Data Locality and Replica Aware Virtual Cluster Embeddings

Paolo Costa[1], Carlo Fuerst[2], Maciej Pacut[3], Stefan Schmid[4]
[1] Microsoft Research, UK; [2] TU Berlin, Germany; [3] University of Wroclaw, Poland; [4] TU Berlin & T-Labs, Germany

ABSTRACT

Virtualized datacenters offer great flexibilities in terms of resource allocation. In particular, by decoupling applications from the constraints of the underlying infrastructure, virtualization supports an optimized mapping of virtual machines as well as their interconnecting network to their physical counterparts: essentially a graph embedding problem.

However, existing algorithms often ignore a crucial dimension of the embedding problem, namely *data locality*: the input to a cloud application such as MapReduce is typically stored in a distributed, and sometimes redundant, file system. Since moving data is costly, an embedding algorithm should be data locality aware, and allocate computational resources close to the data; in case of redundant storage, the algorithm should also optimize the *replica selection*.

This paper initiates the algorithmic study of data locality aware virtual cluster embeddings on datacenter topologies. We show that despite the multiple degrees of freedom in terms of embedding, replica selection and assignment, many problems can be solved efficiently. We also highlight the limitations of such optimizations, by presenting several NP-hardness proofs; interestingly, our hardness results also hold in uncapacitated networks of small diameter.

1. INTRODUCTION

Distributed cloud applications, such as batch-processing applications or scale-out databases, generate a significant amount of network traffic [25]. For instance, MapReduce consists of a network intensive shuffle phase, where data is transferred from the mappers to the reducers. In order to ensure a predictable application performance, especially in shared cloud environments, it is important to provide isolation and bandwidth guarantees between the virtual machines [36], e.g., by making explicit network reservations [5]. Accordingly, modern batch-processing applications provide the abstraction of entire *virtual networks* [25], defining both the virtual machines as well as their interconnecting network. The most prominent virtual network abstraction is the *virtual cluster* [5, 33].

Virtualized datacenters offer great flexibilities on where these virtual networks can be instantiated or *embedded*. In order to maximize the resource utilization in the datacenter, it is in principle desirable to map the virtual machines of a given virtual network as close as possible in the underlying physical network, as this minimizes communication costs (respectively, bandwidth reservations) [5, 33].

However, existing systems often ignore a crucial dimension of the virtual network embedding problem: the fact that the input data for a cloud application, consisting of atomic *chunks*, is typically distributed across different servers and stored in a distributed file system [6, 17, 31]. In order to properly minimize communication costs, an embedding algorithm should hence also be *data locality aware* [4, 23, 35], and allocate (or *embed*) computational resources close to the to be processed data. Moreover, in case of redundant storage (batch processing applications often provide a 3-fold redundancy [31]), an algorithm should also be aware of, and exploit, *replica selection* flexibilities.

Our Contributions. This paper initiates the formal study of data-locality and replica aware virtual network embedding problems in datacenters. In particular, we decompose the general optimization problem into its fundamental aspects, such as assignment of chunks, replica selection, and flexible virtual machine placement, and answer questions such as:

1. Which chunks to assign to which virtual machine?
2. How to exploit redundancy and select good replicas?
3. How to efficiently embed virtual machines and their inter-connecting network?
4. Can the chunk assignment, replica selection and virtual machine embedding problems be jointly optimized, in polynomial time?

We draw a complete picture of the problem space: We show that even problem variants exhibiting multiple degrees of freedom in terms of replica selection and embedding, can be solved optimally in polynomial time, and we present several efficient algorithms accordingly. However, we also prove limitations in terms of computational tractability, by providing reductions from 3-D matching and Boolean satisfiability (SAT). Interestingly, while it is well-known that (unsplittable) multi-commodity flow problems are NP-hard in capacitated networks, our hardness results also hold in *uncapacitated* networks; moreover, we show that NP-hard problems already arise in small-diameter networks (as they are widely used today [2]), and even if the number of replicas is bounded by two.

Organization. The remainder of this paper is organized as follows. Section 2 introduces our formal model in detail. Algorithms are presented in Section 3 and hardness results are presented in Section 4. After discussing related work in Section 5, we conclude our work in Section 6. Some technical details as well as additional hardness results appear in the Appendix.

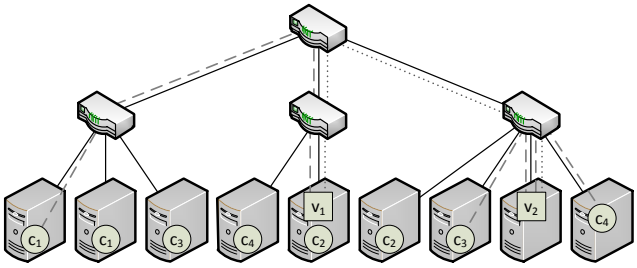

Figure 1: Overview: a 9-server datacenter storing $\tau = 4$ different chunk types $\{c_1, \ldots, c_4\}$ (depicted as *circles*). The chunk replicas need to be selected and assigned to the two virtual machines v_1 and v_2; the virtual machines are depicted as *squares*, and the network connecting them to chunks (at bandwidth b_1) is *dashed*. In addition, the virtual machines are inter-connected among each other at bandwidth b_2 (*dotted*). The objective of the embedding algorithm is to minimize the overall bandwidth allocation (sum of *dashed* and *dotted* lines).

2. MODEL

To get started, and before introducing our formal model and its constituting parts in detail, we will discuss the practical motivation. Figure 1 gives an overview of our model.

2.1 Background and Practical Motivation

Our model is motivated by batch-processing applications such as MapReduce. Such applications use multiple virtual machines to process data, initially often redundantly stored in a distributed file system implemented by multiple servers. [11] The standard datacenter topologies today are (multi-rooted) fat-tree resp. *Clos* topologies [2, 20], hierarchical networks recursively made of sub-trees at each level; servers are located at the tree leaves. Given the amount of multiplexing over the mesh of links and the availability of multi-path routing protocol, e.g. ECMP, the redundant links can be considered as a single aggregate link for bandwidth reservations [5, 33].

During execution, batch-processing applications typically cycle through different phases, most prominently, a mapping phase and a reducing phase; between the two phases, a shuffling operation is performed, a phase where the results from the mappers are communicated to the reducers. Since the shuffling phase can constitute a non-negligible part of the overall runtime [8], and since concurrent network transmissions can introduce interference and performance unpredictability [36], it is important to provide explicit minimal bandwidth guarantees [25]. In particular, we model the virtual network connecting the virtual machines as a virtual cluster [5, 25, 33]; however, we extend this model with a notion of data-locality. In particular, we distinguish between the bandwidth needed between assigned chunk and virtual machine (b_1) and the bandwidth needed between two virtual machines (b_2); in practice, for applications with a large "mapping ratio" where the mapping phase already reduces the data size significantly, it may hold that $b_2 \ll b_1$.

2.2 Fundamental Parts

Let us now introduce our model more formally. It consists of three fundamental parts: (1) the substrate network (the servers and the connecting physical network), (2) the to be processed input (the data chunks), and (3) the virtual network (the virtual machines and the logical network connecting the machines to each other as well as to the chunks).

The Substrate Network. The substrate network (also known as the *host graph*) represents the physical resources: a set S of $n_S = |S|$ servers interconnected by a network consisting of a set R of routers (or switches) and a set E of (symmetric) links; we will often refer to the elements in $S \cup R$ as the *vertices*. We will assume that the interconnecting network forms an (arbitrary, not necessarily balanced or regular) tree, where the servers are located at the tree leaves. Each server $s \in S$ can host a certain number of virtual machines (available server capacity $cap(s)$), and each link $e \in E$ has a certain bandwidth capacity $cap(e)$.

The Input Data. The to be processed data constitutes the input to the batch-processing application. The data is stored in a distributed manner; this spatial distribution is given and not subject to optimization. The input data consists of τ different *chunk types* $\{c_1, \ldots, c_\tau\}$, where each chunk type c_i can have $r_i \geq 1$ instances (or replicas) $\{c_i^{(1)}, \ldots, c_i^{(r_i)}\}$, stored at different servers. A single server may host multiple chunks. It is sufficient to process one replica, and we will sometimes refer to this replica as the *active* (or selected) replica.

The Virtual Network. The virtual network consists of a set V of $n_V = |V|$ virtual machines, henceforth often simply called *nodes*. Each node $v \in V$ can be placed (or, synonymously, *embedded*) on a server; this placement can be subject to optimization.

Depending on the available capacity $cap(s)$ of server s, multiple nodes may be hosted on s. We will denote the server s hosting node v by $\pi(v) = s$. Since these nodes process the input data, they need to be assigned and connected to the chunks. Concretely, for each chunk type c_i, exactly one replica $c_i^{(j)}$ must be processed by exactly one node v; which replica $c_i^{(k)}$ is chosen is subject to optimization, and we will denote by μ the assignment of nodes to chunks.

