大量データをリアルタイムに処理する「InfoFrame Table Access Method」

大沢 英紀・宮田 剛

要　旨

ビッグデータ時代では、大量に発生するデータをリアルタイムに処理することにより、新しい価値やビジネスを創出するニーズが高まっています。大量データをリアルタイムに処理するためには、個々のデータ処理を高速化することに加え、高スループットを実現することが必要です。本稿では、並列データの高速処理が可能で、大量データのリアルタイム処理に適したメモリDB製品「InfoFrame Table Access Method」について紹介します。

キーワード

●メモリDB ●リアルタイム処理 ●高速並列データ処理 ●大量データ

1. まえがき

昨今、センサデバイスの増加やあらゆる情報の電子化によってデータが爆発的に増加しています。そして、これらの大量に発生するデータをリアルタイムに処理し、ビジネスに活用したいというニーズが高まっています。

本稿では、こうしたビッグデータ時代において、従来にない新しい価値を創造することができるメモリDB製品「InfoFrame Table Access Method」（以下、TAM）を紹介します。

2. 製品コンセプト

メモリDB製品TAMの製品コンセプトは、「大量データの高速・リアルタイム処理」です。

このコンセプトを実現するためには、個々のデータ処理性能を高速化することに加え、大量データを並列処理させてもスループット性能を低下させないことが重要となります。

TAMは、高速性と高スループットを実現する高速並列データ処理をキーワードに製品化を行っています。

3. 製品概要

3.1 特長と製品構成

TAMは次のような特長を持ち、システムの高い可用性と高トラフィック状況下のリアルタイム処理を実現しています。

(1) 高速な検索・更新処理

データアクセスに必要な情報をメモリ上に保持することによって、高速なデータアクセスを実現しています。

(2) 優れた同時実行性

一般に、並列処理におけるスレッドの同時実行性を阻害する要因となるのは、リソースへのアクセスの競合です。このリソース競合を抑制することで、並列処理の多重度（スレッド数）を増やしてスループット性能向上を可能にしています。

(3) 高速復旧と多重障害対策を組み合わせた復旧機能

障害発生時の高速復旧を可能とするレプリケーション機能と、復旧に必要な情報をストレージ上に保持するジャーナル機能を持つことにより、多重障害時のデータ消失リスクを解消しています。

TAMでは、メモリ上に保持している表データを「メモリテーブル」と呼びます。TAM自体は、**図1** のようにデータアクセスの中核機能である「テーブルアクセス機構」と「ユーティリティ」、「運用支援ツール」で構成されています。このうち、テーブルアクセス機構は、前述したレプリケーション機能とジャーナル機能の他、メモリテーブルアクセスのためのAPI（Application Program Interface）を含んでいます。またユーティリティは、生成や保存といったメモリテーブルのメンテナンス、レプリケーション機能やジャーナル機能に対するオペレーションなど、システム運用で使用するツール群です。そして運用支援ツールは、データ量の増加傾向といったトレンドを把握し、将来のハードウェア増強（CPUやメモリ

