The Unified Wire Engine

Introducing Terminator 3

A Chelsio Communications White Paper

Abstract

Today's enterprise networking landscape is highly fragmented. A typical enterprise operates multiple independent infrastructures for storage (usually a Fibre Channel storage area network), computing (usually InfiniBand or Myrinet high speed interconnects), and TCP/IP networking (over Ethernet). Consolidating these infrastructures into one unified network has many rewards, both monetary and in ease of administration and management. To realize this vision of a **unified wire**, Chelsio Communications introduces its latest protocol engine, the Terminator 3 (T3) ASIC. T3 is architected to be a unified wire enabler, capable of simultaneously supporting the different applications, while individually matching the performance and capabilities of each specialized network. In addition, T3 introduces unique features to enhance Ethernet's capabilities, and to support a large number of different uses in line card and intermediate box configurations.

Introduction

Many IT departments have to deploy multiple independent physical infrastructures, in order to support the various applications needed by the enterprise. Thus, a typical enterprise network would consist of a separate storage area network (SAN), usually using Fibre Channel, a separate interconnect network for computing (using a specialized fabric such as Myrinet or InfiniBand), in addition to the ubiquitous Ethernet network for internal TCP/IP networking and Internet connectivity.

Besides the inevitable historical legacy aspects, this separation was essentially dictated by two factors:

First, the different applications have different characteristics and requirements, which need to be properly supported by the underlying networking technologies. For instance, storage requires a high bandwidth, reliable infrastructure, while computing requires a very low latency interconnect.

Second, the interfaces to the different networks differed. Storage may use SCSI for block access over a SAN, while computing applications use remote direct memory access (RDMA) to communicate over a computing cluster.

Since none of the available communication technologies was capable of adequately supporting all the applications, the separation remained even as the technologies evolved. However, there are clear benefits to converging all applications onto one, unified network. This convergence has direct payback in terms of saved hardware infrastructure, as well as indirect –and sometimes more significant– savings through simplified management and network administrator training. An ideal candidate for this convergence would be Ethernet, a simple, pervasive, well understood and liked technology with plenty of knowledgeable administrators available. However, even at 1Gb speeds, Ethernet lacked the capabilities to address the main problems above.

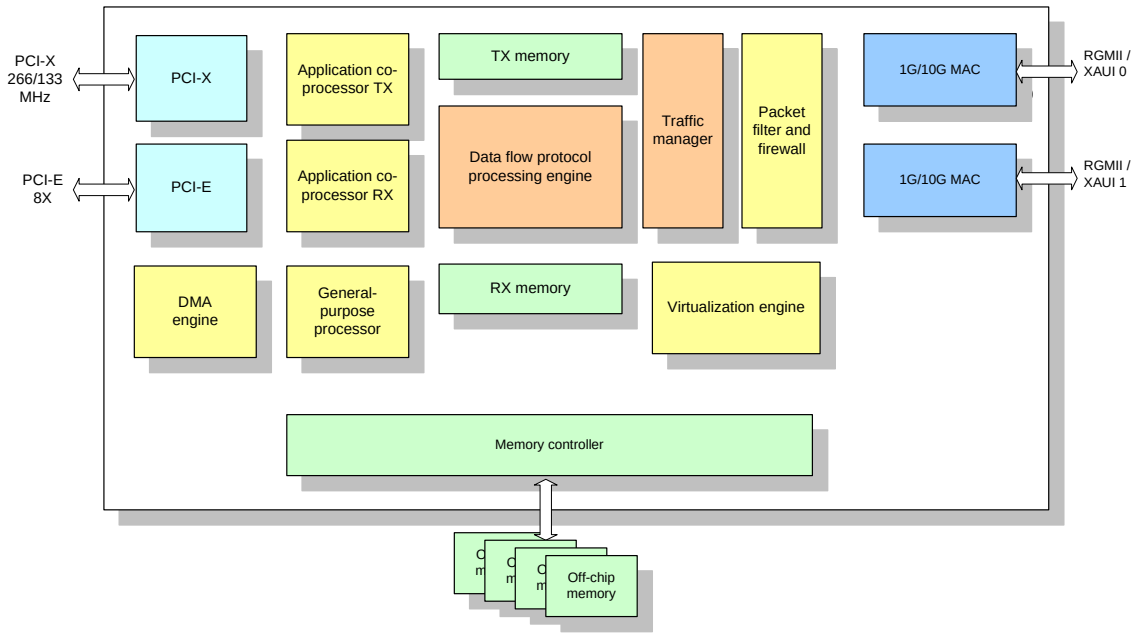
Today, the picture has finally changed. The newest generation of Ethernet, running at 10Gb, has become readily available at price points which have traditionally triggered a major ramp-up in the adoption of new Ethernet speeds. Unlike at 1Gb, 10Gb Ethernet can contend head-to-head, in terms of bandwidth, with the specialized storage and clustering fabrics. In addition, storage and computing application interfaces, namely iSCSI and iWarp Remote Direct Memory Access, have been standardized for TCP/IP over Ethernet, and are rapidly gaining ground in deployment and installed base. In this context, this paper shows how the Terminator 3 Unified Wire ASIC from Chelsio Communications solves the problems which remain in the way of converging the communication fabrics onto Ethernet, thereby opening the way for a **unified wire**.

