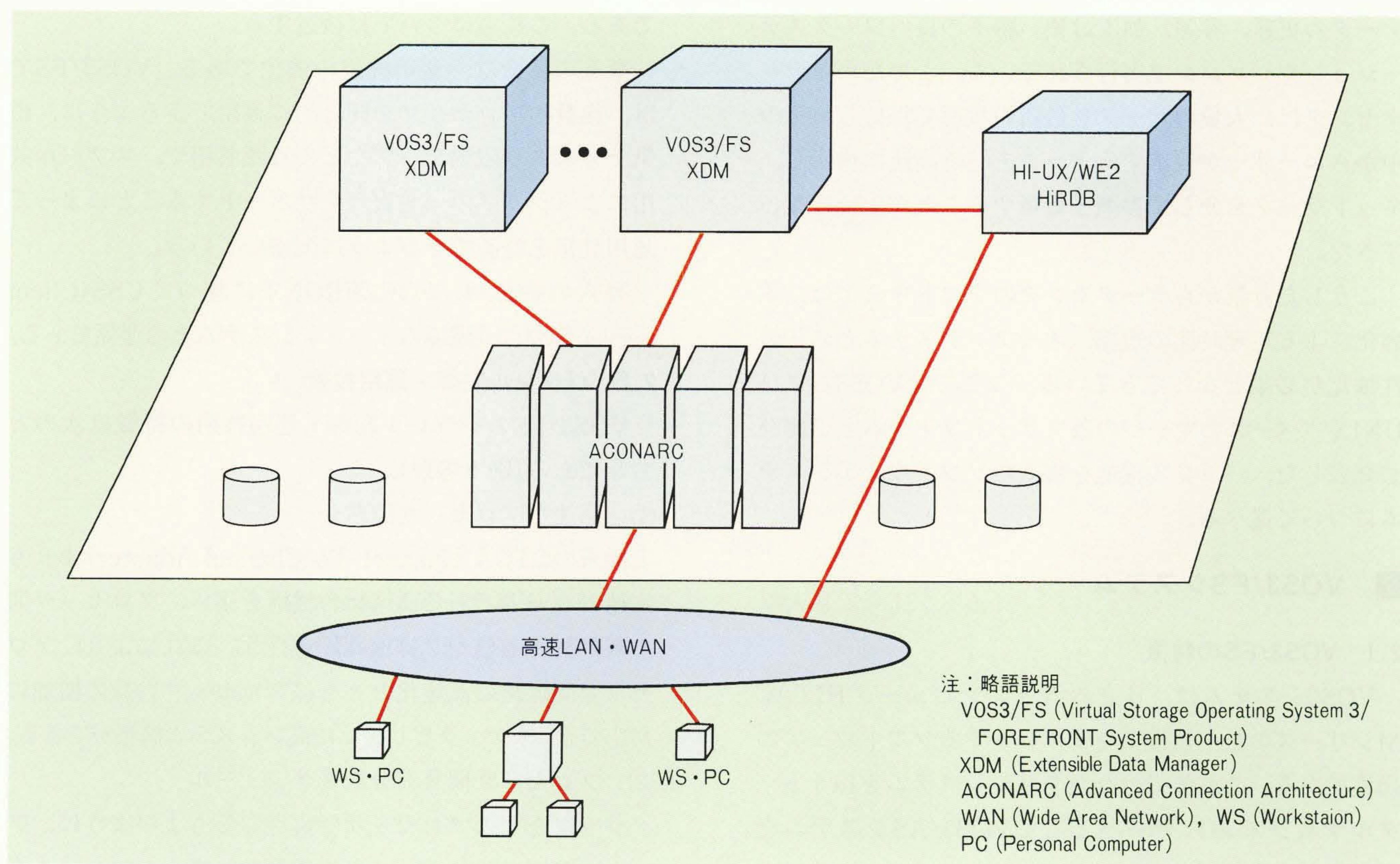


新世代システムによるデータセンタを実現する パラレルソフトウェア

Parallel Software for New Data Center

尾山 壮一* Sôichi Oyama 阿部 淳* Jun Abe
正井 一夫* Kazuo Masai



パラレルソフトウェアによるデータセンタのイメージ

VOS3/FSシステムとRISC向け並列データベース“HiRDB”を接続したデータセンタのイメージを示す。これらの基盤製品により、新たな戦略的
情報活用が実現できる。

今日、企業の基幹情報を支えるデータセンタには、大型汎(はん)用計算機やサーバに格納された大量の情報を、机上のあるいは世界の各拠点のワークステーション・パソコン(パーソナルコンピュータ)からリアルタイムに必要な部分だけを参照・更新し、ビジネスに活用できる機能が求められている。このため、データセンタとそれを取り巻くさまざまなネットワーク上のワークステーション・パソコン上のオープンなアプリケーションが連携するオープン連携機能、およびビジネスチャンスにスピーディにこたえるための高性能化機能としての並列処理機能の二

つが望まれている。

日立製作所は、“FOREFRONT with Open Environment”のコンセプトに従って、VOS3/AS(Virtual Storage Operating System 3/Advanced System)の後継でMパラレルシリーズをサポートするVOS3/FS(FOREFRONT System Product)、その上で並列データベース機能を実現するXDM(Extensible Data Manager)、およびサーバ上で並列データベースを実現するHiRDBを開発し、情報の戦略的活用を図るデータセンタシステムの実現を目指す。

