Power Reduction Techniques for an 8-core Xeon® Processor

Stefan Rusu, Simon Tam, Harry Muljono, Jason Stinson, David Ayers, Jonathan Chang,
Raj Varada, Matt Ratta, Sailesh Kottapalli, Sujal Vora
Intel Corporation
Santa Clara, CA, USA
stefan.rusu@intel.com

***Abstract* — This paper presents the power reduction and management techniques for the 45nm, 8-core Nehalem-EX processor. 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. Clock and power gating minimize power consumed by disabled blocks. 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 un-terminated I/O links and shedding phases in the voltage regulator to improve the power conversion efficiency.**

I. INTRODUCTION

Power efficiency is a key design goal for this 45nm Xeon Processor with 8 cores and 16 threads. Total thermal design power remains at 130W, in line with previous processor generations (as shown in Figure 1), even though the number of cores per socket increased by two additional cores every year (as described in Figure 2). The processor has 2.3B transistors and is implemented in a 45nm CMOS process technology using high-K metal gate dielectric transistors and nine copper interconnect layers [1].

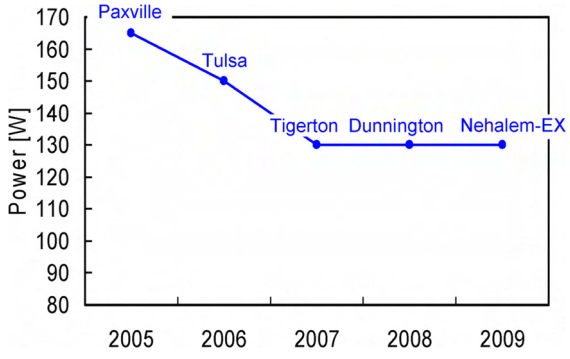

Figure 1 - Power dissipation trends for several generations of Xeon®Processors

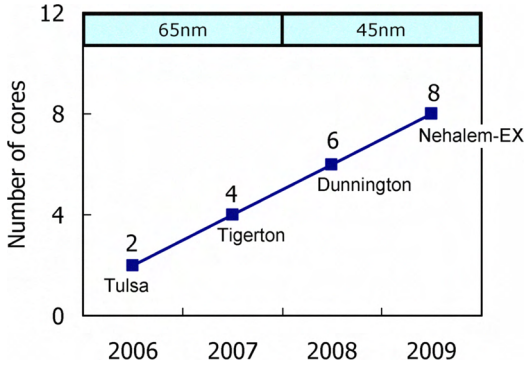

Figure 2 – Core count trend for several generations of Xeon®Processors

The 45nm high-K metal gate process technology reduces the gate leakage by a factor of 25x for NMOS devices and 1000X for PMOS transistors, compared to 65nm process generation. The processor is packaged in a 14-layer (5-4-5), 1567 lands, 40mil pitch organic land grid array package with an integrated heat spreader. The processor supports multiple platform configurations, from dual processor options to quad and even 8-socket options. Figure 3 shows the processor die photo with the major blocks marked out [2].

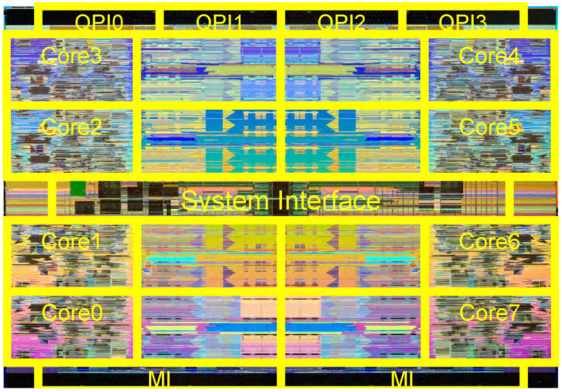

Figure 3 – Processor Die Photo

978-1-4244-4353-6/09/$25.00 ©2009 IEEE

The shared L3 cache is split in eight slices. Even though each cache slice is aligned with a processor core, the entire L3 cache is seen as one large, shared cache by all cores. The top side of the floorplan includes four point-to-point Quick Path Interconnect links (QPI) running at 6.4GT/s, while the bottom side houses the memory interface. 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 and the DFT control box. The uncore clock generator and fuse box also reside in the center channel area.

