ETH zürich

ETH Library

Many-query join: efficient shared execution of relational joins on modern hardware

Journal Article

Author(s):
Makreshanski, Darko; Giannikis, Georgios; Alonso, Gustavo; Kossmann, Donald

Publication date:
2018-10

Permanent link:
https://doi.org/10.3929/ethz-b-000192365

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
The VLDB Journal 27(5), https://doi.org/10.1007/s00778-017-0475-4

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

The VLDB Journal (2018) 27:669–692
https://doi.org/10.1007/s00778-017-0475-4

Many-query join: efficient shared execution of relational joins on modern hardware

Darko Makreshanski[1] · Georgios Giannikis[2] · Gustavo Alonso[1] · Donald Kossmann[3]

Received: 15 January 2017 / Revised: 26 May 2017 / Accepted: 25 July 2017 / Published online: 30 August 2017
© Springer-Verlag GmbH Germany 2017

Abstract Database architectures typically process queries one at a time, executing concurrent queries in independent execution contexts. Often, such a design leads to unpredictable performance and poor scalability. One approach to circumvent the problem is to take advantage of sharing opportunities across concurrently running queries. In this paper, we propose many-query join (MQJoin), a novel method for sharing the execution of a join that can efficiently deal with hundreds of concurrent queries. This is achieved by minimizing redundant work and making efficient use of main-memory bandwidth and multi-core architectures. Compared to existing proposals, MQJoin is able to efficiently handle larger workloads regardless of the schema by exploiting more sharing opportunities. We also compared MQJoin to two commercial main-memory column-store databases. For a TPC-H-based workload, we show that MQJoin provides 2–5× higher throughput with significantly more stable response times.

✉ Darko Makreshanski
darkoma@inf.ethz.ch

Georgios Giannikis
georgios.giannikis@oracle.com

Gustavo Alonso
alonso@inf.ethz.ch

Donald Kossmann
donaldk@microsoft.com

[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland

[2] Oracle Labs Zurich, Zurich, Switzerland

[3] Microsoft Research, Redmond, WA, USA

Keywords RDBMS · OLAP · Analytics · Join · MQJoin · Shared join · Main memory · TPC-H · Xeon Phi · MCDRAM

1 Introduction

In recent years, increased connectivity and availability of information have changed the requirements for databases. Systems catering to large user bases must provide robust performance with strong guarantees. This, together with the trend toward real-time data analytics, has put a strain on database architectures. Under these circumstances, systems must be designed to provide guaranteed response times for complete workloads, rather than the fastest performance for individual queries. For instance, reservation systems used in the airline industry need to execute hundreds of decision support queries per second with tight latency guarantees while sustaining high update rates [47].

An emerging approach to deal with such requirements is to exploit the sharing opportunities available in these workloads. Various techniques for sharing query execution have been explored to date, ranging from exploiting common subexpressions in multi-query optimization [44]; simultaneous pipelining in QPipe [20]; sharing of scans in MonetDB [49], Blink [39,41] and Crescando [47]; sharing global query plans in CJoin [11], Datapath [3] and SharedDB [16].

As one of the most expensive relational operations, efficient join processing is crucial for performance. Exploiting sharing opportunities in joins across multiple queries is important to sustain throughput in highly concurrent workloads.

In this paper, we present MQJoin, a method for sharing join execution that is able to efficiently exploit sharing opportunities and provide high performance for up to hundreds of

concurrent join queries. Similarly to CJoin [11], Datapath [3] and SharedDB [16], MQJoin shares query execution by annotating intermediate results with additional information. What differentiates our approach is the use of several techniques that enables a significantly higher degree of sharing and an efficient use of main-memory bandwidth and CPU resources. This allows MQJoin to outperform state-of-the-art commercial analytical main-memory databases for workloads with high concurrency.

To evaluate MQJoin, we first present a series of microbenchmarks to illustrate the benefits and overhead of the approach with respect to a query-at-a-time counterpart. We analyze how much overlap should intermediate relations of queries have so that sharing pays off. Using an existing shared scan implementation as a storage engine, we then compare MQJoin integrated into a complete system to commercial databases and related work. Performance-wise, we compare to two leading main-memory analytical databases, namely VectorWise and another popular commercial database which we refer to as System X, on a TPC-H-based workload. We show that our system outperforms its commercial counterparts in terms of throughput when the load grows beyond 60 clients. Furthermore, it provides significantly more stable and predictable response times, having a lower 99th percentile even for a handful of clients. In terms of scalability we also compare to CJoin, the closest approach to ours, and show that for the star schema benchmark [37] for which CJoin was designed, MQJoin is able to provide up to an order of magnitude more throughput while maintaining lower response times.

The main contributions of the paper are: (1) We present a method for sharing joins for highly concurrent workloads that supports one order of magnitude more concurrent queries than the best published result to date; (2) we provide an analysis of the impact of sharing on main-memory joins showing how to adapt existing join algorithms to support sharing; and (3) we validate the potential of the idea through a comparison of a shared scan/join system to leading main-memory analytical databases demonstrating 2–5× higher performance.

The rest of the paper is organized as follows: Sect. 2 discusses related work on join algorithms and shared query execution systems; Sect. 3 gives a model of the shared join execution approach that we use; Sect. 4 analyzes effects of sharing on state-of-the-art join algorithms; Sect. 5 explains the two-way join algorithm in detail; Sect. 6 explains how multi-way joins are handled; Sect. 7 explains the system architecture, including integration with shared scans; Sect. 8 provides extensive analysis on the effects of sharing and the performance of MQJoin; Sect. 8.5 investigates an alternative to MQJoin where the shared scans are executed after the join; Sect. 9 concludes the paper.

2 Background and related work

2.1 Main-memory join execution

The performance of a join is very important in a relational database. Due to the availability of systems with large main memories, recent research has focused on optimizing in-memory joins. Shatdal et al. [45] proposed partitioning the relations so that they fit in cache to avoid high random access latencies. Manegold et al. [32] partition the relations in two steps to avoid expensive TLB misses during partitioning. Chen et al. [13,14] propose software prefetching techniques to hide the memory latencies of random accesses. More recently, there has also been discussions on whether sort–merge join or hash-based join are better suited for modern architectures [5,7,24], on complexity and similarity of hashing and sorting [35], as well as whether it is worth tuning to the underlying architecture [6,9]. There is also a line of work that investigates main-memory joins for NUMA architectures [2] and Xeon Phi coprocessors [22] and memory efficient hash joins [8]. We have carefully evaluated these results to design a join that is as efficient as possible but supports shared execution.

2.2 Shared query execution

Several techniques for shared query execution have been developed to date. Sharing execution was initially proposed in the form of multi-query optimization (MQO) [44]. MQO detects and jointly executes common subexpressions in multiple queries including execution of join operations. The idea of MQO has been further extended in various forms, for instance for query result caching [12]; as part of the Volcano optimizer [42] in the presence of materialized/cached views; or for reusing intermediate results [21,34]. StagedDB [19] and QPipe [20] use a simultaneous pipelining technique to share execution of queries that arrive within a certain time-frame. Using a system based on these techniques, Johnson et al. [23] show that there is a trade-off between sharing and parallelism. A limitation in these systems is that they rely on temporal overlap for sharing. Typical results show sharing for a few tens of queries [20].

Sharing data and work for scans has been shown to be effective in various forms and use cases. MonetDB [49] optimizes disk bandwidth utilization by doing cooperative scans where queries are dynamically scheduled according to their data requests and the current status of the disk buffers. Similarly, systems such as IBM UDB [26,27] perform dynamic scan grouping and ungrouping as well as adaptive throttling of scan speeds to increase buffer locality. Blink [39,41] and Crescando [47] go one step further and answer multiple queries in one table scan, independently of the query predicates, thereby sharing disk bandwidth and memory

bandwidth. In those systems, the degree of sharing is between a few hundred to several thousand queries.

Recently, several systems propose shared execution of complex operations such as joins, for queries without common subexpressions. CJoin [11] achieves high scalability, handling up to 256 concurrent queries, by using a single always-on plan of operators that executes all queries. The approach is tailored to star schemas. Datapath [3] makes the case for a data-centric approach to analytical databases, advocating a push-based model to query processing instead of the traditional pull-based. They work with a more general TPC-H schema and show sharing for up to 7 concurrent queries. A push-based, data-flow model for query processing was also used in the Eddies project [4]. While Eddies are similar to sharing, they were designed to provide runtime adaptivity of query execution where a static query plan generation is not sufficient. They cannot provide high throughput for concurrent workloads.

SharedDB [16,17] shows that a shared query execution system based on a global query plan and batching can give robust performance for highly concurrent workloads of up to thousands of queries. SharedDB, however, uses single-threaded operators.

Psaroudakis et al. [38] integrate the approaches of CJoin and QPipe. This work shows that a combination of global query plans with shared operators and simultaneous pipelining is better suited for high concurrency, while traditional query execution with simultaneous pipelining is better suited for low concurrency workloads. Like CJoin, the authors also focus on star schema workloads.

Ebenstein et al. [15] propose FluxQuery, a model for representing the likelihood of queries in highly interactive workloads and a method for handling queries from such workloads. FluxQuery is designed to handle thousands of concurrent queries, where the concurrency comes from the large amount of possible queries derived from an ambiguous query intent of a user input, for instance when the join predicate between two relations is not yet specified. The query processing method is based on a cyclic scan-based approach that combines nested loop joins and hash joins. This provides a balance between the performance of hash joins and responsivity of nested loop joins.

3 Shared join model

This section presents a model for the input and output characteristics of the shared join algorithm. The algorithm itself is described in Sect. 5. For simplicity, this model represents only sharing of two-way inner joins. Handling of other join types is described in Sect. 5.5, while Sect. 6 covers multi-way joins. Before defining a shared join, we will describe a join across two relations. We then formally define a shared scan and then define a shared join as the join between two shared scans.

Let R and S be two relations, and $t_R \in R$ and $t_S \in S$ be tuples of the corresponding relations. A scan and select operation on the relation R is then defined as a function $\sigma^R : R \rightarrow \{\top, \bot\}$, and the output of this scan is noted as σ^R for brevity. A join on selections σ^R, σ^S of the two relations is then defined as:

Definition 1 Join

$$\sigma^R \bowtie \sigma^S = \{(t_R, t_S) \mid \sigma^R(t_R) \wedge \sigma^S(t_S) \wedge f_{\bowtie}(t_R, t_S)\}$$

□

where $f_{\bowtie} : R \times S \rightarrow \{\top, \bot\}$ is the join predicate function and (t_R, t_S) is a concatenation of the attributes t_R and t_S.

A shared join for a set of queries $Q = \{q_1, q_2, \ldots q_n\}$, where $q_i = \sigma_i^R \bowtie \sigma_i^S$ for $i \in \{1, 2, \ldots n\}$, is defined as the join between the result of the shared scans σ_Q^R, σ_Q^S. The result of a shared scan σ_Q^R can be defined as:

Definition 2 Shared Scan

$$\sigma_Q^R = \left\{\left(t_R, \left(b_{q_1}^R, b_{q_2}^R, \ldots b_{q_n}^R\right)\right) \mid \left(b_{q_i}^R = \top \iff \sigma_i^R(t_R)\right) \wedge \exists i . b_{q_i}^R = \top\right\}$$

□

Thus, a shared scan outputs intermediate relations with an extended schema that has one extra Boolean attribute $b_{q_i}^R$ for every query q_i. The attribute $b_{q_i}^R$ for a tuple t_R holds a value of true if and only if the query q_i is interested in that tuple, i.e., $\sigma_{q_i}^R(t_R) = \top$. Furthermore, a tuple t_R is output by the shared scan if at least one query is interested in t_R. The set of the attributes $b_{q_i}^R$ for all queries $q_i \in Q$ is denoted as b_Q^R and a set of values of these attributes for a particular tuple t_R is called the *set of query IDs* for t_R. Having the output of a shared scan defined, we define a shared join as the join of the output of two shared scans or:

Definition 3 Shared Join

$$\sigma_Q^R \bowtie \sigma_Q^S = \{ (t_R, t_S, (b_{q_1}^{R\bowtie S}, b_{q_2}^{R\bowtie S}, \ldots b_{q_n}^{R\bowtie S})) \mid (b_{q_i}^{R\bowtie S} = \top \iff (b_{q_i}^R = \top \wedge b_{q_i}^S = \top)) \wedge \exists i . b_{q_i}^{R\bowtie S} = \top \wedge f_{\bowtie}(t_R, t_S)\}$$

□

CID	Name	NID	Q1	Q2	Q3
1	Laura	1	1	1	0
2	Noah	1	0	1	0
3	Emma	2	1	0	1
4	Pierre	3	1	0	1
5	Marion	3	1	0	0
6	Hans	2	1	1	1

⋈

NID	Nation	Q1	Q2	Q3
1	Switzerland	0	1	0
2	Germany	1	1	1
3	France	0	1	1

CID	NID	Name	Nation	Q1	Q2	Q3
1	1	Laura	Switzerland	0	1	0
2	1	Noah	Switzerland	0	1	0
3	2	Emma	Germany	1	0	1
4	3	Pierre	France	0	0	1
6	2	Hans	Germany	1	1	1

Fig. 1 Sample shared join on attribute NID

In other words, a shared join outputs a relation with extended schema that also contains one extra attribute $b_{q_i}^{R \bowtie S}$ for each query q_i. This attribute is the result of the conjunction of the corresponding attributes of the input relations: $b_{q_i}^{R} \wedge b_{q_i}^{S}$. Similarly to the shared scan, the shared join outputs only tuples for which at least one query is interested. As for the single query join (Definition 1), the shared join (Definition 3) also filters out tuple combinations using a join predicate function $f_{\bowtie}$. The function $f_{\bowtie}$ is the same for all queries. One thing to note is that this shared join model assumes that all queries ask for the inner join. Section 5.5 discuses ways of sharing the join for queries with other join types.

