システムソフトウェアと
オペレーティング・システム 77－13
マルチメディア通信と分散処理 87－13
（1998. 2. 26）

オブジェクトグループ通信の階層的配送機構による実装

和田智仁　吉田隆一
九州工業大学 情報工学部

本論文では，現在我々が開発中の分散オブジェクト指向計算環境DOOCEにおけるグループ通信機構の実装方法について述べる．DOOCEグループ通信機構では階層的な配送構造を採用し，グループメンバの局所性を利用したメッセージの配送を行う．グループ通信メッセージの不可分性と原子性を実現するための順序制御および2相コミット操作も，階層構造を利用してそのオーバヘッドを抑制している．また，メッセージの配送にはpoint-to-pointとブロードキャストの通信路を併用し，通信コストを削減している．

A Hierarchical Implementation of Object Group Communication

Tomohito WADA　Takaichi YOSHIDA
Faculty of Computer Science and Systems Engineering,
Kyushu Institute of Technology

In this paper, we describe an implementation of object group communication in our distributed object-oriented computing environment DOOCE. DOOCE employs a hierarchical delivery structure for group communication and delivers messages using locality of group members. Message ordering and two-phase commit protocol to ensure indivisibility and atomicity of group communication messages are also realized hierarchically to decrease the overheads. Messages for groups are transmitted by using not only point-to-point communication but also broadcasts in order to reduce the costs of communication.

1　はじめに

現在，ネットワーク環境上での分散計算を支援する分散オブジェクト指向技術が広まっている．これを用いたアプリケーションでは，複数のオブジェクトを複数の計算機に分散配置し，これらを協調動作させることによって目的の計算を行う．ここで複数のオブジェクトを一まとめにし，オブジェクトグループとして取り扱うことを可能とすれば，これらのオブジェクトを利用するユーザの負担を軽減できると考えられる．

オブジェクトグループに対するメッセージ送信は，複数のオブジェクトが受信者として通信に関与するため，point-to-point 通信とは異なった性質を持っている．このためオブジェクトグループを対象とする通信，すなわちグループ通信には，新たな機構が必要になると考えられる．そこで，現在我々が開発中の分散オブジェクト指向計算環境DOOCEに，グループ通信機構を導入することとした．

DOOCEでは記述言語におけるオブジェクトがグループを構成する．従って，OSで提供されるプロセスグループのためのグループ通信プリミティブなどと異なり，グループの利用はアプリケーションによって多様にわたると考えられる．このため，DOOCEではグループメンバの分布やメンバ数について，拡張性が高く，かつ効率的な実装が必要になる．そこでDOOCEでは階層的なメッセージの配送機構によってグループ通信を実装し，グループメンバの局所性を利用したメッセージの配送を行う

こととした．本論文では，DOOCEにおけるグループ通信機構の設計および実装について述べる．

2 グループ通信の導入

本節ではグループ通信を実装するベースとなるDOOCEについて簡単に説明したのち，DOOCEにおいてグループ通信をどのようにプログラマに提供するかについて述べる．

2.1 分散オブジェクト指向計算環境DOOCE

DOOCEはネットワークで接続された計算機群の上に，OSの提供するアドレス空間をまたがる仮想的なオブジェクト空間を提供する (図1)．DOOCE記述言語によって記述されるオブジェクトは全てこの空間上に生成され，オブジェクト識別子によって同一の方法でアクセス可能となる．これらのオブジェクトはそれぞれ並行に動作することが可能であり，これらはメッセージを受信することによって活性化され，受信したメッセージを順次受理/実行していく．

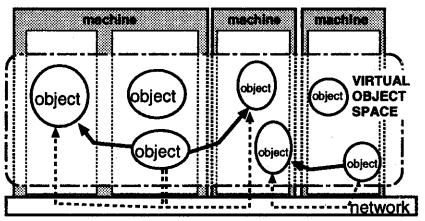

図1: DOOCE仮想オブジェクト空間

DOOCEでは基本的なメッセージ送信の手段として，メッセージ送信後にその返答を待つ同期通信と，返答を待たずに次の動作を開始できる非同期通信の2種類を提供している．

2.2 グループ通信の提供

一般にオブジェクト指向計算では複数のオブジェクトの協調動作によって全体の計算を行う．このため，アプリケーション中で複数のオブジェクトに対して同一のメッセージを送信する場合も多く存在すると考えられる．このような場合に複数のオブジェクトに対して一回の記述でメッセージを送信できるならば，アプリケーションの記述は容易になる．そこでDOOCEでは，オブジェクトの集合をグループとして表現し利用するための枠組を導入している [3]．

