Scalable and Adaptive Online Joins

Mohammed Elseidy, Abdallah Elguindy, Aleksandar Vitorovic, and Christoph Koch
{firstname}.{lastname}@epfl.ch
École Polytechnique Fédérale de Lausanne

ABSTRACT

Scalable join processing in a parallel shared-nothing environment requires a partitioning policy that evenly distributes the processing load while minimizing the size of state maintained and number of messages communicated. Previous research proposes static partitioning schemes that require statistics beforehand. In an online or streaming environment in which no statistics about the workload are known, traditional static approaches perform poorly.

This paper presents a novel parallel online dataflow join operator that supports arbitrary join predicates. The proposed operator continuously adjusts itself to the data dynamics through adaptive dataflow routing and state repartitioning. The operator is resilient to data skew, maintains high throughput rates, avoids blocking behavior during state repartitioning, takes an eventual consistency approach for maintaining its local state, and behaves strongly consistently as a black-box dataflow operator. We prove that the operator ensures a constant competitive ratio 3.75 in data distribution optimality and that the cost of processing an input tuple is amortized constant, taking into account adaptivity costs. Our evaluation demonstrates that our operator outperforms the state-of-the-art static partitioning schemes in resource utilization, throughput, and execution time.

1. INTRODUCTION

To evaluate joins with *arbitrary* predicates on very large volumes of data, previous works [32, 26] propose efficient partitioning schemes for *offline* theta-join processing in parallel environments. The goal is to find a scheme that achieves load balancing while minimizing duplicate data storage and network traffic. Offline approaches require that all data is available beforehand and accordingly perform optimization statically before query execution.

However, online and responsive analysis of fresh data is necessary for an increasing number of applications. Businesses, engineers and scientists are pushing data analytics engines earlier in their workflows for rapid decision making. For example, in algorithmic trading, strategy designers run online analytical queries on real-time order book data. Order books consist of frequently changing orders waiting to be executed at a stock exchange. Some orders may stay in the order book relatively long before they are executed or revoked. Orders are executed through a matching engine that matches between buyer and seller trades using sophisticated matching rules. A broad range of applications, including fraud-detection mining algorithms, interactive scientific simulations, and intelligence analysis are characterized as follows: They (i) perform joins on large volumes of data with complex predicates; (ii) require operating in real-time while preserving efficiency and fast response times; (iii) and maintain large state, which potentially depends on the complete history of previously processed tuples [9, 5].

Previous work on stream processing has received considerable attention [2, 4], but is geared towards window-based relational stream models, in which state typically only depends on a recent window of tuples [9]. Although this simplifies the architecture of the stream processing engine, it is ineffective for emerging application demands that require maintaining large historical states. Only recently, has this concern been acknowledged and interest been raised in devising scalable *stateful* operators for stream processing [9].

This motivates our work towards *full-history* theta-join processing in an *online* scalable manner. In this context, the traditional optimize-then-execute strategy is ineffective due to *lack of statistics* such as cardinality information. For pipelined queries, cardinality estimation of intermediate results is challenging because of the possible correlations between predicates [23, 33] and the generality of the join conditions. Moreover, statistics are not known beforehand in streaming scenarios, where data is fed in from remote data sources [13]. Therefore, the online setting requires a versatile dataflow operator that adapts to the data dynamics. Adaptivity ensures low latency, high throughput, and efficient resource utilization throughout the entire execution.

This paper presents a novel design for an *intra-adaptive* dataflow operator for stateful online join processing. The operator supports arbitrary join-predicates and is resilient to data skew. It encapsulates adaptive state partitioning and dataflow routing. The authors of [17] point out the necessity of investigating systematic adaptive techniques as current ones lack theoretical guarantees about their behavior and instead rely on heuristic-based solutions. Therefore, to design a *provably*-efficient operator we need to characterize the optimality measures and the adaptivity costs of the operator. This requires theoretical analysis and addressing several systems design challenges which we discuss while outlining our main contributions.

Proceedings of the VLDB Endowment, Vol. 7, No. 6

1. Adapting the partitioning scheme requires state relocation which incurs additional network traffic costs. Our design employs a *locality*-aware migration mechanism that incurs *minimal* state relocation overhead.

2. We present an online algorithm that efficiently decides when to *explore* and *trigger* new partitioning schemes. An aggressively adaptive approach has excessive migration overheads, whereas a conservative approach does not adapt well to data dynamics which results in poor performance and resource utilization. Our presented algorithm balances between maintaining optimal data distribution and adaptation costs. It ensures a constant *competitive ratio* (3.75) in data distribution optimality and *amortized linear* communication cost (including adaptivity costs).

3. Previous adaptive techniques [31, 24, 29] follow a general *blocking*-approach for state relocation that quiesces input streams until relocation ends. Blocking approaches are not suitable for online operators that maintain large states because they incur lengthy stalls. Our design adopts a *non-blocking* protocol for migrations that seamlessly integrates state relocation with *on-the-fly* processing of new tuples while ensuring *eventual consistency* and result correctness.

4. Statistics are crucial for optimizing the partitioning scheme. The operator must gather them on-the-fly and constantly maintain them up-to-date. Traditionally, adaptive solutions delegate this to a centralized entity [31, 24, 19, 40] which may be a bottleneck if the volume of feedback is high [17]. Our approach for computing global statistics is decentralized requiring no communication or synchronization overhead.

Next we discuss related work; §3 introduces the background and concepts used throughout the rest of the paper and it outlines the problem and the optimization criteria; §4 presents the adaptive data-flow operator and its design in detail; §5 evaluates performance and validates the presented theoretical guarantees; and §**??** concludes.

2. RELATED WORK

Parallel Join Processing. In the past decades, much effort has been put into designing distributed and parallel join algorithms to cope with the rapid growth of data sets. Graefe [18] gives an overview of such algorithms. Schneider *et al.* [30] describe and evaluate several parallel equi-join algorithms that adopt a *symmetric partitioning method* which partitions input on the join attributes, whereas Stamos *et al.* [32] present the *symmetric fragment-and-replicate* method to support parallel theta-joins. This method relies on replicating data to ensure result completeness and on a heuristic model to minimize total communication cost.

MapReduce Joins. MapReduce [11] has emerged as one of the most popular paradigms for parallel computation that facilitates scalable and parallel processing of large data. There has been much work done towards devising efficient join algorithms using this framework. Previous work focuses primarily on equi-join implementations [3, 8, 28] by partitioning the input on the join key, whereas Map-Reduce-Merge [41] supports other join predicates as well. However, the latter requires explicit user knowledge and modifications to the MapReduce model. Recently, Okcan *et al.* [26] proposed techniques that supports theta-join processing without changes to the model. Finally, Zhang *et al.* [42] extend Okcan's work to evaluate multi-way joins. All of the aforementioned algorithms are offline and rely on static optimization. They have a blocking behavior that is attributed either to their design or to the nature of the MapReduce framework. In contrast, this paper sets out to build an online dataflow operator for scalable theta-join processing which allows for early results and online adaptivity.

Online Join Algorithms. There has been great interest in designing non-blocking join algorithms. The symmetric hash join SHJ [39] is one of the first along these lines to support equi-joins. The family of ripple joins [20] generalize block nested loop join, index loop join, and hash join to their online counterparts as well. However, these algorithms require that relations fit in memory. Subsequently, many works [37, 34, 14, 25] appeared which present memory overflow resolution techniques that allow parts of the input stream to be spilled out to disk for later processing. All the previous are local online join algorithms, and thus, are orthogonal to our data-flow operator. In the presented parallel operator, each machine can freely adopt any flavor of the aforementioned non-blocking algorithms to perform joins locally on its assigned data partition.

Stream Processing Engines. Distributed stream processors such as BOREALIS [2] and STREAM [4] focus on designing efficient operators for continuous queries. They assume that data streams are processed in several sites, each of which holds some of the operators. They are geared towards window-based relational stream models, in which state typically only depends on a recent finite set of tuples [9]. In contrast, this paper is concerned with the design of a scalable operator, as opposed to a centralized approach and it targets stateful streaming queries which maintain large states, potentially full historical data [9]. Castro *et al.* [9] introduce a scale-out mechanism for stateful operators, however they are limited to stream models with key attributes (equi-joins).

Adaptive Query Processing. Adaptive query processing AQP techniques cope their behavior, at run-time, to data characteristics. There has been a great deal of work on centralized AQP [7, 13, 21, 16] over the last few years. For parallel environments, [17] presents a detailed survey. The FLUX operator [31] is the closest to our work. FLUX is a *general* adaptive operator that encloses adaptive state partitioning and routing. The operator is *content-sensitive* and suitable for look-up based operators. Although in [31] the authors focus on single-input aggregate operators [24], it can support a restricted class of join predicates, e.g. equi-join. FLUX supports equi-joins under skewed data settings but requires *explicit user knowledge* about partitions before execution. In [19, 38], the authors present techniques to support multi-way non equi-joins. All these approaches are mainly applied to data streaming scenarios with finite window semantics. On the other hand, this paper presents an adaptive dataflow operator for *general* joins. It advances the state of the art in online equi-join processing in the presence of data skew. And most importantly, the operator runs on long running *stateful*, potentially full-history, queries [9, 5].

