RC24667 (W0810-081) October 14, 2008
Computer Science

IBM Research Report

Characterizing, Constructing and Managing Resource Usage Profiles of System S Applications: Challenges and Experience

Kirsten W. Hildrum, Deepak Rajan, Sujay Parekh, Joel L. Wolf, Kun-Lung Wu
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Characterizing, Constructing and Managing Resource Usage Profiles of System S Applications: Challenges and Experience

Kirsten W. Hildrum
hildrum@us.ibm.com

Deepak Rajan
drajan@us.ibm.com

Sujay Parekh
sujay@us.ibm.com

Joel L. Wolf
jlwolf@us.ibm.com

Kun-Lung Wu
klwu@us.ibm.com

IBM T.J. Watson Research Center, Hawthorne, NY 10532

ABSTRACT

We describe the challenges of and our experience in characterizing, constructing and managing the usage profiles of System S applications. A System S application deployed at runtime is a directed graph with software processing elements (PEs) as vertices and data streams as edges connecting the PEs. The resource usage of each PE is a critical input to the runtime scheduler for proper resource allocation. We represent the resource usage of PEs in terms of resource functions (RFs) that are used by the System S scheduler, with one RF per resource per PE. The first challenge is that building good RFs that can accurately predict the resource usage of a PE can be difficult because the PEs perform arbitrary computations. A second set of challenges arises in managing the RFs and data so that we can apply them for PEs that are re-run and/or reused by the same or different applications or users. We report our experience in overcoming these challenges. Specifically, we present an empirical characterization of PE RFs from several real streaming applications running in a System S testbed. This justifies our simple, yet effective, models of resource usage that build on the data-flow nature of the underlying application. We show that simple piecewise linear models are generally effective in practice, even for complex PEs. To illustrate our methodology, we evaluate and analyze the performance of several real System S applications as a function of the quality of our resource profile models. To obtain these resource profiles, the system automatically learns the models from the raw metrics data collected from running PEs. We describe our approach to managing the metrics and RF models, which allows us to construct generalizable RFs and eliminates or reduces the learning time for new PEs by intelligently storing and reusing the metrics data.

1. INTRODUCTION

System S [2, 20, 8, 19], under development at the IBM T. J. Watson Research Center, and many other distributed stream processing systems [22, 4, 11, 1] are in a class of scalable distributed systems which are geared toward processing long-running queries on continuous streams of data. Streaming applications in such systems are typically organized as data-flow graphs. The runtime deployable unit in System S is a PE. Each application is essentially a directed graph with the software processing elements (the PEs) as verticies and the data streams as directed edges connecting the PEs.

A key resource allocation problem faced by the runtime scheduler in such systems is to map the PEs in the applications to compute resources in a way that utilizes the available CPU, network and memory resources efficiently without overloading any individual node or network link. In essence this is a complex bin-packing problem. A critical input to the runtime scheduler for solving such a problem is the "size" of a PE, or, equivalently, the resource usage of a PE. The resource demand r_p of a PE p for resource r is given by the functional form $r_p = f_{p,r}(d_1, d_2, ...)$ where $d_1, d_2, ...$ are the factors on which the resource usage depends. The function $f_{p,r}$ is called a resource function (RF) and is a model of the PE's resource usage.

There are several challenges that arise in the context of RFs. First and foremost, accurately predicting the resource usage of PEs can be difficult. A PE can be the result of fusing multiple unrelated SPADE operators by the optimizer of the SPADE compiler [8].[1] A PE can also implement a user-defined, arbitrarily complex data analytic algorithm. As a result, the PE sizes are not fixed or even known a-priori. Furthermore, the usage of one resource of a PE can depend on characteristics of the input streams (such as volume/rate and data content) which can change dynamically, potentially causing the resource usage of a PE to change.

A second set of challenges arises in the context of managing the RFs, driven by how the PEs are used. A PE from

[1] SPADE is the development front-end for System S. It is a language for composing a streaming application, which is typically an operator-based data-flow graph. It is also a compiler. After compilation, multiple operators can be fused into a single PE and the operator-based flow graph is coalesced into a PE-based flow graph for deployment [8].

a job may be resubmitted at a later time, either as part of the same job or a different job. Users may share PEs (eg, a classifier) in their own applications. The same PE may run in the system under a different set of circumstances, such as different parameterizations or context (ie, with different upstream or downstream PEs). Since even PEs that have never been run before do need to be scheduled, it is very helpful for good resource allocation to have *some* initial estimate of the PE's resource usage. Hence, one question is how do we identify a PE so we can associate its RF with it? In a similar vein, how should RFs be shared between instances of a PE, and how can observing one instance of a PE yield clues about another, slightly different PE instance's resource usage? For PEs that may not have been run before – what sensible initial RF can be provided?

In this paper, we describe our experience in addressing these challenges. Specifically, we present our practical approach to characterizing, constructing and managing the resource usage profiles of stream PEs for the purpose of providing a critical input to the SODA scheduler [19, 18] in System S. (SODA is the runtime scheduler for System S and it stands for Scheduling Optimizer for Distributed Applications.) Note that we do not claim that our approach to addressing these challenges is the best one. Other alternatives certainly exist, and we continue to explore some of them. Moreover, we focus only on learning the resource usage of CPU and network, and rely on matching resource constraints associated with memory usage specified by the application developer.

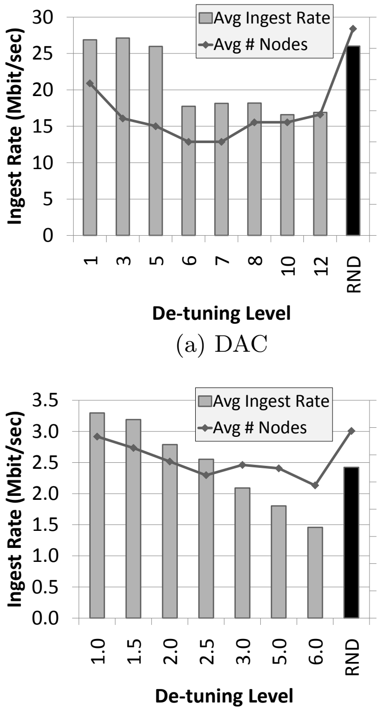
To highlight the impact of resource model inaccuracy on system performance, In Figure 1, we compare the total ingest rate achieved for two System S applications: DAC [20] and SKA [6] when SODA has increasingly incongruous resource models. The higher the total ingest rate, the better the system performance. The applications are described in Section 2.2 and the experiments are detailed in Section 3.5.1. Here, an incongruity level of 1 represents the best application performance when it is scheduled with SODA using the actual trained RFs, i.e., the most congruous/accurate RFs. (Note that, from now on, we will use RF, or Resource Function, to describe the resource usage model of a PE.) The other cases represent increasing levels of "incongruity" of the RFs, namely the RFs are increasingly out of place. We see that the RFs need not be completely accurate to achieve good performance in practice (see the incongruity levels of 3 and 5 for DAC and the incongruity level of 1.5 for SKA). However, they cannot be too far out of place, either. For example, with severely incongruous RFs, the application performance can reduce significantly, e.g., by over 30% for DAC and over 50% for SKA. In extreme cases, the application may even fail to start or run.

