特集

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

Flexible IO Improving Flexibility and Reliability of Mission-Critical IA Server “PRIMEQUEST”

あらまし

基幹IAサーバ“PRIMEQUEST”では，CPUとメモリを搭載したシステムボード（SB）と呼ばれるユニットと，HDD（Hard Disk Drive），PCIスロットなどのI/O関連ブロックを搭載したIOユニット（IOU）を物理的に分離して実装し，両ユニットをクロスバで結合した構造を採用している。この構造により，任意のSBと任意のIOユニットを組み合わせてパーティションを構成することが可能である。つまり，任意のIOユニットは任意のSBと組み合わせることが可能であり，この機能を「フレキシブルI/O機構」と称している。このフレキシブルI/O機構を利用することにより，種々の利点を実現することが可能となる。

本稿では，フレキシブルI/O機構の概要と，フレキシブルI/O機構によって実現される利点を紹介する。

Abstract

The PRIMEQUEST of mission-critical IA servers consists of the system board, which mounts the CPU and memory, and the IO unit (IOU), which mounts IO-related blocks such as hard disk drives and PCI slots. The system boards and IOUs are physically independent of each other and are interconnected by a crossbar. This feature is referred to as the “flexible IO” because it enables partitions to be configured by combining any system boards with any IOUs. This paper gives an overview of the flexible IO and discusses some of 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。

まえがき

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，柔軟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機能を装備している。

本稿では，PRIMEQUESTの特長の一つであるフレキシブルI/O機構について説明する。最初にフレキシブルI/O機構の概要について述べ，つぎにフレキシブルI/O機構によって実現される利点について例を挙げて説明する。

なお，本稿では，PRIMEQUEST 480モデルを例に挙げて述べるが，筐体内に搭載可能なシステムボード（SB）やIOユニット（IOU）の数が異なる以外，PRIMEQUEST 440モデルについても同様である。

フレキシブルI/O機構の概要

PRIMEQUESTは，本誌掲載の「基幹IAサーバ“PRIMEQUEST”の高信頼システムを実現する技術」で述べたように，クロスバにより，CPUとメモリを搭載したSBと呼ばれるユニットと，HDDやPCIスロットを搭載したIOUから柔軟な構成を持つパーティションを構成することができる。

1台のPRIMEQUEST 480モデルでは，最大構成時SB#0からSB#7までの8個の独立したSB群，IOU#0からIOU#7までの8個の独立したIOU群で構成されており，このSBとIOUは物理的に分離したユニットで構成されている。これらSB，IOUとそれらを結合するクロスバで，PRIMEQUESTを構成している。

クロスバによりフレキシブルI/O機構を実現している。フレキシブルI/O機構によりSB#0～SB#7から選択された任意個のSB群と，IOU#0～IOU#7から選択された任意個のIOU群から，1個のパーティションを構成することができる。

パーティションの構成例を**図**-1に示す。

パーティション#Aは同数のSBとIOUから構成される例，パーティション#Bとパーティション#Cは異なる個数のSBとIOUから構成される例である。いずれもフレキシブルI/O機構を利用した例で，柔軟なパーティション構成を定義することが可能となる。

図-2に示すように，従来のサーバでは，一つの物理ユニットにCPU，メモリ，I/Oインタフェース部を搭載していた。このような構成のシステムでパーティションを構成する場合，以下のような問題が発生していた。

CPU，メモリ，I/Oインタフェース部が一つのユニットになっていたため，CPU資源やメモリ資源を多く必要とし，I/O資源を多く必要としないパーティションにも，不必要なI/Oインタフェース用の資源を割り当ててしまっていた（図-2左のパーティション#A）。また逆に，I/O資源を多く必要とするパーティションに，不必要なCPU資源，メモリ資源を割り当ててしまうこともある（図-2左のパーティション#B）。そのため，無駄な投資を強いられるという欠点があった。