* 日立製作所 ソフトウェア開発本部

1 はじめに

エンドユーザーコンピューティングが盛んになるにしたがって、データセンタの役割も変化してきた。従来、データセンタは、センタ内のアプリケーションによって業務に集中的に対応する役割を担ってきた。今日では、データの更新、参照、加工は使い勝手の良いワークステーションやパソコンで実行されている。このため、データセンタは、大量のデータを最新の状態で保持し、その中からユーザーが要求するデータを随時高速に検索し、ネットワークを介して参照・更新することが重要になってきた。

こうした背景からデータセンタのソフトウェアは、並列化による処理時間の短縮、オープンシステムとの接続性強化が必須となってきている。ここではVOS3およびUNIX^{※1)}ベースのサーバの各プラットフォーム上で開発し発表した、パラレル機能を備えたソフトウェアシステムについて述べる。

2 VOS3/FSシステム

2.1 VOS3/FSの特徴

VOS3システムは、日立の汎用コンピュータHITAC Mシリーズの最上位機種をサポートするソフトウェアの総称である。これまでは企業の基幹システムを担うトータルマネジメントサーバとしてVOS3/ASシステムを開発し、提供してきた。

今回このVOS3システムで、バッチはもとよりデータベース処理も並列化してMパラレルシステムをサポートするVOS3/FSシステムを開発し、発表した。

今日、さまざまな種類の並列システムが開発され、利用されている中であってVOS3/FSの第一の特徴は、VOS3に蓄積された膨大なユーザー資産を継承できる点である。ここで言う資産とはプログラムやデータベース、ファイルをはじめ、システム設計・性能設計のノウハウや運用のスキルといったものまでを含む。

第二の特徴はスケーラビリティである。幅広い性能レンジのプロセッサをVOS3/FSの単一のプラットフォームで利用できる。

第三の特徴は、最新のハードウェア機器のサポートで

ある。MP5800システムをはじめとする最新プロセッサをサポートする予定である。また、周辺機器もディスクアレーシステムや、磁気テープライブラリ装置をサポートし、高速・大容量なデータ保管庫としての条件を備えている。

第四の特徴は、バッチ、データベースの並列処理技術である。これらについては後述する。

第五の特徴は、運用機能の強化である。VOS3/FSでは、複数のプロセッサを統合的に運用できるように、磁気ディスクや磁気テープなどの資源共用や、スプール共用によるジョブの一元管理をサポートすることによって運用負担を軽減できるように配慮している。

第六の特徴は、FOREFRONTに基づくCSS(Client Server System)環境のオープンシステムとの関係である。

2.2 パラレル制御・運用技術

VOS3/FSのパラレル制御・運用技術の特徴は次のとおりである(図1参照)。

(1) 高速なプロセッサ間結合

従来のCTCA(Channel-To-Channel Adapter)よりも高速のハードである高速結合機構を用い、プロセッサ間の制御情報通信や、排他制御を行う。これにより、プロセッサ間連絡の高速化とともにプロセッサ台数の増加に対応でき、スケーラビリティの高いシステム構築ができる。

(2) プロセッサ間負荷分散スケジュール

ジョブをどのプロセッサで実行してもよいように、プロセッサ間で磁気ディスクや磁気テープ上のファイルを共用し、システムの負荷状態に応じたジョブスケジュールを可能としている。

