Report on NRAO CASA and CASA-based pipelines

2013 April 12

Review held at NRAO-Socorro, 2013 March 5–6

Written & endorsed by: Bob Hanisch (STScI), Jeff Kantor (LSST), Lee Mundy (U.Maryland), Bob Tawa (NEON), Rick White (STScI, chair), David Wilner (CfA), Michael Wise (ASTRON)

Additional committee members: Bill Cotton, Eric Greisen (NRAO)

1. Introduction

Our review committee met at NRAO in Socorro on March 5–6, 2013, to review the progress and plans for future development of the Common Astronomy Software Applications (CASA). We were presented with a thorough discussion of the current status of CASA, the most important issues, interactions and use of CASA for various NRAO projects (including ALMA, the VLA and GBT), and options for future work on CASA. The committee appreciates and thanks the presenters, authors of white papers, and all who contributed to the discussions.

Our overall impression is that the CASA project is on a productive path and has a lot to show for their efforts over the past few years. The current system has a good breadth of functionality and is actively being used by ALMA/VLA communities. The development of data processing pipelines (including pipelines that produce VLA calibrated data) are an important advance for the current NRAO user community and also are a step toward expanding the NRAO user base beyond its current population of "black belt" interferometrists. The adoption of CASA and CASAcore by external projects including MeerKAT and LOFAR is an encouraging development and demonstrates that the CASA project has produced a valuable system that is recognized as a good basis of support for future radio astronomy projects.

The Challenge: Balancing Current Focus and Forward-looking Projects

Many options for future work on CASA were discussed during the review. There are clearly plenty of important and useful improvements that could be made in the

system. The challenge is to determine the appropriate balance between shorter-term projects that will have an impact now and longer-term projects that position CASA for future radio astronomy data.

We attribute much of the recent success of CASA to a strong focus on the requirements of VLA and ALMA data processing. The CASA team should maintain that focus on this core mission going forward:

- Continue to fill in the functional gaps and missing features for ALMA/VLA data processing to enable more users to have a complete solution in CASA.
- Continue improving CASA performance. VLA users broadly agree that speed is the number one bottleneck issue in using CASA. CASA tasks should have their performance tested; the aim should be to have performance that equals or exceeds that of similar tasks in competing packages.
- Make the necessary architectural changes to enable CASA to handle the increasing volume and complexity of ALMA and VLA data over the next few years.

A focus on the existing NRAO facilities is not at all inconsistent with forward-looking developments in CASA. CASA should take advantage of opportunities to enable new science with the VLA and ALMA, while designing toward the future of radio astronomy. Supporting the increasingly challenging requirements of the current observatories is excellent positioning for the future.

Expanding the user base of ALMA and the VLA is central to NRAO's future success. Decisions on priorities of development for, e.g., imaging pipelines should reflect not just the needs of the current NRAO users but also those of the broader astronomical community that could do science using these powerful new telescopes. As CASA provides both higher-level science-ready data products and a more seamless user data analysis environment, the observatories will find wider use. That is surely the best positioning for the future of both NRAO and CASA.

The sections below address various general areas for CASA:

- Governance (§2)
- Software Development Processes (§3)
- Architectural Issues (§4)
- Pipelines (§5)
- Visualization (§6)
- Virtual Observatory (§7)

- Low Frequency & SKA (§8)
- VLBI (§9)
- GBT (§10)
- Remote Processing/Cloud (§11)

The final section (§12) reprises the questions from the charge to the panel and indicates where in the document the issues are discussed. **Recommendations are highlighted in boldface.**

2. Governance

The major programmatic sources of CASA requirements (ALMA and VLA) are changing with respect to their operational status and priorities, and this creates a dynamic requirements environment for the foreseeable future. The review panel observed clear gaps and inefficiencies in the process of soliciting, communicating, and prioritizing requirements for CASA. Meetings were not attended by all stakeholders, communications did not go to all stakeholders, and necessary technical and scientific representatives were not consistently involved in impact analyses and tradeoff processes. It has been difficult in certain cases to achieve scientific consensus and there does not appear to be a defined process to break deadlocks. The degree of community input appears to be insufficient to represent that perspective adequately.

Requirements management needs a well-defined, documented, and director-approved governance process:

- It should be relatively easy to make a request and to pull a team together for initial analysis.
- Subsequent in-depth impact analysis, approval, prioritization, resource allocation and commitment require a more formal process of management consent.
- This process needs to be executed on a regular, scheduled basis, and should provide a mechanism for ad hoc or emerging requests to "jump the queue" with management consent.
- There needs to be a defined way to move forward in the event of deadlocks.
- Requirements should be captured in a queryable database, and tracked through the process of request, impact analysis, approval, implementation, and verification.
- There should be clearly identified, explicitly authorized, individuals within

CASA, ARDG, ALMA and VLA who represent the requirements and priorities of those entities. These representatives must jointly define requirements.

