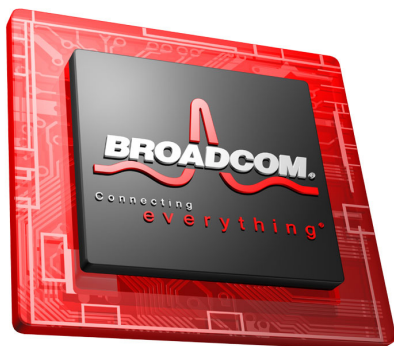


## Windows® TCP Chimney: Network Protocol Offload for Optimal Application Scalability and Manageability

The new TCP Chimney Offload Architecture from Microsoft enables offload of the TCP protocol stack to a TCP offload engine (TOE)-based network interface card (NIC). This results in full integration of TOE capabilities within the Windows® operating environment, while preserving customer investments in applications, manageability, and security.

NetXtreme II™ converged network interface controllers (C-NICs) allow data center administrators to maximize the value of available server resources, which allows servers to share Gigabit Ethernet (GbE) network ports for data networking, clustering, and storage traffic, while removing network processing overhead and delivering optimal application scalability. These benefits are provided with no changes to application software, while maintaining the existing user and software management interfaces.

March 2005



## Unified Ethernet Communications

An emerging approach to improving data center return on investment (ROI) is focused on evolving GbE TCP/IP networking to address the requirements of client/server, clustering and storage communications through the deployment of a unified Ethernet communications fabric. The vision of such a network architecture makes it non-disruptive to an existing data center infrastructure, while providing significantly better performance at a fraction of the cost—all while preserving the existing investment in the server and network infrastructure.

At the root of the emergence of unified Ethernet data center communications are the following three networking standards:

- TCP/IP over GbE
- Remote Direct Memory Access (RDMA) over TCP
- iSCSI

### TCP/IP Over GbE

TCP/IP over GbE is the predominant protocol suite for enterprise data networking. TCP/IP local area networks (LANs) have primarily been implemented using Ethernet, which has seen steady growth from 10 Mbps (megabits per second) in 1990, through 100 Mbps in the mid-1990s, to 1 Gbps (gigabit per second) today. Specifications for 10-Gbps Ethernet were ratified, and the technology is in the process of being deployed by early adopters.

### RDMA Over TCP

RDMA over TCP technology allows the converged network interface controller, under the control of the application, to bypass the networking stacks and place data directly into and out of a remote application memory. Bypass of the networking stack along with direct data placement removes the need for data copying and enables support for low-latency communications involving clustering and storage.

Given the significance of this technology, a number of networking industry leaders, including Broadcom, Cisco, Dell, EMC, Hewlett Packard, IBM, Microsoft, Intel, and Network Appliance, have come together to support the development of an RDMA over TCP protocol standard. RDMA over TCP provides for standardized and highly affordable clustering, storage, and other application protocols, replacing proprietary and more expensive technologies with existing GbE infrastructure.

## iSCSI

iSCSI is designed to enable end-to-end block storage networking over TCP/IP Gigabit networks. iSCSI is a transport protocol for SCSI, which operates on top of TCP through encapsulation of SCSI commands in a TCP data stream. iSCSI is emerging as an alternative to parallel SCSI or Fibre Channel within the data center as a block I/O transport for a range of applications, including SAN/NAS consolidation, messaging, database, and high-performance computing.

While the processing power of server CPUs that run TCP/IP protocol software has increased over time, the demands of existing TCP/IP-based applications are growing faster. For clustering and storage networking, low CPU utilization and low latencies are required. TCP/IP over Ethernet can replace the alternative technologies (i.e., Fibre Channel or proprietary clustering hardware), but a significantly better approach, other than software-based protocol processing, is required.

## TCP Offload Engines for Network Protocol Processing Efficiency

Historically, networks have been slow (relative to the processing power of the CPU), and handing off the TCP/IP functionality to the CPU was adequate. Network technology and traffic have evolved at a much faster rate than CPU performance to the point that processing TCP/IP at Gigabit speeds can overwhelm even the most modern CPUs. A common rule is that it takes 1 GHz of server CPU capacity to process 1 Gbps of TCP/IP network traffic (in one direction or half-duplex). To address this shortcoming, companies are building new devices with TCP/IP offload engines.

The TOE technology offloads the TCP/IP protocol stack to a dedicated controller in order to reduce TCP/IP processing overhead in servers equipped with standard Gigabit network interface controllers. While TOE technology has been the focus of significant vendor engineering investment, a number of obstacles have been in the way of broad-based TOE deployment.

The first generation TOE products have been based on an all-inclusive approach for TCP/IP offload, including offload of connection setup, data path offload, and support for ancillary protocols such as Dynamic Host Configuration Protocol (DHCP), Address Resolution Protocol (ARP), Internet Control Message Protocol (ICMP), and Internet Group Management Protocol (IGMP). Vendors were attempting to ship the offloaded TCP/IP stack as a parallel protocol stack due to the lack of standard operating systems' interfaces that enable the integration of TOE capabilities.

