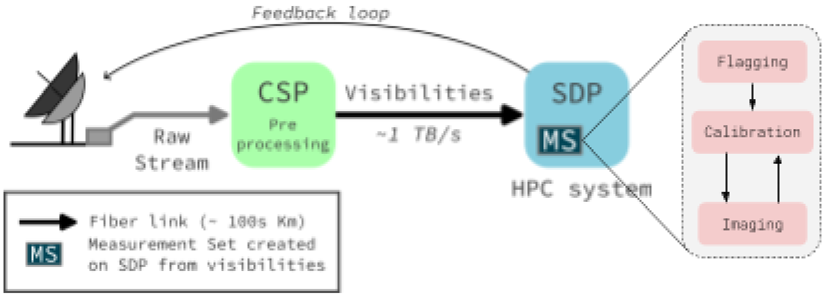
Staging In and Out Astronomical Data
for Large-Scale Radio Astronomy Processing Pipelines

Participants

- François Tessier (Inria) - francois.tessier@inria.fr
- David Guibert (Eviden)
- Cyril Tasse (Observatoire de Paris)

Context, positioning and objectives

The international SKA[1] project aims to create the largest telescope in the world in order to observe a part of the universe. A very large volume of data is generated at the telescope level, pre-processed on local clusters (filtering, reduction) in real time and sent to a supercomputer (SDP) at a rate of 1TB/s. This data feeds numerical simulation, generating 1PB of daily output data that needs to be saved. At this stage, the computing power and storage resources required are such that machines capable of reaching the **exascale**[2] become necessary. However, the efficient use of these systems raises new challenges, especially regarding **data management**.



One clearly identified area for optimization concerns the way in which data is stored and processed on the SDP. The workflow can be described as follows: the data captured by the telescopes is written into **Measurement Sets (MS)**. The MS files, which weigh from a few dozen MB to several hundreds of GB each, then pass through the **flagging phase**, during which interference (due to satellites, for example) on the captured radio frequencies is eliminated. In the next step, the resulting data enters a loop consisting of a **calibration phase** to contain systematic errors, followed by an **imaging phase** to transform data into the image domain. The final output is a visualization that can be interpreted by astronomers. Several software solutions exist to perform this processing, but one potential candidate for SKA is the **DDF pipeline**. During the execution of this workflow, a large volume of data is transmitted or stored, **significantly impacting the time to solution and limiting its ability to scale up**.

[1] https://www.skao.int/en
[2] A measure of computing power representing one trillion (10^18) floating point operations per second

In this context, we would like to recruit a student for a 6-month internship to develop a tool for **distributing the Measurement Sets** used by the various components of the pipeline. Scripts already exist that allow radio telescope observation datasets to be retrieved from remote databases, but these scripts are very basic and difficult to generalize. We want to offer a new tool, called **MS-Broker**, which will allow **datasets to be retrieved online**, **distributed** to shared or node-local storage on the computing cluster according to the pipeline's needs, and all or part of this data to be **repatriated** (some fields may have been added or modified by certain components of the pipeline). The aim of this internship will be to develop such a tool, in collaboration with astronomers working in the ECLAT joint laboratory. The first phase will be devoted to familiarizing the student with the **pipeline by running it in an HPC environment**. Then, after analyzing the requirements, the student will be responsible for **developing an initial prototype of MS-Broker**. **Validation** through the generation of sky images will be expected.

Skills and Abilities

- Programming skills (Bash, Python, MPI)
- Knowledge of computer networks and distributed systems
- Familiarity with parallel and distributed computing is a plus
- A good level of English is required

Organization and Implementation

The student will join the **KerData research team at Inria Rennes**[3] and will be co-advised by **Inria, Eviden** and **the Paris Observatory**. The main supervisor will be **François TESSIER**, from Inria Rennes.

- Inria is the French leading public research institute in computer science. KerData is a research team focusing on data management across the computing continuum, i.e. all computing resources from the Edge to HPC/Cloud infrastructures. Part of the team has a particular focus on I/O and storage issues on large-scale systems.
- Two teams from Eviden will be involved: the Center for Excellence in Performance Programming (CEPP) for the application and performance aspects and the Data Management R&D team which develops products for IO monitoring and optimization. People are based in Rennes and Echirolles.
- The Paris Observatory brings together and manages Earth and Universe sciences research activities. Its missions are research, observation, training and the dissemination of knowledge in these fields.

The internship is expected to **start around February 2026 (flexible).**
The internship is remunerated on the basis of the civil service internship income (~600€/month).

If you are interested, please contact François Tessier - francois.tessier@inria.fr.

[3] Campus de Beaulieu, 263 Av. Général Leclerc, 35042 Rennes