大規模・追記型データベースにおけるデータレプリケーション方式

山岸 義徳† 中村 隆顕† 菅野 幹人†

†三菱電機株式会社 情報技術総合研究所 〒247-8501 神奈川県鎌倉市大船 5-1-1

E-mail: †{ Yamagishi.Yoshinori@cw, Nakamura.Takaaki@dy, Kanno.Mikihito@bc}.MitsubishiElectric.co.jp

あらまし データレプリケーションは、大規模データベースにおける負荷分散や障害対策を実現する手段として有効である．データレプリケーション性能の改善においては、複製データベースに配布する差分データの縮小化とデータ反映時間の短縮が求められる．我々は、センサデータやログのようにデータ更新が発生しない追記型データに適したデータレプリケーション方式を実現した．また、プロトタイプシステムを構築し、性能モデルに基づき大規模データのデータレプリケーション性能を評価し、概ね期待通りの結果を得た．本稿では、本データレプリケーション方式の実現、および性能評価結果について報告する．

キーワード データレプリケーション，負荷分散，性能，追記型データ

A Data Replication Scheme for a Large Write-Once Database.

Yoshinori YAMAGISHI† Takaaki NAKAMURA† and Mikihito KANNO†

†Information Technology R&D Center, Mitsubishi Electric Corporation

5-1-1 Ofuna, Kamakura-shi, Kanagawa, 247-8501 Japan

E-mail: †{ Yamagishi.Yoshinori@cw, Nakamura.Takaaki@dy, Kanno.Mikihito@bc}.MitsubishiElectric.co.jp

Abstract Data replication is useful for load balancing and disaster recovery planning in a large-scale database system. To improve performance of the data replication, it is required to minimize partial data size that is copied to replicas and to reflect the partial data to the replicas in a short time. We have implemented a data replication scheme for write-once data that cannot be modified such as sensor data and logs. The data replication scheme for the large-scale database system has been evaluated based on a performance model on a prototype system. The results were generally as expected. In this paper, we present implementation of the data replication scheme and the performance evaluation on the prototype system.

Keyword Data Replication, Load Balancing, Performance, Write-Once Data

1. はじめに

IT技術やセンサの進歩により、各種機器が出力するセンサデータやログは爆発的な増加傾向にある.近年、これら膨大に発生するデータをデータベースに収集・蓄積し、分析しようというニーズが高まっている．大規模化するデータベースにおいて、データレプリケーションはロード、検索などの処理競合に伴う負荷分散やサーバ故障等の障害対策を実現する手段として有効である．これまでにデータレプリケーションの方法としては様々な方法が提案されている[1][2].

データベースにおけるデータレプリケーションでは、マスタデータベースに対する更新ログを複製データベース側にネットワーク経由で配布し、更新ログに基づき複製データベースへ順番に適用する方式（ログ配布方式）が一般的である．しかし、大量データが発生するデータベース環境では、追加されるレコード件数の増大に伴い更新ログも大きくなる．このため、データ配布に伴うネットワークおよび複製データベースへのデータ反映がボトルネックとなり、データレプリケーションに時間がかかるという問題がある.

一方、センサデータやログは定期的、あるいはイベントによって不定期に発生し、後から更新されることがない追記型データの特性を持っている.

我々は、追記型データの特性を考慮した大規模データの高速処理を可能とする追記型データベースの開発に取り組んできた[3][4][5]．今回、この追記型データベースを用いて、データベースのページをデータ配布の単位とするデータレプリケーション方式を実現した.また、本データレプリケーション方式によるプロトタイプシステムを構築し、大規模データを対象に単純な性能モデルに基づく性能評価を実施した.

本稿では、大規模・追記型データベースに適したデータレプリケーション方式の実現、および本方式に基づくプロトタイプシステムによる性能評価結果について報告する.

2. 前提条件

2.1. システム構成

図 1はロード専用のマスタデータベース(マスタDB)を備えたマスタサーバ 1 台と検索専用の複製データベース(複製DB)を備えた検索サーバN台がネットワークを介して接続された 1:Nのシステム構成となっている.

