マルチメディア通信と分散処理 95－14
（1999. 11. 19）

データベース移動に基づく分散データベースシステム DB–MANα の性能評価

海貝 明道　酒井 仁　秋山 豊和　原 隆浩
塚本 昌彦　西尾 章治郎

大阪大学大学院工学研究科情報システム工学専攻
E-mail: {umigai, sakai, akiyama, hara, harumoto, tuka, nishio}@ise.eng.osaka-u.ac.jp

概要　近年，ネットワークの帯域幅の拡大に伴い，分散システムでは，データの転送遅延よりもデータの伝播遅延の方が処理時間に影響するようになってきている．筆者らの研究グループでは，分散データベースシステムにおいて，通信回数を削減し，伝播遅延の影響を小さくするために，データベースを移動してトランザクションを処理する手法（データベース移動）を提案している．さらに，データベースへのアクセスの状況を考慮して，トランザクションの処理方法を適応的に選択する手法を提案し，これらの提案に基づいた分散データベースシステム DB–MANα を実装している．本稿では，実装した DB–MANα システムの性能を実測評価し，実環境での有用性を検証した．その結果，従来の分散データベースシステムに比べて，DB–MANα システムはトランザクションの平均応答時間を大幅に短縮できることを確認した．

キーワード：　データベース移動，分散データベースシステム，トランザクション処理

A Performance Evaluation of the DB–MANα System Based on Database Migration

Akimichi UMIGAI　Shinobu SAKAI　Toyokazu AKIYAMA　Takahiro HARA
Masahiko TSUKAMOTO　Shojiro NISHIO

Dept. of Information Systems Engineering, Graduate School of Engineering,
Osaka University

Abstract　Due to the recent expansion of network bandwidth, the data propagation delay is becoming a significant factor for the system performance rather than the data transmission delay. Based on this fact, we have proposed a new technology to reduce the bad influence of propagation delay on the distributed database system by relocating dynamically the database through networks, which we call *database migration*. Furthermore, we have proposed a database relocation method to choose the transaction processing method between the conventional database fixed method and the proposed database migration method by giving consideration to the transaction access pattern, and implemented the DB–MANα system based on these proposals. In this paper, we evaluate the performance of the DB–MANα system, and verify the effectiveness of this system in a practical environment. The results show that the DB–MANα system improves the transaction response time compared with the conventional systems.

Keywords:　database migration, distributed database system, transaction processing

1　はじめに

近年，ATM（Asynchronous Transfer Mode: 非同期転送モード）方式を中心としたネットワーク技術の発展に伴い，ネットワークの帯域幅が急速に拡大している．帯域幅の拡大により，大量のデータを短時間で転送できるようになるため，広域ネットワーク上での分散データベースシステムにおいては，データを転送するのにかかる時間（転送遅延）よりも，データがネットワークを伝わる際の遅延（伝播遅延）の方が処理時間に影響するようになる [2, 3]．

従来の分散データベースシステムでは，データベースは特定のサイトに固定され，処理依頼と処理結果のメッセージ交換によって処理を行っていた．広帯域ネットワークを利用すれば，処理を行うサイトへデータベースを移動してローカルに処理を行う

ことで，サイト間通信の回数を減らすという手法が可能となる．筆者らの研究グループでは，このような処理手法としてデータベース移動を提案した [5].

さらに，筆者らの研究グループは，文献 [4] において，移動機能を有する分散データベースシステム DB-MAN を提案し，文献 [6] においてそのプロトタイプシステム DB-MANα の実装について述べた．データベース移動を用いた処理（移動処理）が従来のデータベース固定型の処理（固定処理）に比べて常に処理時間が短いとは限らないことを考慮して，DB-MANα システムでは，文献 [1] において提案した固定処理と移動処理を適応的に選択する手法を用いてトランザクションを実行するように設計を行った．

ここで，文献 [6] では，システムの動作確認を目的として，システム内のデータベースサイズが非常に小さいといった単純な環境での実測評価は行っているが，これは DB-MANα システムの実環境での有効性を示すものではなかった．そこで本稿では，実環境での利用を想定したシステムのパラメータに基づいて，DB-MANα システムの性能評価実験を行う．さらに，その結果から，システムの問題点や今後の拡張について考察する．

