# Technology Impacts from the New Wave of Architectures for Media-rich Workloads

Samuel Naffziger
Advanced Micro Devices, Inc.
2950 E Harmony Rd
Fort Collins CO 80525, USA
samuel.naffziger@amd.com

## Abstract

As the growth in rich graphical and multi-media workloads begins to dominate the compute cycles of our next-generation processors, a revolution in architecture is taking place to efficiently deal with them. This revolution involves the synergistic combination of parallel and serial computation elements on-die. This co-location makes for a rapidly evolving set of technology challenges. With power limits front and center, the need for efficient, dense logic with high-bandwidth interconnect makes the computing industry more dependent than ever on continuing VLSI technology improvements. This talk will explore these trends and the implications for next-generation process development.

## Introduction

It is easy for most of us to remember a scant two decades ago when personal computing was still in its infancy. Laptops had many severe limitations: slow modem-based wired connections, simplistic graphics capabilities and displays, and very limited battery life. Improvements began in the 2000s as wireless capabilities and graphics technologies came to market while battery life slowly improved. Today, with HD displays, ubiquitous connectivity, and touch screen interfaces, we are entering the third decade of mobile computing when widespread connectivity, intuitive user interfaces and visual content combine to place a different sort of burden on processors.

## Drivers of the new computation demands

Some fundamentals of human nature are driving these changes in usage and capability in the next generation of processors. These amount to three basic factors:

1. The primacy of the visual experience
2. The importance of social networking
3. The value of mobility

Our ability to process language is limited to about 150 words per minute, whereas the human brain can process information from visual perception at between 400 and 2,000 times that rate [1]. The value of connecting with others visually can be summarized with some simple statistics: 24 hours of video uploaded to YouTube every minute; 50 million digital media files added to personal content each day; 1,000 images uploaded to Facebook every second [2]. Ten years ago, these capabilities weren't even imagined; today, they are ubiquitous. People need to connect with each other, and doing so visually is vastly more rewarding than doing so textually or verbally. This is driving a change to new computation workloads:

Massive data mining: There is a need to process, catalog, and analyze the vast amount of data being generated. Figure 1 illustrates the growth in digital information created and the impending gap in storage capability. This drives a need for processing power to perform analysis, recognition and compression functions to file this information, much of which is visual in nature.

Visualization: Science and engineering can be much more productive with real-time immersive physics-based rendering. These capabilities are the same ones that enable next-generation gaming. The thirst for real-time computation and image processing for these functions appears boundless.

Natural user interfaces: Keyboards are functional, but people naturally want to interface with computers using voice, gestures or even facial emotions. These capabilities require the computer to process information from the physical world in the form of face and object recognition. These complex associative tasks require vast amounts of parallel computation which have been the realm of supercomputers in the past.

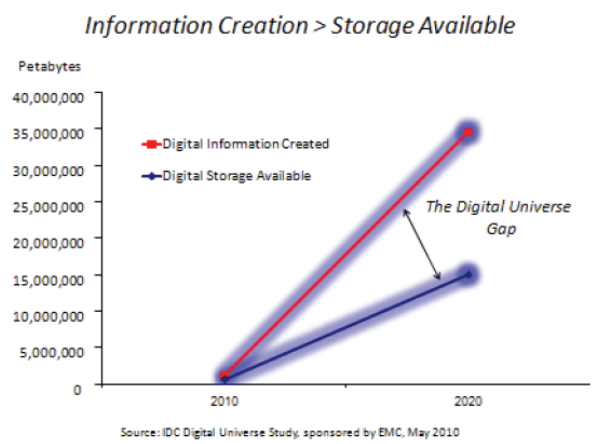



**Figure 1: The Expanding Digital Universe**

## Implications for processor architecture

There are two different forms of computation in our computers today: serial and parallel. Serial computation has historically been the province of the CPU, and the mainstay of programmers since computer science began. Parallel computation has typically been relegated to high-performance scientific computing and graphical

rendering. Two factors are driving a merger of these modes of computation:

