IDC
Analyze the Future

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

WHITE PAPER

IBM eServer 325 with AMD Opteron Processor: New Contender in the High-Performance Server Market

Sponsored by: IBM

Christopher G. Willard, Ph.D.

August 2003

FROM DESKTOPS TO SERVERS: THE NEXT FRONTIER FOR STANDARD PROCESSORS

Standard or volume computer processors are CPUs designed and manufactured by chip companies, such as AMD, Intel, and Transmeta, and incorporated into computer systems by computer manufacturers. The AMD Opteron processor is the new contender in the standard processor market. This CPU extends the x86 architecture into the 64-bit area. It is important to note that the *standard* designation refers to the availability of a technology and not necessarily its wide acceptance in the market. The Opteron processor is one such product vying for a place as a true long-term standard.

This white paper provides a perspective on the progress of standard processors, an overview of the AMD Opteron processor, and IBM's Opteron processor-based eServer 325 system, as well as an analysis of issues raised. Those issues include the technology transition process, processor and computer system balance, and opportunities and challenges that new processors and systems face when entering the market.

A BRIEF HISTORY OF STANDARD PROCESSORS

Standard processors were initially developed to meet the requirements for personal desktop computing systems. Over time, however, standard processors have also been used to address some server workloads. This move has been driven largely by improvements in processor performance as summarized by Moore's law. Clock speeds have improved dramatically, and by providing on-chip caches, processors can help mediate the difference between memory speed and processor speed. These improvements in performance allow standard desktop-oriented processors to be used as the basis for server products and cluster processing nodes under the following conditions:

- Naturally parallel applications where multiple processors can be applied most effectively

- Low-end throughput workloads (i.e., workloads that can be processed in parallel and do not have rigorous response-time requirements)

- Limited demand applications (e.g., print servers, email routers, and general computing resources for smaller organizations)

- Small memory applications (i.e., applications that do not require extensive addressable memory)

SERVER-ORIENTED STANDARD PROCESSORS

For a time, desktop processors could succeed in the server world because IT customers could match less costly servers to just those applications where desktop processor performance was sufficient. Server requirements not normally addressed by desktop processors include:

- Support for multiple users and large, demanding workloads
- Need for large memory (and strong memory) system performance
- Demand for scalability, with minimal restrictions of use as the system is extended
- Support for high-performance I/O and large file systems

Processor vendors have responded to these requirements by developing 64-bit processors specifically to address high-end server markets. AMD's Opteron processor is one such example.

DOES MOORE'S LAW ULTIMATELY SOLVE ALL PROBLEMS?

To a certain degree Moore's law describes raw performance growth that can help mask underlying architectural problems. However, Moore's law is a tide that lifts boats at different rates. Different technologies benefit from Moore's law at different rates leading over time to a growing imbalance among components in a computer system, for example:

- Processors, memory, and storage media have benefited the most. These components can benefit both from higher transistor densities and larger die sizes.
- Communication among components tends to suffer. Pin-outs cannot be reduced in size at the same rate as transistors shrink. Thus the number of wires carrying data between components cannot grow proportionally, and bandwidth among components becomes a limiting factor. Distances among components can also be limiting because latency is directly related to distance.
- Physical environmental factors have also suffered. The power necessary to send a signal increases with signal frequency, and the density of wires increases as components are pressed together to reduce latency. More power in a dense environment can lead to problems with heat dissipation.

It is important to note that Moore's law describes exponential growth in performance. Mathematically, the difference between two exponential curves is itself an exponential curve. Assuming different system components are described by different growth curves, maintaining system balance becomes increasingly difficult. Unfortunately, overall computer system performance will be limited by the slowest components.

OPTERON PROCESSOR TECHNOLOGY

The flagship feature of AMD's new Opteron processor is its implementation of the x86 instruction set with 64-bit memory addressing. The Opteron processor is capable of running 32-bit and 64-bit programs concurrently. The compatibility for 32-bit jobs is implemented natively — both 32-bit and 64-bit applications operate at full hardware speeds (i.e., at the same clock rate).