2 システムの概要

本章では，文献 [6] で提案した DB-MANαシステムの概要について述べる．

DB-MANα システムでは，文献 [1] の手法に基づいて，トランザクション開始時に直前のアクセス情報からアクセスの偏りを検出し，固定処理と移動処理を適応的に選択して処理を行う．また，DB-MANα システムでは，データベース移動の高速化のために，主記憶データベースを用いているため，移動を考慮したバックアップ管理を行う．

図 1に DB-MANαシステムのシステム構成を示す．以下では，図中のそれぞれのモジュールの機能について簡単に述べる．これらのモジュールのうち，手法選択部とバックアップ管理部以外は従来の分散データベースとほぼ同様のものである．

インタフェース部: クライアントからトランザクションを受け取り，トランザクション処理手法選択部に問合せを一つずつ渡す．一方，部分問合せや移動要求などのシステムから受け取った処理要求は，直接トランザクション処理部に渡す．

トランザクション処理手法選択部: トランザクション処理手法選択部は，次の 3 つの各部からなる．

解析部: クライアントから文字列で渡された問合せを解析し，問合せ木を構成して最適化部に渡す．

最適化部: 問合せ木を，実際のデータベース操作を記述したプラン木に変換し，トランザクション処理部へ渡す．

手法選択部: トランザクションを固定処理と移動処理のどちらで処理するかを決定する．移動処理を選択した場合，データベース所持サイトのシステムコマンド実行部に対してデータベース移動要求を行う．

トランザクション処理部: トランザクション処理部は，次の 3 つの各部からなる．

トランザクション実行部: トランザクション処理手法選択部からプラン木を受け取り，指定された処理手法でプラン木を実行する．

システムコマンド実行部: 自サイトおよび他サイトのトランザクション実行部およびシステムコマンド実行部から，データベース移動，部分問合せの実行，カタログ情報の更新などのシステムによる処理要求（システムコマンド）を受け取り，実行する．

バックアップ管理部: トランザクション実行部から更新操作の内容を受け取り，更新ログを作成して物理記憶に格納する．また，バックアップの送信要求を受けたときには，バックアップと更新ログの内容から最新のバックアップを作成

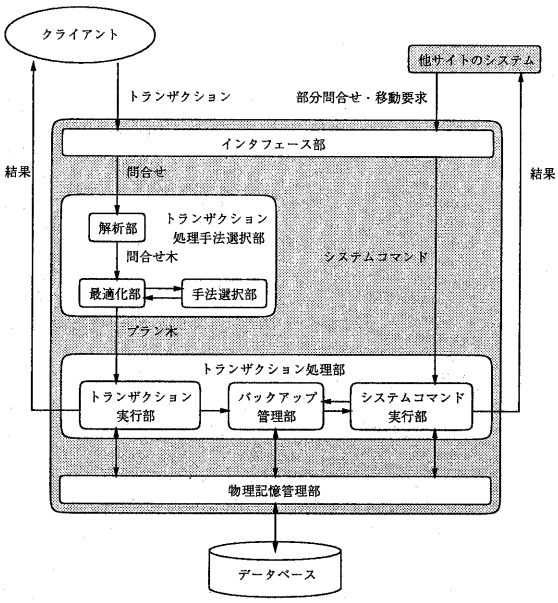

図 1: システム構成

し，これを送信する．このとき，自サイトにバックアップがない場合は，まず，バックアップを持つサイトのバックアップ管理部と通信して，これを受信する．

物理記憶管理部: トランザクション処理部からの要求に応じて，物理記憶に対してデータの入出力を行う．

なお，システムの実装は，カリフォルニア大学バークレー校で開発されたアカデミックフリーのリレーショナルデータベースシステムである POSTGRES [7, 8] をベースに行った．

3 実測評価

本章では，実装した DB-MANα システムの性能評価のために行った実験とその結果について述べ，さらに結果について考察する．

3.1 実験環境

実験では，DB-MANα システムのトランザクション処理にかかる平均応答時間を測定し，固定処理のみで処理した場合，および，移動処理のみで処理した場合と比較する．

実験を行ったシステム環境と，各実験で用いたパラメータを表 1に示す．実験は，表中の*印をつけたパラメータについて，それぞれ括弧内に示した範囲で値を変化させて行った．変化させるパラメータ以外は括弧外の値に固定した．

