ATCA: Its Performance and Application for Real Time Systems

Alexandra Dana Oltean Karlsson and Brian Martin

***Abstract*—The Advanced Telecom Computing Architecture (ATCA), describes a high bandwidth, high connectivity, chassis based architecture designed principally to appeal to the telecommunications industry. The object of the exercise was to closely connect compute engines within the chassis to multiple user services brought in at the front panel. This maps closely to the needs of real time systems and the main points of the architecture are reviewed and discussed in that light. The performance of an ATCA backplane has been tested and measured using a Backplane Tester developed within a 10 Gb/s Ethernet switch project that was an early adopter of the ATCA standard. Some results from these tests are presented.**

***Index Terms*—ATCA, backplane, data acquisition, multiprocessor interconnection, real time systems.**

I. Background

DRIVEN by the 'need for speed' the trend has increasingly been away from the shared resources of a bus and more towards a point to point connectivity between processors and data sources. There are several reasons for this. Firstly, multiplexing between data sources is much easier and faster to do within a silicon bridge chip, switch or processor than it is across a backplane or cable. Secondly the data rates achievable across a bus are limited by the difficulties of maintaining equal round-trip times for each line of the bus and compensating for the signal integrity issues of driving a partially loaded bus equally as well as a fully loaded one. Then the wider the bus becomes, the more pins are required on the silicon chip. The pads on a chip are the most expensive part of the device in terms of silicon real-estate, power consumption and package size. Finally, bussed systems do not scale as extra processors are added since the available I/O bandwidth is both limited and shared.

The migration to point to point switch based systems is now taking place because improvements in signal processing have overcome many of the difficulties of digital transmission through copper interconnects. Commercial SerDes (Serialiser/Deserialisers) are today delivering upwards of 3 Gb/s per differential pair, 10 Gb/s has been demonstrated and more is on the way. Using today's technology, eight pairs, four in each direction, will deliver 10 Gb/s full duplex point to point over either a cable or across a printed circuit. Exactly the same medium can deliver two or three times this speed, once improvements in the packaging technology, which is the limiting factor at the moment, becomes available in commercial volumes of high-speed transceivers.

Manuscript received June 4, 2005; revised February 2, 2006. This work was supported in part by the European Union under Grant IST-2001-33185.

A. D. Oltean Karlsson is with the European Organization for Nuclear Research, Geneva, Switzerland, and also with the "Politehnica" University of Bucharest, Bucharest, Romania (e-mail: alexandra.oltean@cern.ch).

B. Martin is with the European Organization for Nuclear Research, Geneva, Switzerland (e-mail: brian.martin@cern.ch).

Digital Object Identifier 10.1109/TNS.2006.873404

II. Market Analysis

The CompactPCI bus [1] had been developed by the PICMG (PCI Industrial Computer Manufacturers Group) group [2] but it had not achieved the expected market share. The main market for chassis based systems is in telecommunications. CompactPCI had failed to make much impact because the boards are too small, too close together, under-powered and bandwidth limited for this application. It is fairly simple to address questions of power and form factor but in making the move towards greater speed the issue of what choice of serial technology needs to be resolved.

There are already entrenched markets for Ethernet, Infiniband, PCI Express, and more may come in the foreseeable future. The only thing in common between the different technologies is the use of 100 Ohm balanced differential pairs for the transmission lines. By providing enough of these pairs in their new standard, the PICMG group hopes to offer an infrastructure that will attract all comers.

III. ATCA Standard

There is a family of PICMG 3.x standards of which PICMG3.0 is the base specification. This defines the mechanical form factor, power and cooling parameters, backplane interconnects and the system management architecture necessary to construct a compliant backplane, chassis and plug-in boards. It also defines base fabrics for system control and management. Subsidiary specifications define fabric protocols for control and data plane communication. These include PICMG 3.1 for Ethernet, PICMG 3.2 for Infiniband and PICMG 3.3 for Star-Fabric technologies.

