

Real vs. Synthetic Measurements of Performance

Net Forecasts – Peter J. Sevcik
BCR Volume 34, Number 11
November 2004

Application performance is getting more scrutiny as people discover the link between high performance IT and a high performance business. However, performance remains a complex topic that clearly suffers from the “eye of the beholder” syndrome—it means all things to all people. A year ago in this column, I proposed “A Framework for Enterprise Application Performance,” to give this broad topic a structure for constructive dialogue and standards (see BCR, November 2003, pp. 8-10).

Despite progress in getting organizations to better define the specific performance function they are measuring or improving, there remains a lot of confusion. The greatest confusion centers on how to measure the response time of a networked application. The questions surround what to measure and how to measure it.

Defining What To Measure

At least four levels of transaction detail could be measured:

- First, there is the session or process level, for example, purchasing a book or entering an employee record.
- Second there is the series of tasks required to accomplish the preceding process(es).
- Third, there are the client-server and back-end interactions or program threads that occur behind the scenes in order to deliver a response to the user.
- Finally, there are the very detailed protocol interactions required to deliver each part of the process.

Most people agree that delivering the proper application response time falls into second level of detail in the above decomposition.

Within that second area, a “task” is defined as the elapsed time between when a user makes an entry (finishes a field, presses the enter key, clicks on a dialog box, etc) and when the application responds with sufficient information so that the next task can proceed. The task is the level at which a human user feels the responsiveness of the application; the task times are the periods when the user is waiting for the system to respond.

In contrast, the “think times” are all the other periods of time required to complete the process—it’s basically the end user’s share of the work. Thus the sum of all task times plus the sum of all the think times equals the total process time.

In order for a business process to operate at a faster pace, it can either make the user think (including type) faster or make the computer system (applications and networks) run faster. Tasks are the only unit of measure that clearly describe that requirement.

How To Measure Response Time

The above definition precludes many tools that purport to supply information on response time. Element managers, tools that gather data via SNMP from network devices, and RMON probes can only count how much traffic is going by each measurement point. Although valuable information, these types of measurements cannot reconstruct the individual user session flows and then time each task. Also, detailed reports on how an application is performing within the servers are often at the thread level—again, not integrated into complete task timing.

However, many tools are available that do report task-level timings. These tools come in two major categories—real and synthetic (also respectively called passive and active measurement techniques).

The “real” measurement tool can instrument the individual user application flow, and can detect the start and stop of each task. This can occur at a key measurement point where all of the packets of the flow can be detected, or it can be measured directly on the user’s desktop. The latter method requires adding a small measurement applet to the desktop.

It is clear that these measurement techniques actually see the user’s task-level interactions. These real user measurements see all the traffic all the time. The results are not ambiguous.

The “synthetic” measurement tool operates as a proxy for the real user. Software operating on a strategically located computer runs a pre-defined script that is a facsimile of a user task or series of tasks. Often the tools have a feature that records a

real user performing a task(s) in a test environment and then generates the script automatically. The script is loaded into the synthetic agent machines and programmed to operate at a pre-defined pace much like a robot. The results are then reported by each synthetic agent into a central repository where statistical analysis and summaries are performed.

Synthetic agent tools come in two forms—private agents or public services. Private agents are placed within the enterprise network to cover key locations which house the users that are being simulated. Public services place agents globally around the Internet in locations that an enterprise cannot instrument. For example, an online store may want to see the quality of its service across the U.S., Europe and Asia. But such a store has no facilities in any of the cities where its customers reside. The two leading synthetic agent service providers, Keynote and Gomez, have agents strategically located throughout the Internet. Their agents can serve as a proxy for the real users in many geographic locations.

The difference between the two approaches can be seen from an automobile safety analogy. Synthetic agents are the crash test dummies that help designers create a safe car. But real measurements are like studying accident reports of traffic fatalities. Both approaches are valuable, but each provides different insights.

The synthetic agent approach is easy to implement and is often required in order to deploy a new application, just like the crash test dummy is required to test the basic safety of a new car. The problem comes when enterprises fall into the trap of thinking that they know all that they need to know from the synthetic agents. This is especially true of the synthetic test services that keep providing “ongoing” data on the application performance. However, the synthetic agent services are only as good as their ability to accurately simulate all of the users of an application. The question is: How well do the crash test dummy and the test lab simulate all drivers, human bodies and road conditions?

Investigating The Differences

NetForecast was fortunate to get access to data from three major on-line businesses that used both synthetic agent services and real user measurement tools. The businesses were a shopping site, financial services firm and a hosting site from which many online businesses operate. All of these businesses take response time very seriously, and each subscribed to both the Keynote and Gomez services. We investigated a simple question—how well do the synthetic agent services match the universe of real user experience?

We compared data over many days where synthetic agents were testing while real users were operating on the sites. In all these cases, task time is Web page load time.

