U.D.C. 681.325:681.32.067

リレーショナル データベース プロセッサ “RINDA”の開発

Relational Data Base Processer “RINDA”

谷口伸博* *Nobuhiro Taniguchi*
矢田 潔* *Kiyoshi Yata*
木下 理* *Osamu Kinoshita*

RDB(Relational Data Base)は，操作性，保守性の良さから導入顧客数が増加傾向にあるが，用途によっては性能面で顧客要求を満足していない。

RDB専用プロセッサRINDA(Relational Data Base Processor)は，日本電信電話株式会社(NTT)の汎(はん)用計算機DIPS V/5E(Denden Information Processing System V/5E)シリーズのRDB処理を高速化し，その適用業務拡大をねらって開発された。

RINDAを導入することにより，ベンチマークテストで従来に比べ10～100倍の性能向上を実現できた。

1 緒 言

近年，情報処理システムで中央処理装置，外部記憶装置の高速化・大容量化を背景に，意思決定支援，企業戦略立案支援など処理業務の多様化が進み，加えてEDP要員の不足もあり，データベースは操作性，保守性に優れたRDB(Relational Data Base)の導入が増加しつつある[1]。しかし，性能面の問題から実運用上は使い方が制約されており，これがRDBの普及を抑制する一因となっている。今回，日本電信電話株式会社(以下，NTTと言う。)仕様に基づきNTT汎(はん)用計算機DIPS V/5E(Denden Information Processing System V/5E)シリーズ専用のRDB処理プロセッサRINDA(Relational Data Base Processor)を開発し，RDB処理の高速化を実現した。

2 開発の背景

RDBではデータは表形式で表現され，表の任意の行と列に対する集合演算でデータ操作が実現できる。RDBの演算処理例を**図1**に示す。このようにRDBでは「表」という一般概念でデータ構造が表現されており，ユーザーがデータの格納構造を意識する必要がないため，データベースの操作性，保守性に優れている。RDBに対するデータ操作は，要求された所定の条件を満足する行と列の抽出や複数の表の結合のために，関係演算の組み合わせとして実行される。

従来の汎用計算機上に構築されたRDB管理システムでは，

(1) 外部記憶に格納された表形式の大容量データの入出力アクセス時間が大きい。

(2) 表の結合時に複数の表の所定の列どうしの比較のためのソート処理でのCPU負荷増大

という性能上の問題がある。このため，使用頻度の高い列にインデクスを付加して高速化し，表の合成は他の処理との競合が少ない深夜だけ実行といった対応をしているが，インデクス作成のためのエクスパートが必要で，RDB本来の特長であるデータを多面的にアクセスするには性能不足という問題が残っている。

RINDAは，RDBの特長である自由度の高い検索処理(検索条件があらかじめ決まっていない非定型処理)の高速化を目的として開発した。RINDAシステムの外観を**図2**に示す。

3 RINDAの構成

RINDAは，指定された表の条件に一致する行の抽出(選択処理)，所定の列の抽出(射影処理)の高速化を目的としたCSP(内容検索処理機構)，複数の表を結合するための列値の大小順に並べ替え(ソート処理)，不用な行の除去(ふるい落とし処理)，および列値による行のグループ化，の高速化を目的としたROP(関係演算処理機構)から構成される。

CSP，ROPの機能一覧を**表1**に示す[2]。

CSPは磁気ディスクに格納された表の読み出しと並行して選択・射影処理を実行することにより，従来CPUで実行していた選択・射影処理時間を削減することをねらっている。

* 日立製作所 神奈川工場

在庫表

商品番号 (I.GNO)	商品名 (GOODS)	単 価 (PRICE)	在庫量 (QTY)
210	カーテン	15,000	230
311	じゅうたん	30,000	250
410	事務用トレー	10,000	2,500
420	事務用いす	15,000	135
431	事務用机	28,000	175
501	ドレッサー	50,000	35
541	長いす	100,000	70
550	応接セット	150,000	45
554	キャビネット	30,000	50

売上表

伝票番号 (SNO)	顧客番号 (CNO)	商品番号 (S.GNO)	売上量 (SQTY)
311	132	410	500
312	132	420	50
313	132	431	35
314	132	541	40
315	132	550	10
321	111	431	70
322	131	311	15
323	131	410	450

射影 {GOODS.QTY}
＋
選択 {QTY≦50}

商品名 (GOODS)	在庫量 (QTY)
ドレッサー	35
応接セット	45
キャビネット	50

射影 {PRICE,SQTY}
＋
結合 {I.GNO＝S.GNO}
＋
グループ化 {GNO}

単 価 (PRICE)	売上量 (SQTY)
30,000	15
10,000	950
15,000	50
28,000	105
100,000	40
150,000	10

