
THE IA-64 ITANIUM PROCESSOR CARTRIDGE

THE DRIVING FORCE BEHIND THE ITANIUM PROCESSOR CARTRIDGE IS ELECTRICAL BUS PERFORMANCE FOR THE EXTERNAL SYSTEM BUS AND THE INTERNAL L3 BACKSIDE BUS.

..... The Itanium processor cartridge is a packaging optimization for electrical and thermal performance in a server environment. The 3-in. × 5-in. cartridge contains the Itanium CPU, up to 4 megabytes of level-3 (L3) cache, an innovative power delivery scheme, and an integrated vapor chamber thermal spreading lid for removing power. Cartridges and a chip set can be ganged electrically by means of a glueless bidirectional, multidrop system bus. Power is delivered through a custom connection with separate voltages for the 0.18-micron CPU and 0.25-micron custom cache devices.

An I²C serial connection provides access to system management features such as temperature monitoring and cartridge identification information.

Electrical performance

In a departure from other Intel cartridge products, I/O signals in our cartridge are arranged into a pin grid array (PGA) package. This lets the cartridge lay flat on the motherboard. By exploiting double-sided motherboard mounting, the external system bus can have minimal overall length. Using this scheme, a four-way, five-load multiprocessing implementation can achieve 266 megatransfers per second (MT/s) of system bus performance.

The processor's L3 cache is external to the CPU, but the cache devices and the cache/CPU backside bus (BSB) are completely contained and isolated inside the cartridge. The 128-bit BSB cache data bus operates at full CPU core speed to minimize first-access latency and to achieve continuous throughput of approximately 13 gigabytes-per-second (Gbytes/s).

Physical considerations

More than any other factor, the server system electrical performance (system bus, BSB) influenced the physical arrangement of the cartridge and the internal die placement. As seen in Figure 1, the intended system topology promotes two cartridges placed next to each other and duplicated on both sides of the motherboard to enable a minimal length system bus. As shown in Figure 2, the L3 cache silicon is clumped to one end, making the on-cartridge BSB electrically short and minimizing the cartridge width dimension to approximately three inches, which further contributes to short system bus electrical connections between cartridges.

One of the most critical physical dimensions is the electrical stub length of the system bus. This is the distance from the system bus interconnect on the motherboard to the I/O buffer in the CPU. The cartridge topology

William A. Samaras
Naveen Cherukuri
Srinivas Venkataraman
Intel

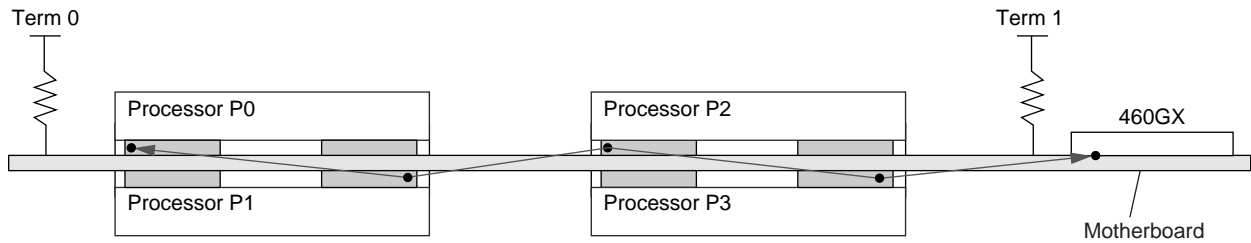


Figure 1. Double-sided system bus topology.

minimizes system bus stub length by positioning the CPU die directly over the pin field array and carefully positioning the on-die CPU I/O buffers to align with external signal pins. Short system bus stubs contribute directly to system bus performance.

Packaging

The processor cartridge contains three substrates: base, CPU package, and cache multichip module (MCM). The base substrate is conventional FR4-type material consisting of 12 alternating layers of impedance-controlled interconnects and reference planes. The CPU is attached with Controlled Collapsed Chip Connection (C4) to a 42.5²-mm, 10-layer organic land grid array (OLGA) package. This CPU OLGA is soldered to the base substrate as a ball grid array (BGA). The custom cache devices are packaged as two- or four-chip MCMs, depending on whether a 2- or 4-Mbyte cache is assembled (each cache size option has a dedicated MCM).

