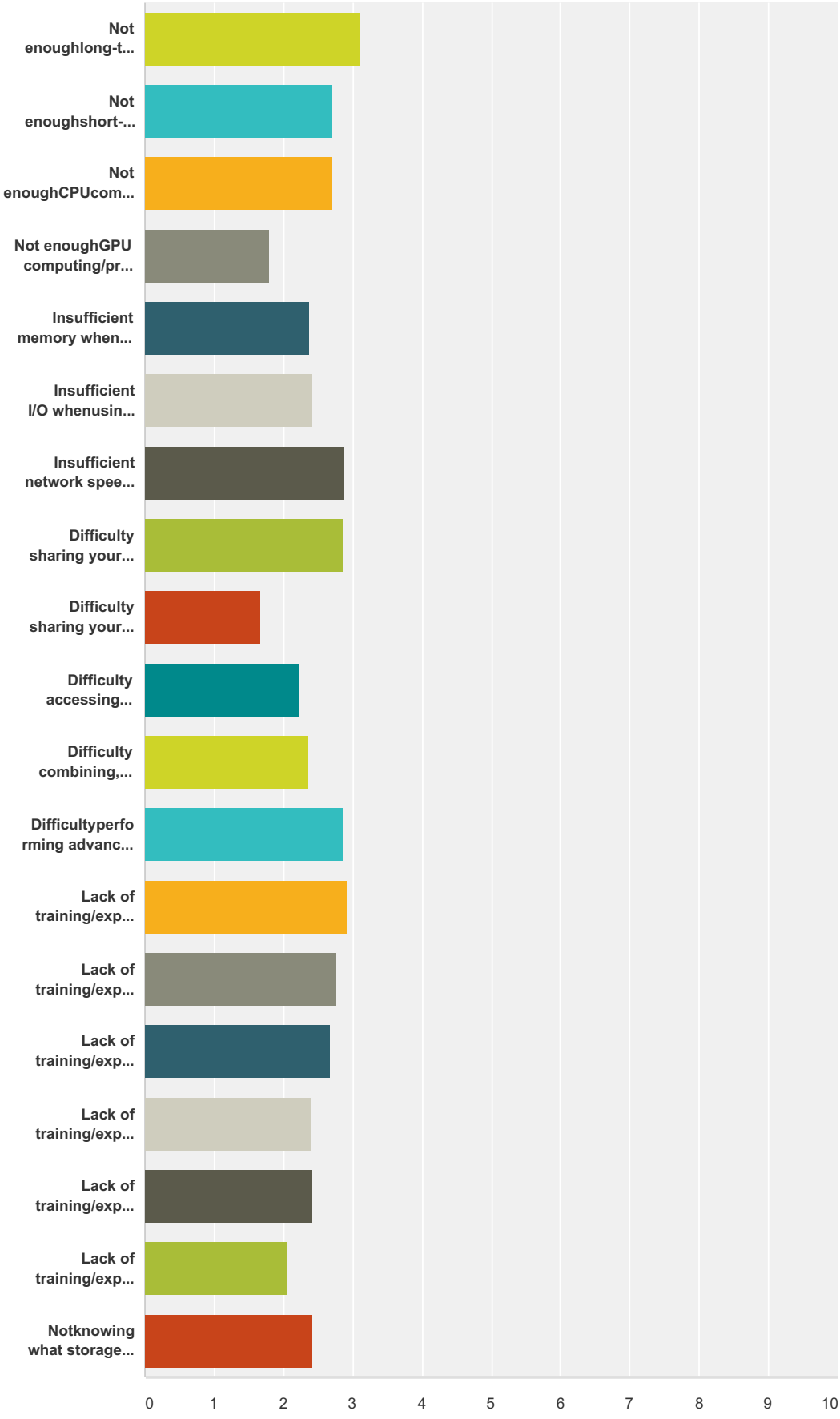
Answered: 119 Skipped: 1



	Never	Once	2-5 times	6 or more times	Total	Weighted Average
National Computational Infrastructure (NCI)	**75.86%** 88	**3.45%** 4	**8.62%** 10	**12.07%** 14	116	1.57
Pawsey Supercomputing Centre	**65.81%** 77	**2.56%** 3	**7.69%** 9	**23.93%** 28	117	1.90
Swinburne Supercomputer (gSTAR & SwinSTAR)	**78.45%** 91	**3.45%** 4	**2.59%** 3	**15.52%** 18	116	1.55
NeCTAR Research Cloud	**86.36%** 95	**4.55%** 5	**5.45%** 6	**3.64%** 4	110	1.26
Research Data Services (RDS, formerly RDSI)	**86.73%** 98	**4.42%** 5	**1.77%** 2	**7.08%** 8	113	1.29
