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†

中村章人†† 滝沢誠††

本論文では，低信頼な放送通信サービスを利用して，分散型システム上の複数の実体に対して，高信頼な放送通信サービスを提供する方法について論じる．複数の実体から構成される分散型システムでは，協調動作を行うために，複数の実体間での通信が必要となる．分散型の応用において，各実体は，あるデータを必ずしもシステム内の全実体に送信する必要はなく，むしろその一部に送信する場合がある．また，各実体は，送信された PDU (protocol data unit) の中で，自分宛のものを，送信された順序で正しく受信する必要がある．グループ内のこのような放送通信サービスを，選択的半順序放送通信 (SPO) サービス (service for Selectively Partially Ordering PDUs) として，本論文で新たに定義する．本論文では，PDU の紛失が発生する複数のチャネルから構成される放送通信網を利用し，SPO サービスを提供するためのプロトコルを示し，その設計について論じる．これまでの放送通信プロトコルは，主制御実体の存在する集中型のものであったが，本論文では，完全分散型の制御に基づいたグループ内の選択的放送通信プロトコルを示す．

1. はじめに

分散型データベースシステム，グループウェアシステム[2]，分散型人工知能システム[5]等の分散型応用システムを実現するためには，複数実体間での協調動作が必要となる．協調動作を行う複数の実体 (OSI のプロトコル・エンティティ) の集合を群[21),22)]とする．これらの応用では，複数の実体同士は，下位層の通信システムを利用してデータの送受信を行う．特に，各実体では，群内の複数の実体にデータを送信することが必要となる．現在のローカルエリア網 (LAN) の MAC 層では，ビット誤り率の低い (一般に 10^{-8}〜10^{-11}) 放送通信サービスが提供されている[6]．しかし，上位層の実体では，バッファの不足や，LAN の高速高帯域化によるバッファ・オーバラン[7]等により，必ずしもすべての PDU を受信できるとは限らない．複数の実体間での，信頼性のある放送通信サービスを実現することにより，これらの応用の実現が容易かつ効率的に行える．

高信頼放送通信プロトコルについて，文献 1)，3)，4)，8)，11)，12)，14)〜25)等で論じられている．これらのプロトコルでは，各 PDU はあるグループ内の全実体に送信される．分散型応用では，各実体は PDU を群内の全実体に対してだけではなく，その一部に送信できればよい．また，各実体は自分宛の PDU のみを送信順に受信する必要がある．本論文では，このような放送通信サービスを選択的半順序放送通信 (SPO) サービスとする．選択的放送通信については，一対一通信に基づいた，PDU の経路制御にスパニング木を用いる方法[27]が議論されている．また，プロトコルを設計する上で，複数の実体での PDU の正しい受信を，どの実体が判断するかという問題がある．これまでに設計された多くの放送通信プロトコルは，ある一つの実体 (指揮者) が，複数の実体 (関係者) での PDU の正しい受信の判断を行うものである．このような制御方式を，集中型制御方式という．集中型制御方式では，各関係者は指揮者の判断を待たねばならず，指揮者の障害が全体障害を引き起こす場合がある．一方，分散型制御方式では，指揮者と関係者の区別がなく，各実体が受信した PDU に基づいて正しい受信の判断を行う．本論文では，SPO サービスを新たに定義し，PDU の紛失が発生する低信頼な放送通信サービスを用いて，分散型制御により，SPO サービスを提供するためのプロトコルを示し，その設計について述べる．

本論文の構成は以下のようである．2章では，SPO サービスを定義する．3章では，SPO プロトコルのデータ転送手続きについて述べる．最後に，4章で SPO プロトコルの正しさと性能について論じる．

2. 選択的放送通信 (SPO) サービス

本章では，本論文で示す，複数の実体に対する SPO サービスとは何かを考える．

† Data Transmission Procedure of Selective Broadcast Protocol on Multi-Channel by AKIHITO NAKAMURA and MAKOTO TAKIZAWA (Department of Information and Systems Engineering, Faculty of Science and Engineering, Tokyo Denki University).

†† 東京電機大学理工学部経営工学科

2.1 正しい受信

通信システム Mは，m個の実体$\{D_1, \cdots, D_m\}$から構成される．各実体は，PDU の送受信によって互いに通信を行う．$s_i[p]$ と $r_i[q]$ を，D_i による PDU p の送信事象と q の受信事象とする．ここで，M内の事象集合上に二つの半順序関係$\rightarrow_i$ と$\rightarrow$[9] を定義する．D_i 内の二つの事象 e_1 と e_2 について，e_1 が e_2 以前に生起するならば，$e_1 \rightarrow_i e_2$ である．(1)ある D_i について $e_1 \rightarrow_i e_2$，または，(2)ある実体 D_i と D_j について，$e_1 = s_i[p]$ で $e_2 = r_j[p]$ である pが存在するとき，$e_1 \rightarrow e_2$ である．ここで，$e_1 \rightarrow^* e_2$ は，$e_1 \rightarrow e_2$ または $e_1 \rightarrow e_3 \rightarrow^* e_2$ なる事象 e_3 が存在するとき成り立つとする ($\rightarrow^*$ は$\rightarrow$の推移的閉包)．$\rightarrow_i^*$ も同様である．

群 $C(\subseteq M)$ を n 個の実体集合 $E_1, \cdots, E_n$ とする．各 PDU p について，$p.DST(\subseteq C)$ を p の宛先実体集合とする．ここで，各 E_i が送信する PDU には，E_i がそれまでに受信した PDU に対する確認通知が含まれる．E_i が送信した pは，以下の三相を経て E_k で正しく受信されたと判断される ($k=1, \cdots, n$)．

第1相 (受理)：E_k が E_i から受信した各 q について，$s_i[q] \rightarrow^* s_i[p]$ であるならば $r_k[q] \rightarrow^* r_k[p]$ であるとき，pは E_k で受理されるとする．

第2相 (前確認)：$p.DST$ 内の各 E_j について，$s_i[p] \rightarrow r_j[p] \rightarrow^* r_k[q]$ である q が存在するならば，pは E_k で前確認されたという ($s_i[p] \Rightarrow_P r_k[q]$ と書く)．