The cache devices are also attached using C4 to a 10-layer OLGA. Like the CPU package, the cache MCM is soldered to the base substrate as a large BGA. The cache MCM configuration contributes to minimizing the BSB interconnect length to the CPU (Figure 3). All of these substrates contain a significant

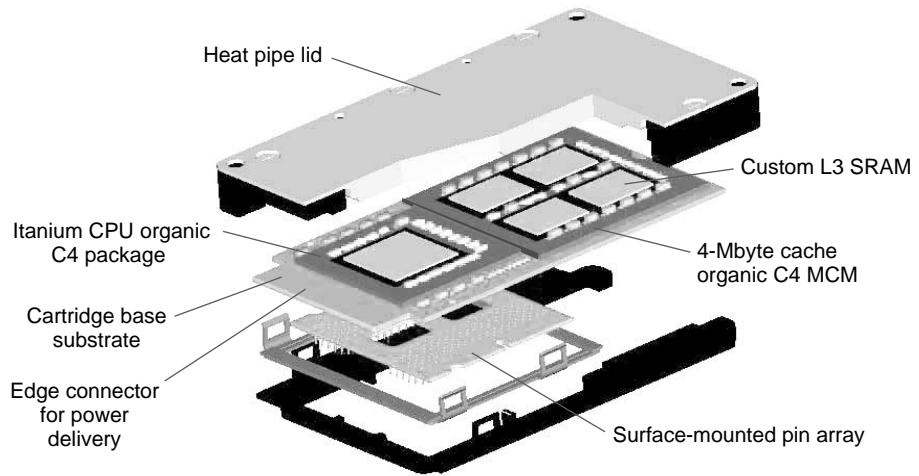


Figure 2. Itanium processor cartridge.

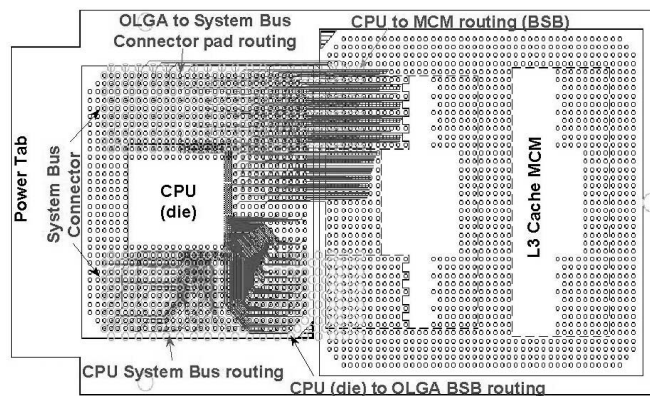


Figure 3. Cartridge interconnect.

number of capacitors for voltage decoupling purposes.

A separate PGA carrier is soldered opposite the CPU side of the base substrate as a large BGA device and serves as the external connection to the system bus (Figure 2).

ITANIUM CARTRIDGE

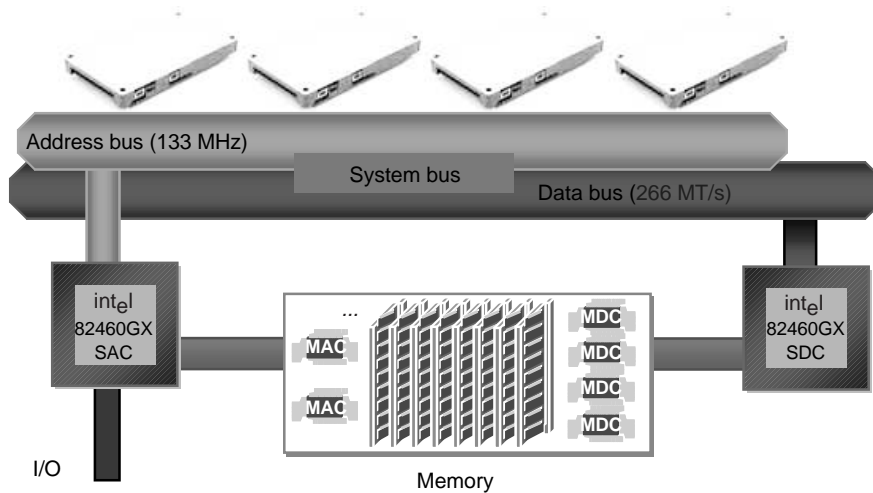


Figure 4. System bus topology.