ただし，DOOCEのようにオブジェクト間に並行性がある環境でのグループ通信は，複数の受信者オブジェクトが通信に関与するためpoint-to-point通信では起こり得ない不都合が生じる場合がある．そこで，DOOCEではプログラマの負担を軽減し，グループ通信を直観的で利用しやすくするために2つの性質を提供している．一つは，不可分性 (indivisibility) と呼ばれる性質である．これはグループ通信が他のグループ通信によって分割されないことを保証するもので，DOOCEグループ通信はこの性質を常に保証している．この性質によって，プログラマはグループに送信されるメッセージがメンバ間で常に一貫した順序で受信されると仮定することができる．もう一つの性質は，原子性 (atomicity) と呼ばれるもので，グループ通信に原子性が指定されると，そのメッセージは全てのメンバが受信するか，あるいは一つも受信しないかのどちらかとなることが保証される．原子性はグループ通信を利用する際にメッセージ毎に指定できる．この性質を指定することで，プログラマは幾つかのメンバだけがメッセージを受信しないといった場合について考慮する必要がなくなる．

3 グループ通信の設計

グループ通信機構の設計にあたり，以下の点について考慮した．

DOOCEオブジェクトグループ

DOOCEでは記述言語のオブジェクトがグループを構成する．この言語レベルのオブジェクトは環境中の計算機やプロセスなどに比べ一般に数が多いため，グループ中のメンバ数も大きくなる場合がある．また，DOOCE仮想オブジェクト空間は広範囲に広がることが可能で，空間中に存在するグループ数も大きくなる可能性がある．従って，グループ通信機構はメンバ数やグループ数などの規模に依存しない方法によって実現される必要がある．

グループメンバの局所性

一方，DOOCE仮想オブジェクト空間の広がりとは無関係に，グループのメンバは物理的な局所性を持つと考えられる．これは，一つのDOOCEプログラムは一つのアドレス空間上で実行されること

や，DOOCE アプリケーションが (仮想オブジェクト空間に対して) 局所的な計算機/サブネットワーク上で実行されることが多いことから推測できる．グループ通信機構は，グループメンバの物理的な局所性を利用したメッセージの配送を行うことで，配送コストを低減できる．

メッセージの送信方法

グループ通信では同一のメッセージを複数のメンバに配送する必要がある．ブロードキャストのような低レベルのグループ通信プリミティブが利用可能な場合には，それらを利用することによってネットワーク帯域の消費を低減できる．

不可分性の実現

不可分性は，グループの全メンバが一貫した順序でメッセージを受信するように，メッセージ配送順序の一貫性制御を行うことで実現できる．順序制御は一箇所で集中的に行うことで容易に実現できる．しかし，この方法は集中型システムの欠点を持つ．分散的な手法による順序制御は一般に高価であり，さらにその計算コストはメンバ数に比例して大きくなる．グループ通信機構では，効率よくかつ強固な順序制御機構を実現することが望ましい．

原子性の実現

原子性は 2 相コミットを用いたメッセージの配送によって実現できる．2 相コミットを用いた配送では，送信者は全ての受信者からのコミット可否メッセージを集める必要があるため，メンバ数が多くなった場合にこれを収集するコストが問題になる．このコストを削減する方法として，グループを階層化して，階層別にコミットの可否を調査し，集計する方法が考えられる．

以上を考慮して，DOOCE では階層的な配送機構を採用した．DOOCE グループ通信機構は，ユーザオブジェクトを葉とし，グループ通信のためのシステムオブジェクトを節とするツリー構造で表現できる (図 2)．ユーザオブジェクトから発信されるグループ通信メッセージは，必要に応じてこの木をたどり，グループメンバに配送される．

ここでの階層は，アドレス空間 / 計算機 / ブロードキャスト到達可能なサブネットワーク / システム全体 という DOOCE 環境の物理的な構成に対応している．これによってメンバの局所性を利用した配送が可能となる．また，ブロードキャスト到達可能なサブネットワークを一つの階層としたため，ここでのメッセージ交換にブロードキャストを利用できる．順序制御と 2 相コミット操作は，部分木を単位として局所的に行われる．すなわち，言語レベルの個々のオブジェクトを直接対象とするのではなく，限られた数のシステムオブジェクトを対象とした制御が行なわれる．順序制御は下位の層ではコストの低い集中的方法によって実現され，最上位の階層でのみ分散的手法によって実現される．最上位では制御対象の数が (メンバが存在するサブネットワーク数まで) 少なくなるためこのコストもそれほど大きくならないと考えられる．また，分散的手法を採用することで木のルートにシステムオブジェクトを置く必要がなくなるため， scalability を保存し，かつ単点故障によるグループ通信の停止を防ぐことができる．

