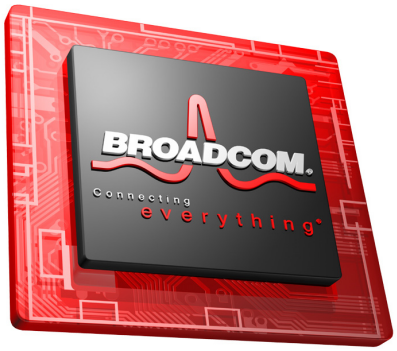
White Paper

From Ethernet Ubiquity to Ethernet Convergence: The Emergence of the Converged Network Interface Controller (C-NIC)

The focus of this paper is on the emergence of the *converged* network interface controller (C-NIC) providing accelerated client/server, clustering, and storage networking and enabling the use of unified TCP/IP Ethernet communications. The breadth and importance of server applications that can benefit from C-NIC capabilities, together with the emergence of server operating systems interfaces enabling highly integrated network acceleration capabilities, promise to make C-NICs a standard feature of volume server configurations.

C-NIC deployment will provide dramatically improved application performance, scalability and server cost-of-ownership. The unified Ethernet network architecture enabled by C-NIC will be non-disruptive to existing networking and server infrastructure, while providing significantly better performance at reduced cost vis-à-vis alternatives.

March 2004



It is widely recognized that the single greatest factor impacting data center application performance and scalability is the server network I/O bottleneck. The fact that network bandwidth and traffic loads for client/server, clustering and storage traffic have outpaced and will continue to consistently outpace CPU performance increases results in a significant and growing mismatch of capabilities that has become one of the data center manager's largest problems.

A common solution to this challenge has been to use different networking technologies optimized for specific server traffic types, for example, Ethernet for client/server communications and file-based storage, Fibre Channel for block-based storage, and special purpose low-latency protocols for server clustering. However, such an approach has acquisition and operational cost difficulties, may be disruptive to existing applications, and inhibits migration to newer server system topologies, such as blade servers and virtualized server systems.

Enter Unified Ethernet Communications

An emerging approach is focused on evolving ubiquitous Gigabit Ethernet TCP/IP networking to address the requirements of client/server, clustering, and storage communications through deployment of a *unified Ethernet communications fabric*. The vision of such network architecture is that it is non-disruptive to existing data center infrastructure and provides significantly better performance at a fraction of the cost – all the while preserving the existing investment in server and network infrastructure. Upgrading of data centers will be enabled in an evolutionary fashion on a rack-by-rack basis using existing management tools and interfaces. Additionally, the approach promises *no* modifications to existing applications.

At the root of the emergence of unified Ethernet data center communications is the coming together of three networking technology trends: *TCP Offload Engine (TOE), remote direct memory access (RDMA) over TCP, and iSCSI.*

TOE refers to the TCP/IP protocol stack being offloaded to a dedicated controller in order to reduce TCP/IP processing overhead in servers equipped with standard Gigabit network interface controllers (NICs). While TOE technology has been the focus of significant vendor engineering investment, a number of obstacles remain for broad-based TOE deployment.

The first generation of TOE products has suffered from two related drawbacks: lack of standard interfaces within server operating systems for TOE integration and an all-inclusive approach to architecting TCP offload as a parallel server network protocol stack. The recent emergence of Microsoft's Chimney Offload Architecture for enabling TOE integration within Windows server operating systems will spur the development of TOE products that are tightly coupled with server operating systems for optimum server performance gains and lower cost of ownership.

An additional disadvantage of existing TOE products, which typically are an embedded component of TOE network interface cards (T-NICs), is the expense. Each one of these T-NICs is needed in each server, and two are needed for high-

availability configurations. Costs are likely to be high, approximately $700 to $1000 for a T-NIC.

RDMA is a technology that allows the network interface controller (NIC), under the control of the application, to place data directly into and out of application memory, removing the need for data copying and enabling support for low-latency communications, such as clustering and storage communications.

RDMA networking has until now consisted of specialized interconnects which represent an expensive and (for existing data centers) disruptive alternative due to their proprietary and monolithic architectures. In essence, these require a special RDMA-enabled NIC with a proprietary interface to a high-speed, proprietary fabric. For organizations with a large investment in standard servers and Gigabit networking and various commercial and custom software applications, or for organizations with tight budgets, this is an unattractive option because it requires abandoning existing infrastructure rather than building upon it.

