Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Computer Networks 52 (2008) 1343–1364
Computer Networks
www.elsevier.com/locate/comnet

Practical large-scale latency estimation

Michal Szymaniak [a,*], David Presotto [b], Guillaume Pierre [a], Maarten van Steen [a]

[a] *Vrije Universiteit Amsterdam, Department of Computer Science, De Boelelaan 1081A, 1081HV Amsterdam, The Netherlands*
[b] *Google Inc., Mountain View, CA, United States*

Received 17 July 2006; received in revised form 21 May 2007; accepted 20 November 2007
Available online 29 January 2008

Responsible Editor: I.F. Akyildiz

Abstract

We present the implementation of a large-scale latency estimation system based on GNP and incorporated into the Google content delivery network. Our implementation employs standard features of contemporary Web clients, and carefully controls the overhead incurred by latency measurements using a scalable centralized scheduler. It also requires only a small number of CDN modifications, which makes it attractive for any CDN interested in large-scale latency estimation.

We investigate the issue of coordinate stability over time and show that coordinates drift away from their initial values with time, so that 25% of node coordinates become inaccurate by more than 33 ms after one week. However, daily re-computations make 75% of the coordinates stay within 6 ms of their initial values. Furthermore, we demonstrate that using coordinates to decide on client-to-replica re-direction leads to selecting replicas closest in term of *measured* latency in 86% of all cases. In another 10% of all cases, clients are re-directed to replicas offering latencies that are at most two times longer than optimal. Finally, collecting a huge volume of latency data and using clustering techniques enable us to estimate latencies between globally distributed Internet hosts that have not participated in our measurements at all. The results are sufficiently promising that Google may offer a public interface to the latency estimates in the future.

Keywords: Large-scale distributed systems; Network modeling; Latency estimation

1. Introduction

Modern large-scale distributed applications can benefit from information about latencies observed between their various components. Knowing such latencies, a distributed application can organize its operation such that the communication delays between its components are minimized [1–3]. For example, a content delivery network can place its hosted data such that its clients are serviced at their proximal datacenters [4,5]. In addition to improving the client-experienced latency, reducing the overall length of client-to-replica network paths allows one to localize the communication, leading to lower backbone and inter-ISP link utilization. Analogous benefits can be achieved for other large-scale distributed applications such as peer-to-peer overlays or online gaming platforms.

---

[*] Corresponding author. Tel.: +31 20 598 7748; fax: +31 20 598 7653.
E-mail address: michal@cs.vu.nl (M. Szymaniak).

1389-1286/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2007.11.022

The effectiveness of latency-driven techniques in improving the application performance depends on the accuracy of the latency information. A simple solution consists of periodically probing each latency the application needs to know [6]. However, such an approach makes sense only in relatively small systems, as continuous probing of pair-wise latencies is clearly not feasible when the number of nodes is very large. For example, re-directing clients to their nearest datacenters would require Google to maintain latency information from virtually every Web client in the Internet to each of its datacenters [7]. Also, the high dynamics of the Internet causes recently measured latencies to not always be a good indication of their current counterparts, as one latency measurement result is not a good predictor of a subsequent identical measurement. These two problems drive the need for scalable and accurate latency estimation techniques.

A promising approach to the problem of scalable latency estimation is GNP, which models Internet latencies in a multi-dimensional geometric space [8]. Given a small number of "base" latency measurements to a number of dedicated "landmark" nodes, GNP associates each node with its coordinates in that space. The latency between any pair of nodes can then be approximated with the Euclidean distance between their corresponding coordinates. What makes GNP scalable is the constant low number of measurements necessary to position each machine, which enables GNP to estimate all-pair latencies between a large number of machines at low cost.

The attractiveness of GNP has resulted in its various aspects being investigated for several years. However, whereas numerous theoretical properties of GNP have been described in detail [9–14], no GNP implementation has been demonstrated to work in a large-scale environment of a commercial content delivery network.

The common property of existing GNP implementations that hinders their deployment is active participation of positioned nodes, which are responsible for measuring and propagating their own base latencies [15–18]. Such an approach has several disadvantages. First, it introduces problems with malicious nodes lying about their base latencies. Handling such nodes is usually very hard, and typically comes at the expense of increased system complexity. Second, independent measurements of base latencies performed by many active nodes might overload both the network and the landmarks. This, in turn, might lead to numerous measurement inaccuracies affecting the GNP performance. Finally, active participation typically requires that some special positioning software is deployed on a significant fraction of positioned nodes. This condition might be infeasible to meet, for example, in content delivery networks, where most nodes are unmodifiable third-party Web browsers.

This article presents a GNP implementation that addresses all these issues. Our solution is based on two key observations. First, instead of relying on remote nodes to measure and report their base latencies, one can simply trigger some standard application-level communication between these nodes and the landmarks, allowing the latter to measure latencies passively on their side. This eliminates the need for customizing the remote nodes and ensures the integrity of measurement results. Second, instead of allowing remote nodes to independently perform their measurements, one can trigger measurements individually using a central, yet scalable, scheduler. This prevents landmarks from overloads and reduces the overall network overhead in general, as the scheduler triggers only the measurements that are really necessary. We demonstrate the feasibility of our approach by incorporating GNP into the content delivery network operated by Google, which enables us to position millions of Google clients.

Compared to the previous GNP implementations, our approach has several advantages. First, it greatly facilitates system deployment, as only the landmarks and the scheduler need to be instrumented. Second, it removes the problem of malicious nodes, as all the instrumented nodes are kept under full control of Google. Third, it eliminates the risk of overloading the landmarks, as the scheduler effectively adjusts the measurement volume to the landmark capacity.

Implementing our system at the scale of millions of clients requires one to address a number of subtle issues. For example, it is necessary to transparently and efficiently schedule measurements such that they do not affect the client-perceived browsing performance. Also, implementing a centralized scheduler is far from trivial when millions of Web clients are serviced by thousands of globally-distributed Web servers [19]. Finally, producing GNP coordinates that can remain representative for a long time requires that some special preprocessing techniques are applied to base latencies.

Within the first 2 months of operation, our positioning system performed more than 75 million