Ubiquitous Ethernet

By Kim Barnhill

There are many interesting areas of development involving Ethernet in the embedded world. Considering this technology is not new, indeed its obituary has been written many times. What are the primary contributors that are sustaining its existence? From the rollout of Gigabit and 10 Gigabit, through the introduction of TCP Offloading to the concept of using Ethernet on the backplane (PICMG 2.16) for blade servers and other high-availability applications, this article examines why Ethernet has survived the storm and why it is still the physical layer of choice for embedded connectivity.

According to the International Data Corporation (IDC, 2000), Ethernet monopolizes over 87 percent of all installed network connections worldwide. For over a quarter of a century, Ethernet has proven itself to be a reliable standard: expanding from the original speed of the 4800 bits/sec transmission system to today's capacity of transmitting 10 billion bits per second across the network.

However, it's possibly the IEEE's determination to progress the standard rather than upgrading with a new and potentially better physical layer that has really encouraged Ethernet's acceptance. Ethernet has maintained a great degree of backward compatibility: making the initial investment in the LAN infrastructure still usable – thus reducing overall costs.

Although other technologies have, at times, been poised to replace Ethernet, they have, in fact, only caused extension to the standards. For example, ATM was speculated to be the replacement technology in many environments. Instead what has happened is that the features that ATM brought to the table (more raw bandwidth and Quality of Service (QoS)) have been adopted into the Ethernet world.

The fact that Ethernet (of various speeds) dominates the networking world has made it a prime communications mechanism in the Embedded Systems Community. However, when considering the effect of Ethernet in networking trends, should we be asking if it's the developments in the protocol stacks and software which sit above the signaling that are encouraging this acceptance?

The medium or the protocol?

Within any area of network infrastructure where data is being sent from point A to point B, the decision as to which transport mechanism to use is based as much on the networking protocol stack capability as it is on the physical medium. Within the last 10 years, Ethernet is effectively synonymous with TCP/IP and, concurrent to the developments of the Gigabit standard itself, there have been significant enhancements of the TCP/IP protocol stack and implementations.

TCP over IP over Ethernet has proved to be a powerful answer to the problems of data transmission. It provides an efficient data delivery mechanism, handling error detection/correction, data routing, and session control.

Version 4 of the TCP/IP stack was released back in the 1980's – before the explosive impact the Internet would have on everyday life was fully realized. More recently, rumors were rife that the world would run out of IP addresses and, immediately, the TCP/IP protocol and its inherent transport medium – Ethernet – were predicted the kiss of death.

The addressing limitations are solved in IP version 6 (IPv6) – and along with that, it advances the QoS features required for applications such as Voice over IP (VoIP) and security features (IPSec) which are required for commercial use of the network. IPv6 was slow to take off – but with all the major OS vendors now claiming compliance, it looks as if it might be providing the solution the Ethernet world had been hoping for.

Connectivity within the networking environment

With more people gaining access to the Internet and higher level computing applications becoming available to the public, increased information within decreased response times was being demanded across the network. By 2001, over 85 percent of the desktop market was estimated to be running at Fast Ethernet speeds of 100 Mbits/sec.

Solving the front-end usage, however, necessarily then put the pressure on the server technologies being accessed by the LAN interfaces. As users fought for and acquired the higher access speeds of 100 Mbits/sec, the servers found themselves needing to upgrade to even higher Ethernet connections to satisfy the ever-hungrier bandwidth demands. Many application and database servers already require multiple Gigabit connections.

The introduction of the IEEE-802.3ab Gigabit Ethernet over copper (enabling Gigabit speeds to be transmitted over Cat-5 cabling) fully instigated the use of Gigabit Ethernet within the computing network backbone and, today, 1000 Mbits/sec connectivity is widely deployed – including switches, servers, and desktop connectivity – using cabling that was already in place in most buildings. Concurrently, the software protocol enhancements were supplying the QoS demands to prevent latency problems derived from voice, video, and data applications vying for network space: with existing traffic management techniques in place (prioritization and MPLS), the Next Generation Network had been born (see the RAMiX PMC698TX: Intelligent Dual 1000Base TX Ethernet in Figure 1).

Figure 1

Failover and redundancy at the NIC level

There is a further benefit in using Gigabit NICs (Network Interface Cards) within storage applications (NAS, SAN) consequential to progress within the failover software developments of the last few years. Since downtime equates to lost money, it is becoming increasingly important to ensure that the system can constantly access, process, and remit information. The loss of one Gigabit link at any time can seriously impact this throughput which is why redundancy, even at the NIC level, is essential.