全ての実験で，サイト数は 3 に，データベース数は 1 に固定した．これは，DB-MANα の手法選択機構では，選択条件はサイト数に依存せず，データベース毎に独立して処理手法を選択するため，これらのパラメータを変化させても結果に影響しないものと考えられるからである．ネットワークは，100 Mbps のイーサネットを用いており，その実効帯域幅は約 80 Mbps である．また，伝播遅延を変化させる実験を行うため，伝播遅延はプログラム内で疑似的に発生させた．

現時点の DB-MANα システムでは，データベース移動とデータベース操作の並行処理制御機構の実装を行っていないため，データベース移動中に他のトランザクションを実行できない．そこで実験では，トランザクションを直列に実行した場合の平均応答時間を測定する．各サイトにおけるトランザクションの発生間隔は，指数分布に基づいて，平均発生間隔をパラメータとして与えて決定する．次のトランザクションの発生時刻になっても，前のトランザクションが終了していない場合は，前のトランザクションが終了するのを待つ．ただし，この待ち時間は応答時間には含まない．

表 1: 実験環境およびパラメータ

ネットワーク環境	
サイト数	3
実行帯域幅	約80 [Mbps]
伝播遅延*	200 [ミリ秒] (50〜400 [ミリ秒])

データベース	
データベース数	1
データベースサイズ*	31.5 [メガバイト] (3.5〜94.5 [メガバイト])

トランザクション	
問合せ/トランザクション*	10 (1〜30)
アクセスするデータ量*	1000 [タプル] (600〜15000 [タプル])
発生間隔変更周期*	400 [秒] (30〜1600 [秒])
発生間隔比*	5 (1〜20)
集中時平均発生間隔	30[秒]
散発時平均発生間隔	150 [秒] (30〜600[秒])

特定サイトからの集中的なアクセスは，ある一つのサイトの平均発生間隔を，それ以外のサイトの平均発生間隔に比べて小さな値に設定することで発生させる．集中してアクセスが発生するサイトの平均発生間隔を集中時平均発生間隔，それ以外のサイトの平均発生間隔を散発時平均発生間隔と呼ぶ．集中的にデータベースにアクセスするサイトは一様分布の乱数によって決定し，発生間隔変更周期ごとにそのサイトを変化させる．また，発生間隔比は (散発時平均発生間隔)/(集中時平均発生間隔) で与え，この値が大きい程，特定サイトから集中的にアクセスが発生していることを表す．

以下の実験結果のグラフでは，次の表記を用いる．

DB-MANα: DB-MANα システムの手法選択機構を用いて，トランザクション処理方法を適応的に選択した場合の平均応答時間．

MIG: 全てのトランザクションを，移動処理のみで処理した場合の平均応答時間．トランザクション開始時にローカルサイトにないデータベースを，全てローカルサイトに移動してから処理する．

FIX: 従来の分散データベースシステムと同様に，全てのトランザクションを，固定処理のみで処理した場合の平均応答時間．

3.2 実験 1 : データベースサイズの影響

データベースのサイズを変化させたときの平均応答時間を図 2 に示す．

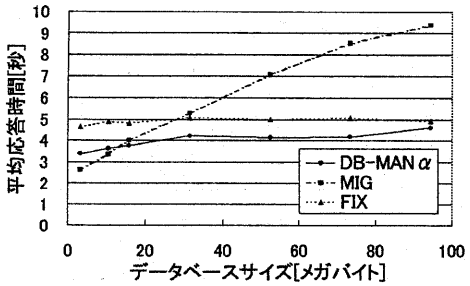

図 2: データベースサイズを変化させた結果

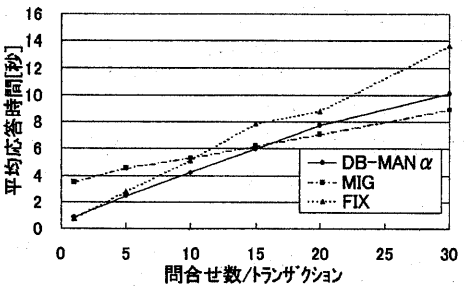

図 3: 問合せ数を変化させた結果

データベースサイズが小さいときには，データベースの移動にかかる時間が短いため，固定処理に比べて通信回数が少ない移動処理が良い結果を示す．DB-MANα では，ほとんどのトランザクションで移動処理を選択するため，移動処理のみの場合の性能と近くなる．しかし，DB-MANα では，同じサイトから連続してトランザクションが発生した場合のみ移動処理を選択するため，トランザクション発生サイトが変化した直後のトランザクションでは必ず固定処理を選択する．そのため，データベースサイズが非常に小さいとき，移動処理のみの場合より平均応答時間が少しだけ長くなる．

