Enabling Knowledge Discovery in a Virtual Universe

Jeffrey Gardner
(U. of Washington)

Abstract

Virtual observatories will give astronomers easy access to an unprecedented amount of data. Extracting scientific knowledge from these data will increasingly demand both efficient algorithms as well as the power of parallel computers. Nearly all efficient analyses of large astronomical datasets use trees as their fundamental data structure. Writing efficient tree-based techniques, a task that is time-consuming even on single-processor computers, is exceedingly cumbersome on parallel or grid-distributed resources. We have developed a library, Ntropy, that provides a straightforward way of developing tree-based data analysis algorithms for parallel platforms. It achieves its high performance and scales to thousands of distributed memory nodes. It does this by providing a combination of two parallelization strategies, remote method invocation (RMI) and distributed shared memory (DSM) that are tailored to the problem domain. This demonstrates the potential value of lightweight domain-specific libraries in data-intensive scientific computing.