(3) 並列環境での自動運転・統合運転

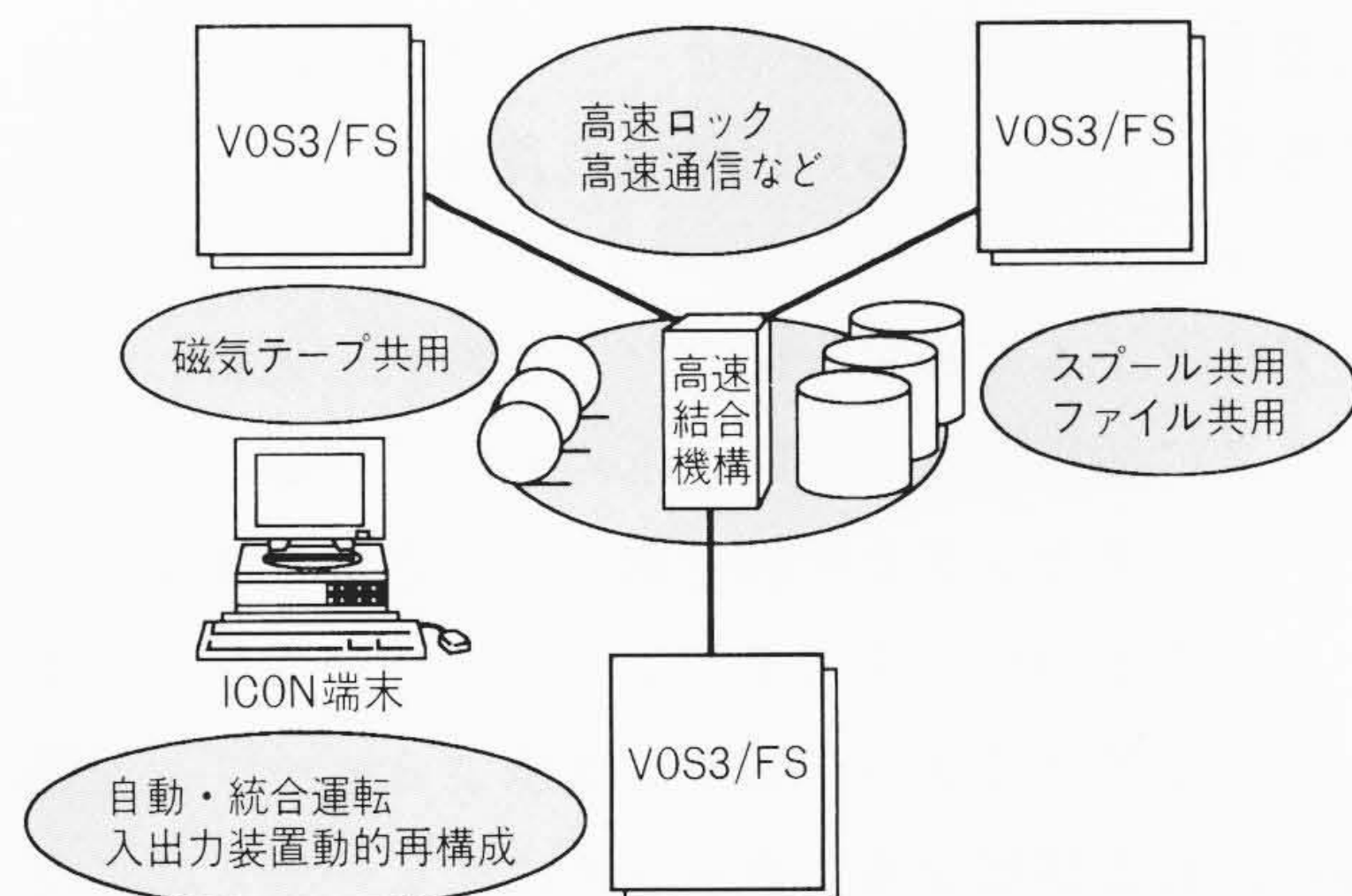


図1 VOS3/FSのパラレル制御・運用技術

高速結合機構による高速ロックや通信機能が実現され、高いスケーラビリティが期待できる。また、Mパラレルシステムの統合運転・自動運転を可能にしている。

※1) UNIXは、X/Open Company Limitedがライセンスしている米国ならびに他の国における登録商標である。

2.3 バッチパラレル機能

2.3.1 バッチパラレルの背景と概念

バッチ実行時間の長さは、今日のデータセンタが抱える課題の一つである。

VOS3/FSではバッチプログラムを改変することなく、バッチ実行時間短縮を図るバッチパラレル機能をサポートする。

バッチパラレル機能の基本原理は、ジョブステップを並列実行させ、実行時間を短縮させることである。これを可能にするのは次の二つの機能である。

(1) ジョブ間、ジョブステップ間でのデータの逐次渡し機能(PREST: Parallel Reference and Synchronous Transfer Facility)

(2) ジョブを分割し、ジョブステップを並列実行する機能

これらを適用したときに得られる効果は、バッチの処理内容やジョブネットの形態に大きく依存する。ある事例に基づく試算では、バッチ総実行時間を30~60%削減できるという結果が出ている。

2.3.2 PREST機能強化

VOS3/ASではジョブ間、ジョブステップ間でのデータの逐次渡し機能としてPRESTを提供してきた。ところが、PRESTを実際に適用する場合、効果の高いジョブ群を選定するのが困難であるという問題があった。そこでVOS3/FSでは、PREST適用対象の候補をシステム稼動履歴などから分析するツールを開発し、提供することとした。また、PRESTに以下の機能を追加した。

