

HIGH-SPEED



PHOTO COURTESY SKY COMPUTERS INC

DATAPATHS

BYPASS BUS BOTTLENECKS

WARREN WEBB, TECHNICAL EDITOR

For years, experts have predicted the end of conventional bus architecture for real-time systems because of problems with bandwidth, scalability, and determinism (the ability of a real-time system to react to an external interrupt in a known time period). Yet, when you consider today's high-performance systems, you find that most are based on standard bus technology because designers have developed sophisticated techniques to bypass bandwidth problems and increase system performance far beyond backplane transfer rates. Raceway, Skychannel, and Myrinet for the VMEbus and the Sebring Ring for the PCI bus exemplify high-speed interconnection techniques. Each provides alternative parallel connections to meet the data-transfer requirements of mission-critical applications, such as sonar, radar, and medical imaging.

With high-performance computing requirements now exceeding 1 Gflops and individual computing elements operating at 200 to 300 Mflops, more multiple-processor systems are becoming available. The bandwidth required to transfer data between processing nodes increases proportionally with the speed of the processor. Yet, a conventional bus system has a fixed maximum bandwidth. As you add boards to a bus system to increase performance, you reduce the bandwidth available to each function. High-performance systems need scalable bandwidth, in which the bandwidth increases as you add nodes to the system. The 80-Mbyte/sec speed of the current VME-

Today's standard backplane bus systems offer low-cost, off-the-shelf hardware and design flexibility but have severe bandwidth limitations. To increase data rates and retain the advantages of the bus system, designers are bypassing the bus and directly transferring data between subsystems.

bus and even the 132-Mbyte/sec of the PCI bus easily become bottlenecks when your design requires multiple data flows to keep up with today's processors.

Other real-time problems are the fact that external events happen whenever

they want, and a bus can respond to only one at a time. If a data transfer is occurring when a higher priority transfer needs the bus, circuitry or software must suspend the lower priority data until the high priority data finishes and then retransmit the blocked data. This

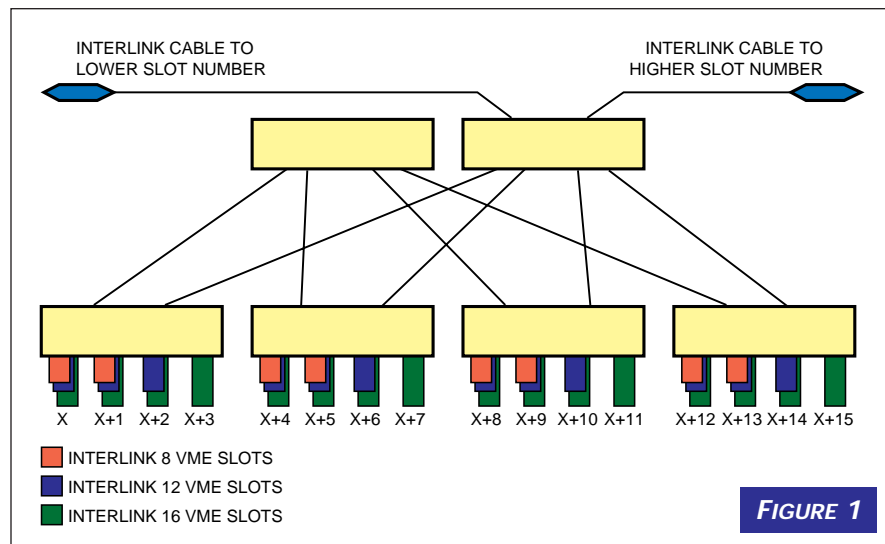


FIGURE 1

Raceway crossbar chips form a switched fabric network that plugs over the P2 pins of as many as 16 VMEbus slots.

INTERCONNECTION TECHNIQUES

contention for the bus is proportional to the number of nodes in the system. The latency of suspended data may become intolerable in some real-time situations.

Raceway, Skychannel, and Myrinet use direct point-to-point connections or crossbar switches, whereas the Sebring ring, as its name implies, employs a ring architecture. Regardless of the architecture, all of the techniques multiply their basic bandwidth by allowing several data transfers to occur simultaneously. VMEbus technologies attach to the user-defined P0 or P2 connector pins, whereas the Sebring ring replaces standard PCI-to-PCI bridge chips. All of the systems continue to operate with standard off-the-shelf boards.

