Parallel Join Algorithms

(Hashing)

@Andy_Pavlo // 15-721 // Spring 2018
TODAY’S AGENDA

Background
Parallel Hash Join
Hash Functions
Hashing Schemes
Evaluation
PARALLEL JOIN ALGORITHMS

Perform a join between two relations on multiple threads simultaneously to speed up operation.

Two main approaches:
→ Hash Join
→ Sort-Merge Join

We won’t discuss nested-loop joins...
Many OLTP DBMSs don’t implement hash join. But an **index nested-loop join** with a small number of target tuples is more or less equivalent to a hash join.
HASHING VS. SORTING

1970s – Sorting
1980s – Hashing
1990s – Equivalent
2000s – Hashing
2010s – Hashing (Partitioned vs. Non-Partitioned)
2020s – ???
### PARALLEL JOIN ALGORITHMS

| SORT VS. HASH REVISITED: FAST JOIN IMPLEMENTATION ON MODERN MULTI-CORE CPUs | ORACLE | VLDB 2009 |
| → Hashing is faster than Sort-Merge. |
| → Sort-Merge is faster w/ wider SIMD. |

| DESIGN AND EVALUATION OF MAIN MEMORY HASH JOIN ALGORITHMS FOR MULTI-CORE CPUs | WISCONSIN UNIVERSITY OF WISCONSIN-MADISON | SIGMOD 2011 |
| → Trade-offs between partitioning & non-partitioning Hash-Join. |

| MASSIVELY PARALLEL SORT-MERGE JOINS IN MAIN MEMORY MULTI-CORE DATABASE SYSTEMS | HyPer | VLDB 2012 |
| → Sort-Merge is already faster than Hashing, even without SIMD. |

| MAIN-MEMORY HASH JOINS ON MULTI-CORE cpus: TUNING TO THE UNDERLYING HARDWARE | Systems of ETH Zürich | ICDE 2013 |
| → New optimizations and results for Radix Hash Join. |

| MASSIVELY PARALLEL NUMA-AWARE HASH JOINS | HyPer | IMDM 2013 |
| → Ignore what we said last year. |
| → You really want to use Hashing! |

| AN EXPERIMENTAL COMPARISON OF THIRTEEN RELATIONAL EQUI-JOINS IN MAIN MEMORY | UNIVERSITÄT DES SAARLANDES | SIGMOD 2016 |
| → Hold up everyone! Let's look at everything for real! |

→ Ignore what we said last year.
→ You really want to use Hashing!
JOIN ALGORITHM DESIGN GOALS

Goal #1: Minimize Synchronization
  → Avoid taking latches during execution.

Goal #2: Minimize CPU Cache Misses
  → Ensure that data is always local to worker thread.
IMPROVING CACHE BEHAVIOR

Factors that affect cache misses in a DBMS:
→ Cache + TLB capacity.
→ Locality (temporal and spatial).

Non-Random Access (Scan):
→ Clustering to a cache line.
→ Execute more operations per cache line.

Random Access (Lookups):
→ Partition data to fit in cache + TLB.

Source: Johannes Gehrke
PARALLEL HASH JOINS

Hash join is the most important operator in a DBMS for OLAP workloads.

It’s important that we speed it up by taking advantage of multiple cores.
→ We want to keep all of the cores busy, without becoming memory bound.
% of Total CPU Time Spent in Query Operators
Workload: TPC-H Benchmark

- HASH JOIN: 49.6%
- SEQ SCAN: 25.0%
- UNION: 19.9%
- AGGREGATE: 3.1%
- OTHER: 2.4%
CLOUDERA IMPALA

% of Total CPU Time Spent in Query Operators

- HASH JOIN
- SEQ SCAN
- UNION
- AGGREGATE
- OTHER

Workload: TPC-H Benchmark

CMU 15-721 (Spring 2018)
HASH JOIN (R⨝S)

Phase #1: Partition (optional)
→ Divide the tuples of R and S into sets using a hash on the join key.

