J. King Saud Univ., Vol. 13, *Comp. & Info. Sci.*, pp. 77-103 (A.H. 1421/2001)

Skew-insensitive Parallel Algorithms for Relational Join

Khaled AlSabti* and Sanjay Ranka**
* *Department of Computer Science, College of Computer & Information Sciences*
King Saud University, P.O.Box 51178, Riyadh 11543, Saudi Arabia
** *University of Florida, Florida, USA*

(Received 22 November 1999; accepted for publication 05 September 2000)

Abstract. Join is the most important and expensive operation in relational databases. The parallel join operation is very sensitive to the presence of the data skew. In this paper, we present two new parallel join algorithms for coarse-grained machines, which work optimally in presence of arbitrary amount of data skew. The first algorithm is sort-based and the second is hash-based. Both of these algorithms employ a preprocessing phase (prior to the redistribution phase) to equally partition the work among the processors. These algorithms are shown to be theoretically as well as practically scalable. Experimental results are provided on the IBM SP-2.

1. Introduction

Join is the most important and expensive operation in relation database [1]. Natural join, the most popular form of join, of relation R on attribute x with relation S on attribute y is the set of all tuples t such that t is the concatenation of a tuple r belonging to R and a tuple s belonging to S and $r.x = s.y$. Parallel join has been a widely studied problem in the literature. Most of the parallel join algorithms are based on the uniprocessor join algorithms. The uniprocessor join algorithms can be categorized into three major paradigms: nested-loop, hash-based, and sort-based based [1]. Further, these algorithms can be roughly divided into two groups. One group of the algorithms is *skew-sensitive* where the performance significantly deteriorates with the presence of data skew, while the other group is *skew-insensitive* which alleviates the presence of data skew to some degree. Database research shows that the data skew exists in many real and realistic datasets[2 - 4].

In this paper, we present two new parallel join algorithms, which work optimally in presence of arbitrary amount and any type of skew. The first algorithm is sort-based while the second is hash-based algorithm. Both of these algorithms employ a preprocessing phase (prior to the redistribution phase) to equally partition the work among the processors using *perfect information* of the join attribute distribution. The cost of this preprocessing step is relatively small in case of uniform distribution. Further, it is shown to generate perfect or near-perfect load balancing for datasets with a varying degree of data skew. These algorithms are shown to be theoretically as well as practically scalable. Experimental results are provided on the IBM SP-2. Our algorithms are relatively architecture independent and are designed for memory-resident (in-core) data.

The proposed algorithms have been designed for memory resident-data. In the new generation of coarse-grained machines, the main memory size can be as large as one GBytes/processor. For a 128-processor machine, the aggregate memory available can be as large as a few hundred gigabytes. This can accommodate relations of reasonable sizes in today relational database applications. Further, our algorithms can be easily extended for disk-resident relations.

The rest of this paper is organized as follows. Section 2 describes the parallel machine model and a set of communication primitives. Sections 3 present our notations and assumptions. Section 4 presents the conventional join algorithms. We review some of the proposed parallel join algorithms and discuss some of the important characteristics, which are used in classifying these algorithms in Section 5. Section 6 presents and analyses the two new algorithms. Experimental results are presented in Section 7. Conclusions are presented in Section 8.

2. Coarse-Grained Parallel Machine

Coarse-Grained Machines (CGMs) consist of a set of processors (tens to a few thousand) connected through an interconnection network. The memory is physically distributed across the processors. Interaction between processors is either through message passing or through a shared address space. CGMs have cut-through routed networks which will be the primary thrust of this paper and will be used for modeling the communication cost of the algorithms.

Our analysis will be done for hypercube and two-dimensional meshes networks. The analysis for permutation networks, such as CM-5 and IBM SP Series, and

hypercube is the same in most cases. These cover nearly all commercially available machines. Although the algorithms are analyzed for two types of interconnection networks, they are architecture independent and can be efficiently implemented on other interconnection networks.

Parallelization of applications requires distributing some or all of the data structures among the processors. Each processor needs to access all the non-local data required for its local computation. This generates aggregate or collective communication structures. Several algorithms have been described in the literature for these primitives and are part of standard textbooks [5,6]. We model the cost of sending a message from one node to another as $O(\tau+\mu m)$, where m is the size of the message, τ represents the latency, and μ represents the inverse bandwidth of the communication network.

Table 1 describes the collective communication primitives used in the development of our algorithms and their communication time requirements on cut-through routed hypercube and meshes. In what follows, p refers to the number of processors. A brief description of the primitives is as follows:

1. **All-to-All broadcasting**: In all-to-all broadcast, every node has a message of size m to be sent to all other processors. For more details see [6].
2. **Global combine and prefix scans**: Each processor has a vector of size m. In the global-combine operation, an element-wise sum (or some other operation) is computed on the input vector such that the resultant vector will be stored on all the processors. In the global vector prefix-sum, an element-wise prefix-scan is used instead of the sum. For more details see [6].
3. **Transportation primitive**: It performs many-to-many personalized communication with possibly high variance in message size. Let r be the maximum of outgoing or incoming traffic at any processor. The transportation primitive breaks down the communication into two all-to-all communication phases where all the messages sent by any particular processor have uniform message sizes [7]. If $r \geq p^2$, the running time of this operation is equal to two all-to-all communication operations with a maximum message size of $O(r/p)$. For more details see [7].
4. **Non-order maintaining data movement**: Each processor i has m_i elements. The objective is to redistribute the elements such that each processor will be assigned approximately equal number of elements (m'). Let m be the maximum difference between the m_i and m'. For more details see [8].
5. **Random access write**: Let M be the number of elements distributed across p processors. Each processor is initially assigned approximately $m(=M/p)$ elements. In a Random Access Write (RAW) each of the M elements may need to write data to another element [9]. Each element has, in array P, the index of the element to which it has to send its data. It is possible to have collisions during a RAW. This

happens when two or more data elements are written to the same destination. When collisions occur, one of the following can *n* be done: (i) choose one of the colliding values using a pre-defined rule (ii) combine the colliding data values using a pre-defined binary associative operator. Details of the algorithm for *n* writes on an array of size *n* are given in [10,11].

6. **Merging two sorted lists**: Merging of globally sorted lists has been widely studied problem in the literature. We have chosen a merging algorithm presented in [12]. The size of first list R is N elements and the size of second list *S* is *M*. Each processor has approximately the same number of elements, $n=N/p$ and $m=M/p$, of lists *R* and *S* respectively. The computation time needed by the algorithm is $O(\delta(n+m))$ when *n* and *m* are sufficiently large.
7. **Parallel sort**: In the parallel sort algorithm, each processor initially has $m(=M/p)$ elements. The objective is to globally sort all the elements across all the processors such that each processor will be assigned approximately equal number of elements. There are several well-known algorithms for sorting on coarse-grain parallel machines. We have chosen a parallel sampling-based sort for our problem [6]. The total computation time required is $O(\delta(m \lg m+p^2 \lg p+p \lg m + m \lg p))$. For m ≥ p2, this reduces to $O(\delta (m \lg m+p^2 \lg p))$.

