## 4.8 A 28nm x86 APU Optimized for Power and Area Efficiency

Kathryn Wilcox[1], David Akeson[1], Harry R. Fair III[1], Jim Farrell[1], Dave Johnson[2], Guhan Krishnan[1], Hugh McIntyre[3], Edward McLellan[1], Samuel Naffziger[2], Russell Schreiber[4], Sriram Sundaram[4], Jonathan White[1]

[1]AMD, Boxborough, MA, [2]AMD, Fort Collins, CO, [3]AMD, Sunnyvale, CA, [4]AMD, Austin, TX

Carrizo (CZ, Fig. 4.8.7) is AMD's next-generation mobile performance accelerated processing unit (APU), which includes four Excavator (XV) processor cores and eight Radeon™ graphics core next (GCN) cores, implemented in a 28nm HKMG planar dual-oxide FET technology featuring 3 $V_t$s of thin-oxide devices and 12 layers of Cu-based metallization. This 28nm technology is a density-focused version of the 28nm technology used by Steamroller (SR) [1] featuring eight 1× metals for dense routing, one 2× and one 4× for low-RC routing and two 16× metals for power distribution.

At 250.04mm$^2$, CZ fits 29% more transistors (3.1 billion) into a die with approximately the same area footprint as the Kaveri APU (KV) [2]. Excluding the L2 cache, XV consumes 14.48mm$^2$ with a total transistor count increase to 102M compared to 86M in SR. Much of the XV growth was due to IPC improvements, including doubling the data cache from 16KB to 32KB per core.

The density increase enabled a larger area allocation for graphics, multi-media offload and integration of a system controller into a single BGA package. Increased area in the graphics IP has allowed CZ to enable a fully heterogeneous system architecture (HSA)-compliant part and the increased area in multimedia has enabled a new high-performance video decoder and double the video compression engines compared to KV. This multimedia area increase has allowed CZ to transcode 9 real-time 1080p video streams, a 3.5× improvement over KV.

CZ moves the eight GCN cores to a separate conditionally enabled power supply (VDDGFX in Fig. 4.8.2) to enable the graphics core to operate at the optimal voltage, which gives a 20% improvement over KV with six GCN cores. The XV design moved to a high density, 9-track standard cell library that was more consistent with AMD's graphics IP library. This provided significant area and power reduction, similar to a technology shrink, while staying in a 28nm process (Fig. 4.8.1). Early experiments showed an average area reduction of approximately 24% and a frequency impact of ~10% at constant voltage across designs. At constant power, frequency is increased because the design power reduction allows a higher voltage. The smaller XV core poses the challenge of higher thermal density when consuming the same power as its predecessor. In addition to the lower leakage profile of the technology, CZ mitigates the higher thermal density through floorplan placement of the cores away from the edge of the die and away from IPs that have high power densities (Fig. 4.8.2).

The L2 cache contains two 6T macros, which employ read/write-assist techniques to improve $V_{min}$. Timing differences between the two macros (phase-bound wordline vs. cycle-bound wordline) require different write-assist techniques. The L2 data macro asserts the wordline for a full cycle and uses a combination wordline underdrive/wordline boost for read/write assist. The wordline uses a voltage somewhat lower than $V_{DD}$ for the first phase of the access. This gives the bitlines of half-selected cells time to discharge enough to reduce susceptability to read disturb before returning the wordline to full $V_{DD}$. During the second phase of the access, the pFET pulldown is turned off, allowing the wordline to reach full $V_{DD}$. After the wordline returns to $V_{DD}$, the shared power header for a set of 16 wordlines is turned off and the virtual supply is boosted above $V_{DD}$ via an nFET used as a capacitor. Turning off the pFET header turns on an nFET to ensure the wordline never leaks further than a $V_t$ below $V_{DD}$. The circuit (Fig. 4.8.3) can be configured via BOOSTEN and WLUDCTL to allow any combination of the following: 1) first-phase underdrive, 2) no underdrive with a second-phase boost, 3) continued underdrive, 4) return to full $V_{DD}$.

