The 24-Core POWER9 Processor With Adaptive Clocking, 25-Gb/s Accelerator Links, and 16-Gb/s PCIe Gen4

Christopher Gonzalez, Michael Floyd, Eric Fluhr, *Member, IEEE*, Phillip Restle, *Senior Member, IEEE*, Daniel Dreps, Michael Sperling, Rahul Rao, *Senior Member, IEEE*, David Hogenmiller, Christos Vezyrtis, Pierce Chuang, *Member, IEEE*, Daniel Lewis, *Member, IEEE*, Ricardo Escobar, Vinod Ramadurai, Ryan Kruse, Juergen Pille, Ryan Nett, Pawel Owczarczyk, Joshua Friedrich, Jose Paredes, Timothy Diemoz, Saiful Islam, Donald Plass, and Paul Muench

***Abstract*— The POWER9TM family of chips is fabricated in 14-nm silicon-on-insulator finFET technology using 17 levels of copper interconnect. The 695-mm^2 24-core microprocessor features a new core based on an execution slice microarchitecture. The chip contains 8 billion transistors and has 120 MB of eDRAM L3 cache. The processor features an adaptive clock strategy to reduce timing margin needed during power supply droop events by embedding analog voltage-droop monitors that direct a digital phase-locked loop to immediately reduce clock frequency in response to a droop event. The scale-out chip IO subsystem supports up to 300-GB/s accelerator bandwidth using new 25-Gb/s links, 48 lanes of PCIeGen4 totaling 192 GB/s, eight ports of 2667 MT/s DDR4, and 256 GB/s of symmetric multiprocessor (SMP) interconnect. The scale-up chip adds additional SMP bandwidth and replaces the DDR4 memory interface with eight ports of differential memory interfaces with 230 GB/s of bandwidth resulting in 12.9 Tb/s of total off-chip bandwidth.**

***Index Terms*— Adaptive clocking, clock distribution, microprocessor, resonant clocking, voltage domains, voltage regulator.**

I. Introduction

THE dominant computing paradigm traditionally centered around homogeneous CPUs is no longer ideally suited for competing needs between major segments of the evolving IT industry. High-performance computing (HPC) is moving into the exascale era thereby demanding exaFLOPS of computation. Meanwhile, hyper-scale data centers, the backbone for the cloud computing, require node cost and power consumption to be balanced against high bandwidth and compute performance needs. Traditional enterprise-level systems rely on increases in core throughput, large memory footprints, and exceptional reliability, availability, and serviceability (RAS). Emerging within these segments are new analytics algorithms based on the artificial intelligence and cognitive processing. These workloads can run effectively on general-purpose CPUs but heterogeneous computing enables orders-of-magnitude increase in processing.

The POWER9 family of chips meets this challenge with multiple options for core configuration, memory bandwidth, and accelerator attachments. In this paper, Section II summarizes the different chip offerings and primary features of POWER9 microprocessors. Next, Section III details the power delivery and voltage regulation that supports flexible core performance. Section IV presents the chip clock infrastructure leveraged by the adaptive clocking technique described in Section V. Section VI explains the IO subsystem, featuring new 25-Gb/s accelerator links, and enhanced PCIe Gen4 at 16 Gb/s, which enable off-chip data bandwidth up to 12.9 Tb/s. Finally, Section VII touches on power optimizations made within the processor and shows processor hardware data before Section VIII concludes this paper.

II. POWER9 Microprocessor Overview

The POWER9 family of chips is designed to meet the differing needs of scale-out (SO) computing versus enterprise-class scale-up (SU) servers, while offering high-bandwidth accelerator links for analytics and HPC workloads. SO computing demands low-latency, power- and cost-efficient, industry standard memory optimized for 1 or 2 sockets. Traditional SU systems leverage terabyte memory capacity with high-bandwidth interfaces along with flat expansion processors. The POWER9 processor variant optimized for SO systems features eight direct-attach DDR four ports, two socket-so-socket interfaces, 48 lanes of PCIe Gen4, and up to 48 lanes of 25-Gb/s acceleration links. A complementary processor optimized for SU computing replaces the DDR4 ports with higher bandwidth differential memory interfaces (DMI) providing larger memory footprints, higher memory bandwidth, and enterprise-level RAS. The SU adds an additional socket-to-socket interface allowing full connectivity to four sockets, and a further 48 lanes of 25-Gb/s

Manuscript received May 17, 2017; revised July 20, 2017; accepted August 20, 2017. Date of publication December 15, 2017; date of current version December 26, 2017. This paper was approved by Guest Editor Muhammad M. Khellah. *(Corresponding author: Christopher Gonzalez.)*

C. Gonzalez is with IBM Systems, Yorktown Heights, NY 10598 USA (e-mail: jgonzalz@us.ibm.com).

M. Floyd, E. Fluhr, D. Dreps, D. Hogenmiller, R. Escobar, V. Ramadurai, R. Kruse, R. Nett, J. Friedrich, J. Paredes, and S. Islam are with IBM, Austin, TX 78758 USA.

P. Restle, C. Vezyrtis, and P. Chuang are with IBM VLSI Design, Yorktown Heights, NY 10598 USA.

M. Sperling, P. Owczarczyk, T. Diemoz, D. Plass, and P. Muench are with IBM Systems, Poughkeepsie, NY 12601 USA.

R. Rao and D. Lewis are with IBM, Bangalore 560045, India.

J. Pille is with IBM, 71032 Boeblingen, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2748623

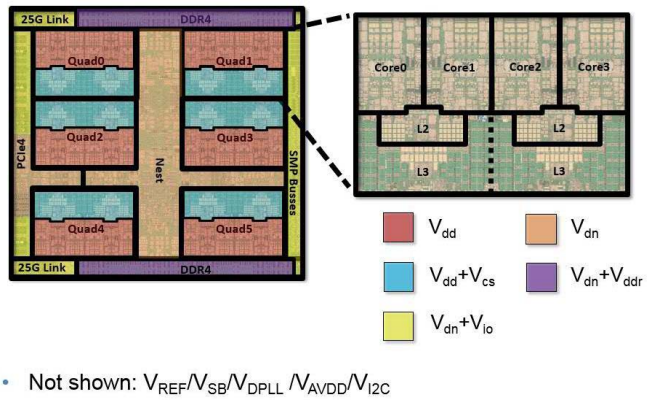


Fig. 1. Location of voltage domains.

IO extends fully connected symmetric multiprocessors (SMPs) to 16 sockets. Two variants of a newly architected POWER ISA v3.0 core can be mated with either I/O subsystem: 1) up to 24 cores with four threads (SMT4) each or 2) up to 12 cores with eight threads (SMT8) each. The 24-core chips are optimized for the Linux ecosystem and virtualization granularity; the 12-core chips maintain continuity and strong thread performance in the PowerVM ecosystem, optimized for large partitions.

The 695-mm^2 SO processor in Fig. 1 is implemented in the14-nm silicon-on-insulator finFET technology [2]. It contains 8 billion transistors with a 17 layer copper interconnect back end of line. The bottom 15 levels of metal are available for signal routing while the top two metal layers are reserved for global clock and power distribution throughout the chip. Three thin-oxide logic transistors V_{ts} were used to balance power and performance requirements. Unlike previous designs, all analog and I/O circuits used stacked FETs to support voltages greater than the technology V_{max} instead of relying on thick oxide devices to enable a simpler manufacturing process. Since finFETs offer increased current per unit area compared to planar transistors, the base standard cell image shrunk to ten tracks from 18 tracks per bit. This maximized area scaling for logic common with POWER8 [3] and increased SAPR efficiency for non-critical circuits.

