InfiniBand and TCP in the Data Center

1.0 Preface

The InfiniBand Architecture is designed to allow streamlined operation of enterprise and internet data centers by creating a fabric that allows low latency, high bandwidth clustering, communication and storage traffic. The TCP protocol stack is ubiquitous as the Internet and is vital to the continued success of LANs, MANs, WANs and the Internet worldwide. While ideal for WAN and LAN applications, TCP imposes serious operating penalties when used within a controlled environment such as today’s Internet and enterprise data centers. This paper explores TCP limitations and discusses how they can be seamlessly overcome by exploiting the low latency hardware transport of InfiniBand to facilitate creation of N-Tier data centers.

2.0 What is an “N-Tier” Data Center?

Data centers are frequently described as having a ‘3-tier architecture’ with little or no explanation of what this really means. A 3-tier architecture is basically an extension of a standard client-server (two-tier) architecture which introduces a set of applications servers between the client and the backend database servers. This middle tier of applications servers provides the data center with the ability to track users and insures the reliability of transactions, even in the presence of equipment and software process failures. This basic 3-tier architecture can be extended to what is called an ‘n-tier’ structure by partitioning the infrastructure to separate independent components. This modularity provides better performance, security, scalability, and provides support for heterogeneous hardware and database environments. In addition, n-tier architectures are able to improve reliability, availability, and serviceability versus a simple two-tier architecture.

3.0 Transport Performance within the N-Tier Data Center

3.1 TCP Performance has not Kept Pace with Data Center Requirements

The benefits of an n-tier architecture do not come for free. First of all, each tier of the data center requires additional equipment, software, and creates complexity due to the larger number of units. The TCP protocol suite is designed to provide an end to end connection between the client and the ultimate server responding to client requests. However, inherent limitations in the IPV4 address space require that TCP is terminated by proxies running Network Address Translation (NAT) that results in multiple TCP connections being spliced together. Also, the modular nature of the N-Tier architecture spawns additional separate TCP/IP connections between each tier. Thus, what was originally implemented as a simple end to end connection between client and server now consists of many spliced TCP connections between client and ultimate endpoint.

Again, TCP is critical in accommodating the uncontrolled and heterogeneous nature of the Internet that connects the client to the data center. However, TCP is not optimized to provide a low latency fabric within the controlled environment of the data center and TCP imposes severe penalties on system cost and performance. These penalties include increased burden on CPU processing resources, increased latency, and inferior RAS capabilities.

These additional tiers necessitate the need for the data to be transported into and out of each system within the tiers, twice. That is, once moving from the first tier to the last and then back again, from the last tier to the point of origin, the first tier. A formula to calculate the total number of transports or hops can is easily derived: 2(n-1), where n is the number of tiers and 2 represents the data moving in and out of the tiers.

It is easy to see that as the numbers tiers in the data center expand, so does the latency. For example, a 7 tier data center would require 12 hops to complete the round trip. This illustrates that in today’s n-tier data centers the implementation of a fast and efficient connection among the many application servers, data bases and load balancers is vital because every hop in the n-tier costs latency.

3.2 Limitations of Software Transports

TCP implements the transport layer of the seven layer OSI model protocol stack which ensures a reliable, in-order connection oriented protocol, and connects between specific user processes running on separate machines.

TCP was designed as a software stack for LAN and WAN applications and was not designed to be implemented in silicon for the following reasons:

- Flexibility is highly desirable for TCP, as network is wide, unknown and heterogeneous.
- The ability to make changes (“upgrades”) without replacing hardware.
- The vast number of different environments prevent a single adoption into silicon.

While providing flexibility, software transport implementations, such as TCP, result in significant CPU utilization, long latencies, and large memory bandwidth requirements. The high CPU utilization derives from memory to memory copies, user context switches, user to kernel transitions, interrupts, and protocol state processing.

