

# Investment Plans and Platforms for Collaboration

## Introduction

An earlier summary of progress reports provided a capability by capability view of issues.

This document takes a perspective on each specific PfC issue in turn and provides a view across all investments. The key PfC issues are:

- *Data storage management, access, discovery and curation* to improve interaction and collaboration;
- *Grid enabled technologies and infrastructure* to enable seamless access to the facilities and services required in various research fields;
- *Support skills* to assist researchers in developing and using this infrastructure effectively;
- *High performance computing* to allow analysis, modelling and simulation; and
- *High quality network access through high capacity bandwidth* to permit interaction with diverse data and computing resources.

## Overall Comments

Generally speaking the power of advanced cyber-infrastructure<sup>1</sup> as a corner stone of future research is not strongly represented across all the investments. The potential result is a piecemeal approach to building cyber-infrastructure which is a risk that should be managed.

Some capabilities have proposed that funding from 5.16 be applied to specific and dedicated activities within their own sphere of activity; which conflicts with the guidelines that have been provided. Funds from 5.16 could only be supplied to another platform in this way at the direct expense of the generic services on which all the investments depend, and on which other research interests also depend.

A productive framework for future discussion involves identifying the services that 5.16 could provide from the 5.16 investment plan that would meet the needs of each capability. A difficulty is that the investment plans relate to specific investments and do not cover the more general research needs of the various communities, to which 5.16 must pay attention. Also, where a service turns into infrastructure permanently dedicated to a single facility, its priority within 5.16 funding is necessarily very low.

In the shared service space, significant budgetary concerns are apparent.

- Many network requirements are left implicit and the assumption of shared funding may be invalid. For instance, multi-site integrated platforms (5.1, 5.2 and 5.3) and the data rates for astronomy and characterisation may require dedicated circuits and upgrades on backbone segments.
- All of the requirement for the theoretical modelling that might be associated with the effective use of the capabilities has been left implicit. Only 5.10 and 5.13 provide an estimate of demand and the 5.13 figure exceeds the total computing power available across the APAC partnership.
- The focus on data management, integration and inter-operation, will also appear as urgent areas for support in other research communities, placing additional demands on the 5.16 capability.

The current budget for Platforms for Collaboration will not be able to support all these requirements.

The intention to develop multi-site integrated platforms, enhanced data accessibility and sharing, and higher levels of collaboration, all require significant growth in levels of available ICT expertise.

---

<sup>1</sup> Cyber-infrastructure encompasses the nationwide services and expertise needed to support effective eResearch within and across all disciplines; including: data capture and management; data publication, discovery and re-use; data analysis; computational modelling; collaboration systems, grid inter-connectivity; and networking.

## Highlighted Items

Because the text covers a broad agenda, this section brings forward items that need attention.

There is unfunded, significant, explicit and implicit pressure on services from 5.16.

The model for 5.16 capabilities will need to include scope for generic and dedicated services, so that enhanced service levels can be provided on a co-funded basis where the generic service levels prove insufficient. This may represent a variation to the NCRIS access and pricing model.

Also, a policy towards fee-for-service in institution's infrastructure support may lead to conflict with NCRIS access and pricing goals.

### Cyber-infrastructure investment

About \$10M pa of NCRIS funds may be being directed towards specific elements of cyber-infrastructure (from the current investment plans). Aggregating with 5.16 funding yields a funding rate about half that of the SII funding rate over the last three years. The result may be a reduced rate of general cyber-infrastructure development and significant difficulty in providing additional services.

The activities within that \$10M pa are mostly focussed on data access and platform integration across multiple sites, with a small component in modelling, and hardly any funding in HPC or networking.

### Data Management

Overall significant attention has been paid to the capture, curation, and access, for "community" data.

A high degree of contention is apparent over the data management practices that should be applied and required for researcher specific or private data. The implications of public funding need to be agreed.

### Grid

Because key services are not yet operational (such as national authentication) detailed plans for integration of the proposed investments with those services is not yet possible.

Hence, some compliance pressure may be needed to ensure independently developed technology plans are changed to integrate with the developing cyber-infrastructure, as it matures.

### Expertise

A large fraction of the implicit cyber-infrastructure expenditure will relate to expertise, which will be in short supply. Mechanisms to grow and share key human resources need to be provided.

### High Performance Computing

An estimate is required of the overall needs of modelling and theoretical research within each capability area, and any increase arising from research activity levels generated by the use of the new facilities.

The benefits of speed vs capacity and the best means of user engagement need to be investigated.

### Networking

The goal should be subscription funded research traffic from all NCRIS funded facilities to all Australian researchers; from shared facilities through to every researcher's desktop.

As investments are made, the network backbone is likely to become less homogeneous (some links with higher speeds), so that a revised pricing framework may need to be developed and agreed.

### Specifics

The 5.2 Atlas proposal should include a scoping study to better define the priority of effort.

The 5.16 funding itemized in 5.3 is a cash transfer that would reduce general research support.

## Cyber-infrastructure Investment

For simplicity, these comments relate to NCRIS funds (exclusive of co-investments) and are rounded percentages. The majority of the investment across the capabilities in cyber-infrastructure is related to accessibility to data and within that the dominant expense would be expertise.

### 5.1 Evolving Biomolecular Platforms

The investment in cyber-infrastructure is hard to estimate in this activity as it is embedded in the 'omic platforms. Taking the bioinformatics activity and some of the 'omic activities as an investment in data accessibility and cross platform systems suggests about 30% in cyber-infrastructure.

