自律ディスククラスタの階層化構成におけるリクエスト転送先推測による性能改善

A Performance Improvement Method of Presuming a Request-Forwarding Target in a Hierarchical Storage Cluster

花井 知広♡ 渡邊 明嗣♢ 小林 大♢
山口 宗慶♢ 田口 亮♠ 林 直人♠
上原 年博♠ 横田 治夫♣

Tomohiro HANAI Akitsugu WATANABE
Dai KOBAYASHI
Munenori YAMAGUCHI Ryo TAGUCHI
Naoto HAYASHI Toshihiro UEHARA
Haruo YOKOTA

我々はストレージシステムにおいて負荷分散，故障対策，障害回復などの高度な機能を実現するためのアプローチとして，自律ディスクを提案している．しかし，近年のコストや性能に対する様々な要求に応えるためには，多種のデバイスを組み合わせることが有効である．我々はこれらの要求に応えるために，半導体メモリと磁気ディスクの各デバイスから構成されるクラスタを組み合わせた，自律ディスククラスタの階層的構成手法を提案している．

本稿では自律ディスククラスタの階層的構成手法において，ディレクトリ探索と通信回数の削減による高速化手法を提案する．また試作システムにより性能評価を行い，我々の提案する構成手法の有効性を示す．

We have proposed autonomous disks to import advanced functionality into storage systems, which balances loads, tolerates faults, and recovers data within a cluster. However, since the higher performance is strongly required for the storage systems, a combination of various devices is significant. To solve these problems, we have proposed hierarchical architecture of autonomous storage constructed from a magnetic autonomous disk cluster and a solid state autonomous disk cluster.

In this paper, we propose a method of presuming a request-forwarding target in the hierarchical configuration to eliminate message passing and directory traversing costs. The evaluation results of our experimental system demonstrate the proposed method sufficiently improve the performance.

1. はじめに

大規模データ処理システムにおけるストレージ管理コストの増加に伴い，ストレージをシステムの中心に据えるストレージセントリックシステムが注目されている．我々はストレージシステムにおける負荷分散，故障対策，障害回復などの高度な機能を実現するためのアプローチとして，ディスク装置内の制御用プロセッサとキャッシュ用メモリを利用する自律ディスク [1] を提案してきた．しかし，ストレージシステムで扱われるデータ量は未だ増加の一途をたどり，その性能に対する要求もより高いものが求められている．例えば典型的なトランザクションシステムの処理性能向上のために，より高速なレスポンスがストレージに対して求められている．しかし，現在のストレージシステムの基礎となる記憶媒体は磁気ディスクであり，その構造から飛躍的な速度向上は望めない状況である．

これらの問題に対して我々は，性能の異なるデバイスを組み合わせて階層的な自律ストレージを構成する手法を提案し，性能評価を行ってきた [2, 3].

本稿ではこの自律ディスククラスタの階層的構成手法において，ディレクトリ探索とメッセージ通信の回数削減による高速化手法を提案する．さらに模擬実装を用いた実験により性能評価を行い，提案手法の有効性を示す.