The board form factor is 7.25 U high by 230 mm deep and a pitch of 30.48 mm housed in a chassis that is from 10 to 12 U high depending on the choice of air flow for cooling. The cooling is designed to support up to 200 W of power per slot. The chassis width depends on the host rack that could be either a 19" instrumentation rack or 23", which is more common for telecom racks. In the first case there are 14 slots per chassis and in the second there can be either 14 or 16 slots. Two of the slots are redundant copies of each other and are the centre points for control switching and one of the data switching topologies, as illustrated later in this chapter. These are called the logical slots 1 and 2 and their physical position is not defined by the standard. Common practice puts them adjacent to each other, either

TABLE I
COMPARISON OF ATCA WITH BUS BASED STANDARDS

	ATCA	PCI (long)	VME 6U
Board Area cm2	995	316	373
Power Watts	200	10/25	30
Bandwidth I/O Gb/s	20 full duplex	4.3 66MHz 64 bits	2.4 VME 2eSST
Front panel H * W cm	30 * 2	8 * 1.2	21.5 * 2
Component Height mm	21.33	14.48	13.72

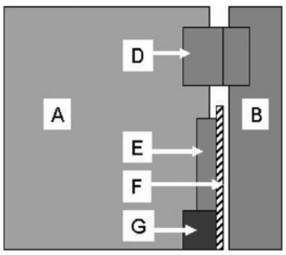

Fig. 1. ATCA Board Form Factor.

at the centre or extreme left of the backplane. Table I compares the main parameters of the ATCA standard with current bus systems.

A major departure from previous instrumentation chassis implementations is the power distribution, which is dual redundant -48 V. This results from the fact that there is no longer a single dominant voltage requirement for the electronics of choice, plus the telecommunications market long ago standardized on -48 V. Individual board voltages are therefore generated by DC-DC converters on each board. This obviously subtracts from the useful board area.

The format of the board is shown in Fig. 1. The main board, A, has space for up to four of the popular PMC daughter-board footprints although these are not part of the specifications.

There is also an optional rear transitional module B which allows for the mounting of external connectivity from the rear of the chassis. Access to the transitional module is via the connectors D. These connectors do not make contact with the backplane F but pass over the top of it. The main board connects to the backplane data and control transport connections through the connectors E. Power is drawn through connector G. The board height allows for a chassis variant where the boards are mounted horizontally within a 19" rack and having a limited number of slots for more compact applications.

The backplane carries the following interconnects.

1) Shelf Management: Management of the chassis contents is a major part of the specification since it is understood that the chassis may be housing equipment from various vendors not all of whose I/O is compatible and therefore needs to be verified before power is applied. In addition many of the boards will be running full processor operating systems with their attendant needs of booting remote IP management and environmental monitoring. This is achieved over an I2C bus.
2) Base Interface: Logical Slots 1 and 2 are dedicated to being the redundant hubs for a dual star interface using a 10/100/1000 BASE-T Ethernet interconnect to every other slot. The base interface offers a medium speed control path that parallels the higher speed Fabric Interface.
3) Synchronisation Clock Interface: There are three clocks that are bussed across each slot, two of them are Sonet/SDH clocks at 8 Khz and 19.44 Mhz. The third is user definable. The clock sources can be in any user defined slot.
4) Update channel Interface: Each board has 10 differential pairs connecting it to its neighbour. These are expected to be used for proprietary uses with proprietary protocols.
5) Fabric Interface: The standard defines two different transport architectures and variants on the theme for special purposes. The first is the Dual Star and the second the Full mesh.

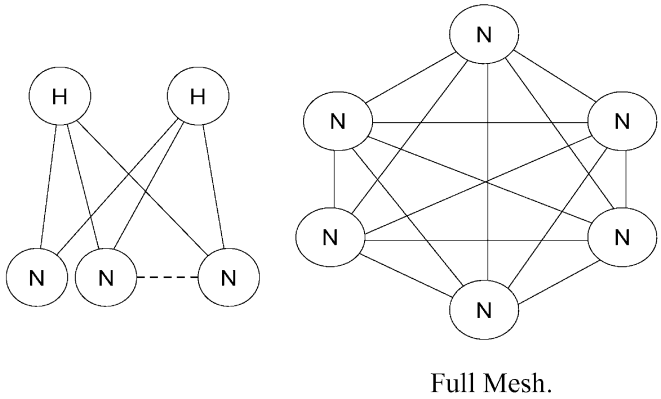

Fig. 2. Dual Star Interconnect.

