WHITE PAPER

Express5800/ft サーバ
White Paper

ご注意

本書の内容は、予告なしに変更されることがあります。

日本電気株式会社は、本書の技術的もしくは編集上の間違い、欠落について、一切責任をおいません。また、お客様が期待される効果を得るために、本書に従った導入、使用および使用効果につきましては、お客様の責任とさせていただきます。

日本電気株式会社は、NEC Express サーバシリーズ製品保証書で保証する内容以外には、一切の保証は致しません。

Copyright NEC Corporation 2001

- Microsoft、Windows、Windows NT は米国 Microsoft Corporation の米国およびその他の国における登録商標または商標です。
- Intel はアメリカ合衆国および他の国におけるインテルコーポレーションおよび子会社の登録商標または商標です。
- ESMPRO,CLUSTERPRO は日本電気(株)の登録商標です。
- その他、記載の会社名および製品名は各社の登録商標または商標です。

目次

IT をとりまくビジネス環境

近年、IT を取巻くビジネス環境には 2 つの大きな流れがあります。

- インターネットの拡大によるビジネスのグローバル化。
- ビジネス クリティカルなアプリケーションの IA サーバへの移行。

この 2 つの流れに共通しているシステムへの要求は、ダウンタイムを最小限に抑えるあるいは完全に防止するという高い可用性の実現にあります。また、IA サーバの特徴である安価で高性能なシステムを構築できるというポイントも損なってはなりません。

NEC では、この要求に答えるため、米国 Stratus Technologies 社と共同で IA サーバをベースとしたフォールト・トレラントサーバ「Express5800/ft サーバ」を開発しました。Express5800/ft サーバは、主要コンポーネントの二重化を行っています。それぞれのコンポーネントで単一故障が発生してもシステムの連続運用が可能で、障害コンポーネントの交換もシステムを止めることなく行えるように設計されています。このように Express5800/ft サーバはハードウェア障害によるビジネスへの損害を最小限に抑えるためのベスト・ソリューションです。

NEC の高可用性システム戦略

一般的にシステムの可用性を高めるためには、通常サーバでは Disk Array や冗長電源の使用があります。さらなる高可用性へのアプローチには、ハードウェアの主要モジュールを多重化するアプローチであるフォールト・トレラント・システムとソフトウェアによるコンピュータの多重化するアプローチであるクラスタシステムの 2 つがあります。

フォールト・トレラント・システムでは、ハードウェアで二重化を実現しているので、障害発生時の切り替えは瞬時に行われます。しかし、ソフトウェアの障害によるシステムのダウン（OS パニック、アプリケーション・エラー）には対応出来ません。通常のサーバと同様にリソース管理ソフトなどで、監視する必要があります。

一方、クラスタシステムは OS パニックやサービス、プロセスの消失を監視して、これらソフトウエァの障害時にも待機ノード（クラスタを構成するコンピュータ）への切り替えが可能です。システムの拡張もクラスタを構成する個々のコンピュータに対して機器を増設するほかに、クラスタを構成するコンピュータ自体を追加するという拡張ができます。したがって、クラスタシステムの方がスケーラビリティの富んでいるといえます。しかし、処理を別のノードへ引き継ぐために、数十秒から数分のシステムの停止時間が必要です。また、クラスタソフトによっては処理を引き継ぐために切り替えスクリプトの作成や専用 API を使用しなければならなかったり、クライアント側のアプリケーションで処理の状況の保持などをしなければならない場合があります。

高可用性システムの構築は、そのシステムの許容ダウンタイム、障害監視対象、拡張性、導入費用、TCO など様々な要件と照らし合わせて、フォールト・トレラント、クラスタのどちらかを選択することが重要です。

NECでは、これら両方のアプローチに対するソリューションをご提供いたします。フォールト・トレラント・システムの製品としては「Express5800/ft サーバ」、クラスタリング・ソフトとしては「CLUSTERPRO」と「Microsoft Cluster Server」を商品化して高可用性の幅広い要求に

