Performance Tuning of Streaming Applications via Search-space Decomposition

Authors: Shobana Padmanabhan, Roger D. Chamberlain, and Yixin Chen

High-performance streaming applications are typically pipelined and deployed on architecturally diverse (hybrid) systems. Developers of such applications are interested in customizing components used, so as to benefit application performance. We present an efficient and automatic technique for design-space exploration of applications in this problem domain. We solve performance tuning as an optimization problem by formulating cost functions using results from queueing theory. This results in a mixed-integer nonlinear optimization problem which is NP-hard. We reduce the search complexity by decomposing the search space. We have developed a domain-specific decomposition technique using topological information of the application embodied in the queueing network models. Our analysis includes when our decomposition preserves optimality. Our preliminary empirical results confirm two-fold benefits--solving a problem that is currently not solvable using state-of-the-art solvers and in some problem instances, improving initial solution value from the solver by over two orders of magnitude.

... Read complete abstract on page 2.
Performance Tuning of Streaming Applications via Search-space Decomposition

Complete Abstract:

High-performance streaming applications are typically pipelined and deployed on architecturally diverse (hybrid) systems. Developers of such applications are interested in customizing components used, so as to benefit application performance. We present an efficient and automatic technique for design-space exploration of applications in this problem domain. We solve performance tuning as an optimization problem by formulating cost functions using results from queueing theory. This results in a mixed-integer nonlinear optimization problem which is NP-hard. We reduce the search complexity by decomposing the search space. We have developed a domain-specific decomposition technique using topological information of the application embodied in the queueing network models. Our analysis includes when our decomposition preserves optimality. Our preliminary empirical results confirm two-fold benefits—solving a problem that is currently not solvable using state-of-the-art solvers and in some problem instances, improving initial solution value from the solver by over two orders of magnitude.
Performance Tuning of Streaming Applications via Search-space Decomposition

Authors: Shobana Padmanabhan, Roger D. Chamberlain, Yixin Chen

Corresponding Author: shobana@arl.wustl.edu

Abstract: High-performance streaming applications are typically pipelined and deployed on architecturally diverse (hybrid)systems. Developers of such applications are interested in customizing components used, so as to benefit application performance. We present an efficient and automatic technique for design-space exploration of applications in this problem domain.

We solve performance tuning as an optimization problem by formulating cost functions using results from queueing theory. This results in a mixed-integer nonlinear optimization problem which is NP-hard. We reduce the search complexity by decomposing the search space. We have developed a domain-specific decomposition technique using topological information of the application embodied in the queueing network models. Our analysis includes when our decomposition preserves optimality. Our preliminary empirical results confirm two-fold benefits--solving a problem that is currently not solvable using state-of-the-art solvers and in some problem instances, improving initial solution value from the solver by over two orders of magnitude.

Type of Report: Other
Performance Tuning of Streaming Applications via Search-space Decomposition

Shobana Padmanabhan, Roger D. Chamberlain, and Yixin Chen
Dept. of Computer Science and Engineering, Washington University in St. Louis

Abstract

High-performance streaming applications are typically pipelined and deployed on architecturally diverse (hybrid) systems. Developers of such applications are interested in customizing components used, so as to benefit application performance. We present an efficient and automatic technique for design-space exploration of applications in this problem domain.

We solve performance tuning as an optimization problem by formulating cost functions using results from queueing theory. This results in a mixed-integer nonlinear optimization problem which is NP-hard. We reduce the search complexity by decomposing the search space. We have developed a domain-specific decomposition technique using topological information of the application embodied in the queueing network models. Our analysis includes when our decomposition preserves optimality. Our preliminary empirical results confirm two-fold benefits—solving a problem that is currently not solvable using state-of-the-art solvers and in some problem instances, improving initial solution value from the solver by over two orders of magnitude.

