

## リアルタイムシステムにおけるデータ管理の永続性実現

これまでに、データの高い信頼性、永続性を達成するためにインメモリのデータ記録方式が発表されました。この文書ではこれまでとは異なるアプローチでデータの永続性を満たす方法について検討します。

メインメモリのデータ管理システムは、リアルタイムの世界で必要とされる高性能と予測可能なレスポンスが提供できるので、組み込みの世界では広く使用されています。これまでに、データの高い可用性、永続性を達成するためにインメモリのデータ記録方式が発表されました。その選択肢の中には、NVRAMデータベースのサポートやオンラインバックアップ、トランザクションログ、データベースの複製などが含まれます。組み込みシステムのエンジニアは、これらの手段をうまく使って、伝統的なディスクベースの製品を使う場合と同程度のデータ保護を獲得しました。同じくらいに重要なことは、これらのアプローチは開発者が永続性のレベルと彼らが望むスループットやアプリケーションの性能を調整することが可能になることです。組み込みシステムのエンジニアはアプリケーションの特定のニーズに基づいて適切な選択ができます。

近年、リアルタイムの組み込みシステムは急激に増加しています。それと同時に拡張された特長や能力に対する要求の結果として、より複雑な構造を持ったより多くのデータを管理しなければならなくなりました。開発者は、組み込みデバイスにデータを保存するためのデータベース管理システムの必要性に気付いています。開発時間の短縮と機能や信頼性の向上を行うために、開発者が内製のデータ保持ソフトウェアから商業的に入手可能な既製のデータベースソフトウェアへ切り替えることが著しく増えてきています。

リアルタイム処理の性質を考えたとき、データベースへの要求はユニークです。リアルタイムデータベースは、時間的な制限とデータの論理的な一貫性を維持しつつトランザクションを行います。リアルタイムの環境下では、デスクトップやエンタープライズアプリケーションとは対照的に、データ処理の遅延はしばしば間違ったり危険を伴います。従って、データベースがアプリケーションの制限内で十分密接に存在し、状態を表すことは極めて重要です。例えば、原子力発電所において、自動化されたコントローラに組み込まれたデータベースは、プラントの至る所の数百ものセンサから受け取られたデータを集めて処理します。もし、データ（外の放射能レベルやリアクタ温度等）が決められた時間制限内に処理されないならば、悲惨な結果を招くでしょう。更に、予測可能性と高性能に加えて組み込みシステムのデータ保持機能は、人的な介入を必要としないこともあげられます。自動的に故障から回復しデータアクセスを継続できることが要求されます。

これらのシステムのリアルタイムパフォーマンスと予測可能性の要求は、インメモリデータ

ベースシステムの採用を決定付けます。それは、データベース自身が完全にメモリ上で実行され決してディスクにアクセスしないからです。機械的なディスクアクセスを完全に排除することは、オーバーヘッドの原因を取り除き、キャッシングの様なディスクアクセスに関係するロジックも不要です。インメモリ実行は、CPUサイクルをより効率よく利用できる最新式のアーキテクチャです。ディスクI/Oのための待ちを回避することにより、インメモリデータベースはスピードを要求されるトランザクションの応答時間を予測・保証することができます。しかしながら、完全にRAM上で動作するデータベースは性質上壊れ易いと言えます。もし、RAMの内容が破壊されたら、データベースも破壊されます。では、どのようにしてメモリーのみで動作するデータベースの永続性を上げたり、ソフトウェアが動作する環境にあるハードウェアデバイスの故障から生き残る能力を与えることができるでしょうか？

要求された永続性を満たすために、インメモリデータベースシステムはいくつかの解決法を提供します。インメモリデータベースでは、しばしばデータのコピーを保持する方法を提供し、それによりRAMの内容が失われても、データアクセスやデータそれ自体の消滅を意味しなくなります。このデータベースの複製と言われる方法では、処理が失敗した場合でもスタンバイデータベースを使用し続けることを許してしまいます。他の方法では、システム障害が発生した場合にデータの回復を助ける不揮発性の記憶装置を使う必要があります。同期・非同期の複製を含む、これらの異なるアプローチはトランザクションログや不揮発性の記憶装置を用いる場合、それぞれのデバイスは固有の性能とリソース活用方法を持っているので、開発者はそれらをよく理解し、リアルタイムのデータ管理をデザインする場合に考慮しなければなりません。

現実的なアプリケーションで利用される基本的な概念では、データ管理はトランザクションと呼ばれる処理の単位によって運用されるということです。トランザクションとはデータの一連のリード・ライト操作で、ACID特性(atomicity:原子性, consistency:一貫性, isolation:独立性, durability:永続性)と呼ばれるものに対応しなければなりません。特にトランザクションの永続性は、いったんトランザクションが問題なくコミットされたらデータベースへの全ての変更は恒久的で引き続き起こるシステム障害や誤動作から守られねばなりません。このようにデータベースの永続性について論じることは、トランザクションの永続性とデータベース管理システムの維持方法について言及することになります。

データの複製を維持することはデータベースのトランザクション永続性を達成するのに効果的です。複製されたデータベースは故障から独立したノードから成り立っています。データベースシステムは、あるノードで故障が起きたときでもデータを喪失していないことを確認できるように、複数のノードに分散しているデータを管理します。複製されたデータベースを管理する伝統的な手法では、データの更新情報をレプリカへ伝える方針により同期・非同期と分類されます。同期レプリカでは、全てのレプリカはオリジナルのトランザクションの一部として全てのレプリカにアップデートを適用します。トランザクションの永続性を保証しているので、レプリカがノード障害の場合に処理を奪取するケースでは、同期レプリカは資源の拘束時間が長くなり、その結果ネットワーク遅延のためにトランザクションのスループットが落ちてしまうことがあります。

