AMD Fusion™ Family of APUs: Enabling a Superior, Immersive PC Experience

Table of Contents

ABSTRACT:

The AMD Fusion™ Family of Accelerated Processing Units (APUs) is coming. These processors' compatibility with Microsoft® Windows® 7 and DirectX® 11 is designed to ensure that they will provide an outstanding out-of-the-box experience for those who use PCs built on them. Working in conjunction with the advanced x86 processor cores, an APU's multiple vector cores enable software developers to create innovative new applications that expand usage models and make PCs fast, easy to use, and more intuitive. The recent introduction of new tools (DirectCompute and OpenCL™) for thread-level and data-parallel applications development simplifies the task of creating these applications. The development platforms and tools are scheduled to be widely available. The rest is up to entrepreneurial developers and PC manufacturers.

Insight 64

Nathan Brookwood, Research Fellow, March 2010

Insight 64 thanks AMD for its financial and technical assistance in the creation of this white paper.

AMD
The future is fusion

INTRODUCTION

Imagine a PC that:

- Recognizes your face to login
- Recognizes your gestures without a remote
- Responds to your touch or voice to do your bidding
- Supports bi-directional hi-definition video chat over links with limited bandwidth
- Finds and tags the photos and videos in your library that contain particular faces, places or objects
- Helps you sort through your photo libraries to eliminate duplicates saved with different file names
- Enhances the videos you've created with regard to color, focus and image stability
- Takes you to the right page in IMDB or Wikipedia when you point to an actor in a movie or a figure in a news program
- Up-scales even low-quality content to seamlessly match the capabilities of your HD display
- Adds stereoscopic 3D realism to 2D content
- Supports immersive, multi-monitor 3D gaming experiences
- Sells at price points well within reach of the mainstream consumer.

Many of these capabilities exist today piecemeal in labs, running on expensive, workstation-class computers that cost as much as tens of thousands of dollars. Why haven't we progressed further, faster in delivering these capabilities to the mainstream? The semiconductor industry prides itself on rapid improvements in system performance, but hardware that runs fast enough to enable these advanced capabilities still costs far too much to enable high-volume deployment. Software developers, always tuned to market realities as well as technology, have focused their efforts on applications that run well on the dual- and quad-core x86 processors that comprise the bulk of today's mainstream system offerings. But change is in the air; in 2011, affordable mainstream systems that can support these advanced capabilities are set to enter the market.

You've probably heard this story before. Every two years, advances in semiconductor technology allow chip architects to double the number of transistors they can fit in a given area of silicon. Over the past decade, these extra transistors have been used to increase the size of on-chip caches and add more x86 processor cores to designs, making today's CPUs the fastest processors ever. Even the slowest contemporary CPUs have more than enough performance to handle traditional office productivity, Internet browsing and e-mail applications, which long ago ceased to be limited by CPU speed. But as fast as they are, today's CPUs lack the performance to deliver a vivid, modern computing experience on their own. The latest applications require CPUs that can deal with vast amounts of data and require hundreds, if not thousands of individual threads to manipulate the massive databases needed to recognize an object in a scene, the meaning in a sentence, or an anomaly in an x-ray image. Not surprisingly, traditional CPU architectures and application programming tools optimized for scalar data structures and serial algorithms fit poorly with these new vector-oriented, multi-threaded data-parallel models.

Fortunately, innovative architectures and tools better suited for these new workloads have emerged. Graphics processing units (GPUs), originally intended to enhance 3D visualization, have evolved into powerful, programmable vector processors that can accelerate a wide variety of software applications. Software tools like DirectCompute and OpenCL permit developers to create standards-based applications that combine the power of CPU cores and programmable GPU cores, and run on a wide variety of hardware platforms. A few ambitious independent software vendors (ISVs) have already added support for these new vector capabilities into their most advanced products, even if they had to structure their code around proprietary hardware and software interfaces to get the job done.

Advanced Micro Devices' (AMD's) forthcoming Accelerated Processing Units (APUs) build upon this momentum and take PC computing to the next level. These new processors are being designed to accelerate multimedia and vector processing applications, enhance the end-user's PC experience, reduce power consumption, and offer a superior visual graphics experience at mainstream system price points. Insight 64 would expect no less. More importantly, these APUs will

enable ISVs to create new generations of applications and user interfaces limited perhaps only by the inventiveness of their developers, rather than by the constraints of the traditional CPU architectures that have dominated the computer industry for decades.