ed into the cartridge's metal lid. The vapor chamber works like a heat pipe (we call it the heat pipe lid) by alternately condensing and evaporating a sealed liquid. This phase change process results in a temperature differential of just a few degrees Celsius across the entire top surface of the cartridge lid. The heat pipe lid effectively magnifies the CPU and cache die area to almost 13 square inches. A conventional passive, removable heat sink can be mounted on the cartridge for use in an air-cooled environment.

Power delivery

Processor cartridge power is delivered through a custom edge connection on the base substrate located at the CPU end of the cartridge (Figure 2). The base substrate edge connection plugs into a custom DC-DC power converter called the Power Pod. Due to the planar topology of this power connection, the power planes of the cartridge base substrate essentially extend into the power converter. With careful design, the entire power delivery loop inductance is reduced to just a few hundred picohenrys, enabling us to move some decoupling capacitance off the cartridge and into the power converter. Although the system bus PGA contains many signal reference (ground) connections, the larger power supply current flows through the low-inductance power connector. Power and signal currents are separated.

Since the CPU and cache silicon are manufactured with different silicon fabrication processes, each require different core voltages. The Power Pod contains two power converters—one for each core voltage. Both voltages are routed into the cartridge through the single edge connector.

Thermal solution

Removing thermal energy from the CPU and caches while maintaining reliable die temperatures proved to be a formidable design challenge. The thermal solution uses an integral vapor chamber heat spreader incorporat-

System bus overview

The processor cartridge is a self-contained processing element (the CPU) with a glueless integrated system bus interface. As an example, combining up to four Itanium processor cartridges with Intel's 460GX chip set¹ forms a robust high-performance multiprocessing system. With custom chip set designs,² other system configurations are possible.

The processor's system bus connects processing elements to each other and to the chip set. Multiprocessing snoops, memory, and I/O traffic flow through this bus, but the chip set isolates the memory and I/O subsystems from the system bus, as shown in Figure 4.

The bidirectional, multidrop system bus provides 44 bits (plus 2 parity bits) for addresses and 64 bits (plus 8 bits of error-correcting code, or ECC) for data. As address references are less frequent than data transfers, the address bus operates at the system bus frequency (133 MHz) in a conventional "common clock" mode of one transfer per clock cycle.

The data bus can operate in a double-pumped (source synchronous^{3,4}) transfer mode. That is, data transfers occur twice for every bus clock cycle (see the example in Figure 5). This means that the system bus can transfer a 64-byte cache line in a single transaction with eight consecutive data bus transfers (8 bytes/transfer \times 8 transfers) in just four system clock cycles. The 133-MHz system bus permits data transfers of up to 266 MT/s or 2.1 Gbytes/s. The

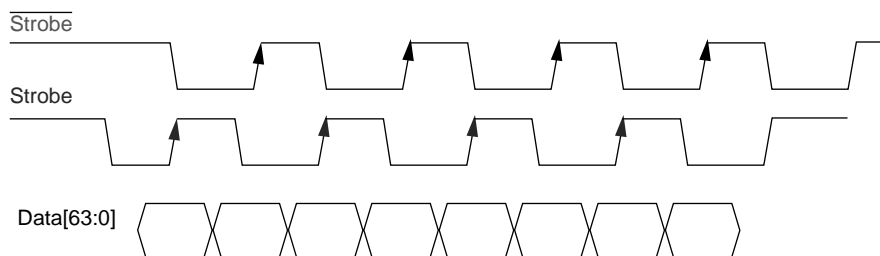


Figure 5. Source synchronous timing.

system bus supports continuous full-rate data transfers without wait cycles.

