関係データベース処理エンジンのソータの試作と評価†

岩田和秀†† 神谷茂雄††† 酒井浩†††
柴山茂樹††† 伊藤英則†† 村上国男††††

ソーティングは非数値処理分野における基本演算の1つであり，ソフトウェアの高速処理アルゴリズムの研究と同時にソータと呼ばれる専用ハードウェアによる高速処理の研究が活発に行われている．ソータの応用分野としては種々のものが考えられるが，その代表的なものの1つに関係データベース処理がある．しかし，これまでに提案されたソータは，アルゴリズムの検証や基本機能の実装が中心であり，ソータを実際に応用する場合の諸条件や処理時間の解析等が十分に検討されていない．そこで，筆者らはソータを関係データベースの処理に応用する場合に必要となる，レコード長，レコード数，キー長等のパラメータ，null 値の扱い，重複レコードの検出等に柔軟な対応機能を持つソータの実現方式を考察した．次に，パイプライン化された2ウェイ・マージソート・アルゴリズムを採用して，上記の諸機能を盛り込んだセルの設計とゲートアレイによる実現を行い，これを用いてソータの試作と性能評価を行った．本ソータはセルを12個使用しており，3 MByte/sec のデータ転送速度に同期して4,096個の同一形式のレコードをソートする．本稿では，関係データベースの処理にソータを使用する場合にソータが具備すべき諸機能の考察，その機能を盛り込んだセルの設計例，ソート処理時間の詳細な解析および解析結果と実測値の比較評価について報告する．

1. まえがき

第5世代コンピュータ・プロジェクトの前期（昭和57年度～昭和59年度）では，知識ベースマシン研究の第1歩として，Prolog のファクトを格納する関係データベースマシン Delta の開発が行われた[1]．筆者らは，このプロジェクトに参加し，Delta の特徴の1つである，関係代数演算等をパイプライン方式で実行する関係データベース処理エンジンの開発を行った[2),3)]．本稿では，このエンジンの基本構成要素の1つであるソータの特徴と試作結果について述べる．

ソーティングは非数値処理分野における基本演算の1つであり，古くから数多くのアルゴリズムが提案されて，目的に応じた使い分けが行われてきた[4)]．

1970年代に入ると，ホスト計算機のソート処理ルーチンを付加プロセッサのファームウェアで実行して，ホスト計算機の負荷を軽減する方式が提案された[5)]．その後，VLSI 技術の進歩と関係データベースマシン研究の興隆により，ソータの提案が活発になった．

これまでに提案されたソータの主要なものには，磁気バブルメモリのループ構造を利用したソータ[6)]，n 個のレコードを n 個のセルを用いてソートするバイトニックソータ[7)]や並列計数ソータ[8)]，$\log n$ 個のセルを用いて n 個のレコードをソートするパイプライン方式のソータ等がある．これらを現状の技術を用いて試作するという観点からみると，$\log n$ 個のセルを用いたパイプライン方式のソータが，ハードウェアが小型化できるので実用的と考えられる．

パイプライン方式のソータとしては，マージソータ[9)]，ヒープソータ[10)]，シストリックソータ[11)]等が発表されている．ヒープソータはレコード列の入力を終了してから結果が出力され始めるまでの出力遅れ時間がなく，かつメモリの使用効率が良いという優れた特徴を持っている．しかし，データの入出力端が同一のため，ソータの容量を単位とした連続的なパイプライン処理ができない．シストリックソータは入出力端を独立にしてヒープソータの問題点を解決し，さらに比較結果を移動歴として記憶することによりマルチウェイ・マージを可能とした大容量ソータである．しかし，マルチウェイ・マージを行う時のデータ転送制御方式等に未検討の課題が残されている．これらに対して，マージソータは若干の出力遅れ時間があることおよびメモリの使用効率が悪い欠点を持つものの，ソータの容量を単位とした連続的なパイプライン処理が可能で制御が簡単なため，汎用計算機の付加プロセッサとするのに適している．

---

† Implementation and Evaluation of the Sorter in a Relational Database Engine by KAZUHIDE IWATA (Institute for New Generation Computer Technology), SHIGEO KAMIYA, HIROSI SAKAI, SHIGEKI SHIBAYAMA (Toshiba Research and Development Center), HIDENORI ITOH (Institute for New Generation Computer Technology) and KUNIO MURAKAMI (NTT Communications and Information Processing Laboratories).

†† 新世代コンピュータ技術開発機構
††† (株)東芝総合研究所
†††† NTT 情報通信処理研究所

マージソータについては，そのアルゴリズム[12]およびシミュレーション[9]による動作確認等がすでに発表されている．しかし，これらの研究はアルゴリズムの検証が中心であり，ソータを実際に応用する場合の諸条件が十分に検討されていない．すなわち，ソータを関係データベースの処理に応用する場合には，レコード長，レコード数，キー長等のパラメータ，null 値の扱い，重複レコードの検出等に関する柔軟な対応が必要であるが，現段階ではこれらの実現方式を盛り込んだソータの設計は行われていない．

そこで，筆者らは関係データベースの処理にソータを使用する場合にソータで具備すべき諸機能を考察し，それらの機能を盛り込んだソータの設計を行った．次に，3 MByte/sec のデータ転送速度に同期して 4,096 個の同一形式のレコードをソートできるソータ（固定長ソータ）の試作を行い，処理時間の解析結果と実測値の比較評価を行ったので，その概要を報告する．

2. ハードウェア化の検討

ソータを用いて関係代数演算を効率よく処理するためには，ソート処理と関係代数演算が連続して行われるよう配慮する必要がある．このため，ソート・アルゴリズムには連続的なパイプライン処理が可能な2ウェイ・マージ法を採用し，入力レコードを語（2バイト）単位で処理する方式を検討した．本章では，アルゴリズムの概要と特徴，ソータを関係代数演算の前処理に使用するために必要となる機能とその実現方式について述べる．