Australian National Data Service (ANDS, e.g data discovery, project registry)	**94.69%** 107	**0.00%** 0	**4.42%** 5	**0.88%** 1	113	1.12
Overseaslarge-scale supercomputing facility	**76.32%** 87	**0.88%** 1	**11.40%** 13	**11.40%** 13	114	1.58
Commercial cloud services (e.g. Amazon Web Services, Azure)	**77.19%** 88	**3.51%** 4	**5.26%** 6	**14.04%** 16	114	1.56

Q6 To what extent do you expectthe following factors will limit yourresearch activities inthe next year?

Answered: 100 Skipped: 20

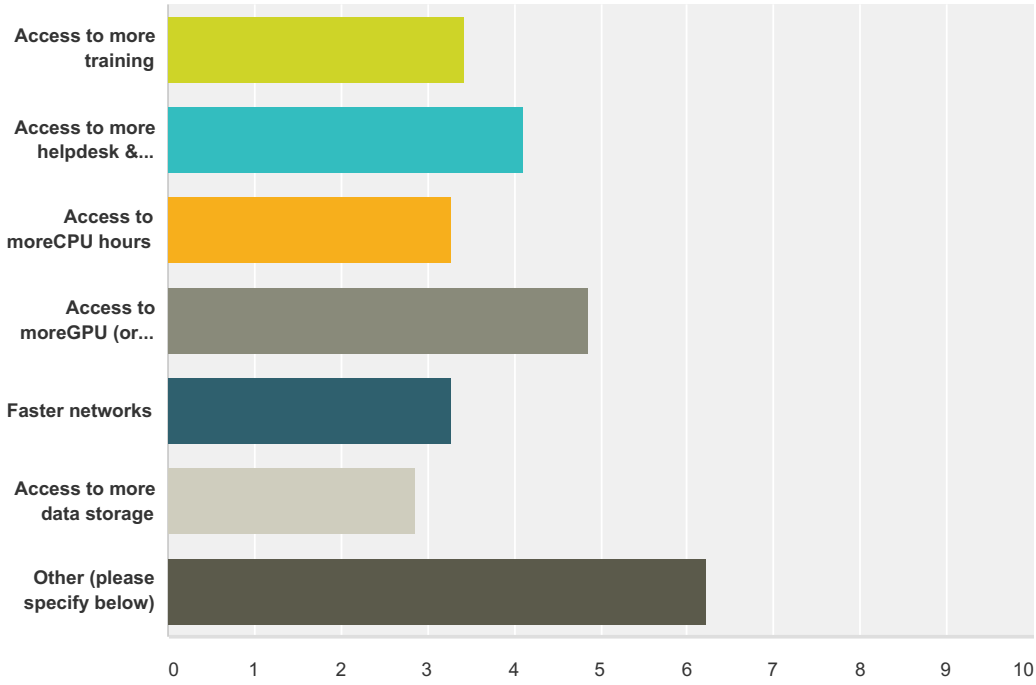


	Not at all	Moderately	Severely	Slighty	N/A	Total	Weighted Average
Not enoughlong-term data storage	**26.26%** 26	**40.40%** 40	**19.19%** 19	**14.14%** 14	**0.00%** 0	99	3.12