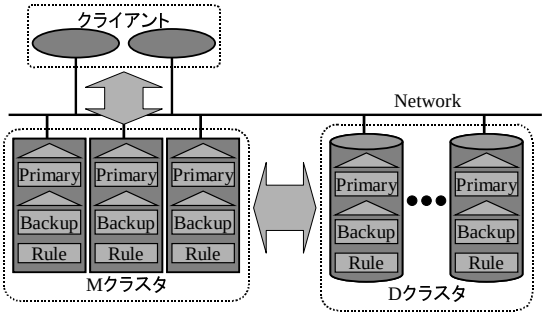

図 1: 階層化アーキテクチャ
Fig.1 Hierarchical architecture

2. 自律ディスクとその階層化構成

2.1 自律ディスク

はじめに自律ディスクの概要について説明する．自律ディスクはネットワーク環境でディスククラスタを構成することを前提としている．クライアントはデータにアクセスするためにクラスタ内の任意のノードにリクエストを発行し，クラスタ内のノードはノード間の局所的な通信を行うことで，協力してクライアントからの要求に対処する．標準的な構成では，クラスタ内の各ノードはプライマリディレクトリと他ノードのバックアップを行うバックアップディレクトリを持つ．データに対するアクセスは分散ディレクトリをトラバースし，リクエストを適切なノードに転送することにより行われる．このような前提のもとで，自律ディスクはデータ分散，ホストからの均質的なアクセス，同時実行制御，偏り制御，耐故障性，異種性といった性質を持つ．自律ディスクの詳細については文献 [1] を参照されたい.

2.2 自律ディスクの階層化構成

次に我々の提案する自律ディスクの階層化構成手法の概要を説明する．我々の提案する手法では，磁気ディスクのみを用いて構成された自律ディスククラスタ (D クラスタ) と，半導体ディスクのみを用いて構成された自律ディスククラスタ (M クラスタ) をそれぞれ構成し，M クラスタが D クラスタのキャッシュとして動作する．このアーキテクチャの概要を図 1 に示す.

我々はこのアーキテクチャにおいて Write-Through-Sync(WTS) プロトコル，Write-Through-Async(WTA) プロトコル，Delayed-Write(DW) プロトコルという 3 種類の挿入・更新プロトコルと，読み出しプロトコルを提案している．3 種類の挿入・更新プロトコルはいずれも耐故障性を持つように設計されており，主な違いはクライアントのリクエストに対して同期的に書き込みが行われる

♡ 正会員 (株) 日立製作所 中央研究所 t-hanai@crl.hitachi.co.jp
♢ 学生会員 東京工業大学 大学院 情報理工学研究科 計算工学専攻 {aki,daik,muu}@de.cs.titech.ac.jp
♠ NHK 放送技術研究所 {taguchi.r-cs,hayashi.n-gm,uehara.t-jy}@nhk.or.jp
♣ 正会員 東京工業大学 学術国際情報センター yokota@cs.titech.ac.jp

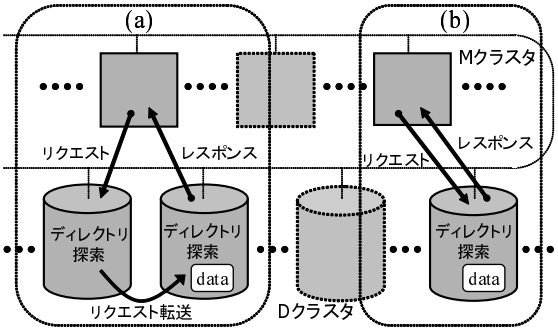

図 2: リクエストの転送
Fig.2 Request forwarding

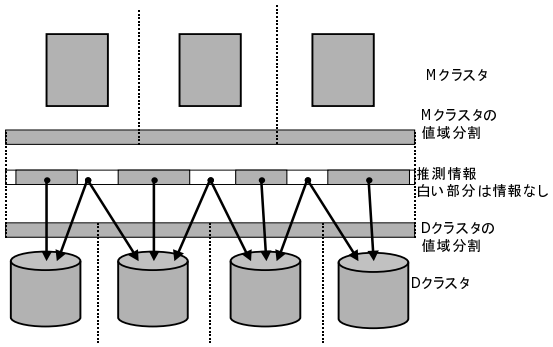

図 3: リクエスト送信先の推測
Fig.3 Request-forwarding target presuming

箇所である．以下に各プロトコルの概要を述べる．詳細は [2, 3] を参照されたい．

WTS プロトコル M クラスタのプライマリ，D クラスタのプライマリとバックアップへ同期的に書き込みが行われる．

WTA プロトコル D クラスタ内のバックアップ生成を非同期的に行うことにより WTS を高速化したもの．

DW プロトコル M クラスタ内でのみ書き込みを行い，クライアントのリクエストとは非同期に D クラスタへ書き込みを行うため高速な書き込みが行える．

読み出しプロトコル M クラスタ内でリクエストされたデータを探索し，存在すればそれをクライアントに返す．存在しない場合は D クラスタからリクエストされたデータを読み出す．

