Presented at Workshop on Distributed Data & Structures - WDAS-2002, University Paris 9 Dauphine, March 20-23, 2002

High-performance GRID Database Manager for Scientific Data

Tore Risch, Milena Koparanova, and Bo Thide

Tore.Risch@dis.uu.se, Milena.Koparanova@dis.uu.se, bt@irfu.se

Uppsala University
Sweden

Jan 14, 2002

Abstract

The GRID initiative provides an infrastructure for distributed computations among widely distributed high-performance computers. This will allow for exchanging and processing very large amounts of data. The LOFAR project (www.nfra.nl/lofar) is an international initiative to build a versatile, geographically distributed, multi-point radio facility for astrophysics, space physics, atmospheric physics, and radio research, utilizing very high performance GRID computing. LOIS is a proposed Swedish outrigger to LOFAR providing a software radar. As the volume of processed data by LOFAR/LOIS is very large and dynamic there will be need for very high performing data management systems. For this a high-performance stream-oriented distributed data manager and query processor is being developed that allows very efficient execution of database queries to streamed data involving numerical and other data. Very high performance is attained by utilizing many object-relational main-memory database engines running on PCs and connected through the GRID. The project leverages upon a high-performance, extensible, and object-oriented database engine, the Amos II kernel, developed in the Uppsala Database Laboratory. A very high performing stream-oriented DBMS is being developed for representing and querying non-relational data representations extracted from the data flows used in space and environmental physics applications. Of particular interest is the development of new distributed data population and query processing techniques for this kind of applications and thereby utilizing distributed and scalable data structures for high-performance stream data processing.

1 Introduction

We are developing a new kind of database manager utilizing the evolving GRID infrastructure for distributed computations [FK1999], called the *GRID Data Manager (GDM).* GDM will have very high performance and support for customizable representation of streamed data in distributed data and computation servers. The target application area is space physics, in particular the LOFAR/LOIS project described below, whose purpose is to develop a distributed software space telescope and radar utilizing the GRID. LOFAR/LOIS will produce extremely large amounts of distributed data streams by sensors receiving signals from space. One these streams various numerical selection and transformation algorithms are applied before the data is delivered to client workstations for visualizations and other processing. This is a new environment very different from conventional server-oriented database environments. A research challenge is to make GDM able to handle very large amounts of dynamically produced distributed data. The application requires orders of magnitude better data processing performance than conventional DBMSs and support for queries involving customized computations. We will achieve this by utilizing cheap and large main-memories (on PCs) connected through the GRID to form clusters of main-memory based object-relational databases. Each node in such a system is a main-memory database managed by a fast main-memory object-relational DBMS engine. Sections of the streams are materialized in the node databases and numerical algorithms are run on the nodes to select, combine, and transform the data. GRID-based clusters of main memory database nodes combined with extensibility through user-defined data representations provide very high performance. The databases are scalable by dynamically incorporating new GRID nodes as the database grows.

The environment will require the integration of both data and computations. Research is needed on how to effectively combine high-performance distributed stream databases with distributed computation techniques. A framework will be developed for general extensibility of the parallel GDM through user-defined data representations and query optimization algorithms distributed over many nodes. Efficient customized data representations and query optimization algorithms can thus be distributed to each node in the cluster.

It will be impossible to store all data in the distributed nodes, but rather only moving windows of the streams are stored in the nodes and the queries are executed over these windows. Managing the very high data flow though the nodes requires support for very high insert and delete frequencies.

It will be virtually impossible to manually configure subtasks in this dynamic environment. New tools and algorithms will be needed for automatically configure and adapt data and computations servers based on the resources available.

We are developing a prototype illustrating the new database architecture and how to utilize such an architecture for the very demanding LOFAR/LOIS application. Through it we will evaluate the benefits from using GDM for this kind of demanding applications.

The prototype will combine high performance through main-memory data representations, extensibility through object-orientation and plug-in of external code, and scalability through the use of scalable and distributed data structures [LNS1996] for storing distributed node data.

2 LOFAR/LOIS

Public concern about and natural and anthropogenic effects on Earth and its space habitat, Man's existential questions about the origin and fate of Universe, and the quest for better and more resource conserving information processing, are challenges that society put on the Earth and Space Science communities. These communities accept these challenges and also formulate their own.