図1 RDB(Relational Data Base)の演算処理例 関係演算の組み合わせで必要な表を生成できる。

ROPは表の列ごとの並べ替え，グループ化に必須(す)なソート処理をハードウェア化することによって，CPU負荷の大幅な削減をねらったものである。

RINDAには最大CSP，ROP各2台を搭載することができ，おのおのDIPS複合インタフェースを介してDIPS V/5Eシリー

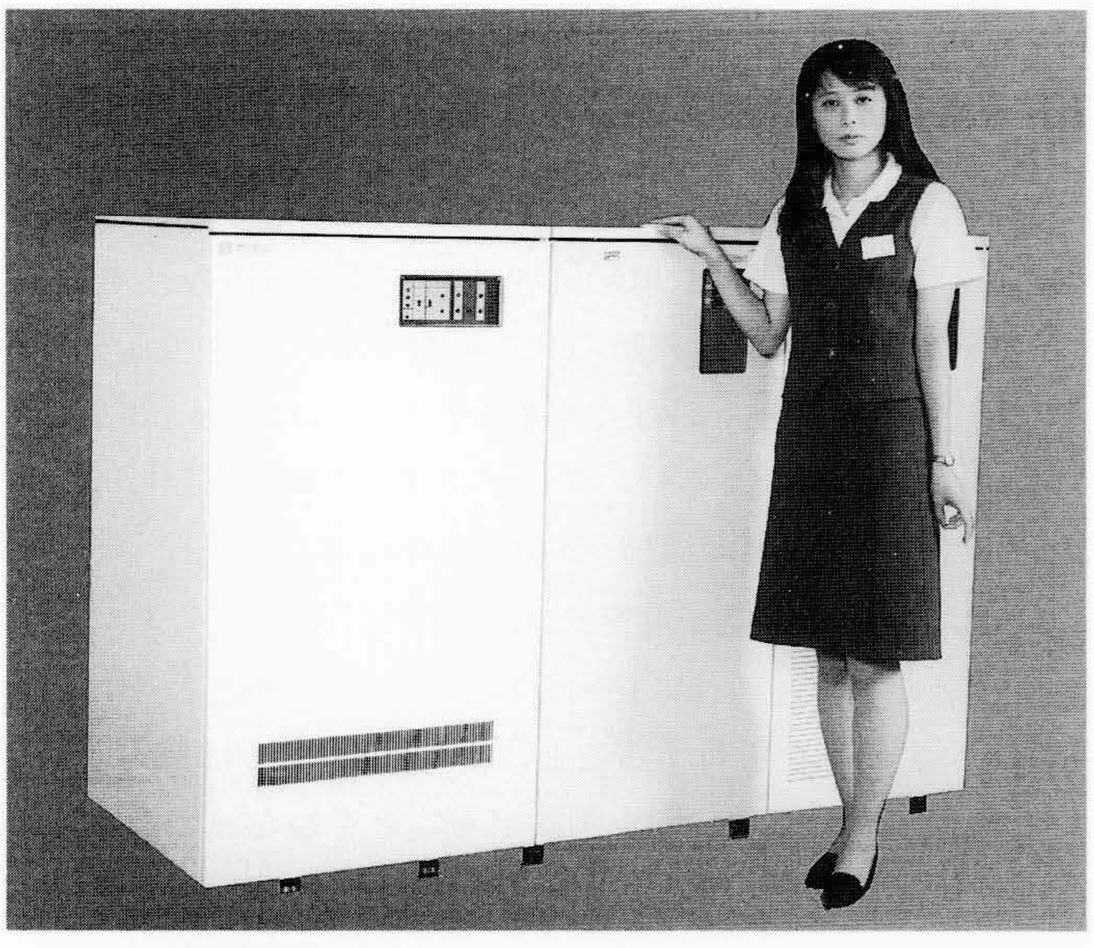


図2 RINDAシステムの外観 左端がRINDA，右端はディスク駆動装置，中はディスク制御装置である。

表1 CSP(内容検索処理機構)，ROP(関係演算処理機構)の機能一覧 RDBデータ操作の選択，射影，結合に必要な機能を備えている。

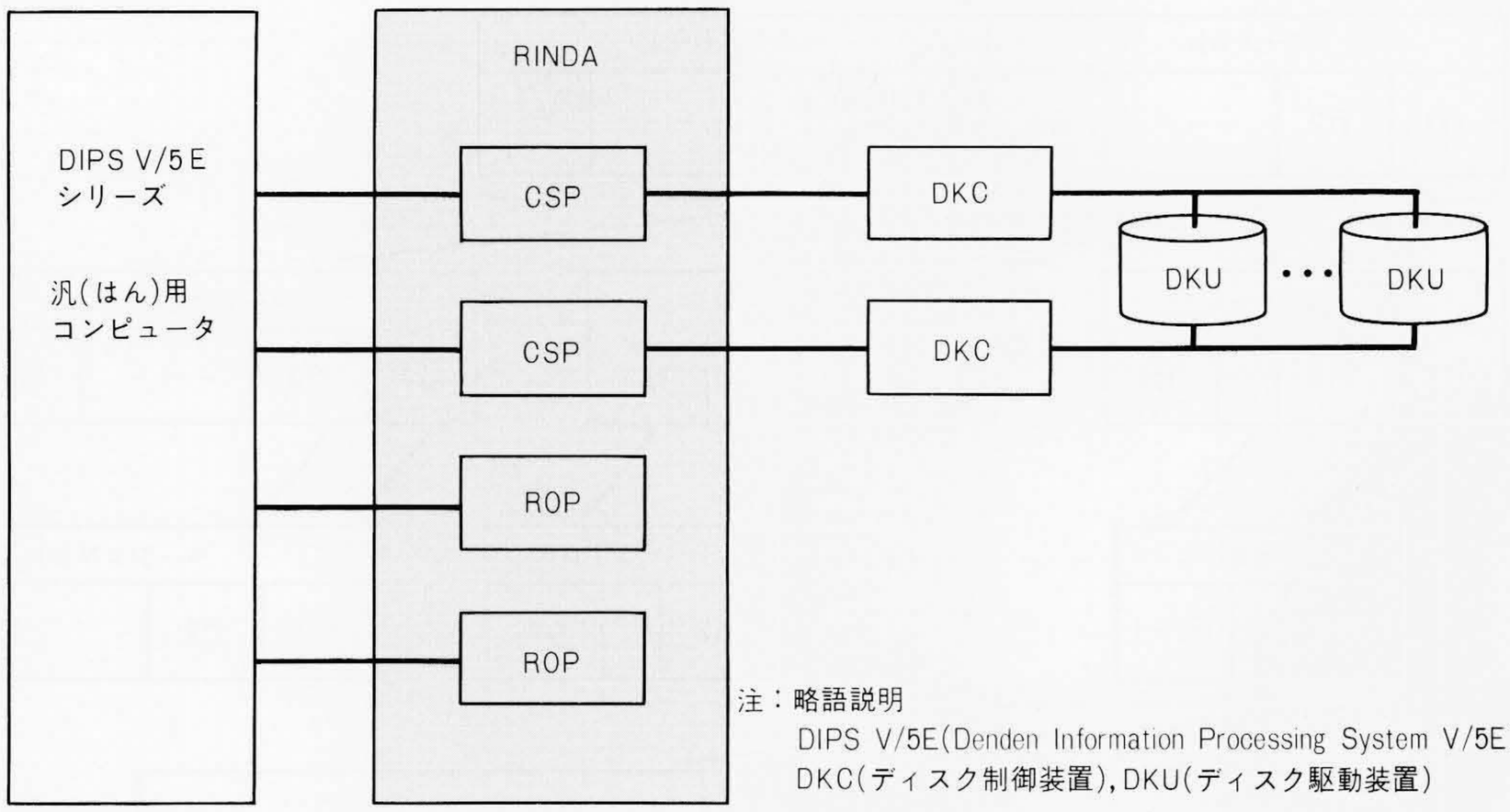
(a) CSP

機能		説明
述語判定	比較述語	(列指定)，(比較演算子)，(定数)の判定
	IN述語	(列指定)，〔NOT〕，IN(定数リスト)の判定
	LIKE述語	(列指定)，〔NOT〕LIKE(パターン)の判定
	NULL述語	(列指定)IS〔NOT〕NULLの判定
検索条件判定		述語のAND/ORによる任意の論理式判定
出力列の抽出		上記検索条件を満足する行からの任意の列の出力
集合関数演算		上記検索条件に合致する行数のカウント (COUNT(＊))

注：用語，記号はSQLに準拠

(b) ROP

機能	説明
ソート	指定された列の値(ソートキーと呼ぶ。)による昇順または降順への行の並べ替え ソートキーは任意の順序の複数列で構成可能
ふるい落とし	ソートキーの値による二つの表からの結合可能性のない行の除去 一方または両方の表からの除去が可能
重複除去	ソートキーの値が重複する2番目以降の行の除去