Raceway on VME P2

One of the first interconnection schemes was a switched fabric or network architecture in which datapaths between computing nodes change dynamically to support multiple simultaneous data transfers. In a network architecture, you can connect any node to any other node through datapaths that resemble the threads in a cloth; hence the name "switched fabric." Engineers at Mercury Computer Systems Inc exploited the network



FIGURE 2

Pentek's \$1795 model 7106 Raceway master/slave PMC interface board exemplifies third-party Raceway-compatible products.

architecture to develop Raceway Interlink for the VMEbus. Raceway Interlink is an approved ANSI/VITA (VME International Trade Association) standard (ANSI/VITA 5-1994) using the undefined pins on the VME Pin 2 connector to interconnect boards and provide simultaneous high-speed datapaths. Interlink printed circuit modules come

in various sizes that you can combine to create any-length Raceway interconnections up to all 21 slots in a VME chassis.

Raceway uses a custom CMOS ASIC crossbar switch to build the interconnect fabric. The crossbar, which Mercury developed, contains six data ports allowing three simultaneous paths between any possible pair of ports. The crossbar can also broadcast data from one input port to as many as five output ports. Each port has a 32-bit-wide datapath plus 5 control bits. Data transfers through the crossbar are 4 bytes wide at 40 MHz, yielding a 160-Mbyte/sec maximum data rate per path. You can combine several identical crossbars in various topologies to create a larger interconnect fabric (Figure 1).

A master node initiates a Raceway data transfer by sending a two-word packet header to the slave node. The first word in the header

contains as many as nine 3-bit codes indicating the preferred exit ports at each crossbar in the path and the packet priority. The second word defines the DRAM block starting address at the destination node and the data direction. The slave node then responds with an acknowledgment signal, indicating that the master node has established the transfer path. Finally, the master node transfers as many as 2048 bytes of data from source to destination. Byte count is not part of the header; the master node activates an end-of-packet signal to indicate the last byte and to disconnect the path.

Although the data within each packet travels at 160 Mbytes/sec, the effective bandwidth of a Raceway path depends on the packet length and the number of crossbar switches between the master and slave. Raceway requires a latency of approximately 150 nsec plus 125 nsec for each crossbar to establish a path and start a data transfer. Because Raceway must re-establish the

@ a glance

- Traditional bus architecture is a severe data-transfer bottleneck in high-performance multiprocessor systems. The available bandwidth decreases as you add to the system.
- Today's real-time VMEbus systems use switched crossbar networks to bypass the bus and increase effective data-transfer bandwidth to gigabytes per second.
- New developments in ring architecture promise to propel the PCI bus into the exclusive domain of VMEbus. New features include multiple datapaths, 512 slots and fault tolerance.
- Reacting to the competition, the VMEbus has increased its bandwidth from 80 to 320 Mbytes/sec and is on track for more than 1 Gbyte/sec.

INTERCONNECTION TECHNIQUES

transfer path for each packet, the effective transfer rate is proportional to packet size. At a maximum 2048-byte packet size and five crossbar switches, the effective transfer rate is approximately 130 Mbytes/sec. The transfer rate falls slightly to 120 Mbytes/sec with a 1024-byte packet size but drops to less than 50 Mbytes/sec when the packets are 128 bytes. Contention at any of the crossbar ports along the path increases latency.

