高速レスポンス，高スループットを実現する超高速データ管理ミドルウェア

Ultra-high-speed In-memory Data Management Software Achieving High-speed Response and High Throughput

● 橋詰保彦　● 高崎喜久夫　● 山崎　毅　● 山本昌司

あらまし

ネットワークの進化により，従来の常識やレベルをはるかに超える「膨大なデータの超高速処理」が求められるようになってきている。高速にデータ処理ができるようになったために，これまでにはなかった新しいサービス形態が生まれ，そのサービス形態が更に新しいICTの使い方を生み出すという連鎖が広がっている。

このような膨大なデータを高速に処理するためには，高速なレスポンスと高スループット化が重要であるが，信頼性との両立もシステムにとって必要不可欠である。富士通は，ミッションクリティカルシステムへの豊富な経験と先進のテクノロジを強みとしており，実際に東京証券取引所様arrowheadシステムの安定稼働を超高速データ管理ミドルウェア“Primesoft Server”で支えている。Primesoft Serverは，完全ディスクレスを考え方の原点とし，大規模データ管理，信頼性，接続性，拡張性を同時に実現した。

本稿では，高速レスポンス，高スループットの実現に向けたアプローチ，およびPrimesoft Serverの考え方と適用した新技術について紹介する。

Abstract

The evolution of the network has driven a demand for ultra-high-speed processing of huge amounts of data that far exceeds existing levels and what has commonly been considered possible. This data processing capability is giving birth to completely new modes of service, which in turn are giving birth to new ways of using information and communications technology (ICT). For a system that must process huge amounts of data rapidly, both high-speed response and high throughput are of course important, but high reliability must also be achieved at the same time. Leveraging its extensive experience in mission critical systems and its strength in advanced technologies, Fujitsu supports stable operation in the “arrowhead” trading system of the Tokyo Stock Exchange through its Primesoft Server ultrahigh-speed data management software. Based on the concept of diskless operation, Primesoft Server achieves large-scale data processing together with superb extensibility, flexibility, and reliability. This paper describes Fujitsu’s approach to achieving high-speed response and high throughput, explains the Primesoft Server concept, and introduces new technologies in this field.

ま え が き

金融取引，クレジット，流通，旅行，テレコムなど，様々な業界で，ビジネスモデルの革新により，システム利用人口，トランザクション処理数の飛躍的な増大が起きている。情報量増大の一途をたどるこれからの情報化社会では，ミッションクリティカルなシステムにおいても高い信頼性はもちろんのこと，ハイレベルな「高速レスポンスと高スループット」の実現が求められる。

このような状況から，富士通は高速レスポンス，高スループットの実現に加え，大規模なデータの管理を可能とし，高い信頼性，接続性，拡張性を兼ね備えた超高速データ管理ミドルウェア“Primesoft Server”を提供している。

本稿では，高速レスポンス，高スループットの実現に向けたアプローチ，Primesoft Serverでの実現に向けた考え方と適用した新技術について紹介する。

高速レスポンス，高スループットの実現アプローチ

本章では，ICTの進化とデータ処理モデル変遷の関係，および高速レスポンス，高スループット実現において着目したシステムアーキテクチャについて述べる。

● データ処理システムの抽象化モデル

富士通には，POS管理システムを原点とした「古典的なデータ処理システム」と呼ぶ抽象化モデルがある。このモデルでは，業務系のアプリケーションは，「エントリ処理→マスタ処理→通知処理→応答処理」という流れの形態をとっており，この流れの中に，データ処理に関する作業も組み込んでいる。エントリ処理の方法は，ネットワークの飛躍的な進歩に伴いリアルタイム化が進み，インターネットの活用が一般化している。

このように，ICTの素材やソフトウェアを含めた利用ツールは時代とともに飛躍的に進歩しているが，エントリ処理とマスタ処理のデータ連携，あるいは情報系とのデータ連携などの一連のデータ処理プロセスは，古典的なモデルが継承され続けている。そして，データ連携の多くが，ディスク利用を前提とした技術である。

