単一チャネル上の選択的全順序放送型通信プロトコルのデータ転送手続き

立川敬行† 中村章人†† 滝沢誠†

グループウェア等の分散型応用では，複数エンティティ間での高信頼なグループ通信が必要となる．グループ通信の中で，各エンティティが各データ単位をグループ内の一部のエンティティに送信する選択的グループ通信が必要となる．さらに，複数のエンティティから送信されたデータ単位を，これらの共通の宛先エンティティが，同一の順序で受信せねばならない場合がある．これを，選択的全順序放送型通信 (ST) サービスとする．本論文では，高速な放送型通信網を利用して，ST サービスを提供するプロトコルについて述べる．本プロトコルは，データ単位の紛失が生じるもとで，紛失したデータ単位のみを再送し，選択的全順序性を保つために，受信キュー内のデータ単位を一定の順序で整列するものである．また，ST プロトコルの性能を，データ単位の紛失に対する再送 PDU 数により評価する．

Data Transmission with Selectively Totally Ordered Broadcast Protocol on Single Channel Network

TAKAYUKI TACHIKAWA,† AKIHITO NAKAMURA†† and MAKOTO TAKIZAWA†

In this paper, we discuss how to provide a kind of group communication for multiple entities in distributed systems by using a high-speed broadcast communication channel. In the group communication, a protocol data unit (PDU) sent by each entity has to be delivered atomically to all the destinations in the group. In distributed applications, each entity rather sends a PDU to only a subset than all the entities, and each entity receives only PDUs destined to it from every entity in the sending order. Furthermore, every common destination of the PDUs sent by multiple entities receives them in the same order. We name such a broadcast service a *selectively totally ordering broadcast* (ST) service. This paper discusses the ST protocol by using a high-speed broadcast network in the presence of PDU loss. While the conventional broadcast protocols adopt the centralized control scheme, this protocol is based on the distributed control scheme. In the ST protocol, PDUs lost are selectively retransmitted and the PDUs are sorted by every entity so as to order the PDUs received in the ST service. We discuss the performance of the ST protocol.

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は，複数のエンティティ間での協調動作が必要となる．こうした応用では，エンティティのグループ内での信頼性のある通信が必要となる．高信頼放送型通信は，グループ内で送信されたデータ単位が，全宛先で，ある定められた順序で原子的[2),13),14),18)]に受信されるものである．エンティティ間で交換されるデータ単位をプロトコルデータ単位 (PDU) とする．放送型通信プロトコルについては，文献 1)～4), 9)～16) 等で論じられている．これらは，各 PDU をグループ内の全エンティティに原子的に送信するプロトコルである．

データベースサーバとクライアントからグループが構成されている場合を考えると，クライアントは，必要とするデータを含むサーバに対してのみ，データ操作要求を送信する．このように，各エンティティから送信された PDU が，グループ内の全エンティティに送信される必要のない分散型応用がある．グループ内の一部のエンティティに放送することを，選択的放送とする．選択的放送の一種として，選択的半順序放送型通信 (SP) サービス[9)~11)]が論じられている．ここで

† 東京電機大学理工学部経営工学科
Department of Computers and Systems Engineering, Faculty of Science and Engineering, Tokyo Denki University
†† 電子技術総合研究所
Electrotechnical Laboratory

は，各エンティティから送信されたPDUは，全宛先で送信順に受信されるが，異なるエンティティからのPDUの受信順序は，エンティティ間で同一とは限らない．これに対して，例えば，二つのエンティティ A と B が，C と D にPDUを送信したとき，C と D が同じ順序でPDUを受信するサービスがある．これを選択的全順序放送型通信（ST）サービスとする．ISIS等では，各エンティティが，あらかじめ定められた複数のグループにPDUを送信し，グループ間での受信順序を一定にできる．これに対して，STプロトコルでは，各エンティティはグループ内の任意のエンティティに，任意の時に送信できるものである．

文献19)では，網は高信頼であり，網障害はないものとしている．高速網[6]では，転送速度が処理速度より速いために，エンティティがPDUの受信に失敗する場合がある．これをモデル化したものを単一チャネル（1C）[13),14)]とする．本論文では，1Cサービスを用いて，STサービスを提供するSTプロトコルを示す．STプロトコルは，各エンティティが非同期にPDUを送受信しながら，全宛先での原子的受信の判断を，分散型制御により行うものである．ISIS[19]とDelta-4(xAMp)[20]，AMOEBA[21]は，それぞれ，原子的受信の判断を送信元エンティティとあるエンティティが指揮エンティティとなる集中型である．集中型制御のアルゴリズムは単純であるが，指揮エンティティの障害に対して頑強でなく，負荷の集中，PDU待ちによる遅延時間の増加の問題がある．

本論文の2章では，STサービスを定義する．3章では，STプロトコルのデータ転送手続きを論じる．4章では，障害に対する復旧手順を述べる．5章では，STプロトコルの正しさと性能について述べる．

2. 選択的全順序（ST）サービス

本章では，複数のエンティティのグループに対するSTサービスを定義する．

2.1 群

通信システムは，**図1**に示す三階層から構成される．システム層のエンティティは，網層が提供する網サービスを利用して，応用層に高信頼な放送型通信を提供する．群とは，二つのサービスアクセス点（SAP）間で定義されている従来のコネクションを，n（≧2）個のSAP間に拡張した概念[13]である（図1）．すなわち，群 C は，n（≧2）個のシステムSAP $S_1,\cdots,S_n$ の組である．システムエンティティ E_i は，網サービスを利用し，S_i を通じて応用エンティティ A_i にサービスを提供する（$i=1,\cdots,n$）．このとき，C は，$E_1,\cdots,E_n$ によって提供されるとし，$C=\langle E_1,\cdots,E_n\rangle$ と書く．応用エンティティ $A_1,\cdots,A_n$ のグループが群を利用して，グループ通信を行える．

