A 45 nm 8-Core Enterprise Xeon® Processor

Stefan Rusu, Fellow, IEEE, Simon Tam, Senior Member, IEEE, Harry Muljono, Member, IEEE, Jason Stinson, Member, IEEE, David Ayers, Member, IEEE, Jonathan Chang, Senior Member, IEEE, Raj Varada, Member, IEEE, Matt Ratta, Sailesh Kottapalli, and Sujal Vora

Abstract—This paper describes a 2.3 Billion transistors, 8-core, 16-thread, 64-bit Xeon® EX processor with a 24 MB shared L3 cache implemented in a 45 nm nine-metal process. Multiple clock and voltage domains are used to reduce power consumption. Long channel devices and cache sleep mode are used to minimize leakage. Core and cache recovery improve manufacturing yields and enable multiple product flavors from the same silicon die. The disabled blocks are both clock and power gated to minimize their power consumption. Idle power is reduced by shutting off the unterminated I/O links and shedding phases in the voltage regulator to improve the power conversion efficiency.

Index Terms—Circuit design, clock distribution, computer architecture, core recovery, 45 nm process technology, leakage reduction, microprocessor, voltage domains.

I. INTRODUCTION

The next-generation Xeon® EX processor, code-named Nehalem-EX, has eight dual-threaded 64-bit cores and a 24 MB shared L3 cache. The processor has 2.3 B transistors and is implemented in a 45 nm CMOS technology using metal gate high-K dielectric transistors [1] that reduce the gate leakage by a factor of 25× for nMOS devices and 1000× for pMOS devices, compared to the 65 nm process generation. Fig. 1 shows a transistor cross-section. The metal interconnect system has nine copper layers with a thick (7 μm), low resistance M9 interconnect that enables the on-die power gates. The thermal design power is 130 W. Fig. 2 shows the trend of the number of cores per socket increasing by two additional cores every year for the Xeon® EX processor line. Fig. 3 shows the processor die photo with the major blocks marked out.

II. PROCESSOR ARCHITECTURE

Fig. 4 shows the block diagram of the processor, including the top stripe holds four Quick Path Interconnect (QPI) links, while the bottom stripe houses the Scalable Memory Interconnect (SMI) links. The center channel holds the system interface that includes two memory controllers, two hub interfaces to the last level cache, an 8-port router, the power control unit (PCU) and the DFX control box. The processor supports multiple platform configurations, from dual processor options to 4-socket and 8-socket glueless options.

III. CORE DESIGN

Fig. 5 shows the core floorplan with its major units highlighted. The core adds new SSE4.2 instructions, improved lock support, enhanced memory ordering and execution, improved loop streaming, deeper buffers, simultaneous multi-threading (SMT) with two threads per core, improved virtualization, and better branch prediction to improve overall throughput and performance. All domino data-path circuits were replaced with static CMOS with no additional pipeline stages added or frequency loss [2]. The only remaining precharge/discharge circuitry is in the memory arrays and register files. The core memory arrays were re-designed to enable a lower core-level VMIN (0.85 V) for low power operation. Additional buffers were added in the branch prediction logic to improve performance on large code size (e.g., database applications).

Each core is equipped with on-die power gates as shown in Fig. 6. The thick M9 layer is used to minimize voltage drop in on-die power distribution. The core architecture supports turbo mode for the cores to enable higher voltage/frequency operation on a given core when other cores are idle to fit within the overall...
power envelope. The power gates allow zero power for an idle core at the same time enabling additional performance from the active cores.

IV. L3 CACHE DESIGN

The L3 cache is built from eight 3 MB slices, each having 2048 sets and 24 ways [3]. The slices are interconnected so they are seen by the processor cores as a single, distributed 24 MB L3 cache. Each cache line is 64 bytes constructed in two chunks. There are a total of 48 sub-arrays in each slice [4] as shown in Fig. 7. To reduce the L3 cache slice active power, only 3.125% of the arrays are powered up for each access. The data array is built with 0.3816 $\mu$m$^2$ 6T cell and protected by inline double-error-correction and triple-error-detection (DECTED) ECC scheme with variable latency. Tag arrays are built with 0.54 $\mu$m$^2$ 6T cell and are protected by inline single-error-correction and double-error detection (SECDED) ECC scheme with fixed latency. Data arrays have both column and row redundancy, while the Tag arrays have only column redundancy. Since the number of fuses required is too large to be stored on die, the L3 cache redundancy fuses are stored in an on-package serial EEPROM that provides much larger capacity. The fuse values are shifted into on-die radiation-hardened storage elements during the reset sequence.