The new POWER9 core is optimized for cognitive workloads, with an extensible microarchitecture built around execution slices. Each SMT4 core has a 32-KB L1 instruction cache and a 32-KB L1 data cache which supports up to four load or store instructions per cycle. The pipeline eliminates five stages between instruction fetch and execution [2] versus POWER8. The fixed and floating point units merge for improved data exchange, and a new adaptive-prefetch algorithm dynamically pulls either 64 or 128 B cache lines from memory depending on program characteristics and available memory bandwidth. Core pipeline efficiency also improved, increasing utilization and reducing pipeline disruptions via improved branch prediction and local instruction recycling in the load/store unit (LSU). The fundamental execution building block is a 64-bit slice containing dataflow plus associated control of the instruction sequencing unit, the vector scalar unit, and the LSU. To optimize area and power, the design combines two 64-bit slices together to form one 128-bit super-slice. From this physical design building block, two super-slices are assembled into a 256-bit wide SMT4 core, or four super-slices are assembled to create a 512-bit wide SMT8 core.

Differing performance and design efficiency requirements in various regions of the chip requires seven different memory families, each balancing port count, area, frequency, and power against design effort. Structures critical to chip area, power, or frequency such as the L1 data cache, the larger L2 cache, and the 120-MB L3 eDRAM are custom designed with ground-rule optimized technology cells. Multi-ported, critical register files, such as the GPR, rely on ground-rule clean cells for their custom design. The set of custom memories is built on four different technology-offered RAM cells: 1) a performance optimized 0.102-μm^2 cell for core SRAMs; 2) a leakage optimized 0.102-μm^2 cell for large SRAMs; 3) an 8T 0.143-μm^2 cell for compilable SRAMs; and 4) an 0.0174-μm^2 eDRAM cell supporting the L3 cache. The remaining memories leveraged one of the three compilable systems: 1) growable eight-transistor technology cell for density; 2) growable ground-rule legal register file for multi-port support; and 3) a new "soft structured" array built with synthesizable subcells.

POWER9 contains an embedded PowerPC 405 on-chip controller (OCC) centrally located to monitor and adjust the energy efficiency according to firmware and operating system directives. To enable improved power management algorithms with a 2× increase in core count versus POWER8, 20 additional microcontrollers distributed around the chip assist the OCC. These smaller engines communicate with the OCC via a sideband network are placed locally to improve response time for cores to switch between defined power and idle states, and provide a 10× faster wakeup latency from core power-off state compared to POWER8. The chip includes 63 digital thermal sensors and 30 voltage-droop monitors feeding these engines for dynamic voltage and frequency scaling (DVFS) decision making. A primary design goal was to obtain performance improvement through workload-optimized frequency. For any set of workloads running, this power management subsystem monitors current and thermal headroom and adjusts core voltage and frequency to maximize performance or power consumption.

III. Power Distribution

POWER9 has a total of ten input voltages: core/cache logic (V_{dd}), cache arrays (V_{cs}), nest logic (V_{dn}), PCIe/25G/SMP (V_{io}), DDR (V_{DDR}), I2C/SPI (V_{I2C}), digital phase-locked loop (DPLL) voltage supply (V_{DPLL}), analog circuitry (V_{AVDD}), stand-by logic (V_{sb}), and a high-precision reference (V_{ref}) as shown in Fig. 1 and contains 48.5 μF of deep-trench decoupling capacitance. The core and L2/L3 cache regions are divided into "quads," which consist of four cores plus associated cache and can be individually power gated via distributed PFET headers.

Each quad re-uses its PFET power headers as low-dropout (LDO) integrated voltage regulator macros (iVRMs) to provide a cost-effective path to realizing per-quad DVFS. Lowering frequency and voltage on cores that do not need

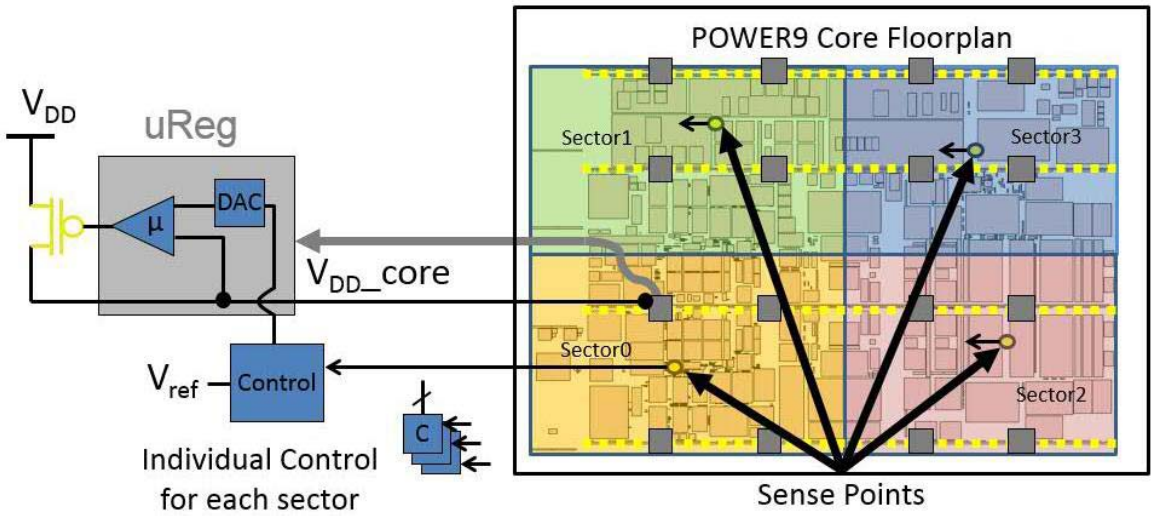


Fig. 2. Location of uReg and sense points.

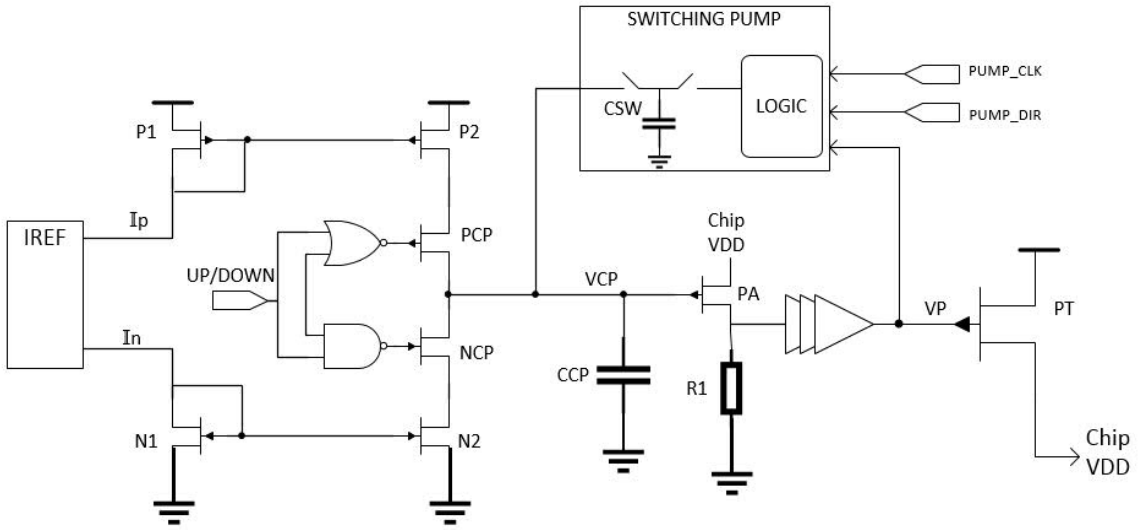


Fig. 3. Fast switching iVRM circuit.

to run at full speed allows power to be reallocated to more active cores which can then operate at higher frequencies. Thus, the overall performance of the system is maximized within a given power envelope. The POWER9 iVRM design made numerous improvements to achieve approximately a 50% reduction in the effective dropout voltage versus POWER8 [4]. This was accomplished through optimization of the PFET power headers to distribute the high-speed digital LDO signals throughout the header row. Instead of a single voltage sense point, the improved iVRM uses an optimized multi-sector, multi-sense scheme that divides each core into four virtual regions (sectors) which are separately monitored and adjusted to the target voltage. Fig. 2 shows the location of the sense points in the core floorplan. The central controller receives the differential analog sense voltages from these sense points and uses a common reference circuit to generate UP/DOWN signals for the micro-regulators (uREGs) in each of the four sectors. By centralizing in this fashion, there is less mismatch in the resultant sector voltages and the quad benefits from less area and power overhead of the iVRM system.

