The Chelsio Terminator 4 ASIC

Enabling Next-Generation Converged Network Interconnects

Abstract

This paper examines Chelsio's Terminator ASICs, focusing on the fourth-generation device known as Terminator 4 (T4). After reviewing key market trends, an overview of the chip's architecture is presented. Key features and capabilities are then discussed, with an emphasis on additions to T4. Finally, the paper looks at applications for the T4 ASIC as well as competing architectures for these applications.

Market Drivers for Network Convergence

With the proliferation of massive data centers, equipment density and power consumption are more critical than ever. At the same time cloud computing and server virtualization are driving the need for more uniform designs than traditional three-tier data-center architectures offer. In this traditional structure, servers are separated into web, application, and database tiers. These tiers connect with one another using Ethernet, typically 10Gb Ethernet (10GbE) for new installations. For storage, the web and application tiers typically use file storage provided by network-attached storage (NAS) systems connected over Ethernet.

The database tier, or "back end," has traditionally used block storage provided by a dedicated storage-area network (SAN) based on Fibre Channel. Database servers, therefore, have required both FC HBAs connected to FC switches and NICs connected to the Ethernet LAN. In addition, clustered databases and other parallel-processing applications often require a dedicated low-latency interconnect, adding another adapter and switch. Clearly, installing, operating, maintaining, and managing as many as three separate networks within the data center is grossly inefficient.

With the introduction of the third-generation Terminator (T3), Chelsio enabled a unified wire for LAN, SAN, and cluster traffic. This unified wire was made possible by the high bandwidth and low latency of 10GbE combined with storage and cluster protocols operating over TCP/IP (iSCSI and iWARP, respectively). In parallel, operating systems and hypervisors have incorporated native support for iSCSI and database applications are now supporting file-based storage protocols such as NFS as an alternative to SANs. Over the longer term, these trends make a homogeneous data-center network a reality.

Still, there exists a large installed base of Fibre Channel SANs, which must be accommodated by the evolving data-center network. Fibre Channel over Ethernet (FCoE) provides a transition path from legacy SANs to converged networks. Expanding its unified wire approach, Chelsio has added FCoE hardware support to the new T4.

Introduction to Chelsio's Terminator Architecture

Terminator is a highly integrated 10GbE controller chip built around a programmable protocol-processing engine. The T4 ASIC represents Chelsio's fourth-generation TCP offload (TOE) design, third-generation iSCSI design, and second-generation iWARP (RDMA) implementation. In addition to full TCP and iSCSI offload, T4 supports full FCoE offload. Much of the processing for these protocols is implemented in microcode running on a pipelined VLIW engine. To minimize latency, the pipeline supports simultaneous cut-through operation for both transmit and receive paths. This data-flow processing architecture also provides for wire-speed operation at small packet sizes and regardless of the number of TCP connections. Although the T4 pipeline is faster than that of T3, the new chip can run the same microcode that has been field proven in very large clusters. Chelsio provides production-level firmware, shielding customers from the details of the Terminator programming model.

Figure 1 shows a block diagram of the T4 internal architecture. For the server connection, the chip includes a PCI Express v2.0 ×8 host interface. With support for the 5Gbps Gen2 data rate, the PCIe interface provides up to 32Gbps of bandwidth to the server. T4 also adds support for PCIe I/O virtualization. On the network side, T4 integrates four Ethernet ports that support GbE as well as 10GbE operation. Using 10Gbps embedded serdes, all four ports offer direct support for 10GBASE-KR. Two ports also support the four-lane 10GBASE-CX4/KX4 interfaces. For GbE operation, all four ports offer a choice of SGMII or 1000BASE-KX. New to T4 is a DMTF-standard

