AMD Targets Accelerating Energy-Efficiency Gains

Whitepaper Sponsored by AMD
June 2014

Executive Summary

AMD products have a strong record of steadily improving power efficiency and increasing performance, and the company is pursuing an ambitious plan to accelerate these advances. Although AMD is driving energy-efficiency gains in servers, discrete graphics, and the other markets it serves, this paper focuses on the energy-efficiency technology in the mobile-computing segment, which consists of chips for notebook and tablet computers. A common energy-efficiency metric is work done per unit of energy consumed; a more efficient laptop, for instance, will enable the user to perform the same tasks with less battery drain and cooler operation.

For mobile computing, a reasonable approximation of typical energy consumption is an approach defined by Energy Star[1] that, perhaps surprisingly, centers on "short idle" power. Generally, after accessing documents or rendering web pages, the user spends time viewing the results. During these idle periods, which in modern systems can be as short as the time between keystrokes or between frames in a video, the processor enters a low-power state. Thus, AMD defines the typical-use efficiency of its mobile-device silicon as the compute capability divided by the typical energy use. For example, when comparing two similarly performing laptops, users prefer the model that exhibits longer battery life – or equivalently, if two laptops deliver the same battery life, users prefer the model with higher performance and better responsiveness. These two cases are embodied in the typical-use energy efficiency metric.

Using this metric, the company tabulated data for the energy efficiency of its mobile-processor line over the last six years and calculated a 10x improvement[2]. This feat sounds remarkable except that we have become accustomed to the rapid pace of semiconductor innovation — according to Koomey's Law[3], we should actually *expect* at least this level of improvement in energy efficiency. But the last 10 years have seen a significant deceleration in raw efficiency improvements, and many pundits think Moore's Law is slowing. That's what makes AMD's plan to *accelerate* the typical-use energy-efficiency gains of its mobile-computing products so impressive. *The company is targeting a 25x typical-use energy-efficiency improvement[2] over the next six years,* significantly outstripping the 10x progress of the last six.

To achieve this remarkable goal, AMD plans to tap a variety of resources, mostly in the areas of architecture, design, and software, in addition to silicon process technology. In particular, the company points to three main resources, citing evidence from existing products:

- **Improvements in intelligent, real-time power management.** These advances help drive down idle power and exploit the "race to idle" benefits of finishing a job quickly to

enable a faster return to the low power state. This approach has been a primary factor behind the 10x gains that AMD has achieved to date, and the company is constantly working to improve this capability.

- **Heterogeneous-compute capabilities.** The Heterogeneous System Architecture (HSA) enables AMD's accelerated processing units (APUs) to raise performance for common workloads, (as demonstrated using industry standard benchmarks such as PCMark 8 v2.0) as well as emerging visually oriented and interactive workloads like natural user interfaces along with image and speech recognition. AMD is a founding member of the influential HSA Foundation, an organization which promotes standards for heterogeneous computing, and is well positioned to regularly deliver processor improvements by taking advantage of its leadership in GPU design.
- **Innovation in power-efficient implementation.** AMD aims to advance the efficiency of intellectual property (IP) in its APUs by applying such technologies as advanced power gating, low-voltage operation, and further integration of system components. These technologies can help the company achieve its goal of a 25x efficiency gain over the next six years. And AMD's roadmap indicates that it has numerous technology innovations in development to deliver even smarter and more-power-efficient implementations.

Combining the previous six years of improvement (10x) with that of the next six (25x), the typical-use energy efficiency of AMD's mobile-computing silicon is calculated to increase by an *astounding 250x* (10x previous improvement multiplied by 25x estimated future improvements) — a surprising feat even to market veterans familiar with the semiconductor industry's remarkable progress.

AMD Innovation in Power Efficiency

AMD boasts a long history of processor innovation. It has responded to the rapid changes in computing over the last decade by applying its own unique inventiveness.

Although the company is known mostly for its PC and server processors, it employs top-notch graphics talent as well. AMD's world-class GPU designs are critical to advancing its vision of combining the CPU and GPU on the same silicon to revolutionize the microprocessor business—an approach called "heterogeneous computing".

AMD's history of innovation includes the first PC processor to reach 1GHz, the first PC and x86 server processors to adopt 64-bit architecture, and the first multicore x86 processor. The company also collaborated on the first supercomputers powered by graphics processing units. Its GPUs deliver industry-leading features and performance, and are inside all three of the latest, major game consoles. Today's AMD enjoys not only a history of performance leadership, but also the ability to capitalize on these two processing elements (CPU and GPU) to deliver cutting-edge performance per watt through heterogeneous computing.

Because power consumption now limits system performance scaling, AMD developed a form of heterogeneous computing that allows highly parallel tasks to shift seamlessly between the CPU and GPU. This architecture offers the opportunity for greater power efficiency while simultaneously increasing performance and maintaining programmability.

This commitment to power efficiency also led the company to negotiate a license to the 64-bit ARM architecture for its next generation of power-efficient dense servers. Expanding on its work with this architecture, AMD is applying its experienced design team to the creation of its own cores optimized for a unique blend of performance and power. In addition, it is also designing APUs to run in low-power server applications, garnering similar energy-efficient benefits and Heterogeneous System Architecture (HSA) compute capability.