Eddies. Eddies [6, 12] are among the first adaptive techniques known for query processing. Eddies act as a tuple router that is placed at the center of a dataflow, intercepting all incoming and outgoing tuples between operators in the flow. Eddies observe the rates of all the operators and accordingly makes decisions about the order at which new tuples visit the operators. In principle, Eddies are able to choose different operator orderings for each tuple within the query processing engine to adapt to the current information about the environment and data. Compared to our work, this direction seeks adaptations at an orthogonal hierarchical level. It is concerned with *inter*-operator adaptivity as opposed to our work on *intra*-operator adaptivity. More-

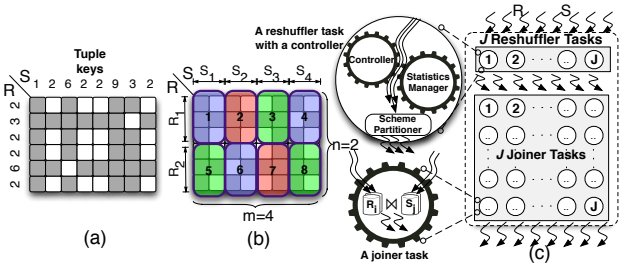


Figure 1: (a) $R \bowtie S$ join-matrix example, gray cells satisfy the $\neq$ predicate. (b) a (2,4)-mapping scheme using $J = 8$ machines. (c) the theta-join operator structure.

over, the original Eddies architecture is centralized and cannot be applied to a distributed setting in a straightforward manner [17]. The work in [35, 43] leverages the Eddies design to the distributing setting but assumes window semantics; tolerates loss of information; and does not investigate adaptations on operators that hold internal state.

3. PRELIMINARIES

This section defines notations and conventions used throughout the rest of this paper. It describes the data partitioning scheme used by the dataflow operator, outlines the operator's structure, and defines the optimization criteria.

3.1 Join Partitioning Scheme

We adopt and extend the join-matrix model [26, 32] to the data streaming scenario. A join between two data streams R and S is modeled as a join-matrix $\mathcal{M}$. For row i and column j, the matrix cell $\mathcal{M}(i,j)$ represents a potential otput result. $\mathcal{M}(i,j)$ is *true* if and only if the tuples r_i and s_j satisfy the join predicate. The result of any join is a subset of the cross-product. Hence, the join-matrix model can represent any join condition. Fig. 1a shows an example of a join-matrix with the predicate $\neq$.

We assume a shared-nothing cluster architecture. J physical machines are dedicated to a *single* join operator. A *partitioning scheme* maps matrix cells to machines for evaluation such that each cell is assigned to exactly one machine. This ensures result completeness and avoids expensive post processing or duplicate elimination. There are many possible mappings [26], however, we present a grid-layout partitioning scheme which (i) ensures *equal* and *minimum* join work distribution among *all* machines, (ii) incurs *minimal* storage and communication costs, (iii) and has a symmetric structure that lends itself to adaptivity. We refer the interested reader to [15] for bounds, proofs, and comparison with previous partitioning approaches [26]. The scheme can be briefly described as follows: the join-matrix $\mathcal{M}$ is divided into J regions of equal area and each machine is assigned a single region. As illustrated in Fig. 1b, the relations are split into equally sized partitions $R_1, R_2, \ldots, R_n$ and $S_1, S_2, \ldots, S_m$ where $n \cdot m = J$. For every pair $1 \leq i \leq n$ and $1 \leq j \leq m$, there is exactly one machine storing both partitions R_i and S_j. Accordingly, each machine evaluates the corresponding $R_i \bowtie_\theta S_j$ independently. We refer to this as the (n,m)*-mapping scheme*.

3.2 Operator Structure

As illustrated in Fig. 1c, the operator is composed of two sets of tasks. The first set consists of joiner tasks that do the actual join computation whereas the reshufflers set is responsible for distributing and routing the tuples to the appropriate joiner tasks. An incoming tuple to the operator is randomly routed to a reshuffler task. One task among the reshufflers, referred to as the controller, is assigned the additional responsibility of monitoring global data statistics and triggering adaptivity changes. Each of the J machines run one joiner task and one resuffler task.

The reshufflers randomly divide incoming tuples uniformly among partitions. Under an (n,m)-mapping scheme, an incoming $r(s)$ tuple, is assigned a *randomly* chosen partition $R_i(S_j)$. This routing policy ensures load balance and resilience to data skew, i.e., *content-insensitivity*. For a large number of input tuples, the numbers in each partition are roughly equal. Thus, all bounds, later discussed, are meant to approximately hold in expectation with high probability.

Exactly m joiners are assigned partition R_i and exactly n joiners are assigned partition S_j. Therefore, whenever a reshuffler receives a new $R(S)$ tuple and decides that it belongs to partition $R_i(S_j)$, the tuple is forwarded to $m(n)$ distinct joiner tasks. Any flavor of non-blocking join algorithm, e.g., [39, 37, 34, 14, 20], can be independently adopted at each joiner task. Local non-blocking join algorithms traditionally operate as follows: when a joiner task receives a new tuple, it is stored for later use and joined with stored tuples of the opposite relation.

3.3 Input-Load Factor

Theta-join processing cost, in the presented model, is determined by the costs of joiners receiving input tuples, computing the join, and outputting the result. Under the presented grid-scheme, the join matrix is divided into congruent rectangular regions. Therefore, the costs are the same for every joiner. Since all joiners operate in parallel, we restrict our attention to analyzing one joiner.

The join computation and its output size on a single joiner are independent of the chosen mapping. This holds because both quantities are proportional to the area of a single region, which is $|R| \cdot |S| / J$. This is independent of n and m. However, the input size corresponds to the semi-perimeter of one region and is equal to $size_R \cdot |R| / n + size_S \cdot |S| / m$, where $size_R(size_S)$ is the size of a tuple of $R(S)$. This also represents the storage required by every joiner since each received tuple is eventually stored. We refer to this value as the **input-load factor** (ILF). This is the only performance metric that depends on the chosen mapping. *An optimal mapping covers the entire join matrix with minimum ILF.* Minimizing the ILF maximizes performance and resource utilization. This is extensively validated in our experiments (§5) and is attributed to the following reasons: (i) there is a monotonically increasing overhead for processing input tuples per machine. The overhead includes demarshalling the message; appending the tuple to its corresponding storage and index; probing the indexes of the other relation; sorting the input in case of sort-based online join algorithms [14, 25], etc. Minimizing machine input size results in higher local throughput and better performance. (ii) Minimizing storage size per machine is also necessary, because performance deteriorates when a machine runs out of main memory and begins to spill to disk. Local non-blocking algorithms perform efficiently when they operate within the memory capacity. Although, they employ overflow resolution strategies for large data, they persist to experience performance hits and long delayed join evaluation [13]. (iii) Overall, minimizing the ILF results in minimum global duplicate storage and replicated messages ($J \cdot ILF$). This maximizes overall opera-

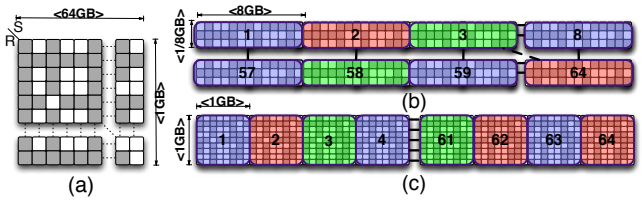


Figure 2: (a) join-matrix (b) $(8,8)$-mapping scheme (c) $(1,64)$-mapping scheme.

tor performance and increases global resource utilization by minimizing total storage and network traffic and thus preventing congestion of resources. This is essential for cloud infrastructures which typically follow *pay-as-you-go* policies.

Fig. 2 compares between two different mappings of a join-matrix with dimensions 1GB and 64GB for streams R and S respectively. Given 64 machines, an $(8,8)$-mapping results in an $(8\frac{1}{8})$GB ILF and a total of 520GB of replicated storage and messages. Whereas a $(1,64)$-mapping results in a 2GB ILF and a sum of 128GB of replicated data. Since stream sizes are dynamic and not known in advance, maintaining an optimal (n,m)-mapping scheme throughout execution requires adaptation and mapping changes.

4. INTRA-OPERATOR ADAPTIVITY

The goal of adaptive processing is, generally, dynamic recalibration to immediately react to the frequent changes in data and statistics. Adaptive solutions supplement regular execution with a control system that monitors performance, explores alternative configurations and triggers changes. These stages are defined within a cycle called the *Adaptivity Loop*. This section presents the design of an adaptive dataflow theta-join operator that continuously modifies its (n,m)-mapping scheme to reflect the optimal data assignment and routing policy. We follow a discussion flow that adopts a common framework [13] that decomposes the adaptivity loop into three stages: (i) The *monitoring* stage that involves measuring data characteristics like cardinalities. (ii) The *analysis and planning* stage that analyzes the performance of the current (n,m)-mapping scheme and explores alternative layouts. (iii) The *actuation* stage that corresponds to migrating from one scheme to another with careful state relocation.

4.1 Monitoring Statistics

In this stage, the operator continuously gathers and maintains online cardinality information of the incoming data. Traditional adaptive techniques in a distributed environment [31, 24, 19, 40] either rely on a centralized controller that periodically gathers statistics or on exchanging statistics among peers [35, 43]. This may become a bottleneck if the number of participating machines and/or the volume of feedback collected is high [17]. In contrast, we follow a decentralized approach, where reshufflers gather statistics on-the-fly while routing the data to joiners. Since reshufflers receive data that is randomly shuffled from the previous stages, the received local samples can be scaled by J to construct global cardinality estimates (Alg. 1 lines 2,3). These estimates can be reinforced with statistical estimation theory tools [22] to provide confidence bounds. The advantages of this design are three-fold: *a*) A centralized entity for gathering statistics is no longer required, removing a source of potential bottlenecks. Additionally, it precludes any exchange communication or synchronization overheads. *b*) This model can be easily extended to monitor other data statistics, e.g., frequency histograms. *c*) The design supports fault tolerance and state reconstruction. When a controller task fails, any other reshuffler task can take over.

Algorithm 1 Controller Algorithm.

