特集

基幹IAサーバ“PRIMEQUEST”の高信頼システムを実現する技術

High-Reliability Technology of Mission-Critical IA Server: PRIMEQUEST

あらまし

基幹IAサーバ“PRIMEQUEST”は，基幹業務に適用するために必要な高信頼性，高可用性を実現するために，チップセットレベルから，ユニットレベル，システムレベルにわたり各種RAS（Reliability, Availability, Serviceability）機能を装備している。

本稿では，PRIMEQUESTの概要を説明し，その後，システムの各階層で採用している高信頼性技術を中心に，高速化技術，スケーラビリティ，高信頼性・高可用性技術について述べる。

Abstract

The PRIMEQUEST of mission-critical IA servers are equipped with various RAS (Reliability, Availability, and Serviceability) functions from the ASIC (Application Specific Integrated Circuit) levels, unit levels, and up to the system level. These RAS functions provide the high reliability and high availability required for mission-critical operations on these servers. This paper first gives an overview of the PRIMEQUEST. It then discusses the speed-enhancement technology, scalability, high-reliability technology, and high-availability technology of the PRIMEQUEST, focusing on the high-reliability technology implemented in the system hierarchies.

浜田王才（はまだ　おうさい）

基幹IAサーバ事業部 所属
現在，PRIMEQUESTシリーズの開発に従事。

まえがき

基幹IAサーバ“PRIMEQUEST”は，富士通が独自に開発したASIC（Application Specific Integrated Circuit：特定用途向け集積回路），およびインテル社製64ビットCPUであるItanium2プロセッサを搭載したミッションクリティカル業務を指向した富士通の新しいサーバである。

本稿では，PRIMEQUESTのハードウェアの概要について述べ，開発した高速化技術，高信頼性・高可用性技術，スケーラビリティ，高保守性デザインについて述べる。

また，以下のトピックスについても概説する。

(1) システムを構成している主なユニットを接続するクロスバシステム
(2) 一つのシステムを分割するパーティショニング機構
(3) CPU/メモリ資源とI/O資源の自由な組合せを実現するフレキシブルI/O機構
(4) ハードウェアコンポーネントに故障が発生してもシステムをノーダウンで走行することを可能とするシステムミラー機構
(5) 信頼性，保守性の向上を実現するケーブルレスデザイン
(6) システムの可用性を向上させる各種冗長化機能，活性保守機能

ハードウェア概要

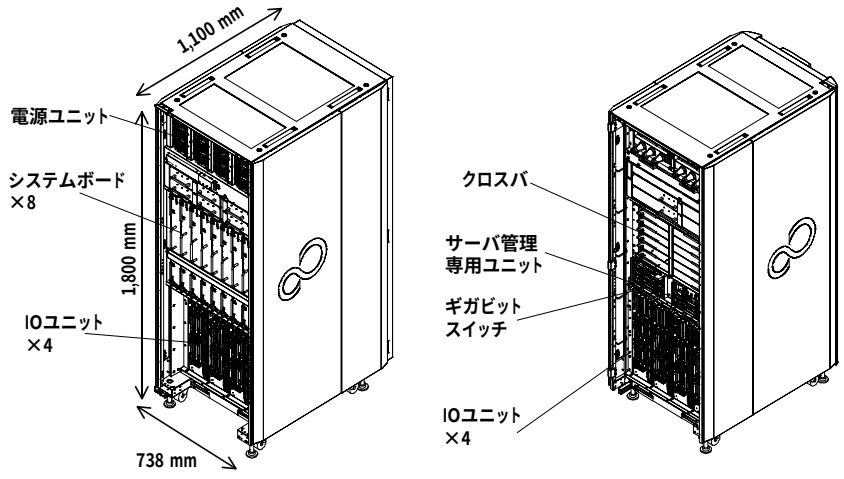
PRIMEQUESTのラインアップは，最大32個のCPUを搭載可能なPRIMEQUEST 480モデルと，最大16個のCPUを搭載可能なPRIMEQUEST 440モデルから成る。OSとしては以下をサポートしている。

(1) Red Hat Enterprise Linux AS（v.4 for Itanium）
(2) Novell SUSE LINUX Enterprise Server 9 for Itanium Processor Family
(3) Windows Server 2003, Enterprise Edition for Itanium-Based Systems
(4) Windows Server 2003, Datacenter Edition for Itanium-Based Systems

PRIMEQUESTの実装を**図**-1に，サーバ全体の論理ブロックを**図**-2に示す。

PRIMEQUEST 480モデルとPRIMEQUEST 440モデルは構造的に同一だが，CPUとメモリを搭載するシステムボード（SB：System Board），HDD（Hard Disk Drive）やPCIスロットを搭載するIOユニットの搭載可能な最大数が異なる。PRIMEQUEST 480モデルには，SBとIOユニットをそれぞれ最大8個，また，PRIMEQUEST440モデルには，それぞれ最大4個搭載可能である。

PRIMEQUESTは図-2の論理ブロック図に示すように，SBとIOユニットをクロスバによって結合す



（a）前面　　（b）背面

図-1　PRIMEQUESTの実装図
Fig.1-Layout of PRIMEQUEST.

