Big Data and Extreme Scale Computing, 2nd Series (BDEC2) White Paper - Statement of Interest from the Square Kilometre Array Organisation (SKAO)

The Square Kilometre Array (SKA) and the High Luminosity - Large Hadron Collider (HL-LHC) are Landmark projects in the area of Big Data and Extreme Scale Computing. In 2017 we launched a formal collaboration on extreme scale computing to work together on areas of mutual interest.

The SKA project will build a new astronomical observatory spanning three continents - Europe, Africa and Australasia - and provide the scientific community with a step-change in radio astronomy capability with two new telescopes: one operating in the frequency range 50-350 MHz - SKA1-LOW in Western Australia - the other operating in the frequency range 350 MHz to 15 GHz - SKA1-MID in South Africa. Raw data rates in the petabit/s range need to be reduced through a number of processing stages to output science data products in the rate of GB/s.

The existing Large Hadron Collider at CERN will undergo a major upgrade in the early 2020s. In the HL-LHC phase the data rates will grow by a factor of 10 with respect of the current running conditions and the complexity of the physics events will also increase considerably. We foresee the need to store and process exascale-level of data while we do not foresee an increase of the budget for computing hardware.

Both projects are large and complex big data challenges. In addition to enormous on-line real-time data and processing challenges, several hundred PB/year will be transported around the globe to a network of centres for further analysis. The SKA Regional Centres (SRCs) will be similar to the Worldwide LHC Computing Grid operated by CERN and its partners. The data centres will likely be realised by a network of cyberinfrastructure shared with other science communities.

The specific application requirements are unique, but there is also commonality with other disciplines. Below we discuss the three areas highlighted in the call for contributions, give our perspective on these matters, highlighting where the our computing requirements perhaps differ, but also where we share the concerns with other data-intensive sciences communities.

1. Novel models of integrated inquiry.

   Neither the LHC's existing applications or those envisaged for the SKA once in operation can be classed as typical HPC workloads. We do recognise the and concur with the recommendations laid out on p4 of the BDEC "Pathways to Convergence" report regarding decentralized edge systems and centralised facilities and the need for these infrastructures to support a variety of complex workflows generating considerable volumes of data, some with demanding real-time characteristics, over wide geographical areas. Both projects are keen to further understand how developments in machine learning and deep learning might be applied to problems ranging from new end-user analysis methods and tools to aid visualization and comprehension of huge data sets, to more efficient algorithms for data reduction and ways of improving the reliability, availability and maintainability of very complex scientific instruments and their supporting computing infrastructure across the globe. A recently held workshop at the Alan Turing Institute in London brought together members of the high-energy physics and astronomy communities with a view to identifying areas of commonality and avenues for future collaboration. Extending these conversations further to include representatives from across the communities represented in BDEC2 is of real interest to us.

2. Support for advanced data logistics.

   The ‘extreme scale’ data production of the likes of SKA and HL-LHC will require ways of developing new cost-effective and user-friendly cyber-infrastructures that are able to capture, analyse and store vast quantities of data for decades. The challenge will be to ensure that these infrastructures will span continents and be able to utilise a heterogeneous pool of resources: HTC, HPC, clouds, CDNs etc. We refer to this model as a ‘data lake’ (but this is perhaps a misnomer in that it differs to how the term is interpreted in a commercial ICT context). The LHC experience shows that delivering data to processing cores in typical HPC systems in an efficient way is a concern. The problems are twofold: efficient ingest of large volumes of data into the receiving nodes within the HPC platform; and the subsequent efficient and timely placement of data on to processing nodes when using shared parallel file systems and typical HPC batch scheduling methods. Traditional HPC techniques do not work well at petascale and certainly won’t at exascale. The way the convergence of techniques from the HPC and HPDA worlds evolves will be key to data-intensive science communities such as ours being able to fully exploit the full range of cyber-infrastructures expected to emerge in the coming decade.

   The infrastructure we will use within the SRCs and WLCG will most likely be shared with peers in other communities: we do not envisage member states funding new infrastructure exclusively for use by the SKA or LHC science community. Therefore it is vital to work with other projects to look at how common tools and techniques around virtualization, containerization in particular (e.g. the “hour glass” model), along with common analytical front-end platforms for driving varied workflows can be developed to meet our needs.

3. Interfaces to commercial cyberinfrastructure.

   SKA and HL-LHC view commercial cyberinfrastructure as part of the overall landscape but we are not concerned with users working at the ‘long tail’ since the user communities will be highly organised around the SRC and WLCG modes of operation. We see the use of commercial clouds as potentially complementary to dedicated resources within the regions around the globe. Current analysis of the overall cost of using commercial cloud suggest that it is not an affordable way of providing the majority of cycles our communities require. (But of course the long-term costs of commercial clouds vs. on-premise provision are hard to predict.) The LHC community has carried out large scale demonstrator projects including US labs (BNL and FermiLab) in order to reach these conclusions. Regardless of the uncertain long-term financial picture, both the SKA and LHC communities would like to fully explore how infrastructure owned by the projects might be seamlessly integrated with publicly funded shard cyberinfrastructures and privately operated public clouds. It should be noted that the appetite for using the leading commercial cyber-infrastructures will likely vary from region to region. For example, the European Commission is clearly concerned about developing deeper dependencies on US-based tech monopolies, and is therefore funding programmes that might lead to greater “technological sovereignty” such as the European Processor Initiative and the European Open Science Cloud. Similarly, in China there are number of exascale initiatives that will be based on homegrown technology, and the country’s science community will have its own commercial cloud infrastructure to use such as Alibaba. As a global project, the SKA needs to develop strategies that bear in mind these regional differences and what drives them, and develop its SRC network accordingly.

The BDEC2 community might also wish to consider whether a collective approach could be adopted to more effectively lobby governments and argue for higher funding levels to meet the challenges we face.