Empowered by Innovation NEC

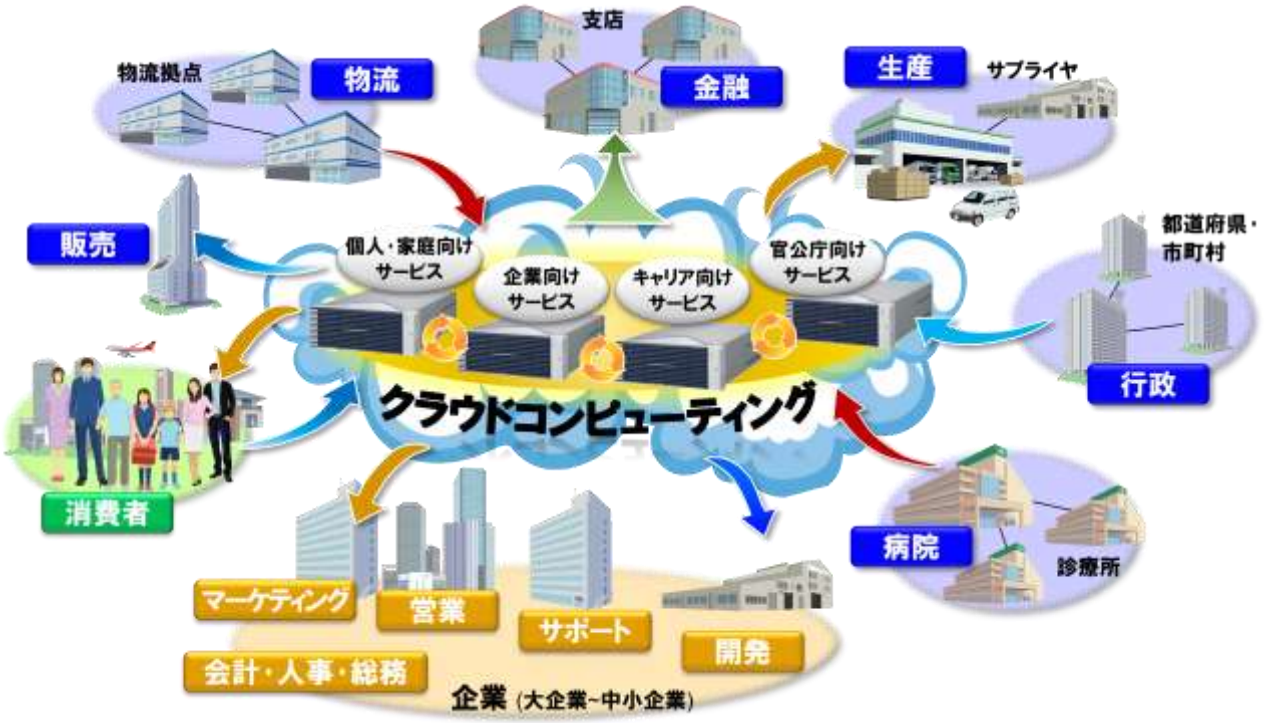
Express5800/ft サーバのご紹介

White Paper

目次

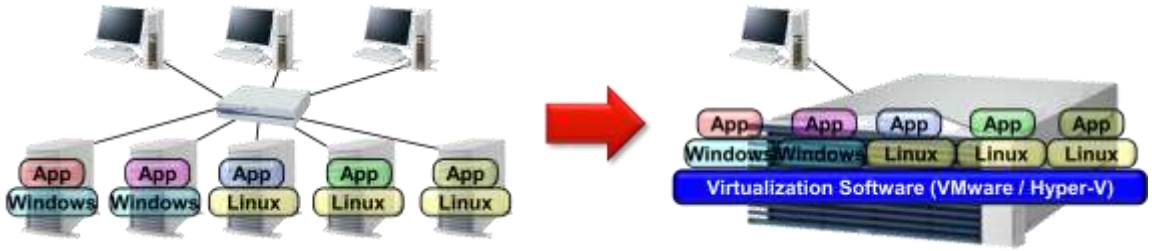
はじめに

IT 技術の驚異的な進歩は、私たち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。近年ではクラウド･コンピューティング技術を活用したクラウド･サービスが大きく成長しており、社会のあらゆるサービス、ビジネスがネットワークを介して結合されています。また、様々な IT 機器やネットワーク･インフラが私たちの生活空間の至るところに浸透しており、これらなくしては我々の生活が成り立たなくなってきています。IT 機器をライフラインとして利用する上で、それらが使い易いだけでなく、安全･確実であることが強く求められており、社会インフラとしてのサーバの信頼性が益々重要になってきています。



一方、サーバ･ベンダーの競争によりサーバ製品が安価に供給されるようになったことで、新たな問題も発生しています。それは、サーバが容易に導入できることで、企業の各部門が必要なときに、必要なシステムを構築し、その都度サーバを増やし続けたことです。結果的に企業は大量のサーバを抱える事になり、維持･管理に要する費用、TCO (Total Cost of Ownership)が増大するという問題に直面しています。

これを解決する手段として、VMware®や、Microsoft® Hyper-V™ といった仮想化技術を用い、一台のサーバで複数の OS を稼動させ、それぞれで複数のサービスを提供する、サーバ統合が盛んに行われています。これによりサーバの管理だけでなく、内部統制も容易になり、大幅な TCO の削減が



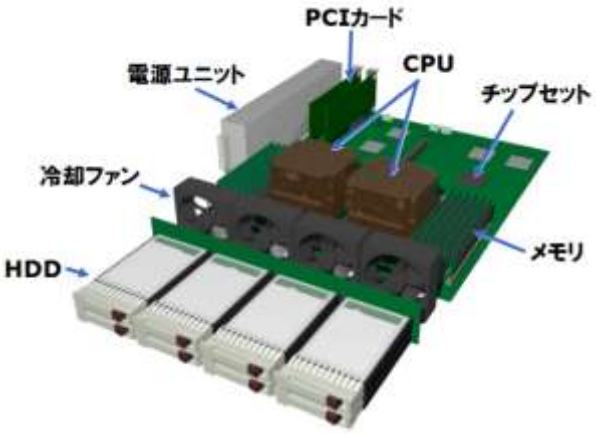
可能となります。

しかし、サーバ統合は別の課題も提起しています。それは一台のサーバが担うサービスが増加し、それに伴い、そのサーバに依存する利用者も大きく増加するということです。つまり一台のサーバ・トラブルによって引き起こされる障害や、損失の規模が、従来に比べ飛躍的に高くなっているのです。これは、リスク分散の格言「卵は一つのカゴに盛るな」とは逆の方向に進んでいることを意味します。

「卵は一つのカゴに盛るな」
卵を複数のカゴに分散して盛っておけば、万が一どれか一つのカゴを落としてしまっても、他の卵は無事だという意味。
もともとは分散投資が大切であることを教える、イギリスの格言。

この様な状況から、仮想化技術が広く普及するにつれ、耐故障性に優れた、高可用性サーバの要求が高まっています。

Express5800 シリーズの信頼性



スペックの高度化に伴い処理能力が飛躍的に向上し、PC サーバがビジネスに果たす役割も拡大し続けています。一方、高度な処理を行うためにサーバの構成コンポーネントには大きな負荷がかかり、故障の発生確率も高まっています。