重複計数	ソートキーの値ごとに重複する行数をカウント



注：略語説明
DIPS V/5E(Denden Information Processing System V/5E)
DKC(ディスク制御装置),DKU(ディスク駆動装置)

図3　RINDAシステム構成例　　V/5Eシリーズ,DKCとはおのおのの DIPS複合インタフェースで接続される。V/5Eシリーズとの間の転送速度3Mバイト/秒，DKCとの間は最大転送速度6Mバイト/秒である。

ズ中央処理装置のブロック マルチプレクサ チャネルに接続される。

システム構成を**図3**に示す。DIPS入出力装置接続インタフェースとして標準的なDIPS複合インタフェースを採用し，各種CPU,DKC(磁気ディスク制御装置)との接続を可能とした。

CSP，ROPはおのおのオプション化されており，ユーザーは業務量，データ量に応じてフレキシブルな構成をとることができる。

3.1　CSPの構成

CSPはアクセスすべき表のディスク上の格納位置，選択・射影処理を実行する行・列に関する条件などの制御情報をチャネルから受け取り，制御情報に従ってディスクから表を読み出し，行ごとに選択・射影処理を実行し，指定された出力形式に従ってRDB管理システムに検索結果を戻す。

CSPの構成を**図4**に示す。

DKC制御部はDKCに対し入出力命令を発行するチャネルとして動作する。つまり，制御情報を基に作成した一連の指令(コマンド)をDKCに発行して，表の格納位置に位置づけ，複数の行から構成されるページ単位に表を読み出す。

選択処理部は行ごとに選択条件が指定された複数の列に対し列値と定数，定数リストとの大小および等価の比較演算を高速で実行し，検索条件との一致・不一致を評価する。同時に行内の射影対象列のアドレスを計算する。

