White Paper

Realizing the Full Potential of Server, Switch & I/O Blades with InfiniBand Architecture

1.0 Introduction

The notion of blade-based servers is a relatively recent phenomenon with both startups and the well-established server vendors having already announced server blade products or the intent to develop such products. Typically the early products have focused on power and density as the primary benefits of server blades. This focus on form factor rather than function misses the true benefit of blade based technology. Focusing only on the server component also misses the other critical parts of a system area networks: I/O connectivity and switching. In fact the true benefit that server blade *and* more importantly I/O blade technology provides is the ability to deliver a highly available, easy to manage and scalable infrastructure for both computing and I/O. Such an infrastructure, when implemented optimally will result in a much lower total cost of ownership (TCO) for the IT manager.

To fully realize the TCO benefits of blade based computing the interconnect technology must deliver (at a minimum) the following core features:

- Transport level connections delivered in hardware
- Reliability
- Scalability
- Fault tolerance
- Hot swappability
- Quality of service (QoS)
- Clustering capabilities (RDMA and message passing)
- Multiple generations of compatibility to 10Gb/s
- I/O connectivity
- Multi-Protocol capabilities

- Inherent Redundancy
- Failover and active stand-by
- In-Band Management
- Interconnect manageability
- Error detection
- Software Backwards Compatibility

This paper will explore the differences between LAN based server blades vs. InfiniBand based blades. It will be shown that, fundamentally, bus and LAN based implementations fall seriously short of meeting the stated requirements. The InfiniBand Architecture delivers the critical capabilities to realize all of these attributes while significantly improving overall system performance.

2.0 What is a Server Blade?

It is important to define what is meant by the term “server blade”. First of all as a practical definition a server can be viewed as simply a re-purposed desktop computer with the distinction of utilizing higher-grade components, additional memory and hard drive capacity, and packaging enabling rack mounting. In the most general sense a server blade then is an industry standard computer delivered on a single PC card, which can be plugged as a module into a chassis. Practically such a chassis may accept from 8 to 24 such cards. As defined this server blade is not able to operate standalone but in fact requires the chassis to provide power, cooling, rigidity, and a disk drive. In this sense a server blade is simply a different form of packaging of a traditional 3U or 1U server (without a lot of the PC baggage).

As defined above a broad class of products satisfy this definition of “server blades”, including single board computers for VME and Compact PCI (CPCI) passive backplanes, as well as, newer proprietary blade form factors utilizing 100Mb and Gigabit Ethernet for connectivity. Even though each of these technologies deliver servers in a blade form factor, none effectively provide the key benefits of high availability, ease of management, and scalability. For example, CPCI is able to provide a solid structure for I/O connectivity, however fails to provide scalability, robustness, and chassis-to-chassis interconnects. The shared bus architecture of CPCI fundamentally limits hot swap, and scalability capabilities while the overall PCI software architecture does not facilitate I/O sharing. Further extremely limited error detection capabilities (a single parity bit) make CPCI insufficient to deliver a platform capable of detecting and correcting errors. In addition the low-level load/store architecture of PCI does not address the requirements of clustering and channel based computing. Thus CPCI does not provide the framework allowing it to deliver a true fault tolerant, highly available, computing platform.

Ethernet based solutions fall short of delivering most of the core requirements. Ethernet interconnected systems can provide some scaling capabilities however they are fundamentally limited by delivering an unreliable service that requires the burden of the TCP stack to ensure the reliable receipt of data. Ethernet is also hampered by its inability to share I/O, its ineffective Quality of Service mechanisms (see Mellanox’s white paper on link level QoS and congestions spreading), its lack of RDMA or messaging capabilities, high latencies, higher power, bandwidth limitations of 1Gb/s and the fact that there is no compatible path[1] to 10Gb/s. Also, no copper connectors are planned for 10Gb/s second Ethernet, making it require the huge expense of fiber optical connections for all links. Also stop gap efforts such as the PICMG 2.16 standard are quickly being superseded by more robust fabric solutions such as the PICMG 3.2 standard that provides interconnect support through InfiniBand.

