RADEON™

Dissecting the Polaris Architecture

AMD

Introduction

Fifty years into the evolution of Moore's Law, semiconductors continue to shrink and each new generation of GPU offers more transistors and functionality. At the same time, the power and energy benefits of new process technology, known as "Dennard scaling", have slowed down. Power has become the limiting factor for performance and user experience. High-end graphics products such as the Radeon™ R9 300 Series are generally limited by power delivery to 300W, while notebook graphics must use as little power as possible to deliver excellent battery life and enable compact form factors.

The Graphics Core Next (GCN) architecture is a solid foundation for high performance across the entire graphics ecosystem, from integrated notebook solutions, to leading edge game consoles, and high-end discrete graphics cards for VR and PC gaming (**Learn More:** 1st-generation Graphics Core Next whitepaper). Polaris builds on the success of GCN, systematically increasing performance, delivering a more responsive experience, enabling high-dynamic range media and display pipelines, all while increasing energy efficiency.[9,10]

The three critical building blocks powering the Polaris generation of GPUs are a new process technology, novel architecture, and creative circuit design techniques that draw on AMD's long expertise in CPU design. The new 14nm FinFET process technology reduces active power consumption and provides more transistors to allow for more compute units and cache.[1] The Polaris architecture also leverages these additional transistors for new intelligent features such as "aggressive primitive culling", which helps improve performance and energy efficiency, and quality-of-service to reduce contention between graphics and compute shaders. The Polaris architecture is implemented using custom and adaptive circuit designs that dynamically run the silicon at the highest frequency and lowest voltage possible, further boosting the energy efficiency of Polaris-based GPUs. The combination of all these innovations is a next-generation graphics architecture that is up to 2.8X more power efficient[9] while helping improve visual quality with high-dynamic range display and media encoding and decoding capability for the latest content.

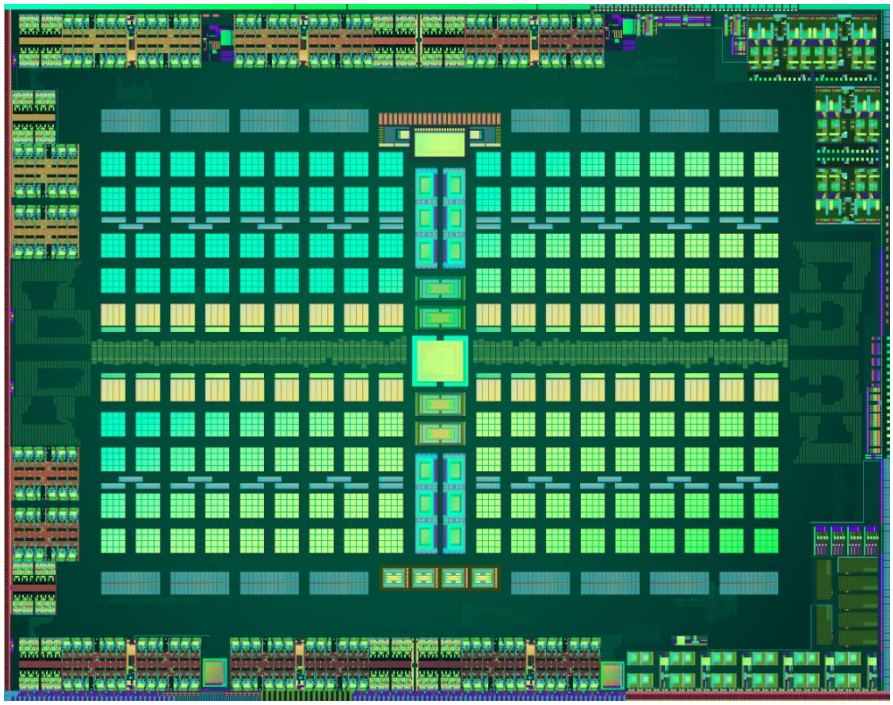

Figure 1: *Die plot of the Radeon™ RX 480 GPU, which is based on the Polaris architecture.*

Process Technology

Underlying Polaris architecture is the choice of process technology, which determines what is physically possible. Active (or dynamic) power consumption increases linearly with the number of computational units, but cubically when boosting frequency through higher voltage (e.g., 15% higher frequency and voltage increases power consumption by 52%). As a result, graphics processors tend to prefer lower frequencies and use greater density to deploy more computational units that operate in parallel. For the last five years, graphics processors have relied on 28nm high-k/metal nodes (see fig. 2).

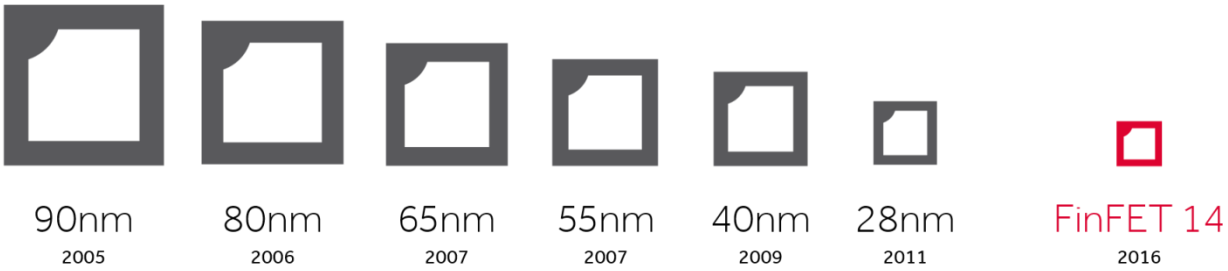

Figure 2: *Evolution of process nodes utilizes by Radeon™ graphics since 2005.*

For Polaris GPUs, AMD selected and Global Foundries' 14nm FinFET-based process technology, which is the densest foundry GPU process available for production to date. FinFET transistors are crucial to reducing power consumption and enable operating

voltages that are 150mV lower than the previous generation, thereby cutting active power by approximately 30% from a 1V baseline.

	Contacted Gate Pitch	SRAM Cell Size
14nm	78nm	0.064µm^2
16nm	90nm	0.070µm^2

Table 1: *Key geometries for FinFET processes*

Table 1 contains publicly available details on key dimensions for modern FinFET process nodes. For example, the table illustrates that the 14nm transistor spacing (i.e. contacted gate pitch) is approximately 15% smaller than TSMC 16nm spacing, while the SRAM used for caches and register files is 10% smaller. Overall, these process technology advantages translate into GPUs with more compute units, which allows for parallelism and better power efficiency.

Graphics Architecture

The Polaris graphics architecture is responsible for taking graphics and compute workloads and executing them as efficiently as possible. The GCN compute units (CUs) are already extremely efficient and well optimized.[1] Polaris builds on GCN, increasing the number of CUs per area to improve the raw computational throughput, but retaining the same overall CU design. Generally, Polaris emphasizes efficiency and focuses on improving control logic and fixed function graphics hardware to make the best use of the available compute units. Polaris enhances the command processing, geometry engines and memory subsystem to achieve even higher performance and greater power efficiency versus just GCN.

The command processor receives high-level API instructions (e.g., DirectX® or OpenCL™) from the driver and transforms them into compute shaders, graphics shaders, or DMA copy commands. Compute tasks are mapped onto several asynchronous compute engines (ACEs). Each ACE receives a separate command stream from the host and has eight queues for tasks. The ACE can dispatch from the head of any of the eight queues. A graphics pipeline contains queues for each type of shader (e.g., pixel shaders, texture shaders, and synchronous compute shaders) and two dedicated DMA engines handle copy commands to and from the GPU's memory. The ACEs dispatch asynchronous compute shader work-groups into the massively parallel shader array, while the graphics command processor dispatches graphics shaders and also coordinates fixed function hardware such as the rasterizers.

AMD pioneered a technique known as asynchronous compute, which enabled the ACEs, graphics command processor, and DMA engines to all simultaneously dispatch work to the GPU without context switching. Dispatching workgroups in parallel substantially reduces execution latency and increases throughput, thereby improving overall performance and

responsiveness. New low-level APIs such as DirectX® 12 and Vulkan™ expose the parallel control logic to developers, taking full advantage of asynchronous shading and enabling higher performance than earlier APIs.