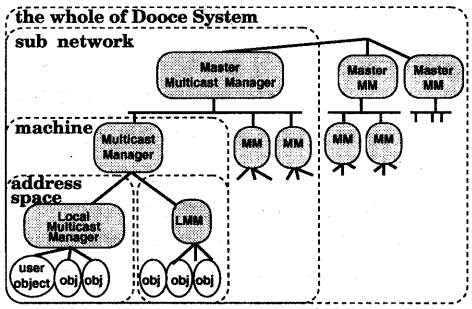

図 2: DOOCE グループ通信機構階層図

4 グループ通信の実装

DOOCE グループ通信機構は，DOOCE システム中のシステムオブジェクトによって実現した．

今回の実装では，DOOCE を利用する各アドレス空間上にローカルマルチキャストマネージャ (LMM) を，また各計算機上にマルチキャストマネージャ (MM) を配置することにした．さらに，ブロードキャスト到達可能なサブネットワーク内の MM からは，それぞれ一つのマスタ MM が選出され，MM としての作業の他に，サブネットワーク内の MM の管理および他のマスタ MM との通信を行う．グループ通信メッセージは，これらのシステムオブジェクトによって転送され，最終的にユーザオブジェクトに配送される．

以下，まずこれらのシステムオブジェクトの基本

的な動作を解説し，次に個々のオブジェクトについて説明する．

4.1 メッセージの転送

システムオブジェクトは，下の階層のオブジェクトからグループ通信要求を受け付けると，まず宛先となるグループメンバの所在を調べる．管理下のメンバ以外がこのグループに含まれている場合には，そのメッセージを上の階層のオブジェクトにグループ通信要求として転送する．

上の階層のオブジェクトからグループ通信配送依頼が到着した場合と，自身が受け付けたグループ通信要求が管理下のメンバだけに宛てたものであった場合には，システムオブジェクトは宛先となるメンバが存在する下の階層のオブジェクトにそれを配送依頼する．

また，全てのシステムオブジェクトはメッセージを逐次的に配送する．すなわち，一つのグループ通信メッセージの配送が終了してから，次の配送を開始する．これによって，グループ通信メッセージの不可分性を保っている．

ここで，先に上の階層に要求したメッセージの宛先に自身が管理するオブジェクトが含まれていて，さらにこのメッセージの配送が未だ終了していない場合に，自身で局所的なメッセージの配送を行うと，メッセージの配送順序に追い越しが発生する．そこで，このような場合にはメッセージの配送を一時保留し，メッセージの追い越しを防いでいる．

4.2 アドレス空間内グループ通信機構

DOOCEにおいては，アドレス空間内にユーザ空間とDOOCEシステム空間が存在する[1]．システム空間にはユーザオブジェクトの管理やメッセージの送受信を行う幾つかのシステムオブジェクトが存在し，ユーザ空間にはユーザが生成するオブジェクトが存在する．今回，システム空間内に新たにLMMシステムオブジェクトを追加した(図3)．このオブジェクトは，アドレス空間内の全てのユーザオブジェクトを担当し，これらからのグループ通信要求の受け付けと，これらへのメッセージの配送を行う．

[1]DOOCEでは仮想的なオブジェクト空間が提供され，プログラマはアドレス空間やこれらのシステムオブジェクトを意識する必要はない．

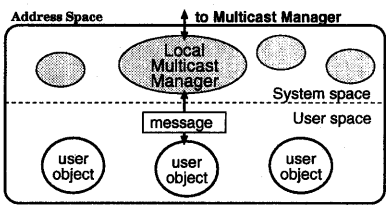

図 3: アドレス空間内グループ通信通信機構

4.3 計算機内グループ通信機構

DOOCEオブジェクトが存在する全ての計算機上には，MMが配置される (図4)．このオブジェクトは，計算機内の全てのLMMを管理する．

MMがマスタMMにグループ通信要求を送信する際には，ブロードキャストを用いる．これによってグループ通信メッセージは，マスタを経由せずに直接全ての(サブネットワーク内)MMに転送される．ただし，このメッセージはマスタから配送可能であることが伝えられるまでは，LMMに配送されない．

