Data Structures and Algorithms for Big Databases

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
Big data problem

Data ingestion

Data indexing

Query processor

Queries + answers

Oy vey

365

42

???

???
For on-disk data, one sees funny tradeoffs in the speeds of data ingestion, query speed, and freshness of data.
Funny tradeoff in ingestion, querying, freshness

- “Select queries were slow until I added an index onto the timestamp field... Adding the index really helped our reporting, BUT now the inserts are taking forever.”
  - Comment on mysqlperformanceblog.com
Funny tradeoff in ingestion, querying, freshness

- “Select queries were slow until I added an index onto the timestamp field... Adding the index really helped our reporting, BUT now the inserts are taking forever.”
  - Comment on mysqlperformanceblog.com
- “I'm trying to create indexes on a table with 308 million rows. It took ~20 minutes to load the table but 10 days to build indexes on it.”
  - MySQL bug #9544
Funny tradeoff in ingestion, querying, freshness

- “Select queries were slow until I added an index onto the timestamp field... Adding the index really helped our reporting, BUT now the inserts are taking forever.”
  - Comment on mysqlperformanceblog.com
- “I'm trying to create indexes on a table with 308 million rows. It took ~20 minutes to load the table but 10 days to build indexes on it.”
  - MySQL bug #9544
- “They indexed their tables, and indexed them well, And lo, did the queries run quick! But that wasn’t the last of their troubles, to tell– Their insertions, like treacle, ran thick.”
  - Not from *Alice in Wonderland* by Lewis Carroll
This tutorial

- Better data structures significantly mitigate the insert/query/freshness tradeoff.
- These structures scale to much larger sizes while efficiently using the memory-hierarchy.
We don’t define Big Data in terms of TB, PB, EB.

By Big Data, we mean

- The data is too big to fit in main memory.
- We need data structures on the data.
- Words like “index” or “metadata” suggest that there are underlying data structures.
- These underlying data structures are also too big to fit in main memory.
In this tutorial we study the underlying data structures for managing big data.
But enough about databases...

... more about us.
A few years ago we started working together on I/O-efficient and cache-oblivious data structures.

Along the way, we started Tokutek to commercialize our research.
Tokutek sells open source, ACID compliant, implementations of MySQL and MongoDB.
Tokutek sells open source, ACID compliant, implementations of MySQL and MongoDB.

**TokuDB**
- Application
  - MySQL Database
    - SQL processing,
    - query optimization
  - libFT
- File System
- Disk/SSD

**TokuMX**
- Application
  - Standard MongoDB
    - drivers,
    - query language, and
    - data model
  - libFT
- File System
- Disk/SSD
Tokutek sells open source, ACID compliant, implementations of MySQL and MongoDB.

**TokuDB**
- Application
  - MySQL Database
    - SQL processing,
    - query optimization
  - libFT
- File System
- Disk/SSD

**TokuMX**
- Application
  - Standard MongoDB
    - drivers,
    - query language, and
    - data model
  - libFT
- File System
- Disk/SSD

libFT implements the persistent structures for storing data on disk.
Tokutek sells open source, ACID compliant, implementations of MySQL and MongoDB.

**TokuDB**
- **Application**
  - MySQL Database
    - SQL processing,
    - query optimization
- **libFT**
- **File System**
- **Disk/SSD**

**TokuMX**
- **Application**
  - Standard MongoDB
    - drivers,
    - query language, and
    - data model
- **libFT**
- **File System**

libFT implements the persistent structures for storing data on disk.

libFT provides a Berkeley DB API and can be used independently.
Our Mindset

• This tutorial is self contained.
• We want to teach.
• If something we say isn’t clear to you, please ask questions or ask us to clarify/repeat something.
• You should be comfortable using math.
• You should want to listen to data structures for an afternoon.
Topics and Outline for this Tutorial

I/O model.

Write-optimized data structures.

How write-optimized data structures can help file systems.

Cache-oblivious analysis.

Log-structured merge trees.

Indexing strategies.

Block-replacement algorithms.

Sorting Big Data.
Data Structures and Algorithms for Big Data
Module 1: I/O Model and Cache-Oblivious Analysis

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
• If we want to understand the performance of data structures within databases we need algorithmic models for understanding I/O.

• There’s a long history of memory-hierarchy models. Many are beautiful. Most have found little practical use.

• Two approaches are very powerful, the Disk Access Machine (DAM) model and cache-oblivious analysis.

• We’ll present the DAM model in this module to lay a foundation for the rest of the tutorial.

• Cache-oblivious analysis comes later in the tutorial.
I/O in the Disk Access Machine (DAM) Model

How computation works:

- Data is transferred in blocks between RAM and disk.
- The # of block transfers dominates the running time.

Goal: Minimize # of block transfers

- Performance bounds are parameterized by block size $B$, memory size $M$, data size $N$.

[Aggarwal+Vitter ’88]
Question: How many I/Os to scan an array of length $N$?
Question: How many I/Os to scan an array of length $N$?

Answer: $O(N/B)$ I/Os.
Question: How many I/Os to scan an array of length $N$?

Answer: $O(N/B)$ I/Os.

Example: Scanning an Array

$N$ blocks

$B$ blocks

scan touches $\leq N/B + 2$ blocks
Example: Searching in a B-tree

Question: How many I/Os for a point query or insert into a B-tree with \( N \) elements?

\[ O(\log_B N) \]
Example: Searching in a B-tree

Question: How many I/Os for a point query or insert into a B-tree with $N$ elements?

Answer: $O(\log_B N)$
Example: Searching in an Array

Question: How many I/Os to perform a binary search into an array of size $N$?
Question: How many I/Os to perform a binary search into an array of size $N$?

Answer: $O\left(\log_2 \frac{N}{B}\right) \approx O(\log_2 N)$
Example: Searching in an Array Versus B-tree

Moral: B-tree searching is a factor of $O(\log_2 B)$ faster than binary searching.

\[
O(\log_2 N) = O\left(\log_B N = \frac{\log_2 N}{\log_2 B}\right)
\]
The DAM model is simple and pretty good

The Disk Access Machine (DAM) model

- ignores CPU costs and
- assumes that all block accesses have the same cost.

Is that a good performance model?

- Far from perfect.
- But very powerful nonetheless.
- (We’ll discuss more later in the tutorial.)
Data Structures and Algorithms for Big Data
Module 2: Write-Optimized Data Structures

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
Important and universal problem.

Data ingestion

Data indexing

Query processor

Queries + answers

Oy vey

365

42

???
For on-disk data, one sees funny tradeoffs in the speeds of data ingestion, query speed, and freshness of data.
Funny tradeoff in ingestion, querying, freshness

- “Select queries were slow until I added an index onto the timestamp field... Adding the index really helped our reporting, BUT now the inserts are taking forever.”
  - Comment on mysqlperformanceblog.com

![Diagram showing data ingestion, data indexing, query processor, and queries + answers]
Funny tradeoff in ingestion, querying, freshness

- “Select queries were slow until I added an index onto the timestamp field... Adding the index really helped our reporting, BUT now the inserts are taking forever.”
  - Comment on mysqlperformanceblog.com
- “I'm trying to create indexes on a table with 308 million rows. It took ~20 minutes to load the table but 10 days to build indexes on it.”
  - MySQL bug #9544
Funny tradeoff in ingestion, querying, freshness

- “Select queries were slow until I added an index onto the timestamp field... Adding the index really helped our reporting, BUT now the inserts are taking forever.”
  - Comment on mysqlperformanceblog.com
- “I'm trying to create indexes on a table with 308 million rows. It took ~20 minutes to load the table but 10 days to build indexes on it.”
  - MySQL bug #9544
- “They indexed their tables, and indexed them well, And lo, did the queries run quick! But that wasn’t the last of their troubles, to tell– Their insertions, like treacle, ran thick.”
  - Not from Alice in Wonderland by Lewis Carroll
This module

- Write-optimized structures significantly mitigate the insert/query/freshness tradeoff.
- One can insert 10x-100x faster than B-trees while achieving similar point query performance.
How computation works:
• Data is transferred in blocks between RAM and disk.
• The number of block transfers dominates the running time.