In order to ensure a predictable application performance, both the connection to the chunks as well as the interconnection between the nodes may have to ensure certain minimal bandwidth guarantees; we will refer to the first type of virtual network as the *(chunk) access network*, and to the second type of virtual network as the *(node) inter-connect*; the latter is modeled as a complete network (a *clique*). Concretely, we assume that an active chunk is connected to its node at a minimal (guaranteed) bandwidth b_1, and a node is connected to any other node at minimal (guaranteed) bandwidth b_2.

Note that we do not explicitly model the process of copying back the final results from the reducers to the file system: this step offers many flexibilities, and can be done based on the current network state; [7] it does not require explicit reservations.

2.3 Optimization Objective

Our goal is to develop algorithms which minimize the *resource footprint*: the guaranteed bandwidth allocation (or synonymously: *reservation*) on all links of the given embedding; note that only the resource allocation at the links but not at the servers depends on the replica selection or embedding. Thus, we on the one hand aim to embed the nodes in a locality-aware manner, close to the input data (the chunks),

but at the same time also aim to embed the nodes as close as possible to each other.

Formally, let $dist(v, c)$ denote the distance (in the underlying physical network T) between a node v and its assigned (active) chunk replica c, and let $dist(v_1, v_2)$ denote the distance between the two nodes v_1 and v_2. We define the *footprint* $\mathrm{F}(v)$ of a node v as follows:

$$\mathrm{F}(v) = \sum_{c \in \mu(v)} b_1 \cdot dist(v, c) + \underbrace{\frac{1}{2} \cdot \sum_{v' \in V \setminus \{v\}} b_2 \cdot dist(v, v')}_{\text{only for inter-connect}},$$

where $\mu(v)$ is the set of chunks assigned to v. Our goal is to minimize the overall footprint $\mathrm{F} = \sum_{v \in V} \mathrm{F}(v)$.

2.4 Problem Decomposition

In order to chart the landscape of the computational tractability and intractability of different problem variants, we decompose our problem into its fundamental aspects, namely replica selection (RS), multiple chunk assignment (MA), flexible node placement (FP), node interconnect (NI), and bandwidth constraints (BW), as described in the following. In this paper, we will consider all possible 32 problem variants, where each of these five aspects can either be enabled or disabled.

***Replica Selection* (RS).** The first fundamental problem is replica selection: if the input data is stored redundantly, the algorithm has the freedom to choose a replica for each chunk type, and assign it to a virtual machine (i.e., *node*). In the following, we will refer to a scenario with redundant chunks by RS; in the RS-only scenario, the number of chunk types is equal to the number of nodes. Otherwise, we will add the +MA property discussed next.

***Multiple Assignment* (MA).** If the number of chunk types τ is larger than the number of nodes, each node needs to be assigned multiple chunks. We will refer to such a scenario by MA, and will denote the number of chunks per node by the integer value $m = \tau / n_V$ (a constant).

***Flexible Placement* (FP).** The third fundamental degree of freedom, besides replica selection and assignment, regards the flexibility in the placement (or synonymously: *embedding*) of nodes on physical servers. We will refer to this aspect by FP.

***Node Interconnect* (NI).** We distinguish between scenarios where bandwidth needs to be reserved both from each node to its assigned chunks as well as to the other nodes (i.e., $b_1 > 0$ and $b_2 > 0$), and scenarios where only the (chunk) access network requires bandwidth reservation (i.e., $b_1 > 0$ and $b_2 = 0$). We will refer to the former scenario where bandwidth needs to be reserved also for the inter-connect, by NI. We also note that the case $b_1 = 0$ and $b_2 > 0$ has been studied in prior work [5, 25, 33], and refer the reader to our related work section.

***Bandwidth Capacities* (BW).** We distinguish between an uncapacitated and a capacitated scenario where the links of the substrate network come with bandwidth constraints, and will refer to the bandwidth-constrained version by BW; the capacity of servers (the number of nodes which can be hosted concurrently) is always limited. Note that capacity constraints introduce infeasible problem instances, where it is impossible to allocate sufficient resources to satisfy an embedding request.

3. POLYNOMIAL-TIME ALGORITHMS

Despite the various degrees of freedom in terms of embedding and replica selection, we can solve many problem variants efficiently. This section introduces three general techniques, which can roughly be categorized into *flow* (Section 3.1), *matching* (Section 3.2) and *dynamic programming* (Section 3.3) approaches. As we will see, sometimes multiple approaches can be used to solve a given problem variant, however, runtimes can differ. For each approach, we will summarize the solvable problem variants using a *Venn diagram*. First, let us make a simplifying observation:

OBSERVATION 1. *In problems without flexible placement* (FP)*, the bandwidth required for the inter-connect network* (NI) *can generally be allocated* upfront*, as it does not depend on the replica selection and assignment. Accordingly, we can reduce problem variant* RS + MA + NI + BW *(as well as all its subproblems) to* RS + MA + BW *(resp. its subproblems).*

3.1 Flow Algorithms (RS + MA + NI + BW)

We first present an algorithm to solve the RS + MA + NI + BW problem. Recall that in this problem variant, we are given a set of redundant chunks (RS) and a set of nodes (the *nodes*) at fixed locations (no FP). The number of chunk types is larger than the number of nodes (MA), and each node needs to be connected to its selected chunks as well as to other nodes (NI), while respecting capacity constraints (BW). Our goal is to minimize the resource footprint F, consisting of the bandwidth reservations in the (chunk) access network and the (node) interconnect. As we will see in the following, we can use a flow approach to solve this problem variant.

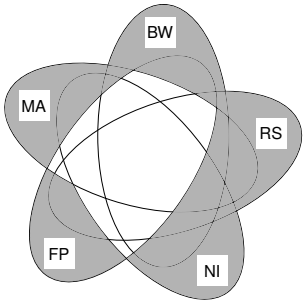

Figure 2: Variants solved by flow approach.

Construction of Artificial Graph. In order to solve the RS + MA + NI + BW problem, we first remove the NI property using Observation 1. We then construct an artificial graph T^*, extending the substrate network T and normalizing bandwidth capacities, as follows. For T^*, we normalize the bandwidth of T to integer multiples of b_1, i.e., for each link $e \in E(T)$, we set its new capacity in T^* to $\lfloor cap(e)/b_1 \rfloor$. After this normalization, we extend the topology T by introducing an artificial vertex for each chunk type. These artificial vertices are connected to each leaf (i.e., server) in T where a replica of the respective chunk type is located, connecting the replica of the respective chunk type by a link of capacity 1. In addition, we create a *super-source* s^+, and connect it to each of the artificial chunk type vertices (with a link of capacity 1). Moreover, we create an artificial *super-sink* s^- and connect it to every leaf containing at least one node; the link capacity represents the number of nodes x hosted on this server, times the multi-assignment factor m: $x \cdot m$. We additionally assign the following costs to edges of T^*: every edge of the original substrate network costs one unit, and all other artificial edges cost nothing.

A solution to the RS+MA+BW problem can now be com-

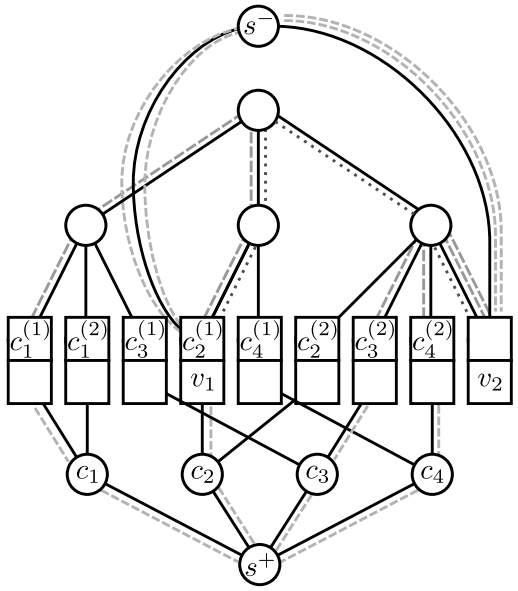

Figure 3: Example of flow construction: Problem instance with two nodes, four chunk types, and two replicas per type. The min-cost-max-flow is indicated by the dashed lines: each line represents one unit of flow.

puted a from a solution to the *Min-Cost-Max-Flow* problem between super-source s^+ and super-sink s^- on the artificial graph T^*.

Example. Figure 3 shows an example of the extended substrate network T^*: The source s^+ is connected to the two leaves, which host the nodes. The artificial nodes are depicted below the leaves, are labeled with their respective chunk types (e.g., c_1), and are connected to the sink s^- as well as to the leaves which contain replicas of their chunk type. The maximum flow with minimal costs is indicated by the dashed lines: each line represents one unit of flow. The dotted lines indicate links which have reduced capacity due to NI.

