The Interaction of Web Content and Internet Backbone Performance

Keynote Systems, Inc.
2855 Campus Drive
San Mateo, CA 94403
(650) 522-1000
www.keynote.com

Introduction

Optimizing Web site performance involves designing and optimizing hardware, software and page design. Optimizing page design is especially important for sites with complex or dynamic content—such as media, news and electronic-catalog shopping sites.

When designing fast-loading Web pages, Web architects follow some common principles, including:

- Smaller pages download faster
- Smaller graphics download faster than larger graphics
- Users don't like to wait for very large graphics to download

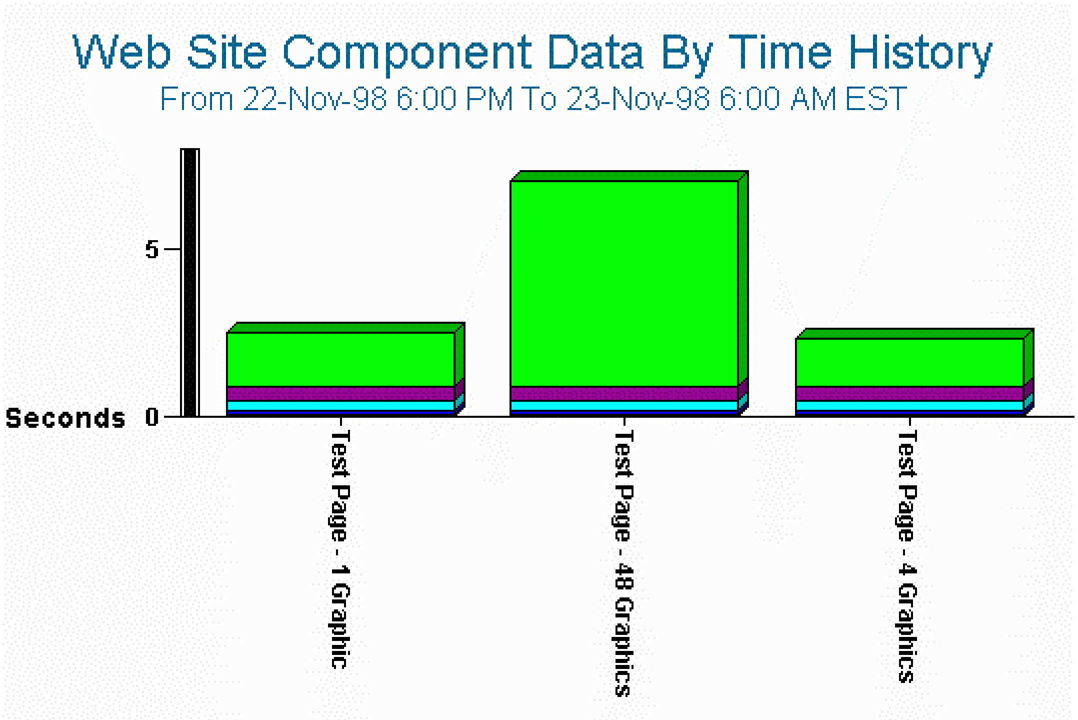
While these considerations are important, they don't capture all the ramifications of page design. Going a step further and understanding how Internet backbone performance, Web browser behavior and page design interact together can help you achieve dramatic performance improvements from your site.

Keynote Web Site Perspective Full Page Component Data helps Web architects optimize page design by providing data and analysis about the relationship between Web content and Internet network delays. By analyzing content download time into its constituent elements, Component Data helps you identify the exact source of delays, as well as provides guidelines for developing and verifying a performance improvement strategy.

The Effect of Graphics on Web Site Performance

The graph below shows the average download time for three pages between 6pm and 6am, which are ideal Internet hours for fast Web site performance due to low traffic volume. All three pages have almost identical benchmark page sizes (i.e., the HTML portion of the page or index.html file). The main difference in page composition is the size and number of the graphics. The first page had one large graphic, the second page had 48 small graphics and the last page had 4 medium-sized graphics. The total size of the graphics on all three pages is almost identical; the only significant difference is in the number of graphics on each page.

Each bar in the graph represents the total download time for one of the test pages. The different colored bands show the individual components of the download, starting with the DNS lookup time and ending with the time to download all the graphics. Download times were measured from over 45 locations on the Internet in different U.S. cities and on different backbones, and are averaged together to provide a composite time that represents the download time experienced by the average Internet user in the U.S.



Comparison of download time for three pages with different page composition.

The following table shows the graphics, bytes and download times for each page:

Page	**# of Graphics**	**Total Bytes**	**Download Time**
Test Page - 1 Graphic	1	67,081	2.5 seconds
Test Page - 4 Graphics	4	64,733	2.34 seconds
Test Page - 48 Graphics	48	60,545	7.03 seconds

This data shows that the page with the most graphics had the worst performance, even though it had the fewest total number of bytes. The reason is that Web browsers request each graphic individually, which results in many more packet round trips between the browser and the Web site compared to sending just a couple of requests for a small number of graphics. Roughly the same number of bytes are being downloaded in both cases, but retrieving a small number of larger graphics is more efficient than retrieving a lot of smaller graphics.

Keynote Component Data

Keynote component data can be used to show how Web page design is interacting with Internet delays and affecting end-user performance. This analysis can be done on your own pages as well as on any of your competitors' pages.

The table below shows the level of detailed information available in the Keynote Component service. For every page, each component of the download process is individually measured and shown. In addition to the summary view shown here, this level of detail is also available on a city-by-city or hour-by-hour basis.