2.1 アルゴリズム

2ウェイ・マージソート・アルゴリズムは，2つのソートされたレコード列（以下，ストリングと呼ぶ）のマージ操作を，繰り返し実行してソート処理を行うものである．本ソータは，このマージ操作を，1次元に 12 個配置したセルにより連続的に行う．すなわち，対象となる入力レコード列(以下，ストリームと呼ぶ)を1段目のセルから入力すると，i 段目 ($i=1, 2, \cdots$) のセルは前段セルからの 2^{i-1} 個のレコードからなる2つのストリングをマージして，2^i 個のレコードからなるストリングを次段のセルに出力する動作を繰り返す．したがって，ストリングの長さはセルで処理されるごとに2倍され，最終段セルからは 2^{12} 個のレコードからなるストリングが得られる．

レコード数 C が5，レコード長 L が2バイトのストリーム（50, 10, 20, 30, 40 の順に入力されるとする）が入力された時の本ソータの動作原理を図1に，パイプライン処理の様子を図2に示す．ソータの各段は，2つのストリングを格納するメモリ（UM と LM）とセルで構成され，奇数番目の入力ストリングを UM に，偶数番目の入力ストリングを LM に格納する．ソータは，入力ストリームを長さ1のストリングが並んだものと見なして，これを入力順に2個ずつマージし，各段でストリング長を倍々にしていく．各段は，偶数番目のストリングの先頭データが入力されるごとに処理を開始し，図2に示すようなパイプライ

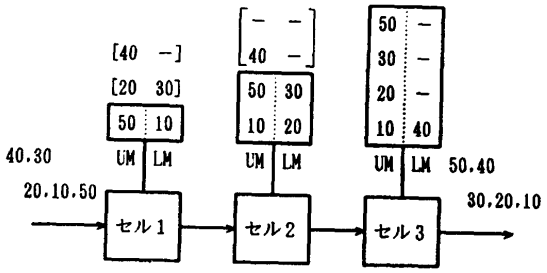

図 1　2ウェイ・マージソータの原理
Fig. 1　Principle of the 2-way merge sorter.

時刻	入力	セル1 UM / LM reg.	セル2 UM / LM reg.	セル3 UM / LM reg.	出力	処理時間
1	50	50 / —				
2	10	50 / — (10)*				
3	20	20 / — (50)	−10 / −−			CLT
4	30	— / 30 (20)*	5010 / −−			
5	40	40 / — (30)	−50 / −20 (10)*			
6		— / — (40)	−50 / −30 (20)*	−−−10 / −−−−		
7			4050 (30)* / −−	−−2010 / −−−−		(2^a−C)LT
8			−40 / −− (50)	−302010 / −−−−		
9			−− / −− (40)	50302010 / −−−−		(a−L/2)2T
10				−503020 / −−−40 (10)*		
11				−−5030 / −−−40 (20)*	10	
12				−−−50 / −−−40 (30)*	20	
13				−−−50 / −−−− (40)*	30	CLT
14				−−−− / −−−− (50)	40	
15					50	

図 2　パイプライン処理の様子
Fig. 2　Overview of pipeline processing.

ン動作を行う．図2において，reg. は，2つのストリングのレコードの比較によって選択されたレコード（＊印）または無条件に選択されたストリングの残りのレコード（無印）を，次段へ出力するために保持するレジスタである．

図2より，ストリームが入力され始めた時点から結果が出力され終るまでのソート処理時間 $T(S)$ を求めると，次のようになる．

$$T(S)=2CLT+(2^a-C)LT+(a-L/2)2T \quad (1)$$

ただし，$a=\langle \log_2 C \rangle$ で $\langle\ \rangle$ は小数点以下を切り上げた整数値，T は1バイト当りの処理時間である．上式において，第1項はストリームの入力時間と出力時間の和，第2項はCの2の累乗からのずれによる遅れ時間，第3項は入力ストリームがソータを通過するために必要な遅れ時間を示す．

本例の場合は $C=5$, $L=2$, $a=3$ であるから，$T(S)=30T$ となる．なお，本ソータの1ステップは2バイト単位の処理を行うので，$T(S)=30T$ をステップ数に換算すると，図2に示されているように，15ステップとなる．

次に，図2より明らかなように，各セルは偶数番目のストリングが入力され始めると，2バイト受け取ると同時に2バイト出力する動作を繰り返すので，セル i で必要になるメモリ容量は1ストリング分の $2^{i-1}\times L$ である．しかし，メモリは2つのストリングを区別してしかもキュー機能を持つ必要があるので，本ソータでは $2^i\times L$ の容量のメモリをセル i に実装してセルの制御回路を簡単にする方式を採用した．

2.2 レコード形式とデータタイプ

関係データベース処理では，レコードをキーとして扱う場合とレコードのあるフィールドをキーとして扱う場合がある．本ソータでは，セルに後述するパス・モード機能を持たせて，ストリーム内のレコードのレコード長とキー長をともに2バイト単位で最大4 kバイトの長さまで指定できるようにした．

フィールドをキーとした演算を効率よくパイプライン処理するには，キーがレコードの先頭にある状態で入力される必要がある．そこで，ソータの入力側に1レコード分のメモリを用意して，レコードのフィールドを回転させ，出力側で元に戻す方式を採用した．

データベースの処理では様々なデータタイプが扱われるが，それをそのままソータで扱うことは困難である．そこで，ソータの入力側でキーを絶対値数に変換し，出力側で逆変換する方式を採用した．本ソータで扱う入力データのタイプは，絶対値数，整数および正規化された IBM フォーマットの浮動小数点数である．

