マルチテナント環境におけるデータベースのポストコピーライブ移送

村 並 勇 紀† 山 田 浩 史†

ライブデータベース（DB）移送は，DB を稼働させたまま別の物理マシンへ移動するテクニックである．クラウド環境などの複数テナントの DB を１つのマシンに集約するマルチテナントプラットフォームにおいては，ライブ DB 移送を用いる事で物理マシンのメンテナンスや負荷分散が容易化することが知られている [1)2)3)].

これまでにライブ DB 移送についての研究が行われてきた．Albatross[1)]ではプレコピー方式によるライブ DB 移送の提案を行なっている．プレコピー方式では，移送元でトランザクション処理を実行したまま，バックグラウンドでデータを転送する．その間に更新があったメモリページを転送し，この工程を繰り返す．これを繰り返した回数，もしくは更新のあったページ数が設定した閾値より小さくなったら，トランザクション受付を停止する．そして更新のあったページを移送先に転送し，移送先でトランザクション受付を再開する．この手法では移送中に多くのメモリページが更新された場合，移送期間の長期化してしまう．またメモリページ転送量も増大し，システム全体のパフォーマンスに悪影響を与え，負荷分散効果が薄くなってしまう問題も考えられる．Zephyr[2)]では移送先でトランザクション受付を開始できるようになった後も完全に移送が終了するまでは移送元でのトランザクション受付を停止する事なく，２台のマシンでトランザクション実行し続けるデュアルモードが存在する．これによりアボートされるトランザクションの数は少なくなるなどの利点は存在するが，負荷分散効果は薄くなってしまう欠点が挙げられる．Squall[3)]ではメインメモリ DB を対象にクラスタ内でのパーティションの再構成の最適化を行なった．リーダーノードを設定し，そのリーダーノードが全体の移送の動きを把握する事で適切に再構成が行われるようにする．ダウンタイムをなくすという目的は達成しているが，再構成にかかる時間は長期化してしまい，負荷分散効果は薄くなってしまう．

本研究ではライブバーチャルマシン（VM）移送で行われているポストコピー方式 [4)]をライブ DB 移送に用いる．ポストコピー方式では最初にトランザクション実行に必要な最小限度の情報（ワイヤフレーム）を移送先に転送し，移送先でトランザクション処理を受付ける．転送されていないデータの実体は順次転送されるが，トランザクション処理でまだ転送されていない領域の実体が必要になると該当ページを優先的に移送元からフェッチする．プレコピー方式と比較すると，ポストコピー方式は全てのメモリページの転送を１度しか行わないため移送期間の長期化を避ける事ができる．また，早期に移送先でトランザクション受付を開始するため，即座に負荷分散効果を得る事ができる．

トランザクション実行を即座に切り替えられるようにするため，対象とする DBMS の調査も行なった．その結果，ログを辿る事で現在の状態が再現できるような機能が実装されている事，レプリケーションモードを標準で搭載している事などから PostgreSQL で実装を行なっていく事を決定した．

現在，性能の比較を行う際に必要な Stop & Copy 方式を実装した．これは移送元でのトランザクション処理を完全に停止し，移送先へデータの移送を完了させてから移送先でトランザクション受付を再開するものである．最も基本的な移送方式のため提案手法との比較に使用される事が多い．またその後提案手法を実現するためにメモリ上のデータ構造を調査し，移送先へコピーする機能を実装している．

今後は，実装を完了させて単純なワークロードを稼働させるところから始まり，TPC-C などのベンチマークソフトを用いた実験を行い，Stop & Copy などと性能の比較検討を行う予定である．

参 考 文 献

[1] Sudipto Das, Shoji Nishimura, Divyakant Agrawal and Amr El Abbadi. Albatross: Lightweight Elasticity in Shared Storage Databases for the Cloud using Live Data Migration. In Proceedings of the VLDB Endowment, Volume 4 Issue 8, pages 494-505. May 2011.

[2] Aaron J. Elmore, Sudipto Das, Divyakant Agrawal, Amr El Abbadi. Zephyr: live migration in shared nothing databases for elastic cloud platforms. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, Pages 301-312. June 12 - 16, 2011.

†東京農工大学 Tokyo University of Agriculture and Technology

[3] Aaron J. Elmore, Vaibhav Arora, Rebecca Taft, Andrew Pavlo, Divyakant Agrawal, Amr El Abbadi. Squall: Fine-Grained Live Reconfiguration for Partitioned Main Memory Databases. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Pages 299-313. May 31 - June 04, 2015

[4] Michael R. Hines and Kartik Gopalan. Post-Copy Based Live Virtual Machine Migration Using Adaptive Pre-Paging and Dynamic Self-Ballooning. Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments. Pages 51-60. March 11 - 13, 2009.