● 高速レスポンス，高スループットを実現するシステムアーキテクチャ

従来のディスクを利用したデータ処理に着目し，これをすべてインメモリのデータ処理に置き換えることが高速レスポンス，高スループット実現のベースになると考えた。ポイントは，高速レスポンス，高スループットを重視するフロントエンドと，データ蓄積・活用を主とするバックヤードを整理・区分化することである。その上で，高速レスポンスを求めるフロントエンド向けのデータをインメモリに展開する。フロントエンドで処理されたインメモリデータの更新情報のうち保存・蓄積が必要なデータは，バックヤードに位置付けた各種マスタファイルに随時反映する。各種マスタファイルは，汎用DBMSを用いて，ディスク利用による永続性を確保する。

このように普遍的な抽象化モデルをベースにしつつ，マスタファイルのデータと同等の2次元表形式のインメモリテーブルと，トランザクションデータをプロセス間で受渡しできるインメモリの非同期メッセージキューを提供することが，高速レスポンス，高スループットの実現に有効との結論に至った。

データ処理システムの抽象化モデルに基づいた実現アプローチを**図-1**に示す。

Primesoft Serverの考え方

Primesoft Serverの開発では，高速レスポンス，高スループットの実現だけでなく，ミッションクリティカルシステムでの利用を配慮し，以下の五つの特長を同時に実現することに取組んだ。

(1) 考え方の原点：完全ディスクレス

高速レスポンス，高スループットを実現するため，完全ディスクレスのデータ処理を採用した。データをハードディスクドライブではなく，すべてメモリ上で処理することで，I/O処理を発生させることなく，処理速度を飛躍的に向上させる。

すべてのデータをメモリ上に置くことで格段の高速処理を実現するミドルウェア製品は，一般に「インメモリデータベース」と呼ばれており，Primesoft Serverもインメモリデータベースの一種と言うことができる。

一般のインメモリデータベースは，ログだけは

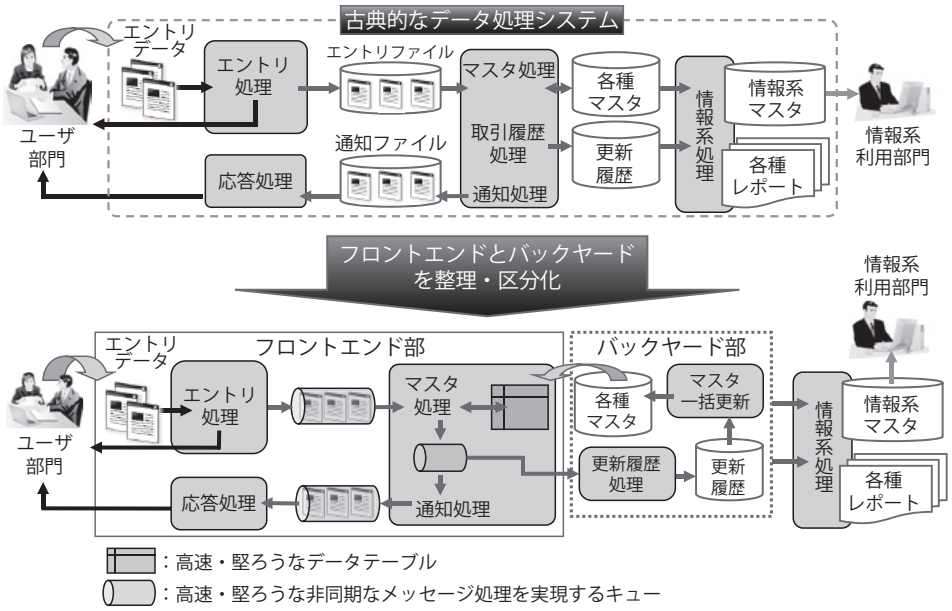


図-1　データ処理システムの抽象化モデルに基づいた実現アプローチ
Fig.1-Realization based on abstract model of data processing system.

