ループを用いた大規模分散処理システム†

鶴保征城††　木ノ内康夫†††　星子隆幸††††
仲谷元††††　宮川順治††††

高い信頼性と拡張性を満足するシステムの構成方法として，分散処理方式が注目されている．本論文で報告するDIPS複合システムは，ハードウェアから応用ソフトウェアに至るまで一貫した構想のもとに，大規模分散処理を指向し，汎用大型機を用いて構成したシステムである．90年代にわたって有用な高い拡張性と信頼性，徹底した自動化の実現をめざしている．本論文では，以下の方式を中心に大規模分散処理システム構成上の問題とその対処法，導入実績を踏まえた評価を報告する．①高位レイヤまでの通信制御機能の分散など，高度のプロセッサ分散方式，②高性能光ループによる，拡張性が高く高信頼なプロセッサ間通信方式，③システム制御用プロセッサを用いた，大規模分散処理システムの自動運転方式，④プロセッサ間の連携により，端末に対してホストプロセッサ障害を隠ぺいするなど，分散処理の特徴を生かした高信頼化方式，⑤オンラインサービス中のシステム増設方式．

1. はじめに

情報処理システムは利用分野の拡大とともに社会生活上必須の基盤となり，信頼性が以前にも増して重要になってきている．また経済環境の変化に即応した多様なサービスの展開が求められており，システム処理能力の拡大やサービス，機能の柔軟な拡張が強く要求されている．センタシステムにおいて，これら高い信頼性と拡張性を同時に満足する有力な手段として，分散処理構成が採用されてきている．

この方向は大きく二つに分類される．一つはミニコンレベルの小型機を多数台接続し，長時間の連続運転や容易な拡張を特徴とするものである[1)~4)]．ここでは分散処理という観点から，ハードウェアとOSを統一した考えのもとに整備してきているが，これまでのところ，主な適用領域をトランザクション処理に設定するなど，用途を限定したシステムが多い．

他の一つは汎用大型機を対象とした動きである．信頼性の向上をねらった疎結合マルチプロセッサ(Loosely Coupled Multi Processor；LCMP)などの負荷分散[5)]や通信制御装置による機能分散があり，最近ではシステム間の連結によるサービスの拡張，統合が行われている[6)]．しかし，プロセッサ間の接続方法としては，種々の方式が使い分けられており，幅広く共通に使用でき，拡張性の高いものはでていない．また通信制御装置の分担機能も，OSI参照レベルのネットワーク層，あるいはトランスポート層相当に留まるなど，分散のレベルは高くない．さらに高信頼化については，プロセッサ切り替え時間の短縮に関するプロトコルの提案もみられるが[7)]，商用のホットスタンバイ方式を採用するシステムでは，プロセッサ障害時の端末における数分間のサービス中断が避けがたい状況にある[8)]．

一方，システム構成の大規模化とともに，運転・操作は複雑化する傾向にある．人件費の節減だけではなく，人手によるミスを防止するためにも省力化，自動化，さらには無人化が望まれ，システム運転の自動化が試みられている[9)~11)]．しかし，大規模なオンラインシステムの立上げから停止まで，媒体操作を含んだすべてのオペレーションを自動化したシステムは現れていない．

本論文で報告するDIPS複合システム[12)~14)]は，ハードウェアから応用フソトウェアに至るまで一貫した構想のもとに分散処理を指向し，大規模オンラインセンタを汎用大型機を用いて構成したシステムである．90年代にわたって有用な，高い拡張性と信頼性，徹底した自動化の実現をめざしている．主な特徴は以下のとおりである．

① 高位レイヤまで通信制御機能を分担する前置プロセッサ(Front End Processor；FEP)[15)]の導入など，高度のプロセッサ分散構成の採用

† Large-scale Distributed Processing System using Loop Networks by SEISHIRO TSURUHO (NTT Software Laboratories), YASUO KINOUCHI (NTT Communication Switching Laboratories), TAKAYUKI HOSHIKO, HAJIME NAKATANI and JUNJI MIYAKAWA (NTT Communications and Information Processing Laboratories).

†† NTTソフトウェア研究所
††† NTT交換システム研究所
†††† NTT情報通信処理研究所

② 高性能光ループ[16),17)]による，幅広い用途に共通に使用可能で，拡張性が高いプロセッサ間通信機能の実現

③ システム制御用プロセッサ (System Control Processor ; SCP)[18)] や磁気テープ操作機構による大規模分散システムでの自動運転の実現[19)~23)]

④ 業務処理を分担するホストプロセッサ(HOST)の障害を，プロセッサ間の連携により端末に対して隠ぺいするなど，分散処理の特徴を生かした信頼性の大幅な改善[24)~26)]

⑤ オンラインサービス中のシステム増設の実現

DIPS 複合システムは，82 年より検討を本格化し，87 年に最初のシステムでのサービスを開始した．これまでに十数システムに適用しており，良好な結果を得ている．

本論文では，はじめに DIPS 複合システム開発の考え方と全体構成を示す．次に上記特徴を中心に，分散化がもたらすシステム全体にわたる設計上の問題と対処法について論じ，最後に導入実績を踏まえた評価結果を明らかにする．

2. システムの開発方針

DIPS 複合システム（以下複合システムと略称）は，分散処理構成を基本に以下の 3 項目を主な開発目標とした．

（a） 高い業務処理能力と拡張性の実現

（b） システム運転の自動化

（c） 端末利用者から見えるシステム障害の大幅低減

これら各項目の設定理由と開発の進め方に関する基本的な考えを以下に示す．

（1） 高い業務処理能力と拡張性の実現

従来，プロセッサ間の接続は，チャネル間接続装置 (Channel-to-Channel Adaptor ; CTCA) を介して，1 対 1 接続あるいは高々十数台規模のスター接続により実現されてきた．しかし，プロセッサ接続台数が増えると接続用装置が多数必要になること，同軸ケーブルを用いた並列転送であることから，プロセッサ間接続距離が短くて分散設置が制約されることなど，経済性，拡張性に問題がある．

一方，サービス別に建設されたシステムを相互接続して，より大きな統合されたシステムとして動作させる要求がでてきている．このため，高速のプロセッサ間接続装置が開発されてきているが，これらは主としてファイル転送など大容量データ転送への適用をねらったものである．短い電文が高頻度で交信される大規模トランザクション処理にこれらの接続手段を適用するには，この交信のためのオーバヘッド削減が重要な課題となる．

