SMA Technical Memorandum

Title Data Reduction Computing for the SMA
Number 99
Date 15 April 1996
From Colin Masson

Summary

An analysis of our data reduction load and benchmark testing of DEC Alpha machines shows that the data processing needs of the SMA can be met by a small group of the fastest current workstations. There is no need for any single larger machine, and the processing power can be added in a modular fashion, as required.

On the assumption that iterative calibration and imaging are performed about 6 times on a subset of 10% of the data, then 2 Alpha processors with a total 30 Gbyte of disk could handle this task. If we then make maps of every single channel, then the averaged load requires 6 Alpha processors with about 100 Gbyte of disk space. For individual reduction, a workstation with 2 Gbyte of disk is the minimum to handle 10% subsets of the data.

Backup storage does not have to be very fast, but should be large and convenient. 0.2 Terabyte is sufficient for 2 weeks of data at 4 bytes per point, and is a minimum requirement. If we condense the data to 2 bytes per point, then 2 Terabytes will handle about 1 year's data. We should aim for a jukebox handling at least 1 year's data with an access time of < 2 hours for any 4 Gbyte dataset, while handling a sustained throughput of 8 Gbytes input per day (12 Gbyte max.), and 24 Gbytes output per day (average of 3 downloads of each complete dataset).

Introduction

This memo is an update of Mark Reid's memo of July 1991. I present some current benchmarks for data reduction on our latest 266 MHz Alphas, such as rgalp0. I then compute the required capacity for CPU power and data storage to keep up with the SMA data, and discuss some possible configuration issues.

In making the calculations below, I have not explicitly put in safety factors for algorithmic complexity, but instead, have calculated the requirements for doing the work on every channel of every dataset. Since most observations will probably use less than 1000 channels, this represents a substantial safety factor. In the scaleable processing scheme described, it is easy to add more capacity.

Data Rate

At maximum specified data rate, the SMA will produce a full dump of data every 10 seconds. That is 15 baselines x 6144 channels x 2 sidebands = 184320 complex numbers or 1.474×10^6 bytes, assuming that we use simple 4 byte floating point numbers. For a full 8 hour track of a typical observation, this dataset will be about 4 Gbyte, with a maximum of 12 Gbytes/day if we observe at 100% efficiency. For the rest of this memo, I assume that we will operate at an average 70% efficiency, or 2 x 4 Gbyte datasets per day.

The sampling rate of 10 seconds is not required by imaging, even at the longest baselines of the A configuration, where 20 seconds would probably be adequate, but may be helpful for self-calibration. At the smaller configurations, 1 minute integrations would be fine. While we may often take advantage of these longer integrations, particularly in the choice of data subsets for calibration and analysis, the processing system should be sized for the worst case.

A significant amount of data compression will usually be possible, since our data are very noisy. With a system temperature of 100 K, the noise level in a 3 MHz channel in 10 seconds will be 0.18 K, or 2.7 Jy,

so 2 byte numbers would be fine for signals up to about 180 K, or 2700 Jy, without doing anything clever at all, and most observations could be fitted into 1 byte numbers. I assume that we will store all data as 2 byte values on backing storage, but expand them to standard 4 byte values when they are restored to disk. Since our data are random, this compression will not be achieved by any standard software, but must be performed explicitly by SMA archiving programs.

In sizing the computer system, I have chosen to use the worst-case 10 second sampling, though in many observations we will use longer periods. The computers must be capable of handling the peak loads, and we need a substantial safety factor in any case.

Mapping

As described by Mark, the dominant cost is in the final creation of maps of many channels. Operations such as self-calibration will be performed on a much smaller dataset and will be negligible in the big picture, although they will require that much of the data be accessible on-line for a long time while the data reduction is carried out. In doing this calculation, I have worked out the cost of making clean maps of every channel in both sidebands. Many observations will do less than this, typically by a factor of 10, but this should be considered as a safety factor to cover the extra cost of iterative data reduction..

Maps in configurations B, C, and D will easily fit in 256 x 256, while A could require 512 x 512 and mosaiced A configuration maps might need 1024 x 1024. The datacubes for 12288 channels are 3.2, 12.8, and 51.2 Gbyte, respectively. It is extremely unlikely that we would make a mosaic of all 12288 channels in the A array, so I will ignore the 51.2 Gbyte requirement, but it is apparent that we will commonly need to be able to handle datasets of several Gbytes in size. Since sensitivity is a serious limitation in A array, it is unlikely that we will often make 512 x 512 maps of all channels. Realistically, 3.2 Gbyte is something to plan for as a standard, while we might need to be able to handle larger sizes occasionally.

