システムソフトウェアと
オペレーティング・システム 64－2
マルチメディア通信と分散処理 65－2
（１９９４．５．１９）

グループ通信プロトコルにおける合意レベルの評価

立川 敬行　滝沢 誠

東京電機大学理工学部経営工学科

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協調動作を支援する高信頼なグループ通信が必要となる．高信頼なグループ通信では，各エンティティが送信したメッセージを，グループ内の一つ以上の宛先に届ける必要がある．このとき，グループ通信が提供するサービスとして，全宛先で受信されたか，また，どのような順序で受信されたかといった種々のサービス品質(QOS)がある．グループ通信では，これらのサービスを提供するために複数のエンティティ間での合意が必要となるが，どのような種類のサービスを提供するかにより，合意方法が異なってくる．本論文では，様々な分散形態を持つグループにおける合意方法と，その合意のレベルを評価することにより，分散型システムにおける合意手順について考察する．

Evaluation of Agreement Level in Group Communication Protocols

Takayuki Tachikawa　Makoto Takizawa

Department of Computers and Systems Engineering
Tokyo Denki University
Ishizaka, Hatoyama, Hiki-gun, Saitama 350-03
e-mail {tachi, taki}@takilab.k.dendai.ac.jp

In distributed systems, group communication among multiple entities is required in addition to the conventional one-to-one communication. Group communication protocols provide multiple entities with some quality of services (QOS) on data transmission in the group. For example, messages have to be delivered to all the destination entities in the group, and every application entity can receive messages in a well-defined order in the presence of multiple entities sending messages. The entities in the group have to make a level of agreement on how to support QOS required by the applications. This paper discusses QOSs of the group communication, what kind of agreement is required to support QOSs, and evaluates the agreement protocols.

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，信頼性のある通信が必要となる．高信頼放送型通信プロトコルは，グループ内で送信されたメッセージを，全宛先で，一定の順序で原子的に受信させるものである [4,13,15]．放送型通信プロトコルについては，文献 [3–5, 13] 等で論じられている．エンティティ間で通信されるメッセージをプロトコルデータ単位(PDU)とする．マルチメディアデータのグループ通信では，各PDUの原子的受信が保障されてなくても，なるべく早く送信順に宛先に届ける必要がある．以上のように，送信されたPDUが原子的に受信されるかどうか，どのような順序で受信されるかといった種々のサービス品質(QOS)が，応用から要求される．グループ内のエンティティが行う合意方法は，提供しようとするQOSに依存したものとなる．本論文では，様々なグループの形態，合意方法，合意レベルを示し，それらの評価を行う．また，QOSに適した合意手順を示す．

第2章と3章では，グループ通信のモデルを示し，合意レベルを定義する．第4章と5章では，実現方式を論じ，合意手順について述べる．

2 グループ通信

2.1 システム階層

通信システムは，図1に示す三階層から構成される．群Cは，$n(\geqq 2)$個のシステムSAP S_1, ..., S_n の組である．システム層のエンティティ E_i は，網層のサービスを利用して，応用層のエンティティ A_i に高信頼な放送型通信サービスを S_i を通じて提供する．ここで，E_1, ..., E_n は Cに含まれるとし，$C = \langle E_1, ..., E_n \rangle$ と書く．

2.2 通信システムの障害

高速網 [1] の通信は高信頼であるが，各 E_i の処理速度よりも転送速度が速い．このためにバッファのオーバーランにより，E_i が PDU を受信できない場合がある．従って，高速通信システムでは，PDU 紛失が主要な障害である．