2.2 放送型通信サービスの性質

群に対する放送型通信サービス[15]を，PDUの系列であるログの集合としてモデル化する．ここで，m 個のPDUから成るログ L を $<p_1\cdots p_m]$ と書く．$top(L)$ は先頭の p_1，$last(L)$ は最後尾の p_m，p_h は h 番目のPDUとする．$L[h]$ は p_h を示し，h を**索引**とする．$h<k$ のとき，p_h は p_k に L 内で先行する（$p_h\rightarrow_L p_k$）とする．p_{h+1} は，p_h の直後にあるとする．$L|^h$ は $<p_1\cdots p_h]$，$L|_h$ は $<p_h\cdots p_m]$ を示す（$h\leq m$）．

群を提供する各 E_i は，それぞれ送受信したPDUの履歴である送信ログ SL_i と受信ログ RL_i を持つ（$i=1,\cdots,n$）．E_i が，p の後に q を送信したならば，SL_i 内で p は q に**先行**する（$p\rightarrow_{SL_i} q$）．p の後に q を受信したならば，RL_i 内で $p\rightarrow_{RL_i} q$．副送信ログ SL_{ij} を，E_i が E_j に送信したPDUから成る SL_i の部分系列とする．

受信ログ間の関係を，以下に定義する．以下，p と q は任意のPDUを示すとする．

[定義] RL_i と RL_j の両方に含まれる p と q について，$p\rightarrow_{RL_i} q$ ならばかつそのときに限り $p\rightarrow_{RL_j} q$ であるとき，RL_i と RL_j は**順序同値**である．RL_i と RL_j が同一のPDU集合を持つとき，RL_i と RL_j は**情報同値**である．□

RL_i と RL_j が順序同値かつ情報同値であるとき，両者は**同値**である．RL_i と RL_j が順序同値である場合，E_i と E_j の両方が受信したPDUの受信順序は同一であるが，両者が同じPDUを受信しているとは限らない．RL_i と RL_j が情報同値である場合，E_i と E_j は同じPDUを受信しているが，受信順序は同一とは限らない．RL_i と RL_j が同値である場合，E_i と E_j は同じPDUを同一順序で受信する．

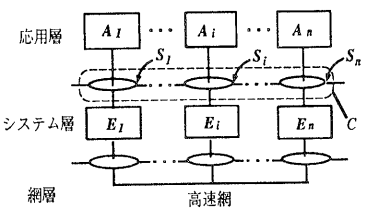


図1 群 C
Fig. 1 Cluster C.

送信ログに対する受信ログの性質を定義する．

[定義] E_i が，各 E_j から受信した p と q について，$p \rightarrow_{RL_i} q$ ならばかつそのときに限り $p \rightarrow_{SL_j} q$ であるとき，RL_i は**順序保存**である．$SL_1, \cdots, SL_n$ 内の PDU の和から RL_i が構成されるとき，RL_i は**情報保存**である．□

RL_i が順序保存でかつ情報保存であるとき，RL_i は**保存**する．RL_i が順序保存であるならば，E_i は各 E_j が送信した PDU を送信順に受信する．RL_i が情報保存である場合，E_i は各 E_j が送信した全 PDU を受信する．

各 PDU が群内の一部を宛先とするとき，PDU は，群内で選択的に放送されるとする．

[定義] RL_i の PDU 集合が $SL_{1i}, \cdots, SL_{ni}$ の PDU の和であるとき，RL_i は**選択的情報保存**である．□

RL_i が選択的情報保存ならば，E_i は，自分宛の全 PDU を受信している．

[定義] RL_i が順序保存かつ情報保存であるとき，RL_i は正しい．正しい RL_i が選択的情報保存であるとき，RL_i は選択的に正しい．□

通信サービス S 内の各受信ログが正しいか，または選択的に正しいならば，S は信頼性があるとする．そうでないとき，S を低信頼とする．

2.3 原子的受信の分散型決定

群 $C=\langle E_1, \cdots, E_n \rangle$ を考える．各 E_i が送信する PDU には，E_i が各 E_j から受信した PDU に対する受信通知が含まれるとする．E_i が C 内で送信した PDU p が，全宛先で正しく受信されたかどうかを，どのように判断していくかが問題となる．p が全宛先で受信されたときのみ，各宛先で受信することを原子的受信とする．原子的受信の決定方法として，集中型，非集中型，分散型の三種[18]がある．集中型と非集中型では，ある指揮エンティティが決定を行う．これに対して，分散型制御では，群内の各エンティティが決定する．分散型制御で，各 E_k が原子的受信の判断を行う基準として，以下の(1)～(3)の三段階[13),14)]がある．ここで，p の宛先を $p.DST$ と書く．$p.DST=\{E_{d_1}, \cdots, E_{d_m}\}(\subseteq C)$ を明示するときに，$p_{\{d_1, \cdots, d_m\}}$ と書く．

(1) **受理**：E_k が，E_i から p を受信し，$E_k \in p.DST$ ならば，p は E_k で受理されたという．

(2) **前確認**：E_k が，全 $E_j \in p.DST$ から，p の受信通知を含む PDU を受信したとき，p は E_k で前確認されたという．

(3) **確認**：E_k が，全 $E_j \in p.DST$ から，p を前確認する PDU の受信通知を含む PDU を受信したとき，p は E_k で確認されたという．

p の受信通知を含む PDU は，p を**前確認する**とする．E_k で p が前確認されていても，E_j からの p を前確認する PDU が紛失した場合に，他の E_h は，p が E_j で受理されていることがわからない．よって，分散型制御下では，各エンティティが「全宛先が p を前確認している」ことを知る必要がある．各宛先は，受理した p が確認されたとき，p は群 C 内の全宛先で原子的に受信されたと判断する．

