Boosting Data Transfer with TCP Offload Engine Technology

on Ninth-Generation Dell PowerEdge Servers

TCP/IP Offload Engine (TOE) technology makes its debut in the ninth generation of Dell™ PowerEdge™ servers, and is expected to boost data transfer performance for many key enterprise applications. This article explores the anticipated benefits of TOE technology being developed by Broadcom, Dell, and Microsoft.

BY PANKAJ GUPTA, ALLEN LIGHT, AND IAN HAMEROFF

Related Categories:

Broadcom

Data networking

Dell ninth-generation servers

Internet Protocol (IP)

Microsoft Windows

Network fabric

Network interface card (NIC)

Scalable Networking Pack (SNP)

TCP/IP Offload Engine (TOE)

Visit www.dell.com/powersolutions for the complete category index.

The online economy—particularly for e-business, entertainment, and collaboration—is dramatically increasing the amount of Internet traffic to and from enterprise servers. Most of this data is going through the TCP/IP stack and Ethernet controllers. Additionally, many IT organizations are moving storage, high-performance computing (HPC), and database cluster networks to Ethernet. As a result, Ethernet controllers are experiencing heavy network traffic, which requires increased system resources to process network packets. Because TCP/IP consumes a significant amount of host CPU processing cycles, a heavy TCP/IP load may leave few system resources available for other applications. This situation can create a bottleneck for applications involving significant network traffic or limit the amount of CPU cycles available for actual application processing. Applications that exhibit this behavior include file-oriented storage, block-oriented storage, backups, database transactions, and tightly coupled distributed applications such as HPC.

The TCP/IP Offload Engine (TOE) model is designed to improve data transfer performance over IP networks by relieving much of the overhead when processing TCP/IP from the host CPU. TOE allows the OS to offload all TCP/IP traffic to specialized hardware on the network adapter while leaving TCP/IP control decisions to the host server. By relieving the host processor bottleneck, TOE can help deliver the performance benefits administrators expect from applications running across high-speed network links. TOE is also cost-effective because it processes the TCP/IP stack on a high-speed network device that requires less processing power than a general-purpose high-performance CPU.

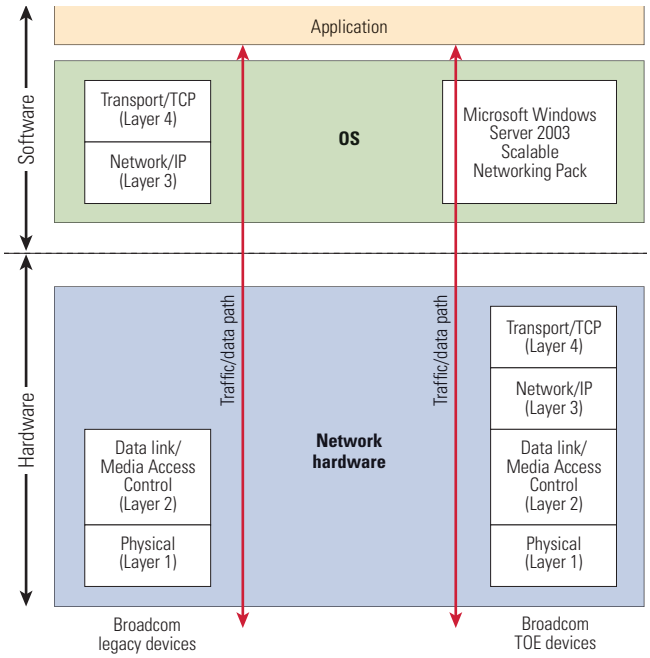

Figure 1. Networking layers implemented in a TOE NIC and a standard Gigabit Ethernet NIC

TOE technology from Dell, Broadcom, and Microsoft

Dell, Broadcom, and Microsoft have developed a TOE implementation that will be available on the Dell PowerEdge server product line. This implementation is based on Broadcom® NetXtreme® II network devices—such as the BCM5708C and BCM5708S 10/100/1,000 Mbps Ethernet controllers, which are integrated on the system board in dual-socket PowerEdge servers and available as optional add-in PCI-e adapters on other platforms—and the Microsoft® Windows Server® 2003 Scalable Networking Pack (SNP). Both include TOE and stateless offload functionality. The Broadcom devices implement Open System Interconnection (OSI) Layers 1, 2, 3, and 4 functionality and in the future may include some Layer 5 functionality. NetXtreme II devices handle the data processing of TCP/IP connections offloaded to them by the OS. The OS handles control and management of TCP/IP processing such as connection setup, connection termination, system resource allocation, prevention of denial-of-service attacks, and error and exception handling. Figure 1 provides a comparison of the various OSI layers implemented in a TOE network interface card (NIC) and a standard Gigabit Ethernet NIC.

