高可用性（HA）システム技術と動向

Overview of and Trends in High-Availability System Technologies

江口　和俊　K. Eguchi
森　良哉　R. Mori

オープンサーバによる基幹業務システムの構築が急激に増えている。オープンなプラットフォームやミドルウェアによって，クライアント／サーバ型アーキテクチャの分散処理を行うコンピューティング環境において，いかにシステムとして高い可用性（HA：High Availability）を実現するかが重要な技術となっている。オープン環境でのシステムの可用性を高めるため，分散化と多重化を組み合わせたクラスタ型HAシステム技術が主流となっている。“Easy Engineering/Simple Operation”を指向する当社の技術的取組みを紹介する。

Open servers are now in widespread use in many mission-critical application systems. In an open, client/server distributed environment, one of the most important technologies is how to construct a high-availability (HA) system.

This paper provides an overview of high-availability clustering technologies and summarizes the trends in this field, especially in relation to our “easy engineering/simple operation” technology development policy.

1 まえがき

計算機システム（特に基幹業務システム）では，そのシステムがユーザに提供する機能はもとより，そのシステムの可用性（Availability）が重要となる。

可用性とは，特定の時間区間にシステムが稼働している時間の割合を言う。可用性は，平均故障間隔時間（MTBF），平均故障修理時間（MTTR）を用いると次のように表すことができる。

可用性＝MTBF／(MTBF＋MTTR)

また，システムとして考えた場合は，

可用性＝実際に正しくサービスを提供する時間
　　／サービス提供が期待される時間

となる。

計算機システムは，従来のベンダ固有の環境から，オープンなプラットフォームやミドルウェアで構築され，クライアント／サーバ型のアーキテクチャをとり，集中処理から分散処理に変化してきている。このような新しいオープンなコンピューティング環境で，いかにシステムとして高い可用性を実現するかが重要な技術となっている。ここでは，オープンサーバでのHAの技術動向について述べる。

2 高可用性（HA）システム技術とは

オープンサーバでは，コンポーネントの信頼性を越えて，業界標準ハードウェア／ソフトウェア上で，目的とするサービスの提供をシステム全体で保証するために，HAシステム技術が必要となる。

システム全体として可用性を高めるために，多くの場合に適用できる根本原則は“分散化”と“多重化”である。

まず，ネットワークコンピューティング，“分散処理”の広がりによる“分散化”自体が，単一障害の影響をシステム全体に及ぼしにくくするという点で，システムの可用性を高めるのに有効である。しかし，障害発生部分のサービスは停止してしまう。もう一つの重要な原理が“多重化”である。すなわち，同じサービスを提供する機能を冗長的にもつことにより，可用性を高めることが可能となる。

付加的なコストを押えてシステムの可用性を高めるのが，“分散化”と“多重化”を組み合わせたクラスタ型HAシステムである。クラスタとは，複数のサーバをネットワーク

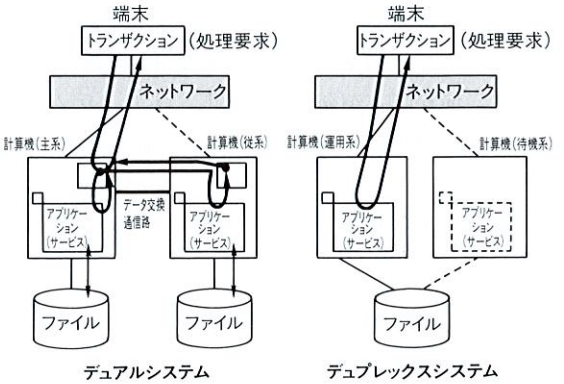

図1．デュアルシステムとデュプレックスシステム　デュアルシステムでは主系，従系で同一処理を行い，障害が起きたときに障害発生系を切り離す。デュプレックスシステムでは運用系で障害が起きたときに待機系に切り換える。

Dual system and duplex system

や高速通信路によって結ぶ，あるいはネットワーク分散された複数のサーバを連携するように制御して，一つのシステムとして機能を実現し運用することを言う。クラスタのタイプには，サーバを追加することによりスケーラビィリティをもたせ，負荷分散，並列実行により性能向上をねらうタイプと，クラスタ内のあるサーバで障害が発生したときに他のサーバがバックアップして，業務を継続する可用性向上をねらうタイプがある。

