To: Distribution

From: T. H. Van Vleck

Date: December 17, 1975

Subject: New Storage System Long Range Plans (revised)

This document supersedes MTR-095. Although the installation of the new storage system on the CISL machine's mini-service was a few weeks later than planned, we remain confident that the final release date can be met. In fact, the date for installation at MIT has been advanced to January 18 to accomodate MIT's inter-term schedule.

OVERVIEW

The following table shows the major phases of the implementation of the new storage system.

Phase		Date
I	Command Level One user at command level	May 75
II	Prototype Running Several users	June 75
III	Design Review Error recovery, backup, mount/demount	Oct 75
IV	Installable System Run mini-service at CISL	Nov 75
V	Initial Installation at MIT No mount/demount	Jan 76
VI	Follow-up Installation at MIT Operational enhancements	March 76
VII	MR 4.0 Installed at MIT With mount/demount	April 76
VIII	Release MR 4.0	June 76
IX	Further Enhancements	-

Administrative Improvements

CHANGES SINCE LAST REPORT

Since MTR-095 was published, the following events have occurred:

1. Salvager. The interim salvager has been integrated into the running system and checked out.

2. Volume salvager. The volume salvager has also been written, checked out, and integrated.

3. New configuration mechanism. The system now bootloads on only the RPV until it crosses out into ring 1, so that all other volumes can be checked by the registration software.

4. Very large configurations. The support for a paged FSDCT was put into page control earlier than planned. The number of drives in the configuration is no longer constrained by the amount of wired core.

5. Error recovery. Improvements were made to system error recovery strategies in several modules. The most important of these was the change to allow a segment to be moved from one physical volume to another if the physical volume ran out of space.

6. The schedule for the new backup has been rethought. Investigations of the performance of the interim backup system show that it will not be unacceptable. And although the hardcore and the dump and reload driving programs for the new backup are not difficult in principle, the complete new backup will require a significant amount of additional work to convert it into a total facility. For example, programs to manipulate and interrogate the system's dump-tape logs, programs to translate retrieval requests from pathname form to unique ID form, and operator documentation describing the new procedures must all be produced. In release 4.0, we may be able to eliminate SAVE in favor of new-style complete dump and reload done while the system is up; but it seems unnecessarily risky to replace the entire backup/reload subsystem with a new one if it has not had sufficient exposure.

7. The new storage system was installed on the CISL machine mini-service on December 1, 1975.

WORK COMPLETED

1. Statement of Problem.

   MTB-017, November 1973.

2. Preliminary Design.

   MTB-055, April 1974
   MTB-060, May 1974
   MTB-065, April 1974
   MTB-095, June 1974
   MTB-110, August 1974
   MTB-167, February 1975
   MTB-203, June 1975
   MTB-206, June 1975
   MTB-213, July 1975
   MTB-220, September 1975
   MTB-221, September 1975
   MTB-229, October 1975
   MTB-233, November 1975
   MTB-237, November 1975
   MTB-238, November 1975
   MTB-239, November 1975
   MTB-243, December 1975

3. Preliminary Task Schedules.

   MTR-068, October 1974
   MTR-081, March 1975
   MTR-084, April 1975
   MTR-095, Sept 1975

4. Phase I: Command Level.

   When this benchmark is reached, the system can be bootloaded from a Multics system tape, either cold or warm, come up to Initializer command level, and shut down. Only one disk need be used; but it has a standard label, VTOC, and volume map. Paged I/O is used for the VTOC. A new version of BOS is required to support the new configuration deck.

   Target date: May 1975. Finished.

5. Phase II: Demonstrable System.

   When this benchmark is reached all functions of the current Multics work in the new system, with the exception of minor bugs and certain metering tools. Since the VTOC is still

accessed by means of paged I/O, 1K per volume of page table for VTOC image is wired, plus the 512 words per volume of volume map. No backup or salvager is implemented. Although much more interesting in terms of function, this stage is not very difficult to accomplish because the system initialization path checks out almost all of the storage system.

Target date: June 1975. Finished May 1975.

6. 64-word I/O Facility.

This facility is used to transport VTOC information and volume map data between core and the disk. Changes were made to the disk DIM, and the module vtoc_manager_ was written for the management of the memory devoted to 64-word data.

Finished.

7. New VTOC Manager.

The new VTOC manager uses the 64-word I/O facility. This change frees 1K per volume of wired core, and decreases the I/O channel time and latency time for requests for data in the VTOC.

Finished.

8. Smaller VTOC Entry.

The interim VTOC entry in use up to this point has 256 words instead of 192, in order to simplify the code for the deactivation of 256K segments. This stage complicates the code but reduces the size of the VTOC by 25%.

Finished.

9. Study and Definition of Backup Problem.

This activity defines the key variables to be optimized in subsequent backup design. See MTB-203 for details.

Finished.

