グループ通信による高信頼分散オブジェクトサービスの設計

増田　大樹　　村山　和宏　　落合　真一

三菱電機(株) 情報技術総合研究所

1. はじめに

レーダ等の無停止連続運転を必要とする分散リアルタイムシステムでは、障害に備えて装置を多重化するだけでなく、データを失わずに処理を継続する仕組みが必要である。

我々はこれまでに、計算機やネットワークの多重化を行うグループ通信ミドルウェアを開発してきた[1]。今回、グループ通信に多重化した計算機間の整合を行う機能等を付け加えることで、上記のようなシステムを実現するミドルウェアの設計を行った。本稿では設計の概要について述べる。

2. 基本構想

2.1. 従来のシステム

本稿で想定するシステムは、図 1のように複数の計算機が連動し、受信したデータを処理していくシステムである。各計算機は入力されたデータを解析し、次の計算機(後段)に処理結果を送信する。データ解析結果の一部は、次回のデータ処理時に必須な情報になるため、計算機内のメモリ上に保持して繰り返し利用している。

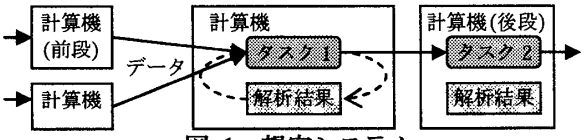

図 1　想定システム

2.2. 多重化システムの構想

前記システムに対し図 2のように計算機を多重化したシステムを構築する。データをグループ通信で後段の計算機に同報配信し、1台の計算機(常用系)のみ処理を行う。常用系に障害が発生したときは、他の計算機(待機系)の1台が新たな常用系となり、処理を引き継ぐ。

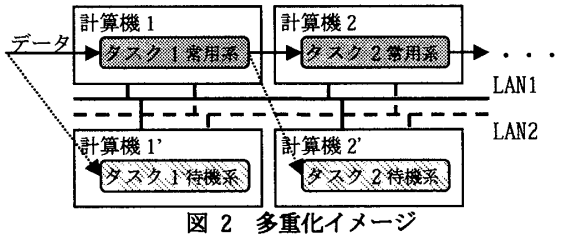

図 2　多重化イメージ

3. 課題

図 2に示すシステムを実現するためには、常用系の持つ解析結果を待機系に整合すると共に、次の課題を解決しなければならない。

(1) 待機系におけるデータの保持と破棄
(2) システム停止及び多重処理の防止
(3) 計算機の再参入

Design of fault-tolerant middleware using reliable multicast.

Hiroki Masuda, Kazuhiro Murayama, Shinichi Ochiai
Information Technology R&D center,
Mitsubishi Electric Corporation

以下に詳細を述べる。

3.1. 待機系におけるデータの保持と破棄

障害が発生した時に、待機系が処理済となったデータを再度処理しないように、待機系は常用系が処理したデータを破棄する必要がある。

このとき、図 1中央の計算機のように、複数の計算機からデータが入力される箇所では、常用系と待機系でデータの到着順序が異なる場合がある。また、常用系がデータを処理した時点で、データが待機系に未到着の場合もある。

そのため、待機系はどのような順序でデータを受信しても、常用系が未処理の間はデータを保持し、処理が完了した順にデータを破棄する仕組みを検討する。(図 3)。

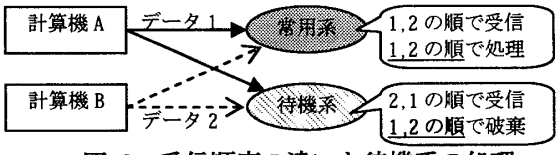

図 3　受信順序の違いと待機系の処理

3.2. システム停止及び多重処理の防止

3.2.1. 不正データによるシステム停止

本システムでは障害として、計算機の故障だけでなく、不正なデータを処理した事によるアプリケーションタスクの停止も想定している。不正なデータが原因の場合、待機系が同じ不正データを処理すると、システム全体の処理が停止する事になる。

常用系と待機系が共に停止した場合、再起動等により復旧に多くの時間を要する。そのため、本稿では、不正データのみを除去して待機系が処理を引き継ぎ、システム全体の停止を防止する仕組みを検討する。

3.2.2. 前段の障害による多重処理

後段に処理結果を送信した直後に障害が発生した場合、待機系が処理を引き継いで動作し、後段が同一データを多重に受け取る可能性がある(図 4)。このような場合に後段で多重処理を防止する仕組みを検討する。

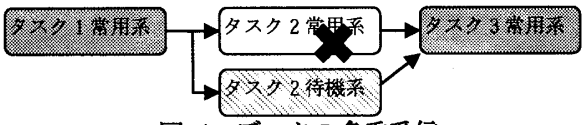

図 4　データの多重受信

3.3. 計算機の再参入

計算機が故障した場合、待機系の計算機台数が減少するため、本システムでは障害から復旧した計算機を新たな待機系としてシステムに再参入させることで、システムの多重度を保つ。

再参入を行うためには、解析結果を稼働中の常用系と整合する必要がある。このとき、障害発生時に常用系の計算機が変更されているため、再参入の時点で常用系となっている計算機から解析結果を取得する方法を検討する。

