The Impending Data Distribution Catastrophe

S. G. Shepherd and R. A. Greenwald

Applied Physics Laboratory Johns Hopkins University

abstract. The increased number of SuperDARN radars, the development of STEREO-mode operation, and the increased use of high temporal resolution data are rapidly bringing the data distribution institutions to their knees. The data throughput has increased by factors approaching four and additional increases are likely. The daily throughput may exceed 5 GB in some cases. Ultimately, we will be limited by our ability to push bits through the system. Within this context, Exabyte tapes are proving to be increasingly cumbersome and do not have the long-term reliability that is necessary considering the total data storage requirement. Our hope that we may eventually be able to go to DVD technology may no longer be the best or most cost effective solution. In this context, we must consider ways in which we might reduce our storage demands, including approaches to limit data collected at the sites. We must also consider cost, reliability, and manpower factors in our data distribution. In this context, we should discuss how we are currently processing data at APL and suggest a new alternative for data distribution.

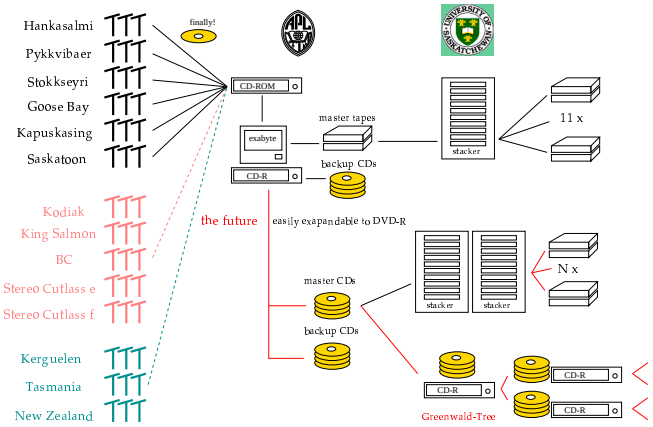

Figure 1: Diagram of the data distribution process at APL and Saskatoon. The radars in the southern hemisphere are not being sent to APL and APL no longer makes master Exabytes tapes, but rather mater CD sets.

1 Introduction

The SuperDARN groups at the Johns Hopkins University, Applied Physics Laboratory (APL) and the University of Saskatchewan, Saskatoon are responsible for distributing the collective data from all of the SuperDARN radars in the northern hemisphere. Figure 1 diagrams the data distribution process. The original CDs recorded at the radars, or copies of these CDs, are sent to APL for collating and processing into a 'master set'. The master set of CDs are sent to Saskatoon for duplication and distribution to the rest of the SuperDARN community.

Currently the distribution is a pair of Exabyte tapes (FIT and DAT) containing compressed DAT files, uncompressed FIT and INX files, and daily summary files (SMR). The Exabyte tapes hold ~4 GBytes of data each and are generally filled to roughly equal capacities. Because data for any given day is prevented from being split between distributions tapes often contain <3 GBytes of data, particularly during extended periods of high time-resolution (HTR) mode.

Timely delivery of the data to the community can only be acheived with prompt and frequent receipt of the data CDs at APL. Delays from any radar affect the entire distribution and compound throughout the process. A month lag in sending the master set of data to Saskatoon is currently the quickest that the data can be pushed to that point in the system. Problems at radars and delays in receiving data CDs have frequently stretched this lag to several months. The groups at APL and Saskatoon are working closely to reduce any additional delays in the system and to improve the efficiency in order to deliver the distribution data set as quickly as possible.

While significant improvements in the distribution system have been made in recent years, the amount of data that must be pushed through this system is rapidly increasing. Several factors including the addition of two new radars, increased use of the HTR mode (doubles data), recording angle-of-arrival information (XCF) every scan (doubles data), increased use of special modes that collect more data, and increased scatter during solar maximum all lead to a significant increase in the amount of SuperDARN data that must be handled and stored. The anticipation of Stereo-CUTLASS and an additional radar in Alaska will further impact the system which is already being stressed to its limits.

While improvements in the system have been made further strategies must also be devised to increase the amount of data that can be processed and distributed to the SuperDARN community. The groups at APL and Saskatoon are currently working towards this goal. SuperDARN users and operators should also be aware of the ever increasing impact that high data rate modes have on the data distibution system. The benefits of these modes should be weighed against their increasing impact.

2 Discussion

The following is a discussion of the amount of data that has been received at APL from the northern hemisphere SuperDARN radars for the past year.