To take advantage of the TOE feature in Dell PowerEdge servers equipped with Broadcom NetXtreme II devices, administrators must install the Microsoft Windows Server 2003 SNP. The SNP supports networking hardware enhancements designed to promote scalability and performance improvements on servers running Windows Server 2003. The architectural innovations provided in the SNP include TCP Chimney Offload, which enhances the existing software implementation of TCP/IP version 4 currently in Windows Server 2003.

TCP Chimney Offload creates a direct connection between the top of the protocol stack and the software drivers to enable partial offload of the protocol stack. The term *chimney* was used because the data is transferred to the top of the protocol stack without moving through the intermediate protocol layers, like smoke rising through a chimney to the top of a building without having to travel through the intermediate floors. Figure 2 shows the TCP Chimney Offload architecture.

The SNP also supports Receive Side Scaling (RSS), a network technology that enables the efficient distribution of network receive processing across multiple CPUs in multiprocessor systems.

The NetXtreme II devices are designed to support 1,024 simultaneous offloaded connections. The decision to offload the connection or pass it through the software protocol stack is made by the OS and is seamlessly handled by a software switch, as shown in Figure 2. A connection can be uploaded—that is, moved back from the hardware to the OS—for a variety of reasons. For example, the system automatically switches to the software protocol stack after 1,024 connections are offloaded per Ethernet port. The host or offload target—in this case, the NetXtreme II device—can also request uploading or termination of an offloaded connection. In addition, the Network Driver Interface Specification (NDIS) 5.2 intermediate driver can upload a connection indirectly by requesting the host TCP/IP stack to upload the connection. If the link fails, the event is reported to the host stack through an NDIS Plug and Play (PnP) event and the host determines the most appropriate course of action.

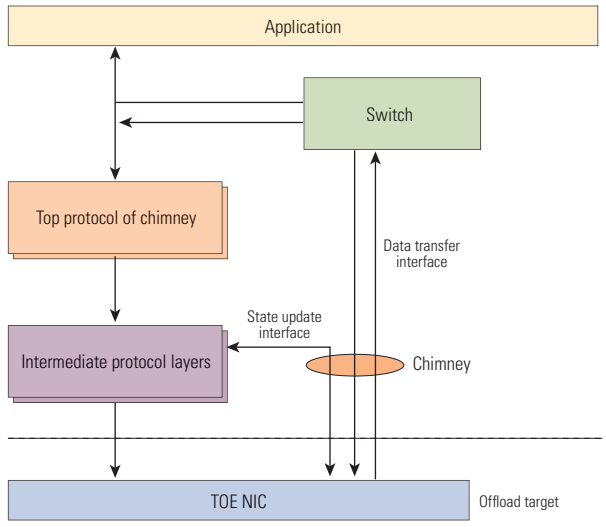

Figure 2. TCP Chimney Offload architecture

TOE components and dependencies

The minimum system requirements for TOE functionality are as follows:

- Microsoft Windows Server 2003 with Service Pack 1 (or later)
- Microsoft Windows Server 2003 SNP (KB91222)—factory installed on supported Dell PowerEdge servers for a TOE LAN on Motherboard (LOM) or installed as part of a TOE add-in NIC customer kit
- Broadcom BCM5708C or BCM5708S add-in NIC or integrated LOM
- TOE license key installed prior to boot on LOMs (the TOE license is built into add-in NICs and provided by default on a system with Broadcom BCM5708C or BCM5708S LOMs)
- Broadcom NetXtreme II drivers (NDIS 5.2 drivers)

The NIC operates in non-TOE mode if any of the preceding components are missing or if an OS other than Microsoft Windows Server 2003 with Service Pack 1 (or later) is running.

Dell platforms such as the PowerEdge 1900, PowerEdge 1950, PowerEdge 1955, PowerEdge 2900, and PowerEdge 2950 servers have integrated TOE capability with the LOM ports. TOE add-in NICs are also available for these and other Dell PowerEdge systems.[1] TOE licensing for Dell PowerEdge servers with Broadcom BCM5708 LOMs is automatically enabled via a hardware key plug-in module. The TOE enablement key is factory installed in these PowerEdge servers and add-in NICs. No additional work is required to enable the hardware feature. TOE software components—the Microsoft Windows Server 2003 SNP and Broadcom NetXtreme II drivers—are preinstalled with factory-installed Microsoft Windows Server 2003; otherwise, the software is provided for installation. Administrators should refer to the user's guide included with the network driver package for information about installing the necessary TOE software components.