II. Power Reduction Techniques

The processor has four voltage domains shown in Figure 4 with level shifters used across domain boundaries. Table I lists the operating voltage ranges. To reduce both switching and leakage power, we strive to operate each domain at the lowest possible voltage level.

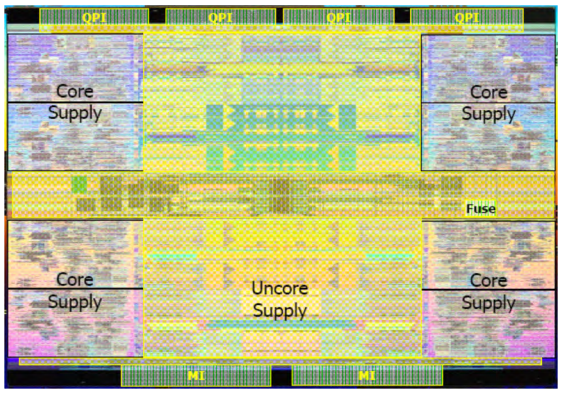

Figure 4 – Processor voltage domains

The processor uncore operates at a fixed voltage, typically 0.9V, although some slow parts may be binned as high as 1.1V. The core voltage is variable from 1.1V going as low as 0.85V. The highest voltage is used for the turbo mode, while the lowest voltage is used when all cores are active. The I/O links operate at 1.1V to enable a solid eye opening. The PLL clean supply uses a 1.8V supply from the motherboard which is passed through an on-die filter to power the multiple PLLs and thermal sensors.

To minimize leakage power, long channel devices are used in non-critical circuit paths. These devices trade off speed for lower leakage: about 10% lower speed provides a 3x leakage reduction. The processor cores include 58% long channel devices, while the uncore logic uses 85% low-leakage transistors. Overall, leakage accounts for about 16% of the total power at the typical process corner and nominal voltage.

TABLE I. Operating Voltages and Power Breakdown

Domain	Voltage	Power
Core supply	0.85 – 1.1 V variable	71 W
Uncore supply	0.90 – 1.1 V fixed	44W
I/O supply	1.1 V fixed	14 W
PLL clean supply	1.8 V fixed	1W

All circuits use static CMOS logic to minimize the active switching power. Domino circuits are used only in the read path of large register files.

The L3 cache implements both sleep and shut-off modes, that reduce leakage by 35% and 83% respectively [3], compared to the active mode as shown in Figure 5. In active state, the large pull-up is turned on and connects the virtual array supply to the real VCC. In sleep mode, the large pull-up is turned off and the small multi-leg programmable pull-up acts as a resistor that drops the supply voltage into the sleep state. In shut-off mode all pull-up devices are turned off and the virtual VCC drops to a low voltage determined by the residual leakage of the array. In shut-off state the array is functionally disabled and the logic state of the array bits is lost.

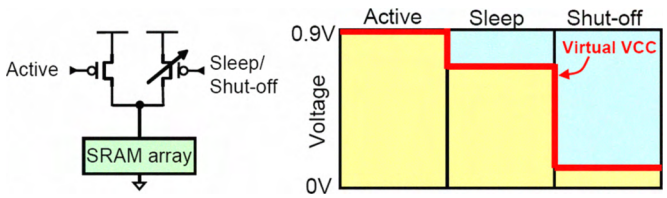

Figure 5 – Cache sleep and shut-off circuit implementation and voltage diagram

Each core is equipped with on-die power gates that cut the leakage in average by 40X in the shut-off state. The power gate implementation details are shown in Figure 6. A thick M9 metal layer is used to minimize voltage droop in on-die power distribution [4].

