

“Beyond Gigabit” Networking and Next-Generation Network-System I/O Standards

Comparative Positioning of InfiniBand® and RDMA over TCP

November, 2002



A Margalla Communications Special Report

About Margalla Communications

Margalla Communications provides the following strategic and technical marketing consulting services:

Custom Consulting

Margalla Communications houses vendor-independent networking domain experts who provide counsel on product planning and strategy, product positioning, market validation, target market selection, business development, custom market research, competitive positioning, demand creation, and other issues of concern to anyone in the networking industry. Margalla has consulted for vendors, end-users, and venture capitalists, and can be contracted on a retainer, daily or project basis. Contact info@margallacomm.com for further details.

Technology White Papers

Margalla Communications can provide technology and marketing white papers to suit a variety of enterprise needs. Contact info@margallacomm.com for further details.

About the Author

Saqib Jang is Principal at Margalla Communications, an Atherton, CA-based firm providing strategic and technical marketing consulting services to storage and media networking, and network security markets. He has more than three years of experience as a successful high-technology marketing consultant and industry analyst. Prior to independent consulting, Saqib held senior management and marketing roles with several public and private high-technology companies, including holding the positions of CEO/Co-founder of a startup developing provider-grade systems for IP videoconferencing services delivery, and Director of Marketing at NEC Systems (the system integration and software arm of NEC Corporation). Previously, Saqib was responsible for software product management and marketing at Auspex Systems, including Auspex’s award-winning entry into the CIFS market, and spent over 7 years at Sun Microsystems/SunSoft in a range of executive marketing roles developing and implementing product marketing strategies for industry-leading network security and storage networking products. Mr. Jang holds a B.S. degree in Electrical Engineering from Massachusetts Institute of Technology (MIT) and an M.B.A. in Marketing from the Wharton School, University of Pennsylvania. He can be reached at saqibj@margallacomm.com

COPYRIGHT NOTICE

Copyright © 2002 by Margalla Communications, Inc. All rights including that of translation into other languages are specifically reserved. Margalla Communications, 3301 El Camino Real, Suite 220, Atherton, CA 94027 (Tel: 650 274 8745; Fax 650 368 8198) www.margallacomm.com

InfiniBand is a registered trademark of the InfiniBand Trade Association. Other product and company names mentioned herein may be the trademarks of their respective owners.

NOTE: The material presented in this report is based on publicly available information coupled with our professional interpretation of the facts. We believe that the basic information and recommendations in this study provide a basis for sound business decisions, but no warranty as to completeness or accuracy is implied. All market estimates and forecasts are those of the author, except as noted. We welcome your comments on this report.

Executive Summary

Current network-system I/O architectures are constraining the effectiveness of rapid advances in CPU and networking technologies. This issue is becoming significant as the market deploys Gigabit Ethernet, and will be a critical issue as multi-gigabit networking begins to be widely adopted. To solve this problem at the root cause, Remote Direct Memory Access (RDMA)-enabled networking standards must be employed.

The InfiniBand Architecture (IBA) is an established RDMA-based 2.5-30 Gbps networking standard enabling switch fabric-based I/O communication for data centers. The vision was for IBA to be the next-generation system area network offering unconstrained scalability, low communications latency, and high availability for connecting multiple servers to both themselves and I/O devices. IBA has been developed by the InfiniBand Trade Association, which comprises a steering committee (Dell, Hewlett-Packard, IBM, Intel, Microsoft, Network Appliance, and Sun Microsystems) and over 200 sponsoring members.

Initial versions of 10Gb/s InfiniBand components are currently available with volume availability of components and switches slated for first-half of calendar '03. Volume availability of PCI-X InfiniBand 10Gbps server Host Channel Adapters (HCAs) is expected in early calendar '03, while higher-end, higher-performance Sun, IBM, and Dell server models are expected to start shipping with native 30 Gbps InfiniBand support in the 2004 timeframe.

The ubiquity of GbE TCP/IP networking, cost and operational efficiencies being a key imperative for the low-end/mid-range data center markets, the emergence of the IP storage market opportunity and other related factors are generating renewed interest in a TCP-based RDMA standard. The RDMA Consortium which was recently formed by leaders in the server, networking, and storage markets, including Adaptec, Broadcom, Cisco, EMC, HP, IBM, Intel, Microsoft Corp., and Network Appliance, has rapidly created specifications for the RDMA over TCP protocol set, and has submitted these in October, 2002 to the Internet Engineering Task Force (IETF) for standardization.