Fig. 3 illustrates the details of one of the uREGs, which are located throughout the power header rows. The general operation of the uREG is to use a high-speed comparator to modulate a PMOS pass gate PT ON and OFF in a bang–bang fashion at node VP. The comparator trip point is tuned for high dc accuracy with a local charge pump composed of transistors PCP and NCP, whose output (VCP) serves as a reference voltage for an error amplifier (common-gate stage power amplifier). Instead of the current steering a digital to analog converter that was used in previous iVRMs, a delta-sigma modulated signal UP/DOWN is used, which simplifies both the integration of the power headers within the processor core as well as the design of the central voltage controller. This simplification is most evident in the use of a single bit comparator (not shown) instead of the multi-bit flash comparator that was previously used. Offsets in this comparator as well as

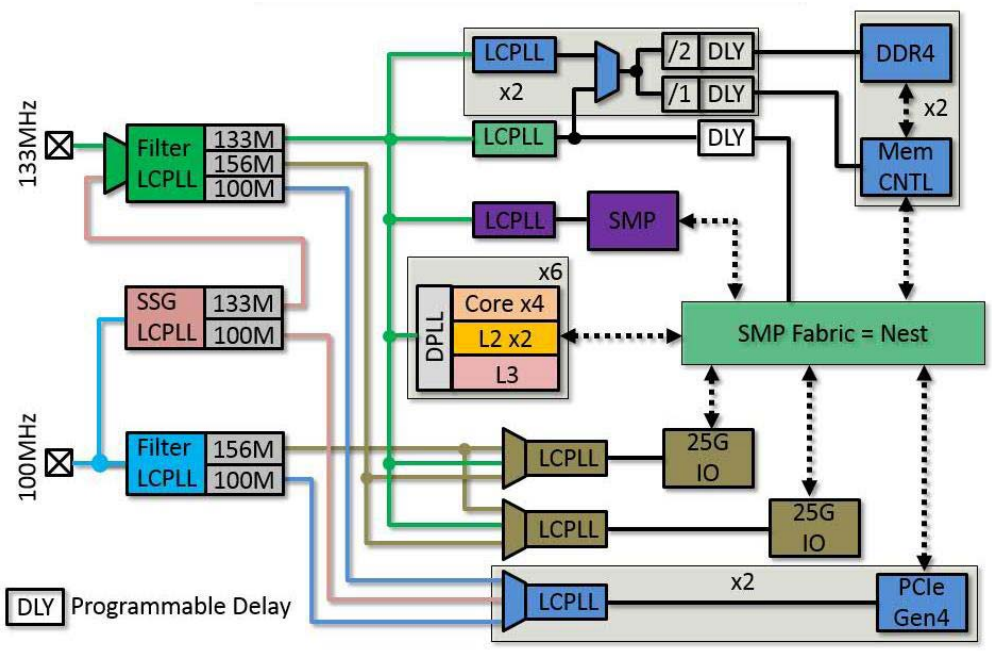


Fig. 4. Global clock distribution.

the analog reference generation are cancelled out using auto-zero techniques which result in a maximum dc error of 5 mV from the ideal set point.

An additional improvement for POWER9 is the addition of a fast switching pump that can transition from one voltage level to another 500% faster than just relying on the action of the local charge pump. Capacitor CSW is sized to a small fraction of capacitor CCP in order to ensure approximately 2 mV of voltage change per pump. Signals PUMP_CLK and PUMP_DIR, which come from the central voltage controller, control the pumping frequency, and direction. The central controller also communicates its status back to the power management logic such that the chip knows when the voltage has reached the target. Extra logic, including feedback from VP, is added to the uREG to ensure that the voltage pump rate does not exceed the system's response rate. Hardware measurement indicates that technique enables the iVRM to reach a controllable 0.6-mV/ns slew rate.

IV. Clocking Architecture

The POWER9 processor clock topology is shown in Fig. 4. It has two system reference clocks each of which can optionally be made redundant for higher RAS. Alternatively, the chip can also run using only one reference clock. There are 17 phase-locked loop (PLL)/DLL's controlling 58 independent clock meshes across the chip. The clock structure enables the ability to run DDR either synchronously or asynchronously with the SMP coherency fabric. The data path between the memory controller and the SMP fabric goes through an FIFO. When operating in synchronous mode, the FIFO is bypassed to reduce memory latency by several nano-seconds. Five programmable delay buffers are statically set at power-on time to ensure proper clock skew between the SMP fabric, memory controller, and DDR. The actual DDR mesh runs at half the clock frequency of the memory controller to reduce power.

In total there are six quads on chip, each of which contains seven clock meshes driven from a single DPLL. The four core and two L2 clock meshes run at 1:1 and use resonant clocking with pulsed sector buffers to reduce global clock mesh power. The L3 clock mesh operates at 2:1 compared to the core clock mesh and hence does not run in resonant-clock mode, as resonant clocking is less attractive at lower frequency operating points. The POWER8 design used non-pulsed buffers [5]. By comparison, the z13 resonant clock used pulsed buffers with a single pulsewidth, which was appropriate for processors optimized for a single frequency [6]. For the POWER9 design, a single pulsewidth is not be optimal over the full voltage and frequency range desired. As a result, the pulsed clock mesh driver was designed with four programmable pulse widths that are selected based on the clock frequency.

In a previous design [6], the pulse-mode buffer used an inverter chain and multiplexers to select the desired pulsewidth. This topology inherently introduces a glitch hazard when the control signals switch, potentially resulting in a single short clock cycle or reduced signal quality. On-the-fly pulse-mode changing is required for POWER9 applications, so an alternative delay schematic was used that eliminates the hazard of glitches. Fig. 5 shows the new programmable pulsewidth circuit, where programmable-strength inverters are used for the pulse delay generation; the circuit is glitch free since control signals only result in different dc settings for the programmable inverters. They do not propagate to the delay output and do not generate glitches as delay modes are changed on-the-fly.

The POWER9 resonant-clock design includes pulsewidth control, buffer-strength control, and the ability to change between resonant and non-resonant modes (for very

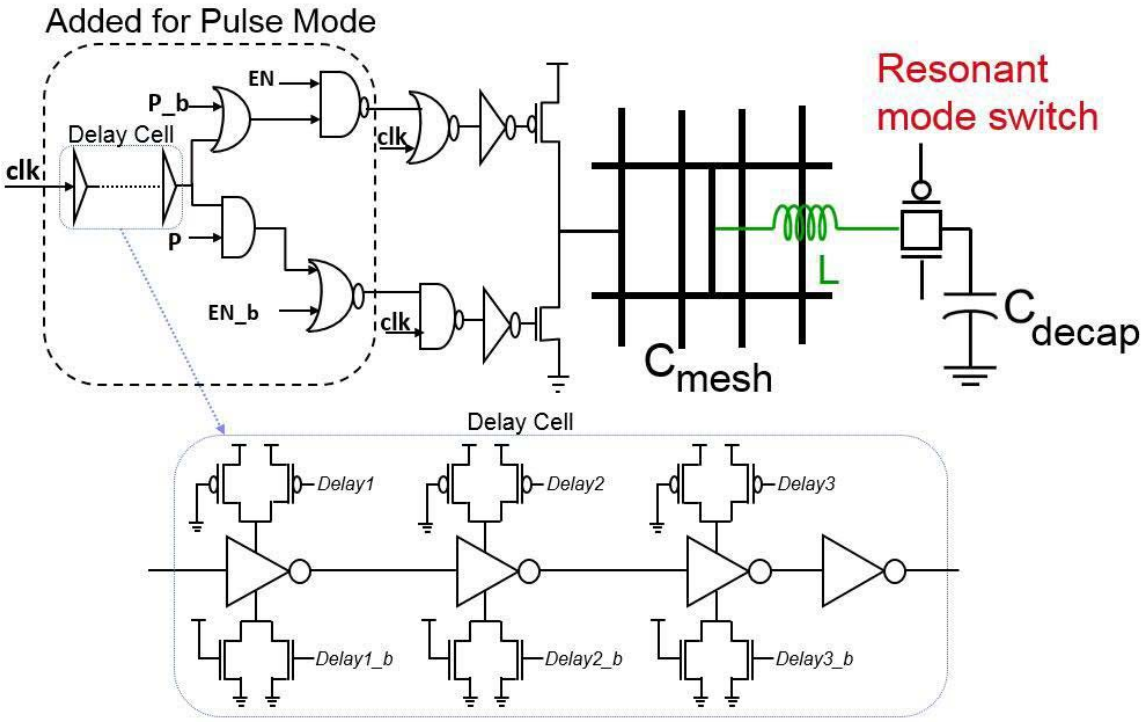


Fig. 5. Pulse-mode clock buffer driving resonant global clock mesh.