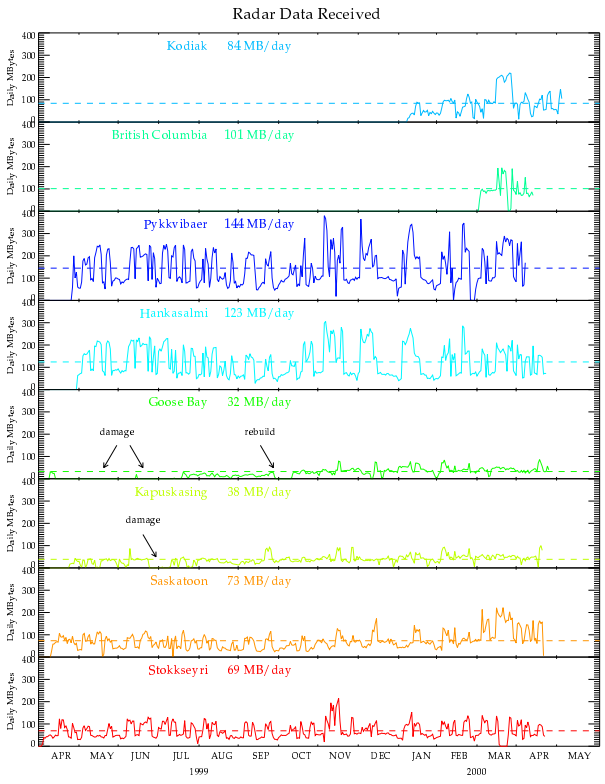

Figure 2: Total MBytes of compressed data received at APL from each of the 8 currently operating radars in the northern hemisphere.

Figure 2 shows the amount of compressed data (in MBytes) for each day that have been received at APL from the 8 operating northern hemisphere radars. A comparision of the amount of data that is generated by each radar can be made from the individual panels. Several factors influence the amount of data that a given radar collects during a single day including the prevailing ionospheric conditions, the mode in which the radar is operating, the operating frequency, and the noise level at the site.

Several interesting features can be seen in Figure 2. Within each month there are periods lasting up to ~1 week where the radars are running the HTR mode. During these periods the amount of data per day roughly doubles due to the 1 minute scan rate. Some radars run the HTR mode during all common time periods.

A possible seasonal effect can be seen starting near October 1999 when the amount of data collected during HTR periods appears to nearly double from the values in the preceding months. During the early part of November 1999 the CUTLASS radar located at Pykkvibær, Iceland collected nearly 400 MBytes of compressed data.

A trend in the amount of data the radars generate is evident in Figure 2. The two CUTLASS radar generate the most data, about twice the amount of Saskatoon (SAS) and Stokkseyri (STOK), which produce roughly twice as much again as Goose Bay (GB) and Kapuskasing (KAP). The two new radars, British Columbia (BC) and Kodiak (AK), appear to generate more data than SAS and STOK but less than the two CUTLASS radars. The average of the daily amount of data is shown in each panel, further illustrating this general trend.

The amount of compressed data is shown in Figure 2 rather than uncompressed data because it was easier to access. Showing the compressed data somewhat obscures the comparison of the amount of data each radar collects since the amount of compression acheived on data files can vary. However, a general comparison can still still be made using the compressed data.

Finally, periods when individual radars are not operating can be seen in the various traces, particularly at GB and KAP during the Summer and early Fall of 1999.

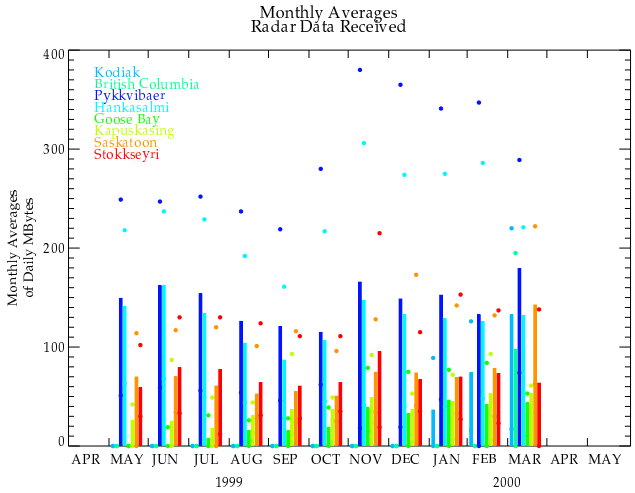

Figure 3: Monthly averages of the daily amount of compressed data produced at the individual radars. Bars show the monthly averages and dots indicate the maximum and minimum daily amount for each month.

