White Paper

Secure Distributed and Parallel File Systems Based on Network-Attached Autonomous Disk Drives

Professor Garth Gibson
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
garth@cs.cmu.edu
(412)-268-5890 (phone)
(412)-268-5576 (fax)

Abstract

The network-attached, autonomous disk drive is a revolutionary advance in disk drive and file system technologies that removes the server workstation bottleneck from networked file systems, embeds the file system into the disk drive and promotes the disk to a first class network device. By integrating security into the drive, network-attached disk drives provide file system integrity that is not dependant on the trustworthiness of the client. Minimizing the latency between disk drive and network is achieved by utilizing a low-power, low-cost, high-bandwidth embedded architecture specifically designed for data movement. Direct attachment to a network means that the disk drive can serve as a fundamental building block for highly-reliable, highly-available and fault-tolerant file systems providing low-cost, high-performance storage to meet the rapidly increasing needs of digital data storage applications.

Goal of this white paper

Our goal in distibuting this white paper is to identify industrial partners for this research agenda. We believe that raising the level of the storage system interface to that of a simple file system, treating the device as a first class network entity and integrating a security protocol at the drive level can open substantial opportunities for drive value-added features. Toward the exploitation of these opportunities we seek the experience and expertise of the storage industry. We would like to build close relationships with particular partners, leveraging the relationships that we currently enjoy through CMU's Data Storage Systems Center and CMU's Parallel Data Laboratory. Further, because widespread impact of the research we are embarking on will require industry consensus, we seek to establish a colaboration with a research group in the National Storage Industry Corporation.

This document was created with FrameMaker 4.0.4

Next Generation Disk Drive Interfaces

Our vision of scalable distributed and parallel storage architectures is guided by these themes:

- to exploit new opportunities enabled by technology trends,
- to bring applications closer to their data and give them greater control over its performance, availability, and consistency,
- and to design security into the most basic levels of the system.

We propose to develop a scalable, secure, high performance file system interface appropriate for commodity disks endowed with first class network ports. We will demonstrate this interface for stand-alone personal computers, traditional distributed file systems, and highly scalable parallel file systems for parallel computers. We will further develop the drive computing architecture capable of meeting the demands of this new role but still sufficiently cost-effective and power-meager to meet the competitive demands of the high-density, commodity drive marketplace. The employment of these drives as a basic system building block in scalable multicomputing environments reconceptualizes network architecture to eliminate file server workstations, endowing the network with data transfer characteristic of a system backplane and a trust model characteristic of an insecure heterogeneous environment.

Enable new classes of performance optimizations by raising storage's system interface

The US digital data storage industry leads the world's storage markets. It has survived the move from being able to rely on captive customers to having to compete in the price sensitive personal computing marketplace. It has responded to the pressure from DRAM density growth and increased its areal growth rates from 25% per year to 60% per year. It has developed perhaps the cheapest, most widely supported and highest bandwidth local area network, the SCSI bus, whose technology hiding aspects have enabled rapid introduction of new data rates and geometries. Moreover, SCSI has induced drive vendors to endow each drive with a device-optimized operating system implementing geometry-dependent scheduling and caching; this has been a particularly effective as added-value that distinguishes one product from another because of the slow evolution of client operating systems.

But the physical layers of SCSI have been pushed to their limit; the next major evolution in disk architecture is the adoption of high-speed serial interconnects such as Fibrechannel, SSA, and Firewire. High-speed is needed because media data rates are approaching 10 MB/s and may grow at up to 20% per year. Serial is needed to lower cost, extend server-disk physical separation, and increase the number of ports on one server adapter. For example, Fibrechannel-attached drives will be available in 1995 and will burst transfer at 1 GB/s over a 2 km cable distance with up to 126 disks and servers attached.

While these serial interconnects will free disk technology from the physical limitations of SCSI, they will not change the logical interfaces: SCSI's abstraction of a drive as a single linear collection of fixed size sectors. In order to continue to add value through drive-embedded optimizations such as aggressive prefetching, dynamic allocation, or on-the-fly compression, we propose to

develop a higher level, file-oriented, performance-enabling interface based on our experience with scalable, aggressively prefetching and caching parallel file systems (SPFS).

Moreover, to exploit the serial interface's high bandwidth for off-loading data transfer from file servers and clients, file systems need to increase their use of drives in peer to peer transfers for copy, drive supported RAID, and continuous time media delivery. To this end, peripheral interconnects must evolve to scalable switched networks such as ATM and drives must behave as first class network nodes. Because of drive marketplace's requirements for increasing packaging density and the sensitivity of drive mechanics to heat, the architecture of a network-attached drive processing system must be very power conscious; today's drives dissipate less power in total than today's fast microprocessors. We propose to develop hardware and software architectures appropriate for high-bandwidth, low-power, network-attached file-managing disk drives.

