大規模網送信順序保存放送通信プロトコルの評価

1F－1

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分散型応用システムを実現するために、複数のエンティティに対して高信頼な通信サービスを提供するためのグループ通信プロトコルが必要とされている。これまでのグループ通信プロトコル(例えば、[1, 2])は、グループ内のエンティティ数nに対して$O(n^2)$の通信負荷があり、大規模なシステムでは性能面で有効でない。本論文では、同一の通信チャネルに接続された複数のエンティティを、小グループに分割し、各グループの代表(ゲートウェイ)エンティティがグループ間通信を行うことで、処理負荷を減少させる方法を示す。また、従来のプロトコルとの処理負荷の比較により、本方式の評価を行う。

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[2]の概要を示す。3章では、システムのモデルを示す。4章では、OPを拡張した大規模群送信順序保存放送通信(COP)プロトコル[3]を述べる。5章では、群構成アルゴリズムを述べる。6章では、本プロトコルの評価について述べる。

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[2]は、高速通信網で提供されるサービスを抽象化したものである。$1C$サービスを利用した場合、各エンティティは、同じ順序でPDUを受信するが、バッファの溢れ等により、PDUを受信し損ねる場合がある。ここで、エンティティのグループを群とする。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。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$E_1,\ldots,E_n$から構成される群Cで、各PDUpは、以下によって受信される。

受理	pが各E_jで受信される
前確認	各E_jが、「全宛先がpを受信した」ことを知る
確認	「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

$p.F$は、PDUp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の送信元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、PDUの送受信履歴を示す受信ログRL_{kj}と送信ログSL_kを持つ。E_kは、送受信用の変数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、$SEQ$がこれ以下である($E_j$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。SEQが、$min(PAL_{1j},\ldots,PAL_{nj})$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、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(平群)でのOPプロトコルでは、処理時間が$O(n^2)$である。このために、エンティティ集合を分割し、それらの代表者が隣接している群に、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、分割された群(要素群)から構成される群を、複合群とする。要素群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]。

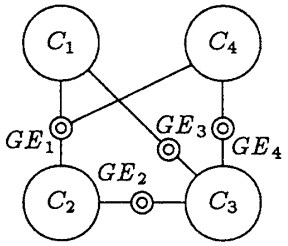

図1: 複合群CとゲートウェイGE

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($C = <C_1,\ \ldots,\ C_h,\ GE_1,\ \ldots,\ GE_g>$)。各$GE_i$は、一般に$k(\geq 2)$個の要素群$C_{i1},\ \ldots,\ C_{ik}$を相互接続する。この$k$を$GE_i$の結合度とする。ここで、複合群は、以下の性質を持つ。

Evaluation of Order Preserving Broadcast Protocol for Large Group
Masaoki Takamura, Akihito Nakamura, and Makoto Takizawa
Tokyo Denki University
e-mail : {taka, naka, taki}@takilab.k.dendai.ac.jp

[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。□

各 C_j 内で、PDUは、OPプロトコルに従って処理される。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。

4 複合群でのOPプロトコル

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 p は、OPプロトコルのPDU形式に加えて、$LSRC$を持つ。$p.LSRC$ は、PDUを放送した送信元エンティティである。E_{is} がゲートウェイでないとき、$LSRC = SRC$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。

5 群構成アルゴリズム

各通信エンティティは、同程度の処理速度を持つとする。このとき、$\mathcal{E} = \{E_1, \ldots, E_n\}$ から、各エンティティが同じ処理時間を持つように、複合群を構成することを試みる。まず最初に、$\mathcal{E}$ を、h $(1 \leq h \leq n)$ の要素群 $C_1, \ldots, C_h$ に分割する。このとき、$m = \lfloor n/h \rfloor$ に対して、m 個の通信エンティティを持つ $h_1(\geq 1) = h - h_2$ 個の要素群と、$m + 1$ 個の通信エンティティを持つ $h_2(\geq 0) = n \ mod \ h$ 個の要素群から構成される。そして、h 個の要素群から、1つの要素群を取り出し (C_0)、残りの $h - 1$ 個の各要素群で、$k - 1$ 個の要素群から構成される副グループを p 個と、q $(= (h - 1) \ mod \ (k - 1))$ 個の要素群から構成される副グループを $r(\leq 1)$ 個 $(G_1, \ldots, G_{p+r})$ 作成する。そして、以下の手続きにより、ゲートウェイによって接続される。

[ゲートウェイ接続手続き]