By adding a processor to a Gigabit NIC with two ports, we invest it with enough intelligence to run an embedded link monitoring software that monitors and manages the dual interfaces. The embedded software on the Intelligent Ethernet Card (IEC) monitors several user definable parameters as well as link status. When the NIC detects that a port is down (e.g. a cable break), it can switch all the traffic to the second port – which then takes over (using the same MAC address). This failover time can be as low as 20 milliseconds. There are configurable ways of letting switches know that it has happened and for how the fail-back is handled.

One key aspect of this method of failover is that it can be transparent to the applications. To the local host, the device appears as a single network port, totally compatible with existing applications and protocol stacks. As the MAC address is migrated to the active port, all systems that have connections are unaffected by any failover. Figure 2 shows an Example of failover at the NIC level – for LAN, server, or clustering environments.

Connectivity within the system

With the demands for higher bandwidth, it was only a matter of time before the industry derived a way of transmitting data at speeds of 1000 (even 10,000) Mbits/sec. But this has helped only in transmitting data across LAN/WAN medium or between the components within a system. The transport bottleneck moved again and it was now the central system architecture that was holding us up.

Whereas five years ago, one processor card supported by a set of I/O cards might have been the basic building block of an embedded system, current architectures depend more upon a large number of more powerful subsystems interconnected to create a distributed computing environment. This has put a huge demand on the control plane of the system itself.

The blade server

The introduction of PCI in the early 1990’s and its ensuing technology, CompactPCI, caused great excitement in the industry. But PCI has its limitations, both in its throughput capacity and as the appropriate solution for the evolving capability of processor modules. PCI is also generally considered to use up too many pins and be poor in its scalability. Furthermore, there was a dawning that the evolutionary path of PCI was not as a faster VME but as a mechanism to implement blades that would be interconnected by a switched fabric. There was one simple solution: Ethernet.

The heart of the blade server concept is the use of a switch fabric that links all the boards (or blades) in the system. It is the Ethernet Switch that is managing the movement of data between the boards. The ubiquitous adoption of Ethernet has encouraged the development of integrated (i.e. low chip count) implementations with aggregate throughput far in excess of PCI.

PICMG took this concept and developed it into the 2.16 specification which established the IEEE-802.3 Ethernet standard as the basis for the new CompactPCI Packet Switching Backplane (cPSB). The cPSB was designed, from its conception, to sup-

