特集

基幹IAサーバ“PRIMEQUEST”とPRIMECLUSTERの連携

Linkage of PRIMECLUSTER and Mission-Critical IA Server “PRIMEQUEST”

あらまし

ユビキタス時代のITシステムは，大規模な負荷変動への対応や24時間365日連続稼働への要求が高く，基幹IAサーバ“PRIMEQUEST”はこのようなITシステムの中核サーバとして，「オープン，ミッションクリティカル，グローバル」の三つのキーワードを基本に開発された。また，PRIMECLUSTERは，サーバ，ストレージ，ネットワークの冗長構成を活用して可用性を向上し，連続稼働時間を最大化する基盤ソフトウェアであり，富士通製ミドルウェア，サーバ，ストレージとの組合せで，他社との競争力を強化してきた。このPRIMEQUESTとPRIMECLUSTERの連携に，これまで培ってきたUNIXサーバの高信頼化技術を結集し，オープンシステムで最高の信頼性と連続稼働時間を実現する。

本稿では，PRIMEQUESTとPRIMECLUSTERの連携と高信頼化テクノロジについて紹介する。

Abstract

Information Technology (IT) systems for today's ubiquitous age must address the urgent feature requirements for flexibly accommodating changes in large-scale workload and for 24/7 continuous operation. The “PRIMEQUEST” mission-critical IA server was developed as a core server in line with “open,” “mission-critical,” and “global” as the keywords. PRIMECLUSTER is basic software designed to maximize continuous operation time by enhancing availability through a redundant server, storage, and network configuration. PRIMECLUSTER has extended its competitiveness by incorporating a combination of Fujitsu's advanced middleware, servers, and storage. By combining Fujitsu's vast experience in high-reliability technology in collaboration with PRIMEQUEST and PRIMECLUSTER, we can provide cluster systems offering the highest reliability and longest continuous operation time in an open system. This paper introduces the technology employed and approach taken to achieve maximum availability through the linkage of PRIMEQUEST and PRIMECLUSTER.

酒井　勝（さかい　まさる）

自律システム基盤開発統括部 所属
現在，PRIMECLUSTERの開発推進に従事。

まえがき

ミッションクリティカルシステムは，従来，メインフレームによる金融勘定系システムに代表される限られたものであったが，ユビキタス時代の到来とともにオープンシステムによるミッションクリティカルシステムが拡大，一般化し，24時間365日連続稼働への要求が高まっている。

PRIMECLUSTER(1)は，富士通が得意とするミッションクリティカルシステムへのこだわりとして，ミドルウェア，UNIXサーバ，ストレージの組合せで，高可用性技術を培ってきた。可用性を向上させるために，サーバを複数台使用して，冗長化することによりシステムの停止時間を最小限に抑えるクラスタシステムが有効であり，クラスタシステムにおいては，サーバの状態を正確かつ迅速に認識し，高速に切り替える動作が必須となる。

富士通は，基幹IAサーバ“PRIMEQUEST”とPRIMECLUSTERの組合せにおいて技術を結集し，オープンシステムで最高の連続稼働を実現する。

本稿では，PRIMEQUESTとPRIMECLUSTERの連携と高信頼化テクノロジについて紹介する。

連続稼働への課題

24時間365日連続稼働への要求に関しては，業務サービスが停止しないことと，万一停止しても短時間のうちに復旧することが必要である。クラスタシステムはハードウェア故障など不慮の停止に備えてサーバを冗長構成にしておく技術であり，サーバや業務の切替えを，迅速かつ確実に実行することが要件となる（**図**-1）。

クラスタシステムに求められる要件を下記に示す。

（1） 異常を迅速に検出すること

（2） 確実に切り替えること（2重起動防止）

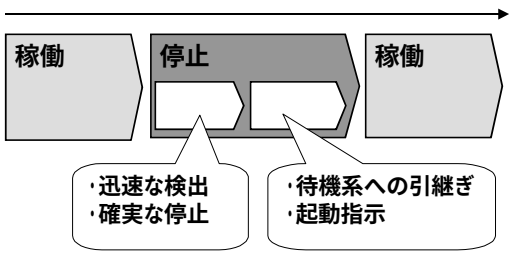

図-1　高速切替えの機能要件
Fig.1-Function requirement of fast switching.