対応します。

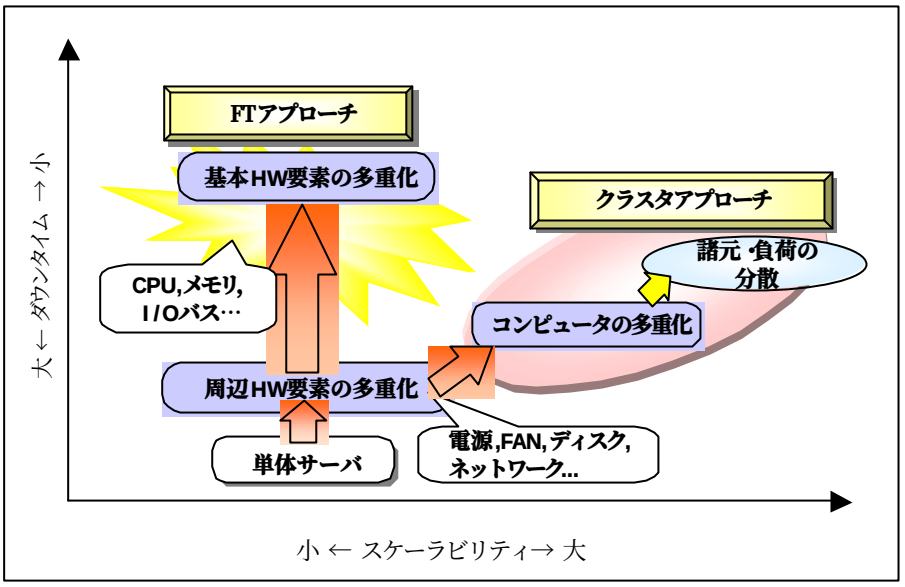

図 1 可用性向上のコンセプト

Express5800/ft サーバ ・システム概要

Express5800/ft サーバはシステムの稼働に必要なコンポーネントを二重化して耐障害性を向上させた Intel Architecture(IA)の フォールト・トレラントサーバです。

OS には IA サーバで標準的なオペレーティングシステム Microsoft Windows2000 に対応しています。Windows2000 に対して HAL（Hardware Abstraction Layer）、デバイスドライバ、ft制御ドライバによってハードウェアの違いを吸収していますから、アプリケーションは二重化を意識することなく、動作することが可能です。

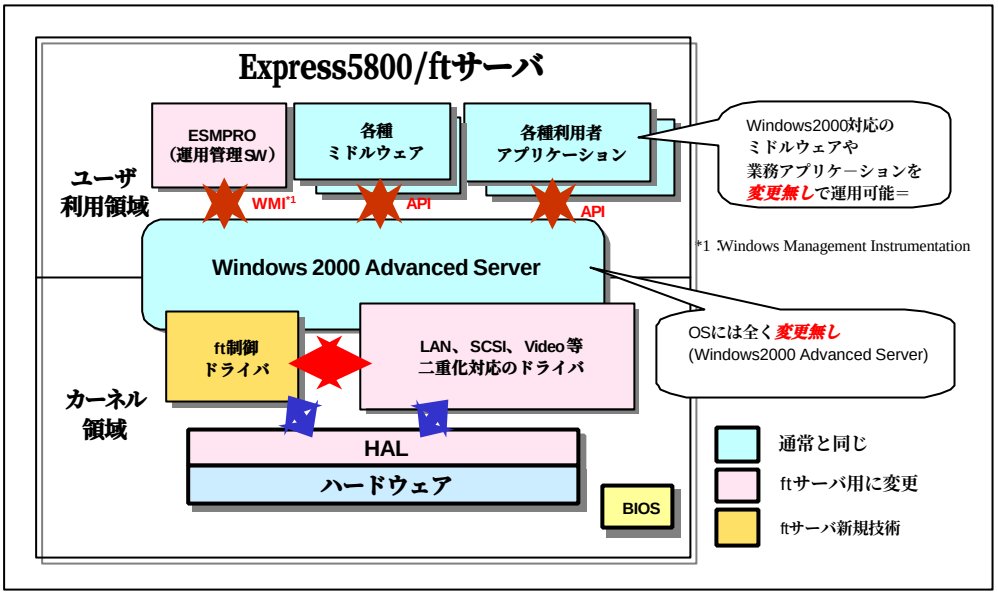

図 2 ソフトウェア・アーキテクチャ

Express5800/ft サーバでは耐障害性を向上させるため、システムの稼働に必要なコンポーネントを二重化しています。個々のコンポーネントはモジュールと呼ばれる単位にまとめられており、CPU モジュール、PCI モジュール、電源モジュール、ハードディスクで構成されます。

