計算機アーキテクチャ　90－9
（１９９１．１０．７）

高可用性の商用ＤＢＭＳ

小島國照
日本タンデムコンピューターズ

ＮｏｎＳｔｏｐ-ＳＱＬは、疏結合マルチプロセッサーアーキテクチャーのハードウェアとメッセージベースのＯＳ（ＧＵＡＲＤＩＡＮ）の上に構築された分散ＤＢＭＳである。リリース２では特に並列処理とより高可用性をサポートする機能が追加されている。並列処理の機能としては、水平分割されたテーブルに対するＪＯＩＮ、アグリゲート評価、ＩＮＳＥＲＴ、ＵＰＤＡＴＥの並列化、ＩＮＤＥＸ保守の並列化がある。また、可用性を上げる機能として、稼働中にテーブル、インデックスの再編成、再構成、およびリプリケートテーブルや災害時のバックアップシステムを可能とする遠隔複製ＤＢ機能（ＲＤＦ）などがある。

CASE STUDY OF A HIGHLY AVAILABLE COMMERCIAL DBMS

Kuniteru Kojima

TANDEM Computers Japan, Ltd.
4-3 Kojimachi, Chiyoda-ku, Tokyo 102 JAPAN

NonStop-SQL is a distributed DBMS based on loosely coupled multi-processor architecture hardware and message based operating system GUARDIAN. In its release 2, several new enhancements are implemented to the product which increases parallelism and availability of the system. Join, insert and update operations, aggregate evaluation against horizontally fragmented tables and index maintenace can be processed concurrently. High availability of the system is facilitated by online reorg and reconf of tables and indices and the Remote Duplicate Database Facility.

1. まえがき

ＮｏｎＳｔｏｐ　ＳＱＬ　リリース1はＴａｎｄｅｍ　ＮｏｎＳｔｏｐシステム上にＡＮＳＩ　ＳＱＬ標準を完全分散ＤＢＭＳとして製品化された。これはリモート、ローカルのデータへのアクセスを可能とするだけではなく、分散ネットワーク上のデータのアップデートに関してトランザクション保護の機能を実装している。さらに、ＳＱＬをファイルシステム、ディスクプロセス、Ｔａｎｄｅｍ　Ｇｕａｒｄｉａｎ　９０　ｏｐｅｒａｔｉｎｇ　ｓｙｓｔｅｍに一体的に組み入れることにより、現実のオンライントランザクションシステムの使用に耐えうるパフォーマンスを実現した。リリース２におけるＮｏｎＳｔｏｐ　ＳＱＬの機能拡張は以下の３点に重点がおかれた。

A．パフォーマンス

B．操作性と管理性

C．ＡＮＳＩ　ＳＱＬへの準拠の強化

ここでは、A.、B.のなかで並列処理とオンラインＲＥＯＲＧなどの可用性機能について概説する。

2．ＮｏｎＳｔｏｐシステムの概要

ＮｏｎＳｔｏｐシステムは、それぞれ独立したメモリーを持った最大１６のプロセッサーから構成されます。透過なネットワークＯＳの機能を通してＬＡＮあるいはＷＡＮ接続で４０００以上のプロセッサーを接続できる。この場合、全プロセッサーが単一システムであるように機能する。

ディスクボリュームはディスクユニットのミラードペアで構成されている。各ディスクユニットはディスクコントローラーのペアに接続され、それぞれのディスクコントローラーはプロセッサーのペアに接続されている。したがって、システムは各ディスクユニットに対し4つの物理的パスを通してアクセスできる。それぞれのディスクボリュームは隔離されたプロセッサー内にあるディスクプロセスペアによって制御される。しかし、ある一時点ではプロセスペアの片側だけがプライマリーとしてミラードディスクボリュームの双方のユニットを制御する。

