Mellanox Technologies

White Paper

PICMG 3.2 Advanced Telecommunications and Computing Architecture

InfiniBand Technology Delivers Performance, Scalability, and Fault Tolerance for Next Generation Platforms

1.0 Introduction

The Advanced Telecommunications Architecture (ATCA) defines an open switch fabric based platform delivering an industry standard high performance, fault tolerant, and scalable solution for next generation telecommunications and data center equipment. The ATCA standard is expected to be finalized in mid 2002 and the development is being carried out within the PCI Industrial Computer Manufacturers Group (PICMG) - the same group that created the highly successful Compact PCI standard. The Advanced TCA 3.0 base specification defines the physical and electrical characteristics of an off the shelf, modular chassis based on switch fabric connections between hot swappable blades. The Advanced TCA base specification supports multiple fabric connections including InfiniBand technology as defined by the PICMG 3.2 sub specification. This white paper provides an overview of the ATCA base specification as well as details of how InfiniBand technology delivers performance, scalability, and fault tolerance for next generation carrier grade telecommunications and computing platforms.

2.0 Why a new Architecture? (Business Justification)

During the 1990's the Telecommunications market experienced two extremely favorable market factors of quickly increasing bandwidth requirements combined with fast market growth. These factors were great for the industry but led to a hodgepodge of proprietary standards designed to meet the immediate market needs that each company was working to service. Despite the advent of the highly successful Compact PCI standard, the fact remains that the majority of the telecommunications market has no standard form factor, backplane or fabric interconnect that can meet

the 10Gb/sec bandwidth requirements of today. There continues to be tens, if not hundreds, of different chassis designs in the telecommunication industry that drive the cost of this equipment higher, prevents multi-sourcing and interoperability across a common backplane or fabric.

Significant changes in the telecommunications market have occurred in the last few years with the meteoric growth, leveling off and competition increasing accordingly. In this cost-cutting environment of constrained budgets it is more important than ever for telecommunications equipment providers to leverage off the shelf components and sub-systems, thereby minimizing investment and maximizing the breadth of product environment. The Advanced TCA platform is particularly attractive in this environment since OEM's can minimize their own and investments and yet by leveraging off-the-shelf sub-systems, can actually outstrip the competition in the performance, and breadth of product offerings.

In parallel the server market has made changes that mirror those of the telecommunications industry, offering a chassis with centralized power and field replaceable units (FRUs). The server vendors generically call this modular computing architecture: "server blades". Today all the major server OEMs have announced plans for server blades, but each are moving toward proprietary designs rather than converging on a single common form factor. Although server blades will lower IT costs, again there will be a hodgepodge of blade form factors that will not provide second sourcing opportunities for the IT manager.

With the advent of the InfiniBand architecture there is now the opportunity for Telecommunications and computing to share an industry standard, high bandwidth, low latency interconnect fabric. Delivering this in a single, cost effective industry standard design that speeds time to market will allow racks to be seamlessly linked together with an InfiniBand subnet. This provides Data Center and IT managers with the peace of mind brought by utilizing industry standards that help to enable lower costs, faster time to market, interoperability and multiple sources for FRUs.

3.0 Next Generations Platforms Demand Switch Fabric Technology

High speed fiber, ATM, IP, and DSL broadband connections to data centers and central offices have increased the bandwidth requirements beyond a few gigabits/second that can be supported by shared bus architectures such as CompactPCI. Next generation platforms require aggregate bandwidth that can scale to 2.5 Terabits/second with individual interfaces supporting upwards of 40 gigabits/second. System availability requirements in excess of 99.999% remains a core requirement. Furthermore the broad range of interfaces and services required by such platforms greatly benefit from an open architecture enabling multi-vendor support with the ability to deliver on the promise of converged voice, data, server, storage, video, and wireless functions.

