# GRANPOWER 7000 クラスタシステムの設計思想と新技術*

1D－1

冨川　ミユキ†、田崎　英明‡、藤原　隆♦
富士通（株）§
{miyuki, tazaki, fujiwara}@yk.fujitsu.co.jp

## 1 はじめに

複数の UNIX サーバをネットワークで接続した、低コストでかつ高可用性の UNIX クラスタシステムが利用されているが、最近では、各UNIXサーバ（以下、ノードと呼ぶ）を高速で接続するネットワークの登場により、スケーラビリティも同時に得られるクラスタシステムが注目されはじめてきた。すでに我々は、GRANPOWER 7000 シリーズで高可用性とスケーラビリティを実現するためにクラスタ技術を導入してきたが、クラスタシステムでの更なるスケーラビリティ向上のために高速なノード間接続機構（以下、インタコネクトと呼ぶ）の採用と可用性の向上を目的とし開発を行い、実現することができた。

本稿では、クラスタシステムの設計思想とその結果実現した機能の概要を報告する。

## 2 スケーラビリティの向上

クラスタシステムを構成する各ノードで、一つのデータベースを共用してトランザクション処理性能の向上を図るためには、高バンド幅で低レイテンシのインタコネクト技術が必要である。このトランザクション処理の並列処理環境として、代表的な並列データベース・ソフトウェアに Oracle 社の OPS (Oracle7 Parallel Server Option)がある。OPSは、データベースを共用ディスクに入れるために、各ノードからの整合性を保つための排他制御機構であるDLM (Distributed Lock Manager)が必要となる。このDLMは、ノード間で共有ディスクへのアクセス権の排他を行うため、ノード間の通信性能が良くないと各ノードで排他するための時間がネックとなり、充分なスケーラビリティを確保することができない。つまり、ノード間の通信速度を飛躍的に向上させることが大きなスケーラビリティを確保する鍵と考えられる。我々は、並列データベースとしてOPSを使用したトランザクション処理性能の向上を図るべく、科学技術分野向けに当社が開発した並列サーバAP3000で実績のある超高速ネットワーク装置（AP-Net）をインタコネクトに採用し、AP-Net の通信制御用ソフトやユーザ通信インタフェースを使用して高速処理を可能としたDLMを開発した。

さらに、OPS を使用して大規模なトランザクション処理を行うには、各ノードから共用できるディスクの台数を多くする必要がある。しかし、ハード的に接続可能な台数に制約があるため、各ノードにローカルに接続されたディスクを高速なインタコネクトを使用してソフト的に共用できる分散共用ディスク(DSD : Distributed Shared Disk)を実現した。この機能により、大規模データベースの構築が可能となった。

## 3 高可用性の向上

大規模構成システムになっても可用性を維持するために、運用系のシステムが停止した場合、運用系の大規模な資源（ハード的な共用装置やシステム構築に必要なプロセス等）をいち早く待機系の資源に切り替え、迅速に業務を継続することが求められている。

我々は、これらの資源の状態を管理・監視し、制御する機能を開発した。この機能の実現により、故障発生時の切替え時間（業務継続にかかる時間）を従来の当社クラスタシステムより2倍向上させ、高可用性を実現した。

さらに、ノード間通信路の信頼性と可用性を保証

---

*Design center and new technology for GRANPOWER 7000 Cluster System
† Miyuki Tomikawa
‡ Hideaki Tazaki
♦ Takashi Fujiwara
§ Fujitsu Limited

するために、ノード間通信路の２重化構成を実現した。また、AP-Net の異常による通信路の切断を防ぐため、AP-Net を資源として監視し、すみやかに正常な AP-Net に切り替えることによって、通信路を確保する機能を実現した。

## 4 新機能の概要

GRANPOWER 7000 クラスタシステムで実現した新機能について概要を示す。

［ノード間通信に高速インタコネクト装置（AP-Net）を採用］

AP-Net は、ノード間のメッセージの中継を行う複数の RTC（ルーティングコントローラ）から構成されている。RTC は、メッシュ接続されているので最大４ノードのルート選択を高速に制御でき、片方向 200MB/S の高速なデータ転送を実現している。

［高速インタコネクト制御ソフト］

AP-Net は、一本の通信路に対し３本の仮想論理チャネルを装備している。ソフトウェアでは、それぞれに TCP/IP、DLM、分散共有ディスク（DSD）として通信する機能を提供する。また、AP-Net の異常を監視する機能を提供する。

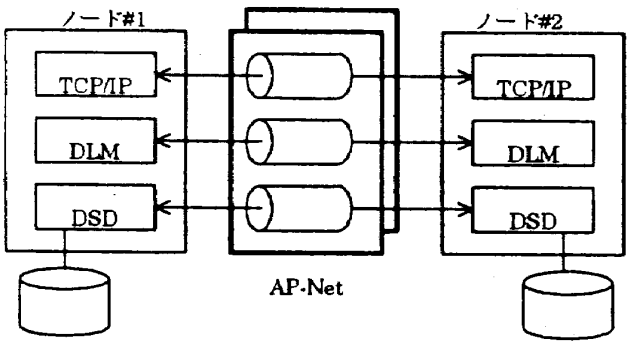


図 1 高速インタコネクトの利用

［資源管理機構とイベント管理機構］

クラスタシステムを構成するハードウェアやソフトウェアを資源として定義し、資源の状態を管理および監視する資源管理機構を提供する。また、資源の状態変化をイベントとして管理するイベント管理機構を提供する。

［分散共有ディスク（DSD）］

クラスタシステムを構築する各ノードに接続されたローカルな物理ディスクを、AP-Net を使用して共用ディスクとして見せる機能を提供する。

［分散ロック機構（DLM）］

通信として TCP/IP を使用すると、７つのプロトコル層を通るためにソフトウェアのオーバーヘッドが発生する。オーバーヘッドの大きい TCP/IP の代わりに、AP-Net のユーザ通信インタフェースを使用することで、高速な分散ロック機構（DLM）を提供する。

## 5 まとめ

OPS において、インタコネクトに AP-Net を使用した場合と、Fast Ethernet を使用した場合を比較すると、スループット、レイテンシ、各ノードの CPU 負荷がともに数倍向上した。また、高可用性という観点では、資源監視機構により大規模システムのようなディスクを多数台接続した場合の切替え性能を、3 倍以上向上させた。

我々は、インタコネクトに AP-Net を採用し、かつ新機能を使用することにより、高可用性とスケーラビリティを得るクラスタシステムを実現できた。

[参考文献]

(1) 阿部他，“高信頼性を実現する資源管理機構とイベントサービス”，第 56 回情報処理全国大会論文集 1D-2，1998

(2) 福井他，“ビジネスソフト向けの高速インタコネクト制御ソフト”，第 56 回情報処理全国大会論文集 1D-3，1998

(3) 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-4，1998

(4) 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-5，1998

(5) Oracle7 Parallel Server 概要および管理 R7.3

(6) 白木他，“高並列計算機 AP1000+のメッセージハンドリング機構”，情報処理学会論文誌，1996