TOE performance expectations

Network performance improvements gained from TOE technology can be determined by measuring either the increase in absolute network throughput or the reduction in system resources such as CPU utilization. TOE performance benefits vary with the type of applications being run. Applications with a small network packet size may experience gains in network throughput, while applications with a large network packet size may not show significant network throughput improvements with TOE but may experience a significant reduction in CPU utilization—thereby helping to keep CPU processing cycles available for other business-critical applications such as database, backup storage, media streaming, and file server applications. Applications that require extensive network utilization—such as network backups, network attached storage, file servers, and media streaming—typically benefit the most from TOE technology.

Applications with a small network packet size may experience gains in network throughput, while applications with a large network packet size may not show significant network throughput improvements with TOE but may experience a significant reduction in CPU utilization—thereby helping to keep CPU processing cycles available for other business-critical applications.

Configuring the test environment

A team of engineers from Dell, Broadcom, and Microsoft tested the TOE implementation during the product development process from May 2005 to May 2006. The test configuration included a Dell PowerEdge 2900 server equipped with 1 GB of double data rate 2 (DDR2) memory and a dual-core Intel® Xeon® processor at 3.2 GHz with a 1,066 MHz frontside bus. Intel Hyper-Threading Technology was enabled on the processor. Tests were performed with on-board Broadcom BCM5708C and BCM5708C-based add-in NICs.

The test results in this article highlight the advantages of TOE; however, real-world performance will vary based on the specific configuration running in a given enterprise environment. To demonstrate the potential performance benefits of TOE, Dell and its partners ran tests with TOE enabled using the Broadcom TOE NIC and without TOE enabled using an otherwise comparable industry-standard Layer 2 Gigabit Ethernet[2] adapter. The same configuration was used for both the TOE and non-TOE test environments.

Testing the impact of TOE on application performance

Network throughput and CPU utilization are common ways to measure a network adapter's performance. Network throughput is the rate in megabits per second that packets are sent and received, while CPU utilization is the percentage of CPU capacity required to process the packets. The Dell, Microsoft, and Broadcom

[1] The Dell PowerEdge 830, PowerEdge 850, PowerEdge 1800, PowerEdge 1850, PowerEdge 2800, PowerEdge 2850, PowerEdge 6800, and PowerEdge 6850 servers also support TOE add-in NICs.

[2] This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

team used benchmarking tools such as IxChariot 6.20 from Ixia[3] and WebBench 5.0[4] to measure network throughput and CPU utilization in the test environment.

The ratio of network throughput to CPU utilization can be expressed as the performance efficiency (PE) index, which was originally developed by *PC Week* (now *eWEEK*) in 1995. The PE index is a commonly used performance metric for evaluating network adapters. For example, high PE indexes indicate high throughput with low CPU utilization, suggesting favorable overall system performance. As Figure 3 indicates, the PE index of the TOE NIC in the test environment described in this article exceeded the PE index for the legacy Layer 2 Gigabit Ethernet NIC for every network I/O size from 1 KB to 64 KB.

TOE-enabled network adapters are designed to achieve high throughput at small network packet sizes and low CPU utilization at large network packet sizes. Figure 4 shows results from the Ixia IxChariot 6.20 performance benchmarking tool using the IxChariot's high-performance script on the test configurations described in this article.

As the number of TOE-enabled ports in a system increases, network throughput can be expected to scale more effectively using TOE-enabled network adapters than standard Layer 2 network adapters. At the same time, the CPU utilization can be expected to remain low enough to allow other processes or applications to run. As shown in Figure 5, when using only four Gigabit Ethernet network controllers that are not TOE-enabled, nearly 60 percent of the PowerEdge 2900 server's CPU was utilized just processing network traffic. But with TOE enabled on the four Gigabit Ethernet network controllers, the CPU utilization was less than 20 percent in the Figure 5 test scenario.

