A query optimization strategy for distributed databases on all-optical networks.

S. Bandyopadhyay*, J. Morrissey* and A. Sengupta**

* School of Computer Science, University of Windsor
Windsor, Ont N9E 4L7, CANADA

** Dept of Computer Science, University of South Carolina
Columbia, SC 29208, U S A

Abstract

In the last ten years, optical communication has become a viable technology and is rapidly replacing computer communication based on copper wires. The speed of optical communication is much higher than the speed of electronic devices and we now need to re-examine the cost criteria for data communication in distributed computing. We have shown that the basic assumptions in query optimization for distributed databases are not valid in all-optical networks.In this paper, we have proposed a new strategy for query optimization appropriate for such networks.

1 Introduction

To process a query on a distributed database, it is necessary to access information residing, in general, at a number of different sites in the computer network and to perform a sequence of database operations. The most important parameter in query processing is the total communication cost to process any given query. To minimize the total communication cost there has been considerable research in the areas of semijoin strategies [1], [2] joins [1], [3], [4] and combined strategies [4]. As in most work, we use the SPJ type of query [1] in this paper.

Research on distributed databases started in the late 70's when computer communication was based on copper cables which are slow by present standards and have important differences from current networks based on optical communication. The primary goal of all research in distributed query processing is to reduce the amount of data transferred. This was a reasonable goal in the 70's since the amount of data communicated was the main component of cost to use a network. One major development in fibre-optic communication is the fact that the bandwidth for optical communication is 4 to 5 orders of magnitude faster than the speed of electronic processing.

In this paper, we have examined the cost of computer communication for the emerging technology of all-optical networks. It is expected that this type of networks will dominate in the future[5, 7, 8].

We have analysed the performances of available components for all-optical networks and have concluded that, for moderate volumes of data typical in semi-join based query optimizations, the communication cost in such systems depends primarily on the number of messages rather than the volume of data involved. This is a major change in the way we need to view the parameters to be optimized in distributed query processing.

We feel that, for all-optical networks, our strategy should reduce, to a minimum, the number of messages needed to process the query. In most situations, semi-join strategy cannot be used in all-optical networks since each semi-join requires two additional messages. We propose the following strategy for query optimization :

- Each site participating in a query should receive exactly two messages. The first is a request for characteristics of the relation at the site after selection and projection[1]. The second message has two parts. Part I is a request to carry out a join operation of the relation at the site (after selection and projection on that relation) and all relations sent to that site. Part II of the message is a request to communicate the result to some other site.

- Each site should send exactly two messages - one in response to each of the two messages.

Our strategy is to use join as the only reducing tactic.

0-7803-3143-5 /96/$4.00

Relatively little work[1], [3] has been done in the area of using join for query processing since a semijoin strategy or a combination of joins and semijoins is quite effective in the cost model based on total volume of data.

2 Review

2.1 Optical communication: The design and implementation of light wave networks have gained significant importance in recent years with the advent of optical devices (tunable optical transmitters, receivers, amplifiers, switches etc.) and optical fibers with enormous bandwidth[7]. Wavelength division multiplexing (WDM) technique has been used as an effective tool to achieve a bandwidth of the order of 10 Gb/s and accommodate large number of channels (operating at electronic processing speed) in the bandwidth of optical fibers[7, 8]. Conceptually, each channel might be considered an acceptable wavelength that can be used to transmit information over a fiber.

In an optical communication network, each node, in general, has several transmitters and receivers. Typically, the transmitters and receivers might be either tuned to a fixed wavelength or might be tunable. Each node needs to tune to one of the channels to transmit (or to receive) to (from) the communication medium. One typical arrangement for a local network is a passive star implementation where each node transmits its message to an optical star which, conceptually, is a broadcast mechanism. All communications, received at the star, are broadcast to all the nodes connected to the star. The nodes having receivers tuned to a particular channel can receive only the signals that were transmitted to the star using that channel. Though the physical interconnection is a star connection, any logical interconnection can be achieved by using appropriate channels for the transmitters and the receivers; that is, if the node u has a transmitter using the channel given by the wavelength λ and the receiver of a node v, connected to the same star, is tuned to the same channel, then from a logical standpoint, the star physical topology provides a direct communication from the node u to the node v. In a multihop network, each node can communicate directly with a few nodes[7]. Communication to other nodes must use multiple hops via forwarding through other nodes. This, as pointed in [5], has the disadvantage of some of the "electronic bottlenecks reintroduced". Another approach is the usage of all-optical networks[5], where the messages are translated from electronic to optical domain at the source node and it remains in the optical domain until the destination node is reached, where the message is translated back into electronic domain. The source node chooses a wavelength to communicate with the destination using a path given by the routing algorithm. Without any wavelength conversion, the message must pass through intermediate nodes and links using the same wavelength. Since two different messages cannot pass the same link using the same wavelength, a communication between a source-destination pair using a predefined route must use a wavelength that is not being used in any existing communication sharing a link of the path[5, 6]. Wavelength routing techniques were used in [5] to find a close-to-optimal number of wavelengths for routing in various different topologies.

