特集

WindowsNT® サーバの高可用性（HA）システム技術

High-Availability System Technology for WindowsNT® Servers

金子　哲夫　T. Kaneko
滝本　秀明　H. Takimoto

PCサーバは急激な普及をみせ，最近では基幹システムにも使われるようになった。それに伴い，PCサーバにも高い信頼性が要求されるようになり，サーバ自体の高信頼技術とともに，サーバの構成を多重化することにより耐障害性に優れ，システムの可用性（Availability）を高めるクラスタシステムが脚光を浴びている。Microsoft®（注1）WindowsNT®（注2）をオペレーティングシステム（OS）とするグローバルネットワークサーバGSシリーズで実現したHAシステムは，高可用性（HA：High Availability）を目的としたクラスタシステムであり，他社にはない多くの特長をもっている。また，クラスタシステムでは，ミドルウェアに対する配慮も重要な要素である。

The price-performance ratio of PC servers has recently dramatically improved, with the result that they have begun to play an important role in corporate applications. On the other hand, higher reliability is now required in the PC server market. As part of this market trend, cluster systems are being spotlighted. A cluster system contains two or more PC servers in one system and provides higher tolerance of software and hardware malfunctions than a single-PC system.

The "HA (high availability) system for WindowsNT®" is a cluster system that runs on the GS series global network server, offering high availability to WindowsNT® servers. The "HA system for WindowsNT®" has several unique features that are not found in any other cluster system for WindowsNT®.

1 まえがき

PCサーバは，パソコン（PC）を出発点としているために，低価格を最大の武器にして市場を拡大してきた。それとともに，PCサーバをベースにした基幹システムが本格化しており，従来のPCの延長にある考えかたでは解決できない問題も表面化してきた。そのなかでも，最大の課題がシステム障害の発生による業務停止への対策である。耐障害性を高めるためにはフォールトトレラントシステムなどの技術は有名であるが，PCサーバシステムでは複数台のPCサーバをネットワークなどで疎結合したクラスタシステム方式が一般的に採用されている。実際に，当社を含めた国内外の有力PCサーバベンダは，WindowsNT®を搭載したPCサーバの高可用性，耐障害性要求にこたえて，クラスタシステムの提供を積極的に展開している。クラスタシステムは，もともと可用性と拡張性の二つの特徴を備えているが，現在のクラスタシステムは可用性を高めることを主たる目的としている。

ここでは，当社がUNIX（注2）サーバで培ってきたHAシステムを，当社のPCサーバであるGSシリーズに適用して実現した“HAシステム for WindowsNT®”（以下，HAシステムと略記）を具体例として取り上げ，PCサーバにおけるHAシステム技術について紹介する。

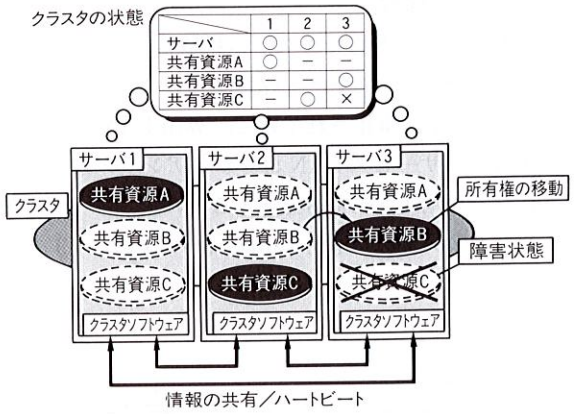

図1．クラスタHAシステムの動作概念　各サーバの状態がどのようになっているか，また各共有資源の所有権がどのサーバにあるかをお互いに監視しながら動作する。

Basic mechanism of cluster HA system

2 HAシステム動作原理

HAシステムそのものの説明の前に，高可用性を目的とし

(注1)　Microsoft, WindowsNTは，Microsoft社の商標。
(注2)　UNIXは，X/Openカンパニーリミテッド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。

たクラスタシステム（以下，クラスタHAシステムと呼ぶ）の動作原理について簡単に述べる。

クラスタHAシステムでは，まずクラスタに含まれる各サーバが相互に情報を交換しクラスタを形成する。通常，サーバ間の通信にはLANが使用される。サーバ間の情報のやりとりは相手のサーバが正常に稼働しているかどうかの判断に用いられ，これをハートビートと呼ぶ。クラスタが形成されたら，各サーバはクラスタ内に存在する資源の管理を共同で行う。管理対象になる資源には，一般的に①共有ディスク，②プロセス，③ネットワーク上のアドレス，などが挙げられる。