Goal: Minimize # of block transfers
• Performance bounds are parameterized by block size $B$, memory size $M$, data size $N$. 

[Aggarwal+Vitter '88]
B-tree point queries: $O(\log_B N)$ I/Os.
Write-optimized data structures performance

Data structures: [O’Neil, Cheng, Gawlick, O’Neil 96], [Buchsbaum, Goldwasser, Venkatasubramanian, Westbrook 00], [Argel 03], [Graefe 03], [Brodal, Fagerberg 03], [Bender, Farach, Fineman, Fogel, Kuszmaul, Nelson’07], [Brodal, Demaine, Fineman, Iacono, Langerman, Munro 10], [Spillane, Shetty, Zadok, Archak, Dixit 11].

Systems: BigTable, Cassandra, H-Base, LevelDB, TokuDB.

<table>
<thead>
<tr>
<th>Insert/delete</th>
<th>B-tree</th>
<th>Some write-optimized structures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$O(\log_B N) = O\left(\frac{\log N}{\log B}\right)$</td>
<td>$O\left(\frac{\log N}{B}\right)$</td>
</tr>
</tbody>
</table>

- If $B=1024$, then insert speedup is $B/\log B \approx 100$.
- Hardware trends mean bigger $B$, bigger speedup.
- Less than 1 I/O per insert.
### Optimal Search-Insert Tradeoff

<table>
<thead>
<tr>
<th>Optimal tradeoff (function of $\varepsilon = 0...1$)</th>
<th>$\varepsilon = 1/2$</th>
<th>$\varepsilon = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>insert</strong></td>
<td>$O\left( \frac{\log_{1+B^\varepsilon} N}{B^{1-\varepsilon}} \right)$</td>
<td>$O\left( \frac{\log B N}{\sqrt{B}} \right)$</td>
</tr>
<tr>
<td><strong>point query</strong></td>
<td>$O \left( \log_{1+B^\varepsilon} N \right)$</td>
<td>$O \left( \log_B N \right)$</td>
</tr>
</tbody>
</table>

#### B-tree
- $O \left( \log_B N \right)$
- $O \left( \log_B N \right)$

10x-100x faster inserts

[10x-100x faster inserts]
Illustration of Optimal Tradeoff [Brodal, Fagerberg 03]

- Fast
- Slow

Point Queries

Inserts

- B-tree
- Optimal Curve
- Logging
Illustration of Optimal Tradeoff [Brodal, Fagerberg 03]

- Insertions improve by 10x-100x with almost no loss of point-query performance.

Target of opportunity:

B-tree

Optimal Curve

Logging

Point Queries

Slow

Fast

Inserts

Slow

Fast
Illustration of Optimal Tradeoff [Brodal, Fagerberg 03]

- **Target of opportunity**
  - B-tree
  - Insertions improve by 10x-100x with almost no loss of point-query performance

- **Optimal Curve**
- **Logging**

**Axes:**
- **Point Queries** (slow to fast)
- **Inserts** (slow to fast)
One way to Build Write-Optimized Structures

(other approaches later in tutorial)
A simple write-optimized structure

**O**(log \(N\)) queries and \(O((\log N)/B)\) inserts:

- A balanced binary tree with buffers of size \(B\)

**Inserts + deletes:**

- Send insert/delete messages down from the root and store them in buffers.
- When a buffer fills up, flush.
A simple write-optimized structure

O(log \(N\)) queries and O((log \(N\))/\(B\)) inserts:

- A balanced binary tree with buffers of size \(B\)

Inserts + deletes:

- Send insert/delete messages down from the root and store them in buffers.
- When a buffer fills up, flush.
A simple write-optimized structure

\(O(\log N)\) queries and \(O((\log N)/B)\) inserts:

- A balanced binary tree with buffers of size \(B\)

Inserts + deletes:

- Send insert/delete messages down from the root and store them in buffers.
- When a buffer fills up, flush.
A simple write-optimized structure

\[ O(\log N) \text{ queries and } O((\log N)/B) \text{ inserts:} \]

- A balanced binary tree with buffers of size \( B \)

\[ \text{Inserts + deletes:} \]

- Send insert/delete messages down from the root and store them in buffers.
- When a buffer fills up, flush.
A simple write-optimized structure

\[ O(\log N) \text{ queries and } O((\log N)/B) \text{ inserts:} \]

- A balanced binary tree with buffers of size \( B \)

\[ \text{Inserts + deletes:} \]

- Send insert/delete messages down from the root and store them in buffers.
- When a buffer fills up, flush.
A simple write-optimized structure

$O(\log N)$ queries and $O((\log N)/B)$ inserts:

- A balanced binary tree with buffers of size $B$

Inserts + deletes:

- Send insert/delete messages down from the root and store them in buffers.
- When a buffer fills up, flush.
Analysis of writes

An insert/delete costs amortized $O((\log N)/B)$ per insert or delete

- A buffer flush costs $O(1)$ & sends $B$ elements down one level.
- It costs $O(1/B)$ to send element down one level of the tree.
- There are $O(\log N)$ levels in a tree.
Difficulty of Key Accesses
Difficulty of Key Accesses
To search:

- examine each buffer along a single root-to-leaf path.
- This costs $O(\log N)$. 
Obtaining optimal point queries + very fast inserts

Point queries cost $O(\log_{\sqrt{B}} N) = O(\log_B N)$
- This is the tree height.

Inserts cost $O((\log_B N) / \sqrt{B})$
- Each flush cost $O(1)$ I/Os and flushes $\sqrt{B}$ elements.
What the world looks like

Insert/point query asymmetry

• Inserts can be fast: >50K high-entropy writes/sec/disk.
• Point queries are necessarily slow: <200 high-entropy reads/sec/disk.

*We are used to reads and writes having about the same cost, but writing is easier than reading.*
The right read-optimization is write-optimization

The right index makes queries run fast.

- Write-optimized structures maintain indexes efficiently.
The right read-optimization is write-optimization.

The right index makes queries run fast.

- Write-optimized structures maintain indexes efficiently.

*Fast writing is a currency we use to accelerate queries. Better indexing means faster queries.*
The right read-optimization is write-optimization

Add selective indexes.

(We can now afford to maintain them.)
The right read-optimization is write-optimization

Add selective indexes.
(We can now afford to maintain them.)

Write-optimized structures can significantly mitigate the insert/query/freshness tradeoff.
Implementation Issues
Optimal read-write tradeoff: Easy

Full featured: Hard

- Variable-sized rows
- Concurrency-control mechanisms
- Multithreading
- Transactions, logging, ACID-compliant crash recovery
- Optimizations for the special cases of sequential inserts and bulk loads
- Compression
- Backup
Some inserts/deletes have hidden searches.

Example:

- return error when a duplicate key is inserted.
- return # elements removed on a delete.

