ELSEVIER
Procedia Computer Science 7 (2011) 136–138
Procedia Computer Science

The European Future Technologies Conference and Exhibition 2011

Affordable Supercomputing for Data Mining Applications☆

András A. Benczúr

Informatics Laboratory, Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)

Abstract

Business intelligence, e-science and Web mining are rapidly growing sources of extreme large scale problems. We demonstrate that emerging means of supercomputing-for-the-masses, distributed or many-core architectures may provide feasible and affordable solution to many applications. We give a brief overview in four areas:

- Web data processing to federate Future Internet Research (FIRE) facilities in a recently started project;
- Image retrieval over tasks of the PASCAL and PROMISE networks of excellence;
- Efficiency of distributed data warehouses for log processing and entity resolution;
- Network analysis, visualization and navigation, in the VAST Challenges and European security projects.

We pass beyond existing technologies in implementing new frameworks, combining distributed and multicore technologies, and seeking new, breakthrough applications in an interdisciplinary research with application partners.

Selection and peer review under responsibility of FET11

Extreme large computational problems arise in social network mining, customer relationship management, personalized and similarity search, recommendation, spam filtering, e-science as well as security problems. Flagship consumers are the so-called Web 3.0 applications envisioned for the next 10-year period of Web development, fast and personalized applications accessible from any device, aided by *data mining* solutions handling *distributed data* by *cloud computing*.

Scalability issues are central in the applicability of data mining and arise from two main sources. Certain problems such as Web or transactional log processing are *data intensive* where reading vast amounts of data itself forms a bottleneck. Others such as machine learning or gene sequencing are *computational intensive* as they require complex algorithms run on large data that may not fit into the internal memory of a single machine.

Until recently, scaling could rely on the exponential growth of hardware capabilities. Moore's law, after an astonishing 40 years of predicting the doubling of capacities over the same circuit size, has recently invalidated by reaching the physical limits to increase the clock speed for single-core processing. The recent answer to the breach in processor capacity increase is to pursuit technologies that provide *affordable supercomputing*, including multicore processors and parallel programming environments. A cluster of commodity machines is ideal for data intensive problems. For computational intensive problems, a graphical processor (GPU), although limited in built-in memory, may even be less expensive than traditional processors. Programmers are however not yet capable of utilizing the full potential of these architectures; in addition, both data and computational intnsive tasks may require yet unknown new architectures.

☆ Support from OTKA NK72845, NKFP-07-A2 *TEXTREND* and EU FP7 projects *LAWA* and *SCIIMS*. In collaboration with researchers of the Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI).

Selection and peer review under responsibility of FET11 doi:10.1016/j.procs.2011.09.009

We illustrate our solution for very large scale computational problems by our four applications below. Although specific technologies marked by the names of Cray or Thinking Machines date back to the early 70's, we had to wait until recently for their widespread use over new, affordable hardware-software architectures.

Web processing. Nowadays data in general and Web data in particular is becoming increasingly local and thus distributed in nature. It has become necessary to move more analysis to the data, not the reverse. In our work [1] we processed the ClueWeb09 corpus of half a billion of English language Web pages and, as the main organizers, prepared a 200 million page data set for the ECML/PKDD Discovery Challenge[1] 2010 over a Hadoop cluster.

As another initiative, in the LAWA project we will build an Internet-based experimental testbed for large-scale data analytics. Its emphasis is on developing a sustainable infrastructure, scalable methods, and easily usable software tools for aggregating, querying, and analyzing heterogeneous data at Internet scale.

Image processing. Image processing forms the natural area of use for graphics processors. While most image processing steps are already implemented on the GPU, global analysis of an image corpus does not translate to a graphics processor in the same smooth manner. Image classification consists of assigning one or multiple labels to an image based on its semantic content. The task remains challenging despite of much progress, in particular in the context of the EU FP7 Networks of Excellence PASCAL (Pattern Analysis, Statistical Modeling and Computational Learning) and PROMISE, which aims at advancing the experimental evaluation of complex multimedia (ImageCLEF) and multilingual information systems. We give an accurate yet very fast GPU implementation[2] of a visual word generation method, a key tool in image classification [2].

Efficiency in large databases. In the last decade, data storage and processing technologies that extend or substitute traditional relational databases gain more and more ground. Certain products including Hadoop are open source, others are proprietary (Oracle, IBM, Greenplum, etc.). Most provide distributed and cost-effective scalability, but only for a limited class of problems. We developed successful applications and tools for reporting and data mining, with effective and scalable compression, nearline storage, data mining and data stream processing methods [3].

A core information integration and fusion task is Entity Resolution, the identification and merging of distinct entity representations from heterogeneous source databases. In an ongoing project we extend and combine known techniques and our results [4] via new classification methods including hybridization.

Very large graph processing. The World Wide Web, the largest network ever processed, relies on graph algorithms including Google's PageRank. Telephone call graphs form another class of large and practically important networks with typical problems of visualization, classification [5] and fraud detection.

Several graph algorithms can be parallelized by splitting the graph into pieces for many servers where they fit into internal memory. Such an architecture can serve shortest path, "give me the nearest community" queries and similarity search. Path fingerprinting [6] is a flexible tool that fit well to parallel architectures. In these applications, partitioning huge distributed graphs raises an additional challenge. We build on our spectral partitioning results, a method that on real networks is thought to perform bad [7]. Hadoop, an open source implementation of the Map-Reduce framework, is considered SoA in distributed Web processing. Unfortunately, large graph and matrix processing exceed Hadoop and Map-Reduce capabilities. For matrix processing may rely on Pegasuscollection of generalizations and optimizations on Hadoop; Direct node messaging such as OpenMPI, or incubatory implementations of graph processing frameworks in the Bulk Synchronous Parallel model similar to Google's Pregel also exist.

Conclusions. Affordable supercomputing can play a key role in scaling data and computational intensive problems in Database Technologies, Applied Statistics, Theory of Algorithms, Information Retrieval and Machine Learning. In the future we seek new applications such as Web 3.0, and consider the combination of distributed and many-core computing for problems that are both data and computational intensive.

References

[1] A. Garzó, D. Nemeskey, R. Pethes, D. Siklósi, A. Benczúr, SZTAKI @ TREC 2010, TREC working notes.
[2] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: in: IEEE CVPR, 2007, pp. 1–8.
[3] B. Rácz, C.I. Sidló, A. Lukács, A.A. Benczúr, Two-phase data warehouse optimized for data mining, in: in: BIRTE in conj. VLDB, 2006.
[4] C.I. Sidló, Generic entity resolution in relational databases, in: in: ADBIS, 2009.

[1] http://www.ecmlpkdd2010.org/
[2] http://datamining.sztaki.hu?=en/GPU-GMM

[5] M. Kurucz, D. Siklósi et al., KDD Cup 2009 @ Budapest: Feature partitioning and boosting, in: KDD Cup and Workshop, 2009.
[6] D. Fogaras, B. Rácz, Scaling link-based similarity search, in: Proc 14th WWW, 2005, pp. 641–650.
[7] M. Kurucz, et al., Large-Scale Principal Component Analysis on LiveJournal Friends Network, in: in: SNAKDD in conj. KDD, 2008.