Load-balancing Remote Spatial Join Queries in a Spatial GRID

Anirban Mondal Masaru Kitsuregawa

Institute of Industrial Science
University of Tokyo, Japan
{anirban,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract. The explosive growth of spatial data worldwide coupled with the emergence of GRID computing provides a strong motivation for designing a spatial GRID which allows transparent access to geographically distributed data. While different types of queries may be issued from any node in such a spatial GRID for retrieving the data stored at other (remote) nodes in the GRID, this paper specifically addresses spatial join queries. Incidentally, skewed user access patterns may cause a disproportionately large number of spatial join queries to be directed to a few 'hot' nodes, thereby resulting in severe load imbalance and consequently increased user response times. Hence, an effective load-balancing strategy becomes a necessity to prevent performance degradation. In this regard, the main contributions of our proposal are as follows. First, we present a dynamic data placement strategy involving online data replication in GRIDs, the objective being to bring the data closer to the node from which it is frequently queried. Second, we propose a novel load-balancing strategy for speeding up spatial joins in GRID environments. Our performance evaluation demonstrates the effectiveness of our proposed approach in reducing the response times of spatial joins in GRIDs.

1 Introduction

The explosive growth of spatial data worldwide coupled with the prevalence of spatial applications has made efficient management of geographically distributed spatial data a necessity. Spatial applications often arise in town planning, cartography, resource management, GIS (Geographic Information Systems), CAD (Computer-Aided Design) and computer vision. Incidentally, the emergence of GRID computing[6], which is associated with the massive integration and virtualization of geographically distributed computing resources, provides a strong motivation for designing a spatial GRID[13] which allows transparent access to geographically distributed data. While different types of queries (e.g., spatial select queries[1], nearest neighbour queries, similarity search queries and spatial join queries) may be issued from any node in the GRID for retrieving the data stored at other (remote) nodes of the GRID, this paper specifically addresses spatial joins on remote data since such queries constitute a typically expensive as well as popular class of query in spatial databases. Incidentally, a spatial join query retrieves from two spatial relations all the tuple pairs satisfying a given spatial predicate.

Now let us understand the importance of optimizing remote spatial joins with the help of an example. This year's Olympic Games is expected to attract tens of thousands of visitors to Athens and many of these visitors would possibly wish to find a hotel near a bus station for the purpose of convenient transportation. Such visitors may issue the following query from their respective home countries which may be quite far away from Athens: *Find all the hotels in Athens which are near to any bus station.* Assuming there are two relations *'Hotels'* (containing details, such as location, rental charges, of all the hotels in Athens) and *'Bus Stations'* (containing information concerning all the bus stations in Athens), this translates to a remote spatial join operation. Interestingly, the above scenario is also *equally* applicable to any major

[1] Our previous work [13] studied load-balancing of spatial select queries in a spatial GRID.

international event that attracts people from various countries. Notably, the recent trend of increased globalization has significantly increased the importance as well as the performance demands of such global applications. Unfortunately, the current state-of-the-art does *not* allow a user to perform this kind of operation efficiently.

Skews in initial data distributions, skewed user access patterns and changing popularities of data regions may cause a disproportionately large number of spatial join queries to be directed to a few 'hot' nodes, thereby resulting in severe load imbalance and consequently increased user response times. From our example, we can intuitively understand that given the huge number of queries from potential visitors to Athens, the nodes containing the data of Athens would quickly become overloaded, thereby necessitating a load-balancing mechanism for processing remote spatial joins efficiently. There are several factors which make the problem of load-balanced spatial join processing in GRIDs significantly different as well as more complicated than that of load-balanced spatial join processing in traditional distributed environments such as clusters. First, unlike in the case of clusters, wide-area communication overheads arise for GRIDs, thereby inducing higher data shipping costs and necessitating re-evaluation of existing cost-benefit formulae concerning spatial joins. Second, spatial join techniques for clusters often assume centralized control by a master node and use intelligent initial static declustering techniques (e.g., tile method[14]) to distribute spatial data over several cluster nodes, but for GRIDs, distributive ownership and inherent lack of centralized control makes it practically infeasible to control the initial data declustering. Third, heterogeneity (in terms of processing power, available disk space, communication bandwidths, indexing mechanisms, cross-domain administrative policies) being typical in GRID environments, spatial join strategies for GRIDs must also take such heterogeneity into consideration[2]. However, we believe that the time has come to deal head-on with this problem. The main contributions of our proposal are as follows.

- We present a dynamic data placement strategy involving *online* data replication in GRIDs, the objective being to bring the data closer to the node from which it is frequently queried.
- We propose a novel load-balancing strategy for speeding up spatial joins in GRID environments.

Our performance evaluation demonstrates the effectiveness of our proposed approach in reducing the response times of spatial joins in GRIDs. To our knowledge, this work is one of the earliest attempts at addressing the load-balancing of remote spatial joins via *online* data replication in GRID environments. The remainder of this paper is organized as follows. Section 2 discusses related work, while Section 3 presents the context of the problem. Issues concerning load-balancing in spatial GRIDs are presented in Section 4. The proposed strategy for load-balancing spatial joins in GRIDs is discussed in Section 5, while Section 6 reports our performance evaluation. Section 7 discusses the limits of our solution. Finally, we conclude in Section 8 with directions for future work.

2 Related Work

Important ongoing GRID computing projects such as the European DataGrid[4], the Grid Physics Network (GriPhyN)[15], Earth Systems Grid (ESG)[7] and the NASA Information Power Grid (IPG)[10] aim at efficient distributed handling of huge amounts of data (in terabyte or petabyte range). The Sloan Digital Sky Survey (SDSS) Project [19] and the Earth Observing System Data and Information System (EOSDIS) Project [16] deal with spatial datasets which are expected to be in the multi-terabyte range. Keeping in mind the demanding I/O requirements of GRID applications, the work in [20] describes a data-movement system (Kangaroo) which makes opportunistic use of resources (disks and networks), while hiding network storage devices behind memory and disk buffers such that background processes handle data movements.

[2] Existing spatial join strategies for clusters do *not* need to consider such heterogeneity.