1 Introduction

In streaming applications, application data is fed forward over a pipeline of processing elements to achieve high performance. Examples of streaming applications abound in the areas of biosequence analysis, computer networking, signal processing, video processing, image processing, and computational science. Streaming applications are sufficiently widespread that several programming languages have been developed for them, including Brook [4], StreamIt [19], and X [11].

We define performance tuning as finding the best values for the different performance-related design parameters, given performance goals and constraints. Examples of design parameters are sizes of buffers in the processing elements, delivery size of messages communicated between processing elements, mapping of processing and communication elements onto physical resources and choices of algorithms. Performance tuning is hard because:

- The number of possible configurations is exponential in the number of performance parameters. For our example streaming sort application there are over $10^{21}$ possible configurations.
- Performance goals are often multiple and conflicting. For instance, it is typical to minimize application latency (end-to-end execution time) while simultaneously maximizing application throughput (input data rate), but these goals often conflict with each other.

Performance tuning has been an active research topic in numerous application areas. For embedded applications, performance tuning has been researched using search heuristics as well as standard and modified optimization techniques [9, 10, 17]. A recent trend has been to model the application’s performance analytically (mathematically) and use the model with search heuristics. Examples of such models include predictive models [12, 14] and examples of search heuristics include gradient ascent [14]. Predictive models are based on regression or machine learning and trained with empirical experimentation through simulation or direct execution. Such models are general and hence can be applied to any design space.

Our approach is to solve performance tuning as an optimization problem; for modeling cost functions, we prefer queueing networks (QNs) over regression-based models. This is because pipelining in the streaming applications gives the application special structure and frequently involves queueing between the processing elements. Examples of streaming applications that have been modeled using queueing networks include computer systems, computer networking, software architecture, biosequence search, and web servers [5, 13, 21].

The resulting optimization problem formulation is a mixed-integer nonlinear problem (MINLP). A characteristic of MINLP problems is that the search time is known to grow exponentially with respect to the number of variables [3]. Further, the nonlinear functions in our optimization problem are neither convex nor quasiconvex which means there is no theory that guarantees finding a global optimum. In practice, state-of-the-art solvers failed to find even a feasible solution to our problem.
A popular approach with MINLP problems is to decompose the search space in order to obtain a better solution from optimization [8]. However, the state-of-the-art solvers [16] such as Bonmin, FilMINT, KNITRO, and MINLP, fail to even find a feasible solution for our problem formulation. It is a common practice to use domain-specific information to guide decomposition [20]. However, no such technique exists for our problem domain.

We have developed a domain-specific decomposition technique that:

- Identifies and characterizes topological information in the domain and uses it to guide the decomposition.
- Identifies the specific MINLP form of problems in the domain.
- Results in a heuristic to order variables and constraints so as to improve decomposability and handle complications. Our analysis includes when our decomposition preserves optimality.

We describe a preliminary empirical validation of the benefits of our techniques. The two-fold benefits are: we solve a problem that is currently not solvable using state-of-the-art solvers, and for a relaxed version of the problem that is solvable, we improve the initial solution value by over two orders of magnitude.

2 Domain-specific decomposition

In general, a problem is considered decomposable if we can arrange the variables and constraints such that a matrix showing the presence of the variables in the different constraints has the canonical Jordan block form [18]. This form is illustrated in Figure 1. In this form, the objective function also decomposes to correspond to the different blocks (i.e., we can solve $Obj_1$ along with the variables and constraints forming block$_1$ as an independent subproblem). Note that decomposing this way preserves optimality.

A variable that prevents decomposition into blocks is traditionally referred to as a “complicating variable” (CV) and a constraint that prevents decomposition is called a “complicating constraint” (CC). The effects of these complications on the Jordan canonical form are shown in Figure 2 and Figure 3 respectively. Decomposition techniques exist for dealing with CVs and CCs [8]. For MINLP problems in particular, integer variables are treated as CVs and if the resulting nonlinear program (NLP) is convex at least locally, Benders decomposition is known to converge to an optimal solution (or to some small duality gap). Another approach with MINLP problems is to consider the nonlinear constraints as CCs and apply the outer linearization algorithm, provided the nonlinear constraints are inequalities and the objective function is linear (both of which can be usually achieved through simple transformations).