メモリ外部のディスクへ一定周期で書き込んでいるが，Primesoft Serverは，ログまでもインメモリで高速処理している。どの側面から見てもディスクレスであり，インメモリであることによって，超高速処理を追求している。

(2) 大規模データ管理：数百GバイトからTバイトオーダの管理

すべてのデータをインメモリ処理するために，Primesoft Serverは，数百GバイトからTバイトオーダのデータをメモリ上に展開することが必要になる。このとき，1台の物理サーバでの搭載メモリ量の限界を超える場合には，複数のサーバのメモリ上に展開せざるを得なくなる。しかし，データ処理としては，あたかも一つのサーバ上のデータのように扱うことが可能な分散処理が必要となる。

(3) 信頼性：汎用的なハード素材を冗長化して信頼性を確保

Primesoft Serverは，高可用性，堅ろう性，高信頼性，一貫した運用性，能力拡張性などの要件を高度に追求しつつ，半導体メモリなどには特殊な製品を使わず，あくまでも汎用的な素材を利用し，素材が持つ機能の限界まで近づくことを目指した。汎用的な素材を利用することで，経済効率を求めるとともに，今後の可能性を柔軟に広げていくことができるからである。

また，インメモリで超高速処理を実現するデータ処理ミドルウェアは，性能要求が非常に高度な金融や産業の最先端システムで使われることになる。そのシステムが万一故障した場合には，影響は企業内にとどまることなく，社会・経済・産業界まで大きく広がってしまう。システム停止が起こった場合の現実世界での影響範囲の予測が難しく，信頼性に社会的責任まで伴うのが，インメモリのデータ管理ミドルウェアならではの特質である。

このため，障害発生時に速やかに業務が再開できるようにする技術も，さらに高度化する必要がある。

(4) 接続性：大量の同時接続時にも高速アクセスを実現

高速レスポンス，高スループットが求められるアプリケーションとは，極めて短い時間に多数のクライアントからのアクセスが集中することがあるアプリケーションである。例えば，一般消費者が利用するWebアプリケーションであれば，世界中からのアクセスが短時間に集中する場合がある。あるいは，クライアントを操作するのは人間ではなく，多数のコンピュータがミリ秒間隔で自動的に応答する場合もある。

したがって，膨大な数のアクセスを，機会均等，かつ並列に無駄なく動作させて，レスポンスを平

準化する技術が不可欠である。

（5）拡張性：トランザクション増加へ迅速に対応

インメモリのデータ処理を必要とする金融・産業界の最先端のシステムでは，予測できない急激なトランザクションの増大が発生することが多い。したがって，「30分以内の負荷分散・能力拡張」のように，これまでは考えられなかったほど，短時間で確実な能力拡張を行うことが必須である。

高速レスポンス，高スループットを実現した新技術

完全ディスクレスを原点として，大規模データ管理，信頼性，接続性，拡張性を同時に実現するために，Primesoft Serverでは，以下に述べる五つのテーマで新規技術開発を行った。

● 物理メモリ量の限界を超越

物理メモリ量の限界を超越するための技術を**図-2**に示す。メモリにデータを展開するということは，データ量が一つの物理サーバの搭載メモリ量より多いと，システム構築ができないということを意味する。例えば最大256 Gバイトのメインメモリを用いた場合でも，信頼性を高めるために二重化すると，半分の128 Gバイトであり，プログラムなども格納するため，128 Gバイトすべてをデータ格納に使うわけにはいかない。メインメモリの物理量には限界がある。したがって，Tバイトオーダのデータ展開を可能にするには，分散配置が不可欠である。

Primesoft Serverでは，データの分散配置を行うために，主に三つの技術を用いた。

（1）テーブルパーティショニング

一つのテーブルのデータを細分化して複数のサーバに配置する技術である。本技術により，一つのテーブルを1台のサーバの物理的なメモリ量限界を超えて構成可能とした。

（2）仮想化

アプリケーションに対して物理サーバとデータ所在の関係を意識させない技術である。物理的な分割を意識させないことで，アプリケーション側では特別なデータの扱い方を考える必要がなく，アプリケーション開発の負荷を軽減する。

