<table>
<thead>
<tr>
<th>Title</th>
<th>A Concurrency Control Method for Parallel Btree Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Tomohiro Yoshihara, Dai Kobayashi, Ryo Taguchi, Haruo Yokota</td>
</tr>
<tr>
<td>Journal/Book name</td>
<td>Proc. of International Special Workshop on Databases For Next Generation Researchers (SWOD 2006), Vol. , No. , pp. 71-76</td>
</tr>
<tr>
<td>Issue date</td>
<td>2006, 4</td>
</tr>
<tr>
<td>DOI</td>
<td><a href="http://doi.ieeecomputersociety.org/10.1109/ICDEW.2006.7">http://doi.ieeecomputersociety.org/10.1109/ICDEW.2006.7</a></td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://www.ieee.org/index.html">http://www.ieee.org/index.html</a></td>
</tr>
<tr>
<td>Copyright</td>
<td>(c)2006 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.</td>
</tr>
<tr>
<td>Note</td>
<td>このファイルは著者（最終）版です。</td>
</tr>
<tr>
<td></td>
<td>このファイルは著者（最終）版です。</td>
</tr>
</tbody>
</table>

*This file is author (final) version.*
OPT and the ARIES/IM\(^1\) on an autonomous disk system [11] using the Fat-Btree as the distributed directory structure and measured the system throughput with changing update ratio. These experimental results indicate that the proposed protocols are effective.

The remainder of the paper is organized as follows. First, the concept of the Fat-Btree and the concurrency controls for parallel Btrees are introduced as background in section 2. We propose the new concurrency control protocols for parallel Btree structures in section 3. The experimental results are reported in section 4. The final section presents the conclusions of this paper.

2 Background

2.1 Fat-Btree Structure

The Fat-Btree [12] is a form of parallel Btree in which the leaf pages of the B*-tree are distributed among PEs and each PE has a subtree of the whole Btree, which contains the root node and intermediate index nodes between the root node and leaf nodes allocated to the PE. Figure 1 shows an example of a Fat-Btree using four PEs.

Although the number of copies increases with proximity to the root node in a Fat-Btree, the update frequency of these nodes is relatively low. On the other hand, leaf nodes have a relatively high update frequency but have no copy. Consequently, the nodes to be updated more frequently have lower overhead for updating with respect to the synchronization between duplicated nodes.

Moreover, in the Fat-Btree, index pages are only necessary for searching for the leaf pages stored in each PE. Therefore, the Fat-Btree can have a high cache hit rate if the index pages are cached in each PE. Because of the high cache hit rate the update processes and the search processes can be processed quickly, compared with a conventional parallel Btree structure.

\(^1\)We implemented ARIES/IM based on [9], but did not recovery mechanism of it because we do not focus on recovery in this paper.

2.2 Concurrency Control Methods

Some concurrency control method for the Btree is necessary to guarantee its consistency. Instead of locks, fast and simple latches are usually used for concurrency control during traversing index nodes in a Btree [5]. A latch is a form of semaphore and the latch manager does not have a deadlock detection mechanism. Therefore, concurrency control for a Btree node should be deadlock free.

In this paper, a latch is assumed to have five modes: IS, IX, S, SIX, and X as shown in Table 1 [5]. The symbol of “∅” means that the two modes are compatible.

Because a parallel Btree structure, including the Fat-Btree, has duplicated nodes a special protocol for the distributed latch manager is required to satisfy the latch semantics. Requested IS and IX mode latches can be processed only on a local PE, whereas the other modes have to be granted on all the PEs storing the duplicated nodes to be latches. That is, the IS and IX modes have much smaller synchronization cost than the S, SIX and X modes, which require communication between the PEs. The S, SIX, and X mode latches on remote copies are acquired by using their pointers. In addition, such latches have to be set in linear order to avoid deadlock. This means synchronization cost grows in proportion to the number of PEs related to latches.

The following conditions of the concurrency controls for parallel Btrees are proposed in [8].

**Condition 1.** A concurrency control method for parallel Btrees should satisfy the following conditions:

(a) No concurrency control protocol method for index nodes, which cause deadlocks, should be used.

(b) Use of S, SIX, and X mode latches on index nodes at upper levels of the Btree should be avoided as much as possible.

(c) The entire tree should not be latched, even for a short duration.

B-OPT [1], OPT-DLOCK [10], and ARIES/IM [9] are excellent concurrency control methods for a Btree on a single machine. However, they do not satisfy Condition 1: B-OPT does not satisfy Condition 1-(b), OPT-DLOCK does not satisfy Condition 1-(a), and ARIES/IM does not satisfy...
3. It never uses a tree latch.

It is easy to prove that the MARK-OPT satisfies the physical consistency requirement for Btrees. When an up-dater realizes that it does not acquire all required X latches for the SMOs, the up-dater releases all the latches without modifying any data. Thus, the MARK-OPT essentially follows the two phase locking (2PL). The 2PL ensures the physical consistency of the Btree structure for each up-date [2].

