7H－7

UNIXワークステーションによるネットワークを用いた高可用性システム

藤井 誠司、上村 ホゼ、菅 隆志

三菱電機(株) 情報システム研究所

1 はじめに

近年、コンピュータシステムのダウンサイジング化の流れの中で、安価なUNIXワークステーションをネットワークに接続して、アプリケーションを利用するクライアント・サーバ型のシステムが増加している。

このようなシステムを予約システム、在庫管理システムのような オンライン・アプリケーションを使用するビジネス分野に適用する場合、高い信頼性または可用性が要求される。

そのための対策として、複数の計算機を冗長にネットワークで接続してシステムを構成し、高い可用性を実現する方法が存在する。

このシステム構成では、長時間の連続稼働を要求される高い信頼性は保つことは難しいが、標準的なハードウェアおよびソフトウェアによって構成することが可能であるので、安価に、オープンなシステムで可用性の高いシステムを実現できる。

我々は、特殊なハードウェアを使用しない高可用性システムを考案中であり、データベースをアプリケーションとして、複数の安価なワークステーションをネットワークに接続した高可用性システムのプロトタイプシステムを試作した。本稿では、この高可用性システムプロトタイプについて報告する。このプロトタイプの特徴は、ネットワークによるデータ複製機構である Network Disk [1] を使用して構成している点にある。

2 高可用性システムプロトタイプの課題

高可用性システムでは、障害が発生した時にいかに迅速に修復できるかが課題である。これを解決するためには、発生する障害とその障害に対する対策技術を検討することが必要である。

Network Based High Availabilty UNIX System
Seiji Fujii, Joe Uemura, Takashi Kan
Computer and Information Systems Laboratory, Mitsubishi Electric Corp.

クライアント・サーバ型システムにおいて可用性を損なう障害として、表1に示す障害が考えられる。

障害区分	障害
ネットワーク障害	経路故障
H／W障害	ノード故障、ディスク故障
S／W障害	OS障害、アプリケーション障害

表 1: クライアント・サーバ型システムの障害

また、その対策技術は処理中に障害を監視する検出技術と、サービスを維持し障害から回復するための構成・回復技術に分けることができる。我々のプロトタイプシステムでは、ノード故障とアプリケーション障害について、それぞれの対策技術を、以下のような方法で実現した。

1. 検出技術
   - ノードおよびアプリケーションの動作の異常を検知するためにノード間のハートビートメッセージの交換によるヘルスチェック
2. 構成・回復技術
   - ノード間でデータを共有するためのS／Wミラーおよび Network Disk [1]
   - ノード障害発生時に、アプリケーションを速やかに他ノード上で起動するための冗長ホストペア構成および自動再起動 (フェイル・オーバー)
   - 障害発生時のアプリケーションの強制終了
   - 履歴データに基づくアプリケーション内部状態回復のためのアプリケーション再起動

3 高可用性システムのプロトタイプ概要

我々が構築したプロトタイプシステムを図1に示す。このシステムは、全く同一なハードウェア構成である標準的な2台のUNIX サーバ・マシンと、一台のクライアント・マシン、このシステムの稼働状態を表示する状態表示用のマシンをネットワークによって接続した構成である。2台のUNIX サー

バ・マシンには、アプリケーションのデータを記憶するための外部記憶装置(ハードディスク)がそれぞれに接続されている。そして、システムの稼働中には、一方が稼働系(Primary Server Machine)、もう一方が待機系(Backup Server Machine)として動作する。

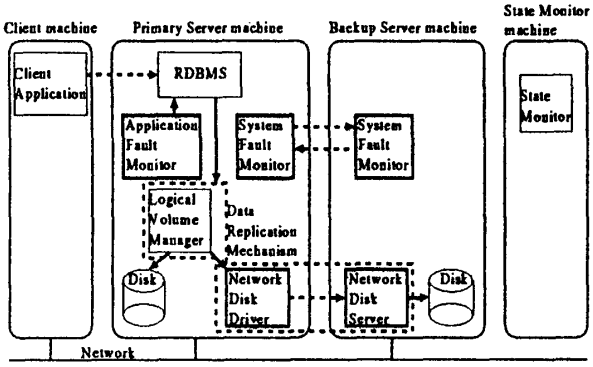

図 1: 高可用性プロトタイプシステムの構成

そして、ソフトウェアは標準的なＯＳ(UNIX)を使用し、サーバ・アプリケーションとしてはビジネス用途で利用されることの多いデータベースを使用している。また、サーバ・アプリケーションの高可用化を実現するために、Network Disk、System Fault Monitor、Application Fault Monitor を含むソフトウェア・パッケージを開発した。

稼働系から待機系へサービスを移行するためには、サーバ・マシン間でサービスを回復するために必要なデータを共有する必要がある。この機能を実現するために、[1]によって提案されたデータ複製機構を使用した。このデータ複製機構は、Logical Volume Manager および Network Disk によって構成される。これにより、稼働系サーバマシン上のアプリケーションがこのデータ複製機構に対してＩ／Ｏを行うことによって、稼働系サーバマシンのディスクへのアクセスがネットワーク経由で待機系のディスクへ複製され、データの共有が行なわれる。

これにより、以下のような特徴を持つ。

- ＯＳおよびサーバ・アプリケーションを修正せず、標準的なハードウェアを使用して、データの共有が実現できる。
- ネットワークの使用により、マルチポートディスクを使用するシステムに比べてケーブルの長さによる制約が小さいため、二つのサーバを離れた場所に設置できる。

System Fault Monitor は、以下のような機能を持つ。

- サーバに存在し、互いにハートビートメッセージを交換することによって、サーバ・マシンの稼働状態を監視する。
- 稼働系サーバで障害が発生した時には、稼働系サーバマシンのサービスを待機系サーバマシンへ移行するための回復処理を行なう。
- システムの稼働状態を State Monitor に通知する。

Application Fault Monitor は、以下のような機能を持つ。

- サーバ・マシンのアプリケーション稼働状態を監視する。
- アプリケーションの異常を検知した場合には、アプリケーションを停止し、アプリケーションを再起動する。
- System Fault Monitor のサーバ回復処理の一貫としてアプリケーションの起動を行なう。

このシステムの可用性を知るために、稼働系サーバの電源を落し、稼働系で障害が発生してから待機系で障害を検知し、サービスの移行処理を開始するまでの時間、スイッチオーバ時間を計測した。System Fault Monitor のハートビート間隔を１秒に設定した時、スイッチオーバ時間は５秒以内であった。これより、スイッチオーバ時間は従来方式[2]と同等な性能を提供しているといえる。

4 おわりに

本稿では、我々が構築した高可用性プロトタイプシステムについて報告した。このプロトタイプシステムは特別なハードウェアの追加やサーバ・アプリケーションの変更をせずに、高可用化のためのソフトウェア・パッケージを付加することによってサーバ・アプリケーションの高可用化を実現している。

現在、ATM などの高速ネットワークの使用、多重サーバへの適用、システムのクラスタ化構成等の検討を進めている。

参考文献

[1] 岡田英明、坂倉隆史、上村ホゼ、菅隆志: UNIX ワークステーションによる高可用性システムのデータ複製機構 情報処理学会第４８回(平成６年前期)全国大会(1994).

[2] Jim Gray and Andreas Reuter: TRANSACTION PROCESSING: CONCEPTS AND TECHINQUES (1993).