複合システムにおいては，汎用大型計算機を用いて高い業務処理能力と拡張性を同時に確保するため，以下の方針を設定した．

① 大規模トランザクション処理および大容量データ転送に統一的に適用できる高速のプロセッサ間接続方式を実現する．

② 任意のプロセッサと直接に交信可能な完全群接続系を経済的に構成可能とする．これとプロセッサ間の柔軟な接続制御により，高いシステム拡張性を実現する．

具体的な完全群接続系の構成手段としては，高負荷環境でも能力低下の少ないトークンパッシング方式の光ループを採用し，これをチャネルインタフェースでプロセッサと接続することとした．

（2） システム運転の自動化

複合システムでは，通常運転の操作から異常時の操作までを自動化の対象とした．通常運転の操作とは，システムの立上げから停止までの媒体操作を含めた運転操作を指し，異常時の操作とは，異常の監視から再構成とサービスのリカバリ操作までを含む．これらの自動運転を実現するため，以下の方針を設定した．

① システム運転に関するハードウェアとソフトウェアの運転制御インタフェースを SCP に集中し，SCP から，全サブシステム* を運転監視可能とする．

② 集中制御を行う SCP の障害によるサービスへの直接的な影響を防止するために，自動化主体を階層化する．サブシステムに閉じた運転は各サブシステムで自動化し，SCP では，サブシステム障害時の再構成処理など，システム全体にわたる制御を分担する．

③ サブシステム追加やサービス拡張などにともなう運用の変更をサービス中に可能とし，システム増設を容易にする．

（3） 端末利用者から見えるシステム障害の大幅低減

業務処理を分担する HOST では，障害時のリカバリ時間を短縮するために，現用系と予備系でデータベースを共用し，予備系のソフトウェアを立上げ済

---

* サブシステムとは，複合システムを構成する単位であり，一つの OS により制御されるハードウェアとソフトウェアを指す．

みの状態としておくホット予備方式が採用されている[6]. しかし, ホット予備方式では端末との交信再開に必要な通信パスの再設定処理が残ることから, リカバリ時間の短縮には限界がある. また, この時間は収容端末数の増加に比例して大きくなるため, この削減は重要な課題となる.

複合システムでは, 全国規模の大規模オンラインシステムへの適用を前提として, 端末利用者からみた複合システムの信頼度を左右する HOST, FEP, SCP, 光ループの信頼度設計において以下の方針を設定した.

① システム障害の大半を占める HOST 障害については障害発生を端末利用者に隠ぺいする.

② 端末は複数の FEP に分散収容され, FEP 障害の影響範囲は HOST に比べ限定される. このため, FEP 障害についてはサービス中断を許容し中断時間の短縮を図る.

③ 光ループ, SCP など集中化される部分については, 新たな信頼性上のボトルネックとならないよう, 個別に高信頼化を徹底する.

3. システム構成

前章の方針に基づき開発した複合システムの構成概要を示す.

3.1 ハードウェア構成

複合システムのハードウェア構成を図-1に示す. 本システムは, 100 メガビット／秒の光ループ（データループ）と 48 キロビット／秒の同軸ループ（制御ループ）の2種類のループをもち, データループは各プロセッサ間のデータ交信のため, 制御ループはシステム内全装置（周辺装置を含む）の電源投入・切断, 接続切り替えなどのハードウェア制御のための通信パスとして使用する.

データループには, プロセッサ間結合装置 (Processor-to-Processor Communication Interface Unit; PCI) を介して HOST, FEP, SCP などのプロセッサが接続され, 相互に直接交信可能となる. SCP は, 集中的なシステム制御を行うために開発したプロセッサである. プログラム制御または配下のコンソール（集中コンソール）により, データループを介した各プロセッサの運転管理および制御ループを介した各装置のハードウェア制御を行う.

3.2 ソフトウェア構成

複合システムのソフトウェア構成を図-2 に示す. HOST, FEP, SCP は同一アーキテクチャで, 同一の OS および通信管理プログラム (Communication Management Program; CMP) が走行する. データループを介した各プロセッサ間の通信は, ループアクセス部 (Loop Access Method; LAM) が実現し, 制御ループを介したハードウェア制御は, ハードウェアアクセス部 (Hardware Access Method; HAM) が行う.

端末との通信は HOST 上の CMP (HCMP) と FEP 上の CMP (FCMP) で機能分担している.

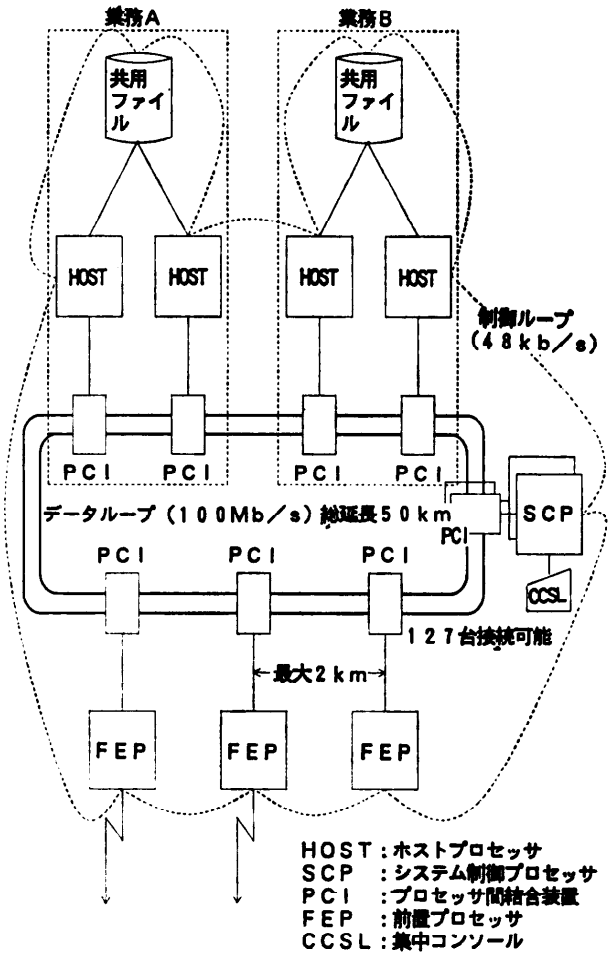

図-1 DIPS 複合システムのハードウェア構成概要
Fig. 1 Hardware configuration of DIPS large-scale distributed processing system.

