The Advantages of Intel® Itanium™ Architecture for Cache Server Software

Information for IT Managers
and System Integrators

White Paper

The Internet provides an optimum e-Commerce environment as it is the first medium to enable automation of both marketing-on-demand and sales-on-demand functions without requiring human intervention for the customer. In general, a successful e-Commerce system must have three primary characteristics: scalability, availability and responsiveness. Instituting cache technology provides a quick way to significantly improve the performance of server applications, increase response time and throughput, and reduce bottlenecks. There are three main types of caching technology: Web server, proxy server and LAN cache file servers.

Web Server Caching

Web servers can become a bottleneck in either intranet or Internet infrastructures. Web servers that must handle heavy traffic—whether a persistently large stream of visitors or a sudden spike in usage—tend to quickly bog down. A great number of incoming requests create large numbers of threads or processes, which must all contend for the same system resources. The resultant reduced throughput and slow response times decrease user satisfaction, potentially causing customers to leave—a death knell for e-Business Web sites.

To avoid this problem, many companies and ISPs have improved network bandwidth and accelerated Web server response times with Web-caching technology. Web caches store static Web pages and the data used to generate dynamic Web pages. For highest performance, Web caches use physical memory to store frequently used pages. The processor sends information straight from memory to the network, thus avoiding the need for disk access. This technique dramatically improves server response times.

Environments with a large amount of content may need to configure Web servers with more than 3–4 GB of cache. Typically, configuring systems with large memory caches will allow the servers to store much of Web site in memory, avoiding the performance penalty of going to disk. Such large physical memory caches can be installed on each Web server, or the cache can form a separate tier of "cache servers" in front of the Web servers. Popular sites such as eToys.com* use such caching techniques to improve the number of hits they can accept and to increase throughput. Even less active commercial sites with smaller amounts of content still increase response times with this caching technique.

Proxy Server Caching

The explosion of e-Business over the last five years has sparked a voracious increase in the demand for bandwidth. Due to the high price of leased lines, many corporate sites have set up Web server proxy caches to reduce external bandwidth requirements and to tighten security.

Proxy servers sit on the border of the LAN and the Internet to keep copies of the most recently accessed or most frequently accessed Web pages. Then, corporate desktops fetch pages from this local cache server on the LAN to avoid a remote network bottleneck and to improve response time.

LAN Cache Fileservers

Overall, LAN environments have not adopted memory caching technology en masse because many of these environments retain a strong need to ensure that their file server data gets written to disk. For environments with a strong percentage of writes to the file server, caching technology does not make sense. However, for environments running applications off the server, rather than off the desktop hard drive, caching at the file server can greatly improve response time for starting up applications.

Intel® Itanium™ Processor Benefits for Caching

The Intel® Itanium™ architecture provides a rich set of functionality that should enable larger, more powerful cache server software. To be more specific, features and benefits enabled by the Intel® Itanium™ processor include:

- **Large 64-bit address space:** Traditional 32-bit architectures only support up to 4 GB of physical memory; often, 2 GB of that is reserved for the typical system kernel.

  The Intel Itanium processor supports a large, flat address space and physical memory addressability that will increase application performance for caching techniques.

- **Increased instruction parallelism:** Intel Itanium compilers package instructions in bundles (three instructions per

bundle) that can be executed in parallel; the Intel® Itanium™ processor provides multiple execution units to execute these instructions concurrently. Increased instructions per clock will benefit a broad range of software routines.

Cache server environments will benefit from parallelism, especially for improving throughput and increasing response times.

- Very large set of registers: For about a decade, the IA-32 architecture has relied on eight registers (temporarily holding places for variables used by an application). Code with heavy data manipulation routines often require more than eight short-term holding places to avoid excessive and unproductive loads and stores. This overhead moves and stores between registers and memory increases the likelihood for processor stalls and reduces performance.

  The Intel Itanium processor provides 128 integer, 128 floating-point and eight branch registers as well as an innovative rotating register model. This increased register set enables cache server applications to rapidly shift between threads and enhance the overall performance to the end user.

- Speculation: Microprocessor clock rates have historically grown much faster than the speed and latency of system memory, leading to a variety of branch prediction and caching techniques. Many competing RISC microprocessors can pre-fetch only after the last code branch, but the Intel Itanium processor does not impose this limitation. The Intel Itanium processor's speculative loads take this one step further, allowing the compiler to schedule pre-fetching speculative loads from memory well in advance of the need for the data, thus removing the latency of the load operations and reducing processor stalls.

  Speculation reduces memory latency and increases performance in cache server applications by allowing the system to retrieve data so it is ready when the processor is ready to use it.

- Predication: Predication is the conditional execution of instructions, which allows code to avoid using branch or "if" statements. Instead, the chip executes both paths of the branch at the same time as the "if" statement is being run and then discards the unwanted path once the results of the "if" statement have been determined. Like speculation, predication pays off for code with heavy branching, but unlike speculation, predication focuses on the process of executing parallel code paths rather than on speeding memory access.

  Portions of caching Web server code experience heavy branching and "if" statements. Predication allows the microprocessor to run both paths of a branch in parallel, avoiding unnecessary processor stalls while computing the results of an "if" statement.

The table below breaks down the unique Intel Itanium architectural features for specific tasks performed by cache server software.

Table 1: Intel® Itanium™ Processor Benefits for Cache Server Tasks

Predication (cond exec)	Speculation (spec load)	Increased Number of Registers	Parallelism (incr. IPC)	Types of Caching Technology
X	X	XX	X	**Web Server Caching:** Represents the primary application in today's Internet-centric business environment; this type of caching stores static Web pages and/or data for dynamic Web pages in physical memory, avoiding relying on slower disk access and therefore increasing throughput
X	X	X	X	**Proxy Caching:** Proxy caches are traditionally distributed to the edges of the network, e.g., a corporate gateway, to improve delivery of the multitude of Internet documents to the group of users connected to that cache; the cache stores and delivers the most frequently accessed content from the millions of documents on the Web, thus providing better quality of service for the end user and saving on the cost of bandwidth necessary to retrieve these documents from the origin server; proxy caching reduces external bandwidth requirements to tighten security
X	X	X	–	**LAN Cache File Servers:** For environments running applications off the server, as opposed to off the desktop hard drive, caching at the file server can greatly improve response time for booting applications

X = a task that will likely benefit from faster performance due to a specific Intel Itanium optimization feature.
– = areas that are unlikely to contain code that will optimize with a particular Intel Itanium feature.

Summary

The Intel® Itanium™ processor is ideal for all caching servers, including Web caching, proxy caching and LAN caching servers. There is a need for headroom and scalability in servers, since today's servers are required to handle unpredictable loads of e-Business solutions, fast processing of Web page requests and large physical memory support. Increased server availability and reliability due to parity protection, ECC protection and enhanced memory check architecture make the Intel Itanium processor and future IA-64 implementations an ideal platform for e-Business. Software developers porting their applications to the Intel Itanium processor and IT Managers integrating Itanium-based servers into their current environments will quickly increase their competitiveness in the marketplace by delivering fast response times and immediate service.

Printed in USA
order number: 283583-001
0101/CMD/JH/PDF