フォールトトレランスな群制御プロトコルの設計*

5V−2

鈴木等　中村章人　滝沢誠†

東京電機大学‡

1　はじめに

通信技術の発展と標準化に伴って、システムは分散型となってきている。グループウェアシステム等の新しい分散型応用では、二つの通信実体間の通信に加えて、複数の通信実体間のグループ通信が必要とされている。

これまでに、EthernetのMAC副層や無線システムが提供する低信頼な放送通信サービス上で、群と呼ばれる$n(\geq 2)$個のSAPに対して高信頼な放送通信サービスを提供する放送通信プロトコルが検討されている[1-4]。下位層でのプロトコルデータ単位(PDU)の紛失に加えて、実体の停止等の障害を考察する必要がある。

本論文では、実体が停止、復旧する動的群に対する高信頼な放送通信サービスを提供するためのプロトコルを検討する。本論文では、第2章で基本定義について、第3章で動的群制御の手続きについて論じる。

2　基本定義

2.1　1チャネル(1C)サービス

本論文では、Ethernet MAC副層が提供する低信頼な放送通信サービスを抽象化したものを1チャネル(1C)サービスとする。

[定義] 1Cサービスとは、PDUの紛失の可能性はあるが、ある実体が放送したPDUは、全実体において、同一の順序で受信されるサービスである。□

2.2　群

群とは、従来の2つのSAP間のコネクションの概念を、$n(\geq 2)$個のSAP間に拡張した概念である。<N−1>群スキーム$\underline{C}$とは、$n(\geq 2)$個の<N−1>SAP $S_1, \ldots, S_n$の組$< S_1, \ldots, S_n >$である。S_jでSDUを送受信している時、S_jは活性(A)である。そうでなければ、非活性(I)とする。$state(S_j)$をS_jの状態とする。S_jはE_jにより提供されているとする。このときE_jはC内にあるとする。群はスキームとその例で与えられる[図1]。

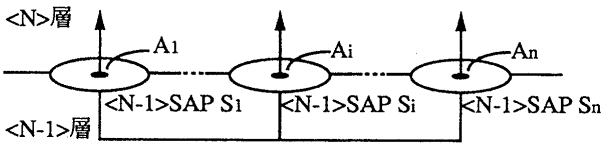

図1. <N−1>群

$state(S_i) = E_i$が提供するS_jの状態$\in \{A, I\}$

$state(C) = < state(S_1), \ldots, state(S_n) >$

各群スキーム$\underline{C}$に対して、同時に高々1つの例Cが存在する。群には静的群と動的群の2種類がある。静的群では、群内のSAPが全て活性である。動的群では、SAPの状態が変化する。

2.3　ログ

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としてモデル化する。ログLとは、PDUの系列$< p_1, \ldots, p_m > (m \geq 0)$であり、各実体の通信履歴である。群内の各実体$E_k$は、送信ログ$SL_k$と受信ログ$RL_k$を持つ$(k = 1, \ldots, n)$。L内で$PDU$ pがqに先行するとき、$p \rightarrow q$とする。

[ログの同値関係]2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。

(1) 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、一方で$p \rightarrow q$ならば、他方でも$p \rightarrow q$である。

(2) 情報同値:RL_i内のPDUの集合と、RL_j内のPDUの集合が同一である。

(3) 同値:RL_iとRL_jが順序かつ情報同値である。

[ログの性質]RL_iについて、以下の性質を定義する。

(1) 順序保存: 各E_jと、RL_i内のE_jから受信した任意の二つのPDU pとqについて、SL_j内で$p \rightarrow q$ならば、RL_iでも$p \rightarrow q$である。

(2) 情報保存:RL_i内のPDUの集合が、$SL_1, \ldots, SL_n$内に含まれるPDU集合の和である。

(3) 正しい: 順序保存でかつ情報保存である。

2.4　高信頼放送通信サービス

放送通信サービス内の活性な各SAP S_jを提供するE_jのRL_jが正しいならば、このサービスを信頼放送通信サービスとする。

[全順序放送通信(TO)サービス] サービス内の全受信ログが正しく、かつ互いに順序同値である。

TOサービスでは、各実体から受信したPDUの順序は、放送順序と同じであり、PDUの紛失がない。また、異なる実体から受信したPDUの受信順序も各実体で同一である。図2にTOサービスの例を示す。