An example for the input and output relations of a shared join is shown in Fig. 1. Here we show a shared join for three queries: $Q1$, $Q2$ and $Q3$, on two relations Customers ($CID, Name, NID$) and Nations ($NID, Nation$) with each query having different predicates on each relation. The upper part of Fig. 1 shows the two input relations of the join or, in other words, the output relations of the shared scan. As explained previously, intermediate relations have an additional Boolean attribute for each query, which has a value of 1 if the corresponding tuple belongs to the query or 0 otherwise. The set of query IDs in this case is the set of values of all Boolean attributes for a particular tuple. The bottom part of Fig. 1 shows the output of the shared join, where the set of query IDs of an output tuple is simply an intersection of the sets of query IDs of the matching pair of input tuples.

3.1 Query ID set representation

As explained above, shared query execution introduces an additional attribute for each intermediate tuple in the system. This attribute keeps information on which queries are interested in each tuple. There are several ways to store and handle this attribute, each with advantages and disadvantages. One way to represent this attribute is as an array of integers each of which represents the ID of the query that is interested in the tuple. The impact of this is that the size of the attribute is $N_i \cdot s_{int}$ bytes where N_i is the number of queries interested in the tuple and s_{int} is the size of the integer.

Another way of representing the set of query IDs is to use a bitset where each query in the system has a dedicated bit position in the bitset. For a certain tuple with a bitset B, and a query Q whose bit position is i if the ith bit in B is 1, then the query is interested in the tuple, and vice-versa if the bit is 0. The size of the attribute is then the size of the bitset which is $\frac{N_s}{8}$ bytes where N_s is the total number of concurrently running queries.

One factor affecting the performance trade-off between the two methods is the ratio $\frac{Avg(N_i)}{N_s}$ of average number of queries per tuple $Avg(N_i)$ to the number of concurrent queries N_s. Bitsets work well when this ratio is high or N_s is sufficiently low. Arrays work well in cases where this ratio is low and N_s is very high. In practice, we found that, for analytical workloads, bitsets perform better.

3.2 Query scheduling

Our model for sharing joins assumes a fixed set of queries during the join operation. To satisfy this property, the join runs in cycles for batches of queries where arriving queries wait in a queue if the system is busy executing a join, similarly to SharedDB. One might argue that this waiting causes response times to increase. The waiting, however, is more than compensated by allowing the system to organize the join execution in a way the whole join operation, that is both the build and probe phases of the hash join algorithm, is shared for all queries currently being executed. The result is, as we will show, higher throughput and more predictable performance.

The other alternative is to schedule queries immediately as they arrive. This is common for systems that share scans such as IBM DB2 [26,27] and MonetDB [49], and is also used by CJoin [11] and Datapath [3]. The effect on join execution is that, for a hash join algorithm, the build and probe phases need to be executed concurrently in a pipeline fashion. The problem is in the redundant work to be done for a set of concurrently running queries, which is the first stage of the pipeline, i.e., the build phase. The effect on performance depends on the relative cost between the build and probe phases and the amount of sharing missed. The effect is aggravated in cases of multi-way joins where there are multiple stages in the pipeline.

4 Choice of join algorithm

Before we introduce the algorithm for MQJoin, we analyze existing query-at-a-time join algorithms focusing on properties affected by sharing. In particular, in the model of sharing described above a key difference of the shared join compared

to single query join is the handling of additional boolean attributes represented as bitsets.

As column-store database engines have become prevalent for analytical query processing, the current state of the art in main-memory join processing is focused on joins for column-store databases with narrow tuples (around 8 to 16 bytes). It is unclear how existing algorithms will be affected by the wider tuples (reminiscent of row-oriented storage) created by the additional query IDs. In particular, for a set of 1024 queries, the bitsets will add an additional 128 bytes to each tuple significantly increasing its size. For this reason, we provide an analysis of existing hash join algorithms focusing on the impact of wider tuples.

First we analyze hash join algorithms in terms of the amount of main-memory bandwidth consumed per tuple for various key (tuple) sizes, and then, we benchmark state-of-the-art hash join algorithms. Note that the model of bandwidth consumption we use resembles cost models for in-memory join processing [25,30,33]. In this case, we are not concerned with accurately predicting the performance, but rather explaining the impact of larger tuples (keys), caused by the additional query ID bitsets, on efficient state-of-the-art join algorithms.

4.1 Memory bandwidth usage model

Join processing is inherently a data movement operation that stresses, and is often bottlenecked by, the memory subsystem. For this reason, calculating the memory bandwidth consumption is an effective way to model the upper bound and predict the join performance.

To analyze the effect of wider tuples, we first provide a model of join algorithms in terms of the amount of memory bandwidth consumed per tuple. We focus on state-of-the-art hash join algorithms, as they have been shown to have highest performance on modern architectures [5,24]. In particular, we focus on main-memory join processing in a single shared-memory system. The model consists of several components which are building blocks of hash join algorithms: building a hash table, probing a hash table and partitioning a relation.

Table 1 shows the bandwidth consumption model for various operations used in hash join algorithms. For simplicity, the model makes the following assumptions: (1) build and probe tuples are identical in size, (2) the entire tuple is read during both building and probing, (3) keys in the build relation are unique and (4) there are no hash collisions in the hash table. Note that this is just the bandwidth, not looking at the different costs of whether data are cache resident or not.

In the model, CL represents the cacheline size, t is the tuple size and m is the extra metadata stored in every hash table bucket. The first two rows correspond to building and probing a hash table out-of-cache where the CPU cache is smaller than the hash table. In this case, it is assumed that hash table accesses incur cache misses and cause one or more cacheline transfers. Building causes one cacheline read and write to access and modify the hash bucket's metadata and store a part (or the whole) tuple in the hash bucket's cacheline. Probing causes as many cacheline reads as it takes to read the metadata and the building tuple. There is an additional cost of t bytes to read an input building/probing tuple.

In the case of in-cache building and probing of a hash table, the only memory bandwidth consumed is reading the building and probing tuples. For partitioning, the model assumes that streaming stores are used, also referred to as partitioning with software-managed buffers [5].

Based on the model from Table 1 and assuming 64-byte cachelines and a 16-byte metadata structure, Fig. 2 shows bytes transferred per tuple for several hash join variants. The upper and lower plots show bytes transferred per building and probing tuple, respectively, for a hash join algorithm: (1) without partitioning where the number of bytes transferred per build and probe corresponds to the model for out-of-cache

Table 1 Model for memory bandwidth consumed per tuple for subop-erations of hash join algorithms

	Bytes read	Bytes written
Out-of-cache build	$CL + t$	CL
Out-of-cache probe	$CL \cdot \lceil \frac{t+m}{CL} \rceil + t$	0
In-cache build	t	0
In-cache probe	t	0
Partition	t	t

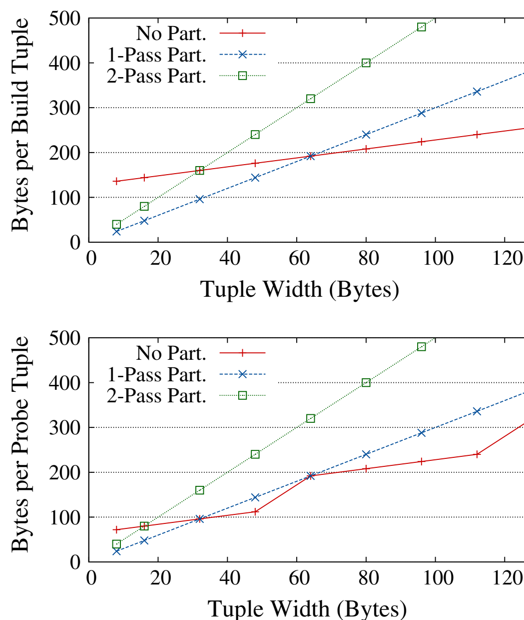


Fig. 2 Model-based memory bandwidth consumption for multiple variants of hash join algorithms and different tuple sizes

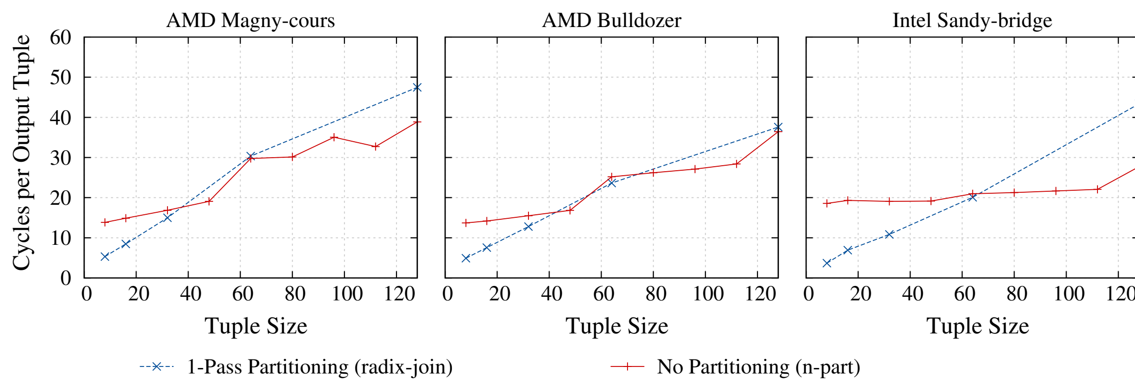


Fig. 3 Hash join performance for various tuple sizes with- and without partitioning (code taken from [5])

hash table building and probing, (2) with a single partitioning where the bytes transferred is consisted of one partition step and one in-cache hash table building and probing and (3) with two partitioning steps which corresponds to two partitioning steps and an in-cache hash table building and probing (3). Based on the results, performing an out-of-cache hash join without partitioning uses significantly more memory bandwidth than a partitioning-based approach for narrow tuples, characteristic for joins in column stores. The reason is that the out-of-cache hash table accesses transfer entire cache-lines while only using a small part of them. Partitioned hash join algorithms use sequential accesses to memory, thereby making efficient use of the bandwidth.

For larger tuples, corresponding to join operations for large join predicates (such as our query IDs) or row stores, the join without partitioning makes more efficient use of the memory bandwidth it consumes.

4.2 Experimental results

Next, we verify the conclusions from the model. We took state-of-the-art algorithms in hash join processing that contain partitioning and both in-cache join processing and out-of-cache join processing from Balkesen et al. [5].[1] We ran microbenchmark experiments with two algorithms, one without partitioning (n-part), which is an optimization of the algorithm from Blanas et al. [9] and one with partitioning (radix join). We measured the execution time when running a join of two relations, each containing 64 million tuples. The order of keys in both relations are randomized to avoid sequential accesses to the hash table. The join is a full table join and for every probing tuple there is exactly one tuple with a matching key in the building relation. The result of the join is not materialized. Instead, only the number of output tuples is counted, which is also 64 million. All available hardware threads are used for running the join and we report the number of CPU cycles spent per output tuple. This is obtained using the formula:

$$\frac{execution_time \cdot cpu_frequency}{\#output_tuples}. \quad (1)$$

We tested performance on three hardware platforms: (1) AMD Magny-Cours machine with four Opteron 6174 processors, (2) AMD Bulldozer machine with four Opteron 6276 processors and (3) Intel Sandy Bridge machine with four Xeon $E5-4640$ processors. The results shown in Fig. 3 confirm the hypothesis in the model. First, for 8-byte tuples the hash join without partitioning is slower than the partitioned one. As the size of the tuples grow, its performance curve resembles a step function that increases whenever the number of cachelines needed for a building and probing operation increase. The performance of the partitioned hash join, on the other hand, is proportional to the tuple size and for significantly wide tuples of more than 64 bytes it performs slightly worse than its counterpart. Since we aim at running several hundred queries in a single shared join, the space we require for a tuple is far more than a cacheline. Hence, for our MQJoin algorithm, we opted for an algorithm that does not partition the input relations.

5 Two-way join algorithm

We faced two key challenges when designing MQJoin: Minimize the time spent per tuple and minimize the number of tuples processed for a set of queries. To address the first challenge, we combine approaches from related work on optimizing joins with techniques to efficiently reuse data structures over multiple join sessions and to minimize the overhead imposed by sharing, such as handling query IDs. To address the second challenge, we use techniques that schedule queries in a way that minimizes redundant work and develop ways to share execution of queries that require different types of joins.

[1] Code can be downloaded from: https://www.systems.ethz.ch/node/334.

5.1 Algorithm overview

From a high level perspective, the algorithm is a parallel hash join running on a single multi-core machine based on state-of-the-art hash join algorithms [6,9,13]. During the build step, multiple threads consume the build relation to populate the hash table. In the next phase, the threads consume the probe relation and probe the hash table to find matches of tuples. Unlike a traditional hash join algorithm, the threads do an additional step of computing the intersection of the query ID sets of all matching pairs of tuples and filtering out tuples with an empty intersection.

The algorithm we use inherits several features from recent work on main-memory hash joins. Similarly to Blanas et al. [9], threads during the build step synchronize using spin locks, where there is one lock per hash entry in the table. Similarly to Balkesen et al. [6], we optimize the algorithm by minimizing the number of random accesses per tuple. As discussed in the previous section, we do not partition the input relations so that they fit in cache. Instead, the hash table is built and probed directly from main memory. To reduce the latency of random accesses, we use a grouped software prefetching technique as proposed by Chen et al. [13,14]. One novelty in our approach, relevant also to query-at-a-time processing, is the introduction of a *session ID* attribute to each hash entry to provide an efficient reset operation of the hash table.

5.2 Hash table structure

The hash table is structured in a way that each bucket is aligned at the $64B$ cacheline boundary, guaranteeing that each hash bucket lookup will access a single cacheline. The structure of a bucket is shown in Fig. 4. A hash bucket consists of the following fields: **Lck: Lock** is used to synchronize between threads during building of the hash table; **SID: session ID** is used to identify the last session when this hash bucket was updated. This is needed in order to reuse the memory of a hash table for multiple join cycles without the need of an expensive `memzero` operation; **Record Ptr** points to the address in memory where the record is located. This can be either in the buffer space of the hash bucket or somewhere else; **Query ID Set Ptr** points to the address in memory where the set of query IDs for the tuple are located; **Next Bucket Ptr** points to the next hash bucket in cases of overflow. Each thread has its own dedicated pool of overflow buckets; **Key: Join Key** is the attribute(s) used in the join predicate that is cached in the hash bucket for quick access. The hash bucket structure shown in Fig. 4 corresponds to an example where the key is a 4-byte integer.; **Buffer space** is the extra memory located on the cacheline that is used to store the record and/or the query ID set in cases when they are small enough. In this example where the key is 4 bytes, the buffer space is 32 bytes.

