WHITE PAPER

クラスタシステム概説

ご注意

本書の内容は、予告なしに変更されることがあります。

日本電気株式会社は、本書の技術的もしくは編集上の間違い、欠落について、一切責任をおいません。また、お客様が期待される効果を得るために、本書に従った導入、使用および使用効果につきましては、お客様の責任とさせていただきます。

日本電気株式会社は、NEC Express サーバ Express5800/100 シリーズ製品保証書で保証する内容以外には、一切の保証はいたしません。

Windows NT®クラスタリングテクノロジー

基幹業務システムは、一般的に膨大なデータ処理を行うため、性能の確保が重要な使命の一つです。また、基幹業務におけるシステムダウンは、その企業の全ての業務を停滞させ多大な損失を招くため、信頼性の確保も重要な使命となっています。

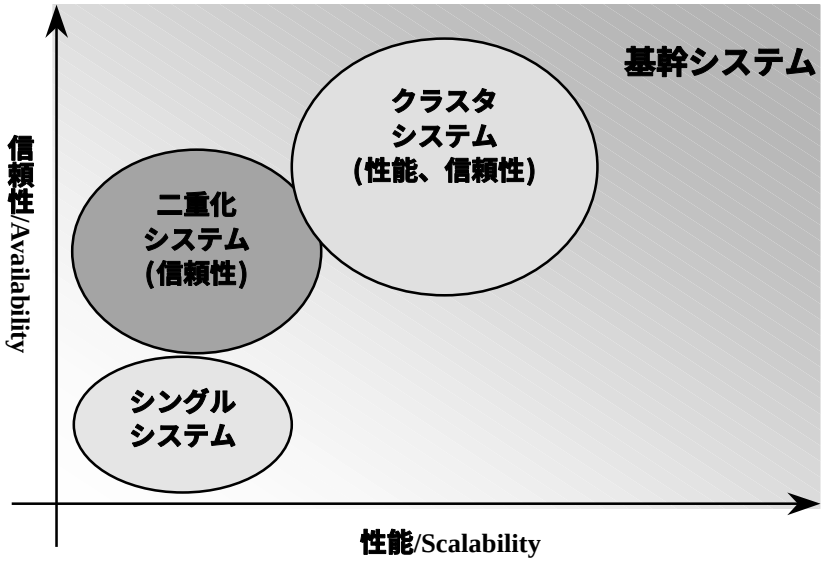
クラスタリングのあゆみ

信頼性を上げるために、システムを二重化するデュアルシステム/デュプレックスシステムは、汎用機の世界では以前から存在していました。しかし旧来、これらのシステムを実現するためには、専用のハードウェアとソフトウェアが必要であったため、高価なシステムとなり、それが信頼性への代償となっていました。

1990 年代になって、クライアント・サーバシステムが主流となり、安価なサーバを LAN と SCSI で接続した、安価な二重化を実現する UNIX®のソフトウェアが現れました。さらに、信頼性を確保するための二重化だけではなく、複数台の SMP(Symmetric Multi Processing)サーバで並列処理を行い、高い性能を実現するクラスタリング技術が市民権を得てきたのです。今後、さらに並列処理の台数は増加していくことでしょう。

Windows NT®は、オペレーティングシステムとしての歴史は浅いものの、最近では日本市場においても単なる PC サーバ・情報サーバとしてだけではなく、業務サーバとしての領域に足を踏み入れようとしています。今後、基幹業務への Windows NT®適用のために、クラスタリング技術は欠かせない技術となっていくでしょう。

クラスタとは、語源である"房"という言葉から想像されるとおり、複数のサーバからなるサーバ群を言い、かつそのサーバ群を１つのサーバのように扱う技術のことです。
クラスタには、サーバを冗長化することで信頼性を向上させる側面と、負荷を分散させ並列処理することで性能を向上させる側面があります。冗長構成によるシステム MTTR(Mean Time To Recover:平均修復時間)の短縮を『可用性(Availability)の向上』といい、クラスタのノードを追加することで性能向上が実現できることを『拡張性(Scalability)の向上』といいます。

