Engineers develop techniques to improve efficiency of cloud computing infrastructure by as much as 20 percent

March 7 2013

(Phys.org) —Computer scientists at the University of California, San Diego, and Google have developed a novel approach that allows the massive infrastructure powering cloud computing to run more efficiently. The new approach can make these warehouse-scale computers run as much as 15 to 20 percent more efficiently. This novel model has already been applied at Google. Researchers presented their findings at the IEEE International Symposium on High Performance Computer Architecture conference Feb. 23 to 27 in China.

Computer scientists looked at a range of Google web services, including Gmail and search. They used a unique approach to develop their model. Their first step was to gather live data from Google's warehouse-scale computers as they were running in real time. Their second step was to conduct experiments with data in a controlled environment on an isolated server. The two-step approach was key, said Lingjia Tang and Jason Mars, faculty members in the Department of Computer Science and Engineering at the Jacobs School of Engineering at UC San Diego.

"These problems can seem easy to solve when looking at just one server," said Mars. "But solutions do not scale up when you're looking at hundreds of thousands of servers."

The work is one example of the research Mars and Tang are pursuing at the Clarity Lab at the Jacobs School, their newly formed research group.

Clarity is an acronym for Cross-Layer.

Architecture and Runtimes. Mars will be presenting some of their work April 18 on the UC San Diego campus at the Jacobs School of Engineering's Research Expo.

"If we can bridge the current gap between hardware designs and the software stack and access this huge potential, it could improve the efficiency of web service companies and significantly reduce the energy footprint of these massive-scale data centers," Tang said.

Finding the NUMA score

Researchers sampled 65 K of data every day over a three-month span on one of Google's clusters of servers, which was running Gmail. When they analyzed that data, they found that the application was running significantly better when it accessed data located nearby on the server, rather than in remote locations. But they also knew that the data they gathered was noisy because of other processes and applications running on the servers at the same time. They used statistical tools to cut through the noise. But more experiments were needed.

Next, computer scientists went on to test their findings on one isolated server, where they could control the conditions in which the applications were running. During those experiments, they found that data location was important, but that competition for shared resources within a server, especially caches, also played a role.

"Where your data is versus where your apps are matters a lot," Mars said. "But it's not the only factor."

Servers are equipped with multiple processors, which in turn can have multiple cores. Random-access memory is assigned to each processor,

allowing data to be accessed quickly regardless of where it is stored. However, if an application running on a certain core is trying to access data from another core, the application is going to run more slowly. And this is where the researchers' model comes in.

"It's an issue of distance between execution and data," Tang said.

Based on these results, computer scientists developed a novel metric, called the NUMA score, that can determine how well random-access memory is allocated in warehouse-scale computers. Optimizing the NUMA score can lead to 15 to 20 percent improvements in efficiency. Improvements in the use of shared resources could yield even bigger gains—a line of research Mars and Tang are pursuing in other work.

Provided by University of California - San Diego