Table 1. Time complexity of communication time of the primitives on different interconnection networks

Primitive	Hypercube	Mesh (wraparound,square)
All-to-all broadcast	$O(\tau \lg p + \mu m(p-1))$	$O(\tau(\sqrt{p}-1)+\mu m(p-1))$
Prefix-sum	$O(\tau \lg p+ \mu m)$	$O(\tau \sqrt{(p}-1)+\mu m)$
Global-combine	$O(\tau \lg p + \mu m)$	$O(\tau(\sqrt{p}-1)+\mu m)$
Transportation	$O(\tau p+ \mu m)$	$O((\tau+ \mu m)\sqrt{p})$
Non-Order maintaining data mov.	$O(\tau p+ \mu m)$	$O((\tau+ \mu m)\sqrt{p})$
RAW	$O(\tau p + \mu m)$	$O((\tau+ \mu m)\sqrt{p})$
Circular q-shift	$O(\tau+ \mu m)$	$O((\tau+ \mu m)(\sqrt{p}+1))$
Merge two lists	$O(\tau p + \mu(m+n))$	$O((\tau+ \mu(m+n))\sqrt{p})$
Sample sort	$O(\tau p +\mu(p \lg^2 p + m))$	$O(\tau \sqrt{p} + \mu(p^{1.5} + m\sqrt{p}))$

3. Notations and Assumptions

Table 2 shows the notations and assumptions that are used for the presentation and the analysis of algorithms described in the next few sections. We assume that each processor has approximately n/p and m/p tuples of relations *R* and *S* respectively. This is not necessarily realistic, especially if some other database operation is performed prior to the join operation. However, this non-uniform tuple distribution can always be handled by the non-order maintaining data movement primitive (see Section 2).

In analyzing the join algorithms, we assume that the local partitions of the two relations at each node are memory resident. Since the total memory capacity in a large parallel system is expected to be high, reasonably large relation partitions can be accommodated in the main memory. Also, no CPU and communication overlap is considered.

Table 2. Notations and assumptions

p - the number of processors in the system
There are two relations R and S where R is the smaller relation
n - the number of tuples in relation R
t_R - the tuple size of relation R
M - the number of tuples in relation S
t_S - the tuple size of relation S
μ - the data transfer rate
τ - the communication start-up overhead
h - the time taken by the hash function
F - the fudge factor
δ - the cost of a unit computation local to a processor
O_R - the largest ratio of the cumulative size of outgoing messages of relations R ($O_R < 1$)
O_S - the largest ratio cumulative size of outgoing messages of relations S ($O_S < 1$)
Q_R - the largest ratio of the relation R which is assigned to some processor to perform the join ($Q_R < 1$)
Q_S - the largest ratio of the relation S which is assigned to some processor to perform the join ($Q_S < 1$)
J - the total join output size produced by all the processors
J_{max} - the maximum join output size produced by some processor

4. Conventional Parallel Join Algorithms

In this section, we briefly present the parallelization of conventional *sort-based* and *hash-based* algorithm. Both of these algorithms are sensitive to the presence of the data skew. The main purpose of this section is to illustrate the communication costs inherent in the join algorithms independent of the data skew. Moreover, we analyze these algorithms under the assumption that no data skew is present, i.e., the amount of required work by the join attribute values has a uniform distribution. Each parallel join algorithm consists of a global and local join method. The global join method refers to the implementation of the join operation across all the processors. The local join method refers to the join method, which is used locally to carry out the join operation between the local fragments of both relations. We restrict ourselves to algorithms in which the global and the local join methods are the same method.

4.1 Sort-based algorithm

We adopt the following version of the sort-based method. This algorithm consists of two phases: sorting and merging phases. In the first phase, both relations ($R + S$) are sorted as one big relation such that any tuple of processor i has a join attribute value less

than the join attribute value of any tuple of processor j, where $i < j$. The sample sort algorithm carries out this phase. As a result of the sorting phase, processor i will receive $O(n/p)$ and $O(m/p)$ tuples of relations R and S, respectively. The received tuples of relations R and S are independently merged to obtain sorted lists for the local fragments of both relations. Each processor then produces the join output by merging its local fragments of both relations. The total time requirement of the sort-based algorithm is given in Table 3. We expect that J_{max} to be very close to J/p with high probability. This algorithm is highly parallel since it achieved a good load balancing (under the uniformity assumption) in two ways. Firstly, the join output is produced almost equally by all the processors. Secondly, the sizes of the local fragments (after sorting) are approximately equal. However, the sort-based method is very sensitive to the data skew. The above algorithm will perform poorly in the presence of the data skew due to the following two reasons. Firstly, J_{max} might be as high as J and secondly the variant of the sizes of the local fragments (after sorting) can be very high.

Table 3. The time total requirements of the sort-based algorithm on different interconnection networks

Network	Complexity of the sort-based algorithm
Hypercube	$O(\tau p + \mu(p \log^2 p + (t_R n + t_S m)/p) + \delta (m/p \lg m/p + n/p \lg n/p + (n+m)/p \lg p + t_R n/p + t_S m/p + (t_R + t_S) J_{max}))$
Mesh	$O(\tau \sqrt{p} + \mu (p^{1.5} + (t_R n + t_S m)/\sqrt{p}) + \delta (m/p \lg m/p + n/p \lg n/p + (n+m)/p \lg p + t_R n/p + t_S m/p) + (t_R + t_S) J_{max}))$

4.2 Hash-based algorithm

There is reasonable consensus that parallel hash-based algorithm is the most efficient algorithm for the join operation in case that the join attribute has a uniform distribution [1]. The hash-based has two phases: the partition and the join phases. In the partition phase, each processor applies a common hash function on the join attribute values for its local fragments of relations R and S and determines the destination processors for the tuples based on predetermined assignment of the hash values into processors number. The expected partition (fragments) sizes of relations R and S are n/p and m/p, respectively.

Each processor may have a set of tuples to send to every other processor. The communication phase can be performed by using the transportation primitive for the two relations with $O(t_R\ n/p\ )$ and $O(t_S\ m/p)$ as the maximum outgoing/incoming message sizes, respectively.

In the join phase, each processor builds a local hash table for its local fragment of one of the relations, i.e. R, using different hash function. Then, each processor probes its local hash table for its local fragment of the other relation, i.e. S. The total time

requirement of the hash-based algorithm is given in Table 4. Like the sort-based algorithm, the hash-based method is very sensitive to the data skew and it is expected to perform poorly in the presence of the data skew. As it is noted in the literature, its performance significantly deteriorates with the presence of the data skew (single or double skew) [13,1]. We will discuss different types of skew in the next section.