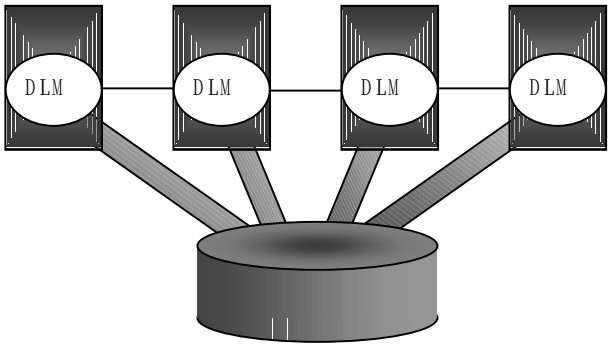


クラスタのモデル

クラスタモデルは、ディスクに対するアクセス制御方法によって、次のように分けることができます。

・シェアード・エブリシングモデル

複数のノードから、同時に同一ディスクに対するアクセスができるモデル。
ディスク内のデータを複数のノードから同時にアクセスできるため、ノードを追加しても、追加されたノードで同一の業務ができます。
ただし、複数ノードから同時アクセスされるため、書き込みの競合が起きた場合には、排他制御を行う必要があります。この排他制御を行うのが、分散ロックマネージャ(DLM:Distributed Lock Manager)です。

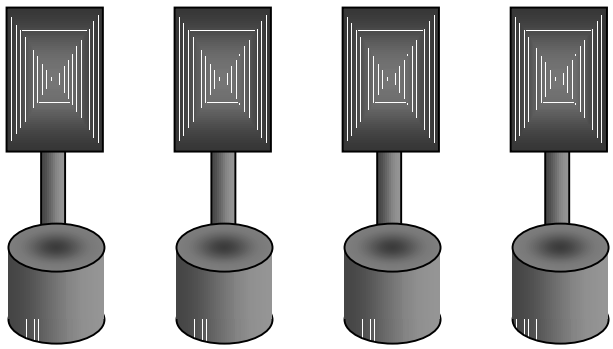


・**シェアード・ナッシングモデル**

あるディスクに対するアクセスを、１つのノードからのみであるモデル。

並列処理は、処理すべきデータを持つノードで行わなければならないが、データの分散が適切になされていれば、ノード数を増やしてもディスクへの書き込み時にノード間での排他制御が必要なく、性能がリニアに増加する特徴をもちます。

しかし、処理すべきデータがどのノードにあるか意識する必要があり、これをアプリケーションに意識させないためには、トランザクションモニタなどのミドルウェアによって、データ配置に対応した分散を実現させなければなりません。



NT クラスタの現状

Windows NT®がサポートするファイルシステムは、DLM の機能を持っていません。そのため、NTFS や FAT でフォーマットされたパーティション上のファイルに複数ノードから同時アクセスした場合、ファイルシステムレベルでの排他制御ができずデータ破壊を招きます。Microsoft® Cluster Server(MSCS)は、このファイルシステム上で動作するため、複数サーバから同時にアクセスすることはできません。

Windows NT®上で DLM の機能を持つものは、Oracle Parallel Server のみです。Oracle Parallel Server は、RAW パーティション(未フォーマットのパーティション)をデータベースファイルとして使用し、ファイルシステムを使わない物理 I/O を行うことで、OS のファイルシステムやキャッシュ等に依存しない、独自の排他制御を実現しています。

MSCS もフェーズ 2 では、ノード数を 2 台から 16 台へ拡大するとともに、シェアード・ナッシングでの負荷分散の実現が予定されていますが、そのほかの他社クラスタリングソフトウェアを含めても、Windows NT®上での負荷分散を実現している製品は、まだ数少ないようです。

クラスタのスタンバイ方式

可用性の向上に関しては、現在の Windows NT®用クラスタリングソフトウェアほぼ全てが実現しています。一般的に二重化されたサーバは待機の形態によって、以下の 3 つの方式に大別されます。

