オペレーティング・システム　56－1
(1992. 8. 19)

分散システムにおける送信データ予約方式による転送方法

山根　義孝　　高橋　義造
四国化工機(株)　　徳島大学

階層構造をした分散システムにおける放送機能を用いたデータ転送方法について提案する．他のプロセスのデータが必要になったプロセスは，予めデータの送信予約を行ない，送信予約されたデータを生成したプロセスは送信予約したプロセスを含む最小グループに対しデータを放送する．データを受信したPEは，接続されているPEで転送範囲内にあるPEへデータを転送する．また，送信予約したデータだけを取り込む．この方法により，分散システムの状態に応じた最短時間経路でデータが転送される．また，各PEは，システム構造によらず接続されているPEが分かればデータ転送を行なうことが可能となる．

DATA REGISTERING METHOD FOR COMMUNICATION ON A DISTRIBUTED SYSTEM

Yoshitaka Yamane　　Yoshizo Takahashi
Shikoku Kakoki Co.Ltd. *　　Univercity of Tokushima **
* 10-1 Nishinokawa Tarohachisu Kitajima-cho Itano-gun Tokushima 771-02, Japan
** Minami-Josanjima-cho Tokushima 770, Japan

In this paper, a data transmission method using broadcast on a distributed system is discussed. When a process requires data produced by another process, the former sends a request to the latter specifying the desired data. When the requested data is generated, it is sent to a minimum group of processes which includes the requesting process. The processing element (PE) which receives the data distributes it through the PEs in its minimum group. Each PE in the minimum group delivers the data to its own processes which have requested data. By the proposed communication method, 1) the data transmission time is reduced, and 2) each PE does not require an overall knowledge of the distributed system architecture, as only the connected PE information suffices.

1.はじめに

階層構造の分散システムが大規模化，高度化するにつれ，システムの信頼性・対故障性，柔軟性・拡張性が重要になってくる．

1.信頼性，耐故障性

システムの一部が故障した場合，システム全体を停止するのでなく，故障の影響を故障したシステムの近傍にとどめ，システムとしては稼働する必要がある．また，故障したシステムの機能を他のシステムが替わりに実行し，故障の影響を最小限に抑える．

2.柔軟性，拡張性

システムが大規模になると，拡張/縮小が頻繁に行われる．この時，システムを止める事無くシステムの変更が行われる必要がある．

信頼性・対故障性，柔軟性・拡張性を満足するシステムとして，自律分散システム[1]が考えられている．自律分散システムとは，システムを構成する各要素（サブシステム）が個々に自律して行動し，互いに協調しながらシステム全体として秩序を生成するシステムの事をいう．

対故障性，拡張性を有する自律分散システムにおけるデータ転送方法について考える．

2.分散システムにおける通信方法

分散システムの通信方法は，次の項目で区分できる．

1)送信データの決定者

送信するデータをだれが決定するかを示す．

a)プログラム作成時に送信データを決定する．

b)データを生成するサブシステムが決定する．

c)データを使用するサブシステムが決定する．

2)データ転送方法

送信サブシステムから受信サブシステムへどのようにデータを転送するのかを示す．

a)送信システム → 受信システム

（直接転送：１対１）

b)送信システム → 受信システム

（直接転送：１対N）

c)送信システム → サーバ → 受信システム

（サーバ経由）

3)データ送信時期

データの送信時期には次の２種類がある．

a)送受信要求がそろった時．

b)データが生成された時．

自律分散システムの通信方式として，日立製作所(株)森らの**内容コード通信方式**がある．この方式は，他のサブシステムが必要とするデータにその内容コードを付加し，ネットワークに接続されている全てのサブシステムに放送する．内容コード付きデータを放送されたサブシステムでは予め登録されている内容コードのデータだけを選択して受信する方式である．この内容コード通信方式を上記区分に当てはめると次のようになる．

送信データの決定者:データを生成するシステム

データ転送方法　：直接転送：１対N
（すべてのPE)

