分散環境におけるリアルタイムな故障回復のためのプロセスのフォールトトレラント化

7H－4

横山　和俊　　遠城　秀和
ＮＴＴデータ通信（株）　技術開発本部

1．はじめに

近年、オンラインシステムのようなリアルタイム処理を分散環境で扱うことが必要となってきている。オンラインシステムのように無停止運転が要求されるシステムでは、故障発生時にもサービスを継続させるフォールトトレラント技術を分散環境で実現することが重要である。このとき、正常なサービスの滞りを防ぐため、短時間でサービスを再開することを可能にする故障対処が求められる。

分散環境におけるフォールトトレラント技術には、チェックポイント・ロールバック方式とプロセスの冗長構成による方式の２つに分類できる。前者は、ロールバックする時間が大きいという欠点がある。後者の場合、故障発生時の回復処理が比較的高速であり、リアルタイムな故障回復には有利である。

本稿では、分散環境において、プロセスの冗長構成を用いたリアルタイムな故障回復について述べる。

2．冗長構成によるフォールトトレラント化

プロセスの冗長構成を用いる方式は、サービス処理を提供するマスタプロセス（以降マスタと呼ぶ）と、マスタと同じ処理を実行する複製バックアッププロセス（以降複製と呼ぶ）を同期実行させることによりフォールトトレラントを実現する。これらの同期実行するプロセス群をグループと呼ぶ。故障発生時には、マスタから複製へ切り替える。しかし、通信路が低速なため、マスタと複製間で厳密な同期をとると、オーバヘッドが大きくなる欠点がある。そのため、分散環境においては、メッセージの送受信制御によりマスタと複製が矛盾なく実行させる手法がとられている。例えば文献[1]では、受信側ではマスタが順序通知を行うことで、送信側では送信済通知を用いることで、矛盾のない実行を制御している（図１参照）。文献[1]のメッセージ制御は以下の性質を満たすものである。

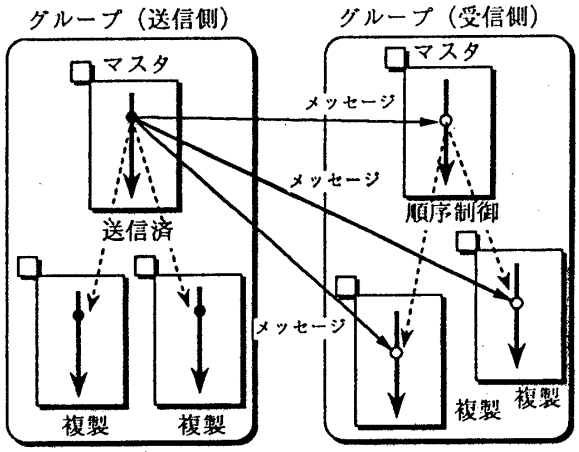

図１　冗長構成によるフォールトトレラント

（１）メッセージ受信に関して、メッセージの受信順序を同一にすることで、同じ実行履歴を持つことを保証する。

（２）メッセージ送信に関して、複製をマスタより先行させないことで、メッセージの重複や欠落がないことを保証する。

しかしながら、複製はマスタより先行することはないが、マスタから大きく遅れる可能性がある。すなわち複製がサービスを提供開始するまでの時間が大きくなる。そのため、リアルタイムな故障回復には不十分である。これを解消するには、複製がマスタとほぼ同じ実行状況にあることが必要となる。

3．冗長構成における遅延制御

文献[1]によって実現されているメッセージ制御方式に加え、複製がマスタから大きく遅れないように制御することが必要である。具体的には、マスタと複製が終了したメッセージ送信数の差を一定値以内に抑制する。この方式を実現するためには、以下の２つの機能が必要である。

（１）複製の遅延を検出する機能

（２）複製の遅延を解消する機能

Fault-tolerant Process for Real-time Recovery in Distributed Environment.

Kazutoshi YOKOYAMA and Hidekazu ENJO
NTTDATA COMMUNICATIONS SYSTEMS CORPORATION.

