Using Cray REELlibrarian in Production Environment

Beata Sarnowski, Arctic Region Supercomputing Center,
University of Alaska, Fairbanks, Alaska

ABSTRACT: *Users at the Arctic Region Supercomputing Center (ARSC) can use Cray/REELlibrarian (CRL), a volume management system, to control a centrally stored library of tapes and to validate requests for tape mounts by communicating with the UNICOS tape subsystem.*
CRL is an important element in data management at ARSC and CRL provides a screen interface for user-maintained backups and archiving. ARSC User Services has been training users in CRL over the past year. This paper will discuss basic CRL usage and some common problems/questions from users.

1 Introduction

Scientific databases consisting of large numbers of huge datasets, such that they cannot be stored on disk practically, are not amenable to today's database systems. The Arctic Region Supercomputing Center, ARSC, utilizes a Storage Technology Automated Tape Cartridge System for near-line data storage.

The number of the users of CRL in ARSC is growing and usage has been increasing. However, slow development in user-support capability and reaching silo capacity has discouraged beginning users from experimenting with various features until CRL is relatively problem-free.

ARSC staff have also been experiencing difficulty with CRL and have been made aware of additional difficulties through user assistance. Sharing these problems with others will contribute to rapid development of a better massive storage system.

2 What is CRL?

Cray REELlibrarian Features

Cray Reel Librarian 2.0 (**CRL**) is an on-line catalog system that allows a large number of tapes and volumes to be managed, maintaining their media type, data format, and many other characteristics. CRL manages all the silo tapes. In addition, CRL provides tape archiving and file backup to user-designated silo tapes. Active CRL tapes remain readily accessible in the silo. Inactive CRL tapes can be moved out of the silo to off-line tape racks.

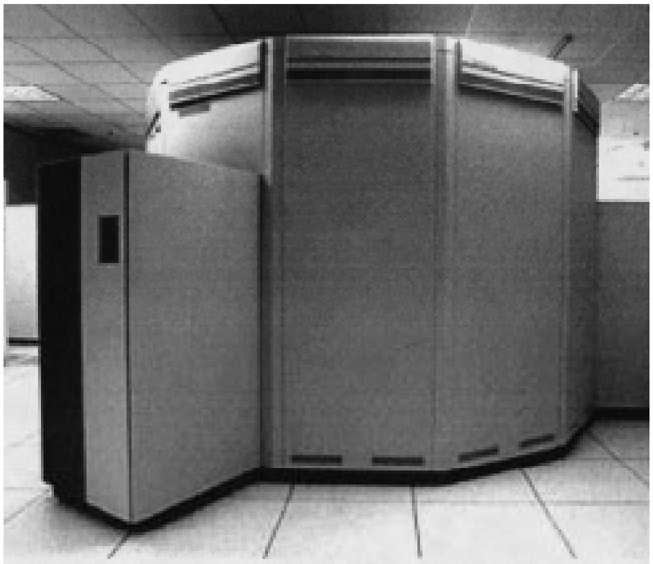

Figure 1: StorageTek *Powderhorn*

One silo, shown in Figure 1, houses 5,546 tape cartridges, each with a capacity of at least 200MB (some will have more than this when data compression is used), for a total storage space of more than 1.1TB. Another silo is planned for 1995, and will probably contain higher density media and more storage capacity.

CRL provides full-screen and command-line interfaces for users and administrators. Most of CRL 2.0 functions can be accomplished using either interface, affording an easy-to-use, menu-based interface for less proficient UNICOS users, or a more powerful command line interface for more experienced users.

Files archived through CRL cannot be accidentally deleted from disk because CRL is not connected to the disk. Moreover, if a disk crash occurs, a CRL backup assures that data won't be lost. User access to tape files is by filename or volume set name. Users are not required to supply any form of volume identification if unambiguous file name references are supplied (e.g. volume name, volume IDs or volume serial numbers).

Data Migration Facility

The Cray Data Migration Facility 2.1 (**DMF**) is a UNICOS facility that ensures the availability of file system space by automatically moving files larger than some base size from on-line disk to an off-line storage medium. This change in residence is almost transparent to users because the files remain cataloged in their original directories and usually act as if they were still on disk.

