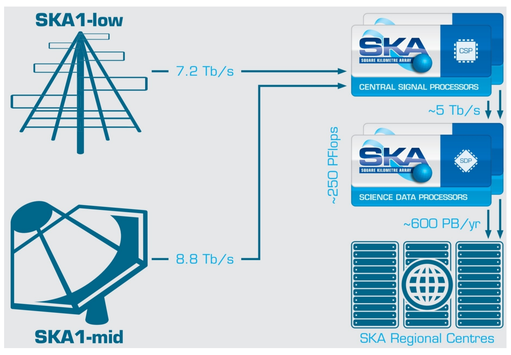
ADVANCED COMPUTING

SKA Science Data Challenges (SDCs)

Bruno Coelho (IT), Domingos Barbosa (IT), Sonia Antón (CIDMA, Dep Física-UA), Jorge Bruno Morgado (FCUP), Valério A.R. M. Riberiro (IT; CIDMA, Dep Física-UA), Dzianis Bartashevich (IT), João Paulo Barraca (IT, DETI-UA), Miguel Bergano (IT), Dalmiro Maia (FCUP), Tjarda Boekholt (IT)

Background

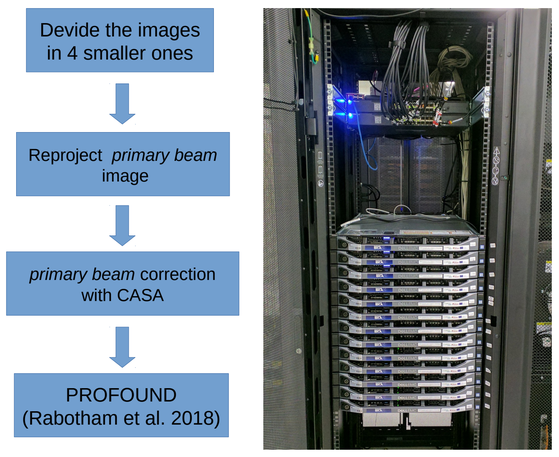
The Square Kilometre Array (SKA) will be the world's largest radio telescope, to be built in Australia and South Africa, it is a global effort of several countries around the world. The large amount of data will push the frontiers of technology not only in terms of processing capabilities, but also in terms of data storage, data transfer and data analytics requiring a large degree of automation relying mostly on machine learning and Artificial Intelligence algorithms.

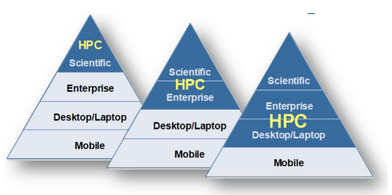


SKA organisation (adapted)

Methodology

The first SDC consisted of nine, very dense and large images, each with 32768x32768 pixels and 4GB in size. They are simulations of how the SKA's mid-frequency array, to be located in South Africa, would see the radio sky at 3 different frequencies (560 MHz, 1.4 GHz and 9.2 GHz), and 3 depths: 8, 100 and 1000 hours of observing time. The objective was to find, calculate physical properties of, and classify the sources in those images, in an automated way to demonstrate the capabilities and measure compute scalability of a processing workflow.



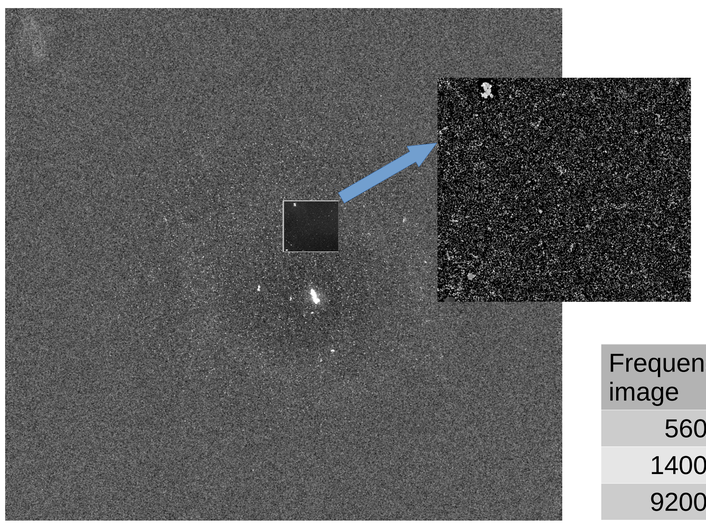


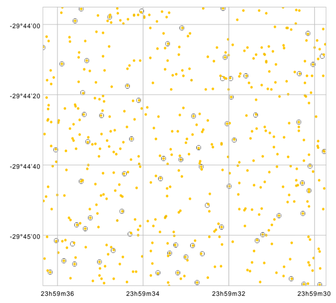
Big data – HPC convergence
...Through the Cloud !

We adopted a direct approach using open source tools, based in R and python, and we made use of the Engage SKA Computer Cluster (12 servers Dell Power EDGE R630) currently in use for SKA1 Software Prototyping.

Results

We participated as team Engage SKA – Portugal on the first SDC, obtaining the 3rd place in an international competition. Overall, team Engage SKA got excellent results in retrieving the cosmic catalogue information. However, there is room for improvements and a more precise morphological determination and automated data pipelines developed in collaboration with the software industry will rely specifically on the application of tuned machine learning techniques in a near future under an HPC environment.





Frequency of the image	No. of sources detected
560 MHz	447 756
1400 MHz	152 003
9200 MHz	469

Impact/Conclusions

In this SDC we found that although some analysis steps may be parallelized, the main limitations were on the memory consumption when dealing with such large images, implying that the challenge will be greater by several orders of magnitude when analyzing large multidimensional data-cubes that will be produced by SKA spectroscopic polarization observations.

Both the development of more sophisticated algorithms with a high degree of automation to astronomical source classification, and the data flow/work flows within the SKA Regional Centres (SRC) in a near future constitute good examples of science cases of advanced computing techniques, and may much benefit from large and powerful machines, such as "BOB" of the Centro de Computação Avançada do Minho. The SRCs federated architecture is currently being investigated through the H2020 AENEAS project to leverage the European processing of more than 600 PB/year. This constitute also a preparation for Portugal to deal with its share of 10 PB/year of the SKA Science archive.