InfiniBand delivers a complete transport providing reliable in-order connections between user processes. The difference is, InfiniBand operates over an inherently reliable link layer whereas TCP typically operates over the unreliable Ethernet link layer. In fact, TCP actually uses packet loss as a form of “implicit congestion notification” to perform end to end rate control. InfiniBand, on the other hand, uses explicit end to end flow control. These differences in the protocol mechanisms lead to fundamental differences in how hardware can be built to deliver a reliable transport. In the case of InfiniBand, the protocol was designed with hardware implementations in mind and thus InfiniBand transport connections offer significant latency and performance advantages relative to TCP implementations.

InfiniBand effectively addresses all of the stated limitations of TCP and can provide seamless connectivity to applications designed to work with TCP. Many counter that the industry will develop TOEs (TCP Offload Engines) that will overcome these limitations and lower both latency and CPU utilization. But TCP has some major implementation issues that can only be overcome with “more” silicon than InfiniBand. TCP in hardware must continue to allow out-of-order retries which means that large caches are needed to hold the data, extra bandwidth is needed to receive, store, and re-order this data. These fundamental limitations hinder the efficiencies of TOE silicon, and even though the goal may be to enable TCP in a more homogenous environment like the data center, it will continue to be limited by necessary “flexibility” requirements of TCP. Also, since Ethernet is a LAN technology and designed to enable 100 meters of distance over copper, the power requirements for its links are necessarily higher than those of InfiniBand (designed for 17 meters over copper). InfiniBand silicon is shipping today that integrates the physical layer (or SerDes) into the core silicon. Mellanox’s InfiniScale device includes eight 10 Gb/s (4X) ports with the corresponding 32 2.5Gb/s serializer/deserializers (SerDes). It is not expected that Gigabit Ethernet devices will be able to achieve this level of integration due to the nearly 10X power requirement of the PHYs necessary to support LANs.

3.3 End to End Latency

The connections between tiers of the data center introduce latency to the typical transaction. The latency consists of three components: flight time, forwarding time, and transport processing. Flight time is the time the data takes to travel from one node to another on the wire and to get from the card into the system. The flight time component is directly tied to the bandwidth of the link therefore, a 10Gb/s link would reduces this time by a factor of ten relative to a 1 Gb/s link. However, the flight time component of latency is insignificant relative to the forwarding and transport processing component.

The second latency component consists of the time the data spends in switches and routers. Typically the forwarding time is minimal, however, congestion can result in significant packet buffering requirements translating directly to latency. InfiniBand link level flow control and virtual lane quality of service mechanisms offer significant advantages over Ethernet to prevent link level congestion, long latencies, and packet loss (see Mellanox’s white paper on link level QoS).

The third component of latency is the transport processing itself and typically it is the most significant of the three. Delivering a reliable transport is actually fairly difficult, and thus, historically, has been implemented by software to provide flexibility. TCP/IP is the most common form of software transport and typically runs on a CPU which is a single, centralized resource. The CPU can only perform one task at a time and because several tasks are required to implement the transport stack they must be serialized (performed sequentially one after the other) and thus latency is introduced. Latency contributions consist of several components:

1. Data Demultiplexing - Datagrams which arrive at a common destination are associated with a particular process to create a connection

2. Integrity Checks - Cyclical redundancy check or checksum verifies that no errors have occurred
3. Reliability and re-ordering - Packet sequence number checking, acknowledgement, and re-ordering
4. Data Movement - Copying of data from kernel space to the user space of a particular process (and vice versa).
5. Segmentation and re-assembly
6. Management - slow start windowing, congestion management, and connection state management
7. OS Scheduling - Operating System thread scheduling

A key observation is that all of these steps are a function of bandwidth - meaning the greater the bandwidth of the link, the greater the amount of data that must be received, stored, reordered and sent. This is important is since as bandwidth increases from 10 Mb to 10Gb, there is a thousand fold increase in the buffer and processing requirements placed on the host.

Some IT managers have had to limit the expansion of their data centers as the latency and performance impact introduced by additional tiers reached a point of diminishing returns.

3.4 Memory Protection, Address Translation, and Performance

A key requirement of data center connections is to insure that processes are protected from one another. Protection requires that data intended for one process does not interfere with the memory space of another process. This is vital to prevent a single rogue process from crashing the entire system, as was common in the days before protected mode operating systems became common.

