氏名（本籍）	岸本光弘（静岡県）
学位の種類	博士（情報科学）
学位記番号	情博第136号
学位授与年月日	平成１２年３月２３日
学位授与の要件	学位規則第４条第１項該当
研究科，専攻	東北大学大学院情報科学研究科(博士課程)情報基礎科学
学位論文題目	高速通信機構を備えた並列計算機に関する研究
論文審査委員	（主査） 東北大学教授　白鳥則郎　東北大学教授　亀山充隆 東北大学教授　阿曽弘具　東北大学助教授　木下哲男 （工学研究科）

論文内容要旨

1　緒論

近年の，インターネットの爆発的な普及拡大によって，日常生活に関わる様々な経済活動がインターネットを通して利用できるようになってきている．インターネットにより提供されるサービスは，利用者にとっては，時間や場所の制約なしに，世界中から２４時間いつでも，常に最新の情報が利用できる非常に便利なサービスである．また、サービスの提供側にとっては，エンドユーザと直接のやりとりでき，従来のような大規模な営業／販売組織なしに，多数のユーザを獲得し維持できるという魅力がある．

しかし，このようなインターネットベースのサービスを実現するサーバシステムは，従来のサーバシステムにはなかった次のような難しい要求を満たさなければならない．

サービスはさまざまな時間帯のユーザから利用されるため，システムには一日２４時間の無停止運転を可能にする高信頼性が要求される．利用者の業務がシステムが提供しているサービスに深く依存している場合には，サービスの停止は訴訟問題となる場合がある．

また，新規のインターネットサービスは，通常，小規模構成からスタートする．実際かなりの数のサービスが，ディスクトップ PC１台でサービスを開始する．そして，運良くサービスが好評になると，短期間で利用者が急増する．従来のクライアント／サーバシステムとのもう一つの大きな違いは，クライアント数の上限を設定できないことである．そのため，これら新しいサーバシステムには，従来とは桁違いの処理能力のスケーラビリティが要求される．

さらに，サービス提供者間での競争があるので，人気のあるサービスであっても，他者

優位性を維持するために，不断の機能拡張とサービスメニューの追加を行う必要がある．しかも機能拡張や追加は，現在提供中のサービスを停止せずに実施する必要がある．

このような，サーバシステムに求められる高信頼性，性能のスケーラビリティ，継続的な機能の追加拡張性を実現できるシステム構成として，クラスタシステムがある．

本研究では，オープンシステムの代表である UNIX クラスタシステムにおける信頼性向上と分散処理性能向上について論じる．

2　プロセスの二重化による OS の高信頼化

従来，OS の高信頼化は、試験やレビューを網羅的に行ない，内在するソフトのバグ等の故障を排除することで追求されてきた（故障回避技術）．しかし現実には，OS はその規模や複雑さが膨大なため，故障回避技術によって故障を完全に取り除くことは不可能であった．

一方，誤りの発生を内部で検出し回復する耐故障技術が知られており，幾つかの特殊用途の OS が耐故障技術に基づいて設計実現されている．しかし，耐故障技術を UNIX のようなオープンな OS に適用するためには，様々な問題を解決しなければならない．そこで，耐故障技術の一つであるプロセスペア方式を，既存のオープンな OS の代表である UNIX に対し適用する方法を検討した．

本研究では、マイクロカーネル化されたマルチサーバ UNIX をサーバ単位にプロセスペア方式を適用して耐故障化する方式を提案する．本方式はハードウェアおよびソフトウェアの故障による誤りに対処することができる．プログラムの保守性や拡張性を損なわずにサーバを耐故障化するため，サーバとは独立に故障管理の機構を用意する．これにより、各サーバは故障管理機構に対し，誤りの発生を通知したり，故障管理機構からの指示に従って回復処理をするだけで済み，個別の誤りの詳細を知る必要がない．