入力列アドレススタッカは選択処理部で計算した射影対象列のアドレスを保持する。

文字列検索部はLIKE述語判定の専用高速処理部であり，1

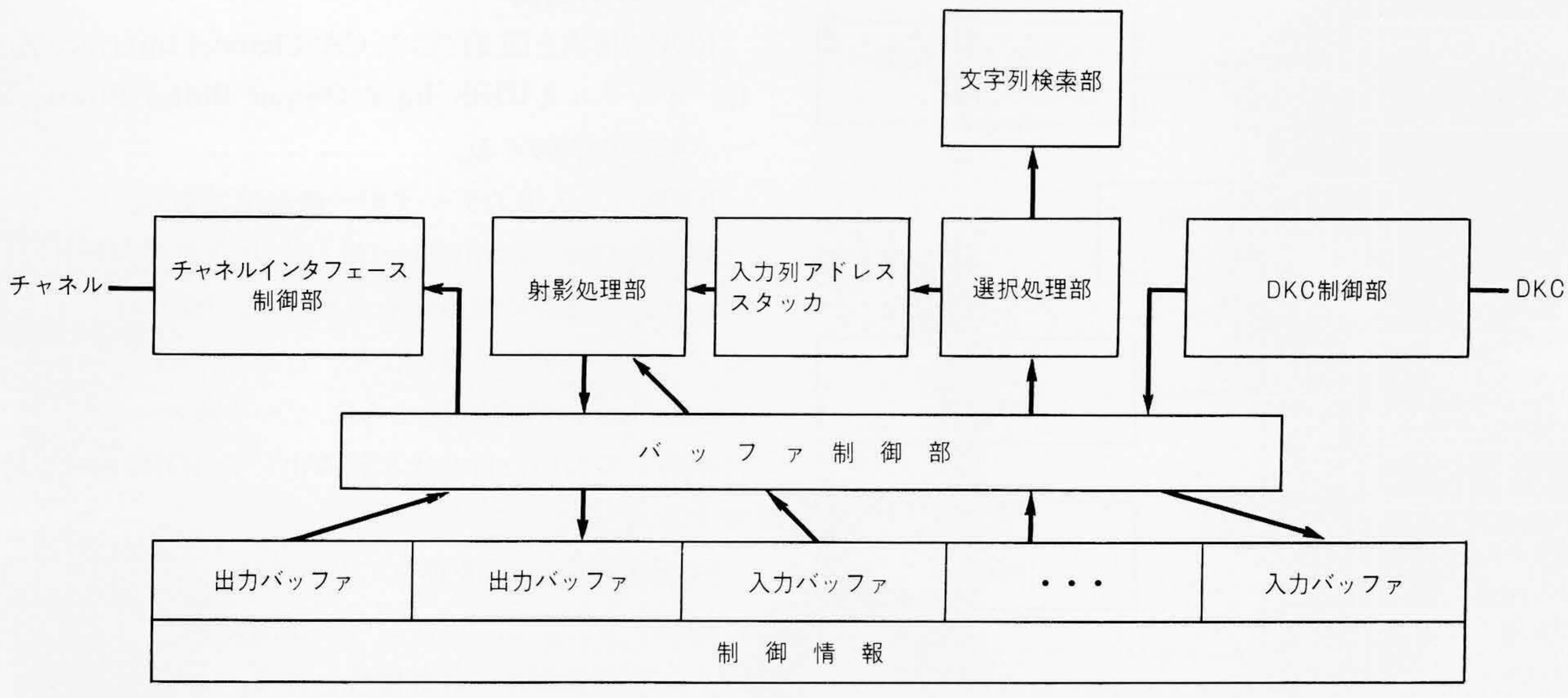


図4　CSPの構成　　複数の入力，出力バッファを時分割でアクセスすることにより，並列動作を実現している。

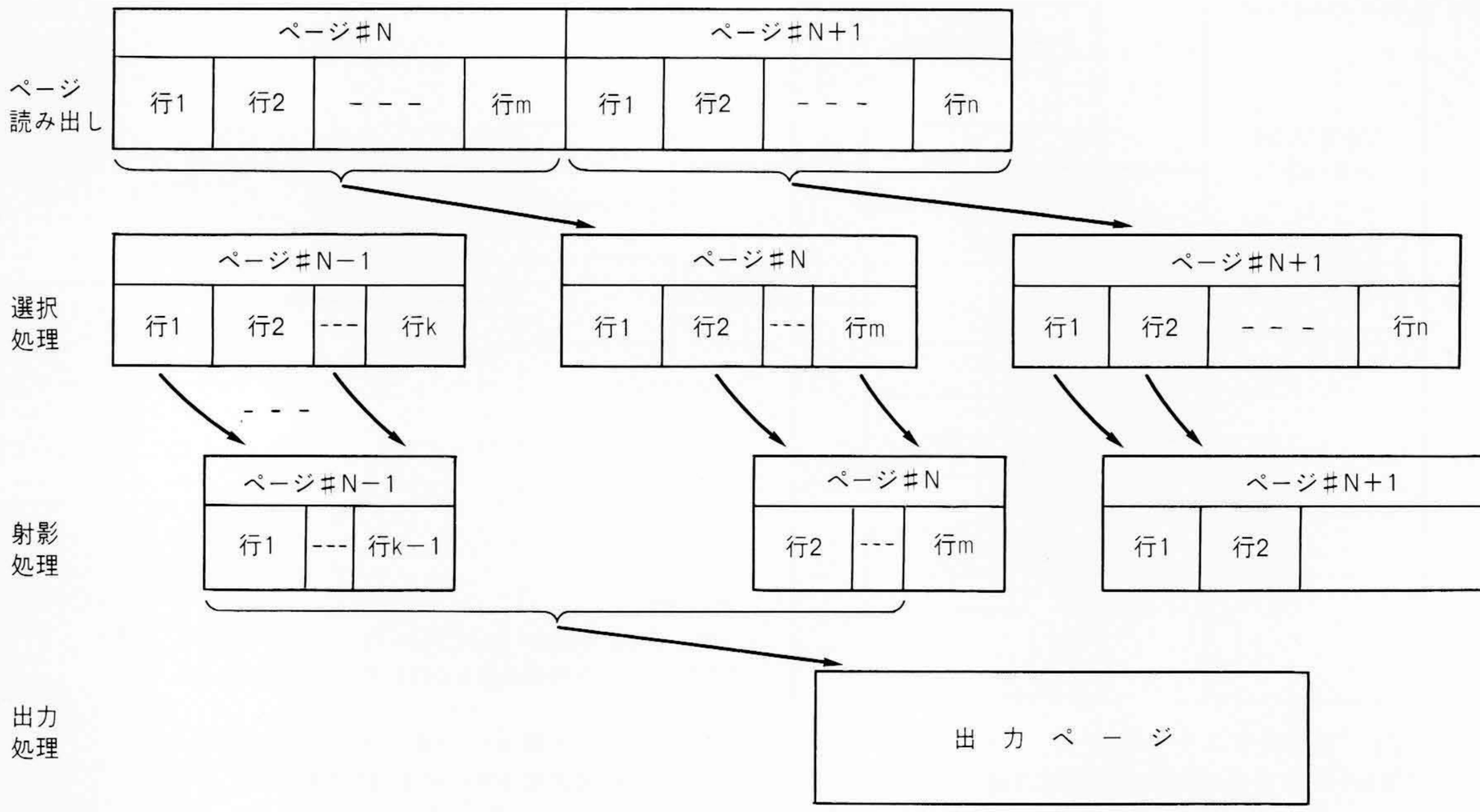


図5 CSPパイプライン制御方式 ページ読み出し，選択処理，射影処理および出力処理を並行実行することで，性能の向上を実現している。

バイト・2バイト文字で構成する列の文字列中の一部分が，指定パターンと一致するかどうかの評価を複数のパターンとの間で高速に実行する。

射影処理部は選択条件に一致した行の射影対象列を，選択処理部で計算した射影対象列のアドレスを基に入力バッファから読み出し，出力バッファに書き込む。このように選択処理中に射影対象列のアドレスを計算しておくことで射影処理の高速化を図っている。

