ALMA Correlator Upgrade : Potential impact on CASA - V.2

Urvashi R.V. (on behalf of the CASA Team)
13 Sept 2017

Summary

The expected effects of the main upgrade parameters on CASA are summarized below.

Increased number of channels :

Specifications : The upgraded ALMA correlator will increase spectral resolution by a factor of 8 and double the bandwidth. The number of channels can reach 65K, 32K or 16K depending on polarization mode.

Impact on CASA : The 8 fold increase in the number of channels will effect all modules.

– Spectral-domain data parallelization schemes exist and are expected to scale although significant work would be needed over the next couple of years to ensure that it does.

– Spectral domain image cube parellelization schemes for image storage/analysis/display modules currently do not exist and will need new ideas and development.

Details are given on pages 2 and 3. All development items are generic enough that they can and should be folded into CASA development over the next few years irrespective of the proposed ALMA upgrade, in order to more efficiently support even the current ALMA and VLA telescopes.

Time estimate : 4-6 FTE yrs for data/image parallelization and performance improvmenets (coordinated effort between ~6 people), 1-3 FTE yrs for image cube visualization (1 NRAO member working with the external ACDC group), 1-2 FTE yrs for other items within CASA, 1-3 FTE yrs for pipeline-specific development and optimization for production use. Given the scale of the work, and overlap between members, a team-wide focus on HPC over the next few years would be required.

Increased Bandwidth Ratio :

Specifications : Increase bandwidth by a factor of two from 8 to 16 GHz x 2 polarizations and process the entire 4-12 GHz IF bandwidth in the correlator.

Impact on CASA : The bandwidth increase is not expected to be a problem. At 100GHz, the bandwidth ratio moves from 8% to 16%, with an expected increase in continuum sensitivity of 1.4 times. Wideband imaging algorithms as currently used by ALMA are expected to scale as is. The VLA currently handles much larger bandwidth ratios, so all basic algorithms and software are already in place. Some commissioning of modes specific to ALMA may be required

Time estimate : 0 - 0.5 FTE yrs to commission and debug algorithms for *each* new mode.

Increased Time resolution :

Specifications : Increase time resolution capability from 16msec to 1msec.

Impact on CASA : Existing data parallelization (and averaging) schemes are expected to scale as is. Proper handling of large volumes of meta data and calibration solution tables may need to be addressed

Time estimate : 1 FTE yr to address issues different from generic performance/parallelization for large datasets

High Performance Computing impact on CASA.

The following are some areas in which the ALMA correlator upgrade will require significant work from CASA, in particular in the context of parallelization and high throughput. The following are what we expect to need for the proposed ALMA upgrade, but they are all also reasonable directions for CASA to move in over the next few years. They are all areas where we are already encountering bottlenecks both from ALMA as well as the VLA (and the VLASS survey requirements) and which we will have to think about for the ngVLA as well.

– Schemes for partitioning data : CASA currently has multiple schemes of partitioning data across frequency (lists of Mss, Multi-Mss containing subMSs split on frequency, multiple spectral windows within each MS, channel chunking to limit memory use during imaging, partitioning based on user defined spectral coordinate systems, etc) along with complicated data selection requirements that optimize on the mapping between data channels on disk and image channels requested by the user/pielines. The interaction between these parts of the software is already very complex, and will need significant streamlining in order for it to scale to 8-16 times more channels and more fine-grained frequency selections. Consideration for such use cases must also be included in discussions about the Measurement Set V3 format for the storage and access of very large datasets.

– Communication across frequency partitions : Data parallelization schemes being commissioned currently operate by treating partitions completely separately for all analysis modules (except for continuum imaging where frequencies are combined during gridding). More channels will require more fine-grained partitioning on the frequency axis, which will increase the need for cross-communication across spectral window and partition boundaries for tasks computing statistics, calculating/storing/applying calibration solutions and uv-continuum fits. Data access mechanisms and algorithms that need to combine data across these boundaries will need some focused work over the next couple of years in order to handle this well. Options of always splitting on axes other than frequency may also need to be discussed.

– Image cube storage/analysis : It is unclear whether CASA image formats will scale well in terms of data access efficiency and ease of parallelization. Significant work will be needed over the next few years to evaluate the limits of the existing formats (native CASA images as well as reference-concatenated images that store sub-images tied together with a lightweight wrapper) and to develop or adopt optimized designs/schemes (for example, hierarchical image formats).

– Visualization of large cubes : The CARTA image viewer is supposed to handle very large image cubes in a scalable way along with support for remote viewing and exploration. However, as of 2017 it cannot handle even modest cubes efficiently and the implementation is undergoing a rework. The basic design itself is scalable, but it remains to be seen how much work will be needed before it can be used for the proposed new ALMA cubes. CARTA-based plugins and all analysis modules will also need parallelization in their design which will also depend on the ability of the CASA image format to handle it. As of now, CARTA development is being done entirely outside of NRAO.

– Data displays : Data plotting modules will need new development in order to scale appropriately. Support for parallelized data reads and distributed data formats will need to be added, as well as cacheless operation to limit instantaneous memory use. Data averaging is already built into the underlying code but work may be required to ensure that it scales appropriately.

– Conversion of ASDM to Multi-MS : Work will be required on performance improvements, introducing parallelization, and performing a direct conversion of the ASDM to the multi-MS format instead of having to go via a regular MS first.

– Handling channel-based flag commands optimally : If the recently proposed channel-based flag command syntax is used excensively within the ASDM's Flag.xml, the flagging module in CASA could see a significant increase in the number of 'manual' flag commands. The basic design of the flagger module is expected to scale in this regard, but work may be required to ensure that it does.

– Limiting memory use : An increased focus on resource estimation and controlling the memory use of all CASA modules will be required in order to support the increased parallelization needs (and related overheads). Efforts towards this are already under way on the 2017-2018 timescale.

– Pipeline modifcations : The pipeline currently has some modules that do channel based operations in python and these will need to move into C++ and/or work done to introduce parallelization. Currently two obvious candidates are 'fintcontinuum' scripts and channel based flagging heuristics and the creation of flagging views. The pipeline also manages domain objects that hold time based meta data and there could be issues similar to what has recently been faced via the VLASS in its OTF observing mode. Work on optimizing throughput for production pipelines on dedicated hardware (or external clusters) would also be required.

<u>Technical details that CASA will need</u> **(in order to plan the development) :**

To ensure that CASA development planning for the ALMA Correlator Upgrade proceeds rationally and proactively (rather than reactively), it is important that an organized discussion of the anticipated calibration and imaging heuristics occur.

In particular, our planning would benefit from details of how observations using the enhanced bandwidth, spectral resolution, and/or time sampling will be organized w.r.t. spectral window setups and intents (including mappings), combined use of wide bandwidths _and_ high spectral resolution, what the dominant use cases are expected to be in terms of spectral/time setups, etc.

Ideally, this discussion should include ALMA support scientists, ALMA pipeline experts, as well as CASA developers. The outcome of this discussion will help tremendously in setting the scope of any required development work, and get it on the todo list well in advance of actually needing it.