データベースサイズが大きいときは，データベース移動にかかる時間が大きいため，移動処理より固定処理が良い結果を示す．データベースサイズが大きくなると，DB-MANα ではほとんどのトランザクションで固定処理を選択するため，固定処理のみの場合の性能と近くなる．

全般的には，DB-MANα は，データベースサイズを考慮して処理手法を選択するため，データベースサイズに関わらず良い結果を示している．

3.3 実験 2：トランザクションに含まれる問合せ数の影響

トランザクションに含まれる問合せ数を変化させたときの平均応答時間を図 3に示す．

DB-MANα では，トランザクションに含まれる問合せ数が少ないとき，特定のサイトでアクセスが集中していると判断されないため，ほとんどのトランザクションで固定処理を選択する．このため，DB-MANα と固定処理のみの場合の結果が近い値を示している．しかし，稀に特定サイトから長期にわたって集中的なトランザクションが発生することにより，あまり有効でないデータベース移動を行うため，トランザクションに含まれる問合せ数が極めて少ないときには，DB-MANα は固定処理のみの場合よりわずかに悪い結果を示している．

また，トランザクションに含まれる問合せ数が多いときは，固定処理では通信回数が多くなるため，移動処理が良い結果を示す．このとき，DB-MANα では，ほとんどのトランザクションで移動処理を選択するため，移動処理のみの場合に近い結果を示している．トランザクションに含まれる問合せ数が非常に多い場合には，常に移動処理を行った方が有効であるため，アクセス集中時にのみ移動処理を行う DB-MANα が移動処理のみの場合よりも悪い結果を示している．しかし，このように非常に多くの問合せを含むトランザクションは，特定のアプリケーションから発生するバッチ処理である場合が多いため，手法選択機構にトランザクションを発生したアプリケーションを判別する機能をもたせることで，性能を改善できる．

3.4 実験 3：伝播遅延の影響

伝播遅延を変化させたときの結果を図 4に示す．

伝播遅延が小さいとき，通信回数は処理時間にそれほど影響しないため，移動処理より固定処理が良い結果を示す．このとき，DB-MANα はほとんどのトランザクションで固定処理を選択するが，実験 2 の場合と同様に，稀に長期にわたって集中的なトランザクションが発生することによってあまり有効でないデータベース移動を行うため，伝播遅延が非常に小さいときには，固定処理のみの場合よりもわずかに悪い結果を示している．

伝播遅延が大きいときは，通信回数が処理時間に大きな影響を与えるため，固定処理より移動処理が良い結果を示している．このような場合，DB-MANα では移動処理が選択されることが多くなるため，移動処理のみの場合に近い結果となる．また，伝播遅延が非常に大きくなると，常に移動処理を行った方が有効であるため，DB-MANα よりも移動処理のみの場合の方がわずかに良い結果を示す．しかし，実環境において，伝播遅延が 400 ミリ秒を越えることは稀なので，DB-MANα の有効性が生かせる状

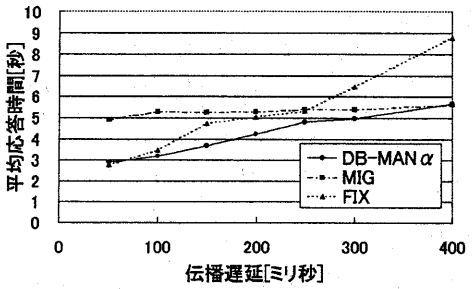

図 4: 伝播遅延を変化させた結果

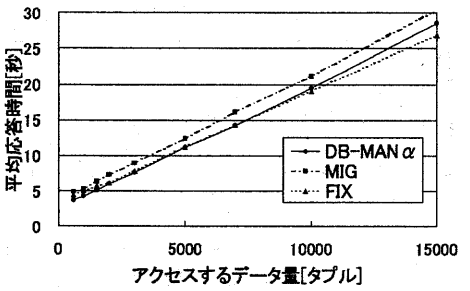

図 5: アクセスするデータ量を変化させた結果

況は多いと考えられる.