Parallel spatial join processing has been extensively researched in the traditional domain. The proposals in [8] and [2] discuss synchronous traversal of generalization trees and R*-trees respectively. The work in [2] is extended in [3], where parallel *load-balanced* spatial join processing using R*-trees on a shared-virtual-memory architecture is investigated. The PBSM (Partition Based Spatial-Merge) algorithm [14] first partitions the inputs into smaller chunks and uses a computational geometry based plane-sweeping technique to obtain a set of candidate pairs and then the tuples corresponding to the candidate set are fetched from disk to determine whether the join condition is actually satisfied. The proposal in [11] uses seeded trees to perform spatial joins. The work in [12] proposes a parallel non-blocking spatial join algorithm which uses duplicate avoidance and addresses main memory issues. Techniques for efficiently parallelizing and dynamically load-balancing spatial databases have been proposed in [18], while several static[1] as well as dynamic[17] load-balancing techniques for clusters have also been proposed specifically for *clusters*. Notably, neither the the existing spatial join techniques nor the existing load-balancing strategies consider GRID-related issues such as heterogeneity and wide-area communication overheads essentially because these issues do *not* arise in traditional environments.

3 Problem Context

This section establishes the context of the problem by specifying an overview of our proposed system and how search is conducted in the system.

System Overview

We envisage the spatial GRID as comprising several clusters, where each cluster comprises nodes that belong to the same Local Area Network (LAN)[13]. This facilitates the separation of concerns between intra-cluster and inter-cluster load-balancing issues. Given that intra-cluster issues have been extensively researched, this paper specifically focusses on inter-cluster issues. Each cluster is assigned a unique identifier *c_id* and each node in a cluster is assigned a unique identifier *n_id*. The global identifier (GID) for any node is obtained by concatenating *c_id* and *n_id* i.e., c_id . n_id, thereby making it possible to uniquely identify any node in the system. Each query is also assigned a unique identifier *Q_id* by the node N_i where it is issued from. *Q_id* is formed by concatenating the GID of N_i with n, where n is a unique integer generated by N_i. For each cluster, the most reliable and best administered node is selected as the cluster leader, any ties being resolved arbitrarily. A cluster leader's job is to coordinate the activities (e.g., load-balancing, searching) of the nodes in its cluster.

We define distance between two clusters as the communication time τ between the cluster leaders and if the value of τ for two clusters is less than a pre-defined threshold, the clusters are regarded as *neighbours*. A cluster C_i is considered to be *relevant* to a query Q_i if C_i contains at least a non-empty subset of the answers to Q. Given that the number of queries waiting in node N_i's job queue is W_i and taking the heterogeneity in node processing capacities into account, we define L_{N_i}, the load of N_i, as follows:

$$L_{N_i} = W_i \times (CPU_{N_i} \div CPU_{Total}) \qquad (1)$$

where CPU_{N_i} denotes the CPU power of N_i and CPU_{Total} stands for the total CPU power of the cluster in which N_i is located. The *load of a cluster* is calculated as $\sum L_{N_i}$ i.e, the sum of the loads of its individual members. Given that the loads of two clusters C_i and C_j are L_{C_i} and L_{C_j} respectively and assuming without loss of generality that $L_{C_i} > L_{C_j}$, the normalized load difference Δ between C_i and C_j is computed as follows:

$$\Delta = ((L_{C_i} \times TC_i) - (L_{C_j} \times TC_j))/(TC_i + TC_j) \qquad (2)$$

where TC_i is the sum of the CPU power of all the members of C_i. Similarly, TC_j is the sum of the CPU power of all the members of C_j.

Indexing mechanism

Spatial indexing mechanisms may vary across clusters. Hence, we propose a generalized indexing scheme which is built on top of the existing index at a cluster node. The indexing scheme in each cluster comprises three index structures, namely IID (Index structure for Internal Data), IED (Index structure for External Data) and IRD (Index structure for Replicated Data). Note that we distinguish between a cluster's *internal data* (the data which is originally stored at a cluster) and its replicated data because it may *not* be possible to integrate the replicated data smoothly into the existing index structure (for the internal data) of the cluster as the replicated data may be far apart in space from the cluster's internal data. Such separation of concerns between internal data and replicated data also makes it easier to periodically delete infrequently accessed replicas for optimizing disk space usage.

IID is a *generalized* two-tier indexing mechanism, the first-tier of which resides at the cluster leader C_i and is essentially a list, each entry of which is of the form (*region*,*node_id*), where *region* represents a specific region and *node_id* stands for the node in the cluster at which the region is located. At the second tier, every node has its own independent index structure for the data allocated to it. For example, in case of R-trees[9], rectangular-shaped regions would be stored in the first-tier at C_i, while the second-tier would comprise R-trees at the individual nodes. IED and IRD are hierarchical tree-based index structures, which reside at the cluster leader. In our example, IED/IRD would be R-tree-like structures except that their leaf nodes would contain cluster IDs of neighbouring clusters' data regions instead of pointers to objects in the database. Updates to IED/IRD are periodically exchanged between neighbouring cluster leaders preferably via piggybacking onto other messages.

Given the above indexing scheme, search is conducted as follows. When a cluster leader C_i receives a query Q, it first checks its IID and IRD to ascertain whether any of its cluster nodes is relevant to Q. If C_i finds that *none* of its nodes is relevant to Q, it checks its IED and sends Q to its neighbouring cluster leaders which are relevant to Q. In case *none* of its neighbouring cluster leaders contain the answers to Q, C_i broadcasts Q to *all* of them. This process continues till either the answers to Q are retrieved or Q is timed-out.

Our Focus

Assume the existence of n clusters in the GRID, $C_1, C_2, C_3, ...C_n$. Now suppose cluster C_i issues a spatial join query Q_i for the data in cluster C_j. A straightforward solution would be for C_j to process Q_i and return the results to C_i, but this solution would *not* be efficient in scenarios where many spatial join queries, which attempt to retrieve data from the same regions in C_j, are issued from C_i to C_j. Our primary focus is to reduce the response times of such spatial join queries on remote data via replication.

4 Issues concerning Load-balancing of Spatial Joins via Replication in GRIDs

Several important issues need to be addressed when supporting load-balancing of spatial joins via replication in GRIDs. This section discusses how we address these issues.