low-frequency operation if desired). The addition of pulsewidth control removed the requirement for two resonant frequencies allowing for half of the inductors and mode switches to be removed compared to the POWER8 design. In the POWER9 design, each parallel pair of inductors was replaced with a smaller single inductor with slightly more than half the inductance of each of the P8 inductors. These results in a single resonant frequency located between the high-frequency and low-frequency resonances of the POWER8 design. Reducing both the number and size of the inductors on POWER9 enabled the use of higher-Q two-layer inductors [6], [7] which reduces the inductor parasitic capacitances by almost 4×. This reduction more than offset the increase in clock power when operating the clock frequency far above or below the single resonant frequency. The measured resonant clock power for the four pulse widths as well as non-resonant mode is shown in Fig. 6. For frequencies near or above the resonant frequency the narrower pulse widths are used to optimize power and clock signal quality, which also results in lower clock power due to reduced power consumption in the first stages of the local clock buffer (LCB) circuits. Below the resonant frequency, wider pulse widths are needed to achieve required signal quality. As a result, the addition of pulsed buffers reduces power by 10% over the POWER8 resonant design over a wide frequency range.

Active de-skew circuitry continuously aligns all of the running meshes, while accommodating DVFS, powering on/off the resonant and pulsed meshes, and other adaptive clocking frequency adjustments. Clock mesh nodes at the sensors are aligned with a skew of less than 15 ps.

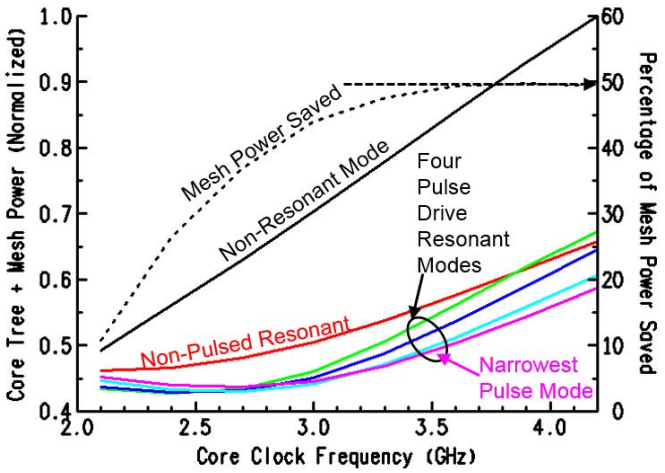


Fig. 6. Measured tree and mesh clock power in all resonant modes at constant VDD.

V. Adaptive Clocking

Increasing transistor counts in modern processors create larger changes in current, causing nanosecond-speed supply voltage droops that require large guard bands that cost power. The POWER9 processor uses an adaptive clock strategy that preserves timing margin during power supply droop events by embedding analog voltage-droop monitors (VDMs) that direct a DPLL to quickly reduce clock frequency in response to droops.

As shown in Fig. 7, a voltage sense point in each of the five clock/power domains in the quad (circles) is sampled (solid lines) by each of the five VDMs (triangles), whose outputs (dashed lines) are combined to feed the DPLL.

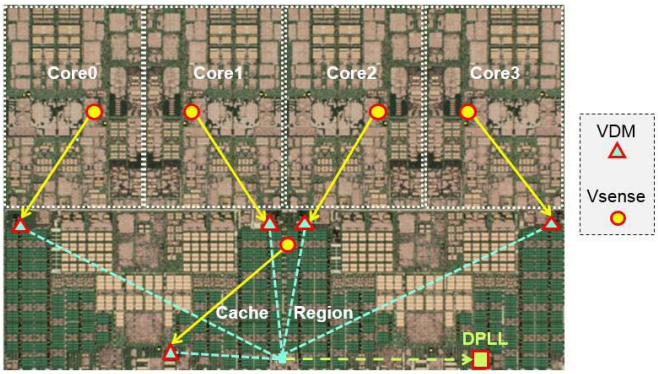


Fig. 7. Die photograph of the POWER9 quad showing the five independent power regions and the associated adaptive clocking network.

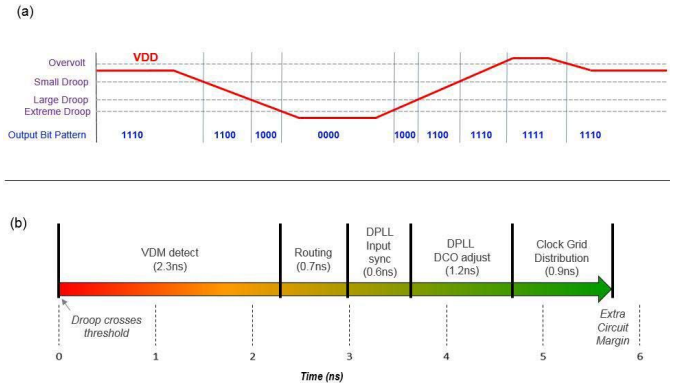


Fig. 8. (a) VDM digital output response to sensed input voltage (VDD). (b) Response latency of each element in adaptive clock sense and response.

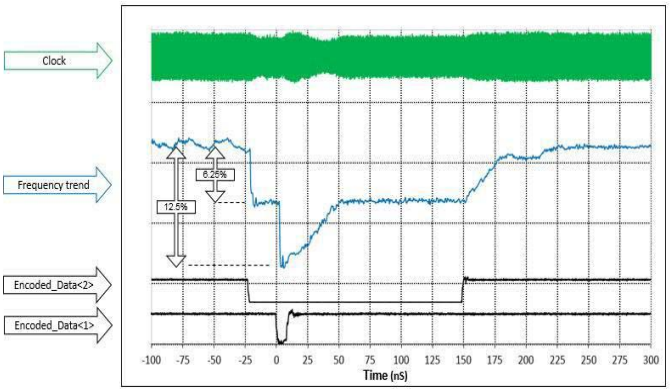


Fig. 9. Scope trace of 2-GHz mesh clock showing two-level frequency jumps followed by a gradual recovery in response to VDM detection. The gradual recovery from a large droop can be seen between 10 and 50 ns, while the small droop recovery occurs between 155 and 180 ns. The DPLL then slews back to full target frequency.

The quad uses an asynchronous interface to communicate to the nest (a.k.a. uncore), allowing per-quad DVFS. Other adaptive clocking schemes, support only fixed clock ratio reduction [8], [9], have higher response latency, use canary circuits with significant overhead [10], or rely on full error detection with retry [11]. These schemes result in greater performance loss to achieve similar guard band reduction. The quad's asynchronous design and fractional-N DPLL allow frequency to be adjusted quickly with fine granularity (3.125%) to maintain timing margin. A previous DLL mitigation scheme [12], similarly reduced frequency by less than a factor of 2, but with a significant increase in jitter, requiring additional frequency reduction to maintain timing guard band. Additionally, the POWER9 mitigation technique has negligible area impact, since the VDMs, routing overhead, and added DPLL circuitry together occupy just 0.12% of the 65-mm^2 quad area.