ＮｏｎＳｔｏｐ-ＳＱＬにおいてはテーブルやインデックスは最大１００のディスクボリュームにパーティションされる。パーティションされたテーブルやインデックスでは複数のディスクボリュームに、キー範囲（キー順ファイルの場合）または、パーティションのサイズ（順編成、相対アドレスファイルの場合）によって物理的に（ネットワーク上も含めて）分散している。つまり、水平分割することが可能である。ＮｏｎＳｔｏｐ-ＳＱＬにおける並列処理はそのシステム構成に依存している。理想的にはプロセッサー、コントローラー、プロセッサーがパーティションされたテーブルやインデックスへの同時アクセスを可能とするようにそれぞれがバランスされることが望ましいが、そのような理想的な状況でなくとも並列処理はある程度の効果をもたらす。

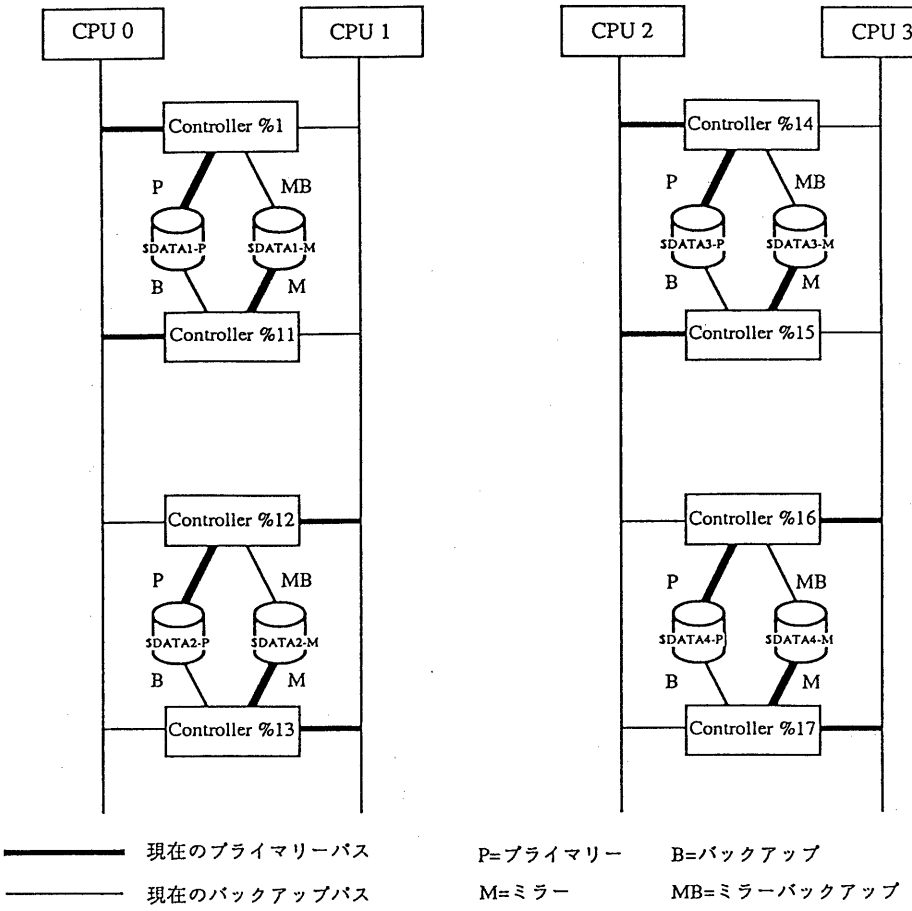

図 1

3. ＮｏｎＳｔｏｐ-ＳＱＬの構成

ＳＱＬステートメントの並列処理を可能とするため、リリース２からイグゼキューターサーバープロセス（ＥＳＰ）とプロセスが導入された。マスターイグゼキューターはＥＳＰの起動、管理を受け持ち、ＥＳＰとの通信によって各ＥＳＰがステートメントの一部を並行に処理することを可能とする。

オプティマイザーにとって並列処理はひとつのオプションであって、それが最適であるときのみ並列処理をプランとして選択する。テーブルや、そのインデックスをパーティション化することは一般的には並列処理を選択しやすい条件をつくる。パーティションは一つのシステム上でもネットワーク上の複数システム上にでも存在できる。

