UNIXサーバの高可用性（HA）システム技術

High-Availability System Technology on UNIX Server Computer

末永　司　T. Suenaga
香川　弘一　K. Kagawa

近年，UNIX(注1)サーバの性能や信頼性が飛躍的に伸び，企業の基幹システム構築などへの採用の動きが高まっている。当社は，UNIXサーバを用いてオープン環境で高可用性（HA：High Availability）を実現するミドルウェアとして，“複合系サポートソフト”を提供し，他の高信頼ミドルウェアと組み合わせて多くの実績を上げている。オープン環境でのHAシステム構築技術のポイントとして，①確実・迅速な障害検知機構の実装，②ユーザアプリケーションとの独立性の確保，③多様なシステム形態への対応，④簡単な環境設定，運用監視，⑤各種ミドルウェアとの密接な連携，⑥ノウハウの蓄積とユーザ支援などを考慮している。

This paper describes a technology for implementing a high-availability (HA) system on an open system. Many companies have recently been building mission-critical systems on UNIX computers, because of the improved reliability and availability of UNIX servers. We have developed an HA middleware on a Toshiba UNIX server. This HA middleware provides a development and management environment for a failure takeover system with other Toshiba middleware.

1 まえがき

近年，UNIXなどのオープンなコンピュータ環境を用いて，企業の基幹システムを構築する動きが高まっている。それに伴い，基幹システム構築の要件である高可用性を実現する技術がますます重要となっている。しかし，オープン環境では従来の汎(はん)用機やミニコンのように，ハードウェアやオペレーティングシステム（OS）だけでの高可用性の実現は限界があるため，プラットホームにできるだけ依存しないミドルウェアで，信頼性や可用性（Availability）を高める方式が一般的となっている。

当社では，Sun(注2)社から提供されるUltra Enterprise Server（UX7000シリーズ），およびUltraSPARC(注3)チップを使用した自製サーバ（UX2000シリーズ）の，高信頼，高性能なサーバコンピュータを2台疎結合し，障害発生時に業務処理を引き継いで運用を継続する機能（Failover機能）を実現する，HAシステム構築ミドルウェア“複合系サポートソフト”を提供し，他の高信頼ミドルウェアと組み合わせて数多くの実績を上げている。

2 複合系サポートソフトの仕組み

複合系サポートソフトを用いると，図1のような二重化構成のシステムを構築できる。ハードウェアとしては同一機種・仕様（CPU数，メモリ容量）のサーバ計算機を2台，共有ディスク，計算機間の監視パスとして当社独自のRAS（Reliability, Availability, and Serviceability）機能カード，およびLANインタフェースカードで直結する。

おのおののサーバ上の複合系サポートソフトの監視デーモンは，二重化された監視パスを用いて相手計算機を監視する。

稼働中のシステムに異常を検出した場合には，今まで待機していたサーバ上で，あらかじめ決められたスクリプトに従って，自動的にディスクやネットワークアドレス（IP：

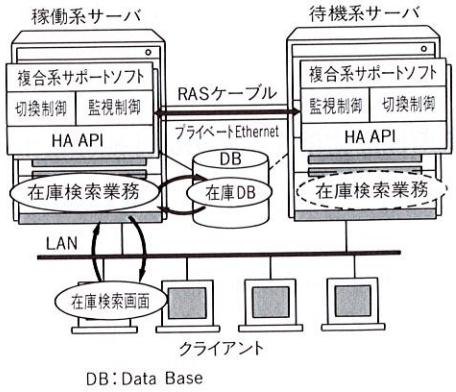

図1．HAシステムの構成例　在庫管理アプリケーションによるHAシステムの構成例。クライアントは，稼働系サーバ上のアプリケーションにアクセスする。

Configuration of HA system

(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) Sunは，SunMicrosystems社の登録商標。
(注3) SPARCは，SPARC International Inc.の米国およびその他の国における商標または登録商標。