Table 4. The total time requirements of the hash-based algorithm on different interconnection networks

Network	Complexity of the hash-based algorithm
Hypercube	$O(\tau p + \mu((t_R n+t_S m)/p) + \delta (t_R n+t_S m)/p + n/p (h+ \delta) + m/p (h+ F\delta + \delta) + \delta (t_R+ t_S) J_{max})$
Mesh	$O(\tau + \mu((t_R n+t_S m)/\sqrt{p})+ \delta (t_R n+t_S m)/p + n/p (h+\delta)+ m/p (h+ F\delta + \delta)+ \delta (t_R +t_S) J_{max})$

5 Join Algorithms with Data Skew

In this section, we describe different types of data skew and review some of the proposed parallel join algorithms and their characteristic. As it is observed by the database research, data skew exists in several real or realistic data sets [2,3,4]. The simple parallelization schemes described in the previous section will have poor performance in the presence of data skew mainly due to resulting load imbalances in the amount of local computation required.

There are four main characteristics which can be used to classify the methods used for achieving load balance in parallelization of join methods [1]. These characteristics are as follows:

1. *Types of data skew:* The data skew may exist in one relation (*single skew*) or in both relations (*double skew*). The data skew and different types of data skew have been defined and modeled in [14]. The data skew types are *tuple placement skew*, *selectivity skew*, *redistribution skew* and *join product skew*. The tuple placement skew occurs when the initial partitions of a relation have different sizes. The selectivity skew results from performing other database operations prior to the join operation. Whereas, the redistribution skew occurs a result of the redistribution phase of the join algorithm, which may generate partitions with high variance in term of sizes. The product skew occurs when the variance of the output sizes produced by each processor is high. The first two types of the data skew can be handled by the non-order maintaining data movement primitive, which is presented in Section 2. In the rest of this paper, we assume that both of the relations are

approximately partitioned among the p processors. We also define work *skew*. This skew combines the redistribution and product skews, which can be measured by the variance of the amount of work performed by each processor.

2. *Load Metrics:* Load metric is the criterion that is used to balance the load across the processors. There are mainly two criterion which have been used as a load metrics:
   - *Cardinality*: This load metric uses the partition/bucket sizes of one or both relations to balance the load across the processors. An approach that can be used for load balancing is to ensure that partition sizes are approximately of the same size. Another approach, which uses the cardinality of the output as a load metric, assigns tuples among processors such that each processor will approximately produce equal number of output tuples.
   - Estimated execution time: The join operation is divided into tasks. The time required to perform each task is estimated using some cost model. The tasks are then assigned to processors such that each processor will finish its tasks in approximately the same time.
   - *Statistical Measures:* Several statistical measures have been used for load balancing:
   - *Bucket-based*: One or both relations are decomposed into buckets. The sizes of the buckets of one or both relations are used in the assignment process.
   - *Class-based*: Join attribute values are organized into *equivalence* classes using some deterministic function. For each class, a set of statistics is maintained, e.g., the number of distinct join attribute values and the number of tuples from both relations.
   - *Perfect information*: This method is an extreme case of class-based method when each class contains only one distinct join attribute value.
3. *Task Allocation*: There are mainly two allocation strategies: *static* and *adaptive*. In the former, a task is statically assigned to one of the processor for the entire computation. The latter allows for immigration of the tasks to other processors during the join process.

Several parallel join algorithms have been proposed to alleviate the presence of the data skew, e.g, [15 - 25]. Most of the proposed algorithms are hash-based algorithms. In the rest of this section, we review a few of the proposed algorithms. *Bucket tuning* was introduced in [15]. In this strategy, the number of the buckets of one of the relations is chosen to be very large. In the later phases, smaller buckets are combined to form large size joins buckets. In Bucket *Spreading* strategy [16], buckets are spread across all processors. These are then reassigned to appropriate processors based on their sizes using a special *Omega* network. A similar algorithm to bucket spreading strategy, which

uses a software control instead of Omega network, has been designed in [17]. A *Partition Tuning* strategy was presented in [17]. This strategy organizes a relation as a set of data cells, and reassigns these data cells from overflow processors to underflow processors using a *best fit* decreasing strategy to balance the load among processors. Three algorithms that use the partition tuning and best fit decreasing strategies have been presented in [26]. Based on their simulation results, they recommended that the *adaptive load balancing parallel hash* (*ABJ*) is the algorithm of choice for most the cases. The bucket tuning, bucket spreading and partition tuning strategies use the cardinality of the partitions/buckets as a load metric. The above algorithms are sensitive to the output skew and expected to perform poorly in the presence of mild or high output skew.

An incremental hash-based algorithm has been proposed and improved in [21] and [22], respectively. In this approach, the join process proceeds in several steps. A *checkpoint* strategy is used after each step to either evaluate the load degree (at the end of the first step) or apply partition tuning using the cardinality of the partially full buckets to change the buckets assignment.

A sampling-based approach has been proposed in [24]. Their approach uses a random sample to estimate the degree of the data skew. Based on this estimation, an appropriate join algorithm is invoked. The invoked algorithm is one of four hash-based algorithms, which have been proposed in [24] along with the conventional hash-based algorithm. Two of these algorithms have been designed to deal with the presence of the redistribution skew and the other two deal with the presence of the product skew. Further, the random sample is, once again, used in the partitioning phase of the join algorithms. Their main assumption is that the skew degree is not very high. Based on their experiments, the *virtual processor range partitioning* (*VPP*) is the algorithm of choice in case of mild skew. Two algorithms have been proposed which use an estimated execution time as a load metric to alleviate the presence of double or single skew [19]. The first algorithm is sort-based and the second algorithm is hash-based. The sort-based algorithm uses a divide-and-conquer approach to address the data skew and a heuristic scheduling phase to balance the load across the processors. The hash-based algorithm uses a two-level hierarchical hashing. The results from the hierarchical hashing are used in a heuristic scheduling phase to balance the load across processors.

A hash-based algorithm (*HISH*), which uses a histogram-based technique to estimate the data distribution and the amount of work, has been proposed in [23]. A cost model has been designed to estimate the amount of work contributed by a join attribute value. The estimated work is then used in the partitioning phase to balance the work among the processors. They also use *virtual processors* approach in assigning the work among the processors. Their histogramming technique, which is based on sampling, produces an approximation of the frequency distributions of both relations and the

output sizes. In most real database systems, the histograms are generally precomputed which makes the preprocessing step of this algorithm is of negligible cost. This algorithm has been compared against VPP algorithm [24], conventional hash-based and *ABJ* algorithm [26]. For mild product skew, the performance of the *ABJ* and *VPP* are comparable. However, *HISH* is superior to all the three algorithms. A PRAM algorithm, which is similar to our new algorithms in spirit, has been proposed in [25]. The proposed algorithm uses the exact total join output size as well as the join output size contributed by each join attribute value to balance the load across the processors.

6. Our Algorithms

In this section, we present and analyze two new parallel algorithms, which deal with arbitrary amount of skew as well as different types of skew. One of these algorithms is a sort-based while the other is a hash-based. Both of these algorithms employ a preprocessing phase (prior to the redistribution phase) to collect *perfect information* of the join attribute distribution.