資源を管理するため，クラスタソフトウェアには資源に対する障害検出機能および各資源を有効化（組込み），無効化（切離し）する機能が必要になる。クラスタを構成する各サーバが，現在のそれぞれのサーバ状態や共有装置状態などのクラスタ状態についてすべてのサーバが同じ情報を所有することにより，サーバ内だけでなくサーバ間での資源の管理が可能になる。障害が発生したときに，クラスタソフトウェアは障害が生じた資源を切り離した後，待機している別の正常な資源を代替することで可用性の向上を図る。

クラスタHAシステムの動作の仕組みを図1に示す。

3 HAシステムの概要

3.1 システム構成

HAシステムの標準的なシステム構成例を図2に示す。

サーバを接続するLANは，専用（プライベート）LANとサービスLANで二重化されていて，専用LANはHAシステムが使用し，サービスLANはクライアント／サーバ間で使用される。専用LANを用意することにより，サーバが相互に監視をするために交信する情報（ハートビート）を，クライアント／サーバ間のLANの負荷の影響を受けずに高速に行うことができる。また，専用LANとサービスLANによってLANが二重化されているので，専用LANに障害が発生してもサービスLANを使用してハートビートを交信することができる。

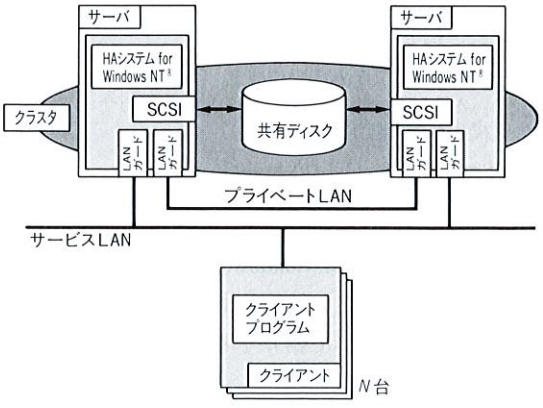

図2．HAシステムの構成例　2台のサーバがLANで接続され，ディスクを共有している。

Example of HA system configuration

共有ディスクは，2台のサーバにSCSI（Small Computer System Interface）で接続されていて，それぞれのサーバから排他的にアクセスされる。排他的とは，通常稼働系から使用している場合，他系のサーバからのアクセスを禁止することを意味する。

3.2 ソフトウェア構成

HAシステムのソフトウェア構成を図3に示す。

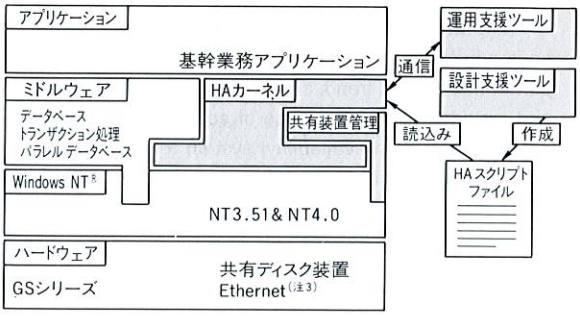

図3．HAシステムのソフトウェア構成　HAシステムはHAカーネルと設計・運用支援ツールで構成される。

Block diagram of HA system

3.2.1 HAカーネル

HAカーネルは，HAシステム全体の動作をつかさどる基本的なモジュールであり，各サーバにインストールされる。HAカーネルは，起動時にHAシステムの動作（HA戦略と呼ぶ）を記述したHAシナリオ（詳細は3.2.2項）を読み込み，各サーバのHAカーネルどうしでHAグループと呼ばれるクラスタを形成し，HAグループ内で共通して使用される資源の管理を行い可用性の向上を図る。

HAグループ内で共通して使用される資源は，HAシステムでは共有装置として定義されていて，共有ディスクなどの物理的な共有資源のほかに，プロセスやネットワーク上のアドレスなども仮想的な装置として扱う。HAシステムでは，①サーバNetBIOS，②共有ディスク装置，③プロセスを共有装置として定義する。

共有装置の所有権は，同時には同じHAグループ内の1台のサーバしか得ることはできない。実際の運用では，相互に関係のある共有装置を一つにまとめてサービスとして定義し，共有装置の所有権はサービス単位でサーバ間を移動

（注3）　Ethernetは，富士ゼロックス㈱の商標。

する。

3.2.2 設計支援ツール　HA カーネルが起動時に読み込む HA シナリオは，HA 戦略を記述するための専用言語で記述され，HA シナリオの存在が HA システムの大きな特長になっている。

