スケーラブルなパラレルシステムを実現する高速結合機構

High-speed Connection Control Feature for Scarable Systems

渡部眞也*　*Masaya Watanabe*
香田克也**　*Katsuya Kôda*
吉岡正壱郎***　*Masaichirô Yoshioka*

H-6710高速結合装置
高速結合機構を搭載するH-6710高速結合装置(後方)と2台のMP5600(前方)が接続されてMパラレルシステムを構築している。

HCCF(High-speed Connection Control Feature：高速結合機構)は,「Mパラレルコンセプト」に基づいて開発された，CPN(Central Processing Node)間の通信や負荷バランス，共用データの排他制御などを高速に行う処理機構である。HCCFは専用の処理装置(H-6710高速結合装置)に搭載され，ISCH(Inter-System Connection Channel：システム結合チャネル)によってMパラレルシリーズの各プロセッサと結合される。また，各CPN上で稼動するオペレーティングシステム“VOS3/FS”(Virtual Storage Operating System 3/FOREFRONT System Product)にも，スケーラブル(段階的)なパラレル処理を実現するための新たな機能をサポートした。

HCCFを利用したMパラレルシステムは，磁気ディスク装置などを介して行っていた従来のパラレル制御とは異なり，共有データの排他制御などのパラレル処理をHCCFにオフロードするため，CPN側の負荷を小さくすることができる。さらに，CPN上のOS(Operating System)は，常にHCCFとの1対1の処理だけを意識すればよいため，パラレル多重度が増してもCPN側の負荷はほとんど増加しない。これによってMパラレルシステムでは，スケーラブルなシステムを容易に構築することができる。

* 日立製作所 汎用コンピュータ事業部　** 日立製作所 ソフトウェア開発本部　*** 日立製作所 システム開発研究所

1 はじめに

近年のコンピュータシステムには，業務量の増加に伴う高速大容量処理，24時間365日連続運転のための高信頼性などが要求されており，さらにコストパフォーマンスの向上が求められ，また，パラレルシステムへのニーズが高まっている。

「Mパラレルシステム」は，Mシリーズの豊富なソフトウェア資産の継承と，性能・信頼性に優れた大型汎用コンピュータ製品の特徴を生かした高性能・高信頼なパラレルシステムである。これにより，幅広い分野でのニーズにこたえてスケーラブルな拡張が可能なシステムを提供することができる。

HCCF(高速結合機構)は，高性能・高信頼なパラレルシステムを実現するためのパラレル処理機構であり，専用の処理装置であるH-6710高速結合装置に搭載される。

ここでは，Mパラレルシステムと，HCCFを搭載するH-6710高速結合装置の特徴について述べる。

2 Mパラレルシステムの特徴

Mパラレルシステムの構成を**図1**に示す。

Mパラレルシステムでは，パラレル制御を行う処理装置としてH-6710高速結合装置を導入している。H-6710高速結合装置は各CPNと結合され，ACONARCチャネルなどで接続された共有データの排他制御や負荷バランス制御などを行う。また，それぞれのCPN(Central Processing Node)にはVOS3/FSが搭載され，HCCFが持つ機能を利用してその連携制御の下でパラレル処理を行う。

