

Web Performance – Not a Simple Number

Net Forecasts – Peter J. Sevcik

BCR Volume 33, Number 1

January 2003

We are often asked to analyze the performance of an Internet application by supplying a single answer to a single question: "How fast is it?" Whenever we explain that there is no simple single number, a common response is: "Yeah, OK, but what is *the* number?"

Performance measurement vendors, on the other hand, make their living by supplying a single number. The result, however, can be misleading; decisions founded on simple data, even if it's impartial, often wind up badly.

The problem with the "simple answer" is that by definition it is an aggregation of all the data gathered on a site. That's why you often see results that repeatedly contain the word "average." For example, "This is the average of measurements from each agent by hour, then averaged across all agents, then averaged across the week." When you see that kind of sentence beware: The interesting or meaningful information often has been ground out.

In the real world, users experience a range of performance that is often much worse than the average would indicate. The average or mean is influenced by the preponderance of time samples around a measurement peak. A more realistic view of performance is the histogram or statistical distribution of time samples. This method often shows what is known as a "long tail" to the distribution of time samples.

Getting Detailed Measurements

We were fortunate to get the cooperation of the three leading measurement services in order to study this issue. The companies are:

Matrix NetSystems (Austin, TX) has been measuring the Internet since 1990. It utilizes thousands of beacons around the Internet that can be controlled to measure latency, packet loss and reachability.

Keynote Systems (San Mateo, CA) has an extensive set of tools for benchmarking and testing of Web sites on the Internet. It operates agents on servers in key locations on the Internet. The data supplied shows load times of full Web pages and their components.

Gomez (Waltham, MA) also operates server agents at key Internet locations. In addition, it has a desktop service based upon "normal" users permitting their PCs to be agents under the control of Gómez. The company also provides an advisory service to understand and improve non-performance aspects of the Web site and the total quality of the user experience.

NetForecast used the measurement tools of each firm to gather data on the performance of many Web pages under a variety of conditions. However, in each case we asked the service to supply us with raw measurement data so we could investigate the distribution of time samples.

It is important to note that the goal of this analysis is to show the insights that can be learned from seeing all the data rather than the mean. It is *not* a comparative assessment of the measurement services.

The result of the analysis is shown in Figure 1. The curves represent the distributions of page load times within the continental United States based upon the following data sources:

Keynote BK40 is the Keynote Business 40 Index of 40 business sites across a week of measurements. The measurements use Keynote's Web Site Perspective Business Edition operating in about 21 cities where the servers are directly connected by 45-Mbps connections to major backbone Internet service providers (ISPs).

Gomez PN Financials is the Gomez Performance Network (GPN) service measurements of 40 major financial sites over a week. The GPN service operates in 25 cities using agents connected to multiple backbone ISPs by 10-Mbps connections.

Matrix of KB40 is the Matrix Internet Average service used to supply round-trip time (RTT) and packet-loss measurements using five beacons testing to 109 destinations that are connected to the Internet by typical access lines and access ISPs. We used this low-level Internet data as input to our performance model set with the application profiles of all the KB40 sites. This is therefore a model-

generated result of page load times at the network edge based upon Internet measurements (see *BCR*, October 2001, pp. 28-36).

Gomez DM of KB40 is the Gomez Desktop Monitoring (GDM) service, which measured the KB40 sites from hundreds of user desktops over two weeks. This service was set to only use broadband connected desktops (1.5 Mbps or faster). This is the most "edge-oriented" measurement since it uses desktops that are connected via access networks or corporate networks.

The curves in Figure 1 provide very interesting insights. First, it is clear that there really are two distinct set of curves: the Keynote and Gomez backbone-measurement pair and the Matrix and Gomez edge-measurement pair. Most surprising is the fact that although the Keynote and Gomez backbone services measured different sites (only 2 of the 80 are in common) using different

measurement agents, the two curves are almost identical.