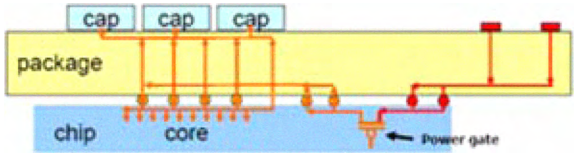

Figure 6 – Power gate implementation details [4]

III. Core and Cache Recovery

The core and cache recovery is a yield improvement technique made possible by having multiple identical instances of cores and cache on a single die. If one of the cores has a manufacturing defect, we disable the core and recover that part as a lower line item. The same applies for a massive defect in the cache which cannot be repaired using the built-in cache redundancy. The core and cache slices are disabled in horizontal pairs, but disabled cache and core slices do not have to be aligned, as illustrated in the example shown in Figure 7. There are multiple combinations possible that cover a majority of the die area and enable a good recovery. To enable testing all eight cores in parallel, the chip is reconfigured such that each core sees the aligned cache slice as its own dedicated cache. This way all eight cores execute the test patterns simultaneously and the tester can make a quick decision on which cores are good and what needs to be disabled. On-die electrically programmable one-time fuses are used for the core and cache recovery. The initial die triage and fusing is

performed at wafer sort and confirmed again during the final test.

The disabled cores are clock and power gated to avoid burning any power in the customer's system on a functional block that is logically disabled and does not deliver any value to the user. Disabled cache slices are also clock gated and placed in the shut-off state described earlier in Section II.

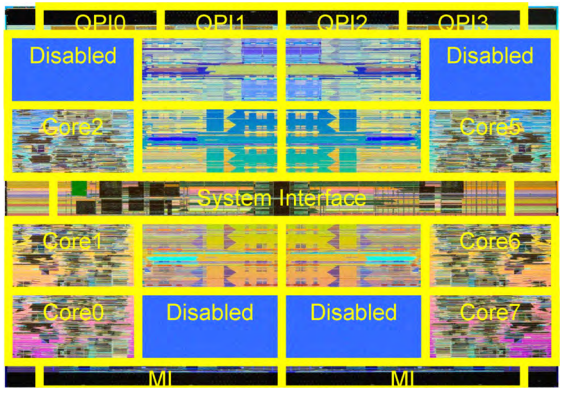

Figure 7 – Core and cache recovery example

Figure 8 shows the backside infrared emission from a die with only six functional cores. Brighter shades of gray indicate higher switching and leakage power that causes increased photon emission. The two cores that are disabled show dark in the picture, since there is no infrared emission coming from them. The one bright spot in each core is due to the thermal sensor, which is powered by the clean PLL voltage domain and is not affected by the core-level power gates.

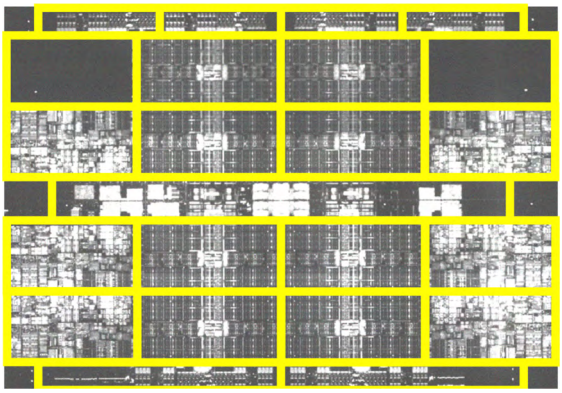

Figure 8 – Backside infrared image of a die with the top two cores disabled

IV. Power Management

Figure 9 shows the block diagram of the Power Control Unit (PCU). The PCU contains a micro-controller and several state machines that control the core voltage regulator output and the core power gates, manage the transitions between the different power states and control the detection and response to thermal events. The PCU receives the output of core-level voltage and temperature sensors, as well as the desired power state for each core. 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 or operating voltage 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.