To obtain maximum benefit from the source synchronous scheme, the data bus is divided into four groups of 20 signals (16 data bits, 2 ECC bits, and a differential strobe pair per group). Each data group and its differential strobe pair are matched for interconnect length with tight tolerances, are routed on the same signal layer, and have the same number of vias to minimize skews between data and strobes. Grouping makes electrical-length matching of the data bus manageable and minimizes on-die delay variations.

Data transfers are clocked by two strobe edges per bus clock cycle (double pumped) at a receiver, as Figure 5 shows. In this source synchronous transfer mode, the active bus agent drives both data bits and the strobes, with strobe transitions centered with respect to data timing.

Short signal lengths inside the cartridge maximize source synchronous performance, and the double-sided topology shown earlier in Figure 1 minimizes the delay impact on address and control signals in the conventional clocking mode. Propagation delay limits the timing of common clock latched signals, and the timing of source synchronous data signals is much less dependent on signal length.

From the cartridge's careful package design, we got past a major performance limiter of package return path inductance, which contributes to simultaneous switching output (SSO) noise. Switching noise generated on the I/O reference (ground) pads during data transmission affects timing and signal quality of subsequently driven strobes. Inductive signal return current loops are minimized by proper placement of return vias for image currents propagating through reference planes inside the multilayer packages.

A simple but very effective external termination scheme at the extreme ends of the bus uses resistors mounted on the motherboard. The 460GX chip set¹ architecture splits the address and data paths into separate packages, adding just a single electrical load for both data and address.

GTL+ bus signaling

All system bus signals use Intel's GTL+ signaling scheme, an enhanced version of standard GTL.⁵ The GTL+ voltage swing of 0.5 V to 1.5 V is a larger swing than standard GTL and contributes to extra noise margin. As another departure, GTL+ drivers contain active pull-up devices to improve rise and fall time symmetry. To control signal slew rate and impedance across variations in process, voltage, and temperature, we compensated the processor's output drivers. Compensated output drivers minimize bus reflections and control transmission variations in source synchronous signal groups.

L3 cache and the BSB

Physically, the 4-Mbyte cache size is implemented using four discrete, 1-Mbyte custom SRAM (CSRAM) devices; each with a built-in tag array (see Figure 6, next page). The cache devices are organized in two address banks, each containing 2 Mbytes of data. The CSRAMs in each bank are arranged as two 64-bit slices for a total data width of 128 bits. The processor's L3 cache line size is 64 bytes and is accessed as a four-cycle burst from the CSRAM array. This organization reduces the number of electrical loads on the data bus to allow continuous cache line data bursts at the CPU core frequency.

The 128-bit data bus is divided into eight 16-bit, source synchronous, length-matched data groups. Each data group includes 2 bits for ECC and a pair of differential strobes for latching data at the receivers. The lower data

ITANIUM CARTRIDGE

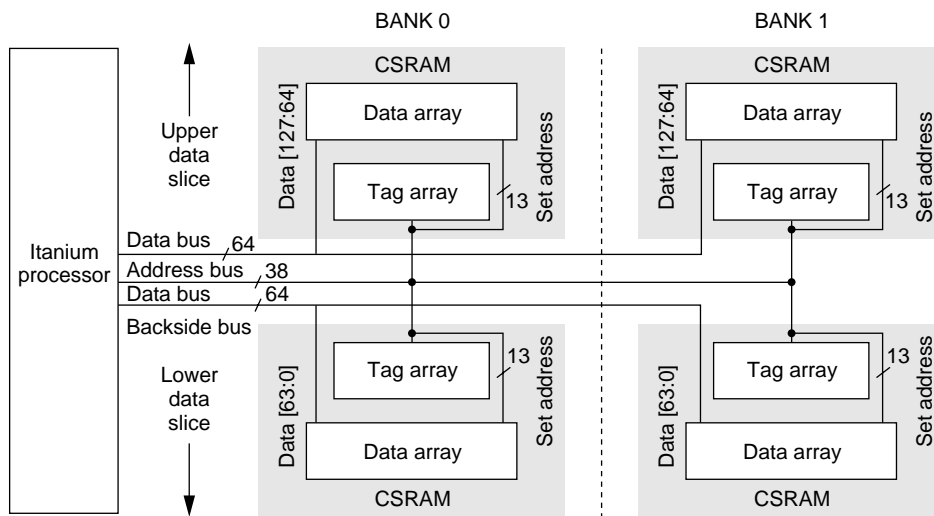


Figure 6. Cartridge L3 cache organization.