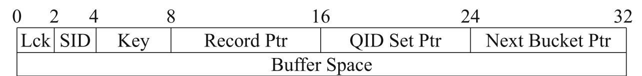


Fig. 4 Structure of a hash bucket

Algorithm 1 Build Phase

```
for group ∈ relation do
S1 { for tuple ∈ group do
       bucket ← COMPUTEBUCKETADDRESS(tuple)
       PREFETCH(bucket)
     end for
S2 { for tuple ∈ group do
       LOCKBUCKET
       if bucket.SID != currentSID then
         POPULATEBUCKET(tuple, bucket)
         bucket.SID = currentSID
       else
         ofbucket ← GETOVERFLOWBUCKET
         SWAPNEXTBUCKETPTRS(bucket, ofbucket)
         POPULATEBUCKET(tuple, ofbucket)
       end if
       UNLOCKBUCKET
     end for
end for
```

5.3 Join procedure

Next we explain the build and probe phases of the hash join algorithm in more detail. For clarification purposes, we also provide an example shown in Fig. 5 and work through the example as we explain the algorithms. The figure shows a simple setup with a building relation of 5 tuples, probing relation with 10 tuples and the populated hash table.

Both the build and probe algorithms divide their input intermediate relations in small groups and iterate over these groups multiple times before proceeding to the next group. The reason for this is to avoid high memory access latencies to memory locations not present in the cache. In each iteration, addresses to non-cache resident memory to be accessed in the next iteration are calculated and prefetch instructions are issued. We experimented with different group sizes and found that groups of about 50 tuples are enough to hide random main-memory access latencies. For simplicity, in the example in Fig. 5 we use group sizes of 5 tuples, so there is only one group for the building relation and two groups for the probing relation.

The **build phase** (Algorithm 1) consists of two iterations over the small groups of tuples. In the first iteration (S1), the hashes of the join keys are precomputed and prefetch statements to the addresses of the corresponding hash buckets are issued. In the second iteration (S2), the hash buckets are

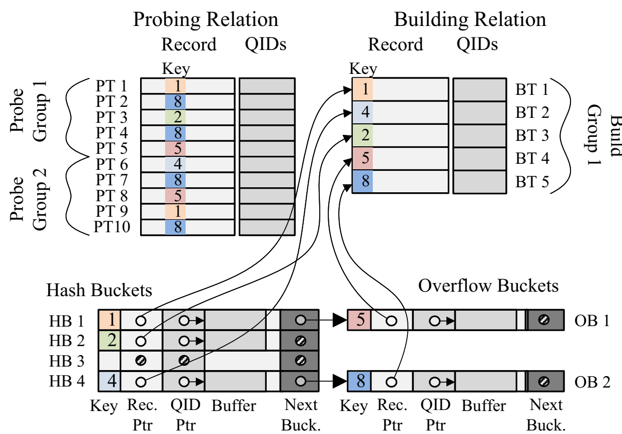


Fig. 5 Two-way join algorithm example

populated with the input tuples. If a hash bucket is already populated, the input tuple is populated in an overflow bucket and the pointer of the hash bucket is updated to point to the overflow bucket.

The input tuples to the join consist of pointers to the corresponding query ID bitset and the record. The query ID bitset for each tuple is generated either by a shared scan operator or by a probe part of a shared join operator. When populating a hash bucket with an input tuple, the bitset and/or record is copied into the buffer space of the hash bucket if they fit.

For the example join, in the first step key hashes of tuples BT{1,2,3,4,5} are computed and the prefetch instructions are issued on the hash buckets HB{1,4,2,1,4} in the corresponding order. In the second step, building tuples BT{1,2,3} are populated in hash buckets HB{1,2,4}, respectively, while tuples BT{4,5} are populated in new overflow tuples OB{1,2} that are linked to hash buckets HB{1,4}. In this example, the buffer space in the buckets is enough to hold the QIDs bitset, but not enough to store the record as well. Therefore, only the bitsets are copied into the buffer space of buckets and QID pointers are set to point to the buffer space, while the record pointers are set to point to the original memory of the input relation.

The **probe phase** (Algorithm 2) is a bit more involved. Step 1 (S1) computes hashes and prefetches the hash buckets. Step 2 (S2) evaluates join predicates and prefetches the set of query IDs and the records. If the set of query IDs and the record span multiple cachelines, it issues a prefetch statement for each cacheline. In case of overflow, it also issues a prefetch statement to the overflow buckets. Step 3 (S3) computes the set intersection of the query ID sets and materializes the output tuple if the set intersection is not empty. The output tuple is projected to contain the union of all attributes requested by the batch of queries that is currently being executed. In cases of overflows, the procedure from step 2 to step 3 is repeated as many times as the length of the longest bucket chain.

Algorithm 2 Probe Phase

```
     for group ∈ relation do
S1  {  for tuple ∈ group do
           bucket ← COMPUTEBUCKETADDRESS(tuple)
           PREFETCH(bucket)
       end for
       while group != [ ] do
           otherGroup ← [ ]
           keyMatchGroup ← [ ]
S2  {      for tuple ∈ group do
               if bucket.SID == currentSID then
                   if bucket.key == tuple.key then
                       ADDTOKEYMATCHGROUP(tuple)
                       PREFETCH(bucket.queryIDs)
                       PREFETCH(bucket.record)
                   end if
                   if HASNEXTBUCKET(bucket) then
                       ADDTOOTHERGROUP(tuple)
                       PREFETCH(bucket.nextBucket)
                   end if
               end if
           end for
S3  {      for tuple ∈ keyMatchGroup do
               resQIDs ← INTERSECT(tuple.QIDs, bucket.QIDs)
               if !EMPTYSET(resQIDs) then
                   OUTPUTTUPLE(tuple, bucket, resQIDs)
               end if
           end for
           SWAP(group, otherGroup)
       end while
     end for
```

In the example in Fig. 5, the probing procedure for the first probing tuple group will be the following. In the first iteration (S1), key hashes of probing tuples PT{1,2,3,4,5} are computed and prefetch instructions are issued on the memory addresses of hash buckets HB{1,4,2,4,1} in the corresponding order. In the next iteration (S2), key comparison of corresponding tuples and hash buckets are performed and the record and QID pointers are prefetched for the buckets with matching key comparison (HB{1,2}). In the same iteration (S2), prefetch instructions are also issued for the overflow buckets OB{1,2}. In the next iteration (S3), QID set intersection is performed and output tuples are materialized for tuple pairs: {(BT1,PT1), (BT3,PT3)}. Since there are overflow buckets, steps S2 and S3 are repeated for the overflow buckets. In the next iteration (S2), key comparison is performed for tuple–bucket pairs: {(OB2,PT2), (OB2,PT4), (OB1,PT5)}, and since all keys match, record and QID pointers are prefetched for both overflow buckets OB{1,2}. No overflow buckets are prefetched in this S2 iteration since all next bucket pointers are null. Finally, the algorithm proceeds with the last iteration of the first probing group of tuples (S3) with performing set intersection and output tuple materialization of tuple pairs: {(BT5,PT2), (BT5,PT4), (BT4, PT5)}. As the algorithm processed all overflow buckets, it exits the while loop and continues on with the next group of probing tuples in a similar manner.

5.4 Discussion

MQJoin is similar to CJoin and Datapath in using bitsets to handle the query ID set for each tuple, as well as the use of a hash-based algorithm.

An important difference to CJoin and Datapath is that in those systems the build and probe phases are executed concurrently. As a result, the hash table is constantly updated for all incoming and outgoing queries, which introduces extra work per query. In CJoin, for instance, the hash table is updated for each query individually, both when the query enters and exits the system. This works for star schemas and low concurrency cases where the effort required in the build phase is significantly lower than during the probe phase. In Datapath, colliding hash table entries from newly arrived queries will be placed in the next available entry in the hash table. This causes extra work to be done during probing and it is unclear how the hash table is purged when queries finish. To minimize the work required for building, updating and clearing the hash table, our join algorithm runs in cycles, performing build and probe one after the other in each cycle. This allows us to share the build operation for all queries that are being executed in the current cycle. At the end of each cycle, we clear all data in the hash table by simply incrementing the session ID number. This avoids replaying build subqueries to clear data from the hash table.

Another difference is in the microarchitectural properties of the algorithm. SharedDB uses single-threaded join operators which is inefficient for analytical workloads on multi-core systems. Similarly, CJoin builds and updates the hash table in a single thread, which is inefficient if the build relations are not insignificantly small. This is another reason why CJoin is restricted to star schemas. Datapath uses a single hash table for all joins which is divided in 64 regions with exclusive locks. To avoid contention on these locks, it needs to update the hash table in two phases. Our algorithm parallelizes both build and probe phases and synchronizes build operations on a per-bucket basis. We optimize for both CPU and bandwidth efficiency by hiding random access latencies through software prefetching and minimizing the number of random cachelines accessed per tuple. Thus, the performance MQJoin is comparable to that of the state-of-the-art query-at-a-time join algorithms, something that none of the competing versions can do.

5.5 Sharing execution for other join types

The algorithm we just described works for queries that require inner joins. It can be extended to share execution for queries that also require (left, right, full) outer joins and (anti) semi-joins, provided that they are all on the same equality predicate. Consider a simple schema of relations $S(Aint, Bint)$ and $R(Aint, Cint)$, and a shared join operator which builds the hash table with the relation S and probes with the relation R. Then consider the set of queries shown in Fig. 6. To handle these queries, the algorithm is extended to perform additional operations on the bitsets. These additional operations depend on the join type and can be setting bits of individual queries to 1, conditional set to zero, or bit-wise OR. This extension allows to answer the queries 2, 4, 5, 6 and 7 from Fig. 6. Another extension is to modify tuples' bitsets in the hash table during probing, and iterate over the build relation again after probing to output tuples which did not have a match in the probing relation. This extension allows to answer the queries 3, 4 and 8.

```
1. SELECT * FROM S, R WHERE S.A = R.A
2. SELECT * FROM S LEFT OUTER JOIN R ON S.A = R.A
3. SELECT * FROM S RIGHT OUTER JOIN R ON S.A = R.A
4. SELECT * FROM S FULL OUTER JOIN R ON S.A = R.A
5. SELECT * FROM S WHERE S.A IN (SELECT R.A FROM R)
6. SELECT * FROM S WHERE S.A NOT IN
                                (SELECT R.A FROM R)
7. SELECT * FROM R WHERE R.A IN (SELECT S.A FROM S)
8. SELECT * FROM R WHERE R.A NOT IN
                                (SELECT S.A FROM S)
```

Fig. 6 List of queries whose join can be shared

The specific additional bitset operations and modifications are shown in Table 2. In particular, it includes the needed operations to join execution of standard inner joins, left (and right) outer, semi and anti joins. Full outer joins would contain the modifications for both the left and right outer joins. To distinguish left and right relations, we assume that the hash table is built with the left relation and probed with the right relation. Furthermore, we distinguish cases where the join attributes in the left (building) relation are unique or not.

The columns in the table correspond to the following bitset operations:

– the output query ID bitset during probing when a matching tuple is found;
– the output query ID bitset during probing when a matching tuple is not found (for a building relation with unique keys) or when finished with a probing tuple (for a building relation with non-unique keys);
– the modification of the left (building tuple) or right (probing tuple) bitset when a matching tuple is found during probing;
– the output query ID bitset when the left (building) relation is scanned again after probing.

The bitsets in the formulas are constructed using the following abbreviations:

– L and R correspond to the input query ID bitsets of the left (building) and right (probing) tuple, respectively;

Table 2 Bitset operations for sharing execution of different join types

			During probing			Output bitset on post-probe left side scan
			Output bitset on match	Output bitset on non-match; after probe	Left; right bitset modification	
Inner join			$(R \& L)$	0; 0	L ; R	
Right	Unique	Outer join	$(R \& L) \mid ROJ$	ROJ; 0	L ; R	
		Semi join	$(R \& L)$	0; 0	L ; R	
		Anti join	$(R \& L) \& !RAJ$	RAJ; 0	L ; R	
	Non-unique	Outer join	$(R \& L)$	0; $!(R \ll 1) \& ROJ$	L ; $R \mid ((L \& R \& ROJ) \gg 1)$	
		Semi join	$(R \& L)$	0; 0	L ; $R \& !(R \& L \& RSJ)$	
		Anti join	$(R \& L) \& !RAJ$	0 ; $R \& RAJ$	L ; $R \& !(L \& R \& RAJ)$	
Left		Outer join	$(R \& L)$	0; 0	$L \mid ((L \& R \& LOJ) \gg 1)$; R	$!(L \ll 1) \& LOJ$
		Semi join	$(R \& L)$	0; 0	$L \& !((L \& R) \& LSJ)$; R	
		Anti join	$R \& (L \& !LAJ)$	0; 0	$L \& !((L \& R) \& LAJ)$; R	$L \& LAJ$

– ROJ, RSJ, RAJ, LOJ, LSJ, LAJ correspond to bitsets which hold the information of the particular type of each query. For instance, if the join for a query with an ID 3 is a right outer join, the third bit in the ROJ bitset will be set to 1. These six bitsets are specific to each join operator and are static for a set of queries.

In general, most of the bitset operations are composed of simple bit-wise `and` (&), `or` (|) and `not` (!) operations. Exceptions are the left shift ($\ll$) and right shift ($\gg$) operations present in left outer join and right outer join with non-unique keys in the building relation. In this case, the system needs to store an extra bit of information per tuple and query that can be achieved by allocating one more bit in the bitset for queries with joins of this type. In particular, for the left outer join, the adjacent bit is used to identify tuples in the building relation that had a match in the probing relation. The same mechanism is used for the right outer join.

The corresponding instructions of the bitset operations used (and, or, not, left shift and right shift) are all simple and are executed efficiently on modern processors. Therefore, they would not cause significant performance degradation in the case when the join is shared for queries with other non-inner join types.