It is vital that the basic platform architecture is able to scale to meet these increasing demands for scalability, bandwidth, availability and performance. Delivering high availability and fault tolerance requires redundancy to be designed into every subsystem including the I/O chassis itself. Not surprisingly the shared bus architecture (originally designed as a local interconnect) is simply not capable of meeting these requirements. In order to overcome these limitations switch fabric architectures have emerged where each I/O input can make temporary connections to any of the I/O outputs. These connections are made through sophisticated switches which can offer advanced

features including quality of service, flow control, integrated management, fault recovery, and scalability.

3.1 ATCA Base Architecture

The ATCA 3.0 base specification defines the frame (rack) and shelf (chassis) form factors, core backplane fabric connectivity, power, cooling, management interfaces, and the electromechanical specifications of the boards. The electromechanical specification is based on the existing IEC60297 EuroCard form factor enabling equipment from different vendors to be incorporated in a modular fashion and be guaranteed to interoperate.

ATCA Terminology

The base specification uses the following semantics to describe elements of the ATCA platform:

Boards: The individual I/O or computing blades that are hot-insertable and removable in a *shelf*

Shelf: The 12U tall chassis that provides power, cooling, backplane connectivity, and the slots to accept up to 16 *boards*

Frame: A rack (typically 46U high) provides a rigid framework accepting up to 3 *shelves* and commonly deployed within enterprise and Internet data centers and telco central offices.

3.1.1 Passive Backplane and Switch Fabric Connectivity

The ATCA defines a passive backplane and details the location of high speed fabric connectors, management connections, power connectors, alignment keys, as well as electrical keying to define the type of cards plugged into the chassis. The backplane is entirely passive and delivers high levels of reliability as there are no active components to fail. The backplane defines the overall PCB board layout and connector location, high speed and power connectors, alignment key structures for mechanical integrity, and electrical keying to support different types of boards. The ATCA specification also defines the backplane connectivity which can support either a full mesh architecture or a dual star connection.

In the full mesh architecture (shown in Figure 1) every board is identical and connected to every other board in the shelf. By using non-blocking switches all boards can be simultaneously communicating with each other. Utilizing 4X (10Gigabit/sec) InfiniBand connections in the configuration shown with 14 boards, the backplane supports an aggregate bandwidth of 1.8 Terabits/sec full duplex (3.6 Terabits/sec total bidirectional bandwidth). A backplane supporting 16 boards would provide about 2.1 terabits/sec.

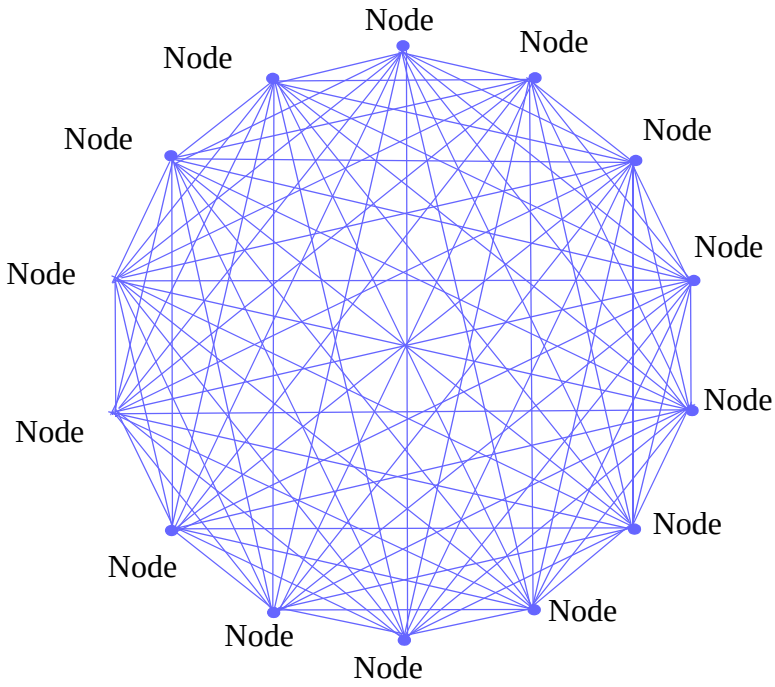


Figure 1. Full Mesh Topology

The dual star topology (shown in Figure 2) specifies two "fabric" (or switch) nodes which connect to all of the other slots. Thus, unlike the mesh configuration, with the dual star architecture there are two different types of boards (and backplane slots): fabric boards and node boards. The configuration shown has two fabric boards and 14 node boards. An implementation of this topology based on InfiniBand 10Gb/s links would provide 140Gb/s raw bandwidth.