(1) データ多重転送機能

先行ジョブの出力をPREST内部でコピーし、複数の後続ジョブに同時転送する機能である。同一データを利用する複数の後続ジョブの同時実行が可能である。

(2) データ分割転送機能

先行ジョブの出力を分割し、複数の後続ジョブに転送する機能である。後続の長大ジョブを短縮させる効果がある。

(3) 転送データバックアップ機能

PRESTによるデータ転送中、同時にバックアップを採取することができる。

(4) システム間データ転送機能

PRESTをシステム間で可能にすることにより、ジョブの負荷分散が図れる。

データ分割転送機能の動作概要を図2に示す。

2.3.3 バッチパラレル機能

ジョブ ステップ パラレル実行機能とは、ジョブが複

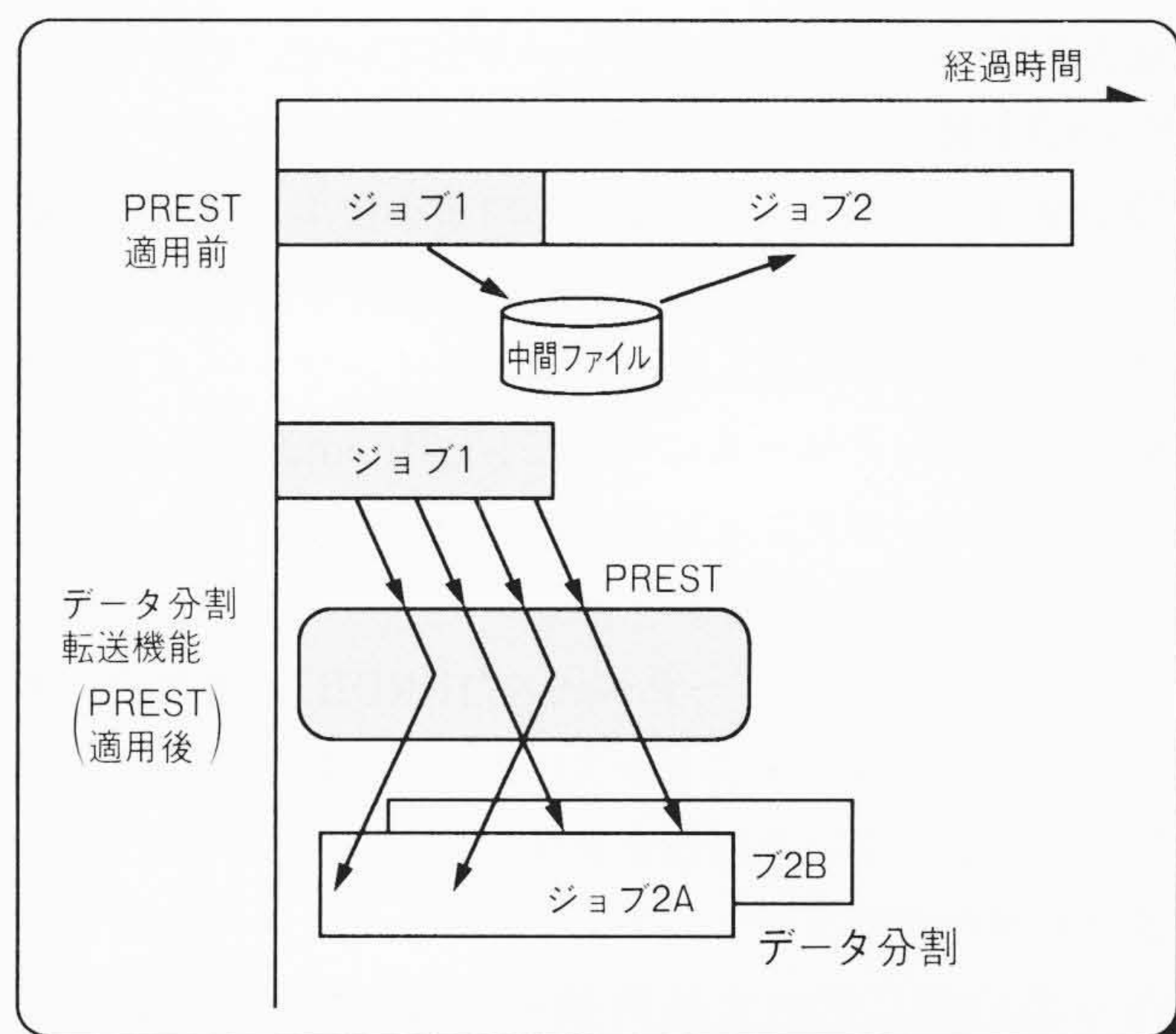


図2 PRESTのデータ分割転送機能

ジョブ1の出力する中間データを分割転送することにより、長大ジョブであったジョブ2を並列実行することが可能になり、ジョブの実行時間の短縮が図れる。

数のジョブステップで構成されているときに、簡単なパラメータの追加により、ジョブステップを並列実行させる機能である。

2.4 オープンシステムとの連携

VOS3/FSでは、パソコン、ワークステーション上の表計算などの著名クライアントソフトからスーパーサーバ上のデータベースの検索更新が行える機能に加えて、次の機能をサポートする。これにより、スピーディなビジネスプロセスが構築できる。

- (1) グループウェア(日立のGroupmax)との連携によるメールの相互受発信や掲示板の共用
- (2) 異種メールとの相互接続
- (3) サーバ・ワークステーションからのMパラレルシリーズでのバッチジョブ、TSSコマンド、アプリケーションの実行
- (4) Mパラレルシリーズと日立の3500サーバとの高速接続による性能向上
- (5) 3500サーバやワークステーション上のWWW(World Wide Web)サーバ上のAP(Application Program)と連携することにより、インターネットからホストのデータベースへのアクセス

3 並列データベースシステム

業務追加などによってCPUの処理能力が限界に達する場合、ユーザーは現行業務のアプリケーション、データベースなどに対する影響を回避するため、より高い処理

能力を持つCPUにアップグレードしていた。しかし、1台のCPU能力には限界があるので、統合型データベース“XDM E2”では既存資産をそのまま使用しながら段階的にCPUを追加しスループットを向上できる並列データベース機能を開発した。既存のメインフレーム上に構築されたデータベースに対してI/O(Input/Output)負荷の高いランザクション処理を並列化することにより、実行時間が短縮できる。

一方、RISC並列データベース“HiRDB”の並列データベース機能ではCPU台数がXDM E2の数台に対して、数十台とより高い並列性を重視して開発しており、高い並列処理が必要な新しいアプリケーションへの適用をねらっている。データ連携機能により、XDM E2とHiRDBが業務特性に応じて連携して使用できる。

3.1 XDM E2のMパラレルサポート

XDM E2は構造型データベースのXDM/SD E2 (Structured Database Extended Version 2), RDB (Relational Database)のXDM/RD E2 (Relational Database Extended Version 2), およびDC (Data Communication)機能のXDM/DCCM 3 (Data Communication and Control Manager 3)など、各種サブシステムを配した統合型データベース/DCシステムである。

3.1.1 XDM E2パラレル機能の概要

(1) 機能分散パラレル

Mパラレルプロセッサは複数のCPUを結合することによって、より高い性能を得ることができる。しかし、一般に既存の大規模オンライン業務、データベースアクセスを伴う大量バッチジョブ業務は分割することが困難である。