第3相 (確認)：$p.DST$ 内の各 E_j について，$s_j[p] \Rightarrow_P r_j[q] \rightarrow^* r_k[q]$ である q と qが存在するならば，pは E_k で確認されたという．

pが E_k で前確認されたとき，E_k は「$p.DST$ 内の各実体でpが受理された」ことが分かったことになる．pが E_k で前確認されても，E_k は「pが宛先内の各実体で正しく受信された」と判断できない．なぜならば，ある E_j でpが前確認されていない場合があるからである．つまり，E_j がpに対する確認通知を，ある E_h から受信しなければ，E_h はpを受信していないと判断するからである．pが E_k で確認されたとき，E_k は p が宛先内の各実体で正しく受信されたと判断する．つまり，E_k は「pが宛先内の各実体で前確認された」ことが分かったことになる．

2.2 信頼放送通信

各実体が利用するサービスを，PDU の系列であるログの集合[23),25)]としてモデル化する．各実体 E_i は，送信ログ SL_i と受信ログ RL_i を持つ ($i=1, \cdots, n$)．SL_i と RL_i は，各々 E_i が送信および受信した PDU の系列である．E_i が送信した p と q について，$s_i[p] \rightarrow^* s_i[q]$ ならば，SL_i 内で $p \rightarrow_{SL_i} q$ である．同様に，E_i が受信した PDU p と q について，$r_i[p] \rightarrow^* r_i[q]$ ならば，RL_i 内で $p \rightarrow_{RL_i} q$ である．ここで，m個の PDU から成るログ Lを $<p_1 \cdots p_m]$ と書くことにする．p_1 と p_m は，各々 E_i が最初と最後に送信または受信した PDU を示し，各々 $top(L)$, $last(L)$ と書く．各実体 E_i と E_j について，副送信ログ SL_{ij} を，E_i が E_j から受信した PDU から成る SL_i の部分系列とする．また，副受信ログ RL_{ij} を E_i が E_j から受信した PDU から成る RL_i の部分系列とする．

最初に，受信ログ間の関係を定義する．任意の二つの受信ログ RL_i と RL_j について，以下の同値関係が存在する．

- L1. 順序同値：RL_i と RL_j の両方に含まれる任意の二つの PDU p と qについて，$p \rightarrow_{RL_i} q$ ならば $p \rightarrow_{RL_j} q$である．
- L2. 情報同値：RL_i 内の PDU 集合と RL_j 内の PDU 集合が同一である．
- L3. 同値：RL_i と RL_j が順序同値でかつ情報同値である．

RL_i と RL_j が順序同値である場合，E_i と E_j は PDU を同一の順序で受信する．しかし，E_i または E_j で，PDU が紛失している場合もある．RL_i と RL_j が情報同値である場合，E_i と E_j は同一の PDU を受信するが，その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 は同一の PDU を同一の順序で受信する．

次に，送信ログに対する受信ログ RL_i の性質を定義する．

- L4. 順序保存：各 SL_j 内の任意の二つの PDU p と qについて，p と qが RL_i 内に存在し，$p \rightarrow_{SL_j} q$ であるならば，$p \rightarrow_{RL_i} q$である．
- L5. 情報保存：RL_i に含まれる PDU の集合が，$SL_1, \cdots, SL_n$ に含まれる PDU の和である．

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 を送信順に受信している．RL_i が情報保存であるとき，E_i は各実体が送信した全 PDU を受信している．各 PDU が各々の宛先を持つとき，L5 は以下のように一般化できる．

- L5′. 選択的情報保存：RL_i に含まれる PDU の

集合が，$SL_{1i}, \cdots, SL_{ni}$ に含まれる PDU の和である．

RL_i が選択的に情報保存であるならば，E_i は自分宛の全 PDU を受信している．L4，L5，L5′ から，RL_i について以下の性質を定義する．

L6.　正しい：RL_i が順序保存かつ情報保存である．

L6′.　選択的に正しい：RL_i が順序保存かつ選択的情報保存である．

サービス S 内の各受信ログが選択的に正しいならば，S には信頼性があり，S を信頼サービスという．つまり，各実体は自分宛の全 PDU を送信された順序で受信する．信頼性のないサービスを低信頼サービスとする．

2.3 MC サービスと SPO サービス

まず，二つの低信頼な放送通信サービスを定義する．

S1.　1チャネル (1C) サービス：サービス内の各受信ログが順序保存でかつ互いに順序同値である．

S2.　多チャネル (MC) サービス：サービス内の各受信ログが順序保存である．

1C サービスを利用した場合，各実体は PDU を同一の順序で受信できる．MC サービスでは，各実体ごとに PDU を送信順に受信できる．1C と MC では，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．1C サービスは，Ethernet の MAC 副層[6),13)]のサービスをモデル化したものである．MC サービスは，複数の Ethernet で計算機が接続されたシステム（多チャネルシステム）で提供されるサービスを抽象化したものである．

図1に，多チャネルシステムの構成例を示す．各計算機は，R (≧1) 個のチャネル $CH_1, \cdots, CH_R$ で接続されており，各 CH_i は，1C サービスを提供する (一本の Ethernet に対応)．ここで，各実体 E_i は，ある一つのチャネルを利用して PDU の送信を行うとする．$CH_1, \cdots, CH_R$ によって提供されるサービスは，MC サービスである．つまり，各実体からの PDU の到着順序は全実体で同一であるが，PDU が紛失する可能性がある．

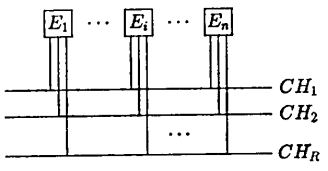

図 1　多チャネルシステム
Fig. 1　Multi-channel system.

選択的放送通信 (SBC) サービスとは，各受信ログが選択的に情報保存であるサービスである．つまり，各実体は自分宛の PDU のみをすべて受信できるサービスである．PDU の受信順序によって，SBC サービスは以下の二種類に分けられる．

S3.　選択的半順序放送通信 (SPO) サービス：サービス内の各受信ログが選択的に正しい．

S4.　選択的全順序放送通信 (STO) サービス：サービス内の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．

SPO と STO は，共に信頼性のある放送通信サービスである．次の章では，低信頼な MC サービスを利用し，SPO サービスを提供するためのプロトコルの設計について述べる．STO サービスの実現は，他の論文で述べる[17)]．

[例 2.1]　MC と SPO サービスの例として，三つの実体 E_1, E_2, E_3 のログを考える．MC サービスを利用した場合，各実体は PDU を送信順に受信できる．しかし，情報保存性が保障されていないので，全 PDU を受信できるとは限らない．**図2**は，MC サービスを利用した場合の，各実体のログを示している．ここで，送信ログ内の各 PDU p について，記法 $p_{i \cdots j}$ は p の宛先が $\{E_i, \cdots, E_j\}$ であることを示すとする．例えば，宛先が E_2 と E_3 である PDU a を a_{23} と記す．MC サービスを利用するので，PDU は全実体に送信され，各実体はこれらを送信順に受信する．例えば，各実体は E_1 が送信した a，b，c，d をこの順序で受信する．また，E_2 と E_3 はそれぞれ c と q を受信していない．**図3**に，SPO サービスを利用して，各実体が図2の場合と同一の PDU を送信した場合のログを示す．各実体は，自分宛の全 PDU を送信順に受信している．例えば，E_1 から受

