AMD's Commitment to Accelerating Energy Efficiency

Sam Naffziger, AMD Corporate Fellow / IEEE Fellow

The remarkable performance advances in computing since the birth of the modern microprocessor can be largely attributed to Moore's Law—the doubling of the number of transistors on a chip about every two years as manufacturing technology advances allow for smaller and smaller transistors. Less well known is that this exponential rise in available processing power has been accompanied by corresponding improvements in energy efficiency and processor speed. A study in 2010 found that the amount of computation that could be done per unit of energy doubled about every 18 months, an observation known as Koomey's Law.[1]

However, the energy-related benefits that flow from Moore's Law are slowing down. This stems from the fact that the miniaturization of transistors is now bumping against physical limits, affecting the design parameters of processors. Historically, as transistors became smaller, the processor speed improved in tandem with the power efficiency. Now, however, scaling has slowed, forcing the industry to consider alternative ways to improve processor performance and efficiency.[2] How well the industry responds may have profound implications for the global economy and the environment as society further relies on digital technologies. AMD is at the forefront of devising technical solutions for improved performance and energy efficiency. Rather than only draw on the historical method—that is, manufacturing technologies for greater transistor density—AMD is developing new processor architectures, power efficient technologies, and power management techniques toward the goal of accelerating energy efficiency of its Accelerated Processing Units (APUs) 25x by 2020.

I. The need for energy efficiency in electronics

With the explosion of computing over the last 20 years and the resulting societal benefits across business, education, research, health care, and other sectors, the energy and environmental footprint from computing has correspondingly grown. The 3 billion personal computers in the world account for more than 1 percent of all the energy consumed annually; 30 million servers worldwide represent an additional 1.5 percent of all electricity used annually, at a cost of between $14 billion and $18 billion per year.[3]

---

[1] http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5440129
http://www.technologyreview.com/news/425398/a-new-and-improved-moores-law/

[2] Jon Koomey, Samuel Naffziger, 2015.
http://spectrum.ieee.org/computing/hardware/moores-law-might-be-slowing-down-but-not-energy-efficiency

[3] https://mitei.mit.edu/news/energy-efficient-computing

As more people around the world get online, the amount of space dedicated to data centers is projected to grow from 1.5 billion square feet globally in 2013 to nearly 2 billion square feet in 2018.[4] The servers in these computing centers will be connecting not only to PCs, phones, and tablets but also to an array of new connected devices and systems. Estimates vary, but conservatively twenty six billion units—everything from wearable computers to industrial sensors—could be connected to the Internet by 2020.[5] That means a sharp increase in Internet traffic, which is predicted to increase from about 245 exabytes in 2010 to about 1,000 exabytes in 2015.[6]

Compounding the need for energy efficient performance, smart phones, tablets, and gaming consoles are being used for more compute-intensive tasks, such as streaming videos, playing visually rich games, and augmented reality. Demands are climbing for mobile and desktop PCs as well, including video-editing, voice and gesture recognition, and data security based on biometric information. The result of all these factors is a strong market pull for technologies that improve processor performance while reducing energy usage.

II. Current state of the art in energy efficiency

Energy efficiency is one of the crucial enablers of the digital mobile revolution. Since the 1940s, the efficiency of computing has increased by several orders of magnitude,[7] making it possible to run laptops, tablets, and mobile phones for hours on a fully charged battery. With battery technology improving far slower than compute performance, however, mobile device makers need to employ a number of techniques to ensure users can operate for long periods on battery power. Smart phones and notebooks, for example, go into sleep mode after a few idle moments. These improvements can have a dramatic impact: If all computers sold in the U.S. were Energy Star certified, it would save $1 billion and reduce greenhouse gas emissions by 15 billion pounds annually, the equivalent emissions of 1.4 million vehicles.[8]

---

[4] http://www.datacenterknowledge.com/archives/2014/11/11/idc-amount-of-worlds-data-centers-to-start-declining-in-2017/
[5] Gartner Inc. "Forecast: The Internet of Things, Worldwide, 2013." 12/2013
[6] Cisco Visual Networking Index (2011)
[7] Jon Koomey, Samuel Naffziger, 2015. http://spectrum.ieee.org/computing/hardware/moores-law-might-be-slowing-down-but-not-energy-efficiency
[8] http://www.energystar.gov/products/certified-products/detail/computers