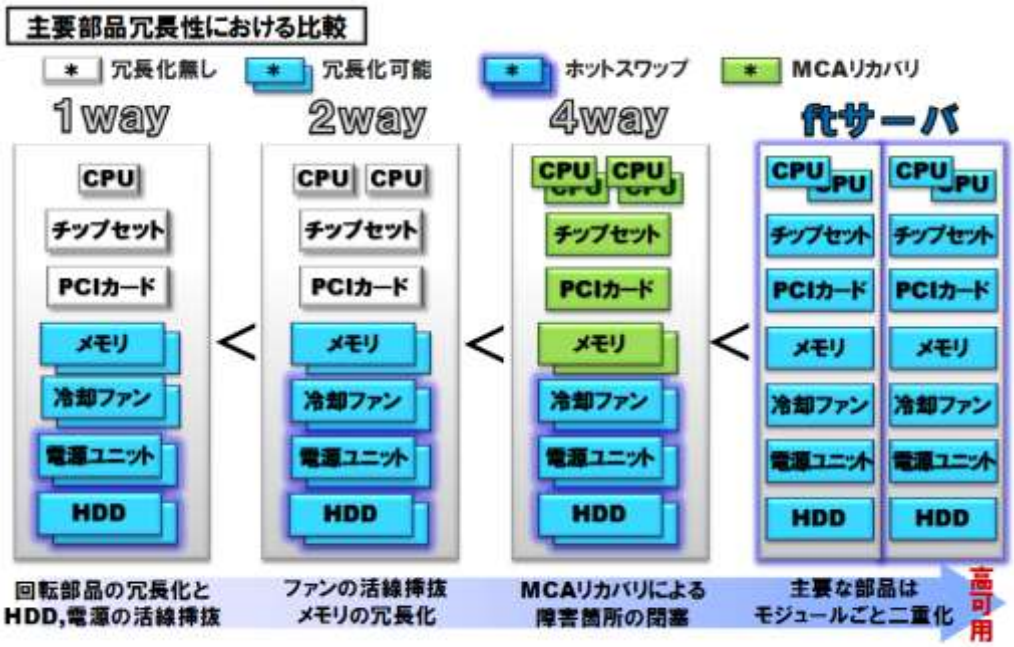
一般的に PC サーバは、右図に示すように、HDD、冷却ファン、電源ユニット、PCI カード、CPU、チップセット、メモリなどのコンポーネントから構成されています。NEC の PC サーバ Express5800 シリーズは製品セグメント毎に、信頼性に対するターゲットが設定されており、上位機種になるにつれ、各コンポーネントの冗長性が向上していきます。

例えば、一番廉価な 1-Way(CPU を 1 個搭載可能)モデルの Express5800/R110 では、HDD(RAID 構成)、電源ユニット、冷却ファンなど回転部品およびメモリの冗長化、さらに HDD、電源ユニットのホットスワップ(活線挿抜)に対応していますが、それ以外のコンポーネントの故障時は、システム・ダウンを引き起こす可能性があります。

2-Way モデルの Express5800/R120 では、回転部品とメモリの冗長化に加え、冷却ファンのホットスワップに対応し、耐故障性を高めています。

4-Way モデルの Express5800/R140 では、1-Way/2-Way ではシステム・ダウンとなってしまうような CPU、メモリ、チップセットの訂正不能なハードウェア不具合時にも、OS と連携した故障箇所の閉塞を行う機能（MCA リカバリ）に対応することで、システムの安定稼動を確保できます。

可用性という観点では、さらに上位に位置するのが ft サーバです。ft サーバでは、主要なコンポーネントを全て二重化しており、2-Way サーバ丸ごと二台分に相当する冗長化を実現しています。



Express5800/ft サーバの開発

以前から銀行の勘定系システムや、ライフラインの制御システムのように非常に重要なシステムでは、メインフレーム・コンピュータやクラスタ・システムによって、高い可用性を実現しています。これらは今後も利用され続けていくでしょうが、社会のあらゆるサービス、ビジネスがネットワークを介して結合されている現在では、これらの限定された範囲に限らず、より身近なところにある IT 機器にも高可用性が求められるようになります。しかし、高価なメインフレームや、運用の複雑なクラスタ・システムはそのような用途には必ずしも向いておらず、低価格で誰にでも簡単に扱える、高い可用性を有する製品が求められていました。

この期待に応えて NEC は、2001 年 6 月に米国 Stratus 社と共同で以下のコンセプトに基づき、IA サーバ（インテル® アーキテクチャに基づくサーバ)をベースとして可用性を飛躍的に高めた FT サーバ（Fault-Tolerant Server)を製品化しました。

(1) 無停止型運用

ハードウェアを二重化することで、何れか一方が故障しても動作継続。

(2) 無停止保守

システムの稼動を継続したまま、故障した部品を交換。

(3) 汎用 OS/ソフトウェアの利用

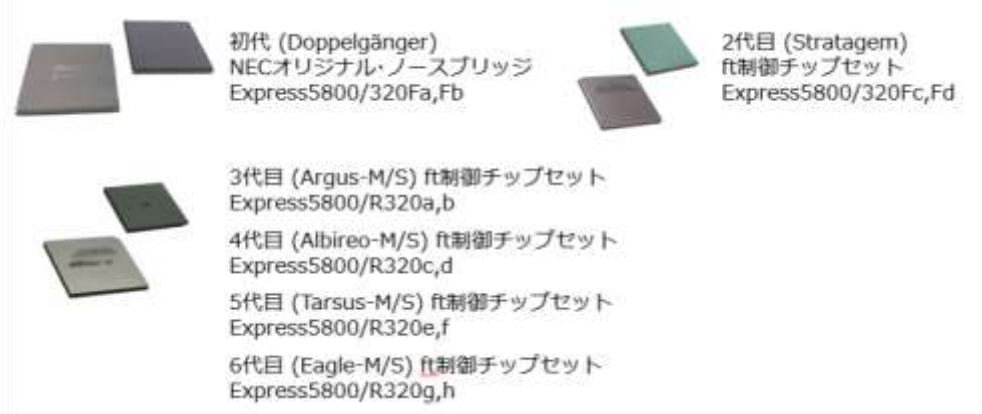
誰でも容易に使えるように、Windows® / Linux® / VMware®といった汎用の OS を搭載し、一般サーバと同様の運用操作性を実現。

ft サーバの市場での認知が進むにつれ、NEC の IA サーバのフラグシップとして ft サーバを推す声が高まり、この勢いをさらに加速するために、2003 年から自社技術による ft サーバの研究開発をスタートしました。それまでは、Stratus 社との協業契約のもと、共同で開発した技術をベースに製品化してきましたが、最新の技術トレンドへの追従、価格低減、お客様からの多様な要求に応えるために、NEC の強みであるハードウェアの開発技術力を活かした自社開発製品が必要との判断によるものです。

2 年半の開発期間を経て、2006 年 2 月に ft 制御 LSI「GeminiEngine™」を搭載した自社開発 ft サーバ、Express5800/320Fa が出荷されました。 これ以後、NEC はインテルの最新 CPU / チップセットにタイムリーに追随した ft サーバを開発しており、Stratus 社[1]へもハードウェアを提供しています。 また、NEC の開発する ft サーバ制御用 LSI は、2021 年現在、6 代目 GeminiEngine™が出荷されております。

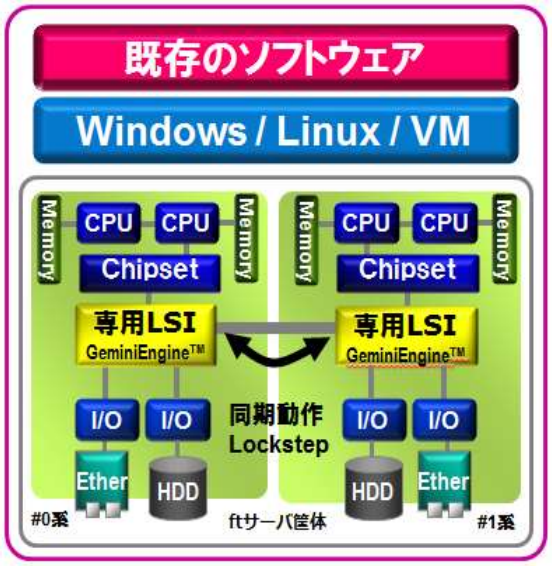
GeminiEngine™

HW の二重化を実現する中核 LSI で、NEC によって開発されています。



[1] 現在 NEC は Stratus 社と ft サーバを共同開発しており、ハードウェアは NEC が開発し、ソフトウェアは Stratus 社が開発しています。