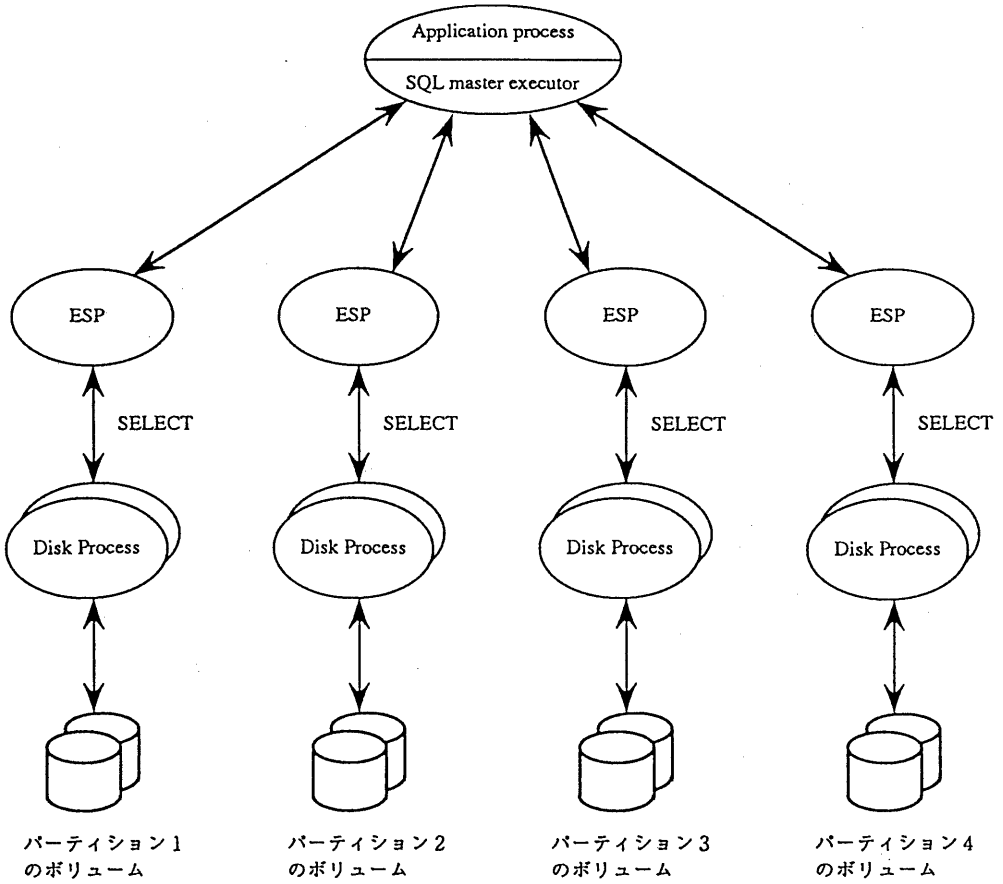

図2

オプティマイザーが並列処理を最適プランと判断した場合、マスターイグセキューターは、アクセスプランによりアクセスされるべき各パーティションに対して一つのＥＳＰを割り当てる。実行時に、すでにＥＳＰが存在しない場合には各パーティションのプライマリーのディスクプロセスのあるプロセッサーにＥＳＰを起動する。各ＥＳＰは割り当てられたパーティションにのみアクセスをおこなう。

通常、ユーザはテーブルをパーティションすることによって並列処理の恩恵に浴することができる。しかし、しばしば、テーブルが直接並列処理を行なうには適していない場合もある。このような場合、オプティマイザーはデータを複製することによって再パーティションし、その再パーティションされたテーブルのうえで並列処理ができるかどうかを検討対象とする。図３は、このような場合の再パーティシ