III. Power challenges in microprocessors

The 1980s and 1990s can be seen as a golden age for microprocessors in terms of both performance and compute efficiency. As designers leveraged smaller transistors and put more onto a chip, the clock speed, or frequency, of processors also increased, which allowed users to enjoy better-performing computers. But even as transistors got smaller, the power density, or the power required per given area, remained roughly constant – a phenomenon known as Dennard Scaling[9]. This meant that with each new processor generation, the energy use per unit of computing power went down by about a factor of four, with both voltage and capacitance decreasing.

However, in the early 2000s, the energy benefits of ever smaller transistors on a chip began to slow.[10] The primary reason has to do with fundamental physical limits at the transistor level. As transistors get smaller, leaking current becomes a greater engineering challenge because transistor threshold voltages have been reduced to the point where the devices don't completely shut off. This breakdown in Dennard scaling has resulted in higher power consumption for the highly integrated, high-performance devices that consumers want, requiring more complex cooling and creative power management techniques.[11]

The net result is that semiconductor makers can no longer rely solely on manufacturing improvements to achieve better energy efficiency. And going forward, even if engineers keep Moore's Law on pace with its historic performance trajectory, they will need new methods to further improve energy efficiency at the rates we've seen in the past.

IV. The AMD 25x20 Initiative

AMD engineers have observed these trends and the market demand for reducing the impact from information technology on the environment, as well as the need to improve battery life and increase the performance of ever-smaller form factors. They have responded with significant improvements in the efficiency of AMD's processors over the last several years. Recognizing the need to do even more, AMD

---

[9] Dennard, Robert H.; Gaensslen, Fritz; Yu, Hwa-Nien; Rideout, Leo; Bassous, Ernest; LeBlanc, Andre (October 1974). "Design of ion-implanted MOSFET's with very small physical dimensions". *IEEE Journal of Solid State Circuits* SC–9

[10] http://www.lithoguru.com/scientist/CHE323/Lecture2.pdf

[11] Koomey, J; Naffziger S (April 2015) "Moore's Law Might Be Slowing Down, But Not Energy Efficiency", http://spectrum.ieee.org/computing/hardware/moores-law-might-be-slowing-down-but-not-energy-efficiency

announced a goal in June 2014 to deliver a 25-times improvement in the energy efficiency of its Accelerated Processing Units (APUs) by 2020, or "25x20".[12]

AMD uses a "typical-use" efficiency index of platform performance divided by typical-use energy consumption to deliver a single measurement of work performed per unit of energy used. Usage profiles[13] clearly show that typical use is actually dominated by idle power, not peak computation power. There are an abundance of power related innovations that maximize idle time and reduce idle power without sacrificing performance. Performance is, of course, a key parameter – users want fast response times, quick calculations, and seamless video playback. They also want longer battery life, thinner and smaller form factors, and reduced environmental impact. Optimizing on typical use energy efficiency delivers on all of these vectors.

Meeting the 25x20 target will require stepping up the pace of typical use efficiency gains substantially by using new technologies and methods. With this goal, the reduced power consumption of AMD's products will outpace the historical efficiency trend predicted by Moore's Law by at least 70 percent from 2014 and 2020. That means that in 2020, a computer could accomplish a task in one fifth of the time as today's PC while consuming on average less than one fifth the power. Using a car analogy, this rate of improvement would be like turning a 100-horsepower car that gets 30 miles per gallon into a 500-horsepower car that gets 150 miles per gallon in only six years.[14]

V. Achieving 25x20

a. Architectural innovation

For decades, CPUs have been designed to run general programming tasks. They excel at running computing instructions serially using a variety of complex techniques, such as branch prediction and out of order execution, in order to improve speed. By contrast, graphical processing units (GPUs) are specialized accelerators originally designed for painting millions of pixels simultaneously across a screen. GPUs do this by performing calculations in parallel using a relatively simple execution pipeline. Historically, CPUs and GPUs have run as separate, though increasingly integrated processors.