2.3 Null 値と重複キーの取り扱い

データベースでは，フィールドの値が不明または定義されていない時，その値は null 値と呼ばれ，特別な扱いがされる．本ソータでは，入力ストリームに null 値が含まれている場合，まず正常なキー値を持つレコードを出力し，その後に null 値のキーを持つレコードを入力順に出力するようにした．実現方式としては，セルの制御回路を簡単にするため，ソータの入力側で後述する null 信号を発生させる方式を考案した．

等しいキーを持つレコードを検出することは，レコードをキーとして扱う場合は unique 演算等で，フィールドをキーとして扱う場合は join 演算等で重要となる．等しいキーの検出はセルで容易にできるが，タグの操作がセルの制御回路を複雑にする．本ソータでは，最終セルの出力にチェッカと呼ぶ専用回路を付加して，ソート結果のチェックと重複信号の発生を行うようにした．なお，重複したキーを持つレコードは，入力された順に出力してマルチ・キー・ソートに対応できるようにした．

3. システム構成の概要

3.1 システム構成

本ソータは関係データベース処理エンジン（以下，RDBE と略記する）の基本構成要素の1つとして開発されたので，ソータの性能評価は**図3**に示す RDBE のシステム構成で行った．RDBE は入力用アダプタ (HMA-IN), IN モジュール，ソータ，関係代数演算モジュール (RAPM), 出力用アダプタ (HMA-OUT) および制御用マイクロプロセッサ (ECP) で構成され，チャネルを介して階層構造メモリ (HM) に接続されている．

RDBE は ECP の制御により，HM からのデータを HMA-IN, IN モジュール，ソータ，RAPM, HMA-OUT を介してパイプライン処理し，結果を HM に転送する動作を行う．

3.2 各モジュールの機能

（1） HMA-IN と HMA-OUT: RDBE と HM 間のインタフェースの制御を行う．

（2） IN モジュール: フィールドの回転操作，データ・タイプの変換操作および null 信号を発生す

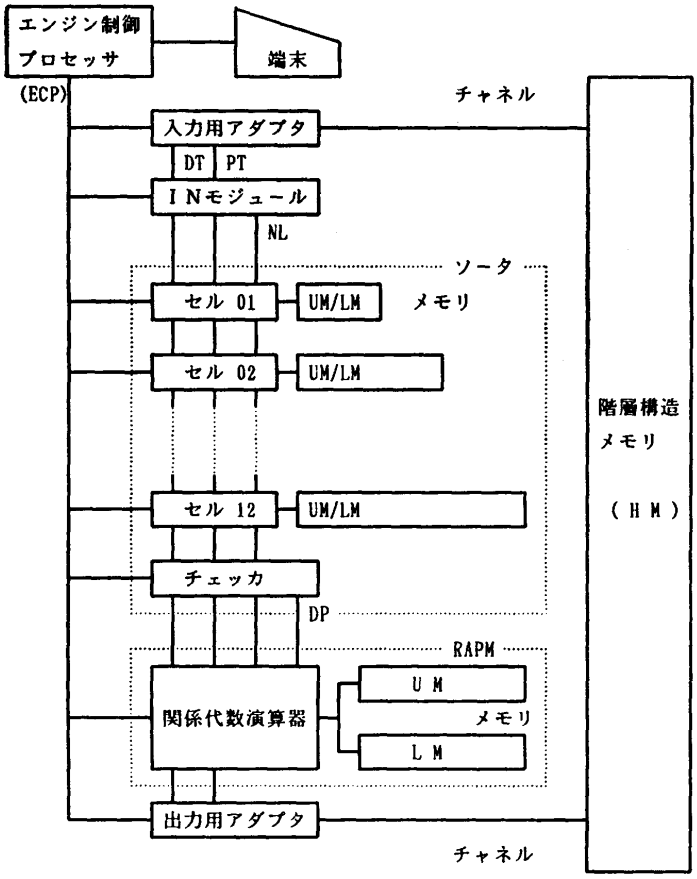

図 3 RDBE のシステム構成
Fig. 3 RDBE system configuration.

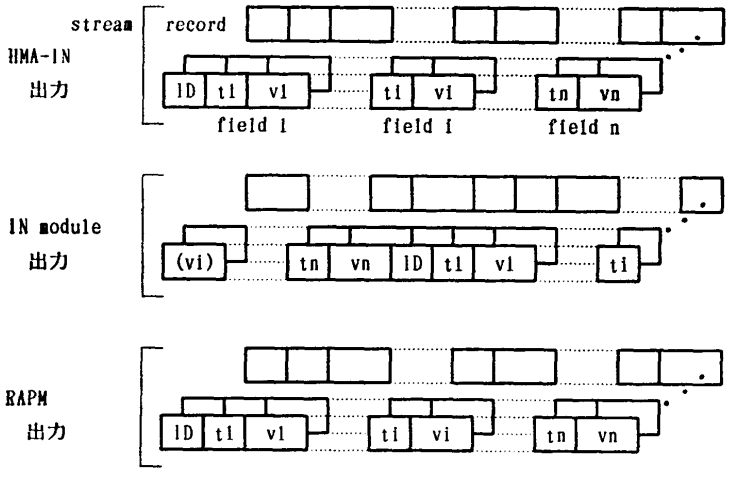

図 4 レコード形式とフィールドの回転操作
Fig. 4 Record format and field ordering.

