

A 1.5GHz Third Generation Itanium[®] 2 Processor

Jason Stinson and Stefan Rusu
 Intel Corporation
 2200 Mission College Blvd., M/S SC12-506
 Santa Clara, CA 95052
 1-408-765-5739
 {jason.stinson, stefan.rusu}@intel.com

ABSTRACT

This 130nm Itanium[®] 2 processor implements the Explicitly Parallel Instruction Computing (EPIC) architecture and features an on-die 6MB, 24-way set associative L3 cache. The 374mm² die contains 410M transistors and is implemented in a dual-Vt process with 6 layers copper interconnect and FSG dielectric. The processor runs at 1.5GHz at 1.3V and dissipates a maximum of 130W. This paper reviews circuit design and package details, power delivery, RAS, DFT and DFM features, as well as an overview of the design and verification methodology. The fuse-based clock de-skew circuit achieves 24ps skew across the entire die, while the scan-based skew control further reduces it to 7ps. The 128-bit front-side bus supports up to 4 processors on a single bus with a bandwidth of up to 6.4GB/s.

Categories and Subject Descriptors

K.1 [Intel Itanium[®] Processor]: Design and verification methodology, reliability, test and manufacturability features.

General Terms

Design, Reliability, Verification.

Keywords

Processor, test, reliability, on-die cache, design methodology.

1. INTRODUCTION

This Itanium[®] 2 processor is implemented in a dual-Vt 130nm process with six copper interconnect layers and FSG dielectric and features a 6MB, 24-way set associative on-die L3 cache. Table 1 summarizes the process characteristics [1,2].

The design implements a 2-bundle 64-bit Explicitly Parallel Instruction Computing (EPIC) architecture and is fully compatible with the previous implementations [3,4], including hardware support for in-order IA-32 binary execution. The 374mm² die contains 410M transistors. The processor operates at 1.5GHz from a 1.3V core supply and is plug-in compatible with the existing Itanium[®] 2 platforms. The worst-case power dissipation is 130W, while the power dissipation on a typical server workload is 110W.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DAC 2003, June 2-6, 2003, Anaheim, California, USA.
 Copyright 2003 ACM 1-58113-688-9/03/0006...\$5.00.

Table 1. 130nm Process Characteristics

| Attribute | Value |
|-------------|---------------------|
| Lgate | 60nm |
| M1 pitch | 350nm |
| M2 pitch | 448nm |
| M3 pitch | 448nm |
| M4 pitch | 756nm |
| M5 pitch | 1120nm |
| M6 pitch | 1204nm |
| Dielectric | FSG, K=3.6 |
| Memory cell | 2.45μm ² |

The front-side bus is 128-bit wide and provides a total bandwidth of 6.4GB/s in a 4-way multi-drop bus configuration. Table 2 summarizes the main attributes of this processor compared to the previous implementations. The on-die cache size (6MB) and the transistor count (410M) for this processor are the largest ever reported for a microprocessor. While the frequency of this processor is 50% higher than the previous generation, we kept the total power dissipation flat at 130W to ensure backward platform compatibility. Figure 1 shows the die photo and highlights the location of the main functional blocks.

Table 2. Itanium[®] Processor Comparison

| Attribute | Itanium [®] Processor | Itanium [®] 2 Processor | This work |
|-----------------|---|----------------------------------|-----------|
| Architecture | Explicitly Parallel Instruction Computing | | |
| Process | 180nm | 180nm | 130nm |
| Device Count | 25M | 221M | 410M |
| On-die L3 cache | 0* | 3MB | 6MB |
| Frequency | 800MHz | 1.0GHz | 1.5GHz |
| Supply Voltage | 1.6V | 1.5V | 1.3V |
| Max. Power | 130W* | 130W | 130W |

