Contribution ID: **161** Type: **Presentation**

Data Challenges at the Square Kilometre Array (SKA)

Tuesday, 9 October 2018 12:15 (15 minutes)

The Square Kilometre Array (SKA) will be a radio telescope distributed over two continents: In South Africa approx. 190 parabolic antennas will be built, in Australia more than 100,000 dipole antennas. The computing at SKA has to cope with next-generation big data analytics challenges: So many data will be taken that only a tiny fraction can be stored in long term archives. Extracting the relevant astronomical information out of huge data streams has to be done nearly in real-time. Due to the complexity of the workflows an enormous computing power is needed. Moreover, the fantastic resolution of the antennas result finally in 3D-images of the universe, which may become as large as one petabyte: Traditional computing architectures are not designed for analyzing objects of such size.

The signals from the antennas are "interfered" in local stations and sent to two computing centers in South Africa and Australia, respectively. The antennas generate a 24/7-stream of "raw data" of the order of 2 Pb/s, which is more than the global internet traffic (~360 Tb/s, Cisco 2016). In both computing centers the incoming data are analyzed iteratively by complicated workflows to reduce the data volumes strongly. The outcome of the central data centers are called *science data products* and will be transported to a few "Regional Centres ". In Europe there will be one virtual Regional Centre that is physically distributed over the European SKA member states. The community of astronomers can access SKA data only via the Regional Centers.

The project AENEAS (Advanced European Network of E-infrastructures for Astronomy with the SKA) is developing a design for the European Regional Center. The talk will give an overview of the current status of AENEAS.

Huge data objects (~1 PB / object) can only be analyzed sufficiently fast if they are stored "in-memory". This needs a radical change in the design of computing infrastructures away from a "processor-centric computing" to a "memory-driven computing". The talk will give an overview of the results of two recent workshops, where the need of a paradigm shift was discussed by big data analytics experts:

- *Exascale Data Center*, Berlin, Jan. 30, 2018
- *Memory-driven Computing for Big Data Analytics*, Berlin, May 30, 2018

see http://bigdata.htw-berlin.de.

Finally it will be indicated that the big data analytics challenges at SKA are not just a "do it bigger and do it faster business" (G. Longo). Almost all data (more than 99.999 % of the raw data) are already rejected, before a human researcher will have had the chance to start his analysis. This needs in particular a development of highly parallelizable machine learning techniques, which are currently not available. Suitable statistical procedures are needed for evaluating the quality of the remaining data, as done exemplary in high-energy physics at the Large Hadron Collider (LHC). Moreover, developing a scalable distributed memory-driven computing infrastructure is an interdisciplinary challenge, where scientists of different disciplines and industry have to cooperate.

Type of abstract

Presentation

Summary

The computing requirements of the Square Kilometer Array (SKA) are most challenging and require paradigm shifts. Only a tiny fraction of the raw data collected by the antennas can be provided for scientific analyses. Novel distributed machine learning methods have to be developed for reducing the data volumes and suitable indicators are necessary for estimating the effectiveness of the data reduction. The extremely high resolution of the antennas result in huge 3D-images of the universe (up to 1 PB/image) that may be analyzed in-memory if the current "processor-centric computing" is replaced by a "memory-driven computing".

Primary author: Prof. HESSLING, Hermann (Univ. of Applied Sciences (HTW) Berlin)

Presenter: Prof. HESSLING, Hermann (Univ. of Applied Sciences (HTW) Berlin)

Session Classification: Thematic Services

Track Classification: Area 3. Computing and Virtual Research Environments