FROM ANTENNAS TO MULTI-DIMENSIONAL DATA CUBES: THE SKA DATA PATH

Andreas Wicenec

University of Western Australia
ICRAR
35 Stirling Hwy., Crawley
6009 Western Australia

ABSTRACT

The SKA baseline design defines three independent radio antenna arrays producing vast amounts of data. In order to arrive at, still big, but more manageable data volumes and rates, the information will be processed on-line to arrive at science ready products. This requires a direct network interface between the correlators and dedicated world-class HPC facilities. Due to the remoteness of the two SKA sites, power as well as the availability of maintenance staff will be but two of the limiting factors for the operation of the arrays. Thus the baseline design keeps just the actual core signal processing close to the center of the arrays, the on-line HPC data reduction will be located in Perth and Cape Town, respectively. This paper presents an outline of the complete digital data path starting at the digitiser outputs and ending in the data dissemination and science post-processing, with a focus on the data management aspects within the Science Data Processor (SDP) element, responsible for the post-correlator signal processing and data reduction.

Index Terms— Radio Astronomy, data management, data processing, Square Kilometre Array

1. INTRODUCTION

The SKA will consist of a low-frequency aperture array with up to 256,000 dipoles, a mid-frequency dish array equipped with phased array feeds and a mid-frequency dish array equipped with cooled high sensitivity single pixel feeds. The first two will be located in Western Australia, the latter array in South Africa. The following section describes the SKA data flow, starting with a very brief outline of the front-end part, which will be located close to the centre of the arrays. More detailed description of this part is outside the scope of this paper and is given elsewhere.

Thanks to University of Western Australia, Curtin University and Western Australian Government for funding of ICRAR.

2. SKA DATA FLOW

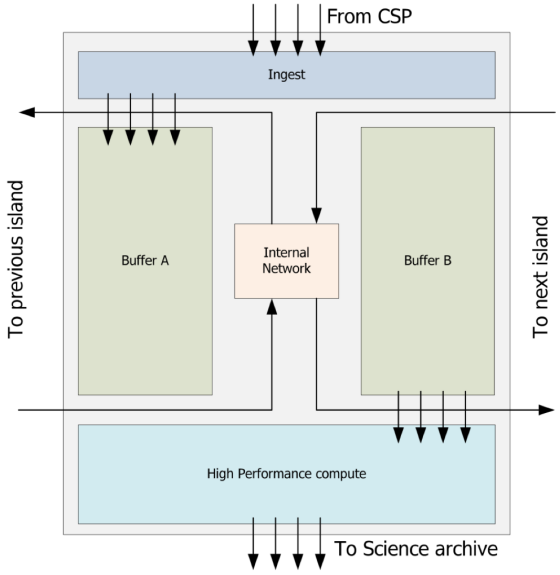
2.1. Digitiser to Correlator

The digitisers will either be located very close to the receivers, or in the central building on the two sites. This depends on whether RF over fibre technology turns out to be a viable option or not. The data-rate after the digitisers is of the order of 10 Gbit/s for the SKA-low stations (1 beam), up to 90 Gbit/s per antenna for SKA-mid and 2.3 TBit/s per antenna for the SKA-survey PAFs and this data will be fed into beamformers and filter-banks and from there into the central correlator. There will be a correlator for each of the arrays. This part of the signal processing is collectively called the Central Signal Processor (CSP).

