Appears in *Proc. of the 30th Hawaii International Conf. on Systems Sciences*, Jan. 1997

Task Force on Network Storage Architecture: Abstracting the storage interface

Garth A. Gibson
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3891
garth.gibson@cs.cmu.edu http://www.cs.cmu.edu/Web/Groups/NASD/

Position Statement

The US digital data storage industry leads the world's storage markets. It has survived the move from being able to rely on the captive markets of single-vendor computing systems to having to compete in the price-sensitive open market of personal computers. 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 current major evolution in disk architecture is the adoption of high-speed serial interconnects such as Fibrechannel, SSA, and Firewire. High-speed is needed because current media data rates exceed 10 MB/s and may grow at up to 40% per year. Serial is needed to lower cost, extend host-disk physical separation, and increase the number of devices connected to one host adapter.

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, I contend that we must develop a higher level, file-oriented, performance-enabling interface for next-generation storage. Promoting the drive interface to the level where drive-understood storage objects correspond closely to the files managed by file system will allow the disk drive to employ more effective optimizations and self-management within the device, significantly improving disk drive performance, easing administration, decoupling the evolution of file system and storage system and lowering the cost of ownership.

Moreover, to exploit the high bandwidths of future disk drives and serial interfaces in local area network file systems, disk data should be transferred directly between devices and clients rather than stored and forwarded through file server machines. This "third party transfer" emphasis on peer-to-peer transfers will reduce latency and increase bandwidth for all operations, and it will enable more effective optimizations such as device-to-device copy, drive supported RAID, continuous time media delivery and network-striped transfers. To this end, device (peripheral) interconnection buses must evolve to scalable switched networks such as gigabit Ethernet or ATM and drives must behave as first class network nodes. Coupled with scalable switched networks, communications-oriented operating systems, application access to network adapters, and network adapter support for checksums and early demultiplexing (scatter/gather) DMA, applications in a large scale multicomputing environment will be "closer" than ever to their data.

Attaching storage devices directly to local area networks that are linked to wider networks such as the internet exposes storage to commands from incompatible, malfunctioning, hostile and malicious machines. Storage security, currently left up to server machines, will need to be insured at the network-attached device without sacrificing the latency or bandwidth advantages of direct communication with clients. I contend that this can be done efficiently with only a few long-term keys in the drive while providing insured authentication of commands, privacy for data transfers, and independence from physical security constraints.

This work was supported primarily by Defense Advanced Research Projects Agency contract N00174-96-0002 and in part by donations from Data General, Symbios Logic, Hewlett-Packard and Compaq. The United States government has certain rights in this material. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any of the funding agencies.