（3）ローカライズ

細分化した単位に独立して実行できる運用技術である。マスタの一部の修正など，日々の保守・運用管理作業を，システム全体を止めることなく実行できる。

● 障害に対する速やかな業務再開

障害の原因は，ハードウェア障害，ネットワーク障害，OS障害，ミドルウェア障害，業務アプリケーション障害など様々であり，これをゼロにすることはできない。そこで必要になるのは，確実な異常検知と，迅速な切替えである。

Primesoft Serverでは，豊富な適用実績を持つ富

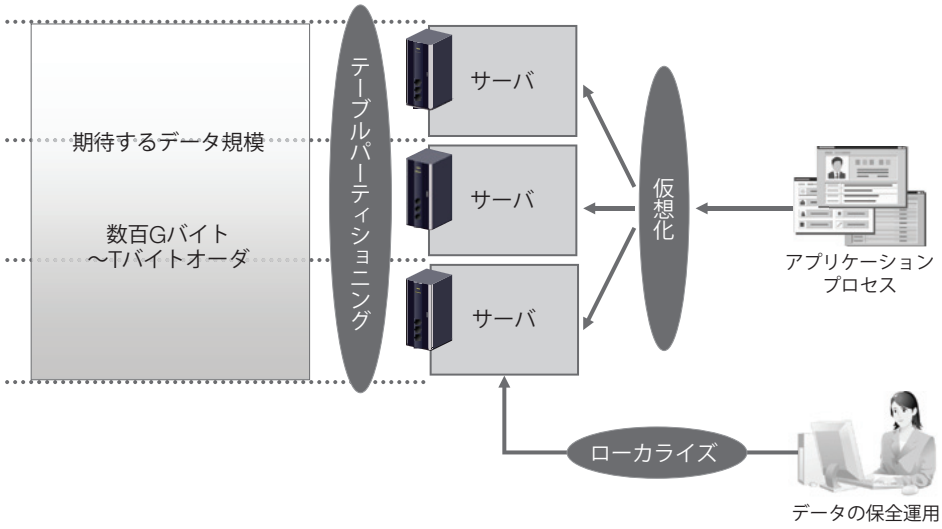


図-2　物理メモリ量の限界を超越するための技術
Fig.2-Technology to exceed limit of physical memory.

士通のクラスタ製品であるPRIMECLUSTERをベースに，新技術を導入して「秒レベル」の検知と切替えの要件に対応した。まず，秒レベルで異常を100％検知するために，総合的ハートビート(注1)診断技術を新規開発した。これは，従来の管理LAN(注2)を通したハートビート診断に加えて，業務LAN,(注3) 同期LAN(注4)の通信状況を加えた総合診断を行う技術である。

さらに，万が一の障害発生時も秒オーダでサービスを再開するために，データミラーリング技術を駆使した。これは，発生したデータをすべてミラーリングして，データの最新性を待機系で常に保証する技術である。また，一つの運用系に，複数個の待機系を用意して，可用性をより高めている。

総合的ハートビート診断技術およびデータミラーリング技術を支えているのは，主に次のような技術である。

(1) クラスタシステム

複数のコンピュータを使って高い可用性を実現する技術。

(2) フェイルオーバ

メモリテーブル操作の秒オーダでの高速再開技術。

(3) インメモリデータのミラーリング

運用系と同じ状態のメモリテーブルを待機系で維持する技術。

速やかな業務再開を支える技術を**図-3**に示す。

● 障害箇所に応じて影響波及範囲を局所化

信頼性を高めるためには，影響波及範囲を局所化することも重要である。システム全体ではなく，障害箇所に応じた業務継続措置を行って，過剰な設備投資を防ぐ意味もある。影響範囲を局所化する技術を**図-4**に示す。