The Polaris architecture enhances the command processor with two new quality-of-service (QoS) techniques designed to increase system responsiveness and performance. The first is known as Quick Response Queue and enables developers to designate a compute task queue as high-priority through APIs. Both high-priority and regular priority tasks co-exist and share the GPU's execution resources, but the ACEs dispatch workgroups from the high-priority task ahead of normal tasks. This prioritization scheme ensures that high-priority tasks will use more resources and complete first, without the command processor context switching out other lower-priority tasks. For example, this technique is used in the AMD LiquidVR™ SDK to prioritize 'time warping', which is a latency and jitter sensitive task, and ensure that the time warping occurs immediately before the vertical sync.

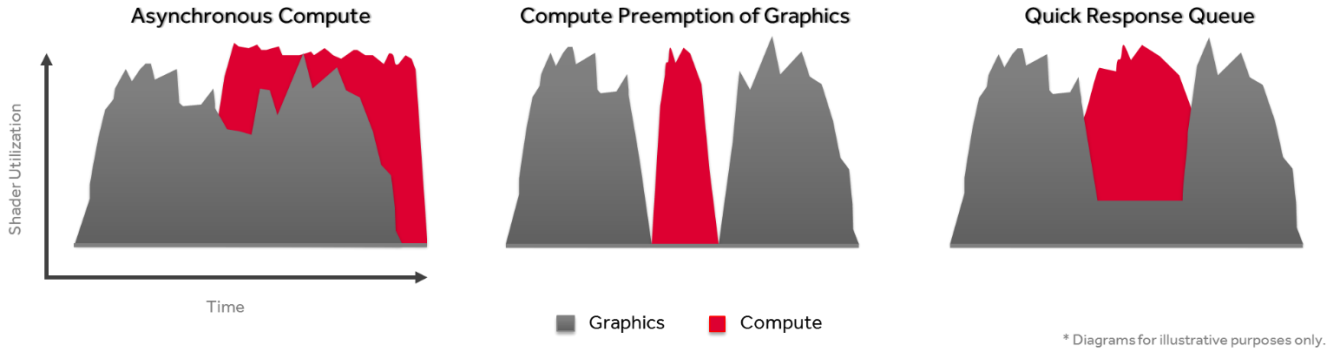

Figure 3: *The Polaris architecture can use its advanced asynchronous compute capabilities in many ways, including compute/graphics concurrency (left), compute preemption of graphics (middle), and compute QoS (right) for latency-sensitive tasks.*

The second quality-of-service technique, compute unit reservation, is even more potent and general-purpose. As the name suggests, programmers can partition the execution resources of the Polaris GPU for compute tasks using API extensions. Specifically, compute units (CUs) in the shader array are reserved for a queue in one of the ACEs, ensuring dedicated resources are available for work-groups from the queue. This is a powerful tool for developers to avoid contention between multiple tasks.

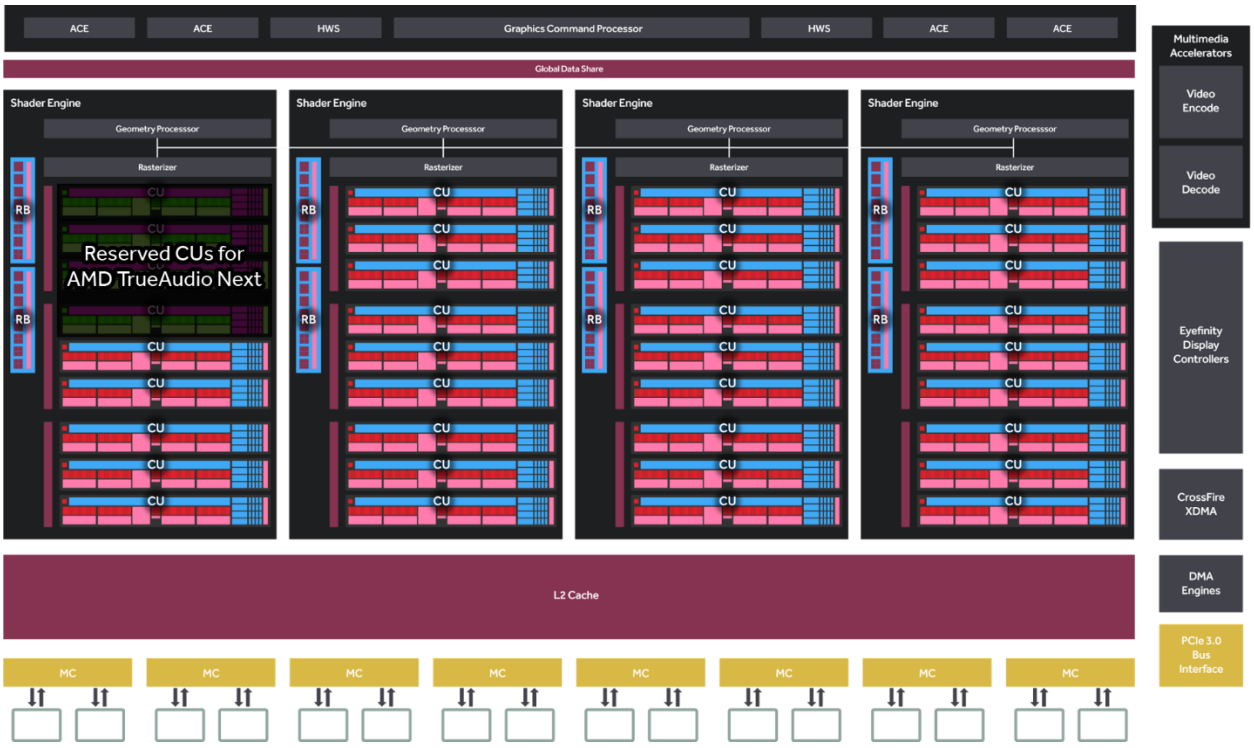

Figure 4: *Compute unit (CU) reservation on the Radeon™ RX 480 ("Polaris 10") graphics card, with four CUs reserved by the application for real-time audio raytracing.*

Figure 4 illustrates an example, where a queue for audio tasks is assigned 4 CUs, while the remaining compute units are available to all tasks. Partitioning the CUs ensures that audio tasks will have the lowest possible latency and jitter, although the CUs are no longer available for other tasks.

Hardware Scheduler (HWS)

In "Hawaii," and other GPUs based on the Graphics Core Next ISA, the hardware was designed to support a fixed number of compute queues (up to 8 per ACE). Starting with 3rd and 4th-gen GCN, however, the HWS makes it possible to virtualize these compute queues. This means that any number of queues can be supported, and the HWS will assign these queues to the available ACEs as slots became available.

Each ACE block in the Polaris and GCN Architecture diagram(s) represent a single wavefront/workgroup dispatcher. Accordingly, the "Fiji" GPU and GPUs based on the Polaris architecture can dispatch up to four wavefronts/workgroups to the shader engines from any compute queue at any time.

The HWS units are dual-threaded microprocessors capable of handling two scheduling threads, and their behavior can be tuned with microcode updates by AMD.

Geometry, Controllers, and Caches

Turning to graphics, the Polaris architecture enhances the geometry engines and tremendously improves both performance and energy efficiency of the rasterization stage. Polaris-based GPUs have 1-4 geometry engines, depending on overall performance targets (e.g. the Radeon™ RX 460 GPU has two, while the Radeon™ RX 480 GPU has four). The screen space is partitioned to load balance between the geometry engines, which can each rasterize a triangle per clock.

The Polaris geometry engines use a new filtering algorithm to more efficiently discard primitives. As figure 5 illustrates, it is common that small or very thin triangles do not intersect any pixels on the screen and therefore cannot influence the rendered scene. The new geometry engines will detect such triangles and automatically discard them prior to rasterization, which saves energy by reducing wasted work and freeing up the geometry engines to rasterize triangles which will impact the scene. The new filtering algorithm can improve performance by up to 3.5X (fig. 6), and the benefits are more pronounced in scenes with many polygons.

