EPYC:
A Study in Energy Efficient CPU Design

[How EPYC Does More Work with Fewer Watts]

Nathan Brookwood
Research Fellow
Insight 64

AMD sponsored the creation of this white paper, but the opinions and analysis are those of the author.

The lights in neighborhoods surrounding the University of Pennsylvania dimmed in 1946 whenever the engineers at Penn fired up ENIAC[1], the world's first digital computer. The monster machine performed over 18,000 calculations per second, an amazing speed at the time, but ENIAC's 17,000 vacuum tubes dissipated over 160 kilowatts. Hence the angry neighbors in their darkened homes.

In the 71 years since ENIAC, computer designers have increased the performance of systems by a factor of approximately 50 *billion*. Fortunately, they improved the performance per kilowatt hour of these systems by a factor of 50 *trillion*, thus allowing a two-socket AMD EPYC™ system to handle roughly one trillion operations per second while consuming less than 450 watts of power. Jonathan Koomey, a computer scientist at Stanford who focuses on this area, has tracked improvements in peak system performance per watt since the dawn of the computer era. He captured this trend in his eponymously-named Koomey's law that states "The number of computations per joule of energy dissipated doubles approximately every 1.57 years."[2] This trend held over transitions from vacuum tubes to discrete transistors to VLSI microprocessors. Since the date of his original publication in 2010, the doubling time has increased from 1.57 years to between 2.4 and 2.7 years, but Koomey argues that by adjusting for actual system usage models, the original trend still holds.[3]

So-called "laws" like Koomey's (or the more famous Moore's) don't magically enable product improvements but serve as guideposts that drive designers to push the limits of technology and set long-term goals based on accumulated industry experience. In 2014, AMD declared its intent to improve the energy efficiency of its client APU products by a factor of 25 by 2020 – its 25X20 program – a pace roughly fifty percent faster than the historic factor of 1.57 every two years. Rather than simply relying on advancements in silicon technology to drive these improvements, AMD took a holistic approach and enhanced energy efficiency through a combination of processor architecture, advanced power management, optimized physical design and process-tuned device libraries. This methodology enabled it to improve the power efficiency of its 2016 "Bristol Ridge" APU by a factor of four over its 2014 "Kaveri" design, while using essentially the same 28nm manufacturing process in both. In its latest Ryzen™ and EPYC™ processors, AMD applied this same methodology to GlobalFoundries' advanced 14nm FinFET technology to achieve additional improvements in performance per watt.

CPUs play a key role in determining overall performance in balanced systems, but also account for more than half the power those systems consume. Since performance per watt is computed as a ratio of system performance (by some measure) divided by power consumed, a power-optimized CPU contributes to both the numerator and the denominator of the equation. Optimizing this aspect of its design became a key goal for AMD's architects and engineers, and this focus shows up in almost all aspects of the resulting products.

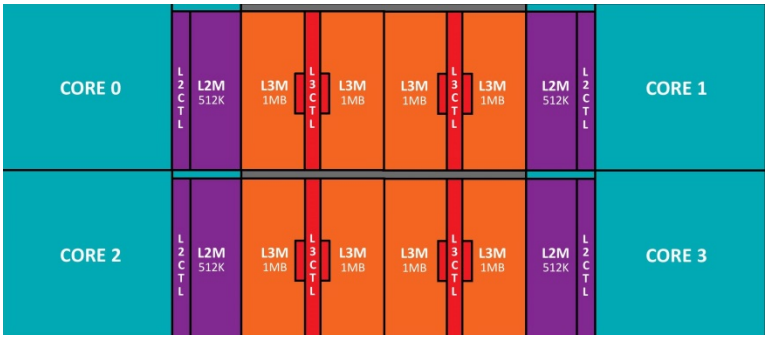
The AMD EPYC Architecture: Designed to Optimize Performance per Watt

Chip architects seldom get opportunities to create entirely new processor microarchitectures. Often, they must extend a prior version of a chip or a core or adapt it to a new technology node. This incremental approach constrains the types of changes they can incorporate in new designs and limits their ability to create breakthrough products.