Primesoft Serverでは，故障・障害の部位に対応したフェイルオーバを行うために，主に以下の二つの技術を駆使した。

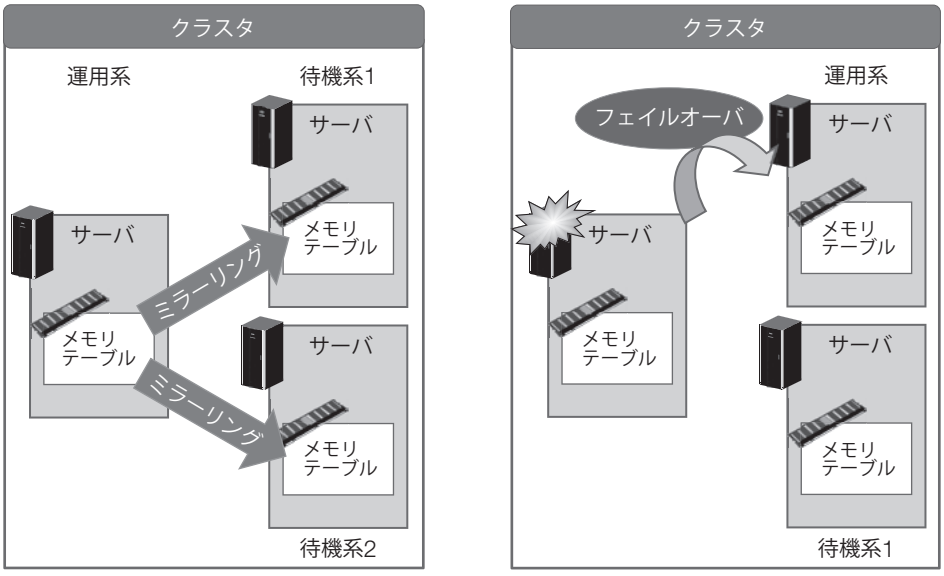
(1) 3層クラスタ

アプリケーション層，ミドルウェア層，物理サーバ層という階層ごとに，故障部位に適したフェイルオーバを行う技術である。

(2) HA-AP（High Availability - Application）環境

アプリケーションプロセスの冗長化による高可用性技術である。業務ロジックを実行するアプリケーションプロセスが異常終了すると，一般には，アプリケーションプロセスの再起動による処理能

(注1) コンピュータやネットワーク機器が正常に稼働していることを外部に知らせるために，定期的に発信しているネットワーク信号。心臓の鼓動の意味。
(注2) クラスタ構成内のサーバ間での制御用LAN。
(注3) 業務データの送受信を行うために利用するLAN。
(注4) データのミラーリングに使用するLAN。



(a) インメモリデータのミラーリング　(b) フェイルオーバ

図-3　速やかな業務再開を支える技術
Fig.3-Technology to support reopening of rapid business.

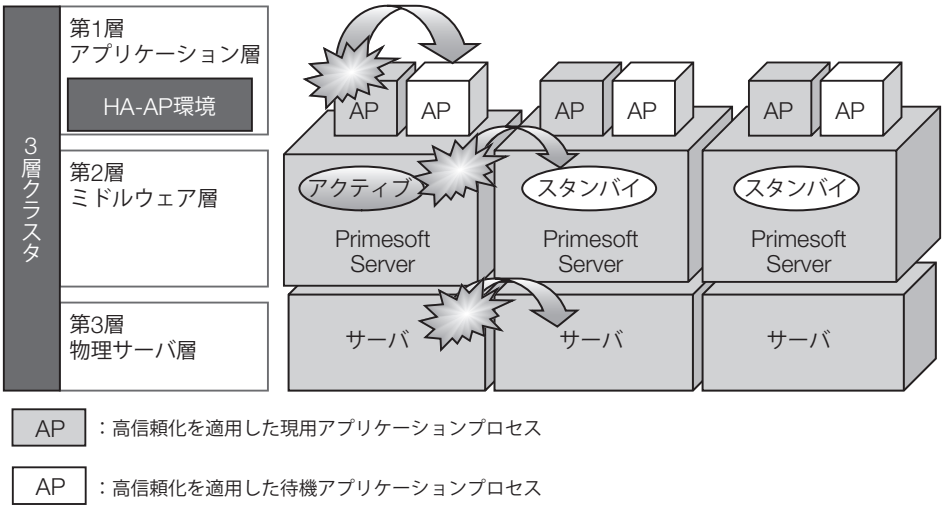


図-4　影響範囲を局所化する技術
Fig.4-Technology to localize influence range.