Requirements that include both research and development components need a sanctioned way to create a combined scientist/development team for impact analysis (as well as implementation and verification).

- The ALMA scientist's assessment is that the blue items on the ALMA list need more scientists for definition (and to support implementation and verification). We observed this shortage of scientific and algorithmic specialists in several areas. For example, it was stated that there are no practicing radio astronomers on the pipeline team. Additional resources of this type to support requirements definition and impact analysis should be added.
- The method of chartering the ARDG in support of these analyses, and the degree to which the ARDG is bound to undertake these analyses should be clarified. A key question is: Should the ARDG be required to focus only on responding to programmatic requests, or do they have a more fundamental research charter to "anticipate" capabilities that may be needed or desired in the future?
- The staffing level of ARDG was stated as 0.8 FTE, this sounds light, depending on their charter and the answer to the question above.

As ALMA is now entering operations, the emphasis and prioritization of requirements across the programs needs some rebalancing.

- Up until now the emphasis of ARDG on EVLA has been appropriate; in the construction phase ALMA had a different mechanism for these sort of issues. That should change going into operations, so that ARDG is properly balanced between EVLA and ALMA.
- The existing ALMA compliance matrix has served its purpose, and it is now time to revisit each program's requirements and manage those requirements across the programs uniformly.

The CASA steering committee is composed entirely of internal NRAO people. There has been insufficient dialog with the community up until now. Creating an external CASA users group is a good idea, although we note that getting effective input from such a group will probably be harder than it appears.

- The role of the CSSC and the new CUC in the governance process should

also be clarified.

- It would be helpful to talk with some other groups that have set up users committees (inside and outside NRAO) to get some ideas about how to make them most effective.

Recommendations:

2.1 Requirements management needs a well-defined, documented, and director-approved governance process.

2.2 Requirements that include both research and development components need a sanctioned way to create a combined scientist/development team for impact analysis (as well as implementation and verification).

2.3 The emphasis and prioritization of requirements across the programs needs some rebalancing.

3. Software Development Processes

The panel found that standard best practices relative to software engineering processes were not being followed or were just not implemented by the CASA software development teams. In addition, the panel found that there was some concern over the CASA architecture and whether it can handle the tasks that are envisioned for it in the near future. Lastly, the panel found the decision making process, especially as it related to the prioritization of CASA functionality, to be opaque and poorly understood by the staff.

The panel recommends that the CASA project define, implement, and refine a formal software engineering process that includes requirements generation and management, code design and implementation, test, release, and deployment planning.

The panel also recommends that a decision making body, such as a Change Control Board (CCB), be stood up to manage changes and enhancements to the CASA baselines, and to make its decisions and schedules public, at least within the CASA project. The CCB should include a representative for the relevant stakeholders of CASA, such as Algorithm developers, architects, developers, testers, and Configuration Management.

The panel recommends adding a software architect, software testers and possibly a

software configuration management specialist to fill out a very capable development team. It was apparent to the panel that developers were multitasking, performing tasks that should be domain of specialists such as testers and configuration managers. The panel recommends that prior to any major CASA enhancements/ initiatives, the architect conduct an independent architectural review of CASA to assess which architectural features will retard or stunt future growth.

Recommendations:

3.1 The CASA project should define, implement, and refine a formal software engineering process.

3.2 A decision making body, such as a Change Control Board (CCB), should be established to manage changes and enhancements to the CASA baselines, and to make its decisions and schedules public.

3.3 A software architect, software testers, and possibly a software configuration management specialist should be added to the development team.

3.4 Prior to any major CASA enhancements or initiatives, the architect should conduct an independent architectural review of CASA.

4. Architectural Issues

The review was fairly wide-ranging and not surprisingly did not give us the opportunity to delve deeply into the architecture of CASA and its implications for the current and future system extensibility or performance. In fact, most of the review committee does not feel sufficiently expert to comment knowledgeably about the issues.

We do not want to sidetrack CASA into a major rewrite just when they are looking to go into main-line usage. But the structure and performance of CASA is potentially limiting as CASA looks to the future, and this needs to be on their radar for a review in a year or two. Underlying design limitations of CASA are especially important if NRAO is planning to scale the current system up to an HPC level. This clearly is an important area of concern that should be addressed as a high priority by the (proposed) CASA architect.