（3） 切替え機構そのものの信頼性が高いこと

この切替えにおける停止時間が短いほど可用性が高い。

迅速に検出し通知するためには，OS停止をイベントドリブンで通知する方法があるが，ソフトウェアだけでは限界があるため，ハードウェアやファームウェアとの共同開発で解決していく必要がある。例えば，サーバ停止の検出には，ソフトウェアの通信を繰り返す“ハートビート”を使用するのが一般的であるが，ハートビート間隔を短くするとシステム負荷が高い場合にも検出してしまうため数十秒間隔で通信することになる。そのため，OSがパニックに陥りサーバが停止した場合に，検出するまで数十秒かかることになる。

従来のIAサーバには，OSの停止を検出/通知するハードウェアが実装されていなかったが，PRIMEQUESTではこのハードウェア機構を実装することで解決した。

PRIMEQUESTとの連携

クラスタシステムとしての完成度を高めるために，PRIMEQUESTの設計段階から連続稼働を実現するための課題解決に取り組み，ハードウェア・ファームウェア・LinuxカーネルとPRIMECLUSTERとの連携で，高可用性・高信頼性を実現した。

● サーバ管理専用ユニット（MMB）による高速切替えの実現

PRIMEQUESTは，システムボード（OS）と独立したMMBが，OSの状態を監視してクラスタソフトウェアに通知する機能，またクラスタソフトウェアからの指示で強制的にOSを停止させる機能を実現している（**図**-2）。これにより，PRIMECLUSTERとの連携で，迅速で確実なサーバ切替えが可能になった。

一般的なオープンシステムのクラスタソフトウェアは，ソフトウェアのハートビート処理を使い，無応答になることでOS停止を検知するために，切替えの初動はハートビート間隔に依存し数十秒かかるが，MMBからの通知は1秒程度である。

さらに，ハートビートとMMB通知の併用によって，それぞれの故障が判別できるため，より正確な故障状態が認識でき，信頼性を向上させている（**図**-3）。以下に判別の動作を述べる。

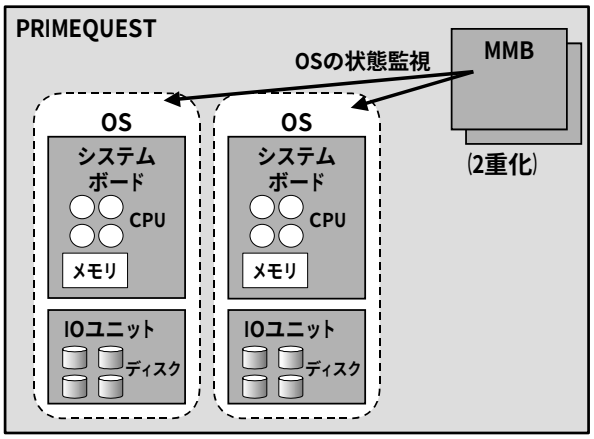

図-2　MMBとシステムボードの構成
Fig.2-Configuration of MMB and system board.

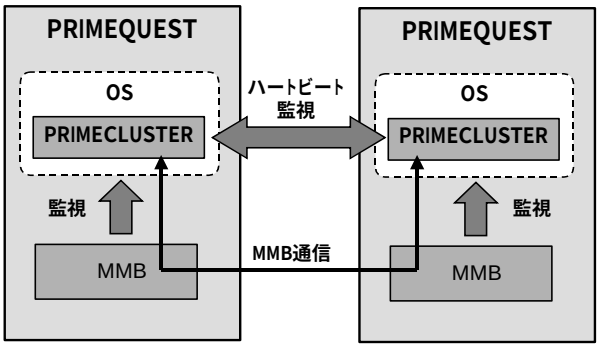

図-3　MMBとPRIMECLUSTERの連携
Fig.3-Linkage of MMB and PRIMECLUSTER.