### 5.2 Integrated Biology Systems (5.2a Phenomics and 5.2b Atlas)

The investments in phenomics will have minor components contributing to cyber-infrastructure, but largely support laboratory services. The entire investment in the Atlas is a contribution and therefore this capability is applying about 20% of its investment towards cyber-infrastructure.

### 5.3 Characterisation (5.3a X-ray/Neutron and 5.3b Microscopy/Imaging)

In characterisation, the need for cyber-infrastructure is recognised in platform integration (for microscopy and imaging) and services integrated to research environments (for x-ray and neutron imaging). However, while a cyber-infrastructure component (some of which is in-kind) equal to about 25% of total NCRIS 5.3 funding is identified, an allocation of only 4% of NCRIS 5.3 funds is made.

### 5.4 Fabrication and 5.5 Biotechnology products

These investments relate directly to laboratory capabilities and support for improved researcher use of those facilities. Cyber-infrastructure that might be needed for the related communities in property modelling for materials or system modelling in biological cases are not provisioned (cash or in-kind).

### 5.8 Biosecurity

The delineation of expenditure in this capability is hard to draw as all the expenditure is aimed at improving services that interface or could interface into the cyber-infrastructure. However the direct contribution is the collaboration environment which represents about 25% of the proposed investment.

### 5.10 Astronomy

The Astronomy community has been an early adopter of virtually all aspects of cyber-infrastructure and views cyber-infrastructure as integral to the mission. Unsurprisingly, the proposed investment is self contained but the portion that might be considered cyber-infrastructure is also unclear.

### 5.12 Integrated Marine Observing System

The marine discipline is accustomed to handling common data sets and has the advantage of close ties to earth systems science which has a strong cyber-infrastructure orientation. The majority of investments are in enhanced data gathering, with about 10% in work directed towards data access; however this is a lower bound as the all the investments will include some data accessibility effort.

### 5.13 Structure and Evolution of the Australian Continent

Researchers in geosciences have been working towards cyber-infrastructure goals for some time. The proposed investment delivers additional or improved data services, grid support and the development of synthesising information products building on those data services. Treating the last two as direct investments in cyber-infrastructure yields 30% of the total investment.

## Data management, access and discovery

The path forward for a coordinated national approach to research data management is still unclear.

The Investment Plans are neutral, in that they appear to provide no additional or different problems to those already present in the system. However the NCRIS process does provide a valuable opportunity to review the direction and can assist the establishment of a national framework.

In general the investments leave responsibility for data with the facility (5.1, 5.2-phenomics, 5.10) or with the institutions related to collections (5.2-atlas, 5.12, 5.13) and the researcher otherwise (so that data is then managed according to the work practice of the researcher's environment).

This last category is where the greatest uncertainty exists. Different parties have different views on the best way to integrate data generating facilities into a national data management system where the data is 'private' to a specific research project, and its potential value outside of that project is undetermined.

*Specifically: there is not yet a consensus on the data access obligations that should flow from public funding, that can be reasonably implemented, and which can lead to more frequent data re-use.*

### State of play

It seems inevitable that research institutions supporting research projects will need to provide, or provide access to, data retention services as part of their research support function (noting that some economies of scale and shared services may arise where large volumes of data are concerned).

Initial surveys and interviews leading up to a planned consultation by 5.16 in this area, show a number of pertinent factors:

- Many research communities are identifying key underpinning data and establishing community agreed collection, curation and access standards
- Most researchers are reluctant to build their research on derived information products, ie the analysis of this underpinning data, preferring to redo the analysis from the base data assets
- Many researchers keep data in 'desk draws', and most expect an increasing scale of data gathering, which suggests there may be a rapidly growing level of unmanaged data
- While the leading proponents of the value of long lived data have convincing strategic arguments, more generally, researchers are reluctant to share their analysis data and are sceptical of the benefits that might arise from access to other researcher's analysis data
- Many institutional IT directors report that enterprise data holdings are growing at high rates and that centrally held research data is a minor component of the overall volume of enterprise data. However, the visibility of non-centrally managed enterprise data holdings to institutional IT directors can be uneven.

### Categories of data

It seems likely that the consultation process will identify different categories of data, and will reach agreement around the practice required to sustain data in at least some of the categories.

For instance, many communities have travelled a similar journey to identify and agree the collection, curation and access standards for the basic data around which a community exists (eg. weather, geoscience, bioinformatics, astronomy and others).

On the other hand, there is little agreement on the appropriate treatment for data generated within research projects where traditionally data exists solely for the use of the directly related researchers.

Consequently, the value of the consultation is likely to be in an understanding that different categories of data have different cost/benefits for re-use, different private/public status, and also are in different stages of readiness for the adoption of more stringent data management practices, and that this can inform our data management goals and the systems we deploy against those goals.

## Public Access to data

An issue arising directly within the 5.3 investment, and mentioned to varying degrees in the others, concerns the perspective that data collected through public funded activities is public data.

While it is tempting to use the case of Astronomy, Marine and others, to illuminate this discussion, very different categories of data are involved. As stated above, understanding the implications of the differences between categories of data is the key to a national data retention policy.

