
The future is fusion.

The Industry-Changing Impact of Accelerated Computing.

Computer processing has reached a crossroads where the relationship between hardware and software must change to support the increased processing needs of modern computing workloads, including virtualized environments and media-rich applications. Accelerated computing uses specialized hardware to increase the speed of certain processing tasks, offering commercial and consumer users simultaneous energy efficiency, high performance, and low cost. From a hardware perspective, accelerated computing provides a general platform that supports the addition of specialized processing units, or accelerators, to a multicore computer. These accelerators may take the form of off-core chips such as media or network packet accelerators, or they may be additional cores designed for specific types of processing tasks. From a software perspective, the accelerated computing framework will simplify writing parallel applications: developers will have high-level tools that support parallel programming, and applications will not have to specify which hardware should be used for a given processing task.

AMD has been building expertise and capability in general- and specific-purpose computing for the last several years, including how to effectively couple general- and specific-purpose processors. This activity has built a solid foundation for further research and development in accelerated computing. We are actively collaborating with large research institutions, independent software vendors, the open source community, and other organizations to develop an open ecosystem around accelerated computing. Our commitment to accelerated computing includes designing hardware accelerators, building developer tools, and driving thought leadership by participating in standards development bodies. We believe that accelerated computing embodies a rapidly approaching paradigm shift, and we intend to offer commercial and consumer users flexible computing solutions that meet their needs based on the components available in the growing ecosystem.

This paper explores the accelerated computing framework. It provides examples of accelerated computing in action today, describes potential uses of accelerators in the near- to mid-term, and makes a clear case that accelerated computing is the natural outcome of the computer industry's current integration trend. It further illustrates that, in full bloom, accelerated computing will create many opportunities to customize computer systems by providing targeted, precise solutions for different kinds of workloads to satisfy the need for application performance, energy efficiency, and low cost.

CONTENTS

A NEW INFLECTION POINT	3
ACCELERATED COMPUTING: THE COMING PARADIGM SHIFT	3
MARKET TRENDS: DRIVING THE SHIFT TO ACCELERATED COMPUTING.....	3
AMD: AT THE FOREFRONT OF ACCELERATED COMPUTING.....	5
X86 PROCESSOR GENERATIONS AND APPLICATION PERFORMANCE	5
GENERATION 1: A FOCUS ON FREQUENCY AND ARCHITECTURE	5
GENERATION 2: HOMOGENEOUS MULTICORES.....	6
GENERATION 3: HETEROGENEOUS MULTICORES	7
APPROACHES TO ACCELERATED COMPUTING.....	7
ACCELERATED COMPUTING: LIGHTS, CAMERA, ACTION!.....	8
NOW PLAYING ON A COMPUTER NEAR YOU	8
COMING SOON TO A COMPUTER NEAR YOU	9
SIDEBAR: THE LOOMING CRISIS	9
THE CASE FOR ACCELERATED COMPUTING NOW.....	10

A NEW INFLECTION POINT

The computing industry is poised on the edge of an inflection point where the interaction of hardware and software compute capabilities will change radically. This inflection point is being driven by the confluence of the following factors:

- Evolving commercial and consumer workloads (including virtualization and increasingly media-rich applications) that demand simultaneous energy efficiency, high performance, and low cost.
- Newer applications with parallel algorithms whose processing can be done across multiple cores, resulting in better performance.
- A scarcity of programmers with the knowledge and skills needed to write parallel applications, and a shortage of tools to help programmers of all ability levels develop parallel applications.

Commercial and consumer demand for increased computing performance has consistently driven technological innovation in the computing industry, creating advances in processor power, core microarchitecture, and multicore designs. However, the emergence of graphically intense applications, virtualized environments, and parallel hardware architectures, combined with a lack of skilled programmers and development tool sets, has provoked intense scrutiny of how hardware and software should interact to provide satisfying computing experiences in this new era.

