リモートバックアップ機能を有する クラスタデータベースシステムの実装と性能評価

理学専攻・情報科学コース　細谷 柚子

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．そこで本研究では，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検討した．これを Pangea**と呼ぶ．Pangea**を TPC-W ベンチマーク [2] を用いて評価することで，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　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．

3　Pangea**

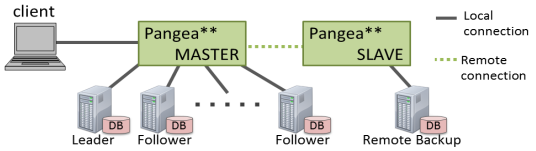

図 1: Pangea**構成

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)．ローカルデータベースサーバ用をマスタ，バックアップサーバ用をスレーブと呼ぶ．クライアントからの処理を分担するローカルのデータベースサーバは，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．データベースの実行処理を担当しないバックアップサーバは，スレーブを介して更新処理のみ行われるようにした．

3.1　Pangea**：アーキテクチャ

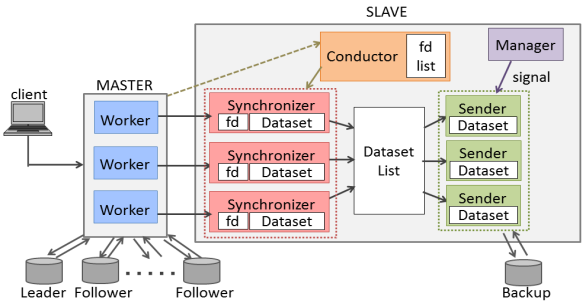

図 2: Pangea**モデル

STS	ETS
FirstQuery	
WriteQuery 1	
⋮	
WriteQuery n	
Commit	

図 3: トランザクションのフォーマット

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．本手法で扱うトランザクションは図 3 のようなフォーマットになる．Worker スレッドは，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．Dataset の構造は，図 3 に示したものと同等となる．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．

3.2　Pangea**：タイムスタンプ

タイムスタンプを記録するために，マスタの時刻を表すカウンタ (Master Logical Clock : MLC) を用意す

る．MLC は，トランザクションが commit したらインクリメントされる．Worker スレッドが各トランザクションの最初のクエリ実行時に，その時点の MLC の値を STS に記録し，commit 実行時には，その時点の MLC の値を ETS に記録する．タイムスタンプの例を図 4 に示す．MLC の初期値は 0 である．トランザクション 1(T1)，トランザクション 2(T2) が開始されると，それぞれの STS に MLC の値の 0 を記録する．その後 T1 が commit され，T1 の ETS に MLC の値の 0 を記録する．MLC はインクリメントされて 1 となり，次に commit された T2 の ETS には 1 を記録する．その後も同様にしてタイムスタンプを記録する．

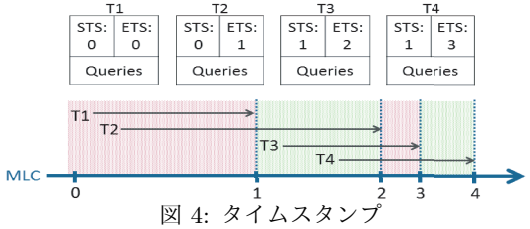

図 4: タイムスタンプ

3.3 Pangea**：並列転送プロトコル

並列転送プロトコルを下記に示す．このプロトコルによりバックアップへの処理を並列化でき，効率的な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．

表 1: 変数の定義

SLC	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．

並列転送プロトコル

1 STS が SLC 以下である全 Dataset のうち最初のクエリ実行全クエリの終了を待つ
2 STS が SLC 以下である全 Dataset のうち commit を除いた全クエリ実行
3 NSTS 取得
4 ETS が NSTS より小さい全 Dataset のうち commit 実行全クエリの終了を待つ
5 SLC 更新

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[3] 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[4] 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．以上の実験環境において，Pangea で遠隔バックアップを行う場合 (Pangea RTT256ms) と Pangea**との性能比較を行った．

4.2 実験結果

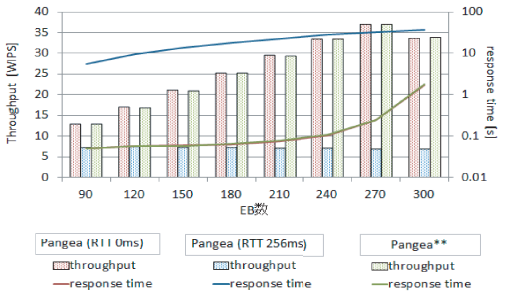

図 5: Pangea と Pangea**の性能評価

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．

5 まとめと今後の課題

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，高信頼性，低コストの各特性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．

今後は，スレーブの評価や既存のバックアップシステムとの比較を行い，本手法の有用性を示していきたい．

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] TPC-W http://www.tpc.org/tpcw
[3] Tomcat http://tomcat.apache.org/
[4] PostgreSQL https://www.postgresql.jp/