IT@Intel Brief

Intel Information Technology

Computer Manufacturing

Server and Blade Platforms

July 2006

Server Rightsizing: Dual-socket Systems Cut Costs

By establishing a policy of using two-socket servers—except for large scale-up applications—Intel has saved more than USD 8 million in capital costs, power and cooling, and data center infrastructure.

Intel operates a worldwide computing environment with about 8,000 servers; buying the most cost-effective servers translates into major savings. Due to rapidly increasing processor power, we found that dual-socket servers can run the vast majority of business applications, including many previously handled by four-socket servers costing three to four times as much.

Profile: Server Rightsizing

- 2-socket performance matches 4-socket in 12 to 24 months
- Dual-socket servers handle up to 95% of business workloads
- USD 8 million in savings and cost avoidance over 3 years

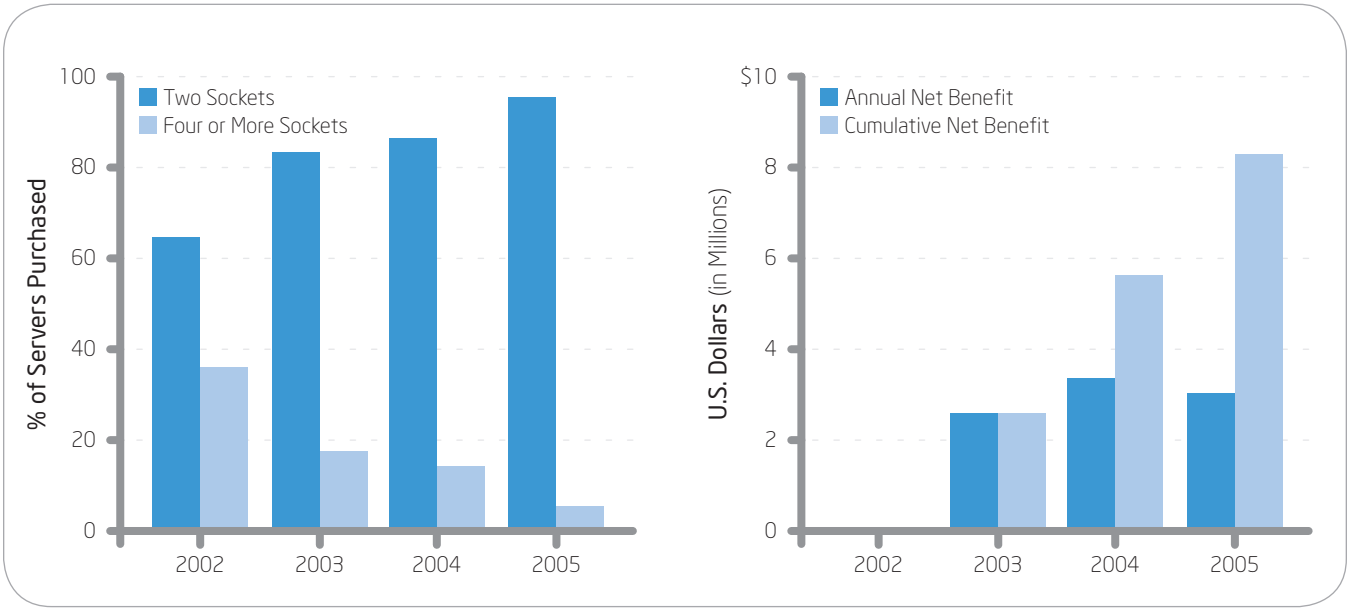

Figure 1. The shift to dual-socket servers provides significant financial benefit to Intel.

IT@Intel

Background

Until 2003, about a third of the servers we purchased were high-end machines with four or more processor sockets. These large systems were historically needed to run many typical business applications, or were thought necessary to provide headroom for expected workload growth.

Since we continually invest in server technology, we regularly analyze server workloads to determine the most cost-effective configurations for our needs. In late 2003, our analysis led to a significant policy change and ongoing analysis continues to support that policy.

