Evaluating Internet servers

White paper: Performance tests rate AlphaServer systems over Sun Netra i20
<table>
<thead>
<tr>
<th>Page</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Highlights</td>
</tr>
<tr>
<td>4</td>
<td>Executive summary</td>
</tr>
<tr>
<td>5</td>
<td>The Internet’s World Wide Web: mushrooming “hot spot” for doing business</td>
</tr>
<tr>
<td>5</td>
<td>A robust Internet server: business-critical for customer service</td>
</tr>
<tr>
<td>6</td>
<td>Evaluating Internet servers</td>
</tr>
<tr>
<td>6</td>
<td>Performance comparison tests: Digital AlphaServer 1000 4/266 and 400 4/233 vs. Sun Netra i20</td>
</tr>
<tr>
<td>7</td>
<td>Critical factors in evaluating Internet server capabilities</td>
</tr>
<tr>
<td>7</td>
<td>The future: interactivity, multimedia, secure transactions</td>
</tr>
<tr>
<td>7</td>
<td>Frequently asked questions</td>
</tr>
<tr>
<td>9</td>
<td>Glossary</td>
</tr>
</tbody>
</table>
In this white paper, you will read about:
• The meteoric growth of the Internet’s World Wide Web as today’s “hot spot” for doing business.
• The importance of having the most robust Internet server possible.
• Recent performance testing in which Digital’s AlphaServer 1000 4/266 and AlphaServer 400 4/233 surpassed Sun Microsystems’ Netra i20 in performance and price/performance comparisons.
  - Digital’s uniprocessor AlphaServer systems served 2x to 1.5x the number of files accessed by Sun’s more expensive dual-processor Netra i20.
  - The AlphaServer 400 4/233 achieved a maximum of 109,680 requests per hour without reaching saturation, and yielded an average service time of 1.1 seconds. The Netra i20 handled a maximum of 66,220 requests with an average service time of 1.2 seconds.
  - The maximum number of requests processed by the AlphaServer 400 was 66 percent greater than the number achieved by the Netra i20, yet the AlphaServer 400 achieved 12 percent faster service times.
  - The AlphaServer 1000 4/266 with 512 MB of memory successfully serviced a maximum of 136,052 requests per hour with an average service time of 0.6 seconds — twice as fast as the Netra i20.
Executive summary

The commercial potential of the Internet is evident everywhere today. The World Wide Web is doubling in size every few months, making it the fastest-growing service on the global Internet, which itself is doubling in size each year. The potential for substantial revenue generation is commensurate.

Offering text, images, and (increasingly) video and audio from tens of thousands of commercially maintained databases, the Web gives companies the chance to reach customers graphically and immediately. From general-interest consumers who browse electronic “shopping malls” to specialized customers who, for example, can download a vendor’s white paper on a forthcoming technology product, the Web is today’s “hot spot” for doing business.

Companies who put up a Web page naturally count on the consumer appeal of this opening salvo to establish an interested customer base that will make repeat visits to their site on the Internet.

Whether they realize it or not, they are also depending on the robustness of their Internet server to do the job of answering the thousands of information requests that can typically “spike” during the initial launch of an appealing new site or new offering — or with any unexpected surge in user demand.

Business-critical point of contact

The business-critical importance of having a robust Internet server cannot be emphasized too strongly. Customer interest and good will can be killed permanently when prospects see the inscrutable “host unavailable” message on their client screens instead of the Web page they were led to expect.

Worse, dropped connections, which result in the frustrating “host unavailable” message, are never picked up again, resulting in dissatisfaction and erosion of the very customer base that was to be established. Worst of all, the Web site never knows the connections have been dropped.

It thus becomes highly critical to “survive the spike” — the unpredictable surge of stress on the server when a Web site is swamped with file requests.

Evaluating Internet server performance

Enterprises must carefully weigh performance characteristics before choosing an Internet server with which to reach their customers.

But the criteria offered by vendors can be elastic, making comparison difficult. When a user requests a “page,” this is commonly thought to mean a single file. Yet a Web page can, and usually does, consist of multiple files. And how many “pages” does a user download during a “visit?” There is no rule of thumb to correlate these terms. Further, the size of the working set within the information hierarchy can drastically change results because of caching effectiveness.

