NOS 2.5.3
FEATURE NOTES

SMD131655

NOS 2.5.3 Feature Notes

Table of Contents

Introduction

This document contains descriptions of most of the features that are part of the NOS 2.5.3 Level 688 release. The articles in this document are targeted for the site analyst, however, some of the topics may be of interest to operations or the end user so we have organized this document such that each chapter may be easily copied and distributed. The Feature Note Audience Matrix should help in distribution of the articles.

These Feature Notes were developed and written by CYBER Software Support. Questions or comments regarding them may be addressed to:

Control Data Corporation
CYBER Software Support - ARH213
4201 North Lexington Avenue
Arden Hills, MN 55126-6198
USA

800-345-9903 (USA and Canada)
612-851-4131 (International)

FEATURE NOTE AUDIENCE MATRIX

Article Title	Site/Analyst	Operations	End User
PP Breakpoint and CPUMTR Breakpoint Package	X		
Automatic Firmware Reload	X		
ISHARE Error Recovery Improvements	X	X	
Tape Alternate Storage	X	X	

CHAPTER 1

PP Breakpoint and CPUMTR Breakpoint Package

NOS 2.5.3 provides two new debugging tools, PP Breakpoint Package and CPUMTR Breakpoint Package, to aid system analysts in debugging their programs. PP Breakpoint Package allows an analyst to set a breakpoint within executable PP code, display and alter PP memory. CPUMTR Breakpoint package allows the console user to breakpoint CPUMTR in both program mode and monitor mode and displays hardware registers for examination.

These packages are standard released products under NOS, there are no special steps to install it. However, a site can disable these breakpoint packages by entering "CPB." in the IPRDECK at deadstart time and save about 180 words in low core.

Detailed information on how to use these packages can be found in the NOS 2 Analysis Handbook.

(This page left intentionally blank.)

CHAPTER 2

Automatic Firmware Reload

NOS has been enhanced to automatically reload firmware whenever the state of a globally downed channel is changed to up or the state of a downed control module is changed to on. The reload is performed by LOADBC using the binaries of the appropriate firmware from the system file. The following disk subsystem components have this capability: 7054, 7154, 7155 controllers; 7165, 7255 adaptors; and 834/836 control modules.

(This page left intentionally blank.)

CHAPTER 3

ISHARE Error Recovery Improvements

The Reliability, Availability and Maintainability (RAM) of ISHARE processing has been improved with the NOS 2.5.3 level 688 release. Now, the same error recovery processing that NOS uses for nonshared RMS devices is used for ISHARE devices.

3.1 Glossary

3.1.1 Device Index Table (DIT)

The DIT is maintained in the label sector of each ISHARED device. It is cleared when the device is PRESET or UNLOADed and entries are made as each mainframe recovers or MOUNTs the device. The position of a mainframe's entry determines the mainframe's index, which is used for interlocking, and the position of its MRT (Machine Recovery Table) sector. The DIT is also used to track permanent file activity and device recovery and/or initialization.

3.1.2 Inaccessible Device

There are two conditions that are defined as "inaccessible" in an ISHARED environment. The standard efinition isa device that has the suspect bit set in the MST, which includes DOWN and OFF states. This definition is used when CPUMTR checks the device status and when tables are being written to the device.

When 1RU is reading tables, a different definition is used. If 1RU gets errors when attempting to read the label track with "ENDMS enabled," it considers the device inaccessible. When ENDMS is enabled, the mass storage error processor can retry the operation using different access paths and can even reload controlware before returning an error status to 1RU.

3.1.3 Label Sector

The label sector is the first sector of the label track. It contains a copy of the MST, the DIT, and other information needed to interlock and recover the device. A copy of the label sector is written as part of the second copy of all the tables on the label track.

3.1.4 Label Track

The label track is a logical track that contains the disk label information. It is usually the first logical track on the device, but if that track is flawed, another one will be selected. Following the label sector is the Track Reservation Table (TRT) for the device, the sector of local areas, the device information sector, and, if the device is an ISHARED device, the Machine Recovery Tables (MRTs) for the mainframes sharing the device. To enhance recoverability, a second copy of all of the above information is written to all devices except extended memory.

3.1.5 Machine Recovery Table (MRT)

Each mainframe that shares an ISHARED device has an MRT in the label track of that device. The mainframe updates its MRT as it reserves logical tracks on the device and as it sets and clears interlocks on logical tracks. The MRT information is used by MREC running on another mainframe to clear the interlocks held by a down mainframe and to release any local file space it has on the device.

3.1.6 Mass Storage Table (MST)

Status and control information for each mass storage device is maintained in the MST in central memory. There is a copy of the MST in each copy of the label sector. In an ISHARED environment, the copy of the MST is updated whenever it or the TRT is changed. Two words are of particular importance for ISHARED devices. Word SDGL is global to all the mainframes sharing the device. When a mainframe accesses a device, SDGL (along with TDGL and ACGL) is copied from the MST in the label sector to the central memory MST. It contains the interlocking information used to control accesses to the device. The second word, MCLL, is not copied to the MST in the label sector. It is local to each mainframe and contains information about the requests pending on the mainframe.

3.1.7 Track Reservation Table (TRT)

Allocation and interlocking information for each logical track is maintained in the TRT in central memory. There are two copies of the TRT in the label track of each device. For an ISHARED device, these copies are updated whenever the TRT is changed.