障害発生時には、障害が発生したコンポーネントを含むモジュールを切り離すことで運用を継続します。

また、各モジュールはホットスワップが可能で、交換する場合にはシステムを停止させる必要がありません。

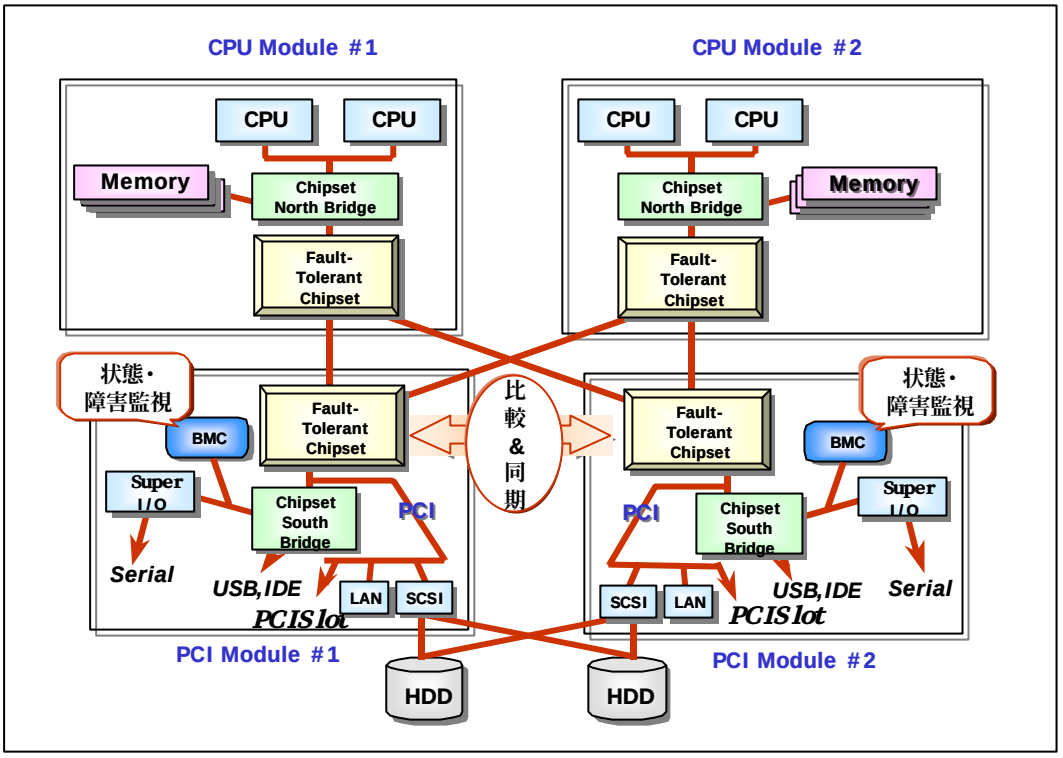

図 3 ハードウェア・アーキテクチャ

その他、次のようなダウンタイムを低減する工夫がほどこされています。

クイックダンプ機能

OS ストップが発生した場合、一方の CPU モジュールはメモリデータを保持したまま切り離されます。もう一方は通常と同じように再起動を行い、OS 起動後、切り離されたＣPUモジュールから障害の発生したメモリの内容をダンプファイルへコピーします。コピーが終了するとその CPU モジュールをリセットしてシステムに組み込み二重化された状態に戻します。この機能によりシステム回復時間を短縮しかつ確実にメモリデータを保存できます。

[参考]通常のメモリダンプ

OS ストップ時、ページファイルにメモリデータをコピーします。OS のリブートをおこないます。ＯＳ再起動後、ダンプファイルをページファイルに保存されたメモリデータから作成します。

Hardened 化したデバイスドライバ

Wdindows2000 における OS ストップの原因の大半はデバイスドライバの不具合によって引き起こされます。この問題は、雑多なデバイスが増えれば増えるほど発生しやすくなります。Express5800/ft サーバでは、正常系動作及び異常系動作の十分なテストを実施