Algorithm. Our algorithm to solve RS+MA+NI+BW consists of three parts: *First*, we construct the normalized and extended graph T^* described above and compute a min-cost-max-flow solution, e.g., using [18, 32]. *Second*, we have to *round* the resulting, possibly fractional flow, to integer values. Due to the *integrality theorem* [1], there always exists an optimal integer solution on graphs with integer capacities. However, while algorithms like the successive shortest path algorithm [24] directly give us such an integral solution (in polynomial time), the fastest min-cost-max-flow algorithms (e.g., based on double-scaling methods [18] or minimum mean-cost cycle algorithms [32], may yield fractional solutions which need to be rounded to integral solutions (of the same cost). In order to compute integral solutions, we proceed as follows: we iteratively pick an arbitrary (loop-free) path currently having a fractional allocation of value f ($f > 0$), and distribute its flow f among all other fractional paths of the same length; due to the optimality of the fractional solution and due to the integrality theorem, such paths must always exist. After distributing this flow, the total allocation on this path will be 0, and we have increased the number of integer paths by at least one. We proceed until we constructed the perfect matching. *Third*, given an integer min-cost-max-flow solution, we need to decompose the integer flow into the paths representing matched chunk-node pairs: The assignment can be obtained by decomposing the flow allocated in the original substrate network. In order to identify a matched chunk-node pair, we take an arbitrary (loop-free) path p carrying a flow of value ≥ 1 from s^+ to s^-: the first hop represents the chosen chunk type, the second hop the chosen replica, and the last but one hop represents the server: we will assign the replica to an arbitrary unused node on this server. Having found this pair, we reduce the flow along the path p by one unit. We continue the pairing process until every chunk type is assigned.

Analysis. The correctness of our approach follows from our construction of T^*, using integer capacities (in our case $\lfloor cap(e)/b_1 \rfloor$), and the fact that cost optimal integral solutions always exist [1]. The runtime of our algorithm consists of four parts: construction of T^*, computation of the min-cost-max-flow, flow rounding, and decomposition. The dominant term in the asymptotic runtime is the flow computation. Using the state-of-the-art min-cost-max-flow algorithms [18, 32] we get a runtime of $\mathcal{O}(n_S^2 \cdot \log\log \min\{U, \tau\})$ where U is the maximal link capacity; note that in networks with high capacity and uncapacitated networks, we can simply set $U = \tau$.

3.2 Matching Algorithms (RS + MA + NI and MA + NI + BW)

This section presents faster algorithms to solve the two problem variants RS + MA + NI and MA + NI + BW which can also be solved with the flow approach introduced above. In general, we refer to the algorithms presented in this section as matching approaches.

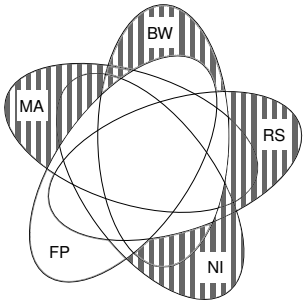

Figure 4: Variants solved by matching approaches.

3.2.1 RS + MA + NI

Let us first consider the RS + MA + NI variant. Recall that in this problem, we are given a set of redundant chunks (RS) and a set of nodes at fixed locations. The number of chunk types is larger than the number of nodes (MA), and each node needs to be connected to its chunks as well as to other nodes (NI). Our goal is to minimize the resource footprint F, consisting of the bandwidth reservations in the access network and the inter-connect.

Algorithm. Due to Observation 1, RS+MA+NI degenerates to RS + MA. In order to solve the RS + MA problem variant, we construct a bipartite graph between the set V of nodes and the set of chunks. Concretely, we clone each node m times, as each node needs to process m chunk types, and we collect all copies of a given chunk type in a single "supernode". We connect each node to all chunk types using the *lowest hop count* to one of the copies as the cost metric (the link weight). On the resulting bipartite graph, we can now compute a *Minimum Weight Perfect Matching* [16]: the resulting matching describes the optimal assignment of chunks to nodes.

Example. Before analyzing our algorithm, let us consider

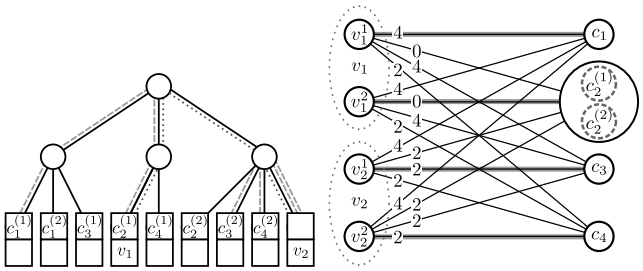

Figure 5: The RS + MA problem on the *left* is converted into a matching problem on the *right*. Since each node has to process two chunks, the nodes are replicated in the matching representation. The two replicas of each chunk type are represented by a single node, and all edges connecting to this node have a weight according to the shorter distance to one of the replicas. This is visualized for c_2.

a small example. Figure 5 illustrates an instance where two nodes are cloned into $m = 2$ nodes each, resulting in a total of four nodes in the matching problem representation. The two replicas of each chunk type are aggregated into a single chunk type vertex c_j in the matching problem; this gives a total of four chunk type vertices in the matching graph. The costs on the links between all clones of a specific vertex and a chunk type are set to the minimum distance. We can observe this for instance at the edges connecting the two clones of v_1 to c_2: both weights are 0.

Analysis. The correctness of our algorithm follows from the construction and the optimal solution of the minimum matching. The runtime consists of two parts: the construction of the matching graph and the actual matching computation. The constructed graph consists of $m \cdot n_V \cdot \tau$ many edges, and for each edge we need to compute its cost, i.e., the shortest distance which in a tree we can computed in time n_S; thus, the overall construction time is $\mathcal{O}(n_S \cdot \tau^2)$. The state of the art algorithm to compute matchings are based on scaling techniques [14]. The runtime translates to $\mathcal{O}(\tau^{5/2} \cdot \log(\tau \cdot n_S))$; recall that $\tau = m \cdot n_V$.

3.2.2 *Faster* MA + NI *and* MA + NI + BW

We now show that we can solve MA + NI even faster, by exploiting locality. Moreover, we will show that we can even solve MA + NI + BW problem variants by simply verifying feasibility. . In the following, due to Observation 1, we can focus on the MA resp. MA + BW problem.

We first introduce the following definition.

Definition 1 (Local Assignment (LA)). *We define an assignment μ to be* local in a specific subtree T', *iff μ assigns the maximum number of chunks in the subtree to nodes in the same subtree. We define μ to be* local *when it is local with respect to all possible subtrees of the substrate network.*

Example. Figure 6 illustrates the concept of local assignment: The closest chunk to v_2 is c_1, and the closest node to c_1 is v_2. However, a subtree T' exists such that $v_1 \in T'$ and $c_1 \in T'$, but $v_2 \notin T'$. Therefore, a local assignment cannot assign c_1 to v_2.

We will see later that optimal solutions to MA have a local assignment. We exploit this in our algorithms described in

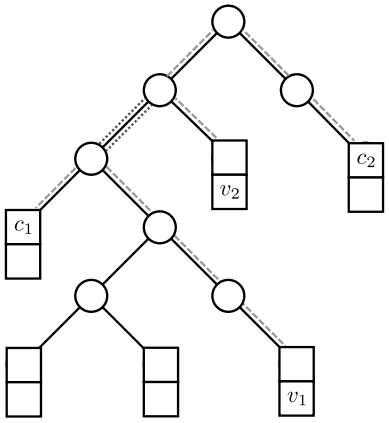

Figure 6: Illustration of local assignment: The dashed lines indicate bandwidth allocations, which occur independently of the chosen assignment. The dotted lines indicate bandwidth allocation which occur only if c_2 is assigned to v_1.

the following.

Algorithm. Our proposed algorithm for MA proceeds in a bottom-up fashion, traversing the substrate network T from the leaves toward the root. For each subtree T', we maintain two sets S_1, S_2 in order to match unmatched chunks S_1 in the subtree T' to unmatched nodes S_2 in T'. Both sets are initially empty.

We first process all the leaves, in an arbitrary order; subsequently, we process arbitrary inner vertices of T, whenever all their children have been processed. We process any leaf ℓ by adding any nodes or chunks which are located on ℓ to the corresponding sets S_1 and S_2. A non-leaf vertex u is processed in the following way: we take the union of the sets of u's children, i.e., the sets contain the unmatched chunks and nodes in this subtree. For both leaves and inner nodes, whenever both sets are non-empty, we greedily match an arbitrary chunk in S_1 with an arbitrary node in S_2, and remove them from the sets.

Analysis. On a given vertex u, emptying one of the sets, results in a *local assignment* (cf Definition 1) in the subtree rooted at u. The bottom-up strategy ensures that this works for every subtree in the substrate, rendering the resulting assignment local. The complexity of this construction is low: For each vertex in the substrate graph, we build the union of the children's sets, and since each vertex can only be the child of one vertex, the amortized runtime per vertex is constant; and hence the overall runtime $\mathcal{O}(n_S)$. The sum of all remove operations, is equal to the number of chunk types $\mathcal{O}(\tau)$. Hence the overall complexity of this construction amounts to $\mathcal{O}(n_S + \tau)$.

It remains to prove optimality of such local assignments. We first characterize the bandwidth allocation on uplinks of subtrees.

