EDOC305

EUROGAM DATA ACQUISITION SYSTEM
EUROBALL DATA ACQUISITION SYSTEM

Tape Server - Program Implementation

Edition 1.0
Sept 1995

Nuclear Physics Support Group
Central Laboratory of the Research Councils
Daresbury Laboratory

Introduction

This document describes the specific implementation details as used by the EUROGAM Data Acquisition System of the Tape Server RPC Program Specification formally defined in the document EDOC304. Where appropiate areas where the EUROBALL implementation may or should differ from the EUROGAM implementation are mentioned The order of procedure descriptions in this document has been chosen as far as possible to be in the order that applications will make use of them.

Overview

The Eurogam Tape Server controls the routing of event-by-event data received from the Event Builder (via ethernet or fddi) to one or more data storage devices (normally Exabyte tapes).

The data acquisition system permits a number of data streams to be generated within the Event Builder (currently this number is limited to 4). As an example of the use of multiple data streams it is possible for one stream to contain all good data; a second stream to contain a selected set of that data already output to the first stream (say events having multiplicity greater than 10) and a third stream to contain all bad events and data rejected by the event builder. This final stream can be used by engineers to diagnose faults in the hardware.

A number of storage devices are permitted. The control graphics interface is currently limited to 8 active devices but this is easily extended. The Tape Server implementation is at present limited to a total of 14 tape devices but can in addition support data devices to file on disc plus any number of test/diagnostic devices. Tape devices are normally exabyte tape but other scsi tape devices such as DAT and DLT have also been used.

The input data streams are routed to output devices by a method allowing each data stream to be routed to one or more output device clusters (if more than one this is data duplication). Each device cluster consists of one or more output devices (if more than one this is data striping).

Procedures are provided to inquire and manage the state of each input data stream, each output device and the state of the server within the data acquisition system. The tape server procedures are the atomic units of access from the user interface. UNIX command line programs are available which correspond to each procedure. However these are intended mainly as a test and diagnostic aid. The primary user interface is from the control graphics which can as required implement macro actions (for example detect and automatically change tapes as the end is reached).

The Tape Server monitors the error retry rate for Exabyte drives as a function of the output data rate and issues warnings if this rate exceeds specified limits.

Procedure Actions

Procedure 24. - Claim EGTS

This procedure must be used by applications wishing to access the Tape Server in order to obtain an access key to be used in all subsequent procedures. The access keys are 64 bit fields issued by the server. The server may use any algorithm it wishes to generate the key. However once issued keys should never be reused. The Eurogam server uses the current number of milliseconds since a fixed date (Jan 1 1970) as the key. Using this algorithm all access keys issued are obviously unique.

The Claim procedure fails if an access key has previously been issued and not released (see the Free EGTS procedure). Security of the Tape Server can be implemented by only accepting Claim requests from a restricted set of clients. Although only one client actually makes the Claim request a group of workstations may subsequently control the Tape Server if the access key is shared between them by any method they wish.

The Eurogam server implements this procedure using the algorithm described to generate the access keys. However it does not currently actually check that the access key supplied in subsequent procedures is valid. However checking could be implemented as long as the server saves details of issued access keys (for example on a disc file) which could then be restored if the server program is reloaded.

The Tape Server at this point has no system resources allocated and next expects an Allocate Device procedure (procedure 23)

The server logs the result of the procedure into the System Log recording the client name to which access has been granted.

Procedure 25. - Free EGTS

This procedure is used by applications to return an access key obtained by a previous Claim procedure to the server. The access key is now invalid and any attempt in the future to access the server using this key will fail.

The Eurogam server only accepts the Free request when in the halted state (that is data acquisition is not in progress to the server). All open files are closed, all tapes currently loaded into tape drives are unloaded and all system resources are released. The Tape Server thus reverts to an initial state and next expects a Claim request.

The server logs the result of the procedure into the System Log.

Procedure 16. - Inquire Available Devices

This procedure returns a list of the data storage devices available to the server. Since this is a read-only procedure it is available to any client and no access key is needed. The actual information returned is system dependant. Each entry consists of a device status, a real device name and a generic device name.