V. CORE AND CACHE RECOVERY

If a processor core has a defect, we can recover that die by de-featuring the defective core. Since each core has its dedicated power gating device, disabled cores do not burn leakage power in the customer’s system. For typical operating conditions, a core in shut-off mode achieves about 40x leakage reduction compared to the ungated case. This technique is especially suited for recovering parts that have a VCC-to-VSS short in the core, since the power gate will remove this short condition.

Similarly, if a defect occurs in the non-repairable area of a cache slice (such as the tag data-path), the die can be recovered by disabling that defective cache slice. For the data arrays, all SRAM cells, word line drivers and peripheral circuitry (sense amplifiers, column multiplexers and write drivers) are placed in shut-off mode to minimize the leakage of disabled cache slices. For the tag arrays, shut-off mode is implemented in the SRAM cells and word-line drivers only. During shut-off, the power supply of SRAM drops from 0.90 V nominal operating voltage
to 0.36 V, providing an 83% leakage power saving. By comparison, the leakage power saving for the SRAM cells at the target sleep voltage (0.75 V) is about 35%.

Fig. 8 shows an example of the core and cache recovery scheme. The cores and cache slices are disabled in horizontal pairs, but disabled cache and core pairs do not have to be aligned, as shown in this example. There are many combinations possible that cover a majority of the die area. The core and cache recovery technique is a virtual die chop mechanism that allows us to serve different market segments (4, 6 and 8 cores) with the same die, without bearing the costs of additional tape-outs and multiple packages.

Fig. 9 shows the backside infrared emission of the same die. Notice that all infrared emission disappeared in the disabled cores. The one little spot left in each core represents the thermal sensor that is connected to the clean PLL supply and is not controlled by the core supply power gates. The leakage reduction in the bottom two cache slices is more difficult to see because there is very little leakage left in the cache due to the high-K metal gate process technology. Notice that the vertical bus that interconnects the cache slices must stay enabled such that any core can access any cache slice.

VI. CLOCK GENERATION AND DISTRIBUTION

The processor has three primary clock domains [5], as shown in Fig. 10: a core clock domain (MCLK), an uncore clock domain (UCLK) and an I/O clock domain (QCLK). There is an independent clock generator for each of the eight cores, one central clock generator for the uncore that includes the system interface and the eight LLC cache slices, and an independent clock generator for each of the QPI and SMI I/O interfaces. The processor architecture ensures that the MCLKs in the eight cores are operating at the same frequency. On the other hand, each of the four QPI I/O interfaces could operate at different rates to maximize the system flexibility. The reference clock to the core, uncore and I/O clock generators is derived from a dedicated on-die filter-PLL that interfaces to the system clock input. The reference clocks of the filter-PLL are distributed to the aforementioned clock generators in a balanced fashion to minimize the skew.

The uncore UCLK domain is a unified clock domain that serves the system interface functional blocks, the eight 3 MB last-level cache slices, and the on-die interconnect fabric. Since the UCLK domain has the largest span and serves the largest number of transistors, the UCLK distribution has to exhibit low skew and low power consumption. To achieve this, the UCLK distribution relies on vertical and horizontal clock spines with embedded clock compensators. Within the system interface, horizontal clock spines located at the top, middle, and bottom sections of the region are used to form three UCLK grids each individually serve the left, center and right regions of the area. Multiple vertical clock spines serve the LLC and the on-die interconnect-fabric. Due to the regular and predictable loading pattern in the LLC arrays, a point-to-point clock network serves the clocking needs in the arrays. The vertical clock spines also drive local grids to serve the on-die interconnect fabric. Fig. 11 shows the UCLK skew profile simulated with post-layout extraction based model containing over 34 K grid receivers that exhibits less than 21 ps skew and was achieved with the clock compensators turned off.