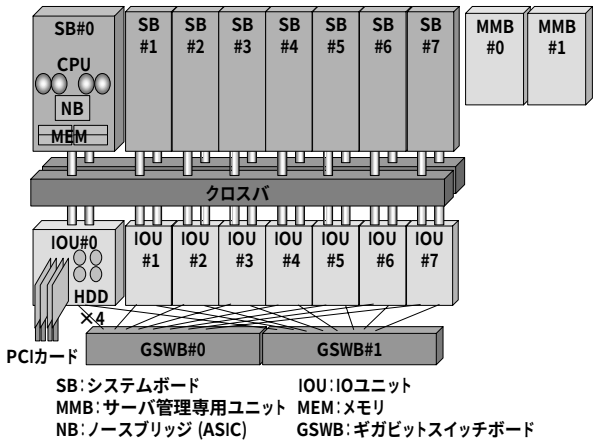


図-2　PRIMEQUESTのブロック図
Fig.2-Block diagram of PRIMEQUEST.

る構成をとっている。このクロスバは後述のミッドプレーンと，アドレス/データ情報の経路制御，伝送を行うASICを搭載したプリント板ユニットから構成されている。

PRIMEQUESTでは，筐体の中央にミッドプレーンを搭載し，このミッドプレーンの両面に各ユニットを搭載する構造を採用している。このミッドプレーン上にクロスバを含む各種高速インタフェースを配線して，ケーブルレスデザインを実現している。

ミッドプレーン上にはSBやIOユニット以外に，筐体内のシステム管理をつかさどるサーバ管理専用ユニット（MMB：Management Board）を搭載している。MMB上にはファームウェアが搭載されており，筐体内の各ユニットの電源制御，筐体内の温度，電圧などの環境監視などのシステム管理を行っている。また，システム管理者に対して管理を行うための操作ビューを提供するためにWebサーバを搭載しており，LAN経由で汎用PCを接続して，そのPC上のWebブラウザを使用してシステムの管理を行うことができる。

なお，システム管理の詳細については本誌掲載の「TCO削減を実現する基幹IAサーバ“PRIMEQUEST”の運用管理」を参照されたい。

さらに，PRIMEQUESTではオプションとしてギガビットスイッチボード（GSWB：Gigabit switchboard）を内蔵可能である。このGSWBには，各IOユニットに搭載されているギガビットイーサネット（GbE）ポートからのLANインタフェースがミッドプレーンを経由して接続されており，筐体外部へLANインタフェースを出力している。

PRIMEQUESTの諸元を**表-1**にまとめる。

Dual Synchronous System Architecture

PRIMEQUEST用に，“Dual Synchronous System Architecture”（2重化同期アーキテクチャ）と呼ぶ新アーキテクチャを開発した。これは，高速同期パラレルバスを構築する技術をベーステクノロジとして，サーバを構成するハードウェアコンポーネントを2重化して同期運転する「システムミラー機構」，高性能な大規模マルチプロセッサシステムを構築する技術の総称である。

このDual Synchronous System Architectureを採用することにより，世界で初めて本PRIMEQUESTクラスの大規模マルチプロセッサシステムでミラー動作を適用することが可能となった。

高速化技術

PRIMEQUESTは，前述のようにミッドプレーン構造を採用して，このミッドプレーンの両面に各ユニットを高密度に実装しており，ミッドプレーン上にはクロスバを構成する高速同期パラレルバスを配線している。

● 高速同期パラレルバス

PRIMEQUEST用に高速同期パラレルバスを開発して，PRIMEQUEST内の主要バスに採用している。本バスは，800 MHz/1.33 GHzの超高速クロックで動作する同期型パラレルバスである。同期型パラレルバスとしては世界トップレベルのスピードを実現している。また，伝送方式としてシングルエンド伝送方式を採用している。シングルエンド伝送方式は差動伝送方式と比較して同一のビット数のデータを伝送するために，半分の信号線しか必要としない。そのため，同一コストのプリント板を使用した場合，倍のビット幅のバスを実装することが可能である。伝送スピードとバスの広ビット幅により，世界トップレベルのバンド幅を実現している。本バスを実現する超高速I/Oインタフェース回路には，富士通の最先端半導体テクノロジであるCS101テクノロジを採用している。PRIMEQUEST用に開発したASICの詳細は本誌掲載の「基幹IAサーバ“PRIMEQUEST”の高性能・高信頼を実現する