Input: Tuple t
Initialize: $|R|, |S|, |\Delta R|, |\Delta S| \leftarrow 0$;
1: **function** UPDATE STATE(t)
2: **if** $t \in R$ **then** $|\Delta R| \leftarrow |\Delta R| + J$ ▷ Scaled increment.
3: **else** $|\Delta S| \leftarrow |\Delta S| + J$
4: MigrationDecision($|R|, |S|, |\Delta R|, |\Delta S|$)
5: Route t according to the current (n,m)-scheme.
6: **end function**

4.2 Analysis and Planning

Given that global statistics are constructed in Alg. 1, the controller is capable of monitoring the efficiency of the current mapping scheme, and thus, determining the overall performance of the operator. Furthermore, it checks for alternative (n,m)-mapping schemes that minimize the ILF (Alg. 1 line 4). If it finds a better one, it triggers the new scheme. This affects the route of new tuples and impacts machine state. Adopting this dynamic strategy reveals three challenges that need careful examination: *a*) Since the controller is additionally a reshuffler task, it has the main duty of routing tuples in parallel to exploring alternative mappings. Thus, it has to balance between the ability to quickly react to new cardinality information against the ability to process new tuples rapidly (the classic *exploration-exploitation dilemma*). *b*) Migrating to a new mapping scheme requires careful state maintenance and transfer between machines. This incurs non-negligible overhead due to data transmission over the network. The associated costs of migration might outweigh the benefits if handled naïvely. *c*) An aggressively adaptive control system might suffer from thrashing and excessive migration overheads (quadratic costs) while a conservative system does not adapt well to data dynamics. Thus, the controller should be alert in choosing the moments for triggering migrations.

In this section, we describe a constant-competitive algorithm that decides when to *explore* and *trigger* new schemes such that the total adaptation costs are amortized.

4.2.1 1.25-Competitive Online Algorithm

Alg. 2 decides the time points that explore and trigger migration decisions. Right after an optimal migration, the operator has received $|R|$ and $|S|$ tuples from the respective relations. The algorithm maintains two counts $|\Delta R|$ and $|\Delta S|$, denoting the newly arriving tuples on both relations respectively after the last migration. If either $|\Delta R|$ reaches $|R|$ or $|\Delta S|$ reaches $|S|$[1], the algorithm explores alternative mapping schemes and performs a migration, if necessary.

The two metrics of interest here are the ILF and the migration overhead. The aim of this section is to demonstrate the following key result.

THEOREM 4.1. *Assume that the number of joiners J is a power of two, the sizes for $|R|$ and $|S|$ are no more than a factor of J apart, and that tuples from R and S have the same size. For a system applying Alg. 2, the following holds:*

[1]These conditions can be generalized to $|\Delta R| = \varepsilon |R|$ or $|\Delta S| = \varepsilon |S|$, where $0 < \varepsilon \leq 1$ trading off competitiveness with costs. Further details and analyses can be found in [15].

Algorithm 2 Migration Decision Algorithm.

Input: $|R|$, $|S|$, $|\Delta R|$, $|\Delta S|$
1: **function** MIGRATIONDECISION($|R|$, $|S|$, $|\Delta R|$, $|\Delta S|$)
2: **if** $|\Delta R| \geq |R|$ or $|\Delta S| \geq |S|$ **then**
3: Choose mapping (n, m) minimizing $|R|/n + |S|/m$
4: Decide a migration to (n, m)
5: $|R| \leftarrow |R| + |\Delta R|$; $|S| \leftarrow |S| + |\Delta S|$
6: $|\Delta R| \leftarrow 0$; $|\Delta S| \leftarrow 0$
7: **end function**

1. *The ILF is at most* 1.25 *times that of the optimal mapping at any point in time. $ILF \leq 1.25 \cdot ILF^*$, where ILF^* is the input-load factor under the optimal mapping at any point in time. Thus, the algorithm is* 1.25*-competitive.*

2. *The total communication overhead of migration is amortized, i.e., the cost of routing a new input tuple, including its migration overhead, is $O(1)$.*

Input-Load Factor. We hereby analyze the behavior of the ILF under the proposed algorithm. Since we assume that $size(r) = size(s)$, it follows that minimizing the ILF is equivalent to minimizing $(|R|/n + |S|/m)$.

LEMMA 4.1. *If J is a power of two and it holds that $1/J \leq |R|/|S| \leq J$, then under an optimal mapping (n, m),*

$$\frac{1}{2}\frac{|S|}{m} \leq \frac{|R|}{n} \leq 2\frac{|S|}{m} \quad \textit{and} \quad \frac{1}{2}\frac{|R|}{n} \leq \frac{|S|}{m} \leq 2\frac{|R|}{n}.$$

PROOF. An optimal mapping minimizes $|R|/n + |S|/m$, under the restriction that $n \cdot m = J$. This happens when $|R|/n$ and $|S|/m$ are closest to each other. Since J is a power of two, by assumption, (thus, also n and m), it follows that under the optimal mapping $|R|/n \leq 2|S|/m$. Assume it were not the case, then $|R|/n > 2|S|/m$. Under the mapping $(2n, m/2)$, both $|R|/n$ and $|S|/m$ are closer, yielding a lower input-load factor, contradicting the optimality of (n, m). Choosing such a mapping is possible, assuming that $1/J \leq |R|/|S| \leq J$. The other inequality is symmetric. □

This lemma is useful in proving all subsequent results. The first important result is that the ILF is within a constant factor from that of the optimal scheme. This is due to the fact that Alg. 2 does not allow the operator to receive many tuples without deciding to recalibrate. The following theorem formalizes this intuition.

LEMMA 4.2. *If $|\Delta R| \leq |R|$ and $|\Delta S| \leq |S|$ and (n, m) is the optimal mapping for $(|R|, |S|)$ tuples, then the optimal mapping for $(|R| + |\Delta R|, |S| + |\Delta S|)$ is one of (n, m), $(n/2, 2m)$, and $(2n, m/2)$.*

PROOF. Without loss of generality, assume that $|\Delta S| \geq |\Delta R|$. It must be that an optimal mapping will not decrease m (since $|S|$ grew relative to $|R|$). Therefore, the optimal is one of (n, m), $(n/2, 2m)$, $(n/4, 4m)$, ..., etc. To prove that the optimum is either (n, m) or $(n/2, 2m)$, it is sufficient to prove the following inequality

$$\begin{aligned} \frac{|R| + |\Delta R|}{n/2} + \frac{|S| + |\Delta S|}{2m} &\leq \frac{|R| + |\Delta R|}{n/4} + \frac{|S| + |\Delta S|}{4m} \\ \frac{|S| + |\Delta S|}{m} &\leq \frac{8(|R| + |\Delta R|)}{n} \end{aligned}$$

which means that the ILF under an $(n/2, 2m)$-mapping is smaller than that under an $(n/4, 4m)$-mapping. This holds because $|S|/m \leq 2|R|/n$ (lemma 4.1), even if $|\Delta S| = |S|$ and $|\Delta R| = 0$. The case $|\Delta R| \geq |\Delta S|$ is symmetric. □

Alg. 2 decides migration once $|\Delta R| = |R|$ or $|\Delta S| = |S|$. Therefore, lemma 4.2 implies that while the system is operating with the mapping (n, m), the optimum is one of (n, m), $(n/2, 2m)$, and $(2n, m/2)$. This implies the following.

LEMMA 4.3. *If $|\Delta R| \leq |R|$ and $|\Delta S| \leq |S|$ and (n, m) is the optimal mapping for $(|R|, |S|)$ tuples, then under Alg. 2, the input-load factor ILF never exceeds $1.25 \cdot ILF^*$. In other words, the algorithm is* 1.25*-competitive.*

PROOF. By lemma 4.2, the optimal mapping is either (n, m), $(n/2, 2m)$ or $(2n, m/2)$. If the optimal mapping is (n, m) then $ILF = ILF^*$. Otherwise, the ratio can be bounded as follows. Without loss of generality, assume that the optimum is $(n/2, 2m)$ then

$$\frac{ILF}{ILF^*} \leq \frac{(|R| + |\Delta R|)/n + (|S| + |\Delta S|)/m}{(|R| + |\Delta R|)/(n/2) + (|S| + |\Delta S|)/(2m)}$$

where the constraints $|\Delta R|/n \leq |R|/n$, $|\Delta S|/m \leq |S|/m$ and those in lemma 4.1 must hold. All cardinalities are non-negative. Consider the ratio as a function of the variables $|R|/n$, $|S|/m$, $|\Delta R|/n$ and $|\Delta S|/m$. The maximum value of the ratio of linear functions in a simplex (defined by the linear constraints) is attained at a simplex vertex. By exhaustion, the maximum occurs when $|\Delta R| = 0$, $|\Delta S| = |S|$ and $|S|/m = 2|R|/n$. Substituting gives 1.25. □

Migration Overhead. It remains to show that, under the described algorithm, the migration overhead is amortized. This requires showing that the migration process can be done efficiently and that when a migration is triggered, enough tuples are received to "pay" for this migration cost.

The migration of interest is the change from the (n, m) to $(n/2, 2m)$-mapping (symmetrically, (n, m) to $(2n, m/2)$). Migration can be done naïvely by repartitioning all previous states around the joiners according to the new scheme. This approach unnecessarily congests the network and is expensive. In contrast, we present a *locality*-aware migration mechanism that minimizes state transfer overhead. To illustrate the procedure, we walk through an example. Consider a migration from a $(8, 2)$ to a $(4, 4)$-mapping scheme $(J = 16)$ as depicted in Fig. 3. Before the migration, each joiner stores about an eighth of R and half of S. After the migration, each joiner stores a quarter of R and only one quarter of S. To adapt, joiners can efficiently and deterministically discard a quarter of S (half of what they store). However, tuples of R must be exchanged. In Fig. 3, joiners 1 and 2 store the "first" eighth of R while joiners 3 and 4 store the "second" eighth of R. Joiners 1 and 3 can exchange their tuples and joiners 2 and 4 can do the same in parallel. Joiners 5 and 7, 6 and 8, and so forth operate similarly in parallel. This incurs a total overhead of $|R|/4$ time units which is the bi-directional communication cost of $|R|/8$. This idea can be generalized, yielding bounds on the migration overhead.