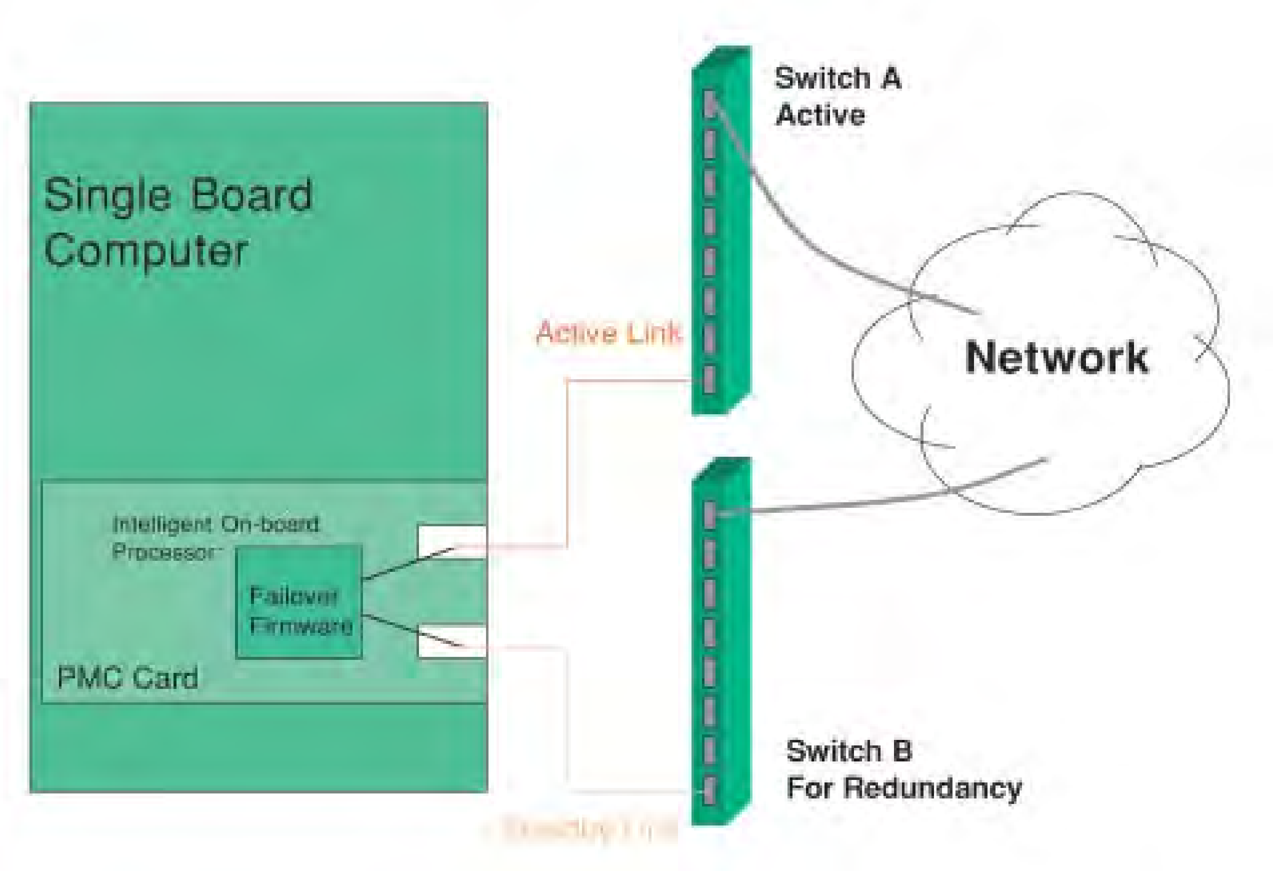

Figure 2

port the full range of Ethernet speeds (10, 100 or 1000 Mbits/sec) on each slot (up to two links per slot) with up to 40 Gbits/sec bandwidth over the backplane. Each of the two links on a blade is connected to an independent switch fabric to support an infrastructure intended towards fault tolerance or higher throughput.

Furthermore, in an industry striving for true 5-nines (or even 6-nines) availability, the need for high availability and hot swap within the system architecture was another driving force towards the adoption of Ethernet as the new backplane. Ethernet (or TCP/IP to be more exact) has built-in reliability.

Switching at Gigabit speeds

It is true that, until recently, most of the Ethernet developments within the cPSB arena were based on 10/100 Base Ethernet. RAMiX was the first manufacturer to recognize that Gigabit Ethernet transfer and switching would address their customers' needs within the near future. The CP924, RAMiX's first all Gigabit 24 port switch, was introduced in May 2002 to address these needs. With a managed version (the CP920) available now, the Gigabit Switching revolution is complete. Figure 3 shows PDSI's Gigabit switching CompactPCI chassis.

Figure 3

Ethernet enabling fault tolerance within the cPSB

While the PICMG 2.16 standard provided for the hardware duplication required for fault tolerance, it did not address the concerns of how to utilize the system infrastructure. This is a significant issue, as the requirements for fault tolerance within the chassis can be substantially different from those used in LAN solutions. In particular very rapid (order of ten's of milliseconds) response is desirable, along with a minimal intrusion on the design of the applications using the fabric.

The switch fabric connecting the blades in a 2.16 environment forwards packets based upon the destination address. The number of ports to which a received packet will be forwarded can be limited to the minimum necessary (in the case of point to point traffic, a packet can be directed out a single destination: for multicast, either static or dynamic information can be used to reduce the number of destination ports to the minimum). By restricting forwarding to only the necessary ports, the switch fabric can maintain multiple simultaneous connections, delivering an aggregate fabric data rate much greater than that of each single connection. Thus applications on blades 1 and 2 can be exchanging data at full wire speed at the same instant that applications on blades 3 and 4 are communicating.

Intercommunication over the fabric: IP protocol stack

An application identifies the target blade with the cooperating application by use of the IP address. IP addresses can be grouped into different subnets and, within a subnet, communication between applications can occur directly; communication across subnets requires the intervention of a router or gateway. The protocol stack executing on the blade chooses the Ethernet port based upon the IP address.