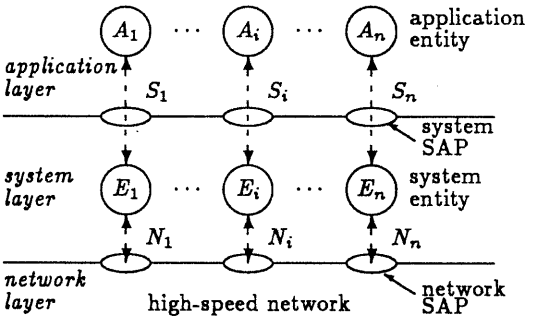

図 1: 群 $C = \langle E_1, ..., E_n \rangle$

2.3 サービスのモデル

群の放送型通信サービスを，送受信 PDU の系列であるログの集合としてモデル化する．以下，p と q は任意の PDU を示す．ログ L を $<p_1 ... p_m>$ と書く．$h < k$ のとき，p_h は p_k に L 内で先行する ($p \rightarrow_L q$)．群を提供する各 E_i は，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．E_i が，p の後に q を送信，受信したならば，各々 $p \rightarrow_{SL_i} q$，$p \rightarrow_{RL_i} q$ である．

[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 が同じ PDU を含むとき，両者は**情報同値**である．□

[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, ..., 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．□

PDU 間の因果関係 [2, 3] を考える．$s_i[p]$ と $r_i[p]$ は，それぞれ E_i の p の送受信事象とする．任意の事象 e_1 と e_2 に対して，$e_1 \rightarrow e_2$ [8] は，e_1 が e_2 よりも前に起きたことを示す．

[定義] 任意の p と q について，$s_i[p] \rightarrow s_j[q]$ ならば，$p \rightarrow_{RL_k} q$ であるとき，RL_k は**因果関係保存**である．□

2.4 網サービスのモデル

網層を利用して，E_1，E_2，E_3 が各々 a b，p q，x y z を送信した場合を図 2 に示す．単一チャネル (1C) サービスでは，各受信ログ RL_i は順序保存で順序同値である．これは，高速網のモデルである．各 E_i は，PDU を同一順序で受信できるが，紛失する場合がある．図 2(a) では，各 E_i は同一の順序で PDU を受信しているが，E_1 は p を，E_3 は z を紛失している．

多チャネル (MC) サービスでは，各 RL_i は順序保存である．これは，複数の高速リンクで計算機が接続されたシステムのモデルである．図 2(b) で，各 RL_i は送信順序を保存しているが，E_2 は y を，E_3 は x を紛失している．

多経路チャネル (MRC) サービスでは，各 RL_i の保存性と同値性が保障されない．各 E_i が，PDU を異なるリンクを用いて送信する ATM サービスをモデル化したものである．図 2(c) では，E_2 と E_3 は，p と q を、異なった順に受信し，E_1 は p を紛失している．

高信頼な放送型通信 (R) サービスでは，各 RL_i は、情報および順序保存で，かつ同値である．図 2(d) では，各 E_i は同一の順序で PDU を受信しており，かつ，PDU の紛失がない．

RL_1: $< a\ x\ b\ y\ z\ q\ ]$
RL_2: $< a\ x\ b\ p\ y\ z\ q\ ]$
RL_3: $< a\ x\ b\ p\ y\ q\ ]$
(a) 1C service

RL_1: $< a\ b\ x\ p\ y\ q\ z\ ]$
RL_2: $< a\ x\ b\ z\ p\ q\ ]$
RL_3: $< a\ b\ p\ q\ y\ z\ ]$
(b) MC service

RL_1: $< b\ x\ a\ q\ z\ y\ ]$
RL_2: $< x\ b\ p\ a\ y\ z\ q\ ]$
RL_3: $< a\ x\ b\ z\ y\ q\ p\ ]$
(c) MRC service

RL_1: $< a\ x\ b\ p\ y\ z\ q\ ]$
RL_2: $< a\ x\ b\ p\ y\ z\ q\ ]$
RL_3: $< a\ x\ b\ p\ y\ z\ q\ ]$
(d) Reliable service

SL_1: $< a\ b\ ]$ SL_2: $< p\ q\ ]$ SL_3: $< x\ y\ z\ ]$

図 2: 通信サービス

3 合意

群 $C = \langle E_1, ..., E_n \rangle$ のエンティティ間で，通信された PDU の，原子的受信と受信順序についての合意が必要となる．

