

Planning Considerations for Multicore Processor Technology

The need to achieve higher performance without driving up power consumption and heat has become a critical concern for many IT organizations, given the density levels at which industry-standard servers are being deployed and the power and thermal constraints in today's data centers. Forthcoming multicore processor architectures will be designed to boost performance and minimize heat output by integrating two or more processor cores into a single processor socket. This article introduces the multicore concept and discusses key factors that IT organizations should consider when determining how best to take advantage of multicore technology.

BY JOHN FRUEHE

Related Categories:

Dell PowerEdge servers
Dell Precision workstations
Intel EM64T processors
Multiprocessor (MP)
Performance
Planning
Processors
Scalable enterprise
System architecture

Visit www.dell.com/powersolutions for the complete category index to all articles published in this issue.

Server density has grown dramatically over the past decade to keep pace with escalating performance requirements for enterprise applications. Ongoing progress in processor designs has enabled servers to continue delivering increased performance, which in turn helps fuel the powerful applications that support rapid business growth. However, increased performance incurs a corresponding increase in processor power consumption—and heat is a consequence of power use. As a result, administrators must determine not only how to supply large amounts of power to systems, but also how to contend with the large amounts of heat that these systems generate in the data center.

As more applications move from proprietary to standards-based systems, the performance demands on industry-standard servers are spiraling upward. Today, in place of midrange and large mainframe systems, tightly packed racks of stand-alone servers and blade servers can be clustered to handle the same types of business-critical

application loads that once required large proprietary systems. Organizations are using databases such as Microsoft® SQL Server, Oracle® Database 10g, and MySQL to enhance business decision making along with enterprise-wide messaging applications such as Microsoft Exchange. Meanwhile, network infrastructure, Internet connectivity, and e-commerce are growing at tremendous rates. Altogether, the result is a steady increase in performance demands as user loads and processing loads grow, driving a steady increase in the density of systems in the data center, which is intensified by ever-faster processors—and in turn this can create power and cooling challenges for many IT organizations.

Current options to address power and cooling challenges

Historically, processor manufacturers have responded to the demand for more processing power primarily by delivering faster processor speeds. However, the challenge

of managing power and cooling requirements for today's powerful processors has prompted a reevaluation of this approach to processor design. With heat rising incrementally faster than the rate at which signals move through the processor, known as clock speed, an increase in performance can create an even larger increase in heat.

IT organizations must therefore find ways to enhance the performance of databases, messaging applications, and other enterprise systems while contending with a corresponding increase in system power consumption and heat. Although faster processors are one way to improve server performance, other approaches can help boost performance without increasing clock speed and incurring an attendant increase in power consumption and heat. In fact, excellent overall processing performance may be achieved by *reducing* clock speed while increasing the number of processing units—and the consequent reduction in clock speed can lead to lower heat output and greater efficiency. For example, by moving from a single high-speed core, which generates a corresponding increase in heat, to multiple slower cores, which produce a corresponding reduction in heat, enterprises can potentially improve application performance while reducing their thermal output.

Balancing performance across each platform. The first step is to optimize performance across all platform elements. Designing, integrating, and building complete platforms that balance computing capabilities across processor, chip set, memory, and I/O components can significantly improve overall application performance and responsiveness. By integrating flexible technologies and balancing performance across all platform components, administrators can help provide the headroom required to support business growth (such as increases in users, transactions, and data) without having to upgrade the entire server. This approach can help the systems in place today support increased business demands, enhancing scalability for future growth. At the same time, this strategy can help extend the life of existing data center components by enabling administrators to optimize the performance of repurposed platforms when next-generation applications are deployed.

Harnessing multithreading technology. The second step is to improve the efficiency of computer platforms by harnessing the power of multithreading. Industry-standard servers with multiple processors have been available for many years, and the overwhelming majority of networked applications can take advantage of the additional processors, multiple software threads, and multitasked computing environments. These capabilities have enabled organizations to scale networked applications for greater performance. The next logical step for multiprocessing

advancements is expected to come in the form of multiple logical processing units, or *processor cores*, within a single chip. Multicore processors—coupled with advances in memory, I/O, and storage—can be designed to deliver a balanced platform that enables the requisite performance and scalability for future growth.

Optimizing software applications. The third step, software optimization, can be an efficient way to enable incremental performance gains without increasing power consumption and heat. Many of today's leading software tools, along with Intel® compilers, can enable significant performance improvements over applications that have not been compiled or tuned using such optimization tools.¹ Actual performance gains will depend on the specific system configuration and application environment. To get the most performance from existing data center components, administrators must not overlook potential gains from optimizing software applications during the infrastructure planning processes.