DMF essentially extends the file system. DMF is useful since the user does not need to copy the file to another archival service or file system. Running DMF on the users' home file system gives users a much larger virtual disk space with no worries about which files need to be specially moved to another place for archival. Under DMF, silo space appears as an extension of ARSC's 89GB Cray disk file space and the access procedure is almost invisible to the user. There are also disadvantages. Because there is no backup of DMF at ARSC, a disk crash or incorrect use of the *rm* command can cause permanent loss of the data. Also, because DMF is so user-friendly, users become very discouraged when they see CRL for the first time.

3 CRL Usage at ARSC

CRL provides library management for all silo tapes, i.e. both DMF silo tapes and user and system silo tapes are managed under CRL. Subsequent references to CRL tapes refer to the user created CRL tapes.

The percentage of CRL users to ARSC total users from the installation of CRL to the current time is shown in Figure 2. The number of CRL users was growing in October, but has not grown fast since. The reason for this fast growth was the activation of a disk quota system so that the users could no longer store massive amounts of their data on the on-line disk anymore and the activation of a charge for DMF file storage.

The above reasons motivated users with massive amounts of data to use CRL, but the unfriendliness of using CRL relative to DMF deters the users with moderately sized data files. The slow increase of the number of CRL users after some months shown in Figure 2, can be explained by this reason.

Figure 3 shows the number of tapes used by CRL in ARSC. Currently we have on the system 6,742 tapes, where 2,742 tapes are under active management of CRL, and the remaining 4,000 are DMF tapes (they are in a pool called DMF).

The curve of *Total Tape* is the sum of *Tape Usage by Users* and *Tape Usage by System*. The number of tapes used is increasing rapidly since CRL has been installed, while the number of users using CRL is growing slowly. This means that the users of CRL are becoming more active in using it once they get used to it. It becomes a necessary tool for archiving data.

It is a challenge to ARSC User Services to assist users beginning to use CRL and to work with CRAY Research Inc. to improve its ease of use.

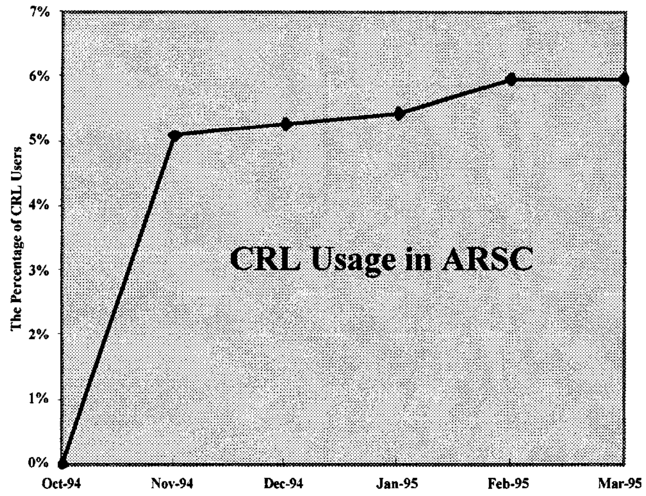

Figure 2: The Percentage of CRL Users at ARSC

4 CRL Problems & Solutions

Many of the problems we experienced have been fixed now by CRL code modifications performed by CRAY Research Inc. in collaboration with our site. The major problems and solutions that occurred at ARSC are summarized in the list below.

4.1 Silo filled up

The silo tape system does not have the capability to balance tapes automatically between DMF and CRL. Thus, a configuration of CRL for DMF/CRL distribution and off-line tapes is necessary. Many users have been using a large number of tapes. For instance, there are a several groups sharing 829 tapes of weather data in various climate studies.

How much data will users need to store and over what length of time? What is the best way to manage huge amounts of data? What will happen when the silo fills up? We need this information in order to allocate enough tapes within the silo.

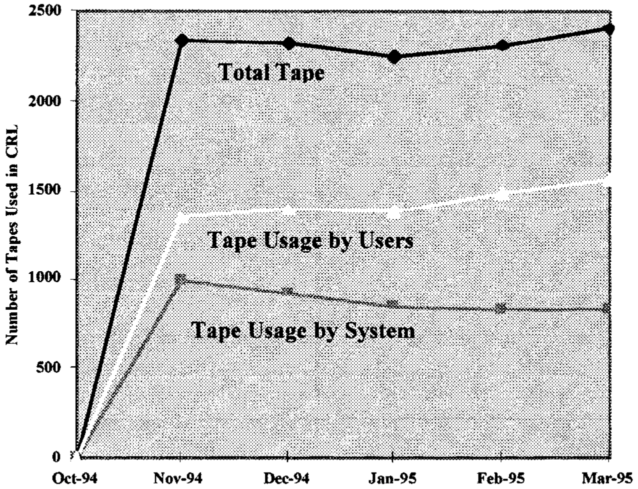

Figure 3: The Number of ARSC CRL Tapes

Tape Distribution

Last October we ran into a situation where we had fewer than 50 tapes in CRL *SCRATCH* pool. A temporary solution was implemented at that time. Some empty DMF tapes were flagged as *read only* and removed from the silo, and CRL *SCRATCH* tapes were put into the silo in those slots.