Phase #2: Build
→ Scan relation R and create a hash table on join key.

Phase #3: Probe
→ For each tuple in S, look up its join key in hash table for R. If a match is found, output combined tuple.
PARTITION PHASE

Split the input relations into partitioned buffers by hashing the tuples’ join key(s).
→ Ideally the cost of partitioning is less than the cost of cache misses during build phase.
→ Sometimes called hybrid hash join.

Contents of buffers depends on storage model:
→ NSM: Either the entire tuple or a subset of attributes.
→ DSM: Only the columns needed for the join + offset.
PARTITION PHASE

Approach #1: Non-Blocking Partitioning
→ Only scan the input relation once.
→ Produce output incrementally.

Approach #2: Blocking Partitioning (Radix)
→ Scan the input relation multiple times.
→ Only materialize results all at once.
→ Sometimes called *radix hash join*. 
NON-BLOCKING PARTITIONING

Scan the input relation only once and generate the output on-the-fly.

**Approach #1: Shared Partitions**
- Single global set of partitions that all threads update.
- Have to use a latch to synchronize threads.

**Approach #2: Private Partitions**
- Each thread has its own set of partitions.
- Have to consolidate them after all threads finish.
## SHARED PARTITIONS

### Data Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## SHARED PARTITIONS

### Data Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>hash&lt;sub&gt;p&lt;/sub&gt;(key)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>#&lt;sub&gt;p&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#&lt;sub&gt;p&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#&lt;sub&gt;p&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#&lt;sub&gt;p&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#&lt;sub&gt;p&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#&lt;sub&gt;p&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

**hash<sub>p</sub>(key)**
**SHARED PARTITIONS**

**Data Table**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>hash_p(key)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>#_p</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#_p</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#_p</td>
</tr>
</tbody>
</table>

**Partitions**

- \( P_1 \)
- \( P_2 \)
- \( \ldots \)
- \( P_n \)
SHARED PARTITIONS

Data Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Partitions

\[ \text{hash}_p(key) \]

\[ P_1 \]

\[ P_2 \]

\[ \vdots \]

\[ P_n \]
# Shared Partitions

**Data Table**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Data Table" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Partitions**

\[
\text{hash}_p(key)
\]

<table>
<thead>
<tr>
<th>( P_1 )</th>
<th>( P_2 )</th>
<th>( \ldots )</th>
<th>( P_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.png" alt="Partitions" /></td>
<td><img src="image2.png" alt="Partitions" /></td>
<td><img src="image2.png" alt="Partitions" /></td>
<td><img src="image2.png" alt="Partitions" /></td>
</tr>
</tbody>
</table>
PRIVATE PARTITIONS

Data Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
</table>

Partitions

\[ \text{hash}_p(key) \]

\[ P_1 \]
\[ P_2 \]
\[ P_n \]
PRIVATE PARTITIONS

Data Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
</table>

hash_p(key)

Partitions

\[
P_1 \
\vdots 

P_n
\]

Combined

\[
P_1 \rightarrow \text{Table} \\
P_2 \rightarrow \text{Table} \\
\vdots \\
P_n \rightarrow \text{Table}
\]
PRIVATE PARTITIONS

Data Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Partitions

\[ \text{hash}_p(key) \]

Combined

\[ P_1 \]
\[ P_2 \]
\[ \vdots \]
\[ P_n \]
RADIX PARTITIONING

Scan the input relation multiple times to generate the partitions.

Multi-step pass over the relation:

→ **Step #1:** Scan $R$ and compute a histogram of the # of tuples per hash key for the *radix* at some offset.

→ **Step #2:** Use this histogram to determine output offsets by computing the *prefix sum*.

→ **Step #3:** Scan $R$ again and partition them according to the hash key.
The radix is the value of an integer at a particular position (using its base).