For the Eurogam server the generic device name MTH is used for magnetic tape drives (Exabytes) and there are a number of real devices using the names MTH0, MTH1 etc. In addition to the magnetic tape devices for which there is one entry in the device list for each real device there are two special entries (SINK and FILE) which have these names as both the generic and real names. These special entries in fact imply a number of available devices limited by server implementation. FILE is a device in which the data is stored in the filestore and SINK is a device similar to /dev/null which can emulate the performance of any real tape drive by using a controlled time delay before discarding the data block.

Where the tape server has available tape drives of different capability (for example Exabyte, DAT and DLT) then each drive type can be assigned a different generic device name while using simple names of the form MTHx (where x is 0 => 9 and A => Z) to physical identify the real drives.

For each entry the status indicates if the real device is free (that is currently not in use) or allocated (see Claim procedure) and in use by this server.The Eurogam tape server is implemented within a multi-user real-time system which allows interactive user sessions to claim tape drives and so those available for use by the tape server may change with time. It also permits multiple copies of the tape server program to be run (connected to different experiments) which share a pool of tape drives and this also causes the list of free devices to change as drives are allocated to another server. Finally a system command interface allows the drive resource table to be dynamically modified. This is useful if a drive develops a fault and becomes unavailable for further use.

Procedure 23. - Allocate Device

This procedure is used to claim a resource currently in the free device list and allocate it to the current experiment. The device can be allocated by supplying either the real name of a device or a generic name. If a generic name is supplied the server will select any free device. The procedure response returns the allocated real device name which is useful if the request supplied a generic name. For the special devices SINK and FILE the real device name returned is always SINK or FILE.

The client supplies a 32 bit identifier which the server associates with the allocated device. All subsequent requests from the client for this device will supply the identifier which the server will use as a reference for the device. The client identifier can have any form but the Eurogam control software uses the names TAP0, TAP1 etc.

The server logs the result of the procedure into the System Log recording the client identifier and real device name allocated.

Procedure 10. - Deallocate Device

This procedure reverses the action of a previous Allocate Device procedure (procedure 23). The device must be in the allocated state (that is there must not be a tape currently loaded). The server returns the device to its free list for use by a subsequent Allocate Device request.

The server logs the result of the procedure into the System Log recording the client identifier and real device name.

Procedure 2. - Mount Volume

Two procedures are provided (Mount Volume and Identify Volume) which supply the volume name to the server which is used when writing file labels. For tape devices either can be used. For the special devices FILE and SINK however only the Mount Volume procedure should be used. For the device SINK it is permitted that a null string be supplied as the volume name.

The server records the volume name supplied in the control table for the drive. In most cases no further action is required and the drive state changes to the mounted state. It is assumed that the user has inserted the correct tape cassette into the drive (this is checked later during the Open procedure). However for magnetic tape devices allocated as a result of the client supplying a generic name in the Allocate procedure the server will request confirmation that the correct tape cassette has been loaded into the required drive and the drive state will be mounting until this confirmation is received. This option is intended for remote operation of the system and assumes that system operators are available to load the requested tape cassette rather in the manner of batch mainframe computer systems. Currently the Eurogam control software does not use this option.

Procedure 14. - Identify Volume

This procedure requests that the server examine the tape volume loaded into the drive allocated to the supplied identifier for a valid standard volume label. Acceptable volume labels are those which conform to the IBM or ANSI standards for tape labels. If a volume label is found the server records the name in the control table for the drive and the drive state changes to the mounted state as would be the result after a successful Mount Volume procedure. This procedure is invalid for the special devices FILE and SINK which should use the Mount Volume procedure.

The server logs the result of the procedure into the System Log recording the drive name and the volume name found.

Procedure 9. - Dismount Volume

This procedure reverses the action of the Mount Volume or Identify Volume procedures. It is only valid if the device is in the mounted state (that is a tape is loaded but no file is open). If the device is a magnetic tape drive any tape cassette currently loaded is unloaded. For the special devices FILE and SINK no action is necessary.

13. - Initialise Procedure Volume

The tape server writes data tapes which are formatted according to either the IBM specified standard or the ANSI standard (as defined by ANSI X3.27 - 1978). Only VOL1, HDR1, EOF1 and EOF2 labels are used by the tape server. Other labels defined by the standards (such as UHLn and UTLn) are not used.

The Eurogam control software by default uses ANSI labels. The option to use IBM standard labels arises from a time when data tapes were analysed using IBM mainframe computers and may be omitted.