Scalability potential of multicore processors

Processors plug into the system board through a socket. Current technology allows for one processor socket to provide access to one logical core. But this approach is expected to change, enabling one processor socket to provide access to two, four, or more processor cores. Future processors will be designed to allow multiple processor cores to be contained inside a single processor module. For example, a tightly coupled set of dual processor cores could be designed to compute independently of each other—allowing applications to interact with the processor cores as two separate processors even though they share a single socket. This design would allow the OS to “thread” the application across the multiple processor cores and could help improve processing efficiency.

A multicore structure would also include cache modules. These modules could either be shared or independent. Actual implementations of multicore processors would vary depending on manufacturer and product development over time. Variations may include shared or independent cache modules, bus implementations, and additional threading capabilities such as Intel Hyper-Threading (HT) Technology.

A multicore arrangement that provides two or more low-clock-speed cores could be designed to provide excellent performance while minimizing power consumption and delivering lower heat output than configurations that rely on a single high-clock-speed core. The following example shows how multicore technology could manifest in a standard server configuration and how multiple

¹ For example, in January 2005 Intel conducted benchmark tests showing that 64-bit Intel Xeon processor technology with Intel Hyper-Threading Technology enabled can provide up to 33 percent improvement in application and server performance compared to the same configuration with Hyper-Threading Technology disabled. For more information about Hyper-Threading Technology performance tests, visit www.intel.com/performance/server/xeon/ht_perf.htm. Please note that Intel performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Administrators should consult other sources of information to evaluate the performance of specific systems or components. For more information about performance tests and the performance of Intel products, visit www.intel.com/performance/resources/benchmark_limitations.htm or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

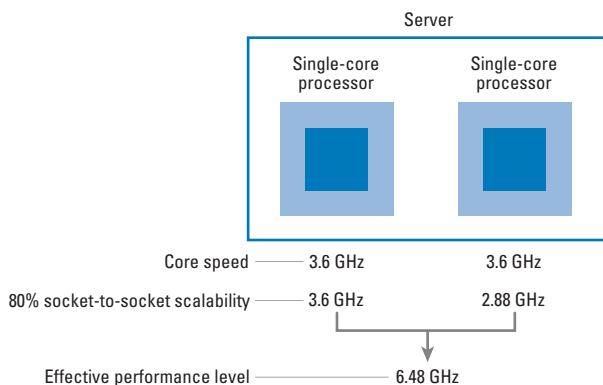


Figure 1. Sample core speed and anticipated total relative power in a system using two single-core processors

low-clock-speed cores could deliver greater performance than a single high-clock-speed core for networked applications.

This example uses some simple math and basic assumptions about the scaling of multiple processors and is included for demonstration purposes only. Until multicore processors are available, scaling and performance can only be estimated based on technical models. The example described in this article shows one possible method of addressing relative performance levels as the industry begins to move from platforms based on single-core processors to platforms based on multicore processors. Other methods are possible, and actual processor performance and processor scalability are tied to a variety of platform variables, including the specific configuration and application environment. Several factors can potentially affect the internal scalability of multiple cores, such as the system compiler as well as architectural considerations including memory, I/O, frontside bus (FSB), chip set, and so on.

For instance, enterprises can buy a dual-processor server today to run Microsoft Exchange and provide e-mail, calendaring, and messaging functions. Dual-processor servers are designed to deliver excellent price/performance for messaging applications. A typical configuration might use dual 3.6 GHz 64-bit Intel Xeon™ processors supporting HT Technology. In the future, organizations might deploy the same application on a similar server that instead uses a pair of dual-core processors at a clock speed lower than 3.6 GHz. The four cores in this example configuration might each run at 2.8 GHz. The following simple example can help explain the relative performance of a low-clock-speed, dual-core processor versus a high-clock-speed, dual-processor counterpart.

Dual-processor systems available today offer a scalability of roughly 80 percent for the second processor, depending on the OS, application, compiler, and other factors.² That means the first processor may deliver 100 percent of its processing power, but the second

processor typically suffers some overhead from multiprocessing activities. As a result, the two processors do not scale linearly—that is, a dual-processor system does not achieve a 200 percent performance increase over a single-processor system, but instead provides approximately 180 percent of the performance that a single-processor system provides. In this article, the single-core scalability factor is referred to as external, or *socket-to-socket*, scalability. When comparing two single-core processors in two individual sockets, the dual 3.6 GHz processors would result in an effective performance level of approximately 6.48 GHz (see Figure 1).

