Economics Legislation Committee
ANSWERS TO QUESTIONS ON NOTICE
Industry, Innovation and Science Portfolio
2016-17 Supplementary Budget Estimates
20 October 2016

DEPARTMENT: DEPARTMENT OF INDUSTRY, INNOVATION AND SCIENCE

TOPIC: SKA and Big Data

REFERENCE: Written Question – Senator Carr

QUESTION No.: SI-68

1. Is one of the major challenges facing the SKA project the storage and processing of data?
2. Astronomers have estimated that the project will generate 35,000-DVDs-worth of data every second. This is equivalent to "the whole world wide web every day". Is that an accurate summary of the data challenge?
3. How much of the required processing capacity will the Pawsey centre provide? Where do you plan on doing the rest?
4. Are there possibilities in the UNSW led quantum computing project that could inform the SKA project's data challenges?
5. How will the data be moved? Will it involve the National Broadband Network or AARNET?

ANSWER

1. Yes. Although the SKA will be one of the world's most data-intensive projects, engineers expect that computers and storage systems of the necessary size and capability will be available at the time the SKA becomes operational. The systems involved will be amongst the largest in the world and cost in the region of several hundred million dollars.

This means the cost and computational/data storage burden will need to be shared and managed in a hierarchical manner (i.e. some processing will be done centrally by the SKA Observatory and some will be done in a distributed network using other facilities available to the member states). Similar systems have been developed for projects like the Large Hadron Collider at CERN.

As well as being large in capacity terms, the SKA's data system will be particularly complex. Developing the automated and intelligent systems to manage SKA-class data and processing is a major and potentially rewarding challenge that is likely to have significant spillover benefits to other data using sectors. SKA Pre-construction consortia are currently working on developing these systems.

2. The raw data flow from the SKA Phase 1 antennas/receivers will be extremely large – around 3 Petabits/second – which is about 10 Zettabytes/year. For comparison the internet in 2016 is about 1 Zettabyte/year. In terms of 'DVDs-worth of data', this raw data flow is approximately equivalent to 100,000 DVDs-per-second. It will possibly be the largest single source of raw data in human history.

It should be noted that not all the raw data will be stored or processed into 'science products'. The actual science data flow from SKA1 will be about 130 Petabytes/year. That's 0.0001 Zettabytes/year. The science data stored by SKA1 corresponds to filling 1 DVD per second.

- *Note that comparisons between SKA and the internet vary depending on when they are made (while also taking into account that the internet is growing rapidly), whether the full SKA concept or Phase 1 is being referred to and whether raw data or science data flow is being compared. The above comparisons are accurate in relation to the current SKA Phase 1 design and the current size of the internet.*

3. The Pawsey Centre currently has a processing capacity of approximately 1 Petaflop and a storage capacity in the order of 50 Petabytes. This is well matched to the needs of the Murchison Widefield Array and Australian SKA Pathfinder. SKA Phase 1 in Australia will need around 100 times the processing power and between 100-200 Petabytes of storage per year.

As noted above, the current plans are for SKA1 to use a distributed data architecture. There will be a significant element of central processing by the SKA Observatory of raw data into science products, along a processing chain that will commence at the telescope site and finish at the Pawsey Centre. These science products will undergo further processing outside the SKA Observatory at processing centres in Australia and internationally. However exactly what quantity of processing will be done where is yet to be determined in detail.

4. Quantum computing deals with relatively small data sets but with large computational requirements. As such, it is not necessarily immediately relevant to SKA challenges. However the development of intelligent computing systems to handle complex work flows is relevant to SKA and there may be some opportunity for quantum computing in that space.

5. Data will be moved from the MRO to Perth by a combination of dedicated and national network infrastructure. The resources of both the NBN (namely the NBN backbone from Geraldton to Perth) and AARNET (which operates the CSIRO link from the MRO to Geraldton) are involved. From the processing centres the data will flow globally at around 100 Gigabytes/second (continuous) using again a combination of international and national networks including AARNET.