UC San Diego Computer Scientists Propose New Data Center Architecture Based on Commodity Network Elements

Co-Author of Data Center Proposal Also Receives HP Labs Innovation Research Award

August 20, 2008

Tiffany Fox

Computer scientists at the UC San Diego's Jacobs School of Engineering have proposed a new way to build data centers that could save companies money and deliver more computing capability to end-users.

"Large companies are putting together server farms of tens of thousands of computers - even approaching 100-thousand, and the big challenge is to interconnect all these computers so that they can talk to each other as quickly as possible, without incurring significant costs," said Amin Vahdat, a professor of Computer Science and Engineering (CSE) in UC San Diego's Jacobs School of Engineering. "We are proposing a new topology for Ethernet data center connectivity."

The innovation is outlined in a paper, titled "A Scalable, Commodity Data Center Network Architecture," presented today by Vahdat at the annual meeting of SIGCOMM, the Special Interest Group on Data Communications. SIGCOMM is the premier academic conference for researchers in the fields of communications and computer networks, and the event runs through Friday in Seattle, Washington.

Vahdat, who also directs UCSD's Center for Networked Systems (CNS), co-authored the paper with two CSE graduate students, Mohammad Al-Fares and Alexander Loukissas.

It was also announced this week that Vahdat is one of only 41 researchers worldwide to be awarded a newly-created Hewlett-Packard Labs Innovation Research Award. The award will allow Vahdat and his team to develop further their proposed new networking architecture outlined in their SIGCOMM paper.

The researchers' work addresses problems inherent to current data center networks found in scientific computing, financial analysis, social networking, or any industry with large-scale computation or storage needs. Explained Vahdat: "Our work addresses the problem of data center network connectivity in a world where consolidation is increasingly taking place in data centers."

Typically, computers are connected by a network architecture that consists of a "tree" of routing and switching elements regulated by specialized equipment, with expensive, non-commodity switches at the top of the hierarchy. But even with the highest-end IP switches and routers, the networks can only support a small fraction of the combined bandwidth available to end hosts. This limits the overall cluster size, while still incurring considerable costs. Application design is further complicated by non-uniform bandwidth among data center nodes, which limits overall system performance.

The UC San Diego researchers' envision creating a data center that will have scalable interconnection bandwidth, making it possible for an arbitrary host in the data center to communicate with any other host in the network at the full bandwidth of its local network interface. Their approach requires no modifications to the end-host network interface, operating system or applications, and is fully backward compatible with Ethernet, IP

and TCP. Ideally, the data center would also use inexpensive, off-the-shelf Ethernet switches as the basis for large-scale data center networks, thereby replacing high-end switches in much the way that commodity personal computers have displaced supercomputers for high-end computing environments.

"The history of computing and technology has an innumerable set of examples where commodity parts take over the functionality of more specialized pieces of equipment," said Vahdat. "But that commoditization hasn't taken place on the communications side. So we do have these specialized components still living in the network infrastructure that incur significant costs and complexity."

"The other issue is that people are treating the data center as a mini-Internet," he continued. "That's fine, but they are then forced to use the same components they might use in the wide-area network with a bunch of concerns that don't come up in the data center environment: adversarial environments, attacks, parties that might not trust each other, etc. So you have a lot of functionalities in these specialized switching components that you really don't need in your data center, and you wind up paying for it in dollars and in complexity."

"Independent of whether our technique is successful or someone else's is, five years from now, switching and communication infrastructure in the datacenter will be based on these small commodity building blocks."

Furthermore, once 10GigE switches become the norm, Vahdat and his colleagues expect that their approach will be the only way to deliver full bandwidth for large clusters.

"One of the big benefits of our approach is that it scales trivially to 10 GigE at the edge, whereas competing techniques rely on aggregation to ever-faster links to achieve higher speeds," said Vahdat. "If you were to go to 10 GigE at the edge for servers today, you would need to aggregate to 40-GigE or 100-GigElinks moving up the tree topology. Unfortunately, the 40Gbps Ethernet standard isn't even out yet, and when it does become available, it will be very expensive. Because we're using identical network elements in our topology, today, 48-port 10-GigE Ethernet switches are becoming relatively inexpensive, so we would scale trivially to that environment as well."

As for the cost differential between his team's technique and those in current use, Vahdat says it's significant: "From a cost perspective, to build out a 25,000 node cluster today using current techniques with 100 percent bandwidth, just the switching equipment would cost somewhere in the order of $28 million, whereas with our technique using the identical network elements, would deliver the same performance but incur costs of maybe $4 million. That's a factor-of-seven difference."

Added Vahdat: "Going back to the question of 10GigE, there is no way you could interconnect 25,000 nodes using 10GB Ethernet today. There just isn't the equipment available for that. We could do so at some cost, even today, but it's not insurmountable. For somebody who has significant communication and computation needs and the appropriate budget, they could use our techniques to deliver 10GigE all the way out to the edge servers."

Media Contact: Tiffany Fox, 858-246-0353