Clear metrics are therefore necessary in setting performance test conditions. Service time is a function of both the rate at which the requests arrive, and the size of the actual file requested.

Evaluating Digital’s Internet AlphaServer systems

Such conditions were set during recent benchmark performance tests by the Application Systems Engineering Performance Group at Digital’s Equipment Performance Laboratory.

In these tests, Digital’s AlphaServer 1000 4/266 and AlphaServer 400 4/233 surpassed Sun Microsystems’ Netra i20 in performance and price/performance comparisons.

Digital’s uniprocessor AlphaServer systems served 2x to 1.5x the number of files accessed by Sun’s more expensive dual-processor Netra i20.

Performance testing was designed to saturate the CPU with requests while eliminating as far as possible any other system components from becoming a binding resource.

The AlphaServer 400 4/233 achieved a maximum of 109,680 requests per hour without reaching saturation, and yielded an average service time of 1.1 seconds. The Netra i20 handled a maximum of 66,188 requests with an average service time of 1.2 seconds.

The maximum number of requests processed by the AlphaServer 400 was 66 percent greater than the number achieved by the Netra i20, yet the AlphaServer 400 achieved 12 percent faster service times.

At a higher-capacity level, the AlphaServer 1000 4/266 with 512 MB of memory successfully serviced a maximum of 132,052 requests per hour with an average service time of 0.6 seconds — twice as fast as the Netra i20.

For test details, please refer to Section 4.

The future: interactivity, multimedia, secure transactions

As Internet commerce continues its rapid expansion and becomes increasingly content-rich, Internet servers will be required to handle more demanding processing tasks. The typical size of requested files will become significantly larger, making Internet server robustness and reliability still more business-critical. In addition, servers will increasingly have to perform interactive and transactional functions (such as handling secure transactions) that place a significant drain on CPU cycles. Digital’s family of Internet AlphaServer systems offer the ability not only to survive the “spike,” but to continue satisfying customers and prospects using the fast-growing World Wide Web.
1. The Internet’s World Wide Web: mushrooming “hot spot” for doing business

The global Internet is doubling in size each year. On the Net today, the fastest-growing service is the World Wide Web, which itself is doubling in size every few months.

The Web — a network of tens of thousands of linked sites — allows enterprises to publish electronic documents containing pictures and sound as well as text, more and more of them user-interactive. Special browsing software enables Web users to move from document to document by pointing and clicking.

Maintaining a site or database on the Web gives a commercial enterprise a unique “presence,” an opportunity to reach customers graphically and immediately. From financial investors who access immediate information and perform instant transactions over the Web to general-interest consumers who browse electronic “shopping malls,” the Web is today’s “hot spot” for doing business.

Operating an Internet server is the basis for maintaining a Web “page” that can deliver text and graphics, and even audio and video in response to requests from customers’ desktop clients.

2. A robust Internet server: business-critical for customer service

Companies who “put up a Web page” (maintain a Web site on a server) naturally count on the consumer appeal of their opening salvo — a uniquely appealing set of messages and service and product offerings — to establish an interested customer base who will make repeat visits to this particular Web site out of the tens of thousands that exist worldwide.

Whether these hopeful companies realize it or not, they are also dependent on the robustness of their Internet server to do the job of answering the thousands of information requests that can typically “spike” during the launch of an appealing new site or new offering, or with any unexpected surge in user demand.

Among many current examples of “spikes” caused by concentrated user demand, Digital offered live returns via the Internet during California’s November 1994 state elections; at peak periods, the Internet server drew more than 300,000 “hits” (requests) per hour.

As an example in the commercial sector, a financial institution serving customers via the Web will experience a large “hit rate” on the day before an anticipated change in federal interest rates.

The business-critical importance of having a robust Internet server cannot be emphasized too strongly. Customer interest and good will can be killed permanently when prospects see the inscrutable “host unavailable” message on their client screens instead of the Web page they were led to expect.

Spiralling damage from dropped connections

What is still worse, dropped connections that result in the frustrating “host unavailable” message are never picked up again; the fact that the request was lost is not recorded or reported anywhere on the server. It literally disappears. It thus becomes critical to “survive the spike” during the initial period when a Web site is offered to the public, or during any period of concentrated demand.

