2.4 POWER10™: A 16-Core SMT8 Server Processor with 2TB/s Off-Chip Bandwidth in 7nm Technology

Rahul M. Rao[1], Christopher Gonzalez[2], Eric Fluhr[3], Abraham Mathews[3], Andrew Bianchi[3], Daniel Dreps[3], David Wolpert[4], Eric Lai[3], Gerald Strevig[3], Glen Wiedemeier[3], Philipp Salz[5], Ryan Kruse[3]

[1]IBM, Bengaluru, India; [2]IBM, Yorktown Heights, NY; [3]IBM, Austin, TX
[4]IBM, Poughkeepsie, NY; [5]IBM, Boblingen, Germany

The POWER10™ processor designed for enterprise workloads contains 16 synchronous SMT8 cores (Fig. 2.4.1) coupled through a bi-directional high-bandwidth race-track [1][2]. A SMT8 core with its associated cache is called a core chiplet, and a pair of core chiplets forms a 39.4mm^2 design tile. Designed in a 7nm bulk technology, the 602mm^2 chip (0.85× of POWER9™ [3]) has nearly 18B transistors, 110B vias and 20 miles of on-chip interconnect distributed across 18 layers of metal: 8 narrow-width layers for short range routes, 8 medium-width layers for high performance signals and two 2160nm ultra-thick metal (UTM) layers dedicated for power and global clock distribution. There are 10 input voltages as shown in Fig. 2.4.1: core/cache logic (V_{dd}), cache arrays (V_{cs}), nest logic (V_{dn}), two PHY voltages (V_{io}, V_{pci}), stand-by logic (V_{sb}), a high-precision reference voltage (V_{ref}), DPLL voltage (V_{DPLL}), analog circuitry voltage (V_{AVDD}), and an interface voltage (V_{3P3}). The C4 array contains 24477 total connections (1.25× of [3]) with 10867 power, 11879 ground and 1731 signal connections. A core and its associated L2 cache are power-gated together, while the L3 cache is power-gated independently.

Scaling from 14nm SOI to 7nm bulk technology forced numerous design innovations. Power grid robustness is bolstered with 19μF (0.4× of [3]) of on-chip capacitance along with additional sensors for droop detection. The change from SOI's sparse buried insulator ties to bulk's denser nwell/substrate contacts resulted in 32× as many placed cells consuming 14× as much area as the previous design, driving new algorithms for cell insertion around large latch clusters and voltage regions. A new hierarchical antenna infrastructure was created, aware of macro dimensions, pin locations, and multi-sink nets, with path-aware insertion of wire jumpers and diodes to enable concurrent hierarchical design. Increased wire parasitics and 1× metal layer rule constraints drove extensive use of via meshes, particularly on wide-wire timing-critical nets. Place and route innovations coupled with a 9 tracks-per-bit library cell image improved silicon utilization and routability without sacrificing performance. SER-resilient latches with redundant state-saving nodes were used, trading 2.5× latch size in <0.5% of architecturally critical instances.

As shown in Fig. 2.4.2, the clock infrastructure uses a pair of redundant reference clocks with dynamic switch-over capability for RAS. 34 PLL/DPLLs control 90 independent meshes across the chip. A complex network of 20 differential multiplexers allow a multitude of reference clocks with varying spread-spectrum and jitter capability into the PLL/DPLLs. A single DPLL generates synchronous clocks for the 8 design tiles, each of which contains four 1:1 resonant meshes (tuned to 2.8GHz), four 2:1 non-resonant meshes, and a chip-wide nest / fabric mesh. Within each design tile, 4 skew-sensors, 4 skew-adjusts, and 4 programmable delays continuously align all running meshes to the chip-wide nest mesh within 15ps across the entire voltage range. Another DPLL generates groups of synchronous Power Accelerator Unit (PAU) clocks, where each PAU portion can be independently mesh-gated. The clock design can also choose between an on-chip-generated PCIe reference clock or two off-chip PCIe reference clocks to further improve its spread-spectrum and jitter. Similar to [3], the programmability of clock drive strengths, pulse widths, and resonance mode reduces clock power by 18% over traditional designs.

Four variants of customized SRAM cells constitute over 200MB of on-die memory. A performance-optimized lower threshold voltage (V_T) 0.032μm^2 6 transistor (6T) SRAM cell with single- and dual-port versions is used in 7 high-speed core arrays and single-port compilable SRAMs. The dense L2/L3 caches use a leakage-optimized 6T 0.032μm^2 cell with dual supply, while a 0.054μm^2 8 transistor (8T) SRAM cell is used in two-port compilable arrays. A larger menu (1.5× of [3]) of different ground rule clean cells is used in 10 custom plus several compilable multi-port register files, 3 content-addressable memories (CAM) and synthesized memories. SRAM arrays have optional write-assist circuitry applying negative bitline boost or local voltage collapse for banked 6T designs to support operation down to 0.45V. Most of the array peripheral logic (decoder, latches, IO and test) is synthesized, with structured placement enabling in-context optimization, logic and latch sharing and simplified custom components [4].