The main idea of the new algorithms is to compute a weight for each distinct join attribute value. These weights are generated using the perfect information of the join attribute distribution. In the partitioning phase of the join algorithm, p partitions of approximately equal weights are generated. These partitions are then assigned among the processors using static allocation strategy. Further, these sets of weights can be defined in different ways to alleviate different types of skew, i.e. define a weight function for each skew type.

For an in-core parallelization of the join operation, we expect that the product skew can affect the performance of the algorithm more than the redistribution skew. For this reason, we will investigate two weight functions, *output* function (for the output skew) and *work* function (for the work skew). The proposed algorithms have been designed using a set of primitives by which they are relatively architecture independent. Below, we describe the new algorithms.

6.1 The sort-based algorithm

The sort-based algorithm presented in Section 4 is expected to perform poorly with the presence of data skew. The new sort-based algorithm has been designed to alleviate the effect of the presence of the data skew (double or single). The algorithm consists of several phases:

- **Sorting phase**: To sort the local fragments of both relations locally, followed by sorting the join attribute values globally.
- **Preprocessing phase**: To collect the perfect information of the join attribute and generate the set of weights.
- **Splitters phase**: To decide the decomposition strategy and generate a set of splitters.
- **Redistribution phase**: To create the partitions and redistribute them among the processors.
- **Merging phase**: Similar to the conventional sort-based algorithm.

Our sort-based algorithm first sorts relations R and S using parallel sample-based algorithm. In the sample sort, each processor first sorts its local fragments of both relations using a sequential sorting algorithm. However, in the subsequent steps of the sorting phase, the join attribute values are projected and used instead of the whole record. In the preprocessing phase, the perfect information of the join attribute is collected as follows. Each processor scans its local fragments (of the join attribute) of both relations and counts the number of duplicates of each distinct value of the join attribute. The last and the first elements of the local lists might cause an interprocessor communication.

Let $Hist_R$ and $Hist_S$ *be* the results of the counting step for both relations R and S, respectively. Each element of these lists consists of two fields: (1) the value of the join attribute and (2) the number of duplicates. It should be noted that the sizes of $Hist_R$ (n_H) and $Hist_S$ (m_H) are smaller than or equal to n and m, respectively. Merging primitive is performed on $Hist_R$ and *HistS* to obtain a combined list *Hist* of both relations. The merging primitive guarantees that all the elements of the same values are assigned to one processor. In the next step, the set of weights is generated using some weight function. We define two such functions: *work* and *output* weight functions. These functions have been defined to assign a weight to some join attribute value using only the information of that value, i.e., the frequency of the join attribute values in each relation (f_R and f_S). They are defined as follows:

- *Output* weight function F_O: $F_O(f_R, f_S) = f_R \times f_S$
- *Work* weight function F_W:[1] $F_W(f_R, f_S) = f_R \times f_S + f_R + f_S$

Potentially, one can define more complicated functions, which include the (exact or estimated) cost of the next phases of the join algorithm, i.e., the cost of the

[1] This function has been defined in [24]. However, they use it to estimate the cost of join buckets.

interprocessor communication and cost of processing a tuple during the local join method.

One advantage of the above weight functions is that the weights set can be computed locally. Scanning Hist list and applying the weight function for each distinct value can achieve this. Let W be a list of the weights. It can be easily shown that the size of W is less than or equal to $min(n_H, m_H)$. During the previous step, each processor keeps track of its local sum of its weights w_i The total sum of the weights w is computed by performing a global-combine-sum primitive of unit size on w_i

Our assignment technique needs the ranks of the weights. These ranks *Rank* can be computed by performing global exclusive-prefix-sum on W list in two steps. In the first step, the rank of the first weight of each processor ($Rank_0$) is computed by performing global exclusive-prefix-sum of unit size on w_i's. The remaining ranks are computed locally by each processor using sequential prefix-sum on local W list with $Rank_0$ as the starting value.

Our objective is to generate p partitions of approximately equal weights. There are two approaches which have been used for assigning tuples to processors: *full-fragmentation* and *fragmentation-replication*. In the full-fragmentation approach, both relations are partitioned into disjoint fragments; these fragments are then assigned among the processors. The fragmentation-replication approach might partition one or both relations into non-disjoint fragments, i.e., replicates some of the data among more than one fragments. Ideally, one would like to use the full-fragmentation approach because it incurs fewer overheads than the fragmentation-replication approach. However, the full-fragmentation approach is not applicable for some cases. This can happen if the maximum value of W_i's, call it w_{max}, has value greater than $c\ w/p$, for some constant c. We call c a *load factor*. In that case, we switch to the fragmentation-replication approach. w_{max} is found by finding the local w_{max} of each processor followed by performing global-combine on the local w_{max}'s. To decide between the two approaches, we use the *Approach* function:

$$\text{Approach}(w_{max}) = \begin{cases} \textit{full-fragmentation} & : \ w_{max} < c\ os/p \\ \textit{fragmentation-replication} & : \ \text{Otherwise} \end{cases}$$

In case of full-fragmentation approach, the algorithm processed as follows. p-1 splitters are chosen to create p partitions. Each partition is assigned to different

processor. These splitters are selected such that the sum of the weights of each partition is $c\ w/p+\ \epsilon$ and all the join attribute values in partition i are smaller than the values in partition i+1. Each processor locally determines which element of its local *Hist* is a splitter by using its ranks *Rank* and weights W lists . Element i is a splitter j if its rank less than $Rank_0$ of processor j and its rank plus its weight greater than or equal to the $Rank_0$ of processor j. After finding the splitters, many-to-all broadcast is performed (with potentially different message sizes) on the splitters.

In the redistribution phase, the local fragments of relations R and S are partitioned using the splitters list using binary search (the local fragments are already sorted). The required inter-processors communication for both relations is performed using the transportation primitive. The merging phase is exactly similar to the conventional sort-based (Section 4).

In case of fragmentation-replication approach, a more complicated assignment procedure is needed. Each processor locally finds a set of splitters as discussed above. Each splitter might be assigned to multiple adjacent processors. For each splitter i with join attribute value t, we need to determine the replicated relation, first destination, the number of destinations and a set of weights (these are another set of weights), which are used in the redistribution phase, call these w^{dist}. The replicated relation is the relation having smaller number of tuples with join attribute value t. This choice will generally have less communication overhead. The first destination (d^i_j) is computed using the splitter's rank ($Rank^i$) and the number of destinations (n_i) is computed using the splitter weight W^i and $Rank^i$; i.e. $(Rank^i + W^i)\ div\ w/p - d^i_i + 1$.

w^{dist}_{ij} is used by all the processors to determine how many number of tuples having the join attribute value t (the value of splitter i) of the fragmented relation to be sent to processor j. These weights are computed locally as follows.

$$w^{dist}_{ij} = \begin{cases} (((j+1) \times w)/p - Rank^i)/ W^i & : \quad (j \times w)/p < Rank^j \\ w/(p \times W^i) & : Rank^j < (j \times w)/p < Rank^{(i-1)} \\ (Rank^{(i+1)} - (j \times w)/p) / W^i & : Otherwise \end{cases}$$

After finding the splitters, many-to-all broadcast (with potentially different message sizes) is performed on the splitters along with other information. It can be easily shown that the size of all the weights set is at most $2p$.

