13.1 A Sub-1W to 2W Low-Power IA Processor for Mobile Internet Devices and Ultra-Mobile PCs in 45nm Hi-κ Metal Gate CMOS

Gianfranco Gerosa, Steve Curtis, Mike D'Addeo, Bo Jiang, Belliappa Kuttanna, Feroze Merchant, Binta Patel, Mohammed Taufique, Haytham Samarchi

Intel, Austin, TX

This paper describes a low-power Intel® Architecture (IA) processor specifically designed for Mobile Internet Devices (MID) and Ultra-Mobile PCs (UMPC) where average power consumed is in the order of a few hundred mW (as measured by MobileMark'05 OP @ 60 nits brightness) with performance similar to mainstream Ultra-Mobile PCs. The design consists of an in-order pipeline capable of issuing 2 instructions per cycle supporting 2 threads, 32KB instruction and 24KB data L1 caches, independent integer and floating point execution units, ×86 front end execution unit, a 512KB L2 cache and a 533 MT/s dual-mode (GTL and CMOS) front-side-bus (FSB); a block diagram is shown in Fig. 13.1.1. The design contains 47M transistors in a die size under 25mm^2 manufactured in a 9-metal 45nm CMOS process with optimized transistors for low leakage [1] packaged in a Halide-Free 441 ball, 14×13mm^2 μFCBGA. Thermal Design Power (TDP) consumption is measured at 2W using a synthetic power-virus test at a frequency of 2GHz.

Features in this new micro-architecture are selected with low power and high performance per watt efficiency in mind. The pipeline is tailored to execute IA instructions as single atomic operations consisting of a single destination register and up to three source-registers and adheres to the Load-Op-Store instruction format. Further, using power efficient algorithms in areas like instruction decoding and scheduling (traditionally complex circuits that are power hungry) achieves high performance per watt efficiency. In addition, support for Hyper-threading technology (HT) is added; the instruction scheduling logic can find a pair of instructions from either the same thread or across threads in a given cycle to dispatch. HT is a feature that provides high performance per watt (typically 30% increases in performance for a 15% increase in power) efficiency in an in-order pipeline. Moreover, the use of specialized execution units is minimized. For example, the SIMD integer multiplier and Floating Point divider are used to execute instructions that would normally require a dedicated scalar integer multiplier and integer divider respectively. Finally, other features like activity-based control of instruction issue and dispatch of operations on the Front Side Bus are added for power reduction.

Several features support extended battery life including the new Intel® Deep Power Down Technology [2], which allows for a majority of the CPU functionality to be powered down except for an array that holds the micro-architectural state with very fast entry/exit times (<100us). Other micro-architecture features include Intel® 64 Architecture support, Intel® Digital Media Boost (SSE3), and Intel® Virtualization technology.

This design uses a "sea-of-Functional-Unit-Blocks (FUB)" methodology whereby all cluster hierarchies as well as all unit-level hierarchies are flattened at the chip-level even though the logical partitioning of the chip-level RTL model is comprised of several logical clusters (Floating Point, Integer Execution, Memory Execution, Front-End, Bus Interface, and L2 Cache). This methodology essentially removes all cluster and unit-level hierarchical boundaries resulting in a physical hierarchy where an object-based parallel editing scheme is used for physical design convergence; Fig. 13.1.7 shows the die photo with the logical unit partitions. The physical database consists of 205 unique FUBs (not including repeater stations) and 41K FUB-to-FUB interconnects. 91% of the FUBs are designed using cell-based design techniques using pre-characterized standard cells with 45% using "structured data-path" design techniques and 46% fully synthesized random logic blocks. The remaining 9% are full-custom blocks.

The clock distribution is architected with power saving as a top priority. Clock recombination in global distribution is limited to a few critical stages after carefully studying the power/skew trade off. Global clock routing is further reduced by: (1) implementing a grid-less topology that routes clock to locations only if required, (2) alignment of clock receivers to major clock spines, and (3) reducing the number of receivers on the clock network by minimizing clock cell cloning where clock tree synthesis is used. Flow automation and a cell-based design approach are used to minimize implementation effort and achieve fast clock convergence.

This processor uses a Register File (RF) design style for all core arrays including the 32KB L1 instruction cache, the 24KB L1 data cache and the 10.5KB C6 array which holds micro-Architecture state during the deep power down state (VCC=0.3V). These use 8T memory cells to attain better SER characteristics, high performance cache access time and lower voltage operation than traditional 6T SRAM cells. The L1 caches implement 1 bit per byte parity and no ECC. RF arrays comprise over 50% of the total core area and are an important contributor to overall chip power. Several power saving techniques are employed: fine granularity sleep on word-line drivers with double stacked PFETs with near zero wakeup times, sleep controls to allow floating of read bit-lines on the ROM arrays (22Kuops × 60 bits/uop) during non-access cycles, and fine granularity power gating on non-accessed array banks. In addition, array bits on the "read-side" are architecturally pre-disposed to a value of '0' to reduce leakage power through the read stack.

