



# Ethernet Interconnects in ATCA Systems

Enabled by Low Latency Switching

**White Paper**

---

*September, 2005*

**Notice:** This document contains preliminary information on new products. The information is subject to change without notice.

**Order Number:** WP ATCA CM

## Overview

High performance embedded systems form the basis for telecommunications and networking equipment as well as for enterprise server applications, storage and high performance computing platforms. Independently designed and optimized for their respective applications, these embedded systems have historically been expensive. The common drive toward cost reduction across all these markets has led toward standardization in the physical and electrical aspects of system design. One of the leading examples of this movement is the ATCA (Advanced Telecom Computing Architecture) specification developed by PICMG (the PCI Industrial Computer Manufacturers Group). In the ATCA specification the backplane fabric interconnect however, remains open to a number of standards, a situation that has led to a splintering of the ecosystem, in conflict with the basic principles of driving toward commoditization through uniformity. This activity has inhibited the ATCA industry growth.

This white paper makes the case that among the candidate fabric technologies Ethernet, with its continually evolving combination of features and performance, is the most consistent with the goals of the ATCA specification and will emerge as the one right interconnect. The unique attributes of Ethernet switch chips soon to enter the market from Fulcrum Microsystems, Inc. will speed the adoption of Ethernet by improving congestion management, its Achilles' heel.

## ATCA

### Driving Toward an Architecture Standard

In the continuing effort to lower the cost of communications equipment, Telecommunications Equipment Manufacturers (TEMs) are migrating away from proprietary designs toward the use of standards-based designs. Standardization can be implemented at several levels in the system's design while preserving the ability to differentiate through software and services. Standards at the level of physical layer protocols, silicon, blades, chassis mechanical characteristics, power, system management and software have all been leveraged to lower the costs of system design, manufacture, maintenance and application development.

The next logical step has been taken by the PICMG in the development of the ATCA standard, perhaps better described as a unified set of standards, forming the basis for an entire system platform. Targeted primarily at carrier grade applications, PICMG 3.0 is the foundational specification for systems featuring scalable capacity, five 9's reliability, manageability, high availability, modularity and serviceability. A well thought out standard attracts a multitude of component vendors creating a rich ecosystem of building blocks from which system designers may choose. Competition among vendors and economies of scale in manufacturing cause price reduction to quickly ensue.

The PICMG 3.0 specification for ATCA systems addresses numerous aspects of an open architecture modular platform: mechanical design, shelf management, power distribution, thermal management, and data transport. Figure 1 gives a physical overview of a basic system.

In order to achieve broad market applicability, elements of the architecture have not been rigidly specified, but left flexible enough to accommodate a variety of usage models. One key aspect of the architecture on which the specification remains agnostic is the data interconnect, or fabric, technology. PICMG 3.0 defines the electrical signaling and

physical layout parameters for a shelf of up to 16 slots containing node boards and hub (switch) boards. The specification provides for four kinds of data transport among boards. A base interface is used primarily for the exchange of management information, although it may also be used for data transport. An update interface is used for redundancy between boards or for specialized fabric bypass. A clock synchronization interface can be used for sharing synchronous clock information throughout the shelf. The data transport fabric provides for the exchange of high volumes of data and ultimately defines the system capacity. It is this last interconnect type that remains open to a number of different data transmission protocols.