3.0 Stateless Server Blades

In order to deliver on the benefits enabled by blade technology we need to extend our definition of a server blade to a Stateless Server Blade. This means that the server blade, rather than being essentially a re-purposed desktop computer, is reduced to its bare essence: CPU, memory, and I/O connectivity. Such a device is effectively a pure computational element and is not encumbered by components and connectivity to support keyboard, video, mouse, hard drive (and the incumbent OS support), plus the myriad of other functionalities required by more general-purpose desktop computers. Such a blade is said to be “stateless” because the parameters which define its identity and application are not stored on the blade itself, but rather is determined intelligently when the blade is initialized. To be truly stateless, the functions of OS booting, network provisioning, loading drivers, and mounting root and application file systems must be done using remote services and storage. The storage contains all of the OS and applications images that reside within the data center and provisioning services automatically direct the blades to their booting resources and I/O interfaces based on the system administrator’s input. This enables a single system admin to provision and manage thousands of stateless server blades from a single management console.

Removing the hard drive from the server blade not only reduces both cost and power but also greatly lowers operating and system management costs. Providing simple computational server blade elements access to an application and operating system software through a single unified and coherent storage repository greatly simplifies application management. Simplicity and lower cost of management is the fundamental advantage that drives IT managers to buy complex and expensive symmetric multi processing (SMP) 8-Way and 16-Way servers. Clusters of stateless server blades can leverage economies of scale to deliver the same management cost benefits at a fraction of the cost of many-way servers.

4.0 I/O Blades and I/O Sharing

The other key aspect of realizing the true benefits of blade based computing is delivering an I/O blade that utilizes the same form factor as the Server Blade. This gives deployment flexibility to the system admin as they face trade-offs between I/O and compute power requirements. Applying this flexibility allows resources to be independently deployed within the same chassis or across multiple chassis and load balanced to meet the current computing needs.

Furthermore InfiniBand extends the concept of the I/O fabric to enable I/O sharing. The unique ability of InfiniBand to create many to many relationships between CPUs and I/O elements pro-

[1] 1Gb/s and 10 Gb/s Ethernet do not share the same signaling at the physical layer.

vides significant cost, reliability, and flexibility benefits. For example, multiple server blades can share a single Fibre Channel blade. Previously I/O cards have always required a one-to-one relationship to a single computer. The cost and power of under utilized I/O cards in multiple systems can now be eliminated. This also provides a significant management benefits in that a driver for a shared I/O blade need be installed only once (into the storage pool), instead of for each server.

5.0 InfiniBand "The Enabler" of the Backplane

The InfiniBand Architecture[1] is the key to enabling the full potential of stateless server blades and I/O blade data centers. Another feature that InfiniBand offers is its use as the backplane interconnect for blade computing. The backplane connects the server blades, I/O blades and switch blades into a single unified chassis fabric. In this architecture the backplane contains 16 1X InfiniBand blade connectors that are redundantly connected (dual star) to each of the two switch blade connectors. Each switch aggregates the 16 blades and connects the chassis to the InfiniBand fabric within the data center.

6.0 InfiniBand Blade Chassis and Switching

As mentioned not only does the chassis accept both server and I/O blades on a common (InfiniBand) backplane but connects to the fabric through Switch Blades, of a similar form factor. The InfiniBand architecture offers the ability to connect either through traces on a board or through copper or fiber optic cabling. This allows the server blades, I/O and switch blades all to interconnect in a chassis through a single InfiniBand backplane to other chassis', storage devices, gateways and any other InfiniBand devices on the fabric.

The ideal switching solution for this chassis is a 16 + 4 switch blade. Such a switch blade has sixteen 2.5 Gb/s (or 1X) InfiniBand ports that connect to the backplane plus four 10 Gb/s (or 4X) InfiniBand ports that connect as uplinks to the fabric. This switch configuration is a full-wire speed non-blocking implementation. This blade is ideal for the aggregation of the 16 1X ports for server or I/O blades in the that connects to the fabric with up to four 10 Gb/s links. InfiniBand is providing 10 Gb/s over copper connections (today) and these connectors cost hundreds of dollars less than the fibre optic transceivers. Mellanox Technologies currently shipping InfiniBridge and InfiniScale devices make it straightforward to implement a full wire speed, non-blocking 16 + 4 switch enabling dual redundant star fabric topologies.

