マルチメディア通信と分散処理 66−26
（１９９４．７．８）

分散型メッセージ通信システムにおける高信頼化機能の実現

藤長 昌彦　加藤 聰彦　鈴木 健二

国際電信電話 (株) 研究所

本稿では、分散処理技術により結合された複数のサーバから構成される分散型メッセージ通信システムの高信頼化機能の実現手法について述べる。本手法では、X.400 MHS プロトコルに基づく MHS メッセージの交換処理を行なう MHS 交換サーバに加えて、障害回復に必要な情報を安全に保持するログサーバと、システムの状態を監視するシステム監視サーバ を導入し、これらのサーバのレプリケーションを行なうことにより信頼性を高める。ソフトウェアのバグやハードウェアの故障等の原因で一部のサーバに障害が発生した場合にその回復を試み、回復できない場合には残りのサーバによりシステム全体としての動作を継続する。

An Implementation of Highly Reliable Distributed MHS System

Masahiko FUJINAGA　Toshihiko KATO　Kenji SUZUKI

KDD R&D Laboratories
Ohara 2-1-15, Kamifukuoka, Saitama 356, JAPAN

A distributed MHS system is a message handling system that is composed of multiple servers on a cluster of computers using distributed processing technologies. In order to make such a system reliable, it is required to recover servers from their failure. Even when the recovery is not possible, the system should continue to operate with possibly degraded performance. This paper describes an implementation scheme of such highly reliable distributed MHS system. The scheme introduces special purpose servers, such as log severs and system monitoring servers, for recovery from the failure, uses distributed processing technologies such as server replication technique, and provides the rollback and dynamic re-configuration function.

1 はじめに

筆者らはこれまでに、クライアントサーバ・モデルに基づく RPC (Remote Procedure Call) 等の分散処理技術を用いて、通信システムを構築する技法 [1, 2] や、複数のサーバにより MHS メッセージの中継交換を行うメッセージ通信システム (以下、分散型メッセージ通信システムと呼ぶ) の実装手法を検討してきた [3]。本システムでは、障害が発生した場合にも、他のシステムから受信したメッセージを紛失しない、同一のメッセージを誤って複製しない等の高い信頼性が要求される。

本稿では、分散型メッセージ通信システムにおける高信頼化機能の実現方法について述べる。本方法では、障害に対処するための特別なサーバを導入し、サーバのレプリケーション等の分散処理技術を用いて、ソフトウェアのバグやハードウェアの故障により一部のサーバに障害が発生した場合にはその回復を試み、回復不能な場合には残りのサーバによりシステム全体としての動作を継続する。以下、第 2 章では対象とする分散型メッセージ通信システムに対する高信頼化機能の設計方針について述べ、第 3 章でその実現方法を詳述する。第 4 章では筆者らの採用した実現法について考察し、第 5 章に結論を示す。

2 高信頼化機能の設計方針

2.1 分散型メッセージ通信システムの概要

分散型メッセージ通信システムは、高速な LAN (Local Area Network) で結合された計算機上において、RPC を用いて機能分散や負荷分散を行う複数のサーバから構成されるメッセージ通信システムである。その基本単位は、MHS 用通信ボード [4] を制御する MHS 交換サーバであり、MHS メッセージの受信、ディスパッチ、送信の各処理を行う。MHS 交換サーバは要求される処理性能に応じて複数導入され、それらの協調動作によって、広域データ網で結ばれた他のメッセージ通信システムとの間で X.400 MHS プロトコルに基づく通信を行う。

2.2 設計方針

分散型メッセージ通信システムの高信頼化機能を実現するために、以下の設計方針を採用した。

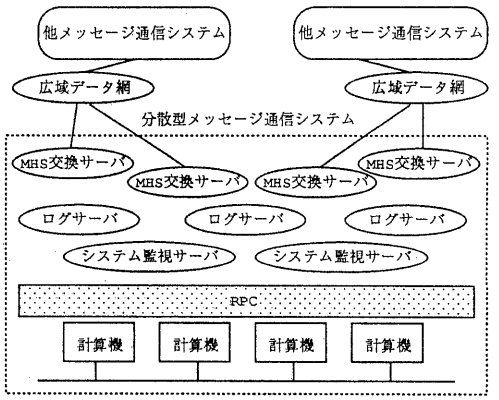


図 1: 分散型メッセージ通信システムの構成

(1) 本稿では、以下の障害を想定する。

- ソフトウェアのバグ等によるサーバ障害
- MHS 用通信ボードのハードウェア故障等によるサーバ障害
- オペレーティングシステムのソフトウェア障害や一時的な電源断等による計算機障害
- ハードウェア故障等による計算機障害

