大規模網のための送信順序保存放送通信手順

4P−3

高村 昌興, 中村 章人, 滝沢 誠

東京電機大学

1 はじめに

グループウェア等の新しい分散型応用で、*OSI*といった従来の1対1通信プロトコルに加えて、複数エンティティ間のグループ通信プロトコルが研究開発されてきている。これらは、通信エンティティ数が数十程度までの網を対象としており、数百以上のエンティティを含んだ大規模網では、負荷が大きくなり、実用的でない。例えば、放送通信プロトコル[1, 2, 3]では、エンティティ数nに対して、PDU数と処理負荷は、$O(n)$から$O(n^2)$になる。本論文では、同一通信チャネルに接続された複数のエンティティを、論理的に分割することにより、処理負荷を減少させることを試みる。ここで、エンティティのグループを群とする。

2章では、群内で、受信の原子性と順序を保証する**送信順序保存放送通信**(*OP*)**プロトコル**[3]の概要を示す。3章では、群を論理的に分割した要素群とそれらの間を取り持つゲートウェイエンティティのモデルについて述べる。4章では、*OP*を拡張した**大規模群送信順序保存放送通信**(*LOP*)**プロトコル**について述べる。

2 送信順序保存放送通信(*OP*)プロトコル

2.1 サービス

単一チャネル(*1C*)**サービス**[3]は、*Ethernet MAC*や、無線システムで提供されるサービスを抽象化したものである。ここでは、各エンティティは、同順序で全PDUを受信するが、バッファ溢れ、オーバーランにより、PDUを受信し損ねる場合がある。*OP*プロトコルは、*1C*サービスを利用し、まず、複数エンティティ$E_1,\ldots,E_n$間に群Cを確立する。群は、二つのエンティティ間のコネクション概念を、$n(\geq 2)$個のエンティティ間に拡張した概念である。Cが確立された後、一意な群識別子が得られる。群識別子を持ったPDUが、C内のエンティティに送られる。*OP*プロトコルでは、各PDUは、C内の送信元の順序を保存しながら全エンティティに送られる。

2.2 三相による正しい受信概念

エンティティ$E_1,\ldots,E_n$が提供する群Cで、E_kが放送したPDU pは、**受理**、**前確認**、**確認**といった三相手続きにより正しい受信が行なわれる。

受理　pがE_jで受信される

前確認　E_jが、「pの全宛先がpを受信した」ことを知る

確認　「pの各宛先は、『全宛先がpを受信した』ことを知っている」ことを知る□

2.3 データの転送と受信

pはPDUを示し、$p.F$により、pの属性Fを示すとする。$p.CID$は、Cの群識別子である。$p.SRC$は、pの送信元E_kである。$p.SEQ$は、E_kが送信したPDUの送信シーケンス番号である。$p.ACK_j$は、E_kがE_jから次に受信予定のPDUの受信シーケンス番号である$(j=1,\ldots,n)$。$p.DATA$は、データである。

E_kは、各E_jに対して、送受信したPDUの履歴を示す受信ログRL_{kj}と送信ログSL_kを持つ。これは、各々、受理、前確認、確認されたPDUが記録される副受信ログRRL_{kj}、PRL_{kj}、ARL_{kj}から構成される。E_kは、送信ログSL_kを持つ。E_kは、PDUを放送するために変数SEQとREQ_jを持つ。E_kが、pを放送するとき、$p.SEQ{:=}SEQ$、$p.ACK_j{:=}REQ_j$とし、$SEQ{:=}SEQ{+}1$とする。

E_kは、各E_jに対して、受信用変数REQ_jを持つ。E_jからのpが、E_kに届いたとき、$REQ_j{=}p.SEQ_k$ならば、E_kはpを受理し、REQ_jに1が加えられる。

E_kは、$n{\times}n$行列ALとPALを持つ。ALは、E_kが「E_iが、SEQが、AL_{ij}以下であるPDUをE_jから受信した」ことを知っていることを示す。E_iからのpが、E_kで受信されたとき、$AL_{ij}{:=}p.ACK_j$となる。$min(AL_{1j},\ldots,AL_{nj})$は、$SEQ$がこれ以下のシーケンス番号を持つ$E_j$からのPDUが、全エンティティで受信されたことを、即ち、E_kで前確認されたことを示す。

PALは、E_kが「E_iが、SEQが、PAL_{ij}以下であるE_jからのPDUを前確認した」ことを知っていることを示す。E_iからのpが、E_kで前確認されたとき、$PAL_{ij}{:=}p.ACK_j$となる。$min(PAL_{1j},\ldots,PAL_{nj})$以下である$E_j$からのPDUが、全エンティティで前確認されたことを、即ち、E_kで確認されたことを示す。

