

Maximum Flows in Parametric Graph Templates

Tal Ben-Nun¹[0000-0002-3657-6568], Lukas Gianinazzi (✉)¹[0000-0001-5975-4526],
 Torsten Hoefler¹[0000-0002-1333-9797], and Yishai Oltchik²

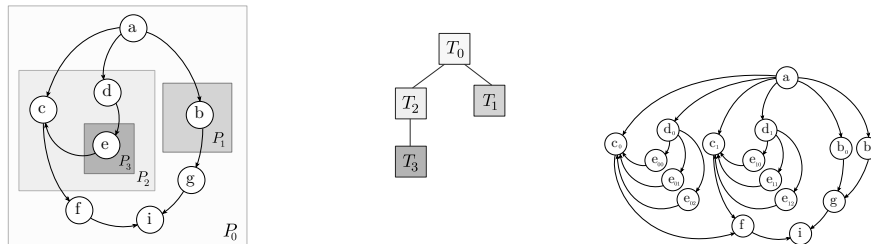
¹ ETH Zurich, Universitätstrasse 6, Zürich
 {talbn,glukas,htor}@inf.ethz.ch
² NVIDIA yoltchik@nvidia.com

Abstract. Execution graphs of parallel loop programs exhibit a nested, repeating structure. We show how such graphs that are the result of nested repetition can be represented by succinct parametric structures. This *parametric graph template* representation allows us to reason about the execution graph of a parallel program at a cost that only depends on the program size. We develop structurally-parametric polynomial-time algorithm variants of maximum flows. When the graph models a parallel loop program, the maximum flow provides a bound on the data movement during an execution of the program. By reasoning about the structure of the repeating subgraphs, we avoid explicit construction of the instantiation (e.g., the execution graph), potentially saving an exponential amount of memory and computation. Hence, our approach enables graph-based dataflow analysis in previously intractable settings.

Keywords: Graph algorithms · Graph theory · Maximum flow

1 Introduction

Parallel program analysis approaches to optimize data movement and program transformation commonly rely on graph algorithms [8, 24, 25, 29]. These problems concern an execution graph, where vertices model computation and edges



(a) Param. Graph Template \mathcal{G} . (b) Template tree of \mathcal{G} . (c) Instantiation of \mathcal{G} .

Fig. 1: Illustration of the parametric graph template \mathcal{G} with templates $T_0 = \{a, b, c, d, e, f, g, i\}$, $T_1 = \{b\}$, $T_2 = \{c, d, e\}$, $T_3 = \{e\}$; parameters $P_0 = 1$, $P_1 = 2$, $P_2 = 2$, $P_3 = 3$; and $h = 2$.

model data movement. Maximum flows provide an algorithmic measure of the overall data movement. Such execution graphs contain highly repetitive substructures. Other application areas also face repeating graph structures, for example, computational biology [19] and network topology [6, 23, 30].

The naive approach is to directly work on the graphs and apply classic algorithms. However, this is prohibitively slow. For example, execution graphs can have billions of vertices or have a parametric size. Another approach is to design domain-specific representations and solutions [17, 31]. Having a more general-purpose framework would allow sharing progress across domains.

We observe that many application-relevant graphs follow a model of *nested repetition*, where a small *template graph* is repeated a *parametric* number of times [5]. In this work, we propose a representation of such hierarchically repeating graphs, which we call *parametric graph templates*, and provide algorithms for extensions to the classical graph problem of maximum *s-t* flow.

The main challenge lies in avoiding the naive solution of materializing the full graph (which we call *instantiation*) and using a classic algorithm, which would negate any time savings. Instead, we carefully study the structural relationship between the template and the potentially exponentially larger instantiated graph. We discover and exploit symmetries in the instantiation process. This allows us to answer graph problems with a runtime that only depends on the size of the succinct representation, enabling asymptotic time and space savings compared to a naive approach that explicitly performs the nested repetition.

1.1 Parametric Graph Templates

Next, we introduce our model and give some examples. Our goal is to represent graphs with a hierarchically repeating structure, where the number of repetitions depends on some parameters. This will allow us to represent parallel loop programs and their executions. A parametric graph template with k parameters $\mathcal{G} = (G, \mathcal{T}, \mathcal{P})$ contains a (potentially weighted) and directed *template graph* $G = (V, E)$ with n vertices V , m edges E and edge weights $w : E \mapsto \mathbb{R}$, a list of *templates* $\mathcal{T} = T_0, T_2, \dots, T_{k-1}$, each with $\emptyset \neq T_i \subseteq V$, and a list of positive integer *parameters* $\mathcal{P} = P_0, \dots, P_{k-1}$ (see Figure 1a). The templates follow a nested structure, meaning that for every pair of templates they are either disjoint or one of them is strictly contained in the other one (for all $i \neq j$, $T_i \cap T_j = \emptyset$ or $T_i \subset T_j$ or $T_j \subset T_i$). In particular, the templates form a laminar set family [7].

We assume that there is a *root template* $T_0 = V$. Hence, the subset relation on the templates induces a *template tree* (see Figure 1b). We denote its height by h . If a template T is contained in another template T' (i.e., $T \subset T'$), then T is a descendant of T' (and T' is an ancestor of T). A template T is a parent of T' (and T' is a child of T) if T' is the direct descendant of T .

To create an *instantiation* of a parametric graph template \mathcal{G} , repeatedly rewrite it as follows (see Figure 1c). As long as there is more than one template, pick a leaf template T_i . For each vertex v in T_i create P_i copies v_1, \dots, v_{P_i} called instances of v , replacing v in V . The set of vertices with the same subscript are called an *instance* of T_i . For each edge $e = (u, v)$ with both endpoints in T_i , create

P_i instances $e_1 = (u_1, v_1), \dots, e_{P_i} = (u_{P_i}, v_{P_i})$, replacing e in E . For each edge $e = (u, v)$ with one endpoint u in T_i , create P_i instances $e_1 = (u_1, v), \dots, e_{P_i} = (u_{P_i}, v)$, replacing e in E . Proceed symmetrically for each edge $e = (u, v)$ with one endpoint v in T_i . Then, remove the template T_i and its parameter P_i .

In Section 3 we will represent nested parallel loop programs as parametric graph templates. In such a representation, cuts and flows correspond to data movement in a parallel execution.

