ethernet alliance

Facing the Challenges of Developing 100 Gbps Platforms

June 2010

Author:
Bill Weisinger, Bay Microsystems

1. This work represents the opinions of the authors and does not necessarily represent the views of their affiliated organizations or companies.

Ethernet Alliance | 3855 SW 153rd Drive | Beaverton, OR 97006
www.ethernetalliance.org

Executive Summary

The communications infrastructure across the globe is being transformed to deliver any service, any where, at any time. Consequently, services are being expanded, and the method for controlling these services is changing as the bandwidth required to deliver them is exploding due to ever-increasing video content. Aging broadcast-based cable structures and circuit-based telecom structures are being systematically replaced with packet-based networks (IP and Ethernet). In addition, wireless access is presenting new complexity to the way consumers and businesses are accessing these networks.

This paper will address the challenges of scaling packet networks from the current state-of-the-art, 10 Gbps limited packet-aware, to next generation 100 Gbps all packet-aware, any-to-any networks. The areas covered will include optical/electrical interfaces, interface adaptation, data plane processing, control plane processing, as well as switching, system power and printed circuit board requirements. These challenges apply to platforms that aggregate lower speed interfaces, up to 100 Gbps, as well as platforms with pure 100 Gbps interfaces.

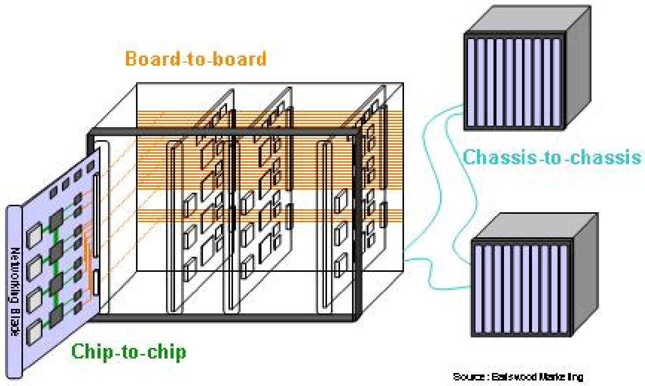
Introduction

There are a number of disruptive global market trends driving the deployment of 100G (Gbps) networks. More people want more of everything, more often and in more places than ever before. This is leading to demand for incremental services and bandwidth across the entire communications infrastructure. The telecommunications downturn of the early 2000s is behind us. Across wireless and wireline networks and across service provider and operator territories, the number of subscribers is increasing steadily. Overall, quadruple-play services – the convergence of voice, video, data, and wireless – demand higher bandwidth and improved quality of service support. There is the rapid growth of broadband subscribers with increased access speeds via xDSL, FTTransit, WiMax, and 3G/4G. Most dramatically, network bandwidth is exploding due to demand for consumer-based video and multimedia applications. These include applications such as IPTV, VoIP, online gaming as well as a growing number of Internet users accessing video sites. Who hasn’t heard of YouTube these days? In the past, traffic growth was steady and predictable. Now, not only is traffic explosive, it's unpredictable. This is a new paradigm for service providers and operators to address.

The downturn in the telecom industry in the early part of this decade resulted in network equipment manufacturers “band-aiding” existing platforms to offer an incremental bandwidth boost to their customer base. The market is dramatically different today, as numerous leading service providers and operators have announced plans to overhaul their networks to address the explosive demand for high-bandwidth, high-quality multimedia packet-based services – AT&T, Verizon, and BT are just a few examples. Having access to more bandwidth opens an entirely new avenue of revenue-generating opportunities for the carriers.

There’s no question that the market demand exists today for high capacity network elements. However, most would agree that deploying 100G networks is no simple task. Next generation networks must deliver the bandwidth, quality of service, and advanced packet-based services required to support future requirements that may still be in their infancy. In addition, the long term viability and future scalability of these new deployments must be carefully considered.

