

3.1 A 45nm 8-Core Enterprise Xeon® Processor

Stefan Rusu, Simon Tam, Harry Muljono, Jason Stinson, David Ayers, Jonathan Chang, Raj Varada, Matt Ratta, Sailesh Kottapalli

Intel, Santa Clara, CA

The next-generation enterprise Xeon® server processor consists of eight dual-threaded 64b Nehalem cores and a shared L3 cache. The system interface includes two on-chip memory controllers and supports multiple system topologies. Figure 3.1.1 shows the processor block diagram. This design has 2.3B transistors and is implemented in 45nm CMOS using metal-gate high- κ dielectric transistors and nine Cu interconnect layers [1]. The thermal design power is 130W.

The L3 cache is built from eight slices, each having 2048 sets and 24 ways. Each cache line is 64 bytes constructed into 2 chunks. There are a total of 48 sub-arrays in each slice [2], as shown in Fig. 3.1.2. To reduce the L3-cache-slice active power, only 3.125% of the arrays are powered up for each access. The data array uses a $0.3816\mu\text{m}^2$ cell and is protected by inline double-error-correction and triple-error-detection (DECTED) ECC scheme with variable latency. Tag arrays are built with a $0.54\mu\text{m}^2$ cell and are protected by inline single-error-correction and double-error detection (SECDED) ECC scheme with fixed latency. Data array has both column and row redundancy, while the tag arrays have only column redundancy. The L3-cache-redundancy fuses are stored in an on-package serial EEPROM and their values are shifted into on-die radiation-hardened storage elements during reset. If any defect occurs in the non-repairable area (such as tag data-path) of a cache slice, the die can still be recovered by disabling the defective slice. Similarly, if a defect falls onto a processor core, the die is recovered by de-featuring the defective core.

Figure 3.1.3 shows an example of this core-and-cache recovery scheme. All disabled cores and cache slices are clock gated to reduce active power and are also placed in shut-off mode to minimize leakage. For the data array, SRAM cells and peripheral circuitry (sense amplifiers, column multiplexers and write drivers) and wordline drivers can be placed in the shut-off mode. For the tag arrays, a shut-off option is implemented in the SRAM cells and wordline drivers only. During shut-off, the power supply of SRAM drops from the 0.90V nominal operating voltage to 0.36V, providing an 83% leakage-power saving. By comparison, the leakage-power saving for the SRAM cells at the target sleep voltage (0.75V) is about 35%.

The processor clocking architecture includes 16 PLLs, 8 DLLs, and independent clock domains for each of the cores, the un-core including the cache and system interface, as well as the 6 independent I/O regions [3]. Figure 3.1.4 highlights the clock-distribution domains and the clock generators' location in the floorplan. The un-core PLL generates a unified un-core clock that serves the system-interface logic and the last-level cache. The processor cores and independent I/O ports are served by a dedicated PLL that enables independent operating frequencies for each region. The filter-PLL interfaces to the system clock buffer (BCLK) at 133MHz to generate the low-noise on-die reference clocks for all the PLLs. The PLLs are distributed in a balanced fashion to ensure high synchronicity between the independent clock domains at the common BCLK reference edges even when the domains are operating asynchronously at different frequencies. To balance low clock power consumption with good skew performance, the un-core clock distribution relies on vertical and horizontal clock spines with embedded clock compensators controlled by the on-die compensation fuses and state machine to a mixed point-to-point and multiple independent grids for the un-core clock distribution. The post-layout simulated skew across all the local clock receivers is under 19ps without any clock compensation engaged. A zone-to-zone skew-budgeting methodology is employed in the un-core timing-verification flow to achieve fast design convergence without compromising design margin.

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 power to the PLLs and on-die thermal sensors. Level shifters are used between voltage domains. The design

uses longer-channel devices in non-timing-critical paths to reduce the sub-threshold leakage. About 58% of the transistor width in the cores and 85% of the transistor width in un-core (excluding cache arrays) are long-Le. Overall, leakage accounts for about 16% of the total power at the typical process corner.

The processor is flip-chip (C4) attached to a 14-layer (5-4-5), 40mil pitch organic land-grid-array package with an integrated heat spreader. A rectangular land-side cavity matching the die outline contains decoupling caps for cores, un-core and I/O links. The chip-level power distribution consists of a uniform M9-M8 grid aligned with the C4 power and ground bump array.