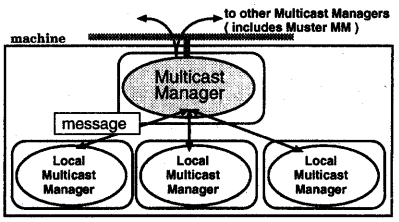

図 4: 計算機内グループ通信通信機構

4.4 サブネットワーク内/間グループ通信機構

サブネットワーク内のMMからはマスタMMが一つ選出され，サブネットワーク内の全てのMMを管理し，またこのネットワークの代表として他のマスタMMとメッセージの交換を行う (図5)．

マスタMMへのグループ通信要求はブロードキャストによってサブネットワーク内のMMから届けられる．マスタは，これらのメッセージと外部のネットワークから受信したメッセージを一意に順序付け，この順序でメッセージの配送を行うことをMMに指示する．この指示はブロードキャストを使って各MMに伝えられるが，この2フェーズの

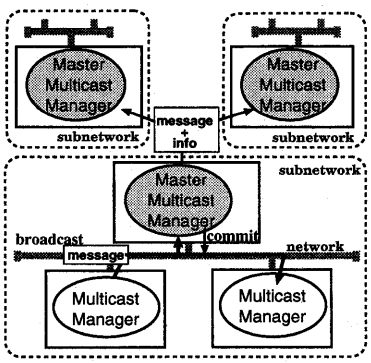

図 5: ネットワーク内/間グループ通信通信機構

通信は，ブロードキャストメッセージの紛失も防止している [2]．また，マスタ MM 間で交換されるグループ通信メッセージは total ordering [1] の手法を用いて順序付けされる．

4.5 原子性の保証

グループ通信に原子性を保証することが指定された場合には，2 相コミットを用いた配送が行われる．DOOCE では階層構造を利用して，コミット可否の確認を分散的に行う．

ここでの 2 相コミット操作は，グループ通信メッセージが到達した最上位の階層のオブジェクトをコーディネータとして行われる．さらに，サブオーディネートとなるオブジェクトにおいてもそれ自身がコーディネータとなりコミット操作を行う．これによって，グループメンバ数が増加した場合にコーディネータにかかる負荷を分散することができる．

5 実験

今回実装した DOOCE グループ通信機構を用いて，グループ通信に必要となるコストを調査する実験を行った．ここでは，グループ通信を用いてオブジェクトグループのメンバにメッセージを送信した場合と，point-to-point 通信を用いてこれらのメンバにメッセージを送信した場合について測定した結果を示す．さらに，グループメンバの存在に局所性が存在する場合に，グループ通信にかかる時間がどの程度短縮されるかについても実験した．

計算機は Sun SS10 と SS20 を用いた．それぞれの計算機は 10M の Ethernet によって接続されている．なお，数値は全てのメンバから返答を受け取るまでの時間を，ユーザプログラム側で測定したものである．

実験 1: メンバを 3 つのサブネットワークにおける 6 台の計算機上に分散し，そのうちの 1 つのサブネットワーク内からメッセージを送信した．結果を図 6 に示す．

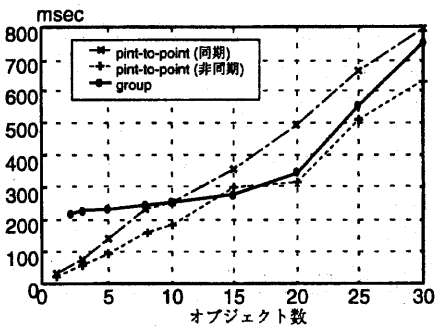

図 6: 広域に分散する場合

DOOCE グループ通信は不可分性を保証しているにもかかわらず，point-to-point 通信を複数回繰り返すのと大差ない時間で通信が完了していることがわかる．これは，マルチキャストマネージャ間の通信にブロードキャストを利用することで，必要になる通信の回数を削減しているためと考えられる．

