An Overview of High Frequency Processor-System Interconnects

This is the first installment in a two part article about the IBM Elastic I/O interconnect for the PPC970, presented at MPF 2002. This first piece examines the physical specifications for the Elastic I/O interconnect, while the second will deal with the protocol that drives the interconnect.

Overview

In the recent Microprocessor Forum, IBM released preliminary details about its upcoming PowerPC processor, the PowerPC 970. The PowerPC 970 processor is targetted at the high performance desktop, workstation and low-end server markets, and inherits much of its technical heritage from the server oriented POWER4 processor. One of the most interesting aspects of the PowerPC 970 processor is the high frequency and high bandwidth processor interconnect planned for this processor by IBM. In Figure 1, we show a diagram of the PowerPC 970 processor connected to its companion chip via the high frequency Elastic I/O interconnect.

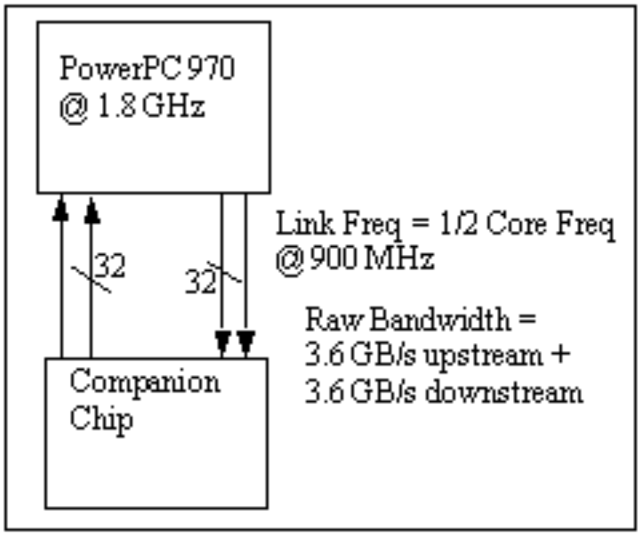

Figure 1: Elastic I/O interconnect of the PowerPC 970 processor

Figure 1 shows the basic details of the Elastic I/O on the PowerPC 970 processor as announced by IBM. The Elastic I/O operates at an integer fraction of the CPU core frequency. In this case, the Elastic I/O operates at 1/2 of the frequency of the CPU core, and for the 1.8 GHz version of the PowerPC 970 CPU, the Elastic I/O would operate at a frequency of 900 MHz. The Elastic I/O consists of two unidirectional point-to-point interconnects that is 4-byte wide in each direction. Simple arithmetic reveals that the Elastic I/O could support a raw bandwidth of 3.6 GB per second in each direction. However, unlike more traditional processor busses with separate address, command, control and data channels, the address, command and data bits for each and every transaction request are all directed onto the same set of wires. As a result, the cycles devoted to the transmission of the address and command requests must be subtracted out in a computation of the available data bandwidth of the Elastic I/O interface. In the PowerPC 970 presentation, IBM claims a peak data bandwidth of 6.4 GB per second from the raw bandwidth of 7.2 GB per second. The Elastic I/O interface of the PowerPC 970 processor also includes some separate control signals that allow the companion chip and CPU to communicate cache snoop acknowledgment and response without occupying the full width of the 32 bit wide Elastic I/O channels.

In this article, we will attempt to cover some of the basic concepts that enable the 900 MHz operation of the Elastic I/O connection interface. We will also make references to other more classical processor interconnects such as Intel processor busses and the Alpha EV6 processor interconnect found on Alpha EV6 and AMD Athlon processors.

Bus Basics

What is a Bus?

In Figure 2, we show four basic combinations of uni-directional, bi-directional, point-to-point and multi-drop bus interconnects.