Specifically, based on our experience, we make the following contributions in this paper:

- We demonstrate a simple data flow-based approach to modeling the resource usage of PEs. We show that simple piecewise linear models are generally effective in practice, even for complex PEs. We validate this approach empirically against PEs from several System S applications, including both simple and complex PEs.

- We show a practical scheme for managing and building these models in a way that maximizes the usage of raw input (training) data and enables the learned resource models to be generalized to new PEs.

- We validate that the RFs need not be very accurate. Schedules generated by SODA degrade, but gracefully, as the incongruity in RFs grows. The sensitivity to RF incongruity appears to depend on the spare capacity in the system as a whole.



(a) DAC

(b) SKA

Figure 1: Effect of scheduling with bad resource profiles

The rest of the paper is organized as follows. In Section 2 we introduce System S and describe the testbed and applications used in the experiments in this paper. Next, we present in Section 3 the resource model for CPU and network that we use, along with an evaluation of the resource model sensitivity. The issue of model management and a description of our approach are given in Section 4. Related work is reviewed in Section 5, and we conclude in Section 6.

2. BACKGROUND

2.1 System S

System S [2, 20, 8] is a large-scale distributed stream processing middleware being developed at IBM Watson Research. It is designed for supporting complex analytics on large volumes of streaming data, both structured and unstructured. System S has two main components: SPADE and the System S runtime. SPADE is a rapid application development front-end for System S [8]. It consists of a language, a compiler, and auxiliary support for building distributed stream processing applications. The SPADE language provides a stream-centric, operator-level programming model. The operator logic can optionally be implemented in a lower-level language, like C++, whereas the

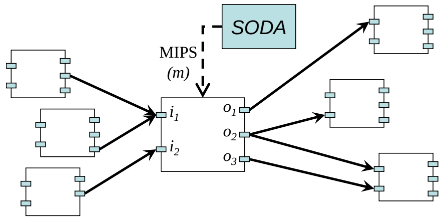


Figure 2: Illustration of a PE showing input ports i_k, output ports o_k and streams.

SPADE language is used to compose these operators into logical data-flow graphs.

However, a SPADE operator implementing a simple logic, like filtering, can be too “small” to be efficiently deployed to a compute node at runtime. The SPADE compiler can fuse several operators into a *processing element* (PE), which is the unit of deployment in System S. As a result, the SPADE compiler is able to coalesce logical data-flow graphs (operator-based) into physical graphs (PE-based) that are more appropriate for deployment.

At runtime, the processing of stream applications is organized in terms of one or more *jobs* that consist of PEs organized into data-flow graphs. PEs can perform arbitrary processing on their input data. A general PE is depicted in Figure 2. PEs consume and produce *streams* which carry data in terms of strictly-typed *stream data objects* (SDOs). A PE receives and sends data through *ports*, which represent attachment points for streams. A PE can read from multiple ports, write to multiple ports, and multiple streams may originate or end in a single port.

2.1.1 Processing Elements

In System S, PEs are a runtime deployable unit. Physically, a PE is a process, and may contain of one or more threads of execution. The PEs of an application are distributed across the nodes of the System S cluster. Each node in the cluster can run multiple PEs and divides its CPU resource between them according to fractions dictated by the scheduler SODA (Section 2.1.2).

The PEs can be generic programs, and are not limited to standard streaming relational operators. As a specific example, a PE may carry out several operators [8] and thus its behavior is a conglomeration of the behavior of these operators and the specifics of how they are connected inside the PE.

2.1.2 SODA

Resource allocation in System S is performed by a centralized, epoch-based scheduler called SODA [19], which is a subcomponent of the System S runtime. In streaming systems such as System S, the jobs are usually long-running, and often continue to run until they are terminated by the user who submits them. Therefore, metrics such as completion time and response time, which are traditionally optimized in batch processing systems are not relevant. Instead, SODA maximizes a utility-theoretic measure known in System S as *importance* subject to a variety of real-world constraints. In doing so, it analyzes a vast number of PE resource allocation alternatives for different job admission and template choices, using importance as a black box objective function.

The importance metric is a weighted sum of selected *value functions* at key (typically terminal) streams in the data-flow graphs. The notion is that these streams represent the “final” products of the various jobs, and the weighted value functions translate these stream rates into measures of goodness of the work done in System S.

The SODA scheduler performs three major functions.

- *Job admission:* It chooses a subset of jobs to execute from a potentially huge collection of jobs submitted to the system.

- *Template selection:* For those jobs that will be admitted it chooses one of potentially several alternate approaches, known as *templates*. Each such template represents a distinct method of performing the job functionality, and is represented by a different data-flow graph. Templates are intended to express alternate approaches which are computationally equivalent but may trade off achieved benefit with required resources for performing the job.

- *Node assignment and fractional allocations:* SODA also chooses flow-balanced resource allocations for all the PEs in the chosen templates of the admitted jobs. By flow balance we mean that PEs are given the *right* level of resource allocation relative to their predecessors. To understand this, consider one particular PE. Giving relatively too many resources to the PE’s predecessors will “flood” the PE, while giving relatively too few resources will starve the PE. The allocation levels need to be balanced throughout the entire data-flow graph. Then, it assigns a processing node to each of these PEs. Finally, combining these decisions, SODA chooses *flow-balanced* fractional allocations of PEs to processing nodes. The fractional allocation prescribes how a node’s CPU resources are divided among the PEs assigned to that node.

MacroQ, one of the mathematical components of SODA, is responsible for the first two of these functions and a portion of the third: specifically, MacroQ performs job admission and chooses job templates. It also computes flow-balanced target resource allocations of PEs. Then, the MacroW component of SODA finds the best processing nodes for each PE, simultaneously load balancing the nodes and minimizing network traffic. Finally, the MicroW component computes fractional allocations of the PEs to those processing nodes to meet the target resource allocations as closely as possible. In this paper, we focus on the MacroQ component of SODA since it uses the RFs as input.

For MacroQ to do its job, the RFs must provide estimates of the CPU requirements of the PEs and the network traffic between PEs. A static estimate of resource usage would perform poorly, because these requirements could (and do) vary according to several factors. For example, a classification PE will likely need resources that depend upon its input rate, and will have an output rate that also depends on its input rate. Resource usage could also depend on properties of the input to the PE – if the classifier above is classifying speech based on audio clips, the length of the audio clips could effect the resources used.

Without going into further details of the MacroQ algorithm (see [17]), there are some requirements imposed on the RFs:

- Scope: A PE should have a common RF across the nodes in the cluster, even for heterogeneous clusters. This is for tractability. Otherwise the MacroQ search space becomes too large.

- Accuracy: since the required accuracy depends on the resource granularity chosen by SODA, we may, in practice, get away with less than perfect RFs as long as they are not dramatically off.

- Form: The RFs should be monotonic increasing in their input parameters. This is followed by a section which is flat, and the function does not increase anymore after that. For most PEs, given more compute resources and/or a higher input rate the output rate does not decrease. This particular requirement is an implicit assumption in the scheduler algorithm, and was discovered during troubleshooting, as discussed in Section 3.6.2.