AMD's accelerated processing units (APUs) place both CPUs and GPUs on the same piece of silicon. This has a number of benefits, including yielding better efficiency by sharing the memory interface, power delivery and cooling infrastructure. Many

---

[12] http://www.amd.com/en-us/press-releases/Pages/amd-accelerates-energy-2014jun19.aspx
[13] Annex B of the Energy Star standard development document http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-383.pdf
[14] http://www.amd.com/Documents/The-Future-of-Energy-Efficient-IT.pdf

workloads, such as natural human interfaces and pattern recognition, benefit from the parallel execution of the GPU and execute many times more efficiently when both GPU and CPU are used cooperatively. Optimizing concurrent GPU and CPU operation allows a machine to deliver its maximum performance, finish the task earlier, and go into a power-saving mode more frequently.

A long-standing challenge is the difficulty for software developers to write applications that take full advantage of both the CPU and the GPU. Traditionally, both processor types have each had their own separate memory systems. That means whenever the CPU wants to take advantage of the GPU, it has to copy the data from its memory to the GPU's memory. This makes it both inefficient and difficult to write applications, and limits GPU usage to only applications that have large data sets.[15] Having separate memories also adds to the power consumption, since the processor is often moving cached data between the CPU and GPU.

With the new Heterogeneous Unified Memory Access (hUMA) from AMD, the CPU and GPU use the same memory. Both can access all the platform's memory and allocate data to any location in the system's memory space. This coherent-memory architecture makes programming far easier because it doesn't require software developers to indicate where data is cached, an error-prone practice that can result in bugs that are difficult to detect and fix.[16]

The unified memory architecture provides a huge advantage in that it allows software programmers fluent in high-level languages such as Java, C++ AMP, and Python, to take advantage of the parallel processing capabilities of GPUs for better performance and efficiency. Recent results running a leading video and photo editing application show as much as a 17 times boost in performance for certain features when using the GPU's parallel processing in concert with the CPU.[17] But since it's a shared power/thermal infrastructure, the power demands are equivalent to using the CPU alone.

hUMA is part of AMD's of Heterogeneous Systems Architecture (HSA) implementation. These power and performance gains extend to other fixed-function

---

[15] http://developer.amd.com/resources/heterogeneous-computing/what-is-heterogeneous-system-architecture-hsa/

[16] http://arstechnica.com/information-technology/2013/04/amds-heterogeneous-uniform-memory-access-coming-this-year-in-kaveri/

[17] Path Blur within Adobe Photoshop obtained a speed up of up to 17x on mobile AMD FX-7600P APU. AMD tests are performed on optimized AMD reference systems. PC manufacturers may vary their configuration yielding different results. Test project used the "Path Blur" feature of the beta version of Adobe® Photoshop® CC (June 2014 release), test image: 4032x6048, .arw format. A notebook PC with AMD FX-7600P APU with AMD Radeon™ R7 Series graphics, 4GB DDR3-1866RAM, video driver 13.350.0.0 - 10-Mar-2014, Window 8.1 build 9600, took 12 minutes 25 seconds with OpenCL off, 41 seconds with OpenCL on.

devices, such as digital signal processors (DSP) or security processors when designed and programmed in accordance with the HSA architecture.

The AMD processor codenamed "Carrizo" is the first in the industry designed to be compliant with the Heterogeneous System Architecture 1.0 specification, developed by the HSA Foundation. The architecture leads to dramatically easier programming and greater application performance at low power consumption.

b. Power efficient silicon technology

The changing workloads of a computer affect the power usage of microprocessors. With more demanding workloads, such as complex server transactions or video rendering, processors draw more electrical current, which then falls during periods of lower demand. The sudden changes in current cause significant fluctuations in the chip's supply voltage. To deal with drops in voltage, which is known as droop, microprocessor designers typically supply excess voltage on the order of ten to fifteen percent to ensure the processor always has sufficient voltage. But over-voltage is costly in terms of energy because it wastes power at a rate that is proportional to the square of the voltage increase (i.e. 10% over-voltage means about 20% wasted power[18]).

