The Chelsio Terminator 6 ASIC

Next-Generation Converged Secure Network Interconnects

Abstract

Chelsio Communications, Inc. a leading provider of Ethernet Unified Wire adapters and ASICs, has announced Terminator 6, the sixth generation of its high performance Ethernet silicon technology. The Terminator 6 (T6) 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 800,000 ports shipped worldwide.

T6 is a highly integrated, hyper-virtualized 10/25/40/50/100GbE 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 NVMe over fabrics support. T6 supports low-latency line rate TLS/SSL, IPsec, and SMB 3.X crypto. T6 supports overlay networks VXLAN, NVGRE, GENEVE and NAT. The T6 has an integrated Layer-2 to Layer-7 switch. T6 provides no-compromise performance with both low latency (1µsec end-to-end) and high bandwidth, limited only by the PCI bus. Furthermore, it scales to true 100GbE 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 overlay data networks, the T6 enables fabric consolidation by simultaneously supporting wire-speed TCP/IP and UDP/IP socket applications, RDMA applications and SCSI applications, thereby allowing InfiniBand and Fibre Channel applications to run unmodified over standard Ethernet. The API used for the complete software suite (on Linux, Windows and FreeBSD) for current T5 installations is the same for the T6, leveraging all the software investment that has been made in T5 deployments.

This paper provides a close examination of the T6. 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 T6. Finally, the paper looks at applications for the T6 ASIC as well as competing architectures for these applications.

Market Drivers for Network Convergence

With the proliferation of public and private clouds, 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 offering. In this traditional structure, servers are separated into secure web 2.0 applications and database tiers. Representative security requirements include TLS everywhere, data encryption at rest and SMB 3.X file server security. The tiers connect with one another using the switchable/routable IP protocol over Ethernet, typically 10/25/40/50/100GbE Ethernet for new installations. There exist lossless Ethernet enhancements that are effective in

limited deployments and closed environments, but TCP/IP fabrics are ubiquitous high bandwidth, low latency and have superior scaling features. For storage, the web 2.0 and application tiers typically use file storage provided by network-attached storage (NAS) systems connected with the TCP/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 Terminator (T5), 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 10/40GbE 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 SMB 3.X and NFS as an alternative to SANs. To enable iSCSI in HA Enterprise applications, T6 adds comprehensive and flexible T10-DIF/DIX support and increases 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. To aid in this conversion, the T6 supports FCoE at 100Gbps line rate.

The iSCSI support in T6 includes 100Gbps line rate and in addition has been enhanced and has improved CPU utilization compared to T5. The iSCSI support for NVMe and other SSD devices benefits from the low latency and high IOPS rate of the T6.

Introduction to Chelsio's Terminator Architecture

Terminator 6 (T6) is a tightly integrated 2x10/2x25/2x40/2x50/2x100GbE 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 flexible pipelined proprietary data-flow engine. The flexible 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 T6 ASIC represents Chelsio's sixth-generation TCP offload (TOE) design, iSCSI design, and iWARP RDMA implementation. In addition to full TCP and iSCSI offload, T6 supports full FCoE offload and NVMe over RDMA fabrics (NVMe-oF) offload. T6 supports failover between different ports of the T6 chip, as well as between different T6 adapters. Any TOE, iSCSI, iWARP RDMA, and NVMe-

oF connection can fail over between different ports or between different adapters. Although the T6 pipeline is more than twice as fast as the T5 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 T5 and T6, and shields customers from the details of the Terminator hardware programming model.