Moreover, it is clear that *where* within the 'Net (i.e., backbone vs. edge) you measure is much more important than how you measure. *What* you measure is defined by the individual page profiles that are diverse; in this case, we averaged results across all Web sites measured by Keynote and Gomez. The detailed view by site, while important, would have resulted in Figure 1 having 160 curves. The two "edge" curves of Matrix and Gomez DM are not as close a match, but are still similar enough to be called a pair.

Second, the edge view consistently shows much slower performance. The backbone performance is fast and consistent. There is a distinct peak of performance in the 2 to 3 second range, with a very fast drop-off towards longer times. However, the edge curves show a much more sloppy performance zone across the 3 to 12 second range.

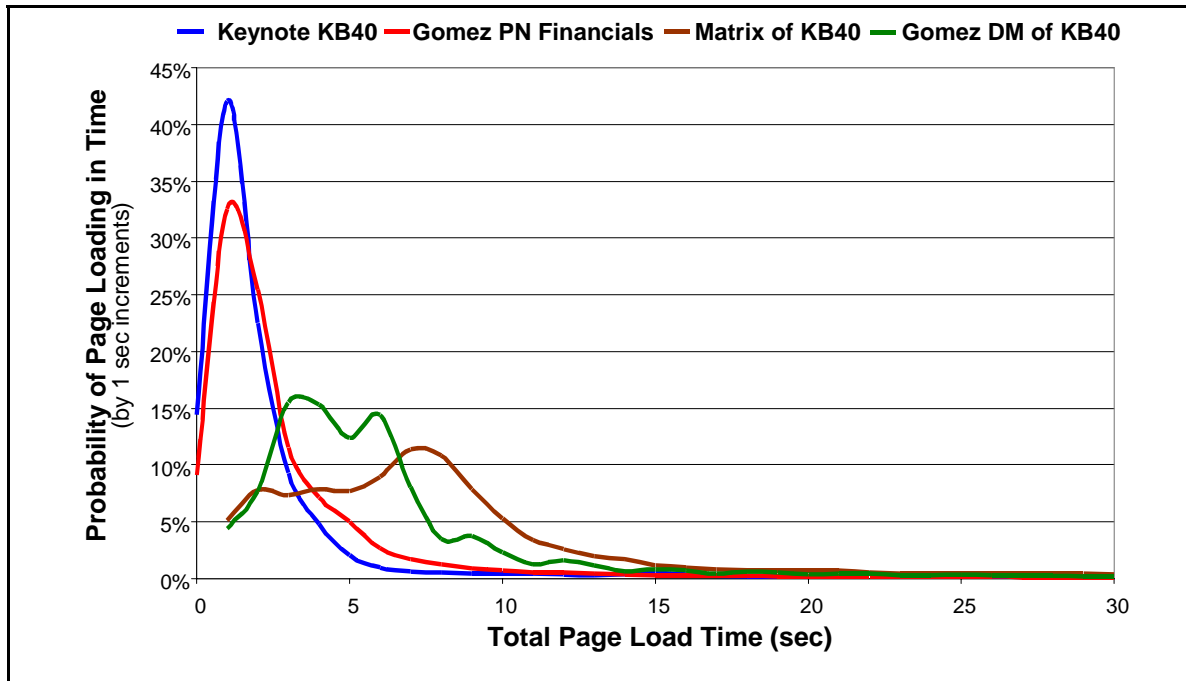


Figure 1 – Distribution of Page Load Times Within the US

Investigating the Long Tail

The four curves in Figure 1 look different in the first 15 seconds, and very similar after 15 seconds with hardly any users occupying the region. But this doesn't mean that there is an almost zero probability of seeing times after 15 seconds. Indeed, while the probability of seeing each unique second after the 15 second mark looks slight, there are many of them stretching far out to 60 seconds and beyond. These small probabilities add up. To really understand the effect of the "long tail," you have to perform in-depth statistical analysis of user groups out on the tail as show in Table 1.