1.2 Related Work

Graph Grammars [4, 10, 14, 15, 27] describe a (possibly infinite) language of graphs compactly with a set of construction rules. There is a wide variety of such ways of constructing a graph, differing in expressive power. A classic problem for graph grammars is to decide whether a graph can be constructed from a given grammar (parsing). In contrast to graph grammars, we are not primarily concerned with expressing an infinite set of graphs, but instead with a succinct representation of a graph and algorithmic aspects of solving graph problems efficiently on this succinct representation.

Hierarchical Graphs [13] model graphs where edges expand to other, possibly hierarchical graphs. They are a variant of context-free hyperedge replacement grammars that incorporate a notion of hierarchy. The authors consider graph transformations (i.e., replacing subgraphs within other subgraphs). However, their method does not include parametric replication. This makes it unsuitable for modeling variably-sized execution graphs.

Nested graphs [28] allow “hypernodes” to represent other nested graphs. The authors focus on the case where a node represents a *fixed* nested graph. This precludes nested graphs from effectively representing graphs of a parametric size.

Edge-Weight Parametric Problems Several graph problems have been generalized to the edge-weight parametric case, where edge weights are functions of one or several parameters μ_i . This includes maximum s - t flow / minimum s - t cut [3, 18, 20], (global) minimum cut [2, 21] and shortest paths [16, 22]. The solution is then a piecewise characterization of the solution space. Usually, only linear (or otherwise heavily restricted) dependency of the edge weights on the parameters have been solved. For *edge-weight parametric minimum s - t cuts*, the problem can be solved in polynomial time when each edge e has weight $\min(c(e), \mu)$ for constants $c(e)$ and a single parameter μ [3]. Granot, McCormick, Queyranne, and Tardella explore other tractability conditions [20].

1.3 Problem Statement

We approach parametric graph templates from an algorithmic perspective. The goal is to solve classical graph problems for *fixed parameters*, but in time that is strongly polynomial in the size of the parametric graph template. We focus on the classic problem of maximum s - t flow, which has an interpretation in terms of data movement for program-derived graphs and operations research [1]. For an

execution graph, a maximum s - t flow corresponds to an upper bound on the data movement between vertices s and t when they are placed on different processors.

An s - t flow f assigns every edge e a nonnegative real flow $f(e) \leq w(e)$. The sum $\sum_{e=(u,v)} f(e) - \sum_{e=(v,w)} f(e)$ is the *net flow* of the vertex v . A flow has to have net flow 0 for all vertices except s and t . The value of the flow is the net flow of the source. A maximum flow is a flow of maximum value.

The maximum s - t flow problem has a natural generalization to parametric graph templates when s and t are vertices in the root template: Instantiate the graph and compute a maximum flow between the only instance of s and the only instance of t . There are multiple possibilities for how to interpret the case when s and t have multiple instances. One interpretation is as a multiple-source and multiple-target flow problem, where all instances of s are treated as sources and all instances of t as sinks. We call this a *maximum all- s - t flow*. Another interpretation considers the maximum flow between a fixed instance of s and a fixed instance of t . We call this a *maximum single- s - t flow*.

1.4 Our Results

We show how to efficiently represent a class of parallel loop programs as parametric graph templates and how properties of data movement in the parallel loop programs relates to cuts and flows in the parametric graph templates.

Then, we demonstrate that maximum s - t flow can be solved asymptotically faster than instantiating the parametric graph template. In particular, it is possible to obtain a runtime that is similar to the runtime on the template graph.

For maximum all- s - t flow, our algorithms match the runtime of a maximum s - t algorithm such as Orlin's $O(mn)$ time algorithm [26]. We solve this problem using a technique called *Edge Reweighting*. It observes that scaling the edge weights in the graph template solves the problem. For maximum single- s - t flow and minimum cuts, there is an overhead proportional to the height h of the template tree. In addition to Edge Reweighting, we use a technique called *Partial Instantiation*. We observe that a carefully chosen part of the instantiated graph can give sufficient information to extrapolate the result to the rest of the graph. How this part is chosen depends on the problem.

2 Preliminaries

We proceed to introduce definitions, notation, and assumptions that we use throughout this work.

Template a vertex belongs to. If a vertex v is in a template T_i and v is in no other template that is a descendant of T_i , then v *belongs to* T_i . We denote the unique template that v belongs to by $T(v)$.

Template an edge belongs to. If both endpoints of an edge belong to a template T_i , then this edge belongs to template T_i . We denote the number of vertices and edges that belong to a template T_i by n_i and m_i , respectively.

Cross-template edges. An edge (u, v) where u and v belong to different templates is *cross-template*.

No Jumping. We assume there are no edges that ‘jump’ layers in the template hierarchy. Specifically, if (u, v) is a cross-template edge, then $T(u)$ is a parent or child of $T(v)$. This rule ensures that a path in the graph corresponds to a walk in the template tree. It comes without loss of generality for cut and flow problems, as an edge that jumps layers can be split into multiple edges (all of weight ∞ except the edge connected to the vertex that belongs to the deeper template in the template tree). For graphs that model programs, this assumption corresponds to disallowing jumps to arbitrary program locations.

Boundary Vertices. Consider a vertex u and v where $T(v)$ is a parent of $T(u)$. If there is an edge from u to v or from v to u in the template graph, then v is a *boundary vertex* of $T(u)$.

Template graph of a template. The subgraph of the template graph G induced by a template T_i is called the *template graph of T_i* .

Instance tree. We extend the nomenclature of templates to instances. The template hierarchy can be transferred onto the instances, where an instance I is a descendant of an instance I' if the template T that instantiated I is a descendant of the template T' that instantiated I' . Similarly, we extend the notions of ancestor, parent, and child to the instances, creating an *instance tree*. Two instances that have the same parent instance are *siblings*. If a vertex v is contained in an instance I , but it is not contained in any other descendant of I , the vertex v belongs to the instance I . If b_i is an instance of a boundary vertex b of a template T , then b_i is a boundary vertex of the instance that b_i belongs to. The instance of the root template is the *root instance*. For a vertex v in the instantiation, we write $T(v)$ for the template of the instance that v belongs to.

Isomorphism. Two parametric graph templates \mathcal{G}_1 and \mathcal{G}_2 are *isomorphic* if they instantiate isomorphic graphs. Two isomorphic \mathcal{G}_1 and \mathcal{G}_2 can have different parameters, templates, and their template graphs need not be isomorphic.