Introducing T3

Chelsio Communications is announcing the third generation of its award winning Terminator ASIC, the high performance Terminator 3 (T3) Unified Wire Engine, which introduces numerous improvements in performance and features. T3 is designed with the expressed goal of addressing the Ethernet deficiencies which have so far hampered network convergence, including the main ones mentioned earlier. The following sections discuss how this goal is achieved, and present T3's unique features.

T3 Architecture

The following diagram shows T3's internal block level architecture. The T3 engine integrates two 10Gb and two 1Gb Ethernet MACs. It also natively supports both PCI-X 2.0, (running at 266MHz with backward compatibility to PCI-X 1.0) and PCI-E x8. With a theoretical aggregate bi-directional bandwidth of 17Gb and 40Gb respectively, both host busses allow T3 to achieve line rate 10Gb operation.



T3 Unified Wire Engine Architecture

Terminator 3 is built around a cut-through data-flow processor core, in conjunction with high bandwidth, ECC protected external memory interfaces. Most packets are fully processed without requiring off-chip storage, resulting in ultra-low latency, while the pipelined core allows line rate operation even with small packet sizes. These architectural features provide no-compromise wire speed storage, computing and networking performance. In addition, the programmable processor loads instructions at boot time, facilitating field upgrades. The diagram also shows a number of unique features, which are described next.

High Performance Unified Wire over Ethernet

The key to the high performance demonstrated by the Terminator architecture is an ideal work distribution between the programmable pipelined core and its application co-processors, and a general purpose on-chip processor. These components work together to provide:

A richly featured server adapter engine, including all stateless offloads such as IP/UDP/TCP checksum, large send/segmentation and virtualization, in addition to unique advanced features, which are described below.

A full TCP/IP sockets application offload engine, including handling of connection establishment and teardown, as well as all exceptions including retransmission and re-ordering. The implementation rivals the most popular stacks in RFC compliance, while providing much higher performance, thanks to a large reduction in packet and interrupt rate, and zero copy capability, all without requiring any application changes. The zero-copy "direct data placement", a feature unique to offload engines, allows dramatic reductions not only in CPU load, but also in memory subsystem usage, which is oftentimes the actual bottleneck in today's systems. In addition, hardware based retransmission and re-ordering virtually eliminate the impact of packet loss on performance, transforming the normally lossy Ethernet network into a high performance reliable fabric. Such performance is impossible to match with stateless offload, as well as with a number of other offload approaches. Note that the same T3 engine is used in partial offload mode, to support Microsoft's TCP Chimney offload.

A high performance storage engine, thanks to the acceleration of the iSCSI protocol. T3 provides hardware handling of the expensive byte touching operations, such as CRC computation and checking, and direct data placement of payload in host memory. This results in line rate operation with large savings in CPU load, while retaining full flexibility for the higher level protocol. In addition, the cut-through processing allows record IOPS numbers to be achieved.

An ultra-low latency cluster computing engine, with full support of the iWarp RDMA and IETF RDDP protocols for remote direct memory access. The cut-through capability of the architecture is leveraged to perform on-the-fly processing of RDMA over TCP, resulting in InfiniBand beating latency, while the core retains its line rate processing capacity for high throughput. The end result is a high performance interconnect, with very low CPU usage.

A number of supporting engines handle unique value added features which benefit all the above, including:

Wire virtualization functionality to allow the sharing of the 10Gb conduit by multiple guest entities, to each of which the wire appears as a dedicated adapter, with segregated resources.

Traffic management implemented in hardware to provide simultaneous low latency/high bandwidth operation, as well as per-flow, per-class and per-virtual adapter fine grain rate control and bandwidth allocation. The traffic manager operates through controlling the internal processing rate at which the protocol is executed. A critical piece of the unified wire engine, it is a required brick of the unified network. Note that the speed at which the traffic manager operates precludes software or firmware implementation.

Traffic filter, sniffer and firewall capability is built in, using flexible rule-based processing of incoming packets at line rate, with support for tens of thousands of rules.

Traffic steering at line rate with configurable hash-based (including full support for Microsoft's Receive Side Scaling) and rule-based mapping supporting tens of thousands of rules.

Support for Storage, Computing and Networking

The emergence of the iSCSI protocol for block storage over TCP/IP and its increasing adoption offer an alternative to specialized and difficult to use storage fabrics. With its support of iSCSI, T3 solves the performance problem introduced by iSCSI's expensive payload processing functions, thereby allowing storage applications to migrate to 10Gb Ethernet, and to benefit from the increased bandwidth available, along with T3's unprecedented IOPS rate.

Similarly, transport-independent RDMA programming interfaces initiatives are starting to bear fruit, enabling cluster computing applications to run unmodified over the high performance T3 engine.