実験 2: メンバを同一サブネットワーク内の 6 台の計算機上に分散し，サブネットワーク内の他の計算機からこれらにメッセージを送信した．結果を図 7 に示す．

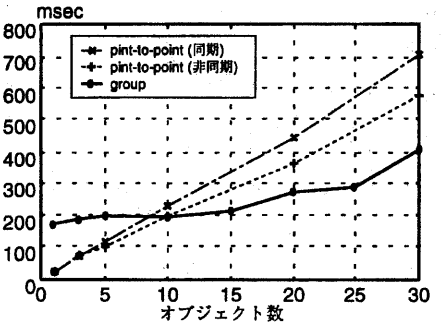

図 7: ネットワーク内に分散する場合

ネットワーク内部にのみメンバが存在する場合

には，マスタ間での total ordering による順序付けアルゴリズムを実行する必要がない．このため，実験1の結果よりも15%程度短い時間で通信が完了している．

実験3: メンバを一台の計算機上の6つの異なるアドレス空間に分散し，別のアドレス空間からこれらにメッセージを送信した．結果を図8に示す．

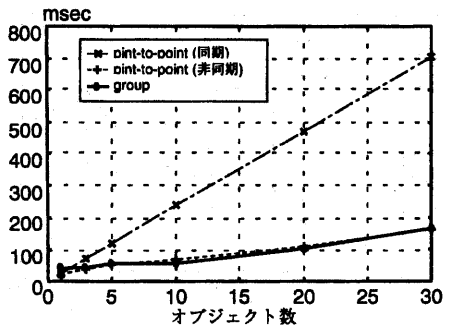

図8: 計算機上に分散する場合

ネットワークを経由する通信が全く行われないため，前の実験に比べ非常に短い時間で通信を終了している．また，グループ通信ではメンバに一斉にメッセージが送信されるため，非同期通信と同様，メンバ間の並行性による効果が表れている．

実験4: メンバを全て同一アドレス空間上に置いた場合．結果を図9に示す．

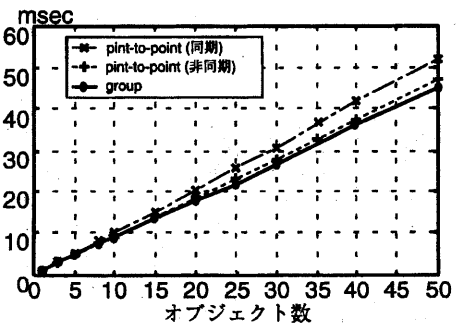

図9: 単一アドレス空間内に存在する場合

ここではプロセス間通信が全く不要なため，非常に短い時間で通信を終了している．また，これらは全て単一プロセス内で処理が行われるため，メンバ間の並行実行による効果は見られない．

結果:

これらの実験全てにおいて，グループ通信は point-to-point 通信をメンバの数だけ繰り返すのと同程度の時間で終了することがわかった．また，メンバの存在が局所的な場合には，その局所性に応じて通信時間が短くなることも確認できた．このため，不可分性は必要としないがグループを用いて記述を簡略化する，といった利用であっても，アプリケーションに不利益はないと言える．ただし，今回の実験では同時に別のグループ通信が発生していないため，システムオブジェクトでの処理がボトルネックとなることはなかった．同時多発的にグループ通信要求が発生した場合には，この実験の結果よりも多少時間が必要になることが予想される．

6 おわりに

グループ通信を階層的な配送構造により実現することで，拡張性 (scalability) が高く，通信時間の小さいグループ通信を実現できた．特に，グループのメンバに局所性がある場合には短い時間でメッセージを配送できることを確認した．

階層的な実装では，上位の階層の故障によって通信が行えなくなる可能性がある．今回の実装では最も上位には管理機構を設置せず，分散的な手法によってメッセージの伝達/順序制御を行うようにして，単点故障によるグループ通信の停止を防いでいる．途中の階層についても同様のことが言えるが，これらの対応は行っていない．耐故障性を高めるために，今後これらの故障への対応について考える必要がある．具体的には，マルチキャストマネージャの故障の検出に関する実装，およびマスタマルチキャストマネージャの交替/選出に関する実装が必要になると考えられる．

参考文献

[1] M. Dasser, "TOMP A Total Ordering Multicast Protocol", *Operating Systems Review*, vol.26, no.1, pp.32-40 Jan. 1992

[2] M.F. Kaashoek, A.S. Tanenbaum, S.F. Hummel, H.E. Bal "An Efficient Reliable Broadcast Protocol", *Operating Systems Review*, SIGOPS, vol23, pp.5-10, Oct. 1989

[3] 和田, 吉田, "分散オブジェクト指向計算環境におけるグループ通信", 情報処理学会九州支部研究会報告, pp.143-152, Mar. 1997