

WHITE PAPER

The Next Evolution in Enterprise Computing: The Convergence of Multicore x86 Processing and 64-Bit Operating Systems

Sponsored by: Advanced Micro Devices Inc.

Kelly Quinn

Jessica Yang

Vernon Turner

April 2005

IDC OPINION

Over the past few years, the x86 server market has seen significant gains in several areas. Key among these are substantial improvements in price/performance ratios, increased economies of scale by deploying scale-out architectures in the datacenter, and continuous increases in processing approximately every 24 months, as described by Moore's law.

However, these changes — although they have proven to be significant in their time — will pale in comparison to the changes that lie before us in the next stage of the evolution of the x86 server market. In the past year, this market has seen the technologies of multicore and 64-bit processing and operating systems develop from [esoteric] idea to tangible product. When assessed discretely, multicore and 64-bit technologies introduce into the x86 server market tremendous performance improvements that will require users to reexamine the ways in which they deploy servers in their datacenters. But together, multicore and 64-bit technologies will change our understanding of the function of servers and open up new opportunities for the market to develop entirely new architectures.

IN THIS WHITE PAPER

This white paper provides an overview of the market progression toward 64-bit computing and multicore processing to satisfy the need for greater computing performance in the x86 market. The convergence of these two technologies will deliver mainframe-class power and performance to mainstream platforms, enabling applications that traditionally run on higher-end systems to also run on lower-priced, industry-standard server platforms. IDC believes that this ability to employ scale-out platforms for scale-up workloads will bring about the next evolution in enterprise computing. This paper discusses the implications of this new computing platform on existing IT infrastructure and, more important, on the future of computing.

SITUATION OVERVIEW

In recent years, many IT managers have struggled to satisfy opposing demands with regard to computing environments — namely, the need to reduce costs while delivering higher levels of service and performance to end users. This business problem plagues companies of all sizes. IDC research shows that as the server installed base continues to expand over time, management and administration expenses of these systems will account for an increasing share of server spending, outpacing growth in new hardware acquisition costs. Many IT organizations today continue to operate with reduced IT budgets, and therefore they are constantly looking for ways to drive costs out of the datacenter and improve efficiency and utilization.

One means to lowering capital and operational costs is through server consolidation, whereby organizations centralize compute resources that can then be dedicated to specific users. This approach reduces the number of server systems and eliminates inefficiencies that exist in traditional compute silos. Companies will also be able to benefit from lower ongoing maintenance and management costs because fewer systems will need to be managed by IT administrators.

The proliferation of server systems today can be attributed to the common practice of deploying one application per server. Additionally, organizations often overprovision IT resources for peak usage, yielding low utilization rates. To combat the inefficiencies inherent in such IT environments, IT managers are using virtualization technology to pool hardware resources and share compute cycles. In doing so, many enterprises are able to recognize additional savings from not only physical server consolidation but also better use of currently deployed hardware assets.

While organizations continue to pursue consolidation and virtualization initiatives to drive down costs and improve efficiency, IT managers are battling to stay ahead of the ever-increasing demands from end users. The large amount of data sets being created is a catalyst for the increasing needs for greater processing power and performance. For example, the data being created for the Internet from personal information to corporate databases, combined with the accelerating number of users online, requires infrastructure environments capable of handling heavy transaction processing. As software applications grow more complex and rich digital media data becomes more commonplace over time, users will require sophisticated and higher-performing microprocessors capable of executing large data sets in a shorter amount of time with superior quality. In fact, the next generation of software development and the future world of lifelike digital multimedia will depend on a tremendous performance enhancement in compute processor technology and a fundamental shift in the existing IT infrastructure.

The industry needs a solution that will not require corporations to increase their datacenter footprints or result in higher levels of power consumption. This solution cannot rely solely on next generations of faster processors because of greater thermal requirements, the inevitable by-products of higher clock speed processors, especially as datacenter environments become increasingly dense over time. The evolution of processor technology to multicore design will address the market's need for a solution that provides performance gains within fixed power and physical space limitations.

THE NEXT GENERATION OF PLATFORM COMPUTING

Four decades ago, Gordon Moore voiced what was to become the fundamental rule of microprocessor engineering for the subsequent computing revolution: the number of transistors on a chip doubles every 24 months. As human engineering and physics proved the validity of his assertion, the rule came to be considered law. But in recent years, a question has arisen: What happens when the conditions implied in Moore's law reach their limit as determined by the fundamental laws of physics?