In response to these challenges, novel scientific methodologies and innovative research facilities are continually being developed. Generally speaking, the research facilities are of two kinds: space-borne and ground-based. While instruments on-board satellites can detect and monitor processes and phenomena which cannot be observed from Earth, the rapid motion of the spacecraft---of the order of 5-10 km/s---precludes long-term observations of given, fixed regions in Earthspace. For such observations, of utmost importance for the discovery of long-term trends, observatories on the ground are ideally suited. Therefore, instruments in space-borne and ground-based instruments complement each other in an almost ideal manner.

An example of the former is the upcoming ENVISAT (envisat.esa.int) advanced polar-orbiting Earth observation satellite which will provide measurements of the atmosphere, ocean, land, and ice over a five year period, and the ongoing atmospheric physics and astrophysics mission Odin (www.snsb.se/Odin/Odin.html) and space physics mission Cluster (sci.esa.int/cluster), in which Swedish science and technology play prominent roles.

An example of the latter is the radio observatory LOFAR (www.nfra.nl/lofar), an international initiative to build, during the 2002-2006 time-frame, in northern Europe and Scandinavia, a versatile, geographically distributed, multi-point radio facility for astrophysics, space physics, atmospheric physics, and radio research which in a unique way addresses the above challenges and more.

For a giant radio and radar research project such as LOFAR, Swedish physics, technology and industry has a very interesting profile. Therefore, LOFAR proper, primarily designed for radio astronomy, will be augmented by a Swedish ``outrigger" called LOIS (www.wavegroup.irfu.se/LOIS) which will add atmospheric, ionospheric, solar and planetary space physics capabilities by providing a radar sub-facility, including an absolute top-notch infrastructure, in the region surrounding Växjö in Southern Sweden.

LOFAR/LOIS will be at least a hundred times more sensitive than any comparable facility anywhere and is best viewed as a huge, geographically distributed network of electromagnetic sensors and emitters, fully digitized and operated entirely in software. The data from Earth's space environment and Universe will be broadcast in real time over the World Wide Grid. This will be the first facility of this kind where public outreach issues, including immediate, direct access to scientific data as they are being produced, will be considered primary design criteria.

At full operation in the 2006 time-frame and beyond, LOFAR/LOIS will consist of 31,000 antennas with sensors and emitters, organized in hundreds of clusters distributed within a circular geographical region of about 350 km diameter. The total data rate will be circa 25 Tbits/s. This means that LOFAR/LOIS will utilize technology at or even beyond the current leading edge and, hence, will advance antenna, radio, detector, and data handling technologies far beyond their current limits and therefore be an ideal Swedish DataGRID test-bed.

3 Related work

The area of stream data management has gained increased attraction recently [BGS2001] [GKS2001] [GMMO2000] [LPT1999] [Sul1996] [TGNO92] and results from that area is expected to be very important for this project. A good recent overview of the area can be found in [BW2001].

Another modern development in the database area is to use large main memories to represent the database [GS1992]. The main memory of modern computers can be cheap and large enough to entirely store many databases. Furthermore, many applications can be simplified by embedding a lightweight DBMS that manages its data. DBMS vendors are therefore developing lightweight main memory relational databases [KLLP2001] [Syb2001] [Ora2001] that often interoperate with their regular relational server DBMSs. The Amos II DBMS engine [RJ2001] is also such an embeddable main-memory DBMS. However, different from commercial embedded databases, it has a complete and extensible query optimizer for object-oriented queries, i.e. it is an object-relational main memory DBMS. Amos II supports extensibility of both the storage manager and the query processor, but the engine itself is limited to a single main-memory. It has been used for integrating high performance query capabilities in engineering analysis systems [Ors1996] [OR1996][FOR1998].

The performance and scalability of main memory databases can be substantially further enhanced by relying not only on one main memory in one computer but also on clusters of main-memory databases. Examples of relational DBMSs using such architectures are ClustRa [HTBH1995] and NDB [Ron1997][Ron1999]. Those systems are very high performing systems, e.g., for telecom applications. They also have the additional property of very high reliability by automatic data replication, which is required for their applications.

The development of very high performance and scalable distributed data structures, SDDSs [LNS1996] provides distributed dynamic data structures that scale well by utilizing large main memories on distributed computers connected through fast communication networks. Some important properties of SDDSs are that they do not rely on a central server node and gracefully scale to use more nodes as the database grows. There are many variants of SDDSs depending on their indexing capabilities, high availability, etc. [KLR1996] [LNS1994] [LS2000] [LR2001]. The Amos II kernel has successfully been coupled to a manager of SDDSs developed at Dauphine University (Paris) [NDLR2001] indicating very good performance and scalability. In the project we will investigate how SDDSs can be utilized for high-performance query processing of stream data.