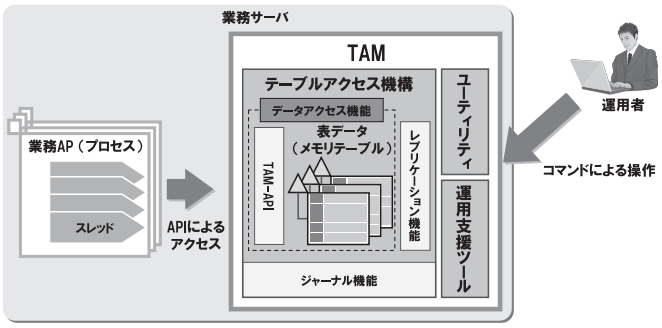

図1　InfoFrame Table Access Method製品構成

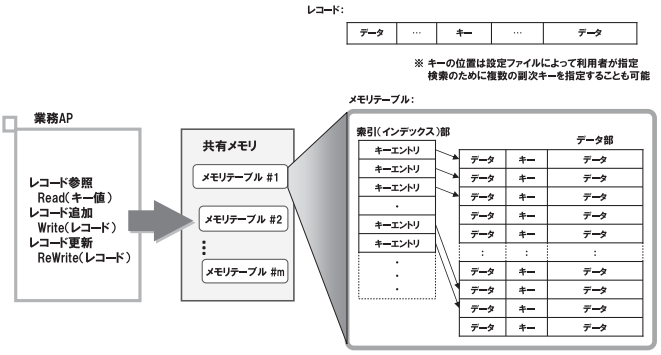

図2　メモリテーブル構造

増設など）の判断材料とするための稼働統計情報を採取する機能や、導入初期の性能分析のために使用するトレース機能などを持っています。

次節では、TAMの特長となり、中核機能でもあるテーブルアクセス機構の概要について説明します。

3.2 データアクセス機能

大量データをリアルタイムに処理するためには、次の2つの点が重要となります。

まず1つ目は、個々のデータを処理するのに掛かる時間が極めて短時間でなくてはならないという点です（高速処理）。そして2つ目は、単一のサーバが扱う処理データの量が急増しても、システムアーキテクチャを大幅に変更することなく、CPU増強などの容易な方法で対応できなくてはならないという点です（データ量増加に対するスループット性能の維持）。

大量のデータ処理において、スループット性能を向上するためには、処理の多重度を上げる（並列度を高める）という手法を採ることが一般的です。しかし、多重度を上げただけでは処理の並列度に限界が生じます。それは、リソース競合が多く発生してしまうような処理方式であることが原因です。このような処理方式では、リソース競合を調停するためのオーバーヘッド（排他制御自体の処理コスト、排他取得待ち/排他取得に伴うコンテキストスイッチのコスト）が大きくなっているために、多重度を上げてもスループット性能が上がらないのです。

このため、大量に発生するデータを高速に処理し、データ量の増加にあわせて処理の並列度を上げることで対応できるようにするには、データアクセス自体に掛かる処理性能を高速化するだけではなく、発生したデータを並列に処理するうえで、その並列動作がリソース競合によって阻害されないように考慮することが重要となります。

TAMでは、高速なデータアクセス機能を提供しており、更にリソース競合によるオーバーヘッド発生を抑制するデータ構造やアクセス処理方式を採用しています。

(1) データアクセスの高速化

TAMのデータアクセスは、レコードの指定された部位をキーとし、キーを使って探索したレコードにアクセスするという、シンプルな処理になっています（**図2**）。すなわち、キーと値（バリュー）を対応づけて格納、管理するKVS（キーバリューストア）型のデータ管理方式とみることができます。

アクセスに必要な情報（索引部、データ部）は、メモリテーブルとしてメモリ上に配置します。そのメモリ上のデータへ、APIを介してダイレクトにアクセスすることによって、データアクセスを高速化しています。

(2) 排他制御のオーバーヘッドの抑制

TAMでは、優れた同時実行性という特長を最も引き出すための処理モデルを採用しています。それは、1つの表データを1つのスレッドのみが更新するという処理モデルです（**図3**）。TAMは、APIやロック区間、ロック粒度を本モデルに最適化することで、更新対象リソースへのアクセス競合を排除し、排他制御のオーバーヘッドを抑制しています。

なお、1つの表データを更新するスレッドは1つですが、参照スレッドについては1つに制約されません。参照を主体とする表データに対しては、複数スレッドから表

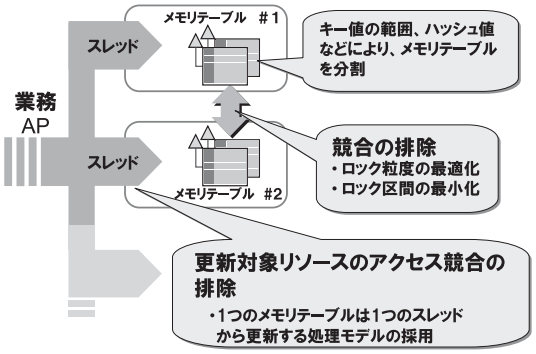

図3　処理モデル

データを参照するための機能を提供しています（共有アクセス機能）。

TAMは、前述したデータ構造、アクセス処理方式となっています。これにより、処理すべきデータ量の増加に合わせて、CPU増設や表データの分割を行い、処理の並列度を向上させて、スループット性能の向上を実現しています。

3.3 復旧機能

TAMには複数のデータ復旧機能があります。障害によるデータ消失のリスクに備え、ネットワークを介して別サーバに表データを複製するレプリケーション機能と、ストレージに保存した更新ログとメモリテーブルイメージからメモリテーブルを復旧するジャーナル機能です。