In the Dual Star every Node Slot, N, supports one channel (four pairs in each direction) to each of two Hub slots, H, that reside in logical slots 1 and 2. Each Hub Slot supports up to the maximum of 15 Channels. In a Full Mesh all slots, N, are equal peers and provide one channel to every other board in the backplane. This is shown graphically in Fig. 2. It is also possible to have Dual-Dual Star configurations in which all Node Boards/Slots support one Channel to each of four Hub Boards/Slots.

A clear advantage for this approach over the bus based systems is that the devices that interface the custom electronics of the application are no longer low volume vendor specific bridges. Now the links can be driven and switched by the competitively sourced transceivers and switches appropriate for the technology of choice. For example in the case of Ethernet a node or hub board would employ integrated Ethernet transceivers and SerDes, as well as single chip switches, all of which have been developed for a mass market and are independent of any particular processor vendor.

IV. REAL-TIME APPLICATIONS

Clearly this architecture offers a powerful message passing platform for applications that can take advantage of it. Consider

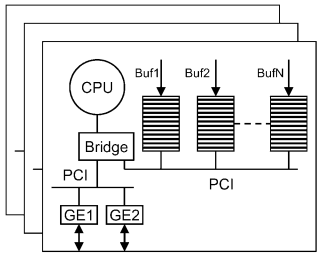

Fig. 3. Data Acquisition Element.

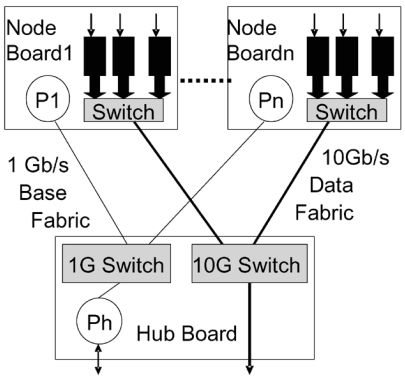

Fig. 4. ATCA implementation.

this example in Fig. 3 of a typical data acquisition tree where multiple incoming data streams are buffered and adapted to the PCI interface.

The buffered data is made available to filter processors (CPU) before being rejected and cleared or accepted and sent on to the next stage of filtering. The processor, CPU, is master of the PCI bus and requests data blocks from individual buffers (Buf1, Buf2, . . ., BufN) and dispatches them to the requesting processors over one or maximum two Gigabit Ethernet links. Typical implementations today are housed in multiple instances of a PC housed PCI bus where the system is fairly well balanced provided that the processor can manage the constant stream of requests and that the aggregate average data flow does not exceed the output rate of the Gigabit link(s).

The PCI bus is multiplexing both command and data streams and the processor is managing the message passing protocols as well as the data flow. There is very little headroom either in the processor or the bus if data rates should increase substantially beyond those predicted when the system was designed. Adding another processor cannot help since the bus is already close to its limit, the only solution would be to reduce the number of buffers managed by any one processor and increase the number of PC housings to cope.

This functionality can be mapped onto the ATCA standard to achieve an increase in performance as shown in Fig. 4.

The buffers on the Node Boards are now interfaced to 10 Gb/s Ethernet SerDes, which are multiplexed through a 10 Gb/s layer 2 Ethernet switch. Single chip solutions exist for this with typically 10 to 16 ports. Two ports are used to transport the buffer output over the Data Fabric backplane to the Hub Boards. For simplicity only one Hub Board is shown in Fig. 4. The fan-in ratio of Node boards to a Hub Board is determined by the expected traffic. The greater the aggregate traffic per Node board, the less Nodes can be supported by any one Hub. In Fig. 4 we show a fan-in of n to 1.

Being in a switched environment means that load balancing is a fairly simple process. The lowest data rates may be handled by one output link from one Hub Board. If not, or as rates increase, it is possible to allocate more than one output link from the Hub Board to the outside world. Then one can add the second Hub Board. If this is still insufficient then the Dual-Dual Star option is available by just changing the backplane and adding two more Hub Boards.

Although the transport standard is Ethernet, there is very little protocol involved here. It is merely being used as a data pump from source to destination. The processors at the buffer sources do not have the time to implement complex protocols like TCP and, even if they did, it is unlikely that the application would have the time to wait for them to work.

Error handling is thus an issue. The real-time solution to errors is at best to rapidly retry the failed transfer or more likely, to just discard the event. Rather than manage this over the 10 Gb/s Data Fabric we separate out data and control functions by sending all command and control over a completely separate network using the Base Fabric at 1 Gb/s. This relieves the processors of any data flow load and allows them to dedicate all their available CPU power for management functions.