Express5800/ft サーバの特長



前ページのコンセプトに示した通り、ft サーバとは

- ハードウェア故障でも止まらない
- 止めずに修理、交換が可能
- 既存 OS/アプリがそのまま動作

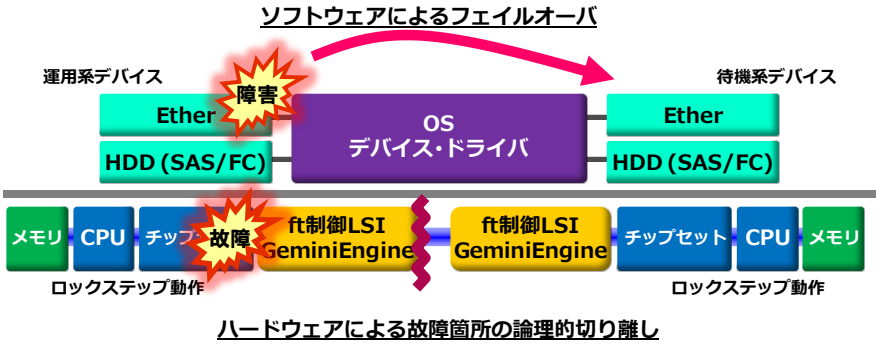
を達成する、ノンストップを追及したサーバです。このため、一般的なPCサーバとは大きく異なる構造を持っています。

右は ft サーバの模式図ですが、筐体の中に全く同一の固まり(これを CPU/IO モジュールと呼びます)が 2 個入っています。それぞれの CPU/IO モジュールは、中央に位置する専用 LSI 以外は、一般サーバとほとんど同じ部品で構成されており、モジュール単体でサーバとして動作することが可能です。

中央の専用 LSI は、NEC の開発する GeminiEngine™であり ft サーバを特徴づける最も重要な部分です。ft サーバは、これらハードウェアによる二重化機能と、ソフトウェアによる二重化制御の双方の技術を組み合わせ、ノンストップを実現しています。以下、その特徴を詳しく説明します。

故障しても止まらない ft サーバ

右図は ft サーバ・システムにおけるソフトウェアとハードウェアの関係を概念的に示しています。



CPU やチップセットといった主要コンポーネントがシステムの土台として全体を支えており、その上に OS が動作しています。また Ether(イーサーネット)や SAS (シリアル・アタッチ SCSI)、FC (ファイバ・チャネル)で制御される HDD といった I/O コンポーネントもこの土台上にあり、OS やドライバから命令を受けて動作しています。

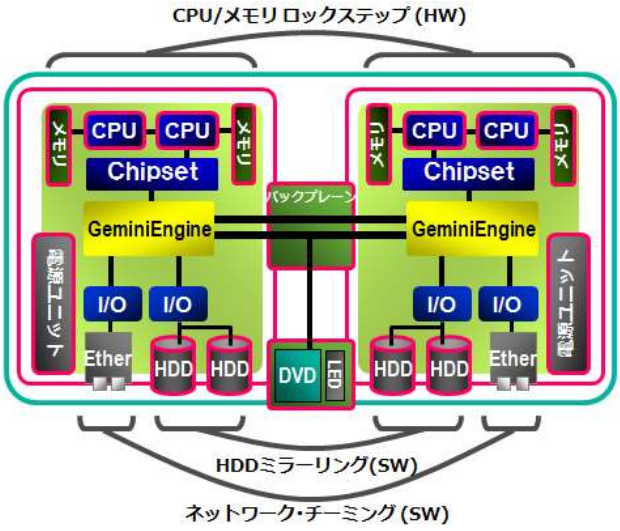
土台として CPU やメモリが二組存在していますが、これらは後述するロックステップ技術により完全に同一の動作を行っており、OS はどちらのハードウェアで動作しているかを意識しておらず、またシステムとしては一つの OS インスタンスのみが動作しているのと等価です。

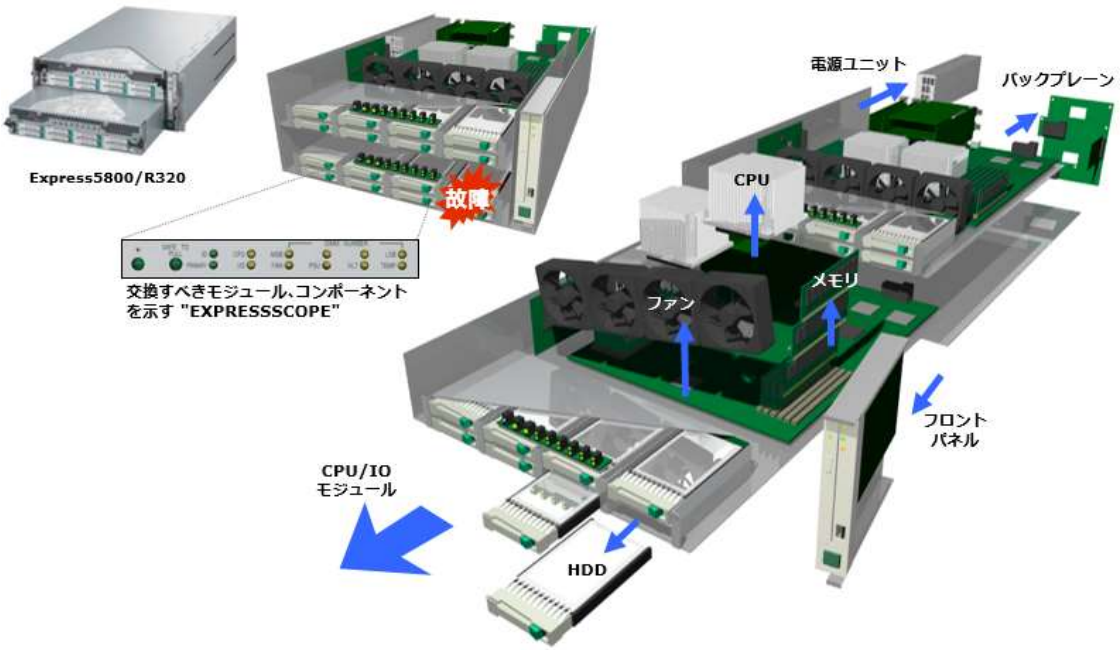
ft サーバでは全てのコンポーネントが二重化されており、I/O コンポーネントの故障発生時にはソフトウェアによって使用デバイスの切り替え(フェイルオーバ) が行われ、動作を続行します。一

方、CPU やチップセットは、ソフトウェア自身が動作する土台であるため、通常時は、両者をロックステップ技術により全く同一の動作をさせ、故障時にはハードウェアによって故障箇所が論理的に切り離され、動作を続行します。以上の仕組みにより、ft サーバは故障しても止まらない運用を可能としています。

システムを止めない修理交換

右図は最新の ft サーバ Express5800/R320g, h の模式図です。赤い枠で示される、CPU/IO モジュールと CPU、メモリ、電源ユニット、HDD、バックプレーン、フロントパネルの各コンポーネントは交換可能単位であることを示しています。またこのうちバックプレーン[2]を除く全てのコンポーネントは、システムを止めずに交換することが可能です。この実現のため、ft サーバの筐体は下図に示す通り、徹底したモジュール構造となっており、各コンポーネントを容易に交換することが可能です。なお、CPU/IO モジュールには EXPRESSSCOPE® と呼ばれる LED が搭載され、故障発生時には、どのコンポーネントを交換すべきかが、一目で分かるようになっており、迅速な故障箇所の判断、交換、復旧を可能としています。