This procedure is used to write a volume label to a new tape or to reinitialise a tape by rewriting its volume label. In the second case the current volume name must be provided as a security check since rewriting the volume label will cause all existing data to be lost. The drive may be initially in the allocated state or in the mounted state (useful when reinitialising a tape) but at the end of the operation will always be in the allocated state. A mount procedure request is required to change to the mounted state.

The drive is in the initialising state during the initialise sequence.

1) rewind label and attempt to read the current label - if the tape is new (contains no data) or does not have a valid volume label goto (3)

2) if the tape does contain a valid volume label check that the volume name on the tape is the same as the current volume name supplied by the initialise procedure request. If these are not the same then abort the sequence.

3) rewind the tape and select the requested write density or mode.

4) write the volume label

5) write 2 file marks to denote end of information - these also allow data to be appended in the case of Exabyte drives. Note - the ANSI standard specifies that a dummy HDR1 block (all nulls) should be written after the volume label which will be overwritten by the first file labels. The Exabyte hardware does not allow this dummy block to be overwritten and so it is ignored.

The server logs the initialise action into the System Log recording the client identifier, real device name, new volume name and possibly old volume name for reinitialise operations.

Procedure 3. - Open File

This procedure is used to prepare the device referred to by the supplied client identifier to accept event-by-event data. The device must currently be in the mounted state when the request is received. The procedure specification allows for a number of options to be supplied by the client. However only the default options need be provided.

For the Eurogam server these are :-

access mode	= 2	write
label type	= 2	ANSI labels
record length	=16384	F format with 16 Kbyte blocks
block length	=16384	

The Euroball server should implement F and FB file format and be prepared to accept data blocks greater than 16K bytes in length. Note however that some UNIX systems find it hard to read magnetic tapes with physical data blocks greater than 64 Kbytes.

The volume name supplied during the Mount or Identify procedure and the file name supplied by the Open File request are used to write the file header labels. The Eurogam server protects data already on the tape by only appending new files after any existing files already on the tape. Existing data files are never overwritten by the Open procedure. The only method provided by the server to reuse a data tape is by the Initialise procedure which will remove all files on the tape.

The server locates the double file mark which marks the end of information on the tape (note - care must be taken not to confuse the 2 consecutive file marks which occur as a result of an empty file for the double file mark which marks the end of information). The second of the two file marks is then overwritten with the new file header labels which consist of a HDR1, HDR2 and file mark. The device status then changes to the Open state and is ready to accept event data.

In the case of the SINK special device no action need be taken and the device immediately changes to the Open state.

The FILE special device simulates the structure of tapes on disc by creating a directory with the name of the tape volume and then creates files within that directory with the name of the tape file.

The server logs the result of the procedure into the System Log recording the drive name, volume name and file name.

Procedure 8. - Close File

This procedure reverses the action of the Open File procedure. It is only valid when the device is in the Open state and if the EGTS system is in the Going state (see Set EGTS Procedure - procedure 26) then the device should not be Associated with a EG Data Stream.

The server will close the tape file by writing the file trailer labels which consist of a file mark, EOF1 and EOF2 and then writing the double file marks which indicate the end of information on the tape. The server should backspace over the second of these file marks and is then positioned correctly to accept a new Open File request.

The device status changes from the Open to the Mounted state.

The server logs the result of the procedure into the System log recording the drive name, volume name and file name. Additionally the maximum recorded write error rate is also recorded. This is the maximum value of the recovered error rate as a percentage of the number of physical blocks written averaged over a 10 second period.

Procedure 7. - Inquire Device Status

This procedure returns status information about the real device implied by the supplied client identifier. Since much of the information returned is device specific the generic device name (e.g. EXB-8500 for Exabyte 8500, Exabyte 8505 and similar devices) is returned. The important information which should be returned is the length of the tape, the length of the tape remaining to be written, the number of (recovered) i/o errors and the number of i/o errors expressed as a percentage of the total number of data blocks written. If any of these items cannot be obtained then the value -1 may be returned. The percentage i/o error rate is returned in units of 0.1% and is averaged over the preceeding 10 seconds.

For SCSI devices the device specific data returned by the device in response to SCSI Request Sense, Mode Sense and Inquiry commands is also returned if possible.

