社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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提案

細谷　柚子†　　小口　正人†

† お茶の水女子大学
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: †yuzuko@ogl.is.ocha.ac.jp, ††oguchi@is.ocha.ac.jp

あらまし　近年，ビジネスフィールドにおけるデータベースに求められる処理性能がより高まり，そのためのシステム構成としてスケールアウトによるクラスタデータベースなどのシステムが提案されてきている．過去に我々の研究では，Pangea システムを検討してきた．一方で，災害対策としてデータベースの遠隔バックアップの必要性も問われている．ローカルのリソースだけでシステムを運用していると，大規模災害時にはバックアップも失ってしまうからである．日本のような地震国では特に，遠方にバックアップを置いておくことがますます重要になっている．本研究では，データの高い処理性能とデータ保護を両立させることを目標に，Pangea をベースにした，遠隔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 Pangea**の提案，実装を行った．そのシステムを検証した結果，クライアントからのトランザクション処理においてはパフォーマンスにほぼ影響を与えずに，リアルタイムにバックアップができる手法であることがわかった．

キーワード　リモートバックアップ，分散データベース，災害対策

Proposal of Cluster Database System with a Remote Backup Function

Yuzuko HOSOYA† and Masato OGUCHI†

† Ochanomizu University
Otsuka 2-1-1，Bunkyo-Ku，Tokyo 112-8610 Japan
E-mail: †yuzuko@ogl.is.ocha.ac.jp, ††oguchi@is.ocha.ac.jp

1. はじめに

近年の SNS や新しいデバイスの普及から，企業が管理しなければならないデータ量が日々刻々と増え続けている．そのため，ビジネスフィールドで多く使われている DBMS に高い処理性能が求められるようになり，そのためのシステム構成としてスケールアップではなくスケールアウトによるクラスタ DBMS などのシステムが注目されている．過去に我々の研究では Pangea システム [1] を提案，検討してきた．

一方で，２０１１年３月の東日本大震災では大量のデータが失われた教訓から，ローカルバックアップだけでなくリモートバックアップの必要性が重要視されてきた．情報社会においてデータを失うことは，システムやサービスの停止に直接つながり，企業にも顧客にも大きな損害を及ぼしてしまう．ゆえに日本のような地震国では，遠方にバックアップを置いておくことがますます重要になっている．

本研究では，Pangea をベースにした，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を検討し，データの高い処理性能とデータ保護を同時に実現することができる手法の確立を目標とする．

2. Pangea

Pangea は，LAN 環境を前提としたデータベース同期ミドルウェアである．データベースサーバの 1 台を Leader，その他は Follower として，クライアントから Pangea を介してデータベースサーバにアクセスすることで同期をとる．クライアントからの処理が照会処理の場合は，1 台のデータベースサーバで実行される．更新処理の場合は，全てのデータベースサーバで実行されるが，このときは Leader への処理が完了した後に，Follower に対しても同様に処理を行う．アプリケーションやデータベースを修正することなく，サーバを増やすことで性能を向上させることが可能である．

3. Pangea の遠隔バックアップ確立

まず，Pangea をそのまま用いて遠隔バックアップを確立さ

せる予備評価を行った．このとき，Follower を遠方に置くことで遠隔バックアップとした．データベースの実行処理を分担する Follower のデータベースは同期されていることから，これがバックアップとして機能する．サーバマシンは Leader，Follower 用に 1 台ずつ用意した．マシンのスペックを表 1 に示す．バックアップは海外にあることを想定し，Dummynet を使用して RTT256ms の遅延を挿入した．データベースは PostgreSQL9.2.6 [3] を使用し，全てのマシンに配置させた．Web サーバとアプリケーションサーバには Tomcat6.0.37 [4] を用いた．

