

Accelerating E-Commerce

Net Forecasts – Peter J. Sevcik
BCR Volume 32, Number 11
November 2002

Internet e-commerce is going mainstream, as evidenced by a recent analysis we performed for a multi-national manufacturer headquartered in Europe. It operates Web-based supplier management, partner relations and customer ordering via the Internet.

The system operates in full production mode, and everything was working smoothly until users in Asia started to really rely on the services for daily business interaction. Suddenly, our client's web-based systems were no longer merely "interesting;" they evolved into a new way to do business and had become essential to the income of both the manufacturer and all its partners. Since the new systems enable faster interaction with much more information than was previously available, all the players want to leverage the new knowledge. But the service is terribly slow, and performance had to be addressed.

E-Commerce Application Profiles

A profile of the e-commerce applications showed that even though all interactions are performed by a browser and therefore are web-based, the traffic per interaction does not look like a typical Web page.

Specifically, the content starts with a base page, which is much larger and more complex than a typical Web page. This base page does not direct the browser to fetch and render many small

elements; instead, the browser has to execute complex style sheets or Java software. The user interacts with this rich content that comes with the base page, and which is used to develop a complex query to the system. The user then submits search, compare or reconcile requests to servers, which respond with large tables of answers or the next set of choices in the process.

The key is that the browser is no longer a simple, thin client; instead, we're moving to a thick-client/thin-server model. The amount of data that must be loaded into the browser varies widely, depending on the size of the data set that is being processed. This trend of sending unprocessed data for processing at the desktop will only keep making the content larger.

We use the Keynote Business 40 (KB40) Index to benchmark general business-to-business interactions. Its current application profile – 120,000 bytes of payload and 49 turns – is nearly the same in size as the e-commerce app (115,000 bytes) we were working with, but the complexity of the page is much lower with only 13 turns (see my article in *BCR*, October 2001, for an explanation of "payload" and "turns"). In this particular case, the browser is rarely redirected to other servers; there are no ads, no eye candy. This is serious business.

Table 1 – Response Times for Alternative Acceleration Techniques
Typical Page Load Time (sec)

	Default Service	CDN Service	Route Control	Compress Content
General B-B (KB40)				
Good Hours	19	9	16	17
Poor Hours	34	27	20	30
E-Commerce B-B				
Good Hours	12	10	10	7
Poor Hours	21	20	13	12

Broadband connected Asian user accessing server in Europe.

In addition, while the content we were studying is similar in size to the KB40, it is very different in composition. There are very few graphics (GIFF, JPEG) being loaded, which leaves almost all of the content amenable to compression. Indeed, about twice as much content per page is compressible in the e-commerce applications relative to the general Web page.

Acceleration Alternatives

While there are many ways to accelerate a Web page, it's important to distinguish between two basic classes of solutions – those that improve throughput and those that shorten time. Since the users complaining about poor performance were on the other side of the world – i.e., there was a lot of Internet latency between themselves and the servers – we focused on three techniques that impact what is going on in the network portion of the page-load time: A content delivery network (CDN), applying route control at the server farm and adding dynamic content compression at the server.

The CDN service delivered static content much faster, because there were servers in Asia. The latency difference between Europe and Asia was an order of magnitude better. However, the pages in the e-commerce app had very little static content, and most of what they had – e.g., the company's logo – could be cached on the user's desktop. Furthermore, the CDN service under evaluation did not provide consistently better latency to all the users in Asia and the Pacific. In fact, sometimes content was being delivered from the U.S., a big ocean away!

We performed tests from various critical cities in the region to determine the basic Internet performance that the users had to deal with back to Europe. The performance in most cities was bi-modal – generally acceptable with expected high latency, or very poor with extremely high latency and packet loss. The periods of poor performance lasted for hours and impacted a significant portion of the workday.

That seemed to fit the job description for a route controller; the problems weren't local and the server site was already multi-homed, providing good connectivity diversity to Asia.

It also was clear that the high level of textual data and Java software (also represented as text) were candidates for compression. Browsers delivered after 1998 (Explorer v4, Netscape v6 and Opera v5) can support HTTP 1.1, which includes the ability to perform “content encoding,” the W3C term for compression.

However, there was a complication with compression, because each page in this e-commerce application is unique; the pages supply information on data elements that are part of the user's profile and interest at the exact time the user makes the inquiry. And, sometimes the pages are designed dynamically based on the options that relate to the data being presented. Therefore, the compression must be performed dynamically on each page just as it is ready to leave the server.

While few web-server software platforms support dynamic compression, this void has been filled with third-party software and a new crop of server-side acceleration engines. These boxes sit in front of the web server and operate many communications functions – e.g., off-load the server for TCP termination, perform dynamic compression and add SSL processing if needed.

Performance Impact of the Solutions

We modeled three alternative approaches to four scenarios as shown in Table 1. The model is based upon the e-commerce profiles and the measured Internet performance between Asia and Europe. The results show that each solution has a place where it is better than the others (the yellow regions represent the best time for each scenario).

We also applied the analysis to more general B-B web interaction, as represented by the KB40. The reason for adding this analysis is to show how differently the profiles behave. Most people would say that since the difference in page size is only 4 percent (120,000 bytes in the KB40 vs. 115,000 in the e-commerce app being studied), the time to load either page should be nearly identical. However, as shown in the table, even under good conditions, the time to load these two types of pages is 12 or 19 seconds. There is a 58-percent penalty for having more turns in the general B-B page.

In the case of a general B-B site with lots of static content, the CDN is best when the Internet is

operating as expected. However, during a period of poor performance, although a CDN improves performance, the route controller delivers the page faster. This is because a lot of interaction still must be processed back to the home server.

The best solution, however, turned out to be content compression, because our client's e-commerce pages have such a large amount of compressible content. Interestingly, contrary to conventional wisdom, compression had little effect on the general Web page, despite the fact that it has half as much content to compress as the e-commerce page.

The reasons have to do with TCP windowing; simply put, in the general Web design, small, one- or two-packet elements get sent before TCP has a chance to properly set windows. So, no flow operates. Compressing small elements only makes for many smaller elements, so TCP flow control continues to not engage and, as shown in the table, compression does not have the desired effect.

But the performance of the e-commerce application was entirely different story – compression worked great. Those who argue that compression has no value in broadband-based B-B applications paint compression's disadvantages with too broad a brush.

Conclusion

Performance can and should be accelerated to insure the success of any e-commerce application. However, there is no single, simple solution to *all* performance problems. While both the solutions discussed above and others not discussed can be applied in combinations to improve response time even further than shown in Table 1, some combinations conflict with each other and there is often diminishing return on the larger investment.

Enterprises need to carefully select an acceleration strategy based upon measured profiles and network performance. There is no substitute for gathering hard data and performing careful analysis before making a decision about performance.

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.