A related issue which came up in side conversations and only briefly during the panel sessions themselves is what is the future of the underlying CASAcore libraries. These are the libraries that implement the Measurement Set structure and table query system. The primary developer for these libraries is not actually at NRAO (and is not so far from retirement). NRAO's stewardship and control of these key libraries is already pretty loose. The future of CASAcore is at risk and needs a mitigation plan.

We note that NRAO has considerable in-house expertise on the design of high performance data processing software (including the NRAO staff on this panel) and expect that the future system architect will take advantage of that experience when the current CASA design is assessed.

Recommendations:

4.1 The structure and underlying design of CASA are important areas of concern that should be reviewed and assessed as a high priority by the (proposed) CASA architect.

4.2 Future support for CASAcore is at risk and needs a mitigation plan.

5. Pipelines

Data reduction pipelines are an increasingly important component of modern observatories. Pipeline processing of the raw data enable broad monitoring of the data quality by the observatory and enable scientist rapid access to the data. With the increasing capability and data rates of ALMA and the VLA, these pipelines will be essential to the scientific productivity of the observatories.

Data calibration and preliminary imaging pipelines are included in the baseline plan for ALMA operations. These pipelines are processing the ALMA Cycle 0 data and continuing development for ALMA Cycle 1. The VLA is implementing a data calibration pipeline for the purpose of monitoring instrumental performance and data quality. This overall focus in data pipelines is a very positive development which we highly commend. The VLA pipeline is a particularly impressive development in these times of limited resources; it reflects well on the dedication and capability of the individuals involved..

The present pipelines are an excellent start. The CASA team should commit to supporting development of a suite of robust, well-tested pipelines that start from

the raw data and follow through to imaging. The intent should be to cover a wide range of observation types: continuum, polarization, and spectral line. Development of the pipelines to maturity, where they can process 90–99% of the data reliably and automatically, should be a priority, despite the significant effort involved.

As outlined by the presentation, capturing and incorporating the scientific intent of the observations is an important final step. The committee is strongly in favor of holding science quality imaging as the end-goal of the pipelines. We submit that the survey presented to us, which indicated a weak interest in science quality images by the user community, was a biased sampling of the true target community for ALMA and VLA. Science quality imaging for a non-expert user should be the long term goal of the pipelines because this is how to reach out to the broad astronomy community and maximize the scientific impact of the instruments.

We anticipate that developing robust imaging pipelines for the VLA will in fact be easier than the already developed calibration/flagging pipelines. Once the flagging is complete, a reference image usable for many science projects should be relatively straightforward to generate. Such an image may not reach the ultimate instrumental limits achievable by careful, expert hand-tuning, but it will definitely be valuable for science assessment of the data. In fact, we would argue that the creation of an image is a necessary step to confirm that the calibration and data flagging were done correctly. The CASA team should not hesitate to start the development of an imaging pipeline as soon as possible.

With this emphasis on pipelines, CASA development must not lose sight of the need by the experts to use the instrument in creative ways and push the instrument to its technical limits. Thus, pipelines should not be the only path. CASA software should continue to support interactive reduction and maintain the flexibility to allow expert users to maximize the capability of the instrument. Since prototyping and development of pipelines relies on access to general-purpose tools for data processing, we think continuing support and enhancement of the general CASA tools will be a natural consequence of a focus on pipeline development.

We recommend that additional science and testing resources be allocated to the pipeline effort. Robust, verified imaging pipelines should be a short term priority for CASA as these pipelines will be dividends in instrument performance and user satisfaction, and increase the community reach of the ALMA and VLA.

"Real-time" (Correlator-attached) Pipelines

The "real-time" (pre-archive) and "SKA-mode" pipeline work is understood as correlator-attached processing, i.e., connecting a processing cluster to the correlator back-end and processing correlator output directly before archiving.

The need to parallelize the pipelines to handle the data volume implicit in this operational mode makes this capability a significant departure from the existing CASA architecture and operational mode. There are questions as to whether the underlying CASA architecture is a suitable starting point for building on to achieve this capability (§4).

Furthermore, given an already fairly full docket of near- and medium-term CASA requirements, this capability should be considered only if driven by specific ALMA or VLA requirements. We do not see any present-horizon requirements for this work, so any effort in this area is considered more forward-looking, and would presumably require proposing this to ALMA and/or NRAO for additional funding.