In order to address the challenge of delivering 100G networks, it is important to develop an entire ecosystem. It is also important to learn from past efforts where well intentioned standards were little-used, such as OIF's SPI-5 and SFI-4.2. There is a need to think in terms of, what do carriers require of the network equipment manufacturers? What do the equipment manufacturers require of their software and hardware component suppliers? How can these groups work together to accelerate the deployment of high bandwidth networks? Can the industry as a whole, as it pushes technology to 100Gbps, meet the challenge to provide end-to-end solutions (see *Figure 1*)? Let's take a look at these challenges…



Source : Ealswood Marketing

Figure 1. End-to-End Solution

Optical/Electrical Interfaces

The IEEE 802.3ba standard has recently been approved to support 100G optical and electrical interfaces. Earlier solutions supported up to 10G optical wavelengths with expansion via multiplexing (n x 10G) techniques. Although demand is growing, 40G wavelengths are not yet widely deployed. To further complicate the matter, until now 40G systems deployed by different manufacturers are not compatible.

It may seem obvious that initial implementations of 100G optical interfaces will be 10 x 10G but beyond this, nothing is final. Early momentum around 25G (x4) signaling has developed. However, 20G (x5) signaling may prove to be more practical, since intermediate steps of 40G and 80G are more easily implemented using 20G increments. In addition, 20G increments could address the areas of 120G and 160G that are already being discussed. Eventually 1 x 100G will evolve but, this is a longer term goal for the industry at large.

Multi-vendor compatibility for optical interfaces is a firm requirement for a successful ecosystem of network equipment products. In addition to deciding what the "atomic unit" (10G, 20G or 25G) of the interface will be, uniform modulation techniques must be used to enable interoperability between different Original Equipment Manufacturer (OEM) systems, such as on switch/routers and the client side of transport equipment.

These requirements may remain different for DWDM line side interfaces where traditionally the industry has delivered non-standard solutions with differing performance capabilities. DWDM interfaces are usually interconnected between equipment from the same manufacturer and hence the specifics of the interface remain the internal property of the equipment manufacturer. However, for high bit-rate Ultra-Long Haul optical transmission, more and more signal processing is being performed in the electrical domain. Some suppliers today are implementing coherent transmission systems at 40G which places heavy requirements on electrical signal processing at the receiver in order to decode and 'clean up' the signal before deciding whether a '1' or a '0' was transmitted. Such coherent transmission systems may offer significant optical performance advantages over traditional optical technologies, such as Dispersion Compensating Fiber, which have been utilized over the last decade to improve transmission quality. Expanding coherent transmission capabilities to 100G places a significant burden on the electrical signal processing required at the receiver. For example, with a complex optical modulation format such as Dual-Polarization Quadrature Phase Shift Keying, where the 100G optical signal is effectively split into 4 optical data streams contained in the same wavelength, the Analog to Digital Conversion at the receiver is required to operate in the region of 50 Giga-samples per second. This is similar to the performance of state-of-the-art test equipment solutions available on the market today. In addition anywhere from 4 to 8 ADCs per 100G wavelength are required depending on the chosen implementation. The requirements on the subsequent DSP architecture where the signal correction algorithms are housed are equally significant.

On the electrical interface, as shown by the dotted line in Figure 2, one of the important considerations will be whether the "backplane technologists" and the "line-side technologists" can cooperate on common standards. The industry as a whole would benefit greatly if the same SerDes (Serializer/Deserializer) technology was on both sides of the line card. This would reduce board development time and component costs, as well as research and development resources.

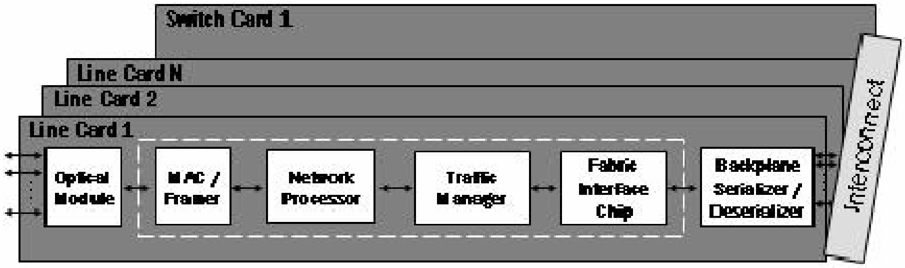


Figure 2. Chassis Line Card

Standards groups, such as IEEE, ITU-T, OIF and others have started formal processes to solve some of these issues. The industry as a whole needs to ensure that we don't end up with incompatible or competing standards, as we have in the past. These groups need to coordinate so that these new standards are useable for both Ethernet and WDM technologies.