5.6 Further optimizations

The approach of MQJoin can also benefit from other techniques that optimize execution of single query join operation. For instance, bloom filters can be used to perform early filtering of tuples during the probing phase. This technique is used by several systems, including IBM DB2 [40], VectorWise [43], SQL Server [29] and HyPer [28]. Bloom filters provide the benefit of reducing expensive accesses to non-cache resident hash tables at the cost of probing the bloom filter for every tuple. Bloom filters for in-memory hash joins are beneficial when the join is highly selective and the hash table does not fit into CPU caches. In MQJoin, the input (and output) relations represent the union of input (and output) relations for all concurrent queries, meaning that hash table is less likely to fit into CPU caches, but the selectivity of the shared join operation is also lower compared to a single query join.

Another optimization is to use auxiliary index data structures, such as join indexes [36,48] or indexes on join attributes (for the purpose of using indexed nested loop join). These data structures may reduce the execution time of join operations at the additional maintenance cost when handling updates. Shared execution of joins is also possible in the presence of indexes. Similarly as the bloom filters, the lower selectivity of a shared join compared to a single query join would affect the decision of the optimizer whether to use indexes.

6 Multi-way joins

In this section, we describe how multiple join operations are handled. As with the query-at-a-time approach, the shared join approach also requires an optimization decision on how to create query plans involving multiple joins. The shared join optimizer not only needs to decide the order of a multi-way join but also which queries' execution should be shared. It can be undesirable to share the execution of some queries either for performance isolation reasons or if sharing would hurt overall performance.

Building such an optimizer is actually a non-trivial task and is out-of-scope of this paper. Recent work by Giannikis

et al. [17] has addressed this problem and these techniques can also be applied to our approach. In this paper, we focus on one end of the spectrum, that is, to maximize sharing across all concurrent queries. While this might not be always optimal or desirable, it provides a lower bound for the performance of our approach. The ordering of the joins is currently done by hand. Next we explain how we maximize sharing for queries with multi-way joins. The key challenge to address is handling queries without common subplans. In this regard, we use two techniques: query plan equalization and global query batching.

6.1 Query plan equalization

A key property of our shared join approach that minimizes the number of tuples processed in the join operators is that a join operation is shared even for queries without common subplans. We illustrate this with an example. Figure 7 shows a sample database with three relations: Orders, Customers and Nations together with 3 queries with various joins and predicates on the three relations. The three queries share almost no common subexpressions: Q1 asks for all the orders from John, Q2 asks for all orders after December 21, 2012, from USA, and Q3 asks for all customers from Germany. The query execution plan that we create to execute all 3 queries together is shown in Fig. 7c.

Relation Orders:

OID	CID	Date
1	2	2012.JUL.01
2	3	2013.JAN.01
3	1	2011.NOV.31
4	3	2013.FEB.01
5	2	2012.NOV.15
5	2	2012.DEC.14

Relation Customers:

CID	Name	NID
1	John	1
2	Maria	1
3	Dieter	2

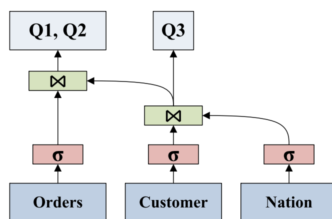
Relation Nations:

NID	Name
1	USA
2	Germany

(a)

Q1	SELECT * FROM Orders O, Customers C WHERE O.CID = C.CID AND C.Name = 'John'
Q2	SELECT * FROM Orders O, Customers C, Nations N WHERE O.CID = C.CID AND C.NID = N.NID AND O.Date > 2012.DEC.21 AND N.Name = 'USA'
Q3	SELECT * FROM Customers C, Nations N WHERE C.CID = N.NID AND N.Name = 'Germany'

(b)



(c)

Fig. 7 Sample database with queries. **a** Sample database. **b** Sample queries. **c** Shared join plan for Q1, Q2, Q3

One can notice that the organization of join operations in the plan does not directly correspond to the join operations required by all queries. In particular, Query 1 does not require the join of Customers ⋈ Nations; however, it is still included in its plan. The reason is that it allows for the Orders ⋈ Customers join to be shared for Query 1 and 2.

We considered the two following techniques to exploit this type of sharing opportunity: (i) modify the query plan of Query 1 to include an unnecessary full table scan of Nations followed by a join with Customers and (ii) have the Customer ⋈ Nations join always forward tuples for which Query 1 is interested. Both methods have advantages and disadvantages, and could be used interchangeably depending on the situation. The advantage of the first method is that it requires less computation per tuple; however, it might process more tuples. This would not be a problem, if for instance a full table join is already required by the union of all joins that are processed. The advantage of the second method is that it might process less tuples, but would require more computation per tuple. Currently we modify the query plans to include extra full table scans and joins. This technique can be referred to as query plan equalization where query plans are modified to increase sharing opportunities. The approach differs the ones in Datapath and SharedDB where, for join queries with different subplans, the join operations are either replicated or process intermediate relations from separate query plans, thereby creating redundant work.

6.2 Global query batching

As explained earlier, to exploit more sharing opportunities, our algorithm requires a fixed set of queries during its execution. This means that queries or subqueries need to be coscheduled (or batched), so that they are executed together. To facilitate this, we use query queues where there is one queue for each class of queries or subqueries that need to be coscheduled. If the system is busy executing a certain class of queries, arriving queries of that class will wait in the queue. The decision of how many queues to have involves making a trade-off between the amount of sharing exploited and isolation of query performance. To show the effects of sharing compared to query-at-a-time execution, we focus on one end of this spectrum which maximizes sharing. Therefore, we use a single query queue which gives the system maximum flexibility in how to schedule execution of concurrent queries. We refer to the use of a single query queue as global query batching. The effect of this can also be shown with the example in Fig. 7. Assume that there are many clients connected to the system where each client repeatedly executes one of the three queries at random. Although the instances of Query 3 do

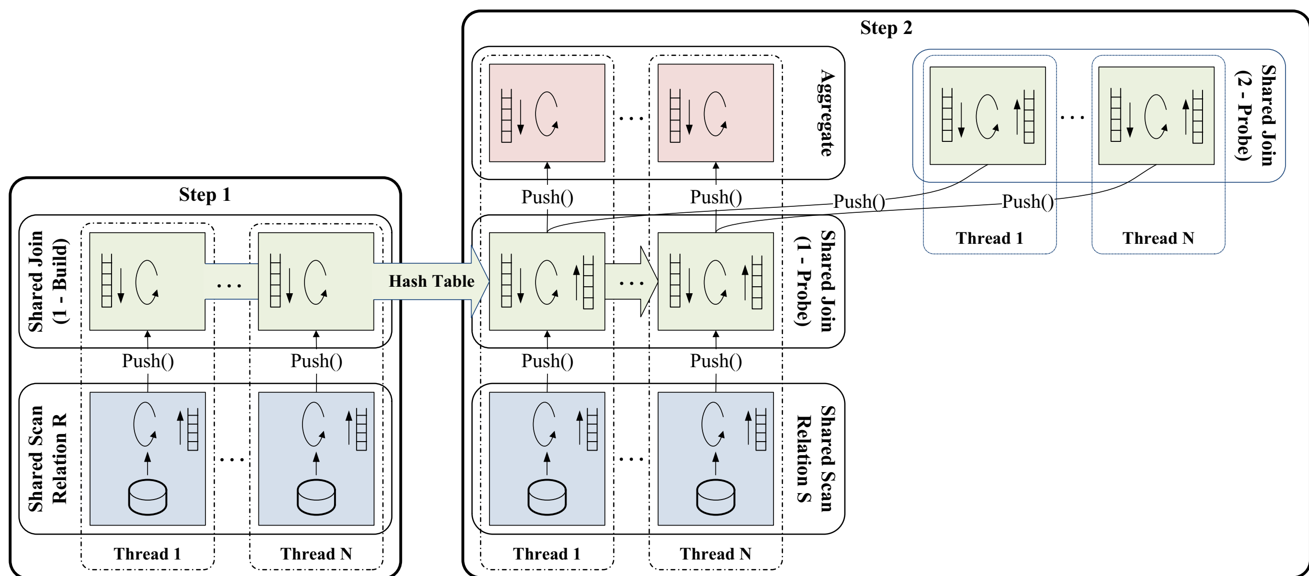


Fig. 8 Integration with shared scans

not require the Orders⋈Customers join like Query 1 and 2, all queries will be coscheduled together. This means that query execution can be organized in a way that the Customers⋈Nations join is executed only once for all queries that are being executed at the same time, minimizing redundant work.

7 System architecture

For an efficient end-to-end query execution, a shared join must run on top of a storage engine that supports shared scans. Sharing computation and bandwidth in scans is a common technique used in many systems. Some examples include: Blink [39], MonetDB [49] and Crescando [16]. To provide a clearer picture of the performance of MQJoin when it runs as part of a complete system, we integrated it with Crescando. Crescando is a row-oriented storage engine where relations are partitioned across cores and fully reside in main memory. Therefore, neither the scan nor the join need to fetch data from disk to execute queries.

An example of how this integration works is shown in Fig. 8. This example depicts two shared scan operations, one join operator divided into build and probe parts, an aggregate operator, and the probe part of another join operator.

Query execution in our architecture is performed in one or more steps. In each step, a set of threads work on separate partitions, and execution progresses to the next step only when all threads are done processing their partition. In the example in Fig. 8, there are two steps, where in step 1, n threads scan the R relation and build a hash table, and in step 2 they scan the S relation, probe the hash table, perform an aggregation and probe another hash table (the building of this hash table is not depicted).

For optimizing data and instruction cache locality, tuples in each step are processed in vectors in a pipelined fashion similarly to MonetDB/X100 [10]. Within a step, threads switch contexts between operations when they fill up a buffer of 1000 tuples. For instance in step 1, thread 1 will scan its partition of relation R until it fills up the buffer with 1000 tuples. Consequently, it will push the buffer of tuples to the operator on top, which in this case is a function call to the build operator.

We use a push-based query processing pipeline where data producing operators call a *Push()* function of the consumer operators. This is unlike the pull-based approach of Volcano [18] where consumer operators perform a function call *next()* of the producer operator. This push-based approach allows for intermediate results to be consumed by multpile operators and still be processed in a pipeline fashion without the need for materialization. This is shown in the example in Fig. 8 where both the *aggregate* and *shared join (2 - Probe)* operators consume the result of the *shared join* $(1 - Probe)$ in a pipeline fashion.

7.1 Interpreting the results of the shared join

The shared join outputs intermediate results in a similar way as it gets the input, by including a bitset to every tuple representing the query ID sets. These intermediate relations can be used by any other operator or can be sent directly to the clients. In the absence of mechanisms to support the execution of more complex queries, we only add a simple aggregate operator on top of the join operator, which can evaluate arbi-

trarily complex expressions for each query individually. This aggregate operator iterates over the bits in the bitset for each query, evaluates the expressions of the corresponding queries and updates their states. This way we avoid sending large materialized relations over the network.

7.2 Tuple format

Since we built MQJoin on top of a row-oriented storage engine, the system uses a row-wise format (NSM) throughout the whole query execution process. Nevertheless, MQJoin could also be integrated with column-oriented storage engines that use a column-wise format (DSM) to store relations. In this case, we would need to employ an on-the-fly conversion between DSM and NSM to be able to benefit from both formats. Zukowski et al. [50] showed that this kind of conversion can be efficiently implemented and that it enables significant performance improvement for in-memory analytical query processing.

7.3 Handling transactional and analytical workloads

While the description of the system so far focused on static datasets, the design principles of MQJoin, such as not relying on index data structures for query processing and batch scheduling of queries, make it suitable for handling large update-intensive transactional workloads in addition to the read-only analytical workloads. Handling of both transactional and analytical workloads is possible with a primary–secondary form of replication where the primary replica is dedicated for transactional workloads and the secondary replica is used by the system described so far to handle analytical workloads. Additional techniques such as lightweight update extraction and propagation, and batch scheduling of queries and updates provide high performance and minimum load interaction between the transactional and analytical workloads, as well as high level of data freshness and snapshot isolation consistency guarantees for the analytical queries [31].

7.4 Main-memory footprint

One may argue that sharing the execution of multiple join operations increases the main-memory requirements of a database. The reason behind is based on a traditional trade-off between the number of concurrent queries and the available memory. While our shared join does take more memory than a single join, we run many queries through it and exploit the sharing opportunities that arise, thereby reducing the overall demand for memory when considering how many queries are being executed concurrently.

7.5 Scaling out

As main memory gets cheaper and larger, an increasing number of datasets can fit in the memory of a single machine. For this reason, our algorithm assumes that relations and intermediate data structures are memory resident. Nevertheless, although our system is designed for single-node join processing, the techniques we use can also be applied in a distributed setting. As a memory intensive operation, network bandwidth is a limiting factor when running a join across multiple machines. Therefore, our approach of sharing the bandwidth for multiple queries will also be beneficial in this case. Similar rationale can also be applied to disk-based joins that spill intermediate relations to disk.

8 Evaluation

To evaluate the performance of MQJoin, we first run a series of microbenchmarks where we investigate the microarchitectural effects of sharing the join. We then evaluate MQJoin running on top of a shared scan with a TPC-H-based scan and join workload and compare the performance to main-memory, column-store databases optimized for analytical workloads such as TPC-H, namely VectorWise [51] and System X. We also compare our approach with CJoin using the star schema benchmark and based on the code provided by the authors of CJoin. Unless otherwise noted, the experiments were done on a machine with 4 × twelve-core AMD Opteron 6174 'Magny-Cours' processors clocked at 2200 MHz. The machine has 8 NUMA nodes each with 16 GB of memory, for a total of 128 GB of RAM.

8.1 Microbenchmarks

The purpose of these microbenchmarks is to show how sharing affects the performance of a join between two relations. We evaluate performance and compare it to a query-at-a-time join for various relevant factors, such as number of concurrent queries, hash table size, tuple size and workload type. All experiments refer to an equi-join between relations $R(int\ A, int\ B)$ and $S(int\ A, int\ C)$. Unless otherwise noted, both relations have 100 million tuples, each of which is 8 bytes, where the first 4 bytes contain the join key. The join key ranges from 1 to 100 million and is randomly distributed across tuples in both relations. The relationship between R and S is a primary key–foreign key relationship; thus, the join key in R is always unique. There is no skew in the workload, so keys in S are evenly distributed. The reason this workload is chosen is that it resembles workloads used to evaluate query-at-a-time join algorithms in related work on joins [6,9,24], giving us a fair base for comparison.