ただし、ディスクについては、ミラーディスクや RAID[5] などの高信頼化ディスクを想定し、ディスクに正しく格納された情報は紛失しないものとする。

(2) MHS 交換サーバに加えて、ログサーバとシステム監視サーバを導入する (図 1参照)。ログサーバは、高信頼化ディスクを用いて、受信した MHS メッセージや処理の進行状況などの MHS 交換サーバの障害回復に必要となる情報を格納する [6]。システム監視サーバは、分散型メッセージ処理システムを構成する計算機、MHS 交換サーバ、ログサーバが、正常に動作していることを定期的に監視する。

(3) これらのサーバを、複数の計算機上でレプリケートすることにより、サーバのアベイラビリティを向上させる。

(4) 上記の各サーバに対して、次のような障害回復機能を実現する。

- サーバに障害が発生した場合はそのサーバを再起動し、障害発生以前の状態に復帰させて処理を再開させる。

- ハードウェア故障等により、そのサーバによる処理の再開が不可能な場合には、他のサーバに処理を代替させる。

3 高信頼化機能の実現法

3.1 高信頼化のための処理概要

上記の設計方針に基づき、MHS 交換サーバ、ログサーバ、システム監視サーバの各サーバに以下のような処理を行わせる。

(1) MHS 交換サーバは、通常動作時において、受信した MHS メッセージや、メッセージ処理の進行状況など、障害からの回復に必要な情報 (以下、障害回復情報と呼ぶ) をログサーバに書き込む。

(2) 書き込み時にログサーバの障害を検出すると、MHS 交換サーバは、他のログサーバを選択し、以降そのログサーバを用いてメッセージ処理を継続する。

(3) MHS 交換サーバに障害が発生し再起動された場合は、まず、その MHS 交換サーバの最新の障害回復情報を保持するログサーバを決定する。次に、そのログサーバに保持された情報から MHS 交換サーバの内部状態を回復し、処理を継続する。

(4) システム監視サーバは、自身が監視すべき MHS 交換サーバとログサーバに対して定期的に動作確認のための RPC(以下、ヘルスチェック RPC と呼ぶ) を発行し、そのサーバが動作していることをチェックする。障害を検出すると、サーバ障害か計算機障害かを切り分ける。サーバ障害の場合は、そのサーバの再起動を試みる。サーバが再起動できない場合あるいは計算機に故障が発生している場合には、そのサーバあるいは計算機をメッセージ通信システムから切り離す。

(5) システムから切り離された MHS 交換サーバまたはログサーバが再起動された場合は、その旨をシステム監視サーバが各サーバに通知する。

(6) システム監視サーバの障害を検出するために、そのレプリケーションを行い相互に監視させる。

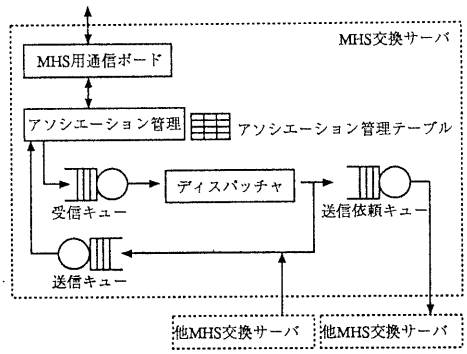


図 2: MHS 交換サーバの内部構成

3.2 MHS 交換サーバの高信頼化処理

3.2.1 通常動作時の処理

(1) ログサーバへの障害回復情報の格納

MHS 交換サーバは、図 2に示すように、MHS 用通信ボードを制御するアソシエーション管理部と、MHS メッセージの解析と更新を行うディスパッチャ部からなる。MHS メッセージの、これらの処理モジュール間における受渡しと他の MHS 交換サーバへの移動のために、受信キュー、送信キュー、送信依頼キューを設けている。また、通信中の各アソシエーションについて、その状態や、関連するセションコネクション識別子等を保持するアソシエーション管理テーブルを設けた。MHS 交換サーバは、MHS 用通信ボードから MHS メッセージを受信すると、メッセージ本体をログサーバに書き込み、受信キューに格納する。以降、メッセージ処理の進行に応じて各キューとアソシエーション管理テーブルを更新するつど、それらの内容をログサーバに書き込む。