In this white paper, Nathan Brookwood of Insight 64 explores the world of these new APUs for AMD. How do they differ from today's CPUs and GPUs? Which applications benefit the most from this technology? Can mere mortals harness their power? Even more importantly, what steps should PC manufacturers and ISVs take today to ensure they will be ready for the new wave of accelerated processing units when it begins to crest, currently expected sometime in 2011?

WHAT IS AN ACCELERATED PROCESSING UNIT?

At the most basic level, AMD's new Accelerated Processing Units combine general-purpose x86 CPU cores with programmable vector processing engines on a single silicon die. AMD's APUs also include a variety of critical system elements, including memory controllers, I/O controllers, specialized video decoders, display outputs, and bus interfaces, but real appeal of these chips stems from the inclusion of both scalar and vector hardware as full-fledged processing elements. Others have lashed a CPU and a basic graphics unit together in a single package, but none have attempted this feat with truly programmable GPUs like those in the AMD Fusion designs, let alone GPUs that can be programmed using high-level industry-standard tools like DirectCompute and OpenCL. AMD is best situated to address this engineering challenge, as it is currently the only company which has access to extensive IP resources (e.g. patents and engineering expertise) in both x86 processor technology and industry-leading GPU technology. In fact, AMD's recognition that it needed proven GPU technology for future converged products drove its 2006 acquisition of ATI Technologies.

AMD's APUs are set to arrive in a variety of shapes and sizes adapted to the requirements of their target markets. AMD has disclosed that its first APUs, code-named "Llano" and "Ontario," are designed for mainstream desktop and notebook platforms and "thin and light" notebooks, and netbooks and slates. Both of these APUs will combine multiple superscalar x86 processor cores with an array of programmable SIMD engines leveraged from AMD's discrete graphics portfolio.

Figure 1. To enable the immersive PC usage models that users demand requires both CPU and GPU collaborative computing

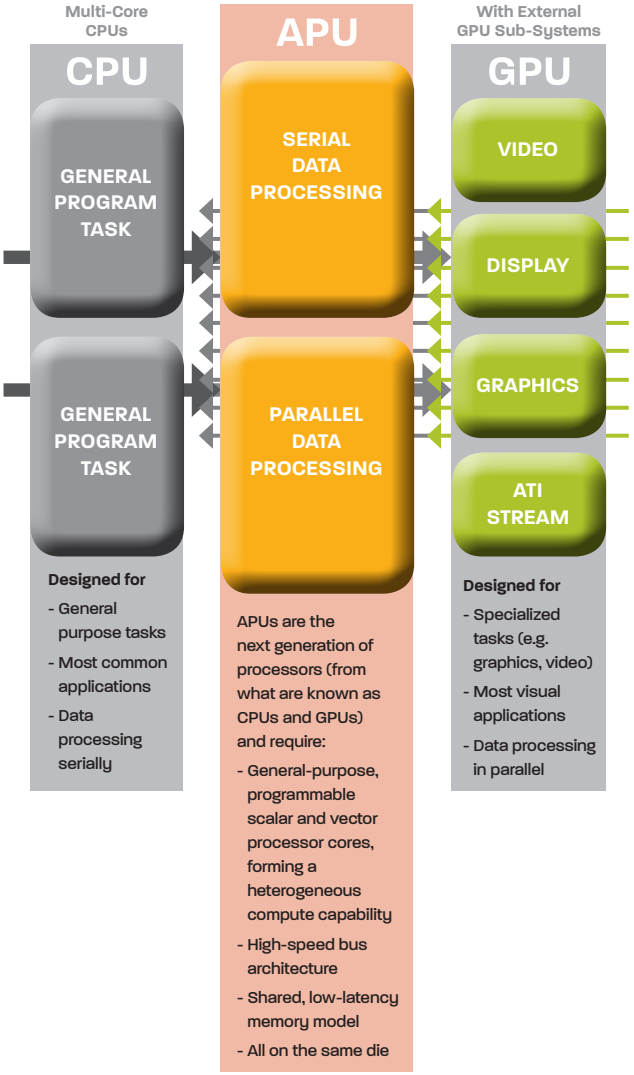
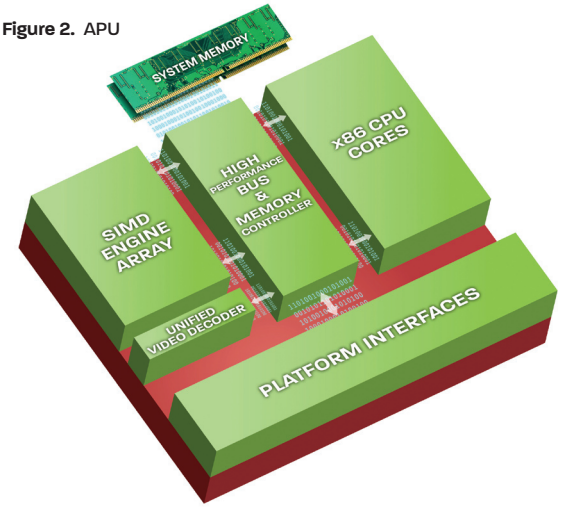