The processor’s timing analysis model employs a zone-to-zone based skew budgeting methodology that identifies the source clock domain and the receiver clock domain and
assigns the skew margin accordingly. Although the skew profile from Fig. 11 could be directly incorporated in the timing analysis model, the domain based skew budgeting methodology is simpler and exhibits adequate design margin to enable a quick design convergence without compromise to the design robustness. Moreover, the embedded clock compensators incorporated in the global clock distribution coupled with the delay-tunable regional clock buffers (RCB) incorporated within the local blocks provide the additional clock delay profile tuning flexibility to maximize the post-silicon clock performance. Specifically, each tunable regional clock buffer has 4 delay control bits partitioned into 2 bits for the rise edge delay adjust and 2 bits for fall edge delay adjust. A dedicated shift chain served by an on-package non-volatile memory stores the delay settings for all the RCBs to enable post-silicon clock delay optimization.

VII. VOLTAGE DOMAINS

The processor uses four voltage supplies: one for the eight cores, a separate supply for the L3 cache and system level interface, and the third supply for the I/O circuits. The fourth supply provides clean voltage to the PLLs and on-die thermal sensors. Level shifters are used between voltage domains. Table I shows the operating voltages and TDP power levels. While the uncore supply is fixed, the range in the table indicates the voltage window that parts will be binned to for different process corners. The design uses longer channel devices in non-timing critical paths to reduce the subthreshold leakage. About 58% of the transistor width in the cores and 85% of the transistor width in the uncore (excluding cache arrays) are long-Le. Overall, leakage accounts for about 16% of the total power at the typical process corner.

VIII. PACKAGE

The processor is flip-chip (C4) attached to a 14-layer (5–4–5) organic land grid array package with 1567 lands at 40 mil pitch and an integrated heat spreader, as shown in Fig. 12. A rectangular land-side cavity underneath the die contains decoupling caps for cores, uncore and I/O links. The top side of the package holds the system management components—an EPROM for processor information and a serial flash memory for storing external fuse values. The package is lead-free and halogen-free. The chip-level power distribution consists of a uniform M9-M8 grid aligned with the C4 power and ground bump array.

IX. I/O LINKS

All I/O links can run asynchronously at 6.4 GT/s, providing up to 25.6 GB/s per port. Each link supports data and clock fail over RAS features, as well as half-width and quarter-width lanes. The link transmitter requires the output swing to meet minimum voltage requirement to guarantee minimum eye-opening at the receiver side. The TX also needs to meet maximum output swing requirement to save power and reduce electro-magnetic interference (EMI) and cross-talk. To meet these requirements, a precise PVT compensation scheme is added to control the TX output swing variation. This is accomplished by comparing the pad voltage to a reference voltage for each lane during the calibration state for both clock and data lanes. The link receiver utilizes the receiver (RX) equalization (EQ) architecture based on Continuous Time Linear Equalization (CTLE) design with complete offset cancellation in order to mitigate the Inter-Symbol Interference (ISI) associated with high speed switching. The CTLE consists of a linear amp with gain greater than one to amplify the incoming signal along with adjustable R and C components to control DC/AC gain, as shown on Fig. 13. The data path contains two CTLE amplifiers, each with its own offset cancellation capability, to capture even and odd data. As the result, the residual offset is minimal and the receiver CMRR and PSRR stays high.

The high speed link interface requires an accurate reference current (Iref) to bias the current-mode analog circuits. This is accomplished by compensating each Iref through a compensation loop that compensates for process variations and continuously tracks for voltage and temperature (VT) variations. Each Iref has its own local IDAC and a digital counter which controls the amount of current to be added or subtracted in the local IDAC. In addition, there are current multiplier, external

