

5.5 A Wire-Speed Power™ Processor: 2.3GHz 45nm SOI with 16 Cores and 64 Threads

Charles Johnson¹, David H. Allen², Jeff Brown², Steve Vanderwiel², Russ Hoover², Heather Achilles³, Chen-Yong Cher⁴, George A. May⁵, Hubertus Franke⁴, Jimi Xenedis⁴, Claude Basso⁶

¹IBM Research, Rochester, MN

²IBM Systems and Technology Group, Rochester, MN

³IBM Research, Bedford, NH

⁴IBM Research, Yorktown Heights, NY

⁵IBM Systems and Technology Group, Essex Junction, VT

⁶IBM Systems and Technology Group, Raleigh, NC

An emerging data-center market merges network and server attributes into a single wire-speed processor SoC. These processors are not network endpoints that consume data, but inline processors that filter or modify data and send it on. Wire-speed processors merge attributes from 1) network processors: many threaded low power cores, accelerators, integrated network and memory I/O, smaller memory line sizes and low total power, and from 2) server processors: full ISA cores, standard programming models, OS and hypervisor support, full virtualization and server RAS & infrastructure. Typical applications are edge-of-network processing, intelligent I/O devices in servers, network attached appliances, distributed computing, and streaming applications.

Figure 5.5.1 shows the chip, which consists of 4 throughput (AT) nodes, 2 memory controllers, 4 accelerators, 2 intelligent network interfaces, and a 4-chip coherent interconnect. The architecture supports upward and downward scalability. A low-end chip scales down to a single AT node (4-A2 64b cores, 2MB L2) and at the high end 4 full-function chips are fully interconnected. The first-generation chip provides a compute range of 4 cores/16 threads to 64 cores/256 threads. As Ethernet standards progress from 10Gb to 100Gb this architecture scales to achieve a 10× throughput gain using the same compute elements and interconnects.

A >50% reduction in chip power is due to architectural power-performance choices. Small highly threaded processors have 2 to 4× better throughput power-performance than a general purpose processor (GPP) but with reduced single-thread performance [1]. The number of execution units per mm² is 2 to 4× higher, simultaneous multi-threading leverages these under-utilized execution facilities [2]. Hardware accelerators are often 10× more power-performance efficient than processors for well-defined applications. The wire-speed workload has many standards-based functions for accelerators and packetized data for the highly threaded cores. Integrating highly threaded processors, accelerators and system I/O onto a single chip provides the high throughput, low system power solution for wire speed computing. Figure 5.5.2 takes a GPP and incrementally adds 1) accelerators, 2) integrated system I/O, and 3) highly threaded cores. In this morphing of a GPP to a wire speed processor, accelerators contribute to overall performance, integrated I/O and highly threaded processors contributed to a 50% system power reduction at constant performance. In addition to the architectural choices, the chip employs aggressive power reduction and power management techniques. These include static voltage scaling, multi-voltage design, multi-clocking, extensive clock gating, multiple threshold voltage devices, dynamic thermal control, eDRAM caches and circuit design to enable low voltage operation. In aggregate the implementation techniques reduce power by >50%, resulting in an SoC with a worst-case power of 65W and typical power of 55W at 2.0GHz at voltages as low as 0.75V (Fig 5.5.3).

Power is minimized by operating at the lowest voltage necessary to function at frequency. Variations in wafer fabrication electrical parameters result in a distribution of hardware performance and power, which allows customizing the main voltage (V_{DD}) for each chip. The minimum V_{DD} required to run at frequency is determined at test and stored in on-chip non-volatile memory, it is retrieved at power on and used to program off-chip voltage regulators. Every chip can operate at several power/frequency combinations and a unique voltage is stored for each. This adaptive power-supply (APS) technique reduces worst-case and average power across the process distribution. Using APS to minimize power requires enabling low-voltage operation, including identifying and redesigning circuits that have functional and timing problems at low voltages. All circuits on the V_{DD} APS are designed to operate at 0.7V or lower across the entire process distribution. This provides margin, as only the fastest hardware is actually operates at as low as 0.7V. The chip uses circuits that require specific voltages to operate at optimal power/performance efficiency. Multi-voltage design is a power-management technique that supplies the minimum voltage required by

each circuit [3]. Most of the logic uses APS V_{DD} , SRAMs use a higher APS V_{CS} . Fixed supplies V_{IO} , V_{MEM} , V_{TT} , and DV_{DD} are required for I/O; A_{VDD} is for the PLL and analog circuits; and V_{FS} is for fuse programming.