Failover using subnets

It is quite possible to have a backplane with two subnets (A and B in Figure 4 illustrate

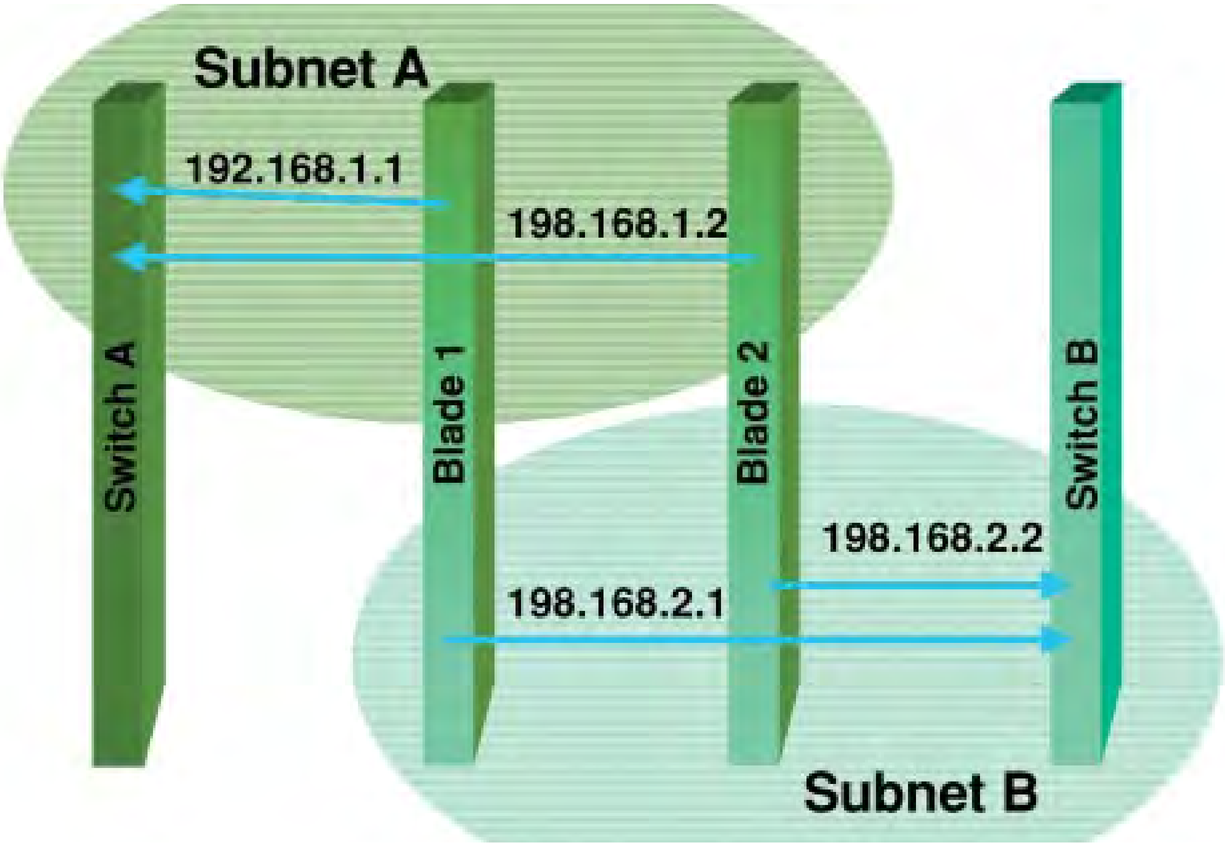

Figure 4

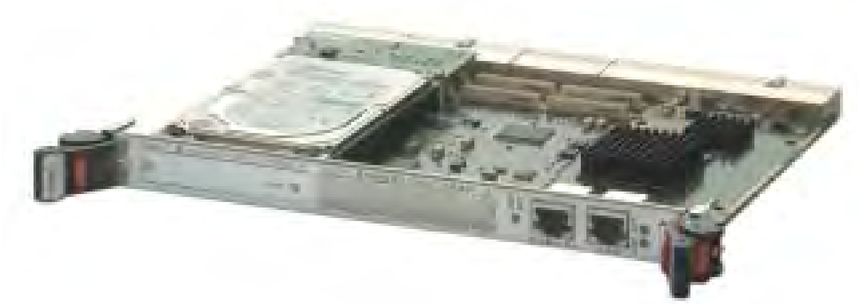

Figure 5

two subnets over the backplane showing one switch handling subnet A and the other subnet B). Every blade has two controllers, each with an IP address on each subnet.