AMD has developed a number of technologies to optimize voltage. Its latest processors have a circuit to track voltage, comparing the average voltage to droops on the order of nanoseconds, or billionths of a second[19]. It recovers most of the wasted power by operating at the average voltage and then quickly reducing frequency for brief periods to counteract dips in voltage supply. Since the frequency adjustments are done at the nanosecond level, there's almost no compromise in computing performance, while power is cut by 10 to 20 percent. Starting with the "Carrizo" APU, this voltage adaptive operation functions in both the CPU and the GPU.

Another power technology that debuts with the "Carrizo" APU is called adaptive voltage and frequency scaling. What this technology involves is the implementation of unique, patented silicon speed capability sensors, and voltage sensors in addition to traditional temperature and power sensors. Silicon speed capability and voltage control can vary significantly part-to-part and platform-to-platform, which is widely understood in the industry. These variations are traditionally dealt with by adding margin, or "guard-band," to the silicon operation since the exact behavior can't be known in advance. Such guard-bands cause significant efficiency losses over what a theoretically optimal system could achieve. With the introduction of AMD's new adaptive sensors and associated control algorithms, much of this efficiency loss can

[18] Switching power = $C*V^2*F$, so at a new voltage V2 that is 10% higher than V1, $P \alpha (V2/V1)^2 = (1.1/1)^2 = 1.21$

[19] Grenat, Aaron et al, International Solid State Circuits 2014 digest of technical papers pp 106-107

be mitigated. The speed and voltage sensors enable each individual APU to adapt to its particular silicon characteristics, platform behavior, and operating environment. By adapting in real-time to these parameters, the APU can dynamically optimize its operation for maximum efficiency, squeezing up to 20 percent power savings at a given performance level[20].

Finally, to reduce power use on the CPU specifically, AMD has leveraged a high-density library more similar in design style to a GPU. This choice allows AMD to squeeze more standard cells — the building blocks that make up the processor — into less area, reducing area and routing distance between cells and reducing power significantly. Using a high-density library generally means somewhat lower speed at the same voltage, but when implemented correctly, can save power and area by 30 percent[21], which means that in power constrained situations (which is almost always the case), the frequency and performance are actually higher than with a traditional high performance library implementation. In addition, it frees up space on the chip to allow AMD to place the GPU, a multimedia processor, and the system controller all on the same chip.

c. Power management

Most computing platforms run at peak power only a small fraction of the time. To minimize power consumption while maximizing performance, AMD has designed power management algorithms to optimize for typical use, rather than peak computation periods that only occur (briefly) during the most demanding workloads. The result is a number of race-to-idle techniques to put a computer into sleep mode as frequently as possible to reduce the average energy use.

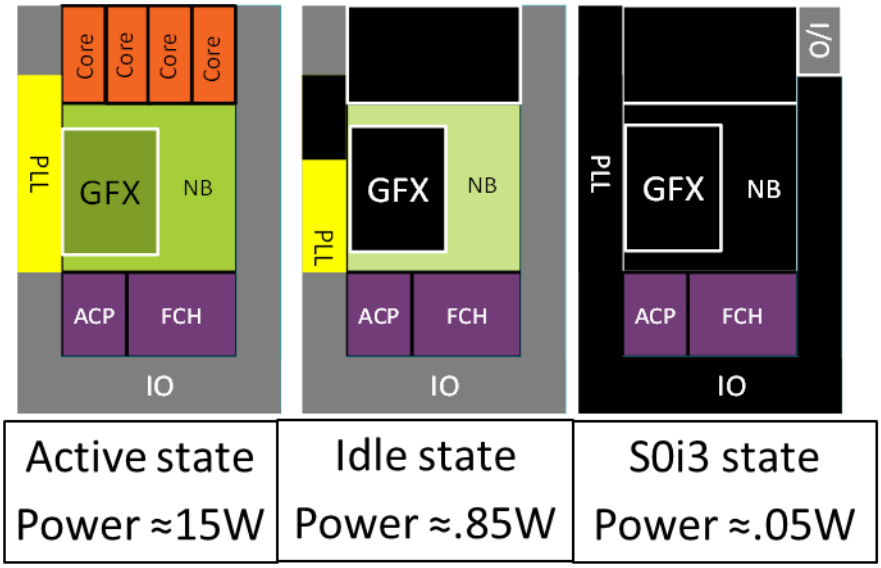
AMD integrates system components on a single die, including the GPU, memory controller, I/O controller, and peripheral buses. This allows for very fine-grained monitoring and management of power, temperature, and activity of all the system components. The power controller directs processing between CPU and GPU to optimize performance and efficiency. With this level of control, it can put the processor in idle mode as frequently as between frames of a video playback or keystrokes while typing, or after a Web page loads. As the performance of the integrated components is increased, tasks finish sooner and can therefore spend more time in idle modes – a "virtuous cycle" of higher performance and lower power that synergistically improves efficiency.
AMD's power management also monitors the temperature of the silicon and the end user machines. Based on the activity of system components, it can determine the temperature of a PC or mobile device and whether it feels too hot to the end user. With this understanding, the APU can briefly increase the power output during