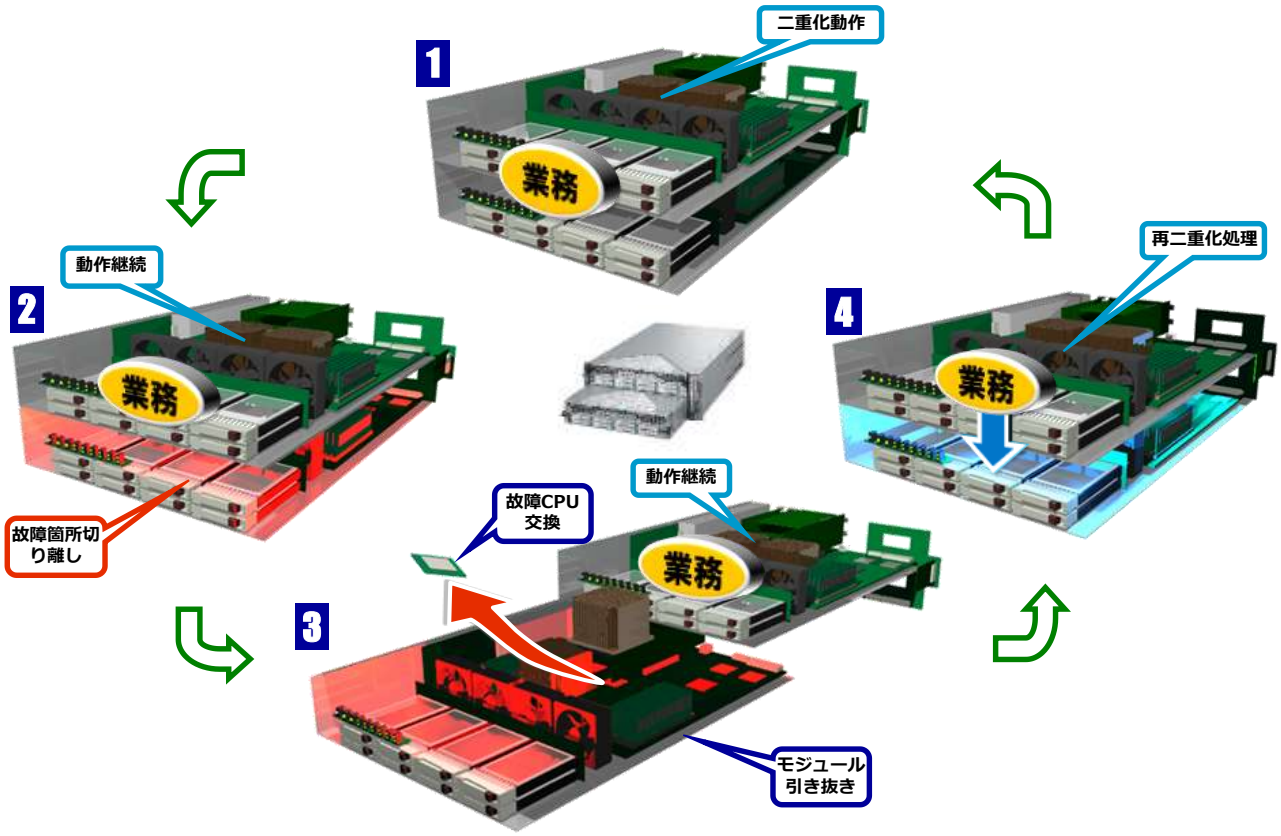




以下に、CPU が故障した場合を例にとって、実際のモジュール交換の流れを示します。

[2] バックプレーンにはコネクタと配線しか存在しておらず、故障の可能性はほとんどありません。

1. 正常運転時、ft サーバは二重化状態で動作。
2. 故障が発生(仮に CPU が故障)すると、GeminiEngine™が故障モジュールの CPU 部位を切り離し、残ったモジュールで動作を継続(この際、故障モジュールの EXPRESSSCOPE®には CPU が故障したことを示す LED がアンバー点灯)。
3. 故障モジュールを引き抜き、故障コンポーネント(この場合 CPU)を交換。ft サーバは残ったモジュールで引き続き動作を継続。
4. 交換修理済みのモジュールを ft サーバに戻すと、GeminiEngine™が自動的に再同期化処理を行い、二重化状態に復旧。以下、[1]に戻る。



既存 OS/アプリがそのまま動作

無停止型のフォールト・トレラント(FT)サーバは過去にも幾つかのメーカーが開発しており、非常に高い可用性を有していましたが、いずれも専用 OS を搭載した FT 専用装置であり、限られた領域でしか使用されませんでした。その後 Windows®や Linux®など、いわゆるオープン・システムを使用する、クラスタ・システムが登場しました。クラスタ・システムはハードウェアだけでなく、ソフトウェアの障害にも対処できるため、高い可用性を有するシステムを構築可能ですが、故障時の切り替えはソフトウェアに負うところが多く、また使用するアプリケーションも、バックアップ・サーバへ切り替わって継続稼動することが前提となるため、既存のどんなアプリケーションもそのままクラスタ・システムで使用できるわけではありません。クラスタ・システムの使用が前提では

ないアプリケーションでは、ほとんどの場合、処理引継ぎのための改造が必要となります。

一方、Express5800/ft サーバは、システム上は 1 つの OS が稼動する 1 台のサーバとして動作しています[3]。そのためシステム構築は、二重化されたハードウェアを意識する必要がなく、通常のサーバと同様に行えます。ミドルウェアやアプリケーションに特別な設定を行わずに、そのまま利用可能ですので、通常のサーバから Express5800/ft サーバに置き換えるだけで、システム全体の可用性を向上させることができます。

既存 OS/アプリケーション動作のポイント

- ネットワーク上では、通常の 1 台のサーバとして存在し、二重化を意識することなく利用可能。
- アプリケーションの二重化設定は不要。利用アプリケーションの制御もなし。
- シングル･サーバのように運用できるので、管理は容易かつ低コスト。

99.999%を超える可用性

可用性とはシステムが継続して稼動できる能力のことをいいます。混同されやすい言葉に信頼性がありますが、厳密な意味では可用性と信頼性は異なります。信頼性は故障する頻度が少なく、結果として故障している期間が短いことを指します。一方、可用性は利用者がシステムを利用し続けられる能力のことを指します。一般には故障が多ければ信頼性が低く、可用性も低くなります。しかし故障が発生しても冗長化されて、実際の使用には影響が無い場合、可用性は高く保たれます。

よく ft サーバは故障しないという誤解がありますが、実際には各種コンポーネントがほぼサーバ 2 台分あるため、故障率は一般サーバの約 2 倍あります。ft サーバは故障しないのではなく、故障してもシステムを使用し続けることが可能、つまり高可用性サーバを目指しています。

可用性を数値として表す場合、稼働率を用います。稼働率とは修理可能なモジュール、コンポーネントが、規定の時間内に機能を維持している確率のことをいいます。

---

[3] VMware®及び Windows Server®2016,2012R2,2008R2 Hyper-V™ 対応モデルでは仮想化による複数 OS の稼動が可能です。

稼働率	年間停止時間
99.9999%	32秒
99.999%	5分15秒 ⇦
99.99%	52分34秒
99.9%	8時間46分
99%	3日15時間36分

NEC の Express5800/ft サーバ･シリーズは稼働率 99.999%以上を達成しており、適切に運用された場合の年間停止時間は 5 分 15 秒以下と、極めて高い可用性を実現しています。