Cycles. Acyclic graphs are easier to handle for many algorithmic problems. In parametric graph templates, we consider two different notions of what constitutes a cycle. The simplest notion of cycles comes from considering cycles in the template graph. If it does not contain any cycles, then the instantiation does neither (and vice versa). A path p_1, \dots, p_k in the template graph that contains three vertices p_i, p_j, p_k with $i < j < k$ and $T(i) = T(k)$ but $T(i) \neq T(j)$ is a *template-cycle*. We say a parametric graph template is *template-acyclic* if it does not contain a template-cycle. This notion is incomparable to the notion of acyclic parametric graph templates. There are acyclic parametric graph templates that are not template-acyclic (consider a path whose nodes alternate between belonging to some template and its child). Note that a template-acyclic graph can have cycles (consider a cycle whose vertices belong to the same template).

3 Templates of Parallel Loop Programs

We show that a broad class of parallel programs can be modeled as parametric graph templates, such that the parametric graph template corresponds to the source code of the program and an instantiation of the parametric graph template corresponds to an execution of the program. This allows us to analyze properties of the execution of a program by considering a parametric graph template of a size comparable to the source of the program.

The parametric graph templates we consider can model nested loop programs, for example Projective Nested Loops [12] and Simple Overlap Access Programs [24]. The program receives a set of multi-dimensional input arrays A_1, \dots, A_k . The goal is to output a multi-dimensional array B . The program can use several multi-dimensional temporary arrays $C_1, \dots, C_{k'}$. For any array D , its size in the i -th dimension is $\text{size}_i(D)$.

Roughly speaking, we allow any composition of elementary operations and parallel nested loops where the loop bounds only depend on the sizes of the input arrays. We allow parallel reduction to aggregate the results of a loop. We do not allow data-dependent control flow, but we allow the locations of memory accesses to be data-dependent. Examples of algorithms that can be represented this way include matrix multiplication, convolution, and cross-correlation. See Figure 2a for a parametric graph template of matrix multiplication and Figure 2b for a parametric graph template of cross-correlation.

We call the resulting parametric graph templates *parallel loop graph templates*. Next, we describe their syntax. Then, we describe a semantic for these parametric graph templates. Finally, we relate the data movement of the parallel loop programs with their templates' instantiations.

3.1 Syntax

The vertices of the template graph are annotated with types corresponding to their function in the program. Each template graph contains the *input memory vertices* A_1, \dots, A_k , the *output memory vertex* B , and the *temporary memory vertices* $C_1, \dots, C_{k'}$. The memory vertices can have arbitrary in-degree and out-degree and belong to the root template. Other vertices have out-degree 1, except if stated otherwise. The outgoing edge is called the *output edge*. To disambiguate the inputs to a vertex, the incoming edges are numbered consecutively. We refer to inputs in this *input order*. We consider the following control flow constructs. These are boundary vertices.

Parfor. A PARFOR (parallel for loop) vertex has no input edge. Its output edge leads to a vertex in a child template.

Reduce(Op), where OP is an associative and commutative operator. Has a single input edge from a vertex in a child template. The output edge leads to a non-memory vertex.

Copy. A COPY vertex v has arbitrary outdegree. For each of its output edges (v, u) , the template $T(u)$ is not a parent of $T(v)$ and u is not a memory vertex.

We consider the following memory constructs, which are boundary vertices.

‘ \rightarrow ’. A (pass-through) \rightarrow vertex has in-degree and out-degree 1. At most one of the two neighbors can be a memory vertex.

Read. A Read vertex has a first input edge from a memory vertex or a \rightarrow vertex and one or more other input edges from a non-memory vertex. Its output edge leads to a non-memory vertex.

Write. A Write vertex has two or more input edges from non-memory vertices. Its output edge leads to a memory or \rightarrow vertex.

We consider the following types of operator vertices. They cannot be connected to memory vertices and are not boundary vertices.

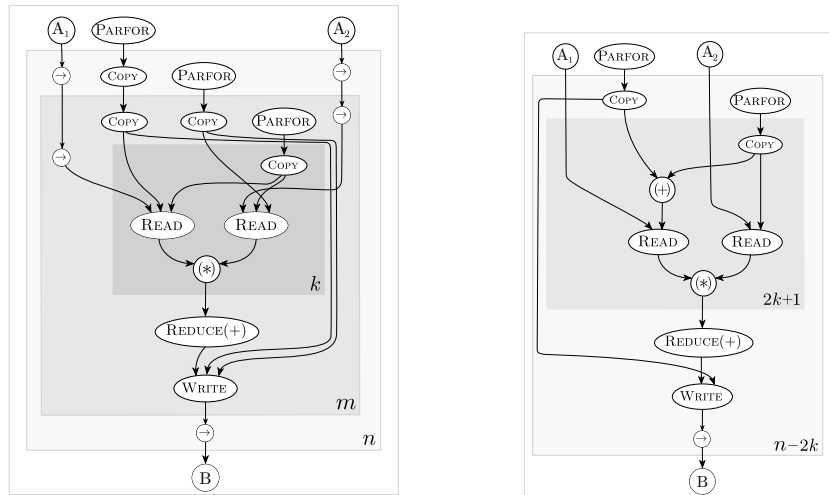
(Op), for $OP \in \{+, -, *, \div\}$, which has in-degree 2.

[c], for any representable constant c .

3.2 Semantics

A *well-formed* program has an acyclic template graph. For a well-formed program, a *serial execution* is any topological order of the instantiation of the parametric graph template. Each d -dimensional input array A_i initially contains some current value $A_i[j_1] \dots [j_d]$ at each position (j_1, \dots, j_d) , where the d -th dimension of A_i has size $\text{size}_d(A_i)$. All other arrays contains 0 at each of their positions. The arrays do not alias each other.

The semantics of a serial execution is given by applying the following rules to each vertex in the serial execution. Before evaluating the rules, contract all



(a) Multiplication of an $n \times k$ matrix A_1 and a $k \times m$ matrix A_2 .

(b) 1-D cross-correlation of a size n array A_1 and a size k array A_2 .

Fig. 2: Example parametric graph templates for execution graphs. The edges are drawn in their input order left-to-right.

edges which have at least one \rightarrow vertex neighbor (these exist to transfer values from inside the template hierarchy to the memory vertices in the root template).

Parfor. A PARFOR vertex with k children outputs one value of the permutation of $\{0, \dots, k-1\}$ to each child in an injective way.

Copy. Given input x , COPY outputs x to all its children.