In the redistribution phase, a destination processor for each tuple is determined as follows. A tuple with join attribute value less than the value of splitter i and greater the value of splitter i-1 is assigned to processor with address equals to d^l_{i-1}. For tuples having a join attribute values equal to the value of splitter i, we perform the following. Assigns those tuples belonging to the replicated relation to all processors with addresses equal to d^l_i, ..., $d^l_i + n_i - 1$. For those tuples belonging to the fragmented relation, computes the number of duplicates of that value, n_{dup}, and assigns $n_{dup} \times w^{dist}_{ij}$ tuples to processor with address equals to $d^l_i + j$. The cost of counting the number of duplicates is linear since the local fragments of both relations are already sorted. The required inter-processors communication and the merging phase are exactly the same as in the full-fragmentation approach.

The overall time requirement of the new sort-based algorithm is the sum of the time required by all the phases. It can be simplified to the time taken by the sorting phase, transportation primitive, merging the R's tuples, merging the S's tuples and the final merging. The computation requirement is:

$$O(\delta(n/p \lg n/p + t_R n/p + m/p \lg m/p + nQ_R \lg p + t_R n Q_R + mQ_S \lg p + t_S mQ_S + (t_R + t_S) J/p)).$$

The communication requirement is given in Table 5 for the two interconnection networks.

Table 5. The communication time requirement for the sort-based algorithm on different interconnection networks

Network	Requirement of the sort-based algorithm
Hypercube	$O(\tau p + \mu(p \lg^2 p + t_R n \max(Q_R, O_R) + t_S m \max(Q_S, O_S)))$
Mesh	$O(\tau \sqrt{p} + \mu \sqrt{p}(p + t_R n \max(Q_R, O_R) + t_S m \max(Q_S, O_S)))$

6.2 The hash-based algorithm

The new hash-based algorithm is very similar to the new sort-based algorithm in the sense that they both collect the same types of information and generate the same set of splitters. However, they mainly differ in the following:

1. Counting the number of duplicates of each distinct join attribute value is done differently. In the hash-based algorithm, a Random Access Write (RAW) primitive with the addition as a collision resolution strategy is used in this process.
2. The local join methods are different. The hash-based algorithm uses a hash-based method as opposed to sort-based method in the sort-based method.

Our hash-based algorithm consists of several phases:

- **Preprocessing phase**: To collect the perfect information of the join attribute using RAW and generate the set of weights.
- **Splitters phase**: To decide the decomposition strategy and generate a set of splitters.
- **Redistribution phase**: To create the partitions and redistribute them among the processors.
- **Join phase**: Similar to the conventional hash-based algorithm.

The distributed memory is *viewed* as a global shared memory with addresses in the range *[0..b]*, where $b=m \times p$ and m is the size of the local available memory of each processor and p is the number of the processors. Location i of this global shared memory resides at processor i *div m* and it corresponds to location *i mod m* of the local memory of that processor. First, the algorithm hashes the join attribute values to integers using some hash function *hash*. The result of *hash* is used as an address of the global shared memory. As necessary requirement of the hash function *hash*, its range should be less than or equal to the size of the global shared memory. We choose to use the RAW algorithm of [10]. This algorithm is very scalable as it shown in Table 1.

Another requirement of the hash function *hash*, to ensure that the number of duplicates computed in the counting step is accurate, is that the hash function *hash* should satisfy the following condition[2]:

- For all join attribute values x and y in both relations, $hash(x) = hash(y) \leftrightarrow x = y$ [25].

We apply the random access write RAW on both relations where the addresses are the hash values and the values are ones. The result of RAW operation is the $Hist_R$ and $Hist_S$ lists of both relations R and S, respectively. For the details about RAW algorithm see Section 2. Computing the set of weights W is straightforward. Since $Hist_R$ and $Hist_S$ have same size, this process does not need any interprocessor communication and it can be done by applying the weight function on each pair of $Hist_R$ and $Hist_S$ lists.

Computing the rank list *Rank*, total sum of the weights set w and finding the splitters phase are exactly similar to the sort-based algorithm. However, the redistribution phase

---

[2] In case that this condition is not satisfied ‹ our hash-based algorithm is still complete (produce all the joinable tuples). Using a function which does not satisfy the condition affects the performance of the algorithm but it does not affect its complicity.

is quite different because the local fragments of the relations R and S are not sorted. In case of full-fragmentation, the assignment of tuples is done by searching the splitters list for each tuple using binary search. While in the fragmentation-replication case, instead of counting the number of duplicates of the tuples having a join attribute value equals to some splitter value as in the sort-based algorithm, we use a weighted round-robin method to assign those tuples belonging to the fragmented relation to destination processors. The required inter-processors communication is carried out by the transportation primitive. The join phase is very similar to the conventional hash-based algorithm.

The overall time requirement of the new hash-based is the sum of the time required by all the phases. The overall time requirement of the hash-based algorithm can be simplified to the time taken by the RAW primitive, transportation primitive, assigning phase, building and probing the hash table and producing the output tuples. The computation time is:

$$O((n+m)/p\ (\delta \lg p + h) + \delta(t_R n/p + t_S m/p\ ) + nQ_R(h+\delta) + \delta mQ_S F + \delta(t_R+t_S)\ J/p + mQ_S h).$$

The communication time is given in Table 6 for the two interconnection networks.

Table 6. The communication time requirement of the hash-based algorithm on different interconnection networks

Network	Requirement of the hash-based algorithm
Hypercube	$O(\tau p + \mu(t_R\ n\ \max(Q_R,O_R) + t_S\ m\ \max(Q_S,O_S)))$
Mesh	$O(\tau \sqrt{p} + \mu \sqrt{p}\ (t_R\ n\ \max(Q_R,O_R) + t_S\ m\ \max(Q_S,O_S)))$

6.3 Scalability considerations

In real database applications, the sizes of the tuples t_R and t_S are generally a few hundreds of bytes. In the two new algorithms, the total cost of the preprocessing step is proportional to the cardinality of the relations times the size of the join attribute. Whereas, the over all costs of the both algorithms are proportional to the cardinality of the relations times the size of the tuples. The size of the join attribute is generally smaller than the size of the tuples by an order to two orders of magnitude. Hence, we expect that the cost of the preprocessing step is relatively small in case of uniform distribution.