し、二重化対応を行ったドライバのみをサポートしています。これにより、デバイスドライバが原因で起こる OS ストップによるシステム停止を最小限にとどめます。

高可用性を維持するハードウェアの仕組み

Express5800/ft サーバでは多重化されたコンポーネントの制御や監視をフォールト・トレラント・チップセットと ft 制御ドライバによって行っています。障害は個々のコンポーネント単位に検知しますが、切り離しはモジュール単位に次の順序で行われます。

まず、障害が発生するとモジュールを切り離し、再組み込みを試みます。

再組み込みを複数回繰り返しても動作が回復しない場合は故障と判断して、モジュールを切り離します。モジュールの切り離しは瞬時に行われ、アプリケーションから意識することはありません。また、縮退運転時も性能が劣化しません。（ハードディスクの復旧時には劣化する場合があります。）

では、各モジュールの機能と動作についてさらに詳しく説明します。

CPU モジュール

CPU モジュールには CPU とメモリ、チップセット(ノースブリッジ)、BIOS プログラム、ファン、ft 機能チップセットが含まれます。

ft 機能チップセットは CPU モジュール同士をクロック単位で同期して命令実行させ、読み書きしたメモリデータの比較を行います。このようにそれぞれのモジュールが全く同じ動作をするように制御しているので、二重化されている CPU やメモリが OS からは１つであるようにみえます。

[障害発生時、復旧時の動作]

障害時にはその種類により2種類の動作を行います。

まず、CPUモジュール内のコンポーネント（CPU、メモリ）の障害をExpress5800/ftサーバの自己監視機能にて発見した場合、Express5800/ftサーバは障害が発生したコンポーネントを含むCPUモジュールを切り離します。

次にデータを入出力した結果が両CPUモジュールで異なった場合には、各モジュールの動作時間を比較し、動作時間の短いCPUモジュールを信頼性が低いと判断して切り離しを行います。

復旧時には、障害の起きた CPU モジュールを正常なモジュールと交換すると自動的に同期動作を行い復旧します。同期動作のために、稼動中の CPU モジュールのメモリデータを復帰させる CPU モジュールのメモリにコピーします。この間(数秒～数十秒)は CPU が完全に停止するため、クライアント側の TCP/IP のタイムアウト値の変更が必要な場合があります。

PCI モジュール

PCI モジュールにはチップセット(サウスブリッジ)、LAN コントローラ、SCSI コントローラ、GA(Graphic Accelerator)、BIOS 設定保存メモリ、BMC (Baseboard Management Controller)、ft 機能チップセットが含まれます。

PCI モジュールにはプライマリとセカンダリという実行優先順位があり、プライマリPCI モジュールのデバイスのみ稼働してセカンダリは障害に備えて待機し、障害発生時に切り替わります。また、CPU モジュールとは異なりそれぞれのコンポーネントが OS からは個々のデバイスとして認識されます。それらのコンポーネントは hardened ドライバを使用することにより、障害時のコンポーネントの切り替えを実現しています。

[障害発生時、復旧時の動作]

障害発生時、プライマリPCI モジュールのコンポーネントに障害が発生した場合は、モジュールを切り離し、セカンダリPCI モジュールをプライマリに切り替えます。このとき CPU モジュールは同期動作を維持しています。PCI モジュールを切り替える時、GA も切り替えるので一瞬画面が消えますが、正常な動作です。セカンダリPCI モジュールのコンポーネントに障害が発生した場合は、モジュールを切り離し、稼動をつづけます。復旧時には故障したモジュールを修理・交換すると自動的にシステムに組み込まれ、セカンダリとして待機します。

但し、増設した PCI カードに障害が発生した場合は対象となるカードのみをバスから切り離します。これらの PCI カードを修理・交換する場合には、手動でモジュールの切り離しを行います。

SCSI コントローラ・ハードディスク

ハードディスクは Windows2000 で標準サポートされているミラーボリュームを使用してデータを二重化しています。障害検知もOS が行います。

ディスクベイに搭載するハードディスクはモジュール単位に異なるSCSI バスへ接続されており、ミラーボリュームのペアが同一の SCSI バスを使用しないように設計されています。片方のSCSI バスに障害が発生してもディスクの読み書きが可能です。

