

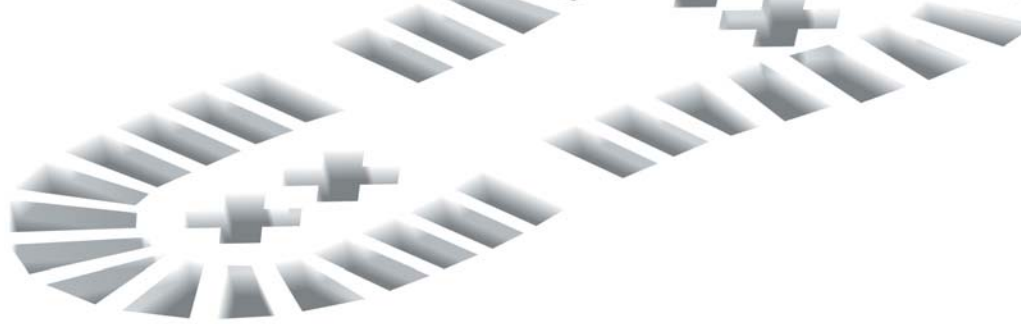
BY MAURY WRIGHT • EDITOR AT LARGE

WHICH **INTERFACE** WILL GET TRACTION?

FIBRE CHANNEL RULES IN THE SAN WORLD, AND INFINIBAND IS PROGRESSING IN COMPUTER CLUSTERS, BUT A FASTER FLAVOR OF THE VENERABLE ETHERNET LOOMS AS A POTENTIAL JACK-OF-ALL-TRADES.



ILLUSTRATION BY JAMES GARY



Typical enterprise IT departments now support three networks. Ethernet links PCs and servers. FC (Fibre Channel) connects SANs (storage-area networks) that connect servers to storage arrays. And proprietary interconnects, Myrinet, or emerging IB (InfiniBand) technology generally link clusters of servers that focus on evolving multithreaded tasks, such as database access. For board and system designers, however, that landscape may change significantly. Proponents believe that IB will first prevail in clusters and then move into storage applications, whereas proponents of 10 GBE (10-Gigabit Ethernet) believe the LAN technology will win in each segment. And you aren't immune to this changing landscape if you don't happen to design servers or storage products. The interconnects that win in these applications will find use in all sorts of specialty applications driven by the low-cost ICs that volume deployment in the enterprise invariably delivers.

A little more than a year ago, Broadcom launched what it termed C-NIC (converged-network-interface-controller) ICs (Figure 1) and squarely targeted Ethernet to replace FC, IB, and proprietary technologies in storage and cluster applications. The theory goes that iSCSI (Internet Small Computer Systems Interface), an IETF (Internet Engineering Task Force) standard, provides the means for moving block-storage data across a TCP/IP (Transfer Control Protocol/Internet Protocol) network and remotely managing storage resources. Meanwhile, RDMA (remote-direct-memory-access) extensions to TCP/IP, an RDMA Consortium standard, support both iSCSI block operations and the low-latency copy of data from one system to another that clusters require. Ethernet still may lack the performance these applications require, but TOEs (TCP/IP-offload engines) and similar technologies can fix that problem.

This time is far from the first that anyone has applied a flavor of Ethernet and TCP/IP to an application in which it seemingly doesn't belong. Embedded systems regularly use low-cost Ethernet chips to link relatively unsophisticated sensors and control systems. EPON (Ethernet-

passive-optical-network) technology is one flavor of fiber-based broadband that carriers are deploying to support the so-called triple play and IPTV. System builders have even used Ethernet for board-to-board communications over backplanes in systems based on buses such as CompactPCI. Another fabric, the Advanced Switching Interconnect, may replace Ethernet in the system-fabric role (see sidebar "Understanding the Advanced Switching Interconnect"). All these interconnects have their own places in the computer-architecture puzzle (Figure 2). Rick Maule, chief executive officer of start-up Ethernet-chip vendor NetEffect, states, "We believe that there is truth to the adage: 'If Ethernet can Ethernet will.'" In the case of storage and cluster interconnects, Ethernet is an attractive option for several reasons. For starters, IT departments could potentially use one set of network-management tools if the departments base data, storage, and clustering all on Ethernet—even if they actually implement them on physically separate Ethernet LANs. Assuming that the ramp in Ethernet is true to form and 10 GBE becomes a high-volume technology, then surely it will be cheaper than FC, IB, or

other options. And Ethernet might carry a mix of data, cluster, and storage traffic on a single network.