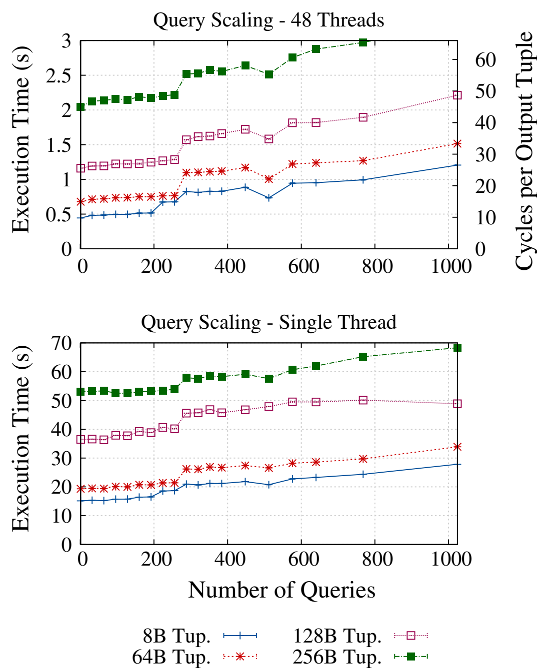


Fig. 9 Performance of a 100 M × 100 M MQJoin for various number of queries and record size

8.1.1 Scaling with the number of concurrent queries and record size

In the first experiment (Fig. 9), we measure absolute performance of our join algorithm in isolation and investigate how performance is affected by the size of the bitset, i.e., the number of concurrently running queries, as well as, the record size. We measure the time it takes to execute all join queries while varying the number of queries. To keep the number of tuples constant and to minimize the effect of the scan, all queries ask for a full table join of the two relations. We show performance for two cases: one where the join runs on 48 cores and one where it runs on a single core. This indicates how the algorithm performs in both memory-bound and compute-bound scenarios.

The first thing to note is the absolute performance of the algorithm compared to state-of-the-art join algorithms. From the 48-thread graph, we see that the maximum performance obtained for small-sized tuples and few queries is a little bit less than 0.5 s, which for our 100 M × 100 M join corresponds to around 200 million tuples per second. Next, we analyze the performance effect of bitset and record size. Dealing with bitsets is an important source of overhead as it is not present in query-at-a-time join algorithms. We note that the difference in performance of sharing join execution for around 500 concurrent queries instead of 1 is in general small and at most a factor of 2, the main reason being that 500 bits can still be accommodated in a single cacheline of 64 bytes.

The overhead of dealing with larger record sizes is also important, since queries might be interested in different attributes requiring for larger records to be projected and processed by the join operators. Similarly to the bitset size, the results show that increasing the record sizes from 8 bytes to a cacheline size of 64 bytes has a marginal overhead. Enlarging the records to sizes bigger than a cacheline of up to 256 bytes, however, adds a significant overhead and performance starts degrading linearly with the record size as more non-cache resident memory has to be accessed per probe operation. To avoid this type of overhead, several techniques can be used. One way is to employ standard techniques used in current databases to avoid processing large records in the join such as data compression and late materialization. Another way is to compress individual records and have record-specific projection using only the attributes that are of interest to the queries the record belongs to. This technique prevents the increase in record size at the expense of having more complex data-dependent code. In our case, we found that such techniques are not necessary, since for the workloads we used the tuples did not exceed 64 bytes. And as mentioned before, the impact on performance in this case is negligibly small.

8.1.2 Effect of hash table size

The join is an operation which scales supralinear with relation size. Typical breaking points are when the hash table no longer fits in cache or no longer fits in main memory. When sharing a join, the input relations are a union of all relations required by the queries. Thus, knowing how exactly does a join scale with the size of a hash table is important to understand the effect of sharing the join.

In this paper, we focus on main-memory databases. Thus, we consider only the cases when a hash table fits in main memory. We vary the size of relation R from 1000 tuples to 100 million tuples which covers the cases from when the hash table fits in `L1` cache until it is much larger than `L3` cache. As before, we measure performance of full table joins for a join on 1 core and 48 cores. Since the size of the build relation is not constant, we only measure the performance of the probe operation. We take measurements for 3 different join cases. The first one is a query-at-a-time join for which we used our join algorithm without sharing support. The second one is a shared join with only few queries (<64). Finally, the third one is a shared join with 512 queries which is already enough to feel the impact of the bitset size.

The most important thing to get from these graphs is the ratio between the lowest performance of the shared join and the highest performance of a query-at-a-time join. The reason this is important is that it depicts a worst-case scenario where the hash tables of each individual query fits in cache, but the union of all hash tables does not fit in cache. The highest

ratio in this case is around 6, which means in the worst case a probe operation will cost 6 times more for a shared join. However, it is important to note that since this corresponds to shared execution of 512 concurrent queries, the extra cost is compensated by the sharing.

Another observation to make is that the single-threaded case is less sensitive to hash table size, and does not experience a performance drop as the hash table grows larger than the L3 cache. This means that the software prefetching technique we use is able to successfully hide the large random main-memory access latencies which occur when each hash table access is a cache miss.

8.2 Performance on Xeon Phi Knights landing

As discussed above, a disadvantage of shared query execution is that the working set size, which in the case of join is the hash table, is larger when sharing. The benchmark results in Fig. 10, which use a standard CPU with a cache of few dozen megabytes, show that probe performance degrades when the building relation is about 100 thousand tuples.

This problem can be alleviated with modern microarchitectures such as the Xeon Phi Knights Landing (KNL) [46] that feature a high bandwidth on package memory called multi-channel DRAM (MCDRAM) in addition to standard DDR4 memory. MCDRAM in the KNL chips contains capacity of up to 16 GB and, according to Intel, 4x higher performance than DDR4 memory. In our experiments, we were able to verify this for a system with Xeon Phi KNL 7210 with 64 Cores (256 hardware threads) @ 1.3 GHz with 16 GB of MCDRAM and 196 GB of DDR4. In particular, with a simple bandwidth measuring benchmark with 256 threads, we measured around 76 and 320 GB/s sequential access bandwidth to the DDR4 and MCDRAM memory, respectively. The measured random access bandwidth was 68 and 240 GB/s, respectively. Thus, the MCDRAM can be seen as a last-level (software-managed) cache.

In the case of sharing hash joins, this means that the problem of larger hash tables is less of a concern since we now have a fast cache of 16 GB which is almost 3 orders of magnitude larger than standard CPU caches. We tested this effect by benchmarking our MQJoin algorithm on the Xeon Phi system. In particular, we varied the hash table size, and measured probing speed for query-at-a-time join, sharing with a single query and sharing with 512 queries. We performed the same experiment in two cases: (1) when the hash table resides in the normal DDRAM memory and (2) when it resides in the MCDRAM memory.

The results of this experiment are shown in Fig. 11. The upper and lower plots depict the cases when the hash table resides in standard DDRAM and MCDRAM memory, respectively. From the results, we can see the benefits of MCDRAM where the probing performance starts degrading for significantly larger build relation size, or about 3 orders of magnitude as also suggested by the size of the

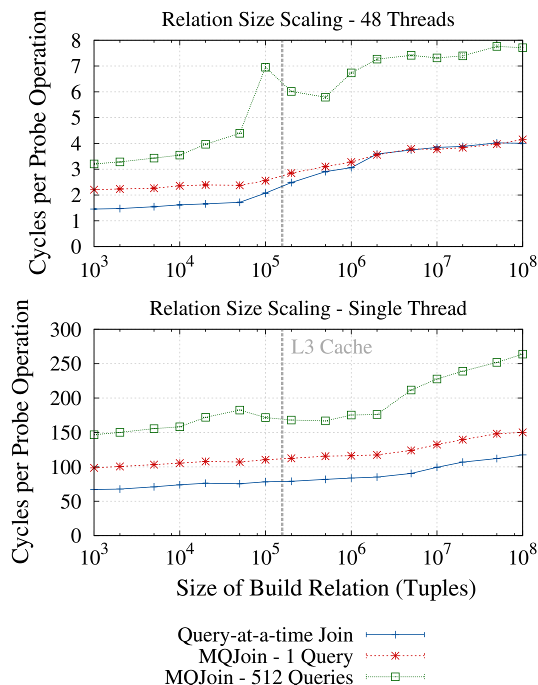


Fig. 10 Probe performance versus build relation size

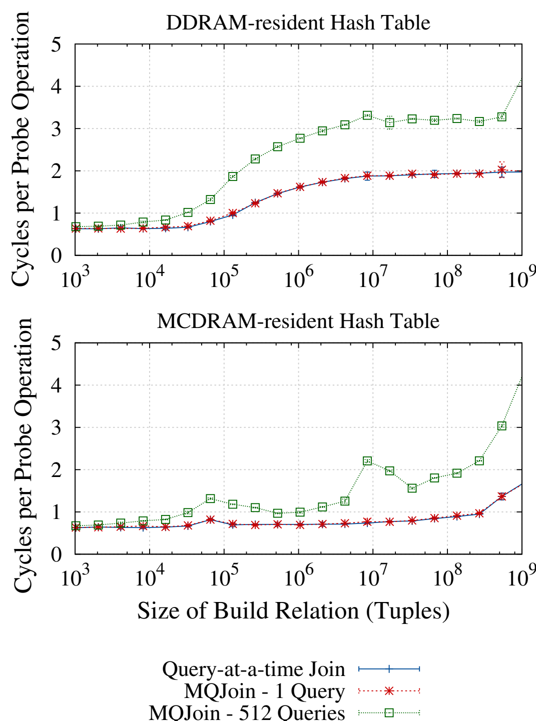


Fig. 11 Probe performance versus build relation size on Xeon Phi Knights landing

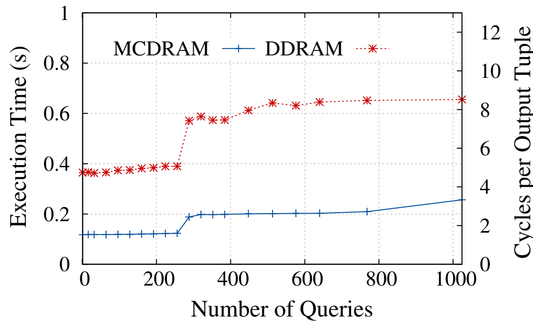


Fig. 12 Performance of a 100 M × 100 M MQJoin on Xeon Phi Knights landing on for various number of queries and hash table storage mediums

MCDRAM memory. Hence, we can conclude that for these emerging platforms the increasing working set sizes imposed by shared execution of queries is less of a concern. One thing to note is the less predictable performance in the MCDRAM case exhibited by isolated performance drops for 10^5 and 10^7 tuples. This may come from the interaction between variations in memory placement and the complex NUMA architecture of KNL with a 2D mesh interconnect.

We also show the general performance of MQJoin on this platform where we vary the number of queries for a 100Mx100M tuple join with 8B tuples. We also show the performance when the hash table resides in DDRAM or MCDRAM. The results are shown in Fig. 12. In particular, for a single query join with a hash table in the MCDRAM MQJoin is able to achieve a speed of more than 800 million tuples per second.

8.2.1 Effects of sharing the join

In the following experiment, we illustrate the effect of sharing the join for queries with predicates. We vary the selectivity as well as the location of the predicates, and we measure how much time it would take to execute a set of queries if they were to be executed with a shared join or one by one with a query-at-a-time join. We consider three types of queries with predicates: one where all queries have a predicate on one of the relations; one where all queries have predicates on both relations; and one where half of the queries have a predicate on one relation and the other half have a predicate on the other relation. We use the default relations R and S both with 100 million tuples. A selectivity of a predicate of, for instance 0.001% on relation R, means that that query selects randomly around 1000 tuples from R. To avoid measuring the effects from scanning and evaluation of predicates, the input relations are precomputed for both the shared join and the query-at-a-time join. Results are shown in Fig. 13.

One important thing to emphasize from these results is that the execution time of the shared join has a ceiling. This represents the point where the union of all tuples is the whole relation, and thus, shared join does a full table join. The performance then is constant until the size of the bitset gets high enough to make an impact.