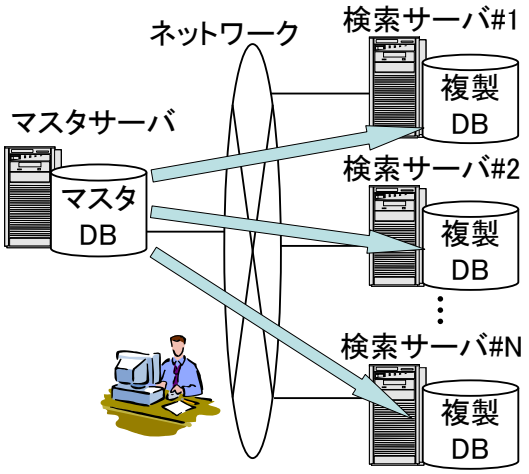

図 1　システム構成図

2.2. データ配布周期

レプリケーションによる最短のデータ配布周期は分単位から時間単位のオーダとする. これは分析系システムの要件に基づく. また、追記型データでは、更新に伴う同一レコードの競合は発生せず整合性を維持できるため、非同期方式によるレプリケーションを可能とする.

3. 従来の課題

データベース操作の更新ログを転送・反映する従来のログ配布方式では、大量データが発生すると更新ログのネットワーク転送量、複製 DB へ適用する更新件数（トランザクション数）とも増大するため、転送・反映に時間がかかる.

4. 実現方式

4.1. 追記型データベース

追記型データベースとして、センサデータやログなどの追記型データを対象とする専用データベース「高速集計検索エンジン」を開発している[3]. 高速集計検索エンジンは追記型データの集計・検索に適したブロック化トランスポーズドデータ配置方式[4]により、メモリ上に格納できない大規模データでも安定した高速処理を実現する.

高速集計検索エンジンのデータベース構造を図 2 に示す. データベースに格納する二次元構造のデータはページという単位に分割される. ページは列単位のブロックから構成され、複数行のデータをまとめた単位である. ページを単位にシェアード・ナッシングな構造とすることでデータ処理の分散・並列化を可能としている. また、ブロックをデータ圧縮[6]により縮小することでI/O量削減による高速化と合わせストレージ容量削減(概ね 1/10～1/50)を可能としている.

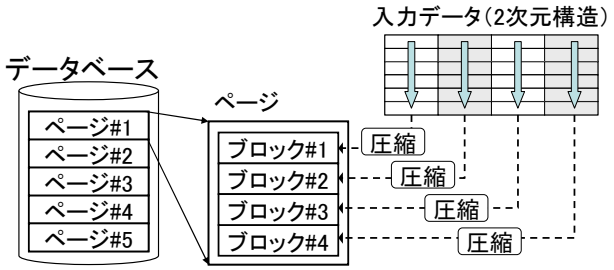

図 2　データベース構造

4.2. ページ配布方式

高速集計検索エンジンでは、更新が発生しない追記型データの特性を踏まえ、マスタ DB に追加された増分ページだけを抽出し、複製 DB に転送・反映するページ配布方式をとる. ページ配布方式では、圧縮ページが転送対象となるため、データ圧縮率に比例した実効的なデータ転送速度の向上が見込まれる. また、ログ配布方式では更新ログに従ってレコード単位の追加トランザクションが発生するが、ページ配布方式ではページを複製 DB へ追加するだけのデータコピーに近い速度でデータ反映が可能となる. これにより、大量データであっても高速なデータレプリケーションが可能となる.

4.3. 基本処理手順

ページ配布方式では、マスタ DB から複製 DB に対して増分ページを配布することにより実現する. 図 3 に基本処理手順を示す.

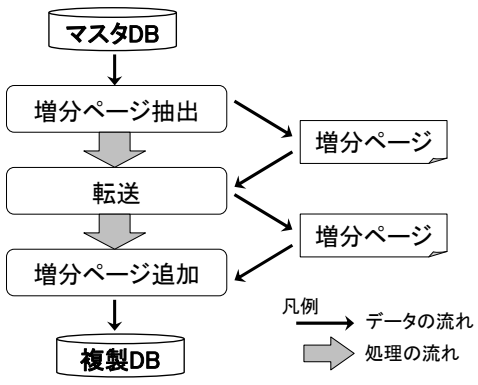

図 3　基本処理手順

マスタ DB から前回のデータレプリケーション以降に追加された増分ページを抽出し、ネットワーク経由で検索サーバへ転送する．検索サーバ側では受信した増分ページを複製 DB に追加することによりデータの反映を実現する．マスタ DB からの増分ページの抽出、および複製 DB への増分ページの反映はアーカイブ機能を活用する．