それに対して，PRIMEQUESTではフレキシブルI/O機構を装備しているので，CPU資源やメモリ資源を多く必要とするパーティションにはSBを多く割り当て，I/O資源を多く必要とするパーティションにはIOUを多く割り当てることが可能となる（図-2右）。

このように，フレキシブルI/O機構の採用によってハードウェア資源の効率的な活用（コスト最適化）を図り，無駄なハードウェア資源を最小限に抑え，最適なパーティションを構成することが可能となり，必要なSBとIOUを装備するために最小限の投資で済むことになる。

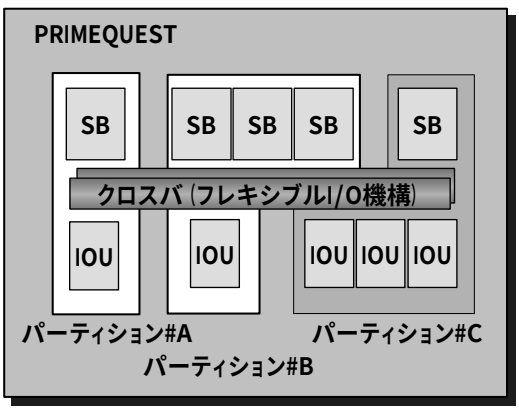

図-1　PRIMEQUESTのパーティション構成
Fig.1-Partition configuration of PRIMEQUEST.

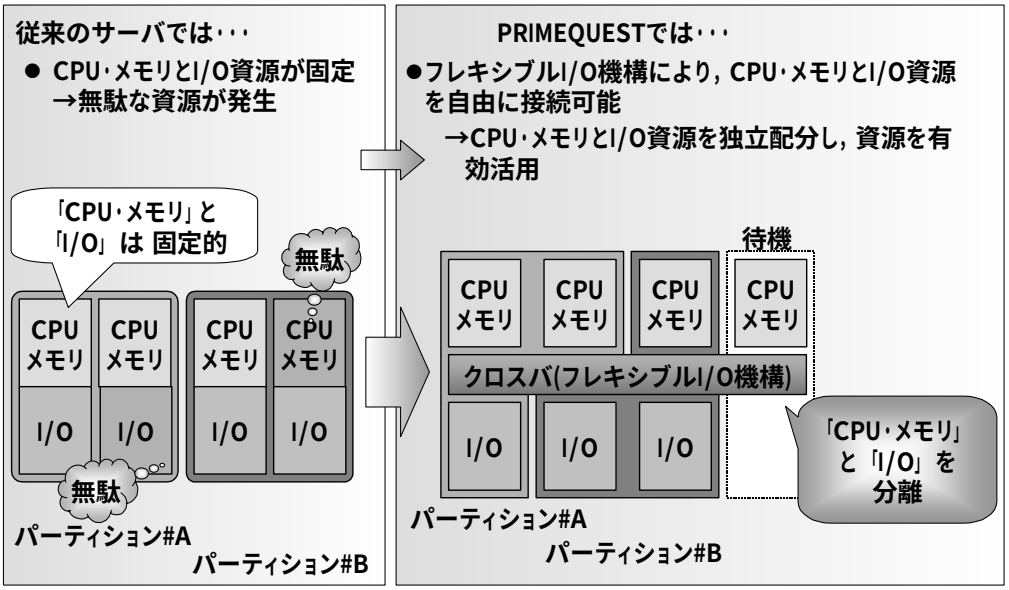

図-2　PRIMEQUESTと従来サーバのパーティション構成
Fig.2-Partition configurations in PRIMEQUEST and conventional server.