2.2 Streaming Applications

For our tests, we study PEs from four applications running on System S; these represent different but typical uses of streaming systems.

- **DAC** [20] represents an insurance claims fraud detection and alerting system involving some heavy streaming analytics, i.e., CPU-intensive complex stream mining algorithms. Consisting of six jobs and 51 PEs. DAC provides some scheduling challenges because its PEs have a wide range of processing requirements.

- **SKA** [6] is a radio astronomy application which reconstructs images from data received by radio telescope antennas using interferometry [7]. SKA involves performing processor and memory intensive computations on large amounts of streaming data.

- **Fab** [12] is an application that processes streaming data from automated tests in a chip manufacturing plant, with a goal to monitor and alter the process to improve yields.

- **VWAP** [3] represents a financial markets scenario where real-time quotes are processed to detect bargains and trading opportunities.

2.3 Testbed

The experiments and data discussed in this paper are collected using a System S deployment on a cluster consisting of IBM BladeCenters running Linux 2.6.9. We run our applications on 14 blades with dual-CPU, dual-core 3GHz Intel Xeon processors with 8GB RAM. The blades are in the same rack; they communicate over 1GB/s links and are inter-connected using a high-speed 20GB/s backplane.

For the specific set of applications and system software used in our experiments, the network, backplane or network interface card (NIC) is almost never a bottleneck. Thus, the only disadvantage of placing two PEs that communicate with each other on separate blades is the additional processing overhead involved in sending data to a different node. Although SODA allocates PEs to the nodes while trying to minimize the traffic across blades, this feature is not critical for the specific combination of infrastructure and applications described here. Workloads that will stress this aspect of the system are currently being developed. Nevertheless, accurate RFs are a key factor in SODAs ability to balance the load on the processing nodes.

The system is operating in reliable transport mode (no packets dropped between PEs in the system). We collect the system metrics such as CPU and network traffic rates in terms of averages for 1 minute intervals. The applications are configured to run for 30 minutes with throughput-oriented workload generators which push as much throughput as is possible. Each run (for a particular setting) is repeated 4 times, and averages collected.

3. PE RESOURCE USAGE MODELING

A PE's RF represents a resource usage model for that PE. In this section we present a model for streaming PEs, and validate it against actual PEs from the test applications described above. Our RF models focus on the usage of CPU and network by the PE. For memory, we currently assume that application developers specify maximum memory requirements of PEs, if necessary. The allocations returned by SODA are such that all PEs assigned to a node will fit within the available memory on that node.

3.1 Model Parameters

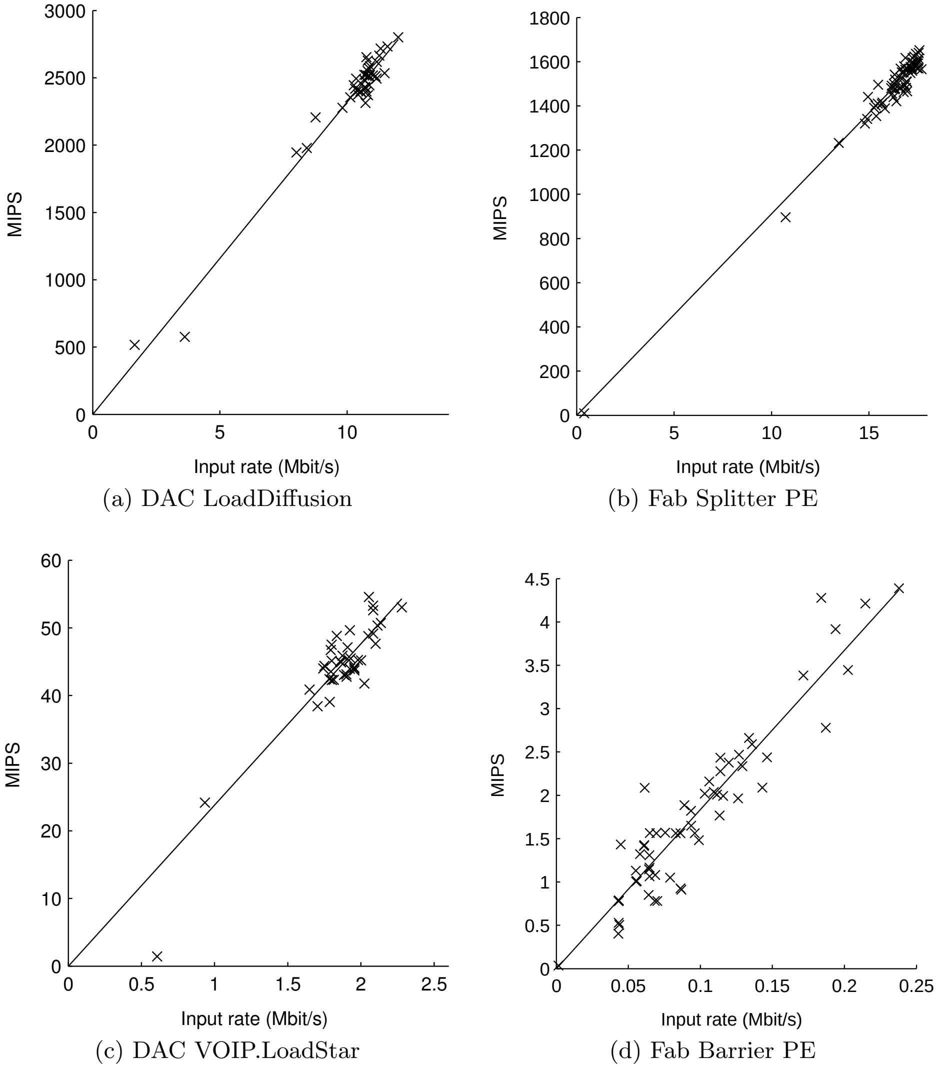
In order to predict the resource usage of a PE, we must consider two broad categories of factors: *dynamic* and *static*, depending on whether they vary at runtime. Dynamic factors include the input data rates to the PE, the distribution of input data types, and data content. Static factors include the PE code, the PE arguments, stream flowspecs and system configuration such as the communication model. One factor which could be regarded as dynamic but is treated as static for modeling purposes is the nature of physical resources given to the PE, for example, whether it is running on generic x86 hardware or special-purpose processors. Also, the data content is difficult to characterize/represent in the general case, so in the interest of simplicity we do not consider it. PEs whose behavior is adaptive to available resources may further depend on the time history of system load and traffic.

In this section, we focus on modeling CPU and output rates as a function of input rates for runtime PE instantiations. The use of the static PE attributes are discussed later, in Section 4.

We also remark that RFs are further classified as *source*, *sink*, or *transform* RFs, depending on whether the PE is a source PE (feeding from primal streams only), a sink PE (not writing any streams, except maybe to disk), or a transform PE (everything else), respectively. Our discussion in this section assumes that the PE is a transform PE (with both input and output streams), but the methodology and techniques naturally carry over to the other types. The notion of PE types will re-appear when we address RF model management in Section 4.