$E_1\,RL_1$: $<a\ x\ b\ c\ p\ y\ d\ z\ q]$　　SL_1 : $<a_{23}\ b_3\ c_{13}\ d_{12}]$
$E_2\,RL_2$: $<p\ a\ b\ x\ d\ y\ q\ z]$　　SL_2 : $<p_{123}\ q_1]$
$E_3\,RL_3$: $<a\ x\ y\ b\ c\ p\ z\ d]$　　SL_3 : $<x_{12}\ y_{123}\ z_3]$

図 2　多チャネル (MC) サービスの例
Fig. 2　An example of the MC service.

$E_1\,RL_1$: $<x\ c\ p\ y\ d\ q]$　　SL_1 : $<a_{23}\ b_3\ c_{13}\ d_{12}]$
$E_2\,RL_2$: $<p\ a\ x\ d\ y]$　　SL_2 : $<p_{123}\ q_1]$
$E_3\,RL_3$: $<a\ y\ b\ c\ p\ z]$　　SL_3 : $<x_{12}\ y_{123}\ z_3]$

図 3　選択的半順序放送通信 (SPO) サービスの例
Fig. 3　An example of the SPO service.

信した PDU について，各実体でそれぞれ $RL_{11}=<cd]$, $RL_{21}=<ad]$, $RL_{31}=<abc]$ である．□

3. MC サービス上での SPO プロトコルのデータ転送手続き

本章では，MC サービスを利用した SPO プロトコルのデータ転送手続きについて述べる．群Cは，n個の実体 $E_1,\cdots,E_n$ から構成されているとする．各副受信ログ RL_{ij} は，三相受信に基づいて，三つの部分系列 APL_{ij}, PPL_{ij}, RPL_{ij} に分割される．それぞれ，E_j から受信した PDU のうち，確認，前確認，受理された PDU を含む．

3.1 変　数

記法 p^i は，pが E_i によって送信されたことを示す．SPO PDU p^i は，以下に示す項目から構成される．

$$p^i=\langle SRC;\ DST;\ TSEQ;\ \langle PSEQ_1\cdots PSEQ_n\rangle;\ \langle ACK_1\cdots ACK_n\rangle;\ BUF;\ DATA\rangle$$

ここで，$p^i.F$ は p^i 内の項目Fを示す．$p^i.SRC$ は，p^i を送信する実体 E_i である．各 $p^i.DST$ は，p^i の宛先を示しており，各実体はこれに基づいて p^i を受理するかどうか判断する．E_j が p^i を受信し，$p^i.DST$ 内に E_j が含まれていないならば，p^i を廃棄する．p^i は，主通番と副通番の二種類の通番を持つ．主通番 $p^i.TSEQ$ は，E_i が送信した全 PDU 系列内の p^i の位置を示している．また，副通番 $p^i.PSEQ_j$ は，E_k が送信した PDU のうち，E_j を宛先に含む PDU 系列内の p^i の位置を示している．$p^i.ACK_j$ は，E_j から受信した PDU の主通番を示し，E_i が既に $q^j.TSEQ<p^i.ACK_j$ である全 PDU q^j を，既に E_j から受信していることを示す確認通知である ($j=1,\cdots,n$)．$p^i.BUF$ により，E_i が利用可能なバッファ数が各実体に通知され，各実体はこれを基にフロー制御を行う．

各実体 E_i は，以下の変数を持つ ($h,j=1,\cdots,n$)．

$TSEQ$ $=E_i$ が，次に送信予定である PDU の主通番．

$PSEQ_j=E_i$ が，次に E_j 宛に送信予定である PDU の副通番．

$TREQ_j=E_i$ が，次に E_j から受信予定である PDU の主通番．

$PREQ_j=E_i$ が，次に E_j から受信予定である PDU の副通番．

AL_{jk} =「E_k が次に E_j から受信予定である」と E_i が認識している PDU の主通番．

PAL_{jk} =「E_k が E_j から受信した PDU のなかで，E_k が次に前確認予定である」と E_i が認識している PDU の主通番．

$F_j=E_i$ が知っている E_j 内の利用可能バッファ数．

$min\,AL_j$ を，$AL_{j1},\cdots,AL_{jn}$ のなかの最小値とする．これは，群内の各実体が，$q^j.TSEQ<min\,AL_j$ である全 PDU q^j を，既に E_j から受信していることを示している．ISS_i と IBF_i をそれぞれ，E_i の主通番と利用可能バッファ数の初期値とする．E_i の初期状態では，$TSEQ=PSEQ_i=ISS_i$, $TREQ_j=PREQ_j=AL_{jk}=ISS_j$ ($j,k=1,\cdots,n$) である．ここで，各実体は群開設手続き[20),21)]によって，各 E_j の ISS_j と IBF_j についての合意が取られているとする．$min\,F$ を，$F_1,\cdots,F_n$ のなかの最小値とする．

3.2 送 受 信

各実体 E_i は，n個の副受信ログ $RL_{i1},\cdots,RL_{in}$ を持ち，各 E_j から受信した PDU を RL_{ij} に記録する ($j=1,\cdots,n$)．E_i は，放送するデータがあり，かつ以下のフロー条件が充足されるならば，p^i を送信動作に従って送信する．

[フロー条件] $min\,AL_i\leqq TSEQ<min\,AL_i+min\,(W,\ min\,F/(H*n^2))$．(ここで，$W$は最大ウィンドウ幅，$H(\geqq 1)$ は定数) □

[送信動作]

(1) $p^i.TSEQ:=TSEQ$, $TSEQ:=TSEQ+1$.

(2) $p^i.PSEQ_j:PSEQ_j$ ($j=1,\cdots,n$).

(3) 各 E_j について，E_j が p^i の宛先であるならば，$PSEQ_j:=PSEQ_j+1$ とし，$p^i.DST:=p^i.DST\cup\{E_j\}$ とする．

(4) $p^i.ACK_j:=TREQ_j$ ($j=1,\cdots,n$).

