July 15, 1976

VLA COMPUTER MEMORANDUM NO. 139

VLA POST-PROCESSING: AN INITIAL DISCUSSION AND PROPOSAL

W. R. Burns

I. Introduction

The VLA will produce an enormous mass of data which, because of its size and its relation to the various desired observables, will be virtually unintelligible. The calibrating, editing, and Fourier transforming of these data represent a significant computation effort. The synchronous and asynchronous computing systems are now being developed with VLA construction funds to carry out these tasks. However, particularly in the polarization and spectral-line cases, the maps produced by these stages of the data manipulation are a beginning rather than an end. Before the maps may be interpreted in astrophysical terms and published, the user must be able to comprehend them, to extract global and detailed information from them, and to describe and display them in formats suitable to the interpretations. These stages of the data manipulation also represent a significant computation effort and will eventually constitute the major portion of the VLA data processing. (The experience at Westerbork is that editing, calibrating, and first-pass mapping now constitute only 20-25% of their computing load and that this percentage is continuing to drop.) The asynchronous computer system, as presently planned, will be able to support only a very small amount of post-initial-synthesis processing ("post-processing"). Thus, if the VLA is to reach its full potential, significant additional facilities must be made available.

In the first part of this memorandum, we describe some general considerations regarding these additional facilities and outline some of the questions which will have to be answered. In the second part, we propose a specific plan to use NRAO's IBM 360 as an interim post-processing center. The intent of proposed project is to provide some limited post-processing capability prior to 1980 and to put ourselves in a position around 1980 to make an intelligent and informed decision regarding the later stages in developing a post-processing center including the replacement of the IBM 360.

II. VLA Post-Processing Considerations

A. Definition of post-processing

Post-processing is defined as VLA data processing which is done after the observation and initial reduction session. The "site product" has been defined to be a single map for each data channel based on well-edited and calibrated data. When the full VLA is in operation, it is assumed that this site product will be produced in an essentially "pipeline" fashion during the observation time of the current and the following observer. Post-processing is, thus, any data processing beyond that required to produce the site product (or that portion of the site product which can be produced by the on-site computer systems while keeping up with the incoming data flow). The categories of data manipulation which will normally comprise post-processing include:

1. Additional editing and calibrating, if required,
2. u-v plane data display,
3. further map synthesis,
4. map correction,
5. map analysis and interpretation, and
6. map display for user and for publication.

Post-processing will be required for many reasons including, primarily, (1) that there will not be sufficient computer power to complete the processing required for publication during the limited time allotted for processing at the site and (2) that most observers will require periods of weeks, months, or, even, years to digest and to interact with the large quantities of data produced by the VLA.

B. Why the NRAO?

There will have to be close interaction between the on-site facilities and any post-processing center. It would indeed be difficult if the two were under separate management. The NRAO will continue to operate a good-size computer center to provide computing for the NRAO staff and to support NRAO's other telescope systems. It would be more economical for this center to be a portion of the VLA post-processing center.

C. How would this post-processing center relate to NRAO's present center?

The overhead cost of a processing center makes it very unattractive for NRAO to support two centers each of which is separated from the observing sites. Thus, the proposed center would replace NRAO's present computation facility. The present facility has three special tasks which NRAO should continue to support.

1. The daily support of the telescope systems in Green Bank is tied closely to the IBM 360. However, if the present trends continue, there should be adequate computer power at the Green Bank site by 1980 to support almost all single-dish work. Observers may still choose to use the main processing center for either routine or special single-dish data reduction procedures, but they would not normally be obliged to do so. The Green Bank interferometer data reduction is not evolving in this direction and represents a special case.

2. The IBM 360 is heavily used by the NRAO staff for computing other than data analysis (e.g., theory, telescope design). This support should be relatively easy to maintain even if the VLA post-processing facility is not located in Charlottesville. The small amounts of input/output required by such computing tasks can easily and economically be handled by remote terminal techniques.