The SMT8 core was optimized for both single thread and overall throughput performance on enterprise-scale workloads. Core microarchitecture enhancements [5] include 1.5× instruction cache, 4× L2 cache, and 4× TLB, all with constant or improved latency, larger branch predictors and significantly increased vector SIMD throughput compared to [3]. The core additionally includes four 512b matrix-multiply assist (MMA) units. The MMA unit shares the core clock mesh to reduce overall instruction latency, yet has a separate power domain that is dynamically power-gated off when not in use to optimize energy efficiency. The core physical design consists of fully abutted physical hierarchies, with the instruction and execution control units flanking the load store and the arithmetic units, placed below the MMA (Fig. 2.4.1). Traditional logical unit boundaries were dissolved, and content was redistributed into large floorplanned blocks. Each resulting block utilized all metal levels except the UTM layers. This improved methodology removed 2 levels of physical hierarchy, enabled efficient area and metal usage in the core, and resulted in 13 synthesized blocks (0.1× of [3]) with an average of 800k nets, 750k cells and a total of 404 hard array instances. Each core chiplet includes 6 digital thermal sensors (DTS), 2 digital droop sensors (DDS) with associated controllers for thermal and voltage management, along with 4 process-sensitive ring oscillators (PSRO). The DTS and DDS are located at high thermal and voltage stress locations in the design as shown in Fig. 2.4.3, superimposed on a thermal map of a 10-core enabled processor running a core-heavy workload. Additionally, 14 DTS, 8 high-precision analog thermal diodes, 16 PSROs (distributed across two voltage rails), 12 skitters for clock jitter monitoring and a composite array of process monitors are distributed across the rest of the chip.

The high-bandwidth performance-critical race track is structured through the design tiles on metal10 and above, enabling improved silicon efficiency and reduced latency. The synchronous portion of the nest also includes power management, 2 memory management units, interrupt handling, a compression unit, test infrastructure and system configuration and control logic. Additionally, 6 accelerator functions, 4 memory controllers, 2 PCIe host bridge units, data and transaction link logic, all on the V_{DN} rail, operate asynchronous to the core, while tied to the I/O speeds. These components, 12 of which can be selectively power gated (Fig. 2.4.4) are built in a modular fashion (identical across the east / west regions), primarily at metal8 ceiling to prioritize race-track wiring.

POWER10™ features 144 lanes of high-speed serial I/O capable of running 25-to-32.5Gb/s at 5pJ/b supporting the OpenCAPI protocol and SMP interconnect providing 585GB/s of bandwidth in each direction as shown in Fig. 2.4.4. To support up to 30dB of channel loss an Rx architecture using 3-tap decision feedback equalizer (DFE) with continuous time-linear equalization (CTLE) and LTE was used, as well as a series-source terminated style Tx utilizing 2 taps of feed-forward equalization and duty-cycle correction circuitry. A dual-bank architecture on the receiver enables complete recalibration of analog coefficients during runtime. 16 OpenCapi Memory Interface (OMI)/DDIMM 8-lane busses capable of running at 21-to-32GB/s designed with the same PHY architecture as the SMP interconnect provides 409.6GB/s of bandwidth. 32 lanes of industry-standard 32Gb/s PCIe Gen5-compatible PHYs are implemented with vendor IP. 16 of the 32 lanes are limited to Gen4, providing a total bandwidth of 96GB/s.

Energy efficiency was significantly improved through micro-architectural and design changes including improved clock gating and branch prediction, instruction fusion and reduced memory access [5]. Power consumption of functional areas and components is illustrated in Fig. 2.4.5. Splitting the V_{IO} domain into multiple islands (Fig. 2.4.1) allows for system-specific power-supply enablement of only used interfaces. System-specific supply voltage modulation enables nest power to be maintained at less than 5%. Leakage power is maintained at less than 20% via aggressive multi-corner design optimization and intelligent usage of three different threshold voltage (V_T) logic devices, with less than 3% usage of the fastest device type. A 25% increase in latches connected to a local clock buffer through improved library design, placement algorithms, and less than 3% usage of high-power latches enables power of sequential components of a core chiplet to be less than ~20%. The clock network consumes ~10% of the total power. These enhancements enable 65% of the power to be allocated to the core chiplets.