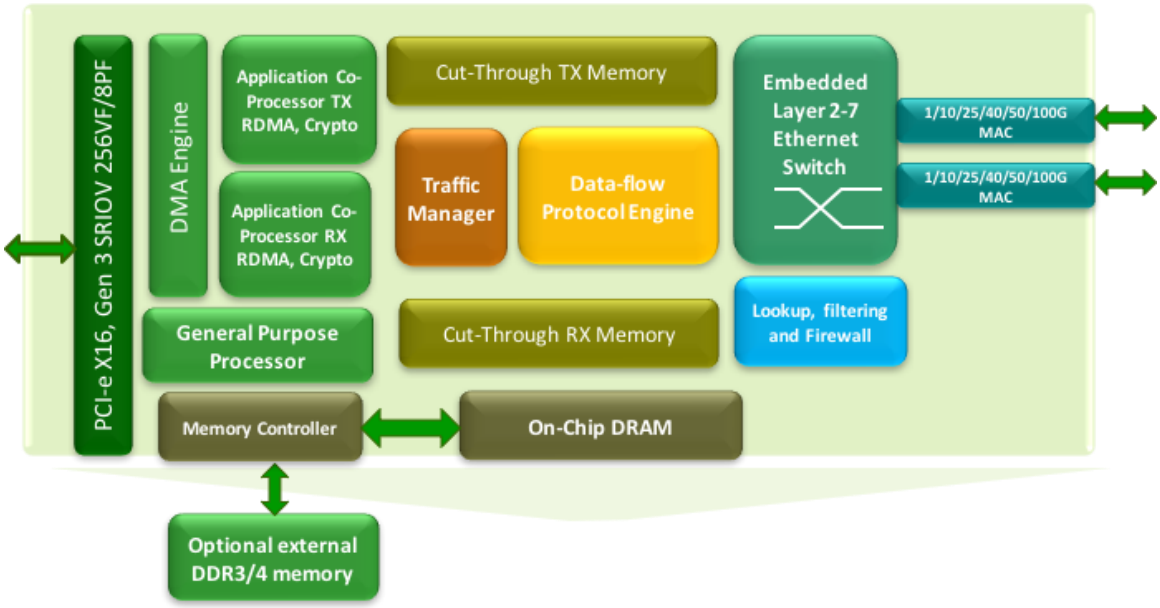

Figure 1 - Terminator 6 Architecture

The figure above shows a block diagram of the T6 internal architecture. For the server connection, the chip includes a PCI Express v3.0 x16 host interface. With support for 8Gbps Gen3 data rate, the PCIe interface provides up to 110Gbps of bandwidth to the server. On the network side, T6 integrates two Ethernet ports, both of which support the whole suite 1/10/25/40/50/100GbE operation. Using 25Gbps embedded serdes, the ports offer direct support for 10GbE standards SFP+ limited mode, XFP, 10Gbase-KR, 10Gbase-KX4, and 1GbE standards, 1000base-KX, and SGMII.

All T5 features carry over to T6, including stateless offloads, packet filtering (firewall offload), IEEE 1588 timestamping, traffic shaping (media streaming), and the embedded switch. T6's internal blocks benefit from enhancements in performance and features over the corresponding versions found in T5. One major new block within the processing engine is an AES/SHA encryption and decryption block.