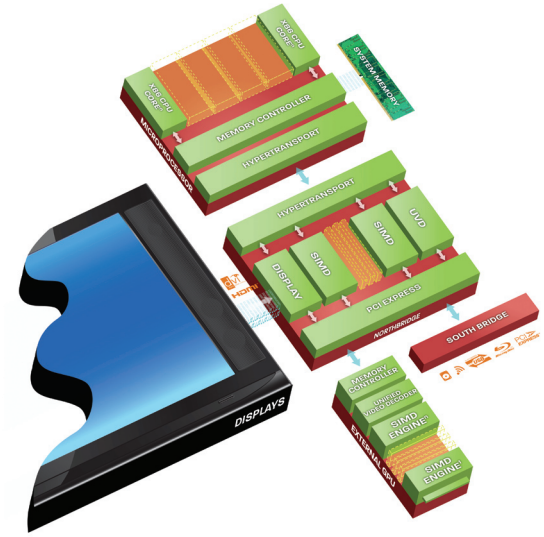
Figure 2 illustrates the arrangement of these first APUs. The key aspect to note is that all the major system elements – x86 cores, vector (SIMD) engines, and a Unified Video Decoder (UVD) for HD decoding tasks – attach directly to the same high speed bus, and thus to the main system memory. This design concept eliminates one of the fundamental constraints that limits the performance of traditional integrated graphics controllers (IGPs).

Figure 2. APU



Until now, transistor budget constraints typically mandated a two chip solution for such systems, forcing system architects to use a chip-to-chip crossing between the memory controller and either the CPU or GPU as shown in Figure 3. These transfers affect memory latency, consume system power and thus impact battery life. The APU's scalar x86 cores and SIMD engines share a common path to system memory to help avoid these constraints.

Figure 3. Typical current AMD system configuration



Total system performance can be further enhanced through the addition of a discrete GPU. The common architectures of the APU and GPU allow for a multi-GPU configuration where the system can scale to harness all available resources for exceptional graphics and enable truly breathtaking overall performance.

Although the APU's scalar x86 cores and SIMD engines share a common path to system memory, AMD's first generation implementations divide that memory into regions managed by the operating system running on the x86 cores and other regions managed by software running on the SIMD engines. AMD provides high speed block transfer engines that move data between the x86 and SIMD memory partitions. Unlike transfers between an external frame buffer and system memory, these transfers never hit the system's external bus. Clever software developers can overlap the loading and unloading of blocks in the SIMD memory with execution involving data in other blocks. Insight 64 anticipates that future APU architectures will evolve towards a more seamless memory management model that allows even higher levels of balanced performance scaling.

Just as AMD's architects have woven x86 cores and GPU cores into a single hardware fabric, astute software developers can now begin to weave high performance vector algorithms into programs previously constrained by the limited computational capabilities of conventional scalar processors, even when arranged in multi-core configurations. In just a few years, machines equipped with programmable GPUs are expected to comprise a meaningful portion of the installed base of PCs. Software coming from ISVs who take advantage of these enhanced capabilities will have the ability to execute well beyond the capability of packages that lack support for these features.

WHY ALL THE FUSS ABOUT "VECTOR PROCESSING"?

