Measurement Strategies to Manage Web Performance

By Peter Sevcik and John Bartlett
December 2002

Speed is one of the many factors that make a Web site successful. Fast response times and reliable performance are fundamental elements of successful Web-based applications. There are many studies that discuss the importance of responsiveness in the human-computer interface. James Gleick [1] makes the case in his book *Faster* that humans have become conditioned to expect all their interactions with technology to speed up over time. This means that a successful Web site must continually improve its responsiveness.

Where is your Web site on the march to fuse your users to your business? How fast is fast enough for today? How fast will you need to be next year?

Understanding the performance of your Web site from the perspective of end-users is essential to managing your service. This applies to all types of Web-based applications: consumer retail sites, business transaction sites, content sites, intranet services for employees, and non-profit public service sites. Speed retains users, increases traffic, and improves transaction completion rates. The bottom line is that Web performance directly impacts usage, business results, and revenue.

Managing the responsiveness of a Web application is a complex problem. Complete data and insightful diagnostic tools are needed to understand the performance users receive and to design performance improvement solutions. This report describes a comprehensive strategy that enables enterprises to manage performance. It begins by laying out a performance management strategy along with a measurement methodology critical to the strategy and concludes with examples showing how the methodology is applied.

The primary objective of the strategy and measurement methodology is, happy users more of the time and performance alarms less of the time. In fact, the simplest test of success is a reduction in performance-related alarms. The business test of success is an increase in revenue or user retention.

Performance Management Strategy

The beauty of the Web is the organic nature of the content, users, and underlying infrastructure. This distributed evolutionary process is one of the essential aspects that permit growth on a global scale. The downside to this arrangement, however, is that everything is always changing. A business manager trying to ensure positive customer experience is managing many moving puzzle pieces.

All businesses that transact over the Internet must develop a performance strategy. The strategy should be developed with the input of all parties in the organization who can affect performance (developers, server managers, network managers) and the managers that set performance targets (business, marketing). The strategy should be documented as a policy with clear goals and a set of procedures by which the policy is implemented.

The strategy begins with knowing the users and the Response Time Agreement (RTA) they require, institutes continuing measurement processes, uses the data gathered to baseline performance, and, finally, correlates data from various measurement points to diagnose problems.

NetForecast Report 5064

Know Your Users

Performance management starts with knowing the application and the user. A Web site manager must know the few tasks a typical user performs on the site on a continuing basis and understand the maximum time it should take to perform those tasks. Better still is a detailed profile by the importance of each major function performed (e.g., browsing vs. checkout). These data can be acquired by benchmarking the site once or on a continuing basis using a measurement system.

The second part of a specification entails defining performance by user class. The strongest correlation to performance is a classification is by the user's access method to the Internet – consumers with dial-up, consumers with broadband, or businesses on a corporate network. The second level of classification should be by distance from the Web site – locally within a metropolitan area, nationally across a country or globally across the world. Some organizations group users by the level of business they perform on the site – premium service to high volume customer or modest service to first-time visitors.

It is also very valuable to interview or observe real users navigating through your site. This is the best way to identify the most frequently performed tasks and their target response times. The classifications and observations of users must be used to define specific performance targets for each class as defined by your organization. Setting user experience targets is the foundation for understanding performance. Not knowing how good the performance should be leads to not caring about performance.

Once the performance requirements of a Web site are understood, it is best to document the performance characteristics and targets in the form of an RTA. An RTA is typically established with an organization that is striving to properly manage assets, vendor relationships, and user expectations. These agreements are a way to ensure that each part of the organization is assigned its proper role in maintaining acceptable performance.