2.4 1C サービスと ST サービス

高速網[5]では，伝送速度がエンティティの処理速度よりも速いことから，バッファオーバーランにより，PDU を受信できない場合がある．単一チャネル(1C)サービスとは，各受信ログが順序保存で，かつ互いに順序同値であるサービスである．1C サービスは高速網をモデル化したものである．1C サービスを利用した場合，各エンティティは PDU を同一順序で受信できる．しかし，PDU が紛失する場合があり，各受信ログは情報保存とは限らない．

次に，選択的放送型通信サービスを考える．

[定義] **選択的順序保存放送型通信** (SP) サービス[9]とは，各受信ログが選択的に正しいサービスである．**選択的全順序放送型通信** (ST) サービスとは，各受信ログが選択的に正しく，かつ互いに順序同値であるサービスである．□

[例 1] 群 $C=\langle E_1, E_2, E_3 \rangle$ を考える．E_1 は PDU a, b, c を，E_2 は p と q を，E_3 は x, y, z を送信するとする．

(1) 1C サービス (**図 2**) では，各 E_i は同一の順序で PDU を受信するが，ある PDU の受信に失敗する場合がある．例えば，E_2 は c，E_3 は q の受信に失敗している．

(2) SP サービス (**図 3**) では，各 E_i は，各 E_j から自分宛の PDU を送信順に受信できる．しかし，E_i は同じ順序で PDU を受信できるとは限らない．例えば，E_2 は x より先に a を受信しているが，E_3 は逆の順に受信している．

RL_1: $\langle a\ x\ b\ c\ p\ y\ z\ q \rangle$ SL_1: $\langle a\ b\ c \rangle$
RL_2: $\langle a\ x\ b\ p\ y\ z\ q \rangle$ SL_2: $\langle p\ q \rangle$
RL_3: $\langle a\ x\ b\ c\ p\ y\ z \rangle$ SL_3: $\langle x\ y\ z \rangle$

図 2 1C サービスの例
Fig. 2 Example of 1C service.

RL_1: $< x\ c\ p\ z]$ SL_1: $< a_{\{23\}}\ b_{\{2\}}\ c_{\{13\}}]$
RL_2: $< a\ x\ b\ y\ q]$ SL_2: $< p_{\{1\}}\ q_{\{23\}}]$
RL_3: $< x\ a\ c\ z\ q]$ SL_3: $< x_{\{123\}}\ y_{\{2\}}\ z_{\{13\}}]$

図 3 SP サービスの例
Fig. 3 Example of SP service.

RL_1: $< x\ c\ p\ z]$ SL_1: $< a_{\{23\}}\ b_{\{2\}}\ c_{\{13\}}]$
RL_2: $< a\ x\ b\ y\ q]$ SL_2: $< p_{\{1\}}\ q_{\{23\}}]$
RL_3: $< a\ x\ c\ z\ q]$ SL_3: $< x_{\{123\}}\ y_{\{2\}}\ z_{\{13\}}]$

図 4 ST サービスの例
Fig. 4 Example of ST service.

(3) ST サービス (**図 4**) では，各 E_i は，各 E_j から自分宛の PDU を送信順に受信でき，かつ，受信順序は同一である．例えば，E_1 と E_3 からおのおの送信された $a_{(23)}$ と $x_{(123)}$ は，共通宛先 E_2 と E_3 を持つ．E_2 も E_3 も同じ順序，すなわち，a は x に先行して受信される．□

3. 基本データ転送手続き

本章では，1C サービスを利用した選択的全順序放送型通信 (ST) プロトコルのデータ転送手続きについて述べる．以下，記号 p, q, r は PDU を示す．

3.1 変　数

記法 p^i は，E_i が p を送信したことを強調するときに用いる．$p.F$ は p 内の項目 F を示す．$p.SRC$ は，p の送信元 E_i (すなわち，p^i) である．$p.DST$ は，p の宛先エンティティの集合である．p は，主通番 $p.TSEQ$ と副通番 $p.PSEQ_j$ の二種類の通番を持つ．E_i が送信した任意の異なった p と q について，$p \rightarrow_{SL_i} q$ ならば，$p.TSEQ < q.TSEQ$ である．$p \rightarrow_{SL_i} q$, $E_j \in p.DST \cap q.DST$ のとき，$p.PSEQ_j < q.PSEQ_j$ である．$p \rightarrow_{SL_i} r \rightarrow_{SL_i} q$ で，$E_j \in r.DST \cap q.DST$ なる E_j が存在しないとき，$r.PSEQ_j = p.PSEQ_j$ である．すなわち，E_i は p を送信するごとに，$TSEQ$ を一つ増加させるが，p が E_j 宛であるときだけ，$PSEQ_j$ を一つ増加させる．$p.ACK_j$ は，E_i が E_j から次に受信予定の PDU の主通番である ($j=1,\cdots,n$)．これは，E_i が，$q.TSEQ < p.ACK_j$ である q を，E_j から既に受信していることを示す．$p.BUF$ は，E_i が利用可能なバッファ数を示す．