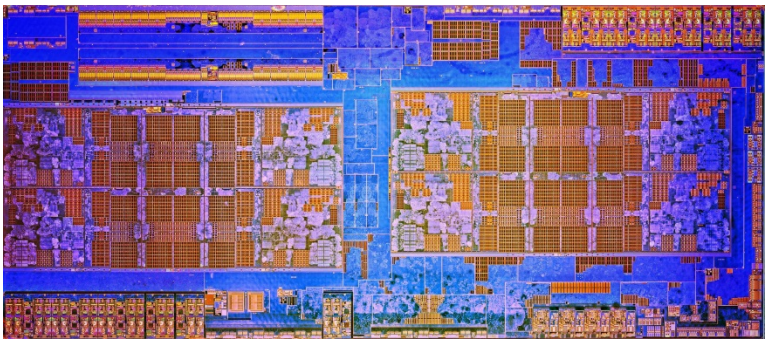
In 2012 Lisa Su, now the CEO of AMD, challenged a team of the company's senior architects and engineers to develop a new CPU core that could enable AMD to regain its competitive position in the server market. Those engineers recognized that performance per watt plays a key role in all the company's markets. In servers, it enables greater compute density for scale-out environments, and lower operating expense (OPEX). In laptops it provides cooler operation and longer battery life. In desktops, it provides higher performance within acceptable thermal limits. The team's initial server concept presumed the use of a large monolithic die, but early on they concluded the economics of manufacturing a die large enough to accommodate all the cores, caches, I/O and memory interfaces needed would have been marginal at best. [4] They investigated a variety of schemes to partition their design over two or four dies, and found that four identical chips linked via a high-speed fabric, each containing eight cores along with their associated memory and I/O systems, would provide the most

competitive solution. AMD adopted that approach, and now brands the products it spawned EPYC™ processors. AMD's single die offerings (Ryzen™) and dual-die offerings (Ryzen™ Threadripper™), aimed at desktop markets, are byproducts, albeit highly valuable byproducts of its EPYC strategy.

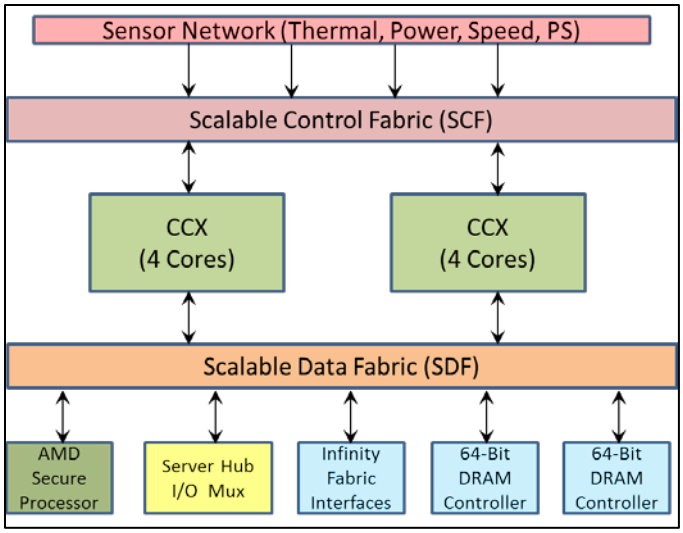
To understand how EPYC achieves its ground-breaking performance per watt, it helps to understand what's inside an EPYC multi-chip package. AMD starts with a quad core plus cache logic block known as a "CPU Complex" or CCX. It combines two of these CCX blocks with a variety of "uncore" elements to create a die codenamed "Zeppelin" that is used singly in Ryzen or packed four to a multichip module in EPYC. In addition to the CCX's, "Zeppelin" contains two 64-bit DRAM controllers, three Infinity Fabric interfaces (that connect to other chips in the package), and 32 PCIe Gen3 lanes. It also includes a Server Control Hub (SCH) that supports basic I/O features, and an AMD Secure Processor that handles secure boot operations and manages the keys for Secure Memory Encryption (SME) and Secure Encrypted Virtualization (SEV) operations. All these elements talk to one another over a coherent Scalable Data Fabric (SDF) controlled by a platform-wide Scalable Control Fabric (SCF). This all fits on a 4.8 Billion transistor, 14nm FinFET chip that measures only 213mm^2. Amazing!