It is beyond the scope of this paper to describe the Opteron architecture in detail. However, three features of the Opteron processor — its integrated memory controller, on-chip HyperTransport technology, and 32- and 64-bit dual-mode capability — are of particular importance for server-class systems operations. Note that the first two features help to address component-level imbalances brought about by Moore's law. Details on these key features are as follows:

- The Opteron processor has an integrated memory controller that provides a dedicated pathway from the processor to memory. Ordinarily, access to memory is managed by an off-processor chip over a shared bus. The Opteron processor's on-chip memory controller implements a cell or NUMA-type memory structure, with each process directly connected to a subset of memory. Incorporating the memory controller on the processor chip provides several advantages:

    - ❑ Speed in terms of bandwidth — The most immediate effect the memory controller onto the processor is that the controller runs at processor clock speeds – roughly an order of magnitude speed up. This provides each Opteron processor with 5.3GBps of memory bandwidth.

    - ❑ Speed in terms of latency — Latency is improved by a combination of factors. Chip-to-chip communications delays are eliminated and contention for access to the memory controller in multiprocessor systems is eliminated in cases of local memory accesses. Interestingly, increases in processor frequency act to reduce controller level latencies. In essence, the memory controller does a fixed amount of work for each memory access. As the chip clock rate increases, the amount of "wall clock" time provided for the controller to do its work decreases.

    - ❑ Scalability — As processors and memory are added to a system, the amount of memory bandwidth available scales proportionally.

    - ❑ Reduced chip count — Moving the memory controller onto a processor eliminates the independent controller chips, reducing to some extent costs and power consumption.

- The Opteron processor incorporates on-chip HyperTransport I/O technology. HyperTransport was developed by a consortium of semiconductor, component, and system suppliers led by AMD. HyperTransport provides the Opteron processor with connectivity to high-speed interconnects, such as Fibre Channel, Gigabit Ethernet, and with PCI-X standard components. In addition, HyperTransport connects processors together for multiprocessor system configurations. HyperTransport provides several advantages:

    - ❑ I/O performance — Each Opteron processor has three HyperTransport links, and each link provides up to a total of 6.4GBps of I/O bandwidth (i.e., 3.2GBps input, 3.2GBps output). Total I/O bandwidth grows in proportion to the number of processors. In addition to raw bandwidth, the on-chip HyperTransport links provide a more direct path between processors and I/O devices. On-chip HyperTransport eliminates a hop to an off-processor I/O device and contention and data collisions associated with independent I/O busses.

    - ❑ Interprocessor communications — HyperTransport links between processors provide contention free, direct point-to-point, dual unidirectional (duplex) communications paths between processors for dual-processor system configurations. The multiple direct paths between processors support processor access to non-local memory.

- The Opteron processor can be run in two modes: long mode for 64-bit processing and legacy mode for 32-bit processing. In addition, long mode has two submodes: 64-bit mode and compatibility mode. In legacy mode, programs previously written for 32-bit x86 processors need not be recompiled in order to run. These programs see the Opteron processor as a 32-bit x86 system. In 64-bit mode, programs must be recompiled and will have access to the Opteron processor's 64 bit capabilities. The Opteron processor can run a mix of 32-bit and 64-bit programs and switch modes with each change in program. Advantages of dual-mode systems tend to fall into operational categories as opposed to technical performance categories and include:

  - ❑ Transition management — Dual-mode operations allow organizations to transfer entire workloads to the new system running in legacy mode. Applications can then be recompiled and tested for 64-bit mode on an as-needed basis.

  - ❑ Job-mix support — For organizations with a mix of applications requiring both 64-bit and 32-bit capabilities, dual-mode systems can support the organization's entire workload. In addition to potentially reducing the number and type of systems purchased, such operational flexibility can provide benefits in systems scheduling and utilization.

  - ❑ Operations efficiencies — The ability to run mixed workloads on a single systems architecture translates into a less-complex computing environment and can reduce staffing costs associated with additional programming and system management skills.

In addition to these three primary features, the Opteron processor also has twice as many registers as its predecessor. The number of general-purpose registers (GPR) was doubled to total 16, and all GPRs were expanded to 64 bits. The number of 128-bit streaming SIMD extension (SSE) registers was doubled to total 16 as well. Additional registers help improve systems performance by reducing requirements to access data from cache. This leads to fewer load and store instructions being required, since more data can be held in registers and leads to fewer processor stalls waiting on cache or memory.