In real-world streaming applications however, the number of CVs and CCs tend to exceed the number of non-complicating variables and constraints making it very hard to achieve decomposability. Hence, we have developed a domain-specific decomposition technique that exploits the pipelining structure of the streaming applications.

Domain-specific topological information We characterize the topological information about streaming applications modeled by queueing networks (QNs) as follows.

- $ip = \{ip_i | ip_i \in \mathbb{R}_+ \cup \{0\}\}$ is the set of input parameters whose optimal settings are to be determined. In practice, most of the parameters are integer- or binary-
valued.

- \( \text{var} = \{ \text{var}, |\text{var}| \in \mathbb{R}_+ \cup \{0\} \} \) is the set of design variables in the optimization problem formulation that correspond to the input parameters. \( n_v = |\text{var}| \) and \( \mathbf{v} \in (\mathbb{R}_+ \cup \{0\})^{n_v} \) is the vector formed by the variables in \( \text{var} \). In most cases, the design variables are the same as the input parameters. An example of design variables that are distinct from their corresponding input parameters arises in the following context. Let \( m \in \mathbb{Z}_+ \) be an input parameter that controls mapping choices. In our optimization model, we transform \( m \) into a set of binary design variables, one for each mapping choice, and constrain them so that only one of them is selected.

- \( \text{top} \subseteq \text{var} \) is the set of design variables that result in distinct alternative QNs. \( n_t = |\text{top}| \) and \( \mathbf{t} \) is the vector formed by the variables in \( \text{top} \). We call these variables topological variables, since they impact the topology of the queueing network.

- \( \text{der} = \{ \text{der}, |\text{der}| \in \mathbb{R}_+ \cup \{0\} \} \) is the set of derived variables that depend on one or more elements in \( \text{var} \). \( n_d = |\text{der}| \) and \( \mathbf{d} \) is the vector formed by the variables in \( \text{der} \).

- \( \text{met} \subseteq \text{der} \) is the set of performance metrics (such as application latency, throughput, power consumption, etc.) that are being optimized. \( n_m = |\text{met}| \) and \( \mathbf{m} \) is the vector formed by the variables in \( \text{met} \).

- \( z \in \text{der} \) is the cost function, also known as the objective function. 
  \[
  z = \sum_{k=1}^{n_m} W_k \times \mathbf{m}_k : (\mathbb{R}_+ \cup \{0\})^{n_m} \rightarrow \mathbb{R}_+ + \{0\} \quad \sum_{k=1}^{n_m} W_k = 1
  \]

- \( \text{ivar} \subseteq \text{der} \) is the set of intermediary variables (IVs). \( n_q = |\text{ivar}| \) and \( \mathbf{q} \) is the vector formed by the variables in \( \text{ivar} \). \( \text{ivar} = \text{der} - \text{met} - \{z\} \).

Intermediate variables may arise for a number of reasons: (1) application developers maybe interested in the values of IVs for debugging purposes; (2) IVs can codify abstractions in the performance models such as QNs; and (3) IVs may help some search heuristics. For example, MINLP solvers using cutting-plane based algorithms, such as FilMINT [1], tend to work better with linear cost functions while solvers using the interior-point algorithm tend to work better with linear constraints [3].