3. ディレクトリ探索とメッセージ転送回数の削減

自律ディスクは負荷分散や偏り制御のためにデータ配置を分散ディレクトリによってクライアントから透過的に管理しており，クラスタ中のいかなるノードに格納されているデータに対するリクエストも任意のノードが受け付ける．クライアントはデータ配置について意識することはなく，データにアクセスする際にはクラスタ中の任意のノードを選択してリクエストを送信する．このため，あるノードが他のノードに格納されているデータに対するリクエストを受け付けた場合は，図 2(a) に示すように，そのリクエストは分散ディレクトリを探索することによってデータを格納しているノードに自動的に転送される．これまで研究における性能評価はクライアントがリクエストを送信する際にクラスタ中の各ノードをランダムに選ぶ条件の元で行われてきた．自律ディスククラスタの階層的構成においては，M クラスタは D クラスタのクライアントとして動作する．この構成において M クラスタが D クラスタへリクエストを送信する際に，図 2(b) に示すように D クラスタ内のデータを格納しているノードへリクエストを直接送信することができれば，ランダムに送信先ノードを選択する場合に比べ D クラスタ内でのディレクトリ探索処理やリクエストの転送処理を削減することができる．

3.1 提案手法

この節では，M クラスタから D クラスタへのリクエスト送信時に送信先を推測することにより，D クラスタ内でのディレクトリ探索とメッセージ転送回数を削減し，D クラスタの負荷を低減することによる高速化手法を提案する．

自律ディスクではデータの配置管理を行う分散ディレクトリとして，Fat-Btree[4] などの分散 Btree の利用が想定されている．この場合，データの名前空間は値域分割され，自律ディスクの各ノードに値域が割り当てられる．また，自律ディスクのアーキテクチャでは挿入・更新や読み出しが行われたノードがクライアントにレスポンスを返すため，クライアント (M クラスタ) はどのノードで実際の処理が行われたかを知ることができる．そこで，M クラスタが D クラスタからレスポンスを受信した場合に，どのキー (ストリーム ID) に対して D クラスタのどのノードで処理が行われたかを記録しておく．そして，M クラスタが D クラスタへリクエストを送信する場合には，その記録を元にリクエストが処理されるノードを推測することにより，D クラスタ内でのメッセージ転送回数を削減することができる．また，D クラスタ内でのディレクトリ探索処理が効率的に行われるため，D クラスタの負荷を低減することができる．これにより，十分な処理結果が記録された後には，非常に高い精度で M クラスタから処理を行う D クラスタのノードへリクエスト送ることができ，リクエストの処理時間を短縮することが可能となる．

実際にはリクエストの処理結果を全て記録する必要はなく，D クラスタの各ノードに対して割り当てられている値域を記録できればよい．この手法の概略を図 3 に示す．過去に処理結果が得られていないため推測が不可能な場合（図 3 で推測情報が白い部分）は，その左右のノードどちらかを等確率で選ぶものとする．また，D クラスタ内での偏り制御により D クラスタの値域分割の割り当ては変化するが，偏り制御による値域分割割り当ての変化は時間的に緩やかであり，記録された推測情報との差は推測情報の更新により修正されてゆくので，大量の推測ミスを引き起こすことはない．推測ミスの際にも自律ディスクの通常動作として D クラスタ内で適切にリクエストが転送されるので，本提案手法によらない場合に比べて処理の増大を引き起こすことはない．