3.1.8 Hardware Reserve

Hardware reserve is circuitry or firmware that prevents two accesses to the same equipment at the same time. Disk controllers and disk drives each have a hardware reserve which

is set and cleared by function codes. The drive reserves are subordinate to the controller reserves. If the controller reserve is cleared, all the drive reserves are also cleared. When a PP completes a data transfer and issues an ENDMS in an ISHARED environment, MTR responds with a clear-controller-reserve only when there is no activity in progress on the other drives connected to the controller. Otherwise, the PP is directed to clear only the drive reserve. This has the effect of locking out other mainframes until there is no activity on the mainframe that has the controller reserved. To avoid an unacceptably long lockout, MTR limits the number of requests that can be processed without releasing the controller reserve.

Both reserves are critical to the integrity and performance of ISHARED devices. They must be maintained the whole time that disk tables are being read until they are interlocked or completely updated. Keeping the reserve until the tables are interlocked prevents getting tables that were modified by another mainframe in the middle of the read. Keeping the reserve until the tables are completely updated prevents another mainframe from repositioning the drive to access a file while the updating mainframe needs to have the drive positioned over the label track to complete the updating.

However, the error recovery options are very limited if a PP must hold the hardware reserve. Most importantly, a retry cannot be made using an alternate channel. This release enhances recoverability by allowing the reserves to be lost while the tables are being read, and then restarting the request after error recovery has completed. Since a software reserve is set after the tables have been read, the loss of the hardware reserve during error processing can be tolerated during the table rewriting.

3.1.9 Software Reserve

A software reserve mechanism is used in addition to the hardware reserve to maintain the integrity of the tables during the update process. Word SDGL in the MST contains the interlocking information. The MST/TRT interlock contains the mainframe index of the mainframe that is updating the tables. Before a monitor function that updates the MST or TRT is processed, the MST/TRT interlock is set in SDGL and written back to the device. The nonzero value in the interlock informs other mainframes that the device is reserved.

After the function is processed, the tables are rewritten to the device. The MST is written with the interlock still set. After the tables have all been rewritten, the label sector is written once more to clear the MST/TRT interlock. Since all of this writing is normally done with the hardware reserve maintained, another mainframe will rarely see the interlock set. But the hardware reserve may be released during error processing. So when the interlock is read and is set, the reader has either interrupted error processing or has read a device that was being written by a mainframe that went down during the write. In either case, the reader must release the device either to let the error processing complete or to let MREC run to clean up the interlocks.

3.2 History

NOS Multimainframe (MMF) software allows more than one mainframe to share access to a disk. The MMF/ISHARE software permits the sharing without the use of an extended memory link device.

NOS keeps track of the current state of a mass storage device in the MST and the TRT. For nonshared devices, these tables are kept in central memory and are periodically copied to the disk for recovery purposes. When a device is being shared using MMF/ISHARE, the tables are kept on the disk and are read to central memory when they must be changed.

In all cases the table updates are initiated by monitor functions. When a monitor function that updates the tables is issued for an ISHARED device, a PP program called lRU is loaded into the PP that issued the function. lRU obtains the hardware reserve, reads the tables into central memory, and issues the

functio , which CPUMTR processes. If there are no other functions to process, the tables are written back and the hardware reserves released. If there are other functions to process, the software reserve is written to the disk, the reserves are released, the other functions are processed, and then the tables are written back, clearing the software reserve.

3.2.1 Problems

This approach has several inherent problems.

3.2.1.1 Dependence on the Hardware Reserve

ISHARE devices depend on the hardware reserves when updating the tables. There are problems with depending on the hardware reserve:

- A hardware problem might cause the hardware reserve to be lost.
- The driver error processor cannot perform all of its normal error processing, such as switching channels, since the reserves must be maintained for ISHARE processing.
- Diagnostics cannot be executed on the device since reserves must be maintained for ISHARE processing.

3.2.1.2 Inconsistent Tables

Disk or system errors that occur while writing the MST/TRT/MRT back to disk after performing an update can result in inconsistent tables on the disk. For example, if an error occurs after writing two sectors of the TRT, part of the TRT will be the new copy and part of it the old one. This problem seems to occur fairly frequently on ISHARE devices, probably because a large percentage of the I/O on ISHARE devices is for the label track. Note that this problem is worse than a missed checkpoint because it results in inconsistent mass storage

tables whereas missed checkpoints result in obsolete but consistent tables. It can, of course, also occur on nonshared and MMF-shared devices. When this happens, device recovery detects a label error and halts recovery.

3.2.1.3 1RU Deadlocks

When a machine holding the device software reserve goes down, copies of 1RU on the machines that are still up get stuck in a loop waiting for the software reserve to be cleared. If all of the control points or PPs are in this wait state, MREC cannot be run and there is no way to clear the software reserve.

Similarly, when 1RUs are looping attempting to recover from a mass storage error, it may not be possible to schedule the diagnostics to run because the system is deadlocked.

3.3 NOS 2.5.3 RAM Changes

3.3.1 Overview

The primary RAM change for MMF/ISHARED is to improve faults tolerance. It tolerates access path errors by allowing channel switching, inaccessible devices by permitting the job to be rolled and write errors by prohibiting other mainframes from accessing the device and by rewriting the tables after the device has been repaired.

The changes do not require a site to initialize and reload mass storage devices. However, they do require a PRESET of the device, which is normally done on the first machine to be deadstarted.

3.3.2 Software Reserve

When the tables are read to process a function, a software reserve is set on the device. The software reserve is always set, regardless of the number of requests for this device. The old MMF/ISHARE only set the software reserve if there was more than one request for a device. NOS reads the label sector and TRT, then rewrites the label sector with the software reserve set.

1RU attempts to read the label sector and TRT and rewrite the label sector to set the software reserve with ENDMS disabled (as it previously did. If an error is encountered, 1RU executes an ENDMS to release the hardware reserve and channel, then restarts the read at the label sector again with ENDMS disabled. Issuing the ENDMS allows the channel switch that was set up by the driver error processor to occur and also allows 1MV to test the device.