Measurement Processes

Measurement in support of a performance strategy is defined as processes that must be instituted on an ongoing basis. A performance strategy has to include the following capabilities to ensure that users receive consistent or improving response time:

> **Baseline** – A formal set of tests to determine the foundation numbers that will be the reference in any comparisons. These are often performed when a new site is deployed, significant design changes occur, or when making a significant change to the infrastructure. A baseline is the starting point for predicting performance problems.
>
> **Trend** – Trend analysis shows how performance is changing over time. Performance changes may come from changes to the content, to network utilization, to the server farm configuration, or to many other factors. These are a limited subset of the baseline tests that are performed on an ongoing basis. Typically these tests are summarized on a sliding timeline: Hourly for the last day, daily averages and top percentiles over the last many months.
>
> **Evaluate** – Evaluation requires a series of tests to determine if a design change or implementation of a new service fits within the established performance goals. Any significant new technology or service should be evaluated for its performance impacts. These tests are best performed in two stages, starting with

controlled conditions, then using reference servers, then moving to operational traffic.

Diagnose – Diagnostics determine which part of the complex system of users, networks, and servers is creating a performance problem. With so many "moving parts" to the problem, diagnosis is really a matter of deductive reasoning. The goal is to eliminate non-offending parts of the system in order to see what is left as the remaining troublemaker.

Correlation Analysis is Critical

If the system were all contained in a lab, all the individual pieces could be instrumented to determine how much time each portion of the system is using. Since the environment is instead spread across the world through multiple services over which there is little control, a different approach is required. This problem is tackled by testing the system with different types of users and network connections, and by then looking for correlations between good or poor performance.

The goal is to identify a strategic set of reference points to which the tests are conducted in order to perform the correlation analysis productively. Here are some to consider:

Application – Known pages that do not change over time. Some may be dummy synthetic transactions to servers behind the primary Web server.
Server – A reference server that is not subject to site load. This can be any administrative server with a few reference pages.
Campus Network – This is typically supplied by pinging the site's edge router. However, it is better to test to yet another reference server in the DMZ.
Wide Area Network – Testing from a wide range of well-connected (broadband or T1) users to a reference server provides a regional view of performance.
User Access – Measuring across many access line rates and comparing them to the wide-area data shows the effects of a slow line.

A proper measurement strategy has the reference points selected and data being gathered continuously in order to be able to quickly correlate among known points and old/new data. Table 1 shows a summary of the techniques that can be employed to diagnose Web performance.

Table 1 – Using Correlation to Zero In on the Problem

Poor Performance Correlates With:	Performance Problem Caused By:
Access line speed	Too much payload, or access line is too slow
Geography	Application Is latency sensitive, has too many elements, or there is too much distance between the client and server
Backbone Provider	Poor carrier, or overloaded peering points
Server	Slow server, or too complex an application

End-to-End Performance Testing Is Essential

The performance management strategy described above completely depends on having accurate test data. Enterprises must purchase a service or product to test their Web services in order to gather such data. There are several alternative sources of such information available. However, selecting the proper measurement resource is critical to the success of the performance management strategy.

Components of Web Load Time

The process of transmitting content or executing transactions between a Web server and browser are very complex. A good description of how it works can be found in *Understanding Web Performance* [2]. The fastest the process can operate is when the server and browser are directly connected with no network to get in the way. In this case, task time is simply driven by computing time. But there is a network that adds time to the interactions. Figure 1 shows the basic elements that must be traversed by the typical server-browser interaction.