As shown in the previous section, the T3 engine transforms Ethernet into a **reliable, high speed, traffic-managed fabric**. It provides the required support for handling the different applications, as well as the critical pieces needed to preserve their performance levels while sharing the same infrastructure. This effectively completes the required elements for a successful transition to a unified wire.

Building Systems with T3

T3 can be used in host bus adapter, line card or intermediate box configuration, with 2x10GbE or 2x1GbE ports. Additionally, it supports up to 16 1GbE ports using an external MAC, for line card or intermediate box operation.

Competitive Landscape

Looking at the 10Gb Ethernet space today, one can identify two categories of vendors:

1. The first have focused their energy solely on solving the problem of TCP processing load at 10Gb speeds.
2. The second category includes vendors who are looking beyond TCP processing. Chelsio has been and remains a pioneer in this area.

Ironically, the first category consists of vendors of "stateless offload" cards, who assert that it is possible to obtain high performance over 10Gb Ethernet by means of simple tricks. Stateless offloads are not without their merits, and for this reason, T3 includes numerous such features as described earlier. However, they do have significant limitations and their use come with caveats as to their unstudied effects on network stability. More importantly, the view of this category is based on the belief that the only problem to be solved is TCP protocol processing load. Therefore, the end result is some gains in CPU usage, which turn out to be limited and specific to certain applications and network environments. Incidentally, because stateless offload solutions have narrow applicability, vendors who have restricted themselves to this approach are still battling with the processing load problem. Clearly, the approach is fundamentally limiting, and has low chances of succeeding in the unified wire arena, as it ignores and therefore doesn't address the real obstacles to unifying communications over Ethernet.

Another fundamental fact, which is missed by the proponents of the first category, is that TCP is not "just another protocol": it is the layer in the TCP/IP networking stack at which reliable communication is provided. Therefore, a device which implements TCP offers a reliable interface to the network, in contrast to a NIC which still exposes IP's unreliable datagram service. For this reason, the dramatic reduction in processing load is not T3's only or most valuable benefit. Rather, it is the essential transformation it introduces in Ethernet's capabilities where it provides a reliable hardware interface, to which applications can tap in directly.

Chelsio solved the TCP/IP problem early on, and because its engine has access to the reliable application byte stream, it has now moved on to richer, more interesting and more rewarding grounds. The unique architecture of Chelsio's solution results in a contained, proven building block which can be used and re-used in systems supporting higher-level and more complex applications. The iSCSI and RDMA support discussed here showcases the flexibility this architecture brings in for supporting different applications.

Additionally, unlike multi-RISC based designs, the offload problem does not have to be re-solved every time a new application is to be supported. In this regard, multi-RISC CPU designs – chosen by vendors mainly to take a shortcut by porting available software stacks- are often claimed to allow unlimited programmability. This claim is quickly debunked with a closer

inspection. Indeed, the illusion of programmability is dissipated knowing that, in their minimal protocol support configuration, these engines have trouble performing better than stateless offload NICs. First, they suffer from high latency due to firmware-based processing in relatively low speed on-chip processors. Second, in terms of throughput, they require a minimum connection count to distribute load among the on-chip processors, and quickly run out of steam with a larger connection count, due to caching effects and memory contention. Therefore, this architecture hits performance limits at both low and high connection counts, and exhibits poor scaling. Typically, vendors who have followed this approach may demonstrate good performance at a sweet spot, but actual users will find that the real performance curve is disappointing. Clearly, adding more functionality is only going to degrade this reality, essentially negating the benefits of offload.

In contrast, Chelsio's performance leading Terminator 3 processes all connections in a specialized programmable data-flow processor to deliver line-rate 10Gb throughput for a single connection, up to thousands of connections, as well as low latency thanks to its cut-through processing. Moreover, as a unified wire enabler, it includes advanced features specifically designed to satisfy the requirements of each application individually, while they all share the same wire, therefore securing a successful transition to the unified network.

Conclusion

Chelsio's T3 Unified Wire Engine heralds an era where the familiar TCP/IP over Ethernet technology is equipped to support all of today's demanding applications unmodified, both in terms of performance and in application interfaces, while preserving the technology's flexibility and ease of use.

Since its introduction, the original Terminator architecture has demonstrated performance levels which remain unmatched by any other vendor. T3 further cements this leadership position with even higher throughput and IOPS numbers, and much lower latency, all the while reducing CPU and memory system usage to minimal levels. The latter is achieved thanks to direct data placement in application buffers, for storage, computing as well as unmodified sockets applications.

In conclusion, along with the standardized iSCSI and iWarp RDMA interfaces, T3 allows **storage** and **computing** applications which required specialized, expensive fabrics to run **simultaneously, unmodified and with higher performance**, over the same Ethernet infrastructure used for TCP/IP networking: the **unified wire**.

For more information about Chelsio Communications and the Terminator 3 ASIC, visit the Chelsio web site at www.chelsio.com or send an e-mail to info@chelsio.com.