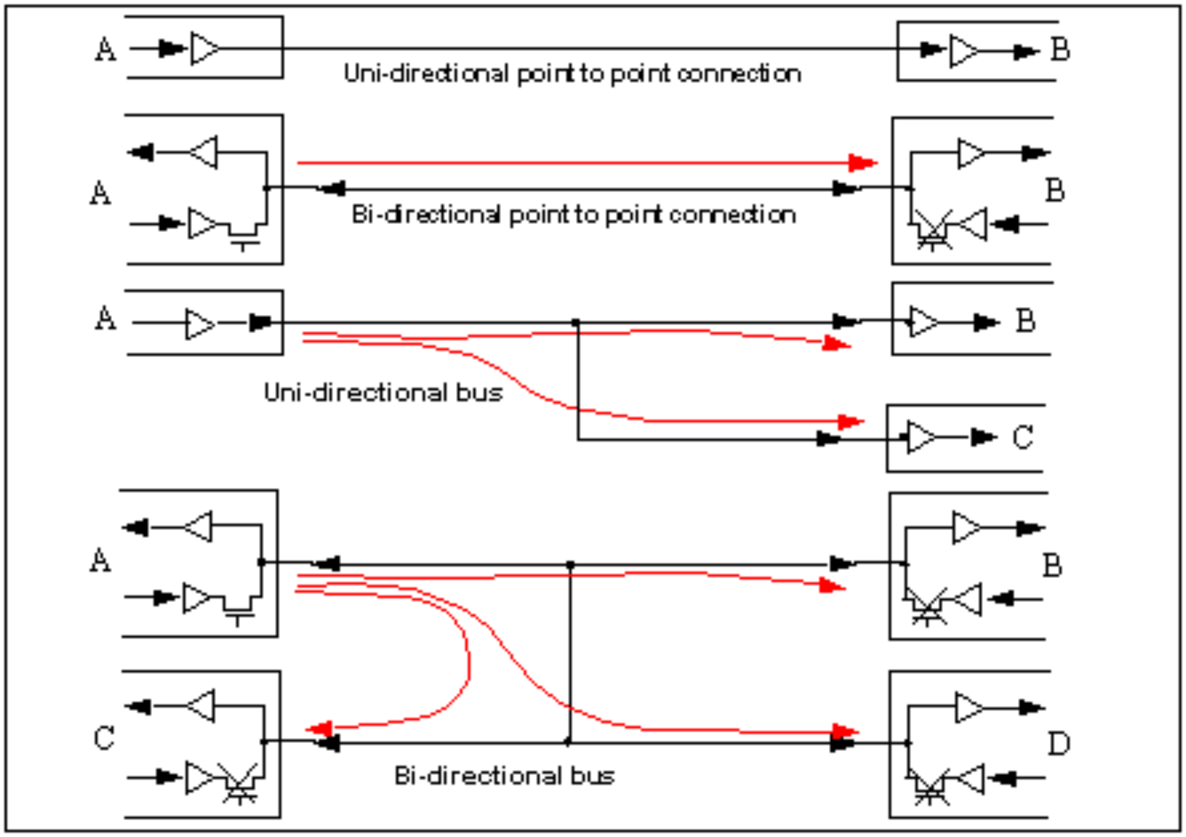
A
B
Uni-directional point to point connection
A
B
Bi-directional point to point connection
A
B
Uni-directional bus
C
A
B
C
D
Bi-directional bus

Figure 2: Uni-Directional and Bi-Directional Point to Point Connections and Multi-Drop Busses

The Elastic I/O used by IBM in the PowerPC 970 processor is a uni-directional point to point interconnect, and signals only travel in one direction, with drivers on one end of the interconnects, and receivers on the other end of the interconnects. The data channel of the Alpha EV6 processor bus interconnect uses a bi-directional point to point interconnect, and a signal may be driven from point A to point B or from point B to point A depending on the dynamic requirement of the data flow between the processor and the support chipset. Finally, Intel processors such as the venerable Pentium !!!, Pentium 4, Xeon, and Itanium series of processors use the more traditional multi-drop bus. In this configuration, each processor can potentially drive signals onto the bus, and when one agent on the bus drives the bus, all of the other agents on the bus can observe the signals as asserted by the driving agent.

In general, multi-drop busses are more difficult to push to higher clock frequencies, but they are capable of supporting small-scale SMP configurations at a lower cost of pin count. Note: While topology has a contributing effect on the operating frequency of the interconnection scheme, it is not the sole limiting factor. Signaling technology and protocol are also important factors that determine the limits of operating frequency.

Bus Issues

It Takes Time for Signals to Propagate from Point A to Point B

In Figure 3, we show that the source at point A drives a signal onto the interconnect, and after some time, the signal reaches the destination at point B.

