We want to join three tables, $S \bowtie R \bowtie T$. Assume $S$ is large but $R$ and $T$ are relatively small (but larger than CPU cache). Assume the two joins are on different join keys. Would you use non-partitioned hash join or radix join for this query? Please justify your choice.
An Evaluation of Buffer Management Strategies for Relational Database Systems

Hong-Tai Chou12 and David J. DeWitt1

Abstract. In this paper we present a new algorithm, DBMIN, for managing the buffer pool of a relational database management system. DBMIN is based on a new model of relational query behavior, the query locality set model (QLSM). Like the hot set model, the QLSM has an advantage over the stochastic models due to its ability to predict future reference behavior. However, the QLSM avoids the potential problems of the hot set model by separating the modeling of reference behavior from any particular buffer management algorithm. After introducing the QLSM and describing the DBMIN algorithm, we present a performance evaluation methodology for evaluating buffer management algorithms in a mult-user environment. This methodology employs a hybrid model that combines features of both trace-driven and distribution-driven simulation models. Using this model, the performance of the DBMIN algorithm is compared with that of the hot set algorithm and four more traditional buffer replacement algorithms.

Key Words. Buffer management, Database systems, Page replacement strategies, Hybrid simulation, Performance evaluations.

1. Introduction. In this paper we present a new algorithm, DBMIN, for managing the buffer pool of a relational database management system. DBMIN is based on a new model of relational query behavior, the query locality set model (QLSM). Like the hot set model [Sage 81], the QLSM has an advantage over stochastic models due to its ability to predict future reference behavior. However, the QLSM avoids the potential problems of the hot set model by separating the modeling of reference behavior from any particular buffer management algorithm. After introducing the QLSM and describing the DBMIN algorithm, the performance of the DBMIN algorithm in a mult-user environment is compared with that of the hot set algorithm and four more traditional buffer replacement algorithms.

A number of factors motivated this research. First, although Stonebraker [Stein 82] convincingly argued that conventional virtual memory page replacement algorithms (e.g., least recently used (LRU)) were generally not suitable for a

1 This research was partially supported by the Department of Energy under Contract No. DE-AC02-76CH00016 and by the National Science Foundation under grant MCB-8119155.
2 Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, USA.
3 Current Address: Microelectronics and Computer Technology Corporation, Austin, Texas, USA.
4 Received March 15, 1986; revised July 7, 1986. Communicated by Dale Skene.

Parts of this article have been reprinted with permission by the "Very Large Data Base Endowment."
Agenda

Buffer management basics
Query locality set model (QLSM)
DBMIN algorithm
Evaluation
Buffer Management Basics
A database management system (DBMS) manipulate data in memory
- Data on disk must be loaded to memory before processed

The unit of data movement is a page

Page replacement policy (what pages should stay in memory?)
- LRU (Lease recently used)
- Clock
- MRU (Most recently used)
- FIFO, Random, …
LRU Replacement

Replace the *least-recently used* (LRU) item in the buffer

Intuition: more recently used items will more likely to be used again in the future
LRU Replacement Example

Example: memory contains 4 buffers. LRU replacement policy

Memory

| 0 | 1 | 2 | 3 |

Incoming requests

0, 1, 2, 3, 0, 1, 2, 4, 0, 1, 2, 5, ...
LRU Replacement Example

Example: memory contains 4 buffers. LRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th>Incoming requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>0, 1, 2, 3, 0, 1, 2, 4, 0, 1, 2, 5, …</td>
</tr>
<tr>
<td>Page 1</td>
<td></td>
</tr>
<tr>
<td>Page 2</td>
<td></td>
</tr>
<tr>
<td>Page 3</td>
<td></td>
</tr>
</tbody>
</table>