<table>
<thead>
<tr>
<th>Domain</th>
<th>Voltage</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core supply</td>
<td>0.85 – 1.15 V variable</td>
<td>71 W</td>
</tr>
<tr>
<td>Uncore supply</td>
<td>0.85 – 1.175 V fixed</td>
<td>44 W</td>
</tr>
<tr>
<td>I/O supply</td>
<td>1.1 V fixed</td>
<td>14 W</td>
</tr>
<tr>
<td>PLL clean supply</td>
<td>1.8 V fixed</td>
<td>1 W</td>
</tr>
</tbody>
</table>
The guiding principle for idle power reduction is to minimize the power consumption in unused blocks. One example is cutting the power dissipation of the unused I/O links. There are multiple platform configurations that leave some I/O links unterminated at the other end. Dual processor platforms typically use only two or three links (out of the total four available on each die), leaving the rest unused. Partially populated quad processor platforms also have several links unconnected. To detect these conditions, we implemented a link detect circuit shown in Fig. 15 that senses the presence of the Rx termination resistor at the other end of the link. On the left side of the figure, the Rx term is present and it pulls down the link below the half VCC level (Rx termination resistor on the receiving chip is much lower in value than the pull-up resistor on the driving die). On the right side of the figure, the Rx termination resistor is missing (i.e., there is no chip at the other end of the link) and the pull-up on the driving side brings the comparator input to a full VCC. The Link Detect signal shuts off the link PLL and the analog bias circuit, such that there is no active power dissipated in the unterminated link. This saves about 2 W per disabled link. The remaining leakage power is relatively low, since most I/O circuits use long channel devices to minimize the impact of process variations. Therefore, power gating each individual link would have only minimal benefits.

Another technique used to reduce the idle power is called Phase Shedding. This improves light load voltage regulator efficiency by dropping phases to avoid switching loss contributions from all phases. A typical on-board voltage regulator distributes its current equally amongst all phases. When the total current drawn by the microprocessor drops (e.g., in the idle state) each phase needs to supply a lower current that degrades its power conversion efficiency. This is illustrated by the 4-phase curve in Fig. 16, with the efficiency dropping fast below 15 A. To avoid this situation, this processor detects an idle state and reduces the number of active phases to retain the same average current per phase. In this way, we maintain an optimal efficiency of ~83% as shown in Fig. 16. This technique is applied to both core and cache voltage regulators, using an internal algorithm to detect the idle state for each supply. Overall, this technique reduces the idle power by about 2 W per socket.

X. IDLE POWER REDUCTION

The processor has nine thermal sensors, one for each core and one in the center of the uncore next to the router. Temperature data from each sensor, as well as their average and peak values can be read over the platform environment control interface (PECI) bus for system-level thermal management. Fig. 18 shows a thermal map with all eight cores enabled, while Fig. 19 has only two cores running and the others power gated. The microcontroller computes the voltage ID bits that program the external voltage regulator and the multiplier ratio for the core PLLs. The PCU also manages thermal alerts by reducing voltage and frequency to keep the die temperature within safe reliability limits. In case the temperature exceeds the reliability limit, the PCU will shut down the PLL and the external voltage regulator to protect the processor chip from a catastrophic failure.

XI. POWER MANAGEMENT

The processor has nine thermal sensors, one for each core and one in the center of the uncore next to the router. Temperature data from each sensor, as well as their average and peak values can be read over the platform environment control interface (PECI) bus for system-level thermal management. Fig. 18 shows a thermal map with all eight cores enabled, while Fig. 19 has only two cores running and the others power gated. The cache area in the middle of the die acts as a heat spreader for the cores located on the left and right edges. In the two core case, the processor operates in turbo-mode, where we take power
budget from the disabled cores and allocate it to the operating cores such that the top temperature matches the 8-core operation (about 80°C–82°C). In this way, we can increase the frequency by a few bins and provide more performance from the operating cores.

XIII. DESIGN FOR TEST AND MANUFACTURING FEATURES

The processor has a wide range of design for test, debug and manufacturing capabilities. There are nine TAP controllers, one for each core and one for the uncore. All blocks implement scan and observability registers. The L3 cache DFX features include a built-in pattern generator and direct array test. The I/O links have loopback and test generator support. Manufacturability features include extensive redundancy for the L3 cache, within-die process monitors, low-yield analysis and stability test mode in the L3 cache.