3. The IBM 360 is also heavily used to support the VLB processor. Here again, however, the trend is to simplify the processing by placing more and more of the computational steps into the minicomputers attached to the processor. Support for VLB post-processing will probably be required at the VLA post-processing center. Some significant economies could be realized by combining these post-processing areas since VLBI and the VLA share many similar mapping, map analysis, and map display problems.

D. Where should NRAO's post-processing facility be located?

Currently, there appear to be three reasonable locations for the facility: Charlottesville, the VLA Site, and Albuquerque.

1. Although Charlottesville offers the attraction that it is the current location of NRAO's scientific staff and major computer center, the large distance between Charlottesville and the VLA site seems to make it an unattractive location for the post-processing center. The interdependency between the on-site and post-processing facilities will force considerable movement of people and data between the locations. Travel between the site and Charlottesville would be time consuming, onerous, and expensive both for NRAO staff and, particularly, for VLA users. Of equal importance is the need for on-site and off-site personnel to communicate and to remain functionally part of a single organization. The current VLA (and Tucson) experience shows that close proximity is required.

2. The VLA Site is not an attractive location for the post-processing center because the expense of supporting a major computer center involving many simultaneous users at such a remote location is too high. The costs of people support (e.g., motels, restaurants, transportation, etc.) and of maintenance and rapid repair of a large variety of independently manufactured hardware devices rise very sharply with the remoteness of the location.

3. Considering these points, we feel that an Albuquerque location would be attractive. It is within easy driving distance of the site and almost all observers in route to and from the site will pass through Albuquerque.

It should be stressed here that this discussion is not comprehensive and has omitted many important factors which affect the selection

of the location of a post-processing center. In particular, the development of the proposed 25-meter telescope or the proposed VLBI array could have some impact on this question.

E. How much post-processing is required?

The required post-processing capabilities depend on the answers to the questions:

1. how much processing will the observer be able to accomplish during his limited on-site processing time,
2. what data processing is required for publication, and
3. how much of the data processing should NRAO provide?

Unfortunately, because of our lack of experience with VLA data processing and because we have not yet done the analysis of these areas properly, we are not now able to give definitive answers to these questions. The answer to the first question depends heavily on the as yet unknown performance characteristics of the VLA hardware. Since there appears to be a wide variety of opinion on these considerations, we will address these questions only briefly in this memorandum.

1. Present plans call for attaching a very substantial amount of computer hardware to the DEC-10. The amount of hardware which may be so attached is limited not only by budgetary considerations but also by considerations of diminishing returns. At some stage in the growth of a complex system, a point is reached where the increased capacity resulting from additional pieces becomes unattractively small. The present plans call for expanding the DEC-10 system close to this point. Given the planned hardware, the amount of computer power which will be available to the user for map making and map processing will depend on how much computation is required to perform calibration, editing, data bookkeeping, array monitoring,

and other necessary functions. This is simply not known at this time. Both attractive and very frightening scenarios have been envisioned. For the present, however, we must simply assume that the site product (defined earlier) is realistic and attainable, but that negligible additional capacity will be available.

2. The subsequent processing required for publication naturally depends very strongly on the particular science being done and on how long it takes for the observer to extract and absorb interesting information from the mass of VLA data. It is impossible to predict what kinds of science will be done during what fractions of time. Based on the Westerbork experience, it is reasonable to assume that, in computer effort, the site product represents roughly 25% of what will be required for average observations. However, estimates of this percentage vary from nearly 100% to ε. Considerably more work and experience are required before a better answer can be given.

3. For its present instruments, the NRAO has mostly provided rather extensive post-processing facilities for a number of standard reduction methods. The amount of computer power and software assistance provided by the NRAO was understood to be fixed, however, and, if these proved not to be sufficient, the user was expected to turn to his home computing facility. The small data rates from the present instruments have made such an understanding reasonable and, for most users, practical. However, even on the present instruments, many users conduct their observations in a less than optimum fashion in order to be able to use existing, NRAO-provided software. With the complexity and high data rates of the VLA, software construction and

execution become very time consuming and expensive and, probably, beyond the reach of most users. The scientific output of the VLA will be diminished if the NRAO does not provide adequate facilities for all stages of the data manipulation and analysis.

