Operator Scheduling in a Data Stream Manager*

Don Carney[†], Uğur Çetintemel[†], Alex Rasin[†], Stan Zdonik[†]
Mitch Cherniack[§], Mike Stonebraker[±]

[†]{dpc,ugur,alexr,sbz}@cs.brown.edu, Department of Computer Science, Brown University
[§]mfc@cs.brown.edu, Department of Computer Science, Brandeis University
[±] stonebraker@lcs.mit.edu Laboratory for Computer Science & Department of EECS, M.I.T.

Abstract

Many stream-based applications have sophisticated data processing requirements and real-time performance expectations that need to be met under high-volume, time-varying data streams. In order to address these challenges, we propose novel operator scheduling approaches that specify (1) which operators to schedule (2) in which order to schedule the operators, and (3) how many tuples to process at each execution step. We study our approaches in the context of the Aurora data stream manager.

We argue that a fine-grained scheduling approach in combination with various scheduling techniques (such as batching of operators and tuples) can significantly improve system efficiency by reducing various system overheads. We also discuss application-aware extensions that make scheduling decisions according to per-application Quality of Service (QoS) specifications. Finally, we present prototype-based experimental results that characterize the efficiency and effectiveness of our approaches under various stream workloads and processing scenarios.

1 Introduction

Applications that deal with potentially unbounded, continuous streams of data are becoming increasingly popular due to a confluence of advances in real-time, wide-area data dissemination technologies and the emergence of small-scale computing devices (such as GPSs and micro-sensors) that continually emit data obtained from their physical environment. Example applications include sensor networks, position tracking, fabrication line management, network management, and financial portfolio management. All these applications require timely processing of large numbers of continuous, potentially rapid and asynchronous data streams. Hereafter, we refer to such applications as *stream-based* applications.

The Aurora data stream manager [1, 8] addresses the performance and processing requirements of stream-based applications. Aurora supports multiple concurrent continuous queries, each of which produces results to one or more stream-based applications. Each continuous query consists of a directed acyclic graph of a well-defined set of operators (or *boxes* in Aurora terminology). Applications define their service expectations using Quality-of-Service (QoS) specifications, which guide Aurora's resource allocation decisions. We provide an overview of Aurora in Section 2.

A key component of Aurora, or any data-stream management system for that matter, is the scheduler that controls processor allocation. The scheduler is responsible for multiplexing the processor usage to multiple queries according to application-level performance or fairness goals. Simple processor allocation can be performed by assigning a thread per operator or per query. Such standard techniques do not scale since no system that we are aware of can adequately deal with a very large number of threads. More importantly, any such approach would abdicate the details of scheduling to the operating system, thereby making it impossible to account for application-level constraints (QoS). We demonstrate this quantitatively in Section 6.

This paper shows that having finer-grained control over processor allocation can make a significant difference to overall system performance by reducing various system overheads associated with continuous query execution. We propose a set of novel scheduling techniques that address scheduler overheads by batching (of operators and tuples), approximation, and pre-computation.

* This work has been supported by the NSF under grant IIS-0086057.

**Proceedings of the 29th VLDB Conference,
Berlin, Germany, 2003**

In particular, we describe the design and implementation of the Aurora scheduler, which performs the following tasks:

1. *Dynamic scheduling-plan construction*: The scheduler develops a scheduling plan that specifies, at each scheduling point, (1) which boxes to schedule, (2) in which order to schedule the boxes, and (3) how many tuples to process at each box execution.
2. *QoS-aware scheduling*: The Aurora scheduler strives to maximize the overall QoS delivered to the client applications. At a high level, our scheduling decisions are based on a novel box priority assignment technique that uses the latencies of queued tuples and application-specific QoS information. For improved scalability, we also use an approximation technique, based on bucketing and pre-computation, which trades scheduling quality with scheduling overhead.

We also evaluate and experimentally compare these algorithms on our Aurora prototype under various stream processing and workload scenarios. Through the implementation of our techniques on the prototype rather than a simulator, we were better able to understand the actual system overheads.

The rest of the paper is organized as follows: Section 2 provides an overview of the Aurora data stream manager. Section 3 describes the state-based execution model used by Aurora. Section 4 discusses in detail Aurora's scheduling algorithms. Section 5 discusses our prototype-based experimental study that provides quantitative evidence regarding the efficiency and effectiveness of Aurora's scheduling algorithms. Section 6 extends our basic approaches to address QoS, describing queue-based priority assignment and an approximation technique for improving the scalability of the system. Section 7 describes related work, and Section 8 concludes the paper.

2 Aurora Overview

2.1 Basic Model

Aurora data is assumed to come from a variety of data sources such as computer programs that generate values (at regular or irregular intervals) or hardware *sensors*. We will use the term *data source* for either case. In addition, a data *stream* is the term we will use for the collection of data values that are presented by a data source. Each data source is assumed to have a unique source identifier and Aurora timestamps every incoming tuple to monitor the prevailing QoS.

The basic job of Aurora is to process incoming streams in the way defined by an *application administrator*. Figure 1 illustrates Aurora's high-level system model. Aurora is fundamentally a data-flow system and uses the popular *boxes and arrows* paradigm found in most process flow and workflow systems. Hence, tuples flow

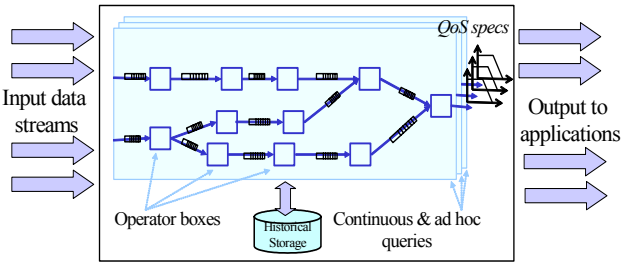

Figure 1: Aurora system model

through a loop-free, directed graph of processing operations (a.k.a. *boxes*). Ultimately, output streams are presented to *applications*, which must be programmed to deal with output tuples that are generated asynchronously. Aurora can also maintain historical storage, primarily in order to support ad-hoc queries.

Tuples generated by data sources arrive at the input and are queued for processing. The *scheduler* selects boxes with waiting tuples and executes them on one or more of their input tuples. The output tuples of a box are queued at the input of the next box in sequence. In this way, tuples make their way from the inputs to the outputs. Each output is associated with one or more QoS specifications, which define the utility of stale or imprecise results to the corresponding application.

The primary performance-related QoS is based on the notion of the *latency* (i.e., delay) of output tuples—output tuples should be produced in a timely fashion, otherwise, QoS will degrade as latencies get longer. In this paper, we will only deal with latency-based QoS graphs; for a discussion of other types of QoS graphs and how they are utilized, please refer to [2, 8]. Aurora assumes that all QoS graphs are normalized, and are thus quantitatively comparable. Aurora further assumes that the QoS requirements are *feasible*; i.e., under normal operation (i.e., no peak overload), an idealized scheduler will be able to deliver maximum possible QoS for each individual output.

Aurora contains built-in support for a set of primitive operations for expressing its stream processing requirements. Some operators manipulate the items in the stream, others transform individual items in the stream to other items, while other operators, such as the aggregates (e.g., moving average), apply a function across a window of values in a stream. A description of the operators is outside the scope of this paper and can be found in [2, 8].