In the financial example above, the overwhelming majority of customers will try to access the institution’s Web page for instant information on changes to their investment portfolios. If the investors trying to access their records can’t “get in,” the institution will experience a communication crisis bearing real monetary consequences.

When an Internet server begins “dropping connections” (failing to service client requests), they are lost without a trace. The system manager has no way of determining whether connections are being dropped. The manager can track the load statistics to estimate whether the server is approaching saturation, but the actual point of saturation remains undetectable.

Because there is no single point at which a dropped request is ever recorded, the enterprise cannot know if the server dropped one, 20, or 2,000 requests. These become so many lost leads, aggravated customers, and lost revenue.

It is possible to measure satisfied demand by determining how many pages were sent out. But there is no mechanism for measuring latent demand — no way of knowing how many pages could have been sent out, but weren’t.

The business impact of not being able to “survive the spike” is especially devastating because it affects the user’s perception of the company, not the technology.

In addition, Internet servers also have to be robust because they perform many demanding functions besides displaying “pages” and delivering data to users. A growing percentage of Web pages today are dynamic pages offering search results, database lookup, and other information exchange functions.

What is more, electronic commerce and “tunneling” (for secure transactions) require complex algorithms to be calculated on the server to encrypt and decrypt data, a significant drain on CPU cycles. Other demanding functions performed by an Internet server can include APIs, server-parsed HTML, logging, lookup, access control, alias handling, multiple IP, proxy and caching, forms processing, and other Internet applications such as mail, news, and FTP all running on the same server.
3. Evaluating Internet servers

As a result, enterprises must carefully weigh performance characteristics before choosing the best Internet server with which to reach their customers. But the criteria offered by vendors can be elastic, making comparison difficult. When a user requests a “page,” this is commonly thought to mean a single file. Yet a Web page can, and usually does, consist of multiple files. And how many “pages” does a user download during a “visit”? There is no rule of thumb to correlate these terms. What is more, the size of the working set within the information hierarchy can drastically change results because of caching effectiveness.

Clear metrics are therefore necessary in setting test conditions. A single file of a given size (as distinct from a variable-size “page”) is the base unit. Professionals who test Internet servers at the system level express performance in terms of the rate at which the requests arrive, and the size of the file requested. (It takes longer for the server to deliver a big file than a small one.)

4. Performance comparison tests: Digital AlphaServer 1000 4/266 and 400 4/233 vs. Sun Netra i20

Such conditions were set during recent benchmark performance tests by the Application Systems Engineering Performance Group at Digital’s Equipment Performance Laboratory in Nashua, New Hampshire.

In these tests, Digital’s AlphaServer 1000 4/266 and AlphaServer 400 4/233 surpassed Sun Microsystems’ Netra i20 in performance and price/performance comparisons. Digital’s uniprocessor AlphaServer systems served 2x to 1.5x the number of files accessed by Sun’s more expensive dual-processor Netra i20.

Performance testing was designed to saturate the CPU with requests while eliminating as far as possible any other system components from becoming a binding resource.

The systems tested ran Netscape Communications Server® V1.1. The workload used consisted of a 16K file, a typical file size requested and serviced on the World Wide Web.

![Performance test results for Digital’s Internet AlphaServer systems vs. Sun Netra. File size was 16K.](image)

Test conditions were representative of what could be achieved in a production environment. To generate the “hits,” the group set up a test bed using an actual Web browser and an infrastructure emulating interactive users. The server contained 32,768 files, an amount representative of a mature Web site.

The AlphaServer 400 4/233 achieved a maximum of 109,680 requests per hour without reaching saturation, and yielded an average service time of 1.1 seconds. The Netra i20 handled a maximum of 66,188 requests with an average service time of 1.2 seconds.

