The ALMA Data Processing System

Sandra Castro

ESO – ALMA – CASA Team

Outline

The existing ALMA Data Processing System: CASA + ALMA Pipeline

The challenges coming from the WSU for the data processing system

The plan for a new data processing system to support ALMA-WSU

Contributing with European Development Studies in this area

The current CASA + ALMA Pipeline system

- The current CASA + Pipeline system was designed in the early 1990’s and early 2000’s.
- The combination of CASA with the ALMA Pipeline already has challenges processing today’s data with 19% of 12m data not having all sources and spectral windows imaged and/or the images having less than optimal properties (A. A. Kepley, Lipnicky, et al. 2023).
- The most challenging use case by far is imaging higher spatial resolution cubes with many channels.
- The significant increase in number of channels due to WSU will further exacerbate these challenges.

The WSU data processing challenges

- Estimates of the ensemble of WSU data properties predict that 15% of data sets will have visibility data volumes greater than the largest ALMA BLC/ACA project today.
- 20% of projects will have products larger than the current product size limit (A. Kepley et al. 2024).
- Tests of the current data processing software (NRAO 2019; A. A. Kepley, Madsen, et al. 2023) show that fundamentally the current data processing system architecture will not scale to support processing ALMA WSU data.

The most fundamental limitations for the current system

- The inability to horizontally scale jobs across multiple nodes with the existing parallelization framework.
  - This inability limits the sizes of cubes that can be imaged to those that can fit on a single node.
- A fixed resource allocation (typically 8 cores, 256GB of RAM) for all jobs, no matter how demanding the job is.
  - This limitation is due to a combination of how the Pipeline and CASA are launched in operations.
- Failures require re-starting jobs from the beginning.

To process WSU data, we require the ability to:

- Horizontally scale jobs across nodes so that we can process cubes with 80,000 channels.
- Dynamically allocate compute resources appropriate to the job.
  - This change will enable ALMA to more efficiently use compute resources and decreasing the overall cost of computing for a given data volume compared to today.
- Re-start jobs from where they failed rather than the beginning.
  - For WSU data processing, more jobs will fail stochastically just due to the increase in size of data processed.

Contributing to the next generation data processing system within this CfP for European Development Studies

RADPS – The new Data Processing System

Radio Astronomy Data Processing System - RADPS

- RADPS is an NRAO-led program to develop a system with the primary objective to support production of high-level data products for ALMA-WSU and next generation VLA.
- Current Status:
  - System Requirements Review passed in February 2025.
  - Conceptual System Design being validated
  - **RADPS-ALMA** is a **project** within the program with a Development Study proposal with the **goals to** demonstrate that the RADPS design meets ALMA's requirements at the level of a Preliminary Design Review and prepare the project to enter the construction phase post-PDR.
  - Approx Timeline: mid-April 2025 to March 2026
  - Currently prototyping a data processing stack → VIPER

VIPER

Visibility and Image Parallel Execution Reduction

Prototype Packages Being Developed:

These prototype packages are still under development and will be rapidly changing, however, some have progressed to the stage where tutorials are available.

Package	Description	Tutorial Available
astrohack	Antenna panel and position corrections.	Yes
casagui	CASA GUI desktop.	Yes
graphviper	Dask based MapReduce for Multi-Xarray datasets.	Yes
astroviper	Radio interferometry data processing.	No
xradio	Xarray radio astronomy data IO.	No
cloudviper	Cloud-native container orchestration system configurations	No
toolviper	Radio astronomy processing tools using the VIPER framework	Yes