The network scaling effect demonstrated in Figure 5 indicates that TOE-enabled Gigabit Ethernet controllers can help servers handle larger workloads than they can handle using a standard Layer 2 Gigabit Ethernet controller. For example, backup applications can benefit from significant performance improvements with TOE enabled, as shown in Figure 6. In this test scenario, a TOE-enabled network adapter allowed the PowerEdge 2900 server to back up the same amount of data (80 GB per stream) requiring approximately half the CPU utilization than when the PowerEdge 2900 was equipped with a standard Layer 2 network adapter, which means a backup server with a TOE-enabled NIC can support additional backup streams. This is an example of how TOE can enable Dell PowerEdge servers to deliver enhanced application performance.

RSS allows applications such as Web servers to process more requests with a large number of connections. For example, Figure 7 shows test results using the WebBench 5.0 benchmark tool. These

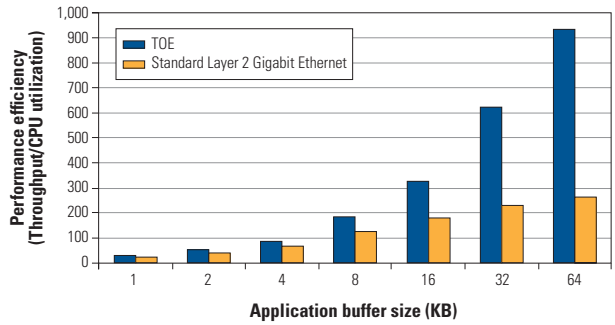

Figure 3. Performance efficiency: TOE-enabled Gigabit Ethernet NIC versus standard Layer 2 Gigabit Ethernet NIC

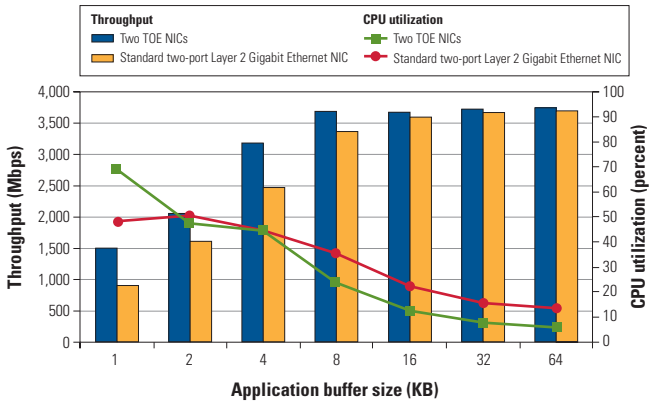

Figure 4. Two-port performance measured by the IxChariot 6.20 benchmark: TOE-enabled Gigabit Ethernet NIC versus standard Layer 2 Gigabit Ethernet NIC

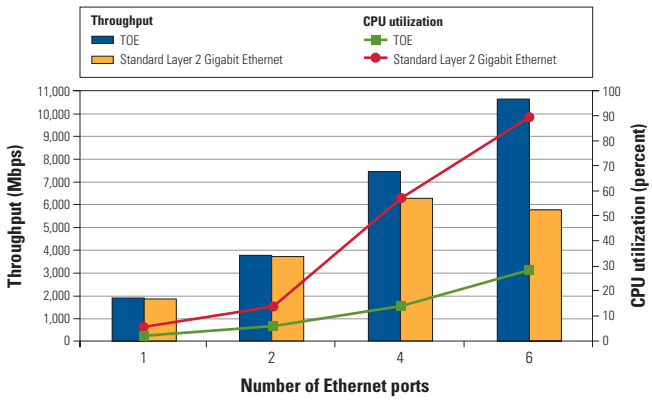

Figure 5. Throughput scaling and the effect on CPU utilization: TOE-enabled Gigabit Ethernet NIC versus standard Layer 2 Gigabit Ethernet NIC (without TOE enabled)

[3] For more information about Ixia IxChariot 6.20, visit www.ixiacom.com/ixchariot.

[4] For more information about WebBench 5.0, visit www.veritest.com/lionbridge.