EPYC's CPU Core Complex (CCX)
Four Cores and 8MB L3 Cache



EPYC's Single Die ("Zeppelin")
Eight Cores and 16MB L3 Cache in 213mm^2!



Zeppelin's major logic blocks

The "Zeppelin" die used in AMD's EPYC and Ryzen offerings benefit from Moore's Law and the move to a 14nm FinFET technology from the previous generation's 28nm process, but that only attributes to some of the efficiency gains. A 52% increase in instructions per clock (IPC) from the "Zen" core vs. "Excavator" delivers another significant increase in performance-per-watt. New power management techniques and power optimized circuity also contribute to the gains in energy efficiency on EPYC. No single silver bullet accounted for all the gains. In the next section, we will dive more deeply into each of these areas and see how each enhances EPYC's performance per watt. Then we will examine the role AMD's Infinity Fabric plays in connecting EPYC's four "Zeppelin" die in an energy-efficient manner.

"Zen" Architecture: A huge leap forward

Chip architects rarely get a chance to design all the elements of a new device from scratch. New architectures (or more accurately, microarchitectures[5]) must accommodate the manufacturing technology constraints that limit initial implementations but must also fit with new technologies that may become relevant during the architecture's lifetime. This complicated balancing act, part art and part science, entails great responsibility. The decisions these architects make can impact their products' competitiveness for a decade or more.

Many "Zen" features improve performance and save power, contributing to both the numerator and the denominator in the performance per watt equation. For example, "Zen" is AMD's first CPU architecture to incorporate a cache that stores instructions after they have been decoded into micro-ops. The next time the CPU encounters that same instruction, if it can find it in the micro-op cache, it can skip the decoding stage, saving time (two clock cycles) and the energy needed to decode it again. "Zen's" large Level 3 caches (8MB per four CPU complex) reduce cache misses, and thus help eliminate time- and energy-consuming accesses to main memory, especially when compared with "Excavator", which had no on-chip L3 cache.

Speaking of caches, "Zen's" Level One (L1) caches use a write-back management policy, so changes in L1 data don't incur a power penalty until the entire cache must be written back to the L2 cache. Similarly, moving the contents of one CPU register to another is accomplished by simply renaming the target register, rather than loading and saving the register. A picoJoule here, a picoJoule there, it all adds up.

Modern speculative out-of-order processor pipelines need to decode instructions and load data long before they know how upcoming branch instructions will impact program flow. If they guess wrong, all their speculative work goes up in smoke. EPYC uses perceptrons, a simple form of a neural network, to enhance its branch prediction accuracy. Better branch prediction translates into higher performance and less wasted power, and helps with both sides of the performance per watt equation.

"Zen" is also AMD's first CPU architecture to support simultaneous multi-threading (SMT), a feature that enables a core to execute two threads concurrently so that it can continue to do useful work when either thread stalls. This feature increases performance with little impact on power consumption, and helps the numerator of the equation.

Together, "Zen's" architectural innovations substantially improve its performance per watt over the earlier Excavator implementation.

"Zen's" Process Technology: Closing the gap with Intel

Semiconductor manufacturers characterize their technology in terms of process nodes, where each node is loosely defined by the dimensions of the smallest transistors it can produce. Over the past two decades, the industry has moved from 65nm to 45nm, 28nm, 22nm, and 14nm features at roughly three-year intervals, although the move to the next generation nodes (10nm and 7nm) has slowed. For most of the time AMD has battled Intel in the processor market, Intel has had a decided process technology advantage; Intel moved from 32nm to 22nm and then 14nm nodes while AMD continued to rely on mature 28nm technology from its manufacturing partners, GlobalFoundries and TSMC.