III. Proposal for an Interim Post-Processing Center

At this time we have neither the money, the manpower, the knowledge, nor the need to propose the immediate commencement of construction of a major VLA post-processing center. Thus we propose, for the moment, a small-scale, relatively inexpensive effort in the post-processing area designed to enhance our knowledge and experience in such matters without significantly detracting from the manpower and funds needed at the VLA site. This go-slow approach has some other advantages. It allows us enough time to find out how well the VLA and on-site computer facilities will perform in order to determine the magnitude of the post-processing problems. It allows us to better understand what role, if any, an optical processor might play in an off-site facility. And it allows computer technology to advance considerably before we make major investments in a full post-processing center. (In fact, it is estimated that by 1980, it should be possible to increase our processing capacity by a factor of ~10 with no increase, beyond inflation, in our rental costs.)

A. Proposed manpower and equipment

The proposed manpower plan calls for one "astronomer" and one "programmer" beginning in or before 1977. During 1978 a second astronomer and a second programmer should be added to the project and a third programmer should be added in 1979. The software effort in the asynchronous group is very large by NRAO standards and we feel that it would be unwise to duplicate such an effort at this time. However, we feel that this proposed manpower level is the minimum required in order to accomplish the goals of this proposal. By programmers, we mean thoroughly competent, but not necessarily astronomically-oriented computer programmers. By

astronomers, we do not mean members of the NRAO Scientific Staff, but rather, astronomically trained individuals of PhD caliber hired as regular members of the NRAO technical divisions specifically for this project. We hope and expect that members of the NRAO Scientific Staff would also contribute their suggestions and their expertise to this project on a part-time basis.

The proposed hardware plan revolves about the NRAO (Charlottesville) Computer Development Laboratory. The minicomputer system of the Development Lab will be directly interfaced to, and used in conjunction with, the IBM 360/65 system. We propose to continue to make yearly improvements and alterations to the Development Lab systems at approximately the rate which has become customary. In addition to the general developments, a number of specific devices will be required. In 1977 we propose to add interactive graphics capability in the form of a television-like image device together with subsidiary display and control devices. In 1978 we propose to add to the system an array processor which would be used principally for the taking of Fourier transforms. Thus, in the 1977-1978 time frame, we propose an additional capital expenditure of $150K. The required additional capital funding for the project after 1978 is not yet clear and will depend on the progress at the VLA site, the progress of this project, and the rate at which the VLA becomes a heavily used scientific instrument. Conservative, but not improbable, expenditure levels would be $75K in each of 1979 and 1980, with a truly major capital development in 1981.

We do not anticipate any significant increase in rental for the principal portions of the NRAO computer center. Although the present basic contract on the IBM 360 will terminate during 1978, we doubt that a sufficient number of the important questions will have been answered by then to allow us to effect major hardware changes. Some increases in disc and/or core capacities may be expected without substantial increase in rental when the present basic arrangement is renegotiated and extended.

B. Tasks and areas of investigation for the proposed project.

It is easy to identify a number of specific tasks which must be accomplished in order to carry out this proposal. These tasks do not constitute the main goals of the proposal but rather are the principal tools with which the main scientific and technical questions may be approached. Among these tasks are:

1. to interface VLA data to the present NRAO computer packages particularly the Green Bank interferometer package,
2. to link the Computer Development Lab computer to the IBM 360 (the major part of the software for this has already been developed),
3. to develop an extensive display capability for the Dicomed/television display system,
4. for some selected experiments, to link the 360 to another computer center, and
5. to develop VLA-oriented data simulation routines.

In carrying out these tasks, the NRAO will acquire significant new knowledge and experience in areas relevant to the VLA post-processing problem. However, these tasks are only the tools with which the important questions are to be approached. Those questions, discussed below, lie in both technical and scientific areas. We imply no ordering of priorities by the order in which these questions are treated.

1. The principal area of study will almost certainly be the determination of algorithms to carry out known data reduction steps. With the low data rates of present instruments, inefficiency, incompleteness, and redundancy have been tolerable in the data reduction packages. For the VLA, however, the algorithms must be efficient as well as scientifically correct and psychologically pleasing. Considerable experimentation with the algorithms will probably be required.

