PentiumProサーバによるスタンバイシステム[1]

1U－9

木村俊之、山本整、水野正博、松本利夫、魚住光成、高橋伸一[2]
三菱電機株式会社、三菱電機東部コンピュータシステム

1　はじめに

PentiumProサーバは、基幹業務やイントラネット、生産管理のサーバとして適用されている。こうした背景のもと、可用性向上を目的に、電源装置、ディスクの冗長化等が行われてきた。今回、サーバ本体の冗長化を図る目的でスタンバイシステムを開発した。このスタンバイシステムは、2台のPentiumProサーバを拡張ディスクキャビネットに接続する構成をとった。異常の監視は、サーバ内蔵のサーバ管理装置と、OS上の監視プログラムで行う。このスタンバイシステムは、異常発生時、OS、アプリケーションの定義情報、データベース、ネットワークアドレスまで待機系に引継ぐ。このため、市販のアプリケーションもスタンバイシステムでそのまま適用できる。

2　サーバ切替の概要

2台のサーバは、業務モード、待機モード、故障モードの3つのいずれかのモードをもっている。また、このシステムはコールドスタンバイ構成をとっており、1台が業務モードの時には、残る1台は必ず待機モードとなっており、業務は実行していない。業務モードのサーバ（以下業務サーバと呼ぶ）は故障を検知すると、故障モードとなる（以下故障サーバと呼ぶ）。一方、待機モードのサーバ（以下待機サーバと呼ぶ）は業務モードに切り替わって、業務サーバになりかわり、業務システムが起ち上がることになる。

例えば、DB処理中にサーバが切り替わったとしても、DBロールバック処理は、サーバが切り替わる前と後では物理的にも同じDBに対して行われることとなり、サーバの外側（例えばネットワーク上のクライアント）から見れば、単にサーバがリブートした時と同じ状態で復帰する。

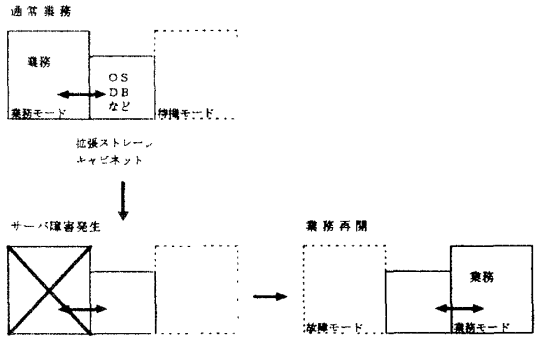

図1　サーバ切替の概要

3　実現のための課題

このように、業務をサービスするサーバは、同時には常に1台としている。これはサーバ本体自体の冗長性を追求し、なおかつ、既存のアプリケーションにインパクトを与えずに基幹業務を構築可能とするような二重系システムを実現するためである。すなわち、使用しているOSおよびDBなどはいわゆるオープンプラットフォームとよばれるものを使用し、系間切替で、業務の連続性を保証するためには、OS、データは単一とし、一元的に管理、簡単な形態として、回復処理を安全に行えるようにするのが最良と考えた。

このようなデータを単一とする方式としたコールドスタンバイシステムを実現するためには、主として次の4点が課題であった。

① 両系サーバの監視および系切替制御方法
② 拡張ストレージキャビネット中のディスク系間制御
③ サーバ切替時のネットワークアドレスの切替
④ サーバ異常検知処理

4　実現方式

3で述べた課題に対する解決方策を以下の節でそれぞれ述べる。

4.1 両系間の監視および障害時の系切替制御

両系のサーバには常時動作可能なようにバッテリバックアップされた専用ハードウェアであるサーバ管理装置を内蔵し、専用ケーブルで接続することによって系間通信を行い、両系の連携した制御を可能としている。業務サーバにおいてハードウェアまたはソフトウェアによる障害が検出された場合、業務サーバのサーバ管理装置は(1)外部に障害を通知する、(2)業務サーバのShutdown処理を起動する、(3)Shutdown処理終了後に拡張ストレージキャビネットを切離す、(4)専用ケーブルを介して待機サーバの状態チェックを行った後、待機サーバのシステム管理装置に切替要求を出す、(5)自系を故障サーバに変更する、という手順で切替制御を行う。

待機サーバのサーバ管理装置は切替要求を受けて、(1)自系を業務サーバに変更する、(2)拡張ストレージキャビネットへ接続要求を出す、(3)待機サーバの電源投入を行いOSを起動する（この際OSおよびアプリケーションによる回復処理が行われる）、という手順で切替制御を完了する。もし、待機サーバの状態チェックで切替不能と判断した場合は切替を行わずシステム全体の障害として外部に通知を行う。

なお、両系のサーバ管理装置は相互に専用ケーブルを介して定期的に通信をすることにより互いの監視を行っている。一定時間相手系から通信がなければ相手系サーバのサーバ管理装置の故障と判断するが、業務に直接の影響はないためこの事象による系切替は行わず外部への故障通知のみを行う。

---

[1] PentiumPro Server Standby System

[2] Toshiyuki Kimura, Hitoshi Yamamoto, Masahiro Mizuno, Toshio Matsumoto, Mitsunari Uozumi(Mitsubishi Electric Corp.), Shin'ichi Takahashi (Mitsubishi Electric Computer Systems(TOKYO))