External requests are picked up by a supervising processor, Ph, in the Hub Board and fanned out to individual Node processors, Pn, in each Node Board. The Node processor can be either a single processor if it is sufficiently powerful or, in the limit, each buffer could have its own control processor and the Node Board could then carry a switch that interconnects them all to the Hub processor. The switched system is thus free of the constraints of the bussed PC motherboard and can be scaled to position bandwidth and CPU where it is needed.

V. Limitations

Bus based systems migrating to ATCA will have to abandon interrupt and DMA driven methods in favour of message passing for both data and control flow. However real time systems frequently also require global signals such as Resets, Triggers and GPS clocks. There is only one free clock line which is inadequate for GPS since that usually supplies two clocks, one low frequency and one high frequency.

The easiest way to expand global signalling is over the Update Channel which is user definable and can be wired through from one slot to the next on each node board. This however requires that consecutive slots are occupied, either with a working board or a jumper board to ensure continuity. Alternatively one could opt for the full mesh connectivity and employ one Node Board as a distribution point for global signals to every other board. Applications that use a fully occupied compact chassis may be constrained to using an ad-hoc cable harness that interconnects boards over the rear transitional module or the front panel.

The front panel itself is the source of some concern. There are two mandatory LEDS that occupy defined positions and Ethernet application developers are currently attempting to identify an RJ45 design that will allow for up to 40 sockets to be mounted per front panel.

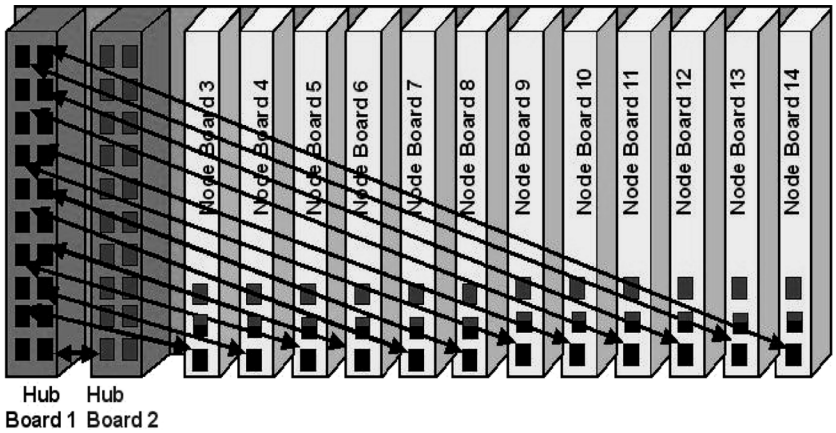

Fig. 5. Backplane Tester.

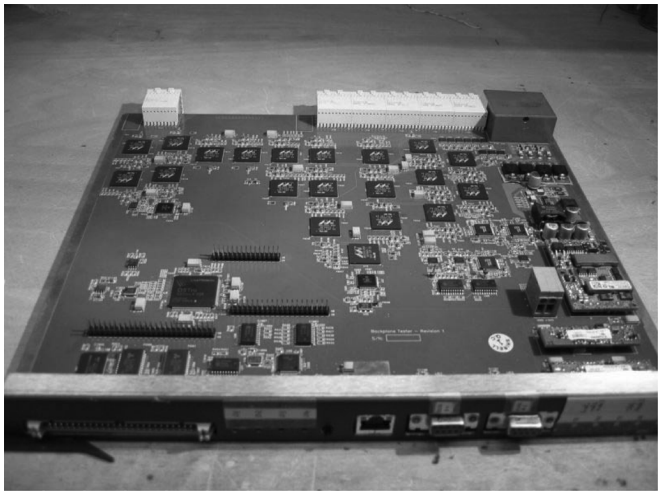

Fig. 6. Master Hub Board.

Current real-time systems often employ daughter boards to carry replicated subsystems such as DSP's or specialized I/O ports. The ATCA board is dimensioned to allow up to four of the popular PMC form factor mezzanines mounted on the left of the board and occupying about 2/3 of the board space.

The choice of full mesh over dual star topologies is essentially one of connector and channel driving costs. Mapping any of the classical real-time topologies such as the tree, ring, pipeline or multidimensional cube is possible through the dual star topology provided that the switching technology employed is non-blocking for the maximum allowable bandwidth. Choosing the full mesh option involves not only the extra backplane connector costs but in addition every node board must connect to every channel so that even if its connectivity is limited in practice it can continue to function in any slot.