4.4. 複数の複製 DB へのページ配布管理

障害等によりデータレプリケーションを中断していた複製DBであってもマスタDBからの 1 回のページ配布でデータ反映を可能とするレプリケーション方式を実現する．この実現のため、配布ページ管理表(表 1)を設けて複製DBごとに最終配布ページ番号を管理する(図 4)．

表 1　配布ページ管理表

DB 名	最終配布ページ番号	ステータス
複製 DB1	ページ＃4	OK
複製 DB2	ページ＃2	OK
:	:	:

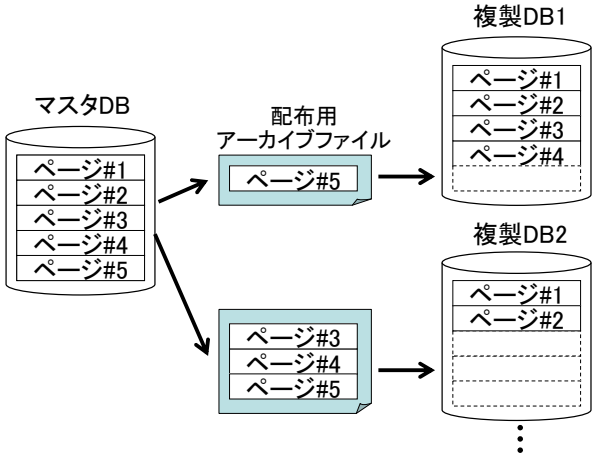

図 4　配布用アーカイブファイル

アーカイブ機能は配布ページ管理表を参照し、配布先の複製 DB に応じた配布用アーカイブファイルを生成する．これにより、データレプリケーションを中断していた複製 DB であっても他の運用中の複製 DB と同様に 1 回のページ配布でデータレプリケーションが可能となる．

4.5. ページ配布方式による性能モデル

本方式によるデータレプリケーション時間は式(1)で表される．ここで T_i は i 番目の複製 DB へのデータレプリケーション時間であり、N 台の複製 DB に対して合計する．

$$T = \Sigma T_i \qquad (1)$$

T_i は式(2)で表される．ここで、T_b は配布用アーカイブファイル生成時間、T_n はデータ転送時間、T_r はアーカイブ復元時間である．

$$T_i = T_b + T_n + T_r \qquad (2)$$

T_b 、T_n 、T_r は式(3)、(4)、(5)で表される．ここで、S は配布ページ合計サイズ、V_{br}、V_{bw} はそれぞれマスタサーバ上でのディスク読み出し、ディスク書き込み速度、V_n はネットワークの転送速度、V_{rr}、V_{rw} はそれぞれ検索サーバ上でのディスク読み出し、ディスク書き込み速度である．配布ページ合計サイズは、増分のデータサイズとデータ圧縮率に依存する．

$$T_b = S/V_{br} + S/V_{bw} \qquad (3)$$

$$T_n = S/V_n \qquad (4)$$

$$T_r = S/V_{rr} + S/V_{rw} \qquad (5)$$

本レプリケーション方式によるデータレプリケーション時間は配布ページ合計サイズ、ディスク I/O 速度、ネットワーク速度によって決まり、ネットワーク間の単純なファイルコピーと同等の速度が得られることが期待できる．

5. 評価

5.1. 評価方法

マスタ DB に追加されるデータ規模をパラメータにデータレプリケーションの速度性能を評価した．評価データには Web アクセスログ(レコード長 708B)を用い、1000 万件単位に 1 億件までのデータレプリケーション時間を測定した．

5.2. 評価環境

評価に用いたハードウェア構成を表 2に示す．ここでは基本構成としてマスタサーバと検索サーバ(1 台)がスイッチングハブ(ギガビット・イーサネット対応)により接続されている．

表 2　ハードウェア構成

	仕様	
マスタサーバ	CPU	Intel Xeon X6470 3.33GHz×2
	メモリ	42GB
	HDD	RAID5 構成(300GB,15000RPM×8)
	LAN	1000Base-T
	OS	Windows Server 2008 R2
検索サーバ	CPU	Intel Xeon X5450 3GHz×2
	メモリ	18GB
	HDD	RAID5 構成(1TB,7200RPM×8)
	LAN	1000Base-T
	OS	Windows Server 2008 SP2