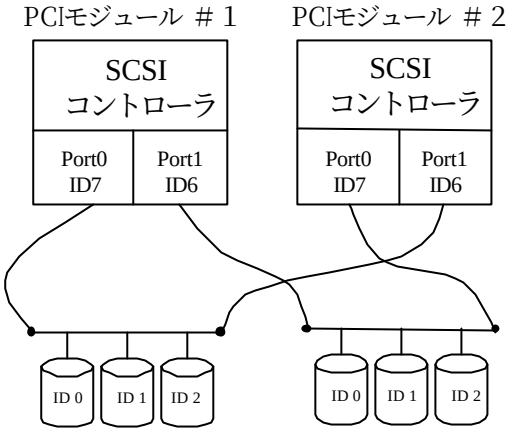

図 4 SCSI 接続イメージ

また、それぞれの SCSI バスは両方の SCSI コントローラへ接続されていますが ft 機能チップセットからの情報を元にどちらの SCSI コントローラを使用するかドライバが選択します。このようにして、2 つのコントローラから同時に同じハードディスクへ書き込むことを防止しています。
プライマリモジュール上のコントローラに障害が発生したときは ft 機能チップセットが検知してもう一方のコントローラへハードディスクの制御を移し、ミラーの状態を保持したまま稼働を続けます。
ところで、起動時にはプライマリの PCI モジュールに近い SCSI バス(#1 の時はスロット1～3、#2 の時はスロット4～6)が最初に認識されます。
プライマリが切り替わると OS のディスク管理ツールで表示順番が変わることに注意する必要があります。

LAN コントローラ

標準 LAN コントローラはそれぞれの PCI モジュールに実装されており、AFT (Adapter Fault Tolerance)により、それらの二重化を実現しています。AFT とは、Intel 社製 LAN コントローラの拡張機能です。複数の LAN コントローラを１つのグループ(この動作をチーミング:teaming と呼ぶ)にまとめ、１つの仮想的な LAN コントローラを構成します。この仮想的な LAN コントローラに対し OS は通信を行いますので、グループ内の１つのコントローラに障害が発生しても通信をつづけることができます。