2.2 Architecture

Figure 2 illustrates the architecture of the basic Aurora run-time engine. Here, inputs from data sources and outputs from boxes are fed to the router, which forwards them either to external applications or to the storage manager to be placed on the proper queues. The storage manager is responsible for maintaining the box queues and managing the buffer, properly making tuple queues available for read and write by operators. Conceptually, the scheduler picks a box for execution, ascertains *how*

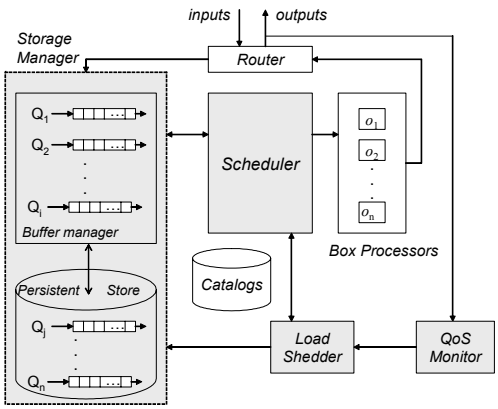

Figure 2: Aurora run-time engine

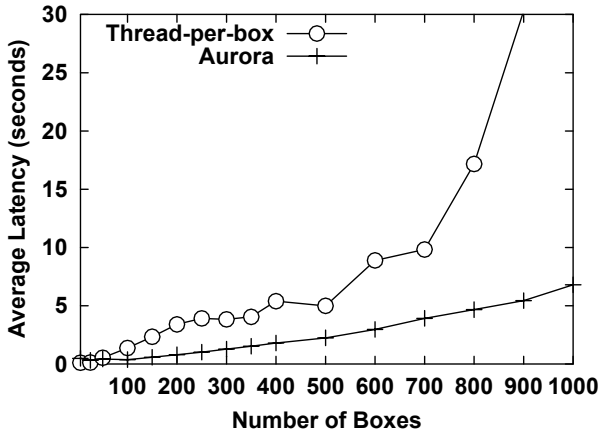

Figure 3: High-level comparison of stream execution models

many tuples to process from its corresponding input queue, and passes a pointer to the box description (together with a pointer to the box state) to the multi-threaded box processor. The box processor executes the appropriate operation and then forwards the output tuples to the router. The scheduler then ascertains the next processing step and the cycle repeats.

The QoS monitor continually monitors system performance and activates the load shedder when it detects an overload situation and poor system performance. The load shedder sheds load until the performance of the system reaches an acceptable level. The catalog contains information regarding the network topology, inputs, outputs, QoS information, and relevant statistics (e.g., selectivity, average box processing costs), and is essentially used by all components.

3 Basic Execution Model

The traditional model for structuring database servers is *thread-based execution*, which is supported widely by traditional programming languages and environments. The basic approach is to assign a thread to each query or operator. The operating system (OS) is responsible for providing a virtual machine for each thread and overlapping computation and I/O by switching among the threads. The primary advantage of this model is that it is very easy to program, as OS does most of the job. On the other hand, especially when the number of threads is large, the thread-based execution model incurs significant overhead due to cache misses, lock contention, and switching. More importantly for our purposes, the OS handles the scheduling and does not allow the overlaying software to have fine-grained control over resource management.

Instead, Aurora uses a *state-based execution* model. In this model, there is a single scheduler thread that tracks system state and maintains the execution queue. The execution queue is shared among a small number of worker threads responsible for executing the queue entries (as we discuss below, each entry is a sequence of boxes). This state-based model avoids the mentioned limitations of the thread-based model, enabling fine-grained allocation of resources according to application-specific targets (such as QoS). Furthermore, this model also enables effective *batching* of operators and tuples, which we show has drastic effects on the performance of the system as it cuts down the scheduling and box execution overheads.

In order to illustrate the basic performance benefits of Aurora's state-based model over the traditional thread-based model (where each operator is assigned a single thread), we ran a simple processing network consisting of multiple queries, each consisting of a chain of five filter operators (see Section 5.1 for a description of our experimental testbed). Figure 3 shows the tuple latencies observed as a function of the total number of operators. As we increase the system workload in terms of number of operators, the performance degrades in both cases, however much less in the Aurora case. In fact, performance degrades almost linearly in Aurora and exponentially in the thread-per-box case, a result that clearly supports the aforementioned scalability arguments.

An important challenge with the state-based model is that of designing an intelligent but low-overhead scheduler. In this model, the scheduler becomes solely responsible for keeping track of system context and deciding when and for how long to execute each operator. In order to meet application-specific QoS requirements, the scheduler should carefully multiplex the processing of multiple continuous queries. At the same time, the scheduler should try to minimize the system overheads, time not spent doing "useful work" (i.e., processing), with no or acceptable degradation in its effectiveness.

4 Two-Level Scheduling

Aurora uses a two-level scheduling approach to address the execution of multiple simultaneous queries. The first-level decision involves determining *which* continuous

query to process. This is followed by a second-level decision that then decides *how* exactly the selected query should be processed. The former decision entails dynamically assigning priorities to operators at run-time, typically according to QoS specifications, whereas the latter decision entails choosing the order in which the operators involved in the query will be executed. The outcome of these decisions is a sequence of operators, referred to as a *scheduling plan*, to be executed one after another. Scheduling plans are inserted into an execution queue to be picked up and executed by a worker thread.

In order to reduce the scheduling and operator overheads, Aurora heavily relies on batching (i.e., grouping) during scheduling. We developed and implemented algorithms that batch both operators and tuples. In both cases, we observed significant performance gains over the non-batching cases. We now describe in detail our batching approaches for constructing scheduling plans.

4.1 Operator Batching - Superbox Processing

A *superbox* is a sequence of boxes that is scheduled and executed as an atomic group. Superboxes are useful for decreasing the overall execution costs and improving scalability as (1) they significantly reduce scheduling overheads by scheduling multiple boxes as a single unit, thereby decreasing the number of scheduling steps; (2) they eliminate the need to access the storage manager for each individual box execution by having the storage manager allocate memory and load all the required tuples at once[1].

Conceptually, a superbox can be an arbitrary connected subset of the Aurora network. However, we do constrain the form of superboxes such that each is always a tree of boxes rooted at an *output box* (i.e., a box whose output tuples are forwarded to an external application). The reasons that underlie this constraint are twofold. First, only the tuples that are produced by an output box provide any utility value to the system. Second, even though allowing arbitrary superboxes provide the highest flexibility and increase opportunities for optimization, this will also make the search space for superbox selection intractable for large Aurora networks.

The following subsections discuss the two key issues to deal with when scheduling superboxes, namely *superbox selection* and *superbox traversal*.

4.1.1 Superbox Selection

The first-level scheduling issue involves determining which superboxes to schedule. Fundamentally, there are two different approaches to superbox selection: *static* and *dynamic*. Static approaches identify potential superboxes statically before run time, whereas the dynamic approaches identify useful superboxes at run time.

