6L-05

遠隔バックアップ機能を伴う クラスタデータベースシステムの提案と実装

細谷 柚子† 三島 健‡ 小口 正人†
†お茶の水女子大学 ‡NTT ソフトウェアイノベーションセンタ

1 はじめに

金融や証券などミッションクリティカルなビジネスではDBMSに対して高性能化と高信頼化の両立が求められる．また長引く不況により低コスト化も重要な要件となり，更に，東日本大震災等を教訓に遠隔バックアップの需要も高まっている．これらの4つの要件を同時に満たせる手法の確立が本研究の課題である．まず高性能化と高信頼化のためには，レプリケーションを導入し，複数レプリカによる負荷分散が必要である．そして低コスト化のためには，汎用IAサーバ上でOSS等の利用が望ましい．更に，遠隔バックアップによる性能低下を防ぐために非同期レプリケーションも必要となる．

そこで我々は，OSSのDB同期ミドルウェア中では最も高性能であるPangea[1]に着目し，Pangeaが前述の4つの要件を満たすために必要な非同期レプリケーションによる遠隔バックアップ機能を加えた新たなミドルウェアを検討した．これをPangea**と呼ぶ．我々はTPC-Wベンチマーク[3]を使いPangea**を評価することで，高性能化，高信頼化，低コスト化を損なわない遠隔バックアップの実現可能性を議論する．

2 既存手法：Pangea

Pangeaはサーバの1台をLeader，その他はFollowerとして，クライアントからミドルウェアを介してサーバにアクセスすることで同期をとる．照会処理は1台のサーバで，更新処理は全てのサーバで実行され，更新処理の場合はLeaderに対して更新をした後に，Followerに対しても同様に処理を行う．Pangeaでは全てのDBサーバが同期されていることから，そのうちの1台を遠方に配置させることで，遠隔バックアップの実現は可能である．しかし，Pangeaからバックアップサーバへの通信による大きな遅延の影響により，大幅な性能低下を招く．また，我々は過去に，非同期的にバックアップを行う手法のPangea++の検討を行った．しかし，Pangea++はバックアップへの処理をシリアルに行う手法であり，バックアップを行うのに多くの時間を要した[2]．そのため，本研究ではバックアップへの処理を並列に実行することで効率化を図る．

Proposal and Implementation of a cluster database system with a remote backup function
† Yuzuko Hosoya, Masato Oguchi
‡ Takeshi Mishima
Ochanomizu University (†)
NTT Software Innovation Center(‡)

3 提案手法：Pangea**

3.1 Pangea**概要

本研究では，Pangeaに遠隔バックアップ機能を加えたPangea**を提案する(図1)．ローカルDB用サーバをマスタ，バックアップ用サーバをスレーブと呼ぶ．クライアントからの処理を分担するローカルDBは，従来のPangea同様，1台をLeader，その他をFollowerとしている．クライアントからの処理はマスタを介して行われる．DBの実行処理を担当しないバックアップサーバは，スレーブを介して非同期的に更新処理のみ行われるようにした．

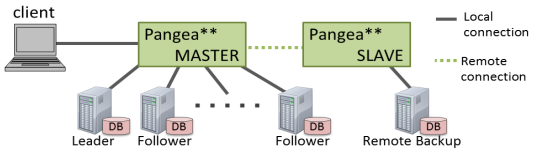

図 1: Pangea**構成

3.2 Pangea**実装

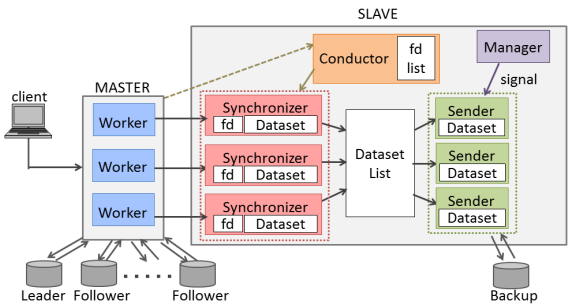

図 2: Pangea**モデル

Pangea**の実装を図2に示す．マスタでは複数のWorkerスレッドが，スレーブではConductorスレッド，Managerスレッドが1つずつと，Synchronizerスレッド，Senderスレッドが複数動作している．Workerスレッドと Synchronizerスレッドは1対1に対応付けられており，トランザクション1つを，Workerスレッド1つが処理するようになっている．Workerスレッドは，クライアントからのトランザクションを受信し，ローカルサーバにおいてクエリを実行する一方で，マスタでのトランザクション順序を明確にするためにトランザクションごとにタイムスタンプを付与する．タイムスタンプはトランザクションの開始を表すSTSと終了を表すETSの2種類を使う．そして，クエリとタイムスタンプの値をSynchronizerスレッドに転送する．Conductorスレッドは Workerスレッド起動時にマスタからの接続

を受付け，ファイルディスクリプタ fd を Synchronizer スレッドに渡す処理のみを行う．Synchronizer スレッドはクエリとタイムスタンプを受信し，トランザクションごとに保存する．これを Dataset と呼ぶ．全クエリを保存後，DatasetList に繋げる．Manager スレッドは，並列転送プロトコルに従って Sender スレッドに指示を出す．Sender スレッドは，DatasetList の先頭から Dataset を取り出し，Manger スレッドの指示に応じてバックアップサーバにてクエリを実行する．

