

同時実行 B⁺ 木におけるロックフリー手続きの改善と実装

平野 匠真^{1,a)} 杉浦 健人^{1,b)} 石川 佳治^{1,c)} 陸 可鏡^{1,d)}

概要:

単一コアの性能向上が限界を迎え、複数コアを活用するマルチスレッド処理が主流となってきた。索引技術においてもその傾向は同様であり、CPU のメニーコア化に伴い、マルチスレッド処理を意識した索引構造が多く提案されている。マルチスレッド処理におけるロックを用いた同時実行制御は、スケーラビリティが悪いため、B⁺ 木を基にした Bw 木や Bz 木といったロックフリー索引に対して注目が集まっている。しかし、これらの索引はマルチスレッド環境下では性能が向上せず、ロックフリー索引には更なる改善が必要であると考えられる。そこで、本研究では B⁺ 木をロックフリー化させた新たな索引構造である B^c 木を提案し、その構造および操作について述べる。

1. はじめに

ムーアの法則が限界を迎え、CPU のコア単位の性能向上はほとんど止まっている。現在のコンピュータ技術では CPU に搭載された複数のコアを活用するマルチスレッド処理が主流となっており、マルチスレッドを効果的に使用するソフトウェアやデータ構造などが提案されている。索引技術においてもその傾向は同様であり、近年ではメニーコア CPU 上でのマルチスレッド処理を意識した索引構造が多く提案されている。

代表的な索引構造である B⁺ 木 [5] では、ロックを用いた同時実行制御が行われている。しかし、マルチスレッド処理においてロックによる同時実行制御は多数の待ちスレッドが発生するため、スケーラビリティが悪化する。そこで、B⁺ 木を基にした Bw 木 [3] や Bz 木 [1] といったロックを取得しないロックフリー索引が提案されている。しかし、これらの索引は書き込みの際に競合や多数のスレッドの失敗が発生することによって、マルチスレッド環境では性能が向上していない。そのため、ロックフリー索引には改善の余地が残されていると考えられる。

本研究では B⁺ 木をロックフリー化させた新たな索引構造である B^c 木を提案し、その構造および操作について述べる。そして提案した索引構造を実装し、Bw 木や Bz 木といったロックフリー索引と比較し、その性能を検証する。

本稿の構成は以下の通りである。2 章では、ロックフリー索引の関連研究として Bw 木および Bz 木について概説する。次に、3 章で B^c 木の構造について説明し、4 章で B^c 木の操作について述べる。最後に、5 章で本稿のまとめと今後の方針を述べる。

2. 関連研究

近年では、ロックを取得しないロックフリー索引に対して注目が集まっている。本節では、代表的なロックフリー索引である Bw 木および Bz 木について説明する。

2.1 Bw 木

Bw 木は、直感的にはロックフリーの単方向連結リストと B⁺ 木を組み合わせた索引構造であり、更新内容が記述された差分レコード (delta record) をノードの前に連結リストとして挿入し、索引に対する全ての更新を表す。差分レコードのリストが一定以上の長さを持つとき、ノードの統合走査が行われ、統合後は B⁺ 木のノードと同様のものが生成される。また、単一の CAS 命令を用いてノード間のポインタをインストールできるようメモリ内のデータ構造を設計することにより、木のロックフリー化を実現している [4]。

Bw 木が持つ他の木と異なる点として、差分レコードの他にマッピングテーブルがある。マッピングテーブルはノード間のリンクを仮想化し、メインメモリ上の差分更新を可能とする。各ノードは自身の子ノードや兄弟ノードへのポインタを直接持つ代わりにマッピングテーブル上の ID (logical page ID, LPID) を持つ。つまり、マッピングテ