ENTERPRISE NIRVANA

Tim Golden, director of PowerEdge-server marketing at Dell, states, "The notion of fat pipes and a single fabric—it's a very promising thing." Golden claims that, from an IT perspective, nirvana might be "a cloud with all resources that can be managed from a single remote point." Broadcom's C-NIC marketing material was the first to make that pitch.

So, Ethernet wins, right? Not so fast. The best effort packet-delivery and collision-based MAC (media-access-control) scheme of Ethernet and TCP/IP isn't a perfect match for storage and cluster applications that demand low latency and guaranteed quality of service. Moreover, low cost—when it comes to 10 GBE—isn't a sure thing.

First, consider cost. Through 1 GBE (1-Gigabit Ethernet), you could easily identify applications that would almost ensure broad market adoption of the next-generation technology. Today, vendors ship almost all clients with 1-GBE ports, although enterprises haven't performed wholesale upgrades of their switches and routers. Still, applications such as video delivery will lead to ubiquitous 1 GBE. But it's tougher to make a case for 10 GBE even if the chip vendors can drive down the cost of 10 GBE using process-geometry reductions.

Most of the Ethernet proponents point to the "if-you-build-it-they-will-come" nature of the tech industry when it comes to data rates and memory. But for 10 GBE, it may be tough to identify any application that requires the sheer data rate the technology affords. Instead, the justification may ultimately be how quickly you can move a large data file—say, a full-length movie—across a network that drives adoption rather than just the ability to stream live movies.

Broadcom sees the process as self-prop-

AT A GLANCE

With volume sales driving down Ethernet's implementation costs, customers are adopting it in applications in which it seemingly doesn't belong.

Ethernet and TCP/IP lack the performance characteristics that storage and clusters require, but extensions such as TOE, RDMA, and iSCSI can close the performance gap with interfaces such as IB.

Unlike earlier generations of Ethernet, 10 GBE (10-Gigabit Ethernet) may not ramp quickly to volume deployment because it's difficult to find applications that require the faster data rate. Then again, the tech industry always seems to find a need for more memory and bandwidth.

Fibre Channel may well remain the dominant choice in storage networks, despite the attack from both 10 GBE and IB.

agating. Allen Light, senior product-line manager for C-NIC, acknowledges that his company is just now getting people to move to 1 GBE. But Light claims that virtually all servers that companies ship today have dual 1-GBE ports. He claims that IT managers will soon discover and start using the available excess bandwidth for storage applications. The transition of storage traffic will lead to greater traffic demand and ultimately to volume deployment of 10 GBE. Maule of NetEffect claims that server designers adopt each succeeding Ethernet generation when the price premium is two or three times the cost of the previous generation—essentially “future-proofing” the design. Maule states, “The server deployment gets you the several-million-unit volume that drops the premium to 30 or 40%, and that then gets you into clients.”

UNDERCUTTING PRICES

Ironically, IB is far cheaper today despite the fact that low cost is Ethernet's normal calling card. Mellanox Technologies is the only true merchant supplier of IB chips. Topspin Communications had developed

both IB chips and system-level products, such as IB switches. But Cisco acquired Topspin and uses the company's ICs only internally in IB switches and other products. PathScale also developed an IB ASIC that it uses on board-level products. It's unclear whether the company will pursue chip-level business. Still, Mellanox has driven chip prices to less than \$100. The InfiniHost III Lx chip it announced in March costs as little as \$69 in high volume, and dual-port chips sell for approximately \$200. Ted Rado, vice president of marketing at Mellanox, claims that his company has shipped 500,000 IB ports and that it shipped 300,000 of those 500,000 in the past year. Rado puts 10-GBE shipments at less than one-tenth that volume, pointing out that IB is enjoying the volume advantage for now.