For over four decades, architects have pursued two competing concepts to enhance system computational performance. One group argues that designs that emphasize a single machine instruction operating on a single data item ("SISD") work with the broadest array of application and lead to the most cost-effective solutions. The other group counters that vector designs that emphasize single instructions operating on multiple data items ("SIMD") deliver relatively higher performance for computationally intensive applications that need to process large arrays of highly structured data, a model that applies to many modern computing tasks.

After all those years of debate, it has now become clear that both sides were correct; it is not a question of "either/or," but rather of "both/and." Some numerically intensive problems lend themselves to parallel algorithms, and others don't. When a machine optimized for parallel computation encounters a problem that cannot be computed in a parallel manner, the machine operates as an inefficient scalar processor, and most of its parallel computing resources sit idle. Conversely, a processor optimized for scalar calculations cannot exploit the parallelism in many algorithms, and thus is limited by its scalar processing speed.

Affordable PCs that offer hundreds of gigaFLOPS performance are not equipped to cause spreadsheets to recalculate faster or e-mail to get to its destination sooner, but they can impact the way users interact with their system to set up those spreadsheets or compose that mail. A new era of Natural User Interfaces (NUIs) can help users to communicate with their system using visual and verbal inputs instead of mice and keyboards. Step-function increases in system capabilities and usability have a way of shaking up the status quo in existing markets, benefiting those who recognize and react to the impending changes, and negatively impacting those who fail to recognize the shift, or react too slowly to it.

THE ABCs OF VECTOR PROCESSING

The technology that allows vector processors to achieve very high levels of computational performance is easy to understand, but difficult to implement. This partially explains why so few companies have fielded successful products based on this approach. Scalar processors operate on arrays of data one element at a time. When a program needs to add one thousand elements in vector A to a separate list of one thousand elements in vector B, and store the results in vector C, it typically sets up pointers to each vector, loads the values pointed to by A and B into CPU registers, adds those registers, and stores the result into the location pointed to by C. Then the program updates all its pointers and repeats the process one thousand times. The actual time the CPU spends handling the one thousand "add" operations pales in comparison to the time it spends handling the looping operations – updating pointers and iteration counts.

Vector processors like those used in advanced GPUs have dozens, and sometimes hundreds of calculating units that operate simultaneously. When an application wants to add two one-thousand element vectors using ten of the system's available processing units, the vector software restructures the work so that each calculation executes simultaneously on ten separate elements, and thus completes the work in as little as one-tenth the time. Sounds easy, doesn't it? Of course, one must account for the time needed to set the operations up, the time needed to ensure they all complete successfully, and the time needed to move data between the system's scalar and vector memory regions. One must also ensure that the operations applied to any element in each vector can be performed independently of operations applied to other elements in that same vector. It is easy to see that implementing these features in hardware might be more complicated than one might initially think.

Given the restrictions and overheads enumerated above, along with a few others, it should come as no surprise that vector processing techniques cannot boost the performance of all programs equally. For small data arrays, the overhead associated with setting up vector operations can outweigh the time saved through parallel execution. Many problems and algorithms have proven a poor fit for this technology, and are best handled using scalar approaches. Some early vector-oriented architectures excelled on vector workloads but failed in the market because their designers neglected these scalar workloads.

AMD's Accelerated Processing Unit designs are constructed to help avoid these pitfalls. These APUs cut through scalar workloads using AMD's proven x86 core technology and through vector workloads using enhanced versions of its GPU technology. Although AMD had to overcome many technical challenges to merge its vector and scalar technologies in a manner that preserves the advantages of each, having the core IP for both processing elements provides this AMD hardware with a significant advantage over other hardware designs which are missing one or the other.

HOW HARD IS IT TO ACCELERATE SOFTWARE USING GPU COMPUTING RESOURCES?

The recent emergence of two important development tools – OpenCL and DirectCompute – enables developers (especially those who have already mastered the art of writing software for single-threaded scalar environments) to more easily create highly dynamic multi-threaded data-parallel software applications.

Over the past three years, the tools AMD provides to developers who want to use GPU hardware to accelerate their applications have undergone a dramatic evolution. Prior to November 2006, ISVs were limited to applying ATI Radeon™ graphics cards or ATI FirePro™ graphics accelerators to anything other than traditional 3D applications. Then AMD launched the "Close To Metal" (CTM) initiative that gave early GPU-computing pioneers a set of low-level, proprietary interfaces they could use to craft GPU-accelerated applications like Folding@Home, a distributed computing project run by Stanford University that puts the idle time normally wasted by screen savers in millions of PCs to good use in the study of protein folding, aggregation, and related diseases.