A variety of policy and framework decisions need to be made and some form of agreement reached around them, and indeed actions are already underway in this regard. Issues that need further addressing directly related to propositions put in the investment plans include:

- The use of a data source is only one contribution (and fund source) in the overall contributions (and fund sources) that support a research activity. Therefore, the principles that guide the decisions over access rights to data from these data sources needs to be agreed.
- The data generated by a data source may be only part of the data generated in a research project. Hence the preservation of research data needs to be addressed from the context of the researcher, and preservation at data sources may need to integrate with relevant data management regimes.
- The rationale, cost/benefit and even the means for permanently retaining limited access data at data sources needs to be determined.
- However, as the benefits in providing a reliable and readily identifiable source for any publicly accessible data is almost self-evident, some national framework for determining responsibilities is desirable.

Further thinking on this topic should arise in the 5.16 Data Management consultation.

## 5.1 Evolving Biomolecular Platforms and Informatics

The 'omic platforms need to import and generate data for use across each platform and need to curate that data according to their own standards and requirements. Therefore data management is a central function of these platforms and a key integrative activity in the bioinformatics investment.

In these platforms, some data will be generally accessible and some will be confidential.

Overall data sizes can be expected to be in the tens of terabytes, but with extensive churn, so that the costs will be in labour and expertise, again a reason to manage the data 'in-house'.

### 5.2a Phenomics

As a key objective in phenotyping is to share the phenotype information across many researchers and research interests, the data will need to be maintained and curated in perpetuity.

Also, the data sets contemplated in phenotyping are diverse in their content and yet their value is in their integrated interpretation, implying a need for specialist expertise.

These factors support the data being held, curated and made accessible as part of the operation of the phenomics facilities, as it is those facilities that will have the relevant interpretative expertise.

Additional specialist expertise will be needed if the standards of practice and access methods for this data are to support provenance requirements of research performed using the mice and plant variants.

### 5.2b Biological Collections

The Atlas of Living Australia represents a major investment in metadata capture, data linkage and distributed information search and retrieval, and supposes the inclusion of a large number of major collections; including metadata, imagery, geo-referencing, and potentially DNA data.

While there are no technological barriers to delivering the Atlas, there are many competing technological and adoption considerations, so that a scoping study as the first phase (as in 5.8) seems advisable. Apart from the technological options, the scoping study should address:

- The basis for making cost benefits decisions on distributed, mirrored, and/or remote management of data, as it is likely that some mix of services will be required.
- An analysis of the network bandwidth and data movement requirements needed for data mining and image processing to be performed using remote computing resources.

### **5.3a X-Ray Techniques (Synchrotron) and Neutron Scattering (ANSTO)**

As users are not co-located with the image sources and as the data is private to each researcher; permanent retention of all data on site is not required for operational or community support reasons. However, there is a detectable movement towards retaining all data at the synchrotron and the stated intention is to retain data in perpetuity and to make it publicly accessible at ANSTO.

As one of the largest sources of large volumes of research data in the country, the synchrotron will have a very significant impact on network requirements and the national data management philosophy.

For neutron imaging, data volumes are expected to be around 2-3 TB per year, related to 400 experiments and a smaller number of principle investigators, with each image generation taking many hours and sometimes days. This image acquisition rate easily permits data to be transmitted off-site on completion of an experiment, within a typical 1Gbps research connection.

### **5.3b Microscopy and Imaging**

Both microscopy and imaging intend to develop a co-operative platform of instruments operating from within participating groups. In this case, as users are co-located with the image sources the full range of activities including retention, curation, analysis and re-use are possible.

However, the data is private to each researcher; and of course the facility needs to service researchers 'outside' of the facility. So unlike the 'omic platforms which maintain community data that all researchers can make use of, the microscopy and imaging platforms need to make the case for retaining private data, and solving the problem of identifying future data owners in perpetuity.

Operationally, data volumes are 10-50 TB per year, image generation is time consuming, and access to very significant compute resource is required to process images quickly (in minutes rather than hours).

## **5.8 Networked Biosecurity Framework**

The collaboration environment would need to grow to include the data, analysis systems and staff of around twenty major centres in the biosecurity area, some of which have up to 400 staff.

At present the data in all these sites is managed according to the requirements of each discipline; no overarching metadata schema exists; each discipline has its own preferred data analysis and classification schemes. As the goal is to provide the means to access data sources and associated analysis tools by other members of the network, the data can remain separately managed at each site.

Hence no major data management issues exist, though effort will be needed to supply data in agreed standard formats for both content and metadata.

## **5.10 Optical and Radio Astronomy**

The Astronomy proposal highlights the way many research communities are moving to collect, curate and preserve base data collections of common value to the community. The data sets are significant being hundreds of Terabytes, and this scale is expected to grow quickly over time. The costs of the data retention for the extensions to the instruments proposed are included in the 5.10 budget.

Overall, a better understanding is needed across the broader astronomy data collections, of the way in which the data holdings can be associated with the scale of computing that could be needed to perform significant datamining or image processing on those data sets.

### **5.12 Integrated Marine Observing System**

The analysis of the data requirements suggest an annual acquisition rate of over 120 TB per year in primary data, subject to scale up should additional variables, additional sensors or resolution enhancements be deployed.

The data will be held and made accessible through significant installations that are funded external to the NCRIS process.

An obvious question concerns the rationale for two separate data management activities, the eMII and the AO-DAAC. The objective should be a single point of access to all data.

### **5.13 Structure and Evolution of the Australian Continent**

This capability supports a diverse set of data generating activities, which while categorised into classes, are relatively independent within and between classes in terms of obtaining their data. As the access standards are already agreed and as resource has been invested in the data management activities, no specific problem arise.