**Input** 8 9 1 2 2 3 0 8 4 1 6 4

**Radix** 8 1 2 0 4 6
The radix is the value of an integer at a particular position (using its base).
The radix is the value of an integer at a particular position (using its base).
The prefix sum of a sequence of numbers 
$(x_0, x_1, ..., x_n)$
is a second sequence of numbers 
$(y_0, y_1, ..., y_n)$
that is a running total of the input sequence.
The prefix sum of a sequence of numbers $(x_0, x_1, ..., x_n)$ is a second sequence of numbers $(y_0, y_1, ..., y_n)$ that is a running total of the input sequence.
The prefix sum of a sequence of numbers \((x_0, x_1, \ldots, x_n)\) is a second sequence of numbers \((y_0, y_1, \ldots, y_n)\) that is a running total of the input sequence.

**Input**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
</table>

**Prefix Sum**

|   | 1 | 3 | 6 | 10 | 15 | 21 |
# RADIX PARTITIONS

**Step #1: Inspect input, create histograms**

<table>
<thead>
<tr>
<th>hash_p(key)</th>
<th>07</th>
<th>18</th>
<th>19</th>
<th>07</th>
<th>03</th>
<th>11</th>
<th>15</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>#p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Spyros Blanas
### RADIX PARTITIONS

**Step #1: Inspect input, create histograms**

<table>
<thead>
<tr>
<th>hash_p(key)</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>#p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td>07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td>07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td></td>
<td>03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#p</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Spyros Blanas
RADIX PARTITIONS

Step #1: Inspect input, create histograms

- **Partition 0**: 2
  - 0
  - 7

- **Partition 1**: 2
  - 1
  - 8
  - 9

- **Partition 0**: 1
  - 0
  - 3

- **Partition 1**: 3
  - 1
  - 1
  - 5
  - 0

Source: Spyros Blanas
**RADIX PARTITIONS**

*Step #2: Compute output offsets*

<table>
<thead>
<tr>
<th>#p</th>
<th>hash_p(key)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>07</td>
</tr>
<tr>
<td>18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td></td>
</tr>
<tr>
<td>03</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

- **Partition 0:** 2
  - CPU 0
- **Partition 1:** 3
  - CPU 1
- **Partition 0:** 2
  - CPU 1
- **Partition 1:** 0
  - CPU 0
- **Partition 1:** 1
  - CPU 1

Source: Spyros Blanas
RADIX PARTITIONS

Step #3: Read input and partition

Partition 0, CPU 0
Partition 0, CPU 1
Partition 1, CPU 0
Partition 1, CPU 1

Source: Spyros Blanas
RADIX PARTITIONS

Step #3: Read input and partition

Partition 0: 2
Partition 1: 2

Partition 0: 1
Partition 1: 3

Partition 0, CPU 0
Partition 0, CPU 1
Partition 1, CPU 0
Partition 1, CPU 1

Source: Spyros Blanas
Step #3: Read input and partition

```
<table>
<thead>
<tr>
<th>P</th>
<th>07</th>
<th>18</th>
<th>19</th>
<th>07</th>
<th>03</th>
<th>11</th>
<th>15</th>
<th>10</th>
</tr>
</thead>
</table>
Partition 0: 2
Partition 1: 2
```

```
<table>
<thead>
<tr>
<th>P</th>
<th>07</th>
<th>07</th>
<th>03</th>
<th>18</th>
<th>19</th>
<th>11</th>
<th>15</th>
<th>10</th>
</tr>
</thead>
</table>
Partition 0, CPU 0
Partition 0, CPU 1
Partition 1, CPU 0
Partition 1, CPU 1
```

Source: Spyros Blanas
RADIX PARTITIONS

Recursively repeat until target number of partitions have been created

<table>
<thead>
<tr>
<th>#p</th>
<th>0</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>#p</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>#p</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>#p</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>#p</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>#p</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>#p</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>#p</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Partition 0: 2
Partition 1: 2

Partition 0: 1
Partition 1: 3

Partition 0
Partition 1

Source: Spyros Blanas
## RADIX PARTITIONS

Recursively repeat until target number of partitions have been created

<table>
<thead>
<tr>
<th>#p</th>
<th>07</th>
<th>07</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>07</td>
<td>03</td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>07</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>03</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Partition 0: 2
Partition 1: 2