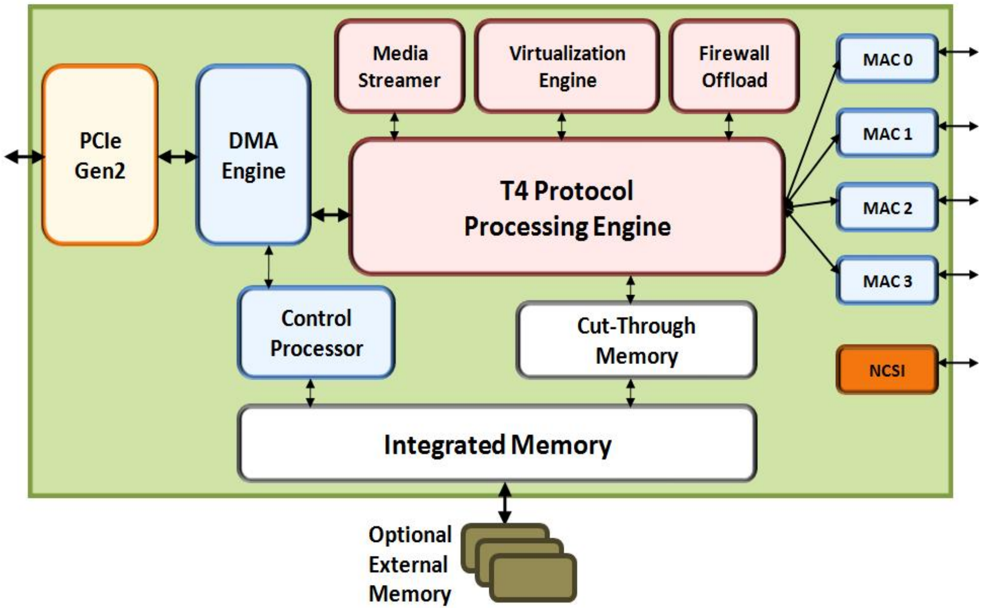

Figure 1. T4 Internal Architecture

Network Controller Sideband Interface (NC-SI) for IPMI pass through in LAN-on-motherboard (LOM) designs.

Most of T4's internal blocks are enhanced versions of those found in T3. Features that carry over from T3 include stateless offloads, packet filtering (firewall offload), and traffic shaping (media streaming). One major new block within the processing engine is an embedded switch, which is important to virtualized servers and is discussed in more detail in the next section. Another major difference between T4 and its predecessor lies in their respective memory architectures. The T3 design requires external DRAM to support its maximum throughput and external TCAM to support TCP/iSCSI/iWARP offloads. The new T4 chip integrates a TCAM and large buffer memory on chip, enabling memory-free designs with full performance and features.

New Attributes and Features of T4

Virtualization

Technologies that improve the performance and features of virtualized servers have been rapidly evolving. With T3, Chelsio supported hypervisor-offload techniques such as VMware's NetQueue. The next step in performance improvements involves hypervisor-bypass architectures, which remove hypervisor overhead by enabling virtual machines (VM) to directly access I/O resources. With multiple VMs sharing a single physical NIC or HBA, isolation becomes a critical function to ensure one VM cannot interfere with another VM.

The PCIe single-root I/O virtualization (SR-IOV) specification standardizes the sharing of one PCIe device by multiple VMs. All SR-IOV devices have at least one physical function (PF) used for device management and multiple virtual functions (VFs). In the case of T4, the chip's PCIe v2.0 interface supports 128 VFs. This means a physical server can have up to 128 VMs sharing one T4, which provides a 40Gbps unified wire for LAN and storage traffic, limited by PCIe Gen 2 x8 to 30Gbps full duplex. For backward compatibility with servers and operating environments that do not support SR-IOV, T4 also supports 8 PFs, which will make the chip look like eight physical devices using traditional PCI multifunction enumeration.

With hypervisor-based NIC sharing, the hypervisor implements a virtual L2 switch (or vSwitch) to handle local VM-to-VM traffic. When direct-access architectures are introduced, this vSwitch is bypassed and VM-to-VM traffic must be handled by the NIC hardware. This is the reason for T4's new embedded switch, which forwards traffic between the 128 virtual-NIC instances (mapped to VFs) as well as to the four physical GbE/10GbE ports. In addition to forwarding unicast packets between its 132 logical ports, the embedded switch handles multicast replication and VLAN filtering (ACLs). Additional ACLs can be implemented and prioritized through the filtering capability of the TCAM.