Major computing industry players, including vendors, standards bodies, and educational institutions, are exploring ways to improve workload-specific processing performance for commercial and consumer computing applications while maintaining a focus on energy efficiency and cost. And they are grappling with the challenge of unlocking the performance improvements possible in parallel architectures and algorithms without provoking a massive rewrite of current applications. All of this activity has created the current inflection point, and on the other side of that point is a paradigm shift that may fundamentally redefine how processing power translates into application performance.

ACCELERATED COMPUTING: THE COMING PARADIGM SHIFT

Accelerated computing is a framework that improves how hardware and software interact to support emerging application areas and evolving workloads while providing energy efficiency, great performance, and low cost. In the future, with accelerated computing:

- A computer will include general and specialized processors.
- A processor will be chosen to handle a given task based on its ability to provide optimal performance in terms of compute power, media processing, and presentation, as well as energy efficiency and cost.
- Software will take advantage of the best hardware for a given processing task without specifying how or where the processing is actually done.
- Developers will have high-level tools for writing parallel applications.

From a hardware perspective, accelerated computing provides a framework that supports the addition of specialized processors, or accelerators, to a multicore computer. These accelerators may take the form of off-core chips, such as media or network packet accelerators, or they may be additional cores designed for specific types of processing tasks. From a software perspective, the accelerated computing framework will simplify writing parallel applications. Developers will have high-level tools that support parallel programming, and the applications will not have to specify which hardware should be used for a given processing task. Accelerated computing will provide a software framework that includes platform abstractions, a specialized runtime environment, application programming interfaces (APIs), and libraries to support this simplified, flexible interaction with hardware accelerators.

Overall, accelerators provide several important benefits over general purpose central processing units (CPUs) alone:

- They speed up certain serial tasks as well as supporting parallel algorithms.
- They tend to have lower latency and higher throughput for processing tasks.
- They are more energy efficient: within a given power envelope, special-purpose hardware accelerators will run better and faster than general-purpose CPUs.

MARKET TRENDS: DRIVING THE SHIFT TO ACCELERATED COMPUTING

Several market trends are driving the shift to accelerated computing:

- Demand for excellent application performance on energy-efficient, competitively priced computer systems.
- The equalization of hardware and software costs.

- The move to virtualization in server and data center computing.
- Diverging system requirements and unique needs.

Performance, Energy Efficiency, and Cost

Commercial and consumer users alike have come to expect high-performance computing experiences with the lowest possible cost and highest energy efficiency. These expectations have been fueled by the consumer electronics industry, which consistently delivers devices that provide these benefits. For example, an emerging grass-roots metric for laptop performance is the number of movies a person can watch on one battery. Consumers want to watch multiple movies with a high-quality picture on one battery, and they don't want to spend \$2,500 on the laptop to get that kind of quality and performance.

Commercial users are also driving the demand for low-cost, energy-efficient systems that can support high-performing virtualized environments as well as applications that deliver media-rich content. As these concepts become significant drivers in the computing industry, using specialized hardware to improve performance while keeping costs and energy use low will become increasingly important.

Equalizing Hardware and Software Costs

In the past, hardware development costs were much greater than software development. However, software development costs (e.g., salaries, testing, and support) have increased to the point where costs in both areas are roughly the same. Consequently, attempts to improve hardware while ignoring software overlook a large component of the total system cost. Accelerated computing addresses this trend by providing a framework that supports faster development of applications that can take advantage of hardware accelerators.

The Move to Virtualization

Early efforts in virtualization focused on server consolidation to achieve better returns on investment in system management and resource utilization in data centers. Virtualization is rapidly becoming a mainstay of enterprise computing as its uses expand into storage, application delivery, and other areas. The trend toward virtualization is changing the data-center computing environment, with storage hosted in one node, workloads on another, and communication overhead distributed across the network.

Virtualization is actually a key component of an intelligent accelerated computing infrastructure because it supports a tight coupling between CPUs and hardware accelerators. The traditional approach to offloading work to an accelerator

uses a loosely coupled connection between the CPU, the operating system, a device driver, and the accelerator. In a virtualized environment, using traditional device drivers that rely on dedicated system resources is not an option. All virtual machines must have access to all accelerators in the system, which means that the accelerators must be virtualized and look like a peer processor rather than a device. This requires a closer relationship between the CPUs and the accelerators. The accelerated computing framework facilitates the move to tightly coupled silicon in virtual environments and creates greater flexibility and better energy efficiency in data centers.