このようにリクエストの処理結果から推測情報を構築することにより，D クラスタが能動的に値域分割の情報を配布する必要がなくなる．D クラスタ内の値域分割は偏り制御のために頻繁に変更されるので，変更の度に M クラスタへ値域分割の情報を配布すると通信量や処理負荷が大きくなる可能性がある．また能動的に値域分割の情報を配布することにより各クラスタの独立性が減少することになり，階層化構成における柔軟性が減少する．さらに D クラスタのノード数が多い場合には値域分割の情報が大きくなる可能性があるが，提案手法では M クラスタの各ノードが独立した推測情報を持ち，M クラスタの各ノードに割り当てられた値域に関する推測情報のみを保持するため，推測情報の量を D クラスタ全体の値域分割情報の量に比べて非常に少なく抑えることができる．関連研究として [5] で提案されている RP^*_C や RP^*_S があり，これらはクライアントが各ノードの値域分割割り当てを記録してリクエスト送信時の送信先決定に利用するが，値域分割したデータの配置方法については述べていない．これらの手法を分散 Btree へ適用した場合に比べ，本稿で提案する手法は D クラスタの変更が必要なく，推測情報の管理が単純であるという特徴を持つ．

3.2 推測情報管理アルゴリズム

推測情報は以下のアルゴリズムで管理する．前節に加えて記法を以下のように定義する．

- M クラスタの各ノードには 0 から $M-1$ まで，D クラスタの各ノードには 0 から $N-1$ までのノード ID が割り当てられている．
- D クラスタの値域分割において，ノード i の値域 $R_{Di} = [K_i, K_{i+1})$ とする．ここで K_i は値域分割における境界のキー値．

M クラスタのノード m が持つ推測情報は範囲のリスト $G_m = (r_i, r_j, r_k, \ldots)$ である．$r_i = [k_{iL}, k_{iR}]$ であり，範囲（値域）を表す．

表 1: 実験システムの構成
Table 1 The experimental system

	D クラスタ	M クラスタ	クライアント
ノード数	6 台	3 台	1 台
CPU	Intel Pentium 3 933MHz	Intel Xeon 2.4GHz, HT	Intel Pentium4 2.0GHz
メモリ	PC133 SDRAM 32MB	PC2100 DDR SDRAM 1024MB	PC2100 DDR SDRAM 1024MB
記憶媒体	Seagate Barracuda IV 20.4GB 7200rpm	Linux ramfs	Seagate Barracuda IV 20.4GB 7200rpm
OS	Linux 2.2.17 glibc 2.1.3	Linux 2.4.20 SMP glibc 2.2.5	Linux 2.4.18 glibc 2.2.5
Java 環境	Sun JDK 1.4.1		
Network	1000BASE-SX, Switching Hub		

$k_{iL} \leq k_{iR}$ とする．G_m 中の各 r_i は範囲が重ならず，小さい順にソートされているものとする．i, j, k は D クラスタのノード ID である．

D クラスタへキー k に関するリクエストの送信先を推測する場合は，G_m から k を含む r_i を探索し，存在した場合は D クラスタのノード i を選択する．存在しなかった場合は k の左側の範囲 r_i と右側の範囲 r_j を探索し，ノード i または j を等確率で選択する．

送信したリクエストのレスポンスがノード d から返ってきた場合は，以下のように推測情報を更新する．

1. 初期状態では $G_m = ()$ である．
2. G_m から r_d を探索する．r_d が存在する場合は 3 へ．存在しない場合は 4 へ．
3. $k < k_{dL}$ ならば $r_d = [k, k_{dR}]$ として 5 へ．$k > k_{dR}$ ならば $r_d = [k_{dR}, k]$ として 6 へ．その他の場合は終了．
4. $r_d = [k, k]$ として G_m に追加して終了．
5. r_d の右の r を r_R とする．$k_{RR} < k$ ならば r_R を削除して終了．$k_{RL} < k$ ならば $r_R = (k, k_{RR}]$ として終了．その他の場合は何もせず終了．
6. r_d の左の r を r_L とする．$k_{LL} > k$ ならば r_L を削除して終了．$k_{LR} > k$ ならば $r_L = [k_{LL}, k)$ として終了．その他の場合は何もせず終了．

4. 性能評価

次に前章で述べた高速化手法を実験により性能測定を行い評価する．まず本稿で提案する高速化手法を，我々が現在 PC 上に Java を用いて実装を行っている試作システム上に実装した．その上で挿入・更新操作，読み出し操作それぞれについて，一定回数のリクエスト処理にかかる時間を測定し，性能を評価する．