Requirements for the vSwitch also extend to how the embedded function interacts with the adjacent physical switch outside the server, and multiple competing approaches have been deployed or proposed. These include Cisco's VNTag, HP's VEPA and FLEX-10, and IBM's VEB. Thanks to the flexible design of the embedded switch combined with the parsing/lookup block for protocol classification, T4 supports all of these specifications.

Ultra-Low Latency iWARP and UDP

Representing Chelsio's second-generation RDMA design, T4 builds on the iWARP capabilities of T3, which have been field proven in numerous large, 100+ node clusters, including a 1300-node cluster at Purdue University. For Linux, Chelsio supports MPI through integration with the OpenFabrics Enterprise Distribution (OFED), which has included T3 drivers since release 1.2. For Windows HPC Server 2008, Chelsio shipped the industry's first WHQL-certified NetworkDirect driver. The T4 design reduces RDMA latency from T3's already low six microseconds to about two microseconds. Chelsio achieved this three-fold latency reduction through straightforward increases to T4's pipeline speed and controller-processor speed, demonstrating the scalability of the RDMA architecture established by T3.

To substantiate T4's leading iWARP latency, Chelsio will publish benchmarks separately. At two microseconds, however, T4's RDMA latency is expected to be lower than that of InfiniBand DDR solutions. Furthermore, independent tests have shown that T3's delivered latency increases by only 1.2 microseconds in a 124-node test. By comparison, InfiniBand and competing iWARP designs show large latency increases with as few as eight connections (or queue pairs). This superior scaling with node count suggests T4 should offer latencies comparable to InfiniBand QDR in real-world applications.

Although MPI is popular for parallel-processing applications, there exists a set of connectionless applications that benefit from a low-latency UDP service. These applications include financial-market data streaming and trading as well as IPTV and video-on-demand streaming. Chelsio has added UDP-acceleration features to T4 and is supplying software that provides a user-space UDP sockets API. As with RDMA, Chelsio expects T4 will deliver two-microsecond end-to-end latency for UDP packets. Application software can take advantage of T4's UDP acceleration using a familiar sockets interface.

Storage Offloads

Like T3, T4 offers protocol acceleration for both file- and block-level storage traffic. For file storage, T3 and T4 support full TOE under Linux and TCP Chimney under Windows. T4's fourth-generation TOE design adds support for IPv6, which has become a requirement for many government and wide-area applications. For block storage, both T3 and T4 support partial iSCSI offload, where the ASICs offload processing-intensive tasks such as PDU recovery, header and data digest, CRC generation/checking, and direct data placement (DDP). The third-generation iSCSI design implemented by T4 adds support for full iSCSI offload, which enables support under VMware ESX.

Broadening Chelsio's support for block storage, T4 adds support for both partial and full offload of the FCoE protocol. Using an HBA driver, full offload provides maximum performance as well as compatibility with SAN-management software. For customers that prefer to use a software initiator, Chelsio supports the Open-FCoE stack and T4 offloads certain processing tasks much as it does in iSCSI. Unlike iSCSI, however, FCoE requires several Ethernet enhancements that are being standardized by the IEEE 802.1 Data Center Bridging (DCB) task group. To enable lossless transport of FCoE traffic, T4 supports Priority-based Flow Control (PFC), Enhanced Transmission Selection (ETS), and the DCB exchange (DCBX) protocol.

When combined with iWARP, which enables NFSRDMA, LustreRDMA and similar protocols, T4 makes for an ideal Unified Target adapter, simultaneously processing iSCSI, FCoE, TOE, NFSRDMA, LustreRDMA, CIFS and NFS traffic.

Flexible Network-Port Options

With four multispeed network ports and embedded 10Gbps serdes interfaces, T4 supports a variety of configurations. For example, an upgradeable LOM design can use two ports for GbE as the base configuration and the remaining two ports as a 10GbE upgrade option. In a blade-server design, T4 provides redundant 10GBASE-KR or -KX4 backplane connections without requiring external PHY devices. All four network ports support direct connection to SFP+ modules using SFI. The chip supports link aggregation across 4×10GbE ports, providing a 40Gbps unified wire in advance of 40G Ethernet availability.

