Taming the Dynamics of Distributed Data

Krithi Ramamritham
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
krithi@iitb.ac.in

The Internet and the web are increasingly used to disseminate fast changing data such as sensor data, weather information, stock prices, sports scores, and even health monitoring information. These data items are highly dynamic, i.e., the data changes continuously and rapidly, streamed in real-time, i.e., new data can be viewed as being appended to the old or historical data, and aperiodic, i.e., the time between the updates and the value of the updates are not known a priori. Increasingly, more and more users are interested in monitoring such data for on-line decision making. Traditional dissemination methods involve a pull or a push of data between a source of the data and a client. However, resource limitations at the source limits the number of users that can be served directly by it.

A natural solution to this is to have a set of repositories which replicate the source data and serve it to geographically closer users. Services like Akamai and IBM's edge server technology are exemplars of such networks of repositories, which aim to provide better services by shifting most of the work to the edge of the network (closer to the end users). Although such systems scale quite well, when the data changes rapidly, the quality of service at a repository farther from the data source will deteriorate. In general, replication can reduce the load on the sources, but replication of time-varying data introduces new challenges. Unless updates to the data are carefully disseminated from sources to repositories (to keep them coherent with the sources), the communication and computation overheads involved can result in delays as well as scalability problems, further contributing to loss of data coherence.

In situations where the data is to be used for on-line monitoring or online decision making, users specify the bound on the tolerable imprecision associated with each requested data item, this can be viewed as *coherence* requirement associated with the data. The coherence requirements associated with a time-varying data item depend on the nature of the item and user tolerances. For example, a user involved in exploiting exchange disparities in different markets or an on-line stock trader may impose stringent coherence requirements (e.g., the stock price reported should never be out-of sync by more than one cent from the actual value) whereas a casual observer of currency exchange rate fluctuations or stock prices may be content with a less stringent coherence requirement.

What is needed is a dynamic data distribution system that is *coherence-preserving*, i.e., the delivered data must preserve associated coherence requirements, and *resilient*, i.e., the system should be resilient to failures. Needless to say, it should work effectively with the minimal provisioning of resources.

Given such a data dissemination network, users can execute queries over distributed data by obtaining the data required for the query from one or more data repositories in the network. How this mapping – between query needs and data repositories – is done, depends on (a) the coherency of the data available at each repository and the precision requirements associated with the query, (b) the dynamics of the data, etc. Solving this optimization problem turns out to be challenging, especially since the mapping must be revisited as data characteristics change.

In this talk we will present work done at IIT Bombay towards solving these problems and also relate to ongoing work on sensor networks and stream processing systems.