Interface Adaptation to the Data Plane

The next area of concern in a typical system is the interface to the data plane, where compatibility is the key issue. For example, there is a standard 7dB gain Forward Error Correction (FEC) code defined for 10G and 40G in ITU G.709 standard; however, many systems require better gain, requiring an enhanced FEC (eFEC). Currently, there is no standard for an eFEC. Therefore, *de facto* standards are the only way to address this need. If there is a consensus on a *de facto* standard, then this is not a problem, especially when compared to an environment where multiple standards exist.

In the area of electrical interfaces multiple standards are more common. For example, the varying constituents of certain standards groups have led to differing data-striping schemes in Ethernet interfaces versus optical interfaces, such as byte-alignment algorithms in XAUI and bit-alignment algorithms in SFI. This requires component suppliers to build separate devices for each application or interface and has led some suppliers to create custom interfaces to accomplish a mix of interface and protocol support. Common interfaces can help leverage volume, bring the cost down, and reduce development time.

Finally, packet payload size compatibility is critical to protocol interoperability. The networks being built today will have to handle legacy protocols, as well as evolving protocols and data types. The mismatch that took place between 10GE (10.3125 Gbps) and OC-192 (9.5846 Gbps) is something that the industry can ill afford to repeat. These networks are complex enough without building foreseeable incompatibilities into the architecture.

Data Plane Processing

Similarly, compatibility is the key issue in achieving high performance data plane processing solutions. In this case compatibility is required between components supplied by different manufacturers, as opposed to the standards compatibility described above in the Interface Adaptation section. Chip-to-chip interfaces involve a variety of technologies, speeds and protocols. These chip-to-chip interfaces can be broken down into two categories, data path interfaces and side-band interfaces, as shown in *Figure 3*. Data path interfaces include the MAC/framer to network processor (NP), the NP to traffic manager (TM), and the TM to switch fabric or fabric interface chip (FIC). Although there may be more or less components in the data path (i.e. – FPGA, no separate TM), the three functions listed above represent the main components and interfaces. Side-band interfaces include everything from search engines and co-processors to memory for instructions and buffering. Additional challenges include clear definitions of processing functions in the data plane, which protocols need to be supported, and Quality of Service (QoS) functionality. Standard metrics to demonstrate that 100G rates can be achieved and maintained are also required.