フレキシブルI/O機構の利点

フレキシブルI/O機構により，柔軟なパーティション構成を組むことができ，この機構を使用してフローティングSB運用を実現することができる。ここでは，いずれのパーティションの構成にも組み込まれていないSBをフローティングSBと称する。

以下にフローティングSBによって実現される利点を挙げる。

（1） SB故障時の高速切替え
（2） 業務負荷変動時のパーティション再構成
（3） 故障の予兆監視とパーティション再構成

本章では各利点について説明する。

● SB故障時の高速切替え

PRIMEQUESTは，故障したSBを自動的に切り離し，あらかじめフローティングSBとして設定しておいたSBを組み込み，パーティションを再起動する機能をサポートする。フローティングSBがない場合には，あるSBの故障によりそのSBが属するパーティションが停止し，再起動される。再起動時に故障したSBをパーティションの構成から切り離すことで業務を再開できるが，SBの数が減少するため，パフォーマンスの低下が発生する。

フローティングSBが設定されている場合には，パーティションの再起動時にフローティングSBを自動的に組み込んで，パーティションを再立上げする。これにより，そのパーティションのSBの数が維持される。つまり，CPUの数やメモリの容量が維持されるため，パフォーマンスの低下を防ぐことができる。

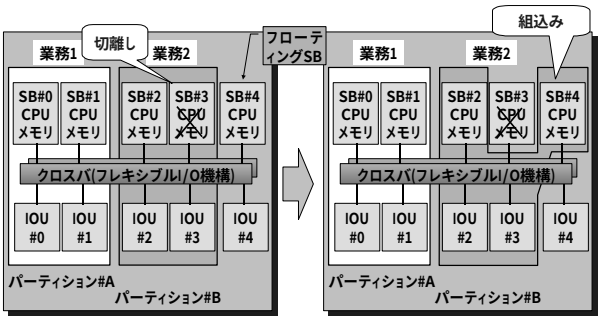

図-3　SB故障時の高速切替え
Fig.3-Rapid reconfiguration on SB failure.

SB故障時の高速切替えの例を**図**-3に示す。

図-3では，SB#0～#1とIOU#0～#1でパーティション#Aを，SB#2～#3とIOU#2～#3でパーティション#Bを構成しており，SB#4がフローティングSBとして設定されている。SB#3が故障した場合，パーティション#Bが停止し，サーバ管理専用ユニット（MMB）ファームウェアの指示によりパーティション#Bが再起動される。この再起動の過程で，故障したSB#3をパーティション#Bの構成から外し，続けて，代わりにフローティングSB#4をパーティ

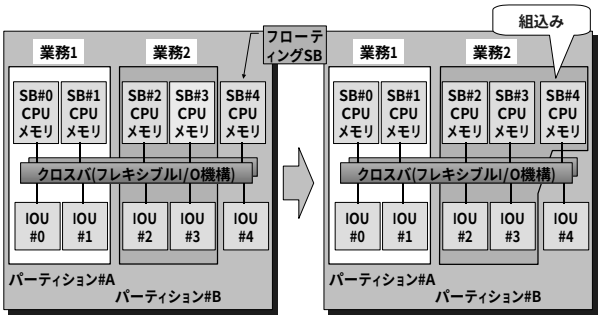

図-4　業務負荷変動時のパーティション再構成
Fig.4-Partition reconfiguration at workload change.

ション#Bに組み込んでパーティション#Bを再起動する。このプロセスをMMBファームウェアからの指示で実施し，SB#3の故障の発生前後で，パーティション#Bを構成するCPUやメモリの資源量を維持することができる。このように，フローティングSBを備えておくことにより，SBの故障が発生しても短時間で故障SBをパーティションから切り離し，フローティングSBを組み込むことができ，システムのダウン時間を最小限にすることが可能となる。

●　業務負荷変動時のパーティション再構成

あるパーティションで処理している業務の負荷が増大した場合，フローティングSBをパーティションに組み込むことにより，業務負荷の変動に対応することができる。

業務負荷変動時のパーティション再構成の例を**図**-4に示す。

図-4では，SB#0～#1とIOU#0～#1でパーティション#Aを，SB#2～#3とIOU#2～#3でパーティション#Bを構成しており，SB#4がフローティングSBとして設定されている。

パーティション#Bの業務負荷が増大してきたと仮定する。このとき，フローティングSB#4をパーティション#Bに組み込むことにより，CPUやメモリ資源を追加することができる。これにより，パーティション#Bの処理能力を増大させ，業務負荷の増大に対処することができる。