These “cryptosearches” throttle insertions down to the performance of B-trees.
Uniqueness checking has a hidden search:

\[
\text{If Search(key) == True} \\
\quad \text{Return Error;}
\]

\[
\text{Else} \\
\quad \text{Fast_Insert(key,value)};
\]

In a B-tree uniqueness checking comes for free

- On insert, you fetch a leaf.
- Checking if key exists is no biggie.
Cryptosearches in uniqueness checking

Uniqueness checking has a hidden search:

\[
\text{If Search(key) == True}
\]

\[
\text{Return Error;}
\]

\[
\text{Else}
\]

\[
\text{Fast_Insert(key, value)};
\]

In a write-optimized structure, that crypto-search can throttle performance

- Insertion messages are injected.
- These eventually get to “bottom” of structure.
- Insertion w/Uniqueness Checking 100x slower.
- Bloom filters, Cascade Filters, etc help.

[Bender, Farach-Colton, Johnson, Kraner, Kuszmaul, Medjedovic, Montes, Shetty, Spillane, Zadok 12]
One implementation of pessimistic locking: *maintain locks in leaves*

To insert a new row $v$, determine whether there is already a lock on $v$ at a leaf.

This is also a cryptosearch.
Performance
iiBench Insertion Benchmark

Write performance on large data

iiBench Benchmark (throughput)
TokuMX vs. MongoDB
(higher is better)

MongoDB

MySQL
TokuMX runs fast because it uses less I/O

iiBench Benchmark (Average Write IOPs)
TokuMX vs. MongoDB
(lower is better)

100M inserts into a collection with 3 secondary indexes
Compression (up to 25x)
TokuDB on rotating disk beats InnoDB on SSD.
Write-optimization Can Help Schema Changes

**InnoDB**

**Index Creation**

00:31:34

**TokuDB**

**Hot Indexing**

00:00:02

**InnoDB**

**Column Addition**

17:44:41

**TokuDB**

**Hot Column Addition**

00:00:03

**Platform**: CentOS 5.7; 2x Xeon L5520; 72GB RAM; 8x 300GB 10k SAS in RAID10.

Hot Schema – Schema Changes in Seconds, not Hours

TokuDB v5.0 introduced Hot Column Addition (HCAD). You can add or delete columns from an existing table with minimal downtime — just the time for MySQL itself to close and reopen the table. The total downtime is seconds to minutes. We detailed an experiment that showed this in this blog.

TokuDB v5.0 also introduced Hot Indexing. You can add an index to an existing table with minimal downtime. The total downtime is seconds to a few minutes, because when the index is finished being built, MySQL closes and reopens the table. This means that the downtime occurs not when the command is issued, but later on. Still, it is quite minimal, as we showed in this blog.

**Platform**: CentOS 5.5; 2x Xeon E5310; 4GB RAM; 4x 1TB 7.2k SATA in RAID0.

**Appendix – Software Configuration Details**
Scaling into the Future
Write-optimization going forward


- log high-entropy data sequentially versus index data in B-tree.

<table>
<thead>
<tr>
<th>Year</th>
<th>Size</th>
<th>Bandwidth</th>
<th>Access Time</th>
<th>Time to log data on disk</th>
<th>Time to fill disk using a B-tree (row size 1K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1973</td>
<td>35MB</td>
<td>835KB/s</td>
<td>25ms</td>
<td>39s</td>
<td>975s</td>
</tr>
<tr>
<td>2010</td>
<td>3TB</td>
<td>150MB/s</td>
<td>10ms</td>
<td>5.5h</td>
<td>347d</td>
</tr>
<tr>
<td>2022</td>
<td>220TB</td>
<td>1.05GB/s</td>
<td>10ms</td>
<td>2.4d</td>
<td>70y</td>
</tr>
</tbody>
</table>

Better data structures may be a luxury now, but they will be essential by the decade’s end.

- log high-entropy data sequentially versus index data in B-tree.

<table>
<thead>
<tr>
<th>Year</th>
<th>Size</th>
<th>Bandwidth</th>
<th>Access Time</th>
<th>Time to log data on disk</th>
<th>Time to fill disk using a B-tree (row size 1K)</th>
<th>Time to fill using Fractal tree* (row size 1K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1973</td>
<td>35MB</td>
<td>835KB/s</td>
<td>25ms</td>
<td>39s</td>
<td>975s</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>3TB</td>
<td>150MB/s</td>
<td>10ms</td>
<td>5.5h</td>
<td>347d</td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>220TB</td>
<td>1.05GB/s</td>
<td>10ms</td>
<td>2.4d</td>
<td>70y</td>
<td></td>
</tr>
</tbody>
</table>

Better data structures may be a luxury now, but they will be essential by the decade’s end.

* Projected times for fully multi-threaded version

- log high-entropy data sequentially versus index data in B-tree.

<table>
<thead>
<tr>
<th>Year</th>
<th>Size</th>
<th>Bandwidth</th>
<th>Access Time</th>
<th>Time to log data on disk</th>
<th>Time to fill disk using a B-tree (row size 1K)</th>
<th>Time to fill using Fractal tree* (row size 1K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1973</td>
<td>35MB</td>
<td>835KB/s</td>
<td>25ms</td>
<td>39s</td>
<td>975s</td>
<td>200s</td>
</tr>
<tr>
<td>2010</td>
<td>3TB</td>
<td>150MB/s</td>
<td>10ms</td>
<td>5.5h</td>
<td>347d</td>
<td>36h</td>
</tr>
<tr>
<td>2022</td>
<td>220TB</td>
<td>1.05GB/s</td>
<td>10ms</td>
<td>2.4d</td>
<td>70y</td>
<td>23.3d</td>
</tr>
</tbody>
</table>

Better data structures may be a luxury now, but they will be essential by the decade’s end.

* Projected times for fully multi-threaded version
Write-optimization can solve many problems.

- There is a provable point-query insert tradeoff. We can insert 10x-100x faster without hurting point queries.
- We can avoid much of the funny tradeoff between data ingestion, freshness, and query speed.
- We can avoid tuning knobs.
Data Structures and Algorithms for Big Data
Module 3: (Case Study)
TokuFS--How to Make a Write-Optimized File System

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
Algorithms for Big Data apply to all storage systems, not just databases.

Some big-data users store use a file system.

The problem with Big Data is Microdata...
HEC FSIO Grand Challenges

Store 1 trillion files

Create tens of thousands of files per second

Traverse directory hierarchies fast (ls -R)

B-trees would require at least hundreds of disk drives.
**TokuFS**

[Esmer, Bender, Farach-Colton, Kusmaul HotStorage12]

- A file-system prototype
- >20K file creates/sec
- very fast `ls -R`
- HEC grand challenges on a cheap disk (except 1 trillion files)
**TokuFS**

A file-system prototype

- >20K file creates/sec
- **very fast** `ls -R`
- HEC grand challenges on a cheap disk (except 1 trillion files)

- **TokuFS offers orders-of-magnitude speedup on microdata workloads.**
  - Aggregates microwrites while indexing.
  - So it can be faster than the underlying file system.
Big speedups on microwrites

We ran microdata-intensive benchmarks

- Compared TokuFS to ext4, XFS, Btrfs, ZFS.
- Stressed metadata and file data.
- Used commodity hardware:
  - 2 core AMD, 4GB RAM
  - Single 7200 RPM disk
  - Simple, cheap setup. No hardware tricks.
- In all tests, we observed orders of magnitude speed up.
Create 2 million 200-byte files in a directory tree

