



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

FY06 LDRD Final Report Data Intensive Computing

Ghaleb M. Abdulla

February 15, 2007

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

FY06 LDRD Final Report
Data Intensive Computing
Tracking code: 06-ERD-058
Ghaleb M. Abdulla

Abstract

The goal of the data intensive LDRD was to investigate the fundamental research issues underlying the application of High Performance Computing (HPC) resources to the challenges of data intensive computing. We explored these issues through four targeted case studies derived from growing LLNL programs: high speed text processing, massive semantic graph analysis, streaming image feature extraction, and processing of streaming sensor data. The ultimate goal of this analysis was to provide scalable data management algorithms to support the development of a predictive knowledge capability consistent with the direction of Aurora.

Introduction

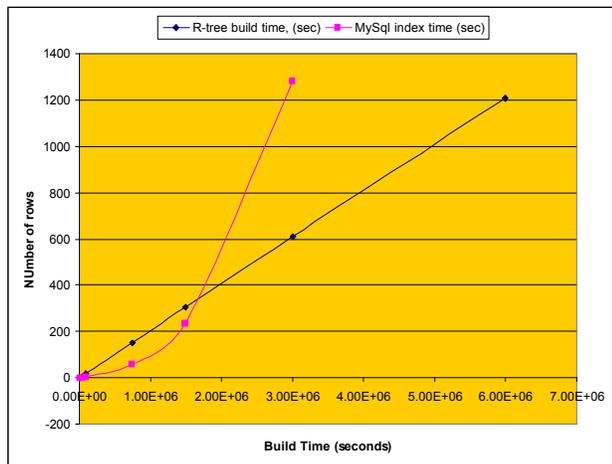
We focused on exploring the underlying computer science approaches, models, and technologies that enable high-performance systems to be applied to data intensive computing. We identified a use case that represents an example of a data intensive computing application where large amounts of streaming images need to be analyzed and transient objects identified in real time. We characterized the requirements, designed an efficient algorithm for indexing objects, implemented it, and compared it to traditional database approaches. Our results were encouraging.

Research activities

We developed a spatial data indexing algorithm based on the SaIL's library. Our implementation supports point indexing and distributed indexes.

We performed a set of experiments to compare the performance of our approach to using a relational database system. Our implementation scales linearly with the number of data points (see figure); on the other hand it grows exponentially using a regular DBMS. We implemented an incremental clustering algorithm for spatial objects and we started characterizing the query performance using our distributed indexing algorithm.

We implemented an efficient text indexing algorithm that supports efficient disk access while compressing the indexed data. Using our implementation we were able to fully construct an index for a 4GB collection of 750,000 Usenet articles on a Linux workstation in 50 minutes. We compared our results to the official



TREK conference results and we are comparable to the best published algorithm. We developed a use case for transient image detection for the large scale synoptic telescope (LSST) and we documented it. We published the results of evaluating our approach to help solve the LSST use case to using traditional database approach.

Exit Plan

This project was funded for 5 months only, however, the research ideas and results were useful to other projects and PKS. The main developer joined Proteus and he is helping the project solve related problems. The PI has moved to work on PKS and he is utilizing the spatial and temporal indexing experience to help PKS.

References

Evaluation of Potential LSST Spatial Indexing Strategies, Sergei Nikolaev, Ghaleb Abdulla, Robb Matzke, 2006, UCRL-TR-225827