Server Workload Analysis

We use several types of data to understand the size and growth of server workloads, including:

- Server utilization data from production systems running various business applications including e-mail, file and print services, directory services, and enterprise resource planning (ERP)
- Historical trends in workload growth collected by the owners of different applications
- Published benchmark data for different server classes

We have found that most business and productivity application workloads consistently grow at modest rates. ERP applications, such as order management, generally grow 10 to 20 percent a year, as shown in Figure 2. E-mail growth, barring filtered spam, is similar. Exceptions to this trend are some newer workloads, such as business intelligence and XML processing, which can grow more quickly.

We have also found that server utilization across our business computing environment remains generally low. The vast majority of servers use less than 20 percent of CPU capacity, on average, during the business day, as shown on the next page in Figure 3. Utilization is generally moderate even at peak times. Like many organizations, Intel historically has dedicated each server to a single application for reliability, manageability, flexibility, and other reasons.

Solutions

We realized that by taking advantage of Moore's law, we could make substantial savings. Relentless increases in processor power mean that the performance provided by high-end four-socket servers at any given time is matched by lower-cost dual-socket machines

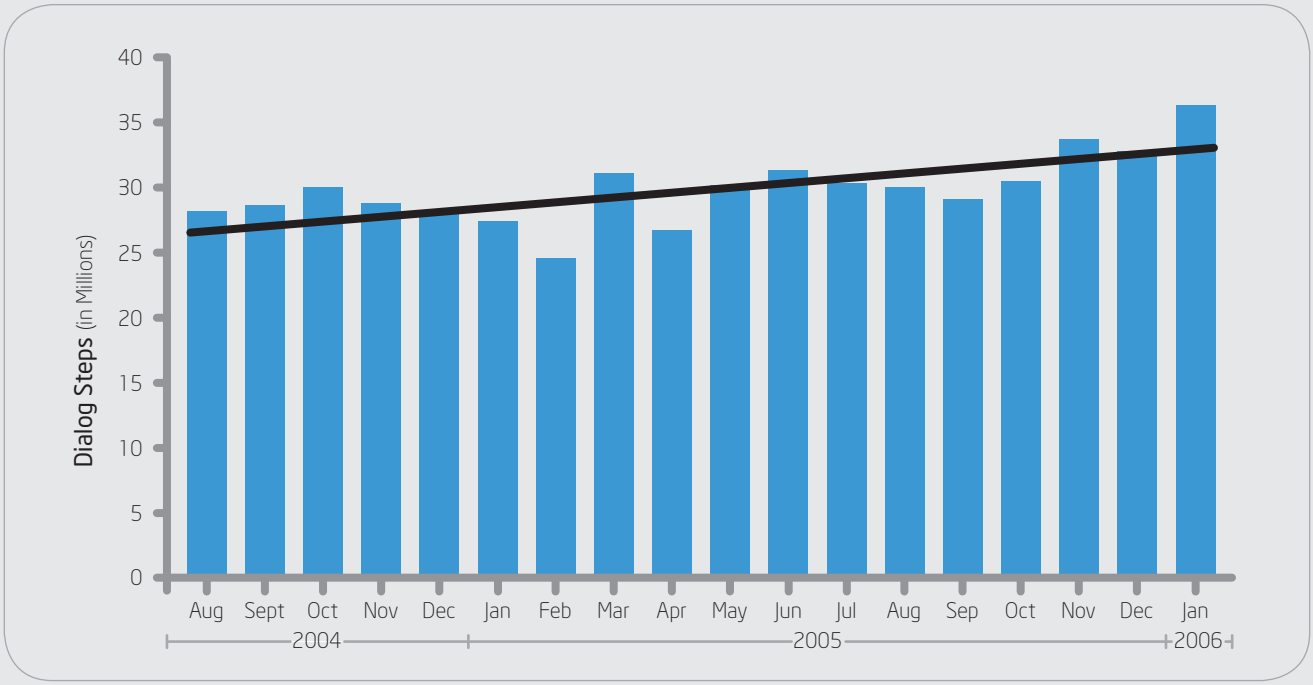

Figure 2. Most Intel business workloads grow at modest rates. Intel order management server workload grows about 10 to 20 percent each year.

12 to 24 months later, as shown in Figure 4. These performance increases easily outstrip the growth of most business applications. Because of our low average server utilization and modest workload growth, dual-socket machines are adequate for most applications, including many workloads that were previously handled by four-socket machines.