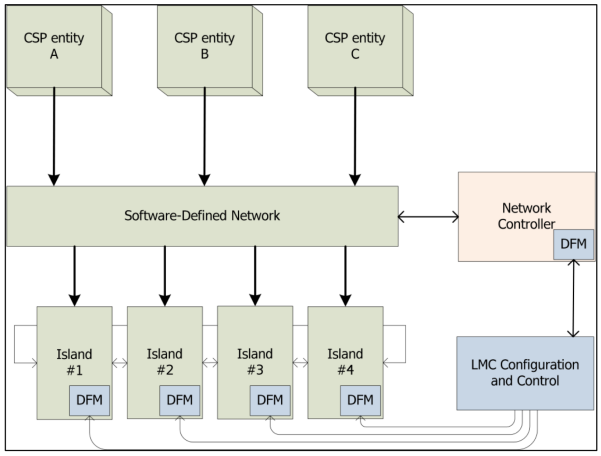
2.2. Correlator to HPC Centre

The signals are then correlated, i.e. cross-multiplied and fourier transformed. The data rate after the correlator is still very high at around 7.3 TByte/s for SKA-low, 3.3 TByte/s for SKA-mid and 4.6 TByte/s for SKA-Survey. This amount of data, plus some 10% of additional meta-data has to be transported from the observatory sites to the HPC facilities for subsequent calibration and imaging by the Science Data Processor (SDP). Most of the SDP processing has to be carried out on a per-observation basis and currently some of the available algorithms require multiple iterations over the whole data set of an observation and thus the data has to be buffered. Typically observations are taken in many hour time intervals and thus the buffer sizes are very big (100 PBytes/array). Moreover, in order to be able to carry on with the next observation while the previous one is being processed, the current SDP design foresees a double buffer scheme (see Figure 1a). The performance requirements for the buffer are quite extreme as well, as can be seen from the input data rates above and the fact that data will be read and written concurrently during the processing. Additional complexity is introduced by the requirements to run up to 16 observations in parallel on sub-arrays of each of the arrays and to be able to reduce the same data in a commensal way

Figure 1



(a) Schematic data flow within a compute island. The double buffering scheme is shown processing from buffer B while buffer A is being used to ingest data from CSP.



(b) Relation between Local Monitor and Control (LMC), the Data Flow Manager (DFM) and the software defined network connecting multiple CSP entities with SDP compute islands.

using different reduction pipelines.

2.3. Internal to HPC

The data from the correlator arrives in a certain order, which is dictated by the data collection. The processing on the other side requires a complete re-ordering of the data, a so-called corner turning. The current SDP design assumes that this corner turning can be performed using software defined networking (see Figure 1b), i.e. essentially the network switch on the input side to the SDP will distribute the data packets in such a way that the individual compute islands receive their assigned chunks in the correct order. The concept of the compute islands has been introduced in order to provide both failure isolation and scalability for the HPC implementation. The current straw man design of the compute islands defines a rack containing 56 nodes with GPGPU accelerators and Infiniband low-latency interconnect. External networking is provided by multiple 40 Gbit ethernet connections. This straw man is based on current day technology and is mainly used for the development of the SDP cost-model. The fast evolution of compute hardware makes it very hard to predict which technologies would be the most viable and economic at the time when the SKA has to procure the compute hardware. The whole observatory, including the SDP will be controlled by the Telescope Monitor and Control (TMC) subsystem. The design of the TMC is hierarchical in the sense that each of the operational subsystems implement one or more Local Monitor and Control (LMC) instances, which in turn implement the actual interface to the TMC. The LMC is comprised of the following major functional elements:

- *Control (CTRL):* Controls the SDP capabilities including availability and schedule planning and exposes this control to the TM.
- *Data Flow Manager (DFM):* Manages the construction of the internal execution graphs.
- *Health and Monitoring (HM):* Performs health, alarms and status monitoring and provides this information to the TM and internal elements as required.
- *Quality Assurance (QA):* Provides aggregate quality assurance metrics to the end user to allow verification of the quality of science data as it is processed through the SDP.