3.3 Pangea**：並列転送プロトコル

並列転送プロトコルを図 3 に示す．このプロトコルによりバックアップへのクエリ実行を並列化させ，効率的にバックアップが可能となる．説明のための変数を表 1 に定義する．また，図 4 に Dataset の構造を示す．

表 1: 変数の定義

LC	スレーブの時刻を表す． commit 1 回につきインクリメントする．
NSTS	これから実行される Dataset の STS のうち 2 番目に小さい STS の値．

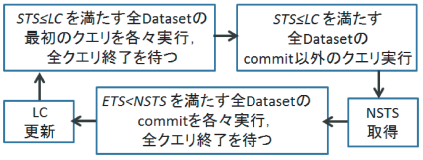

図 3: 並列転送プロトコル

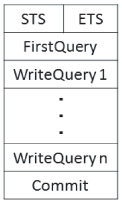

図 4: Dataset の構造

4 評価実験

4.1 実験環境

Pangea と Pangea**を用いて，遠隔バックアップ機能がトランザクション処理に与える影響を調査した．実験環境は，Web サーバとアプリケーションサーバに Tomcat6.0.37[4] を用いて，ローカル DB サーバ，バックアップ DB サーバは 1 台ずつ用意，それぞれに PostgreSQL9.2.6[5] を配置させた．バックアップは海外にあることを想定し，Dummynet を用いて RTT256ms の遅延を挿入した．Pangea**用マシンのスペックは，1.60GHz Intel(R) Xeon(R) E5310 の CPU，4 つの core，メモリ 2GB で，DB サーバ用マシンのスペックは，3.60GHz Intel(R) Xeon(TM) の CPU，1 つの core，メモリ 4GB のものである．OS はどちらも Ubuntu14.04 である．TPC-W は仮想的なブラウザ (EB) が、それぞれ照会処理と更新処理の割合が異なる browsing mix，shopping mix，ordering mix の 3 種のワークロードで DB にトランザクションを発行する．本稿ではその中で最も更新トランザクション処理の多い ordering mix で評価を行った．性能評価指標は，スループット (1 秒あたりの Web 画面表示数) とレスポンス時間 (1 画面データの転送時間) とした．遠隔バックアップをしない Pangea の通常動作（Pangea RTT0ms）を性能のベースラインとした．そして，Pangea で遠隔バックアップを行う場合（Pangea RTT256ms）と Pangea** とで性能比較を行った．

4.2 実験結果

実験の結果を図 5 に表す．Pangea，Pangea**の遠隔バックアップ時と Pangea の通常動作時の最大スループットを比較すると，Pangea で遠隔バックアップを行った場合は約 80 %の性能低下がみられた．他方，Pangea**では殆ど差が見られなかった．レスポンス時間については，Pangea で遠隔バックアップを行う場合には全ての EB 数で 5 秒以上となってしまっていた．他方，Pangea**で遠隔バックアップを行う場合には Pangea の通常動作時とほぼ変わらなかった．

この結果から，提案した Pangea**は，クライアントからのトランザクション処理に殆ど影響を与えずにバックアップ可能であることがわかった．

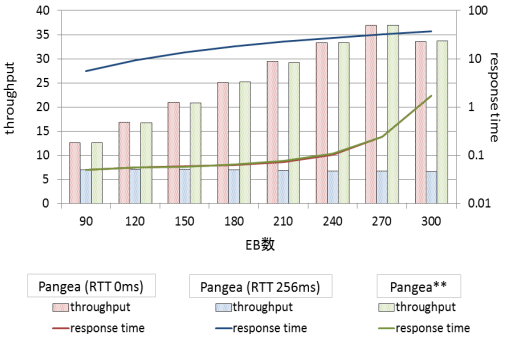

図 5: Pangea と Pangea**の性能評価

5 まとめ

遠隔バックアップ機能を伴うクラスタ DBMS の Pangea**を提案した．Pangea**は，非同期レプリケーションによるアプローチでバックアップを行う手法である．また，バックアップへのクエリ実行を効率化するために，並列転送プロトコルを導入した．評価を行った結果，クライアントからの処理に殆ど影響を与えずバックアップ可能であったことから，本手法は，高性能化，高信頼化，低コスト化を損なわない遠隔バックアップ手法であることが示せた．

参考文献

[1] T.Mishima,and H.Nakamura,”Pangea:An Eager Database Replication Middleware Guaranteeing Snapshot Isolation without Modification of Database Servers”,PVLDB2009,424-435.

[2] 細谷柚子, 三島健, 小口正人, ” データベース同期ミドルウェアによる遠隔バックアップの活用手法の検討 ” DEIMForum2015,B6-5,2015 年 3 月

[3] TPC-W http://www.tpc.org/tpcw

[4] Tomcat http://tomcat.apache.org/

[5] PostgreSQL https://www.postgresql.jp/