New Attributes and Features of T6

Virtualization

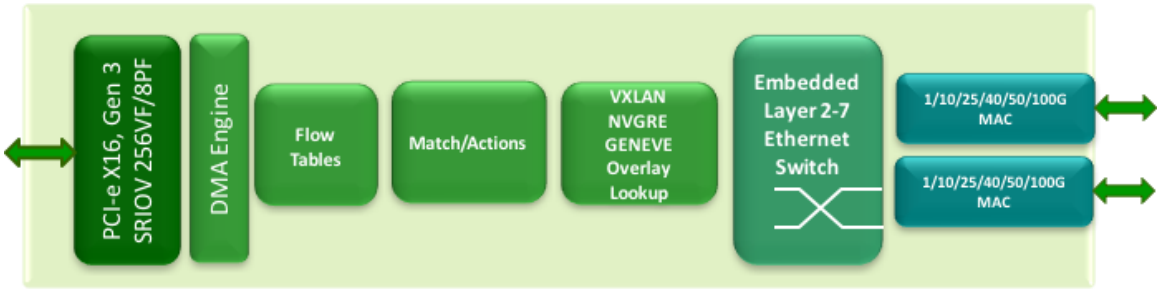

Figure 2 - Virtualization Block Diagram

Secure Cloud and Web 2.0 technologies such as SDN and OVS offloads with crypto have been evolving rapidly, and T6 integrates several technologies that improve the integration and performance of these features. Like T5, T6 supports Layer-2 based overlay networks based on Cisco's VNTag; HP's VEPA, FLEX-10 and FlexFabric; and IBM's VEB and DOVE. In addition, the T6 offloads Layer-3 overlay technologies VXLAN, NVGRE, and GENEVE header encapsulation and de-capsulation, as well as stateless and iSCSI, iWARP, NVMe-oF, TLS/SSL and TOE state-full offloads of inner packets. In addition, the flexible pipeline design implements Match/Actions flow tables for SDN and OVS offloads. The capacity and performance is summarized in the following tables:

Table 1 - OVS Offload Performance

OVS Offload	Performance
Matching Rules	Up to 500K matching rules
Rule Update Rate	Up to 200K updates/sec
Rule Lookup Rate	Up to 40M/sec@1%CPU

Table 2 - SDN Offload Performance

SDN	Capacity
Flow Table-1	• Overlay network L2/VLAN and inner L2/VLAN • Line rate replication • 512 entries
Flow Table-2	• L2/L3/L4/L5 • Up to 500K entries
Chaining	Flow Table-2 back into Flow Table-1

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 case of T6, the chip's PCIe v3.0 interface supports 256 Virtual Interfaces (VI) with dedicated statistics and configuration settings that can optionally be mapped to 256 VFs when SR-IOV is used. This means that a physical

server can have up to 256 VMs sharing one T6, which provides up to 2x100Gbps unified wire for LAN and storage traffic, limited by PCIe Gen 3 x16 to 110Gbps full duplex bandwidth. For backward compatibility with servers and operating environments that do not support SR-IOV, T6 also supports 8 PFs, which will make the chip appear as 8 physical devices using traditional PCI multifunction enumeration, and the 256 VIs can be mapped to the 8 PFs to support up to 256 VMs. Furthermore, T6 can support up to 1K VMware NetQueue and Hyper-V VMQueue instances.

The T6 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; or IBM's VEB and DOVE, or with OVS or SDN such as OpenFlow within OpenStack.

Traffic Management and QoS

The T6 enhances QoS capabilities, and in addition to supporting ETS, 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 fourth-generation iWARP RDMA design, T6 builds on the RDMA capabilities of T3 and T5, which have been field proven in numerous large, 100+ node clusters, including a 1300-node cluster at Purdue University. The RDMA capabilities have also been proven with GPUDirect installations. The T6 adds support for SRQ (Shared Receive Queues), 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 and T5 drivers since release 1.2. For Windows HPC Server 2008, Chelsio shipped the industry's first WHQL-certified Network Direct driver. For Windows Server 2012, Chelsio shipped the industry's first WHQL-certified SMB 3.X driver. The T6 design reduces RDMA latency from T5's already low 1.5 µs to 1 µs. Chelsio achieved this two-fold latency reduction through increase to T6's pipeline speed and controller-processor speed, demonstrating the scalability of the RDMA architecture established by T3 and T5.

To substantiate T6's leading iWARP latency, Chelsio will publish benchmarks separately. At 1 µs, however, T6's RDMA user-to-user latency is expected to be on par with InfiniBand solutions. Furthermore, independent tests have shown that latency for previous versions of the Terminator ASIC increases by only 1.2 µs 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 T6 should offer latencies comparable to InfiniBand EDR 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 enhanced UDP-acceleration features to T6 and is supplying software that provides a user-space UDP sockets API. As with RDMA, Chelsio expects T6 will deliver 1 µs end-to-end

latency for UDP packets. Application software can take advantage of T6 UDP acceleration using a familiar sockets interface.

Storage Offloads

Like T5, T6 offers protocol acceleration for both file and block-level storage traffic. For file storage, T5 and T6 support full TOE in FreeBSD and Linux, as well as the crypto features, and RDMA feature (SMB Direct) of the Windows SMB 3.X protocol.

