クラスタソフトウェア DNCWARE™ClusterPerfect™

DNCWARE™ClusterPerfect™ Cluster Software

金子　哲夫　KANEKO Tetsuo
森　良哉　MORI Ryoya

PCサーバの適用業務範囲は急激に広がり，また業務のインターネット化やグローバル化に伴い，基幹系システムにも多く使われるようになった。そのため，PCサーバの予期しない障害によるダウンがもたらす影響も深刻になってきている。最近では，PCサーバの高信頼性を表現する指標に，可用性（サーバ稼働率）という基準が使われ始め，可用性の優劣がPCサーバを選択する大きなポイントとなっている。高い可用性を持ったシステムを構築するためには，障害が発生した時にダウン時間をいかに短くするかがポイントになる。

クラスタソフトウェアDNCWARE™ClusterPerfect™は，高可用性を実現するキープラットフォームであり，業界一の機能を数多く備えている。

PC servers have recently been extensively introduced in business operations and are being used in mission-critical systems.

A consequence of this trend is that the damage caused by an unexpected malfunction in a PC server can be serious. As work becomes more network-connected and globalized, the availability and reliability index of servers has recently begun to be used as the standard. High availability is the key point of quality in a PC server. MTBSD (meantime between system down) is more important for a high-availability system than MTBF (meantime between failures).

Toshiba's DNCWARE™ClusterPerfect™ cluster software is the best solution to achieve high availability, offering superior functions compared to other manufacturers' software.

1 まえがき

クラスタシステムとは，複数のサーバでサービスシステムを構成し，一部のサーバが障害を起こしてもサービスを引継いで全体を停止させないシステムのことを言う。

PCサーバが使われる業務の拡大とともに，サーバダウンによる影響度をいかに最小にするかが大きな課題になってきている。従来，信頼性について語るときに，ダウンが多いか少ないかに目が注がれがちであったが，PCサーバが顧客に提供するビジネスサービスに注目した場合，実はビジネスサービスが停止している時間の長短の方がより重要であることに気が付く。

ここでは，基本的な考え方や機能面において多くの特長を持つクラスタソフトウェアDNCWARE™ClusterPerfect™について述べるが，高可用性を実現するためにクラスタソフトウェア自身の信頼性（品質）も重要である。

2 なぜクラスタシステムか（高可用性とは）

図1に示すように，PCサーバクラスタシステムでは，あるサーバがダウンした場合に実行していた業務を他のサーバに引継ぐことにより，業務停止時間を短くする。この業務停止時間を短くすることが，可用性（業務稼働率）を高めることになる。

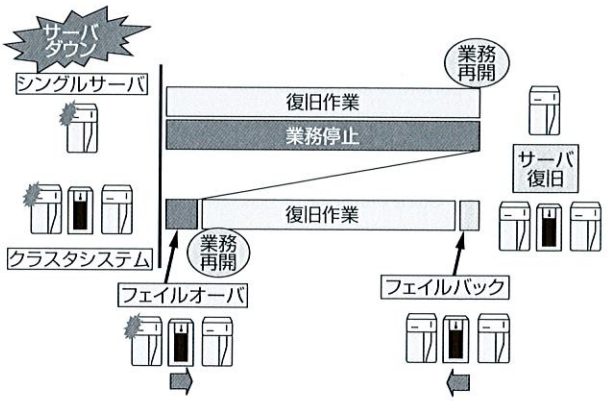


図1．クラスタシステムの効果　1台のサーバがダウンしても他のサーバに業務を引継ぐことにより，業務停止時間を大幅に短縮する。
Effect of cluster system

可用性とは「通常の使用に対してコンピュータシステムがアクセス可能な時間の割合」を言い，期待される稼働時間を100%にして，予期しない障害による業務停止時間を引いた稼働時間の率で表す。

PCサーバを使った基幹システム構築を推進するユーザーの協議会であるECA（Enterprise Computing Association）の活動を通して，高可用性99.90%が一つの目標になっている。可用性99.90%とは，24時間365日稼働のシステムで，許容されるダウン時間が約8時間以内を意味する。PCサーバ

ベンダー各社とも高可用性の実現には，クラスタ構成を中核に位置づけている。当社はClusterPerfect$_{TM}$を中心にしたロバストPCサーバとサービスを組み合わせて高可用性を実現する。

3 HAクラスタシステム動作原理

クラスタシステムの目的には高可用性と拡張性があるが，ここでは高可用性(High Availability)クラスタシステム(以下，HAクラスタと略記)の動作原理について簡単に述べる。