3.2 Modeling CPU Usage

A challenge in modeling the CPU needs of a program is that this demand will vary depending on the specific CPU being used. To enable us to construct a general model that can be used across all nodes in a heterogeneous cluster, we use MIPS as a measure of CPU demand. The MIPS used here is the processor BogoMips [16] reported by the Linux kernel. The MIPS consumed by a PE is calculated by multiplying



(a) DAC LoadDiffusion

(b) Fab Splitter PE

(c) DAC VOIP.LoadStar

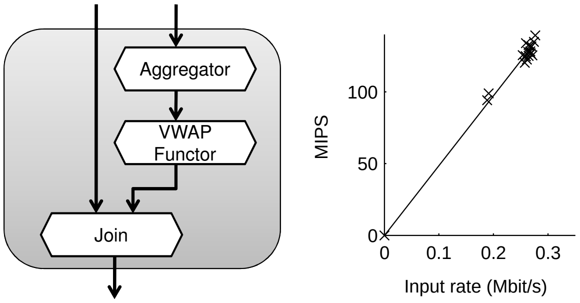
(d) Fab Barrier PE

Figure 3: Examples of linear MIPS profile

(a) the CPU time fraction used by a PE on a specific node (reported by the OS) and (b) the total BogoMips of the cores on a processor. Inspite of the known limitations [16] of the BogoMips measure, this approach is still useful as it allows us to generalize across CPU speed variability within the same processor family.

For processing resources, a natural model is to consider a processing cost per incoming data object. In a queueing theory sense, this is the "service time" per request. Even though there is not necessarily a well-defined "service time" for each incoming data object in streaming systems, in the aggregate, one may expect the CPU requirements of a PE to scale proportionally with input rate. For simplicity we consider a linear scaling, and further it is based on the *total* data rate into the PE ; we do not distinguish the rates on individual ports.

To validate this model, we study the behavior of PEs in our test applications. Some examples are shown in Figure 3. For a majority of PEs, the MIPS-per-datum model seems



(a) VWAP PE Internals

(b) VWAP PE RF

Figure 4: Internal structure and RF of a fused PE from VWAP

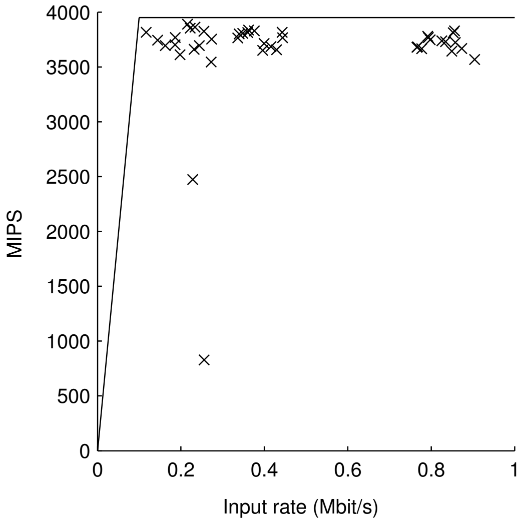


Figure 5: Example of maximum MIPS profile (DAC LoadDiffusion.JoinOperator)

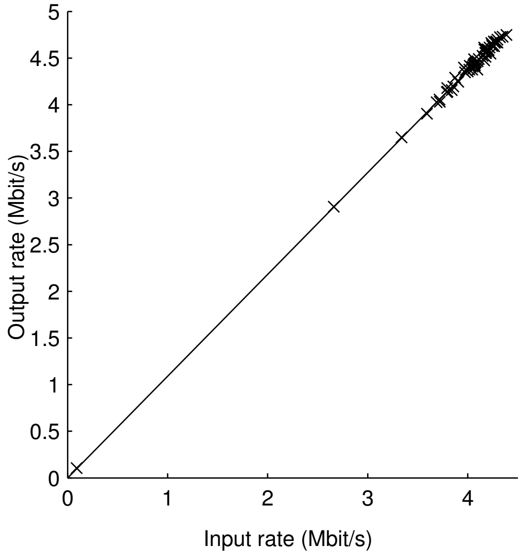


Figure 6: Example of linear output rate profile (Fab PLLTest PE)

to hold true. This seems to be the case for both heavier, computationally more demanding, PEs (Figure 3(a), Figure 3(b)) as well as lighter PEs that use few MIPS per datum (Figure 3(c), Figure 3(d)). In the case of the SPADE-based applications, this model applies even for those PEs which are composed (fused) from simpler SPADE operators. Another surprise is that even some of the join PEs show this linear behavior. As an example, the PE shown in Figure 4(a) is composed of an *aggregator*, a *functor* and a *join*. Yet, the MIPS profile follows the same linear MIPS-per-tuple pattern as seen in 4(b).

Based on our experience, we present our initial candidate model for CPU.

$$m_p = \min(M_p, a_p \sum_{i \in \text{IPorts}(p)} r_i^I) \qquad (1)$$

where M_p represents the maximum MIPS that can be allocated to the PE p. This limit often occurs due to system considerations: the maximum MIPS on any processing node in the system, for instance. Even when this is not the case, this limit allows us to place bounds on the search space explored by SODA. In this model, the term a_p represents the best-fit slope of the MIPS needed by the PE as a function of its input rate.

To illustrate, consider the DAC Load Diffusion PE illustrated in Figure 3(a). For this PE, the coefficients M_p and a_p in (1) can be derived from fitting the best linear model to the data, yielding $M_p = 3000$, and $a_p = 200$.

Some PEs, on the other hand, notably from the DAC application, consume a lot of CPU even when the input rates are not very large, effectively saturating the processing node. One example is shown in Figure 5. The PEs shown are single-threaded, and 4000 MIPS represents the BogoMips capacity of one core of the node. However, even such PEs can be modeled using the linear model described in (1). By recognizing that the resource usage of such PEs is largely determined by the maximum MIPS available, we can set a_p sufficiently large (in this case larger than 50000), and set $M_p = 4000$ (from the observed data). Now the first term in (1) dominates the second, resulting in a MIPS prediction m_p of 4000 for any sufficiently large non-zero data input rate.

3.3 Modeling Output Rates

Output rates for most streaming PEs are relatively easy to model, since in many cases they have a 1:1 relationship to the inputs, meaning that for every incoming tuple, an outgoing tuple is produced. Operations that do not fit this pattern include filtering, aggregation and timer-based data output. We use the following model for output rate:

$$r_p^O = \min(R_p, b_p \sum_{i \in \text{IPorts}(p)} r_i^I) \qquad (2)$$

Thus, for PEs that have a 1:1 relationship to the inputs, b_p is equal to 1, and R_p is set to the largest data rates seen at the output port. Consider the PLLTest PE from Fab, illustrated in Figure 6. From the data, we observe that $R_p = 5$. Filtering PEs will have $b_p < 1$. Tuple count-based operations (that produce an output for every k input tuples) will have the slope $b_p = 1/k$. PEs that produce output based on timers are the main exception requiring a different model since they are independent of the input rate. For such PEs, the model takes the form $r_p^O = 1/T$ where output is every T time units.