7.0 High-Performance Clustering Server Blades

Server blades within the chassis or between chassis can be clustered together for both high performance and fail over. With up to 16 server blades per chassis and multiple chassis connected through the InfiniBand fabric, clustering is now simple, and can easily be scaled to 64 or 128

[1] More details about the specific attributes of InfiniBand can be found in Mellanox's white papers: Introduction to InfiniBand and InfiniBand in the Data Center at http://www.mellanox.com/products/whitepaper.html.

nodes or in theory can be scaled to thousands of server blades (although the practical limit will be determined by the latency requirements and number of switch hops the application can bare).

Clustering within the OS is the ability to view multiple CPUs (in this case blades) as either one single compute image or as multiple compute images. This allows the OS to configure, for example, 16 blades as either a single 16-way compute node or as eight dual-clustered nodes (or other combinations). The benefit occurs when an OS can support the repartitioning of it's resources through clustering; this allows stateless server blades to be configured into the needed images that best provide the optimal performance the IT manager desires.

Also there are a number of key applications that recognize clustering natively. One of them is DB2 for which IBM has also announced the availability of DB2 Universal Database to customers on the InfiniBand fabric[1]. Also in an IBM white paper: IBM DB2 Universal Database and InfiniBand Technology; IBM discusses performance advantages with clustered processors as compared to processors in a SMP server[2]

8.0 Message Passing and Remote Memory

The InfiniBand Architecture supports both message passing and remote memory semantics to provide flexibility for communications and cluster development. The "send" capabilities of the InfiniBand architecture enable support for messaging. Message passing enables clustering applications that do not require exposing server node memory. Instead server-to-server communication relies on a few well-known and high performance communication channels between nodes.

The fundamental technology provided by the InfiniBand Architecture for memory semantics is the ability to support Remote Dynamic Memory Access or RDMA. RDMA allows for systems on the InfiniBand fabric to share specific (protected) locations in memory with other systems and pass data (read and write) between the locations at either the 1X (2.5Gb/s) or 4X (10Gb/s) bandwidth with extremely low latencies *without significant involvement of the CPU*. This enables very tight clustering for performance and the ability to access cached data from storage devices in record times. Frequently applications exploit both message passing and RDMA to achieve optimal performance

9.0 I/O Connections

Over time InfiniBand will drive a completely native InfiniBand data center. In such a data center the storage connects to the fabric from either SCSI to InfiniBand, or Serial ATA to InfiniBand, or Raid to InfiniBand, or other Native InfiniBand NAS or SAN units. "Storage Blades" can even be added to the chassis to replace Direct Attached Storage or remote SAN connected storage. Until native InfiniBand storage devices are available, Fibre Channel to InfiniBand gateways will be the

[1] See: http://www.ibm.com/Press/prnews.nsf/jan/F30E4D2988601A8B85256B1200701B98 for the press release.

[2] See http://www-3.ibm.com/software/data/pubs/papers/infiniband/infiniband.pdf for the white paper.

primary vehicle for remote storage. These devices exist in the market today and future versions will take on the blade form factor.

Communications within the data center run on the fabric, but externally, the communication uses Ethernet via an Ethernet to InfiniBand router. These routers terminate TCP at the edge of the data center and pass on the requests into the data center as native InfiniBand packets. These routers exist today in 1U form factors, are under development as blades and they provide the benefit of removing the overhead of TCP from the entire data center for SDP (Sockets Direct Protocol will be discussed later).

10.0 Cable Management

Another benefit of the InfiniBand blades and chassis is that nearly all cables are eliminated. This is possible since InfiniBand provides both basic I/O connectivity as well as offering in-band management. This means that clustering, communication; storage and management functionality that would have normally been accomplished using MANY cables, can now be done over the fabric with the assistance of the InfiniBand Subnet Manager (SM). With a stateless server blade utilizing the InfiniBand backplane to provide automatic I/O connectivity the cable requirements for computing is now reduced to zero. That is correct, there isn't a single cable needed for clustering up to 16 server blades together. Providing I/O connectivity or expanding the fabric beyond 16 nodes does still need cables, but the requirement is greatly reduced. For example in a typical non redundant fabric the cables can be reduced to a single 4X cable running from the 16 + 4 switch, plus the cables running from the Ethernet blade (typically 2) and the Fibre blades (either one or two). A typical, fully redundant fully I/O connected implementation shouldn't contain more than 8 cables, well less than one cable per server blade. And fewer cables improve reliability and reduce the chances for mistakes.