The SDP LMC will provide the Data Flow Manager (DFM) which essentially implements a configuration management interface between the scheduled observation and the available hardware. The DFM functionality is based on graph based scheduling systems (see [1], [2], [3]). The DFM will instantiate Data Manager modules on the compute islands and provide them with the information about the actual observation and planned processing. The Data Managers in turn will instantiate data objects which will provide the required interfaces (e.g. sockets) to accept the data from their respective parent data objects. The data objects will trigger their assigned consumer modules, which represent the actual processing. In this way the processing is designed to be fully

data driven, i.e. the data managers and the data objects themselves are aware of the processing modules to be launched next in the processing pipeline. The information about the required data objects and their associated processing modules originates from a static processing graph which has to be developed for each of the processing pipelines and their individual modes. These graphs will then be scaled to the actual observation requirements and then mapped onto the available hardware. During the mid-term scheduling phase of an observation the DFM will check for the availability of the required resources and, if available, provisionally accept the execution of the observation. Just before the actual execution a similar check will be carried out and the resources will be allocated.

Processing SKA observations will require several hundred Petaflops of computing, distributed across the three telescopes. The actual amount is both dependent on the science goals and the processing modes and algorithms. The algorithms are still under investigation and while some require less FLOPs they will require far more memory and/or I/O and might thus be far less efficient. Also the overall and specific efficiency of the pipelines and the individual modules is being studied in terms of actual FLOPs achieved, FLOPS/Watt and Amdahl number (FLOPs/byte). The latter is quite important, since quite a number of the key algorithms used in radio astronomy are not very compute intense, i.e. they only require between 10 and 200 FLOPs per byte of I/O. Current day supercomputers and HPC systems are optimised towards far higher compute intensities [4].

2.4. HPC to Science Archive

The concept of a Science Archive is a quite traditional view on how data products are managed and presented to the outside world in some sort of a back-end, detached archive system. The SDP architecture revisits this concept and attempts to generalise the persistent storage requirements from the ingest, processing and archiving stages into a single design, allowing to control the life-cycles of all persistent data objects in a homogenous way. The Science Archive storage in this design is just an integral part of the overall storage hierarchy featuring certain reliability and data safety specifications, but also enabling access to data from the outside world. It is mainly the visibility to scientific access mechanisms, like the IVOA standard protocols [5], and the meta-data models describing the data objects, which promotes a particular data objects to be part of the Science Archive. In fact data objects can be both promoted and demoted from the Science Archive for instance to cover cases when they become obsolete due to the availability of higher quality products. Moreover, due to the expected growth rate of the Science Archive (hundreds of PB/year) it is even conceivable that the operational policy will put a cost cap on the operations of the Science Archive, which in turn translates to a finite total size and thus would require to retire data products from the archive. Technical and economic considerations will also require to constantly migrate data from one technology or platform to the next and thus promoting the data life cycle management of archived data objects into a similar regime as for shorter term storage tiers. With this kind of overall dat management the actual physical ingest into the Science Archive is limited to generating a number of meta data records in the Science Archive database. The actual data objects will at the same time enter the transition to a higher persistence state and then the data management system will ensure that the appropriate actions are being taken to ensure that the requirements for that persistence state are satisfied. In practice this may mean to create another copy in some remote location, or to create a copy on an off-line long-term medium. The design of the system is as far as possible technology and policy neutral, in order to allow for frequent and continuous changes and migration needs.

2.5. Science Archive to Data Centers

The data products produced by the standard pipelines will be 'science ready'. This means that in general these are the only data products scientists will have access to. This is quite a paradigm shift in radio astronomy, since typically the scientists nowadays are performing their own data reductions, using far less advanced products. With the amounts and rates of data produced by the SKA, this approach is not sustainable. Even the highly reduced data products of the SKA will pose significant challenges to the science teams when it comes to scientific analysis. Even for smaller scale projects, the usage of HPC type facilities will be required for the analysis. Typically this will mean that data products will not be transferred to individuals, but rather to data centres, which are able to store and process the data. Thus the scientists will potentially need to apply for such resources and, once granted, the SKA Science Archive can then organise the delivery process. For the big surveys it is quite likely that a few dedicated data centres will take over this role and the SKA Science Archive together with those data centres will optimise the data transfer. The details of such arrangements as well as possible alternative solutions are currently under discussion and evaluation. Alternative solutions also include public cloud resources, which might indeed be quite attractive and even economic at least for the post-processing stage.