このように，フレキシブルI/O機構を使用することにより，業務負荷の変動に対処することが可能となる。

●　故障の予兆監視とパーティション再構成

サーバで発生するハードウェアの故障は，大きく二つに分類することができる。一つ目は各種バス上で伝送されるデータやメモリに格納されたデータの1ビットエラーに代表される訂正可能なエラーである。この場合には，ハードウェアに装備されているECC（Error Correcting Code）回路において，ハードウェアが自動的に1ビットエラーを訂正し，アプリケーションの実行には影響を与えない。二つ目は，データの多ビットエラーに代表される訂正不可能なエラーである。この場合は，ECC回路で多ビットエラーが発生したことを検出するが訂正不可能である。ハードウェアは，訂正不可能なエラーが発生したことをOSに通知し，OSは緊急停止してデータの破壊を防いでいる。

このように，訂正可能なエラーが発生しているうちは，アプリケーションの実行には影響を与えないので，業務が停止することはない。訂正可能なエラーがある程度の頻度以下で発生している間は問題ないが，そのうち多ビットエラーに進展する確率が上がってくる。そこで，データの1ビットエラーなどの訂正可能なエラーがある程度以上の頻度で発生していることを検出した場合には，将来訂正不可能な多ビットエラーが発生する確率が高いと予測することができる。このような処理を「故障の予兆監視」と呼ぶ。

この故障の予兆監視機構とフローティングSBを組み合わせることにより，システムの可用性を高めることができる。

故障の予兆監視とパーティション再構成の例を**図**-5に示す。

図-5では，SB#0～#1とIOU#0～#1でパーティション#Aを，SB#2～#3とIOU#2～#3でパーティション#Bを構成しており，SB#4がフローティングSBとして設定されている。

パーティション#Bを構成しているSB#2で，あるデータの1ビットエラーが頻発した場合には，予兆監視機構により将来多ビットエラーに進展する可能性が高いと判断できる。

データの1ビットエラーが多ビットエラーに進展しないうちに，その故障しかかっているハードウェア資源を搭載しているSB#2をパーティション#Bの構成から外して，フローティングSB#4をパーティション#Bに組み込むことにより，多ビットエラーの発生を未然に防ぎ，パーティションのダウンをあらかじめ防止することが可能となる。

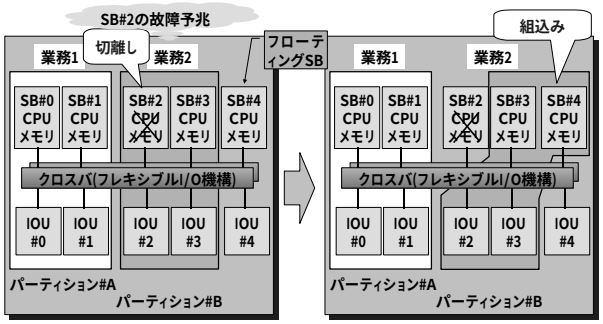

図-5　故障の予兆監視とパーティション再構成
Fig.5-Fault prediction and partition reconfiguration.

以上，フローティングSBの利点を三つ挙げたが，いずれもフレキシブルI/O機構を装備することにより，任意のSBをフローティングSBとして設定することができ，柔軟な運用が可能となる。

む　す　び

本稿では，PRIMEQUESTの特長の一つであるフレキシブルI/O機構について述べ，柔軟性に富んだパーティション構成が可能で，業務に必要な最小限のハードウェア資源を搭載することができ，TCO（Total Cost of Ownership）の削減につながること，フローティングSBを使用することにより，ダウンタイムを最少にできることを述べた。

インターネット時代を担う情報システムには，24時間365日の運用に耐え得る高い信頼性と，急激なワークロードの変動に対応できる柔軟性が要求される。フレキシブルI/O機構は柔軟性に富んだシステムを実現するためにキーとなる機能である。今後もフレキシブルI/O機構の機能の充実を図り，より運用管理のしやすいシステムをお客様に提供していく。

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。