XIV. PHYSICAL DESIGN INTEGRATION

The die has long routes to connect between the I/O links on the top and bottom and the system interface functions in the center of the chip. This requires significant planning and implementation of signal repeating solutions for both buffer/inverter repeaters, as well as cycle stage repeaters. The entire system interface was built with a single repeatable interconnect tile with single pitch for each metal layer; this allowed for regularity and repeatable interconnect structures. The synthesized logic was implemented as large blocks with the block size being limited only by the repeating distance of the highest metal layer; all the long distance interconnect was half-shielded (power or

Fig. 15. Link detect circuit details: Terminated link on the left side, open link on the right side.

Fig. 16. Voltage regulator efficiency for different number of phases.

Fig. 17. Power Control Unit connectivity to the processor cores.

Fig. 18. Thermal map with all eight cores turned on.

Fig. 19. Thermal map with only two cores turned on.
ground on one side of the signal wire) and this enabled further stretching the repeating distance. With these design optimizations, we were able to limit the area and power overhead for the repeating solutions. The repeaters were inserted using a two step process [6]. During design convergence, virtual repeating solution was used to drive timing convergence and finally the physical repeaters were inserted without any impact to the design. The die has a total of 106 k buffer/inverter repeaters (65% of the repeaters are buffers and 35% inverters) implemented in 2100 buffer/inverter repeater blocks and about 8 k sequential cycle repeaters implemented in 125 sequential cycle repeater stations.

The uncore, except for the cache arrays, is implemented using a cell-based design methodology and uses a single repeatable interconnect tile with single pitch for each metal layer. With the exception of the arbiter block in the router, the system interface was designed using low leakage cells. Each block is initially mapped to low leakage cells and selected cells are converted later to nominal cells on timing critical paths. The register files are also assembled from a separate library of cells that use low leakage transistors. Most of the register files in the system interface are 1-read 1-write and use flat auto-routing methods [7].

VX. SUMMARY

We presented a 45 nm 8-core, 16-thread Xeon processor with a 24 MB L3 cache. Multiple voltage and clock domains are used to minimize the power consumption for each domain. Power gating is implemented at both the core and cache level to control leakage. The core and cache recovery techniques improve manufacturing yields and enable multiple product flavors from the same silicon die. Disabled core and cache blocks do not burn power in the customers’ systems. The I/O links operate at 6.4 GT/s, providing up to 25.6 GB/s per port. An on-die microcontroller manages voltage and frequency operating points, as well as power and thermal events. Idle power is reduced by shutting off the unteminated I/O links and by turning off phases to improve the voltage regulator conversion efficiency at low compute loads.

ACKNOWLEDGMENT

The authors gratefully acknowledge the work of the talented and dedicated Intel team that implemented this processor.

REFERENCES


Stefan Rusu (M’85–SM’01–F’07) received the M.S.E.E. degree from the Polytechnic Institute in Bucharest, Romania.

He first joined Intel Corporation in 1984 working on data communications integrated circuits. In 1988 he joined Sun Microsystems working on microprocessor design with focus on clock and power distribution, packaging, standard cell libraries, CAD and circuit design methodology. He rejoined Intel Corporation in 1996 working on the clock and power distribution, cell library, I/O buffers and package for the first Itanium® processor. He is presently a Senior Principal Engineer in Intel’s Enterprise Microprocessor Group leading the technology and special circuits design team for the Xeon® Processors Family. His technical interests are high-speed clocking, power distribution, I/O buffers, power and leakage reduction, and high-speed circuit design techniques. He has authored or co-authored more than 80 papers on VLSI design methodology and microprocessor circuit technology. He holds 33 U.S. patents with several more pending.

Dr. Rusu is a member of the Technical Program Committee for ISSCC, ESSCIRC, and A-SSCC conferences and an Associate Editor of the IEEE JOURNAL OF SOLID-STATE CIRCUITS. He received the ISSCC 2009 Beatrice Winner Award for Editorial Excellence and is a Distinguished Lecturer of the Solid-State Circuits Society.

Simon Tam (M’80–SM’07) received the B.S., M.S., and Ph.D. degrees, all in electrical engineering and computer sciences, from the University of California at Berkeley.

