SDP Memo: Regional Centres

Document number SKA-TEL-SDP-0000060
Document type ..... REP
Revision 01C
Author ..... R. Simmonds and the DELIV team
Release date 2016-04-07
Document classification ..... Unrestricted
Status ..... Draft

Lead author:

Name	Designation	Affiliation
Rob Simmonds	SDP.DELIV Lead	University of Cape Town
Signature & Date:	RWJ Simmonds (Apr 7, 2016) robert.simmonds@uct.ac.za	

Released by:

Name	Designation	Affiliation
Paul Alexander	SDP Project Lead	University of Cambridge
Signature & Date:	Paul Alexander (Apr 8, 2016) pa@mrao.cam.ac.uk	

Version	**Date of issue**	**Prepared by**	**Comments**
1.0	2015-02.09	Verity Allan	Version submitted to PDR panel
1A	2015-05-29	Rob Simmonds	Updates based on comments from PDR panel
01C	2016-04-07	Rob Simmonds	Released for ΔPDR

ORGANISATION DETAILS

Name	Science Data Processor Consortium

Table of Contents

List of Abbreviations

ALMA	Atacama Large Millimeter/submillimeter Array
ARC	ALMA Regional Centre
LHC	Large Hadron Collider
PI	Principal Investigator
PoP	Point of Presence
RC	Regional Centre
SDP	Science Data Processor
SKA	Square Kilometre Array
SLA	Service Level Agreement
WAN	Wide Area Network
WLCG	Worldwide LHC Computing Grid

Summary

This document is intended to suggest ways in which geographically distributed Regional Centres may be used to support SKA activities. At the time of writing, the SKA Data Flow Advisory Panel was preparing a report, which will contain complementary information about Regional Centres.

Scope of this document: This document only deals with issues related to Regional Centres (RCs). RCs are not funded by the SKA project.

Assumptions made in this document: This document discusses ways that the RCs could be used, but makes no assumptions about where funding to support the centres would come from.

Layout of the document: This document starts with an overview of the possible uses for RCs. It then has sections providing more detail on potential processing and storage capabilities, use in optimising international WAN links, and as off-site backup facilities. From there the use of RCs for regional outreach and support is discussed, and the need for agreements between RCs and the SKA is presented. The final section gives a brief summary of the document.

Applicable and Reference Documents

Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, *the applicable documents* shall take precedence.

Reference Number	Reference
AD01	SKA-TEL-SDP-0000013 – SDP Element Architecture Design

Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, *this document* shall take precedence.

Reference Number	Reference
RD01	SKA-TEL-SDP-0000075 – SDP Standard Data Products

1 Introduction

The SDP is responsible for the processing of observed data into science-ready data products, referred to as Science Products. It is also responsible for the long term preservation of these Science Products and the delivery of these to trusted external sites [AD01]. The scientific analysis of these Science Products requires further processing which is out of the scope of the SDP. Therefore there is a need for external processing centres, able to receive and host SKA Science Products and facilitate astronomers' access to and ability to analyse these Science Products. These external processing centres are referred to as Regional Centres (RCs) in this document.

Although not an SDP funded area and currently not included within the central SKA funding model, the RCs will play a key role in the scientific exploration of the SKA data. Therefore, although the SDP architecture design does not include the whole software stack for these centres, it does provide ways of moving Science Products to them and for replication of the Science Catalogue to them.

2 Role of Regional Centres

We envision that SKA Regional Centres (RCs) will fulfil a number of roles within the SKA project. They will facilitate easy and fast access to data, computing and e-science tools for the science community, both for globally distributed major survey project consortia and for individual science projects. They will support common collaboration infrastructure and middleware tools to provide uniform access to data and computing resources, independent of location.

The RCs will provide an international network of centres of excellence for data science and provide regional points of contact and outreach for SKA activities. Using the presence of SKA data and research within a country with an active user community will enable hosting organisations to make a business case for funding enhanced resources. SKA has been used to support resource investments in several countries already. It is not thought that the RCs will be owned by the SKA project itself.