3.2 Extensions of the MARK-OPT

We propose three variations of the concurrency control protocol, INC-MARK-OPT, 2P-INT-MARK-OPT and 2P-REP-MARK-OPT, which are extensions of MARK-OPT.

In each protocol only the second phase of the update protocol is different. MARK-OPT does not change the process even if the tree structure is changed by other transactions. On the other hand, the extension protocols look at the state of the node first latched with the X mode in that phase and checks the change from the previous phase of a subtree relating to SMO. If the extension protocols judge that the tree structure has been changed, each protocol executes a different process. The difference in the processes is shown in Table 2. The columns indicate the process phases, because the protocols judges if the tree structure has changed, and the rows indicate the presence of a restart at that time.

Because these extension protocols mark the height of the tree as does MARK-OPT, they execute very similar process as MARK-OPT except for phase change. All of these also satisfy Condition 1.

3.2.1 INC-MARK-OPT Protocol

The incremental marking optimistic (INC-MARK-OPT) protocol restarts when it judges that the tree structure has changed in the second phase. In this case, the height marked for next phase is not complete because traverse does not reach a leaf node. However, the INC-MARK-OPT decides the range of the X latch based on that information.

The second phase of the INC-MARK-OPT protocol for update is as follows. The INC-MARK-OPT acquires the X mode latches on the node marked in the previous phase. If the node is not full, it executes a process similar to the MARK-OPT to the leaf node. If the node is full, it releases all latches and restarts. This process continues until all the nodes involved by SMOs are protected by X latches.

If the INC-MARK-OPT judges that the tree structure has been changed, it restarts at once. Therefore, the INC-MARK-OPT does not acquire more needless X latches than MARK-OPT when SMOs have actually spread. On the other hand, the INC-MARK-OPT may judge that the tree structure has changed when SMOs have contracted. In that case, the INC-MARK-OPT acquires more needless X latches than MARK-OPT. Moreover, INC-MARK-OPT increases the frequency of restarts compared with MARK-OPT. However, the number of maximum phases in the INC-MARK-OPT is $H$ at most the same as MARK-OPT because it spreads the range of the X latch at each restart.

3.2.2 2P-INT-MARK-OPT Protocol

The 2-phase integrated marking optimistic (2P-INT-MARK-OPT) protocol performs the latch-coupling with IX latches below the node when it judges that the tree structure has changed in the second phase. That is, it returns to the first phase. Because the MARK-OPT marks the tree state in the

---

**Table 2. Comparison of protocols by handling for a tree structure change between phases**

<table>
<thead>
<tr>
<th></th>
<th>continue</th>
<th>shift to</th>
</tr>
</thead>
<tbody>
<tr>
<td>the 2nd phase</td>
<td></td>
<td>the 1st phase</td>
</tr>
<tr>
<td>non-restart</td>
<td>MARK-OPT</td>
<td>2P-INT-MARK-OPT</td>
</tr>
<tr>
<td>restart</td>
<td>/ INC-OPT</td>
<td>2P-REP-MARK-OPT</td>
</tr>
</tbody>
</table>

**Figure 2. The MARK-OPT protocol**
second phase, it shifts to the first phase without the problem of marking that is incomplete.

The second phase of the 2P-INT-MARK-OPT protocol for update is as follows. The 2P-INT-MARK-OPT acquires the X mode latches on the node marked in the previous phase. If the node is not full, it executes a process similar to MARK-OPT to the leaf node. If the node is full, it shifts to the first phase on the node.

If the 2P-INT-MARK-OPT judges that the tree structure has changed, it shifts to the first phase. Therefore, the 2P-INT-MARK-OPT decreases the frequency of restarts to a greater extent than the INC-MARK-OPT and it does not acquire more needless X latches than MARK-OPT. Moreover, 2P-INT-MARK-OPT does not acquire needless X latches when SMOs have contracted. On the other hand, the 2P-INT-MARK-OPT may require more restarts than H, the height of the tree. However, this will happen infrequently. In the worst case, the 2P-INT-MARK-OPT cannot complete the process but the 2P-INT-MARK-OPT requires only a small number of restarts because this will not happen in practice.

3.2.3 2P-REP-MARK-OPT Protocol

The 2-phase repetitive marking optimistic (2P-REP-MARK-OPT) protocol restarts when it judges that the tree structure has changed in the second phase. The 2P-REP-MARK-OPT returns to the first phase after it restarts.

The second phase of the 2P-REP-MARK-OPT protocol for update is as follows. The 2P-REP-MARK-OPT acquires the X mode latch on the node marked in the previous phase. If the node is not full, it executes a process similar to MARK-OPT to the leaf node. If the node is full, it releases all latches and it executes the process in the first phase on the root node.

If the 2P-REP-MARK-OPT judges that the tree structure has changed, it restarts at once and shifts to the first phase. Therefore, the 2P-REP-MARK-OPT acquires the least needless X latches of the proposed protocols although it increases the frequency of restarts more than any of the proposed protocols. As well as 2P-INT-MARK-OPT, the 2P-REP-MARK-OPT may require more restarts than H, the height of the tree. However, the 2P-REP-MARK-OPT requires only a small number of restarts because this is unlikely to happen in practice.