The proposal includes a requirement for distributed and centralized storage, with an identified demand for 750 TB of hierarchical storage with an additional 20 PB of off-line storage required to archive all raw geospatial data

## Grid-enabled technologies

Few of the capabilities reference the need for system wide authentication, authorisation and accounting support or provide a detailed view of their needs against specific compute and data grid capabilities (5.13 being the exception).

This is most likely a reflection of inexperience with grid technologies and that the communities have yet to develop a service level approach to cyber-infrastructure.

Most research groups represented in the investment plans appear to be strongly wedded to a traditional 'in-house' and 'make-do' approach to IT.

However, requirements for grid-like capabilities arise as follows:

Grid service	5.1	5.2a	5.2b	5.3a	5.3b	5.4	5.5	5.8	5.10	5.12	5.13
Integrated multi-site platform	X	X			X						
Distributed shared view of data	X	X	X		X			X	X	X	X
Remote access to compute power	X		X	X	X	X	X	X	X		X
Real-time multi-site data analysis									X		
Virtual presence and control				X	X				X		
Access Grid like collaboration						X	X	X		X	X

In general:

- Existing technology is capable of providing a seamless view by a community across a range of data collections separately managed by multiple organisations. The provision of fine grain access control within such distributed collections is less well developed.
- Limited scale solutions exist for providing research groups direct access into their private data holdings across multiple participating institutions. Limited scale means that these technologies involve a rapidly escalating level of administration effort as the number of users and sites grows, and generally assume a range of enterprise security permissions can be managed collectively.
- Remote access to compute power can be provided. However real-time access, is a continuing technical challenge especially where the remote resources are strongly scheduled. The virtual machine technology introduced into commodity chips this year, is expected to provide the capability for near instant access to remote compute resources (on a pre-agreed basis of course).
- Real-time applications that link distributed systems into a single analysis can be achieved using grid-like components. However, at present, such applications require dedicated network bandwidth and fixed system availability to be reliable.
- The existing examples of production quality remote presence (such as the virtual critical care unit) require dedicated network bandwidth and systems. This is unlikely to change in the near future.
- Access grid technology has existed for many years. The capability for user management of access grid sessions remains a problem; especially in the presence of transient network, system or configuration faults.

Overall, a seamless view of remote instruments, combined with integrated analysis using compute and data resources drawn on-demand from around the country, is a vision that will be hard to achieve as a generic service in the NCRIS investment timeframe.

It will be possible to do so for exemplar cases. It is important to appreciate that such developments are *leading edge* and that a wide range of specialist expertise will be needed to build and operate them.

The state of technologies also suggests that some staging be involved: as it is possible to gain immediate benefits from sharing data and supporting collaborative activity; whereas benefits from dynamic real-time access to instruments or compute platforms will take longer to secure.

## 5.1 Evolving Biomolecular Platforms and Informatics

The Proteomics, Genomics and Metabolomics Facilities have the full range of grid potential from shared data through to operating seamlessly across multiple sites and multiple organisations.

These are strong candidates for exemplar grid enabled platforms.

Each facility includes a separate investment in informatics infrastructure which needs to be managed in a flexible manner as the technology can be expected to change significantly over the next 5 years.

### 5.2a Phenomics

The phenomics facilities are more likely to deploy standard enterprise technologies than grid-like capabilities.

### 5.2b Biological Collections

As a case of general access to public good data assets, the Atlas of Living Australia presents few technological difficulties and can be expected to deploy relatively basic grid capabilities.

It may, however, lead to important opportunities for datamining, correlation and image processing which could use grid capabilities to bring compute power to the data as a second stage development.

### 5.3a X-Ray Techniques and Neutron Scattering

From a grid point view, these facilities should integrate with the national authentication and associated authorisation standards (as they develop) so that remote resources can be accessed from the facilities to assist researchers while on site.

Support for grid-like access to data (ie automatically from remote compute capabilities) would need to be provided in the event the facilities maintain data for researchers.

The development of remote presence should be undertaken in stages (as is planned) to 'learn as we go' and to work with similar overseas developments to assist the maturation of the required technologies.

### 5.3b Microscopy and Imaging

These platforms like the 'omics are strong candidates for exemplar grid enabled platforms. They both involve researchers and instruments co-located and at multiple sites, with a wider off-site research community, leading to plans for integrated instrument access and multi-site compute and data systems.

## 5.4 Fabrication and 5.5 Biotechnology products

While neither of these capabilities has invested in modelling support, it is still the case that researchers at the proposed sites may wish to be able to access their home research environment and undertake some modelling from the sites. This would suppose high speed network connectivity and might require some grid capabilities depending on the interaction level required.

Also, it might be the case that some research interaction could be valued, so that collaboration supported from the sites to the rest of the community might be worthwhile. Again the level of interaction that might be valued would need to be assessed.

## 5.8 Networked Biosecurity Framework

The Networked Biosecurity Framework is a high profile exemplar collaboration goal within the proposed investments. An essential capability for its success involves the creation of a shared work space, able to hold and distribute any data from a wide array of sources, and which limits access to nominated persons within a broad community.

That goal is most likely to be achieved in urgent situations if the technology is deployed in routine use as part of every day work practice, so this investment aims at establishing and managing a common use national collaborative environment within the bio-security sector.