Lemma 1. *Given an* MA *problem and a subtree T' containing x chunks and y nodes, the minimal bandwidth allocation of any assignment μ on the uplink of T' is $|x - y \cdot m| \cdot b_1$.*

Proof. In case the number of chunk types equals the processing capacities of the nodes in the given subtree, the bandwidth allocation inflicted by the chunk access network

on the uplink can be zero, since we can assign all chunks to nodes in the same subtree. Otherwise, we distinguish between two cases: Recall, that in instances without RS, all chunks have to be processed. In case there are more chunks in the subtree, at least all of the excess chunks have to be transferred to a different subtree, which will inflict costs b_1 per excess chunk on the uplink connecting T' with the remaining parts of T, which will inflict costs b_1 per excess chunk on the uplink of root of T'. Similarly, if the processing capabilities exceed the amount of available chunks, excess chunks from other subtrees will have to be transferred to nodes in the subtree T', inflicting bandwidth costs of b_1 each. Hence, the minimum bandwidth allocation for the chunk access on the uplink is the difference between the number of chunks and the processing capabilities of the subtree $|x - y \cdot m|$ times the amount of bandwidth needed, for a single transfer b_1. □

THEOREM 2. *Given an* MA + NI *problem instance, a feasible assignment μ is optimal iff it is local.*

PROOF. Local assignments generate exactly the minimal allocations on all links, as the assignments which generate the minimal bandwidth allocations described in the proof of Lemma 1 are local in the given subtree. Hence each local assignment has to be optimal. A non-local assignment, has at least one subtree, in which it is not local. This subtree will have a higher allocation on the uplink. Since the local assignment has minimal allocations on all other links, the non local assignment has a larger footprint. □

Combined with a simple postprocessing step, this approach can also solve MA + BW. The central idea of this extension, is that *local* assignments allocate the minimal bandwidth on each individual edge. In consequence, each bandwidth constraint which is lower than the allocation of a local assignment on one link, renders the problem infeasible. Hence, it is sufficient to temporarily omit the bandwidth limitations, compute an optimal assignment for an MA instance, and verify that the resulting allocations do not violate any capacities. The postprocessing step scales linearly with the number of edges in the substrate graph.

3.3 Dynamic Programming (MA+FP+NI+BW)

We now show how to solve the MA + FP + NI + BW problem variant in polynomial time. Note that this problem variant requires to find a tradeoff between the desire to place nodes as close as possible to each other (in order to minimize communication costs), and the desire to place nodes as close as possible to the chunk locations.

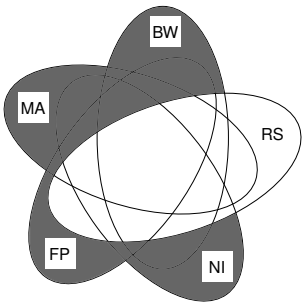

Figure 7: Variants solved by dynamic programming approach.

Example. Figure 8 shows an example: one extreme solution is to minimize the distance between chunks and nodes, see mapping π_1 in Figure 8 *(left)*: the four nodes are all collocated with chunks, resulting in a zero-cost

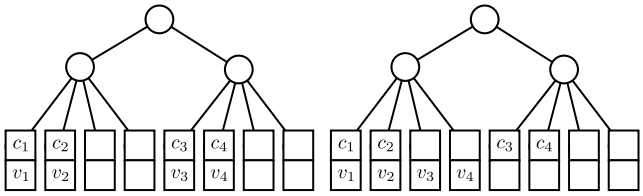

Figure 8: Two different node placements for the same substrate graph and chunk locations. For $b_1 = b_2$, both solutions have an identical footprint. In other cases, one solution outperforms the other.

chunk access network. As a result, the paths between the individual nodes are longer than in alternative node placements: each node has a distance of two hops to one other node, and four hops to two other nodes. Hence the resulting allocations for the node interconnect sum up to $20 \cdot b_2$. Figure 8 *(right)* shows a different node mapping π_2, which seeks to minimize the communication costs between the nodes, and places all nodes in one subtree. The distance between all nodes is two, which results in a total bandwidth allocation of $12 \cdot b_2$ for the interconnect. However, this reduced price comes at additional costs in the access network: c_3 and c_4 have to be communicated to v_3 and v_4, which requires a total bandwidth allocation of $8 \cdot b_1$.

Basic ideas. Our proposed approach is based on dynamic programming, and leverages the *optimal substructure property* of MA + FP + NI + BW: as we will see, optimal solutions for subproblems (namely subtrees) can efficiently be combined into optimal solutions for larger problems. Indeed, the MA + FP + NI + BW problem exhibits such a structure, and we show how to exploit it to compute efficient embeddings, even in scenarios where multiple chunks need to be assigned to flexibly placeable nodes.

For ease of presentation we will transform the substrate network T into a binary tree, using binarization: we clone every higher-degree node, iteratively attaching additional clones as right children and original children as left descendants.

As usual in dynamic programs, we define, over the structure of the tree, a recursive formula formula f for the minimal cost solution *given* any possible number of nodes embedded in a given subtree. The actual set does not matter, due to symmetry arguments. Our approach is to evaluate this function in a bottom-up manner. To finally compute the actual optimal embedding, we traverse the computed minimal-cost path backwards (according to the optimal values found for f during the bottom-up computation).

Concretely, the first argument to function f is a subtree T', containing a given number of chunks $y(T')$, and the second argument is the number of nodes to be embedded in the subtree. Function f is evaluated in a bottom up manner. We initialize the function at each leaf ℓ, by $f(T_\ell, x) = \infty$ for all numbers of nodes x which are larger than the server capacity $cap(\ell)$; to calculate $f(T_\ell, x)$, for $x \leq cap(\ell)$, we compute the bandwidth allocation on the uplink of T_ℓ, referred to by the function $bw(T_\ell, x)$: $bw(T_l, x) = b_1 \cdot |x - y(T_\ell)| + b_2 \cdot (n_V - x) \cdot x$, which accounts for the bandwidth allocation on the uplink of T_ℓ. The first term represents the required bandwidth for the communication between the x nodes on ℓ, and the $n_V - x$ nodes in the remaining parts of the substrate network. The

second term represents the bandwidth, which is necessary to transport the chunks from their location to the node which should process the data (see Lemma 1 for more details).

After initialization, we proceed to compute f for non-leaf nodes in a bottom-up manner: We split the x nodes into two positive integer values, and we put r on the right and $x-r$ on the left subtree. That is, we take the optimal cost (given recursively) of placing r nodes in the right subtree $\text{RI}(T')$ of T' and $x-r$ nodes in left subtree $\text{LE}(T')$ of T'. Given the cheapest combination, we add the bandwidth requirements on the uplink of T' to generate the overall costs for placing x nodes in T'.

$$f(T',x) = \min_{0\le r\le x} \{f\left(\text{LE}(T'), x-r\right) + f\left(\text{RI}(T'), r\right)\} + bw(T',x)$$

Again, we set $f(T',x)$ to infinity if the required bandwidth bw exceeds the capacity cap of the uplink of T'.

Analysis. The correctness and optimality of our dynamic program is due to the decoupling of the costs induced by the tree structure of T and the substructure optimality property. The substructure optimality follows from the observation that costs can be accounted on the uplink, and the fact that we check each possible node distribution. For each substrate vertex (n_S many) we have to check the cost of all possible splits, resulting in an overall complexity of $\mathcal{O}(n_S \cdot n_V^2)$. The runtime to binarize T is asymptotically negligible.

Remark: Simple Problems. To conclude, for the sake of completeness, we also observe that there are several problems which allow for a trivial solution. Concretely, problems with FP plus any combination of RS and BW (but without MA and NI) can easily be solved by mapping nodes to chunk locations.

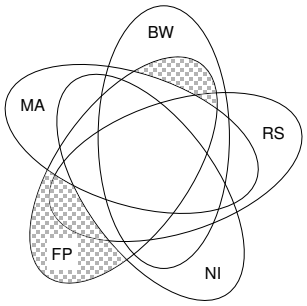

Figure 9: Trivially solvable problem variants.

4. NP-HARDNESS RESULTS

We have seen that even problems with multiple dimensions of flexibility can be solved optimally in polynomial time. This section now points out fundamental limitations in terms of computational tractability. In particular, we will show that problems become NP-hard if multiple chunks have to be assigned to a flexibly placeable node (FP+RS+MA is proved NP-hard in Section 4.2), and if inter-connects have to be established between flexible nodes (FP + RS + NI is proved NP-hard in Section 4.3); both results hold even in uncapacitated networks, and even in small-diameter substrate networks (namely two- or three-level trees [2]). The hardness of FP + RS + MA and FP + RS + NI imply the hardness of four additional, more general models, as summarized in Figure 10:

Additional NP-hardness results appear in the Appendix.