・コールドスタンバイ方式

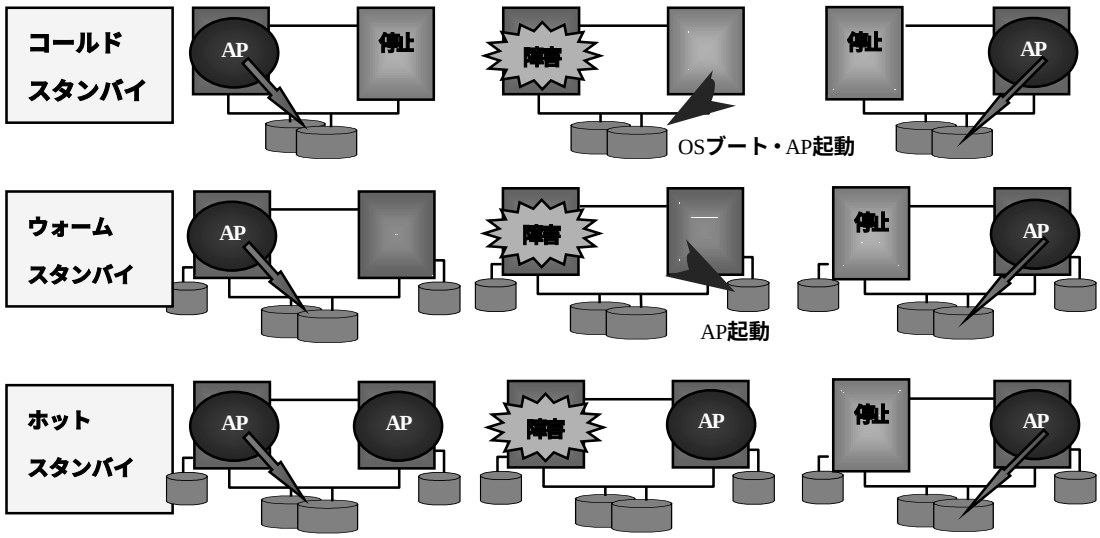
待機ノードは、Windows NT®の起動前で待機しています。待機サーバの電源 OFF の状態か OS ブートの直前で、稼動サーバのサーバダウンを検出すると、待機サーバの OS ブートが開始され、待機サーバが立ち上がりサーバの切替えが行われます。したがって、サーバダウン時の切替え時間がかかります。

・ウォームスタンバイ方式

クラスタ内のノードは、Windows NT®が起動しており、その OS 配下でクラスタリングソフトウェアが互いに他のノードを監視しています。監視は、ノード間のハートビートと呼ばれる系間通信により行い、このハートビートが途絶えたことによりサーバダウンの検出を行います。サーバダウンを検出すると、待機サーバでは、引き継ぐディスクデータや IP アドレスを移動し、アプリケーションを起動します。

・ホットスタンバイ方式

従来、汎用機で実現されていた、デュアルシステム/デュプレックスシステムの形態です。OS のみではなく、ミドルウェアもしくはアプリケーションまで起動した状態で待機、または現用系と同一の動作をしています。クラスタであることを意識したソフトウェアが必要ですが、サーバダウン時に短時間での切替えが可能です。



現在の NT クラスタリングソフトウェアの殆どがウォームスタンバイ方式で、ホットスタンバイ方式に比べサーバの切替え時間は長いが、上位のミドルウェアやアプリケーションが、クラスタであることを意識する必要がないという特徴があります。

