AdvancedTCA™

PICMG® 3.2 R 1.0

InfiniBand® for AdvancedTCA™ Systems Short Form Specification

January 22, 2003

FOR INFORMATION ONLY; DO NOT ATTEMPT TO DESIGN FROM THIS DOCUMENT

NOTE: This short form specification is a subset of the InfiniBand® for AdvancedTCA™ Systems specification, PICMG 3.2 R 1.0. For complete guidelines on the design of InfiniBand® for AdvancedTCA™ implementations, the full specification is required.

To order a full copy of the PICMG 3.2 specification, go to www.picmg.org, or contact the PCI Industrial Computer Manufacturers Group at 401 Edgewater Place, Suite 500, Wakefield, Mass., 01880. Phone 781-246-9318, fax 781-224-1239, email info@picmg.org.

Introduction

The complete PICMG 3.2 Specification is a member of the PICMG 3.0 AdvancedTCA™ series of specifications. The base specification in the family, PICMG 3.0, defines a Board and Shelf architecture for modular computing and communications components. The Boards share a common Backplane for point-to-point differential connectivity that supports many common high-speed switched fabric technologies. PICMG 3.2 specifies the usage of InfiniBand® over this backplane fabric.

Scope

The scope of this effort includes design rules and guidelines for implementation of InfiniBand Boards based on the PICMG 3.0 specification.

Objectives

The specification in the full 3.2 document is wholly derived from, and is dependent upon, the PICMG 3.0 Base Specification. It is not intended to stand alone or to be used separately from the Base Specification.

The complete PICMG 3.2 document builds upon the PICMG 3.0 Base Specification and the InfiniBand Specifications to meet the following objectives:

- Define how InfiniBand signaling is to be used over the Fabric Interface defined in the PICMG 3.0 Specification
- Provide guidelines for the use of 1X, 4X, and 12X InfiniBand links
- Provide guidelines for the use of InfiniBand in-band management
- Provide recommendations and guidelines for utilization of advanced InfiniBand features that increase performance and interoperability

The guidelines and design rules set forth in the complete specification shall be consistent with the applicable sections of the InfiniBand Specification version as referenced in Section 1.3 of the complete 3.2 document. PICMG 3.2 products will comply with InfiniBand Signal, Link, Transport, and Management layers in order to maximize interoperability with other InfiniBand hardware and software products.

Overview

The complete PICMG 3.2 document specifies an open architecture for modular computing and communications components based on InfiniBand technology. The full PICMG 3.2 is a subsidiary of the PICMG 3.0 Base Specification, and together they specify an InfiniBand-based architecture for converging communications and data networking applications. There are several advantages to using InfiniBand as a backplane fabric:

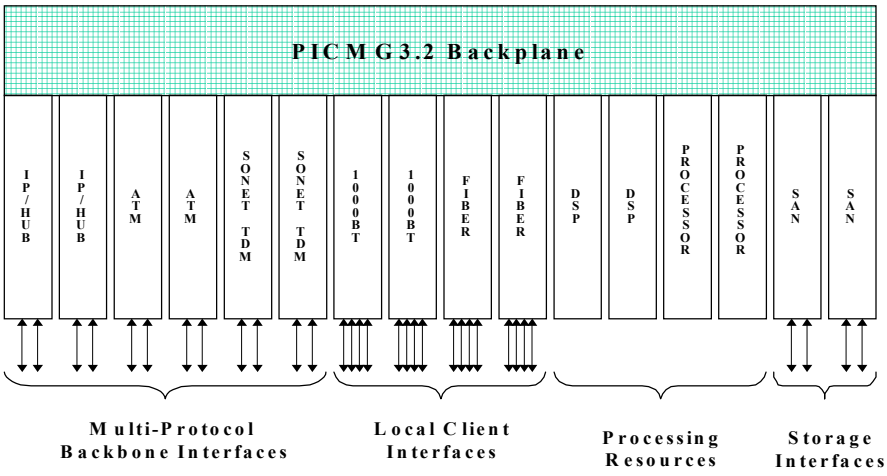
- Improved backplane capacity: InfiniBand supports 10Gb/s bi-directional connections today with standards-based silicon. This reaches a capacity of 2.4Tb/s in a single PICMG 3.2 Shelf using 16 Boards in a full mesh backplane topology.
- Improved Quality of Service: The virtual lane and service level mechanisms of InfiniBand allow the strict segregation of different classes of traffic. This permits the creation of equipment that supports different applications without risk of lower priority traffic delaying higher priority traffic.
- Hardware Transport: InfiniBand provides reliable connections in hardware which eliminates the software overhead of a transport protocol stack. This not only enables low-latency and high throughput for server clustering and storage, but enables simplified management and multi-protocol support for real time traffic.
- Multi-protocol support: Since many different protocols can be easily mapped onto and transported by InfiniBand, a PICMG 3.2 system can support IP, ATM, TDM, data acquisition protocols, and mixed protocol systems.
- Integration of servers and storage: InfiniBand was developed as a standard RDMA-based clustering interconnect for servers and Storage Area Networks (SANs).
- Simplified multi-Shelf systems: The same simplified capabilities for interconnect with servers and networked storage also facilitate the interconnection of multiple PICMG 3.2 shelves with InfiniBand cables to create very large, multi-Shelf systems that behave as single logical systems.
- Improved management: The extensive network management capabilities available in InfiniBand work in concert with the PICMG 3.0 Shelf Management infrastructure to provide a simplified and highly available management framework.

These advantages bring significant improvements to the development of telecommunications and data processing systems.

Application Example

Figure 1 shows an example application of PICMG 3.2 to a multi-service telecom / datacom network element intended for deployment in a telecommunications central office, data center, or workgroup machine room. Its functions include uplinks to several backbone network types (IP, ATM and SONET / TDM), several different user interfaces (copper Ethernet and fiber), signal processing, server processors, and storage management.

Figure 1. Shelf Level Block Diagram



Eight types of Boards are deployed in a redundant manner to enable high availability and reliability. Boards that have special features and performance requirements would likely be custom designed by a Telecommunications Equipment Manufacturer (TEM). However, many of the Boards will be available off-the-shelf from commercial vendors in the computing and telecommunications industry. Server, storage, and Ethernet Boards, for example, will be widely available for enterprise and datacenter applications. This offers a significant time-to-market and cost savings advantages to the TEM.

The following briefly describes the different Board types:

1. The IP/HUB Boards provide switching and routing between an external IP WAN and the InfiniBand fabric. A typical configuration might include an InfiniBand switch for 15 backplane ports with multiple Gigabit Ethernet ports on the face plate. Wire speed packet processing between Ethernet and InfiniBand would support the IP-over-IB protocol for routing IP to other Boards in the system, and the Sockets Direct Protocol for terminating the connection to an internal InfiniBand server or DSP Board. Firewall and security processing can also be supported by on-board processors or other processor Boards in the system. If a Dual Star Topology is used, this Board would also provide the PICMG 3.0 Hub functionality.

2. The ATM Boards adapt the InfiniBand protocol found on the Backplane to ATM, and provide ATM interfaces (probably mapped over SONET) to ATM core networks.

3. SONET/TDM Boards support Time Division Multiplexed, channelized traffic such as voice trunks. They include the protocol conversion logic needed to map timeslots to the InfiniBand Backplane interconnect. Voice traffic may be sent to other voice Boards for call switching functions, or may be sent to the DSP or server Boards for VOIP packet processing and messaging applications.

4. The 1000BASE-T local client interface Boards support LAN/MAN traffic to local destinations such as customer premise last mile or desktop computers within the building. The architecture might be similar to the IP/HUB interface Board, except without the firewall and security functions.

5. Fiber interface Boards perform a similar functions as the 1000BASE-T interfaces except for the use of fiber to provide longer reach for applications like Fiber to the Home or Metropolitan Optical Networks.