各 E_i は，以下の変数を持つ ($j,k=1,\cdots,n$)．

- $TSEQ$＝次に送信予定の PDU の主通番．
- $PSEQ_j$＝次に E_j 宛に送信予定の PDU の副通番．
- $TREQ_j$＝E_j から次に受信予定の PDU の主通番．
- $PREQ_j$＝E_j から次に受信予定の PDU の副通番．
- AL_{kj}＝「E_j が，E_k から次に受信予定である」と E_i が認識している PDU の主通番．
- $minAL_j = AL_{j1},\cdots,AL_{jn}$ の最小値．
- PAL_{kj}＝「E_j が E_k からの PDU で，次に前確認予定である」と E_i が認識している PDU の通番．
- BUF_j＝E_j 内の利用可能なバッファ数．

AL_{kj} は，p^j の確認通知 $p.ACK_k$ を記録する．これにより，E_i は，「$q.TSEQ < AL_{kj}$ である q^k を E_j が受理している」ことがわかる．$minAL_j$ は，群内の全エンティティが，$g.TSEQ < minAL_j$ である g^j を既に受理していることを示す．群確立手続き[13),14)]により，$TSEQ_j$ と BUF_j のそれぞれの初期値 ISS_j と IBF_j についての合意がとられている．群確立時に，各 E_i で，$TSEQ = PSEQ_j = ISS_j$, $TREQ_j = PREQ_j = AL_{jk} = ISS_j$, $BUF_j = IBF_j$ である ($j,k=1,\cdots,n$)．

各 E_i の受信ログ RL_i は，三つの副ログ ARL_i, PRL_i, RRL_i から成る．それぞれ，確認，前確認，受理された PDU が格納される．

3.2 送 受 信

各 E_i は，上位層からデータ送信要求 R を受けたとき，以下のフロー条件を充足するならば，送信手続きに従って，R に対する PDU p を送信する．ここで，W は最大ウィンドウ幅，H (≥ 1) は定数である．一つの PDU が確認されるまでに，$O(n)$ 個の PDU の送受信が必要である[9)]ので，E_i は，$O(n)$ 個の PDU を記憶できるだけのバッファを持たねばならない．

[フロー条件] $minAL_i \leq TSEQ < minAL_i + min(W, minBUF/(H*n))$. □

[送信手続き] (E_i による p の送信)

(1) $p.TSEQ := TSEQ$, $TSEQ := TSEQ+1$.
(2) $p.PSEQ_j := PSEQ_j$ ($j=1,\cdots,n$).
(3) 各 E_j について，E_j が p の宛先であるならば，$PSEQ_j := PSEQ_j+1$ とし，$p.DST := p.DST \cup \{E_j\}$ とする．
(4) $p.ACK_j := TREQ_j$ ($j=1,\cdots,n$), $p.BUF := BUF_i$.
(5) p を SL_i に追加し，p を放送する．□

放送型通信網を用いているので，各 E_j から送信された p は全エンティティで受信される．ここで，E_i が p を受信したとする．p は以下の受理手続きに従って，受理される．E_i は p が E_i 宛でなくても通信の制御情報は獲得する．p が E_i 宛のときのみ受理し，受信ログに追加する．

[受理条件]

(1) $p.TSEQ = TREQ_j$ または,

(2) $E_i \in p.DST$ かつ $p.PSEQ_i = PREQ_i$. □

[受理手続き]

(1) $TREQ_j := p.TSEQ_j + 1$.

(2) $AL_{kj} := p.ACK_k$ $(k=1,\cdots,n)$.

(3) $E_i \in p.DST$ ならば, $PREQ_j := p.PSEQ_i + 1$ とし, p を RRL_i の最後尾に追加する. そうでなければ, p を破棄する. □

E_i は, 各 E_j から自分宛の PDU を送信順に受信する. E_j から p を受理するごとに, AL_{kj} が, $p.ACK_k$ に変更される $(k=1,\cdots,n)$.

3.3 前確認

$AL_j(p^j)$ を $\{AL_{jk} | E_k \in p.DST\}$, $min\ AL_j(p)$ を $AL_j(p)$ 内の最小値とする. これは, p^j の各宛先が, $g.TSEQ < min\ AL_j(p)$ であるすべての g^j を, 受理していることを示す. よって, 受理した p が以下に示す前確認条件を充足するならば, E_i は p を前確認する.

[前確認条件] $p^j.TSEQ < min\ AL_j(p^j)$. □

[前確認手続き] $p^j = top(RRL_i)$ が前確認条件を充足する間, p を RRL_i から PRL_i に移動し, $PAL_{kj} := p.ACK_k$ $(k=1,\cdots,n)$. □

前確認条件は, AL 内のある値が受理動作によって変更されたときに調べられる. また, p の送信元 E_j は, 群内の全エンティティから p の受信通知を含む PDU を受信したとき, すなわち, $p.TSEQ < min\ AL_j$ ならば, p を SL_j から削除する.

3.4 確認

1C サービスでは, p を前確認する PDU の受信順序は, 各 E_i で同一である. ここで, q が受信されたときに, p が前確認されたとする. p を前確認する各 E_j からの PDU は, q 以前に受信されている. 従って, q が前確認されれば, p を前確認している全 PDU が前確認されている. 以上から, PRL_i 内の PDU は, 以下の確認手続きにより確認される.

[確認条件] $p^j.TSEQ < min\ PAL_j(p^j)$. □

[確認手続き] $p^j = top(PRL_i)$ が確認条件を充足する間, p を PRL_i から取り出し, ARL_i に追加する. □

4. 障害復旧

1C サービスでは, PDU が紛失する場合がある. 紛失復旧では, まず, 紛失の検出方法が問題となる. 次に, 各エンティティで, 紛失 PDU についての合意をとる必要がある. この後に紛失 PDU の再送と, 受信した PDU の選択的全順序付けの方法が必要となる. 以下, p, q, g は PDU を示すとする.