Mパラレルシステムは以下に述べる特徴を持つ。

(1) パラレル処理によるスケーラビリティの向上

パラレル制御をH-6710高速結合装置へオフロードし，CPN上のOS負荷を抑えることによってスケーラビリティを向上した。

(2) システム可用性の向上と増設の容易化

各CPNはすべて稼動中増設・切り離しが可能な設計とし，システムの可用性を向上させた。

(3) オペレーションの一元化

統合コンソールから各CPNと連携のとれた，最適なオペレーション環境が提供できる。

3 パラレル処理専用のH-6710高速結合装置

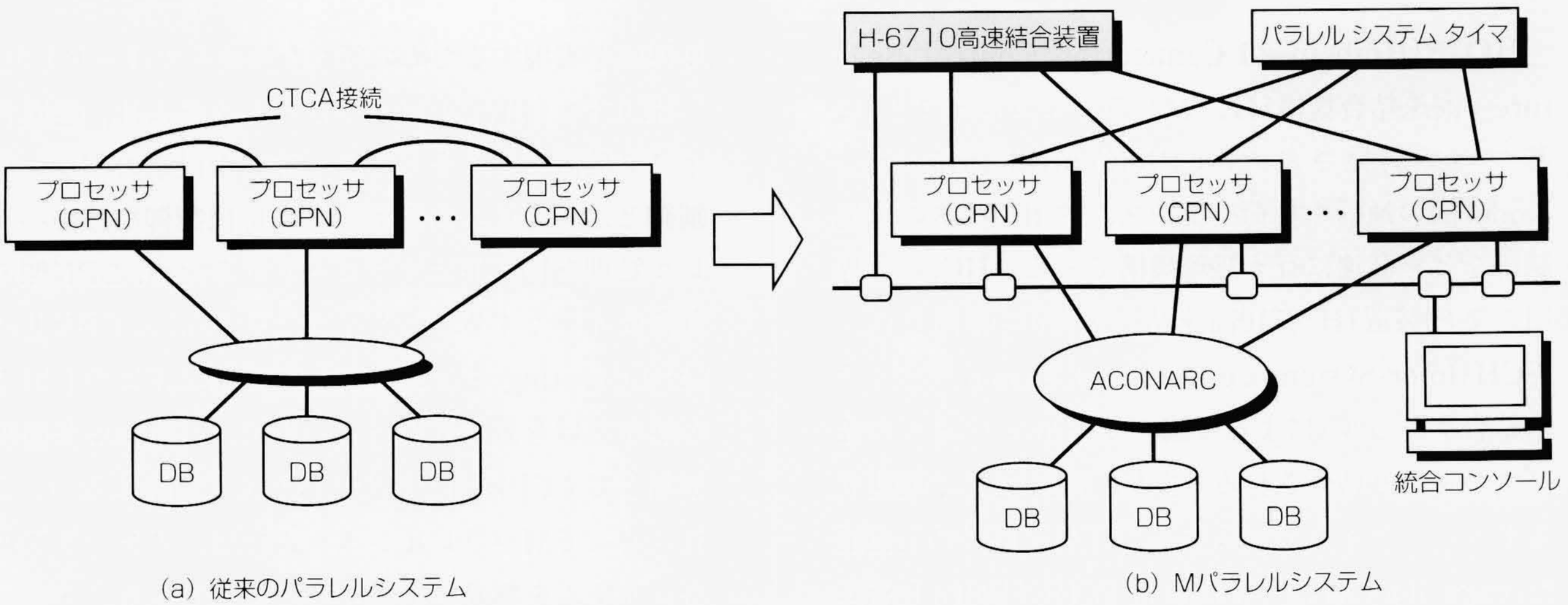
HCCFを搭載し，パラレル処理を専門に行う処理装置としてH-6710高速結合装置を開発した。

3.1 H-6710高速結合装置の特徴

H-6710高速結合装置は，HCCFおよびISCH(システム結合チャネル)を搭載するパラレル処理専用の処理装置として以下に述べる特徴を持つ。

(1) パラレル制御専用処理装置

パラレル制御専用の処理装置として，自動立ち上げやシステム構成の自動認識機能を持ち，簡易なシステム運用を実現した。



(a) 従来のパラレルシステム　　(b) Mパラレルシステム

注：略語説明　CTCA(Channel to Channel Adapter)，ACONARC(Advanced Connection Architecture)

図1　Mパラレルシステムの構成

Mパラレルシステムは，複数のプロセッサ，プロセッサを接続するH-6710高速結合装置，ファイル共有のためのACONARCチャネルなどで構成される。

(2)　システム結合チャネルによる高速データ転送

チャネルにはISCHだけを搭載し，チャネルサブシステムをISCH専用に特化させて，CPNとの高速な通信機能を実現した。

(3)　さまざまなパラレル環境に適合できる性能レンジ

マルチプロセッサ対応によってユニプロセッサから8プロセッサまでをサポートした。また，パラレル処理専用にメモリの最適化などを行って高速制御を実現した。

3.2　実現方式

H-6710高速結合装置は，高速CMOSプロセッサ，HCCF，および高速なデータ転送を行うシステム結合チャネルで構成する。

3.3　ISCH(システム結合チャネル)

ISCHはH-6710高速結合装置とこれに接続されるCPNのそれぞれに搭載され，両者を光ファイバケーブルで接続する。ISCHは，Mパラレルシステム内の複数のCPNどうしがH-6710高速結合装置を介してメッセージ交換するためのメッセージ通信専用チャネルである。

Mパラレルシステムでは，システムの中心に位置するH-6710高速結合装置に多くのCPNを接続し，排他制御などのために大量のメッセージを頻繁に通信するので，ISCHには遠距離接続と高速転送が要求される。このため，ISCHファイバ チャネル スタンダードに準拠した光送受信モジュールを適用し，最大3 kmの接続距離と最大100 Mバイト/sの転送速度を実現している。さらに，ISCHは従来の高速光チャネルやACONARCと異なり，メッセージの多重かつ双方向の通信を実現して，性能と使用率を向上させた。