* Includes 4MB cache on cartridge

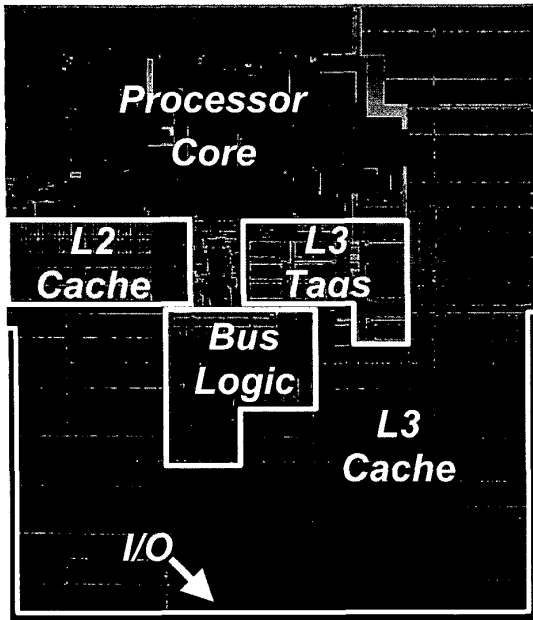


Figure 1. Die Photo

2. ARCHITECTURAL OVERVIEW

The integer execution unit has a 128x65b 20-port integer register file combined with six integer and six multi-media units (1-cycle and 2-cycle respectively) with full symmetric bypassing between each other and the L1D cache. The floating-point unit has a 128x82b register file combined with two FMAC units (4-cycle) that are fully bypassed with each other. The cache is organized in three hierarchical levels. The first level instruction and data caches are 4-ways 16kB each. The second level is a unified, 8-ways, 256kB array, while the third level is a unified, 24-way set associative, 6MB cache. While the latencies of the first two levels of cache are the same as in the previous generation, interconnect constraints increased the L3 latency from 12 to 14 cycles.

3. POWER REDUCTION

Figure 2 shows the power distribution of the current design and its 180nm predecessor. Both products fit in the same 130W system power envelope. However the current design has a 50% higher frequency, a 2x larger L3 cache and a 3.5X increase in transistor leakage. In order to stay within the same power envelope, the active power had to be reduced from 90% to 74% of the total. This was accomplished through an aggressive management of the dynamic power, focused on three areas: reduced clock loading, lower contention power and better L3 cache power management. To reduce the clock loading, we replaced about half of the approximately 50k clocked deracers with static buffers. We also redesigned the dual-rail domino library cells to reduce their clock loading by sharing the clocked NMOS device between the complementary pull-down stacks without affecting performance while preserving the footprint and port locations.

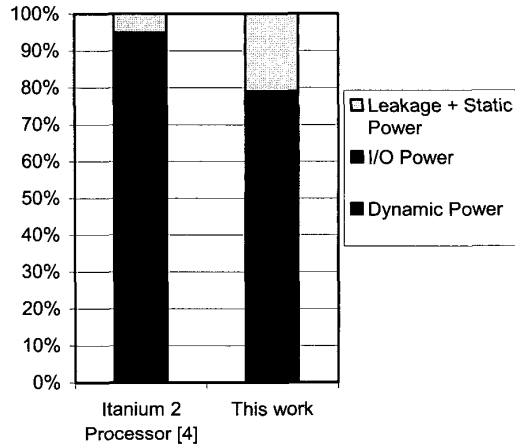
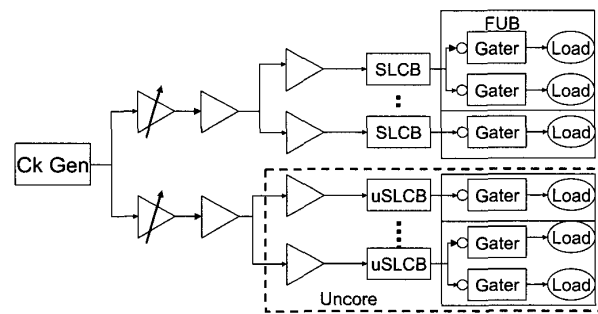


Figure 2. Power Breakdown

4. CLOCK DISTRIBUTION

Figure 3 illustrates the microprocessor's clock distribution architecture that consists of the global and zonal distributions. The global clock distribution network is a multi-level tree structure with each branch individually tuned to achieve equal delay. It includes the clock generator, the main primary driver (MPD), the primary driver (PD), the global clock repeater (RPTR) and the second-level-clock-buffer (SLCB). The global distribution is a full swing differential clock designed to minimize the impact of supply and coupling noise. The global clock uses the top two metal layers exclusively. Shielding for the differential global clock routes is implemented with metal in the same metal layer and by the underlying adjacent co-linear metal layer. The zonal distribution consists of the SLCB driver and the accompanying SLCBO clock tree. There are 32 zonal clocks each served by a dedicated SLCB buffer. The SLCB buffer converts the differential global clock into the single ended SLCBO clock and drives the local clock buffers. Similar to the global clock network, each SLCBO tree is delay matched to attain low skew.



PLL → MPD → PD → RPTR → SLCB → Gater

Figure 3. Clock Distribution Hierarchy

To facilitate post-silicon speed path balancing, the clocking architecture includes a deskew function that can be accessed through a fuse array or the scan chain. Each SLCB buffer has a 5-bit delay control register that is fully addressable via a dedicated scan chain. Additionally, three of the five register bits are addressable by dedicated fuses embedded within the fuse unit. Reducing the number of bits addressable via fuses economizes the fuse usage. The design incorporates 23 deskew zones and 69 dedicated deskew fuses. Two additional fuses are designated as mode control bits that are read during the initial microprocessor power-up. Depending on the status of these mode control bits, either the default deskew settings or the fuse settings will be used in the SLCB delay register. Figure 4 shows the skew reduction after applying the scan- and fuse-based deskew adjustments. The worst-case skew without applying any deskew technique would be about 60ps. The average step sizes in the scan mode and the fuse mode are 7ps and 27ps respectively. The total skew adjustment range is about 210ps. The corresponding global skews achieved are 7ps and 24 ps respectively. The scan mode is used for product debug and speed path analysis, while the fuse mode is used for permanent production settings.