2.6. Science Archive to Scientists

As outlined above, the traditional data delivery of standard SKA products directly to scientists will be the exception rather than the rule for SKA data, at least for the foreseeable future. However, some of the products as well as some more advanced services might still enable scientists to directly access small regions of interest around their favourite astronomical objects or catalogs of such objects, which will also be

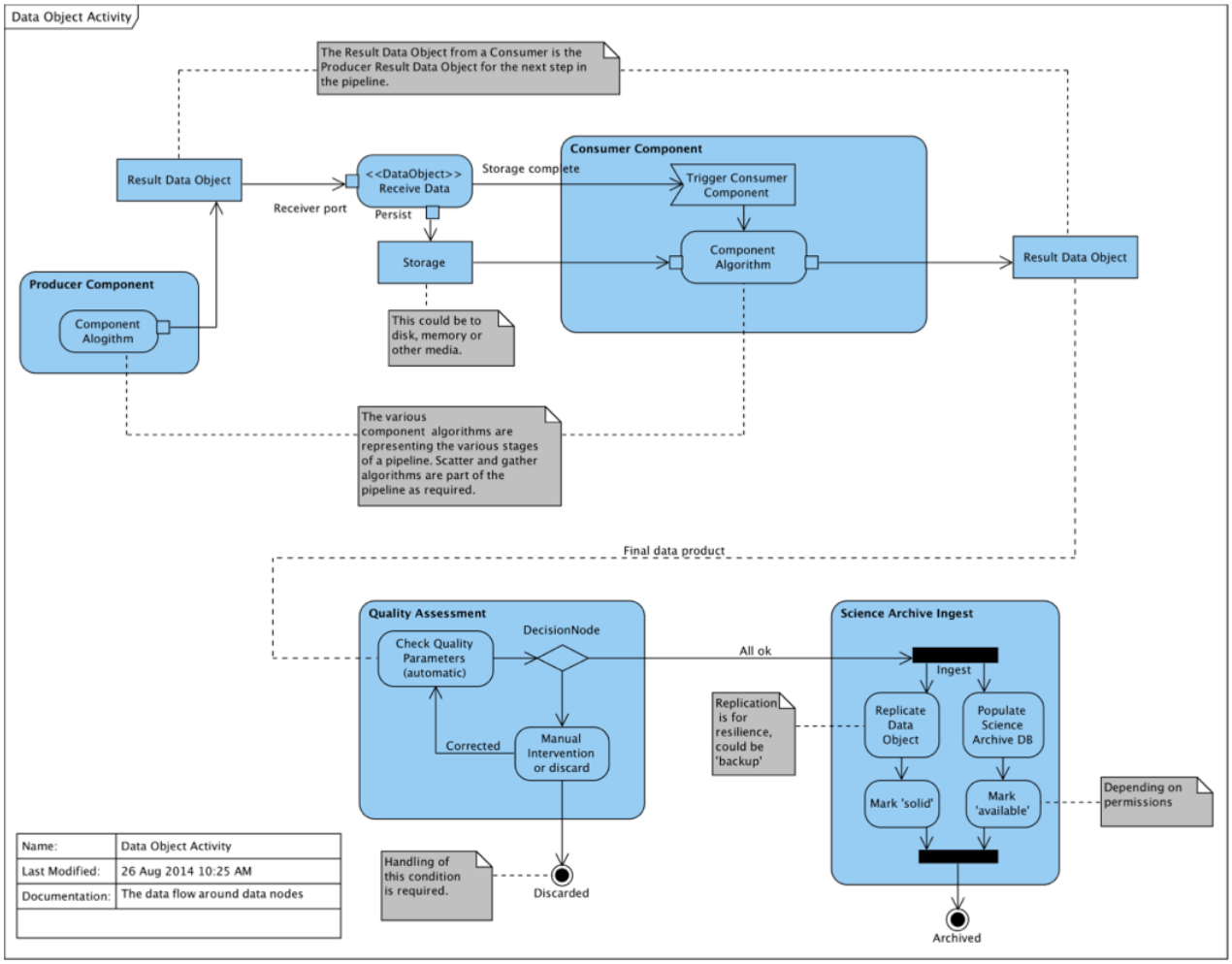


Figure 2: Data driven processing, quality assessment and science archive interaction. The data objects are active components, receive the data and trigger the processing appropriate to them to reach the next data product level. Once the final product for a standard processing pipeline is reached, there will be a quality assessment stage and, if successful, the product will be elevated to the archive state. Products in that state will be searchable and accessible to external scientists, subject to data distribution policies, defined by the SKA organisation.

produced as part of the standard processing. There are also investigations on-going to use advanced, multi-resolution, lossy and lossless compression data formats, such as JPEG2000 [6] or remote data visualisation techniques [7], [8] to enable more seamless interaction with the vast SKA data space.

3. CONCLUSION

The SKA processing and data handling imposes new challenges on both the computer engineers and scientists as well as radio astronomers. The amount and scale of the SKA data problem truly pushes the boundaries of the currently existing compute environments and economics. Very likely, SKA operations will be limited by the available power budget, both for the actual telescopes at the remote sites, but even more for the computing, networking and storage. In order to make full use of the SKA, potential computer technologies like processors, storage and network have to provide 'more for a lot less': More processing for less watts, more I/O for less FLOP, more bandwidth for less connections or in general more for less money. In addition astronomers have to learn how to work with all remote data access and processing