AMD Ryzen 7040 Series

Mahesh Subramony, David Kramer, and Indrani Paul, *AMD, Austin, TX, 78735, USA*

The AMD Ryzen 7040 Series processors are designed to strike the right balance between domain-specific accelerators and general-purpose compute. The system on chip features "Zen 4," RDNA 3, AMD XDNA architecture, and additional accelerators with a focus on delivering power-efficient performance.

Modern mobile PCs, given their acute focus on power efficiency, are fast reaching an inflection point where the significance of domain-specific accelerators is as important as general-purpose compute. Although traditional compute (CPUs and GPUs) is still a key driver of the mobile user experience, achieving high levels of power efficiency for domain-specific tasks requires the right mix of accelerators in the design (see Figure 1). Select AMD Ryzen 7040 Series processors introduce the first dedicated artificial intelligence (AI) engine inference processing unit (IPU) on an x86 processor,[a] and improved accelerators for evolving usage scenarios, which results in energy-efficient performance gains.[b] With the right balance of traditional compute and accelerators, an intelligent power management framework and an effective memory/input–output (I/O) subsystem, the Ryzen 7040 Series processors deliver exceptional performance for an ultrathin mobile device.[c]

[a]PHX-3: As of May 2023, AMD has the first available dedicated AI engine on an x86 Windows processor, where "dedicated AI engine" is defined as *an AI engine that has no function, other than to process AI inference models, and is part of the x86 processor die*. For detailed information, visit https://www.amd.com/en/products/processors/consumer/ryzen-ai.html.
[b]GD-220: Ryzen AI technology is compatible with all AMD Ryzen 7040 Series processors except the Ryzen 5 7540U and Ryzen 3 7440U; original equipment manufacturer enablement is required. Please check with your system manufacturer for feature availability prior to purchase.
[c]PHX-10: based on testing by AMD as of 23 December 2022. Testing results versus Apple are demonstrated in DaVinci Resolve BlackMagic, V-Ray, Blender, Cinebench R23 nT, and Handbrake 1:5:1. The testing results versus Intel are demonstrated in Cinebench R23, PCMark 10, Passmark 10, Kraken, 7Zip, Hand Brake, Blender. The Ryzen 9 7940HS system: The AMD reference motherboard is configured with a 4 × 4-GB LPDDR5, 1-TB solid-state drive (SSD), and Radeon 780 M graphics on a 64-bit Windows 11 Pro. The Apple M1 Pro system: The MacBook M1 Pro 18 is configured with a 32-GB LPDDR5 on a 1-TB SSD MacOS Monterey (12.6.1). The latest available 12th-generation Intel Core i9 12900HK: A Dell XPS 15, with GeForce RTX 3050Ti graphics, 16-GB random-access memory (RAM), and Samsung 1-TB SSD NVMe on a Windows 11 Pro. System manufacturers may vary configurations, yielding different results.

0272-1732 © 2024 IEEE
Digital Object Identifier 10.1109/MM.2024.3394479
Date of publication 29 April 2024; date of current version 6 June 2024.

"PHOENIX" SYSTEM ON CHIP

Figure 2 shows the "Phoenix" system on chip (SoC), identifying the major components that make up the design. At the heart of the SoC is the up to eight-core "Zen 4" core complex (CCX), with each core capable of running in single-threaded or simultaneous multithreaded (SMT) mode. Each core instance has 1 MB of private L2, doubling that of the previous generation. The CCX employs a 16-MB shared, unified level-3 (L3) cache, which is a 16-way set associative, and is populated by the L2 cache victimizations, both the clean and dirty lines. The CCX is connected to the Infinity Fabric with a 32 B per-cycle interface for both reads and writes. The RDNA 3 graphics (GFX) engine consists of a single shader engine (SE); two shader arrays (SAs), each with up to three physical workgroup processors (WGPs); and two render back ends (RBs). The engine features a 2-MB shared graphics level-2 cache (GL2). In total, the SoC features a 1-SE, 2-SAs, 6-WGPs, 4-RBs, 2-MB GL2 graphics engine and it is connected to the Infinity Fabric via four 32 B per-cycle data ports for increased bandwidth. The third compute pillar of the SoC is the IPU based on the XDNA architecture that consists of a 5 × 4 mesh of AIE2 digital signal processors (DSPs) with memory tiles for shared memory processing between them. It is connected to the Infinity Fabric with a 32 B per-cycle data port.