Cold start misses: load pages 0–3 to memory

Disk
LRU Replacement Example

Example: memory contains 4 buffers. LRU replacement policy

Memory

| Page 0 | Page 1 | Page 2 | Page 3 |

Incoming requests

0, 1, 2, 3, 0, 1, 2, 4, 0, 1, 2, 5, …

Disk

Cache hits on pages 0–2
LRU Replacement Example

Example: memory contains 4 buffers. LRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>Page 1</td>
<td>Page 2</td>
<td>Page 3</td>
<td>Page 4</td>
</tr>
</tbody>
</table>

Incoming requests:
0, 1, 2, 3, 0, 1, 2, 4, 0, 1, 2, 5, …

Page 4 replaces page 3 in the buffer since page 3 is the least-recently used page.
LRU Replacement Example

Example: memory contains 4 buffers. LRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th>Incoming requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>0, 1, 2, 3, 0, 1, 2, 4, 0, 1, 2, 5, …</td>
</tr>
<tr>
<td>Page 1</td>
<td>Cache hits on pages 0–2</td>
</tr>
<tr>
<td>Page 2</td>
<td></td>
</tr>
<tr>
<td>Page 4</td>
<td></td>
</tr>
</tbody>
</table>
LRU Replacement Example

Example: memory contains 4 buffers. LRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th>Incoming requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>0, 1, 2, 3, 0, 1, 2, 4, 0, 1, 2, 5, …</td>
</tr>
<tr>
<td>Page 1</td>
<td>Page 5 replaces page 4 in the buffer since page 4 is the least-recently used page</td>
</tr>
<tr>
<td>Page 2</td>
<td></td>
</tr>
<tr>
<td>Page 5</td>
<td></td>
</tr>
<tr>
<td>Page 4</td>
<td></td>
</tr>
</tbody>
</table>

Disk
A Different Access Pattern

Example: memory contains 4 buffers. LRU replacement policy

Memory

Incoming requests
0, 1, 2, 3, 4, 0, 1, 2, 3, 4, …
A Different Access Pattern

Example: memory contains 4 buffers. LRU replacement policy

Memory

| Page 0 | Page 1 | Page 2 | Page 3 |

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...

Cold start misses: load pages 0—3 to memory
A Different Access Pattern

Example: memory contains 4 buffers. LRU replacement policy

Memory

<table>
<thead>
<tr>
<th>Page 4</th>
<th>Page 1</th>
<th>Page 2</th>
<th>Page 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...

Page 4 replaces page 0 since page 0 is the *least-recently used* page
**A Different Access Pattern**

Example: memory contains 4 buffers. LRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th>Page 4</th>
<th>Page 0</th>
<th>Page 2</th>
<th>Page 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...

Page 0 replaces page 1 since page 1 is the *least-recently used* page
A Different Access Pattern

Example: memory contains 4 buffers. LRU replacement policy

Memory

<table>
<thead>
<tr>
<th>Page 4</th>
<th>Page 0</th>
<th>Page 2</th>
<th>Page 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...

Page 0 replaces page 1 since page 1 is the **least-recently used** page

Each future access will replace the page that will be immediately accessed, and all accesses are misses.
MRU Replacement

Replace the **most-recently used** (LRU) item in the buffer

Intuition: avoid the cache thrashing problem in the previous example
MRU Replacement Example

Example: memory contains 4 buffers. MRU replacement policy

Memory

| Page 0 | Page 1 | Page 2 | Page 3 |

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, …
MRU Replacement Example

Example: memory contains 4 buffers. MRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th>Incoming requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...</td>
</tr>
<tr>
<td>Page 1</td>
<td></td>
</tr>
<tr>
<td>Page 2</td>
<td></td>
</tr>
<tr>
<td>Page 4</td>
<td></td>
</tr>
<tr>
<td>Page 3</td>
<td></td>
</tr>
</tbody>
</table>

Page 4 replaces page 3 since page 3 is the **most-recently used** page.
MRU Replacement Example

Example: memory contains 4 buffers. MRU replacement policy

Memory

| Page 0 | Page 1 | Page 2 | Page 4 |

Disk

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...