The incident response objective implies the use of grid-like solutions to achieve appropriate levels of security when required, and supposes a high speed network, and authentication and authorisation systems (for this purpose) reaching outside of the research community along with unknown requirements on data security and retention policy.

The resulting collaboration environment could have an extended reach (beyond the nominated centres) by hosting or accessing data for 3<sup>rd</sup> parties through standard web services and portal technologies.

### **5.10 Optical and Radio Astronomy**

These investments can be expected to be appropriately web or grid enabled due to the leading position this community has demonstrated in developing and adopting grid technologies over many decades.

However, the development of a national authentication system and the provision of a growing pool of data management expertise should provide the community with a less fragile resource base.

### **5.12 Integrated Marine Observing System**

A key output of the Marine Observing system is a coherent view of the gathered data and derived information products. Sufficient grid middleware and domain specific browsing and data trawling tools already exist to support the goals. Some effort (in addition to that proposed) will be required to integrate with national authentication standards as they develop, especially as some of the community tools do not integrate at present.

The Ocean Observing Node involves several data gathering systems supported by a dozen organisations in various combinations. There appears to be no need to consider an integrated 'grid' platform across the participants in order to implement the Ocean Observing Node.

The Coastal Ocean Observing Nodes also involve multiple data gathering systems, each supported by a large number of organisation. Again, few significant technical issues are likely to arise in a federation where participants serve their own data against public queries in agreed standards.

### **5.13 Structure and Evolution of the Australian Continent**

This proposal demonstrates the strongest grasp of grid possibilities and technologies, along with the service orientation implicit in the concept of cyber-infrastructure.

The investment intends to rely on nominated grid services being supplied through 5.16 investments.

While the 5.16 plan is yet to be detailed, these services are expected to be provided.

The 5.13 plan also supports staging of benefits in two levels:

- Establishing sources of data that can be combined by virtue of common directory and community agreed data access standards and interfaces
- Developing composite services which provide value adding products but which also require more grid-like capabilities for combining analysis services with data access (on demand)

## Expertise

The earlier analysis of investments in cyber-infrastructure across the NCRIS capabilities, identifies a level of spending that will ultimately translate in to a demand for expertise.

The Investment Plans themselves show that the various communities are at different stages of development towards a cyber-infrastructure perspective; which means they will necessarily have access to very different levels of such expertise.

What is perhaps less clear is that the expertise needed will be in multiple and entirely unrelated specialisations, such as curation of data, advanced networking, or parallel software for supercomputing. Added to this, grid capabilities and middleware are a rapidly evolving set of specialisations in their own right (such as searching, authentication and authorisation).

It is important to understand that

- Expertise management will be enhanced by building groups of specialists in each of these areas rather than relying on unrelated individuals
- Different expertise and different levels of expertise are required during different stages of a communities migration toward cyber-infrastructure
- Eventually some expertise needs to be embedded in communities (eg. data curation) and some needs to be embedded in service providers (eg. network management)
- Along the way, flexible collaborative teams will need to be established and managed so that the infrastructure can evolve as the requirements are better understood

It is unlikely that each capability can or should manage this issue independently.

Also, as communities become more cyber-infrastructure oriented, they tend to co-evolve services for data generating and gathering, with services for information analysis and re-use. This happens because each community needs to develop a consensus on the standards required for inter-operation. This process inevitably requires long iterative processes, and often layers of standards (so that what is agreed at any point can be agreed), and consequently bespoke software development as people continue their research within this evolving context.

Further difficulties then arise as reliance is placed on the software and services and the need for software engineering expertise and particularly software productisation expertise becomes apparent. At present it would be fair to say that the manner in which community developed software can be most effectively moved into a long term support framework is uncertain, as is the manner in which community developed services can best migrate to service providers.

Many of these issues were evident in the report of the eResearch Co-ordinating Committee and the basic perspective of that report remains valid and is reflected here.

In summary:

- The communities most able to support their own expertise requirements are those which have been working with cyber-infrastructure the longest, in the NCRIS context this includes 5.13 and 5.10 and to lesser extent 5.12. However even in these cases, the expertise pool will need to grow.
- Additional expertise will be required to support the investments in 5.1, 5.2, 5.3 and 5.8
- As a rough estimate, if about \$10M pa NCRIS cash is likely to be devoted to this area across the current investment plans, and then allowing for co-investment, more than 50 new experts in various areas of cyber-infrastructure may need to be found

While the plan for Platforms for Collaboration has yet to be formed, addressing this issue will be vital. Some mechanisms for building on existing pools of expertise and providing arrangements where expertise can be accessed and re-assigned over time will be required. If resource pooling is not the preferred solution for managers of the various facilities, a policy approach may also be needed.

## High performance computing

Any given facility and any given researcher could have access to a wide variety of compute and storage capabilities.

These might include large systems, operated on a merit basis; mid-range systems at many research institutions, operated on a shared access basis; and a large number of usually smaller and dedicated departmental clusters and desktop systems.

Consequently the HPC needs of a research community are difficult to estimate, and the degree to which central or co-ordinated provision is relevant, is even more difficult to estimate.

### Scale

The comments on HPC made during the APAC review (which has yet to finalise) provides several observations, perhaps most importantly related to the scale of investment.

- Additional expenditure is required if the peak facility is to be retained at the historical level
- A more frequent purchase rate is needed to improve the return in Tflops delivered against dollars, under the assumption that a significant overlap in the operational periods of systems is manageable within the machine room infrastructure
- Given the already competitive nature of access, more resource overall should be provided to allow for the broader clientele envisaged under NCRIS

Clearly the latter has not emerged, although the one capability to nominate a possible demand figure (5.13) suggested a number that exceeds the capacity of APAC and APAC partner resources in total.

