5.3 A Dual-Core Multi-Threaded Xeon® Processor with 16MB L3 Cache

Stefan Rusu, Simon Tam, Harry Muljono, David Ayers, Jonathan Chang

Intel, Santa Clara, CA

This Xeon® MP processor consists of two 64b cores and a 16MB unified L3 cache. Each core has two threads and a unified 1MB L2 cache. Figure 5.3.1 shows the block diagram. The existing core to front-side bus (FSB) connection is replaced with a simple direct interface to minimize the L3 cache and external bus latencies. The caching FSB controller handles the core arbitration, L3 cache accesses, and external bus requests.

The 435mm² die has 1.328B transistors. The processor operates at more than 3.0GHz from a 1.25V core supply. The worst-case power dissipation is 165W while the power dissipation on a typical server workload is 110W. The processor is implemented in a 65nm process technology with eight copper interconnect layers and low-k carbon-doped oxide (k=2.9) inter-level dielectric. Figure 5.3.2 summarizes the process characteristics [1].

The 16MB L3 cache is built using 256 data sub-arrays (64kB each) and 32 redundancy sub-arrays (68kB each). Each data sub-array stores 32 bits while the redundancy ones store 34 bits. The 6T memory-cell bit size is 0.624µm² [2] and the physical address is 40b wide. To reduce the L3 cache active power, only 0.8% of all array blocks are powered up for each cache access. To reduce the L3 cache leakage, NMOS sleep transistors are implemented in the SRAM sub-arrays and PMOS power gating devices in the cache periphery. Figure 5.3.3 shows the sub-array sleep transistor circuit and its three operating modes. In *active* mode, the virtual V_{SS} is shorted to the real V_{SS} to provide the full voltage swing for array read or write. In *sleep* mode, the virtual V_{SS} floats up to about 250mV to reduce the leakage by about 2X while preserving the logic state of the sub-array. In *shut-off* mode, the NMOS shut-off device is turned off and causes the virtual V_{SS} to float up to about half V_{CC}, reducing the leakage an additional 2X beyond the sleep state. Processors that ship with only half the L3 cache enabled (8MB) use this shut-off mode to suppress leakage in the disabled cache section, saving about 3W of leakage.

Figure 5.3.4 shows the clock distribution map. Separate PLLs and clock distribution trees drive each core and the associated L2 cache. A third PLL drives the uncore half-frequency clock. The FSB uses the external bus clock (200MHz) and the quad-pumped version (800MHz). The three PLLs are grouped together on the left side of the die and the differential clock input is routed to three pairs of C4 bumps inside the package. The uncore clock is distributed through a balanced tree embedded in nine vertical spines. De-skew circuits controlled by on-die fuses [3] reduce the uncore clock skew to less than 11ps. To ensure that the uncore logic is not in the full chip critical timing path, a 5% margin is added to the uncore timing-verification flow.

The processor uses three voltage supplies: one for the two cores, a separate supply for the L3 cache together with the associated control logic, and the third one for the FSB I/O circuits, as shown in Fig. 5.3.5(a). Level shifters are used between voltage domains. A custom tool is developed to check for the presence and correct connectivity of level shifters on all signals that cross voltage-domain boundaries. Figure 5.3.5(b) shows the power breakdown across the major blocks, with three quarters of the power consumed in the two cores. The design uses longer Le devices (10% longer than nominal) in non-timing-critical paths to reduce sub-threshold leakage. About 54% of the transistor width in the cores and 76% of the transistor width in the uncore (excluding cache arrays) are long-Le. Overall, leakage accounts for about 30% of the total power at the typical process corner.

The processor is flip-chip (C4) attached to a 12-layer (4-4-4) organic package with an integrated heat spreader. The package has 604 pins, out of which 238 are signal pins and the rest are power and ground. The processor die has 13164 C4 solder bumps arrayed with a single uniform bump-pitch. The chip-level power distribution consists of a uniform M8-M7 grid synchronized with the C4 power and ground bump array.