VDMs were chosen over other sensors [13] that are more difficult to calibrate over the wide range of DVFS required. In contrast, the VDM needs only the voltage set point required at each frequency target, a relationship stored in a DVFS table for each chip during manufacturing test. The VDM detects over- and under-voltage conditions by comparing the VDD supply grid to this desired voltage set point. By using an 8-b voltage identification code corresponding to the target frequency, as well as multiple digital threshold compare values, the VDM derives multiple threshold compare voltages. Each of these compare voltages can be independently tuned in 8-mV steps, which then feed comparators against the differentially sensed VDD. These produce a negative-active thermometer code, where a zero signifies a lack of voltage margin, as shown in Fig. 8(a). A code of "1111" is an overvolt condition, and "1110" means that VDD is within the "nominal" (acceptable) range for the target frequency when no droop is present. The codes "1100," "1000," and "0000" indicate small, large, and extreme droops, respectively. The VDM filters out high-frequency noise over a 1-ns timescale and responds with a droop indication within 1.5 ns, using digital comparators that output a new code every cycle. The output codes from the five VDMs are combined at the quad level and routed to the DPLL, which accounts for metastability and determines if a response is needed. When a droop code activates, the DPLL jumps to a lower frequency by instantly turning off a percentage of active devices in its digitally controlled oscillator (DCO). Each small or large jump amount is selectable in 3.125% increments. The DPLL also supports intermediate jumps between small and large droop indications. Two options are available to quickly recover frequency as the droop subsides: either gradually add devices back to the DCO every programmable number of reference clock cycles or instantly add the devices. Both cases are selectable in 3.125% increments, and overshoots are minimized by adding back fewer active devices compared to the original set point. This allows the natural DPLL dynamics to add the remaining devices back to the DCO as it naturally slews and locks to the original target. There is also the option to start this natural slewing immediately, which produces a smooth, though slower, return to the proper frequency. Regardless of the jump amounts, the total latency from supply voltage threshold crossing until the clock is slowed at the quad and core level circuits is 6 ns at typical core frequencies as shown in Fig. 8(b), which is sufficient to recover timing margin before the droop can induce a circuit failure.

While the adaptive clock function was verified above 4 GHz, cycle-accurate frequency measurements of the core mesh clock in the test setup are limited to 2 GHz. Fig. 9 shows both

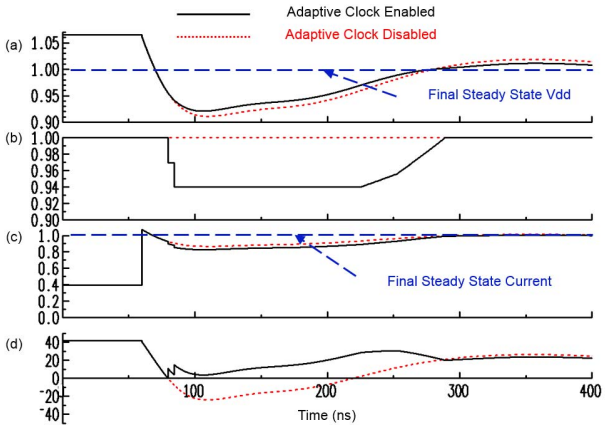


Fig. 10. Full system noise model of maximum activity-step droop event on all cores with adaptive clocking enabled and disabled. (a) V_{dd} (normalized). (b) Frequency (normalized). (c) Current (normalized). (d) Noise timing margin(ps).

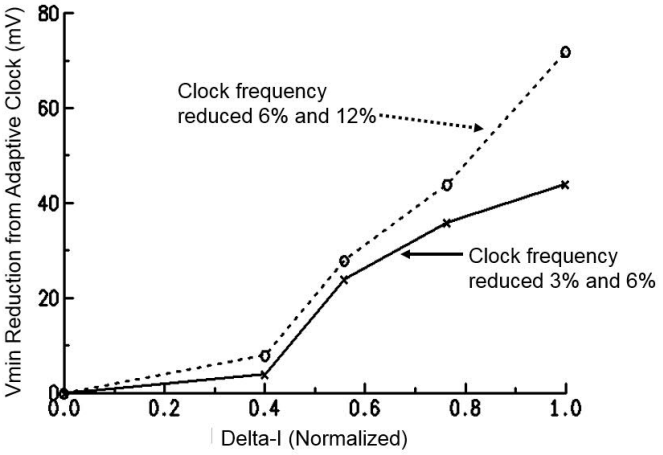


Fig. 11. Measured V_{min} reduction as a function of Delta-I created with artificial workload, for different adaptive clock two-level frequency reductions.

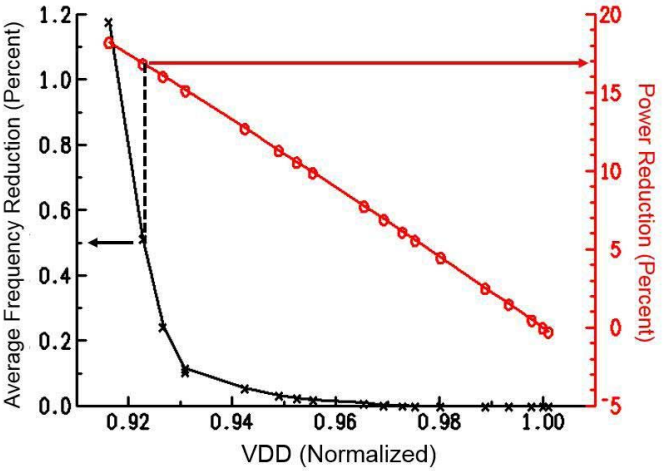


Fig. 12. Measured average frequency reduction and power savings versus normalized V_{dd}. Without adaptive clock, chip fails below normalized $V_{dd} = 1$. Adaptive clock saves approximately 17% power with 0.5% frequency loss.

small- and large-frequency jumps as the corresponding VDM thresholds are crossed in response to a real voltage droop event, followed by a gradual recovery back to the target frequency. While the models and initial testing indicate that 3.125% and 6.25% frequency responses should be sufficient, these traces were collected using 6.25% and 12.5% due to measurement jitter.

Fig. 10 contains a modeled voltage droop for a worst case activity-step synchronized across all cores, as well as the modeled adaptive clock response. The instantaneous demand for current in Fig. 10(c) triggers the two-step frequency reduction in response, as shown in Fig. 10(b). The second-level 6.25% frequency response reduces the power supply noise margin, in Fig. 10(d), required for the largest possible droops by 50% (−20 ps for the dotted line versus 0 ps, as compared to +20 ps at final steady state). While the reduction amount is primarily chosen to maintain timing margin during the droop event, the resultant voltage also lowers active power by the same amount, reducing droop magnitude by 17% seen by the dotted line in Fig. 10(a). When the adaptive clock droop mitigation is enabled, this allows the chip to run safely at lower power supply voltages, saving significant power. To measure this benefit, an artificial test sequence was used that can create a large programmable step in chip activity in a few cycles, and thus a large change in current (Delta-I). For each Delta-I the power supply voltage was reduced until a chip failure was detected. The minimum functional voltage (V_{min}) was then recorded for each case. Fig. 11 shows that for smaller Delta-I values and thus smaller noise, the two-step frequency reduction of 3% and 6% yielded similar results to the larger two-step response of 6% and 12%. However, for the largest artificial droops, the larger response would be necessary. In practice, the two-step response percentages, as well as the VDM thresholds, are optimized to maximize power savings while minimizing the performance cost from these rare noise events.

