To: Distribution

From: T. H. Van Vleck

Date: November 18, 1975

Subject: Contents of Initial New Storage System at MIT

INTRODUCTION

This memorandum describes the initial version of the new Multics Storage System, as it will be installed at MIT in early 1976. Our current target date is the weekend of January 18.

Over 140 programs are affected by this installation. In this memorandum we describe eleven functional changes to the system, and their implementation.

The consequences of the changes are discussed briefly. A final section describes planned functions which will not be in the initial version.

NEW FEATURES

1. Branch split into branch and VTOCE.

   Segment control is modified to obtain the file map for a segment from the VTOC entry instead of the branch when activating a segment. Boundfault is greatly simplified since no reference to the branch to reallocate the filemap is required. Deactivate need not touch the branch either. Segment control calls a new subroutine, vtoc_man, to request all I/O on the VTOC.

   Many directory control programs are modified to know that some items formerly in the branch were moved to the VTOC. A new program, vtoc_attributes, was created to get and set all such attributes. It has entry points for handling quota, date-time used and modified, maximum length.

---

2. Change to interpretation of disk address; new per-volume table; reorganized FSDCT.

Each volume now has its own address zero. All pages of a segment reside on the same volume. The AST entry for a segment contains an index into a new wired table, the Physical Volume Table (PVT), which contains disk subsystem and drive number, location of free map, and so forth for each volume in use. The FSDCT is now an array of bit maps for volumes, pointed to by PVT entries.

Volumes are identified in the branch by a 36-bit volume unique ID, which is looked up in the PVT to yield a PVT index when necessary. Volumes also have a 36-bit logical volume ID, which is used to select which physical volume to place a new segment on and to select the volume to which a segment will be moved if it cannot be grown on its current volume.

3. FSDCT now paged; pages withdrawn at fault time.

Making the FSDCT paged allows for very large disk configurations. To permit the unwiring of the FSDCT, page control was modified to check whether the needed page of the FSDCT was in core and to read it in first if necessary.

4. Directory locking changes; improve verify_lock; remove many cleanup handlers.

The lock primitive was rewritten to improve the speed and reliability of directory locking. The directory lock is no longer a location in the directory: instead, a new segment, dirlockt_seg, contains an array of lock entries showing the unique ID's of locked directories. The directory header page now is not stored into when the directory is only being read, which decreases paging. The module sum$dirmod must now be called before unlocking. The module modify_dir has been eliminated: programs which used to call modify_dir must now set and reset dir.modify themselves.

The dozens of places in directory control where verify_lock was established as a cleanup handler have been removed. Instead, signal_ has been modified to call verify_lock when crawling out of ring zero. Verify_lock has been rewritten to know about almost all locks explicitly and to attempt the salvage of databases before unlocking.

5. Simplification to AST locking

Besides the changes described above, special entrypoints in lock for locking and unlocking the AST eliminate the sst declaration from several programs which only referenced sst.astl.

6. Meaning of partition changes; volumes labeled.

The idea of the MULT partition disappears. Other partitions are constrained to reside entirely on one physical volume. The configuration deck is reorganized to have a DISK card for each disk subsystem and a PART card for each partition. PART cards now just name the disk drive; the partition extents come from the data on the disk volume's label.

Each volume has a label which identifies the volume and lists the partitions contained on the volume and their extents. Each volume also contains a free storage map showing which pages of the paging area are free and which are assigned, and storage for the VTOC for the volume.

7. Hardcore partition; volume accepting via ring 1; elimination of cold boot.

A special partition is required on the root physical volume (RPV), called the hardcore partition. Multics paging on hardcore segments and so-called "deciduous segments" (those recreated at every bootload) is done in this partition before the RPV free storage map is loaded into the FSDCT. This method allows the system to salvage the RPV without requiring the salvager to fit in core.

The system uses only the RPV until it has attained ring 1 Initializer command level. New commands to system_startup_ cause the initialization and mounting of the rest of the root logical volume and the other volumes in the configuration.

The concept of "cold boot" is whittled away to almost nothing. If, in collection 2, the system discovers that the RPV has a special indicator for the value of the root's VTOC index, and if the RPV is completely empty, a new root will be created. A BOS program is available to label an empty disk as a RPV.

