Tapes and Robots for the CS6400 SuperServer

Phil Stringer, Manchester Computing Centre,
The University of Manchester, England

1 Introducton

The CS6400 runs the Solaris operating system which in its earlier versions only ran on single user workstations. As a result some aspects of that background are still present, for example the support for tape devices expects that the device is in front of you and there is no demand for multi-user access. This is not what is required of a powerful multi-user system and facilities are required that give the level of functionality that historically was always provided with mainframe systems. This session will attempt to show how this was provided on the CS6400 at MCC.

2 Background and history

MCC has always provided two distinct services. One to provide computing facilities to the University of Manchester, and also a centrally funded National service providing specialized facilities to other UK universities and related institutions. The national facilities comprise two components, supercomputing and the datasets service, now called MIDAS, which provides access to over 80 databases including National census government economic and other statistical data.

At present we run MIDAS on the CS6400 and the supercomputing service on a Fujitsu VPX 240/10, both running Unix. There is also the usual mix of a whole host of other machines but they are not relevant to this discussion. The previous machines running these services were an Amdahl 5890 300 running VM/CMS and a Fujitsu VP1200 running MVS/XA. Both these machines shared 2 Memorex 5400 tape libraries, which were retained and are now shared by the current systems. Each Memorex ATL has 4 IBM 3480 compatible tape drives and robotics for storing and mounting 5000 cartridges.

The two points that stand out is that we have moved from an IBM operating system environment to Unix, and that we still have an important part of our service based on equipment using IBM channel architecture.

3 Hardware

Now a few details of the hardware for which tape support is required. Our CS6400 has 12 processors, 768 Mb of memory and 170Gb of disc storage. The tape devices are 2 x 1/2" open-reel decks, 1 x QIC150, 2 x Exabyte, 4 x 3480 (Memorex ATL) and in the near future 1 x DAT.

The decks on the Memorex ATL have IBM channels and we have a CNT Channelink box which converts the block multiplexor protocol to SCSI and vice versa. The CNT box is in fact a more general purpose converter, normally used as a channel extender, allowing certain peripherals to be sited at other locations connected via communications lines. The SCSI option is a fairly new introduction to the box.

The Memorex ATL used to be controlled via mainframe software down the IBM channel with a database of cartridge locations held on one of the mainframes. Those of you with robot systems will know that this can be a little restrictive operationally. The Unix implementation from Memorex is a significant improvement. The database and control functions have been moved to a PC and the host communication to get cartridges mounted etc. is via a TCP/IP link. This is a much more flexible arrangement and removes a lot of operating system dependencies. It can now be connected to any system that supports 3480 and TCP/IP.

Access to the control functions on the host is via a set of Unix commands. These communicate with a daemon which does the message passing between the host and the PC. Data access is via the standard device drivers supplied with the operating system. E.g. /dev/rmt/<nn>.

4 Software

The tape devices are used for a number of distinct and competing functions. Firstly the two systems functions, backup provided by Sun's Online Backup, and data archive and restore provided by Amdahl's A+Unitree. These two are based on the ATL as we require fully automated unmanned operation of these services.

Then there is user tape access usually for data interchange, which can be any type of media and which requires operator interaction for all but ATL access.

5 Facilities required for tape access

The basic facility required is data transfer to the medium selected, and this is performed by the device driver in the operating system.

Security has two requirements. Firstly making sure that the owner of the data is the only one who can access it once the tape is mounted. The Unix permissions on the device driver provide this. However it is too inflexible for a multi-user system as the only real options are to set it up so that all users may access the device or that only one particular user can. The second requirement is to ensure that the correct media (tape) has been mounted on the deck.

On a single user system such as a workstation there is no problem caused by contention for decks. The deck is always available and you can always use the same one. On a large system there has to be a mechanism to decide which decks are free, and provide a simple mechanism for access. Traditionally you used the same physical deck which was coded into your job as the device address. This is no longer good enough as the job should refer to the tape by name and the operating system should automatically direct all operations to the appropriate physical deck.

A secure defined operator interface is required for requesting tapes to be mounted either by the operator or by a robotic system. It is especially important with a robotics system that all requests are passed through a system to check ownership etc. for the media requested.

If the system is to be effective then it must have simple easy to user user commands. There must be sensible defaults for non-expert users and the user must be able to easily find out what is going on. When the user can't see the tape drive and the tape on it the software provided must give that level of information.

6 Addressing these needs

The major component the we are using to address these needs is Amdahl's tape daemon which came with A+Unitree. Although it meets our needs it was in fact the only software that was available in our timescale to provide the service we require. It consists of two parts, the tpdaemon which provides a centralized control function and user commands to communicate with it.

The tpdaemon accepts requests for tape mounts, assigns an appropriate deck (or queues the request), requests the operator to mount the tape, and once mounted performs any label checking required. The user is than informed that the tape is available by a message to his screen, and he can then access it via a special device /dev/rmt/<volser> which has permissions set so that only he owns that device. When he has finished with the tape a command is sent to the tpdaemon to unload the tape, delete the special device and return the deck to the pool for other users.

7 Deficiencies in this system

The supplied system however still does not quite fully met our needs and there were a few areas where still further additional functionality had to be provided locally.

The full range of device drivers provided with Solaris are not available with the tape daemon. The missing parts are that the new devices are no-rewind devices which actually causes little problems and only need documenting. The more important is that the bsd tape movement options are not available. This required a System V version of local software and the provision of a System V tcopy command as Sun only provide a bsd version.