To evaluate the trade-off of frequency loss versus power savings benefit, a high-coverage high-power noisy workload was further modulated by throttling and un-throttling instructions in all cores to create Delta-I events that roughly halve and double the current every 570 μs. First the functional V_{min} was found with the adaptive clock disabled. Next the adaptive clock was enabled, and VDD was reduced until the average frequency reduction exceeded 1% as shown in Fig. 12. For this noisy workload, V_{dd} can be reduced by 7.7% (corresponding to approximately 17% power reduction) with an average frequency reduction of only 0.5%.

VI. High-Bandwidth IO

The POWER9 chip family offers two separate off-chip interface silicon solutions to address the unique needs of the SO versus the SU server markets. Common to both solutions are 48 lanes of 16-Gb/s PCIeGen4 totaling 192 GB/s of bandwidth. This provides a 2× bandwidth improvement over POWER8 with a 15% power increase per lane. The PCIeGen4 connection is also used for the next generation CAPI2.0 interface. Both chips also have SMP interconnect consisting of variable speed differential links with a top speed of 16 Gb/s. The SMP socket-to-socket bus is an extension

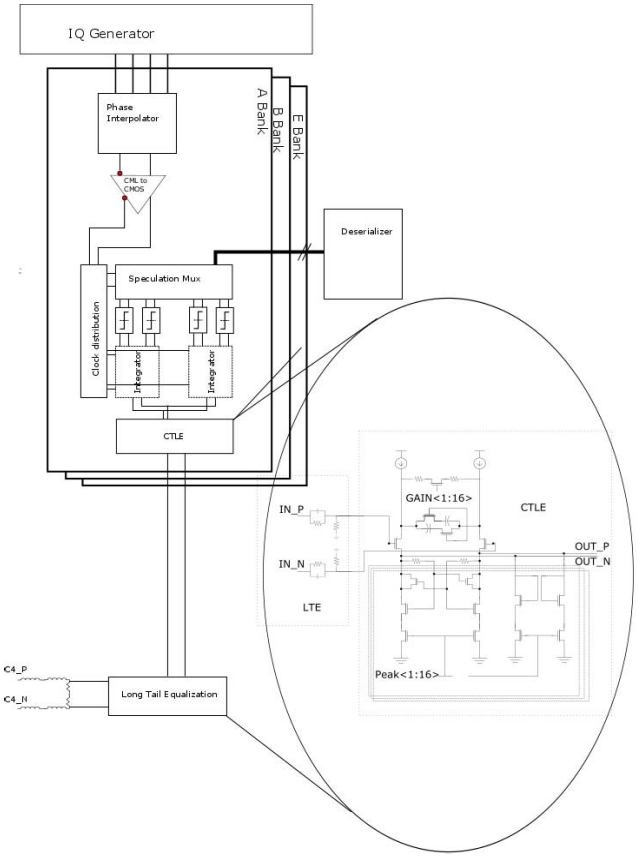


Fig. 13. 25-Gb/s RX topology.

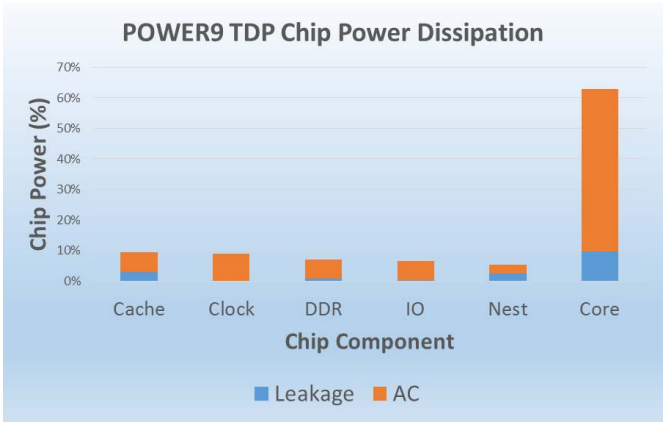


Fig. 14. Chip power distribution running a TDP workload.

of the elastic differential interface first used in Power8 at 9.6 Gbit at 5 pj/b with one tap of DFE [4]. Extending this to 16 Gbit improved CTLE inclusive of long-tail equalizer (LTE) and DFE modes of 1, 5, or 12 taps while maintaining 5 pj/bit with 1-V nominal voltage rail. The SO chip has 256 GB/s of 16-Gb/s SMP interconnects bandwidth while the SU optimized chip has 384 GB/s of 16-Gb/s bandwidth. The POWER9 SO chip memory interface uses eight ports of direct attach 2.667-GB/s DDR4 connecting up to 16 RDIMMs per socket while the SU chip replaces the DDR4 channels with eight DMI channels running at 9.6 Gb/s interfacing to memory buffer chips for a total of 230 GB/s of memory bandwidth.

The POWER9 SO chip features 48 lanes of 25-Gb/s PHY to support the next generation NVLink protocol enabling GPU acceleration with 7–10× the bandwidth of an industry standard PCIeGen3 connection. It also supports the OpenCAPI protocol which has up to 4 X8 interfaces to enable field-programmable gate array or application-specified integrated circuit acceleration across an open interface with minimal latency and up to 200 GB/s of bi-directional bandwidth. The SU chip adds 48 additional lanes 25-G link channels that can function as either acceleration links or SMP busses increasing total off-chip bandwidth to 12.9 Tb/s.

As shown in Fig. 13, the 25-Gbit receiver (RX) uses a three bank architecture where banks A and B ping pong and bank E acts as the edge path. This architecture allows dynamic calibration of all parameters such as gain, peaking adjust, offset cancellation, DFE adjust, and horizontal centering, without affecting integrity of live data. The RX includes a passive LTE common to all three banks of the design. After the initial 1:2 deserialization, the data are synchronized to the edge clock domain before a final 2:16 DEMUX. Clocking for the RX macro is provided by a DDR clock distributed from the PLL via a CMOS resonant-clock distribution. The RX internally generates quadrature phases which feed three independent octagonal CML phase rotators, one per bank. The outputs of the phase rotators are amplified to CMOS levels using an active inductor peaking buffer and an ac coupled CML to CMOS circuit to drive the DDR clocking used in each bank. Each bank consists of a CTLE, two track-and-hold circuits, two current integrators, four sampling latches, and two DFE speculation MUXes. The CTLE is designed to allow both dc gain adjustment using programmable source degeneration and peaking gain adjustment using variable strength active inductors and can be powered down when the bank is not being recalibrated. The current integrator provides approximately 12 dB of gain while consuming 7.5-mW per bank. The four sampling latches have a sensitivity of less than 10 mV and consume a combined power of 10.4-mW per bank. The per bank clocking power is approximately 25 mW including the phase rotator and CML to CMOS power. The RX's CDR is one pole placed between 1–10 MHz. Each lane of the RX macro occupies 0.072 mm^2 and operates at a power efficiency of 5.5 pJ/b at the 25.78125-Gb/s per lane data rate.

VII. Chip Power and Frequency Results

The estimated per component power breakdown for the POWER9 chip is shown in Fig. 14. Under nominal process conditions and while running a thermal design point (TDP) workload, the chip power distribution is 80% ac and 20% dc, with nearly 60% of the power consumed in the cores. To reduce IO power, the interface units are placed in special low-power modes during periods of low-off chip communication activity. The clock power is maintained at less than 10% due to the combination of improved resonant clocking (with pulsed mode operation), optimized LCBs that could drive greater than 35% more latches, thereby reducing the load on the clock mesh, and efficient latch design and clustering. Fig. 15 shows the comparison of chip power

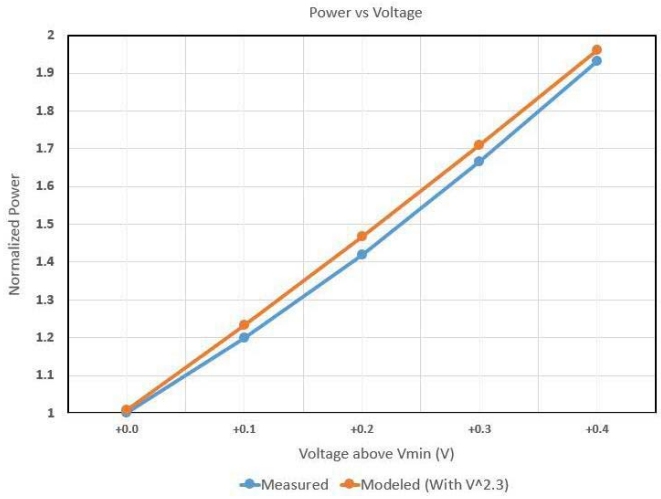


Fig. 15. Model to hardware correlation chip ac power versus voltage.

