高信頼UNIX「風雅」のプロセス間通信機構*

6M－3

今井祐二†, 小村 昌弘‡
（株） 富士通研究所§
{kimai,komura}@flab.fujitsu.co.jp

1 はじめに

我々は高信頼UNIX「風雅」を開発した．風雅は，Chorusマイクロカーネルと，その上で動作するマルチサーバUNIXサブシステム(MiX)[1]をベースに採用している．MiXは複数のモジュールに分割されており，クライアント／サーバモデルで動作している．風雅ではMiXサブモジュールにプロセスペア手法を適用し，ソフト，ハードの故障からのリカバリを可能にしている．PA(Port Alias)は耐故障性を備えたプロセス間通信機構(Inter Process Communication(IPC))である．本稿ではPAの機能とその実現方法に関して述べる．

2 RPCの問題

マイクロカーネルが提供するIPCは，サブモジュールが持つport間で，メッセージを受け渡す機構である．portにはUnique Identifier(UI)と呼ばれるIDがついており，これが宛先として用いられる．

風雅では故障時に現用系から待機系への引き継ぎが起こるので，カーネルが提供するIPCを単純に使用しただけでは，次のような問題が起こる．

宛先の変化

引き継ぎによって，IPCを受け付けるサーバのポートが，故障した現用系から待機系に変化し，宛先UIが変化する．

メッセージの消失

故障発生時に，クライアントから送信中だったメッセージが，現用系の故障によって失われる．

冗長なメッセージの発生

メッセージの消失は，メッセージを再送することで再実行することができる．ただし，故障のタイミングによっては，すでにサーバで処理が行なわれている場合もありえる．処理が行なわれたにもかかわらず，再送されるメッセージを冗送と呼ぶ．処理の種類によっては，冗送メッセージを単純に再実行すると，誤った結果を引き起こす場合がある．典型的なケースは，ファイルの消去操作時の現用系故障である．クライアントからのメッセージにしたがって消去作業を行ない，クライアントに返信メッセージを送る段階で故障が起こった場合，引き継ぎ後にクライアントからの冗送メッセージを単純に処理すると，消すべきファイルが存在しない旨のエラーを起こしてしまう．

3 PAの機能

上記の問題は，プロセスペア化するすべてのサーバに共通なので，風雅では耐故障型プロセス間通信機構Port Alias(PA)を開発しプロセス間通信の処理を集約した．

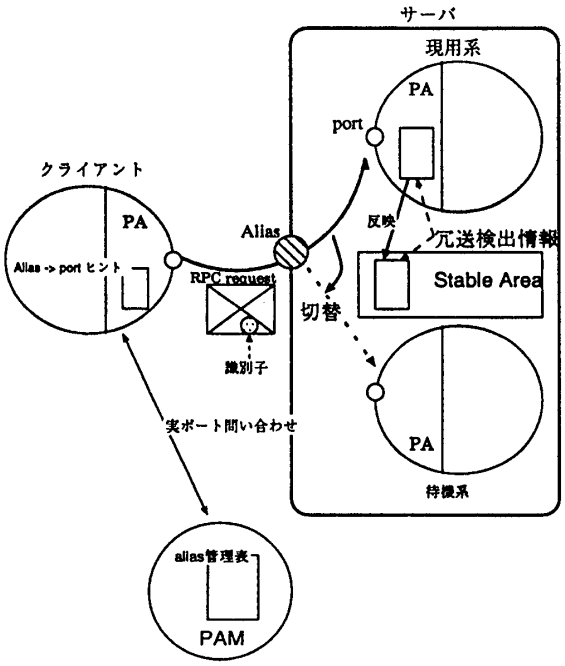
PAは，引き継ぎを越えて存在する永続的なIPC送受信ポート「alias」を用意する．aliasもUIを名前として持つ．サーバはalias間を耐故障性IPC(FT-IPC)を用いて、リクエストを受け渡すことができる．故障時には新現用系がaliasを引き継ぎ，PAがリクエストを正しく新現用系へ届ける．FT-IPCはマイクロカーネルのIPCの上に実現されており，故障時にはマイクロカーネルのIPCメッセージは消失が起きる．しかし，FT-IPCのリクエストはクライアントPA上に残っており，PAによって新しい現用系に再送される．再送は冗送となる可能性があるが，サーバ側のPA内で自動的に処理されるので，サーバからは隠蔽される．

*Inter Process Communication Mechanism of Highly Available UNIX "FUGA"

†Yuji Imai

‡Masahiro Komura

§Fujitsu Laboratories Ltd.



4 実現

PA は FT-IPC を実現するために，以下の 3 つの処理を行なう．

宛先の自動切替え

alias の作成・消滅・引き継ぎによって変化する，alias とその実際のポートとの関係を，Port Alias Manager(PAM) で管理する．

PA は FT-IPC リクエスト送信時に，宛先 alias の現在の port を PAM に問い合わせ，得られた port に対してリクエストを送信する．1 度問い合わせた alias と port の関係はヒントとして保管し，次回からはヒントを用いて IPC を行なう．故障時には，まず PAM が alias の port の情報を更新する．クライアントの PA は，古いヒントを用いて送信するため，一旦宛先不明で失敗するが，PAM に対して新しい port を要求して再送信する．

自動再送信

サーバの現用系が故障すると，リクエスト依頼中だったクライアントの PA に IPC の異常が通知される．PA は上記の宛先の切替処理を行なった後，自動的にメッセージを再送する．

冗長なメッセージの検出と処理

クライアント側の PA は FT-IPC のたびに，ユニークな識別子を作成し，メッセージに添付する．サーバは受信したリクエストの処理を進めていき、冗送の検出が必要になた時点で，PA に対して現在処理中のリクエストの冗送を検出するように依頼する．その際に，冗送が来た時に返すべき返信メッセージを同時に PA に登録する．以後，PA はメッセージを受信するたびに，メッセージから識別子を取り出して，冗送であるかどうかを検査し，冗送であった場合には登録された返信メッセージを自動的に返す．

待機系が引き継いだ後も，冗送検出と自動返信が継続しておこなわれる必要があるので，冗送検出情報と返信メッセージは Stable Area(SA) に保管される．この際，PA の状況とサーバの状況の整合性を保つために，サーバの進捗状況変化と，PA の冗送処理情報を，atomic write を用いて SA に格納する．

FT-IPC の結果メッセージがクライアントまで到着し，冗送が起こらないことが確定すると，クライアントから FT-IPC の終了が通知される．この通知を受けてサーバ側で冗送検出情報と返信メッセージ領域を回収する．FT-IPC 終了通知は，通常は，同じクライアントサーバ間の次の FT-IPC メッセージが送られる時に，メッセージに添付して送られるので，終了通知のためだけの余分な通信は起こらない．

参考文献

[1] batlivala et al., “Experience with SVR4 Over CHORUS”, USENIX Workshop on Micro-Kernels and Other Kernel Architectures April 27, 1992, pp. 223-242

[2] 岸本他,” 高信頼 Unix「風雅」”, 第 52 回情報処理全国大会論文集 6M1-2,4-6, 1996