4.1 障害検出

E_k が g^j を紛失したとき, g が E_k 宛の場合だけ, 復旧を行う必要がある. したがって, g の紛失とともに, g が E_k 宛かどうかを確認する必要がある. 以下の障害条件によって, g の紛失を検出できる.

[障害条件] (図 5)

(1) p^j を受信したとき, $PREQ_j < p.PSEQ_k$ ならば, $PREQ_j \le g.PSEQ_k < p.PSEQ_k$ で $E_k \in g.DST$ である g^j を紛失している.

(2) q^h を受信したとき, ある j $(\neq h)$ について $TREQ_j < q.ACK_j$ ならば, $TREQ_j \le g.TSEQ < q.ACK_j$ である g^j を紛失している. □

ここで, 以下のように障害点を定義する.

[定義] 受信ログ RL_i に対する索引 f_i に対して, $RL_i|^{f_i-1}$ は選択的に正しいが, $RL_i|^{f_i}$ は選択的に正しくないとき, $RL_i[f_i]$ を直接的障害点とする. □

すなわち, E_i は, f_i 番目に受信すべき PDU の受信に失敗している.

ここで, E_i は, g^j の受信に失敗しているとする $(g = RL_i[f_i])$. このとき, g は再送されるが, g を RL_i 内のどの位置に記憶するかが問題となる. RL_i 内の g の位置により, 他の受信ログ間での選択的全順序性を保てなくなる場合がある. 例えば, 図 4 において, E_3 が c を紛失した場合を考える. E_3 が, 再送された c を受信ログの最後尾に追加したとする. E_3 では $z \rightarrow_{RL_3} c$ となるが, E_1 では $c \rightarrow_{RL_1} z$ であるため, 選択的全順序性が成り立たない. このために, 各受信ログ内のどの PDU までが, 互いに選択的に正しいかを示す障害線を以下に定義する.

[定義] $RL_1,\cdots,RL_n$ の障害線は, 以下の手順により得られる PDU 索引の組 $\langle f_1,\cdots,f_n\rangle$ である.

(1) f_i を RL_i の最後尾の PDU の索引とする $(last(RL_i) = RL_i[f_i])$ $(i=1,\cdots,n)$.

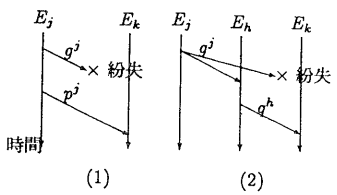


図 5 障害の検出
Fig. 5 Detection of PDUs lost.

(2) 各 E_h について，直接的障害点 g_h を持ち，$g_h < f_h$ ならば，(3)～(6)を行う．

(3) $f_h := g_h$. p を E_h が受信に失敗した PDU とする．

(4) 各 E_i について，RL_i が p を含めば，p の PDU 索引を g_i とする．$g_i < f_i$ ならば $f_i := g_i$ とする．E_i をマークする．

(5) 非マークのエンティティがあれば，RL_h 内で索引が g_h の PDU を p とする ($p = RL_h[g_h]$)．全エンティティがマークされていれば，マークをはずして，(2)へ．

(6) E_j がマークされていないとき，RL_j が p ($= RL_j[l_j]$) を含むとする．$g_j := l_j + 1$. $g_j < f_j$ ならば，$f_j := g_j$. E_j をマークする．(5)へ．□

[例 2] 図4で，E_3 が c を紛失した場合の例を**図 6** に示す．$\|p$ は，p が障害点であることを示す．ここで，E_3 は，E_2 が送信した q によって，c の紛失を検出したとする．□

4.2 障害線合意手順

E_k が，E_h からの PDU で，$p.PSEQ_h \geq PREQ_k$ である p を紛失している．まず，以下により，全エンティティで，障害線について合意をとる．

[障害線合意手順]

(1) E_k は，障害が発生したことを全エンティティに通知する．このために $r.A_h := PREQ_h, r.D_h := PSEQ_h$ ($h = 1, \cdots, n$) なる RST(reset) PDU r を放送する．

(2) r を受信した各 E_j は，データ転送を停止する．$s.A_h := PREQ_h$, $s.D_h := PSEQ_h$ ($h = 1, \cdots, n$) なる RST_PAK PDU s を放送する．このとき，RST_PAK の受信順序を変数 $ORDR$ に記憶する．

(3) 各 E_k は，RST_PAK s を全エンティティから受信後，障害点を確定する．すなわち，RL_k 内で，$p.PSEQ_h < min\{s^i.A_h \mid i = 1, \cdots, n\}$, $q.PSEQ_h \geq min\{s^i.A_h \mid i = 1, \cdots, n\}$, かつ $p \rightarrow_{RL_k} q$ である q の直前の p を障害点とする．$t.A_h := PREQ_h$, $t.D_h := PSEQ_h$, $t.ORDR := ORDR$ ($h = 1, \cdots, n$) なる RST_ACK t を放送する．

RL_1: $< x_{\{123\}} \| c_{\{13\}}\ p_{\{1\}}\ z_{\{13\}} >$
RL_2: $< a_{\{23\}}\ x_{\{123\}}\ b_{\{2\}}\ y_{\{2\}} \| q_{\{23\}} >$
RL_3: $< a_{\{23\}}\ x_{\{123\}} \| z_{\{13\}}\ q_{\{23\}} >$.

図 6 障害点の例
Fig. 6 Example of failure point.

