Thesis Overview:

Design, Implementation, and Evaluation of a New Journaling File System with Data and Meta-data Separation

Juan Piernas Cánovas
Universidad de Murcia (Spain)
Departamento de Ingeniería y Tecnología de Computadores
Advisors: Toni Cortes and José M. García Carrasco
October 29, 2004
piernas@ditec.um.es

This thesis describes the design, implementation and evaluation of a journaling file system called *DualFS*. Our main goal is to improve disk I/O by designing a new file system which must have two features: a better performance than that of current file systems, and a fast consistency recovery after a system crash.

The design of the new file system is based on both a total separation of data and meta-data blocks, and a specialized treatment of meta-data blocks. Separation is achieved storing blocks in two different devices: the *data device* and the *meta-data device*. These devices are commonly two adjacent partitions of the same disk, although they can be partitions of different disks.

The data device is divided into several groups where data blocks of "regular" files are spread out. Grouping is performed in a per directory basis: data blocks of regular files created in the same directory are put together in the same group assigned to the directory (or in near group if the corresponding group is full). The meta-data device, however, is a *log-structured file system* where meta-data blocks are efficiently written in a sequential manner. The meta-data log used by DualFS does not only allow it to recover its consistency quickly after a system crash (as it occurs in traditional journaling file systems) but also it greatly improves meta-data operations and, hence, the overall file system performance.

Exploiting particular features of the new design, we have added three mechanisms to DualFS which are aimed at improving its performance even more. These mechanisms are: *directory affinity*, *meta-data prefetching*, and *on-line meta-data relocation*. The first mechanism reduces the number of seeks in the data device, and therefore the time taken by data read/write operations. The second one improves meta-data read operations which in turn allows DualFS to improve the read of regular files. In order to be I/O-time efficient (and to avoid extra I/O requests which can produce long seeks), prefetching is carried out when the file system starts a disk I/O request for reading a meta-data block which is not in memory; DualFS prefetching also involves reading a group of consecutive meta-data blocks from disk where it is the requested meta-data block. Finally, the third method is used to avoid degradation of prefetching efficiency by increasing spatial and temporal locality in the meta-data device. This is obtained by putting together on disk all meta-data blocks of a file, and meta-data blocks of files which are read at the same time.

After implementing and tunning the new file system in Linux, we have compared its performance with that of other general purpose file systems which are extensively used nowadays: Ext2, Ext3, XFS, JFS and ReiserFS. Experimental results show that DualFS can reduce time of I/O operations up to 98 %, and that it is, in most cases, the best file system to manage workloads which are very common in current environments, such as web, mail and news servers, program development, etc. For example, in the PostMark benchmark, which models the workload seen by Internet Service Providers under heavy load, DualFS achieves 60 % more transactions per second than Ext2 and Ext3, twice as many transactions per second as ReiserFS, almost three times as many transactions per second as XFS, and four as many transactions per second as JFS.