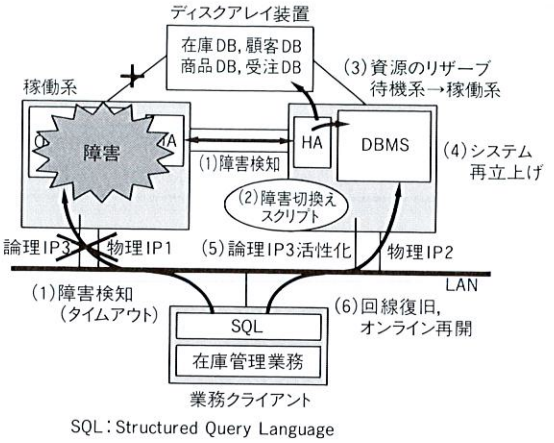

図2．HA切換処理の流れ（障害発生による切換）　HAは障害の発生を検知し，スクリプトに従って待機系でシステムの再立上げを行う。

Flow of takeover

Internet Protocol）などの共有資源の引継ぎ処理や，ユーザシステムの再起動，復旧処理を行う（図2）。

また，保守のために，手動操作によって，稼働系をシャットダウンし，意図的に待機系に処理を移動するという使いかたも可能である。

3 高可用性を実現するための仕組みとポイント

オープン環境で，HAシステムの構築基盤を提供するにあたり，当社では次の六つのポイントを考慮している。

(1) 確実で迅速な障害検知機構の実装
(2) ユーザアプリケーションとの独立性の確保
(3) 多様なシステム形態への対応
(4) 簡単な環境設定，運用監視
(5) 各種ミドルウェアとの密接な連携
(6) ノウハウの蓄積とユーザ支援

3.1 確実で迅速な障害検知機構の実装

複合系サポートソフトでは，誤った障害検知による切換処理を発生させることは決して許されない。そこで，UNIXサーバの提供する標準のハードウェアやOSの上に，複合系サポートソフトを実装するにあたり，次の三つの技術施策を実行した。

(1) 異なる監視パスでの確実な障害監視　相手計算機の障害を，専用のプライベートEthernet(注4)と，当社で開発したRAS機能カードという異なる方式の監視パスで確実に監視できる（図3(a)）。

(2) アプリケーション監視モニタによる監視　システム内のアプリケーションやミドルウェアを監視する機能を，マルチスレッドで実現できる（図3(b)）。

(3) HAデーモンプロセスのページ固定化　CPUやメモリ割当てのために処理の遅延が発生し，障害検知を誤らないように，HAデーモンプロセスをリアルタイムプロセスとして定義し，使用するメモリのページ固定化を実施する。

また，HAシステムでは，切換時間をできるだけ短縮する必要がある。しかし，オープン環境であるために，従来のミニコンや専用機などと比較して，一般的に切換え完了までの時間が長くなってしまう。

通常，HAの切換処理は，三つのフェーズで処理が進められる（図4）。

(注4)　Ethernetは，富士ゼロックス㈱の商標。
(注5)　Solsticeは，米国SunMicrosystems Inc.の商標。

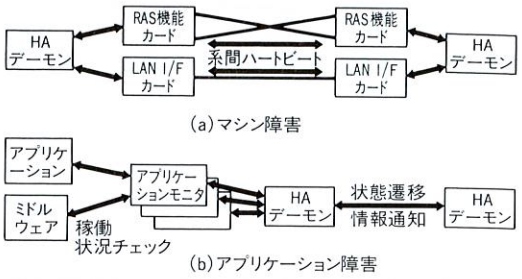

図3．障害監視機構（マシン障害およびアプリケーション障害）　(a) RASカードとLANカードの二重化された監視パスで，相手計算機の状態を監視する。(b)アプリケーションモニタは，各種アプリケーションの稼働状況を監視し，問題発生時には状態遷移を発生させる。

Machine failure monitoring (top), and application failure monitoring (bottom)

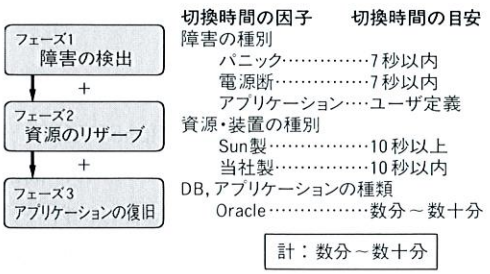

図4．HA切換えのステップと所要時間　HAシステムの切換時間は，三つのステップの因子により決まる。

Steps of takeover

(1) フェーズ1：障害の検出　ハートビート間隔の短縮により，約7秒以内での検出を可能とした。

(2) フェーズ2：資源のリザーブ　切換後のファイルシステムの整合性チェック処理の短縮に，ボリューム管理ツールであるSDS（Solstice(注5) Disk Suite）のジャーナル機能を利用した。