データ転送時期　：データが生成されたとき

3.送信データ予約方式

森らの内容コード通信方式には次のような問題点がある．

1.大規模システムでの通信コストの増加

内容コード通信方式は，送信データを全てのサブシステムに放送する方式であるため，分散システムが大規模になりサブシステム数が増加すると，それに伴い通信コストも増加する．

2.サブシステムの増加によるシステムの再構築

データを生成するサブシステムが送信データを決定するため，サブシステムの追加により送信データが増加するとデータを送信するサブシステムを再構築する必要がある．

これらの問題点を解決するために，**送信データ予約方式**による転送方法を提案する．この送

信データ予約方式による転送は，次のように行なう．

データを必要とするサブシステムはデータを生成するサブシステムを含むグループに対しデータの送信要求を放送する．この時，データの中継経路を確定する．送信要求をうけたサブシステムが送信要求のデータを生成した時，送信要求を行ったサブシステムを含む最小グループに対し確定された中継経路を介してデータを放送する．データが転送されたサブシステムは，予め送信予約したデータだけを選択して受信する．

この送信データ予約方式によるデータ転送は，2の区分に当てはめると次のようになる．

送信データの決定者:データを利用するシステム
データ転送方法　：直接転送：1対N
　　（サブグループ）
データ転送時期　：データが生成されたとき

この送信データ予約方式によるデータ転送は，次の利点がある．

1)送信要求されたデータを生成するサブシステムは，自分と送信要求を出したサブシステムを含む最小のグループに対してデータを送信する．そのため，転送範囲内でのサブシステムの増加による通信コストの増加は無い．

2)データの中継経路を確定しデータを放送するため，中継経路上に存在するPE/ネットワークが故障しても，他の経路が存在すれば、その経路を介して目的のPEにデータが自動的に転送される．そのため，データ転送時のシステム状態に応じた最短時間経路でデータが転送される．

3)グループに対しデータを放送するため，各PEは分散システム全体の構造がわからなくても，接続されているPEのPE番号/グループ番号を知る事によりデータ転送が可能である．

4.転送方法

送信データ予約方式によるデータ転送をFig.1に示すように，3ブロックにわけて説明する．

1.サブシステム間の通信方法
2.PE間の通信方法
3.通信路における通信方法

送信データ予約方式によるデータ転送では，通信路の形態を特に定めていない．例として，リング，バス，スター状のネットワーク等が考えられる．よって，ここでは，サブシステム間の通信方法，PE間の転送方法について以下に示す．

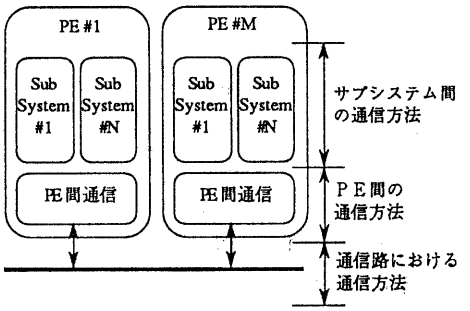

Fig.1 転送方法のレイヤー図

4.1.サブシステム間の通信方法

サブシステム間の通信はFig.2に示すような手順で行なう．

1.送信予約Requestの送信
　他のサブシステムの生成データを必要とするサブシステムは送信予約確認Ackを受信するか，受信タイムアウトになるまで，最下位グループから最上位グループへ順に転送範囲を広げながら送信予約Requestを放送する．
2.送信予約確認Ackの送信
　送信予約Requestのデータを生成するサブシステムは，送信予約Requestの内容を送信予約リストに登録し，送信予約確認Ackを送信予約Requestの転送範囲に送信する．
3.データの送信
　サブシステムがデータを生成した場合，送信

予約リストで送信予約データかチェックし，送信予約データの場合，送信予約要求グループにデータを放送する．複数のサブシステムから送信予約が行われた場合，送信予約要求グループは送信要求Requestの転送範囲を全て含む最小のグループとなる．