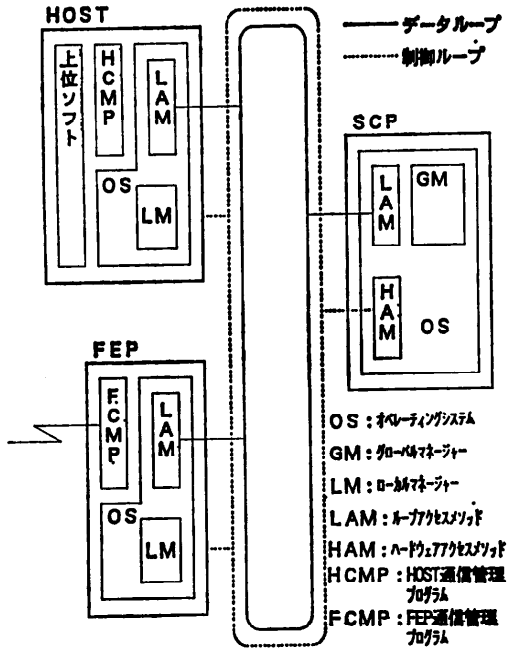

図-2 DIPS 複合システムのソフトウェア構成概要
Fig. 2 Software for DIPS large-scale distributed processing system.

複合システムの運転は，システム全体の運転を分担する SCP 上の GM (Global Manager) と各サブシステムに閉じた運転を分担する LM (Local Manager) によって階層的に管理される．なお，GM は SCP の役割を実現するソフトウェアであり，特に明記する場合を除き，以下は GM を含めて SCP と総称する．

4. 主な方式技術

複合システムで採用した特徴的な方式に関し，主な設計上の問題点とその対処について示す．

4.1 プロセッサ分散方式

(1) HOST 系の構成

HOST 系の構成は，HOST 障害の影響範囲の限定，予備 HOST コストの低減などをねらって，現用系 n 台に対し共通に 1 台の予備系を置く $n+1$ 予備の LCMP 構成を採用した．また，この LCMP 構成を単位とする複数の HOST 群を同一データループ上に接続し，HOST 群間で FEP の共用を実現した．

(2) HOST-FEP 間機能分散構成

従来は FEP に OSI 参照モデルのレイヤ 3 または 4 以下のプロトコル処理を分担させるものが多いが，複合システムでは以下の考え方の下に，同モデルのレイヤ 7 相当までを FEP で機能分担することとした．

① HOST 障害の隠ぺいによる高信頼化

レイヤ 7 相当までを機能分担することにより，FEP で端末との通信パスを保持可能となる．これにより，HOST 障害を隠ぺいできる (4.4 参照)．

② 通信制御用プロセッサによる経済化

FEP は，HOST に比べてチャネルや浮動小数点演算機能などを簡素化でき，単位性能当たりのハードウェアコストを HOST より低く抑えることができる．このため，高位レイヤまで FEP で分担したほうが，システムとしての経済化効果が大きい．

(3) 監視系の構成

HOST, FEP 全体の運転管理および障害処理は，専用の独立したプロセッサで実現することとした．これにより，制御するプロセッサと制御されるプロセッサとを分離し，ソフトウェア制御の簡素化を図った．

4.2 サブシステム間通信方式

(1) データループの通信性能

複合システムでは，プロセッサ間のすべての通信がデータループを介して行われるため，データループの通信性能がシステム性能上のネックとならないようにする必要がある．この性能を規定する主な要因として，データループ上の単位時間当たりのデータ転送量と，PCI での単位時間当たりのデータ転送回数がある．

まず，最大データ転送量を決定するループ伝送速度については，以下の理由により 100 メガビット／秒に設定した．

① トランザクション処理用の短電文 (数百バイト程度) では短時間のピークトラフィック時でも伝送路ネックとならないこと．

② トランザクション処理と並行して大容量のファイルデータの多重転送が可能なこと．

③ 光モジュールとして経済性に優れた発光ダイオード (送信側) ／ピンフォトダイオード (受信側) が適用可能なこと．

次に，PCI でのデータ転送回数については，通常データ転送制御のためのプロセッサ走行時間 (T_{CPU}) と，PCI-PCI 間のデータ転送時間 (T_{PCI})* が主な要因となるが，プロセッサのデータ転送制御と PCI-PCI 間のデータ転送を並行して独立に行うことを可能とし，T_{PCI} のみでデータ転送回数を決定可能とした．

* T_{PCI} は，PCI 内部の伝送制御時間，チャネルおよびループ伝送路とのデータ送受信時間，ループ伝送遅延時間，トークン獲得待ち時間から構成される．

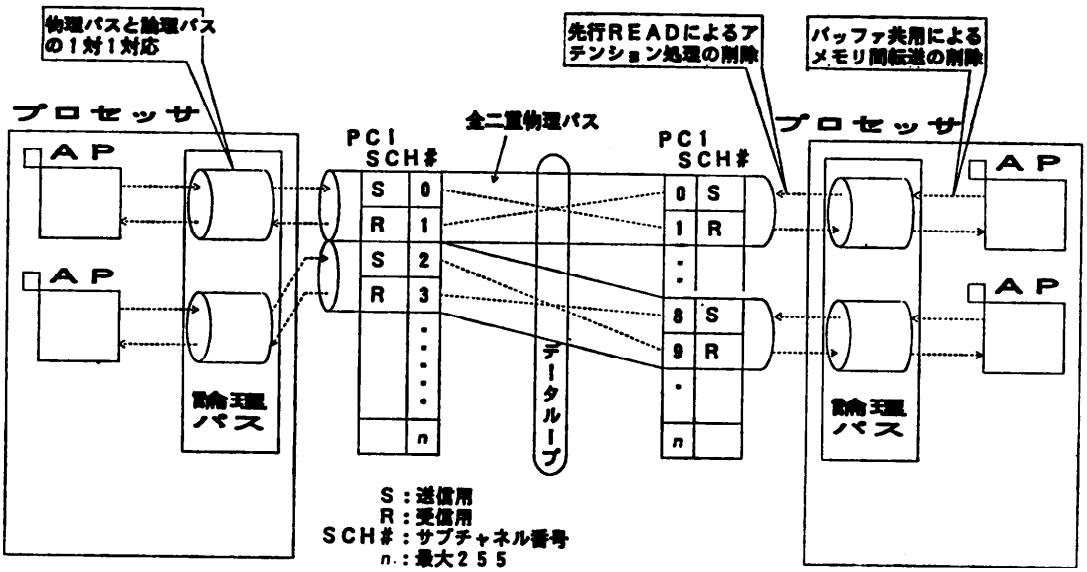

図-3 サブシステム間通信の基本方式
Fig. 3 Basic concept of transmission between subsystems.