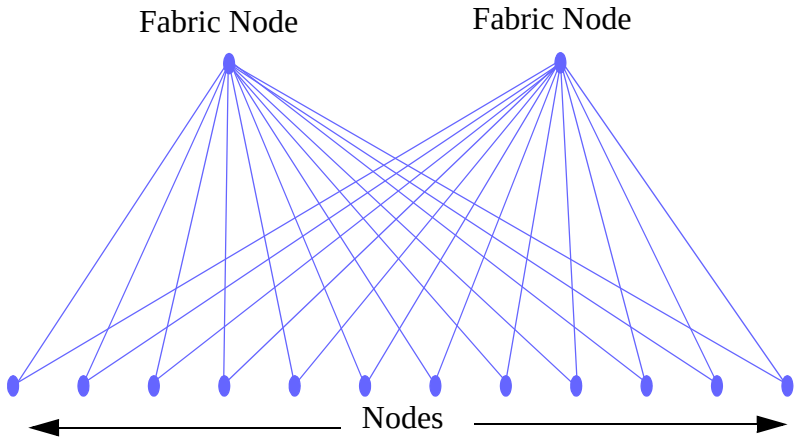


Figure 2. Dual Star Topology

Since the dual star topology is a subset of the full mesh connection, a backplane providing full mesh connectivity can be used in a dual star topology simply by plugging fabric boards into the two designated fabric slots and node boards into the remaining slots. Utilizing two fabric boards eliminates any single point of failure and enables fault tolerance and high availability, without incurring the complexity and expense of full mesh applications.

The ATCA specification requires the backplane to support the dual star topology at a minimum and may optionally support dual-dual star or full mesh topologies. The specification utilizes the Tyco/Erni HM-ZD 2.0mm high speed connector in a 4x10 differential pair configuration.

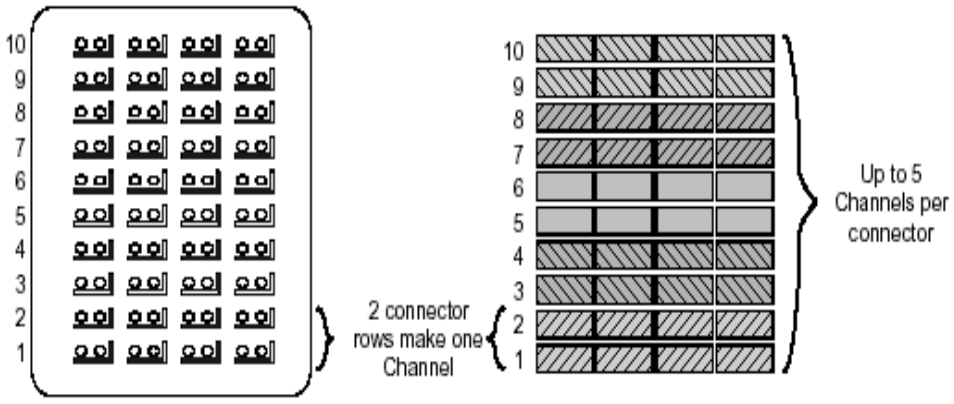


Figure 3. Front View of Backplane ZD Connector and Diff Pair Assignment

There are five such connectors supporting up to 200 signal pairs. Each signal pair includes a differential pair plus a dedicated ‘L’ shaped ground shield. Eight of these signal pairs define a “channel” which is the unit of connectivity between nodes in the ATCA fabric. Figure 3 shows the channel organization within the connector. InfiniBand offers forward and backward compatibility since a single channel can operate as a four 1X (2.5Gb/s) link or a single 4X (10Gb/s) link.