3.5 実験 4 : 各問合せがアクセスするデータ量の影響

1つの問合せがアクセスするデータ量を変化させたときの平均応答時間を図 5に示す.

アクセスするデータ量が少ないときは, 固定処理において問合せ結果を転送する時間が短くて済むため, 移動処理のみの場合よりも固定処理のみの場合が良い結果を示している. また, このような場合には, DB-MANα では固定処理が多く選択されるようになる. 移動処理の処理時間とリモートサイトに対する固定処理の処理時間との差が小さいため, アクセスするデータ量が非常に小さい場合も, 散発的なデータベース移動による結果への影響は小さく, その結果, DB-MANα が最も良い結果を示している.

アクセスするデータ量が多くなるにつれて, 固定処理における問合せ結果のデータ転送量は増加するが, 平均応答時間は移動処理に比べて短くなっている. これは, 各サイトにおける問合せの処理, および, リモートサイトとトランザクション発生サイト間のデータの受け渡しを並行に実行できるためである. つまり, リモートサイトから結果を転送している間に, トランザクション発生サイトで処理を実行でき, また, リモートサイトで部分問合せを実行している間に, トランザクション発生サイトでは, 既に受け取った結果の処理を行える. このようなことから, 固定処理が移動処理よりも良い結果を示している.

各サイトでの処理や結果の転送は非同期に実行されるため, 処理時間の見積もりが難しいことから, DB-MANα では, それらが直列に処理される場合の処理時間を見積もり値として計算し手法の選択を行う. そのため, アクセスするデータ量が多くなると, DB-MANα では, 移動処理を選択しやすくなり, 固定処理のみの場合よりも悪い結果を示している.

したがって, アクセスするデータ量が多くなる場合には, 各サイトでの処理や結果の転送などの非同期な実行を考慮するように, 手法選択機構を拡張する必要がある.

3.6 実験 5 : 発生間隔変更周期の影響

トランザクションの発生間隔変更周期を変化させたときの平均応答時間を図 6に示す.

発生間隔変更周期が短いときには, 特定サイトからの集中的なアクセスがそれほど継続しないため, 移動処理より固定処理が良い結果を示す. このとき, DB-MANα ではほとんどのトランザクションで固定処理を選択する. 実験 2 および実験 3 の場合と同様に, 稀に起こる長期にわたる集中的なトランザクションの発生により, 有効でないデータベース移動が行われるため, 発生間隔変更周期が非常に短いときには, 固定処理のみの場合の方がわずかに良い結果を示している.

発生間隔変更周期が長くなるにつれて, 特定サイトからのアクセスが長くなるため, DB-MANα と移動処理のみの場合の平均応答時間が短くなる. しかし, 移動処理のみの場合では, 集中的なアクセスの間に他サイトから散発的に発生するトランザクションよってあまり有効でないデータベース移動を行うため, 平均応答時間は固定処理のみの場合よりも長くなる.

DB-MANα は, アクセスの集中する度合いを考慮して適応的にデータベース移動を行うので, 全般的に良い結果を示している.

3.7 実験 6 : 発生間隔比の影響

トランザクションの発生間隔比を変化させたときの平均応答時間を図 7 に示す.

発生間隔比が小さいときは, 特定サイトからの集中したアクセスがほとんど発生しないため, 移動処理より固定処理が良い結果を示す. このとき, DB-MANα はほとんどのトランザクションで固定処理