- \( \text{mu} \subseteq \text{ivar} \) is the set of mean service rates at each queueing station. \( n_u = |\text{mu}| \) and \( \mathbf{u} \) is the vector formed by the variables in \( \text{mu} \).

\[
\mu_j = u_j(\mathbf{v}) : (\mathbb{Z}_+ \cup \{0\})^{n_u} \rightarrow \mathbb{R}_+ \text{ where } \mathbf{u} \text{ is the vector of functions that define } \mu.
\]

- \( \text{lam} \subseteq \text{ivar} \) is the set of mean job arrival rates at each queueing station. \( n_l = |\text{lam}| \) and \( \mathbf{t} \) is the vector formed by the variables in \( \text{lam} \). \( \lambda \) are related by a system of linear expressions with a unique solution (in terms of the input parameter \( \lambda_{in} \in \text{var} \)). 
  \[
  \lambda_j = I_j(\lambda_{in}, \mathbf{t}) : (\mathbb{R}_+ \cup \{0\})^{n_l} \rightarrow (\mathbb{R}_+ \cup \{0\})^{n_l}, \lambda_{in} \in \mathbf{v} \text{ where } \mathbf{I} \text{ is the vector of functions that define } \lambda.
  \]

- \( \text{sq} \subseteq \text{ivar} \) is the set of Single-QS (SQ) variables. Each element of \( \text{sq} \) is in the domain of only one \( u_j(\cdot) \). \( n_s = |\text{sq}| \) and \( \mathbf{s} \) is the vector formed by the variables in \( \text{sq} \).

- \( \text{mq} \subseteq \text{ivar} \) is the set of Multi-QS (MQ) variables. Each element of \( \text{mq} \) is in the domain of more than one \( u_j(\cdot) \). \( n_z = |\text{mq}| \) and \( \mathbf{m} \) is the vector formed by the variables in \( \text{mq} \).

\( \mathbf{ul} \) is a vector of model constraints that restricts every \( \lambda_j < \mu_j \) for the system to be stable.

Topologies of streaming applications are restricted to directed acyclic graphs while QN topologies are annotated digraphs. Annotations on each digraph include, at the minimum, expressions for each of \( \mu, \lambda, \) and \( o \). Each node of the digraph represents a queueing station (QS) which is a service facility with its queue. Each edge represents the communication link between two nodes.

**Domain-specific MINLP form** The general form of the optimization function of a streaming application modeled using queueing networks is as follows. The nonlinear functions may not be continuous or differentiable.

\[
\begin{align*}
\min_{\mathbf{z}} & \quad \sum_{k=1}^{n_m} W_k \times o_k(\mathbf{v}), \sum_{k=1}^{k} W_k = 1 \\
\text{subject to} & \quad \mathbf{u}(\mathbf{v}) = 0 \\
& \quad \mathbf{I}(\lambda_{in}, \mathbf{t}) = 0 \\
& \quad \mathbf{ul}(\lambda, \mu) \leq 0
\end{align*}
\]

The presence of \( \mathbf{t} \) is a discovery from our experimentation described in Section 4. There may be multiple QNs for a given application topology. If \( \text{top} \neq \emptyset \), we use a set of binary variables to model the QNs resulting from every variable in \( \text{t} \). In that case, there exist (equality) constraints to choose only one of the binary variables from each set.

**Domain-specific complicating variables and optimality** It is our observation that the topological variables \( \mathbf{t} \) are special complicating variables in that decomposing by them results in completely independent subproblems by definition. That means resolving solutions from the subproblems
is trivial because we only need to identify the minimum valued solution. That also means decomposing by these variables will preserve optimality.

On the other hand, $\lambda$ and multi-QS variables ($\mathbf{m}$) are complicating variables that will have to be resolved if decomposed over. Because $\lambda$ are constrained by $\mu$ through $\mathbf{u}$, $\mu$ have the same effect on decomposition as the $\lambda$. In our ongoing work, we handle this case by implementing branch-and-bound where we use the variable-constraint ordering described below to determine the branching order.

**Domain-specific variable-constraint ordering** To identify decomposability, we have developed the following heuristic to order the variables and constraints our variable-constraint matrix gets closer to the canonical Jordan block form as mentioned at the beginning of this section.