（年間停止時間は設計値から算出される平均停止時間であり、稼働時間が保証されるものではありません）

Express5800/ft サーバの基本アーキテクチャ

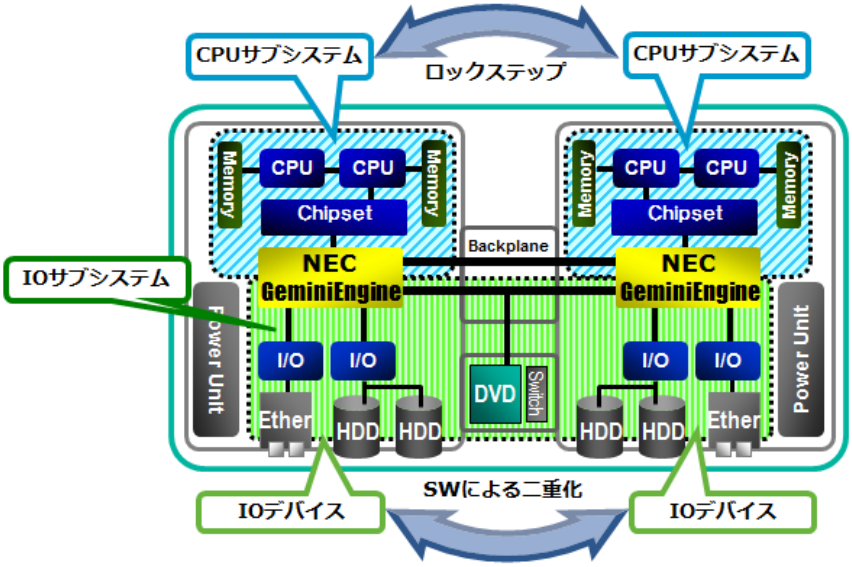
ft サーバに求められる機能は非常に多岐に渡っていますが、基本的な思想は至って単純です。それは「2 つのハードウェアを用意し、一方が故障して動作を停止しても、正常な方が動作を継続する」というものです。これを実現するために、ハードウェア、ソフトウェア共に多くの機能が必要となりますが、大きく分けて以下の 3 つの機能が基本となっています。

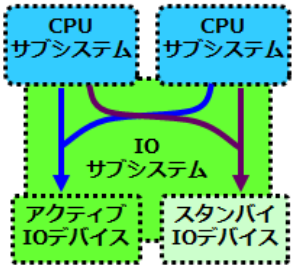
- I/O フェイルオーバ
- サプライズ･リムーバルとエラーの隠蔽
- ロックステップ

ft サーバの 2 つの CPU/IO モジュールは、CPU やメモリ、チップセットを含む CPU サブシステム部と、各種 IO デバイスを含む IO サブシステム部に分かれており、それぞれ二重化の方式が異なります。

次ページの図は CPU サブシステム、IO サブシステムの範囲と概念を示しています。CPU サブシステムは二つのモジュールで全く同じ動作をしており、双方から発行される 2 つのリクエストは GeminiEngine™で 1 つにまとめられて処理されるため、システムとしては実質 1 個の CPU サブシステムが動作しているのと等価になります。一方 IO サブシステムは、バックプレーンを介してモジュール間をまたがって存在しており、二重化状態では 2 モジュール分の IO デバイスが存在します[4]。また、CPU サブシステムからは両モジュールの IO デバイスが見えており、このことから IO サブシステムにおいては、ソフトウェアによる二重化制御が可能となっています。

[4] R320c、d、e、f、g、h では、DVD や USB などの一部デバイスは、利便性を考慮してシステムで一つしか存在しません。なお、これら単一デバイスが故障してもシステムが停止することはありませんし、やはり稼動したままで交換が可能です。

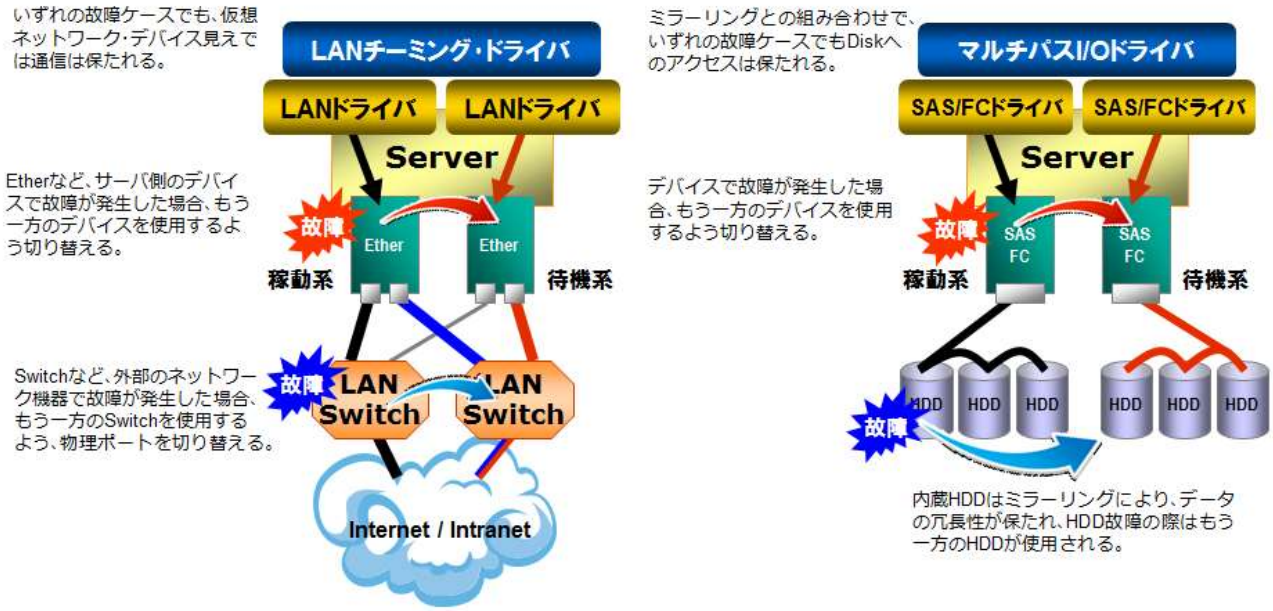




CPUサブシステムからは、両モジュールのIOデバイスが見えており、ソフトウェアによる二重化制御が可能。

I/O フェイルオーバ

I/O デバイスは両モジュールで同一の構成となっており、どちらか一方の I/O デバイスがアクティブ側として通常使用され、もう一方はスタンバイ側として待機状態になります[5]。使用しているアクティブ側デバイスで故障が発生した場合、これをソフトウェア(デバイス･ドライバ)で検出し、直ちにスタンバイ側に切り替えます。この代替処理(フェイルオーバ)方式は、一般的な PC サーバの I/O デバイス冗長化技術として開発されたものであり、さらに ft サーバでは、システムの稼働中にも故障デバイスを交換可能とする独自のモジュール構造により、一層の機能強化がなされています。



