The Chelsio Terminator 5 ASIC

Enabling Next Generation Converged Network Interconnects

Abstract

Chelsio Communications, Inc. a leading provider of Ethernet Unified Wire Adapters and ASICs, has announced Terminator 5, the fifth generation of its high performance Ethernet silicon technology. The Terminator 5 (T5) ASIC is built upon the latest iteration of Chelsio's industry-proven, protocol-rich high speed network processing architecture, which has been widely deployed with more than 200 OEM platform wins and more than 400,000 ports shipped worldwide.

T5 is a highly integrated, hyper- virtualized 10/40GbE controller built around a programmable protocol-processing engine, with full offload of a complete Unified Wire solution comprising NIC, TOE, iWARP RDMA, ISCSI, FCoE and NAT support. T5 provides no-compromise performance with both low latency (1µsec through hardware) and high bandwidth, limited only by the PCI bus. Furthermore, it scales to true 40Gb line rate operation from a single TCP connection to thousands of connections, and allows simultaneous low latency and high bandwidth operation thanks to multiple physical channels through the ASIC.

Designed for high performance clustering, storage and data networks, the T5 enables fabric consolidation by simultaneously supporting wire-speed TCP/IP and UDP/IP socket applications, RDMA applications and SCSI applications, thereby allowing InfiniBand and FibreChannel applications to run unmodified over standard Ethernet. The API used for the complete software suite (on Linux, Windows and FSBD) for current T4 installations is the same for the T5 chip and future 100Gb capable versions, leveraging all the software investment that has been made in T4 deployments.

This paper provides a close examination of the T5. 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 T5. Finally, the paper looks at applications for the T5 ASIC as well as competing architectures for these applications.

Market Drivers for Network Convergence

With the proliferation of massive data centers, equipment density and total 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 the switchable/routable IP protocol over Ethernet, typically 10/40GbE Ethernet for new installations. For storage, the web and application tiers typically use file storage provided by network-attached storage (NAS) systems connected with the IP protocol 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 Ethernet NICs connected to the IP network. In addition, clustered databases and other parallel-processing applications often require a dedicated low-latency interconnect, adding another adapter and switch, perhaps even a new fabric technology. Clearly, installing, operating, maintaining, and managing as many as three separate networks within the data center is expensive in terms of both CAPEX and OPEX.

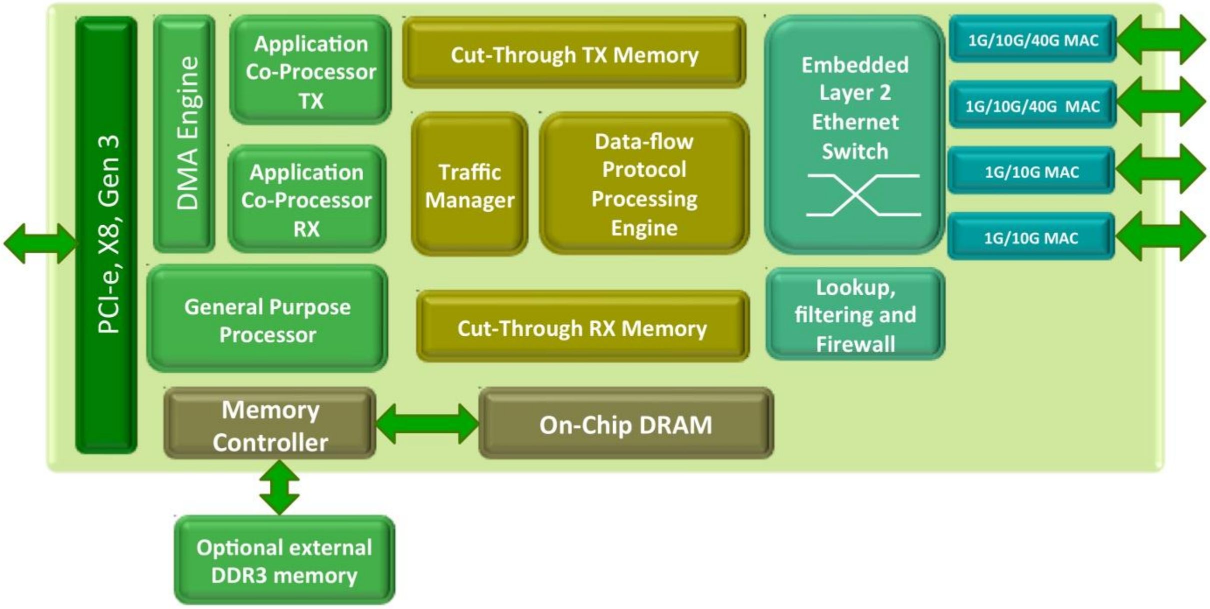
With the introduction of the Terminator 4 (T4) in 2010, Chelsio enabled a unified IP protocol over Ethernet wire for virtualized LAN, SAN, and cluster traffic. The virtualization features are implemented using a Virtual Interface (VI) abstraction that can be mapped onto the SR-IOV capability of PCIe or can use regular PCIe. This unified wire is made possible by the high bandwidth and low latency of 10GbE combined with storage and cluster protocols operating over TCP/IP (iSCSI and iWARP RDMA, 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. To enable iSCSI in HA Enterprise applications T5 adds comprehensive and flexible T10 DIF/DIX support and increasing the maximum IOPS rate. Going forward, these trends make a homogeneous IP Ethernet data-center network a reality.