1. The explosion of visual content and processing requirements described in the previous section
2. The plateauing of serial computation capability

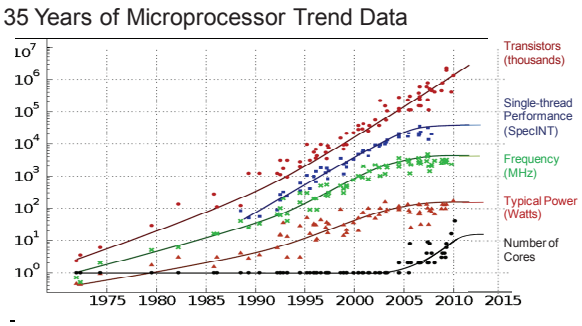



**Figure 2: Serial Microprocessor Trends**

As illustrated in Figure 2 [8], gains in frequency and single-thread performance have markedly leveled off in recent years due to fundamentals of physics and the nature of the serial compute model. Symmetric multiprocessing architecture designs are useful for some workloads, but their capabilities are eclipsed by a different trend, illustrated in Figure 3.

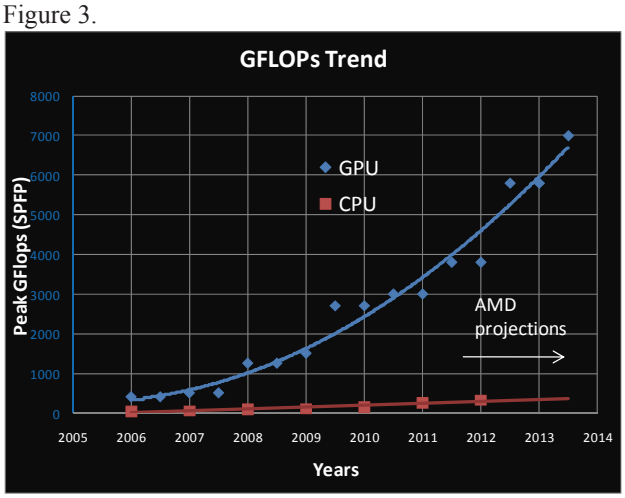



**Figure 3: Flops-per-second Trend for Single-chip High-end Designs**

This disparity in the growth of number-crunching capability is truly remarkable, and explained by a few simple factors:

- GPU (or vector) computation essentially removes the serial dependencies present through branches and conditionals in traditional computing. This change limits the flexibility of the coding model, but still enables application to a broad set of emerging workloads.
- The lack of serial loops in code means latency is not a factor in the design. This opens the door to utilizing slower, cooler devices that mitigate the power density issues that plague speed-demon serial processors.
- Since these designs are by definition parallel, they benefit linearly with density improvements, which is the one thing that Moore's Law [7] has continued to deliver in recent years even as the frequency improvements from silicon scaling have flattened.

The combination of trends toward insatiable demand for high bandwidth, visual processing, and the continued exponential growth in vector processing capability are the forces behind the merging of CPU and GPU designs on a single piece of silicon, as demonstrated in recent AMD and Intel designs [3,4,5]. In summary, as illustrated in Figure 4, having passed through the single core and SMP eras, we are entering a third era of computation due to these factors: that of the heterogeneous systems.