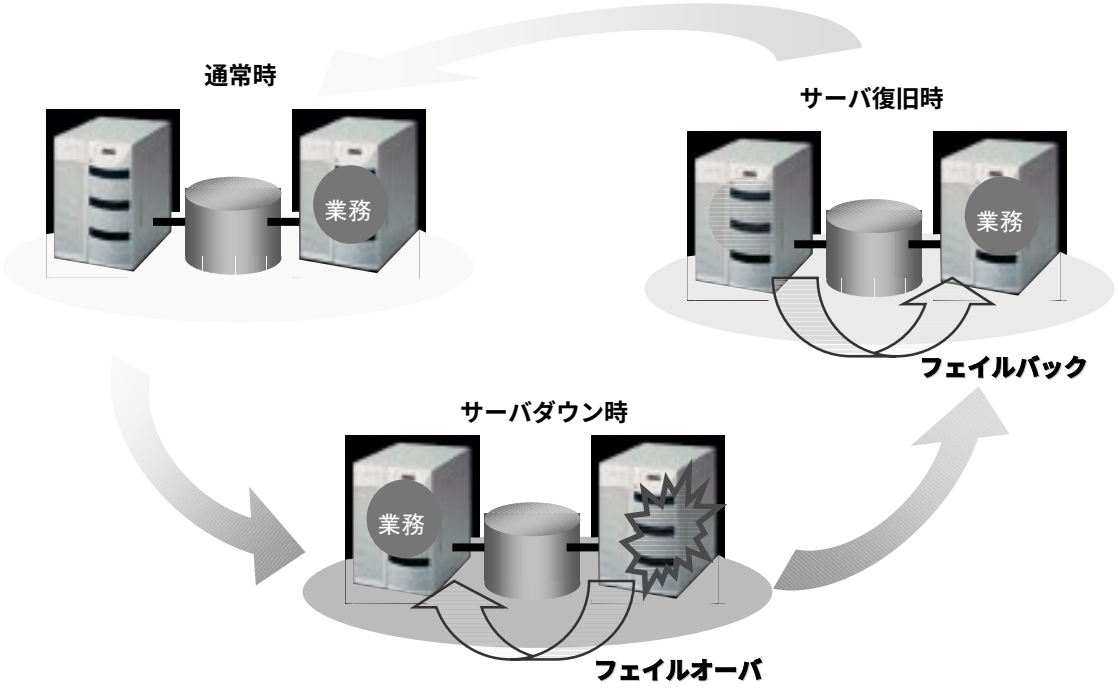
CLUSTERPRO™ の概要

CLUSTERPRO は、最大 4 台の Windows NT®サーバでシェアード・エブリシングモデルとシェアードナッシングモデルの両方を兼ね備えたウォームスタンバイクラスタを提供するミドルソフトウェアです。
CLUSTERPRO は、NEC のメインフレームやオフコンで培ったノウハウをベースに、NEC 独自の技術で、Windows NT®のクラスタシステムを実現しました。

1996年10月	ESM/ActiveRecoveryManager Ver1.0
1997年1月	ESMPRO/ActiveRecoveryManager Ver1.1
1997年4月	ESMPRO/ActiveRecoveryManager Ver2.0
1997年7月	CLUSTERPRO/ActiveRecoveryManager Ver3.0
	CLUSTERPRO/ServerNet™オプションVer3.0
	CLUSTERPRO/Oracle Parallel Server対応オプションVer3.0
1998年4月	CLUSTERPRO/ActiveRecoveryManager Ver4.0
	CLUSTERPRO/Oracle Parallel Server 対応オプション Ver4.0

現在の CLUSTERPRO のファーストバージョンは、1996/10 に ESM シリーズとしてリリースされました。Microsoft® Cluster Server が現れる 1 年以上も前のことです。
1997/7 には、世界で最初の Windows NT®版 Oracle Parallel Server と連携するクラスタ製品としてリリースし、製品名も CLUSTERPRO™ シリーズとなりました。
1998/4 リリースした Version 4.0 は、ノード数が 4 ノードまでのスケーラビリティをもった本格的クラスタシステム製品です。

CLUSTERPRO™ のアベイラビリティ機能



・クラスタのライフサイクル

クラスタ構成されたサーバは、相互にサーバ監視を行い、サーバダウンが発生した場合、待機ノードが通常運用していたサーバ上の業務を引継ぎ、業務を継続します。この引継ぎの動作をフェイルオーバと呼んでいます。双方のサーバで異なる業務を行い、互いに相手サーバの待機ノードとなる構成も可能です。また、4 ノード構成の場合、待機ノードは、残りの 3 ノードが待機サーバとなることができます。

ダウンしたサーバが復旧したのち、フェイルオーバしていた業務を元のサーバへ戻すことをフェイルバックと呼んでいます。フェイルバックによって、完全に元の状態に戻すことができますが、業務を行っているサーバが一方のみの場合は、フェイルバックを行わなくても、復旧したサーバが直ちに新たな待機サーバとなることが可能です。

・サーバ監視

サーバの監視は、サーバ間の LAN(インタコネクトと呼ぶ)を通して、ハートビートと呼ばれる通信で行います。サーバ間のインタコネクトには、クライアントが接続している LAN(Public LAN と呼ぶ)とサーバ間専用に設ける LAN があり、インタコネクトの二重化を行っています。さらに、インタコネクトの増設も可能であり、1 本のインタコネクトが切断してもサーバの監視を継続します。