¹ 名古屋大学大学院情報学研究科
Graduate School of Informatics, Nagoya University

a) hirano@db.is.i.nagoya-u.ac.jp

b) sugiura@i.nagoya-u.ac.jp

c) ishikawa@i.nagoya-u.ac.jp

d) example.com

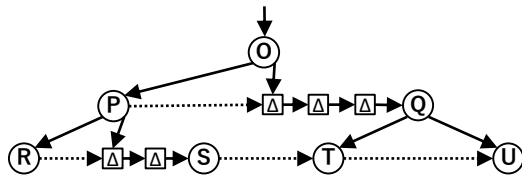


図1 Bw 木の簡単な木構造

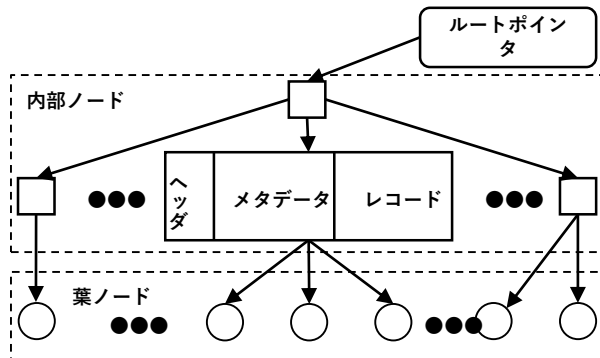


図2 Bz 木の簡単な木構造

ブルを用いて LPID を実際のポインタへと変換することでノードを辿る。

2.2 Bz 木

Bz 木は、 B^+ 木の葉ノード内部にロックフリーの固定長配列を持つ索引構造であり、図 2 のような構造を持つ。Bz 木は、Bz 木の根を示すルートポインタと、複数の内部ノードと葉ノードからなる。ルートポインタが木の根となるノードへのポインタを持ち、根ノードおよび中間ノードなどの内部ノードが子となる内部ノードまたは葉ノードへのポインタを持つ。Bz 木は単一の CAS 命令ではなく、メモリ上の複数のワードを対象とする multi-word compare-and-swap (MwCAS) 命令を用いて木のロックフリー化を実現している。

Bz 木の各ノードにはヘッダ、メタデータ、レコードが格納されている。ヘッダにはステータスワードというフィールドが格納されており、ステータスワードは CAS 命令や構造変更操作の制御、ノード内のレコード数などノード更新時に必要な情報を格納している。Bz 木はノードに対する書き込みおよび構造変更操作をステータスワードを更新することによってロックフリーに実現している。

3. B^c 木の構造

B^c 木は Bw 木と Bz 木を組み合わせたロックフリー索引である。 B^c 木の概形を図 3 に示す。木の論理的な構造は B^{link} 木 [2] に則っており、各ノードが同じ階層の右兄弟への参照リンクを持つ。各ノードへの参照はマッピングテーブルを用いた間接参照を採用し、マッピングテーブル内の物理ポインタを差し替えることでそのノードへの参照を一

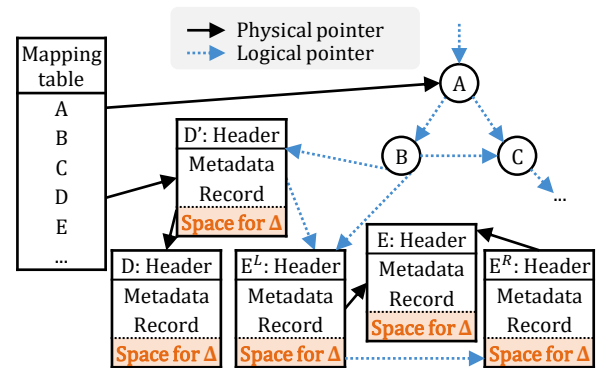


図3 B^c 木の概形

括で変更する。

B^+ 木と同様に、 B^c 木は索引層およびデータ層によって構成される。索引層のノード（中間ノード）は分割キーと子ノードへのポインタの組を格納し、木の下方向への検索を補助する。

各ノードの領域は immutable 領域と mutable 領域に分けられる。immutable 領域はノードヘッダおよびソート済みのレコードを格納する。ヘッダは immutable 領域の情報を管理し、構造変更時のみその値が変更される。mutable 領域はステータスワードの格納と差分レコードを挿入するための書き込みバッファの役割を果たす。ステータスワードは mutable 領域の情報を管理し、ノードの現在の状態や残容量などを管理する。各ノードへ構造変更操作を行う際は、構造変更後のノードから構造変更前のノードへ物理リンクを張り、古いノードへの参照を可能にする。なお、古いノードは構造変更が済み次第、ガベージコレクションへ追加され、他のスレッドから参照されないことが保証された時点で削除される。