4.データの受信

送信予約確認Ackを受信したサブシステムはデータの受信状態になる．受信状態のサブシステムがデータを受信すると，送信予約したデータがチェックし，送信予約データの場合，サブシステム内に取り込み，その他のデータは無視する．

5.送信予約終了Requestの送信

他のサブシステムのデータを利用しているサブシステムで，データが不要になった場合，送信予約終了Requestを送信予約Requestの転送範囲に放送し，送信予約を解除する．

6.送信予約の解除

送信予約終了Requestを受信したサブシステムは，送信予約リストをチェックし送信予約があると送信予約リストから転送範囲を除く．転送範囲が無くなった時，送信予約リストから送信予約データの項目を削除する．

7.中継リストの更新

送信予約リストで転送範囲が変更になった場合，その旨を変更前の転送範囲に放送する．

4.2.PE間の通信

PE間の通信は，各PEの通信路の情報，中継リスト，パケット情報をもとに行われる．ここでは，パケットの転送方法で，パケットの中継処理，通信バッファについて考える．

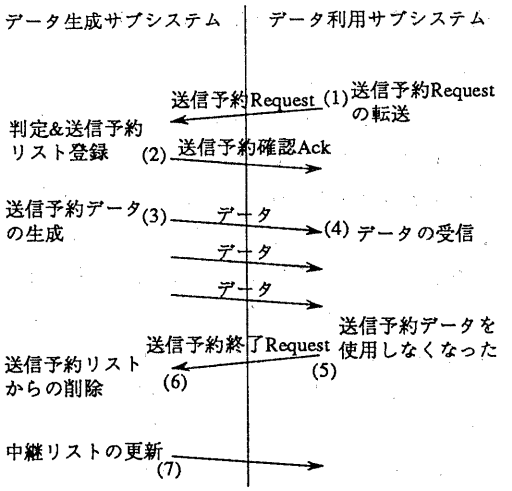

Fig.2 サブシステム間の通信方法

4.2.1.パケット構造

ここで扱うパケットはFig.3で示す構造をしている．

パケット種類	転送範囲	中継情報	データid	データ

Fig.3 パケット構造

1)パケット種類

パケットには，送信予約Request，送信予約確認Ack，データ，送信予約終了Request，中継リスト更新Requestがある．

2)転送範囲

転送範囲はパケットのどの範囲に放送するかを示す．

3)中継情報

パケットをどの通信バッファ(4.2.3で示す)を使用して転送するかを示す．

4)データid＋データ

4.2.2.パケットの中継処理

各PEにおけるパケットの中継処理は，通信路の結合リスト，データの中継リストをもとに行なう．

通信路の結合リストは，自分と結合しているPEのリストで，PE番号/グループ番号を示す．データの中継リストは，データid毎に転送する通信路，転送範囲及びデータ転送の有無を示す．各PEは，パケットの中継処理を次のように行う．

1)中継リストへの中継登録

受信パケットが送信予約Requestの場合，中継リストにデータid，転送範囲，転送通信路を仮登録する．受信パケットが送信予約確認Ackの場合，仮登録データを正式登録する．この時，同じデータidで転送範囲が異なる場合，転送範囲が最小になるように中継リストを更新する．

