特集

基幹IAサーバ“PRIMEQUEST”の高信頼システムミラー機構

Highly Reliable System Mirror Function of Mission-Critical IA Server: PRIMEQUEST

あらまし

基幹IAサーバ“PRIMEQUEST”で採用したシステムミラー機構（オプション製品）は，ハードウェアコンポーネントを2重化し，2重化したハードウェアコンポーネントをクロック同期で動作させる。この2重化したハードウェアコンポーネントの片側で故障が発生した場合には，ソフトウェアの介入を一切必要とせずに，残りの一方で処理を継続してシステムの可用性を大幅に高める機能である。このシステムミラー機構はハードウェアのみで実現しているので，ソフトウェアには一切手を加える必要がなく，従来のままのソフトウェアを使用して，高い信頼性と可用性を得ることができる。

本稿では，システムミラー機構の概要，システムミラー機構によって実現される利点について説明する。

Abstract

The system mirror function is an option for the PRIMEQUEST of mission-critical IA servers that dualizes the hardware components and operates them in clock synchronization basis. When this function is used and a failure occurs in one side of a dualized hardware component, the function continues processing in the other side. This function, therefore, considerably enhances system availability without the need for software intervention. Therefore it is implemented only via the servers' hardware, highly reliable and available system operation can be achieved without any modifications to the software available in a market.

This paper gives an overview of the system mirror function and discusses the benefits that it provides.

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。

基幹IAサーバ“PRIMEQUEST”の高信頼システムミラー機構

ま　え　が　き

基幹IAサーバ“PRIMEQUEST”では，高信頼性・高可用性を実現する主要な機構としてシステムミラー機構を採用している。システムミラー機構導入の第一の目的は，お客様の業務を停止させないことである。サーバを構成するハードウェアコンポーネントを2重化して同期運転させることにより，一方のハードウェアコンポーネントが故障しても残りのハードウェアコンポーネントで処理を継続して，この間アプリケーションはサービスを停止することなく運用を継続することができる。この仕組みによりサーバの信頼性を大幅に向上させ，可用性を高めることができ，企業ビジネスの基盤となる，止まることが許されないようなシステムへの適用を可能とする。

従来のサーバでもミラー機構を導入しているものがあったが，これらは主にエラー検出機構として導入しているものであった。これらのシステムでは，内部エラー検出能力が十分でない，あるいは全くないマイクロプロセッサを2個導入して，その2個のマイクロプロセッサの出力信号を外部の比較回路で逐次比較していた。両者の出力信号の不一致を検出した場合には，そのプロセッサ・ペアを停止し，OSを停止させて，データ化けを防止していた。ここで業務を停止させないようにするためには，マイクロプロセッサを3個用意して多数決論理を採用したり，マイクロプロセッサ・ペアを2組，つまり4個のマイクロプロセッサを用意する方式を採用したりしていた。

PRIMEQUESTでは，片系のハードウェアだけでも完全なエラープロテクション機能を装備している。このようにハードウェアを2重化して，万が一片側の系でリカバリ不可能なエラー（メモリの多ビットエラーなど）が発生した場合でも，正常に動作している残りの系で動作を継続して，システムとして業務を止めないようにすることを目的としている。

本稿では，PRIMEQUESTで採用したシステムミラー機構の概要について紹介する[1]

なお，PRIMEQUESTではオプション製品としてシステムミラー機構が提供される。

システムミラー機構の概要

一般にサーバシステムは最下位層のハードウェアからファームウェア，OS，ミドルウェア，アプリケーションまでの階層構造で構成されている。クラスタシステムのようにミドルウェア，OSなど上位層の仕組みによって，業務の停止を最小限に抑えることも可能であるが，そのためには上位アプリケーションにクラスタ対応の変更などが必要となる場合がある。それに対して，PRIMEQUESTで採用したシステムミラー機構は，ハード部品（メモリ，クロスバなど）を2重化し，2重構成で処理を同期実行することでハードウェア故障に対する信頼性・可用性を向上している。