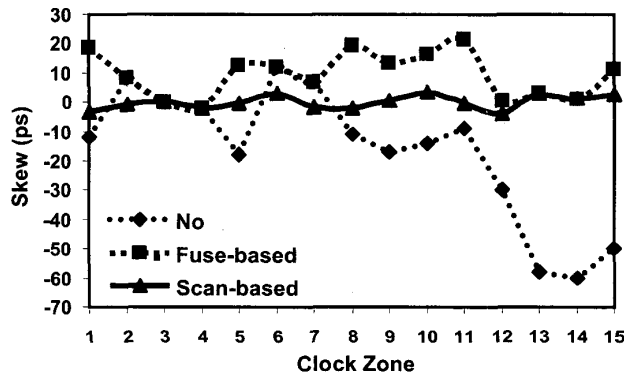


Figure 4. Clock Zone Skew Plot

5. PACKAGE DESIGN

Figure 5 is a photo of the processor package. The processor is flip-chip mounted to a 12-layer organic ball-grid array (BGA) package with an integrated heat spreader measuring 42.5mm on each side. The BGA is attached to an 8-layer interposer that has an edge connector for direct power delivery from the power pod and houses additional components for server management functions and decoupling. Interdigitated caps are placed on both the BGA substrate (underneath the heat spreader) as well as on the interposer, for both the core voltage and the front-side bus termination voltage. Inductive signal return current loops are minimized by proper placement of return vias for image currents propagating in the reference planes inside the multi-layer package.

Core power is delivered to the processor interposer assembly through an edge connector that provides lower impedance compared to traditional power delivery using pins through the socket. The I/O circuitry is powered through the interposer pins by a separate supply located on the motherboard. The chip implements a hybrid pitch C4 bump pattern that allows the direct shrink of the M6 routing from the previous implementation and

enables maximum C4 connections at the package 213 μ m C4 bump pitch. Maximizing C4 bump density addresses the higher current densities and increased core frequency of this design. About 95% of the 7877 package C4 bumps are devoted to power delivery connections.

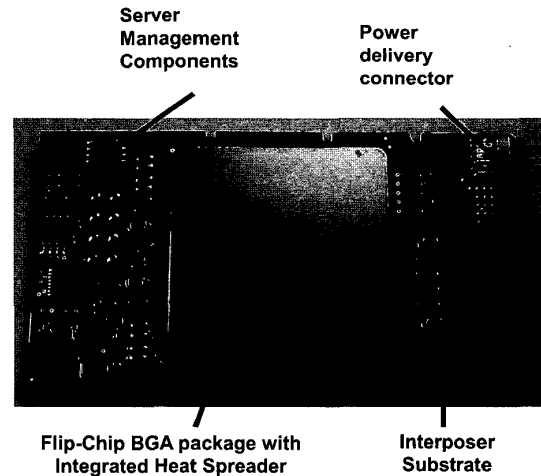


Figure 5. Package Implementation

6. DATABASE MANAGEMENT

The hierarchical circuit and layout database is broken into 370 large macro blocks ranging in size from 10K devices up to about 300K devices. In addition, approximately 400 unique small gate level devices are placed directly at the full chip level (typically in the routing channels) to optimize routing and timing constraints. To enable a stronger consistency across the design, a central validation archive is built around the entire design database to capture all validation data inputs, flows and expected results for every block in the design. With automated runs for each block in the database, all physical verification flows for the design can be turned within less than a week, with no intervention from the designers. This enables faster convergence by keeping the entire database up to date and helping to free the designers to focus on real issues rather than running the tools. Additionally, the central archive provides an invaluable mechanism to verify robustness of new tool releases and measure the impact of potential design rule changes. Lastly, the archive makes the design database highly portable by capturing all of the critical design flow information for each block and recording input stimulus and expected outputs. This is particularly important in enabling future proliferations to turn new versions of the design more quickly.

7. TIMING VERIFICATION

The timing validation was one of the most significant tasks, representing close to 30% of our overall design effort. The product goal was to push the design to 1.5X the frequency of the 180nm Itanium® 2 processor. Additionally, because of the extensive use of self-timed and pulsed circuitry, as well as complex circuit topologies, min-delay and pulse margin scaling to

130nm were a significant concern for robustness. Timing analysis was critical in verifying electrical reliability.