<table>
<thead>
<tr>
<th>AlphaServer 1000 4/266</th>
<th>AlphaServer 400 4/233</th>
<th>Sun® Netra i20 Internet Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of and CPU type</td>
<td>1 21064A 266 MHz</td>
<td>1 21064 233 MHz</td>
</tr>
<tr>
<td>Cache</td>
<td>2 MB</td>
<td>512 KB</td>
</tr>
<tr>
<td>RAM</td>
<td>512 MB</td>
<td>64 MB</td>
</tr>
<tr>
<td>Network controller</td>
<td>Ethernet</td>
<td>Ethernet</td>
</tr>
<tr>
<td>Disk controller</td>
<td>SCSI</td>
<td>SCSI</td>
</tr>
<tr>
<td>Operating system</td>
<td>Digital UNIX® V3.2c</td>
<td>Digital UNIX V3.2c</td>
</tr>
<tr>
<td>WWW server software</td>
<td>Netscape Communications Server V1.1</td>
<td>Netscape Communications Server V1.1</td>
</tr>
</tbody>
</table>

This table shows the configurations tested.
The maximum number of requests processed by the AlphaServer 400 was 66 percent greater than the number achieved by the Netra i20, yet the AlphaServer 400 achieved 12 percent faster service times.

At a higher-capacity level, the AlphaServer 1000 4/266 with 512 MB of memory successfully serviced a maximum of 132.052 requests per hour with an average service time of 0.6 seconds — twice as fast as the Netra i20.

Note that Digital offers still higher power with multiprocessor configurations in the AlphaServer 2xxx series. The AlphaServer 2000 can be configured with one to two processors; the AlphaServer 2100 can be configured with one to four processors.

5. Critical factors in evaluating Internet server capabilities

These excellent results will come as little surprise to those who know that Digital’s AlphaServer family recently took the majority of the 1995 Hot Iron awards from AIM Technology. Yet while the test results demonstrate clearly superior performance by Digital’s AlphaServer systems on the basis of CPU performance alone, the choice of an appropriate system should be balanced by other critical factors in addition to relative performance.

• Price/performance. Uniprocessor AlphaServer systems outperform dual-processor Netra i20 configurations, yet the Alpha systems are less expensive.

• Configuration options. Digital Internet AlphaServer systems come as desktop workstations, neatly packaged servers, or space-saving rackmount servers with integrated telecommunications, storage, and CPU.
  – The Internet AlphaServer comes with a suite of applications including news, mail and a proxy caching server, as well as administrative tools.
  – Digital’s Internet software kits let you create the Alpha family configuration that suits your needs best.

• Future expansion capabilities. With expandable memory, disk, and I/O slots, Digital’s Internet AlphaServer systems are ready for processing-intensive tasks such as:
  – Handling dynamic pages created on the fly by CGI scripts.
  – Performing complex algorithms required by electronic commerce and secure transactions.
  – Proxy caching, forms processing, and other Internet applications running on the same server.

6. The future: interactivity, multimedia, secure transactions

Waves of innovation have made highly appealing Web pages possible for today’s mushrooming electronic marketplace. And technology is being developed, for example, to enable dynamic new browsers that will provide customers with interactive Web pages that can reply to queries and confirm orders placed.

As enterprises increasingly provide more complex offerings on the Web, including more interactive pages, multimedia attractions, and secure transactions, the typical size of requested files will steadily become larger. Robustness and reliability will become even more business-critical factors in selecting the best Internet server.

Today, Internet commerce is poised to continue its rapid expansion and to exchange increasingly content-rich information with users. Digital’s family of production-level AlphaServer systems will continue to meet the needs and expectations of commercial enterprises developing themselves into a full-service World Wide Web presence.

7. Frequently asked questions

Q. “How does client/server computing support Internet connectivity?”
A. Internet server performance is measured in terms of single exchanges with individual clients — a massive many-to-one relationship to the server. The presented load from any one client is minuscule — a few files at a time, at most. The server has to have it on disk and send it over the network. The amount of work that any one client does is extremely small compared to the work performed by the server. Thousands of clients each want a single thing from the server, which has to serve them all at the same time.

Q. “What causes a Web site to crash?”
A. Web sites don’t “crash” in the usual sense; they don’t typically have functional problems. The way a performance problem manifests itself is that the node appears to some percentage of the client population to simply not be there. Worse, there is no notification that this has happened.