Roughly a year later, AMD augmented its GPU software toolkit with the release of the ATI Stream SDK and Brook+, an open-source, high-level C-like language that simplified the use of AMD GPUs for computational tasks. This brought a few more developers on board the GPU-computing bandwagon, but most ISVs held out for industry-standard based tools that would allow them to address a broader range of hardware configurations. That wait officially ended last fall, with the release of OpenCL, a cross-platform standard for parallel computing coordinated by the Khronos Group; and DirectCompute, a new set of Windows DirectX APIs that facilitates GPU computing applications. Thus ISVs have moved in short order from having no industry-standard options for GPU computing to having two.

ISVs who focus on Windows and use DirectX APIs for graphics will most likely gravitate toward DirectCompute. It features data structures compatible with DirectX 10 and DirectX 11 application programming interfaces (APIs), and thus helps simplify the process of adding GPU acceleration for physics or Artificial Intelligence tasks to DirectX 11 applications. Microsoft has further encouraged this by releasing DirectX 11 and DirectCompute APIs as "platform extensions" to Windows Vista, which allows software that uses these APIs to target a huge installed base of systems, along with all the new Windows 7 systems being sold now.

ISVs wishing to address a broader market may find OpenCL the better development platform, especially if they already use OpenGL APIs to handle graphics interfaces in existing applications. The Khronos Group, an industry consortium that drives a variety of open API standards, serves as the development coordinator for both OpenCL and OpenGL languages. OpenCL includes support for both data-parallel (i.e., SIMD) and task-parallel execution models. It uses data structures compatible with OpenGL APIs, thus helping to simplify the task of adding GPU compute acceleration to OpenGL applications. Key processor suppliers, including AMD, ARM, Intel and IBM, and key GPU suppliers, including AMD and Nvidia, support OpenCL and have released OpenCL drivers that work in a number of OS environments for many of their devices. AMD's OpenCL compiler supports both its ATI Radeon™ HD 4000 and HD 5000 series GPUs and its multi-core x86 processor offerings.

ISVs who want to get a head start on developing software for AMD's forthcoming APU offerings can begin their work today on any AMD platform that includes an ATI Stream enabled ATI Radeon HD 4000 or HD 5000 series discrete GPUs. The new APUs will appear just like today's DirectCompute and OpenCL platforms from the viewpoint of the software that runs on the platform. This means investments ISVs make today to support current platforms will remain useful on future platforms addressing a much wider breadth of the PC marketplace.

DOES THIS STUFF REALLY WORK?

The tools needed to accelerate applications via GPU computing have only been around for a few years, but already a few innovative ISVs have used this technology to enhance their applications. A few of the more interesting examples include:

- Adobe's ubiquitous Flash Player now uses GPU hardware to decode video streams. This innovation helps improve the quality of the video playback on enabled GPUs, reducing the processing load on the CPU, and thus uses less power, extending system battery life. The release candidate of Flash Player 10.1 is available for download as of this publication date.
- ArcSoft has added a GPU-enabled SimHD™ plug-in to its TotalMedia Theatre package. The new plug-in enhances video quality by intelligently up-scaling standard DVD video from 480 vertical lines to 720.
- Cyberlink has enhanced its line of media software applications to use GPU acceleration whenever possible. Its Power Director 8 package takes advantage of enabled GPU resources to speed up video editing, video encoding and video effects rendering. Its MediaShow line uses enabled GPU hardware to accelerate video format conversion (transcoding) and encoding, as well as to implement an automated "face tagging" feature that sorts the user's photo collection based on the faces in the photos. Its PowerDVD offering takes advantage of GPU resources to enhance Blu-Ray playback; the company demonstrated a future version of PowerDVD for Blu-Ray 3D playback at the 2010 CES show. Cyberlink started its GPU-acceleration efforts in 2008, and used the proprietary tools then available from AMD and Nvidia to develop its software. Now it is converting its software to use DirectCompute in order to increase the range of supported platforms and get to market with new features more quickly.
- One Silicon Valley startup uses GPU resources to clean up video files, compensating for noise, pixilation, graininess, poor focus, low contrast, and shaky images due to shaking cameras. The package works just like the fictional ones you might see in a film where the hero zooms in on a satellite image and reads the villain's license plate, but this package relies on GPU hardware, rather than Hollywood gimmicks.
- Another startup has demonstrated facial recognition software that finds individual faces in photos or videos and matches them to faces in its database. This obviously requires a tremendous amount of computational horsepower, but with GPU assistance, it can accomplish this task virtually in real time. It's not hard to imagine that GPU computing could be employed by civil protection organizations to help make the world a safer place.

