RESEARCH STATEMENT

Hyoung-Kee Choi

In past few years, we have witnessed an explosive growth of the World Wide Web (Web) traffic in terms of the number of hosts, users, and application-level services. This fast growth rate of the Web has led to a dramatic increase in Internet traffic, as well as a significant degradation in user-perceived latency while accessing the Web. This has spawned significant research efforts aimed at improving Web performance. Fundamental to improved Web performance is a solid understanding of Web traffic characteristics. Systems designed without a proper understanding of traffic patterns can result in unpredictable outputs and biased performance under certain situations. Consequently, it is important that we have a deep and broad understanding of this important source of traffic.

1. Graduate Research

In the doctoral thesis, I address three key steps in the understanding process of the Web: Web traffic measurement, Web traffic characterization, and Web traffic modeling. These three steps are equally important as some of the results used in each step are incorporated into analysis in the other steps. I attempt to monitor it for a sufficient amount of time, analyze the collected data to characterize the behavior, and come up with an accurate model. The traffic model developed in this study must behave similarly to real Web traffic. To reach this stage, it is required to characterize the current types of interactions over the Web. I looked at the long-term behavior of Web traffic to infer distinct characteristics of Web traffic. The results obtained can be used in generating synthetic Web traffic for simulations or providing a better understanding of Web traffic to improve its performance.

A set of requirements must be satisfied in this measurement for an interpretation of the characterization and model to be meaningful. Among numerous requirements the most critical one is the accuracy of empirical data. A number of studies including our study are based on a Web page or a similar entity to study the Web traffic. Many observations take place within a Web page to characterize Web traffic: for example, the duration to download Web pages and the duration a user views Web pages. Hence, a method to separate Web pages in one user’s browsing session greatly affects the accuracy of the Web characterization. I took advantage of HTTP header information in addition to TCP/IP header information to separate Web pages. About 97 percent of accuracy was attained in determining the boundary of Web pages.

Web traffic is characterized in two different layers; HTTP and TCP. I examined transactions associated with a single Web-page retrieval and selected a set of primary parameters that could be used to define the Web traffic model and secondary parameters that helped to understand the behavior of Web traffic. A “keep-alive” connection is studied in the HTTP layer. A keep-alive connection in HTTP was developed as a means for reducing latency in transferring a Web page. Performance improvement of HTTP by adapting keep-alive connections has been shown in previous work. However, the extent to which users actually benefit from a keep-alive connection is not measured yet. This study evaluates HTTP performance with real traffic by characterizing how many objects are actually delivered on keep-alive connections. A transient behavior of TCP is studied in the TCP layer. Before reaching to a steady state, TCP alternates between active and
inactive periods of transmitting data segments. After a TCP sender transmits an entire window-size worth of segments in a burst, it pauses until the first ACK in the burst returns. I developed a method to distinguish different bursts in a connection. Based upon the burst boundary TCP is characterized by: 1) defining a period between the starts of adjacent bursts and 2) measuring the number of segments transmitted and the time spent in this period.

A Web traffic model is constructed based upon knowledge learned from the Web traffic characterization. With an accurate traffic model we can simulate the performance of new system designs and experiment with alternative designs. One feature of Web traffic is “self-similarity”: that is, traffic can show significant variability over a wide range of time scales. Self-similarity in traffic has been shown to have a significant impact on network performance, so it is important feature to capture in the Web traffic model. I examined the closeness between the synthetic traffic and the real traffic in a number of ways. In particular, the synthetic traffic exhibited strong self-similar characteristics.

The study of Web traffic was used to evaluate performance of two Medium Access Control (MAC) protocols for Internet access. The first application of the model is a performance evaluation of a satellite MAC protocol to deliver two-way Web traffic. The second application is for a MAC protocol of a Wireless Local Loop (WLL) system. The performance evaluation was implemented in a simulation with our model. Our Web traffic model provided detailed information that evaluated the proposed MAC protocols to be optimized. In particular, the satellite MAC protocol was developed into a product after examining its performance through the simulation.

Six specific research questions that I try to answer throughout the thesis are:
1) How accurate can we parse empirical data from the measurement of Web traffic?
2) What are the primary components of Web traffic?
3) How much do users benefit from a keep-alive connection in a new version of HTTP?
4) How has the Web traffic changed over time?
5) What is the accurate manner to generate Web traffic synthetically?
6) Where can this study of Web traffic be applied?

2. Future Research Interests

In the short term, I plan to continue working on problems in accurate measurement, rigorous characterization, and detailed modeling of the Web traffic. I would also like to extend the modeling technique to other popular source of traffic in Internet.

The Internet is evolving rapidly, giving rise to new challenges and opening up exciting areas of research such as Internet QoS, IP telephony, Web caching, contents delivery, multimedia transmission and so on. My goal in the long term would be to expand my research to such emerging areas, develop an understanding of the issues involved, and then build scalable and practical solutions.