3.1 合意の種類

C内のどのエンティティ間で合意されるかにより，以下のような合意の種類がある．ここで，(i, j) は C内の $i\ (\leq n)$ 個のエンティティの中で，$j\ (\leq i)$ 個で合意が取られることを示す．

(1) (n, n) **合意**: C内の全エンティティ間で同一の決定を行う必要がある合意である．

(2) $(n, 1)$ **合意**: C内のエンティティが一つでも同一の決定を行ったときに合意される．

(3) (n, r) **合意**: C内の r 個のエンティティの合意が必要な場合である．

(4) **意味的合意**: 合意が必要であるエンティティ集合を，意味的に決める場合である．

コミットメント制御 [6] が (1) の例である．(2) は，冗長なエンティティの中の少なくとも一つが受信すればよい場合である．過半数 ($r > n/2$) 合意が (3) の例である．(4) は，C内で宛先を指定して送信した場合に用いられる．

3.2 原子的受信

原子的受信とは，C内で送信された PDU が全宛先で受信されるか，全くされないかのいずれかであることである．E_i が送信する p 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．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．E_k が原子的受信の判断を行なう基準として，以下の (1) ～ (3) の三段階 [13] がある．

(1) **受理**: E_k が自分宛の p を受信したとき，E_k で受理されたという．

(2) **前確認**: E_k が，「p の受信の確認通知を必要とする全 E_j が p を受理した」ことを知ったとき，p は E_k で前確認されたという．

(3) **確認**: 全 E_j が「全 E_j が p を受理したこと」を知ったことを，E_k が知ったとき，p は E_k で確認されたという．

(2) で，p は，原子的に受信されたことになる．しかし，「他の宛先が，全宛先で p が受信されたことがわかっている」かはわからない．(3) では，このことが全宛先でわかる．即ち，原子的受信が全宛先で確認される．

3.3 受信順序

各 E_i がどのような順序で PDU を受信するかが重要であり，以下の順序がある [4, 15].

(1) **送信順序保存** (LO): 順序保存である.

(2) **因果関係保存** (CO): 順序保存でかつ，因果関係保存である.

(3) **全順序** (TO): 順序保存でかつ，互いに順序同値である.

4 グループ通信の形態

4.1 通信の方向性

エンティティ間で通信されるデータの流れについて考える.

(1) 1 : n: 群 C内の各 E_i は，指揮エンティティ E_C を通じて，PDU の送受信を行う.

(2) **対等**: C内の各 E_i が，全エンティティと PDU の送受信を行うことができる.

(3) **選択的**: 各 E_i は，PDU を，C内の一部に選択的に送信できる.

(1) の例はクライアントとサーバ間の通信で，サーバが指揮エンティティとなる. (2) では，電子会議のように，エンティティ間で通信が行われる. (3) は，PDU が宛先を持つ場合である. 選択的通信には，静的と動的との二種類がある. 静的選択的通信では，Cの部分集合が副群 C' としてあらかじめ定義されている. 各 E_i は C' に対して，PDU の送信が行える. 副群が階層的に構成される場合 [12] もある. ISIS [2] では，静的選択的通信が行われている. 一方，動的選択的通信 [9] では，C内の各 E_i は，任意の時に，任意のエンティティの部分集合に送信できる.

4.2 合意を指揮するエンティティ

群内のどのエンティティが，合意についての決定を行うかについて考える.

(1) **分散型制御**: C内の各 E_i が，他の E_j との通信により，合意についての決定を行う.

(2) **非集中型制御**: p を送信した E_i が，全宛先で受信されたかどうかの判断を行う.

(3) **集中型制御**: 群開設時に決定された指揮エンティティ E_cが，決定を行なう.