LEMMA 4.4. *Migration from (n, m) to $(n/2, 2m)$-mapping can be done with a communication cost of $2|R|/n$ time units. Similarly, migrating to $(2n, m/2)$ incurs a cost of $2|S|/m$.*

PROOF. Without loss of generality, consider the migration to $(n/2, 2m)$. No exchange of S state is necessary. On the other hand, tuples of R have to be exchanged among joiners. Before migration each of the J joiners had $|R|/n$ tuples from R, while after the migration, each must have $2|R|/n$. Consider one group of n joiners sharing the same tuples from S (corresponding to a "column" in Fig. 3). These

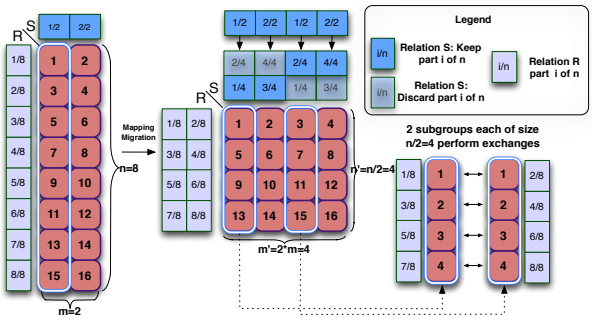


Figure 3: Migration from a $(8, 2)$- to a $(4, 4)$-mapping.

joiners, collectively, contain the entire state of R. They can communicate in parallel with the other $m-1$ groups. Therefore, we analyze the state relocation for one such group and it follows that all groups behave similarly in parallel.

Divide the group into two subgroups of $n/2$ joiners. Number the joiners in each group $1, 2, \ldots, n/2$. Joiner pairs labeled i should exchange their tuples together. It is clear that each pair of joiners labeled i ends up with a distinct set of $2|R|/n$ tuples. Fig. 3 describes this exchange process. Each of the pairs labeled i can communicate completely in parallel. Therefore, the total migration overhead is $2|R|/n$, since each joiner in the pair sends $|R|/n$ tuples to the other. □

LEMMA 4.5. *The cost of routing tuples and data migration is linear. The amortized cost of an input tuple is* $O(1)$.

PROOF. Since all joiners are symmetrical and operate simultaneously in parallel, it suffices to analyze cost at one joiner. Therefore, after receiving $|\Delta R|$ and $|\Delta S|$ tuples, the operator spends *at least* $\max(|\Delta R|/n, |\Delta S|/m)$ units of time processing these tuples at the appropriate joiners. By assigning a sufficient amortized cost per time unit, the received tuples pay for the later migration.

By lemma 4.2, the optimal mapping is (n, m), $(n/2, 2m)$ or $(2n, m/2)$. If the optimal mapping is (n, m), then there is no migration. Without loss of generality, assume that $|\Delta S| \geq |\Delta R|$ and that the optimal mapping is $(n/2, 2m)$. Between migrations, $\max(|\Delta R|/n, |\Delta S|/m)$ time units elapse, each is charged 7 units. One unit is used to pay for routing and 6 are reserved for the next migration. The cost of migration by lemma 4.4 is $2(|R|+|\Delta R|)/n$. The amortized cost reserved for migration is $6\max(|\Delta R|/n, |\Delta S|/m)$. Since a migration was triggered, either $|\Delta R| = |R|$ or $|\Delta S| = |S|$. In either case, it should hold that the reserved cost is at least the migration cost, that is,

$$6\max(|\Delta R|/n, |\Delta S|/m) \geq 2(|R| + |\Delta R|)/n.$$

If $|\Delta R| = R$, then by substituting, the left hand side is $6\max(|\Delta R|/n, |\Delta S|/m) \geq 6|R|/n$ and the right hand side is $2(|R| + |\Delta R|)/n = 4|R|/n$. Therefore, the inequality holds. If $|\Delta S| = S$, then the left hand side is

$$6\max(|\Delta R|/n, |\Delta S|/m) \geq 2|\Delta R|/n + 4|S|/m.$$

Therefore, the left hand side is not smaller than the right hand side, since $2|S|/m \geq |R|/n$ (by lemma 4.1). Thus, the inequality holds in both cases. The cases, when $|\Delta R| \geq |\Delta S|$ or when the optimal is $(2n, m/2)$, are symmetric. □

Lemmas 4.3 and 4.5 directly imply Theorem 4.1.

4.2.2 Generalization and Discussion

This section outlines the relaxation of the assumptions in Theorem 4.1 while preserving the *constant* competitiveness and the amortized adaptivity costs. This is followed by a discussion about including *elasticity* into the framework while maintaining strong guarantees.

Relation Cardinality Ratio. Without loss of generality, assume that $|R| > |S|$. Previously, the analysis assumed that $|R| \leq J|S|$. This can be relaxed by continuously padding the smaller relation with dummy tuples to maintain the ratio less than J. This requires padding the relation S with *at most* $|R|/J \leq T/J$ tuples, where T is the total number of tuples $|R| + |S|$. Therefore, the total number of tuples the operator handles, including dummy tuples, is *at most* $T + T/J = (1 + 1/J)T$ tuples. The ratio of the relation sizes still respects the assumption. Therefore, the analysis in the previous section holds except that the ILF now gets multiplied by a factor of $1 + 1/J$. This factor is *at most* 1.5 (since $J \geq 2$). This factor *tends to one* as the number of joiners increases. Therefore, the algorithm is still constant-competitive, with the constant being $1.25 \cdot 1.5 = 1.875$. Similarly, adding the dummy tuples multiplies the migration costs by *at most* 1.5. Hence, the overhead remains linear.

Relative Tuple Sizes. To facilitate analysis we assumed that tuples from R and S have the same size. To generalize this, let the size of an $R(S)$ tuple be $\tau_R(\tau_S)$. An input R tuple can be viewed as the reception of τ_R "unit" tuples. Similarly an S tuple is τ_S unit tuples. The previous analysis holds except that migration decisions can be slightly delayed. For example, if the migration decision is supposed to happen after the reception of 5 unit tuples and a tuple of size 1000 units is received, then the migration decision is delayed by 995 units. Therefore, the ILF is increased by *at most* an additive factor of $\max(\tau_R, \tau_S)$, i.e., $ILF \leq K \cdot ILF^* + \max(\tau_R, \tau_S)$.

Number of Joiners. Until this point, we have assumed that the number of available machines is a power of 2. Under this scheme, *at most* half of the resources ($\mathcal{R} \in \mathbb{N}^+$) need to be discarded, and hence, both the joiner workload and the ILF are *at most* doubled. Nevertheless, we describe a simple extension to the scheme that exploits *all* resources while providing strong guarantees. Assume that $J \in \mathbb{N}^+$, then J has a unique decomposition into a sum of powers of two. Let $J = J_1 + J_2 + \ldots + J_{\mathcal{G}}$ where each J_i is a power of two. Accordingly, the machines are broken down into $\mathcal{G}$ groups, where group i has J_i machines. There can be *at most* $\lceil \log J \rceil$ of such groups. Finally, each group operates exactly as described previously. Fig. 4a illustrates an example, given a pool of $J = 20$ machines, it is clustered into two groups of sizes 16 and 4 which operate independently. An incoming tuple is sent to all $\mathcal{G}$ groups to be joined with all stored tuples. Only one group stores this tuple for joining with future tuples. This is determined by computing a pseudo-random hash such that the probability that group i is chosen is equal to $P_i = J_i/J$. To guarantee correctness, it is essential that if a pair of tuples are sent to two machines, each belonging to different groups, that this pair of tuples is received in the same order by both machines. With very high probability, the mappings of two groups will be similar. More specifically, for two groups with sizes $J_1 < J_2$, it will hold that n_2 (m_2) is divisible by n_1 (m_1). Blocks of machines in the bigger group correspond to a single machine in the smaller group (see Fig. 4a). In each such block, a single machine does the task of forwarding all tuples to machines within that block as well as the machine in the smaller group.

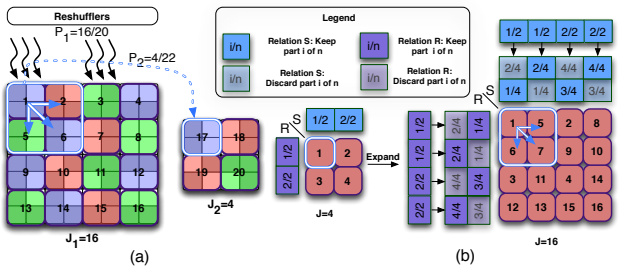


Figure 4: (a) decomposing $J = 20$ machines into independent groups of 16 and 4 machines. (b) Elastic expansion from 4 to 16 machines.

This ensures that machines get tuples in the same order at the cost of tuple latency proportional to $\log J$, since tuples have to be propagated serially among $\log J$ groups of machines. Under this modified scheme, the total join work is equally distributed among all joiners, the competitive ratio of storage is *at most* doubled (3.75), and the total routing cost, including migrations, is $O(T \log J)$. For further details about these results and their proofs, we refer the interested reader to the technical report [15].

Elasticity. In the context of online query processing, the query planner may be unable to a-priori determine the number of machines J to be dedicated to a join operator. It is thus desirable to allocate as few joiners as possible to the operator while ensuring that the stored state on each machine is reasonably maintained to prevent disk spills and performance degradation. We hereby briefly describe a simple extension that allows the operator to elastically scale-out as required while maintaining all the theoretical bounds. For joiners, designate a maximum capacity C of tuples (ILF) per joiner. At the migration checkpoints (Alg. 2) and after migration, if any joiner's storage exceeds $C/2$, the operator scales-out by splitting the state of every joiner to 4 joiners. Each joiner distributes state to 3 new joiners as described in Fig. 4b. Under the described extension, the competitive ratio remains unchanged as before and the amortized costs of scaling-out and adaptation remain linear. Due to space constraints we refer the interested reader to [15] for further details, analyses, and proofs.

4.3 Actuation

The previous section provides a high-level conceptual description of the algorithm. Migration decision points are specified to guarantee a close-to-optimal ILF and linear amortized adaptivity cost. This section describes the system-level implementation of the migration process.

