NetForecast

Real vs. Synthetic Web Performance Measurements, a Comparative Study

By John Bartlett and Peter Sevcik
December 2004

Enterprises use today's Internet to find customers, provide them information, engage them in business transactions and in many cases to deliver products. The Internet has become an important business conduit for today's Enterprise. To manage our businesses, we need to insure our conduits perform well, and the first step in quality management is always to measure.

Two distinct methods of testing web-based performance exist in the market today, synthetic test services and user measurement appliances. The first approach distributes agents around the Internet, which are used to test an application or web site at periodic intervals, simulating the behavior of users. The appliance approach installs a measurement box at the data center, which measures and tracks the performance of each and every user visiting the site.

NetForecast has performed a comprehensive analysis of these two methodologies. While both approaches can provide IT executives and business managers with visibility into their application service levels from an end user's perspective, they differ significantly in their ability to deliver accurate response time information to represent the experience of application users.

Synthetic test services, by design, simulate a relatively small number of user transactions on a full-scale e-business web site. Synthetic measurements are based on simulated user transactions and are valuable for pre-production response time testing before real users are available to exercise applications. They also have value in a production environment when public benchmarking against competitive web sites is important for marketing purposes. However, IT support organizations should not base the management of application service delivery on small simulated user samples when it is possible to measure the real experience of every user.

Executive Summary

With respect to accuracy, our analysis shows that synthetic measurements randomly and frequently understate the actual response time seen by real users by as much as 50% and overstate the response time of other applications by as much as 350%. These error margins are much too broad to serve as the basis for infrastructure management or SLA enforcement. When comparing the results of real versus synthetic measurements over time, we also found poor correlation of response time trends. Establishing accurate baselines is an essential first step in proactive detection of service level degradation or traffic anomalies.

With respect to coverage, our analysis shows that synthetic agents usually interact with less than 10% of total website URLs and less than 15% of the ISPs connecting real users to the site. It is dangerous to extrapolate the performance characteristics of an entire application environment and its user communities from this small sample. In one site analyzed, we found that 190 synthetic agents were used to represent 1.7 million total website visitors.

The principle risk of using synthetic measurement for service level management is that service problems can impact an unknown quantity of users in unknown locations with an immediate and negative impact on the business. This can leave the IT organization without the data they need to detect and analyze service problems so they can take the necessary corrective actions to restore service.

NetForecast Report 5077

Application Performance

When we talk about application performance we are not referring to server load, hits per second or transaction counts. We are instead referring to the user experience; the user's view of the application as fast, slow, available, buggy or useful. Poor availability makes a user reluctant to waste time trying the application. Errored pages create user frustration. And poor response time at best reduces productivity, and at worst frustrates the user until he/she quits trying to use it.

A Network World survey done in August 2003 of companies with revenue above $1B annually found that 82% of responses indicated performance incidents have impacted user productivity. Postponing application launches because of poor application performance was reported by 51%. Many legacy business applications are being converted to web-based versions, which often have poor performance. Whether applications are supporting internal business processes, business to business interactions with partners or vendors, or reaching out directly to customers, application reliability and response time are critical to keeping business flowing smoothly.

Part of sorting out application performance in today's distributed world is understanding where users are located, and what are the performance problems of their specific situation. If critical users in Germany are getting poor response time, it doesn't help to know that the average customer is getting 2 second response. User demographics are a key part of a successful measurement tool.

Best Practices Measurement Techniques

The measurement goal is to measure the performance of web content and applications, to insure they are available, reachable and providing appropriate performance to all constituents.

Synthetic Test Services

One approach used to make these measurements is to use the services of a synthetic measurement service such as Keynote Systems and Gomez. These services have testing agents distributed throughout the Internet that are programmed to interact with an enterprise web site or application.