XDM/SD E2, XDM/RD E2の機能分散パラレルではアプリケーションを実行するノード(APノード)でデ

ータベース処理を抽出し、バックエンドのデータベース処理用ノード(DBノード)に転送する。データベースノードではデータベース処理を実行した後、検索結果など処理結果をAPノードに返す。CPU負荷の高いデータベース処理をバックエンドの別CPUで実行することにより、アプリケーションの変更なしにシステム全体のスループットが向上できる(例：各ノードのCPU負荷が均等になるようにデータベース配置した場合、4ノードで約3倍のスループット)(図3参照)。

(2) データベースパラレル

データベースパラレル機能とは、データベース処理を並列に実行することによって実行時間を短縮するものであり、バッチ・ユーティリティパラレル機能、入出力パラレル機能、およびSQLパラレル機能がある。

バッチ・ユーティリティパラレル機能は、データベースの再編成ジョブなどをエリア単位に独立して並列実行するものであり、XDM/SD E2, XDM/RD E2としてすでに製品化している。

一方、XDM/RD E2の入出力パラレル機能では、複数のエリアに分割格納された表を検索する場合に、エリアごとに入出力を並行処理する。このため、条件指定によってデータベースの広い範囲から目的のデータを絞り込むような、I/O負荷の高いSQLを高速に実行できる(図4参照)。またSQLパラレルは、ジョインなど負荷の高いSQLサブSQLに分解し、並列実行する機能である。