The L2 cache is a phase-based, 8-way 512KB design supporting single cycle throughput for both read and write operations in a 9-cycle pipeline including tag lookup, data read, in-line ECC and transmit to the L1. Figure 13.1.2 shows the L2 cache timing diagram. Tag, LRU and State bits are combined together into one array to minimize area and power. Tag and data sub-arrays are 4.5KB and 17.5KB, respectively with 256 cells (bit-cell area = 0.3816μm^2) per column and column redundancy for optimum array efficiency. Average power reduction techniques include: power-gating (PG) transistors for the word-line drivers, PG and sleep transistors (ST) for the memory arrays (see Fig. 13.1.3 and [3]), floating bit lines and tri-state-able write drivers. The ST implementation is based on PFETS, creating a virtual VCC which can go as low as 750mV reducing bit-cell leakage power up to 2.5×. The 512kB L2 cache is dynamically chop-able down to 2-ways. For less demanding applications, sub-arrays of the unused ways are powered down via PG and ST resulting in 10× leakage power reduction per sub-array.

A dual mode IO buffer is implemented where both legacy Gunning-Transistor-Logic (GTL) signaling and a full CMOS swing can be supported with a fuse-able option. In CMOS mode, the buffer can reliably transmit data at 400 to 533MT/s while reducing the total Front-Side-Bus (FSB) power up to 2.5× (data-dependent) as compared to GTL in an ISO-slew rate comparison. Essentially the resistance-compensated NFET pull-down impedance is reprogrammed to 55Ω and the on-die-termination (using R-compensated 55Ω PFETs) is turned off to eliminate DC power. Figure 13.1.4 shows the dual mode buffer and receiver. IO leakage power is further reduced by "splitting" the 1.05V IO power supply (VCCP) and only keeping 21 pins "alive' during the deep power down state resulting in ~10% reduction of average power; Fig. 13.1.5 shows a platform implementation.

An IA microprocessor specifically micro-architected for low power and implemented in 45nm CMOS technology is presented. This processor is suitable for MIDs and UMPCs given the range (~0.6W to 2.0W) of measured TDP power on real world applications. 2 GHz core frequencies are achieved at 1.0V and 90°C; Fig. 13.1.6 shows measured power at different power states and a shmoo of the processor running several worst-case workloads.

Acknowledgments:
The authors gratefully acknowledge the support of project manager Elenora Yoeli and the work of the talented and dedicated Intel design and product teams in Austin, Texas 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", *Tech. Dig. IEDM*, Dec. 2007.
[2] V. George, et al., "PENRYN: 45-nm Next Generation Intel® Core™ 2 Processor", *accepted to ASSCC Dig. Tech. Papers*, Nov. 2007.
[3] F. Hamzaoglu, et al., "A 153Mb SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm Hi-K Metal Gate CMOS Technology", *ISSCC Dig. Tech. Papers*, pp. 376-377, Feb. 2008.

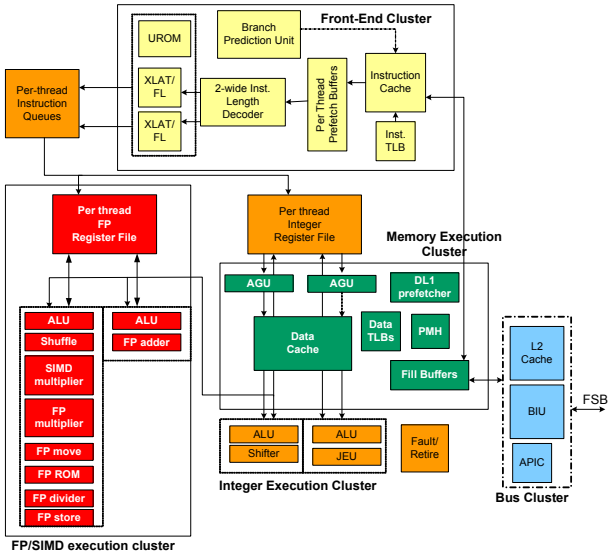

Figure 13.1.1: Low-power IA processor architecture block diagram.

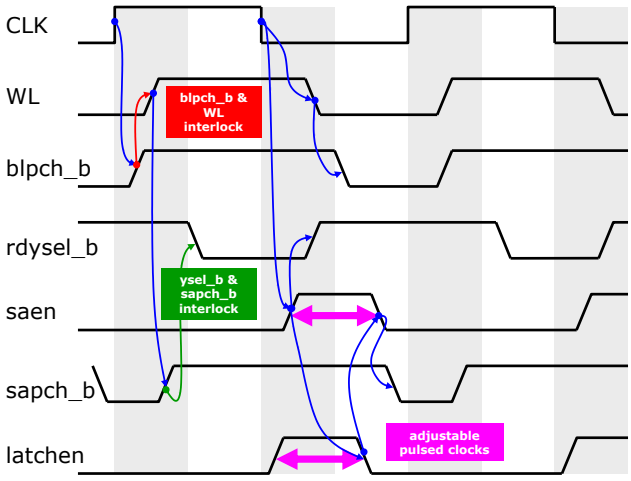

Figure 13.1.2: L2 timer phase-based timing diagram.

