Big Data Analytics 2013

Invited talk

Big Data on Small Machines

Guy Blelloch, *Carnegie Mellon University*

Although the largest data sets can only fit on large clusters of thousands of machines, there are many applications of "large data" that with appropriate techniques can fit on much more modest machines.

Graphs with more than 100 Billion edges (significantly larger than the twitter graph), for example, can fit in the primary memory on a modest rack mounted server and on secondary memory of a modest desktop machine.

Furthermore multi or many-core shared memory machines can be much more efficient way of taking advantage of parallelism in terms of throughput per core or per watt than large distributed memory machines.

I'll describe our experience mapping relatively large graph problems onto modest machines, including approaches for in-memory compression, for processing from secondary memory, and for using multicore parallelism within a single system.