DISTRIBUTING A B+ TREE IN A LOOSELY COUPLED ENVIRONMENT

by

G. Matsliach and O. Shmueli

Technical Report #546

March 1989
DISTRIBUTING A B+ TREE IN A LOOSELY COUPLED ENVIRONMENT

Gabriel Matsliach       Oded Shmueli

TECHNION — Haifa, Israel.

abstract

We consider the problem of maintaining a data file which must be distributed among disks, each controlled by a processor and residing in a different site. The pairs, of disks and processors, are connected via a local network. A simple and practical method for maintaining such a file is presented.

Keywords: Distributed data structures, B+-trees, loosely coupled environment.
1 Introduction

There has been an immense amount of work on data structures for secondary memory [3], [8], [10], [15], [16]. This work was mainly done in the context of traditional systems, in which memory and processing power were limited. Modern systems possess an abundance of resources: memory and processors. The latter appear as either loosely coupled or tightly coupled.

The idea of using parallelism for achieving higher performance is not new, for example see work on parallel database machines [2], [5], [6], [14]. In the context of data structures, there has been work on parallel access to shared data structures such as B-trees and Hash tables, for example [7], [12], [13]. However, little work was carried out on distributing the data structures themselves. We consider a special case in which a large file (that does not fit into one disk) must be distributed among several disks. Each disk is controlled by a processor, and each pair of disk and processor, resides in a different site; the sites are connected via a local network.

Many file systems use variants of B-trees, usually B+-trees. Our goal is to design a file system, based on a B+-tree, in a multiprocessor environment with P processors which are located in different sites and connected via a local network. We consider the following standard operations:

- **SEARCH**: Searching for a record given its key.
- **UPDATE**: Updating an existing record.
- **INSERT**: Insertion of a record.
- **DELETE**: Deletion of a record.
- **SUBRANGE**: Producing a list, in key order, of all records in a subrange of keys.

The *Straight Forward Method* for distributing a B+-tree (SFM for short), is to partition its records among the processors. Each processor uses a local B+-tree. (As usual, each tree node spans a disk page). Such a distribution of keys among the processors obeys the following *distribution rule*:

\[ i > j \text{ then the key values of all records of processor } i \text{ are greater than the key values of the records of processor } j. \]

This rule ensures that no merge operation is necessary to produce a result for a SUBRANGE operation. In order to equally partition the records among the processors so that the distribution rule is obeyed, statistics on the distribution of keys is needed. Such statistics may be difficult to obtain in practice. GAMMA, a parallel Database machine [8], consists of 20 processors which are connected via a token ring network. GAMMA distributes the indices among its processors according to the straightforward method described above but without necessarily obeying the distribution rule.

A family of methods for utilizing many small disks, RAID levels 0-5, is reported in [9]. One scheme involves 32 disks, each storing one bit of each 32 bit word. In addition, 8 more disks contain error correcting bits and one additional disk is on standby to replace a failing disk. Each access to the simulated large disk is performed by accessing in parallel all 32 small disks, reading their pages to an array of pages in cache. This is essentially a simulation of a large disk containing large pages by small disks containing small pages. A sequence of improvements on this general idea is presented in [9].

The methods of [9] can obviously be used to distribute a B+-tree among processors. However, in a loosely coupled environment, this results in a communication overhead since each time a tree node should be accessed, all disks are requested to send their parts of the node to the host.
computer, which operates on the node and distributes it back to the disks. We later refer to this method as PDM (Physical Distribution Method). From the above discussion, it seems that PDM is more appropriate for a tightly coupled environment in which special hardware can handle the simultaneous accesses to the disks.