4.1 実験方法

性能測定に用いた実験環境を表 1 に示す．M クラスタの記憶媒体としては，Linux の ramfs ファイルシステムにより M クラスタノード上の主記憶を用いた．M クラスタにおける最大キャッシュエントリ数は 1 ノードあたり 600 とした．また，D クラスタについては OS のバッファキャッシュの影響を小さくし，実際にディスクアクセスが発生するように，主記憶の容量を 32MB に制限するとともに，模擬実装上でのページサイズを M クラスタの 4KB に対して 16KB と大きく設定した．実装の簡便化のために自律ディスクの分散ディレクトリとして aB⁺-tree[6] を用いている．

この環境に対して，以下の手法で性能測定を行う．挿入・更新操作，読み出し操作をそれぞれ一定回数行い，処理完了までの所要時間を計測する．リクエストは 6000 個のキー集合を用い，連続して 30000 回行う．格納されるデータのキーは 16 バイトのランダムな文字列であり，データサイズは 1K バイトとする．値域分割は値域の範囲が均等になるように行い，性能測定中に偏り制御は行わないものとする．本稿で提案する推測手法を使用した場合，推測手法を使用せずに D クラスタのノードをランダムに選択した場合，階層化を行わずに D クラスタのみで構成した場合に対して測定を行う．推測手法を用いた場合の推測情報は，空の状態から測定を始める．

表 2: M, D クラスタの性能測定
Table 2 Performance of M and D clusters

操作	Insert		Retrieve	
クラスタ	M	D	M	D
合計所要時間 [sec]	133.8	1602.1	64.41	329.9
1 リクエスト処理時間 [ms]	4.46	53.4	2.15	11.0
処理所要時間比	1 : 11.97		1 : 5.12	

表 3: 偏りに対するキャッシュヒット率
Table 3 Cache hit ratio for skew

偏りの度合い θ	0.0	0.5	1.0	1.5	2.0
キャッシュヒット率 r [%]	24.7	34.6	63.5	77.4	74.7

4.2 予備実験

各操作の性能測定を行う前に，M クラスタと D クラスタ各々の性能を測定する．各々のクラスタで独立に従来の自律ディスククラスタを構成し，偏りのない状態で Insert，Retrieve 操作の性能を計測した結果が表 2 である．この結果より，この実験システムにおいて M クラスタは D クラスタに比べ，Insert で約 12 倍, Retrieve で 5.1 倍程度高速であることがわかる．Insert と Retrieve で性能比が異なる理由としては，Insert 処理では実際に書き込みを行うために記録媒体の性能差が大きく現れるのに対し，Retrieve 処理の場合は D クラスタにおいてもディスクキャッシュが利用されることにより記録媒体の性能差が現れにくいためである．また Insert 処理ではプライマリデータとバックアップデータの両方にアクセスするのに対し，Retrieve 処理ではプライマリデータにのみアクセスするので D クラスタにおけるディスクキャッシュの利用率が向上する．

4.3 挿入・更新性能

前述の手法によって挿入・更新操作の性能を測定した結果を図 4 に示す． 各プロトコルともに提案手法により処理時間が短縮されており，本稿で提案する推測手法が有効であることがわかる．提案手法により，WTS プロトコルでは 16%，WTA プロトコルでは 19%，DW プロトコルでは 15%の時間短縮になっている．特に WTS プロトコルでは提案手法の適用により階層化なしの場合に比べて高速になっている．また DW プロトコルでは D クラスタへの挿入・更新処理がリクエストと非同期に行われるため提案手法の効果は小さいはずであるが，提案手法により高速化されている．これは，今回の実験環境では M クラスタと D クラスタの性能差が大きいため，M クラスタであふれたキャッシュの追い出し処理による D クラスタへの挿入・更新処理がリクエストの処理時間に影響を与えているためである．