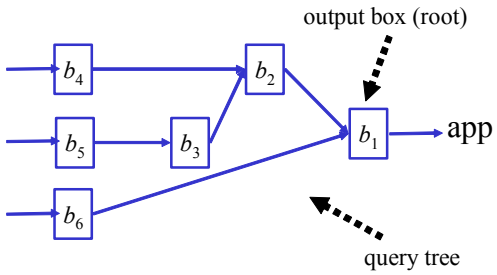

Figure 4: Sample query tree

In Aurora, we implemented a static superbox selection approach, called *application-at-a-time* (*AAAT*). AAAT statically defines one superbox for each query tree. As a result, the number of superboxes is always equal to the number of continuous queries (or applications) in the Aurora network. Figure 4 illustrates a simple query tree that consists of six boxes (the tree is rooted at box b_1). Once the superboxes are identified, they can be scheduled using various scheduling policies (e.g., round-robin).

We also implemented a dynamic approach, called *top-k spanner*, which identifies, at run-time, the operator tree that is rooted at an output box and that spans the *k* highest priority boxes for a given application (see Section 6.1 to see how we compute box priorities). However, we do not study dynamic approaches in this paper and rely only on static AAAT scheduling.

4.1.2 Superbox Traversal

Once it is determined which superboxes need to be executed, a second-level decision process specifies the ordering of component boxes. This is accomplished by *traversing* the superbox. The goal of superbox traversal is to process all the tuples that are queued within the superbox (i.e., those tuples that reside on the input queues of all boxes that constitute the superbox).

We investigate three traversal algorithms that primarily differ in the performance-related metric they strive to optimize: *throughput*, *latency*, and *memory consumption*.

Min-Cost (MC). The first traversal technique attempts to optimize per-output-tuple processing costs (or *average throughput*) by minimizing the *number of box calls per output tuple*. This is accomplished by traversing the superbox in *post order*, where a box is scheduled for execution only after all the boxes in its sub-tree are scheduled. Notice that a superbox execution based on an MC traversal consumes all tuples (available at the start of execution) while executing each box only once.

Consider the query tree shown in Figure 4 and assume for illustration purposes that a superbox that covers the entire tree is defined. Assume that each box has a processing cost per tuple of *p*, a box call overhead of *o,* and a selectivity equal to one. Furthermore, assume that each box has exactly one non-empty input queue that

[1] Another benefit of superbox scheduling, which we do not address in this paper, is that it improves effective buffer utilization by consuming as many tuples as possible once the tuples are in memory. This potentially reduces the number of times each tuple is swapped between memory and disk.

contains a single tuple. An *MC traversal* of the superbox consists of executing each box only once:

$$b_4 \rightarrow b_5 \rightarrow b_3 \rightarrow b_2 \rightarrow b_6 \rightarrow b_1$$

This traversal consists of six box calls. A simple back-of-the-envelope calculation tells us that the total execution cost of the superbox (i.e., the time it takes to produce all the output tuples) is $15p + 6o$ and the average output tuple latency is $12.5p + 6o$.

Min-Latency (ML). Average latency of the output tuples can be reduced by producing initial output tuples as fast as possible. In order to accomplish this, we define a cost metric for each box b, referred to as the output cost of b, *output_cost*(b). This value is an estimate of the latency incurred in producing one output tuple using the tuples at b's queue and processing them downstream all the way to the corresponding output.

This value can be computed using the following formulas:

$$o\_sel(b) = \prod_{k \in D(b)} sel(k)$$

$$output\_cost(b) = \sum_{k \in D(b)} cost(k) / o\_sel(k)$$

where $D(b)$ is the set of boxes *downstream* from b and including b, and *sel*(b) is the estimated selectivity of b. In Figure 4, $D(b_3)$ is $b_3 \rightarrow b_2 \rightarrow b_1$, and $D(b_1)$ is b_1. The output selectivity of a box b, *o_sel*(b), estimates how many tuples should be processed from b's queue to produce one tuple at the output.

To come up with the traversal order, the boxes are first sorted in increasing order of their output costs. Starting from an empty traversal sequence and box b with the smallest such value, we can then construct the sequence by appending $D(b)$ to the existing sequence.

An *ML traversal* of the superbox of Figure 4 described above is:

$b_1 \rightarrow b_2 \rightarrow b_1 \rightarrow b_6 \rightarrow b_1 \rightarrow b_4 \rightarrow b_2 \rightarrow b_1 \rightarrow b_3 \rightarrow b_2 \rightarrow b_1 \rightarrow b_5 \rightarrow b_3 \rightarrow b_2 \rightarrow b_1$

The ML traversal incurs nine extra box calls over an MC traversal (which only incurs six box calls). In this case, the total execution cost is $15p + 15o$, and the average latency is $7.17p + 7.17o$.

Notice that MC always achieves a lower total execution time than ML. This is an important improvement especially when the system is under CPU stress, as it effectively increases the throughput of the system. ML may achieve lower latency depending on the ratio of box processing costs to box overheads. In this example, ML yields lower latency if $p / o > 0.22$.

Min-Memory (MM). This traversal is used to maximize the consumption of data per unit time. In other words, we schedule boxes in an order that yields the maximum increase in available memory (per unit time).

$$mem\_rr(b) = \frac{tsize(b) \times (1 - selectivity(b))}{cost(b)}$$

The above formula is the expected memory reduction rate for a box b (*tsize*(b) is the size of a tuple that reside on b's input queue). Once the expected memory reduction rates are computed for each box, the traversal order is computed as in the case of ML.

Let's now consider the *MM traversal* of the superbox in Figure 4, this time with the following box selectivities and costs: b_1 = (0.9, 2), b_2 = (0.4, 2), b_3 = (0.5, 1), b_4 = (1.0, 2), b_5 = (0.4, 3), b_6 = (0.6, 1). Assuming that all tuples are of size one, *mem_rr* for all the boxes, b_1 through b_6 respectively, are computed as follows: 0.05, 0.3, 0.5, 0, 0.2, 0.4. Therefore, the MM traversal is:

$b_3 \rightarrow b_6 \rightarrow b_2 \rightarrow b_5 \rightarrow b_3 \rightarrow b_2 \rightarrow b_1 \rightarrow b_4 \rightarrow b_2 \rightarrow b_1$

Note that this traversal might be shorter at run time: for example, if b_5 consumes all of its input tuples and produces none on the output, the execution of b_3 after b_5 will clearly be unnecessary. In this example, the average memory requirements for MM, MC, and ML turn out to be approximately 36, 39, and 40 tuples, respectively (memory requirements are computed after the execution of each box and averaged by the number of box executions).

4.2 Tuple Batching - Train Processing

A *tuple train* (or simply a *train*) is a sequence of tuples executed as a batch within a single box call. The goal of tuple train processing is to reduce overall tuple processing costs. This happens due to several reasons: First, given a fixed number of tuples to process, train processing decreases the total number of box executions required to process those tuples, thereby cutting down low-level overheads such as scheduling overhead (including maintenance of the execution queue and memory management), calls to the box code, and context switch. Second, as in the case of superbox scheduling, train processing also improves memory utilization when the system operates under memory stress (see Section 4.1). A third reason, which we do not directly explore in this paper, is that some operators may optimize their execution better with larger number of tuples available in their queues. For instance, a box can materialize intermediate results and reuse them in the case of windowed operations, or use merge-join instead of nested loops in the case of joins.

