高信頼UNIX「風雅」の耐故障ファイルシステム*

6M－5

金沢裕治†，村上岳生‡，　小暮淳§
（株）富士通研究所¶　富士通株式会社‖
kanazawa@flab.fujitsu.co.jp

1　はじめに

風雅システムは，Chorusマイクロカーネルベースの UNIXをプロセスペア方式により高信頼化したシステムである[2]．既存UNIXの高信頼化という性格上，本システムの実装においては，以下の要件を満たさなければならない．

・特別なハードウェアを必要としない．
・通常実行時の性能の劣化を可能な限り抑える．

プロセスペア方式によるＦＴ化では，現用系から待機系へのデータ受渡しオーバーヘッドによる性能劣化が問題になる．従来のシステムでは，専用メモリなど，特別なハードウェアを用いてきた[1]．本論文では，風雅システムに実装された耐故障ファイルシステムの概要と，通常実行時の性能劣化を抑える技法について述べる．

2　風雅ファイルシステムの概要

図１は，風雅の耐故障ファイルシステムマネージャ(FTFM)の説明図である．現用系とは別のサイトに待機系が存在しており，現用系と待機系は同じディスクにアクセスすることができる（あるいは，デバイスドライバレベルでミラーリングを行なっている）．ユーザプロセスが発行したシステムコールは，プロセスマネージャ(PM)が受け，PM がFTFM のクライアントとしてリクエストを送ることによって処理される．

現用系に故障が発生したときは，待機系のファイルシステムが，現用系の状態を復元し，以後，クライアントからのリクエストを受け付ける．引き継ぎデータは風雅システムが提供する基本サービスであるStable Area（SA）を通じて待機系に送られる．クライアントからのリクエストは，Port Alias（PA）を通じて送られるので，クライアントがFM の引き継ぎを意識することはない．

FTFM の既存 FM からの主な変更点は，現用系の故障後に待機系がサービスを引き継ぐ処理が追加されることと，現用系の通常実行時の処理に，引き継ぎデータを SA に格納する処理が追加されることである．ここでは，まず，引き継ぎ作業について述べることにする．

引き継ぎ作業で復元しなければならない状態には，２種類存在する．第１に，クライアントとの間で共有している状態である．例として，クライアントに渡したファイルの識別子とそれに対応する vnode，または，サーバ引き継ぎ後にクライアントから送られてくる再送リクエストを検出して同じリクエストを２回実行しないようにするためのデータ（冗送検出用データ）などが存在する．第２に，ディスク上のデータの一貫性である．例えば，ファイルのリンクカウントと，ファイルシステム内に存在しているディレクトリエントリの数は一致していなければならないが，これらのデータを１回のディスクへの書き込みですべて書き換えることはできないため，引き継ぎ時に何らかの方法で一貫性を回復する必要がある．

他サーバと共有している状態については，２種類の回復方式が考えられる．第１の方式は，引き継ぎデータとして SA を通じて待機系に送っておき，引き継ぎ時に回復する方式であり，第２の方式は，引き継ぎ時に他サーバと話しあい，複数サーバ間で矛盾のない内部状態を再構築する方式である．風雅では，基本的に第１の方式を使用し，第２の方式は，故障によるクライアントの消滅の有無を確認するだけにとどめている．

ディスク上のデータの一貫性についても，２種類の回復方式が考えられる．引き継ぎ時に fsck を行なって修復する方式と，一貫性を保つためのデータを引き継ぎデータとして待機系に送る方式である．ここでは紙面の都合上詳細な議論は省く

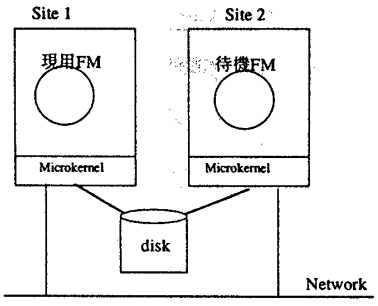

図１　FTFM の構成

*File System of Highly Available UNIX "FUGA"
†Yuzi Kanazawa, ‡Takeo Murakami, §Jun Kogure
¶Fujitsu Laboratries Ltd., ‖Fujitsu Ltd.