(4) 全 E_j から RST_ACK t を受信したならば，E_k は PRL_k 内と RRL_k 内で障害点以前に受信した PDU を順々に ARL に移す．各 E_h は，$PSEQ_i := t^i.A_h$, $PREQ_i := t^i.D_h$ とする ($i = 1, \cdots, n$)．以下に述べる再送手続きにより，紛失した PDU の再送を行う．□

各 E_h は，E_k からの障害通知 RST を受信すると，障害点を確定するための RST_PAK を放送する．これらにより，E_h は障害点 f_k を確定する．障害点に先行して受信した PDU 系列は，選択的に正しく，確認がとれているので，ARL に移す．

4.3 再送手順

PDU の紛失に対する復旧方法として，後退復旧と選択的再送がある．前者では，障害線以降の PDU を破棄し，再送を行う．後者では，紛失した PDU のみが再送される．前者は後者よりも再送 PDU 数が多くなる．このために，本論文では選択的再送を用いる．

選択的再送では，再送 PDU p を受信ログのどの位置に追加するかが問題となる．一つは，p の追加点についての合意をとり，その位置に p を追加する方法である．この方法は，複数の PDU が紛失した場合，それぞれの追加点についての合意をとる必要があるため，手続きが複雑なものとなる．このため，本論文では，再送 PDU をログに追加し，障害点以降の PDU を，一定の規則に従って整列する方法を用いる．本方式では，紛失 PDU が複数あっても，紛失 PDU の送信元は，他のエンティティと独立に再送を行える．

[再送手続き]

(1) 各 E_j は，SL_j で $p.PSEQ_i > PSEQ_i$ かつ $E_i \in p.DST$ である p を再送する ($i = 1, \cdots, n$).

(2) 各 E_j は，再送された p^i を受信したとき，$PREQ_i < p.PSEQ_j$ ならば，p を受信していないので，p を RRL の最後尾に追加する．E_j は，一定時間内に自分が紛失した全 PDU を受信しなければ，再送を要求する．□

全紛失 PDU を受信した E_j の受信ログは，情報保存性を満たしている．しかし，障害点以降の PDU の順序は全エンティティで同じとは限らない．すなわち，E_j は，p を紛失しているならば，再送された p を受理するが，既に受理しているときは，p を無視する．このために，順序保存性と順序同値性の性質を満たすために，各 E_i の RRL_i 内にある PDU の順序を一定とする必要がある．

各 RRL_j の障害点を f_j とする．このとき，

$RRL_j|_{f_j}$ を**整列対象連** S_j とする．S_j は，障害点以降に受信された PDU の系列である．各 S_j が，選択的情報保存であることは明らかである．ここで，E_i と E_k からの RST_PAK s^i と s^k の受信順序が $s^i \rightarrow_{RL_j} s^k$ ならば，$E_i < E_k$ と順序付ける．障害線合意手順により，RST_PAK の受信順序は $ORDR$ に記憶されている．1C サービスを用いているので，RST_PAK の到着順序は全エンティティで同じである．従って，各エンティティは同一の順序 $<$ を持つ．E_j は，RST_PAK の受信順序により各エンティティを順序付け，これにより S_j を整列する．

[整列規則]

(1) $p.SRC=E_i$, $q.SRC=E_k$, $E_i<E_k$ ならば，$p \rightarrow_{RL_j} q$.

(2) $p.SRC=q.SRC$, $p.PSEQ_j<q.PSEQ_j$ ならば，$p \rightarrow_{RL_j} q$. □

[例 3] 群 $C=\langle E_1, E_2, E_3\rangle$ で，各 E_i が全紛失 PDU の受信後の PDU の整列例を**図 7** に示す．ここで，$p^k_{(u)}$ は，$p.PSEQ_i=u$ である PDU p^k を示す．各 E_i $(i=1,2,3)$ での RST_PAK s の受信順序を $s^2 \rightarrow_{RL_k} s^1 \rightarrow_{RL_k} s^3$ とする．各 E_i は整列規則(1)に従い，PDU を整列する．例えば，E_1 では $b^1_{(4)}$, $a^1_{(3)}$ を $q^2_{(6)}$ の後に，$x^3_{(5)}$ を $b^1_{(4)}$, $a^1_{(3)}$ の後に置く．同じ送信元からの PDU が複数ある場合，整列規則(2)に従い，$PSEQ_i$ により整列する．□

5. 評　　価

本プロトコルの性能を，PDU 紛失に対して整列される PDU 数と，再送される PDU 数により評価する．群内のエンティティ数を n，PDU の平均宛先数を d $(\leq n)$，各 PDU がある一つのエンティティで受信に失敗する確率を f とする．ここで，各 PDU は任意の宛先に，ランダムに送信されるとする．各 PDU が全宛先で受信される確率 F は，$(1-f)^d$ である．各 PDU が少なくとも一つの宛先で受信に失敗する確率は，$1-F$ である．各 PDU の宛先は均一に n 個のエンティティに分散しているとする．

$$
\begin{array}{lll}
E_1 & RRL_1: & <\ldots \| b^1_{(4)}\ x^3_{(5)}\ a^1_{(3)}\ q^2_{(6)}\ ] \\
E_2 & RRL_2: & <\ldots \| a^1_{(5)}\ q^2_{(4)}\ b^1_{(6)}] \\
E_3 & RRL_3: & <\ldots \| y^3_{(8)}\ x^3_{(7)}\ c^1_{(5)}\ p^2_{(2)}\ ] \\
 & & \Downarrow \\
E_1 & RRL_1: & <\ldots \| q^2_{(6)}\ a^1_{(3)}\ b^1_{(4)}\ x^3_{(5)}\ ] \\
E_2 & RRL_2: & <\ldots \| q^2_{(4)}\ a^1_{(5)}\ b^1_{(6)}] \\
E_3 & RRL_3: & <\ldots \| p^2_{(2)}\ c^1_{(5)}\ x^3_{(7)}\ y^3_{(8)}\ ]
\end{array}
$$

図 7 PDU 整列の例
Fig. 7 Example of PDU sorting.