4 Experiments

To show that MARK-OPT and its extension protocols are effective, we implemented them on an autonomous disk system [11] on a blade system, which uses the Fat-Btree, and evaluated their performance under a number of conditions.

<table>
<thead>
<tr>
<th>Table 3. Experimental Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Nodes: 128 (Storages), 16 (Clients)</td>
</tr>
<tr>
<td>CPU: AMD Athlon XP-M 1800+ (1.5GHz)</td>
</tr>
<tr>
<td>Memory: PC2100 DDR SDRAM 1GB</td>
</tr>
<tr>
<td>Hard Drive: TOSHIBA MK3019GAX (30GB, 5400rpm, 2.5inch)</td>
</tr>
<tr>
<td>OS: Linux 2.4.20</td>
</tr>
<tr>
<td>Java VM: Sun J2SE SDK 1.5.0_01 Server VM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4. Parameters used for the experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page size: 4KB</td>
</tr>
<tr>
<td>Tuple size: 3KB</td>
</tr>
<tr>
<td>Max No. of entries in an index node (fanout): 8</td>
</tr>
<tr>
<td>Max No. of tuples in a leaf node: 1</td>
</tr>
</tbody>
</table>

4.1 Experimental Environment

We used an experimental system of the autonomous disk distributed storage technology we proposed. The experimental system was implemented on a 144 node blade system using the Java programming language on Linux. We used 128 nodes for storing data and 16 nodes as clients sending requests. A preliminary experiment showed that the backbone network switch had adequate performance. The experimental environment is summarized in Table 3. Table 4 shows the basic parameters we set for the experiments. These parameters were chosen to distinguish clearly the differences between the protocols.

4.2 Experimental Result

Sixteen clients (thirty-two threads in parallel per blade) sent requests to PEs containing the Fat-Btree with 256 tuples per PE, for 60 seconds. The access frequencies were uniform. Figure 3 shows the performance of the six concurrency controls as the update ratio changes from 0% through 100%. The solid lines show the performance of the four proposed protocols, the dotted line shows the performance of INC-OPT and the dashed line shows the performance of ARIES/IM. The horizontal and vertical axes are the update ratio and throughput, respectively.

When the update ratio was 0%, the results of all protocols were virtually the same. This is because the concurrency controls used to retrieve data are basically the same. But, the throughput of ARIES/IM decreases sharply even though the increase in the ratio of update operations is small. This is because the cost of global synchronization by using the tree latches for SMOs caused by the update operations. On the other hand, when the update ratio is low,
the results of all other protocols were better than that of ARIES/IM and almost the same. However, the throughput of the INC-OPT decreases as the update ratio increases. In contrast to the INC-OPT, the proposed protocols can provide reasonable throughput although the update ratio increases. The decline in the throughput of the proposed protocols is much slower than that of INC-OPT even when the update operations are included. This is because the proposed protocols reduce the frequency of restarts compared with the INC-OPT when SMOs occur; although the increase in the update ratio increases the occurrence of SMOs.

In the comparison of the proposed protocols, the throughput of 2P-REP-MARK-OPT is the highest, when the update ratio is 100%. Then, the throughput of 2P-INT-MARK-OPT is next highest and then MARK-OPT is higher than INC-MARK-OPT. The throughput of the INC-MARK-OPT is the lowest. Therefore, the 2P-REP-MARK-OPT is the most effective when many SMOs occur.

5 Conclusion

We propose a new concurrency control, MARK-OPT, for parallel Btrees, for the shared-nothing environment. When SMOs occur, the proposed protocol marks the node for which the X-latch should be acquired first and it acquires the X latch nodes below the marked height after it restarts. We also propose three extensions of the MARK-OPT protocol. In addition, we have experimented on an autonomous disk system implemented on a large-scale blade system to compare the four proposed protocols and the conventional protocols. The experimental results indicated that the proposed protocols were effective and 2P-REP-MARK-OPT was the superior protocol.

In future studies we plan to apply the B-link [7, 6] to the Fat-Btree. It is known that the B-link can achieve excellent concurrency control. The B-link uses links to chain all nodes at each level together. In the B-link algorithm, neither readers nor updaters latch-couple on their way down to a leaf node and they acquire the latch only on one node at a time. Moreover, the B-link algorithm does not require restarts and they complete processing during one traverse. We need to examine how to apply the B-link to the Fat-Btree and compare this to the conventional protocols.

Acknowledgments

We thank Dr. Jun Miyazaki of NAIST for his advice on concurrency control for the Fat-Btree. This work is partially supported by CREST of JST (Japan Science and Technology Agency), SRC (Storage Research Consortium), a Grant-in-Aid for Scientific Research from MEXT Japan (#16016232) and the Tokyo Institute of Technology 21COE Program “Framework for Systematization and Application of Large-Scale Knowledge Resources”.

References