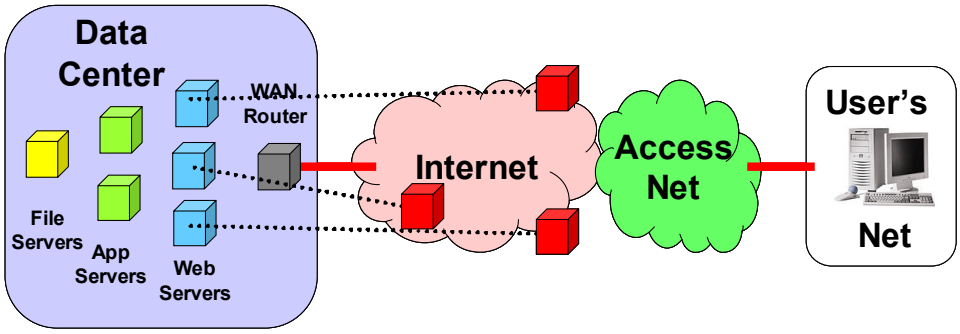

Figure 1 - Synthetic Measurement Service Deployment

Agents managed by the measurement service company are located inside primary Internet switching centers, distributed across the Internet (see Figure 1.) These agents often have 100Mbps connectivity to many carriers at that one location. Each server runs a scheduler, a multi-user browser or browser emulator, and data reduction software to compact the results and report them to a central collection site.

For each customer of the synthetic service a transaction script is built to emulate a user interaction. Selected servers then run this script at predetermined intervals, capture

metrics of the performance experienced, and report this data to a central site. Servers are chosen by geography or Internet connectivity to best serve the enterprise goals.

The central site then performs analysis on the data from all agents testing the customer's site. It creates averages, checks for thresholds and creates reports. These reports show detail by location, overall performance, detailed breakdown of page element timing, reliability and other metrics. Reports are available either via email or a web portal.

Real User Measurement Appliances

The second approach to application performance monitoring is to measure the performance of each user interacting with the site. Real user measurement tools such as the ITvisibility solution from Adlex, use appliances to passively monitor all client and data center interactions by attaching to a mirror port on the edge router at the data center. (See Figure 2) Each client transaction is captured and analyzed by the measurement tool. Furthermore, the tool can monitor and parse the standard HTTP and HTML protocols to reconstruct page and session details. The tool can then recreate the user experience by taking the page context and the round trip delay into account, and monitor response times of the client and server during the web transaction.

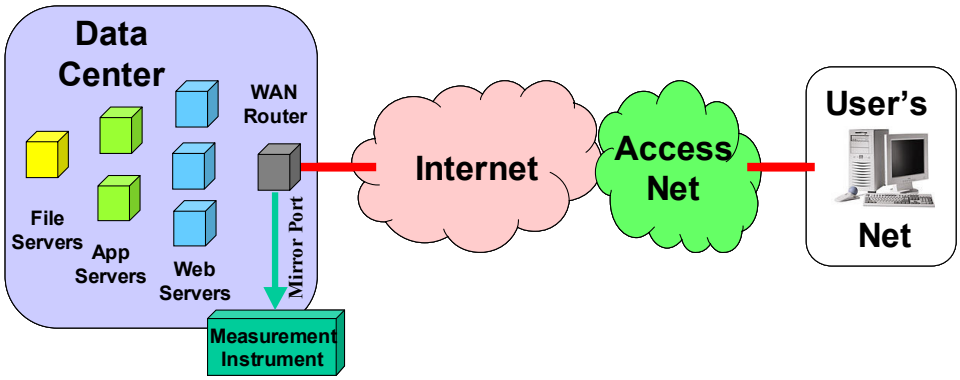

Figure 2 - User Measurement Appliance Deployment

Sophisticated data reduction algorithms are required for these tools since the quantity of data being collected is very large. Data is reduced and stored in a local database available for creating reports or doing specific queries to find specific performance metrics. Reports and charts are also derived from this database. Trending thresholds and alarms are implemented to provide early warning for performance or availability problems occurring in the infrastructure. Many real-user measurement tools, like the Adlex solution, take the process to the next level with drill down diagnostics to isolate and analyze the cause of performance problems.

Comparing Real and Synthetic Measurement

Given these two distinct approaches to application performance measurement, how do they compare? Is it possible to determine how well each approach meets business goals? NetForecast undertook a study to find these answers.

To collect data for this study we worked with Adlex and collected data from three active sites, a shopping site, a financial services site and a web-hosting site managing an online magazine. At all three sites one or more synthetic measurement services were in use during the measurement period. The synthetic agents were identified by their IP addresses so we could distinguish between real users and the synthetic agents. Measurement data for at least a week of time was collected from each site, and then analyzed. By separating the measurements of synthetic agents from measurements of real users we were able to predict the values the synthetic service would provide, and

compare that to measurements of actual users. The synthetic measurement services were assumed to accurately gather and report on their agents measurements.

Study Results

Results are summarized in two sections. The first section, Response Time Measurement, looks at the response times experienced by both users and synthetic agents, and compares the results each one reports. The second section, URL and User Coverage, examines how well the test methodology covers the functions of the web site being tested and how well the synthetic agents match the volume and distribution of actual users.