Some users are concerned that even 200 W per board is a limitation given the consumption of today's most performant processors and that even the allowable component height may not be enough for the heat sinks needed to cool them. However 200 W per board is already at the limit of the air flow cooling capacity and just raising the consumption by only 50 W would mean that a typical rack with four such chassis would move from a possibly manageable 12.8 Kw to a possibly unmanageable 16.8 Kw.

VI. Practical ATCA Design

The EU funded ESTA (Ethernet Switching at Ten gigabits and Above) project [3] examined backplane technology for a 10 Gb/s Ethernet switch fabric before the ATCA standard was ratified. The resulting design was so similar to the ATCA, yet without the extensive management services, that it was deemed not worth building a proprietary backplane but preferable to exploit the standard one. The standard itself is supported by extensive simulation but it is clearly not possible to simulate the entire backplane and obtain some quantitative metric of the cumulative background noise that could be generated. In addition, even if such a study could be done it would not be able to take into account the negative effects of poor manufacturing techniques. It is possible for example for a board to be accepted following a measurement of a test coupon on one part of a board yet other parts of the same board are out of specification. We were particularly concerned that a substandard prototype board being used by prototype silicon would yield the kind of low level system error whose diagnosis and cure could easily exceed the development time and budget of the whole project.

We therefore developed a backplane tester [4] that would exercise every connection simultaneously on the backplane using the same driver and receiver technology, and at the same speed, as would be employed by the switch fabric used in the 10 Gb/s switch design.

The tester consists of one board for every slot in the chassis, two of them are Hub Boards and twelve are Node Boards. Each Hub Board drives one channel to every Node Slot and each Node Board drives one channel to each Hub Board. Each channel is driven with a Marvell Alaska SerDes [5] used in standalone test mode. The SerDes has built in circuitry to generate self test data pattern sequences to test high or low frequency jitter effects and Pseudo Random Bit Sequences (PRBS) [6]. The SerDes are controlled over a low speed MDIO (Management Data Input/Output) bus [6] which is used to select the channels of interest for any given test, initialize the pattern type required, clear the error counters and start the test. This is done for every board in the chassis and once the test has been launched the same control bus is used to poll the registers of every SerDes to monitor the error counters. The connectivity, corresponding to the primary star implemented on the ATCA backplane, is shown

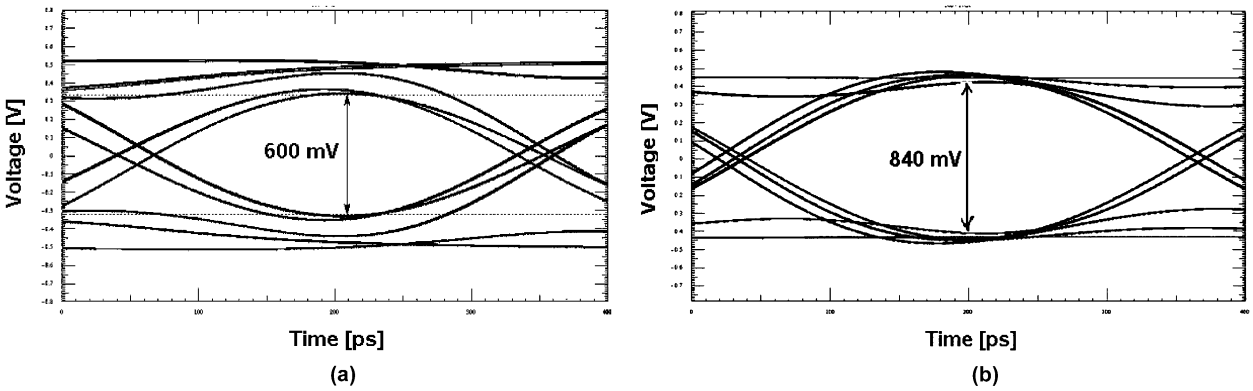

Fig. 7. (a) 0% Pre-emphasis Eye Opening 600 mV; (b) 33% Pre-Emphasis Eye Opening 840 mV.