(3) フェーズ3：アプリケーションの復旧　現在，切換時間にもっとも大きな影響を与えるものである。例えば，データベースを使用する場合に待機系でダウンリカバリ処理を行って再立上げを行う必要があるため，切換時間にかなりの時間を費やしている。

したがって，短時間での切換えを要求するシステムでは，アプリケーションの作りかたをくふうする必要がある。

3.2 ユーザアプリケーションとの独立性の確保

HAシステムであることは，ユーザアプリケーションにはできるだけ隠蔽(ぺい)することが望ましい。当社の複合系サポートソフトは，障害発生時に待機系でアプリケーションの再起動を行うため，ユーザアプリケーションは，二重系の構成であることを意識してプログラミングする必要はない。障害発生時のスクリプトの定義，データベースやロードモジュールなどの共有資源を共有ディスク上に配置するなど，HA構成に合わせた環境設定を行うだけである。

また，クライアントアプリケーションは，障害発生時にネットワークが切断されることになるが，待機していたサーバで同一ネットワークアドレスが活性化されるので，同じネットワークアドレスで接続を繰り返せば，切換え処理の完了後，同一システムに接続することができる。

3.3 多様なシステム形態への対応

HA構成を制御するためのAPI (Application Program Interface) を公開することにより，ユーザの要望に合わせたHAシステムの構築ができる。

複合系サポートソフトにより，次の四つの2台系HAシステムを構築できる。

(1) ディスク非共有型片系スタンバイ
(2) ディスク非共有型相互スタンバイ→ゲートウェイシステムなど。切換時間7秒以上
(3) ディスク共有型片系スタンバイ
(4) ディスク共有型相互スタンバイ→オンラインシステムなど。切換時間1分以上

相互スタンバイとは，おのおののサーバで異なるシステムを実行し，障害発生時は相互に相手のシステムを引き継ぐ構成である。しかし，障害発生時に複数のシステムが片系に縮退して実行することになるので，二つのシステムを処理するだけの資源や性能の見積りが必要となる。

3.4 簡単な環境設定，運用監視

3.4.1 環境設定　HAシステムの環境設定ミスによる，障害発生時の切換え失敗を防止するため，環境設定をマクロ化，コマンド化して標準部品として提供している。特に，次の三つの監視スクリプトを提供する。

(1) データベース　Oracle(注6)のデーモンの監視から，ユーザテーブルへのアクセスまで3レベルの監視機構を提供する。
(2) LANカード　指定されたLANカード故障を，ネットワークに負荷をかけずに監視する。
(3) ユーザ指定プロセス　ユーザが指定した特定プロセスの生存をチェックする。

3.4.2 運用管理　HAシステムの監視も非常に重要であり，次の三つの監視機能を提供する。

(1) GUI (Graphical User Interface) 監視 (図5)　ネットワークで接続されたワークステーションから，GUIでHAシステムの各種資源 (ディスク，監視パス) や，アプリケーションの稼働状況を監視する。
(2) HA管理コマンド　状態表示，構成制御を行う。
(3) TME (Tivoli Management Environment) との連携
複合系サポートソフトの出力するシステム状態の記録 (syslog) メッセージを，イベントアダプタ経由で収集し，集中コンソール上で，障害のレベルに合わせた表示を行う。

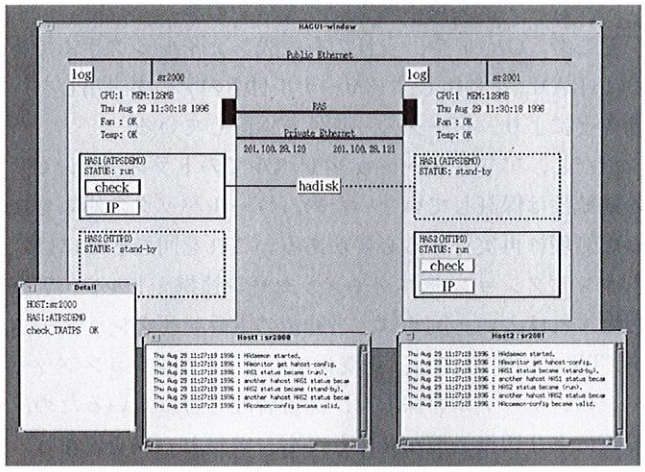

図5．監視GUI　HAシステムの稼働状況をGUI画面で監視する。
Example of HA system monitoring display

3.5 各種ミドルウェアとの密接な連携

複合系サポートソフトは，障害時の切換え機構を提供するが，基幹システムを構築する場合，データベース管理システム (DBMS) などの高信頼ミドルウェアと組み合わせて，実行環境の可用性をさらに高めることが望ましい。また，開発環境から運用環境まで，統合的にサポートすることも重要である。