6. The DSP Boards provide signal processing capacity for applications requiring vocoding, compression, encryption, conference bridging, and video encoding.

7. The processor Boards provide standard general-purpose RISC or CISC computers that are often used as servers. They can provide high performance database applications as well as run call processing, protocol processing, or network management applications.

8. Finally, the SAN Boards are the interfaces to Storage Area Networks. These might be external disk array shelves providing many terabytes of storage using InfiniBand's ability to bridge easily to Fibre Channel. If InfiniBand is used to interconnect the disk arrays with the PICMG 3.2 Shelf, the SAN Boards have very simple designs (since there is no need to adapt the native InfiniBand protocol from the Backplane). An alternative implementation could mount several low profile disk drives directly on the Boards, providing capacity of hundreds of gigabytes integral to the PICMG 3.2 Shelf. Boards of similar architecture to the SAN Board provide Shelf-to-Shelf IB links supporting multi-Shelf systems.

Topologies and Bandwidth

The PICMG 3.2 Backplane is the core of this reference architecture and offers a choice of topologies, link speeds, protocols, and services supported on the system.

The two most common fabric topology choices are Dual Star and Mesh. A dual star system uses two Hub Boards in Logical Slots 1 and 2 (labeled IP/HUB in Figure 1) and Node Boards (each with two InfiniBand links) installed in the remaining 14 slots. A 16 Slot Dual-Star System (14 Node Slots) using 4X InfiniBand links uses 28 links to create the fabric, or a total bandwidth of [28 * 10Gb/s =] 280Gb/s. Using 1X links, this mesh provides a raw capacity of [28 * 2.5Gb/s =] 70Gb/s.

A Mesh topology utilizes all available links in all its slots. Mesh Enabled Boards have InfiniBand switching capacity for all possible links. For example a Mesh Enabled Board in a 16 slot system will have 15 InfiniBand links to the Backplane (one to each of the other Boards). A 4X Mesh reaches the theoretical maximum capacity of [16 * 15 * 10Gb/s =] 2.4Tb/s. A 1X Mesh is capable of [16 boards * 15 links * 2.5Gb/s =] 600Gb/s theoretical maximum capacity.

InfiniBand uses 8B/10B encoding, therefore usable bandwidth is 80% of the above quoted line rates. Also, usable payload bandwidth will have to be derated to account for the InfiniBand protocol overhead.

PICMG 3.0 Backplanes also include two Base Interface Channels, three Synchronization Clock Interfaces, two IPMB channels, and the Update Channel Interface to all slots for use by advanced telecommunications Boards. These features will not be discussed in detail in this document, but PICMG 3.2 systems can make use of them as needed.

Board Design Example

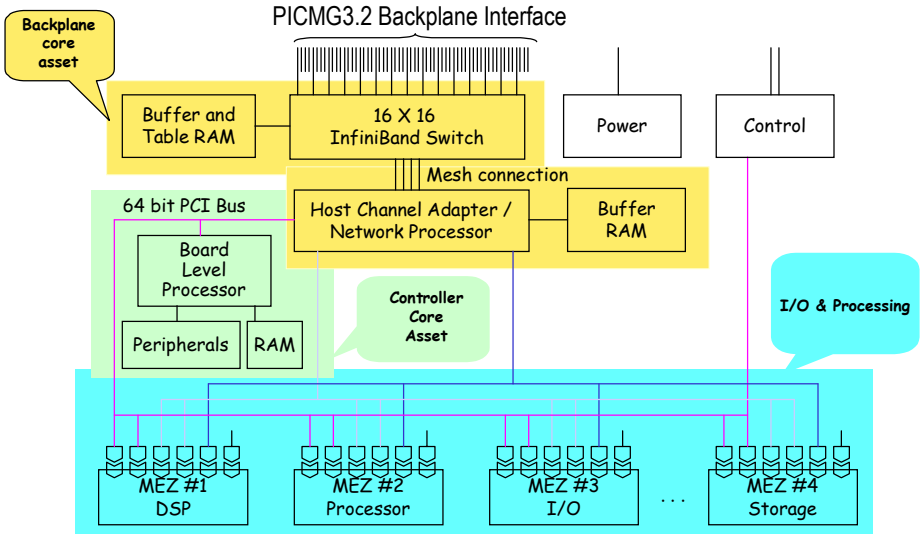
Figure 2 is a block diagram of a PICMG 3.2 Board using a 1X mesh topology capable of supporting 2.5Gb/s of raw throughput to each of 15 other Boards. This design is a superset of the designs required for the Hub Boards or Node Boards needed with a Dual Star Topology. A four mezzanine design is shown to illustrate the flexibility of supporting multiple kinds of I/O and processing resources on a universal board type. The design would be similar if the mezzanines were deleted and the prime functions (processors, DSPs, I/O, etc.) were mounted directly on the Board.