7 Experimental Results

We have implemented the four algorithms, namely the conventional hash-based (*SSH*) and sort-based (*SSS*) algorithm, the new hash-based (*SIH*) and the new sort-based

(*SIS*) algorithms, on an IBM SP-2 with 16 processors. The clock speed of the processors is 66.7 MHz, the memory size is 256 MB per processor, and the operating system is AIX version 4.1.4. Our experiments were targeted to study the effect of the weight functions, the load factor, the tuple size, and the size of the relations.

Datasets

We have evaluated the algorithms for dataset generated using three distributions:

1. *Uniform* distribution: The join attribute values has a uniform distribution in [0..256K].
2. *Scalar skew* distribution: This distribution has two parameters (one_R and one_S). Relation R (S) has one_R (one_S) tuples with join attribute of value "1" and the rest of the tuples are generated randomly from *[2..n]* (*[2..m]*) [24]. The default value for one_R and one_S is 1000.
3. *Zipf* distribution: The Zipf distribution has two parameters that determine the degree of the skew of the data [27]. The first parameter z is between zero and one. The dataset corresponds to a uniform distribution when z is set to zero. The level of skew increases as the value of this parameter increases. The second parameter determines the number of distinct values d. For all experiments, d is fixed to 128K. The default value of z is 0.75.

The default values of the sizes of both relations and the tuple size are 256K and 100 bytes, respectively.

For each experiment, the algorithms were executed three times and the median is reported. This is done to alleviate the effect of the randomization of the communication of the underlying network.

In the first experiment, we ran the new algorithms using the two weight functions. Tab. 7 shows the overall execution time in seconds of the new algorithms for different datasets. Clearly, the work function captures the cost more effectively as it results in better load balance. We have set the weight function to the work function for the rest of our experiments.

Table 7. The overall execution time (in seconds) for the two new algorithms using different weight functions for the three distributions on 16 processors

Distribution	Hash output	Based work	Sort output	Based work
Uniform	0.425	0.423	0.794	0.798
Scalar	1.043	0.701	2.150	1.294
Zip-f	6.940	5.936	7.443	6.818

We also ran the new algorithms using different load factors (1, 1.25 and 1.5), the overall performance is almost independent of the load factor for the three datasets. The load factor is fixed to 1 for the rest of the experiments.

Figures1 through 3 show the total execution times for the four algorithms using different tuple sizes (52, 100 and 200 bytes). We can draw the following conclusions from these figures:

1. For uniform distribution the absolute cost of the preprocessing phase is independent of the tuple size. However, its relative cost decreases with the increase of the tuple size. Further, the preprocessing step is relatively small comparing to the overall cost.
2. Our algorithms substantially outperform the conventional algorithms for mild (Scalar Skew) and high (Zip-f) skews. Further, we expect the improvement to be significantly better for large number of processors. This is due the fact that that the conventional algorithms are not scalable, whereas our algorithms are.
3. The new hash-based algorithm is the clear winner for small levels of skew. For high degree of skew, the new sort-based outperforms all the other algorithms. This is due to the fact that the sequential sort-based (in-core version) outperforms the other algorithms for large and highly skewed relations. This can attributed to the high cost of probing the hash table.

The speed-up[3] of the four algorithms on 4,8 and 16 processors for datasets and tuple sizes of 256K and 100 bytes, respectively, are shown in Fig. 5. When the amount of the work required is not high, our algorithms do not achieve any speed-up. For example, the hash-based algorithms do no achieve any speed-up for uniform distribution and small number of processors. This is because that the amount of the required work is about 1.11 seconds. This is comparable to the overhead of these algorithms. However, our algorithms achieved almost similar speed-up as of the conventional algorithms for these cases.

For mild and high skew, the speed-up achieved by our algorithms is significantly better than the conventional algorithms. Further, the new algorithms achieved almost linear speed-up.

[3] The speed-up of the hash-based (sort-based) algorithms are measured against the sequential hash-hased (sort-based) algorithm.

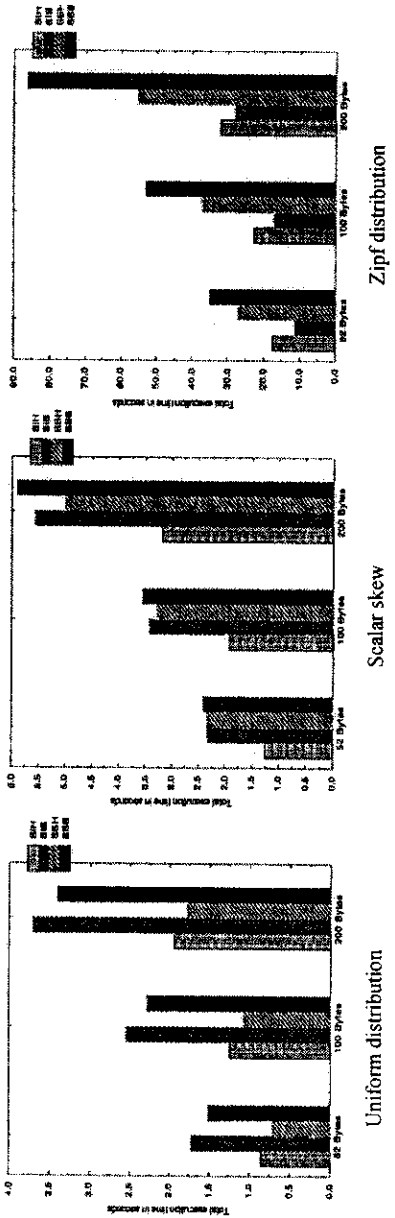

Fig. 1. Comparison of different algorithms for different sizes of tuple on 4 processors.

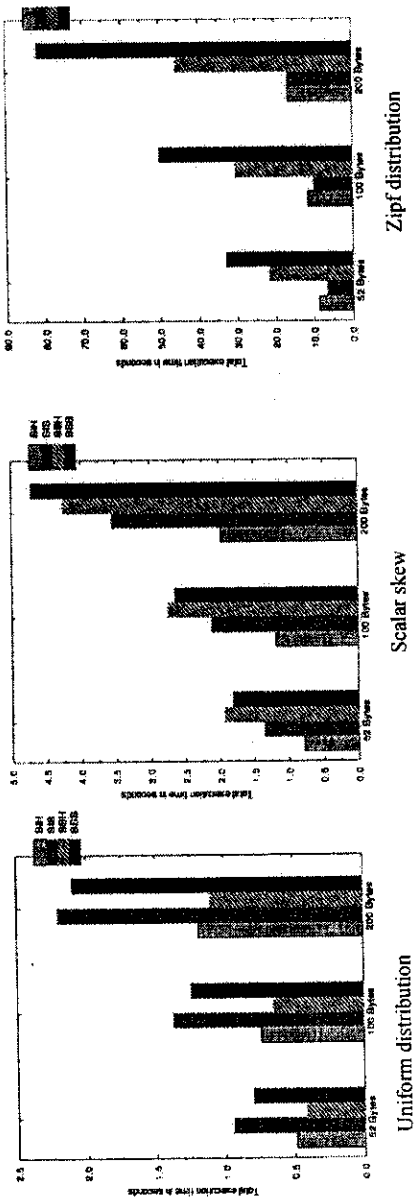

Fig. 2. Comparison of different algorithms for different sizes of tuple on 8 processors.