This situation can be directly attributed to the focussing of funding and in-kind contributions in the investment plans towards instrumentation and resources that support data acquisition, data access and material production.

The value proposition of NCRIS overall therefore supposes a balanced investment in HPC, to support theoretical and modelling work around this focus on data and material production facilities, presumably entirely from within the 5.16 funded investment plan

Greater clarity from the communities on their need for HPC would be helpful, and specifically from the facilitators, some indication is needed on expectations that the new facilities would or would not lead to increased modelling and other theoretical research that might then need support from 5.16.

### Mission

The different requirement for rate of computing (peak Tflops/s) versus the amount of research computing that can be supported (Tflop-years) is also left unclear.

While the annual average for both, and the corresponding rates per dollar spent, are all significantly improved if the frequency of purchases can be increased (as suggested during the APAC review), which of the two measures one is aiming at dominates the strategy for provision.

Overall, the implicit need in the NCRIS investments appear more likely to benefit from capacity compared to speed, however this is untested.

It would not be particularly surprising, as a peak capability requirement, by virtue of using scarce and expensive resource, can only ever be instantiated in a limited number of cases. For the same reason, those cases that do substantiate the need (such as climate modelling) have very high value.

Nevertheless, a change of emphasis through a decision to more frequently upgrade with reduced peak speed at any given purchase, may have merit and needs to be evaluated.

## High capacity networks

The NCRIS investments suggest a generalisation of the existing research network that needs to be better understood, both in terms of detailed requirements and in provisioning options.

Overall, many research needs are likely to exceed the bandwidth and/or quality of service available at reasonable cost from commodity suppliers for some time to come. Therefore the existing investment in infrastructure operated through AARNet should continue to provide significant cost benefits for services targeting research demands.

We can note overall that NCRIS investments:

- Include many examples of grid like systems providing shared access to infrastructure operating across multiple sites with a national distribution
- Includes in 5.3 and 5.10, examples of very high real time data generation rates that may require special analysis and treatment
- Represent a substantial move to develop federated views of widely dispersed data (in all but 5.4 and 5.5), which may lead to service level requirements designed to support rapid search and access
- Nearly exclusively treat communications in the operating budget, thus assuming the infrastructure investment is already in place, or will be provided by Platforms for Collaborations or through some other fund source

The generalisation of the research network to encompass the large range of government agencies and state based research institutions envisaged (see Appendix I for a list of participants in NCRIS investments) appears to lead inevitably to the use of commodity suppliers as part of the research infrastructure. The specific challenge here is management and policy, to ensure differential charging doesn't arise within a shared collaborative environment based on source and destination of traffic.

Apart from 5.3 and 5.10, each specific platform proposed under the NCRIS investments might reasonably operate over a quality of network connectivity that could be met with 1Gbps tails. To support multiple investments with that kind of connectivity, some of the Research Intensive institutions will need significantly higher aggregate connectivity.

From the table it would seem this would include (but not necessarily be limited to) the ANU, CSIRO, Monash University, University of Melbourne and University of Sydney.

At present, measured gridftp file transfers provide the following data (in MB per second):

ANU to (NatFac 28.5, UQ 25.0, Monash 22.1, VPAC 10.5, SAPAC 9.6, ac3 6.8, iVEC 6.4, JCU 0.6)

At these speeds, transferring a terabyte would take 0.4, 0.5, 1.1, 1.2, 1.7, 1.8 and 19.2 days respectively, and just over a year to move a petabyte across the ANU machine room. The results demonstrate the large gap between raw bandwidth and application level end-to-end performance, and the difficulty posed by bulk data access across shared network links.

Finally, the actual user experience in access to services operating over the research network, depends ultimately on institutional infrastructure and institutional policy. An assessment should be made to determine that the level of connectivity and access pricing required by the various platforms will be provided by the host institutions.

### 5.1 Evolving Biomolecular Platforms and Informatics

It would be best if each of the Genomics, Proteomics and Metabolomics platforms were provided with independent 1Gbps layered services to allow full flexibility in multi-site interactions.

The Bioinformatics platform is provided through co-ordinated expertise and data access within the other platforms and can leverage the network capabilities without further dedicated capacity.

## 5.2 Integrated Biological Systems

The data rates and integrative requirements for the mouse and plant phenotyping investments can be achieved without specific bandwidth enhancement.

The Atlas of Living Australia represents a very significant data federation activity, which would be best supported if the 'research network' could be extended to a large variety of institutions through subscription rather than volume pricing. Until access rates are established, it is unlikely that any bandwidth enhancement could be justified.

## 5.3 X-Ray Techniques (Synchrotron)

If remote presence is required, proposed data rates may be 10 MB/sec data, plus 2-3 video channels and two way audio, per beam line, to arbitrary sites around Australia. VICCU currently delivers remote presence within a 10MB/sec capacity.

Taking that as an upper bound, and allowing remote presence at the full set of initial beamlines, gives a bound of perhaps 200MB/s from the Synchrotron for distribution nationwide, growing over time as additional beamlines are installed and as image resolutions increase.

Current data rates achieved over the national backbone show that this is a very significant problem.

A detailed network dimensioning analysis that supports the Synchrotron data and remote presence requirements is required; especially as image resolutions scale up over the next 10 years.