Figure 3 shows the same data as Figure 2 but as a monthly average. That is, the monthly average of the daily compressed data from each radar. The dots indicate the maximum and minimum during each month.

The general ranking of the data produced by the radars (mentioned above) can be more clearly seen by the size of the bars. The apparent seasonal trend is not evident in the histogram, but is clear when looking at the dots marking the monthly maximums.

During March, 2000, the amount of data produced by SAS appears to nearly double, with no corresponding increases of this magnitude in the other radars. Most likely this increase is due to this radar either recording XCFs on every scan or running HTR mode during all common time periods. Either activity doubles the amount of data collected by any given radar.

Finally, the new radars (BC and AK) are seen to contribute in the latter part of the plot.

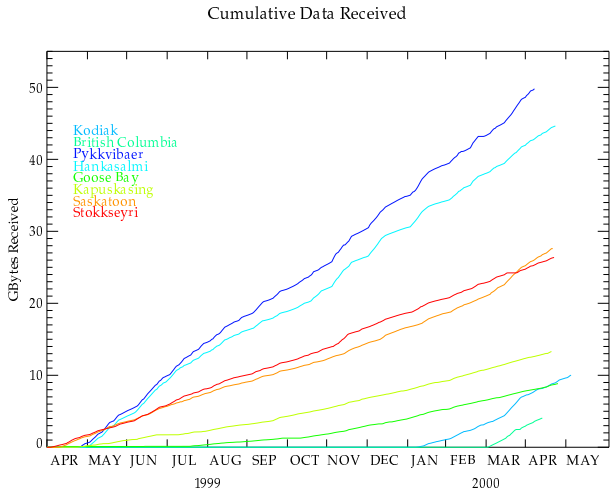

Figure 4: Cummulative compressed data from each radar beginning from April, 1999 or the initial date of operation.

Figure 4 is yet another depiction of the data from Figure 2 and Figure 3. In this figure the cummulative data from each radar received at APL is plotted. The slope of the lines indicates the rate at which data is being generated by the various radars. The yearly amount of data is given by the final position of each trace for 6 of the 8 radars. For instance, the CUTLASS radar at Pykkvibær produced ∼50 GBytes of compressed data during this period of one year. It is estimated that the 8 operating radars produce ∼250 GBytes of compressed data per year.

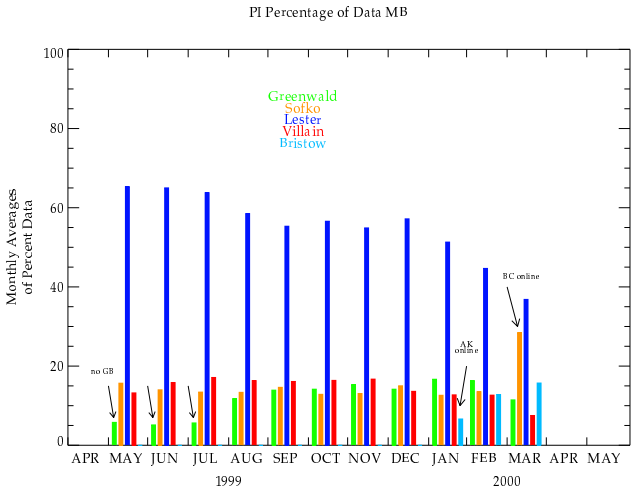

Figure 5: Percentage of total monthly averaged compressed data produced by each principle investigator (PI).

Figure 5 shows the percentage of the monthly averaged compressed data that is produced by each principle investigator (PI). Those PIs with multiple radars or radars that produce more data have larger percentages. For example, the two CUTLASS radar produced between 60% and nearly 70% of the total SuperDARN data from the northern hemisphere during 1999. March, 2000 is the only month when all 8 radars were operating, but may or may not be indicative of the long-term distribution of data production between all 8 radars.

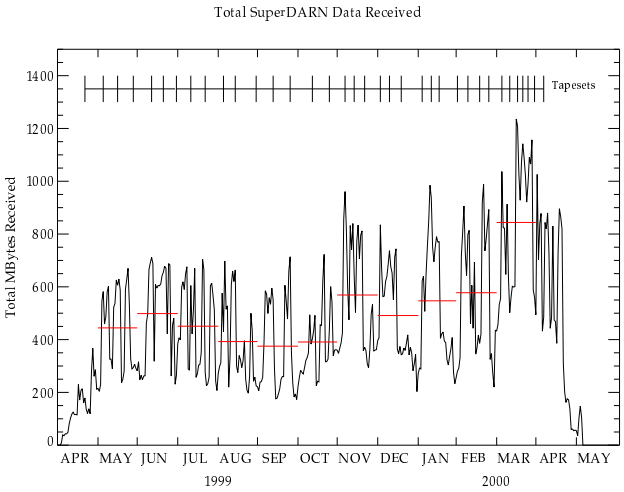

Figure 6: Total amount of compressed data from the northern hemisphere radars.