Frequency vs. voltage shmoo is shown in Fig. 2.4.6 for cores within a chip, and across process splits. The shallower slope at higher voltage and faster process can be attributed to wire-dominated lowest V_T paths being limiters. Frequency is boosted based on workload [6] up to an all-core product maximum of 4.15GHz. A POWER10™ processor can be packaged in a single-chip, as well as dual-chip module, enabling up to a maximum of 256 threads per socket.

References:

[1] Samsung 7nm Technology.
[2] W. C. Jeong et al., "True 7nm Platform Technology featuring Smallest FinFET and Smallest SRAM cell by EUV, Special Constructs and 3rd Generation Single Diffusion Break," *IEEE Symp. VLSI Tech.*, pp. 59-60, 2018.
[3] C. Gonzalez et al., "POWER9™: A Processor Family Optimized for Cognitive Computing with 25Gb/s Accelerator Links and 16Gb/s PCIe Gen4," *ISSCC*, pp. 50-51, 2017.
[4] P. Salz et al., "A System of Array Families and Synthesized Soft Arrays for the POWER9™ Processor in 14nm SOI FinFET technology," *ESSCIRC*, pp. 303-307, 2017.
[5] W. Starke, B. Thompto "IBM's POWER10™ Processor", *IEEE HotChips Symp.*, 2020.
[6] B. Vanderpool et al., "Deterministic Frequency and Voltage Enhancements on the POWER10™ Processor," *ISSCC*, 2022.

2022 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731594

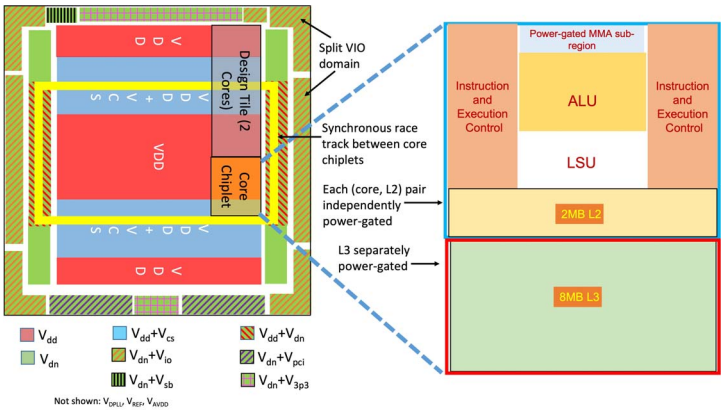

Figure 2.4.1: POWER10 floorplan and voltage representation.

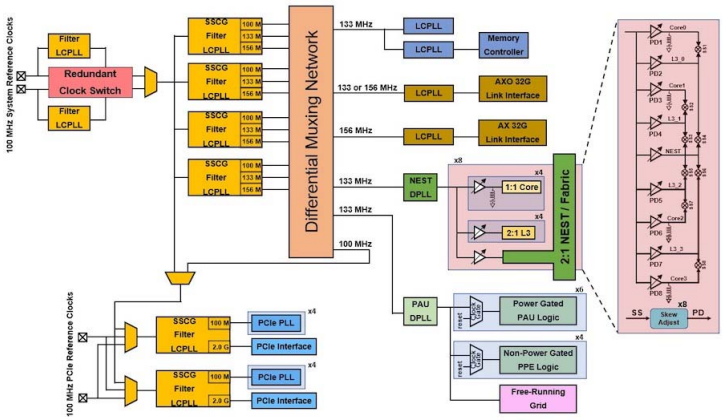

Figure 2.4.2: POWER10 clock distribution.

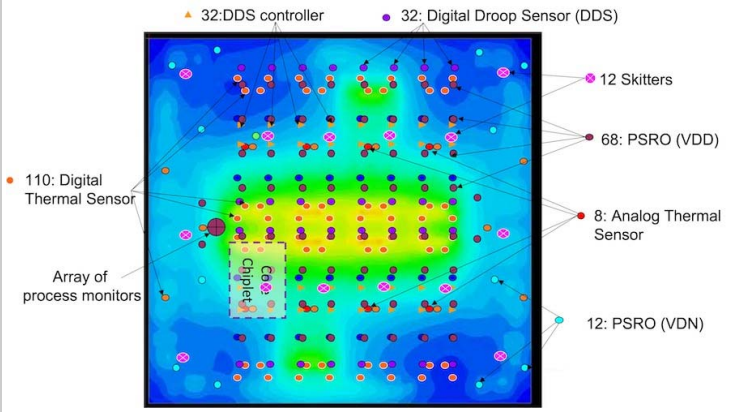

Figure 2.4.3: POWER10 sensors on 10 core thermal map for a core sensitive workload.