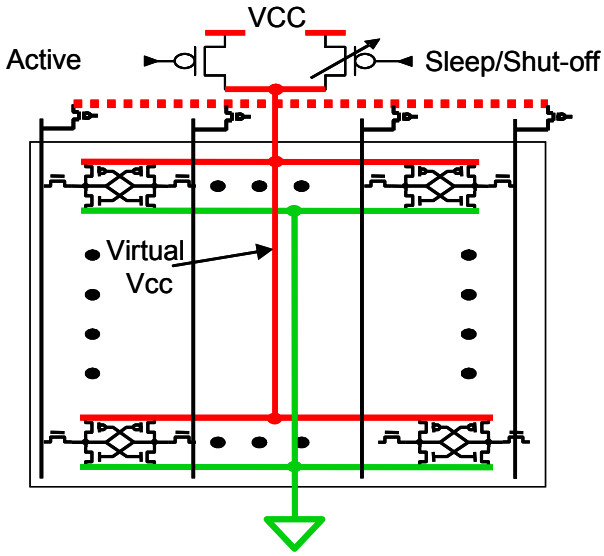

Figure 13.1.3: L2 cache sleep circuit and shut-off (Idle) mode.

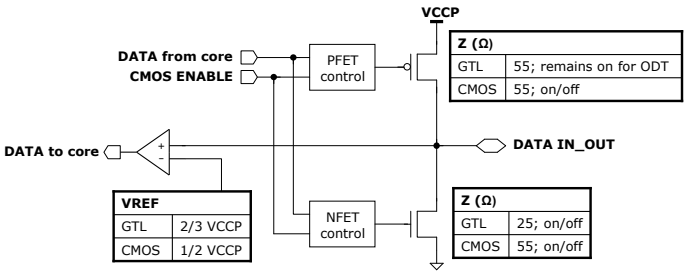

Figure 13.1.4: Dual-mode Front-Side-Bus IO driver.

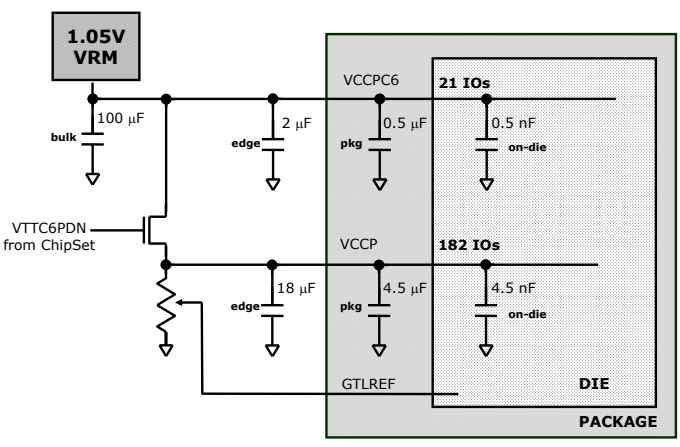

Figure 13.1.5: Split IO Power Supply to reduce IO leakage power in SLEEP state.

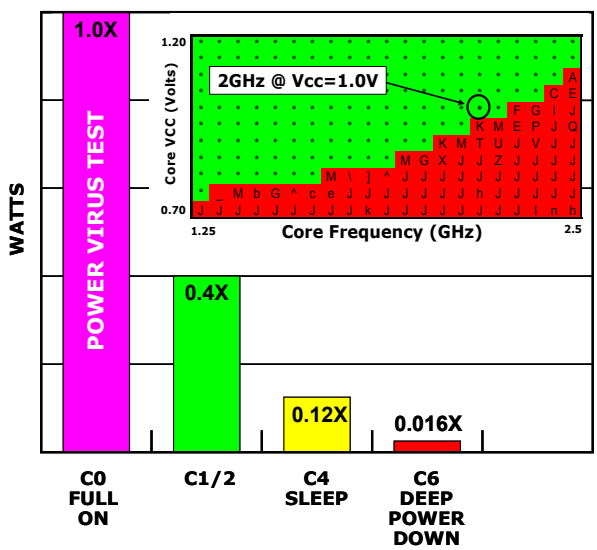

Figure 13.1.6: Power for several power states and VCC-Fmax shmoo @ 90°C.

Continued on Page 611

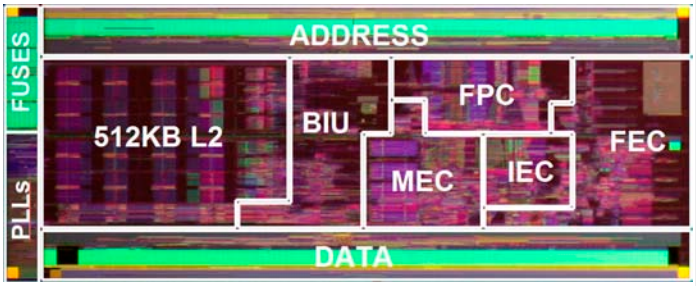

Figure 13.1.7: Low-power IA processor die micrograph.

 978-1-4244-2010-0/08/$25.00