る操作等を行う．Null 信号はタグを調べて1語ごとに生成するもので，パリティ・ビットと同様に専用線で転送される．

（3） ソータ：12 個のセルと1個のチェッカにより構成され，入力ストリームをソートした後，その結果のチェックとキーの重複を検出して重複信号を生成する．

（4） RAPM：2つの 64k バイトメモリ (UM, LM) と関係代数演算器で構成され，関係代数演算，マージ演算および IN モジュールで行った各種変換の逆変換等を行う．

（5） ECP：上記のモジュールの制御，RAPM では処理できない算術演算の実行およびソータの性能測定用データの収集等を行う．

（6） HM：汎用コンピュータで実現された大容量メモリで，RDBE とは3 MByte/sec のデータ転送速度を持つ2つのチャネルで結合されている．

図3において，DT, PT, NL, DP はそれぞれデータ線 (16 ビット)，パリティ・ビット線 (2ビット)，null ビット線 (1ビット)，重複ビット線 (1ビット) を示す．Null ビット線と重複ビット線は，データ線上のデータの属するレコードのキー値が，null 値であることおよび前のレコードのキー値と等しいことをそれぞれ示す．

図4に，RDBE に入力されるストリームのレコード形式と IN モジュールおよび RAPM におけるキー・フィールド i の回転操作の様子を示す．同図において，ID はレコード識別子，ti と vi はそれぞれ i 番目のフィールドのタグおよび値，(vi) は vi の絶対値数表現を示す．

RDBE のソフトウェアは，ハードウェアを制御する機能とハードウェアでは実行できない算術演算等を行う機能を持つが，詳細は文献[1]で述べた．

4. ソータのハードウェア構成

4.1 ソータの仕様

試作したソータの概略仕様は次のとおりである．

ソート・レコード数　　：4,096 個

レコード長L : $2 \leqq L \leqq 4,096$ バイト
キー長K : $2 \leqq K \leqq 4,096$ バイト
セル数N : $N=12$
メモリ素子 : 8×8kビット SRAM
セルiのメモリ容量: 2^{i+4} バイト
セル12のメモリ容量M
: $M=64$kバイト
処理速度T : $T=330$ nsec/バイト

本ソータはレコード長が16バイト以下の時，4,096個のレコードをソートする能力を持つが，セルのメモリ容量が固定されているので，この値はレコード長により変化する．そこで，セルには2つの入力ストリングをマージするソート・モードのほかに，入力されたデータをメモリを介さずにそのまま出力するパス・モードを用意し，レコード長が大きい時はそれを格納できるメモリ容量を持つセルまでパス・モードを用いることにした．したがって，本ソータの実効ソート容量Eは，次式のようになる．

$$E=L\times 2^{e} \qquad (2)$$

ただし，eは実効動作セル数で$e=\min([\log_2 M/L], N)$，Nは実装セル数，Mは最終段セルの実装メモリ容量，[ ]は小数点以下を切り捨てた整数値を示す．

4.2 セルの制御方式

セルの2つの動作モードは，入力ストリームのレコード数Cとレコード長Lにより，次式に従って使い分ける．

$$J=\min([\log_2 M/L], N, [\log_2 2(C-1)]) \qquad (3)$$

すなわち，1から$(N-J)$番目のセルがパス・モードで，$(N-J+1)$から12番目のセルがソート・モードで動作するよう制御する．例えば，$L=100$バイトで$C=200$の時，$J=8$となるので，1から4番目のセルがパス・モードで，5から12番目のセルがソート・モードで動作する．なお，入力データを直接RAPMに転送する時は，全セルをパス・モードで動作させる．

ソート・モードの動作は，1語単位の繰り返し処理，レコード・レベルの繰り返し処理，ストリング／ストリーム・レベルの繰り返し処理を行う3階層構造で制御される．一方，パス・モードの動作は，1語単位の繰り返し処理を行う階層だけで制御される[3]．

4.3 セルの構造

試作したセルの構造を**図5**に示す．セルは演算部，アドレス生成部，制御部およびインタフェース部より構成され，メモリに接続される．

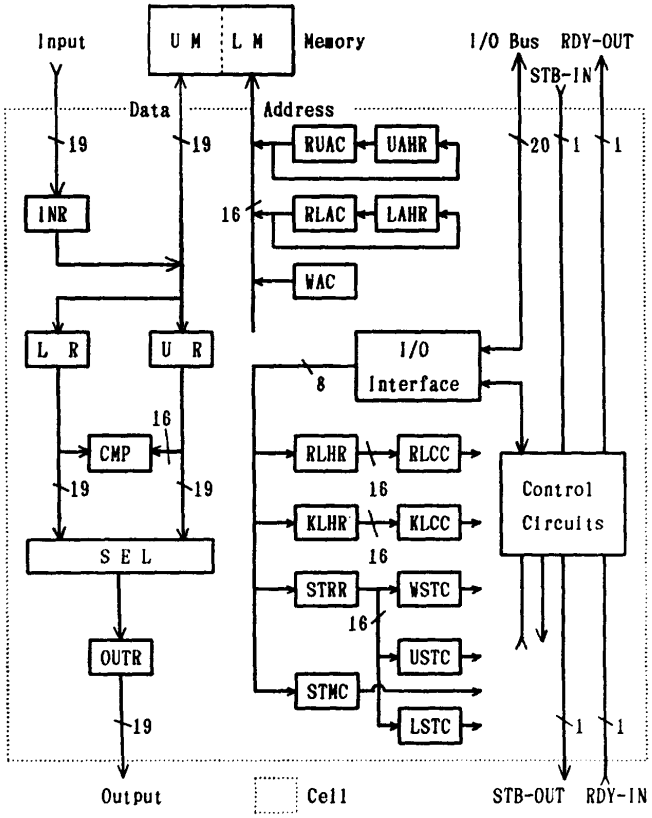

図5 セルの構造
Fig. 5 Structure of the cell.