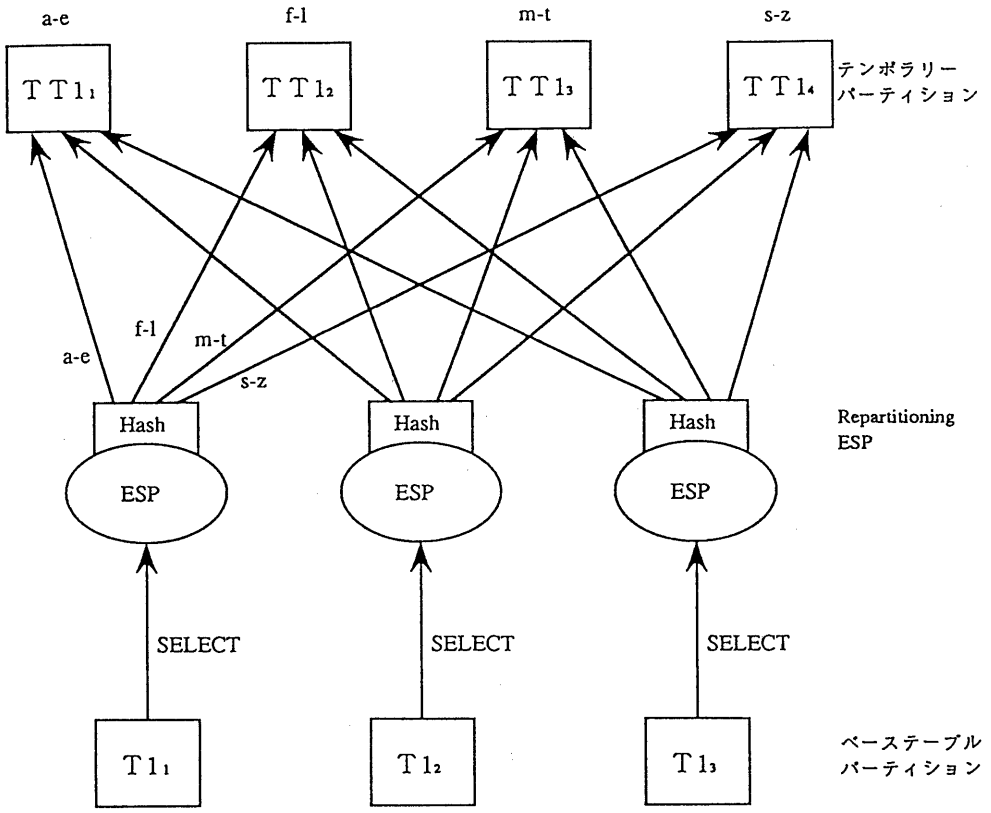

図3

ョンの処理のされかたを図式化したものである。実行時に、マスターイグゼキューターはＥＳＰを起動し、プレディケイトに応じてレコードの取り込みを開始させる。取り込まれたデータは臨時テーブルが均等に分配されるように作成されたハッシュ関数によって割り振られる。必要に応じてＳＯＲＴプログラムが起動されることもある。これらのコストは最適プランの選択時の考慮対象となる。

4．並行処理機能

ユーザはＳＱＬオプティマイザーが並列処理を選択するかどうかをコンパイル時に指定できる。ユーザが指定した場合でオプティマイザーが最適プランと判断したときのみ、並列処理が選択される。ただし、インデックス保守の並列処理は自動的に行なわれる。

4-1. 並列アグリゲート評価

ＣＯＵＮＴ、ＳＵＭ、ＡＶＧなどのアグリゲート関数はサマリーリポートなどで頻繁に使用される。並列ＥＳＰは各パーティションにたいしてアグリゲート関数の評価を実行する。結果はマスターイグゼキューターに戻され、集計されてアプリケーションに渡される。ＮｏｎＳｔｏｐ　ＳＱＬはパーティション化されていないテーブルにはアグリゲート関数評価の並列処理を行なわない。しかし、ＧＲＯＵＰ　ＢＹクローズがあってアグリゲート関数が各グループにたいしておこなわれる時にはリパーティションを行なって並列処理をおこなう。