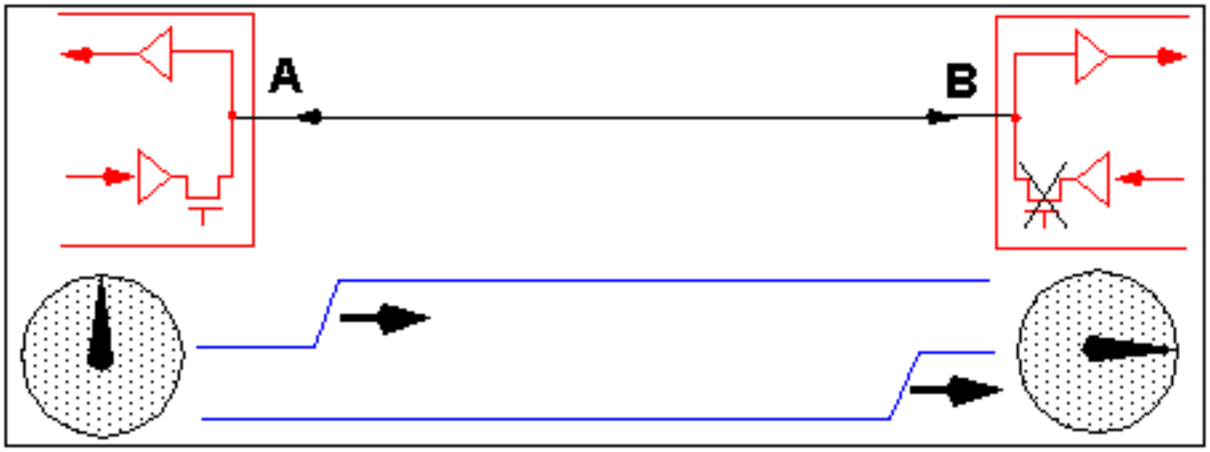

Figure 3: Signal propagation from Point A to Point B

An Idealistic figure of signal propagation speed may be obtained by using an approximation of 0.66c, or roughly two-thirds of the speed of light. With this approximation, we find that it would take 1 ns for the signal to travel a distance of 20 cm. However, traces on modern 4 layer PCB boards and through module interconnects are far away

Propagation delay on modern computer system boards manufactured on multi-layer PCB boards are strong functions of the from ideal lossless transmission lines. Effects of non-ideal transmission lines, especially mismatched load and line impedance characteristics often combine to reduce actual signal propagation speed.

dielectric material used on the board. The propagation delay may be approximated by computing the wave velocity of a strip line transmission line. For more detailed discussions, please consult the relevant text. Quick discussions on this topic may be found online from Rambus Inc. and lecture notes from Stanford Computer Systems Laboratory. These discussions are cited in the reference section.

Signal Waves Bounce Back and Forth on Transmission Lines with Mismatched Load Impedance

In Figure 4, we illustrate that a transmission line will have a given characteristic impedance, and on an ideal transmission line, a signal wave can travel from one end of the transmission line to the other end very rapidly.