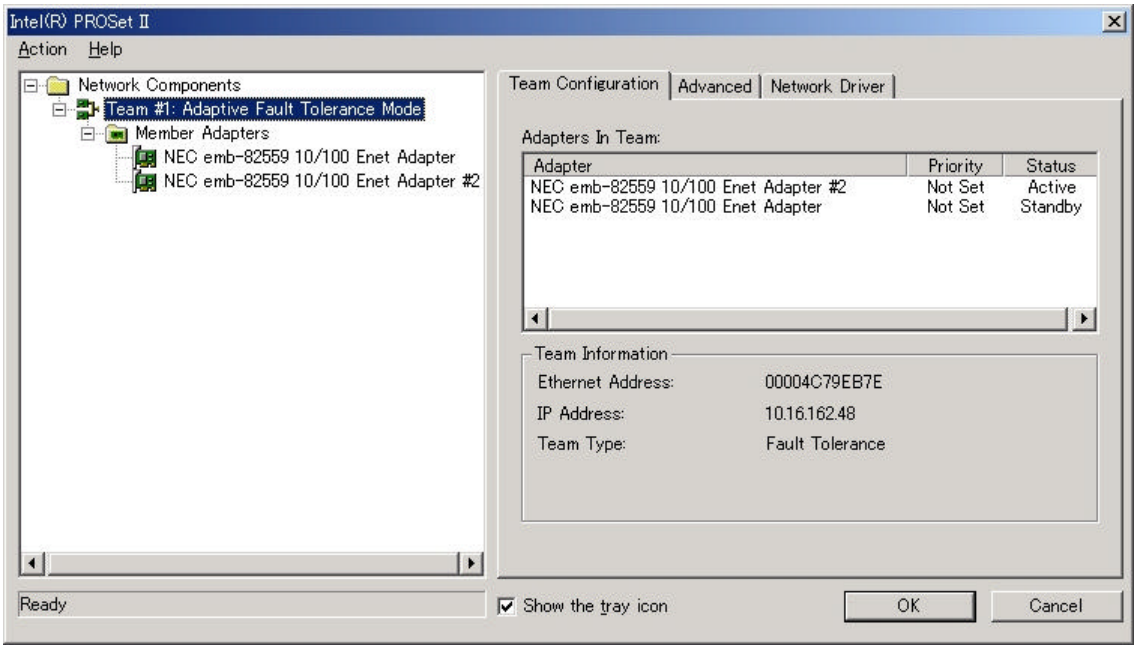

図 5 AFT 設定画面

通常時は、プライマリで動作するPCI モジュール上の LAN コントローラがプライマリポートとして通信を行っていますが、ドライバが障害を検知するとセカンダリポートへ通信を引き継ぎます。このとき、リンクも引き継ぎますので、コンピュータ間の接続がとぎれることはありません。

この機能は、Express5800/100 シリーズの通常のサーバでも、採用されています。(ただし、LAN コントローラの増設が必要)通常のサーバとの違いは、オンボードLAN コントローラはコントローラの情報（MACアドレスなど）をコントローラ上のLSI に格納するのではなく、**図 6 フロントパネルボード**にて四角で囲った筐体裏にあるフロントパネルボードに格納しています。したがって、PCI モジュールが障害などで新しいものに変更されても、LAN コントローラのMACアドレスが変わることはありません。

尚、増設のLAN ボード同士で二重化を構成した場合はExpress5800/100 シリーズの通常サーバと同様、LAN ボード側で個別に Mac アドレス等の情報を持ちます。

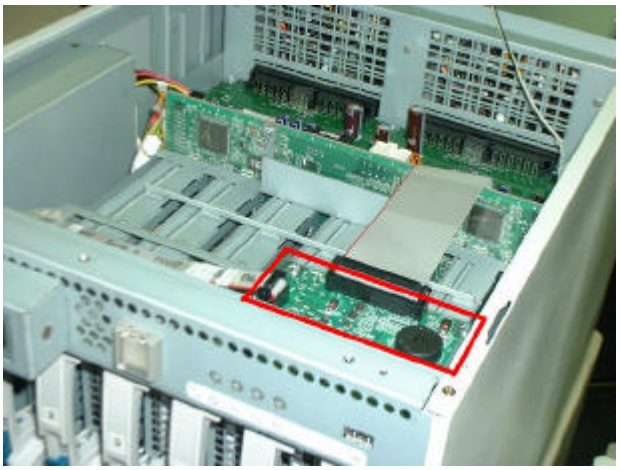

図 6 フロントパネルボード

電源モジュール

Express5800/ftサーバは、電源モジュールについても冗長構成をとっています。通常２台の電源モジュールで電力を供給していますが、一方のモジュールの故障やAC供給停止といった障害が発生すると、もう一方の電源モジュールだけで引き続き電力を供給します。障害からの復旧もホットスワップ可能なモジュールを使用していますので障害の起きた電源モジュールはシステムの運用を継続しながら、取り外し修理・交換が可能です。なお、電源モジュールの冗長化については、通常のExpressサーバも同じ技術で実現しています。

高可用性を維持するソフトウェア

Windows2000 Advanced Server

Windows2000 ファミリーは WindowsNT4.0 の後継に当たる製品で、完全なマルチタスクを実現する堅牢な NT カーネルをベースにさらなる安定性を追求したオペレーティングシステムです。再起動無しに IP アドレスなどの各種設定変更を可能にするなど、システムの可用性を向上させる工夫も行われています。

このほかに可用性を向上させる以下のような特長があります。

・プラグアンドプレイ (Plug and Play)
Windows9X では既にサポートされていましたが NT カーネルでは初めてサポートしました。
これにより再起動無しにハードウェアを追加・削除することが可能です。
この機能が PCI モジュールの二重化実装に必要不可欠であることは周知の通りです。

・システムファイルプロテクション(System File Protection)
以前の Windows ではアプリケーションと共にインストールした共有ライブラリ(DLL)が原因でシ

ステムが不安定になったり既にインストールされていたアプリケーションが動作しなくなるという問題がありました。
システム ファイル プロテクションは、保護されたシステム ファイルが移動、削除、不正なファイルで上書き等が行われると自動的に修復してシステムは安定した状態を保ち続けます。