Reduce(Op), for $OP \in \{+, *\}$. Given the inputs x_1, \dots, x_k , REDUCE(OP) outputs the result of applying OP repeatedly in an arbitrary order to the inputs.

Read. Given inputs A_i, j_1, \dots, j_d , if A_i has dimension d and for all $j_k \in \{j_1, \dots, j_d\}$ we have $0 \leq j_k < \text{size}_k(A_i)$, a READ vertex outputs the current value of $A_i[j_1] \dots [j_d]$. Otherwise, the result of the serial execution is undefined.

Write. Given inputs x, j_1, \dots, j_d , a WRITE vertex outputs x . This has the side effect of updating the current value of the array into which the output edge leads: Say it leads to A_i . Then, if A_i has dimension d and for all $j_k \in \{j_1, \dots, j_d\}$ we have $0 \leq j_k < \text{size}_k(A_i)$, the current value of $A_i[j_1] \dots [j_d]$ becomes x . Otherwise, the result of the serial execution is undefined.

(Op), for $OP \in \{+, -, *, \div\}$. Given x, y , outputs $x \text{ OP } y$.
[c] outputs the constant c .

In a serial execution, we say that two reads or writes u, v are *totally ordered* if there is a path from u to v or from v to u in the instantiation. A *data race* occurs if there are two writes W_1, W_2 with the same right input (i.e., index) that connect to the same memory vertex and W_1 and W_2 are not totally ordered. The output of a well-formed program is *well-defined* if none of its serial executions has an undefined result or contains a data race.

3.3 Applications of Flows and Cuts

To model dataflow in the parallel loop programs, we set the weight of the PARFOR edges to 0 and the weight of the other edges to 1. Loop indices can be recomputed and thus do not cause data movement. The parallel loop graph template encodes all the data movement in its edges. However, it cannot resolve the aliasing of array locations. Hence, the weight of the edges going across a partition of the vertices provides an *upperbound* on the data movement:

Observation 1 *Consider a partition (V_0, \dots, V_p) of the vertices in an instantiation of a parallel loop graph template. The value total weight of the edges with endpoints in different partitions is an upper bound on the data movement incurred when the partitions are allocated to distinct processors.*

Note that in our formulation of parallel loop graph templates, vertices corresponding to arrays are placed on a single processor. Thus, to model the distribution of an array across multiple processors, a vertex must be created for each processor that holds its subarray (this subarray can be discontinuous though).

Since a maximum s - t flow equals the value of a minimum s - t cut, the maxflow provides a partition of the loop program with small data movement:

Observation 2 *If a parallel loop graph template has a maximum all- s - t flow of value x , then there is a partition of the parallel loop program which incurs at*

most x data movement and in which all instances of s are executed on a different processor as all instances of t .

We can get a similar statement for maximum single- $s-t$ flows.

4 Template Maximum Flows

Next, we turn to the first algorithmic question on parametric graph templates. Our goal here is to solve the maximum $s-t$ flows problem on a parametric graph template *without explicitly instantiating it*. Instead, the goal is to get a runtime that is polynomial in the size of the graph template. Our algorithms use a series of observations on the structure of maximum flows in parametric graphs which allow us to produce transformed parametric graph templates, on which the answer can be efficiently computed.

We will approach the problem by considering the case where s and t are in the root template first. Then, we show how to reduce both the maximum all- $s-t$ flow and the maximum single- $s-t$ flow problem to an instance of this simpler problem. Throughout, we assume that all vertices are reachable from s and can reach t , as otherwise they cannot carry flow.

In the template-acyclic case, the maximum single- $s-t$ flow is trivially zero *except* when s and t are in the same instance of the least common ancestor of $T(s)$ and $T(t)$ in the template tree. Therefore, in the acyclic case it makes sense to restrict our attention to this case where the flow is not trivially zero. In the case where there are template-cycles, it matters which instances of s and t are picked. These can be identified by numbering the instances they belong to.

4.1 Edge-Reweighting

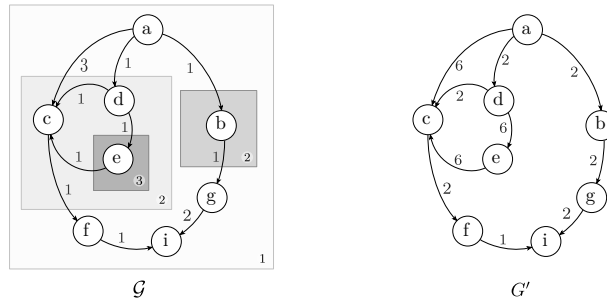


Fig. 3: Edge Reweighting turns a parametric graph template \mathcal{G} into a graph G' by scaling the weights of each edge in its template graph by all the parameters of the templates that contain an endpoint of the edge.

An efficient way to solve a problem on parametric graph templates is to show how it relates to a problem on the template graph with *scaled weights*. The idea

is that an edge that intersects template T_i can be used P_i times and can therefore be used to carry P_i times the amount of flow. We will see that this observation holds as long as s and t are in the root template or if we consider the maximum all- s - t flow problem. We call this approach *Edge-Reweighting*. See Figure 3 for an example of Edge Reweighting.

Algorithm: Edge-Reweighting. Transform the parametric graph template $\mathcal{G} = (G, \mathcal{T}, \mathcal{P})$ with edge weights w into a graph G' with edge weights w' . The *reweighted graph* G' has the same vertex and edge set as the template graph G , but the weights are scaled as follows: Multiply the weight of an edge in the template graph by the product of the parameters of the templates that contain at least one endpoint of the edge. That is, let $I(e)$ be the index set of all templates that contain at least one of the endpoints of e . Then, the weight of $w'(e)$ is $w(e) \prod_{i \in I(e)} P_i$. To implement this in linear time $O(m)$, precompute in a pre-order traversal of the template tree for each template the product of all the ancestors' parameters.

4.2 Source & Sink Belong to the Root Template

Our goal is to show that when the source s and sink t belong to the root template, then a maximum s - t flow in the reweighted graph equals the value of a maximum all- s - t flow. If s and t belong to the root template (which is instantiated once), then a maximum single- s - t flow equals a maximum all- s - t flow and we call it a maximum s - t flow for short.

The linear programming *dual* of a maximum s - t flow is a *minimum s - t cut* [11]. We will use strong duality [9] in our proof, which means that it suffices to identify an s - t flow and a minimum s - t cut of equal value to prove that they are optimal. We argue that Edge Reweighting preserves the value of the dual minimum all- s - t cut. Hence, it also preserves the maximum all- s - t flow value.