Diverging System Requirements

Computing environments today are diverging in several areas, making integrated solutions unfeasible:

- **DIVERGING WORKLOADS:** The workloads in today's data centers are rapidly splitting into two categories: basic productivity applications and functions (like file, print, and storage) and performance-sensitive mission-critical applications (like financial transaction processing and technical computing). Administrators can increase the performance of mission-critical applications by using hardware with specialized accelerators.
- **POWER AND FORM FACTOR REQUIREMENTS:** From handhelds to servers, power and form factor requirements are diverging rapidly. Using appropriate accelerators can increase performance on many of these platforms. For example, placing a video accelerator in a laptop would let the user watch a high-definition movie without consuming all of the CPU and graphics processing unit (GPU) cycles. Also, using special-purpose accelerators to process high-bandwidth network communications (e.g., 10 GbE) in data centers would be more efficient and create better performance than passing that traffic through CPU cores.
- **DENSITY:** Some newer computing approaches, including cloud computing, provide value around optimized packaging and cooling for applications and hardware. Accelerated computing supports these approaches by providing integrated processors that help further optimize the use of computing resources.

Accelerated computing provides a general platform that allows off-core chips like network and media accelerators to be added to general-purpose multicore systems. The approach reduces the cost of individualized, integrated solutions.

AMD: AT THE FOREFRONT OF ACCELERATED COMPUTING

All of these market trends are driving the move toward accelerated computing, and AMD is committed to influencing the direction of this movement. We have always maintained a forward-thinking approach to computing, which has created a track record of wise choices regarding future direction and development:

- Evolving from 32-bit to 64-bit processing with AMD64.
- Developing the Direct Connect Architecture to reduce bottlenecks and increase processing throughput with integrated memory controllers and HyperTransport™ technology.
- Designing and producing dual and multicore AMD Opteron™ processors.

In keeping with that approach, we recognized early on that silicon and software development costs would be too high to support individualized processing solutions for emerging application areas. Consequently, we have invested considerable time and resources into understanding the factors leading to the current inflection point.

We have been building expertise and capability in general- and specific-purpose computing for the last several years. For example, our initiative code-named Torrenza and our Stream Computing work produced important research findings about how to effectively couple general and specific purpose processors, which built a solid foundation for further research and development in accelerated computing.

The key to accelerated computing lies in effectively leveraging processing power from general-purpose CPUs and specific-purpose processors like GPUs, media accelerators, and others. The computing industry knows that GPUs will eventually be integrated with CPUs, and we acted on that future trend by acquiring ATI Technologies. This move added significant GPU and multimedia expertise to our already extensive experience with general-purpose processors, creating a platform from which to drive innovation and development for future computing needs around data, voice, photo, video, and face recognition, among others.

Just as importantly, we are actively collaborating with large research institutions, independent software vendors (ISVs), the open source community, and other organizations to develop an open ecosystem around accelerated computing. Our commitment to accelerated computing includes designing

hardware accelerators, building developer tools, and driving thought leadership by participating in standards-development bodies. We believe that accelerated computing will be the result of the coming paradigm shift, and we intend to offer commercial and consumer users flexible solutions that meet their computing needs based on the components available in the growing ecosystem.

The rest of this paper provides a closer look at accelerated computing, including a brief history of innovations in processing power to improve application performance, different approaches to accelerated computing, and the value that this framework provides to commercial and consumer users alike.

X86 PROCESSOR GENERATIONS AND APPLICATION PERFORMANCE

For the last 20 years, the focus of general-purpose processing (the x86 platform) has been on improving application performance through technological advancements in processor frequency, core microarchitecture, and multiple cores. The baseline assumption for defining “performance” and measuring increases has been that frequency times the amount of work per clock cycle equals application performance. As the computing industry has matured, vendors have taken different approaches to improving application performance using this assumption (Figure 1).