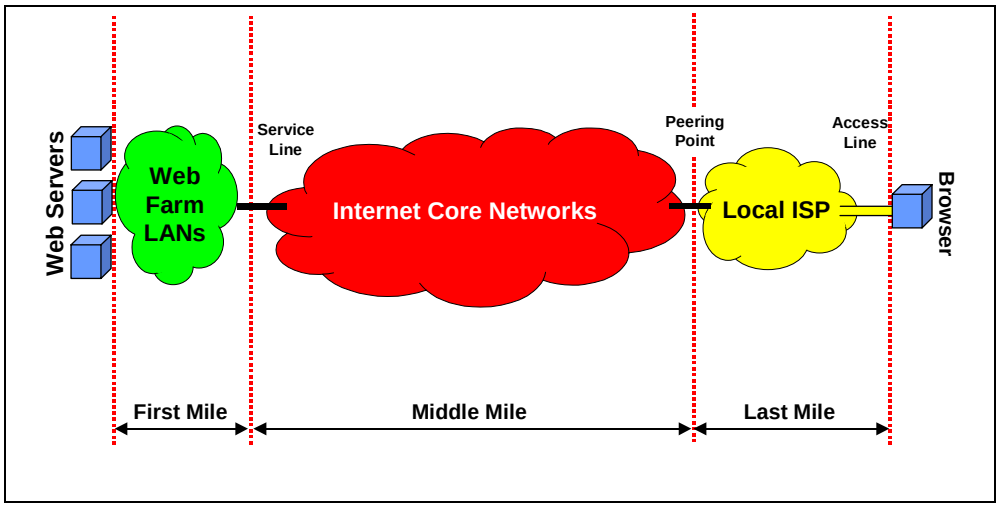

Figure 1 – Breaking Down the Path from Server to Browser

In Figure 1, the First Mile indicates the connection from the server to the campus network. It may be a hosting site or the corporate data center. In either case, the data must traverse a significant number of switches, load balancers, firewalls, etc., before it reaches the border router and then the access line and the edge of the Internet.

The Middle Mile represents the region where traffic will pass over one or many networks (autonomous systems or ASs) before arriving at the destination ISP. The formal definition of the Middle Mile is the portion of the Internet where path diversity exists. This portion is also typically referred to as the Internet backbone or core.

The Last Mile is the destination ISP that serves the user and then the access line going to the user. This may be a complex regional network or it may be a simple direct line if the user is on a significant corporate location. However, more than a direct line will separate even the corporate user from the Internet core. That user will typically have to traverse a large corporate network to get to the Internet.

Figure 2 shows the percentage of total page time that is attributed to the major elements of the path from the Web site to the user. The figure is based on the model [2] and methodology developed by NetForecast for predicting the performance of a typical Web page from profiles gathered in 2001.

