Chapter 9

A Tutorial on Network Latency and Its Measurements

Minseok Kwon
Rochester Institute of Technology, USA

ABSTRACT

Internet latency is crucial in providing reliable and efficient networked services when servers are placed in geographically diverse locations. The trend of mobile, cloud, and distributed computing accelerates the importance of accurate latency measurement due to its nature of rapidly changing locations and interactivity. Accurately measuring latency, however, is not easy due to lack of testing resources, the sheer volume of collected data points, the tedious and repetitive aspect of measurement practice, clock synchronization, and network dynamics. This chapter discusses the techniques that use PlanetLab to measure latency in the Internet, its underlying infrastructure, representative latency results obtained from experiments, and how to use these measure latencies. The chapter covers 1) details of using PlanetLab, 2) the Internet infrastructure that causes the discrepancy between local and global latencies, and 3) measured latency results from our own experiments and analysis on the distributions, averages, and their implications.

INTRODUCTION

Internet latency is crucial in providing reliable and efficient networked services such as online retails (e.g., Amazon), multimedia streaming (e.g., Netflix), and social networking (e.g., Twitter). For example, Netflix runs its servers on the Amazon cloud in geographically diverse locations, and provides video streams from the server that can deliver the content to a client in the shortest time. In order to support this server selection in distributed computing, measuring accurate latency becomes extremely important. The trend of mobile computing like iPhone and Android-based smartphones only accelerates the importance of accurate latency measurement due to its nature of rapidly changing locations and interactivity. Accurately measuring network latency, however, is not an easy task due to lack of testing resources, the sheer volume of collected data points, the tedious and repetitive aspect of measurement practice, and

DOI: 10.4018/978-1-4666-8662-5.ch009
clock synchronization. In addition, the time that latency is measured affects measurement results significantly due to network dynamics, volatile traffic conditions, and network failures. Hence, it is critical to measure latencies over a wide span of time and days to acquire a representative and comprehensive view.

In the literature, latency measurement has been studied extensively. For accurate measurements, end hosts at strategic locations are used (Francis, et al., 2001), latency is estimated based on coordinated using landmark hosts (Ng & Zhang, 2001) or DNS queries (Gummadi, Saroiu, & S., 2002), or routing topology (Dabek, Cox, Kaashoek, & Morris, 2004) (H. Madhyastha, 2006). Measured latencies also help select target servers to achieve specified optimization goals in routing, content distribution networks, and cloud computing (Wendell, Jiang, Freedman, & Rexford, 2010) (Khosla, Fahmy, & Hu, 2012) (Ding, Chen, T., & Fu, 2010). Their objective function is to minimize the total execution time and select best replica servers considering server loads, cost and locations. Recently, mobile-cloud computing has gained considerable attention in which a mobile device offloads its computation to servers in the cloud seamlessly and transparently. Again, choosing right servers is critical to lower execution time, and we can use profiling that estimates execution time as a function of latency, loads, program, and network conditions.

The primary focus of this chapter is to discuss the techniques that use PlanetLab (PlanetLab, 2014) in order to measure latency in the Internet, its underlying infrastructure, representative latency results obtained from experiments, and how to use these measure latencies. As PlanetLab provides a world-scale network testbed, it has become a popular platform for network measurement. Based on our own experience, we discuss details regarding PlanetLab: 1) the fundamentals of PlanetLab, 2) how to establish a PlanetLab site and connect to its network, 3) how to manage an account, nodes, and slices, 4) how to run a server and manage tens or hundreds of servers, and 5) how to collect measurements like latency. Our results imply significant discrepancy between global and local latencies. Here, global latencies refer to latency of a long distance connection like transoceanic network link while local latencies are network delay between regional end-hosts. We look into the causes that make such discrepancy; this helps understand the underlying cause. Additionally, we present measured latency results from our own experiments and analysis on the distributions, averages, and their implications.

This chapter is organized as follows: First, we discuss the basics of PlanetLab and using PlanetLab for network latency measurement purposes. Second, we discuss the underlying network infrastructure that causes different latency measurements for global and local connections. Third, we report measured latencies from our experiments and analyze their causes and effects. Finally, we conclude this chapter with discussion on using measured latencies for applications and systems.

**LATENCY VS. THROUGHPUT**

**What Is Latency?**

It would be ideal if data in the Internet services can move from one node to another instantaneously at the speed of light. This is, unfortunately, not possible in reality. As a packet travels from a host (the sender) to subsequent routers or another host (the receiver), the packet experiences different types of delays at hosts, routers, and network links. Those delays include 1) processing delay, 2) queuing delay, 3) transmission delay, and 4) propagation delay. A host or a router needs time to process an incoming packet (processing delay) for packet forwarding such as reading a packet header and searching the routing table to determine the next hop. The packet also often needs to wait in the queue to be transmitted