〈ハードウェア〉

Express5800/ftサーバ

Express5800/ftc Server

岡田政彦* 森川 誠**
Masahiko Okada Makoto Morikawa

要 旨

情報機器に対する高い信頼性の要求に応えるため，NECはStratus Technologies社との技術提携により，2001年6月にIAベースのフォールトトレラントサーバ「Express5800/ftサーバ」を製品化して以来，ftサーバの製品強化およびラインナップの拡充に取り組んでいます。ここでは，Express5800/ftサーバの狙いや特長の説明に加え，最新のLinuxモデルを例に挙げて，ftサーバを実現するための重要なテクノロジについて紹介します。

In order to bring secure computing to the business, NEC has developed a fault tolerant computer “Express 5800/ftc Server” which is based on IA server architecture in cooperation with Stratus Technologies, Ltd. and has been carrying on enhancement of the product and expanding lineup. This paper describes the target and features of the Express5800/ftc server and key technologies to implement ftc Server taking brand new Linux model as a example.

1．まえがき

IT技術の驚異的な進歩は，われわれの生活やビジネス基盤にも大きな変化をもたらしてきました。情報機器の社会インフラへの浸透が進むにつれ，これまで以上に高い信頼性を求められてきています。この高い要求に応えるため，NECはStratus Technologies社との技術提携により，2001年6月にIA32（Intel Architecture）ベースのフォールトトレラントサーバ「Express5800/ftサーバ」を製品化して以来，精力的にftサーバ製品の強化およびラインナップの拡充に取り組んでいます。

また，近年の官公庁や自治体を中心としたLinuxや高可用性へのニーズの高まりに応え，Linuxを搭載したftサーバを世界に先駆けて2002年7月に製品化し，2004年10月には，Linuxモデルの大幅な強化を行い，**図1**に示すような充実したラインナップを揃え，多様な要求に応えています。

2．ftサーバの狙いと特長

システムの堅牢性，可用性を高める手段として，これまでにも専用設計のフォールトトレラントシステムやクラスタシステムが採用されてきましたが，これらのシステムは，非常に高価なことや，運用に特別な技術を要するなど，やや敷居の高いシステムでした。ftサーバは，これらの課題を克服し，より広く高可用性のメリットを享受し，活用していただけるよう汎用のソフトウェアやオプション製品を利用できることを前提に開発されました。ここでは，ftサーバの特長と狙いについて概説します。

(1) 止まらないシステム

情報機器が業務の基盤として活用されている昨今，システムダウンによる影響は計り知れず，ダウン時間の短縮が重要な命題となっています。この要求に応えるため，ftサーバは**図2**のとおりCPU，メモリ，ディスク，電源などの大部分の構成要素を二重化し，仮にいずれかの部品が故障した場合にも，システムの稼働を継続できるようにしています。

同じ二重化方式としてクラスタシステムがありますが，ソフトウェアによって障害の検出と他系への切り替え，デ

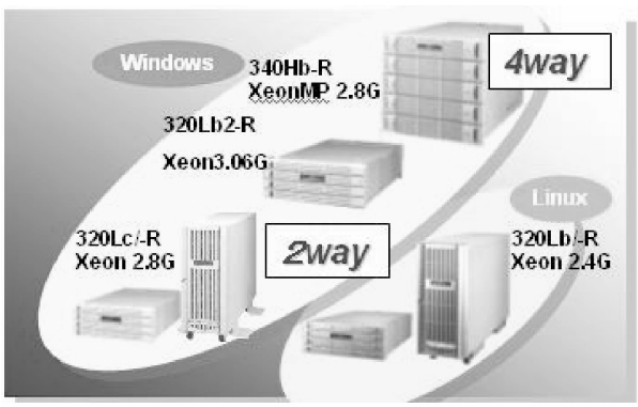

図1 ftサーバの製品ラインナップ
Fig.1 ftc Server Product lineup.