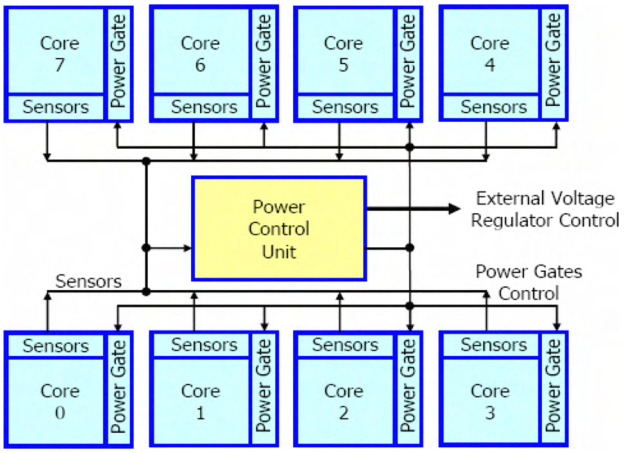

Figure 9 – Power Control Unit connectivity to the processor cores

V. Idle Power Reduction

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 I/O links un-terminated 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 the link detect circuit shown in Figure 10 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 has a much lower value than the pull-up resistor on the driving die).

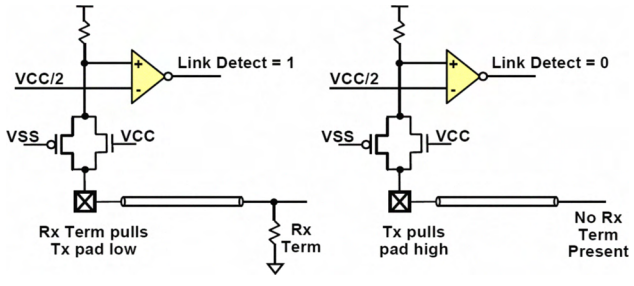

Figure 10 – Link detect circuit details: terminated link on the left side, open link on the right side

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 un-terminated link. This saves about 2W 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 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 Figure 11, with the efficiency dropping fast below 15A. 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 Figure 11. 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 2W per socket.

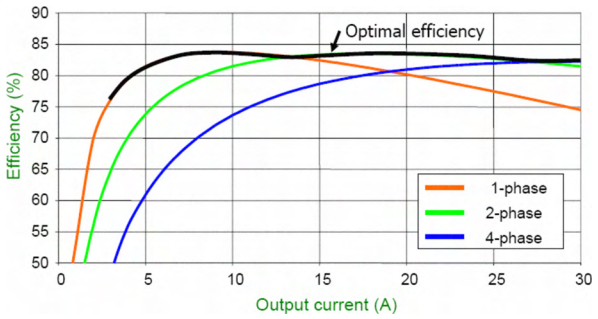

Figure 11 – Voltage regulator efficiency for different number of phases

VI. SUMMARY

This paper reviewed the power reduction and management techniques for the 45nm 8-core Xeon® processor. 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 technique improve manufacturing yields and enable multiple product flavors from the same silicon die. Disabled core and cache blocks do not burn any power in the customers' systems. 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 un-terminated I/O links, while the voltage regulator phase shedding technique improves the power 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

[1] K. Mistry, et al., "A 45nm Logic Technology with High-k+ Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging", IEDM Tech. Digest, December 2007

[2] S. Rusu, et al., "A 45nm 8-Core Enterprise Xeon® Processor", ISSCC Tech. Digest, February 2009

[3] F. Hamzaoglu, et al., "A 153Mb-SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm High-K Metal-Gate CMOS Technology ", ISSCC Tech. Digest, February 2008

[4] R. Kumar, G. Hinton, "Nehalem: a family of 45 nm IA processors", ISSCC Tech. Digest, February 2009