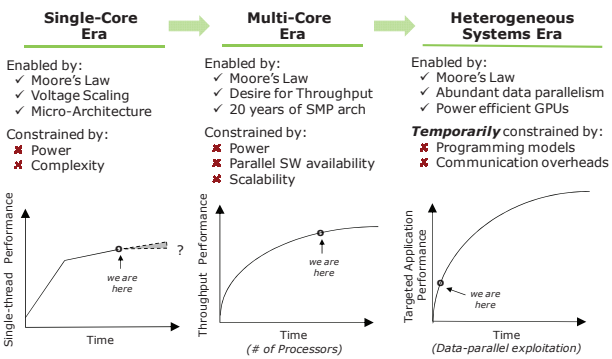



**Figure 4: Three Eras of Processor Performance**

## AMD Fusion™ architecture

The goal of the AMD Fusion architecture is to optimize the computation engine of the next generation of processors for power efficiency on both parallel and serial workloads. The need for high-speed serial computation will always be central to mainstream computing: good response time requires fast serial computation, and many workloads simply aren't amenable to parallel algorithms.
So an effective merged architecture must efficiently enable both high-performance x86 cores and leading-edge GPU technology co-located on the same piece of silicon. Tight coupling enables data sharing among all computation elements on the die. An initial installment of these capabilities from AMD is illustrated in Figure 5 with the accelerated processing unit (APU) code-named Llano.

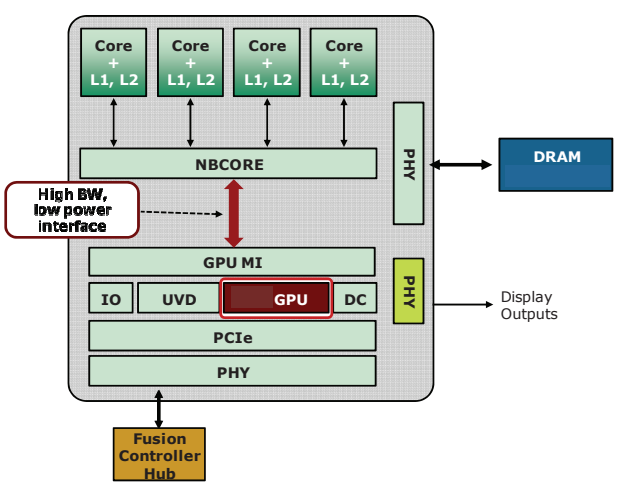
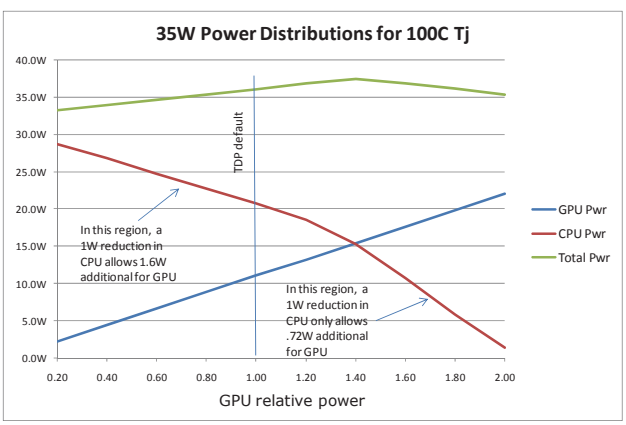



**Figure 5: AMD Fusion APU Architecture ("Llano")**

Key synergies come from reduced communication overhead through leveraging high-bandwidth on-die wires, sharing the memory interface, and the ability to efficiently use a single thermal solution for greater power efficiency, which introduces the next topic.

## Power management

Mobile computing demands extended battery life and a thin, light form factor. These are difficult goals to achieve while delivering the high-performance computation required for natural human interfaces, visualization, and data mining. An important step in this direction is achieved by fine-grained power sharing among the compute elements on the die. Most workloads emphasize either the serial CPU or the GPU and do not heavily utilize both simultaneously. By dynamically monitoring the power consumption in each CPU and the GPU [2, 6], and tracking the thermal characteristics of the die, watts that go unused by one compute element can be utilized by others. This transfer of power, however, is a complex function of locality on the die and the thermal characteristics of the cooling solution. As illustrated in Figure 6, the efficiency of sharing is a function of where the hot spot is and will vary across the spectrum of power levels.