The L2 tag macro uses a combination of wordline underdrive/negative bitline for its assist techniques due to the phase-bound wordline of the tag, which does not leave adequate time for the bitlines of half-selected cells to discharge prior to the write assist technique starting. The negative-bitline circuitry (Fig. 4.8.3) uses a single cap per logical bit column rather than a traditional two. The capacitor is coupled to the bitline via an nFET passgate using the same signal that controls the write driver's pFET pull up. The circuit drives the bitlines and asserts NEGBLEN after a self-timed delay tuned to align the negative bitline event after the bitline is fully at ground. When NEGBLEN is asserted, it deasserts the signal driving the gate of the nFET pulldown of the driver and an OR gate delay later, driving the bitline below ground via an nFET capacitor. Since the two techniques used extend the voltage beyond the $V_{DD}$-$V_{SS}$ operating window, care was taken to avoid damaging the devices due to excessive voltage. A microcode-controllable signal called superVminEnable turns the assist functions off at high voltage.

XV supports AMD's first implementation of adaptive voltage-frequency scaling (AVFS), which has similar characteristics to other adaptive voltage approaches [3,4,5], with two key improvements: 1) infrastructure to enable replica paths to function as a statistical sample of the full set of $F_{max}$ limiting paths resulting in improved tracking, and 2) explicit disambiguation of the voltage impact on delay from the intrinsic circuit speed by coupling path margin assessments with a voltage reading from the integrated power supply monitors (PSMs). The key motivation is to provide an accurate assessment of intrinsic silicon speed capability across process, voltage and temperature. AVFS allows each part to self-calibrate and determine the optimal voltage for current operating frequency and conditions. Benefits include removing voltage uncertainty and guard-bands, common in traditional test and binning flows, and the potential to minimize/eliminate costly system level test in production.

To achieve self-calibration, AVFS relies on a set of timing-critical replica paths. Core gate-dominated, wire-dominated and macro (cache array) critical paths are used. AVFS employs a special shadow flop that compares the outputs of a native replica path with its data-delayed version (Fig. 4.8.4). The shadow flop monitors each path for a late transition, indicating a "near miss" in timing. A critical path accumulator (CPA) steps through a programmable delay element and collects near miss information. During operation, near misses accumulated by the CPA across voltage-frequency-temperature (VFT) are sent to the on-chip system-management unit (SMU). The SMU summarizes the statistics of the data and creates a VFT table, essentially the part-specific optimal voltage for any frequency-temperature combination. P-state changes during normal operation reference the VFT table to determine the optimal voltage.

Robust sampling of the chip-wide critical path variations is provided by 10 dispersed instances of the CPA. Each CPA exercises 50 critical paths for a total of 500 paths (300 gate-dominated, 100 wire-dominated and 100 macro replica paths). AVFS extracts Gaussian distribution statistics for the paths and infers the timing margin for the actual core path using sampling statistics (Fig. 4.8.5). Gates, wires, and macros are treated separately to differentiate the distribution, and appropriate guard-bands added for sampling uncertainty. Timing-margin prediction vs. actual timing margin indicates the ability of AVFS to set the minimum voltage required across the entire voltage range, resulting in up to 30% power savings.

As mentioned, the system can automatically adjust for local voltage noise by having the CPAs use the PSM indicated voltage to compensate for the voltage difference (Fig. 4.8.6) by modifying the critical-path timing margin assessment. The AVFS system can be triggered either by microcode or the SMU and is transparent to normal core operation. The full implementation cost is under one percent of the core area.

XV achieves the program goals by reducing the power by 40% and area by 23% compared to SR. By reducing the overall core power consumption across all voltages (from $V_{min}$ to $V_{max}$), XV shifts its operating power range lower at the expense of frequency scaling at higher power ranges. This tradeoff is balanced to allow CZ to get power-constrained frequency uplifts on single, multi-threaded and gaming use cases for mobile platforms designed around a 12-to-35W SoC. The area savings has been instrumental to CZ to be cost neutral and repurpose the area towards enhancing gaming and multimedia features.