群内で送信された全 PDU で，前確認されていない受信列 L を考える．L の長さ (PDU 数) は $O(n)$ である[15]．実際には，各 E_i は，L 内で，自分宛の $(d/n)L$ 個の PDU を受信する．ここで，先頭から l 番目の PDU までは，全宛先で受信されており，$l+1$ 番目の PDU を紛失した場合を考える．この PDU が自分宛である確率は (d/n) である．この場合が起こる確率 P_l は $F^l(1-F)(d/n)$ である．$l+1$ 番目以降の PDU が整列対象となるエンティティの数は，紛失した PDU の宛先数 $(=d)$ である．この整列により，新たに $l+2$ 番目以降の PDU が整列対象となるエンティティ数は $((n-d)/n)d$ である．このように，$l+m$ 番目以降の PDU が整数対象となるエンティティ数は，$((n-d)/n)^{(m-1)}d$ である．以上より，直接的障害点が $l+1$ 番目の PDU であるとき，各エンティティで整列対象となる PDU 数 T_l は次式で与えられる．

$$
T_l=\left\{(L-l)+\left(\frac{n-d}{n}\right)(L-l-1)+\left(\frac{n-d}{n}\right)^2(L-l-2)+\cdots+\left(\frac{n-d}{n}\right)^{(i-1)}(L-l-i)+\cdots+\left(\frac{n-d}{n}\right)^{(L-l-1)}\right\}*(d/n) \qquad (1)
$$

T_l と，$l+1$ 番目の PDU が直接的障害点である確率 P_l の積は，直接的障害点が $l+1$ 番目の PDU であるときの，各エンティティで整列対象となる平均 PDU 数である．この中で，実際に整列されるのは，自分宛の PDU であるので，平均 PDU 数 U_l は次式により与えられる．

$$
U_l=T_l*P_l*(d/n) \qquad (2)
$$

以上より，各エンティティで整列される平均 PDU 数 U は，次式より求まる．

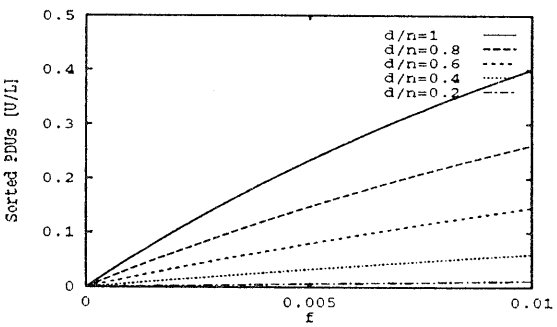


図 8 整列対象 PDU 数
Fig. 8 Number of PDUs sorted.

$$U=\sum_{l=0}^{L-1} U_l \qquad (3)$$

L は $O(n)$ であることから，$L=n$ とする．**図8**に，f について，整列対象 PDU 数 U の L に対する割合 (U/L) を示す．これを**平均整列率**とする．$d/n=1$ は，PDU が群内の全エンティティに送信され，$d/n=0.5$ は，PDU がその半数に送信されることを示す．図8から，宛先数にほぼ比例して，整列される PDU 数が増加することがわかる．

次に，再送 PDU 数について考える．PDU が紛失した場合の再送方式としては，(1)後退復旧と(2)選択的再送がある．(1)では，障害点以降の全 PDU が再送される．これに対して，(2)では，紛失 PDU のみが再送される．(1)では，(3)式で与えられる数の PDU が再送される．選択的再送を考える．障害線が $l+1$ 番目の PDU である場合を考える．この確率は P_l である．このとき，$l+1$ 番目以降の PDU の中で，紛失し再送される PDU の平均数は，$1+(L-l-1)(1-F)$ である．このときの平均再送 PDU 数 R_l は次式で与えられる．

$$R_l=P_l*\{1+(L-l-1)(1-F)\} \qquad (4)$$

平均再送 PDU 数は以下となる．

$$R=\sum_{l=0}^{L-1} R_l \qquad (5)$$

受信ログの長さ L で R を割ったものを，**平均再送率**とする．ここで，$d/n=1$ の場合は，選択的な放送型通信ではなく，全宛先に放送される TO プロトコル[15]に対応する．**図9**に，ST $(d/n=0.5)$ と TO ($d/n=1$ の ST) での平均再送率を示す．また，ST で，後退復旧を行った場合には，整列される PDU が再送されることになり，平均再送率は U/L となる．選択的再送を行った場合，後退復旧を行った場合のそれぞれの場合について示す．

図9から，ST プロトコルで選択的再送を行った場合の再送 PDU 数は，TO での後退復旧の 1/10，ST での後退復旧の 1/2 に相当することがわかる．

6. おわりに

本論文では，高速放送型通信網を利用して，信頼性のある放送型通信サービスを提供する ST プロトコルのデータ転送手続きについて述べた．ST プロトコルにより提供される ST サービスを利用することにより，各エンティティから送信された PDU は，群内の全エンティティに選択的全順序に届けられる．ST プロトコルは，分散型制御と群の概念に基づいている．ST プロトコルを利用することにより，グループウェア等での協調動作を容易に実現し，効率的に実行できる．

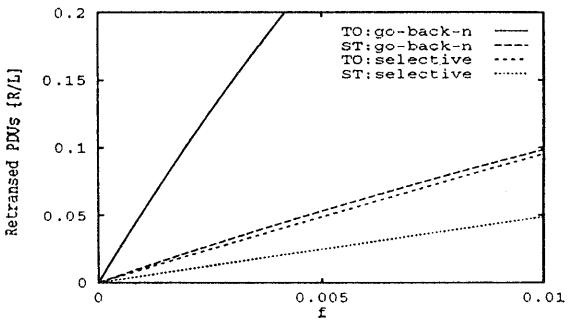


図9 再送 PDU 数
Fig. 9 Number of PDUs retransmitted.