1. Start with $\mathbf{sq}$. If multiple elements in $\mathbf{sq}$, include them in the order of the pipeline which is also the order of the queueing stations in the QN model. Include first all the variables from $\mathbf{sq}$ that concern a given QS. For example, include all SQ variables that affect $QS_0$ and then include all SQ variables that affect $QS_1$ etc.

2. Include MQ variables relating, increasingly, to one or more QS variables. When relating to multiple other QS variables, they are ordered in the direction of the application’s pipeline.

3. Include variables from $\mathbf{top}$, $\mathbf{ivar}$, $\mathbf{met}$, and finally $\mathbf{z}$.

4. Order the constraints, starting with $\mathbf{u}$ and then $\mathbf{l}$. Include the constraints, again, in the direction of the pipeline. For example, include the constraints that are isolated to $QS_0$ before including those for $QS_1$ and so on. If all the constraints in $\mathbf{u}$ and $\mathbf{l}$ span QSs, then, include them so that the constraints for relating fewer QS variables are before the ones relating more QS variables.

5. Follow with the equality and inequality constraints of $\mathbf{ul}$, $\mathbf{o}$, and then the binary constraints for $\mathbf{t}$.

We use the ordering of variables and constraints to guide our decomposition of the search space. For the complicating variables that are elements of $\mathbf{top}$, our decomposition works as follows.

1. Start with the CV that appears in the most number of constraints and evaluate each of its values as an independent subproblem (i.e., branch on the CV).

2. For each subproblem at this step, we decompose the subproblem by the next CV.

3. Repeat until we have decomposed all the CVs.

By the nature of applications in our problem domain, we do not expect all CVs to be from $\mathbf{top}$. Dealing with CVs not in $\mathbf{top}$ is our ongoing work. We are implementing a branch-and-bound algorithm guided by the ordering of variables and constraints discussed above.

### 3 Example Application

Our example application is a streaming (parallel) sort. Although sorting is a simple application, it is prototypical of how streaming applications get parallelized for high performance. In addition, the number and type of performance-related parameters for streaming sort on our deployment platform is also prototypical.

In streaming sort, input data is split into parts and each part is sent to a sort instantiation. This application topology is shown in Figure 4. The sort instantiation is referred to as a sort “block” in the parlance of Auto-Pipe [11]. The sort blocks execute in parallel and when done, each block sends its output to a merge block. The merge block then merges its inputs and sends out the sorted data. The edges between the blocks are communication links. A “column” refers to all the blocks or links at the same level in Figure 4.

The set of design variables, $\mathbf{var}$, along with their characterization, bounds and constraints, for streaming sort are enumerated in Figure 5. Note that the number of sort blocks, denoted by $2^N$, controls the degree of parallelization which in turn controls the tradeoff between application latency and throughput, reflected in a change in the application topology as illustrated in Figure 6.

*Mapping* of the application blocks and edges onto available physical resources is an important concern during application deployment. Here, we consider only a few interesting and prototypical mappings. The choices are represented by the binary variables $m_0, m_1, m_{CHR}, m_{IR}$ which are constrained to have only one of them be true. $m_0$ models
when every block in the application gets its own set of resources, $m_1$ models when all compute blocks share a single resource, $m_{CB}$ models when all split blocks share a single computation resource, and $m_{IR}$ models when the communication links into and out of the sort blocks are shared. The variables controlling both the number of sort blocks and the mapping are topology variables, elements of $top$.

Queueing network model  Queueing network models represent a system as an interconnected set of queueing stations and customers (jobs) serviced by those queueing stations. Each queueing station has one or more servers. Queueing stations are conventionally labeled with the notation $a/b/s$, where $a$ and $b$ represent the distribution of interarrival and service times respectively; and $s$ represents the number of servers. The model we use initially is the classic $M/M/1$ model where $M$ is an exponential (memoryless or Markovian) distribution.