There exists a large installed base of Fibre Channel SANs, which is accommodated by the evolving data-center network. Fibre Channel over Ethernet (FCoE) provides a transition path from legacy SANs to converged IP Ethernet networks. Expanding its unified wire approach, Chelsio has enhanced FCoE hardware support in the new T5 adding T10 DIF/DIX support and increasing the maximum IOPS rate.

In addition to adding T10 DIF/DIX support, and improving the IOPS rate, the iSCSI support in T5 benefits from offering HA failover capability between ports of the T5, and also between different T5 adapters. The T10 DIF/DIX capability can optionally be employed per connection between the host and a T5 and/or on the Ethernet wire. The adapter can therefore be used to implement HA iSCSI storage targets, with T10 DIX protection for all transfers between the target host and T5 adapter, and optional T10 DIF support on the Ethernet wire, and to/from T10 DIF compatible HDD.

Introduction to Chelsio's Terminator Architecture

Terminator T5 is a tightly integrated 4x10/2x40 GbE controller chip built around a highly scalable and programmable protocol-processing engine. Much of the processing of the offloaded protocols is implemented in microcode running on a pipelined proprietary data-flow engine. The pipeline supports cut-through operation for both transmit and receive paths for minimum latency. The transport processor is designed for wire-speed operation at small packet sizes, regardless of the number of TCP connections. The T5 ASIC represents Chelsio's fifth-generation TCP offload (TOE) design, fourth generation iSCSI design, and third generation iWARP RDMA implementation. In addition to full TCP and iSCSI offload, T5 supports full FCoE offload. T5 supports failover between different ports of the T5 chip, as well as between different T5 adapters. Any TOE, iSCSI, or iWARP RDMA connection can fail over between different ports or between different adapters. Although the T5 pipeline is twice as fast as the T4 pipeline, the new chip can run the same microcode that has been field proven in very large clusters. Chelsio provides a uniform firmware interface across T4 and T5, and shields customers from the details of the Terminator hardware programming model.


PCI-e, X8, Gen 3
DMA Engine
Application Co-Processor TX
Application Co-Processor RX
Cut-Through TX Memory
Traffic Manager
Data-flow Protocol Processing Engine
Embedded Layer 2 Ethernet Switch
1G/10G/40G MAC
1G/10G/40G MAC
1G/10G MAC
1G/10G MAC
General Purpose Processor
Cut-Through RX Memory
Lookup, filtering and Firewall
Memory Controller
On-Chip DRAM
Optional external DDR3 memory

Terminator 5 Block Diagram

The figure above shows a block diagram of the T5 internal architecture. For the server connection, the chip includes a PCI Express v3.0 ×8 host interface. With support for the 8Gbps Gen3 data rate, the PCIe interface provides up to 60 Gbps of bandwidth to the server. On the network side, T5 integrates four Ethernet ports, two of which support 40GbE, while all support 10GbE and 1GbE operation. Using 10Gbps embedded serdes, two ports offer direct support for

KR/XFI/SFI, XAUI, CX4/KX4, KX1 40GbE/10GbE interfaces, and two ports offer direct support for 10GbE KR/XFI/SFI, KX1 interfaces. For 1GbE operation, all four ports offer a choice of SGMII or 1000BASE-KX.