Previous work on designing adaptive operators [31, 24, 29] follow a general theme for state relocation. The following steps give a brief description of the process: (i) Stall the input to the machines that contain state to be repartitioned. The new input tuples are buffered at the data sources. (ii) Machines wait for all in-flight tuples to arrive and be processed. (iii) Relocate state. (iv) Finally, online processing resumes. Buffered tuples are redirected to their new location to be processed. This protocol is not suitable for large state operators. Its blocking behavior causes lengthy stalls during online processing until relocation ends.

4.3.1 *Eventually Consistent Protocol*

It is essential for the operator to continue processing tuples *on-the-fly* while performing adaptations. Achieving this presents new challenges to the correctness of the results. When the operator migrates from one partitioning scheme $\mathcal{M}_i$ to another $\mathcal{M}_{i+1}$, it undergoes a state relocation process. During this, the state of all machines, within the operator, does not represent a state that is consistent with either $\mathcal{M}_i$ or $\mathcal{M}_{i+1}$. Hence, it becomes hard to reason about how new tuples entering the system should be joined. This section presents a non-blocking protocol that allows continuous processing of new tuples during state relocation by reasoning about the state of any tuple circulating the system with the help of *epochs*. This ensures that the system (i) is consistent at all times except during migration, (ii) eventually converges to the consistent target state $\mathcal{M}_{i+1}$, and (iii) produces correct and complete join results in a continuous manner. The operation of the system is divided into *epochs*. Initially, the system is in epoch zero. Whenever the controller decides a mapping change, the system enters a new epoch with incremented index. For example, if the system starts with the mapping $(8, 8)$, later migrates to $(16, 4)$ and finally migrates to $(32, 2)$, the system went through exactly three epochs. All tuples arriving between the first and the second migration decision belong to epoch 1. All tuples arriving after the last mapping-change decision belong to epoch 2. Reshufflers and joiners are not instantaneously aware of the epoch change, but continue to process tuples normally until they receive an epoch change signal along with the new mapping. Whenever a reshuffler routes a tuple to joiners, it tags it with the latest epoch number it is aware of. It is crucial for the correctness of the scheme described shortly to guarantee that all machines are at most one epoch behind the controller. That is, all machines operate on, at most, two different epochs. This is, however, guaranteed theoretically and formalized later in Theorem 4.4.

The migration starts by the controller making the decision. The controller broadcasts to all reshufflers the mapping change signal. When a reshuffler receives this signal, it notifies all joiners and immediately starts sending tuples in accordance to the new mapping. Joiners continuously join incoming tuples and start exchanging migration tuples. Once a joiner has received epoch change signals from *all* reshufflers, it is guaranteed that it will receive no further tuples tagged with the old epoch index. At that point, the joiner proceeds to finalize the migration and notifies the controller once it is done. The controller can only start a new migration once all joiners notify it that they finished the data migration. The subsequent discussion shows how joiners continue processing tuples while guaranteeing consistent state and correct output.

A migration decision partitions the tuples into several sets. During a migration, τ is the set of all tuples received before the migration decision. μ is the set of all tuples that are sent from one joiner to another (due to migration). The set of new tuples received after the migration decision timestamp are either tagged with the old epoch index, referred to as Δ, or with the new epoch index, referred to as Δ'. Notice that $\mu \subset (\tau \cup \Delta)$. To simplify notation, no distinction is made between tuples of R or S. For example, writing $\Delta \bowtie \Delta'$ refers to $(\Delta_R \bowtie \Delta'_S) \cup (\Delta_S \bowtie \Delta'_R)$, where σ_R (σ_S) refers to the tuples of $R(S)$ in the set σ.

During the migration, joiners have tuples tagged with the old epoch and the new epoch. Those tuples tagged with the new epoch are already on the correct machines since the

Algorithm 3 Joiner-Epoch Algorithm.

Input: s signal
Initialize: Use HANDLETUPLE$_1$ to handle incoming tuples.
1: **procedure** MAIN(s)
2: **if** First Reshuffler Signal Received **then**
3: SEND τ for migration.
4: **else if** All Reshuffler Signals Received **then**
5: Use HANDLETUPLE$_2$ to handle incoming tuples.
6: **else if** Migration Ended **then**
7: Run FINALIZEMIGRATION.
8: Use HANDLETUPLE$_1$ to handle incoming tuples.
Input: t an incoming tuple
9: **procedure** HANDLETUPLE$_1$(t)
10: **if** $t \in \mu$ **then** OUTPUT $\{t\} \bowtie \Delta'$; $\mu \leftarrow \mu \cup \{t\}$
11: **else if** $t \in \Delta'$ **then** OUTPUT $\{t\} \bowtie (\mu \cup \Delta')$; $\Delta' \leftarrow \Delta' \cup \{t\}$
12: OUTPUT $\{t\} \bowtie$ KEEP$(\tau \cup \Delta)$
13: **else if** $t \in \Delta$ **then** OUTPUT $\{t\} \bowtie (\tau \cup \Delta)$; $\Delta \leftarrow \Delta \cup \{t\}$
14: **if** $t \in$ KEEP(Δ) **then** OUTPUT $\{t\} \bowtie \Delta'$
15: **if** $t \in$ MIGRATED(Δ) **then** SEND $\{t\}$ for migration
Input: t an incoming tuple
16: **procedure** HANDLETUPLE$_2$(t)
17: **if** $t \in \mu$ **then** OUTPUT $\{t\} \bowtie \Delta'$; $\mu \leftarrow \mu \cup \{t\}$
18: **else if** $t \in \Delta'$ **then** OUTPUT $\{t\} \bowtie (\mu \cup \Delta')$; $\Delta' \leftarrow \Delta' \cup \{t\}$
19: OUTPUT$\{t\} \bowtie$ KEEP$(\tau \cup \Delta)$
20: **procedure** FINALIZEMIGRATION
21: SEND(Ack) signal to coordinator
22: $\tau \leftarrow$ KEEP$(\tau \cup \Delta) \cup \mu \cup \Delta'$
23: $\Delta \leftarrow \emptyset$; $\Delta' \leftarrow \emptyset$; $\mu \leftarrow \emptyset$

reshuffler sent them according to the new mapping. Joiners should redistribute the tuples tagged with old labels according to the new mapping. The set of tuples tagged with the old label is exactly $\tau \cup \Delta$. Joiners discard portions and send other portions to the other machines. The discarded tuples are referred to as DISCARD$(\tau \cup \Delta)$. For convenience, $(\tau \cup \Delta) -$ DISCARD$(\tau \cup \Delta)$ is referred to as KEEP$(\tau \cup \Delta)$. The migrated tuples are MIGRATED$(\tau \cup \Delta)$ which coincides exactly with μ. KEEP(τ) refers to tuples in KEEP$(\tau \cup \Delta) \cap \tau$. The same holds for DISCARD, MIGRATED and the set Δ.

DEFINITION 4.2. *A migration algorithm is said to be correct if right after the completion of a migration, the output of the system is exactly* $(\tau \cup \Delta \cup \Delta') \bowtie (\tau \cup \Delta \cup \Delta')$.

During the migration, the output may be incomplete. Therefore, completeness and consistency are defined only upon the completion of the migration. The complete output is the join of all tuples that arrived to the system before (τ) and after the migration decision $(\Delta \cup \Delta')$. Alg. 3 describes the joiner algorithm whose output is provably correct. For the proof of correctness, we provide an alternative characterization of the output.

LEMMA 4.6.

$$(\tau \cup \Delta \cup \Delta') \bowtie (\tau \cup \Delta \cup \Delta')$$

is equivalent to the union of (1) $\tau \bowtie \tau$*, (2)* $\Delta \bowtie \Delta$*, (3)* $\tau \bowtie \Delta$*, (4)* $\Delta' \bowtie \mu$*, (5)* $\Delta' \bowtie$ KEEP(Δ)*, (6)* $\Delta' \bowtie$ KEEP(τ)*, and (7)* $\Delta' \bowtie \Delta'$.

PROOF. Since set union distributes over join, the result can be rewritten as,

$$(\tau \bowtie \tau) \cup (\tau \bowtie \Delta) \cup (\tau \bowtie \Delta') \cup (\Delta \bowtie \Delta) \cup (\Delta \bowtie \Delta') \cup (\Delta' \bowtie \Delta').$$

Subsets (1), (2), (3) and (7) appear directly in the expression. It remains to argue that $\Delta' \bowtie (\tau \cup \Delta)$ is equal to $\Delta' \bowtie (\mu \cup$ KEEP$(\tau \cup \Delta))$. This follows directly from the correctness of the migration. $\tau \cup \Delta$ is the set of tuples labeled with the old epoch, while $(\mu \cup$ KEEP$(\tau \cup \Delta))$ is the same set distributed differently between the machines according to the new mapping. □

Alg. 3 exploits this equivalence to continue processing tuples throughout migration. Informally, parts (1), (2) and (3) are continuously computed in HANDLETUPLE$_1$ whereas, (4), (5), (6) and (7) are continuously computed in both HANDLETUPLE$_1$ and HANDLETUPLE$_2$.

THEOREM 4.3. *Alg. 3 produces the correct and complete output and ensures eventually consistent state for all joiners.*

PROOF. First, it is easy to see that the data migration is performed correctly. τ is sent immediately at the very beginning (line 3). Tuples of Δ are sent as they are received (line 15). Finally, the discards are done once the migration is over (line 22). By lemma 4.6, the result is the union of:

1. $\tau \bowtie \tau$. This is computed prior to the start of migration.

2. $(\Delta \bowtie \Delta) \cup (\tau \bowtie \Delta)$. Δ is initially empty. Tuples are only added to it in line 13. Every added tuple gets joined with all previously added tuples to Δ and to all tuples in τ (also in line 13). It follows that this part of the join is computed. τ never changes until the migration is finalized.

3. $\Delta' \bowtie (\mu \cup$ KEEP$(\tau \cup \Delta))$. Whenever a tuple is added to Δ' (in lines 11 and 18), it gets joined with $\mu \cup$ KEEP$(\tau \cup \Delta)$ (lines 11, 12, 18 and 19). Whenever a tuple is added to μ (lines 10 and 17), it gets joined with Δ'. Furthermore, tuples added to Δ are joined with Δ' if they are in KEEP(Δ) (line 14). τ never changes until the migration ends.