We also use parallelism by distributing the tree among the processors. The tree is called the conceptual tree and its distributed implementation is called the physical tree. The method is called a Totally Distributed B+ tree (TDB for short). TDB is based on the following idea: the entries (keys and pointers) of each node in the conceptual tree (called a conceptual node) are partitioned among the processors. In TDB, each disk is accessed by a different processor. In other words, each processor accesses only a part of each large simulated page and, in order to work properly, does not need to “see” the other parts of the page. Therefore, TDB is more appropriate in a loosely coupled environment in which each processor-disk pair resides in a different site.

We show that the communication overhead of TDB, in terms of bits sent, is reasonable and much smaller than the respective overhead for PDM. Therefore, TDB is more appropriate for a loosely coupled environment. Also, TDB is better than SFM in terms of maximum retrieval time and minimum, average and maximum capacities. A full analysis can be found in [11]. Finally, the possibility of superimposing the algorithms in [12] on TDB is currently being investigated.

2 The physical environment

The physical environment consists of a host computer (host for short), P processors (numbered 0 ... P-1) and P disks. Each disk is controlled by exactly one of the P processors. Each pair, of processor and disk, resides in a different site, and all pairs are connected via a local, Ethernet type, network. The interaction between the host and the processors, and among the processors, is via the local network. In this network, each message sent by a processor, can be “heard” by all processors. Therefore, the cost of a message that is directed to a specific processor (by writing its identification number on the message), and the cost of a message that is sent to all processors, are essentially the same. A diagram of the physical environment is presented in figure 1.

Each processor mainly operates on its own disk. In some cases a processor might need to access another disk. This is done by requesting the desired information from the processor in charge of that disk. When an operation is requested from the host, the host sends it to all the processors, and the processors execute in parallel (with some interactions) in order to produce the results. The processors send the produced result(s) back to the host.

3 The conceptual tree

The B+ trees considered in this paper are defined as follows. In a B+ tree of degree M, each internal node contains no less than M/2 pointers and at most M pointers. The number of keys in an internal node is one less than the number of pointers in it. The leaves contain an equal number of pointers and keys. This number is no less than (M/2)-1 and at most M-1. (Note that the precise leaf content is usually left unspecified in defining a B+ tree. We have fixed a convention to facilitate calculations). The pointers in the leaves point to the actual data records. Thus, the tree is a dense index [15]. All leaves are doubly linked from the leftmost leaf (which contains the lowest key in the
FIGURE 1: A local network.
tree) to the rightmost leaf (which contains the highest key in the tree). This description implies that one can operate on the internal nodes and on the leaves in, essentially, the same way.

The conceptual tree is a B⁺-tree of degree $E$ (as defined above) with the following exceptions; each internal conceptual node contains two additional keys, which are the minimum possible key value and a key larger than the maximum possible key value in the subtree rooted at the conceptual node, and each conceptual leaf contains one additional key which is larger than the maximum possible key of a record which is pointed from this conceptual leaf. (These additional keys may be omitted in practice. We add them so that all parts of the processors in a conceptual node have identical structure). Therefore, each conceptual node contains up to $E+1$ keys and $E$ pointers, and each leaf contains no less than $E/2$ pointers and at most $E$ pointers which point to the actual data records. To simplify the presentation, we assume that $E$ and $P$ are even.

4 TDB

In TDB, the conceptual tree is distributed among the processors by assigning each processor an exclusive part of each node in the conceptual tree. Therefore, each conceptual node is partitioned among the processors holding the physical tree. All processors, together, implement the physical tree. Each part of a node in the tree is called a cut because it can be thought of as obtained from the conceptual node by cutting it at two points. Each processor's cut spans a disk page and therefore a conceptual node spans $P$ disk pages.

All the processors hold their cut of a particular node at the same actual address. Each processor can access only its cut of each node. During key search, this enables a processor, holding a portion of a conceptual node $v$, to provide an address for continuing the search through a child of $v$, even though the cut, through which the search should proceed, is in a different processor. The requirement of same actual physical address can be relaxed. This results in experiencing non-uniform seek time. However, by providing an interrupt mechanism, with which a processor that "knows" how to proceed modifies the others to stop searching their cuts, performance is dominated by the time it takes, for the "right" processor, to determine where to "go down".