演算部は入力データ用レジスタ (INR)，データ保持用の2つのレジスタ (UR, LR)，比較器 (CMP)，出力データを選択するセレクタ (SEL) および出力データ保持用レジスタ (OUTR) より構成され，2つのデータを比較して条件に合致したデータを出力する動作を行う．

アドレス生成部は書込みアドレスカウンタ (WAC)，2つの読み出しアドレスカウンタ (RUAC, RLAC)，2つのアドレス保持用レジスタ (UAHR, LAHR) より構成され，入力データの書込みアドレスと比較データの読み出しアドレスを生成する．UAHR (LAHR) は比較中のレコードの先頭アドレスを保持するもので，選択されなかったレコードを次のサイクルで読み出す時に用いる．

制御部は入力ストリームのレコードを計数するカウンタ (STMC)，レコード長を保持するレジスタ (RLHR)，レコード長制御カウンタ (RLCC)，キー長保持レジスタ (KLHR)，キー長制御カウンタ (KLCC)，

ストリングのレコード数保持レジスタ (STRR), 入力ストリングのレコードを計数するカウンタ (WSTC), メモリに格納されたストリングの未処理レコードを計数するカウンタ (USTC, LSTC), 状態遷移を制御する15個のフリップ・フロップおよび外部インタフェースより構成され, セル全体の制御を行う.

メモリは入力ストリングを区別するため UM と LM に論理的に分割されており, 入力ストリングを UM, LM の順に交互に格納する.

4.4 セルの外部インタフェース

セルは, セル間のデータ転送と ECP の入出力バスに関するインタフェース機能を持つ.

セル間のデータ転送は, 転送準備完了信号 STB-IN (OUT) と転送要求信号 RDY-IN (OUT) により前後のセルの準備が完了した時に起動され, 一方でも準備が完了していなければ待ち状態に入るよう制御される. このため, ソータの入力側または出力側でデータ転送が一時的に中断されてもパイプライン動作は乱れない.

ECP からセルにセットされる制御情報には, レコード長, キー長, ストリームとストリングのレコード数, セルの動作モード指定がある.

4.5 セルの処理タイミング

本セルは, サイクルタイム 220 nsec で動作し, 3サイクルで1語の処理を行う. したがって, 処理速度 T は 330 nsec/バイトである. ソート・モードの処理タイミング・チャートを**図6**に示す. 第1サイクルは, セル間の制御信号 STB-IN (OUT) と RDY-IN (OUT) の4つが共に ON の時に開始され, 入力データの INR へのセット, UM のデータを UR に読み出す動作, および OUTR データの出力動作を行う. 第2サイクルは LM または INR のデータを LR に読み出す動作を, 第3サイクルは UR と LR の比較および INR データのメモリへの書き込む動作を行う.

なお, パス・モードでは, 第1サイクルで入力データの INR へのセットと OUTR データの出力動作が, 第2サイクルで INR データの LR への転送が, 第3サイクルで LR データの OUTR への転送が行われる.

4.6 セルのゲートアレイ化

セルの回路規模は2入力 NAND 回路換算で約5,000 ゲートであり, 汎用の IC 回路で実装するとソータの1段分で 30 cm×30 cm の基板1枚を必要とする. そこで, ソータの小型化と高信頼化を計るため, セルは CMOS ゲートアレイで実現し, 上記の基板に4段分を実装した[1]. ゲートアレイの大きさは, セル内のレジスタやカウンタのテストを容易化する回路の追加, ゲート当りの遅延時間を 2 nsec 以内にするための配線長の短縮が行えるよう 8,000 ゲートのものを使用した. パッケージは, メモリとの信号線が40本, 前段セルからの入力信号線と後段セルへの出力信号線が各々21本, セルの制御用信号線が20本必要となることから, 179 ピンのものを使用した. 開発期間は, 論理設計が5か月, 論理シミュレーションと遅延時間解析が2か月, 製造が3か月, テストが1か月で, 合計11か月である.

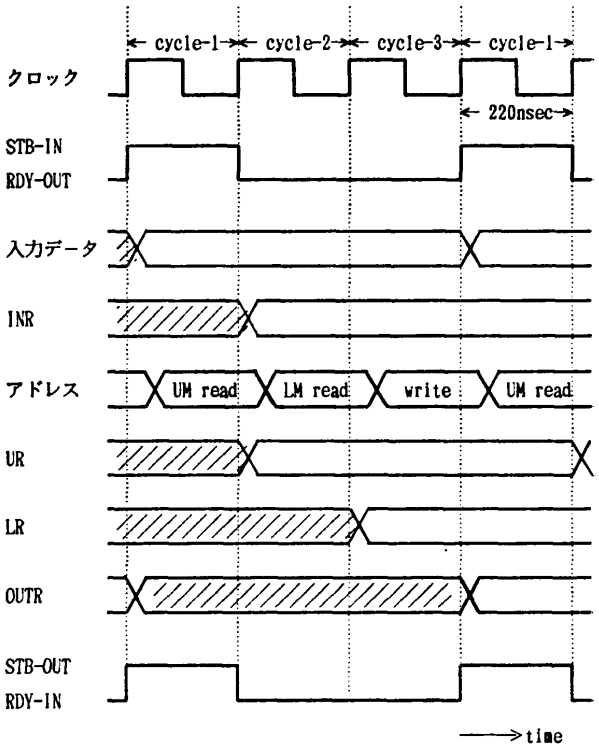

図 6 タイミング・チャート
Fig. 6 Timing chart.