ただし，データループがシステムの信頼性，性能上のボトルネックとなるのを回避するため，PCI レベルで確実に送達確認を行い，上位レイヤでの送達確認を軽減可能とした．このため，T_{PCI} には PCI での送達確認のための時間が含まれている．

以上のことを踏まえ，PCI での所要データ転送回数（N_{PCI}）は，プロセッサ当たり1台の PCI 接続を原則として，800 回／秒（転送データ長：200 バイト）に設定した．この理由は以下のとおりである．

① プロセッサ当たりの最大トランザクション処理件数は，最大性能のプロセッサで処理されるトランザクション件数と，トランザクション当たりのデータ転送回数の積で求まり，800 回／秒となる．

② $N_{PCI}=800$ 回／秒で，$T_{PCI}=1.25$ ミリ秒となる．PCI は1回のデータ転送時間に送達確認の時間を含むが，後述の伝送制御手順の採用，ファームウェア制御法の工夫などにより，この時間内に実現可能との見とおしを得た．

（2） 通 信 方 式

データ転送制御におけるソフトウェアのダイナミックステップ（DS）の削減に有効な方式と，所要転送能力を満足する効率の良い伝送制御の実現が必要となる．

（a） データ転送制御における DS の削減

従来の CTCA などを介した方式では，上記 DS が全トランザクションの処理 DS の 5～15% 程度を占めている．この DS 削減および PCI-PCI 間のデータ転送とプロセッサ上のデータ転送制御を並行して行うため，図-3 に示す手法を採用した．本手法の特徴は，以下のとおりである．

① PCI 内に，サブチャネルを最大 256 個用意した．このサブチャネルは，相手 PCI サブチャネルと対を成して物理的な通信パス（物理パス）を設定し，物理パス対応に論理的に独立したデータ送受信動作を実現する．LAM が，論理パスとこの物理パスとを1対1に対応づけることにより，異なる論理パス対応のデータ転送制御と PCI-PCI のデータ転送動作を同時に実行可能とした．また，物理パス1本に複数の論理パスを対応づける従来方式と比べると，電文の振り分け，メッセージバッファ間の電文移送などの処理を不要とした．

② 二つの物理パスを全二重パスとして使用することにより，送信データの衝突による再送や，これを避けるための送信権制御などの処理を不要とした．

③ 全二重パスの受信側サブチャネルを，LAM が先行して受信可能な待合せ状態に設定することにより，相手からの通信起動時の割込み処理を不要とした．

（b） 効率の良い伝送制御の実現

大容量データ転送を，特定のデータ転送が長時間

ループを保留する沈み込み現象の発生なく，短データ転送と混在して行うため，以下の方式を採用した．

① 1キロバイト (KB) を超える長データは PCI ハードウェアが自動的に1KB に分割して転送する．プログラムは最大転送データ長を意識しない．

② 各 PCI はトークン獲得後，最大1KB の1フレームを送信するとただちにトークンを解放する．

以上①と②を組み合わせることにより，一定の短い周期内に各 PCI が均一の送信機会を獲得でき，沈み込みを回避できる．また，チャネル転送速度（3メガバイト／秒）と光ループ伝送速度（12.5 メガバイト／秒）に約1対4の性能差があるため，最大4対の PCI 間でチャネル転送速度を落とすことなく同時にデータ転送可能となる．なお，高速 LAN の国際標準である光ファイバ分散データインタフェース (FDDI) においても，②と類似した手順を採用しているが，最大データ長制限があり，大容量データ転送時は，通信制御プログラムでデータの分割，組立を必要としている[27),28)]．

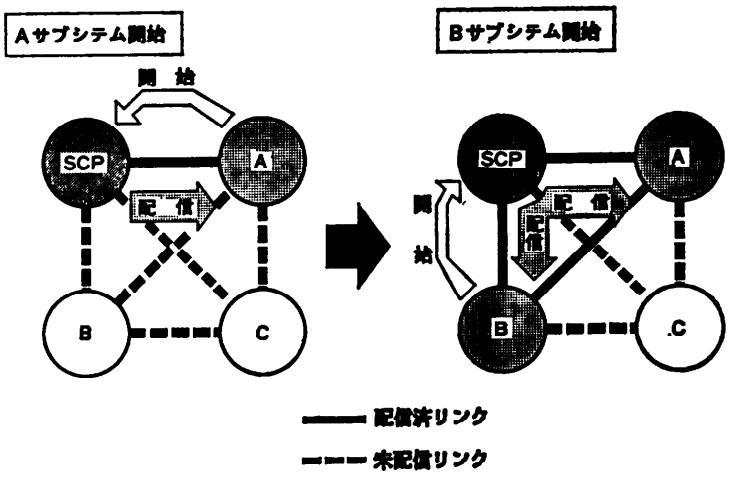

図-4 サブシステム間接続制御方式
Fig. 4 Link control between subsystems.

（3） 接続制御方式

高いシステム拡張性を確保するためには，ループの採用による接続構成の単純化に加え，拡張性のあるソフトウェアでの接続管理が必要となる．ソフトウェアの管理する接続情報であるサブシステム間の物理パス情報の追加変更を容易にすることをねらいに，以下の方式を採用した．

① サブシステム間の物理パス情報を SCP で一元管理する．

② 物理パス情報は，サブシステムの立上げを契機に SCP から配信し，サブシステム間の物理パスを図-4 に示すように段階的に拡大していく．

本方式を実現するため，PCI に物理パスを複数設定できる点を利用し，サブシステム立上げ時に SCP へ開始通知を送るためのパスと，SCP から物理パス情報を送るためのパスを，専用の物理パスとして用意した．

4.3 システム運転の自動化方式[22),23)]

多数のプロセッサから構成される分散システム運転の自動化に向けて，その前提となる集中化方式，自動化に当たっての考え方と主な技術を以下に示す．

（1） システム運転の集中化

SCP からシステム全体を運転制御するため，各サブシステムのコンソール機能および各プロセッサと周辺装置のハードウェア制御機能を SCP に集中した．コンソール機能は，各サブシステムに論理的なコンソールを定義し，データループを介して SCP の管理する集中コンソール画面に対応づけて実現した．ハードウェア制御機能は，制御ループを介して集中化するが，装置ごとに仕様が異なっているため，装置の違いを SCP のハードウェアアクセス部 (HAM) で吸収し，汎用的な電源制御や切り替え制御コマンドとして実現した．また，サブシステムのソフトウェアとハードウェアの障害を統一して早期に検出できるように，制御ループを介してサブシステムから SCP へ通信する命令を各プロセッサに付加した．サブシステム停止処理の過程でこの命令を実行することにより，自動的に SCP への停止通知を送ることができる．

（2） システム運転の自動化方式

通常運転の操作から異常時の操作まで自動化するため，以下の機能を実現した．

① コンソール操作の手順化

システム運転におけるオペレータ操作は，コンソール操作と媒体操作からなる．コンソール操作手順を簡

単に記述できる運転手順記述言語 (Operation Procedure Language; OPL) と，OPL で記述された運転手順 (Operation Procedure; OP) に従って，メッセージまたは時刻を契機にコマンド投入を行う機能を開発した．本機能は SCP を含む全プロセッサ上で動作可能とし，サブシステムに閉じた運転はサブシステムの OP で自動化した．また，運転手順はサービスの追加などに応じて変更する必要が生じるため，OP の実行方式として，OP 間のリンクが不要なインタプリタ方式を採用した．これにより，動作中に変更の生じた OP の部分入替えを可能とした．

