Task Force on Network Storage Architecture: Internet-attached storage devices

Rodney Van Meter, Steve Hotz and Gregory G. Finn
University of Southern California/Information Sciences Institute, Marina Del Rey, CA
rdv@isi.edu

Position Statement

Networks such as HiPPI, SSA and Fibre Channel are becoming the access technology of choice for peripherals such as disk drives, tape drives and disk arrays. These networks scale better than traditional I/O channels, connecting more devices over greater distances and providing greater aggregate bandwidth. More complex protocols are required for network interfaces than for channels. In most cases, specially-developed protocols are used, rather than existing standards such as TCP/IP, due to perceived differences in functionality, focus, complexity and especially performance. We reason that most of these concerns either reflect misunderstanding of the IP suite or are being met as the suite evolves. We further argue that the benefits of using IP, including wide-area connectivity, cross-media bridging and reduced R&D, are substantial. Therefore, we feel that IP is an appropriate choice for a storage device and should be the protocol of choice for systems implementers.

The wide area connectivity that is IP's strength opens up new functionality for peripherals, for example, remote mirroring of disk drives and remote backup. Cross-media bridging can be useful in heterogeneous computing environments, allowing transparent interoperation different types of networks. Using IP makes use of the large existing body of research and development in routing, congestion control, flow control and reliability. This reduces R&D effort, as well as allowing quick integration of emerging features such as resource reservation and real-time protocols. It also alleviates the problem of committing to a protocol suite which is more or less tied to a choice of physical media, such as Fibre Channel or ATM, providing a growth path unconstrained by the future development of a particular technology.

Many of TCP/IP's known performance problems come from outdated implementations or restrictions imposed from outside, such as Berkeley's 128-byte mbuf memory management and ethernet's small packet size limit. Changing to a larger units (buffers or packets) can dramatically reduce the CPU load at high data rates. On media that support large packet sizes, the first step in reducing CPU utilization is to increase the packet size, which IP supports. At the transport layer, a significant source of latency and CPU consumption is TCP checksum calculation. Solutions such as transport-level trailers and zero-pass checksumming can reduce this performance penalty. At the peripherals, where virtual memory management and user buffer alignment and filling are not concerns, implementation can be simple and efficient. A modern scatter-gather data transfer engine could move data from buffers to disk or network without copying or reorganizing the data.

TCP poses two possible problems for I/O transactions -- lack of application framing and excessive generality. We believe a variant of TCP, similar to Transaction TCP, could address these concerns. If TCP ultimately proves untenable, I/O operations could be supported with UDP and application-supplied reliability, or as last resort, a new transport protocol could be developed within the IP framework.

Our conclusion is that as we move to larger, more complex switched networks for I/O, some sacrifice of performance is the inevitable result. However, IP has no inherent performance penalty relative to other protocol choices; the loss is entirely attributable to managing the additional complexity. In this environment, IP offers significant advantages and few drawbacks. IP, therefore, should be network protocol of choice for developers of network-attached peripherals.

This research was sponsored by the Advanced Research Projects Agency under Contract No. DABT63-93-C-0062. Views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of ARPA, the U.S. Government, or any person or agency connected with them.