Computer scientists scale 'layer 2' data center networks to 100,000 ports and beyond

August 17 2009

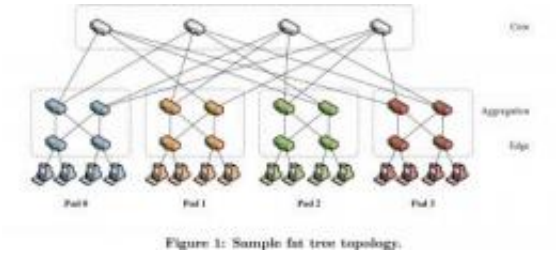

Figure 1: Sample fat tree topology.

A full prototype of PortLand, illustrated in Figure 1 from the paper, is currently running on a network in the Department of Computer Science and Engineering at UC San Diego's Jacobs School of Engineering. PortLand is a fault-tolerant, layer 2 data center network fabric capable of scaling to 100,000 nodes and beyond. PortLand is fully compatible with existing hardware and routing protocols and holds promise for supporting large-scale, data center networks by increasing inherent scalability, providing baseline support for virtual machines and migration, and dramatically reducing administrative overhead. Credit: UC San Diego Jacobs School of Engineering

University of California, San Diego computer scientists have created software that they hope will lead to data centers that logically function as single, plug-and-play networks that will scale to the massive scale of modern data center networks. The software system -- PortLand -- is a fault-tolerant, layer 2 data center network fabric capable of scaling to 100,000 nodes and beyond.

PortLand is fully compatible with existing hardware and routing

protocols and holds promise for supporting large-scale, data center networks by increasing inherent scalability, providing baseline support for virtual machines and migration, and dramatically reducing administrative overhead. Critically, it removes the reliance on a single spanning tree, natively leveraging multipath routing and improving fault tolerance. The computer scientists report this advance in data center networking on August 18, 2009 at SIGCOMM, the premier computer networking conference.

"With PortLand, we came up with a set of algorithms and protocols that combine the best of layer 2 and layer 3 network fabrics," said Amin Vahdat, the senior author on the SIGCOMM paper and a computer science professor at UC San Diego's Jacobs School of Engineering. "Today, the largest data centers contain over 100,000 servers. Ideally, we would like to have the flexibility to run any application on any server while minimizing the amount of required network configuration and state."

As mega data centers handle more and more of the world's computing and storage needs, data center networking is becoming increasingly important, the computer scientists say. Loading the front page of any active Facebook user, for example, typically involves over 1,000 servers in 300 milliseconds or less.

Looking for ways to improve data center networking, Vahdat and his team of graduate students from the Jacobs School of Engineering revisited the long-standing trade-offs between layer 2 or Ethernet networks—which route on MAC addresses—and layer 3 networks—which route on IP addresses.

Their result: PortLand, a system of algorithms and protocols that eliminates the scalability and routing-path limitations of existing layer 2 approaches and avoids the administrative and virtualization headaches

caused by implementing layer 3 networks in data center environments.

Today's data centers are often run on layer 3 networks, but this demands huge numbers of person-hours to set up and maintain the networks. Layer 3 networks also prohibit straightforward implementation of virtual machine migration— limiting flexibility and efforts to reduce energy and cost in the data center.

"Our goal is to allow data center operators to manage their network as a single fabric," said Vahdat, who directs the Center for Network Systems at UC San Diego. "We are working toward a network that administrators can think of as one massive 100,000-port switch seamlessly serving over one million virtual endpoints."

Location Discovery

One of PortLand's key innovations is its location discovery protocol, which opens up the possibility of a scalable layer 2 network. Switches automatically learn their location within the data center topology without any human intervention. These switches, then, assign "Pseudo MAC" (PMAC) addresses to each of the servers they connect to. These PMAC addresses—rather than MAC addresses—are used internally in the network for packet forwarding.

Server behavior remains the same in networks running PortLand. When a server wants to talk to a server on the other side of the data center, that first server still sends out an "ARP," which is a request for the MAC address of the computer with which it wants to communicate, based on its IP address.

But now, instead of broadcasting this request to the entire network, the switch that received the ARP talks to a directory service which returns a PMAC address, rather than the traditional MAC address.

"We have replaced broadcast with a server lookup. And we are forwarding based on PMAC addresses rather than MAC addresses. On the last hop, the egress hop, the switch rewrites the PMAC to be its actual MAC address," said Vahdat, the current Science Applications International Corporation (SAIC) Chair at the Jacobs School of Engineering. "We in effect transparently leverage the built-in hierarchy of data center networks."

When new machines are added, or when virtual machines are moved, new PMAC addresses are automatically generated.

"An important thing here is that all the switches are off the shelf—unmodified 'merchant silicon'," said Vahdat.

"I think PortLand is something that will be useful in the real world. The goal is to create a network fabric that allows you to buy any server or switch, plug it in and have it just work," said Radhika Niranjan Mysore, a UC San Diego computer science graduate student and the first author on the SIGCOMM paper. Mysore presented this work at SIGCOMM 2009 in Barcelona, Spain on August 18, 2009.

A full prototype of PortLand is currently running on a network in the Department of Computer Science and Engineering at UC San Diego's Jacobs School of Engineering.

"The students are getting good jobs and internships coming out of this project because they have data center networking skills. Companies are looking for this skill set," said Vahdat.

More information: "PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric," by Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat from the

Department of Computer Science and Engineering at the Jacobs School of Engineering at the University of California San Diego.
Download a copy of the paper at: cseweb.ucsd.edu/~vahdat/papers … rtland-sigcomm09.pdf

Source: University of California - San Diego (news : web)