Furthermore, modern operating systems employ virtual addressing in order for systems to support application memory requirements larger than their physical (DRAM) memory space. An application that tries to access a virtual address that is not in physical memory causes the page to be swapped from disk to physical memory. Thus, address mapping between virtual and physical memory is also required.

Both, memory protection and virtual address translation, are normally taken for granted as they are handled by the operating system. Nonetheless, address translation and protection checks consume a significant amount of CPU resources and thus seriously impact performance. InfiniBand implements these functions in hardware so that they need not be performed by the kernel of the operating system. This “kernel bypass” frees up the CPU, thus making cycles available for applications rather than for low level operating system functions.

3.5 RDMA and Message Passing and Performance

Two other key mechanisms that can accelerate the performance of data center interconnects are Remote Direct Memory Access (RDMA) and message passing. RDMA enables a local system to expose its memory to a remote system. Further RDMA allows the remote device to get/put data directly from/to the memory of the local system (remote and local are relative terms as the process is symmetric).

Message passing on the other hand does not rely on exposing local memory to remote devices., and exchanges messages between devices instead. In this case, the local device decides where to put incoming messages in memory. InfiniBand supports both RDMA and message passing in hardware, again greatly accelerating data center performance.

Of course, enabling RDMA potentially compounds the protection issues described in the previous section. Without the necessary precautions, RDMA allows not only rogue processes on your own system to corrupt other processes, but also allows *every* remote system potentially dangerous access to local system memory. Thus RDMA definitely demands the protection mechanisms described in the previous section supported by InfiniBand in hardware. Otherwise, a rogue process on a remote system could inadvertently send data and corrupt multiple applications. InfiniBand's protection mechanisms support these accelerated RDMA and message passing mechanisms, while still providing hardware memory protection.

3.6 Ethernet Performance in the Data Center

Typically, today's data centers are connected with 100 Mb/sec or 1Gb/sec Ethernet and much of the processing for the transport of data takes place in the host processor in software; namely the TCP stack. Since TCP is designed for LAN & WAN environments, it must allow for dropped packets and for packets which are received out of order. Although this has great advantages for communication over long distance and in a heterogeneous environment, it does have its price. Not only do all of the steps mentioned above need to be executed, but there must be enough transport or Network Interface Card (NIC) bandwidth to store, re-order, acknowledge, and maintain state information for all of the incoming data. This generally means that the NIC throughput capabilities must be three or even four times the link bandwidth to receive, reorder and resend simultaneously, which is a crucial requirement for many application including real time transmissions, like video streaming.

For 100 Mb/s Ethernet, this isn't too much of an issue, but as the test results provided by eTesting Labs indicate, a huge amount of the server's CPU power is required to move 1Gb/s of data with TCP. In October, 2001 eTesting Labs published[1] the results of a test they conducted on the Alacritech 1000x1 TCP Offload Engine (or TOE) which compares the results of the Alacritech 1Gb/s TOE versus standard 1GB/s 3COM and Intel Ethernet devices. The comparison of the TOE and the standard NIC clearly shows the