性能評価は TPC-W ベンチマーク [2] を使用した．TPC-W は仮想的なブラウザ（以下 EB とする）が，データベースにトランザクションを発行する．TPC-W には 3 種類のワークロードがあり，それらの違いは表 2 に示すように照会処理と更新処理の割合が異なる．性能評価指標はスループット（1 秒当たりの Web 画面表示（WIPS））とレスポンス時間（1 画面データの転送時間（秒））とした．実験環境を図 1 に示す．

表 1　マシンのスペック

OS	Ubuntu14.04
CPU	Intel(R) Xeon(R) CPU 1.60GHz (4cores) / Intel(R) Xeon(R) CPU 1.60GHz (1core)
Memory	2GByte / 4GByte

表 2　ワークロード

ワークロード	Read-only	Update
Browsing mix	95 %	5 %
Shopping mix	80 %	20 %
Ordering mix	50 %	50 %

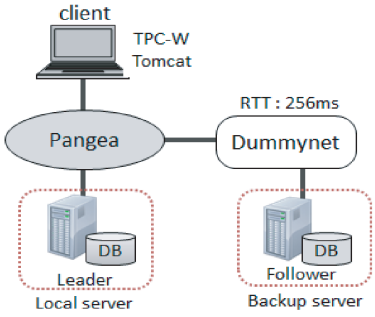

図 1　Pangea の実験環境

TPC-W の 3 種のワークロードのうち，照会処理の多い browsing mix と，更新処理の多い ordering mix の 2 つで実験を行った．Follower をローカルに配置した場合と，遠方に配置した場合で性能を比較した．結果を図 2，3 に示す．

棒グラフはスループット，線グラフはレスポンス時間を示している．横軸は EB 数である．Follower がローカルにある場合の Pangea の最大スループット値は，browsing mix では，EB が 150 のとき 15.71WIPS，ordering mix では，EB が 500 のとき 64.8WIPS となり，Follower が遠方にある場合の Pangea の

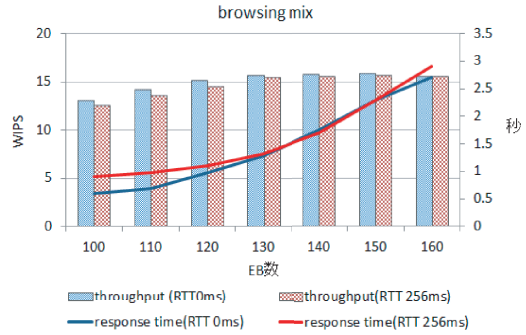

図 2　Pangea の browsing mix の性能

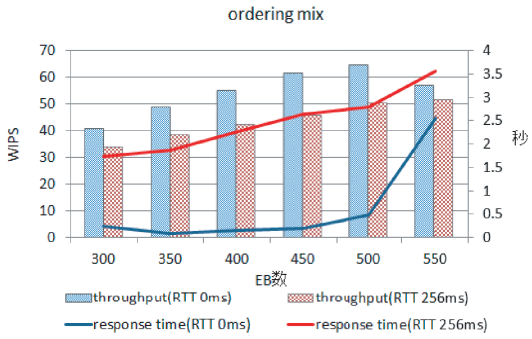

図 3　Pangea の ordering mix の性能

最大スループット値は，browsing mix では，EB が 150 のとき 15.6WIPS，ordering mix では，EB が 550 のとき 51.57WIPS となった．Follower を遠方に置くことによる最大スループットの低下率は，browsing mix では 0.7 %，ordering mix では 20 %あった．

レスポンス時間は，browsing mix では最大で 0.3 秒の増加 ordering mix では最大で 2.4 秒の増加がみられた．このように，両方のワークロードで性能が低下するのは，遠方にある Follower でクエリを処理するためである．また，ordering mix でより性能が悪化した理由は，他の mix と比較して遠方にある Follower で実行するクエリが最も多いからである．

4. Pangea**

3 節の実験より，Pangea を無修正でそのまま使うとパフォーマンスの低下を招くことが分かった．本稿ではこの問題を解消するために，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として，Pangea**を提案する．