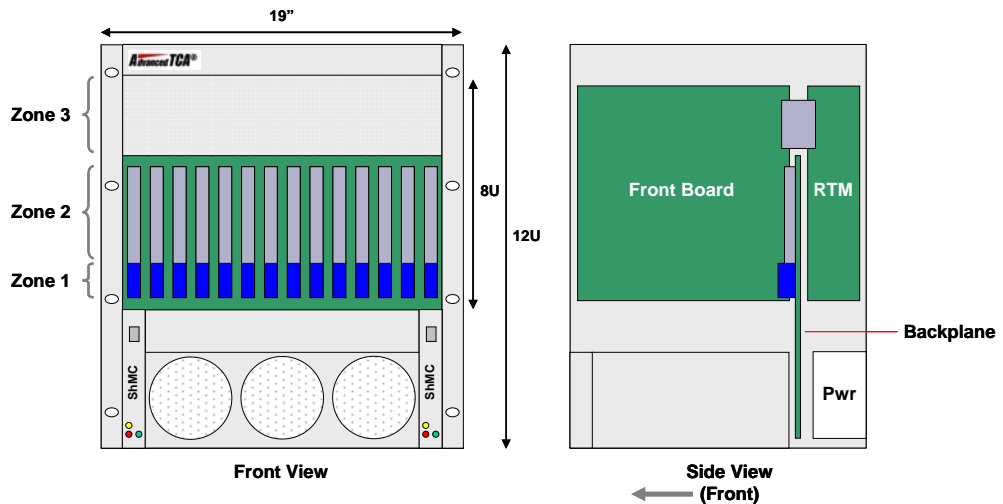


Figure 1: Elements of the ATCA modular architecture.

## Unintentional Fragmentation

Ideally, one interconnect technology would have all the attributes required for use across the wide range of applications targeted by ATCA systems - equipment spanning edge, core, transport and data center applications. While the base system fabric is defined to be 10/100/1000BASE-T Ethernet, five subsidiary specifications, PICMG 3.1 through PICMG 3.5, define the use of Ethernet/Fibre Channel, Infiniband, StarFabric, Advanced Switching and Serial Rapid IO protocols, respectively, for the critical data fabric interconnect. Moreover, there are potential additions to this list as additional proposals to PICMG are received.

Each protocol has a set of advantages and disadvantages associated with it, and the PICMG 3.0 specification allows the designer to choose from among the available options the one that best matches his particular set of requirements. Switching elements and fabric interface chips (FIC) on node and hub cards implement the interconnect protocol and the cards are therefore not interchangeable with other systems using different protocols. While the existence of numerous interconnect protocols does address the issue of broad applicability, it also has the undesirable effect of fragmenting the market for ATCA systems, thereby inhibiting the development of a robust component ecosystem, consequently slowing its market acceptance and increasing cost. Clearly, the existence of one single interconnect with all the functionality required would accelerate the adoption of the ATCA specification.

## Ethernet as a Fabric Interconnect

### The Case for Ethernet

Due largely to its ubiquity, Ethernet offers compelling features that are consistent with the drive toward low cost system architectures. Initially developed as a LAN/WAN networking protocol, it has been successfully applied as a box-to-box interconnect and as a backplane interconnect, bringing it to a pre-eminent position in terms of number of installed network connections. This adaptability is a direct result of the formation of working groups within the IEEE 802 LAN/MAN Standards Committee for the purpose of creating extensions to the standard. Its very large installed base has resulted in the highest level of software compatibility, the lowest implementation and support cost, and the broadest applicability of all the combined interconnect technologies. If Ethernet adequately addressed all the critical functional requirements of ATCA switch fabrics, it would be the obvious choice to standardize upon, but there is one area where Ethernet is perceived to be deficient.

### Congestion Management in Ethernet

ATCA-based systems are being designed for demanding compute and communications applications, requiring blade to blade movement of vast amounts of data. Data transmission through the backplane switch fabric must support carrier-class performance, in turn requiring lossless transmission, flow control and congestion management, essential elements in enabling Quality of Service (QoS). It is in this related set of requirements where Ethernet is seen as deficient.

Presently, Ethernet's fundamental mechanism for providing congestion management (other than frame discard, which is unacceptable) is the PAUSE function provided by IEEE 802.3x. This mechanism provides XON/XOFF functionality via the transmission of MAC control frames in the upstream direction, from a receiving port experiencing congestion to its transmitting partner originating the congesting traffic. The reception of PAUSE control frames causes the transmitting device to halt transmission for a time period specified within the frame – enough time for the congested device to relieve its buffer overflow. It was conceived as a link layer flow control mechanism and can be effective when applied to a pair of directly connected devices and for transitory congestion, where peak traffic conditions exist for short time periods and can be buffered.

However, IEEE 802.3x flow control as currently defined does not distinguish between multiple data flows that use the same path, meaning all data flows through a path are paused when only one flow is congested. Because of this, when applied to more complex data paths involving traffic from multiple end points to multiple end points through a switch, for example, this mechanism may cause head-of-line (HOL) blocking, ultimately resulting in congestion spreading. Whereas 802.3x PAUSE operation can be effectively used on a link level, it is not deemed effective for this end to end congestion management.