10. New Directory Locking.

This step creates a wired hardcore table with one entry per active directory. The directory lock is kept in this table. Directories need not be modified to be locked and unlocked as a result of this change; this reduces the paging traffic significantly.

Finished.

11. Interim Version of BOS SAVE and RESTOR

An interim version of BOS SAVE and RESTOR has been written which handles single volumes only. It is now being used to support performance testing.

Finished.

12. Hardcore Partition.

The system bootloading sequence is altered by this task to use a special area of the root physical volume for all paging needed before directory control initialization, as described in MTB-213.

Finished.

13. Multics Utilities for Pack Maintenance.

These programs initialize a disk, set up VTOC entries, volume map, and write and check labels.

Finished.

14. Consistent Directory Locking.

The lock primitive and page control have been modified so that so that modified pages of directories which are locked are flushed from core when the directory is unlocked. This strategy couples with the double-write for directories mechanism to keep the disk copy of directories as consistent as possible. See MTB-239 for details.

Finished.

15. New Configuration Strategy.

The new configuration deck mechanism described in MTB-213 is now operational. The mechanism chosen appears to integrate nicely with the eventual RCP strategy.

Finished.

16. Performance Measurements.

Initial performance measurements showed that the new system's performance was the same as the installed system's, to within the accuracy of the test. Further measurements pointed out some inefficiencies in the new system which were remedied. An MTB describing the strategy used for performance measurements will be published.

18. Interim Salvager.

The current system's salvager has been integrated with the system startup sequence and modified for the new structure of directories. When salvaging is necessary, The salvager is invoked to salvage crucial directories only from ring 0 during startup, and invoked again by operator command after ring 1 has finished mounting the RLV. See MTB-221 for a discussion of the salvager.

Finished.

19. Design for Backup.

This design presents the long-term plan for the evolution of the system's backup capabilities. Two modes of data recovery are required, one for reconstruction of the contents of a complete physical volume (or group of physical volumes), and another mode for retrieval of the contents of a single segment. MTB-233 describes the new backup.

20. Design for System Recovery Modes.

A monolithic salvager subsystem becomes more and more unwieldy as the size of the configuration increases. The proper solution is to improve the system's in-line error detection and dynamic salvaging code. The emergency shutdown, backup, crawlout, on-line salvager, and directory locking facilities will be redesigned into a coherent and complete package. See MTB-220 for more information.

21. Design for Volume Mount/Demount.

The design of the Resource Control Package (RCP) has been extended to handle the management of logical volumes as well as physical disk packs and disk drives. Both logical and physical volumes will require registration data which must be consulted before a logical volume can be mounted for a user. MTB-229 presents an overview of the design.

22. Phase III: Design Review.

This review, held in October 1975, covered the design of the error recovery modes, backup, and the volume mounting and demounting modules. Overall reaction was very good, and several useful suggestions were made about details of the design.

23. Dynamic Physical Volume Mounting and Accepting.

This step finished the implementation of the new configuration strategy described in MTB-213. Privileged calls from ring 1 allow the operator to add volumes to the

storage system configuration while the system is running.

Finished.

45. Pageable Volume Maps.

Page control has been modified to allow for the possibility that the page of the FSDCT which contains the volume map for a given volume may not be in core. This change allows the use of very large numbers of disk drives without requiring large amounts of wired core.

Finished.

27. Specifications for Command Changes.

Many minor changes will have to be made to the command system. Quite a few of these can be done ahead of time if a document setting forth the standards for system commands and subroutines is published. MTB-243 describes the necessary changes.

28. Error and Exception Handling Improvements.

The system is able to recover from the cases of "no more VTOC entries on the volume" and "no more pages on the volume." To handle the second case the supervisor must move the segment to another volume in the logical volume which has sufficient room.

48. Disk DIM Error Handling Improvements.

Improvements have been made to the disk error handling programs. The system's retry strategy now takes account of the type of error and the previous error history for the drive. These improvements are described in MTB-239.

30. Run Mini-Service at CISL.

Until we actually run a "service" of some sort, we will not know what the performance is really like and what operational improvements are required. Installation at CISL also allows other projects to integrate with the new storage system and decreases the number of changes which must be made to two versions of Multics.

The mini-service can be started without the availability of the new backup or the new salvager.

Service, about 5 hours a day 5 days a week, began on December 1, 1975.

CURRENT TASKS

17. Interim Backup.

    This task modifies the current backup programs to dump and reload the new directory quota cell and the logical volume ID for a directory. This change allows the current incremental/catchup/complete dumper to be used for backup until a new version is designed and built.

24. Implementation of Hardcore Primitives for Backup.

    The hardcore primitives to support the new backup system must be able to maintain the list of modified segments on each physical volume for the use of incremental dumping; and to activate and dump or reload a segment by volume ID and VTOC index without referencing the branch.

25. Implementation of Backup Dumping Programs.

    The new complete and incremental dumping programs can be much simpler than the current dump programs, since all hierarchy walking and access forcing code is eliminated. The hardcore primitives do most of the work. These programs are easy given the format of the output records to be produced.