4. $\Delta' \bowtie \Delta'$. Initially Δ' is empty. Tuples get added to it in lines 11 and 18. Whenever a tuple gets added, it gets joined with all previously added tuples (lines 11 and 18).

Therefore, all parts are computed by the algorithm (completeness). Since the analysis covers all the lines that perform a join, it follows that each of the 4 parts of the result is output exactly once (correctness). Thus, the result of the algorithm is correct right after migration is complete. Tuples tagged with the old epoch (τ and Δ) are migrated correctly. Tuples tagged with the new epoch (Δ') are immediately sent to machines according to the new scheme. Therefore, at the end of migration, the state of all joiners is consistent with the new mapping. □

4.3.2 Theoretical Guarantees Revisited

The guarantees given in Theorem 4.1 assume a blocking operator. During migration, it is required that no tuples are received or processed. However, Alg. 3 continuously processes new tuples while adapting. We set the joiners to process migrated tuples at twice the rate of processing new incoming tuples. We show that, under these settings, the proven guarantees hold. It is clear that the amortized cost is unchanged and remains linear because incoming tuples continue to "pay" for future migration costs. The results for competitiveness, on the other hand, need to be verified.

THEOREM 4.4. *With the non-blocking scheme Alg. 3, the competitive ratio ensured by Theorem 4.1 remains* 1.25[2].

PROOF. We prove that the numbers of tuples, received during migration, $|\Delta R|$ and $|\Delta S|$, are bounded by $|R|$ and $|S|$, respectively. Hence, the 1.25-competitiveness follows immediately (by lemma 4.3).

Consider a migration decision after the system has received $|R|$ and $|S|$ tuples from R and S. Let the current mapping be (n, m). Lemma 4.2 asserts that the optimal

[2]Notice that Theorem 4.4 is based on the assumptions made in Theorem 4.1. However, it naturally follows, that if any of the assumptions are relaxed the competitive ratio is changed accordingly as described in §4.2.2.

Query	Join	Predicate
E_{Q_5}	(Region⋈Nation⋈Supplier)⋈Lineitems	Equi-join
E_{Q_7}	(Supplier⋈Nation)⋈Lineitems	Equi-join
B_{NCI}	Lineitems⋈Lineitems	Band-join
B_{CI}	Lineitems⋈Lineitems	Band-join
Fluct-join	Lineitems⋈Orders	Equi-join

Table 1: Join Queries

	E_{Q_5}			E_{Q_7}		
Zipf	SHJ	DYNAMIC	STATICMID	SHJ	DYNAMIC	STATICMID
$Z=0$	79	168	838*	98	192	210
$Z=1$	79	176	851*	159	183	301
$Z=2$	2742*	158	1425*	191	369	462
$Z=3$	4268*	212	2367*	5462*	334	2610*
$Z=4$	5704*	203	2849*	6385*	415	3502*

* Overflow to disk.

Table 2: Runtime in secs.

mapping is one of (n, m), $(n/2, 2m)$ and $(2n, m/2)$. This is trivially true for the first migration. Since we prove below that $|\Delta R|$ and $|\Delta S|$ are bounded by $|R|$ and $|S|$, this also holds for all subsequent migrations, inductively. Without loss of generality, let the chosen optimal mapping for a subsequent migration be $(n/2, 2m)$. The migration process lasts for $2|R|/n$ time units (by lemma 4.4). Alg. 3 processes new tuples at half the rate of processing migrated tuples. Thus, during migration, the operator receives *at most* $1/2 \cdot (n/2)$ new tuples from R and $1/2 \cdot (2m)$ from S per time unit. Hence, it holds that,

$$|\Delta R| \leq \tfrac{2|R|}{n} \cdot \tfrac{n}{4} < |R| \text{ and } |\Delta S| \leq \tfrac{2|R|}{n} \cdot m \leq \tfrac{|S|}{m} \cdot m = |S|$$

where the last inequality holds by lemma 4.1 (with the optimal being $(n/2, 2m)$ instead of (n, m)). □

5. EVALUATION

Environment. Our experimental platform consists of an Oracle Blade 6000 Chassis with 10 Oracle X6270 M2 blade servers. Each blade has two Intel Xeon X5675 CPUs running at 3GHz, each with 6 cores and 2 hardware threads per core, 72GB of DDR3 RAM, 4 SATA 3 hard disks of 500GB each, and a 1Gbit Ethernet interface. All blades run Solaris 10, which offers Solaris Zones, a native resource management and containment solution. Overall, there are 220 virtual machines available exclusively for our experiments, each with its own CPU hardware thread and dedicated memory resources. There are 20 separate hardware threads for running instances of the host operating system.

Datasets. For the evaluation setup, we use the TPC-H benchmark [1]. We employ the TPC-H generator proposed by [10] to generate databases with different degrees of skew under the *Zipf* distribution. The degree of skew is adjusted by choosing a value for the *Zipf* skew parameter z. We experiment on five different skew settings Z_0, Z_1, Z_2, Z_3, Z_4 which correspond to $z = 0$, $z = 0.25, z = 0.5, z = 0.75$ and $z = 1.0$ respectively. We build eight databases with sizes $8, 10, 20, 40, 80, 160, 320$, and 640GB.

Queries. We experiment on four join queries, namely, two equi-joins from the TPC-H benchmark and two synthetic band-joins. The equi-joins, E_{Q_5} and E_{Q_7}, represent the most expensive join operation in queries Q_5 and Q_7 respectively from the benchmark. All intermediate results are materialized before online processing. Moreover, the two band-joins depict two different workload settings. *a*) B_{CI} is a *high*-selectivity join query that represents a computation-intensive workload, and *b*) B_{NCI} is a *low*-selectivity join query that corresponds to a *non*-computation-intensive workload. The output of B_{CI} is three orders of magnitude bigger than its input size, whereas the output of B_{NCI} is an order of magnitude smaller. Both join queries are described in [15] and all query characteristics are summarized in Table 1.

Operators. To run the testbed, we implement SQUALL[3], a distributed online query processing engine built on STORM[4], Twitter's backend engine for data analytics. The engine is based on Java and runs on JRE v1.7. Throughout the discussion, we use four different dataflow operators: (i) STATICMID, a static operator with a fixed $(\sqrt{J}, \sqrt{J})$-mapping. This scheme assumes that both input streams have the same size and lies in the center of the (n, m)-mapping spectrum. (ii) DYNAMIC, our adaptive operator, initialized with the $(\sqrt{J}, \sqrt{J})$-mapping scheme. (iii) STATICOPT, another static operator with a fixed optimal mapping scheme. This requires knowledge about the input stream sizes before execution, which is practically *unattainable* in an online setting. (iv) SHJ, the parallel symmetric hash-join operator described in [18]. This operator can only be used for equi-join predicates and it is *content-sensitive* as it partitions data on the join key. STATICMID, assumes as a best guess, that the streams are equal in size; hence it has a square grid partitioning scheme, i.e., $(\sqrt{J}, \sqrt{J})$. Comparing against STATICOPT shows that our operator does not perform much worse than an omniscient operator with oracle knowledge about stream sizes, which are unknown beforehand. Joiners perform the local join in memory, but if it runs out of memory it begins spilling to disk. For this purpose, we integrated the operators with the back-end storage engine BERKELEYDB [27]. We first experimentally verify that, in case of overflow to disk, machines suffer from long delayed join evaluation and performance hits. Then, for a more fair comparison, we introduce more memory resources, such that all operations fit in memory if possible. The heap size of each joiner is set to 2GB. As indexes, joiners use balanced binary trees for band joins and hashmaps for equi-joins. Input data rates are set such that joiners are fully utilized.