Dual Star Node Boards would be similar in design to this example, but simplified, as the InfiniBand switch function and much of the network processing capability can be deleted. Two Host Channel Adapter functions are required on Dual Star Node Boards, one for each link to the duplicated Hub Boards.

Dual Star Hub Boards would also be similar to this example, but their emphasis would be on the InfiniBand Switch and Network Processor functionality. The Controller Core Asset and I/O and Processing blocks are optional in the case of Hub Boards, and are needed only in high density systems that provide Board functionality beyond simple hub functions on Logical Slots 1 and 2.

The Mesh reference design can also be extended to support 4X and 12X InfiniBand links by using appropriate InfiniBand switch components with higher capacity Backplane connections, and increasing the Host Channel Adapter and network processing throughput accordingly. Also, using a 12X links requires that the Backplane routing supports replicated meshes, per Section 3.1.1.1.

Figure 2. Example PICMG 3.2 Mesh Board Design



The PICMG 3.2 Backplane traffic enters a 16 X 16 InfiniBand switch component that performs the distributed switching functions for the Shelf. Its primary purpose is to select and direct the backplane traffic that is to be processed by this Board. If this Board acts as the switch Hub for a dual star backplane topology (installed in logical slot 1 or 2), this switch component represents the main Hub function for an entire Shelf. Because of the focused load on the sixteenth link between the switch and the Host Channel Adapter, this link should support higher traffic capacity. For a Dual Star topology this link may be scaled down depending on the outgoing traffic requirements. Since this interface often represents the system level bandwidth bottleneck for a PICMG 3.2 Shelf, it is advisable to make it as high a capacity as practical.

Next in the signal processing chain is an InfiniBand Host Channel Adapter function. This terminates the InfiniBand protocol from the backplane links, and makes the payload available to the rest of the Board.

A network processor function is associated with the HCA. Its purpose is to perform protocol adaptation, routing, and quality of service operations. The network processor throughput also can represent a capacity bottleneck for PICMG 3.2 systems, so care must be taken to insure it has adequate throughput. In some cases, the network processor functions are not required, and the HCA is adequate for the interface.

A board level processor, labeled the Controller Core Asset, provides board level control functions. It includes a microprocessor, RAM and peripherals. This complex will often run call processing, bandwidth allocation, slowpath data switching functions, and network management applications software.

Finally, the four mezzanine board positions permit the installation of an appropriate mix of mezzanine boards onto this carrier Board as required to deliver its intended functionality. A mix of DSP, processor, I/O, and Storage mezzanine boards are shown in Figure 2, but in practice, all four slots may be equipped with the same board type, or

different types than those shown could be required. One great advantage of the mezzanine board approach is its improved modularity. PICMG 3.2 has a maximum of sixteen Board slots, and that often is not quite enough to permit the exact customization of the Shelf's functions needed for a particular application. By using a mezzanine rich architecture as shown in this reference design, up to 64 mezzanine positions are available per 16 slot Shelf, providing much more modularity. Exactly the required complement of processing, DSP, I/O and storage are installed in the mezzanine position to meet the needs, and no more, leading to very dense, cost effective systems.

###