Up until now, it has been acceptable to limit operational modes and PI experiments based on the inherent capability and performance of the processing system. It remains to be seen if this approach will continue, or if the scientific demand will drive CASA in the direction of such pipelines. NRAO needs a systems engineering approach to the data rate problems, including limiting at the proposal end by restricting users, scaling up the ability to collect and process the data, and understanding the rate at which data can be delivered to users. There does not appear to be any model or projection of data rates expansion, and the panel feels that such a model should be developed, even if it is only a model for exploring such scenarios (see also the discussion in §4 above).

Finally, given the very preliminary state of analysis and relative lack of experience on the CASA team in this sort of parallel processing architecture, the panel is skeptical of the preliminary, rough FTE estimate presented for this capability. The panel understands that the estimate was not made via a rigorous analysis process. Any estimates should be made on the basis of a detailed exploration of the work required to employ CASA in this mode, including scaled prototype implementations. There are many big-science projects within and outside of radio astronomy that regularly process large data volumes in parallel, and the CASA team should only attempt this in collaboration or consultation with teams experienced in this area.

In summary, we see this work as low priority because the effort is ill-defined and

the need for the capability is not obvious. Those combined suggest the risk of a significant waste of resources on this project. Developing a model for future data rates would at least provide a sounder basis for understanding the need.

Recommendations:

5.1 The CASA team should commit to supporting development of a suite of robust, well-tested pipelines that start from the raw data and follow through to imaging, and should not hesitate to start the development of an imaging pipeline as soon as possible. Additional science and testing resources should be allocated to the pipeline effort.

5.2 CASA software should continue to support interactive reduction and maintain the flexibility to allow expert users to maximize the capability of the telescopes.

5.3 Parallelization of the calibration pipelines should be considered only if driven by specific ALMA or VLA requirements. The current FTE estimates for supporting parallel processing need to be vetted through more detailed study, included scaled prototype implementations.

5.4 A model of the operational modes of ALMA and the VLA should be developed in order to better project data rates and anticipate potential processing bottlenecks.

6. Visualization

Visualization tools are essential to realize the scientific potential of the ALMA and VLA. However, this is one of the most difficult software problems facing CASA due to the potential for large datasets and the wide range of user capabilities. The CASAviewer is a good platform for meeting the immediate user needs. There has been excellent progress in its capabilities. We recommend that incremental improvements in its capabilities driven by ALMA and VLA requirements continue. We recommend against exploring a major rewrite of the core architecture.

In our view, CASAviewer provides the breathing space for CASA participation in building collaborations with external groups to develop a next generation viewer. The visualization challenges facing CASA are shared by astronomical efforts ranging from the LSST to NOAO to the JWST. They are also shared by other fields and are a well recognized field of study in computer science. We strongly

encourage CASA to engage other groups in collaborations. This may require NRAO to take a leadership position in coordinating/organizing the effort; but, depending on the interests and abilities of the groups, NRAO should consider taking a secondary role in the actual effort. The key to success here is to engage the broader community in the intellectual problem, and to bring in experience and innovation from outside of NRAO/ALMA.

Recommendations:

6.1 Incremental improvements in the capabilities of the CASAviewer, driven by ALMA and VLA requirements, should continue. We recommend against exploring a major rewrite of the core architecture.

6.2 Collaborations with other groups sharing interests in visualization are encouraged.

7. Virtual Observatory

The panel supports the idea of integrating Virtual Observatory data access protocols into CASA, particularly the Simple Image Access Protocol (SIAP) V2, which will support discovery, access, and dynamic subsetting ("slice and dice") of data cubes. CASA users come from a broad range of expertise, and ALMA and VLA science involves multi-wavelength analysis. Providing VO capabilities within CASA, such that a user does not need to leave the CASA environment to perform such analyses, will be a significant added value. Provision of such capabilities in CASA can take advantage of software developed by the U.S. Virtual Astronomical Observatory (VAO) project, such as the SIAP V2 protocol, SIAP V2 reference implementations in the VAO DALServer package, a suite of pure Python VO service bindings, and the higher level VOClient package which also provides Python bindings to VO services. VOClient has in common with CASA that it is implemented in C with a Python interface, although that is no guarantee that integrating it into CASA will be trivial. The pure-Python version of VOClient may be easier to include within the CASA Python environment. VOClient development and support is funded by the VAO project with staff located at NRAO and NOAO.