(1)　ハートビートの切断で，MMBがOS稼働を認識している場合は，ハートビート通信経路故障か通信処理のハングアップなどが発生している。
(2)　MMB通信がエラーの場合は，ハートビート通信経路を使用して状態を確認し，正常ならMMBが故障している。

なお，PRIMEQUESTではMMBおよびクラスタソフトウェアへのインタフェースは完全に2重化しており，信頼性を高めている。

●　高速切替えを実現するI/Oフェンスの実装

サーバ切替えで待機側が起動できる条件は，確実に引継ぎ資源が停止していることである。共用ディスクやIP切替えなど資源が確実に停止していないと，引継ぎ後のアクセスで，双方のサーバから同時アクセスが発生してデータ破壊を引き起こすという重大問題が発生する可能性がある。

このため，PRIMEQUESTのOSやドライバ，ファームウェアは，OSパニックなどの停止トリガに対してI/Oを瞬時に停止するI/Oフェンス機能を実現している。MMBが他サーバのクラスタソフトに異常を通知した時点で，I/Oフェンスが保証されていることで，速やかに待機側の起動が可能となる。

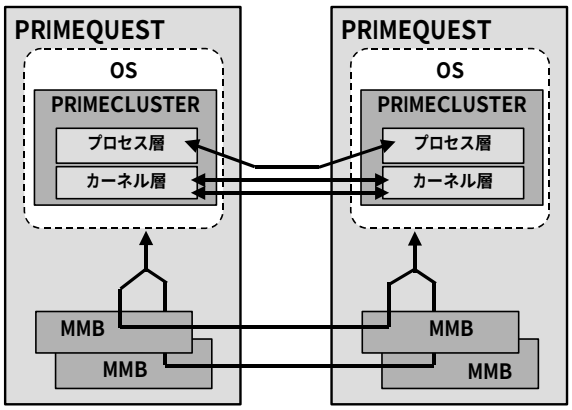

図-4　PRIMEQUESTとPRIMECLUSTERの2重化構造
Fig.4-Duplicated configuration of PRIMEQUEST and PRIMECLUSTER.

●　2重化構造の追求

クラスタ切替機構そのものの信頼性が失われると，万一に備えたクラスタ構成であるにもかかわらず，クラスタ切替えが失敗して業務が停止するという，甚大な損害が発生することになる。

このため，**図**-4に示すようにPRIMEQUESTのMMBとその通信経路は完全に2重化されており，PRIMECLUSTERのノード間通信経路は2重化し，かつカーネル層とプロセス層の2層で通信する構造になっている（プロセス層はカーネル層の2重化された通信経路を使用）。

PRIMECLUSTERの概要

PRIMECLUSTERはサーバ，ストレージ，ネットワークの冗長構成により可用性を向上させ，システムの稼働時間を最大化する高信頼性基盤ソフトウェアである。

●　サーバの冗長化

クラスタを構成するサーバ間で，LANを使用したハートビート監視を行い，応答がなくなった時点で待機サーバに切り替え，業務を引き継ぐ。また，ハードウェアの監視機構を併用することで高速切替えと信頼性を確保している（**図**-5）。

