Aurora - The next OS/2 Warp Server:
File System Services Technical White Paper

Introduction

Remember the wonders of the 10-megabyte (MB) hard drive on the first IBM PC XT? And how long it took to outgrow?

As files and partitions grew larger, new file systems like IBM's High Performance File System (HPFS) provided faster disk access and improved overall performance. Now that high-capacity drives--with storage capacity in the gigabytes--fit in your hand, audio and video files consume many megabytes of space, and PC databases approach terabytes of data, new demands on the files system are driving changes. Research shows that 625 new terabytes of data are added to the Internet each month. Technologies are evolving to meet the challenge with rapid expansion of hardware capacity. These large amounts of data and the growing networks interconnecting computers worldwide impact today's file system dynamics.

IBM is bringing a Journaled File System (JFS) to the next version of OS/2 Warp Server, code named Aurora. Designed for faster performance, enhanced Web and Lotus Notes serving capability, and improved scalability, Aurora with JFS positions customers to better meet their business objectives as they transition to a network computing environment. They can now implement JFS and use the latest high-capacity DASD that their hardware supports. Designed for the high throughput and reliability requirements of corporate servers, JFS has incredibly quick recovery times that are an essential factor in improving server availability. Logical Volume Manager (LVM), also new to Aurora, helps administrators balance dynamic file requirements. Now a partition can be expanded dynamically and a volume can span physical disks. While maintaining compatibility with existing applications that use FAT, HPFS and HPFS386, JFS and LVM combine to provide:

- A significant reduction in file restore time
- Increased file and partition size limitations
- "Sticky" and dynamic drive letter assignments for enhanced support of removable media
- Enhanced performance scalability on Symmetrical Multiprocessor (SMP) systems

Boost Server Availability

An e-business environment can't endure the lengthy file system recovery times required after a system crash with a non-journaled file system. Using proven database journaling techniques taken from IBM's AIX system, the OS/2 JFS can restore a file system to a consistent state in a matter of seconds or minutes, compared to non-journaling file systems like FAT and HPFS, which can take hours or days depending on the partition size.

FAT, HPFS, and other file systems are subject to corruption in the event of a system failure, since a logical file operation, for example a "create," often takes multiple I/Os to accomplish and may not be reflected in the data at any given point in time. These file systems rely on time-consuming and tedious processes like CHKDSK that examine all of a file system's directories and disk-addressing structures to restore a file system to a reliable state. In contrast, JFS creates a continuous log of transactions performed on the file system. In the event of a system failure, the file system is restored quickly and accurately by replaying the log and rewriting records for the appropriate transactions. JFS is not a bootable partition. FAT or HPFS still must be used for the bootable file system. We recommend that HPFS be used as the boot partition to support the Java Virtual Machine's long file names.

Break File System Limitations

As businesses expand and new opportunities emerge, hard drives and their supporting files begin bursting with data and quickly exceed the defined boundaries. Remember when the industry said that no one would ever need more that 10 megabytes of hard disk capacity? New technology continuously grows the hard file and removable media capacity to new heights. All these changes render the old file and partition sizes too limiting. The new file and partition sizes supported by JFS significantly strengthen OS/2's server capability. JFS raises the previous file size limitation of 2 gigabytes to 2 terabytes. Partition size is raised from 64 gigabytes to 2 terabytes, so a file can now be as big as a partition. The maximum number of files and directories within a single JFS file system is over 4 billion.

JFS offers an alternative to pre-allocating all hard file space in blocks. Sparse files only occupy the amount of disk space that is consumed by the data. The actual allocation of any given block in the file doesn't occur until a write operation is performed to that block. In practice you could define a 1 petabyte file inside a 1 megabyte partition, as long as you don't write more data than can fit on the physical partition.

Maximize Hard File Resources

As companies, their departments, and supporting databases become larger, LVM will make it easier to grow the server with new physical disk drives (DASD). The growing volume of files and databases require expandable system DASD. Logical drives can now span multiple physical hard drives. With permanent or sticky drive letter assignments, hard drives can be moved or added without changing drive letter designation, thus keeping program path information consistent. Partitions can dynamically grow without reformatting or rebooting the system. IBM enabled both GUI and command line APIs to access LVM, allowing maximum flexibility in implementation methods. New with LVM is LVMDISK, which replaces FDISK. LVM makes it easier for I/T to manage dynamic DASD requirements.

JFS supports the expansion of a mounted and actively accessed file system. Should a file run out of space, the system administrator can increase the volume size without disrupting ongoing transactions. This eliminates the need to stop, back up, reformat the partition, and then restore the data. With JFS and LVM, a partition can be extended without reformatting the file system that it contains. Additionally, it is possible to write a routine that would automatically increase a volume as long as the space is available.

- For example, IBM's DB2 on OS/2 returns an error message that indicates an "out-of-space" condition for a given partition. A routine could be written to intercept the error message and kick off another routine calling on LVM's APIs to dynamically increase the partition by a predetermined size -- an efficient way to automatically handle the error condition.

Based on the application environment, DASD space utilization can be optimized with block sizes of 512, 1024, 2048, and 4096 bytes. While smaller block sizes allow more efficient use of disk space, they can increase access time. The default block size is set at 4096 bytes since performance, rather than space utilization, is generally the primary consideration for server systems.

Free space may be defragmented on DASD while it is active. Once the free space becomes fragmented, defraging the file system allows JFS to provide more I/O-efficient disk allocations and avoid some out-of-space conditions.

Consistent with HPFS, JFS supports upper and lower case for names. Case is preserved in the storage and retrieval of file names, but is ignored during directory searches. All file and directory names are stored and managed as strings of unicode characters. Like HPFS, JFS limits file or directory names to 255 unicode characters.

Network File System

Enabling business data to be shared horizontally across the enterprise generally increases its value. Aurora delivers a cost-effective way to manage information interdepartmentally or company-wide. With Network File System (NFS), a feature now included with Aurora, a RISC (AIX/UNIX) DASD can be mounted and made a sharable resource to the server's users. Sharing data -- files and databases -- reduces the number of pieces of software that must be maintained, which simplifies system administration and helps contain costs.

Summary

Aurora provides a proven mission-critical foundation for the transformation to the Network Computing and e-business environments, while supporting legacy applications and integrating OS/2 more tightly into the IBM family of business servers. Aurora's Journaled File System provides a state-of-the-art, quickly recoverable, transaction-oriented, log-based, scaleable file system. The increased performance from scalability and the enhanced reliability that JFS provides make Aurora better suited to Web and Lotus Notes serving in support of network computing and e-business. With LVM and its ability to span multiple physical drives for a single partition, sticky drive letter assignments, and increasing partitions dynamically, you have a strong team. Bring in the added value of NFS and its network flexibility and you have a winning combination of hard disk and file management that's unbeatable in the Intel PC server market.

Notice

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

References in this publication to IBM products, programs, or services do not imply that IBM intends to make them available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's products, programs, or services may be used.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give any license to those patents. License inquiries can be sent, in writing, to the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594, U.S.A.

Trademarks

IBM, AIX, and OS/2 are Registered trademarks of International Business Machines Corporation in the United States and/or other countries.

Windows and Windows NT are registered trademarks of Microsoft Corporation.
UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

Other company, product, and service names may be trademarks or service marks of others.