Naturally, there will always be exactly one processor which can determine the address of the next node while searching in the physical tree. This implies that a search in the physical tree is done cooperatively by all the processors. In figure 2 we present the general idea: figure 2a presents a conceptual tree of degree 4, figure 2b presents the physical tree which is held by two processors. The root of the conceptual tree is held in the host's main memory, thus, it is not partitioned among the processors.

Reading a node in the conceptual tree means that each processor reads its cut of the node in the physical tree. We would like that all accesses to the disks will be done simultaneously as if only one large disk is accessed. The environment assures us that indeed all accesses can be done almost simultaneously because each processor reads from its dedicated disk.

It is possible that an actual data record is not in the disk of the processor whose cut, of a leaf, contains the pointer that points to that record. This has to do with a decision not to move large data records between processors; see below.

The algorithms which implement TDB (detailed in the next section) operate on the physical tree. Certain features are needed. The following are the important properties of the physical tree:
FIGURE 2a: The conceptual tree (of degree 4).
(a) The root (kept in the host’s main memory).

(b) First cut (processor 0).

(c) Second cut (processor 1).

FIGURE 2b: The physical tree (two processors).
1. The root node of the tree is kept in the host's main memory.

2. Each processor's cut of a conceptual internal node contains the minimum possible key and a key larger than the maximum possible key in its subtree. Therefore, the number of keys in a cut of a node is one more than the number of pointers in that cut. This implies that the leftmost key is the minimum possible key in the conceptual subtree rooted at this cut of the node and the rightmost key is greater than the maximum possible key in the conceptual subtree rooted at this cut of the node. A key and the pointer to its left (in a cut of a node) are called an entry. (This implies that the minimum key in a cut of a node does not belong to any entry.) Each processor's cut of a conceptual leaf contains the minimum key and a key larger than the maximum possible key of a record which is pointed from the cut.

3. Let \( p \) be a pointer in a cut of an internal node held by a processor. Let \( n \) be the key to its left and \( m \) the key to its right. The possible key values in the subtree pointed by \( p \) are such that: \( n \leq \text{value} < m \).

4. Let \( p \) be a pointer in a cut of a leaf held by a processor, and let \( n \) be the key to its left. The key of the record pointed by \( p \) is \( n \). The rightmost key value in a cut of a leaf is larger than the maximum possible key of a record which is pointed from the cut.

5. Invariant: The maximum key in a cut of a node held by processor \( i \) is equal to the minimum key in the cut of processor \( i+1 \) of this node. \( 0 \leq i \leq P-2 \).

6. The left and the right links of a conceptual leaf are kept in the cut of processor \( P-1 \). These are used to implement the doubly connected list of conceptual leaves.

7. The cut addresses of the leaves are marked in order to distinguish between an address of an internal node and an address of a leaf. This is done by adding a special indicator to a leaf address. (Another way to identify leaves is by keeping track of the height of the tree).

8. Each processor holds in its cut of a node up to \( \frac{E}{P} \) entries.

9. Each processor has a counter in each cut, containing the number of entries in the cut.

10. A pointer in a leaf's entry (pointing to a record) also contains the identification of the processor that holds the record in its disk, and, of course, the record's address in that disk.

Point 10 above needs an explanation. In some cases, entries need to be transferred from one processor to another. For example, when trying to insert an entry into a conceptual node which is not full, we may have a problem if the processor that should insert the entry to its cut of the node, has a full cut. In this case we have to "make room", for this new entry by transferring at least one entry out of this processor's cut. This is neither complicated nor expensive when dealing only with a transfer of entries, but it may be an expensive operation when dealing with records since a record may be massive. Therefore, we prefer to avoid moving records pointed by transferred entries. This implies that processor \( i \) will request processor \( j \) to retrieve a record if the entry of the record is in processor \( j \)'s cut and the record was originally inserted into processor \( j \)'s cut or moved, for some reason, to processor \( j \)'s cut.