4.1 Pangea**の概要

Pangea**は，Pangea にリモートバックアップ機能を付加させたミドルウェアで，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 (図 4)．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，マスタを介してデータベースサーバにアクセスし，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

み行うようにした．

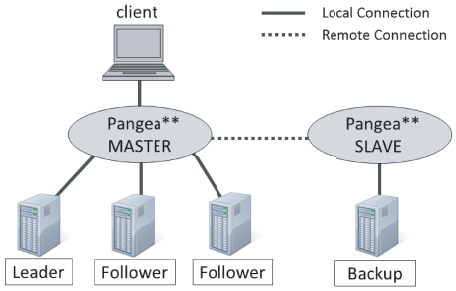

図 4　Pangea**構成図

4.2　Pangea**の実装

Pangea**の実装を図 5 に示す．

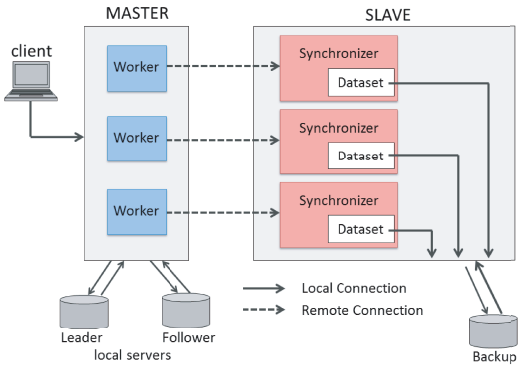

図 5　Pangea**実装例

マスタでは，複数の Worker スレッドがクライアントからのトランザクションを受け取り，クエリを解析した後，ローカルサーバにクエリを実行する．1 つのトランザクションに関して，1 つの Worker スレッドが処理を行うようにしている．送られてきたクエリが更新処理の場合は，スレーブに転送する．このとき，ローカルサーバでのトランザクションスケジューリングを再現するためのタイムスタンプを記録し，クエリと共にスレーブに転送する．

スレーブでは，Worker スレッドと関連付けられた複数の Synchronizer スレッドが，マスタから送られてきたクエリとタイムスタンプをトランザクションごとに保管する．これを Dataset と呼ぶ（図 6）．Synchronizer スレッドは commit を受け取ったら，一貫性を崩さないよう，送信アルゴリズムに従い，Dataset をバックアップサーバに送ることでローカルサーバと同期をとる．

タイムスタンプや送信アルゴリズムに関して，この次の節で詳しく述べる．

4.2.1　Pangea**：タイムスタンプ

タイムスタンプは，ローカルサーバでトランザクションがどの順で処理されたかというスケジューリングをバックアップサーバで再現するためのものである．トランザクションの始まりを表す Start Time Stamp(STS) と，終わりを表す End

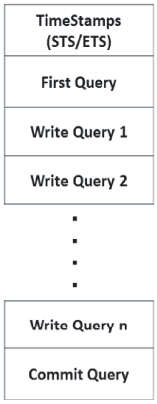

図 6　Dataset

Time Stamp(ETS) の 2 種類用意する．タイムスタンプを記録するため，マスタに Master Logical Clock（MLC）というカウンタを用意する．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．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 の値が STS として記録され，commit 実行時には，その時点の MLC の値が ETS として記録される．

タイムスタンプの例を図 7 に示した．MLC の初期値は 0 である．トランザクション 1（T1）が開始された時，STS に MLC の値の 0 が記録される．commit された時も同様に，ETS に MLC の値である 0 が記録される．T1 が commit されたため，MLC はインクリメントされて 1 となる．その後に実行されるトランザクション 2（T2），トランザクション 3（T3）も同様にして，それぞれの STS と ETS を記録する．