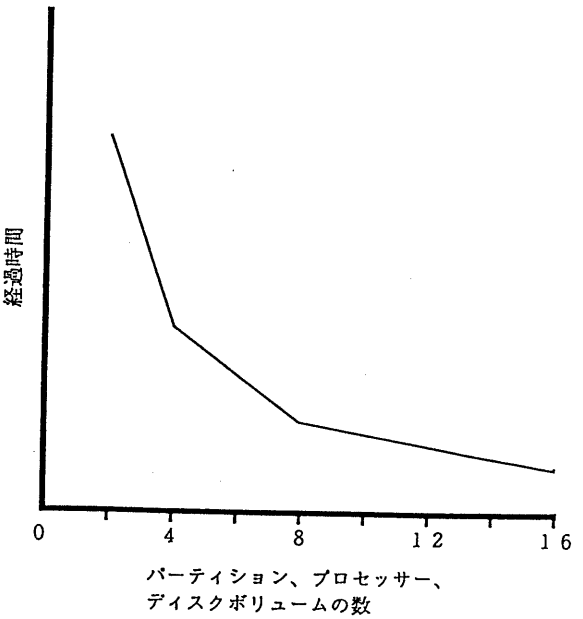

図４

4-2. 並列ＵＰＤＡＴＥ、ＤＥＬＥＴＥ

ＵＰＤＡＴＥ、ＤＥＬＥＴＥに関して最適プランは実際に複数のディスクプロセスへアクセスされるかどうかの評価の結果を反映する。テーブルがパーティション化されていない場合でも、インデックスがパーティションされており、複数のインデックスパーティションへのアクセスを必要とする場合は実行コストのよっては並列処理を選択する。さらに並列処理がプランの選択の対象となった場合は、ＥＳＰを起動するためのコストも考慮される。通常、コミュニケーションコストは無視できる。図４は１.６ギガバイトのテーブルを２、４、８、１６プロセッサー（ＶＬＸプロセッサー）に分割した場合の測定結果である。これにより、適切にシステムが構成されている場合、並列処理は経過時間の短縮に大きな効果を持つことがしめされている。

4-3. 並列Ｊｏｉｎ

Ｅｑｕｉｊｏｉｎについて以下の項目を考慮にいれて最適プランは作成される。

＊　すべてのテーブルがＪｏｉｎのカラムについて等しくパーティションされているか。

＊　一つのテーブルがパーティションされている場合、他のテーブルは比較的小さく、Ｊｏｉｎカラムにインデックスを持っているか。

＊　すべてのテーブルが等しくＪｏｉｎカラムについてパーティションされておらず、Ｊｏｉｎカラムの上にインデックスが存在しない。

これらのそれぞれのケースに応じて異なるプランが作成される。

4-4. 並列インデックス保守

リレーショナルデータベースはインデックスの存在によってはじめて、しかるべきパフォーマンスを達成することができる。また、インデックスはカラムのユニーク性を維持するためにも重大である。しかし、一方ではインデックスの保守に要するコストは大きく、インデックスを多く付け過ぎるとパフォーマンスに悪影響を与える。これはデータベースデザイナーに多くの制限を与えていた。

NonStop SQLはテーブルのUPDATEがインデックスのUPDATEを必要とする場合で、インデックスが他のディスクボリュームにある場合はインデックスへの要求は非同期に行なわれる。これにより、すべてのインデックスがそれぞれ異なるディスクボリュームに割当てられている場合、UPDATEに要する時間は1つのインデックスを持つ場合と多数のインデックスを持つ場合は実質的に同じとなる。図5は4つのインデックスを持つテーブルに10、000レコードを挿入した場合と、更新した場合の、並列インデックス処理の効果を測定したものである。