チャネルインタフェース制御部は，チャネルとのインタフェース制御をつかさどり，制御情報の受け取り，出力バッファの送出を制御する。

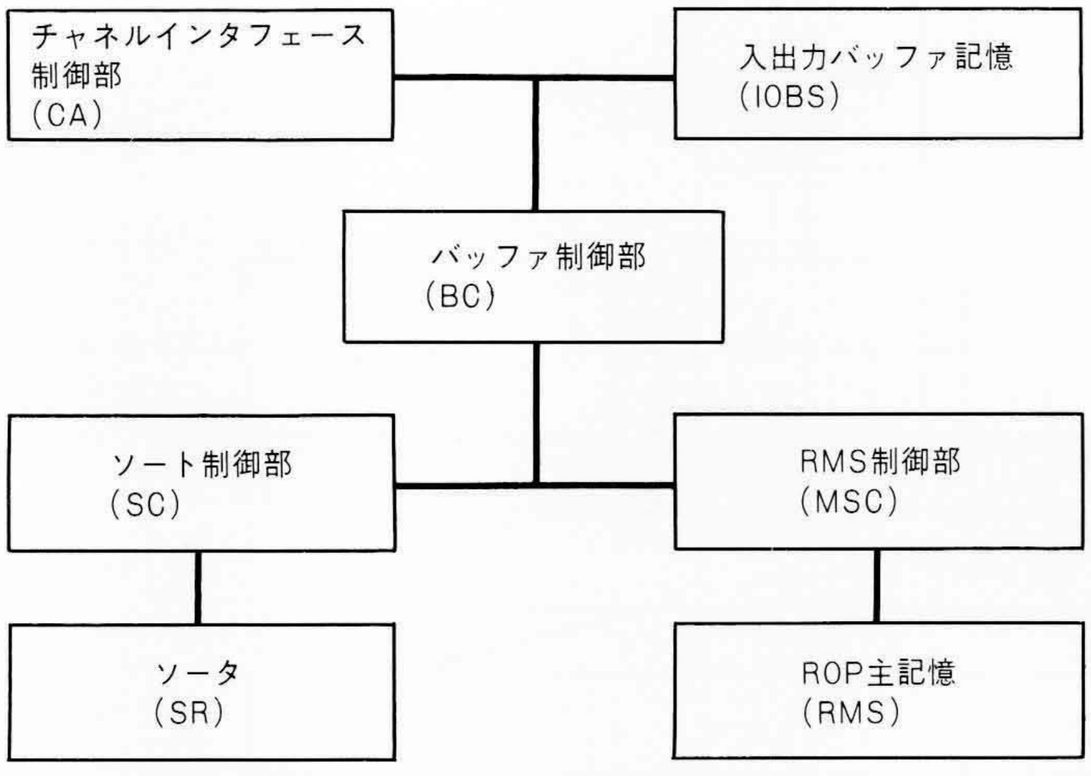


図6 ROPの構成 大量のレコードをソートするために大容量のROP主記憶(RMS)を備えている。レコード数量がソータのハードウェア量に依存しない方式を採用している。

以上の各制御・処理部は文字列検索部を除いて，1～2個のLSIで実現している。

バッファ制御部はバッファとDKC制御部，選択処理部，射影処理部，チャネルインタフェース制御部間の時分割データ転送を制御する。

CSPはディスクからのデータ転送に追随して(以下，オンザフライと言う。)検索処理を行うためパイプライン制御を採用し，性能向上を図っている。

パイプライン制御方式を**図5**に示す。

このため入出力バッファを複数面用意して，ページ読み出し，選択処理，射影処理をパイプラインで実行し，さらに出力処理も含めて並行動作することで高速化を図っている。

3.2 ROPの構成

ROPの構成を**図6**に示す。CA(Channel Interface Adapter)は，チャネルとIOBS(Input Output Buffer Storage)間のデータ転送を制御する。

IOBSは，入出力データの一時記憶である。

BC(Buffer Controller)は，入力された行をRMS(ROP Main Storage)に転送するとともに，行内の複数列で構成するソート対象データ(以下，キーと言う。)を抽出して，SC(Sort Controller)に供給する。さらに，ソート後キー値に従ってRMSに格納しておいた行を逐次読み出して，IOBS上に出力表を形成する。

SCはSR(Sorter)を使ってキーのソートを行い，結果をRMSに格納する。また，SRからキー抽出時に同値キーの計数を行う。これを重複キー計数機能と言う。

MSC(RMS Controller)は，RMSに対するECC(Error

Checking and Correction)機能を実現する。さらに，ふるい落とし機能を実現するために，キーに対するハッシングを実行し，RMS上のハッシュビットアレーの設定・参照を行う。

以上の各制御部は1個のLSIで実現している。

RMSは，前記のハッシュビットアレー，ソート済みキー群，入力行群が格納されるメモリである。

3.2.1 k-Wayマージソータ

ROPには10万行オーダのソートを，入出力データ転送速度に追随して実行できる性能が要求された。この実現のために，k-Wayマージソート法を採用した。

SRは，数十個の未ソートキーを1個ずつ入力し，次に1個ずつ出力させると，その入出力時間内に各キー相互の大小比較を行い，降順(あるいは昇順)にソートする機能を持つ。言い換えれば，SRから出力されるキーは常にそのときSR内にあるキーの最大値(最小値)であるということである。このようにしてできたソート済みキー列を，ソートすべき全体キー列の一部分であるという意味から部分列と言う。

k-Wayマージソートの概念図を**図7**に示す。

SRは1回の動作でm個のキーをソートして部分列を作ることができるものとする。S個(S＝m×k)のキーに対しk回の動作を実行すると，k個の部分列ができる。これを同図では小さな三角形で示している。k≦mならば，この後1回のk-Wayマージソートで S個のキーのソートが完了することになる。その動作手順を次に述べる。