The VO Simple Applications Messaging Protocol (SAMP) provides the means to utilize software outside of CASA in a seamless manner. For example, sophisticated plotting and cross-matching capabilities of TOPCAT, use of the VAO Data Discovery Tool and spectral energy distribution builder/analyzer Iris, and other visualization

tools such as Aladin and DS9, can all be used in conjunction with CASA if CASA, and particularly the CASA Viewer, is SAMP-enabled. SAMP could also be used to couple CASA with the NRAO archive access user interface. VOClient provides embedded SAMP support; the pure Python VO bindings do not (though the SAMPy toolkit could be used for this purpose).

Providing VO-compliant access to NRAO data is also a high priority, but it is not a direct responsibility of the CASA software team. However, we would expect these archive-related tasks to be implemented using CASA tools. For example, the SIAP V2 protocol supports dynamic operations on data cubes, such as subsetting, extraction of specific image planes (along any two axes of the cube), and averaging (also on any axis). These operations are already inherently supported within CASA, however in an era of TB-scale data cubes it is necessary to move to a client-server architecture providing direct access to arbitrarily large cubes stored remotely; SIAP V2 provides this capability. NRAO and the ALMA project need to determine how to facilitate utilization of their data by VO users (and other archival users). Easy access to NRAO data holdings will be appreciated by the community. Note that this is related to the production of imaging pipelines, since images are certainly the products of most general use (see §5).

The panel ranks the priorities of VO-related work in CASA as follows:

1) Provide SIAP V2 support in the CASA Viewer. Extensions to VOClient to support SIAP V2, funded by the VAO Project and implemented by NRAO staff, would make the CASA Viewer (also a C program) enhancement straightforward. Other VO services could also be exploited, such as cone search services for catalog overlays and SIAP V1 services for simple image comparisons.

2) Add SAMP capabilities to the CASA Viewer so that users can easily take advantage of other SAMP-enabled applications (Aladin, TOPCAT, VAO Data Discovery Tool, etc.). Since SAMP support is included in VOClient, this is also straightforward.

3) Provide the VAO pure Python bindings to VO services and the VOClient package as CASA add-ons, allowing users to integrate VO capabilities into CASA scripts or other Python scripts.

4) Incorporate VOClient capabilities into other CASA applications for higher-level VO-enabled tasks and asynchronous processing.

The VAO project is providing funding for 12 staff-months of effort at NRAO in time

period May 2013 – September 2014 specifically to assist NRAO in implementing VO capabilities in CASA and in deploying VO-compliant services on its data holdings. The panel urges NRAO to use these resources to address as many of the priorities listed above as possible. If the external funding proves to be inadequate for this work, the CASA project should opt for those efforts that give the most benefit for the least effort and should not divert major internal resources to this work.

Recommendations:

7.1 Provide SIAP V2 support in the CASA Viewer.

7.2 Add SAMP capabilities to the CASA Viewer.

7.3 Provide the VAO pure Python bindings to VO services and the VOClient package as CASA add-ons.

7.4 Incorporate VOClient capabilities into other CASA applications.

7.5 Focus on those VO-related efforts that give the most benefit for the least effort and without a major diversion of internal resources.

8. Low Frequency & SKA

With the opening of several new low-frequency arrays around the world such as LWA, MWA, PAPER, and LOFAR, low-frequency radio astronomy is currently the subject of significant renewed interest in the community. While the scientific potential of these new arrays is high, working at these frequencies also brings a unique set of challenges in terms of calibration and imaging. These challenges translate into increased complexity and cost in terms of the data processing and storage. Providing comprehensive support for processing low frequency data in CASA would likely require significant changes to the underlying codebase. These changes might include modifications to the imaging to better support wide-field imaging with variable beams, improved calibration routines to account for ionospheric fluctuations, and support for highly parallelized processing.

Although interesting from a research perspective and potentially useful to the larger community interested in using CASA to reduce data from these other facilities, such major modifications run the risk of disrupting ongoing efforts to support EVLA and ALMA. The panel did not feel that investing significant effort in this area was currently well motivated and would recommend against it at this

time. The noted exception would be work associated with supporting a low-frequency upgrade to the VLA assuming this upgrade does go forward. Even in the event such a low frequency upgrade does go forward, a better assessment of the necessary changes to CASA to support it would be required and such a census should be conducted before proceeding.

The panel noted that synergies and ongoing collaborations with other low and high frequency projects do obviously exist. The collaboration with the LOFAR project on modifications to the CASA imager codebase to support wide-field, wide-band imaging for aperture arrays is one obvious example. The continued support for the CASAcore libraries used by various external projects (LOFAR, MeerKAT, ASKAP, etc.) is another. While such collaborations are desirable, additional work in CASA on behalf of these projects should be undertaken only if additional development resources, whether internally at NRAO or externally from the projects themselves, can be provided. With this said, if NRAO can make small scale changes in CASA to support these other telescopes at low-cost, it should do so since this support can only benefit the broader community and increase adoption of CASA. Support for the CASAcore libraries for example should be continued.