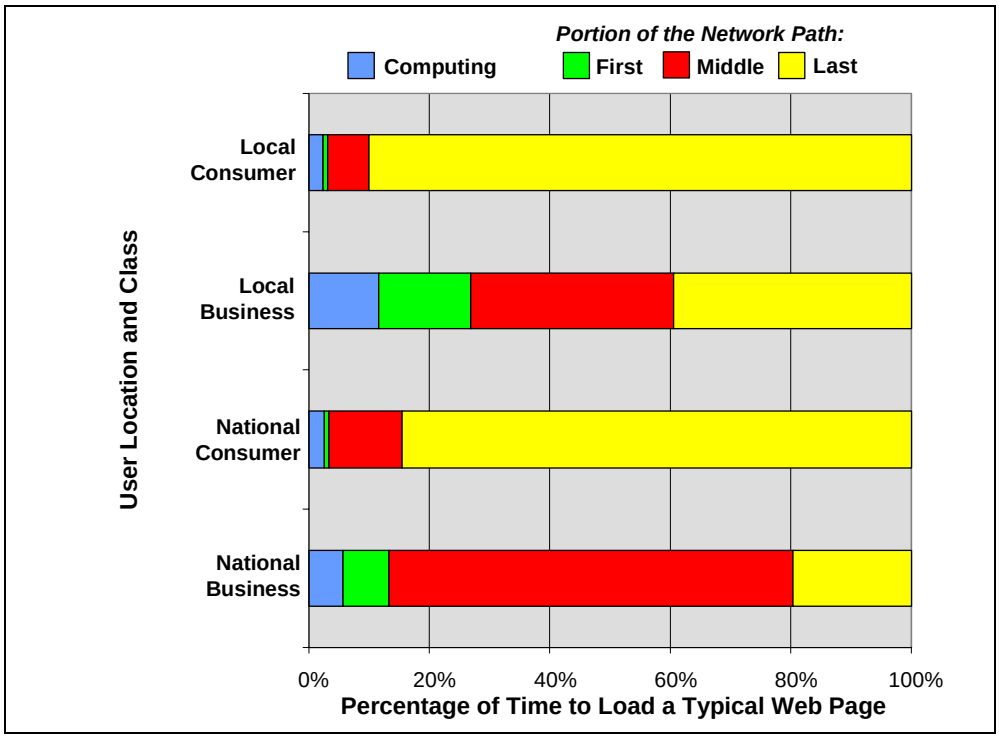

Figure 2 – Components of Web Response Time From User's Perspective

Clearly, all of these parts of the path must be measured. Measurement services that operate from a few locations on the Internet core are missing the data associated with the Last Mile. Although the user's local ISP looks insignificant in Figure 1, its impact can be dramatic. Figure 2 shows that the Last Mile represents 20% to 90% of the response time! Any comprehensive measurement service must include this element of time.

Figure 2 also shows the shift of element times from a transnational (across the US) session compared to a local (within a metropolitan area) session. It is important to note that the Last Mile represents an even more significant portion of transaction time when the user is closer to the Web site.

It is interesting to note that when a site employs content delivery network (CDN) service, then the measurement from a backbone-based agent will likely be very unrealistic. Only the Last Mile measurements can be used to evaluate the impact of a CDN service.

Where to Measure Performance

There are many vendors in the Web performance testing marketplace offering services that look similar but that are fundamentally different views of performance as described in Figure 1.

> **First Mile** – Test equipment or software that instruments Web servers and the local data center or Web farm LANs. These tools are useful to ensure that the

Web pages or transactions are created in a properly and timely manor. They are primarily used to manage performance behind the Web server.

Middle Mile – Services that operate instrumentation points on Internet backbone networks with agents directly in carrier switching nodes or in third-party data centers that are directly connected to carrier nodes. These services typically perform tests among themselves to study the performance of backbone networks and the peering points among backbone networks. Some of these services can also test back to a Web server. However, such a test is incomplete for managing the user experience and does not provide sufficient value to this strategy.

Last Mile – Services that operate agents on user desktops. These services address the true end-to-end performance of Web applications – providing visibility across the first, middle, and last mile to enable companies to develop a truly comprehensive performance management strategy. These services are essential to this strategy.

The Gómez Desktop Monitoring Service

Gómez Desktop Monitoring offers a practical Last Mile measurement solution that can support this strategy. The company's Desktop Monitoring service lets customers design and submit performance testing and maintenance programs by entering the URLs (or multi-URL transactions) that they want to monitor, choosing the frequency of testing, and selecting the desired testing coverage (including parameters such as geography and connection types). After submitting tests, Desktop Monitoring allows customers to view their site's performance data in real-time, analyze trends, diagnose problems, and receive automated alerts whenever performance falls outside configurable thresholds.

This network of machines is carefully managed by Gómez to offer a representative slice of the Internet population. Since Gómez gathers measurements through active testing agents installed on real user desktops, the results truly represent the real user experience. This ensures that the measurements include all the elements of response time as described in Figure 2. The distribution of agents represents the industry's largest collection of geographies, ISPs, and access line types.

The results from this testing includes data by specific types of users and various reference points. Results can then be correlated across the different parameters to determine where performance problems exist.

Examples of Improving the User Experience

The following are specific examples of cases with performance concerns that could be managed using this approach:

Example I – Web-Based Retailer

A Web-based retailer (e.g. department store, financial services, electronics, automotive) is providing information via their Web sites, and will likely have the following concerns:

Browsing speed – Will users continue to page around my Web site until they find the products / services they want? **Solution** – Test page response time for key pages, across a wide geography, and range of defined user classes. Compare the results of these tests with the baseline that was considered satisfactory. Results that are consistently outside the target must be investigated

further. Correlate the sub-page timing data across several reference points on the path to determine the problem.