Software organization

Much of our computing task would be made easier with good software organization. Some of the possibilities are:

Calibration
Disk space may be a limitation during the relatively slow, iterative process of calibration. This will be eased if the software can easily keep the small flagging and calibration tables on disk long term while parts of the whole dataset are swapped in and out from time to time. In a sense, the flag and calibration tables are the output of this process, while the data to which they are applied can easily be retrieved from long term storage when needed. This type of operation is possible, but difficult, in AIPS, where the calibration tables are bound to the datafiles, are lost if the datafile is deleted, and are usually saved along with a whole new copy of the data.

Maps
Similar considerations apply to multiple copies of maps. The ridiculous distinction in AIPS between multichannel maps and datacubes should be avoided, and the physical transposition of datacubes could probably be done away with, just as UVSRT has been consigned to history. With a little bit of care in programming, it should be possible to keep just a single copy of the map dataset.

Data Subsets
The software should make it easy to work with subsets of the data during the exploratory phases, followed by reduction of larger amounts of data when the parameters have all been determined. Typical subsets may include averages over several channels or time samples, or selections of channels. After working with some initial channels, the user should easily be able to bring down a new bunch of channels add them to the existing ones and process them the same way. This would require replacing fixed size

datasets, where space is allocated for all channels at the start, with expandable ones where the channels are added as required and there is a simple index in the header for the programs to look up the physical locations of the channels.

By making these types of operations easier, we will reduce the burden on the astronomers at the same time as we reduce the burden on our computing system.

Benchmarks

I have run some benchmarks on rgalp0, a 266 MHz DEC Alpha machine, with a continuum dataset, using MX. In most cases I used 17874 visibilities to match Mark Reid's earlier tests, but I also ran some larger datasets. The results are shown in the Table below:

Size	NITER	Major Cycles	Visibilities	CPU/channel	CPU/track	Real-time/track
256	0	0	17874	1.5 sec	5 hr.	6 hr.
"	1000	2	"	11	37	44
"	1000	9	"	37	123	148
512	0	0	"	2.7	9	11
"	1000	2	"	31	103	124
"	1000	10	"	53	177	212
256	0	0	253000	11	37	44
"	1000	2	"	37	123	148
"	1000	9	"	118	393	472
512	0	0	"	12	40	48
"	1000	2	"	44	147	176
"	1000	9	"	157	523	628
512	10000	4	253000	120	400	480

Notes:
1) CPU/track is the time required to MAP and CLEAN all the channels.
2) Real times were about a factor of 1.2 greater than CPU times on a lightly loaded machine.
3) The CPU times are about a factor of 13 faster than the Convex C1 tested in 1991,
or about 100 Mflops vs. 6 Mflops.

Required Computing Power

Production Processing

This applies to final production of all maps once the parameters have been determined from a small sample of channels. The load will vary seasonally, as we move between our small (B, C, & D) configurations and the larger (A) configuration. I assume that B, C, & D data will all be analyzed with 256 x 256 maps, while A array uses 512 x 512. While larger mosaics may need 1024 x 1024, I assume that they will have correspondingly fewer channels, so the 512 x 512 x 12288 (channels) is the largest case considered. I assume that we will cycle regularly between configurations, spending 25 % of our time in A configuration, and I make no allowance for the lower probability of successful observations in A array, or the likelihood that lower sensitivity will reduce the number of usable channels.

The closest case to our B, C, & D arrays is the second row in the table, with a CPU time of 11 seconds, and a total processing time of 44 hours per track. To keep up with 2 observed tracks per day, then, we would need 4 Alphas if we map and CLEAN every channel. Since we need 10 Gbyte per project, if each of these machines worked on one project at a time, it would need 10 Gbyte of disk space. Mapping of a subset of 1000 channels could be done in only a few hours.

In A configuration, we might need 512 x 512 maps which are 3 times slower, and we would need 12 Alphas to keep up with 2 tracks per day, each with 20 Gbyte of space in a worst case. If we size our capacity for the A configuration, then we should have plenty to spare in the smaller configurations. The average capacity required is (4 x 75 + 12 x 25)/100 = 6 Alphas, if the computation load is evenly spread.

The most important piece of information from this analysis is that complete mapping of even a large dataset is well matched to a single fast, well-equipped workstation. And the time will only get shorter in two more years as these machines develop. We do not need any large mainframe or multiprocessor machine to cope with these tasks. We can therefore work with a conveniently distributed system, rather than requiring a single central computer, and we can easily adjust the amount of power by adding machines as required.

Calibration and Image Analysis

The stages of calibration and image processing are very labor intensive and require a number of screens for observers to sit in front of, as well as a large amount of disk space to hold the data. Much of this work may be done on each observer's individual workstation, but I will attempt to estimate the resources required if the SMA has to provide all computing centrally. During calibration, where there may be relatively frequent accesses of the backup data store, it may be convenient to work on a central workstation, even if the data are displayed in the observer's own office.