[20] "A 28nm x86 APU optimized for power and area efficiency." http://ieeexplore.ieee.org/Xplore
[21] Ibid

compute-intensive jobs by increasing processor frequency to deliver strong performance without overheating a notebook PC or convertible laptop. When the task is complete, the power is dynamically reduced, thus lowering the device temperature. This practice yields better overall energy efficiency because tasks are performed quicker and a machine can rapidly shift to idle mode while delivering a responsive experience.

The power management micro-controller also tracks the behavior of specific applications in real time and determines how much they will benefit from higher processor frequency. Applications that don't benefit from higher frequency, which requires more energy, operate at lower frequency than the processor's maximum capability, avoiding wasted energy.

Another capability incorporated in recent AMD APUs is around run-time entry of the processor into the extremely low-power S0i3 state. The use of this low power state is OEM/platform-specific (i.e. can be connected standby, modern standby or suspend-to RAM), but this state has almost all of the APU silicon power-gated and all relevant I/O devices in their low-power states as well, driving platform power to extremely low levels. The diagram below illustrates the power gating of the APU under these conditions. The S0i3 state enables the platform to achieve the same power level as the legacy S3 state, traditionally known as "standby," which is very time-consuming to enter and exit because it requires operating system intervention. By doing this on the fly, under the control of the integrated power management micro-controller, the APU can achieve standby equivalent power levels transparently at sub-second time frames if activity levels are low enough. This translates directly to lower average power consumption for typical use conditions.



There are also a number of other efficiency-oriented features in recent AMD products, such as video and audio acceleration, and the AMD development pipeline

includes adaptive I/O optimization and compression, more fine-grained voltage management, and energy optimization based on workloads.

Summary

The energy consumption incurred by society's growing use of information technology is colliding with the slowdown in Moore's Law efficiency gains. In the future, the bulk of the energy-efficiency gains will stem from new circuit designs and power-management techniques, rather than transistor scaling as it has in the past. In response, AMD has built a portfolio of intellectual property around architectural innovation, power-efficient technology and power management techniques. These innovations are being incorporated into all of AMD's products and will allow its processors to outpace the historical efficiency trend by 70 percent[22], despite the slowdown in silicon scaling. The company's resulting goal is to improve typical use energy efficiency 25 times between 2014 and 2020 in its APUs to enable new digital experiences while limiting energy use.

Sam Naffziger *is a Corporate Fellow at AMD responsible for low power technology development, and has been the key innovator behind many of AMD's low power features. He has been in the industry twenty seven years with a background in microprocessors and circuit design, and has been at AMD since 2006. Sam received a BSEE from CalTech in 1988 and MSEE from Stanford in 1993, and holds 115 US patents in processor circuits, architecture and power management. He has authored dozens of publications and presentations in the field, and is a Fellow of the IEEE.*

[22] AMD Accelerates Energy Efficiency of APUs, Details Plans to Deliver 25x Efficiency Gains by 2020. http://www.amd.com/en-us/press-releases/Pages/amd-accelerates-energy-2014jun19.aspx