For this reason, 802.3x flow control has not been widely implemented. Class-based congestion management mechanisms, those that can distinguish between multiple data flows sharing a path, are capable of mitigating higher levels of congestion before congestion spreading becomes an issue. They can flow control specific data flows without slowing other data flows sharing the same path. A Task Force within the IEEE 802.3 Working Group is now working to add increased congestion management capabilities to Ethernet that will resolve these issues, and efforts outside the task force may result in more immediate ad hoc solutions.

## Enabling Effective PAUSE-Based Congestion Management

Nevertheless, when considering a model for general ATCA data transmission from an NPU on one card to one on another card (or traffic manager to traffic manager), there are certain attributes that, as present in Fulcrum’s Ethernet switches, enable effective PAUSE-based end to end congestion management. Straightforward enhancements to the basic operation of 802.3x PAUSE operation, when coupled with very low latency through the switch, form the basis for this operation.

A generalized data flow from an NPU on line card 1 to another NPU on line card 2 through an Ethernet switch fabric is illustrated in Figure 2.

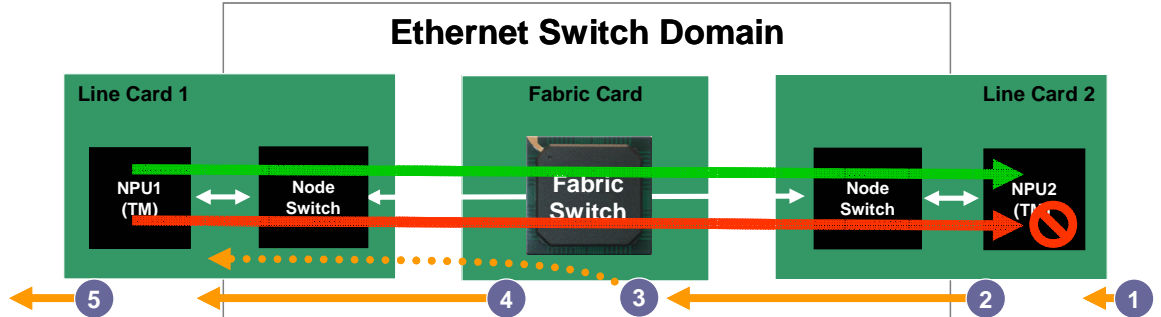


Figure 2: Data flows across an Ethernet ATCA backplane.

In this scenario, two data flows traverse the backplane from NPU1 to NPU2 through the switch – a higher priority flow (green) and a lower priority flow (red). The different data flows constituting the 10 Gbps traffic are distinguished by the different MAC addresses assigned to them. Detecting congestion, NPU2 chooses to flow control the low priority data flow, accomplished by the following steps as numbered in Figure 2.

1. NPU2 detects a congested condition.
2. NPU2 issues a PAUSE control frame toward NPU1 bearing the source address (SA) of the low priority data flow.

[NPU’s have the ability to assign and manage multiple MAC SA’s for data flows using the same physical interface. In this way the NPU’s are able to communicate between themselves as a collection of logical end points.]

3. The switch fabric chip, with its flexible port logic, may optionally respond to the PAUSE frame.
4. The switch fabric chip, again through the use of its flexible port logic, passes the PAUSE frame on toward NPU1.
5. NPU1 responds to PAUSE by throttling the queue for low priority data flow without affecting the throughput of the higher priority flow.

NPU2 must have enough queue buffer capacity to absorb all the frames it receives until NPU1 throttles the low priority data flow. Frames partially transmitted, partially received and “in flight” must be stored in NPU1’s input buffers during the PAUSE operations “latency”. The switch latency, as a contributor to the overall latency of the PAUSE mechanism, must be kept to an absolute minimum. Low latency switching ensures that end point buffer requirements are minimized, and Fulcrum’s industry leading latency performance is a key enabling factor in this congestion management solution

## **Conclusion**

Ethernet's long history of adaptability has led to its enormous installed base. Its ubiquity as a network connection has contributed to its unparalleled low cost and ease of use. For these reasons as well as its specified use as a base fabric, Ethernet should be considered for use as the high capacity data fabric in ATCA systems. Its weak point - congestion management - is overcome in the near term by switch fabric products from Fulcrum Microsystems, Inc. that feature flexible port logic and industry leading low latency. In the long term it will be overcome by class-based flow control to be defined by the IEEE and designed into future Fulcrum products.