力の確保措置がとられる。しかし，マイクロ秒オーダの高速レスポンスや高スループットが求められる状況での業務アプリケーションプロセスの再起動は，各種のシステムリソースの負荷を高め，結果として全体の処理能力を低下させる。

そこで，待機アプリケーションへ予備プロセスを事前に起動してプール化しておき，現用アプリケーションで予期しないアプリケーションプロセスの異常終了が発生したときには，待機アプリケーションからプロセスを補充して，瞬時に業務サービスを継続するアプリケーション高信頼化技術を開発した。

● 大量の同時接続時にも均等・高速なアクセスを実現

高度なアクセシビリティを実現するWebアプリケーションでは，膨大な数のクライアントに対して，マイクロ秒レベルのレスポンスを実現しなければならない。大量の同時接続時における高速なアクセス性能の実現は，半世紀に及ぶ富士通のDBMS開発の歴史の中で培った高度な技術を，インメモリデータ管理向けに進化・チューニングして実現した。

しかし，Primesoft Serverは，多数のサーバをネットワーク化した分散サーバ構成をとるため，インメモリでのアクセス性能をいかに強化しても，最終的には，各サーバ間の制御通信性能やデータ連携性能がレスポンスやスループットを大きく左右することになる。そこで，新たに高信頼なUDP（User Datagram Protocol）通信技術を開発した。

UDPとは，TCP/IPにおけるトランスポート層プロトコルの一つであり，一般にUDP通信は，TCP通信よりも高速だが信頼性が低い。そこで，複数の受信先サーバ（運用系，待機系など）に対する多彩な送達確認パターンを実装し，局面に応じて使い分けることにより，送達確認のオーバヘッドを必要最小限に抑制しながらUDP通信でも高信頼な通信を実現する方法をとった。

また，Primesoft Serverでは，UDP通信で発生するパケットロストを検知して再送するのではなく，同一パケットを冗長化して複数の経路に一度に転送するという独自の送達確認技術により，経路異常が発生した場合にも，性能劣化のない通信を実現し，高速通信と高信頼性を両立させることに成功した（**図-5**）。

なお，高信頼UDP通信の技術は，インメモリデータのミラーリングにも適している。

● 処理量の増加に対するスケーラビリティ

突発的なトランザクションの増大に対して，極めて短い許容時間以内でシステムキャパシティ強化を実現するために用いた主な技術は，下記の二つである。

(1) 動的スケールアウト

オンライン条件下で，業務サービスの実行環境をほかのクラスタへ移動する技術である。例えば，

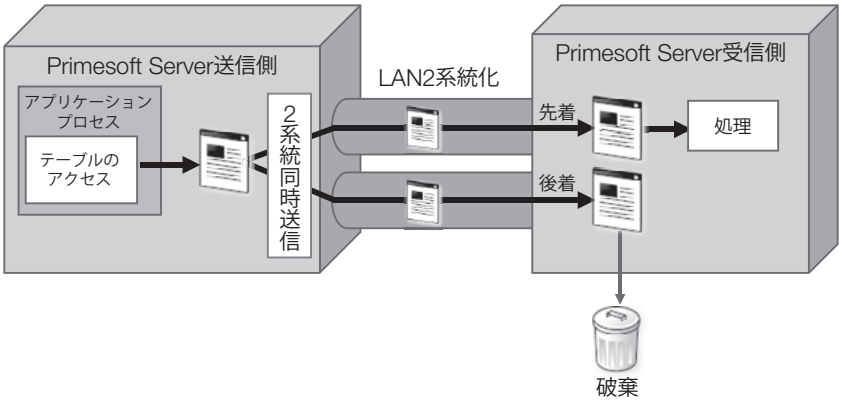


図-5　高信頼UDP通信
Fig.5-High reliable UDP communication.