and far less data transfer. As a short conclusion one could say that the full-scale SKA compute infrastructure is technically achievable, but currently not affordable. However, even if we will not be able to fully utilise all the potential of the SKA receiver deployment, still the SKA will enable unique and ground-breaking science and the next generation of computer technology will unleash even more.

4. REFERENCES

[1] Yu-kwong Kwok and Ishfaq Ahmad, "Static Scheduling Algorithms for Allocating Directed Task Graphs to Multiprocessors AND," *ACM Computing Surveys*, vol. 31, no. 4, pp. 406–471, 2000.

[2] Josef Weinbub, Karl Rupp, and Siegfried Selberherr, "A lightweight task graph scheduler for distributed high-performance scientific computing," in *Applied Parallel and Scientific Computing*, Pekka Manninen and Per ster, Eds., vol. 7782 of *Lecture Notes in Computer Science*, pp. 563–566. Springer Berlin Heidelberg, 2013.

[3] Manar Qamhieh, Frédéric Fauberteau, Laurent George, and Serge Midonnet, "Global EDF scheduling of directed acyclic graphs on multiprocessor systems," *Proceedings of the 21st International conference on Real-Time Networks and Systems - RTNS '13*, p. 287, 2013.

[4] Alexander S. Szalay, Gordon C. Bell, H. Howie Huang, Andreas Terzis, and Alainna White, "Low-power amdahl-balanced blades for data intensive computing," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 1, pp. 71–75, Mar. 2010.

[5] "IVOA Standards," http://www.ivoa.net/documents/, 2014.

[6] V.V. Kitaeff, A. Cannon, A. Wicenec, and D. Taubman, "Astronomical imagery: Considerations for a contemporary approach with JPEG2000," *Astronomy and Computing*, , no. 0, pp. –, 2014.

[7] Cameron Kiddle, A. R. Taylor, Jim Cordes, Olivier Eymere, Victoria Kaspi, Dan Pigat, Erik Rosolowsky, Ingrid Stairs, and A. G. Willis, "Cyberska: An on-line collaborative portal for data-intensive radio astronomy," in *Proceedings of the 2011 ACM Workshop on Gateway Computing Environments*, New York, NY, USA, 2011, GCE '11, pp. 65–72, ACM.

[8] A.H. Hassan, C.J. Fluke, and D.G. Barnes, "Interactive visualization of the largest radioastronomy cubes," *New Astronomy*, vol. 16, no. 2, pp. 100 – 109, 2011.