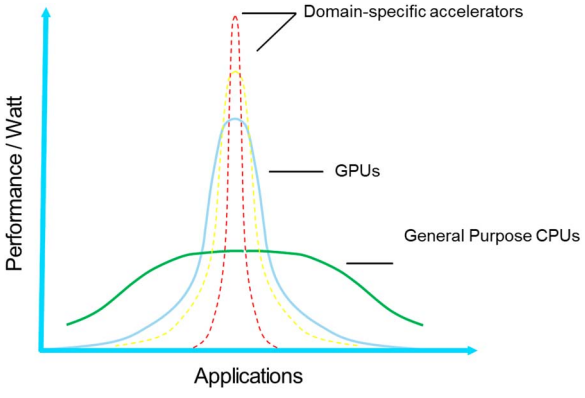

FIGURE 1. General-purpose compute versus domain-specific accelerators (in performance per watt and breadth of application).

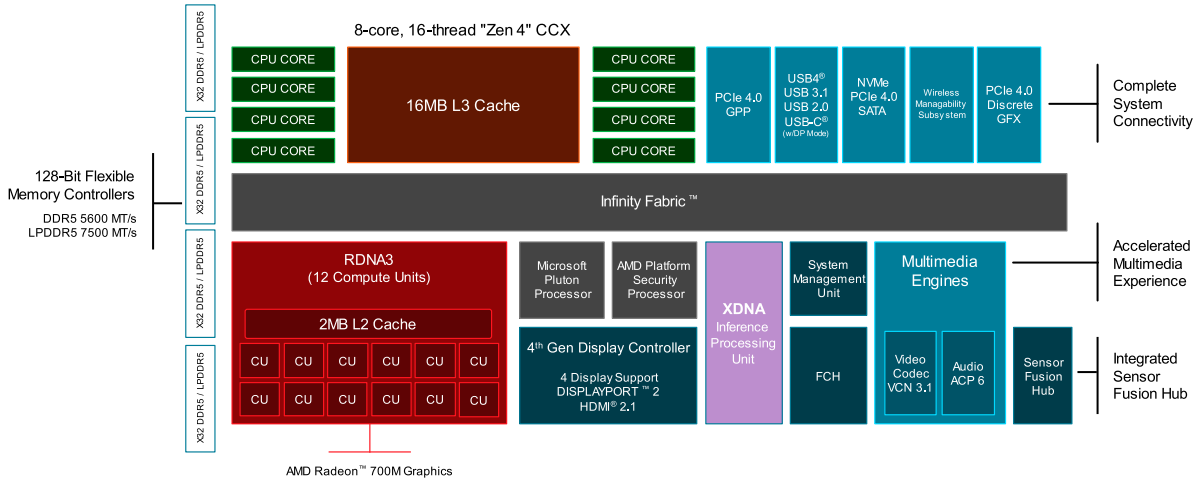

FIGURE 2. "Phoenix" SoC block diagram. L2: level 2; L3: level 3; PCIe: PCI Express; CUs: compute units; HDMI: High-Definition Multimedia Interface; SATA: serial advanced technology attachment; GPP: general purpose ports; FCH: fusion controller hub; VCN: video encode engine; GFX: graphics; ACP: audio coprocessor.

The display engine consists of four display pipes supporting a maximum of four independently timed displays simultaneously. The system management unit consists of a cluster of micro controllers that orchestrate various key functions within the SoC like security, power management, wireless manageability, and core frequency management. The SoC supports 20 lanes of Gen4 PCIe and a USB subsystem that supports eight ports with a combination of USB4, USB3 Type-C, Type-A, and USB2.0 ports. The memory subsystem consists of four unified memory controllers each driving a 32 b memory interface or channel. They support both LPDDR5 and DDR5 memories with data rates up to 7500 MT/s and 5600 MT/s, respectively. The dataflow and coherency between the various blocks and the memory is managed by the Infinity Fabric with improved quality-of-service (QoS) features and a system probe filter for reduced probe traffic.