The stature of the vendors involved in driving the RDMA over TCP effort gives the emerging protocol set strong momentum and a number of vendors have already announced plans to ship first-generation RDMA over TCP products utilizing pre-standard implementations of the protocol set during 2003. An additional benefit furthering momentum is that the technology can be easily integrated under Sockets Direct APIs (originally developed to enable sockets-based applications to take advantage of InfiniBand and other RDMA transports) including Microsoft's Winsock Direct technology shipping on Win2K and .NET --giving existing user-mode binary applications access to the benefits of RDMA over TCP. Further, the deployment model for RDMA over TCP technologies is envisioned to be non-disruptive with the benefits of RDMA acceleration accruing proportionally by the number of connections/applications that require or use RDMA mode.

There are also a number of challenges regarding volume deployment of RDMA over TCP. First, the timing of volume deployment of RDMA over TCP NICs depends upon relatively synchronized availability of such products from a significant number of network interface hardware vendors. In addition, it will take operating system vendors implementing special network programming interfaces that export native RDMA semantics for optimal application use of RDMA over TCP. Finally, it will take some time for RDMA NICs to make it through the server delivery cycle and be ready for enterprise deployment.

Based on the above, we anticipate initial availability of 1GbE and 10GbE TOE/RDMA NICs by late 2003 and late 2004 respectively. Volume deployment in enterprise environments of RDMA-enabled 1/10 GbE network hardware is anticipated to follow initial availability by a year.

In our estimate, InfiniBand will be a higher bandwidth, lower latency and increased functionality RDMA fabric when compared to Gigabit Ethernet for the foreseeable future. From an application standpoint, InfiniBand will be used as a high-performance, low-latency I/O expansion network interconnecting IB-enabled servers, and native IB Network-Attached Storage (NAS) systems (e.g. Direct Access File System (DAFS) systems) within and among data center server tiers. This is consistent with IB gaining momentum among early adopters as the Inter-Process Communications (IPC) interconnect of choice for mid-range/high-end database and HPC clustering deployments. IB is expected to coexist with non-IB environments (e.g. FC for block storage and 1/10 GbE for NAS, data center backbone, and inter-data center networking) via shared I/O adapters. *From a market segment standpoint, IB is expected to gain traction in the middle to high-end enterprise and service provider data center environments having performance scalability as the driving requirement.*

The goal behind the RDMA over TCP standard is to enable the ubiquitous base of enterprise applications, servers, and network hardware to leverage the performance benefits of RDMA networking while enabling acquisition and operational efficiencies. We envision non-disruptive deployment of GbE RDMA over TCP networking in low-end/mid-range data center environments beginning in 2004 and ramping in 2005/2006 timeframe. Intra-data center IP-based file and block storage (i.e. iSCSI) and low-end/mid-range server blades are envisioned to be the early adopter applications for RDMA over TCP, with high-performance enterprise-wide TCP-based client/server communication and inter-data center networking applications following next. RDMA over TCP is expected to co-exist with Fibre Channel for block storage access and InfiniBand-based server networking (especially in high-end environments). *From a market segment standpoint, RDMA over TCP is expected to gain momentum in the low end-to-middle enterprise and service provider data center environments having acquisition and operational cost efficiencies as the driving requirement.*

Table of Contents

1.0	Introduction	6
2.0	TCP Receive Buffer Copy Problem and Remote Direct Memory Access (RDMA)	6
2.1	Remote Direct Memory Access (RDMA)	8
3.0	InfiniBand Overview	9
3.1	InfiniBand Status	10
3.2	InfiniBand and RDMA-native Protocols	11
3.3	InfiniBand and TCP Protocol Offload	11
4.0	RDMA over TCP	12
4.1	RDMA over TCP Status	14
5.0	Conclusions	15
	Appendix A – Acknowledgements	17
	Appendix B – References	18

1.0 Introduction

The capacity of applications running within enterprise or Internet data centers is limited by server CPU/memory architectures. To increase the capacity of these data center applications, either faster CPU/memory architectures need to be deployed or improved network-system I/O mechanisms need to be explored, or both. Among network-system I/O mechanisms, offload techniques work well for software modules that consume large portions of CPU or memory bandwidth and have standard interfaces. Networking protocol stacks, for example, are considered prime candidates for offloading.

Protocol stack overhead can directly affect the performance of server applications that are constrained by CPU or memory bandwidth. For example, 1GbE networking protocol stacks can consume as much as one third of the total CPU utilization for servers running database applications [WSD]. If that consumption were reduced to 5%, then additional CPU resources would be available to the database application. The I/O capacity available to the database application could potentially be 42% ($0.95/0.66$) greater. Another way of thinking of it is that an 8-processor system is transformed into a 12-processor system by adding a networking adapter that supports protocol offload.