3.1.2 パラレル機能の特長

XDM E2のパラレル機能の特長を以下に述べる。

(1) スケーラビリティ

機能分散パラレルによって段階的なハードウェアリソース(ノード)の追加ができる(最大8ノード)。

(2) 移行性の確保

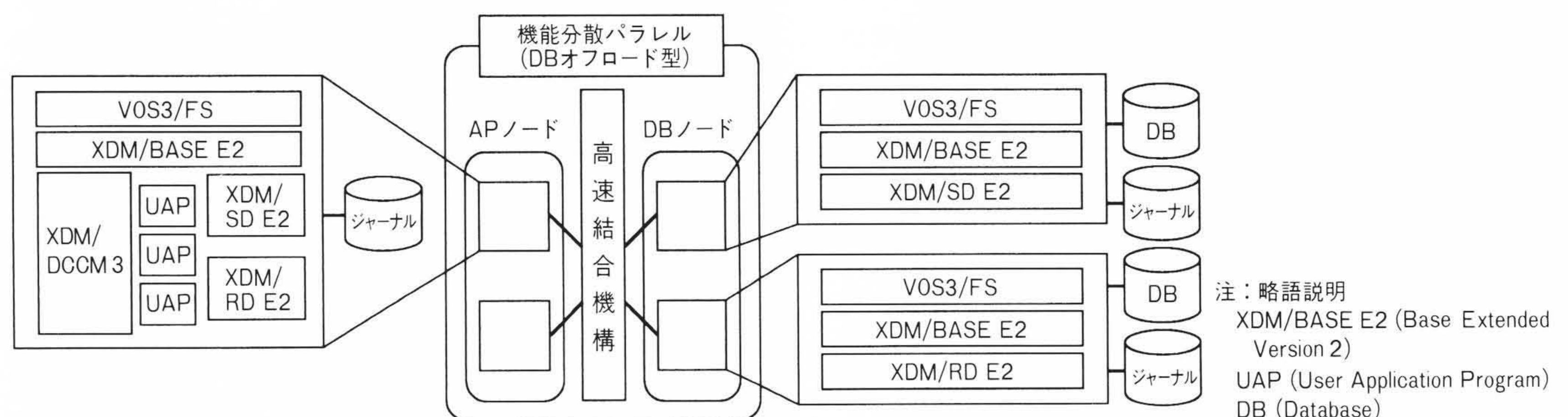


図3 XDM E2の機能分散パラレル

XDM E2では、APやデータベースの各ノードに機能分散してデータベースシステムを実現することにより、システム全体のスループットを高めることができる。

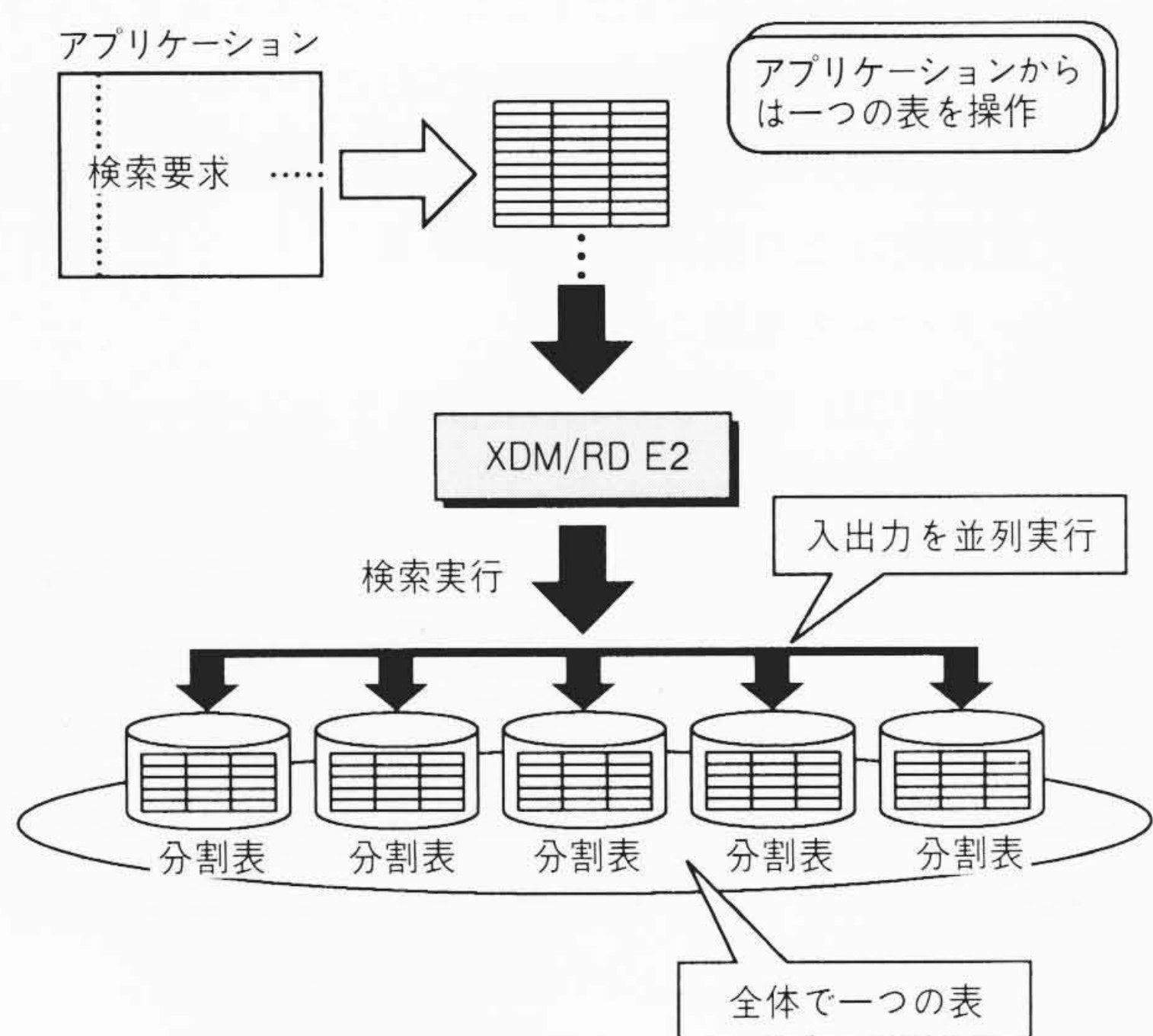


図4 XDM/RDのI/Oパラレル

XDM/RD E2では、一つの表を物理的に最大256の分割表に分けて並列実行することができる。

機能分散パラレルでは、アプリケーションはAPノードで一元的に実行されるため、プログラム、運用手順の移行性が高い。

(3) 1システムイメージ運用

統合コンソールによって一元的な運用が可能である。運用コマンドの振り分けはシステムがブロードキャストによって実現するため、システムの開始、終了や状態表示など1システムイメージで運用できる。

(4) 実行時間短縮

データベースパラレル機能によってメインフレームに構築したデータベースに対してそのままパラレル処理を適用することにより、実行時間が短縮できる(例：12分割の1,000万件データベースから1,000件抽出で経過時間約 $\frac{1}{10}$)。

3.2 RISC並列データベースHiRDB

HiRDBは、UNIXベースのスケラブルデータベースサーバであり、3500サーバシリーズ上で動作する。HiRDBは、シェアドナッシングという、メモリもディスクも共有しないアーキテクチャに基づいて設計されており、非常に高いスケラビリティを実現している。高い並列性の引き出しが可能であり、新しいアプリケーションを創造することが期待できる。

HiRDBは、複数の独立したデータベースアクセスサーバとSQLを受け付けるSQLサーバから成る(図5参照)。また、CPUの不足を補うためのフLOATブルSQLサーバを持つ。HiRDBはクラスタ構成の3500群に分散して

存在し、論理的に分割した表に対して並列のアクセスを可能としている。HiRDBの構成を図5に示す。

3.2.1 HiRDBの特長

(1) 高いスケラビリティ

HiRDBは、複数のプロセッサ上で動作できる。各プロセッサ間は、ネットワークで接続されている以外は共有しているリソースを持たない。したがって、4台で3.9倍といった非常に高いスケラビリティが実現できる(図6参照)。処理量やデータ量が増えてもプロセッサを増加することによって対応できる。

(2) 徹底した並列化を実現

(a) HiRDBでは、検索性SQL(SELECT)だけでなく更新系SQL(UPDATE, INSERT, DELETE)も並列化して実行できる。

(b) 運用関連機能(データロード、バックアップ・リカバリ、リランなど)を並列で実行できる。

(3) 高信頼性

(a) HiRDBの基本的回復機能は、XDM/RD E2の技術を踏襲している。

(b) 並列化に伴ってリカバリを強化した。シェアドナッシング方式のため、自然にリランも並列に実行できる。各データベースアクセスサーバは、自立的に並行してリランを行う。

(c) データベースはミラー化(OS機能)できる。

(d) 各プロセッサ(ノード)の障害時は、ノードを切り替えて処理を続行できる。シェアドナッシングであるが、障害対策用に2台のノードでディスクを共有しておき、HA(High Availability)モニタと呼ばれる監視機