4. 設計

4.1. 設計概要

図 5は本方式のシステム概要である。

アプリケーションがメモリ上に保持する解析結果は解析結果格納領域に保存する。常用系は、データの処理を完了した時点で、解析結果格納領域の内容を待機系に送信し、整合を行う。また、データ受信順序の不整合に対応するために、データ受信キューを持つ。さらに、障害検出タスクにより、他の計算機とアプリケーションタスクの動作を監視する。

以下に設計の詳細を述べる。

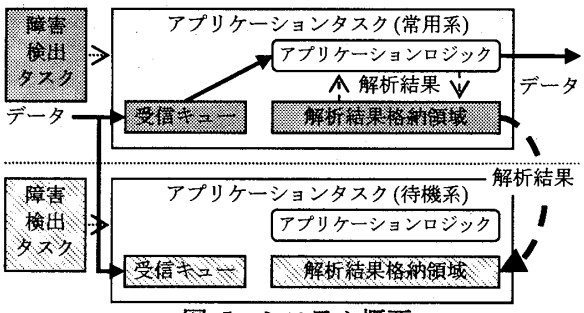

図 5 システム概要

4.2. 解析結果の整合と待機系のデータ破棄

常用系が処理したデータを待機系が特定するため、データを後段に送信する際にデータに ID を付ける。常用系はデータの処理が完了した時に、処理したデータの ID と格納領域の内容をグループ通信で全待機系に通知する(図 6 完了通知)。待機系は受信キューから該当データを破棄し、格納領域の更新を行う。

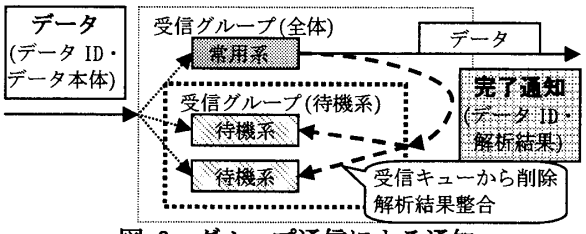

図 6 グループ通信による通知

このとき、3.1で述べたように、常用系が処理した時点ではデータが待機系に未到着の場合がある。そこで、図 6の完了通知で受信したデータ ID を記録し、データが遅れて到着しても受信キューに挿入しないようにする。

これらの処理により、待機系はどのような順序でデータを受信しても、常用系が処理したデータのみ受信キューから破棄するため、障害発生時にデータの消失や多重処理を防ぐことができる。

4.3. 障害発生時の動作

4.3.1. 常用・待機切り替え処理

各計算機には待機系としての優先順位を予め設定しておく。常用系計算機の故障を検出すると、稼働中の待機系の中で最も優先順位が高い計算機の障害検出タスクがアプリケーションタスクに通知を行い、常用系・待機系の切り替えを指示する。

4.3.2. 不正データの除去

3.2.1で述べた不正データを除去するために、障害検出タスクは同一計算機上のアプリケーションタスクを監視する。障害検出タスクはアプリケーションタスクに障害が発生した時に OS より通知を受けるように設定し、アプリケーションタスクが強制終了した場合を、不正データによる障害発生とみなす。

アプリケーションタスクは、データ処理時にデータ ID を障害検出タスクに登録する。アプリケーションが強制終了した場合は、常用系の障害検出タスクから待機系にデータ ID を通知し、各待機系は受信キューから該当データの除去を行う。これにより、不正データによるシステム停止を防止できる。

4.3.3. 多重受信への対応

3.2.2で述べた後段計算機の多重処理を防止するため、後段に送るデータ ID を解析結果の一部として格納領域に保存し、整合させる。データ ID を常用系と待機系で整合することで、4.2で述べた受信キューの動作により、後から来たデータはキューに入らないようになる。

このとき、待機系が処理を引き継いだ時に、常用系と異なるデータを処理すると、異なるデータに同一 ID が付き、後段は本来処理すべきデータを破棄する事になる。

そのため、常用系はデータの処理を開始する前に、処理するデータの ID を待機系に通知し、待機系は該当データを受信キューの先頭に移動する。これらの処理により、障害発生時の多重処理と誤ったデータ破棄を防止することができる(図 7)。

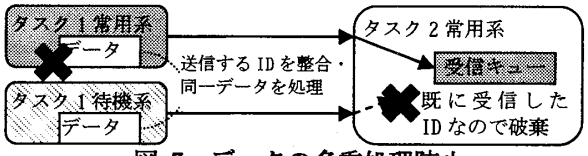

図 7 データの多重処理防止

4.4. 再参入処理

計算機が再参入する際は受信グループ全体に格納領域のデータ転送要求を送信する。この要求は常用系のみが処理し、格納領域の内容を要求元に返信する。この処理により、再参入した計算機は、稼働中の常用系と解析結果を整合し、再参入後に受信したデータから待機系として動作することができる。

5. おわりに

本稿では無停止連続運転を行うシステムに向けて、グループ通信で複数の計算機にデータを配信し、障害発生時に処理を継続する方式の設計を行った。現在、我々は本方式を実現するグループ通信ミドルウェアの開発を行っている。今後、このグループ通信を、我々が開発しているCORBAミドルウェア[2]の通信層として組み込み、分散オブジェクトの多重化を行うミドルウェアとする予定である。

参考文献

[1] 増田他「グループ通信ミドルウェアの冗長化設計」, 情報処理学会第 130 回 DPS 研究会、pp7-12、2007/03

[2] 大谷他「組み込みシステムにおける CORBA の機能要件」, 情報処理学会第 66 回全国大会、2004/03