また，階層化構成の各挿入・更新プロトコルと階層化しない場合を比較した場合，WTS プロトコルでは 12%，WTA プロトコルでは 38%，DW プロトコルでは 72%の時間短縮になっている．これにより階層化構成の全ての挿入・更新プロトコルで階層化しない場合に比べて高速になり，階層化構成の有効性が確認された．

4.4 読み出し性能

読み出し操作の性能測定では，リクエストされるキーの偏りによる性能の変化を評価するために，各キーの利用頻度を Zipf 分布とし，その偏りの度合いを表すパラメータ θ を 0.0〜2.0 に変化させて測定を行った．このパラメータは $\theta = 0$ の場合には偏りがなく，θ が大きくなるに従って偏りも大きくなるものである．

読み出し操作の性能を測定した結果を図 5 に示す．また，この測定において M クラスタでのキャッシュヒット率 r を測定した結果を表 3 に示す．読み出し処理についても提案手法により処理時間が短縮されていることがわかる．偏りの度合い $\theta = 1.0$ の場合で提案手法により 10%の時間短縮が見られた．しかし，偏りの度合

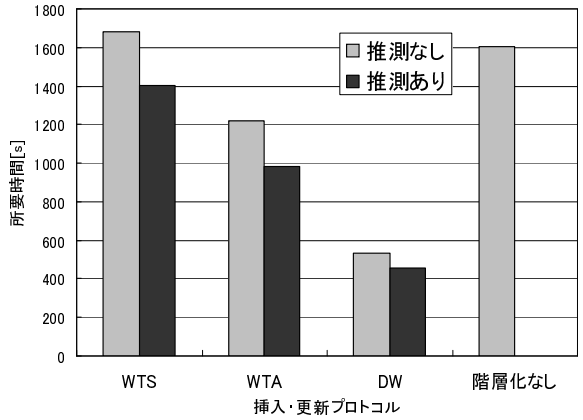

図 4: 挿入・更新操作の実行時間
Fig.4 Execution time for Insert operations

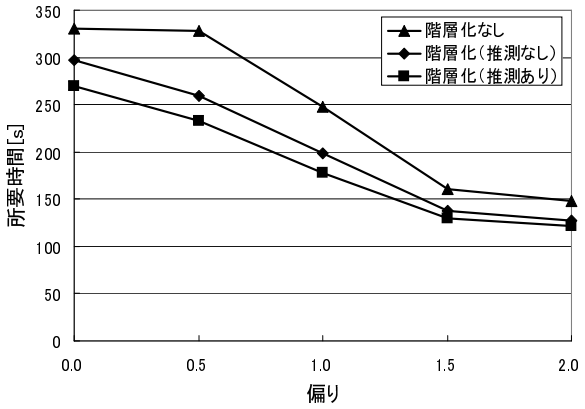

図 5: 読み出し操作の実行時間
Fig.5 Execution time for Retrieve operations

いが大きくなるにつれて M クラスタでのキャッシュヒット率が向上するため，D クラスタへのリクエストの割合が少なくなり，提案手法の効果は小さくなっている．

階層化手法と階層化しない場合を比べた場合，提案手法を利用した読み出しプロトコルは階層化しない場合に比べ $\theta = 0.5$ の場合で 29%，$\theta = 1.0$ の場合で 28.5%の時間短縮になっている．これにより，読み出しの場合についても階層化構成の有効性が示された．

5. まとめと今後の課題

本稿では，自律ディスククラスタの階層的構成手法において，D クラスタ内のディレクトリ探索とメッセージ転送の回数を低減することによる高速化手法を提案した．これは，高速なクラスタである M クラスタから低速なクラスタである D クラスタへリクエストを送信する際に，過去のリクエスト処理の記録から送信先ノードを推測する手法である．さらに実験による性能測定を行い，本稿で提案する推測手法の有効性を確認した．

本稿の評価実験では単一のクライアントを想定して測定を行ったが，複数のクライアントが存在する場合についても階層化構成の有効性を検証する必要がある．またクライアントのアクセスパターンが時間的に変化するような場合の性能についても検討する必要がある．