The synthetic agent approach is a sampling methodology. The agents operate scripts that hit some of the websites’ pages from some of the customer geographic locations some of the time. It is impossible for the agents to completely recreate a full user population, since that would by definition double the traffic on the websites. Such a high level of instrumentation would adversely affect the sites’ performance and be prohibitively expensive. So the proof of how well sampling covers the real user population is to measure the degree of coverage that actually existed for these sites.

Extent Of Coverage

Each of the three business sites operates a vast website with a large amount of content, and each is accessed by millions of users across most of North America. The test of coverage is to see just what percentage of unique Web pages (URLs) accessed by real users was tested by the synthetic agents, and how much of the geographic regions (autonomous systems or ASes) were in common between users and synthetic agents.

The percentages of Web pages tested by the synthetic agents relative to the pages viewed by real users were:

Shopping	8%
Financial	9%
Hosting	1%

Many of the hosting site’s URLs were uniquely tailored to each visitor of the site such that the synthetic agents could only test the basic pages of the site.

The percentages of networks or autonomous systems from which the synthetic agents operated, compared to all of the networks represented by all the users were:

Shopping	9%
Financial	9%
Hosting	14%

In general, synthetic agents represent less than a tenth of the networks where real users originate or pages real users visited. It would appear that this is a small fraction of the user experience upon which to rely when making judgments about service delivery. Remember that these are not 1-in-10 samples of the identical product coming off an assembly line. This is a sampling of 1-in-10 products from a catalog of millions of different products.

However, the common coverage result immediately opens the next question—are a few well chosen samples good enough to provide an accurate measure of the user experience?

Response Time Measurement Accuracy

The detailed task timing data from all the users and all the synthetic agents is presented in the graphs shown in Figures 1-3. Each of the points in the charts is a 10-minute average across all the measurements for each synthetic agent service and all the real users. Even though the synthetic services were set to have each agent run its script in approximately 15-minute intervals, there were so many agents involved that each 10-minute interval had many samples for each curve.

All of the response times are normalized to the overall average response time seen by the real users. The elapsed time for each of the sites is different based upon the data sets available. However, each site provided at least a week of useful data.

The financial site purchased three services from the two measurement service providers as shown in Figure 1. The top line in Figure 1 is the real user response baseline. Notice that it shows a definite pattern of nominal performance followed by about 20 percent longer response times during each day. The big exception is Sunday, July 11, 2004, when the load on all of the systems dropped sufficiently to permit a 30 percent faster response time. All three of the synthetic services reported dramatically faster response times and generally did not track the typical daily cycle of performance. The one exception is that one of the three services does see the Sunday dip. It is clear that the financial site is getting a very misleading view of response times.

The shopping site has an even more serious synthetic reporting error, as shown in Figure 2. Again, the synthetic services present an overly optimistic and dramatically divergent view of reality. First, Synthetic service #2 is extremely optimistic, showing times that are often a fifth of the real user experience. Secondly Synthetic #2 did not notice a big degradation in service that occurred in the afternoon of July 1 (Synthetic #1 does see this event but not as severely as it should). Finally, Synthetic #2 shows a false alarm of a nine-fold jump in response time when actual response time was below average and steady.

The hosting site has yet a different story to tell, as shown in Figure 3. Here, Synthetic #1 is always reporting somewhat slower performance than the real users see. However, Synthetic #2 is reporting a four-fold pessimistic error until July 14, during which the errors swing wildly and settle on a generally accurate view of performance only to drift towards optimism by July 17. The way the hosting company copes with these dramatic changes is to first investigate the synthetic service to see if there really is a problem that should be dealt with.