4.2 ディスクアクセス系間制御

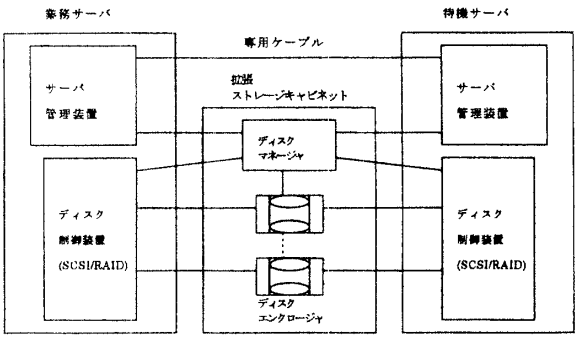

図2　二重系におけるハードウェア接続概要

各サーバの PCI バスを経由して接続されているディスク制御装置は、拡張ストレージキャビネットに実装されている複数台のディスクエンクロージャ（デュアルポートのディスク装置）に対して２つの系からアクセスできるように接続されている。系間制御を行うためのディスクマネージャは、拡張ストレージキャビネットに実装されており、各々のサーバに内蔵されたサーバ管理装置からの接続要求信号を受け取る。

ディスクマネージャは、業務サーバからの指示によって該当する系へのアクセスのみが実行されるように、ディスクエンクロージャにアクセス権要求を発行する。ディスクエンクロージャはそれによってアクセス権要求のあった系にのみアクセス権を与える。これによって業務サーバのみがディスクにアクセスでき、業務 OS やデータは業務サーバからのみアクセスするように保証し、両系からの同時アクセスによるデータの破壊を防止している。この機構によって、故障したサーバからのディスクアクセスも防止することができるため、故障サーバ修理の際の安全性も確保できる。

ディスクマネージャはこの他、ソフトウェアとの通信機能、各イベントのログ機能を持ち、さらにディスク制御装置との通信機能を備え、系間切替やディスクアクセス上のエラー等のログをタイムスタンプとともに残すことができ、これらの情報を参照することでシステム管理に役立てることができる。

4.3 ネットワークアドレスの引継ぎ

待機サーバが業務サーバに切り替わると、業務 OS を使用して、TCP/IP をネットワークプロトコルとして使用する場合でも、IP アドレスは、もともとの業務 OS で使用していたアドレスを使用できることは予想できる。ところが、もともとのネットワークアドレスは、２枚の別々のネットワークボードのために、異なるアドレスとなっている。そのため、ARP テーブルがもつ、IP アドレスとネットワークアドレスとの対応関係に不整合が発生し、IP 上での通信の保証が不可能となる。

このような問題に対処するため、通常業務に使用しているサーバの LAN 制御装置のネットワークアドレスを業務 OS 上に記録しておき、通常待機しているサーバが業務モードとなったときに、業務 OS 上にあるネットワークアドレスを、LAN 制御装置のネットワークアドレスを上書きするようにした。すなわち、もとあった LAN 制御装置固有のネットワークアドレスに替わり、業務用のネットワークアドレスを使用することとなる。これにより、ネットワークアドレスの上でも一致し、システム、ディスクデータもそれまで使用していたものをそのまま使用するので、サーバの外側から見ても、一貫性・整合性が保たれることとなる。

4.4 異常の検出

通常、二重系でも、ハードウェアによる系間通信のみによって相手系のサーバの異常を検出する方式では、OS が異常となって、サーバとしての活動を停止しても、系間通信自体は生きているために、サーバ自体を異常とは検出しない。また、RAID 制御装置や、LAN 制御装置などの拡張ボードの異常に対しても検出することも不可能である。

本スタンバイシステムでは、これらの異常をも検出して、異常検出範囲を広げてサーバ可用性を向上させている。これらのサーバの異常検出は、大別すると、サーバ管理装置(ハードウェア)による検出と、ソフトウェアによる検出の２つに大別される。

1. ハードウェアによる検出
   - ◆ OS 起動失敗タイマ監視
     サーバ電源ＯＮから、ブート処理の各段階にチェックポイントを設定してそれぞれをタイマ監視し、規定の時間までにそれぞれのイベントが終了しない場合には、異常と判定する。
   - ◆ ウォッチドッグタイマ
     サーバ管理装置にはウォッチドッグタイマを用意し、OS 起動完了後、OS 上のサービスプログラムが、サーバ管理装置に対して、定期的にリセットをかけるようにしている。OS に異常が発生すると、このリセットの動作が停止するので、ウォッチドッグタイマのタイムアウトが発生した時点で、サーバ管理装置はサーバ異常と判定する。
2. ソフトウェアによる検出
   - ◆ RAID/LAN 制御装置の監視
     OS 上のサービスプログラムから、定期的に RAID 制御装置および LAN 制御装置に対して実際に入出力を発生させることにより、異常か否かを判定している。この入出力による性能の低下はほとんど無視しえるほど軽微なものである。これによって、サーバ管理装置では検知できない、拡張ボードの異常をも検知し、検知したときには、サーバ管理装置に対してサーバ切替を要求することによって、OS のシャットダウンおよびサーバの切替を実施する。
   - ◆ アプリケーションサービスプログラムの監視
     これも OS 上のサービスプログラムから、定期的にアプリケーションサービスプログラムの状態を調べることにより、停止したサービスの再起動などを実施する。このソフトウェア異常は、サーバ本体の異常が原因とは考えにくいので、たとえこの異常が検知されたとしても、サーバの切替は実施されず、リブートやサービスの再起動を実施することとなる。

5　まとめ

以上のように、ハードウェアの拡張機能の追加、ソフトウェアからも追加の監視・制御を行うことで、比較的安価で、対障害性の大幅向上を実現することができた。今後は、待機サーバを、別の業務にも使用できるような、サーバの有効利用にもつながるシステムを考案・構築する予定である。