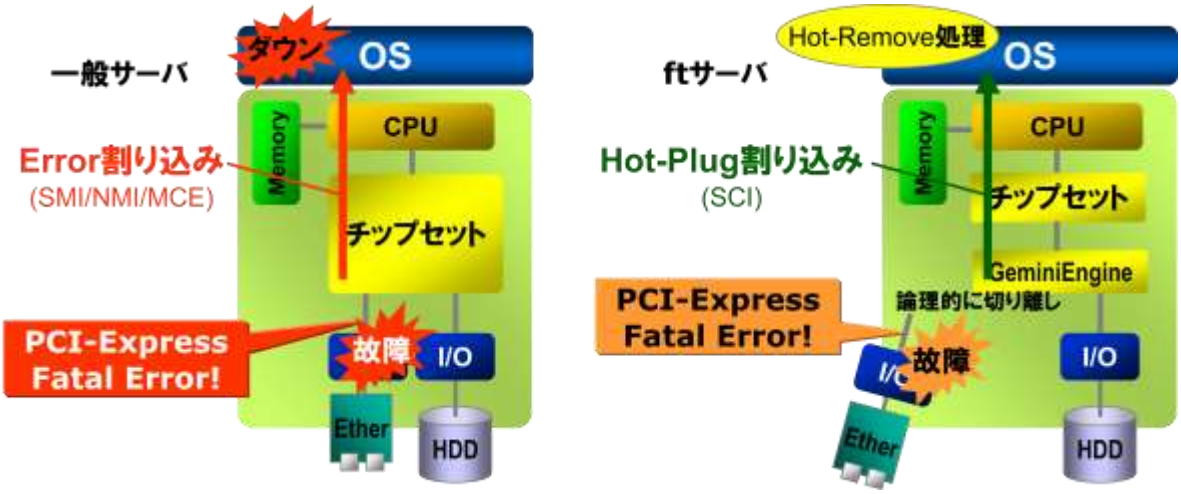
[5] 設定により、両方の I/O デバイスを使用することは可能ですが、二重化動作時と単体動作時の I/O 性能を平準化するために、アクティブ/スタンバイでの使用を想定しています。

例えば、ネットワークにおいては、チーミング(Windows®)やボンディング(Linux®)と呼ばれる方法が用いられます。これは、複数の物理ネットワーク・ポートを束ねて一つの仮想ネットワーク・ポートを生成し、ネットワークの障害発生時には、別の物理ポートに切り替えて動作を継続するため、仮想ネットワーク・ポートとしては障害の影響が及びません。

同様にファイバ・チャネルや SAS 等のストレージ系ではマルチパス I/O ドライバと HDD のミラーリングによって I/O フェイルオーバを実現します。

サプライズ・リムーバルとエラーの隠蔽

デバイスの故障は、予期せぬ動作を引き起こし、システムを巻き込んだ障害に発展する場合があります。典型的な故障の例として、I/O デバイスとチップセットとの接続に使用されている PCI Express で何らかの訂正不可能な致命的なエラー(Fatal Error)が発生した場合を考えます(下図-左)。この場合、一般的な PC サーバでは致命的なハードウェア・エラーとして OS に通知され、そのままリカバリできずにシステム・ダウンに至ってしまいます。



一方、ft サーバでは、全ての I/O デバイスは GeminiEngine™に接続されており、I/O デバイスで発生している事象を全て把握しています。仮に同様に PCI Express で致命的なエラーが発生した場合、GeminiEngine™は該当部分を論理的に切り離し、該当 I/O デバイスをシステムから見えない状態にします。また、OS へはそのままエラー通知をせずに、Hot-Plug[6]で使用される割込みを使用し、デバイスが突然引き抜かれたことを示す、サプライズ・リムーバルとして通知を行います。これにより、OS には実際のハードウェア・エラーが隠蔽され、システム・ダウンを防ぐことが可能となっています。

なお、サプライズ・リムーバル通知を受けた OS は該当デバイス・ドライバにその旨通知を行い、デバイス・ドライバ側はその通知により I/O フェイルオーバを発生させ、代替デバイスで運用を続行します。何の前準備、通知もなく突然デバイスが抜かれた場合を想定した「サプライズ・リムーバル」

[6] Hot-Plug とはシステムを稼動させたまま、PCI カードや PCI Express カードを抜き挿しすることを可能にする方式。

は Windows®や Linux®で使用されている Hot-Plug で規定された機能の一つですが、オプション扱いのため、全てのデバイス・ドライバやアプリケーションがこの機能をサポートしている訳ではありません。ft サーバではこのサプライズ・リムーバル機能のサポートが必須のため、どんな I/O デバイスも使用可能という訳にはいかず、一般サーバに比べサポートできるデバイスに制限があります。

現状下記のデバイス・ドライバに対してサプライズ・リムーバル機能を追加し、サポート対象 I/O デバイスとしています。

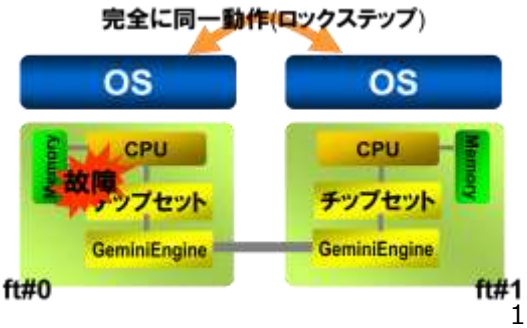
- イーサーネット
- SCSI / SAS
- ファイバ・チャネル
- ビデオ・ディスプレイ
- USB (但し I/O フェイルオーバの際は、一旦デバイスの挿抜が発生します)

サプライズ・リムーバルをサポートしない、不適切なドライバを使用した場合、正しく I/O フェイルオーバが行われず、システム障害を引き起こす場合があります。特定のアプリケーションにおいてはドライバにフィルターをかけたり、ハードウェアに直接アクセスするものがあり、使用には注意が必要です。ご不明の際は、NEC ファースト・コンタクト・センターへお問い合わせ下さい。
http://www.nec.co.jp/products/express/question/top_sv1.shtml

ロックステップ

ロックステップは ft サーバの最も重要な機能であり、NEC が世界に誇るオンリー・ワン技術でもあります。

CPU やチップセット、メモリなどのサーバの基幹コンポーネントが存在する CPU サブシステムは、それ自身上で OS や制御ソフトウェアが動作しています。このため CPU サブシステム内のコンポーネントが故障してしまうと OS は動作続行不可能な状態となり、さらにサブシステム内のデータは全て不正状態、または消失してしまいます。従って、CPU サブシステムにおいては、I/O サブシステムのような稼動系/待機系によるフェイルオーバは不可能です。



ft サーバでは 2 つのモジュール間の CPU サブシステムをクロック単位で完全に同期させて動作させており、これをロックステップと呼びます。両方とも全く同じ

動作をしているので、故障発生時は対象となる CPU サブシステムを論理的に切り離し、正常な方で動作を続行させます。従い、CPU サブシステムには稼動系/待機系の概念はありません。

ロックステップの実現には最先端のテクノロジが必要となります。NEC のハードウェア開発陣は、様々な独創的アイデアを GeminiEngine™に搭載することにより、常に最新ハードウェアでのロックステップを実現させています。その困難さから、現在ではインテル®アーキテクチャを使用したロックステップ型 FT サーバのハードウェア開発は、世界でも NEC のみとなりました。