GENERATION 1: A FOCUS ON FREQUENCY AND ARCHITECTURE

In the 1980s and 1990s, vendors focused on designing single-core CPUs that would run single-threaded applications. At the time, increasing the frequency of the CPU was the clearest path to enhancing application performance because the higher the frequency, the more work that could be done per second. Vendors also worked on improving the core microarchitecture of their CPUs as a method to deliver better application performance. These architectural improvements were driven in part by the realization that continuing to increase a processor’s frequency to enhance performance was not a viable option. Focusing on frequency alone led to increased design complexity and size as well as unacceptable thermal limits and power consumption. This processor generation ended with a dramatic increase in power consumption and a significant decrease in the potential for future frequency increases (this was due to pushing the underlying microarchitecture to its limits and the declining contributions for frequency in new process technologies). Ultimately, these factors set an impending limit on single-threaded application performance and opened the door for a new era of parallel computing.

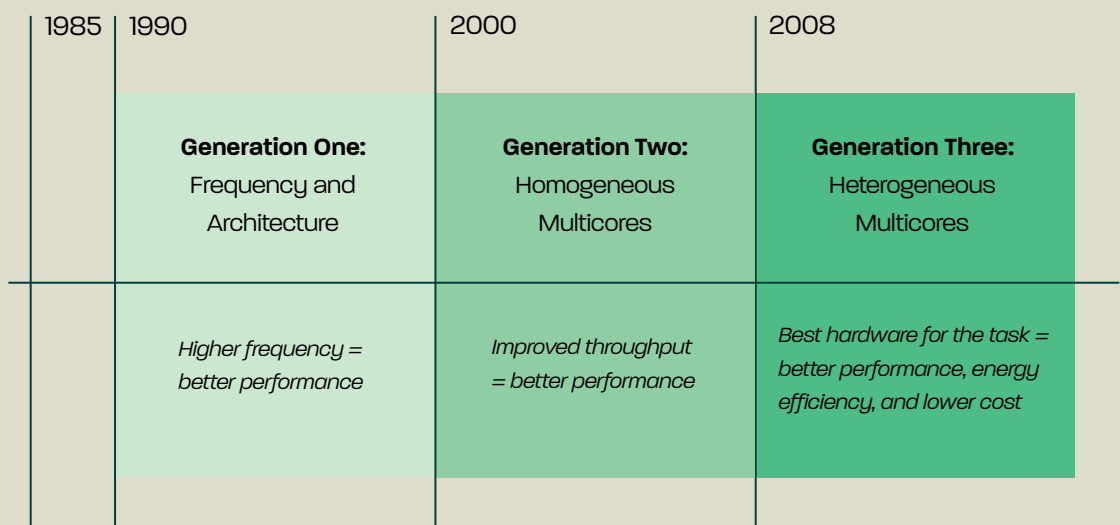


FIGURE 1. X86 PROCESSOR GENERATIONS FOCUSED ON DIFFERENT METHODS FOR IMPROVING APPLICATION PERFORMANCE

GENERATION 2: HOMOGENEOUS MULTICORES

In the early 2000s, the industry began to introduce homogeneous multicore computer systems—multiple cores of the same type built on one processor die. The goal of this processor generation was to create better application performance through improved throughput. The hardware technology that enabled multicore processors has evolved to the point where computer systems can now scale up to a very large number of CPU cores, although this generally results in untenable power envelopes and very high solution costs.

The Achilles' Heels of General Purpose Multicore Systems

While adding more cores to a system is one option for continuing to improve application performance, it isn't always the best one. When it comes to enhancing application performance, general-purpose, homogeneous multicore computer systems suffer from two Achilles' heels:

- They cannot keep up with the performance demands of modern and emerging computing applications while still meeting the criteria of energy efficiency and low cost.
- They are limited by current serial programming models that cannot take full advantage of the parallel hardware architectures inherent in multicore systems.