Cache hits on pages 0–2
MRU Replacement Example

Example: memory contains 4 buffers. MRU replacement policy

Memory

<table>
<thead>
<tr>
<th>Page 0</th>
<th>Page 1</th>
<th>Page 3</th>
<th>Page 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Incoming requests

0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...

Page 3 replaces page 2 since page 2 is the most-recently used page
MRU Replacement Example

Example: memory contains 4 buffers. MRU replacement policy

<table>
<thead>
<tr>
<th>Memory</th>
<th>Incoming requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>0, 1, 2, 3, 4, 0, 1, 2, 3, 4, …</td>
</tr>
<tr>
<td>Page 1</td>
<td></td>
</tr>
<tr>
<td>Page 3</td>
<td></td>
</tr>
<tr>
<td>Page 4</td>
<td></td>
</tr>
<tr>
<td>Page 2</td>
<td></td>
</tr>
</tbody>
</table>

Disk

Page 3 replaces page 2 since page 2 is the most-recently used page.

LRU: all accesses are misses
MRU: 25% of accesses are misses

Selection of replacement policy depends on the data access pattern.
Insight

The optimal buffer replacement and allocation policies depend on the data access pattern.

The data access pattern is relatively easy to predict in a DBMS compared to hardware or OS.
Query Locality Set Model (QLSM)
Query Locality Set Model

Observations

– DBMS supports a limited set of operations
– Data reference patterns are regular and predictable
– Complex reference patterns can be decomposed into simple patterns
Quantitative Set Model

Observations
- DBMS supports a limited set of operations
- Data reference patterns are regular and predictable
- Complex reference patterns can be decomposed into simple patterns

Reference pattern classification
- Sequential
- Random
- Hierarchical

Locality set: the appropriate buffer pool size for each query
QLSM – Sequential References

Straight sequential (SS): each page in a file accessed only once
- E.g., select on an unordered relation
- Locality set: one page
- Replacement policy: any
QLSM – Sequential References

Straight sequential (SS): each page in a file accessed only once
- E.g., select on an unordered relation
- Locality set: one page
- Replacement policy: any

Clustered sequential (CS): repeatedly read a “chunk” sequentially
- E.g., sort-merge join with duplicate join keys
- Locality set: size of largest cluster
- Replacement policy: LRU or FIFO (buffer size ≥ cluster size), MRU (otherwise)
QLSM – Sequential References

Straight sequential (SS): each page in a file accessed only once
- E.g., select on an unordered relation
- Locality set: one page
- Replacement policy: any

Clustered sequential (CS): repeatedly read a “chunk” sequentially
- E.g., sort-merge join with duplicate join keys
- Locality set: size of largest cluster
- Replacement policy: LRU or FIFO (buffer size ≥ cluster size), MRU (otherwise)

Looping Sequential (LS): repeatedly read something sequentially
- E.g. nested-loop join
- Locality set: size of the file being repeated scanned.
- Replacement policy: MRU
QLSM – Random References

Independent random (IR): truly random accesses
- E.g., index scan through a non-clustered (e.g., secondary) index
- Locality set: one page or $b$ pages ($b$ unique pages are accessed in total)
- Replacement: any
Independent random (IR): truly random accesses
- E.g., index scan through a non-clustered (e.g., secondary) index
- Locality set: one page or $b$ pages ($b$ unique pages are accessed in total)
- Replacement: any

Clustered random (CR): random accesses with some locality
- E.g., join between non-clustered, non-unique index as inner relation and clustered, non-unique outer relation
- Locality set: size of the largest cluster
- Replacement policy:
  LRU or FIFO (buffer size $\geq$ cluster size)
  MRU (otherwise)
QLSM – Hierarchical References

Straight hierarchical (SH): single traversal of the index
- Similar to SS

Hierarchical with straight sequential (H/SS): traversal followed by straight sequential on leaves
- Similar to SS

Hierarchical with clustered sequential (H/CS): traversal followed by clustered sequential on leaves
- Similar to CS

