論文 / 著書情報
Article / Book Information

題目(和文)	並列分散システムにおける分散B-tree 並行性制御の最適化に関する研究
Title(English)	A Study on Optimization of Concurrency Control for Distributed B-trees on Parallel and Distributed Systems
著者(和文)	吉原朋宏
Author(English)	Tomohiro Yoshihara
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10875号, 授与年月日:2018年3月26日, 学位の種別:課程博士, 審査員:横田 治夫,宮崎 純,権藤 克彦,DEFAGO XAVIER,金子 晴彦
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10875号, Conferred date:2018/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	計算工学 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of
学生氏名： Student's Name	吉原朋宏	指導教員（主）： Academic Supervisor(main)	横田治夫
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Recent IT systems are composed of many core CPUs, many CPU sockets, storages, main memories, and networks. Increasing the performance of mainstream multicore processors is becoming difficult, microprocessor vendors are developing “manycore” CPUs with over 60 cores on a single-chip. Although Ethernet as networks among servers with CPUs is mainstream, the performance and cost for Ethernet are not improved easily. Prevailing fast SSDs and large main memories resolves bottleneck of storages, and it is very important for deriving the high performance to consider networks and many cores.

Various applications such as RDBMSs, key-value stores, graph databases, and file systems are processed on IT systems. Indexes are important for improving the performance of their applications. Their applications use parallel and distributed processing using distributed indexes to improve the performance. Distributed storage systems and parallel database systems are key components in their applications, their indexes are mainly distributed B-trees. Since its concurrency control greatly affects the system performance, it is important for their optimization to consider the characteristics of many cores and networks.

Concurrency control for distributed B-trees is one of broad parallel and distributed processing. An important issue for improving the performance of parallel and distributed processing is an optimization of “communications” and “skew handling” among parallel processing. Parallel and distributed processing use mainly optimistic approaches for improving the performance, an important issue is an optimization of “retry processing” due to optimistic processing. Moreover, many applications like HDFS manage distributed redundant data not only for high performance but also for reliability and availability. An optimization of “access to redundant data” is also an important issue.

This Ph.D. dissertation challenges four above issues about concurrency control for distributed B-trees. This thesis proposes four novel optimization of concurrency control techniques for distributed B-trees focusing on the characteristics of many cores and networks in parallel and distributed systems.

At first, this thesis proposes MARK-OPT optimizing retry with optimistic methods, and MARK-OPT is capable of reducing the cost of structure-modification-operation (SMO) compared to the conventional protocols. Since it marks the lowest SMO occurrence point during optimistic latch-coupling operations. The marking reduces middle phases for spreading an X latch and removes needless X latches. In addition, this thesis proposes three extensions of the MARK-OPT, which focus on tree structure changes by other transactions. This thesis implements and evaluates proposed protocols on distributed B-trees. Experimental results in various environments indicate that the proposed protocols always improve system throughput, and the most optimistic extension of the MARK-OPT is the most useful protocol in high update environment.

Next, this thesis proposes LCFB optimizing latch-coupling suitable for communications, and LCFB requires no latch-coupling in optimistic processes. LCFB reduces the amount of communications between servers during B-tree traversal. To detect access path errors in LCFB caused by removal of latch-coupling, LCFB assigns boundary values and delete bit on each index page. This thesis proposes LCFB-link with the B-link algorithm within each server to reduce the cost of SMOs in a server, as a solution to the difficulty of consistency management for the side pointers in distributed B-trees. This thesis implements and evaluates proposed protocols on distributed B-trees. Experimental results in various environments indicate that the system throughput of the proposed protocols is always superior to those of the other protocols, especially in large-scale configurations, and LCFB-link is effective at higher update ratios.

Thirdly, this thesis proposes SARP optimizing of redirections suitable for skews of workloads. SARP is suited to manycore processor platforms. SARP address the problem that it is difficult for the conventional partitioning approaches on manycore processors to provide consistent performance for unpredictable workloads such as OLTP. SARP selects a strategy for distributing the workload to many cores, using a distribution that is appropriate to the characteristics of each processor. This treats effectively any large skews in the workload and minimizes the redirection communication among the cores. SARP-abc is an extension of SARP which introduces three techniques for reducing the more frequent latch collisions from large-scale workload skew when using manycore processors. To improve performance stability, this thesis also proposes SARP-ab'c. Because the latch collisions on a B-tree subroot depend on the height of the B-tree, SARP-ab'c uses this height to control optimally the number of subroot index pages. To evaluate our proposed methods and the conventional method which is known as one of the fastest approach on multicore processors, this thesis performs experiments with a manycore processor. Experimental results demonstrate that SARP-abc can achieve a throughput of 44 times that for the conventional method in the maximum-skew case and can maintain the throughput at 66% of a throughput for uniform workloads, the fluctuation in throughput for SARP-ab'c was least during gradual increases in workload skew. This indicates that the proposed methods can achieve expected distributing workloads to appropriate cores for processors.

Finally, this thesis proposes RA-Btree optimizing of access methods to replicas, and RA-Btree treats instantly the temporary skews of workloads in short term. For the load balance against the skew of search operations among each partitions with servers, RA-Btree utilizes replicated data. RA-Btree balances loads to servers with backup data by permitting search operations to access backup data. For the load balance against the skew of insert operations, RA-Btree utilizes servers with low loads. The servers with low loads synchronize among replicated data to balance loads to servers with low loads. Moreover, RA-Btree switches multiple modes for insert operations according to the load situation of networks. Experimental results using server clusters demonstrate that the proposed method can achieve a throughput of 3.5 times that for the conventional method in maximum-skew case.

Four novel optimization techniques in this dissertation contribute not only storage systems and relational database systems but also various kinds of applications using indexes such as key-value stores and graph databases in parallel and distributed systems. Although this thesis proposes and evaluates their techniques on combination of them, each technique contribute independently parallel and distributed systems. And their techniques improve continuously the performance of many applications on current and future hardware configurations.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).