MULTICS TECHNICAL BULLETIN MTB-055

TO: Distribution

From: T. H. VanVleck, S. H. Webber, A. Bensoussan

Date: March 12, 1974

Subject: The New Storage System--Overview

The purpose of this memo is to record the most important points that have been discussed during a few meetings on a possible new design for the storage system. The problems in the current system and the goals for the new system are described in MTB-017.

1. The new disc organization

The significant points in the new disc organization are the following:

a. All pages of a given segment are on the same volume. This is the first step toward grouping physically data that is logically closely related. Also, it is absolutely required for the implementation of removable segments.

b. Each volume begins with a "Volume Table of Content" or VTOC, containing the list of all the segments stored in the volume with their physical attributes. These attributes are: the primary entry name of the segment, the unique identifier, the date and time used, the date and time modified, the current length, the file map and the AST entry pointer. These items, which, in the current system are part of the branch in the directory, will be located in the same volume as the segment itself. They will be replaced, in the branch, by a pointer to the VTOC entry where they are stored. This pointer consists of a disc uid and the VTOC entry number.

The existence of the VTOC makes it possible to tell what is stored on a given volume, even though the full pathnames are not recorded in the VTOC entry. Also, it makes it possible to check the consistency of the volume, without the need for the directory hierarchy.

c. Each page of a volume has a descriptor containing the uid of the segment to which the page belongs, with the page number within the segment. All page descriptors are collected together at the beginning (or at a standard location) of the volume in an array called the volume map.

The volume map makes it possible to detect any file map inconistency. Most important, it makes it possible to correct some reused address conflicts, particularly those conflicts caused by the hardware where a bit is "dropped".

2. Volume Assignment

When creating a segment, one can specify the volume on which this segment is to be stored.

A default value kept in the directory will be used if no volume is specified.

Each volume is identified by its 36-bit volume uid, and also by its symbolic name. A catalog of volumes, similar to the catalog of tapes, will be maintained by ring 1 procedures, in order to associate a volume name with its uid, its location, its owner and the list of persons that can use it.

It is also possible to designate by a symbolic name, a group of physical volumes. In this case, the catalog entry contains, under the same name, the list of all the volume uid's that are part of the volume group.

A user can create a segment in any volume, provided that the volume be on line and the user have the right to use it. The idea of mapping a subtree of the hierarchy into a volume was considered and rejected because it does not correspond to the natural way one likes to group segments on a volume in practice.

3. Removable Volumes

The new disc organization and the volume assignment capability provide the basic mechanisms for implementing the removable volume facility.

An attach/detach command will be provided to put a volume on line and off line. These operations have to be explicitely requested; the system will not be responsible for issuing the attach command in response to a missing volume detection. The attach and detach operations are done on a volume group basis, i.e., all physical volumes that are part of the same volume group will be on line or off line at the same time.

In order to be able to use a segment at all, the segment and all its parent directories must be on line. To guarantee that all on-line segments will be reachable, it would be desirable to force all directories to remain on line. The proposed manner by which this rule would be enforced is to allocate all directories in the same volume, that the system would always keep on line. Keeping directories always on line also makes it less difficult to validate a detachable volume at attach time. Keeping all directories in the same volume makes it easier to implement the directory duplication, method that is proposed as a substitute for incremental back-up.

Even if directories always remain on line, the attach operation requires some validation of the volume being mounted. The VTOC and the volume map can be checked for consistency; some inconsistencies may even be corrected. However, it is not clear how one can detect a modification that has been done off line and which could not have been done on line. One may use check sums as a partial solution. One may also consider making the distinction between system volumes, which never leave the computer room, and user volumes. User volumes that have been detached would be, when reattached, imposed some restrictions. For example, they would not be allowed to contain segments with ring number smaller than 4. There does not seem to exist a simple solution to the validation problem.

4. The Quota Problem

A new method to manage quota and accounting will be documented in a subsequent MTB. The primary reason for the redesign is that, in the current system, the system administrator needs too much privilege in order to perform quota and accounting function.

The new scheme will not be described in this memo, but it has to be noted that it does not assume the existence of the new storage system; it could very well be implemented with the current storage system.

More consideration has to be given to quota and accounting with regard to removable volumes.

5. Back-up

The incremental backup definitely consumes too much computer resources. It seems that Multics spends more time than other systems in back-up and yet looses more data. It is therefore questionable whether or not the incremental back up is the answer.

In the new design, the following approach is proposed:

- Directory hierarchy duplication
- No incremental back up
- Still daily and weekly back up
- Back up of individual segment on explicit request

All directories are in the same volume (or volume group). Page Control maintains on an auxiliary volume, an exact copy of the volume containing the directory hierarchy. Each time a page of a directory is written out, page control also writes it on the auxiliary volume, at the same disc relative address. If the directory hierarchy is damaged by a system crash, the system can switch to the copy stored on the other volume.

6. Advantages and Disadvantages

The most important advantages that the new storage system provides over the current one are the following:

- Each disc is self describing
- One can group a collection of segments on the same volume
- Provides removable volumes
- Less data will be lost since reused address is eliminated
- Even if reused address occurs, the conflict is internal to the Volume
- Disc consistency can be checked off line and corrected in some cases.
- Less overhead for back up since no incremental back up
- Directories are smaller
- Directory duplication is expected to reduce the number of segments lost where a directory is damaged or badly salvaged

The disadvantages that can be identified are the following:

- More complicated and very big job
- VTOC addressability must be provided (1000K words per disc)
- More page faults or disc I/O can be expected when accessing segments attributes in the VTOC than in the directory
- More work has to be done at activation-deactivation for avoiding reused address and for accounting. That is, more overhead and more page faults
- Overhead in directory duplication

7. Conclusion

It seems that the new design would increase the reliability of the system, would make physical dumps of the discs useful while in the current system only logical dumps are useful and would add the capabilities of physical segment grouping and removable volumes. On the other hand it seems that the new design introduces additional overhead in the management of segment attributes, mainly in segment creation, segment deletion, activation and deactivation. A study is being done to try to quantify this additional overhead.