To take advantage of these trends in server performance, we established a server rightsizing policy that made dual-socket servers the standard for business computing workloads across Intel. Our engineering computing group uses a similar approach for most applications.

When authorizing server purchases, senior managers grant exceptions to this policy only when benchmarks, pilots, or other factors clearly justify a different class of server.

In practice, we have found that we can comfortably accommodate up to 95 percent of workloads with dual-socket servers. These include our small to medium application workloads.

Dual-socket servers also efficiently handle large applications that are suited to a scale-out approach. These include e-mail, which can be partitioned between several servers, and web servers that can be dynamically balanced across multiple machines. Other workloads that suit this approach are directory services, security servers, and other IT infrastructure services. Dual-socket machines have proved to be extremely effective for server consolidation, offering better price-performance and increased density compared with four-socket machines.

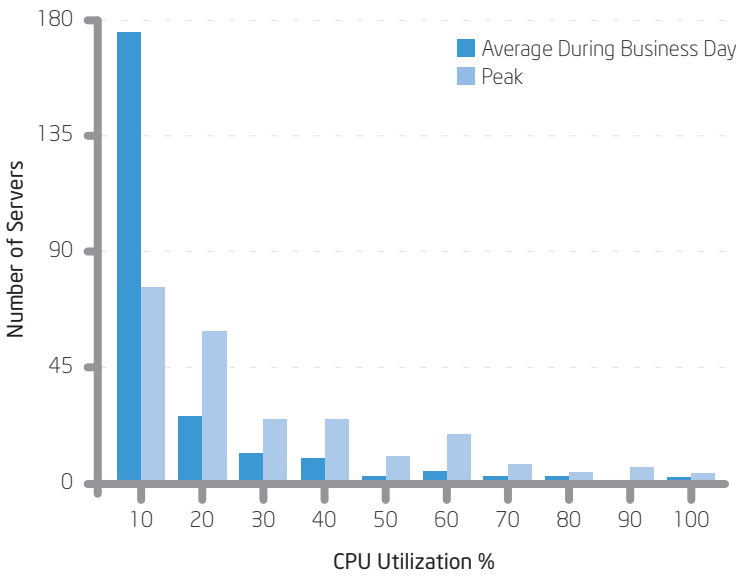

Figure 3. Business computing server utilization is generally moderate.