The Aurora scheduler implements train processing by telling each box when to execute and how many queued tuples to process. This approach contrasts with traditional blocking operators that wake up and process new input tuples as they arrive. This somewhat complicates the implementation and increases the load of the scheduler, but is necessary for creating and processing trains, which significantly decrease overall execution costs.

Aurora allows an arbitrary number of tuples to be contained within a train. In general, the size of a train can be decided by constraining a specific attribute such as the number of tuples, variance in latencies, total expected processing cost, and total memory footprint. Intelligent train construction is a research topic on its own and is outside the scope of this paper.

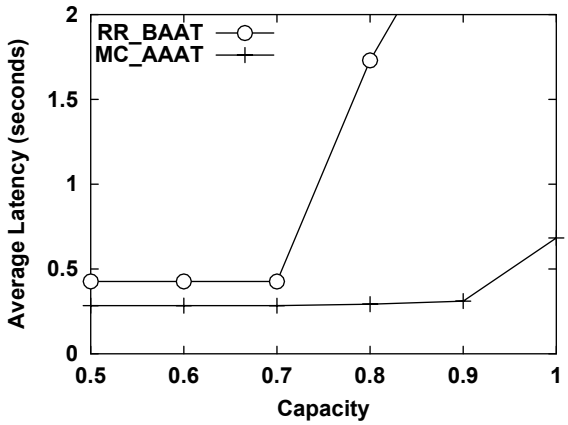

Figure 5: Box vs. application scheduling

5 Experimental Evaluation

5.1 Experimental Testbed

We use the Aurora prototype system to study our operator scheduling techniques. The reference run-time architecture is defined in Section 2.2.

The prototype is implemented on top of Debian GNU/Linux using C++. In the experiments, we used a dedicated Linux workstation with 2 Ghz Pentium IV processors and 512M of RAM. The machine was isolated from the network to avoid external interference.

Due to the fact that the domain of stream-based applications is still emerging and that there are no established benchmarks, we decided to artificially generate data streams and continuous queries to characterize the performance of our algorithms, as described below.

We generated an artificial Aurora network as a collection of continuous queries, each feeding output tuples to individual applications. We modeled a continuous query as a tree of boxes rooted at an output box (i.e., a box whose outputs are fed to one or more applications). We refer to such a query tree as an *application tree*. Each query is then specified by two parameters: *depth* and *fan-out*. Depth of a query specifies the number of levels in the application tree and fan-out specifies the average number of children for each box.

For ease of experimentation, we implemented a generic, *universal* box whose per-tuple processing cost and selectivity can be set. Using this box, we can model a variety of stateless stream-based operators such as filter, map, and union. For purposes of this paper, we chose not to model stateful operators as their behavior is highly-dependent on the semantics they implement, which would introduce another dimension to our performance evaluation and restrict the generality of our conclusions. This would complicate the understanding of the results without making a substantial contribution to the understanding of the relative merits of the algorithms.

An Aurora network consists of a given number of query trees. All queries are then associated with *latency-based* QoS graphs, a piece-wise linear function specified by three points: (1) maximum utility at time zero, (2) the latest latency value where this maximal utility can be achieved, and (3) the deadline latency point after which output tuples provides zero utility.

To meaningfully compare different queries with different shapes and costs, we use an abstract *capacity* parameter that specifies the overall load as an estimated fraction of the *ideal* capacity of the system. For example, a capacity value of .9 implies that 90% of all system cycles are required for processing the input tuples. Once the target capacity value is set, the corresponding input rates (uniformly distributed across all inputs) are determined using an open-loop computation. Because of various system overheads, the CPU will saturate typically much below a capacity of one.

The graphs presented in the rest of the paper provide average figures of six independent runs, each processing 100K input tuples. Unless otherwise stated, the fan-out parameter is set to three; the depth is set to five; the selectivities of the boxes are set to one; and the per-tuple processing costs are selected from the range [0.0001 sec/tuple - 0.001sec/tuple]. Furthermore, unless otherwise stated, we use the round-robin scheduling policy to arbitrate among boxes and superboxes.

5.2 Operator Batching – Superbox Scheduling

We investigate the benefits of superbox scheduling by looking at the performance of the round-robin (RR) algorithm, run in the default box-at-a-time (BAAT), and the MC traversal algorithm applied to superboxes that correspond to entire applications (i.e., application-at-a-time or AAAT, which is described in Section 4.1.1).

Figure 5 shows the average tuple latencies of these approaches as a function of the input rate (as defined relative to the capacity of the system) for five application trees. As the arrival rate increases, the queues eventually saturate and latency increases arbitrarily. The interesting feature of the curves in the figure is the location of the inflection point. RR-BAAT does particularly badly. In these cases, the scheduling overhead of the box-at-a-time approach is very evident. This overhead effectively steals processing capability from the normal network processing, causing saturation at much earlier points. On the other hand, the MC_AAAT algorithm performs quite well in the sense that it is resistant to high load. This technique experiences fewer scheduler calls and, thus, have more processing capacity and is able to *hang on* at input rates of over 90% of the theoretical capacity.

5.3 Superbox Traversal

We first investigate the performance characteristics of the Min-Cost (MC) and Min-Latency (ML) superbox traversal algorithms. In this experiment, we use a single application tree and a capacity of 0.5.

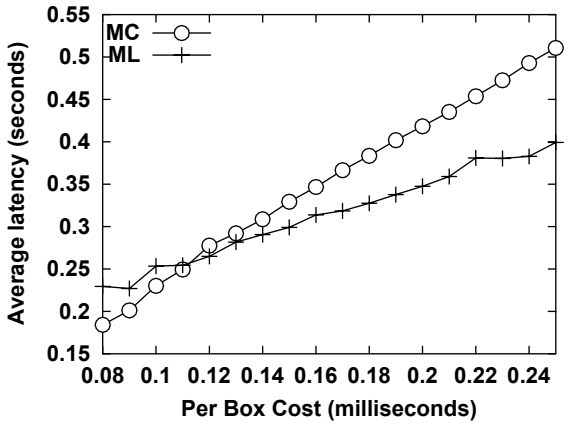

Figure 6: Min-cost vs min-latency traversals

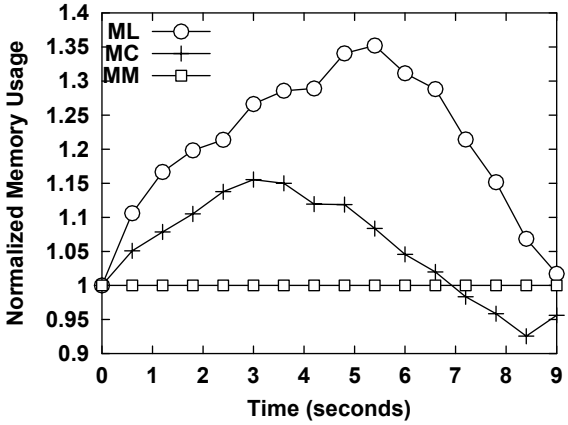

Figure 7: Memory requirements of traversal algorithms

Figure 6 shows the average output tuple latency as a function of per-tuple box processing cost. As expected, both approaches perform worse with increasing processing demands. For most of the cost value range shown, ML not surprisingly performs better than MC as it is designed to optimize for output latency. Interestingly, we also observe that MC performs better than ML for relatively small processing cost values. The reason is due to the relationship between the box processing cost and box call overhead, which is the operational cost of making a box call. The box call overhead is a measure of how much time is spent outside the box versus inside the box (processing tuples and doing real work). As we decrease the box processing costs, box call overheads become non-negligible and, in fact, they start to dominate the overall costs incurred by the algorithms. As we explained in Section 4.1.2, an MC traversal always requires less number of box calls than ML does. We thus see a cross-over effect: for smaller box processing costs, box call overheads dominate overall costs and MC wins. For larger processing costs, ML wins as it optimizes the traversal for minimizing output latency.