k個の部分列からおのおの最大キーをSRに入力する。次に1個出力すると，それはSR中のk個のキーの中の最大値であるのでこれをRMSに格納する。

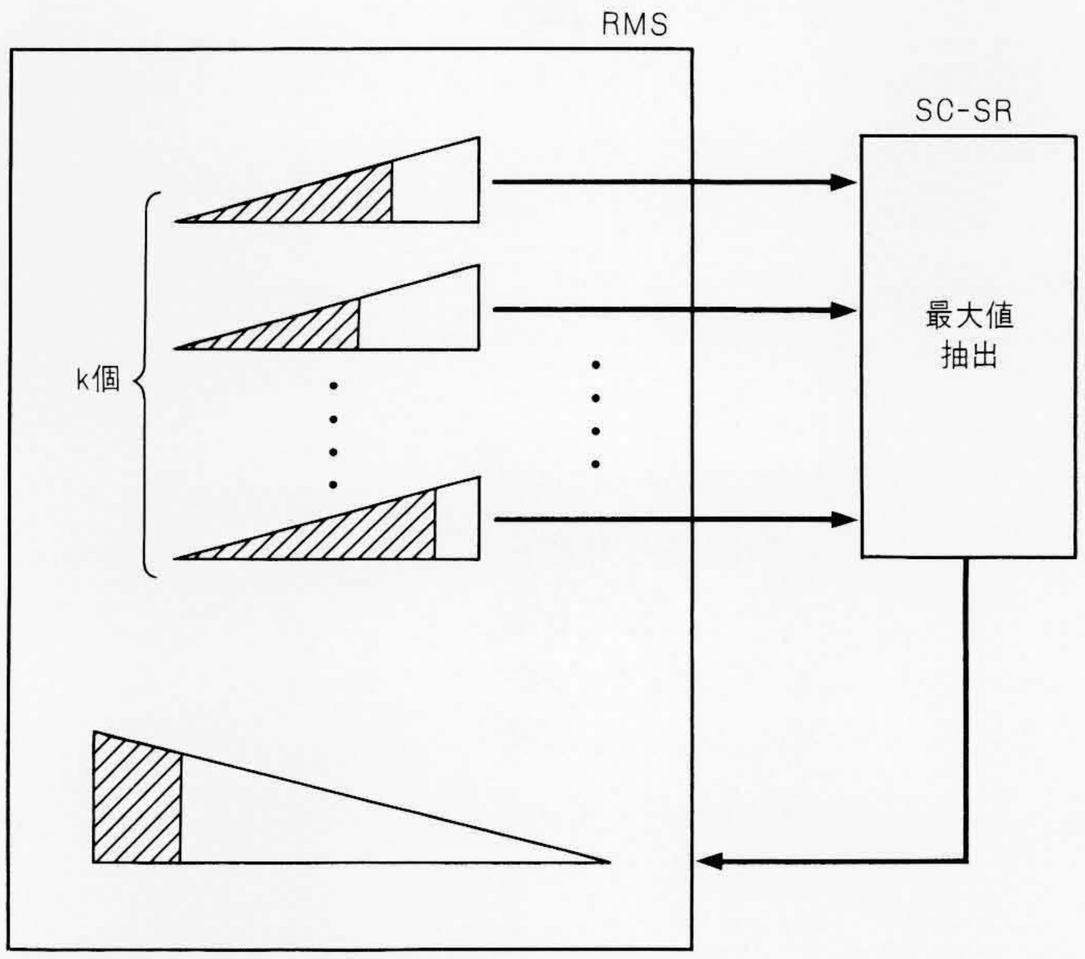


図7 k-Wayマージソート概念図 k個の部分列から一度に1個ずつキーを取り出し，その中の最大値を抽出することを繰り返してマージソートを行う。

いま出力したキーの属していた部分列から次のキーを1個取り出して入力し，1個出力する。これを先に出力したキーの次に並べてRMSに格納する。この1個ずつのキー入出力を，すべての部分列の最後のキーがなくなるまで繰り返すことで，S個のキーのソート列がRMS上に完成することになる。

以上の例では，マージソート段階が1段であったが，これを多段にすることで，より多くのキー数に対応可能である。

この方法によれば，原理的にはソート可能なキー数は，ハードウェアの比較器の数に依存せず，RMSの容量だけに依存することになる。

3.3 実装技術

RINDAは以下の半導体・実装技術を採用し，コンパクトで高性能なプロセッサを実現した。

(1) 論理LSI

1.3 μmデザインルールの6,000～4万ゲートのCMOS(Complementary Metal Oxide Semiconductor) LSIを採用し，消費電力をきわめて小さくした。

(2) RAM(Random Access Memory)

(a) CSPのBS(Buffer Storage)，ROPのIOBSにアクセスタイム25 nsの64 kビットBiCMOS(Bipolar CMOS) RAMを採用し，低消費電力と高集積密度を実現した。