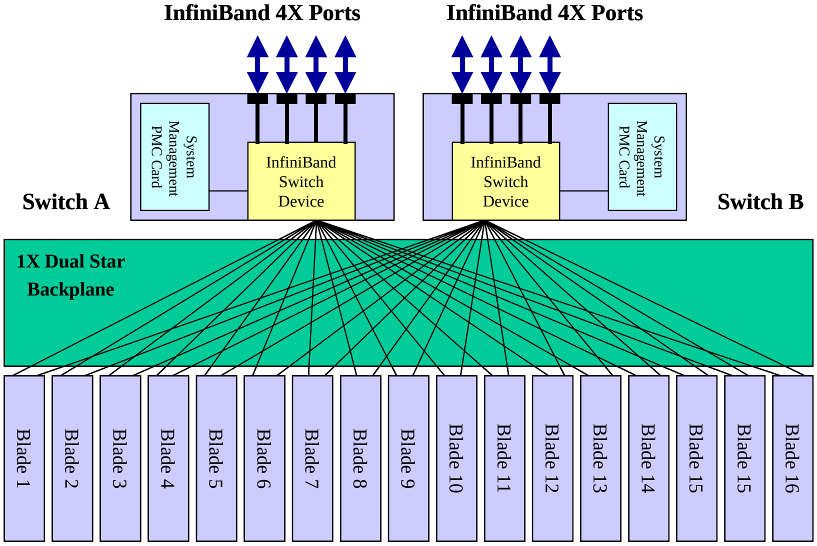
11.0 Fail Over and Provisioning

Many IT managers relate the term “clustering” as a way to achieve fail over, meaning they often implement the deployment of a complete second unit to cover the failure of the first. This is an expensive proposition as it not only requires twice the hardware, but also more than twice the management since the cluster needs software to ensure failover to the second system. With stateless server blades and I/O blades two or more blades can be deployed for a single task and provide their combined compute resources to perform the task. Should one fail the other(s) picks up the load with no loss in operation. In the event of such a failure, a few “spares” can be assigned out of the complete rack (maybe 2 or 3 out of a 100) and one of them can then be quickly remotely booted to assume the place of the failed unit. Again this is accomplished thought the tightly coupled InfiniBand fabric that allows I/O or compute sharing throughout the fabric. So the failed blade can be replaced by another, even if its replacement is in another rack within the fabric. Also, the replacement of the failed unit is very straightforward. Remove the blade from the chassis (at the convenience of the system admin); quickly swap out the DRAM between the blades, replace and then remotely reboot.

IT managers today have a difficult time determining what their real compute and I/O needs are going to be. For a small, but fast growing company running a data base program does the IT manager need a 2-way, 4-way or 8-way server to perform the task? The IT manager often doesn't know but buys the 8-way server because he or she knows that the lack of performance of the deployed system or disruption of upgrading to a bigger server, in the future, could cost the company much more than the price difference between a two way system and 8-way system. This is why so many server OEMs are now offering "on-demand" upgrades of their multi-way systems to their IT customers. InfiniBand based blades remove the guess work from the process and allow the IT managers to either scale or re-deploy compute or I/O resources to meet growing demands or to satisfy short term peak demands.

Returning to the database program example, the IT manager could deploy two blades to the program. Then once it is running, measure the performance of the two blades and then add a third or fourth blade and measure that performance. In this manor the IT manager can quantitatively measure where there is diminishing returns and then scale back one blade to achieve optimal performance. Once the company grows and the performance is determined to diminish, other blades can be deployed to maintain the desired performance level. Another aspect of this is the load balancing of compute needs. For example if the IT manager knows there is high web activity during the day and heavy data base use (internally) during the night. Stateless server blades could be redeployed (re-initialized) from the n-tier data center at night to run the database and then return to the n-tier application during the day. Or another way to solve this issue is to keep a few "spare compute units" on the side. These spares would really be utilities players that are deployed only when needed to meet peak demands (quarter end for example) or when an individual blade fails.