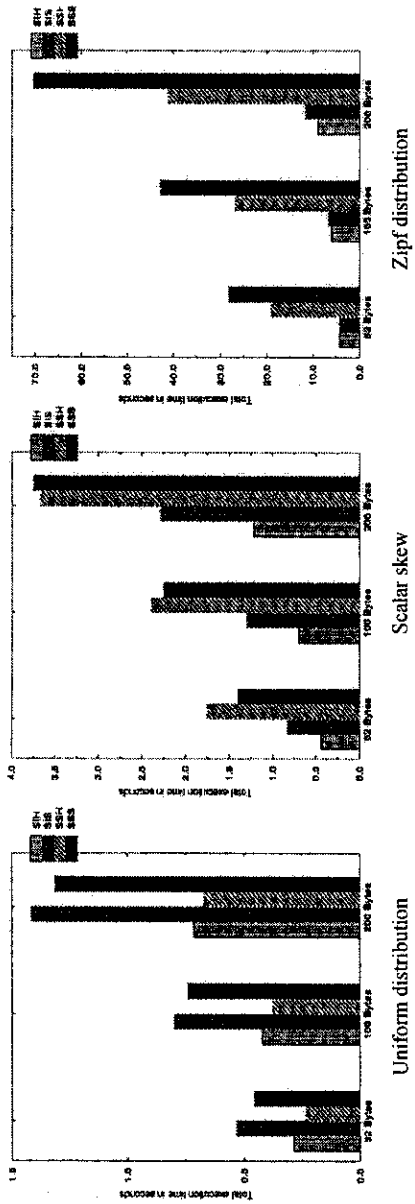

Fig. 3. Comparison of different algorithms for different sizes of tuple on 16 processors.

Fig. 4 shows that our algorithms have excellent size-up properties.

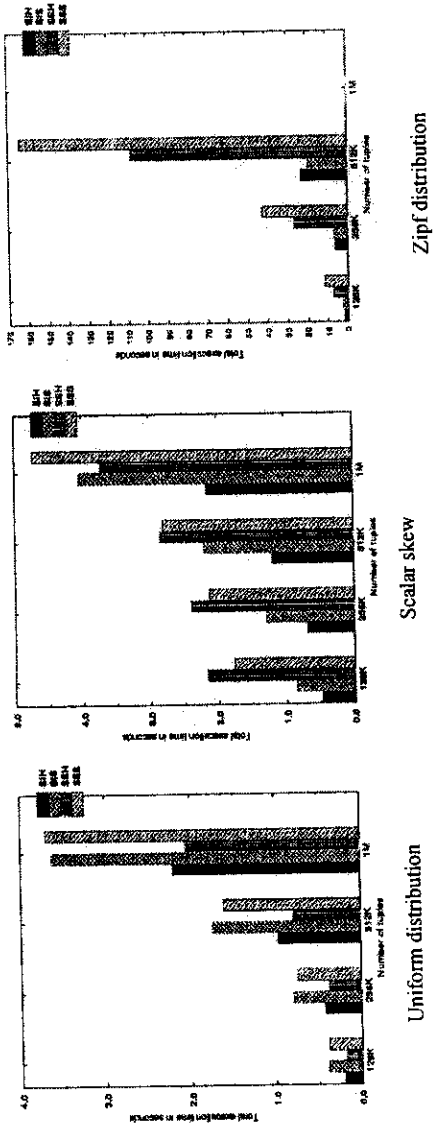

Fig. 4. Comparison of different algorithms for different sizes of relations 16 processors.

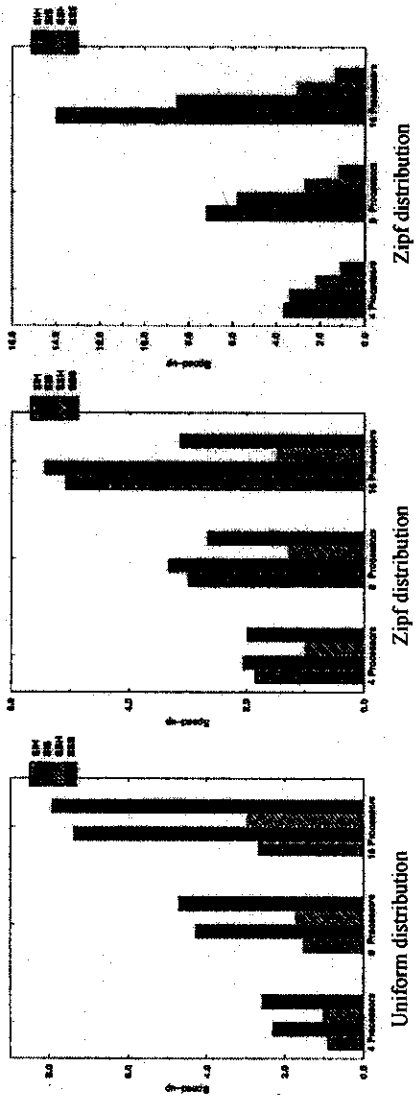

Fig. 5. Speed-up of different algorithms on datasets of size 256K and tuples of size 100 bytes.

Conclusion

We have presented two new parallel join algorithms for coarse-grained machines, which work optimally in presence of arbitrary amount of data skew. The first algorithm is sort-based and the second algorithm is hash-based. Both of these algorithms employ a preprocessing phase (prior to the redistribution phase) to equally partition the work among the processors using perfect information of the join attribute distribution. The cost of this preprocessing phase is relatively small in case of uniform distribution. These algorithms are shown to be theoretically as well as practically scalable. The hash-based algorithm achieved almost perfect speed-up for highly skewed data on different number of processors. It was also better than the other algorithms except for high degree of skew and large relations for which the new sort-based algorithm performs slightly better. Clearly, one can design a hybrid algorithm, which estimates the amount of skew to trigger the appropriate join algorithm.

The proposed algorithms can be easily extended to disk-resident relations. In case that the join attribute values of both relations can be accommodated in the main memory, one can project these values and apply our techniques on them to compute the set of splitters. Then, one can apply the state of the art sequential join algorithm for the local disk-resident fragments in the join/merge phase. This will require one extra sequential I/O (Read) of both relations. As we discussed earlier, the total size of the aggregate main memory across coarse-grained machines can be as large as few hundred gigabytes. Assuming that the tuple size is larger than the join attribute size by a factor of more than 25, this technique can handle relations with sizes proportional to a few terabytes.