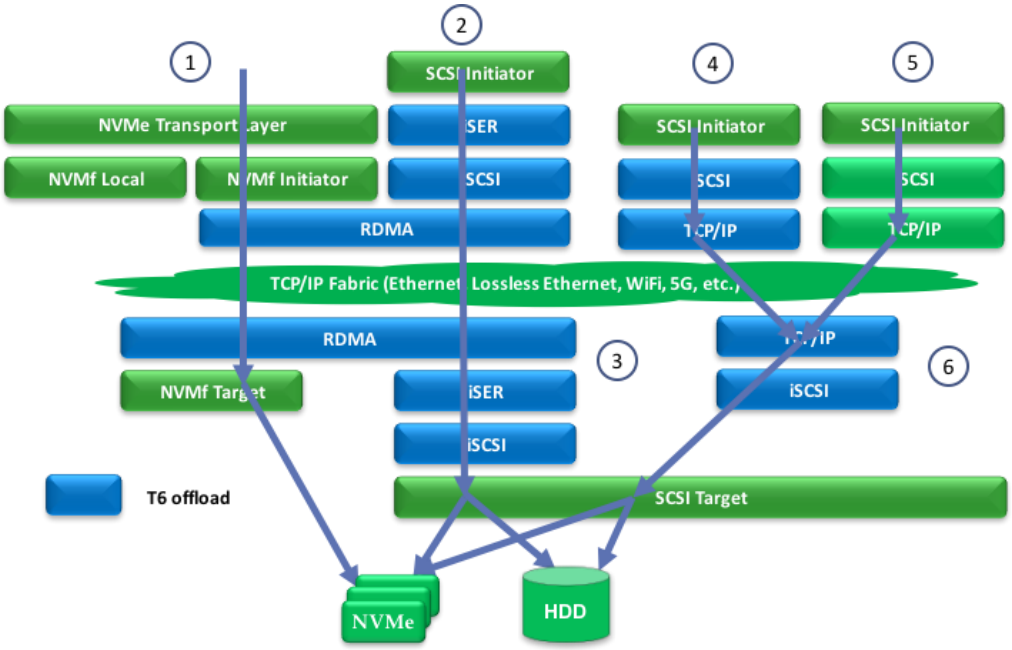

Figure 3 - Storage Offload using T6

For block storage, T6 preserves the existing product investment while at the same time being future proof by supporting emerging SSD and PM (Persistent Memory) technologies. The T6 is designed to achieve optimal performance for existing iSCSI products and optimal performance for emerging technologies such as NVMe over RDMA fabrics.

For block storage, both T5 and T6 support partial and full protocol offload of iSCSI and NVMe over RDMA fabrics (NVMe-oF), where the ASICs either offload all aspects of the protocols on the initiator/client and target/server sides, or the processing-intensive tasks such as PDU recovery, header and data digest, CRC generation/checking, and direct data placement (DDP). Like T5, the T6 supports T10-DIF/DIX optionally between the host computer and T6 and/or on the Ethernet wire.

Refer to the above diagram: The T6 supports offload of NVMe-oF on the client (1) and server (3) sides, supports iSCSI over RDMA fabrics iSER (2) (3), and also iSCSI over TCP/IP (4)(6). The iSCSI over TCP/IP operates with the same efficiency as iSER when there are iSCSI offload devices on the initiator and target sides (4) (6): same bandwidth, same IOPS, and same CPU usage. In addition, T6 iSCSI over TCP/IP operation is efficient for use cases with software initiators (5) and T6 offloaded targets (6).

T6 supports 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. FCoE protocols like RoCEv1 and RoCEv2 require several Ethernet enhancements. To enable lossless transport of FCoE traffic, T5 and T6 support Priority-based Flow Control (PFC), Enhanced Transmission Selection (ETS) and the DCB exchange (DCBX) protocol. The T6 TCP/IP offloads for iSCSI and RDMA are compatible with these enhanced features, as well as IP ECN congestion avoidance, but T6 TCP/IP offloads do not require these for high bandwidth, low latency operation in congested environments.