HAクラスタでは，まずクラスタに含まれる各サーバがお互い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。サーバ間での情報のやり取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プロセスや共有ディスク，ネットワークアドレスなどの資源管理を共同で行う。

ハートビートが途絶えたら，途絶えたサーバに障害が発生したと判断され，障害サーバの業務を含む資源を他のサーバに引継ぐフェールオーバ処理が行われる。これをサーバ連携と呼ぶ。

クラスタHAシステムの動作概念図を図2に示す。

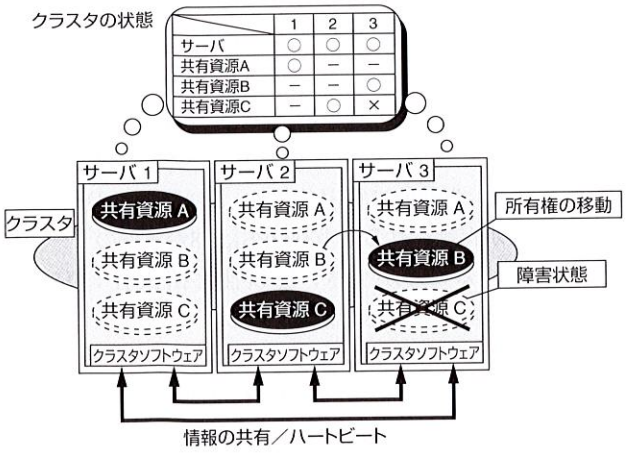


図2．クラスタHAシステムの動作概念　各サーバの状態がどのようになっているか，また各共有資源の所有権がどのサーバにあるかを，お互いに監視しながら動作する。
Basic mechanism of cluster high-availability system

4 ClusterPerfect$_{TM}$の概要

4.1 DNCWARE$_{TM}$とは

DNCWARE$_{TM}$のDNC(Distributed Nodes Cooperation)は，分散ノード連携技術を意味し，DNCWARE$_{TM}$は分散コンピューティング環境で高可用・スケーラブルシステムを構築するためのソフトウェア群である。

DNCWARE$_{TM}$の開発アプローチは図3に示すように，一般的なクラスタソフトウェアの開発アプローチとは異なり，最初に分散ノード連携アーキテクチャを確立し，そのアーキテクチャの上に市場ニーズに合わせた製品を段階的に市場投入するアプローチを採っている。これにより機能拡張がスムーズに行われ，結果として高い品質を維持しつつ機能拡張が可能になっている。

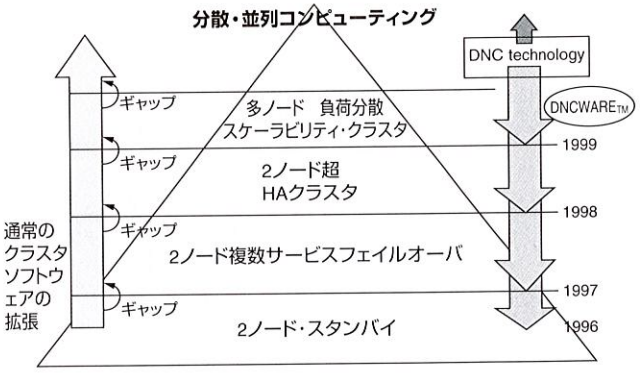


図3．DNCWARE$_{TM}$の開発アプローチ　究極のDNCアーキテクチャを順次製品化していく。
Development approach for DNCWARE$_{TM}$

4.2 システム構成

高可用性を目的としたClusterPerfect$_{TM}$の標準的な構成例を図4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で二重化されており，専用LANはClusterPerfect$_{TM}$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により，ハートビートをクライアント／サーバ間のLAN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に障害が発生してもサー

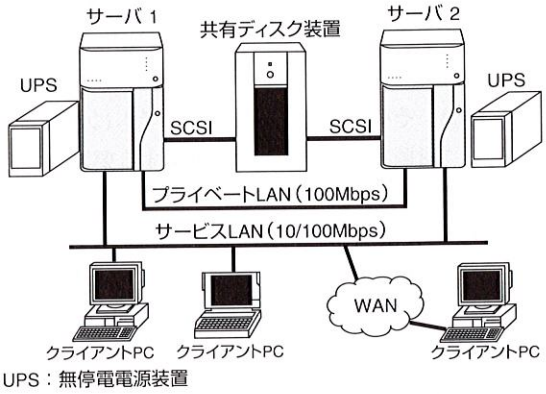


図4．ClusterPerfect$_{TM}$の構成例　2台のサーバがLANで接続され，ディスクを共有している。
Example of high-availability system configuration using Cluster Perfect$_{TM}$