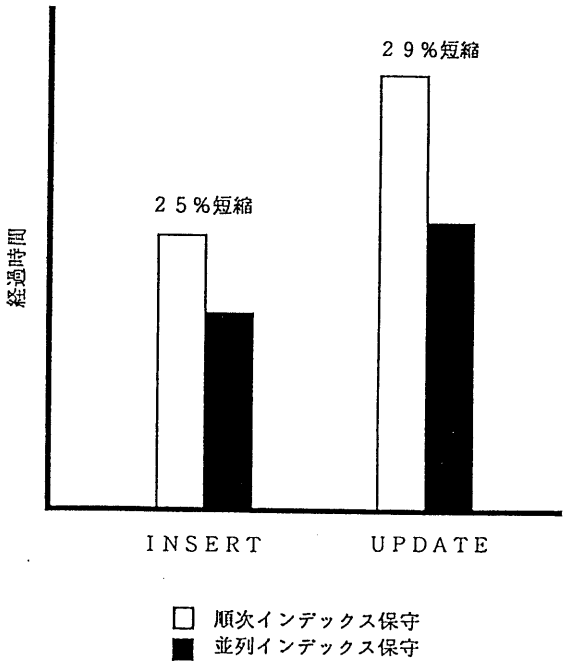

図5

5. 可用性機能

5-1. オンラインREORG

キー順編成テーブルは二つの条件を満たす時オーガナイズされていると呼ばれる。第一に、データレコードを含むブロックは物理的に連続している。第二に、インデックスブロック、データブロックは、適正な量のレコードを均等に含んでいる。しかし、時間が経つにつれて、ランダム挿入更新削除などの結果として、オーガナイズされた状態からはずれてくる。適切な空き領域がブロックないに保たれていないとレコードの挿入や更新によってブロックのスプリットが発生する。つまり、新しいブロックが作成されてデータの複製がおこなわれ、結果としてパフォーマンスに悪影響をもたらす。また、論理的に連続したレコード群にアクセスする場合、物理的にも連続していれば1回の物理I／Oで処理可能な場合

でも、連続性が失われていれば複数Ｉ／Ｏとなり、パフォーマンスは劣化する。

オンラインＲＥＯＲＧ機能は、アプリケーションを止めることなく、テーブルを再オーガナイズする機能である。アプリケーションのパフォーマンスへの影響を減らすためプライオリティをしてしたり、ＲＥＯＲＧを一時停止したり、再開することも可能である。また、ＲＥＯＲＧ中のシステムエラーに対するデータの保護も考慮されている。

5-2．リプリケーテッドテーブル

ＲＤＦ（遠隔複製データベース機能）は、アプリケーションによらずに遠隔地に複製データベースを作成する機能である。複製はトランザクションログから更新データを摘出して遠隔地のシステムに送ることによって実現されているので少ないコストで実行される。通信回線の障害時の迂回、回復は自動的である。また、複製データベースの更新は一時的に停止したり、再開することも可能であるので、この機能を使用してスナップショットデータベースとして使用することもできる。また、それぞれのシステムにプライマリー／バックアップの関係がないので、互いにバックアップシステムとなることもできる。

6．まとめ

ＮｏｎＳｔｏｐ　ＳＱＬリリース２の並列処理機能はバッチ処理、レポート生成、ＯＬＴＰ等の処理のレスポンス時間を短縮することができる。並列処理はアグリゲート、ＳＥＬＥＣＴ、ＵＰＤＡＴＥ、ＤＥＬＥＴＥ、Ｊｏｉｎの他、インデックスの保守にも応用されている。これらの並列処理のメリットを最大限に活用するためには、ディスク、パーティション、コントローラー、プロセッサーなどのシステムの各要素が適切に構成されていることが望ましい。オンラインＲＥＯＲＧの機能は、アプリケーションを停止することなくディスクスペース、ブロック割付などを最適化してパフォーマンスの劣化を防ぐことを可能とする機能でクリティカルなアプリケーションには不可欠な機能である。ＲＤＦはリプリケーテッドデータベースをアプリケーションに透過に実現する機能であって、災害対策、スナップショット上のバッチ処理を容易にし、可用性を高める役目をはたす。

[参考資料]

Cassidy, J. and Kocher, T. 1989. NonStop SQL: The Single Database Solution. Tandem Systems Review. Vol.5, No 2. Part no. 28152. Tandem Computers Incorporated.

Date, C.J. 1984. An Introduction to Database Systems. Volume 1. Addison-Wesley.

Englert, S. and Gray, J. 1990. Performance Benefits of Parallel Query Execution and Mixed Workload Support in NonStop SQL Release 2. Tandem Systems Review. Vol 6, No. 2. Part no. 46987. Tandem Computers Incorporated.

Mark Moore, Amardeep Shdhi. 1990. Parallelism in NonStop SQL Release 2. Tandem Systems Review. Vol 6, No. 2. Part no. 46987. Tandem Computers Incorporated.