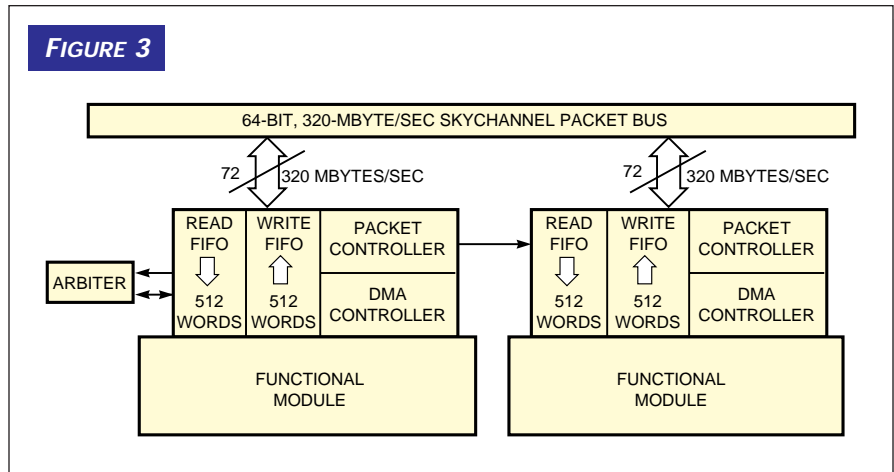
Raceway uses a priority scheme to arbitrate contention and maintain determinism for critical data transfers. If two data transfers attempt to use the same crossbar port, Raceway examines the packet priority bits to select the higher priority and suspend the lower priority transfer. If the priorities are equal, the crossbar selects the transfer with the highest entering port number and blocks the other. Blocked transfers may attempt to regain the path at the end of each packet.

Several VMEbus-board manufacturers have adopted the Raceway Interlink for high-bandwidth applications. Echotek, Myriad Logic, Pentek, and Vmetro are just four of at least 15 manufacturers supplying VME boards or mezzanine cards adhering to the standard (Figure 2). Raceway Interlink is an open standard that requires no licensing.

Skychannel packet bus

A more recent development than Raceway is Skychannel, which Sky Computers developed. Sky based the communications technique on packet-bus architecture. Skychannel is also an ANSI standard, which ANSI approved in October as ANSI/VITA 10-1995. With the same 40-MHz clock as Raceway, Skychannel transfers data at 320 Mbytes/sec over a wider, 8-byte path. You can use a bus or a multipath crossbar network to connect Skychannel nodes. As a bus, Skychannel solves some of the VMEbus bottlenecks by increased bandwidth and time-shared bus transmissions; however, high-performance multiprocessing systems rely on crossbar networks to provide simultaneous transfers for a scalable communications structure.

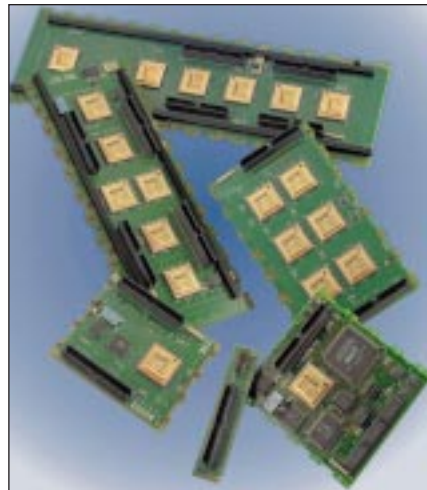
Packet-bus architectures attempt to



Each Skychannel node interface includes a packet controller, FIFO buffers, and an optional DMA controller.

reduce transfer latency and blocking in a Skychannel system by requiring packet controllers and bidirectional FIFO buffers at each node. This additional complexity allows several data transfers to interleave communications and maintain Skychannel's bandwidth of 320 Mbytes/sec per path. The FIFO buffers collect data at any rate and transmit at the full 320 Mbytes/sec to avoid tying up the communications path with slow data transfers (Figure 3).

Although circuit-based architectures must request a datapath, wait for verification, and then send the data, packet-based architectures may initiate a transmission when only the first seg-



The effective bandwidth increases as you add Raceway slots.

ment in the path is available. If the data packet encounters an obstacle along the path, the data packet waits in FIFO buffers and advances toward the destination, as each segment of the path becomes available. Forward progress occurs even in the face of heavy traffic. If no contention occurs, the FIFO design allows data transmission to begin before the destination receives the entire packet.

Skychannel packets contain an address word, a variable-length data payload, and an error-detecting checkword. The address word contains 28 control bits plus a 44-bit address field (16 Tbytes). Each of the 72-bit-wide data-transfer words contains 8 data bytes plus 1 control byte. The maximum data size is 1024 bytes, or 128 words. A control bit in the last data word signifies the end of packet. Finally, the hardware computes a checkword and appends it to the packet for error checking at the destination node.