表-1　PRIMEQUESTの諸元

モデル	PRIMEQUEST 480	PRIMEQUEST 440
CPU	Itanium2プロセッサ（1.60 GHz/9 Mバイト L3キャッシュ，1.50 GHz/4 Mバイト L3キャッシュ）	
CPU数	最大32CPU（4CPU/SB×8）	最大16CPU（4CPU/SB×4）
メモリ★1	最大512 Gバイト（64 Gバイト/SB×8）	最大256 Gバイト（64 Gバイト/SB×4）
システムボード	最大8個	最大4個
IOユニット	最大8個	最大4個
クロスバ	最大102.4 Gバイト/秒	最大51.2 Gバイト/秒
ディスク	最大147 Gバイト×32台	最大147 Gバイト×16台
PCIスロット	最大128スロット	最大64スロット
GbEインタフェース	最大32ポート	最大16ポート
SCSIインタフェース	最大16ポート	最大8ポート
パーティション分割	最大8	最大4
冗長構成	ディスク，電源，ファン，サーバ管理専用ユニット（MMB），ギガビットスイッチボード★2 クロスバ★2 メモリ★2	
基本筐体外形寸法（mm）	幅738×奥行1,100×高さ1,800	
質量（kg）	720	600
OS（サポート時期ほか）	Red Hat Enterprise Linux AS（v.4 for Itanium）（2005年6月末） Novell SUSE LINUX Enterprise Server 9 for Itanium Processor Family（2005年9月末。海外市場向け） Windows Server 2003, Enterprise Edition for Itanium-Based Systems（2005年9月末） Windows Server 2003, Datacenter Edition for Itanium-Based Systems （2005年9月末）	

★1：標準時。ミラーモード時は最大256 Gバイト×2重化（PRIMEQUEST 480），最大128 Gバイト×2重化（PRIMEQUEST 440）
★2：オプション

チップセット」を参照されたい。

● クロスバ

各SBと各IOユニット間を接続するためにpoint-to-pointクロスバを採用している。クロスバは最大8個のSBと最大8個のIOユニットを接続するために，各ユニット専用のポートを装備している。SB用ポートは最大12.8 Gバイト/秒（＝16バイト×0.8 Gbps）～21.3 Gバイト/秒（＝16バイト×1.33 Gbps）のバンド幅を実現している。クロスバのSB用ポートは全部で8ポートあるので，総合バンド幅は102.4 Gバイト/秒（＝12.8 Gバイト/秒×8）～170（＝21.3 Gバイト/秒×8）Gバイト/秒である。

フレキシブル運用可能技術とスケーラビリティ

PRIMEQUESTでは，システム運用の柔軟性を高めるために，パーティショニング機構，フレキシブルI/O機構を装備している。また，サーバ導入後に業務負荷が増加した場合にも対応できるように高いスケーラビリティを有している。本章ではこれらの機構について述べる。

● パーティショニング機構

PRIMEQUESTは，一つの筐体内のシステムを複数の独立したシステムに分割するパーティショニング機構をサポートする。システムを分割してできた「独立したシステム」を「パーティション」と称する。パーティションの最小の単位はSB1個とIOユニット1個である。PRIMEQUEST 480，440モデルは，それぞれ最大8個，4個のパーティション構成が可能である。

パーティショニング機構によって，異なるOSの混在利用や，本番業務に使用するパーティションと開発業務に使用するパーティションなどの混在利用が可能となり，柔軟なシステムを組むことが可能となる。また，曜日ごとにパーティションの構成を変更することにより，必要最小限のSBやIOユニットを用意すればよく，適切な投資が可能となる。さらに，PRIMEQUESTの大きな特長であるフレキシブルI/O機構を利用することにより，SBとIOユニットの自由な資源配分が可能となる。

● フレキシブルI/O機構

PRIMEQUESTは，任意のSBとIOユニットからパーティションを構成することができるフレキシブルI/Oと呼ばれる機構を装備する。本機構の詳細は本誌掲載の「基幹IAサーバ“PRIMEQUEST”の柔軟性・信頼性を増すフレキシブルI/O機構」を参照されたい。