Processors are tangible, physical items that make use of the flow of electrons across transistors to construct a platform to build computational systems. However, no matter how sophisticated the applications they enable, these processors are ultimately bound by the basic restrictions of the physical world. In current designs, a logical boundary exists around a 16nm transistor size, whereby electron transmission can no longer be controlled predictably.

What does this mean from a business perspective? Obviously, the laws of physics are not top of mind for IT buyers in 2005. However, as transistor size decreases and transistor density and computing power increase, heat generation becomes a major issue, and the control of this heat and solutions for heat dissipation become top problems for IT buyers. These challenges also directly affect the viability of Moore's law, and we find that they translate to substantial concerns.

This is where multicore becomes important. Moore's law assumed single-core processors as the underlying chip archetype — and it holds up for those chip types. However, through further advancement in the manufacturing process, chip makers are now able to produce dies with more than one core, which in turn enables the manufacture of chips with more than one core each in the same physical space where only one core could previously be found.

So, rather than invalidate Moore's law, multicore's design approach operates outside the implied conditions that are present in Moore's original assertion. And, although in this light it appears to be a technology that's tangential to Moore's law, it is an area of development that may help manufacturers tackle some of the most difficult challenges in the industry today.

Multicore Processing Creates a New Business Solution

From an IT buyer's perspective, multicore processing has inherent advantages. First and foremost is the increased economy of the chip design, an element that enables greater computing power without introducing an unwieldy level of increased heat generation, power consumption, and so forth.

To date, the majority of processing performance improvements have come from innovations to cache, clock speed, memory access, and I/O. However, each of these advancements has required an additional increase in power consumption. As IT managers well know, the problems of heat dissipation and power consumption are already significant issues in the datacenter. Increasing heat and power to produce

incrementally improved performance will raise the question of diminishing returns at best. Additionally, today's offerings restrict the ability of IT facilities planners to maximize the infrastructure and therefore force them to add real estate and conduit, among other elements, to run systems that were supposed to reduce the cost level of ownership and safeguard long-term capital investments. A key benefit of multicore chips is the improved performance without a stratospheric spike in power or heat. The architecture of dual- and multicore processors is different from that of single-core chips. Rather than each chip having separate architectural components, such as memory or I/O, the chips share some of these resources. As a result, the power required to support these shared resources is minimally higher than that required for identical resources on a single-core processor. The only increase in power consumption comes from the addition of the extra execution cores to the processor, something that results in a relatively small addition to the overall power demands and heat generation. The result is a processor that provides greater computing power than a single-core processor, without doubling the power demands.

The reduced levels of heat generated by these systems enable vendors to manufacture servers that are denser than existing servers. This ability — combined with the reduced power consumption for greater processing power — comes as good news to IT managers who are facing increasing rates of physical hosting space as well as the ever-increasing costs of energy to power datacenters. IDC believes this technological advantage will be of particular benefit to both vendors and buyers of blade servers.

An additional benefit that may be derived from multicore processors lies in the processing power available for multithreaded 32-bit applications. As operating system and software vendors work their way toward a 64-bit world, compute-intensive workloads can take advantage of the increased processing power to support 32-bit applications and operating systems. Because of the architecture of the dual- and multicore chips, discrete threads can be allocated to the distinct cores within the processor, thereby increasing speed and reducing latency in data and transaction processing.

One final consideration is that of multicore processors in support of virtualization efforts, a trend that is a hot area for users across different industries and an approach that is intrinsically tied to the current progress toward greater consolidation in the datacenter. As stated in the previous point, specific applications and/or operating systems can be dedicated to distinct cores. In virtualization, this presents a uniquely compelling argument for the adoption of these new chips. One concern that is ever present in today's x86 virtual machine strategies is that of isolating applications and operating systems from one another while they run on the same hardware. With multicore chips, the different processors within the chips can "host" the different applications and operating systems in separate physical states. The architecture of these chips enables them to draw from a common memory pool, one that can be virtualized in a manner already in deployment across virtual machines. The resulting solution leverages the advantages of both hardware virtualization (one dedicated core per application or operating system) and software virtualization (access to shared resources that can be scaled according to demand).

Benefits of Multithreaded Environments

Before the market reaches ultradense servers, x86 users have an opportunity to derive significant benefit from multicore servers in multithreaded computing environments.