Table 1 starts with the overall mean of the total sample set for each measurement service. This is the "one number" answer that most people like to hear. Not surprisingly, the means for the two backbone measurements are very close (2.2 and 2.7 seconds). But notice that even though the curves for the edge measurements look different in Figure 1, their means are actually the same at 8.4 seconds.

However, when the means are calculated for subsets of the total population, the values tell a different story. Not surprisingly, when the samples are cut in half by speed, the faster half is *much* faster than the overall mean, while the slower half

is understandably slower. The slowest 20th and 10th percentiles of the samples are 3- and 4-times slower than the overall means. This is a consistent pattern for both the backbone and edge measurements. The long tail is confirmed and shows the same pattern in all measurement methods. This is an important finding, and the implication is that regardless of the measurement service or testing location, large numbers of the user population see significantly poorer performance than the one-number answer.

Keynote also supplied us with a week of raw measurements of the Keynote Consumer 40 (KC40) sites. It was expected that these sites, which are measured using dial-up modems, would show a more significant distribution tail.

Surprisingly, although the KC40 had a long overall mean of 22.5 seconds, the ratio of the slow group to the overall mean was a very modest 1.8 for both the slowest 20th and 10th percentiles. This shows that the consumer test results are much more flat with no performance peak or tail. We think that the payload for the KC40 Web sites trying to pass over the constrained dial-up link governs this flat curve. Therefore, the effect of the distribution of latency across the Internet is removed from the picture.

Table 1 – Mean Response Times (sec) by User Percentile

	Backbone Measurements		Edge Measurements	
	Keynote KB40	Gomez PN Financials	Matrix of KB40	Gomez DM of KB40
Overall Mean (0-100%)	2.2	2.7	8.4	8.4
Fastest Half (0-50%)	0.7	1.0	3.8	2.7
Slowest Half (51-100%)	3.4	4.3	12.5	13.2
Slowest Fifth (81-100%)	6.1	7.1	19.6	25.6
Slowest Tenth (91-100%)	9.2	9.8	26.8	41.9

Practical Example

Why should you care about all these complicated details? Simple: It matters to your users. For example, consider a Web site that has a 3-second target for page-load time, which is reasonable for a typical business-to-business site.

If the Web manager were to subscribe to one of the more popular backbone measurement services, he or she may believe that they had achieved their goal with a mean score of less than 3 seconds (see Table 1). And given the above lesson in the distribution tail, even the slowest 20th and 10th percentiles of users experience about 7 and 9 seconds mean response times. Not good, but not too bad. In some on-line business situations, it may be worth having a discussion with business managers with the question, "How bad is 3-times the target for a tenth of our users?"

However, if the manager thought through the problem a bit more, he/she might realize that very few users are directly connected to major Internet backbone nodes at 10 or 45 Mbps. In fact, the broadband edge user profile may be much more typical of the real user population.

Furthermore, 4-times slower has a significant impact to the user's interaction with a computer (see *BCR*, July 2002, pp. 8-9). In this example, 12 seconds is 4-times the target, but we can clearly see that half of the users on the Internet edge are seeing

performance that is slower than 12 seconds. Now how about a conversation with management that starts with, "How bad is 4-times the target for half of our users?"

A shift in measurement point and looking deeper into the data creates a very different perspective on how the users feel.

Conclusion

Use these measurement services, but use them wisely. Most importantly, avoid the simple index or single-number answer. Demand insightful distribution analysis of performance from these vendors. One size does not fit all in clothing, so why would one number represent the performance seen by *all* your users?

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

NetForecast Inc. is a network technology consulting firm based in Andover, Massachusetts. Our seasoned consultants draw on decades of experience to help clients worldwide choose new technologies, improve performance, and align infrastructure to business. We have helped leading enterprises, service providers, and vendors navigate the changing competitive landscape of the Internet economy. Please call us to discuss how we can help your information network succeed.