All T4 features carry over to T5, including stateless offloads, packet filtering (firewall offload), and traffic shaping (media streaming), and the embedded switch. T5's internal blocks benefit from enhancements in performance and features over the corresponding versions found in T4. One major new block within the processing engine is a T10 DIF/DIX generation and validation block.

New Attributes and Features of T5

Virtualization

Technologies that improve the performance and features of virtualized servers have been rapidly evolving. With T4, Chelsio added support for SR-IOV and 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 T5 uses the SR-IOV capability of PCIe, T5 support for Access Control Lists (ACL), T5 support for packet filtering, replication, and steering, and finally T5 support per VM QoS bandwidth provisioning, to realize the VM separation.

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 T5, the chip's PCIe v3.0 interface supports 128 Virtual Interfaces (VI) with dedicated statistics and configuration settings, which can optionally be mapped to 128 VFs when SR-IOV is used. This means a physical server can have up to 128 VMs sharing one T5, which provides a 2x40Gbps unified wire for LAN and storage traffic, limited by PCIe Gen 3 x8 to 60Gbps full duplex bandwidth. For backward compatibility with servers and operating environments that do not support SR-IOV, T5 also supports 8 PFs, which will make the chip look like up to eight physical devices using traditional PCI multifunction enumeration, and the 128 VIs can be mapped to the 8 PFs to support up to 128 VMs. Furthermore, T5 can support up to 1K VMware NetQueue and Hyper-V VMQueue instances.

With hypervisor-based NIC sharing, the hypervisor implements a virtual L2 switch (or vSwitch) with the following capabilities:

- applies anti-spoofing ACL to outbound packets
- filters and steers inbound packets to one or more VM
- replicates multi-cast and broadcast packets to different VM
- handles local VM-to-VM traffic subject to ACL processing

When direct-access architectures are introduced, parts of the vSwitch functionality are offloaded to the T5 L2 switch. The embedded T5 L2-L7 switch implements a superset of the above vSwitch capabilities and has the following functions:

- applies anti-spoofing ACL to outbound packets
- switching/routing inbound traffic at line rate to one or more VM with low latency and in addition between the four 10/40Gbe Ethernet ports
- replicates multi-cast and broadcast packets to different VM and is in addition capable of unicast mirroring/replication, subject to VLAN filtering (ACLs)
- compatible with IEEE 802.1Qbg/h requirements and forwards unicast packets between its 140 logical ports
- supports additional ACLs such as Firewall filtering, NAT, and full TCP proxy switching with the T5 filtering capability

The T5 embedded switch can be configured from the network or from the host to be compatible with any of Cisco's VNTag, HP's VEPA, FLEX-10 and FlexFabric, IBM's VEB and DOVE, or with SDN such as OpenFlow.

Traffic Management and QoS

The T5 enhances the T4 QoS capabilities and in addition to supporting ETS it supports SLAs that limit each VM to a fixed allotment of the available bandwidth, and connections within the VM e.g. MPEG-4 connections to a fixed rate with low jitter.

Ultra-Low Latency iWARP and UDP

Representing Chelsio's third-generation iWARP RDMA design, T5 builds on the RDMA capabilities of T3 and T4, which have been field proven in numerous large, 100+ node clusters, including a 1300-node cluster at Purdue University. The T5 adds support for atomic RDMA operations, immediate RDMA data, and advanced end-to-end completion semantics for HA applications. For Linux, Chelsio supports MPI through integration with the OpenFabrics Enterprise Distribution (OFED), which has included T3/T4 drivers since release 1.2. For Windows HPC Server 2008, Chelsio shipped the industry's first WHQL-certified NetworkDirect driver. For Windows Server 2012 Chelsio shipped the industry's first WHQL-certified SMB 3.0 driver. The T5 design reduces RDMA latency from T4's already low three microseconds to 1.5 microseconds. Chelsio achieved this two-fold latency reduction through increases to T5's pipeline speed and controller-processor speed, demonstrating the scalability of the RDMA architecture established by T3/T4.