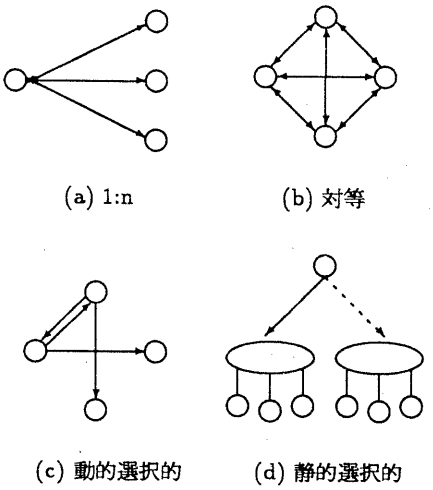

図 3: グループ通信の形態

ISIS [2] と Delta-4(xAMp) [16] は，非集中型制御を用いている. AMOEBA [7] は，集中型制御を用いている. [9, 10, 13, 14] は，分散型制御を用いている. 集中型制御のアルゴリズムは単純であるが，E_c の障害に対して頑強でなく，E_c への負荷の集中，E_c からの PDU 待ちによる遅延時間の増加の問題がある. 一方，分散型制御では，制御負荷が複数のエンティティに分散される. さらに，下位層に放送型通信サービスを用いた場合には，互いに他のエンティティの送信した PDU を受信できるので，あるエンティティの決定を待たずに自分自身で判断を行うことができる.

5 合意手順

原子的受信と受信順序の合意手続きについて考える. 群 $C = \langle E_1, ..., E_n \rangle$ とする.

5.1 原子的受信

A. 集中型制御

集中型と非集中型制御での合意方式として，二相コミットメント制御 [6] がある. まず，PDU p を受信した各 E_i は，指揮 (非集中型制御では送信元) エンティティ E_C に対して受信通知を送信する. E_C は，各 E_i から受信通知を受信したならば，確認通知である PDU を各 E_i に送信する. これにより，各 E_i は各エンティティで p が受理されたことがわかる.

B. 分散型制御

次に，分散型の制御 [13] を考える [図 4]. ここで，p の宛先を $p.DST$ と書く.

(1) **受理**: E_i からpを受信し，$E_k \in p.DST$ ならば，p は E_k で受理する.

(2) **前確認**: E_k が，全ての $E_j \in p.DST$ 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 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. (2) の段階で，各 E_k は p の原子的受信を決定できる. p を前確認する PDU が届かないなら，E_k は E_j に問い合わせることにより，p の前確認を確認できる. 応用層に PDU をある順序で届けるためには，前確認の段階ではまだ PDU を応用層に送れない.

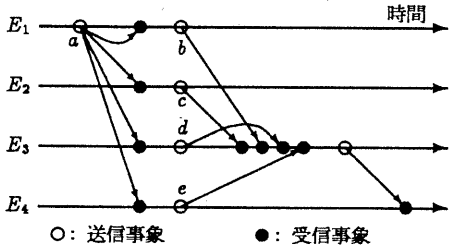

図 4: 三相手順

5.2 受信順序

PDU の受信順序についての合意を考える. ここで，送信される各 PDU は，シーケンス番号 seq と，各 E_k から受信予定の PDU のシーケンス番号(受信確認) ack_k $(k = 1, ..., n)$ を持つ [14]. $p.SRC$ は p の送信元を示す.

A. 無順序 (N)

順序を全く考慮しない場合には，C内の各エンティティ間での合意は必要としない.

B. 送信順序保存 (LO)

LO では，各 RL_i は順序保存，即ち，RL_i 内の任意の p と q について，$p.SRC = q.SRC$ で $p.seq < q.seq$ ならば $p \rightarrow_{RL_i} q$ である. 各 E_i がこの判断を行うために，他の E_j と合意を得る必要はなく，E_i が判断を行うことができる.

C. 因果関係保存 (CO)

CO で，各 RL_i 内の任意の p と q について，(1) $p.SRC = q.SRC$ で $p.seq < q.seq$ であるか，(2) $p.SRC(= E_j) \neq q.SRC$ で $p.seq < q.ack_j$ ならば，$p \rightarrow_{RL_i} q$ でなければならない [11]. 従って，各 E_j からの ack が必要となる.