When combined with iWARP, which enables SMB 3.X, NFSRDMA, Lustre RDMA and similar protocols, and combined with the T6 QoS SLA capabilities, the T6 makes for an ideal Unified Target adapter, simultaneously processing NVMe-oF, iSCSI, FCoE, TOE, SMB 3.X, NFSRDMA, Lustre RDMA, CIFS and NFS traffic.

Integrated L2-L7 Switch

The T6 integrated switch enables storage rack designs and NFV racks with servers connected directly without involving a Top of Rack (ToR) switch.

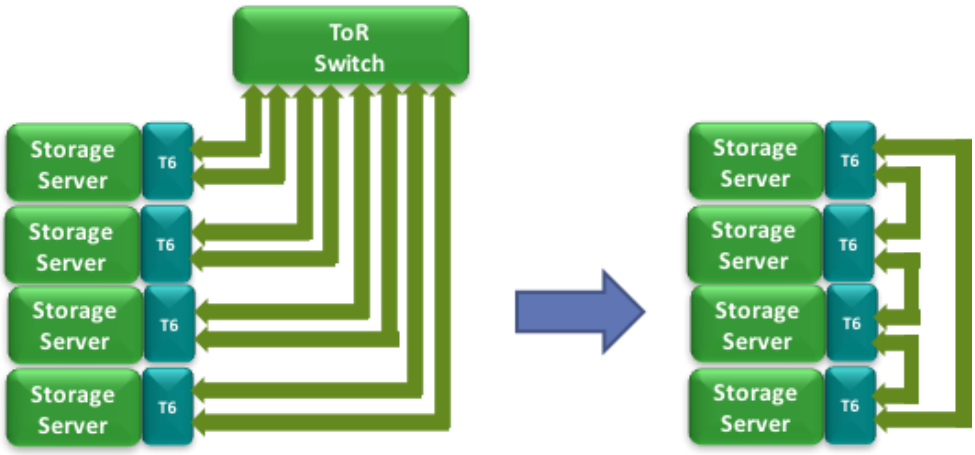

Figure 4 - Storage and NFV Rack Designs with T6

The T6 integrated switch has a full suite of L2-L7 features including ACL with support for L2 Ethernet switching, L3 routing, NAT, TCP Proxy, iSCSI proxy and L7 payload tag switching. These features enable flexible high bandwidth, low latency and high IOPS designs, while at the same time lowering TCO (Total Cost of Ownership), reducing CAPEX (Capital Expenses) by reducing switch port expenses and lowering OPEX (Operating Expenses) by lowering power usage by employing offload technologies.

Security Offloads

The T6 supports all the most popular AES/SHA cipher suites in TLS/SSL in-line mode with 100Gbps bandwidth and less than 2 µs end-to-end latency. The typical T6 SKU supports 32K simultaneous TLS sessions and the T6 has support for up to 1M simultaneous sessions. The in-line mode achieves TCP/IP processing and TLS/SSL AES/SHA processing in cut-through fashion to achieve optimal bandwidth and latency. A co-processor mode of operation is supported for

TLS/SSL, SMB 3.X, IPsec, data at rest, encryption/decryption, authentication and data de-dupe fingerprint generation. The TLS/SSL session key negotiation is performed by software on a host computer. The T6 is therefore ideal for TLS/SSL encryption authenticated media streaming applications and also has efficient support for SMB 3.X, IPsec, data at rest, encryption/decryption, authentication and data de-dupe fingerprint generation. The performance of the AES and SHA protocol suites is summarized in the following table:

Table 3 - AES & SHA performance

Cipher	BW	Latency
AES-CBC	Encryption=30Gbps/Decryption=100Gbps	< 10µs
SHA1	40Gbps	< 10µs
SHA224/256/384/512	25-40Gbps	< 10µs
AES-GCM/CTR/XTS	100Gbps	< 1µs