With its new generation of "Zen"-based processors, AMD has narrowed the gap between its manufacturing technology and Intel's considerably. Purists can (and do) argue about the merits of Intel's 14nm technology (recently renamed "14++" to demonstrate progress) versus the similarly numbered process from GlobalFoundries. Both companies now use three-dimensional FinFET[6] transistors instead of the planar (flat) transistors that the industry relied on from the mid-1970s until the early part of this decade. The technologies are now close enough that at the 2016 Solid-State Circuits Conference (ISSCC) AMD's engineers could tout that the "Zen" core in their new designs was about ten percent smaller than the comparable core Intel had revealed for its 14nm "Skylake" design ($44mm^2$ for AMD versus $49mm^2$ for Intel).[7]
The move from 28nm to 14nm gives "Zen" processors the ability to have larger L1 and L2 caches than the earlier AMD "Excavator" based part and adds an 8MB L3 cache to each CCX. This enhanced cache architecture boosts performance, while the FinFET technology lets "Zen" operate at lower voltages and save power. Together, these changes enable significant improvement in performance per watt compared to "Excavator".

Of course, semiconductor technology continues to move forward. To paraphrase Mark Twain, rumors of the death of Moore's law are a bit exaggerated. In June, GlobalFoundries announced its 7nm process will be ready for prime time in 12 to 18 months. GF claims its 7nm process will be 40 percent faster, use 60 percent less power, and be denser as well. TSMC, AMD's other foundry partner, is racing GlobalFoundries to market with its own 7nm technology. AMD has already taped-out its next-generation 7nm server processor, dubbed "Rome". Rome will use the same SP3 socket and infrastructure as the 14nm version of EPYC ("Naples"), which should make it easy for customers and system partners to migrate to the new version.

PicoJoules matter! Advanced Power Management

While most of AMD's engineers focused on developing "Zen" in all its forms, the company still needed to keep its 28nm products competitive to pay the bills. During that period, its engineers refined several approaches to optimize power consumption, knowing those same techniques could be used in its upcoming 14nm products as well. To squeeze every last joule out of each compute operation, "Zen" employs many of the techniques AMD tried out in its earlier "Kaveri", "Carrizo" and "Bristol Ridge" APUs, along with other saving techniques that are new to "Zen", including Precision Boost and workload aware power management. Several new features give customers the ability to fine tune the performance per watt characteristics of their systems. Let's examine these more closely.

EPYC has the ability to fine tune the voltage and frequency of each core in a four-die processor module. Each "Zeppelin" die includes over 1300 sensors that monitor temperature, voltage, power and silicon speed once each millisecond, and deliver this data over its Scalar Control Fabric. It knows

which parts of the chip are working and which are loafing, and uses this information to fine tune the voltage and frequency of each core in the system. AMD's fine-grained power management, known as Precision Boost Technology, adjusts frequencies in small 25MHz increments, so cores don't need to stop to synchronize their PLLs. Of course, knowing which core needs more power, and delivering that power is no easy feat, but here too, AMD has a solution. Based on the reports from each core, the system sets the minimum voltage needed to allow the busiest core on the chip to operate properly. LDOs (Low Drop Out linear regulators) attached to the power grid adjust the voltages and frequencies of all other cores. To absorb what can be almost instantaneous surges in power demand, AMD embeds insulators between two of the top layers of the chip's power distribution network to form Metal-Insulator-Metal capacitors (MIMcaps) that source current to reduce short term voltage drops. All these features improve EPYC's performance per watt. But AMD didn't stop there.