For multicore processors, administrators must take into account not only socket-to-socket scalability but also internal, or *core-to-core*, scalability—the scalability between multiple cores that reside within the same processor module. In this example, core-to-core scalability is estimated at 70 percent, meaning that the second core delivers 70 percent of its processing power. Thus, in the example system using 2.8 GHz dual-core processors, each dual-core processor would behave more like a 4.76 GHz processor when the performance of the two cores—2.8 GHz plus 1.96 GHz—is combined.

For demonstration purposes, this example assumes that, in a server that combines two such dual-core processors within the same system architecture, the socket-to-socket scalability of the two dual-core processors would be similar to that in a server containing two single-core processors—80 percent scalability. This would lead to an effective performance level of 8.57 GHz (see Figure 2).

To continue the example comparison by postulating that socket-to-socket scalability would be the same for these two architectures, a

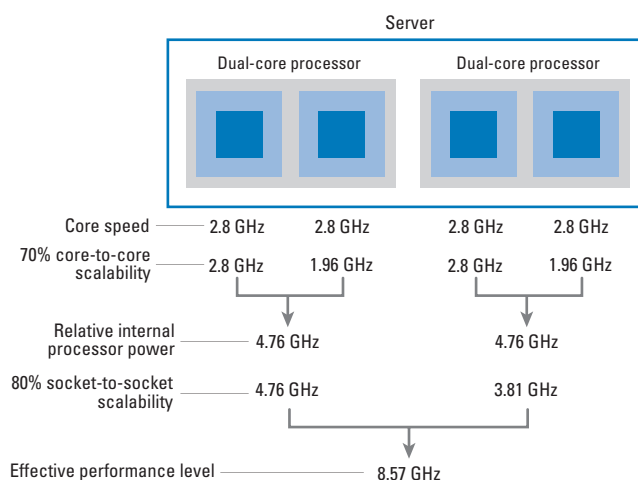


Figure 2. Sample core speed and anticipated total relative power in a system using two dual-core processors

²While 80 percent scalability for the second processor is a representative approximation for the example set forth in this article, even higher scalability has been achieved. For example, in January 2005 Intel conducted benchmark tests of 64-bit Intel Xeon processor scaling based on a two-processor configuration versus an otherwise comparable one-processor configuration. For more information, visit www.intel.com/performance/server/xeon/scaling.htm.

UNDERSTANDING HYPER-THREADING TECHNOLOGY

Today's 64-bit Intel Xeon, Pentium® 4, and Celeron® processors include HT Technology, which enables the processor to execute multiple threads of an application simultaneously. Multithreaded applications perceive a single physical processor as two separate, logical processors and will execute threads independently on each logical processor to help speed overall processing execution. Recent benchmark tests by Intel of 64-bit Intel Xeon processor-based platforms have shown a performance gain of up to 33 percent by enabling HT Technology on applications that are HT Technology-aware as compared to running the same applications with HT Technology disabled.*

Today, individual Intel NetBurst® microprocessors appear to the OS as two logical processors. On a dual-processor system supporting HT Technology, the application perceives four processor threads (two physical processors and two logical processors). Equipped with multicore processors, that same dual-socket system could have a total of four processor

cores. Through the effective use of HT Technology, those four processor cores could appear to the application as eight total processors.

By leveraging HT Technology, a properly compiled application can achieve performance increases because of the improved utilization of the existing system processors, compared to the same application running with HT Technology disabled. Most multiprocessor-aware applications can take advantage of HT Technology, and applications that have been specifically designed for HT Technology have the potential to achieve a significant performance increase.

By combining multicore processors with HT Technology, Intel aims to provide greater scalability and better utilization of processing cycles within the server than is possible using single-core processors with HT Technology. The addition of HT Technology to multicore processor architecture could present an excellent opportunity to help improve the utilization and scalability of future processor subsystems.

* For more information about Intel HT Technology performance tests, visit www.intel.com/performance/server/xeon/ht_perf.htm. Please note that Intel performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Administrators should consult other sources of information to evaluate the performance of specific systems or components. For more information about performance tests and the performance of Intel products, visit www.intel.com/performance/resources/benchmark_limitations.htm or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

multicore architecture could enable greater performance than a single-core processor architecture, even if the processor cores in the multicore architecture are running at a lower clock speed than the processor cores in the single-core architecture. In this way, a multicore architecture has the potential to deliver higher performance than a single-core architecture for enterprise applications.

Power and cooling advantages of multicore processors

Although the preceding example explains the scalability potential of multicore processors, scalability is only part of the challenge for IT organizations. High server density in the data center can create significant power consumption and cooling requirements. A multicore architecture can help alleviate the environmental challenges created by high-clock-speed, single-core processors.