(2) ログサーバの障害検出時の処理

MHS 交換サーバが障害回復情報の書き込み時にログサーバの障害を検出すると、システム監視サーバにその事実を通知するとともに、レプリケートされた複数のログサーバの内ひとつを選択し、ログサーバの切替えを行う。この時 MHS 交換サーバは、自分自身の障害回復時に使用するログサーバを決定可能とするために、新たに使用を開始するログサーバのサーバ名と現在の時刻の組 (以下、ログサーバ切替え情報と呼ぶ) を、その時点で動作しているすべ

てのログサーバに書き込む。

3.2.2 障害回復時の処理

(1) 最新のログサーバの決定

MHS 交換サーバに障害が発生して再起動された場合には、以下の手順により障害が発生した時点で使用していたログサーバを決定する。

(1) 導入されているすべてのログサーバに対して、自分自身に関する最新のログサーバ切替え情報を問い合わせる。

(2) ログサーバは、保持している最新のログサーバ切替え情報を通知し、更に、その情報が書き込まれた時刻以降にログサーバ自身に障害が発生していた場合には、障害の発生時刻及び回復した時刻を MHS 交換サーバに通知する。

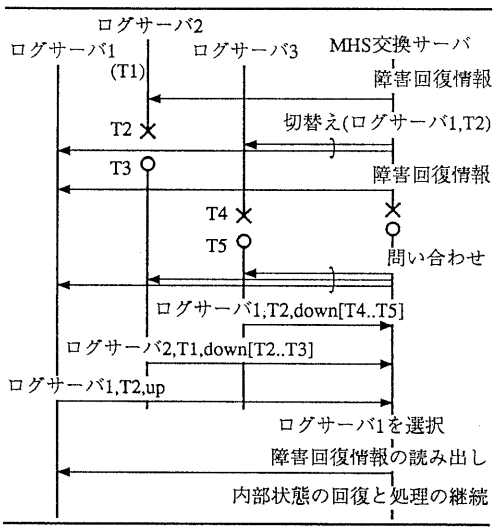
(3) すべてのログサーバから応答があった場合には、その中の最新のログサーバ切替え情報により障害発生時に使用していたログサーバを決定できる。

(4) 問合せに応じない (障害中の) ログサーバがある場合には、得られたログサーバ切替え情報の内最新のもの以降に、問合せに応じたすべてのログサーバが障害中であった時間帯があるか確認する。

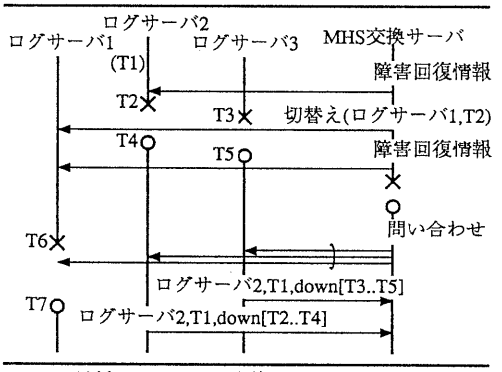
(5) 問合せに応じたログサーバがすべて障害中であった時間帯がなければ、最新のログサーバ切替え情報によりログサーバを決定できる。

(6) もしあれば、障害発生時に使用していたログサーバを決定することができないため、障害中のログサーバが回復するまで、問合せを繰り返す。

図 3に、この手順の例を示す。図 3(a) ではすべてのログサーバから応答がある場合であり、最新のログサーバ切替え情報により、障害発生時にログサーバ 1 を使用していたことを決定できる。一方、図3(b) では、問合せに応じたログサーバ 2 及びログサーバ 3 の情報は最新のログサーバとしてログサーバ 2 を示しているが、時刻 T1 以降に両者が障害中であった時間帯 [T3..T4] が存在する。この期間中にログサーバの切替えを行なった可能性があるため、障害中のログサーバ 1 の回復を待つ必要がある。