3.1.2 Boards

The ATCA board (shown in Figure 4) has dimensions 8U tall (14 in/355.6mm) by 280mm deep (11.02 in) and 6HP wide (1.2”/30.48mm). Boards may be any integer multiple of the basic 6HP width. The 8U board height allows up to 4 standard PMC cards to be accessed via the front panel. The board has approximately 140 square inches of area allowing multi processor designs accommodating large amounts of memory. The back side of the board has power connectors, alignment keys, rear I/O access, and high speed connections to the passive backplane.

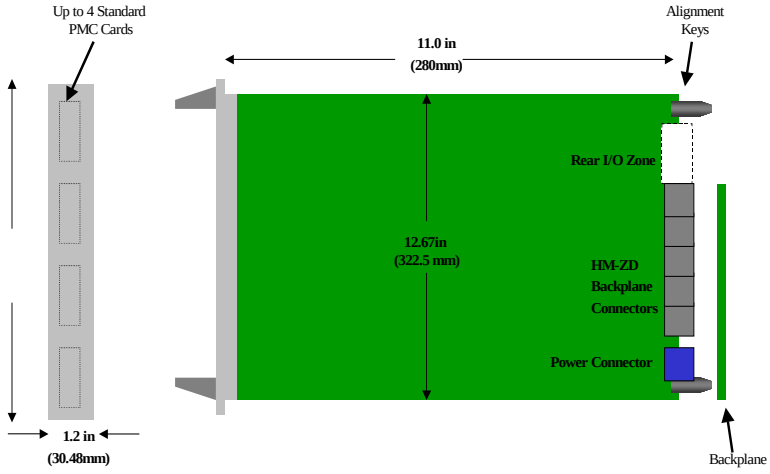


Figure 4. ATCA 3.X Board Diagram

3.1.3 Shelf Specification

Figure 5. ATCA 3.0 Shelf Prototype
(photo courtesy Rittal Corporation)

The ATCA shelf form factor defines a chassis which is a total of 12U high (21 inches) supporting 8U high boards, an 1.75U air plenum below, and 2.25U space above for redundant hot swappable synchronized blowers. Figure 5 shows a prototype of ATCA 3.X shelf illustrating blowers, backplane, air plenum, etc. Taking into account the sidewalls a standard 19 inch shelf can support up to 14 boards, while a 23 inch shelf supports up to 17 boards. The 12U shelf height allows three shelves in a standard 42U frame and still leaves 6U for power conditioning panels.

3.1.4 Frame Specification

The frame (or rack) specification includes requirements to insure NEBS, ETSI, and IEC compliance. The ATCA specification does not provide frame mechanicals but instead supports existing standards including:

- ETSI 600mm 600mm
- Standard 19” and 23” frames
- Universal Telecom Framework
- NEBS 2000
- Universal Telecom Framework IEC

These are defined in ETSI IEC 300-119-(1-4), EIA 310-D, IEC 60297-(1-2), T1E1.8 Project 41, and GR-063-CORE specifications. These frame form factors are all readily available off-the-shelf from multiple vendors.

3.1.5 Power

The base specification defines a power budget of 200 Watts per board enabling high performance servers with multi processor architectures and multi gigabytes on-board memory. The frame power is delivered by redundant -48 VDC feeds. These dual frame power feeds are typically fused and multiple sub feeds generated allowing each shelf to remain electrically isolated. Local DC-DC conversion is accomplished per board. Redundant local power feeds are normally attached through either diode *or'ed* connections to a single on board DC-DC converter (shown in Figure 6) or via on-board dual redundant load sharing DC-DC converters (not shown).