Rearchitecting high performance client server file systems

Network file systems have always depended heavily on dedicated file server machines. For systems such as AFS that are optimized for very large numbers of clients, file server costs are minimized by extensive use of caching on client disks. Though this scales well, client performance is limited to less than its local storage bandwidth so high performance client computing requires expensive duplication of high-performance storage at all clients. Alternatively, more recent systems such as Sprite Zebra and IBM Vesta, exploit high bandwidth networks to deliver high performance client computing at the expense of a much larger ratio of high performance servers per client machine.

Not only is the scaling of such systems put at risk by the cost of file servers, the technology trends in commodity workstations renders them increasingly inefficient as file servers. Fundamentally, workstations increase performance per dollar by upgrading peripheral technology more slowly than microprocessor technology. For example, Digital's cost-effective UNIX workstation line has seen microprocessor performance grow by a factor well over 10 since the introduction of the DECstation 5000 while its system bus bandwidth has only recently been increased by as little as 30%. This is the worst possible trend for high-bandwidth file servers and it has caused Wisconsin and Berkeley groups to express an interest in eliminating dedicated file servers by using clients as file servers.

Our approach is more radical still; with network-attached drives, dedicated workstation file servers can be eliminated altogether and the requisite file management functionality repartitioned between clients and drives. This will eliminate throughput-limiting and latency-inducing copies into and out of a workstation whose primary function is as a traffic manager; in this way, drive bandwidth is brought close to client machines. Coupled with scalable switched networks, path-oriented operating systems, application access to network devices, and network adapter support for checksums and scatter/gather DMA, applications in a large scale multicomputing environment will be "closer" than ever to their data.

Strengthen file system security at the most basic and most protectable level

In addition to the problem of performance impact of one client on another, the approach of dis-

tributing storage over client machines decreases the security of a system's most protection needy resource, its files. In the distributed network of workstations model, this problem is at its worst. Every client, physically accessible to a wide range of individuals and providing network user login shells, exposes the global file system to security threats.

In our approach, dedicated workstations acting as a file servers are eliminated, but disks are kept in the physically secure server room, accessed directly by authenticated clients, and employ the environment's authentication service and onboard tamper-resistant encryption to validate authorization, protect media security and defeat network eavesdropping. While enabling systemic security, our approach rides on a small amount of mechanism. Like today's shrinkwrap software, a drive comes with a unique key (serial number or password) and a little bit of client software that uses this key to encrypt a token that the drive can decrypt. On this rests the drive's notion of authentication; any machine authenticated by this key can establish other temporary keys which the drive uses to identify and distinguish authenticated clients. A drive's master key is protected on the drive because it is stored in a tamper-resistant encryption device, a PCMCIA card or drive-mounted chip. This device also provides simple but effective encryption that is used for key manipulation, protecting network traffic from eavesdropping, and on-media exposure to theft.

An important aspect of our approach is that the file system integrity is not dependent on the trustworthiness of client operating system or client applications. Any entity participating in the authentication protocol can obtain access to authorized data and only to authorized data.

Repartitioning network file systems for flexibility, scalability, availability and performance

Network file systems are usually partitioned into a low-level that manages storage, memory caching, and intra-machine consistency, and a high-level that manages the users' namespace, authorization and authentication, inter-machine consistency, and network transport. While namespace management and cache consistency control form the "personality" of file system, authorization and authentication, network transfer, storage management, and prefetching and caching form the "kernel" of a secure, high-performance file system. This organization is emphasized in parallel file systems for I/O intensive parallel programs such as out-of-core scientific simulation or data mining systems. To maximize these applications' performance, low-level data access must employ aggressive asynchronous strategies for disk access and network transfer. Further, for applications that employ sharing synchronized by application specific communication, high-level consistency management may rely on application assistance to avoid inefficient, conservative caching policies.

We propose to develop a highly asynchronous, transfer optimized, secure, low-level file system interface for network-attached, autonomous disk drives. We will also develop flexible library routines (middleware) employed by untrusted client operating systems, client file system personalities, or high-performance applications to bind autonomous disk file systems into a scalable, network-wide parallel file system. These client-side routines will also provide caller-configurable redundancy for file system fault-tolerance and availability, and caller controls for cache consistency and resource management hints. We will demonstrate this partitioning for I/O intensive parallel applications, and traditional distributed file systems such as NFS or AFS.