5. 性能評価

5.1 処理時間の解析

入力ストリームは, 実効ソート容量 E 以下のサブストリームに分割して処理される. この時, サブストリームの個数により処理方式が異なる. サブストリームの個数 F を $F=\langle CL/E\rangle$ と表すと, $F=1$ の時はソータを, $F=2$ の時はソータと RAPM を, $F>2$ の時は $F=2$ の処理と RAPM を用いた処理方式と

なる．

4.1 節で述べたように，本ソータはストリームの大きさによりセルの動作個数が変化するので，$F=1$ の場合にその詳細な解析を行い，それ以外はストリームの大きさが 64 k バイトの整数倍で $E=$ 64 k バイトとして解析する．

（1） $F=1$ の場合の処理時間 T(1)：ストリームは HMA-IN, IN モジュール，ソータ，RAPM, HMA-OUT を介して HM に送られる．したがって，$T(1)$ は次式で表される．

$$T(1)=t0+t(S)+(N-J)2T+3LT+t2 \quad (4)$$

ただし，$t0$ はコマンドの解釈とエンジンを起動するための時間，$t(S)$ は(1)式の $T(S)$ で $a=J$ としたソータでの処理時間，$(N-J)2T$ はパス・モードで動作するセルを通過する時間，$3LT$ は IN モジュールとチェッカと RAPM における 1 レコード分のバッファリング遅れ時間の和，$t2$ は HM への終了報告時間を示す．

（2） $F=2$ の場合の処理時間 $T(2)$：ストリーム S は 2 つのサブストリーム（$S1$, $S2$）に分割されてソートされ，RAPM でマージされる．この場合の各モジュールにおける処理時間を**図 7** に示す．同図において，$t0$, $t2$ および LT は $F=1$ の場合と同様であり，$t1$ は $S1$ から $S2$ への切替えとエンジンにパラメータをセットする時間を示す．したがって，$T(2)$ は次のようになる．

$$T(2)=t0+t2+(t1+2t(s)+ET+5LT) \quad (5)$$

（3） $F>2$ の場合の処理時間 $T(2^i)$：$F=4$ の例を検討して $F=2^i$ の場合に拡張する．$F=4$ の時，まずストリーム S を 2 つのサブストリーム（$S1$, $S2$）に分割し，これらを $F=2$ の方式で処理して，ソートされたサブストリーム（$S10$, $S20$）を求める．次に $S10$, $S20$ を E の単位に分割，すなわち（$S11$, $S12$）と（$S21$, $S22$）にして，これら 4 つのサブストリームのマージ操作を RAPM で行う．したがって，$T(4)$ は次のようになる．

$$T(4)=t0+t2+2A+2(2t1+3LT+2NT+2T+3ET) \quad (6)$$

ただし，$A=(t1+2t(s)+ET+5LT)$ である．

上式の右辺において，$2A$ は $F=2$ の処理を 2 回行う時間を，括弧内は $S11$ と $S21$ をマージする場合の最悪処理時間を示す．すなわち，$S11$ をソータをパス（遅れ時間 $(N+1)T$）して RAPM の UM に入力（入力時間 ET）し，次に $S21$ を同じくソータをパスして RAPM の LM に入力しつつ，UM と LM のレコードのマージ（処理時間 $2ET$）を行う時間を示す．次に，$F=2^i$ の場合の処理は，まずストリームを $2E$ 単位でソーティングするため $F=2$ の処理を 2^{i-1} 回繰り返し，その後上記のマージ操作を $(2^i E/2E)\log(2^i E/2E)$ 回繰り返すことにより実行する．したがって，$T(2^i)$ は次のようになる．

$$T(2^i)=t0+t2+A2^{i-1}+B(i-1)2^{i-1} \quad (7)$$

ただし，$B=(2t1+3LT+2NT+2T+3ET)$ である．

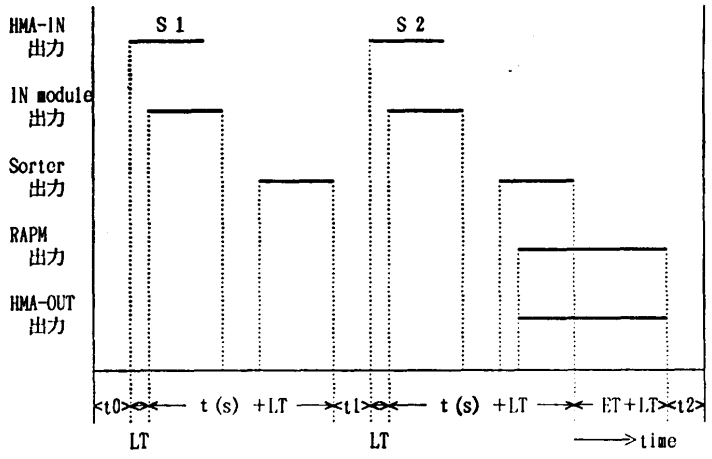

図 7 各モジュールにおける処理時間
Fig. 7 Processing time in each module.

5.2 測定結果

測定は図 3 のハードウェア構成で，端末より RDBE コマンドを入力し，ECP 内の評価プログラムにより，コマンドの解釈からレスポンスの作成までの時間を計測した．なお，評価の対象としたストリームは，ECP で乱数を発生させ，それをすべて HM のバッファ上に格納して行った．$F=1$ の場合の本ソータの処理時間の計算値と測定値（■）を**図 8** に，それ以外の場合を**表 1** に示す．

5.3 考察

ソート処理時間の計算値には，$t0$〜$t2$ の値として，それぞれ実測平均値 10 ms, 3 ms, 5 ms を用いたが，測定値は計算値よりデータ量が増加するにつれて 12% から 35% ほど大きくなっている．この原因は，図 3 の評価システムが RDBE から HM にデータ要求を出して処理する方式のため，上記の $t0$〜$t2$ 以外に HM が RE からの内部コマンドを解釈してバッ