Object-Relational DBMSs [SB1999] allow the definition of abstract data types whose instances are stored in relational tables and user defined functions that are executed in the database server. Cost hints can be defined to guide the query optimizer how to place calls to the UDFs in the optimized query execution plans. Modern relational DBMSs have object-relational extensions and extensibility is now part of the SQL standard through the SQL-MED standard [M+2001]. However, commercial object-relational databases are still disk based, and computationally intensive applications still require customized main-memory based data representations.

The difference between GDM and previous approaches is the development of a DBMS that supports advanced queries over both regular data and data streams and is extensible while at the same time having very high performance. This is achieved by developing a distributed, main-memory, stream-oriented, object-relational DBMS. We are extending an existing main-memory object-relational DBMS engine with capabilities to utilize large distributed main memories. Main-memory, distribution, and query processing will provide high performance, while object-relational functionality will provide extensibility capabilities.

4 Project Status

The GDM project has just started. Based on studies of the problem area we are developing the system architecture based on Amos II and modern research on scalable data structures, stream oriented data management, extensible data managers, etc. A first prototype on a PC cluster will be developed during 2002. In parallel LOFAR/LOIS subsystems mainly for interactive selection and visualization of some of the space data are being developed. These applications will be used as benchmarks for evaluation and evolution of GDM.

GDM is expected to have significant practical impact not only on LOFAR/LOIS, but also on other applications requiring high performance data processing, while at the same time providing basis for state-of-the-art database research.

An interesting possible further development is to combine GDM with our previous work on integration of heterogeneous data sources [FR1997] [JR1999a] [JR1999b] [JR2002][LRK2001].

This would provide a high-performance extensible database engine that also could query data from many different external data sources, e.g. on the Internet and in relational databases.

5 References

[BGS2001] Ph. Bonnet, J. Gehrke, P. Seshadri: Towards Sensor Database Systems. In *Proc. of the 2nd Intl. Conf. on Mobile Data Management*. Hong Kong, January 2001.

[BW2001] S. Babu and J. Widom: Continuous Queries over Data Streams. *SIGMOD Record*, 30(3), 109-120, 2001.

[FK1999] I. Foster, C. Kesselman (eds.): *The Grid: Blueprint for a new Computing Infrastructure*, Morgan-Kaufmann, 1999.

[FOR1998] S.Flodin, K.Orsborn, T.Risch: Using queries with multi-directional functions for numerical database applications. *2nd East-European Symposium on Advances in Databases and Information Systems (ADBIS'98)*, Poznan, Poland, 7-10 September 1998. At http://www.dis.uu.se/~torer/publ/

[FR1997] G. Fahl, T. Risch: Query Processing over Object Views of Relational Data, *The VLDB Journal*, 6(4), 261-281, 1997. At http://www.dis.uu.se/~torer/publ/

[GKS2001] J. Gehrke, F. Korn, D. Srivastava: On computing Correlated Aggregates over Continual Data Streams, In *Proc. of the 2001 ACM SIGMOD Intl. Conf. on Management of Data*, pages 13-24, Santa Barbara, CA, May 2001.

[GMMO2000] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering Data Streams. In *Proc. of the 2000 Annual Symp. on Foundations of Computer Science*, pages 359-366, 2000.

[GS1992] H. Garcia-Molina, K. Salem: Main Memory Database Systems: An Overview. *IEEE Transactions on Knowledge and Data Engineering*, 4(6), 509-516, 1992.

[HTBH1995] S-O.Hvasshovd, Ö.Torbjörnsen, S.E.Bratsberg, P.Holager: The ClustRa Telecom Database: High Availability, High Throughput, and Real-Time Response, In *Proc. of the 21st VLDB Conference*, 1995.

[JR1999a] V. Josifovski, T. Risch: Functional Query Optimization over Object-Oriented Views for Data Integration, *Journal of Intelligent Information Systems (JIIS)*, 12(2-3), 1999. At http://www.dis.uu.se/~torer/publ/

[JR1999b] V. Josifovski, T. Risch: Integrating Heterogeneous Overlapping Databases through Object-Oriented Transformations, *25th Conference on Very Large Databases (VLDB'99)*, 435-446, 1999. At http://www.dis.uu.se/~torer/publ/

[JR2002] V. Josifovski, T. Risch: Query Decomposition for a Distributed Object-Oriented Mediator System. To be published in *Distributed and Parallel Databases J.*, Kluwer, May 2002. At http://www.dis.uu.se/~torer/publ/

[KLLP2001] J.S. Karlsson, A. Lal, C. Leung, T. Pham: IBM DB2 Everyplace: A Small Footprint Relational Database System, *17th International Conference on Data Engineering*, 2001. At http://www.dis.uu.se/~torer/publ/

[KLR1996] J.S. Karlsson, W. Litwin, T. Risch: LH*lh: A Scalable High Performance Data Structure for Switched Multicomputers, *Intl. Conf. on Extending Database Technology (EDBT'96)*, 1996.