Workload Aware Power Management: One of the classic techniques designers use to reduce processor power consumption is to follow a HUGI (Hurry Up and Go Idle) approach, in which cores run as fast as they can to complete a task, and then drop into an idle low-power state. In practice, this often turns into a Hurry Up and Wait (HUW) scheme, where non-latency-sensitive tasks run at maximum speed, using more energy than they need to meet critical timing requirements. AMD developed an algorithm that detects this type of behavior and lets non-latency-critical tasks run in a more leisurely and less power-hungry manner. This feature further improves EPYC's performance per watt.

Configurable Power Limits: FinFET technology allows EPYC processors to operate over a wide range of thermal design points (TDPs). Chips spec'd at 180 Watts can run at TDPs ranging from 165 to 200 Watts. Parts spec'd at 155 Watts can run on as little as 140 Watts or as much 175 Watts. Low power 120 Watt versions can actually operate on as little as 105 Watts. EPYC's ability to run at lower power than its TDP label indicates means a customer can qualify the part once, and then deploy it in a variety of applications where datacenter power constraints may force systems to run at reduced power levels.

Deterministic Performance or Power: Datacenter operators must sometimes contend with the reality that not all chips in their servers will exhibit identical operating characteristics, even if all share a common part number and meet the supplier's specs. Ambient temperatures vary from rack to rack; airflow varies from chassis to chassis in a rack, or even within a chassis. Minor variations in a chip's recipes cause some devices to be more impacted by current leakage than others. As a result, customers may find their software programs run faster on some supposedly identical systems than on others. Some might see these minor speed-ups as an added bonus, but others demand repeatable, deterministic performance. Not to worry. With EPYC, customers can configure their systems to run in a power-capped mode, where the processors deliver all the performance they can within that cap, or they can opt for a deterministic performance mode, where system power consumption is allowed to vary. As Burger King might say, "Have it your way."

Physical Design Still Matters

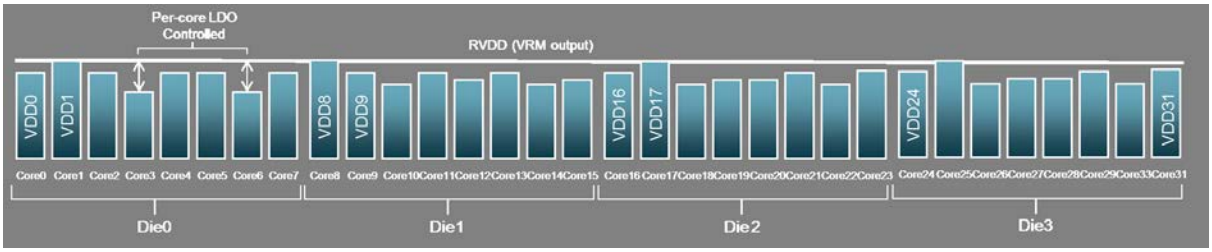
These days, fewer and fewer engineers who call themselves "chip designers" worry about the physical design of their chips. They create incredibly complex logic blocks and toss them over the wall to a group that translates their logic into silicon. In effect, their skills are more aligned with those of software developers, rather than the hardware developers of yore. But for companies like AMD and Intel, not to mention the dying breed of analog designers, the floor-planning of a chip and the custom circuitry that makes that chip go faster still matter.

AMD devoted considerable effort to improving its cache performance, since caches play a huge role in overall computational performance. They moved from their earlier write-through L1 data caches to a write-back model that saves energy. They made their L2 and L3 caches bigger and dramatically reduced latencies, thus minimizing the performance impact of misses in lower-level caches.

The gospel of saving power was thoroughly diffused throughout AMD's design organization. Aggressive clock gating was the norm. Clock distribution networks, notorious for their power-hungry ways, were particularly suspect. AMD's design methodology paid close attention to optimizing the size of the

transistors used in its circuit library with regard to performance requirements and power. This can be a bit tricky with FinFET transistors, since the fin in a "FinFET" is actually a fixed-size vertical gate, and to increase drive current, circuit designers vary the number of fins. Using transistors that are too big wastes energy and creates thermal problems but using transistors that are too small leads to performance issues that can break the chip's timing or limit frequency scaling. As far as we can tell, AMD's engineers hit upon a design that would make even Goldilocks happy.