4.1 Introduction to 3D Perfect Matching

Both the hardness of FP+RS+MA and FP+RS+NI are shown by a reduction from the NP-complete problem of *3D Perfect Matching* [10], which we can see as a generalization of bipartite matchings to 3-uniform hypergraphs. We will refer to this problem by 3-DM, and for completeness, review it quickly: 3-DM is defined as follows. We are given three finite and disjoint sets X, Y, and Z of cardinality k, as well as a subset of triples $T \subset X \times Y \times Z$. Set $M \subseteq T$ is a 3-dimensional matching if and only if, for any two distinct triples $t_1 = (x_1, y_1, z_1) \in M$ and $t_2 = (x_2, y_2, z_2) \in M$, it holds that $x_1 \neq x_2$, $y_1 \neq y_2$, and $z_1 \neq z_2$. Our goal is to decide if we can construct a $M \subseteq T$ which is *perfect*, that is, a subset which covers all elements of $X \times Y \times Z$ exactly once.

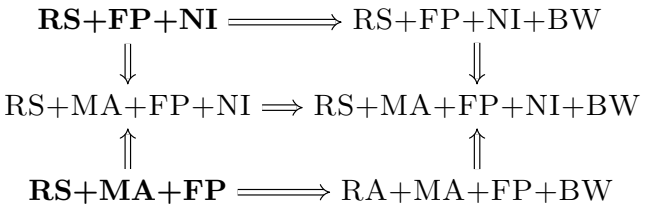

Figure 10: The NP-hardness of 2 variants, implies that 4 other variants are also NP-hard.

4.2 Multi-Assignments are hard (FP+RS+MA)

Our proof that FP + RS + MA is NP-hard is based on the following main ideas. We encode a 3-DM instance as an FP + RS + MA instance as follows:

- For every element in the universe $X \cup Y \cup Z$, we create a chunk type. Intuitively, in 3-DM, each element must be covered, which corresponds to the requirement of FP + RS + MA that each chunk type is processed.
- We will encode each triple as gadget with three leaves in a substrate tree T. The three leaves are close to each other in T, and the placement of chunk replicas in FP + RS + MA corresponds to the elements of the triples in these leaves.
- The node placement will correspond to the choice of triples, independently of which leaf the node is mapped to. A node will process its collocated chunk, as well as the chunks in other two leaves of the same gadget.
- In order to turn the optimization problem into a decision problem, we will use a cost threshold Th. The cost threshold will be met by all assignments which assign all three chunks of each triple to a node which is collocated with one of the chunks. Assignments which connect a chunk to a node in a different triple, will have a larger footprint, and are considered to be infeasible.

Construction. Given an instance I of 3-DM in which k triples have to be choosen, we construct an instance I' of FP + RS + MA as follows:

- *Tree Construction:* We create a tree consisting of a root, and for each triple, we create a gadget which we directly attach as child of the root. The gadget is of height 2, and has the following form: The gadget of each triple consists of an inner node (a router) and three leaves.

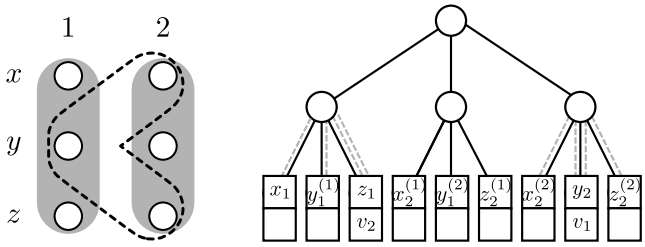

Figure 11: ***Left:*** **A 3-DM instance with three triples:** (x_1, y_1, z_1)**,** (x_2, y_1, z_2)**, and** (x_2, y_2, z_2)**. The solution is indicated by the grey triples; the dashed triple is not used for the solution.** ***Right:*** **The corresponding problem and solution of FP + MA + RS.**

- *Chunks and chunk replicas:* For each element in X, Y and Z, we create a chunk type ($3 \cdot k$ in total). Every gadget contains three chunk replicas, corresponding to the elements of the triple. Each leave in a gadget, contains exactly one replica.
- *Other properties:* We set the number of to-be-embedded nodes to k, b_1 to 1, and the number of chunk slots in each node to the multi-assignment factor $m = 3$. We use a threshold $Th = 4 \cdot k$.

Example. Figure 11 shows an example of our construction: An instance I of 3-DM is given: The disjoint sets X, Y and Z have a cardinality $k = 2$. We will refer to the two elements in X as x_1 and x_2, and use the same notation for the other two sets. T contains the three triples (x_1, y_1, z_1), (x_2, y_1, z_2), and (x_2, y_2, z_2). The goal of 3-DM is to find a subset $M \subseteq T$, which contains each element in each of the three sets exactly once. This instance only has one solution: $M = \{(x_1, y_1, z_1), (x_2, y_2, z_2))\}$.

To construct the corresponding instance I' of FP + RS + MA, we create a gadget for each triple in T. For each variable which occurs in a triple, the corresponding gadget contains a chunk of the type of the variable. The triple (x_2, y_1, z_2) of the instance is represented by the middle gadget in Figure 11. The objective of I' is to spawn $k = 2$ nodes, with the smallest possible footprint. If the total footprint is $\leq 2 \cdot 2 \cdot k$, we can construct a solution to I from the solution to I'. The footprint consists of the costs which occur when a node is embedded in a gadget, and the three chunks of that gadget which are assigned to that node: one of the chunks is collocated with the node, the other two have to be transferred via two hops, inflicting unitary costs on each hop.

Correctness. Given these concepts, we can now show the computational hardness.

THEOREM 3. FP + RS + MA *is NP-hard.*

PROOF. Let I be an instance of 3-DM and let I' be an instance of FP + RS + MA constructed as described above. We prove that I' has a solution of cost $\leq Th$ if ($\Rightarrow$) and only if ($\Leftarrow$) I has a matching of size k.

($\Rightarrow$) Let us take a solution to 3-DM. We place a node in every gadget that corresponds to the chosen triples. In each of the corresponding gadgets, we match every chunk to the node in this gadget. This solution has cost exactly Th. As every element of the universe is covered, every chunk type is processed.

($\Leftarrow$) Let us take a solution to FP+RS+MA of cost $\leq Th$. We choose triples that correspond to gadgets where there are nodes. Since all chunks are processed, every element of X, Y and Z is matched. Each node must process chunks that correspond to the triple, otherwise the cost must be larger than Th (high costs for chunk transportation). □

4.3 Inter-connects are hard (FP + RS + NI)

Next, we prove that the joint optimization of node placement and replica selection is NP-hard if an inter-connect has to be established between nodes. In our terminology, this is the FP + RS + NI problem.

The proof is similar in spirit to the proof of FP + RS + MA, however, we modify the construction to account for the absence of MA: we choose a high value for b_1, such that nodes will be directly collocated with their assigned chunks. We leverage the fact that any solution which does not assign 0 or 3 chunks to each gadget, will have higher communication costs.

Construction. Let I be an instance of 3-DM. We will create an instance I' for FP + RS + NI as follows:

- We will construct the same tree as in previous reduction with chunk replicas placed in the same way.
- The communication cost in the inter-connect is set to $b_2 = 1$.
- The number of nodes (virtual machines) is $n_V = 3 \cdot k$, where k is the set cardinality.
- Only solutions which place a node in each leaf of k gadgets, can be converted into solutions for the 3-DM problem. We use the cost threshold $Th = 6 \cdot k + 18 \cdot (k-1) \cdot k$, to verify whether a solution achieves this, transforming FP + RS + NI into a decision problem. A detailed explanation of this value can be found in the proof of Theorem 5.
- We set the access cost b_1 to a chunk replica to a high value W. This will force nodes to be collocated with the replica. One example of sufficient (and polynomial but not necessarily minimal) W is the value of the threshold $Th + 1$. Any solution not assigning chunks to collocated nodes, have cost $> Th$: communicating a chunk inflicts costs $W = Th + 1$ over every link.

We focus on instances with unit server capacities.

Proof of correctness. Intuitively, in order to minimize embedding costs, nodes should be placed on near-by replicas. We use the following helper lemma.

LEMMA 4. *In every valid solution of I' of cost $\leq Th$, each gadget falls in one of two categories: k gadgets have exactly 3 nodes, and $n - k$ gadgets remain empty.*

PROOF. Since W is large enough, the $3 \cdot k$ nodes have to be placed directly on different chunks, resulting in 0 costs for the access network. Consider any pair of nodes communicating over the inter-connect; due to our construction, the communication cost for each such pair is either 2 hops (if they belong to the same gadget) or 4 hops (if they belong to different gadgets). The lemma then follows from the observation that Th is chosen such that it is never possible to distribute nodes among more than k gadgets. □

THEOREM 5. FP + RS + NI *is NP-hard.*

PROOF. Let I be an instance of 3-DM and let I' be an instance of FP + RS + NI constructed as described above. We prove that I' has solution of cost $\leq Th$ if ($\Rightarrow$) and only if ($\Leftarrow$) I has a solution.

($\Rightarrow$) In order to compute a solution for I' given a solution for I, we proceed as follows. Given an exact covering set of triples $S = \{t_1, t_2, \ldots, t_k\}$, we place three nodes in each gadget that corresponds to every triple of S. Chunks are matched to the nodes which are located on the same server.