*References*:

[1] K. Gillespie, *et al.*, "Steamroller: An x86-64 core implemented in 28nm bulk CMOS", *ISSCC Dig. Tech. Papers*, pp. 104-105, Feb. 2014.

[2] D. Bouvier, *et al.*, "Applying AMD's "Kaveri" APU for Heterogeneous Computing", *Hot Chips*, 2014.

[3] M. Floyd, *et al.*, "Introducing the Adaptive Energy Management Features of the Power7 Chip", *IEEE Micro*, vol. 31, no. 2, pp. 60-75, 2011.

[4] D. Blauuw, *et al.*, "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance", *ISSCC Dig. Tech. Papers*, pp. 400-401, Feb. 2008.

[5] K. Bowman, *et al.*, "Resilient Microprocessor Design for Improving Performance and Energy Efficiency," *IEEE International Conf. on Computer-Aided Design*, pp. 85-88, 2010.

 978-1-4799-6224-2/15/$31.00 

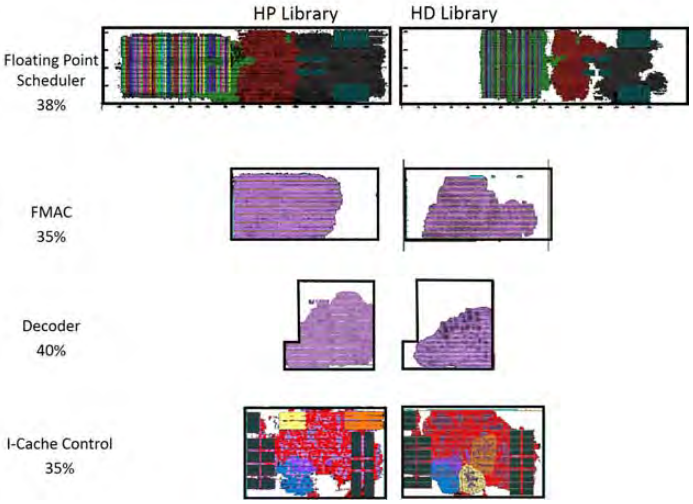



Figure 4.8.1: High-performance vs. high-density cell examples.

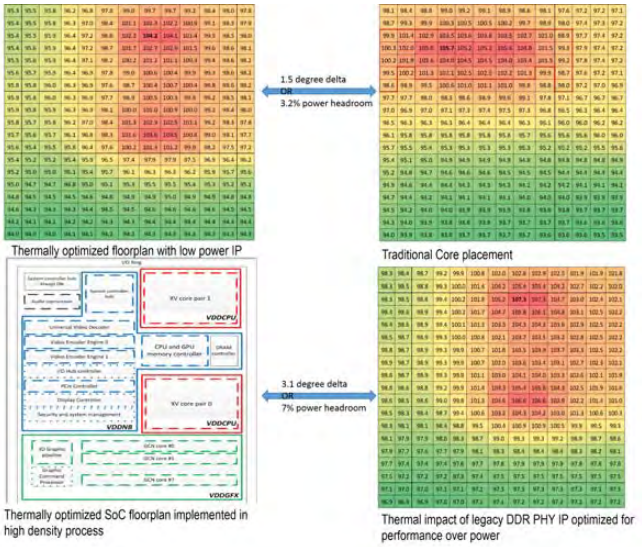



Figure 4.8.2: Temperature differentials and CZ optimized floor plan.