ENERGY EFFICIENCY WITHIN BUDGET

Area is a key constraint for all SoC designs, and its significance is elevated in mobile SoCs due to the impact it has on power efficiency. The SoC employs a 15-metal-layer (ML) stack of the Taiwan Semiconductor Manufacturing Company (TSMC) N4P technology, which is designed for density on the lower layers and speed on the upper layers. As a result, "Phoenix" was able to achieve higher performance efficiencies and fit within the bounding box of power, package fit, and cost. Although the "Phoenix" SoC contains nearly twice the number of transistors compared to the previous-generation "Rembrandt" SoC, it managed to achieve this in a 15% smaller die footprint. This allowed the "Phoenix" SoC to fit into the same package size as that of the previous generation (25 mm × 35 mm) and still deliver significant performance and efficiency gains.

AREA IS A KEY CONSTRAINT FOR ALL SoC DESIGNS, AND ITS SIGNIFICANCE IS ELEVATED IN MOBILE SoCs DUE TO THE IMPACT IT HAS ON POWER EFFICIENCY.

PERFORMANCE AT A GLANCE

By driving higher instructions per cycle and pushing single-threaded boost frequency [peak (maximum) frequency] through a combination of clocking, thermal, and power management algorithm techniques, the "Phoenix" SoC delivers up to a 16% single-threaded performance increase compared to that of the previous generation. By driving the dynamic (switching) and static (leakage) power lower and using the tailwind provided by the technology shrink, the SoC delivers up to a 15% higher multithreaded performance over that of the previous generation. This addresses the traditional compute needs of a mobile PC, like gaming, creative, and scientific endeavors. Power-aware design in the rest of the chip and efficiency improvements in the RDNA 3 design

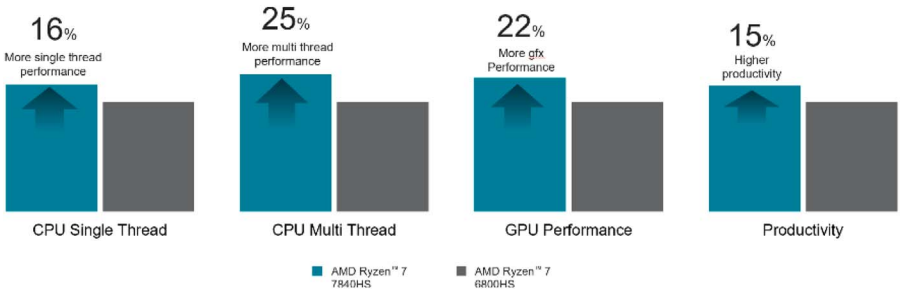

FIGURE 3. Ryzen 7040 Series generational performance uplifts.

led to a higher frequency at a given voltage, which enabled up to a 22% higher generational uplift in graphics (GFX) performance (see Figure 3).[d]

"ZEN 4" CORE ARCHITECTURE

Figure 4 shows the block diagram of the "Zen 4" architecture. The design includes significant improvements to the front end which consists of the branch predictor and the OPcache. Improved branch predictor accuracy, larger branch target buffers, and a 68% larger OPcache allow the design to get more instructions in the pipeline faster to leverage the wider design. In the execution engine, deeper buffers and structures for the out-of-order window allow for extraction of more instruction level parallelism in single-threaded code and drive higher throughput in SMT mode. The 25% larger instruction retire queue and larger integer/floating point (int/FP) register files facilitate this. In the load/store unit, a larger load queue and larger data translation lookaside buffers (DTLBs) allow leveraging the "Zen 3" throughput while increasing the data cache's efficiency, reducing translation look-aside buffer misses and increasing outstanding cache requests. "Zen 4" features a fast, private 1-MB L2 and supports more outstanding misses from L2 to L3 per core and from L3 to memory. The improvements in the cache hierarchy enable larger footprint workloads in L2. The floating-point unit adds power-efficient support for 512-bit advanced vector extension floating-point instructions using a 256-bit data path.