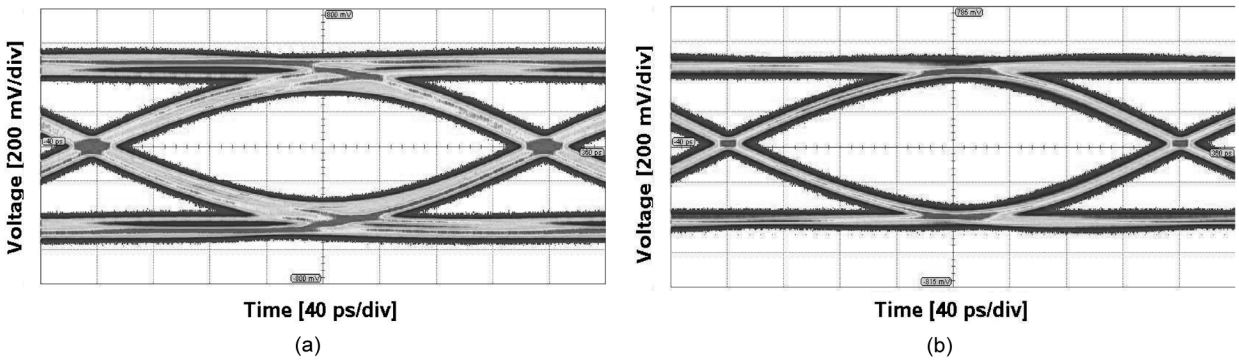

Fig. 8. (a) 0% Pre-emphasis Eye Opening 575 mV; (b) 33% Pre-Emphasis Eye Opening 740 mV.

graphically in Fig. 5. A similar redundant star (not shown in Fig. 5.) has the start point in the secondary Hub board 2.

The whole system is controlled from a simple controller microprocessor [7] mounted on one of the hub boards. It communicates with a remote PC using TCP/IP. A simple spreadsheet is used on the PC to define which channels on which boards will be participating in any given test as well as the test patterns of choice. This spreadsheet is parsed by a small application program that converts the spreadsheet into a sequence of MDIO commands that are sent to the microcontroller, which executes each in turn by emulating the MDIO bus through its parallel I/O port.

Fig. 6 shows the fully populated master hub board.

The SerDes have programmable levels of voltage swing and pre-emphasis [8] to compensate for losses in the transmission medium. We were able to exercise these to measure not just the losses, but also, by exaggerating the settings beyond the normal operating point, how much margin was available without incurring errors.

We performed a full simulation of the results expected from using pre-emphasis in HSPICE, employing the Marvell models for the transmitter/receiver and chip package. This was built into a circuit that also contained models for the connectors and a W-element model for the differential stripline trace on the line card and on the backplane. The output of the simulation was then post processed to impose a 5 Ghz cutoff frequency so that the result could be meaningfully compared to the eye diagrams that were measured with a 6 Ghz differential probe [9] feeding a 5 Ghz Serial Data Analyser [10].

Fig. 7 shows the results of the processed simulations for 900 mV with respectively 0% and 33% pre-emphasis.

These simulations can fairly be compared with the measured results which are given in Fig. 8. These show slightly reduced amplitudes but well within the expected range of performance.

Even with no loss compensation at all there is a substantially better eye opening that the required 400 mV and this is improved even further by the use of a moderate amount of pre-emphasis.

The Alaska device is actually programmed by defining the level of 'de-emphasis' which means that the swing of the post transition bit is set to a known value, and the nominal level of subsequent bits defined according to the chosen level of pre-emphasis. In the limit, we can require up to 300% of de-emphasis, which for a first bit swing of 1100 mV yields a nominal bit swing of only 366 mV, as shown in Fig. 9.

The eye diagram at the receiver corresponding to the signal in Fig. 9 is shown clearly in Fig. 10, where the eye opening has been thus artificially reduced to only 220 mV together with an obvious signal overshoot. Thus, the de-emphasis effect reduces the eye-opening to much less than the nominal 400 mV required by the standard.

Only under these extreme conditions was it possible to force the system to start generating errors which clearly demonstrates that not only is the backplane technology well defined and manufactured but that there is clearly the possibility of achieving much higher bit rates as are described in the ATCA roadmap.

VII. Summary

The ATCA standard has been briefly presented as has its possible application in real time systems. A systematic test of the performance of the backplane shows that it meets the specifications with the potential to achieve much more.