Partition 0: 1
Partition 1: 3

Source: Spyros Blanas
RADIX PARTITIONS

Recursively repeat until target number of partitions have been created

<table>
<thead>
<tr>
<th>#_p</th>
<th>07</th>
<th>18</th>
<th>19</th>
<th>07</th>
<th>03</th>
<th>11</th>
<th>15</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>hash(p(key)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Partition 0: 2
Partition 1: 2

Partition 0: 1
Partition 1: 3

Source: Spyros Blanas
RADIX PARTITIONS

Recursively repeat until target number of partitions have been created

hash_{p}(key)

<table>
<thead>
<tr>
<th>#_p</th>
<th>07</th>
<th>18</th>
<th>19</th>
<th>07</th>
<th>03</th>
<th>11</th>
<th>15</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition 0: 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partition 1: 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Spyros Blanas
BUILD PHASE

The threads are then to scan either the tuples (or partitions) of $R$.

For each tuple, hash the join key attribute for that tuple and add it to the appropriate bucket in the hash table.

→ The buckets should only be a few cache lines in size.
Design Decision #1: Hash Function
→ How to map a large key space into a smaller domain.
→ Trade-off between being fast vs. collision rate.

Design Decision #2: Hashing Scheme
→ How to handle key collisions after hashing.
→ Trade-off between allocating a large hash table vs. additional instructions to find/insert keys.
HASH FUNCTIONS

We don’t want to use a cryptographic hash function for our join algorithm.

We want something that is fast and will have a low collision rate.
HASH FUNCTIONS

MurmurHash (2008)
→ Designed to a fast, general purpose hash function.

Google CityHash (2011)
→ Based on ideas from MurmurHash2
→ Designed to be faster for short keys (<64 bytes).

Google FarmHash (2014)
→ Newer version of CityHash with better collision rates.

CLHash (2016)
→ Fast hashing function based on carry-less multiplication.
HASH FUNCTION BENCHMARKS

Intel Core i7-8700K @ 3.70GHz

Throughput (MB/sec) vs. Key Size (bytes)

- std::hash
- MurmurHash3
- CityHash
- FarmHash
- CLHash

Source: Fredrik Widlund
HASH FUNCTION BENCHMARKS

Intel Core i7-8700K @ 3.70GHz

Throughput (MB/sec) vs. Key Size (bytes)

- std::hash
- MurmurHash3
- CityHash
- FarmHash
- CLHash

Source: Fredrik Widlund
HASHING SCHEMES

Approach #1: Chained Hashing
Approach #2: Linear Probe Hashing
Approach #3: Robin Hood Hashing
Approach #4: Cuckoo Hashing
CHAINED HASHING

Maintain a linked list of “buckets” for each slot in the hash table.

Resolve collisions by placing all elements with the same hash key into the same bucket.
→ To determine whether an element is present, hash to its bucket and scan for it.
→ Insertions and deletions are generalizations of lookups.
CHAINED HASHING

\[ \text{hash(key)} \]
LINEAR PROBE HASHING

Single giant table of slots.
Resolve collisions by linearly searching for the next free slot in the table.
→ To determine whether an element is present, hash to a location in the table and scan for it.
→ Have to store the key in the table to know when to stop scanning.
→ Insertions are generalizations of lookups.
LINEAR PROBE HASHING

\[\text{hash(key)}\]

<table>
<thead>
<tr>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>F</td>
</tr>
</tbody>
</table>

\[
\text{hash}(A) | A
\]
LINEAR PROBE HASHING

hash(key)

A
B
C
D
E
F

hash(A) | A

hash(B) | B
LINEAR PROBE HASHING

\[
\text{hash(key)} \\
A \\
B \\
C \\
D \\
E \\
F
\]

\[
\begin{align*}
\text{hash(A)} & \mid A \\
\text{hash(B)} & \mid B \\
\text{hash(C)} & \mid C
\end{align*}
\]
LINEAR PROBE HASHING

hash(key)