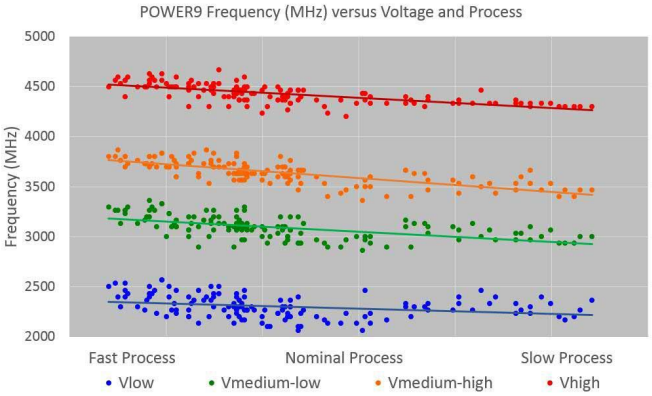


Fig. 16. POWER9 hardware frequency data.

(normalized to the power consumption at $V_{\min}$), as the operating voltage is increased in steps of 100 mV, at the same frequency of operation. The power consumption shows a greater than quadratic dependence on the supply voltage at nearly close to the $V^{2.3}$ coefficient used during modeling, due to the voltage dependent gate capacitance and the contribution of short-circuit currents. Finally, Fig. 16 shows POWER9 hardware measurement data. In addition to a top end frequency in excess of 4.5 GHz, full functionality across the required voltage and process range was achieved.

VIII. Conclusion

The POWER9 microprocessor includes an all-new core microarchitecture with advanced I/O topologies and demonstrates a new adaptive clocking technique coupled with improvements in resonant clocking and iVRM response time. Two core variants, SMT4 or SMT8, allow configurations tuned for Linux or PowerVM ecosystems and optimized for traditional and emerging workloads. An adaptive clocking loop improves power efficiency over previous microprocessor generations. Alternate I/O subsystems within the POWER9 family address both SO and SU needs, featuring CAPI2.0 over PCIe Gen4 at 16 Gb/s and OpenCAPI 3.0 over advanced 25-Gb/s links. Finally, hardware measurements on the SU chip yield 4.5-GHz core frequencies and 12.9 Tb/s of off-chip bandwidth.

References

[1] C.-H. Lin *et al.*, "High performance 14 nm SOI FinFET CMOS technology with 0.0174 μm^2 embedded DRAM and 15 levels of Cu metallization," in *IEDM Tech. Dig.*, Dec. 2014, pp. 3.8.1–3.8.3.

[2] B. Thompto, "POWER9: Processor for the cognitive era," in *Proc. Hot Chips*, Aug. 2016, pp. 1–19.

[3] E. Fluhr *et al.*, "POWER8: A 12-core server-class processor in 22 nm SOI with 7.6 Tb/s off-chip bandwidth," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 96–97.

[4] E. J. Fluhr *et al.*, "The 12-core POWER8 processor with 7.6 Tb/s IO bandwidth, integrated voltage regulation, and resonant clocking," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 10–23, Jan. 2015.

[5] P. Restle *et al.*, "Wide-frequency-range resonant clock with on-the-fly mode changing for the POWER8 microprocessor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 100–101.

[6] D. Shan *et al.*, "Resonant clock mega-mesh for the IBM z13," in *Proc. VLSI Circuits Symp.*, Jun. 2015, pp. C322–C323.

[7] R. Groves, P. Restle, A. Drake, D. Shan, and M. Thomson, "Optimization and modeling of resonant clocking inductors for the POWER8 microprocessor," in *Proc. CICC*, Sep. 2014, pp. 1–4.

[8] C. Takahashi *et al.*, "A 16 nm FinFET heterogeneous nona-core SoC complying with ISO26262 ASIL-B: Achieving 10-7 random hardware failures per hour reliability," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 80–81.

[9] K. A. Bowman, C. Tokunaga, T. Karnik, V. K. De, and J. W. Tschanz, "A 22 nm all-digital dynamically adaptive clock distribution for supply voltage droop tolerance," *IEEE J. Solid-State Circuits*, vol. 48, no. 4, pp. 907–916, Apr. 2013.

[10] D. Ernst *et al.*, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proc. IEEE Int. Symp. Microarchitecture*, Dec. 2003, pp. 7–18.

[11] D. Blaauw *et al.*, "Razor II: *In situ* error detection and correction for PVT and SER tolerance," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 400–401.

[12] A. Grenat, S. Pant, R. Rachala, and S. Naffziger, "Adaptive clocking system for improved power efficiency in a 28 nm x86-64 microprocessor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 106–107.

[13] A. Drake *et al.*, "A distributed critical-path timing monitor for a 65 nm high-performance microprocessor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 398–399.

Christopher Gonzalez received the B.S. degree in electrical and computer engineering from Rutgers University, New Brunswick, NJ, USA, and the M.S.E.E. degree from Columbia University, New York, NY, USA, in 2007.

He joined IBM in 2001. He is currently a Senior Engineer with IBM Systems Group. He is currently working on a future POWER processor. While working on POWER7, POWER8, and POWER9, he has held several lead roles in circuit design and power analysis. His current research interests include power-efficient and high-performance circuit design, power modeling, and high-performance microarchitectures.

Michael Floyd received the bachelor's (Hons.) degree in computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1995, and the master's degree from Stanford University, Stanford, CA, USA, in 2000.

He was with IBM Research pioneering the adaptive clocking techniques used on the POWER7, POWER7+, and z13 microprocessors. He has been with IBM Server development, Austin, TX, USA, since 1995. His previous experience with IBM server development includes hardware bring-up, test, debug, and reliability, availability, and serviceability (RAS) in addition to design, lead, and microarchitect roles on the POWER4, POWER5, and POWER6 processors and support chips. He is currently the Architect and leads the POWER9 EnergyScale design, a role he previously held on POWER7, and an IBM Master Inventor with more than 60 issued U.S. patents.

Eric Fluhr (M'14) received the B.S.C.S. and M.S.E.E. degrees from the Georgia Institute of Technology, Atlanta, GA, USA.

In 1996, he joined IBM. He began in circuit design on the POWER3 and POWER4 series processors. He switched to microarchitecture and logic design on POWER5, followed in POWER6 by load/store-unit circuit lead and chip characterization lead. He followed as POWER8 core circuit lead, also responsible for implementing statistical timing for IBMs 32- and 22-nm server processors, and is currently a Chip Circuit Lead for IBM's POWER9 family.

Phillip Restle (M'87–SM'14) received the Ph.D. degree in physics from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1986.

He is currently a Research Scientist with the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. He concentrates on tools and designs for VLSI clock distribution networks, contributing to many game and server processors, including Blue Gene/Q, the 5.5 GHz zEC12, and POWER4 through POWER9. He has co-authored 74 papers including two best paper awards, and holds 26 patents.

Dr. Restle was a recipient of the two IBM Corporate Awards for VLSI Clock Distribution and Methodology.

Daniel Dreps received the B.S.E.E. degree from Michigan State University, East Lansing, MI, USA, in 1983.

He is currently a Distinguished Engineer with the IBM Systems Group. During his IBM career, he has designed and developed transistor models, fiber optic links, ASIC technology custom elements, and high-speed serial links for IBM servers. He has authored multiple papers and holds more than 100 patents in broad areas of interconnect and server design. His current research interests include high-speed link development and applications in the entire range of IBM servers.