参考文献

1) Florin, G. and Toinard, C.: A New Way to Design Causally and Totally Ordered Multicast Protocols, *ACM Operating Systems Review*, Vol. 26, No. 4, pp. 77-83 (1992).
2) Garcia-Molina, H. and Spauster, A.: Message Ordering in a Multicast Environment, *Proc. of IEEE ICDCS-9*, pp. 345-361 (1989).
3) Garcia-Molina, H.: Ordered and Reliable Multicast Communication, *ACM Trans. Computer Systems*, Vol. 9, No. 3, pp. 242-271 (1991).
4) Melliar-Smith, P. M., Moser, L. E. and Agrawala, V.: Broadcast Protocols for Distributed Systems, *IEEE Trans. Parallel and Distributed Systems*, Vol. 1, No. 1, pp. 17-25 (1990).
5) Abeysundara, B. W. and Kamal, A. E.: High-Speed Local Area Networks and Their Performance: A Survey, *ACM Computing Surveys*, Vol. 23, No. 2, pp. 221-246 (1991).
6) American National Standards Institute: FDDI Token Ring Media Access Control (MAC), ANSI X3.139 (1987).
7) 松下 温 (編著)：図解グループウェア入門，オーム社 (1993).
8) 滝沢 誠，中村章人(訳)：OSI プロトコル技術解説，ソフト・リサーチ・センター (1993).
9) Nakamura, A. and Takizawa, M.: Reliable Broadcast Protocol for Selectively Partially Ordering PDUs (SPO Protocol), *Proc. of IEEE ICDCS-11*, pp. 239-246 (1991).
10) Nakamura, A. and Takizawa, M.: Design of Reliable Broadcast Protocol for Selectively

Partially Ordering PDUs, *Proc. of IEEE COMPSAC 91*, pp. 673-679 (1991).
11) 中村章人，滝沢 誠：多チャネル上の選択的放送通信プロトコルのデータ転送手続き，情報処理学会論文誌，Vol. 33, No. 2, pp. 223-233 (1992).
12) Nakamura, A. and Takizawa, M.: Priority-Based Total and Semi-Total Ordering Broadcast Protocols, *Proc. of IEEE ICDCS-12*, pp. 178-185 (1992).
13) Takizawa, M.: Cluster Control Protocol for Highly Reliable Broadcast Communication, *Proc. of the IFIP Conf. on Distributed Processing*, pp. 431-445 (1987).
14) Takizawa, M.: Design of Highly Reliable Broadcast Communication Protocol, *Proc. of IEEE COMPSAC 87*, pp. 731-740 (1987).
15) 滝沢 誠，中村章人：1チャネル上の全順序放送通信プロトコルにおけるデータ転送手続き，情報処理学会論文誌，Vol. 31, No. 4, pp. 609-617 (1990).
16) Takizawa, M. and Nakamura, A.: Partially Ordering Broadcast (PO) Protocol, *Proc. of IEEE INFOCOM '90*, pp. 357-364 (1990).
17) International Standards Organization: OSI—Connection Oriented Transport Protocol Specification, ISO 8073 (1986).
18) 滝沢 誠，中村章人：放送型通信アルゴリズム，情報処理，Vol. 34, No. 11, pp. 1341-1349 (1993).
19) Birman, K., Schiper, A. and Stephenson, P.: Lightweight Causal and Atomic Group Multicast, *ACM TOCS*, Vol. 9, No. 3, pp. 272-314 (1991).
20) Verissimo, P., Rodrigues, L., Baptista, M.: AMp: A Highly Parallel Atomic Multicast Protocol, *ACM SIGCOMM '89*, pp. 83-93 (1989).
21) Kaashoek, M.F. and Tanenbaum, A.S.: Group Communication in the Amoeba Distributed Operating System, *Proc. of IEEE ICDCS-11*, pp. 222-230 (1991).

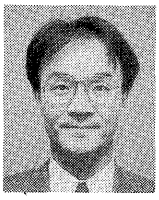
(平成5年11月18日受付)
(平成6年7月14日採録)

立川 敬行（学生会員）

1971年生．1994年東京電機大学理工学部経営工学科卒業．現在，同大学院理工学研究科修士課程在学中．分散型システム，通信網，プロトコル等に興味を持つ．

中村 章人（正会員）

1966年生．1989年東京電機大学理工学部経営工学科卒業．1991年同大学院理工学研究科修士課程修了．1994年同大学院理工学研究科博士課程修了．現在，通商産業省工業技術院電子技術総合研究所研究員．工学博士．「OSIプロトコル技術解説」ソフトリサーチセンタ(共訳)．分散型システム，通信プロトコル等に興味を持つ．



滝沢 誠（正会員）

1950年生．1973年東北大学工学部応用物理学科卒業．1975年同大学院工学研究科応用物理学専攻修士課程修了．同年(財)日本情報処理開発協会入社．1986年東京電機大学理工学部経営工学科講師，1987年より同助教授，工学博士．1989年9月より1年間ドイツ国立情報処理研究所（GMD）客員教授．1990年7月よりKeele大学（英国）客員教授．分散型データベースシステム，通信網，分散型システム，知識ベースシステム等の研究に従事．電子情報通信学会，人工知能学会，ACM，IEEE各会員．「知識工学基礎論」オーム社（共著），「データベースシステム入門技術解説」ソフトリサーチセンタ，「分散システム入門」近代化学社（共著），「OSIプロトコル技術解説」ソフトリサーチセンタ（共訳）．