HPC APPLICATIONS AND THE OPTERON PROCESSOR

Virtually any computer processor can be used to solve any computational problem. However, some architectural characteristics allow a CPU to address some classes of problems more efficiently than the *average* computer. IDC notes four characteristics of the Opteron that should be of particular interest for technical computing organizations:

1. Memory bandwidth — Many technical applications stream through data in memory performing the same calculations on all data. This aspect of high-performance computing (HPC) computation was the basis for vector computers. These technical applications are often limited by how fast data can be moved to and from memory. The Opteron processor's high local-memory bandwidth and memory scalability features should be of interest to organization with data streaming requirements.

2. Memory latency — Some applications work on sparse matrices (i.e., data sets where most of the values are zero). Other applications contain significant amounts of scalar code (i.e., calculations that stream data due to such factors as logical branching or dependencies in calculations). In these cases, memory latency becomes the critical path. The faster individual data items can be accessed, the faster the calculations proceed. The Opteron processor's on-chip memory controller and support for HyperTransport links directly address latency issues.

3. Large memory — Memory size affects many applications in two ways. First, as applications become more complex, they generate larger data sets that expand beyond the limits of 32-bit memory systems. A 64-bit memory space allows users to expand data sets without rewriting applications. Second, parallel applications tend to work better when running on more powerful nodes. The more work that can be done on a single node, the less time is wasted in coordinating efforts between nodes. Larger memory spaces allow vendors to produce more powerful nodes. This is of particular importance for clustered systems. Servers based on Opteron will allow memory capacity to scale with processors.

4. I/O performance — An ancient cybernetic proverb says that *a supercomputer is a system that turns a compute-bound problem into an I/O-bound problem*. Application data must be moved in and out of memory and, as applications grow in size, the requirements to get data in and out of the server also increase. In addition, no matter how large a system memory is available, HPC users will generate problems that extend beyond the server's capabilities requiring the processor to read and write to disk drives. The Opteron processor's on-board HyperTransport links work to address I/O performance issues.

Applications that require these four features cover the entire spectrum of the HPC market, from finite element analysis, to computational fluid dynamics, to visualization, to stochastic financial analysis, to genomic sequencing and matching, to cryptography, and the list goes on.

It is interesting to note that the above list does not include peak floating-point performance as measured by the number of floating-point functional units multiplied by clock speed. Issues around getting data in and out of the floating-point functional units often overshadow raw computational speed and this is addressed by major innovations in the Opteron architecture. The Opteron processor can perform two floating-point operations per cycle, and the theoretical peak performance of the 2GHz processor is 4Gflops.

IBM OPTERON PROCESSOR STRATEGY

In August 2003, IBM announced the eServer 325, which is based on the AMD Opteron processor and targets the HPC segment of the server market. The eServer 325 is a two-way symmetric multiprocessor (SMP) packaged in a 1U rack configuration. Table 1 shows the eServer 325 technical features.

The eServer 325 will also be available as nodes for the eServer Cluster 1350 solution. IBM is positioning the 1350 as a customizable "ready to run" cluster system. Clustered systems are products engineered and configured by systems vendors. In the case of 1350, IBM provides tested support for such system components as adapters, operating systems, compilers, and management software. The flexibility of cluster technology allows vendors to provide systems on a build-to-order basis. Vendors can leverage experience gained over time to develop strong cluster integration teams that can pass on their experience to customers.

