Millionths of a Second Can Cost Millions of Dollars: A New Way to Track Network Delays

August 20, 2009

Daniel Kane

Computer scientists have developed an inexpensive solution for diagnosing delays in data center networks as short as tens of millionths of seconds-delays that can lead to multimillion-dollar losses for investment banks running automatic stock trading systems. Similar delays can delay parallel processing in high performance cluster computing applications run by Fortune 500 companies and universities.

Computer scientists have developed an inexpensive solution for diagnosing networking delays in data center networks as short as tens of millionths of seconds-delays that can lead to multimillion-dollar losses for investment banks running automatic stock trading systems.

University of California, San Diego and Purdue University computer scientists presented this work on August 20, 2009 at SIGCOMM, the premier networking conference.

The new approach offers the possibility of diagnosing fine-grained delays-down to tens to microseconds-and packet loss as infrequent as one in a million at every router within a data center network. (One microsecond is one millionth of a second.) The solution could be implemented in today's router designs with almost zero cost in terms of router hardware and with no performance penalty. The UC San Diego and Purdue University computer scientists call their invention the Lossy Difference Aggregator.

"A lightweight network monitoring approach such as ours allows you to pinpoint the source of the performance degradation and identify the problem routers," explained SIGCOMM 2009 paper author Kirill Levchenko, a UC San Diego post-doctoral researcher who recently earned his Ph.D. in computer science at UC San Diego.

"This is stuff the big traders will be interested in," said George Varghese, a computer science professor at the UC San Diego Jacobs School of Engineering and an author on the SIGCOMM paper, "but more importantly, the router vendors for whom such trading markets are an important vertical."

If an investment bank's algorithmic stock trading program reacts to information on cheap stocks from an incoming market data feed just 100 microseconds earlier than the competition, it can buy millions of shares and bid up the price of the stock before its competitors' programs can react, the computer scientists say.

While the network links between Wall Street and investment banks' data centers are short, optimized and well monitored, the performance of the routers within the data centers that run automated stock trading systems are difficult and expensive to monitor. Delays in these routers, also known as latencies, can add 100s of microseconds, potentially leading to millions of dollars in lost opportunities.

"Every investment banking firm knows the importance of microsecond network delays. Because routers today aren't capable of tracking delays through them at microsecond time scales, exchanges such as the London Stock Exchange use specially crafted external boxes to track delays at various key points in the data center network,"

said Alex Snoeren, a computer science professor at the UC San Diego Jacobs School of Engineering and an author on the SIGCOMM paper.

But these external systems are generally too large and expensive to be added to every router in a data center network running an automated stock trading system. This makes it difficult for the network managers to identify and locate problematic routers before they cost the company large amounts of money, the computer scientists say.

"Our hope is that this approach will allow router vendors to add fine scale delay and loss tracking, at almost zero cost to router performance, perhaps obviating the desire for expensive external network monitoring boxes at every router," said Ramana Kompella, the first author on the SIGCOMM paper and a computer science professor at Purdue University. Kompella earned his Ph.D. in computer science at UC San Diego in 2007.

The SIGCOMM 2009 paper presents simulations and proof-of-concept code for measuring latencies down to tens of microseconds and losses that occur once every million packets.

"The next step would be to build the hardware implementation, we are looking into that," said Kompella, who plans to continue pioneering research in fault diagnosis at Purdue.

This work highlights a fundamental shift happening across the Internet. As computer programs-rather than humans-increasingly respond to streams of information moving across computer networks in real time, millionths of seconds matter. Algorithmic stock trading systems are just one example. Extra microseconds of delay can also mean slower response times across clustered-computing platforms, which can slow down computation-intensive research, such as drug discovery projects.

"When it comes to fault isolation, networks are a big black box. You put packets in on one side and you get them out the other side," explained SIGCOMM paper author Kirill Levchenko, a UC San Diego post-doctoral researcher who recently earned his Ph.D. in computer science at UC San Diego. "A lightweight network monitoring approach such as ours allows you to pinpoint the source of the performance degradation and identify the problem routers."

Lossy Difference Aggregator

Simple counters and clever thinking are at the heart of the Lossy Difference Aggregator.

The classical way to measure latency is to track when a packet arrives at and leaves a router, take the difference of these times, and average over all packets that arrive over a fixed time period, such as one second. However, a typical router may process 50 million packets in a second, and keeping track of each packet's arrival and departure is a daunting piece of bookkeeping. It may seem that a simple approach is to sum all the arrival times in one counter, sum all the departure times in another counter, subtract the two counters and divide by number of packets. Unfortunately, this simple "aggregation" idea fails when a packet is lost within a router (which commonly happens). In that case, the lost packet arrival time is included but its departure time is not, throwing the whole estimate wildly out of whack.

Instead of summing the arrival and departure times of all packets traveling through a router, the computer scientists' system randomly splits incoming packets into groups and then adds up arrival and departure times of each of the groups separately. As long as the number of losses is smaller than the number of groups, at least one group will give a good estimate.

Subtracting these two sums (from the groups that have no loss) and dividing by the number of messages provides an estimate of the average delay with very little overhead-just a series of lightweight counters.

With this invention built into every router, a data center manager should be able to quickly pinpoint the offending router and interface that is adding extra microseconds of delay or losing even a few packets in a million, explained Levchenko.

"This is diagnostic tool, a potentially extremely important one. You don't want to just know that you have a network problem, you want to know which router and which application is causing the problem," said Snoeren.

The network manager can then upgrade the router or link, or reassign an offending application that is sending message bursts to another processing path.

By contrast, today's routers can be made to log messages; but looking through logs of millions of messages to pinpoint delay problems is like looking for a needle in a haystack.

"If implemented, this kind of approach should enable investment bankers to turn their attention to tuning their algorithmic trading programs to make more intelligent investments, instead of worrying about delays through obscure routers," said Varghese.

Media Contact: Daniel Kane, 858-534-3262 or dbkane@ucsd.edu