The solution has the following cost: (1) the communication cost inside a gadget is $2 \cdot \binom{3}{2}$, as every pair contributes two hops; (2) the communication cost from each gadget to all other gadgets is $4 \cdot 3 \cdot 3 \cdot (k-1)/2$, where the factor 4 is for the communication over 4 hops, the factor 3 corresponds to the number of nodes per gadget, and $3 \cdot (k-1)$ is the number of nodes in remote gadgets; as we count each pair twice, we need to divide by two in the end. Summing up over all k gadgets, we get exactly Th.

($\Leftarrow$) Given a solution for I', we can exploit Lemma 4 to construct a solution for I. We know that in any solution of cost at most Th, k gadgets contain exactly 3 nodes. These gadgets correspond to a valid 3D Perfect Matching: exactly one replica of every chunk type is processed and hence every element is covered exactly once. □

Finally, let us conclude by remarking that all NP-hard problems discussed in this paper are obviously also in NP (and hence NP-complete): solutions can be verified efficiently.

NP-hard	5 combinations	RS + MA + FP + NI + BW
	4 combinations	RS + MA + FP + NI; RS + MA + FP + BW; RS + FP + NI + BW
	3 combinations	RS + MA + FP; RS + FP + NI
Flow	4 combinations	RS + MA + NI + BW
	3 combinations	RS + NI + BW; RS + MA + BW
	2 combinations	RS + BW
DP	4 combinations	MA + FP + NI + BW
	3 combinations	MA + FP + NI; MA + FP + BW; FP + NI + BW
	2 combinations	MA + FP; FP + NI;
Matching	3 combinations	RS + MA + NI; MA + NI + BW
	2 combinations	RS + MA; RS + NI; MA + NI; MA + BW; NI + BW
	1 combinations	RS; MA; NI; BW
0 Cost	3 combinations	RS + FP + BW
	2 combinations	RS + FP; FP + BW
	1 combinations	FP

Table 1: Fastest algorithms for different respective problem variants.

5. RELATED WORK

There has recently been much interest in programming models and distributed system architectures for the processing and analysis of big data (e.g. [3, 11, 34]). The model studied in this paper is motivated by MapReduce [11] like batch-processing applications, also known from the popular open-source implementation *Apache Hadoop*. These applications generate large amounts of network traffic [8, 25, 36], and over the last years, several systems have been proposed which provide a provable network performance, also in shared cloud environments, by supporting relative [26, 27, 30] or, as in the case of our paper, *absolute* [5, 21, 28, 29, 33] bandwidth reservations between the virtual machines.

The most popular virtual network abstraction for batch-processing applications today is the *virtual cluster*, introduced in the Oktopus paper [5], and later studied by many others [25, 33]. Several heuristics have been developed to compute "good" embeddings of virtual clusters: embeddings with small footprints (minimal bandwidth reservation costs) [5, 25, 33]. The virtual network embedding problem has also been studied for more general graph abstractions (e.g., motivated by wide-area networks). [9, 15]

From a theoretical perspective, the virtual network embedding problem can be seen as a generalization of classic VPN graph embedding problems [19, 22], in the sense that in virtual network embedding problems, also the embedding endpoints are flexible and subject to optimization. In this sense, the virtual network embedding problem can also be seen as a generalization of the classic NP-hard Minimum Linear Arrangement problem which asks for the embedding of guest graphs on a simple *line topology* (rather than tree-like topologies as studied in this paper) [12, 13].

However, to the best of our knowledge, we are the first to provide an algorithmic study of the virtual cluster embedding problem which takes into account data locality as well as the possibility to select replicas—aspects which so far have only been studied from a best-effort perspective and using coarse-grained metrics (e.g., same rack or same server), thus limiting the flexibility of the system [4, 23, 35].

6. SUMMARY AND CONCLUSION

At the heart of locality and replica aware virtual cluster embeddings lie fundamental algorithmic problems. This paper has shown that despite the multiple dimensions of flexibility in terms of chunk assignment and node placement, many problems can be solved in polynomial time. However, we have also shown that several embedding problems are NP-hard already in two- and three-level trees—a practically relevant result given today's datacenter topologies [2])—and even if the the number of replicas is bounded by two.

Our results are summarized in Figure 1. One interesting takeaway from this figure regards the question which properties render the problem NP-hard. For instance, we see that, BW does not influence the hardness of any problem variant, while RS is crucial for NP-hardness. MA only affects hardness if combined with RS. NI is trivial without FP, and FP requires more sophisticated algorithms when combined with NI or MA; in combination with RS and MA or NI, FP renders the problem NP-hard.

We believe that our paper opens several interesting directions for future research. In particular, it would be interesting to shed light on the possible approximation algorithms for our NP-hard problems.

7. REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc., 1993.

[2] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In *Proc. ACM SIGCOMM*, pages 63–74, 2008.

[3] I. Alagiannis, R. Borovica, M. Branco, S. Idreos, and A. Ailamaki. Nodb: Efficient query execution on raw data files. In *Proc. ACM SIGMOD*, pages 241–252, 2012.

[4] G. Ananthanarayanan, S. Agarwal, S. Kandula, A. Greenberg, I. Stoica, D. Harlan, and E. Harris. Scarlett: Coping with skewed content popularity in mapreduce clusters. In *Proc. EuroSys*, pages 287–300, 2011.

[5] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron. Towards predictable datacenter networks. In *Proc. ACM SIGCOMM*, 2011.

[6] R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. Scope: Easy and efficient parallel processing of massive data sets. *Proc. VLDB Endow.*, 1(2), 2008.

[7] M. Chowdhury, S. Kandula, and I. Stoica. Leveraging endpoint flexibility in data-intensive clusters. In *Proc. SIGCOMM*, pages 231–242, 2013.

[8] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica. Managing Data Transfers in Computer Clusters with Orchestra. In *Proc. ACM SIGCOMM*, 2011.

[9] N. M. K. Chowdhury and R. Boutaba. A survey of network virtualization. *Comput. Netw.*, 54(5):862–876, 2010.

[10] P. Crescenzi, V. Kann, M. Halldorsson, M. Karpinski, and G. Woeginger. Maximum 3-dimensional matching. *A Compendium of NP Optimization Problems*, 2000.

[11] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proc. USENIX OSDI*, pages 137–150, 2004.

[12] N. R. Devanur, S. A. Khot, R. Saket, and N. K. Vishnoi. Integrality gaps for sparsest cut and minimum linear arrangement problems. In *Proc. ACM STOC*, 2006.

[13] J. Díaz, J. Petit, and M. Serna. A survey of graph layout problems. *ACM Comput. Surv.*, 34(3):313–356, 2002.

[14] R. Duan and H.-H. Su. A scaling algorithm for maximum weight matching in bipartite graphs. In *Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1413–1424, 2012.

[15] A. Fischer, J. Botero, M. Beck, H. DeMeer, and X. Hesselbach. Virtual network embedding: A survey. 2013.

[16] H. Gabow. A scaling algorithm for weighted matching on general graphs. In *Proc. IEEE FOCS*, 1985.

[17] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. In *Proc. ACM SOSP*, pages 29–43, 2003.

[18] A. V. Goldberg and R. E. Tarjan. Finding minimum-cost circulations by canceling negative cycles. *J. ACM*, 36(4):873–886, 1989.

[19] N. Goyal, N. Olver, and F. B. Shepherd. The VPN conjecture is true. *Proc. ACM STOC*, 2008.

[20] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: a scalable and flexible data center network. In *Proc. ACM SIGCOMM*, 2009.

[21] C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang. SecondNet: A data center network virtualization architecture with bandwidth guarantees. In *Proc. 6th CoNEXT*, 2010.

[22] A. Gupta, J. Kleinberg, A. Kumar, R. Rastogi, and B. Yener. Provisioning a virtual private network. In *Proc. ACM STOC*, New York, New York, USA, 2001.

[23] M. Isard, V. Prabhakaran, J. Currey, U. Wieder, K. Talwar, and A. Goldberg. Quincy: Fair scheduling for distributed computing clusters. In *Proc. ACM SOSP*, pages 261–276, 2009.

[24] Z. Kiraly and P. Kovacs. Efficient implementations of minimum-cost flow algorithms. In *ArXiv Technical Report 1207.6381*, 2012.

[25] J. C. Mogul and L. Popa. What we talk about when we talk about cloud network performance. *ACM SIGCOMM CCR*, 2012.

[26] L. Popa, A. Krishnamurthy, S. Ratnasamy, and I. Stoica. Faircloud: Sharing the network in cloud computing. In *Proc. HotNets-X*, 2011.

[27] L. Popa, P. Yalagandula, S. Banerjee, J. C. Mogul, Y. Turner, and J. R. Santos. Elasticswitch: Practical work-conserving bandwidth guarantees for cloud computing. In *Proc. ACM SIGCOMM*, pages 351–362, 2013.

[28] B. Raghavan, K. Vishwanath, S. Ramabhadran, K. Yocum, and A. C. Snoeren. Cloud control with distributed rate limiting. In *Proc. SIGCOMM*, 2007.