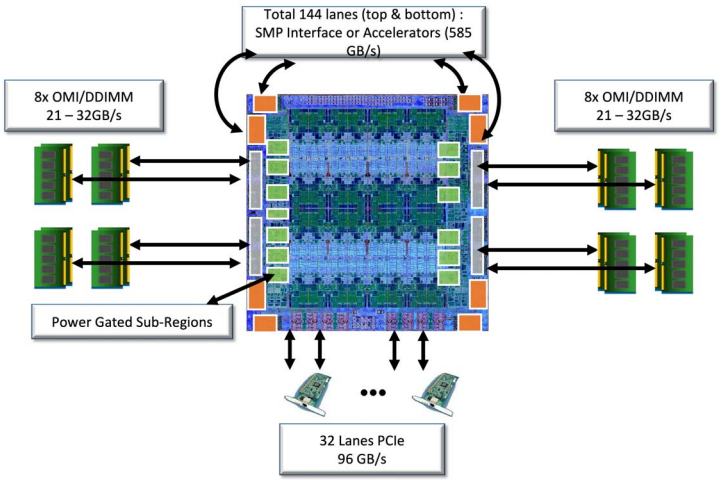

Figure 2.4.4: POWER10 interfaces and power gating.

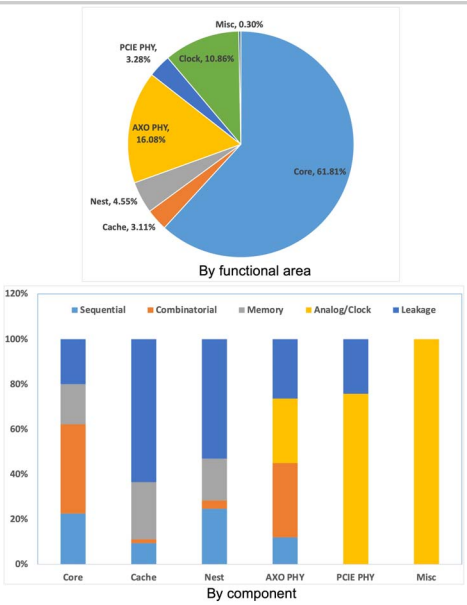

Figure 2.4.5: Chip power components.

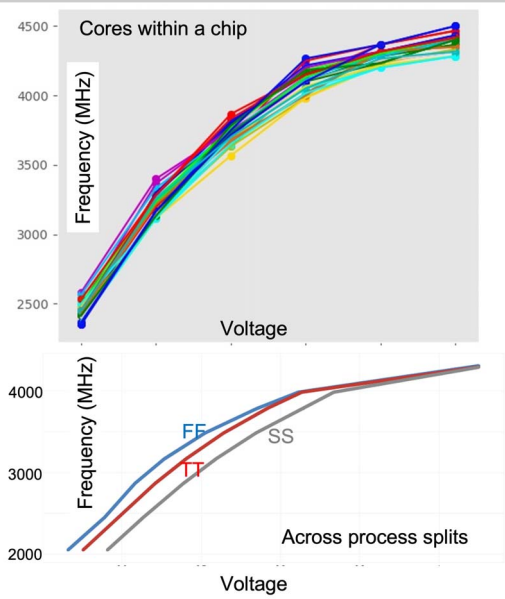

Figure 2.4.6: Frequency sensitivity across voltage and process.

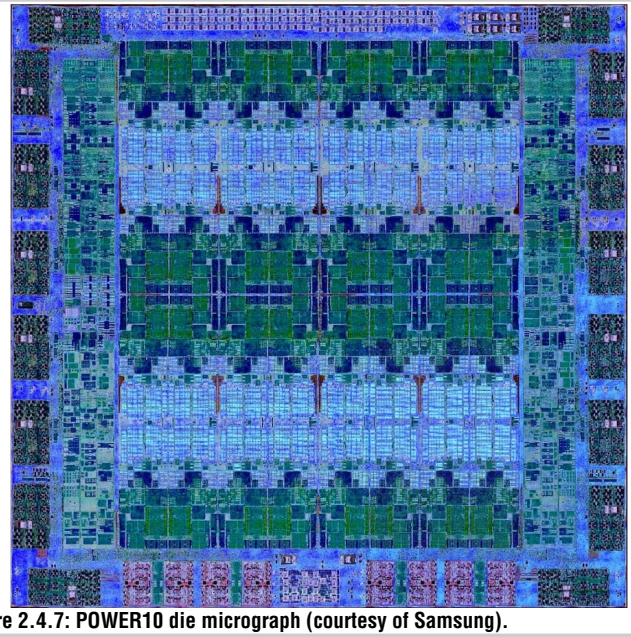

Figure 2.4.7: POWER10 die micrograph (courtesy of Samsung).

978-1-6654-2800-2/22/$31.00