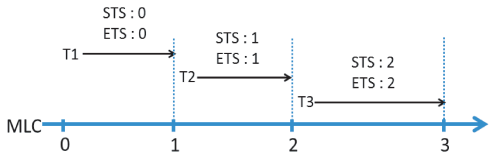

図 7　タイムスタンプ

スレーブは転送されたタイムスタンプを参照することで，各トランザクションのスケジューリングを把握することができる．

4.2.2　Pangea**：送信アルゴリズム

送信アルゴリズムは，スレーブがバックアップ処理を行うための手順である．このアルゴリズムのため，スレーブに Slave Logical Clock(SLC) というカウンタを用意する．SLC は前節の MLC 同様，トランザクションが commit したら，インクリメントされる．送信アルゴリズムを図 8 に示す．

Synchronizer 1 はトランザクション 1（T1）に，Synchronizer2 はトランザクション 2（T2）に関連付けられている．マスタから commit が来たら，Dataset に保存し，その時点の SLC と Dataset に保存していた STS とを比較する．STS が SLC 以下の場合はバックアップにクエリを実行し，それ以外の場合はスリープ状態となり，シグナルを待つ．バックアップに

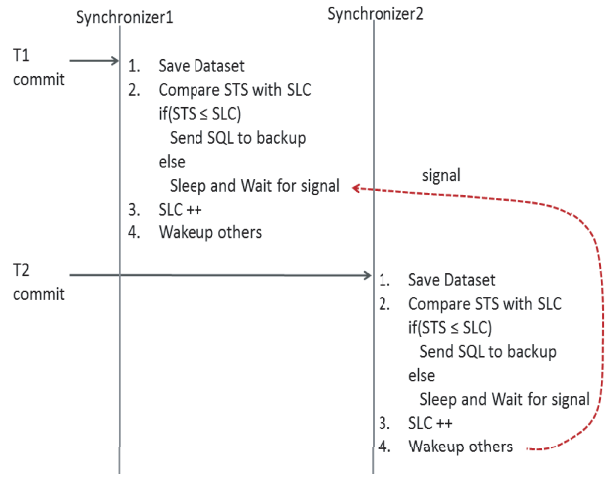

図 8　送信アルゴリズム

commit を実行した後，SLC をインクリメントし，スリープ状態になっている他のスレッドにシグナルを送る．

Pangea**はマルチスレッド実装を想定しているため，ローカルサーバとバックアップサーバの一貫性が崩れてしまう問題が懸念される．そのため，スレーブでは，この送信アルゴリズムを用いて，タイムスタンプを参照しながらバックアップ処理を行うことで一貫性を保つことができる．

5.　Pangea**の評価実験

5.1　実 験 環 境

提案手法 Pangea**の性能評価実験を行った．ローカルサーバ用にマシン 2 台，バックアップサーバ用にマシン 1 台を用いた．バックアップは海外にあることを想定し，Dummynet を使用して RTT256ms の遅延を挿入した．マシンや，データベース，Web サーバ，アプリケーションサーバ，ベンチマークは 3 節と同様のものを用いた．実験環境を図 9 に示す．

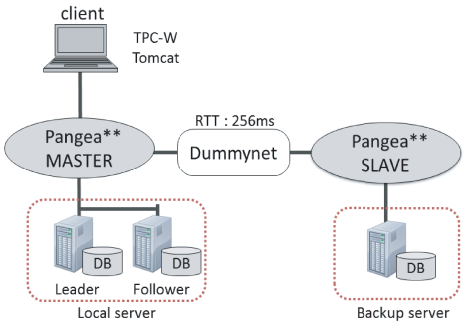

図 9　Pangea**実験環境

5.2　Pangea**の性能

3 節と同様，TPC-W の browsing mix と，ordering mix の 2 つで評価実験を行った．結果を図 10，11 に示す．

棒グラフはスループット，線グラフはレスポンス時間を示している．横軸は EB 数である．最大スループットは browsing mix では，EB が 150 のとき 15.76WIPS，ordering mix では，