8.3 TPC-H benchmark

For the first type of queries where all predicates are on one relation, a shared join almost always performs better than a query-at-a-time approach. This is true even in the cases when

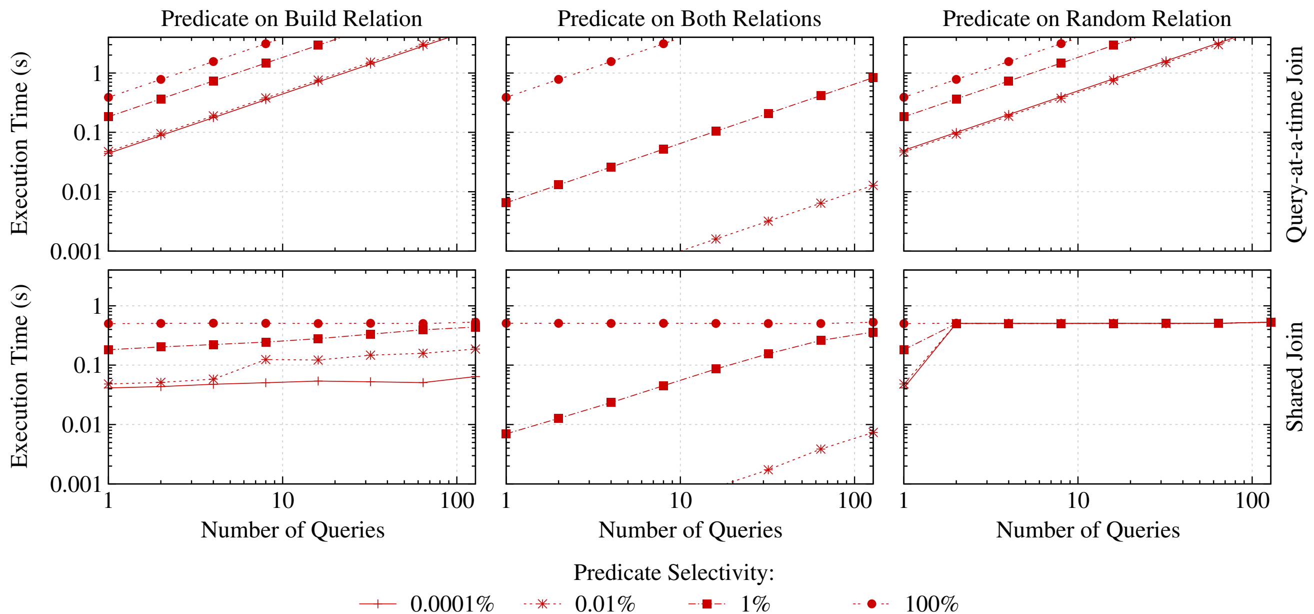


Fig. 13 Performance of MQJoin versus query-at-a-time join

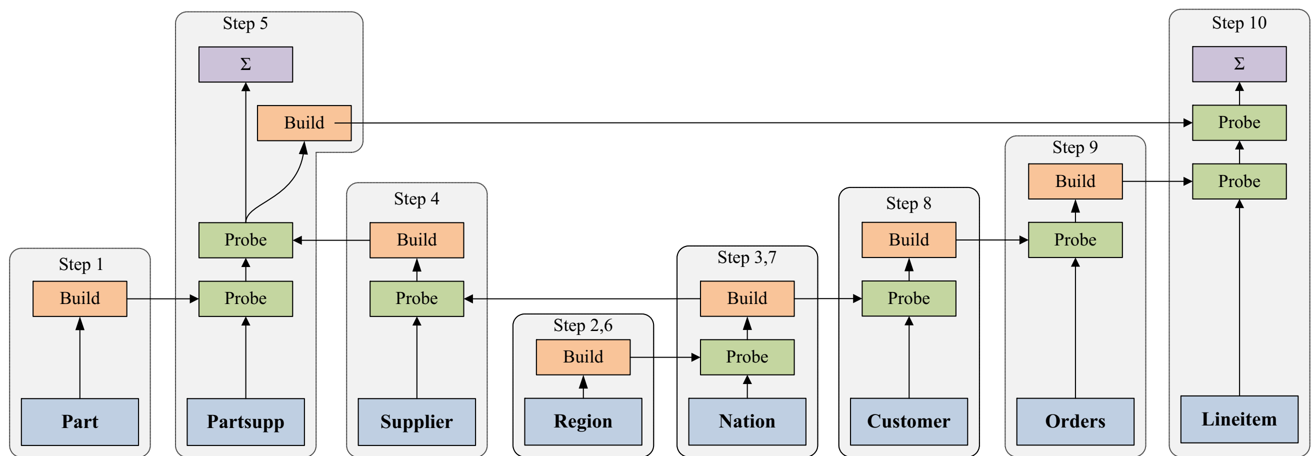


Fig. 14 Global join plan for the TPC-H workload

the predicates are mutually exclusive for all queries, making the output relations mutually exclusive as well. The benefit in this case comes from sharing the probe relation, where every probing tuple is shared for all queries.

For the second type of queries where each query has a predicate on both relations, we can see that a shared join is only beneficial if there are some common tuples between the queries. Due to the randomness of the predicates in this setup, only the queries with lower selectivity predicates share tuples as the number of queries increase. As the bitsets increase with more queries, the performance of the shared join will suffer. However, if there are no common tuples between queries, then the bitset will contain mostly zeros so it will be easily compressible.

While the previous two cases were interesting to point out, we expect that a realistic workload will consist of more diverse sets of queries. The worst-case scenario for shared join is when there are two queries one with a predicate on one relation, while the other has a predicate on the other relation. The shared join in this case will do a full table join, and the number of queries in this case required for the shared join to do better than the query-at-a-time join depends the impact of the size of the hash table on the join, which is what we see in Fig. refvaryspsbuildspssize.

The TPC-H benchmark suite [1] consists of 22 analytical queries most of which require heavy scans and joins on large portions of data. As we focus only on the scan and join operations of the queries, we took the TPC-H queries and extracted their scan and join subqueries. In order to avoid sending large materialized relations over the network, we included simple `SUM` aggregate on top of every query. To ensure that all attributes are projected during the joins as required in the original queries, we added all necessary attributes to the `SUM` aggregate expression. For instance the transformed version of Query 9 is shown in Listing 1. The rest of the queries used are shown in Appendix 1

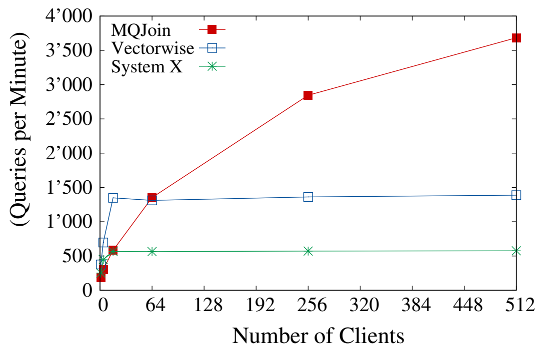
Listing 1 Transformed SQL version of Query 9

```
SELECT SUM(
    ps_supplycost + l_extendedprice +
        l_discount
    l_quantity + s_nationkey +
        o_totalprice)
FROM lineitem, part, supplier, partsup,
     orders
WHERE l_orderkey = o_orderkey
AND l_partkey = p_partkey
AND l_suppkey = s_suppkey
AND l_partkey = ps_partkey
AND l_suppkey = ps_suppkey
AND p_name LIKE '%[COLOR]%';
```

Furthermore, we removed any queries that required no joins, queries that contain more complex predicates which our shared scan implementation does not yet support and queries that required joins other than equi-joins, which are not currently supported by our system. The final set of queries include 13 query templates that contained the scan and join subqueries of the following TPC-H queries: 2, 3, 5, 7, 8, 9, 10, 11, 14, 16, 17, 19, 20. For comparison, related work uses a smaller subset of TPC-H. Both QPipe and Datapath work with only 8 queries. The global operator plan that is used to process these queries is shown in Fig. 14. This plan was created as described in Sect. 6 to maximize sharing for a batch of queries. The scale of the TPC-H data used was 10.

8.3.1 TPC-H execution

To compare our approach to a query-at-a-time system, we run an experiment with multiple clients where each client executes the TPC-H-based queries one by one, in randomized order, and with randomized parameters as per the TPC-H specification. The clients execute the queries without think time for a particular period of time, while we vary the number of clients and measure throughput and response time.



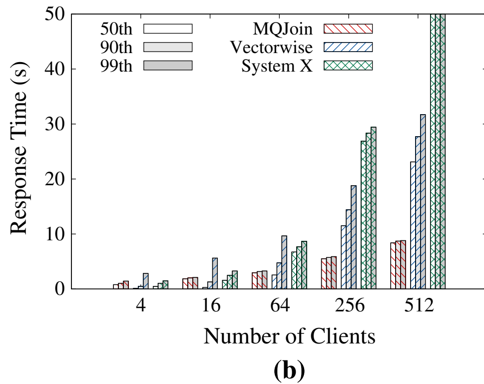


Fig. 15 Throughput and response time for TPC-H-based workload on a database of scale 10. **a** TPC-H throughput. **b** TPC-H response time percentiles

Our shared join system in this case is running on all 48 cores, with every TPC-H relation partitioned across the cores. The memory of each partition is bound to the NUMA node of the core, while the memory of the hash tables is interleaved across all NUMA nodes. As mentioned before, the system executes queries in batches where queries are queued up in a batch while the current batch is being executed. For VectorWise, we have one connection per client for up to 64 clients. With more than 64 clients, the clients start sharing 64 connections in a FCFS fashion.

In Fig. 15a, we show the throughput comparison of our system, VectorWise and System X as we increase the number of clients. The results indicate that our system outperforms both commercial counterparts for more than 60 clients and gives 2-5x higher throughput for 256 clients. Although this might not seem to be a large improvement, it should be taken into account that we are comparing to systems that are leading TPC-H benchmark performers for single-node main-memory processing. There are many optimizations in VectorWise and System X that are orthogonal to shared query processing, in particular column-wise processing. The reason why performance increase slows down from 256 to 512 queries is explained in the next section where we profile the performance of our system.

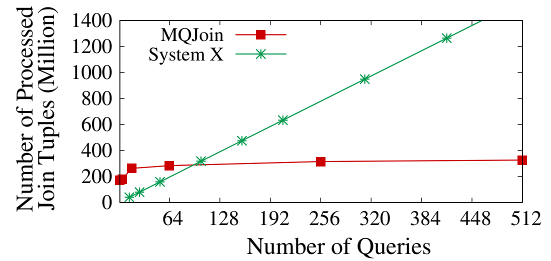


Fig. 16 Number of tuples processed in join operations in MQJoin versus system X

In Fig. 15b, we show response time percentiles for the same experiment. This graph shows that our system has significantly more stable and consistent response time. Both VectorWise and System X are slower for the most expensive queries for as low as 4 clients. This graph also shows the equalization effect of the single query queue on response time. This might not always be a desired property for all queries in the system; however, it provides predictable performance that is important for systems which must provide response time guarantees.

Figure 16 illustrates the advantage of MQJoin over its query-at-a-time counterparts. For this experiment, we measured the total number of input and output tuples processed by hash join operators for different sized sets of concurrently running queries. Due to technical limitations we only collected these data from MQJoin and System X. Nevertheless, the results clearly show the difference between MQJoin and a traditional query-at-a-time approach. For small number of concurrent queries, System X processes significantly less join tuples than MQJoin since it can optimize the join order of each query individually. Furthermore, System X makes use of bitmaps to prefilter join tuples, further reducing them at additional cost to the scan operation. On the other side, the number of processed join tuples in MQJoin grows significantly slower reaching a plateau due to the sharing effect as shown in Fig. 13. This enables MQJoin to process larger workloads more efficiently reaching higher performance.

8.3.2 Performance profiling

In the performance results in the previous experiment, the throughput no longer increased for MQJoin after a certain point. In the following experiment, we show the reason behind this. Figure 17 shows the breakdown of CPU time spent in our system per operator class while varying the number of concurrently running queries. The concurrent queries are a multiple of the set of 13 TPC-H-based queries with randomized parameter values. The results show that, as the number of queries in the batch increase, the scan operators take most of the CPU time. The reason for this is that, as shown previously, a shared join scales almost constantly with

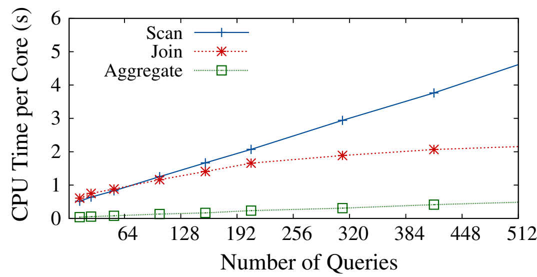


Fig. 17 CPU time spent per operator class for TPC-H-based workload

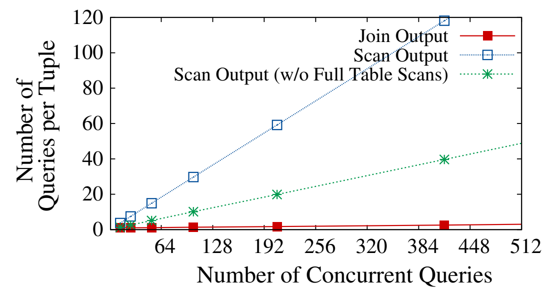


Fig. 18 Amount of data overlap in intermediate results

Table 3 Workload properties: selectivity and number of possible parameters per query template

	Select. (%)	#Param.		Select. (%)	#Param.
Q2	0.08	1250	Q11	4	25
Q3	0.465	155	Q14	0.277	60
Q5	4	25	Q16	2	3750
Q7	0.064	625	Q17	0.1	1000
Q8	0.0053	18750	Q19	0.0028	250
Q9	5.26	92	Q20	0.043	2300
Q10	4.16	24			

the number of queries as soon as the point of doing full table joins is reached. On the other hand, scaling the evaluation of predicates is more difficult and depends on workload parameters such as complexity and selectivity of the predicates.

8.3.3 Workload properties

Since we are running a workload with hundreds of concurrent queries, it is important to understand the amount of overlap in the queries and its effect on performance. For this reason, we performed both a static analysis of each of the 13 query templates and a dynamic analysis on the workload as it is being executed in the system. The results show little overlap in the amount of data queries are interested in and demonstrates the reason why MQJoin is able to benefit from sharing opportunities in this case.

Table 3 shows the summary from the static workload analysis with two key properties for each of the 13 query templates. The number of possible predicate parameters indicate how many unique queries there are in a certain workload. For the largest workload of 512 concurrent clients, this corresponds to around 40 query instantiations per template. As the table shows, only 3 of the 13 templates have less than 40 possible parameter values, the smallest one having 24. The rest have many more possible parameter values, meaning that even in a set of 512 concurrent queries, the expected amount of identical queries will be marginally small.

The selectivity values show the combined selectivity of the predicates for each query template. The results show that the majority of the queries have a selectivity of less than 1 percent, which indicates possibly little overlap in the data of interest even for several hundred queries. This is confirmed by the results of our dynamic workload analysis shown in Fig. 18. In this experiment, we measured the average number of queries per tuple for different types of intermediate results. The solid red line corresponds to the intermediate results, which are the output of join operators and input to aggregate operators. The very small amount of queries per tuple of around 1.4 for 100 queries and 2.6 for 400 queries confirms the small overlap in data mentioned before.

Unlike the output, the input to the join operators contains a larger overlap in data with an average of 120 queries per tuple for 400 concurrent queries. For this case, we measured the average number of queries per tuple in the intermediate results that are the output of scan operators and input to join operators. As is also shown in Fig. 18, the majority of this overlap comes from full table scans. This experiment demonstrates the benefits of sharing join execution even for queries with a disjoint set of predicates, since there is still a large overlap in the data that needs to be processed.