* クライアント・サーバ事業部
Client And Server Division

** 第二コンピュータソフトウェア事業部
2nd Computers Software Division

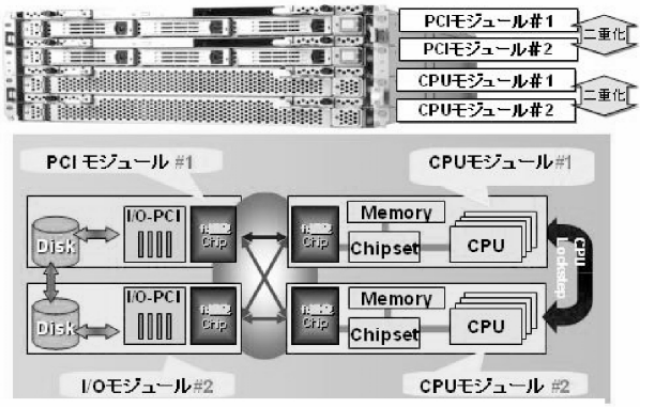

図2　ftサーバのアーキテクチャ
Fig.2　ftc Server architecture.

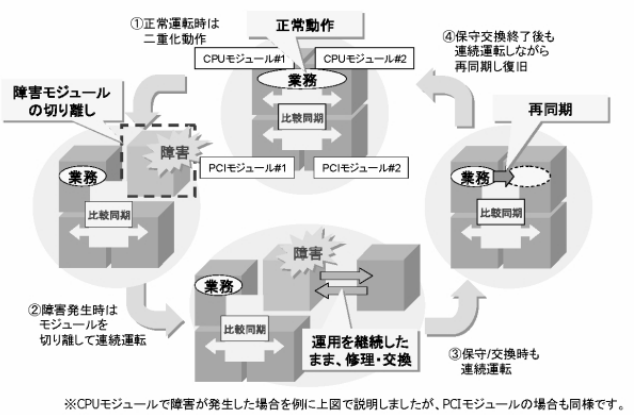

図3　故障時の動作
Fig.3　Behavior in case of failure.

ータの復旧と，非常に複雑な手順と長い時間を必要とします。これに対し，ftサーバは瞬時に切り替えが行われ性能も変わらないため，ほとんどのお客様は故障したことすら気づかずに運用できます。この点が，ftサーバの特に優れている点だといえます。

(2) 無停止保守

ftサーバのもう1つの大きな特長は，システムを稼働したまま，部品を交換修理できることです。これをExpress5800/320Lc（**写真**）を例に挙げて説明します。このシステムは，主にCPUモジュールとPCIモジュールの2種類のモジュールから構成され，さらに各モジュールを2つずつ備えた二重構成となっています。各モジュールは1Uサイズのモジュールであり，引き出しを出し入れする要領で，簡単にラックに出し入れすることができます。通常，これらの作業は保守員が行いますが，お客様でも簡単に交換できるほど簡単なものになっています。

図3は，CPUモジュールが故障した場合の動作と修理までの仕組みを概念的に示したものです。いずれかのCPUモジュールが故障した場合，そのCPUモジュールは瞬時に切り離され，残りの正常なCPUモジュールのみで継続して動作します。さらに，システムを稼働したままCPUモジュールを交換後，自動的に二重化状態に復旧します。

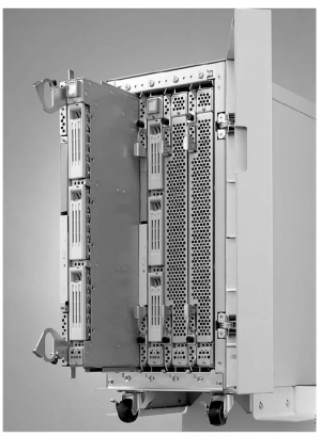

写真　モジュール構造
Photo　Module structure.