3.4 Building RFs

A parametric RF can be built for each output port of a PE. Such an RF first describes the type of model considered, followed by all the relevant parameters.

This study also reveals a methodological challenge: the stable flow balance during this workload generator driven data collection can result in very narrow input rates observed at the PEs. This can result in skewed data for the resource function learning. One way to address this is to explicitly "tweak" the input rates to each PE to ensure coverage over some reasonable range, so that the function fitting will be valid and generalizable outside the range observed during the calibration step.

3.5 Model Evaluation

To evaluate the RFs, we run the applications as described in Section 2.3, and look at the following application and infrastructure level metrics:

- *Ingest rate*: This is a measure of how much data (in Mbps) could be processed by the system. It is intended as a measure of the system's "effective capacity" and should be correlated with importance. In stream processing systems such as System S, flow-balanced resource allocations for the PEs and a load-balanced allocation of PEs to processing nodes minimizes bottlenecks, thus maximizing the amount of data processed at the source PEs (ingest rate).
- *Stream affinity*: One way to measure the quality of the placement is in terms of the traffic load on the system. We compute the amount of traffic that is sent between PEs on the same node divided by the total traffic. The higher this quantity, the better, since PEs which share a stream should be put on the same node (or nearby) to minimize network utilization.
- *Maximum node utilization*: This is a measure of how well SODA distributes the processing load across various machines. This metric is especially interesting when evaluated in conjunction with stream affinity; SODA attempts to maximize stream affinity while simultaneously minimizing maximum node utilization.

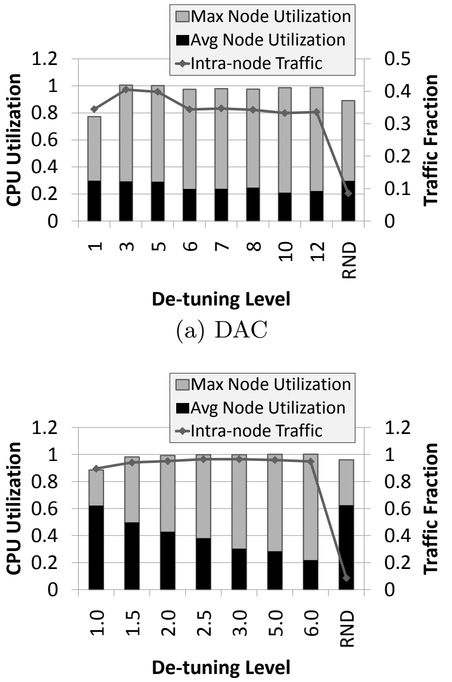
The first metric (ingest rate) is the most tangible measure of system performance for streaming systems. The latter two metrics, in conjunction, illustrate the quality of the placement of PEs to processing nodes. Given the same stream affinity, smaller the maximum node utilization the better. Given the same maximum node utilization, higher the stream affinity the better. These metrics are computed from the raw system metrics such as CPU usage per PE and traffic consumed and produced by each PE.

3.5.1 *Making* RF*s Less Congruous*

In these experiments, we modify the RFs learned from an expert placement in an attempt to systematically degrade their quality, and analyze the resulting deterioration in the quality of SODA placement. The expert placement corresponds to an allocation of PEs to processing nodes by someone with significant knowledge about the application, and is determined by a trial-and-error process until the domain expert is satisfied with the performance of the application. Note that the RFs are not tuned by hand; their parameters are learnt automatically by the system using the models described in Sections 3.2 and 3.3.

To analyze the deterioration of RFs carefully, we parametrize the de-tuning of the RFs using a parameter κ. (When $\kappa = 1$, the RFs are not modified.) Since our modifications involve generating random numbers, to ensure that these modifications are consistent across runs and values of κ, we pregenerate a sequence of random numbers, and use the same sequence for all the runs. This is equivalent to seeding our random number generator with the same value each run.

We now describe precisely how each term in an RF is changed, given κ. Fix a particular value of κ. We modify each term t in an RF by a factor α, as follows. We draw two random numbers, say r_1 and r_2. The first number r_1 determines how much the term t gets modified, and the second number determines whether t is increased or decreased. Let $round(a, b)$ denote the value of a rounded to the nearest multiple of b. We set $\alpha = 1 + round(r_1 * (\kappa - 1), 0.1)$; Thus, α is some multiple of 0.1 between the values 1 and κ, determined by the value of the random number r_1. If $r_2 < 0.5$, we multiply t by α. Otherwise we divide t by α. Observe that when $\kappa = 1$, $\alpha = 1$, and t is not modified. Furthermore, as κ is increased, the amount of de-tuning increases probabilistically. This allows us to modify each term in the RFs in a controllable fashion, analyzing the performance of SODA as the RFs degrade. Since we use the same sequence of random numbers, each term is modified using the same r_1, r_2 in all the runs. Different κ values will result in different values of α that are strongly correlated with κ, as we desired. As we increase κ to ∞, α also increases to ∞ (for positive r_1), and the RF term t either increases to ∞ or decreases to 0, depending on whether $r_2 < 0.5$.



(a) DAC

(b) SKA

Figure 7: Effect of scheduling with bad resource profiles on node utilization and intra-node traffic

We illustrate using the RF modeling the CPU usage of the DAC Load Diffusion PE (see Figure 3(a)). The RF for this PE has two terms $M_p = 3000$ and $a_p = 200$. Let us consider the first term. Suppose the two random numbers r_1 and r_2 are 0.84 and 0.39. When $\kappa = 3$, $\alpha = 2.7$, and M_p is modified to the new value 8100. On the other hand, when $\kappa = 5$, $\alpha = 4.4$, and M_p is modified to a new value of 13200. If $\kappa = 1$, the value of M_p is unchanged. Observe that larger values of r_1 result in a larger perturbation for a given κ, and also that if $r_2 > 0.5$, the value of M_p would have decreased instead.

3.5.2 *Impact on Application Performance*

The effect of making the RFs less congruous is to make

SODA's estimates of the resource requirements of the various PEs more inaccurate. In the experiments we present here, we do not consider job admission or template selection decisions. Therefore, the MacroQ component of SODA determines the CPU requirements of the PEs (and traffic between PEs) in the submitted job in the process of maximizing the net importance of the system. These numbers are then used as input by the MacroW component to allocate the PEs to nodes in an intelligent manner (load-balancing the nodes and minimizing inter-node traffic simultaneously). Less congruous RFs results in worse CPU estimates in MacroQ and less intelligent placement by MacroW.

These effects are shown in Figure 1 and Figure 7 for two applications: DAC and SKA. We also show the performance of the RND scheduler that assigns PEs to nodes in the cluster randomly. Given our cluster size, the Fab and VWAP applications did not present enough work, so we do not use them in the evaluation here. We see that for SKA there is a systematic drop in performance (ingest rate) as the RFs are made more incongruous, whereas for DAC, the performance is relatively insensitive to RF perturbation. In SKA, we see that RND performance is considerably worse than SODA with congruous RFs. On the other hand, the performance of the RND placement on DAC is close to that of SODA with congruous RFs (see Figure 1), even though RND uses more nodes than SODA with congruous RFs, thus sending a larger traffic load on the network (smaller stream affinity).