(b) ROPのRMSに1MビットDRAMを採用し，高密度実装を実現した。

(3) 装置実装

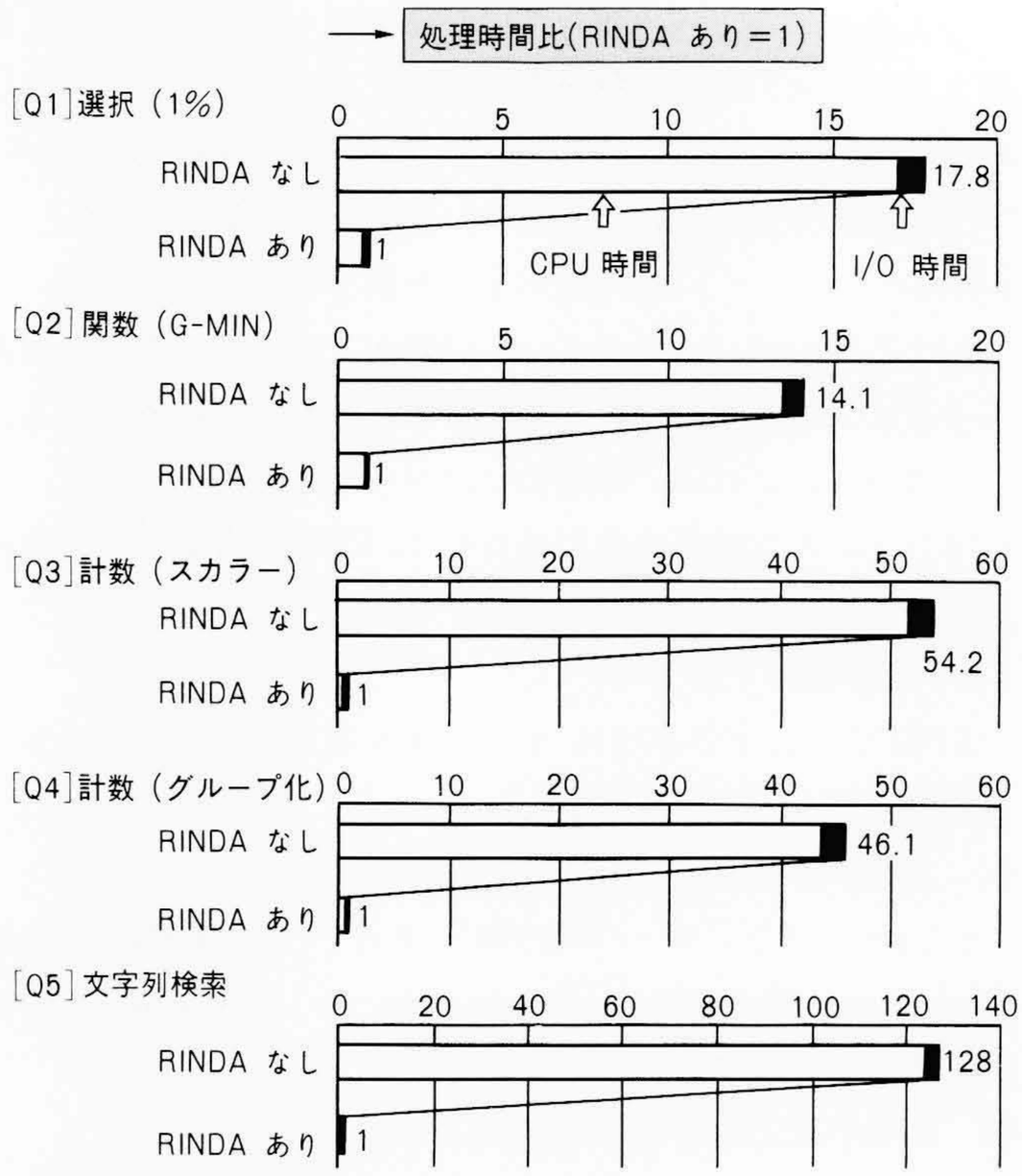
CSP, ROPのすべての論理パッケージを1枚のプラッタに収容し，小形化を図った。さらに，ROPのRMSはROP論理パッケージに挿入する方式を採用し，RMS容量の可変化を実現した。

3.4 性能評価[3)]

評価用問い合わせを**表2**に，評価結果を**図8**に示す。この性能評価結果はNTT情報通信処理研究所の測定結果である。問い合わせはRDB処理の性能評価で用いるWisconsinベンチマ

表2 性能評価用問い合わせ RDB性能評価で用いるWisconsinベンチマークテストの問い合わせである。

問い合わせ		対応するSQL文
Q1	選択(1%)	SELECT * FROM tenk WHERE unique 2 >＝5000 AND unique 2 <5100
Q2	関数(グループ化MIN)	SELECT MIN(unique 2) FROM tenk GROUP BY hundred
Q3	計数(スカラー)	SELECT COUNT(*) FROM tenk WHERE unique 2 >＝5000 AND unique 2 <5100
Q4	計数(グループ化)	SELECT COUNT(*) FROM tenk GROUP BY hundred
Q5	文字列検索	SELECT COUNT(*) FROM tenk WHERE stringu 2 LIKE 'XABCX'



注：入力(10,000行)，行長(208バイト)

図8 性能評価結果 問い合わせにより，10～100倍の性能向上が実現できる。

ークテストの問い合わせである。評価に用いた表は一万行，行長208バイト(制御情報を除く。)である。評価システムは，DIPS-V30CPUにRINDA 1台(CSP×2，ROP×1搭載)，転送速度1.8 Mバイト/秒のディスク駆動装置，ディスク制御装置2台によって構成している。

(1) 問い合わせQ1はCSPのオンザフライ選択・射影処理
(2) 問い合わせQ2はROPの高速ソート処理
(3) 問い合わせQ3はCSPのオンザフライ選択処理
(4) 問い合わせQ4はROPの高速ソート，重複キー計数処理
(5) 問い合わせQ5はCSPの文字列検索専用処理

により，CPU処理時間を大幅に削減している。

問い合わせによって，従来と比較して10～100倍の性能向上が達成できた。

4 結 言

以上述べたように，RDB専用プロセッサRINDAは各種高効率な論理方式を採用し，高性能マシンとして開発することができた。

RDB専用プロセッサはまだ萌(ほう)芽期と言えるが，今後も顧客要求，市場動向にマッチしたRDBマシンの開発を推進していきたい。

最後に，本稿で報告した開発はNTTのRINDAシステム開発計画の一環であり，ご指導いただいたNTT情報通信処理研究所専用システム研究部の関係各位に対し感謝の意を表す次第である。

参考文献

1) 拝原，外：データベースプロセッサの動向，NTT R&D，Vol.38，No.8(1989)
2) 速水，外：データベースプロセッサRINDAのハードウェア機構，NTT R&D，Vol.38，No.8(1989)
3) 福岡，外：DIPSデータベースプロセッサRINDAのアーキテクチャ，NTT R&D，Vol.38，No.8(1989)