Combining the power savings of reduced idle power and more-intelligent power management with the performance boost of heterogeneous computing (HSA) and process improvements, we believe there's a strong argument to be made for *AMD's estimates that it will be able to deliver a 25x typical-use energy-efficiency improvement between 2014 and 2020.*

Designing an Energy-Efficient Processor

There are number of ways to attack energy-efficiency issues, but none are considered the proverbial "silver bullet" in reaching the next level of efficiency. In the past, reductions in

semiconductor process geometries drove most efficiency increases, because smaller transistors (a "shrink") require less energy to switch states and can also function at lower voltages. A process shrink typically affects two important elements in electronic devices: the supply voltage and the current required to switch transistors (owing to the lower capacitance). Current also derives from the speed of the switching transistors and is directly proportional to clock frequency. For this reason, a higher clock speed leads to higher power. Lower power and lower clock speed alone may be insufficient to save energy if the slower clock extends processing time, thus keeping the system in a higher power state for a longer duration.

Figure 1 shows a simplified graph of processor power consumption. It illustrates two main components: idle power and dynamic power. Idle power is the result of current leakage from the millions (or billions) of transistors on the chip (transistors are imperfect on/off switches—they always leak some small amount of current in either state). Dynamic power is proportional to clock frequency - the faster the clock, the greater the power consumption. In the past, designers ignored leakage to focus on managing dynamic power, but thanks to ultra-fast power management, processors are now spending more time in the idle state.

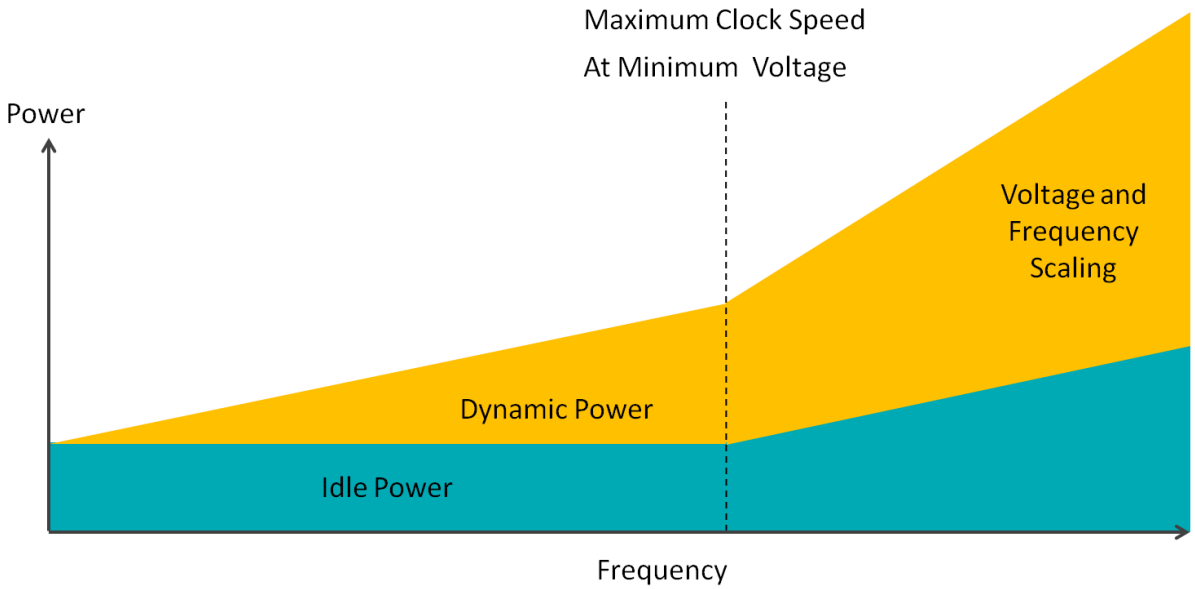

Figure 1 - Simplified plot of processor power versus frequency

Given the extensive reliance on the idle state, reducing idle power has become extremely important to managing energy consumption, because the power consumed in this state is now the primary indicator of typical-use power in mobile computing. Idle power also increases with processor voltage (the region in the figure to the right of the dashed line), further aggravating peak consumption. Some leakage can be controlled using new semiconductor-manufacturing

technology (usually through new transistor structures or materials), but circuit design has an increasingly critical role.

Even though power reduction is an important goal, the semiconductor industry also counts on higher performance to improve productivity. The benefits of reduced power consumption due to process shrinks (reducing the size of transistor structures) allow chip designers to fit more transistors into a similar-size die with a similar power budget.

The industry drive for even more performance has led to larger die sizes and an almost exponential growth in transistor counts, but this growth has outpaced power savings from the adoption of new manufacturing processes. As such, even though each new transistor is more efficient than the last, the total number of transistors has expanded even faster over the previous decade, pushing CPUs to higher and higher power levels.

Then the industry changed: laptops began to replace desktops, and mobility became more important than pure performance. Tablets and smartphones supplanted laptops in some applications. Semiconductor vendors could no longer increase clock speeds and transistor counts without taking extra steps to reduce the chip's active power, and process shrinks alone were no longer the dominant way to save power; the industry required innovative circuit and system design.

In addition, trouble is on the horizon for the continual march of semiconductor-process improvements commonly known as Moore's Law. Newer process geometries are becoming more expensive to manufacture, and the core voltages at which they operate no longer scale down significantly. Unlike a decade ago, process improvements alone are insufficient to deliver power savings. Today, system and chip design play a critical role in power savings. Newer, efficient design methodologies and more-intelligent power management, along with better software distribution of workloads, are delivering greater energy efficiencies while still enabling ever-increasing performance for advanced applications.