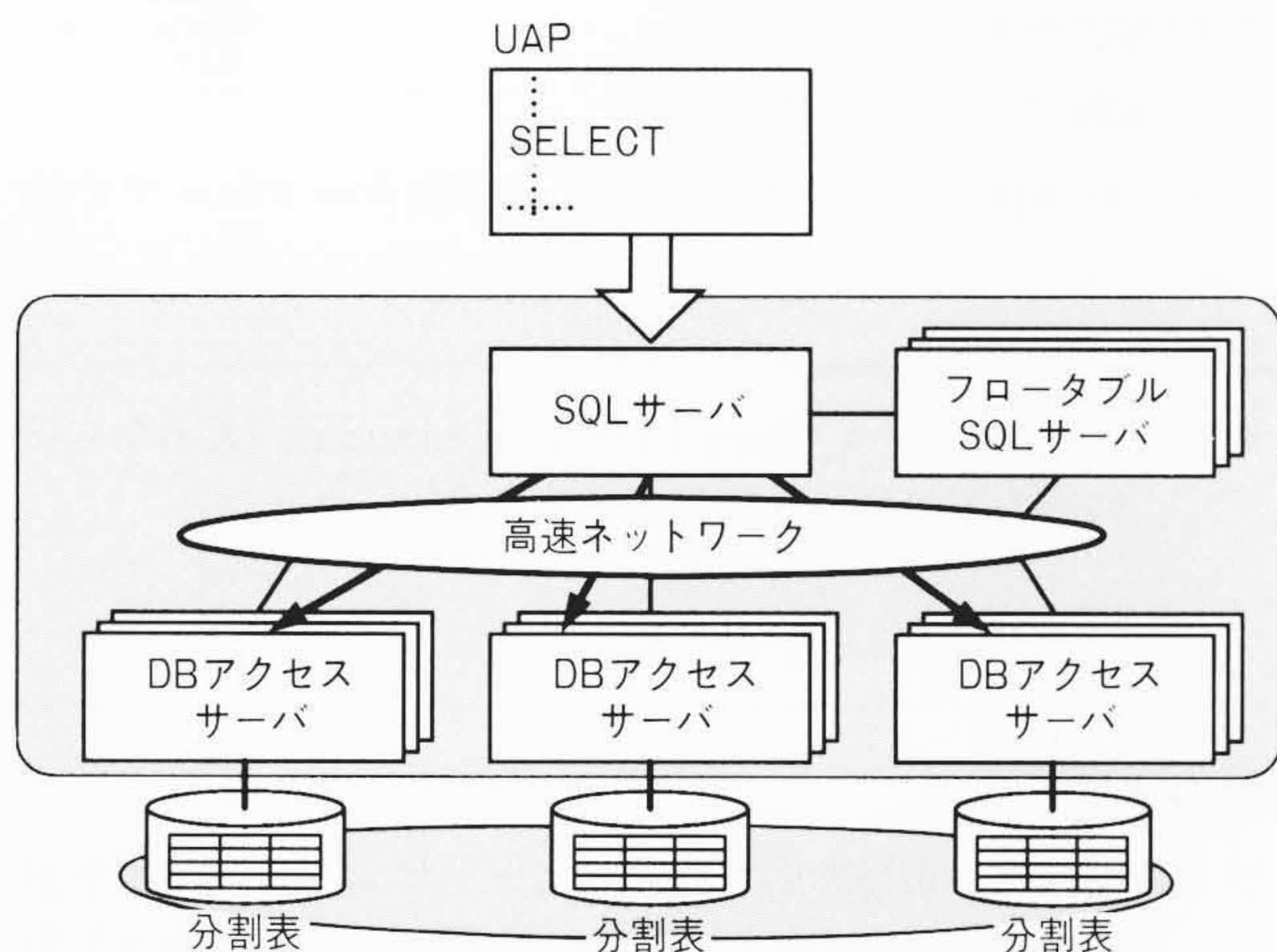


図5 HiRDBの構成

ディスクの接続されたデータベースアクセスサーバとSQLサーバ群を高速ネットワークで接続したクラスタ構成である。

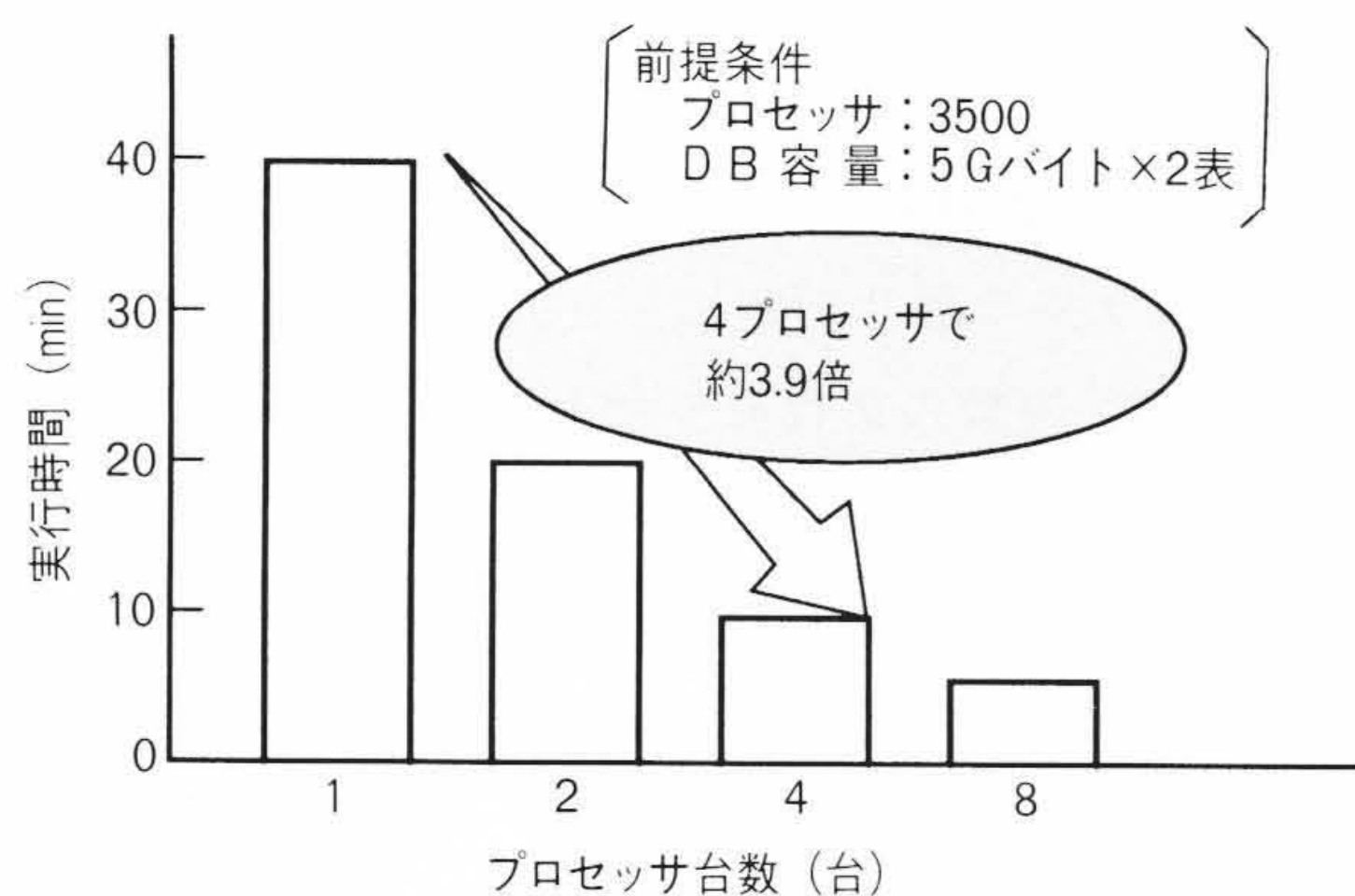


図6 ジョイン性能

1,000万件の表と1,000万件の表とをジョインするモデルでプロセッサ数を変化させた。4プロセッサで3.9倍の性能を実現した。

構を用いて実現する。通常時は片系からだけアクセスするので、シェアドナッシングの特性を失うことはない。

(4) オープン性

- (a) HiRDBはHI-UX/WE 2ベースのデータベースであり、特別なハードウェアに依存していないため、高いポータビリティを持つ。今後サーバについては、他社UNIXプラットフォームへの搭載を予定している。
- (b) 各種標準化(ISO SQL, OSI RDA X/Open^{※2)}XAなどに準拠する。

(5) 一元運用

HiRDBは複数ノードから成るデータベースであり、運用は1台のワークステーションから集中的に行える。運用コマンドは各プロセッサにブロードキャストされ、自動的に全ノードの開始、終了や状態表示が行える。

3.2.2 XDMとHiRDBの連携機能

XDMとHiRDBの連携には、大きく分けて次の二通りの方法がある。

(1) 分散データベース

- (a) HiRDBは、XDM/RD E 2と分散データベースで接

続する。既存のXDM/RD E 2にアクセスするアプリケーションをそのまま利用してHiRDBにアクセスできる。

- (b) トラフィックが高くなく、個々の検索、更新要求が一定規模以上の場合に効果がある。

(2) データベース連携

- (a) XDM/SD E 2やXDM/RD E 2からデータ抽出し、その結果をファイル転送してHiRDBに反映できる。大量のデータ反映には、大量データ分散処理機能Transbatchを用いると並列化ができ、効率が良い。