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。

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する。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。

4.3 ソフトウェア構成

ClusterPerfect$_{TM}$のソフトウェア構成を図5に示す。

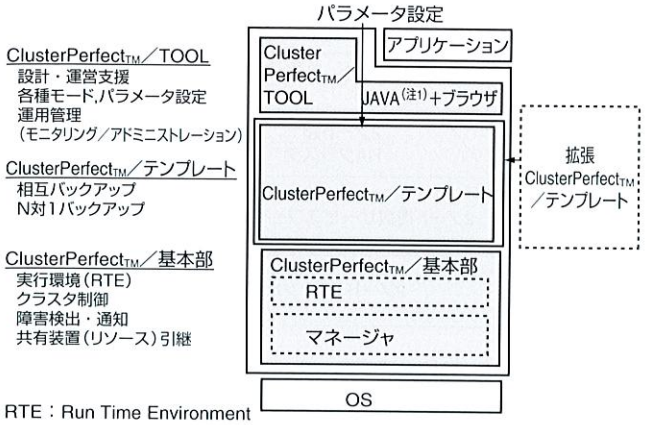


図5．ClusterPerfect$_{TM}$のソフトウェア構成　ClusterPerfect$_{TM}$の実行部は，基本部とテンプレートに分かれている。
Software configuration of ClusterPerfect$_{TM}$

ClusterPerfect$_{TM}$の構造は，TOOL（ツール），テンプレート，基本部の3層に分かれ，TOOLはクラスタシステム設計とクラスタシステム運用を支援するユーティリティ機能である。ClusterPerfect$_{TM}$の構造上の特長は，クラスタ実行制御部がシナリオテンプレート部と基本部に分かれている点にある。シナリオテンプレートとは，サーバ連携全体を制御する部分の原形で，サーバ状態遷移やクラスタ運用形態のノウハウが凝縮されている。実際にはクラスタシステム設計支援ツールを使用して，シナリオテンプレートにユーザー固有情報を付加することによりシナリオに形を変えて使用可能になる。ClusterPerfect$_{TM}$には，このサーバ連携部分で他に例を見ない特長を持つ。その特長とは，分散システムにおける“状態の一貫性”と“耐障害性”という一種相反する要素を，両立してサーバ連携を制御できている点である。例えば，あるサーバがダウンした場合，他のサーバどうしが矛盾することなく協議し，常に動作サーバの認識が一致させるようになっている。このことはClusterPerfect$_{TM}$が非常に品質の高いクラスタリングを可能にしている。

基本部は障害を検出したり，共有リソースの引継ぎなどを実行する。ClusterPerfect$_{TM}$は，テンプレートを取り換えて実装できる方法を採っているため，ユーザーが必要とするテンプレートを選択することにより，多種多様なシステム形態に対応できる。例えると，基本部はOS（Operating System），シナリオテンプレート部はユーザーアプリケーションに対応する。アプリケーションの多様性で様々なユーザー要求にこたえるように，ClusterPerfect$_{TM}$も様々なユーザー要求にこたえられる。また，シナリオを基本部より分離させたことにより，ClusterPerfect$_{TM}$自身の肥大化を防止できている。

5 ClusterPerfect$_{TM}$の展開

現在リリースしているClusterPerfect$_{TM}$は，高可用性を目的としたクラスタソフトウェアであるが，4.1節で述べたように，DNCアーキテクチャをベースに分散環境の可用性向上，拡張性向上を実現するための機能強化が順次計画され，製品化に移っている。以下，それらのうち代表的なものについて述べる。これらの機能が実現できることにより，より高い可用性を保持したPCサーバシステムの大規模化が可能になる。なお，5.1節の共有ディスクロードシェアと，5.2節の分散レプリケーションはリリース済みである。

5.1 共有ディスクロードシェア

ORACLE(注2) DB（データベース）をロードシェアするOPS（Oracle Parallel Server）をリリースしている。

OPSは，クラスタ化されたサーバからアクセスされるDB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。更に，どのサーバで障害が発生しても，そのサーバのユーザー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，DBの論理的な一貫性が保証される。当社OPSは，8台までのサーバでDB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。

5.2 分散レプリケーション

データを引き継ぐ場合，一般的には共有ディスクを構成するが，各サーバのローカルディスク上にあるファイルをレプリケーションすることにより，共有ディスクを必要としないデータ引継ぎを実現できる。多ノードのクラスタ構成でデータ引継ぎを行う場合は，分散レプリケーションが重要機能になる。