If in a BCMP queueing network each station has an infinite queue, it follows from the equivalence property that (under steady-state conditions) each station can be analyzed independently [2]. Here, we restrict ourselves to $M/M/1$ BCMP networks with infinite buffers and FIFO queueing discipline. We make these assumptions only to prevent complicating the analytical expressions. Inherently, our work can handle relaxation of each of these assumptions as long as there are known results in queueing theory to handle the relaxation. See, e.g., [13] for approaches to handling finite queue sizes and/or phase-type service distributions.

For the sort application, we begin with the application

\[
\begin{align*}
B_{SZ} &= K, K \in \mathbb{Z}_+ \text{ (e.g., 20)}, \\
N &= 1, 2, 3, ..., \frac{2^{M-2} - 2}{2^{M-2}}; N \in \text{top} \\
\text{Sort algorithm (only with cpu) mapping) } &= \text{binary: cpu, or fpga}, \\
\text{Communication resource type } &= \text{binary: smem, or gige,} \\
\text{Number of sort blocks } &= \text{m, } j = 0, 2, ..., 4N; \forall j, nRes_j \geq 1 \\
\text{Number of compute resources } &= \text{nRes, } j = 1, 3, ..., 4N - 1; \forall j, nRes_j \geq 1 \\
\text{Number of communication resources } &= \text{nRes, } j = 1, 3, ..., 4N - 1; \forall j, nRes_j \geq 1 \\
\text{Mapping choices } &= \text{binary: m0, m1, mCB, mIR, } m_0 + m_1 + m_{CB} + m_{IR} = 1; \text{ each } m_i \in \text{top} \\
\text{System-wide comm message size } &= \text{2^M, } M = 0, 1, ..., M_{UB} \text{ where } M_{UB} \leq N \text{ and } M_{UB} = K, K \in \mathbb{Z}_+ \text{ (e.g., 14).} \\
\text{Sort algorithm (only with cpu) mapping) } &= \text{binary: alg1 or alg2, } m_1(fpga[0]) + alg1 + alg2 \\
\text{Input mean job arrival rate } &= \lambda_{in} \in \mathbb{R}_+ \\
\text{By solving } u \text{, we derive these expressions based on first principles and}
\end{align*}
\]

Figure 5: Performance-related parameters of streaming sort

Figure 7: Queuing network model

Figure 8: Queueing network has feedback when communication edges to/from sort blocks are shared

Cost functions and validation  Recall from Section 2 that expressions for $\mu$ are given by $u$ and are generally input by the application developer. For our example application, we derived these expressions based on first principles and
Formulate optimization problem We use the standard weighted sum [15] technique to combine the multiple (normalized) performance objectives (met from Section 2) as shown in equation 4. Note that if we optimized only the application’s throughput, given by \( \frac{1}{\lambda} \), the problem degenerates to identifying the bottleneck in the pipeline.

\[
\text{minimize } \quad W_1 \times \text{Latency} + W_2 \times \frac{1}{\lambda_{\text{in}}} \sum_{i=1}^{2} W_i = 1 \quad (4)
\]

We use the equivalence property of M/M/1 BCMP networks with infinite buffers and a FIFO queueing discipline [2] to define the mets. Accordingly, latency is given by:

\[
\text{Latency} = \sum_{j=0}^{4N} \frac{1}{\mu_j - \lambda_j} \quad (5)
\]

As the example equations above illustrate, the equations are highly nonlinear and the state-of-art-solvers we mentioned earlier are unable to find even a feasible, much less optimum, solution. The number of variables and constraints in the original problem range from (50, 30) to (399, 3077) as we increase \( N \) from 1 to 13. This corresponds to \( 2^N = 2 \) to 8192 sort blocks, given a batch size of \( 2^{14} \) elements per sort.