We use the variant of a \( B^+ \)-tree whose leaves contain entries. In the other variant, the leaves contain the actual data records. TDB can be adapted to this variant as well.
5 Algorithms

We first define three basic procedures: (We assume familiarity with B+-tree operations [8], [15]).

1. Given a key, finding the path to the record containing the key.
2. Inserting an entry into a tree node.
3. Deleting an entry from a tree node.

5.1 Searching

The host finds the appropriate entry in the root and sends the actual address contained in this entry to the processors. From now until finding the leaf, in each level, each processor searches its cut of the node for the appropriate entry (i.e. the leftmost entry in the conceptual node whose key is greater than the searched key): This entry is the only entry in the conceptual node, whose key is greater than the searched key and the key to its left is smaller than or equal to the searched key. (This is why properties 3 and 5 in the previous section are needed). Only one processor finds the entry and sends the address of the node pointed by the entry to the other P−1 processors. Then, all processors "go down" to their cut of the pointed (conceptual) node at the next level. Finally, a processor finds that the entry is pointing to a leaf and informs the others that a leaf has been reached. At the end of this procedure, the processor that holds in its cut the entry which points to the desired record can determine the disk in which the record is kept and its address in that disk.

5.2 Inserting an entry into a node

Each processor checks if the new entry should be inserted to its cut of the node. Only one processor will find the cut to which the new entry should be inserted. Call this processor center. There are three possibilities:

1. There is enough free space in the page holding center's cut of the node. In this case center inserts the new entry to its cut of the node, increments the cut's entry counter and sends the others a stop message informing them that this insertion operation has ended.

2. The page holding center is full but there is free space in the conceptual node. (That is, the conceptual node is not full, as the number of entries held by all processors is less than the degree of the conceptual tree. This is checked for (by center) by sending a count message instructing each processor to send the content of its cut's counter to all other processors). In this case, by the result of the count message, center determines a nearest processor whose page (containing a cut of this node) is not full. Call this processor np. Space for the new entry is freed in the cut of center (of this node) as follows. If np>center (respectively, np< center) then entries are shifted left (respectively, right) from processor i to processor i+1 (respectively i−1), center≤i < np (respectively, np< i ≤ center). Figure 3 exhibits this idea.

3. The conceptual node is full. Therefore, it is split into two conceptual nodes. In the physical tree, this means that each processor should have a cut of each of these two conceptual nodes. Thus, each processor sends its cut's entries to the appropriate processors. (Based on its identification number and the counters of all the processors, each processor can determine to which processors it should send which entries). Hence, each processor's cut of the old conceptual node is now changed. As a result of this node-split, recursively, a new entry is inserted into the parent of the split node.
Assume we have a $B^+$ tree of degree 10. (A) presents a typical node (with two free entries). The same node in the conceptual tree is shown in (B). In the physical tree, $P=5$, each processor can hold in its cut up to two entries; (C) presents a possible distribution of entries to cuts. The entry $(41, L_{new})$ should be entered into the conceptual node. $P_1$ is full. Using a count message, $P_1$ determines that there is still room for a new entry and that the closest processor that can accommodate an entry in its cut is $P_3$. Therefore, a shift of the maximum entry from $P_1$ to $P_2$ and from $P_2$ to $P_3$ is performed resulting in (D). Now $(41, L_{new})$ is entered into $P_1$, as shown in (E).

FIGURE 3
5.3 Deleting an entry from a node

Each processor checks if the entry should be deleted from its cut. Only one finds the entry. Call it center; center checks if the conceptual node contains more than E/2 entries. (This is checked for (by center) by sending a count message. Recall that E is the degree of the conceptual tree).