8.4 Comparison to CJoin

As the closest related work, we also compare our approach to CJoin [11]. We use the same star schema benchmark [37] workload used to test CJoin. The dataset has a scale of 100 and we use three workload types. The first two come from the same workload used in the CJoin paper with the predicates on the dimension relations set to 1 and 10%, respectively. The third one uses the queries and selectivity as defined by the star schema benchmark specification. We do not use queries 1.1, 1.2 and 1.3 as they contain predicates on the fact relation which is not supported by CJoin. Since we do not support a `group by` operation, we used a corresponding `sum` operation for CJoin as well. To avoid any disk accesses for CJoin, we placed the underlying Postgres instance in a temporary in-memory file system. For both systems, we varied the number of clients and measured response time and throughput. Clients issue queries one after another without think time.

The results are shown in Fig. 19. The first thing to note is the large performance difference between the two systems.

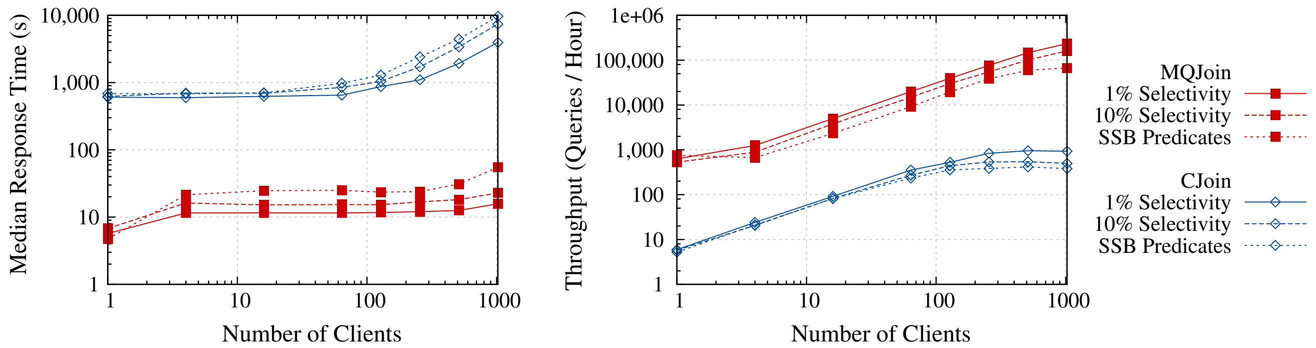
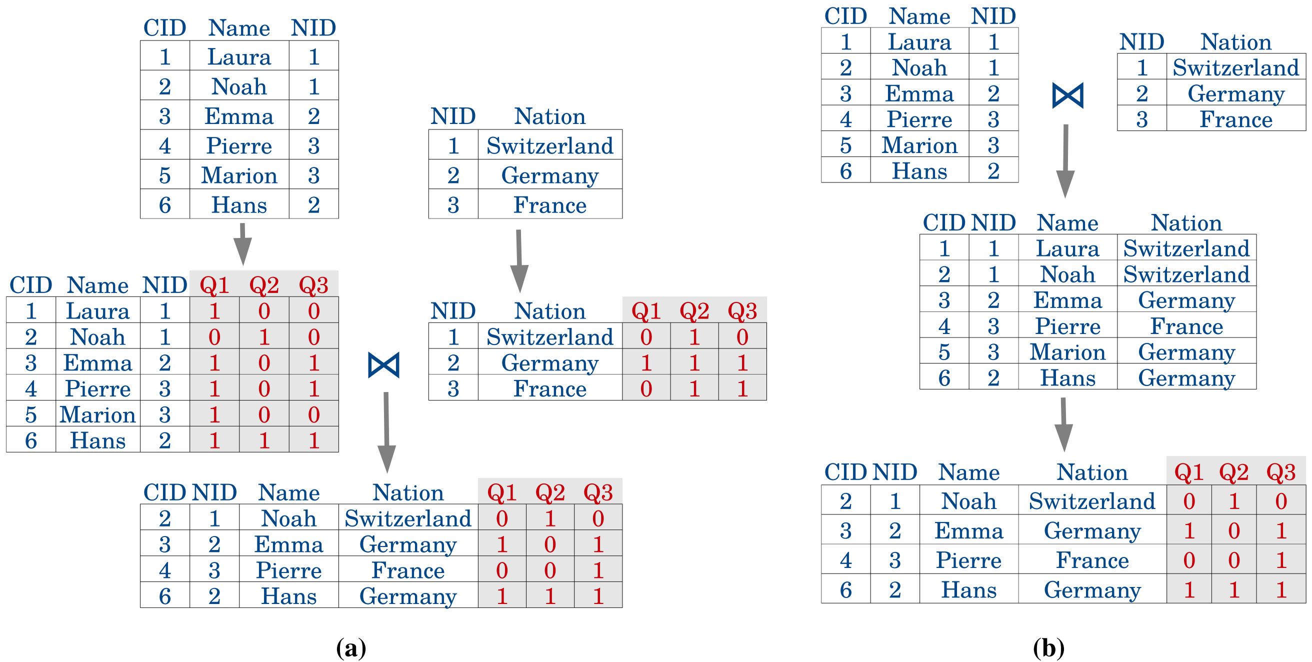


Fig. 19 MQJoin and CJoin performance for star schema benchmark dataset of scale 100



(a)

CID	Name	NID
1	Laura	1
2	Noah	1
3	Emma	2
4	Pierre	3
5	Marion	3
6	Hans	2

NID	Nation
1	Switzerland
2	Germany
3	France

CID	Name	NID	Q1	Q2	Q3
1	Laura	1	1	0	0
2	Noah	1	0	1	0
3	Emma	2	1	0	1
4	Pierre	3	1	0	1
5	Marion	3	1	0	0
6	Hans	2	1	1	1

NID	Nation	Q1	Q2	Q3
1	Switzerland	0	1	0
2	Germany	1	1	1
3	France	0	1	1

CID	NID	Name	Nation	Q1	Q2	Q3
2	1	Noah	Switzerland	0	1	0
3	2	Emma	Germany	1	0	1
4	3	Pierre	France	0	0	1
6	2	Hans	Germany	1	1	1

(b)

CID	Name	NID
1	Laura	1
2	Noah	1
3	Emma	2
4	Pierre	3
5	Marion	3
6	Hans	2

NID	Nation
1	Switzerland
2	Germany
3	France

CID	NID	Name	Nation
1	1	Laura	Switzerland
2	1	Noah	Switzerland
3	2	Emma	Germany
4	3	Pierre	France
5	3	Marion	Germany
6	2	Hans	Germany

CID	NID	Name	Nation	Q1	Q2	Q3
2	1	Noah	Switzerland	0	1	0
3	2	Emma	Germany	1	0	1
4	3	Pierre	France	0	0	1
6	2	Hans	Germany	1	1	1

Fig. 20 An example of sharing of scans and joins with two different approaches. **a** Shared scan before join. **b** Shared scan after join

One reason is that CJoin was designed with a disk resident fact relation in mind, and was run on a smaller machine with 8 cores. Although Postgres resides fully in main memory, the streaming of the fact relation from Postgres to CJoin becomes a bottleneck and is not able to supply the CJoin operator running on 40+ cores. Nevertheless, the main conclusion to draw from these results comes from the relative performance of the two systems as the number of clients increases. CJoin's performance starts degrading significantly sooner as a result of missed sharing opportunities. Since CJoin updates the hash tables for each query as the query arrives to the system, it misses out on sharing the build operation for concurrent queries. As the number of clients increase, updating the hash tables becomes a bottleneck. The per-query cost of building and updating the hash table is also a relevant factor. As shown in the results, workloads with less selective or more complex predicates on the dimension relations aggravate the problem. It is also for this reason why CJoin is suitable only in a star schema scenario.

8.5 Post-join shared scan execution

As we analyzed with the microbenchmarks in Sect. 8.1, the additional bitsets used in MQJoin cause a certain overhead per tuple when processing the join. An alternative which would avoid the use of bitsets is to use a technique where the shared scan predicates are evaluated after the shared join.

In this case, the input to the join will essentially be full table scans without any query ID bitset annotations. The join will be executed as a single common subexpression operation that in essence denormalizes the entire dataset.

An example shared scan execution before and after the join is depicted in Fig. 20. The example shown in Fig. 20a depicts the approach of MQJoin, while the example shown

in Fig. 20b depicts the case where the shared scan is executed after the join.

The problem of this approach is that that the shared execution of predicates becomes more expensive after the join. We explain this effect using the example shown in Fig. 20. In this case, any predicates on the nation table will be twice as expensive when they are evaluated after the join. This is because tuples from the nation relation are duplicated when the join with the customer nation is executed.

To test the effect of this, we performed an experiment where we compare the performance for two setups:

1. shared scan followed by a shared join on normalized `lineitem` and `orders` relations
2. shared scan on a denormalized `lineorders` relation

The first setup resembles our standard MQJoin setup, while the second setup is even stronger than the scan-after-join approach since it does not actually perform any join.

The queries we use are shown in Listing 2 and 3. The `CSTART` and `CEND` parameters are randomized for each query and the difference between them is set such that the queries select around 5% of the `Customer` relation. The database is populated with 3 million `order` and 48 million `lineitem` records.

Listing 2 Normalized Scan and Join Query

```
SELECT SUM( l_extendedprice )
FROM lineitem, orders
WHERE
     l_orderkey = o_orderkey
AND  o_custkey >= [CSTART]
AND  o_custkey < [CEND]
```

Listing 3 Denormalized Scan Query

```
SELECT SUM( l_extendedprice )
FROM lineorders
WHERE
     o_custkey >= [CSTART]
AND  o_custkey < [CEND]
```

Results are shown in Fig. 21. We can see that for a few queries the simple denormalized scan performs significantly better as it does not perform any join between the two relations. As we add more queries, though, the scan on the denormalized relation has a worse scaling factor and at some point becomes even more expensive than scanning and joining two normalized relations. If we would include the join cost to the denormalized scan so that it resembles the scan-after-join setup, it would more quickly become less efficient than the scan-before-join setup of MQJoin. The reason behind this is that the shared scan and shared join scale differently with the number of queries, as we have shown earlier in Sect. 8.3.2 and Fig. 17. In particular, in our case the join scales better, and after a certain point in the CPU time is dominated by the scan.

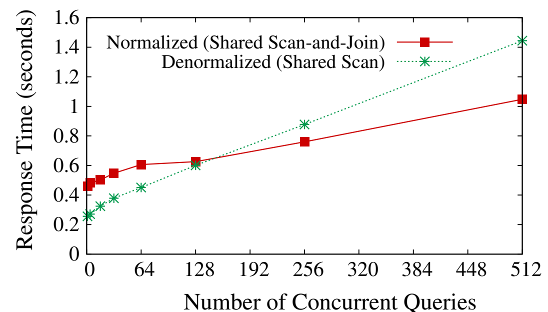


Fig. 21 Effect of number of concurrent queries on: (1) shared scan on a denormalized relation versus (2) shared scan with shared join on two normalized relations

8.6 Key takeaways

Finally, we summarize the key takeaways of this evaluation of our MQJoin approach. The above experiments demonstrate that MQJoin's algorithm for shared execution of joins achieves performance comparable to the state-of-the-art query-at-a-time join algorithms and experiences a per-tuple performance degradation of only a factor of 3 when executing 1000 join operations as opposed to one (Sect. 8.1.1). The additional performance degradation that can occur due to a lower combined selectivity of a shared join operation (leading to larger hash table sizes) is less of a concern in compute-bound cases (Sect. 8.1.2) and can be aleviated by emerging hardware technologies, such as MCDRAM, that introduce an additional layer in the memory hierarchy in between last-level CPU caches and main memory (Sect. 8.2). Also, while the benefits of MQJoin depend on predicate location and selectivity of individual queries, the work done has an upper bound, that is, a full table join between the relations leading to predictable performance and increasing benefits of sharing as the number of concurrent queries increases (Sect. 8.2.1).

In the main experiment based on a TPC-H-based benchmark, where concurrent queries have little overlap (Sect. 8.3.3), a system based on MQJoin outperforms state-of-the-art analytical query processing engines (Sect. 8.3) reaching up to 5 times higher throughput for 512 clients and significantly lower 99th percentile response times for as low as 4 clients (Sect. 8.3.1). The analysis of CPU time spent reveals that for large number of clients the majority of time is spent in predicate evaluation (Sect. 8.3.2), demonstrating the better scalability of join evaluation as opposed to predicate evaluation as the number of concurrent queries increases. This, in turn, shows the need for a shared join execution with predicates evaluated more efficiently on a normalized schema as opposed to evaluating the predicates on a denormalized schema after a join operation (Sect. 8.5).

Finally, the comparison to related work (Sect. 8.4) demonstrates the significantly higher performance of MQJoin due to efficient use of modern hardware as well as better scalabil-

ity with the number of concurrent queries due to exploiting of more sharing opportunities.

9 Conclusions

This paper presented an algorithm that exploits the sharing potential of join execution up to a very high level to meet the demands of such workloads. The goal is achieved by using techniques that minimize redundant work across concurrent queries and efficiently use the hardware resources such as CPU and memory bandwidth. The resulting method handles significantly larger workloads than the state of the art and outperforms leading main-memory analytical databases by providing higher throughput and more stable and predictable response times.