Figure 5: *Batman™: Arkham Origins uses high tessellation factors in the mesh of Batman's cape; some performance can be recovered by pre-rasterization discard of triangles that do not affect any pixels.*

In a similar vein, the Polaris geometry engines can detect triangles that have no area, and discard them during the input assembly stage. As vertex indices are read from the input buffer, the Polaris geometry engine will check if two or more vertices have the same coordinates (i.e., degenerate triangles). The degenerate triangles are culled before they

are passed to the vertex shaders, which increases throughput by reducing the amount of work done and reducing the energy consumed. By eliminating the vertex fetches for degenerate triangles, Polaris can increase throughput by up to 3X for certain scenes.

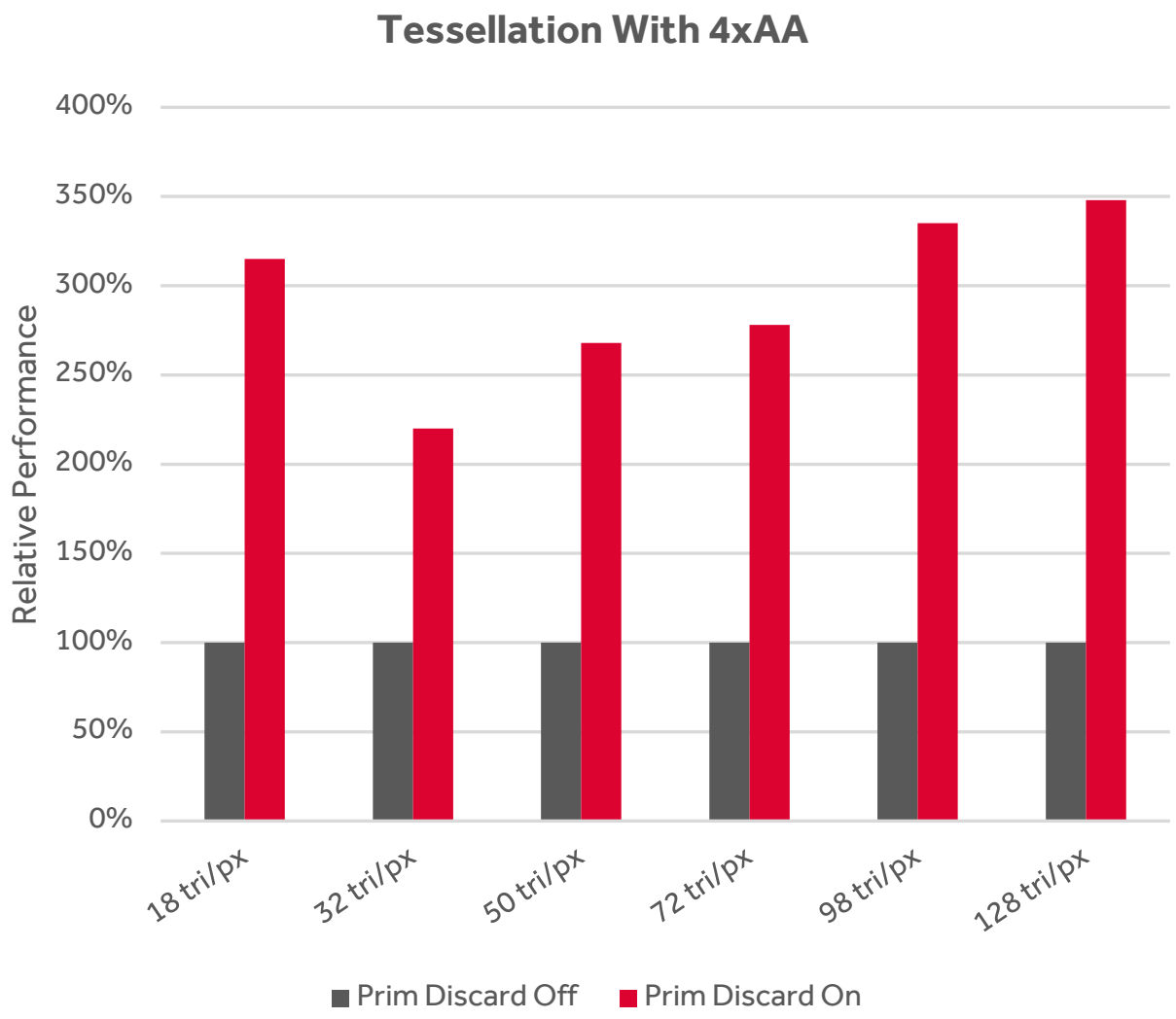

Figure 6: *The primitive discard accelerator for small and degenerate primitives is more effective as the triangle density increases.*[2]

The Polaris geometry engines are also more flexible than in previous generations. Triangles are commonly organized into lists and strips. A list of N triangles is simply a set of 3N vertices where each set of three vertices represents a triangle. A strip of triangles is a more compact data structure that takes advantage of locality to reduce the number of vertices needed. In a triangle strip, the first triangle takes 3 vertices, and each additional triangle shares two vertices and only needs a single additional vertex – so a strip of N triangles has N+2 vertices. The Polaris geometry engine can handle both lists and strips at full rate, which avoids the overhead of converting triangle strips to lists in software.

The geometry engine was also modified to more efficiently cache meshes for geometry instancing. When replicating a mesh (e.g. a tree) across a scene using instancing, the previous generation would store the mesh in the L2 cache and fetch it once for each instance. Polaris can actually cache meshes in the geometry engine itself, using various queues and buffers within the fixed-function hardware, thereby reducing the number of L2 cache requests which helps improve power consumption and performance.

Collectively, the changes in the Polaris geometry engine boost performance by increasing the achievable rasterization throughput, rather than simply relying on brute force and throwing additional geometry pipelines at the problem. Even better, eliminating triangles from the rasterizer not only increases performance, but saves energy, which ultimately translates into more active compute units and higher operating frequencies.

The Polaris memory interface has been updated to both increase bandwidth and also operate more efficiently with compression (fig. 7). The Polaris render back-end is designed to compress color buffers to save power and more effectively use the available memory bandwidth. Delta color compression is a lossless algorithm that dynamically divides a color buffer into several blocks and was first deployed in 3rd-generation GCN solutions (e.g. GPUs codenamed "Tonga," "Fiji," and "Antigua"). A single pixel in each block is written using a normal representation and all other pixels in the block are encoded as a difference from the first value. The block size is dynamically chosen based on access patterns and the data patterns to maximize the benefits. The peak compression ratio is 8:1 for a 256-byte block. Since many objects have large patches of similar colors (e.g. clothing and cars), the delta color compression takes advantage of this locality to improve performance. While 3rd-generation GCN and Polaris use similar algorithms, Polaris is more aggressive and compresses even more blocks, thereby saving more bandwidth and power.

The biggest savings come when the color buffer is read back for subsequent computations, i.e. render to texture mode. The Polaris shader cores can read and transparently decompress the compressed color data thereby saving read bandwidth in the memory and caches as well.

In fact, the compression is so efficient that the Polaris architects were able to reduce the number of render back-ends. Polaris has a compact 256-bit memory interface that uses cost-effective GDDR5 memory, but delivers similar end-user performance to the GPU codenamed "Hawaii", which used a 512-bit memory interface – all while consuming less power (fig. 8).