既存のオープンな UNIX OS の拡張性や保守性を維持したまま，小規模な修正でプロセスペア方式の耐故障性を付与するためには，以下の機能が必要である．

- 誤りの検出から回復までの一連の処理を指示する機能．
- クライアントとサーバ間の通信での，宛先の変化やメッセージの消失や冗送への対処．
- 入れ子になったＮ段通信において，段階毎に再実行を行いサーバ状態を回復する方式．
- 現用系プロセスの内部状態の更新を，アトミック性を保持して効率良く待機系プロセスに伝播すること

本論文では、上記機能を提供する故障管理機構を，個々のサーバプロセスとは独立に用意することを提案した．故障管理機構は，故障管理サーバ（Fault Manager），耐故障通信機構（Port Alias），データ安定格納機構（Stable Area）から構成される（図１参照）．

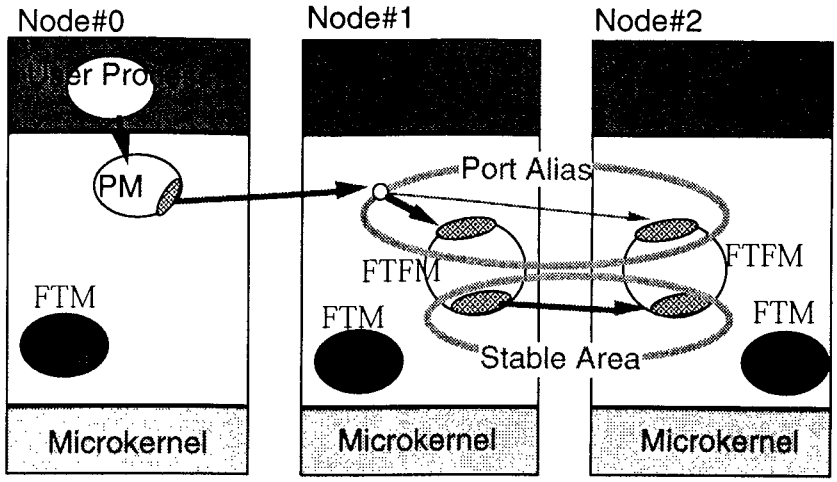

図１．高信頼 OS のアーキテクチャ

提案した耐故障管理機能を作成し，それを利用してファイルサーバの耐故障化を行い，本方式の有効性を実証した．

作成したプロトタイプの耐故障性を確認するために次のような実験を行った．

(1) ハードウェアの故障をエミュレートするため，slave の電源を動作中に落す．

(2) マイクロカーネルのソフト故障をエミュレートするために，動作中に slave のリセットボタンを押す．

(3) 故障発生ツールを使って OS の内部で人工的なソフト故障を起す．

いずれのケースに於いても，耐故障機構が期待通り動作し、エラーはユーザには見えず，引き継ぎ時間の分だけ処理が遅れただけであった．

さらに、高信頼性を付加するために必要となる既存プログラムの修正は軽微で，想定する全ての種類の故障に対する耐故障性を確認した． システムの高信頼化に必要な修正量を客観的に測る指針として，追加／削除／変更したプログラムの行数を調べた．修正前の行数に対する比率（%）で表したものが表 1である．表中の「新規」はファイル単位で追加した新規機能を，「追加」および「削除」は既存のファイルに追加／削除した行数の比率を表している．

表 1 プログラム変更行数の割合（%）

	新規	追加	削除	合計
全体	8.0	0.5	0.02	8.5
ファイルシステム	4.5	0.8	0.05	5.4

UNIX サブシステム全体での修正箇所は，FTM や耐故障通信機構，データ安定格納機構

といった新規機能の開発分が大部分を占めており，それらを考慮に入れても変更量は既存プログラムのステップ数の 8.5%であった．また，既存のサーバである FM に対する変更もプロセスペアの初期化処理や引き継ぎ時の処理など新規開発分が多く，純粋に既存コードを追加／削除した行数は1%以下であることが分かった．