All I/O links can run asynchronously at 6.4GT/s, providing up to 25.6GB/s/port. Each link supports data and clock failover RAS features, as well as half-width and quarter-width lanes. The link transmitter requires the output swing to meet a minimum voltage requirement to guarantee minimum eye-opening at the receiver side. The TX also needs to meet a maximum output swing requirement to save power and reduce EMI and cross-talk. To meet these requirements, a precise PVT-compensation scheme controls 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 uses the receiver (RX) equalization (EQ) architecture based on continuous-time linear-equalization (CTLE) design with complete offset cancellation to mitigate the ISI associated with high-speed switching. The CTLE consists of a linear amp with gain greater than 1 to amplify the incoming signal along with adjustable R and C components to control AC/DC gain, as shown on Figure 3.1.5. The datapath 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 stay high.

The high-speed link interface requires an accurate reference current (I_{ref}) to bias the current-mode analog circuits. This is accomplished by compensating each I_{ref} through a compensation loop that compensates for process variations and continuously tracks for voltage and temperature (VT) variations. Each I_{ref} has its own local IDAC and a digital counter that controls the amount of current to be added or subtracted in the local IDAC. In addition, there is a current multiplier, external discrete resistor and comparator shared by each current output, as shown in Figure 3.1.6. First, one current output (I_{ref0}) is used to compensate the current source (Global IDAC). The I_{ref0} is multiplied through the current multiplier, converted to a voltage through the resistor R_{ext} and compared to a reference voltage V_{ref} . The output of the comparator provides direction to the digital counter that adjusts the Global IDAC accordingly. Once this initial loop is complete, the subsequent compensation loops cycle through each individual copy of the current outputs (I_{ref1} through I_{refn}). Each Local IDAC adjusts the current to match I_{ref0} . Once all outputs are compensated, all copies are delivered to the current mirrors and the initial loop is re-selected to enable VT tracking through I_{ref0} . I_{ref0} adjustments as the result of VT tracking directly adjust I_{ref1} through I_{refn} as well, resulting in all current outputs tracking VT variations.

DFT and debug features include scan, observability registers (scan-out), I/O loopback and I/O test generator (IBIST), on-die clock shrink, within-die process monitors and multiple TAP controllers. Cache DFT features include built-in pattern generators for testing large arrays (PBIST), stability test mode and low-yield analysis support.

Acknowledgments:

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

References:

- [1] K. Mistry, C. Allten, C. Auth, et al., "A 45nm Logic Technology with High- κ Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging", *IEDM Dig. Tech. Papers*, pp. 247-250, Dec. 2007.
- [2] F. Hamzaoglu, K. Zhang, Y. Wang, et al., "A 153Mb-SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm High- κ Metal-Gate CMOS Technology", *ISSCC Dig. Tech. Papers*, pp. 376-377, Feb. 2008.
- [3] N. Kurd, J. Douglas, P. Mosalikanti, et al., "Next generation Intel® micro-architecture (Nehalem) clocking architecture", *VLSI Circuits Symposium*, pp. 62-63, Jun. 2008.

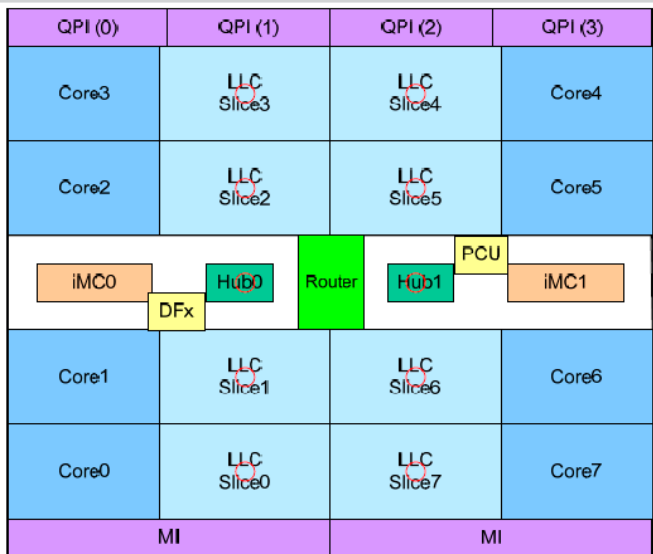


Figure 3.1.1: Processor block diagram.

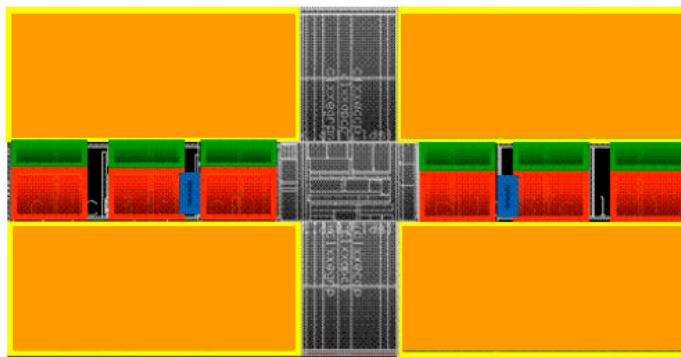


Figure 3.1.2: L3 cache slice.