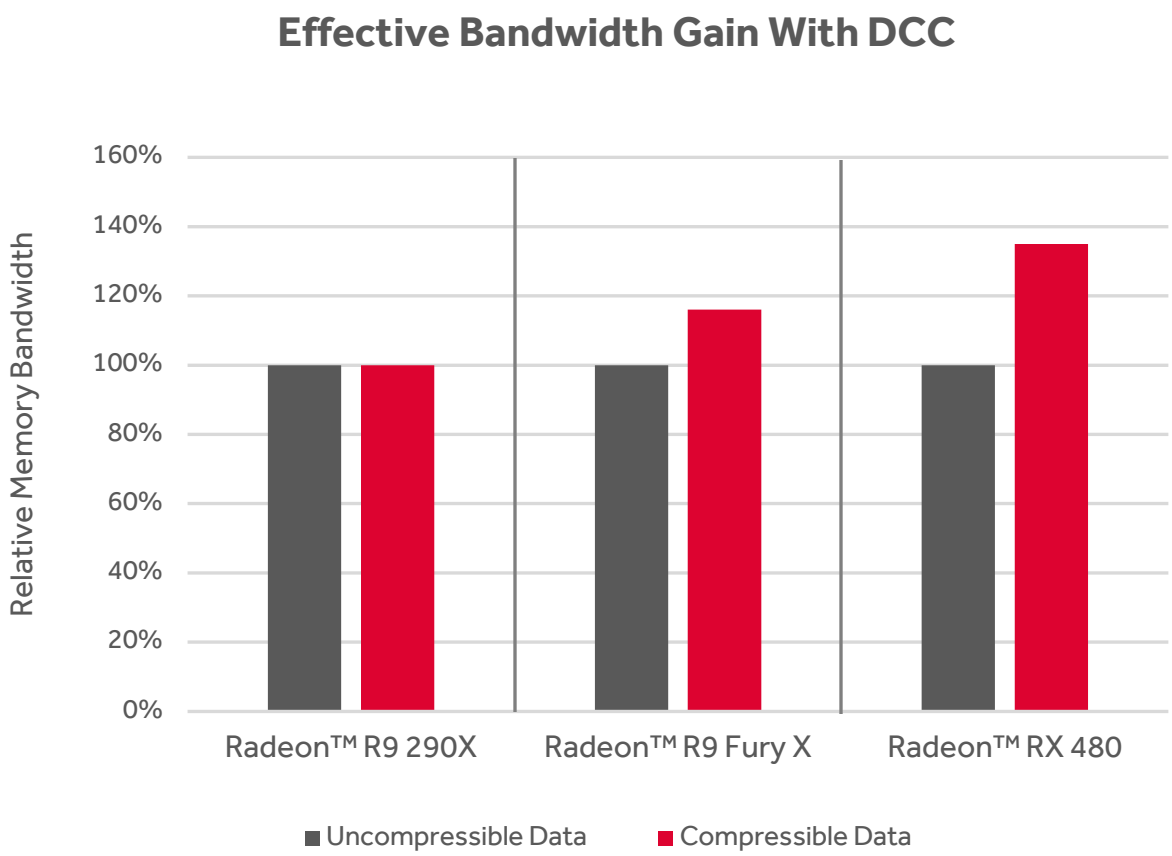

Figure 7: *The efficacy of AMD's delta color compression technology has increased over time, increasing the effective memory bandwidth up to 35% on the Radeon™ RX 480 with the Polaris architecture.*[3]

The Polaris architects also took advantage of the density of the 14nm FinFET process to double the L2 cache, compared to prior designs. The larger cache reduces the number of memory accesses that hit the GDDR5, saving power and decreasing latency. Together, the increased L2 cache and aggressive utilization of delta color compression in the Radeon™ RX 480 GPU ("Polaris 10") save up to 40% power on memory transactions compared to Radeon™ R9 290X.[4]

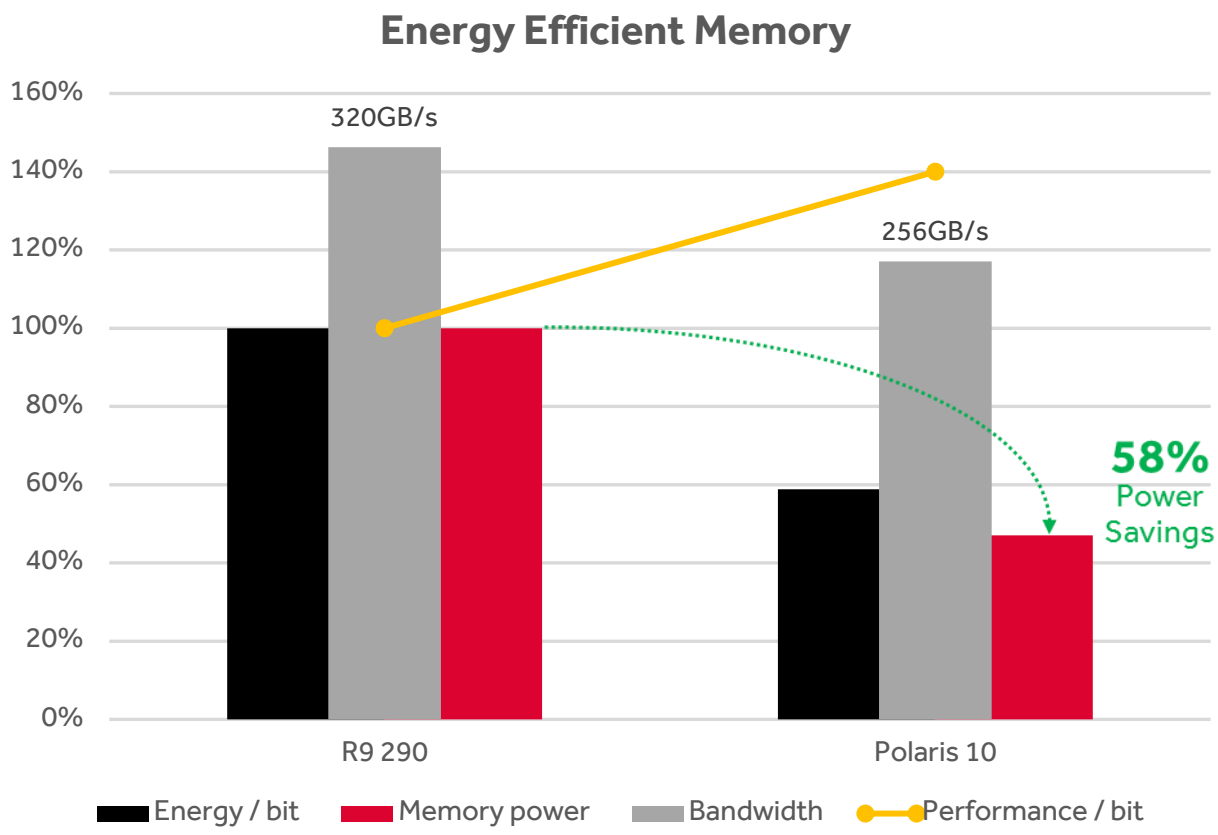

Figure 8: *The Polaris architecture combines aggressive delta color compression, new buffers, and a larger L2 cache size to deliver higher performance/bit at 58% lower power than the Radeon™ R9 290 GPU (codenamed "Hawaii").*[5]

Circuit Design

Fundamentally, circuit design bridges the gap between the logical architecture of a GPU, and the physical embodiment in silicon. Most graphics processors are designed using ASIC design flows that rely on automated tools and emphasize simple and high density circuits, rather than high frequency. In contrast, CPUs are typically designed to achieve the highest possible frequencies at the lowest power using cutting edge circuit design techniques. As a leader in both CPUs and GPUs, AMD has unique expertise in both design styles and can pick and choose the best approach. For example: the 6th-generation AMD A-Series APU ("Carrizo") heavily borrowed from GPU design techniques to increase the density of the "Excavator" CPU core design by 23% while decreasing power consumption by 40% on the same 28nm process technology.[6]

Polaris is AMD's first GPU architecture to take advantage of the advanced power management and circuit design techniques that have been developed for CPUs at AMD. Generally, the goal of power management is to dynamically adapt the GPU to the overall

system, usage model, and operating conditions. Collectively, these techniques enable Polaris-based GPUs to save substantial power, which directly translates into higher frequencies and superior performance.

Adaptive Frequency and Voltage Scaling (AVFS)

The most powerful technique deployed to manage power consumption in the Polaris architecture is AMD's AVFS, which was first developed for the 6th-generation AMD A-Series APUs ("Carrizo"). Modern GPUs operate in an incredibly complex environment with radically different combinations of system configurations (e.g. voltage regulator quality, cooling solution), temperature, and varied and changing workload (e.g. light gaming or the latest AAA games filled with explosions and sophisticated effects). Moreover, even theoretically identical GPUs are subject to subtle variations in silicon manufacturing. Traditional design techniques are fairly pessimistic and account for all these potential differences through guardbands, which reduce the operating frequency and/or increase the voltage – sacrificing performance and increasing power consumption.