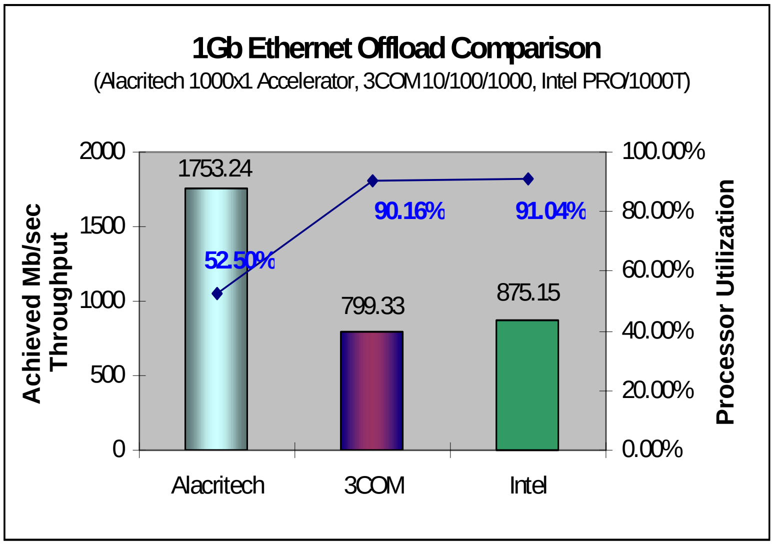

Figure 1. 1Gb Ethernet Offload Comparison

benefits that a TOE can provide to realize both, Ethernet bandwidth and lower host processor utilization. Figure 1, “1Gb Ethernet Offload Comparison,” on page 5 shows the full duplex data (read and write) data bandwidth (the theoretical maximum bandwidth is 2Gb/s) and the corresponding CPU utilization number.

These tests results were achieved on a dual Pentium® III 1Ghz processors systems. In other words, even with an offload engine, TCP still requires 52% of TWO processors to move 1.75 Gb/s of data. It is also important to note that even with two processors, the standard 1Gb/s Ethernet NICs can only achieve around 800-875 Mb/s (only 40-44% of the stated data rate) while using over 90% of the processor cycles offered.

Even though these test results do not mention latency, industry analysts estimate that typical TCP transport latencies are in the hundreds of microseconds to about 50 milliseconds (some state even higher numbers). Therefore, applying the formula previously stated the round trip transport latency can range from just over a millisecond (100 *u*sec*2*6 = 1.2milliseconds) to over 1/2 of a second (50 msec*2*6 = 0.6 seconds)!

3.7 10Gb/s Ethernet

There have been a number of discussions in the press lately about the adoption of 10Gb/s Ethernet, but some interesting details on 10Gb/s ethernet aren’t commonly stated. The 10Gb/s IEEE specification is yet to be ratified, although this is expected to happen some time in 2002. The design goal of 10Gb/s Ethernet is to apply this technology as aggregation pipes in the MAN (Metro Area Network) where it is much needed and likely to be a big success. But to meet this market need, the specification is designing the link only as a fiber optical technology (not copper). This technology also implements a physical layer that is different than that of the 1Gb/s Ethernet technology. This change will not allow 10Gb/s and 1Gb/s links to interoperate, or 10Gb/s is not backwards compatible and can NOT run over CAT 5 cabling. As previously stated, the processing, re-ordering and other TCP software processing requirements grow exponentially in the moves from 100 Mb/s to 1Gb/s to 10Gb/s. For a processor to receive data, store it, re-order it and resend it the processing and memory requirements again jump 10 fold. As shown by the eTesting results, even with a TOE, the processing requirements to try to saturate a 1Gb/s link exceed the performance that a single 1GHz processor can supply.

Because of the target market, the volume of 10Gb/s is not expected to be anywhere near that of today’s 1Gb/s Ethernet ports. An IDC report (#25001R) has projected that 10Gb/s Ethernet ports for LAN connections will expand from 16 thousand ports in 2001 to just 387 thousand ports in 2005. This clearly illustrates that 10Gb/s NICs will not be a predominate desktop or server technology, in the next 4 years, and that it will not share the economies of scale that previous Ethernet technologies have enjoyed. Also, since this is a fiber optical technology only, each port must contain fiber transceivers (which cost hundreds of dollars more than copper connectors per port).

[1] http://www.etestinglabs.com/main/reports/alacspssa.pdf. Names are those of the respective owners.

4.0 Clustering and Virtualization

Another key element in the data center is clustering and virtualization. Clustering involves the use of high speed, low latency connections between compute, storage, and I/O elements in order to build scalable and high availability systems. In order to manage a cluster effectively and efficiently it is necessary to virtualize these individual resources into a single management entity. Virtualization here means that a pool of resources are presented as a single resource. Any particular element in the pool of resources can be deployed on demand, which - in turn - requires the ability to move context (state) from one resource to another very quickly. This puts pressure on both latency and bandwidth, and for these applications a high speed clustering fabric like InfiniBand is ideal. Clustering and Virtualization also creates large amounts of inter-resource traffic to transfer context and explains why internal data center traffic is typically much greater than at the edge.