Application Response Time

First we look at the response time for users and synthetic agents, to see if their experience of application performance is equivalent. Response times were averaged over short intervals (5 or 15 minutes) and plotted for users as a group, and for the individual synthetic agent services. Figures 3, 4 and 5 show response time measurements for each of the three web sites measured, over their respective measurement periods. Normalized performance is shown on the Y-axis, where 1 represents the average performance experienced by users during the whole measurement period.

Figure 3 graphs response time for the financial web site. The upper line on the graph represents response time experienced by the users. The three lower lines show the results from three synthetic agent services. Note the substantial performance difference between synthetic agents and real users. The synthetic agents are indicating approximately half the response rate experienced by users.

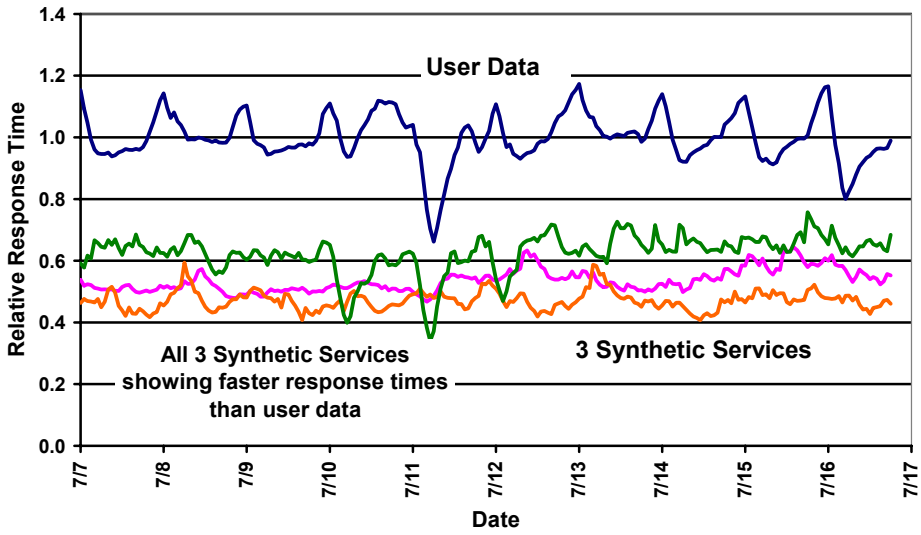

Figure 3 - Financial Site Relative Response Time

Figure 4 graphs the response time measurements from the shopping web site. Here we again see low readings from the two synthetic services. Additionally we note an event on July 1st, where users suddenly were experiencing almost twice the normal latency. Synthetic service #1 also experienced a slowdown, but only by about 30%. Synthetic service #2 shows no indication of a slowdown at all. Looking back at Figure 3 we see a similar story on July 11th, where a sudden drop in user response time is experienced by

one service but not by the other two. Peaks in response time are in some cases mirrored in one synthetic service, but the correlation is not very good.