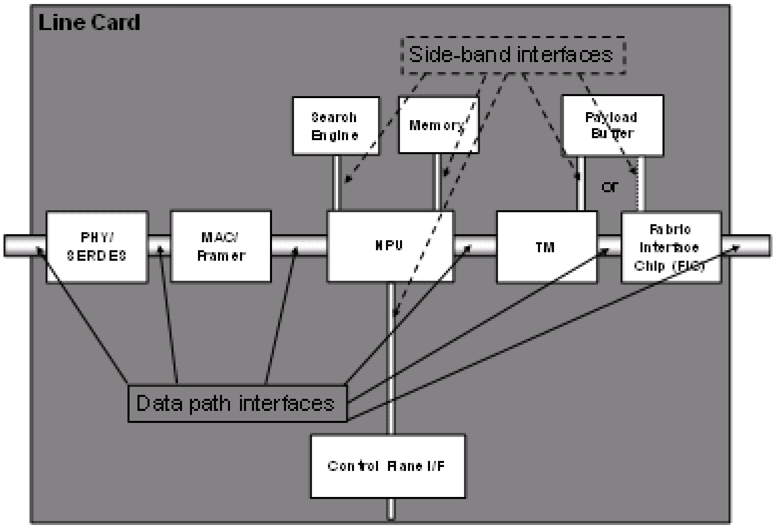


Figure 3. Line Card Interfaces

Data plane interfaces, both data path and side-band, are areas that suffer from too many standards today. The various standards for these interfaces tend to be developed by different organizations, each with its own concerns and priorities. IEEE has developed the XAUI specification for MAC devices, while OIF puts forth Implementation Agreements to the ITU-T for framers. Since both the MAC and framer devices need to interface with the NP, this is an ideal situation where cooperation could solve a problem for NP manufacturers. This opportunity for cooperation flows to all of the other data path interfaces, where the same interface could be used between the NP and TM, as well as the TM and FIC. This common interface across all blocks becomes critical since some designs may omit the NP/TM or TM/FIC interface, but not both. As of October 2007, there are three to four industry standards approved or in process for 40G to 100G interfaces (OIF SPI-5 and SPI-S, as well as IEEE 40GE and 100GE). The side-band, or memory interfaces, has even more standards to contend with in data plane designs. SSRAM has QDR-II/II+/III and DDR-II/II+/III interfaces, while DRAM has DDR3-SDRAM and RLDRAM-2/3. Each memory type has its own advantages with respect to density, latency and determinism, yet technology has led them to similar interfaces as they move to the highest performance nodes. Cooperation amongst memory suppliers leading towards interface convergence would enable everyone in the ecosystem to become more efficient and focus on their own core competencies to be more competitive. In the case of search engines, high-speed parallel interfaces have traditionally been used to interface to Network Processor Units (NPUs) or host ASICs. However, due to technology limitations, the frequency and width of these parallel interfaces cannot be scaled indefinitely to accommodate higher search rates. As a result, serial technology is being investigated as a forward looking path for side-band interfaces. As we move forward in this area, we must be diligent as the evolution to serial from high-speed parallel interfaces will need to be navigated carefully.

Basic data plane processing functions can be broken down into classification, ingress policing, editing, and egress traffic management (scheduling and shaping) blocks. Depending upon the service or application being provided, the details within each block can vary significantly. Setting a baseline of protocol support requirements for 100G will allow the industry to continue its convergence and make measurement and guarantees of service much easier. This is not to say that new protocols will not evolve over time, as we know they will, but a baseline will give the industry a common starting point from which this evolution can continue. New revenue generating services will have to maintain QoS guarantees across the network if they are to be truly successful. This will require clear industry-wide definitions of QoS, including ingress policing requirements and algorithms, egress traffic management requirements, buffer capacity requirements, multicast handling, and more. Measurable standards such as the minimum packet size, number of searches per packet, number of policing instances will need to be defined to ensure data rates are achieved.

The baseline standards should not be a watered down specification which can easily pass through committee approval. Standards should be based upon the needs of the service and true capabilities of the industry as a whole. For example, if traffic shaping can be achieved within 1% of the requested rate, it should not be specified at 10% due to prior sufficiency. Efficiencies in shaping will enable less over-provisioning of trunks to maintain QoS and increase margins for provider and operators. This type of advancement is a tangible return that should be exploited by everyone in the ecosystem, at every opportunity.

Control Plane Processing

Modern networking equipment is expected to have a long lifetime, on the order of 10-20 years. This is because networking infrastructure is increasingly an integral part of the fabric of the society. What does this have to do with the control/management plane? The longevity of equipment can only be satisfactorily ensured if the key components of the software architecture itself are solid, yet adaptable.

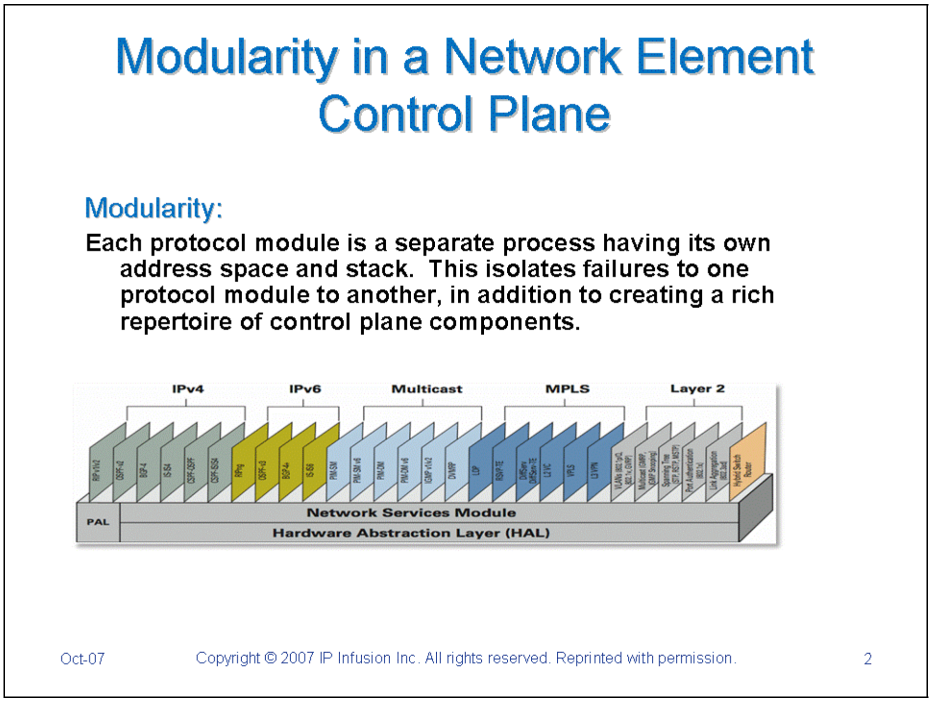
Networking equipment is increasingly called upon to provide a rich variety of functions. Several different access methods must be supported, such as wireless, cable, DSL, fiber, etc. Different levels of functionality are expected including, Layer 2 functionality (e.g., Ethernet LANs, point-to-point Ethernet connectivity, Ethernet VLANs, etc.), Layer 3 connectivity, be they physical or virtual, and Layer 7 functionality. Some subscribers still need some legacy connectivity (e.g., T1/E1, T3/E3), as well. A good control plane implementation must therefore have resilient software architecture to support the functionality illustrated.