3.3.3 Label Sector Checksum

The label sector is critical to the integrity of the device. It contains the software reserve and a set of counters that are used to verify the consistency between the MST, TRT, and MRTs. Counters are set in the MST and DIT when a table update begins. Corresponding values are written in the last sector of the TRT and in the MRT sectors. If the corresponding counters do not agree, the table update has not completed. To assure the validity of the interlock field and the counters, they are checksummed and the checksum is stored in the last byte in the label sector.

3.3.4 Duplicate Table Copies

The ISHARED RAM feature has added code to support the maintenance of two copies of the information recorded in the label track. The second copy ensures that the device always contains a consistent set of tables. 1RU writes both copies of the tables each time that it updates them. The second copy is not written until the first one is known to be written correctly. The software reserve is set whenever the tables are written to an ISHARED device. 1CK also writes both copies of

the tables whenever it checkpoints devices, unless the device is extended memory.

This enhances the availability of files by increasing the recoverability of the device and minimizing the need to reload files. When MSM or 1MR recovers a device, it uses the second copy of the tables if and only if the first copy is inconsistent.

3.3.5 Fault Tolerance

1RU has been enhanced to deal with various errors that can occur while updating the tables on the device. In addition to hardware errors, 1RU deals with label checksum errors and software reserve errors. The enhancements to ISHARED error handling permit the following:

- abandoning the function if the device becomes inaccessible.
- releasing the hardware reserve to switch access paths or to run diagnostics.
- restarting the I/O following a recoverable error.
- retrying following an unrecoverable error.
- responding to an OVERRIDE during error processing.

3.3.6 Automatic Retry

If an error occurs on an ISHARED device while the tables are being read, NOS attempts to use an alternate path to the device. If the device becomes inaccessible, NOS attempts to return to the caller. This is permitted if and only if the function was issued with the return-on- inaccessible-device option. When a PP uses this option and the device is inaccessible, the PP program rolls the job out until the device becomes accessible again. If control cannot be returned to the caller, 1RU waits for the device to become accessible and then retries the request.

If an error occurs while writing the tables, NOS again attempts to use an alternate path to the device. If the device becomes inaccessible, control returns to the caller as though the table update had completed normally. After verifying the device, 1MV forces the tables to be rewritten.

3.3.7 Online Repair

When the previous or current status of a device is unknown, the tables are verified and possibly repaired, before they are used.

3.3.7.1 MSM Verification

When device labels are checked during deadstart or when the operator mounts a removable pack, NOS attempts to repair inconsistencies if the device is nonshared or if this is the first mainframe to access the device. First, it checks for a second copy of the tables. If a second copy does not exist, one is created by copying the current tables. The MST/TRT update counters are added to the appropriate tables to create a consistent set. If a second copy does exist, it is used if the first copy is inconsistent.

In all cases, the suspect flag is set in the MST, which causes 1MV to verify the device and the label.

3.3.7.2 1MV Verification

As 1MV performs label verification of a device, it checks for read errors, incorrect values, checksum errors, and software reserves. If any of these condition exists and this is a nonshared device that is otherwise usable, 1MV requests a device checkpoint. If the device is ISHARED and the tables are being updated, 1MV calls 1RU to rewrite the tables. Otherwise, 1MV sets the device state to OFF or DOWN.

1MV has been modified to omit the label check when device initialization is pending or the device is inaccessible. Label checking is performed on all up channels.

3.3.7.3 lMR Verification

When the operator initiates label verification and repair via the MREC utility, lMR checks the tables for consistency and attempts to repair problems.

3.4 Operator Interface Changes

3.4.1 E,M Display

3.4.1.1 New Status Code

T — The status field in the E,M display contains a T when an ISHARED device has a table update pending. That is, the device's tables in central memory have been changed but they have not yet been copied to the label track on the device. Only one mainframe should have this status set at any one time.

3.4.2 B Display

Message: ERROR ON ACTIVE DEVICE.

Significance: The significance is unchanged but additional actions are specified for ISHARED devices.

Action: Check the E,M display look for a T status on all mainframes. The mainframe having the T status is attempting to rewrite the label. Check its E,E display for error information.

When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has "T" status. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart.

Message: EQest, INTERLOCKED BY id.

Significance: The software reserve is being held by the mainframe whose machine identifier is id.

Action: Check the device for T status in the E,M display on the mainframe that has the interlock. A device status of T indicates that a table update is pending on that device. The mainframe is probably detecting device errors. Check its A,ERRLOG display for error information.

When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear the reserves and interlocks held by the mainframe. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart.

Finally, the job displaying the message may be aborted by typing OVERRIDE,jsn., but this should only be done as a last resort since it may cause other problems.

Message: EQest, LABEL CHECKSUM ERROR.

Significance: During the reading of the label sector, a checksum error was detected. Normally, a checksum error is due to a write error. The mainframe that was writing should attempt to rewrite the label sector correcting the checksum error.

Another less frequent cause of this error is an access path problem that went undetected by the hardware.

Action: Check the device for T status in the E,M display on the other mainframes. A device status of T indicates that a table update is pending on that device. The mainframe having a device with a T status set will be attempting to rewrite the label to repair the checksum error. Check its A,ERRLOG display for error information.

When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has T status. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart.

Finally, the job displaying the CHECKSUM ERROR message may be aborted by typing OVERRIDE,jsn., but this should only be done as a last resort since it may cause other problems.

Message:	EQest, LABEL READ ERROR.
Significance:	During the reading of the label track, a hardware error was detected.
Action:	Check the A,ERRLOG display for error information. When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) this mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has T status. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart. Finally, the job displaying the CHECKSUM ERROR message may be aborted by typing OVERRIDE,jsn., but this should only be done as a last resort since it may cause other problems.