All Time Ranges	Test Page - 1 Graphic		Test Page - 48 Graphics		Test Page - 4 Graphics	
Component	Avg. Time (secs.)	%	Avg. Time (secs.)	%	Avg. Time (secs.)	%
DNS Lookup	.06	2.25	.05	.66	.07	2.79
Initial Connection	.12	4.88	.15	2.09	.10	4.26
Redirection	0		0		0	
First Byte Download	.27	10.66	.28	4.00	.28	12.08
Base Page Download	.42	16.68	.44	6.26	.44	18.64
Content Download	1.65	65.52	6.12	86.99	1.46	62.23
Count	561		560		561	
Average Total Bytes	67081		60545		64733	
Total Measurement Time	2.52		7.03		2.34	

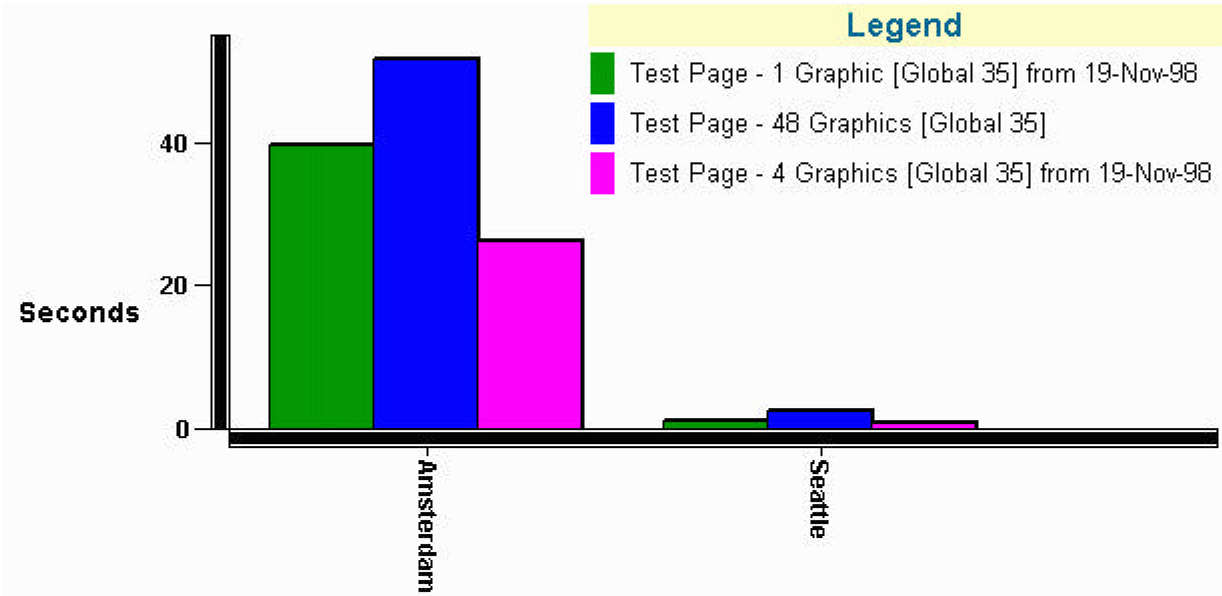
This data analysis can be used to determine how page design, Internet delay and browser operation all interact and affect performance. Web browsers download multiple images in parallel using up to 4 threads. Since there were 48 graphics in the test page, each of the 4 threads, on average, would need to download approximately 12 graphics. Each graphic requires a new network connection and GET request before it is downloaded. Normally this is an additional 2 round trips and only amounts to a total of .40 seconds, but when repeated 12 times within each thread to retrieve each of the small graphics, this results in the additional 4.8 seconds of download time: 12 x .40 = 4.8 secs.

The 4.8 seconds show up almost exactly in the difference in download times between the 48-graphic page and the 1-graphic page: 7.03 - 2.5 = 4.53 secs. To understand how the .40 seconds was calculated, look at the detail table above showing each of the page components. The table indicates that the average time to set up a connection is .12 secs (Initial Connection). The average time to request a new image is .28 seconds (First Byte Download). This shows that the overhead of requesting an additional image is .12 + .28 or .40 seconds.

This example illustrates the detrimental effect of even a small amount of network latency under ideal conditions. When network delays increase, such as when transmitting over international connections or when packet loss rates go up, this effect can significantly impact performance.

The Effect of Graphics and High Latency on International Performance

The following shows a more extreme example of a high latency connection during peak Internet hours when performance is the worst.



Page performance comparison between high latency and low latency conditions

This graph shows download times for the same three test pages from two different cities, Amsterdam and Seattle. The same software is downloading the pages. The only difference between these cities is the Internet delay experienced while downloading the test pages.

This example also shows that downloading 4 graphics is faster than a single graphic, especially over longer latency links such as from Amsterdam. The reason that 4 graphics downloads faster than a single graphic is that you are taking advantage of the multi-thread operation of the browser by downloading 4 graphics in parallel.

Summary

For Web sites with complex or fast-changing content, optimizing page design is critical to delivering fast Web site performance. To achieve the greatest performance improvements, page design optimization must take into consideration Internet backbone performance and Web browser behavior.

Full Page Component service measures the complete user experience of downloading Web site content – including graphics and banner ads - to multiple locations around the world, to show precisely the impact that worldwide Internet delays are having on your pages. Using this comprehensive data, you can modify your pages for optimal performance against internal goals and competitive benchmarks, and see the benefit immediately in your Keynote performance measurement. Using Full Page Component data you can:

- Identify the source of delays in your page download
- Determine how long graphics are taking to download
- Modify page design for optimal performance

Keynote Full Page Component service is ideal for businesses with fast-changing Web sites, helping them ensure that their content is optimally tuned to deliver the highest possible quality of service to their online customers, wherever they may be located on the Internet. Full Page Component service can be used in conjunction with Keynote's Benchmark Page service to precisely determine whether a performance problem is caused by a change in content or by a backbone/infrastructure problem.