Debbie Vogt, vice president of marketing at Ethernet proponent Siliquent, doesn't dispute the current IB cost advantage. Siliquent claims to be the first company shipping 10-GBE chips that can support storage and cluster applications. And Vogt admits that the 10-GBE links—the chip on both ends and the cable that connects the two—cost three to four times as much as an IB link today. Still, Siliquent officials believe that Ethernet is the future for clustering and storage. Vogt states, “Being able to do something with IP throughout an infrastructure is very powerful.”

In reality, the cost comparison also goes far deeper than the NIC price. Ultimately, you must factor in the cost of switches and other infrastructure. As

Broadcom's Light states, “You do TOE, RDMA, and iSCSI only on the ends,” meaning that standard 10-GBE switches will handily carry storage and cluster traffic and will likely cost far less than IB switches.

ENTRENCHED CONTENDERS

Cost is obviously only one point of comparison. You also need to evaluate how Ethernet and the competition stack up from a performance perspective. In the storage segment, FC is the entrenched competitor. Originally a 1-Gbps interconnect, FC is now widely available at 2 Gbps, and the industry is doing early testing of 4-Gbps products. FC's developers designed it with a relatively thin protocol layer for the block-level data-storage and -manipulation requirements of big databases, such as Oracle. The interconnect also finds use in storage-centric applications, such as data mirroring and transparent backup and restoration. For more background information, check out the Fibre Channel Industry Association's Web site.

In the cluster case, the installed base is more diverse. Some of the “big-iron” computer vendors have proprietary interconnects. Mellanox boasts a number of impressive case studies of IB-based clusters; you can find details on the company's Web site. At a signaling rate equivalent to that of 10 GBE, IB delivers 8-Gbps data rates due to the 8B/10B encoding that IB uses. IB is scalable through the addition of signaling lanes, and supporters are planning a doubling of data rates, but the 8-

(continued on pg 64)

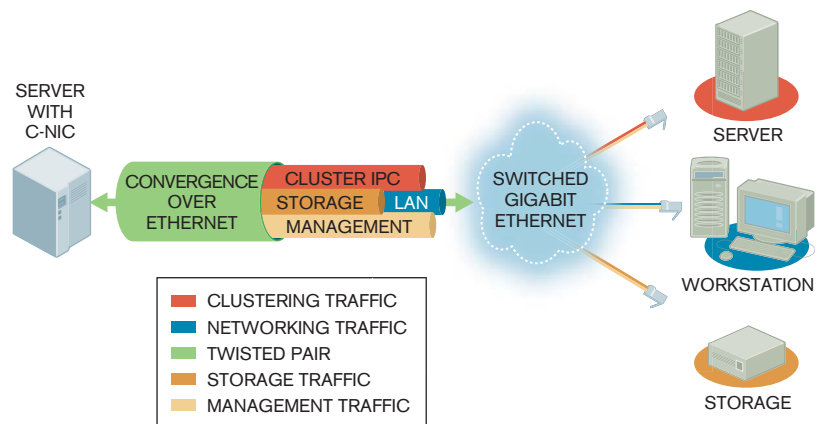


Figure 1 Broadcom has evangelized the convergence of data, storage, and cluster interconnects since the announcement last year of its C-NIC family.

UNDERSTANDING THE ADVANCED SWITCHING INTERCONNECT

By John Chiang, IDT

The ASI (Advanced Switching Interconnect) serial-interconnect technology provides high-performance features of proprietary fabrics and uses the standards-based economy of scale and ecosystem of PCI Express.

Through enhancements in the transaction layer, ASI extends the capabilities of PCI Express to support a variety of converged computation and communication applications (Figure A). Key features of ASI include low latency, peer-to-peer communications, multiple levels of QoS (quality of service), sophisticated congestion management, high reliability and fail-over mechanisms, multiprotocol and multicast support, and built-in security. ASI targets backplane and local onboard-interconnect applications, as well as chassis-to-chassis communications.