To a large extent, the need for wire-speed processing ties system designers to certain hardware architectures. Therefore, a major portion of a system's flexibility is borne by the software in the network equipment. Since the software in network equipment is essentially in the control plane, it is important to ensure that a proper architectural foundation is created in the control plane. At a minimum the following set of attributes, in no particular order, must be engineered from the beginning to create effective and resilient control plane processing software.

1. Modularity. The primarily value of modularity is the mix-and-match property that it enables, as shown in Figure 4. Not all equipment needs to have the same functionality; however, certain basic functions are shared by all equipment.
2. Portability. A key advantage of a portable design is that essentially the same functionality can be offered on a variety of platforms, at various different price-performance points.
3. Distributed architecture. With the cost of computation constantly coming down, it becomes more economical to provide multiple, relatively lower performance, computational elements instead of a single very high performance element. Thus, a software architecture that can exploit distributed

computing power can more effectively scale as new processors become available, which is particularly valuable at high speeds such as 100Gbps.

4. Performance. Performance is particularly acute in the case of real-time communication involving audio and video.
5. Scalability. In many cases when something new becomes available on the Internet, it is rapidly accessed by millions of subscribers within minutes. Without adequate scalability built into network elements' control plane, it is not possible to sustain such large usage changes.
6. Manageability. No deployed equipment is particularly useful without adequate manageability for detecting and correcting faults, configuring the network services effectively based on subscriber's needs, keeping accurate accounts of usage for proper billing, ensuring satisfactory performance based on service level agreements (SLAs), and providing for secure operation and usage of the network. This directly impacts operational expenses (OPEx) for providers and operators.
7. High Availability. Subscribers have come to expect near 100% availability. In practice, 99.999%, or roughly the equivalent to 5 minutes of downtime in a year, is expected. This relates directly to network resiliency (hardware and software). High availability cannot be achieved without adequate resiliency built into the network elements.



Modularity in a Network Element Control Plane

Modularity:

Each protocol module is a separate process having its own address space and stack. This isolates failures to one protocol module to another, in addition to creating a rich repertoire of control plane components.

Oct-07 Copyright © 2007 IP Infusion Inc. All rights reserved. Reprinted with permission. 2

Figure 4. Software Modularity

One of the first ingredients of the architectural essentials is a set of software abstractions. Each of these abstractions addresses an important piece of the system information that must be hidden from the majority of the rest of the network equipment software.

1. A hardware abstraction layer (HAL) serves to hide the details of different kinds of hardware essentially aiming for a similar, or a family of, performances in the hardware.
2. A platform abstraction layer (PAL) assists in making sure that the control plane software is not tied to a particular operating system (OS). Although there has been a reduction in the number of operating system candidates in recent times, this is useful in itself as an element of good design.

3. A service availability abstraction layer (SAAL) is useful to insulate the system designers from various different high availability (HA) middleware.
4. A management abstraction layer (MAL) is needed to make available the management information in various alternative access methods such as TL-1, SNMP, XML, HTML and CLI.