E_1　RL_1: < a x b p y z c q d]　SL_1: < a b c d]
E_2　RL_2: < a x b p y z c q d]　SL_2: < p q]
E_3　RL_3: < a x b p y z c q d]　SL_3: < x y z]

図2. 全順序放送通信(TO)サービス

3　動的群制御

動的群は、群スキームに対する動的群と群スキームの例に対する動的群の2種類について考える。

3.1　群スキームの例に対する動的群

信頼放送通信サービス[1-4]では、群内のあるSAP S_jが非活性になる、即ちE_jが停止した場合、群内の全実体に対する通信サービスを提供できない。例えば、冗長なモジュールを含んだシステムでは、あるモジュールが停

*Design of Fault-tolerant Dynamic Cluster Contorol Protocol

†Hitoshi Suzuki, Akihito Nakamura, and Makoto Takizawa

‡Tokyo Denki University

止しても残りのモジュールにより処理を続行できる。本論文では、実体停止障害をモデル化し、群内のある実体が停止した場合でも、他の実体間にTOサービスを提供するための手続きについて述べる。

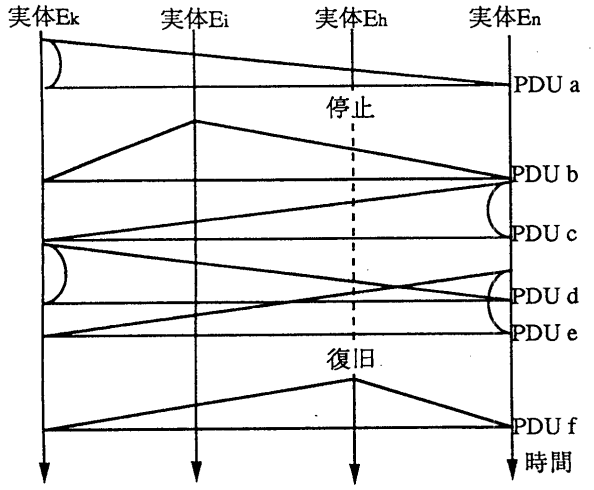

図3. 実体の停止と復旧

[実体停止] ある実体が、一定時間以上動作しないとき、停止しているとする。□
群内の実体は、ある実体の停止を、確認通知待ちでのタイムアウトによって検出する。

3.2 障害手続き

低信頼な1Cサービス上のTOプロトコル[1]を基本として、障害実体を検出し、正常動作している実体間で正しい送受信を行なうための手続きを示す。本方式は、データ転送を停止せずに動作できる。TOプロトコルにおける各PDU pに以下の項目を追加する。また、各実体は停止した実体を示すビットマップ$emask$を持つ。

$p.BMAP$= 停止した実体を示すビットマップ。

[実体停止の検出条件] ある実体E_kが、E_hからのPDUを一定時間以上受信できないとき、E_kは実体停止合意手続きを実行する。

[実体停止合意手続き](1)E_kは、障害実体E_hに対して、$emask$のhビットをONにし、$s.ACK_i$=REQ_i(i=1,...,n), $s.BMAP$=$emask$なる$STOP$ PDU sを放送する。
(2)$STOP$又は$STOP_PACK$ sを受信したE_jは、$emask$=$emask$ ∨ $s.BMAP$, AL_{ji}=$s.ACK_j$(j=1,...,n)として、$sp.BMAP$=$emask$, $sp.ACK_i$ =REQ_i(i=1,...,n)なる$STOP_PACK$ spを放送する。
(3)各E_iは、$emask$のビットがOFFの全実体から$STOP_PACK$又は$STOP_ACK$を受信したなら、$sa.BMAP$=$emask$とした$STOP_ACK$ saを放送する。ここで、REQ_j=AL_{ji}(j=1,...,n), $sp.ACK_j$=REQ_j。
(4)各E_iは、$emask$のビットがOFFの全実体から$STOP_ACK$ saを受信したとき、$emask$のビットがONの障害実体E_hから受信したPDUの中で通番がREQ_h以上のPDU p^hをRL_iの中から削除する。
(5)各E_iは、AL行列内[1-4]のE_hに関する確認情報に対して停止印を付け、以後データPDUの前確認、確認条件に関してE_hの情報を無視する。