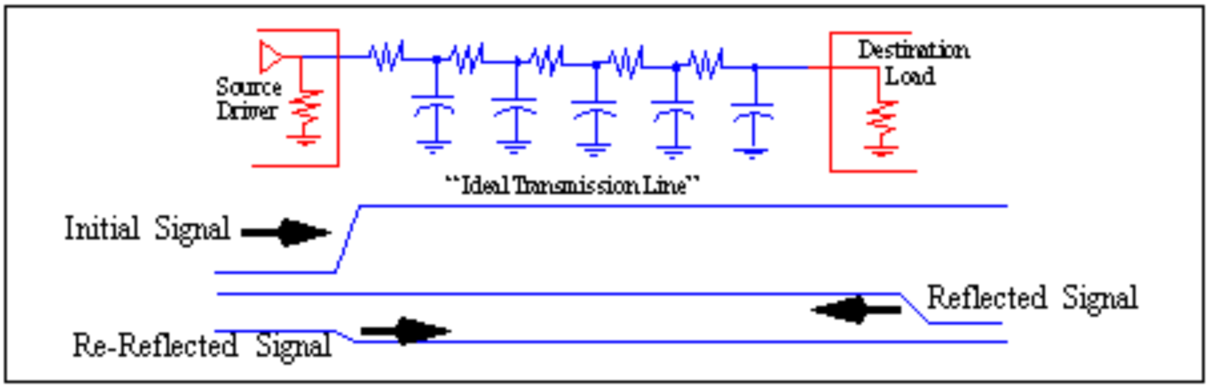

Figure 4: Wave Reflections with Mismatched Load Impedances

However, when the wavefront reaches the end of the transmission line, if the input impedance of the load does not perfectly match the characteristic impedance of the transmission line, then some portion of the input signal wave will reflect back onto the transmission line. The phase and magnitude of the reflected wave are functions of the mismatches between the characteristic impedance of the transmission line and the impedance of the load. In figure 4, we show a reflected wavefront that travels from the destination back to the source with a reduced magnitude. When the reflected wavefront reaches the source driver of the initial signal, if the impedance of the source driver also does not match the characteristic impedance of the transmission line, then another reflected wave will once again reflect from the source to the destination. The voltage on the transmission line will then be a summation of the wavefronts as they reflect back and forth. Theoretically, poorly matched load and source impedances can ensure that a signal wave reflect back and forth for a long time before settling to the final signal value. For this reason, properly matched and terminated signal paths are absolute necessities for high frequency signaling. ("High Frequency" is with respect to the length of the transmission line. For a transmission line that is 100 meters in length, 10 MHz is a very high frequency) For signal paths and loads that are not nearly perfectly matched in the impedance characteristic, the worse case cycle time on the signal bus must be computed to allow the signal reflections time to settle. For a wave pipelined interconnect, where new signal wavefronts are placed by the source driver onto the interconnect even before the previous signal wave front reaches the destination, the reflections on each interface must be negligible by design.

Multi-Drop Busses Introduce Non-Ideal Discontinuities in Signal Paths

In Figure 5, we show that each load on a multi-drop bus becomes a discontinuity on the transmission line. Each discontinuity on the transmission line creates an interface where signal waves can reflect.

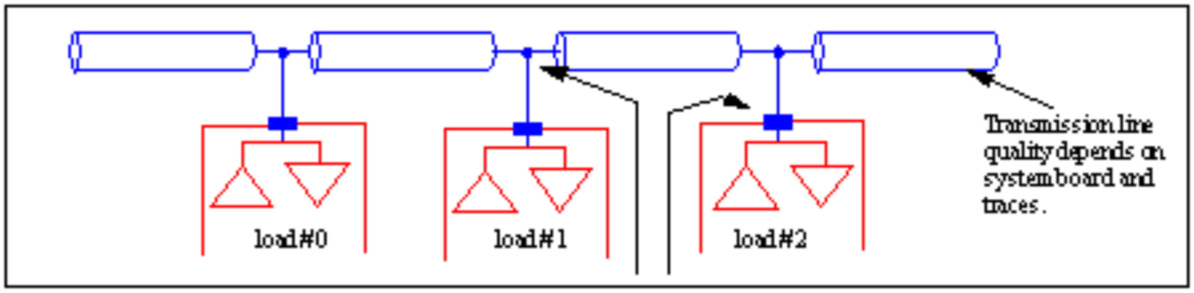
load#0
load#1
load#2
Transmission line quality depends on systemboard and traces.

Figure 5: Each Load Introduces Discontinuity on the Transmission Line

In Figure 6, we show that if we represent each load as a capacitive element, with a larger number of loads, rise time of the signal decreases, signal velocity decreases, and signal ringing continues for a longer period of time.