A similar situation occurred again this year, at the end of February. Rather then reducing the available DMF pool, inactive CRL tapes were taken out of the silo. Right now we are in the process of implementing a solution to the problem of the silo filling up by creating off-line tape storage locations. Also, we are in the process of establishing a long-term plan which will regularly and automatically balance tapes in the silo (DMF and CRL), i.e. available slots in the silo, CRL tapes outside the silo, and available tapes outside the silo.

Off-line Tape System

Another silo is planned for this year. In the meantime we will establish thresholds for each of these pools which will indicate when operator action is required.

What does this mean now for our users?

1. All DMF data will remain in the silo.
2. Some CRL tapes will be removed from the silo, oldest tapes first. When a tape has been removed from the silo a user may still access it, but the tape must be manually moved into the silo by an operator.
3. Because the mount requires operator intervention, the delay will be greater than if the tape were in the silo.
4. Because there is no operator coverage on Saturday nights and Sunday day and night, users will not be able to access those tapes during these periods. Jobs will wait until the tape has been mounted, all night if necessary. We plan full operator coverage later this year.

For the off-line tape system, conditions should be defined to decide which tapes will go to it. The current method for the replacement of CRL tapes to the off-line tape system is that the oldest last access dates of CRL tapes in the silo will go into this off-line tape system.

As a part of the replacement of CRL tapes procedure ARSC may also ask large tape users to specify tapes for long term storage only.

4.2 Changing an owner of CRL files

Change of the owner of CRL files is impossible. It is believed that the problem is caused by the way of defining the file owner in CRL.

This problem occurred when there was a request from a user to load his old tapes from UNICOS tape subsystem to CRL. The job could not be done under root because the change of owner from root to his username was impossible. For that reason, we needed to do the job under the user's account with his password.

Also, when another user requested a change of his username, it was impossible to do it without restoring all his files from CRL, i.e. 182 tapes, back to his directory and then after his username change, storing them back on CRL under the new username.

4.3 Group access

In order to access another CRL user's files, the pool, volume set, and file permissions have to be set. A few users, under the same project, had trouble accessing another person's CRL files with permission. Thus we needed to determine just what permissions were necessary to access those CRL files even though the pool had the proper permissions. There is no good volume access control through owner, group and world permissions yet. The problem has been reported, SPR 83157 filed, but still does not have a solution. Right now, users can only partially share CRL / tape volume list within group.

The volume pool (*pname*) permission can be set with the *rlpedit* command:

rlpedit agacc=group_id pname,

where *group_id* is a group ID added to the volume pool's group access list. The *rlvedit* command can be used to change volume set, *volset*, to set the volume set permission mask, *vs_perm_mask*:

rlvedit vmode=vs_perm_mask volset.

The file, *filename*, itself must have the proper permission set, *f_perm_mask*. It can be done with the *rlfedit* command:

rlfedit fmode=f_perm_mask filename.

If a volume set name is known and if group permission is given, then the following commands list all volume sets that other users (*uname*) in the same group own, and all files on the volume set (*uname/vsname*). Also, the *rlr* command displays the catalog entry (volume set information) for the volume (*uname/vsname*).

rlr volset=uname/vsname vslist
rlr volset=uname/vsname vsflist
rlr volset=uname/vsname vinfo.

Right now, a user cannot list all the files on the volume set of other users in the same group. The volume set of the other users in the same group is supposed to be listed by the below command, but it does not work.

rlr volset=uname/vsname flist.

4.4 Problem accessing the **rlr** *command*

Users entering *rlr flist* command may get a message like *"can't open /usr/tmp/RELL.nnn: No such file or directory*" instead of the list of CRL files. This message is caused by wrong permission settings for the */usr/tmp* directory. The default mount permissions for */usr/tmp* directory are rwxrwxrwt for normal CRL working conditions. For instance, if the CRL server encounters a condition that results in a core dump, the resulting core file will be created in */usr/tmp*.