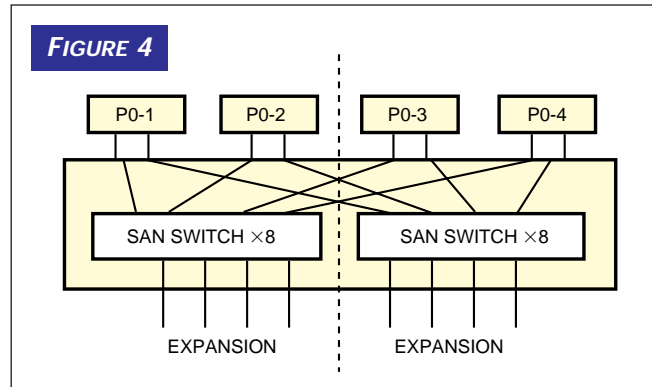
As with Raceway, the Skychannel specification provides for a VME backplane overlay using the user-defined pins on connector P2 to implement a multipath crossbar network. Sky has developed a backplane module containing a 10-port crossbar to connect as many as eight VME boards plus two external paths. In the best case, the crossbar supports five simultaneous data-transfer paths—a whopping 1.6-Gbyte/sec aggregate data rate.

However, the aggregate data rate

INTERCONNECTION TECHNIQUES

does not account for overhead and latency. Even without contention, each segment of the path requires switching overhead plus address and checkword latency. At a maximum packet size of 1024 bytes, the effective transfer rate is approximately 306 Mbytes/sec per path. The effective rate drops to 160 Mbytes/sec when the packet data size is 64 bytes.

Standard FPGA chips for the packet and optional DMA controllers implement the common interface connections to Skychannel. Off-the-shelf FIFO buffers and transceivers complete the multichip node interface. The specification allows partial interfaces that



Myrinet's four-slot P0 backplane overlay module uses eight-port crossbars to provide two independent paths. Myrinet can coexist with P2 overlays, such as Raceway and Skychannel.

support only a subset of the bus-interface protocol. For example, a memory board may not require master capabilities.

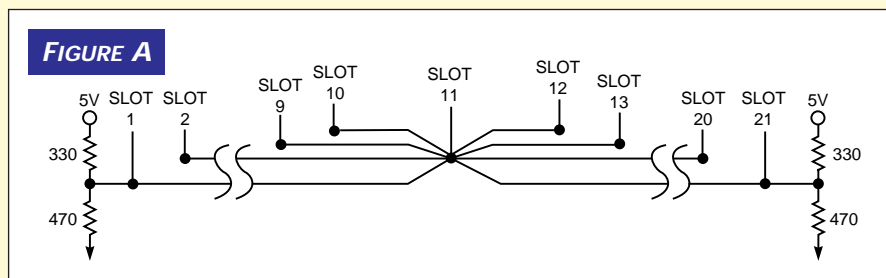
In use for years as a LAN, the Myrinet system-area network (SAN) is another high-speed communications network for the VMEbus. Myricom and CSPI jointly developed the Myrinet SAN for VME, and ANSI/VITA are working on approving the technique as an open standard (VITA 26-199x). The companies designed Myrinet to coexist with any P2 connector protocol, such as Raceway or Skychannel. Myrinet SAN uses the P0 connector or the front panel for access to the communications path. The front panel datapath operates at a maximum of 160 Mbytes/sec/channel, whereas the P0 connector path supports transfers of

VME ZOOMS TO 320 MBYTES/SEC

Data-transfer rates on the VMEbus have increased with each new VME International Trade Association (VITA) standard. VME64, which VITA approved in 1996, increased data rates to 80 MHz by doubling the path width to 64 bits. Another doubling of the data rate to 160 Mbytes/sec reduces the transfer protocol from a four-edge handshake to a two-edge (2eVME). VME320, which will become available this year, is an all new backplane design and backward-compatible protocol to increase data transfers to 320 Mbytes/sec.

Available shared-bus backplanes are wired from slot to slot with a terminating resistor network at each end of the bus. The distributed inductance and capacitance of a fully loaded bus acts like a low-impedance transmission line. The propagation time for a 21-slot backplane can be as long as 8 nsec. Line drivers cannot match the low impedance of the transmission line, and ringing and reflections increase with the data rate.