First-generation TCP protocol offload algorithms provide checksum offload and large send offload. Many vendors are pursuing a second-generation offload mechanism (commonly referred to as TCP Offload Engine (or TOE)) centered on offloading the TCP transport stack to the GbE server network adapter. While studies show that such approaches can lead to improvements in server and application scalability under limited set of traffic assumptions¹, *there is a growing awareness that the emerging deployment of “beyond gigabit” networking in data centers will require a radically different approach to improving system-network I/O performance: requiring not only offloading the transport protocol stack, but also avoiding the work that occurs between the application and the protocol stack.* As we’ll see in later sections, these new approaches include Remote Direct Memory Access (RDMA) and RDMA transport application programming interfaces (APIs), such as Direct Access APIs (DAPL). The use of these techniques not only provides significantly superior system-network I/O characteristics for GbE networking but also is mandatory as data center networking evolves to 10 Gbps levels.

2.0 TCP Receive Buffer Copy Problem and Remote Direct Memory Access (RDMA)

This section details the critical source of TCP/IP protocol stack CPU and memory bandwidth utilization. While offload benefits for the TCP networking stack vary by workload and application, the TCP ‘receive copy’ problem is a key source of CPU and memory utilization across workloads for GbE networking and has the potential of becoming a “showstopper” inhibiting migration to 10Gbps networking.

¹ Existing TOE algorithms for 1GbE networking work well for applications that are posting TCP message sizes greater than 16 KB but almost no difference in CPU utilization is seen when comparing these TOE approaches to pure TCP/IP software stacks for 4-8KB message sizes. [ALACR]

It is typical for a TCP/IP host to copy received data after it arrives in host memory. This copying incurs CPU, memory and bus costs that are substantial and are not masked by advancing hardware technology, including TOE technology. While there is a long history of research and experimental schemes to reduce or eliminate receiver copying overhead for IP networking in general, and for TCP/IP communication in particular, the growing importance of data centers serving applications to growing numbers of TCP/IP-based clients and the opportunity to extend the benefits of IP technologies to high-performance application domains, such as file and block-based storage applications, is driving the need to address this challenge through a generalized, commercially viable solution.

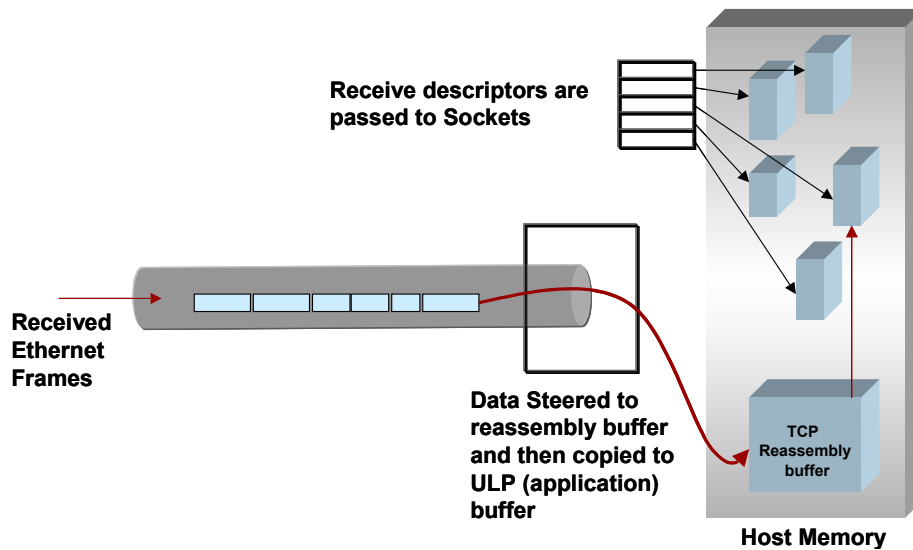


Fig 1. TCP Buffer Copy Problem

Packet loss, and attendant out-of-order reception is a frequent, continuous characteristic of both wide-area, and switched local area networks of almost any size, as TCP adjusts to varying congestion conditions. For example, [TCP] shows that segment reordering in TCP/IP networks is not a rare event and eliminating reordering is a difficult problem. In addition, a range of TCP-based application protocols provide transport-like features such as control and data multiplexing, layered above TCP (e.g., iSCSI). In many such protocols control information contained in the protocol uniquely identifies the destination application buffer of each particular piece of data. If TCP segments are not received in sequence order it may not be possible to unambiguously locate the protocol control information. For example, if the upper-level Protocol Data Unit (ULPDU) length information is in a TCP segment that is delayed or lost in transmission, assuming the ULPDU length is the only means of locating the beginning of the subsequent ULPDU, and it is then impossible to locate ULP control information for ULPDUs in subsequent TCP segments until the lost or delayed TCP segment is received.

The traditional solution to this problem is to build a reassembly buffer into the server network interface as shown in Figure 1. Data received out-of-order can be held in the network interface reassembly buffer until all preceding data is received, and then direct copying to ULP memory can be performed on the reassembled data. Within certain

implementation assumptions, this is a reasonable approach, but, unfortunately there are two important issues that make the reassembly approach undesirable, especially for multi-gigabit networking environments.