Timing validation is done using hierarchical static timing analysis tools. To build a full timing model, the design is broken into two basic hierarchies: block and full-chip levels. The central archive mechanism verifies macro blocks individually. Block information is then abstracted into proprietary models and rolled up into regular full-chip timing runs. Full-chip routes are simulated using complete netlists containing both driver and receiver information. Results from the full-chip runs are then “rolled-down” to the block level as interface timing specs. Full-chip timing analysis and macro block abstract models are designed to significantly reduce the iteration required to converge a design, including complex timing slack allocation and accurate modeling of block transparencies.

Timing analysis uses fully extracted parasitics at all hierarchy levels. At the macro block level, full-chip routing is “shadowed-down” for parasitic extraction to improve accuracy of signal coupling. All capacitive coupling information between signals is preserved in the parasitic output files and is used for both timing and noise analysis. At the full-chip timing level, a dynamic coupling factor is applied to every signal. Bus attackers are collapsed and only the top attacking signals are considered switching in the opposite direction relative to the victim for max-delay analysis and in the same direction for min-delay analysis.

Both min-delay and max-delay analyses are done at a single process corner. Additional timing margin is applied to the min-delay corner to account for PVT variation as well as clock skew. The vast majority of the block design is analyzed at the transistor level. Although this adds some additional runtime compared to a gate level analysis, significant benefit comes from being able to quickly turn process file changes as well as better timing accuracy.

8. SIGNAL INTEGRITY

Signal integrity analysis follows the same hierarchical approach as the timing analysis, sharing common data and database management. The verification is done at a macro block level, abstracted to proprietary models and rolled up to a full-chip analysis. Data from the full-chip analysis is then rolled back down to the block level for re-validation.

At a block level, the signal integrity analysis tool uses a complex algorithm to extract gates into simple equivalence circuits that include models for charge-sharing, input noise amplification, current drive fights, ground bounce and active feedback. Interconnect analysis models capacitive coupling, including time window filtering, bus collapsing and attacking signal switching slopes. All noise sources are assumed to be additive. Noise is dynamically propagated from one stage to the next, roughly modeling electrical amplification. Signal integrity violations are flagged based on dynamically calculated output noise threshold limits for each gate. Since timing delay push-out due to noise events is already modeled in the timing analysis tools (through such mechanisms as dynamic capacitive coupling factors and contention degradation), signal integrity threshold limits are set to

the point of circuit failure. This significantly reduces the number of false violations, without incurring a timing or reliability penalty.

To account for signal integrity impact of the inductive effects, a correct-by-construction approach is taken to eliminate the need for complex inductance extraction and analysis. The power and ground grids for metals 4, 5 and 6 are finely interspersed to minimize inductive loops. All signal wire lengths are kept below 2mm—signals that must travel further than this distance are repeated to prevent excessive inductance loops.

9. ELECTRICAL RELIABILITY

All nets are verified for electro-migration (EM) and self-heating (SH), while all transistor gates are verified for hot-electron (HE) effects. The power and ground grid for metals 5 and 6 are designed to be correct-by-construction, while metals 4 and below are verified using the EM validation tools. In an effort to reduce false violations, logical signal activity factors are applied to all nets to model realistic switching behavior. Additionally, the die thermal map is used to adjust EM and SH violations, relaxing constraints in cooler areas of the die and tightening them in warmer areas.

10. SUMMARY

This 130nm Itanium[®] 2 processor delivers a 50% frequency increase over the previous implementation while maintaining the same power envelope. This processor breaks a new record in on-die cache size and overall total transistor count for a microprocessor. The fuse-based clock de-skew technique achieves 24ps skew across entire die, while the scan-based skew control further reduces it to 7ps. We reviewed the design methodology that enabled high design productivity and guaranteed a reliable circuit behavior.

11. ACKNOWLEDGMENTS

The work of a talented and dedicated team is presented in this paper. The authors feel privileged to represent their work. Our inheritance from the McKinley team of the core micro-architecture and circuits is also gratefully acknowledged.

12. REFERENCES

- [1] S. Tyagi, et al, - “A 130nm generation logic technology featuring 70nm transistors, dual Vt transistors and 6 layers of Cu interconnects”, *IEDM Tech. Digest*, Dec. 2000
- [2] S. Thompson, et al, - “An enhanced 130nm generation logic technology featuring 60nm transistors optimized for high performance and low power”, *IEDM Tech. Digest*, Dec. 2001
- [3] S. Rusu, G. Singer – “The first IA-64 microprocessor,” *IEEE J. Solid-State Circuits*, Nov. 2000, pp 1539–1544
- [4] S. Naffziger, G. Hammond – “The implementation of the next-generation 64b Itanium[®] microprocessor”, *ISSCC Dig. Tech. Papers*, Feb. 2002, pp 344–472