Arizona Digital and Bustronic have jointly developed a backplane based on a star configuration (Figure A). A signal driven from Slot 1 goes to Slot 11 and then radiates out to all other slots. All the capacitance concentrates at Slot 11 instead of having a transmission-line effect as in a conventional shared-bus backplane. The result is that the equivalent circuit of the backplane is a simple lumped 200-pF capacitance. The lumped capacitance grows slightly as you add circuit cards, but the equivalent circuit remains the same. The VME320 specification



The VME320 backplane's star pattern eliminates the transmission-line effects of a standard slot-to-slot layout.

adds undershoot Schottky diodes to kill ringing resulting from trace inductance in series with the lumped capacitor.

The rise and fall times of the VME320 bus signals are clean and monotonic because of the lumped capacitance. A source-synchronized transfer protocol, which eliminates one of the remaining data-strobe edges, increases the data rate to 320 Mbytes/sec. The developers claim that 320 Mbytes/sec signals even look clean with the circuit card on a 14-in. extender board.

Although the backplane and specification are available now, to produce VME320 bus products, vendors have to wait for the release of the Universe III bridge chip, which Tundra expects to introduce this year.

Even 320 Mbytes/sec is not the upper limit of the VMEbus. According to Ray Alderman of VITA, this bus can operate at 533 Mbytes/sec without errors. At the Real Time Conference in Santa Clara, CA, in January, Alderman predicted VMEbus data rates of 1 Gbyte/sec by 2000.

INTERCONNECTION TECHNIQUES

320 Mbytes/sec (**Figure 4**).

Each Myrinet SAN data-link channel contains 8 data bits, 1 data-control bit, and a flow-control bit. The data-control bit determines whether the byte contains data or control information, and the flow-control bit stops the data flow. Two independent channels—one for each direction—exist at each node, and Myrinet can simultaneously transfer data over each path.

The 8 data bits and the data-control bit of a Myrinet SAN are NRZ-encoded. The data-control bit is a one for data, so there is at least one transition for each byte. The path can operate asynchronously at any rate up to the maximum path data rate. The receiver looks for data transitions and clocks in the data as it arrives. The NRZ encoding cuts the frequency of any line to one-half of the data rate. For instance, the data-control bit sustains an 80-MHz frequency while transferring data at 160 Mbytes/sec.

A Myrinet packet comprises a multi-byte header, the data payload, and a 1-byte trailer. The header contains pack-