Table 1. BSB latency components.

External L3 latency component	CPU clock cycles
Cache access time (first access)	5
Protocol requirement	1
Round-trip signal flight time (bus + die)	~3
Misc. (clock skew, jitter, timing margin, etc.)	~1
Total BSB latency	10

groups (data[63:0]) are connected to the lower 64-bit slices of the cache memory banks; the upper data groups (data[127:64]) are connected to the upper 64-bit slices of the cache memory banks. The CSRAM bit-slice architecture lets us implement the data bus using a topology of three electrical loads.

The BSB address bus is 38-bits wide and is protected by 3 bits of parity. Since the data bus accesses a cache line in four successive 128-bit data bursts, the cache line address is specified just once prior to the bursts and is held for four clock cycles. Therefore, the address bus operates at one fourth the processor's core frequency. Since the slower address can tolerate additional electrical loading, this greatly simplifies the signal integrity requirements on the address bus.

The BSB returns data from the CSRAMs to the processor with a 10-cycle latency. Latency in this context corresponds to the number of clock cycles from the time the processor core issues the cache address to the time the proces-

sor core receives the first data burst of a 64-byte cache line.

The components contributing to latency include CSRAM line access time, protocol timing requirements, PLL jitter, clock skews between processor and cache, signal flight time, and on-die signal delays. Table 1 summarizes the contributions of these components of BSB latency. Since a complete latency loop includes both synchronous and asynchronous paths, the absolute contributions of these components are approximately converted to processor clock cycles at 800 MHz. The entire L3 latency loop is synchron-

ized at the processor boundary after 10 clock cycles.

BSB electrical design

The processor's cache bus is a formidable electrical design as the data transfers occur at the full CPU core speed (up to 800 MHz) with a wide, bidirectional, 128-bit data path. The bus crosses several packaging boundaries, with multiple electrical loads and multiple clocking domains. All BSB signals use an on-die parallel termination scheme to minimize reflections and reduce inter symbol interference (ISI). Bus signal quality becomes less susceptible to signal-trace impedance variations across the cartridge substrates. In addition, both the BSB I/O buffer output impedance and the termination resistance are programmable and self-compensating to minimize environmental and process variations.

BSB performance

Packaging the CSRAMs as an MCM improves packaging density and directly contributes to BSB performance. For example, the short bus lengths obtained from closely packaging the cache devices reduces cache latency. Also, OLGA substrate technology with C4 die-attach capability for both the CPU and CSRAM enables dense interconnect routing with excellent signal integrity (low cross talk) and provides exceptional low inductance power delivery.

Electrical performance is a direct function of our ability to match circuits and interconnect delays. A single-source synchronous signal group, for example, can perform at exceptionally high data rates if the data and strobe delays are exactly matched. Layering and matching the lengths of all signals within a source synchronous group eliminate the effects of manufacturing variations within a substrate. For instance, each signal sees the same discontinuity—the trace lengths before and after a T-junction, are identical for all signals. Figure 3's cartridge layout plot shows an example of direct BSB interconnections.

Cartridge electrical modeling

System electrical simulations guaranteed reliable data transfers with excellent signal integrity. All aspects of a multiprocessing hardware system were analyzed concurrently: cartridges, the chip set, the motherboard, and associated packaging. This effort spurred hardware reference designs and elaborate simulation models that accurately predicted system bus performance. We used a similar methodology for the BSB.

Predicting system bus and BSB electrical behavior required the extensive use of electromagnetic modeling and circuit simulation methods. We applied a global system methodology toward bus development by simulating an entire four-way multiprocessing system.

Central to this methodology is complete dependence on full circuit simulators. That is, Spice-type circuit analysis tools can combine nonlinear transistor models (I/O buffers) with complex interconnect structures.