Characteristics of Modern and Emerging Computing: Even with multiple cores, general-purpose processing is hard pressed to provide the performance, cost savings, and energy efficiency needed for today's data-driven computing needs. Modern

applications fall into two performance categories:

- COMPUTE-PERFORMANCE applications involve the collection and manipulation of large volumes of complex data in a single process (i.e., how fast a task can be turned around). Examples of these applications include oil and gas exploration, drug development, and contextual search.
- THROUGHPUT-PERFORMANCE applications process massive amounts of data per second from transactions that generally have short life spans, follow predictable processes, and contain standard data elements. Throughput performance (i.e., how many tasks can be completed within a give time span) is vital for e-commerce sites, financial transactions, and mobile phone call monitoring.

While general purpose CPUs can be used in high-performance computing applications like geophysics, science, and simulation, offloading specific processing tasks to specialized accelerators can dramatically improve power, space, and cost efficiency.

Current Serial Programming Models: Homogeneous multicore systems can support multi-threaded applications, but they do not necessarily enhance single-threaded application performance. In fact, vendors have found that general-purpose multicore systems tend to reach a point of diminishing performance returns as they scale beyond eight cores.

This discovery is supported by Amdahl's Law, which is used to find the maximum expected improvement to an overall system

when only part of the system is improved (Figure 2). In the case of application performance, the hardware has been improved through the addition of multiple cores with parallel processing capabilities. However, the applications themselves are limited by serial processing models and techniques. Regardless of how many cores are available, many sit idle as they wait for data to be processed serially. And even when certain algorithms are parallelized, the majority of the application's processing is still serial. The end result is that application performance on homogeneous multicore systems tends to be lower than the anticipated theoretical peak performance.

in the extra socket available on the motherboard.

In a similar way, this next processor generation focuses on offloading certain processing functions to specialized hardware accelerators to improve overall performance. As an integral part of accelerated computing, these processors provide the hardware foundation for today's data-driven and media-rich computing experiences. Because AMD is taking a holistic approach to accelerated computing, this processor generation will be designed to help provide better power efficiency and lower costs for hardware as well as dramatically improving

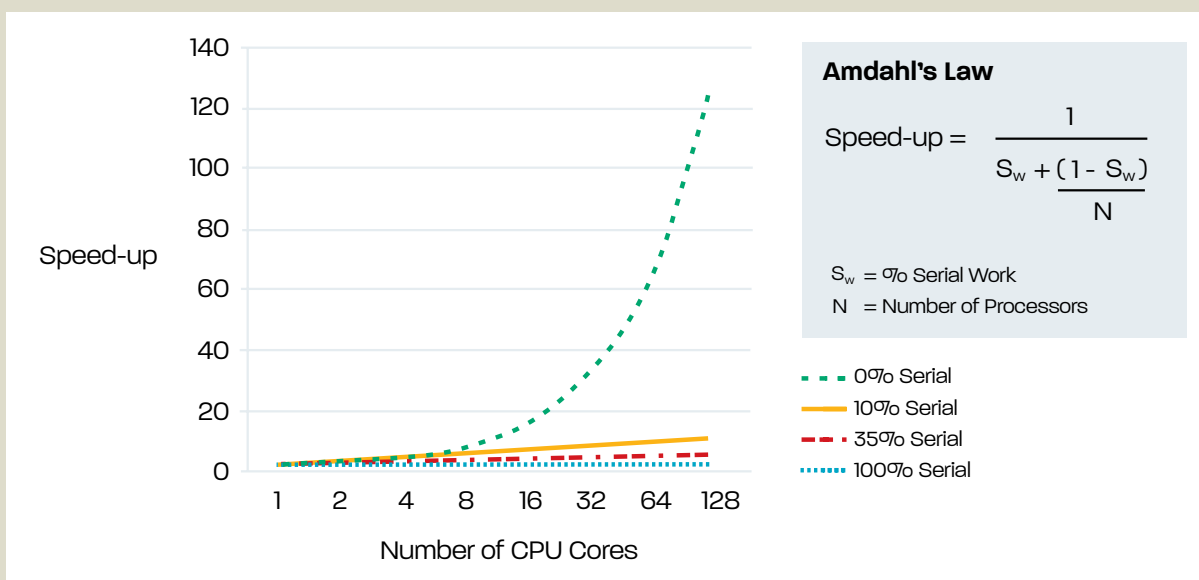


FIGURE 2. AMDAHL'S LAW SHOWS HOW APPLICATION PERFORMANCE IMPROVES THE AMOUNT OF SERIAL WORK DIMINISHES AND THE NUMBER OF PROCESSORS INCREASES

GENERATION 3: HETEROGENEOUS MULTICORES

The next processor generation is focused on creating heterogeneous multicore computer systems: computers with multiple processors that have different capabilities. A heterogeneous multicore system could contain several general purpose CPUs plus several accelerators designed to enhance the performance of a particular function (Figure 3).