(3) トータルコスト

ftサーバは，WindowsやLinuxなどの汎用のオペレーティングシステムが動作し，汎用のアプリケーションや業務パッケージを，一切の変更を加えずに使用することができます。クラスタシステムでは，利用できるソフトウェアがクラスタ対応の製品に限定されたり，フェールオーバ時の手続きを用意する必要がありますが，ftサーバではこのような制限を受けることはありません。

情報システムにも投資効果を求める昨今，コストの削減や開発期間の短縮が重要なテーマになっています。既存の汎用製品を利用できることのメリットは，システム全体の開発期間の短縮や開発コストの削減が図れることにあります。運用や保守に際しても，クラスタシステムを運用できるほどの高いスキルを備えたSEを必要としないため，中小規模の企業や自治体のように人材のリソースが限られているところにも容易に導入できます。

3．ftサーバを支える技術

(1) CPUの冗長化とロックステップ

すでに多くのIAサーバで，ストレージ，LAN，電源などの冗長化が図られていますが，CPUや主要なチップセットの冗長化は，今なお技術的にもコスト的にも難しいものとされています。

コンポーネントを冗長化する手段には大きく2つの方法があります。1つ目は2つのコンポーネントを備え，1つを待機させておき障害時に切り替える方法です。2つ目は複数のコンポーネントを同期して実行し，障害時に一方を切り離す方法です。ftサーバでは，2つのCPUを同期して実行する方法を採用しています。同期の精度は，クロックレベルでの完全な同期になります。これをロックステップと称しています。

(2) I/Oデバイスの仮想化

CPUモジュールとは対照的に，I/Oデバイスについては

片方のデバイスを待機させておき，障害発生時に切り替える方式を採用しています。これは，ハードウェアとデバイスドライバの連携により，2つのデバイスから仮想的な単体のデバイスを構築することで実現しています。故障した場合や，ボードが突然抜かれた場合，電源が切れた場合には，ハードウェアとデバイスドライバの連携により，動作するコンポーネントを切り替えます。この一連のフェールオーバの手続きは，アプリケーションに透過的に，デバイスドライバの範囲内で行われるため，アプリケーションが意識することはありません。これによって，既存のアプリケーションに手を加えず利用することが可能となります。

(3) ft機能チップセット

ftサーバを構成する上で，最も重要なチップセットについて説明します。ftサーバは2組のCPUモジュールとPCIモジュールから構成されており，各モジュールがクロスバーリンクで相互接続されています。クロスバーリンクの制御は，すべてft機能チップセットによって行われます。ftサーバのクロスバーリンクは，1対1の接続に限らず，1対多の接続をサポートしています。たとえば，1つのCPUモジュールと2つのPCIモジュールを接続することもできれば，2つのCPUモジュールと1つのPCIモジュールを接続することもできます。この仕掛けによって，CPUモジュールとPCIモジュールを柔軟に組み合わせ，動的に二重化や縮退の制御を行うことができるようになります。

もう1つの重要な機能は，CPUモジュールが同期して動いていることを，クロックステップごとに常に監視することです。いずれかのCPUモジュールに故障が発生すると，同期ズレとして検出され，故障した側のCPUモジュールが切り離されます。一度切り離されたCPUモジュールから，他のPCIモジュールへのアクセスはftサーバチップセットによって一切禁止され，システムへの悪影響を抑止します。

(4) モジュール化とホットスワップ

ftサーバは，無停止保守を原則としています。たとえ二重化していても，修理するためにシステムを停止する必要があれば，全体としての稼働率は向上しません。一般のサーバでも，DISK，電源，PCIカードなどのホットスワップ機能を備えた製品が増えてきましたが，ftサーバはこれをさらに一歩進め，ほとんどの構成要素をモジュール化し，ホットスワップできるようにしました。

CPUモジュールのホットスワップを実現するためには，動的な二重化や縮退を実現する必要があります。つまり，1つのCPUモジュール上でソフトウェアが稼働している状態を維持したまま，もう1つのCPUモジュールに稼働中のCPUモジュールの状態を完全にコピーするということです。ここでの課題は，CPUやチップセットにはソフトウェアや外部回路から直接操作できないレジスタなどが含まれていることです。これらは，間接的な操作や擬似的な事象を発生させるなどの様々な手段によって一致を図っています。また，メモリ容量が大きくなるほど，コピーに要する時間が長くなることから，この影響を極小化するために，システムを止めずにバックグラウンドでコピーする仕組みを実装しました。システム稼働中にはコピー元のメモリの状態も刻々と変わっていきますが，その状態を逐次把握しコピーする高度な制御を行っています。