He is a Senior Principal Engineer at Intel Corporation’s Enterprise Microprocessors Group, where he has been engaged with the design of high performance microprocessors with special emphasis on high frequency clocking architecture and circuits. Prior to joining the Enterprise Microprocessor Group, he was with the Intel Neural Network Group. He designed electrically programmable neural network chips using analog VLSI techniques and EEPROM technology. He was also with the Intel California Technology Development Division engaged with the development of flash memory and EEPROM. He has been awarded 28 U.S. patents and has authored or co-authored 38 technical papers in the areas of microprocessor clocking, neural networks, flash/EEPROM memories, and VLSI device physics.

Dr. Tam was a member of the Technical Program Committee of the 2007, 2008, and 2009 Symposium on VLSI Circuits and the 2009 Custom Integrated Circuit Conference.


Since 1992, he has been with Intel Corporation designing high-speed I/O circuits for multiple processor products. He holds 30 patents with more pending in the area of I/O circuits and has co-authored eight conference and journal papers.


He is a Senior Principal Engineer in the Digital Enterprise Group, Intel Corporation, and is currently the physical design lead for a next-generation Xeon processor. He joined Intel in 1992 and has split his time between physical design methodologies and post-silicon validation.

ACKNOWLEDGMENT

The authors gratefully acknowledge the work of the talented and dedicated Intel team that implemented this processor.

REFERENCES

David Ayers (M’93) received the B.A. degree in business administration from California State University, Stanislaus, in 1982. He worked in finance, accounting, and consulting before becoming interested in electronics. He joined Intel Corporation in 1992 working in flash memory and then earned an M.S. in electronic engineering from Stanford University in 1994. He subsequently rejoined Intel where he is currently a Principal Engineer with the Enterprise Microprocessor Group in Santa Clara. His focus is on power delivery, power consumption and dissipation, packaging, and signaling.

Jonathan Chang (M’93–SM’06) received the B.S. degree in electrical engineering from National Taiwan University in 1990, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1994 and 1998, respectively. He joined Intel Corporation, Santa Clara, CA, in 1998 and since has been engaged in the design of several high-performance microprocessors with emphasis on large, high-speed, low power cache design. He is currently a technical lead in the area of cache design in the Enterprise Microprocessor Group, Santa Clara, CA, working on the next-generation Xeon® server processors.

Raj Varada (M’89) received the M.S.E.E. degree from the Indian Institute of Science, Bangalore, India. He currently is a Principal Engineer with the Graphics processor design team in Intel Corporation, Santa Clara, CA. He is the methodology lead for cell-based control, data path and register file design styles, and implementation lead for cell-based control logic for next generation Graphics Processors. Prior to Intel, he was with Semiconductor Complex Limited, India, and at IBM EDA Labs, East Fishkill, NY, on physical design and timing convergence tool development and at Texas Instruments, San Jose, CA, on early design planning solutions. His areas of interest include 22nm/15nm process design rules, design physical/timing convergence, physical implementation methodologies.

Matt Ratta received the Bachelors degree in computer engineering from the University of Illinois in 1986. He spent 15 years in the Design Technology group in Intel developing CAD tools and methodologies. In the last seven years, he has been in the Digital Enterprise group working on development of specific microprocessors design focusing on performance verification.

Sailesh Kottapalli co-led the architecture development of Beckton and was the Beckton chief micro-architect. Prior to Beckton, he contributed to a number of Intel Server CPU programs both on the Xeon and IPF side in different roles including architect, micro-architect, logic design engineer, pre- and post-silicon validation engineer.

Sujal Vora received the B.E. degree in electrical engineering from the D.D. Institute of Technology, India, in 1998, and the M.S degree in electrical engineering from Wright State University, Dayton, OH, in 2000. He has been with Intel Corporation, Santa Clara, CA, since 2000. He has worked on low power library design, power distribution and clock distribution for Xeon Processors. He is currently working on post-Si power/thermal validation for the Xeon processor and driving synergy on post-silicon power/thermal methodology within the Enterprise Processor Division of Intel.