The 3-load multi-processor FSB operates at 800MT/s. Figure 5.3.6 shows the symmetric pre-driver design used to control the edge rate to meet timing and signal integrity requirements. This is accomplished by dividing the FSB output swing (from V_{OL} to V_{OH}) into six voltage levels, each driven by an output driver segment with different R_{ON} value. When a segment is enabled, it forms a parallel resistance to the previously enabled segments to generate a new voltage level, thus creating a stair-case-like waveform in every transition. Each segment is triggered by a DLL-controlled delay line that is PVT compensated. Equal rise and fall times are established by reversing the order of the enabling sequence on the driver segments.

To prevent multiple bit errors caused by a single upset event in the same cache line, all caches use bit interleaving for adjacent cache lines in the physical layout. Both L3 data and tag arrays have ECC protection. The L2 data array also has ECC protection while the L2 tag has parity checking. A dynamic 32-entries cache-line disable mechanism protects the L3 cache from erratic bits and infant mortality failures. There are three diodes available for temperature sensing, one located in each core and one located between the two cores. The diodes in the cores are routed to an on-package temperature-monitor chip that provides temperature data to the system for fan speed control. The central diode is routed to pins for system use. There is also a temperature sensor located near the hot spot in each core that provides a digital temperature readout. The temperature values from these sensors are used in conjunction with operating-system power-state requests to make informed throttle and boost decisions. The sensors are calibrated on a per core basis to ensure accuracy.

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 three TAP controllers (one for each core and one for the uncore). Cache DFT features include built-in pattern generators for testing large arrays (PBIST), programmable weak-write test mode, and low-yield analysis support.

Figure 5.3.7 is the die micrograph with the major blocks highlighted.

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

References:
[1] P. Bai, et al., "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57µm² SRAM Cell," *IEDM Technical Digest*, pp. 657-660, 2004.
[2] K. Zhang, et al., "SRAM Design on 65nm CMOS Technology with Integrated Leakage Reduction Scheme," *Symp. VLSI Circuits*, pp. 294-295, Jun., 2004.
[3] S. Tam, et al., "Clock Generation and Distribution for the Third Generation Itanium® Processor," *Symp. VLSI Circuits*, pp. 9-12, Jun., 2003.

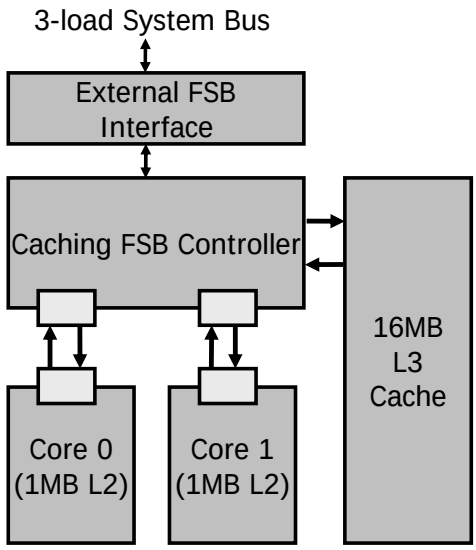

Figure 5.3.1: Block diagram.

Attribute	Value
Contacted Gate Pitch	220nm
M1 pitch	210nm
M2 pitch	210nm
M3 pitch	220nm
M4 pitch	280nm
M5 pitch	330nm
M6 pitch	480nm
M7 pitch	720nm
M8 pitch	1080nm
Dielectric	CDO, K=2.9
Memory cell	0.624μm^2

Figure 5.3.2: 65nm process technology summary.

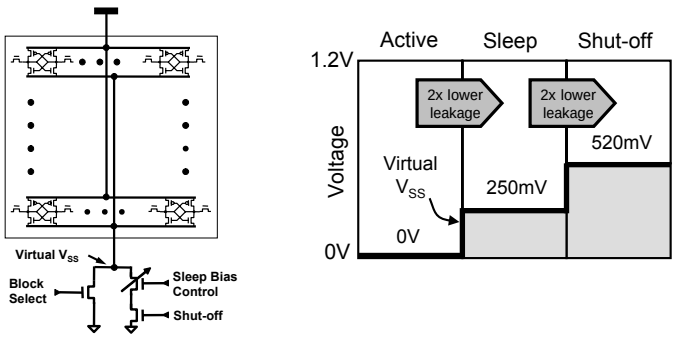

Figure 5.3.3: L3 cache sleep circuit and shut-off mode.