Similar to the proposed low-frequency additions to CASA, significant work to support SKA does not seem appropriate at this time. The panel felt it would be far more profitable at this time for NRAO to focus on supporting the high data rate cases for ALMA and the EVLA. Such support would quite naturally move CASA development in the direction of SKA-level processing support. Scalability issues with CASA are likely to be the most pressing aspect requiring attention and, as was the case with potential low-frequency support, require significant refactoring of the codebase. Again, such a major overhaul to CASA should not be undertaken at this time unless driven by the needs of ALMA and the VLA.

Recommendations:

8.1 Investing significant effort in support of low-frequency arrays is not sufficiently justified at this time. The noted exception would be work associated with supporting a low-frequency upgrade to the VLA, although such efforts would need to be assessed and understood before proceeding.

8.2 Additional work in CASA on behalf of projects such as LOFAR, MeerKAT, and ASKAP, etc., should be undertaken only if additional development resources, whether internally at NRAO or externally from the projects themselves, can be provided. If NRAO can make small scale changes in CASA

to support these other telescopes at low-cost, it should do so since this support can only benefit the broader community and increase adoption of CASA. Support for the CASAcore libraries should be continued.

8.3 Significant work to support SKA does not seem appropriate at this time.

8.4 Supporting the high data rate cases for ALMA and the EVLA is important, and is a sufficient step forward for now toward SKA-scale data processing.

9. VLBI

The data processing for VLBI applications requires a suite of tasks that have not been implemented in CASA yet. These include the import and export of VLBI data, fringe fitting, and algorithms that deal with specific issues such as short coherence times. The significant effort to incorporate full VLBI functionality into CASA has been estimated with care (6 FTE years of software developers plus 1 FTE year of a dedicated scientist). However, the user base for VLBI is currently supported by alternative software packages, in particular AIPS, and the committee was not presented with any critical short term driver to migrate this user base over to CASA. Moreover, users may be inclined to resist a major shift in the reduction software in the midst of large scale programs that rely on a consistent analysis approach. As a result, we conclude that CASA development of full VLBI functionality is a low priority at this time.

There is concern that the current support structure using AIPS effectively relies on a single developer. To mitigate the risk inherent in this situation, it will be prudent to build up foundational VLBI reduction capability in CASA over time. With this in mind, we strongly encourage taking advantage of the overlap of VLBI needs with ALMA and the VLA requirements, such as basic fringe-fitting. Further opportunities for increasing VLBI capabilities in CASA may arise within the context of the ALMA phasing project for VLBI, or through future NRAO proposals for ALMA development funds.

Recommendations:

9.1 Gradually develop VLBI reduction capabilities, such as basic fringe fitting, beginning with areas of overlap with ALMA and VLA requirements.

10. GBT

There does not appear to be a strong desire for GBT data reduction and processing to be integrated into CASA at this time. The GBT reduction software based on IDL for current spectral line observations seems sufficient, and there appear to be paths to handle the challenge posed by the next generation of spectral line backends through extensions of this framework. Fully calibrated GBT image cubes may be imported into CASA through FITS, to make use of CASA viewer capabilities. One clearly identified observatory need is to enable the proper combination of GBT data with VLA observations, to provide short spacings for certain experiments. The basic CASA infrastructure to accomplish this is expected to come from the ALMA project, which is making a concerted effort to implement single dish capabilities in CASA in support of the ACA. We conclude that further additions to CASA driven solely by the GBT requirements are low priority.

ALMA is developing other spectral line and continuum single dish capabilities within CASA for support of the ACA. The first implementation for spectral lines is in-place for Cycle 1 and we expect that ALMA will be motivated to improve the capabilities as the usage of the ACA increases. In the long term, it seems logical that the ALMA/ACA will require the majority of the capabilities also needed for GBT spectral line and continuum observations. Some skepticism was expressed by GBT presenters about the quality and utility of the ACA tools currently in CASA. Nonetheless, the progress within CASA in this area should be monitored by GBT/NRAO staff with an eye toward switching to CASA at a future date in order to take full advantage of the analysis and visualization tools that will be available in CASA.