- (b) 今後、XDM/DSのようなデータ連動ソフトを用いることにより、連続的にXDMのデータベースの内容を反映できるようにする計画である。これにより、オンライン複合処理(OLCP)の支援も可能となる。

3.3 並列データベースの適用

並列データベースは並列検索の力が強く、大容量のデータベースへの適用が可能である。最も有効な応用は、100万件以上のデータベースに対する複雑な検索である。例えば、1,000万件以上ある顧客情報から、ある条件に合うものを取り出すことがあげられる。これは、金融系であれば、特定の金融商品を持ってはいないが一定以上の貯金高を持っている顧客を、流通業であれば、特定の商品を何か月か以内に購入して、かつ一定回数以上の取引のある顧客をそれぞれ検索することなどである。

XDM/RD E 2およびHiRDBに代表されるスケラブルデータベースは、従来現実的に実現が困難なアプリケーションを可能とする機能を持っており、爆発的データの増大に追随するキー技術になると考える。

4 おわりに

ここでは、今後のデータセンタを支えるパラレルソフトウェア製品としてのVOS3/FS, XDM, およびHiRDBの並列機能について述べた。これらの製品群はトータルシステムコンセプトであるFOREFRONTの考え方に沿ってオープン化、分散化に対応している。

今後も情報処理技術を先取りし、顧客のニーズに合った製品を開発していく考えである。

※2) X/Openは、X/Open Company Limitedの英国ならびに他の国における登録商標である。

参考文献

- 1) 松岡, 外: 市場即応型ハイスピードビジネスの実現を目指す新FOREFRONT, 日立評論, 77, 5, 334~338(平7-5)
- 2) 須永, 外: 新世代大型コンピュータ Mパラレルシリーズの開発, 日立評論, 77, 5, 371~374(平7-5)
- 3) 長須賀, 外: 入出力の仮想化と並列処理によるバッチ処理の高速化機能PREST, 情報処理学会論文誌, 第35巻, 第5号, pp.856~864(1995-5)