並列DBのための高速で高信頼な分散ロック機構

1D－5

片山朝子†　岸本光弘†　黒沢崇宏†　鶴見昌弘‡
富士通研究所†　富士通北陸システムズ‡
{archan,kiss,kurosawa}@flab.fujitsu.co.jp,tsuru@so.fjh.se.fujitsu.co.jp

1. はじめに

並列データベースにおいては排他処理の通信オーバヘッドが高速化の妨げとなる．富士通では当社のUNIXサーバGRANPOWER7000を構成ノードとするクラスタ間の通信路に高速結合網であるAP-Netを採用し、通信ハードウェアの高速化と、プロトコルオーバヘッドの大きいTCP/IPの代わりにAP-Netの直接通信を行うことによる高速化を図った。本稿では分散ロック機構がAP-Netを使用する際の高速高信頼な通信を行なう方式について示す。

2. AP-Netの特徴と問題点

AP-Netは当社の並列計算機APシリーズのノード間通信用に開発された高速結合網で、片方向200MB/s、トポロジは二次元メッシュ(トーラス)で最大結合ノード数1024台,ホストアダプタはS-Busを用いる。

AP-Net直接通信にはAP-Net通信ライブラリ(以下APLIB)を使用する．APLIBはノード毎に1プロセスによる利用しかサポートしていない。また通信の信頼性は保障しておらず、使用するアプリケーションに任されている。

従って並列DBのように複数プロセスによるアクセスや信頼性が要求される場合，APLIBを利用するアプリケーションがそのような機能をサポートする必要がある。

3. AP-Net直接通信のための拡張機能

分散ロック機構(Distributed Lock Manager,以下DLM)は今回並列DBMSとして採用したOracle Parallel Server(OPS)のノード間通信モジュールで並列DBの排他処理を行なう。従来当社のDLMはEthernet上などでのTCP/IPを使用して通信を行なっていた。

今回このDLMがTCP/IPプロトコルスタックを使用せず、APLIBを通してAP-Netハードウェアを直接使用して通信を行う(図1参照).

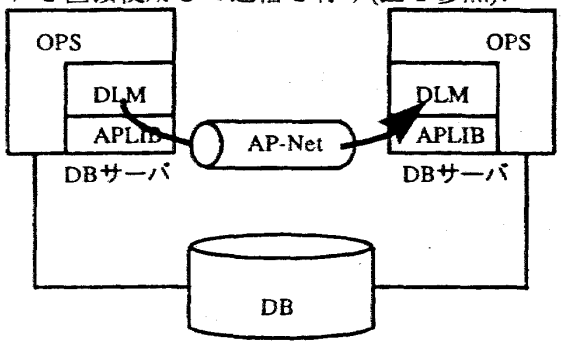

図 1 並列DBの構成図

できるだけ最小の変更で，高速性を維持しつつ高信頼な通信を行なうために複数プロセスからの使用やエラー/フロー制御の機能を以下の方式で実現した。

(1)ソケットセマンティクスのエミュレーション

AP-Net直接通信にはソケットという概念がない。従来はDLMおよびOPSモジュールがソケットを用いたDLMの通信機構を使ってノード間通信を実現していたので、ソケットセマンティクスをエミュレートすることが必要である。

ノード内通信ではカーネルの提供するファイルディスクリプタ(以下 fd)をベースとしたsend/receiveによる通信を行う．AP-Netを使用するユーザレベルのノード間通信ではAPLIBの提供する通信処理を行なうが，通信初期化時に擬似的なfdを作成し、メッセージ送信時に付加する．受信側でfdの種類によって処理を分けることでシングルリードを実装した。

†Asako KATAYAMA,†Mitsuhiro KISHIMOTO,†Takahiro KUROSAWA,‡Masahiro TSURUMI
† Multimedia Systems Laboratories.Fujitsu Laboratories LTD.
‡ Fujitsu Hokuriku Systems Limited.

更にノード内複数プロセスがメッセージ送信を行うために、送信バッファとその制御情報を共有メモリ上に置くことでプロセス間で共有し，AP-Net を利用できるようにした．

(2)エラー制御機能

AP-Net 直接通信で信頼性をサポートするために、一時的なメッセージ脱送などに対処する脱走検出および再送機構と、ネットワーク自体の故障に対処する二重化機構を実装した．

一時的なメッセージ喪失に関しては、送信側でメッセージにノード毎のシーケンス番号を埋め込み、受信側でそれを確認することで脱送検出を行う。脱送が起こると受信側 DLM がメッセージの再送要求を行うようにし、再送の自動化を実装した。

ネットワークの故障に関しては AP-Net ハードウェア二重化し，運用と待機とに分ける(図 2 参照)．運用中の AP-Net が故障すると，DLM は故障した AP-Net を切り離し、待機用 AP-Net の通信初期化処理を行い，新たに運用とする．

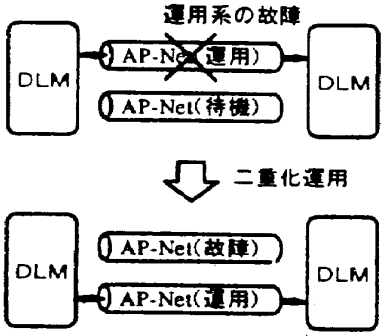

図2 AP-Net 二重化運用

(3) フロー制御機能

AP-Net 直接通信ではメッセージを送信バッファ上の空領域に作成し送信する。メッセージ領域の解放は受信確認後，受信ノード毎に行われるのでノード毎に送信バッファの管理を行う必要がある．

送受信バッファの大きさは全ノードで等しく設定し、送信先ノード毎にバッファ領域を等分割する．あるノードへの送信バッファの合計はそのノードの受信バッファの大きさを超えない。それによって送信バッファの領域管理のみで受信バッファの空き領域を考慮する必要がなくなった。

4. 評価

DLM 操作のうち、リモートノードにロックの変換要求を行ってから応答を得る操作について次のような測定を行った．GRANPOWER7000 を 2 台使用し、ノード間でそれぞれ 100BaseT(TCP/IP)と AP-Net の直接通信を使用してスループットを測定し，性能比較を行った．

その結果，AP-Net 直接通信は FastEther と比較してスループットは 3.85 倍高速になった．

5. おわりに

本稿では DLM について、高速結合網を使用する際に必要な機能の実現について述べ、それらを付与しても FastEther に比べてスループットは約 4 倍高速であることを示した。今回の実装はアプリケーション側であったが、将来高速結合網を利用するアプリケーションが多くなると予想され、通信ライブラリ内でこれらの機能を実現していく必要がある。

[参考文献]

[1] 富川ミユキ.: GRANPOWER7000 クラスタシステムの設計思想と新技術．第 56 回情報処理学会 1D-1 (1998).

[2] 福井恵右.: ビジネスソフト向けの高速インタコネク制御ソフト．第 56 回情報処理学会 1D-3 (1998).

[3] 日本オラクル株式会社 .: Oracle7 Prallel Server 概要および管理 R7.3

[4] O, Shiraki., et al.: Advanced High-Performance Network for Scalable Parallel Server, *Hot-Interconnects4*, Palo Alto,CA,Stanford University (1996)