ASI provides as many as 20 virtual channels, of which eight are bypass channels that can transport

load and store protocols, eight are ordered only, and four are multicast, to facilitate almost any application profile. At the logical level, it supports eight traffic classes per virtual-channel type for QoS and traffic differentiation. ASI uses a credit-based, link-level, flow-control scheme. For congestion management, it supports status-based flow control, injection-rate control, minimum bandwidth, or vendor-defined egress scheduling.

ASI embraces compatibility and provides the mechanisms for more efficiently supporting legacy infrastructure. It realizes this goal through various native data-movement protocols, software semantics, and protocol-agnostic tunneling through a universal fabric technology. PI-2 (protocol interface 2), generic data transport, provides a reliable transport mechanism with built-in segmentation and reassembly for message-passing architectures.

Users can employ it to interoperate various end-point devices, such as NPUs (network-processing units), CPUs, microprocessors, and DSPs. ASI defines PI-0 to PI-95. It leaves PI-96 to PI-127 for vendors' proprietary protocols. It assigns the protocols as follows:

- PI-0: spanning tree,
- PI-1: congestion management,
- PI-2: generic data transport,
- PI-4: device management,
- PI-5: event reporting,
- PI-8: PCI Express tunneling,
- PI-E: Ethernet tunneling,
- PI-9: socket-data transfer,
- PI-10: simple load store, and
- PI-11: simple queue (SQ).

ASI has not yet assigned PI-12 through -95.

Fabric-management capabilities are also part of the ASI protocol to support a number of services, such as connection setup and teardown, event management, performance and

health monitoring, redundant routes, path invalidation, resource allocation, and load balancing.

ASI provides a scalable architecture by using the same PHY (physical) and data-link layers as PCI Express. It supports 2.5-Gbps serial-link technology in one-, two-, four-, eight-, 12-, 16-, and 32-lane configurations. Second-generation, 5-Gbps serial-link technology has also emerged. ASI can flexibly autonegotiate and interoperate to a variety of port bandwidths that different applications require. It also supports lane reversal to prevent a failure of a single lane's bringing down the entire link. ASI supports various fabric topologies, such as meshed, star, dual star, and dual-dual star and can cascade to larger fabric topologies by integrating sophisticated congestion-management and end-to-end flow-control capabilities.

ASI targets enterprise, communications, and embedded systems requiring high-performance fabric features. Typical applications include enterprise storage routers and arrays; blade servers; telecom edge, access, and metropolitan switches and routers; and embedded-system computing, such as in military and medical imaging. ASI works with any protocol and its broad industry support and maturing ecosystem of products provide strong advantages over proprietary- or niche-fabric technologies.

AUTHOR'S BIOGRAPHY

John Chiang is product manager of IDT's Serial Switching Division.

