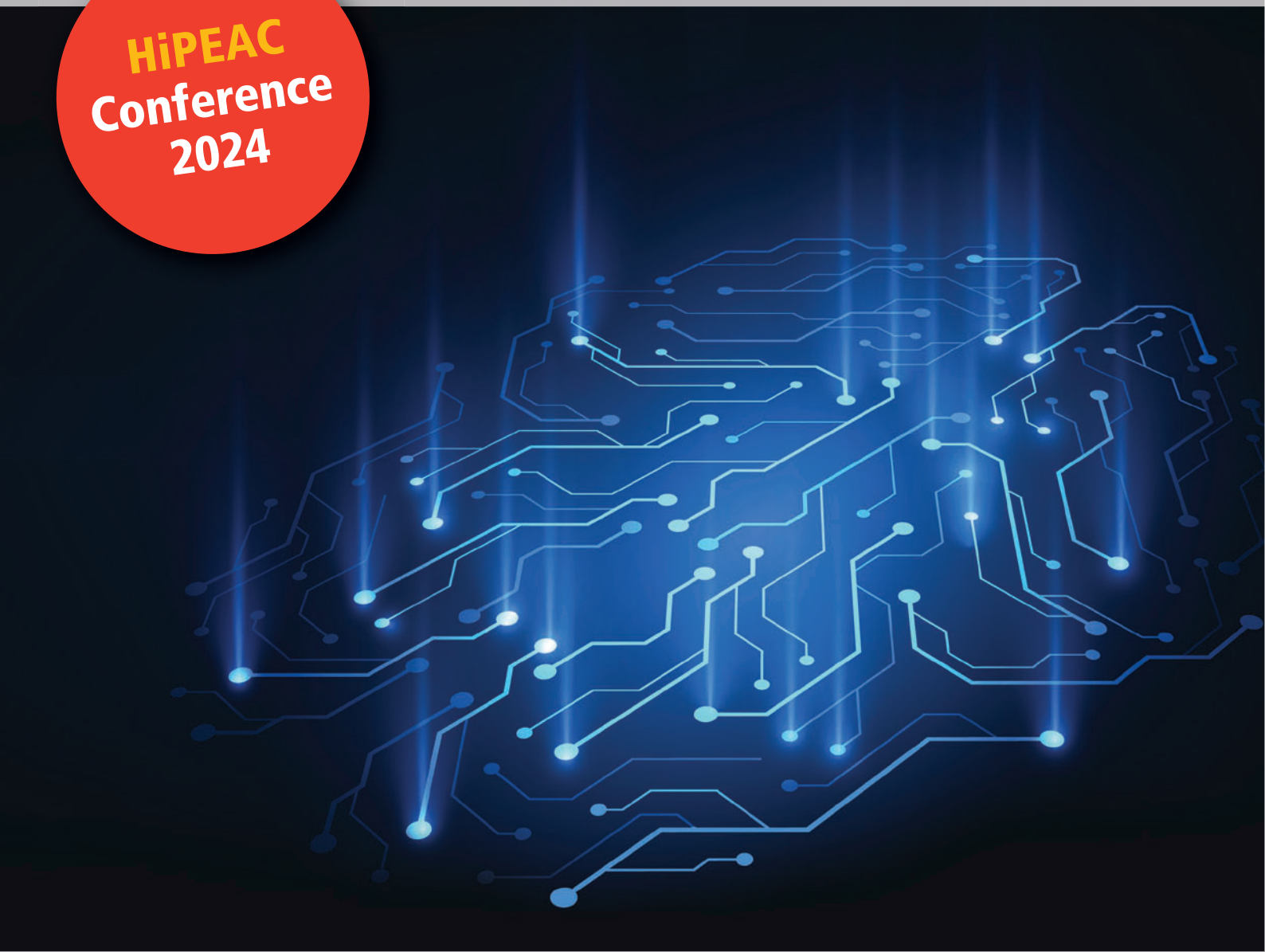
HiPEAC info 71

JANUARY 2024

HiPEAC Conference 2024



Entering the next computing paradigm: The HiPEAC Vision 2024

Spinning out the smarts: Powering edge AI and distributed computing

Lieven Eeckhout on sustainability, Reetuparna Das on data-centric architectures and Mitsuhisa Sato on building a top supercomputer