● スケーラビリティ

拡大する基幹業務に対応するために，PRIMEQUEST 480/440モデルは，それぞれ最大32/16のCPUと，最大512 Gバイト/256 Gバイトのメモリを搭載可能である。また，筐体内に最大32個/16個のPCIスロットを搭載可能であるほか，PCIスロットを追加する場合には，外部にPCI_Boxを接続することにより，最大128/64個まで増設可能である。さらに，各IOユニットに4個のギガビットイーサネットポートを搭載しており，最大32/16ポートのギガビットイーサネットポートを搭載可能である。

このほか，前述のパーティショニング機構を利用することにより，スケールアウト，スケールアップ両方の拡張シナリオに対応することができる。

スケールアウトとスケールアップの例を**図**-3に示す。

AP層（アプリケーション層）のサーバに適用した場合に，AP層の業務負荷が増大した場合はAP層のサーバに割り当てる小さいパーティションの個数を増やすこと（スケールアウト）により対応し，DB層（データベース層）のサーバに適用した場合には，SBやIOユニットを追加して1個のパーティションのサイズを大きくすること（スケールアップ）で対応している。

高信頼性・高可用性技術

24時間365日稼働するミッションクリティカルシステムには，高い信頼性と可用性が要求される。PRIMEQUESTでは，富士通が従来からメインフレームやUNIXサーバで培ってきた各種技術を適用し，「ダウンしないシステム」，「データのインテグリティを保証するシステム」を目標に掲げて開発した。

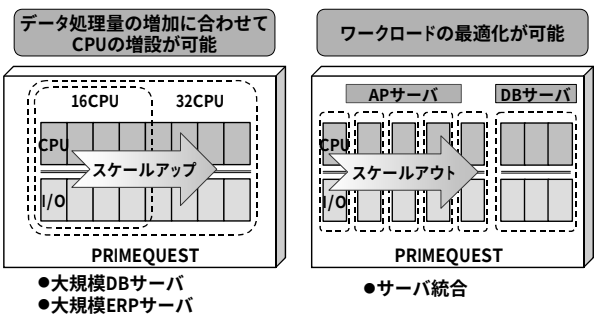


図-3　スケールアウトとスケールアップ
Fig.3-Scale-out and scale-up.

● データインテグリティの保証

PRIMEQUESTでは，処理中にデータ化けが発生していないことを保証するために，各ユニット内部のRAM，データパスには，ECC（Error Correcting Code），またはパリティを付加してデータのインテグリティを保証している。主な例を以下に挙げる。各ASIC内部にある全内蔵RAMには，ECC，またはパリティを付加して，RAMデータの読み出し時にECCチェック，またはパリティチェックを実行してRAMに記憶されていたデータのインテグリティを保証している。

ASIC内部，ASIC間のデータ転送時にもECC，またはパリティを付加してデータ転送時のデータのインテグリティを保証している。とくにクロスバを含むデータ転送径路では，データ転送経路の各ポイントにおいて，ECCチェック，あるいはパリティチェックをすることにより，データのインテグリティを保証するとともに，万が一データエラーが発生した場合には，精度の高いエラー解析を実施して，エラー発生箇所を特定できるようにしている。

● システムミラー機構

PRIMEQUESTは，ほとんどのユニットを2重化して同期運転を実現し，2重化している片系で故障が発生しても残りの系で動作を継続するシステムミラー機構と呼ばれる機構をオプションで装備している。本機構の詳細は本誌掲載の「基幹IAサーバ“PRIMEQUEST”の高信頼システムミラー機構」を参照されたい。

● 各種冗長化機能

PRIMEQUESTは，電源，ファン，HDD，MMBなどほとんどすべてのコンポーネントの冗長構成を可能としている。

(1)　電源供給系

標準で，電源ユニットのN＋1冗長構成を採用している。つまり，電源ユニットが1個故障した場合でも必要な電力は供給可能で業務の運用は継続される。また，オプションで2系統AC受電機構を装備可能なため，AC電源供給元から2重化することが可能である。

(2)　冷却系

筐体内をいくつかの冷却ゾーンに分けており，冷却ゾーンごとに冷却ファンはN＋1冗長化されており，1個の冷却ファンが故障しても残りの冷却ファ

ンで冷却可能な設計をしている。冷却ファンが故障した場合でも，業務を停止することなくその冷却ファンを活性交換可能である。

(3) サーバ管理専用ユニット（MMB）

標準でMMBを2台搭載して冗長化しており，2台のMMBはアクティブ/ホットスタンバイ動作をしている。両MMBは互いに相手側の正常性をチェックしており，ホットスタンバイ側のMMBがアクティブ側のMMBの異常を検出した場合には，ホットスタンバイ動作していたMMBが自動的にアクティブ動作に切り替わる。また，故障した旧アクティブMMBは活性交換可能である。この活性交換の間でも，PRIMEQUESTの各パーティション上で実行されている業務は継続される。同時に，MMBの機能も新アクティブMMBに引き継がれる。