[29] H. Rodrigues, J. R. Santos, Y. Turner, P. Soares, and D. Guedes. Gatekeeper: Supporting bandwidth guarantees for multi-tenant datacenter networks. In *Proc. 3rd Conference on I/O Virtualization (WIOV)*, 2011.

[30] A. Shieh, S. Kandula, A. Greenberg, and C. Kim. Seawall: Performance isolation for cloud datacenter networks. In *Proc. USENIX HotCloud*, 2010.

[31] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In *Proc. IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10, 2010.

[32] E. Tardos. A strongly polynomial minimum cost circulation algorithm. *Combinatorica*, 5(3):247–255, July 1985.

[33] D. Xie, N. Ding, Y. C. Hu, and R. Kompella. The only constant is change: incorporating time-varying network reservations in data centers. *ACM SIGCOMM Computer Communication Review (CCR)*, 2012.

[34] R. S. Xin, J. Rosen, M. Zaharia, M. J. Franklin, S. Shenker, and I. Stoica. Shark: Sql and rich analytics at scale. In *Proce. ACM SIGMOD*, pages 13–24, 2013.

[35] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica. Delay scheduling: A simple technique for achieving locality and fairness in cluster scheduling. In *Proc. EuroSys*, pages 265–278, 2010.

[36] Measuring EC2 system performance. http://goo.gl/V5zhEd.

APPENDIX

A. REPLICA SELECTION IS HARD

We have seen that replica selection flexibilities can render embeddings computationally hard. In the following, we will explore the minimal requirements for rendering replica selection hard. In particular, we will show that already two replicas for each chunk type are sufficient to introduce intractability.

In the following, we allow the number of chunk types to be smaller than the number of nodes. The "idle" nodes however do participate in the inter-connect communication (in practical terms: in the shuffle phase and the reducing phase).

We first give an alternative proof for the hardness of NI + FP + RS + BW. Subsequently, we show that the 2-replica variant RS(2) is hard.

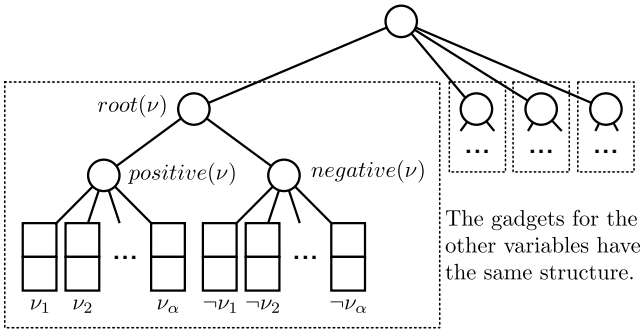

Figure 12: The construction of the variable gadget for ν. If ν appears in the first clause, a chunk c_1, will be located at ν_1. If $\neg\nu$ satisfies the last clause, $\neg\nu_\alpha$ will host c_α.

We prove that NI+RS+FP+BW is NP-hard by reduction from the *Boolean Satisfiability Problem* (SAT). Since SAT is a decision problem, we introduce a cost threshold Th to transform NI + RS + FP + BW into a decision problem too.

Let's first recall that the SAT problem asks whether a positive valuation exists for a formula Ψ with α clauses and β variables. In the following, we will only focus on SAT instances of at least four variables; this SAT variant is still NP-hard.

Construction. Given any formula Ψ in *Conjunctive Normal Form (CNF)* with α clauses and $\beta \geq 4$ variables, we produce a RS+FP+BW instance as follows: First, we construct a substrate tree T_Ψ, consisting of a root and separate gadgets for each variable of Ψ, each of which is a child of the root. The gadget of variable ν consists of $root(\nu)$ and its two children: $positive(\nu)$ and $negative(\nu)$. Child $positive(\nu)$ has α many children labeled $\nu_1, \nu_2, \ldots, \nu_\alpha$, and child $negative(\nu)$ has α many children labeled $\neg\nu_1, \neg\nu_2, \ldots, \neg\nu_\alpha$. Every gadget has the same structure: the same height and the same number of leaves. This construction is illustrated in Table 12.

For all variables ν, we set the bandwidth on the link, connecting $root(\nu)$ to the root of the substrate tree, to $\alpha \cdot (\alpha \cdot \beta - \alpha)$. The bandwidth on the other edges is not limited.

We set the number of nodes to $n_V = \alpha \cdot \beta$. Moreover, we define the inter-connect communication cost to be 1, and the access cost to be a sufficiently large constant W, such that nodes must always be collocated with chunks ($W = Th + 1$ is sufficient).

We set the number of chunk types to be equal to the number of clauses, $\tau = \alpha$. To finish our construction, we place data chunks at leaves, as follows: for the i-th clause we construct as many replicas of chunk c_i as there are literals in the clause. For each literal ℓ (of the form ν or $\neg\nu$) that satisfies clause i, we place a replica of chunk c_i in the leaf labeled ℓ_i.

Note, that in this construction some nodes will be idle. No chunks will be assigned to these nodes, but they will nevertheless participate in the node interconnect.

We set the threshold Th to: $Th = \beta \cdot (\binom{\alpha}{2} \cdot 2 + \alpha \cdot (\alpha \cdot \beta - \alpha))$.

Proof of correctness of construction. We now show that our construction indeed decides SAT. We set the capacities such that in every gadget, at most α nodes can be mapped, where α is the number of clauses of Ψ . We can apply the Bandwidth Lemma (Lemma 11) as follows: We interpret a_i as the number of nodes that are embedded in the i-th gadget, α as the number of clauses, and β as the number of variables. The LHS of the inequality of Lemma 11 is a formula for the communication cost of nodes inside the i-th gadget to nodes outside the gadget. The RHS of the inequality is the bandwidth constraint for the gadget. This implies that any feasible solution must embed exactly α nodes in every gadget. Recall that in our SAT instance, we have at least four variables.

THEOREM 6. *The problem* RS + FP + BW + NI *is NP-hard.*

PROOF. We will prove that formula Ψ is satisfiable iff RS + FP + BW + NI has a solution of cost $\leq Th$.

($\Rightarrow$) Let us take any valuation Val that satisfies Ψ. We will construct a solution to RS + FP + BW + NI using Val in the following way. For each variable ν in Ψ, we embed α many nodes at the leaves of the gadget of ν. We need to choose α out of $2 \cdot \alpha$ leaves to embed nodes. If $Val(\nu) = 1$, we embed nodes at the leaves of $positive(\nu)$, else we embed all nodes at leaves $negative(\nu)$. The solution constructed this way has cost exactly Th, because the nodes are evenly split among gadgets, and nodes are not distributed across $positive(\nu)$ and $negative(\nu)$ subtrees.

We calculate the chunk-node matching μ by assigning every chunk to the node which is collocated with the first chunk replica. This solution is feasible because every clause of Ψ was satisfied and chunks correspond to clauses.

Now we will show that this solution has cost Th. Due to the Bandwidth Lemma (Lemma 11), we only have to consider the communication cost. We sum inner-gadget communication and communication among gadgets to get exactly Th.

($\Leftarrow$) Let us take any solution to RS+FP+BW constructed based on Ψ of cost $\leq Th$. We will construct a positive valuation Val by considering the nodes in the solution to RS + FP + BW.

We make the following observations. In every solution of cost $\leq Th$, every gadget has exactly α many nodes at its leaves. This is due to the Bandwidth Lemma (Lemma 11). Also, inside every gadget either all nodes are in the $positive(\nu)$ subtree of variable ν, or in the $negative(\nu)$ subtree. This is true because the cost of a solution where at least one gadget has nodes distributed across subtrees is always greater than Th.

Now we can construct our valuation Val, as follows (for each variable ν in Ψ): If v_1 hosts a node then $Val(\nu) = 1$, otherwise $Val(\nu) = 0$.

The valuation Val satisfies all clauses, and hence Ψ, as the solution to RS + FP + BW covers all chunks. To see this, consider the leaf which hosts a node which is assigned to any given chunk (i.e., the leaf handling any given clause chunk); it is a witness that the corresponding clause is satisfied. □

We conclude by observing that our construction leverages the fact that the number of nodes may exceed the number of chunk types, e.g., for a clause $(x \vee y \vee z)$ in Ψ, both x and y being true implies the mapping of nodes on vertices labeled x_1 and y_1, and which contain the same chunk c_1.

A.1 Two replicas are hard (RS(2)+FP+NI+BW)

Our results so far indicate that dealing with replication can be challenging. However, all our hardness proofs con-

cerned scenarios with three replicas, which raises the question whether the problems are solvable in polynomial time with a replication factor of 2. (Similarly to, say, the 2-SAT problem which is tractable in contrast to 3-SAT.)

In the following, we show that this is not the case: the problem remains NP-hard, at least in the capacitated network.

The proof is by reduction from 3-SAT. Given a formula Ψ in conjunctive normal form, consisting of α clauses and β variables, we construct a problem instance and substrate tree T_Ψ using two types of gadgets: gadgets for variables and gadgets for clauses. *Nota bene:* unlike in the previous proof we will create three chunk types instead of just one, for every clause.