The following shows us how to construct an s - t cut C' in the transformed graph G' from an s - t cut C in \mathcal{G} of the same value. Together with the other (easier) direction of the proof, this shows that the transformed graph G' has the same maximum s - t flow.

Lemma 1. *In a parametric graph template $\mathcal{G} = (G, \mathcal{T}, \mathcal{P})$, if s and t are in the root template, there is a minimum s - t cut of the instantiation of \mathcal{G} where every instance of every vertex is on the same side of the cut.*

Proof. The proof is by induction on the number of templates in the parametric graph template. If the parametric graph template has only a single template, then (since s and t must be in this template) the claim is trivial because the root is repeated only once, by assumption.

Otherwise, let $C = (V_s, V_t)$ be a minimum all- s - t cut of the parametric graph template \mathcal{G} (i.e., V_s contains the vertices assigned to s and V_t those assigned to t). Consider an arbitrary template T_i that is a *child of the root template* and its graph template G_i . The sets B_s and B_t contain the boundary vertices of T_i that are in V_s and V_t , respectively.

If either of the sets B_s or B_t is empty, then it follows immediately that all instances of the vertices that are in T_i are in the same part of the cut C (i.e., on the side of s if the set B_t is empty and vice versa).

Otherwise, merging all vertices in B_s into a vertex s' and merging all vertices in B_t into a vertex t' does not change the value of the minimum s - t cut in \mathcal{G} . Moreover, if the merged parametric graph template has a minimum s - t cut that puts every instance of every vertex on the same side of the cut, then so does the original parametric graph template (because B_s and B_t contain only vertices that belong to the root template and we can “undo” the merging). We thus further assume w.l.o.g. that B_s and B_t contain a single vertex named s' and t' , respectively. Note that since these vertices belong to the root template, the vertices s' and t' coincide with their only instances.

Every minimum s - t cut must separate s' from t' in the subgraph H given by the instances of T_i and the vertices s' and t' (but without a potential edge from s' and t'). We use our induction hypothesis to show there is a minimum s' - t' cut in this subgraph that puts all instances of a vertex on the same side of the cut.

We construct a parametric graph template \mathcal{G}'' such that a maximum s' - t' flow in \mathcal{G}'' can be extended to a flow for the graph H . The parametric graph template \mathcal{G}'' has the following template graph: take the subgraph of G induced by T_i together with its boundary vertices, then delete any edges going between s' and t' . The boundary vertices s' and t' and all vertices that belong to T_i are put into the root template of \mathcal{G}'' (which has parameter 1). Moreover, \mathcal{G}'' has the templates and parameters of the descendants of T_i in G . The parametric graph template \mathcal{G}'' contains at least one template less than G . Hence, by induction, there is a minimum s' - t' cut C'' of \mathcal{G}'' that puts all instances of the same vertex into the same partition.

Let f'' be the dual maximum s' - t' flow corresponding to C'' in \mathcal{G}'' of value μ . Now, we construct a s' - t' flow f in H and show it is maximum. Along each instance of each edge e that intersects T_i we send $f''(e)$ flow. The capacity constraint on the flow is trivially satisfied. The conservation constraint on the flow is satisfied because in the instantiated graph, the total flow going in and out of an instance of v is the same as for vertex v for f'' in \mathcal{G}'' . The value of the flow f is $P_i \cdot \mu$. Now, consider the cut C' where we put every instance of a vertex v in T_i on the same side as v is in C'' . The value of this cut is $P_i \cdot \mu$. By strong duality, this shows that C' is a minimum s' - t' cut in the graph H . By construction, this cut puts every instance of every vertex on the same side of the cut. We conclude that all children of the root template can be cut such that every instance of the same vertex is in the same part of the cut. Because the root has a single instance, the statement follows for the root as well. \square

Lemma 2. *If a parametric graph template \mathcal{G} has a minimum s - t cut of value μ and s and t are in the root template, then the graph G' constructed by edge reweighting has a minimum s - t cut of value μ .*

Proof. Any cut in the reweighted graph G' corresponds to a cut of the same value in the parametric graph template $\mathcal{G} = (G, \mathcal{T}, \mathcal{P})$: Put every instance of a

vertex into the partition that it has in the cut of G' . Since every edge e is cut exactly $\prod_{i \in I(e)} P_i$ times, this shows that the value of the minimum s - t cut of the graph G' is at least the value of the minimum all- s - t cut of the parametric graph template $(G, \mathcal{T}, \mathcal{P})$. It remains to show that the minimum s - t cut of the reweighted graph G' is at most the value of the minimum s - t cut of the parametric graph template \mathcal{G} . By Lemma 1, there is a minimum s - t cut C of \mathcal{G} that puts every instance of every vertex on the same side of the cut. Now, we construct a cut C' of G' from this cut C by putting every vertex v in G' on the same side of the cut as all the instances of v are in C . The cut C' has the same value μ because every instance of an edge e in \mathcal{G} is crossing $\prod_{i \in I(e)} P_i$ times, which is the amount by which we scaled the weight of edge e in G' . \square

4.3 Instance Merging

We show how to merge all instances of a vertex v in a parametric graph template by transforming it into parametric graph template of almost the same size (the overhead is an additive $O(nh)$). We will use this technique to reduce the general case for maximum all- s - t flow to the case where s and t are in the root.

The idea is that merging all instances of a vertex s is akin to moving the vertex from the template $T(s)$ it belongs to into the root template (so that it belongs to the root template). The *no jumping rule* only allows edges to go from parent templates to children templates (or vice versa), we need to introduce *dummy edges* and *dummy vertices* along the way. The dummy edges have ∞ weight. An original edge (u, s) will be transformed into a path u, d_1, \dots, d_k, s for dummies d_1, \dots, d_k (symmetrically for an edge (s, u)). See Figure 4 for an example.

Algorithm: Instance-Merging. Given a parametric graph template \mathcal{G} and a vertex s , repeat the following until s is in the root template:

1. For any cross-template edge (u, s) , introduce a dummy vertex d in the template $T(s)$ that s belongs to. Replace the edge (u, s) by two edges $e_1 = (u, d)$ and $e_2 = (d, s)$. The weight of the edge e_1 is the same as the weight of the edge e , but the weight of the edge e_2 is set to ∞ . Proceed symmetrically for any cross-template edge (s, u) .
2. Move the vertex s from the template $T(s)$ to the parent of the template $T(s)$ (i.e., remove s from the set $T(s)$).