![Randomized small file creation](chart.png)

- **Ext4**: Blue bars
- **XFS**: Orange bars
- **Btrfs**: Red bars
- **ZFS**: Green bars
- **TokuFS**: Yellow bars

**Chart Description:**
- **Y-axis**: files/sec
- **X-axis**: Filesystem
- **Legend**:
  - 1 thread
  - 4 threads
  - 8 threads
Create 2 million 200-byte files in a directory tree

Randomized small file creation

- 1 thread
- 4 threads
- 8 threads

files/sec

ext4  xfs  btrfs  zfs  tokufs

filesystem

Log scale
Faster on metadata scan

Recursively scan directory tree for metadata

- Use the same 2 million files created before.
- Start on a cold cache to measure disk I/O efficiency

![Recursive metadata scan graph]
Create one million empty files in a directory

- Create files with random names, then read them back.
- Tests how well a single directory scales.
Faster on microwrites in a big file

Randomly write out a file in small, unaligned pieces
TokuFS
Implementation
TokuFS employs two indexes

**Metadata index:**

- The metadata index maps pathname to file metadata.
  - /home/esmet mode, file size, access times, ...
  - /home/esmet/tokufs.c mode, file size, access times, ...

**Data index:**

- The data index maps pathname, blocknum to bytes.
  - /home/esmet/tokufs.c, 0 [ block of bytes ]
  - /home/esmet/tokufs.c, 1 [ block of bytes ]

- **Block size is a compile-time constant: 512.**
  - good performance on small files, moderate on large files
Common queries exhibit locality

**Metadata index keys: full path as string**
- All the children of a directory are contiguous in the index
- Reading a directory is simple and fast

**Data block index keys: 【full path, blocknum】**
- So all the blocks for a file are contiguous in the index
- Reading a file is simple and fast
TokuFS compresses indexes

Reduces overhead from full path keys
  • Pathnames are highly “prefix redundant”
  • They compress very, very well in practice

Reduces overhead from zero-valued padding
  • Uninitialized bytes in a block are set to zero
  • Good portions of the metadata struct are set to zero

Compression between 7-15x on real data
  • For example, a full MySQL source tree
TokuFS is a prototype, but fully functional.

- Implements files, directories, metadata, etc.
- Interfaces with applications via shared library, header.

We wrote a FUSE implementation, too.

- FUSE lets you implement filesystems in user space.
- But there’s overhead, so performance isn’t optimal.
- The best way to run is through our POSIX-like file API.
Microdata is the Problem
Data Structures and Algorithms for Big Data
Module 4: Cache-Oblivious Analysis

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
External-memory model:

- Time bounds are parameterized by $B$, $M$, $N$.
- Goal: Minimize # of block transfers $\approx$ time.
External-memory model:

- Time bounds are parameterized by $B$, $M$, $N$.
- Goal: Minimize number of block transfers $\approx$ time.

Beautiful restriction:

- Parameters $B$, $M$ are unknown to the algorithm or coder.
Cache-Oblivious (CO) Algorithms [Frigo, Leiserson, Prokop, Ramachandran ’99]

External-memory model:
• Time bounds are parameterized by $B, M, N$.
• Goal: Minimize # of block transfers $\approx$ time.

Beautiful restriction:
• Parameters $B, M$ are unknown to the algorithm or coder.

An optimal CO algorithm is universal for all $B, M, N$. 
Overview of Module

- Cache-oblivious definition
- **Cache-oblivious B-tree**
- Cache-oblivious performance advantages
- Cache-oblivious write-optimized data structure (COLA)
- Cache-adaptive algorithms
Traditional B-trees aren’t cache-oblivious

There do exist cache-oblivious B-trees.

- We can still achieve $O(\log_B N)$ I/Os per operation, even without parameterizing by $B$ or $M$.

[Bender, Demaine, Farach-Colton ’00] [Bender, Duan, Iacono, Wu ’02]
[Brodal, Fagerberg, Jacob ’02]
• Technique: divide & conquer (Van Emde Boas layout)

Static cache-oblivious B-Tree (no inserts) [Prokop 99]

Lay out each subtree

Repeat recursively.
Analysis of vEB Layout

Conceptually stop recursion when recursive subtrees $\leq B \Rightarrow$ subtree fits in $\leq 2$ blocks.

$\lg N \geq \frac{T}{\lg B}$

$\Rightarrow A$ search visits $\leq \frac{\lg N}{\lg B} = 2\log_B N$ subtrees

$\Rightarrow \leq 4\log_B N$ memory transfers
Dynamic Cache-Oblivious B-trees

We won’t describe how to dynamize....

After all, the cache-oblivious dynamic B-tree isn’t write-optimized.

We believe that write-optimized data structures win out over B-trees (even cache-oblivious ones) in the majority of cases.
Overview of Module

Cache-oblivious definition

Example: cache-oblivious B-tree

**Cache-oblivious performance advantages**

Cache-oblivious write-optimized data structure (COLA)

Cache-adaptive algorithms
The DAM model is a simplification.

Tracks get prefetched into the disk cache, which holds ~100 tracks.

Fixed-size blocks are fetched.

Disks are organized into tracks of different sizes.
The DAM model is a simplification

2kB or 4kB is too small for the model.
- B-tree nodes in Berkeley DB & InnoDB have this size.
- Issue: sequential block accesses run 10x faster than random block accesses, which doesn’t fit the model.

There is no single best block size.
- The best node size for a B-tree depends on the operation (insert/delete/point query).
## Time for 1000 Random B-tree Searches

[Bender, Farach-Colton, Kuszmaul ’06]

<table>
<thead>
<tr>
<th>$B$</th>
<th>Small</th>
<th>Big</th>
</tr>
</thead>
<tbody>
<tr>
<td>4K</td>
<td>17.3ms</td>
<td>22.4ms</td>
</tr>
<tr>
<td>16K</td>
<td>13.9ms</td>
<td>22.1ms</td>
</tr>
<tr>
<td>32K</td>
<td>11.9ms</td>
<td>17.4ms</td>
</tr>
<tr>
<td>64K</td>
<td>12.9ms</td>
<td>17.6ms</td>
</tr>
<tr>
<td>128K</td>
<td>13.2ms</td>
<td>16.5ms</td>
</tr>
<tr>
<td>256K</td>
<td>18.5ms</td>
<td>14.4ms</td>
</tr>
<tr>
<td>512K</td>
<td>16.7ms</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Big</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO B-tree</td>
<td>12.3ms</td>
<td>13.8ms</td>
</tr>
</tbody>
</table>

There’s no best block size.
The optimal block size for inserts is very different.
Cache-Oblivious Analysis

- Cache-oblivious algorithms work for all $B$ and $M$...
- ... and all levels of a multi-level hierarchy.