Disk space may be the biggest problem during the calibration phase. If we need to keep 2 weeks of UV data on moderately fast access backup (2 hour max.), then we need about 100 Gbyte of space in an on-line jukebox of tapes or CD-ROM's. The on-line disk requirements are more modest if we assume that the software is designed to facilitate working with data subsets. If I assume that 2 weeks (28 projects) will be calibrated simultaneously, with a 25% being worked on each day, and that 10% of the data will be used for calibration, then I find that we need 28 x 0.4 Gbytes for UV data and 28 x 0.6 Gbytes for two simultaneous sets of 256x256 maps. This is a modest total of 28 Gbytes. The processing power to make 25% of the maps every day is 28 x 1/4 x 1/10 x 44 hours = 31 processor-hours per day. This corresponds to making each set of maps up to 7 times over during the calibration period, depending on how much CPU is needed for the calibration and flagging. Two Alphas with 30 Gbyte of disk between them could handle this load, but we would need remote logins to give the needed 7 screens. Note that the CPU power scales with the number of iterations of data reduction, while the data storage depends on the elapsed time for analysis.

The above example is a little bit unrealistic, since most calibration is done over a longer time with a smaller duty cycle, but the resources should not be much different provided that we can easily roll data in and out of backup storage without losing the calibration and flagging information. A backup store with access to 1.5 Terabyte would give access to a whole year's data if it is stored in 16 bit precision. If some of the load is taken by individual workstations, we can see that they would need about 2 Gbyte of disk (1 Gbyte as an absolute minimum to hold the basic datasets) and processing power of about 1/4 Alpha to map a 10 % subset in 4 hours.

Image analysis has similar requirements of modest CPU power, but fairly large disk space. If we assume that final images will consist of 10% of the 12000 available channels, then we need about 0.3 Gbyte for each image set being studied. This may well be done largely on each user's desktop machine, but for the largest projects, we might benefit from the ability to roll subsets of data in and out of backup storage easily during this phase. If we really make maps of all channels in a significant number of projects, then we will need machines with several times 3.2 Gbyte of disk for studying these images, and we would probably want space for them on backup storage also.

Possible Computer Configuration

For computational efficiency, we need close coupling between CPU and the disk where data are stored. It is therefore important that each machine should have a sufficient complement of disk. The best

arrangement may be to have clusters or multiprocessor systems. For example, a dual processor Alpha with 30 Gbyte of disk would be a good basic unit. Such a machine could handle all of our calibration and iterative preliminary imaging, or could process a complete set of 12288 channels at 512 x 512 in 22 hours of real-time. The needs described above could be handled by a set of 4 such dual-processor machines on average, plus another for display and image analysis. Since our load may not be perfectly smooth, it would be wise to have at least one more to handle peaks, for a total of 10 CPUs and 150 Gbyte of disk. This should handle our average load, but may fall behind during peak times.

Data storage on disk should then be sufficient for most analysis and display, and we need a medium speed backup store (< 2 hours access), with a capacity of 1.5 Tbyte to keep a whole year's data on line. This access time should be achievable, while handling an input rate of 8 Gbytes per day and an output rate of 12 Gbytes per day (average 3 downloads of each dataset). For studies of long term array performance and calibrator variability, the storage system should be able to provide a total of 10 Gbyte of data from 100 random locations in < 1 day, again while handling the throughput given above.

Taiwan and Hawaii

We must also consider how the processing is split between Cambridge and Hawaii. There will certainly need to be 2 copies of the data store, but the overall CPU time will not increase. I expect that most of the CPU power will be in Cambridge, but the inefficiency of having 2 sites may require extra machines, for a total of 12 CPUs.

When Taiwan's 2 antennas come on line, the data rate will double, but the mapping task will not be much affected. So we need more disk and backup store, but not much more CPU. The total size of each backup storage unit will increase to 3 Tbyte, and the disk space on all machines will also double.

Conclusions

A data reduction facility with at least 12 fast Alpha CPUs, each with 15-20 Gbyte of disk, should be well able to keep up with our averaged processing needs. A good configuration would be in small clusters or multiprocessors with 2 CPUs sharing 30 Gbyte of disk, with a fast link to the backup jukebox. These machines should be split between Cambridge and Hilo.

The backup jukebox should keep at least 1.5 Tbyte easily available, though access times up to 2 hours would be adequate. Two such systems should be provided, one in Hilo and one in Cambridge. Possibly a third in Taiwan will also be needed.

Software infrastructure is critical. I have assumed that data are archived in 2 byte form, and that the data reduction software avoids data redundancy and facilitates working with data subsets. If these are not possible, than the requirements, especially for disk space, increase several times over.

This analysis has no explicit allowance for repetitions of data reduction, since it is assumed that these will usually be done on partial datasets. However, there is a safety margin built in, since many observations will not need to map and CLEAN all channels.

Since current workstations are about the right size for SMA observations, we should plan to acquire a sufficient number of these workstations, waiting as late as possible to get the best performance per dollar, and gradually adding capacity as required.