クラスタ型HAシステムはこの後者に分類される。

多重化による高信頼／高可用システムは，1960年代から大・中規模データ処理システム，産業用制御システムなどで，各社ベンダ固有の汎(はん)用計算機，ミニコンピュータで構築されてきた。2系列の処理系を組にしたデュアルシステムとデュプレックス システムが基本型である（図1）。

デュアルシステムでは，主系，従系で同じ処理を行い，正常時は両系の処理結果を比較チェックし，主系から処理結果を処理要求元に返しているが，どちらかの系に障害が発生したときは，障害発生系を切り離してサービスを継続する。いわゆるフォルトトレラント コンピュータに類似したシステム構成である。障害発生時からサービスを再開，継続するまでの中断時間を非常に短くすることができるが，同期制御などが難しく，アプリケーション システム構築が複雑となり，コストも高くなる。

デュプレックス システムでは，正常時は運用系で処理をし，待機系は障害に備えて待機している。運用系に障害が発生したときには待機系に切り換えられる。待機系は，正常時には他の優先度の低い業務に用いることもでき，使用効率が高く，アプリケーション システムの構築が容易でコスト効率も良い。ただし，切り換えのためにデュアルシステムに比較して中断時間が長くなる。

クラスタ型HAシステムは，現在，デュプレックス システム型のものが主流で，実用実績が多い。クラスタ型HAシステムのキー技術には，障害監視・検出・通知技術，クラスタ状態制御技術，障害対処制御技術，HAシステム運用管理支援技術，HAシステム構築支援技術がある。

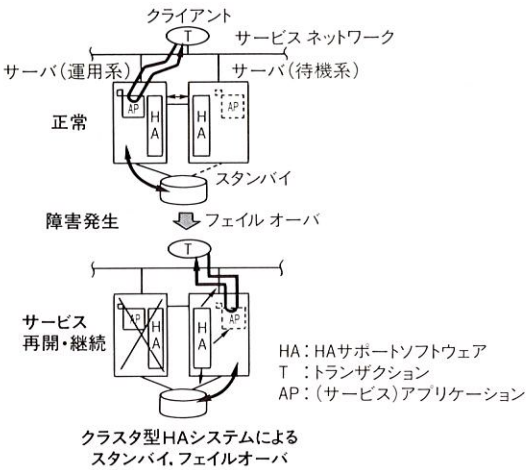

図2．HAサポート製品（クラスタ型HAシステム）　可用性を高めることを第一のねらいとしたクラスタリング技術の応用製品で，デュプレックス システム型である。

HA support system

3 クラスタ型HAシステム

3.1 クラスタ型HAシステムの動作

クラスタ型HAシステムはデュプレックス システム型のHA技術であり，図2のように運用系（サービス中）と待機系からなる。待機系は運用系の障害に備えてスタンバイしており，ハートビート（待機系が運用系の脈を取るイメージ）と呼ばれる通信により運用系を監視している。運用系で障害が発生したとき（脈がなくなったとき）に，共有ディスクやネットワーク識別子などの環境を待機系に引き継ぐ。この引継ぎはフェイルオーバと呼ばれる。フェイルオーバの後，待機系でサービスを提供するアプリケーションを起動してサービスを再開する。

待機系は，スタンバイしているときに他の業務，サービスを行っていてもよく，運用系が複数あり，相互に待機系の役割をする相互バックアップができるのが現在では一般的である。

クラスタ型HAシステムでは，運用系の障害時に処理をバックアップするという観点からは，待機系は何もしないで完全にスタンバイしている形態がHAシステムの信頼性がもっとも高い。また，他系の障害を監視するという観点からは，2台の系が互いに相手系が障害を発生していると判断して，勝手に動いてしまう現象（スプリット ブレインと呼ばれる）を避けるために，他系を強制停止させるとか，共有ディスクへの誤ったアクセスをプロテクトするなどの複雑な処理が必要となるため，3台系（奇数の系）で監視し，多数決判断をするのが好ましい。

保守などのために，オペレータからの指示でサービスを運用系から待機系に切り換える機能をスイッチオーバと呼ぶ。

障害を起こした系が復旧したときに，サービスをバックアップ系から引き戻す場合もスイッチオーバあるいはフェールバックと呼ぶ。この引戻しは自動で行う場合と手動で行う場合があり，システムごとの運用形態により選択する。