Not enoughshort-term data storage (e.g. temporarystorage for data generated on HPC facilities)	**39.39%** 39	**32.32%** 32	**14.14%** 14	**9.09%** 9	**5.05%** 5	99	2.71
Not enoughCPUcomputing/processing time	**35.00%** 35	**33.00%** 33	**13.00%** 13	**17.00%** 17	**2.00%** 2	100	2.71
Not enoughGPU computing/processing time	**61.62%** 61	**14.14%** 14	**5.05%** 5	**8.08%** 8	**11.11%** 11	99	1.80
Insufficient memory when using high-performance computers	**42.42%** 42	**26.26%** 26	**9.09%** 9	**17.17%** 17	**5.05%** 5	99	2.39
Insufficient I/O whenusing high-performance computers	**42.42%** 42	**26.26%** 26	**10.10%** 10	**16.16%** 16	**5.05%** 5	99	2.43
Insufficient network speed or bandwidth	**28.00%** 28	**36.00%** 36	**15.00%** 15	**21.00%** 21	**0.00%** 0	100	2.89
Difficulty sharing your data withother researchers	**27.27%** 27	**35.35%** 35	**14.14%** 14	**23.23%** 23	**0.00%** 0	99	2.86
Difficulty sharing your software/code with other researchers, or developing code collaboratively	**69.70%** 69	**12.12%** 12	**4.04%** 4	**12.12%** 12	**2.02%** 2	99	1.66
Difficulty accessing externaldatasets	**43.43%** 43	**25.25%** 25	**5.05%** 5	**23.23%** 23	**3.03%** 3	99	2.23
Difficulty combining, comparing or cross-matching multiple external datasets	**40.40%** 40	**28.28%** 28	**6.06%** 6	**20.20%** 20	**5.05%** 5	99	2.36
Difficultyperforming advanced visualisation of data/simulations	**27.27%** 27	**33.33%** 33	**13.13%** 13	**18.18%** 18	**8.08%** 8	99	2.86
Lack of training/expertise inadvancedstatistical and informatics techniques	**19.19%** 19	**37.37%** 37	**12.12%** 12	**30.30%** 30	**1.01%** 1	99	2.93
Lack of training/expertise inmachine learning/artifical intelligencetechniques	**28.00%** 28	**29.00%** 29	**13.00%** 13	**21.00%** 21	**9.00%** 9	100	2.76
Lack of training/expertise in programming or software development	**29.00%** 29	**34.00%** 34	**9.00%** 9	**27.00%** 27	**1.00%** 1	100	2.67
Lack of training/expertise in software version control and best-practises in software citation	**36.00%** 36	**21.00%** 21	**11.00%** 11	**28.00%** 28	**4.00%** 4	100	2.41
Lack of training/expertise indatabase management	**32.00%** 32	**23.00%** 23	**9.00%** 9	**31.00%** 31	**5.00%** 5	100	2.43
Lack of training/expertise in creatingvirtual observatory-compliantdatabases and services	**43.00%** 43	**19.00%** 19	**3.00%** 3	**25.00%** 25	**10.00%** 10	100	2.04
Notknowing what storage and computing resources are available, or how to access them	**33.33%** 33	**23.23%** 23	**9.09%** 9	**28.28%** 28	**6.06%** 6	99	2.43

Q7 Over the next 5 years what type of resources will need to grow to support your research? Please rank the following in order of importance (1 = most important):