3 複合群とゲートウェイ

従来の群を、平群とする。平群の*OP*プロトコルでは、処理負荷が$O(n^2)$であることから、大規模システムには適用できない。このために、一つの通信チャネルに接続されたエンティティを、論理的に分割し、それらの代表者が隣接している群に対して、受信状態を知らせることで、処理負荷の減少を試みる。ここで、分割された複数の群(**要素群**)から構成される群を、**複合群**とする。そして、要素群の代表者であり、かつ、それらの間の通信を行なうエンティティをゲートウェイGEとする[図1]。

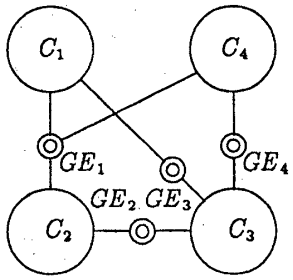

図1: 複合群CとゲートウェイGE

複合群Cは、要素群$C_1,\ldots,C_h$から構成される。$C_1,\ldots,C_h$は、ゲートウェイ$GE_1,\ldots,GE_g$により接続される。従って、複合群$C{=}<C_1,\ldots,C_h,GE_1,\ldots,GE_g>$とする。各$GE_i$は、一般に$k(\geq 2)$個の要素群$C_{i1},\ldots,C_{ik}$を相互接続す

Order-Preserving Broadcast Communication for Large Scale Network

Masaoki Takamura, Akihito Nakamura, and Makoto Takizawa

Tokyo Denki University

る。このkをGE_iの結合度とする。図1の例では、GE_1の結合度3、GE_2、GE_3、GE_4の結合度2である。ここで、複合群は、以下の性質を持つ。

[性質] 任意の二つの要素群C_iとC_j間には、必ず1つのゲートウェイエンティティが存在するとする。□

各C_j内で、PDUが、OPプロトコルに従って、C内に放送される。このとき、各要素群が、Cより少ないエンティティを含むので、各エンティティでの処理負荷を減少できる。更に、もし、各要素群が、一定数m以下の固定エンティティを含むならば、各エンティティのプロトコル処理の負荷を一定値$O(m^2)$に押えられる。

4 複合群でのOPプロトコル

4.1 PDU形式

複合群$C=<C_1,\ldots,C_h,GE_1,\ldots,GE_g>$を考える。各$C_i$は、エンティティを$E_{il}(l=1,\ldots,m_i)$とゲートウェイ$GE_{ij}(j=1,\ldots,g_i)$を持つ。ここで、$GE_{ij}$を$E_{it}(t{=}m_i{+}j)$とも書く。$E_{is}(s{=}1,\ldots,c_i(=m_i{+}g_i))$で放送される各PDU pは、OPプロトコルのPDU形式に加えて、$LSRC$を持つ。これは、$p.SRC$は、C_i内の送信元エンティティ(E_{il})であるのに対し、$p.LSRC$は、PDUを放送した送信元エンティティである。例えば、E_{is}がゲートウェイでないときは、$LSRC=SRC$である。E_{is}がゲートウェイのとき、pを最初に送信したエンティティである。

4.2 手順

要素群C_iを考える。C_i内で、ゲートウェイでないエンティティE_{il}を通信エンティティとする。C_iは、通信エンティティ$E_{i1},\ldots,E_{im_i}$から構成されるとする。更にC_iは、ゲートウェイ$GE_{i1},\ldots,GE_{ig_i}$により接続されているとする。即ち、$C_i=<E_{i1},\ldots,E_{im_i},GE_{i1},\ldots,GE_{ig_i}>$である。通信エンティティ$E_{il}$は、平群の$OP$プロトコルと同一の動作を行なう。ゲートウェイ$GE_{ij}$の動作について考える。

C_i内のあるGE_{ij}が、$C_{ij1},\ldots,C_{ijf_{ij}}$を接続するとする$(j=1,\ldots,g_i)$。$GE_{ij}$は、各要素群$C_r$に対する処理を行なう副エンティティ$SGE_{ijh}$から構成される$(h=1,\ldots,f_{ij})$。

SGE_{ijh}が、E_{is}で放送されたpを受信するとする。このとき、$p.SRC{=}p.LSRC$ならば、SGE_{ijh}は$SGE_{ijs}(h\neq s)$にpを渡す。各SGE_gは、pをC_g内で放送する。$p.SRC{\neq}p.LSRC$ならば、SGE_{ig}に渡さない。これは、$p.SRC{=}p.LSRC$ならば、E_{ij}が、実際の送信元エンティティであり、隣接している要素群内に放送しなければいけない。しかし、$p.SRC{\neq}p.LSRC$ならば、E_{ij}は別のゲートウェイであるので、これ以上、隣接している要素群に放送する必要がない。