RDNA 3 ARCHITECTURE

The RDNA 3 (see Figure 5) features the compute unit pair or WGP, which delivers a significant architectural improvement clock for clock. The design was architected for higher frequencies, which drives a peak performance of 8.6 teraflops (FP16). RDNA 3 has enhanced ray-tracing acceleration with wider nodes and shallower trees. Enhancements in the Shader core including software-managed single-instruction, multiple-data (SIMD) hazard avoidance helped drive higher performance efficiency. An optimized and rebalanced memory hierarchy, along with the higher double data rate (DDR) bandwidth provided by the higher data rates, drive significantly higher performance. Finally, significant advancements in the geometry engine and the pixel pipeline, including variable-rate shading improvements, drive higher performance per bit. With all the aforementioned enhancements, in an ISO-configuration comparison with the previous-generation RDNA 2 engine featured in the Ryzen 6000 Series processor, the 7040 Series delivers double-digit performance improvements across a variety of games.[e]

AMD XDNA ARCHITECTURE

The AMD XDNA architecture (see Figure 6) is a highly scalable spatial dataflow design, consisting of a tiled array of SIMD/very long instruction word custom-designed processors, bringing significantly higher compute density for accelerating deep neural network (DNN) workloads. Each AI engine tile includes a vector

[d]PHX-18: Testing as of February 2023 by AMD Performance Labs using Cinebench R23, 3Dmark Night Raid, PCMark 10 app start, Procyon, Kraken System configuration for a Ryzen 7 7840HS: AMD reference system, 16-GB RAM, Samsung 980 Pro 1TB NVMe, and AMD Radeon 780 M integrated graphics on a Windows 11 Pro. The system configuration for a Ryzen 7 6800HS: an HP EliteBook 845, AMD Radeno 680 M integrated graphics, 16-GB (2 × 8) 4800-MHz RAM, and 1-TB SSD on a Windows 11 Pro. PC manufacturers may vary configurations, yielding different results.

[e]PHX-29: Testing as of April 2023 by AMD Performance Labs using 3DMark Timespy, and the following game titles tested at 1080P on the lowest settings: F1 2021; Far Cry 6; CS:GO; Grand Theft Auto V; World of Tanks Encore; Assassin's Creed Valhalla, Borderlands 3; DOTA 2; League of Legends; Total War: Three Kingdoms Battle; Shadow of the Tomb Raider; Final Fantasy XIV; and Strange Brigade. The configuration for an AMD Ryzen 7 7840U: an AMD Mayan reference board, 16-GB RAM, 1-TB SSD, BIOS RMH0081cA, and integrated Radeon 700 M graphics on a Windows 11 Pro.