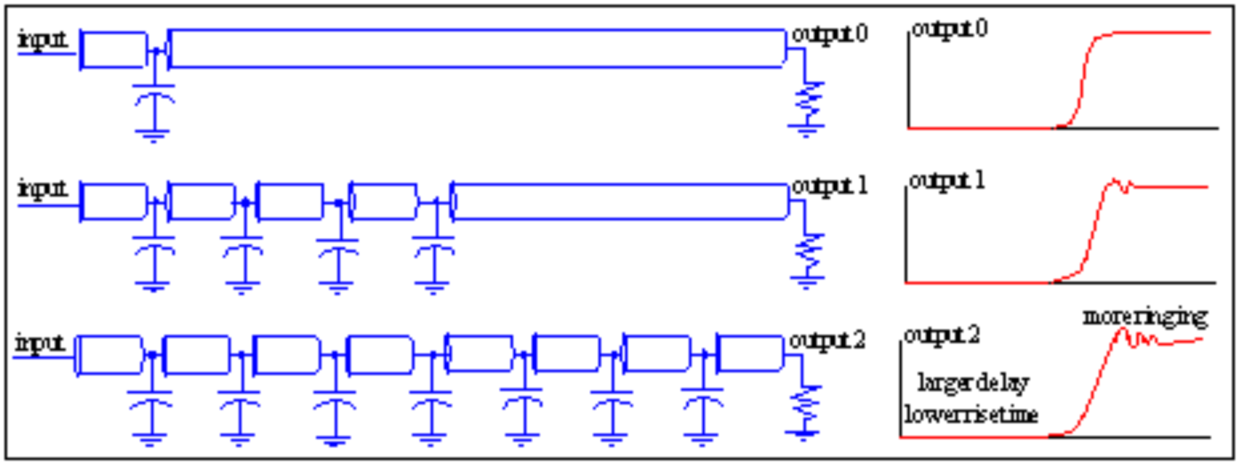

Figure 6: More Loads, Slower Signal Transmission, More Reflections

Bus Limitations

System Bus Frequency Also Depends on Signaling Technology

There are numerous types or classes of electrical signaling technology: TTL, ECL, RSL, SSTL, Rambus Yellowstone, GTL+, AGTL+ and others. There are a large number of signaling technology, some designed for high frequency, some designed for lower power consumption, while still others exist to maintain compatibility with legacy electronic devices. In Figure 7, we show a basic signaling technology where two different voltage levels are used to represent two different logic levels.

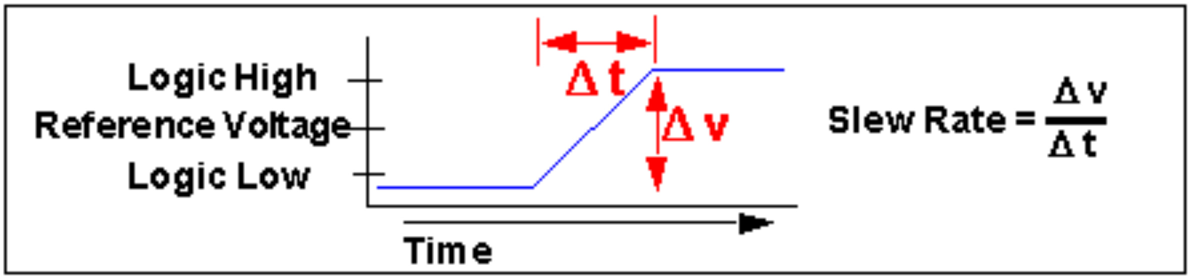

Figure 7: Signaling Technology in a Nutshell

When a single bit of information is to be transmitted from one end of the signal interconnect to the other end of the signal interconnect, the voltage level on the interconnect is made to switch state from one state to another. The voltage switches state, shown in figure 7 as a certain magnitude of voltage swing in a finite amount of time. The change in voltage divided by the time required for the state switch is known as the slew rate of the voltage swing. Since the slew rate cannot often be increased to facilitate a faster state transition, the alternative is to reduce the magnitude of the voltage swing between the voltage states. For this reason, the high frequency signaling technologies all have very small voltage swings between logic states.

There are many different signaling technologies, and the exacting details are too numerous for this brief overview. We will thereby simply assert that the speed of signal propagation and logic state switch on the system interconnect depends not just on the characteristic impedance of the system interconnect and the number of loads on that interconnect, but also the signaling technology. We will hereafter assume a "generic" binary signaling technology not unlike SSTL or RSL.

If You Send the Signals in Parallel, Across Parallel Interconnects, Will They Reach Destination at Same Time?

In figure 8, we show a two poorly matched signal interconnects on a parallel bus. In this figure, we show that it takes longer for the signals from the driver to reach load B as compared with load A. Moreover, we also show that path #2 to load A is longer than path #1 to load A. As a result, the difference in the physical trace lengths could introduce timing skew between signals sent on path #1 and path #2. In such a case, even if signals are sent from the driver at the exact same instance in time, they will not arrive at the interface of the load at exactly the same instance in time.

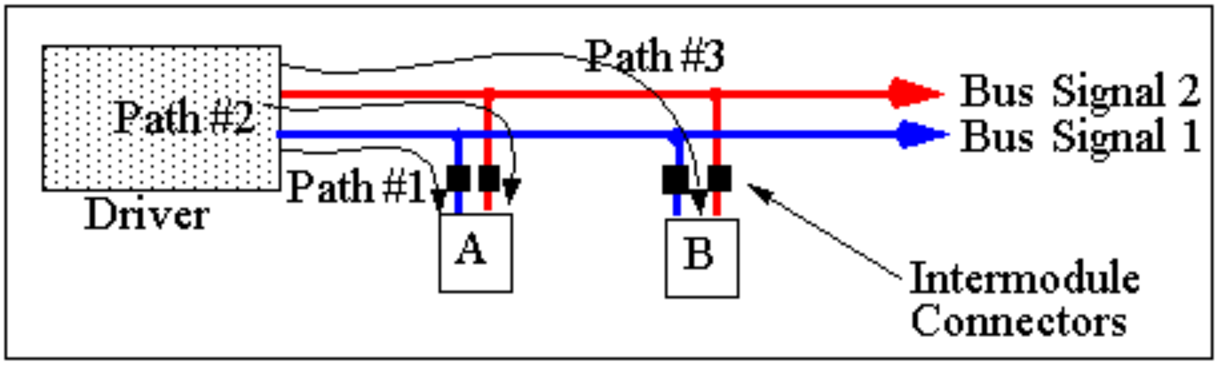

Figure 8: Differences in Trace Lengths and Impedance Characteristics Introduce Skew