There are three possibilities:

1. The conceptual node contains more than E/2 entries. In this case center deletes the entry from its cut.

2. The conceptual node contains exactly E/2 entries but the sibling to its left (or its right) contains more than E/2 entries. (This is also checked using a count message). In this case processor P-1 (0) transfers the rightmost (leftmost) entry of its cut of the conceptual sibling, to processor 0 (P-1) that inserts the entry, as the leftmost (rightmost) entry, in its cut of the conceptual node. (Therefore one entry is transferred from the conceptual sibling to the conceptual node. It is possible that a shift of entries among cuts of this conceptual node is necessary to make room for the transferred entry). In the conceptual parent, the key, between the pointer that points to the conceptual node and the pointer that points to the conceptual sibling, is updated accordingly. Now the conceptual node contains (E/2)+1 entries and therefore center can delete the entry from its cut of the conceptual node.

3. The conceptual node contains exactly E/2 entries and so are the sibling to its left and the one to its right. In this case the conceptual node and its sibling to its left (to its right) are concatenated and become one conceptual node that contains, after the deletion, E−1 entries. Therefore, in order to perform the concatenation in the physical tree, each processor sends the entries of its cut of the conceptual node and of the conceptual sibling, to the appropriate processors. (Using its identification number and the counters of all the processors, each processor can determine to which processors it should send which entries). Following this concatenation, the algorithm deletes an entry, recursively, from the parent of the concatenated node.

We note that the count messages, in cases one and two above, can be omitted if each processor keeps, in each of its cuts, a counter that counts the number of entries in the conceptual node.

To complete the description of the algorithms we show how to implement the required operations using these three basic procedures:

- SEARCH: By using the procedure in 5.1 the processor that contains the entry that points to the searched record, in its cut, requests the proper processor (the number of the disk is part of the pointer in the entry) to retrieve the record, and to send it to the host.

- UPDATE: Same as SEARCH but once retrieving the record, the processor updates it and writes it back. (We assume that no modification to a record's key is made by the update operation).

- INSERT: Using the procedure in 5.1 the proper leaf is found. Using the procedure in 5.2, a new entry, that points to the new record, is inserted to that leaf. We omit details concerning the initial placement of the record.

- DELETE: Using the procedure in 5.1 the proper leaf is found. Using the procedure in 5.3, the entry that points to the record is deleted from that leaf and the record is discarded. We omit garbage collection details.
• **SUBRANGE**\( (k_1, k_2) \): Using the procedure in 5.1 (searching for \( k_1 \)), the leftmost leaf that points to a record whose key is greater or equal to \( k_1 \), is found. Each processor sends to the host all the records that are pointed by its cut's entries and whose keys are in the subrange. Then, all the processors, simultaneously, read their cut of the next leaf (to the right of the current leaf) and redo this process. (Recall that processor P-1 contains, in each leaf's cut, the links to the right and the left sibling). When processor P-1 finds, in its cut of a leaf, an entry that points to a record whose key is greater than \( k_2 \), it sends the others a message that this subrange operation ends in this leaf.

The host collects the messages from the processors. Each group of messages contains the records in the subrange that are pointed by a conceptual leaf. The host concatenates the messages (of such a group of messages) in processor number order. In that way the records are produced in key order.

## 6 Performance

SFM distributes the records among the processors so that the distribution rule is obeyed (see section 1). Each processor keeps its records in its own "local" B*-tree. Each such B*-tree is as defined in the first paragraph of section 3, and thus each node spans one disk page. Recall that each conceptual node of TDB spans \( P \) pages.

TDB is better than SFM in terms of maximum retrieval cost and minimum, maximum, and average capacities. These results are intuitive as the degree of the conceptual tree (of TDB) is almost \( P \) times that of the degree of each of the \( P \) trees of SFM, and the exponential effect of the degree (of the tree) on capacity and height is well known. Therefore, our analysis, regarding these results, is omitted and can be found in [11].