The supported options with the AES and SHA protocols are summarized in the following tables:

Table 4 - AES & SHA protocol: Cipher only modes (encryption/decryption only)

Cipher	Key Sizes supported	Protocol Requirement
AES-CBC	128, 192, 256	TLS, IPSEC, SMB 3.X
AES-CTR	128, 192, 256	IPSEC
AES-XTS	128, 192, 256	Generic Protocol

Table 5 - AES & SHA protocol: Combined cipher modes (authentication and encryption/decryption)

Cipher	Key Sizes supported	Protocol Requirement
AES-GCM	128, 192, 256	TLS, IPSEC, SMB 3.1
AES-CCM	128, 192, 256	SMB 3.X (co-processor only)

Table 6 - AES & SHA protocol: Authentication and generic hash modes

Hash Function	Key Sizes supported	ICV Size	Protocol Requirement
SHA1 **SHA224/256/384/512**	Equal to the output of hashing algorithm, it is expected that longer keys will be hashed to L bits. Refer RFC2104.	Variable	TLS, IPSEC, Generic
SHA2-224-HMAC **SHA2-256-HMAC** **SHA2-256-HMAC** **SHA2-384-HMAC** **SHA2-512-HMAC**	Equal to the output of hashing algorithm, it is expected that longer keys will be hashed to L bits. Refer RFC2104.	Variable	TLS, IPSEC

For a description of T6 crypto use cases we refer the accompanying T6 ASIC crypto white paper.

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 T6, 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 4K connections. By adding commodity DDR3 or DDR4 SDRAM, NIC designs can support up to 1M connections. A typical 2x100GbE NIC/HBA design would use five DDR3/4 devices to support 32K connections and less than 20W. Such a design fits easily within a low-profile PCIe form factor. Aside from the optional DRAM devices, T6 requires less than $2 in external components. For thermal management, the chip requires only a passive heat sink.

T6 Applications

By supporting the newest cloud virtualization and security protocol offloads, T6 delivers a universal design for server connectivity. Thanks to T6's high level of integration, customers can instantiate this universal design as 2x10/2x25/2x40/2x40/2x100GbE 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, T6 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.

Secure multi-tenant clouds are critically important to new server designs and I/O technologies are evolving rapidly in response to the private and public cloud 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, T6 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, T6 preserves precious CPU cycles for application processing, and equally important, lowers the overall power consumption of the datacenter.

With broad and proven support for file and block-level storage, T6 is also ideal for networked storage systems. In NAS filer/head designs, T6 provides full TOE for Linux and FreeBSD-based operating systems. Similarly, T6 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 secure NVMe-oF/TOE/iSCSI/FCoE/ iWARP, T6 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, T6 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 100GbE 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, T6 also opens up horizontal applications like clustered databases that fall outside the traditional HPC space.

Alternative Architectures

Although Chelsio pioneered 100GbE 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 T6 100GbE line-rate bandwidth performance is achieved by a modest increase in the pipeline core frequency. This still leaves ample headroom for scaling beyond 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 T6 ASIC improves latency and integration while maintaining the proven Terminator architecture.

Conclusion

The concept of secure network convergence around 10/25/40/50/100GbE has been discussed in the industry for some time. The security aspect is being addressed by the ubiquitous TLS/SSL, the SMB 3.X crypto support and by IPsec. 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 datacenters. 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 secure IP protocol over 10/25/40/50/100GbE controller offers a superior ROI for the customer. Offloading protocols such as iSCSI and iWARP and TLS/SSL requires a reliable high-performance underlying TCP engine. For secure 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 sixth-generation T6 ASIC, Chelsio has taken the unified wire to the next level. T6 delivers an unmatched feature set combined with a single-chip design. No other vendor offers a single SKU for NVMe-oF, NIC, TOE, iSCSI, FCoE, and iWARP RDMA that concurrently supports in-line TLS/SSL and that supports SMB 3.X crypto and IPsec. Why settle for partial solutions to server connectivity when Chelsio makes a universal solution today?