5.0 InfiniBand in the Data Center

InfiniBand implementations currently offer two levels of bandwidth:1X links (2.5 Gb/s) and 4X links (10 Gb/s)[1], both of which support copper connectors. In addition, the specification defines a 12X link offering 30 Gb/s with devices expected in 2003. InfiniBand implements the transport functions 1-5 (listed above) entirely in hardware with minimal requirements for CPU processing. In addition InfiniBand implements advanced features such as RDMA, message passing, memory protection, and address translation in hardware. Furthermore, InfiniBand greatly accelerates items 6 and 7 which reduces the software involvement to a minimum. Since these tasks are implemented in hardware (silicon), it is possible to process them in parallel simultaneously, since no central resource is involved. This greatly reduces the latency of the transport connection as compared to TCP. Also, Infiniband is designed upon a reliable, in-order, transport so the need for storing data for reordering of the packets is not required. This eliminates a costly and slow step in the process and greatly reduces the buffering required in the design of InfiniBand devices.

Mellanox Technologies InfiniBridge MT21108 HCA device can operate in either the 1X or 4X mode and is able to achieve excellent data rates and low latencies in the field. For example, in a storage environment application on a 1X link Mellanox is able to achieve a data rate of almost 3.8 Gb/s while using only 6.9% of a Pentium III 800 MHz processor. Although the test set-ups are not the same, both the

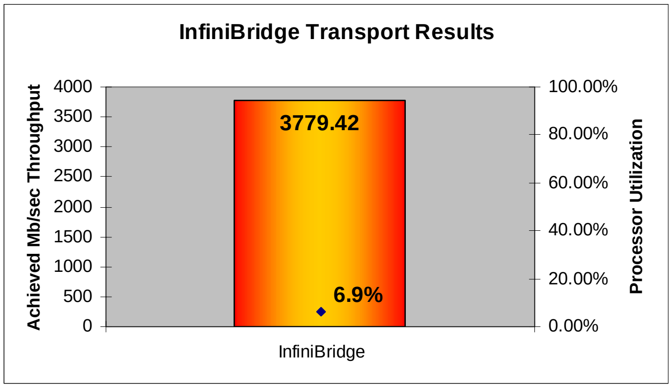

Figure 2. InfiniBridge Transport Results

[1] The InfiniBand architecture embeds the clock in its signal using standard 8b/10b encoding thus baud rates of 2.5 and 10 Gb/s yield corresponding data rates of 2.0 Gb/s and 8.0 Gb/s. Ethernet quotes data rates and both technologies are full duplex; meaning that they can transmit and receive at the same time so, the total aggregate bandwidth is actually twice the data rate.

eTesting results and our results are applications or benchmarks that work to saturate the links with maximum bandwidth between systems.

In addition, the measured round trip latency between ringing a doorbell, the sending of the data to a target and the receipt back of a completion notification is generally less than 6 microseconds. A 7 tier round trip based on this would be 6*2*6 = 72 microseconds or 0.000072 seconds which is significantly (generally orders of magnitude) faster than TCP.

InfiniBand 4X links operating at 10 Gb/sec offer 10-100 times the bandwidth between each tier, without the burden of running TCP on each server. With InfiniBand the CPU cycles that would normally be consumed running the TCP/IP protocol stack become available to run the application within the tier (web, application, or database server, etc.). The benefits of an n-tier data center architecture are many, and Infiniband solves the critical issue of latency which is a by-product of the flexibility that tiers provide.

6.0 Optimal Utilization of InfiniBand and TCP

The n-tier data center has multiple requirements which are different at the core and at the edge. To optimize the performance of the Data Center, both, InfiniBand and TCP, can be utilized in a complimentary manner. TCP/IP is indeed required at the edge of the data center and is the ideal protocol to connect to the uncontrolled environment of the Internet. However, within the controlled environment of the data center the complexity of TCP transport, CPU processing requirements, and long latencies argue for the more streamlined transport offered by InfiniBand. Fortunately, a seamless connection between TCP and InfiniBand transports is available at the application level.