The central concept of AVFS is to avoid guardbands and instead intelligently measure the behavior of each GPU and chose better combinations of voltage and frequency (fig. 9). AVFS uses power supply monitoring circuits to measure the voltage across different parts of a Polaris GPU in real time as seen by the actual transistors. Polaris GPUs also contains small replica circuits that mimic the slowest circuits in the GPU and are continuously monitored. Together these two blocks can measure how close the GPU is to the voltage limit at a given frequency. Similarly, the GPU can dynamically measure the temperature of the silicon in order to choose the right operating point since temperature affects transistor speed and power dissipation.

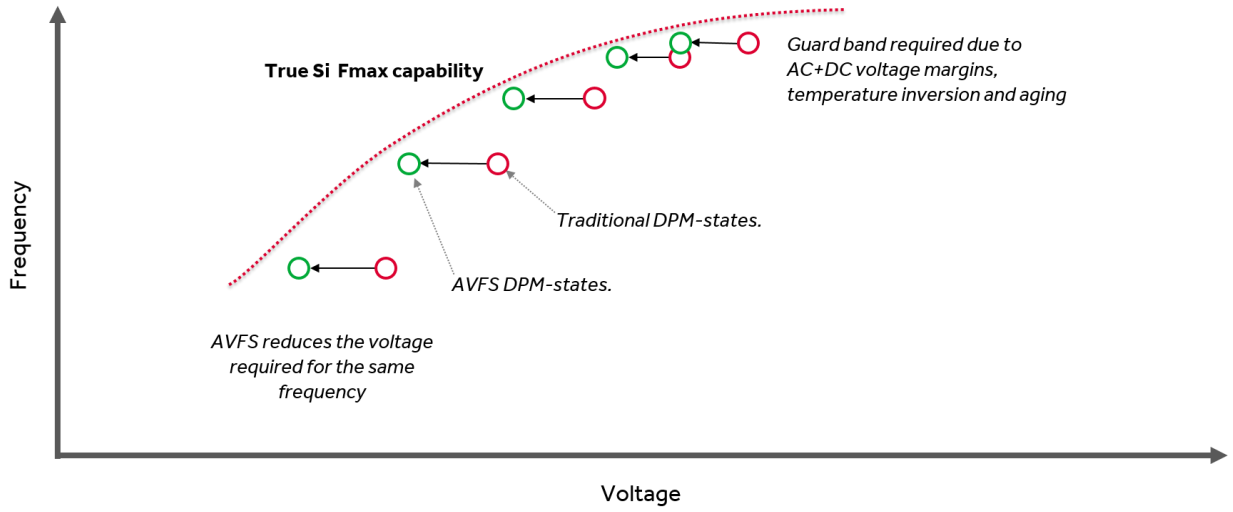

Figure 9: *AVFS in the Polaris architecture permits the reduction of voltage/frequency margin that would "leave performance on the table", improving performance and energy efficiency in products. (Diagram for illustrative purposes.)*

When the GPU boots up, the power management unit performs boot time calibration, which measures the voltage that is delivered to the GPU, compared to the voltage measured during the test and binning process. For example, it is fairly common for a voltage regulator to output 1.15V, but the GPU only receives 1.05V due to the system design. In the Polaris architecture, the power management unit can correct for this static difference very precisely, rather than requesting a more conservative (i.e. higher) voltage that would waste power. As a result, platform differences (e.g., higher quality voltage regulators) will translate into higher frequencies and lower power consumption.

In addition, the boot-time calibration optimizes the voltage to account for aging and reliability. Typically, as silicon ages the transistors and metal interconnects degrade and need a higher voltage to maintain stability at the same frequency. The traditional solution to this problem is to specify a voltage that is sufficiently high to guarantee reliable operation over 3-7 years under worst case conditions, which, over the life of the processor, can require as much as 6% greater power. Since the boot-time calibration uses aging-sensitive circuits, it automatically accounts for any aging and reliability issues. As a result, Polaris-based GPUs will run at a lower voltage or higher frequency throughout the life time of the product, delivering more performance for gaming and compute workloads.

Adaptive Clocking

Another advantage of AVFS is that it naturally handles changes induced by the workload. For example, when a complex effect such as an explosion or hair shader starts running, it will activate large portions of the GPU that suddenly draw power and cause the voltage to "droop" temporarily until the voltage regulators can respond. Conceptually, these voltage droops in a GPU or processor are similar to brownouts in a power grid (e.g. caused by millions of customers turning on their lights when they get home from work around 6pm).

The power supply monitors detect the voltage droop in 1-2 cycles, and then a clock-stretching circuit temporarily decreases the frequency just enough so that all circuits will work safely during the droop. The clock stretcher responds to voltage droops greater than 2.5% and can reduce the frequency by up to 20%. These droops events are quite rare, and the average clock frequency decreases by less than 1%, with almost no impact on performance. However, the efficiency benefits are quite large. The clock-stretching circuits enable increasing the frequency of Polaris GPUs by up to 140MHz.

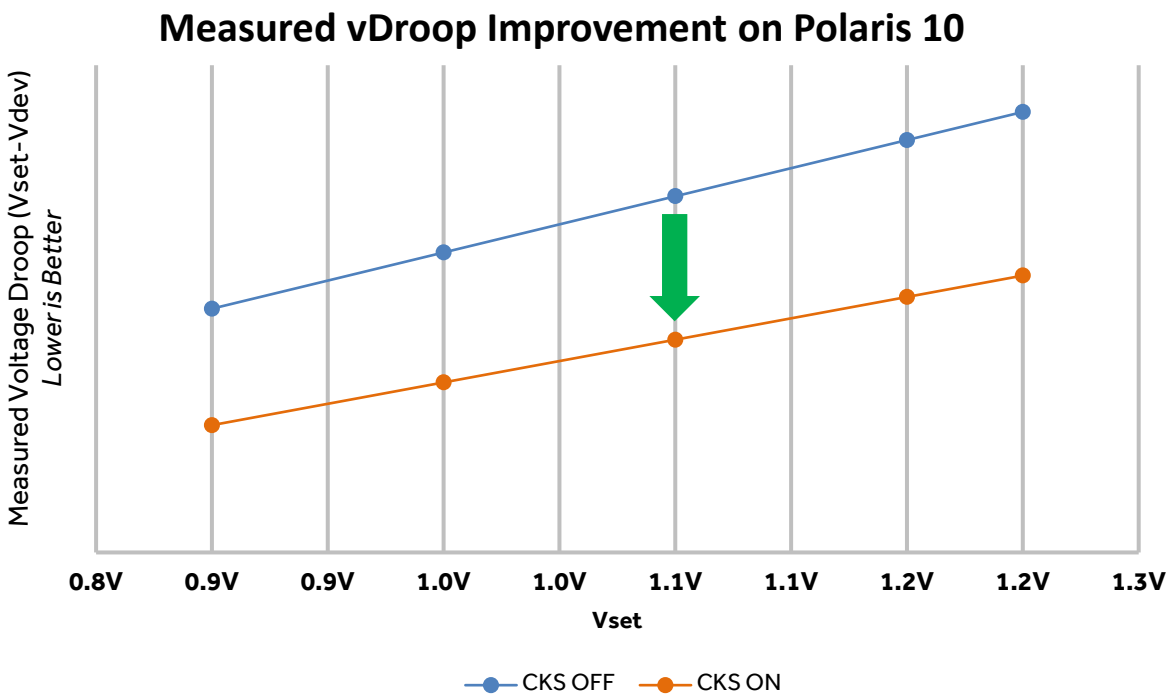

Figure 10: *AMD's adaptive clocking technology mitigates voltage droop events, which enables up to 140MHz higher frequency in the Radeon™ RX 400 Series GPUs based on the Polaris architecture.*

Multi-Bit Flip-Flop (MBFF)

The Polaris circuit designers borrowed a technique known as multi-bit "flip-flops" from the CPU to save power and increase performance. Flip-flops temporarily hold a single bit between computational functions and between different pipeline stages and are one of the most common building blocks in a GPU. For example, a Polaris CU contains roughly twenty-one million flip-flops. Every flip-flop has a clock input, data input, data storage, and data output. The clock input triggers the flip-flop to transit its stored data to the output and receive new data from the input. A clock network runs throughout the entire chip, distributing clock signals that synchronize operation. In active operation, the clock network typically consumes around 20-35% of the total power of a Polaris-based graphics chip.