3.2 HAシステム形態

クラスタ型HAシステムを構成する場合のシステム形態について述べる。

バックアップの形態からみると次のようになる（図3）。

(1) スタンバイ型　待機系は何もしないで，もっぱら

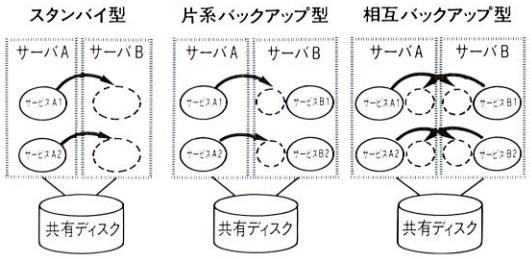

図3．バックアップ形態。　アベイラビリティ クラスタでHAシステムを構成する場合の バックアップ形態

Backup style

運用系の障害に備えている。待機系の使用効率は悪いが，バックアップ後に業務サービスの縮退がない。

(2) 片系バックアップ型　　重要度の高い業務サービスを提供する運用系に対し，待機系は優先度の低い業務サービスを行っている。運用系に障害が発生した場合，待機系は優先度の低い業務サービスを停止，あるいは縮退して，優先度の高い業務をバックアップする。

(3) 相互バックアップ型　　各系が運用系であると同時に待機系でもあり，他系の障害に備えている。どの業務サービスを優先してバックアップするかは，システム設計時に決めておく。

(4) N 対1バックアップ型　　スタンバイ型あるいは片系バックアップ型の組合せで，運用系が N ノードあり，待機系が1ノードの場合（図4）。

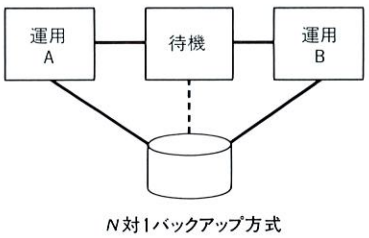

図4．N 対1バックアップ形態　　運用系が N ノードあり，待機系が1ノード。図は2対1バックアップの例を示す。

N to 1 backup style

ディスクデータの引継形態からみると次のようになる（図5）。

(1) 共有ディスク型　　共有ディスクによりデータを引き継ぐ。ただし，クラスタ型HAシステムでは，運用系からだけ直接アクセスできる。フェイルオーバ後，待機系からアクセスする。共有ディスクであるが，複数のノードからの同時，直接アクセスはないので，ノンコンカレント シェアード方式とも呼ばれる。

(2) ディスクのリモート ミラー型　　各サーバのローカルディスクにデータを置き，運用系のデータを待機系に通信路を介して送り，待機系にデータのコピーを置いておく。つまりミラーリングしておく。共有データをもたないので，シェアード ナッシング方式と呼ばれる。

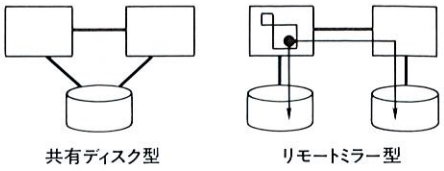

図5，ディスクデータの引継ぎ形態　　共有ディスク型は共有ディスクによりデータを引き継ぐ。ただし，運用系，待機系からの同時アクセスはできない。リモートミラー型は通信路によりデータを待機系にミラーリングする。

Two types of disk data takeover

4 技術動向と展望

70年代までは，アプリケーション ソフトウェアでつくりこんで，高可用性システムを構築していたが，80年代には，クラスタサポート，複合系サポートといった，各ベンダのベンダ固有のプラットホーム上でのHAシステム構築を支援する製品が出てきて，それを使ってシステム構築が行われるようになった。

当社は，産業用コンピュータのTOSBAC$_{TM}$-7/70シリーズ，TOSBAC$_{TM}$-G8000シリーズやVL2000シリーズで共有メモリを使った密結合型の複合系サポート製品を提供して，2台系から8台系までの大規模な制御システムを発電，電力，産業の各分野で数多く構築してきた。また，オフィスコンピュータでもアドオンの形で，HAシステムを事務処理システムに適用してきた。

90年代には，オープンなプラットホームでHAシステム構築をサポートする製品が主にプラットホームベンダから提供できるようになり，現在ではサードベンダもHAサポートソフトウェア製品を提供している。96年までに，主なUNIX(注1)サーバベンダといくつかのサードベンダがアベイラビリティ クラスタによるHAサポート製品を出しており，かなりの実用実績がある。