Hotspot detection

The heat of data regions should be defined with respect to clusters which issue queries for these regions. For example, if cluster C_S issues a large number of queries pertaining to data region D_i of cluster C_T, but cluster C_O issues no queries for D_i, D_i will be considered a 'hot' region only w.r.t. C_S (but *not* w.r.t. C_O). Understandably, many different clusters accessing D_i infrequently would make D_i a 'hot' region in the conventional sense, but replicating D_i at any of these clusters may *not* necessarily be useful for reducing response times and may indeed be counter-productive owing to replication-related overheads.

For hotspot detection purposes, every node within a cluster maintains its own access statistics comprising a list $HotList$, each entry of which is of the form (*data*,ptr_{access}). Here *data*

represents a specific data region and ptr_{access} is a pointer to a three-dimensional array of the form (cluster_id,*num*,*avgtime*), where cluster_id stands for the ID of the particular cluster which accessed *data*, *num* indicates the number of times that cluster accessed *data* and *avgtime* represents the average processing time it took for a node to perform spatial join on *data*. The value of *avgtime* is used in ascertaining the benefit of replicating a specific data region as we shall see in Section 5. This information is periodically sent by every node to its cluster leader, thereby enabling the cluster leader to determine the 'popularity' of different data regions with respect to different clusters.

However, keeping counts of accesses to data regions over time is *not* fair and may *not* give an accurate reflection of the actual hotspots. For example, a spatial join query Q_1 may have queried the system 10000 times in a particular day during the last month, and thereafter it is never asked again. Other queries that query the system 100 times on each day of the current month will require 100 days before they can qualify to be associated with hotspots. This is clearly undesirable because it gives little importance to the current hotspots and attaches high importance to those portions of the data that were hotspots in the past. In order to accurately reflect true hotspots at the current point in time, we propose that $HotList$ should be initialized periodically i.e., after every n_t time units, the information in $HotList$ should be deleted and then $HotList$ should be populated with fresh access information. The value of n_t primarily depends on how quickly access patterns change and is application-dependent. When an overloaded source cluster node N_i has completed offloading some part of its load to a node at another cluster, N_i refreshes its own $HotList$ by deleting those entries in $HotList$ which triggered the replication. This enables $HotList$ to reflect *only* the current hotspots concerning which action has *not* yet been taken. Replication being an expensive operation in GRIDs, true reflection of hotspots is necessary to prevent triggering of spurious and unnecessary replications. In essence, keeping in mind the inherent dynamism of GRID environments, we use *only* the most recent access statistics to determine popular regions.

The granularity at which access statistics concerning regions are maintained essentially represents a trade-off between accuracy of hotspot detection and the overhead of maintaining such statistics. We maintain access statistics information at the granularity of the respective leaf node levels of the index structures at the nodes. Throughout this paper, we shall use the term **data region** to indicate the spatial region corresponding to the Minimum Bounding Rectangle (MBR) of the data stored at a leaf node of the IID at a particular node. In this paper, we do *not* specifically addressed issues concerning the optimal granularity at which access statistics information should be maintained, but we intend to address this issue in detail in the near future. An interesting question which arises here is: *Given that several different kinds of queries can be issued to a real system, how do we know whether the leaf node accesses are being made specifically for a spatial join query?* Incidentally, at the leaf node level, it is *not* feasible to determine the type of query for which the leaf node is being accessed, but this information can be found at the query engine level.

What to replicate?

Once a 'popular' region D_i w.r.t. a specific cluster C_i has been detected, our strategy is to replicate the *results of the spatial join operation on* D_i at C_i. Note that replicating the results (as opposed to replicating the data itself) can significantly benefit those subsequent spatial join queries whose spatial select windows have considerable overlap with D_i since C_i will *not* need to do any processing at all for a significant part of these queries. Additionally, replicating the results can reduce communication cost significantly if join selectivity of D_i is low. Even in case of high join selectivity of D_i, intuitively we can understand that the communication cost of replicating the result tuples can never exceed that of replicating the data itself. As in existing GRID efforts[20], our strategy assumes that the datasets are relatively static. Note that we use 'data replication' and 'replication of result tuples' interchangeably throughout this paper to imply replication of result tuples.

Exploiting overlap between different spatial join queries

Interestingly, every spatial join query has an associated MBR associated with it either explicitly or implicitly. We shall designate this MBR as SPJMBR (Spatial Join's Minimum Bounding Rectangle). For example, the join query *"Find a hotel near a station in Athens within a 5 km radius of X, where X is a certain landmark in Athens"* explicitly specifies the SPJMBR associated with the join query, while the query *"Find a hotel near a bus station in Athens"* implicitly specifies that the SPJMBR for this query corresponds to the MBR of Athens. Intuitively, efficient exploitation of overlaps between different spatial join queries requires a mechanism for storing the replicas in a manner which enables quick identification of overlap between spatial join queries and existing replicas.

In our proposed system, whenever a replica of result tuples is stored at a cluster, the cluster leader also stores the SPJMBR corresponding to that replica. Identification of overlap between an SPJMBR and a spatial join query Q can be classified into 3 cases: (a) Q's MBR does *not* intersect with SPJMBR: This implies that there is *no* overlap between the query and the existing replica. (b) Q's MBR is fully contained within the SPJMBR: This means that all the results tuples requested by Q are already in the stored replica and only a spatial select query using Q's MBR as the spatial select condition should be run on the replicated data to obtain the answers to Q. We propose to run this spatial select condition on the replicated data at the cluster C_{Rep} where the replicated data exists (if C_{Rep} is *not* overloaded and has sufficient disk space) to save communication costs. However, if C_{Rep} is overloaded and/or C_{Rep} has insufficient available disk space, the replica is sent to the cluster C_{Issue} which issued Q and C_{Issue} needs to run Q's MBR as the spatial select query on the replica to obtain the query results. (c) Q's MBR partially intersects with SPJMBR: The implication is that the results of Q already exist for the intersecting part between Q's MBR and SPJMBR, but for the non-intersecting parts, the results need to be computed. In this case, the tuples in the intersecting part are sent to the query issuing peer, while a spatial join operation needs to be run to get the result tuples in the non-intersecting parts between Q's MBR and SPJMBR.