② 異常処理の自動化と高信頼化

複合システムでは異常時の回復操作を自動化するため，プロセッサと周辺装置は，二重化するか，あるいは共通の予備を設置した．また，異常の原因となる障害を，サブシステム内で縮退することで継続可能な障害と予備切り替えを必要とする障害に大別し，前者の処理をサブシステムで，後者の処理は SCP の OP で分担した．

予備構成の異なる多数のサブシステムを集中監視する SCP では，複数の異常通知が同時あるいは，短時間に連続的に生じても，誤動作なく個々の異常処理を正しく完遂できる必要がある．このため，OP の構成を監視部と処理部の二層構成とし，先に通知された異常の処理が完了するまで，次の異常を監視部でキューイングし，先行異常処理を優先して完遂する方式とした．これにより，たとえば，HOST と FEP で同時に障害が発生しても，両障害を順次自動回復できる．

複合システムでは，異常監視自体が自動化されるため，自動監視系の高信頼化が重要となる．このため，SCP，データループ，制御ループをそれぞれ二重化するとともに，両ループの正常性を定期チェックする機能を付加し，自動監視系自身の高信頼化を図った．

③ 媒体操作の不要化

通常運転時の媒体操作としては，ジャーナル取得，全ダンプ，バッチ処理などにともなう磁気テープ操作が大半を占め，かつ運転上の大きな負担となっている．このため，カートリッジテープ (VHS ビデオテープ) を用い，保管庫と運転操作用アクセッサを備えた集合型大容量記憶装置 (Compact Mass Storage System; CMSS)[14),15)] を開発し，これを用いて媒体操作を自動化した．

4.4 システム高信頼化方式[24)~26)]

(1) サブシステムの高信頼化対策

(a) HOST 障害

トランザクション処理において端末に固有な応答監視時間（通常 1 分）以内に応答電文を返却できれば，端末利用者に HOST 障害発生を隠ぺい可能となる点に着目した．このためには，まず HOST リカバリ時間を 1 分以内に短縮する必要がある．リカバリ処理の高速化のために，複合システムでは $n+1$ のホット予備方式を採用した．しかし，障害を隠ぺいするには，さらに以下の問題を解決しなければならない．

① 通信パスの再設定処理にともなう端末の初期化により利用者の処理が中断するほか，収容端末数の増大にともなって，端末との通信パス再設定処理時間が増大する．

② メモリ復旧時間を短縮するためにホット予備 HOST のメモリをオンライン処理と並行して事前に更新しておく方式は，現用系 n 台に対する共通予備を前提とした場合，予備系で走行するバッチ処理への影響があり実現性が乏しい．

複合システムでは，HOST，FEP の機能分散をベースとする図-5 に示すサービス高速再開方式により，これらの問題を解決した．

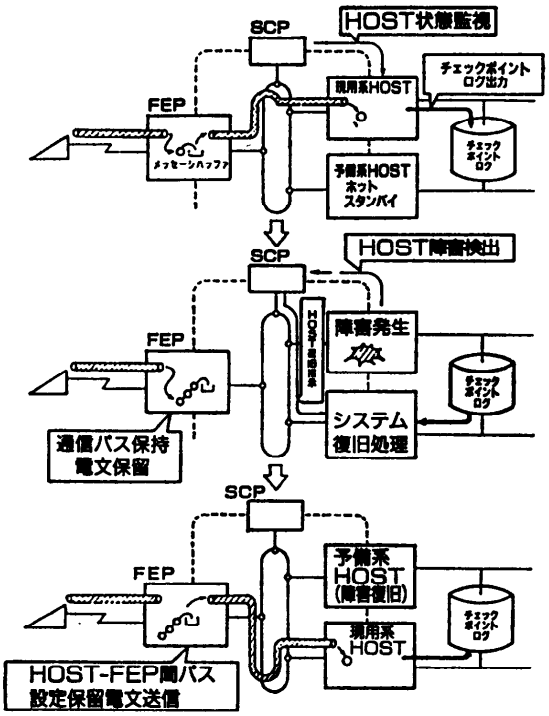

図-5 サービス高速再開方式の概要
Fig. 5 Outline of rapid service recovering method.

① FEP で OSI 参照モデルのレイヤ7相当の機能までを分担させ，HOST 障害発生時に FEP で通信パスを保持し，端末との通信を継続し受信した電文を保留することとした．これにより，端末とのパス再設定を行わずに，FEP と予備 HOST の端末に関する状態の一致処理のみ高速に実施し，保留電文を FEP から HOST へ再送することにより，通信系のリカバリ時間を大幅に短縮した．

② メモリ復旧用とデータベース復旧用のジャーナルデータとを別媒体に分離取得するとともに，半導体メモリ装置などの高速ファイル装置の導入によりメモリ復旧処理時間の短縮を図った．

(b) FEP 障害

これまで，FEP には特定の系を予備系とする固定予備方式を採用していたが，24 時間連続サービスに対応するため FEP 稼動率の向上をねらいとして，FEP と回線群との自由な組合せを実現する完全群スイッチを導入し，任意の FEP を予備系とできる浮動予備方式を採用した．さらに，予備切り替えを高速化するため，予備 FEP のホット予備化を図った．

(c) SCP 障害

SCP はシステムに対し受動的に動作するため，SCP の単一障害は業務処理への影響を及ぼさず，HOST，FEP との重複障害が発生した場合にのみ限られる．このため，SCP では，デュプレックス構成を可能とすることにより SCP 障害の端末利用者への影響を排除した[23]．

(d) PCI／光ループ障害

光ループ障害はサブシステム間の通信が不能となることからシステム動作に致命的な障害となる．また，複数のサブシステムにまたがる障害として検出されるため，障害の特定，復旧に時間がかかる恐れがある．このため，ループ伝送路障害については，永久障害から一時障害に至るまで，予備ループ切り替え，ループバック切り替えなどによりハードウェアで自動復旧し，システム動作への影響を除去した[16),17)]．