A
B
C
D
E
F

hash(A) | A
hash(B) | B
hash(C) | C
hash(D) | D
LINEAR PROBE HASHING

hash(key)

\begin{align*}
&\text{A} & \text{hash(A)} & \text{A} \\
&\text{B} & \text{hash(B)} & \text{B} \\
&\text{C} & \text{hash(A)} & \text{B} \\
&\text{D} & \text{hash(C)} & \text{C} \\
&\text{E} & \text{hash(D)} & \text{D} \\
&\text{F} & \text{hash(E)} & \text{E} \\
\end{align*}
LINEAR PROBE HASHING

\[
\begin{align*}
\text{hash(key)} & \quad \text{hash}(A) & \quad A \\
& \quad \text{hash}(B) & \quad B \\
& \quad \text{hash}(A) & \quad A \\
& \quad \text{hash}(C) & \quad C \\
& \quad \text{hash}(D) & \quad D \\
& \quad \text{hash}(E) & \quad E \\
& \quad \text{hash}(F) & \quad F
\end{align*}
\]
OBSERVATION

To reduce the # of wasteful comparisons during the join, it is important to avoid collisions of hashed keys.

This requires a chained hash table with ~2x the number of slots as the # of elements in $R$. 
ROBIN HOOD HASHING

Variant of linear hashing that steals slots from "rich" keys and give them to "poor" keys.

→ Each key tracks the number of positions they are from where its optimal position in the table.
→ On insert, a key takes the slot of another key if the first key is farther away from its optimal position than the second key.

ROBIN HOOD HASHING
Foundations of Computer Science 1985
ROBIN HOOD HASHING

hash(key)

A
B
C
D
E
F

hash(A)| A [θ]

# of "Jumps" From First Position
ROBIN HOOD HASHING

hash(key)

A
B
C
D
E
F

hash(B)|B[∅]

hash(A)|A[∅]
ROBIN HOOD HASHING

\[\text{hash(key)}\]

\[
\begin{array}{c|c}
\text{hash(B)} & B[\emptyset] \\
\text{hash(A)} & A[\emptyset] \\
\text{hash(C)} & C[1] \\
\end{array}
\]

\[A[0] == C[0]\]
ROBIN HOOD HASHING

hash(key)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>hash(A)</td>
<td>A [0]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hash(B)</td>
<td>B [0]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hash(C)</td>
<td>C [1]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hash(D)</td>
<td>D [1]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C[1] > D[0]
Robin Hood Hashing

\[ \text{hash}(key) \]

- \( \text{hash}(A) \mid A[0] \)
- \( \text{hash}(B) \mid B[\emptyset] \)
- \( \text{hash}(C) \mid C[1] \)
- \( \text{hash}(D) \mid D[1] \)

- \( A[0] = E[0] \)
ROBIN HOOD HASHING

hash(key)

\[
egin{array}{c|c}
A & \emptyset \\
B & \emptyset \\
C & 1 \\
D & \\
E & 2 \\
F & \\
\end{array}
\]

- hash(A) = 0
- hash(B) = 0
- hash(C) = 1
- hash(E) = 2

- \( A[0] = E[0] \)
ROBIN HOOD HASHING

```
hash(A) | A [empty]
hash(B) | B [empty]
hash(C) | C [1]
hash(D) | D [2]
hash(E) | E [2]
```

- A[0] == E[0]
- C[1] == E[1]
ROBIN HOOD HASHING

hash(key)

\[
\begin{align*}
\text{hash}(A) & : A[0] \\
\text{hash}(B) & : B[0] \\
\text{hash}(C) & : C[1] \\
\text{hash}(D) & : D[2] \\
\text{hash}(E) & : E[2] \\
\text{hash}(F) & : F[1]
\end{align*}
\]

D[2] > F[0]
CUCKOO HASHING

Use multiple tables with different hash functions.
→ On insert, check every table and pick anyone that has a free slot.
→ If no table has a free slot, evict the element from one of them and then re-hash it find a new location.