AMD's Response to the Power Challenge

AMD has conducted ongoing design efforts over the last decade to reduce power consumption, but it is preparing to make an important leap with its APUs, GPUs, and dense servers. The company has measured notebook processor energy consumption over the last six years, and observed a drop of almost 60% (for typical use), whereas the computational capability has improved more than fourfold[4]. These results represent more than a 10x net range in energy efficiency between 2008 and 2014 as measured by peak performance per watt of typical-use power.

Because most systems fail to fully utilize their maximum processing capabilities all the time, shifting the processor to the lowest possible power state is critical. But the device must still maintain sufficient processing headroom to handle peak performance demands. AMD is achieving this goal through a number of essential system innovations.

The company has developed an intelligent and dynamic power-management approach that it incorporates into all its APU products. Its recent APUs also integrate numerous system components onto one die, thus reducing power by eliminating interfaces between chips and by manufacturing all components using an advanced process generation. And, of course, AMD has also used process scaling and optimum silicon-design principles to reduce power consumption.

In notebooks, AMD has integrated system components such as the GPU, memory controller, I/O controllers, and peripheral buses, all on a single die.[5] Combining the CPU and GPU on the same die has multiple benefits; for instance, it enables fine-grained power-management technology that can monitor both the CPU and GPU. This technology balances power optimization between the two units, allocating more thermal dissipation to the unit that needs it most. In addition, moving the GPU to the CPU die reduces the number of required memory interfaces, also saving power.

AMD's smart power management wrings additional efficiency out of its APUs by providing a dedicated on-die controller to track the power consumption, temperature, and activity of all major components. This power microcontroller acts as the conductor for the "APU symphony", directing the processing emphasis to the right section at the right time. It enables fast response to thermal events and allows the controller to quickly allocate power to specific parts of the CPU in order to maximize performance and efficiency. It can also determine when units are mostly inactive and then reduce their operations to a minimal state or turn them off completely.

The latest version of AMD's smart power manager incorporates a 32-bit controller, as Figure 2 shows. It calculates system activity and measures thermal power to determine APU voltage and clock frequency. When activity is low, it cuts the voltage and frequency. When activity increases, it raises the voltage and frequency until the processor reaches its thermal limit.