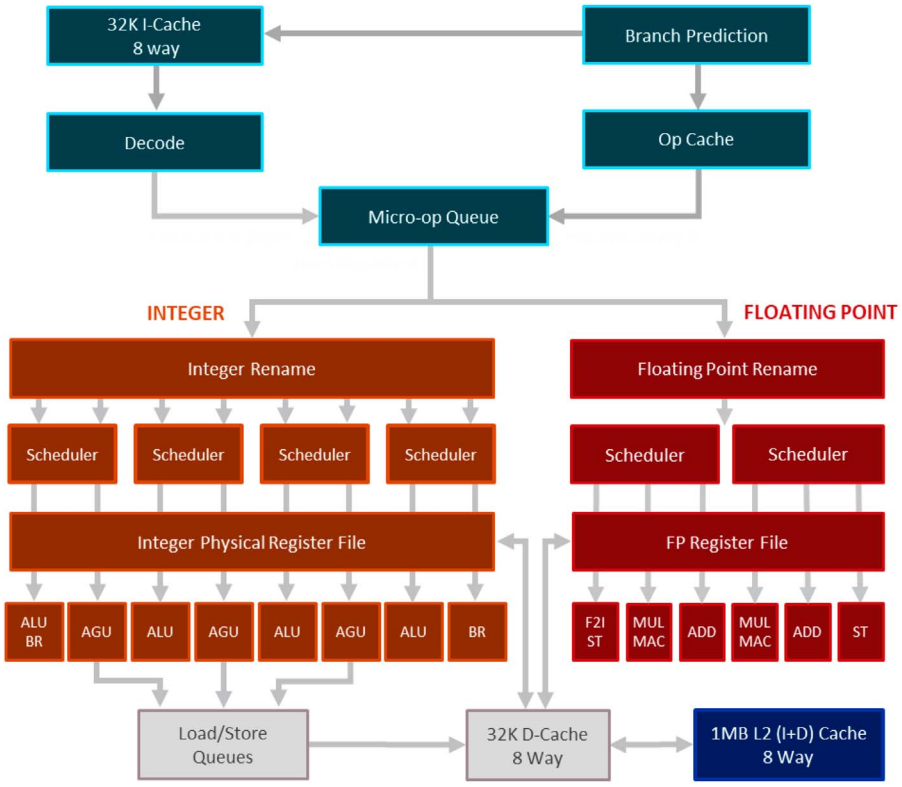

FIGURE 4. "Zen 4" architecture. MAC: media access control; ALU: arithmetic logic unit; AGU: address generation unit; F2I ST: transfer of data or flags from FP to EX; MUL: multiply; ADD: add; I+D Cache: instruction + data cache.

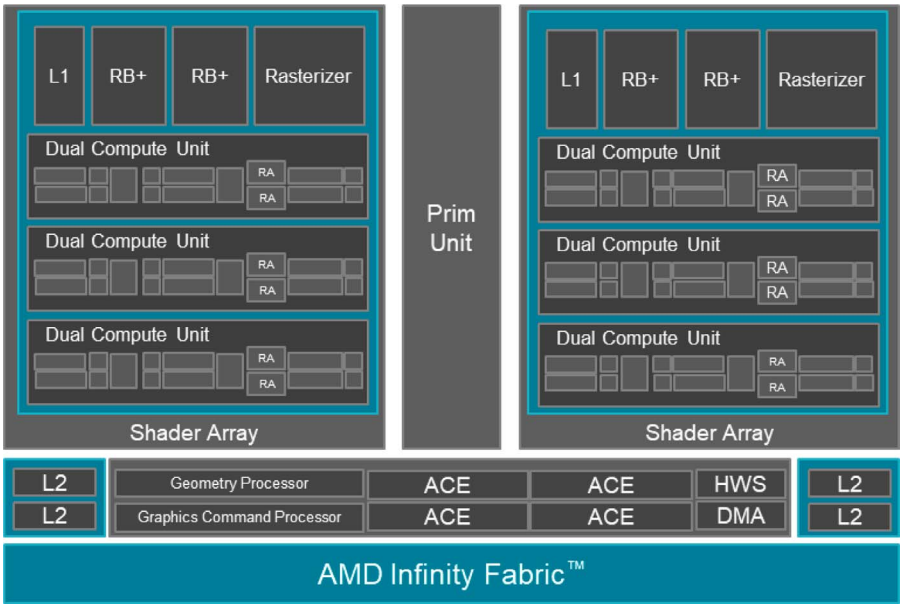

FIGURE 5. RDNA 3 architecture. Prim Unit: primitive assembly unit; RA: resource arbiter; ACE: asynchronous compute engine; HWS: hardware scheduler.