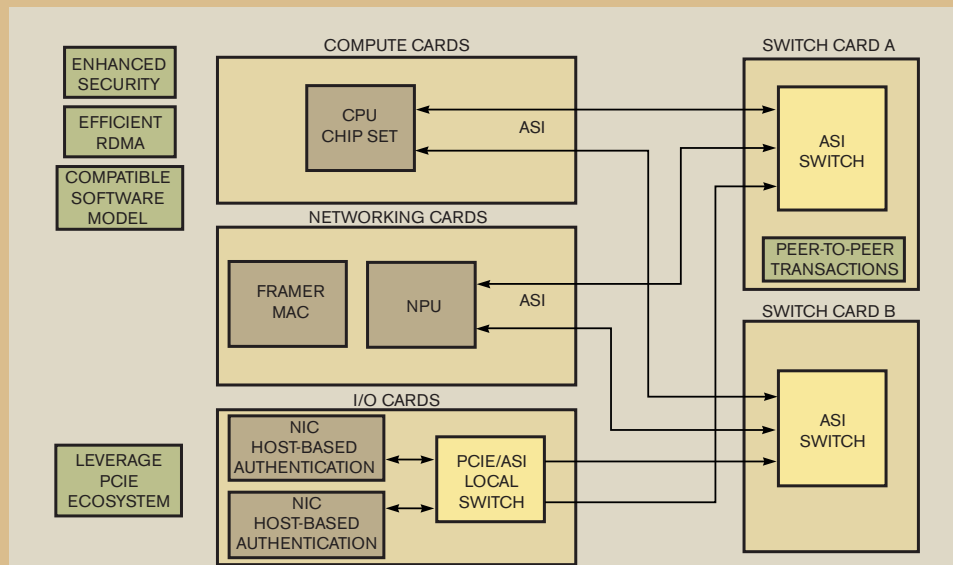


Figure A Through enhancements in the transaction layer, ASI extends the capabilities of PCI Express to support a variety of converged computation and communication applications.

Gbps flavor is likely to find the broadest near-term use. Visit the InfiniBand Trade Association for more background.

Still, the volume of IB cluster installations is relatively small compared with total cluster deployments. Most deployed, nonproprietary clusters now use Myrinet technology from small, privately held Myricom. The company supplies Myrinet host adapters and switches to most of the major server vendors, including Hewlett-Packard, Sun, and IBM. The Myrinet technology that's now available is a full-duplex, 2-Gbps PHY (physical layer) and a set of routing and link specifications that ANSI has standardized. Myricom adds a proprietary set of protocol and software layers for clustering.

You can get a quick look at leading-edge cluster deployments at the Top500 Supercomputer Sites' Web site. The site biannually releases a new list of the most powerful high-performance-computing installations in the world. If you select the database tab, you can sort the data by interconnect type, and you will see the success that Myrinet enjoys. But note that the categorization is imperfect because it classifies at least a few systems that have "IB" in some form as having "mixed" interconnects, and some Myrinet systems may suffer the same fate.

In any event, it's clear that Myrinet, IB, and FC all have data-rate advantages over 1 GBE. Still, Broadcom believes that a lot of 1-GBE business exists in storage and clustering, and the company seems in no hurry to announce 10-GBE products. But most believe that success for Ethernet will require a move to the faster 10-GBE flavor. And the equally challenging issue for Ethernet is the latency and best effort service of both Ethernet and TCP/IP.

LATENCY IS KEY

Many ways exist for defining, specifying, and benchmarking latency, and little consistency exists when it comes to interfaces and types of networks. Still, FC products regularly claim latency of less than 1 μ sec. Both Myrinet and IB also offer latencies of much less than 5 μ sec and into the 1- to 3- μ sec range. Standard Ethernet networks feature latencies greater than 50 μ sec and even up into the hundreds of microseconds.

The latency limitation associated with Ethernet comes from both the collision-based MAC protocol and TCP/IP. TCP/

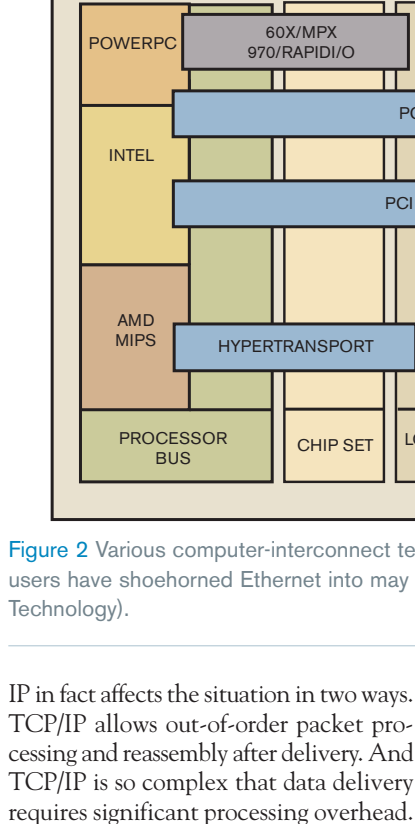


Figure 2 Various computer-interconnect technologies fit into various niches, although users have shoehorned Ethernet into many places other than LANs (courtesy PLX Technology).

IP in fact affects the situation in two ways. TCP/IP allows out-of-order packet processing and reassembly after delivery. And TCP/IP is so complex that data delivery requires significant processing overhead.