2)中継リストの中継削除

受信パケットが送信予約終了Requestの場合，中継リストの転送通信路にパケットを送信後，中継リストからデータidに関する項目を削除する．

3)受信パケットの転送

中継リストの転送通信路で，未転送の通信路に対し受信パケットを転送する．転送後，転送したことを中継リストに登録する．

中継リストの項目を以下に示す．

1)データid

データidは，送信予約をしたデータの種類を示す識別子である．

2)転送範囲

転送範囲とは送信予約を行った全てのサブシステムを含む最小のグループを示す．

3)転送通信路

データを転送する通信路を示す．

4)転送フラグ

通信路毎のデータ転送の有無を示すフラグである．

5)最新転送データの生成時刻

自律分散システムのネットワーク構造内にループ構造が存在する場合，前に生成したデータが後から送信される可能性がある．この事を防止するためのデータである．

4.2.3.仮想通信路

送信データ予約方式は，放送機能を用いてデータを転送する方式で，パケットが消滅するのは，これ以上転送(放送)できなくなった時である．この通信方式では，階層構造をした各レベル毎にそのレベル内に存在するグループ/PE専用の仮想通信路[3]を設け，パケット内の中継情報により示される仮想通信路を通って各PEに転送(放送)する方法をとる．

仮想通信路は，Fig.4に示すように各レベル毎にそのレベルに存在するグループ/PE数だけ存在する．

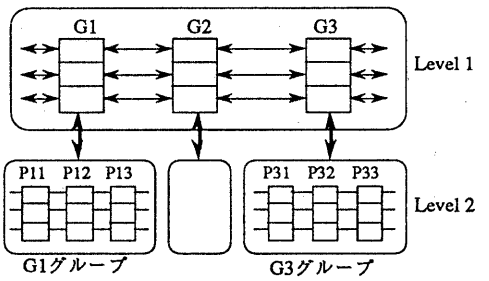

Fig.4 仮想通信路

各PEは通信バッファとして，同一レベル間用通信バッファと異なるレベル間用通信バッファを持つ．それぞれのバッファ数は次のようになる．

1)同一レベル用通信バッファのバッファ数はそのレベルに存在するグループ数/PE数である．

2)異なるレベル間用バッファのバッファ数は，下位レベルに存在するグループ数/PE数－1である．

Fig.5に仮想通信路に対応したバッファを示す。

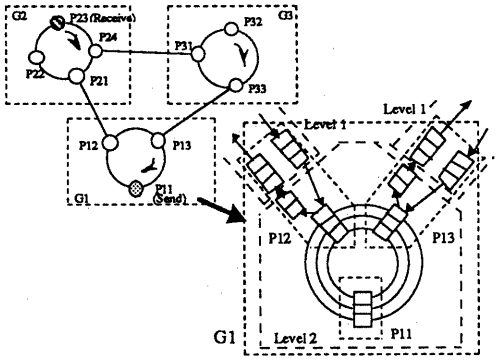

Fig.5 バッファ構成

4.2.4.中継情報

中継情報は，各レベルで使用する仮想通信路を示す。つまり，中継情報の各レベルの値は，パケットが各レベルのグループで最初に受信したPE，または，パケットを生成したPEのPE番号である．

中継情報は，最初はパケットの生成PE番号で，他のPEに転送される毎に以下の手順に従い更新される．

Step1 受信パケットを送信したPEのPE番号と自分のPE番号を各レベル毎に上位レベルから比較し，最初に異なるレベルを決定する．このレベルはPE間の通信路が確定された時点で決定され，通信路が変更がない限り不変である．

Step2 Step1で決定したレベル以下の中継情報の値を自分のPE番号の値で置き換える．

中継情報の更新例として，Fig.6を示す．Fig.6の中継経路の下線部分はPE番号が最初に異なるまでのレベルを示す．

中継経路：P2111 → P2130 → P2132 → P1231
中継情報： 2111 → 2110 → 2110 → 2231

Fig.6 中継情報の更新例

4.3.通信路の追加

通信路が追加された場合、通信路が追加されたPEは，次の処理を行なう．

1)追加された通信路に結合されているPEのPE番号/グループ番号を調べ，通信路の結合リストに登録する．

2)中継リストをもとに追加された通信路に対する中継データを確定する．

3)通常のパケットの中継処理を行なう．

4.4.データ転送例

正常状態でのデータ転送例，中継通信路が故障したときのデータ転送例をFig.7,8に示す．共に，P11 → P23へのデータ転送で，Fig.7の例では4回のPE間の転送で目的のPEに到達する．Fig.8の例は，P12<->P21間の通信路が故障した時の例で8回のPE間の転送で目的のPEに到達する．図中の括弧付き数値はデータ転送時間に相当する値である．