Recommendations:

10.1 Further additions to CASA driven solely by the GBT requirements are low priority.

10.2 NRAO/GBT staff should monitor CASA developments in support of the ACA, in consideration of an eventual adoption of CASA for single dish data processing and analysis.

11. Remote Processing/Cloud

As the scale of radio datasets continues to increase, the traditional NRAO data delivery paradigm whereby essentially unprocessed datasets are transferred to users for offline processing and analysis with CASA on their local systems will become increasingly untenable. With the EVLA and ALMA, data sizes are already such that single-user desktop processing can be prohibitively slow. In response to this growing size and computational cost, many institutions have begun investing in significant local processing clusters with multiple processing cores and multi-terabyte storage. Extrapolating only slightly, one can easily see a day coming when shipment of raw data to the users becomes impractical entirely. How NRAO and CASA should support users and deliver data in these not-too distant scenarios is a key question as we move into the SKA era.

In the simplest terms, NRAO will need to choose whether to provide support to users wishing to process data on their own local clusters or expand its own processing capacity and deliver more science-ready data directly. Neither option is without cost both in terms of software effort and more critically perhaps user support. Supporting a large community running CASA based pipelines on a wide variety of heterogeneous hardware platforms would require a substantial amount of user support by NRAO personnel. Expanding NRAO's own intrinsic capacity to handle user data processing would require additional hardware investment as well as additional overhead for support science overseeing the processing and verifying the quality of the outputs. In both these scenarios, the panel notes that the importance of having robust pipelines capable of automating the data processing is clear.

For the near-term, the panel did not see sufficient urgency for significant, immediate action in these areas nor the need to make any firm decision about how to handle large-scale processing. There does not currently appear to be any model for, or projection of, the expected data rates expansion. The panel feels that such a model should be developed to provide a rational basis for decisions regarding the tradeoffs between local processing, data transfers, and the utilization of supercomputer centers. NRAO needs a systems engineering approach to the data rate problems, including limiting at the proposal end by restricting users, scaling up the ability to collect & process the data, and understanding the rate at which data can be delivered to users. The panel felt that it is appropriate at the present time to perform a low-effort study of these options in preparation for the next hardware

refresh and enable the community to explore local computing options. Such a study should naturally focus on the projected increases in data rates for ALMA and the EVLA.

Given the other pressures on the CASA team's resources, providing blanket support for generic, local user processing clusters seems unfeasible. As a first step, the panel recommends specifying a standard stand-alone workstation system configuration (cpu, memory, disk configuration) that gives a reasonable price/performance return. This specification should include basic benchmarks that indicate the level of performance that can be expected on typical datasets. (Currently available specifications appear outdated; e.g., they give benchmarks for processing Cycle 0 ALMA data with only 11 antennas rather than data from the current antenna complement.) This specification can later be expanded to include a precise small cluster standard to be supported by NRAO. For clusters not adhering to this standard, NRAO would only provide general advice on running CASA processing pipeline software on these systems.

Finally, we note that as mentioned previously, significant development effort is likely to be required in order to improve CASA's scalability to the point where it can fully exploit the power of a moderate or large compute cluster. Such enhancements might include more robust pipelines, improvements to the thread-safety of some CASA core components, or even adapting some of these components to take advantage of new GPU compute hardware. In all cases, the panel felt that such work should for now only be undertaken if driven by the needs of ALMA or EVLA data rates.

Recommendations:

11.1 There is no near-term urgency for significant action in supporting remote or cloud-based computing, nor is there a need to make any immediate decision about how to handle large-scale processing.

11.2 A model for, or projection of, the expected data rates expansion should be developed to provide a rational basis for decisions regarding the tradeoffs between local processing, data transfers, and the utilization of supercomputer centers.

11.3 It is appropriate at the present time to perform a low-effort study of the options for supporting high data rate operations in preparation for the next hardware refresh.

11.4 NRAO should recommend a standard stand-alone workstation system configuration (cpu, memory, disk configuration) that gives a reasonable price/performance return. This specification can later be expanded to include a standard small cluster configuration.

11.5 Work to improve CASA's scalability should for now only be undertaken if driven by the needs of ALMA or EVLA data rates.

12. Questions for the Panel

Below are specific questions that were posed in the charge to the panel. Rather than organize our report around our answers to the questions, we briefly summarize our response and indicate where the topic is discussed in the document above.

1. *Are the CASA development priorities clearly specified and appropriate for the short term (~2 year) needs of ALMA and the Karl G. Jansky Very Large Array (VLA)?*

The processes used for setting priorities, tracking requirements, and resolving conflicts between projects all need improvements to ensure that the requirements of these projects are captured (§2, §3). Architectural issues that limit the future extensibility and/or performance of the system should be addressed by the (proposed) system architect (§4).