The problem occurred on our system last September. The permission set was altered by the */etc/rc* script during an upgrade installation. In the */etc/rc* script, *chmod* seemed to be filtering what it was directed to do through *umask.* The man pages, example #4, for *chmod* showed setting permissions for all (user, group and other) with *a +perm*. The result depended on what the *umask* setting was. After the upgrade installation the */etc/rc* script acted as if *umask* was not involved even though both */tmp* and */usr/tmp* had *chmod +wt* commands given on them.

The change to POSIX compliance for *chmod* required changing our */etc/rc* script to *chmod a+rwx*. Then the */tmp* and */usr/tmp* mount areas took into account the new *chmod* POSIX compliance: the *+perm* command is filtered through the *umask* setting, the *group+perm* is specific enough to give the desired effect. *chmod* man examples need to be updated.

4.5 CRL crashes when user lists more then 600 files

Until last September, there was a hard-coded number, 600, which limited the number of files which could be listed from a volume set using the *flist* command. A modified version of CRL that will allow up to 5,000 file names is now in the system until a general fix can be added to CRL in a regular design change.

The number that we are using is 5,000 because significantly more memory would be required for a higher number. Until a real fix is available, there is still an upper limit of 5,000 at which CRL could crash. The general fix will take longer, but will replace this arbitrary number method with a different way of doing it.

4.6 Ambiguous name of the silo and the cart drives

CART was used for both *SILO* (w/ Improve Data Recording Capability, IDRC) and *CART* (no IDRC). The CART device group represented 3480 type cartridges (square tapes).

There was no way to distinguish between the silo drives with IDRC and the cart drives without IDRC when both were type *CART*. The four drives without IDRC are located outside the silo and they are operated manually by the operator. There are eight cart drives physically located inside the silo and they are operated automatically by the robot autoloader. The *tpmnt* command displays the current status of all tape devices. Since the drives have different capabilities and can cause problems if the wrong one is used, there should be a way to distinguish them.

Now we have *MCART* (for manually operated drives) and *CART* (for silo drives) so that they can be distinguished. The *text_tapeconfig* file, which defines the configuration of tape devices and all of the tape hardware used by the system, has been modified, and it works as expected.

4.7 CRL install aborts if no /usr/spool/crl available

ARSC opened SPR #81501 on this problem last August. Locally the */usr/spool/crl* directory has been created. The */usr/spool/crl* directory is CRL command log directory, used to create and write information pertinent to the command's execution. This directory must be world writable and present for the install of CRL. It can be changed in the */etc/config/reelenv* file.

5 Frequently Asked Questions from Users

5.1 How to make sure that data has been stored?

The *tpmnt* command creates a log file called *tape.msg* in a user's current working directory. The *tape.msg* file shows a record of what was written to the tape, i.e. it consists of a history of tape commands. We advise users always to check the file *tape.msg* for any error messages because all informative and error messages are appended to this file. It is extremely important that they verify that their data has been written to the tape.

5.2 How to operate on CRL tapes from batch session?

To submit a job through a Network Queuing System (NQS), using the CART resource group, the *qsub -lUa limit* along with any other options must be specified. *-lUa limit* specifies the maximum number of tape drives of a device group *'a'* allowed for a batch request. Note that there is an error in the man pages because the *limit* part is not shown, but must be specified.

5.3 What are usage limits?

What is the maximum size of a file which a user can copy? What are the limits on the number of characters for sizes?

There is no restriction on the file size. A tape cartridge holds approximately 200 Mbytes. If the file is larger then 200 Mbytes (with compression, more) then additional tapes are used.

A pool name and a volume set name can be up to 12 characters each and they must be unique among all of the user's other pool and volume set names. The maximum file ID length is 44 characters.

Users do not have to keep careful records in order to execute successfully the *tpmnt* to get the files back. CRL generates many reports that help users in their volume management work. The *rl* command invokes CRL full-screen interface to the catalog. Users can select these reports from the User Tape Access menu through rl or by using the *rlr* command.

5.4 How to delete a file from a volume set?

A file cannot be deleted from a volume set, the volume set must be recreated. However a volume set can be scratched. It can be done through the full-screen interface, *rl,* or through command-line operations.

