TO: MTB Distribution

FROM: Bernie Greenberg

DATE: 03/14/78

SUBJECT: Duplicate Disk Volumes

This memo presents a discussion of the subject known as "Shadow copies", or duplicate on-line disk copies, for recovery purposes. A design discussion will be held in the near future.

The response of Multics software to failures of disk drives and packs is a major remaining reliability problem. Maintaining two or more copies of critical volumes on line, in parallel, has been proposed. In case of failure, the duplicate volume may be switched to, these duplicates being guaranteed by the system to be as up-to-date as the failing unit.

This strategy has several immediate consequences: increased channel usage, heavier usage of disk queue entries, a slight increase in CPU time to check for this double-writing and to perform it, as well as some increase in the complexity of the supervisor.

The disk table management software must be changed to accept willingly two physical volumes of the same Physical Volume Identifier (PVID). Whether or not duplicate packs should have the same physical volume name is an open issue; clearly they must have the same PVID. These issues are open to some debate, but solutions do not seem difficult.

The large problem on which I seek input is that of when to use this duplicate volume. At the time that major-order disk problems are detected, the system is usally choking in error messages. Even were there some way to suppress these messages, the Initializer or other processes would be trapped in error processing, or inability to access the disk drive at issue. Were a decision to be made on the fly to use a "backup copy" of some disk in such a case, it would have to either be made automatically, or some new form of conveying information to the supervisor must be developed.

I have contemplated the construction of a facility to give commands to the hardcore ring via the BOS typewriter, implemented in wired code, running on the operator's console interrupt side. Clearly, among the various commands amenable to such treatment, (e.g., patching, dumping, crashing, ESD'ing), abandoning disks ranks as a prime candidate. The construction of such a facility is reasonable, but a fairly large task in terms of interface design.

It may be reasonable to allow the use of copy disk volumes via abandoning the old through any means now available, such as crashing, and using the new pack, knowing that it is as up-to-date as the old. This approach requires no less work in terms of disk support, but dodges the issue of dynamic switching to the new volume.

I hesitate to allow the system to make the decision to switch to a backup-copy drive. Once the decision is made by any agency to do so, the old drive becomes useless. One envisions any flurry of disk errors as triggering such an automatic switchover. This allows precisely one such flurry per bootload per volume. I cannot see now an acceptable set of heuristics, computable by the disk control routines in real-time, which meet the criteria of utility that this facility must provide. The obvious "x many errors of y type in z time" is both difficult to specify and would probably become a continual source of frustration and near-miss design.

The design and implementation of the ring zero software to maintain duplicate copies is straightforward in terms of the current page control, disk control, and VTOC management.

I need input on the following questions. Either respond to me by phone (HVN-261-9330) or mail (Greenberg.Multics on MIT or System M). Please be prepared to discuss them at a design review:

1. Should the system switch over to a backup-copy volume automatically? Although the only choice during unattended operation, should this be so during attended operation?

2. If so, what criteria should it use to do so? Automatic switchover can be an option, or a set of options as well.

3. If not, how shall we convey to the system to use such a volume during times of extreme non-communicativeness and system stress?

4. If not, is this the time to implement a "talk-to-ring-zero no matter what" facility? If so, what else ought be in it?