・カーネルモードライトプロテクション (Kernel-mode write protection)
Windows2000 はより信頼性のあるプラットファームを実現する機能としてカーネルモードライトプロテクションを実装しています。Windows2000 メモリマネージャは不正なコードに対してカーネルがロードされているメモリ空間を読みとり専用にすることでドライバやカーネルモードで動作しているアプリケーションがシステム領域を破壊して停止、いわゆる「ブルースクリーン」の発生を防ぎます。

ft サーバユーティリティ

ローカルにメンテナンス操作を行うためのユーティリティです。自動インストールを実行すると ESMPRO/SA ともにインストールされます。モジュールの切り離しや自己診断、組み込みなど、ハードウェア・メンテナンスの他に、システム運用中のファームウェア(システム BIOS、BMC など)のアップデート、メモリダンプの採取をこのユーティリティから行います。

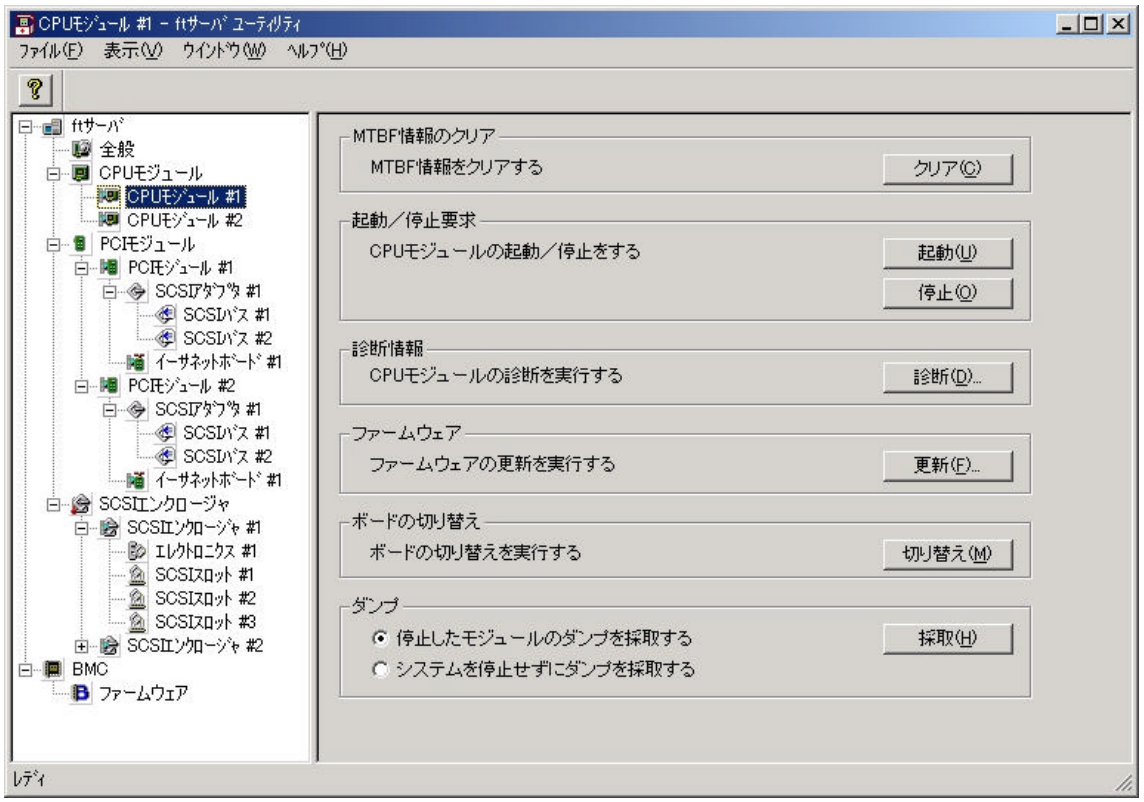

図 7 ft サーバユーティリティ

ESMPRO/Server Manager

Express サーバの運用管理ツールとして標準添付されているソフトウェアです。Express5800/ft サーバだけではなく、ネットワーク上の Express サーバの一元管理が行えます。

Express5800/ft サーバ用の拡張機能として、ft サーバユーティリティで実行できるモジュールの切り離し組み込み、ファームウェアのアップデートなどをリモートの管理端末から可能です。