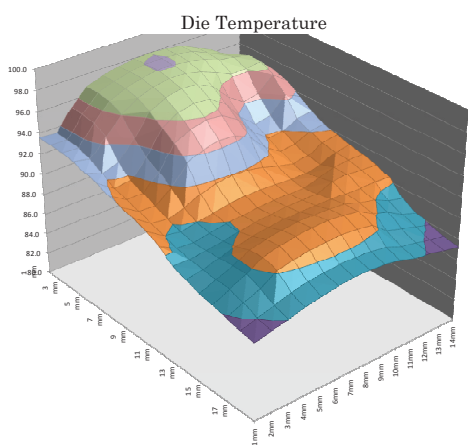
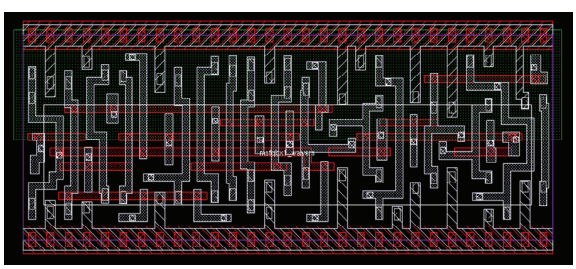




**Figure 6: CPU <--> GPU Power Sharing**

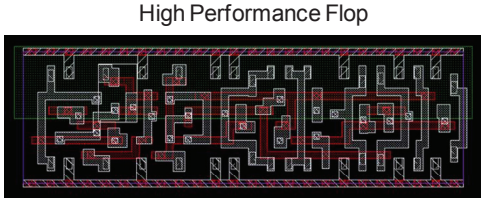
While the CPU is the hot spot on the die, a 1W reduction in CPU power allows the GPU to consume an additional 1.6W before the lateral heat conduction from CPU to GPU heats the CPU enough to be the hot spot again. As the GPU consumes more power, it finally becomes the hot spot on the die, and the reverse situation occurs. A power-management system that maintains repeatable performance must have sophisticated power-tracking capability and thermal modeling to ensure maximum compute capability is extracted from a given thermal solution. Once that is in place, the APU can deliver far more computation within a thermal envelope than either design in isolation.

## Technology implications

With an efficient architecture in place, and optimal power management, the remaining item to optimize is the technology. As already discussed, the GPU requires layout density and lower-power devices to sustain the growth in required compute capability. To facilitate this, separate standard cell libraries are required on an APU, as illustrated in Figure 7, in which the high-performance CPU flop design consumes almost 50% more area than the high-density (but slower) GPU flop design. The reduced power rails in the high-density library facilitate routeability while taking advantage of lower power densities in the parallel compute units. This trade-off makes sense when the flop count differences are assessed: all four CPU cores on Llano instantiate 660,000 flops, while the Llano GPU uses almost 3.5 million of them. Density is essential.



High Performance Flop



High Density Flop

**Figure 7: High-density vs. High-performance Library Comparison**

A density-optimized library benefits from tight metal pitches, while a performance-optimized design wants thicker, lower-RC wires. These conflicting demands place extra stress on the choice of a metal stack that best balances the needs of both. An important point here as regards interconnect issues is that for both serial and parallel computation, quality interconnect is imperative. As shown in Figure 8, we are at the cusp of a dramatic increase in wire RC delays due to a number of factors such as small geometries resulting in increased edge effect scattering, the lack of dielectric improvement and the overhead from copper cladding layers. This trend will seriously reduce

both the density and performance benefits of process scaling, so some sort of revolutionary improvement in interconnect is required.