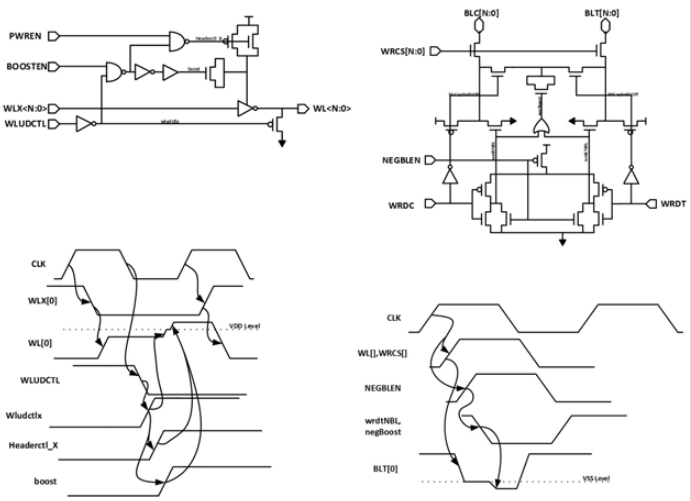



Figure 4.8.3: L2 Data and L2 Tag macro assist.

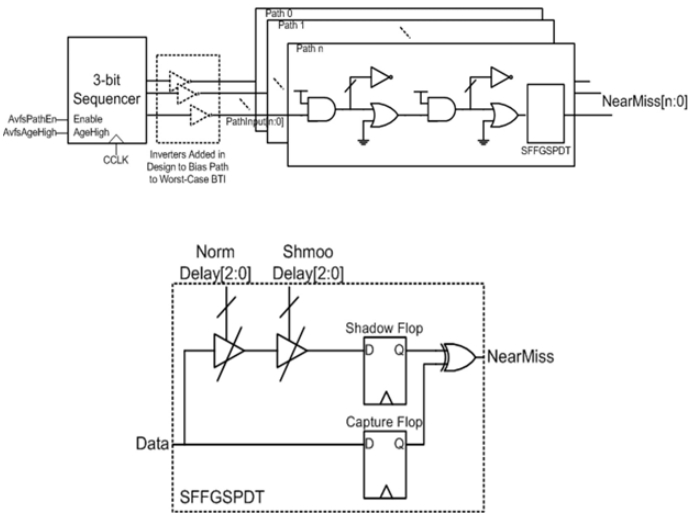



Figure 4.8.4: Shadow flop and replica-path schematic.

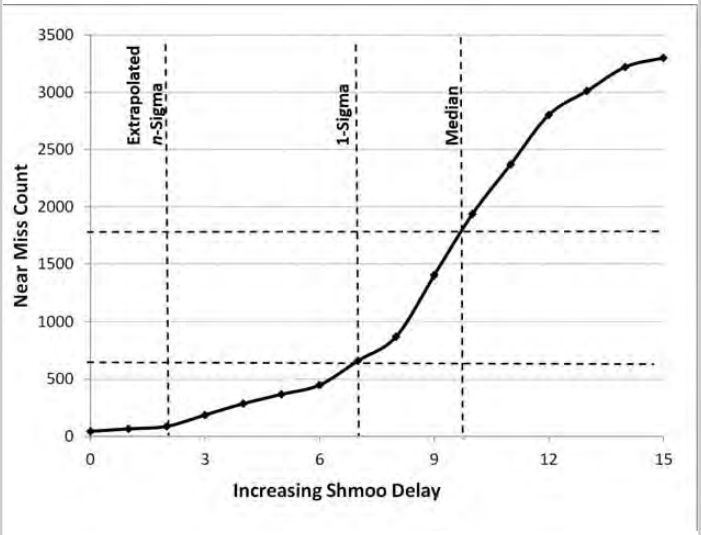



Figure 4.8.5: Near miss statistics and extrapolation to estimate timing margin.

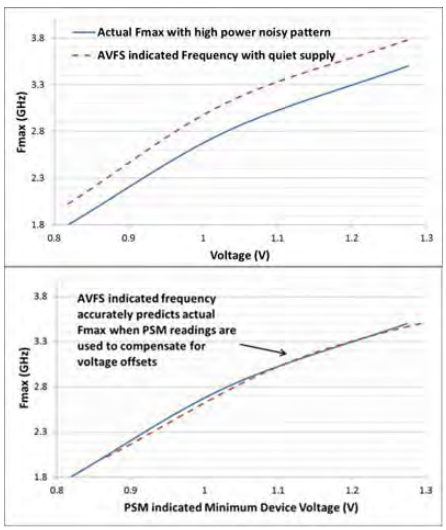



Figure 4.8.6: Measured AVFS results with sampling-theory-based $F_{max}$ prediction and self-measured voltage compensation.

