Indexing the Web - A Challenge for Supercomputers

Monika Henzinger
Google Inc.

Abstract

Since January 2002, the Google search engine has been powering an average of 150 million web searches a day, with a peark of over 2000 searches per second. These searches are performed over an index of over 2 billion documents, over 300 million images, and over 700 million Usenet messages.

To guarantee fast user response time, Google performs these searches on a cluster of over 10,000 PCs. The main challenages with this architecture are fault-tolerance and the quality of search results. Replication solves the former and the PageRank score is used to advance the latter. The PageRank score is based on an eigenvalue computation of a large matrix that is derived from the web graph and is one of the main contributor to very high quality search results.

As Internet use continues to grow, so does the use of the Google search engine. The Google architecture is designed to scale to accommodate the growth in useage as well as the growth of the web.

Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER'02)
0-7695-1745-5/02 $17.00 © 2002 **IEEE**

IEEE COMPUTER SOCIETY