*It’s better to optimize approximately for all $B$, $M$ than to pick the best $B$ and $M$.*

[Frigo, Leiserson, Prokop, Ramachandran ’99]
Cache-oblivious definition

Example: cache-oblivious B-tree

Cache-oblivious performance advantages

**Cache-oblivious write-optimized data structure (COLA)**

- You can even make write-optimized data structures cache-oblivious

  [Bender, Farach-Colton, Fineman, Fogel, Kuszmaul, Nelson, SPAA 07]
  [Brodal, Demaine, Fineman, Iacono, Langerman, Munro, SODA 10]

Cache-adaptive algorithms
Recall optimal search-insert tradeoff [Brodal, Fagerberg 03]

<table>
<thead>
<tr>
<th>Optimal tradeoff (function of $\varepsilon=0...1$)</th>
<th>Insert</th>
<th>Point query</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O \left( \frac{\log_{1+B^\varepsilon} N}{B^{1-\varepsilon}} \right)$</td>
<td>$O (\log_{1+B^\varepsilon} N)$</td>
<td></td>
</tr>
</tbody>
</table>

- **B-tree** ($\varepsilon=1$)
  - Insert: $O (\log_B N)$
  - Point query: $O (\log_B N)$

- $\varepsilon=1/2$
  - Insert: $O \left( \frac{\log_B N}{\sqrt{B}} \right)$
  - Point query: $O (\log_B N)$

- $\varepsilon=0$
  - Insert: $O \left( \frac{\log N}{B} \right)$
  - Point query: $O (\log N)$

10x-100x faster inserts

We give a cache-oblivious solution for $\varepsilon=0$. 
Simplified CO write-optimized structure (COLA)

- Sorted arrays of exponentially increasing size.
- Arrays are completely full or completely empty (depends on the bit representation of # of elmts).
- Insert into the smallest array.
- Merge arrays to make room.

O((logN)/B) insert cost & O(log²N) search cost
Simplified CO write-optimized structure (COLA)
Insert Cost:

- cost to flush level of size $X = O(X/B)$
- cost per element to flush level = $O(1/B)$
- max # of times each element is flushed = $\log N$
- insert cost = $O((\log N)/B)$ amortized memory transfers

Search Cost

- Binary search at each level
- $\log(N/B) + \log(N/B) - 1 + \log(N/B) - 2 + \ldots + 2 + 1 = O(\log^2(N/B))$
Cache-oblivious write-optimized structure (COLA)

$O\left(\frac{\log N}{B}\right)$ insert cost & $O(\log N)$ search cost

- Some redundancy of elements between levels
- Arrays can be partially full
- Horizontal and vertical pointers to redundant elements
- (Fractional Cascading)
Overview of Module

Cache-oblivious definition
Example: cache-oblivious B-tree
Cache-oblivious performance advantages
Cache-oblivious write-optimized data structure (COLA)
Cache-adaptive algorithms
Cache-oblivious algorithms are universal algorithms. They are platform independent.
Cache-oblivious algorithms adapt to changing RAM.
Cache-oblivious algorithms adapt to changing RAM.

External-memory model:
- Time bounds are parameterized by $B$, $M$, $N$.
- Goal: Minimize # of block transfers = time.

Beautiful restriction:
- Parameters $B$, $M$ are unknown to the algorithm or coder.
- An optimal CO algorithm is universal for all $B$, $M$, $N$. 

Interesting
Cache-Oblivious Algorithms

External-memory model:
- Time bounds are parameterized by $B$, $M$, $N$.
- Goal: Minimize # of block transfers = time.

Beautiful restriction:
- Parameters $B$, $M$ are unknown to the algorithm or coder.
- An optimal CO algorithm is universal for all $B$, $M$, $N$.

Cache-oblivious algorithms adapt to changing RAM.

(I don’t know what the heck I’m talking about.)
We had an activity of presenting an abstract for a fictional paper that we wanted to write.

Cache-oblivious algorithms adapt to changing RAM.

We have this concern at --- (social media company).

We have this concern at ---- (database company).
So I proved some theorems

Theorem: Some (but not all) cache-oblivious algorithms adapt to changing sizes of RAM.
Some CO algorithms adapt when RAM changes

Some cache-oblivious algorithms run optimally even when the RAM changes arbitrarily over time.

- Sorting
- Problems with a special recursive structure (matrix multiplication, transpose, Gaussian elimination, all-pairs shortest paths)
In the cache-oblivious model, $B$ and $M$ are unknown to the coder.

(Of course, we still use $B$ and $M$ in proofs.)

It’s remarkable how many common I/O-efficient data structures have cache-oblivious alternatives.

Sometimes it’s better to optimize approximately for all $B$ and $M$ instead of picking the best $B$ and $M$. 
Data Structures and Algorithms for Big Data
Module 5: Log Structured Merge Trees

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
Log Structured Merge Trees

Log structured merge trees are write-optimized data structures developed in the 90s.

Over the past 5 years, LSM trees have become popular (for good reason).

Accumulo, Bigtable, bLSM, Cassandra, HBase, Hypertable, LevelDB are LSM trees (or borrow ideas).

http://nosql-database.org lists 122 NoSQL databases. Many of them are LSM trees.
Recall Optimal Search-Insert Tradeoff

[Brodal, Fagerberg 03]

<table>
<thead>
<tr>
<th>insert</th>
<th>point query</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O \left( \frac{\log_{1+B^\varepsilon} N}{B^{1-\varepsilon}} \right) )</td>
<td>( O \left( \log_{1+B^\varepsilon} N \right) )</td>
</tr>
</tbody>
</table>

Optimal tradeoff
(function of \( \varepsilon = 0 \ldots 1 \))

LSM trees don’t lie on the optimal search-insert tradeoff curve.

But they’re not far off.

We’ll show how to move them back onto the optimal curve.
An LSM tree is a cascade of B-trees. Each tree $T_j$ has a target size $|T_j|$. The target sizes are exponentially increasing. Typically, target size $|T_{j+1}| = 10 |T_j|$. 

$T_0$ $T_1$ $T_2$ $T_3$ $T_4$
LSM Tree Operations

Point queries:

\[ T_0 \quad T_1 \quad T_2 \quad T_3 \quad T_4 \]
LSM Tree Operations

Point queries:

Range queries:
** LSM Tree Operations **

**Insertions:**

- Always insert element into the smallest B-tree $T_0$.

- When a B-tree $T_j$ fills up, flush into $T_{j+1}$.
Deletes are like inserts:

- Instead of deleting an element directly, insert tombstones.
- A tombstone knocks out a “real” element when it lands in the same tree.
An LSM Tree is an example of a “static-to-dynamic” transformation [Bentley, Saxe ’80].

- An LSM tree can be built out of static B-trees.
- When $T_3$ flushes into $T_4$, $T_4$ is rebuilt from scratch.
Let's analyze LSM trees.
Recall: Searching in an Array Versus B-tree

Recall the cost of searching in an array versus a B-tree.

\[ O(\log_B N) = O\left(\frac{\log_2 N}{\log_2 B}\right) \]
Recall the cost of searching in an array versus a B-tree.

\[
O\left(\log_2 \frac{N}{B}\right) \approx O\left(\log_2 N\right)
\]

\[
O(\log_B N) = O\left(\frac{\log_2 N}{\log_2 B}\right)
\]
Analysis of point queries

Search cost:

$$\log_B N + \log_B N/2 + \log_B N/4 + \cdots + \log_B B$$

$$= \frac{1}{\log B} (\log N + \log N - 1 + \log N - 2 + \log N - 3 + \cdots + 1)$$

$$= O(\log N \log_B N)$$
The cost to flush a tree $T_j$ of size $X$ is $O(X/B)$.

- Flushing and rebuilding a tree is just a linear scan.

The cost per element to flush $T_j$ is $O(1/B)$.

The # times each element is moved is $\leq \log N$.

The insert cost is $O((\log N)/B)$ amortized memory transfers.

A flush costs $O(1/B)$ per element.

$T_j$ has size $X$.

$T_{j+1}$ has size $\Theta(X)$. 