2. *CASA is presently charged with providing both a traditional individual researcher data reduction package and a pipeline facility. We believe that Pipelines, although still in their early deployment phase, are key in making pre-reduction of data realistic reducing internal staff effort, while greatly facilitating access to radio-instruments by non-experts. Is the balance between these two appropriate now, and how strongly should it change in the future?*

We strongly support a focus on pipelines in the future, especially imaging pipelines that produce products more readily usable by the wider community outside radio astronomy (§5).

3. *CASA currently executes on laptop through small cluster computational systems, but not in the cloud or supercomputing centers. Should this strategy be revised? If so, what form of external collaboration should NRAO pursue?*

The panel did not see sufficient urgency for significant, immediate action in these

areas nor the need to make any firm decision about how to handle large-scale processing (§11). It is important to first develop a model of the expected data rate expansion to provide a solid basis for decisions regarding the tradeoffs between local processing, data transfers, and the utilization of supercomputer centers (§5). It is also appropriate to perform a low-effort study of processing options in preparation for the next hardware refresh. Such a study should naturally focus on the projected increases in data rates for ALMA and the EVLA.

For local user processing, the panel recommends specifying a standard stand-alone workstation system configuration. This specification should include basic benchmarks that indicate the level of performance that can be expected on typical datasets. This specification can later be expanded to include a precise small cluster standard to be supported by NRAO. For clusters not adhering to this standard, NRAO would only provide general advice on running CASA processing pipeline software on these systems.

Significant development effort is likely to be required in order to improve CASA's scalability to the point where it can fully exploit the power of a moderate or large compute cluster. The panel felt that such work should for now only be undertaken if driven by the needs of ALMA or EVLA data rates.

4. *Two of NRAOs telescopes are not supported by CASA (the Robert C. Byrd Green Bank Telescope and the Very Long Baseline Array). Given that the observing communities have data processing solutions, should CASA be extended to support these telescopes? If so, are the requirements and costs understood?*

CASA development of full VLBI functionality is a low priority at this time, although development of basic VLBI capabilities to support ALMA and the VLA (e.g., fringe fitting) should allow some mitigation of the risk associated with dependence of VLBI on AIPS (§9).

There is not a strong need for CASA to include the capabilities to support GBT data processing at this time (§10). However, GBT/NRAO staff should monitor the progress within CASA in developing tools to support the ALMA/ACA single-dish data.

5. *There are a number of radio telescope projects where NRAO and those projects could potentially enter into mutually beneficial collaborations, notably low-frequency radio interferometry and VLBI. Do you have recommendations for strategies or partnerships that should be pursued?*

CASA should focus on supporting the VLA's low frequency projects (§8). Synergies will likely exist with other projects, but extension of the work should rely on additional resources. Work on high-data rates with ALMA and the VLA will naturally position CASA/NRAO for future SKA activities; specific prep work for the SKA is not appropriate at this time.

6. *CASA presently supports a traditional data processing model, where the raw data sets down on disk several to many times during its processing. This will not scale to very high data rate interferometers (SKA). Should the NRAO pursue alternative data processing strategies on our existing arrays to position NRAO to effectively participate in next-generation telescope construction projects? Would enabling CASA to process data real-time from the ALMA and/or VLA correlators (capable of 1 and 16 GB/s respectively) be a good test platform as well as enabling important science for our user communities?*

This capability should be considered only if driven by specific ALMA or VLA requirements. We do not see any present-horizon requirements for this work, and any effort in this area would presumably require proposing this to ALMA and/or NRAO for additional funding. There does not appear to be any model or projection of data rates expansion, and the panel feels that such a model should be developed. The panel is skeptical of the preliminary, rough FTE estimate presented for this work (§5).

7. *Visualization and high-level (e.g., model fitting) data analysis have received relatively little attention compared to calibration, flagging, and imaging. What strategies should be chosen in this area? Are there existing packages that could be adopted?*

A collaborative approach that engages the wider community is appropriate for developing enhanced visualization capabilities (§6).

8. *There is considerable interest in the VAO development community to promote access to cubes, including VO enabling CASA. What is the priority of this work in CASA compared to other initiatives? Does the answer change depending on the funding source?*

Access to VO resources from within CASA will be useful, as will providing CASA/NRAO data products through VO protocols (§7). We assume most of this work will be performed using resources outside of the CASA project and do not recommend diverting major internal resources to this work.