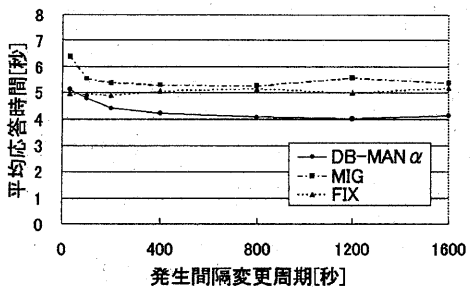

図 6: 発生間隔変更周期を変化させた結果

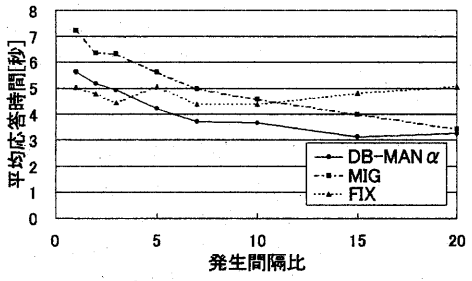

図 7: 発生間隔比を変化させた結果

を選択するが，実験 2，実験 3，および，実験 5 の場合と同様に，稀に起こる集中的なトランザクションの発生によって有効でないデータベース移動が行われるため，固定処理のみの場合よりもわずかに悪い結果を示している．

また，発生間隔比が大きいとき，すなわち特定サイトからアクセスが集中するときは，固定処理より移動処理が良い結果を示す．発生間隔比が大きくなるにつれて，DB-MANα はほとんどのトランザクションで移動処理を選択するようになるため，移動処理のみの場合に近い結果を示している．

全般的には，DB-MANα は発生間隔比によらず良い結果を示している．アクセスが極端に集中する場合や，全く集中しない場合でも，従来の固定処理のみの場合や移動処理のみの場合とほぼ同じ性能を示す．

4 おわりに

本稿では，移動機能を有する分散データベースシステム DB-MANα の実環境における有効性を検証するため，様々なシステム環境における実測評価を行った．DB-MANα では，データベースに対するアクセスの状況から，適応的に処理方法を選択してトランザクション処理を行うため，従来の分散データベースに比べ，多くの場合で処理時間を短縮できることを確認した．

今後は，データベースの移動中に発生したトランザクションに対する並行処理制御について考える必要がある．また，各サイトにおける処理や結果の転送などの非同期な実行を考慮するように，DB-MANα システムの手法選択機構を拡張する必要がある．

謝辞

本研究は，日本学術振興会科学研究費奨励研究 (A)(10780260) および日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」(Project No. JSPS-RFTF97P00501) の研究助成によるものである．ここに記して謝意を表す．

参考文献

[1] 秋山 豊和，原 隆浩，春本 要，塚本 昌彦，西尾 章治郎: “トランザクション系列に基づくデータベース移動を用いたデータベース再配置手法,” 情報処理学会マルチメディア通信と分散処理ワークショップ論文集, pp. 153–159 (Nov. 1998).

[2] S. Banerjee, V.O.K. Li, and C. Wang: “Distributed database systems in high-speed wide-area networks,” *IEEE Journal on Selected Areas in Commun.*, vol. 11, no. 4, pp. 617–630 (May 1993).

[3] S. Banerjee, C. K. Panos: “Network latency optimizations in distributed database systems,” *Proc. of IEEE Data Engineering.*, pp. 532–540 (Feb. 1998).

[4] T. Hara, K. Harumoto, M. Tsukamoto, and S. Nishio: “DB-MAN: A distributed database system based on database migration in ATM networks,” *Proc. of IEEE Data Engineering.*, pp. 522-531 (Feb. 1998).

[5] T. Hara, K. Harumoto, M. Tsukamoto, and S. Nishio: “Database migration: A new architecture for transaction processing in broadband networks,” *IEEE Trans. Knowledge and Data Eng.*, vol. 10, no. 5, pp. 839-854 (Sept./Oct. 1998).

[6] 酒井 仁，秋山 豊和，海貝 明道，原 隆浩，春本 要，塚本 昌彦，西尾 章治郎: “移動機能を有する分散データベースシステム DB-MANα の設計と実装,” 電子情報通信学会第 10 回データ工学ワークショップ (DEWS'99) 論文集 (CD-ROM) (Mar. 1999).

[7] M. Stonebraker and L.A. Rowe: “The design of POSTGRES,” *Proc. of ACM SIGMOD'86*, pp. 340–355 (June 1986).

[8] M. Stonebraker, L.A. Rowe, and M. Hirohama: “The implementation of POSTGRES,” *IEEE Trans. Knowledge and Data Eng.*, vol. 2, no. 1, pp. 125–141 (Mar. 1990).