With multicore technology, highly threaded and parallel processing systems can run on x86 servers by virtue of each contained core's ability to allocate dedicated resources to each thread. Because of the unified I/O, memory, and caching that is shared across the multiple cores on the single chip, individual cores can manage distinct threads while coexisting in a low-latency environment. This significantly helps reduce existing latency when attempting to deploy multithreaded environments on single-core, multichip systems. This improvement also opens up the opportunity for multithreaded operating systems supporting either numerous single-threaded applications or some multithreaded applications common on client computers.

As a result, these significantly improved compute capabilities will provide the advanced technological environments required to drive a shift in workload processing. IDC believes that as multicore technologies develop and are adopted and deployed by server vendors, their ability to enable heavy multitasking environments will be the catalyst for an evolution in server workload management, one that results in the eventual migration of workloads previously deployed only in mainframe environments into the more pricing-aggressive x86 end of the server market.

Overview of x86 64-Bit Processing and Benefits

The advent of 64-bit processing marks another turning point for the x86 server market. A compelling driver for x86 64-bit chips is the demand for significant processing power — which has previously been available mostly through mainframe servers — at a much more reasonable price point, particularly for buyers in the small and medium-sized business (SMB) market segment. 64-bit processing enables smaller companies to support more compute-intensive workloads that, prior to the introduction of x86 64-bit processing, could be supported only through high-end, expensive RISC servers. The appeal for OEMs is clear — new products and a presently untapped market at compelling price points.

As a result of their architecture, x86 64-bit systems also offer the added benefit of investment protection for customers. Because x86 64-bit platforms are designed to enable future expansion in I/O, storage, memory, and processing capacity, buyers, particularly those in the SMB segment, will have minimal need for concern about being able to increase the capabilities of these servers as their applications and workloads grow to demand more and more processing power.

x86 64-bit processors offer compelling answers to problems faced in x86 32-bit addressable memory. Because 64-bit exponentially increases both physical and virtual memory address space, the resultant increase in addressable memory significantly benefits compute-intensive workloads, such as large databases and complex financial modeling. The increase in memory capabilities improves the performance of operating systems and speeds the processing capabilities of

multiapplication environments. An additional benefit of x86 64-bit processors can be seen in the increased ability for customers to deploy larger scale-out and clustering infrastructures for more compute-intensive workloads without requiring significant investment in expensive, high-end SMP systems.

As of 1Q05, Linux and Sun's Solaris 64-bit operating systems are commercially available in the x86 market. Microsoft has followed suit with beta versions of Windows® Server 2005 and XP Professional and has publicly stated that final versions of both will be available in 1H05. While x86 64-bit applications will most likely come to market in a sizeable way in 2006 and beyond, users can expect to receive boosts in performance and reductions in processing times as a result of using 32-bit applications on 64-bit operating systems that run on x86 64-bit chips.

CONVERGENCE OF MULTICORE AND X86 64-BIT TECHNOLOGIES

The improvements available in the move from 32-bit to 64-bit are impressive and become astounding when the improvements in abilities available through multicore processing are combined with those of 64-bit processing. The one-two punch of multicore and x86 64-bit processing represents the next evolution in enterprise computing. The impact of this convergence introduces a fundamental shift in existing IT infrastructure and has revolutionary significance for customers.

The combination of multicore with x86 64-bit processing will make it possible to offer scale-out platforms for scale-up workloads at lower price points. Applications and workloads, which traditionally run on non-x86 higher-end RISC systems, will be able to run on industry-standard server platforms. These platforms consistently make available an ever-increasing portfolio of computing power at consistently lower price points.