当社は，UNIXサーバのUXシリーズにおいて，95年からHAサポート製品を提供し始め，多くのシステムに適用している。

HAシステムを構成する最大サーバ数は4ノード，8ノードの製品が多い。実際のシステム適用は2ノードスタンバ

イのものが主流であるが，2対1，3対1バックアップシステムも増えてきた。

フェイルオーバにかかる時間は，アプリケーション，データベースのリカバリ時間にもよるが，数十秒から数分である。

共有ディスクの引継ぎは，最大ノード数に合わせて，4ノード，8ノードからの接続のものが多い。ネットワーク識別子の引継ぎは，IP (Internet Protocol) アドレスの引継ぎはすべての製品でサポートされており，MAC (Media Access Control) アドレスの引継ぎを行うものもある。引継時の対応処理は，サーバごとに用意されたスクリプトにより行うようになっている。

一方，基幹業務システムをPCサーバで構築するケースも増えてきたために，WindowsNT® (注2)サーバのクラスタリングが急に注目を集めるようになった。UNIXサーバでアベイラビリティクラスタを提供していたベンダがその技術をベースに2台系を中心に提供し始め，97年初頭には主なPCサーバベンダが製品をそろえた。技術的には，UNIXサーバやプロプライエタリ サーバでのHAシステム技術の移行が中心である。

当社も96年末にWindowsNT® サーバであるGSシリーズの2台系を発表している。これは最大4台系まで対応可能である。

現在，オープン環境での実アプリケーション システム構築では，UNIXとWindowsNT® を適材適所に使い分ける必要がある。しかし，HAシステム構築，運用の立場からみると，共通の見えかたをすることが望ましい。

当社は，“Easy Engineering/Simple Operation”を指向し，正確にかつ極力シンプルにHAシステム構築，カスタマイズができるように新しい技術を開発，製品化している。

それは，ノードごとにスクリプトを作成するのではなく，システム全体でのHAの動きを記述し (HAシナリオと呼ぶ)，それを各ノード用のスクリプトに変換する方式である。変換時に設定や記述の誤りをチェック (シンタックス チェック) し，並行動作する各ノードの動きをシミュレーション チェック (セマンティクス チェック) する機能，各ノードへ配布する機能も備えている。

これらの機能をHAシステム構築支援ツールとして提供しており，Java (注3)言語とブラウザベースのGUI (Graphical User Interface) により，HAシステム設計時の各種設定，引継ぎサービスの指定が行える (口絵参照)。さらに，HAシステムの運用管理ツールのGUIも共通化されている。この結果，HAシステム構築のための設計，設定，および運用管理はパソコンあるいはワークステーション のいずれのクライアントでも可能となっている。

HAシステム技術の今後の方向としては，ネットワーク分散された異機種サーバ連携によるバックアップ，高速なサーバ間通信路 (インタコネクト) によるクラスタ共有ファイルなどの実用性向上や，ネットワーク直結型データサーバの連携バックアップ，シングルシステムイメージでのHAシステム構築／運用管理の実現が考えられる。

5 あとがき

UNIXサーバやWindowsNT® サーバなどのオープンサーバによる基幹業務システム構築が増えつつある。

当社は，産業用コンピュータによる制御システム構築やオフィスコンピュータによる業務システム構築で培ってきたHAシステム技術を継承しながら，UNIXサーバにおいてオープンサーバにフィットした新たなHA技術を実現し，さらにその技術をWindowsNT® サーバにも生かしている。

今後も，オープンな時代に，マルチプラットフォームで共通かつ高度なHAシステム技術の研究・開発，商品化で対応することにより，ユーザの要求に合ったソリューションを提供しつづけていく所存である。

(注1) 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。
(注2) WindowsNTは，Microsoft社の商標。
(注3) Javaは，SunMicrosystems社の商標。

江口 和俊 Kazutoshi Eguchi

情報・通信システム技術研究所 開発第三担当グループ長。コンピュータ応用システム基盤技術の研究開発に従事。情報処理学会，IEEE会員。
Information & Communications Systems Lab.

森 良哉 Ryoya Mori

情報・通信システム技術研究所 開発第三担当主査。高可用性システム技術の研究開発に従事。情報処理学会，IEEE会員。
Information & Communications Systems Lab.