この2重化はすべてハードウェア階層だけで実現しているため，OSを含めたソフトウェアの修正を一切必要としない。そのため，お客様はシステムミラー機構を導入することにより，現用のソフトウェアはそのままで一段高い信頼性と可用性を得られる。

また，システムミラー機構を導入したPRIMEQUESTにクラスタシステムを併用することも可能である。

システムミラー機構の仕組み

PRIMEQUESTは，CPUとメモリを搭載したシステムボード（SB）とハードディスクやPCIスロットを搭載したIOユニット，これらのユニットを接続するクロスバから構成される。

SBやIOユニット以外にもギガビットスイッチボード（Gigabit switchboard），MMB（サーバ管理専用ユニット，Management Board）などが存在するがシステムミラーの動作には直接は関与しない。PRIMEQUESTのシステムミラー機構の概要を**図**-1に示す。

システムミラー機構を適用した場合には，メモリ部，SB上のメモリコントローラ部，SB上のチップセット内部，クロスバ部などシステム全体を徹底的に2重化している。以下に各コンポーネントの2重化動作について概説する。

● SB上のコンポーネントの2重化

SB上には，CPUとメモリ，CPUバス，メモリコントローラに接続されるノースブリッジと呼ばれるASIC，メモリコントローラ用ASICを搭載している。システムミラー機構を適用した場合には，「メモリ

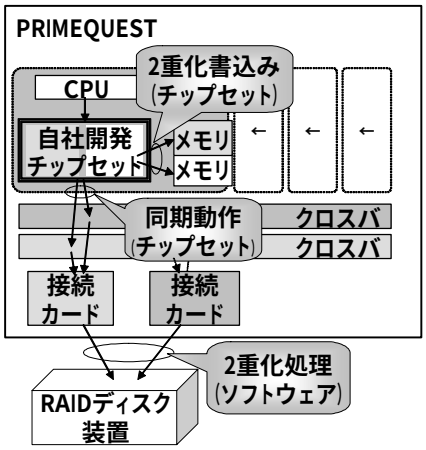

図-1　PRIMEQUESTのシステムミラー機構
Fig.1-System mirror function of PRIMEQUEST.

表-1　メモリミラーのリード時の動作

系0のリードデータ	系1のリードデータ	動作
エラーなし	エラーなし	それぞれの系のデータを使用
	訂正可能エラー	それぞれの系のデータを使用
	訂正不可能エラー	系1も系0のデータを使用して処理続行
訂正可能エラー	エラーなし	それぞれの系のデータを使用
	訂正可能エラー	それぞれの系のデータを使用
	訂正不可能エラー	系1も系0のデータを使用して処理続行
訂正不可能エラー	エラーなし	系0も系1のデータを使用して処理続行
	訂正可能エラー	系0も系1のデータを使用して処理続行
	訂正不可能エラー	システムの緊急停止，リブート処理

コントローラ用ASICとメモリ」,「クロスバとのインタフェース（アドレス部，データ部)」,「ノースブリッジASICの内部」がそれぞれ2重化される。

(1)　メモリコントローラ用ASICとメモリの2重化

SB上にメモリコントローラ用ASICが4個搭載されており，この4個のASICを2個ずつのペアに分割して，ペア間で全く同じ処理を，同じタイミングで実行する。また，それぞれのペアに接続されたメモリ同士も2重化され，全く同じデータを同じタイミングでライト/リードする。メモリリードオペレーションの場合，両方のメモリの同じロケーションからデータをリードする。両方のメモリは，それぞれECC（Error Correcting Code）が付加されているので，それぞれのペアでエラー検出，エラー訂正が可能である。この場合，**表-1**に示すように両方のデータで訂正不可能なエラーが検出された場合のみ，処理を継続できない。そのほかの場合には，少なくとも一方の系で正しいデータを取得できるので，システム処理を継続することが可能となり，可用性を大幅に向上することができる。

一般にサーバシステムの場合，ストレージに次いでメモリ部分が最も故障の発生確率が高い部分であり，メモリのミラー化はシステム全体の可用性を上げるためには，効果的な部分である。