TABLE 1

IBM eSERVER 325 TECHNICAL FEATURES

Processor	AMD Opteron; Options — Model 240 (1.4GHz), Model 242 (1.6GHz) or Model 246 (2.0GHz)
Memory	PS2700 DDR; Six DIMM slots for 512MB or 1GB DIMMs (2GB DIMMs when cost effective)
I/O ports and bays	Two 64bit/100 MHz PCI-X Slots: 1 full slot and 1 half-slot low profile
	1 video, 1 serial, 4 USB, 2 RJ-45
	Two SCSI Hotswap or IDE bays
DASD and DASD support	Hotswap SCSI disk: 10K RPM – 36.4–146GB; 15K RPM – 36.4 GB and 73.4GB
	Integrated Ultra 320 RAID 1
	IDE disk: 7.2K RPM – 40–120GB
	CD-ROM drive (24x slim)
	Single-channel U320 SCSI interface
	ATA/133 dual-channel IDE interface
System management	Temperature, voltage, fan monitor and fan speed control
	IBM SP/System Management Card support, Automatic Server Recovery (ASR)
	Full pre/post Text Console Redirect over serial and LAN
	Secure Remote Power On and Off Standard over Serial and LAN

Source: IBM, 2003

Of particular interest is the system management capabilities built into the eServer 325. System managers using Linux and Unix ordinarily expect to add hardware to a generic server to achieve this functionality. IBM has included industry-standard system management support as a part of its offering because it will be a necessary capability for what IBM believes will be eServer 325's predominant usage.

IBM intends to launch the product with the Linux operating system provided by SuSE. Operating systems from other Linux suppliers and from Microsoft are expected to be available later in 2003. A basic suite of cluster system management software, compilers, mathematic libraries, and message-passing libraries necessary for high-performance computing will be available when the eServer 325 begins shipping, although some functionality will not be available until 2004. IBM's cluster file system, for example, is slated for release in the second quarter of 2004.

POSITIONING FOR THE eServer 325

IBM believes that the primary use of the eServer 325 will be as compute nodes in a Linux cluster. IBM intends to price the eServer 325 aggressively so that it will compete on a price/performance basis. IBM expects that customers will value the eServer 325's dual-mode x86-compatibility.

Although IBM is targeting the eServer 325 specifically to HPC computing, which is predominately Linux and Unix today, IBM will also support Microsoft Windows for customers wishing to migrate to a processor with 64-bit memory access while maintaining the ability to run existing 32-bit Windows and new 64-bit Windows applications.

IDC ANALYSIS

The advent of standard processors for servers provides users with another technology choice along with low-end standard processors and RISC processors offerings. New technologies have the advantages of allowing organizations to expand existing applications, introduce new and more demanding applications, better manage resources through techniques such as resource consolidations, and improve cost management associated with generally improved price/performance characteristics of new technology.

PLANNING A TECHNOLOGY TRANSITION

The decision to move to a new technology is not a simple one. Technology transitions necessarily create some disruption as applications are ported to new systems. In addition transitions generate costs for applications conversion and testing, user and staff educations, operation of both new and old technology during a transition period, and so on. IDC believes that there are five key factors that computing managers need to consider in planning a technology transition.

- Strategic fit considers how well a given computer technology matches up against an organizations applications and workload mix. Strategic issues that affect processor choice include the following:

  - ❑ The memory requirements for critical workloads. Some workloads can run with the processor's cache memory. Other workloads require access to the server's dynamic memory (i.e., "in core"). Finally, some workloads are I/O-bound and depend upon access to data not stored in dynamic memory.

  - ❑ Memory size and processor-to-memory bandwidth requirements for applications and workloads

  - ❑ Configurations of multiple processors in shared memory systems

  - ❑ Workload requirements for 64-bit versus 32-bit calculations

- Road map takes into consideration how well the technology is expected to match user requirements over time. It is important for organizations to have reasonable assurances that the technology will keep pace with both the organization's growing requirements and improvements in competing technologies. Specific processor issues include:

  - ❑ Forecasts for performance and price/performance improvement

  - ❑ Schedule for introducing new expanded capabilities (e.g., larger caches, more sophisticated instruction processing methods, improved memory interfaces)

  - ❑ Plans for introducing new versions of the architecture (e.g., lower power, new processes technology or materials, multicore processors).