Switching

High performance switching (aka - switch fabrics) has long been the domain of network equipment manufacturers. These switch fabric technologies have often been the crown jewel of chassis-based systems, far ahead of the processing/routing blades in performance, and the infrastructure for which many generations of routing blades can be sold into without forklift changes to the chassis. As we move towards 100G, the challenges of next generation switch fabrics are taking on the same characteristics as other challenges described in this paper. New standards are needed for the speed, number and coding schemes of backplane serial links. Common approaches to handling legacy data types, such as TDM, as they converge with new packet-based data types are needed. And, the ever-important power and cost per port/gigabit metrics continue to be driven to lower and lower numbers, to reduce the “carbon footprint” of users around the world.

When discussing backplane serial links, timing and maturity of SerDes technology comes into play, as discussed in the Optical/Electrical Challenges section. Today’s use of 3.125G links with 8B/10B encoding (20% overhead) greatly increases in complexity when moving to 100G bandwidths. The use of 10G, 20G or 25G links need to be considered, along with the new and more efficient coding or scrambling techniques. However, the use of higher backplane speeds cannot come at the price of using exotic and expensive printed circuit board materials and connectors. These limitations present challenges for system designers trying to mitigate crosstalk, signal loss and other issues. Reliability of newer technologies must be proven in order for them to be deployed in long-life products, such as networking infrastructure equipment. Reliability in this environment includes auto-sensing of links for graceful degradation, and tolerance to humidity and temperature, as well as basic life expectancy.

Convergence to a unified packet-based infrastructure brings additional challenges to switch fabrics, especially when looking at converging TDM-based data with IP and/or Ethernet data. Meeting the QoS requirements of these varying traffic types requires a complex architecture to differentiate between unicast and multicast traffic as well as low latency, high bandwidth or guaranteed delivery traffic. While power and cost are important to the entire system, switching technology can be a significant contributor to the problem with the large number of SerDes interfaces and the amount of memory required to provide the QoS characteristics just described.

System Power

Power consumption is a never-ending battle for system designers. However, as the phrase "Carbon Footprint" becomes part of our everyday language, the power problem is one that needs to be addressed up front, not as an after-thought. Developing standards for power consumption is needed moving forward. Is the important number "total power" or a somewhat arbitrary "power density" number? Is it important to measure power per gigabit or power per port (1G or 10G)? Are the metrics different for an ATCA chassis versus micro-ATCA or a proprietary 7 ft. rack? Identifying what is wanted, needed or even possible and how to measure power needs to be determined and standardized so customers can compare solutions across the industry.

Printed Circuit Board (PCB) Assemblies

The standard speed for SerDes in today's chassis-based communications equipment is 3.125G. This can equate to as little as 2.5G of effective bandwidth per serial link, depending upon the protocol/encoding scheme used across the link. For lower density line cards, such as 10G up to 20G, this approach results in a manageable number of link pairs between devices. However, when designing cards above 20G, this approach results in significantly increased cost and power, and becomes unmanageable. Therefore the industry is exploring the use of higher speed serial links, as well as more efficient encoding schemes. This is particularly a challenge for high capacity system backplanes, where route density and connection length issues exacerbate the problem further.

The use of higher speed links introduces a number of challenges for system designers. The choice of serial link speed must consider signal integrity, and thus the overall PCB design. These challenges include the following:

1. Desired bit error rate (BER)
2. PCB material: Some companies wish to use standard FR4 material and back-drill (a technique to counter signal reflections that are caused by the thickness of the PCB) specific vias. Others are migrating to a modified material such as Isola or Nelco
3. Trace length: What is the maximum distance between components? Are switch cards located in the middle of the chassis or are they placed at one end?
4. Connector type: Through hole or surface mount? What type of shielding, if any, is included in the connector (which is where most crosstalk occurs since signals are less closely coupled)?
5. Signal Integrity
    a. Transit and receive termination
    b. Jitter: transit generation and receive tolerance
    c. Voltage swing: transit output swing and receive sensitivity

About the Ethernet Alliance

The Ethernet Alliance was formed by companies committed to the continued success and expansion of Ethernet technologies. By providing a cohesive, market-responsive, industry voice, the Ethernet Alliance helps accelerate industry adoption of existing and emerging IEEE 802 Ethernet standards. It serves as an industry resource for end users and focuses on establishing and demonstrating multi-vendor interoperability. As networks and content become further intertwined, the Ethernet Alliance works to foster collaboration between Ethernet and complementary technologies to provide a totally seamless network environment. To learn more, please go to www.ethernetalliance.org.