Performance Trend – Is my Web site continuing to provide good service as the network and Web site traffic grows over time? Is my Web site providing adequate speed during all hours of the week, through busy periods for my Web site and through busy periods for the network? **Solution** – Test page response on an ongoing, periodic basis, log trends and graph the long term results to predict when upgrades will be necessary to continue to meet accessibility goals. Compare the results of live traffic and a reference server on the site (not loaded) to determine the effects of load. Set alerts to indicate when performance crosses a predetermined threshold.

Example II – Web-Based Financial Service

Web-based business (e.g. banks, insurers, credit card issuers, lenders) are doing transactions via their Web sites, or trying to answer common questions in a cost-efficient way. These firms often have a different set of requirements.

Transaction speed – Will users work through each step of a process in order to finish their transaction? **Solution** – Test response time for multiple steps in a transaction script, across a wide geography and range of access lines and client types. The tests should focus on a critical process such as initiating an order.

Completion Rate – When users are ready to transact, are my secure pages fast enough to keep them connected until the transaction is completed? Another critical aspect of performance is the fact that consistent completion rates engender trust in the service. **Solution** – Test specific transactions, through a Secure Sockets Layer (SSL) interface, across a wide (or specific) geography, range of access line speeds, and client types.

Example III – Information Web Site

Information Web sites are typically supported by outside services such as advertisers and news feeds. The user views a composite from these various sources. Such sites are also often subject to flash crowds so they implement a content delivery network (CDN) service to offload the traffic.

Diagnostics – Are third-party Web servers that participate in my pages slowing my response time? **Solution** – Compare response time of the site's content to response time of other third-party content such as banner advertisements or logos of partners. Trend response times of third parties to ensure that they are meeting performance goals during heavy Web page or network load periods.

Scaling – Does the CDN provide proper scaling for large crowds? Is it cost-effective? Is latency reduced for all major user geographies across the nation / globe? Is performance from each CDN site adequate? **Solution** – Test the site from a wide geography, using common desktop and access line speeds for comparison. Correlate performance issues to geography or CDN address. Compare response time of static content (CDN based) with dynamic content (server based). Evaluation of the CDN service is often an ongoing analysis as both the nature of the Web site content and the CDN service coverage change over time.

Summary

Every network-based service must manage the user experience in order to satisfy customers and grow revenues. Successful organizations develop a performance strategic plan with proper testing tools to ensure that the plan is meaningful. The strategy allows active performance management to solve problems before they become customer dissatisfaction.

Tests performed when a site is launched are flatly not enough. Tests must be continuously monitored and improved to target important users as part of an ongoing maintenance program. Therefore, a comprehensive testing service is a critical element of the strategy.

Gomez's Desktop Monitoring service provides the feature set required for comprehensive testing and performance management. Gómez's large base of agents and distributed geography allow correlation of performance issues and network parameters. Gómez's trending and reporting features enable performance strategies that can be maintained and refined as requirements change. The value of the Gómez offering builds over time as the database of reference measurements increases.

Companies whose Web sites represent a significant source of revenue, customer contact, and resources should develop a performance management strategy and invest in performance testing tools to continuously gauge and improve performance. ¤

References

1. James Gleick, "Faster, The Acceleration of Just About Everything," Random House, 1999.

2. Sevcik and Bartlett, "Understanding Web Performance," NetForecast Report 5055, October 2001. (Available at www.netforecast.com.)

Peter Sevcik is President of NetForecast and a leading authority on network technology. He specializes in performance analysis and holds the patent on application response time prediction. Peter has contributed to the design of more than 100 networks including the Internet. He can be reached at peter@netforecast.com.

John Bartlett is Vice President of NetForecast, and has 24 years of experience designing at the chip, board, system and network levels, with a focus on performance. John led the team that built the first VLAN implementation, one of the first ATM switches, and he is a leading authority on the behavior of real-time traffic on the Internet. He can be reached at john@netforecast.com.