Ever Wonder Why Traces have to Snake Around?

Differences in Path lengths force system designers to go to heroic efforts to match the lengths of the traces of a parallel bus. Most any modern system boards will have more than a few signal lines that snake around parts of the board to attempt to match signal path lengths.

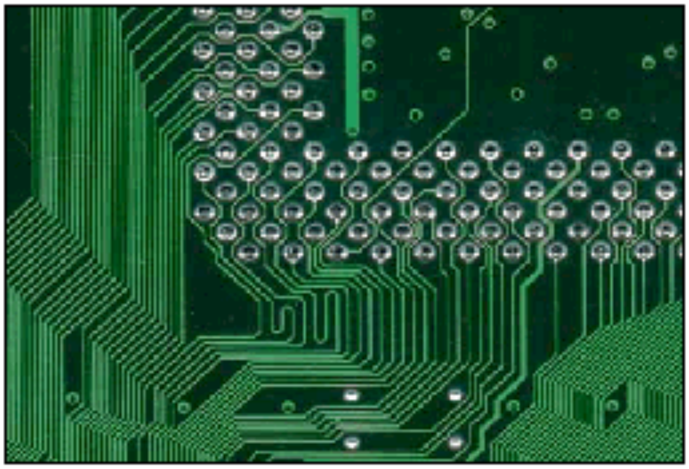

Figure 9: System Board Designers Try Hard to Path Length Match Wires in Parallel Busses

Skew Troubles

Wider Busses Tend to have Terrible Skew Characteristics

In figure 10, we extend the ideas expressed in figure 8 and figure 9, and show that signals in a parallel bus often show varying amounts of skew.

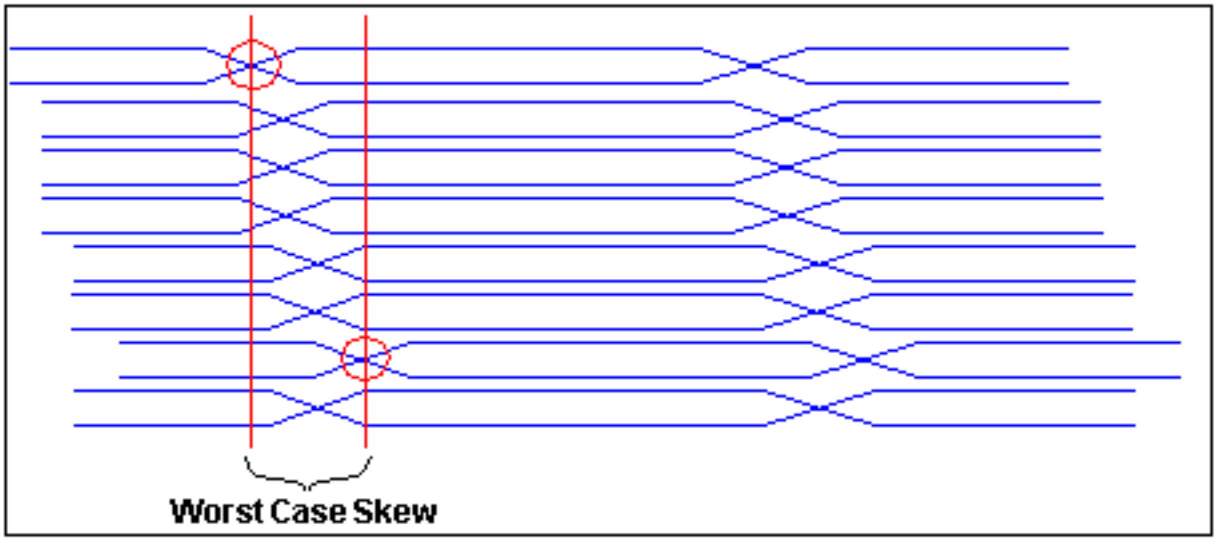

Figure 10: Wide Parallel Busses Generally Show Widely Varying Skew Characteristics