PRIMECLUSTERではサーバの状態を監視する

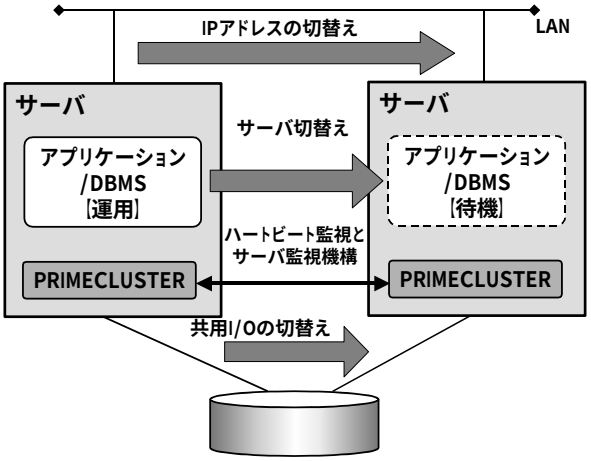

図-5　PRIMECLUSTERの概要
Fig.5-Outline of PRIMECLUSTER.

ために，ハートビートとハードウェアのサーバ監視機構を併用することが可能であり，サーバの故障検出時間の短縮である高可用性と，安全で確実なサーバ切替えという高信頼性を実現している。以下にPRIMECLUSTERの特長を述べる。

(1)　同期監視と非同期監視

複数の経路を使用した監視と，ハードウェアに搭載されたサーバ監視機構との組合せで，高速かつ高精度なサーバダウン検出と安全で確実なサーバ切替え機能を実現している。

・同期監視

専用LAN（クラスタインタコネクト）の経路を使用し，カーネル空間，ユーザ空間の2重の経路を使用し，定周期のハートビート監視でサーバダウン/ハングアップなどの異常を確実に検出する。

・非同期監視

PRIMEQUESTではMMBがサーバダウンを検出すると，専用経路を通じてPRIMECLUSTERに即時通知する。

(2)　故障サーバの切離し

ハートビート応答のないハングアップしたサーバを，ハードウェアのサーバ監視機構を利用して強制停止させることにより，安全で確実なサーバ切替えが可能である。

(3)　ホットスタンバイ

運用系の異常発生後に，待機系でデータ引継ぎや業務アプリケーションの起動を行う一般的なスタンバイ方式とは異なり，待機系で事前に業務再開の準備を整えておくホットスタンバイと呼ばれる方式をサポートしている。これに対応する富士通のデータベース管理ソフトウェアSymfoware Serverやアプリケーション基盤ソフトウェアInterstageは，待機系でデータベース管理ソフトウェアの起動，共用ディスク装置の事前オープン，そしてアプリケーションの起動までを完了し，即座に業務処理を再開できる状態で待機することで，業務再開までの時間を大幅に短縮している。

(4)　パトロール診断（待機パトロール）

一般的な切替え型クラスタシステムでは，運用系のサーバ，ネットワーク，ストレージなどの監視が行われるが，待機系の監視は行われていない。このため，待機系にサーバを切り替えても，待機系の故障により，業務が全面停止する可能性がある。

PRIMECLUSTERでは，こうした最悪の事態を防止するために，「待機パトロール」機能を提供している。待機系においても，サーバ，ネットワーク，ストレージ，業務アプリケーションの監視を行い，業務引継ぎの失敗を未然に防止する。待機系に異常が発生した場合は，故障箇所を切り離し，PRIMECLUSTERのコンソールにアラームが通知され，システム管理者に対処を促すことができる。

●　ストレージとシステムディスクの冗長化

PRIMECLUSTER GDSは，物理ボリュームをミラーリングして論理ボリュームとして上位ソフトウェアに提供する。

(1)　システムボリュームのミラーリング

システムボリュームのディスクが故障した場合，故障ディスクの交換およびリストア処理で長時間のシステム停止が余儀なくされるが，システムボリュームをミラーリングすることで，故障ディスクを切り離し，正常なディスクで運用を継続することができる。また，故障した装置の代替ミラーを自動的に生成するホットスペア機能も備えており，交換が遅れた場合でもシステムの可用性を維持することが可能である。

(2)　RAID装置間のミラーリング

ストレージの可用性を向上させる目的でRAID装置が広く採用されているが，ミッションクリティカルシステムでは更に高いデータの可用性が求められる。RAID装置を更に冗長構成し，RAID装置間でミラーリングすることで，高レベルなデータアクセスの継続性を実現している。