First, the size of reassembly buffer needed in the network interface is a direct function of the bandwidth times delay product of all active TCP connections. Reasonable assumptions on the active bandwidth times delay product can imply a large amount of reassembly memory. For example, assuming 100 ms link delay for a worst-case link length of 20K km, a reassembly buffer size of 125MB would be required for 10 GbE TOE hardware.

Second, this large reassembly memory must run at high speed---more than two times the link speed, to maintain full link bandwidth. Performing reassembly in the network interface requires that the bandwidth from the network interface to host memory be not just equal, but substantially greater than the maximum bandwidth of the network link, to ensure that the reassembly buffer is drained when reassembly is complete. System bus and interconnect bandwidth are particularly scarce and expensive resources in most systems.

Link	Usable Link Rate one direction	Memory Bandwidth For Buffer Copy	CPU's required at 200 MB/s Buffer Copy rate	Total Memory Bandwidth with DMA and Buffer Copy
Gig-Ethernet	125 MB/s	375 MB/s	1.9	500 MB/s
10 Gig-Ethernet	1250 MB/s	3750MB/s	18.8	5000 MB/s

Fig. 2: CPU and Memory Bandwidth Requirements by Link Rate [RDMA]

Figure 2 shows that for GbE network fabrics, memory architectures can barely keep up, and leave little extra bandwidth for the application. As an example, assume that a single CPU/memory combination can maintain a 200 MB/s buffer Copy rate (which is aggressive for today’s high performance systems). In this case, it is difficult for the system to have extra bandwidth for the application while also managing a single gigabit-per-second link. At 10 GbE rates, the problem is essentially not solved—regardless of what the application is trying to do or the protocol offload methodology used. Consuming 19-20 CPUs just to drive the network interface is not an effective solution. Further, 10GbE server networking requires 40Gb/s memory bandwidth or 3 DDR SDRAM banks. An interesting point to note is that a 10GbE adapter can quite effectively serve a 16-32 node server cluster as a shared network adapter (which was part of the rationale for the InfiniBand (IB) model of shared network I/O, as discussed in later sections).

2.1 Remote Direct Memory Access (RDMA)

Because of high protocol processing overhead including end-host overhead due to receiver-side copying, the TCP/IP protocol stack has typically not been used for high-speed low-latency data transfer. Instead, special purpose network fabrics using remote direct memory access (RDMA) have been developed and are widely used. RDMA is a

technology that allows the network adapter, under control of the application, to place data directly into and out of application buffers to address the receive-side copy problem. This capability is also referred to as "direct data placement". The primary examples of standard interconnection fabrics with defined RDMA semantics is the InfiniBand Architecture [IB] developed as an I/O expansion network and the emerging RDMA over TCP [RDMAC] standard being developed to provide RDMA capabilities for GbE/TCP networking, while a range of proprietary RDMA interconnects have been deployed including Compaq Servernet for System Area Networks. In addition, a range of application-specific RDMA-based application-programming interface (API) standards have been developed including SCSI RDMA Protocol [SRP] for block storage transfer, Sockets Direct Protocol [SDP] for general data networking, and Virtual Interface Architecture [VI] for database clusters.

3.0 InfiniBand Overview

The InfiniBand Architecture [IB] is an established RDMA-based 2.5-30 Gbps networking standard enabling switch fabric-based I/O communication for data centers. The architecture uses point-to-point communications channels that scale from 500MB/s (“1X”) to 6 GB/s (“12X”) per device. It offers unconstrained scalability, low communications latency and high availability through redundant connections. The InfiniBand architecture has been developed by the InfiniBand Trade Association, which comprises a steering committee (Dell, Hewlett-Packard, IBM, Intel, Microsoft, Network Appliance, and Sun Microsystems) and over 200 sponsoring members. More information, including the InfiniBand Architecture 1.0 specification, can be found at the Trade Association Website: www.infinibandta.org.

The vision was for InfiniBand to be the next-generation system area network that connected multiple servers to both themselves and I/O devices. But aside from connecting servers within data centers, InfiniBand is also being viewed as enabling a new data center paradigm-- permitting power, Ethernet, Fibre Channel, and other functionality to be taken out of the server box and shared by many more rack mounted smaller form-factor servers, called "server blades," than presently fit in a typical rack. PCI, PCI-X, or PCI –Express (formerly known as 3GIO) would function internal to each server connecting the chips on the motherboard and providing master/slave I/O interconnect capability, with InfiniBand serving as the overall peer-to-peer interconnect. In such a model, all the server "blades" can be plugged or swapped at will using InfiniBand interconnects, along with sharing power and Ethernet. This allows much higher server density in the data center, where space is precious.