6.1 Sockets Direct Protocol - Seamless Application Connectivity

Sockets Direct Protocol (SDP) delivers a backwards compatible interface to applications originally written to work above a TCP software transport and enables these applications to transparently take advantage of InfiniBand hardware transport and operate with greatly accelerated performance. SDP delivers the basic “sockets” mechanism offered by TCP so that applications connect transparently to the underlying hardware. If only Ethernet hardware (NIC or offload engine) is available, then SDP will revert to software transport where the CPU is required to implement some or all of the transport stack. If InfiniBand hardware is available, the SDP stack will detect this and allow the InfiniBand channel adapter to perform the complex transport tasks normally performed by the kernel of the operating system. This kernel bypass is illustrated in Figure 3, “Sockets Direct Protocol Stack,” on page 9. There is no difference from the application viewpoint except that with InfiniBand hardware higher performance is achieved and more CPU cycles are available for the application.

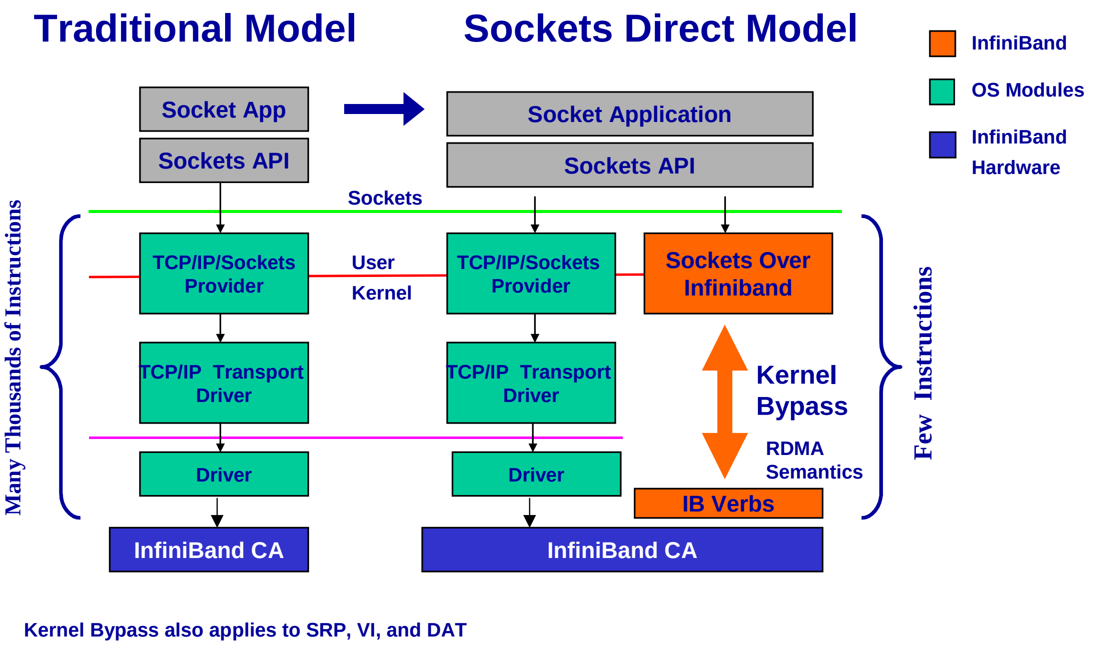

Figure 3. Sockets Direct Protocol Stack

Other software interfaces such as Virtual Interface (VI), Direct Access Transport (DAT), and SCSI RDMA Protocol (SRP) have been specially designed to take advantage of InfiniBand transport to support high performance implementations of database clustering, network attached storage, and block level storage area network applications respectively. All of these software interfaces exploit InfiniBand's hardware transport and "kernel bypass" capabilities in order to deliver higher performance while consuming less CPU cycles. Database applications need to move and process huge amounts of data and the performance is thus hurt most by the server to server latencies introduced by expensive software transports like TCP. The CPU performance penalty caused by software transport (TCP/IP) explains why parallel database providers have been so willing to adopt the Virtual Interface Architecture supported by InfiniBand.