また，稼動中の保守や増設・撤去を実現し，Mパラレルシステムのスケーラビリティと可用性を向上させた。

H-6710高速結合装置は，最大32チャネルのISCHが搭載できる。また，チャネル構成定義にはISCHだけの最大構成が出荷時から組み込まれており，顧客サイトでの増設・撤去の際にはチャネル構成定義の変更作業が不要である。

4　HCCF(高速結合機構)の構成と機能

4.1　パラレル処理の動作原理

HCCFでは，従来のチャネルとは異なるISCHを介して各CPNと結合され，**図2**に示す動作原理によってパラレル処理を行う。

各CPN上のOS(VOS3/FS)はISCHを介し，メッセージ機構と呼ばれるインタフェースを用いてHCCFの機能を利用する。この際，ISCHは従来のチャネルと同様にデータ転送などを行うが，I/O(Input/Output)割込みを発生させることがないため，ソフトウェアに対して余分な負荷をかけることがない。

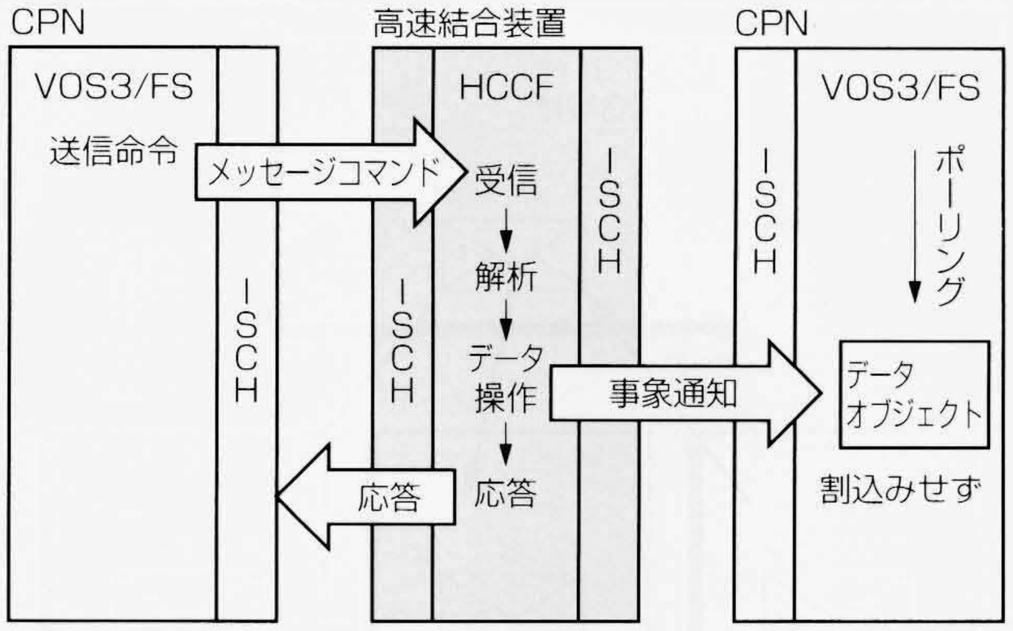


図2　パラレル処理の動作原理
HCCFを利用したパラレル処理はすべてCPN上のOSとは非同期に行われる。

また，HCCFはISCHを介して各CPN上のデータオブジェクトを直接操作することができる。この機能を用いて，HCCFはCPN上のOSに処理の結果や事象の通知を行う。CPN上のOSでは，このデータオブジェクトを定期的に参照し(ポーリング)，HCCFからの通知を受ける。

以上のように，HCCFを利用したパラレル処理はすべてがHCCF上で処理され，CPN上のOSの動きとは非同期に行われるため，スケーラブルなパラレル設計が容易となる。