2. There is a wide range of questions about numerical analysis of map and u-v plane data which need further study both with artificial and with real VLA data. These questions include:

   (a) the effects of cross-overs and weighting in the synthesis process,
   (b) the effects of aliasing on VLA maps,
   (c) the effects of map grid spacing on map displays, map manipulations, source modeling, etc., including the use of the so-called FFT interpolation procedure,
   (d) the methods by which the dynamic range of the VLA may be extended including self-calibration and continuum subtraction,
   (e) the effects of finite bandwidth and integration time on the maps and methods by which these effects may be reduced,
   (f) the effects of the sidelobes of the VLA synthesized beam pattern including methods of source subtraction and inner sidelobe suppression, and,
   (g) the methods by which source components may be modeled and removed from the data.

3. The effects of such things as atmosphere, antenna polarization and total power beam patterns, and various other instrumental properties on u-v plane and map data must also be investigated. In all probability, such studies will be carried out by site personnel primarily. However, the proposed off-site project should provide some support for these studies.

4. An important technological area for study is the determination of a reasonable balance in the use

of dedicated minicomputers and large, more general-purpose systems. The increase in power and decrease in cost of minicomputer systems has made their use attractive. However, one has not seen any large computation centers replace a large general purpose machine solely with minicomputers. It seems clear that the VLA post-processing center will use both a large machine and minicomputers. The degree to which the large and small machines are linked and the division of tasks between the machines remain to be determined.

5. Another problem is the determination of the relative merits of conventional, machine-independent software and special-purpose, machine-dependent software. In a real-time/on-site system in which the hardware turnover is small, the machine-dependent approach is reasonable and tends to produce more efficient and somewhat more flexible code. Such an approach has been, properly, adopted for the asynchronous computer project. However, large data processing centers normally involve hardware from a variety of manufacturers and experience a higher rate of hardware turnover. For these reasons alone, large software packages at such centers tend to be made as machine independent as possible. This approach appears attractive for the VLA post-processing center not only for the above reasons, but also for the likelihood that users will wish to run portions of the post-processing programs, particularly map display routines, on their own computer systems. This appraoch must be evaluated carefully to determine tolerable levels of machine dependency, loss of efficiency, and inflexibility.

6. It is possible to do number crunching by purchasing time at existing large computational centers such as that at Lawrence Berkeley. Since these centers recover only their operational costs from users, they offer a computing service at 1/5 to 1/2 the cost that would otherwise be possible. However, to separate the computational aspects of the VLA data processing from the I/O and display aspects is not easy. In addition, the costs of transmitting large amounts of data between the centers might substantially offset the reduced cost of the computing. In summary, we do not know to what extent links are attractive for the VLA problem and we are not likely to find out without trying.

7. If the present study of the proposed optical Fourier transform processor produces encouraging results, then a large number of questions will have to be answered. Such a processor should fit fairly easily into the "pipeline-processing", one-user environment of the asynchronous computer system. However, the methods by which an optical processor might be most profitably used in the less organized, multi-user environment of the post-processing center are less clear. Some prototyping and simulation will certainly be required.

8. The manner in which non-NRAO users interact with the post-processing facility is another area which requires more study. The size of the VLA data base will force most users to employ the post-processing center even to obtain answers to apparently trivial questions about their data. Should we force users to travel to the post-processing center, as we do now, in order to obtain answers to these simple

questions? One method to avoid excessive travel (and/or excessive demands on NRAO personnel) would be to link the post-processing center to terminals at some universities. The use of such links would have to be limited strictly to the manipulation by NRAO-supplied software packages of data taken on NRAO-supported telescopes. A number of questions exist concerning the policing of this limitation, data rates, access methods, remote terminal oriented software, etc.

IV. Prospective

This memorandum does not present a definitive solution to the VLA post-processing problem. Rather it represents a first step at defining the problem and proposes a conservative initial project designed to provide some of the information needed to determine a definitive solution. Succeeding memoranda shall address the various specific areas of investigation.