(4) クロック回路

筐体内にシステムクロック発振器を2個搭載している。電源投入時に実行される初期診断時にシステムクロック発振器の正常性を診断して，異常を検出した場合にはもう一方のシステムクロック発振器に切り替えて運用を開始する。

● 活性保守機能

PRIMEQUESTでは，PCIカード，HDD，冷却ファンのホットプラグをサポートしている。

また，SB，IOユニット，GSWBなども一定の条件のもとでホットプラグ可能である。このように，主要なユニットがホットプラグ可能で，システムの運用を止めることなく故障ユニットの交換が可能である。

● 高保守性を実現するケーブルレスデザイン

筐体の中央に搭載したミッドプレーン上にすべての高速インタフェースを搭載して，全ユニットをミッドプレーンの両面へプラグイン実装する徹底的なプラグイン方式を採用している。ミッドプレーンには以下のインタフェースをケーブルレスで実装している。

(1) SB-IOユニット間のクロスバ
(2) IOユニット-GSWB間のギガビットイーサネットインタフェース
(3) IOユニット-MMB間の管理LAN（100メガビットイーサネットインタフェース）
(4) IOユニット-KVM（Keyboard/Video/Mouse）インタフェースユニット間のビデオ，キーボード，マウスインタフェース
(5) MMBと各ユニット間の各種管理用インタフェース
(6) 電源供給ライン

すべての配線をケーブルレスにすることにより，サーバ装置設置時やオプション品増設時などの保守作業の容易化・高信頼化，ケーブル接続ミスなどの人為的ミスの防止に寄与している。また，ケーブル配線の追加，変更などのケーブリングコストの削減，確実な信号接続を実現することにより，信頼性・保守性の向上に寄与している。

以下，ミッドプレーン上に実装されている主なインタフェースの概要を述べる。

(1) GSWBの内蔵

GSWBを使用することでケーブル接続にかかる工数を大幅に削減することができ，運用コストの低減，ケーブル接続数の低減による信頼性の向上，保守操作性の向上が図られる。

筐体内のパーティション構成の変更やパーティションの追加など構成を変更した場合でも，物理的にLANケーブルの接続変更，追加をしなくても，GSWBのVLAN（仮想LAN）定義の変更だけで済むため，運用コストを低減することができる。

GSWBにより外部へ接続されるケーブル本数を大幅に削減することができ，信頼性を向上させることができる。

(2) 管理LANの配線のケーブルレス化とLANスイッチの内蔵

主に業務用に使用するギガビットLAN以外に，運用管理に使用する管理LANの配線もミッドプレーン上に配線しており，管理LANのLANスイッチも内蔵している。そのため，サーバ装置導入後管理LANとしてすぐに使い始めることができ，現地でのセットアップ手番が短縮できるだけでなく，接続ミスや設定ミスを最小化することが可能で，TCO（Total Cost of Ownership）削減に貢献する。例えば，新たにSBを増設しパーティションを追加する場合でも，管理LANに接続されているそのパーティション内のイーサネットインタフェースのIPアドレスの設定以外の作業は不要で，増設作業の最小化が可能となる。

(3) KVMインタフェースユニット

各IOユニット上に搭載されているビデオインタ

フェース，USBインタフェース（3本）は，ミッドプレーンを経由してKVMインタフェースユニットに集線されている。MMBファームウェアの制御により，いずれか一つのIOユニットのビデオインタフェースとUSBインタフェースを選択して，KVMインタフェースユニットの外部コネクタに出力されており，ディスプレイ装置，キーボード，マウスを接続する。

3本のUSBインタフェースのうち1本は，筐体内に搭載されているDVD-ROMドライブに接続されており，これもMMBファームウェアの制御により，いずれか一つのIOユニットのUSBインタフェースに切り替えられ，各IOユニットから共用できる。

むすび

本稿では，PRIMEQUESTの高速化技術，スケーラビリティ，高信頼性・高可用性技術について説明した。

あらゆる業務がコンピュータネットワークに接続されているユビキタスコンピューティング時代の基幹業務を担うため，今後もPRIMEQUESTの高信頼性技術，高可用性技術の拡張に努め，お客様の要求に応える製品を開発，提供していく。

この研究に対して「半導体アプリケーションチッププロジェクト」の一環として助成していただいた経済産業省と独立行政法人 新エネルギー・産業技術総合開発機構に感謝します。