Message: EQest, LABEL WRITE ERROR.

Significance: During the writing of the label sector, a hardware error was detected. If a good access path to the device can be found by 1MV, it attempst to rewrite the label sector.

Action: Check its A,ERRLOG display for error information. When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has T status. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart.

Finally, the job displaying the CHECKSUM ERROR message may be aborted by typing OVERRIDE,jsn., but this should only be done as a last resort since it may cause other problems.

Message:	EQest, MACHINE NOT IN DIT.
Significance:	This machine's mid is not in the Device Information Table (DIT) in the label sector. The device has probably been PRESET from another mainframe.
Action:	Perform a level 0 deadstart on the mainframe displaying the error message.
Message:	EQest, PN=packnam, DN=nn-u
Significance:	The pack name recorded in the label of the pack on equipment est is packnam. The device number is nn. If u is displayed, it is the unit number of the pack in a multispindle set. These are not the values the system expects.
Action:	Check the E,F display for the pack name of the pack the system expects on the device, remove the current pack, mount the correct one, and type GO,SYS. If the pack that is mounted should be the correct one, it has been initialized with a different pack name. It has to be reinitialized before it can be used.

Message:	EQest, TABLE WRITE ERROR.
Significance:	During the writing of the label track, a hardware error was detected. If a good access path to the device can be found by 1MV, it attempts to rewrite the tables.
Action:	Check its A,ERRLOG display for error information. When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has T status. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart. Finally, the job displaying the CHECKSUM ERROR message may be aborted by typing OVERRIDE,jsn., but this should only be done as a last resort since it may cause other problems.

Message: EQest, BUSY ON ID nn.

Significance: During device recovery, an interlock that was set by a mainframe with an id of nn is detected in the label of the device.

Action: If this is the only mainframe that accesses the device that is up, enter GO,. or GO,SYS. This causes device recovery to attempt to recover the device from the second copy to the tables. This may be an older copy of the tables and may cause track linkage errors to be detected later in the recovery process. If the device is not necessary, you may enter "PAUSE,." or "PAUSE,SYS.". In that case, a label error is set on the device and recovery continues.

If there are other mainframes already accessing the device, one of them is having problems rewriting the label. The only safe action is to enter" PAUSE,". or" PAUSE,SYS".. This causes a label error to be set on the device, but recovery continues to be attepmted after deadstart has completed.

Check the device for a T status in the E,M display on all the other mainframes to determine which one is failing. Check its E,E display for error information. When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has T status. When the corrective action is complete, the downed mainframe must be recovered with a level 0 deadstart.

Message:	EQest, ID id NOW IN DIT.
Significance:	This replaces the old message MACHINE ALREADY IN DIT. It signifies that the mainframe id of the recovering mainframe is already in use on the device. This is usually because an MREC or PRESET has not been done since this mainframe last used the device.
Action:	If this is the only mainframe that has access to the device, it must preset the device. If the device is removable, simply remount it with MOUNT,est,P. If deadstart recovery has detected the problem, restart the deadstart and enter a PRESET command in the EQP deck display. If there are other mainframes already accessing the device, run MREC on one of them and, if deadstarting, restart the deadstart.

Message:	EQest, TABLES INCONSISTENT.
Significance:	During device recovery, an inconsistency has been detected in the first set of tables on the label track (the MST and TRT.)
Action:	If this is the only mainframe that accesses the device that is up, enter GO,. or GO,SYS. This causes device recovery to attempt to recover the device from the second copy to the tables. This may be an older copy of the tables and may cause track linkage errors to be detected later in the recovery process. If the device is not necessary, you may enter "PAUSE,." or "PAUSE,SYS.". In that case, a label error is set on the device and recovery continues. If there are other mainframes already accessing the device, one of them is having problems rewriting the label. The only safe action is to enter "PAUSE,." or "PAUSE,SYS.". This causes a label error to be set on the device, but recovery continues to be attepmted after deadstart has completed. Check the device for T status in the E,M display on all the other mainframes to determine which one is failing. Check its E,E display for error information. When the problem is corrected, activity will resume normally. If the corrective action will require a long time, the device should be OFFed on the other mainframes. Otherwise, you may want to stop (deadstart) the mainframe and run MREC on one of the other mainframes to clear any reserves and interlocks held by the mainframe that has "T" status. When the corrective action is complete, the downed mainframe will have to be recovered with a level 0 deadstart.

3.5 Requirement

All mainframes using an ISHARE device must be running the same version of the ISHARE software.

3.6 Compatibility

Although this level of NOS has changed the label track information, it will recover existing preserved files if the first mainframe to access the device contains a PRESET EQPDeck entry for the ISHARE device. When a removable ISHARE pack is mounted, a MOUNT command with the PRESET parameter must be used. If these steps are followed, NOS will construct the correct label track for the system it is running. This also holds true if you wish to move back and forth between NOS 2.5.3 and any NOS 2 level system.

CHAPTER 4

Tape Alternate Storage

NOS has been enhanced to support magnetic tape as an external storage medium for permanent files. Referred to as Tape Alternate Storage, the feature provides the capability to copy selected permanent files to magnetic tape (a process called destaging), release the disk space occupied by those files, and restore the file data to disk on demand (called staging) when a job accesses a destaged file. The feature is designed to integrate with existing NOS permanent file utilities and formats. Tape Alternate Storage may be used alone or in conjunction with the cartridge alternate storage subsystem MSE (Mass Storage Extended subsystem). However, unlike MSE, no special equipment is required other than a nine-track tape drive.

Terminology and detailed information including usage examples for the Tape Alternate Storage feature may be found in the NOS Analysis Handbook under a new section titled Tape Alternate Storage. Changes to the permanent file utilities to incorporate this feature, documentation for the new Permanent File Supervisor (PFS) utility PFREL, and examples of using PFDUMP in a tape alternate storage environment can be found in the Permanent File Utilities section of the NOS Analysis Handbook.