Initial processing of SKA data at or near the telescope sites will generate Science Products that will be stored at the SDP sites located in South Africa and Australia. The best model for utilising global resources may vary from one SKA science project to another, but providing common middleware at RCs will enable these projects to use resources at different centres as appropriate. Also utilising the network connections at a subset of RCs would help when optimising the use of the Wide Area Network (WAN) links leading out of the host countries. Other uses such as RCs acting as off-site backup centres could also be explored. The following sections describe some of the potential uses for RCs in more detail.

3 Processing and Data Storage Capabilities

There are large computing resources available throughout the world that could be used for SKA processing. A regional centre would help make best use of these regionally-funded computing and storage systems. These resources will continue to be refreshed over the time that the SKA operates and could eventually provide far more computing and storage resources than is available at the SDP sites. Other large data-processing projects, such as those associated with CERN, have demonstrated that obtaining regional funding for regional resources, often shared between different user communities, is possible even when obtaining funding for

project specific equipment located in other geographic regions is not. These resources could eventually be used for off-site backup storing primary copies of Science Products as well as for reprocessing data as algorithms are improved or other reference models are updated.

With the role of supporting SKA processing activity in a region, the centre would provide technical support for researchers in that region and also support the software and middleware stacks needed for obtaining and processing SKA data and for analysing results using visualisation applications. Supporting a common set of tools across centres processing SKA data will enable different sets of resources to be utilised for different projects in a seamless manner. Particular RCs could support additional tools or extensions to the common tools as needed.

As well as providing computing and hosting Science Products transferred from the SDP sites, RCs could also be used for storing derived Science Products created in that region, allowing for the sharing of these products by researchers in that region. These products could also be shared with or searched by researchers in other regions, if appropriate sharing agreements were put in place.

An extension to the use of RCs would be to allow for data processing initiated by users from other regions. This would make sense if particular Science Products are only stored in a particular region and the processing to be performed would eliminate the need to move these products to a different region for processing there.

One further activity that should be explored is the use of systems for reprocessing data. Reprocessing work is mentioned in some science use cases, but there is currently no support for this. If this reprocessing work was done often and the network bandwidth is available to transfer the data used in this activity to RCs, then work that would otherwise need to have time scheduled on the SDP systems could be performed at RCs. This could eliminate the need for SDP production work to be suspended for reprocessing work to be performed, or for the SDP systems to be scaled up to handle production and reprocessing work concurrently. By default the SDP will deliver a set of standard Science Projects [RD01]. If RCs were used for this type of processing, additional Science Projects would need to be made available by the SDP to be reprocessed at the RCs. This would also facilitate the development of future SDP pipelines at the RCs.

4 Network Optimisation and Data Caching

Distributing large amounts of data efficiently over large distances on Wide Area Networks (WANs) requires careful tuning of the network endpoints and of the storage systems on both the sending and receiving ends. Making an RC responsible for hosting Science Products transferred from the SDP sites to their region will allow for WAN and storage performance tuning for long distance transfers. These RCs would be responsible for monitoring the network and thus help to maintain high performance operation, quickly determining when problems occur. Keeping the networks leading out of the SDP host countries working optimally will also help with making new Science Products available in a timely manner.

By ensuring the RC is connected to a Point of Presence (PoP) for the regional research network, other institutions in the same region will be able to access hosted Science Products efficiently from these RCs without additional endpoint tuning and monitoring. This will help ensure that the main international network links from the SDP sites are used efficiently and thus make best use of the bandwidth available on these links.

In addition, the use of an RC for hosting Science Products could reduce the need to have the same Science Products transferred over the international links multiple times. As long as the products are stored in the RC it should not be necessary to transfer them to that region

again. An extension of this use as regional data caches would be to allow other RCs to access data products from these sites, reducing the load on the network links leading away from the SDP sites. Doing this would require agreements among the different RCs and/or between the RCs and the SKA Observatory.