(5) p^i を SL_i の最後尾に追加し，p_i を送信する．□

E_i が p^j を受信したとき，p^j が以下に示す受理条件を充足するならば，p^j は E_i で受理される．つまり，E_i が PDU pを受信するとは，p が E_i に届くことであり，E_i がpを受理するとは，受信した PDU がある条件を充足することである．

[受理条件]

(1) (1-1) $p^j.TSEQ=TREQ_j$ または

(1-2) $p^j.PSEQ_i=PREQ_j$, かつ

(2) $p^j.ACK_k\leqq TREQ_k$ ($k=1,\cdots,n$). □

E_j が p^j 以前に送信した全 PDU を受理しているな

らば，条件(1-1)は常に真となる．E_i がある p^j を受信できなくても，E_i が p^j の宛先でなければ，E_i にとって p^j の紛失は障害とならない．しかし，E_i が p^j の宛先ならば，E_i は p^j を受理する必要がある．条件(1-2)によって，これを判断できる．p^j が受理条件を充足するならば，E_i は以下の受理動作を行う．

[受理動作]

(1)　$TREQ_j := p^j.TSEQ + 1$.

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

(3)　$E_i \in p^j.DST$ ならば，$PREQ_j := p^j.PSEQ_i + 1$ とし，p^j を RPL_{ij} の最後尾に追加する．そうでなければ，p^j を廃棄する．□

[例 3.1]　三つの実体 E_1, E_2, E_3 から成る群 C におけるデータ転送の例の図を，**図4**に示す．

(1)　各実体は初期状態にあり，まだ PDU を送受信していない．各実体 E_j の主通番の初期値 ISS_j は，それぞれ5，0，3である $(j=1,2,3)$．各変数は群開設手続き[20),21)]によって初期化されている．各実体 E_j は，送信した PDU を記録するための送信ログ SL_j と，各 E_k から受理した PDU を記録するための副受信ログ RL_{jk} を持つ $(j,k=1,2,3)$．最初に，E_1 が $a=\langle E_1;\ \{E_2, E_3\};\ 5;\ \langle 5\,5\,5\rangle;\ \langle 5\,0\,3\rangle;\ buf_a;$

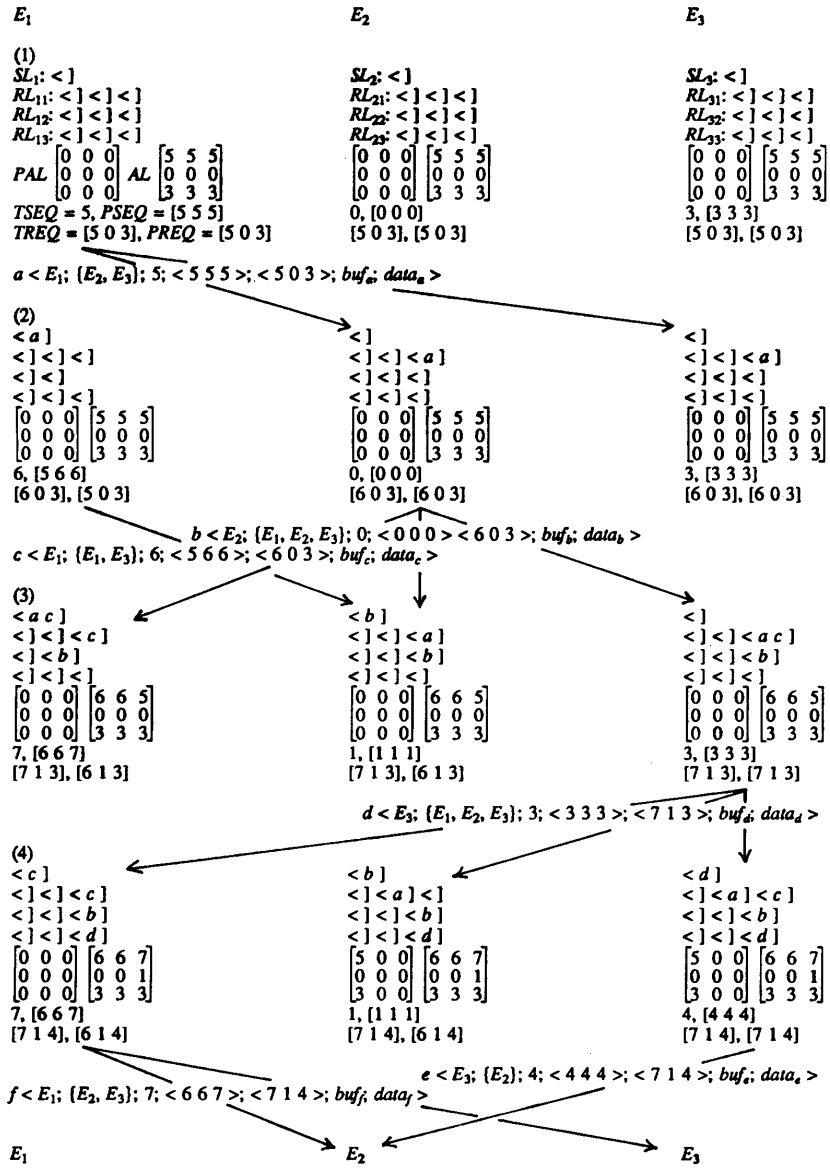

図 4-1　SPO プロトコルのデータ転送手続きの例（1）

Fig. 4-1　An example of the data transmission procedure of the SPO protocol (1).

$data_a$〉を送信する．a の宛先は E_2 と E_3，$TSEQ = PSEQ_j = 5$ $(j=1,2,3)$，$ACK_1 = 5$，$ACK_2 = 0$，$ACK_3 = 3$ である．

（2） E_1 では，a の送信動作により，$TSEQ$ が5から6に，a の宛先が E_2 と E_3 なので $PSEQ_2$ と $PSEQ_3$ がそれぞれ5から6に変更される．つまり，E_1 が次に送信する PDU の主通番（$TSEQ$）は6，各実体に対する副通番（$PSEQ$）はそれぞれ5，6，6となる．各実体で，$a.TSEQ\,(=5) = TREQ_1$ でかつ $a.ACK_j \leqq TREQ_j$ なので，a は受理条件を充足し，受理動作によって $TSEQ_1 := a.TSEQ + 1\ (=6)$，$AL_{j1} := a.ACK_j$ $(j=1,2,3)$ となる．つまり，各実体 E_j は，$TSEQ$ が6，$PSEQ_j$ がそれぞれ5，6，6の PDU を E_1 から受信予定である．E_2 と E_3 は，a をそれぞれ RL_{21} と RL_{31} に追加し，$PREQ_1 := a.PSEQ_k + 1$ $(k=2,3)$ とする．E_1 は，$a.DST$ に