4.1 System Setup Requirements

Following is a checklist of the steps required to enable the Tape Alternate Storage feature on a NOS 2.5.3 L688 system.

- Enable the feature with the DSD command or IPRDECK entry:

```
ENABLE,TAPE PF STAGING.
```

  The Tape Alternate Storage feature is enabled by default in the NOS 2.5.3 L688 system, as this entry is included in the standard IPRDECK released with the system.

- MAGNET must be called with the new SJ (staging jobs) parameter set to a nonzero integer.

  Two parameters have been added to the MAGNET command to control the maximum number of staging jobs that can be active concurrently and the maximum number of staging request VSNs that can be displayed on the DSD E,P display.

```
MAGNET,SJ=msj,SV=mvd.
```

  msj The maximum number of concurrent staging jobs (default value = 0, maximum value = 30B).

  mvd The maximum number of staging VSNs to display on the E,P display (default value = 17B, maximum value = 77B).

NOTE

The SJ parameter MUST be specified on the MAGNET command to allow staging.

- Procedure STAGE must be saved on username SYSTEMX. It is called by the staging job that MAGNET initiates. The standard released version on the system OPL is:

```
.PROC,STAGE,S=OFF/ON.
.*        STAGE - PROCEDURE TO STAGE PERMANENT FILE FROM TAPE.
.*
.*        *STAGE.* - STAGE PERM FILE FROM TAPE ALT. STORAGE.
.*        *STAGE,S.* - INITIATE *PFHELPR* JOB.
.IF,$S$.EQ.$ON$,SIF.
PFHELPR.
.ELSE,SIF.
PFRES.
.ENDIF,SIF.
```

4.2 New Utilities For Tape Alternate Storage

PFREL — A PFS utility used to release the disk space of files that reside on tape or cartridge alternate storage media. PFREL releases the disk space only if a file's backup requirements are met as specified by the user on a DEFINE, SAVE, or CHANGE command via the BR parameter.

A complete description of PFREL parameters and directives can be found in the Permanent File Utilities section of the NOS Analysis Handbook. Backup specifications are referenced in the Tape Alternate Storage section of the NOS Analysis Handbook.

PFRES — PFRES is called by MAGNET in response to a staging request to restore a file to disk residency. PFRES has no parameters or directives and cannot be called directly.

GENPFD Utility used to select files for destaging and disk space release. While general PFDUMP selection parameters can be used to directly select files for a destage dump, GENPFD offers the following advantages:

- Files can be prioritized according to the time they were last accessed or modified.
- Special user index or file name patterns can be special cased according to site requirements.
- The GENPFD selection utility can select files which will reasonably fill a single tape reel without overflowing to additional reels. The creation of single reel tapes is desirable because it avoids the overhead of reel switching when files are restored to disk.

GENPFD is provided as a standard file selection program for NOS. A site may also choose to write their own file selection program based on algorithms tailored to meet their individual requirements.

GENPFD parameters and directives are documented in the Tape Alternate Storage section of the NOS Analysis Handbook.

4.3 Operational Overview

Files are selected, based on parameter values specified, and copied to tape in PFDUMP format. A secondary VSN may be designated with the verify file (VF) parameter to create a backup copy of the destaged files when PFDUMP destages files to tape. Once a file has been destaged, a pointer to the VSN is set in the permanent file catalog (PFC). In general, the procedure to be followed when destaging files is:

1. Execute PFDUMP with the DT (destage to tape) and IP (inhibit processing) parameters. This writes candidates for destaging on a summary file. Enhancements to PFDUMP include new directives for a more specific, focused file selection.

2. Run GENPFD,PO=D using the summary file for input. (Be sure to rewind it first.) Candidates are prioritized and directives are generated for the files selected to be destaged.

3. Execute PFDUMP again with the DT parameter and specify the directives file created by GENPFD as input. (Once again, be sure to rewind the input file first.)

NOTE

The OP=P (purge after dump) option cannot be specified with the DT parameter.

If the backup requirement for a file is met, the disk space occupied by the file may be released to gain tracks on a device. This can be accomplished by simply executing PFREL. However, by using a utility such as GENPFD, a site can determine if enough files have been destaged to release a specific number of tracks for a device. In general, the procedure to be followed for file release processing is:

1. Execute PFREL with the IP parameter to create a summary file of all candidates. Files written to the summary file meet user-specified backup requirements.

2. GENPFD,PO=R may be run to generate directives from the summary file which release the number of tracks required to meet the goal specified by the RL parameter. The summary file is specified as the input file (rewind it first).

3. Execute PFREL again, using the directives file generated by GENPFD as input (rewind it first).

A user is able to identify destaged files in the CATLIST command output. Filenames of destaged files appear in parentheses and summary information for these files is listed. When a request is made for a destaged file via GET, ATTACH, APPEND, or OLD commands, the *CHECK E,P DISPLAY* message flashes to alert the operator. Requests for staging tapes always appear as the first entry (or entries) on the E,P display with a JSN of SYS and a user name of (STAGE). After the tape is mounted by the operator, file data is subsequently restored to disk. If a direct access file is attached in write mode or an indirect access file is replaced, the alternate storage pointers and flags (both tape and MSE) are cleared in the file's PFC entry. Otherwise, the file is once again a candidate for releasing disk space.

4.4 PFDUMP Usage in a Tape Archive Environment

Because a large percentage of the total permanent file data typically does not reside on disk, performing full backup dumps of file data is impractical in a tape archive or MSE environment. Whether PFDUMP dumps the data of a selected file depends on the user-specified backup requirement, the PFDUMP options selected, and the number of alternate storage copies which exist. If these criteria do not require a data dump, only the file's PFC entry and permits are written to the dump file.