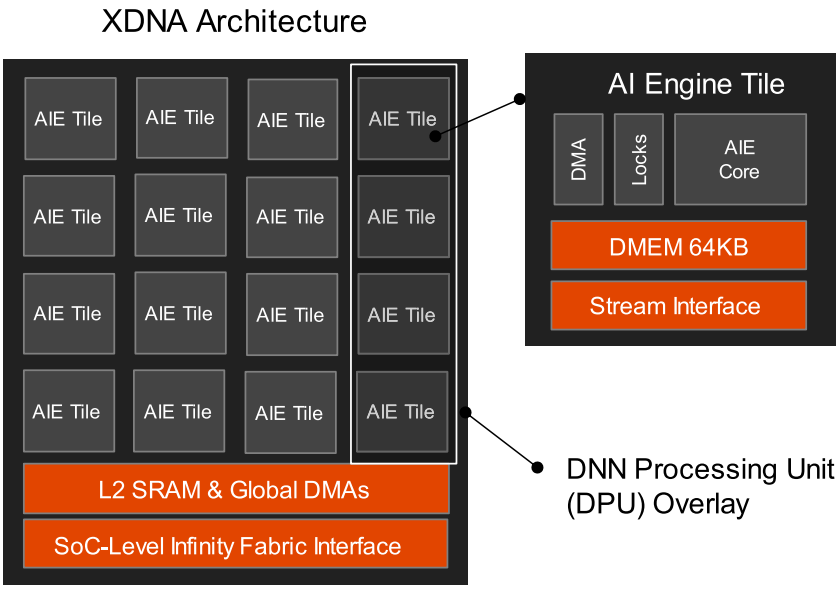

FIGURE 6. XDNA architecture. SRAM: static RAM; DPU: DNN processing unit; AIE: artificial intelligence Engine; DMEM: data memory.

processor for both fixed- and floating-point general matrix multiply (GEMM) and vector operations as well as nonlinear functions, a scalar processor, dedicated program (16-KB) and data memory (64-KB), dedicated advanced extensible interface (AXI) data movement channels, support for direct memory access (DMA) and hardware-managed locks. Optimized for real-time artificial intelligence/machine learning (AI/ML) computation, the AI engine array provides deterministic timing for use in video/audio ML applications as well as the modern and evolving generative AI workloads.

AI ENGINES ARE INHERENTLY SOFTWARE PROGRAMMABLE AND HARDWARE ADAPTABLE TO SUPPORT A WIDE RANGE OF WORKLOADS.

AI engines are inherently software programmable and hardware adaptable to support a wide range of workloads. Data movement is compiler driven, leveraging large scratchpad memories to reduce external memory bandwidth dependencies. Preallocated portions of the array can be isolated and dedicated to handle a specific set of workloads. This preallocation of resources is referred to as a *DNN overlay*. In Figure 6, a single-column DNN overlay is shown, and typically, several DNN overlays run concurrently, allowing users to deliver a range of QoS and performance levels within the same AI engine array.

To maintain the focus of the AMD XDNA architecture on providing performance efficiency, the supported data types are focused on edge-inferencing workloads and include INT4, INT8, and INT16 accumulating into 32/64 bit and BFLOAT16 accumulating into FP32. Also available is structured hardware weight sparsity (INT8), maximizing efficiency on sparse networks.

AI engines are interconnected with an AXI streaming network. AXI-streaming connectivity is predefined and programmed by the AI engine compiler tools based on the dataflow graph. Data transfers over this nonblocking interconnect allow for broadcast and overlapping of compute and data transfers. There are multiple initiator and target interfaces in the grid interconnect in each direction (north, south, east, and west). There are also direct connections between neighboring tiles to facilitate fused operations.

The AMD XDNA complex interfaces to the application processing unit (APU)-coherent Infinity Fabric through a series of memory tiles. These memory tiles serve as a staging buffer for weights and activations during the computation of the consecutive layers. There is one memory tile per column, and they are sized at a 512-KB connect through the AXI streaming network.

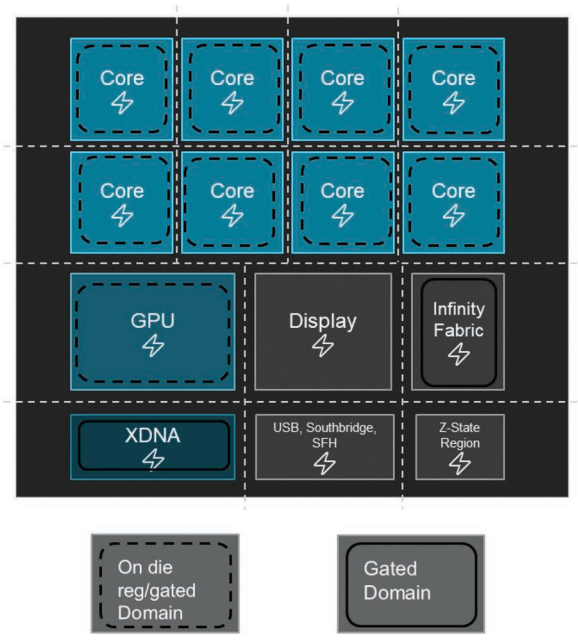

FIGURE 7. Power-efficient design within the SoC. On die Reg: on-die regulator; SFH: sensor fusion hub; Z-State: low power state.

