Hash-Based Algorithms For Operator Load-Balancing In Database Middleware Systems

WALSAIP

www.walsaip.uprm.edu

Angel L. Villalain-Garcia – M.S. Student
ADM Group, ECE Department,
Email: *angel.villalain@ece.uprm.edu*

Prof. Manuel Rodriguez – Advisor
University of Puerto Rico,
Mayagüez Campus

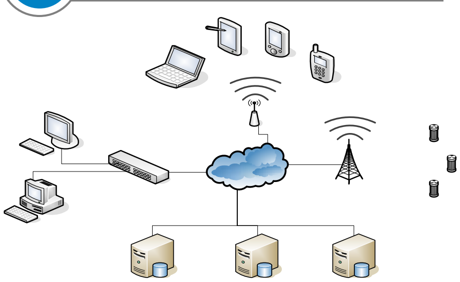
1 Project Description

Scientific applications are becoming highly dependant on Wide Area Network environments to access data and computing resources. Traditional database middleware systems have succeeded on integrating heterogeneous data sources but fail on scaling to WANs because they are focused on centralized architectures with static characteristics.

Our goal is to design a distributed database middleware system that takes into consideration the dynamic characteristics of each host on the WAN to efficiently run queries on it. We call this system NetTraveler

Our goal with NetTraveler is to design an **efficient query execution environment for distributed and parallel query execution** that improve response time based on the characteristics of the hosts on the WAN.

2 WAN Scenario



3 NetTraveler Architecture

•**Query Service Broker (QSB)**

Coordination of the Execution Process.

- •Interface exposed to the client
- •P2P behavior

•**Registration Server (RS)**

Coordinates federations and acts as a Catalog Manager.

•**Data Processing Server (DPS)**

Interface for grid services and sensors

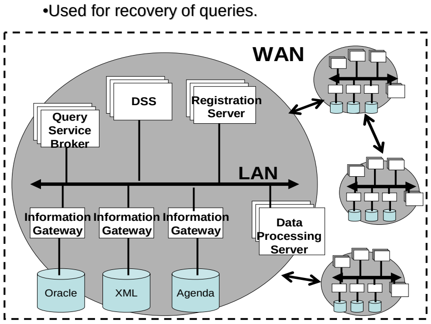
•**Information Gateway (IG)**

Interface for DB access

•**Data Synchronization Server (DSS)**

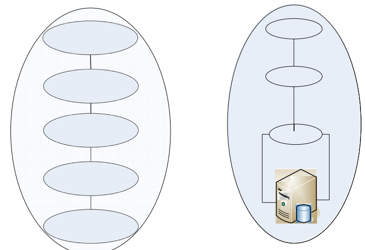
Acts as a data source when needed.

- •Used for recovery of queries.



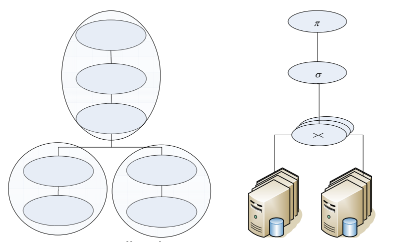
4 Proposed Work

Conventional DBMS are usually modelled as a centralized architecture with several pipelined steps. One important step is the Query Optimizer (QO). A QO is the process that determines the needed actions to solve a query. Image below represents the most important aspects of a QO and its output.



Centralized Query Optimizer

Due to several reasons as cited on [5] centralized QO fails to scale to WAN environment. With that in mind we proposed the idea of a decentralized QO that could create plans that exploits parallel and distributed execution on WAN environment (see image below).

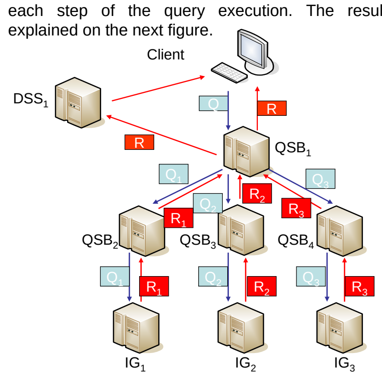


Decentralized Query Optimizer

Distributed and Parallel DBMS Plan

Both the query optimizer and the query parallelization step would need information of the available resources on each host.

Our basic assumption is that data sources are replicated, we just need to select the better host for each step of the query execution. The result is explained on the next figure.



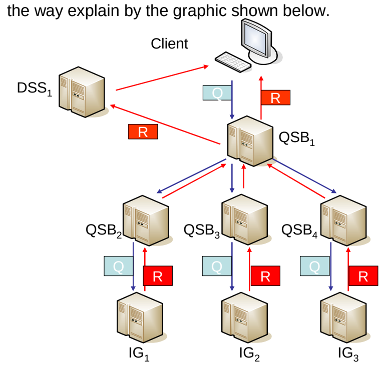
5 Methodology

Use of different Java technology for implementing our proposed architecture:

- •Web Services, Axis
- •Data access using JDBC, Hibernate

6 Project Status

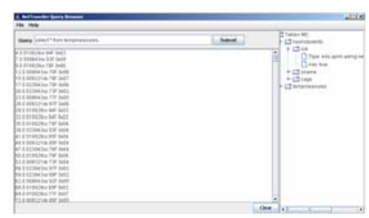
Currently the system provides an environment for remote query execution and the orchestration of the several services to bring up the result for the client in the way explain by the graphic shown below.



Actual system support for parallel query execution of operation and static distribution of the load across participating sites.

7 Future Work & Results

- •Study Scheduling Effect and improvements
- •Hash Join operators and functionalities
- •Second Phase Optimizer



A set of experiments will be performed to compare the number of queries per time unit between similar systems in order to validate our proposed solution.

8 References

1. David J. DeWitt, Shahram Ghandeharizadeh, Donovan A. Schneider, Allan Bricker, Hui-I Hsiao, and Rick Rasmussen. "The gamma database machine project". pages 609–626, 1994.
2. Sumit Ganguly, Waqar Hasan, and Ravi Krishnamurthy. Query optimization for parallel execution. In SIGMOD '92: Proceedings of the 1992 ACM SIGMOD international conference on Management of data, pages 9–18, New York, NY, USA, 1992. ACM Press.
3. Wei Hong and Michael Stonebraker. Optimization of parallel query execution plans in xprs. In PDIS '91: Proceedings of the first international conference on Parallel and distributed information systems, pages 218–225, Los Alamitos, CA, USA, 1991. IEEE Computer Society Press.
4. Manuel Rodriguez-Martinez, Nick Roussopoulos, "MOCHA: A Self-Extensible Database Middleware System for Distributed Data Sources", In Proceeding of the ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, pp. 213-224, May 2000.
5. Waqar Hasan, Daniela Florescu, Patrick Valduriez, Open Issues in Parallel Query Optimization, SIGMOD Rec., Vol. 25 No. 3 pp. 171-179, September 1996.

Supported By