The package environment includes structures such as nonideal power and ground planes (meshes in reference planes is an example), decoupling capacitors, routing via discontinuities, packaging interfaces such as C4, and BGAs. All affect wave propagation. The electrical parasitics associated with these structures are extracted using 3D field solvers that have the ability to translate complex physical geome-

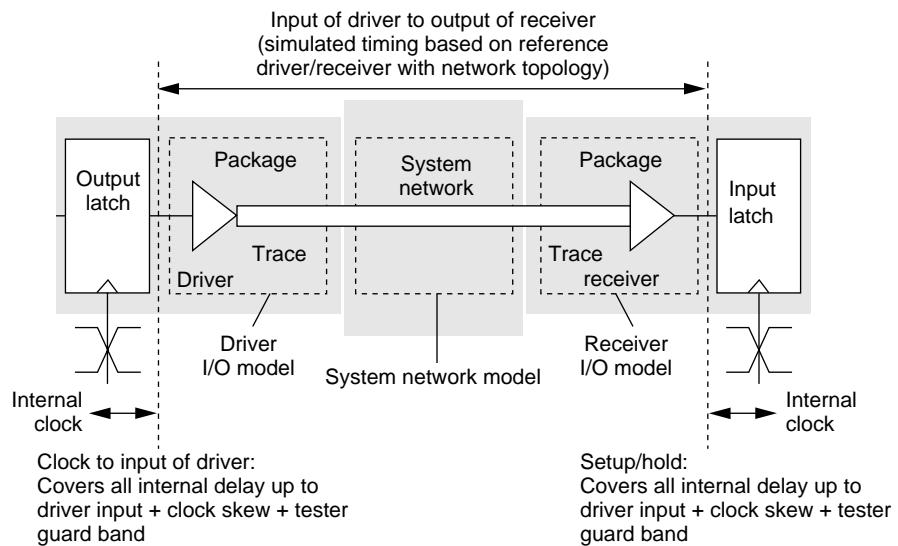


Figure 7. Bus timing partition.

tries into equivalent electrical circuit models. The partial element equivalent circuit (PEEC^{6,7}) approach, combined with 3D extraction, is the building block to model coupled interconnects and discontinuities. The PEEC method allows us to explicitly separate power and ground return paths from the signal path in the simulation model. This deviates from a modeling approach in which power and ground parasitics are lumped into the signal path and can mask the effects of nonideal reference planes. These nonideal plane models provide necessary information about the spatial variation of the power delivery system as seen by the processors, chip set, and cache devices. Because actual measurements correlated well with simulated models, explicit return path modeling proved to be a tremendous benefit for BSB and system bus electrical modeling.

Since electromagnetic signal coupling adversely impacts timing, we modeled a complete source synchronous data group to include electromagnetic system effects between data signals and strobes. The total simulation timing path extends from the input of the output buffer through all package and interface interconnections to the output of the receiver, as shown in Figure 7. Simulation stimuli include patterns to cover ISI effects. These effects are important because the signal reflection behavior on the bus can be a function of previous cycles, possibly as left-over reflections due to