- Availability considers when the product will be available in the market in sufficient volume to guarantee easy access at reasonable prices? In addition, computing managers will need to assess whether the technology has the staying power to remain in the market place through several business cycles. Processor-specific issue include the following:

  - ❑ Complexity of the architecture, which is related to time-to-market

  - Financial resources and partner support for the supplier
  - Option to license the technology to multiple manufactures

- Ecosystem – A technology ecosystem is the set of independent hardware, software, and services companies that support the technology combined with a user base with knowledge of the technology. Processor ecosystem issues include:
  - Number and strength of systems vendors incorporating the processor into their product line, and component vendors providing products for the processor
  - Number of computer peripheral and component vendors shipping products that interface efficiently with the processor
  - Amount of proven system and middleware software available (e.g. compilers, code profiles, debuggers)
  - Support from independent software vendors (i.e., number of third party applications ported to and certified for the new architecture)
  - Availability of systems integrators, specialty programming services, and other service providers qualified to work with the processor

- In-house conversion costs consider the expense to implement the new technology over and above cost associated with simply upgrading existing technology. This is the most immediate issue that user organizations must face when changing technologies. Conversion costs fall into two broad categories:
  - Personnel cost centers around how much of current expertise can be used with the new technology and costs associated with hiring or retraining personnel to bring the overall level of proficiency in the new technology up to acceptable standards.
  - Application cost generally involves recompiling, testing, optimizing the tested code and then retesting. Additional application costs may occur in such areas as updating system calls in programs to account for changes in systems software and modifying an application to take advantage of new technology features (e.g., 64-bit versus 32-bit memory spaces).

Ultimately the decision to move to a new technology is based on comparing the costs and risks of adopting new technology versus the lost opportunity costs of continuing to work with older technology.

PROCESSORS VERSUS COMPUTER SYSTEMS

Baseball lore has it that the legendary pitcher Dizzy Dean's career was ended by a toe injury. The problem with the toe was not thought to be too serious, and Dean started pitching again before it was completely healed. However, the toe injury was just serious enough to cause Dean to change his pitching motion, which in turn caused him to *throw out* his arm. Dean's career was ended, and he retired.

Computer systems — like Dizzy Dean — can fail to perform due to seemingly minor or uninteresting components that create an imbalance in the overall system. Even when run on a server equipped with the fastest available processor, a customer's application many not come close to advertised peak performance due to the slow performance of system components relative to the speed of the processor. Analysis of computer system performance quickly becomes unmanageably complex due to the number of interdependent components that must work in harmony. Those that start with a single measure, such as processor clock speed, and attempt to apply this measure to how the overall system can be expected to perform on a given application quickly discover that the analysis is intricate and complex.

For example, the system clock speed should indicate the number of instructions a processor can perform in a given time period. However, the number of instructions actually performed is also dependent, in part, on the number of functional units available, and whether they are pipelined. In turn, the performance of the functional units is dependent on availability of data. Data availability depends on the number of data registers in the processor and how those registers are connected to the functional units. Register effectiveness is determined, in part, by how well data can be made available to the registers from cache. Cache performance is determined by cache size, hierarchical structure, coherency schemes, and memory performance.

The analysis could go further, but two related points should be clear. First, no single metric, such as clock speed, provides an accurate estimate of systemwide performance. Second, each component in a computer system can impact the overall performance of the overall system. How well the performance of various components is matched to, or balanced with, the performance of other system components can also significantly affect application-level performance.

SYSTEM BALANCE AND MOORE'S LAW

Ironically, one of the effects of Moore's law has been to make processor clock speeds a less meaningful measure of system performance over time. Since processor performance has increased at a faster rate than the performance of other system components, particularly memory, it is likely that other, slower components will limit overall system performance. In addition, to raw processor speed, users need to consider two other variables:

- Expected processor efficiency — The proportion of time that the processor is actually running the application workload

- Incremental cost of greater efficiency — The cost of additional programming, more or faster on-board memory or I/O devices, or other hardware component improvements that increase system efficiency

It is interesting to note that these attributes are not important in desktop computers that can stand idle the better part of the day. Users generally evaluate desktop processors on how well they perform the most demanding applications (e.g., 3D graphics, photo editing, "crunching" a large spreadsheet). In contrast, server processors can be kept busy continuously and efficient operation translates directly into money saved and thus improved return on investment (ROI).

We currently cannot point to any generally recognized efficiency measures for processors. However, this value can be measured for particular workloads by comparing the relative performance for the same workload on systems that differ only with respect to clock speeds for the same processors type.