Looping hierarchical (LH): repeatedly traverse an index
- Example: index nested-loop join
- Locality set: first few layers in the B-tree
- Replacement: LIFO
### Summary of Reference Patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
<th>Locality set</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight sequential (SS)</td>
<td>File scan</td>
<td>1 page</td>
<td>any</td>
</tr>
<tr>
<td>Clustered sequential (CS)</td>
<td>Sort-merge join with duplicate keys</td>
<td>Cluster size</td>
<td>LRU/FIFO</td>
</tr>
<tr>
<td>Looped sequential (LS)</td>
<td>Nested-loop join</td>
<td>Size of scanned file</td>
<td>LRU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt; Size of scanned file</td>
<td>MRU</td>
</tr>
<tr>
<td>Independent random (IR)</td>
<td>non-clustered index scan</td>
<td>1 or b</td>
<td>any</td>
</tr>
<tr>
<td>Clustered random (CR)</td>
<td>Non-clustered, non-unique index as inner relation in a join</td>
<td>Same as CS</td>
<td></td>
</tr>
<tr>
<td>Straight hierarchical (SH)</td>
<td>Single index lookup</td>
<td></td>
<td>Same as SS</td>
</tr>
<tr>
<td>Hierarchical with straight sequential (H/SS)</td>
<td>Index lookup + scan</td>
<td></td>
<td>Same as SS</td>
</tr>
<tr>
<td>Hierarchical with clustered sequential (H/CS)</td>
<td>Index lookup + clustered scan</td>
<td></td>
<td>Same as CS</td>
</tr>
<tr>
<td>Looping hierarchical (LH)</td>
<td>Index nested-loop join</td>
<td>First few layers in the B-tree</td>
<td>LIFO</td>
</tr>
</tbody>
</table>
DBMIN algorithm
For each open file operation
  – Allocate a set of buffers (i.e., locality set)
  – Choose a replacement policy
  – Each open file instance has its own set of buffers
  – If two file instances access the same page, they share the page

Predicatively estimate locality set size by examining the query plan and database statistics

Admission control: a query is allowed to run if its locality sets fit in free frames
Other Buffer Management Algorithms
Simple Algorithms

Replacement discipline is applied globally to all the buffers in the system

- RAND
- FIFO (first-in, first-out)
- CLOCK
Sophisticated Algorithms

Replacement discipline is applied locally to each query or file instance

- **DBMIN**
- HOT (the hot set algorithm): always using LRU
- WS (the working set algorithm)
- Domain separation: LRU within each domain (e.g., an index level)
Evaluation

Except DBMIN and HOT, performance of all the other algorithms thrashes at high concurrency.

DBMIN outperforms HOT.
Q/A – Buffer Management

Modern relational DB buffer management policies are the same?
Relational vs. non-relational buffer management?
How to predict access patterns? Automated with ML?
Need multiple buffers to support multiple users?
Should buffer management be more concerned about heterogeneous memory latency?
How to support updates?
Group Discussion

In a conventional disk-based system, the bandwidth and latency gaps between DRAM and disks are large. Modern storage devices like non-volatile memory (NVM) have (1) bandwidth and latency close to DRAM and (2) byte-addressability. How do NVM devices change buffer management in a DBMS?

1. no page
2. scans directly go to DML.
Wisconsin DB Affiliates Workshop

Time: **Thursday** 8:30am - 4:00 pm.

Location: **Northwoods (Union South 3rd Floor)**

Workshop contents

- Research highlight talk from faculty member
- Research talks from PhD students
- Pitch talks from industry
- Poster session
- Discussion with industry partners including **AWS, Databricks, Google, MatrixOrigin, Microsoft, Oracle, Snowflake, TiDB**

Can also attend on zoom
Before Next Lecture

Submit review for

Viktor Leis, et al., *LeanStore: In-Memory Data Management Beyond Main Memory*. ICDE 2018