Observation 3 *Instance Merging* (\mathcal{G}, s) produces a parametric graph template \mathcal{G}' whose instantiation is, after merging all instances of dummy edges of weight ∞ , isomorphic to the graph that we get by instantiating the original parametric graph template \mathcal{G} and merging all instances of s . *Instance Merging* (\mathcal{G}, s) adds at most $d(s) \cdot h$ vertices and edges, where $d(s)$ is the degree of the vertex s in the template graph.

Proof. In the template graph of \mathcal{G}' , there is a path consisting of ∞ -weight edges from s to every neighbor of s in the template graph G of \mathcal{G} . There are no other

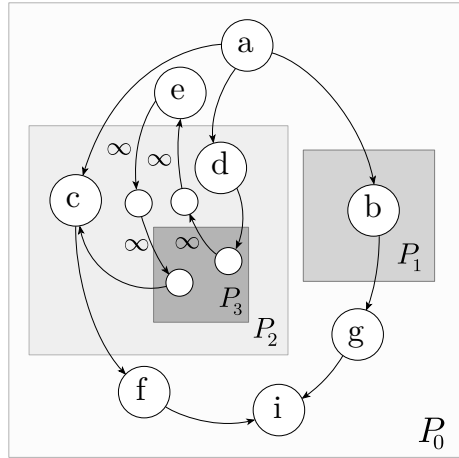


Fig. 4: Instance Merging on the graph \mathcal{G} from Figure 1a with vertex e pushes the vertex e into the root template. This introduces dummy nodes (drawn without labels) and ∞ -weight dummy edges. After contracting all dummy edges, the instantiation of the transformed parametric graph template is isomorphic to the graph we get by merging all instances of e in the instantiation of \mathcal{G} .

∞ weight edges. Hence, there also is such an ∞ -weight path in the instantiation of \mathcal{G}' to every instance of every vertex that is a neighbor of s in G . Contracting these paths gives a graph where the vertex s has an edge to all instances of neighbors of s in G , which is the same graph that we get by instantiating the original parametric graph template \mathcal{G} and merging all instances of s . \square

4.4 Maximum All- s - t Flow

To solve maximum all s - t Flow, all we would need to do is use Instance Merging on s and then on t to ensure that they are both in the root template. Then, we could use the edge reweighing Lemma 2. This approach would cost $O(nm + n^2h)$ time. We can avoid this overhead by observing that edge reweighing works directly for maximum all- s - t Flow (even when s and t are not in the root template).

Lemma 3. *Edge reweighing of a parametric graph template \mathcal{G} produces a reweighted graph G' where the value of the maximum s - t flow of G' equals the value of the maximum all- s - t flow of \mathcal{G} .*

Proof. Instance Merge s and then t in \mathcal{G} to produce a parametric graph template \mathcal{G}' . By definition, all instances of s (and t respectively) must be on the same side of a minimum all- s - t cut, this parametric graph template \mathcal{G}' has the same minimum all- s - t cut value as the original parametric graph template \mathcal{G} . Edge reweighing \mathcal{G}' gives us a graph \hat{G} . From Lemma 2 we know that a minimum s - t

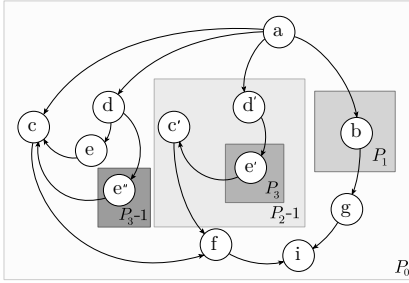


Fig. 5: After running Upwards Partial Instantiation from e on the parametric graph template \mathcal{G} from Figure 1a, the vertex e is in the root template. The transformed parametric graph template instantiates the same graph.

cut of \hat{G} corresponds to the minimum all- s - t cut of \mathcal{G}' (which puts all instances of the same vertex on the same side of the cut).

An ∞ -weight edge never crosses a minimum s - t cut and therefore such dummy edges (introduced by the instance merging) from \hat{G} can be contracted, yielding a graph G' . This graph G' is the same graph that we get from edge reweighting the original parametric graph template \mathcal{G} . \square

Now, the results follows:

Theorem 1. *Computing a maximum all- s - t flow in parametric graph template takes $O(mn)$ time.*

4.5 Partial Instantiation

The technique of *partial instantiation* revolves around instantiating only part of the parametric graph template, depending on the problem at hand. The goal is to choose the partial instantiation such that the remaining problem is solvable by using the symmetry of the problem (e.g., using edge-reweighting). Partial instantiation can be seen as an example of the more general technique of *retemplating*. The intuition of retemplating is that in certain cases, it suffices to change the representation of the parametric graph template into another isomorphic parametric graph template to significantly simplify the problem at hand.

Next, show how to move a single vertex s from deep in the template tree to the root, without changing the instantiated graph. This solves the maximum s - t problems when s (or t) belongs to a template that is deep in the template tree (See Section 4.6). We call this technique *Upwards Partial Instantiation* from s . For simplicity, let us start with the special case of template-acyclic graphs.

In a template-acyclic parametric graph template, once a path goes from an instance of a template T_i to its parent, it never enters another instance of T_i again. This property implies that, when considering the reachable subgraph from a vertex that is an instance of s , we can simply “merge” $T(s)$ and all the templates that are ancestors of the template $T(s)$ in the template tree. Formally,

this corresponds to deleting $T(s)$ and all the templates that are ancestors of $T(s)$ (except the root) from the parametric graph templates' list of templates.

If the parametric graph template has template-cycles, our goal remains to transform the parametric graph template into an equivalent graph where a particular instance of a vertex s is in the root template. See Figure 5 for an illustration of Upwards Partial Instantiation.

Algorithm: Partial Instantiation. Repeat the following until all templates from $T(s)$ to the root have parameter 1:

1. Consider the topmost template T that contains s and has parameter greater than 1. Let P_s be the number of instances of the template T .
2. Instantiate the template T twice. Create a new parametric graph template that has the two instances as templates, where the first template has parameter 1 and the second template has parameter $P_s - 1$. The vertices in the second template are relabeled (s is in the one with parameter 1).

Now, merge $T(s)$ and all the templates that are ancestors of $T(s)$, leaving s in the root template.

