1D－3

ビジネスソフト向けの高速インタコネクト制御ソフト·

福井恵右†，内藤雅行‡
富士通(株)§
{fukui, naito}@yk.fujitsu.co.jp

1　はじめに

弊社並列コンピュータ AP3000 システムで実績を持つ超高速通信路 AP-Net を使用して、クラスタシステム上での並列データベース(DB)の性能向上に必要な通信制御ソフトウェアの開発を行った。通信路二重化を前提とした高信頼 TCP/IP 通信路 HAnet、定周期診断による状態監視、クラスタ資源管理機構と連動した AP-Net リソースモニタ、などの可用性を向上するソフトウェアの開発を行った。これにより、低遅延・高速通信機能と高可用性と両立させることができた。

2　ノード間通信路

ビジネスソフトウェアに要求される可用性/信頼性や高性能を実現されるための方法の一つしてクラスタ構成をとる場合、クラスタノード間の通信路には次のような特性が求められる。1)ノード数増加に対応する性能向上(ノード数スケーラビリティ)、2)あるクラスタノードでの障害が他ノードに波及しないこと(独立性)、3)ノード単位のネットワークへの追加・削除の監視(ノード生存監視)、4)通信ハードウェアの故障への耐久性(耐故障性)、5)運用中故障の統計記録手段や活性保守機能(汎用システムとしての保守性)。一般に、LAN の通信媒体では十分なノード数スケーラビリティを確保することは難しく、SAN(System Area Network)を使用した上で、上記全てを満たす実装が求められている。

クラスタノード間通信路としての制約

AP-Net ハードウェアは通常の LAN で実現できない低遅延・高性能通信を実現する一種の SAN である。このための制約として、送信側と受信側のソフトウェアが緊密に協調して通信を行う必要がある。ノードがネットワークから一旦切り離された後の送信先処理の再開時に、直ちにノード復旧を確認するための手段として、通信参入通知プロトコルを考案した。

3　AP-Net ハードウェア

ノード数に比例したスケーラビリティ(ノード数スケーラビリティ)を実現するためには低遅延・高バンド幅の通信機能を提供することが必要である。AP-Net ハードウェアでは、ユーザ空間にマッピングしたレジスタ領域へのコマンド・プリミティブ(Send/Put/Get 等)書き込みによる転送の実行を可能にすることにより、オペレーティングシステム(OS)のオーバヘッドなしに、ユーザプロセス間でデータ転送を行うことができる。

4　ノード数スケーラビリティの実現

実際のアプリケーションプログラムにおいてノード数スケーラビリティを実現するためには、そのプログラムが AP-Net の論理チャネルを専有して、OS のオーバヘッドを伴わず通信処理を行うことが必要である。このために、直接通信機能を提供するドライバ、および、ライブラリを開発した。これにより並列 DB プログラムで必要とされるノード数スケーラビリティを確保できた[1]。

直接通信機能

Send/Put/Get 等の AP-Net 通信プリミティブをユーザプロセスから直接に発行するために、AP-Net ハードウェアのレジスタをユーザプロセスのアドレス空間にマップしそれを管理するための機構を実現した。また、AP-Net ハードウェアが提供する仮想的に独立した 3 本の論理チャネルを柔軟に対応して使用することにより、任意の論理チャネルに対して直接通信機能を利用し、残りの論理チャネルにおいて、DSD[2]、DLPI プロバイダを、追加のオーバヘッドなしに共存させることが可能である。

通信完了通知(同期)

AP-Net 通信の完了通知はハードウェアから割り込

· High Speed Interconnect Facility designed for business oriented software
† Keisuke Fukui
‡ Masayuki Naito
§ Fujitsu Limited

み、または、指定されたメモリ領域(完了フラグ)の更新の形で行われるが、前者を使用する場合は OS の介在が必要となる。後者を使用する場合はユーザプロセスで完了フラグを監視するビジーウェイトループを組む方法がある。この場合、完了待ちの間に CPU を無駄に消費する。AP-Net ハードウェアはごく小さい遅延(10 マイクロ秒)で実行を完了するため、割り込みのオーバヘッドを考えれば、通常の場合はこの方法で十分である。

