How did MS-DOS decide that two seconds was the amount of time to keep the floppy disk cache valid?

devblogs.microsoft.com/oldnewthing/20190924-00

September 24, 2019

Raymond Chen

MS-DOS 2.0 contained a disk read cache, but not a disk write cache. Disk read caches are important because they avoid having to re-read data from the disk. And you can invalidate the read cache when the volume is unmounted.

But wait, you don't unmount floppy drives. You just take them out.

IBM PC floppy disk drives of this era did not have lockable doors. You could open the drive door and yank the floppy disk at any time. The specification had provisions for reporting whether the floppy drive door was open, but IBM didn't implement that part of the specification because it saved them a NAND gate. Hardware vendors will do anything to save a penny.

But that read cache is crucial for performance. Without it, you have to start from scratch at every I/O operation, re-reading the volume table of contents, finding the directory entries, searching the block allocation tables looking for the next free cluster… And a floppy disk is not exactly the fastest storage medium out there, so all of these operations cost *seconds* of performance.

To avoid having to abandon the cache entire, the MS-DOS developers did some benchmarking: How fast can a human being swap floppies in an IBM PC floppy drive?

Mark Zbikowski led the MS-DOS 2.0 project, and he sat down with a stopwatch while Aaron Reynolds and Chris Peters tried to swap floppy disks on an IBM PC as fast as they could.

They couldn't do it under two seconds.

So the MS-DOS cache validity was set to two seconds. If two disk accesses occurred within two seconds of each other, the second one would assume that the cached values were still good.

I don't know if the modern two-second cache flush policy is a direct descendant of this original office competition, but I like to think there's some connection.

Raymond Chen

Follow