IB proponents, such as Mellanox, claim that, at 10-GBE rates, a host processor can spend as much as 90% of its time on TCP/IP processing. Mellanox's Rado claims that the company's IB chips, running at full wire speed, require only 3 to 4% of the CPU's cycles for overhead. Chuck Seitz, chief executive officer of Myricom, makes an even more astounding claim. According to Seitz, two Myrinet systems can transfer a 1-Mbyte block of data from user memory on one system to user memory on another system, and the two system CPUs cumulatively spend 0.3 μ sec on the transfer.

Engineers and IT specialists have long understood the problem of Ethernet-protocol complexity. In the days of 16-bit CPUs, an Ethernet card dedicated a processor to the TCP/IP tasks. Over time, host CPUs became sufficiently powerful that TCP/IP became a host's insignificant chore. Now, with the move to 1 GBE and then 10 GBE, TCP/IP overhead has again become an unacceptable burden on host CPUs in many applications and an issue that storage or cluster networks must face.

These days, chip vendors are integrating TOEs into their products to reduce the impact of the protocol. In fact, all of the Ethernet proponents, including Broad-

com, NetEffect, Siliquent, and Astute Networks, list TOE as a standard feature. But TOE alone will not make Ethernet an effective storage or cluster fabric. Maule of NetEffect claims that packet processing associated with TCP/IP is responsible for only about 35 to 40% of the total overhead associated with Ethernet. Maule claims that tasks such as intermediate buffer copies are responsible for 20 to 25% of the overhead. And he adds that operating-system overhead—moving into and from kernel space and handling interrupts—can be responsible for as much as 40% of the overhead. Maule states, "You must reduce overhead by 90% or more."

Maule points to the new iWarp, or high-speed-Internet, standard as the answer. The standard builds on RDMA and TOE, and a consortium at the Inter-Operability Laboratory at the University of New Hampshire Research Computing Center is nurturing it. The iWarp standard enables an "operating-system-bypass" technique that allows applications running on two computers to exchange data blocks with no intervention from the operating system. Maule claims that TOE handles the protocol overhead; RDMA, the buffer-copy problem; and iWarp, the operating-system issues.

CAN YOU BUY IT?

The 10-GBE story sounds promising, but can you buy it, and will it work? Siliquent was the first to announce chips



a full year ago. The \$500 SLQ1010 and the 4-Gbps SLQ1004 are available, but Siliquent can't yet claim that equipment vendors are shipping products based on the chips. But the company's Vogt expects customer shipments to begin shortly. She claims that the chips fully support iWarp and that latency with iWarp is now less than 10 μ sec. Moreover, she believes that they will get to 5- μ sec latencies.

Meanwhile, NetEffect last November announced an iWarp product, but Maule admits that the chips have yet to ship but promises them this year. You can now buy 10-GBE board-level products with TOE capabilities. Both Neterion (formerly, S2io Inc) and Chelsio have such products, although neither now claims iWarp support. Hewlett-Packard has begun to ship products based on the 1-GBE Broadcom BCM5706 and to espouse the converged-network philosophy. Broadcom also has a 2.5-Gbps version of the C-NIC family available.

Astute Networks may have been further along the 10-GBE chip-development path when a year ago it demonstrated the Pericles chip based on 10 Tensilica RISC cores. But that first chip included SPI-4 (System Packet Interface, Level 4) interfaces and targeted the system side of the 10-GBE link. The company has since refocused on the storage-appliance side of the SAN application and is working on its next-generation offering. Jon Siann, Astute's vice president of marketing, claims that the company learned the hard way that TOE and RDMA support is insufficient to win customers. The company's experience yielded the decision to focus on SANs in which Astute will deliver complete sets of storage software, such as mirroring and data-migration applications. Siann does not believe that one IC design can succeed in both the cluster and the storage segments. "You are not going to win both," he says.