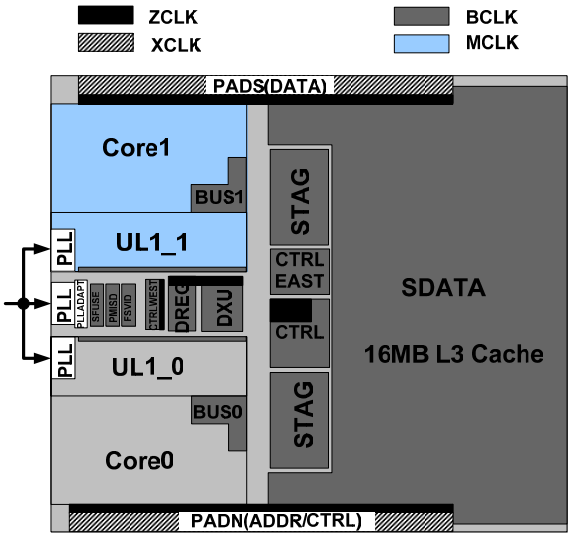

Figure 5.3.4: Clock distribution map.

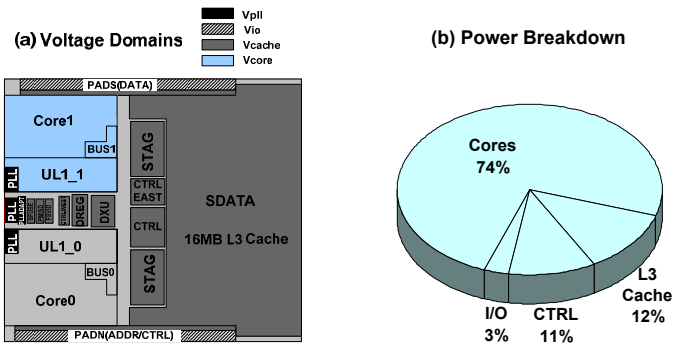

Figure 5.3.5: Voltage domains and power breakdown.

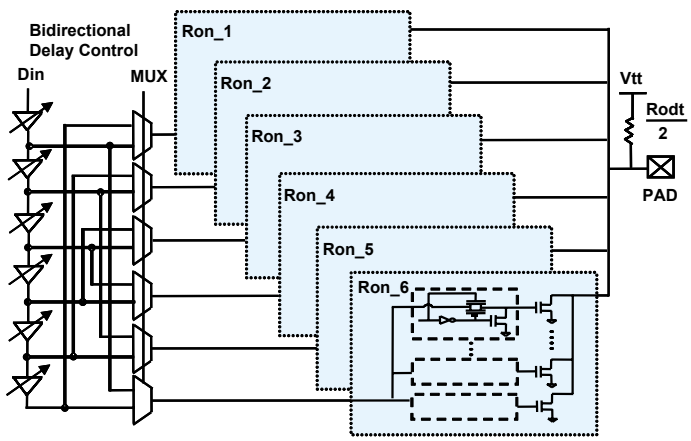

Figure 5.3.6: Symmetric I/O pre-driver circuit.

Continued on Page 641

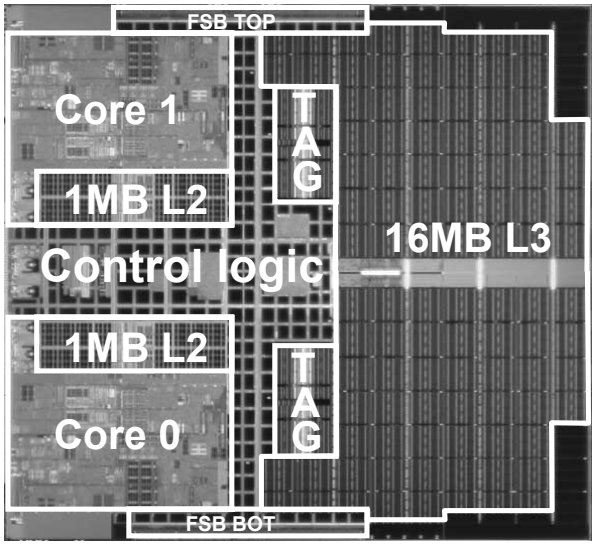

Figure 5.3.7: Die micrograph.