耐故障化修正により，通常実行性能が劣化したが，オーバヘッドを詳細に分析し，削減方法の目処をつけた．

3　システムエリアネットワークを用いた性能向上

クラスタシステム上で，高性能な並列分散プログラムを実現するためには，複数ノード上に分散配置されたプロセス間の通信処理を高速化する必要がある．クラスタ向けの高速結合網として， System Area Network (SAN)が開発され， SAN における高速通信の標準仕様として，Virtual Interface Architecture (VIA)が提案されている．

しかし,VIA およびその関数インターフェースである Virtual Interface Provider Library (VIPL)の現在の仕様(version 1.0)には，特定の CPU アーキテクチャと OS にのみ良く適合する規定が含まれており，様々な CPU や OS 上での効率的な VIA の実現を妨げている．

そこで，本論文では、OS や CPU アーキテクチャの中立性から見た，現在の VIA および VIPL 仕様の問題点を明確化し，他の OS や CPU へ適用するための以下のような仕様拡張を提案した．

- マルチプロセス・プログラミングモデルのサポート
- シグナル処理のサポート
- ビックエンディアンの効率的サポート

本改善案は，標準化作業を経て VIPL の業界標準仕様へ反映される．

また実際に UNIX クラスタ上に高速通信機構を実装して提案の妥当性を確認した．作成した UNIX VIA は、図2に示すようにユーザエージェント（もしくは VIPL）と呼ばれるライブラリと，カーネルエージェントと呼ばれるデバイスドライバ部分から構成される．

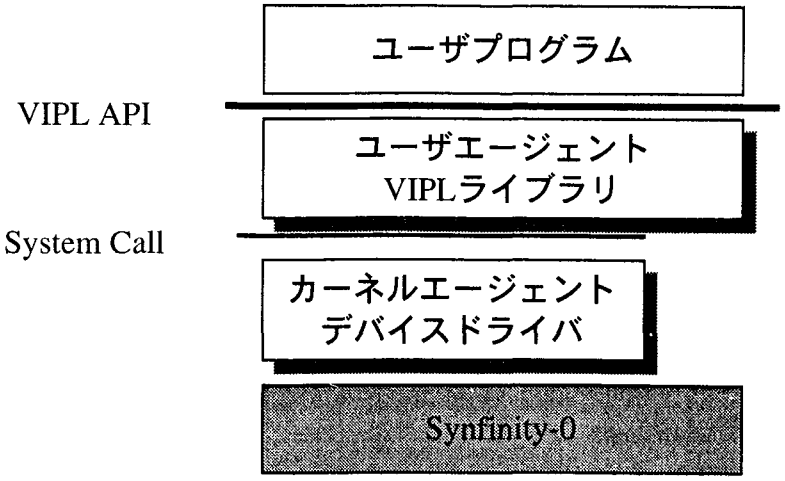

図 2 UNIX VIA の全体構成

デバイスドライバは以下のモジュールから構成されている．作業キュー管理機構（送信キューと受信キューにエンキューされたディスクリプタの指示に従って処理を行い，通信管理モジュールにデータ通信の指示を行う）．メモリ登録管理機構（ユーザプロセスからのメモリ登録／解除要求に従い，物理ページの固定／解除を実行する．またユーザ仮想アドレス，カーネル仮想アドレス，I/O バス仮想アドレス間のアドレス変換を担当する）．接続管理機構（VI の接続開設，切断処理を行う）．通信キュー機構（データの送信指示に従い，ハードウェアの通信起動処理を行う．また，ハードウェアからの割り込みを受け付け，通信完了処理を実行する）．

従来の UNIX VIA の実装は，高性能を実現するプロトタイプレベルのものであり，性能評価もごく小規模のマイクロベンチマークを用いて行っていた．そのため，大規模な商用分散並列プログラムで VIA を使用する上で必要な，大容量の実現や，保護機構，可用性等を可能にする検討はなされていなかった．