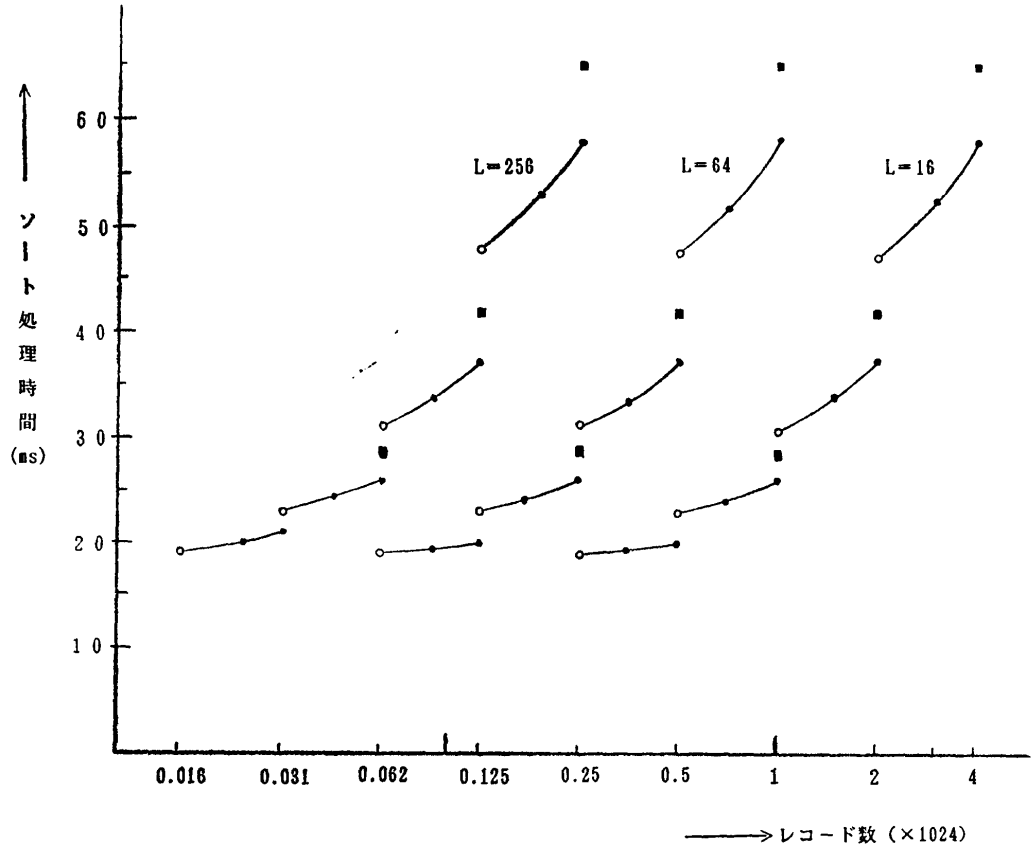

図 8　ソート処理時間 (F=1)
Fig. 8　Sorting time (F=1).

表 1　ソート処理時間 ($F \geqq 2$)
Table 1　Sorting time ($F \geqq 2$).

ストリームのレコード数	ソート処理時間 (msec)	
	計算値	実測値
1	15	16
4,096	57	65
8,192	123	141
16,384	369	460
32,768	999	1,340
65,536	2,535	3,330
132,072	6,159	8,160
262,144	14,511	18,390
524,288	33,423	45,860

(レコード長：16 バイト)

ファ管理等を行うオーバヘッドが存在するためである．したがって，このオーバヘッドを減らすためには，RDBE を HM の入出力機器として HM から制御するか，または RDBE に HM の機能を含めてしまうことが有効であるので，現在後者の検討を進めている．

図8より，本ソータはレコード長が変化した場合，1回にソートできるレコード数は減少するが，ソート容量（レコード数×レコード長）からみた処理時間は一定であり，セルをパス・モードで動作させることによる遅れは無視できることがわかる．しかし，レコード数の2の累乗値に変曲点があることから明らかなように，レコード数が2の累乗値を少し越えた場合の遅れ時間が大きくなる欠点がある．しかし，平均値でみれば 10% 程度の損失であるので，アルゴリズムの改良による効果とセルの複雑度の増加とのトレード・オフを検討中である．

6. むすび

関係データベースの処理に必要となるレコード数，レコード長，キー長等のパラメータの取り扱い，null 値の取り扱い，重複レコードの検出等の諸機能を考慮したソータの設計と試作機の評価を行った．その結果，unique 演算や join 演算等の重負荷となる演算が本ソータでの前処理と RAPM の演算により，$O(n)$

で処理できることが確認できた．

本ソータは，単体で64kバイト，RAPMを連動させた時は128kバイトのストリームを3MByte/secの処理速度でソートする．この処理速度は，現在のチャネルの最高転送速度に相当する．したがって，本ソータは汎用計算機の付加プロセッサとして，磁気ディスクから主記憶へのデータ転送時に，ソート処理や関係代数演算を行う等の応用に適している．後者への応用については別途報告する予定である．

なお，本ソータは関係代数演算の前処理に用いることを主目的に関発したので，

（1） 大量レコードのソートでは性能が低下する．

（2） セルのメモリ使用効率が悪い．

等の欠点を持つ．（1）については，RAPMにマルチウェイ・マージ機能を付加する，汎用計算機の付加プロセッサとして用いる場合は汎用計算機でマルチウェイ・マージを行う等の方法が考えられる．しかし，マルチウェイ・マージを効率よく行うためには磁気ディスクを含めた検討が必要であり，今後の課題である．（2）については，原理的にはセルiは$2^{i-1}\times L$のメモリ容量を持てばよいので，ポインタを用いて削減することも可能である[9]．しかし，セルの制御回路が複雑になることおよびセルのメモリ容量が2倍ずつ増えていく特殊構成のため，本ソータの規模では$2^i\times L$のメモリ容量を実装する方がハードウェアを小型化できる．現在，本ソータの開発経験を基に，可変長キーを処理できるソータの設計を進めている．