Modern parallel busses that do not have de-skewing circuitry for individual bits must account for the worse case skew differential. In a high frequency interconnect, skew differential contributes heavily to limiting the operating frequency of the parallel system interconnect. For high frequency parallel signaling across a system board, signal de-skewing is an important and necessary aspect of the design. There are two important aspects to the signal-skewing problem. One aspect of the problem is that signal skew may be introduced by static variables, such as signal path lengths, or mismatched impedance characteristics. The second aspect of the problem is that there is a dynamic component to the signal skew that varies with the environmental conditions of the source chip, the destination chip, as well as the transmission line itself. The dynamic portion of the skew equation is a function of component variances that is introduced by a change in temperature or voltage. The existence of the dynamic component of the skew equation means that for extremely high frequency designs, circuits deployed to de-skew signals on a parallel system interconnect may have to be dynamically re-calibrated during runtime to ensure that the signal skews are properly compensated for, regardless of the changes in the operating environment. (Minute differences in component characteristics between a cold boot up and a warm runtime can introduce skew differentials in a high frequency parallel system interconnect)

Elastic I/O Designed to Handle Skew

In figure 11, we show that the parallel interconnect actually has a few signal paths that were different in length.

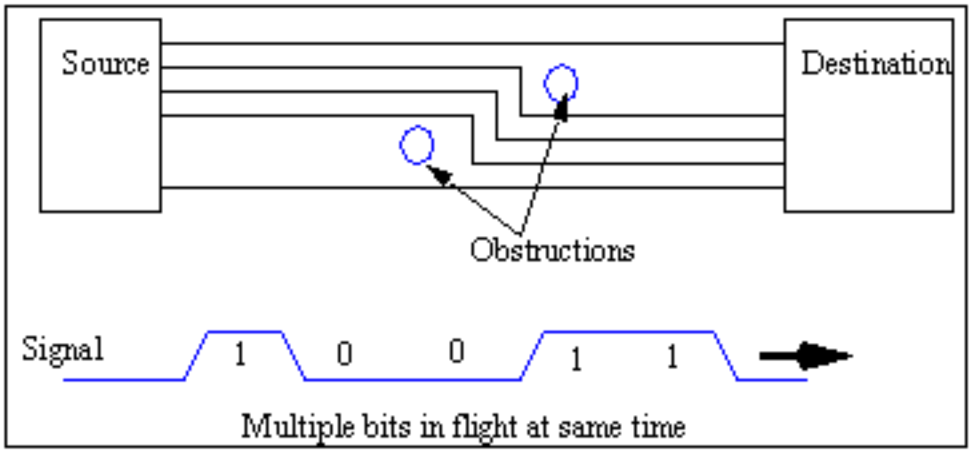

Figure 11: Elastic I/O Designed to Tolerate Skew

IBM's Elastic I/O, as presented in the 1999 Hotchips conference, has been designed to tolerate a large degree of variance between the characteristics of individual signal paths. In this case, we show some intermediate signal lines in the parallel interconnect as having longer path lengths to avoid obstacles in the system board. Furthermore, with properly matched impedances on the interconnects, a driver can begin to place additional bits onto the interconnect even before the previous bits reach the destination. As mentioned previously, this feature must rely on proper matching of the load impedance with the characteristic impedance of the transmission line. In the POWER4 processor, the Elastic I/O runs at 500 MHz, and it is used as a high frequency chip to chip interconnect between the different CPU dies. In the PowerPC 970 processor, the high frequency system interconnect is tie to the companion chip that contains the memory controller that performs the memory access functionalities for the PowerPC 970 processor.

SMP Issues

Dedicated Ports are Needed For SMP Configuration

In a shared memory multi-processor configuration, the companion chip (system controller) must explicitly broadcast cache snoop requests to each CPU, and collect snoop responses from each CPU. Unlike processors connected in a shared bus topology, where addresses for snoop requests can be effectively broadcast through the shared bus, processors connected by point to point connection fabrics must rely on the support chipset to rebroadcast snoop addresses and accept snoop results for the maintenance of cache coherency. In this manner, the point to point "processor bus" found on the PowerPC 970 processor is not unlike the Alpha EV6 "processor bus". However, unlike the connection scheme on the PowerPC 970 CPU, the Alpha EV6 bus use separate channels for data and command+address. The data bus found on the EV6 bus is bi-directional, and must be directed properly for transaction read and writes.