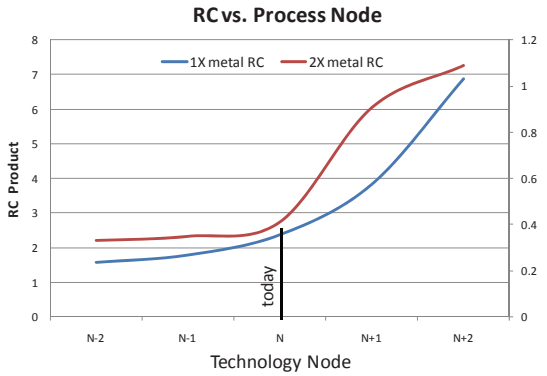



**Figure 8: RC Trends**

Another mis-match in the technology needs of these architectures can be seen in Figure 9: CPU and GPU Vt Mix Differences 9, which plots the percentage of transistor width for each Vt choice in the two designs. The GPU makes heavy use of the least leaky and slowest long-channel HVT (LC-HVT) device, while the same device goes virtually unused in the CPU because speed loss is unacceptable. The CPU makes heavy use of the RVT device for speed, but the GPU cannot afford the leakage. Future implementations will need to engage in careful analysis of optimum Vt offerings for each design style and work with the fab to strike the right balance.

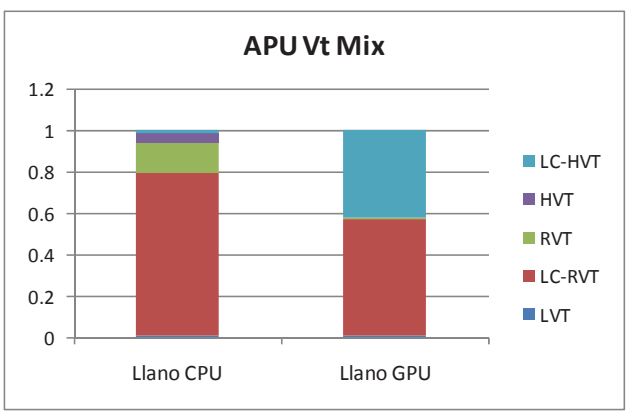



**Figure 9: CPU and GPU Vt Mix Differences**

The next technology challenge driven by an emphasis on parallel computation is shown in Figure 10. If we hold power density constant to avoid increasing the cost of the cooling solution, but take advantage of the density improvements of the process technology generations, the operating voltage necessarily decreases. This puts extreme pressure on device variation control and transistor drive characteristics. Low leakage tends to demand high Vts that enable reduced power density, but this is at odds with the demand for reasonable performance at very low voltages.

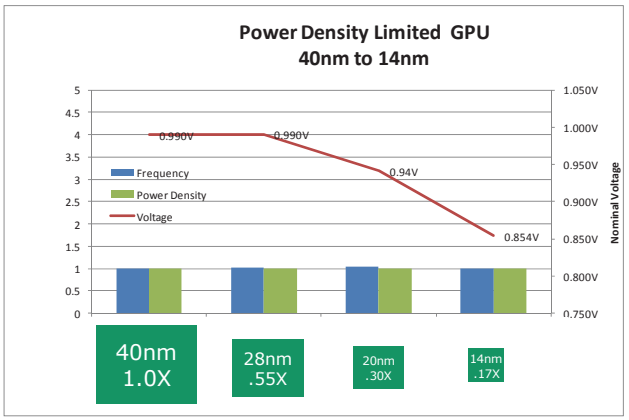



**Figure 10: Voltage Impact of Power Density-limited Designs as Process Scales**

The impact of variation effects on low-voltage performance can be seen in Figure 11, which plots the maximum operating frequency of 40-nm GPUs representing the full range of process characteristics against voltage. The absolute spread in MHz of frequencies increases as the voltage (and median frequency) decreases. When this spread is taken as a fraction of the median frequency, the increase in frequency variation is a dramatic 300% when voltage is reduced from 1.15V to 0.85V.

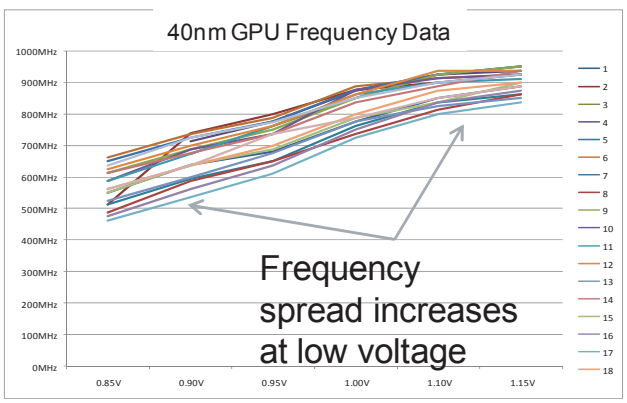



**Figure 11: GPU operating Frequency Spreads vs. Voltage**

This represents a significant technology and circuit design challenge to sustaining the density scaling benefits of Moore's Law for accelerated processing unit (APU) designs.