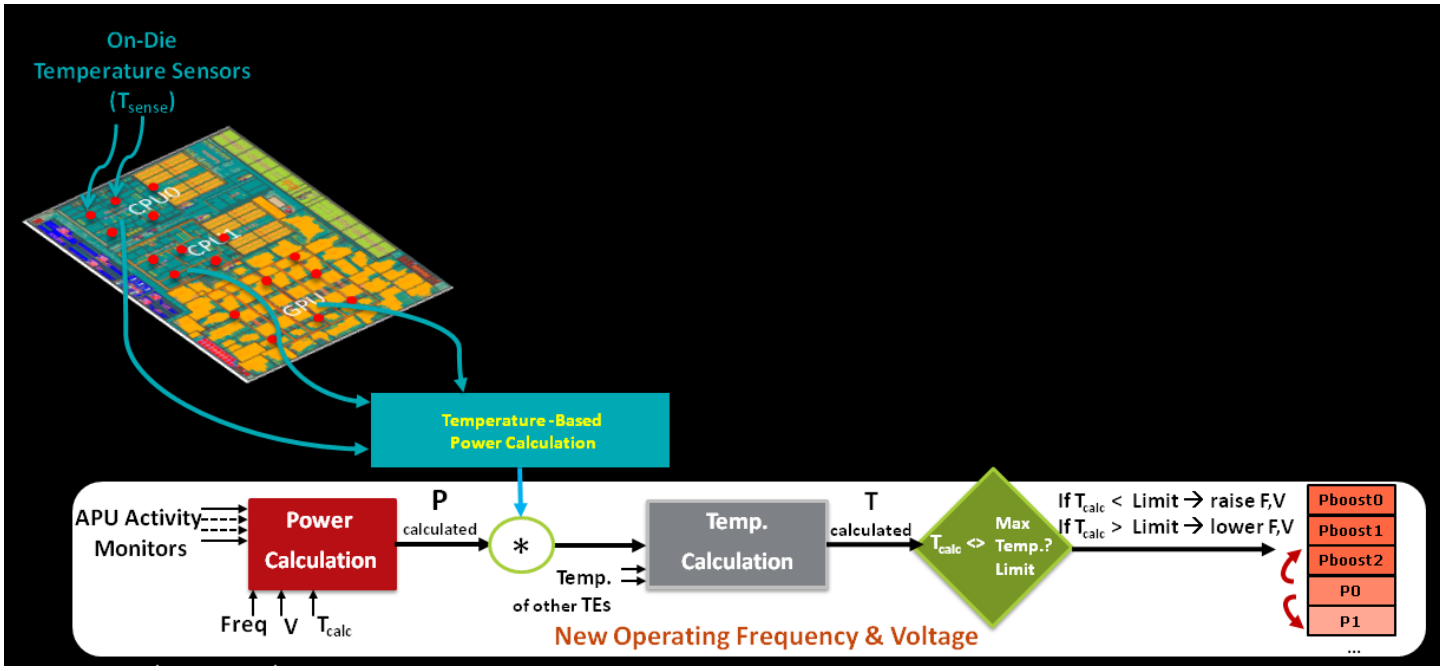

Figure 2 - AMD smart power manager

Often, a processing element achieves its greatest energy efficiency when it can complete its work in as little time as possible and then enter the deepest possible sleep state, as Figure 3 shows. Most consumer-oriented tasks, such as web browsing, document editing, and photo editing, benefit from this "race to idle" behavior. Using the GPU in concert with the CPU can allow the APU to complete its task sooner and then power down, reducing total energy consumption (energy is power multiplied by time). This power-state transition should be quick so the unit can shut down as soon as possible, enabling the processor to enter an idle state between user keystrokes or between frames during video playback.

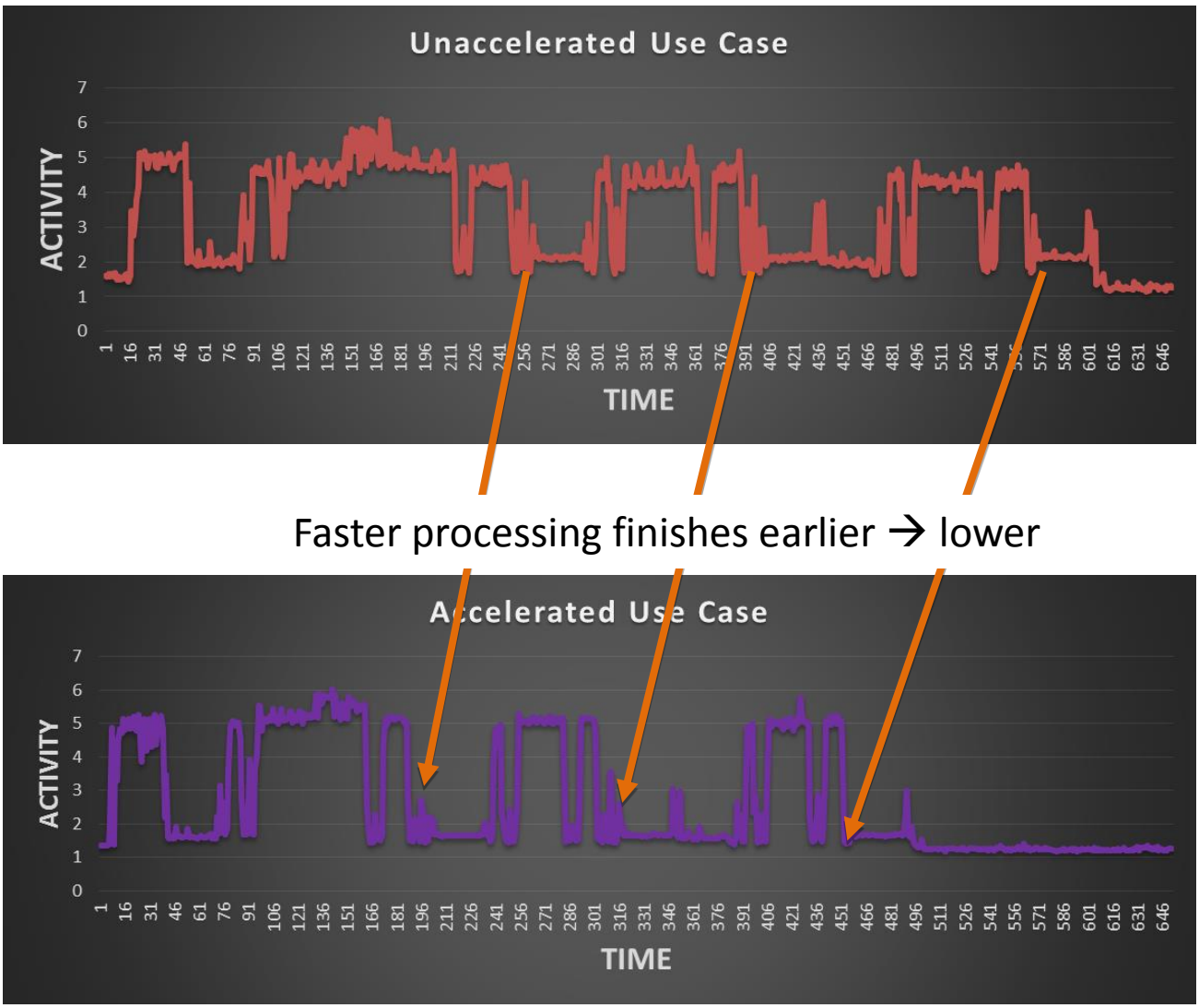

Figure 3 - Unaccelerated versus GPU-accelerated processing

The Future of Energy-Efficient Computing Is Heterogeneous

AMD is designing its leading-edge APUs to be compatible with the Heterogeneous System Architecture (HSA), a specification designed to help software use the power and performance of the GPU and other processing elements. When running highly parallel code on the GPU rather than the CPU, the APU can process workloads using the power efficiency and massive parallelism of its GPU cores, completing tasks faster. The HSA programming architecture enables routing of workloads to the best chip resource, which may, for example, be a specialized accelerator designed for a particular algorithm. HSA is designed to reduce the amount of cycles and power consumed for a fixed workload, and to enable advanced compute-intensive

applications to run within the power constraints of mobile devices. Advanced mobile applications include next-generation user interfaces such as voice recognition, gesture recognition, face recognition, and photo indexing, all of which can garner an order-of-magnitude performance improvement by turning to the GPU rather than the CPU.