6.2 TCP at the Edge

As previously asserted, TCP is indeed ideal "at the edge." This begs several questions: Why does it matter where TCP is terminated? Why is it better to terminate TCP at the edge and not in the final tier? Isn't the number of TCP connections and processing requirements the same, independent of where TCP is terminated?

The answer to these questions requires a better understanding of how information is stored and transactions occur within the data center. This analysis yields the interesting result that in fact terminating TCP at the edge and using streamlined InfiniBand connections within the data center results in the optimal solution, minimizing transport CPU processing, providing minimum latencies, and maximum throughput.

6.2.1 Information Hierarchy in the Data Center

Imagine a user (client) surfing the web across the internet communicating with a data center. Typically, the majority of web browsing activity is fairly cursory in nature - a quick glance at the first few pages of a media or e-commerce web site. These first few pages typically consist of generic information intended for all users and the content is primarily static. The majority of web browsing activity stops at this static content, however some percentage of users are more interested and probe deeper into the site. These users effectively begin to move through the n-tiered data center. As a user progresses the site may begin to provide user dependent, cookie-aware, rich, dynamic content. Ultimately if a transaction needs to be processed (for example purchasing content from an e-commerce site) than the backend tier supporting data base transaction processing becomes involved. The number of users diminishes progressively as one moves through the tiers of the data center.

This is illustrated by the inverted pyramid shown in Figure 4. For example, perhaps 90% of all users do not go past the first few pages of static web content. A few more users start to browse in areas which require an application server to retrieve user specific information and generate custom, dynamic content. Only a few percent of users actually progress to the point of making a purchase that requires transaction processing to access the data base tier of the data center.

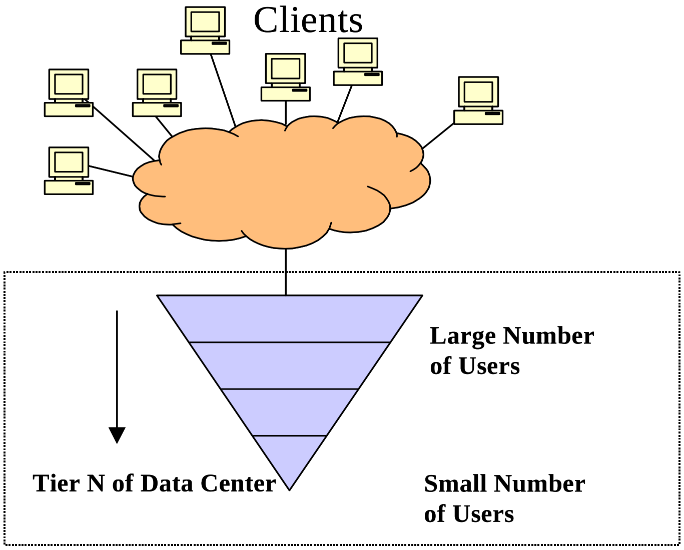

Figure 4. User and Connection Hierarchy in the Data Center

This type of user traffic pattern seems to imply that the front tiers of the data center need to offer higher performance and bandwidth than the backend tiers. In reality this is false. The front end does indeed need to be able to support a large number of connections, however queries tend to be simple and static content is largely cacheable elsewhere, thus decreasing the requirement for bandwidth. Clearly as one moves to application server tiers the custom, dynamic content can not be cached and thus must be delivered new each time. As one progresses to the database tier of the data center an odd thing happens: a trivial client request (“Buy Now”) spawns and enormous amount of intra-datacenter traffic. This is so because in an advanced data center a huge amount of bandwidth and server to server communication is generated by accesses to large clustered databases, transaction monitors, and two phase commit procedures required to insure integrity of the transaction. Thus, Figure 5, “Bandwidth and Processing Hierarchy in the Data Center,” on page 11 shows the pyramid describing the bandwidth and processing requirements of the data center is inverted from that of the number of users.