InfiniBand architecture has many built-in features that ensure every server on the switched communications fabric can communicate concurrently, with high bandwidth and low latency, making it a highly scalable I/O architecture. The InfiniBand architecture assures low-latency, high-bandwidth communications through:

- Increased bandwidth -- A single InfiniBand link supports 2.5 Gbps in each direction; since the links can be bundled into groups of four or 12, data throughput improves to as high as 30 Gbps

- RDMA capabilities
- Reliability features built into the transport

Clustering architectures provide an opportunity for growth with a minimum of application disruption, because multiple servers act as a single system and are managed with the same tools. It also provides exceptional scalability and failover, and management capabilities. InfiniBand stands out in providing the mechanisms necessary to support the demanding requirements of clustering. Using the InfiniBand fabric as the high-bandwidth, low-latency cluster inter-process communications (IPC) interconnect boosts cluster performance and scalability, and improves application response times.

3.1 InfiniBand Status

Initial versions of 4X 10Gb/s InfiniBand components are currently available with volume availability of components and switches slated for first-half of calendar '03. Volume availability of PCI-X InfiniBand 10Gbps Host Channel Adapters (HCAs) is expected in early calendar '03, while higher-end, higher-performance Sun, IBM, and Dell server models are expected to start shipping with native 30 Gbps InfiniBand support in the 2004 timeframe.

Negative publicity recently resulted from Intel [INTEL-IB] and Microsoft [MS-IB] (both leading InfiniBand proponents) recently pulling back from their level of investment and commitment to InfiniBand. Specifically, Intel recently announced that it has stopped investing in the development of InfiniBand silicon, while continuing to invest in IB enablement and software. Microsoft pulled back from its announced intention to implement IB drivers and IB subnet management in the Beta release of .Net

Beyond the PR impact, both these announcements are a reflection that the data center networking industry is converging on the original positioning for InfiniBand as a high-throughput, low-latency I/O expansion network. This is consistent with IB early adopters consisting of high-end data center environment requiring performance scalability for clustered high-performance computing (HPC) and database applications. The inflated expectations for the technology (as a wholesale replacement for Ethernet and Fibre Channel as the all-inclusive data center fabric) came about during the dot.com bubble when everyone expected the market to be building out many new “Greenfield” data centers. The clear and sustainable throughput and performance advantages of IB vis-à-vis other networking fabrics continue to provide it with a compelling value proposition as the intra and inter-server tier interconnect for mid-range and high-end data center environments having performance scalability as the driving requirement.

Intel’s decision regarding its InfiniBand silicon program is widely thought to be due to it being behind IBM and Mellanox in development of 4X 10Gbps IB silicon. In addition, Intel plans to continue to invest in InfiniBand software and support within high-end blade servers. Microsoft’s decision is at a high-level consistent with the way Microsoft has treated other emerging I/O technologies such as Fibre Channel and in our estimate will have manageable consequences for the deployment of Infiniband-based Windows servers. As with Fibre Channel, InfiniBand vendors can offer server-based add-on

InfiniBand drivers for their products. Microsoft has committed to certifying the compatibility of vendors’ InfiniBand drivers with its products. In a nutshell, the case for InfiniBand as an I/O expansion network interconnecting servers and server tiers and enabling sharing of I/O interfaces to non-IB networks remains strong for performance-intensive mid-range and high-end data center environments.

3.2 InfiniBand and RDMA-native Protocols

Applications today use a set of protocols for storage, inter-process communications, and networking - these include Fibre Channel Protocol (FCP) for Fibre Channel SANs, Network File System (NFS) and Common Internet File System (CIFS) for NAS, sockets over Ethernet for IPC, and NDIS over Ethernet for general data networking. A number of these of these have been adapted to take advantage of InfiniBand and other RDMA transports (including the emerging RDMA over TCP transport). These include SRP for block SAN storage, sockets direct protocol (SDP) and remote NDIS (RNDIS. In addition, RDMA-native protocols and APIs are also available. These enable applications to take maximum advantage of InfiniBand, upcoming RDMA over TCP protocol set, and other RDMA transports. Direct Access API (DAPL) is an RDMA-native transport API (developed by the DAT Collaborative) [DAT] which is being incorporated into the standardization efforts of the Interconnect Standard Consortium (Open Group). It is being broadly adopted by the InfiniBand Host Channel Adapter (HCA) vendors as a transport API for upper-level protocols and applications, and by a number of application vendors (parallel databases are furthest ahead) to enable their applications to have no-overhead, wire-speed Inter-process communications.