Partners in the EXTRACT project (EXTReme dAta Across the Compute conTinuum) are working on a data-driven, open-source platform integrating cloud, edge and HPC technologies for trustworthy, accurate, fair and green data mining workflows. In this article, Daniel Barcelona Pons and Enrique Molina Giménez (Universitat Rovira i Virgili) explain the data pipelines aspect of this project.

Taming a universe of data

How the EXTRACT project is parallelizing data-processing pipelines

The Cloud and Distributed Systems Lab (CLOUDLAB) research group from the Universitat Rovira i Virgili (URV) is a multi-disciplinary team that tackles key research lines of distributed systems. This research group has experience in scalable systems (cloud computing, serverless architectures, distributed storage, peer-to-peer) and web-based infrastructures.

As part of the EU-funded EXTRACT project, in which CLOUDLAB participates, project partners are working together to create enhanced workflows that will process extreme data reliably so that it can be used across a variety of scientific disciplines. Extreme data possesses a set of challenging properties such as high volume and speed, but also variability, that make it very hard to manage effectively.

The project's technology is being validated on two use cases:

- a personalized evaculation route (PER) system to guide citizens through a safe route in real time
- the TASKA (Transient Astrophysics with a Square Kilometre Array Pathfinder) use case, driven by l'Observatoire de Paris

This article focuses on the TASKA use case.

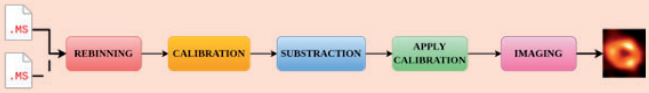
The TASKA data pipeline

Extreme volumes of data (from a few gigabytes to several terabytes) at variable speed are captured by many radio-telescope antennas. This data must be processed to generate high-resolution images of the cosmos that scientists can interpret. To help generate these images, the EXTRACT project is pursuing technical synergies that will help create and improve the workflows used in the TASKA use case by integrating the latest cloud technologies in data-processing parallelization.

The data processing necessary to obtain these images requires several composable steps to prepare and then analyse the data collected by antennas in the MeasurementSet (MS) format established by the Common Astronomy Software Application.

These steps are as follows:

1. rebinning
2. calibration
3. subtraction
4. applying calibration
5. imaging



Example of TASKA pipeline where antennas collect data and generate images

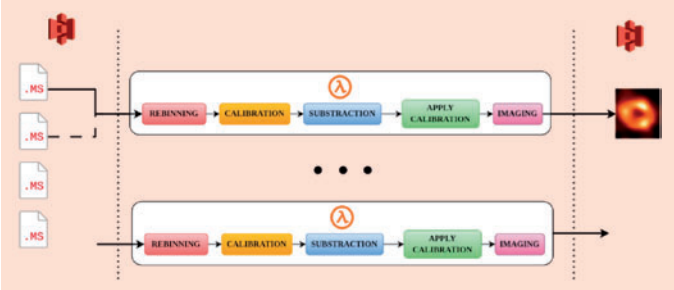
Currently, this pipeline yields poor performance since it is executed manually and monolithically onsite, meaning that it is difficult for scientists to explore these data effectively in a timely manner. Indeed, this process has the potential for many improvements. The CLOUDLAB research group is currently working on two notable improvements to increase the performance of data processing in the TASKA use case: inter-job parallelization and intra-job parallelization.

Serverless technologies create an ideal scenario for executing this pipeline. These technologies can support high scalability while abstracting the underlying infrastructure, which is key to adapting data processing to a variable volume of data and obtain results in real time. CLOUDLAB uses a function-as-a-service approach and its corresponding parallelization (specifically the Lithops cloud framework) to improve the current TASKA pipeline.