The pertinent PFDUMP options are:

OP=Y	Treat all files as though the user had specified a backup requirement of yes (BR=Y).
COS=size	Catalog (PFC) Only Size threshold. Dump only the catalog entry and permits if the file size is greater than or equal to the specified size (in disk PRUs) and the file's backup requirement is met by alternate storage copies.
OP=S	Suppress staging of file data from alternate storage if a data dump is selected and the file is not disk resident.

The decision hierarchy is as follows:

If OP=Y is specified, force the file's backup requirement to Y.

If the file size is less than the value specified by the COS parameter, select a data dump.

If the file size is greater than or equal to the value specified by COS:

- If the backup requirement is MD (media dependent):
  - If no alternate storage copy exists, select a data dump.
  - If at least one alternate storage copy exists on MSE or tape, select a PFC/permits-only dump.
- If the backup requirement is Y (yes):
  - If there are less than two alternate storage copies on MSE or tape, select a data dump.
  - If there are two or more alternate storage copies on MSE or tape, select a PFC/permits-only dump.

If a PFC/permits-only dump was selected, write the file's PFC entry and permits to the dump file.

If a data dump was selected:

- If the file is disk resident, write the file's PFC entry, permits, and data to the dump file.
- If the file is not disk resident:
  - If staging is not suppressed, stage the file data to disk from alternate storage and write the file's PFC entry, permits, and data to the dump file.
  - If staging is suppressed (OP=S), write the file's PFC entry and permits to the dump file.

A typical set of parameters a site would use for full and incremental backup dumps would be:

```
OP=Y   NOT specified.
COS    Set to some reasonably small value or zero
       (default is 0).
OP=S   Specified.
```

The OP=S option would not have to be specified if only tape alternate storage were in use and secondary VSNs had been created on all tape destages. Specifying a nonzero value of COS may reduce the amount of staging activity by users following a permanent file reload, since file data is disk resident for all files smaller than the threshold. Files larger than or equal to the COS value have to be staged back to disk by the user in such a situation since only the PFC and permits are reloaded.

When a partial PFDUMP is performed to move files to a different user name, family, or system, it may or may not be desirable to force a dump of the file data and clear the alternate storage information in the files' PFC entries depending on the situation. If files are being moved to a user name on the same family or to a different family on the same machine, there should be no need to dump the data of files residing only on tape alternate storage since the ability to stage files from tape does not depend on the family or user index under which a file is cataloged. If files are to be moved to a remote machine or if any of the files have an image on MSE, the file data should be dumped and the alternate storage information should be cleared either by PFDUMP or on the subsequent PFLOAD. This can be accomplished by specifying the following parameters:

```
OP=S   NOT specified.
COS=*  Select unlimited PFC only threshold.
OP=Z   Specified.
```

Unlimited COS must be specified to select a data dump for all files having BR=Y. OP=Z could be specified on the PFLOAD instead of the PFDUMP, but it is recommended that it be done at dump time so that the operator does not have to remember to do it at load time.

4.5 PERMANENT FILE UTILITIES - New and Modified Parameters

A number of the permanent file utilities have been enhanced for the tape alternate storage feature. Following is a list of the changes made to the various PFS utilities. The terms "directives" and "parameters" are used interchangeably, as all applicable options may be specified directly on the utility command statement as well as through the K display and new directives input file.

NOTE

In this section, "all utilities" does not include PFRES.

4.5.1 Utility Command I Parameter (all utility commands)

An I parameter has been added to each of the permanent file utility commands to specify a directives file name. Any valid command parameter or K display parameter option can be specified in the directives file with the exception of the I parameter itself. The directives file is typically used to specify multiple user indices and file names for processing.

4.5.2 Summary Output File Parameters (PFDUMP, PFREL, PFCAT, PFATC)

S=file name	Summary file name (default = SUMMARY).
SR=record name	Summary file record name (default = YYMMDD).

If S is specified, a file is written containing the PFC entries of all files processed. GENPFD (or a site written utility) can use this information to generate selection directives for permanent file utility operations such as file destage and disk space release or to obtain permanent file statistics for billing, etc. The SR parameter specifies the record name that is written to the prefix table of the record generated. The default record name is the current date in the form YYMMDD. For PFCAT, a summary file and an output file (L and LO parameters) cannot be specified simultaneously.

4.5.2.1 Summary File Format

The summary file created by the permanent file utilities consists of a sequence of blocks of the following types:

- Prefix table
- System information block
- Archive file identifier block
- Device status block
- Catalog image record (CIR) blocks
- Catalog entry blocks for all files selected for processing

This format allows the use of the CATALOG command to display information about the contents of the file and the environment in which it was created. For a complete description of each block and block layout, refer to the Permanent File Utilities section of the NOS Analysis Handbook.

NOTE

Since additional block types may be defined in the future, any program processing the summary file should be coded to read and ignore any blocks which have a block type other than those documented above.

4.5.3 Utility Command IP Parameter (PFDUMP, PFREL)

If the IP parameter is specified, PFDUMP or PFREL generates an output file and/or summary file, as determined by the L, LO, and S parameters, listing the files that would have been processed by that utility without actually dumping files or releasing disk space. A selection utility (such as GENPFD) can use the summary file to generate a directives file specifying file names to be processed by PFDUMP or PFREL without inhibiting processing.

4.5.4 PFCAT DN Parameter Processing

PFCAT does not require entry of the DN parameter if a summary file (S parameter) rather than an output file (L and LO parameters) is selected. In this case, all devices in the family selected or defaulted are processed.