Comparison with other ongoing research

Scalable file systems have been and remain an active area of research, but none conceptualize the disk drive as a computational entity with security and performance characteristics preferable to file servers. Although considerable consensus has been reached in the design of large-scale, secure, distributed file systems such as AFS, systems emphasizing higher client performance such as Zebra or xFS focus on moving the file system and even the storage to the client. Our work differs from these in its treatment of clients as distrusted. Most of the current efforts in scalable file systems, notably those underway in the context of the Scalable I/O Initiative such as Vesta and Passion, emphasize portability, programmer convenience, and scientific parallel programs. Our proposal derives from our contribution to parallel file systems for parallel computing on multicomputers, the Scotch Parallel File System (SPFS), which emphasizes scalability and performance though client-managed consistency, reliability and availability and from our intra-node resource management and latency reducing strategies, informed prefetching and caching, that exploit application disclosure and highly parallel storage arrays. Our emphasis on reliability and availability is in the spirit of RAID structures applied to storage servers as is done in TickerTaip, Swift RAID, or DEC SRC's parallel block server, but by restricting redundancy encodings to span only data having the same authorization profile we can endow distrusted clients with library routines enabling them to manage their own redundancy and availability needs.

Our emphasis on secure file systems is based on the results of prior work on tamper-resistant, security coprocessors. Our emphasis is not on new security models or protocols as a basic research thrust, but rather on endowing high-performance, scalable computational systems with a stronger, more flexible security model resting on the disk as an indivisible, tamper-resistant storage component. The application of ubiquitous security to network-attached storage architecture is essential for the promotion of drives to first class network citizens and the industry is ripe for adopting a well demonstrated strategy.

In the drive technology domain, we are the principle systems integration component of the largest university research program in storage technology, CMU's Data Storage Systems Center, and we are working closely with the National Storage Industry Consortium and high-performance drive manufacturers such as IBM, HP and Seagate. These collaborative efforts include RAID architectures for online recovery and minimizing redundancy maintenance overhead, rapid prototyping for the development of new RAID architectures, drive support for RAID, system integration of serial interfaces for drives, understanding and exploiting processing power in the drive, and evolving SCSI for higher performance drives. The primary alternative drive interface work in the storage industry is the Scalable Storage Interface, a standards effort focussed on a pinout and control/status register abstraction of storage devices attaching to PCI and SCI buses. Where we envision drives as autonomous units networked away from clients to exploit machine room physical security, they are focussing on chip-mounted drives and industry standard adapter interfaces.

An exciting synergy exists between our bandwidth-optimized drive-embedded secure file system vision and the Arizona communication-oriented operating system design, Scout. We share the vision of a high-performance, customized software architecture moving storage data across high-bandwidth networks. Our penetration of the storage industry may facilitate the rapid evaluation and adoption of their technology.

Summary of Proposal

- **Network-attached autonomous disk drive:** Carnegie Mellon University proposes to research, design and build a network-attached autonomous disk drive that provides:

    - first-class network status for disk drives
    - disk drive embedded file systems
    - secure data access
    - high-performance data transfer from disk to client
    - high degree of scalability, availability, reliability and fault-tolerance

    The network-attached, autonomous disk drive is a revolutionary advance in disk drive and file system technologies that removes the server workstation bottleneck from networked file systems, embeds the file system into the disk drive and promotes the disk to a first class network device. By integrating security into the drive, network-attached disk drives provide file system integrity that is not dependant on the trustworthiness of the client. Minimizing the latency between disk drive and network is achieved by utilizing a low-power, low-cost, high-bandwidth embedded architecture specifically designed for data movement. Direct attachment to a network means that the disk drive can serve as a fundamental building block for highly-reliable, highly-available and fault-tolerant file systems providing low-cost, high-performance storage to meet the rapidly increasing needs of digital data storage applications.

- **Security integrated into the drive:** The drive will provide a high-degree of data security and integrity by integrating authentication and encryption into the system. Implemented with minimal infrastructure, a range of security levels can be provided by the richness of the trusted authentication server, guaranteeing security in network environments with untrusted clients.

- **High-level disk drive interface and functionality:** Currently, disk technology cannot fully utilize aggressive prefetching, dynamic allocation and on-the-fly compression because it does not understand file structure. Promoting the drive interface to the level of a file system will allow the disk drive to employ performance optimizations within the device, significantly improving disk drive performance.

- **In-drive computing architecture:** To meet the demands of high-speed data movement, the disk drive will employ a computing architecture specifically designed to support the file system and network software, focusing on the minimization of data movement while maintaining software modularity and integrity.

- **Scalable, parallel file system:** A set of library software modules will provide portable and flexible protocols that bind collections of disk drive file systems into parallel file systems. The modules will enable fault tolerant and high availability networked file systems, runable alternatively by client kernels, user-level servers or user applications.