3．1　遅延検出

マスタと複製が終了した送信メッセージ数を把握する必要がある　このために各プロセスは以下の情報を持つ必要がある

message_count：自プロセスが終了したメッセージ送信数を示すカウンタである（MCと略す）。

group_info[pid]：グループ内の各プロセスが到達したメッセージ送信数を示す配列である（GI[pid]と略す）

例えば、マスタは図２に示すように動作することで、複製が終了したメッセージ送信の数を知ることができる。

[step1] マスタは、他プロセスへのメッセージ送信が終了したら、自MCを１インクリメントする。

[step2] 複製は、メッセージ送信を終了したら、自MCを１インクリメントする。さらに、マスタへ現在のMCの値を通知する（到達通知）。

[step3] マスタは、複製からの到達通知を受け取った場合、該当プロセスのGI[pid]を更新する。

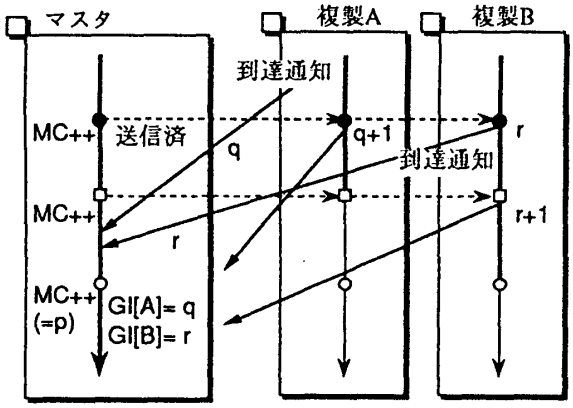

図２　メッセージ送信よる遅延の把握

例えば、マスタのMCをｐ、GI[A]をｑとすると、マスタは複製Ａの遅延がｐ－ｑと把握できる。

3．2　遅延解消機能

複製のメッセージ送信数がマスタより一定値（εとする）以上遅れた場合、複製の遅延を解消する機能が必要である。このためには、以下の項目を決定する必要がある。

（１）解消契機

（ａ）１つの複製がεより遅れたとき。

（ｂ）すべての複製がεより遅れたとき。

(a)は１つのプロセスだけが遅れた場合でも遅延解消を行うため、解消回数が多くなる。しかし、他の複製がε以内を走行しているため、遅延解消を遅らすことができる。(b)は逆に、解消回数は少ないが、遅延解消を即実行しなければならない。さらに、遅延解消が連続的に発生する可能性がある。

（２）解消方式

以下の２つの方式が考えられる。

（ａ）遅延が検出されたプロセスからの到達通知が届くまでマスタが待機する。

（ｂ）遅延が検出されたプロセスをグループから切り離し、マスタと同じ実行状況のプロセスを生成しグループに組み込む。

(a)は、制御が簡単であるが、マスタからε遅れた所までしか解消できない。そのため、連続的に遅延解消が発生しマスタの中断回数が多くなる。(b)はマスタと同じ所まで解消するため、再度遅延する可能性は少ない。解消契機と解消方式の組み合わせで表１に示す４つが考えられる。

表１　遅延解消機能の特徴

	契機	方式	特徴
A	(a)	(a)	解消が多く、かつ、連続。マスタの待機を遅らすことで走行を継続。
B	(a)	(b)	解消が多いが連続的でない。プロセス生成を遅らすことで走行を継続。
C	(b)	(a)	解消が(a)より連続的。マスタが即中断
D	(b)	(b)	解消が連続的。マスタが即中断。

故障発生時の複製への切り替え時間は４つとも余り変わりがない。遅延解消によるマスタの中断を考慮すると、最も少ないＢが有利と考えられる。

4．おわりに

本稿では、冗長構成によるフォールトトレラントにおいて、マスタと複製間の遅延を一定値以内に抑制する方式について述べた。今後は、解消契機と解消方式の特徴をハードウェアの性能を考慮して明確にする。さらに、提案方式を実現し有効性を評価する。

文献

[1] H. Higaki et.al. :"Group to group communications for fault-tolerance in distibuted systems", IEICE Trans. Vol.E76-D, No.11 (1993).