Whenever a spatial join query Q arrives at a cluster leader, the cluster leader traverses its list of SPJMBRs and identifies and exploits overlaps (if any) in the manner stated above. Additionally, in order to optimize disk space usage, each cluster leader keeps track of the replicas in its cluster nodes as well as the number of accesses made to each of the replicas during recent time intervals. Replicas whose access frequency during recent time intervals falls below a pre-defined threshold are deleted because the valuable disk space consumed by such unused replicas can be put to better use by storing 'hot' data, thereby improving system performance.

5 Load-balancing Strategy for Spatial Joins in GRIDs

In our proposed strategy, a cluster leader determines itself to be overloaded if its load exceeds the average loads of its neighbouring clusters by more than 15%. When a cluster leader determines itself to be overloaded, it periodically checks the frequency with which its **data regions**[3] are being queried by other cluster leaders during the recent time intervals. Based on this information, the cluster leader C_i creates a set ψ comprising all cluster leaders which have issued more than η queries for *any* of its regions. (η is a threshold parameter which influences the sensitivity of load-balancing.) C_i sends a message to each member of set ψ informing its own disk space requirement (i.e., the amount of disk space required to store the replicated data) and requesting information concerning their load status, list of neighbours and whether their available disk space is sufficient to store the replicated data. After receiving the necessary status information of ψ's members, C_i evaluates their replies one-by-one. Members of ψ whose available disk space is too low to store the replicated data or whose normalized load difference

[3] Recall that 'data region' refers to the spatial region corresponding to the MBR of the data stored at a leaf node of the IID at a particular node.

```
Algorithm LB_OverloadedSource( )
Check number of queries issued for each of its regions by other cluster leaders
Create a set ψ comprising cluster leaders which have issued more than η queries for any of its regions
if (ψ is an empty set) {
  exit
} else {
  for each element α in set ψ {
     Send message to α informing α its disk space requirement and asking α's disk space, load and neighbours' list
     Receive message from α concerning whether α's disk space is sufficient and if so, α's current load
     and α's list of neighbours List_Neighbours and if not, List_Neighbours
     if  ( ( α's disk space is NOT sufficient ) OR ( Δ ≤ LOAD_THRESHOLD ) ) {
     /* Δ is the normalized load difference between itself and α */
        Delete α from set ψ and Add members of List_Neighbours to ψ
        for each member NG of List_Neighbours {
           Send message to NG informing its disk space requirement and asking NG's disk space availability
           and current load
           Receive message from NG concerning whether NG's disk space is sufficient and NG's current load
           if ( ( NG's disk space is NOT sufficient ) OR ( Δ ≤ LOAD_THRESHOLD ) ) {
             Delete NG from ψ
           }
        }
     }
  }
  for each element X in ψ {
     for each data region H queried by X {
        if ( cost_benefit_analyze ( ) = = TRUE ) {
          Put X into a temporary list designated as 'temp'
        }
     }
  }
  Select_dest_from_temp( )
}
end
```

Fig. 1. Load-balancing Algorithm executed by an overloaded source cluster leader

with C_i (Δ) falls below a pre-specified threshold are deleted from ψ. For such members, C_i adds their list of neighbouring clusters to ψ. For these neighbouring clusters, clusters with low disk space or those with low normalized load difference with C_i are deleted from ψ. The remaining members of ψ are candidates for replication. For each member X of ψ, C_i traverses each hot data region H that has been queried by X and decides whether to replicate the spatial join result tuples associated with H on a case-by-case basis. Now let us see how C_i makes this decision. The total cost C_H of replicating H from C_i at X consists of the cost $Extr_H$ of extracting H from IID at C_i, the communication cost Cm_H of transferring H and the bulkloading cost $Bulk_H$ of integrating H into the IRD of X. Hence, C_H is given by the following formula:

$$C_H = Extr_H + Cm_H + Bulk_H \tag{3}$$

For evaluating $Extr_H$ and $Bulk_H$, we ran experiments on an R-tree for which the number of branches emanating from each node was 64. In the interest of space, we only present a small part of the tables here to indicate how we evaluate $Extr_H$ and $Bulk_H$. Figure 2 depicts the values of $Extr_H$ and $Bulk_H$ for different values of NUM, where NUM is the number of data rectangles to be extracted/bulkloaded. For example, if we need to extract p data rectangles from the IID at a node such that $8192 < p \leq 12288$, we use a value of 918 milliseconds. If we need to bulkload q data rectangles into an R-tree structure such that $16384 < q \leq 20480$, we use a value of 3005 milliseconds.

NUM	$Extr_H$ (ms)
4096	583
8192	688
12288	918
16384	977
20480	1465

(a) Extraction Cost

NUM	$Bulk_H$ (ms)
4096	1459
8192	2226
12288	2467
16384	2636
20480	3005

(b) Bulkloading Cost

Fig. 2. Table for evaluating $Extr_H$ and $Bulk_H$

Recall that every cluster leader maintains information concerning the average processing time and the accesses made to each data region from each of the clusters in the system. Let n_H denote the number of times H has been accessed by X and $avgtime_H$ represents the average processing time of H. Hence, the benefit B_H of replicating H at X can be estimated as follows:

$$B_H = (n_H \times avgtime_H) \tag{4}$$

From (3) and (4), we have the following formula:

$$Decide_H = (B_H - C_H \geq TH_{min}) \tag{5}$$

where TH_{min} is a pre-defined threshold parameter which is essentially application-dependent and on which the degree of load-balancing depends and $Decide_H$ is a boolean variable. In our algorithm, we use a function **cost_benefit_analyze( )** to evaluate the value of $Decide_H$ using the formulae (3),(4) and (5). If $Decide_H$ evaluates to 'TRUE', **cost_benefit_analyze( )** returns 'TRUE', otherwise it returns 'FALSE'. Every member X of ψ for which **cost_benefit_analyze( )** returns 'TRUE' is put into a temporary list data structure which we shall designate as *'temp'*. The data structure of 'temp' is essentially a list structure where for each 'hot' data region H, the corresponding destination candidates (those members of ψ for which **cost_benefit_analyze( )** had returned 'TRUE') for H are stored in a linked list. Using the 'temp' data structure, the overloaded cluster leader uses a function **Select_dest_from_temp( )** which selects (as the destination cluster) the least loaded member in 'temp' (corresponding to each H) for each H. The load-balancing algorithm executed by an overloaded source cluster leader is depicted in Figure 1, while the load-balancing algorithm executed by a potential destination cluster leader is presented in Figure 3.