WHAT CAN APUs DO FOR TRADITIONAL WORKLOADS?

Although it's exciting to look at the new applications that will finally become practical in the "Fusion" era, the fact remains that most users will want their new APU-based systems to handle a mix of traditional applications for office productivity and Internet access, along with those new exciting apps. Fortunately, the changes AMD made to enable new APU-accelerated applications can also help existing applications run better as well.

Many of these improvements stem from AMD's ability to fit the CPU cores, GPU cores and North Bridge (the part of the chip where the memory controller and PCI-express interfaces reside) onto a single piece of silicon. As noted earlier, this eliminates a chip-to-chip linkage that adds latency to memory operations and consumes power. It takes less energy to move electrons across a chip than to move those same electrons between two chips, and the power saved by this small change alone can help significantly increase system battery life. The co-location of all key elements on one chip also allows AMD to take a holistic approach to power management on these APUs. They can power various parts of the chip up and down depending on workloads, squeezing out a few milliwatts here and another few milliwatts there – which in the aggregate can amount to significant power savings.

Finally, some of the improvements can be attributed to the advanced GPU technology AMD embeds in its APU offerings. Although the company has yet to reveal the technical specs of these GPUs, it has disclosed they will be DirectX 11-compliant. These will be the first APU-based systems that can support DirectX 11's enhanced visual experience without a discrete GPU, and thus will represent a cost-effective solution for systems developers.

SUMMARY: WHAT CAN APUs DO FOR PC USERS?

The AMD Fusion family of Accelerated Processing Units is scheduled to arrive in 2011. The expectation is that their compatibility with Windows 7 and DirectX 11 will ensure that they will provide an outstanding experience for those who purchase PCs based on these processors. Their enhanced processing power and power efficiency will enable sharp and clear videos, realistic and responsive games, and notebooks that can run longer between battery charges.

More importantly, compared to today's mainstream offerings, APU-based platforms will possess prodigious amounts of computational horsepower. This processing power will allow developers to tackle problems that lie beyond the capabilities of today's mainstream systems, and will enable innovative developers to step up and update existing applications or invent new ones that take advantage of GPU acceleration. These features will be a standard part of every APU. Over time, even the most affordable PCs can be expected to have the computational performance of yesterday's million dollar mainframes with "all day" battery life.

Of course, few users will want to run the same applications on tomorrow's notebooks that they ran on yesterday's mainframes and supercomputers. They will likely want to run applications that help them in their everyday lives, doing tasks they cannot accomplish on the systems they own today. They may want to use facial recognition software to sort their photos and videos, or even to help them identify people they meet on the street or actors they see in movies. They may want the on-screen appearance of the videos they stream to approach that of the HD content on their TVs, even when bandwidth constrains that content to a low resolution format.

Since the days of the earliest personal computers, each major advance in system capability has enabled innovative software developers to create new products that opened new markets. The Apple II gave us VisiCalc, the first spreadsheet. The original IBM PC led to Lotus 1-2-3, the first spreadsheet with graphics. The Macintosh ushered in an era of desktop publishing that has forever changed the way the world creates and distributes information.

The dramatic increase in performance enabled by AMD Fusion technology can create new opportunities for entrepreneurial developers to innovate and make the world a better and richer place. Along the way, they may enrich themselves as well. That's the way the system is supposed to work.

For the hardware developer, ODM or PC manufacturer, it's time to start thinking about how to incorporate these new APUs into product lines in order to enhance the consumer experience. Software developers should look to this new power to help their software run even better. All developers are encouraged to upgrade their skills and learn about OpenCL and DirectCompute, and to examine current software projects to see how they can be improved in a world where systems have dramatically more power. Because pretty soon, they will.

48423B

AMD
The future is fusion