A set of complementary results (not shown here due to space limitations) demonstrates that MC incurs less overall box overhead as it minimizes the number of box calls. The difference increases as the applications become deeper and increase in the number of boxes. In fact, the overhead difference between the two traversals is proportional to the depth of the traversed tree.

These key results can be utilized for improving the scheduling and overall system performance. It is possible to statically examine an Aurora network, obtain box-processing costs, and then compare them to the (more or less fixed) box processing overheads. Based on the comparison and using the above results, we can then statically determine which traversal algorithm to use. A similar finer-grained approach can be taken dynamically. Using a simple cost model, it is straightforward to compute which traversal algorithm should do better for a particular superbox.

Figure 7 demonstrates the amount of memory used over the time of superbox run. The curves are normalized with respect to the MM values. ML is most inefficient in its use of memory with MC performing second. MC minimizes the amount of box overhead. As a result MC discards more tuples per unit of time than ML.

MM loses its advantage towards the end since all three traversals are executed on a common query network. Even though each chooses a different execution sequence and incurs different overhead, all of them push the same tuples through the same sequence of boxes. The crossover towards the end of the time period is a consequence of the fact that different traversals take different times to finish. In general, MC has the smallest total execution time–the reason why it catches up with MM at towards the end of the shown execution range.

5.4 Tuple Batching - Train Scheduling

Train scheduling is only relevant in cases in which multiple tuples are waiting at the inputs to boxes. This does not happen when the system is very lightly loaded. In order to see how train scheduling affects performance, we needed to create queues without saturating the system. We achieved this by creating a *bursty* (or clustered) workload that simply gathers tuples in our previously studied workloads and delivers them as a group. In other words, if our original workload delivered n tuples evenly spaced in a given time interval T, the bursty version of this delivers n tuples as a group and then delivers nothing more for the next T time units. Thus, the bursty workload is the same in terms of average number of tuples delivered, but the spacing is different.

The graph in Figure 8 shows how the train scheduling algorithm behaves for several bursty workloads. In this experiment, we have a single application tree with a fan-out of two and a depth of five. In order to isolate the effects of operator scheduling, we use round-robin BAAT for this experiment. The train size (x-axis) is given as a percentage of the queue size. As we move to the right, the trains *bite off* increasingly larger portions of the queues. With a burst size of one, all tuples are evenly spaced.

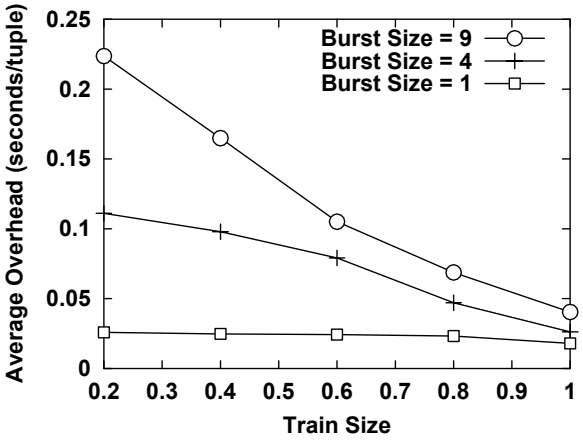

Figure 8: Train scheduling effects

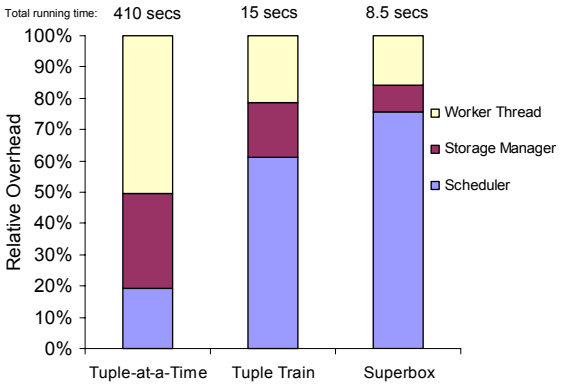

Figure 9: Distribution of execution overheads

This is equivalent to the normal workload. Notice that the curve for this workload is flat. If there are no bursts, train scheduling has no effect. For the other two curves, however, as the burst size increases, the effect gets more pronounced. With small a train size of 0.2, the effect on the overhead (i.e., total execution time less processing time) of increasing the burst size is substantial. For a burst size of 4, we quadruple the average overhead. Now as we increase the train size, we markedly reduce the average overhead for the bursty cases. In fact, when the train size is equal to one (the entire queue), the average overhead approaches the overhead for the non-bursty case. Trains improve the situation because tuples do not wait at the inputs while other tuples are being pushed through the network. It is interesting to note that the bursty loads do not completely converge to the non-bursty case even when the train size is one (i.e., the whole queue). This is because the tuples still need to be processed in order. Since the bursty workload generation delivers *n*-1 of the tuples early, their latency clock is ticking while the tuples in front of them are being processed. In the non-bursty case, the tuples arrive spaced out in time, and a fair amount of processing can be done on queued tuples before more tuples arrive.

5.5 Overhead Distribution

Figure 9 shows a comparison of the relative execution overheads and how they are distributed for TAAT (tuple-at-a-time), BAAT (tuple trains), and MC (superbox), for four application trees. Each bar is divided into three fundamental cost components: the worker thread overhead, the storage management overhead, and the scheduler overhead. The number at the top of each bar shows the actual time for processing 100K tuples in the system.

Looking at the total running times, the first thing to notice is that TAAT is significantly worse than the other two methods, underscoring our conclusion that train and superbox scheduling are important techniques for minimizing various system overheads and improving the overall system throughput. Additionally, this graph shows clearly the benefits of superbox scheduling, which decreased the overall execution time of the system running tuple trains almost by 50%.

Finally, we note the interesting trend in the relative component costs for each approach—while the percentages of the worker thread and storage manager overheads decrease, as we go from the leftmost bar to the right, the percentage of the scheduler overhead increases and starts to dominate the rest. The reason is that, as batching is increased, increasingly more tuples get processed at each scheduling step. In other words, the number of scheduling steps to process a specific number of tuples decreases, but the number of box executions decreases more. Because the worker thread and storage management overheads are primarily associated with the number of box executions, their overheads decrease more relatively to that of the scheduler. Another contributing factor is that, again as we go from left to right, the scheduler algorithms become increasingly more intelligent and sophisticated, taking more time to generate the scheduling plans.