一方、非同期レプリカ方式では、増え続けるレプリカのアップデートを待たずにトランザクションのコミットを行うので、資源を拘束する時間はより短くなります。しかし、非同期レプリカ方式では、アップデートがレプリカに届く前に、もとのトランザクションサイトで故障が

発生するという潜在的なデータ喪失のリスクがあります。

一般的に言えば、非同期レプリカアルゴリズムでは、トランザクションの永続性は高いとはいえません。アプリケーションが必要とする予測可能な応答時間を満たすによりよい選択肢としては、時間を認識する機能を追加した同期レプリカ方式を採用したデータ管理と言えます。組込みシステムは頻繁に非常に厳しい制限が設けられますが、このアプローチではレプリカノードとのトランザクションデータのオンタイムデリバリーを保証できます。

組込みシステムの中にハードディスクやFLASHカードの様な不揮発性のメディアが存在する場合、このメディア上に永続的なインメモリデータベースを作ることができます。もし、システムが停止して再開したときに、DBMSで供給される回復手順に従ってこれらのディスクからデータベースの内容を復元することができます。これらはロールバックとロールフォワードと言われる基本的な2つの方法で行われます。どちらのアプローチにせよ、通常運転時にチェックポイントと呼ばれるインメモリデータベースの定期的なスナップショットがとられ、不揮発性のメディアに格納されます。システム再開時のロールバックリカバリでは、最後のチェックポイントが単純にメモリにロードされます。最後のチェックポイント以降の全てのデータベース変更は失われます。ロールフォワードリカバリでは、チェックポイントは同様に規則正しく格納されますが、チェックポイント間の全てのトランザクションはトランザクションログに書き込まれます。回復時には最後のチェックポイントに対して、ログに入力されたのと同じように、古いものからログが適用されます。

このデータベース情報の永続的な記憶装置に対する書き込みは、IMDBがなくそうしているディスクアクセスのオーバーヘッドを導入しないのでしょうか？例えばMcObjectのeXtremeDBのような、商業的に利用可能なほとんどのメインメモリデータベースのトランザクションログの機能は、トランザクションの永続性のレベルを設定でき、システム設計者がパフォーマンスとトランザクション喪失のリスクの間でトレードオフを決めることが可能です。例えば、トランザクションログを同期または非同期のいずれかで永続的な記憶装置に書き込むことができます。同期のアプローチでは、トランザクションデータはトランザクション内でトランザクションデータをディスクに書きます。このケースでは、全てのトランザクションが記録されますが、このプロセスの間はデータベースがロックされ、リアルタイムプロセスでは予知できない遅延を引き起こす可能性があります。非同期のログ方式では、ログデータの書き込みはシステム全体の負荷が低い時やアプリケーションが決めたり、メインアプリケーションのプロセスに影響を及ぼすパフォーマンスの減少を低くするために個別のプロセスで行われるように設定されます。非同期のログ方式では、トランザクションの永続性は緩和されますが、データベース全体の版の反応性はより高くなります。

一般にトランザクションログはIMDBのインメモリアーキテクチャを変質させません。そして、ディスクベースのデータベースに対するハイパフォーマンスのアドバンテージはなら変わりはありません。IMDBのリードパフォーマンスはトランザクションログに影響されませんし、ライトパフォーマンスは伝統的なディスクベースを遥かに凌駕します。理由は簡単です。トランザクションログは、一つのトランザクションに対して、ファイルシステムに対する書き込みはただ一度だけです。ディスクベースのデータベースは一つのトランザクションに対して、データページ、インデックスページ、トランザクションログなどの沢山の書き込み動作を行います。より大きなトランザクションとより多くのインデックスが修正された場合は、もっと多くの書き込みが必要となります。

今日、コンピュータが停止したり外部電源がなくなった時でも内容を保持できる不揮発性メモリのNVRAMを使用することで、多くの組み込みデバイスが特長をつけています。NVRAMは通常SRAMとバックアップ用電池か、保持したデータを電氣的に消去できるEEPROMで形成されています。もし、NVRAMデータベースの保持に使われるとしたら、データベース管理システムはリブートすることで最後の一貫した状態で回復することができます。このアプローチは、インメモリデータベースにとって非常に魅力的な永続性のオプションです。オーバーヘッドに関連するディスクI/Oを含んだトランザクションログや、通信のオーバーヘッドがある複製のアプローチとは対照的に、NVRAMデータベースはオペレーション中のディスクやネットワークのオーバーヘッドがありません。商用データベースでは、NVRAMへのデータ保持を直接サポートするものはほとんどありません。これは主に、データベースを保持する大容量のNVRAMは非常に高価であることが原因です。いくつかのケースでは、費用は正当化されます。例えばハイエンドルータはデータベースの構成をNVRAMボードに保持していますし、ネットワークストレージデバイスはNVRAM上でいくらかのアーキテクチャが実現されています。そのようなアプリケーションでは、NVRAMデータベースは実に貴重なパフォーマンスのアドバンテージとデータベースの永続性オプションにより増加したNVRAMの費用を調整しています。