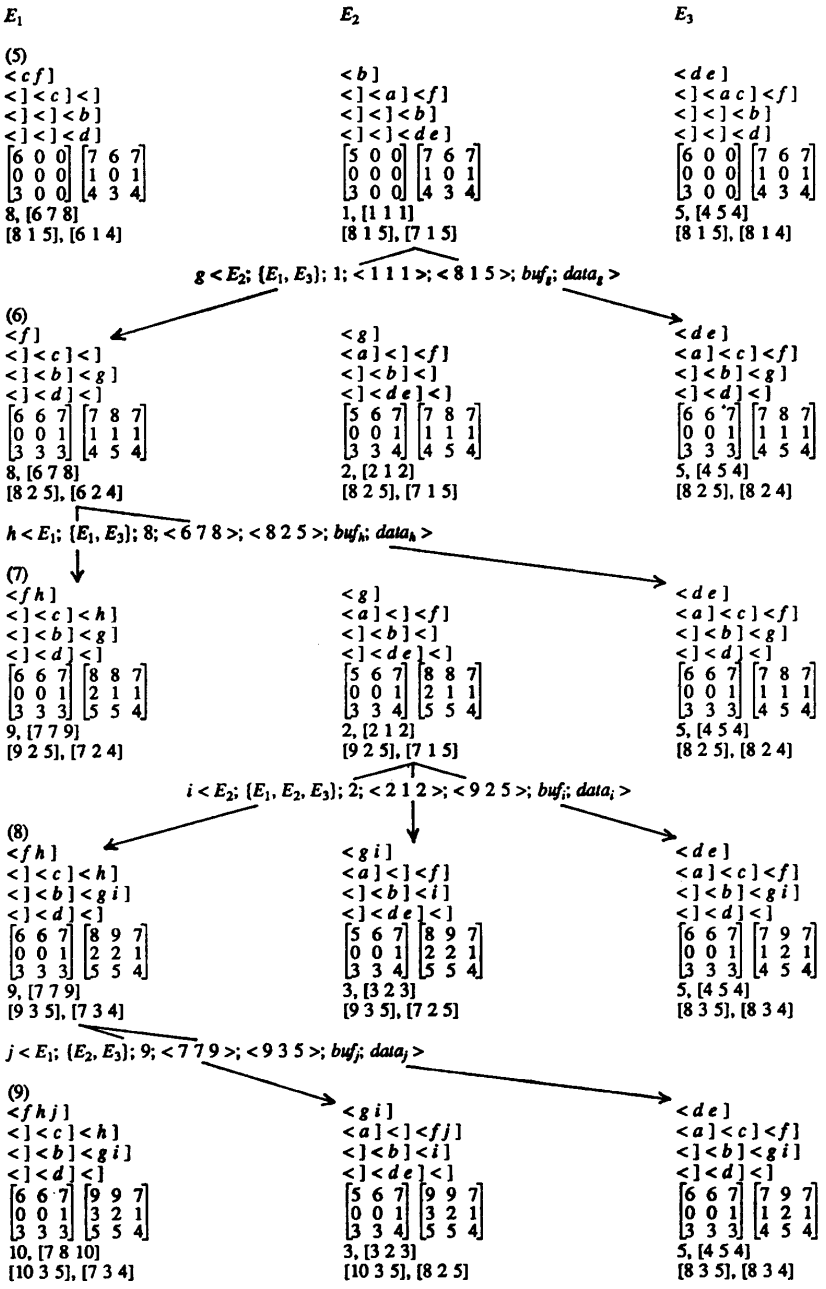

図 4-2 SPO プロトコルのデータ転送手続きの例（2）
Fig. 4-2 An example of the data transmission procedure of the SPO protocol (2).

E_1 が含まれていないので，aを廃棄する．□

3.3 前確認

ここで，$AL_j(p^j)$ を $\{AL_{jk} | E_k \in p^j.DST\}$，$min\ AL_j(p^j)$ を $AL_j(p^j)$ のなかの最小値とする．これは，p^j の宛先の各実体が，$g^j.TSEQ < min\ AL_j(p^j)$ であるすべての g^j を既に E_j から受理していることを意味している．よって E_i は，受理した p^j が以下に示す前確認条件を充足するならば，p^j が各宛先実体で既に受理されていることが分かる．前確認条件を充足する p^j は，以下の前確認動作によって前確認される．

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

[前確認動作] $j=1,\cdots,n$ について，

(1) $p^j = top(RPL_{ij})$ が前確認条件を充足するならば，p^j を RPL_{ij} から PPL_{ij} に移動し，$PAL_{kj} := p^j.ACK_k$ $(k=1,\cdots,n)$ とする．

(2) (1)の操作を，$top(RPL_{ij})$ が前確認条件を充足しなくなるまで繰り返す．□

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る．

[例 3.2] 前確認の例を付図の(2)～(4)に示す．

(2) E_2 は $b=\langle E_2;\ \{E_1, E_2, E_3\};\ 0;\ \langle 000\rangle;\ \langle 603\rangle;\ buf_b;\ data_b\rangle$ を送信し，E_1 は $c=\langle E_1;\ \{E_1, E_3\};\ 6;\ \langle 566\rangle;\ \langle 603\rangle;\ buf_c;\ data_c\rangle$ を送信する．

(3) b と c は各宛先で a と同様に受理され，各変数は受理動作に従って更新される．また，b と c はそれぞれ RL_{j2} $(j=1,2,3)$ と RL_{k1} $(k=1,3)$ に記録される．E_3 は $d=\langle E_3;\ \{E_1, E_2, E_3\};\ 3;\ \langle 333\rangle;\ \langle 713\rangle;\ buf_d;\ data_d\rangle$ を送信する．

(4) 各実体で d が受理された後，$AL_{11}=6$，$AL_{12}=6$，$AL_{13}=7$ となる．ここで，$a.DST=\{E_2, E_3\}$ であり，$a.TSEQ\ (=5) < min\ AL_1(a) = min\{AL_{12}, AL_{13}\}\ (=6)$ であるから，aは前確認条件を充足し，各 E_i で RPL_{i1} から PPL_{i1} に移動される．つまり，各実体は a の宛先である E_2 と E_3 が a を受理したことが分かった．また，E_1 は a が各宛先で受理されていることが分かったので，a を再送する必要がなくなり，a を SL_1 から削除する．このとき，E_2 と E_3 では $a.ACK_k$ が PAL_{k1} に記憶される $(k=1,2,3)$．つまり，$PAL_{11}=a.ACK_1\ (=5)$，$PAL_{21}=a.ACK_2\ (=0)$，$PAL_{31}=a.ACK_3\ (=3)$ となる．□