12.0 Reliability, Availability, and Serviceability



16 Slots accept Server or I/O Blades

Reliability, Availability, and Serviceability (RAS) is a touchstone requirement for modern data centers. The InfiniBand switch fabric architecture provides basic mechanisms that deliver RAS naturally. From a practical perspective failures are inevitable and redundant connections are the key to insuring RAS. InfiniBand hardware offers built in failover mechanisms (automatic path migration) that utilize these redundant connections to maintain data flow even in the presence of link failures. Practically this means that two 16 + 4 switch blades are deployed inside a single chassis to create a redundant star. This redundancy can then be extended across the fabric, by simply connecting in other redundant switches. So for applications requiring high availability the InfiniBand Architecture offers the solution and eliminates any single point of failure.

The InfiniBand Architecture has designed in hot swap capabilities, active and passive standby, scalability without disruption, managed and unmanaged (hardware) failover and other features supporting high availability. IT managers today pay high premiums for highly available systems. The InfiniBand Architecture designs in high-availability that insures that the data center will always be running and achieves this naturally, by design, not by massive (and expensive) over-provisioning.

13.0 Quality of Service

QoS is supported by InfiniBand at the fabric level through the advent of multiple Virtual Lanes (VL). These VLs are separate logical communication links, which share a single physical link. Each link can support up to 15 standard VLs and one management lane (VL 15). VL15 is the highest priority and VL0 is the lowest. As a packet traverses the subnet, a Service Level (SL) is defined to ensure its QoS level. Each link along a path can have a different VL, and the SL provides each link a desired priority of communication. An example of this would be that clustering traffic could be given a high priority while storage backup and email traffic is given a low priority. Should congestion begin, the clustering traffic would continue while the backup and email traffic hold off till credits could be issued.

14.0 Low Latency and Performance Benefits

Many data center, especially those that are large web hosting sites utilize an N-Tier architecture for distributed computing. The key advantages of these n-tier designs are that services (or applications) are layered in tiers so that modifications, expansion or the addition of another tier can take place without the disruption of the other tiers. A web hosting n-tier data center would have as many as 7 tiers (client –> firewall –> load balancer –> SSL –> Web Host –> E-Com –> Data Base and back). Although there are great design benefits to this architecture, there is a big drawback to these designs, and that is LATENCY and the host processor power lost to the TCP stack executing these requests. A single HTTP command from a client must pass through 12 TCP/IP stacks to serve a single request; that is it must move though all the tiers and then back again to service the request.

Industry sources have quoted the latency of a TCP from anywhere from a hundred microseconds to as long as 50 millisecond and even higher. This means in total the latency alone in processing a request inside the data center (not including the client) could be from over a 1 millisecond to as slow as 1/2 second. Either way this is a significant amount of time for Yahoo.com or Amazon.com that are processing millions of request an hour. Compare this to InfiniBand where Mellanox’s InfiniBridge silicon provides the transport in hardware that typically moves the data in just a few microseconds. Multiple tiered data centers or mission critical applications that use the classic 3-tier model can greatly improve performance with InfiniBand’s lower latency and it’s more than 2X bandwidth advantage (2 Gb/s InfiniBand 1X compared to Ethernet at 1Gb/s).

Many estimate that TCP processing can typically steal 30-50% of the processing cycles from the host. Using an average number of 40% an analogy of that could mean that for every 700 MHz low power processor that only around 400 MHz of its cycles are available to run the targeted application. CPU performance continues to increase at a dazzling pace so some IT vendors view these MIPS as “free” however the expensive database applications that run on them are not and typically licensing fees are per CPU. Given the expensive of these database licenses every bit of CPU processing power should be devoted to the application and not devoted to delivering a reliable connection. Thus, TCP is utilizing a major part of a data center’s compute power just for reliable transport. By contrast InfiniBand offers low overhead protocols and hardware transport support that only requires a few processor cycles are required to move data.