## Samples from LSM Tradeoff Curve

<table>
<thead>
<tr>
<th>tradeoff</th>
<th>insert</th>
<th>point query</th>
</tr>
</thead>
<tbody>
<tr>
<td>(function of $\varepsilon$)</td>
<td>$O \left( \frac{\log_{1+B^\varepsilon} N}{B^{1-\varepsilon}} \right)$</td>
<td>$O \left( (\log_B N)(\log_{1+B^\varepsilon} N) \right)$</td>
</tr>
</tbody>
</table>

### sizes grow by $B$
- ($\varepsilon=1$)
  - $O (\log_B N)$
  - $O (\log_B N)$

### sizes grow by $B^{1/2}$
- ($\varepsilon=1/2$)
  - $O \left( \frac{\log_B N}{\sqrt{B}} \right)$
  - $O \left( (\log_B N)(\log_B N) \right)$

### sizes double
- ($\varepsilon=0$)
  - $O \left( \frac{\log N}{B} \right)$
  - $O \left( (\log_B N)(\log N) \right)$
How to improve LSM-tree point queries?

Looking in all those trees is expensive, but can be improved by

- caching,
- Bloom filters, and
- fractional cascading.
When the cache is warm, small trees are cached.
Bloom filters in LSM trees

Bloom filters can avoid point queries for elements that are not in a particular B-tree.
Fractional cascading reduces the cost in each tree

Instead of avoiding searches in trees, we can use a technique called *fractional cascading* to reduce the cost of searching each B-tree to $O(1)$.

Idea: We’re looking for a key, and we already know where it should have been in $T_3$, try to use that information to search $T_4$. 

$$\begin{align*}
    &T_0 &T_1 &T_2 &T_3 &T_4 \\
\end{align*}$$
Searching one tree helps in the next

Looking up $c$, in $T_i$ we know it’s between $b$, and $e$.

$T_i$

```
  b e v w
```

$T_{i+1}$

```
a c d f h i j k m n p q t u y z
```

Showing only the bottom level of each B-tree.
If we add *forwarding pointers* to the first tree, we can jump straight to the node in the second tree, to find c.
Remove redundant forwarding pointers

We need only one forwarding pointer for each block in the next tree. Remove the redundant ones.
We need a forwarding pointer for every block in the next tree, even if there are no corresponding pointers in this tree. Add **ghosts**.
LSM tree + forward + ghost = fast queries

With forward pointers and ghosts, LSM trees require only one I/O per tree, and point queries cost only 

\[ O(\log_R N) \].

[Bender, Farach-Colton, Fineman, Fogel, Kuszmaul, Nelson 07]
This data structure no longer uses the internal nodes of the B-trees, and each of the trees can be implemented by an array.

[Bender, Farach-Colton, Fineman, Fogel, Kuszmaul, Nelson 07]
Data Structures and Algorithms for Big Data
Module 6: What to Index

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
This module explores indexing.

Traditionally, (with B-trees), indexing improves queries, but cripples insertions.

But now we know that maintaining indexes is cheap. So what should we index?
An Indexing Testimonial

Add selective indexes.

This is a graph from a real user, who added some indexes, and reduced the I/O load on their server. (They couldn’t maintain the indexes with B-trees.)
To understand what to index, we need to get on the same page for what an index is.
Row

- Key, value pair
- key = a, value = b, c

Index

- Ordering of rows by key (dictionary)
- Used to make queries fast

Table

- Set of indexes

```
create table foo (a int, b int, c int, primary key(a));
```
Dictionary API: maintain a set $S$ subject to

- **insert($x$):** $S \leftarrow S \cup \{x\}$
- **delete($x$):** $S \leftarrow S - \{x\}$
- **search($x$):** is $x \in S$?
- **successor($x$):** return $\min y > x$ s.t. $y \in S$
- **predecessor($y$):** return $\max y < x$ s.t. $y \in S$

We assume that these operations perform as well as a B-tree. For example, the successor operation usually doesn’t require an I/O.
A table is a set of indexes with operations:

- Add index: `add key (f_1, f_2, ...)`;
- Drop index: `drop key (f_1, f_2, ...)`;
- Add column: adds a field to primary key value.
- Remove column: removes a field and drops all indexes where field is part of key.
- Change field type
- ...

Subject to index correctness constraints.

*We want table operations to be fast too.*
Next: how to use indexes to improve queries.
Indexes provide query performance

1. Indexes can reduce the amount of retrieved data.
   - Less bandwidth, less processing, ...

2. Indexes can improve locality.
   - Not all data access cost is the same
   - Sequential access is MUCH faster than random access

3. Indexes can presort data.
   - GROUP BY and ORDER BY queries do post-retrieval work
   - Indexing can help get rid of this work
Indexes provide query performance

1. Indexes can reduce the amount of retrieved data.
   • Less bandwidth, less processing, ...

2. Indexes can improve locality.
   • Not all data access cost is the same
   • Sequential access is MUCH faster than random access

3. Indexes can presort data.
   • GROUP BY and ORDER BY queries do post-retrieval work
   • Indexing can help get rid of this work
An index can select needed rows

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

count (*) where a<120;
An index can select needed rows

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

\[
\text{count (*) where } a < 120;\]
No good index means slow table scans

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

`count (*) where b>50 and b<100;`
No good index means slow table scans

count (*) where b>50 and b<100;
You can add an index

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>23</td>
<td>206</td>
</tr>
<tr>
<td>43</td>
<td>412</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>92</td>
<td>101</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>

```sql
alter table foo add key(b);
```
A selective index speeds up queries

```
count (*) where b > 50 and b < 100;
```
A selective index speeds up queries

count (*) where b>50 and b<100;
Selective indexes can still be slow

```
sum(c) where b>50;
```
Selective indexes can still be slow

**sum(c) where b>50;**

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>23</td>
<td>206</td>
</tr>
<tr>
<td>43</td>
<td>412</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>92</td>
<td>101</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>
Selective indexes can still be slow

$$\text{sum}(c) \text{ where } b > 50;$$
Selective indexes can still be slow

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>23</td>
<td>206</td>
</tr>
<tr>
<td>43</td>
<td>412</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>56</th>
<th>156</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>92</td>
<td>101</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>

Fetching info for summing c: slow
Selecting on b: fast

\text{sum(c) where } b>50;
Selective indexes can still be slow

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>23</td>
<td>206</td>
</tr>
<tr>
<td>43</td>
<td>412</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>92</td>
<td>101</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>

**sum(c) where b>50;**
Selective indexes can still be slow

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>23</td>
<td>206</td>
</tr>
<tr>
<td>43</td>
<td>412</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>92</td>
<td>101</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>

```
sum(c) where b>50;
```
Selective indexes can still be slow.