1. 要素群 C_0 と各副グループ $G_1, \ldots, G_p$ を、それぞれ、ゲートウェイで接続する $(\langle C_0, C_{i0}, \ldots, C_{i,k-2} \rangle, i = 1, \ldots, p)$。$q > 0$ ならば、G_{p+r} と接続する $(\langle C_0, C_{p+1,0}, \ldots, C_{p+1,q-1} \rangle)$。

2. $q > 0$ ならば、$C_{10}, C_{2i}, \ldots, C_{pi}, C_{p+1,i}$ が、ゲートウェイで接続される $(i = 0, 1, \ldots, q - 1)$。$q \geq 0$ ならば、$C_{10}, C_{2i}, \ldots, C_{pi}$ が、ゲートウェイで接続される $(i = q, q + 1, \ldots, k - 2, \ q \geq 0)$。

3. $q = 0$ ならば、$C_{1i}, C_{2j}, C_{3,j+1}, \ldots, C_{p,j+p-2}$ が、ゲートウェイで接続される $(i = 1, \ldots, k - 2, \ j = i, i + 1, \ldots, i + k - 2)$。ここで、$j = j' \ mod \ (k - 2)$ ならば、$C_{ij} = C_{ij'}$ である。$q > 0$ ならば、$C_{1i}, C_{2j}, C_{3,j+1}, \ldots, C_{p,j+p-2}, C_{p+1,j+p-2}$ が、ゲートウェイで接続される $(i = 1, \ldots, k-2, \ j = i, i+1, \ldots, i+k-2, \ j+p-2 \leq q-1)$。$C_{1i}, C_{2j}, C_{3,j+1}, \ldots, C_{p,j+p-2}$ が、ゲートウェイで接続される $(i = 1, \ldots, k - 2, \ j = i, i + 1, \ldots, i + k - 2, \ \ j + p - 2 > q - 1)$。□

6 評価

複合群 C 内の全通信エンティティ数を n、全ゲートウェイ数を g とする。TEP/n を最小とする h と k に対して、COPプロトコルの処理時間[図2]とPDU長[図3]を示す。

TEP/n^3　各エンティティの処理時間。
CEP/n^3　各エンティティの処理時間。
GEP/n^3　ゲートウェイ全体の処理時間。
$GEP/(gn^2)$　各ゲートウェイの処理時間。

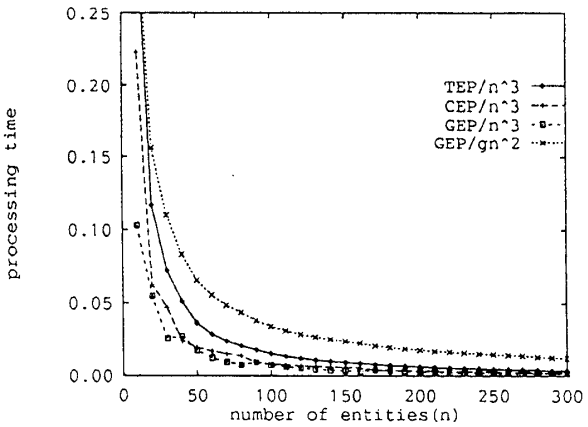

図 2: 処理時間の比率

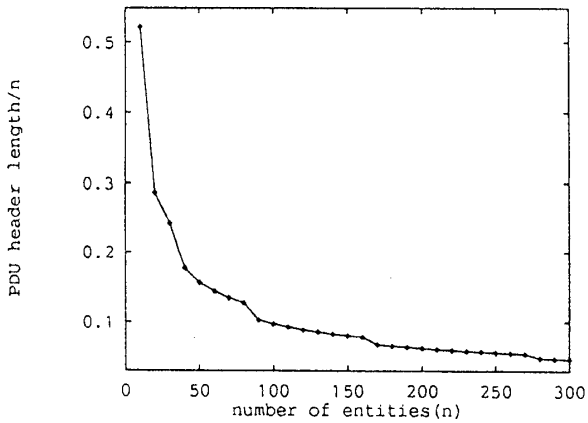

図 3: PDUヘッダ長の比率

各図より、本プロトコルは、従来のOPプロトコルと比較して、処理時間を減少できることが分かる。

7 おわりに

本論文では、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、高速網に接続されたエンティティを、複数の要素群に分割することで、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 IEEE ICDCS-11*, 1991, pp.239-246.

[2] Takizawa, M. and Nakamura, A., "Partially Ordering Broadcast (PO) Protocol," *Proc. of IEEE INFOCOM90*, 1990, pp.357-364.

[3] Takizawa, M. Takamura, M. and Nakamura, A., "Group Communication Protocol for Large Group," *to appear in the proc. of 18th IEEE Conf. on Local Computer Networks*, 1993.