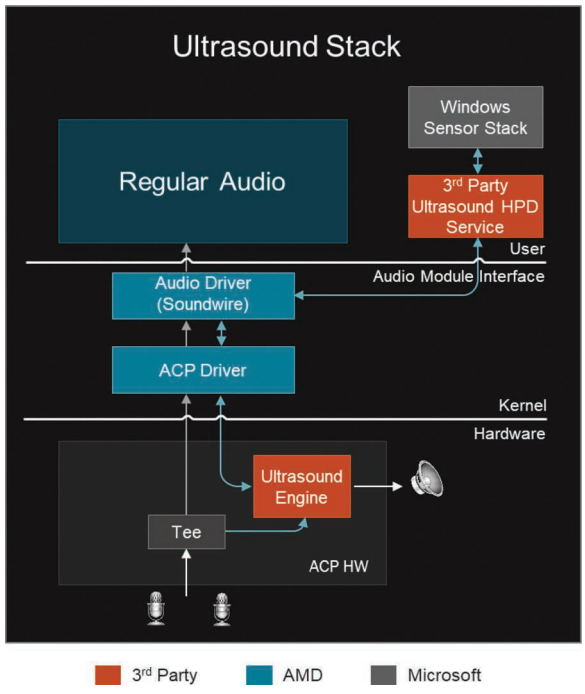

FIGURE 8. Ultrasound presence detection. HPD: human presence detect; ACP HW: audio coprocessor hardware.

High-performance DMA engines interface to the Infinity Fabric and act as the main data movers to and from the off-chip DDR/low power (LP)DDR memories. DMAs are typically scheduled data transfers, which is beneficial for repetitive data movements that are typical in inference workloads.

SoC POWER EFFICIENCY

The Ryzen 7040 processor brings forward new technology advances targeting power efficiency. Maximizing battery life for many common usage scenarios was a priority when designing this generation of APU.

A new hardware-managed power state (P-state) (Z8) was introduced, which offers a deeper P-state for brief periods of idleness or low-intensity activity. Using a combination of root tree clock gating and power gating, this state can be entered and exited in the order of 100 s of μs, providing a seamless user experience.

The combined DDR/LPDDR physical layer employs dynamic voltage and frequency scaling (DVFS) with a discrete voltage regulator, which allows it to adjust voltage given the current bandwidth demands on the system.

The new IPU based on the AMD XDNA architecture was supported with an advanced power management scheme, including DVFS and hierarchical clock gating to provide a power-efficient, rich AI experience. Additionally, when AI workloads are not running, the entire AMD XDNA complex can be power gated using an low dropout-based power gating circuit (see Figure 7). At runtime, an intelligent P-state management algorithm uses the AI workload's QoS requirement to determine the AMD XDNA's frequency operating point.

THE AMD XDNA COMPLEX INTERFACES TO THE APPLICATION PROCESSING UNIT (APU)-COHERENT INFINITY FABRIC THROUGH A SERIES OF MEMORY TILES.

Ryzen power-efficiency technology can also provide system-level benefits. Software tuning that utilizes existing platform power management technology and operating system modes help improve nonvolatile memory express' (NVME) drive-power efficiency. There is support for low-power clocking circuits and fast save restore with resume latency acceleration for modern standby modes. To improve certain use cases, such as video conferencing, at the software level, there is a smart scenario analysis engine running to dial in operational modes that balance power and user experience for the CPU and platform given an application-specific profile.

DOMAIN-SPECIFIC ACCELERATORS (AUDIO COPROCESSOR AND MULTIMEDIA INTELLECTUAL PROPERTY)

In addition to the highly specialized programmable engines in the Ryzen 7040 CPU, there are also multiple domain-specific accelerators that provide high-efficiency fixed functions, which are commonly used in productivity-type applications. In this section, we provide two examples of such an accelerator.