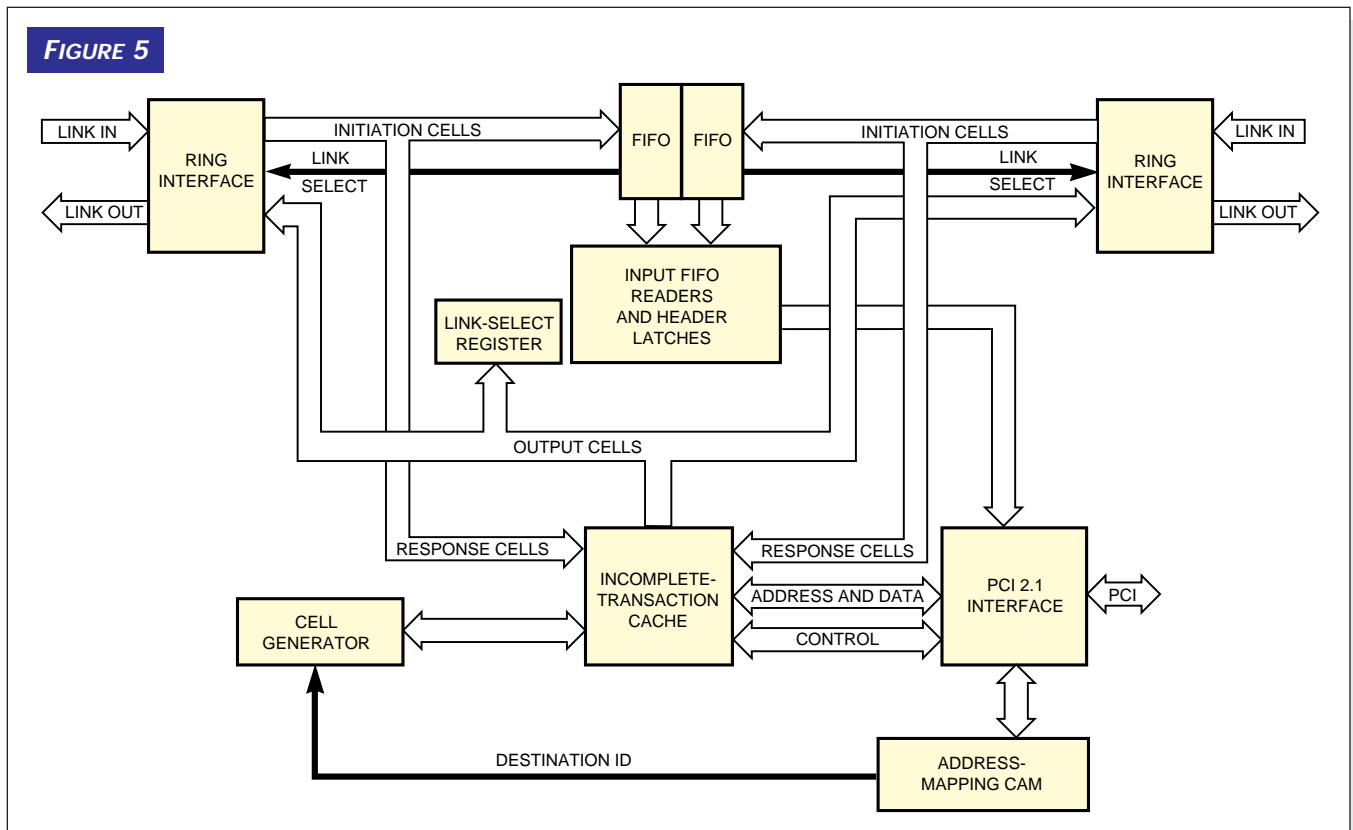
et-routing information that Myrinet examines at each switch and then shifts in preparation for the next switch. A 4-byte packet type to identify the protocol is also a part of the header. The payload can be as long as 4 Mbytes. The final byte is a cyclic-redundancy-check character that the hardware computes from all the previous bytes in the packet.

CSPI and Myricom offer the MAP-4P0, a four-slot overlay switch module that installs over the VME P0 connector. Redundant eight-port crossbar switches create as many as eight simultaneous datapaths to interconnect the two SAN data links from each slot. Myrinet SAN connections can also exit via the front panel for additional bandwidth.

VME has grabbed most of the communications network attention because most high-performance systems employ that bus. Nevertheless, do not rule out the PCI bus. With its huge stock of desktop software, experienced programmers, and low-cost silicon, PCI

deserves consideration when developing a new system. However, PCI has its problems: A limited number of slots and a single master have thus far made it a poor choice for real-time embedded systems. Sebring Systems has introduced an interconnection technique based on ring architecture that extends the number of slots to 512 and supports a best-case sustained bandwidth of 2.5 Gbytes/sec.

The Sebring Ring is a dual, counter-rotating ring network providing traffic flow in either direction (**Figure 5**). Sebring's \$59 (10,000), software-transparent SRC3266 acts as a virtual bridge between the rings and the local PCI bus. PCI modules retain their plug-and-play features and the same software drivers. Each SRC3266 acts as a traffic controller—adding, removing, or passing along blocks of data, depending on address. The data rate over each point-to-point ring segment is 532 Mbytes/sec, and the typical local-bus rate is 133 Mbytes/sec. Input and output 256-byte FIFO buffers synchronize the two rates.



Sebring's SRC3266 network interface is part of a dual counter-rotating ring architecture compatible with PCI 2.1.

INTERCONNECTION TECHNIQUES

The second ring provides four times the data throughput and adds single-fault tolerance. If a link or node in the ring fails, traffic automatically routes around the failure. By using the remaining segments of both rings, the performance loss is only 15%. The second ring also adds hot-swap capability, allowing you to isolate, remove, and replace a node while it is operating.

