



長年にわたって並列データベース処理を研究している東京大学生産技術研究所喜連川研究室では、並列コンピューティングにおける共有ディスクの有効性を確認すべく、SANを用いた実証実験にて従来よりアプリケーション性能が30%向上することを示した。実験に使われたのは32台のパソコンとディスク・アレイで構成された並列コンピューター・ネットワークで、共有ディスクへのアクセスにはFCスイッチを介したSANが利用されています。

SOLUTIONS

概要

●技術的課題：

並列コンピューティングにおける共有ディスクへの有効性を検証すること。

●ソリューション：

最大128MBまでのデータを一気に転送できるファイバーチャネルとSAN接続の効率を高めるBrocade SANスイッチ・ファブリック

●成果：

- ・SANを用いて、従来よりアプリケーションの処理性能を30%は高められる方式を実現
- ・共有ディスクの仮想化を実現
- ・動的デクラスタリングを実現

並列コンピューティングにおける共有ディスクへのアクセスにSANを活用

コンピューターは科学の最先端分野なので、実用化に先立つアイデアの多くは大学などの研究機関から生まれ、地道な基礎研究が続けられています。東京大学生産技術研究所に属する喜連川研究室では、データベース処理を高速化するための研究をここ十数年進めており、その成果は毎年のように学会論文として発表されてきました。その喜連川研究室が最近選んだ研究テーマの1つが、並列コンピューティングにおける共有ディスクの有効性を検証することです。

並列コンピューティングとは、同時に動作する複数のコンピューターに処理を分担させる技術のことで、高速な処理速度と高い可用性が得られるという利点があります。この技術はすでに実用化レベルにあり、小規模なものでは、1つの筐体に収めた複数のCPUに仕事を分散させる対称型マルチプロセッシング（SMP）、大規模なものでは、数十台から数千台のスーパー・コンピューターを超高速ネットワークで接続するグリッド・コンピューティングがよく知られている例となるでしょう。

並列コンピューティングを実現する上での大きな課題の1つは、ディスクに保管されているデータへのアクセスをどのように効率化するかとい

うことです。従来の並列コンピューティングで広く使われていた疎結合（shared nothing）方式では、ディスクは個々のコンピューターに直接に接続され、コンピューター間のデータのやりとりはネットワークを介して行われていました。このため、大量のデータが関係する処理を並列コンピューティングで行おうとすると、ネットワークにかかる高い負荷が並列処理そのものに影響を及ぼすこともありました。喜連川研究室では、並列コンピューティングにおけるディスクアクセスのこうした問題を、SANを用いた共有ディスク方式で解決しようとしたのです。

32台のコンピューター群とディスク・アレイをFCスイッチで接続

SANには、共有ディスクへのアクセスを高速化するのに役立つ特性が2つあります。

まず、LANやWANなどのネットワークとは別の系統になっていることから、互いの影響を被ることがありません。ディスクとのデータ転送性能を最大限に確保しつつ、コンピューター間のデータとコントロールのやりとりについても100%のパフォーマンスを発揮させることができるのです。

SOLUTIONS

次に、SANで使われているファイバー・チャネル（FC）は、TCP/IPでのネットワークに比べて原理的に大容量のデータを高速に転送できます。LANやWANでは比較的短いブロック長（Gbit Ethernetでは1.5KB）が使われているので、ブロックの分解と組み立てをするのに多くのCPU能力を必要とします。これに対し、SANでは転送コマンドで128MBまでのブロック長を指定できますから、CPUの割り込み回数はそれだけ減ることになり、用途に応じて最適なブロック長を選ぶといった高度な使い方も可能です。

こうしたSANの特性を生かして共有ディスクへの高速アクセスをどのように実現するか、喜連川研究所では並列コンピューター・ネットワークを作って実際に試してみました。実験システムの予備的な検討は1999年4月に始まり、2000年7月からスタートした開発作業は2001年4月に完了しました。