This behavior of the RF sensitivity as well as effectiveness of RND placement is explained by considering the overall node utilizations of the various scenarios (see Figure 7). We see that in general, DAC has low average utilization, implying it is very over-provisioned. In this case, the effect of a poor placement (in terms of network traffic load and node utilization) on ingest rate is not going to be significant until the placement gets dramatically tweaked. For SKA, on the other hand, some PEs are quite computationally intensive. Here, the nodes are well utilized in a placement computed with congruous RFs, and the effect of making the RFs less congruous is to create bottlenecks. This increases the maximum node utilization while decreasing average utilization. For SKA, the RND case also seems to utilize the nodes well, but because it does not account for intra-node traffic, its performance is much lower than SODA with congruous RFs.

Thus, we see that in the under-utilized case, the performance can be less sensitive to RF incongruity than under higher node utilizations. On the other hand, with larger workloads (higher node utilizations as in SKA), and with mixed application workloads, the performance of the scheduler SODA deteriorates significantly when using incongruous RFs as input data.

3.6 Experience with Advanced Models

In this section, we describe our experience and lessons learned with more advanced models than the simple models presented in Sections 3.2 and 3.3. We also consider the impact of alternate data transport mechanisms, from the perspective of both the RFs and the scheduler itself.

3.6.1 Unreliable Data Transport

In the preceding discussion, the RF was a function mapping from the input rates to either MIPS or to the output rate. This was because System S, by default, operates in reliable data transport mode. In other words, no packets are dropped, and queues get backed up if the PEs do not get sufficient resources. As a result, in these RFs, input rates are the only independent variables. We also refer to this mode of operation, and the corresponding RFs, as the "no-drop" model.

On the other hand, System S can also operate in an unreliable mode, in the sense that reducing the MIPS allocation to a PE effectively forces packet drops. We call this the "drop" model, in contrast with the "no-drop" model. This affects the RF and the scheduler primarily in the following way.

- Implications to SODA: Since the PEs now drop packets when they do not have sufficient resources (in terms of MIPS), the scheduler has one more knob to twiddle, in its attempt to maximize system importance. It can, intentionally, decide to give less resources to some PEs than they would otherwise need. In some sense, SODA can now decide to "partially" allocate resources to PEs in some jobs, or even to parts of a job. To be able to do so meaningfully, SODA needs to know how the data output rates change as a function of the MIPS allocated to the PEs. In other words, MIPS is no longer an independent variable in the RFs.

- Implications to RFs: Now, the RF is a function mapping from the input rates and MIPS to the output rate. These sorts of functions, where MIPS are input rather than output, are appropriate for the cases in which resource usage limitations result in packet drops. These new kinds of RFs need to be learned from the data.

3.6.2 Non-parametric Approaches

The model we described previously is a parametric model. It assumes a particular formula and then determines parameters that best fit the data. If the formula is wrong, then even the best fit model will be a bad fit. An alternative is to use a non-parametric model. Non-parametric models, as the name suggests, do not assume a particular form of the resource function.

In particular, we elaborate on our experiences using one particular kind of non-parametric model for the RFs: a decision tree [10], with the "no-drop" mode mentioned above. Although the resulting RFs for a decision tree based model cannot be described as compactly as the parametric RFs, we hoped that using these more accurate RFs would result in better scheduling results.

The decision tree is trained using training data, with the CPU usage and input rate as the classification attributes. The output of the decision tree is an output rate, which is used by the scheduler. Given a set of training data, the output would be one of the values from the training set.

However, our initial decision-tree-based RFs resulted in bad scheduling decisions because all the PEs were allocated low amounts of resources. Upon investigation, we learned that the data were noisy. The noisy data resulted in decision tree outputs that did not satisfy the implicit form assumptions by the scheduler (listed in 2.1.2).

As a general trend, when MIPS increases, the output rate also increases. However, because the data were noisy, there were cases where $\mathsf{MIPS}_1 < \mathsf{MIPS}_2$, but $orate_1 > orate_2$. When the SODA scheduler does a series of RF queries and encounters a point where increasing the MIPS decreases the output rate, the scheduler decides that further increasing the

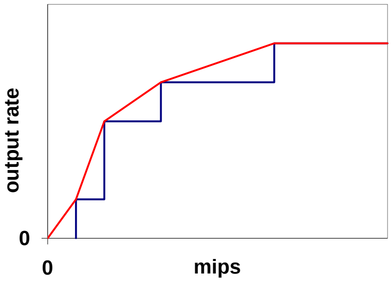


Figure 8: Illustrative decision tree RF and its envelope

MIPS will not produce a higher output. As a consequence, the source PEs (and correspondingly, the downstream PEs) end up with very few MIPS. While stopping under such a condition is a result of the SODA scheduler's specific algorithm, we believe that many algorithms using RFs would encounter the same problem and would probably behave similarly.

Our first response was cleaning up the data by discarding points that caused the problem. Once those outlier points had been removed, the resulting RF had a shape as in Figure 8. Yet, these RFs still had similar problems as the previous ones, though for a different reason. When the SODA scheduler encounters a flat spot (i.e., a place where $\mathsf{MIPS}_1 < \mathsf{MIPS}_2$ and $orate_1 = orate_2$), the scheduler acts as if it were the case that for any $\mathsf{MIPS} > \mathsf{MIPS}_1$, $orate = orate_1$. (Pictorially, it treats the RF as if the first flat spot lasted forever.) Thus, the MIPS assignments were again extremely small.

As a next step, we solved this problem by interpolating between the outer points (see Figure 8), so long as there was only one input. This model is effectively piecewise-linear and not truly non-parametric. The results achieved were not significantly better, so we discarded that approach as well. If there is more than one input, it is difficult to see how to ensure that the surface is increasing in both directions, and we did not test models of that type.

We also considered a parametric model that did not include the maximum function. The problem encountered there was quite interesting: the SODA scheduler never returned results. It kept allocating more and more MIPS, so its running time was very long.

From these different scenarios, we learned two lessons about the RFs:

- The RFs must be strictly increasing. Once they stop increasing, SODA will stop further exploration.
- They must have a flat spot where they stop increasing.

While these lessons are specific to the SODA scheduler, they should be applicable to other schedulers which use RF as their inputs. Furthermore, because of the fundamentally different way in which SODA works in the no-drop case, it may be possible for other model types to be effective in the no-drop case; this is a topic of future research.

4. MODEL AND DATA MANAGEMENT

The study of real PEs in the previous section indicates that an empirical approach based on collecting a few observations from each PE and constructing a model could be a simple and practical approach to obtaining useful RFs. While the previous section addresses the mathematics issues, it does not touch the management issues. It leaves several questions unanswered:

- What should we do about PEs when they are seen by the system for the first time, and there is no empirical data available for them yet?
- How can we identify whether previously collected data (or a model built from it) is applicable for a PE that runs in a slightly different environment than where the data is collected? In addition to the input rate dependency, a PE resource usage can depend on other factors such as: the nature of the input data (which may be a function of the upstream PEs), PE configuration (via, for example, command-line arguments), or systems issues, such as processing node architecture. This issue arises also when the same PE (e.g., a classifier) is reused in multiple applications.
- How should metrics be stored and used to build and update the PE RFs?