Look-ups are always O(1) because only one location per hash table is checked.
Cuckoo Hashing

Hash Table #1

Insert X

$\text{hash}_1(X)$

$\text{hash}_2(X)$

Hash Table #2
Cuckoo Hashing

Hash Table #1

Insert X

\[ \text{hash}_1(X) \]

\[ \text{hash}_2(X) \]

Hash Table #2
Cuckoo Hashing

Hash Table #1

Insert X

\[ \text{hash}_1(X) \]
\[ \text{hash}_2(X) \]

Hash Table #2

Insert Y

\[ \text{hash}_1(Y) \]
\[ \text{hash}_2(Y) \]
CUCKOO HASHING

Hash Table #1

Insert X

$hash_1(X)$

$hash_2(X)$

Insert Y

$hash_1(Y)$

$hash_2(Y)$

Hash Table #2

X

Y
CUCKOO HASHING

Hash Table #1

Insert X
\[ \text{hash}_1(X) \quad \text{hash}_2(X) \]

Insert Y
\[ \text{hash}_1(Y) \quad \text{hash}_2(Y) \]

Insert Z
\[ \text{hash}_1(Z) \quad \text{hash}_2(Z) \]

Hash Table #2

X

Y

Z
CUCKOO HASHING

Insert X
\[ \text{hash}_1(X) \quad \text{hash}_2(X) \]

Insert Y
\[ \text{hash}_1(Y) \quad \text{hash}_2(Y) \]

Insert Z
\[ \text{hash}_1(Z) \quad \text{hash}_2(Z) \]
Cuckoo Hashing

Hash Table #1

$\text{Insert } X$
$\text{hash}_1(X)$ $\text{hash}_2(X)$

$X$

$\text{Insert } Y$
$\text{hash}_1(Y)$ $\text{hash}_2(Y)$

$\text{Insert } Z$
$\text{hash}_1(Z)$ $\text{hash}_2(Z)$

Hash Table #2

$Z$
Cuckoo Hashing

Hash Table #1

Insert X
hash₁(X)  hash₂(X)

Insert Y
hash₁(Y)  hash₂(Y)

Insert Z
hash₁(Z)  hash₂(Z)
hash₁(Y)

Hash Table #2

Z
Cuckoo Hashing

Hash Table #1

Insert X
\[ \text{hash}_1(X) \quad \text{hash}_2(X) \]

Insert Y
\[ \text{hash}_1(Y) \quad \text{hash}_2(Y) \]

Insert Z
\[ \text{hash}_1(Z) \quad \text{hash}_2(Z) \]
\[ \text{hash}_1(Y) \]

Hash Table #2

Z
CUCKOO HASHING

**Hash Table #1**
- Insert X
  - hash_1(X)
  - hash_2(X)
- Insert Y
  - hash_1(Y)
  - hash_2(Y)
- Insert Z
  - hash_1(Z)
  - hash_2(Z)
  - hash_1(Y)
  - hash_2(X)

**Hash Table #2**
- Z
- X
Threads have to make sure that they don’t get stuck in an infinite loop when moving keys.

If we find a cycle, then we can rebuild the entire hash tables with new hash functions.

→ With two hash functions, we (probably) won’t need to rebuild the table until it is at about 50% full.

→ With three hash functions, we (probably) won’t need to rebuild the table until it is at about 90% full.
PROBE PHASE

For each tuple in $S$, hash its join key and check to see whether there is a match for each tuple in corresponding bucket in the hash table constructed for $R$.

→ If inputs were partitioned, then assign each thread a unique partition.

→ Otherwise, synchronize their access to the cursor on $S$. 
PROBE PHASE – BLOOM FILTER

Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table. This will be faster since the filter will fit in CPU caches.

→ Sometimes called *sideways information passing.*
Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table.