## 5.3 Neutron Scattering (ANSTO)

Data rates from neutron Scattering are in the order of 2-3 TB per year. Also the length of time required for imaging (many hours) means remote monitoring is likely to be intermittent and hence less bandwidth hungry, leading to an assessment that a 1Gbps tail should support this investment.

### 5.3b Microscopy and Imaging

With tens of TB generated per year, and image generation taking hours but requiring access to remote sources for visualisation and analysis, and the desire for shared data management, these investments might also benefit from an independent 1Gbps layered service.

## 5.8 Networked Biosecurity Framework

Because this network would need to include the data, analysis systems and staff of a large number of major centres in the biosecurity area, and because the collaboration environment could be required to operate with high levels of security from time to time, the merit in a dedicated layered service or a VPN style solution will need to be evaluated during the proposed scoping study.

The resulting collaboration environment will also need an extended reach (beyond the nominated major centres) by hosting data for other components of the biosecurity network through standard web services and portal technologies.

However in the immediate future, it is likely that a pilot system can be reasonably demonstrated on existing network infrastructure, with perhaps some expenditure on providing tails to any priority institutions not already connected to aarnet3.

## 5.10 Optical and Radio Astronomy

As well as the need to ship high data volumes from telescopes to the research community, the network demand in Astronomy also includes real-time dedicated bandwidth for Very Long Baseline Interferometry involving the ATNF telescopes and remote computing systems.

The general requirement is for dedicated network at 1Gbps between all the telescope sites and key Astronomy research institutes (CSIRO ATNF, University of Melbourne, ANU, Swinburne, UWA) rising to 10Gbps rates by the 2008/09.

International connectivity at 1Gbps is required to enable international access to data products obtained with the MIRT and Australian access to data products from Gemini and other international telescopes.

A more detailed network dimensioning analysis to support the expected data rates as image resolutions scale up over the next 10 years is needed.

### **5.12 Integrated Marine Observing System**

For Ocean Observing, the observation data rates are low, hence the data gathering component presents no network issues despite the number of participating organisations. Generally, large data volumes are generated and processed at sites with access to significant compute resources and which already provide or have planned effective network access as part of their existing functions.

For Coastal Ocean Observing, the loose federation of parties serving data (maintained against their own mission requirement) and supporting public queries using agreed standards across web services does not require any specific additional network support.

The investments in remote sensing data appear to have concluded that existing network services are sufficient.

### **5.13 Structure and Evolution of the Australian Continent**

While some very large data sets are included in this area, the investment plan specifically states satisfaction with 1Gbps network links, although these need to be extended to the Geological Surveys.

The reach of the research network on an equal footing and the quality of networks and the speed of data services within organisation, once reached, have been nominated as potential key limiters.

## Appendix I – Network Reach

Organisation	1	2	3	4	5	8	10	12	13	Type	Location
GMP Mammalian Cell Facility					X						
Australian Virtual Herbarium		X								Association	Various
Antarctic Climate and Ecosystems CRC								X		Aus Gov Agency	TAS
Australian Antarctic Division								X		Aus Gov Agency	TAS
Australian Institute of Marine Science								X		Aus Gov Agency	QLD
Australian Museum								X		Aus Gov Agency	NSW
Bureau of Meteorology								X		Aus Gov Agency	MEL
Geoscience Australia								X	X	Aus Gov Agency	ACT
Grains Research & Development Corp		X								Aus Gov Agency	ACT
AAO – Siding Springs							X			Aus Gov Res Inst	NSW
ANSTO			X							Aus Gov Res Inst	NSW
Australian Animal Health Laboratory						X				Aus Gov Res Inst	VIC
Australian biological Resources Study, DEH		X								Aus Gov Res Inst	ACT
CSIRO									X	Aus Gov Res Inst	?
CSIRO ATNF							X			Aus Gov Res Inst	NSW
CSIRO Clayton					X					Aus Gov Res Inst	VIC
CSIRO Entomology		X								Aus Gov Res Inst	ACT
CSIRO Industrial Physics				X						Aus Gov Res Inst	NSW
CSIRO LI St Lucia						X				Aus Gov Res Inst	QLD
CSIRO MAR								X		Aus Gov Res Inst	TAS
CSIRO Plant Industry		X								Aus Gov Res Inst	ACT
DAFF (APPD)		X								Aus Gov Res Inst	ACT
MIRA – Mileura							X			Aus Gov Res Inst	WA
DSTO								X		Defence	?
RAN Dir of Oceanography and Meteorology								X		Defence	NSW
Australian National University	X	X	X	X			X		X	Higher Ed Inst	ANU
Curtin University of Technology						X		X	X	Higher Ed Inst	WA
Flinders University								X		Higher Ed Inst	SA
James Cook University						X		X		Higher Ed Inst	QLD
La Trobe University	X			X						Higher Ed Inst	VIC
Macquarie University	X			X	X				X	Higher Ed Inst	NSW
Monash University	X	X	X	X	X				X	Higher Ed Inst	VIC
Murdoch University	X									Higher Ed Inst	WA
Queensland University of Technology					X					Higher Ed Inst	QLD
RMIT University				X						Higher Ed Inst	VIC
Southern Cross University		X								Higher Ed Inst	NSW
Southern Cross University (Lismore)	X									Higher Ed Inst	NSW
Swinburne University				X					X	Higher Ed Inst	VIC
University of Adelaide	X	X	X			X			X	Higher Ed Inst	SA
University of Melbourne	X	X	X	X		X			X	Higher Ed Inst	VIC
University of New South Wales	X		X	X	X					Higher Ed Inst	NSW
University of Newcastle	X			X						Higher Ed Inst	NSW
University of Queensland	X		X	X	X				X	Higher Ed Inst	QLD
University of South Australia				X						Higher Ed Inst	SA
University of Sydney	X		X	X	X	X		X	X	Higher Ed Inst	NSW
University of Tasmania		X						X	X	Higher Ed Inst	TAS
University of Western Australia	X		X					X	X	Higher Ed Inst	WA
University of Western Sydney			X							Higher Ed Inst	NSW
University of Wollongong				X						Higher Ed Inst	NSW
Gemini – Cerro Pach, Chile							X			Int Org	Chile
Gemini – Mauna Kea, Hawaii							X			Int Org	Hawaii
NASA									X	Int Org	USA
PILOT – Concordia, Antarctica							X			Int Org	Antarctica
Scripps Institute of Oceanography								X		Int Org	Internet
Marine National Facility								X		Nat Facility	Mobile
Animal Health Australia						X				Not for Profit	ACT
Queensland Cyber Infrastructure Foundation								X		Not For Profit	QLD
Sydney Harbour Institute of Marine Science								X		Not for Profit Assoc	NSW
AuScope Limited									X	Priv Res Org	