D. 全順序 (TO)

TO で，各 RL_i は順序保存でに順序同値である. 各 RL_i が順序同値であることの合意を得るためには，群内の全エンティティと合意を得る必要がある. 即ち，(n,n) 合意が必要となる.

6 評価

6.1 原子的受信

前確認 (PACK) と確認 (ACK) のために必要となる PDU 数と遅延時間について評価する [表 1]. ここでは，m 個のエンティティの合意が必要となる (n, m) 合意について考える $(m \leq n)$. T は，各エンティティが PDU を一つ送信する時間単位 (ラウンド) とする. 集中型制御では，はじめに E_i が E_C に PDU を送信するため，非集中型制御に比べて PDU 数が一つ多い.

6.2 受信順序の合意手順

表 2 に，網層のサービスに対して，受信順序を保障するために必要な合意レベルを示す. ACC, PACK, ACK は各々，受理，前確認，確認の合意レベルを示す.

7 おわりに

本論文では，グループ通信における QOS と，これにより要求される合意レベルについて評価した. また，様々なグループ通信の形態に対する制御方法を示した.

参考文献

[1] 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.221-264 (1991).

表 1: 原子的受信の合意手順

制御	集中型		非集中型		分散型	
方向性	1 : n		1 : n		対等	
	PDU	delay	PDU	delay	PDU	delay
前確認 (PACK)	$m+3$	$3T$	$m+2$	$3T$	$m+1$	$2T$
確認 (ACK)	$2m+4$	$5T$	$2m+3$	$5T$	$2m+1$	$3T$

表 2: 受信順序の合意手順

	無順序 (N)	送信順序保存 (LO)	因果関係保存 (CO)	全順序 (TO)
単一チャネル (1C)	—, ACC	—, ACC	—, ACC	—, ACC
多チャネル (MC)	—, ACC	—, ACC	(n,n), PACK	(n,n), ACK
多経路チャネル (MRC)	—, ACC	$(n,1)$, ACC	(n,n), PACK	(n,n), ACK

[2] Birman, K., Schiper, A., and Stephenson, P.: Lightweight Causal and Atomic Group Multicast, *ACM TOCS*, Vol.9, No.3, pp.272-314 (1991).

[3] 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).

[4] Garcia-Molina, H. and Spauster, A.: Message Ordering in a Multicast Environment, *Proc. of IEEE ICDCS-9*, pp.345-361 (1989).

[5] Garcia-Molina, H.: Ordered and Reliable Multicast Communication, *ACM TOCS*, Vol.9, No.3, pp.242-271 (1991).

[6] Gray, J.: Notes on Database Operating Systems, *An Advanced Course*, *Lecture Notes in Computer Science*, No.60, pp. 393-481 (1978).

[7] Kaashoek, M. F. and Tanenbaum, A. S.: Group Communication in the Amoeba Distributed Operating System, *Proc. of IEEE ICDCS-11*, pp.222-230 (1991).

[8] Lamport, L.: Time, Clock, and the Ordering of Events in Distributed Systems, *CACM*, Vol.21, No.7, pp.558-565 (1978).

[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.: Priority-Based Total and Semi-Total Ordering Broadcast Protocols, *Proc. of IEEE ICDCS-12*, pp.178-185 (1992).

[11] Nakamura, A. and Takizawa, M., "Causally Ordering Broadcast Protocol," to appear in *Proc. of IEEE ICDCS-14*, 1994.

[12] Takizawa, M., Takamura, M., and Nakamura, A.: Group Communication Protocol for Large Group, *Proc. of IEEE LCN*, pp.310—319 (1993).

[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. and Nakamura, A.: Partially Ordering Broadcast (PO) Protocol, *Proc. of IEEE INFOCOM'90*, pp.357-364 (1990).

[15] 滝沢 誠, 中村 章人: 放送型通信アルゴリズム 情報処理学会誌, Vol.34, No.11, pp.325-332 (1993).

[16] Verissimo, P., Rodrigues, L., Baptista, M.: AMp: A Highly Parallel Atomic Multicast Protocol, *ACM SIGCOMM'89*, pp.83-93 (1989).