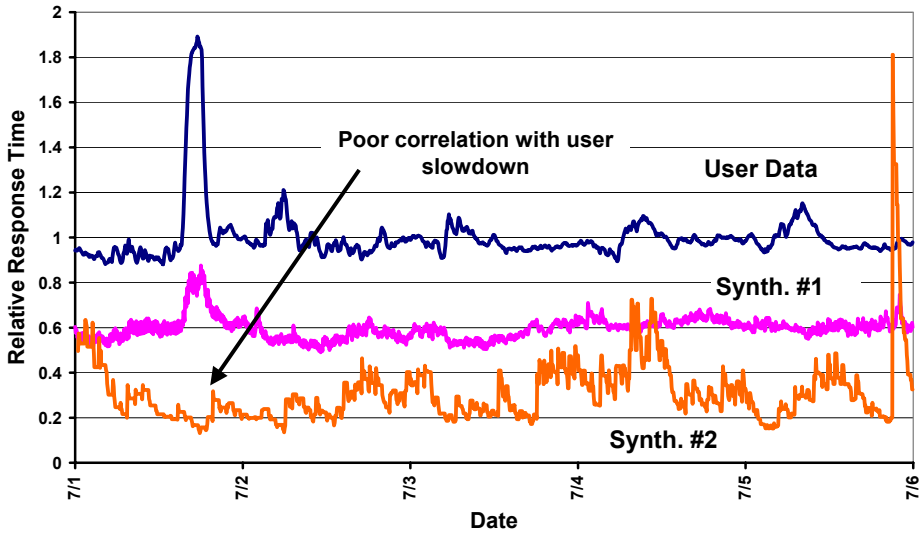

Figure 4 - Shopping Site Relative Response Time

In Figure 5 we see a very different situation where one synthetic service is showing 3½ times the response latency experienced by users for the first half of the test period. On July 14th the situation changes, and the synthetic service drops to be in line with user and other synthetic data. This may be due to a failure of an agent at the synthetic service, or a failure in its connectivity to the Internet site. The synthetic service has generated a false positive, a set of results that would ring alarm bells in the IT department, but turn out to be only a testing error.

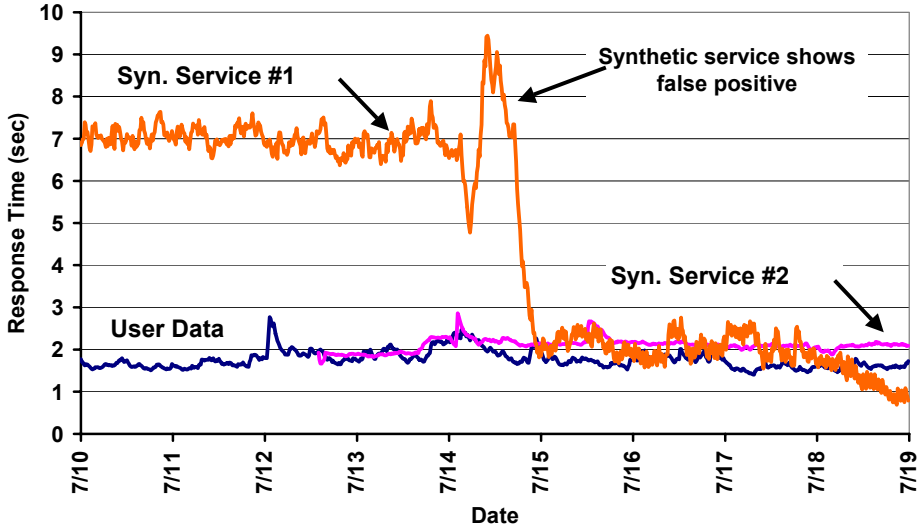

Figure 5 – Magazine Site Response Time (seconds)

Note also in Figure 5 that the second synthetic testing service is well aligned with user response data. Our experience with testing user vs. synthetic measurements is that the synthetic agents can report either faster, slower or correct response times depending on the dynamics of the particular site, connectivity, placement of the test agents and location

of the users. Another explanation for the large variation in the test results is that the synthetic agents may be testing different pages than the ones being delivered to users.

Response Time Distribution

The data in the three figures above (Figures 3-5) show response times averaged across all users for the specific time of day measured. Averaging is a useful way of aggregating large amounts of data, but it can also hide important details. In this section we look at the variation in response time experienced by a subset of users.

To get response time distributions we focused on users and synthetic agents connected to three Internet service providers. User and agent response time were categorized into three zones: satisfied, tolerating and frustrated users (see sidebar, **Categorizing Web Performance**.) Tables 1-3 show the percentage of users in each of these three categories.

Table 1 - Users Connected To Sprint

	Satisfied	*Tolerating*	*Frustrated*
Real Users	52%	40%	8%
Synthetic Agents	98%	2%	0%

Table 2 - Users Connected to SBCIS

	Satisfied	*Tolerating*	*Frustrated*
Real Users	49%	25%	26%
Synthetic Agents	99%	1%	0%

Table 3 – Users Connected to MCI

	Satisfied	*Tolerating*	*Frustrated*
Real Users	85%	14%	1%
Synthetic Agents	97%	3%	0%

These results show the very broad range of user response time experienced. Users connected to SBCIS experienced response times as short as 50 milliseconds, and as long as 35 seconds.

Results in the second line of each table show the performance distribution experienced by synthetic agents. This much narrower distribution of response times gives an overly optimistic view of the experience of users. Ninety nine percent (99%) of synthetic tests from SBCIS finished in less than 2 ½ seconds, whereas this only accounted for 48% of the real user measurements.

URL and User Coverage

URL Coverage

URL coverage measures the number of distinct URLs tested by synthetic agents and compare that count to the number of distinct URLs touched by actual users. URLs that are not tested may have poor performance or errors which would not be seen by synthetic tests, and thus not fixed until some other form of feedback occurs, such as a help desk call or user complaint.