5.5 How to write a file to an existing volume set?

To write a file to an existing volume set, at the end of tape, the user needs to add to his *tpmnt* command the *-q n* option.

6 Common Users' Errors

System error messages are found in either the *tape.msg* file or the user's standard output file.

New CRL users often issue the *rsv* command (reserves tape resources for users) before the release of all previously reserved resources by using the *rls* command which releases reserved tape resources. This mistake will be found in either the *tape.msg* file or the user's standard output file and it generates the TM025 error message (see Appendix). Also, users may not remember to invoke the *rsv* command before doing *tpmnt* (see the TM056 message in Appendix). Trying to use an incorrect volume set name generates the TM086 error message (see Appendix).

Users issuing the *rsv* command, and exceeding their current job limit for tape resources generate the TM109 error message (see Appendix). At ARSC, the users global limit is set to two which means that users cannot use more than two tape drives simultaneously.

Frequently, users had written UNICOS tapes without a label. Our CRL setup does not allow unlabelled tapes, we require ANSI labels. When the system or users try to mount the old tapes without a label it generates the TM173 error message (see Appendix).

7 CRL Future Plans

As far as the future is concerned, the ARSC has short-term and long-term goals. Short-term goals are summarized as follows. We are in the process of discussing a lot of options which we could choose to add to the storage management policy. In some cases, from the users' perspective, tapes will look the same, but we may need to make changes which will actually affect users.

For example, a user may deliberately choose to have his tapes removed from the silo if a known long-term period of inactivity will occur. Since we do not currently charge for CRL tape usage (or silo space), there is no real motivation for a user to do this.

Also we may need to put the out-of-silo tapes in a distinct storage pool which will require users to specify that pool when looking for their tapes.

Our long-term goal is to train all users to use CRL for backups and archives.

8 Conclusion

The ambiguous way of listing other users' volume lists under certain group permission becomes a big barrier for CRL to provide data sharing functions between users.

We are also looking for a solution to a significant group of problems related to the inability to change the owner of a file. The goal is to have CRL be more UNIX compatible in the way it manages files.

Although the mass storage project has taken some time to unfold, we feel that we have kept up with the advances in mass storage and, therefore, are still providing our users with the latest technology. As the current technology moves forward at a rapid pace, there is room for ARSC to expand and enhance the systems we provide to our users.

The storage needs of supercomputer users have increased dramatically and the need for massive data storage systems grows. CRL tries to provide high performance storage capability. However, to be more attractive to the user, the current CRL system should be more user-friendly, easier and more UNIX-like in its commands.

CRL is becoming a more sufficient product. We hope that future CRL releases will eliminate one system crash cause, e.g. the limit on the number of files which can be *listed* from a volume set using *flist* command. Also, we are looking for a good volume access control through owner, group and world permissions which should provide volume-level security.

9 Acknowledgements

I would like to thank Dorothy Corbett, Director of User Services at ARSC who provided support and guidance. Special thanks to Mike Dority of Cray Research Inc. and to Virginia Bedford and Dong Ju Choi of ARSC.

ARSC is supported by the Strategic Environmental Research and Development Program (SERAPE) under the sponsorship of the Army Corps of Engineers Waterways Experiment Station.

References

[1] *Cray/REELlibrarian (CRL) User's Guide (SG-2126 2.0),* Cray Research Incorporated, Minnesota, 1994.

[2] *Cray/REELlibrarian (CRL) Administrator's Guide (SG-* 2127 2.0), Cray Research Incorporated, Minnesota, 1994.

Appendix

TM025 - release previous reservation before issuing reserve

User has issued a *rsv* command, but he must release all previously reserved resources by using the *rls* command.

TM056 - device group not reserved

Either the device group name on the *tpmnt* command does not match the device group name user used on the *rsv* command or he has not issued an *rsv* command.

TM086 - tape daemon error code : *error code*

The tape daemon returned error *error code*.

TM109 - request exceeds job limit; didn't do the reserve

User issued a *rsv* command, exceeding his current job limit for tape resources.

TM173 - bypass or unlabel permission required

User requested nonlabeled or bypass-label processing on the *tpmnt* command, but user does not have permission.

TM346 - CRL file seq gap *seg1* to *seg2* for volset *volset*

CRL daemon encountered missing file sequence records for volume set *volset*. The missing sequence numbers are between *seq1* and *seq2*.