(2) システムの信頼度

上述の高信頼化対策によるシステム信頼度の向上効果を，端末利用者からみたサービス中断時間で予測すると以下となる．

(a) 従来方式

端末利用者からみたサービス中断時間は，HOST 障害による中断と FEP 障害による中断からなる．さらに，固定予備方式では予備 FEP の再組込み時に FEP の停止が必要になることから，FEP 切戻しによる中断が加わる．従来，FEP 切戻し処理はサービスの終了後に実施可能であったが，24 時間連続サービスのシステムではサービス中に実施せざるを得ず，FEP 切戻しは障害発生と同じ影響をもたらすことになった．サブシステム二重障害を無視し，単位時間当たりの HOST および FEP の障害発生件数を F_H, F_F, HOST および FEP のリカバリ時間を T_H, T_F, FEP 切戻しの中断時間を T_S として端末からみたサービス中断時間 T_{down} を求めると(1)式となる．

$$T_{down}=F_H\cdot T_H+F_F\cdot T_F+F_F\cdot T_S \quad (1)$$

(b) 本方式

HOST 障害発生は隠ぺいされること，予備 FEP の浮動化により FEP 切戻しが不要となることから，サービス中断時間は FEP 障害による中断のみとなる．したがって，サービス中断時間を求めると(2)式となる．

$$T_{down}=F_F\cdot T_F \quad (2)$$

デュプレックス方式採用時の HOST リカバリ時間を5分，固定予備方式採用時の FEP リカバリ時間および FEP 切戻しによる中断時間を2分，HOST 障害発生件数を FEP の3倍，FEP リカバリ時間は従来方式の約 1/2 に短縮されると仮定すると，本方式によるサービス中断時間は約 1/20 となる．

4.5 システム増設方式

システム増設は以下の4段階に分けることができる．

ステップ1 (ハードウェア増設／確認)
ステップ2 (基本ソフトウェア情報の変更／確認)
ステップ3 (業務ソフトウェア情報の変更／確認)
ステップ4 (実サービス／運用への組込み)

従来のデュプレックス方式では，ステップ1は，予備系への増設後，系切り替えを行う方式により，ステップ2，3，4は，増設装置を含めたシステムファイルを作成しておき，増設後実装化する方式を採用している．この方式は，コールド予備切り替えのため中断時間が長く，また，将来のサービス条件を正確に予測してシステムファイルを事前に作成しなければならないなどの制約があった．複合システムでは，ホット予備切り替えにより切り替え時間を短縮でき，かつ予備系が別のシステムファイルで動作するため，新しくシステムファイルを作り直すことも可能となる．しかし，サービスで使用中のループへ直接増設工事を行う必要があり，ループ増設工事や増設後の確認試験にともな

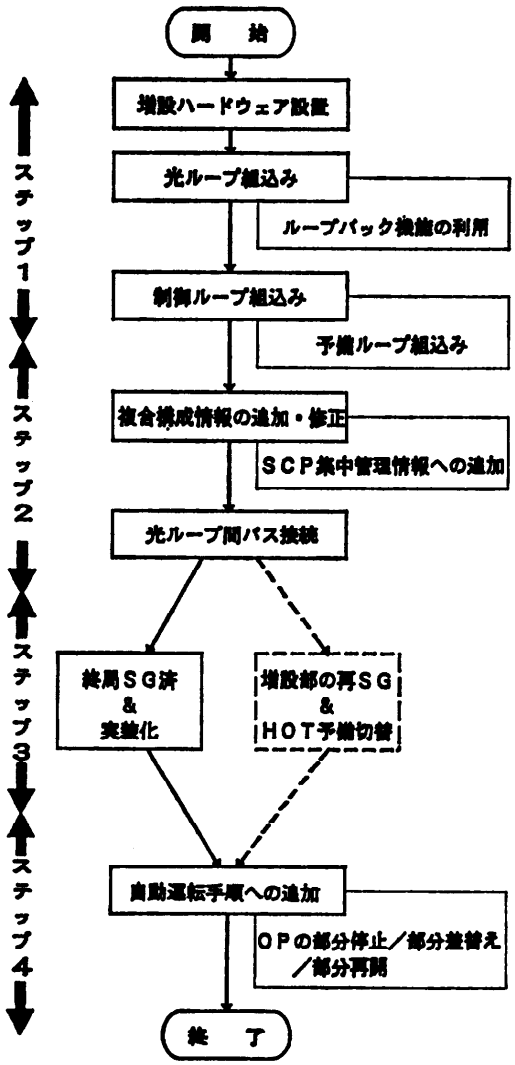

図-6 サブシステム増設手順
Fig. 6 System expansion procedure.

う現用サービス系への擾乱防止が重要となる．このため，以下の機能を実現した．

（a） 光ループ伝送路の二重化と PCI のループバック機能により，ループ切断による現用系への擾乱を防止する．制御ループについては，SCP と対で二重化し，予備ループへ組込み後，SCP の現用／予備切り替えを行う方式とした．図-6 に，本複合システムでのサブシステム増設手順の概要を示す．

（b） 現用系への組込み前に，増設サブシステム群に閉じた試験環境を作るため，図-7 に示すループ分割試験機能を実現した．PCI の複数サブチャネルを用いて1本の光ループを最大4本の独立した論理ループに分割し，試験サブシステム群に閉じたデータループ交信試験を可能とした．

5. 評 価

これまでに述べた主な方式技術の評価を行う．

5.1 プロセッサ間通信能力

PCI を介したデータ転送制御における DS 削減のために採用した通信アクセス法および高効率の伝送制御手順を使って得られたデータ転送能力を示す．

図-8 に上記通信アクセス法の適用効果を示す．本アクセス法を用いてデータ転送制御の高速化を図った結果，従来の CTCA などを介したデータ転送制御用 DS と比べて，約 30%の削減を達成した．これは全トランザクション DS の2～5%程度の削減になる．

次に，データループ性能については，これまで大規模オンラインシステムに数多く適用してきているが，システム性能上のネックにはなっていない．具体的な測定結果として，図-9 に PCI を介したデータ転送回数と転送データ長の関係を示す．転送データ長 200 バイトの場合に所要データ転送回数（800 回/秒以上）を満足していることが分かる．また，長データ域では，転送データ長 10 キロバイトの場合でループ使用率 18% を実現している．この使用率は，チャネル転送速度の 75% に当たる．

5.2 システム運転の自動化・省力化

システムの立上げから停止までの通常運転から，異常処理については，一重障害レベルの障害処理までの自動化を十数システムで実施し，これまで問題なく運転されてきている．また，多重障害などについても，先に発生した異常処理の完遂や，オペレータへの事象引継ぎ情報の作成などの処理を OP に組み込むことによって一部自動化できている．

省力化の効果としては，大規模化によるセンタ集約化，媒体操作とコンソール操作の不要化によるオペレータ要員の削減として，12人で運転していたセンタを，通常2人に省力化できた例がある．また，ネットワークサービス系の 24 時間運転システムでは，SCP による通常運転の自動化と集中コンソールの遠隔設置により，センタの無人化を実現している．