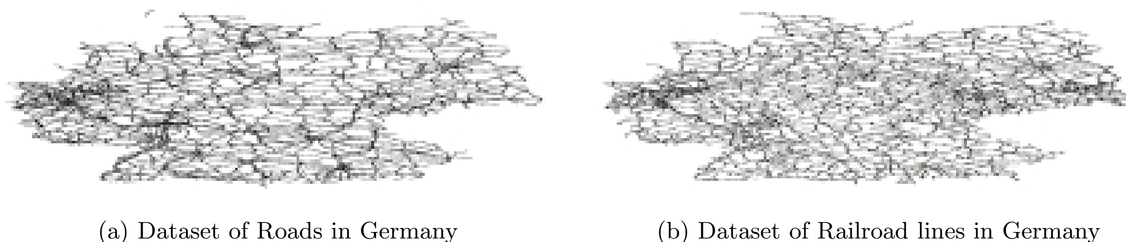
Observe that in contrast with existing works in traditional environments, our strategy does *not* use the value of normalized load difference when deciding upon the amount of data to replicate. This is because in our scenario, the increase in load at X owing to spatial join queries on H is negligible (even in case of spatial select conditions) as compared to the decrease in load for C_i especially since the join has already been computed. Moreover, note that replication is initiated from C_i to X whenever the normalized load difference between C_i and X exceeds a given threshold, irrespective of whether C_i is *really* overloaded or not. Even if C_i is *not* overloaded, we believe it is still reasonable to replicate at X since bringing the data closer to the cluster from where the data are being frequently queried implies a reduction in network overheads (as well as response times) for future spatial joins on the same data.

6 Performance Study

This section reports the performance evaluation of our proposed inter-cluster load-balancing technique via replication of result tuples of spatial join queries. Note that we consider performance issues associated *only* with inter-cluster load-balancing since a significant body of research work pertaining to efficient intra-cluster load-balancing algorithms already exists. Hence, for our experiments, we use a cluster size of 1. The machine used for the experiments

Algorithm LB_PotentialDestination( )
Receive message from overloaded source cluster leader SRC /* *The message contains disk space requirement of* SRC */
Send a Broadcast message to all the nodes in its cluster asking each node for its current load and disk space
Receive replies concerning current load and disk space of each node
Nodes with sufficient available disk space and load below a pre-defined threshold λ are put into a set $Candidate$
if ($Candidate$ is an empty set) {
  Send message to SRC stating that its disk space is insufficient and informing SRC about its list of neighbours
} else {
    Send message to SRC informing SRC about its sufficient disk space, its current load and its list of neighbours
}
Receive reply from SRC
if (SRC has selected it as the destination cluster) {
  Send a Broadcast message to the nodes in $Candidate$ for their current load status
  Receive the corresponding replies and select the least loaded node MIN from $Candidate$
  Send a message to SRC to replicate the data at MIN
}
end

Fig. 3. Load-balancing algorithm executed by each potential destination cluster leader



(a) Dataset of Roads in Germany (b) Dataset of Railroad lines in Germany

Fig. 4. Image of datasets

had processing capacity of 1.7 GHz (Pentium-4), main memory of 768 Mbytes and disk space of 40GB. We ran the experiments under the Redhat Linux (version 7.3) operating system using LAM-MPI (version 7.00) for message-passing. In order to model inter-cluster communication in a wide area network environment, we assigned transfer rates for communication between cluster leaders randomly in the range of 0.8 Megabit/second to 1.2 Megabit/second. We used a maximum of 3 neighbouring cluster leaders corresponding to each cluster leader. The number of clusters simulated in our experiments was 24. The interarrival time between queries arriving at a cluster was fixed at 10 milliseconds and the value of TH_{min} was set to 5 seconds.

We have used two *real-life* datasets [5] for our experiments. The first dataset is the set of roads in Germany, while the second one is a dataset of railway lines in Germany. The first dataset comprises MBRs of 30,674 streets of Germany, while the second one consists of MBRs of 36,334 railroad lines in Germany. Figure 4 depicts the datasets. We had enlarged each of these datasets by translating and mapping the data for the purpose of our experiments. For our experiments, each of the clusters had more than 200000 rectangles for each of the relations. We used two R-trees at each cluster, one for each dataset. We assumed that one R-tree node fits in a disk page (page size = 4096 bytes). Hence, R-tree node capacity is the same as page size in our case. The height of each of the R-trees was 3 and the fan-out was 64.

We generated queries for each cluster by using a spatial select (window) condition in conjunction with the spatial join. Note that this is in consonance with real-world scenarios where spatial joins may be quite often accompanied by certain select conditions. The selectivity of each spatial join query was fixed at 40%. Assuming n queries for a particular cluster C_i, let

us designate the queries as $Q_1, Q_2, ...Q_n$. We generated the n queries for C_i such that the queries had *at least* 75% overlap with each other. This overlap was generated by shifting the respective spatial select query windows in such a manner that each query had $x\%$ (where $x \geq 75\%$) overlap with the other queries.

For performing the spatial join operation at each cluster, we use an existing approach where the data from the smaller fragment is extracted and used to probe the index structure corresponding to the larger fragment. For the sake of convenience, we shall refer to our proposed technique as LBREP (Load-balancing via replication). Since *no* work on load-balanced processing of remote spatial joins in GRIDs exists, we shall compare the performance of LBREP with a technique which performs spatial join without load-balancing. We designate this reference technique as NOLB (No load-balancing). For all our experiments, we had run the system for an initial period of time to obtain access statistics information and once the system had reached a stable state (after the replication of result tuples have been performed), we noted down the results. We only present results associated with the stable state of the system.