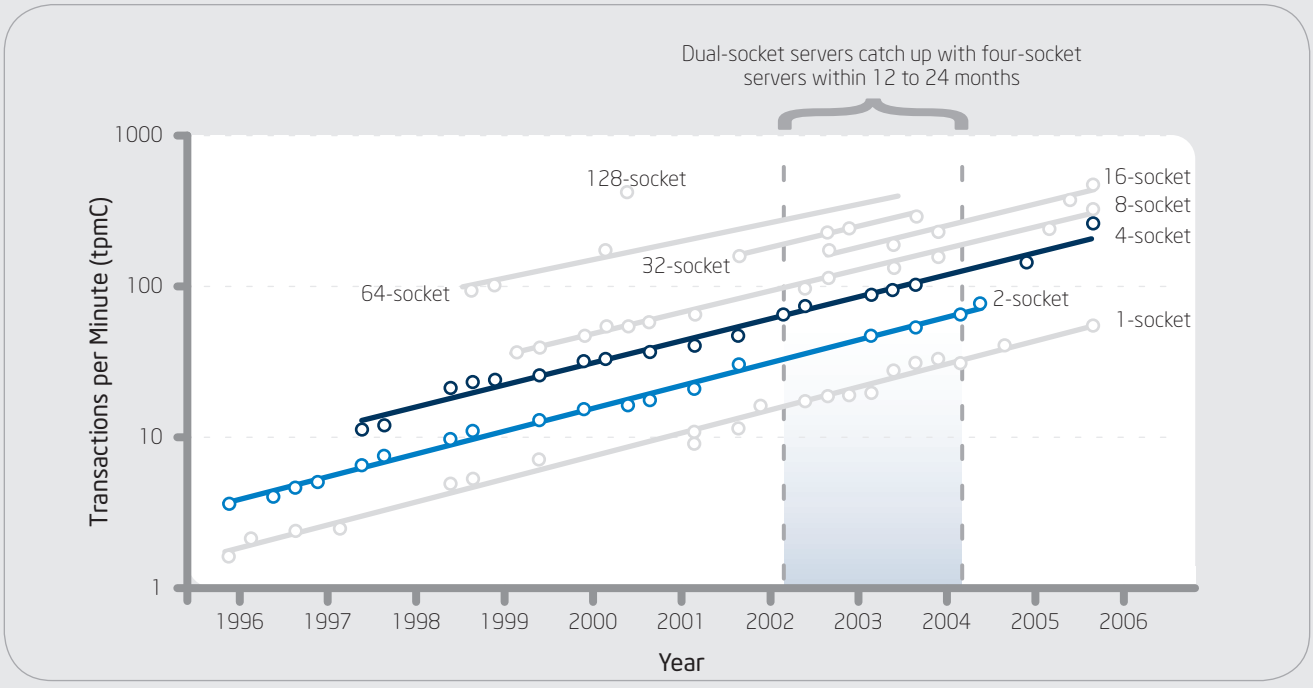

Figure 4. Server performance grows exponentially. Best published transactions per minute (tpmC) for Intel® Xeon® processor-based servers in the past decade.

We make exceptions for large "scale-up" applications that we cannot easily divide among multiple servers and that may need considerable headroom for growth. Examples include large human resources, data warehouse, ERP, and financial applications. For these, we may purchase Intel® Xeon® processor MP or Itanium®-based servers with four or more sockets, based on detailed workload characterization and testing.

Results

The percentage of dual-socket servers we purchase for business computing has increased each year since we established the rightsizing policy. We estimate that this has saved more than USD 8 million between 2003 and 2005. A significant part of this is due to lower capital costs: Four-socket machines are typically three to four times as expensive as dual-socket systems. However, there are also other savings, as shown in Figure 5. Dual-socket servers require less power and generate less heat, so we are spending less on power and cooling. Lower power consumption also means that we need less data center infrastructure, which creates more savings through cost avoidance. As an added benefit, dual-socket servers are smaller and occupy less rack space.

Dual-socket servers are becoming even more capable as processor technology evolves to multi-core chips. Current configurations use two Dual-Core Intel Xeon processors, replacing the two single-core processors in some earlier dual-socket systems.

Because of the additional compute power provided by these new systems, Intel IT is extending its use of dual-socket servers. We have standardized on dual-core dual-socket systems for server consolidation, using virtualization to run multiple applications on a single server, because of the better price-performance and increased compute density that can be achieved with these systems compared with four-socket servers.

In addition, our manufacturing organization plans to adopt dual-socket machines as mid-range servers for uses that previously required four-socket systems. As processor technology advances, dual-socket servers may take on an even broader set of tasks at Intel.

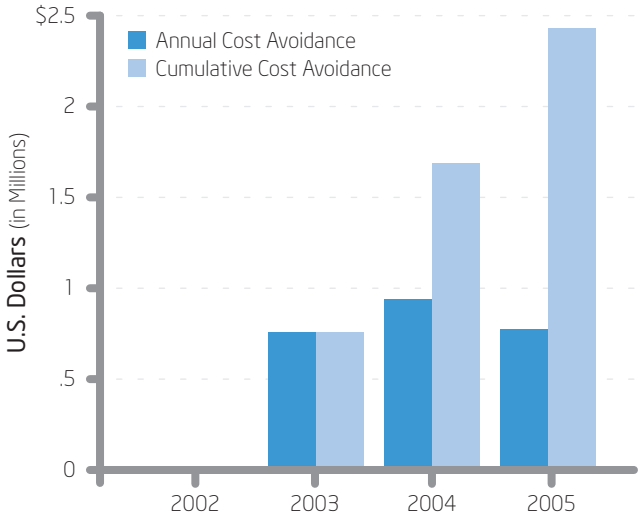

Figure 5. Data center cost avoidance due to dual-socket server policy.

Acronyms

ERP	enterprise resource planning
tpmC	transactions per minute

Printed in USA
0706/ARM/RDA/PDF

♻ Please Recycle
Order Number: 314624-001US