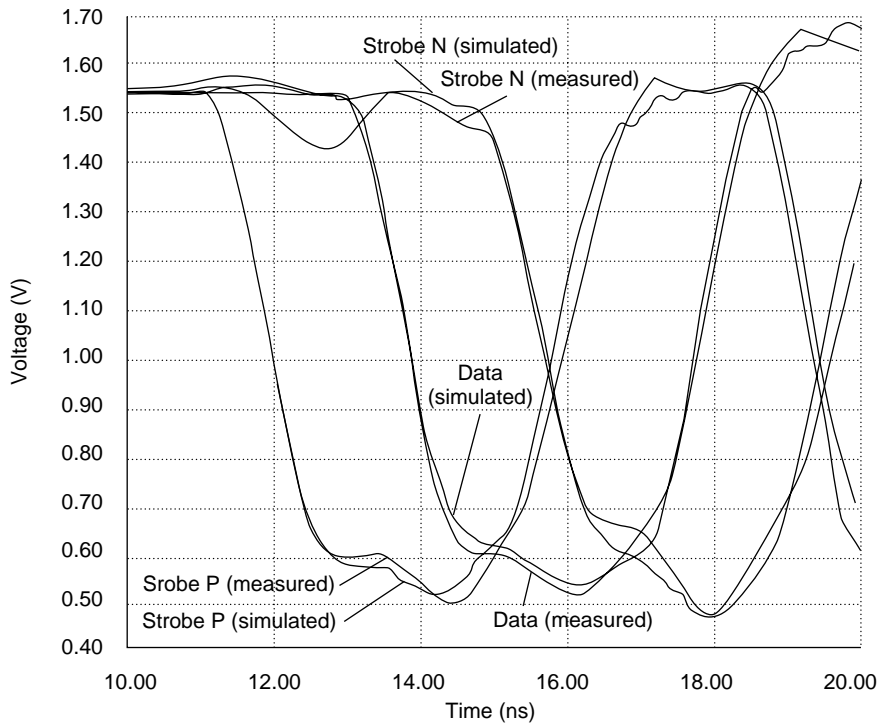


Figure 8. Bus signal waveform correlation.

inadequate settling time.

The simulation stimuli also include even-mode (all signals in phase) and odd-mode (all but one signal in phase) excitations to capture effects of cross talk, and SSO, with just one transient analysis. Simulation transient sweeps cover worst-case conditions: silicon process, voltage, temperature, interconnect geometry, and dielectric variations. These simulation-based analysis methods determined return path requirements and signal-to-ground ratios, and quantify timing skews to develop a robust system bus and BSB electrical design.

Figure 8 is a correlation example of actual physical system bus measurements (with Itanium processors) and full system simulations under similar conditions. This transient response shows one data signal and its differential strobe pair from a group of switching signals. Excellent signal quality correlation is seen between simulation predictions and system measurements.

A complete reference model—which includes the system bus I/O buffer models, package, and process information in Hspice—is available to customers to enable them to design systems using the Itanium processors.

The driving force behind the Itanium processor cartridge is the electrical bus performance for the external system bus and the internal L3 BSB. Careful electrical modeling of these buses allowed us to achieve our ambitious bus performance goals. We also developed advanced power delivery and thermal management techniques to address performance requirements. The resulting Itanium processor cartridge is a compact packaging solution that enables high-performance computing in a multiprocessing system environment.

Our team is now developing electrical and packaging solutions for future IA-64 and IA-32 processors to meet the demands of future high-performance enterprise computing.

MICRO

References

1. E. Dahlen, "The 82460GX Server/Workstation Chip Set," *IEEE Micro*, Nov.-Dec. 2000, pp. 69-75.
2. F. Aono and M. Kimura, "The Azusa 16-Way Itanium Server," *IEEE Micro*, Sep.-Oct. 2000, pp. 54-60.
3. T. Arabi et al., "Modeling, Simulation, and Design Methodology of the Interconnect and Packaging of an Ultra-High Speed Source Synchronous Bus," *Proc. 1998 IEEE Seventh Topical Meeting on Electrical Performance of Electronic Packaging*, IEEE Press, Piscataway, N.J., 1998, pp. 8-11.
4. E.A. Reese et al., "A Phase-Tolerant 3.8 GB/s Data-Communication Router for a Multiprocessor Supercomputer Backplane," *Proc. 1994 IEEE Int'l Solid-State Circuits Conf., Digest of Technical Papers 41st ISSCC*, 1994, pp. 296-297.
5. B. Gunning et al., "A CMOS Low Voltage Swing Transmission Line Transceiver," *Proc. 1992 IEEE Int'l Solid-State Circuits Conf.*, IEEE Press, 1992, pp. 58-59.
6. A.E. Ruehli, "Inductance Calculations in a Complex Integrated Circuit Environment,"

IBM J. Research and Development, Sept. 1972, pp. 470-481.


- P.A. Brennan et al., "Three Dimensional Inductance Communications With Partial Element Equivalent Circuits," *J. Research and Development*, Vol. 23, No. 6, Nov. 1979, pp. 661-668.