それぞれの復旧機能について、システムの可用性、信頼性の面から説明します。

(1) レプリケーション機能の特長

レプリケーション機能は、複数のサーバ間で、メモリテーブルの逐次同期を行う機能です（**図4**）。

業務APは、業務サーバ上にある更新対象のメモリテーブル（マスターメモリテーブル）に対して、データの更新を行います。レプリケーション機能は、そのデータ更新内容（更新ログ）を他サーバ上にネットワークを介して転送し、そこにあるレプリカメモリテーブル（複製されたメモリテーブル）に随時反映します。なお、このとき、転送元のマスターメモリテーブルを保持するサーバをマスターサーバ、転送先でレプリカメモリテーブルを保持するサーバをスレーブサーバと呼びます。メモリテーブルの内容は、業務APがコミットしたタイミングでレプリカ側と同期します。業務APによるコミットの都度、内容が同期されるため、マスターとレプリカとの間のデータ同期ずれにより、マスターサーバ障害時にコミットされたデータが消失することを防止しています。

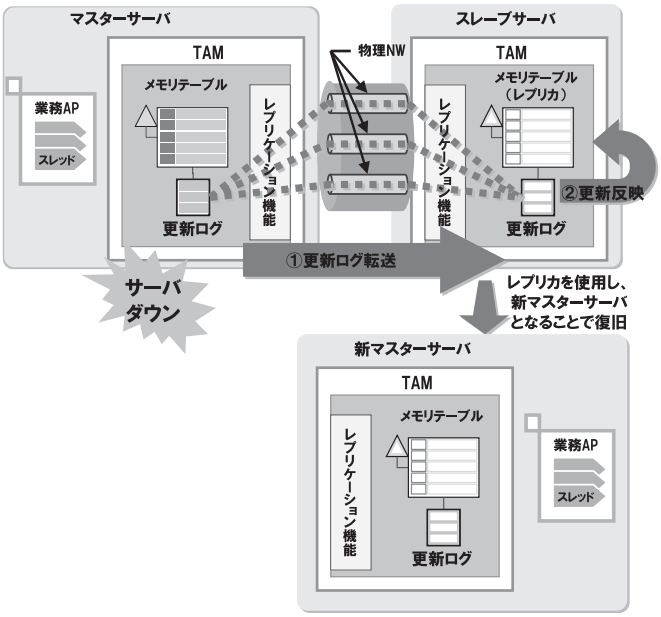

図4　レプリケーション機能による復旧

マスターサーバが障害となった場合は、レプリカのメモリテーブルを利用し、スレーブサーバを新たなマスターサーバとすることで障害復旧を行うことができます。

レプリケーション機能を利用した復旧では、メモリテーブルをコミットのタイミングで常に同期しているため、メモリテーブルのデータ内容は、未確定状態のデータを確定または廃棄するだけで済みます。そのため、約1秒以内[*1]で復旧することができ、システムの可用性を高めることができます。

また、データのレプリケーション先を複数としたり、レプリケーションで使用する回線を複数としたりすること

[*1] 環境や構成、利用方法によって復旧時間は異なります。任意の条件で、記載した復旧時間を保証するものではありません。

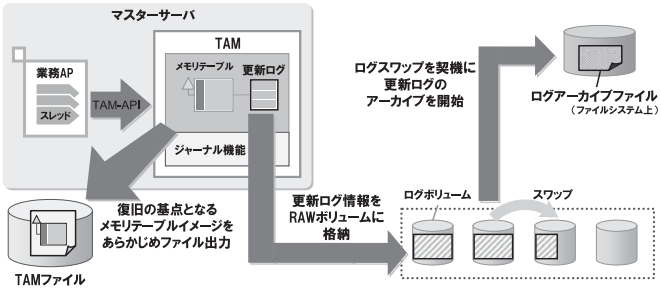

図5　ジャーナル機能による更新ログ情報の保存

もでき、レプリケーション機能の冗長性を高めることもできます。

(2) ジャーナル機能の特長

ジャーナル機能では、データ復旧に必要な情報をストレージに保持します。ジャーナル機能で必要となる情報は、更新ログとジャーナル復旧の基点となるメモリテーブルのイメージファイル（TAMファイル）となります。更新ログ情報は、ファイルシステムを経由するオーバーヘッドを抑制するため、RAWボリュームにサイクリックに書き出します。書き出された情報は、書き出し先のボリュームに空きがなくなり次のボリュームに切り替わった契機に、TAMのアーカイブ制御プロセスによって自動的にログアーカイブファイルとしてファイル出力されます（**図5**）。

TAMファイルは、TAMの運用コマンドにより生成されます。TAMファイルとして保存されるメモリテーブルイメージは、コマンド発行時点でのメモリテーブル内容となります。