AMD developed special "quad-flops", where four flip-flops share a single stronger clock input (fig. 11). A single quad-flop takes about twice the energy compared to a normal flop, but performs the work of four flops – reducing the load on the clock network by a factor of two. Using quad-flops reduces the energy consumed by flip-flops in a compute unit by a factor of two, which in turn saves about 5% of the compute unit's total power consumption. As an added benefit, the quad-flop saves a small amount of area over separate flops.

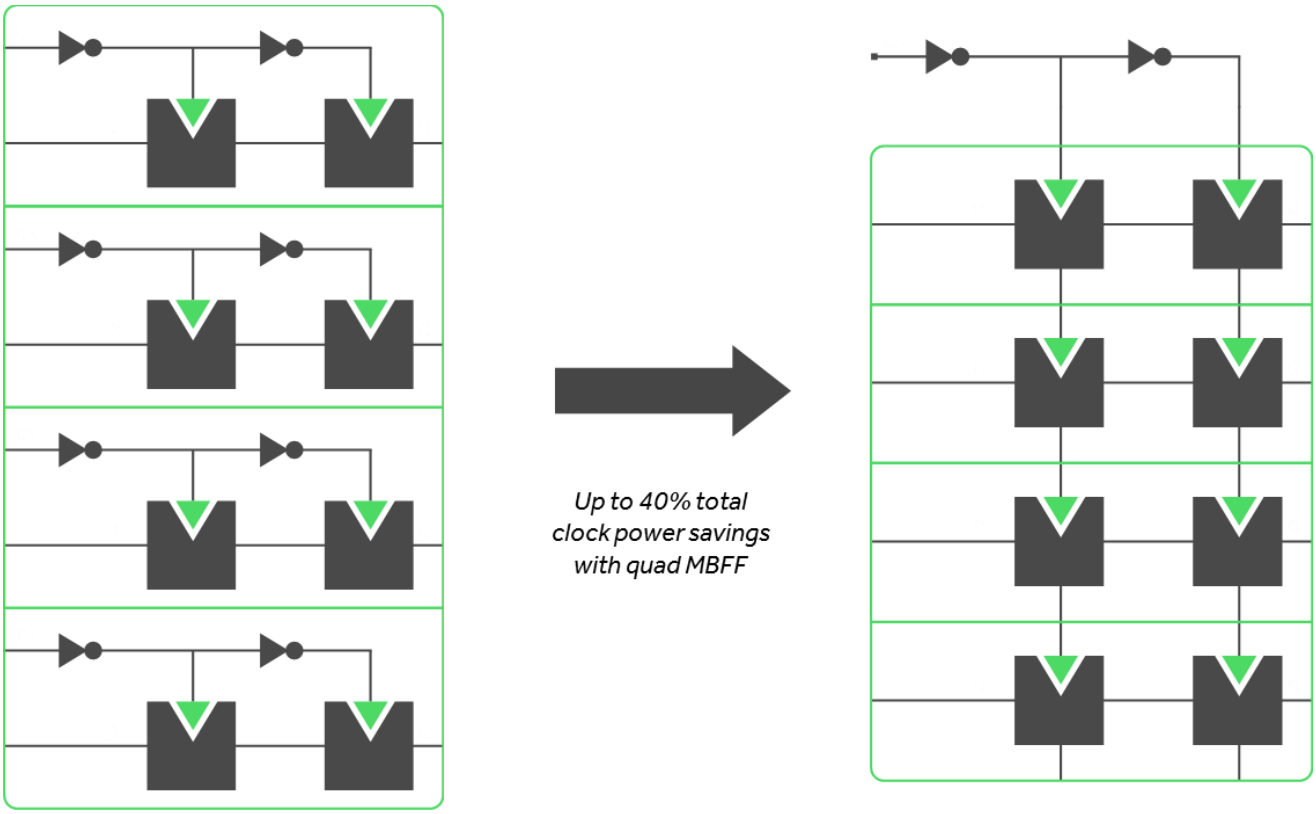

Figure 11: *The structure of a set of normal flops (left) vs. quad MBFF used in the Polaris architecture (right). The quad MBFF reduces die area, reduces clock power by up to 40%; and generally reduces total product TDP by 4-5%.*

Another example of a technique perfected in CPUs and carried over to GPUs is custom circuit design. Typically, GPUs are built by automated design tools that use a small library of standard cells (e.g., logical AND, OR, exclusive OR), whereas CPUs often use highly custom circuit designs. In Polaris, AMD's circuit designers introduced a number of custom cells that are smaller and more power efficiency than standard libraries. The GPU is still built by automated tools, but using these more efficient custom cells saves power and area, improving performance and efficiency.[7]

Last on the long list of circuit design improvements in Polaris is clock gating. Previous generations of GCN have all used clock gating to reduce the number of transistors that switch when active, which drives down power consumption. Polaris has even more fine-grained clock gating than previous GCN to save power and improve performance.

Media and Display Architectures

Some of the most exciting advances in graphics are taking place in display technology. Virtual reality is constantly pushing towards higher resolution and higher frame rates to deliver a seamless user experience. Computer displays are also aiming for higher resolution (e.g. 4K and 5K displays), but are beginning to shift focus to higher picture

quality through high-dynamic range displays that can reproduce more of the visible color spectrum.

The consumer electronics industry encompasses game consoles, movies, television, and home video. The large number of participants means that changes — such as the switch from standard definition (480p) to high definition (720p or 1080p) — occur more slowly, but the benefits and impact can be significant. The UltraHD standard is the next inflection point for mass market displays and will improve nearly every dimension. The standard increases resolution to 3840x2160 (4K) and 7680x4320 (8K). More importantly for consumers, UltraHD targets high dynamic range with 10 bits per color (bpc) channel on each pixel and a frame rates up to 120Hz or 60Hz stereo.

To drive the industry forward, the Polaris architecture upgrades both the display output and the integrated multimedia accelerators. Polaris GPUs offer both DisplayPort™ 1.3 and 1.4-HDR, as well as HDMI® 2.0b to connect with a display. Figure 12 shows the supported resolutions for DisplayPort™.

Figure 12: *The incorporation of DisplayPort™ 1.3 HBR3 and DisplayPort™ 1.4-HDR permits a wide range of new resolution/refresh rate combinations for gamers, in addition to support for high dynamic range (HDR).*

The Polaris architecture is optimized for High Dynamic Range (HDR) content; the display pipe is capable of transmitting output with 10- or 12-bits of color data per channel as illustrated in figure 13.[8]

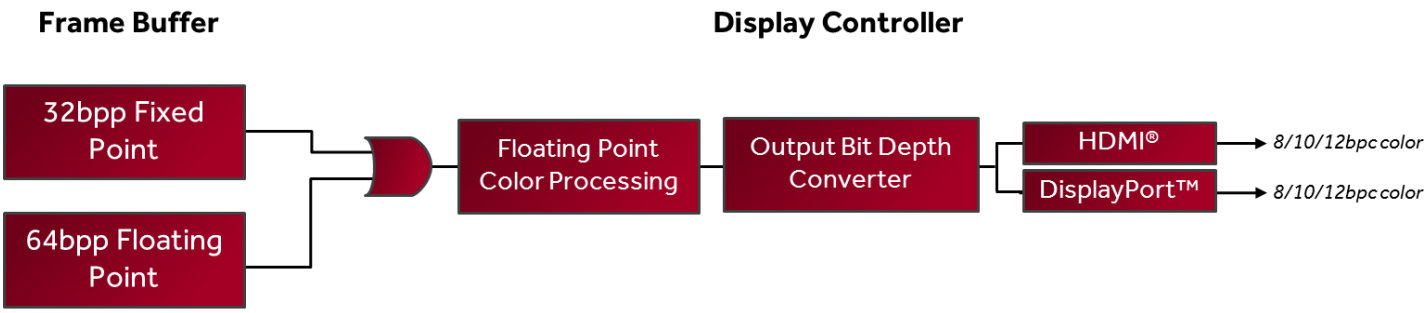

Figure 13: *The display pipe on the Polaris architecture can handle 32bpp fixed point or 16bpp floating point framebuffer data, allowing for 8/10/12bpc output to HDMI® or DisplayPort™ displays.*

HDR refers collectively to the combination of Rec.2020 color space, CTA-861.3 HDR metadata transport and the SMPTE 2084 electro-optical transfer function. A display and graphics system that incorporates these three technologies can display scenes with a tremendous color gamut and contrast ratios that will come very close to the perceptual limits of the human visual system (fig. 14).

