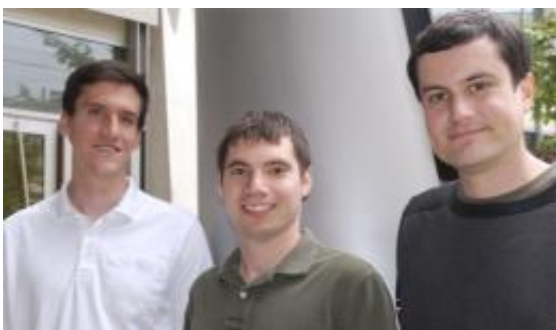


Computer scientists claim world data sorting record for second year

June 21 2011



The TritonSort-MR team included (l-r) Ph.D. students Alex Rasmussen and Michael Conley; CNS assistant research scientist George Porte.

(PhysOrg.com) -- Not content to rest on their laurels, a team of data center researchers from the Center for Networked Systems (CNS) at the University of California, San Diego recently broke two of their own world records. They also set world records in three other categories, including one for their TritonSort-MR system sorting a terabyte (one trillion bytes) of data in 106 seconds.

The competition that they entered, the Sort Benchmark, is the Formula One World Championship and Daytona 500 rolled into one for the world of large-scale data processing world. It attracts competitors from academic and industry labs all over the world, who vie to implement ever-faster data center designs.

“The competition provides excellent feedback on the team’s progress and gives them focus,” said CNS assistant research scientist George Porter. “The Sort Benchmark is like an annual reality check that gives us this objective standard by which we can validate how well we’re doing.” In addition to Dr. Porter, the CNS team included Center Director Amin Vahdat, and Ph.D. students Alex Rasmussen and Michael Conley from the Computer Science and Engineering department of the UCSD Jacobs School of Engineering.

The TritonSort-MR compute cluster is housed in the UCSD division of the California Institute for Telecommunications and Information Technology (Calit2), a close partner of CNS on the La Jolla campus.

Since 1994 the competition has spurred creativity in the realm of data sorting speed, and the number of applications demanding fast data sorting has increased exponentially – making the need for innovation more pressing each year. Massive data centers support processes like searching for tagged pictures of friends on Facebook, checking an order history with Amazon, or typing a term into a search engine. As data centers become faster in retrieving records, the more data-sorting applications can practicably be developed.

This expansion in the use and ubiquity of data centers has resulted in a concomitant explosion in capital expenditures for the enterprises that use them: data centers are expensive to equip, maintain, house, cool and power. Moreover, large-scale data processing tasks remain a significant bottleneck in the efficiency of data center activities. Rather than wait for hardware designers to come up with new equipment, data center architects are looking for better ways to use the equipment that currently exists on the market to achieve new goals in speed and efficiency.

In 2010 the CNS group won in the “Indy” category for the “Gray” and “Minutesort” categories, racing to sort 1 TB of data as quickly as

possible, and as much data as possible in a single minute, respectively. The “Indy” category exists only for this competition, so designing a system to compete here is comparable to constructing a racing vehicle that can only be driven on a track. But building on their successful foray in 2010, the team decided to take their game to a new level by adjusting their system to compete in the “Daytona,” or general purpose, category as well.

Rasmussen says the team had unfinished business from the previous competition. “When we set the record the first time, we had only just gotten TritonSort to go as fast as we thought it could go,” noted the Ph.D. student. “But there were a lot of questions about the system’s performance that we just didn’t have answers to.”

“The key to the TritonSort-MR design is seeking an efficient use of resources, and to build balanced systems,” added Porter. “We made some improvements on the data structures and algorithms, basically to make it a lot more efficient in terms of sending records across the network.”

With the modifications, “Daytona” was successful, and the modifications also allowed the team to upgrade the original specialized system built to compete in the “Indy” category. Showing impressive improvements in performance, the team submitted for and won both categories in the “Gray” and “Minutesort” competitions.

Beyond the achievement of speed, TritonSort-MR also proved remarkable for its efficiency: while the second-place team used 3,500 nodes to achieve their result, the TritonSort-MR team used only 52. If implemented in a real-world data center, TritonSort-MR would therefore allow a company to sort data more quickly, while only making one-seventh of the investment in equipment, space, and energy costs for cooling and operation.

While winning in these four categories exceeded the team's original goals from 2010, they found themselves intrigued by a new category on offer in 2011. The "100 Terabyte Joulesort" competition challenges teams to build systems that can sort the greatest number of data records, while consuming no more than one joule of energy. (By way of illustration, it takes roughly one million joules to watch TV for an hour.) The introduction of this new category reflects the recognition of an increasingly dire challenge facing industry in trying to solve data-intensive computing problems: energy usage. A primary reason why data centers are expensive to operate is the staggering scale of their energy consumption. Any design that increases energy efficiency would have a positive and much needed impact on both the environment and on a company's bottom line.

Though intrigued by this new opportunity, the team was skeptical at first that they could compete in the Joulesort arena. "Typically when you look at systems that set records like this, they're all built out of these incredibly energy efficient pieces," said Alex Rasmussen. "But you'd never see this equipment deployed in an actual data center setting [because of its high cost]."

The TritonSort-MR team, on the other hand, was focused on making a system of direct applicability to enterprises with real-world needs and resources, rather than breaking a record for its own sake. This is reflected, said Rasmussen, in the type of equipment the TritonSort-MR team employs for its system. "The stuff that we're using is kind of commodity server stuff," he said. "We've got machines from HP that are a year and a half old," with multi-core processors and a Cisco Nexus 5596 switch. As an additional challenge to the efficiency of the design, the team elected not to customize their system for energy optimization. Despite placing these limitations upon themselves, the TritonSort-MR group won the Joulesort category handily – proving that the CNS solution was both fast and remarkably energy efficient.

Provided by University of California - San Diego

Citation: Computer scientists claim world data sorting record for second year (2011, June 21)
retrieved 26 April 2024 from <https://phys.org/news/2011-06-scientists-world-year.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.