Using Layer 2 Ethernet for High-Throughput, Real-Time Applications

Robert Blau, Mercury Computer Systems, Inc., bblau@mc.com

Overview

Network-centric HPEC solutions strive for convergence around Ethernet in order to simplify system complexity. However, mainstream Ethernet connectivity often does not work for all of the interconnect functionality needed in these applications. This happens because Ethernet is usually associated with TCP/IP [1,2], a WAN/LAN protocol which requires an inappropriate amount of processing and latency overhead for many real-time functions. Lighter-weight Ethernet protocols such as UDP/IP [3] or TIPC [4] improve matters, but still require an unnecessary level of IP overhead (in addition to other limitations.) Specialized interconnect technologies such as RapidIO and Serial Front Panel Data Port (sFPDP) are often needed to satisfy the strict, real-time performance and latency constraints of the Data Plane and Sensor Plane components of applications.

There are two ways to work toward convergence of the real-time HPEC interconnects. It is possible to approach convergence by tunneling Ethernet over RapidIO (EoRIO [5]). This has the benefit of adding the advantages of network convergence to existing real-time system architectures. The other approach is to tunnel the real-time serial interconnect technologies over Layer 2 (L2) Ethernet, while still maintaining their latency and performance characteristics. This has the benefit of riding the wave of future Ethernet advances such as CEE (Convergence Enhanced Ethernet) [6] and data rates beyond 10Gbaud.

In this presentation, we describe a framework for tunneling lightweight interconnect protocols such as sFPDP, RapidIO, PCI Express (PCIe), and VITA 49 (Digital IF) over Layer 2 Ethernet. This provides an efficient, hardware-implemented, lightweight, lossless, peer-to-peer protocol by using the Layer 2 Ethernet protocol (IEEE 802.1Q [7]) directly. It is also able to simultaneously co-operate with other Ethernet protocols (for example TCP/IP and FCoE [8]). It operates with any speed of Ethernet, ranging from 1Gbit (1GbE), 10Gbit (10GbE), and in the future, 40Gbit Ethernet (40GbE) and 100Gbit Ethernet (100GbE). The design and use of the custom network interface will be presented, covering throughput and latency results. Reliability features will also be described.

Protocols

Serial protocols such as 10GbE, Serial RapidIO (SRIO), PCIe, and sFPDP, all use the same Layer 1 (physical media access) hardware. This enables them to use commodity SERDES and transceivers. However, the different upper layers of the protocols still necessitate protocol-specific switches. It has now become highly desirable to also use L2 Ethernet switches to take advantage of their economies of scale. Ethernet switches have much higher volumes, resulting in faster innovation and lower costs. The combination of high performance and low cost, has pushed the use of high-performance Ethernet down from the WAN and LAN domains onto the backplane.

The tunneling protocol consists of L2 request frames sent by a master to a target, and L2 reply frames returned to the master. The format of these frames is simple, flexible enough to convey sFPDP, VITA 49 (Digital IF), PCIe and SRIO packets, control symbols, and configuration frames. The common format is designed to provide functionality in a compatible and efficient manner [9]. A hardware-implemented Layer 3/4 protocol provides real-time reliability and flow-control features.

Characteristics

Systems can use the tunneling protocol to access an arbitrarily large array of endpoints through the use of commodity L2 Ethernet switches. The tunneling transactions can be between FPGAs on a board, or between blades plugged into a rack-mounted chassis, or between endpoints in separate chassis.

The systems are designed to efficiently run real-time applications using either low-latency DMA, MPI, or socket-based communication mechanisms. Mercury has paid particular attention to throughput, latency, and reliability in the design. The lightweight format and packet aggregation features achieve low overhead. The need for only minimal software overhead and the packet cut-through design, provide low-latency benefits. Robust error detection and automatic retry provide real-time reliability. Credit-based flow control and priority levels provide well-behaved behavior in the presence of contention.

Performance

Figure 1 presents some performance models showing the data (PDU) size versus the peak transmit throughput (GBytes per second) for tunneling over various next-generation Ethernet technologies. Complete throughput and latency performance results will be reported in the presentation.

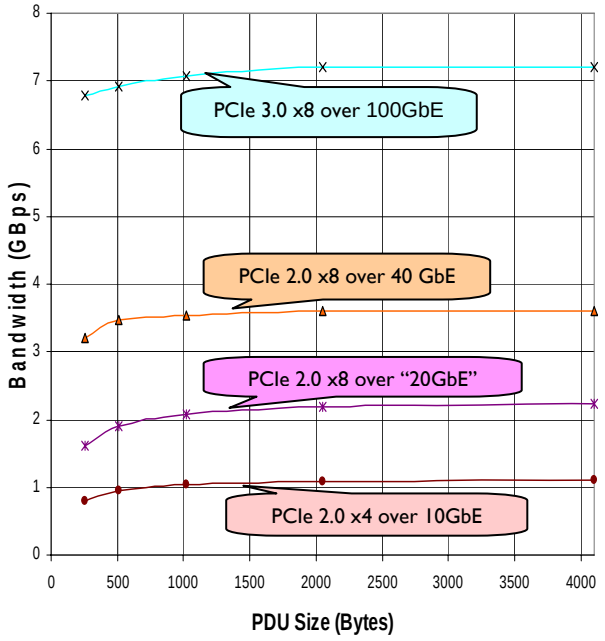

Figure 1: Performance models.

References

[1] Postel, J., "Transmission Control Protocol," STD 7, IETF RFC 793, September 1981.

[2] Deering, S. and Hinden, R., "Internet Protocol, Version 6 (IPv6) Specification," RFC 2460, December 1998.

[3] Postel, J., "User Datagram Protocol," STD 6, IETF RFC 768, August 1980.

[4] Maloy, J., "Telecom Inter Process Communication," http://tipc.sourceforge.net, January 2003.

[5] Dunn, I., Desrochers, M., and Cooper, R., "Network Attached Signal Processing," to be published May 2008.

[6] Merritt, R., "Ethernet upgrades to be sole data center net," EETimes,http://www.eetimes.com/showArticle.jhtml;?articleID=198700949, March 2007.

[7] IEEE Std. 802.1Q-2005, Virtual Bridged Local Area Networks.

[8] INCITS Project 1871-D, Fibre Channel – Backbone - 5 (FC-BB-5).

[9] IEEE Std. 802a-2003, IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture Amendment 1: Ethertypes for Prototype and Vendor-Specific Protocol Development.