Figure 4 shows the potential performance benefits that a heterogeneous architecture can bring as embodied in the exploitation of additional floating point operations per second (FLOPS).

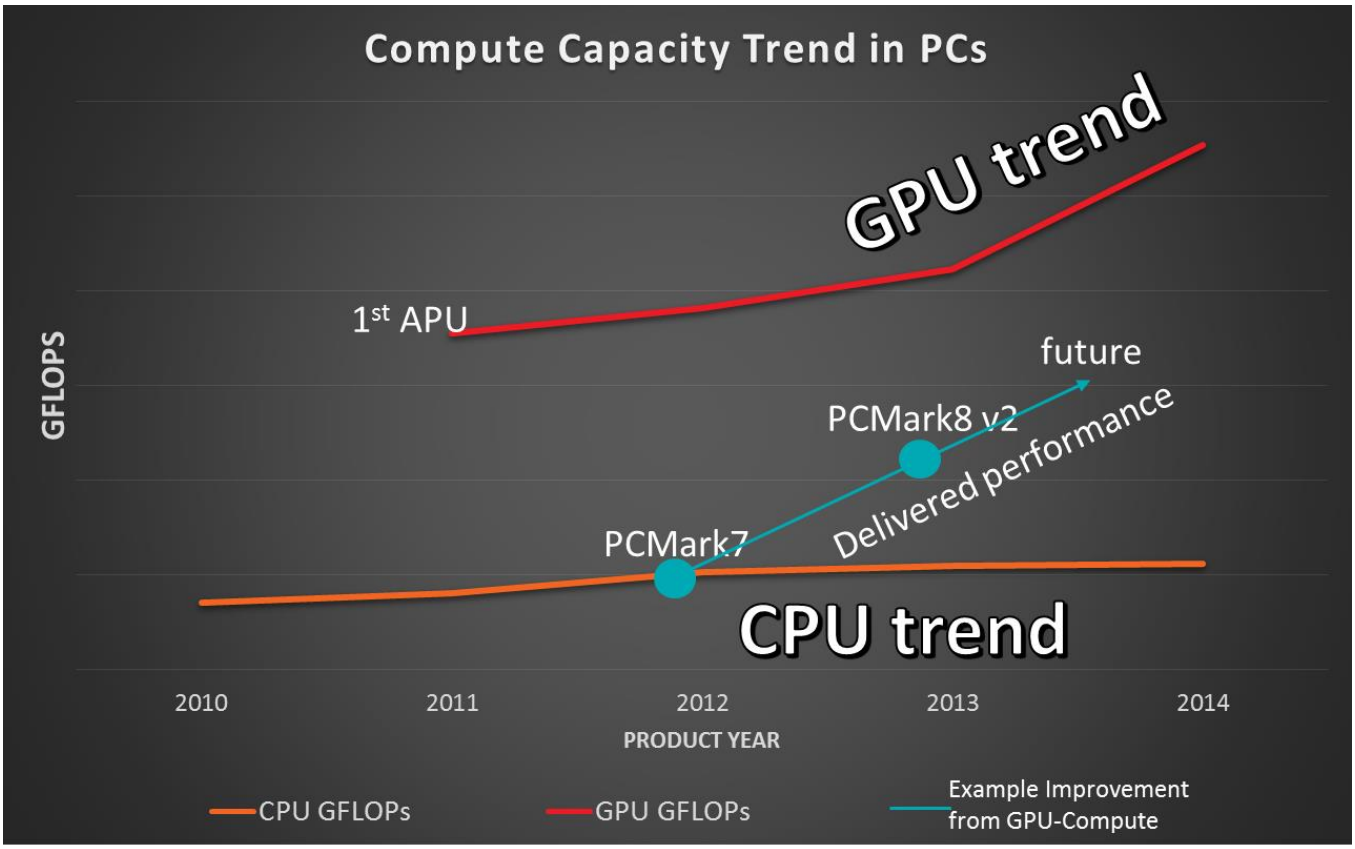

Figure 4 - GPU-compute trend in a typical 35W mobile processor

The performance of PC processors has increased over time, but at a relatively slow rate. On the other hand, GPU performance has ramped quickly as designers devote more silicon area to graphics in order to support displays with up to 4K resolution. Using the HSA architecture, AMD can tap into this GPU performance growth. PC applications and benchmarks are recognizing the benefits of employing the GPU for general-purpose tasks, as Figure 4 illustrates. OpenCL is one of the first industry-standard programming languages to support parallel computing on the GPU, allowing C programs to tap into language extensions that can potentially deliver orders-of-magnitude performance improvements on compute-intensive portions of the

code. The PCMark 8 v2.0 benchmark shows gains of up to 25% when OpenCL 1.x (a precursor to full HSA enablement in OpenCL 2.x) acceleration is employed, as Figure 5 depicts.