4.2　HCCFの機能

HCCFは**表1**に示す機能を用いてパラレル制御を実現している。

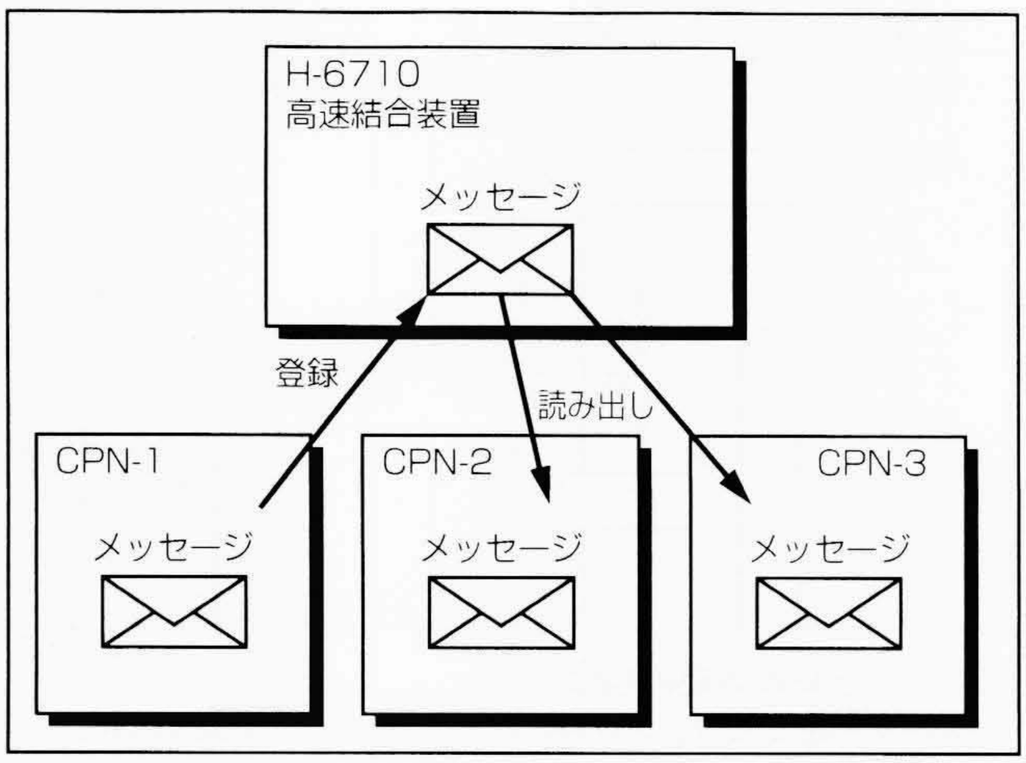
各機能の動作を**図3**に示す。各機能について以下に述べる。

4.2.1　CPN間メッセージ通知

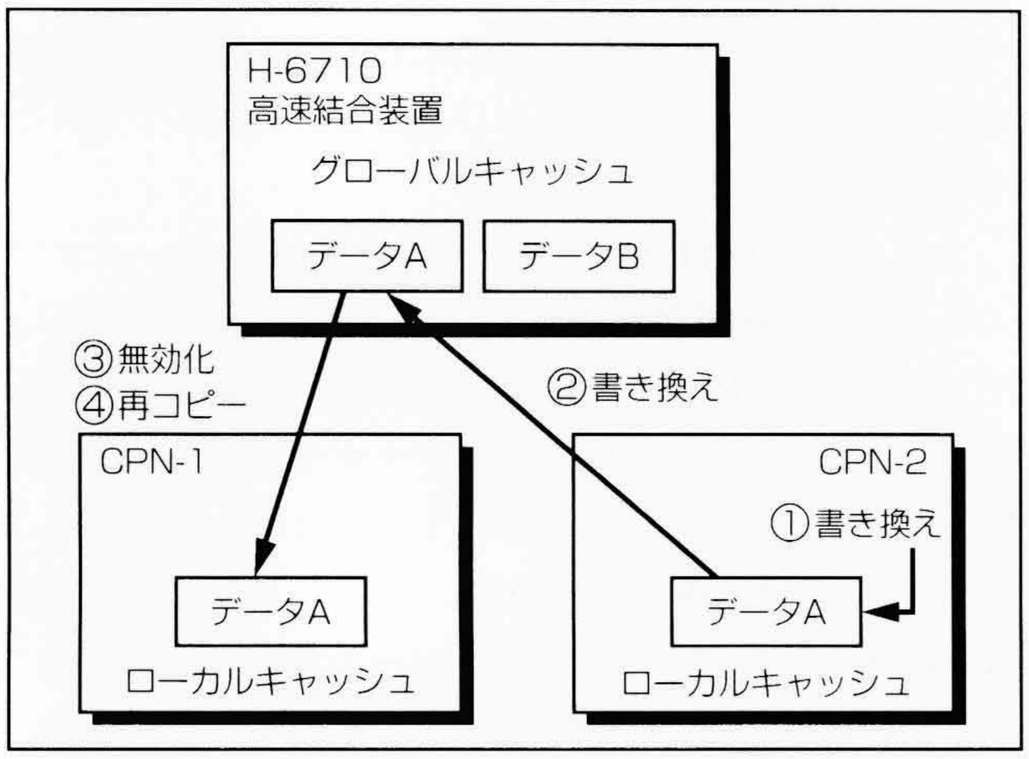
パラレルシステムでは，CPN間の連絡やデータのやり

表1　HCCFの機能
HCCFは表に示す三つの機能を用いて高速なパラレル制御を実行する。

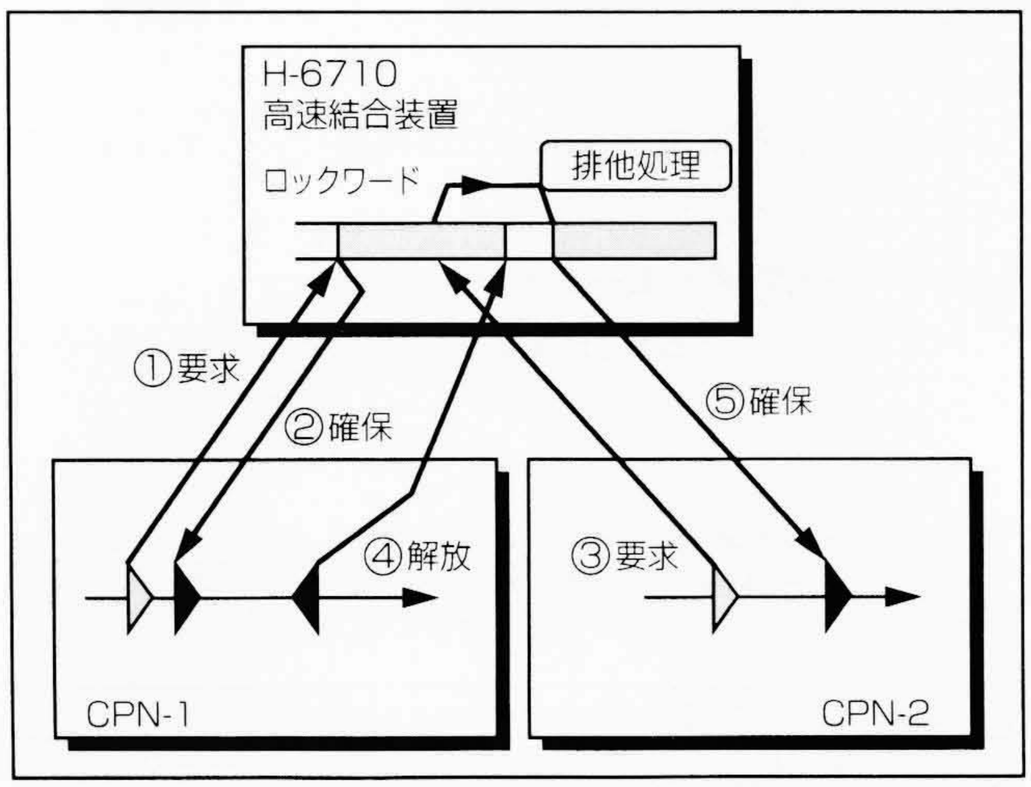
機能名	内　容	特　　徴
CPN間 メッセージ通知	CPN間のメッセージ送受信	・非同期にメッセージ送信が可能 ・複数CPNへの同時送信が可能
CPN間 データ共用	CPN間の共用データアクセス	・キャッシュ機能によって高速アクセスが可能 ・一致制御をHCCFが行う。
CPN間 排他制御	CPN間共用資源の排他制御	・排他制御をHCCFが行う。 ・排他制御単位が任意に決められる。



(a) CPN間メッセージ通知



(b) CPN間データ共用



(c) CPN間排他制御

図3 HCCFの機能

HCCFではCPN間のメッセージ通知，データ共用，排他制御を用いてパラレル制御を行う。メッセージ通知機能では1対多の非同期な通知が可能であり，キャッシュの一致制御と排他制御ではHCCF上で処理が行われる。

取りが重要となる。

CPN間メッセージ通知では，CPN間の連絡やデータのやり取りを並列度を落とさないで高速に実行する。従来のCTCAを用いたメッセージ通信では，送信側と受信側が同期して送受信処理を行う必要があり，しかも1対1の通信しかできなかったため並列度が上がらなかった。