Poor data locality

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
</tr>
<tr>
<td>23</td>
<td>206</td>
</tr>
<tr>
<td>43</td>
<td>412</td>
</tr>
<tr>
<td>56</td>
<td>156</td>
</tr>
<tr>
<td>56</td>
<td>256</td>
</tr>
<tr>
<td>92</td>
<td>101</td>
</tr>
<tr>
<td>202</td>
<td>198</td>
</tr>
</tbody>
</table>

sum(c) where b>50;

Selecting on b: fast

Fetching info for summing c: slow

<table>
<thead>
<tr>
<th></th>
<th>56</th>
<th>156</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>202</td>
<td>198</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>156</th>
<th>56</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
<td></td>
</tr>
</tbody>
</table>

202 198
Selective indexes can still be slow

<table>
<thead>
<tr>
<th>Poor data locality</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sum(c) where b&gt;50;</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
</tr>
</tbody>
</table>
Indexes provide query performance

1. Indexes can reduce the amount of retrieved data.
   • Less bandwidth, less processing, ...

2. Indexes can improve locality.
   • Not all data access cost is the same
   • Sequential access is MUCH faster than random access

3. Indexes can presort data.
   • GROUP BY and ORDER BY queries do post-retrieval work
   • Indexing can help get rid of this work
Covering indexes speed up queries

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>b,c</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>5,45</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>6,2</td>
<td>165</td>
<td></td>
</tr>
<tr>
<td>23,252</td>
<td>206</td>
<td></td>
</tr>
<tr>
<td>43,45</td>
<td>412</td>
<td></td>
</tr>
<tr>
<td>56,2</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>56,45</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>92,2</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>202,56</td>
<td>198</td>
<td></td>
</tr>
</tbody>
</table>

```sql
alter table foo add key(b,c);
sum(c) where b>50;
```
Covering indexes speed up queries

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>101</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>156</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>198</td>
<td>202</td>
<td>56</td>
</tr>
<tr>
<td>206</td>
<td>23</td>
<td>252</td>
</tr>
<tr>
<td>256</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>412</td>
<td>43</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b,c</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,45</td>
<td>100</td>
</tr>
<tr>
<td>6,2</td>
<td>165</td>
</tr>
<tr>
<td>23,252</td>
<td>206</td>
</tr>
<tr>
<td>43,45</td>
<td>412</td>
</tr>
<tr>
<td>56,2</td>
<td>256</td>
</tr>
<tr>
<td>56,45</td>
<td>156</td>
</tr>
<tr>
<td>92,2</td>
<td>101</td>
</tr>
<tr>
<td>202,56</td>
<td>198</td>
</tr>
</tbody>
</table>

```
alter table foo add key(b,c);
sum(c) where b>50;
```
Indexes provide query performance

1. Indexes can reduce the amount of retrieved data.
   - Less bandwidth, less processing, ...

2. Indexes can improve locality.
   - Not all data access cost is the same
   - Sequential access is MUCH faster than random access

3. Indexes can presort data.
   - **GROUP BY** and **ORDER BY** queries do post-retrieval work
   - Indexing can help get rid of this work
Indexes can avoid post-selection sorts

```
select b, sum(c) group by b;
```
Data Structures and Algorithms for Big Data
Module 7: Paging

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
This Module

The algorithmics of cache-management.

This will help us understand I/O- and cache-efficient algorithms.
Goal: minimize # block transfers.

- Data is transferred in blocks between RAM and disk.
- Performance bounds are parameterized by \(B, M, N\).

When a block is cached, the access cost is 0. Otherwise it’s 1.

[Aggarwal+Vitter ’88]
Disk Access Model (DAM Model):
- Performance bounds are parameterized by $B$, $M$, $N$.

Goal: Minimize # of block transfers.

Beautiful restriction:
- Parameters $B$, $M$ are unknown to the algorithm or coder.

[Frigo, Leiserson, Prokop, Ramachandran ’99]
CO analysis applies to unknown multilevel hierarchies:
• Cache-oblivious algorithms work for all $B$ and $M$...
• ... and all levels of a multi-level hierarchy.

Moral:
• It’s better to optimize approximately for all $B$, $M$ rather than to try to pick the best $B$ and $M$.

[Frigo, Leiserson, Prokop, Ramachandran ’99]
Which blocks are currently cached in RAM?

- The system performs its own caching/paging.
- If we knew $B$ and $M$ we could explicitly manage I/O. (But even then, what should we do?)

![Diagram of RAM and Disk with $M=??$ and $B=??$]
Which blocks are currently cached in RAM?

- The system performs its own caching/paging.
- If we knew $B$ and $M$ we could explicitly manage I/O. (But even then, what should we do?)

*But systems may use different mechanisms, so what can we actually assume?*
With cache-oblivious analysis, we can assume a memory system with optimal replacement.

Even though the system manages memory, we can assume all the advantages of explicit memory management.
An LRU-based system with memory $M$ performs cache-management < 2x worse than the optimal, prescient policy with memory $M/2$.

Achieving optimal cache-management is hard because predicting the future is hard.

But LRU with $(1+\varepsilon)M$ memory is almost as good (or better), than the optimal strategy with $M$ memory.

[Sleator, Tarjan 85]
The paging/caching problem

A program is just sequence of block requests:

\[ r_1, r_2, r_3, \ldots \]

Cost of request \( r_j \)

\[
\text{cost}(r_j) = \begin{cases} 
0 & \text{block } r_j \text{ is already cached,} \\
1 & \text{block } r_j \text{ is brought into cache.}
\end{cases}
\]
RAM holds only $k = \frac{M}{B}$ blocks.

Which block should be ejected when block $r_j$ is brought into cache?
Paging Algorithms

LRU (least recently used)
• Discard block whose most recent access is earliest.

FIFO (first in, first out)
• Discard the block brought in longest ago.

LFU (least frequently used)
• Discard the least popular block.

Random
• Discard a random block.

LFD (longest forward distance) = OPT
• Discard block whose next access is farthest in the future.

[Belady 69]
LFD (Longest Forward Distance) [Belady ’69]:

- Discard the block requested farthest in the future.
LFD (Longest Forward Distance) [Belady ’69]:

- Discard the block requested farthest in the future.

Cons: Who knows the Future?!

Page 5348 shall be requested tomorrow at 2:00 pm.
LFD (Longest Forward Distance) [Belady ’69]:
- Discard the block requested farthest in the future.

Cons: Who knows the Future?!

Pros: LFD can be viewed as a point of comparison with online strategies.
An online algorithm $A$ is $k$-competitive, if for every request sequence $R$:

$$\text{cost}_A(R) \leq k \text{cost}_{\text{opt}}(R)$$

Idea of competitive analysis:

- The optimal (prescient) algorithm is a yardstick we use to compare online algorithms.
LRU is no better than $k$-competitive

Memory holds 3 blocks

$$\frac{M}{B} = k = 3$$

The program accesses 4 different blocks

$$r_j \in \{1, 2, 3, 4\}$$

The request stream is

$$1, 2, 3, 4, 1, 2, 3, 4, \ldots$$
LRU is no better than $k$-competitive

There’s a block transfer at every step because LRU ejects the block that’s requested in the next step.
LRU is no better than k-competitive

LFD (longest forward distance) has a block transfer every $k=3$ steps.
In fact, LRU is $k = \frac{M}{B}$-competitive.

- I.e., LRU has $k = \frac{M}{B}$ times more transfers than OPT.
- A depressing result because $k$ is huge so $k \cdot $OPT is nothing to write home about.

[Sleator, Tarjan 85]
On the other hand, the LRU bad example is fragile.

If $k = M/B = 4$, not 3, then both LRU and OPT do well. If $k = M/B = 2$, not 3, then neither LRU nor OPT does well.
LRU is 2-competitive with more memory

LRU is at most twice as bad as OPT, when LRU has twice the memory.

\[ \text{LRU}_{|\text{cache}|=k(R)} \leq 2 \text{OPT}_{|\text{cache}|=k/2(R)} \]

In general, LRU is nearly as good as OPT when LRU has a little more memory than OPT.
LRU is 2-competitive with more memory  

[Sleator, Tarjan 85]

LRU is at most twice as bad as OPT, when LRU has twice the memory.

\[ \text{LRU}_{|\text{cache}| = k(R)} \leq 2 \text{OPT}_{|\text{cache}| = k/2(R)} \]

LRU has more memory, but OPT=LFD can see the future.