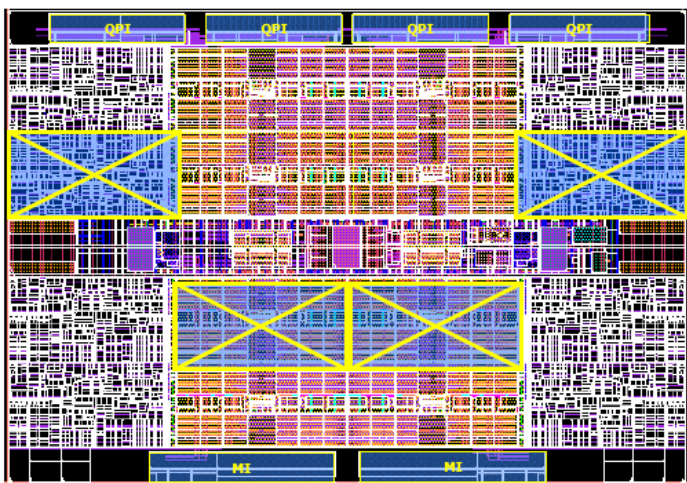


Figure 3.1.3: Core and cache recovery example.

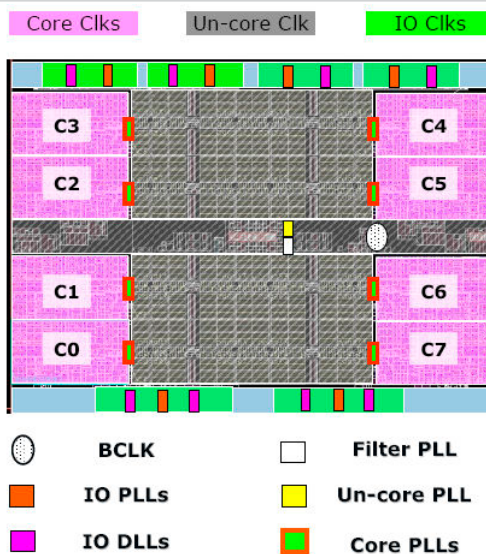


Figure 3.1.4: Clock distribution domains and generators.

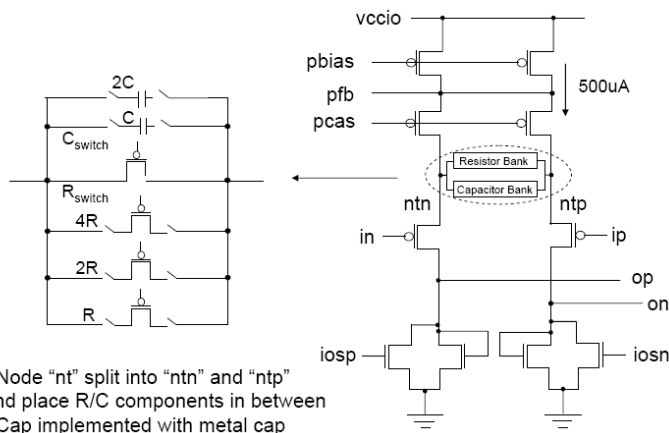


Figure 3.1.5: CTLE amp and R/C bank.

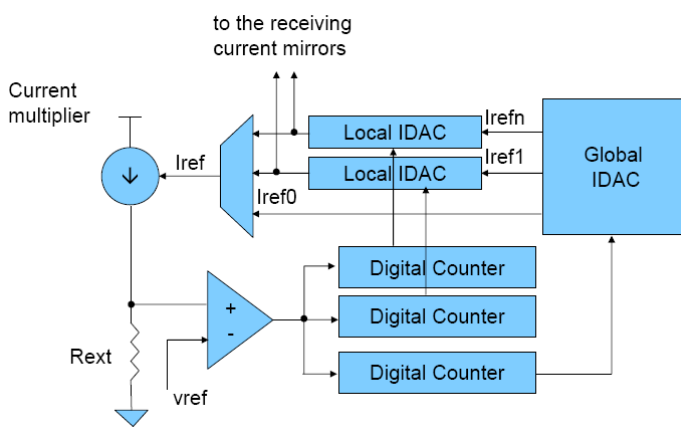


Figure 3.1.6: Reference-current compensation circuit.