Appendix: A modified TPC-H queries

Listing 4: Modified TPC-H Queries

```
Q2:  select   sum(ps_supplycost)
     from     part, supplier, partsupp, nation,
        region
     where    p_partkey = ps_partkey
              and s_suppkey = ps_suppkey
              and p_size = $1
              and p_type like $2
              and s_nationkey = n_nationkey
              and n_regionkey = r_regionkey
              and r_name = $3;
Q3:  select   sum(l_extendedprice + l_discount +
                  l_orderkey)
     from     customer, orders, lineitem
     where    c_mktsegment = $1
              and c_custkey = o_custkey
              and l_orderkey = o_orderkey
              and o_orderdate < $2
              and l_shipdate > $2;
Q5:  select   sum(l_extendedprice + l_discount +
                  c nationkey + s nationkey)

     from     customer, orders, lineitem, supplier
        ,
              nation n1, region r1, nation n2,
                  region r2
     where    c_custkey = o_custkey
              and l_orderkey = o_orderkey
              and l_suppkey = s_suppkey
              and c_nationkey = n2.n_nationkey
              and n2.n_regionkey = r2.r_regionkey
              and r2.r_name = $1
              and s_nationkey = n1.n_nationkey
              and n1.n_regionkey = r1.r_regionkey
              and r1.r_name = $1
              and o_orderdate >= $2
              and o_orderdate < $2 + interval '1'
                  year;

Q7:  select   sum(l_extendedprice + l_discount +
                  c_nationkey + s_nationkey) as
                      revenue
     from     supplier, lineitem, orders, customer
        ,
              nation n1, nation n2
     where    s_suppkey = l_suppkey
              and o_orderkey = l_orderkey
              and c_custkey = o_custkey
              and s_nationkey = n1.n_nationkey
              and c_nationkey = n2.n_nationkey
              and n1.n_name = $1
              and n2.n_name = $2
              and l_shipdate between date '
                  1995-01-01'
                                  and date '
                                      1996-12-31';

Q8:  select   sum(l_extendedprice + l_discount +
                  o_totalprice + s_nationkey)
     from     part, supplier, lineitem, orders,
         customer,
              nation n1, nation n2, region
     where    p_partkey = l_partkey
              and s_suppkey = l_suppkey
              and l_orderkey = o_orderkey
              and o_custkey = c_custkey
              and c_nationkey = n1.n_nationkey
              and n1.n_regionkey = r_regionkey
              and r_name = $2
              and s_nationkey = n2.n_nationkey
              and n2.n_name = $1
              and o_orderdate between date '
                  1995-01-01'
                                  and date '
                                      1996-12-31'
              and p_type = $3;

Q9:  select   sum(ps_supplycost + l_extendedprice
     +
              l_discount + l_quantity +
                  s_nationkey +
              o_totalprice)

     from     lineitem, part, supplier, partsupp
     where    l_partkey = p_partkey
              and l_suppkey = s_suppkey
              and p_name like $1
              and l_partkey = ps_partkey
              and l_suppkey = ps_suppkey;

Q10:  select  sum(l_extendedprice + l_discount) as
      revenue
      from    orders, lineitem
      where   o_orderkey = l_orderkey
              and o_orderdate >= $1
              and o_orderdate < $1 + interval '3'
                  month
              and l_returnflag = 'R';
Q11:  select  sum(ps_supplycost + ps_availqty +
              ps_partkey)
      from    partsupp, supplier, nation
      where   ps_suppkey = s_suppkey
              and s_nationkey = n_nationkey
              and n_name = $1;
Q14:  select  count(*)
      from    lineitem, part
      where   l_partkey = p_partkey
              and l_shipdate >= $1
              and l_shipdate < $1 + interval '1'
                  month
              and p_type like '%PROMO%';
Q16:  select  sum(p_size)
      from    part, (select   *
                     from     partsupp
                     where    ps_suppkey in (
                                (select s_suppkey
                                 from   supplier
                                 where  s_comment
                                     not like '%
                                     Complaints%')
                   ) as suppcnttbl
```

```
       where   p_partkey = ps_partkey
               and p_brand <> $1
               and p_type not like $2
               and p_size = $3;
Q17:   select  sum(l_extendedprice + l_quantity)
       from    lineitem, part
       where   p_partkey = l_partkey
               and p_brand = $1
               and p_container = $2;
Q19:   select  sum(l_extendedprice + l_discount)
       from    lineitem, part
       where   p_partkey = l_partkey
               and p_brand = $1
               and p_container like '%SM%'
               and l_quantity >= $2
               and l_quantity <= $2 + 10
               and p_size between 1 and 5
               and l_shipmode = 'AIR'
               and l_shipinstruct = 'DELIVER_IN_
                   PERSON';
Q20:   select  sum(ps_supplycost)
       from    part, supplier, partsupp, nation
       where   ps_suppkey = s_suppkey
               and s_nationkey = n_nationkey
               and n_name = $2
               and ps_partkey = p_partkey
               and p_name like $1;
```

References

1. TPC-H Benchmark. http://www.tpc.org/tpch/spec/tpch2.17.0.pdf
2. Albutiu, M.-C., Kemper, A., Neumann, T.: Massively parallel sort-merge joins in main memory multi-core database systems. PVLDB **5**(10), 1064–1075 (2012)
3. Arumugam, S., Dobra, A., Jermaine, C.M., Pansare, N., Perez, L.: The DataPath system: a data-centric analytic processing engine for large data warehouses. Proc. SIGMOD **2010**, 519–530 (2010)
4. Avnur, R., Hellerstein, J.M.: Eddies: continuously adaptive query processing. Proc. SIGMOD **2000**, 261–272 (2000)
5. Balkesen, C., Alonso, G., Teubner, J., Özsu, M.T.: Multi-core, main-memory joins: sort versus hash revisited. PVLDB **7**(1), 85–96 (2013)
6. Balkesen, C., Teubner, J., Alonso, G., Özsu, M.T.: Main-memory hash joins on multi-core CPUs: tuning to the underlying hardware. Proc. ICDE **2013**, 362–373 (2013)
7. Balkesen, C., Teubner, J., Alonso, G., Özsu, T.: Main-memory hash joins on modern processor architectures. IEEE Trans. Knowl. Data Eng. **27**(7), 1754–1766 (2015)
8. Barber, R., Lohman, G., Pandis, I., Raman, V., Sidle, R., Attaluri, G., Chainani, N., Lightstone, S., Sharpe, D.: Memory-efficient hash joins. Proc. VLDB **8**(4), 353–364 (2014)
9. Blanas, S., Li, Y., Patel, J.M.: Design and evaluation of main memory hash join algorithms for multi-core CPUs. Proc. SIGMOD **2011**, 37–48 (2011)
10. Boncz, P.A., Zukowski, M., Nes, N.: MonetDB/X100: hyper-pipelining query execution. Proc. CIDR **2005**, 225–237 (2005)
11. Candea, G., Polyzotis, N., Vingralek, R.: A scalable, predictable join operator for highly concurrent data warehouses. PVLDB **2**(1), 277–288 (2009)
12. Chen, C., Roussopoulos, N.: The implementation and performance evaluation of the ADMS query optimizer: integrating query result caching and matching. In: Proc EDBT, pp. 323–336 (1994)
13. Chen, S., Ailamaki, A., Gibbons, P. B., Mowry, T. C.: Improving hash join performance through prefetching. In: Proc. ICDE 2004, pp. 116– (2004)
14. Chen, S., Ailamaki, A., Gibbons, P.B., Mowry, T.C.: Improving hash join performance through prefetching. ACM Trans. Database Syst. **32**(3), 17 (2007)
15. Ebenstein, R., Kamat, N., Nandi, A.: FluxQuery: an execution framework for highly interactive query workloads. In: Proc SIGMOD, pp. 1333–1345. ACM, New York, NY, USA (2016)
16. Giannikis, G., Alonso, G., Kossmann, D.: SharedDB: killing one thousand queries with one stone. PVLDB **5**(6), 526–537 (2012)
17. Giannikis, G., Makreshanski, D., Alonso, G., Kossmann, D.: Shared workload optimization. PVLDB **7**(6), 429–440 (2014)
18. Graefe, G.: Volcano— an extensible and parallel query evaluation system. IEEE Trans. Knowl. Data Eng. **6**(1), 120–135 (1994)
19. Harizopoulos, S., Ailamaki, A.: StagedDB: designing database servers for modern hardware. In: In IEEE Data, pp. 11–16 (2005)
20. Harizopoulos, S., Shkapenyuk, V., Ailamaki, A.: QPipe: a simultaneously pipelined relational query engine. Proc. SIGMOD **2005**, 383–394 (2005)
21. Ivanova, M. G., Kersten, M. L., Nes, N. J., Gonçalves, R. A.: An architecture for recycling intermediates in a column-store. In: Proc. SIGMOD, pp. 309–320. ACM, New York, NY, USA (2009)
22. Jha, S., He, B., Lu, M., Cheng, X., Huynh, H.P.: Improving main memory hash joins on intel xeon phi processors: an experimental approach. PVLDB **8**(6), 642–653 (2015)
23. Johnson, R., Harizopoulos, S., Hardavellas, N., Sabirli, K., Pandis, I., Ailamaki, A., Mancheril, N.G., Falsafi, B.: To share or not to share? Proc. VLDB **2007**, 351–362 (2007)
24. Kim, C., Kaldewey, T., Lee, V.W., Sedlar, E., Nguyen, A.D., Satish, N., Chhugani, J., Di Blas, A., Dubey, P.: Sort versus hash revisited: fast join implementation on modern multi-core CPUs. PVLDB **2**(2), 1378–1389 (2009)
25. Krikellas, K., Inc, G., Viglas, S. D., Cintra, M.: Modeling multithreaded query execution on chip multiprocessors. In ADMS (2010)
26. Lang, C.A., Bhattacharjee, B., Malkemus, T., Padmanabhan, S., Wong, K.: Increasing buffer-locality for multiple relational table scans through grouping and throttling. Proc. ICDE **2007**, 1136–1145 (2007)
27. Lang, C. A., Bhattacharjee, B., Malkemus, T., Wong, K.: Increasing buffer-locality for multiple index based scans through intelligent placement and index scan speed control. In: Proc. VLDB, pp. 1298–1309 (2007)
28. Lang, H., Mühlbauer, T., Funke, F., Boncz, P. A., Neumann, T., Kemper, A.: Data blocks: hybrid OLTP and OLAP on compressed storage using both vectorization and compilation. In: Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16, pp. 311–326. ACM, New York, NY, USA (2016)
29. Larson, P.-A., Birka, A., Hanson, E.N., Huang, W., Nowakiewicz, M., Papadimos, V.: Real-time analytical processing with SQL server. Proc. VLDB **8**(12), 1740–1751 (2015)
30. Liu, F., Blanas, S.: Forecasting the cost of processing multi-join queries via hashing for main-memory databases. In Proc. SoCC, pp. 153–166. ACM, New York, NY, USA (2015)
31. Makreshanski, D., Giceva, J., Barthels, C., Alonso, G.: BatchDB: efficient isolated execution of hybrid OLTP+OLAP workloads for interactive applications. In: Proc. SIGMOD, pp. 37–50. ACM, New York, NY, USA (2017)
32. Manegold, S., Boncz, P., Kersten, M.: Optimizing main-memory join on modern hardware. IEEE Trans. Knowl. Data Eng. **14**(4), 709–730 (2002)
33. Manegold, S., Boncz, P., Kersten, M. L.: Generic database cost models for hierarchical memory systems. In: Proc VLDB, pp. 191–202. VLDB Endowment (2002)
34. Manegold, S., Pellenkoft, A., Kersten, M. L.: A multi-query optimizer for Monet. In: Proc. BNCOD, pp. 36–50. Springer, London, UK (2000)
35. Müller, I., Sanders, P., Lacurie, A., Lehner, W., Färber, F.: Cache-efficient aggregation: hashing is sorting. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management

of Data, Proc. SIGMOD 2015, pp. 1123–1136. ACM, New York, NY, USA (2015)
36. O'Neil, P., Graefe, G.: Multi-table joins through bitmapped join indices. SIGMOD Rec. **24**(3), 8–11 (1995)
37. O'Neil, P., O'Neal, B., Chen, X.: Star schema benchmark. http://www.cs.umb.edu/~poneil/StarSchemaB.PDF
38. Psaroudakis, I., Athanassoulis, M., Ailamaki, A.: Sharing data and work across concurrent analytical queries. PVLDB **6**(9), 637–648 (2013)
39. Qiao, L., Raman, V., Reiss, F., Haas, P.J., Lohman, G.M.: Main-memory scan sharing for multi-core CPUs. PVLDB **1**(1), 610–621 (2008)
40. Raman, V., Attaluri, G., Barber, R., Chainani, N., Kalmuk, D., KulandaiSamy, V., Leenstra, J., Lightstone, S., Liu, S., Lohman, G.M., Malkemus, T., Mueller, R., Pandis, I., Schiefer, B., Sharpe, D., Sidle, R., Storm, A., Zhang, L.: DB2 with BLU acceleration: so much more than just a column store. Proc. VLDB **6**(11), 1080–1091 (2013)
41. Raman, V., Swart, G., Qiao, L., Reiss, F., Dialani, V., Kossmann, D., Narang, I., Sidle, R.: Constant-time query processing. In: Proc. ICDE 2008, pp. 60–69 (2008)
42. Roy, P., Seshadri, S., Sudarshan, S., Bhobe, S.: Efficient and extensible algorithms for multi query optimization. In: Proc. SIGMOD, pp. 249–260. ACM, New York, NY, USA (2000)
43. Răducanu, B., Boncz, P., Zukowski, M.: Micro adaptivity in vectorwise. In: Proc. SIGMOD, pp. 1231–1242. ACM, New York, NY, USA (2013)
44. Sellis, T.K.: Multiple-query optimization. ACM Trans. Database Syst. **13**(1), 23–52 (1988)
45. Shatdal, A., Kant, C., Naughton, J.F.: Cache conscious algorithms for relational query processing. Proc. VLDB **1994**, 510–521 (1994)
46. Sodani, A.: Knights landing (knl): 2nd generation intel(r) xeon phi processor. In: 2015 IEEE Hot Chips 27 Symposium (HCS), pp. 1–24 (Aug 2015)
47. Unterbrunner, P., Giannikis, G., Alonso, G., Fauser, D., Kossmann, D.: Predictable performance for unpredictable workloads. PVLDB **2**(1), 706–717 (2009)
48. Valduriez, P.: Join indices. ACM Trans. Database Syst. **12**(2), 218–246 (1987)
49. Zukowski, M., Héman, S., Nes, N., Boncz, P.: Cooperative scans: dynamic bandwidth sharing in a DBMS. Proc. VLDB **2007**, 723–734 (2007)
50. Zukowski, M., Nes, N., Boncz, P.: DSM versus NSM: CPU performance tradeoffs in block-oriented query processing. In: Proc. DaMoN 2008, pp. 47–54 (2008)
51. Zukowski, M., van de Wiel, M., Boncz, P.: Vectorwise: a vectorized analytical DBMS. Proc. ICDE **2012**, 1349–1350 (2012)