CLUSTERPRO では、インタコネクトの多重障害にも備え、最終的なサーバダウンのチェックを共有ディスク接続に使用している SCSI(または Fibre Channel)を通して行っています。

・フェイルオーバ

サーバダウンが検出されると、ダウンしたサーバが使用していた資源(ディスク、IP アドレスなど)を他のサーバへ移動させ、そのサーバでアプリケーション・サービスを起動します。

移動可能な資源

ディスク	パーティション単位にサーバ間で引継ぐことができます。	
クライアント接続	TCP/IP	IP アドレスをサーバ間で引継ぐことができます。
	NetBIOS	コンピュータ名をサーバ間で引継ぐことができます。
アプリケーション・サービス	CLUSTERPRO が制御するスクリプト(バッチファイル)に起動・停止を記述することで、サーバ間で引継ぐことができます。	
WAN 回線(V.24/X.21)	回線切替え装置を使い、WAN 回線(V.24/X.21)をサーバ間で切替えすることができます。	

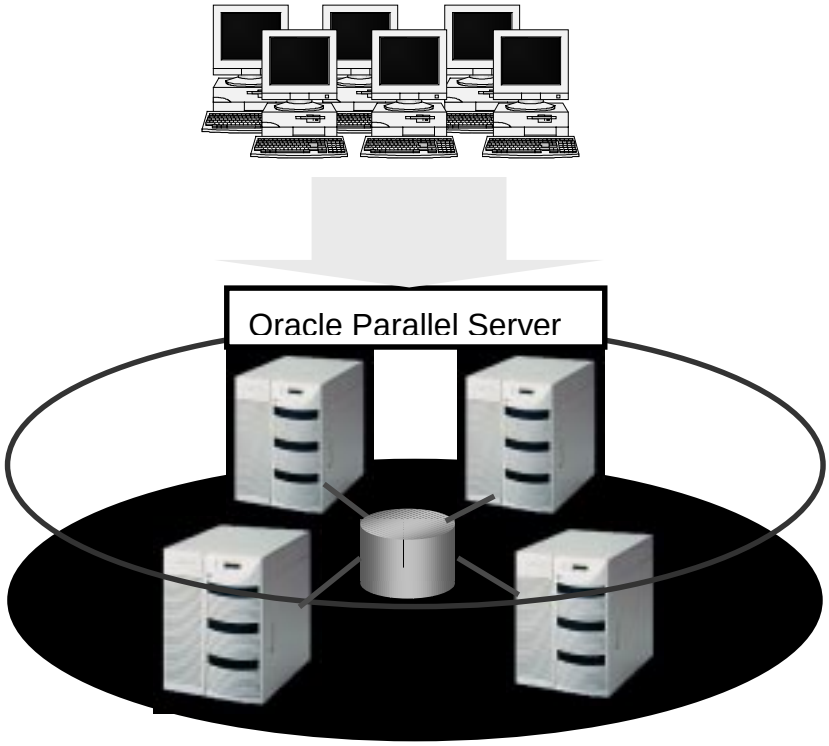
・フェイルオーバ

ダウン等のトラブルが発生したサーバは、クラスタ化されたグループから除外され、サーバの復旧作業中は、クラスタの動作に影響を与えません。

原因を取り除き復旧したサーバは、再度クラスタに組み込みます。この操作は、CLUSTERPROの管理マネージャからマウスのみの操作で行えます。

待機サーバにフェイルオーバしている業務を、フェイルバックさせる(元のサーバへ戻す)操作も CLUSTERPRO の管理マネージャからマウスのみの操作で行えます。

CLUSTERPRO™ のスケーラビリティ機能



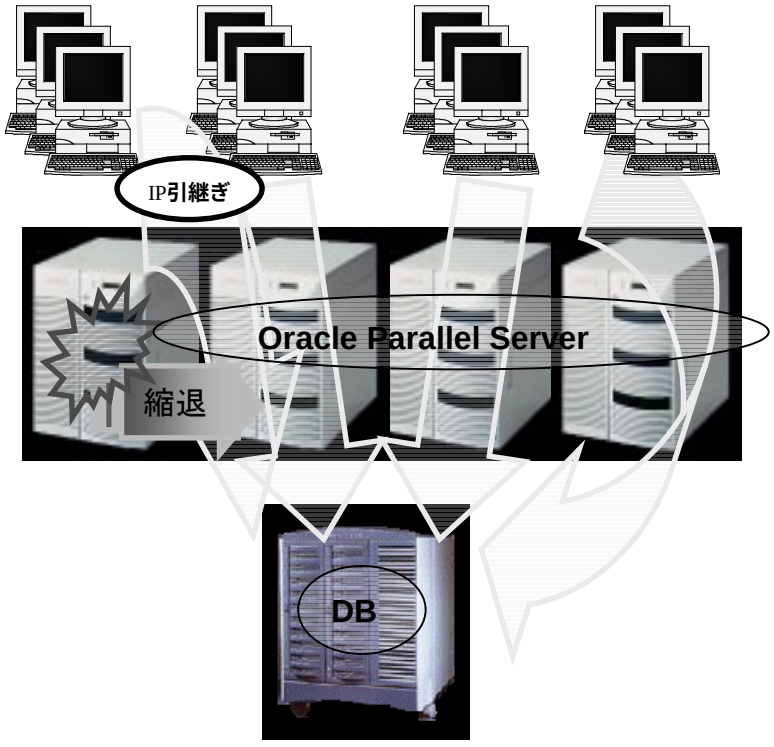
CLUSTERPRO では、Oracle Parallel Server と連携して、シェアードディスク方式のスケーラビリティを持つクラスタシステムを構築できます。このクラスタシステムでは、2 ノードから 4 ノードで同一のデータベース処理を行うことが可能で、負荷分散を行えます。

また、追加したノードを含む全ノードで同じデータベース処理が行えますので、ノード追加時に、クライアント AP を変更する必要がありません。

負荷分散動作中のサーバ障害には、Oracle Parallel Server は縮退動作を行い、CLUSTERPRO のアベイラビリティ機能が、クライアント接続を別のノードにフェイルオーバさせますので、業務が継続できます。

