1R－5

永続分散共有メモリ機能を提供するデータベースサーバ「わかし」のメモリコヒーレンス機構*

Taiyong JIN, Kunihiko KANEKO, Akifumi MAKINOUCHI †

1 Introduction

Distributed Shared Memory (DSM) is defined to be a shared memory area over network, Network Of Workstation(NOW) can act as a parallel machine using DSM. The DSM space in every workstations of NOW always should be consistent .This consistency is maintainced by the continous communication via network.The cost of communication should be decreased,in order to improve the performance of DSM.But with the contrast to the improvement of CPU, and hard disk, the speed of network became the bottle neck of DSM. In this Paper, we will introduce a new mechanism to maintain the memory coherence of WAKASHI, which is a prototype of Persistent DSVM. That mechanism can also decrease the cost of communication in some degree. First, the overview of the memory coherence mechanism of DSVM is introduced in section 2, and WAKSHI is described in section 3 briefly . Next we will introduce the new memory coherence mechanism in WAKASHI in section 4. Finally,the conclusion is given in section 5.

2 Overview of the memory coherence mechanism of DSM

Any processor can access any memory location in DSM space directly.A DSM space is coherent if the value returned by a read operation is always same as the value written by the most recent write operation to the same address[1].The memory coherence of DSM is maintained by the communication between these duplicated imageg in DSM.The speed of network is often a bottleneck of DSM,while the number of these duplications is large. So to decrease the cost of the communication is the most important for improving of DSM.

In Weaken Consistency(WC) models[2,3,4], their performance can be improved a great deal by having users to take part in keeping memory coherence in some degree, such as the explicit locking command, barrier's setting, and so on.We can say that this way may be not a good way to have users share the load that should be imposed on system, and there are a great deal of complexity in the programing on these system.

3 Architecture of WAKASHI

Wakashi is a Distributed Virtual Shared Memory (DSVM)System for Object Orient Database , which is based on NOW enviorment. And there are 2 main functional blocks in each wakashi server process, which are Storage Management,and Transaction Management.

3.1 Storage Management

A DSVM space of Wakashi is a group of heaps . A heap comprises a sequence of bytes,which is mapped from a disk file.A Wakashi heap can be distributed into all of the workstations via local mapping or remote mapping. If a heap in a workstation is created by local mapping , the server process in this worksation is called as the "primary server" of this heap, otherwise it is called as the "mirror server". And the primary server can manage the remote access in reasonable order.

The granularity of heap is a page, whose size is decided by the setting of Operating System, usually as 4K. In order to synchronize the access of each page,We defined two kinds of page locks which are write-lock(WL) and read-lock(RL).The read lock is shared and the write lock is exclusive.

3.2 Transaction Management

As Wakashi is designed for Object Orient Database System,each user of any lock is decided only as a transaction. And we employed 2PL to maintain the serializability, when the transactions run in parallel.

*Memory coherence algorithm for a persistent distributed shared database server 「WAKSHI」

†Graduate School of Information Science and Electrical Engineering,Department of Intelligent Systems, Kyushu University 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-81, JAPAN

4 Memory Coherence Mechanism of WAKSHI

WAKSHI uses Wirte Invalidate(WI) coherence protocol as common protocol.For example, there is a page read-lock-request in remote wakashi server, if the page is valid , then the request is permitted right now. If the page is invalid then the request is forwarded to primary server, the value of this remote page will be refreshed in case that the lock-request can be permitted.In the case of write-lock-request, firstly the lock-request will be forward to primary server. And if there are other copy(s) of this page distributed into the other servers, , an invalidation message should be send to every copy of this page when the lock-request can be permitted. And a page remain valid until a invalidation comes. Totally, the efficient of read is local , and the efficient of update is global in common case of wakashi. If the percentage of update access in DSM application is low and the number of machines is not too large, the performance of write invalidate coherence is better. Or the performance may decrease . So we can say that the presupposition of this protocol is that the percentage of update is low and the number of valid copys of the page is not too large.

A new protocol named "Even Cost protocol"(EC) is introduced as an assist way of "write invalidate protocol".It means that the cost of read and update access should be even. For example, there is a page's reading request in remote wakashi server, although the page is valid , then the request is also forwarded to primary server. If needed, the value of this remote page will be refreshed .In the case of a page write request, the lock-request will be forward to primary server. Although there are other copy(s) of this page distributed into the other servers, , invalidation message need not to be send to every copy of this page. This protocol also has a presupposition . It is that the percentage of update is so high that the extental cost of update is too large.

The two protocol above have advantage and shortcoming respectively.We found that it is good way to hybrid them dynamicly. In other words, in common case the write invalidation protocol can be selected as fist choice.but if the cost of update becomes larger ,then the Even Cost Protocol should be better choice. This protocol , we call it "Hybrid Balance Cost Protocol"(HBC). This name means that this protocol is to keep the balance cost via hybrid of "write invalidate protocol" and "Even Cost Protocol".

We select EC for a page when the page became a Update-Hot-Spot. Update-Hot-Spot is defined as a page that is updated frequently during a period of time. The period of time is named as the "Hot-Last-Period".

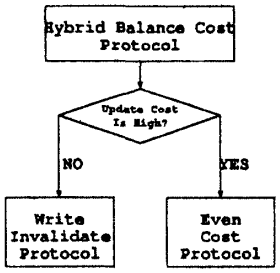

図 1: Structure of Hybrid Balance Cost Protocol

5 Conclusion

Memory Coherence of DSM is the most import and diffcult problem. WAKASHI is a DSM prototype system. WAKSHI employs two kinds of implicit page-based locks(write-lock and read-lock) to synchronize the accss on the DSM of WAKASHI.For the process of lock or unlock is transparent to user's level, it is simpler for user to programing on WAKSHI than on the relaxed consistency system.In order to decrease the cost of communication , we firstly select the "write invalidate protocol "as basic protocol when the percentage of write is low in common case. If the percentage become high , WAKSHI can switch "Even Cost Protocol" as the substitute of the "write-invalidate protocol".This hybrid protocol is named as "Hybrid Balance Cost Protocol".Following is the comparison table of WC,WI, and HBC protocols.

	Consistent Safety	Performance
WC	low	best
WI	high	good
HBC	high	better

参考文献

[1] K.Li and P.Hudak. Memory coherence in shared virtual memory systems. ACM Transactions on Computer Systems,7(4):321-359,November 1989.

[2] B.N.Bershad and M.J.Zekauskas. Midway: Shared memory parallel programming causal distributed shared memory. In Proceedings of the 11th International Conference on Distributed Computing Systems ,Pages 152-164, Octorber 1991.

[3] J.B.Carter,J.K.Bennett,and Willy Zwaenepoel. Implementation and Performance of Munin. In Proceedings of the Thrteenth Symposium on Operating Systems Principles,pages 152-164,Octorber 1991.

[4] P.Keleher,A.L.Cox,and W.Zwaenepoel. Lazy Consistency of Software DIstributed Shared Memory. In Proceedings of the 19th Annual Symposium on Computer Architecture. pages 13-21,May 1992