B^c 木は Bw 木および Bz 木でそれぞれボトルネックとなる箇所を以下の通りに克服している。Bw 木ではデルタレコードによって更新を実現しているが、差分レコードの探索はランダムアクセスでありキャッシュヒット率が低い。そこで Bz 木のノード内の書き込みバッファを採用することで、Bw 木で発生していたデルタレコードによるキャッシュミス削減している。次に、Bz 木は差分レコードの書き込みおよび対応するレコードメタデータの更新を MwCAS 命令を用いたステータスワードの更新によって実現している。また、同時に発生した構造変更操作を検知するためにステータスワードが変更されていないことを MwCAS 命令を用いて確認している。このように Bz 木では、各操作がステータスワードに強く依存している。そこで、差分レコード用の領域を確保する際にステータスワードを更新し、差分レコードの書き込みおよびメタデータの挿入を完了するといった手法でステータスワードの更新するフィールドを少なくする。また、Bw 木のマッピングテーブルを採用することで、構造変更後のノードに他のスレッドが到達でき

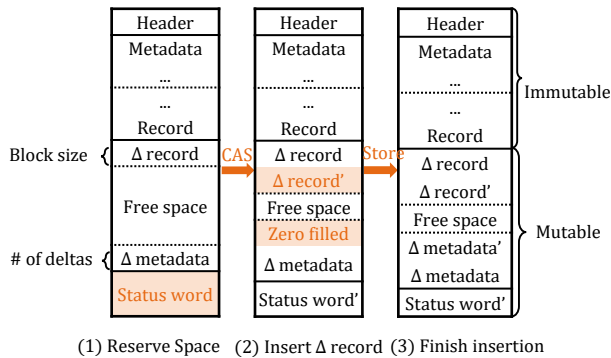


図4 B^c 木における差分レコードの挿入

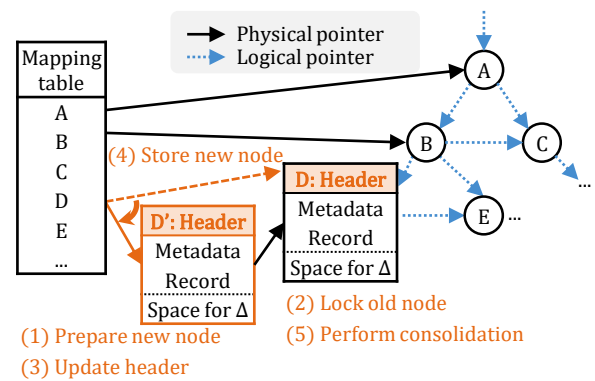


図5 B^c 木における統合操作

る．以上により，B_Z 木のようなステータスワードに対する強い依存を軽減している．

4. B^c 木の操作

本節では，B^c 木のノード操作および構造変更操作について述べる．

4.1 ノード操作

B^c 木は以下の読み取りおよび書き込み操作をサポートする．

4.1.0.1 書き込み

書き込み操作はキーとペイロードを差分レコード領域に挿入する操作である．キーが属する書き込み先の葉ノードを根ノードから二部探索により特定する．挿入先の葉ノードに到達後，その葉ノードの差分レコード領域に値を挿入する．

書き込み操作は差分レコード領域の予約とレコード挿入および可視化の2ステップで行われる．図4にB^c 木における差分レコード挿入を示す．

ノードヘッダには mutable 領域の状態を表すステータスワードを用意し，この中のレコード数と使用済みブロックサイズを加算することで差分レコード用の領域を予約する．確保した領域へ差分レコードを書き込み，対応するレコードメタデータの値を更新することで挿入処理を完了する．

以上の書き込み操作によって差分レコードを挿入していくが，挿入後の差分レコードの総数またはノード容量のしきい値を越える場合，差分レコードの統合操作または構造変更操作が行われる．しきい値の確認はステータスワードの更新時に行われ，しきい値を越えた場合はレコードの書き込み後に統合操作，構造変更操作いずれかの操作を行うか判定する．

書き込み処理は基本的にロックフリーに動作する．書き込み同士の競合は基本的にステータスワードで解決され，差分レコード用領域の予約は CAS 命令によって順序付けられる．

4.1.0.2 読み取り

読み取り (Read) は与えられた対象キーのペイロード返す操作である．対象キーが与えられたとき，根ノードからの二部探索によりキーが属する葉ノードに到達する．葉ノードに到達した後は，葉ノード内の差分レコードを線形探索と mutable レコードの二部探索の2ステップで読み取り操作を行う．対象キーが存在している場合には，対応するペイロードを読み返す．対象キーが存在していない場合には，存在していない旨を返す．また，B^c 木は始点キーおよび終点キーを指定し，そのキー範囲内の全てのペイロードを返す範囲読み取り (Scan) 操作も行う．