Implementing a TOE approach that requires duplicating all the previously mentioned functionality creates a number of important challenges with respect to complexity, manageability, and network security.

The first challenge leads to the administrative complexity of having to manage two separate TCP/IP implementations within the same system, each of which has its own repository for connection-state and protocol-specific information. In addition, continuity of experience for users and continued support for networking features, such as link aggregation, virtual LANs (VLANs), load balancing and failover, and the forwarding of traffic between network interfaces, may be put at risk.

The second challenge is that the offloading connection setup may also have security implications. By having centralized connection management in operating system software that has been hardened and tuned over a number of years, there is a single, mature location to protect against denial-of-service and other attacks along with, typically, many more CPU resources to perform this protection.

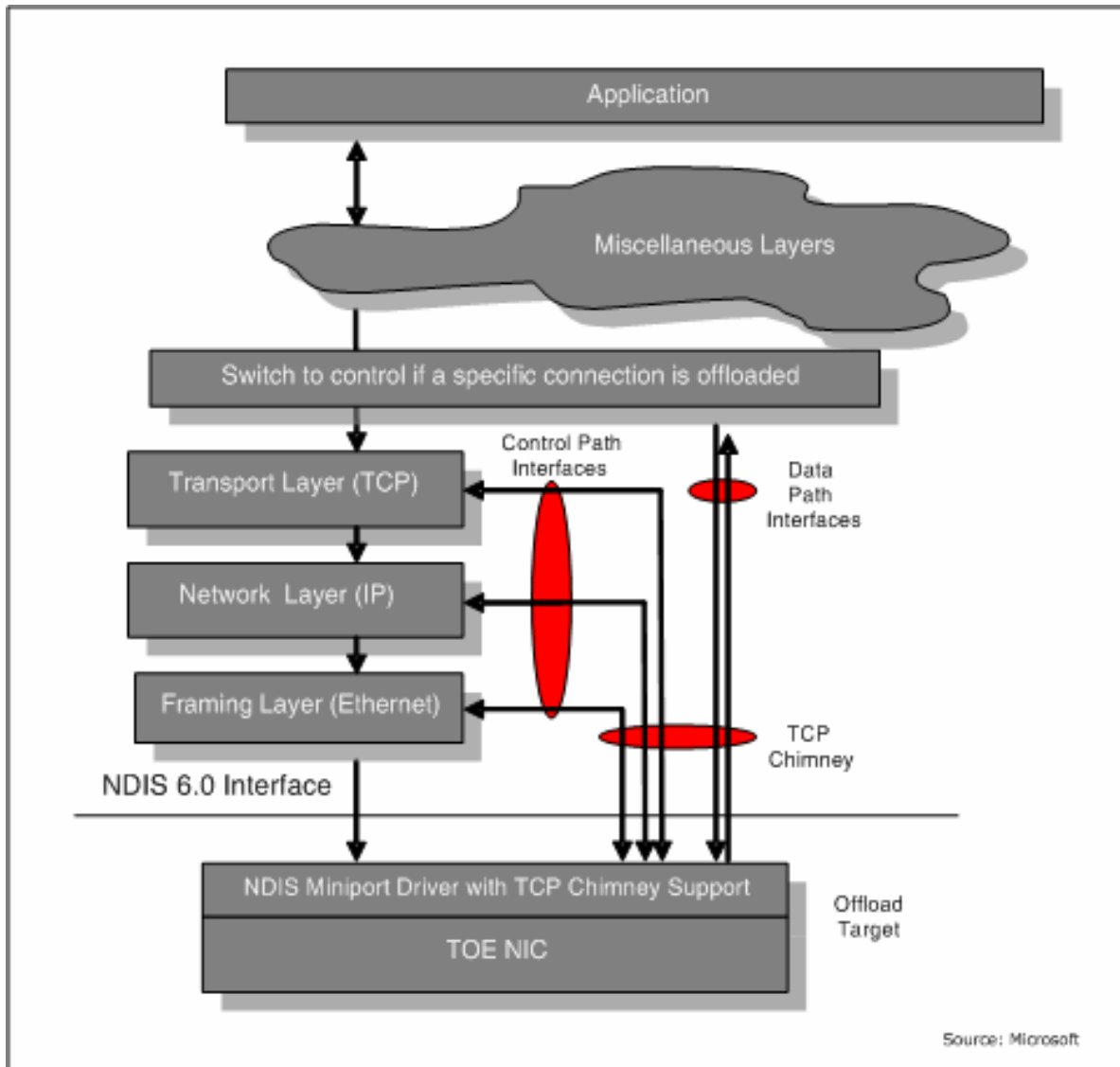
Another challenge is that early TOE implementations used dedicated memory on NICs for data buffering and, in some cases, context buffering. The result is a large subsystem that is more expensive and power hungry. This approach cannot be used for higher speeds as the size of the required memory scales with the network speed. The real estate required for a TOE with external memory makes it harder to offer it as a standard LAN connection for the server.