Michael Sperling received the B.S. and M.S. degrees in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2002 and 2003, respectively.

Since 2003, he has been with the IBM Corporation, Poughkeepsie, NY, USA, where he is involved in the Power family of servers with the System and Technology Group. His current research interests include circuit and architectural contributions to on-chip voltage generation and various analog sensors.

Rahul Rao (M'06–SM'12) was a Research Staff Member with the High Performance Circuit Design Group, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. He is currently a Senior Technical Staff Member with the Processor Circuit Design Team, IBM, Bangalore, India, and a member of the IBM Academy of Technology and an IBM Master Inventor with over 20 field patents and 30 conference and journal papers. He is currently the chip Power and the nest circuit lead with POWER10.

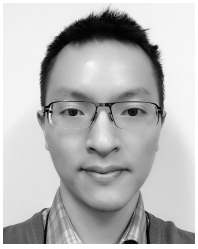
David Hogenmiller received the B.S.E.E. degree from Drexel University, Philadelphia, PA, USA, in 1990.

He started his career in custom circuit design on the early POWER processors. He was the Chip Clock Lead with the POWER 9 design. He has worked on processor clocking, both architecture and implementation, from the latches to the chip distribution.

Christos Vezyrtis received the B.Eng. degree from the National Technical University of Athens, Athens, Greece, in 2006, and the M.Phil., M.S., and Ph.D. degrees from Columbia University, New York, NY, USA.

He is currently a Research Staff Member with IBM T. J. Watson Research Center, Yorktown, NY, USA, and an Adjunct Assistant Professor with Columbia University. He has co-authored 12 patents. His current research interests include high-speed mixed-signal and digital circuits, as well as asynchronous circuits.

Dr. Vezyrtis was a recipient of the Best Paper Award for the logic and circuits track at the IEEE International Conference on Computer Design in 2006.



Pierce Chuang (M'10) received the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada.

Since 2014, he has been with the IBM T. J. Watson Research Center, VLSI Department, Yorktown Heights, NY, USA. His current research interests include power supply noise guard-band reduction techniques, and algorithm-architecture co-optimization to accelerate emerging cognitive workloads.

Daniel Lewis (M'07) received the B.E. degree in electronics and communication engineering from SJCE, Mysore, India.

In 2007, he joined IBM. He began his career as a Circuit Designer with Fabric Unit, Power 7 Microprocessor and then moved to Core and led the physical design efforts of Core Pervasive (PC) unit for Power 8 Microprocessor. For Power 9, he was a Circuit Designer for Load Store Unit before moving to the Laboratory for circuit characterization, where he was responsible for doing voltage and frequency characterization of the Power 9 chip. He is currently the Circuit Lead with the Fabric Unit, POWER processor design, IBM Bangalore.

Ricardo Escobar received the B.S. degree in electrical and computer engineering from the University of Texas at Austin, Austin, TX, USA.

In 2016, he joined IBM, where he was involved in POWER system characterization and processor failure analysis. He is currently a Hardware Developer with the IBM Systems Group.

Vinod Ramadurai received the bachelor's degree in electrical engineering from the University of Arizona, Tucson, AZ, USA, in 1999, and the master's degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 2001.

He is currently a Functional Manager of physical design with the Systems Group, IBM, Austin, TX, USA. Since 2001, he has been a Memory Circuit Designer on various System Z/P/ASICs chips with IBM. He was the Power Lead and most recently was responsible for power-frequency projections for the POWER 9 processor. He has co-authored five conference papers and is currently an IBM Master Inventor with more than 25 patents issued.

Ryan Kruse received the B.S. degree in electrical engineering from the New Mexico Institute of Mining and Technology, Socorro, NM, USA, in 2003, and the M.S. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2005.

He joined IBM Systems, Austin, TX, USA, where he was involved in server board design for POWER6-based systems. In 2006, he joined Sun Microsystems, where he was involved in EMIR and electrical analysis on SPARC series microprocessors. In 2009, he joined IBM, where he has worked on POWER7+, POWER8, and POWER9 microprocessors. He is currently a Senior Engineer with IBM, Austin, TX, USA, where he is involved in the POWER series chip integration and physical design space. He is currently involved in work on the next-generation POWER microprocessor.

Juergen Pille received the M.S. degree in microelectronics from Hanover University, Hanover, Germany.

He was the Array Lead of the CELL BE processor design, and is currently the high-speed SRAM Lead with POWER architecture server processors, IBM. He is a Senior Technical Staff Member with the IBM Systems Group. Since 1990, he has been with IBM, where he is involved in various microprocessor designs, SRAM arrays, and circuits.

Ryan Nett received the B.S. degree in computer engineering from the University of Maryland, College Park, MD, USA, and the M.S. degree in electrical and computer engineering from the University of Texas at Austin, Austin, TX, USA, in 2009.

In 2001, he joined IBM. From 2002 to 2006, he was with the CELL microprocessor, where he was involved in unit and chip physical design convergence. He was one of the Lead PD Integrators for the POWER9 family of chips. He is currently manages a small team of physical designers.

Mr. Nett was a recipient of the Outstanding Technical Achievement Awards for his PD convergence work on the POWER7 and POWER8 designs.

Pawel Owczarczyk received the B.S. degree in computer engineering from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2001.

He has been a part of the IBM analog circuit team. He is currently involved in several mixed signal clocking designs and all levels of integration solutions as part of the Systems and Technology Group.

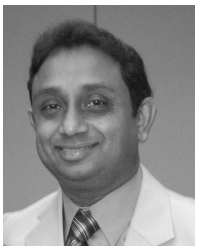
Joshua Friedrich received the B.S. degree in electrical engineering from the University of Texas at Austin, Austin, TX, USA.

He is currently a Distinguished Engineer and the Director of POWER Technology Development with IBM's Server and Technology Group. He is responsible for the design and delivery of IBM's POWER9 processor. He was the Circuit Design Lead for the POWER8 chip and has been part of the POWER Development Team since POWER4. He has led multiple design disciplines including power estimation and reduction, hardware characterization, memory subsystem circuit development, and core execution units.

Jose Paredes is currently a Senior Technical Staff Member with IBM Systems Group. His current research interests include high-performance microprocessor circuit and logic design, including custom digital circuits, SRAM design, CAM design, synthesis, and tools/utilities creation. He has held various circuit leadership jobs, including core circuit physical design leader, load/store unit physical design leader, recovery unit circuit leader, and mask layout design team leader.

Timothy Diemoz received the B.S.E.E. degree from Union College, Schenectady, NY, USA, in 1987.

He is currently an Advisory Engineer with IBM Systems Group. He is a member of the Advanced VLSI Circuit Design Team and owns test and characterization of phase-locked loops, thermal sensors, critical path monitors, and integrated voltage regulator circuits.



Saiful Islam received the bachelor's degree in electrical engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 1987, and the M.S.E.E. and Ph.D. degrees from The University of Texas at Austin, Austin, TX, USA, in 1990 and 1994, respectively.

He was with AMD as a Circuit Designer in 1995. In 2003, he joined the IBM POWER6TM Team, where he has been involved in register files and custom circuits in both P and Z processors. He led the Custom Register File Team for POWER8, POWER9, and z360 processors. He is currently a Functional Manager for the Austin Array Team, involved in both P and Z processors.

Donald Plass is a Distinguished Engineer in IBM Enterprise Systems and Technology Development. He has been responsible for RAM technology and designs for many generations of IBM servers, including both pSeries* and zSeries*, with a focus on the larger arrays. He joined IBM in 1978 at the Poughkeepsie facility. In addition to CMOS SRAM and DRAM, his research and development interests have included gallium arsenide (GaAs), and BiCMOS. His more recent accomplishments include bringing SOI eDRAM and dense SRAMs to the product level for the 22nm and 14nm microprocessors.

Paul Muench received the B.S. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1984.

He was involved in analog circuits during his time at IBM including phase locked loops, voltage regulation and I/O circuitry as part of the Systems and Technology group.