GeminiEngine™とハードウェア二重化技術

GeminiEngine™は ft サーバを実現するための中核 LSI で、主に以下の役割を担っています。

- デターミニズムの確保とロックステップの実現
- CPU コンテキストとメモリ内容の同期化
- エラー検出と切り離し制御

以下、それぞれについて詳細に説明します。

デターミニズムの確保とロックステップの実現

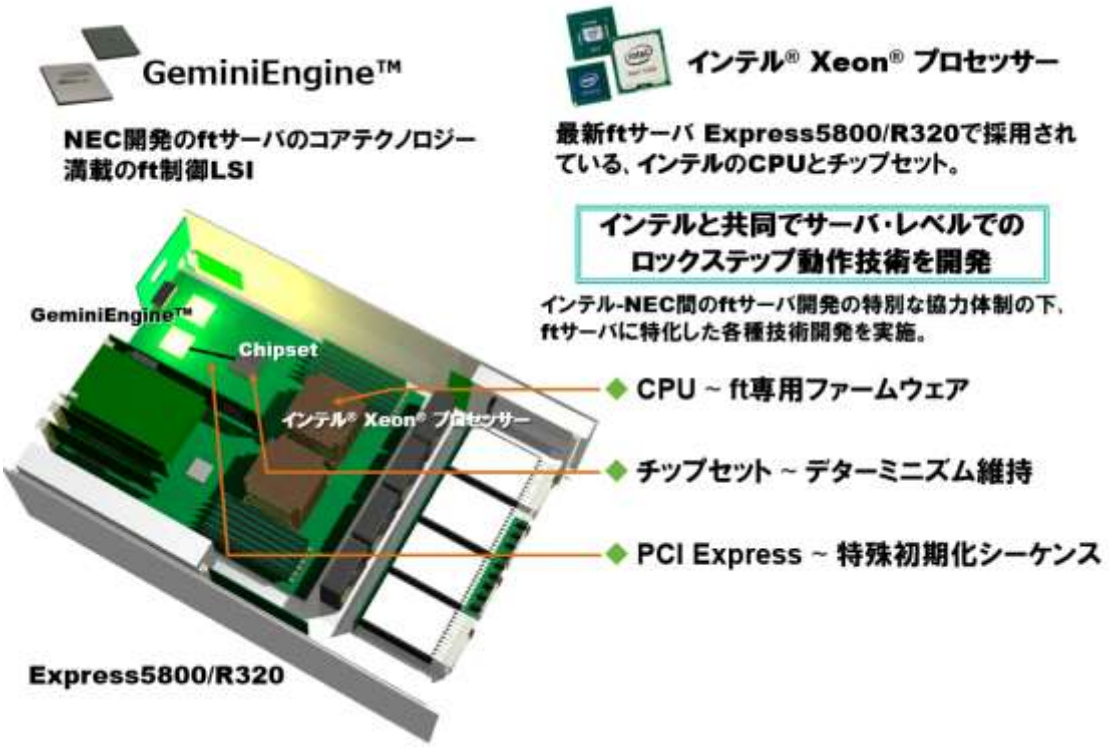
ロックステップは ft サーバ実現の最も重要な要素です。具体的には LSI に外部から全く同じクロック信号を入力し、同じタイミングでリセットを解除すれば、何回やっても毎回必ず同じ動作をするはずです。この特性を「デターミニズム」と呼びます。また、全く同一の LSI を二つ並べ、同一のクロック信号を入力して同時に動かし始めると、デターミニズムが確保されている場合、2 つの LSI は全く同じ動作をします。この状態を「ロックステップ」と呼びます。

過去にはこの概念でロックステップによる FT サーバを開発していたベンダーも多くありましたが、最近ではあらゆるコンポーネント、インタフェースが高速化し、さらにアナログ特性に依存する部分も増大しており、デターミニズムを利用した FT サーバの開発は困難を極めています。

例えば、CPU 動作周波数の高速化に加え、温度/消費電力による CPU 動作周波数、電圧の調整機能などは多分にアナログ的な要素が関与しており、デターミニズムの維持を困難にしています。またインタフェースも、CPU を接続するインテル® QPI (インテル® QuickPath Interconnect ~8.0GT/s 動作)や I/O デバイスを接続する PCI Express (5.0GT/s 動作)など、高速シリアル伝送が主流となっており、僅か数百ピコ秒[7]タイミングがずれただけでもロックステップ出来ないという状況にあり、FT サーバの開発には極めて高度な技術が要求されています。

[7] 1 ピコ秒は、1 秒の 1 兆分の 1 の長さ。0.000000000001 秒、psec と表記。

特にロックステップの実現が難しい、CPU とチップセットにおいては、インテルと協力してロックステップ技術の開発を行っています。ft サーバは一般のサーバと同じ、インテルが開発する CPU、チップセットを採用していますが、これらにはロックステップを可能とする「特殊な動作モード」というものは存在しません。このため、CPU、チップセット内部、PCI Express インタフェースのそれぞれで、ロックステップを可能とする技術をインテルと協力して開発する必要がありました。



また、ロックステップを実現する上で必要なクロック信号の位相調整、クロック源振の二重化などはクロック・チップ・ベンダーとの緊密な連携により開発、実現されています。

このように様々なコンポーネント・ベンダーとの協力体制の下で、ロックステップ技術が開発されており、その技術の全てが GeminiEngine™に搭載されています。Express5800/R320g,h においては、GeminiEngine™により 100 ピコ秒レベルでのクロック位相調整やリセットのタイミング調整が行われており、CPU やチップセットをはじめ、多数の LSI のデターミニズムが確保され、ロックステップを実現しています。

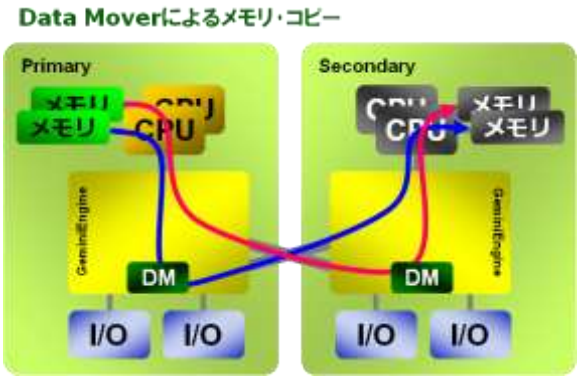
CPU コンテキストとメモリ内容の同期化

ft サーバ起動時や、ボード交換による CPU サブシステムの二重化の際は、全てのメモリ内容を、もう一方の CPU/IO モジュールにコピーします。コピーのほとんどは Brownout Copy と呼ばれる方式で行われ、稼働中のサービスが止まることはありません[8]。コピーの最終段階では、一旦システム

[8] Windows Server® 2016,2012R2、2008 R2 で Hyper-V™ 使用の場合、メモリ・コピーは全て Blackout Copy 方式で行われます。実装メモリ量に応じ、数秒～数十秒のシステム停止時間が発生します。システム構成ガイドを参照のうえご注意願います。

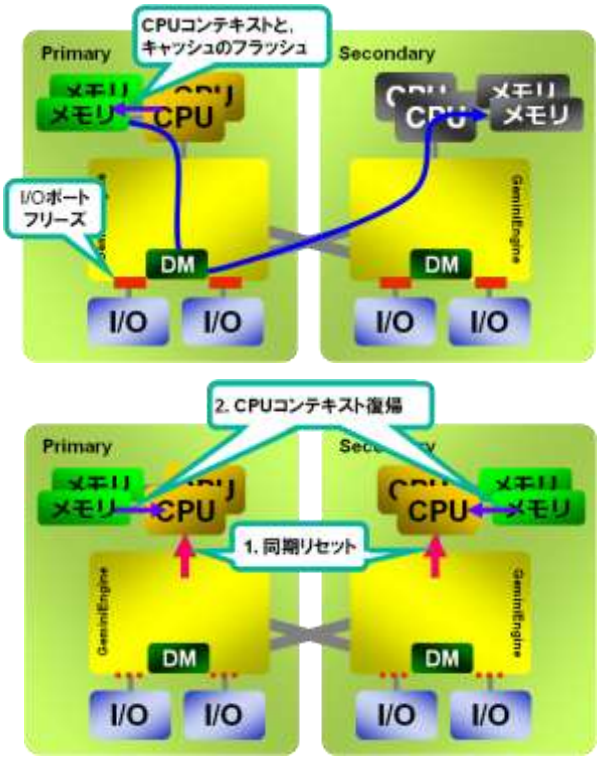
を停止させる Blackout Copy 方式で、CPU の内部情報(コンテキスト)とキャッシュ内容がコピーされます。この際、僅かにサービス停止が発生しますが、極めて短い時間のためシステムに影響を及ぼすことはありません。