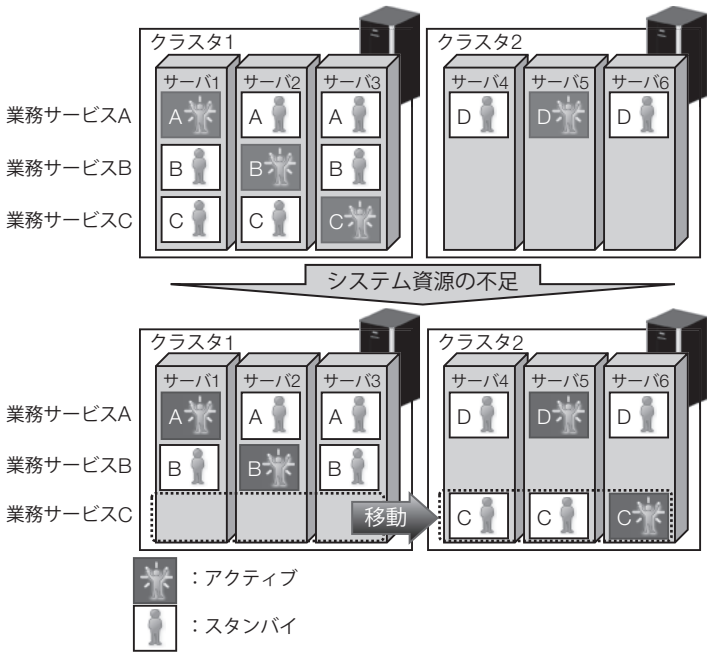


図-6　動的スケールアウト
Fig.6-Dynamic scale out.

取引量増大によって，メモリ不足の発生が予想されるとき，クラスタ1から能力に余裕のあるクラスタ2へ業務サービスの一部を移動することで，クラスタ1に余裕を持たせることができる。

動的スケールアウトの例を**図-6**に示す。

(2) テーブル配置の仮想化

メモリテーブルの分割単位（パーティション）は，負荷分散・危険分散の見地でシステムごとに効果的なデータ機軸を考えて，利用者が設計可能な構造としており，かつ，アプリケーションからは配置先に関する意識を不要としている。これにより，限界のない性能拡張と，容易なスケールアウトを実現し，業務サービスの全体停止リスクを大幅に減少させている。

む　す　び

Primesoft Serverは，高速性と信頼性を徹底的に追求したインメモリデータ管理ミドルウェアである。そのレスポンスは，汎用DBMSの10倍から100倍の高速性を実現している。

これを実現するために，Primesoft Serverには，これらのデータ管理テクノロジ，ネットワークテクノロジ，クラスタテクノロジが縦横無尽に駆使されている。一方では，CPUの平均命令実行時間を左右するキャッシュのミスヒット率の減少に向けたプログラミング技術，あるいは制御テーブルのサイズや配置の最適化など，地道な取組みも基盤となっている。

「従来の常識を超える膨大なデータの超高速処理」を必要とする市場に対するソリューションの一つとしてPrimesoft Serverを提供した富士通は，今後も，Primesoft Serverで培ったテクノロジを富士通のほかのミドルウェアへもフィードバックしていき，お客様のビジネスをしっかりと支えていく方針である。

著者紹介

橋詰保彦（はしづめ　やすひこ）
プラットフォームソフトウェア事業本部第一プラットフォームソフトウェア事業部 所属
現 在，Primesoft Server，ETERNUSソフト開発に従事。

山崎　毅（やまざき　たけし）
プラットフォームソフトウェア事業本部第一プラットフォームソフトウェア事業部 所属
現在，Primesoft Server開発に従事。

高崎喜久夫（たかさき　きくお）
プラットフォームソフトウェア事業本部第一プラットフォームソフトウェア事業部 所属
現在，Primesoft Server開発に従事。

山本昌司（やまもと　しょうじ）
プラットフォームソフトウェア事業本部第一プラットフォームソフトウェア事業部 所属
現在，Primesoft Server開発に従事。