2B-03

カーネルトレーサを用いたPostgreSQLの入出力挙動の観測と一考察

加藤千裕† 早水悠登† 合田和生† 喜連川優†‡
†東京大学生産技術研究所 ‡国立情報学研究所

1 はじめに

データベースシステムの性能を分析する際に，入出力挙動の観測は有効なアプローチの1つである．特にデータベースの性能低下の一原因となる，更新処理を行ったことによる物理的な構造の劣化（エージング）[1][2] に関しては，ストレージシステムへの入出力を観察することにより詳細な分析を行うことができると考えられる．本論文では，オープンソースのデータベースシステム PostgreSQL を対象として，著者らが行ったカーネルのイベントトレーサを用いた入出力挙動の観測を示し，エージングがどのような入出力挙動の変化をもたらすかについて考察する．

2 入出力挙動観測実験と考察

著者らは，エージング度合いの異なるデータベースを用意し，問合せ実行をする際の入出力挙動を観測し，その結果について考察を行った．

2.1 実験環境と実験用問合せ

実験環境のサーバとして，Dell Power Edge R720xd（Intel(R) Xeon(R) CPU E5-2690 v2，メモリ 64GB，CentOS release 5.8（64bit））を用い，磁気ディスクドライブとしては，10Krpm で 900GB の容量を備えたものを用いた．データベースシステムとして PostgreSQL 9.4.0 を用いた．TPC-H 付属の dbgen を用いてスケールファクタを 100 として初期データと更新クエリの作成を行った．PostgreSQL の設定パラメータはすべて初期状態とした．

エージングによる入出力挙動の変化を計測するため，エージングしていないデータベース（**初期状態**）と，90%のデータを，TPC-H の定める更新クエリで更新し

```
SELECT SUM(l_extendedprice) FROM lineitem
```

図 1: 問合せ（A）

```
SELECT SUM(l_extendedprice) FROM part
JOIN lineitem ON p_partkey = l_partkey
WHERE l_orderkey < 1024000
```

図 2: 問合せ（B）

エージングさせたデータベース（**エージング後**）を用意した．更新前後でデータベースのデータサイズは変化しない [1]．なお，どちらのデータベースも，実験前に vacuum コマンドにより削除ページの回収を行った．問合せとしては，図 1 と図 2 に示す lineitem 表の一属性値の総和を求める問合せ（A）と，lineitem 表と part 表の二つを結合する問合せ（B）を用いた．これらの問合せを実行中に，Systemtap を用いて，Linux カーネル内で SCSI 命令をトレースすることにより，入出力挙動を観測した．

2.2 実行計画と実行時間

問合せ（A）を実行した際には，表全てを読込む全表走査が行われた．また，問合せ（B）を実行した場合には，lineitem 表と part 表どちらに関しても，表の一部を索引を使って読込む索引走査が行われ，表の結合にはネステッドループ結合が用いられた．

問合せ（A）の実行には，初期状態のデータベースでは 695 秒を要し，エージングしたデータベースでは二倍近い 1281 秒を要した．問合せ（B）の実行時間は，初期状態のデータベースでは 1805 秒だったのに対し，エージング後は 2419 秒と，問合せ（A）ほど顕著ではないが，増加する傾向がみられた．

2.3 入出力挙動の観測結果

図 3 に，例として，問合せ（A）を実行中の入出力の挙動を示す．横軸は経過時間，縦軸はセクタ単位の

An Observation of Input and Output Behavior of PostgreSQL Using a Kernel Tracer and Its Analysis
Chihiro KATO†, Yuto HAYAMIZU†, Kazuo GODA†, Masaru KITSUREGAWA†‡
†Institute of Industrial Science, the University of Tokyo
‡National Institute of Informatics