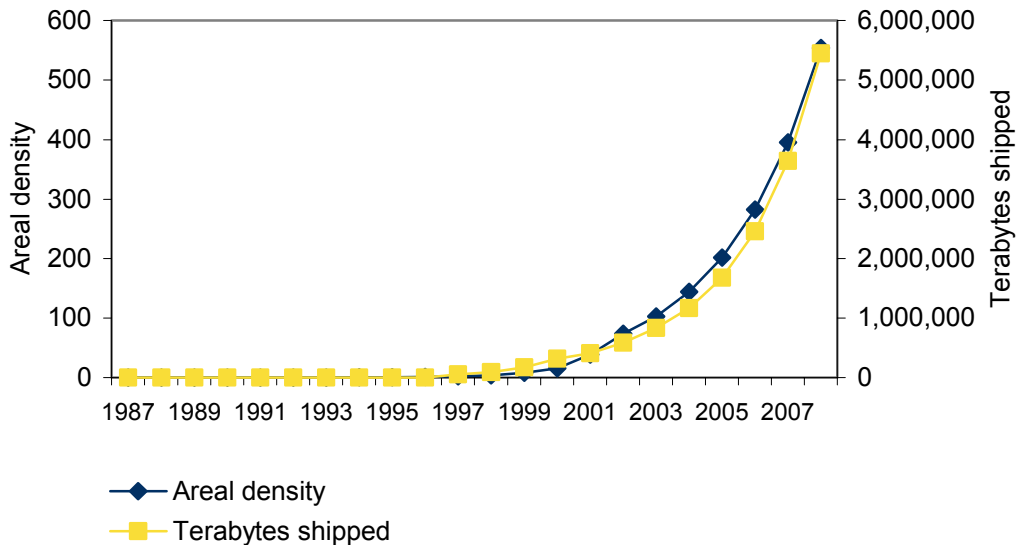
Multithreaded operating systems and applications in particular will benefit from this combination. They will experience increases in speed and processing power and decreases in latency as more resources within the chip become available as dedicated resources for them. Additionally, x86 blade servers will benefit from the reduced power and cooling needs of multicore technology. These new chips will allow for consistent deployment of fully loaded blade chassis and enable blades to take on higher-level workloads than they can today.

Disruptive Opportunity

Disruption of this type is not new. This is a cycle we have seen before in the storage market. In the late 1990s, expanded disk storage capacity allowed for the emergence of new markets — namely, SAN and NAS. Figure 1 illustrates that as drive density increased (areal density/y-axis), hardware prices fell and storage options became commoditized. As a result, the total sum of terabytes shipped spiked at an exponential rate, causing the explosion of what we now know as the storage market. IDC believes this model can be mapped directly to the future growth of the server market with the advent of multicore processing. As the number of cores per socket increases, the power of the server industry will increase at an exponential rate similar to the growth of the storage industry over the past five years.

FIGURE 1

Mapping Disk Storage to Disk Drive Density



Source: IDC, 2005

AMD MULTICORE STRATEGIES

Advanced Micro Devices Inc. (AMD) anticipated the ever-increasing need for processors capable of faster and greater performance when it set out in the late 1990s to design a future processor architecture. First, AMD addressed the needs for 64-bit processing capability with the 2003 introduction of its AMD Opteron™ processor family for both 32- and 64-bit computing in the x86 market. This new class of processors enables customers to deploy x86-based systems to tackle high-performance workloads that traditionally could have been run only on large SMP systems. Furthermore, the x86-64 extension technology offers customers a smooth migration path to 64-bit computing.

Once again, AMD is leading the charge in bringing multicore technology to x86 platforms. The company first presented its x86 dual-core strategy in 1999, setting a clear road map of its design intentions for the industry and its customers. AMD64 technology — AMD Opteron and AMD Athlon™ 64 processors as well as AMD Turion™ 64 mobile technology — is designed to accommodate multiple cores on a single die. In fact, multicore design is a natural extension to the architectural advantages of the AMD64 platform.

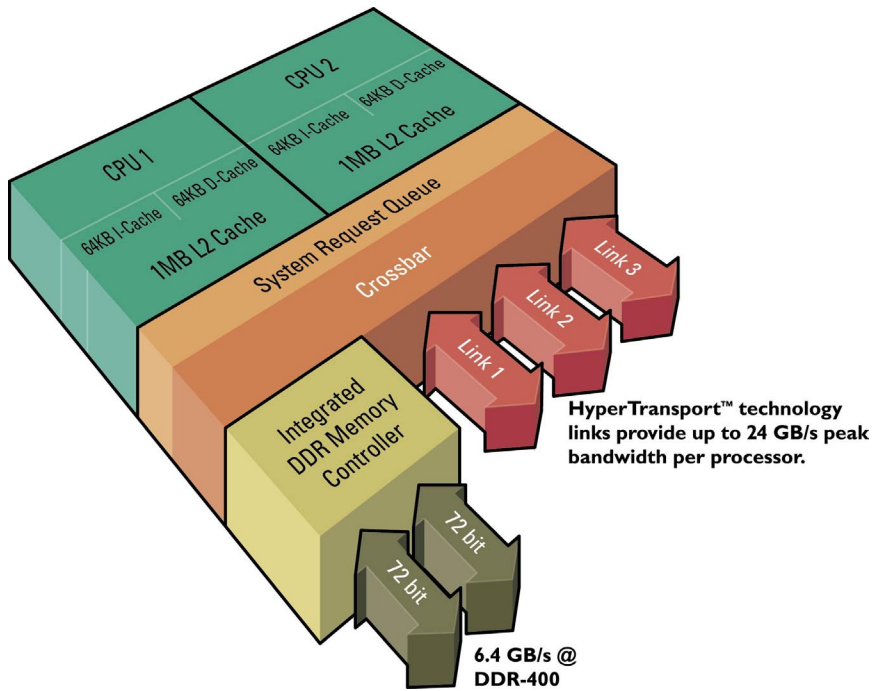
In designing its new generation of processors, AMD sought to eliminate the age-old problem inherent in the legacy processor design of connecting the CPU to memory and I/O through a front-side bus. AMD's engineering innovation resulted in the Direct Connect Architecture. There are several architectural advantages with this technology:

- ☒ Processors are directly connected using HyperTransport™ technology. This high-speed interconnect technology ensures high-bandwidth communication links between processors, resulting in increased scalability.
- ☒ Memory is directly connected to the CPU for optimized memory performance. An integrated memory controller changes how the processors access main memory, therefore reducing latency and improving processor performance.
- ☒ I/O is directly connected to the CPU, resulting in better balanced I/O and increased throughput.

These design points in the Direct Connect Architecture overcome the challenges and bottlenecks at the processor level and provide the scalability customers need to run existing 32- and 64-bit applications at peak performance. Now, AMD is taking Direct Connect Architecture one step farther by directly connecting two cores on a single processor (see Figure 2). IDC believes multicore technology will accentuate the benefits of the Direct Connect Architecture design because of the direct access to memory and I/O so that the additional core will not be bounded by the need to share access. The reduced latency and enabled scalability will allow end users to more fully realize the benefits and performance gains of multicore technology.

FIGURE 2

AMD Opteron Dual-Core Architecture



Source: AMD, 2005

AMD will first introduce dual-core technology in its AMD Opteron processor family. The company is deliberately targeting the server and workstation markets initially because many applications are already multithreaded and therefore ready to take advantage of multicore processing. Dual-core AMD Opteron processor-based systems promise to deliver significant performance boosts, especially for those multithreaded applications optimized for symmetrical multiprocessing. Additionally, many of today's datacenter challenges surrounding power envelope limitations, increased heat dissipations, and the desire for greater server density can be alleviated by AMD64 multicore technology because it will improve on its current industry-leading performance per watt as a result of the lower power consumption. AMD64 multicore technology makes it possible for more customers to take a scale-out approach to running scale-up applications while leveraging the lower price points of industry-standard x86 platforms.

As discussed earlier in this paper, multicore processor technology has significant development implications for server virtualization. Because virtual machines must draw from a common memory, Direct Connect Architecture enables very high-performance memory sharing between virtual machines. This is a critical design advantage as the processors become dual cores and multicores with virtual machines being hosted on multiple cores within a single processor. In addition, AMD is actively developing a technology codenamed "Pacifica" that is targeted at reducing the complexity in virtualization, which in turn will help virtualization software players and, more important, will reduce the performance penalties in virtualization implementation today.

AMD validated the dual-core technology in August 2004, when it became the first company to publicly demonstrate an x86 dual-core processor-based system, further extending its leadership position in the marketplace. AMD also has commitments from its OEM partners, including Cray, eGenera, HP, IBM, and Sun Microsystems, to integrate dual-core AMD Opteron processors into their server and workstation product portfolios. Such partner support is crucial in ensuring the availability of a wide set of products that will enable rapid customer adoption of multicore technology in the x86 market.

Current AMD customers may also take advantage of multicore processors with ease because AMD will use the same manufacturing processes to produce new multicore processors so that they will fit into the current 940 pin sockets. Users are able to easily upgrade existing systems compatible with the 90nm single-core processors to dual-core processors. This clearly illustrates AMD's commitment to provide a level of backward compatibility and offer investment protection for its current customers.

For the client and mobile PC markets, AMD plans to introduce dual-core AMD Athlon 64 processors in the second half of 2005. Multicore technology in the PC market heeds users' desire to multitask and, at the same time, satisfies the need for greater processing performance to handle increasingly complex software and growing data in digital media.

CHALLENGES AND OPPORTUNITIES

Of course, the dual-core AMD Opteron processor is not without its competition. Although multicore technology is new to the x86 market, it is readily available in the RISC server market in IBM's POWER5, Sun's Ultra SPARC IV, and HP's PA-RISC 8800 chips. Customers can already enjoy the same benefits promised by multicore processors by deploying RISC-based systems. Another non-x86 dual-core chip on the horizon is Intel's forthcoming "Montecito" processor in its Itanium family. Montecito was introduced in September 2004 at the Intel Developer's Forum and is expected to ship later this year.