5.3 システム信頼度

（1） HOST リカバリ

約1テラバイトのデータベースを備え，約1万5千端末を収容し，毎秒約100件のトランザクションを処

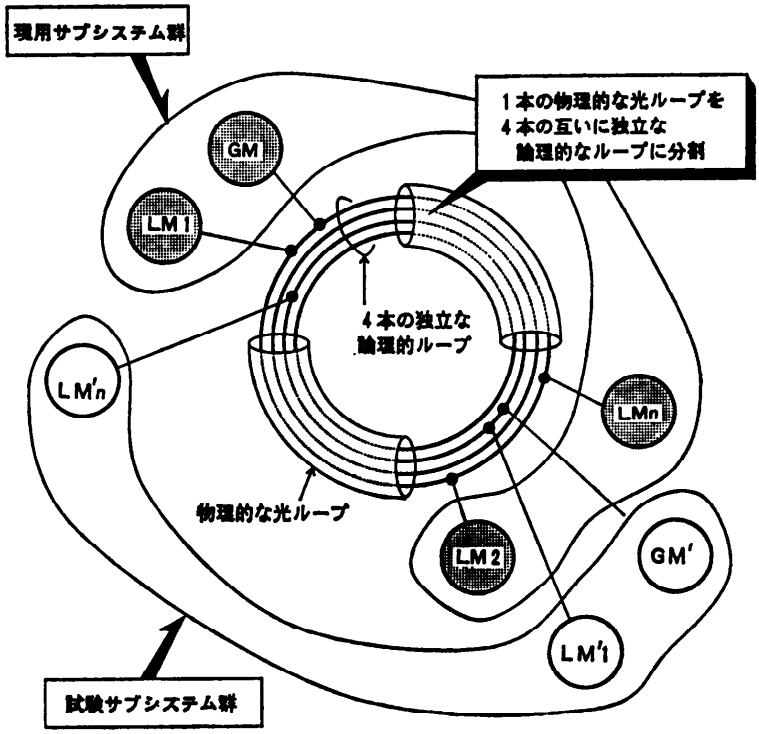

図-7 ループ分割試験機能
Fig. 7 Test on partitioned loop.

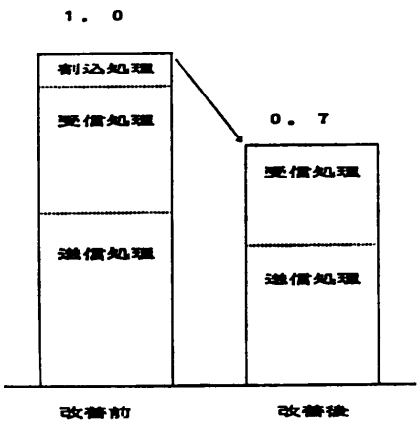

図-8 通信アクセス法の適用効果
Fig. 8 Effect of transmission method via PCIs.

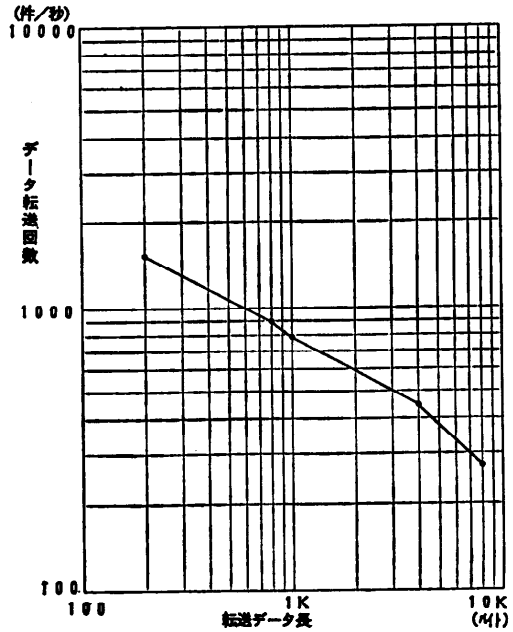

図-9 PCI を介したデータ転送回数
Fig. 9 Number of data transmissions via PCIs.

理する全国規模のオンラインシステムに本方式を適用した結果，以下が確認された．

図-10 に示すようにシステム復旧時間を1分以内とすることができ，ほぼすべての端末利用者に HOST 障害発生を隠ぺいできた[25),26)]．通信系のリカバリ時間

は，FEP と端末との間の通信パス再設定処理が不要で，従来2～3分要していたものが十数秒まで短縮された[25]．本方式の採用によりハードウェア障害のほか，ソフトウェア障害，誤操作が起因となる HOST 障害発生を端末利用者に隠ぺいでき，特に不安定性の残るサービス開始初期から高品質なサービスの提供が可能となった．

（2） FEP リカバリ

前述のシステムで FEP 切り替え時間を評価した結果ホット予備の採用により，プログラムと制御データのローディングに要する時間がなくなったことから，図-10 に示すように，従来の約 1/2 の約1分で FEP 切り替えが可能となった[26]．

（3） その他のサブシステム障害

複合システムはこれまで延べ約20万時間* の運用実績をもつが，SCP 障害がサービスへ影響を及ぼした例はない．また，光ループについても延べ約 200 万時間**の運用実績をもつが，これまでアダプタの光モジュール障害による予備ループ切り替えやループバック発生の事例はあるものの，サービスへの影響は生じていない．

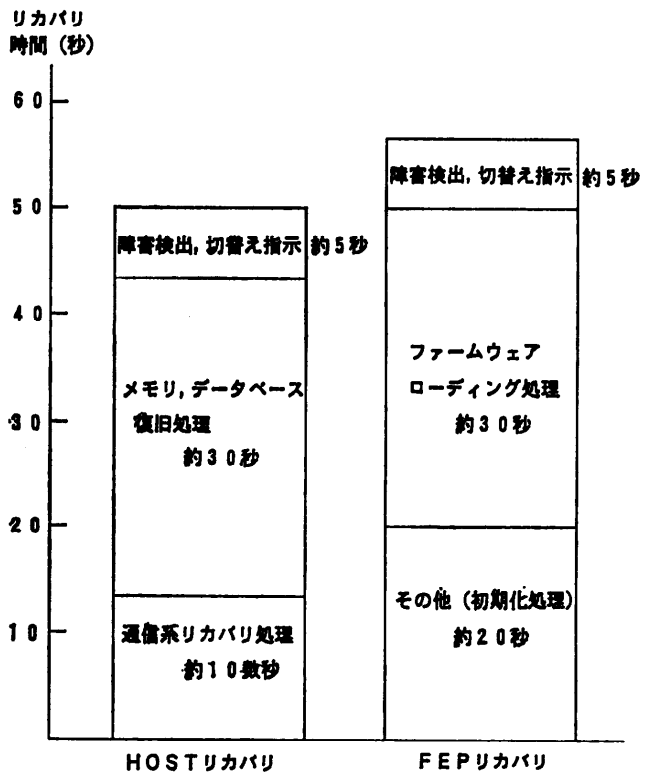

図-10 HOST および FEP リカバリ時間の内訳
Fig. 10 Items of HOST and FEP recovery time.