GE_{ij}は、以下の受理と送信動作を行なう。

(1) C_iのE_{is}が、pを放送したとする。$p.SEQ{=}REQ_j$ならば、SGE_{ijh}は、pを受信する。そして、$REQ_j{=}REQ_j{+}1$となる。$p.SRC{\neq}p.LSRC$ならば、pを$SGE_{l_1},\ldots,SGE_{l_g}$に送る。$p.SRC{=}p.LSRC$ならば、$OP$プロトコルに従って、$p$の確認通知を含んだPDUを$C_i$に放送する。

(2) SGE_{ijh}が、SGE_{ijr}からpを受信したとする。このとき、次のPDU qを作る。$q.CID{:=}C_i$、$q.SEQ{:=}SEQ$、$SEQ{:=}SEQ{+}1$、$q.SRC{:=}p.SRC$、$q.LSRC{:=}GE_{ik}$、$q.DATA{:=}p.DATA$、$q.ACK_j{:=}REQ_j$を行なう$(j=1,\ldots,i)$。この後、qを放送する。qをLSLに、pをLRL_{ijr}に記憶する。LSL内のqに対して、$q.ptr{=}p$なるポインタptrを設ける。

次に、図2の例を用いて前確認と確認動作について考える。まず、C_1内のE_{11}が、pを放送したとする。GE内のSGE_1は、pを受信してSGE_2に渡す。SGE_2は、pからqを作り、qをC_2内に放送する。C_1内で受信したエンティティE_{11}、E_{12}、E_{13}から、pについての受理情報を含んだPDUを受信したならば、SGE_1は、SGE_2にこれを放送する。SGE_1は、GE内の全副ゲートウェイから、この通知を受けたとき、pの受理情報を含んだPDU rをC_1内に放送する。即ち、$r.ACK_1{:=}p.SEQ{+}1(:=REQ_1)$であり、各エンティティがこれを受信したとき、$p$が前確認される。。同様に$r$について、前確認が行なわれたとき、$p$の確認がされる。

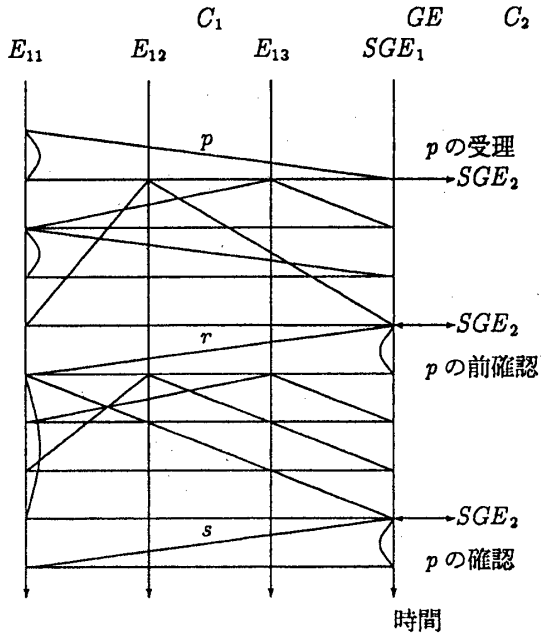

図2: PDU pについての三相手続き

5 結論

本論文で、大規模なシステムの複数エンティティ間での高信頼放送通信を有効に行なうための手順について述べた。本プロトコルでは、$1C$サービス網に接続されたエンティティを、要素群として論理的に分割することで、各エンティティの処理負荷を減少できる。

参考文献

[1] Nakamura, A. and Takizawa, M.: *Reliable Broadcast Protocol for Selectively Ordering PDUs, Proc. of the 11th IEEE International Conf. on Distributed Computing Systems (ICDCS-11)*, pp.239-246 (1991).

[2] Nakamura, A. and Takizawa, M.: *Priority-Based Total and Semi-Total Ordering Broadcast Protocols, Proc. of the 12th IEEE International Conf. on Distributed Computing Systems (ICDCS-12)*, pp.178-185 (1992).

[3] Takizawa, M. and Nakamura, A.: *Partially Ordering Broadcast (PO) Protocol, Proc. of the 9th IEEE Conf. on Computer Communications (INFOCOM)*, pp.357-364 (1990).