PCIモジュールのホットスワップは，既存技術であるディスクやPCIカードのホットスワップ機能の範囲を，PCIブリッジ全体まで広げることにより実現しています。既存の汎用OSではサポートされていない機能ですが，ハードウェアとデバイスドライバによる各デバイスやPCIブリッジ全体の仮想化によって，これを実現しています。

(5) ストレージの二重化

ftサーバの内蔵ストレージは，ソフトウェアミラーリングにより二重化しています。一般のサーバで用いられるハードウェアRAIDも，低コストでDISKを冗長化できる良い方法ですが，アダプタの故障やバスの擾乱に対しては無力です。ハイエンドのストレージ製品の中には，アダプタやパスが二重化された製品がありますが，低価格なサーバ製品で利用するには，まだまだ高価です。このような現状のなかで，コスト，性能，実現性，そして可用性をハイレベルでバランスした方式として，ソフトミラーの方式を選択しています。将来的には，性能を引き出せるハードウェアRAIDの実現が望まれます。

ハードディスクのミラーリングには比較的長い時間がかかりますが，ソフトウェアの障害やPCIモジュールの障害により一時的に同期が外れた場合は，差分データのみを再同期することで早期に二重化状態に復旧できるよう考慮されています（**図4**）。

(6) ネットワークの二重化

ネットワークの二重化は，すでに一般のサーバでも利用されている二重化の技術を活用して実現しています。ネットワークの冗長化の方法には，ホットスタンバイ型とロードバランス型の2つがありますが，可用性という点においては，いずれの方式にも差がありません。ftサーバでは，通常は片側のアダプタだけで送受信を行うスタンバイ型を採

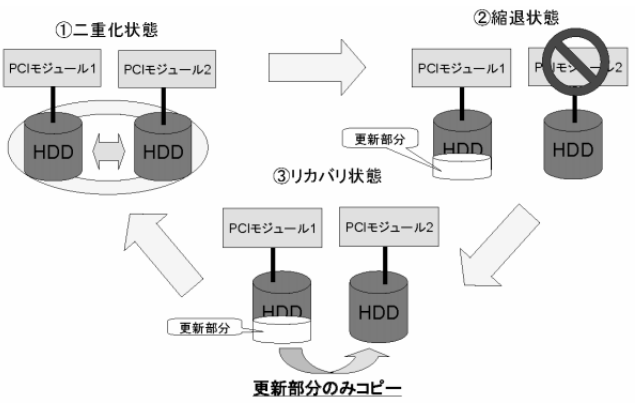

図4　HDDの高速再同期の仕組み
Fig.4　HDD fast re-syncing mechanism.

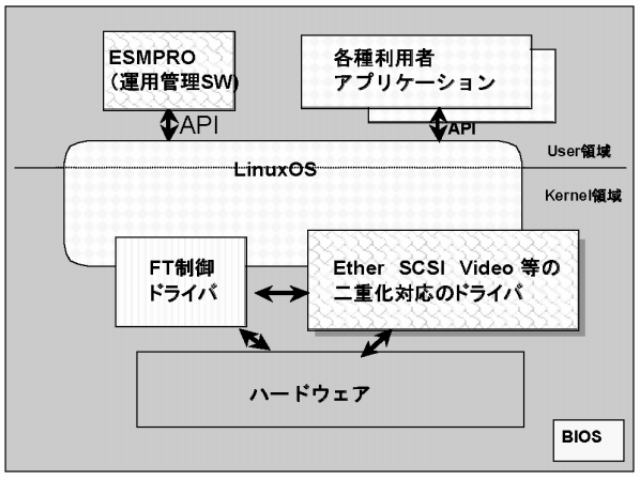

図5　ソフトウェアアーキテクチャ
Fig.5　Software architecture.