EPYC's physical design plays a key role in enabling the Power Boost technology that allows each core in a 32-core single-socket system or a 64-core dual-socket system to get the precise supply voltage it needs to support its particular operating frequency. From an energy-efficiency perspective, it's important to regulate the voltage to each core individually, so that it runs at the lowest possible voltage consistent with its clock frequency[8]. To accomplish this, a platform voltage regulator module delivers the lowest voltage (RVDD) needed to drive the fastest core in the system at any given moment, based on speed monitored by the SCF (Scalable Control Fabric). The power rail of each individual core connects to this RVDD supply via a linear regulator header device which is pretty much just a big transistor. Simple low-dropout linear voltage regulators (LDOs) deliver the fine-tuned voltage to the rest of the logic in the core, distributing this on a cut-out of the thick package power plane. Since the large L3 cache arrays on the chip operate at a constant speed, a separate LDO is used to power these portions of the chip.



A combination of MIMcap and LDO circuitry delivers the precise voltages needed to drive each core

Energy Efficiency is in AMD's DNA

These days most suppliers tout their dedication to energy efficiency. Performance per watt matters in laptops and mobile devices where it affects battery life, but also in data centers, which consumed about 70 billion kilowatt-hours of electricity in 2014, representing two percent of the country's total energy usage. AMD is no Johnnie-come-lately to the energy efficiency party. In 2003, when AMD introduced Opteron™, the industry's first 64-bit x86 server processor, its performance and performance per watt dramatically surpassed competitive products. The differences were so stark that in 2006 AMD erected billboards in Times Square, Silicon Valley and in Austin that continuously displayed how much customers could save by switching to AMD-based servers.

An On-die Server Control Hub makes EPYC a True SoC

Every server requires a control hub to handle the run-of-the-mill I/O interfaces found on almost all systems. "Why not incorporate a hub on the server die itself, rather than use a discrete chip?" AMD's engineers asked. Control hubs often consume five to eight watts, and in the discrete case much of that power just goes into inter-chip communications. The EPYC team put a Server Control Hub (SCH) on the "Zeppelin" die, and tied it into the Scalable Control Fabric, where it can observe the activity of all the chips and links in the system. It provides four USB 3.0 ports, two SMBus ports, platform clock generation, along with a few other goodies everyone notices only when they're absent. This simplifies the design of EPYC motherboards, and saves a few watts as well, as compared with competitive platforms that still require external control hubs.

Parting Thoughts

We began this paper by looking back at ENIAC and the dawn of the electronic computer era. We're going to conclude it by looking back to the dawn of the semiconductor industry[9], and what that history tells us about two of its key players.

The decade from 1965 to 1975 saw an explosion of new semiconductor companies, as entrepreneurial technologists employed by Fairchild Semiconductor became frustrated by its management practices and set out on their own. Soon the Santa Clara valley[10] became home to a plethora of high tech companies with names like Intel, AMD, National Semiconductor, Avantek and others, collectively known as the "Fairchildren." The companies with the smartest engineers and marketers started shipping products. IPOs followed and soon employee parking lots overflowed with Porsches and BMWs. Amazingly, only two of the original Fairchildren still operate independently today – Intel and AMD. Why have these two survived while all the others fell by the boards or were gobbled up in corporate acquisitions?