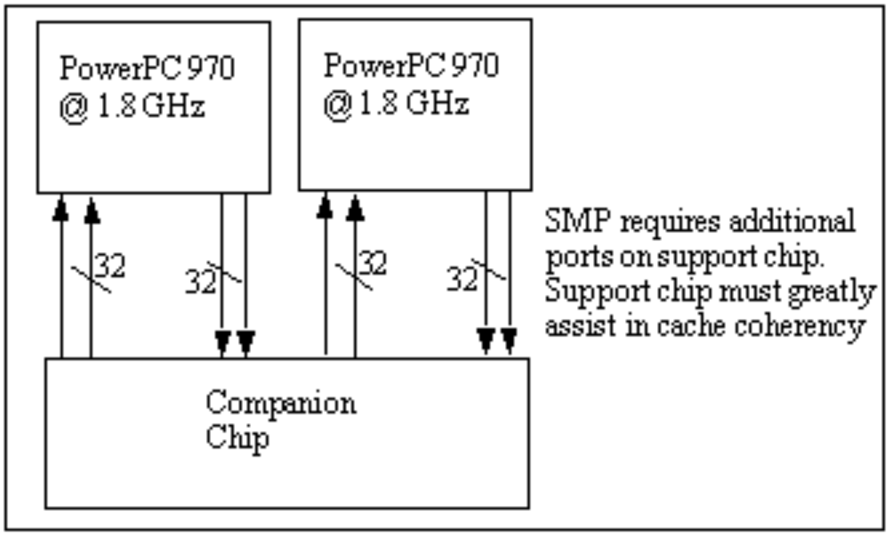
PowerPC 970
@ 1.8 GHz
PowerPC 970
@ 1.8 GHz
32
32
32
32
SMP requires additional ports on support chip. Support chip must greatly assist in cache coherency
Companion Chip

Figure 12: For Point to Point Connections, Each CPU Needs Dedicated Port in SMP Configuration

Wrapup

Summary

In this article, we briefly examined the connection scheme found on the Power4 processor. We expressed in abstraction some reasons why a unidirectional point to point connection scheme can be pushed to higher frequencies as compared to a bi-directional multi-drop bus based connection scheme, even when using comparable signaling technology. Furthermore, we also examined problems associated with signal skews on a high frequency and wide parallel interconnect, and how the PowerPC 970 processor would benefit from its POWER4 lineage in inheriting a high frequency, low point count, wave pipelined interconnect with built-in de-skewing circuitry.

References

[1] Dally, W. "Digital System Engineering Lecture notes", Stanford University.
http://cva.stanford.edu/books/dig_sys_engr/lectures/

[2] RIMM Module Propagation Delay Measurement and Optimization, Rambus Inc.
http://www.rambus.com/downloads/tpd_appnote_r.01.pdf

In this article, we examined some differences between a point-to-point connection scheme versus a multi-drop bus based connection scheme. One fact that may be of interest to the reader is the fact that the discussion contained herein is not specifically limited to processor and chipset interconnects, but are also applicable to connection schemes seen in modern memory systems. Traditional memory systems such as DDR SDRAM that allow end users to upgrade the size of the memory system have the disadvantage that each memory module interface is a discontinuity on the transmission line. The discontinuities increase signal switching times and limits system operating frequencies. Ironically, non-expandable memory

[3] Ferraiolo, F., Cordero E., Dreps D., Floyd M., Gower K., McCredie B., “POWER4: Synchronous Wave-Pipelined Interface”. Presented at Hotchips 1999.

[4] Poulton, J., “Signaling in High Performance Memory Systems”, ISSCC 1999.
http://www.cs.unc.edu/~jp/signaling_tutorial.ps

systems such as those found on high end graphics cards, especially those that have relatively smaller requirement for memory capacity, can operate a single rank of memory at a higher frequency, and could provide higher bandwidth as a result. The reason for this apparent paradox is that for these embedded memory systems, the signal traces are often very short and since there is only one single rank of memory chips, the connection closely resembles a point-to-point connection. As a result, the operating frequency of a tightly coupled and small memory system can be increased to a higher degree as compared to the operating frequency of a larger, user upgradeable memory system. In the not-so-distance future, we may see the onset of on-CPU memory controllers that is limited to the control of a single rank of non-expandable memory directly soldered to the CPU. A secondary memory controller may also exist on-CPU or elsewhere, but relegated to controlling a larger, slower, but user upgradeable memory system.