Clock distribution is implemented using grids suspended from balanced H-trees to minimize skew. Every 1ps reduction in clock skew saves ~0.4W. Multiple clock domains provide flexibility in the operating frequencies of the cores, accelerators, I/O, and other logic blocks. This flexibility allows the chip to meet the requirements of multiple applications with optimized power. The large number of asynchronous clock-domain crossings increases the chip latch-metastability failure rate (MFR). These latches were originally designed for soft-error rate (SER) reduction with triple-stack devices in the feedback inverter to reduce diffusion cross-section. SER was calculated to be orders of magnitude less than MFR, so the latches were redesigned (Fig 5.5.4) to remove the triple-stack to improve the MFR of these latches by >10×.

The chip extensively uses clock gating to reduce AC power. Clock buffers include controls to disable local clocks in groups of 4 to 32 latches. Clock gating is measured as *installed gating* (% of latches logically disabled), *busy gating* (% of latches gated off in operation), and *idle gating* (% of latches gated off when disabled). Targets for busy and idle gating were 70% and 90%, respectively. Achieved numbers varied by function, but installed clock gating numbers up to 95% were common, resulting in an AC power savings of up to 32% (40W). Gating of top-level clock grids is not implemented since physical separation would introduce skew between grids and would only reduce clock power by an additional 12% of the savings realized by local clock gating. In applications where an entire clocking domain can be shut down, a global signal disables all local clock buffers within that domain.

Balancing the delay requirements of timing paths with chip DC leakage budgets requires selective use of transistors' V_T . The technology supports multiple V_T levels but only three were used to limit cost: super-high (SVT), high (HVT), and nominal (RVT). Circuits known to have high switching factors (e.g., clock buffers) are implemented using RVT to minimize device width. Large blocks of random logic are synthesized using HVT, then minimum-sized devices in paths with positive slack are replaced with lower-leakage SVT. RVT is budgeted and used only in paths that could not otherwise make timing. Overall, the chip uses 75% SVT, 20% HVT, and 5% RVT devices.

Temperature monitors are placed in areas of high power density and feed the on-chip power limiter (PL) that reduces AC power if the die temperature exceeds a programmable level. By inserting wait cycles for new operations initiated on the bus (PBUS), the PL reduces switching activity and AC power. Power is reduced by as much as 24W using the PL, resulting in a die-temperature reduction of ~10C under normal system conditions.

Deep trench (DT) embedded DRAM (eDRAM) is used instead of SRAM in the four 2MB L2 caches. The 2MB L2 caches are implemented as 16 1Mb eDRAM macro instances, each is composed of four 292Kb sub-arrays (264 WL × 1200 BL). The eDRAM cell measures 152×221nm² (0.0672μm²). This eDRAM implementation allows substantially larger cache sizes (3×SRAM) with only ~20% of the AC and DC power [4]. Use of eDRAM provides a significant noise reduction benefit in the form of DT decoupling capacitance. DT provides at least a 25× improvement over thick-oxide dual-gate decoupling devices in capacitance per unit area. Coupled with a robust power distribution, DT decoupling reduced AC noise 30mV and reduced power more than 5W.

Double pumping a 2R1W cell provides 2R2W functionality for the 64 instances in the chip. A double-pumped register file (Fig. 5.5.5) is used to reduce area by 45% and power by 15%. Figure 5.5.6 illustrates power by unit at 2.0GHz. Power management and reduction techniques used on this chip reduced the overall power >50%, resulting in a power density over most of the chip of <0.25 W/mm².

References:

- [1] S. Balakrishnan, R. Rajwar, M. Upton, K. Lai, "The Impact of Performance Asymmetry in Emerging Multicore Architectures", *ISCA '05*, pp. 506-517, June 2005.
- [2] D.M. Tullsen, S.J. Eggers, H.M. Levy, "Simultaneous multithreading: maximizing on-chip parallelism", *ISCA '95*, pp. 392-403, June 22, 1995.
- [3] M. Keating, D. Flynn, R. Aiken, A. Gibbons, K. Shi, "Low Power Methodology Manual." Springer, 2007, p. 21-24.
- [4] J.Barth et al, "A 1 MB Cache Subsystem Prototype With 1.8 ns Embedded DRAMs in 45 nm SOI CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, Apr. 2009.