Addresses and data travel around the ring in variable-length asynchronous-transfer-mode-like cells containing PCI commands, addresses, data, and response codes. The Sebring chips recreate PCI transactions from cells transferred the shortest direction around the ring. The source node retains enough information to retry a transaction until it receives a response cell. If traffic is light, cells travel to and from intermediate ring nodes and exit at the destination node. When traffic is heavy, the cell may become delayed in a bypass FIFO buffer at a node as a new cell enters the ring. Additional new cells must wait for the bypass FIFO buffer to empty, assuring forward progress around the ring. Like a network-interconnection system, multiple nodes attempting to transmit to one destination can block transfers. Transmissions then continue around the ring, and the

source node removes and retries them.

The transmission latency of the ring depends on the number of nodes traversed and the amount of traffic. Each node introduces a fixed pipeline delay plus the possibility of a bypass FIFO delay. The transmission delay is less than $2*N*18$ nsec, where N is the average number of nodes in the path. For a 16-node system, the average path distance is four, yielding an average latency of 144 nsec. This latency, along with concurrent datapaths, makes the Sebring Ring a viable candidate for deterministic real-time systems.

The Sebring Ring along with Raceway, Skychannel, Myrinet and are just a few of the techniques designers are using to squeeze more performance from today's technology. Increasing digital data rates and processor speeds will dictate a constant evolution in bus bypass techniques to eliminate bottlenecks. EDN

References

1. Quinnell, Richard A, "The ever-evolving VMEbus adapts to user needs," *EDN*, Feb 17, 1997, pg 82.
2. Einstein, Tom, "Raceway Interlink—a real-time multicomputing interconnect fabric for high-performance VMEbus systems," *VMEbus Sys-*

tems, February 1996.

3. Jaenicke, Richard, "Skychannel—a high-performance communications architecture for embedded multi-processor systems," *VITA Journal*, September 1997.

4. Solomon, Susan S, "Myrinet on VME," *VITA Journal*, March 1997.

Acknowledgments

Thanks to Barry Isenstein and Greg Rocco of Mercury Computers, Richard Jaenicke of Sky Computers, and Ray Alderman of the VME International Trade Association for their assistance and insight.



You can reach Technical Editor Warren Webb at 1-619-513-3713, wwwwebb@cts.com.

VOTE

Please use the Information Retrieval Service card to rate this article (circle one):

High
Interest
594

Medium
Interest
595

Low
Interest
596

FOR MORE INFORMATION...

For information on products such as those described in this article, circle the appropriate numbers on the Information Retrieval Service card or use *EDN's* Express Request service. When you contact any of the following manufacturers directly, please let them know you read about their products in *EDN*.

Arizona Digital
Scottsdale, AZ
www.arizonadigital.com
Circle No. 301

Bustronic Corp
Fremont, CA
1-510-490-7388
www.bustronic.com
Circle No. 302

CSP Inc
Billerica, MA
1-800-325-3110
www.cspi.com
Circle No. 303

Echotek
Huntsville, AL
1-205-721-1911
www.echotek.com
Circle No. 304

Mercury Computer
Systems
Chelmsford, MA
1-508-256-1300
www.mc.com
Circle No. 305

Myriad Logic Inc
Silver Spring, MD
1-301-588-4155
www.myriadlogic.com
Circle No. 306

Myricom Inc
Arcadia, CA
1-818-821-5555
www.myri.com
Circle No. 307

Pentek Inc
Upper Saddle River, NJ
1-201-818-5900
www.pentek.com
Circle No. 308

Sebring Systems Inc
Los Gatos, CA
1-408-358-7827
www.sebringring.com
Circle No. 309

Sky Computers Inc
Chelmsford, MA
1-508-250-1920
www.sky.com
Circle No. 310

Tundra Semiconductor Corp
Kanata, ON, Canada
1-613-592-0714
www.tundra.com
Circle No. 311

VME International
Trade Association
Scottsdale, AZ
1-602-951-8866
www.vita.com
Circle No. 312

Vmetro Inc
Houston, TX
1-713-584-0728
www.vmetro.com
Circle No. 313

Super Circle Number

For more information on the products available from all of the vendors listed in this box, circle one number on the reader service card.

Circle No. 314