PROCESSOR ARCHITECTURE AND SYSTEM PERFORMANCE

At a processor architecture level, the efficiency of a processor is largely affected by how well processor parameters are in concordance with overall system parameters. Key processor/system parameters are as follows:

- Memory size supported by the processor

  - ❑ Memory size is important because I/O systems are virtually always the slowest components in computers. When more information can be maintained in memory, applications will be delayed by slower disk performance less frequently.

  - ❑ A major advantage the Opteron processor is simply that its 64-bit address space will allow systems to be configured with more memory. In addition, each application or process can have access to more than 4GB of virtual and physical memory.

- Memory bandwidth to the processor

  - ❑ Generally, the second slowest activity on computers is moving data between memory and processors. The problem is confounded on multiprocessor systems, as the demand on memory bandwidth tends to increase as processors are added to a system. Multiprocessor designs reach a point of diminishing return where additional processors do not improve performance because memory access bandwidth is exhausted.

  - ❑ The Opteron processor has wider data paths than 32-bit processors (i.e., the number of bits that can be moved in a single operation) because they are designed to work with larger (64-bit versus 32-bit) word sizes. The Opteron processor's on-chip memory controller allows memory bandwidth to scale up with processors.

- Memory latency to the processor

  - ❑ Memory latency is the time that elapses after a processor requests data from memory until that data is available for use in the processor's registers. Because processor performance has continued to outpace improvements in memory latency over the last decade, processors end up spending more and more time just waiting for data. If this trend continues, further increases in processors speed will have relatively little effect on application performance and memory latency will be the limiting factor.

  - ❑ The Opteron processor addresses the latency problem with its on-chip memory controller and its buses that provided dedicated access to memory rather than shared access.

- Cache size

  - ❑ Caches are specialized memory systems either included on the processor chip or on specialized chips designed to run with the processor. Caches can be viewed as bandwidth-and-latency shock absorbers for memory systems. A large and efficient cache that prefetches data read from memory and buffers data to be written to memory can have a significant positive impact on system performance with workloads and applications.

  - ❑ Opteron has a 1MB on-chip L2 cache. Further, as noted, the Opteron processor includes an on-chip memory controller to accelerate access to memory. In addition, AMD doubled the number of registers available on the Opteron processor, thus providing more capacity for on-chip data manipulation.

- System configurability

  - Computer system architects can address memory performance issues with flexible system designs. For example, cell-based NUMA systems consisting of small processor/memory cells with fast local memory performance can be combined under different shared memory schemes. Doing so allows programmers to access all memory in all cells or to treat memory as distributed and use message passing protocols when data are shared between cells. Flexible approaches effectively increase memory bandwidth by partitioning bandwidth within cells so that it can be used more efficiently.

  - In the IBM eServer 325, one of the three HyperTransport links built into the Opteron processor provides dedicated interprocessor communication for two-way cells. Another HyperTransport link is available for whatever configuration of I/O that is needed. For example, the HyperTransport link connects to bridge chips to provide access to PCI-X slots, dual-integrated Gigabit Ethernet network interface cards (NICs), and other I/O interfaces.

SOFTWARE QUALITY AND PROCESSOR DESIGN

The efficiency of a system is also a factor of how much programming effort is required to get top performance on applications. Skillful programming and cunning algorithms can increase a processor's efficiency significantly. Conversely, even the best-designed system can be brought to its knees by poor quality software. The issue here is not whether quality can be achieved, but rather the tradeoff of costs in time and talent to build and maintain excellent programs. Issues related to processor design and software quality are as follows:

- **Architectural transparency.** Effective programming requires processor architectures that are readily understood by programmers, system software developers, and compiler writers. Architectures that are overly complex limit the number of people that are able or willing to learn the processor's idiosyncrasies. Complex processor instruction sets increase the time required to create efficient software tools and applications.

- **Architectural generality.** System architectures can be designed to work well on a small set of application types, and performance can be dramatically improved for these applications. However, high performance many not generalize to other application areas.

- **Availability of compilers and tools.** The availability of compilers, debuggers, profiles, and libraries can mitigate the challenges that programmers face and relax the skill set that programmers need to write efficient, high-quality software. The availability of such tools is determined both the architectural complexity and by the investments made by processor and system vendors.