用しています。このアダプタのリンクが何らかの原因（ポートの障害，ケーブル切断，アダプタの故障など）でダウンした際，残りのアダプタが自動的にリンクを引き継ぎLANの接続を維持します。

（7）ソフトウェアアーキテクチャ

ftサーバのソフトウェアは，**図5**に示すようにシステム全体の二重化や縮退などを制御するft制御ドライバと各デバイスを二重化するデバイスドライバから構成されます。ft制御ドライバは，これらの制御を行うため高い特権レベルで動作するデバイスドライバとして実装されています。

ftサーバはシステム起動時には単体のサーバとして動作しますが，その後，ft制御ドライバにシステム制御の主権が移り，ft機能チップセットや仮想デバイスドライバを制御しながらシステム全体の二重化や縮退を制御します。各デバイスドライバは，両系のデバイスを二重化し，一方でアプリケーションに対しては仮想的な1つのデバイスとして振舞うように拡張しています。さらに，デバイスの動的な二重化と切り離しを行うための機能も付加しています。

このように，ft制御ドライバと，各デバイスドライバの拡張によって，すべての二重化の仕組みを吸収することで，一般のアプリケーションを変更せずに動作できるようにしています。

4．ftサーバLinuxモデル

ここでは，2004年10月に以下のように前機種から大幅に強化した最新のExpress5800/320Lb Linuxモデルを紹介します。強化内容としては，

・CPUをPentium Ⅲ→Xeon 2.4GHzに大幅強化
・最大メモリ容量を2GB→3GBに強化
・大幅な省スペース化（ラック8U→4Uサイズ）

となっています。

（1）Linuxモデルの特長

信頼性の高いLinux OSを採用するとともに，以下のようなExpress5800/ftサーバ特有の機能で高い可用性を実現しています。

1）Kernel2.4ベース

Kernel2.4をベースにExpress5800/ft サーバ専用のドライバを提供することによりフォールト・トレラントを実現しました。これらは，OSのAPIには影響を与えないため，OSS（Apache, sendmailなど）アプリケーションがそのまま動作可能です。

2）デバイスドライバ

Express5800/ftサーバで利用するデバイスドライバは二重化に対応した専用ドライバを提供します。専用ドライバによりPCIボードなどに冗長性を持たせることが可能です。

3）デバイスドライバのハードニング

デバイスドライバにて停止としていた処理も可能な限りフェイルオーバー処理で救済します。ハードウェアに最適化されたより強固なOSとして提供しています。

4）メモリダンプ機能の強化

ハードウェアまたはソフトウェアの要因で万一システムが停止した場合でも，片系のみダンプを採取し，もう片系ではダンプ採取処理を省略しリブートします。これにより再起動時間を短くすることが可能です。

（2）ディストリビューション

Linuxには，Redhat，Suseなどの様々なディストリビューションが存在します。ftサーバでもこれらのディストリビューションを活用することが望ましいのですが，非常に特殊な二重化機構を備えたドライバをディストリビューションに組み込むことや，製品の認証を受けるには多くの課題があります。

さらに，組み込んだソフトウェアをGPLのライセンスに基づいて公開する義務が生じてしまうことも課題です。このような環境のなか，早期製品化とIP保護を前提に，ディストリビューションを受けない独自のLinuxとしての製品化をめざしました。

（3）OSS（アプリケーション）

Express5800/ftサーバは，フォールト・トレラント機能をデバイスドライバのレベルで実現しているので，利用者自身がハードウェアの二重化構造を意識する必要はなく，Linux（Kernel2.4）対応のアプリケーションに手を加えずにそのまま利用できます。

これにより，既存の多様なOSSを自由に活用することができます。

5．むすび

以上，述べてきましたように，Express5800/ftサーバは極めて高い可用性を備えるとともに，価格，操作性，保守性をも高次元でバランスさせた優れた製品です。

今後も可用性とLinuxに対する市場ニーズのいっそうの高まりに応えられるよう，NECはftサーバの継続的な製品強化に取り組んでいきます。

筆者紹介

Masahiko Okada
岡田（おかだ）　政彦（まさひこ） 1987年，NEC茨城入社。現在，NEC 第二コンピュータ事業本部クライアント・サーバ事業部第一技術部技術エキスパート。

Makoto Morikawa
森川（もりかわ）　誠（まこと） 1987年，NEC入社。現在，コンピュータソフトウェア事業本部第二コンピュータソフトウェア事業部エキスパート。