{kato, haya, kgoda, kitsure}@tkl.iis.u-tokyo.ac.jp

[1] それぞれのデータベースは異なる磁気ディスクに格納されている．よって，磁気ディスク内における格納位置の影響は生じないものとする．

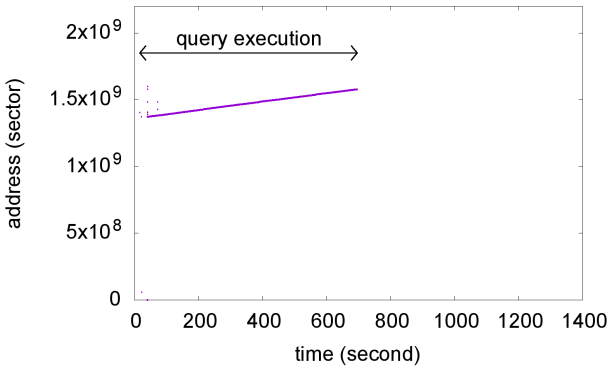

図 3: 初期状態の問合せ（A）の入出力挙動

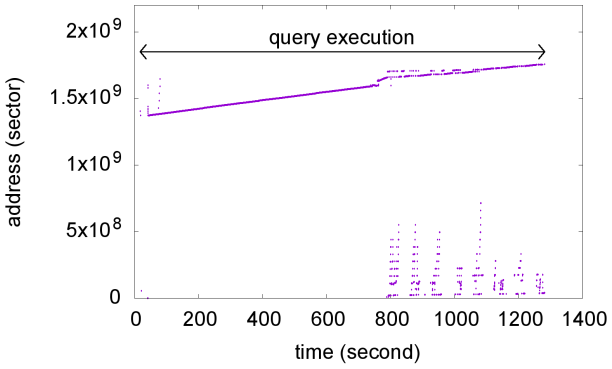

図 4: エージング後の問合せ（A）の入出力挙動

アドレスである．全表走査を実行する問合せ（A）は，シーケンシャルに読込みを行っていることが確認できる．対して，エージングしたデータベースにおいて，問合せ（A）の入出力挙動を観測した結果を，図 4 に示す．図 3 と図 4 を比較すると，実行時間が初期状態と比較して増加しており，初期状態の場合にはアクセスされなかった領域を読んでいることが観測された．データの更新を行った際に，これまでのアドレスの先にデータが格納されたことによって，全表走査で読む範囲が増大し，このような結果になったと考えられる．

2.4 入出力挙動の定量的な比較

二つの問合せについて，トレーサによる観測を基に，実行時間，読込みの入出力発行回数，その総データ量，ディスク上の総シーク距離を比較した．結果を纏めたグラフを図 5 に示す．グラフの数値は全て初期状態のデータベースの測定値を 1 として正規化を行っている．

問合せ（A）では，エージングによってデータが格納されている範囲が広がっているため，入出力発行回数と総データ量が増大する結果となった．入出力発行回数は，初期状態の 70.4 万回から，エージングにより 1.91 倍の 134 万回となり，92.2GB であった総データ量も 175GB と，1.90 倍に増加した．総シーク距離も，73 億セクタから 23.7 倍の 1740 億セクタになった．問合せ（B）に関しては，索引条件により指定された領域のみを読みだすため，入出力発行回数は初期状態の 43.2 万回に比べ，エージング後も 1.00 倍の 43.3 万回に収まっており，総データ量に関しても 3.71GB から 1.01 倍の 3.75GB と，変動が小さい結果となった．一方総シーク距離は 2.02 兆セクタから 195 兆セクタに増加し，96.8 倍となった．データの格納場所が分散したことによる総シーク距離の大幅な増加が，問合せ（B）の実行時間が増加する原因になったと考えられる．

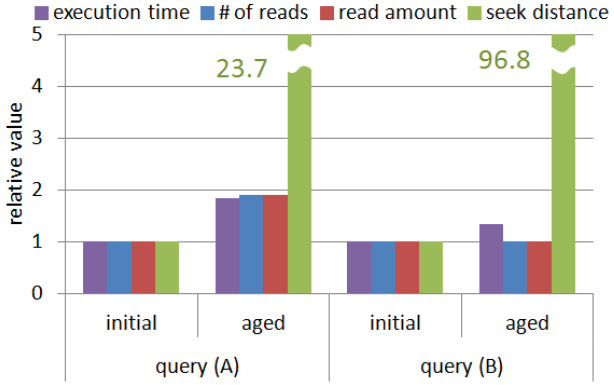

図 5: 問合せ毎のエージングの影響比較

3 終わりに

本論文では，PostgreSQL の入出力挙動をカーネルトレーサを用いて精緻に観測し，その結果に基づきエージングがもたらす影響について考察した．その結果，実際にエージングによってデータの格納領域が拡散し，問合せ実行時間の増加を引き起こす様子が観測された．また，問合せ実行に要する入出力命令数，読込み総データ量，総シーク距離という三つの指標に着目して定量的な評価分析を行った結果，問合せの種類に応じて，エージングの影響の受け方が大きく異なることを確認した．

参考文献

[1] Gary H Sockut and Robert P Goldberg. Database reorganization-principles and practice. *ACM Computing Surveys (CSUR)*, Vol. 11, No. 4, pp. 371–395, 1979.

[2] 合田和生, 喜連川優. データベース再編成機構を有するストレージシステム. 情報処理学会論文誌. データベース, Vol. 46, No. 8, pp. 130–147, 2005.