Formulate optimization problem We use the standard weighted sum [15] technique to combine the multiple (normalized) performance objectives (met from Section 2) as shown in equation 4. Note that if we optimized only the application’s throughput, given by \( \frac{1}{\lambda} \), the problem degenerates to identifying the bottleneck in the pipeline.

\[
\text{minimize } \quad W_1 \times \text{Latency} + W_2 \times \frac{1}{\lambda_{\text{in}}} \sum_{i=1}^{2} W_i = 1 \quad (4)
\]

We use the equivalence property of M/M/1 BCMP networks with infinite buffers and a FIFO queueing discipline [2] to define the mets. Accordingly, latency is given by:

\[
\text{Latency} = \sum_{j=0}^{4N} \frac{1}{\mu_j - \lambda_j} \quad (5)
\]

As the example equations above illustrate, the equations are highly nonlinear and the state-of-art-solvers we mentioned earlier are unable to find even a feasible, much less optimum, solution. The number of variables and constraints in the original problem range from (50, 30) to (399, 3077) as we increase \( N \) from 1 to 13. This corresponds to \( 2^N = 2 \) to 8192 sort blocks, given a batch size of \( 2^{14} \) elements per sort.

Variables other than \( \mu \), \( \lambda \), Latency, and Throughput are integer-valued, making the problem mixed-integer. The expression for latency, handling of multiple mapping choices, and handling the number of sort blocks as a variable all make the problem nonlinear, and some of the nonlinear functions are nonconvex and non-quasiconvex.

4 Preliminary empirical results

We first categorize the variables and use the categories to order the variables and constraints to achieve decomposability per our heuristic described in Section 2. The resulting matrix structure is shown in Figure 9. The matrix shows significant decomposable parts but it also shows the presence of many complicating variables and constraints.

We first decompose the complicating variable \( N \) and then \( m \). \( N \) determines the number of sort blocks, and \( m \) is the set of mapping choices. Let \( p \) be the original problem that evaluates all values of \( N \) and all the mapping choices simultaneously. Let \( p_i \) denote the subproblem obtained by decomposing \( p \) by fixing \( N \) at \( i \) (which implies the number of sort blocks is \( 2^i \)) but considers all the mapping choices. Then, we decompose each \( p_i \) based on \( m \). We denote the subproblem of \( p_i \) that considers only \( m_0 \) by \( p_{i,0} \), only \( m_1 \) by \( p_{i,2} \), only \( m_{CR} \) by \( p_{i,3} \), and only \( m_{IR} \) by \( p_{i,4} \). The relationship among the subproblems in the hierarchical decomposition is shown in Figure 10.

The objective function values (equally-weighted sum of application latency and job interarrivial time to sort a batch of a million 64-bit elements) from the solutions of each of the subproblems, as solved by FilmINT [1], are presented in Figure 11 (up to \( N = 9 \)). Recall that the solution to the original problem is simply the minimum of the solutions from all the subproblems. This minimum value is highlighted
The solution of every subproblem through our heuristic is not necessarily a local optimum because the cost functions and some constraints still remain non-convex in the subproblems. For our example application, an example of solution to a subproblem not being a local optimum is the subproblem $p_{3,0}$. The solution has $M = 4$ but increasing $M$ to 5 in a neighborhood search lowers the objective function value from 299.02 to 298.87 ms. While the difference between these two values is actually not truly significant from the point of view of the application developer (i.e., the performance models are very unlikely to be accurate to that many significant digits), it is significant that the solver is unable to find even a local optimum.

Our solution is indeed sensitive to the application’s performance goals. For instance, rather than optimizing equally for both application latency and throughput, if we increase the weight on throughput to 0.9 and reduce the weight on latency to 0.1, the recommended configuration changes from $(N = 3, m = m_{1R})$ to $(N = 4, m = m_0)$ with a corresponding objective function value of 184.5 ms.

**Problem variation** In the problem instance we have considered so far, the communication architecture depends on the resource type selected for the processing elements. For example, in Figure 4 if a software implementation is used for each of the split and sort blocks, the communication architecture can be shared memory or Gigabit Ethernet, but if one of the ends is mapped instead to an FPGA, the communication architecture is automatically set to PCI-X.

