AeRO Forum: Service Provider Perspectives
A perspective from NCI

Lindsay Botten
Director, NCI

May 2017
## Mission: World-class, high-end computing services for Australian research & innovation

### What is NCI:
- Highly-integrated, e-infrastructure environment: joint HPC/HPD focus
- Comprehensive, integrated and expert service
- Supercomputer + supercomputer-class cloud + highest-performance storage + internationally-renowned expert support team

### National, strategic and values-driven:
- Enabling high-impact research—informing policy, delivering social/env./econ. benefits
- Research- & outcome-driven—serving national priorities and research of excellence
- Designed by deep engagement, collaborations and industry
- Delivering transformative outcomes and national benefits
- Quality and innovation — through scale, experience and expertise
- Value demonstrated through growing co-investment, uptake and impact

### Profile:
- A capability beyond the capacity of any single institution to provide
- Serves: 35x universities, 5x science agencies, 8xNCRIS; 3 MRIs; industry
- ~4,000 users; ~600 projects
- 500+ papers pa; ~200 ARC/NHMRC funded activities—$58M pa
- Capital (NCRIS): $60M infrastructure + $26M data centre building
- Recurrent Costs ($18+ M) — 60 staff; $3.3 M utilities; etc.

### Sustained by:
- Collaboration: agencies/universities/ARC (~$13.4 M p.a.)
- NCRIS (~$5.5 M p.a.)
NCI at a glance in 2017: investment and co-investment

Cash Investments/Co-investment and Expenses

Capital (NCRIS)

Data Centre $26M
HPC and Storage Infrastructure $60M

Recurrent (p.a.)

NCRIS $5.5M
Partners $13.4M

NCRIS investments since 2008:
NCI: $92.5M
Other: ~$8.1M

Partner Co-investment 2008-17: $71.6M
Supercomputer Raijin — *Australia’s highest performance research supercomputer*

- Fujitsu: 1.20 petaflops, 57,492 Xeon/SB cores, 160 TB memory, 10 PB filesystem, FDR IB backplane
- Lenovo: 0.94 petaflops, 22,792 Xeon/BW cores, 144 TB memory, EDR backplane
- Plus GPUs + KNL + IBM Power — small test environments for the future

**HPC Cloud** (NeCTAR & Tenjin/private): 3,200 cores, supercomputer spec. for orchestrating data services

**Global integrated storage** (*highest performance filesystems in Australia*)

- 22 PB (actual) (rising to 36 PB, June 2017)—up to 120 GB/sec bandwidth; 40 PB of tape archive
Service Portfolio

• Services and Technologies portfolio
  – Generalised (high-performance) to meet all needs
  – Comprehensive, integrated compute/storage infrastructure
  – Expert environment tailors for specialised needs

• Research Engagement and Initiatives portfolio ("special sauce" / "glue")
  – Evolved some general capabilities from meeting specialised requirements
  – Driven by research and innovation needs of NCRIS, partner organisations, MRIs, industry
  – Danger in overreaching to generalised solutions rather than attaining valued, specific goals

Copernicus Sentinel Sentinel Regional Hub
Sustainability (NCI) is built on:

- **Sense of shared responsibility**
  - Commonwealth for the major capital; user organisations for the bulk of recurrent funding
- **Demonstrating indispensability** — critical to national capability/research competitiveness
  - Essential in arguing case of for ongoing capital investment — references in the draft RI Roadmap
- **Reputation**
  - Quality/innovative services → impact/excellence, national/inst. priorities, program-scale reach
  - Excellence of expertise; quality of infrastructure (integrated) and operations
  - Services delivered more cost-effectively and better than in-house provision

Confidence comes from:

- **Being run as a business**
  - Service oriented — one-stop shop “your problem is our problem”
    - Quality of service (staff with “business owner” mentality) — drives new and repeat business
  - Driving income/revenue generation to up-scale service and value
- **Strong, “skin-in-the-game” governance** — that cares about the outcomes
- **Major science agencies as partners** — provides confidence to others
- **Crucial support from government** — Agility Fund was a major confidence shot-in-the-arm
Forum issue: Key eResearch learnings of the past ten years

• Importance of Research-driven Planning
  – Services designed through engagement — communities, institutions
  – National science/research priorities; support for research of project-scale and program-scale