A range of networking industry leaders, including Broadcom, Cisco, Dell, EMC, HP, IBM, Microsoft, and NetApp, has come together to support the development of an *RDMA over TCP* protocol standard and provides facilities immediately useful to existing and future Gigabit TCP/IP-based clustering, storage, and other application protocols.

RDMA APIs including *Sockets Direct Protocol* (SDP) and *Winsock Direct* (WSD) give the base of existing TCP/IP sockets applications access to the benefits of RDMA over TCP, while RDMA-native APIs, which allow applications to take maximum advantage of RDMA over TCP, have also been developed. These include the *Direct Access Programming Library* (DAPL) developed by the DAT Collaborative and incorporated into the standardization efforts of the Interconnect Standard Consortium (ISC) of the Open Group, as well as the emerging Windows Server *Named Buffers* API.

iSCSI is designed to enable end-to-end block storage networking over TCP/IP Gigabit networks. It is a transport protocol for SCSI that 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.

Application servers have two available options for supporting iSCSI initiators: a software-only initiator, (such as Microsoft's iSCSI software driver) with a standard Gigabit Ethernet NIC or an iSCSI host bus adapter (HBA). iSCSI HBAs typically use dedicated controllers to fully offload iSCSI and TCP/IP protocols from server host CPU for optimum performance and leverage storage services, such as multi-path I/O provided by the server operating system.

Unleashing Data Center Application Performance Through IP Protocol Suite Offload

The specific performance benefits provided by TCP/IP, RDMA, and iSCSI offload depend upon two broad application-specific factors: the nature of application network I/O and location of its performance bottlenecks. Among networking I/O characteristics, average transaction size and throughput and latency sensitivity play an important role in determining the value TOE and RDMA bring to a specific application. The degree to which an application is storage network-bound is a critical determinant in the benefit that iSCSI block storage networking brings to it.

Margalla Communications analysis of workloads running on "scale-out" 1P/2P servers shows that IP protocol suite offload benefits span the range of applications running within today's data centers. As summarized in *Figure 1*, IP protocol suite offload is especially beneficial for front-end web server applications, back-end high-performance computing and decision support applications, and Common Internet File Service (CIFS) file storage services. These applications taken together comprise a major portion of the overall server market, both in terms of shipments and revenue, making IP protocol suite offload the industry's first real and complete solution to the server network I/O problem that can be introduced in a seamless and non-disruptive manner.

Server Application	**TCP/IP Offload**	**RDMA over TCP Offload**	**ISCSI Offload**	**Comments**
Common Internet File Services (CIFS)	√		√	Persistent, high-bandwidth client to file server connections – Back-end iSCSI storage enables file server consolidation.
Web Services	√	√		TOE/RDMA benefit scales with Web server capacity.
High-Performance Computing (HPC)	√	√	√	RDMA improves HPC cluster scalability – Biosciences and earth sciences applications are network I/O bound.
Decision Support Services	√	√	√	Decision support client-to-server and server-to-storage communications is typically network I/O bound – RDMA impacts server cluster scalability.

Fig. 1: IP Protocol Suite Offload Value by Server Workload
Source: Volume Server Workload Analysis, Margalla Communications, 2003

The breadth and importance of server applications that can benefit from IP suite offload, together with the emergence of server operating systems interfaces enabling highly integrated offload capabilities, promise to make IP protocol suite offload a standard feature of volume server configurations. This promise can only be fulfilled through the emergence of a new converged network interface controller (C-NIC) product category which goes beyond existing monolithic offload approaches

to provide unified TOE, RDMA, and iSCSI offload and dramatically improved application performance, scalability, and server cost-of-ownership.

Converged Network Interface Controller (C-NIC) – The Critical Enabler for Unified Ethernet Communications

The main focus of emerging C-NIC products will be on integrating each of the main hardware and software components of IP protocol suite offload into a unified whole. Specifically, an integrated C-NIC device will comprise the following specific capabilities:

- TCP protocol offload tightly coupled with server operating systems using interfaces such as Microsoft Chimney
- RDMA over TCP offload supported via RDMA interfaces such as Sockets Direct Protocol (SDP) or WinSock Direct (WSD)
- iSCSI offload which is tightly coupled with server operating system storage stack.
- Embedded server management capability with support for OS present and absent states – enabling complete remote systems management for rack, blade, and virtualized server systems