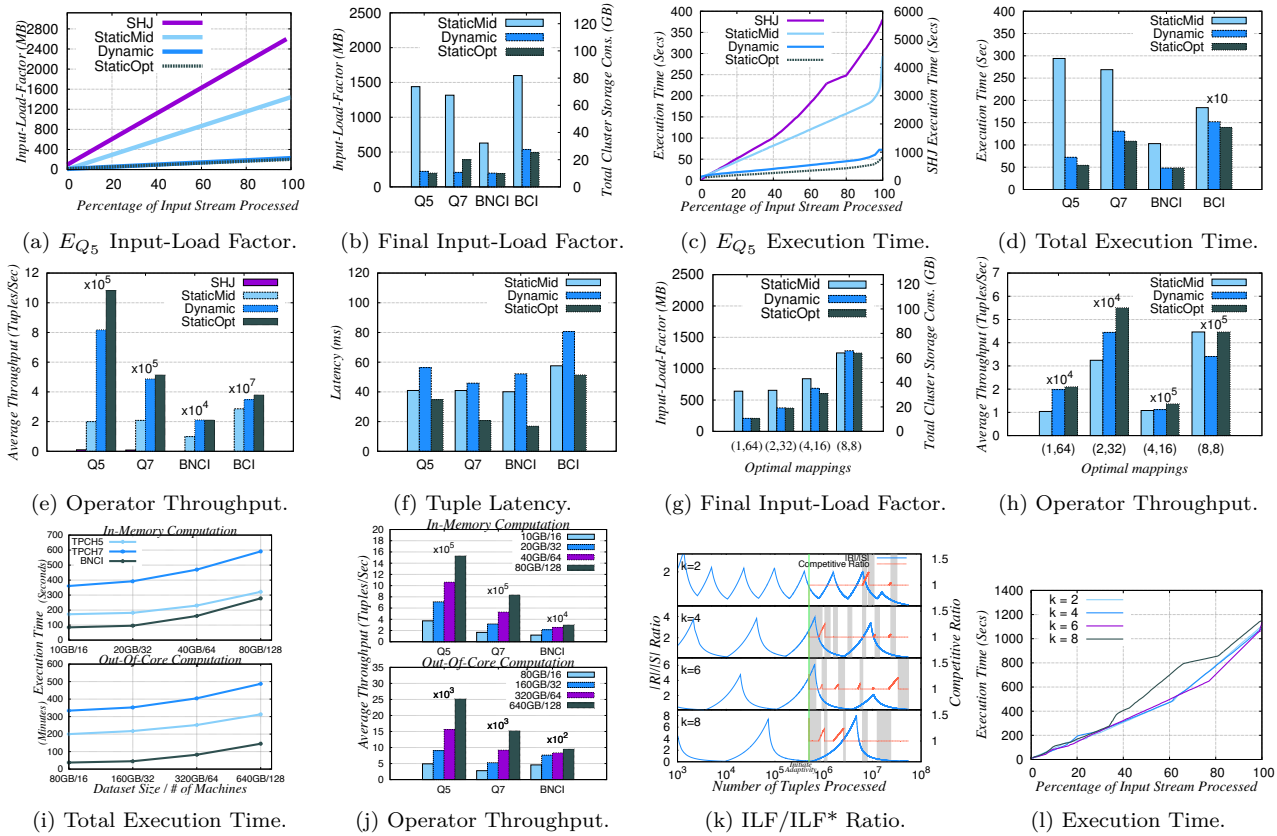
5.1 Skew Resilience

Table 2 shows results for running joins E_{Q_5} and E_{Q_7} with different skew settings of the 10G dataset. It compares the performance of our DYNAMIC operator against the SHJ operator using 16 machines. We report the final execution time. We observe that SHJ performs well under non-skewed settings as it evenly partitions data among machines and does not replicate data. On the other hand, the DYNAMIC operator, distributes workload fairly between machines, but pays for the unnecessary overhead of replicating data. As data gets skewed, SHJ begins to suffer from poor partitioning and unbalanced distribution of data among joiners. Thus, the progress of join execution is dominated by a few overwhelmed workers, while the remaining starve for more data. The busy workers are congested with input data and must overflow to disk, hindering the performance severely. In contrast, the DYNAMIC operator is resilient to data skew and persists to partition data equally among joiners.

5.2 Performance Evaluation

We analyze in detail the performance of static dataflow operators against their adaptive counterpart. We report the results for E_{Q_5} and E_{Q_7} on a Z_4 10G dataset and of B_{NCI}

[3]https://github.com/epfldata/squall/wiki
[4]https://github.com/nathanmarz/storm



(a) E_{Q_5} Input-Load Factor. (b) Final Input-Load Factor. (c) E_{Q_5} Execution Time. (d) Total Execution Time.

(e) Operator Throughput. (f) Tuple Latency. (g) Final Input-Load Factor. (h) Operator Throughput.

(i) Total Execution Time. (j) Operator Throughput. (k) ILF/ILF* Ratio. (l) Execution Time.

Figure 5: Performance Evaluation

and B_{CI} on a uniform (Z_0) 10G dataset. We start by comparing performance using 16 machines. As illustrated in Table 2, DYNAMIC operates efficiently, whereas STATICMID consistently performs worse. For skewed data, the latter suffers from very high values of ILF, and thus, overflows to disk, hindering the performance drastically. For a fair comparison, we increase the number of machines to 64 such that STATICMID is given enough resources. Under this setting, STATICMID has a fixed $(8, 8)$-mapping scheme, whereas the optimal mapping scheme for all joins is $(1, 64)$. Our results show that DYNAMIC behaves roughly the same as STATICOPT. This is attributed to the fact that DYNAMIC migrates to the optimal mapping scheme at early stages. For completeness, we also include the results for E_{Q_5} and E_{Q_7} using SHJ. The operator overflows to disk due to high data skew.

Input-Load Factor. As described in §3.3, different mappings incur different values for the input-load factor. Examining the average input-load factor for each operator shows that the growth rate of the ILF is linear over time. Due to the lack of space, we illustrate this behavior for E_{Q_5} only. Fig. 5a plots the maximum size of ILF per machine against the percentage of total input stream processed. SHJ and STATICMID suffer from a larger growth rate than DYNAMIC. Specifically, their rates are 27, 14 and 2MB per 1% input stream processed, respectively. The graphs depicted in Fig. 5b report on the final average ILF per machine for all the join queries. STATICMID is consistently accompanied with larger ILF values. Its ILF is about 3 to 7 times that of DYNAMIC. The optimal mapping $(1, 64)$ lies at one end of the mapping spectrum and is far from that of STATICMID. And SHJ is up to 13 times that of the other operators.

§3.3 also emphasizes the fact that minimizing the ILF maximizes resource utilization and performance. This is due to the fact that higher ILF values also imply (i) unnecessary replicated data stored around the cluster, (ii) more duplicate messages sent congesting the network, and (iii) additional overhead for processing and housekeeping replicated data at each joiner. In what follows, we measure the impact of ILF on overall operator performance.

Resource Utilization. Fig. 5b also shows the total cluster storage consumption (GB), as shown on the right axis. STATICMID's fixed partitioning scheme misuses allocated resources as it unnecessarily consumes more storage and network bandwidth to spread the data. Moreover, it requires four times more machines (64) than DYNAMIC to operate fully in memory (16 machines used in Table 2). SHJ could not fully operate in memory even with 64 machines. DYNAMIC performs efficiently in terms of resource utilization. This is essential for cloud infrastructures which typically follow *pay-as-you-go* policies.

Execution Time. Fig. 5c shows the execution time to process the input stream for query E_{Q_5}. The other join queries are similar in behavior and we omit them due to the lack of space. Fig. 5d shows the total execution time for all the join queries. We observe that execution time is linear in the percentage of input stream processed. The ILF has a decisive effect on processing time. The rigid assignment $(8, 8)$ of STATICMID yields high ILF values and leads to consistently worse performance. As ILF grows, the amount of data to process, and hence, processing time increases. However, this

performance gap is not large when the join operation is computationally intensive, i.e., B_{CI} in Fig. 5d. The execution time for SHJ, shown at the right axis of Fig. 5c, is two orders of magnitude more, illustrating that poor resource utilization may push the operator to disk spills, hindering the performance severely. In all cases, the adaptivity of DYNAMIC allows it to perform very close to STATICOPT.

Average Throughput and Latency. Fig. 5e shows global operator throughput. For all queries, the throughputs of DYNAMIC and STATICOPT are close. They are at least twice that of STATICMID, and two orders of magnitude more than that of SHJ, except for B_{CI} where the difference is slight. This validates the fact that the ILF has a direct effect on throughput, and that the effect is magnified when overflow occurs. The throughput gap between operators depends on the amount of join computation a machine has to perform (e.g. compare B_{CI} and B_{NCI}). Fig 5f shows average tuple latencies. We define latency as the difference between the time an output tuple t is emitted and the time at which the (more recent) corresponding source input tuple arrives to the operator. The figure shows that the operator latency is not greatly affected by its adaptivity. During state migration, an additional network hop increases the tuple latency. DYNAMIC achieves average latency close to that of STATICMID while attaining much better throughput.

Different Optimal Mappings. So far, the join queries we experiment on capture the interesting case of an optimal mapping that is far from the $(\sqrt{J}, \sqrt{J})$ scheme. As illustrated in Figs. 5g, 5h, we compare performance under various optimal mapping settings. We achieve this by increasing the size of the smaller input stream. In all cases, DYNAMIC adjusts itself to the optimal mapping at early stages. Fig. 5g shows how the input-load factor gap between DYNAMIC and STATICMID decreases as the optimal mapping gets closer to the $(\sqrt{J}, \sqrt{J})$-mapping scheme. Similarly, Fig. 5h illustrates how the performance gap decreases between the two operators. This validates the fact that the input-load factor has a decisive effect on performance. In case of the optimal $(\sqrt{J}, \sqrt{J})$-mapping scheme, STATICOPT has the same mapping as STATICMID, whereas DYNAMIC does not deviate from its initial mapping scheme. However, it performs slightly worse because adaptivity comes with a small cost.

5.3 Scalability Results

We evaluate the scalability of DYNAMIC. Specifically, we measure operator execution time and throughput as both the data-size and parallelism configurations grow. We evaluate weak scalability on 10GB/16 joiners, 20GB/32 joiners, and so forth as illustrated in the in-memory computation graphs of Figs. 5i, 5j. Ideally, while increasing the data-size/joiners configuration, the input-load factor and the output size should remain constant per joiner. However, the input-load factor grows, preventing the operator to achieve perfect scalability (same execution time and double average throughput as the data-size/joiners double). For example, for B_{NCI}, under the 20GB/32 configuration, the input stream sizes are 0.68M (million) and 30M tuples, respectively, yielding a $(1, 32)$ optimal mapping scheme with an ILF of $0.68M + 30M/32 = 1.61M \cdot size_{tuple}$ per joiner. However, under the 40GB/64 configuration, the input stream sizes are 1.36M and 60M, respectively, yielding a $(1, 64)$ optimal mapping scheme with an ILF of $1.36M + 60M/64 = 2.29M \cdot size_{tuple}$. In both cases, the output size per joiner is the same (64K tuples). However, the ILF differs by 42% because of the replication of the smaller relation. The ILF for the other two joins does not grow more than 9%. Accordingly, the execution time (Fig. 5i) and the average throughput (Fig. 5j) graphs show that E_{Q_5} and E_{Q_7} achieve almost perfect scalability. In case of B_{NCI}, a joiner processes more input tuples as data grows. Overall, the operator achieves good scalability taking into account the increase in ILF.

Secondary Storage. Out-of-core computation in Figs. 5i, 5j illustrates performance under weak scalability with secondary storage support. As before, all the queries achieve ideal scalability, taking into account the increase in ILF. This validates the fact that our system can scale with large volumes of data, and that it works well regardless of the local join algorithm. However, compared to the *in-memory* results (Fig. 5i), the performance drops by an order of magnitude which validates the conclusion that secondary storage is not perfectly suited for high-performance online processing.