Our observation is that even though there are a myriad factors (in addition to input rate) that affect a PE's behavior, a PE will not be run in every possible configuration. Thus, rather than building a model that explicitly tries to model all these factors, we build and manage separate models for different combinations of those factors. This multiplicity of models raises the need for managing the models as well as the raw metrics data that is collected from the system. Specifically, two aspects must be addressed: (a) given a specific instance of a PE in a specific job that is submitted to the system, which model should be used by the scheduler? (b) given an observation of the resource usage of a specific PE instance, which models should be updated, and how?

To facilitate the model management, each PE is associated with a multi-part *signature*. We learn and maintain an RF for each signature. For a specific PE, the scheduler uses the signature to decide which RF should be used. In our system, we use the following four parts of the signature, in order of increasing specificity:

- **PE type**: source, transform or sink, as described in Section 3.1. In the future, it is possible to envision a much finer granularity of PE classification into types.
- **Executable:** the second part of the signature is the most general, consisting of an MD5 hash of the PE's executable. If the PE has been run before in any context, a learned model will be available. This will likely be better than a default model based only on PE type.
- **Arguments:** the third part is a MD5 hash of the arguments. A PE's command-line arguments may alter its behavior, so this piece attempts to capture this dependency. For simplicity, the arguments are simply treated as unique, unrelated categories.
- **Flowspec:** the fourth part is a representation of how the PE is connected to its upstream PEs, which is known as the *flow specification* or *flowspec*. In System S, streams themselves can be annotated by operations (such as filtering) which are performed by the

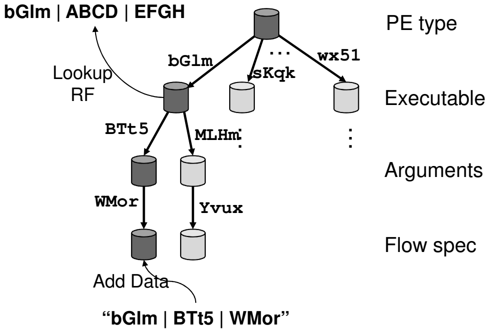


Figure 9: Hierarchical PE signature management

infrastructure rather than the PE at either end of the stream. The flowspec captures the specific interconnection of a PE to its senders and receivers, in terms of both the identity of PEs at the other end of each input/output stream, as well as the in-stream operations. It thus represents the most specific attribute of a particular instance of a PE that is connected in a specific way to other PEs.

The signature captures some key attributes of the PE which are knowable at *job submission time*. This allows the scheduler to choose an RF before the PE even begins execution. After execution, the PE RF may be refined further based on ongoing observations, but the initial lookup addresses the bootstrap problem.

The collection of signatures constitutes a hierarchical organization, as depicted in Figure 9. There is an RF for each node in this hierarchy. The node labeled ‘bGlm’ represents all PEs whose executable hashes to that string. Its two child nodes represent that PE executable being run with two different command-line arguments. Thus each node in the tree represents a “generalization” of all its children. To lookup the RF for a specific PE, we find the most specific node which matches the PE signature. In the figure, the lookup for ‘bGlm | ABCD | EFGH’ stops at the PE level node ‘bGlm’ because the rest of the signature represents entries that are not in the database (yet). Thus, if specific information about the PE in that context was available, that is the information that would be used. If no specific information was available, the system uses information based on only the executable. In the case of a brand new PE that has never been run before in any configuration, the system uses either (a) a generalized RF based on the root node (PE type), or (b) a default type-based RF which is hand-populated by us based on our calibrations from some earlier System S applications.

Analogously, a new data point is used to update the model at each of these three levels, starting at the most specific node and propagating up the tree. Thus, an observation $< r_j^I, m_p, r_k^O >$ for PE p with input port j and output port k, whose signature is ‘bGlm | BTI5 | WMor’ can update not only the model at the most specific node, but at every generalization above it in the path to the root, namely ‘bGlm | BTI5’ and ‘bGlm’. This approach allows the most specific models to reflect observations about the PE in the most specific context, while the more generalized models gather data from several sub-models. When a known PE is encountered in a new context, we can obtain a better approximation of that PEs behavior by using this generalized information.

Application	Level		
	Executable	Arguments	Flowspec
DAC	40	64	81
SKA	27	102	102
VWAP	25	365	365

Table 1: Count of signatures by level for application PEs

To highlight the possibility of reuse in actual applications, Table 1 shows the number of unique paths at each level in the tree for our applications. SKA and VWAP (and most SPADE-based applications) do not use flowspecs on their streams, so each argument-level node has only one child node (the ‘null’ flowspec). However, these applications contain several replicas of the same PE, executing with different arguments. This indicates that maintaining the executable-level information is likely to be useful if we encounter a new PE with a different set of arguments than what is seen before.

5. RELATED WORK

In the literature on resource allocation and scheduling in distributed systems, the resource requirements are typically assumed to be known or given. Much of the other known modeling work has occured in the context of single and multi-tier distributed systems. In [13], the authors develop and use linear models for CPU, disk and memory demands based on incoming workload rates, similar to our models. A more complex, analytical queueing model of multi-tier services is developed in [14]. However, this model is difficult to apply to streaming systems which are not neatly organized into tiers – in general they are directed graphs, may have cycles and little identifiable substructure.

In general cluster environments, [15] use kernel-based monitoring tools to learn application profiles in terms of stochastic token-bucket models of CPU and network usage. This approach builds an application-level workload-independent usage profile, and schedules to the tail of the distribution. In our case, for streaming applications, the workload data rates are expected to vary widely, and the system is expected to be quite dynamic. As we see, PEs are very sensitive to the incoming data rates, so a rate-sensitive model is needed to ensure responsiveness to changes in resource demands.

For streaming systems, a cost-per-tuple model is also proposed [21] in the context of the Borealis system. In their case, however, the operators modeled are much simpler (like SPADE operators) compared to the PEs in our system. We improve on their work by showing that even complex PEs can be modeled using a similar approach, and further we propose a scheme for managing and generalizing these models.

[13] also raises the issues that models may become inaccurate due to interference caused by colocation of PEs on a node. They mention that developing models in heterogeneous clusters is a challenge, but others [9] have suggested a solution involving parametric cross-architecture models.

Model management is an even less discussed topic. For the profiling step, some authors [5, 13] suggest running the applications on idle nodes for accurate measurements.

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented the challenges of predicting the CPU and network usage of PEs in System S applications. An empirical study of PEs from applications in a System S testbed reveals that simple piecewise linear models based on their input rates are sufficient for modeling these PEs, which is encouraging given that some of these PEs perform relatively complex analytics. We have also presented an approach based on hierarchical PE signatures for managing and updating these models that addresses the issues of getting usable models for new PEs and allowing one PE's model to be used effectively for another PE. Although not discussed in this paper, our RFs are dynamically updated based on new observations from the running PEs, this allows the SODA scheduler to respond to dynamic changes in resource demands of the applications.