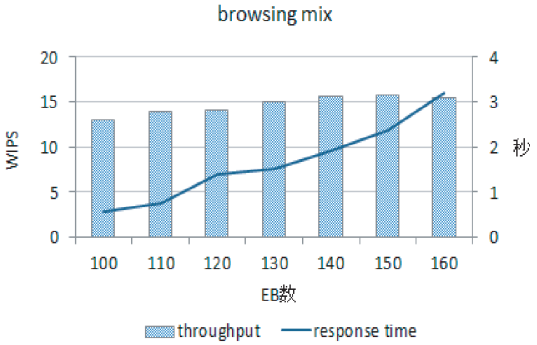

図 10　Pangea**の browsing mix の性能

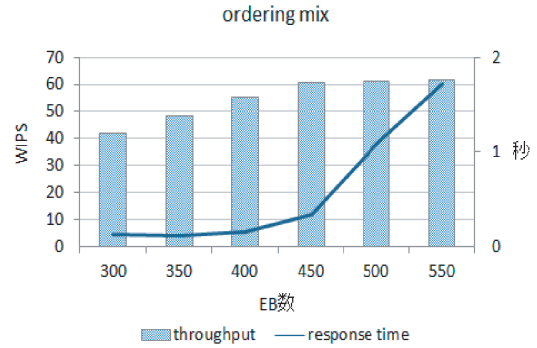

図 11　Pangea**の ordering mix の性能

EB が 550 のとき 61.74WIPS となった．3 章の Pangea の性能と比較すると，browsing mix では最大スループットは 0.3 %向上し，レスポンス時間は 0.08 秒増加した．ordering mix では，最大スループットは 4.7 %低下し，レスポンス時間は 1.23 秒増加した．Pangea**ではマスタがスレーブにクエリを転送することがオーバーヘッドになると考えられ，browsing mix のような照会処理が多い場合には，性能にほぼ影響しなかったが，ordering mix のように更新処理が多い場合には，わずかな性能低下が見られた．

6.　関 連 研 究

ストリーミングレプリケーションに代表されるようなログシッピング型のレプリケーションを使った遠隔バック アップが考えられる．しかし，これはデータベースサーバの種類とバージョンを合わせる必要があるという制約がある．それに対して Pangea**は標準の SQL を使っているため，ヘテロ構成が可能となる．これは，例えばサービスを止めずにバージョンアップする場合などに応用できる．

7.　まとめと今後の課題

本研究では，高いデータ処理性能とデータ保護を同時に実現できるシステムの確立を目標として，遠隔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検討を行った．既存システムの Pangea を用いて遠隔バックアップ確立を行う

と，パフォーマンスの低下がみられたため，それを解消すべく，Pangea をベースとした Pangea**を提案，実装した．Pangea**では，ローカルサーバへの処理はマスタが，バックアップサーバへの処理はスレーブが実行し，バックアップは更新処理のみ行われるようにした．

提案した Pangea**を TPC-W ベンチマークを用いて性能評価を行った．Pangea と Pangea**を比較すると，照会処理が多い場合は性能への影響は殆どなく，更新処理が多い場合は性能低下がわずかにあった．これは，マスタがスレーブにクエリを転送することがオーバヘッドになると考えられるが，スループット低下率は 5 %以内に抑えることができた．

本稿で行った実験は，マスタの処理性能を調査したものだが，スレーブのバックアップへの処理に関しても調査する必要があると考える．今後の課題としては，バックアップへの処理に関しても性能を評価し，課題抽出を行いたい．またその課題から，提案手法の改良，拡張を行いたい．

8. 謝　　辞

本研究を進めるにあたって，NTT の三島 健氏より大変有用なアドバイスをいただきました．深く感謝いたします．

文　　献

[1] T.Mishima and H.Nakamura,"Pangea:An Eager Database Replication Middleware guaranteeing Snapshot Isolation without Modification of Database Servers", Proc.VLDB2009, pp.1066-1077, August 2009. PVLDB2009.
[2] TPC-W http://www.tpc.org/tpcw
[3] PostgreSQL https://www.postgresql.jp/
[4] Tomcat http://tomcat.apache.org/