6 QoS-Driven Scheduling

We first discuss how we compute box priorities and, at a coarser level, output priorities using application-specific QoS information and tuple latencies. After describing our basic approach, we propose and experimentally evaluate an approximation technique, based on bucketing and pre-computation, which is used to improve scalability by trading off scheduling overhead with scheduling quality.

6.1 Computing Priorities

The basic approach is to keep track of the latency of tuples that reside at the queues and pick for processing the tuples whose execution will provide the most expected increase in the aggregate QoS delivered to the applications. Taking this approach per tuple is not scalable. We therefore maintain latency information at the granularity of individual boxes and define the *latency* of a box as the averaged latencies of the tuples in its input queue(s).

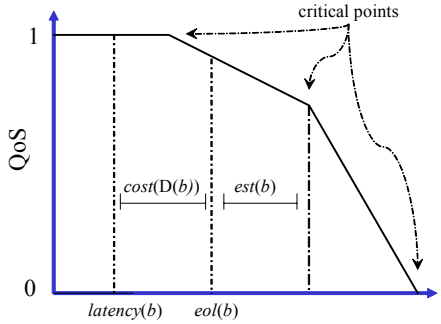

Figure 10: Critical points and expected output delay

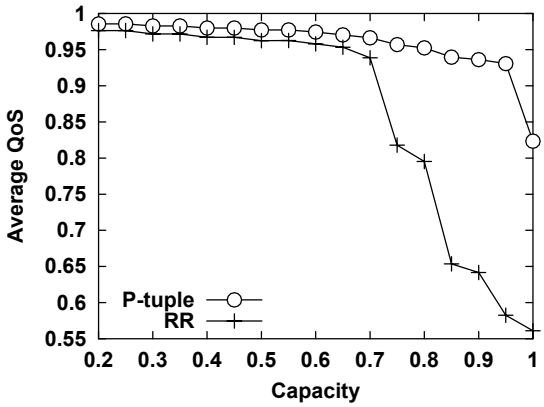

Figure 11: QoS-aware scheduling

Our priority assignment approach is to order the boxes in terms of their *utility* and *urgency*. We define the importance of a box b in terms of its *expected slope value*, *slope*(b), and define its urgency in terms of its *expected slack time*, *slack*(b).

We compute the utility of b as follows:

$$utility(b) = gradient(eol(b))$$

This value is the gradient of the QoS-latency curve for b's corresponding output at the latency value *eol*(b), where *eol*(b) is the *expected output latency* of b. This value is an estimation of where b's tuples currently are on the QoS-latency curve at the corresponding output. In other words, this value provides a lower bound on the expected latency of the corresponding tuples at the output (assuming that the tuples are pushed all the way to the output without further delay). The value *eol*(b) is computed by adding the current latency value to the expected computation time for a given output as follows:

$$eol(b) = latency(b) + cost(D(b))$$

where $D(b)$ is the set of boxes downstream from b and *cost*(D(b)) is an estimate of how long it will take to process the tuples downstream from b. This utility function is a measure of the expected QoS (per unit time) that will be *lost* if the box is *not* chosen for execution.

The *expected slack time*, *est*(b), is an indication of how close a box is to a *critical point*; i.e., a point where the QoS changes sharply. Urgency can be trivially computed by subtracting the expected output latency from the latency value that corresponds to the critical point. If there are multiple critical points, *est*(b) always corresponds to the distance to the closest critical point. These concepts are illustrated in Figure 10, where the QoS is specified as a piece-wise linear function of latency with three critical points.

At each scheduling point in time, we can order the boxes with respect to their *priority tuple*, or *p-tuple*[2]:

$$priority(b) = (utility(b), -est(b))$$

[2] If a box b has multiple downstream applications, *utility*(b) is defined as the sum, and *est*(b) as the minimum value computed across all applications.

In other words, we first choose for execution those boxes that have the highest utility, and then choose from among those that have the same utility, the ones that have the minimum (i.e., least) slack time.

Figure 11 shows a comparison of Aurora's QoS-aware scheduling approach (p-tuple) and a simple round-robin BAAT scheduling policy (RR). In the experiment, there are 20 applications, each with a fan-out of one and a depth of five. Two types of QoS graphs, tight and loose QoS, are modeled (the graphs are specified by the set of points {(0,1), (0.001,1), (1,0)} and {(0,1), (4,1), (5,0)}, respectively) and are randomly assigned to applications.

The graph reveals a significant difference between the average QoS values achieved by the algorithms. The difference is pretty much stable up to a capacity value of 0.7, after which the system becomes overloaded and the performances of both algorithms decrease drastically and will eventually drop to zero (note that they remain above zero due to the finite amount of time experiments were run).

6.2 Approximation for Scalability

A straightforward implementation of the above QoS-driven scheduling approach requires, at each scheduling point, computing the p-tuple for each box and then sorting the boxes with respect to their p-tuples. This is an $O(n \times \log n)$ operation, where n is the number of boxes.

We improve upon the basic algorithm using a combination of (1) *approximation* (via bucketing) and (2) *pre-computation*. Our approach is to partition the utility-urgency space into discrete buckets, and efficiently assign boxes to individual buckets based on their *p-tuple* values at run time. During scheduling, buckets can be traversed in the order of decreasing *p-tuples* (illustrated in Figure 12(a)), and the corresponding boxes are placed in the execution queue. Given a *latency* value, our first goal is to compute the corresponding bucket assignment in $O(1)$. To do this, we make use of two auxiliary graphs, gradient- and slack-latency graphs.

We divide the range of the gradient (i.e., utility) values into g buckets (Figure 12(b) shows an example with four buckets; the cross symbols highlight the latency values

where bucket transitions take place). All gradient values in the same bucket are treated as the same. The width of each bucket, thus, defines a bound on the inaccuracy (or variance) that we are willing to tolerate in terms of the potential deviation from the highest possible gradient value. In other words, the width of a bucket is a measure of the bound on the quantitative deviation from the optimal (with respect to the heuristic) scheduling decision.

Similarly, we divide the slack values into s buckets (Figure 12(c)) and treat all the slack values within a single bucket as equal. Again, the width of a bucket is an indication of the level of approximation we make with regards to the slack values.

Given pre-computed gradient-latency graphs, it is possible to pre-compute the application-specific latency ranges that correspond to each bucket. For example, b_1 will be in $bucket_2$ beyond $latency = 5$ and in $bucket_3$ beyond $latency = 15$; whereas b_3 will be in $bucket_1$ till latency = 12 and in $bucket_4$ afterwards. Slack-latency graphs can be interpreted in a similar fashion as illustrated in the figure: b_1 falls in $bucket_2$ when latency is between 5 and 10, and in $bucket_1$ for other latency values.