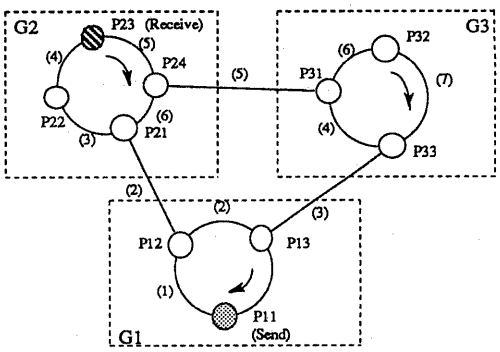

Fig.7 正常状態でのデータ転送

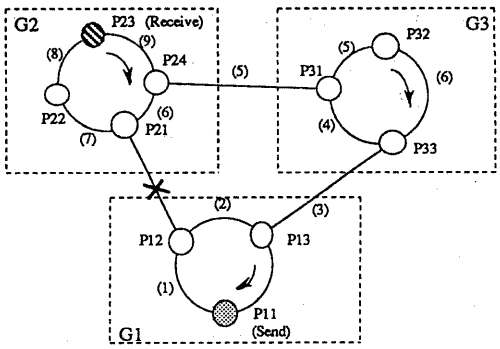

Fig.8 中継通信路が故障したときのデータ転送
(P12 -> P21への通信路)

5.PEのバッファ数

分散システムの個々のPEのバッファ数を求める.

- N　分散システムの全PE数
- L　レベル数
- i_j　プロセッサ・シーケンス番号iの最上位レベルからレベルjまでの
  PE番号= $(i_1,i_2,\cdots,i_j)$, $1 <= j <= L$
  1=最上位レベル,L=最下位レベル
- i_L　プロセッサ・シーケンス番号iの
  PE番号=$(i_1,i_2,\cdots,i_L)$
- P_{i_j}　レベルjのPE番号$(i_1,i_2,\cdots,i_j)$のグループまたはPE
- N_{i_j}　レベルjのPE番号$(i_1,i_2,\cdots,i_j)$のグループまたはPE数
- S_{i_j}　レベルjのグループ/PE(P_{i_L})の送信通信路数
- R_{i_j}　レベルjのグループ/PE(P_{i_L})の受信通信路数
- K_{i_j}　上位グループでの通信路の有無
  K_{i_j}= 0 : 通信路が存在しない
  K_{i_j}= 1 : 通信路が存在する.
- B_{i_j}　プロセッサ・シーケンス番号iのレベルjのバッファ数

PE(P_{i_L})のバッファ数(B_{i_L})は式(1)で計算される.

$$B_{i_L}=\sum_{j=1}^{L}\left(k_{i_j}*(N_{i_j}*(S_{i_j}+R_{i_j})+N_{i_j}-1)\right) \qquad (1)$$

$N_{i_j}-1=0$　上位レベルに通信路がない場合

6.まとめ

階層構造の分散システムにおける送信データ予約方式による転送方法について述べた. この方法には, 1)転送途中のPE/通信路が故障してもその時のシステム状態に応じた最短時間経路でデータが転送される. 2)パケットの中継処理がPE間の結合状況だけで行える利点がある.

今後, シミュレーションにより, 通信量と平均転送時間の関係, PE/ネットワークの故障の影響について調べていく予定ある.

[参考文献]

[1] 伊藤,"自律分散システムはいかにして構成されるか",計測と制御,vol.29,no.10,1990

[2] 森,"自律分散システムと制御分野での実用例", 計測と制御,vol.29,no.10,1990

[3]William.J.Dally,"Virtual-channel flow control",*IEEE Trans. Parallel and Disributed Syst.*,vol.3,no.2,1992