So we can configure a 2.16 system with a fault tolerant communication infrastructure using the fabrics and network connections on separate IP subnets. However this places the responsibility on each application to be aware of any faults so that the operational subnet may be chosen. No standard method is in place to support such a requirement, placing the burden on the system designers to define and implement a unique, proprietary solution. As an example, the application starts talking to IP address 192.168.1.2. When failover occurs, it switches to 192.168.2.2.

The effort can be somewhat mitigated by creating a middleware layer, that is invisible to application programmers. However, this middleware still needs to be created and is likely to add overhead to the network execution.

Failover at the Ethernet driver layer

By using a technique similar to our failover NICs, RAMiX also offers an alternative to the above "failover using subnets." A mechanism uniquely appropriate to the requirements of the 2.16 chassis is to present the dual Ethernet connections as a single port to the higher layer protocols. This can be done either with a special purpose driver or (as in Linux) by inserting a pseudo driver between the hardware and the IP stack. The driver chooses one port as the currently active port, with the other port being an alternative. When a link failure is detected, the driver then transitions to the alternative port which then assumes the attributes of the original port (e.g. MAC Address). By maintaining the same MAC address, external systems (e.g. applications on the other blades) are not aware of the change. The switch fabric will note the change and automatically update its tables (this is analogous to the Ethernet cable on a workstation being pulled from one switch and plugged into another). By arranging for both the switches to keep in sync, both the failover time and the impact of failover can be kept to an absolute minimum.

TCP/IP off-load

One of the issues of using Gigabit Ethernet as a control plane solution is not only the processing power required to drive the Gigabit Ethernet at true "wirespeed" (1000 Mbits/sec input and output data rate), but also cope with the recognized overhead of the protocol stack itself. Moving to Gigabit Ethernet also requires processing a Gigabit of TCP/IP – a rather large processing requirement for any application. Effectively then, the bottleneck has moved: it is now the processor speed that needs to catch up with the interconnect.

In order to achieve full Gigabit functionality within the central system, future Gigabit Ethernet blades will necessarily need to be based on higher processor speeds. RAMiX's CP723 blade is designed around dual 800 MHz MIPS processors – one to off-load the TCP/IP protocol stack and drive a Gigabit Ethernet stream, the other to run the application itself. Figure 5 shows the RAMiX CP721 with one PMC site populated by the PMC233 High Capacity Disk Module

Beyond the Gigabit boundaries

With Gigabit Ethernet now addressing the speed, security, high availability, and fault tolerance issues demanded by high end telecommunications systems, there is little reason to believe that Ethernet as a transport medium will not continue to establish itself into the future. There are arguments of course that Gigabit Ethernet cannot attain the desired speeds of the network, but 10 Gigabit Ethernet is already a reality. Indeed, 10 Gigabit Ethernet is already being introduced into Embedded Products (RAMiX is already designing its first family of switches with 10 Gigabit Uplinks) and further flavors and selections of silicon will be released in the next six to nine months. The industry is only waiting on the applications to justify its use.

We have discussed how other technologies have, at times, been poised to replace Ethernet and others will, undoubtedly, be introduced. Where there are strong arguments for specifications such as StarFabric and RapidIO to succeed in one or other part of the overall topology of the network, no technology, however, possesses the basic essentials available with Ethernet.

Ethernet has a secure foundation – not only in its deployment but also in its acceptance as a reliable and secure technology. Off-the-shelf Ethernet solutions are easy to come by. Engineers, even those now emerging from education, have a strong understanding of both the transport mechanism and software protocols surrounding the technology. New technologies may encroach on some of its established ground, but it will be some time before they secure much stronghold.

With 10 Gigabit now a reality, Ethernet has quite a lifespan left.

Kim Barnhill *joined RAMiX Inc. in 2001 where he is now responsible for Product and Technical Marketing. Kim first got involved with Embedded Systems when he was working within engineering development for Raytheon Company. Since then, he has spent more than 18 years in direct Sales and Marketing within the community. Kim has a BS in Electrical Engineering from the University of Colorado and a BA Political Science degree from Colorado College.*

For more information, contact Kim at:

Kim Barnhill
RAMiX Inc.
1672 Donlon Street
Ventura, CA 93003
Tel: 805-650-2111
Fax: 805-650-2110
E-mail: kim@ramix.com
Web site: www.ramix.com