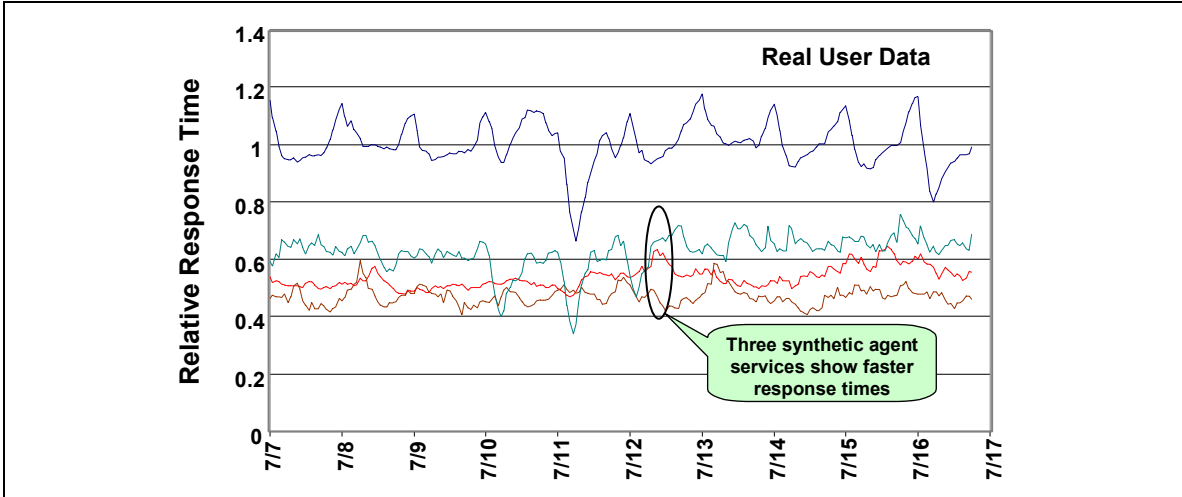


Figure 1 – Financial Site Response Time

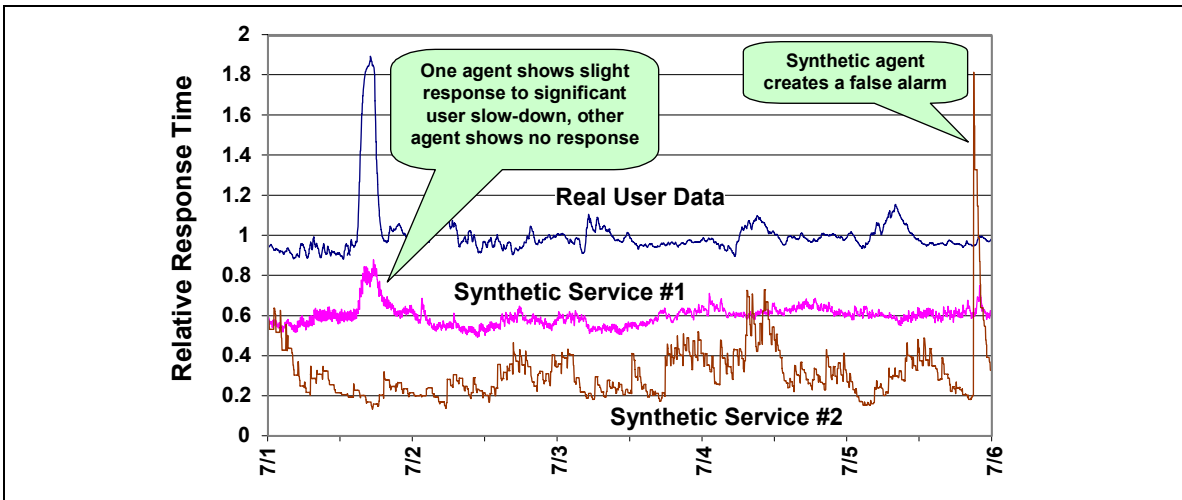


Figure 2 – Shopping Site Response Time

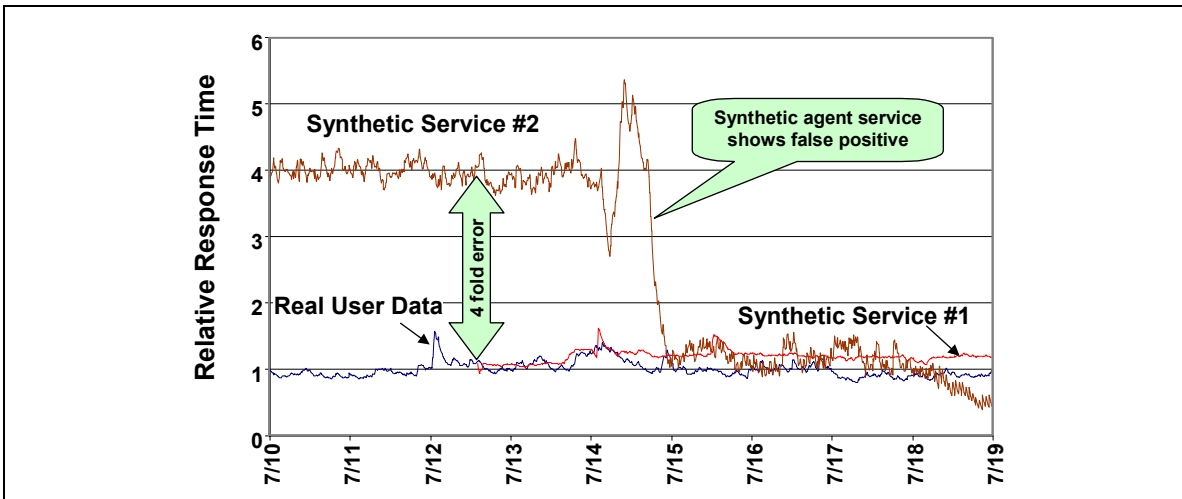


Figure 3 – Hosting Site Response Time

These data show that that the two leading synthetic services provide very unreliable information regarding actual user experience. The errors are often not even consistent enough to develop a strategy on how to use the results.

