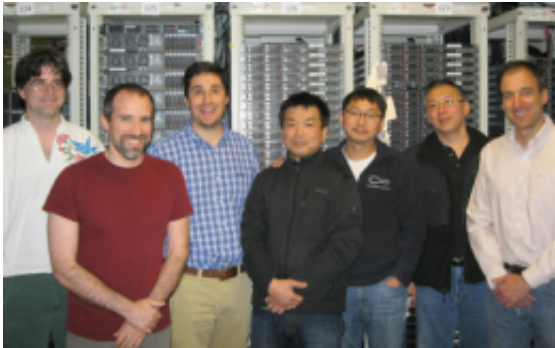# Data in the fast lane

May 22 2012, By Douglas Gantenbein

The record-setting MinuteSort team: (from left) Jon Howell, Jeremy Elson, Ed Nightingale, Yutaka Suzue, Jinliang Fan, Johnson Apacible, and Rich Draves.

A new approach to managing data over a network has enabled a Microsoft Research team to set a speed record for sifting through, or "sorting," a huge amount of data in one minute.

The team conquered what is known as the MinuteSort benchmark—a measure of data-crunching speed devised by the late Jim Gray, a renowned Microsoft Research scientist, and deemed the "World Cup" of data sorting. The MinuteSort benchmark measures how quickly data can be sorted starting and ending on disks. Sorting is a basic function in computing, demonstrating the ability of a network to move and organize data so it can be analyzed and used.

The team, led by Jeremy Elson in the Distributed Systems group at Microsoft Research Redmond, set the new sort benchmark by using a

radically different approach to sorting called Flat Datacenter Storage (FDS). The team's system sorted almost three times the amount of data (1,401 gigabytes vs. 500 gigabytes) with about one-sixth the hardware resources (1,033 disks across 250 machines vs. 5,624 disks across 1,406 machines) used by the previous record holder, a team from Yahoo! that set the mark in 2009.

Two Hundred Bytes for Everybody

To put things in perspective, in one minute, the Microsoft Research team sorted the equivalent of two 100-byte data records for every human being on the planet.

The record is significant because it points toward a new method for crunching huge amounts of data using inexpensive servers. In an age when information is increasing in enormous quantities, the ability to move and deploy it is important for everything from web searches to business analytics to understanding climate change.

In practice, heavy-duty sorting can be used by enterprises looking through huge data sets for a competitive advantage. The Internet also has made data sorting critical. Advertisements on Facebook pages, custom recommendations on Amazon, and up-to-the-second search results on Bing all result from sorting.

The award for the team's achievement will be presented during the 2012 SIGMOD/PODS Conference, an international forum for database researchers, practitioners, developers, and users to explore cutting-edge ideas and results. This year's conference occurs in Scottsdale, Ariz., from May 20 to 24.

The team, formed and led by Elson, included Johnson Apacible, Rich Draves, Jinliang Fan, Owen Hofmann, Jon Howell, Ed Nightingale,

Reuben Olinsky, and Yutaka Suzue.

Their approach was to take a fresh look at a relatively old model for sorting data. More than a decade ago, a network of computers would access data on a single file server, and each computer saw all of the data.

But that model didn't scale up as data centers became larger. Researchers at Google tackled that problem in 2004, creating a data-management scheme called MapReduce. It worked by essentially sending computation to the data, rather than dragging data to a computer. It made possible computation across huge data sets using large numbers of cheap computers. In recent years, the Apache Software Foundation developed an open-source version of MapReduce dubbed Hadoop.

MapReduce and Hadoop greatly advanced the state of data sorting. But, Elson says, they still weren't perfect.

"Some kinds of computations just can't be expressed that way," he says of the drag-computation-to-the-data model. "If you have two big data sets and you want to join them, you have to move the data somehow."

Three years ago, Elson, Nightingale, and Howell had an insight into how new advances in network bandwidth could lead to a simpler model of data sorting—one in which every computer saw all of the data—while also scaling to handle massive data sets.

The solution was dubbed Flat Datacenter Storage. Elson compares FDS to an organizational chart. In a hierarchical company, employees report to a superior, then to another superior, and so on. In a "flat" organization, they basically report to everyone, and vice versa.

FDS takes advantage of another technology Microsoft Research helped develop, called full bisection bandwidth networks. If you were to draw

an imaginary line through a collection of computers connected by a full bisection bandwidth network, every computer on one side of the line could send data at full speed to every computer on the other side of the line, and vice versa, no matter where the line is drawn.

Using full bisection networks, the FDS team built a system that could transfer data at two gigabytes per second on each computer for input, with another two gigabytes for output.

New Techniques Needed

"That's 20 times as much bandwidth as most computers in data centers have today," Elson says, "and harnessing it required novel techniques."

With that, the team was ready to take on the MinuteSort challenge. The contest actually has two parts: an "Indy" category, in which systems can be customized for the task of sorting, and a "Daytona" category, in which systems must meet requirements for general-purpose computing—think super-sleek, open-wheel Indianapolis 500 cars versus Daytona 500 stock cars that look a little like what you see on the street.

In 2011, a team from the University of California, San Diego set a record in the Indy category, sorting 1,353 gigabytes of data in a minute. In the Daytona category, the record had been held by a team from Yahoo!, which sorted 500 gigabytes of data in a minute.

The Microsoft Research team blew past both marks. Moreover, the team beat the standing Indy-sort record using a Daytona-class system. This isn't the first time that has happened, Elson says, but it is rare.

The record represents a total efficiency improvement of almost 16 times. Interestingly, Microsoft Research set the record using a remote file system, which is an unusual choice of architecture for sorting

because it commonly is perceived to be slow. Whereas most sorting systems read data locally from disk, exchange data once over the network, and write data locally to disk, in a remote file system, data is read, exchanged, and written over the network, so each data record crosses the network three times. The team deliberately handicapped the system to demonstrate the phenomenal performance of the new FDS file-system architecture.

Thus far, the Microsoft Research team has worked with the Bing team to help Bing accelerate its search results. The Microsoft Research engineering team is partially funded by Bing and has been actively supported by Harry Shum, the Microsoft corporate vice president who leads Core Search Development.

Exciting Breakthrough

"We are very excited about the MinuteSort breakthroughs made by our Microsoft Research colleagues," Shum says. "I look forward to taking advantage of the FDS technology to further online infrastructure for Bing and for Microsoft—and delivering even faster results to our users."

Nightingale, co-leader of the FDS project with Elson, is working with Bing to integrate FDS to improve Bing's efficiency and speed.

Given the ubiquity of interest in managing "big data," the Microsoft Research work is apt to find a home in several computing fields. It could be used in the biological sciences, managing gene sequencing or helping to create new classes of drugs, or it might help in stitching together aerial photographs to give people better imagery of the planet.

The ability to sort data rapidly also will aid machine learning—the design and development of algorithms that enable computers to create predictions based on data, such as sensor data or information from

databases. [Microsoft](#) Research has a big stake in machine learning, in work ranging from language processing to security applications.

"Improving big-data performance has a wide range of implications across a huge number of businesses," Elson says. "Almost any big-data problem now becomes more efficient, which, in many cases, will be the difference between the work being economically feasible or not."

For now, there's also a lot of celebrating going on.

"Our hands," Howell laughs, "are bruised from high-fiving."

Provided by Microsoft Corporation