HCCFを用いたCPN間メッセージ通知では，送信側CPNはメッセージをHCCFに登録するだけで，メッセージ送信処理が完了する。HCCFは受信側CPNにメッセージが登録されたことを通知する。このとき，一つのメッセージを一度に複数のCPNに通知できるため，並列度を落とさないでメッセージ通知が実現できる。受信側CPNは都合が良いときに，登録されているメッセージを受信すればよい。

また，メッセージのデータ通信には，高速なISCH使用によって高速なメッセージ通信が可能である。

4.2.2 CPN間データ共用

パラレルシステムでは，共用データに高速にアクセスできることが，並列度を上げてトータルスループットを向上させる重要な要因である。

CPN間データ共用では，共用データに高速にアクセスできる。従来の半導体記憶装置や磁気ディスク装置に共用データを置く方法では，共用データにアクセスするたびに入出力処理が発生し，スループットを向上させることができなかった。

HCCFを用いたCPN間データ共用では，HCCF内に共用データ(グローバルキャッシュ)を置く。各CPNはグローバルキャッシュの一部のコピー(ローカルキャッシュ)を持つ。各CPNでは，ローカルキャッシュに対応するグローバルキャッシュが更新されないかぎりローカルキャッシュ参照で共用データが参照可能なので，高速な共用データの参照が実現できる。

HCCFではグローバルキャッシュが更新されると，更新されたグローバルキャッシュに対応するローカルキャッシュを持つCPNに，ローカルキャッシュの無効を通知する。各CPNではローカルキャッシュを参照する際に，ローカルキャッシュの有効性をチェックし，無効であればグローバルキャッシュをローカルキャッシュに再コピーしてから共用データを参照する。

4.2.3 CPN間排他制御

パラレルシステムでは，CPN間で共用する資源を高速に排他制御できることが，並列度を上げてトータルスル

ープットを向上させる重要な要因である。

CPN間排他制御では，CPN間で共用する資源を高速に排他制御する。従来は，半導体記憶装置や磁気ディスク装置上にロックワードを置き，各OSがこれを「取り合い」することで排他制御を行っていた。この方法では，頻繁に入出力処理が発生して排他処理に長時間を要していた。また，排他する単位が外部媒体のアクセス単位と大きく異なるため，排他期間がシステム間で競合するケースが多く，並列度が上がらなかった。

HCCFを用いたCPN間排他機能では，共用資源の排他を管理するロックワードをHCCF内に置き，この「取り合い」制御をHCCFが行うことによって高速に排他制御が実現できる。また，排他する共用資源の単位はCPN側のOSが任意に設定できるので，排他単位を最適に設定することができる。そのため，排他期間がCPN間で競合するケースを少なくすることが可能なので，並列度を上げることができる。

5　利用形態

VOS3/FSを搭載したMパラレルシステムでは，HCCFを以下の分野に適用する。

5.1　スプール共用

DASD(Direct Access Storage Device)に格納していた共用ジョブキュー情報をHCCF上に配置し，高速に排他制御データ転送することによってスケーラブルな処理能力を提供する。

また，共用されるスプールの排他制御にHCCFを利用することにより，排他オーバヘッドの削減と特定システムの沈み込みを防止する。

5.2　負荷分散ジョブスケジューリング

スプール共用しているシステム間では，HCCFを利用したシステム間負荷分散機能により，各システム間のCPU(Central Processing Unit)使用率を平準化してスループット向上を図る機能を提供する。

5.3　システム間データ転送

HCCFを利用したシステム間のデータ転送機能により，システム間の異なる業務の間でデータの転送を行って業務全体を並行に実行する機能を提供する。

5.4　XDM機能分散パラレル

複数システムを結合したパラレルシステムでは，DB処理をDBを割り当てている別のシステムで実行して負荷を分散させることにより，システム全体のスループットを向上させる。

5.5　XDM SQLパラレル

SQL(Structured Query Language)を複数のサブSQLに分割し，複数のDBノードで並列検索を実行することによってSQLの実行時間を大幅に短縮する。また，ノード間のCPU使用率を平準化してスケーラブルな処理能力を実現する。

5.6　XDMセッションパラレル

XNF〔Extended HNA(Hitachi Network Architecture)Based Communication Networking Facility〕と連携して，オンライン端末のセッション接続時に，HCCFを利用したAP(Application Program)ノードの付加情報に基づいて，接続するAPノードを選択する。これによってAPノード間のCPU使用率を平準化し，スケーラブルな処理能力を実現する。