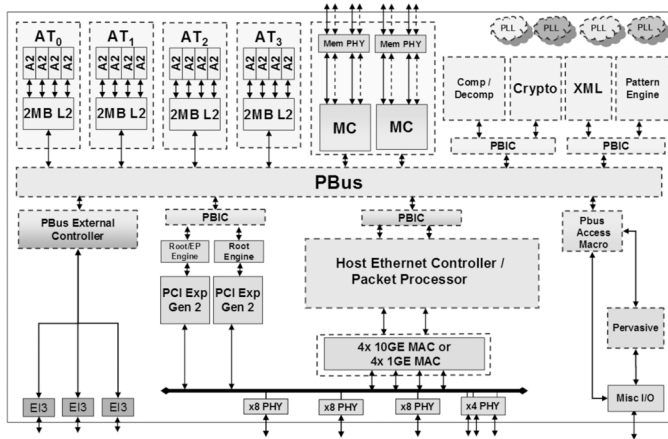


Figure 5.5.1: Wire Speed Processor Chip Diagram.

Technology	IBM 45nm SOI
Core Frequency	2.3GHz @ 0.97V
Chip Size	428mm ²
Chip Power (4-AT nodes)	65W @ 2.0GHz, 0.85V, Max Configuration
Chip Power (4-AT nodes)	20W @ 1.4GHz, 0.77V, Min Configuration
Main Supply Voltage (VDD)	0.7V to 1.1V
Metal Layers	11 Cu (3 1x pitch, 2 1.3x, 3 2x, 1 4x, 2 10x)
Latch Count	3.2M
A2 Cores	16
Threads	64
L1 I & D Cache	16 x (16KB + 16KB) SRAM
L2 Cache	4 x 2MB eDRAM
Hardware Accelerators	Crypto, Compression, RegX, XML
Intelligent Network Interfaces	Host Ethernet Adapter/Packet Processor 2 Modes: Endpoint & Network
Memory Bandwidth	2x DDR3 Controllers 4 Channels @ 800-1600MHz
System I/O Bandwidth	4x 10G Ethernet, 2x PCI Gen2
Chip-to-Chip Bandwidth	3 Links, 20GB/s per link
Chip Scaling	4 Chip SMP
Package	50mm FCPBGA (4 or 6 layers)

Figure 5.5.3: Key Technology and Chip Characteristics.

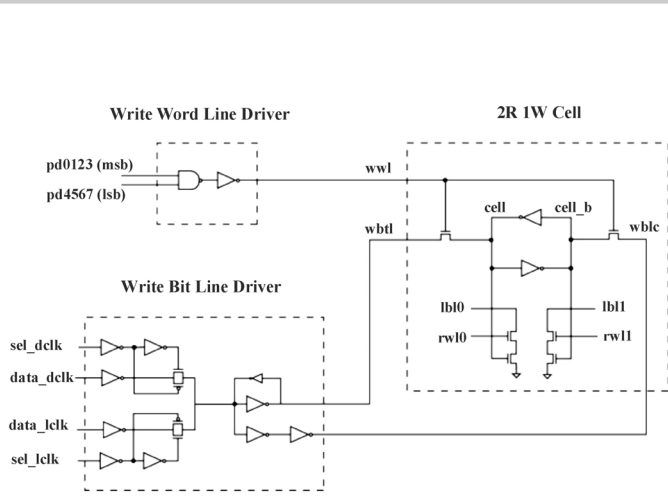
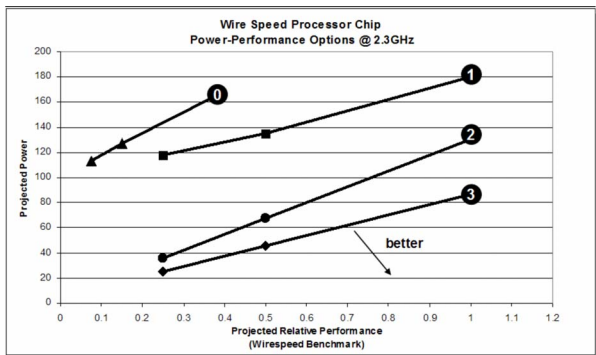


Figure 5.5.5: Double Pumped Register File.