## Future Trends

In summary, as we have implemented the APU as an answer to the emerging workloads discussed at the start of this paper, we have gained insight into the technology needs that result from this architecture. In the meantime, market trends have continued to evolve and a couple of these bear mentioning. As illustrated in Figure 12 [9] there is a literal explosion in the number of mobile devices that both generate information (pictures, video, audio, etc.) and query for information (web access, location, augmented reality). The demand for compute capability in these devices is driven by the processing opportunities afforded by the large volumes of multi-media data including the possibility of using them to create the natural interface capabilities

mentioned earlier. In addition to computation within these devices, they are also driving a surge in cloud server computation to perform all the search functions, processing, and indexing necessary to serve the requests generated. The combined serial + parallel computation capability of the APU matches the computation requirements in the cloud, as well as the power efficient, visually oriented needs of the mobile devices. New challenges in power efficiency and density, but this architecture is well poised for the emerging trends.

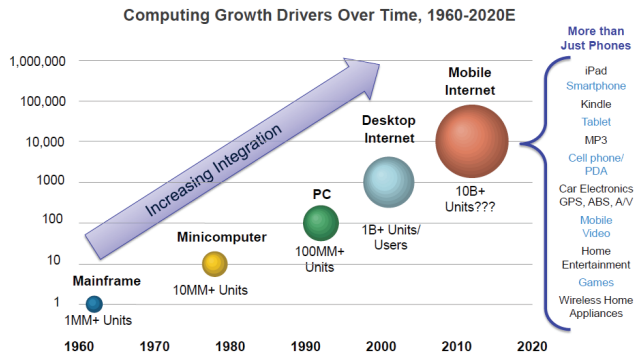



**Figure 12: Massive Growth in Cloud Clients**

## Conclusion

The demand for visually stimulating, interactive user experiences along with the explosion in data- and query-generating mobile devices has driven and continues to drive the development of a new wave of computer architectures. These leverage hybrid serial+parallel compute processing engines that are tightly coupled on the die and share memory, thermal, and electrical resources. These accelerated processing units, or APUs, will be able to ride the wave of continued density gains afforded by Moore's Law even if device speeds are no longer improving. However, this direction adds significant stress to silicon process development in several areas. First, the back end of line must strike a careful balance between density and speed. Next, the transistor suite must span a broader range of Ion/Ioff points than that of CPU-only designs. Finally, the power requirements will force a relentless reduction in operating voltage, which will demand much better variation control.

## References


[1] Encyclopedia of Psychology, "Visual Thinker", 2004

[2]http://socialmediatoday.com/soravjain/237864/fascinating-social-media-facts-year-2010

[3] D. Foley et al, "A Low-Power integrated x86-64 and Graphics Processor for Mobile Computing Devices", IEEE International Solid State Circuits Conference, Feb 2011

[4] R. Jotwani et al, "An x86-64 Core Implemented in 32-nm SOI CMOS", IEEE International Solid State Circuits Conference, Feb 2010

[5] M. Yuffe et al, "a Fully integrated Multi-cpu, Gpu and Memory controller 32nm processor", ", IEEE International Solid State Circuits Conference, Feb 2011

[6]C. Demerjian, "PowerTune" section , http://semiaccurate.com/2010/12/14/look-amds-new-cayman6900-architecture/

[7] Gordon Moore, "Cramming more components onto integrated circuits," Electronics, Volume 38, Number 8, April 19, 1965

[8] Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten

[9] ITU, Mark Lipacis, Morgan Stanley Research