

BEYOND 10 GIG ETHERNET

40 and 100 Gigabit Ethernet: Ready for Real-Time?

With 10 Gigabit Ethernet just ramping up in deployment, the natural question is, “What about 40GbE and 100GbE?” In terms of real-time applications, the needed tweaks are being made to 10GbE, but for now 40 and 100GbE remain in the server space—for now.

by Rob Kraft
AdvancedIO Systems

While 10GbE is barely off the launching pad in terms of broad deployment in just about any market space, recent excitement has turned to 40GbE and 100GbE. It’s not a stretch to acknowledge that real-time embedded engineers, who are often also technophiles, are likely to be seduced by the allure of such high bandwidth technology.

So, the question is: should those of us in the real-time space shelve the just-ordered 10GbE technology and start designing 40GbE or 100GbE into our next-generation bandwidth-hungry applications? Has 10GbE already become passé?

In December 2007, the IEEE P802.3ba 40 Gbit/s and 100 Gbit/s Ethernet Task Force was formed out of the High Speed Study Group (HSSG) that had been working since 2006. At the time of writing this article, the most recent release from the task force was P8023.ba Draft 1.2, on February 10, 2009. The targeted date for a ratified standard is June 2010. Current predictions are that 100GbE will “take off” in 2013. 40GbE is expected to ramp up sooner. For reference, Table 1 shows how 10GbE rollouts compared to predictions and gives some data about 1GbE rollouts at selected points after the respective standards’ ratification dates.

Factors Driving 40GbE and 100GbE

Increases in Internet video traffic are driven by sources such as YouTube, high-definition IPTV, video conferencing, and

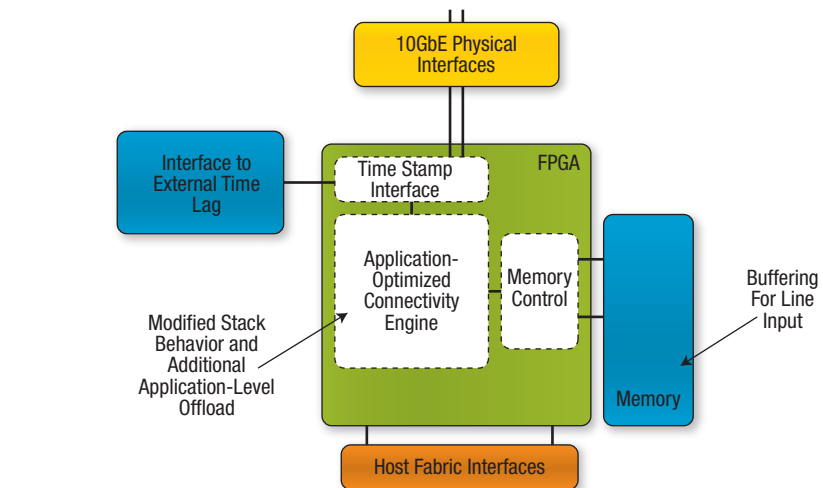


Figure 1 Elements of an FPGA-based solution that implements features required for real-time 10GbE connectivity.

enterprises migrating from private networks to Internet-based VPNs. To adequately service the increasing bandwidth demands of their customers, ISPs need to increase their backbone bandwidth at a ratio that may be 4x to 10x the customer’s needs. In addition, large search engine and social networking companies are eyeing 100GbE as a solution to their bandwidth needs for inter-data-center network aggregation.

Simultaneously, 40GbE (and good “old” 10GbE) are being driven by the growing needs for data movement between servers and other computing applications within data centers. The higher bandwidths not only solve data aggregation

and improve the flow of information, they also can reduce cabling infrastructure—a non-trivial problem given the number of servers in a data center.

A recent report cited the number of servers in 2007 to be 11.8 million in the U.S. and 30.3 million worldwide, up more than 4 times from a decade earlier. Analysts also predict that the U.S.’s 6600 data centers (where most servers are hosted) will need replacing or retrofitting in the next several years. Clearly, there are lots of servers to interconnect and lots of “voltage” behind the market force.