3.4 確認

E_i が受理した PDU を確認する方法について述べる．ここで，$PAL_j(p^j)$ を $\{PAL_{jk} | E_k \in p^j.DST\}$，$min\ PAL_j(p^j)$ を $PAL_j(p^j)$ のなかの最小値とする．E_i 内で，p^j が以下の確認条件を充足するならば，以下に示す確認動作に従って確認される．

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

[確認動作] $j=1,\cdots,n$ について，

(1) $p^j = top(PPR_{ij})$ が確認条件を充足するならば，p^j を PPL_{ij} から APL_{ij} に移動する．

(2) (1)の操作を，$top(PPR_{ij})$ が確認条件を充足しなくなるまで繰り返す．□

前確認の場合と同様に，確認条件は PAL 内のある値が更新されたときに調べる．

[例 3.3] PDU の確認の例を，図4の(4)～(6)に示す．

(4) E_3 と E_1 は，それぞれ $e=\langle E_3;\ \{E_2\};\ 4;\ \langle 444\rangle;\ \langle 714\rangle;\ buf_e;\ data_e\rangle$ と $f=\langle E_1;\ \{E_2, E_3\};\ 7;\ \langle 667\rangle;\ \langle 714\rangle;\ buf_f;\ data_f\rangle$ を送信する．

(5) $c.TSEQ\ (=6) < min\ AL_1 = min\{AL_{11}, AL_{13}\}\ (=7)$ なので，c は E_1 と E_3 で前確認される．E_2 は $g=\langle E_2;\ \{E_1, E_3\};\ 1;\ \langle 111\rangle;\ \langle 815\rangle;\ buf_g;\ data_g\rangle$ を送信する．

(6) g は E_1 と E_3 で受理される．ここで，$b.TSEQ\ (=0) < min\ AL_2(b) = min\{AL_{21}, AL_{22}, AL_{23}\}\ (=1)$ でかつ $d.TSEQ\ (=3) < min\ AL_3(d) = min\{AL_{21}, AL_{22}, AL_{23}\}\ (=4)$ であるので，b と d は前確認条件を充足し，前確認される．このとき，b と d の各 ACK_j は PAL_{j2} と PAL_{j3} に記憶される $(j=1,2,3)$．また，$min\ PAL_1(a) = min\{PAL_{12}, PAL_{13}\} = min\{6,7\} = 6$ となり，$a.TSEQ\ (=5) < 6$ である．よって，a は確認条件を充足し，PPL_{j1} から APL_{j1} に移動される $(j=2,3)$．□

3.5 障害

各実体は，誤動作[10]しないものとし，ある実体が障害を起こした場合，群は異常終了するものとする．MC サービスを利用した場合，PDU の紛失が発生する．紛失した PDU は，以下の障害条件により検出される．

[障害条件] [図5]

(1) p^j を受信したとき，$PREQ_j < p^j.PSEQ_i$ な

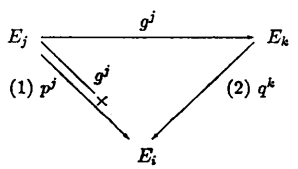

図5 紛失した PDU の検出
Fig. 5 Detection of lost PDU.

らば，E_i は $PREQ_j \leqq g^j.PSEQ_i < p^j.PSEQ_i$ である g^j を受信していない $(j=1,\cdots,n)$.

(2) q^k を受信したとき，ある $j\ (\neq k)$ について，$TREQ_j < q^k.ACK_j$ ならば，E_i は $TREQ_j \leqq g^j.TSEQ < q^k.ACK_j$ である g^j を受信していない $(k=1,\cdots,n)$. □

障害条件により，PDU の紛失が検出されたならば，E_i は以下の復旧動作を行う.

[復旧動作]

(1) 受信した p^j が，障害条件(1)を充足するならば，E_i は後述の再送手続きを行い，g^j を送信した E_j に g^j の再送を要求する.

(2) 受信した q^k が障害条件(2)を充足するならば，E_i は q^k が受理条件を充足するか調べ，充足するならば q^k を受理する．そうでなければ，q^k を廃棄する．E_i は，E_j に対するタイマを設定し始動させる.

(2-1) タイマが時間切れになったとき，E_i は再送手続きを行う.

(2-2) E_i は，E_j から p^j を受信したならばタイマを停止し，p^j が受理条件を充足するか調べる．p^j が受理条件を充足するならば，p^j を受理する．そうでなければ，p^j を廃棄し，再送手続きを行う. □

障害条件(1)が充足されるとき，E_i はある PDU を紛失しており，$PREQ_j \leqq g^j.PSEQ_i < p^j.PSEQ_i$ であるすべての g^j を受理する必要がある．また，障害条件(2)が充足されるとき，E_i はある PDU g^j の紛失を検出するが，g_j が E_i を宛先に含んでいる $(E_i \in g^j.DST)$ かどうかは分からない．E_i は，宛先に E_i を含む g^j のみを受信すれば良い．障害条件(2)が充足されたとき，受信できなかった g^j が E_i を宛先に含んでいないかもしれないので，E_i はある時間だけ E_j から PDU が到着するのを待つ．E_i が，E_j から p^j を受信したとする．このとき，$p^j.PSEQ_i = PREQ_j$ ならば，E_i は g^j を受信する必要がない．$PREQ_j < p^j.PSEQ_i$ ならば，E_i は g^j を受信する必要があるので，再送を要求する．一定時間 E_j から PDU が到着しないときは，再送手続きを行う.

[再送手続き]

(1) E_i は，$rt.ACK_h = TREQ_h\ (h=1,\cdots,n)$ なる *RETRANS* PDU rt を送信する.

(2) E_h は，*RETRANS* rt を E_i から受信したら，$rt.ACK_h \leqq p^h.TSEQ$ かつ $E_i \in p^h.DST$ である p^h を再送する. □

各実体は，PDU を再送する場合があるので，以下の重複処理が必要となる.

[重複処理] $p^j.TSEQ < TREQ_j$ または $p^j.PSEQ_i < PREQ_j$ ならば，p^j を廃棄する. □

[例 3.4] 障害の例を，図4の(6)～(9)に示す.

(6) E_1 は $h=\langle E_1; \{E_1, E_3\}; 8; \langle 6\,7\,8\rangle; \langle 8\,2\,5\rangle; buf_h; data_h\rangle$ を送信する.