エンドユーザ固有の HA シナリオは，あらかじめ用意された HA 戦略が組み込まれているシナリオ テンプレートに対象システム固有の情報，例えばサーバ名称や IP (Internet Protocol) アドレスなどを埋め込んで作成する。作成された HA シナリオは，実際には各サーバごとの HA スクリプト ファイルに展開されている。**図 4** に HA シナリオを作成する一連のファイルの流れを示す。

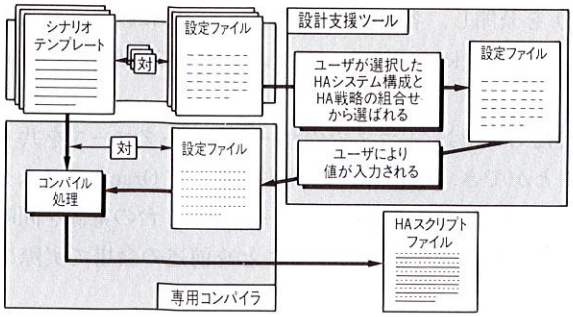

図 4．HA シナリオのファイルの流れ　シナリオ テンプレートにユーザ指定の情報を合わせてコンパイルし，HA スクリプトファイルを作成する。

Flow of HA scenario file

3.2.3 運用支援ツール　運用支援ツールは，HA システムに対する手動の操作や HA システムの状態の表示など運用上の手助けを行うものである。**図 5** に運用支援ツールの画面イメージ例を示すが，先に設計支援ツールでユーザが作成した HA シナリオに合わせた画面を表示する。設計支援ツールも選択した HA 戦略に合った同様の画面を見ながら HA シナリオ作成作業を進めるようになっているため，設計支援ツールと運用支援ツールの基本画面は同じものになる。

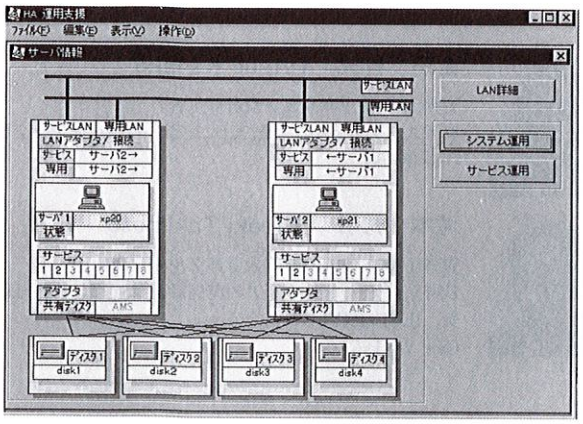

図 5．運用支援画面　HA シナリオに合わせた画面が表示される。

Example of operation support tool display

4 HA システムの要素技術

以上述べた HA システム技術のポイントをまとめると，次の 3 点に集約される。

(1) クラスタを形成する各サーバ間の通信
(2) 共有資源の管理
(3) システム運用の柔軟性

4.1 クラスタを形成する各サーバ間の通信

クラスタを形成する各サーバ間の通信は，クラスタの形成，相互の障害検出，システム状態の共有など，システムの根幹をなすものであるため，高い信頼性が要求される。

例えば，サーバ間で共有している情報の不整合や通信路に障害が発生し，相互で相手サーバがダウンしたと認識してしまうと，場合によってはそれぞれのサーバが勝手にサービスを起動してしまい，もし共有ディスクに書き込むようなサービスだった場合は共有ディスクのデータを破壊してしまう危険性がある。このような状態はスプレットブレインと呼ばれ，HA システムにとって致命的な状況と言える。たとえどのような状況でも，クラスタの状態に関する情報の共有や 相手サーバの障害検出が確実に行えることが重要である。LAN を使用している場合には，通信路の障害への耐性を高めるために，通常 LAN の多重化を行う。HA システムには，LAN 故障時に別の LAN を使った再確認機能，3 台以上構成時のルーティング機能をもつ。

4.2 共有資源の管理

共有資源の管理で重要なことは，資源の所有権をサーバ間で自由に移動できるかどうかである。WindowsNT® では，OS の外から自由に資源を有効化，無効化する方法が十分とは言えず，資源の所有権をサーバ間で移動することが困難であり，共有資源の種類に制限がついてしまう。また，資源が正常に動作しているか判断する資源の障害検出も，共有資源の管理においては重要である。