Inter-job parallelization and intra-job parallelization

The first improvement that CLOUDLAB is implementing is inter-job parallelization. In this case, 'job' is understood as an instance of this process that takes a set of MS files and generates one of the

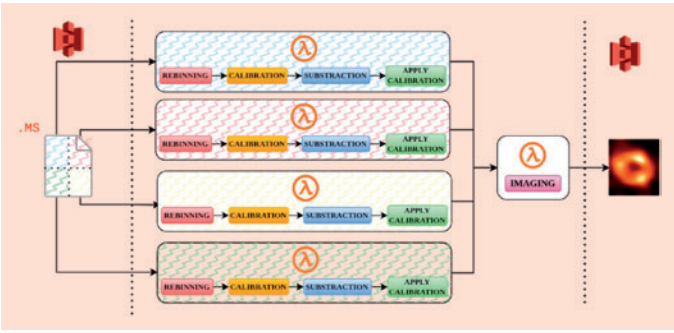
desired images. By running several jobs concurrently on different groups of functions, we already parallelize the execution of this process and generate a pipeline for the creation of images, which helps keep the process to a reasonable timeframe after the arrival of data.



The TASKA use case pipeline showing inter-job parallelization

By analysing the internals of the process, it is also possible to apply intra-job parallelization. This means parallelizing the different steps within the generation of the same image. It is possible to split each of the four first steps into multiple 'workers', with each taking a part of the original input dataset MS. Workers will consume partitions of the data to perform processing cooperatively. This parallelizes a large part of the workload within a job and yields further performance improvements. The last step (imaging), however, operates on the aggregate result of the previous steps and requires synchronization.

Performance improves even more when these two types of parallelism are combined. Combining both parallelization strategies results in the rapid processing of multiple datasets and the continuous generation of astronomical images, which can then be analysed by scientists in a timely manner.



The TASKA use case pipeline showing intra-job parallelization

The parallelization of the TASKA workflow will allow it to behave elastically when dealing with the variability of the extreme data that feeds into it. However, there are still challenges to solve in such a complex and demanding task, such as the correct and efficient partitioning of a complex data format like the MeasurementSet.

The EXTRACT consortium will continue towards further improvements on the TASKA use case. Specifically, they will examine a novel way to ingest data efficiently with smart partitioning. The success of this research will contribute to a better data-staging solution for the EXTRACT platform, which will help reduce data-processing latency thanks to more effective resource utilization.

More EU data projects

In addition to participating in EXTRACT, CLOUDLAB coordinates three EU-funded research projects:

- **NEARDATA**: Extreme Near-Data Processing Platform
  This project creates an extreme data infrastructure to mediate dataflows between object-storage and data-analytics platforms across the compute continuum. The NEARDATA platform is a novel technology for the mining of large and dispersed unstructured data sets that can be deployed in the cloud and in the edge (high-performance computing (HPC), internet-of-things (IoT) devices), that leverages advanced artificial intelligence (AI) technologies and offers a novel confidential cybersecurity layer for trusted data computation.
- **CloudSkin**: Adaptive virtualization for AI-enabled Cloud-edge Continuum
  This project aims to design a cognitive cloud continuum platform to fully exploit the available cloud-edge heterogeneous resources, finding the 'sweet spot' between the cloud and the edge, and smartly adapting to changes in application behaviour via AI.
- **CLOUDSTARS**: Cloud Open-Source Research Mobility Network
  CloudStars is a staff exchange programme that allows the mobility and exchange of researchers between academia and industrial institutions in the fields of cloud computing and AI technologies.

FURTHER READING

CLOUDLAB group cloudlab.urv.cat/web

EXTRACT project extract-project.eu

EXTRACT has received funding from the European Union's Horizon Europe programme under grant agreement number 101093110.