C_{Source}	$C_{Destination}$
1	24, 15
2	23,17
3	22, 14
4	21, 18
6	20, 15
8	19, 14
9	18, 10
11	17,15
12	16, 14
13	14, 10

(a) Table indicating replication

N	C_E	$C_Q(f)$
16	1	24(9),15(7)
12	2	23(6), 17(6)
12	3	22(6), 14(6)
4	4	21(3), 18(1)
4	6	20(3),15(1)
4	8	19(3), 14(1)
4	9	18(3), 10(1)
4	11	17(3),15(1)
4	12	16(3), 14(1)
4	13	14(3), 10(1)

(b) $QD1$

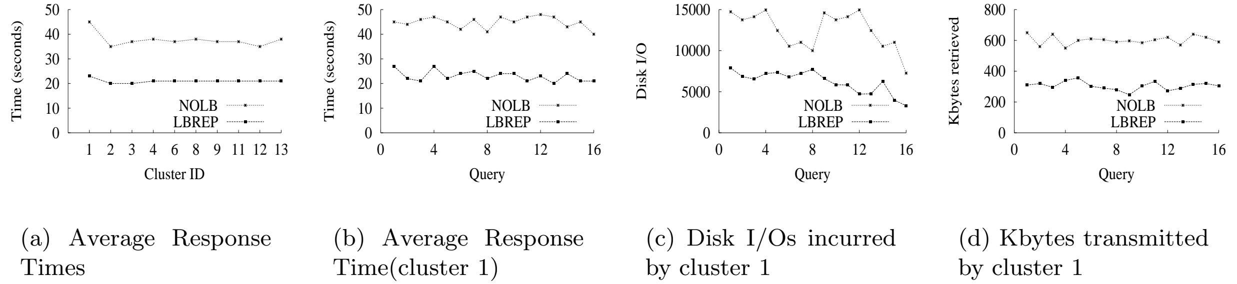
N	C_E	$C_Q(f)$
16	1	15(10), 24(6)
12	2	17(4), 23(8)
12	3	14(3), 22(9)
4	4	18(3), 21(1)
4	6	15(3), 20(1)
4	8	14(3), 19(1)
4	9	10(3), 18(1)
4	11	15(3), 17(1)
4	12	14(3), 16(1)
4	13	10(3), 14(1)

(c) $QD2$

Fig. 5. Replication table and $QD1$ and $QD2$ for a 24-cluster GRID

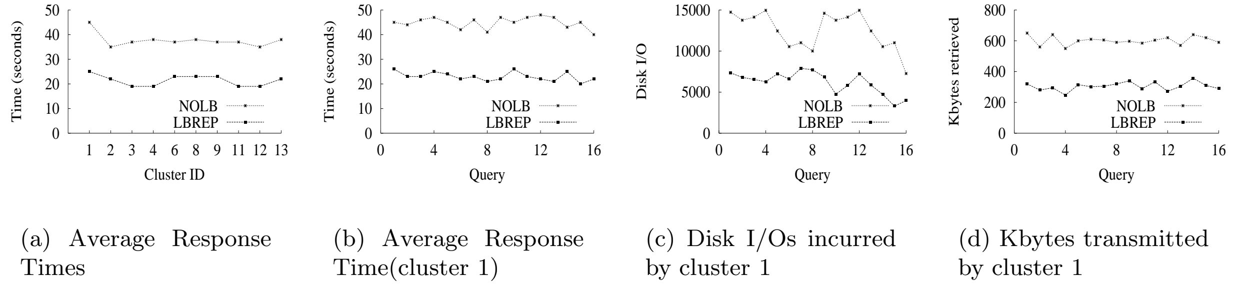
Performance evaluation for a 24-cluster GRID

The replications that have already been performed (based on access statistics information) prior to the system reaching stable state are depicted in Figure 5(a). In Figure 5(a), C_{Source} represents the IDs of the source cluster whose data (spatial join result tuples) have been replicated, while $C_{Destination}$ stands for the IDs of the destination clusters where C_{Source}'s data has been replicated. For example, the first row of the table indicates that a portion of cluster 1's data has been replicated at clusters 24 and 15. Similarly, a part of cluster 2's data has been replicated at clusters 23 and 17 and so on. We shall use this convention to indicate replication information throughout the remainder of this paper. Note that the portions of cluster 1's data that have been replicated at clusters 24 and 15 need *not* necessarily be the same, even though overlap is possible between the replicated data of cluster 1 at cluster 24 and cluster 15. This is because the replication performed was based on previous access statistics, thereby implying that the replicated data at different clusters depends upon the queries that these clusters had issued during the past. Now we shall evaluate the relative performance of LBREP and NOLB by using different query distributions. Even though we had used several query distributions to test the robustness of LBREP, in the interest of space, here we present only 4 such distributions. For the sake of convenience, we shall refer to these query distributions as $QD1$,$QD2$,$QD3$ and $QD4$ respectively. Figures 5b and 5c summarize $QD1$ and $QD2$, while Figure 8 presents $QD3$ and $QD4$. In Figure 5b, N denotes the number of queries, C_E indicates



(a) Average Response Times (b) Average Response Time(cluster 1) (c) Disk I/Os incurred by cluster 1 (d) Kbytes transmitted by cluster 1

Fig. 6. Results on $QD1$ for a 24-cluster GRID



(a) Average Response Times (b) Average Response Time(cluster 1) (c) Disk I/Os incurred by cluster 1 (d) Kbytes transmitted by cluster 1

Fig. 7. Results on $QD2$ for a 24-cluster GRID

the ID of the cluster which processed the queries, C_Q represents the IDs of the clusters which issued those queries and f stands for the number of queries issued by a cluster. Note that the sequence of the queries arriving at each cluster is also specified by Figure 5b. For example, the first row of the table in Figure 5b indicates that 16 queries (let us designate them as $Q1$ to $Q16$) were processed by cluster 1. $Q1$ to $Q9$ were issued by cluster 24, $Q10$ to $Q16$ were issued by cluster 15. In contrast, the first row in Figure 5c indicates that $Q1$ to $Q10$ were issued by cluster 15, while cluster 24 issued $Q11$ to $Q16$. We shall use this convention to indicate query distributions throughout the remainder of this paper. Owing to space constraints, we are *not* able to present the detailed results concerning *all* the queries in the system. Note that the selectivity of each spatial join query in each of the 4 query distributions was fixed at 40%. In all our experiments, cluster 1 is the most overloaded (hot) cluster and also it was the last cluster in the GRID to complete processing. Hence, we shall specifically examine details concerning the processing of queries that were directed to cluster 1.