Procedure 12. - Inquire Stream Status

This procedure returns information about the status and use of the data file implied by the supplied identifier. Information supplied by the Allocate Device, Mount or Identify Volume and Open File procedures for this identifier is returned.The three ‘magic number” fields are currently not used and should be returned as -1. The block count and byte count are related to data written to this device since the last Open File procedure. The data rate is the number of bytes/sec wriiten averaged over the last 10 seconds.

Procedure 28. - Associate Data Stream

This procedure defines the way incoming experimental event-by-event data should be distributed to the available real data device streams. The online event formatting system allows events to be written to up to 4 event streams. The Associate Data Stream procedure is used to route each data stream to one or more lists of device streams.

The associate mode = 2 (modes 0 and 1 were used by earlier versions of the program protocol and are now obselete). Mode 2 allows one or more lists of device identifiers; each list may contain one or more device identifiers.

Each list of client identifiers receives a copy of the incoming data for this Data Stream. This enables online duplication of the experimental data and will generate a number of identical copies of the original raw data. Not only does this give greater security to the original data but is more efficient than offline duplication of the data tapes. Each of these lists may itself be a list of device identifiers. The data is written in a “round robin” manner to each of the devices implied by the client identifiers in the list. This enables experimental data to be handled at data rates which are several times the data rate possible for an individual data device. By using “round robin” scheduling all data devices in a list receive data at the same rate and hence at any time have approximately the same amount of data written to them and all devices in a list will reach end of tape at about the same time. Additionally ,if necessary, since the data blocks have been written to the members of the list in a defined and predicably manner it is possible easily to read the tapes and recover the original data ordering.

Association to a null list cancels any existing association. An Associate Stream procedure overwrites any existing association. This procedure is only valid if the Tape Server is in the Halted state.

Procedure 29. - Inquire Data Stream Association

This procedure returns the current Stream Association for the Data Stream in the same format as supplied by procedure 28.

Procedure 26. - Set GETS State

This procedure is used to control the state of the Tape Server as a part of the data acquisition system.

The state may be :-

halted That is no data should be received from the data processors and no data written to the output devices. If the state is changing to halted from going then any data already received from the data processors may be written to the output devices but the tape server should control the data flow from the data processors so as to halt further data.

going The tape server enables receipt of data from the data processors and is prepared to write any data received to the output devices. It should check that all devices which are associated with input data streams are in the open state and thus in fact ready for writing. Any data stream for which there is an empty association list should not be enabled for receipt of data.

test In this state all data streams are enabled and will receive data. Data flow is handled normally and all normal checking occurs but the data is then discarded without writting to tape. This is a very useful diagnostic option.

Procedure 27. - Inquire EGTS State

This procedure requests that the current state (as set by a preceding Set EGTS State procedure - procedure 26) be returned.

Procedure 30. - Inquire EGTS Stream State

This procedure requests that statistics related to a specific input data stream be returned. The block count and byte count relate to data received on this Event Data stream since the last GO command. The data rate is the number of bytes/sec received averaged over the last 10 second period.

Procedure 18. - Move Tape

This procedure is concerned with management of the Automatic Changer robot. Details are to be supplied.

Procedure 19. - Position Tape Changer

This procedure is concerned with management of the Automatic Changer robot. Details are to be supplied.

Procedure 20. - Inquire Element Status

This procedure is concerned with management of the Automatic Changer robot. Details are to be supplied.

Error Recovery Actions

The server while in the GO state continually monitors the state of the drives. For exabyte drives it monitors the number of recovered i/o errors every 10 seconds and expresses this as a percentage of the number of physical data blocks written in that period. If the error rate exceeds 5% a warning message is sent to the System Log. The number of warning messages is limited to one per minute to prevent excessive messages if a drive exceeds the 5% threshold limit for a long period. Experience shows that an Exabyte drive will continue to record with an error rate of up to 10% but at a reduced maximum performance in bytes/sec. At an error rate of around 15% it is very likely that a fatal permanent error will occur.

If a drive generates a permenent error a suitable message is sent to the System Log which can then distribute it to all interested consoles. The drive is then removed from any Association list which it is a part of. For example if the data stream is being recorded in simple duplicate mode to 3 drives and one of these 3 generates a non recovered error then data recording will continue to the 2 remaining active drives.