2.2 Join query graph: A join query graph is defined as **G** = (**V**, **E**) where each node in **V** represents a relation and each edge in **E** corresponds to an equi-join predicate. The nodes are integers i, j, ... , k and store relations R_i, R_j, ... R_k respectively. An edge (i, j) with a label X connecting two nodes i and j (where relations R_i and R_j are stored) indicates that relations R_i and R_j both have an attribute X that participates in an equi-join predicate. We use $R_i(X)$ to denote the projection of relation R_i on attribute set X, and |X| to denote the cardinality of set X. A typical query from[12] is shown below.

Example 1:

select B, D, F, G, H, I
from R_1, R_2, R_3, R_4, R_5

where R_1. A = R_5. A **and** R_1. B = R_2. B **and**

R_2. C = R_5. C **and** R_2. F = R_3. F **and**
R_3. E = R_4. E **and** R_4. D = R_5. D

schema(R_1) = {A, B}, schema(R_2) = {B, C, F}, schema(R_3) = {E, F}, schema(R_4) = {D, E}, schema(R_5) = {A, C, D}. The corresponding query graph and the database profile are shown in Fig 1 and in tables 1 and 2.

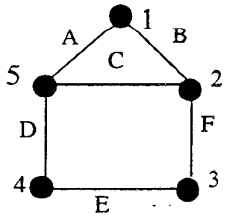

Fig 1 : A query graph

Table 1:

Relation R_i	$\|R_i\|$	Attribute X	$\|R_i(X)\|$
R_1	1190	A B	850 1100
R_2	3440	B C F	900 1000 480
R_3	1180	E F	900 450
R_4	3100	D E	700 800
R_5	2152	A C D	800 1000 720

Table 2:

Attr X	A	B	C	D	E	F
|X|	1000	1200	1200	800	1000	600
Width	2	1	3	1	1	1

We will use R_j => R_i to denote the join of R_i and R_j after R_j has been communicated to the site of R_i.

Theorem 1: Let $G = \{V, E\}$ be a join query graph and $S = \{V_S, E_S\}$ be a connected subquery graph of G. Let $R_1, R_2, \ldots, R_p$ be the relations corresponding to the nodes of V_S and let edgelabels(G) = $\{A_1, A_2, \ldots, A_q\}$. The expected number of tuples in $N_T(S)$ in the relation resulting from joining all the relations in S is given by

$$N_T(S) = \frac{\prod_{i=1}^{p} |R_i|}{\prod_{i=i}^{q} |A_i|^{(m_i - 1)}} \qquad (1)$$

where m_i is the total number of times A_i appears as an attribute in the relations in S and $|A_i|$ represents the number of distinct values of attribute A_i[3].

3. Cost model for all-optical networks

In a typical protocol for setting up the communication between a source-destination pair, the source, using the chosen wavelength, sends a signal using a predetermined path to the destination node until the destination responds[5]. Each node in the path, has its receiver scanning the possible channels to find if any other node wants to establish communication through or to that node. When the receiver at the destination node receives a signal from the source, its receiver gets locked until the communication ends. A communication might get blocked if any of the nodes in the path is already receiving another signal using the same wavelength. The set up protocol might either choose an alternate wavelength to reestablish a communication when the blocking occurs or may back out and try again later with another randomly chosen wavelength.

Typically, a transmitter tuning time is of the order of 1 micro second[8]. In our simplified model, we ignore the fact that, at a high load condition, the blocking probability increases and several wavelengths might have to be tried to establish a communication. Considering the fact that a typical bandwidth is 10 Gb/s using a 4 wavelength multiplexing[9], the set up time is determined by the time for each receiver in the path from the source to the destination to be tuned sequentially to the wavelength used by the transmitter in the source node. We now look at three topologies - 2-grid (or the toroidal bidirectional square lattice), twin shuffle and the de Bruijn graph topology. The data used in this discussion is from [5].

i) For a 2-grid topology with 1024 nodes, the number of wavelengths needed is 512. The maximum length of the path is 32. Thus the tuning time for all the receivers in the path is 32 * 512 = 16 milliseconds.

ii) For a twin shuffle with 1024 nodes, the number of wavelengths is 224 and the maximum length of path is 7. The tuning time is 7 * 224 = 1.5 milliseconds.

iii) For a de Bruijn graph with 1024 nodes, the number of wavelengths is 128 and the maximum length of path is 5. The tuning time is 5 * 128 =.5 milliseconds.