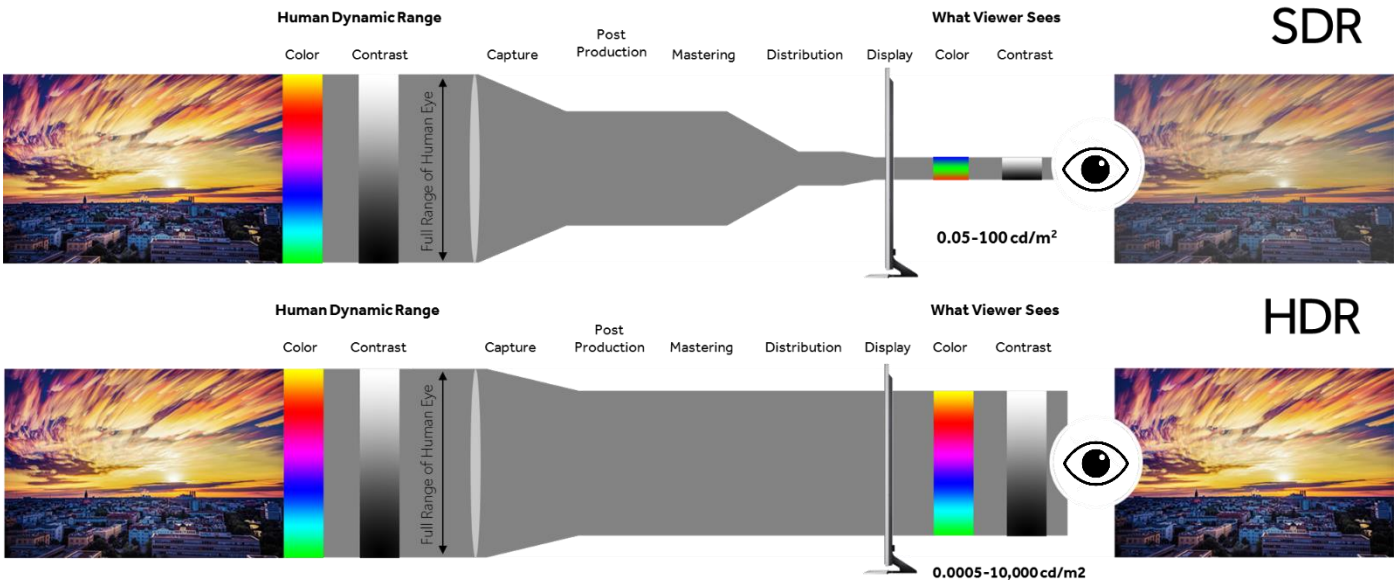

Figure 14: *HDR capture, mastering, distribution and playback attempts to capture a significantly wider portion of the human visual system. The result is a significantly more lifelike image.*
(Screen images simulated.)

The Polaris architecture is designed to display HDR content over both the DisplayPort™ and HDMI® interfaces across all the major resolutions used in gaming, movies, television and streaming video. The display pipeline also has support for HDCP 2.2 to help ensure compatibility with protected HDR platforms, content and services.

HDMI® 2.0b (with HDCP 2.2)	1920x1080 @ 192Hz 2560x1440 @ 96Hz 3840x2160 @ 60Hz (4:2:2)
DisplayPort™ 1.4-HDR (with HDCP 2.2)	1920x1080 @ 240Hz 2560x1440 @ 192Hz 3840x2160 @ 96Hz

Table 2: *Common display resolutions and their maximum refresh rates in HDR mode on the Radeon™ RX 400 Series graphics cards.*

Additionally, AMD is working with game developers to ensure HDR tonemapping is performed in the display pipe of Radeon™ graphics cards, instead of in the panel controller. Shifting the tonemapping to the display pipeline helps reduce the frame latency by ensuring a simple and fast path in the display itself.

The Polaris architecture includes the latest generation of AMD's video encode and decode acceleration engines. The Polaris architecture's decode accelerator has been upgraded to handle HEVC/H.265 main10 profile, with support for 3840x2160 resolution at up to 60Hz with 10-bit color for the HDR content.[9] The Polaris architecture has also been updated to include support for the VP9 codec at up to 4K resolution, which dovetails with YouTube's transition to VP9 encoding.

On the encode side, H.264 encode acceleration is carried forward from previous-generation products at 1080p120, 1440p60 or 2160p30 rates. AMD has worked with a variety of application vendors—including Plays.tv, AMD Gaming Evolved Powered by Raptr, and OBS Studio™—to expose this functionality. As streaming platforms and services transition over to HEVC/H.265 to improve quality and data rates, the Polaris architecture has also been updated to include H.265 encode acceleration at 1080p240, 1440p120 and 2160p60 rates.

The Polaris architecture also improves encoding quality by enabling two-pass variable bitrate encoding. The video encode accelerator performs a pre-encoding pass on a downscaled scene that is analyzed at the frame and macroblock level to determine efficient bitrate budgeting and quantization parameter (QP) selection. The rate control parameters derived from this analysis are used to guide the final encoding, resulting in output with fewer overall macroblock artifacts and notably higher fidelity. The two-pass mode increases latency on the encoding pipeline, but can reduce bitrate requirements and improve the final results for users recording their gameplay to disk.