When the execution queue is about to become empty, the scheduler wakes up and performs *bucket assignment* by going through the boxes and assigning them into their current buckets. A straightforward implementation of bucket assignment takes $O(n)$ time by going through all the boxes, computing the bucket for each box in $O(1)$. This approach has the potential drawback of redundantly reassigning buckets for each box, even if the box's bucket has not been changed since the last assignment. In particular, we want the bucket assignment overhead to be proportional to the number of boxes that made a transition to another bucket. In order to accomplish this, we use a *calendar queue* [7], which is a multi-list priority queue that exhibits $O(1)$ amortized time complexity for the relevant operations (*insertion*, *deletion*, and *extract-min*) under popular event distributions. As a result, we can implement all phases of bucket assignment in constant amortized time.

6.3 Bucketing Results

We ran the slope-slack (p-tuple) algorithm and our bucketing algorithm on a network with 200 non-overlapping straight-line applications, each with five boxes. The results are shown in Figure 13. The x-axis represents the number of partitions for each of the *QoS-gradient* and the *slack time* ranges. We assume that these two dimensions are partitioned equally. Thus, for example, 10 partitions represent 100 buckets.

The slope-slack method produced a measured QoS of 0.796, which is shown for reference on the graph as a horizontal line. When there is only one bucket, the observed QoS is a very poor 0.427. This is because with one bucket all runnable boxes end up in a single grouping which is then equivalent to round-robin scheduling. Notice, however, that as we increase the number of buckets, the QoS rises sharply; until at 20 partitions we reach a maximum QoS value of 0.85. We manage to exceed the slope-slack value (although only by 7%) because the decrease in scheduler overhead dominates the loss of precision in scheduling decisions introduced by bucketing.

Increasing the number of partitions and thus the number of buckets improves the accuracy of scheduling decisions. Working against this effect, though, is the fact that as the number of buckets grows past some moderate level (approximately 30 partitions), the scheduler overhead begins to increase as can be seen in Figure 14. Simply having a very large number of buckets becomes a bookkeeping problem. Thus, the scheduler overhead will begin to dominate the incremental gain in accuracy which we see in Figure 13 as the QoS curve steadily declines from its maximum and eventually drops below the slope-slack technique at about 260 partitions.

7 Related Work

There has been extensive research on scheduling tasks under real-time performance expectations both in operating systems [14, 16, 17, 20] and database systems [3, 11, 12, 18, 19]. To the best of our knowledge, Aurora's scheduling approach that combines priority assignment and dynamic scheduling plan construction is the first comprehensive proposal for scheduling continuous queries over real-time data streams and QoS expectations. Our solutions no doubt borrow a lot from the myriad of existing work on scheduling. Due to lack of

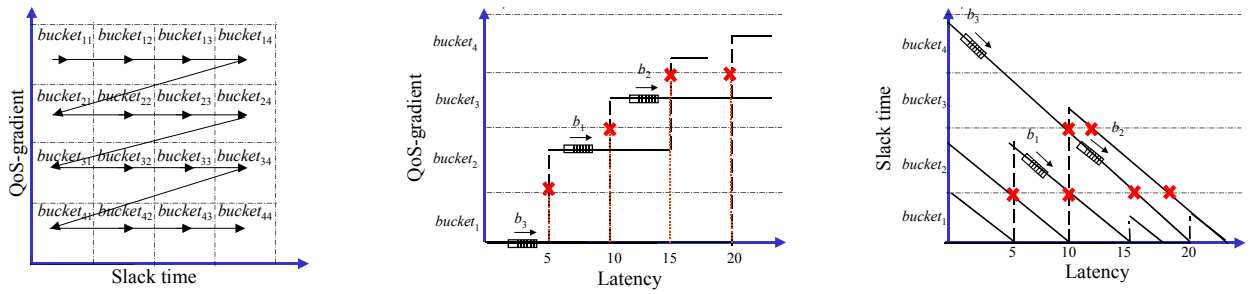

Figure 12: Illustrating (a) bucket traversal, (b) gradient-latency graphs, and (c) slack-latency graphs

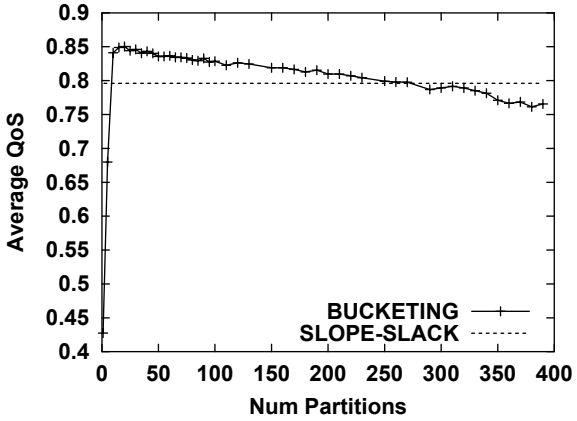

Figure 13: Bucketing effects on QoS

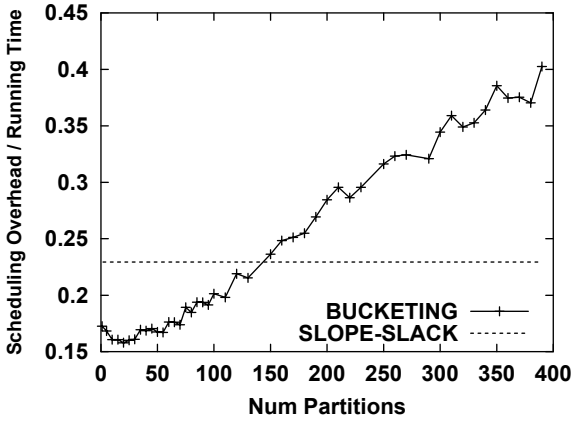

Figure 14: Bucketing overheads

space, however, we only discuss related work that is particularly relevant to our work and highlight the primary differences.

Scheduling proposals for real-time systems commonly considered the issue of priority assignment and consequent task scheduling based on static (table- or priority-driven) approaches or dynamic (planning or best-effort) approaches [19]. Static approaches are inherently ill suited for the potentially unpredictable, aperiodic workloads we assume, as they assume a static set of highly periodic tasks. Dynamic planning approaches perform feasibility analysis at run-time to determine the set of tasks that can meet their deadlines, and rejecting the others that cannot [14]. This decision is based on two key observations: First, our priority assignment algorithm is based on a variation of Earliest-Deadline-First (EDF) algorithm [16], which is well known to have optimal behavior as long as no overloads occur. Second, Aurora employs a *load shedding* mechanism (not described in this paper but can be found in [8, 21]) that is initiated when an overload situation is detected and that selectively sheds load to get rid of excess load in a way that least degrades the QoS. This allows our scheduling algorithm to focus only on underload situations. We note here that Haritsa *et al*. [12] proposed an extension of EDF that is designed to handle overloads through adaptive admission control.

Real-time database systems [3, 11, 12, 15, 18, 19] attempt to satisfy deadlines associates with each incoming transaction, with the goal of minimizing the number of transactions that miss their deadlines. These systems commonly support short-running, independent transactions, whereas Aurora deals with long-running continuous queries over streaming data Leaving aside these differences, of particular relevance to Aurora scheduling is the work of Haritsa *et al*. [11] that studied a model where transactions have non-uniform *values* (or utilities) that drop to zero immediately after their deadlines. They studied different priority assignment algorithms that combine deadline and value information in various ways, one of which is a *bucketing* technique. This technique is similar to ours in that it assigns schedulable processing units into buckets based on their utility. The differences are that (1) we use bucketing to trade off scheduling quality for scheduling overhead and, consequently, for scalability; and (2) we also use bucketing for keeping track of slack values.