Company	Year Founded	Year Acquired	Acquirer
AMD	1968		
Avantek	1968	1991	HP
Intel	1967		
Intersil	1967	1988	Harris Semi
Linear Technology	1981	2017	Analog Devices
LSI Logic	1981	2014	Avago
National Semi	1967	2011	Texas Instruments
Signetics	1961	1975	Philips (Now NXP)
Synertek	1973	1979	Honeywell

The Fairchildren

Over the nearly five decades since their founding, astute technologists at both companies have succeeded in discerning important trends and targeting products that intercepted those trends in a timely manner. Intel has been particularly effective in driving process technology, and in creating the "x86 franchise" that drives much of both companies' revenues today. But AMD has played a key role in preserving that x86 franchise. Intel's original deal with IBM (to put an Intel 8088 in the original IBM PC) required that Intel enable a second compatible CPU source, and Intel turned to AMD to play that role. Twenty years later, Intel aspired to shift the industry away from x86 to its new and incompatible 64-bit Itanium architecture. AMD's x86-64 architecture provided an easier pathway to 64-bit systems and Intel was forced to follow. A few years later, Intel was still trying to improve performance by using ever higher clock speeds that generated more heat than performance. AMD again led the way to energy-efficient multicore processors with its dual-core Opteron™ processor.

We're now witnessing another important divergence in the two companies' strategies. As we noted earlier, AMD's engineers invented the Infinity Fabric and used it to distribute 32 "Zen" cores over four separate chips in an MCM. Intel's engineers took a more conventional approach and crammed 28

Skylake cores onto one reticle-busting 677mm^2 chip. It's clear that Intel knows ultimately it too will have to resort to multi-chip packages. It's even been touting its Embedded Multi-Die Interconnect Bridge (EMIB) technology as a better way to do MCMs.

With the recent launches of Ryzen, EPYC, and Threadripper processors as well as "Vega" based graphics cards, AMD has demonstrated that its culture respects not only technical excellence, but true grit as well. Over the past five years, its engineers ignored the naysayers who doubted the company's capabilities and delivered a new generation of processors and graphics engines that will drive the company's growth for yet another decade.

For almost a decade, Intel has had almost no competition in the x86-based server market. AMD's reentry into that market is sure to get both companies' competitive juices flowing and accelerate innovation that will benefit users and suppliers alike. Let the games begin!

---

[1] Electronic Numerical Integrator and Computer

[2] Koomey, Jonathan; Berard, Stephen; Sanchez, Marla; Wong, Henry (29 March 2010). "Implications of Historical Trends in the Electrical Efficiency of Computing]". IEEE Annals of the History of Computing. 33 (3): 46–54.

[3] Koomey, J.; Naffziger, S. (31 March 2015). "Moore's Law Might Be Slowing Down, But Not Energy Efficiency". IEEE Spectrum.

[4] The number of die on a silicon wafer (known as "candidates") is determined by their dimensions. The number of defect-free die on a wafer (known as the "yield") is a function of the number and size of those die. A small defect on a large die causes that entire die to be scrapped and limits yield. Using smaller dies creates more candidates, and increases the fraction of candidates that emerge defect-free. Yield plays a key role in determining the profitability of most semiconductor products.

[5] "Architecture" defines the software-visible aspects of a machine, independent of how that machine is implemented. "x86" and "x86-64" specify those characteristics for most machines AMD and Intel build today. "Microarchitecture" defines the manner in which the machine processes data and instructions, but (aside from performance considerations) is largely invisible to software running on the machine. To preserve software compatibility, chip designers limit most of their innovation and design effort to the realm of microarchitecture.

[6] The physics of transistor operation are vastly more complex than we can address in a document like this. Suffice it to state that (1) transistor performance depends on the dimension of the gate that switches power on and off, and as dimensions got smaller, the only way to increase gate area was to flip it on its side. (2) The power consumed by a transistor varies with the square of the voltage it uses, and FinFET transistors operate at far lower voltages than their planar predecessors.

[7] http://www.eetimes.com/document.asp?doc_id=1331317&page_number=2

[8] Power varies with the voltage^2, so a 10 percent increase in voltage results in a 21 percent increase in power.

[9] I claim some academic expertise in this area. My 1967 MIT undergraduate thesis, titled "A History of Semiconductor Research," focused on this topic. There was a lot less history then.

[10] The name "Silicon Valley" didn't come until much later.