The formulation for these constraints, however, is highly nonlinear and therefore we relaxed these constraints (and simultaneously changed a number of the constants) to form a variation of the problem that we call “relaxed streaming sort.” The solver does manage to solve this relaxed version of the problem, but the improvement is highlighted. The subproblems and their solutions are indexed using the same convention as above.

### Analysis of empirical results

The solver runtime for the different subproblems ranged from 20 ms to approximately 7 s and hence the overhead from decomposition is not a concern for this problem. Although we are solving many subproblems instead of just one (original) problem, each subproblem is progressively less complex. After decomposing, the number of variables and constraints in any subproblem is no more than 380 and 326, respectively.

The solution of every subproblem through our heuristic is not guaranteed to be necessarily a local optimum because the cost functions and some constraints still remain non-convex in the subproblems. For our example application, an example of solution to a subproblem not being a local optimum is the subproblem $p_{3,0}$. The solution has $M = 4$ but increasing $M$ to 5 in a neighborhood search lowers the objective function value from 299.02 to 298.87 ms. While the difference between these two values is actually not truly significant from the point of view of the application developer (i.e., the performance models are very unlikely to be accurate to that many significant digits), it is significant that the solver is unable to find even a local optimum.

Our solution is indeed sensitive to the application’s performance goals. For instance, rather than optimizing equally for both application latency and throughput, if we increase the weight on throughput to 0.9 and reduce the weight on latency to 0.1, the recommended configuration changes from $(N = 3, m = m_{1R})$ to $(N = 4, m = m_0)$ with a corresponding objective function value of 184.5 ms.

### Problem variation

In the problem instance we have considered so far, the communication architecture depends on the resource type selected for the processing elements. For example, in Figure 4 if a software implementation is used for each of the split and sort blocks, the communication architecture can be shared memory or Gigabit Ethernet, but if one of the ends is mapped instead to an FPGA, the communication architecture is automatically set to PCI-X.

The formulation for these constraints, however, is highly nonlinear and therefore we relaxed these constraints (and simultaneously changed a number of the constants) to form a variation of the problem that we call “relaxed streaming sort.” The solver does manage to solve this relaxed version of the problem. Here, using our decomposition heuristic improves the initial solution by 480-fold as shown in Figure 13. In the figure, solution values are denoted by $s$ and solver runtimes are denoted by $t$. The subproblem $p_{4,1}$ giving the improvement is highlighted. The subproblems and their solutions are indexed using the same convention as above.

---

**Figure 10:** Decomposition of search space of streaming sort application

**Figure 11:** Objective function values (in ms) in the subproblems when optimizing application latency and throughput equally

**Figure 12:** Configuration corresponding to the best solution when optimizing application latency and throughput equally $(N = 3, m = m_{1R})$
5 Conclusions and Future Work

We have presented a heuristic for automatic application-specific performance tuning of streaming applications modeled using queueing networks by developing a domain-specific decomposition technique using topological information embodied in the queueing network models. Our preliminary empirical results show two-fold benefits—solving a problem that is currently not solvable by the state-of-the-art solvers, and for some problem instances improving the initial solution by over two orders of magnitude.

In performing a neighborhood search on the solution of the optimization problem of some problem instances for our example application, we observed that the reported solution was not even a local optimum. This is expected given that some of the functions in our optimization problem are not convex. An easy fix would be to search the neighborhood of the solution reported by the solver. More interesting would be if we can identify topological information that will guide us to partition such that the resulting optimization solution is at least a local optimum. In addition, we are pursuing a wider application set to verify the effectiveness of these ideas across a large domain of problems.

Acknowledgements

We thank the authors and user community of AMPL, FilMINT, and restricted cubic spline package for R, and You Xu of Washington University.

References