- **Compatibility.** Compatibility with previous versions of a processor architecture reduces costs by taking advantage of previous work and skill sets.

OPPORTUNITIES AND CHALLENGES FOR THE NEW STANDARD TECHNOLOGIES

Products based on new standard technologies, such as the Opteron processor, create a number of opportunities both for technology providers and for those using the technology. At the same time, simply creating a new technology does not guarantee its success. Market entry can be difficult with the new standard being relegated to niche markets. Opportunities associated with new standards include:

- **Capturing market share.** The most basic and economically compelling driver for introducing a new product is to gain sales. In the case of standard technologies, these gains can occur from capturing market share from other technologies. This can be a friendly transition, as system vendors such as IBM are able to shift the research development and production costs to an outside organization.

- **Potential for market expansion.** More capable processors always hold the potential for both expanding existing markets by enabling more capable applications and by creating new markets where previous technologies could not address user requirements.

- **Ecosystem-level economies.** Successful standard technologies can encourage a supporting product ecosystem. In the case of processors, the ecosystem can create opportunities for application and middleware software suppliers, as well as service providers.

- **More competitive environment.** From a user perspective, standard technologies create a more competitive market when multiple suppliers develop hardware and software products based on the technology. Competition acts both to decrease prices and to increase product choices.

Challenges for new standard technologies include:

- **Obtaining critical mass.** The primary issue for the entry of new technologies is gaining sufficient market acceptance to allow for an adequate ROI and the creation of a healthy ecosystem. This is a general problem with technology transitions and can be made more difficult when a large number of options exist in the market.

- **Finding a workable price point.** Nonstandard server markets have been able to maintain higher prices for processors based on specialized capabilities. However, nonstandard server markets are relatively small. Standard processors offer a cost advantage based on production synergies, but most suppliers may not be able to cut prices when selling to markets with lower server volumes.

- **Getting it right.** New technology products need to work as advertised, have few defects, and be delivered in volume and on schedule. These issues are mitigated by the staying power of the processor vendor — big companies get more swings of the bat and, unlike Dizzy Dean, are allowed to stub their toes occasionally.

- **Product start-up issues.** Initial offerings of new technology products lack both a track record and a clear roadmap. Although this challenge is a case of the "you can't get a job without experience, and you can't get experience without a job" paradox, it presents important issues for new technologies.

MEETING THE CHALLENGES

To succeed with Opteron processor-based products, both IBM and AMD must first show that the initial systems are solid machines that are delivered on time, do not have any major technical problems, and perform as promised. Second, the companies must provide a clear roadmap for product upgrades and follow-on systems. Finally, they must provide assurance to the market that they committed to the technology over the long term.

IDC believes that the introduction of the eServer 325 is a good first step in meeting these challenges, and that IBM's use of the technology provides some assurances of long-term staying power. (IBM has historically prided itself on its commitment to its products and customers.) In addition, the Opteron processor's compatibility with other x86 architectures anchors the processor in a strong, existing technology ecosystem thus providing the potential for a transparent transition from 32-bit to 64-bit operations.

CONCLUSION

Over its long history, IBM has evolved from a company that provided all components of a system (more or less starting with silicon and ending with the people who installed systems at user sites and polished the CRT screens), to a company that has adapted to and leverages standard technology-based products. Although IBM has aggressively pursued standards-based products over the last several years, the company has continued to support and promote the proprietary product technology it has developed over the last four decades. Thus, IBM is in a relatively unique position of being able to cover all technology bets.

IBM's introduction of the eServer 325, its first server based on the AMD Opteron processor, is an indicator of continuing innovation and competition among suppliers of industry-standard processors. By positioning the eServer 325 as a cluster building block for high-performance computing, IBM continues its commitment to incorporate competing processors that provide different features and functions. Enabled by the Opteron processor, the eServer 325 provides a dual-mode alternative for customers with requirements for running legacy 32-bit programs along with new 64-bit programs. By incorporating chip-set functionality for I/O and memory management on the processor, the Opteron processor design addresses issues of performance balancing between processors, memory, and peripherals.

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.