5.4 システムの増設性

24 時間連続のネットワーク交換型サービスを行う9つのサブシステムからなる複合システムで，処理能力の拡大を目的として 16 サブシステムのオンライン増設を 4.5 に示す手順に従って実施し，サービス無停止での増設を確認した．増設に要した期間は，ステップ1，2の増設工事と基本ソフト組込みが，2日間（延べ8時間）で，その後約1カ月間のオンラインサービスとの並行試験の後，現用サービスへの追加を図った．さらに，サービス品目の追加変更など増設性を向上するために，システム特性の検証機能や試験機能の拡充について検討を進めている．

6. おわりに

本稿では DIPS 複合システムに関し，開発のねらいとシステム構成，主な方式技術とその評価について述べた．

本システムは，大規模分散処理システムを構成する上で有効な方式として，全国規模のオンラインシステムをはじめ，大型のオンラインシステムに数多く適用されている．

分散処理システム特有の問題として，システム開発時のデバッグの複雑化，試験環境設定の難しさがある．これに対処するため，仮想計算機技術を用いてデータループと多数台のプロセッサを含む複合システム全体を一つのマシン上に仮想的に構成する方式など，支援系の整備を並行して進めている．また，多様な計算機接続のニーズに対処できるオープンシステム化についても検討しており，引き続き一層の拡充，展開を図っていく予定である．

謝辞 日頃ご指導をいただく NTT 情報通信処理研究所の石野福彌所長をはじめ，本システム開発にご協力いただいた同研究所の松永俊雄研究企画部長ならびに関係各位に深謝します．

* 各システムの運転時間の総和．
** 各システムの PCI 稼動台数と運転時間との積の和．

参考文献

1) Daniel, E. L.: A Highly Integrated, Fault-Tolerant Minicomputer: The Nonstop CLX, Digest of Papers Compcon Spring, pp. 514-519 (1988).
2) グレイ, J. ほか: フォールト・トレラント・システム, マグロウヒルブック (1986).
3) 飯島: 耐故障機能を強化した疎結合マルチプロセッサ方式汎用コンピュータ, 日経コンピュータ, 1986. 7. 21.
4) Nancy, P. K. et al.: VAX Clusters: A Closely-Coupled Distributed System, ACM Trans. Computer Systems, Vol. 4, No. 2, pp. 130-146 (May 1986).
5) 宮島ほか: 高信頼度オンライントランザクションシステム, 情報処理学会オペレーティングシステム研究会報告, 43-1 (1989).
6) Cristensen, G. S. et al.: Design and Analysis of the Access Protocol for HYPERchannel Networks, 3rd U. S. A-Japan Comput. Conf., pp. 86-93 (1978).
7) 平田ほか: ホスト計算機の高速切替方式とそのプロトコルの提案, 情報処理学会論文誌, Vol. 30, No. 6 (1989).
8) 汎用コンピュータ, 信頼性向上の手法＊システム編 ホットスタンバイシステム, 日経データプロ, 1988. 3.
9) 小口: 真の自動化を目指して進むコンピュータ運用, 日経コンピュータ, Vol. 20, No. 60 (1988).
10) 今井ほか: 日本電気の自動運転システムへの取組み, C & C System Review, No. 9 (1988).
11) 広瀬ほか: 大規模分散ネットワークシステムの自動運転とリモート保守, FUJITSU, Vol. 39, No. 6 (1988).
12) 松永ほか: DIPS 複合構成システムのハードウェア, NTT 電気通信研究所研究実用化報告, Vol. 35, No. 3, pp. 257-266 (1986).
13) 鶴保ほか: DIPS 複合システムのソフトウェア構成法, NTT 電気通信研究所研究実用化報告, Vol. 36, No. 8, pp. 1021-1030 (1987).
14) Tsuruho, S. et al.: Methods for Achieving Integrated Operation in a High-Performance Optical Loop Intercomputer Communications System, FJCC, pp. 1050-1055 (1986).
15) 武井ほか: DIPS 1625 通信制御装置の実用化, NTT 電気通信研究所研究実用化報告, Vol. 35, No. 4, pp. 395-404 (1986).
16) 星子ほか: 100 Mb/s 光トークンリングを用いたプロセッサ間結合システム, 情報処理学会論文誌, Vol. 27, No. 4 (1986).
17) Hoshiko, T. et al.: A 100 Mb/s Optical Token Ring Network Suitable for High-speed Interprocessor Communication, ICDCS 87, pp. 382-389 (Sep. 1987).
18) 森ほか: DIPS 複合構成システムシステム制御装置, NTT 電気通信研究所研究実用化報告, Vol. 35, No. 3, pp. 267-274 (1986).
19) 山口ほか: MT 操作自動化方式に関する一検討, 58 年度電子情報通信学会全国大会 (1983).
20) 板生ほか: 磁気テープ記憶自動化システムの方式構成, NTT 電気通信研究所研究実用化報告, Vol. 35, No. 7, pp. 697-704 (1987).
21) 倉田ほか: DIPS 106 OS 超大容量記憶装置 (CMSS) 制御方式, NTT 電気通信研究所研究実用化報告, Vol. 36, No. 8, pp. 1073-1080 (1987).
22) 鶴保ほか: DIPS 複合システムにおける運転管理方式, NTT 電気通信研究所研究実用化報告, Vol. 36, No. 8, pp. 1031-1040 (1987).
23) 鶴保ほか: 大規模複合システムにおける運転管理方式, 電子情報通信学会論文誌採録決定 (1990).
24) 仲谷ほか: DIPS 複合システムにおけるホット予備制御方式, NTT 電気通信研究所研究実用化報告, Vol. 36, No. 8, pp. 1041-1050 (1987).
25) 川原ほか: システム高速再開における端末無中断方式, 情報処理学会論文誌, Vol. 30, No. 2, pp. 214-225 (1989).
26) 鶴保ほか: 大規模分散処理システムの高信頼化方式について, 電子情報通信学会論文誌, Vol. J73-D-I, No. 2, pp. 235-244 (1990).
27) Joshi, S. et al.: New Standards for Local Networks Push Upper Limits for Lightwave Data, Data Communications, pp. 127-138 (July 1984).
28) Ross, F. E.: FDDI—An Overview, Compcon Spring 87, pp. 434-440 (Feb. 1987).

(平成元年8月31日受付)