Figure 6 is the total daily compressed data from all the radars in operation. The red horizontal lines indicate monthly averages. Periods when HTR is running are seen when the total amount of data roughly doubles. Also evident is an increase in the peaks and valleys of the trace after October, 1999.

March, 2000 is the only complete month during which all 8 northern hemisphere radars were operating. The increase in the total amount of data generated by all the radars is obvious. During this month there were several days in which the total amount of compressed data from all the radars approached or exceeded 1.2 GBytes.

During such periods of high data accumulation the number of days that can included on the data distrbution set gets small (approaching zero!). Figure 7 better shows the decreasing number of days that can be fit onto a set of Exabyte tapes. The problem is clearly evident in the downward trend of the dotted line fit to the data for this one year period.

Concern regarding this issue was raised at the previous SuperDARN meeting in 1999 and a change was made to distribute compressed FIT files, increasing the number of days which fit on a pair of Exabyte tapes. Indeed, at the beginning of 1999 a tape set comprised 10–15 days of data, however, a further increase in the amount of data being collected has reduced this number to as low

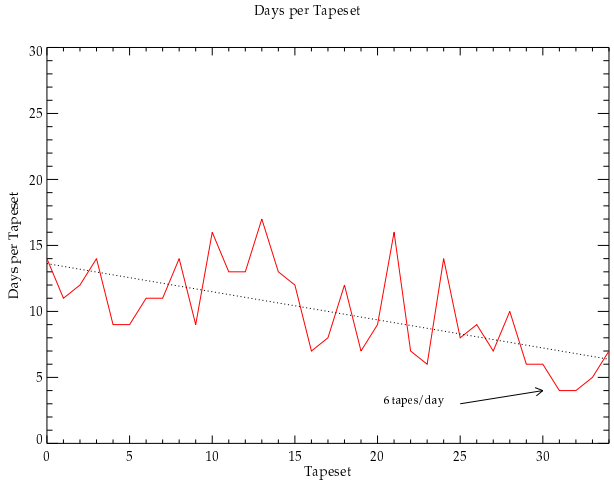

Figure 7: Number of days of radar data that comprise an Exabyte pair of tapes for distrubution to the SuperDARN community.

as 4 days. Further improvements to the system or new technology is necessary.

An example of the impact these high data rates have on our existing system is noted in Figure 7. During periods when 4 days comprise a tape set the group at Saskatoon must produce 6 tapes per day to supply the necessary number of tapes to the community. Considering weekends this number increases and approaches the limits of the duplicating sytsem used at Saskatoon.

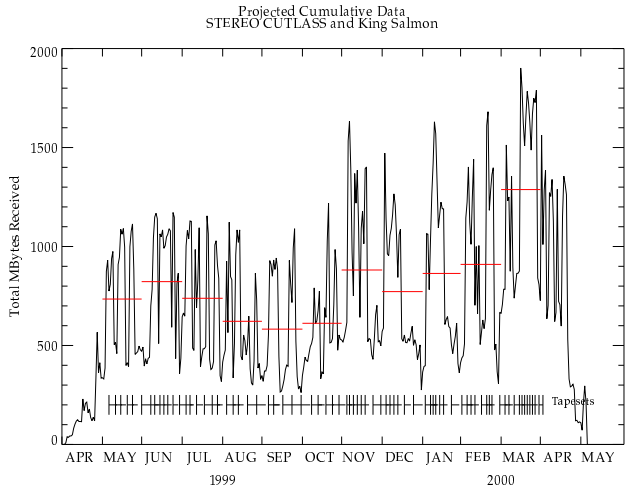

Figure 8: Projected total amount of data for the northern hemisphere radars including Stereo CUTLASS and a second radar in Alaska for the same period as previous figures.

As previously mentioned, the planned addition of Stereo CUTLASS and a second radar in Alaska in 2000 will further add to the amount of data generated by the northern hemisphere radars. Figure 8 shows the estimated amount of total data collected by these radars for the same period as the other figures. The projection is based on the two CUTLASS radars doubling the amount of data generated and the new Alaska radar providing the same as the existing radar in Kodiak.