謝辞 本研究は新世代コンピュータ技術開発機構からの委託研究の一環として実施したものであり，研究に参加された同機構第3研究室，(株)日立製作所，(株)東芝の関係者に深く感謝いたします．

参考文献

1) 角田，宮崎，柴山，横田，伊藤，村上：関係代数演算専用エンジンを備えた関係データベース・マシン Delta，日経エレクトロニクス，No. 378, pp. 235-280 (1985).
2) Sakai, H., Iwata, K., Kamiya, S., Abe, M., Tanaka, A., Shibayama, S. and Murakami, K.: Design and Implementation of the Relational Database Engine, *FGCS '84*, pp. 419-426 (1984).
3) Kamiya, S., Iwata, K., Sakai, H., Matsuda, S., Shibayama, S. and Murakami, K.: A Hardware Pipeline Algorithm for Relational Database Operation and Its Implementation Using Dedicated Hardware, *IEEE 12th ISCA*, pp. 250-257 (1985).
4) Knuth, D. E.: *The Art of Computer Programming*, Vol. 3, *Sorting and Searching*, pp. 11-388, Addison Wesley, Reading (1973).
5) Barsamian, H.: Firmware Sort Processor System, U. S. Patent, 3.713.107 (1973).
6) Chung, K. M., Luccio, F. and Wong, C. K.: On the Complexity of Sorting in Magnetic Bubble Memory Systems, *IEEE Trans. Comput.*, Vol. C-29, No. 7, pp. 553-563 (1980).
7) Nassini, D. and Sahni, S.: Bitonic Sort on a Mesh Connected Parallel Computer, *IEEE Trans. Comput.*, Vol. C-27, No. 1, pp. 2-7 (1979).
8) 安浦，高木：並列計数法による高速ソーティング回路，信学論 (D), Vol. J 65-D, No. 2, pp. 179-186 (1982).
9) 喜連川，伏見，桑原，田中，元岡：パイプラインマージソータの構成，信学論 (D), Vol. J 66-D, No. 3, pp. 332-339 (1983).
10) Tanaka, Y., Nozaka, Y. and Masuyama, A.: Pipeline Searching and Sorting Modules as Components of a Data Flow Database Computer, *IFIP 80*, pp. 427-432 (1980).
11) 土肥：大容量ファイルを整列するシストリック・ソータ，信学論 (D), Vol. J 67-D, No. 3, pp. 281-288 (1984).
12) Todd, S.: Algorithm and Hardware for a Merge Sort Using Multiple Processors, *IBM J. Res. Develop.*, Vol. 22, No. 5, pp. 509-517 (1978).

(昭和61年9月19日受付)
(昭和62年4月15日採録)

岩田 和秀（正会員）

昭和18年生．昭和43年名古屋大学工学部電気工学科卒業．昭和48年同大学院博士課程修了．同年東京芝浦電気(株)（現(株)東芝）入社，総合研究所に勤務．電力変換装置，産業用ロボット，専用プロセッサなどの研究・開発に従事．昭和62年(財)新世代コンピュータ技術開発機構に出向．現在，同機構研究計画部にて海外交流を担当．電気学会会員．

神谷 茂雄（正会員）

昭和24年生．昭和48年早稲田大学理工学部電子通信学科卒業．昭和50年同大学院修士課程修了．同年東京芝浦電気(株)（現(株)東芝）入社．総合研究所情報システム研究所において，マルチプロセッサ，データベースマシンの研究開発を行う．現在，同社半導体技術研究所にてマイクロプロセッサの研究・開発に従事．電子情報通信学会会員．

酒井 浩（正会員）

昭和30年生．昭和53年東京大学工学部計数工学科卒業．昭和55年同大学院修士課程修了．同年東京芝浦電気(株)（現(株)東芝）入社．総合研究所情報システム研究所において，知識ベースマシンの研究・開発に従事．人工知能学会会員．

柴山 茂樹（正会員）

昭和28年生．昭和50年東京大学工学部電子工学科卒業．同年東京芝浦電気(株)（現(株)東芝）入社．総合研究所情報システム研究所において，アレイプロセッサ，CT用画像再構成プロセッサ，データベースマシン，知識ベースマシンの研究・開発に従事．この間，昭和57年～60年(財)新世代コンピュータ技術開発機構に出向．電子情報通信学会会員．

伊藤 英則（正会員）

昭和21年生．昭和44年3月福井大学工学部卒業．昭和49年3月名古屋大学大学院工学系研究科博士課程電気及電子専攻修了，工学博士．昭和49年4月，NTT通信研究所入社．現在，(財)新世代コンピュータ技術開発機構に勤務，研究室長．これまでに，オートマトンと数理言語理論の研究と大型計算機の研究開発に従事．現在，知識情報処理システムの研究開発に従事．電子情報通信学会，人工知能学会各会員．

村上 国男（正会員）

昭和15年生．昭和39年茨城大学工学部電気工学科卒業．同年日本電信電話公社入社．現在，NTT情報通信処理研究所知能処理研究部総括担当主席研究員．工学博士．この間，昭和57年～60年(財)新世代コンピュータ技術開発機構第一研究室長として出向．コンパイラ，ファイル管理方式，DIPS用OS，機能分散型コンピュータ・アーキテクチャ，知識ベースマシンなどの研究・開発に従事．ACM，電子情報通信学会，日本ソフトウェア科学会，人工知能学会各会員．