In general, LRU is nearly as good as OPT when LRU has a little more memory than OPT.
LRU is 2-competitive with more memory \cite{SleatorTarjan85}

LRU is at most twice as bad as OPT, when LRU has twice the memory.

$$\text{LRU}_{|\text{cache}|=k(R)} \leq 2 \text{OPT}_{|\text{cache}|=k/2(R)}$$

LRU has more memory, but OPT=LFD can see the future.

In general, LRU is nearly as good as OPT when LRU has a little more memory than OPT.
Divide LRU into phases, each with $k$ faults.

$r_1, r_2, \ldots, r_i, r_{i+1}, \ldots, r_j, r_{j+1}, \ldots, r_\ell, r_{\ell+1}, \ldots$
Divide LRU into phases, each with $k$ faults.

$$r_1, r_2, \ldots, r_i, r_{i+1}, \ldots, r_j, r_{j+1}, \ldots, r_\ell, r_{\ell+1}, \ldots$$

**OPT[$k$]** must have $\geq 1$ fault in each phase.

- Case analysis proof.
- LRU is $k$-competitive.
Divide LRU into phases, each with $k$ faults.

$\ell_1, \ell_2, \ldots, \ell_i, \ell_{i+1}, \ldots, \ell_j, \ell_{j+1}, \ldots, \ell_\ell, \ell_{\ell+1}, \ldots$

**OPT[$k$] must have $\geq 1$ fault in each phase.**

- Case analysis proof.
- LRU is $k$-competitive.

**OPT[$k/2$] must have $\geq k/2$ faults in each phase.**

- Main idea: each phase must touch $k$ different pages.
- LRU is 2-competitive.
Under the hood of cache-oblivious analysis

Moral: with cache-oblivious analysis, we can analyze based on a memory system with optimal, omniscient replacement.

- Technically, an optimal cache-oblivious algorithm is asymptotically optimal versus any algorithm on a memory system that is slightly smaller.
- Empirically, this is just a technicality.

\[
\text{LRU} \quad (1+\varepsilon) M \quad \text{Disk} \quad \text{OPT} \quad M \quad \text{Disk}
\]

This is almost as good or better...

... than this.
Moral: There’s not much performance on the table for new cache-replacement policies.

- Bad instances for LRU versus LFD are fragile and depend on a particular cache size.

There are still research questions:

- What if blocks have different sizes [Irani 02][Young 02]? 
- There’s a write-back cost? (Complexity unknown.) 
- LRU may be too costly to implement (clock algorithm). 
- The cache-size changes over time.
  
  [Bender, Ebrahimi, Fineman, Ghasemiesfeh, Johnson, McCauley 13]
Data Structures and Algorithms for Big Data

Module 8: Sorting Big Data

Michael A. Bender
Stony Brook & Tokutek

Bradley C. Kuszmaul
MIT & Tokutek
Another way to create an index is to sort

- Sorting creates an index all-at-once.
- Sorting does not incrementally maintain an index.
- Sorting is faster than the best algorithms to incrementally maintain an index.

I/O-efficient mergesort

Parallel sort
How computation works:
  • Data is transferred in blocks between RAM and disk.
  • The # of block transfers dominates the running time.

Goal: Minimize # of block transfers
  • Performance bounds are parameterized by block size $B$, memory size $M$, data size $N$. 

[Aggarwal+Vitter ’88]
To sort an array of $N$ objects

- If $N$ fits in main memory, then just sort elements.
- Otherwise,
  - divide the array into $M/B$ pieces;
  - sort each piece (recursively); and
  - merge the $M/B$ pieces.
Why Divide into $M/B$ pieces?

- We want as much fan-in as possible.
- The merge needs to cache one block for each sorted subinput.
- Plus one block for the output.
- There are $M/B$ blocks in memory.
- So the fan-in can be at most $O(M/B)$
Merge Sort
Question: How many I/Os to sort $N$ elements?

- First run takes $N/B$ I/Os.
- Each level of the merge tree takes $N/B$ I/Os.
- How deep is the merge tree?

$$O \left( \frac{N}{B} \log \frac{M}{B} \frac{N}{B} \right)$$

Cost to scan data  # of scans of data
Question: How many I/Os to sort $N$ elements?

- First run takes $N/B$ I/Os.
- Each level of the merge tree takes $N/B$ I/Os.
- How deep is the merge tree?

\[
O \left( \frac{N}{B} \log_{M/B} \frac{N}{B} \right)
\]

Cost to scan data  # of scans of data

This bound is the best possible.
T(N), the number of I/Os to sort N items, satisfies this recurrence:

\[ T(N) = \frac{M}{B} \cdot T \left( \frac{N}{M/B} \right) + \frac{N}{B} \]

when \( N < M \)

\[ T(N) = \frac{N}{B} \]

Solution:

\[ O \left( \frac{N}{B} \log_{M/B} \frac{N}{B} \right) \]

Cost to scan data \# of scans of data
I/Os to sort $N$ objects:

$$O \left( \frac{N}{B} \log_{M/B} \frac{N}{B} \right)$$

I/Os to insert $N$ objects into a COLA:

$$O \left( \frac{N}{B} \log(N/M) \right)$$

I/Os to insert $N$ objects into a B-tree:

$$O \left( N \log_B(N/M) \right)$$

Sorting can usually be done in 2 passes since $M/B$ is large.
Big data might not fit on one machine.
So use many machines and parallelize.

Parallelizing merge sort is tricky, however.
To sort an array of $N$ objects

- If $N$ fits in main memory, then just sort elements.
- Otherwise
  - pick $M/B$ pivot keys;
  - partition data according to pivot keys; and
  - sort each partition (recursively).
Parallelizing Partitioning

- Broadcast the pivot keys to every processor.
- Compute the local rank of each pivot on each processor.
  - Sort local data to make this fast.
- Sum the local ranks to get global ranks.
- Send each datum to the right processor.
- The final step is a merge, since the local data was sorted.
Engineering Parallel Sort

- **Scheduling:**
  - Overlap I/O with computation and network communication.
  - Schedule network communication carefully to avoid network contention.

- **Hardware:**
  - Use a serious network.
  - Get rid of slow disks. Some disks are 10x slower than average. Probably failing.

- **In memory:**
  - Must compute local pivot ranks efficiently.
  - Employ a heap data structure to perform merge efficiently.

![Diagram of sort operations and data flow](image-url)
Bradley holds the world record for sorting a Terabyte: sortbenchmark.org

- 400 dual core machines with 2400 disks in 2007.
- Ran in 3.28 minutes.
- Used a distribution sort.
- Terabyte sort now deprecated, since it’s the same as minute sort (how much can you sort in a minute).
- Today to compete, you must sort 100TB in less than two hours.
Fast sorting is an important tool for big data.

Sorting provides many opportunities for cleverness.

No one can take my Terabyte sorting trophy!
Closing Words
We want to feel your pain.

We are interested in hearing about other scaling problems.

Come talk to us.

bender@cs.stonybrook.edu
michael@tokutek.com
bradley@mit.edu
bradley@tokutek.com
The problem with big data is microdata.

Sometimes the right read optimization is a write-optimization.

It’s often better to optimize approximately for all $B, M$ than to pick the best $B$ and $M$.

As data becomes bigger, the asymptotics become more important.