Researchers develop cluster management tool that triples server efficiency

March 3 2014, by Tom Abate

Stanford engineers Christina Delimitrou and Christos Kozyrakis have created a software tool that can triple the efficiency of computer server clusters.

(Phys.org) —We hear a lot about the future of computing in the cloud, but not much about the efficiency of the data centers that make the cloud possible. In those facilities, clusters of server computers work together to host applications ranging from social networks to big data analytics.

Data centers cost millions of dollars to build and operate, and buying servers is the single largest expense the centers face. Yet at any given moment, most of the servers in a typical data center are only using 20 percent of their capacity.

Why? Because the workload can vary greatly, depending on factors such as time of day, the number of users logged in or sudden, unexpected demand. Having excess capacity is the usual way to deal with this peak-demand issue.

But as cloud computing grows, so will the cost of keeping such large cushions of capacity. That's why two Stanford engineers have created a cluster management tool that can triple server efficiency while delivering reliable service at all times, allowing data center operators to serve more customers for each dollar they invest.

Christos Kozyrakis, associate professor of electrical engineering and of computer science, and Christina Delimitrou, a doctoral student in electrical engineering, will explain their cluster management system, called Quasar, when scientists who design and run data centers meet for a conference in Salt Lake City, beginning March 1.

"This is a proof of concept for an approach that could change the way we manage server clusters," said Jason Mars, a computer science professor at the University of Michigan at Ann Arbor.

Kushagra Vaid, general manager for cloud server engineering at Microsoft Corp., said that the largest data center operators have devised ways to manage their operations but that a great many smaller organizations haven't.

"If you can double the amount of work you do with the same server footprint, it would give you the agility to grow your business fast," said

Vaid, who oversees a global operation with more than a million servers catering to more than a billion users.

How Quasar works takes some explaining, but one key ingredient is a sophisticated algorithm that is modeled on the way companies such as Netflix and Amazon recommend movies, books and other products to their customers.

How it works

To grasp what's new about Quasar, it's helpful to think about how data centers are managed today.

Data centers run applications such as search services and social media for consumers or data mining and large-scale data analysis for businesses. Each of these applications places different demands on the data center and requires different amounts of server capacity.

The cloud ecosystem includes software developers who run applications, and cluster management tools that decide how to apportion the workload and assign which applications to which servers. Before making such assignments, the cluster managers typically ask developers how much capacity these applications will require. Developers reserve server capacity much as you might reserve a table at a restaurant.

"Today data centers are managed by a reservation system," said Stanford's Kozyrakis. "Application developers estimate what resources they will need, and they reserve that server capacity."

It's easy to understand how a reservation system lends itself to excess idle capacity. Developers are likely to err on the side of caution. Because a typical data center runs many applications, the total of all those overestimates results in a lot of excess capacity.

Kozyrakis has been working with Delimitrou, a graduate student in his Stanford lab, to change this dynamic by moving away from the reservation system.

Instead of asking developers to estimate how much capacity they are likely to need, the Stanford system would start by asking what sort of performance their applications require. For instance, if an application involves queries from users, how quickly must the application respond and to how many users?

Under this approach the cluster manager would have to make sure there was enough server capacity in the data center to meet all these requirements.

"We want to switch from a reservation-based cluster management to a performance-based allocation of data center resources," Kozyrakis said.

Quasar is designed to help cluster managers meet these performance goals while also using data center resources more efficiently. To create this tool the Stanford team borrowed a concept from the Netflix movie recommendation system.

If you liked this application …

Before delving into the algorithms behind Quasar, understand that servers, like some people, can multitask. So the simplest way to increase server utilization would be to run several applications on the same server.

But multitasking doesn't always make sense. Take parenting, for instance. A mom or dad might be able to wash dishes, watch television and still spell a word to help a child with homework. But if the question involved algebra, it might be wise to dry your hands, turn off the TV and look at the problem.

The same is true for software applications and servers. Sometimes differing applications can coexist on the same server and still achieve their performance goals; other times they can't.

Quasar automatically decides what type of servers to use for each application and how to multitask servers without compromising any specific task.

"Quasar recommends the minimum number of servers for each application and which applications can run best together," said Delimitrou.

This isn't easy.

Data centers host thousands of applications on many different types of servers. How does Quasar match the right applications with the right server resources? By using a process known as collaborative filtering – the same technique that sites such as Netflix use to recommend shows that we might want to watch.

Applying this principle to data centers, the Quasar database knows how certain applications have performed on certain types of servers. Through collaborative filtering, Quasar uses this knowledge to decide, for example, how much server capacity to use to achieve a certain level of performance, and when it's OK to multitask servers and still expect good results.

Thomas Wenisch, a computer science professor at the University of Michigan, is intrigued by the Quasar paper, in which Kozyrakis and Delimitrou show how they achieved utilization rates as high as 70 percent in a 200-server test bed, compared with the current typical 20 percent, while still meeting strict performance goals for each application.

"Part of the reason the Quasar paper is so convincing is that they have so much supporting data," Wenisch said.

Next steps

Increasing data center efficiency will be essential for cloud computing to grow. These installations draw so much electricity that escalating demand threatens to overtax power plant output. So throwing more servers into the data center isn't the answer, even if money were no object.

But while they pursue higher efficiency from multitasking servers, data center operators must deliver consistent levels of service. They can't allow some customers to suffer because the servers are processing the wrong mix of tasks, a shortcoming known as "tail latency."

"The explosive growth of cloud computing is going to require more research like this," said Partha Ranganathan, a principal engineer at Google who is on the team that is designing next-generation systems and data centers. "Focusing on resource management to address the twin challenges of energy efficiency and tail latency can have significant upside."

Kozyrakis and Delimitrou are currently improving Quasar to scale to data centers with tens of thousands of servers and manage applications that span multiple data centers.

"No matter how well we manage resources in one data center, there will always be cases that exceed its capacity," Delimitrou said. "Offloading parts of work to other facilities in an efficient manner is key to achieving the flexibility that cloud computing promises."

Provided by Stanford University