Masaryk University
Faculty of Informatics

Comparison of filesystems in Linux

PV208 Advanced Topics of Linux Administration

Bc. Pavol Babinčák

Brno, April 2008

Introduction to filesystems

Main purpose of filesystems is store data and allow operating system to find and access to them. This includes support for hierarchical organization, manipulating and data retrieval. On the top of that filesystems provides unique properties and services like consistency check, snapshot of all data, support for high availability, failure resistence, etc.

Filesystem data structures can be used on various storage media from virtual filesystems in memory through magnetic devices such as hard disk to read only media like DVD ROM. Connection method between storage media and logic which understand format of file system are from local bus in PC to network relayed devices like SAN (Storage area network).

There exist many different filesystems which can be divided into categories by level of suitability for these storage media and connection method. In this paper I have chosen two categories and three file systems from each of them. For every filesystem I will focus on some interesting features which provides that filesystem. All of mentioned filesystems can be used on Linux system with or without additional effort in the form of patching kernel.

Disk file systems

Disk file systems are connected to PC by local interface like (P|S)ATA. This connection method is most widely used for common PC and for some reasons in servers. It is cheap, relative fast but without use of other technologies (like RAID) it easily affected to hardware failures. In same time only one PC (operating system) is supposed to access to this file system so there is no need to consider concurrent access.

Extended file system (ext2, ext3)

Extended file system was designed for Linux needs. Second version of filesystem (abbreviated as ext2) was default on many Linux distributions in past. Now same applies to third version (ext3). Therefore it is used by large amount of computers and is probable that most of bugs in drivers for this filesystem were found. It is very stable filesystem and has good support from developers. Unexpected reboot can hurt filesystem consistency when buffered data are not written to disk. This is found on computer boot and filesystem check is run automatically. In second version this repair is quite slow for large filesystems. Third version of filesystem is ext2 with journaling support. Journaling is used to solve problem of slow consistency check. [from now on all mentioned filesystems has journaling support] Ext3 is backward compatible with ext2 however it has poor performance for some metadata operations. Some of interesting features are immutable files (which can not be deleted only read), append-only files (data can be only appended---good for log files).

ReiserFS

Most mentioned feature about ReiserFS is excelent performance with large amount of small files. Filesystem is not so stable and it has less support from developers. Third version

is in mainline Linux kernel, fourth version not in kernel has features like plugin driven architecture and transaction support.

XFS

XFS has good performance for large files, directories with many entries and big filesystems. Filesystem repair in case of failure is quite slow similarly directory entries creation/deletion. Some of features are delayed allocation for reducing fragmentation and native backup/restore utilities able to make FS dump without unmounting.

Shared disk file systems

Shared disk filesystems are also known as SAN or cluster filesystems. SAN is often used architecture for sharing access to hard disk. This type of access to disk are used in clusters so hardware failure of computer node does not affect availability of data on hard disks because another node can substitute failed node.

Global File System

Global file system (GFS) uses peer to peer architecture where all nodes are equal in controlling access to shared resources. It has support for databases allowing direct I/O operations. With dynamic multi-path routing data can be routed around failed components.

General Parallel File System

General Parallel file system (GPFS) is proprietary filesystem used on very large clusters (up to 2000 nodes) for high performance computing and grids. SQL (structured query language) syntax is used for defining placement and management policies for files. GPFS offers high availability (HA) for NFS (network filesystem) and can work on top of shared disk or network block I/O.

Lustre

Lustre (named from Linux and Cluster) is cluster filesystem on top of modified ext3. It uses architecture with object storage servers (typically 2--8) and clients acessing to data. It has direct support of HA, recovery and transparent reboots of storage servers. Data blocks can be stripped across objects so there no size limit of storage object.

References

- Val Henson. *Choosing and Tuning Linux File Systems*, `http://www.valhenson.org/review/choosing.pdf`

- Rémy Card, Theodore Ts'o, Stephen Tweedie. *Design and Implementation of the Second Extended Filesystem*, `http://e2fsprogs.sourceforge.net/ext2intro.html`
- David Teigland. *Symmetric Cluster Architecture and Component Technical Specifications*, `http://people.redhat.com/~teigland/sca.pdf`

Other resources can be found on Wikipedia pages with keywords: Reiserfs, XFS, IBM General Parallel File System, Lustre (file system), Shared disk file system.