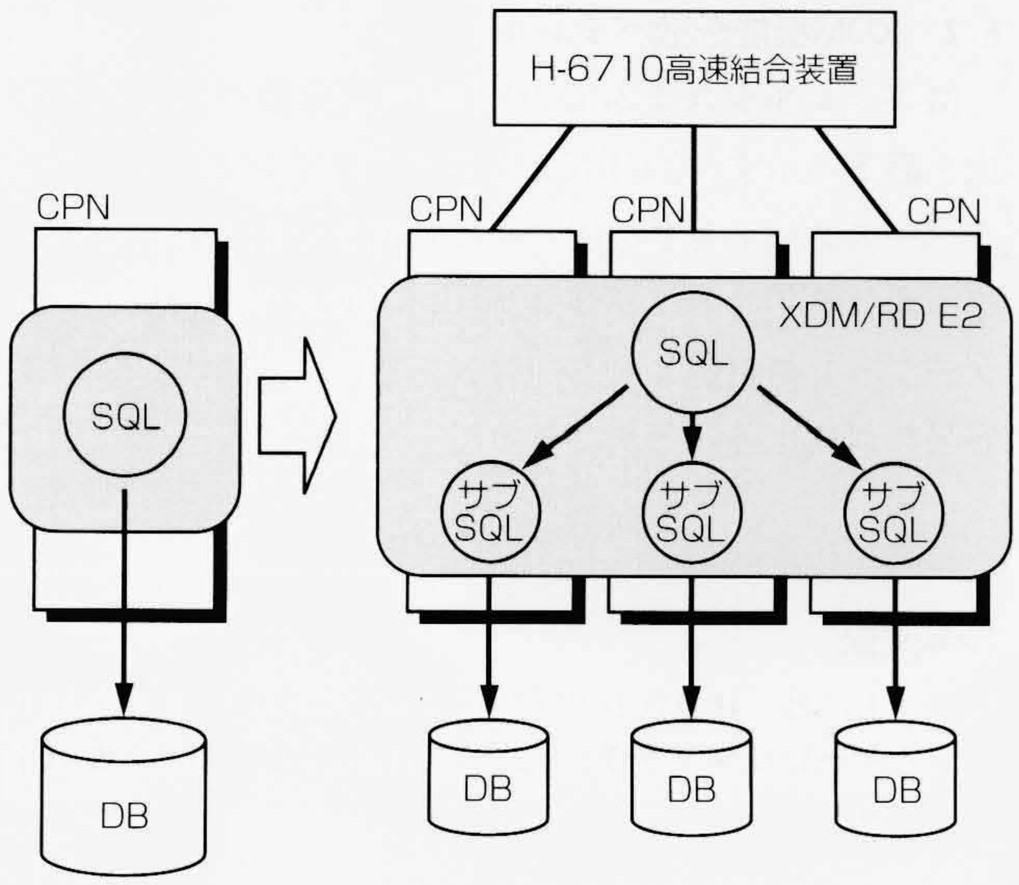
5.7　XNFセッションパラレル

TSS(Time Sharing System)オンライン端末のセッション接続時に，HCCFを利用したシステム間の負荷情報に基いて，接続するシステムを選択する。これによってシステム間のCPU使用率を平準化し，スケーラブルな処理能力を提供する。