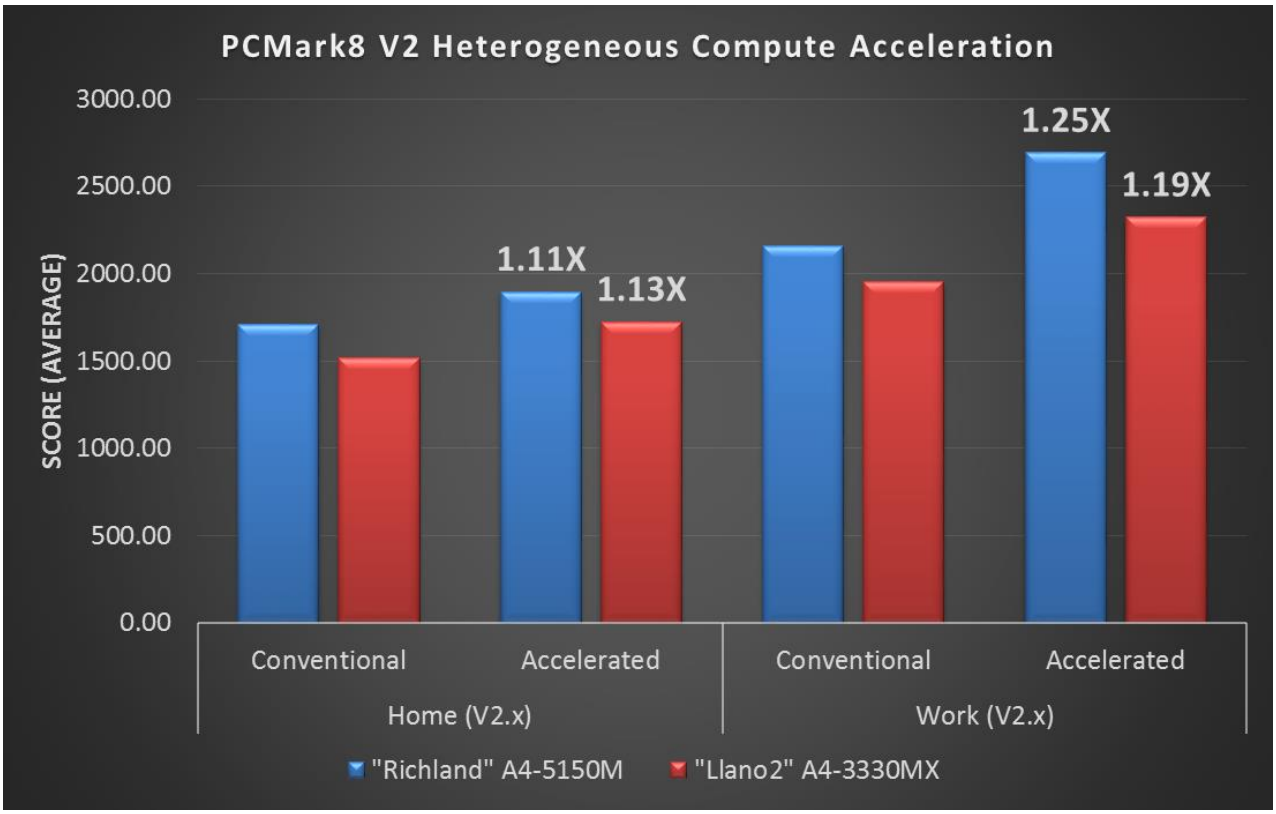

Figure 5 - PCMark 8 v2.0 acceleration through GPU offload in today's platforms

The heterogeneous architecture takes advantage of fast growth in GPU performance, which has far outstripped the performance of recent CPUs, as Figure 4 shows. The GPU will remain critical in allowing future processors to achieve both higher performance and better energy efficiency. Each GPU has multiple "shader" cores (which AMD calls "stream processing units"), each capable of processing integer or floating-point math while maintaining a smaller area and power profile compared with a typical CPU core. And because each shader core is small, a die can integrate many tens or even hundreds of them along with a single-digit number of general-purpose CPU cores. Hence, the GPU can produce orders-of-magnitude computation increases on workloads that are able to exploit these many processing cores. Each of the previously mentioned advanced applications can use the GPU's inherent parallelism to achieve these astounding performance gains, all while consuming a modest amount of power.

Figure 6 shows the decrease in typical-use power[6] starting with AMD's "Puma" CPU processor from 2008 and projecting to the company's APUs in 2020. The dotted grey line indicates the

power trend through 2015; beyond that point we expect slower decreases through 2020 as the power asymptotically approaches zero watts. Although typical-use power reductions will start to level after 2015, we project that heterogeneous computing will take off, driving higher performance levels commensurate with the GPUs' capability.