Results concerning $QD1$ and $QD2$ Figures 6 and 7 depict the results corresponding to $QD1$ and $QD2$ respectively. Figure 6a indicates the average response times of *all* the queries directed to each cluster. The results demonstrate that LBREP is indeed able to decrease the average response times for each of the clusters significantly, especially decreasing the average response time of cluster 1 by upto 48%. The reduction in average response times occurs because of the reduction in disk I/O overhead at the query executing clusters as well as the reduction in communication overhead arising from transmission of result tuples to the clusters which issued the respective queries.

To put things into perspective, we take a closer look at the processing of the 16 queries that were directed to cluster 1. Figure 6b depicts the individual response times of each of the 16 queries that were directed to cluster 1 for $QD1$, while Figure 6c shows the corresponding disk I/Os incurred for each query at cluster 1 for the same experiment. Figure 6d indicates the number of KBytes for each query that cluster 1 had to transmit to the cluster which

had issued the query. Observe that Figure 6b indicates that for *all* the queries directed to cluster 1, LBREP's performance is superior to that of NOLB in terms of response times. Such reductions occur because part of the results of the spatial join have already been replicated at clusters which issued these queries (clusters 24 and 15 in this case). The implication is that cluster 1 did *not* need to process a significant part of each of these queries, thereby resulting in reduction of disk I/O cost incurred by cluster 1. Moreover, since clusters 24 and 15 already had a part of the results associated with the queries that they issued, the number of result tuples that cluster 1 had to transmit to such clusters was also reduced, thereby reducing the communication overhead. Detailed investigation of the experimental results revealed that the reduction in disk I/O cost varied between 45% to 54%, while reduction in the total size of result tuples transmitted to the querying clusters varied between 46% to 52%. However, note that the price LBREP pays for improvements in response time is additional disk space usage since replication causes redundant usage of disk space. We believe that the overhead of additional disk space usage is justifiable because of the significant improvement in response times of spatial joins that LBREP provides.

The explanations for Figure 6 also hold good for the results in Figure 7. Observe that the performance of NOLB remains same in case of Figures 6 and 7 because in case of NOLB, no data has been replicated at the querying clusters, thereby implying that every query is completely processed at the query executing cluster and then the results are sent back to the querying clusters. We also find that the results in Figures 6 and 7 differ to some extent for LBREP. This is because the portions of cluster 1's data replicated at clusters 24 and 15 were *not* exactly the same, even though there was overlap between those portions. Notably, in case of both $QD1$ and $QD2$, the queries are issued by clusters which already had some portion of the results replicated at themselves i.e., $QD1$ and $QD2$ represent cases when the query distribution is in accordance with the access statistics information that had earlier prompted the replications.

N	C_E	$C_Q(f)$
16	1	24(10),7(6)
12	2	23(8),14(4)
12	3	22(8),16(4)
4	4	21(2),17(2)
4	6	20(2),14(2)
4	8	19(2),20(2)
4	9	18(2),7(2)
4	11	17(2),18(2)
4	12	16(2),13(2)
4	13	14(2),24(2)

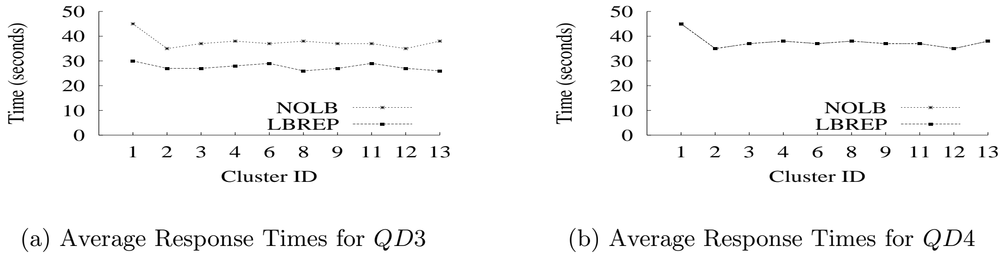
(a) $QD3$

N	C_E	$C_Q(f)$
16	1	14(10),22(6)
12	2	16(4),21(8)
12	3	7(3),19(9)
4	4	17(3),20(1)
4	6	13(3),16(1)
4	8	24(3),18(1)
4	9	7(3),17(1)
4	11	14(3),12(1)
4	12	13(3),17(1)
4	13	7(3),16(1)

(b) $QD4$

Fig. 8. $QD3$ and $QD4$ for a 24-cluster GRID

Results concerning $QD3$ and $QD4$ Now let us evaluate the performance of LBREP when the query distribution is *not* strictly in accordance with the access statistics information that had earlier prompted the replications. $QD3$ (see Figure 8a) is a query distribution which shows some deviation from the previous access statistics, while $QD4$ (see Figure 8b) is a query distribution which is *totally* different from the previous access statistics. The results corresponding to $QD3$ and $QD4$ are depicted in Figure 9. We find that for $QD3$, LBREP still outperforms NOLB significantly, but the gain in average response time for $QD3$ is lower than the gains in case of $QD1$ and $QD2$. This is because in case of $QD1$ and $QD2$, all the



(a) Average Response Times for $QD3$ (b) Average Response Times for $QD4$

Fig. 9. Results on $QD3$ and $QD4$ for a 24-cluster GRID

queries were speeded up owing to replicated data existing at the query issuing clusters. In contrast, for $QD3$, only some of the queries were facilitated by replication, while the other queries (which deviated from the past access statistics) were *not* facilitated by replication. For $QD4$, the performance of LBREP and NOLB is comparable since *none* of the querying clusters had any replicated data of *any* of the query executing clusters.