(a) 最新のログサーバを決定できる場合の例



(b) 最新のログサーバを決定できない場合の例

注) ×：障害の発生, ○：再起動, Tn：時刻

図 3: 最新のログサーバの決定手順の例

(2) 内部状態の回復

MHS 交換サーバは障害発生時に使用していたログサーバを決定すると、そのログサーバに保持された最新の各キュー及びアソシエーション管理テーブルの内容と、処理中であった MHS メッセージの本体を読み出して、処理を再開する。なお、MHS 用通信ボードの故障を検出した場合には、処理途中のすべての MHS メッセージについてディスパッチ処理までを行い、その送信処理を他 MHS 交換サーバに依頼する。

3.3 ログサーバの障害回復処理

ログサーバは、データを強制的にディスクに書き込む機能を提供する。アペンドオンリーにデータを書き込むことによりディスクヘッドのシーク時間を最小とし、高速な強制書き込み機能を実現している。

ログサーバの障害回復処理では、最新のログを検索してその完全性をチェックする。もし書き込みの途中で障害が発生していた (最新のログが不完全である) 場合にはそのログを廃棄する。障害から回復したログサーバは、システム監視サーバに再登録され、MHS 交換サーバからのアクセスを待つ。

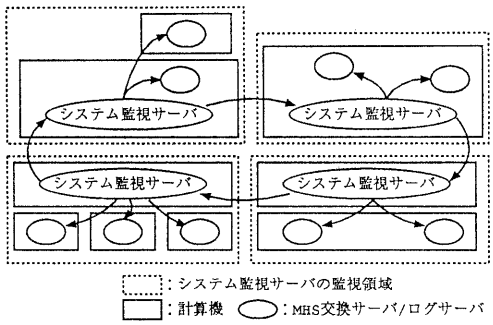


図 4: システム監視サーバによるシステムの監視

3.4 システム監視サーバによる高信頼化処理

(1) 障害の検出と障害回復手順の起動

システム監視サーバは、その起動時にシステム管理者が作成する初期設定ファイルを読み込み、各サーバがどの計算機上で動作すべきかの情報と、各システム監視サーバがどのサーバと計算機を監視するかの情報を記したリストを作成する。このリストに従って各サーバに対して定期的にヘルスチェック RPC を発行する (図 4参照)。

ヘルスチェック RPC が失敗した場合、ICMP (Internet Control Message Protocol) ECHO パケットをその計算機あてに送信し、計算機自体の障害が発生しているか確認する。ICMP ECHO パケットに対する応答があれば、サーバ障害の発生と判断してサーバの再起動を行う。

応答が無い場合、その計算機が再起動中であるか、ハードウェア故障等による停止状態であるかを判断するため一定時間後に ICMP ECHO パケットを再送する。再起動が行われない場合にはシステム管理者に通知するとともに、システムの再構築処理を起動する。

ハードウェア故障と判断された計算機のシステムからの切り離しは、その上で動作するサーバをヘルスチェック RPC の対象から外し、他の MHS 交換サーバに当該サーバが利用不能になったことを通知することにより行う。システム監視サーバはその計算機上で動作していた MHS 交換サーバについては、他の計算機上でその肩代りを行うサーバを起動する。肩代りサーバは、MHS 交換サーバが使用していたログサーバを決定し、その内部状態を回復して処理中であった MHS メッセージのディスパッチ処理までを行い、その送信処理を他の MHS 交換サーバに依頼する。

(2) サーバの再導入

切り離されたサーバの再導入は、システム管理者のコマンド入力により行う。システム監視サーバは、監視対象とするサーバのリストを更新して再導入されたサーバを再びヘルスチェック RPC の対象とし、他のサーバに対してその旨を通知する。

(3) システム監視サーバの障害回復処理

システム監視サーバの障害を検出すると、そのシステム監視サーバの再起動を行う。システム監視サーバは、各時点でシステム内に存在するサーバの動作状況に関する情報を保持しているが、これは、ヘルスチェック RPC の引数として、他のシステム監視サーバから与えられる。このため、システム監視サーバの障害回復処理においては、特別な処理を行う必要はない。

4 考察

(1) 一般的に、障害回復のための基本的情報を特定のサーバに保持させると、そのサーバの高信頼化のために、一層複雑な回復処理を導入する必要が生じる。例えば、MHS 交換サーバの使用するログサーバに関する最新の情報を、システム監視サーバに保持させると、その情報自身をログサーバに格納する必要があり、システム監視サーバが使用する最新のロ