• Importance of service integration vs fragmentation

• Leadership — respected governance/management
  – Direction setting, service delivery, sustainable business model

• Aggregated Infrastructure and Distributed/Aggregated Expertise
  – Infrastructure: greater resilience/robustness, with greater efficiency and diversity
  – Expertise: distributed for broad support; concentrate for depth supporting transdiscipl. research

• Holistic and Balanced Investments
  – Hard infrastructure investments must be balanced with soft infrastructure investments
  – Data methods need to be balanced with computational techniques

• National vs Institutional Conceptualisation of the eResearch landscape
  – Differentiation of national and institutional responsibilities; avoidance of cost shifting
  – Tier 1 vs tier 2 obligations, role of the cloud now and into the future

• Service Focus
  – Facility/service businesses brings a service focus/discipline — engagement with users/institutions
  – “Funding projects” lack the same service focus — indirection and reduced focus
ARDC objectives comprise four components:
- World leading data advantage; Acceleration of innovation; Enhanced translation of Research; Collaboration for borderless research
- NCI endorses their substance, emphasising critical role of HPC in high-end data services (Roadmap)

But is there an overarching goal to drive the vision?
- More than an ARDC as the national custodian/funder of data access/services in isolation

Goal: Advancement of research and innovation of impact/excellence for the benefit of society in a world in which the grand challenges are increasing transdisciplinary (?)

Realisation of the goal requires:
- Harmonised, collaborative service arrangements between ARDC, HPC, institutional capacity, AARNet/AAF, other infrastructure and eResearch providers
  - Converged expert services to support data-driven/dependent and computational research
  - Robust infrastructure platforms (HPC, cloud, storage) and networks
  - Curatorial services framework (data management/provenance) underpinned by FAIR principles
  - National Computational and Data Sciences Capability (presently missing, cf. US/UK national labs)
    • High-end, expert focus in contrast to just isolated, localised and duplicated skills
Forum Question: ARDC Modus Operandi

• ARDC service provision: to be more than the sum of ANDS/NeCTAR/RDS parts
  – Parallel with EU Open Science Cloud is “cloudy” and needs articulation: more than data sharing?
  – National data science expertise development, peak computing collaboration, ?

• Scale/sources of funding ↔ operational model, governance and advisory structures
  – Lightweight model
    • Modest funding → most “data” funds to NCRIS capabilities which commission services
    • Focus is set naturally by NCRIS capabilities
    • But the danger of inefficient, subscale, unambitious, and unaligned outcomes?
  – Heavyweight model
    • How to achieve more with the same/less funding — pre-ARDC is $21 million p.a.
    • Large (central) funding requires a strong vision, strong governance, robust plan, engagement
      – Focus set by domains/NSRPs, excellence/impact dimensions
    • Business model:
      – What is to be funded and by whom?
      – What is to be provided at no cost, subsidised, or charged for?
      – Efficient/effective use of money, avoidance of cost shifting (institutional obligations), attraction of cash co-investment, access/service fees (from institutions) rather than free
  • Setting the balance of program-scale (national) and project-scale (institutional) needs
    – e.g., national/international reference collections vs project datasets
Business model to reflect a world different to that in which ANDS/NeCTAR/RDS originally conceived

• Environment and infrastructure
  – Pervasiveness of commercial providers / cloud
    • Incl. AARNet — fee-for-service storage
  – HPC facilities delivering high-end data services
  – Fee-for-service, subscriptions now common

• Services expected/required
  – Repository (high-level and basic), storage, compute attached to data
  – Value-add expected: more than just data access data
    • Advanced analytics, AI/machine learning, interoperability required
  – Implications for infrastructure & service provision

• Unwinding legacy decisions?
  – Greater infrastructure concentration or comm. cloud
  – Distribution basic skills, but with domain focus
  – Concentration of high-level data science capabilities
Forum Questions: Streamlined access to data/tools from many providers

• Yesterday: moving data, establishing replicas of datasets — with all the problems this brings

• Today, it is about:
  – Using data in situ — ensuring provenance, providing comprehensive data interoperability
    • Virtual desktops — accessing the compute and tools at the location of the data
    • Directly accessing data (files) on Raijin

  – Using data remotely
    • Data services (THREDDS, others) — maintaining provenance

• Access also via the catalogue and DAP portal — but with loss of the provenance chain
Thank you

Questions