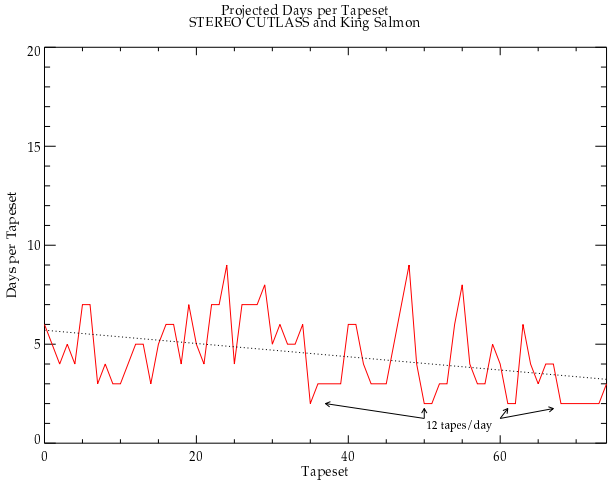

Figure 9: Projected number of days comprising a tape set considering Stereo CUTLASS and a second radar in Alaska, both due in 2000.

It can be seen in Figure 8 that during periods running HTR mode the daily amount of compressed data approaches 2 GBytes per day. The corresponding number of days comprising a tape set drops to two days in Figure 9. During such periods Saskatoon will have to generate more than 12 tapes per day, exceeding their current capacity. It is obvious from this estimate that alternatives must be sought.

Finally, Figure 10 shows that the radars generating the most data do not necessarily provide the most gridded data points. If the number of gridded data points per MByte of data is used as a proxy for the 'efficiency' of the radars, then the Goose Bay radar comes out on top. Of course, the number of gridded data points isn't the only information contained in the data files. For instance, the angle of arrival information and ground scatter is not included in this efficiency calculation and is certainly useful for some studies.

3 Summary

The recent addition of two new radars (BC and AK) and the increased use of high data rate modes (including HTR during common time, those which record angle of arrival every scan, and others) have increased the amount of data that the northern hemisphere radars collect. The anticipated addition of a second radar in Alaska and Stereo CUTLASS mode will overwhelm the

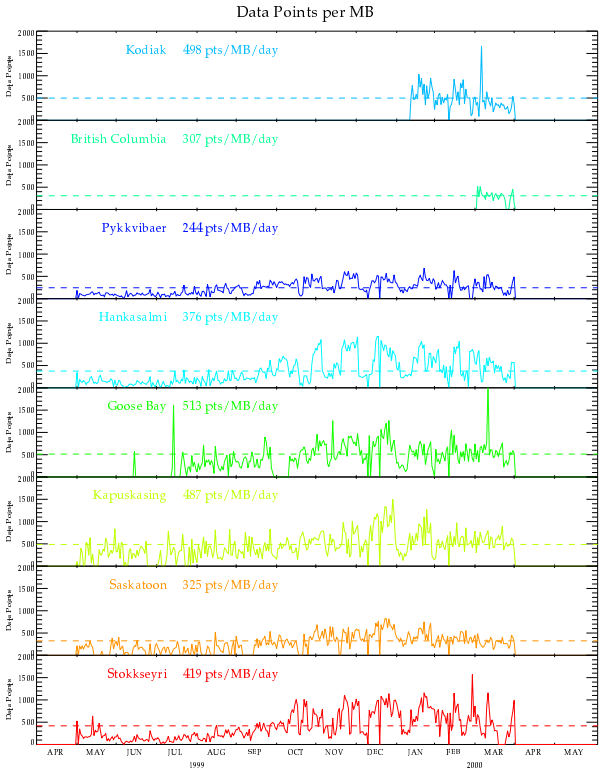

Figure 10: Number of gridded data points per MBytes of data produced by each radar.

current system for collating, processing, and distributing these data to the SuperDARN community.

The groups at APL and Saskatoon are working to further improve the distribution process, however, physical limits may soon be reached with existing hardware and distribution media. The number of Exabyte tapes that can be produced in a day limits the amount of data that can be pushed through this system. We are rapidly approaching this limit and suggest that other distribution media be considered. Unfortunately DVD technology appears to be mired in standardization battles that some experts in the field feel may never be resolved. At present, this media is too expensive and non-standard to use as a substitute. CD-R technology is one possibility that is superior to Exabyte tapes in nearly every aspect, including the amount of data throughput that can be achieved. The SuperDARN community should be aware of other technologies and collectively decide on an alternative to Exabyte tapes.

In addition, the community should be aware of the impact that the high data rate modes have on the groups at APL and Saskatoon. The benefits of these modes should be considered against the increasing impact that will be felt by the entire community.