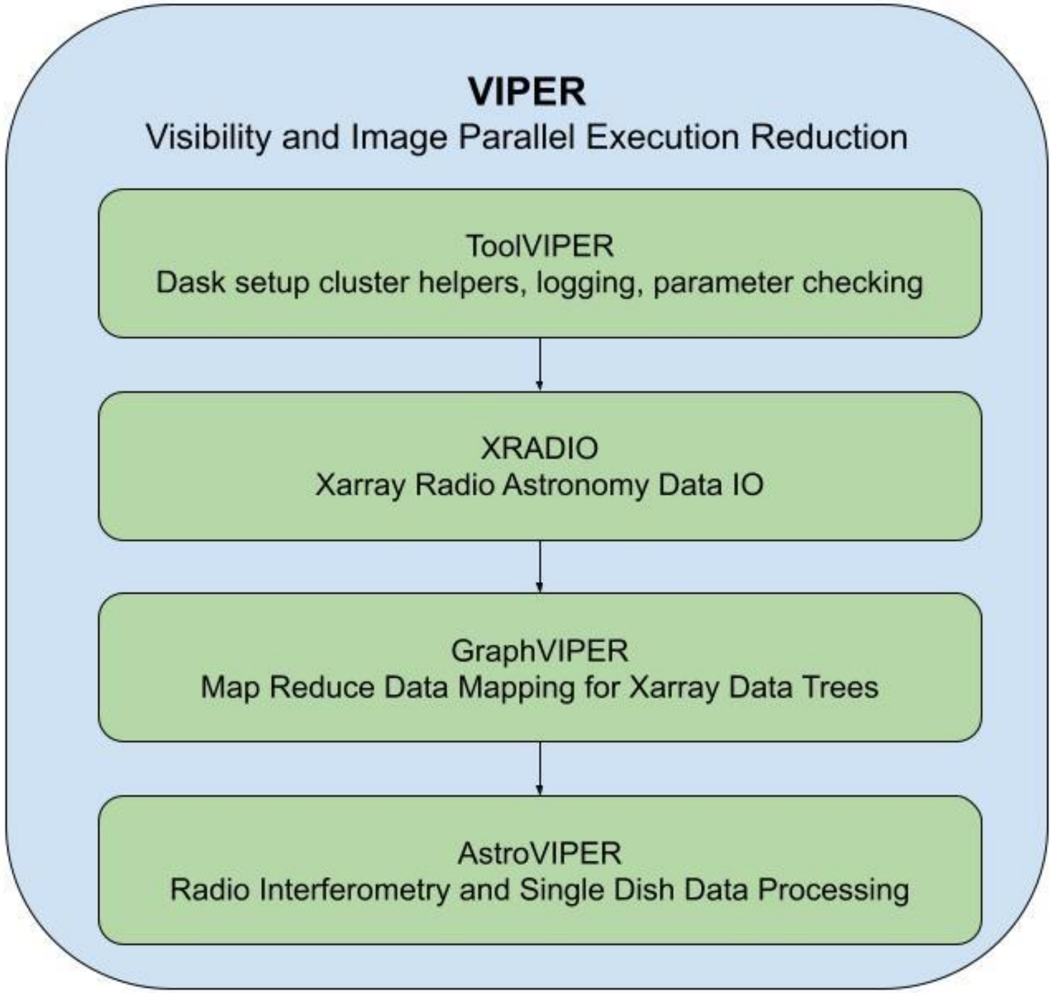
Contributing to XRADIO

https://xradio.readthedocs.io/en/latest/overview.html#Contributing

All repositories are open source developed under the BSD 3-Clause License including a Contribution License Agreement (CLA)

https://github.com/casangi

VIPER Ecosystem



https://github.com/casangi

Documentation and tutorials are available for some of the projects.

The description and selection of data in `XRADIO` is based on `xarray`. To use `XRADIO` effectively, it's crucial to understand the terminc

- xarray terminology
- xarray indexing and selection guide

Search docs

OVERVIEW

Introduction

Contributing

We welcome contributions to XRADIO from the radio astronomy community and beyond!

Preparation

- Read the XRADIO Overview, Development, and the relevant schema section for example Measurement Set v4.0.0.
  - Pay special attention to the Foundational Reading subsection in the Overview.
- Complete the relevant tutorials (for example the measurement set tutorial), which demonstrates the schema and API usage.

Submitting Code

- Any code you submit is under the BSDv3 license and you will have to agree with our contributor license agreement that protects you and the XRADIO project from liability.
- Create an issue on github outlining what you would to contribute XRADIO GitHub repository.
- Once there is agreement on the scope of the contribution you can create a branch on github or in you clones repository:

```
git checkout -b feature-or-fix-name
```

(If you create the branch in your cloned repository remember to link it to the GitHub issue). - Make your code changes and add unit tests. - Run the tests locally using pytest. - After running Black add, commit and push your code changes to the GitHub branch:

```
git add -u :/ #This will add all changed files.
git commit -m 'A summary description of your changes.'
git pull origin main #Make sure you have all the latest changes in main.
git push
```

- If you are making many changes you can break up the work into multiple commits.
- If tests pass and you are satisfied open a pull request in GitHub. This will be reviewed by a member of the XRADIO team.

Contributing to XRADIO

https://xradio.readthedocs.io/en/latest/overview.html#Contributing

All repositories are open source developed under the BSD 3-Clause License including a Contribution License Agreement (CLA)

References and links

1. A. A. Kepley, Lipnicky, et al. 2023: “Mitigation Statistics for ALMA Cycle 7.” ALMA Memo Series 263 (February)".
2. A. Kepley et al. 2024: “Data Processing Working Group Report: ALMA WSU Size of Compute Estimate.” ALMA-05.00.00.00-3064-A-REP”
3. NRAO 2019: “CASA Next Generation Infrastructure.”
4. A. A. Kepley, Madsen, et al. 2023: “Imaging Unmitigated ALMA Cubes.” NAASC Memo Series 121 (August)
5. XRADIO: https://xradio.readthedocs.io/en/latest/index.html
6. GraphVIPER: https://graphviper.readthedocs.io/en/latest/
7. AstroVIPER: https://github.com/casangi/astroviper

Thank you!

Sandra Castro
scastro@eso.org

@ESOAstronomy
@esoastronomy
@ESO
european-southern-observatory
@ESOobservatory