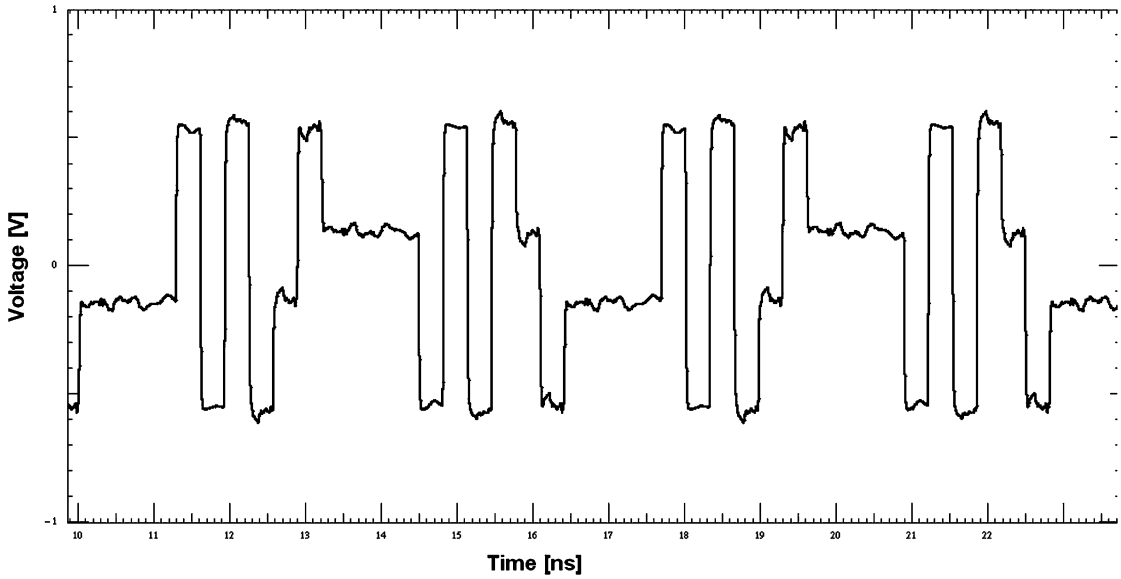

Fig. 9. Transmitted Signal with excessive de-emphasis.

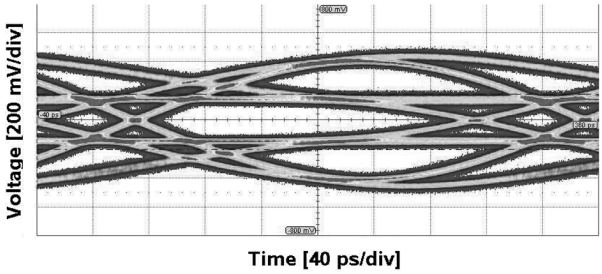

Fig. 10. Eye Diagram with excessive pre-emphasis.

ACKNOWLEDGMENT

The authors would like to thank Dr. A. Lestra for his major contribution in bringing the design through the production cycle to a successful prototype implementation, as well as his useful discussions and technical assistance during the development and testing of the system.

REFERENCES

[1] [Online]. Available: http://www.picmg.org/specdirectory.stm.
[2] [Online]. Available: http://www.picmg.org/specdirectory.stm#_PICMG_3.0.
[3] [Online]. Available: http://www.ist-esta.org/index.cfm?PID=56.
[4] A. Oltean and B. Martin, "AdvancedTCA backplane tester," in *Postgraduate Symp. PGNET*, Jun. 27–28, 2005, Liverpool John Moores Univ.
[5] 88X2040 Datasheet-integrated single chip quad 3.125/3.1875 Gbps Transceiver Sep. 15, 2003, Rev. E.
[6] IEEE Draft P802.3ae /D5.0 May 1, 2002, Annex 48A.
[7] [Online]. Available: http://www.lantronix.com/pdf/DSTni-LX_PB.pdf.
[8] Johnny Zhang and Zhi Wong, "White paper on transmit pre-emphasis and receive equalization," *Mindspeed Technol.*, Oct. 31, 2002.
[9] [Online]. Available: http://www.lecroy.com/tm/products/Probes/Differential/WaveLink/default.asp.
[10] [Online]. Available: http://www.lecroy.com/tm/products/Analyzers/home.asp?mseries=10.