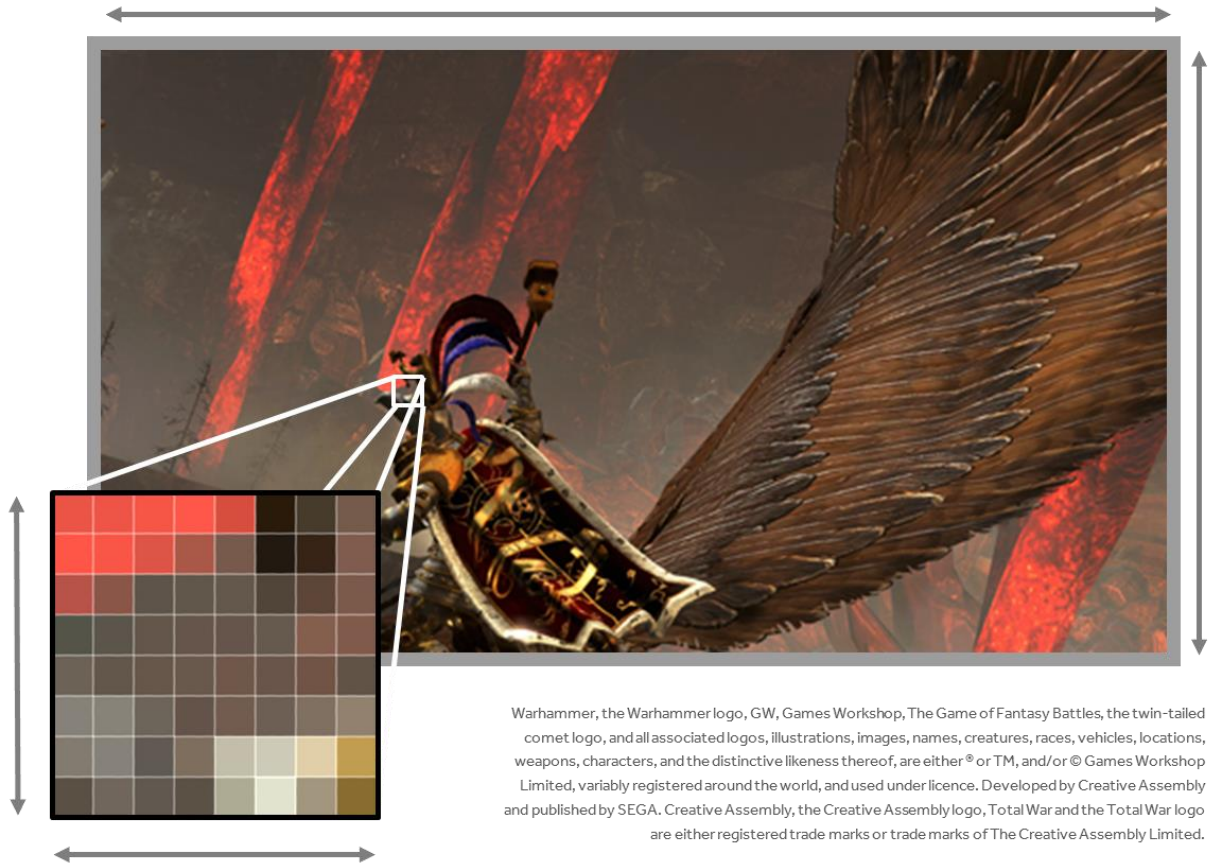

Figure 15: *A new two-pass VBR encoding mode on the Polaris architecture analyzes the frame and macroblocks in the scene during a pre-encode pass to intelligently guide bitrate and QP selection during final encode.*

Polaris Guides the Way Forward

Polaris starts with the excellent GCN architecture, and builds on top of it to deliver one of the most efficient graphics architectures ever. Polaris takes advantage of the latest 14nm FinFET process, which improves performance per watt by 1.7X (fig. 16) and doubles transistor density.[10] Rather than simply building more of the same, these extra transistors are wisely invested in architectural features that boost performance, efficiency, and image quality. For example, each Polaris compute unit achieves 15% more performance than the prior generation Radeon™ R9 290X ("Hawaii") GPU.[11]

Similarly, enhancements to the geometry engines boost both raw performance and power efficiency by aggressively discarding triangles prior to rasterizing a scene, and new quality-of-service features enable compute and graphics workloads to use the GPU simultaneously. The architecture also invests in HDR display support, which is the future of media and graphics, improving the color gamut and intensity to enable even more impressive visuals. All these improvements are tied together with a set of proprietary AMD circuit design strategies to improve frequency and reduce voltage, ultimately achieving an

impressive 2.8X increase in performance per watt in Radeon products (fig. 16).[12] Collectively, these innovations position the Polaris architecture to be the backbone of modern graphics today and tomorrow, whether in PC graphics cards, consoles, or virtual reality.

Figure 16: *the FinFET 14 process proves a solid foundation for the energy efficiency of the Polaris architecture, but AMD proprietary power management and circuit design techniques were pivotal in extracting peak performance per watt.*

Legal Attributions

---

[1] One compute unit contains 64 stream processors on AMD Radeon™ and AMD FirePro™ GPUs based on the Graphics Core Next graphics instruction set.

[2] Based on AMD internal small prim filter test as of 6/14/2016. Primitive assembly rates with prim filter ON vs. OFF: 18 tri/px (3.947 vs. 1.255), 32 tri/px (3.901 vs. 1.773), 50 tri/px (3.760 vs. 1.402), 72 tri/px (3.303 vs. 1.187), 98 tri/px (3.928 vs. 1.171), 128 tri/px (3.870 vs. 1.171). System configuration: Radeon™ RX 480, Core i7-6700K, 16GB DDR4-2666, Windows 10 x64, Radeon™ Software 16.5.2.

[3] Based on AMD internal memory bandwidth test as of 6/14/2016. Radeon™ R9 290X: 263GB/s peak memory bandwidth. Radeon™ R9 Fury: 333 peak GB/s without DCC vs. 387 peak GB/s with DCC. Radeon™ RX 480: 186 peak GB/s without DCC vs. 251 peak GB/s with DCC. System configuration: Core i7-6700K, 16GB DDR4-2666, Windows 10 x64, Radeon™ Software 16.5.2.

[4] Based on measurements of total memory interface power in watts conducted by the AMD performance labs as of 5/21/2016. System configuration: Radeon™ R9 290 vs. Radeon™ RX 480, Core i7-5960X, Gigabyte GA-X99-UD7, 16GB DDR4-2666, Windows 10 x64, Radeon™ Software 16.5.2.

[5] Based on measurements of total memory interface power in watts conducted by the AMD performance labs as of 5/21/2016. System configuration: Radeon™ R9 290 vs. Radeon™ RX 480, Core i7-5960X, Gigabyte GA-X99-UD7, 16GB DDR4-2666, Windows 10 x64, Radeon™ Software 16.5.2.

[6] ISSCC Carrizo, 2015, slide 13: http://www.slideshare.net/AMD/isscc?ref=http://www.amd.com/en-us/who-we-are/corporate-information/events/isscc

[7] Results obtained from AMD internal testing.

[8] HDR content requires that the system be configured with a fully HDR-ready content chain, including: graphics card, monitor/TV, graphics driver and application. Video content must be graded in HDR and viewed with an HDR-ready player. Windowed mode content requires operating system support.

[9] HEVC acceleration is subject to inclusion/installation of HEVC-compatible applications.

[10] Based on AMD internal data generated in AMD performance labs as of May 2016, measurements of capacitance, voltage frequency, leakage and power data show up to 1.7x performance/watt on 14nm vs 28nm FINFET technology. Final performance/watt results on AMD products using 14nm FinFET technology may vary and will depend on various factors including but not limited to clock speed, voltage, and various AMD proprietary technologies. RX-17

[11] Testing conducted by AMD performance labs as of May 18, 2016 on the Radeon RX 480 and Radeon R9 290 on a test system comprising Intel Core i7-5960X, 16GB DDR4-2666, Gigabyte X99-UD4, Windows 10 x64 (build 10586), Radeon Software Crimson Edition 16.5.2using Ashes of Singularity, GTA V, Project Cars, Witcher, and Assassin's Creed Syndicate, All games tested at 1440p. Radeon RX 480 graphics (150W TGP/36 CU) vs. Radeon R9 290 graphics (275W TGP/40 CU) scores as follows: Ashes of the Singularity (44.19 FPS vs 46 FPS); GTA V (66.23 FPS vs. 66.44 FPS); Project Cars (48.99 FPS vs. 45.99 FPS); Witcher 3 (50.78 FPS vs. 50.13 FPS); Assassin's Creed Syndicate (50.51 FPS vs. 45.78 FPS). Average FPS of above game scores: 52.14 (Radeon RX 480) vs. 50.06 (Radeon R9 290). Discrete AMD Radeon™ GPUs and AMD FirePro™ GPUs based on the Graphics Core Next architecture consist of multiple discrete execution engines known as a Compute Unit ("CU"). Each CU contains 64 shaders ("Stream Processors") working together (GD-78). CU efficiency formula = average FPS/# of CUs. Test results are not average and may vary. RX-4

[12] Testing conducted by AMD Performance Labs as of May 10, 2016 on the AMD Radeon™ RX 470 (110w) and AMD Radeon™ R9 270X (180w), on a test system comprising i7 5960X @ 3.0 GHz 16GB memory, AMD Radeon Software driver 16.20 and Windows 10. Using 3DMark Fire Strike preset 1080p the scores were 9090 and 5787 respectively. Using Ashes of the Singularity 1080P High, the scores were 46 fps and 28.1 fps respectively. Using Hitman 1080p High, the scores were 60 fps and 27.6 fps respectively. Using Overwatch 1080p Max settings, the scores were 121 fps and 76 fps respectively. Using Performance/Board power, the resulting average across the 4 different titles was a perf per watt of 2.8X vs the Radeon R9 270X. Test results are not average and may vary. RX-6