## **Windows TCP Chimney Offload Architecture Overview**

Microsoft's new TCP Chimney offload architecture enables offload of the TCP protocol stack to a TOE-based NIC. This enables the full integration of TOE capabilities within the Windows operating environment, while preserving customer investments in applications, manageability, robustness, and security. Windows applications that are currently bound by network processing overhead enjoy significant performance improvements, and scale better when used with TCP Chimney. The result is a reduction in the overhead of network protocol processing for the server CPU. The TCP Chimney offload architecture is slated to ship this year as part of the recently announced Scalable Networking Pack (SNP) for Microsoft Windows, which is named Longhorn.

The Microsoft Chimney offload architecture departs significantly from the all-inclusive first-generation TOE approaches and provides seamless integration of TOE capabilities within Windows. The TCP Chimney specifically offloads the performance-sensitive TCP/IP data path to the offload target TOE NIC (moving data between two peers), while maintaining the control path operations for setup and teardown of accelerated TCP/IP connections and support of ancillary protocols, such as DHCP, RIP IGMP, and ARP, within the Windows TCP/IP stack.

The following figure is a diagram of the Microsoft TCP Chimney block.



## Benefits of the Chimney Approach of TCP Offload

The first benefit of the Chimney approach of TCP offload is the use of the Windows TCP/IP stack for supporting connection setup/teardown and support of ancillary protocols, which enables a simplified implementation and faster deployment model. The benefits of Chimney-based TCP/IP acceleration are transparently available to all Windows-based network applications for all applications using the standard WinSock interface (Microsoft's implementation of the popular Sockets API).

The second benefit of the Chimney approach of TCP offload is that the support of connection setup by the Windows protocol stack addresses the security vulnerabilities posed by offload hardware handling the connection setup.

Windows TCP Chimney is ideally suited to enabling TOE to become a standard function for volume server GbE ports. Broadcom's industry-leading NetXtreme II C-NIC delivers TCP protocol offload designed to match the Chimney TCP offload architecture to provide acceleration for GbE-based data as well as for clustering and storage traffic.

## NetXtreme II Value Proposition Summary

NetXtreme II C-NICs allow data center administrators to maximize the value of available server resources. NetXtreme II C-NICs also allow servers to share GbE network ports for all types of GbE TCP/IP traffic, while removing network overhead and simplifying existing network cabling, and facilitate infusion of server and network technology upgrades.

### Benefits of the NetXtreme II C-NICs

The following sections summarize the benefits provided by NetXtreme II C-NICs.

#### *Increased Server and Network Performance*

Compared to existing Gigabit NICs, NetXtreme II C-NICs significantly reduce the overhead of network I/O processing from the server. Aggregation of networking, storage, and clustering I/O offload into the C-NIC function improves cost-effectiveness versus single-purpose networks by leveraging the existing Ethernet infrastructure. Internal Broadcom benchmarking (using the industry popular NTTTCP benchmark) has shown that NetXtreme II C-NICs reduce server CPU utilization by up to five times, dramatically improving server efficiency.

#### *Lower TCO for Data Center and Simplified Server Additions and Data Center Network Upgrades*

Because data networking, clustering, and storage functions have been aggregated, newly added NetXtreme II C-NIC-equipped servers (such as blade servers) only need to be connected to a GbE connection. Conversely, upgrading to a higher speed (such as 10 GbE) only requires the replacement of one NIC type versus multiple dedicated NICs and HBAs. Low latency clustering addresses the needs of database and other latency-sensitive applications.

### *Full Networked Storage Solution*

NetXtreme II C-NICs provide high performance and low CPU utilization for file (i.e., CIFS and NFS), as well as block (iSCSI) storage, which eliminates the need for a dedicated SAN with specialized equipment (such as Fibre Channel).

### *Improved Data Center Operational Efficiency*

By using NetXtreme II C-NICs, the interfaces to each server and rack are simplified. There are fewer connection points, cables, adapter cards, and easier upgrades to existing networks. Changes are localized to the NetXtreme II C-NIC. Fewer changes translate to improved efficiency. In addition, more productive servers can result in fewer servers for some data centers, which reduce acquisition and ongoing maintenance and management expenses.

Broadcom®, the pulse logo, Connecting everything®, the Connecting everything logo, and NetXtreme® are among the trademarks of Broadcom Corporation and/or its affiliates in the United States, certain other countries and/or the EU. Any other trademarks or trade names mentioned are the property of their respective owners.

Connecting  
everything®



BROADCOM CORPORATION  
16215 Alton Parkway, P.O. Box 57013  
Irvine, California 92619-7013  
© 2005 by BROADCOM CORPORATION. All rights reserved.

Phone: 949-450-8700  
Fax: 949-450-8710  
E-mail: [info@broadcom.com](mailto:info@broadcom.com)  
Web: [www.broadcom.com](http://www.broadcom.com)