How The User Feels About Response Time

All of the interesting data in the figures is averaged across many measurement samples just like most measurement tools or services report. However, I have repeatedly made the case in these pages that users do not view performance as an average but rather as operating within one of three zones: satisfied, tolerating and frustrated (see BCR, July 2002, pp. 8-9 and March 2003, pp. 8-9). Furthermore, any averaging method washes out the long tail of the performance distribution curve (see BCR, January 2003, pp, 8-10). So we used the transaction quality metric, now called Application Performance Index (Apdex), to show the same data more clearly (see BCR November 2003, pp. 8-10). The Apdex normalizes the response times into a uniform number from 0 to 1. Zero is total failure (none of the users are satisfied) and 1 is a perfect score (all the users are satisfied).

A true view of performance should not be across the complete user population. Some users in a nearby network will always see better performance than those on the other side of the globe. Table 1 shows the index of performance for the users of the financial site grouped into meaningful user sets--the major ISP carrier from which they originate. Each of these user groups is large enough to be meaningful to the business, yet focused enough such that the enterprise can do something about their performance.

Table 1 more clearly shows the dramatic difference between the synthetic services and the real user. The users on the SBC, Sprint and MCI networks are receiving performance that ranges from Poor to Good. However, the index reported by the synthetic agents is consistently Excellent. The real user measurements include the user's computer, user's LAN(s), user's access line, an aggregation network, the access network (listed above), peering point(s), financial site's backbone ISP and the datacenter servers. It is a long list of stuff that produces a wide range of response times.

Table 1 – Financial Site Apdex By Carrier				
	Real Users		Synthetic Agents	
Carrier	Apdex	Rating	Apdex	Rating
SBC	0.62 ₃	Poor	0.99 ₃	Excellent
Sprint	0.72 ₃	Fair	0.99 ₃	Excellent
MCI	0.92 ₃	Good	0.99 ₃	Excellent

Clearly, to get such low Index values, many users were often tolerating or frustrated, to use the lexicon we described above. However, the synthetic agents are hard-wired directly into major national backbone networks at 100 Mbps. So the agent tests run over the agent's computer, its backbone network, peering point(s), financial site's backbone ISP and the datacenter servers. Not measuring the "last mile" components has a huge impact on the results.

In fact, the difference in the range of Apdex values among the three carriers is additional evidence that the last mile is important. The SBC network has some dial-up customers while the MCI network is only used by corporate customers that are typically connected at 45 Mbps or higher. This difference in the respective last mile composition explains the 0.3 Apdex difference. However, even the high-end corporate users see the effect of their campus network when checking on their 401k plans, since their index is still not Excellent. The lesson is that many "well connected" users do not get satisfactory service.

Of the two synthetic measurement companies used in this analysis, Gomez does supply a last mile measurement service. However, it was not used by these three online businesses.

Conclusion

So what does all of this mean? We have two vastly different views that really need vastly different labels. It is good to go back to the Performance Framework described last year. It shows that there are two distinct motivations behind performance claims--asset management and experience management. The system's ability to establish new service, or recover failed service, is the relevant Provisioning function under asset management, while the Quality function of transaction

applications is the user's task response time under experience management. Synthetic agent services measure Provisioning, and real user measurements report on Quality.

In other words, synthetic agents are good for finding out if the doors to the store are open in Europe before the European customers wake up. But the real measurements are what you need in order to know if they are happy with the shopping experience.

The provisioning goals of knowing if a service is working, has recovered from a failure, expanding to a new geography that has no real users, rolling out a new service, etc., all need to be understood before the service goes live. The trap that enterprises fall into is thinking that they can then continue to rely on provisioning data as quality data. Remember that the crash test dummy only told you that the car passed minimal safety standards. It did not test if the gas tank catches on fire or the car has a tendency to roll over. You get

that kind of data from reading the accident reports generated by real people in real accidents.

Companies Mentioned

Gomez (www.gomez.com)

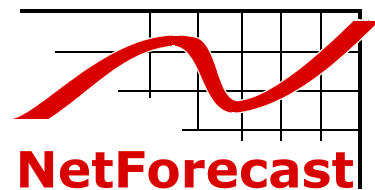
Keynote (www.keynote.com)

Note: The original version of this article in BCR magazine used the name NetForecast Response Index (NFRI) for the transaction quality metric. The name of the index has changed to the Application Performance Index (Apdex) as reflected in this version of the article.

Peter Sevcik is president of NetForecast 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 helps change delivery systems to improve the performance of networked applications. This includes advising enterprises on how to evaluate, improve and manage the performance of business applications, as well as advising vendors about customer requirements, technology issues, and adoption trends.

Smart Strategies From Hard Data



www.netforecast.com