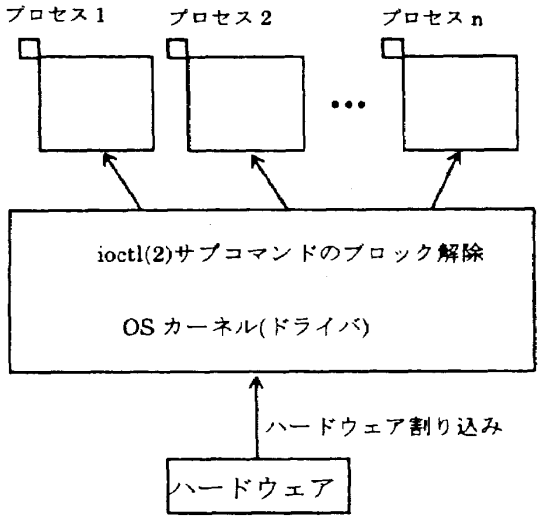

図1 マルチプル・リードでの完了待ちの構造

高速完了通知

送信プロセスと受信プロセスが独立し処理を行う場合、もしくは、複数プロセスが共通の論理チャネル上で受信を行う場合(マルチプル・リードである場合)、受信プロセスが長期間にわたり、ウェイトループを実行するなど不都合がある。これに対処するため、AP-Net ドライバ(APNET)の ioctl(2)サブコマンドとして、次回の完了通知割り込みまでブロックする機能を提供し、効率的に完了通知を行うことを実現した。

5 高可用性の実現

AP-Net 資源監視機能

AP-Net においては特定ノード間の通信のみが失敗する形態の故障も想定される。また、各論理チャネルはそれを使用するプログラムにより専有されるため、全てのエラーがドライバにより検出されるわけではない。このため、通信不達となる故障が発生した場合は故障範囲を特定することが論理チャネル利用者の責任となる。これを可能にするためにクラスタ資源管理機構とイベント管理機構[3] に基づいた、ノード間の故障判定を支援機能を実現した。

高信頼性/高可用性/耐故障性

AP-Net の二重化構成を可能とすることで既に提供済である HAnet と組み合わせて、単一ハードウェア故障への耐故障機能を実現した。HAnet は AP-Net の他、Ethernet / FDDI / FibreChannel で使用可能な TCP/IP 通信路二重化機能である。二重化通信路に対して一つの仮想アドレスを割りつけることにより、仮想アドレスに関して通信路の片側故障を隠蔽する他、定周期診断機能によりノードの生存監視と通信路故障の検出を実現している。

DLPI プロバイダ機能

AP-Net 上で動作する DLPI プロバイダ機能を提供することにより、UNIX SVR4 ストリーム機構により TCP/IP 通信を行う機能を提供している。

製品保守機能

製品保守機能として、ハードウェアレベルでは常時履歴保存を、ソフトウェアレベルでは常時履歴保存、指定による履歴保存の両方を実現した。これにより、ハードウェア・ソフトウェア双方からの履歴資料が、柔軟な単位で、かつ、システムワイドな共通インタフェースにより採取可能である。また、新規導入設置時や定期保守時に実行されるシステム自己診断機能にも対応している。

6 まとめ

以上において説明した機能により、並列 DB において要求される高性能通信機能と高可用性をクラスタシステムにおいて実現することができた。今回、特定の並列 DB での利用を確認したが、他の並列 DB やオブジェクト連携機構の通信機構など、様々な分野への適用を検討したい。

参考文献

[1] 片山他, "並列DBのための高速で高信頼な分散ロック機構", 第５６回情報処理全国大会論文集, 1D-05, 1998
[2] 明石他, "大容量で高信頼な分散共用ディスク", 第５６回情報処理全国大会論文集, 1D-04, 1998
[3] 阿部他, "高信頼性を実現する資源管理機構とイベント管理機構", 第５６回情報処理全国大会論文集, 1D-02, 1998