William (Bill) Samaras is a principal engineer and the engineering manager of the Itanium Processor Cartridge Development Group at Intel in Santa Clara. Previously, he worked for Digital Equipment Corporation in Massachusetts, specializing in high-speed interconnects, bus design, and clocking systems. He was one of the contributors to the electrical section of the original PCI specification and the electrical sections of IEEE Std. 896.2, and a visiting lecturer at the University of Massachusetts, Lowell, as an electronics instructor. Samaras received a BSEE degree from the University of Massachusetts, North Dartmouth.

Naveen Cherukuri, a lead designer on the Itanium processor cartridge at Intel, is responsible for the external L3 bus (BSB) design and overall cartridge electrical and physical design. Earlier, he worked for Fujitsu Microelectronics in MCM design for portable applications and signal integrity analysis on memory boards. He also worked as a software engineer at Tata Consultancy Services, India. Cherukuri has a BSEE degree from Jawaharlal Nehru Technological University, Kakinada, India, and an MSEE degree from the University of Arizona, Tucson.

Srinivas Venkataraman is responsible for the electrical design, interconnect modeling, and analysis of the Itanium processor system bus interface. Earlier, he worked at Tandem Computers in Cupertino, Calif., where he was responsible for the electrical design of high-speed buses on systems based on the MIPS R10K processor. Srinivas received an MS degree in electrical engineering from the University of Arizona, Tucson.

Direct questions about this article to W. Samaras, Intel Corporation, Mailstop SC12-214, 2200 Mission College Blvd., Santa Clara, CA 95052; bill.samaras@intel.com.



Electrical and Computer Engineering

*On the edge and leading the way:
The University of Calgary is an innovative university that leads a path of discovery and equity with dedication, achievement and quality learning experience.*

The Department of Electrical and Computer Engineering has been funded to implement two new undergraduate degree programs leading to the degrees of Bachelor of Science in Electrical Engineering and Bachelor of Science in Software Engineering. The construction of a new building to provide additional space for these programs has begun. Enrollment positions at the Assistant, Associate and Full Professor levels will be available over the next three years.

The Department of Electrical and Computer Engineering invites applications for full time, tenure track faculty appointments at the Assistant, Associate and Full Professor levels in the areas of Computer Engineering, Electrical Engineering and Software Engineering. Rank and salary are commensurate with qualifications and experience. Applicants are encouraged to apply as soon as possible for positions which are currently open and for positions with a starting date of July 1, 2001.

Qualifications: Successful candidates will have excellent academic credentials, the ability to develop strong independent research programs, and to teach effectively at the undergraduate and graduate levels. Depending on the position being applied for a PhD in electrical, computer, or software engineering or a related area is required. Candidates who are nearing the completion of their PhD programs and arranged to apply demonstrate ability in written and oral use of the English language is required. The Department is particularly interested in applicants with a demonstrated background in:

Electrical Engineering <ul style="list-style-type: none"> • microelectronics • communications • telecommunication • signal processing • IC design • systems engineering • biomedical engineering • embedded systems 	Computer Engineering <ul style="list-style-type: none"> • computer architecture • embedded systems • networks • digital systems • digital signal processing 	Software Engineering <ul style="list-style-type: none"> • software architecture • software tools • software methods • software reliability • software quality • empirical software methods
--	---	---

How to apply: Applications including a curriculum vitae and the names and addresses of three confidential references should be sent to:

Dr. Ronald H. Johnston, Dept.
Department of Electrical and Computer Engineering
University of Calgary
2500 University Dr. N.W.
Calgary, AB, CANADA T2N 1N4
Fax: (403)207-4050

Applications may be sent electronically as text documents via e-mail to hr@ece.ualberta.ca

RESEARCH FUNDING OPPORTUNITY: Calgary is noted for a high quality lifestyle and a rapidly expanding high-tech Information and Communication Technology industry. Federal and industrial research funding is available. In addition the Alberta Government has introduced substantial funding for ICT research through CORE and ABERE (<http://www.ualberta.ca>).

The University respects, promotes and encourages diversity. Applicants from diverse and disadvantaged backgrounds are encouraged to apply. Reasonable accommodations for those persons with documented disabilities will be provided upon request and personal information will be kept confidential.

www.ucalgary.ca