8. Integrated salvager.

The salvager now has two parts. The volume salvager eliminates reused addresses and checks the validity of a single volume. The hierarchy salvager walks down the tree structure verifying and repairing directories. These two subsystems are part of the normal boot tape.

When the RPV label is read in collection 2, initialization calls the volume salvager if volume is not normally shut down. If the RPV is salvaged, the hierarchy salvager is also run in collection 2 to salvage the root and system_control_1. The system then crosses out into ring 1. The rest of the volumes in the configuration are volume salvaged before they are accepted for paging, if their label indicates that they were not shut down. The paging device is flushed onto each volume as it is mounted if necessary. Once all physical volumes part of the RLV have been mounted, the operator may type the SALV command to ring 1 to cause the hierarchy salvager to be called through a privileged gate to salvage the entire directory tree. If salvaging is successful, the operator can continue with a normal startup without rebooting.

9. Support for MSU-451 disks.

Changes have been made to disk_control and initialization to allow MSU-451 disks to be used, intermixed on the same controller with MSU-400's.

10. Better disk error recovery.

Certain cases of disk errors will now result in more orderly retry and recovery by the system. A new error may be reflected to the user if his segment has been damaged due to I/O error, where the current system gives the user an old page or a page of zeroes with no warning.

11. New rules for branch appending.

Each directory contains and inherits from its parent the logical volume unique id to be used when creating sons of the directory. A privileged call from ring 1 can set this identifier for an empty directory.

A call to append a branch can fail if no mounted volume has a logical volume id matching the sons_lvid in the parent directory. A call to append a branch can also fail if the access class of the segment being created does not fall within the AIM brackets for the volume on which the segment is being created.

CONSEQUENCES

The changes described above will have many effects on Multics operation. The following paragraphs describe the major consequences.

1. Need for more disk storage.

   As a result of the VTOC allocation strategy chosen, approximately five percent more disk storage is required to support the new storage system. In addition, if storage at MIT is subdivided into many logical volumes, additional breakage costs may arise.

   Handling of per-process hardcore segments is changed so that their pages are withdrawn at creation time. If a system is very close to running out of disk, this change will limit the number of processes which can be created, and will prevent some messy crashes.

2. Improved reliability.

   Because of the improvements to locking, emergency shutdown succeeds in a very high percentage of all system failures. Better handling of I/O errors should also pervent system crashes in some cases. When there is a head crash on a disk pack, it may be possible to avoid a complete reload.

3. Performance changes.

   Some operations in the new storage system will be faster, and some will be slower. The effects do not look too bad right now. An MTB is forthcoming on this subject.

4. Operator interface changes.

   Operations has to learn a lot of new things in order to run this system. First of all, the CONFIG deck is different. Also, the operators should have at least heard of a VTOC, and know which volumes contain directories. A series of MOSN's will be produced.

5. Administrative changes

There are not many administrative changes. If logical volumes are to be used some thought will have to be given to the use of the quota primitives on master directories. The quota for directory pages also needs definitior.

FEATURES NOT PART OF FIRST MIT INSTALLATION

1. Dynamic mounting.

This facility requires changes to RCP, new user commands for mounting, modifications to disk_table_, introduction of the LVRF, and the use of an ACS for logical volume access checking. Until this change is made all volumes will be "permanent" and need not be listed in the KST in order for a user to initiate segments on them. MTB-229 describes these facilities in more detail.

2. Dynamic demounting.

Changes are required for KST management and a method of forced deactivation of all pages of a volume must be invented.

3. Redundant directory structure and improved salvager.

Space may be reserved for the new fields, and they may be maintained, but no programs will depend on them. MTB-220 describes these in more detail.

4. New backup and recovery.

This change requires hardcore and ring 1 programs to be written. Space may be reserved in VTOC for the dumper data and it may be maintained, but no programs will depend on it. MTB-233 gives more information on this subject. Part of this facility may be made available soon after the installation at MIT; the plan is to run both dumping schemes in parallel while gaining confidence in the new method.

5. Master Directory Control.

This subsystem involves several user and administrative commands as well as the code for maintainging the MDCF in ring 1 for each logical volume. MTB-229 describes these facilities in more detail.