Acknowledgment. The work of Sanjay Ranka was supported in part by AFMC and ARPA under F19628-94-C-0057 and WM-82738-K-19 (subcontract from Syracuse University) and in part by ARO under DAAG 55-97-1-0368 and Q000302 (subcontract from NMSU). The content of the information does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

References

[1] Lu, H., Ooi, B. and Tan, K. *Query Processing in Parallel Relational Database Systems.* IEEE Computer Society press, 1994

[2] Christodoulakis, S. "Estimating Record Selectivities". *Information System*, 8,No. 2 (1983), 105--115.

[3] Montgomery, A., D'Souza, D. and Lee, S. "The Cost of Relational Algebraic Operations on Skew Data: Estimates and Experiments". In: *Proc. of Information Process*, 1983.

[4] Lynch, C. "Selectivity Estimation and Query Optimization in Large Databases with Highly Skews Distribution of Columns Values". In: *Proc. of the 14th Int. Conf. Very Large Data Bases*, 1988.

[5] Fox, G. et al. *Solving Problems on Concurrent Processors*: Vol. 1. Englewood Cliffs, NJ: Prentice-Hall, (1988).

[6] Grama, A., Kumar, V., Karypis, G. and Gupta, A. *Introduction to Parallel Computing: Design and Analysis of Algorithms.* The Benjamin/Cummings Publishing Company, Inc, 1994.

[7] Shankar, R., AlSabti, K. and Ranka, S. "Many-to-Many Personalized Communication with Bounded Traffic". In: *Proc. of the Fifth Symposium on the Frontiers of Massively Parallel Computation*, February, 1995.

[8] Al-Furaih, I., Aluru, S., Goil, S. and Ranka, S. "Parallel Construction of Multidimensional Binary Search Trees". *ICS'96, Pennsylvania.*, 1996.

[9] Nassimi, D. and Sahni, S. "Data Broadcasting in SIMD Computers". *IEEE Transactions on Computers*, C-30(2), (1981), 101--107.

[10] Shankar, R. and Ranka, S. "Random Data Accesses on a Coarse-Grained Parallel Machine II. One-to-Many and Many-to-One Mappings". *Journal of Parallel and Distributed Computing*, July 1997.

[11] Bae, S., AlSabti, K., and Ranka, S. "Array Combining Scatter Functions on Coarse-Grained Machines". *CRPC On-line Technical Reports: No. CRPC-TR97732-S.*

[12] Ou, C. and Ranka, S. "Parallel Remapping Algorithm for Adaptive Problems". *Journal of Parallel and Distributed Computing*, 1997.

[13] Schneider, D. and DeWitt, D. "A Performance Evaluation of Four Parallel Join Algorithms in a Shared-Nothing Multiprocessor Environment". In Proc. of the ACM Int'l Conf. Management Data, 1989.

[14] Walton, C. and Dale, A. " Data Skew and the Scalability of Parallel Joins". In: *Proc. of the 3rd IEEE Symposium on Parallel and Distributed Processing*, 1991.

[15] Kitsureqawa, M., Nakayama, M. and Takagi, M. "Query Execution for Large Relations on Functional Disk System". In: *Proc. of the IEEE Fifth Int. Conf. Data Engineering*, 1989.

[16] Kitsureqawa, M. and Ogawa, Y. "Bucket Spreading Parallel Hash: A New, Robust, Parallel Hash Join Methods for Data Skew in the Super Database Computer (SDC)". *Proceedings 16th Int'l Conf. Very Large Data Bases*, 1990.

[17] Hua, K. and Lee, C. "Handling Data Skew in Multiprocessors Database Computers Using Partition Tuning". In: *Proc. of the 17th Int'l Conf. Very Large Data Bases*, 1991.

[18] Omiecinki, E. and Lin, E. "The Adaptive-hash Join Algorithm for a Hypercube Multicomputer". *IEEE Transactions on Parallel and Distributed Systems*, No. 3, May, 1992.

[19] Wolf, J., Dias, D., Yu, P. and Turek, J. "New Algorithms for Parallelizing Relational Database Joins in the Presence of Data Skew". *IEEE Transactions on Knowledge and Data Engineering*, December 1994

[20] Lu, H. and Tan, K. "Dynamic and Load-balanced Task-oriented Database Query Processing in Parallel Systems". In: *Proc. Third Int'l Conf. Extending Data Base Technology.* 1992.

[21] Zaho, X., Johnson, R. and Martin, N. "DBJ- a Dynamic Balancing Hash Join Algorithm in Multiprocessor Database Systems". *Information Systems Journal*, 19(1), 1994.

[22] Hua, K., Tavavapong, W. and Young, H. "A Performance Evaluation of Load Balancing Techniques for Join Operations on Multicomputer Database Systems". *IEEE 11th Int'l Conference on Data Engineering*, 1995.

[23] Poosala, V. *Histogram-Based Estimation Techniques in Database Systems.* Ph.D. Thesis, University of Wisconsin-Madison, 1997.

[24] DeWitt, D., Naughton, J. Schneider, D. and Seshadri, S. "Practical Skew Handling in Parallel Joins". In: *Proc. of the 18th VLDB Conference*, 1992.

[25] Raman, R. and Vishkin, U. "Parallel Algorithms for Database Operations and a Database Operation for Parallel Algorithms". In: *Proc. of the 9th IEEE International Parallel Processing Symposium IPPS*, 1995.

[26] Hua, K., Lee, C. and Hua, C. "Dynamic Load Balancing in Multicomputer Database Systems Using Partition Tuning". *IEEE Transactions on Knowledge and Data Engineering*. 7, No 6, (December 1995).

[27] Zipf, G. *Human Behavior and the Principle of Least Effort*. Reading, MA: Addison-Wesley, 1949.

خوارزميات متوازية غير منحرفة للربط العلائقي

خالد عبدالله السبتي* و سانجاي رانكا**

*قسم علوم الحاسب، كلية علوم الحاسب والمعلومات،
جامعة الملك سعود، ص.ب: ٥١١٧٨ ، الرياض ١١٥٤٣، المملكة العربية السعودية
** جامعة فلوردا، فلوردا، الولايات المتحدة الأمريكية

(قدّم للنشر في ١٩٩٩/١١/٢٢م؛ وقبل للنشر في ٢٠٠٠/٠٩/٠٥م)

ملخص البحث. يعتبر الربط أهم عملية في قواعد البيانات العلائقية ومن أكثرها تكلفة، كما أن الربط المتوازي عملية حساسة لوجود انحرافات بيانية. وفي هذه الورقة نقدم خوارزميتين متوازيتين جديدتين لعملية الربط على الأجهزة المتوازية، فهاتان الخوارزميتان تعملان بشكل مثالي في حالة وجود بيانات منحرفة. حيث إن الخوارزمية الأولى مرتكزة على الترتيب، في حين أن الخوارزمية الثانية مرتكزة على التشتت. وكلتا الخوارزميتان تقسمان العمل على المعالجات في المرحلة التمهيدية. ونوضح جودة هاتين الخوارزميتين نظرياً وعملياً.