(7) E_1 と E_2 は h を受信したが，E_3 は受信していない.

(8) E_3 は，E_2 から $i=\langle E_2; \{E_1, E_2, E_3\}; 2; \langle 2\,1\,2\rangle; \langle 9\,2\,5\rangle; buf_i; data_i\rangle$ を受信したとき，障害点条件(2)によって，$TSEQ=8$ である E_1 が送信した PDU h を紛失していることが分かる．つまり，$TREQ_1\ (=8) < i.ACK_1\ (=9)$ である．ここで E_3 は，紛失した h が E_3 を宛先に含んでいるかどうか分からないので，復旧動作に従い E_1 から PDU が到着するのを待つ.

(9) E_3 は，E_1 から $j=\langle E_1; \{E_2, E_3\}; 9; \langle 7\,7\,9\rangle; \langle 9\,3\,5\rangle; buf_j; data_j\rangle$ を受信し，紛失した h が E_3 を宛先に含んでいるかどうか調べる．$j.PSEQ_3 = PREQ_1\ (=8)$ ならば，h は E_3 を宛先に含んでいない．しかし，$j.PSEQ_3\ (=9) \neq PREQ_1\ (=8)$ なので，h が E_3 を宛先に含んでいることがわかる．E_3 は，再送手続きに従って *RETRANS* を送信し，h の紛失を通知する．E_1 は，$ACK_1=8$ である *RETRANS* を受信したら，$TSEQ \geqq 8$ である h と j を再送する. □

4. 評　価

4.1 正　し　さ

[補題 4.1] E_i が受理した PDU p^j が前確認条件を充足するならば，p^j は E_i で前確認される.

[証明] 前確認条件は，$p^j.DST$ 内の各 E_k について，$p^j.TSEQ < AL_{jk}$ であることを意味している．つまり，$p^j.DST$ 内の各 E_k について，$s_j[p] \rightarrow r_k[p] \rightarrow^* r_i[q]$ となる q が存在する．よって，p^j は E_i で前確認される. □

[補題 4.2] p^j が確認条件を充足するならば，p^j は E_i で確認される.

[証明] q^j を，$q^j.TSEQ = min\ PAL_j(p^j)$ である PDU とする．これは，$p^j.DST$ 内の各 E_k について，E_i が $s_j[q^j] \rightarrow r_k[q^j] \rightarrow^* r_i[g]$ である g を受理していることを意味する．つまり，q^j は前確認される．$p^j.TSEQ < q^j.TSEQ$ なので，$p^j \rightarrow_{SL_j} q^j$ であ

る．よって，p^j は確認される．□

[定理 4.3] SPO プロトコルは，MC サービス上で，群内の各実体に対して SPO サービスを提供する．

[証明] 障害がないときは，補題 4.1 と 4.2 から明らかである．E_i が，ある p^j を紛失したと仮定する．

（1） E_i が p^j を紛失した場合，p^j は $E_k \in p^j.DST$ である各 E_k で前確認されない．なぜならば，E_i は p^j の確認通知を含む PDU を送信しないからである．p^j の紛失は，障害点条件によって検出される．

（2） E_i が p^j を前確認する q^h を紛失した場合，q^h の紛失は障害点条件によって検出される．p^h が E_i で前確認されないので，p^j は $p^j.DST$ 内の各実体で確認されない．□

定理 4.3 から，多チャネルサービスを利用した SPO プロトコルは，上位層に SPO サービスを提供する．

4.2 性　　能

n を群内の実体数，m を各 PDU の宛先実体数の平均値とする（$n \geqq m$）．各実体 E_i について，d_i を PDU の平均送信間隔時間とする．つまり，E_i は平均して，各 d_i 時間ごとに PDU を一つ送信する．t_i を E_i での PDU の平均到着間隔時間とする．Ethernet を下位層のサービスとして利用するので，t_i は一定値 t となり，$t=1/(1/d_1+\cdots+1/d_n)$ である．ここで，各 d_i が一定値 d であるとすると，$t=d/n$ となる．r を平均伝送遅延時間とする．

まず，下位層のサービスが無限の容量を持つ場合，すなわち，伝送路が超高速の場合を考える．つまり，各実体は，待ち時間なしで PDU を送信できる．この場合，受信した各 PDU は，前確認されるまでに平均して $(r+d/2)$ 時間必要となる．この時間内に，各実体は $(r+d/2)/t=(r+d/2)n/d$ 個の PDU を受信する．これは，受信ログ内の PDU 数を示す．伝送路が超高速なので，r は n と独立であり，PPL と RPL の長さは $O(n)$ となる．次に，下位層のサービスの容量が有限であるとすると，r は n に比例し，PPL と RPL の長さはそれぞれ $O(n^2)$ となる．

次に，文献 14), 15), 20)～25) で示した TO プロトコルと PO プロトコルと，本論文で示した SPO プロトコルの性能を比較する．SPO プロトコルは，障害が発生した場合に，その復旧のための処理負荷が，TO プロトコルと PO プロトコルよりも小さい．

表 1 障害時に再送される PDU 数の比較 ($n=16$)
Table 1 Number of PDUs to be retransmitted on failure ($n=16$).

m	N_{TO}	N_{PO}	N_{SPO}
16	1とする	1/16	1/16
8	1	1/16	1/32
4	1	1/16	1/64
2	1	1/16	1/128

PDU の紛失が発生したときに，再送される PDU 数を比較してみる．N_{TO}, N_{PO}, N_{SPO} をそれぞれ，ある PDU p の紛失が検出されたときに，TO, PO, SPO の各プロトコルで再送される PDU 数の期待値とする．TO と PO の場合，各 PDU の宛先は，常に群内のすべての実体である．TO プロトコルでは，受信したすべての PDU は一つの受信ログに記録されるので，紛失が発生した場合，複数の実体から到着する PDU が廃棄された場合がある．PO プロトコルでは，各実体から受信した PDU は，それぞれ異なる受信ログに記録されるので，廃棄される PDU は紛失した PDU の送信元の実体から到着する PDU のみである．よって，$N_{PO}=N_{TO}/n$ となる．TO と PO の場合，紛失した PDU は必ず再送される．しかし，SPO の場合，紛失した PDU の宛先によっては，再送しない場合がある．つまり，E_k がある PDU p を紛失しても，p が E_k を宛先に含んでいなければ，E_k は p を受信する必要がなく，$N_{SPO}=0$ である．p が E_k を宛先に含んでいる場合，$N_{SPO}=N_{TO}/n*m/n$ となる．つまり，$N_{SPO} \leqq N_{PO} \leqq N_{TO}$ である．n を 16，m を 16，8，4，2 としたときの，N_{TO}, N_{PO}, N_{SPO} の例を**表 1** に示す．