The only robotics interface supplied with the Amdahl tape daemon is for Storagetek silos. There are no facilities built in for adding support for other robotics systems.

There is a tape ownership and checking facility built into the tape daemon, but it is more restrictive than normal Unix file permissions and requires a reload of the tape daemon to change the database. This can only be done when there are no tapes loaded, and is therefore unacceptable.

Online Backup is designed to use the same dedicated tape deck on every run using bsd tape movement and can therefore not be made to use the tape daemon. A separate system was required to support it. Unitree is designed to work with the tape daemon, so once it was made to support the Memorex ATL, then Unitree could use it.

8 Adding ATL support to the tape daemon

There is only a small amount of functionality that actually needs adding to the tape daemon to support the Memorex ATL. Data transfer is done via the device driver and so is supported by default. Tapes are automatically returned to their home cell when they are unloaded which is done through the normal device driver. The only mechanism that was therefore required was a means to intercept the operator mount messages and use them to generate atlmount commands to get the tapes mounted.

This was not as easy as it appears as it is not possible in Solaris for a program to filter all messages written to the operator console, and periodically running a command to list pending mounts was not responsive enough. The solution was to use the freely available expect package to execute an rlogin command to the tapeoper userid. This enabled it to read any messages which appeared on the tapeoper screen. Appropriate messages are parsed and used to generate atlmount commands.

Some automation is also required at this point. For example if a tape with the wrong label is mounted, or a readonly tape for a write request, then the normal operator interaction must be performed. This consists of canceling the mount and sending an appropriate message to the user.

This interface has two major plus points. Firstly it requires no changes to the Amdahl software thus guarding against changes in newer software releases. The second advantage is that it is now a very trivial enhancement to support any other other tape robot that has Unix commands to achieve tape mounting.

9 Enhancements to the user interface

A frontend was added to the supplied tape command to make it easier for the user to use. This adds some logic to automatically determine the type of tape device to use when a user has not selected a type. Firstly it uses an atl command to interrogate the PC controlling it and see if a tape cartridge with the label specified by the user is in the ATL. If so it uses a 3480 deck, otherwise it defaults to the second most popular media type of 1/2" open reel tape.

The front end is also used to provide a much more flexible mechanism to determine who owns tapes, and who can mount them. This is provided by using dummy files with names matching the tape labels in /usr/local/lib/tapelib. Standard Unix file permissions are used to determine who can mount the tape for reading or writing. This makes it much easier for normal users to understand. Once a tape is mounted the device permissions are used to determine who can do I/O to the tape.

10 Online Backup

Online Backup is also a user of the tape decks and so a mechanism to let it share them was required. It is also designed to have an operator present to handle tape mounts and is not specifically designed for unmanned running, for example the deck to use can only be given via a full screen configuration program rather than on the backup command itself.

Deck sharing was provided by adding another daemon which both the backup interface and the local tape mounting daemon interface to. This daemon is also actually used to share decks between machines so that both both the Fujitsu VPX and the CS6400 can use them.

Tape mounting is performed by again using the expect package, this time running the backup software, parsing the operator messages, issuing appropriate mounts and replying to operator prompts. It also translates Online Backup's tape label requests into standard IBM cartridge names.

The software to incorporate Online Backup is therefore quite simple, but achieving a fully working system took rather longer to achieve than was originally planned.

11 Problems encountered

As some of the hardware and software interfaces are new a number of unforeseen problems occurred which delayed implementation.

When the machine was first installed with the Memorex ATL software some initial testing was done by manually issuing atlmount commands as required to test basic functionality. No problems were encountered and so step by step various bits of the software were written and implemented to provide the automation required. As the urgent requirement was for user access the first component installed was the tape daemon and the atl tape mounting daemon. As we had it as part of Unitree on a 30 day trial, then Unitree was tested at the same time which showed up the first problem. If the tape daemon was checking a tape label and the tape was a virgin unused tape then the SCSI driver had a problem and the transfer eventually timed out. This could even bring the whole system down at times.

A joint diagnostic effort with local staff to produce the fault, CNT in Minneapolis and Cray in San Diego doing remote diagnostics found that this was caused by a bug in the ISP driver code supplied by Sun. We were told that this would be fixed in Solaris 2.4 which was still some months from release. However the problem could be obviated by connecting the drives via ESP instead. We are therefore using ESP drivers until Solaris 2.4 is installed. 2.4 has now been released, but we do not expect installation until mid summer.

Things went well for a time and more components were tested and installed including the Online Backup interface. However just as soon as we went live with backup to the ATL, a new problem occurred. This manifested itself as the SCSI channel timing out when backup reached the end of tape. This hadn't occurred with test data and only happened after a few tapes had been filled.

On investigation it was caused by a CNT buffering problem with large block sizes and high data transfer rates. After a number of tests CNT produced patches to both their BMX and SCSI interfaces to correct the problem. We were very pleased with the effort that CNT put into solving the problems as they were very responsive, monitoring the tests via a modem into their box and downloading corrective code. This was done even though their engineering development people are in Minneapolis and we had the time difference to contend with.

This sorted out the hardware problems but we then encountered a software problem with Online Backup. If more than 20 tapes were used in a backup run then although backup ran successfully it was impossible to extract location information for files from its database when attempting a recovery. We now have a fix for this problem and at last have everything working as required.

12 Moving forward

We believe that the MCC CS6400 and Solaris now has the tape handling facilities that we have long expected in powerful multi-user systems. Furthermore we now have a platform that can be used to relatively easily add support for new tape devices if required in the future.