Although we find that many PEs can be modeled using these linear models, there will be PEs that do not fit into this scheme. Although our initial attempts with non-parametric approaches were not very successful, we aim to refine them and pursue other advanced techniques that can capture a larger set of PEs. Our current models also are limited to handling CPU and network, and may even be generalizable to disk resources [13]. However, the issue of modeling memory consumption is still an open issue. In general, since PEs can arbitrarily allocate memory at runtime, this is a difficult issue. Currently SODA relies on PE developers to provide hints about the memory needs of their applications, but an automated RF-based approach would allow even the memory demands to be taken into account during scheduling. Finally, our use of the MIPS metric does not generalize well across architectures. For clusters of truly heterogeneous nodes, especially ones containing specialized resources (such as the Cell processor), a more generalizable metric and model (for example, [9]) would be very useful.

Acknowledgment

We would like to acknowledge Nagui Halim, the Principal Investigator, and Lisa Amini, the technical leader, of the System S project for providing us with invaluable support and guidance throughout the development of SODA. Wei Fan provided immense help and guidance on the development of the decision tree based RFs. We also thank the SPC team, Henrique Andrade, Yoonho Park, Richard P. King, Philippe L. Selo, and Chitra Venkatramani, for their numerous help and support. Last, but not least, we thank Louis R. Degenaro, Jim Challenger and James Giles for their help, suggestions and feedback on the integration of SODA into the rest of System S.

7. ADDITIONAL AUTHORS

8. REFERENCES

[1] D. J. Abadi, Y. Ahmad, M. Balazinska, U. Cetintemel, M. Cherniack, J.-H. Hwang, W. Lindner, A. S. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik. The design of the Borealis stream processing engine. In *Proceedings of the Conference on Innovative Data Systems Research*, 2005.

[2] L. Amini, H. Andrade, R. Bhagwan, F. Eskesen, R. King, P. Selo, Y. Park, and C. Venkatramani. SPC: a distributed, scalable platform for data mining. In *Proceedings of the International Workshop on Data Mining Standards, Services and Platforms*, 2006.

[3] H. Andrade, B. Gedik, K.-L. Wu, and P. S. Yu. Scale-up strategies for processing high-rate data streams in System S. In *Proceedings of the IEEE International Conference on Data Engineering, to appear*, 2009.

[4] A. Arasu, B. Babcock, S. Babu, M. Datar, K. Ito, I. Nishizawa, J. Rosenstein, and J. Widom. STREAM: the Stanford stream data manager (demonstration description). In *SIGMOD '03: Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 665–665, New York, NY, USA, 2003. ACM.

[5] M. Aron. *Differentiated and Predictable Quality of Service in Web Server Systems.* PhD thesis, Computer Science, Rice University, Oct. 2000.

[6] A. Biem. Imaging for next-generation radio telescopes. Personal communication, July 2008.

[7] T. J. Cornwell, K. Golap, and S. Bhatnagar. W projection: A new algorithm for wide field imaging with radio synthesis arrays. In P. Shopbell, M. Britton, and R. Ebert, editors, *Proceedings of Astronomical Data Analysis Software and Systems XIV ASP Conference Series*, volume 347 of *2005ASPC*, page 86. San Francisco: Astronomical Society of the Pacific, 12 2005.

[8] B. Gedik, H. Andrade, K.-L. Wu, P. S. Yu, and M. Doo. SPADE: The System S declarative stream processing engine. In *SIGMOD '08: Proceedings of the ACM SIGMOD International Conference on Management of Data.* ACM, 2008.

[9] G. Marin and J. Mellor-Crummey. Cross-architecture performance predictions for scientific applications using parameterized models. In *Proceedings of the 2004 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, New York, NY, USA, June 10–14 2004.

[10] T. M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[11] R. Motwani, J. Widom, A. Arasu, B. Babcokc, S. Babu, M. Datar, G. Manku, C. Olston, J. Rosenstein, and R. Varma. Query processing, approximation, and resource management in a data stream management system. In *Proceedings of the 1st Conference on Innovative Data Systems Research*, 2003.

[12] D. Sankus, R. Redburn, D. S. Turaga, M. C. Johnson, A. Norfleet, O. Verscheure, and W. Fan. Drowning in data - a streaming solution, Submitted to International Test Conference, 2009.

[13] C. Stewart and K. Shen. Performance modeling and system management for multi-component online services. In *Proceedings of the 2nd Symposium on Networked Systems Design and Implementation (NSDI'05)*, Boston, MA, May 2–4 2005.

[14] B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitzer, and A. Tantawi. An analytical model for multi-tier internet services and its applications. In *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, Banff, Alberta, Canada, June 6–10

2005.

[15] B. Urgaonkar, P. Shenoy, and T. Roscoe. Resource overbooking and application profiling in shared hosting platforms. In D. Culler and P. Druschel, editors, *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI'02)*, pages 239–254, Boston, MA, Dec. 9–11 2002.

[16] W. van Dorst. BogoMips mini-Howto. `http://tldp.org/HOWTO/BogoMips/`.

[17] J. Wolf, N. Bansal, K. Hildrum, S. Parekh, D. Rajan, R. Wagle, and K.-L. Wu. Job admission and resource allocation in distributed streaming systems. In *Proceedings of the International Workshop on Job Scheduling Strategies for Parallel Processing*, 2009.

[18] J. Wolf, N. Bansal, K. Hildrum, S. Parekh, D. Rajan, R. Wagle, K.-L. Wu, and L. Fleischer. Scheduling optimizer for distributed applications: A reference paper. Technical Report 24453, IBM Research, 2007.

[19] J. L. Wolf, N. Bansal, K. W. Hildrum, S. Parekh, D. Rajan, R. Wagle, K.-L. Wu, and L. Fleischer. SODA: An optimizing scheduler for large-scale stream-based distributed computer systems. In *Proceedings of the ACM/IFIP/Usenix International Middleware Conference*, 2008.

[20] K.-L. Wu, P. S. Yu, B. Gedik, K. W. Hildrum, C. C. Aggarwal, E. Bouillet, W. Fan, D. A. George, X. Gu, G. Luo, and H. Wang. Challenges and experience in prototyping a multi-modal stream analytic and monitoring application on System S. In *VLDB '07: Proceedings of the 33rd International Conference on Very Large Data Bases*, Vienna, Austria, Sept. 2007.

[21] Y. Xing, J.-H. Hwang, U. Çetintemel, and S. B. Zdonik. Providing resiliency to load variations in distributed stream processing. In *VLDB '06: Proceedings of the 32nd International Conference on Very Large Data Bases*, pages 775–786, Seoul, Korea, Sept.12-15 2006. ACM.

[22] S. Zdonik, M. Stonebraker, M. Cherniack, U. Cetintemel, M. Balazinska, and H. Balakrishnan. The Aurora and Medusa projects. *IEEE Data Engineering Bulletin*, 26(1), 2003.