This processor generation is at the heart of accelerated computing. And interestingly, its emergence is analogous to the introduction of math coprocessors in the early days of personal computing. When the original IBM PC was introduced in 1979, it contained an 8088 CPU, which did not include a math coprocessor because most software at the time was not math intensive. Anyone who needed more advanced math functions bought an 8087 floating point math coprocessor and installed it

programmer productivity to help reduce the cost of software development.

APPROACHES TO ACCELERATED COMPUTING

As accelerated computing gains momentum in the market, three distinct approaches to using accelerators are emerging:

- CPU-CENTRIC APPROACH: This approach seeks to improve application performance by extending the instruction set of the general purpose x86 architecture. While a general-purpose processor can do practically any type of processing task, using only CPUs does not provide the simultaneous cost, performance, and energy efficiency benefits that an accelerator can provide.

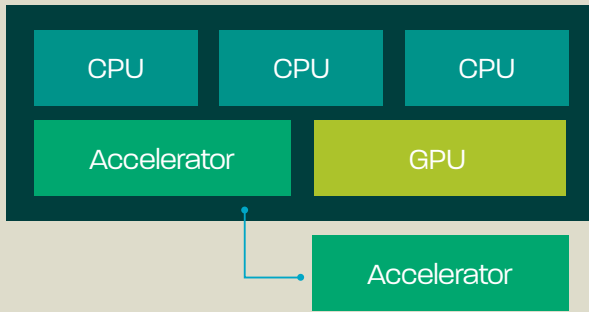


FIGURE 3. A HETEROGENEOUS MULTICORE PROCESSOR CONTAINS MULTIPLE CORES DESIGNED FOR SPECIALIZED TASKS

- GPU-CENTRIC APPROACH: This approach seeks to meet modern computing needs by using only GPUs for all processing tasks. Although GPUs are well-designed for coarse-grained parallel tasks, they do not provide the broad range of capabilities that a general-purpose CPU brings to processing tasks, nor do they easily support fine-grained parallelism.
- BALANCE AND OPTIMIZATION APPROACH: This approach focuses on choosing the best hardware to optimally process a given workload. It uses CPUs to orchestrate system activities, balance workloads, and determine which tasks to offload to appropriate accelerators. Accelerators are added to the system to complement existing CPU abilities and improve application performance. This approach makes it possible for computer systems and applications to be optimized for stream, parallel, sequential, or single-threaded processing, depending on need.

AMD has chosen to follow the balance and optimization approach. Because of our extensive experience developing CPUs, GPUs, and accelerators, we believe that the best approach to accelerated computing is to cooperatively develop the overall ecosystem (software as well as hardware), provide the processors that we have the most expertise in, and empower customers to choose the components that create the best performance for the applications they need to run. We intend to provide a set of accelerated computing solutions targeted toward different markets, including, server, desktop, and laptops. And our work with the *Torrenza* initiative has given us the ability to easily integrate external accelerators so our customers can differentiate and customize their computing solutions as needed.

¹ Waters, "Grid," Bob Giffords, 1 April 2008, (<http://db.riskwaters.com/public/showPage.html?page=788660>)

² Ziff Davis Enterprise CustomSolutions, "Expediting Time to Insight via Accelerated Computing," 2007

ACCELERATED COMPUTING: LIGHTS, CAMERA, ACTION!

Accelerated computing boils down to finding a balance of power, performance, and cost. From commercial customers with enterprise data centers to consumers who use their laptops to watch movies, combining general-purpose processors with hardware accelerators can help create that balance. The following sections provide brief descriptions of how accelerated computing is being implemented now and how it could be used in the future.