そこで，複合系サポートソフトと各種ミドルウェアとの連携による，基幹システム構築を支える統合ミドルウェア環境を実現している (図6)。

(注6) Oracleは，Oracle Corporationの商標。

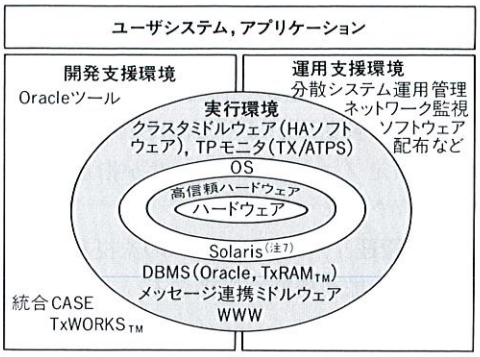

図6．基幹システム構築を支えるミドルウェア群　HAソフトウェアをベースに，各種ミドルウェアの組合せにより基幹システムを構築する。

Middleware of HA system

また，ミドルウェアの選択により，障害復旧のレベルをさらに向上させることができる。

例えば，ユーザが作成するアプリケーションが，UFS (Unix File System) にアクセスしている場合，障害発生後のファイルシステムの整合性は，UFSのチェックにより保証されるが，アプリケーション側で物理ディスクへの書出しを行わない場合，データロス発生のおそれがある。

そこで，Oracleや，当社の有編成ファイル システムであるTxRAM™を利用し，ジャーナル付きのデータ書出しを行うことにより，データの一貫性を保証している。

ただし，アプリケーション レベルでのトランザクションの継続性は保証していないため，ロールバックの対象となった処理は再実行する必要がある。これを回避するには，トランザクション モニタやメッセージ連携ミドルウェアを利用すればよい。例えば，当社製のトランザクション モニタ (TX／ATPS) では，送受信するトランザクション メッセージのジャーナルを採取し，通番管理を行っているため，障害により切換処理が発生しても，メッセージを復元し，脱送や重送がなく処理を継続できる。

3.6　ノウハウの蓄積とユーザ支援

当社では，以上のような複合系サポートソフトをベースに，各種ハードウェア，ミドルウェアとの組合せ検証や，ユーザ支援を実施し，ノウハウの蓄積に努めている。

HAシステムの構築において，特に検討する必要のあるポイントを次に示す。

(1)　システム設計　　ネットワーク機器が，サーバのネットワーク アドレスの切換えに対応しているか，採用するミドルウェアがHA構成で動作可能かの確認

(2)　業務分析　　運用形態の複雑化への対応

(3)　性能見積　　相互スタンバイで縮退運転中の性能予測

(4)　アプリケーション設計　　クライアントから見た場合の障害処理方法

(5)　実機評価　　二重化システムおのおのへの環境設定とプログラム登録。構成制御を意識した評価の実施。

4　システム事例

複合系サポートソフトは，次のようなミッション クリティカルなシステムで実際に利用され，システムの可用性向上に大きく貢献している。

(1)　列車の座席予約システム

(2)　ドキュメント管理システム

(3)　メッセージ受配信システム

(4)　ホスト計算機へのゲートウェイ システム

5　あとがき

以上，UNIXサーバにおける高可用性技術について述べた。今後さらにユーザサポートで得たノウハウや，新しい高信頼新技術を製品へ反映し，環境設定の容易化や，n台系システムなど，システム構成の自由度を高めていく。

(注7)　Solarisは，SunMicrosystems Inc.の商標。

末永　　司　Tsukasa Suenaga

府中工場 電算機ソフトウェア部グループ長。
基本ソフトウェアの開発・設計に従事。
Fuchu Works

香川　弘一　Koichi Kagawa

府中工場 電算機ソフトウェア部主務。
高信頼ミドルウェアの開発・設計に従事。
Fuchu Works