Answered: 100 Skipped: 20



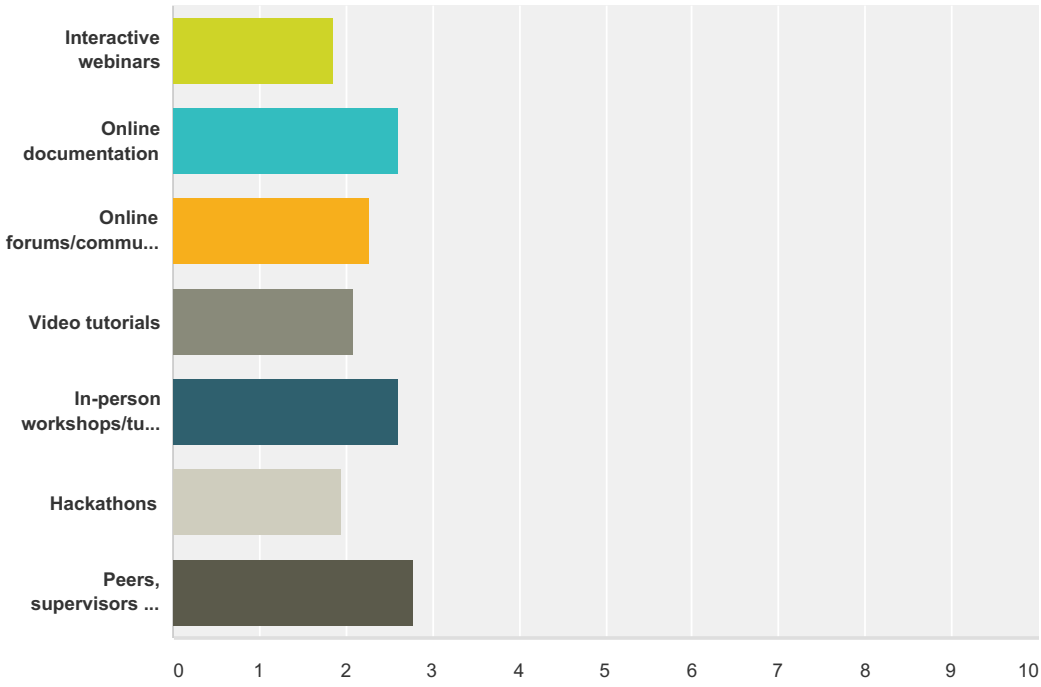
	1	2	3	4	5	6	7	Total	Weighted Average
Access to more training	**22.34%** 21	**14.89%** 14	**15.96%** 15	**10.64%** 10	**19.15%** 18	**14.89%** 14	**2.13%** 2	94	3.43
Access to more helpdesk & expertsupport services	**4.26%** 4	**18.09%** 17	**18.09%** 17	**14.89%** 14	**18.09%** 17	**20.21%** 19	**6.38%** 6	94	4.11
Access to moreCPU hours	**23.66%** 22	**11.83%** 11	**13.98%** 13	**21.51%** 20	**22.58%** 21	**5.38%** 5	**1.08%** 1	93	3.28
Access to moreGPU (or other massively parallel-processing) hours	**4.40%** 4	**6.59%** 6	**10.99%** 10	**15.38%** 14	**9.89%** 9	**43.96%** 40	**8.79%** 8	91	4.87
Faster networks	**12.77%** 12	**17.02%** 16	**30.85%** 29	**17.02%** 16	**15.96%** 15	**5.32%** 5	**1.06%** 1	94	3.27
Access to more data storage	**22.22%** 22	**31.31%** 31	**11.11%** 11	**18.18%** 18	**10.10%** 10	**6.06%** 6	**1.01%** 1	99	2.85
Other (please specify below)	**11.11%** 4	**0.00%** 0	**2.78%** 1	**0.00%** 0	**0.00%** 0	**0.00%** 0	**86.11%** 31	36	6.22

Q8 If you are able to quantify your computinggrowth needs, please provide estimates foryour requirements and specifications in2021 (e.g. storagevolumes, number ofCPU/GPU hours/year, memory per core, I/O bandwidth, T/Pflops, etc).

Answered: 28 Skipped: 92

Q9 How effective andusefuldo you usuallyfind the following training/learning methods?

Answered: 100 Skipped: 20



	Not at alluseful	Somewhatuseful	Veryuseful	Don't know	Total	Weighted Average
Interactive webinars	**18.00%** 18	**45.00%** 45	**8.00%** 8	**29.00%** 29	100	1.86
Online documentation	**4.00%** 4	**31.00%** 31	**64.00%** 64	**1.00%** 1	100	2.61
Online forums/communication tools (e.g. Slack)	**10.00%** 10	**46.00%** 46	**35.00%** 35	**9.00%** 9	100	2.27
Video tutorials	**13.00%** 13	**54.00%** 54	**21.00%** 21	**12.00%** 12	100	2.09
In-person workshops/tutorials	**3.00%** 3	**32.00%** 32	**63.00%** 63	**2.00%** 2	100	2.61
Hackathons	**15.00%** 15	**31.00%** 31	**12.00%** 12	**42.00%** 42	100	1.95
Peers, supervisors or mentors	**2.00%** 2	**18.00%** 18	**79.00%** 79	**1.00%** 1	100	2.78

Q10 What tools/languages would you be interesting in learning?

Answered: 58 Skipped: 62

Q11 Any other feedback or comments?

Answered: 6 Skipped: 114

Q12 Please provide your email below if you are happy for us to contact you, in the (unlikely) event that we would like to follow-up on any of your responses.

Answered: 39 Skipped: 81