グサーバに関する情報を別途管理する必要がある。このような問題点を避けるために、本システムでは、各MHS交換サーバの使用するログサーバに関する情報を、そのMHS交換サーバ自身が、利用可能なログサーバに分散して格納し、障害回復時においても、その時点で利用可能なログサーバに問い合わせて、最新のログサーバを決定するという分散型の手法を用いた。

(2) システム監視サーバは、定期的にMHS交換サーバとログサーバおよび他のシステム監視サーバにヘルスチェックRPCを発行している。このため本システムでは、その時点で利用可能なMHS交換サーバとログサーバの動作状況に関する情報をヘルスチェックRPCの引数に含めている。従って、システム監視サーバによるシステム監視に最低限必要なオーバヘッドにより、サーバの動作状況に関する情報を通知することができる。しかし、ヘルスチェックRPCは一定周期ごとにしか発行しないため、必ずしも各時点での最新のサーバの動作状況を伝えることはできない。このため、MHS交換サーバはヘルスチェックRPCに含まれる情報をヒントとして利用し、MHSメッセージの送信依頼やログサーバの切替えを行う。指示されたサーバにアクセスして障害中であると検出された場合は、その情報の示す別のサーバにアクセスする。

(3) 分散システムにおける障害回復では、相互に依存するサーバ間におけるログ書き込みのタイミングや回復処理における同期の不整合が原因で、障害からの回復時に初期状態にまで復帰するドミノ効果と、回復処理が終了しないライブロックが問題となる[7]。分散型メッセージ通信システムにおいては、MHS交換サーバ間の相互依存関係はMHSメッセージの送信処理の依頼に限定しており、送信の依頼を行った場合には、その直後に依頼元と依頼先の両者で整合性のとれた障害回復情報を書き込んでいる。また、障害が発生したMHS交換サーバは、他のMHS交換サーバをより以前の状態に復帰させることなく、独立に回復可能である。これらによりドミノ効果とライブロックは発生しない。

(4) 本システムでは、MHS交換サーバの内部状態を、MHS交換サーバ自身とログサーバ内とに二重化していると捉えることができる。MHS交換サーバとログサーバが異なる計算機上で動作すれば、同時に双方のサーバに障害が発生する確率は小さい。ログサーバに障害が発生した場合、MHS交換サーバはすぐにログサーバを切替えて、新たなログサーバに障害回復情報を記録し、通常処理に戻る。またMHS交換サーバに障害が発生した場合は、ログサーバに正常に格納された情報から、MHS交換サーバを再起動可能である。

5 むすび

本稿では、分散型メッセージ通信システムの高信頼化機能の実現手法について述べた。本手法は、X.400 MHSプロトコルに基づくMHSメッセージの交換処理を行なうMHS交換サーバに加えて、障害回復に必要な情報を安全に保持するログサーバと、システムの状態を監視するシステム監視サーバを導入し、これらのサーバのレプリケーションを行なうことにより信頼性を高めることを特徴としており、障害からの回復機能と、回復不能な場合の動的な再構築機能を容易に実現している。最後に、日頃御指導戴くKDD研究所浦野所長、真家次長に感謝する。

参考文献

[1] 加藤, 藤長, 鈴木：分散処理技術を用いた通信システムの構築に関する一考察, 情報処理学会第43回全国大会, No. 7T-1 (1991).

[2] Fujinaga, M., Kato, T. and Suzuki, K.: Implementing IN Functional Entities on top of Distributed Operating System, in *Proceedings of the XIV International Switching Symposium*, Vol. 1, pp. 268 – 272 (1992).

[3] 藤長, 加藤, 鈴木：分散処理技術に基づくメッセージ中継処理システムの 実装手法, 信学技報, Vol. IN92, No. 138, pp. 121–126 (1993).

[4] 加藤, 井戸上, 鈴木：MHS P1/P2/RTボードの開発, 1993年電子情報通信学会春季大会 (1993).

[5] Patterson, D., Gibson, G. and Katz, R.: A case for redundant arrays of inexpensive disks (RAID), in *ACM SIGMOD Conference*, pp. 109–116 (1988).

[6] 藤長, 加藤：高信頼アプリケーションのための汎用ログサーバ, 電子情報通信学会秋季全国大会, No. D–115 (1990).

[7] Koo, R. and Toueg, S.: Checkpointing and Rollback-Recovery for Distributed Systems, *IEEE Transactions on Software Engineering*, Vol. SE–13, No. 1 (1987).