Performance Analysis and Models of Web Traffic

Federico Montesino Pouzols
University of Seville, Spain

Angel Barriga Barros
University of Seville, Spain

Diego R. Lopez
RedIRIS, Spain

Santiago Sánchez-Solano
CSIC - Scientific Research Council, Spain

INTRODUCTION

The Internet and, more specifically, Web-based applications now provide the first-ever global, easy-to-use, ubiquitous and economical communications channel. Most companies have already automated their operations to some extent, which enhances their ability to interact with other companies electronically. With the advent of Web services, the interaction between companies becomes easier and more transparent (Khalaf, Curbera, Nagy, Tai, Mukhi, & Duftler, 2005).

Web-based technologies are extensively employed and support core components of virtual and networked organizations. Many of them, including for instance Web-based communities, heavily rely on Web traffic. Additionally, Web technologies play a central role in the technologies for supporting industrial virtual enterprises (VE) being developed by the National Industrial Information Infrastructure Protocols Consortium (NIIIP).

Thus, modelling and analysis techniques for Web traffic become important tools for performance analysis of virtual organizations (Malhotra, 2000; Foster, Kesselman, & Tuecke, 2001). This article overviews current models of Web traffic as well as performance analysis of Web-based systems.

BACKGROUND

Wold Wide Web, (WWW or Web henceforward) traffic is conveyed in HTTP transfers. HTTP is a key component for Web services and is based on the TCP transport protocol (Khalaf, Curbera, Nagy, Tai, Mukhi, & Duftler, 2005). This article overviews models, techniques and tools to analyze Web traffic at network nodes. In particular, we describe the discovery of self-similarity in Web traffic at macroscopic scale which has provided a mathematical framework for analyzing Web and Internet traffic in general.

The Web can be viewed as an example of a very large distributed and dynamic system with billions of pages resulting from the uncoordinated actions of millions of individuals. Despite this complete lack of central control, the graphical structure of the Web is far from random and shows properties shared with other complex graphs found in social, technological, and biological systems (Krisnamurty & Rexford, 2002).

Examples of invariant properties include the power-law distribution of vertex connectivities and the small-world property (any two Web pages are usually only a few clicks away from each other). Similarly, predictable patterns of congestion have also been observed in Web traffic. Many of these characteristics are also shown to a varying extent by other common Internet applications, such as e-mail and FTP. While the exploitation of these regularities allows for performance optimization of Web supporting systems and thus can be beneficial to providers and consumers; their mere existence and discovery has become a topic of basic research (Arlitt, 2000).

Traffic models and workload characterization are frequently necessary in order to better understand the performance of a Web site. Typical workload questions include: What do requests look like? How popular are some documents versus others? How large are Web

transfers? What level of HTTP protocol deployment exists on the Web? (Baldi, Frasconi, & Smyth, 2003).

An important caveat is that any one Web site may not be representative of a particular application or workload. For example, the behavior of a very dynamic Web site which hosts a great deal of rapidly changing content is most likely very different from an online trading site which conducts most of its business encrypted using the secure sockets layer (SSL). Dynamic content (Iyengar, Nahum, Shaikh & Tewari, 2005) has become a central component of modern transaction-oriented Web sites. While dynamic content generation is clearly a very important issue, there is currently no consensus as to what constitutes a representative dynamic workload.

A number of studies on Web server workload characterization (Arlitt & Williamson, 1996; Liu, Niclausse & Jalpa-Villanueva, 1999) have been performed through measurement on academic networks, scientific research organizations and commercial Internet providers. As a result, some invariants have been identified and models for WWW traffic have been proposed (Baldi, Frasconi & Smyth, 2003).

WEB TRAFFIC MODELS

Traffic models and workload characterization for WWW can be considered at various levels: IP packets (network layer), TCP connections (transport layer), HTTP messages and user sessions (transaction and application layers). Since self-similarity and long-range dependence was first discovered in Internet traffic (Leland, Taqqu, Willinger, & Wilson, 1994) a number of application traffic patterns have been identified to show some degree of self-similarity. Self-similarity in WWW traffic can be explained based on the underlying distribution of transferred document sizes, the effects of caching and user preference in file transfer, the effect of user "think time", and the superimposition of many such transfers in a local area network (Crovella & Bestavros, 1997).

Since self-similarity and long-range dependencies were first identified in Internet traffic, a number of empirical studies (Park & Willinger, 2000; Karagiannis, Molle, & Faloutsos, 2004) of high-resolution traffic measurements have provided evidence that network traffic is self-similar or fractal in nature. The dependence structure of Internet traffic in general and Web traffic in particular is, however, such a complex phenomenon that no model is entirely satisfactory. Additional modeling approaches have been proposed and recent studies have developed and validated nonstationary poisson models (Karagiannis, Molle, Faloutsos, & Broido, 2004). Among other modeling approaches, queueing analysis (Daigle, 2004) is a well-established field and complementary engineering tool for network traffic analysis and server load characterization (Rolls, Michailidis, & Hernández-Campos, 2005).

At the HTTP layer, generic patterns for relative frequencies of HTTP methods and response codes have been identified. Document or object popularity and frequency of request have been found to obey Zipf-like distributions. Reference locality has been demonstrated as well (Almeida, Crovella, Bestavros, & de Oliveira, 1996; Iyengar, Squillante, & Zhang, 1999; Molina, Castelli, & Foddis, 2000).

The distributions of document sizes and transfer sizes is hard to characterize in a generic manner. There is consistent agreement that sizes range over multiple orders of magnitude and that the body of the distribution is log-normal. The shape of the distribution tail may, however, obey Pareto, Log-Normal and other distributions (Almeida, Crovella, Bestavros, & de Oliveira, 1996; Crovella & Bestavros, 1997; Iyengar, Nahum, Shaikh, & Tewari, 2005).

Application-level measurements are needed for a clear view of overall application performance, which cannot easily be synthesized from lower level data. They may also offer some insights into the performance of the client and server hosts, and of the network links between. However, although Web transactions may be thought of as a network service, measuring them gives only an indirect view of underlying network behavior. Effects of network performance on WWW response times is currently an important topic of research (Andrews, Cao, & McGowan, 2006). New challenges in modelling Web traffic also arise because Web traffic has moved from delivery of text and image content to include audio and video streaming (Krisnamurty & Rexford, 2002).

At the user session level, numerous invariant session characteristics have been identified. These characteristics include the number of requests per session and the number of different documents requested per session. Statistical distributions of session length and intersession times have been also identified (Arlitt, 2000; Cockburn & McKenzie, 2002). Additionally, a number of methods to utilize these invariant characteristics in