To support the 40GbE and 100GbE data rates, there have been several inno-

variations. Some have been driven by the fact that current technology does not permit the transmission of 40 Gbits/s or 100 Gbits/s in a single stream over any optical fiber or copper wire. For instance, in the current baseline, 100GbE would be transported in parallel over cables consisting of 10 fibers, or 10 wires, or using 4 wavelengths in the case of single-mode fiber. Except for increased data rate, there are no changes proposed at the Ethernet MAC layer (compared to 10GbE). Still, the baseline has introduced the concept of multiple lanes and multi-lane distribution (MLD) at the Physical Coding Sublayer (PCS). This was done to accommodate combinations of differing numbers and speeds of parallel electrical lanes and media lanes (fibers, wires, or light wavelengths), and to decouple those two numbers since electrical and optical technologies will develop at different rates.

Among other changes is an evolution of the now-familiar XAUI (10 Gigabit Attachment Unit Interface), used for on-board signaling, into XLAUI (40 Gigabit) and CAUI (100 Gigabit). The 'XL' and 'C' correspond to Roman numerals for 40 and 100. To accommodate the higher rates (10.3125 Gbaud per lane, compared to XAUI's 3.125 Gbaud per lane), XLAUI and CAUI use 64B/66B encoding, which has a reduced overhead (3%) compared to 8B/10B (20%) used in XAUI. 40GbE is handled in 4 XLAUI lanes, and 100GbE in 10 CAUI lanes.

Application-Level Characteristics

At the moment, the driving applications involve aggregation and distribution of data. But eventually, the data has to terminate at processors or other devices. The termination problem is challenging even at 10GbE (sometimes even at 1GbE), where processors get clobbered by the effort of running the protocol stacks. It stands to reason that the termination problem will be 4x to 10x worse for the case of 40GbE and 100GbE. In the commercial space, the solution at 10GbE typically involves forms of protocol offload, whereby some or all of the protocol processing elements are farmed out to a coprocessing ASIC. However, there are a variety of application characteristics unique to many real-time I/O applications that use 10GbE. These

characteristics are not encountered in the server space, so server space solutions do not address them.

Real-time test and measurement, scientific, defense, hardware-in-the-loop simulation, and other multi-sensor systems rely on incoming data being accurately time stamped. The time-stamping aligns data arriving from multiple sources or sensors to permit detailed off-line analysis, as inputs into models in real-time processing, or to tightly control the release of data in complex simulation systems. At the higher-performance end, the CPU cannot time-tag data with accuracy and precision since, by the time the packets reach this point, they have gone through several non-deterministic interfaces. A solution is to stamp the packets at the 10GbE interface ingress point, before they ever get to the CPU.

Most real-time applications begin as a stream of digitized analog real-world sensor signals in a control, automation, communications or measurement/analysis system. The sampled data passes through signal processing algorithms such as filtering, FFT, decoding and many others, and is subsequently passed on to other processing functions. Many signal processing algorithms correct for or tolerate scattered errors and noise in the signal stream, but choke when faced with a consecutive stretch of missing data.

Ironically, the typical behavior of the standard Ethernet protocol stack can transform benign scattered errors into a swath of missing data that chokes signal processing algorithms. This occurs because the protocol will discard entire packets or messages, which could be 1500, 9000, or up to 64000 bytes long, if an error is detected in a checksum or if a message arrives incomplete. And yet the source of the checksum error may be just one or two data bytes, or in the packet header. The solution is to modify the stack behavior to avoid dropping the packets in the presence of the CRC or checksum errors, and to allow them to pass to the signal processing stage.

High-throughput real-time instrumentation, communication and other sensor processing systems can receive multiple back-to-back packets burst at full line speed on a regular or even a sustained basis. In such cases, there are no spare cycles for flow control or retransmission requests

when the receiving Ethernet interface is momentarily unable to access host system memory to store the incoming data. The result is unacceptable permanent packet loss. This characteristic is relatively uncommon in server systems that have much softer real-time requirements, thus giving opportunities for flow control or retransmission when required. Therefore, the cost-optimized solutions targeted at the server space do not have to be designed to accommodate these regular long-duration bursts.