Australian National Fabrication Facility Limited				X						Priv Res Org	VIC
Australian Synchrotron Research Program Inc			X							Priv Res Org	VIC
Bandwidth Foundry P/L (MNRFP)				X						Priv Res Org	NSW?
MiniFab (Aust) P/L				X						Priv Res Org	VIC?
Queensland Institute of Medical Research		X								Priv Res Org	QLD
Walter & Eliza Hall Institute	X	X								Priv Res Org	VIC
Bresagen Ltd					X					Pty Ltd	SA
Progen Ltd					X					Pty Ltd	QLD
Radpharm Ltd					X					Pty Ltd	ACT
TGR Biosciences	X									Pty Ltd	SA
Australian Synchrotron			X							Pty Ltd?	VIC
Australian Red Cross Blood Services					X					Society	VIC
Royal Perth Hospital					X					State Gov Agency	WA
Royal Prince Alfred Hospital					X					State Gov Agency	NSW
Dep of Agriculture and Food						X				State/Ter Agency	WA
Dep of Economic Development								X		State/Ter Agency	TAS
Dep of Environment and Conservation								X		State/Ter Agency	NSW
Dep of Environment and Heritage									X	State/Ter Agency	SA
Dep of Further Ed, Employment, Sci and Tech				X						State/Ter Agency	SA
Dep of Innovation, Ind & Reg Development				X						State/Ter Agency	VIC
Dep of Land Information								X		State/Ter Agency	WA
Dep of Lands								X		State/Ter Agency	NSW
Dep of Primary Industries					X		X			State/Ter Agency	NSW
Dep of Primary Industries					X					State/Ter Agency	VIC
Dep of Primary Industries and Water								X		State/Ter Agency	TAS
Dep of Primary Industries, Fisheries & Mines					X			X		State/Ter Agency	NT
Dep of Primary Industry and Resources								X		State/Ter Agency	SA
Dep of State Development, Trade, Innovation				X			X			State/Ter Agency	QLD
Dep of Sustainability and Environment								X		State/Ter Agency	VIC
Dep Science and Innovation		X								State/Ter Agency	SA
Geological Survey of NSW								X		State/Ter Agency	NSW
Geological Survey of Western Australia								X		State/Ter Agency	WA
Geological Survey Queensland								X		State/Ter Agency	QLD
Geoscience Victoria								X		State/Ter Agency	VIC
Mineral Resources Tasmania								X		State/Ter Agency	TAS
Northern Territory Geological Survey								X		State/Ter Agency	NT
Office of Scientific and Medical Research				X						State/Ter Agency	NSW
Planning and Land Authority								X		State/Ter Agency	ACT
Queensland Natural Resource Management								X		State/Ter Agency	QLD
South Australia R&D Institute (SARDI)					X		X			State/Ter Agency	SA
Sydney Water and Manly Hydraulics Lab								X		State/Ter Agency	NSW
Tasmanian Partnership for Advanced Computing								X		State/Ter Agency	TAS
Animal resources Centre		X								State/Ter Res Inst	WA
Australian Museum		X								State/Ter Res Inst	NSW
Garvan Institute, St Vincent's Hospital, Sydney	X									State/Ter Res Inst	NSW
Institute for Medical & Veterinary Sciences		X								State/Ter Res Inst	SA
Institute for Medical and Veterinary Science					X					State/Ter Res Inst	SA
Museum Victoria		X								State/Ter Res Inst	VIC
PathWest Laboratory Medicine						X				State/Ter Res Inst	WA
Peter MacCallum Cancer Institute					X					State/Ter Res Inst	VIC
Queensland Institute for Medical Research					X					State/Ter Res Inst	QLD
Queensland Museum		X								State/Ter Res Inst	QLD
Royal Brisbane Hospital (QIMR)	X									State/Ter Res Inst	QLD
Tasmanian Museum & Art Gallery		X								State/Ter Res Inst	
Victorian Infectious Diseases Reference Lab						X				State/Ter Res Inst	VIC
Westmead Hospital?						X				State/Ter Res Inst	NSW