Is Ordering of Disk Updates Required to Maintain File-System Crash-Consistency?

ABSTRACT

On reboot after a crash, the file system should be consistent: e.g., previously correct files should not now contain garbage. In early file systems, getting to a consistent state involved a full scan after reboot. This was very slow, and impractical for large systems. Modern file systems improve upon this by writing updates to disk in a *specific order*: e.g, metadata before commit blocks. This allows them to get to a consistent state without a scan.

However, ordering updates results in certain problems:

1. The file system write order may not be the most efficient order for writing blocks to disk. This reduces performance.
2. The file system has to be very careful about the order; this increases complexity, potentially leading to more bugs and lower reliability.
3. For disks with write caches, commands such as cache flushes are required to ensure correct ordering. If such commands are not properly implemented, consistency is compromised [1].
4. In virtualized stacks, even if *one* of the many layers between the file system and the disk does not enforce ordering, consistency is lost.

The question then arises: can crash-consistency be maintained without ordering updates? Recent work introduced the **No-Order File System** (NoFS) [2], the first file system to provide strong consistency despite not ordering updates. NoFS uses a novel technique called **Backpointer-Based Consistency** (BBC) that establishes consistency via mutual agreement between file-system objects. NoFS performs as well as a comparable journaling file system (ext3) for most workloads, and increases throughput by 20-70% for metadata-intensive workloads.

BODY

Not only is it possible to maintain file-system crash-consistency without ordering updates, but doing so may actually increase performance.

REFERENCES

[1] Rajimwale, A., Chidambaram, V., Ramamurthi, D., Arpaci-Dusseau, A. C., and Arpaci-Dusseau, R. H. Coerced cache eviction and discreet mode journaling: Dealing with misbehaving disks. In *Dependable Systems & Networks (DSN), 2011 IEEE/IFIP 41st International Conference on* (Hong Kong, China, June 2011), IEEE, pp. 518–529.

[2] Vijay Chidambaram, Tushar Sharma, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. Consistency Without Ordering. In *Proceedings of the 10th Conference on File and Storage Technologies (FAST '12)* (San Jose, California, February 2012).

Volume 2 of Tiny Transactions on Computer Science