We now calculate the time needed to send a 100 Kbyte file, once the path has been established. The bandwidth/channel is 2.5 Gbits/s and the time to communicate.8 Mbits is.3 milliseconds. Clearly the time to communicate relatively short files (100 Kbytes or less) is dominated by the set-up time.

4. A heuristic for query optimization in all-optical networks

Query optimization problem is known to be NP-complete[1]. Our approach is a greedy heuristic that takes the largest relation R_i which has not been processed yet and reduces it as much as possible using join. In this process, we look for the following two types of subgraphs in the query graph:

- relations lying on cycles that pass through R_i
- relations adjacent to R_i in the query graph

Due to lack of space, we will informally illustrate our approach by optimizing the query given in Fig 1.

Step 1: We sort the relations according to their size, giving us the sorted list R_2, R_5, R_4, R_1, R_3.

Step 2: We determine the cycles that pass through R_2 are $(R_2 - R_1 - R_5 - R_2)$, $(R_2 - R_1 - R_5 - R_4 - R_3 - R_2)$ and $(R_2 - R_3 - R_4 - R_5 - R_2)$.

Step 3: R_2 is connected to R_5, R_3 and R_1. The number of tuples in the join of $R_5 \Rightarrow R_2$ using equation 1 is $\frac{2152 \times 3440}{1200}$. Since this is larger than 3440, the result of the join is larger than R_2. We draw a similar conclusion for R_3 and R_1.

Step 4: The expected number of tuples when we join the relations in cycle 1 obtained in step 2 is $\frac{1190 \times 3440 \times 2152}{1000 \times 1200 \times 1200} \approx 4$. The width of each tuple is 8. The benefit of this operation is $3440 \times 5 - 4 \times 8 = 17168$. The number of tuples for cycles 2 and 3 are 0 and 47 respectively so that the benefits are 17200 and 17159.

Step 5: Our heuristic picks the cycle which has the highest benefit/(number of relations in cycle). Since our most economic operation is the join of relations in cycle 1, we select this as the operation for reducing R_2. (The rationale for this is the fact that if a cycle has smaller number of relations the relation resulting from the join is likely to be an effective reducer for the remaining relations). We note that after the join operation, the query graph is modified to that shown in Figure 2. The "shrinking of edges" due to a join has been discussed in [4]. The relation R_2' has schema {A, B, C, D, F}.

Step 6: The next largest relation is R_4.

Step 7: The number of tuples after the join operation $R_2 \Rightarrow R_4$ is $\frac{4 \times 3100}{800} \approx 16$. The benefit is 6040. It may be verified that this is the most beneficial operation for R_4.

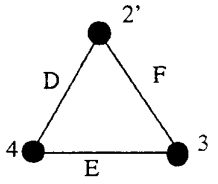

Fig 2: Query graph after first join

Step 8: Similarly we find that $R_{4'} \Rightarrow R_3$ is beneficial.

5. Conclusions

In this paper, we have looked at all-optical networks and have established that decreasing the number of messages is more important than reducing the volume of data. We have described an efficient method of query optimization using the join operation.

6. References

[1] S. Ceri and G. Pelagatti, *Distributed Databases: Principles and Systems*, McGraw-Hill, 1984.

[2] M. T. Ozsu and P. Valduriez, *Principles of distributed database systems*, Prentice Hall, 1991.

[3] P. Legato, G. Paletta and L. Palopli, "Optimization of join strategies in distributed databases", *Information Systems*, Vol 16, No 4, 1991.

[4] M. S. Chen and P. S. Yu, "Interleaving a join sequence with semijoins in distributed query processing", *IEEE Trans. Parallel and Distributed Systems*, Vol 3, No 5, Sept 1992.

[5] .M. Ajmone Marshan, A. Bianco, E. Leonardi and F. Neri, "Topologies for wavelength-routing all-optical networks", *IEEE/ACM Trans. Networking*, vol. 1, pp. 534-546, Oct., 1993.

[6]. K. N. Sivarajan and R. Ramaswami, "Lightwave networks based on DeBruijn graphs", *IEEE/ ACM Trans. Networking*, vol. 2, pp. 70-79, Feb., 1994.

[7] B. Mukherjee, "WDM-based local lightwave networks - part II : Multihop systems", *IEEE Network*, pp 20-32, July 1992.

[8] P. E. Green, *Fiber Optic Networks*, Prentice Hall, 1993.

Acknowledgment:

S. Bandyopadhyay and J. Morrissey acknowledge support from the Natural Science and Engineering Research Council of Canada.