が，FTFMでは，一貫性を保つためのデータを引き継ぎデータとして待機系に送る方式を採用した．

以上をまとめると，FTFMの引き継ぎ時の動作は，以下のように進められる．

1) 他サーバとの間で共有されていたデータと，それに付随するデータをSAから読みだし，復元する．
2) SAから読み出した情報にしたがって，ディスク上のデータの一貫性を回復する．
3) 他サーバの生死を確認し，死んでいるサーバがあった場合は，それに対応するデータを消去する．

引き継ぎ処理をさらに詳しく検討することにより，SAに格納するデータを決定した．これに基づいて，現用系の通常実行時の処理に，引き継ぎデータをSAに格納する処理を追加した．現用系の通常実行時の処理は，この処理以外は，既存コードと同様である．

3 引き継ぎ用データの転送タイミング

風雅システムでは，特殊なハードウェアを前提とせずに通常実行時の性能劣化を抑えるという要件があるため，特に引き継ぎデータの受渡しオーバーヘッドが問題になる．

FTFMにおいては，分散データ共有の概念を応用することにより，特殊なハードウェアを用いずに性能の劣化を極力抑えながら，現用系から待機系への引き継ぎ用データの転送を行なうことを可能にしている．以下で，その概要を述べる．

分散システムにおいて満たされなければならない一貫性として，因果関係に基づくものが考えられる．例えば，プロセス1がデータAを変更した後，それに基づいてデータBを変更したとする．この後，プロセス2がデータB，Aの順に読み出したときに，両方旧内容が返ってきたり，データBだけ旧内容が返ってくるのは，論理的には矛盾がないので許されるが，データBは新内容，データAは旧内容が返ってくるのは因果関係に反するのであってはならないというものである（図2）（シリアライザビリティとは，データ間に因果関係が無い場合に順番を気にしなくてよい点が異なる）．

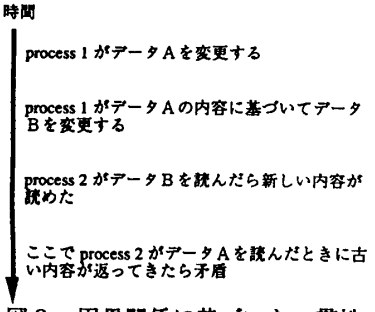

図2 因果関係に基づいた一貫性

これをFMのSAの内容の同期において満たすためには，以下のようにすればよい．SAの変更がデータAの変更，ディスクへの書き出しや他サーバへの通信(以下、外部データと書く)をデータBの変更と考える．現用系が外部データを変更すると，そのデータは待機系に見えるので，この処理を行なう直前には，SAの変更部分を待機系に送らなければならないことになる．逆に，最後に外部データを変更したあと故障が発生するまでの間に行なわれた作業の結果は，現用系のメモリ上だけに存在しており，待機系からは見えないため，このタイミングまで転送を遅延できる．待機系は，現用系が最後に外部データを変更した後の処理を，再実行することになる．

この方式だと，素朴な実装でも，ディスクへの書き込み，他サーバへ通信が発生したときだけ，待機系への通信が発生するため，粗い評価で実行時間が最悪で2倍程度になると予想される．実際には，データの内容を詳しく吟味し，ディスクに書き出すデータがSAの内容と論理的に関係がない場合は転送を省略することにより，さらにSAの同期を行なう回数を減らすことが可能である．

4 まとめ

FTFMの概要を紹介した．

本システムでは，分散データ共有の概念を適用することにより，現用系から待機系へのSA内容の転送回数を減らすことを可能にした．この方式は，ファイルシステム以外のサービスにも適用可能である．しかし，ネットワークモジュールなどの，他サーバとの通信が非常に頻繁に発生するサーバにおいては，今回紹介した方法を用いても，やはり性能劣化が問題になる．このようなシステムでの引き継ぎ方式，データ転送についての検討が，今後の課題になるだろう．

参考文献

[1] 村松他,"システムを止めずに保守・運用が可能なＯＳを開発",日経エレクトロニクス,No. 520, pp.209-223, 1991

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