(2)　クロスバとのインタフェースの2重化

SBは，メモリコントローラ用のASICを経由してクロスバのデータ部に，ノースブリッジASICを経由してクロスバのアドレス部に接続される。

メモリコントローラ用のASICは上述のように2ペアで2重化動作しているので，クロスバのデータ部も2重化されることになる。

ノースブリッジASICは後述のように，内部で2重化しており，クロスバのアドレス部も2重化されている。

(3)　ノースブリッジASICの内部の2重化

ノースブリッジASICには，CPUバスとのインタフェース部，メモリコントローラ部とのインタフェース部を搭載している。これらのインタフェース部は，ノースブリッジASICの内部で2重化されており，完全にクロック同期で同一処理を実行している。

そのため，1重のハードウェアでは不可能であったASIC内部の制御系回路の故障検出，リカバリ処理を実現している。

●　クロスバの2重化

クロスバは，大きくアドレス部とデータ部から構成され，アドレス部とデータ部それぞれが2重化される。クロスバ上では，以下のアクセスが処理される。

(1)　CPUからのメモリライトアクセス，メモリリードアクセス
(2)　CPUからのIOユニット内リソースへのライトアクセス，リードアクセス
(3)　IOユニットからのメモリライトアクセス，メモリリードアクセス

いずれのアクセスを処理する場合でも，アドレス情報がクロスバのアドレス部に送出され，続いてデータがクロスバのデータ部に送出される。その際,

2重化されているクロスバの両系に同じ情報が同一タイミングで送られ，両系で同じ処理が行われる。クロスバの両系は，それぞれの系に閉じた範囲で，パリティ，ECCなどのエラープロテクション機能を備えており，それぞれの系のクロスバ内で情報を転送する節目，節目でエラーチェックを行っている。情報転送の途中で，訂正不可能なエラーを検出した場合には，もう一方の系の情報を使用して処理を継続していく。この間の動作は，途中でエラーが検出されても，されなくてもソフトウェアには完全にトランスペアレントで処理が行われる。つまり，ソフトウェアは一切関知する必要がない。

PRIMEQUESTを構成する主要コンポーネントには，SBやクロスバ以外にIOユニットがあるが，IOユニットはクロスバのインタフェース部を2重化している。IOユニットの下流にはPCIバスが接続され，これらは1重である。IO系は，LANや，FC（Fibre Channel）のPCIカードを複数枚搭載して，ミドルウェア階層のソフトウェアを使用して冗長化する。

なお，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・高信頼を実現するチップセット」を参照されたい。

システムミラー機構によって実現される利点

以上述べたシステムミラー機構によって，お客様の業務にもたらされる利点についてまとめる。

(1) 従来のミラー化していないシステムと比較して，エラープロテクションのレベルが大幅に向上しており，可用性が向上している。そのため，お客様は，サーバ（ハードウェア）の故障時でも業務の停止を回避することができる。

(2) システムミラー機構は，純粋にハードウェア階層で実現しており，ソフトウェアからは完全にトランスペアレントで，ソフトウェアの対処は一切不要である。そのため，お客様の業務で必要とするすべてのミドルウェア，アプリケーションがシステムミラー機構による効用を享受でき，業務の可用性を高めることができる。

いずれの利点も，システムに故障が発生した場合でも，お客様のビジネス機会損失の回避に加え，お客様の業務の運用スケジュールに合わせて保守時間を調整できる柔軟な運用を可能とする。

むすび

本稿では，PRIMEQUESTで導入したシステムミラー機構について説明した。

お客様の基幹情報システムに安心して導入いただけるサーバを提供できるように，今後も更にPRIMEQUESTの信頼性，可用性を向上するテクノロジを開発して，適用するよう努めていく。将来に向けて，さらなる高信頼性，高可用性の向上を追求していき，あらゆる企業，個人がネットワークで結ばれるユビキタス社会のIT基盤の中核をなすサーバを提供していく。

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。

参考文献

(1) 浜田王才：基幹IAサーバ“PRIMEQUEST”の高信頼システムを実現する技術．*FUJITSU*，Vol.56，No.3，p.194-200（2005）.