5. おわりに

本論文では，MC サービスという低信頼な放送通信サービスを利用して，SPO サービスという高信頼な放送通信サービスを提供するプロトコルの設計について述べた．SPO サービスでは，従来の放送通信サービスと異なり，各 PDU は全実体ではなく，宛先実体にのみ届けられる．SPO プロトコルは，分散型の制御と，群という概念に基づいている．群とは，複数の実体の集合である．SPO プロトコルが提供する SPO サービスでは，受信する PDU の半順序性が保障される．つまり，各実体から受信する PDU の順序は，それらが送信された順序と同一である．また，MC サービス上での SPO プロトコルの正しさについて述べ，

SPO プロトコルの性能を，TO と PO という，各 PDU の宛先が常に全実体であるプロトコルと比較し，検討を行った．SPO プロトコルを利用することにより，分散型データベースシステムにおける，コミットメント制御，分散型デッドロック検出，分散型問合わせ処理等の実現が効率的に行える．現在，SPO プロトコルを利用し，オブジェクト指向分散型システムにおけるトランザクション処理機構[26]の実現を行っている．

参考文献

1) Chang, J.-M. and Maxemchuk, N. F.: Reliable Broadcast Protocols, *ACM Trans. Comput. Syst.*, Vol. 2, No. 3, pp. 251-273 (1984).
2) Ellis, C. A., Gibbs, S. J. and Rein, G. L.: Groupware: Some Issues and Experiences, *Comm. ACM*, Vol. 34, No. 1, pp. 38-58 (1991).
3) Garcia-Molina, H. and Kogan, B.: An Implementation of Reliable Broadcast Using an Unreliable Multicast Facility, *Proc. of the 7th IEEE Symp. on Reliable Distributed Systems*, pp. 428-437 (1988).
4) Garcia-Molina, H. and Spauster, A.: Message Ordering in a Multicast Environment, *Proc. of IEEE ICDCS-9*, pp. 354-361 (1989).
5) Huhns, M. N. (ed.): *Distributed Artificial Intelligence*, Morgan Kaufmann Publishers (1987).
6) IEEE Project 802 Local Network Standards-Draft (1982).
7) Jacobson, V.: Congestion Avoidance and Control, *Proc. of the ACM SIGCOMM '88*, pp. 314-329 (1988).
8) Kaashoek, M. F., Tanenbaum, A. S., Hummel, S. F. and Bal, H. E.: An Efficient Reliable Broadcast Protocol, *ACM Operating Systems Review*, Vol. 23, No. 4, pp. 5-19 (1989).
9) Lamport, R.: Time, Clocks, and the Ordering of Events in Distributed Systems, *Comm. ACM*, Vol. 21, No. 7, pp. 558-565 (1978).
10) Lamport, R., Shostak, R. and Pease, M.: The Byzantine Generals Problem, *ACM Trans. Prog. Lang. Syst.*, Vol. 4, No. 3, pp. 382-401 (1982).
11) Luan, S. W. and Gligor, V. D.: A Fault-Tolerant Protocol for Atomic Broadcast, *IEEE Trans. Parallel and Distributed Systems*, Vol. 1, No. 3, pp. 271-285 (1990).
12) 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).
13) Metcalfe, R. M.: Ethernet: Distributed Packet Switching for Local Computer Networks, *Comm. ACM*, Vol. 19, No. 7, pp. 395-404 (1976).
14) 中村章人，滝沢 誠：多チャネル上の全順序放送通信プロトコルについて，情報処理学会マルチメディアと分散処理研究会，39-1, pp. 1-8 (1988).
15) Nakamura, A. and Takizawa, M.: Totally Ordering (TO) and Partially Ordering (PO) Broadcast Protocol, *Proc. of the 4th Joint Workshop on Computer Communication (JWCC)*, pp. 35-43 (1989).
16) Nakamura, A. and Takizawa, M.: Reliable Broadcast Protocol for Selectively Ordering PDUs, *Proc. of the IEEE ICDCS-11*, pp. 239-246 (1991).
17) Nakamura, A., Takizawa, M. and Takamura, M.: STO Protocol: A Reliable Broadcast Protocol for Selectively Totally Ordering PDUs, 準備中.
18) Schneider, F. B., Gries, D. and Schlichting, R. D.: Fault-Tolerant Broadcasts, *Science of Computer Programming*, Vol. 4, pp. 1-15 (1984).
19) 関 俊文，岡宅泰邦，田村信介：知的分散システムにおける高信頼放送通信機構，電子情報通信学会論文誌，Vol. J 73-D-I, No. 2, pp. 117-125 (1990).
20) Takizawa, M.: Cluster Control Protocol for Highly Reliable Broadcast Communication, *Proc. of the IFIP Conf. on Distributed Processing*, pp. 431-445 (1987).
21) Takizawa, M.: Design of Highly Reliable Broadcast Communication Protocol, *Proc. of IEEE COMPSAC 87*, pp. 731-740 (1987).
22) Takizawa, M. and Nakamura, A.: Totally Ordering Broadcast (TO) Protocol on the Ethernet, *Proc. of the IEEE Pacific RIM Conf. on Communications, Computers and Signal Processing*, pp. 16-21 (1989).
23) 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).
24) Takizawa, M. and Nakamura, A.: Partially Ordering Broadcast (PO) Protocol, *Proc. of IEEE INFOCOM '90*, pp. 357-364 (1990).
25) Takizawa, M. and Nakamura, A.: Reliable Broadcast Communication, *Proc. of IPSJ Info-Japan*, pp. 325-332 (1990).
26) Takizawa, M. and Deen, S. M.: Synchronization Problems of Compensate Operations in the Object-Model, *Proc. of the Int'l. Working Conf. on Cooperating Knowledge Based Sys-*

tems, pp. 1-19 (1990).
27) Wall, D. W.: Selective Broadcast in Packet-Switched Networks, *Proc. of the 6th Berkely Workshop on Distributed Data Management and Computer Networks*, pp. 239-258 (1982).

(平成3年5月8日受付)
(平成3年12月9日採録)

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年同大学大学院工学研究科修士課程修了．現在，同大学院理工学研究科博士後期課程在学中．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12月6日生．1973年東北大学工学部応用物理学科卒業．1975年東北大学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）．