5.4 Data Dynamics

In order to validate the proven theoretical guarantees, we evaluate the performance of DYNAMIC under severe fluctuations in data arrival rates. We simulate a scenario where the cardinality aspect ratios keep on alternating between k and $1/k$ where k is the fluctuation rate. Data from the first relation is streamed into the operator until its cardinality is k times that of the second one. Then, the roles are swapped, by quiescing the first input stream and allowing data to stream in from the second until its cardinality is k times that of the first. This fluctuation continues until the streams are finished. We experiment on an 8G dataset using the *Fluct-Join* query defined in [15] on 64 machines. We run the query under various fluctuation factor, specifically, $k = 2$, $k = 4$, $k = 6$ and $k = 8$. We set the operator to begin adapting after it has received at least 500K tuples, corresponding to less than 1% of the total input.

Analysis. The first metric of interest is the ILF competitive ratio of DYNAMIC in comparison to an *oracle* that assigns the optimal mapping, and thus optimal ILF*, instantly with no cost at all times. Fig. 5k plots both the $|R| / |S|$, on the left axis, and the ILF/ILF* ratio, on the right axis, throughout query execution. In the graph, migration durations are depicted by the shaded regions. We observe that the ratio never exceeds 1.25 at all times which validates the result of Theorem 4.4. Even under severe fluctuations, the operator is well advised in choosing the right moments to adapt. Fig. 5l shows the execution time progress under different fluctuation factors. Although DYNAMIC undergoes many migrations, it persists to progress linearly showing that all migration costs are amortized. This verifies the results of Lemma 4.5 and Theorem 4.1.

5.5 Summary

Experiments show that our adaptive operator outperforms *practical* static schemes in every performance measure without sacrificing low latency. They emphasize the effect of ILF on resource utilization and performance. This validates the optimization goal of minimizing ILF as a direct performance measure. Our operator ensures efficient resource utilization in storage consumption and network bandwidth that is up to 7 times less than non-adaptive theta-join counterparts. Non-adaptivity causes misuse of allocated resources leading to overflows. Even when provided enough resources, the adaptive operator completes the join up to 4 times faster with an average throughput of up to 4 times more. Adaptivity is achieved at the cost of slight increase in tuple latency (by as little as 5ms and at most 20ms). Experiments also show that our operator is scalable. Under severe data

fluctuations, the operator adapts to data dynamics with the ILF remaining within the proven bounds from the optimum and with amortized linear migration costs. Additionally, the operator, being *content-insensitive*, is resilient to data skew while *content-sensitive* operators suffer from overflows, hindering performance by up to two orders of magnitude.

6. ACKNOWLEDGMENTS

This work was supported by ERC Grant 279804.

7. REFERENCES

[1] The TPC-H benchmark. `http://www.tpc.org/tpch/`.
[2] D. Abadi, Y. Ahmad, M. Balazinska, U. Çetintemel, M. Cherniack, J. Hwang, W. Lindner, A. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik. The design of the Borealis stream processing engine. In *CIDR*, pages 277-289, 2005.
[3] F. Afrati and J. Ullman. Optimizing joins in a MapReduce environment. In *EDBT*, pages 99-110, 2010.
[4] A. Arasu, B. Babcock, S. Babu, J. Cieslewicz, M. Datar, K. Ito, R. Motwani, U. Srivastava, and J. Widom. STREAM: The Stanford data stream management system. Technical report, Stanford InfoLab, 2004.
[5] A. Arasu, M. Cherniack, E. Galvez, D. Maier, A. Maskey, E. Ryvkina, M. Stonebraker, R. Tibbetts. Linear road: a stream data management benchmark. In *VLDB*, pages 480-491, 2004.
[6] R. Avnur and J. Hellerstein. Eddies: continuously adaptive query processing. In *SIGMOD*, pages 261-272, 2000.
[7] S. Babu and P. Bizarro. Adaptive query processing in the looking glass. In *CIDR*, pages 238-249, 2005.
[8] S. Blanas, J. Patel, V. Ercegovac, J. Rao, E. Shekita, and Y. Tian. A comparison of join algorithms for log processing in MapReduce. In *SIGMOD*, pages 975-986, 2010.
[9] R. Fernandez, M. Migliavacca, E. Kalyvianaki and P. Pietzuch. Integrating scale out and fault tolerance in stream processing using operator state management. In *SIGMOD*, pages 725-736, 2013.
[10] S. Chaudhuri and V. Narasayya. TPC-D data generation with skew.
[11] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In *OSDI*, pages 10-10, 2004.
[12] A. Deshpande and J. Hellerstein. Lifting the burden of history from adaptive query processing. In *VLDB*, pages 948-959, 2004.
[13] A. Deshpande, Z. Ives, and V. Raman. Adaptive query processing. *Foundations and Trends in Databases*, 1(1):1-140, 2007.
[14] J. Dittrich, B. Seeger, D. Taylor, and P. Widmayer. Progressive merge join: a generic and non-blocking sort-based join algorithm. In *VLDB*, pages 299-310, 2002.
[15] M. Elseidy, A. Elguindy, A. Vitorovic, and C. Koch. Scalable and adaptive online joins. EPFL-REPORT 190035 *Technical Report*, 2013.
[16] A. Gounaris, N. Paton, A. Fernandes, and R. Sakellariou. Adaptive query processing: A survey. In *British National Conference on Databases*, pages 11-25, 2002.
[17] A. Gounaris, E. Tsamoura, and Y. Manolopoulos. Adaptive query processing in distributed settings. *Advanced Query Processing*, 36(1):211-236, 2012.
[18] G. Graefe. Query evaluation techniques for large databases. *ACM Computing Surveys*, 25(2):73-169, 1993.
[19] X. Gu, P. Yu, and H. Wang. Adaptive load diffusion for multiway windowed stream joins. In *ICDE*, pages 146-155, 2007.
[20] P. Haas and J. Hellerstein. Ripple joins for online aggregation. In *SIGMOD*, pages 287-298, 1999.
[21] J. Hellerstein, M. Franklin, S. Chandrasekaran, A. Deshpande, K. Hildrum, S. Madden, V. Raman, and M. Shah. Adaptive query processing: Technology in evolution. *IEEE Data Engineering Bulletin*, 23(2), 2000.
[22] J. Hellerstein, P. Haas, and H. Wang. Online aggregation. In *SIGMOD*, pages 171-182, 1997.
[23] Y. Ioannidis and S. Christodoulakis. On the propagation of errors in the size of join results. In *SIGMOD*, pages 268-277, 1991.
[24] B. Liu, M. Jbantova, and E. Rundensteiner. Optimizing state-intensive non-blocking queries using run-time adaptation. In *ICDE Workshop*, page 614-623, 2007.
[25] M. Mokbel, M. Lu, and W. Aref. Hash-Merge join: A non-blocking join algorithm for producing fast and early join results. In *ICDE*, pages 251-262, 2004.
[26] A. Okcan and M. Riedewald. Processing theta-joins using MapReduce. In *SIGMOD*, pages 949-960, 2011.
[27] M. Olson, K. Bostic, and M. Seltzer. Berkeley DB. In *Annual Technical Conference*, USENIX, pages 43-43, 1999.
[28] C. Olston, B. Reed, A. Silberstein, and U. Srivastava. Automatic optimization of parallel dataflow programs. In *Annual Technical Conference*, USENIX, pages 267-273, 2008.
[29] N. Paton, J. Buenabad, M. Chen, V. Raman, G. Swart, I. Narang, D. Yellin, and A. Fernandes. Autonomic query parallelization using non-dedicated computers: an evaluation of adaptivity options. *VLDBJ*, 18(1):119-140, 2009.
[30] D. Schneider and D. DeWitt. A performance evaluation of four parallel join algorithms in a shared-nothing multiprocessor environment. In *SIGMOD*, pages 110-121, 1989.
[31] M. Shah, J. Hellerstein, S. Chandrasekaran, and M. Franklin. Flux: An adaptive partitioning operator for continuous query systems. In *ICDE*, pages 25-36, 2002.
[32] J. Stamos and H. Young. A symmetric fragment and replicate algorithm for distributed joins. *Transactions on Parallel and Distributed Systems*, 4(12):1345-1354, 1993.
[33] M. Stillger, G. Lohman, V. Markl, and M. Kandil. LEO - DB2's learning optimizer. In *VLDB*, pages 19-28, 2001.
[34] Y. Tao, M. L. Yiu, D. Papadias, M. Hadjieleftheriou, and N. Mamoulis. RPJ: producing fast join results on streams through rate-based optimization. In *SIGMOD*, pages 371-382, 2005.
[35] F. Tian and D. DeWitt. Tuple routing strategies for distributed eddies. In *VLDB*, pages 333-344, 2003.
[36] P. Upadhyaya, Y. Kwon, and M. Balazinska. A latency and fault-tolerance optimizer for online parallel query plans. In *SIGMOD*, pages 241-252, 2011.
[37] T. Urhan and M. Franklin. XJoin: A reactively-scheduled pipelined join operator. *IEEE Data Engineering Bulletin*, 23(2):27-33, 2000.
[38] S. Wang and E. Rundensteiner. Scalable stream join processing with expensive predicates: workload distribution and adaptation by time-slicing. In *EDBT*, pages 299-310, 2009.
[39] A. Wilschut and P. Apers. Dataflow query execution in a parallel main-memory environment. In *Parallel and Distributed Information Systems*, pages 68-77, 1991.
[40] Y. Xing, S. Zdonik, and J. Hwang. Dynamic load distribution in the Borealis stream processor. In *ICDE*, pages 791-802, 2005.
[41] H. Yang, A. Dasdan, R. Hsiao, and D. Parker. Map-Reduce-Merge: simplified relational data processing on large clusters. In *SIGMOD*, pages 1029-1040, 2007.
[42] X. Zhang, L. Chen, and M. Wang. Efficient multi-way theta-join processing using MapReduce. *VLDBJ*, 5(11):1184-1195, 2012.
[43] Y. Zhou, B. Ooi, and K. Tan. Dynamic load management for distributed continuous query systems. In *ICDE*, 2005.