ClusterPerfect$_{TM}$の分散レプリケーションは，同期レプリケーションと非同期レプリケーションの併用による性能劣化の防止，各サーバ連携によるデータ不整合の防止といった点が考慮されている。

この機能は，アプリケーションやミドルウェアとNTFS（NT File System）(注3)の間に実装するため，一般ファイルの

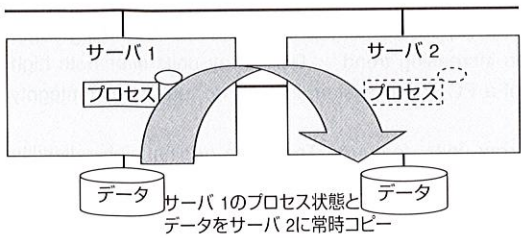
（注1）Java並びにその他のJavaを含む商標は，米国SunMicrosystems社の商標。
（注2）ORACLEは，Oracle社の商標。
（注3）Microsoft®WindowsNT®が採用しているファイルシステム。

レプリケーションやファイル単位のレプリケーションが可能になる。

5.3 高速切換え機能

稼働系サーバのデータだけでなくプロセスの状態もチェックポイントのたびに待機系に常時コピーし，稼働系がダウンした時に待機系でチェックポイントから動かす。

これをORACLEに適用したQuick Recoveryにより，通常フェールオーバが発生すると数分から30分くらい要していたORACLEのDBリカバリが不要になるため，切替えは30秒以内で完了する。クラスタにおけるダウン時間の大幅な短縮が可能になる(図6)。



ORACLEをはじめとするDBシステム障害時の切換え時間を

数十秒に短縮!

(従来は数分)

・DBのリカバリ処理不要
・データレプリケーション(複製)によるデータの一貫性

図6. サーバ高速切換え機能　データ及びプロセス状態がコピーされ，高速引継ぎが可能になる。
Basic mechanism of quick recovery

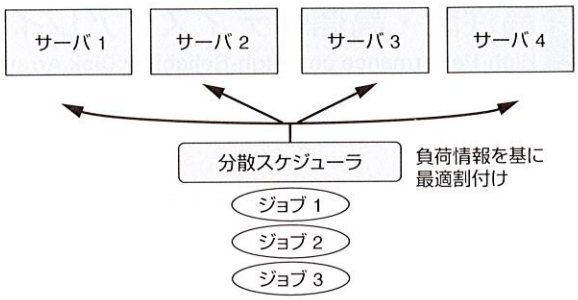
5.4 負荷分散

クラスタシステムを構成しているノード(サーバ)の構成や負荷状態を監視し，その負荷情報を基にクライアントからの要求やフェールオーバ時の業務を最適なノードに動的に割り付けることにより性能向上を図る。

また，サーバを追加するだけで容易にシステム拡張が可能である(図7)。

5.5 データバックアップ機能

ネットワークに分散された各サーバのデータをリアルタイムで，しかも簡単かつ高速にバックアップするサイトを構築できる。バックアップ対象サーバ台数やバックアップサイトの構成も自由で，OSボリュームのバックアップも可能である。高速インターコネクトの採用により高速なバックアップも可能



動的負荷分散による性能向上!

・システム構成／負荷に応じて,最低なノードにジョブを割付け
・サーバを追加するだけで容易にシステム拡張可能

図7. 負荷分散機能　サーバ構成やサーバ負荷により業務を振り分ける。
Basic mechanism of load balancing

である。システム運用面ではバッチ形式のバックアップ時間を設ける必要がなくなり，また世代管理の考えが導入されているので，アプリケーションの任意の時点までさかのぼってファイル回復が可能である。

6 あとがき

ユーザー要求により，障害時のダウン許容時間は異なるが，高可用性システムを実現するためにはクラスタ構成が前提となることは間違いない。ClusterPerfect™は，今後もお客様の期待にこたえる最良のクラスタソフトウェアとして発展強化し，提供していく所存である。

金子　哲夫　KANEKO Tetsuo
デジタルメディア機器社 青梅工場 ミドルウェア設計部参事。クラスタシステムをはじめミドルウェアのプロダクトマネージメントに従事。情報処理学会会員。
Ome Operations

森　良哉　MORI Ryoya
デジタルメディア機器社 コンピュータ＆ネットワーク開発センター 開発第三部グループ長。
分散ノード連携ソフトウェア技術の研究・開発に従事。情報処理学会，IEEE会員。
Computer & Network Development Center