6　効　果

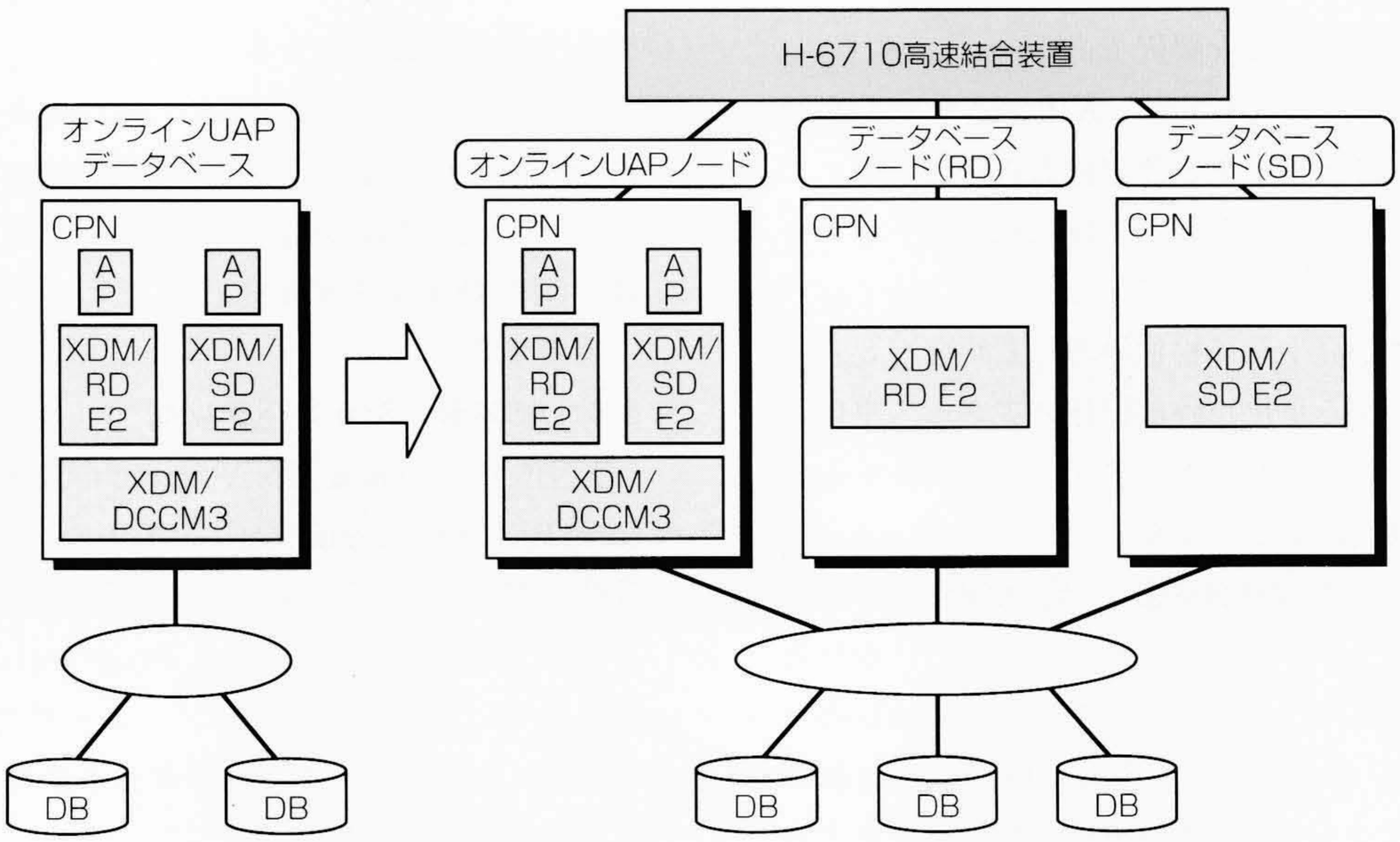
6.1　SQLパラレル

Mパラレルシステムでの XDM SQLパラレルの構成を**図4**に示す。従来，1台のマルチプロセッサ上で行って



注：略語説明
XDM/RD E2(Extensible Data Manager/Relational Database Extended Version 2)

図4　XDM SQLパラレルの構成
各CPN上に分割したサブSQLを配置する。データベースもおのおのに分割して並列実行する。



注：略語説明
UAP(User Application Program), XDM/SD E2(XDM/Structured Database E2)
XDM/DCCM3(XDM/Data Communication and Control Manager 3)

図5 XDM機能分散パラレル構成
各CPN上に機能ごとに分散させる。オンライントランザクションの増大や新規アプリケーションの追加には，CPNを増設することによって容易に対応できる。

いた処理と表データを，おのおののCPN上の複数のサブSQLに分割して実行することによって処理時間を短縮する。

例えば，約30万件の表を三つに分割し，三つのサブSQLで2万件を検索する処理では，処理時間が約$\frac{1}{2.7}$に短縮できる。

6.2 XDM機能分散パラレル

Mパラレルシステムでの XDM機能分散パラレルの構成を**図5**に示す。各CPNごとにオンラインノード，データベースノードなど機能ごとに処理を分散する。この結果，オンライントランザクションの増大や，新規アプリケーションの追加に対し，CPNを増設することによって容易に対応することができる。

7 おわりに

ここでは，HCCFを搭載し，パラレル処理を専門に行うH-6710高速結合装置の概要について述べた。

優れた最新の大型汎用機間の連携処理を実現し，VOS3/FS，XDMなどのパラレル対応により，既存のソフトウェアの継承の下で高信頼でスケーラブルなパラレルシステムの構築を可能とした。また，統合コンソールやパラレル システム タイマの開発により，システムの管理運用性を大幅に向上させた。

今後もオープンシステムとの連携をいっそう強化し，社会ニーズにこたえることのできる充実したシステムを開発していく考えである。

参考文献

1) 尾山，外：新世代システムによるデータセンタを実現するパラレルソフトウェア，日立評論，**77**，5，375～380(平7-5)