並列動作のコンピューター群となるのは32台のx86パソコンで、それぞれの間は100Base-TX仕様のLANとFastEthernetスイッチ（Blackdiamond製）で接続してあります。また、共有ディスクとしては18GB容量のドライブを32基搭載したディスク・アレイ（Ciprico製FibreStore）を用意し、パソコン側に装着したFCホスト・バス・アダプター（Emulex製LP8000）からの光ケーブルを、FCスイッチを経由してディスク・アレイにつなぎました。FCスイッチはすべてプロセードコミュニケーションズシステムズ製で、パソコン側には4基のSilkworm 2050（8ポート）で構成したFCアービトレートッド・ループ（FC-AL）、ディスク・アレイ側にはSilkworm 2800（16ポート）が取り付けられています。

共有ディスクの有効性の検証には、実用的なアプリケーションであるデータマイニング処理を利用して行われています。大量のデータから一定の傾向を読みとるデータマイニングには計算とデータベース・アクセスの量が多に多いという特徴があるので、並列コンピューティングとディスク・アクセスの両方の

性能を試すのに最適です。

仮想化と動的デクラスタリングで

30%の能率向上は達成可能

実験システムを使って得られた研究の成果は、「SAN統合PCクラスタ上の並列データ・マイニングのための動的データ・クラスタリング」（情報処理学会データベースシステム研究報告、2001）などの論文として学会に発表されました。共有ディスクへのアクセスに関しては、複数のデバイスに存在しているファイル群を論理的な1ファイルとして扱う仮想化とディスク・アクセス能力を必要に応じて拡張する動的デクラスタリングの2点が、主要な成果となっています。

一般に、研究室での成果がビジネス界で実用化されるまでには長い年月を必要とします。しかし、その効果について喜連川研究所のスタッフは「SANを用いることで従来のアプリケーション処理性能が30%は高められる」と見ており、将来には大きな期待が持てそうです。

この研究成果を基に、喜連川研究室ではWebサイトのリンク解析を始めようとしています。Webページ間のリンクで成り立っているWorld Wide Webの世界に統制的な組織はなく、Webサイト間のつながりの全貌はだれにも分かっていません。そこで、並列コンピューティングを利用し、6000万ページ以上の個々のWebページに埋め込まれているリンク（HTMLの<a>タグ）を高速に解析して、日本全体のWebサイト構造を明らかにしようとしているのです。トップのWebサイトからリンクをたどるクロウリング処理でディスクに溜め込まれたデータを解析するのに現在は数日かかりますが、これを数時間のレベルにまで短縮しようというのが当面の目標です。多数のドライブで構成された大規模ディスクへのアクセスに十分な能力を持つSANは、この新しい研究テーマでも基礎的な技術となることは間違いありません。



© 2002 Brocade Communications Systems, Incorporated. All rights reserved. GA-CS-280-00-J

Brocade, SilkWorm, Extended Fabrics, Remote Switch, Fabric Aware, Fabric OS, Fabric Watch, QuickLoop, SOLUTIONware, WEB TOOLS, Zoningは、米国またはその他の国におけるBrocade Communications Systems, Inc.の商標または登録商標です。その他のブランド、製品名、サービス名は各所有者の製品またはサービスを示す商標、登録商標、サービスマークである場合があります。

注意:本ドキュメントは情報提供のみを目的としており、Brocadeが提供しているか、今後提供する機器、機器の機能、サービスに関する明示的、暗示的な保証を行うものではありません。Brocadeは、本ドキュメントをいつでも予告なく変更する権利を留保します。また、本ドキュメントの使用に関しては一切責任を負いません。本ドキュメントでは、現在利用することのできない機能について説明している可能性があります。機能や製品の入手可能性については、Brocadeのセールスオフィスまでお問い合わせください。

本ドキュメント中の技術データを輸出する際には、アメリカ合衆国政府の輸出許可が必要になる場合があります。