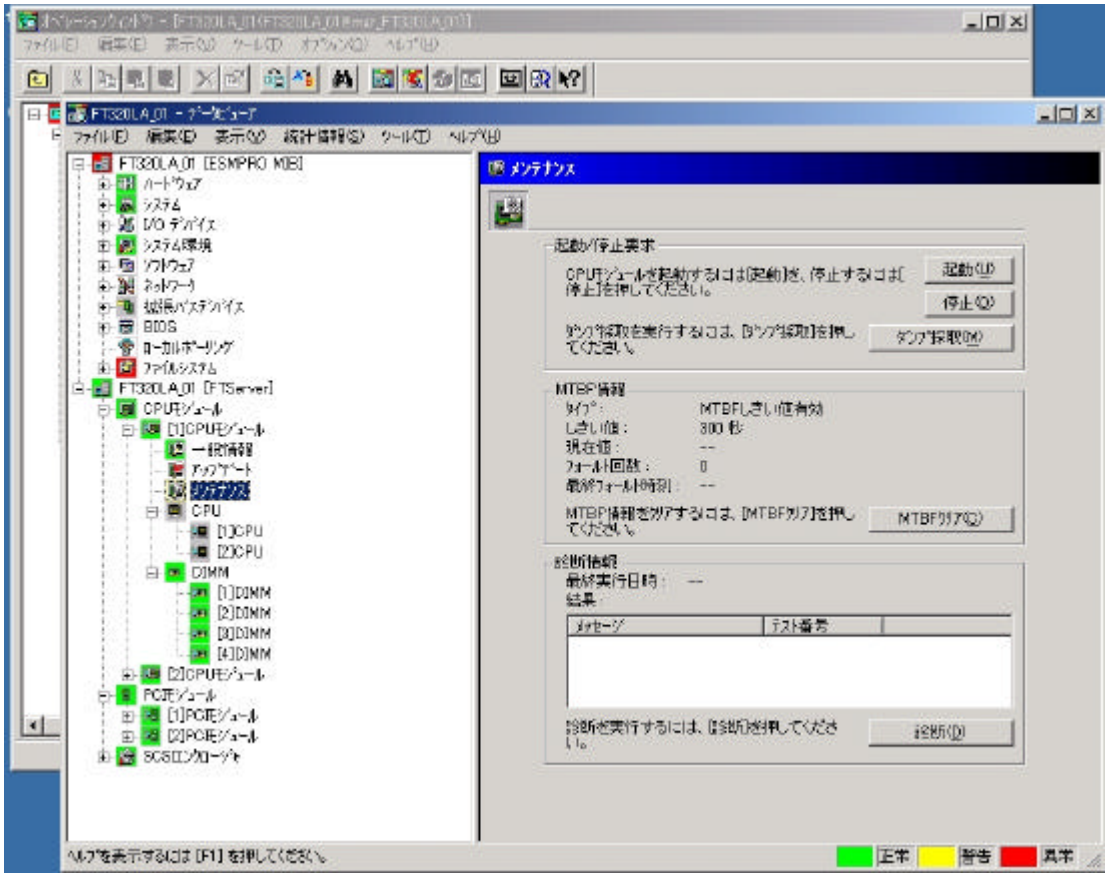

図 8 ESMPRO/Server Manager

まとめ

Express5800/ft サーバの耐障害性向上の仕組みはハードウェアロジックのみに依存せずデバイスドライバ、オペレーティングシステムの機能も利用して実現します。この設計はコストパフォーマンスに優れ、将来発表される高性能なコンポーネントへ柔軟に対応できます。Express5800/ft サーバでは、フォールト・トレラント機能をすべて HAL,Driver で吸収しており、通常の IA サーバと比較してソフトウェアの動作に差異が生じることはありません。HW 障害発生時の切り替え処理・モジュール交換後の復旧処理もアプリケーションからは意識されませんので Express5800/ft サーバを導入するだけでシステムの可用性を向上させることができます。

ハードウェアのトラブルはいつ起きるか予期できません。ハードウェア障害が発生してもシステムの動作に影響を与えない特長を持つ Express5800/ft サーバは、ビジネスクリティカルなアプリケーションを動かすための最適なプラットフォームを提供します。