Processing Parallel Insights

KENNETH ROSS

Professor of Computer Science

The volume of data we want to analyze is growing even faster than computing power. Kenneth Ross is looking for ways to close the gap. "People are coming up with ever-more challenging database projects, like analyzing the differences in genomes, which have billions of base pairs, among thousands of patients," Ross said.

Until now, computer scientists have relied on raw increases in computer power to crunch more data. Today, those advances have been harder to achieve. To keep moving forward, engineers reinvented the microprocessor, dividing it into two or more smaller processors, or cores.

Dividing tasks among cores works best when the answers do not depend on the previous step. Databases are like that. "The work you do on one record is pretty much what you do on another, you can process them in parallel," Ross said.

Yet parallelism comes with its own set of problems, such as cache misses and contention.

Cache misses occur because computer processors have fast and slow memory. They waste hundreds of processing cycles retrieving data from slower memory. Those lost cycles—cache misses—waste half the time needed to perform some tasks.

Ross wants to reorganize data to take up less space in memory. The hard part, he said, is doing this without spending too much time or resources.

"I'm trying to take advantage of relatively recent changes in computer architecture to make database software more efficient," said Ross. "Computer processors are now made up of four to eight smaller processors, or cores. We have to take advantage of those cores by developing code that runs in parallel."

Contention occurs when several parallel jobs all need to update a single item. "Each of those jobs needs exclusive access to the item for a short time to keep them from interfering with one another. If the item is sufficiently popular, those jobs get stuck in line waiting for their turn to access the data rather than working in parallel," Ross explained.

Ross' recent research seeks to automatically detect contention and then create several clones of the busiest data items. "We want to distribute processes among the clones and then combine results. Again, the key is to do this without using more computer resources than we are saving by eliminating contention," he said.

From genomics to climate, the sciences are accumulating data at a faster rate than ever before. Ross' work will help make it possible to analyze that data and see what they really mean.

B.Sc., University of Melbourne, 1986; Ph.D., Stanford, 1991