From Figures 6, 7 and 9, we find that LBREP significantly outperforms NOLB as long as the query distribution is in accordance with the access statistics information based on which replications had been earlier performed. Even when the query distribution is *not* in accordance with the access statistics information, LBREP does *not* perform worse than NOLB in terms of response times. In essence, LBREP is robust to different query distributions and adapts very well indeed.

7 Discussion

This section discusses the limits of our proposed solution, the objective being to identify scenarios in which our solution encounters problems. Our algorithms assume that if a cluster has sufficient available disk space, the cluster leader will be willing to store the replicated data, but in certain situations, the cluster leader may *not* be willing. For example, when there is no collaborative agreement between the organizations that own the clusters, the system administrator of the destination cluster may not be willing due to administrative reasons and/or issues concerning security and trust. Alternatively, there may be situations when the system administrator of a potential destination cluster knows that a huge amount of data (from other sources) will soon be arriving at its own cluster, thereby preventing him from being willing to store another cluster's replicated data. Notably, issues concerning trust/security and the occurrence of external events is well beyond the scope of this paper, but nevertheless, these issues influence the deployment of our proposed solution. These issues are not just specific to the problem of load-balanced spatial join processing in GRIDs, but are applicable to GRIDs in general. Interestingly, several organizations all over the world have already signed agreements between themselves and we hope that more organizations will participate in such agreements because such agreements are of paramount importance for GRIDs to work efficiently and effectively in practice.

Incidentally, the ramifications of load-balancing in GRIDs differ significantly from that of load-balancing in the conventional sense. In traditional cluster environments, nodes typically collaborate with each other, the objective being to distribute the load as uniformly as possible[4]. Incidentally, uniform load distribution across all the clusters in a GRID is *not* always possible and neither is it necessary for providing reduced query response times. For example, even in case of severe load imbalance among the clusters of the GRID, the overloaded clusters may *not* be able to replicate a part of their hot data to the underloaded clusters possibly because the

[4] Uniform load distribution in clusters is possible partly owing to low communication overheads associated with local area networks.

underloaded clusters may be too far away (in terms of communication time) for the replication to be beneficial. Also, if *none* of the underloaded clusters are issuing any queries for the data stored at any of the overloaded clusters, replicating hot data from any of the overloaded clusters to these underloaded clusters will *not* reduce query response times and may indeed be counter-productive owing to high wide-area communication overheads. In particular, we believe that the critical point concerning load-balancing, from the perspective of optimizing response times, is to replicate hot data regions at or near to clusters which issue a large number of queries for those data regions.

8 Conclusion

Huge amounts of available spatial data worldwide and the prevalence of spatial applications, coupled with the emergence of GRID computing, provides a strong motivation for designing a spatial GRID. Skewed user access patterns may cause severe load imbalance in the system, thereby degrading system performance significantly. Our proposal has specifically focussed on speeding up remote spatial joins in this environment via a novel dynamic load-balancing strategy which deploys online replication. To our knowledge, this work is one of the earliest attempts at addressing the load-balancing of remote spatial joins via *online* data replication in GRID environments. Our performance study demonstrates the effectiveness of our proposed approach in reducing the response times of spatial joins in GRIDs. In the near future, we intend to extend this work by investigating issues concerning the optimal granularity of maintaining access statistics information for hotspot detection. Additionally, we plan to address issues concerning the cost-effective integration of LBREP into existing GRIDs.

References

1. H. Boral, W. Alexander, L. Clay, G. Copeland, S. Danforth, M. Franklin, B. Hart, M. Smith, and P. Valduriez. Prototyping Bubba, a highly parallel database system. *Proc. IEEE TKDE*, 2(1), March 1990.
2. T. Brinkhoff, H.P. Kriegel, and B. Seeger. Efficient processing of spatial joins using R-trees. *Proc. ACM SIGMOD*, pages 237–246, 1993.
3. T. Brinkhoff, H.P. Kriegel, and B. Seeger. Parallel processing of spatial joins using R-trees. *Proc. ICDE*, pages 258–265, 1996.
4. European DataGRID. http://eu-datagrid.web.cern.ch/eu-datagrid/.
5. Datasets. http://dias.cti.gr/∼ytheod/research/datasets/spatial.html.
6. I. Foster and C. Kesselman. The GRID: Blueprint for a new computing infrastructure. *Morgan-Kaufmann*, 1999.
7. Earth Systems GRID. http://www.earthsystemgrid.org/.
8. O. Gunther. Efficient computation of spatial joins. *Proc. ICDE*, pages 50–59, 1993.
9. A. Guttman. R-trees: A dynamic index structure for spatial searching. *Proc. ACM SIGMOD*, pages 47–57, 1984.
10. NASA IPG. http://www.ipg.nasa.gov/.
11. M. L. Lo and C. V. Ravishankar. Spatial joins using seeded trees. *Proc. ACM SIGMOD*, pages 209–220, 1994.
12. G. Luo, J. F. Naughton, and C. Ellmann. A non-blocking parallel spatial join algorithm.. *Proc. ICDE*, 2002.
13. A. Mondal, K. Goda, and M. Kitsuregawa. Effective load-balancing via migration and replication in spatial GRIDs. *Proc. DEXA*, 2003.
14. J. Patel and D. DeWitt. Partition based spatial-merge join. *Proc. ACM SIGMOD*, pages 259–270, 1996.
15. GriPhyN Project. http://www.griphyn.org/index.php.
16. NASA's EOSDIS Project. http://eospso.gsfc.nasa.gov/.
17. P. Scheuermann, G. Weikum, and P. Zabback. Disk cooling in parallel disk systems. *IEEE Bulletin of the Technical Committee on Data Engineering*, 17(3):29–40, 1994.
18. S. Shekhar, S. Ravada, V. Kumar, D. Chubb, and G. Turner. Declustering and load-balancing methods for parallelizing geographic information systems. *Proc. TKDE*, 10(4):632–655, 1998.
19. A. S. Szalay, P. Z. Kunszt, and A. Thakar et. al.. Designing and mining multi-terabyte astronomy archives: The sloan digital sky survey. *Proc. ACM SIGMOD*, pages 451–462, 2000.
20. D. Thain, J. Basney, S.C. Son, and M. Livny. The Kangaroo approach to data movement on the GRID. *Proc. HPDC*, 2001.