- 0 : 8- GP cores [base General Purpose Processor chip]
 1 : 8- GP cores + 4 Accelerators
 2 : 8- GP cores + 4 Accelerators + Integrated I/O (DDR3, 10Ge, HEA)
 3 : 16- A2 cores + 4 Accelerators + Integrated I/O (DDR3, 10Ge, HEA) [WireSpeed chip]

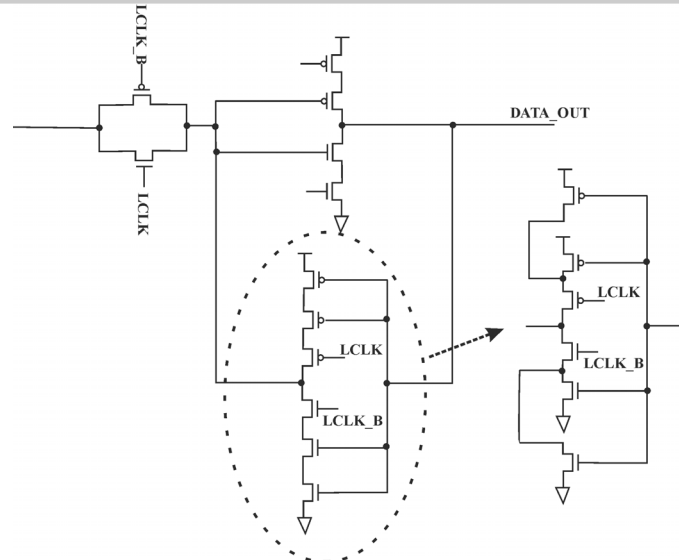


Figure 5.5.4: Latch for Clock Domain Crossings.

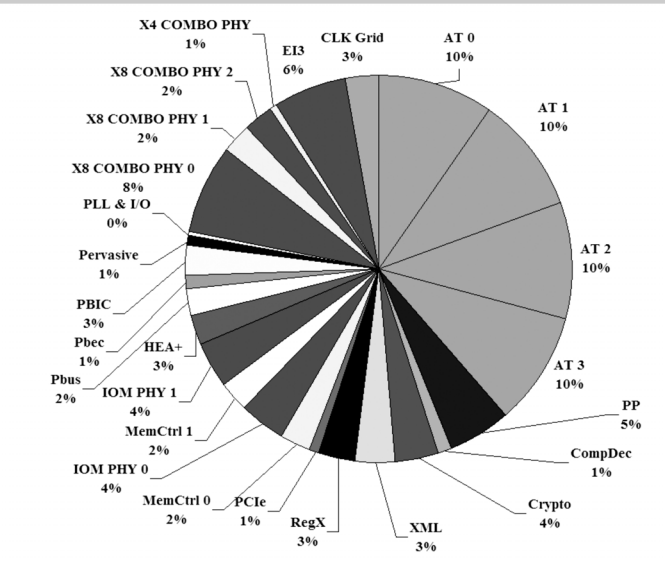


Figure 5.5.6: Power breakdown by function @ 2.0GHz.

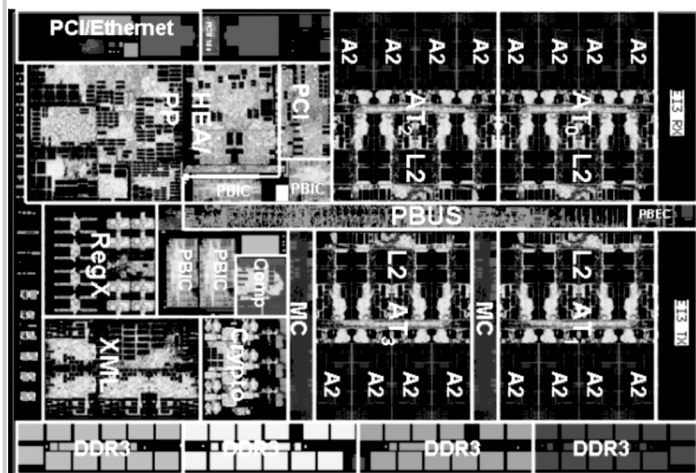


Figure 5.5.7: Wire Speed Die Photo.