Because this process performs the same rewriting of the parametric graph template as instantiation, just in a different order and stopping early, this process creates an isomorphic parametric graph template. Every iteration adds at most n vertices and m edges and there are at most h iterations. We conclude that:

Observation 4 *Upwards Partial Instantiation from s produces an isomorphic parametric graph template with at most h additional templates and $O(nh)$ vertices and $O(mh)$ edges in the template graph.*

4.6 Maximum Single- s - t Flow

We give a partial instantiation and edge reweighting approach to maximum single- s - t flow. For there to be a flow through some instance, it must lie along an s - t path. Hence, we can use Upwards Partial Instantiation twice to ensure that s and t lie in the root template. Then, we use Edge Reweighting.

Algorithm: Single- s - t Flow. We solve maximum single- s - t flow as follows:

1. Perform upwards partial instantiation from s .
2. Perform upwards partial instantiation from t .
3. Construct an edge-reweighted graph G' .
4. Run a maximum s - t flow algorithm on the partially instantiated and reweighted graph G' .

Theorem 2. *Computing a maximum single- s - t flow in parametric graph template takes $O(mnh)$ time.*

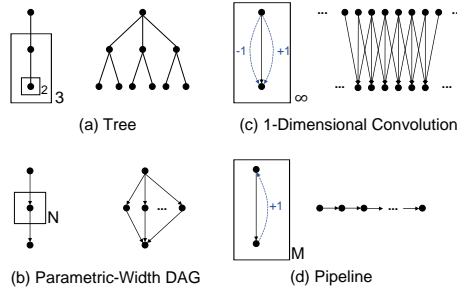


Fig. 6: Graph Templates and their expanded/instantiated counterparts. An edge e labelled with $\pm\Delta$ indicates a sibling edge e with sibling function $f_e(x) = x + \Delta$

5 Allowing Edges Between Sibling Templates

So far, we have disallowed any edges between instances of the same template. This limits the types of graphs which have a small template graph. For example, a path of length n requires a template graph with n nodes. We can extend the model by allowing an instance to have edges *to another instance of the same template*. These edges can, for example, more efficiently model sequential chains (paths), convolutional networks, and grids. We call these edges *sibling edges* (because they connect siblings). A *sibling edge* (u,v) of template T_i connects (in the template graph of T_i) a vertex u that belongs to template T_i to a vertex v that *also* belongs to T_i . Every sibling edge $e = (u,v)$ of template T_i is associated with a *bijective* (and computable) *sibling function* $f_e : \{0, \dots, P_i - 1\} \rightarrow \{0, \dots, P_i - 1\}$ which tells us that if the head of edge e is in instance j of the template T_i , then the tail of edge e is in instance $f(j)$ of the template T_i .

Note that in the model with sibling edges, a path of length n can be represented with two nodes instead of n nodes and a 1-dimensional cross-correlation of two n -dimensional signal can be represented with 2 nodes.

The structural Lemma 1 for edge reweighting still holds with sibling edges.

Lemma 4. *In a parametric graph template $\mathcal{G} = (G, \mathcal{T}, \mathcal{P})$ with sibling edges, if s and t are in the root template, there is a minimum s - t cut of the instantiation of \mathcal{G} where every instance of every vertex is on the same side of the cut.*

Proof (sketch of Lemma 1 with sibling edges). Extend the proof of Lemma 1 as follows: Proceed to construct the flow f as usual. Then, observe that the conservation constraint on the flow f is satisfied because the sibling functions are bijective: In the instantiated graph, the total flow going in and out of an instance of v is the same as for vertex v for f'' in G'' . The value of the flow f is $P_i \cdot v(f)$. \square

Hence, the results on maximum all- s - t flow and maximum single- s - t flow hold analogously in the presence of sibling edges within the same bounds.

6 Conclusion

In this work, we explored the notion of structural parameterization in graphs. We show how graph templates correspond to the computation graphs of parallel programs. Our model leads to a $O(mn)$ time algorithm for a template version of maximum s - t flow (and hence minimum s - t cuts). These flows provide upper bounds on the data movement of partitions of certain parallel loop programs.

Other interesting problems would include partitions into multiple parts and subgraph isomorphism. Moreover, future work could explore lower bounds for parametric graph template algorithms.

Acknowledgements. This work received support from the PASC project DaCeMI and from the European Research Council under the European Union’s Horizon 2020 programme (Project PSAP, No. 101002047), as well as funding from EuroHPC-JU under grant DEEP-SEA, No. 955606.

Bibliography

- [1] Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Inc., USA (1993)
- [2] Aissi, H., Mahjoub, A.R., McCormick, S.T., Queyranne, M.: Strongly polynomial bounds for multiobjective and parametric global minimum cuts in graphs and hypergraphs. *Math. Program.* **154**(1-2), 3–28 (2015). <https://doi.org/10.1007/s10107-015-0944-8>
- [3] Aneja, Y.P., Chandrasekaran, R., Nair, K.: Parametric min-cuts analysis in a network. *Discrete Applied Mathematics* **127**(3), 679–689 (2003). [https://doi.org/10.1016/S0166-218X\(02\)00496-1](https://doi.org/10.1016/S0166-218X(02)00496-1)
- [4] Bauderon, M., Courcelle, B.: Graph expressions and graph rewritings. *Math. Syst. Theory* **20**(2-3), 83–127 (1987). <https://doi.org/10.1007/BF01692060>
- [5] Ben-Nun, T., de Fine Licht, J., Ziogas, A.N., Schneider, T., Hoefler, T.: Stateful dataflow multigraphs: A data-centric model for performance portability on heterogeneous architectures. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. SC ’19*, Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3295500.3356173>
- [6] Besta, M., Hoefler, T.: Slim fly: A cost effective low-diameter network topology. In: *SC ’14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. pp. 348–359 (2014). <https://doi.org/10.1109/SC.2014.34>
- [7] Cheriyan, J., Jordán, T., Ravi, R.: On 2-coverings and 2-packings of laminar families. In: *Algorithms - ESA ’99, 7th Annual European Symposium, Prague, Czech Republic, July 16-18, 1999, Proceedings*. pp. 510–520 (1999). https://doi.org/10.1007/3-540-48481-7_44
- [8] Chu, W.W., Holloway, L.J., Lan, M., Efe, K.: Task allocation in distributed data processing. *Computer* **13**(11), 57–69 (1980). <https://doi.org/10.1109/MC.1980.1653419>