To address the requirements above, 10GbE technology must implement features including interfaces for precision time-stamping, local memory to accommodate large full-rate inbound bursts and outbound data staging, and the ability to customize stack behavior for receiving real-time sensor data (Figure 1).

These real-time adaptations will become even more complicated to implement for 40GbE and 100GbE and need to be solved before solutions are matured. Among the challenges:

- Incoming buffers need to be larger—4x to 10x the data arrives in the same time period as before.
- Likewise, external memory and controllers need to operate at a higher bandwidth.
- Interfaces like CAUI require 2.5x the number of high-speed pins of XAUI, increasing the number of high-speed I/O pins required in devices and complicating PCB routing.
- Internal processing bandwidth needs to increase correspondingly to realize protocol offload and other processing required for Ethernet termination.

Implications for Real-Time 40GbE and 100GbE

Because 40GbE and 100GbE technologies are currently driven by the massive server-type markets, the ASIC-based solutions, out of sheer volume necessity, will target those applications, just as they have in the 10GbE case. As pointed out, those applications have some fundamentally different requirements from the higher-end real-time embedded applications. The latter will therefore need to use solutions based on programmable logic, which allows the solutions to be customized to the problems in the application space, while

Specification	Ratification Year (y0)	Switch Port Shipments				
		Year y0+3	Year y0+4	Year y0+5	Year y0+7	Year y0+9
10GbE (802.3ae) - predicted from 2006	2002		1.8M			12M
10GbE (802.3ae) – actuals and 2008 predictions	2002	150K	320K	<1M		7M
1GbE actuals (1998)	1999	20M			80M	

Table 1 The 10GbE switch port shipment predictions and actual shipments in January 2006, and revised predictions for 2011. The actual switch port shipments for 1GbE are also shown. Source: Dell’Oro, Infonetics

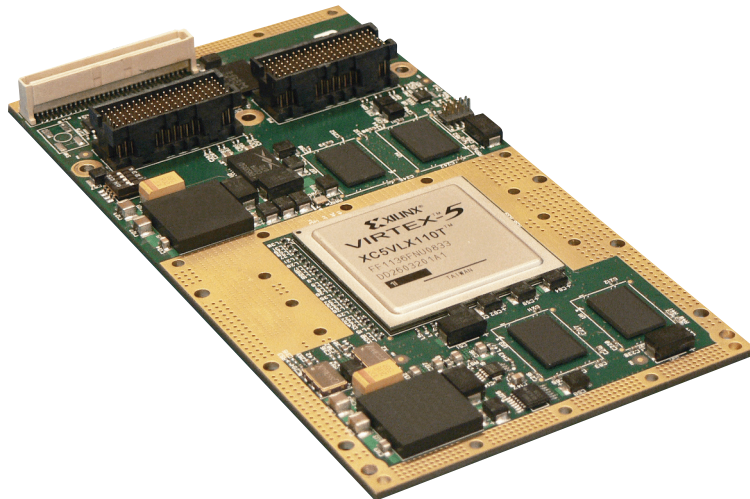


Figure 2 The AdvancedIO Systems V1120 Dual-port 10GbE conduction-cooled rugged XMC module, designed following the new VITA 42.6 standard, implements the architecture shown in the block diagram of Figure 1.

maintaining a standard external software interface for compatibility with the Ethernet ecosystem. Figure 2 shows one such solution for 10GbE systems: AdvancedIO Systems V1120 dual-channel conduction-cooled 10GbE interface module, based on the Xilinx Virtex-5.

Without question, 40GbE and 100GbE will arrive in the real-time embedded space, but the arrival is still some years away and needs to be accompanied by the same kinds of innovation that make 10GbE suitable for the space. The good news is that, at 10 Gbits/s, Ethernet finally has sufficient bandwidth for most real-time high-speed applications, and there are real-time focused solutions existing today that you can use to get on board the Ethernet bandwagon. Once aboard, you can smoothly ride the Ethernet speed curve to 40GbE and beyond as your requirements scale, avoiding the software and architectural upheaval that resulted from previous iterations of integrating different high-speed technologies. ▲

AdvancedIO
 Vancouver, BC.
 (604) 331-1600.
[\[www.advancedio.com\]](http://www.advancedio.com).