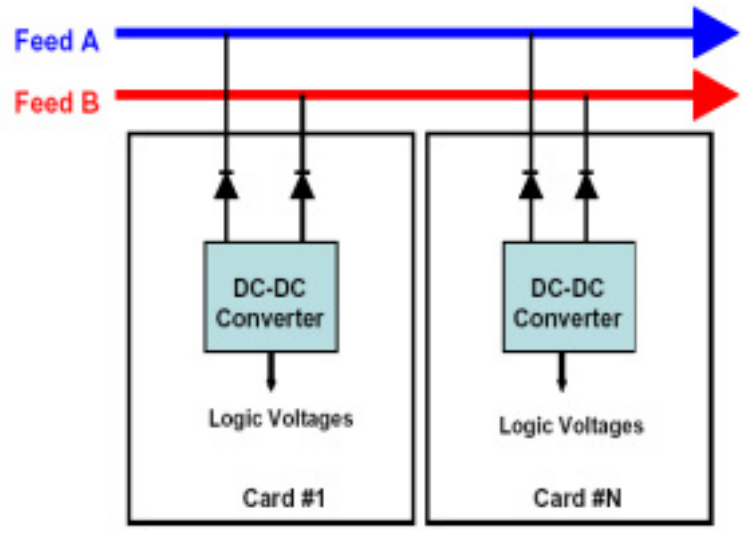
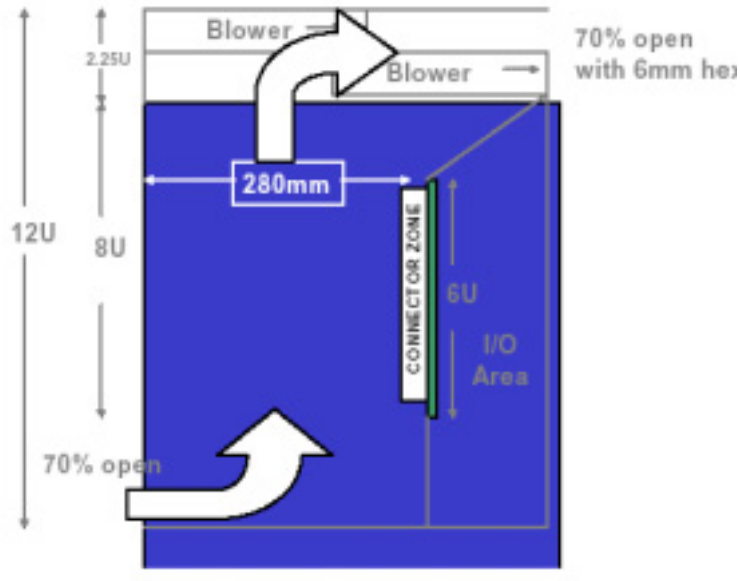


Figure 6. Diode Or'ing of Feeds to Single Card DC to DC Converter

3.1.6 Cooling

The PICMG organization has performed extensive thermal modeling in order to design the ATCA board and shelf form factors to be able to support 200W power dissipation per board. The shelf uses conventional air cooling with blowers pulling air from front to rear and bottom to top. Mechanical fans are typically the element with the lowest MTBF and thus thermal design must incorporate sufficient overhead to accommodate a failed blower. Blowers may be synchronized and include temperature controlled air flow to reduce audible levels.



4.0 Advantages of InfiniBand Technology

The ATCA 3.0 base standard is fabric agnostic and supports various options for I/O fabrics including InfiniBand, Ethernet, and StarFabric technologies. While multiple fabrics are envisioned by the specification it is likely that InfiniBand technology will emerge as the dominant general purpose I/O fabric, while other technologies will address special purpose applications. This is mainly due to the four key advantages that InfiniBand offers over other technologies:

- **Aggregation**: 2.5Gb/s, 10Gb/s, and 30 Gb/s links compatible on the same backplane
- **Hardware Transport:** Eliminates the CPU burden of a software transport stack (i.e. TCP/IP)
- **Low Latency:** Round trip latencies in the microseconds
- **I/O Sharing**: Storage, WAN, and LAN interfaces can be shared across the fabric by all servers

These features and others make InfiniBand the ideal solution for next generation equipment as it delivers high performance, fault tolerance, quality of service, and transport level connections with the highest levels of integration available. The InfiniBand architecture greatly accelerates data movement and offload the CPU from transport processing. Powerful layer 2 features of the InfiniBand Architecture (such as virtual lanes and link level flow control) result in significant advantages over un-reliable link technologies which use dropped packets as a form of implicit congestion notification to higher software levels. Another powerful feature of the InfiniBand Architecture is Automatic Path Migration which enables the fabric to detect errors and fail-over to a pre-defined alternate path. This ability of the fabric to heal itself is particularly critical, where low latency fault recovery is important, such as in many applications supporting real time traffic (voice, video, etc).