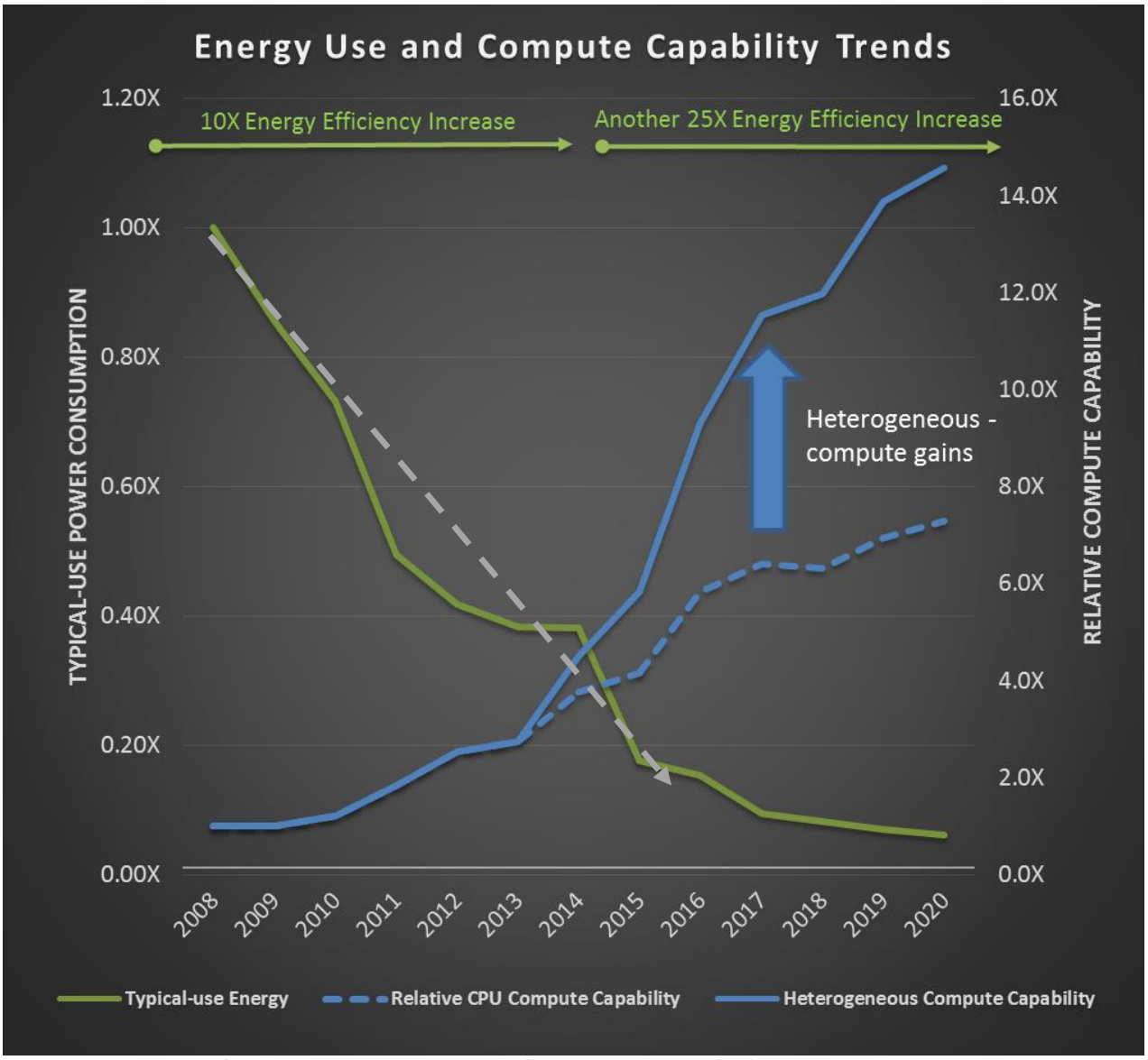

Figure 6 - Energy-use and compute trends for AMD APUs

The idle-power reductions helped reduce typical-use power from the nearly 4W "Puma" CPU in a 2008 platform to the 1.6W "Kaveri" APU of 2014, yielding a modest 2.5x baseline power improvement (both products fit in the same 35W thermal power envelope for notebooks). The

processor performance improvements come from more-numerous CPU cores, higher clock speeds, and the beginning of GPU-compute assist, with "Kaveri" delivering up to 4.5x the compute power of "Puma". "Kaveri" therefore offers more than a 10x increase in typical-use energy efficiency, as maximum performance over typical energy consumption indicates, compared with "Puma". We expect AMD's next major power reductions to arrive in 2015, using a series of improvements well beyond those found in "Kaveri". Future idle-power reduction will inevitably decelerate as consumption asymptotically approaches zero and platform component power starts to dominate. According to our analysis of AMD's product roadmap, we expect the company to demonstrate a remarkable drop in typical-use power between 2008 and 2020, representing a 16x improvement over this 12-year period.

On the basis of projected GPU performance increases and the growing number of applications that will tap into this heterogeneous-compute capability (the solid blue line in Figure 6), we project AMD will reach a parity point by 2020. At this point, approximately half of the delivered peak performance will come from the GPU and half from the CPU. Conservatively, we expect a 4x improvement[4] in compute performance between "Kaveri" in 2014 and a future APU in 2020 owing to both CPU and GPU advances. We expect performance to further improve owing to even larger GPUs and more-parallel applications. The reasonably conservative 4x estimate for the compute-performance increase, combined with at least a 6x reduction in typical-use power, yields a 25x improvement in AMD's ratio of maximum performance to idle power from 2014 to 2020. That's a fantastic dynamic range, offering the best of both worlds: low idle power to reduce energy use and high peak performance for tomorrow's advanced applications.

A New Computing World: Energy Efficiency Everywhere

The rise of cloud computing, mobile computing and big-data analytics has driven the need to process vast quantities of information under tight power, space, and cost constraints. An exponentially growing amount of data, most of it unstructured, requires significant computational capabilities to extract useful intelligence. Unfortunately, data centers that process cloud data cannot easily expand fast enough in both physical size and power to keep pace with this data growth. This dynamic has forced cloud servers to slash their power consumption in dense rack servers.