This will be faster since the filter will fit in CPU caches.

→ Sometimes called *sideways information passing*.
Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table. This will be faster since the filter will fit in CPU caches.

→ Sometimes called *sideways information passing*.
Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table. This will be faster since the filter will fit in CPU caches.
→ Sometimes called \textit{sideways information passing}.

\textbf{PROBE PHASE – BLOOM FILTER}
Create a Bloom Filter during the build phase when the key is likely to not exist in the hash table.

→ Threads check the filter before probing the hash table.
  This will be faster since the filter will fit in CPU caches.
→ Sometimes called *sideways information passing*.
## Hash Join Variants

<table>
<thead>
<tr>
<th></th>
<th>No-P</th>
<th>Shared-P</th>
<th>Private-P</th>
<th>Radix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partitioning</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Input scans</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Sync during partitioning</td>
<td>_</td>
<td>Spinlock per tuple</td>
<td>Barrier, once at end</td>
<td>Barrier, 4 * #passes</td>
</tr>
<tr>
<td>Hash table</td>
<td>Shared</td>
<td>Private</td>
<td>Private</td>
<td>Private</td>
</tr>
<tr>
<td>Sync during build phase</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sync during probe phase</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
BENCHMARKS

Primary key – foreign key join
→ Outer Relation (Build): 16M tuples, 16 bytes each
→ Inner Relation (Probe): 256M tuples, 16 bytes each

Uniform and highly skewed (Zipf; s=1.25)

No output materialization
HASH JOIN – UNIFORM DATA SET

Intel Xeon CPU X5650 @ 2.66GHz
6 Cores with 2 Threads Per Core

Intel Xeon CPU X5650 @ 2.66GHz
6 Cores with 2 Threads Per Core

3.3x cache misses
70x TLB misses

24% faster than No Partitioning

Source: Spyros Blanas
HASH JOIN – SKEWED DATA SET

Intel Xeon CPU X5650 @ 2.66GHz
6 Cores with 2 Threads Per Core

Cycles / Output Tuple

Source: Spyros Blanas
OBSERVATION

We have ignored a lot of important parameters for all of these algorithms so far.
→ Whether to use partitioning or not?
→ How many partitions to use?
→ How many passes to take in partitioning phase?

In a real DBMS, the optimizer will select what it thinks are good values based on what it knows about the data (and maybe hardware).
RADIX HASH JOIN – UNIFORM DATA SET

Intel Xeon CPU X5650 @ 2.66GHz
Varying the # of Partitions

Cycles / Output Tuple

<table>
<thead>
<tr>
<th>Radix / 1-Pass</th>
<th>Radix / 2-Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>4096</td>
<td>4096</td>
</tr>
<tr>
<td>8192</td>
<td>8192</td>
</tr>
<tr>
<td>32768</td>
<td>32768</td>
</tr>
<tr>
<td>131072</td>
<td>131072</td>
</tr>
</tbody>
</table>

Source: Spyros Blanas
RADIX HASH JOIN – UNIFORM DATA SET

Intel Xeon CPU X5650 @ 2.66GHz
Varying the # of Partitions

<table>
<thead>
<tr>
<th>Cycles / Output Tuple</th>
<th>64</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>4096</th>
<th>8192</th>
<th>32768</th>
<th>131072</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radix / 1-Pass</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radix / 2-Pass</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Spyros Blanas
EFFECTS OF HYPER-THREADING

Intel Xeon CPU X5650 @ 2.66GHz
Uniform Data Set

Multi-threading hides cache & TLB miss latency.
Radix join has fewer cache & TLB misses but this has marginal benefit.
Non-partitioned join relies on multi-threading for high performance.

Source: Spyros Blanas
PARTING THOUGHTS

On modern CPUs, a simple hash join algorithm that does not partition inputs is competitive.

There are additional vectorization execution optimizations that are possible in hash joins that we didn’t talk about. But these don’t really help...
NEXT CLASS

Parallel Sort-Merge Joins