The fourth-generation Radeon Multimedia engine brings an improvement in performance. versus the previous generation, facilitating AOMedia Video 1 (AV1) encode and decode rates up to 8000 at 45 frames per second.[f] Most of the video conferencing and video content offers lower resolutions and frame rates; the extra performance here can be used for multistream codecs and can offer a race-to-idle power advantage by entering a low-power gated state between batches of frames.

Covering the common codecs AV1, high efficiency video coding (HEVC), automatic volume control, and VP9 (decode) enables this hardware accelerator to bring power efficiency to many video workloads.

The audio coprocessor also brings new capabilities and features to low-power audio processing. Increased static RAM, a second high-performance DSP, and new low-power clocking modes enable the coprocessor to handle more use cases that offload the higher-power x86 cores. Real-time, low-power audio playback on speakers, headphones, Bluetooth, and USB represents a set of use cases that can operate and help enable long battery life. Additionally, various wakes on audio features can be enabled while the device is in low P-states, such as modern standby. Keyword spotting and ultrasound presence detection are increasingly popular wake sources for modern laptops, both of which can be implemented on this coprocessor (see Figure 8).

PLATFORM OPTIMIZATIONS

Today's laptops offer a rich set of I/O for peripheral attachment. The Ryzen 7040 brings up to three type-C ports featuring USB 3/4, along with USB 3.1 Type A, and multiple USB 2. Twenty lanes of PCI Express allow for the attachment of an x8 dGPU, 2xNVMe, and other single-lane connections, such as a wireless local area network and a wireless wide area network. AMD FreeSync, with a panel self-refresh selective update on the embedded display port, enables power-efficient operation of lower-activity display modes.[g] Continuing the support for DP 2.0 and High-Definition Multimedia Interface 2.1 ensures the latest high-performance connectivity for a variety of external display devices. Finally, an update to USB 3 for the camera attachment allows for the use of a higher resolution greater than five microprocessor sensors. The APU separates this USB interface so that it can be used as secure biometrics, leveraging virtualization-based security when enabled in the system, enabling face recognition as an authentication feature.

CONCLUSION

The "Phoenix" SoC delivers significant improvements in traditional compute with the "Zen 4" CPU cores and RDNA 3 GPU cores, and introduces key accelerators in the design, like IPUs based on AMD XDNA architecture to deliver power-efficient solutions to the evolving mobile use cases.

REFERENCE

1. "AMD Ryzen™ 7040 series: Technology overview," in *Proc. IEEE Hot Chips 35 Symp. (HCS)*, 2023, pp. 1–27, doi: 10.1109/HCS59251.2023.10254701.

MAHESH SUBRAMONY is a senior fellow design engineer at AMD, Austin, TX, 78735, USA. His research interests include architecture and design on systems on chip spanning mobile, desktop, and servers. Subramony received his M.S. degree in computer engineering from the University of Minnesota Twin Cities. Contact him at mahesh.subramony@amd.com.

DAVID KRAMER is a fellow design engineer at AMD, Austin, TX, 78735, USA. Contact him at david.kramer@amd.com.

INDRANI PAUL is a senior fellow design engineer at AMD, Austin, TX, 78735, USA. Paul received her Ph.D. degree in computer engineering from Georgia Institute of Technology. Contact her at indrani.paul@amd.com.

[f]GD-176: Video codec acceleration [including at least the HEVC (H.256), H.264, VP9, and AV1 codecs] is subject to and not operable without inclusion/installation of compatible media players.

[g]GD-127: AMD FreeSync technology requires AMD Radeon graphics and a display that supports FreeSync technology, as certified by AMD. AMD FreeSync Premium technology adds requirements of mandatory low frame-rate compensation and at least a 120-Hz refresh rate at minimum full HD (FHD). AMD FreeSync Premium Pro technology adds requirements for the display to meet AMD FreeSync Premium Pro compliance tests. See www.amd.com/freesync for complete details. Confirm capability with your system manufacturer before purchase.