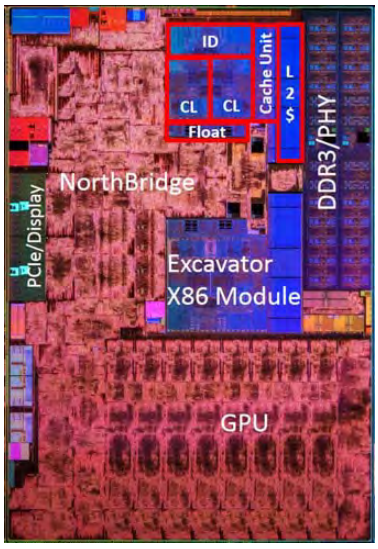
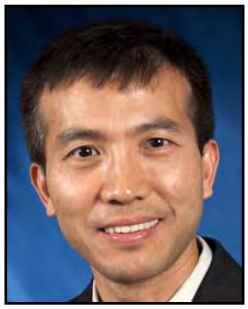



**Figure 4.8.7: Carrizo die photo.**


978-1-4799-6224-2/15/$31.00 

# Session 5 Overview: *Analog Techniques*

## ANALOG SUBCOMMITTEE



**Session Chair:** *Xicheng Jiang, Broadcom, Irvine, CA*

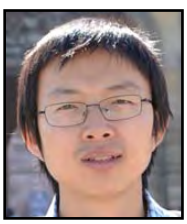
**Session Co-Chair:** *Ed van Tuijl, University of Twente, Enschede, The Netherlands*

This year's session on Analog Techniques continues to defy simple categories. This session illustrates the diversity and vigor of modern analog circuitry. The rise of wearable devices and Internet of Things (IoT) leads to the emergence of nano-power designs of references, oscillators and many other blocks. New frontiers of precision, power, and performance are established.

**5.1 A 60V Auto-zero and Chopper Operational Amplifier with 800kHz Interleaved Clocks and Input Bias-Current Trimming** **1:30 PM**

*Y. Kusuda*, Analog Devices, San Jose, CA

Paper 5.1 from Analog Devices presents an auto-zero and chopper operational amplifier with 6.8nV/√Hz PSD in 0.18µm BCDMOS. The chopper op-amp employs 800kHz interleaved clocks, moving the majority of the switching PSD up to 4.8MHz, above the op-amp's unity gain frequency of 3.2MHz. The maximum input bias current is 200pA. It achieves 152dB CMRR and 1.2V/µs slew rate.



**5.2 A 110dB SNR ADC with ±30V Input Common-Mode Range and 8µV Offset for Current Sensing Applications** **2:00 PM**

*L. Xu,* Delft University of Technology, Delft, The Netherlands

Paper 5.2 by Delft University of Technology presents a ΔΣ ADC with an improved capacitively-coupled high voltage chopper. It achieves ±30V input common-mode voltage range, 110dB SNR and 8µV maximum offset. The sensing AFE consumes 505µW from a 5V supply.

**5.3 A 2-Channel -83.2dB Crosstalk 0.061mm² CCIA with an Orthogonal Frequency Chopping Technique** **2:30 PM**

*Y-L. Tsai,* National Taiwan University, Taipei, Taiwan

Paper 5.3 by National Taiwan University presents a continuous two-channel capacitively-coupled instrumentation amplifier using an orthogonal frequency chopping technique and only one active amplifier. It achieves 26nV/√Hz input-referred noise, -83.2dB crosstalk and consumes 27µA from a 3V supply.