Low Bill of Materials Cost

By integrating memories and making other enhancements to T4, Chelsio has dramatically reduced the system cost of fully featured LOM and NIC designs alike. With T4, external memories are optional and do not affect performance. In a memory-free LOM design, the chip supports its maximum throughput and can offload up to 1K connections. By adding commodity DDR2 or DDR3 SDRAM, NIC designs can support up to a massive 1M connections. A typical 2×10GbE NIC/HBA design would use three DDR3 devices to support 512K connections. Such a design fits easily within a low-profile PCIe form factor. Aside from the optional DRAM devices, T4 requires less than $2 in external components. For thermal management, the chip requires only a passive heat sink.

T4 Applications

By supporting the newest virtualization and protocol offloads, T4 delivers a truly universal design for server connectivity. Thanks to T4’s high level of integration, customers can instantiate this universal design as 10GbE LOM, upgradeable GbE LOM, blade-server mezzanine cards, or PCIe adapters in standard or custom form factors. Chelsio’s unified wire design allows customers to support a broad range of protocols and offloads using a single hardware design (or SKU), reducing the support and operational costs associated with maintaining multiple networking options (See Figure 2). With its support for full FCoE offload, for example, T4 eliminates the need for customers to offer optional converged network adapters (CNAs) specifically for FCoE. Chelsio offers the only design that offloads all types of network-storage traffic plus cluster traffic.

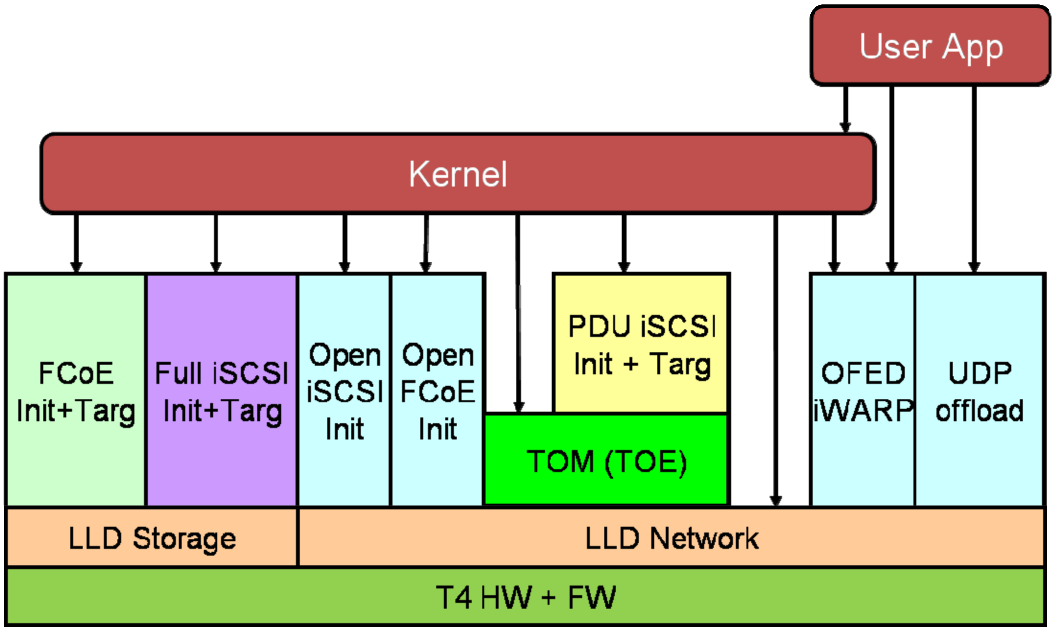

Figure 2. Linux Unified Wire

Virtualization is critically important to new server designs and I/O technologies are evolving rapidly in response to the virtualization trend. New server designs must anticipate future requirements such as offloads under development by operating-software vendors. With support for SR-IOV, a very large number of VMs, and the newest protocols for virtual networking, T4 delivers a state-of-the-art virtualization design. Virtualization is driving dramatic increases in server utilization, which means fewer CPU cycles are available for I/O processing. By providing virtualization-compatible offloads, such as full iSCSI offload, T4 preserves precious CPU cycles for application processing.