Brownout Copy 時の動作



図中、左側が稼働中の CPU/IO モジュールを示し、右側は被二重化対象モジュールを示しています。メモリ・コピーは GeminiEngine™内部のデータ・ムーバ (図中 DM)で行われます。但し、この間も CPU、I/O デバイスは動作し続けており、メモリの内容も刻一刻と変化しています。コピー済みのメモリ領域が更新された場合、再度その部分のみコピーを行います。なお、これらコピー制御は ft 制御ソフトウェアで行われます。

Blackout Copy 時の動作



Brownout により、全メモリ・エリアのコピーが完了すると、全ての I/O 及び、OS の動作を HW 的に一旦停止させます。そして ft 制御ファームウェアにより CPU のコンテキストとキャッシュ内容がメモリ上にフラッシュされ、データ・ムーバは該当するメモリ内容をコピーします。

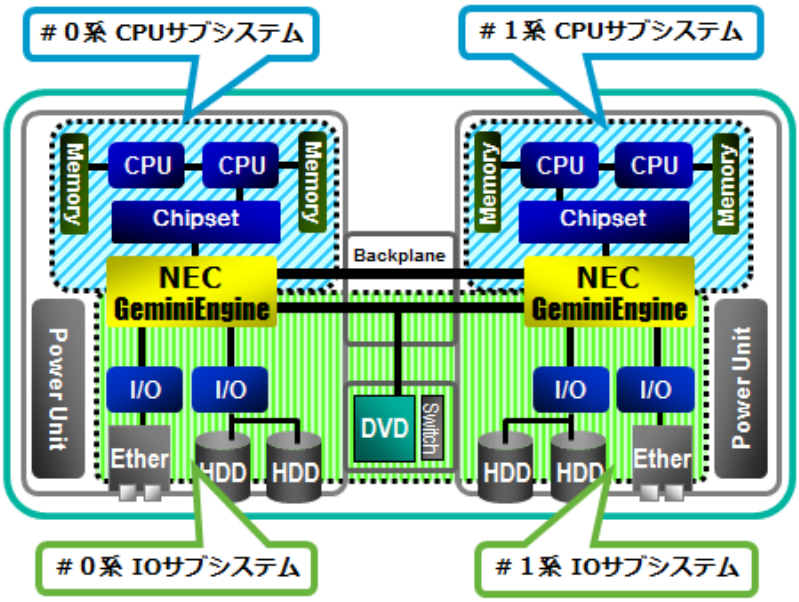
両モジュールをロックステップさせるために、両 CPU に同期リセットを掛けます。これ以降、両モジュールの CPU は全く同じ動作をします。最後に CPU に停止直前のコンテキストを復帰させ、I/O 及び OS の動作を再開させます。

これらの Blackout 処理は短い時間で行われ、稼働中サービスにインパクトを与えることはありません。

エラー検出と切り離し制御

ft サーバにおいては、エラーの検出能力と、エラー発生箇所を特定する分解能を上げることが大変重要です。また、エラー箇所を論理的に切り離して動作を続行することが求められるため、不正動作が伝搬しない工夫も必要となっています。Express5800/ft サーバではシステムを 4 つのサブシ

ステムに区切って管理しており、エラー検出の際はこの 4 サブシステム単位で切り離し制御が行われます。



この各サブシステムにまたがって存在する GeminiEngine™は、システム中の全てのトランザクションと、チップセットからのエラー信号を監視しており、ひとたびハードウェア的なエラーが検出されると即座に該当部位を論理的に切り離します。しかし、切り離したサブシステムを含むモジュールは直ちに修理交換が必要と判断される訳ではありません。それは、ハードウェアのエラー発生原因は多岐に渡っており、一概に故障と判断できないからです。ハードウェアのエラー要因は主に以下のものが挙げられます。

1. 故障によるエラー発生
2. 外部からの電気的ノイズによる一時的な誤動作
3. 宇宙線やその他放射線によるメモリ化け

この中で、[2]は稼働環境に依存して、[3]は通常状態において、ある一定の確率で発生する一時的なエラーであり、交換を要するものではありません。この判断のため、ftサーバではエラー検出で切り離した後、サブシステム内のハードウェアを診断チェックし、明らかな故障が見つからない場合、再度二重化して使用を継続します。但し、診断チェックでは簡単に見つからない故障の可能性も排除できないため、サブシステム毎にエラーの発生回数をカウントしておき、ある閾値を超えてエラーが発生した場合、再組み込みを停止し EXPRESSSCOPE®の LED 点灯などのエラー通知を行なうとともに、通報機能による保守サービス会社への自動通報（要登録）によりモジュールの交換を促します。

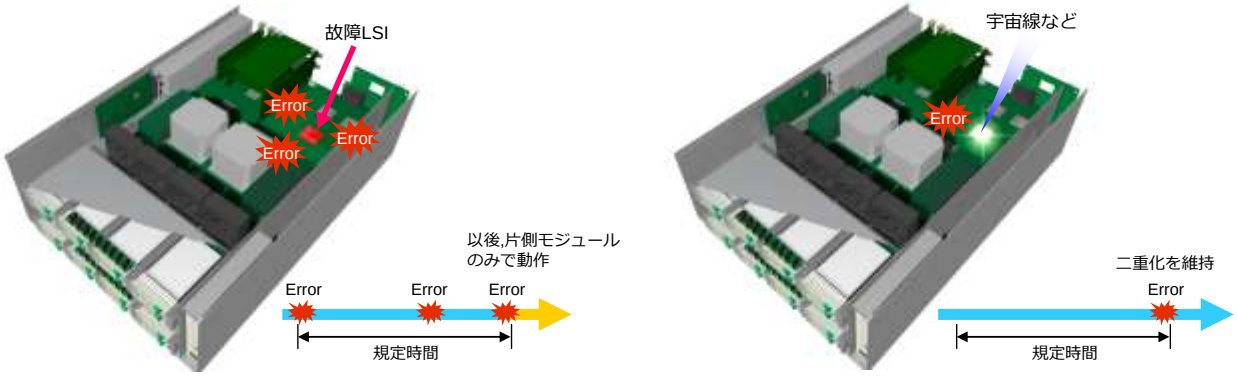
これらは具体的にはMTBF と呼ばれる値を用いて制御しています。MTBF とはコンポーネントで故障(エラー)が発生するまでの時間の平均値を意味し、平均故障間隔とも呼びます。例えば、MTBF

が 10 万時間のシステムの場合、確率的に 10 万時間(約 11 年半)に一度故障が発生することを意味します。一般的に MTBF はサーバを構成する全ての部品から計算して求めることが出来ます。

Express5800/ft サーバでは、計算して求められた MTBF 値にさらに、発生頻度を加味した独自の方式を用いてモジュールの切り離し/再組み込み制御を行っています。

切り離し/再組み込み制御の概要

いずれのケースも、図中下側のモジュールが無停止で動作し続けていることに注意



規定時間内にハードウェア･エラーがしきい値を超えて
発生した場合は故障と判断し、モジュールを切り離し。
以後、残った片側モジュールのみで動作続行。

規定時間内のハードウェア･エラー回数がしきい値を下回る
場合は一時的な誤動作と判断し、モジュールを再組み込み。

このように、エラーが発生した場合でも、実際の使用上問題のない、一時的なエラーと判断できる場合は、可能な限り再二重化を行い、可用性を上げる工夫がなされています。

おわりに

クラウド･コンピューティングや仮想化によるサーバ統合を支える、高可用性プラットフォームである ft サーバの機能詳細について、その中核となる GeminiEngine™を中心に使用されている技術を紹介しました。

NEC は、一般サーバの可用性を飛躍的に高めるクラスタと、ftサーバという２つの高可用性の技術を持つ、数少ない企業の１つです。今後もこれらの特徴を適材適所で活かし、お客様からの様々な要求に応えるソリューションを提供してまいります。