さらに，これまでは 2 種の異なるデバイスを用いて 2 階層のストレージシステムを構築することを対象としてきたが，我々が提案する自律ディスククラスタの階層的構成手法は 3 階層以上のストレージシステムに対しても適用が可能であると考えられる．3 階層以上のストレージシステムに対して本研究の提案手法の適用のためには，さらなる検討が必要である．

[謝辞]

本研究の一部は，文部科学省科学研究費補助金特定領域研究 (15017233)，情報ストレージ研究推進機構 (SRC)，独立行政法人科学技術振興機構 CREST，および 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の助成により行なわれた．

[文献]

[1] Haruo Yokota. Autonomous Disks for Advanced Database Applications. In *Proc. of International Symposium on Database Applications in Non-Traditional Environments (DANTE'99)*, pp. 441–448, Nov. 1999.

[2] 花井知広, 渡邊明嗣, 山口宗慶, 田口亮, 林直人, 上原年博, 横田治夫. 半導体ディスクを用いた自律ディスクの階層化. 情処学会研究会報告, データベースシステム DBS-131-19. 情報処理学会, 2003.

[3] 花井知広, 渡邊明嗣, 山口宗慶, 田口亮, 林直人, 上原年博, 横田治夫. 半導体ディスクによる自律ディスククラスタの階層化構成. 日本データベース学会 Letters, Vol. 2, No. 3, pp. 41–44, Dec. 2003.

[4] Haruo Yokota, Yasuhiko Kanemasa, and Jun Miyazaki. Fat-Btree: An Update-Conscious Parallel Directory Structure. In *Proc. of the 15th Int'l Conf. on Data Engineering*, pp. 448–457, 1999.

[5] Witold Litwin, Marie-Anne Neimat, and Donovan A. Schneider. RP*: A Family of Order Preserving Scalable Distributed Data Structures. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, pp. 342–353, September 1994.

[6] Mong Li Lee, Masaru Kitsuregawa, Beng-Chin Ooi, Kian-Lee Tan, and Anirban Mondal. Towards Self-Tuning Data Plaement in Parallel Database Systems. *SIGMOD Record*, Vol. 29, No. 2, pp. 225–236, Sep. 2000.

花井 知広　Tomohiro HANAI
平 16 東工大大学院・情報理工・計算工・修士課程了．同年より (株) 日立製作所中央研究所．日本データベース学会正会員．

渡邊 明嗣　Akitsugu WATANABE
平 14 東工大大学院・情報理工・計算工・博士前期課程了．同大大学院・情報理工・計算工・博士後期課程在学中．日本データベース学会学生会員．

小林 大　Dai KOBAYASHI
平 15 東工大・工・情工卒．同大大学院・情報理工・計算工・修士課程在学中．

山口 宗慶　Munenori YAMAGUCHI
平 15 東工大・工・情工卒．同大大学院・情報理工・計算工・修士課程在学中．

田口 亮　Ryo TAGUCHI
平 6 慶應義塾大大学院・理工・計測工・修士課程了．同年より NHK 放送技術研究所．映像情報メディア学会会員．

林 直人　Naoto HAYASHI
昭 58 金沢大・工・精密工卒．同年より NHK 放送技術研究所．映像情報メディア学会会員．

上原 年博　Toshihiro UEHARA
昭 56 慶應義塾大・工・電気工・修士課程了．昭 59 より NHK 放送技術研究所．電子情報通信学会，映像情報メディア学会各会員．

横田 治夫　Haruo YOKOTA
昭 55 東工大・工・電物卒．昭 57 同大大学院・情報・修士課程了．同年富士通 (株)．同年 6 月 (財) 新世代コンピュータ技術開発機構研究所．昭 61(株) 富士通研究所．平 4 北陸先端大・情報・助教授．平 10 東工大・情報理工・助教授．平 13 東工大・学術国際情報センター・教授．工博．日本データベース学会, 電子情報通信学会, 情報処理学会，人工知能学会，IEEE，ACM 各会員．