One upper-level protocol, which uses DAPL as its transport layer, is DAFS (Direct Access File System) [DAFS]. This is a high-performance file access protocol derived from NFSv4, which was designed to take maximum advantage of RDMA transports, and is expected to gain significant traction as the shared file storage for InfiniBand-based server clusters (native NAS for InfiniBand). A number of vendors have recently demonstrated DAFS on InfiniBand prototypes working with a number of clustered database applications, including Oracle 9i and IBM DB2 EEE. Network Appliance has been shipping DAFS products using an earlier VI/IP interconnect since April ’02 and is expected to release DAFS products using InfiniBand interconnect during 2003.

3.3 InfiniBand and TCP Protocol Offload

Use of InfiniBand as a high-performance, low-latency cluster interconnect can also accelerate the performance of TCP/IP applications in performance-sensitive mid-range/high-end data center environments. Typically, data center server cluster configurations provide scientific or data base application services to TCP/IP clients (users workstations or other servers). Since there are currently no RDMA acceleration solutions available to these clients, cluster servers must typically handle heavy TCP/IP processing while also providing such application services. Using InfiniBand as an enabling technology, this TCP/IP processing load can be removed from the clustered servers by gateways that forward traffic between TCP/IP and InfiniBand networks. This is accomplished by combining TCP termination with translation into the appropriate InfiniBand protocol (e.g., SDP, SRP). Client TCP/IP connections can be terminated by

the gateway, and the application traffic extracted and sent on to the application servers via an RDMA-capable InfiniBand protocol (e.g., SDP or SRP). Application responses are obtained via RDMA from clustered servers and translated into TCP/IP for return to the client. By removing the TCP overhead from the clustered servers, service latency is reduced and higher client loads can be supported. This type of gateway function is not only useful for communications between clustered servers and their application clients, but also between clustered servers and NAS storage. In addition, by using the IB model of shared network I/O, the more expensive TOE implementations in 1/10GbE adapters can be effectively cost shared over a number of servers.

Data center server clustering applications typically use either Fibre Channel SAN or GbE/IP network-attached storage (NAS) systems for back-end storage, including database storage. The increased throughput and efficiency of communication between application server clusters and NAS systems that is provided by an InfiniBand-GbE gateway function is expected to make NAS storage much more competitive with the more expensive SAN storage subsystems. For example, offload of TCP/IP processing to the gateway enables high-performance access between the application server cluster and a GbE/IP-based NFS/CIFS server.

It is possible to architect this InfiniBand-GbE gateway function so that it provides a value proposition that is economically compelling, even in today’s economy in which the market is reluctant to invest in new infrastructure. This requires combining the InfiniBand-to-InfiniBand IPC capability with the gateway function, and sharing this combined functionality across multiple servers simultaneously. This shared combination of InfiniBand switch and InfiniBand-GbE gateway capabilities could be considered to be the first stage in deployment of RDMA capabilities to address the data center server system-network I/O bottleneck problem. Assuming the RDMA/IP efforts result in viable products near the end of 2004, the next stage in addressing this problem would be adding an InfiniBand-RDMA/IP gateway function to the InfiniBand-GbE gateway function. This would, in effect, provide end-to-end RDMA capability for an entire enterprise.

4.0 RDMA over TCP [RDDP]

The ubiquity of GbE TCP/IP networking, cost and operational efficiencies being a key imperative for the low-end/mid-range data center markets, the emergence of the IP storage market opportunity and other related factors are generating renewed interest in TCP-based solutions to address the buffer copy problem. While emerging TCP Offload Engine (TOE) NICs offload TCP protocol processing up through the transport layer, they cannot address the memory bandwidth problem caused by receive-side copying because the information needed to place the payload directly into application memory is not known to the transport layer. While the problem can be solved one application protocol at a time by implementing the upper level protocol in the NIC (this is being done now by several vendors for iSCSI [iSCSI]), there are so many TCP-based applications (for example, RDBMS, HTTP, NFS, and remote procedure call (RPC)), affected by the memory bandwidth problem that migrating application protocol implementations into the NIC is economically infeasible. Neither a multitude of specialized NICs each

implementing one application protocol, nor a large, complex, expensive, multipurpose NIC implementing many application protocols is attractive to either vendors or end users.

Direct Data Placement (DDP) and *Remote Direct Memory Access (RDMA)* are members of a new IETF protocol family to be called “RDMA over TCP” being developed as generalized TCP-based solutions to the buffer copy problem. The goal of the RDMA protocol is to provide the semantics to enable Remote Direct Memory Access between RDMA over TCP peers in a way consistent with application requirements. The RDMA protocol is not an application protocol in itself, but provides facilities immediately useful to existing and future networking, storage, and other application protocols.