4.3 システム運用の柔軟性

クラスタが 2 台の場合，一般的には**図 6** の形態をとる。相互バックアップ型は，両方のサーバでサービスを実行し，どちらのサーバに障害が発生しても，正常なサーバでサービスを引き継ぐ。一方，スタンバイ型は，稼働系とバックアップ系を明確に分け，稼働系サーバに障害が発生したときにスタンバイ系に引き継ぐ。

サーバが 3 台，4 台と増えていくと，運用形態はさらに 2 対 1 バックアップ，ローテート バックアップなど複雑にな

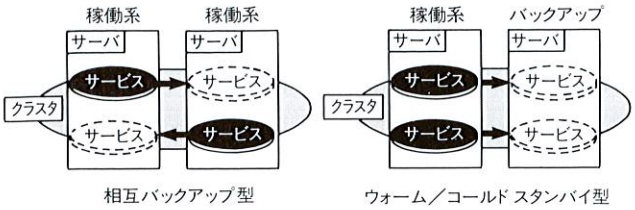

図6．2台クラスタ運用の形態　障害が発生したときに引き継がれるサーバをどのように待機させるかによる。

Operating style for 2 clusters

るが，ここでは割愛する。

以上のように，クラスタHAシステムにはさまざまな運用形態が考えられ，運用形態を柔軟かつ容易に設定できることが非常に重要である。HAシステムは，HAシナリオ方式の導入により多種多様な運用を可能にしている。

5 HAシステムとミドルウェア

HAシステム上でのミドルウェア検証結果およびOPS（Oracle(注4)パラレルサーバ）対応について述べる。

5.1 ミドルウェア検証

単独サーバ障害発生後の再立上げにどれだけミドルウェアが適応できるかでリカバリ精度が決まる。例えば，Oracle，SQL（Structured Query Language）サーバは障害時のデータベース保護がなされ，HAシステム化しやすいミドルウェアと言える。グループウェアでは，Lotus(注5) NotesがHAシステム対応が可能であるのに対し，現行のMicrosoft® Exchange 4.0は，内部構造上HAシステム対応が困難である。ここでいうHAシステム対応とは，サーバ切換え時に引き継ぐデータを共有ディスクに配置することが可能かどうかを意味する。

もう一つ重要なことは，クライアント側からのサーバ接続に何を使っているかである。現行，当社のHAシステムは，WINS（Windows® Internet Name Service）による接続だけをサポートしているので，直接コンピュータ名やIPアドレスを保持しているクライアントソフトウェアは，手動による再接続を必要とする。

以上，ミドルウェアをHAシステム対応させるためには，ミドルウェア側にリカバリ精度と接続の連続性について何らかの仕組みが必要であり，このような仕組みをもったミドルウェアも増えてきている。

5.2 OPS対応

1997年4月に行われた“Oracle Open World”において当社は世界初の4台クラスタ構成によるOPSを出展したので，簡単に紹介する。

OPSは，クラスタ化されたサーバからアクセスされるデータベースファイルを共有ディスクに置き，この共有ディスクのデータを同時に読込み，書込みを可能にする。さらに，どのサーバで障害が発生しても，そのサーバのユーザは別のノードにログインし，アプリケーションを実行し続ける。正常なサーバは，障害が発生したサーバで実行中であった不完全な処理をすべてロールバックし，自動的に復旧させる。これにより，データベースの論理的な一貫性が保証される。OPSは，API（Application Programming Interface）を公開し，各社のクラスタシステムに組み込むことによって，Oracleクラスタシステムを構築できるように作られている。

当社OPSは，4台までのサーバでデータベースを共有することができ，信頼性と拡張性に富んだOracleクラスタシステムを構築できる。4台のサーバ中の3台の電源を同時に落としても稼働を継続できることを前述の会場で実際に実証してみせた。

6 あとがき

PCサーバのクラスタシステムは市場が立ち上がったばかりであるが，高可用性さらには拡張性を向上させるソリューションとして今後も発展し続けるであろう。ただし，クラスタシステムは基本プラットホームとして位置づけられるので，ミドルウェアさらにはアプリケーションのクラスタ対応も含めて考える必要がある。当社はHAシステムをベースにして，今後さらに本質的なクラスタシステムを追求していく。

(注4)　Oracleは，Oracle Corporationの商標。
(注5)　Lotusは，Lotus Development社の商標。

金子　哲夫　Tetsuo Kaneko

青梅工場 ミドルウェア設計部主査。
クラスタシステムの開発設計に従事。情報処理学会会員。
Ome Works

滝本　秀明　Hideaki Takimoto

青梅工場 ミドルウェア設計部グループ長。
UNIX，PCミドルウェアの開発設計に従事。情報処理学会，IEEE会員。
Ome Works