In a recently released TCP offload engine report, it was documented[1] that in order to achieve close to the 1Gb/s maximum bandwidth of Ethernet that VERY significant host CPU cycles are required. In a report by eTesting Labs, it was shown that to obtain about 1.9 Gb/s of (full duplex) Ethernet bandwidth that 52% of dual 1GHz processors were required even with the implementation of a TCP offload engine. They compared this to two conventional NICs that achieved 800 to 875 Mb/s of (full duplex) bandwidth while using over 90% of the cycles of the dual 1GHz processors. Although the test conditions are not the same, Mellanox has a customer application, for storage, that achieves over 90% of the maximum InfiniBand 1X bandwidth (full duplex) of almost 3.8 Gb/s while using only 7% of a 800 MHz.

InfiniBand 2.5 Gb/s (1X) and 10Gb/s (4X) HCAs and switches are available in the market today today. It is important to note the affordability of 10Gb/s switching silicon. There are multiple silicon vendors with 10Gb/s InfiniBand switching devices that have announced, shipped and stated silicon pricing at less than $75 per 10Gb/sec port. This pricing will allow InfiniBand to achieve the lowest 10Gb/sec port costs in the industry that will quickly approach today’s pricing for 1Gb Fibre Channel switches. And as mentioned, these 10Gb/s connections are designed to connect with low cost copper connections (or fiber optics, if desired).

15.0 Software Protocols

The following protocols assist InfiniBand enable its low overhead architecture and all of them have been publicly demonstrated at a number of events since August of 2001.

DAFS (Direct Access File System) protocol enables standard memory-to-memory interconnects, based on VI (Virtual Interface Architecture), between servers and shared filed storage. With DAFS, servers can request data from network attached storage over InfiniBand and utilize the memory protection and RDMA features that allow the NAS filer to DMA data directly back into the server’s memory. These breakthroughs will improve the performance and reliability of Internet and enterprise applications in the data center environments.

SDP (Sockets Direct Protocol) is a lightweight protocol that allows the movement of communication data within an InfiniBand Data Center without excessive assistance from the OS and host processor. SDP technologies takes advantage of InfiniBand’s reliable, in-order delivery (in hard-

[1] See: http://www.etestinglabs.com/main/reports/alacritech.asp for the report

ware), de-multiplexing, RDMA (remote direct memory access), and message passing capabilities, to bypass the OS Kernel and avoid interrupts to achieve low CPU utilization, low latencies and high bandwidth. RDMA transactions with SDP can move data from an Ethernet edge device to a server (or server-to-server) directly between the user's memory space. This means that the requested data needs to be moved only once to get to a location in memory where it can be immediately utilized. This is overcomes many of the inefficiencies of TCP.

SRP (SCSI Remote Protocol) brings the virtues of RDMA transfers for moving SCSI block data from servers to SANs and vise versa. This protocol provides new levels of storage efficiencies within InfiniBand Data Center for Storage Area Networks. InfiniBand provides a superset of Fibre Channel's underlying mechanism and uses the identical SCSI encapsulation protocol. This greatly simplifies the tasks of porting software originally written for Fibre Channel applications and allows InfiniBand to take advantage of SAN management and virtualization software.

16.0 Deployment

In December 2001, Mellanox Technologies publicly disclosed their architecture for InfiniBand Stateless Server, Switch and I/O Blades. This architecture will speed the time to market for future blade products from OEMs supporting InfiniBand server and I/O blades. More details on the expected deployment of this architecture and the resulting follow-on products are available at www.mellanox.com.

17.0 Summary

The true benefit that InfiniBand server blades *and* more importantly I/O blade technology delivers is a highly available, easy to manage, scalable computing and I/O infrastructure, that results in much lower TCO costs for the IT manager. InfiniBand Server, Switch and I/O Blade architectures provide the tight coupling needed for compute power, I/O connectivity, and switching that provides an adaptive environment that is reliable (fully redundant), easy to manage, flexible, available, scalable and has the performance that IT managers demand from data centers. A blade design must include the InfiniBand architecture to provide the IT manager with an adaptive data center environment that scales with the future, and provides the maximum return on the investment.

18.0 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.