Even beyond big data and the cloud, computing is becoming ubiquitous. The present era of "surround computing" is also extremely power sensitive, thanks in part to Internet of Things (IoT) devices and always-on computing products that must "sip" power. The processors needed to support this surround-computing era must also be smart enough to apply the right processor element (i.e., CPU or GPU) to the right workload. In addition, markets will see a proliferation of

contextually aware intelligent devices, or agents, that will communicate with the user through natural interfaces and connect to a global information infrastructure. All this technology requires complex and power-efficient processing.

The goal of an energy-efficient processor is to deliver more performance using less power. For mobile devices, the desired result is longer battery life and lighter weight. For notebook PCs, it is to enable a cooler and quieter computer. Energy-efficient server processors seek to cut overall power consumption and reduce data-center costs—not just for the servers, but also for the cooling equipment. Greater efficiency can lead to lower total cost of ownership for electronic devices, and helps reduce the environmental impact of operations.

Processing these new services, intelligence, and data must use the same amount of power consumed today (or less), owing to the expense and environmental impact of the energy production that supports mega data centers. Despite current more-efficient processors, the power consumption and carbon footprint of these data centers continues to grow. The information- and communications-technology industry is currently responsible for 2% of global carbon emissions according to some estimates.[7] Although this percentage may seem small, the compound annual growth rate (CAGR) of emissions is increasing by 6% each year[8].

Leading the Industry to an Energy-Efficient Future

AMD is rapidly expanding its market breadth to drive energy-efficient computing into all of these new product areas. The company is applying its power-management technology and design methodology to processor development across its businesses. AMD's adoption of ARM cores for upcoming dense-server products should provide outstanding performance per watt for cloud computing. The company will also continue to invest in discrete GPU technology, with the goal of delivering even higher graphics and compute performance. Power-efficiency gains in these GPUs will come from the same technologies AMD developed for APU power management. APUs are where all of AMD's technological achievements combine to tackle the most challenging requirements for power-efficient computation. For example, the mobile-processor arena is where the company expects to see the predicted 25x typical-use efficiency improvement by the year 2020 through the use of heterogeneous computing and advanced power management—a 250x improvement relative to 2008 when AMD first embarked on its rapid push toward power efficiency.

AMD considers HSA to be an essential method both for improving performance on highly parallel workloads and for saving power by relying less on the CPU side of the APU. In addition to the GPU, the company is incorporating single-function accelerators into its APUs, paralleling

developments made by developers of smartphone chips. These dedicated accelerators are designed to deliver the most-energy-efficient performance in the smallest die area and for the least power. The disadvantage of dedicated units is the inability to adapt easily to new algorithms because of limited programmability. Alternative programmable units, such as digital signal processors (DSPs), excel at communications and audio processing, offloading those tasks from the CPU cores. For example, AMD has incorporated a Digital Sound Processor (DSP) in its newest APUs and GPUs to offload audio processing from the CPU. For some workloads, such as audio processing, these small additions to the architecture enable large power-efficiency gains that range from just over 2x to almost 25x compared[2] with processing solely on the CPU.

Looking even further ahead, we expect AMD will continue to invest in power-efficient design as it continues to diversify its business. Building on its success with integrating high-performance graphics, the company will further refine smart power management, heterogeneous computing, and power-efficient CPUs, GPUs, and accelerators; and it will continue to integrate and shrink its processors. As the semiconductor industry faces new challenges, such as the rising transistor costs that are reducing the historical benefit from process shrinks, AMD will be in an excellent position to take advantage of its early focus on power-efficient processor design.

References

[1] Energy Star Program Requirements for Computers Rev 6.0 (October 2013) specifies ETEC—typical energy consumption for notebook computers.
[2] Based on typical-use energy efficiency as defined by taking the ratio of compute capability as measured by common performance measures such as SpecIntRate, PassMark and PCMark, divided by typical energy use as defined by metrics such as ETEC (Typical Energy Consumption for notebook computers) as specified in Energy Star Program Requirements Rev 6.0 10/2013.
[3] Koomey, J.G., Berard, S., Sanchez, M., and Wong, H., "Implications of Historical Trends in the Electrical Efficiency of Computing," *IEEE Annals of the History of Computing*, Vol. 33, No. 3, 2011.
[4] As measured by common performance measures such as SpecIntRate, PassMark and PCMark.
[5] Bouvier, D., Cohen, B., Fry, W., Godey, S., and Mantor, M., "Kabini: An AMD Accelerated Processing Unit System on A Chip," *IEEE Micro*, Vol. 34, No. 2, 2014.
[6] Ibid.
[7] According to the International Energy Agency's (IEA) *2013 World Energy Outlook*, greenhouse emissions could increase 20 percent by 2035 under the agency's "central thesis," leading to Earth becoming about 3.6°C hotter. That's far above the international consensus of 2°C.
[8] The growth rate in greenhouse gas emissions from the ICT industry, i.e. end-user devices, telecommunication networks, and data centers, from 2002 to 2011 was 6.1 percent per year; the growth rate for 2011 to 2020 is expected to slow to 3.8 percent per year. GeSI, SMARTer 2020: The Role of ICT in Driving a Sustainable Future, December 2012; http://gesi.org/SMARTer2020.