THIS SHEET HAS BEEN REPLACED

A B S T R A C T

A fault-tolerant high performance mirrored disk subsystem is described which has an improved disk writing scheme that provides high throughput for random disk writes and at the same time guarantees high performance for disk reads. The subsystem also has an improved recovery mechanism which provides fast recovery in the event that one of the mirrored disks fails and during such recovery provides the same performance as during non-recovery periods.

Data blocks or pages which are to be written to disk are temporarily accumulated and sorted (or scheduled) into an order (or schedule) which can be written to disk efficiently, which in a preferred embodiment is in accordance with the physical location on disk at which each such block will be written. This also generally corresponds to an order which is encountered by a write head during a physical scan of a disk. The disks of a mirrored pair are operated out of phase with each other, so that one will be in read mode while the other is in write mode. Updated blocks are written out to the disk that is in write mode in sorted order, while guaranteed read performance is provided by the other disk that is in read mode. When a batch of updates has been applied to one disk of a mirrored pair, the mirrored pair switch their modes, and the other disk is updated. Preferably the updates are kept in a non-volatile memory, which furthermore advantageously may be made fault-tolerant as well.

During recovery a pair of spare alternating mirrored disks is introduced to which new updates are directed, while a background scan process copies data from the surviving disk to the new mirrored pair.

This Sheet Has Been Replaced