Also related to Aurora scheduling is the work on adaptive query processing and scheduling techniques [4, 13, 22]. These techniques address efficient query execution in unpredictable and dynamic environments by revising the query execution plan as the characteristics of incoming data changes. Eddies [4] tuple-at-a-time scheduling provides extreme adaptability but has limited scalability for the types of applications and workloads we address. Urhan's work [22] on rate-based pipeline scheduling prioritizes and schedules the flow of data between pipelined operators so that the result output rate is maximized. This work does not address multiple query plans (i.e., multiple outputs) or deal with and support the notion of QoS issues (and neither does Eddies).

Related work on continuous queries by Viglas and Naughton [23] discusses rate-based query optimization for streaming wide-area information sources in the context of NiagaraCQ [9]. Similar to Aurora, the STREAM project [6] also attempts to provide comprehensive data stream management and processing functionality. The Chain scheduling algorithm [5] attempts to minimize intermediate queue sizes, an issue that we do not directly address in this paper. Neither NiagaraCQ nor STREAM has the notion of QoS.

8 Conclusions

This paper presents an experimental investigation of scheduling algorithms for stream data management systems. It demonstrates that the effect of system overheads (e.g., number of scheduler calls) can have a profound impact on real system performance. We have run our experiments on the Aurora prototype since simulators do not reveal the intricacies of system implementation penalties.

We show that the naïve approach of using a thread per box does not scale. We further show that our approaches of train scheduling and superbox scheduling help a lot to reduce system overheads. We have also discussed exactly how these overheads are affected in a running stream data manager. In particular, these algorithms require tuning parameters like train size and superbox traversal methods.

We also addressed QoS issues and extended our basic algorithms to address application-specific QoS expectations. Furthermore, we described an approximation technique based on bucketing that trades off scheduling quality with scheduling overhead.

The overriding message of this paper is that to build a practical data stream management system, one must ensure that scheduler overhead be small relative to useful work. We have provided some interesting results in this direction by focusing on batching techniques. We intend to extend these studies in the future by examining self-tuning approaches that dynamically revise algorithm parameters based on workload and resource conditions. We are also considering extending our scheduling techniques to distributed environments and other resources (such as bandwidth) in the context of Aurora* [10].

References

[1] D. Abadi, D. Carney, U. Cetintemel, M. Cherniack, C. Convey, C. Erwin, E. Galvez, M. Hatoun, J. Hwang, A. Maskey, A. Rasin, A. Singer, M. Stonebraker, N. Tatbul, Y. Zing, R.Yan, and S. Zdonik. Aurora: A Data Stream Management System (demo description). In *Proceedings of the 2003 ACM SIGMOD Conference on Management of Data*, San Diego, CA, 2003.

[2] D. Abadi, D. Carney, U. Cetintemel, M. Cherniack, C. Convey, S. Lee, M. Stonebraker, N. Tatbul, and S. Zdonik. Aurora: A New Model and Architecture for Data Stream Management. In VLDB Journal, 2003 (to appear).

[3] R. J. Abbott and H. Garcia-Molina. Scheduling real-time transactions: a performance evaluation. *ACM Transactions on Database Systems (TODS)*, 17(3):513-560., 1992.

[4] R. Avnur and J. Hellerstein. Eddies: Continuously Adaptive Query Processing. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, 2000.

[5] B. Babcock, S. Babu, M. Datar, and R. Motwani. Chain: Operator Scheduling for Memory Minimization in Stream Systems. In *Proc. of the International SIGMOD Conference*, San Diego, CA, 2003.

[6] S. Babu and J. Widom. Continuous Queries over Data Streams. *SIGMOD Record*, 30(3):109-120, 2001.

[7] R. Brown. Calendar Queues: A Fast O(1) Priority Queue Implementation of the Simulation Event Set Problem. *Communications of the ACM*, 31(10):1220-1227, 1988.

[8] D. Carney, U. Cetintemel, M. Cherniack, C. Convey, S. Lee, G. Seidman, M. Stonebraker, N. Tatbul, and S. Zdonik. Monitoring Streams: A New Class of Data Management Applications. In *proceedings of the 28th International Conference on Very Large Data Bases (VLDB'02)*, Hong Kong, China, 2002.

[9] J. Chen, D. J. DeWitt, F. Tian, and Y. Wang. NiagaraCQ: A Scalable Continuous Query System for Internet Databases. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, 2000.

[10] M. Cherniack, H. Balakrishnan, M. Balazinska, D. Carney, U. Cetintemel, Y. Xing, and S. Zdonik. Scalable Distributed Stream Processing. In *Proceedings of CIDR'03*, Asilomar, California, 2003.

[11] J. R. Haritsa, M. J. Carey, and M. Livny. Value-Based Scheduling in Real-Time Database Systems. *VLDB Journal: Very Large Data Bases*, 2(2):117-152, 1993.

[12] J. R. Haritsa, M. Livny, and M. J. Carey. Earliest Deadline Scheduling for Real-Time Database Systems. In *IEEE Real-Time Systems Symposium*, 1991.

[13] J. M. Hellerstein, M. J. 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):7-18, 2000.

[14] M. B. Jones, D. Rosu, and M.-C. Rosu. CPU Reservations and Time Constraints: Efficient, Predictable Scheduling of Independent Activities. In *Symposium on Operating Systems Principles*, 1997.

[15] B. Kao and H. Garcia-Molina, "An Overview of Realtime Database Systems," in *Real Time Computing*, A. D. Stoyenko, Ed.: Springer-Verlag, 1994.

[16] C. D. Locke. Best-Effort Decision Making for Real-time Scheduling, CMU TR-88-33, 1988.

[17] J. Nieh and M. S. Lam. The Design, Implementation and Evaluation of SMART: A Scheduler for Multimedia Applications. In *Proc. 16th ACM Symposium on OS Principles*, 1997.

[18] G. Ozsoyoglu and R. T. Snodgrass. Temporal and Real-Time Databases: A Survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 7(4):513-532, 1995.

[19] K. Ramamritham. Real-Time Databases. *Distributed and Parallel Databases*, 1(2):199-226, 1993.

[20] K. Ramamritham and J. Stankovic. Scheduling algorithms and operating systems support for real-time systems. *Proceedings of the IEEE*, 82(1):55-67, 1994.

[21] N. Tatbul, U. Çetintemel, S. Zdonik, M. Cherniack, and M. Stonebraker. Load Shedding in a Data Stream Manager. In *Proceedings of VLDB*, Berlin, Germany, 2003.

[22] T. Urhan and M. J. Franklin. Dynamic Pipeline Scheduling for Improving Interactive Query Performance. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*, Rome, Italy, 2001.

[23] S. Viglas and J. F. Naughton. Rate-Based Query Optimization for Streaming Information Sources. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Madison, Wisconsin, 2002.