Heat is a function of several factors, two of which are processor density and clock speed. Other drivers include cache size and the size of the core itself. In traditional architectures, heat generated by each new generation of processors has increased at a greater rate than clock speed.

In contrast, by using a shared cache (rather than separate dedicated caches for each processor core) and low-clock-speed processors, multicore processors may help administrators minimize heat while maintaining high overall performance. This capability may help make future multicore processors attractive

for IT deployments in which density is a key factor, such as high-performance computing (HPC) clusters, Web farms, and large clustered applications. Environments in which 1U servers or blade servers are being deployed today could be enhanced by potential power savings and potential heat reductions from multicore processors.

Currently, technologies such as demand-based switching (DBS) are beginning to enter the mainstream, helping organizations reduce the utility power and cooling costs of computing. DBS allows a processor to reduce power consumption (by lowering frequency and voltage) during periods of low computing demand. In addition to potential performance advances, multicore designs also hold great promise for reducing the power and cooling costs of computing, given DBS technology. DBS is available in single-core processors today, and its inclusion in multicore processors may add capabilities for managing power consumption and, ultimately, heat output. This potential utility cost savings could help accelerate the movement from proprietary platforms to energy-efficient industry-standard platforms.

Significance of sockets in a multicore architecture

As they become available, multicore processors will require IT organizations to consider system architectures for industry-standard servers from a different perspective. For example, administrators currently segregate applications into single-processor, dual-processor, and

quad-processor classes. However, multicore processors will call for a new mind-set that considers processor cores as well as sockets.

Single-threaded applications that perform best today in a single-processor environment will likely continue to be deployed on single-processor, single-core system architectures. For single-threaded applications, which cannot make use of multiple processors in a system, moving to a multiprocessor, multicore architecture may not necessarily enhance performance. Most of today's leading operating systems, including Microsoft Windows Server System™ and Linux® variants, are multithreaded, so multiple single-threaded applications can run on a multicore architecture even though they are not inherently multithreaded. However, for multithreaded applications that are currently deployed on single-processor architectures because of cost constraints, moving to a single-processor, dual-core architecture has the potential to offer performance benefits while helping to keep costs low.

For the bulk of the network infrastructure and business applications that organizations run today on dual-processor servers, the computing landscape is expected to change over time. However, while it may initially seem that applications running on a dual-processor, single-core system architecture can migrate to a single-processor, dual-core system architecture as a cost-saving initiative, this is not necessarily the case. To maintain equivalent performance or achieve a greater level of performance, the dual-processor applications of today will likely have to migrate to dual-socket, dual-core systems. As postulated in the Figure 1 example, a system architecture consisting of four processor cores in two sockets can be designed to deliver superior performance relative to a dual-socket, single-core system architecture, while also delivering potential power and cooling savings to the data center. The potential to gradually migrate a large number of older dual-socket, single-core servers to energy-efficient dual-socket, multicore systems could enable significant savings in power and cooling costs over time. Because higher-powered, dual-socket systems typically run applications that are more mission-critical than those running on less-powerful, single-processor systems, organizations may continue to expect more availability, scalability, and performance features to be designed for dual-socket systems relative to single-socket systems—just as they do today.

For applications running today on high-performing quad-processor systems, a transition to multicore technology is not necessarily an opportunity to move from four-socket, four-core systems to dual-socket, four-core systems. Rather, the architectural change suggests that today's four-processor applications may migrate to four-socket systems with eight or potentially more processor cores—helping to extend the range of cost-effective, industry-standard alternatives to large, proprietary symmetric multiprocessing (SMP) systems. Because quad-processor systems tend to run more mission-critical applications in the data center as compared

to dual-processor systems and single-processor systems, administrators can expect quad-processor platforms to be designed with the widest range of performance, availability, and scalability features across Dell™ PowerEdge™ server offerings.

When comparing relative processing performance of one generation of servers to the next, a direct comparison should not focus on the number of processor cores but rather on the number of sockets. However, the most effective comparison is ultimately not one of processors or sockets alone, but a thorough comparison of the entire platform—including scalability, availability, memory, I/O, and other features. By considering the entire platform and all the computing components that participate in it, organizations can best match a platform to their specific application and business needs.

Evolution of software toward multicore technology

Multicore processing continues to exert a significant impact on software evolution. Before the advent of multicore processor technology, both SMP systems and HT Technology motivated many OS and application vendors to design software that could take advantage of multithreading capabilities. As multicore processor-based systems enter the mainstream and evolve, it is likely that OS and application vendors will optimize their offerings for multicore architectures, resulting in potential performance increases over time through enhanced software efficiency.