また、サーバ多重障害などにより、すべてのサーバのメモリテーブルが消失してしまった場合、生成していたTAMファイルをメモリにロードし、そのメモリテーブルに対してアーカイブファイル内の更新ログ情報を適用（ロールフォワード）することで消失したデータを復旧します（**図6**）。

TAMのジャーナル機能では、TAMファイルやログボリューム、アーカイブファイルの出力先を二重化することもでき、多重障害などによるデータ消失リスクを、レプリケーション機能だけを使用する場合に比べて低減することができ、より信頼性を高めることができます。

TAMでは、システムに求められる要件（可用性、信頼性）に応じて、これらの復旧機能を選択、または併用することができます。

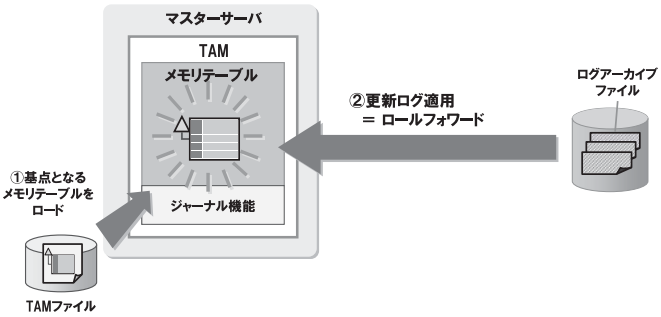

図6　ジャーナル機能による復旧

4. おわりに

メモリDB製品TAMは、キャリア系の大量イベント処理をするシステムや金融系のリアルタイム配信システムなどの基盤製品としての実績を持ち、高信頼、高可用性を備え、大量データのリアルタイム処理を実現できる製品となっています。

今後も、ビッグデータ時代の増加していくデータやこれまでリアルタイムで活用できていなかったデータを利用し、新しい価値やビジネスを創出する要求は、日々高まっていくものと考えています。

これからも継続してお客様の声を取り入れていくとともに、更なる技術強化に努め、より満足いただける製品を提供すべく取り組んでいく所存です。

執筆者プロフィール

大沢 英紀
ITソフトウェア事業本部
第三ITソフトウェア事業部
マネージャー

宮田 剛
ITソフトウェア事業本部
第三ITソフトウェア事業部
主任

関連URL

InfoFrame Table Access Method製品情報:
http://www.nec.co.jp/tam/

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報(日本語)　　NEC Technical Journal(英語)

Vol.65 No.2　ビッグデータ活用を支える基盤技術・ソリューション特集

ビッグデータ活用を支える基盤技術・ソリューション特集によせて
ビッグデータを価値に変えるNECのITインフラ

◇ 特集論文

データ管理 / 処理基盤

超高速データ分析プラットフォーム「InfoFrame DWH Appliance」
SDN技術で通信フローを制御する「UNIVERGE PFシリーズ」
大量データをリアルタイムに処理する「InfoFrame Table Access Method」
大量データを高速に処理する「InfoFrame DataBooster」
ビッグデータの活用に最適なスケールアウト型新データベース「InfoFrame Relational Store」
高い信頼性と拡張性を実現したExpress5800/スケーラブルHAサーバ
大規模データ処理に対するOSS Hadoopの活用
大容量・高信頼グリッドストレージiStorage HSシリーズ(HYDRAstor)

データ分析基盤

ファイルサーバのデータ整理・活用を支援する「Information Assessment System」
超大規模バイオメトリック認証システムとその実現
WebSAMの分析技術と応用例～インバリアント分析の特長と適用領域～

データ収集基盤

スマートな社会を実現するM2Mとビッグデータ
微小な振動を検知する超高感度振動センサ技術開発とその応用

ビッグデータ処理を支える先進技術

多次元範囲検索を可能とするキーバリューストア「MD-HBase」
高倍率・高精細を実現する事例ベースの学習型超解像方式
ビッグデータ活用のためのテキスト分析技術
ビッグデータ時代の最先端データマイニング
ジオタグ付きデータをクラウドでスケーラブルに処理するジオフェンシングシステム
柔軟性と高性能を備えたビッグデータ・ストリーム分析プラットフォーム「Blockmon」とその使用事例

◇ 普通論文

地デジTVを活用した「まちづくりコミュニティ形成支援システム」

◇ NEC Information

NEWS

スケールアウト型新データベース「InfoFrame Relational Store」が2つの賞を受賞

Vol.65 No.2
(2012年9月)

特集TOP