[LNS1994] W. Litwin, M-A. Neimat, D. Schneider: RP*: A Family of Order-Preserving Scalable Distributed Data Structures, *20th Intl. Conf. on Very Large Databases (VLDB'94)*, 1994.

[LNS1996] W. Litwin, M-A. Neimat, D. Schneider: Scalable Distributed Data Structures, *ACM Transactions on Database Systems*, Dec. 1996.

[LPT1999] L. Liu, C. Pu, W. Tang: Continual Queries for Internet Scale Event-Driven Information Delivery, *IEEE Trans. on Knowledge and Data Engineering,* 11(14):610-628, 1999.

[LR2001] W. Litwin, T. Risch: LH*g : a High-availability Scalable Distributed Data Structure by Record Grouping. To be published in *IEEE Transactions on Knowledge and Data Engineering*. At http://www.dis.uu.se/~torer/publ/

[LRK2001] H. Lin, T. Risch, T. Katchaounov: Adaptive data mediation over XML data. To be published in of *Journal of Applied System Studies (JASS)*, Cambridge International Science Publishing. At http://www.dis.uu.se/~torer/publ/

[LS2000] W. Litwin, T. Schwarz: LH*rs: A High-Availability Scalable Distributed Data Structure using Reed-Salomon Codes, *SIGMOD Conference*, 2000.

[M+2001] J. Melton, J. Michels, V. Josifovski, K. Kulkarni, P. Schwarz, K. Zeidenstein: SQL and Management of External Data, *SIGMOD Record*, 30(1), March 2001.

[NDLR2001] Y. Ndiaye, A.W. Diene, W. Litwin, T. Risch:
AMOS-SDDS: A Scalable Distributed Data Manager for Windows Multicomputers. Presented at *14th International Conference on Parallel and Distributed Computing Systems*, Dallas, Texas, August 8-10, 2001.
At http://www.dis.uu.se/~torer/publ/

[OR1996] K. Orsborn, T. Risch: Next generation of O-O database techniques in finite element analysis. *Proceedings of The Third International Conference on Computational Structures Technology (CST 96)*, Budapest, Hungary, 21-23 August, 1996.

[Ora2001] ORACLE Inc.: *Oracle8i Lite: The Internet Platform for Mobile Computing,* http://technet.oracle.com/products/8i_lite/

[Ors1996] K. Orsborn: *On extensible and object-relational database technology for finite element analysis applications*. PhD Thesis, Thesis No. 452, ISBN 91-7871-827-9, Linköping University, Sweden, October, 1996.

[OV1999] M.T. Özsu, P. Valduriez: *Principles of Distributed Database Systems*, 2nd ed. Prentice Hall, 1999.

[RJ2001] T. Risch, V. Josifovski: Distributed Data Integration by Object-Oriented Mediator Servers, *Concurrency and Computation: Practice and Experience J.* 13(11), John Wiley & Sons, September, 2001.
At http://www.dis.uu.se/~torer/publ/

[Ron1997] M. Ronström: The NDB Cluster--A parallel data server for telecommunications applications, *Ericsson Review* No. 04, 1997.
At http://www.ericsson.com/review/1997_04/article58.shtml

[Ron1999] M. Ronström: Database Requirement Analysis for a Third Generation Mobile Telecom System.*Workshop on Databases for Telecom*, Edinburgh, Scotland, September 1999, pp. 90-105.

[SB1999] M. Stonebraker, P. Brown: *Object-Relational DBMSs: Tracking the Next Great Wave*, Morgan Kaufmann Publishers, 1999.

[Sul1996] M. Sullivan: Tribeca: A Stream Database Manager For Network Traffic Analysis, In *Proc. of the 22nd VLDB Conference,* p. 594, Mumbai, India, 1996.

[Syb2001] SYBASE Inc.: *SQL Anywhere Studio,* http://www.sybase.com/mobile/

[TGNO92] D. Terry, D. Goldberg, D. Nichols, and B. Oki: Continuous Queries over Append-Only Databases. In *Proc. of the 1992 ACM SIGMOD Intl. Conf. on Management of Data*, pages 321-330, 1992.