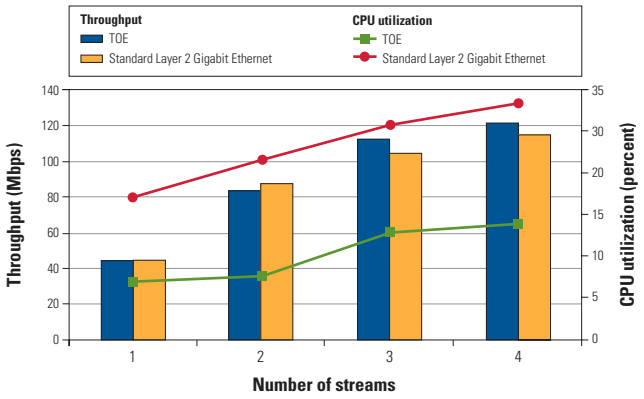

Figure 6. Backup performance: TOE-enabled Gigabit Ethernet NIC versus standard Layer 2 Gigabit Ethernet NIC

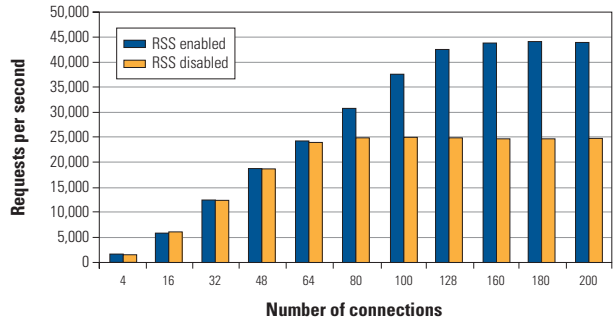

Figure 7. WebBench 5.0 benchmark performance: TOE-disabled Gigabit Ethernet NIC with RSS enabled and standard Layer 2 Gigabit Ethernet NIC with no RSS support

results highlight the performance advantage of enabling RSS on a TOE-disabled adapter compared to a standard Layer 2 Gigabit Ethernet adapter.

TOE can significantly benefit typical Web server environments in which connections are long lived, such as configurations for load-balancing and online shopping. Applications such as ad servers may not show the same benefits of using TOE-enabled NICs because of the overhead for uploading and offloading many short-lived connections. However, the majority of Web applications maintain connections over longer periods of time than the benchmarks used to measure performance in the example scenarios tested in this article.

Increased throughput, reduced CPU utilization for enhanced application performance

Dell PowerEdge servers such as the PowerEdge 1900, PowerEdge 1950, PowerEdge 1955, PowerEdge 2900, and PowerEdge 2950 have integrated TOE capability with the LOM ports. TOE-enabled PCI-e add-in adapters are also available for these and other Dell PowerEdge servers. To enable TOE functionality, servers must run Microsoft Windows Server 2003 with Service Pack 1 (or later) using the Microsoft Windows Server 2003 Scalable Networking Pack and Broadcom NetXtreme II drivers (NDIS 5.2). TOE technology is designed to enhance application performance by increasing network throughput and reducing CPU utilization for applications requiring extensive network activity. Applications such as backup, storage, file servers, and media streaming are likely to benefit most from TOE-enabled systems.

Pankaj Gupta works in the Dell Product Group and led the TOE project. His interests include communication, networking, operating systems, and the development of system software. He has a master's degree in Computer Science from Polytechnic University of New York.

Allen Light is the product line manager for the Broadcom NetXtreme II C-NIC Ethernet controllers at Broadcom Corporation, an Irvine, California–based provider of integrated circuits for wired and wireless communications. He has worked in marketing and application engineering at Intel Corporation and Philips Semiconductors. Allen has a B.S.E.E. from the University of California at Davis, and an M.B.A. from the University of New Mexico.

Ian Hameroff is the senior product manager in the Security and Access Product Marketing Group at Microsoft Corporation, where he is responsible for the product management and marketing of the Microsoft Windows Server platform networking technologies. In this role, Ian drives the go-to-market strategy for Windows Server networking and solutions focused on key security and networking initiatives, such as server and domain isolation, scalable networking, and IPv6 adoption. He has a bachelor's degree in Computer Science with a minor in Mathematics from Plattsburgh State University in New York and is a Certified Information Systems Security Professional (CISSP).

FOR MORE INFORMATION

Senapathi, Sandhya, and Rich Hernandez. "Introduction to TCP Offload Engines." *Dell Power Solutions,* March 2004. www.dell.com/downloads/global/power/1q04-her.pdf

Microsoft Windows Server 2003 Scalable Networking Pack: www.microsoft.com/snp

Broadcom NetXtreme II network devices:
www.broadcom.com/products/Enterprise-Small-Office/Gigabit-Ethernet-Controllers/BCM5708C
www.broadcom.com/products/Enterprise-Small-Office/Gigabit-Ethernet-Controllers/BCM5708S

Network teaming technical brief:
support.dell.com/support/edocs/network/582dx/teaming.htm