4.5.5 PFC-Only Size Threshold Parameter (PFDUMP)

COS=nnnnnnn	PFC (catalog) only size threshold in sectors (default = 0).
COS=*	Set unlimited PFC only size threshold (forces a data dump of files).

This parameter determines whether PFDUMP attempts to process a file as PFC-only or forces a dump of the file data. Processing files as PFC-only reduces the time spent in dump and load operations and reduces the size of the archive file. For files larger than or equal to the specified size, only the PFC entry and permits are dumped if the file's backup requirement is met by a sufficient number of alternate storage copies. A file's backup requirement is met if the backup requirement is media dependent (BR=MD) and one alternate storage copy exists or if the backup requirement is yes (BR=Y) and more than one alternate storage copy exists. For all files smaller than the specified size and for files larger than or equal to the specified size which are not backed up on alternate storage, the data is dumped unless file staging is explicitly suppressed (OP=S) and the file is not disk resident. The COS parameter does not cause files with a backup requirement of none (BR=N) to be dumped. Specification of COS=* has the same effect as the OP=Y option for files with BR=MD, but OP=Y does not force a dump the data of BR=Y files with multiple alternate storage copies. Specification of COS=* is the only way to force a data dump of all files having multiple copies on alternate storage media. The COS parameter can be used on full and incremental dumps to dump the data of small disk resident files which are already backed up on alternate storage. This hopefully reduces staging activity when permanent files are reloaded after a disk failure.

4.5.6 LS and US (Size Limit) Parameters (all utilities)

LS=nnnnnnn	Lower file size limit in sectors (default = 0).
US=nnnnnnn	Upper file size limit in sectors (default = unlimited).
US=*	Set unlimited upper size limit.

These parameters define upper and lower size criteria for file selection. When specified, only files of size greater than or equal to the lower size limit and less than or equal to the upper size limit are processed. For PFDUMP, these criteria are checked before the COS parameter, if COS is specified.

4.5.7 Utility Command OP=T Parameter Option (all utilities)

This option, when specified with the BT, BD, AT, and AD parameters, causes file selection to be made according to the file's data modification date and time. Unlike the OP=M option, the control modification and utility control fields are ignored. OP=T cannot be specified simultaneously with OP=A, OP=C, or OP=M. This option is intended primarily to allow recently modified files to be excluded when creating a summary file for input to a destage file selection utility.

4.5.8 CA and CCA Directives (all utilities)

```
CA=R
CA=N

CCA=R
CCA=N
```

The CA directive selects files according to residence (having a current copy) on cartridge alternate storage. R selects files residing on cartridge alternate storage and N selects files which do not reside on cartridge alternate storage. A CCA directive clears the effect of its corresponding CA directive and is intended for K display use. When generating a summary file for input to a destage dump file selection utility such as

GENPFD, CA=N should be specified if cartridge alternate storage resident files are to be excluded from consideration.

4.5.9 TA and CTA Directives (all utilities)

```
TA=vvvvvv
TA=R
TA=N

CTA=vvvvvv
CTA=R
CTA=N
```

The TA directive selects files according to residence (having a current tape ASA) on the specified tape alternate storage VSN. A value of vvvvvv selects files which reside on VSN vvvvvv. R selects files residing on all VSNs and N selects files which do not reside on tape alternate storage. A CTA directive clears the effect of its corresponding TA directive and is intended for K display use. When generating a summary file for input to a destage-to-tape file selection utility such as GENPFD, TA=N should normally be specified unless some older tapes are to be recycled in which case, TA=vvvvvv entries may also be specified for these tapes.

4.5.10 Utility Command OP=Z Parameter Option (PFDUMP, PFLOAD)

The OP=Z option is valid with PFDUMP as well as PFLOAD. When specified, PFDUMP or PFLOAD clears the cartridge and tape alternate storage information in the PFC entries written to the dump file.

4.5.11 UN, UI, and CUI Directives (all utilities)

```
UN=user name
UI=user index
CUI=user index
```

The UN and UI directives specify a user name or user index to process. If no file selections are in effect for the affected user index at the time of entry, the UN or UI directive selects all files for the user index and specifies the user index to be

assumed for subsequent PF directives. If file selections are in effect for the user index, the UN or UI directive simply sets the assumed user index for subsequent PF directives. The CUI directive clears all file selections for the specified user index.

4.5.12 PF and CPF Directives (all utilities)

```
PF=file name
CPF=file name
```

The PF directive selects the specified file. Use of the PF directive without a preceding UN or UI directive is an error. The CPF directive deletes the selection of the specified file. Reentry of the PF directive for a selected file no longer clears the selection. If the deleted file selection was the only one for the user index, all files are selected for the user index. The CUI directive must be used to clear all file selections for the user index.

4.5.13 Utility Command OP=K Parameter Option (all utilities)

This option indicates that the K display should be brought up if an error is detected when processing input parameters and directives in PFS. If OP=K is not specified, the utility aborts if such an error is detected (which allows exit processing to be programmed within the calling procedure). Note that this is a change in behavior from previous systems; previous systems always brought up the K display on such errors.

4.6 DSD and IPRDECK Changes

4.6.1 E,P Display Changes

Requests for staging tapes always appear as the first entry (or entries) on the E,P display. These requests are displayed with a JSN of SYS and a user name of (STAGE). Requests for staging tapes appear only if one or more staging jobs are actually

waiting for tapes; the requests do NOT appear if all staging jobs already have tapes assigned, even if stage requests for other tapes are pending.