Q. “How many files does a Web page contain?”
A. There is some question as to how many requests comprise a page. The term “page” has more than one definition. Some users identify a requested file with a “page” of information. Deceptively in a sense, a Web page looks like a single page because the browser formats it that way.

However, a single page can comprise many different files. There can be one file that contains text, one file that contains the toolbar, one file that contains a picture that might be the masthead. To compose a single page, the client may have to request a dozen files.

As one progresses through the media types, file sizes increase correspondingly. Still images and sound clips are bigger than text files; moving images are bigger than still images. Even a still picture (for example, the corporate logo that everyone sees on a home page) can be large, from 2K up to 150K at times.

Even though many files are small today, we are rapidly moving toward a future in which increasingly large files will be served.
Q. “One reads about security issues on the Internet. Can real business be transacted securely?”
A. There are mechanisms whereby real commerce can be transacted safely over the Internet. It is possible to do secure transactions such as exchanging a credit card number with the server in a way that precludes any interception between the client and the server. Digital’s Internet Tunnel products provide an example.

Q. “In what way does caching improve performance?”
A. The first time the server responds to a request, it actually has to go to the disk drive to read the file. Once it has recovered the file, it stores it in host memory. The second time a client requests that same file, the server does not have to get it from the disk. It can recover it from memory. Disk latency (time required by the process) can be thousands of times greater than memory latency. The performance testing described in this paper utilized a large number of files (32,768), providing a more realistic test condition than if the server had serviced only a small number of files and been able to keep them all in memory.

Q. “How does a connection actually get dropped?”
A. The client sends the server a request for synchronization, commonly called a SYN packet. If the server does not respond, the client sends another SYN within a few seconds. It sends SYN’s at further “retry” intervals, after which the client waits for a reply to the final SYN request.

There are thus four intervals at which the connection can succeed: immediately upon the initial SYN, after the first retry, after the second retry, or after the third retry. If none of the SYN’s is acknowledged, the connection attempt fails, and the client returns a “Host Unavailable” message to the user.
**Glossary**

*Alias.* A “pointer” used as an alternative, faster means of accessing a file, widely used on networks.

*API.* Application programming interface. Routines that an application uses in executing lower-level operating system services such as accepting input to a document, managing files, and displaying information.

*Caching.* When an Internet server processes the first request for a file from an Internet site, the file is cached (stored in the server’s memory) as well as delivered to the client. Subsequent requests for the same file can then be served from memory, a faster process. See also proxy server.

*Dropped connections.* Also called “timing out,” a critical situation in which the user’s request is ignored by a server that has reached saturation.

*FTP.* File Transfer Protocol. A program allowing a client to download files from a remote server over the Internet.

*Hit rate or request rate.* The rate at which client requests (hits) arrive at the server.

*HTML.* Hypertext Markup Language, enabling Web pages to contain “hot links” to other pages and sites. See also World Wide Web.

*Internet server.* A server dedicated to maintaining one or more sites accessible via the Internet, including World Wide Web sites.

*IP.* Internet Protocol. Part of the official protocol suite for the Internet.

*Logging.* Recording transactions and other activities that take place over the network.

*Lookup.* A lookup function searches a previously constructed table of values for a requested item of information.

*Page.* A Web page can deliver text, graphics, audio, and video in response to requests from customers’ desktop clients. A page consists of more than a single file.

*Proxy server.* A server on the user side of the Internet gateway which caches files as they are requested, so that subsequent requests can be served without returning to the Internet. See also caching.

*Saturation.* Maximum rate at which the server can sustain requests without losing any.

*Search engine.* Powerful research software offering expert guidance through the multitude of Web sites.

*Service time.* The time required for the server to deliver a file to the client. (It takes more time for the server to deliver a big file than a small one.)

*Tunneling.* A method of transporting data by encapsulating it in wrapper packets, thus protecting the encapsulated traffic from intrusion during transmission from source to destination. With tunneling, commerce can be transacted securely over the Internet.

*Visit.* A request or series of requests from a user who is visiting a Web site.

*Web server.* See Internet server.

*World Wide Web.* The fastest growing part of the global Internet. A network of servers whose text files are linked by hypertext, so that a user can navigate from one Web site to the next by clicking on key words serving as “hot links.”