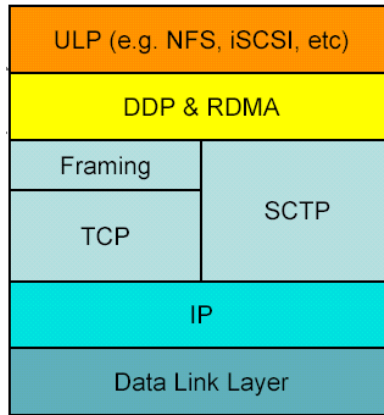


Fig 3. RDMA Over IP Layering [RDMA]

The DDP protocol is being developed within IETF as a standard solution to the problem of receive-side copying of network payload. Just as a network layer protocol such as IP can be thought of as steering data from a source node to a destination node, so DDP steers data from a source buffer to a destination buffer. A protocol stack residing in a NIC and containing all layers up through DDP/RDMA can place incoming payloads directly in the application’s buffer with only one memory bus crossing and can do so for any application.

The DDP and RDMA protocols work together to achieve their respective goals. RDMA provides facilities to an application protocol for identifying buffers, controlling the transfer of data between application peers, and providing completion notifications to the application protocol. RDMA uses the features of DDP to steer payloads to specific buffers at the Data Sink. ULPs that do not require the features of RDMA may be layered directly on top of DDP. Figure 3 shows the relationship between RDMA, DDP, Upper Layer Protocols (ULPs) and Transport

DDP and RDMA protocols are transport independent and direct data placement can in theory be done with either of IP’s reliable transports, the widely used TCP protocol and the Stream Control Transmission Protocol (SCTP).

Direct data placement for TCP-based application protocols requires the use of a *framing* mechanism for the TCP transport protocol. Framing refers to the ability to locate the

Upper Layer Protocol Data Unit (ULPDU) boundary by a hardware network adapter that uses DDP to directly place the data in the application buffer based on the control information carried in the ULPDU header. This may be done without requiring that the packets arrive in order. A major benefit of this capability is the avoidance of the memory copy overhead. IETF’s Transport Area Working Group [TSVWG] has recently begun work on a TCP framing protocol defined as a shim layer protocol between an Upper Layer Protocol (such as RDDP) and TCP.

4.1 RDMA over TCP Status

The RDMA Consortium was formed by leaders in the server, networking, and storage markets, including Adaptec, Broadcom, Cisco, EMC, HP, IBM, Intel, Microsoft Corp., and Network Appliance, with a goal of rapidly creating RDMA, DDP, and TCP framing specifications, and then submitting these to the IETF as Internet Drafts for standardization. These specifications are complete², and have been submitted to the IETF.

The stature of the vendors involved in driving the RDMA over TCP effort gives the emerging protocol set strong momentum towards standardization and deployment. A number of software and network hardware vendors have already announced plans to ship first-generation products utilizing pre-IETF standard implementations of the RDMA over TCP protocol set during 2003. An additional benefit furthering momentum is that the technology can be easily integrated under Sockets Direct [SDP] APIs (originally developed to enable sockets applications to take advantage InfiniBand) including Microsoft’s Winsock Direct technology shipping on Win2K and .NET – giving existing user-mode binaries access to the benefits of RDMA over TCP. The deployment model for RDMA over TCP technologies is envisioned to be non-disruptive and will most likely involve end-to-end discovery/negotiation of RDMA capabilities (i.e. use of RDMA if available or fall back to TCP sockets/TOE mode if not).

There are also a number of challenges regarding volume deployment of RDMA over TCP products and technologies. First, the timing of volume deployment of RDMA over TCP network interface cards depends upon relatively synchronized availability of such products from a significant number of vendors. In addition, while legacy applications using TCP sockets-based network interface may see out-the-box performance advantages through use of APIs such as Sockets Direct Protocol [SDP], it will take operating system vendors implementing special network programming interfaces that export native RDMA semantics (e.g. the Direct Access Transport [DAT] API) for optimal application use of RDMA over TCP.

Finally, it will take some time for RDMA NICs to make it through the server delivery cycle and be ready for enterprise deployment. RDMA NIC vendors will need to create the new NICs, operating system vendors will need to develop the code needed to support the RDMA NICs, and then O/S or server vendors will need to perform the testing required to validate the RDMA NIC functionality. It sometimes takes more than one pass through this cycle before the hardware has the necessary enterprise class functions and those functions have been validated.

² <http://www.rdmaconsortium.org/home/PressReleaseOct30.pdf>

Based on the above, we anticipate initial availability of 1GbE and 10GbE TOE/RDMA NICs by late 2003 and late 2004 respectively. Volume deployment in enterprise environments of RDMA-enabled 1/10 GbE network hardware is anticipated to follow initial availability by a year.

5.0 Conclusions

In our estimate, InfiniBand will be a higher bandwidth, lower latency and increased functionality RDMA fabric when compared to Gigabit Ethernet for the foreseeable future. 10 GbE RDMA-enabled NICs will most likely start shipping in the 2004 timeframe, but will not be ready for enterprise deployment until the 2005/06 timeframe. In comparison, IB will be operating at 30 Gbps/s with increased functions (e.g. transport reliability capabilities). It is very likely that the next generation of IB will be available in the 2006/07 timeframe operating at 6 GB/s to 12 GB/s, whereas the post 10 Gbps/s generation of Ethernet will likely not be available in that same time period.

