

Small Tree's InfiniBand technology

BY CORKY SEEBER

In the broadcast industry, content delivery bandwidth is critical, and today's network wiring typically offers no more than a gigabyte of throughput. As the industry moves toward HD, a network-wiring alternative is needed to handle increased data requirements. The emerging solution is InfiniBand — a one-wire, high-performance interconnect supported by Small Tree Communications.

Moving data quickly

Traditional networking protocols are directed through a TCP/IP stack. TCP was designed more than 20 years ago, at a time when files were smaller and networks were slower and less reliable. Packet sizes were limited to 1500B as a tradeoff between using large packets for efficiency, while keeping them small enough so a small amount of corruption would not require too much data to be retransmitted. Each of these tiny packets is encapsulated with a medium access control (MAC) header, an IP header and a TCP (or perhaps UDP) header, which all have to be stripped off at the final destination. This leads to quite a bit of overhead and protocol traffic, creating additional latency and congestion and making a traditional TCP/IP network less than ideal for editing HD footage in real time.

The idea behind the InfiniBand technology was to create a scalable, extensible fabric that could be used to interconnect systems, I/O devices and storage. The latency to send a message across the fabric needed to be extremely low in order for things like MPI clustering and storage to be effective.

The technology's low-level protocols allow upper level protocols to be set on top of them in such a way as to add

minimum overhead. Sockets Direct Protocol (Expected Q106) is one such example. It allows sockets-based applications to send data directly through the InfiniBand network without using TCP. As a result, there is much less protocol overhead, and updating frames can happen more quickly. Latencies on the TCP stack on Mac have been measured at 60 μ s, even with 10Gb, whereas when using InfiniBand, latencies have been measured at 6 μ s.

A substantial problem with current cluster file systems is latency to and from the metadata server.

So just what is this one-wire technology? It's a switch-based serial I/O interconnect architecture that operates at a base speed of 20Gb/s per port, or 10Gb/s in each direction. It differs from shared bus architectures in that it is a low pin-count serial architecture that connects devices on the PCB and enables bandwidth out of the box. It provides both the high bandwidth one would expect from a 10Gb/s interconnect as well as the low latency advantages one might see when communicating with a device over a local bus. The architecture was designed to simplify and speed server-to-server connections and links to additional server-related systems.

The primary use of this technology is focused in cluster systems — two or more systems working as one. It was originally created to serve as a bus replacement, acting as a switched fabric network in place of 10Gb or other Ethernet networks. Today, more and more broadcast companies are looking carefully at cluster file system technologies. These technologies are viewed as a way of storing all data in one place and helping reduce copies.

A problem with current cluster file systems is latency to and from the metadata server. Somewhere in the organization, there is a server tracking who is working with a file. It also prevents two people from working from the same file at the same time. Unfortunately, almost every one of these metadata servers cannot provide greater than 40 μ s to 60 μ s latency from the time the message is sent out and the other side receives it. This

technology offers the promise of reducing latency by a full order of magnitude to improve the ability of clustered file systems to operate across a group of machines at greater speeds.

Shifting to one wire

As a one-wire network interconnect, the InfiniBand technology also helps reduce infrastructure requirements. Current workstation systems easily have four cables — one or two each for Ethernet and Fibre Channel — plus a disc drive connected to one system.

With Small Tree's technology, each individual workstation in a workgroup of 20 users, for example, would run their workstations via one wire to a relatively small switch that is bridged back to a larger switch in the data center. This gives each user up to 10Gb/s of access to all the data stored on the server. Using bridged I/O, none of these systems need to have Ethernet, and you can get rid of the discs within those systems. **BE**

Corky Seeber is president of Small Tree Communications.