We calculate the communication overhead of TDB and PDM in terms of total size (in bits) of messages sent per operation. Recall that the cost of a message being sent to all processors is essentially the same as the cost of a message sent to a specific processor, as all processors are connected via a broadcast local network. Also recall that in PDM, an access to a conceptual node is done by requesting all processors to send their parts of the node to the host. We examine the communication cost when searching, inserting, and deleting a record. The height of the conceptual tree is denoted by \( H \).

- **Searching a record:** In each level of the conceptual tree used in TDB, one pointer is sent to all processors by the processor that finds the appropriate entry in its cut. Thus, the total communication cost of a record search operation is \( H \cdot \text{pointer.size} \). The respective cost for PDM is \( H \cdot \text{block.size} \cdot P \), as each processor sends to the host its part of each node on the search path.

- **Inserting a record:** After reaching the appropriate leaf, an entry should be inserted into that leaf. Thus, each processor sends a message containing its cut's counter, and receives the messages of all other processors. If the conceptual node is not full but the cut of *center* is full (case 2 of 5.2) then, in the worst case, each processor sends one entry via the network. If the conceptual node is full then it must be split (case 3 of 5.2), and thus each processor sends its cut's entries to the other processors, namely, total size of one block. As a result of this node split, a new entry is, recursively, inserted to the conceptual parent.

We conclude that if \( X \) splits have occurred, then the total communication cost, in the worst
case, is (note that $X \leq H$ and is usually much smaller than $H$):

$$H \cdot (\text{pointer.size}) + X \cdot P \cdot (\text{counter.size} + \text{block.size}) + P \cdot (\text{counter.size} + \text{entry.size}).$$

In PDM, in order to split a conceptual node, all processors send their parts of the node to the host, which splits the conceptual node and distributes each of the two new nodes among the processors. Thus the respective cost of PDM is much greater:

$$H \cdot P \cdot \text{block.size} + 2 \cdot X \cdot P \cdot \text{block.size} + 2 \cdot P \cdot \text{block.size}$$

In both calculations, the first term accounts for searching for the leaf, the second for the $X$ splits and the third for the final, non-splitting, insertion.

- Deleting a record: After reaching the appropriate leaf, an entry should be deleted from that leaf. Thus, each processor sends a message containing its cut's counter of the conceptual node and of the conceptual sibling, and receives the messages of all other processors. If the conceptual node contains exactly $E/2$ entries but its conceptual sibling contains more than $E/2$ entries (case 2 of 5.3), then one entry is sent from processor $P-1$ to processor 0 (or from 0 to $P-1$). If the conceptual node and its conceptual sibling contain $E/2$ entries (case 3 of 5.3), then they are being concatenated and thus, each processor sends the entries of its cuts of both nodes to some of the other processors. As both conceptual nodes are half empty, the total size of messages being sent, in order to concatenate the nodes, is $P$ times block size. As a result of this concatenation, an entry is, recursively, deleted from the conceptual parent. We conclude that if $X$ concatenations have occurred, then the total communication cost, in the worst case, is (note that $X \leq H - 1$ and is usually much smaller than $H - 1$):

$$H \cdot (\text{pointer.size}) + X \cdot P \cdot (2 \cdot \text{counter.size} + \text{block.size}) + P \cdot (2 \cdot \text{counter.size}) + (P/2) \cdot \text{entry.size}.$$  

Here, the first term accounts for searching for the leaf, the second for the $X$ concatenations, the third for the non-concatenating final deletion and the fourth for an entry moved between conceptual nodes in the final deletion.

In PDM, in order to concatenate the conceptual nodes, all processors send their parts of the nodes to the host, which concatenates the conceptual nodes and distributes the concatenated node among the processors. Thus the respective cost of PDM is much greater:

$$H \cdot P \cdot \text{block.size} + 2 \cdot X \cdot P \cdot \text{block.size} + 2 \cdot P \cdot \text{block.size}$$

References