5.3. 評価結果

配布データ件数とデータレプリケーション時間の関係について、実測値と性能モデルに基づく期待値の結果を図 5に示す．ここでは、1000 万件のデータ配布時のレプリケーション時間の期待値を 1 に正規化して表している．また、期待値は評価マシン上のディスク I/Oの基礎データから算出している．また、Webアクセスログのデータ圧縮率は実測値 94.8%(1/20)を利用した．

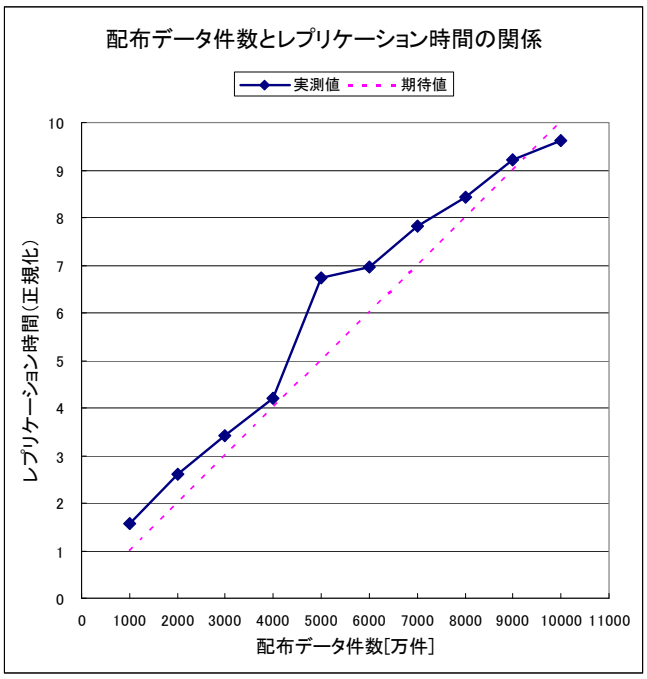

図 5　レプリケーション時間の実測値

この結果、性能モデルと比較し、概ね期待通りの結果が得られた．今回のページ配布方式では、配布用アーカイブファイルを介してデータレプリケーションを実現しているため、配布用アーカイブファイルに対する I/O がマスタ DB、複製 DB でそれぞれ発生する．これをメモリバッファ経由に変更することにより、数10%程度のレプリケーション性能の改善が見込まれる．

6. まとめ

本稿では、大規模・追記型データベースに適したデータレプリケーション方式としてページ配布方式を実現し、プロトタイプシステムによる性能評価を実施した．この結果について性能モデルに基づく期待値と比較し､概ね期待通りの結果が得られることを確認した．

今後は、複数の検索サーバへのレプリケーション中に特定の検索サーバの故障やネットワーク障害等が発生した場合の復旧処理の動作検証や、複数のマスタサーバ構成（M:N 構成）に関する調査を進めていく．

参　考　文　献

[1] F. D. Munoz, H. Decker, J. E. Armendariz, etc., “Database Replication Approaches”, Institute Tecnology of Information, Technical Report TR-ITI-ITE-07/19, October 5, 2007

[2] M. Wiesmann, F. Pedone, A. Schiper, B. Kemme, and G. Alonso. “Understanding replication in databases anddistributed systems”, In Proc. of the 20th International Conference on Distributed Computing Systems (ICDCS),Taipei, Taiwan, Republic of China, Apr. 2000.

[3] 山岸 義徳 , 平井 規郎 , 西村 達夫、”高速集計検索エンジンとセンサデータベースへの応用”,三菱電機技報, Vol.83, No.12, pp.11-14, 2009

[4] 郡 光則, 中村 隆顕, 山岸 義徳, “高性能並列情報検索技術”, 三菱電機技報, Vol.83, No.12, pp.7-10, 2009

[5] 上田尚純, 郡光則, 青野正宏,渡辺 尚, 水野忠則, ”ブロック化転置ファイルを利用したデータウェアハウス向けデータベース管理システムの評価”, 情報処理学会論文誌, Vol.42, No.SIG10, Sep 2001

[6] 加藤守, 谷垣宏一, 郡光則, “環境情報データベース向け高性能センサデータ圧縮方式”, 情報処理学会第 73 回全国大会, 2011