In the short term, designers may have no choice but Myrinet and IB for clusters and IB and FC for storage. Despite many pro-Ethernet statements, Dell's Golden states, "For the next few years, IB may be the best thing going." IB does seem to be on a roll. Mellanox's Rado goes so far as to say, "I challenge the theory that Moore's Law can eliminate the TCP/IP-overhead problem of Ethernet." Mellanox officials believe that, at the very least, the

MORE AT EDN.COM

+ We encourage your comments!
Go to www.edn.com/0500707cs and click on Feedback Loop to post a comment on this article.

+ Read more of what Maury has to say on many topics at his blog, www.edn.com/ontheverge.

silicon needed to offset the disadvantages of Ethernet will keep IB prices in an advantageous position. Moreover, Rado points out that IB will double rates before 10 GBE can come down the cost curve.

An announcement that should occur just before press time, slated for the International Supercomputer Conference in Heidelberg, Germany, might change things, however. Myricom is planning the next generation of Myrinet: Myrinet-10G. The 10-Gbps technology will migrate to the 10-GBE PHY and maintain Myricom's clustering-protocol and software layers. Essentially, the most common clustering technology in the market will be faster than IB.

Myricom plans to launch the technology at aggressive prices. NICs will sell for \$795, and switches will sell for \$400 per port. Moreover, the new ICs that the company developed for the launch can run either the Myrinet protocols or TCP/IP, so the new products can carry Ethernet traffic. Myricom has always assigned Ethernet MAC addresses to its products and uses what appear to be standard Ethernet drivers to support the products. The Myrinet-10G launch will also make Myricom a full-fledged chip supplier.

Chief Executive Officer Seitz claims that Myrinet has a huge advantage over IB. He states, "For InfiniBand, RDMA is not only not the answer; it's part of the problem." He claims that the IB standard is flawed when it comes to how RDMA is implemented and that the flaw both drives up memory requirements and hurts performance because an application can-

not move blocks of data from user space in one machine to user space in another. NetEffect's Maule essentially concurs. NetEffect began life as Banderacom, an IB player, and company officials believe the lessons it learned in IB serve it well in pursuing iWarp. Maule claims that IB has a user-level direct-access problem.

On the storage side, meanwhile, it's a fair question to ask why FC needs a replacement. One theory holds that TCP/IP-connected Network Attached Storage (NAS) appliances are far cheaper than SAN systems, and iSCSI enables remote management of and access to NAS appliances. Still, even in the SAN environment, SAS (Serial Attached SCSI) or SATA (Serial ATA) will shortly replace the native FC-based drives vendors are deploying in SAN boxes today. The FC fabric of a SAN differs from the FC loop in a drive array at the PHY level, but the two share the storage protocols. As the drives move to SAS, momentum will likely emerge to migrate away from FC because it will become yet another bridge interface in enterprises. Still, it's tough to find anyone other than Mellanox that believes IB will take over the SAN, and FC may be a healthy market for years. **EDN**

FOR MORE INFORMATION

- Astute Networks**
www.astutenetworks.com
- Broadcom**
www.broadcom.com
- Chelsio Communications**
www.chelsio.com
- Cisco Systems**
www.cisco.com
- Dell**
www.dell.com
- Fibre Channel Industry Association**
www.fibrechannel.org
- Hewlett-Packard Co**
www.hp.com
- IBM**
www.ibm.com
- IDT**
www.idt.com
- InfiniBand Trade Association**
www.infinibandta.org
- Internet Engineering Task Force**
www.ietf.org
- InterOperability Laboratory at the**
- University of New Hampshire Research Computing Center**
www.iol.unh.edu/consortiums/iwarp/
- Mellanox Technologies**
www.mellanox.com
- Myricom**
www.myricom.com
- NetEffect Inc**
www.neteffect.com
- Neterion**
www.neterion.com
- PLX Technology**
www.plxtech.com
- RDMA Consortium**
www.rdmaconsortium.org
- Siliquent**
www.siliquent.com
- Sun Microsystems**
www.sun.com
- Top500 Supercomputer Sites**
www.top500.org
- Topspin Communications**
www.topspin.com



You can reach Editor at Large **Maury Wright** at 1-858-748-6785, 1-858-679-1861 (fax), and mgwright@edn.com (e-mail).