Most application vendors will likely continue to develop on industry-standard processor platforms, considering the power, flexibility, and huge installed base of these systems. Currently, 64-bit Intel Xeon processors have the capability to run both 32-bit applications and 64-bit applications through the use of Intel Extended Memory 64 Technology (EM64T). The industry is gradually making the transition from a 32-bit standard to a 64-bit standard, and similarly, software can be expected to make the transition to take advantage of multicore processors over time.

Applications that are designed for a multiprocessor or multithreaded environment can currently take advantage of multicore processor architectures. However, as software becomes optimized for multicore processors, organizations can expect to see overall application performance enhancements deriving from software innovations that take advantage of multicore-processor-based system architecture instead of increased clock speed.

Although faster processors are one way to improve server performance, other approaches can help boost performance without increasing clock speed and incurring an attendant increase in power consumption and heat.

In addition, compilers and application development tools will likely become available to optimize software code for multicore processors, enabling long-term optimization and enhanced efficiency for multicore processors—which also may help realize performance improvements through highly tuned software design rather than a brute-force increase in clock speed. Intel is working toward introducing software tools and compilers to help optimize threading performance for both single-core and multicore architectures. Organizations that begin to optimize their software today for multicore system architecture may gain significant business advantages as these systems become mainstream over the next few years. For instance, today's dual Intel Xeon processor-based system with HT Technology can support four concurrent threads (two per processor). With the advent of dual-core Intel Xeon processors with HT Technology, these four threads would double to eight. An OS would then have eight concurrent threads to distribute and manage workloads, leading to potential performance increases in processor utilization and processing efficiency.

Licensing considerations

Another key area to consider in planning for a migration to multicore processors is the way in which software vendors license their applications. Many enterprise application vendors license their applications based on the number of processors, not the number of users. This could mean that, although a dual-socket, dual-core server may offer enhanced performance when compared to a dual-socket, single-core server, the licensing cost could potentially double because the application would identify four processors instead of two. The resulting increase in licensing costs could negate the potential performance improvement of using multicore processor-based systems. Because the scalability of multicore processors is not linear—that is, adding a second core does not result in a 100 percent increase in performance—a doubling of licensing costs would result in lower overall price/performance.

For that reason, software licensing should be considered a key factor when organizations assess which applications to migrate to systems using multicore processors. For example, enterprise software licensing costs can be significantly higher than the cost of the server on which the application is running. This can be especially true for industry-standard servers that deliver excellent performance at a low price point as compared to proprietary servers. Some application vendors have adopted a policy of licensing based on the socket count instead of the number of cores, while others have not yet taken a stance on this matter. Until the industry gains more clarity around this software licensing issue, organizations must factor software licensing costs into the overall platform cost when evaluating multicore technology transitions.

Dell multicore processor plans

Dell plans to begin integrating multicore processor designs into Dell PowerEdge and Dell PowerEdge SC servers during the next 12 to 18 months as multicore processors become available. Through Dell's close relationship with Intel, Dell intends to deliver solutions built on Intel Xeon, Pentium 4, and Itanium® multicore processors. Dell is working closely with Intel to ensure that the next generations of PowerEdge and PowerEdge SC servers are designed to meet both the performance and scalability needs of enterprises in relation to multicore processor architectures.

Shift in focus toward multicore technology

Multicore processors most likely represent the future direction of server architecture, which is expected to enhance application performance and platform power with thermal efficiency. By combining multiple logical processing units within a single processor package as described in this article, multicore processor architectures have the potential to provide superior performance and scalability without a corresponding increase in power consumption and heat, as would be the case by simply increasing the clock speed of existing single-core processor designs.

In readiness for this impending change in processor architecture, system vendors like Dell will begin to address system design in a different manner when determining system performance and scalability. What was once a focus on processor count will become a focus on socket count as the shift occurs in the number of processor cores per socket. However, before adopting this emerging system architecture, IT organizations need to carefully evaluate the software ramifications of migrating applications to multicore processor technology. This consideration can enable enterprises to benefit from the higher performance and lower power consumption expected of multicore processor architecture as compared to single-core processor architecture, while helping ensure that multicore processor platforms are licensed appropriately to control acquisition costs. ➤

John Fruehe is a marketing strategist for the Dell Enterprise Product Group. He has worked at Dell for nine years. Prior to that, John was at Compaq and Zenith Data Systems. John has a B.S. in Economics from Illinois State University and has been in the technology field for 14 years.

FOR MORE INFORMATION

Multicore processor architecture:

www.intel.com/cd/ids/developer/asm-na/eng/201969.htm?page=6

Dual-core and HT Technology:

www.intel.com/cd/ids/developer/asm-na/eng/technologies/threading/199701.htm