26. Implementation of New Reload and Retrieve.

    The reloading and retrieval programs will use the output of the dumping programs to reconstruct volumes and to recover the contents of single segments.

29. Improved Directory Format.

    This task redesigns the directory to be more easily verified for correctness. All storage system modules which reference the directory must be recompiled with the new declaration. The various redundancy checks are not inserted by this task, though. See MTB-221 and MTB-220 for details.

31. New Directory Salvager

    Rewrite salvager to operate on a new expanded directory structure, without reference to the VTOC entry.

32. Directory Control Checking.

    This task adds to directory control new code for maintenance of the various redundancy fields added to the directory structure, and appropriate in-line checks and repair operations. MTB-220 describes the details of this change.

33. Phase IV: Make System Installable.

Once system reliability and performance are acceptable, the new storage system is ready to be installed at MIT.

The first version of the new storage system to be installed at MIT will not have all the functional improvements which will be provided with release 4.0. In particular, the final salvager system is not required, and the interim backup will be used. The ability for a user to request the mounting of a logical volume will not be present in this version of the system. What will be provided is the reformatting of disk storage and directories and the consequent improvements in reliability.

MTB-238 describes the contents of this intallation.

Target date: January 1976.

34. Formalities of Submission.

This step covers filling out submission forms, auditing of all programs, running final performance runs, fixing last-minute problems, etc.

35. Phase V: First Installation at MIT.

Target date: January 18, 1976.

37. Master Directory Operations.

This task adds ring-1 support for operations on master directories. User calls are create, delete and list. The create_dir and delete_dir commands need modification for this case.

The ring-1 programs will use the Logical Volume Registration File (LVRF) and the Master Directory Control Segments (MDCSs). Administrative commands to manage these data bases are necessary. For the volume librarian, we need register, unregister, modify, and list. For the volume administrator we need permit, deny, and list.

38. Ring 1 Volume Mount Module.

When a logical volume is to be mounted, the LVRF must be consulted to find the list of physical volumes to be mounted. Calls must then be made to RCP to mount each of the physical volumes, the volume labels must be checked, and the hardcore must be called to tell it that the volumes are accepted.

39. User Request to Connect Logical Volume.

User requests to connect to a logical volume will be passed through RCP. If the user process is permitted to connect to the logical volume, the hardcore will be informed of the connection, a counter associated with the logical volume will be incremented, and a mount request for all physical volumes will be issued, as described above, if they are not already up. The logical volume will be disconnected when the connection count goes back to zero as a result of users unassigning the resource or logging out, or when the operator forces the unassignment.

40. Hardcore Check on Volume Connection.

The hardcore will be changed by this task to require the connection call from ring 1 before allowing a process to initiate a segment on a demountable volume. (The root logical volume is never demountable and other "public" volumes can be declared not demountable.) This insures that RCP is not bypassed, and makes sure that all programs using segments on removable volumes execute independently of whether some other process has caused a pack to be mounted. The list of demountable physical volumes which the process is connected to will be stored in the PDS or some other per-process data base.

41. Phase VI: Follow-up Installation at MIT.

Operational experience will lead us to make many improvements to the interface and behavior of the storage system. Performance measurements under actual load may also show use where to concentrate our programming effort in order to speed the system up; if these improvements are possible we will install them soon.

Target date: April 1976.

42. Command System Changes.

These changes are the ones specified in paragraph 27. In addition to the changes to handle new error and state conditions, the create_dir command must accept and check the new parameter which specifes the logical volume in which storage will reside, and the list and status commands must be modified to show this attribute.

43. Phase VII: Install MR 4.0 at MIT.

Target date: April 1976.

44. Phase VIII: Release MR 4.0

Target Date: June 1976.

FURTHER ENHANCEMENTS

36. Backup Integration.

This task integrates the new backup mechanisms into the system and ties backup in with salvaging.

Although most of the parts of the new backup system will be available by the time release 4.0 is frozen, the new backup may not have had sufficient testing and operational experience to allow us to depend upon it.

46. Keep Duplicate Copies of Selected Volumes.

Once this task is completed, crucial volumes in the system can be maintained in duplicate; all modified pages will be written out to both devices. In a configuration which places the secondary copy on a different disk subsystem from the primary copy, the cost of maintaining two copies will be very low.

47. Automatic Use of Secondary Volume on Error.

Once the duplicate copy facility is available, the system can be modified so that when a disk record is unreadable, the system automatically switches to the use of the secondary copy.

49. Calls to Initializer Process During Connection.

This step causes RCP to pass all connection requests through the system control process, so that charging can be done, mount messages can be routed, and so that operator commands affecting the request can be issued.

50. Billing.

Modifications must be made to the administrative and billing package to enhance the administrator's ability to manage the system resources. Some of these improvements cannot be specified until we have obtained some operational experience.