3.3 復旧手続き

停止していたE_hが、復旧したとき、E_hを群に戻し、信頼放送通信サービスを再開するための手続きが必要である。本論文では、ある実体の復旧についての合意をとった後、データ転送を再開する手続きを紹介する。E_hを復旧した実体とする。

[復旧手続き](1)障害復旧したE_hは、$emask$の全ビットをOFFにし、$rc.BMAP$=$emask$, $rc.SEQ$=$rc.ACK_h$ = 0なる$RCV(RECOVERY)$ PDU rcを放送する。
(2)E_hからRCVを受信したE_jは、$emask$のhビットをOFFとする。E_jは、E_hからのRCV又はRCV_PACK rc受信したならば、$emask$=$emask$ ∨ $rc.BMAP$とする。E_jが復旧した実体ならば、REQ_h=$rc.SEQ$とする。また、$rcp.BMAP$=$emask$, $rcp.ACK_i$=REQ_i(i=1,...,n)なるRCV_PACK rcpを放送する。
(3)各E_j($j \neq h$)は、$emask$のビットがOFFの全実体からRCV_PACK rcpを受信し、$emask$==$rcp.BMAP$ならば、REQ_h=AL_{hi}=$rc.SEQ$, REQ_j=SEQ_j+1, AL_{hj}=$rcp.ACK_j$として、ALリセット手続きを実行する。また、$rca.BMAP$=$emask$なるRCV_ACK rcaを放送する。 そうでなければ、$emask$=$emask$ ∨ $s.BMAP$なる$RCV_$ $PACK$ rcpを放送し、(3)を実行する。RCV_PACK後に受信したRCVは、無視する。
(4)各E_iは、$emask$のビットがOFFの全実体からRCV_ACK rcaを受信し、各E_jについて、$rca.ACK_j$=REQ_jならば、PRL_iとRRL_i内のPDUの中でrca以前に受信したPDUをARL_iに移し、それ以外のPDUをRL_i内から削除する。
(5)各E_iは、ALテーブル内のE_hに対する停止印を削除し、データPDUの送信を再開する。

[ALリセット手続き]各E_jとp=$last(RL_{kj})$に対して、REQ_j=AL_{jh}=$p.SEQ$ + 1(j = 1,...,n)。

4 まとめと今後の課題

本論文では、データPDUの送受信を停止することなく、実体停止時に他の実体で合意するための手続きを論じた。また、復旧手続きについて論じた。

今後は、データ転送を停止しない復旧手続きを考察する。また、実体の停止と復旧が同時に発生する場合や、実体停止以外の実体障害についても合意手続きを考察すると共に、群を構成する実体数が変化する群スキームに対する動的群制御プロトコルを設計し、フォールトトレランスなサービスを提供し続けるための手続きの検討を行う。

参考文献

[1] Takiazwa, M., "Cluster Contorol Protocol Highly Reliable Broadcast Communication," *Proc. of the IFIP Conf. on Distributed Processing,* 1987, pp.431-345.

[2] Takizawa, M. and Nakamura, A., "Partially Ordering Broadcast Communication," *Proc. of the IEEE INFOCOM,* 1990, pp.357-364.

[3] Nakamura, A. and Takizawa, M., "Reliable Broadcast Protocol for Selectively Ordering PDUs," *Proc. of the 11th IEEE ICDCS,* 1991, pp.239-246.

[4] Nakamura, A. and Takizawa, M., "Priority-Based Total and Siti-Total Ordering Broadcast Protocols," *Proc. of the 12th IEEE ICDCS,* 1992, pp.178-185.