本研究では、大規模な商用分散並列プログラムで利用可能な，高機能性／大容量／高性能／資源保護／高可用性／高品質を実現した，実用的な VIA を実現する以下の手法を検討し．SPARC UNIX 上で実際にシステムを構築した．大規模な商用分散並列プログラムで実際に利用し，拡張提案の有効性を明らかにした．

- Oracle OPS と SymfoWARE は，もともとプロセス間での VI 共有を必要とせず，VI 非共有の仕様拡張のマルチプロセスモデルで対応できた．しかし，CrispORB はプロセス間 VI 共有機能が必要であった．さらに， VI 共有の機能拡張は Oracle OPS の使用 VI 個数の削減にも有効である．
- SymfoWARE と CrispORB は当初からスレッドセイフだったが，Oracle OPS はシングルスレッドプロセスである．そのため VI の利用にあたり，プログラムの慎重な調整

が必要である.

- シグナルに関する仕様拡張は，全てのプログラムで妥当なものであり，既存機能に影響を与えない.
- ビックエンディアン対応は，ハードウェア側で対処し，プログラムは通常どおり代入文を使用することができる.

4 結論

本研究では，クラスタ構成のサーバシステムの信頼性の向上と並列分散プログラムの性能向上に関する検討を行った.

信頼性の向上に関しては，マイクロカーネル構成の既存のオープンな UNIX OS の拡張性や保守性を維持したまま，小規模な修正でプロセスペア方式の耐故障性を付与することが可能な故障管理機構を提案し，耐故障ファイルシステムを試作して有効性を実証し評価を行った.

また，性能向上に関しては，IA アーキテクチャの CPU と Windows NT に適した Virtual Interface Architecture (VIA)を中立化する仕様拡張を提案するとともに，大規模な商用分散並列プログラムで利用可能な実用的なVIA を実現する手法について検討し，SPARC UNIX 上でシステムを構築し提案の有効性を確認した.

論文審査の結果の要旨

インターネット上でサービスを提供する情報システムは，従来よりも高性能で高い信頼性が要求されている．情報システムの性能がスケーラブルに向上し，高い信頼性を実現する手法として，複数のコンピュータを結合したクラスタシステムが重要となっている．そこで著者は，コンピュータ間を高速通信機構で結合したクラスタシステムにおける，信頼性と実行性能の向上に関する詳細な研究を行った．本論文はその成果をまとめたものであり，全編５章からなる．

第１章は序論である．

第２章では，クラスタシステムとコンピュータ間を結合する高速通信機構を概観している．特に高速通信機構の標準として提案されている VIA(Virtual Interface Architecture) について述べている．

第３章では，マイクロカーネル構成に基づく UNIX の各プロセスを二重化する耐故障化法を提案している．二重化に必要となる故障管理方式を，プロセスから独立して実現することにより，既存プログラムの修正量を最小限にとどめ，機能の拡張性および保守性を保存することが可能となることを明らかにした．更に，これの実装法として，故障管理機構，耐故障通信機構，データ安定格納機構を開発した．これらを具体的にファイルシステムの構築に適用し，信頼性の向上を確認している．これはオープンシステムの信頼性の向上に関する有用な成果である．

第４章では，高速通信機構の標準として提案されている VIA の現在の規格を拡張し，特定のアーキテクチャおよび OS に依存しない VIA の規格を提案している．具体的には，同期と非同期事象を効果的に処理する機構を開発し，これに基づいた規格を考案している．更に，提案規格の VIA を UNIX 上で実現し，大規模並列分散プログラムでの利用を通じて提案規格の有効性を示している．これはクラスタによるシステムの性能向上に寄与する実用上重要な成果である．

第５章は結論である．

以上要するに本論文は，UNIX の高信頼化と新しい VIA 規格の研究を行い，高速通信機構を備えた並列計算機の信頼性および処理性能に関する有用な基礎を与えたものであり，情報基礎科学の発展に寄与するところが少なくない．

よって，本論文は博士（情報科学）の学位論文として合格と認める．