4.6.2 New and Changed Commands

The existing bit in SSTL for ENABLE/DISABLE PF STAGING is now used to enable and disable staging of permanent files from cartridge alternate storage. The existing DSD commands/IPRDECK entries:

```
ENABLE,PF STAGING.
DISABLE,PF STAGING.
```

are replaced by the new commands:

```
ENABLE,CARTRIDGE PF STAGING.
DISABLE,CARTRIDGE PF STAGING.
```

A new bit is defined in SSTL to enable and disable staging of permanent files from tape alternate storage. The following new commands/IPRDECK entries are processed by DSD and SET respectively:

```
ENABLE,TAPE PF STAGING.
DISABLE,TAPE PF STAGING.
```

4.7 New Account File Messages

A new SPSP message is issued to the account dayfile by PFM upon issuance of a STAGEPF request. It has the following format:

```
SPSP, filenam, , .
```

filenam The local file name in the calling FET.

A new STBS message is issued to the account dayfile by MAGNET on receipt of a stage request. It has the following format:

```
STBS, filenam, userin, fampack, vsn, r.
```

filenam The permanent file name of the file to be staged.

userin The user index under which the file is saved.

fampack The family name or pack name under which the file is saved.

vsn The VSN of the tape from which the file is to be staged.

r The number of retries for this stage request (zero if this request is not a retry).

A new STES message is issued to the account dayfile by PFRES upon successful completion of a stage request. It has the following format:

STES, filenam, userin, fampack, vsn, r.

filenam	The permanent file name of the file which was staged.
userin	The user index under which the file is saved.
fampack	The family name or pack name under which the file is saved.
vsn	The VSN of the tape from which the file was staged.
r	The number of retries required for this stage request (zero if no retries were required).

A new STTA message is issued to the account dayfile by PFRES when a staging tape has been successfully assigned. It has the following format:

STTA, vsn, fampack, numreq.

vsn	The VSN of the staging tape which has been assigned.
fampack	The family name or pack name to which the file is staged.
numreq	The number of staging requests pending for this tape at the time of assignment.

4.8 Changes To System Program Decks

New code and utilities added by the NOS tape alternate storage feature and reorganization of existing permanent file utility decks resulted in the following new OPL decks:

GENPFD	Select permanent files for destage to tape or disk release.
COMCPFP	Permanent file utility preset common routines.
COMCSRI	Stage request processing interface to MAGNET.
PFAM	Permanent file archive file management utilities.
PFDM	Permanent file disk management utilities.
PFHELPR	Helper job for managing stage requests.
STAGE	SYSTEMX-resident procedure for processing stage requests.

The following decks no longer exist:

PFATC	Replaced by an entry point in the new deck PFAM.
PFCOPY	Replaced by an entry point in the new deck PFAM.

4.9 System Analysis Overview

4.9.1 Stage Processing Overview

Stage processing is accomplished by PFM, MAGNET, PFRES, and RESEX; the basic flow of stage processing is as follows.

- When PFM decides to stage a file from tape, it sends a TDAM request to MAGNET and rolls the user job out (the original PFM request is reissued on rollin).

- MAGNET adds the TDAM request to a queue of stage requests. If the request is for a new tape (not a tape currently being processed by an active stage job) and the number of stage jobs active is less than the maximum, MAGNET calls DSP to initiate a new stage job.

- The stage job calls the STAGE procedure (a permanent file on SYSTEMX). This procedure includes a call to PFRES.

- PFRES does an internal LABEL request (with no VSN specified) to request a staging tape. This loads RESEX (as a DMP= program) on top of PFRES.

- RESEX does the following on a request for a staging tape:

  - Attaches the stage request file (STRQid) in write mode.

  - Gets all the unassigned stage requests from MAGNET and appends them to the stage request file.

  - Makes a list of all the VSNs for which stage requests have been made and sends this list to MAGNET.

  - If none of the tapes on the list is mounted, rolls out waiting for any of the requested tapes.

  - Picks one of the requested tapes and assigns it to the job.

  - Returns to PFRES, leaving both the tape and the stage request file assigned.

- PFRES does the following:
  - Reads, rewrites, and returns the stage request file, extracting all the requests for the VSN which was assigned. Puts the selected requests into a table.
  - Reads the tape, skipping files until it finds one for which it has a stage request. As it skips files, periodically asks MAGNET for more requests for the VSN which was assigned.
  - When a file is found on the tape for which there is a stage request, copies the file onto the appropriate device and does a SETDA (for a direct access file) or a UREPLACE (for an indirect). If this is successful, an event is issued to indicate that the file is now available. If there are no more stage requests for this VSN pending, the tape is rewound and returned.
  - When the end of the tape is reached, checks if any new requests were received from MAGNET which have not yet been processed. If there are, rewinds the tape and reads it again from the beginning.
- The user job rolls back in on the event and automatically reissues the original PFM request, which now succeeds (since the file is now disk resident).

4.9.2 PFDUMP Destage Dump Processing Overview

When destage mode is selected with the DT parameter, PFDUMP writes the PFM special request block information for each file dumped to a scratch file. This processing is similar to that currently used for the OP=P (purge after dump) option. After all selected files have been written to tape, PFDUMP processes the scratch file entries by calling the SETASA PFM function to set the tape alternate storage information in the PFC of each dumped file. Delaying the SETASA processing until after the dump and verify files are successfully written ensures that there are no PFCs with invalid ASAs if PFDUMP aborts or is dropped by the operator.

Because PFDUMP interlocks a catalog track only during the time that files cataloged on that track are being dumped, it is possible that PFM activity could modify a file between the time it was dumped and the SETASA processing. This is prevented by having PFDUMP call the PFU CTSL function while the catalog track is interlocked to set the TFLOK flag in the PFC entries of the affected files. PFM clears this flag whenever a direct access file is attached in a writeable mode or an indirect access file is replaced or appended. The SETASA function does not set the tape alternate storage information in a PFC entry if the flag is clear. This scheme ensures that multiple copies of PFDUMP do not attempt to destage the same file simultaneously.