最新の値は差分レコード領域に書き込まれるため，まず mutable 領域を線形探索し，一致するキーがあるか確認する．このとき，差分レコードの数はステータスワードから読み取り，差分レコードが可視化されていなければスキップする．差分レコード中に対象のレコードが存在しなければ，ノードヘッダから immutable レコードの数を読み取り，二部探索によって対象キーの有無を確認する．先頭ノードが統合操作の途中である場合には，差分レコードを読み終わった時点で物理ポインタをたどり古いノードへ移動し，同様の2ステップを古いノード上で行う．

読み取り処理は wait-free に動作する．読み取り処理は一切のノード状態を変更しないため，読み取った状態に応じて手続きを選択する処理が止まることはない．

4.2 構造変更操作

B^c 木は以下の統合，分割およびマージをサポートする．構造変更操作を始める前に，新たなノードをマッピングテーブルに挿入することで，他のスレッドの待ち時間を削減する．

4.2.0.1 統合

統合操作は差分レコードの数またはノードの容量がそれぞれのしきい値を越えた場合，差分レコードと immutable レコードをソートして immutable 領域に反映する操作である．図5にB^c 木における統合操作を示す．

統合が必要になった場合，新たなノード領域を用意す

る．そして、古いノードのステータスワード更新後、ノードヘッダの情報から統合後に必要な immutable 領域を計算する．次に、新たなノードにおいてステータスワードを含むヘッダ情報のみを更新した後、新規ノードとしてマッピングテーブル上の参照を更新する．その後、古いノード上の差分レコードを immutable レコードへ反映させつつ、新たなノードへレコードをコピーする．最後に、統合後の状態でノードヘッダを更新し、ステータスワードを CAS 命令を用いて更新する．この間に発生した書き込み操作は新規ノードの mutable 領域で受け付ける．これにより統合操作の間も、読み取り操作は新規レコードの mutable 領域、統合前のノードの mutable 領域および immutable 領域の順でレコードを確認することで読み取り操作を実行できる．

4.2.0.2 分割

分割操作では対象ノードをロックした後、レコードコピーを始める前に分割したノードを用意しマッピングテーブルに挿入する．これを実現するために、各ノードは自身の分割キーの位置を統合および分割操作のたびに記録する．

分割したノードの挿入後、統合操作と同様の手続きでレコードをコピーする．分割時には、古いノードの分割キー未満の全てのレコードを左ノードにコピーする．残りのレコード、すなわち分割キー以上の全てのレコードは右ノードにコピーする．レコードのコピー後は、親ノードに右分割ノードへのリンクを挿入することで分割操作が完了する．このとき、中間ノードにおける CPU キャッシュ効率を高めるために、親ノードへはリンクの情報のみを挿入し、反映は親ノードの統合操作時に行う．

分割キーは immutable レコードから決定される．つまり差分レコードは考慮されないため、分割操作時に片方のノードにレコードの偏りが発生しうる．前述した通り、分割操作前にノードを用意するため、immutable 領域および対応するメタデータ数は余裕を持って確保しておく必要がある．そのため、差分レコードが片方のノードに集中し、その分のスペースがもう片方のノードで使用されない、未使用な immutable 領域が残る可能性がある．

5. おわりに

本稿では、 B^+ 木をロックフリー化させた新たな索引構造 B^c 木を提案し、その構造および操作を紹介した．今後は提案した索引構造を実装し、ノードのレコードを他のノードとつなぎ合わせるようなマージ操作など、他の操作を実装する．また、近年提案されている B_w 木や B_z 木といったロックフリー索引との性能を比較検証する．

謝辞 本研究は JSPS 科研費 JP20K19804, JP21H03555, JP22H03594 の助成、および国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP16007) の結果得られたものである．

参考文献

- [1] Arulraj, J., Levandoski, J., Minhas, U. F. and Larson, P.: BzTree: A High-Performance Latch-free Range Index for Non-Volatile Memory, *PVLDB*, Vol. 11, No. 5, pp. 553–565 (2018).
- [2] Lehman, P. L. and Yao, S. B.: Efficient Locking for Concurrent Operations on B-Trees, *ACM TODS*, Vol. 6, No. 4, pp. 650–670 (1981).
- [3] Levandoski, J., Lomet, D. and Sengupta, S.: The Bw-Tree: A B-tree for New Hardware Platforms, *Proc. ICDE*, pp. 302–313 (2013).
- [4] Petrov, A.: 詳説データベース: ストレージエンジンと分散データシステムの仕組み, オライリー・ジャパン (2021).
- [5] 北川 博之: データベースシステム, 昭晃堂 (1996).