**5.4 A 32nW Bandgap Reference Voltage Operational from 0.5V Supply for Ultra-Low Power Systems** **2:45 PM**

*A. Shrivastava,* PsiKick, Charlottesville, VA

Paper 5.4 by PsiKick Inc. presents a 32nW bandgap reference from a 0.5V supply voltage. It achieves 75 ppm/°C over a temperature range of 0 to 80°C and 2% untrimmed 3σ process variation.

 978-1-4799-6224-2/15/$31.00 

**5.5 A Forward-Body-Bias Tuned 450MHz Gm-C 3rd-Order Low-Pass Filter in 28nm UTBB FD-SOI with >1dBVp IIP3 over a 0.7-to-1V Supply** **3:15 PM**

*J. Lechevallier,* University of Twente, Enschede, The Netherlands and STMicroelectronics, Crolles, France

Paper 5.5 by University of Twente presents a 450MHz inverter based, 3rd-order Butterworth Gm-C filter in 28 nm UTBB FD-SOI technology. The cut-off frequency and filter shape are kept constant over a supply ranging from 0.7V-to-1V, while maintaining linearity above 1dBV without any requirement for Q-tuning and supply voltage tuning. It achieves 6.1nV√Hz input noise, and dissipates less than 5.6mW.

**5.6 A 0.13µm Fully Digital Low-Dropout Regulator with Adaptive Control and Reduced Dynamic Stability for Ultra-Wide Dynamic Range** **3:45 PM**

*A. Raychowdhury,* Georgia Institute of Technology, Atlanta, GA

Paper 5.6 by Georgia Institute of Technology presents a fully digital LDO that uses adaptive control and fine grained clock gating to enable >90% current efficiency across a 50× load current range. The design also introduces a technique to enable 8× reduction in transient time in response to large voltage droops.

**5.7 A 29nW Bandgap Reference Circuit** **4:15 PM**

*J. M. Lee,* Pohang University of Science and Technology, Pohang, Korea

Paper 5.7 by Pohang University of Science and Technology presents a bandgap reference circuit with a PTAT generated by leakage current. No start-up circuit is needed. The BGR is fabricated in 0.35µm CMOS and consumes 29nW from a 1.4V supply.

**5.8 A Digitally Assisted Single-Point-Calibration CMOS Bandgap Voltage Reference with a 3σ Inaccuracy of ±0.08% for Fuel-Gauge Applications** **4:30 PM**

*G. Maderbacher,* Infineon Technologies, Villach, Austria

Paper 5.8 by Infineon Technology presents a digitally assisted single-point-trim CMOS bandgap reference. The key idea is to keep the analog bandgap core simple and only compensate non-PTAT related effects by using chopping techniques. A 3σ inaccuracy of ±0.08% from –40°C to +120°C is achieved. The temperature drift is 7ppm/°C.

**5.9 A 37µW Dual-Mode Crystal Oscillator for Single-Crystal Radios** **4:45 PM**

*D. Griffith,* Texas Instruments, Dallas, TX

Paper 5.9 by Texas Instruments presents a 24MHz crystal oscillator and a 31.25kHz derived sleep timer for a single crystal wireless node application. A dedicated state machine enables mode transitions without losing clock pulses. It achieves 9ppm/V of frequency variation with voltage and consumes 37µW in sleep mode.

**5.10 A 4.7MHz 53µW Fully Differential CMOS Reference Clock Oscillator with –22dB Worst-Case PSNR for Miniaturized SoCs** **5:00 PM**

*J. Lee,* Institute of Microelectronics, Singapore, Singapore

Paper 5.10 by Institute of Microelectronics Singapore presents a 4.7MHz CMOS reference clock oscillator for SoCs with severely digital noise contaminated supply and ground. It employs a fully differential supply- and ground-regulating frequency-locked loop architecture, a differential period detector with a supply-insensitive period reference, and a differential integrator generating a virtual 0V reference. It achieves a worst-case power supply noise rejection of -22dB and consumes 53µW from a 1.4V supply.