From an application standpoint, IB will be used as high-performance, low-latency I/O expansion network interconnecting IB-enabled servers, and native IB NAS storage systems (e.g. DAFS systems) within and among data center server tiers. This is consistent with IB gaining momentum among early adopters as the IPC interconnect of choice for mid-range/high-end database and HPC clustering deployments. IB is expected to coexist with non-IB environments (e.g. FC for block storage and 1/10 GbE for NAS, data center backbone, and inter-data center networking) via shared I/O adapters. *From a market segment standpoint, IB is expected to gain traction in the middle to high-end enterprise and service provider data center environments having performance scalability as the driving requirement.*

The goal behind the RDMA over TCP standard is to enable the ubiquitous base of enterprise applications, servers, and network hardware to leverage the performance benefits of RDMA networking while enabling acquisition and operational efficiencies. We envision non-disruptive deployment of GbE RDMA over TCP networking in low-end/mid-range data center environments beginning in 2004 and ramping in 2005/2006 timeframe. Intra-data center IP-based file and block storage (i.e. iSCSI) and low-end/mid-range server blades³ are envisioned to be the early adopter applications for RDMA over TCP, with high-performance enterprise-wide TCP-based client/server communication and inter-data center networking applications following next. RDMA over TCP is expected to co-exist with Fibre Channel for block storage access and InfiniBand-based server networking (especially in high-end environments). *From a market segment standpoint, RDMA over TCP is expected to gain momentum in the low end-to-middle enterprise and service provider data center environments having acquisition and operation cost efficiencies as the driving requirement.*

³ RDMA over TCP support in low-end/mid-range server blades will allow high performance disk access while removing the requirement of the local disk (thereby reducing cost, MTBF, and power requirements).

Appendix A – Acknowledgements

Many people participated in this study and review of the report. They include Harish Ghadia (Adaptec), Renato Recio and Greg Pfister (IBM), Robert Simcoe (InfiniSwitch), Jim Pinkerton (Microsoft), David Dale and Tom Talpey (NetApp), Steve Hauser (Paceline), and Ted Compton (RDMA Consortium), in addition to others who prefer to remain anonymous.

Appendix B – References

[WSD]

“Winsock Direct – The Value of System Area Networks” A Microsoft Corporation White Paper
<http://www.microsoft.com/windows2000/techinfo/howitworks/communications/winsock.asp>

[ALACR]

eTesting Labs: Alacritech® 1000x1 Single-Port Server and Storage Accelerator: Chariot 4.0 Performance Testing
http://www.veritest.com/clients/reports/alacritech/alac_chariot.pdf
See Figure 8 for a discussion of server CPU utilization by TCP file (IO) size.

[TCP]

[On Making TCP More Robust to Packet Reordering](#), Ethan Blanton and Mark Allman, IEEE/ACM Transactions on Networking, January 2002, <http://www.acm.org/sigcomm/ccr/archive/2002/jan02/ccr-200201-allman.pdf>

[IB]

InfiniBand Trade Association
<http://www.infinibandta.org>

[RDMAC]

RDMA Consortium
<http://www.rdmaconsortium.org>

[RDMA]

“The Case for RDMA”, RDMA Consortium, 5/02
http://www.rdmaconsortium.org/home/The_Case_for_RDMA020531.pdf

[SRP]

SCSI RDMA Protocol
<http://www.t10.org/drafts.htm#SRP>

[VI]

Virtual Interface Architecture
<http://www.vidf.org/info/04standards.html>

[SDP]

“Winsock Direct – The Value of System Area Networks”, A Microsoft White Paper
<http://www.microsoft.com/windows2000/techinfo/howitworks/communications/winsock.asp>
Sockets Direct Protocol V1.0 Specification
<ftp://download.intel.com/technology/infiniband/data/IBAS83.htm>

[INTEL-IB]

http://www.byteandswitch.com/document.asp?doc_id=16572
http://www.byteandswitch.com/document.asp?doc_id=23233

[MS-IB]

http://www.byteandswitch.com/document.asp?doc_id=18800

[DAT]

Direct Access Transport
<http://www.datcollaborative.org>

[DAFS]

<http://www.dafscollaborative.org>

[RDDP]

<http://www.ietf.org/html.charters/rddp-charter.html>
<http://www.rdmaconsortium.org/home>

[ISCSI]

<http://www.ietf.org/html.charters/ips-charter.html>

[TSVWG]

<http://www.ietf.org/html.charters/tsvwg-charter.html>