NOW PLAYING ON A COMPUTER NEAR YOU

Several industries are already implementing solutions based on accelerated computing. Each industry has specific performance requirements that demand processing power only available through hardware accelerators.

Financial Services

One critical performance driver for the financial services industry is low-latency trading. The amount of available market data that financial services organizations must process is overwhelming: for example, the full data feed from the Options Price Reporting Authority (OPRA) surpassed one million messages per second this year. The ability to process this data feed with the lowest possible latency (measured in microseconds) is key to success for these organizations:

Companies like Exegy and ACTIV Financial directly connect specialized off-chip accelerators called field programmable gate arrays (FPGAs) to homogeneous multicore processors using HyperTransport™ technology to create ticker plant appliances that process these huge streams of financial market data. Exegy's accelerated appliance has been benchmarked at processing one million messages per second (the full OPRA feed) with an average latency of 80 microseconds¹.

Telecommunications

As the demand for video on TVs, laptops, cell phones, and portable media players increases, telecommunications providers need systems that can transcode huge volumes of streaming video data to multiple devices simultaneously. In March 2007, Sun Microsystems introduced Sun Streaming System, a video delivery platform that transcodes video data and delivers personalized streams to end users. This platform relies on multiple integrated network cards and general-purpose CPUs to provide high performance and energy efficiency in this video-processing solution².

Security

Accelerated computing provides the processing power needed for advanced security functions such as deep content inspection, which is used to analyze data in intrusion prevention and antivirus applications. For example, antivirus engines can be required to inspect all data that enters a computer and then compare their findings against databases with many thousands of threats. Using specialized hardware accelerators that are tightly coupled with multicore CPUs can increase performance for deep-packet inspection and other advanced security activities.

COMING SOON TO A COMPUTER NEAR YOU

The emergence of media-rich applications, as well as increasingly data-driven computing tasks, are creating many new opportunities to implement the accelerated computing framework. Here is a brief sampling of the broad spectrum of possible ways that accelerated computing could be applied in the near- to mid-term.

Future High Bandwidth Data Centers

As very high bandwidth Ethernet becomes a reality, current processing methods will not be able to keep up with network communication. For example, a server running the full TCP/IP stack uses most of its CPU power to process the traffic generated on a 1-GB Ethernet network. Because the network processing alone consumes so many CPU cycles, implementing

10-, 40-, or 100-GB Ethernet will require another approach.

Network packet accelerators can provide an alternative model that balances power and performance for the data center.

Video Acceleration

Many consumer computer users use the number of movies they can watch on a single battery as a criterion for evaluating laptop performance. The accelerated computing framework provides several options for meeting this performance requirement, including using video accelerators to meet the criteria of high performance, energy efficiency, and low cost.

Video Conferencing Systems

High-end video conferencing systems available today deliver an experience so realistic that it looks like the person you are conferencing with is actually sitting on the other side of the table. However, this type of conferencing capability is generally limited to enterprise organizations with financial resources that can support this level of conferencing. The accelerated computing framework provides a pathway toward creating high-quality video conferencing on the desktop at a reasonable cost to end users.

Image Recognition

As our society produces increasing quantities of digital images—both still and video—the need to automatically recognize and classify images is becoming more urgent. For example, right now it is difficult to find images with specific content unless someone

THE LOOMING CRISIS

The growing gap between hardware and software parallel processing capabilities is creating a looming crisis in the computer industry. While hardware technology continues to advance in the area of data parallel processing, the parallel programming tools and skilled developers needed to extract maximum performance from multicore systems are not readily available. According to Stanford University, by 2010 software developers will face systems with heterogeneous multicore processors and application specific accelerators that run hundreds of hardware threads and leverage deep memory hierarchies to create over one teraflop of computing power.

While some data parallel applications are being written to address the increasing role of media-rich applications and virtualized environments, developers have realized that writing these kinds of applications at a low level is very challenging—especially in the areas of synchronization, communication, and scheduling. And for the moment, higher-level tools and languages to support data parallel development efforts do not exist. Another complication is that the computer science education system is not prepared to teach parallel computing skills.