The competing non-x86 dual-core processors came to market quite a bit ahead of either AMD's or Intel's dual-core products. Subsequently, these products can boast proven track records of reliability and solid performance. However, a key benefit of the x86 platform is the industry-standard architecture, as opposed to vendor-specific, proprietary chip architectures. Moreover, the introduction of dual-core processors to the x86 market will begin to apply price/performance pressure to non-x86 products from below. IDC believes the increased processing capacity and improved performance derived from the combination of x86 64-bit processing and multicore computing present an opportunity for x86-based systems to take on more workloads. Organizations will be able to take a scale-out approach on x86 platforms for scale-up workloads.

Although no competing x86 dual-core processors exist in the market today, Intel has plans to bring multicore technology to the Xeon processor family in 2006. As expected, Intel has an opportunity to leverage its installed base of more than 90% of the total x86 market as a target for this new processor. This represents a potential threat to AMD in the future. It's worth noting that AMD Opteron processor-based

systems have a larger profile of compute-intensive and high-performance applications due to the low latency and scalability benefits from the Direct Connect Architecture design. These same benefits are realized in varying degrees under all workloads and should be important considerations for customers as both chip vendors bring multicore technology to market.

One significant challenge that may arise to delay the adoption of multicore technologies is that of software licensing. In 2004, Microsoft made the decisive move of announcing to the market that it would charge on a per-processor basis rather than devising its pricing on a "per-core" approach. The software giant's decision came as welcome news to chip makers and end customers alike. What remains to be seen, however, is how other software vendors will approach the dilemma. Will they follow Microsoft's lead and come to market with pricing plans that are advantageous to end customers, or will they go after the near-term gains that may be available through per-core pricing schemes? Although IDC believes the long-term, sustainable model for software vendors is that of basing licensing on the number of sockets in a system — just as Microsoft has declared it will do — the market will not have a clear picture of the outcome of this problem until the end of 2005. It is a reasonable conclusion to state that, although licensing issues may construct some roadblocks in the initial adoption rates of multicore processing, these issues will be resolved to the advantage of chip makers and end customers in the long term.

CONCLUSION

The convergence of multicore and 64-bit technologies will deliver mainframe-class power and performance to mainstream platforms, enabling applications that traditionally run on higher-end systems to also run on lower-priced, industry-standard server platforms. The evolution of processor technology to multicore design will address the market's needs for a solution that provides performance gains within fixed power and physical space limitations. IDC believes that this ability to employ scale-out platforms for scale-up workloads will bring about the next evolution in enterprise computing.

With multicore technology, highly threaded and parallel processing systems can run on x86 servers by virtue of each contained core's ability to allocate dedicated resources to each thread. Individual cores can manage distinct threads while coexisting in a low-latency environment. This improvement opens up the opportunity for multithreaded operating systems supporting either numerous single-threaded applications or some multithreaded applications common on client computers.

As a result, these significantly improved compute capabilities will provide the advanced technological environments required to drive a shift in workload processing. IDC believes that as multicore technologies develop and are adopted and deployed, their ability to enable heavy multitasking environments will be the catalyst for an evolution in server workload management. The result will be the eventual migration of workloads previously deployed only in mainframe environments into the more pricing-aggressive x86 end of the server market.

Because 64-bit exponentially increases both physical and virtual memory address space, the resultant increase in addressable memory significantly benefits compute-intensive workloads, such as large databases and complex financial modeling, and increases customers' ability to deploy larger scale-out and clustering infrastructures for more compute-intensive workloads without requiring significant investment in expensive, high-end SMP systems. The combination of multicore technology with x86 64-bit processing will make it possible to offer scale-out platforms for scale-up workloads at lower price points. Applications and workloads, which traditionally run on non-x86 higher-end RISC systems, will be able to run on industry-standard server platforms.

Because AMD's Direct Connect Architecture is designed for direct access to memory and I/O, IDC believes multicore technology will accentuate its benefits, providing reduced latency and enabled scalability. For end users, the result will be greater realization of the benefits and performance gains of multicore technology.

The one-two punch of multicore and x86 64-bit processing represents the next evolution in enterprise computing. The impact of this convergence introduces a fundamental shift in existing IT infrastructure and has revolutionary significance for customers.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2005 IDC. Reproduction without written permission is completely forbidden.