Table 4- URL Testing Coverage

	Synthetic Agents	*User Measurement*
Shopping Site	9%	100%
Financial Site	10%	100%
Magazine Site	1%	100%

Table 4 shows that two sites test about 10% of their URLs, presumably the ones they consider the most important pages. The magazine site only shows 1% of its pages tested. This may be due to a different testing methodology, but may also result from a site that uses distinct URLs for individual users, generating user specific data for each client.

User Coverage

This comparison shows the number of testing agents and the number of distinct clients using the sites during the test period. The first column shows total site users during the measurement period, the second shows users of the most popular (top) application, and the third shows the number of distinct IP addresses associated with synthetic agents discovered during the measurement period.

Site Users	*Top App. Users*	*Syn. Agents*
1.7 M	762 K	190

Table 5 - User and Agent Counts

The ratio of synthetic agents to site users is about 1 to 10,000, whereas the ratio of synthetic agents to top application users is about 1 to 4,000.

User and Synthetic Agent Page Hits

The ratio of user transactions with the web site (hits) to synthetic agent transactions is shown in Table 6, for three source service providers. This means that the user and agent count is for users connected to that specific service provider only. The ratio of synthetic to user hits is shown in the fourth column of the table.

Table 6 - User and Synthetic Agent Page Hits From Three Service Providers

Service Provider	*User Hits*	*Syn. Agent Hits*	*Ratio Syn/User*
Sprint 1239	1,804,487	2,970	1/625
SBCIS 7132	4,623,641	7,737	1/588
MCI 11486	971,884	5,306	1/185

Summary and Conclusions

Statistical sampling methods are well understood and can report valid information about a large population. However, statistical methods rely on choosing samples randomly so that all user groups are represented in the sampled subset. Using a synthetic testing service at first appears to be sampling, but the analysis above shows otherwise.

URL and user coverage results show there is a major disparity in the way real and synthetic measurements represent the true delivery application service levels. Hidden problems in the provided information are not recognized or fixed without the visibility that full coverage supplies. User location average is not an accurate sample because test agents are located by geography and availability of good Internet connectivity in hosting locations, not by random sampling of user locations. These sites may reflect user connectivity for all Internet users, but probably don't accurately reflect user connectivity for your specific business. Furthermore, most agents do not reflect last mile speed limitations because of their backbone-connected location.

Response time inaccuracies shown in this report reflect the coverage issues described above, and further show how an agent's location and connectivity hide the broad range of user response time actually experienced. Only knowing that the average user is experiencing good performance may mask critical information about high value constituents in specific locations experiencing frustrating results.

When choosing a measurement strategy, first define your requirements for how these measurements will be used. If response time metrics will be used to detect and correct performance problems for any and all affected users, then real user data is essential. Similarly, if service level agreements (SLAs) are keyed to application response time, SLA infractions should be based on highly accurate and credible information. Real user measurement systems provide the only way to collect, analyze and report data with this level of accuracy.

Creating Application Performance Metrics

NetForecast uses a response time index based on research on how users perceive the speed of their interactions with a computer. Users are sensitive to the response time of each interaction rather than the complete process (many interactions). These interactions, called Tasks, are the time between when the user sends "enter" and the next useful or actionable response appears. In Web-based applications, tasks are Web pages. Users group their judgment of application responsiveness into the following three task response time zones.

Satisfied *– All responses in the satisfied zone are fast enough to satisfy the user, who is able to concentrate fully on the process, with no concern about performance. When asked about the application experience, the user may have a variety of comments or criticism but slow speed is not mentioned.*

Tolerating *– Responses in the tolerating zone are longer than those of the satisfied zone, exceeding the threshold at which the user notices how long it takes to interact with the system, and potentially impairing the user's productivity.*

Frustrated *– As response times increase, at some threshold the user becomes unhappy with slow performance and enters the frustrated zone. When frustrated a casual user may abandon the process and a production user is likely to stop working on the process.*

The threshold between satisfied-and-tolerating or maximum response time (seconds) permitted in order to keep the user satisfied, is determined by estimating the application's process repetitiveness and user interest. The next threshold between tolerating-and-frustrated is four-times the value of the previous threshold.

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.

Peter Sevcik is President of NetForecast and a leading authority on Internet traffic and performance. He has contributed to the design of more than 100 networks. Peter led the project that divided the Arpanet into multiple networks in 1984, which was the beginning of today's Internet. He can be reached at peter@netforecast.com.