With broad and proven support for file- and block-level storage, T4 is also ideal for networked storage systems. In NAS filer/head designs, T4 provides full TOE for Linux and FBSD-based embedded operating systems. Similarly, T4 fully offloads iSCSI and FCoE processing in SAN targets such as storage arrays. Full offload has the dual benefits of minimizing host-processor requirements and easing software integration. By simultaneously supporting TOE/iSCSI/FCoE/iWARP, T4 is the ideal Unified Target adapter, enabling NAS/SAN systems that adapt to and grow with end-customer needs.

For high-performance computing (HPC) applications, T4 combines industry-leading iWARP latency with robust production-level software. Chelsio's support for various commercial and open MPI variants—including HP MPI, Intel MPI, Scali MPI, MVAPICH2, and Open MPI—means that many parallel-processing applications will run over 10GbE without modification. This software compatibility plus excellent RDMA performance eliminates the need for a dedicated interconnect, such as InfiniBand, for cluster traffic. By bringing RDMA to LOM designs, T4 also opens up horizontal applications like clustered databases that fall outside the traditional HPC space.

Alternative Architectures

Although Chelsio pioneered 10GbE TOE and iSCSI, a number of competitors now offer 10GbE controllers with TOE and/or iSCSI offload. These competing designs, however, use a fundamentally different architecture from that of Terminator. Whereas Chelsio designed a data-flow architecture, competitors use a pool of CPUs operating in parallel. These CPUs are typically simple 32-bit RISC designs, which are selected for ease of programming rather than optimal performance in packet processing. An incoming packet must be classified to identify its flow and it is then assigned to the CPU responsible for that flow.

Implementing TCP processing across parallel CPUs introduces a number of architectural limitations. First, performing complete protocol processing in firmware running on a single CPU leads to high latency. Because iSCSI and iWARP operate on top of the TOE, processing these protocols only adds to total latency. Second, these designs can exhibit problems with throughput scaling based on the number of TCP connections. For example, some designs cannot deliver maximum throughput when the number of connections is smaller than the number of CPU cores.

At the other end of the spectrum, performance may degrade at large connection counts due to how connection state is stored. Assuming each CPU can store state (or context) for a small number of connections in local cache, connection counts that exceed this local storage will create cache misses and require high-latency external-memory accesses.

These parallel-CPU designs can demonstrate adequate throughput when benchmarked by a vendor using a controlled set of parameters. For the reasons discussed above, however, their performance will vary in real-world testing based on connection counts and traffic patterns. Although some of these vendors claim their designs support RDMA, none has demonstrated acceptable iWARP latency.

By contrast, third parties have demonstrated Terminator's deterministic throughput and low latency. Chelsio's unique data-flow architecture delivers wire-speed throughput with one connection or tens of thousands of connections. Furthermore, Terminator provides equal bandwidth distribution across connections. The T4 ASIC improves latency and integration while maintaining the proven Terminator architecture.

Conclusions

The concept of network convergence around 10GbE has been discussed in the industry for some time. But changes of this magnitude do not happen overnight. While iSCSI adoption has grown rapidly, a large installed base of FC SANs reside in data centers. To bridge the gap between today's reality and tomorrow's unified network, FCoE has emerged as an alternative to iSCSI for these deployed SANs. Unlike FC, however, Ethernet was not designed for reliable delivery. As a result, FCoE requires enhancements to the Ethernet protocol that are not yet widely deployed in data-center infrastructures. In parallel, server and network virtualization are orthogonal trends driving system-architecture changes.

Against this backdrop of diverse and dynamic requirements, creating a universal 10GbE controller is a daunting task. Offloading protocols such as iSCSI and iWARP requires a reliable high-performance underlying TCP engine. For storage and cluster traffic alike, latency is increasingly important. Virtualization requires significant new hardware functions to support both VM isolation and VM-to-VM communication. Finally, a universal design must deliver a high level of integration to meet the space and cost requirements of LOM and mezzanine designs.

With its fourth-generation ASIC, Chelsio has taken the unified wire to the next level. T4 delivers an unmatched feature set combined with a single-chip design. No other vendor can offer a single SKU for TOE, iSCSI, FCoE, and iWARP. Why settle for partial solutions to server connectivity when Chelsio makes a universal solution today?