InfiniBand technology is architected as an I/O fabric and does not carry the overhead required of a LAN technology, such as a complex software protocol stack necessary to insure reliable delivery. InfiniBand is optimized for backplane and short reach applications (~30 inches on standard FR4 PC board backplanes and up to 17 meters over copper cables) and thus, is not burdened by the higher power and complex signalling requirements of LAN technologies which need to span to distances greater than 100 meters. InfiniBand offers multiple protocols to carry IP based traffic (IPoverIB, Sockets Direct Protocol, RNDIS, etc) as well as protocols for transporting storage and other cell and packet based standards. In fact, one of the real strengths of InfiniBand is its ability to support a multi-protocol environment while still providing quality of service and latency assurances.

Table 1 further lists some of the advantages that the InfiniBand Architecture offers:

Table 1. Key Advantages of InfiniBand Architecture

Feature	InfiniBand Architecture	Ethernet
Defined and Approved Specification	2.5, 10, and 30 Gb/sec	1Gb/s (10Gb/s not yet finalized)
HCA/NIC Availability as of Q4/01	2.5 & 10Gb/s	1Gb/s only
Link Compatibility to 10Gb/s	Yes: Compatible 2.5 - 10Gb/s	No: 10/100/1000BT inompatible with 10Gb/s Ethernet
10Gb/sec Copper Ports	Yes	No (optical only currently being defined)

Table 1. Key Advantages of InfiniBand Architecture (Continued)

Feature	InfiniBand Architecture	Ethernet
Link Level Flow Control	Round robin credit based with priority	xon/xoff
Multiple Layer 2 virtual fabrics with Quality of Service	Yes: Multiple Virtual Lanes with independent flow control	No: Class of Service shares single xon/xoff flow control
Reliable Link	Yes	No
Hardware Transport Connections	Yes	No
Kernel Bypass	Yes	No
Remote Direct Memory Addressing (RDMA) Capabilities	Yes	No
I/O Sharing Capabilities	Yes	No
Hardware Segmentation and Re-assembly	Yes	No
Hardware End to End Flow Control	Yes	No
In-Band Management	Yes	No
Automatic Path Migration	Yes	No
Switches Available with Integrated Physical layer	Yes: 32 @ 2.5Gb/s	No, Multiple switch and PHY devices required

5.0 Summary

The Advanced TCA 3.2 architecture defines a next generation platform for high performance, fault-tolerant, scalable, telecommunications and computing equipment. The open platform defined by the PICMG 3.0 specifications is expected to be finalized mid 2002 and will provide a framework for multiple protocols, but it is clear that only the 3.2 InfiniBand specification provides the bandwidth (10Gb/s today and 30Gb/s in 2003), scalability, availability, performance, and cost advantages to meet the needs of the telecommunications market of the future.

6.0 About Mellanox

Mellanox is the leading supplier of InfiniBand semiconductors, providing switches and channel adapters to the server, communications and data storage markets. In January 2001, Mellanox Technologies delivered the InfiniBridge MT21108, the first 1X/4X InfiniBand device to market, and is now shipping second-generation InfiniScale silicon. The company has raised more than $89 million to date and has strong corporate and venture backing from Bessemer Venture Partners, Dell Computer, Intel Capital, Raza Venture Management, Sequoia Capital, Sun Microsystems, US Venture Partners, Vitesse and others. Mellanox currently has more than 200 employees in multiple sites worldwide. The company's business operations are headquartered in Santa Clara, CA; with the design, engineering, software, system validation, and quality operations based in Israel. For more information on Mellanox, visit www.mellanox.com.