To substantiate T5's leading iWARP latency, Chelsio will publish benchmarks separately. At 1.5 microseconds, however, T5's RDMA user-to-user latency is expected to be lower than that of InfiniBand DDR solutions. Furthermore, independent tests have shown that latency for previous versions of the Terminator ASIC 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 T5 should offer latencies comparable to InfiniBand QDR and FDR 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 T5 and is supplying software that provides a user-space UDP sockets API. As with RDMA, Chelsio expects T5 will deliver 1.5 microsecond end-to-end latency for UDP packets. Application software can take advantage of T5's UDP acceleration using a familiar sockets interface.

Storage Offloads

Like T4, T5 offers protocol acceleration for both file- and block-level storage traffic. For file storage, T4 and T5 support full TOE under Linux and TCP Chimney under Windows, as well as the latest RDMA-based SMB Direct protocol, a key part of Windows Server 2012's major SMB 3.0 update. For block storage, both T4 and T5 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 T5 adds support for T10 DIF/DIX optionally between the host computer and T5 and/or on the Ethernet wire. The iSCSI design implemented by T5 improves support for full iSCSI offload, which enables support under VMware ESX.

Broadening Chelsio's support for block storage, T5 improves the 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 T5 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 and T5 support 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, and combined with the T5 QoS SLA capabilities, the T5 makes for an ideal Unified Target adapter, simultaneously processing iSCSI, FCoE, TOE, NFSRDMA, LustreRDMA, CIFS and NFS traffic.

Low Bill of Materials Cost

By integrating memories and making other enhancements to T5, Chelsio has reduced the system cost of fully featured LOM and NIC designs alike. With T5, 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 1M connections. A typical 2×40GbE NIC/HBA design would use five DDR3 devices to support 32K connections. Such a design fits easily within a low-profile PCIe form factor. Aside from the optional DRAM devices, T5 requires less than two dollars in external components. For thermal management, the chip requires only a passive heat sink.

T5 Applications

By supporting the newest virtualization and protocol offloads, T5 delivers a universal design for server connectivity. With the T5’s high level of integration, customers can instantiate this universal design as 2x40GbE 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. With its support for full FCoE offload, for example, T5 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.

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-system software vendors. With support for SR-IOV, a very large number of VMs, and the newest protocols for virtual networking, T5 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, T5 preserves precious CPU cycles for application processing, and equally important, lowers the overall power consumption of the data center.

With broad and proven support for file- and block-level storage, T5 is also ideal for networked storage systems. In NAS filer/head designs, T5 provides full TOE for Linux and FBSD-based operating systems. Similarly, T5 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, T5 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, T5 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, T5 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 generic 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 RDMA 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 iWARP RDMA, none has demonstrated acceptable iWARP latency or scalability when the number of Queue Pairs (QP) is increased.

By contrast, third parties have demonstrated Terminator's deterministic throughput and low latency. The T5 40GbE line-rate bandwidth performance is achieved by a modest increase in the pipeline core frequency, which still leaves ample head room for scaling to 100GbE performance. 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 T5 ASIC improves latency and integration while maintaining the proven Terminator architecture.

Conclusions

The concept of network convergence around 10/40GbE has been discussed in the industry for some time. But changes of this magnitude do not happen overnight. While iSCSI adoption has grown rapidly, there is a large installed base of FC SANs that 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 legacy SANs. Unlike FC, however, Ethernet is not designed for reliable end-to-end delivery. As a result, FCoE requires enhancements to the Ethernet protocol that are not widely deployed in data-center infrastructures, and to complicate deployment multi-hop FCoE implementations use incompatible "standards."

Against this backdrop of diverse and dynamic requirements, creating a universal IP protocol over 10/40GbE controller offers a superior ROI for the customer. Offloading protocols such as iSCSI and iWARP requires a reliable high-performance underlying TCP engine. For storage and cluster traffic alike, low-latency/high-IOPS is increasingly important. Virtualization requires significant new hardware functions to support both VM isolation and VM-to-VM communication. Finally, a universal design delivers a high level of integration to meet the total space and cost and power budget requirements of LOM and mezzanine designs.

With its fifth-generation T5 ASIC, Chelsio has taken the unified wire to the next level. T5 delivers an unmatched feature set combined with a single-chip design. No other vendor offers a single SKU for NIC, TOE, iSCSI, FCoE, and iWARP RDMA. Why settle for partial solutions to server connectivity when Chelsio makes a universal solution today?

###