- [9] Chvátal, V.: *Linear Programming*. Series of books in the mathematical sciences, W. H. Freeman (1983)
- [10] Courcelle, B.: An axiomatic definition of context-free rewriting and its application to NLC graph grammars. In: STACS 88, 5th Annual Symposium on Theoretical Aspects of Computer Science, Bordeaux, France, February 11-13, 1988, Proceedings. pp. 237–247 (1988). <https://doi.org/10.1007/BFb0035848>
- [11] Dantzig, G.B., Fulkerson, D.R.: *On the Max Flow Min Cut Theorem of Networks*. RAND Corporation, Santa Monica, CA (1955)
- [12] Dinh, G., Demmel, J.: Communication-optimal tilings for projective nested loops with arbitrary bounds. In: Scheideler, C., Spear, M. (eds.) SPAA '20: 32nd ACM Symposium on Parallelism in Algorithms and Architectures, Virtual Event, USA, July 15-17, 2020. pp. 523–525. ACM (2020). <https://doi.org/10.1145/3350755.3400275>
- [13] Drewes, F., Hoffmann, B., Plump, D.: Hierarchical graph transformation. *J. Comput. Syst. Sci.* **64**(2), 249–283 (2002). <https://doi.org/10.1006/jcss.2001.1790>
- [14] Ehrig, H., Pfender, M., Schneider, H.J.: Graph-grammars: An algebraic approach. In: 14th Annual Symposium on Switching and Automata Theory, Iowa City, Iowa, USA, October 15-17, 1973. pp. 167–180 (1973). <https://doi.org/10.1109/SWAT.1973.11>
- [15] Engelfriet, J.: Context-free NCE graph grammars. In: *Fundamentals of Computation Theory, International Conference FCT'89*, Szeged, Hungary, August 21-25, 1989, Proceedings. pp. 148–161 (1989). <https://doi.org/10.1007/3-540-51498-8>“15
- [16] Erickson, J.: Maximum flows and parametric shortest paths in planar graphs. In: *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010*, Austin, Texas, USA, January 17-19, 2010. pp. 794–804 (2010). <https://doi.org/10.1137/1.9781611973075.65>
- [17] Feautrier, P.: Some efficient solutions to the affine scheduling problem. i. one-dimensional time. *Int. J. Parallel Program.* **21**(5), 313–347 (1992). <https://doi.org/10.1007/BF01407835>
- [18] Gallo, G., Grigoriadis, M.D., Tarjan, R.E.: A fast parametric maximum flow algorithm and applications. *SIAM J. Comput.* **18**(1), 30–55 (1989). <https://doi.org/10.1137/0218003>
- [19] Ginsburg, A., Ben-Nun, T., Asor, R., Shemesh, A., Ringel, I., Raviv, U.: Reciprocal grids: A hierarchical algorithm for computing solution x-ray scattering curves from supramolecular complexes at high resolution. *Journal of Chemical Information and Modeling* **56**(8), 1518–1527 (Aug 2016)
- [20] Granot, F., McCormick, S.T., Queyranne, M., Tardella, F.: Structural and algorithmic properties for parametric minimum cuts. *Math. Program.* **135**(1-2), 337–367 (2012). <https://doi.org/10.1007/s10107-011-0463-1>
- [21] Karger, D.R.: Enumerating parametric global minimum cuts by random interleaving. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016*, Cambridge, MA, USA, June 18-21, 2016. pp. 542–555 (2016). <https://doi.org/10.1145/2897518.2897578>

- [22] Karp, R.M., Orlin, J.B.: Parametric shortest path algorithms with an application to cyclic staffing. *Discrete Applied Mathematics* **3**(1), 37–45 (1981). [https://doi.org/10.1016/0166-218X\(81\)90026-3](https://doi.org/10.1016/0166-218X(81)90026-3)
- [23] Kim, J., Dally, W.J., Scott, S., Abts, D.: Technology-driven, highly-scalable dragonfly topology. In: 2008 International Symposium on Computer Architecture. pp. 77–88 (2008). <https://doi.org/10.1109/ISCA.2008.19>
- [24] Kwasniewski, G., Ben-Nun, T., Gianinazzi, L., Calotoiu, A., Schneider, T., Ziogas, A.N., Besta, M., Hoeffler, T.: Pebbles, graphs, and a pinch of combinatorics: Towards tight I/O lower bounds for statically analyzable programs. In: Agrawal, K., Azar, Y. (eds.) SPAA '21: 33rd ACM Symposium on Parallelism in Algorithms and Architectures, Virtual Event, USA, 6–8 July, 2021. pp. 328–339. ACM (2021). <https://doi.org/10.1145/3409964.3461796>
- [25] Lattner, C., Adve, V.: LLVM: a compilation framework for lifelong program analysis transformation. In: International Symposium on Code Generation and Optimization, 2004. CGO 2004. pp. 75–86 (2004). <https://doi.org/10.1109/CGO.2004.1281665>
- [26] Orlin, J.B.: Max flows in $o(nm)$ time, or better. In: Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1–4, 2013. pp. 765–774 (2013). <https://doi.org/10.1145/2488608.2488705>
- [27] Pavlidis, T.: Linear and context-free graph grammars. *J. ACM* **19**(1), 11–22 (1972). <https://doi.org/10.1145/321679.321682>
- [28] Poulouvasilis, A., Levene, M.: A nested-graph model for the representation and manipulation of complex objects. *ACM Trans. Inf. Syst.* **12**(1), 35–68 (1994). <https://doi.org/10.1145/174608.174610>
- [29] Shen, C., Tsai, W.: A graph matching approach to optimal task assignment in distributed computing systems using a min-max criterion. *IEEE Trans. Computers* **34**(3), 197–203 (1985). <https://doi.org/10.1109/TC.1985.1676563>
- [30] Valadarsky, A., Shahaf, G., Dinitz, M., Schapira, M.: Xpander: Towards optimal-performance datacenters. In: Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies. p. 205–219. CoNEXT '16, Association for Computing Machinery, New York, NY, USA (2016). <https://doi.org/10.1145/2999572.2999580>
- [31] Vasilache, N., Zinenko, O., Theodoridis, T., Goyal, P., DeVito, Z., Moses, W.S., Verdoolaege, S., Adams, A., Cohen, A.: Tensor comprehensions: Framework-agnostic high-performance machine learning abstractions. *CoRR* **abs/1802.04730** (2018)