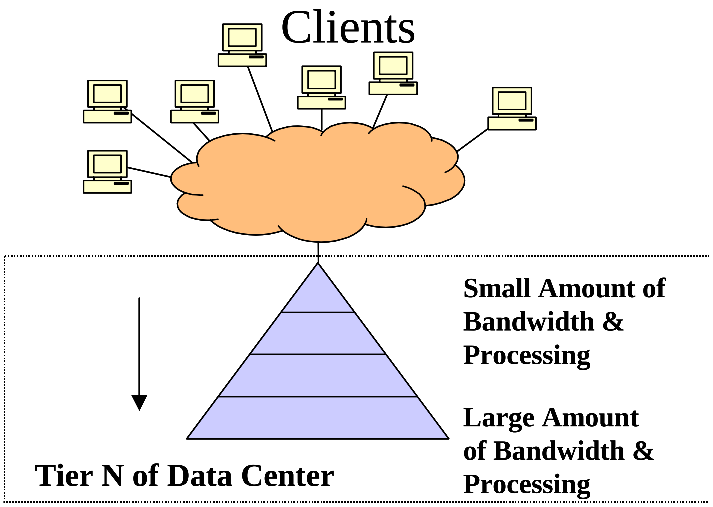
Clients
Small Amount of Bandwidth & Processing
Large Amount of Bandwidth & Processing
Tier N of Data Center

Figure 5. Bandwidth and Processing Hierarchy in the Data Center

In other words, the number of users/connections in the first tier of the network is large however the bandwidth (and thus processing) requirements are relatively small. As one moves deeper into the data center the number of users/connections decreases while the amount of bandwidth increases. The TCP transport connection is ideal to connect to clients across the internet to the edge of the data center. Fortunately, the bandwidth requirements of each request at the edge are relatively small compared to those deeper within the data center. Thus terminating the TCP at the edge of the data center and making a transparent connection to InfiniBand delivers optimal performance as the cost of the termination can be amortized for a rack or data center at the edge and NOT on each server.

6.3 Virtualization Advantages

Combining the concept of TCP termination, n-tiered data centers, SDP and InfiniBand are the perfect combination to enable even more efficient data centers. Termination of TCP creates a virtual pool of requests for servicing, the InfiniBand fabric allows these requests to move anywhere on the fabric (even the ability to “skip unneeded” tiers for specific requests), and SDP is the protocol that enables the quick movement of context from each tier. Data centers with architectures that includes InfiniBand will have much greater flexibility in designing and implementing efficient interchanges of their data.

7.0 Summary

TCP and Infiniband are complementary technologies designed for different environments. TCP thrives in providing reliable connections for heterogeneous environments often over long distances. InfiniBand thrives by providing a reliable, in-order, fast, low latency, high bandwidth, scalable transport in a homogenous data center environment. These two are a natural fit for each other:

- TCP connects the clients and the roadways of the Internet highway
- InfiniBand will serve as the compute and I/O fabric that fuels data generation for the Internet and Enterprise engines from the data center

Together these two technologies will continue to advance the performance, and reliability of the Internet.

8.0 About Mellanox

Mellanox is the leading supplier of InfiniBand semiconductors, providing Switches, Host Channel Adapters, and Target Channel Adapters to the server, communications, and data storage markets. In January 2001, Mellanox Technologies delivered the InfiniBridge^{TM} MT21108, the first 1X/4X InfiniBand device to market, and is now shipping second generation InfiniScale silicon. The company has raised more than $33 million to date and has strong corporate and venture backing from Intel Capital, Raza Venture Management, Sequoia Capital, and US Venture Partners.

In May 2001, Mellanox was selected by the Red Herring Magazine as one of the 50 most important private companies in the world and to Computerworld Magazine Top 100 Emerging Companies for 2002. Mellanox currently has more than 200 employees in multiple sites worldwide. The company's business operations, sales, marketing, and customer support are headquartered in Santa Clara, CA; with the design, engineering, software, system validation, and quality and reliability operations based in Israel. For more information on Mellanox, visit www.mellanox.com.