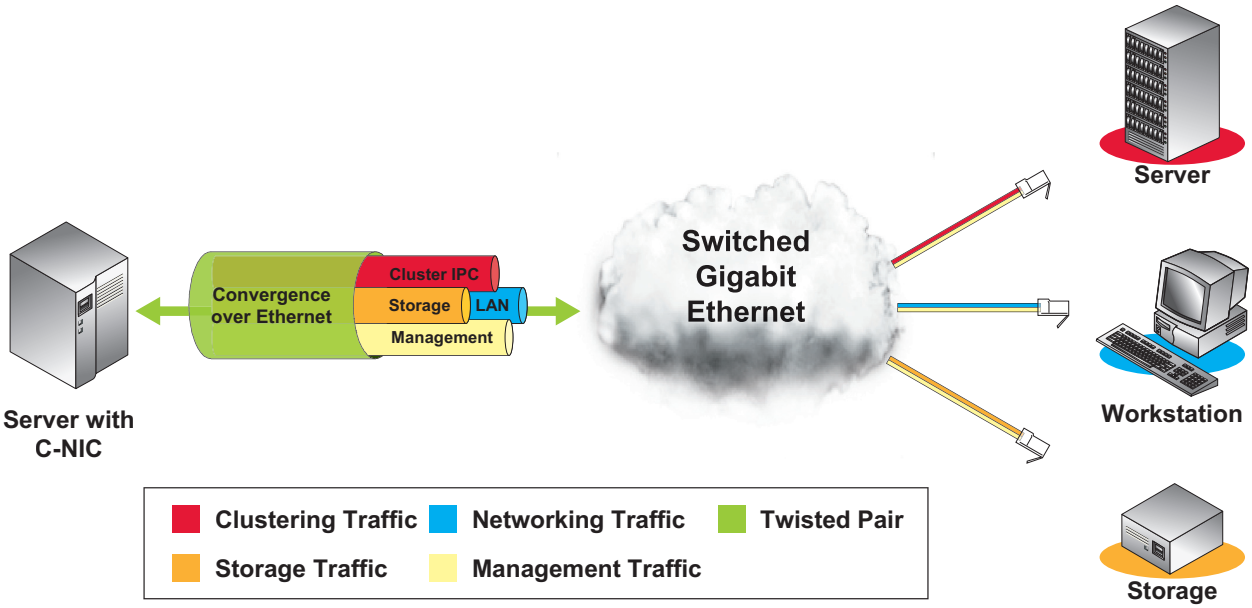


Fig. 2: C-NICs Concurrently Accelerate Ethernet TCP/IP-based Networking, Clustering, and Storage Traffic.

The requirement for seamless integration of the multiple facets of C-NIC functionality with server operating system environments is a critical product requirement; 'out-of-the-box' acceleration of existing and new applications using standard TCP/IP and SCSI interfaces will be a prerequisite for C-NIC deployment.

A single integrated C-NIC device with associated software will support accelerated performance for general data networking, clustering, and storage communication. As shown in Figure 2, C-NIC-equipped servers will have the capability to concurrently support the three different traffic types across one or multiple server Gigabit ports, truly enabling unified end-to-end Gigabit Ethernet-based communications.

C-NIC Value Proposition Summary

C-NICs will allow data center administrators to maximize the value of available server resources. It will allow servers to share GbE network ports for all types of traffic, remove network overhead, simplify existing network cabling, and facilitate infusion of server and network technology upgrades. These benefits are provided with no changes to application software and while maintaining the existing user and software management interfaces.

In summary, C-NICs provide the following important benefits:

- **Increased server and network performance.** Compared to existing Gigabit NICs, C-NICs allow full overhead of network I/O processing to be removed from the server. In addition, aggregation of networking, storage, and clustering I/O offload into the C-NIC function removes network overhead and significantly increases effective network I/O, while improving cost-effectiveness versus alternative discrete offload approaches.

- **Simplified server additions and data center network upgrades.** Since network I/O functions have been aggregated, newly added C-NIC-equipped servers, such as blade servers, only need to be connected to a single Gigabit Ethernet connection. Conversely, upgrading to 10 GbE only requires the addition of a module to the server.

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

Connecting everything®

BROADCOM®

BROADCOM CORPORATION
16215 Alton Parkway, P.O. Box 57013
Irvine, California 92619-7013

NGST-WP100-R 09.03.2004

Phone: 949-450-8700
Fax: 949-450-8710
E-mail: info@broadcom.com
Web: www.broadcom.com