Given this state of affairs, Stanford University has created the Pervasive Parallelism Lab (PPL). This project focuses on creating high-level

programming tools and providing educational opportunities so that current developers can more easily create parallel programs. PPL is a combined effort between top Stanford researchers in applications, languages, systems software, and computer architecture and leading companies in computer systems and software. In addition to developing parallel algorithms, development environments, and runtime systems that scale to 1000s of hardware threads, PPL members are also working on making parallel programming a core component of undergraduate computer science education. This educational goal will develop a new class of scientists and engineers who can close the current parallelism gap and avert the looming crisis.

has manually tagged the images and added the content to a searchable database. Large image sites pay people to do this to provide a better customer experience, but the average person with thousands of digital pictures does not take the time to tag all of their photos for easy searching.

The same is true of video: the only way to find your favorite actor in a given movie and scene is to consult a source that has viewed the movie and manually recorded the information. Automating this type of image and video indexing would be a welcome addition to modern computing, and putting the capability in a \$50 application would make commercial and consumer computer users alike very happy.

Of course, this type of software is computationally complex and involves an incredible amount of processing. Applications that can process images and video, look for patterns, and create metadata that represents what is contained in the images will require dedicated hardware accelerators to handle these tasks without bogging down the entire computer system.

Healthcare and Network Security

The added emphasis on information privacy in the healthcare industry has prompted new levels of network security. Instead of simply controlling who has access to healthcare information, IT departments in hospitals, insurance companies, and other healthcare institutions must conduct analyses of the information that comes into their networks. Common types of security analysis include packet filtering, but more advanced techniques like deep packet analysis, which looks beyond the packet header to analyze the payload for potential threats, take many processor cycles. By offloading deep packet and other security analyses to a specialized hardware accelerator, healthcare IT departments could significantly increase application performance without reducing available system resources and still meet energy efficiency and cost criteria.

THE CASE FOR ACCELERATED COMPUTING NOW

Accelerated computing is not some passing fancy that will disappear as quickly as it appeared. Quite the opposite: accelerated computing is a continuation of the current trend toward integration at the level of silicon, consumer computers, and data centers. In terms of silicon integration, advancing process technology is letting vendors place increasingly complex designs on a single piece of silicon. This makes it

possible for them to put different capabilities (general- and specific-purpose processors, for example) on the same piece of silicon, which increases processor utilization, enhances energy efficiency, and assimilates capabilities that used to reside on separate chips.

This integration passes on performance improvements to applications that support media-rich experiences for commercial and consumer users. And in the data center, it supports the trend toward virtualization as a means of increasing resource utilization and reducing management costs. This entire trend focuses on extracting more processing power per unit (e.g., performance per dollar) from all areas of the computing system, which translates into excellent performance for emerging applications that require more processing power.

Accelerated computing also makes the x86 instruction set more affordable and applicable to many different uses by allowing computer systems to split and balance workloads between CPUs and accelerators. For example, CPUs excel at coordinating and processing linear tasks, while GPUs are designed to process large amounts of data in parallel. The more media-rich or data-intensive the workload, the more parallel processing is needed to provide a great computing experience, so the presence of dedicated GPUs to improve performance is critical. Accelerated computing provides the framework for easily creating or modifying applications so their workloads can be split into serial and parallel tasks and then sent to the processor that can provide the best performance.

The trend in modern computing toward visually rich content spans all levels of computing because people tend to remember visual content better. High-end workstations are by nature graphic-intensive, and standard client computer systems in commercial business are being called on to present more media-rich content, including presentations and data mining applications that present data visually. From a consumer computing perspective, we are experiencing a distinct shift from text to multimedia content. And finally, current and emerging operating system user interfaces are becoming more graphically rich.

In full bloom, accelerated computing will create many opportunities to customize computer systems by providing targeted, precise solutions for different kinds of workloads. And it will create those opportunities while satisfying the need for application performance, energy efficiency, and low cost.