・通常運用時

・サーバ障害時



NT クラスタの今後

可用性と拡張性の向上に向けて進歩しつづけるクラスタリング技術は、今後どのように強化されていくのでしょうか。

・ノード数の拡大

現在、一部の 4 ノードおよび 16 ノード製品を除き、(MSCS に代表されるように)2 ノードが主流です。今後は、スケーラビリティの有効性を打ち出すために、ノード数の拡大が図られます。CLUSTERPRO は、現在 4 ノードを実現しており、次のターゲットとして 16 ノードへの拡張を目指しています。

・負荷分散機能の本格化

MSCS のフェーズ 2 では、負荷分散機能の実現が計画されています。ノード数の拡大とともに負荷分散機能が実現されなくては、クラスタの持つ 2 つの側面を満たすことができません。未だ、出そろっていない負荷分散機能にターゲットをおいた製品強化が進むでしょう。

・Fibre Channel Disk の採用

ノード数の増加に従い、共有ディスクの性能を確保するため、Fibre Channel Disk の採用が進みます。Fibre Channel は潜在能力では 100MB/Sec の転送速度を持ち、Ultra SCSI の 40MB/sec をはるかに上回る性能を実現します。

「CLUSTERPRO™」では、すでに 4 ノードを製品化し、さらに Oracle Parallel Server と連携することで、負荷分散によるスケーラビリティを実現しています。今後は Fibre Channel Disk のサポート、また、ノード間通信を高速に行う技術を採用することで、シェアード・エブリシングモデルの排他制御によるオーバヘッドの短縮、さらにシェアード・ナッシングモデルデータベースへの対応等を積極的に行い、Windows NT®クラスタの適用範囲を拡大していくことにしています。

・SCSI(スモール・コンピュータ・システム・インタフェース)
ANSI(米国規格協会)が定めた、コンピュータと周辺装置を接続するためのインタフェース。
・デュアルシステム
2 台のシステムで同期を行い、故障が発生した場合は、故障したシステムを切り離して運用を続行するシステム。
・デュプレックスシステム
2 台のシステムのうち、一方のシステムを予備とするシステム。稼動しているシステムが故障した場合、予備のシステムに切り替わる。
・Fibre Channel Disk
ANSI(米国規格協会)が定めた、コンピュータと周辺装置を接続するためのインタフェース。光ファイバーや同軸ケーブルなどを用いる。