5 Off-Site Backup

Another possible use for RCs is to act as off-site backup facilities for Science Products generated at the SDP sites. If there is network bandwidth available to transfer all Science Products for a particular project to a specific RC, that RC could become the backup site for that data, thus eliminating the need to keep a second copy of the data at the SDP site. If this backup capability was part of carefully considered data management and distribution plan for the SKA, then an agreement would need to be put in place such that the RC assured a suitable level of access to these products by the SDP site when needed. Given that we can foresee that regions are likely to want to have complete sets of Science Products for particular projects and that the type of centre that is likely to act as an RC will have its own data resilience strategy, providing this capability could be done with minimal additional cost.

6 Non-Technical Requirements

In addition to providing technical resources, RCs could support regional coordination and community building for SKA activities. Utilising existing centres of excellence in astronomy and data science will provide a basis for these activities. In this role they would coordinate participation in SKA science programmes and foster interaction between the local scientific community, providing collaboration tools for the region's SKA scientists and engineers. They could also lead outreach to government, educators, industry and the general public in their region. In this role they could organise the creation of grant proposals and organise regional meetings, workshops and summer schools associated with the SKA and data science. It should be noted that developing these types of use cases is not considered to be part of the Science Data Processor work.

7 Case for Project Wide Agreements

The model of tiered data processing adopted by CERN's Worldwide LHC Computing Grid (WLCG) utilises agreements to determine which regional centre performs particular processing. In this, regional tier-1 centres are responsible for processing part of the data for a particular experiment and for archiving the allocated data sets. In the case of the ATLAS experiment these tier-1 centres distribute data to tier-2 centres in their region to perform part of the work and make the data available for tier-3 centres which are owned by and perform work for individual scientists. In this model the tier-1 and tier-2 centres have agreements with ATLAS to perform specific processing for the experiment. Resource pledges and 0 Service Level Agreements (SLAs) are made with the WLCG to support this processing. The SLAs include uptime, responsiveness and capacity requirements. It is not clear that this type of central management will make sense for survey type work performed by radio astronomers. However, as pointed out in the previous sections, resources supporting a common tool-set could be used in a flexible way appropriate for particular projects. Agreements could enable more effective sharing of

data and thus decrease the load on key WAN connections. Also as noted above, agreements to store particular Science Products could allow RCs to act as off-site backup facilities.

8 ALMA Use Case

The support of scientists provided by ALMA is organised via a network of regional support nodes. There are three so-called ALMA Regional Centres (ARCs), located in Germany, Japan and the USA. Each centre supports several so-called ARC nodes that provide a sub-set of the ARC functions. The role of the ARCs is twofold. First, each ARC contains a full copy of the ALMA archive, allowing users to access data from a geographically close location. ALMA provides images processed using default parameters that can be used for first inspection as well as raw data that can be reprocessed if specific parameters are needed for the science goal of the observer. Delivering data by shipping hard drives was considered, but now raw data is distributed over the network. It is also possible to contribute advanced data products back to the archive. The second role of the ARCs, together with the nodes, is to provide support to scientists. Examples of such support are helping PIs to write research proposals and supporting PIs whose proposals have been accepted to request an instrument configuration that fits the science goals best, data reduction help and help with data mining on the archive. The node's staff provide face-to-face support and also have systems available where a user can perform post-processing, directly supported by a local support scientist. The utility of these centres should be examined further and it needs to be determined what features should be supported by the SKA RCs.

9 Conclusions

Regional centres could play an important role in maximising what can be achieved by the SKA as a whole. They will facilitate easy and efficient access to data and common tools as well as support data mining on the archive and analysing the data. They could also be used for developing future pipeline algorithms and testing new pipeline implementations. They could provide regional points of presence enabling SKA project outreach to government, industry and the general public. As an extension to their regional role, regional centres could also be used for optimising global data distribution, to provide off-site backup for SKA data, and for providing international access to locally produced Science Projects. As the capabilities of international WAN links increase, so will the ability to move increasing amounts of data from the SDP sites to RCs, allowing RCs to take over a larger role in SKA data processing.