● ネットワークの冗長化

PRIMECLUSTER GLSは，複数のIPアドレスを一つの仮想IPアドレスとして上位ソフトウェアに提供する。

(1) NIC切替え方式（異機種/マルチベンダ環境）

NIC（Network Interface Card）切替え方式では，2重化したNICを同一ネットワークに接続し，排他使用して伝送路の切替えを制御する。通信相手が限定されないため，多くのサーバおよびネットワーク機器ベンダ製品との通信が可能である。

(2) 高速切替え方式

正常運用時は複数の伝送路を並列に使用することで，帯域を拡大する。また，監視用プロトコルの使用により故障を即時検出し，故障した伝送路を高速に切り離すことで，アプリケーションにネットワーク切断を意識させることなく業務を継続することができる。

● 計画停止時間の最小化による高可用化

情報システムの稼働時間を最大化するためには，システムトラブル発生から業務再開までの計画外のシステム停止時間を最小化するだけでなく，システム保守・増設時の計画的なシステム停止時間も最小化する必要がある。

従来の情報システムでは，計画的に業務を停止し，システム保守やアップグレードを行ってきた。しかし，24時間365日連続稼働が要求されている今日の情報システムにおいては，こうした停止を伴う保守は許されない状況となっている。

PRIMECLUSTERシステムでは，以下の機能により，計画停止時間を最小化することが可能である。

(1) 活性交換

冗長化されたネットワークまたはストレージは，故障箇所を自律的に業務から切り離すことにより，業務に影響を与えることなく交換作業および冗長化状態への復旧が可能である。

(2) ローリングアップデート

クラスタシステムを構成するサーバを1ノードずつ順番に停止し，ほかのノードに業務を引き継ぎながらハードウェアおよびソフトウェア保守を行うことで，保守時の業務停止時間を最小化することができる。

(3) 活性増設

ビジネスの急激な拡大に伴い，処理能力やファイルシステム容量が不足した場合は，サーバ追加による処理能力増強や，オンラインでのファイルシステム容量の拡張が可能である。また，PRIMEQUESTやETERNUS[(2)]の活性増設機能により，CPU，メモリ，ディスクを増設することも可能である。

今後の取組み

PRIMECLUSTERとPRIMEQUESTは，ミッションクリティカル分野へのこだわり，高信頼性への追求とともに，クラスタシステムだけでなく，IT基盤「TRIOLE」[(3)]への対応を図っていく。

TRIOLEは，複雑化したシステム資源を可視化（構成，故障箇所，性能などをセンタから一元的に把握）し，自律制御（必要に応じてサーバ資源などを自動的に再配置）する機能を展開しており，クラスタ技術で実現してきた，故障検出や業務引継ぎの機能を，TRIOLEの可視化や自律制御のコアテクノロジとして活用し発展させていく。

む　す　び

PRIMECLUSTERはPRIMEQUESTと連携してオープン系基幹サーバによるクラスタシステムの高可用性と高信頼性を実現した。

ユビキタス時代を迎え，ITシステムの高信頼性や最適化・運用性がますます求められている。PRIMECLUSTERおよびPRIMEQUESTは，富士通のDNAである高信頼性への追求を継承し，お客様の期待や信頼に応えるために，TRIOLEを軸にして更に発展させていく。

参 考 文 献

(1) 阿部敏浩：UNIXサーバソフトウェア：PRIMECLUTER. *FUJITSU*, Vol.53, No.6, p.456-462（2002）.

(2) 松本一志ほか：経営資源としてのデータの安全・柔軟な運用を支えるストレージ：ETERNUSとSoftek. *FUJITSU*, Vol.56, No.1, p.41-46（2005）.

(3) TRIOLEホームページ. http://triole.fujitsu.com/jp/