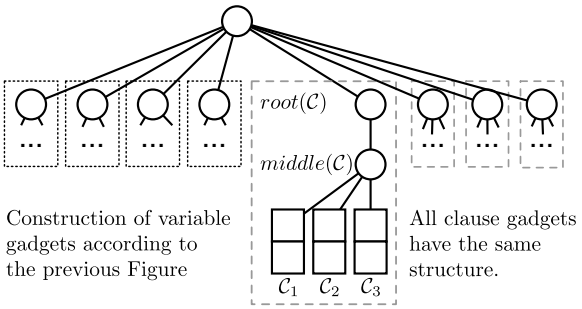

Figure 13: Structure of clause gadgets.

Construction. We build upon the construction for variable gadgets introduced (see also Figure 12).

1. *Tree Construction*: In addition to the variable gadgets known from the previous construction, we introduce *clause gadgets*. The clause gadget for a clause $\mathcal{C}$ (illustrated in Figure 13) has two inner vertices: $root(\mathcal{C})$, $middle(\mathcal{C})$[1] and three leaves $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$. We connect leaves to the middle vertex, and the middle vertex to $root(c)$. We attach the gadget to the tree by linking directly the global root to $root(\mathcal{C})$. We construct our tree out of β variable gadgets and α clause gadgets.
2. *Chunk Distribution:* For each clause $\mathcal{C}$, we generate 3 chunks types with 2 replicas each. Each server in the clause gadget of $\mathcal{C}$ holds a replica of a different chunk type. The remaining replicas of chunk types, are placed in the variable gadgets of the variables which satisfy the clause, similarly to the previous proof. Thus, in total, $6 \cdot \alpha$ variable chunks are distributed in the substrate network. We will consider a setting where $\alpha \cdot \beta + 2\alpha$ nodes need to be mapped. Our intention is that in every variable gadget, there will be α nodes, and in every clause gadgets there will be two nodes.
3. *Bandwidth Constraints:* The available bandwidth of the top edge of the gadget of each variable ν is set to $cap(\nu) = \alpha(\alpha(\beta - 1) + 2 \cdot \alpha)$. This value, results from α nodes in the gadget for ν, which each have to communicate to $\alpha \cdot (\beta - 1)$ nodes in other variable gadgets and $2 \cdot \alpha$ nodes in clause gadgets. The available bandwidth for the top edge of each clause gadget is set to $cap(\alpha) = 2(\alpha \cdot \beta + 2(\alpha - 1))$. This value allows both of the 2 nodes in a clause gadget, to communicate to the $\alpha \cdot \beta$ nodes in variable gadgets and the the other $2(\alpha - 1)$ nodes in clause gadgets.
4. *Additional Properties:* We set the threshold in a similar fashion as in previous proofs. That is, the threshold depends on the intra-clause communication cost (2 hops), and the inter-clause communication cost (6 hops). We set the number of nodes to be placed to $\alpha \cdot \beta + 2 \cdot \alpha$. We set the hosting capacity of each server to 1, and set $b_1 = Th + 1$ to disallow remote chunk access. We set $b_2 = 1$.

[1]The only purpose of the middle vertex is to maintain the balanced tree property.

Proof of correctness. We first prove the following helper lemma.

LEMMA 7. *Every valid solution to* FP+RS(2)+BW+NI *with cost at most Th has the property that there are exactly α nodes in each of the β variable gadgets and exactly two nodes in each of the α clause gadgets.*

Correctness follows from the extended bandwidth lemma (Lemma 12).

THEOREM 8. FP + RS(2) + BW + NI *is NP-hard.*

PROOF. We show that FP + RS(2) + BW + NI has a solution of cost $\leq Th$ if and only if $\Psi \in$ 3-SAT is satisfiable.

($\Rightarrow$) If we have a positive valuation of Ψ, we fill variable gadgets with nodes like in the proofs before. Then we place 2 nodes in each of the α clause gadgets as follows: Given a clause $\mathcal{C} = \ell_1 \vee \ell_2 \vee \ell_3$, we pick an arbitrary literal which satisfies the clause. Subsequently we place nodes at the leaf nodes in the clause gadget, which correspond to the other two literals. This strategy ensures that all chunks can be assigned to collocated nodes, as the only chunk type, which cannot be assigned to a collocated node in the clause gadget, has a node collocated with its second replica in the variable gadget.

We will then assign chunks to nodes in the following way: For chunk type we assign the replica in the variable gadgets to a collocated node. If this node does not exist, we assign the replica in the clause gadgets, to its collocated node.

Thus, we have produced a feasible solution of cost Th.

($\Leftarrow$) Let us take any solution SOL to RS(2) + FP + BW of cost $\leq Th$. Similar to the proof of Theorem 6 and Lemma 11 all nodes which are placed in a variable gadgets, will be located in either the *positive* or the *negative* subtree. Then we can compute a positive valuation by setting each variable ν as follows: $Val(\nu) = 1$ iff there is a node at the first leaf on the positive side of the ν gadget in SOL, and $Val(\nu) = 0$ otherwise.

The theorem now follows from the following two additional lemmas.

LEMMA 9. *For every clause there exists a node in a variable gadget that processes one of three chunks that correspond to that clause.*

PROOF. Each of the three chunks that correspond to each clause, is assigned a collocated node. At least one of those three nodes is not idle in a variable gadget; otherwise, those two nodes in the clause gadgets would not suffice in satisfying all chunk types. □

Observe that it might happen that in SOL, two nodes in clause variables are idle, and three nodes in variable gadgets are processing those 3 chunk types. In this case, arbitrary nodes can be taken for the rest of the proof.

LEMMA 10. *Val satisfies* Ψ.

PROOF. Let us consider the matching M of SOL, and let us consider an arbitrary clause of Ψ as well as its three chunk types: Due to bandwidth constraints, at most two of the chunks types, can be processed by nodes in the clause gadgets. We identify any chunk type, which is not assigned to a replica in the clause gadgets. The processed replica of that chunk type was located in a variable gadget. Depending on whether the replica was located in the positive or the negative subtree, we set the value of the according variable to 1 (positive subtree) or 0 (negative subtree). □

B. THE BANDWIDTH LEMMAS

LEMMA 11 (BANDWIDTH LEMMA). *Let α and $\beta > 4$ be two arbitrary positive integers. Let $a_1, a_2, \ldots, a_\beta$ be a sequence of β integers which adds up to $\alpha \cdot \beta$. Also, for each i we have $a_i \leq 2 \cdot \alpha$. Then it holds that if*

$$\forall_i : \quad a_i \cdot (\alpha \cdot \beta - a_i) \leq \alpha \cdot (\alpha \cdot \beta - \alpha),$$

then for each i: $a_i = \alpha$.

PROOF. By contradiction. Let us assume that there exists an index k such that $a_k \neq \alpha$. Then we can distinguish between two cases: either $a_k < \alpha$ or $a_k > \alpha$.

Case $a_k < \alpha$: If there exists a k with $a_k < \alpha$, due to the fact that the sequence adds up to $\alpha \cdot \beta$, there must also exist a k' such that $a_{k'} < \alpha$ (by a simple pigeon hole principle). Thus, this case can also be reduced to the second case (Case $a_k > \alpha$) proved next.

Case $a_k > \alpha$: Since it also holds that $a_k < 2\alpha$, a_k must be of the form $\alpha + x$ for $x \in [1, \ldots, \alpha]$. Let us consider the (bandwidth) inequality:

$$(\alpha + x) \cdot (\alpha \cdot \beta - \alpha - x) \leq \alpha \cdot (\alpha \cdot \beta - x)$$

This can be transformed to:

$$0 \leq x(x - (\alpha \cdot (\beta - 2)))$$

The equation holds for $x \leq 0$ or $x \geq \alpha \cdot (\beta - 2)$, and no positive $x \leq \alpha$ can satisfy this inequality for $\beta > 4$. Contradiction. □

LEMMA 12 (EXTENDED BANDWIDTH LEMMA). *Let α and $\beta > 4$ be two arbitrary positive integers. Let $a_1, a_2, \ldots, a_\alpha$ and $b_1, b_2, \ldots, b_\beta$ be two sequences of integers (numbers of nodes in clause and in variable gadgets). The sum of all elements in a and b adds up to $\alpha \cdot \beta + \alpha \cdot 2$ (number of nodes). Also we have $a_i \leq 2 \cdot \alpha$ (variable gadget node hosting capacity – equal to number of leaves), and $b_i \leq 3$ (clause gadget node hosting capacity). If uplink of variable gadget does not exceed bandwidth constraints*

$$\forall_{i \leq \beta} : \quad b_i \cdot (\alpha \cdot \beta + 2 \cdot \alpha - b_i) \leq \alpha \cdot (\alpha \cdot \beta - \alpha + 2 \cdot \alpha),$$

and uplink of clause gadget does not exceed bandwidth constraints

$$\forall_{i \leq \alpha} : \quad a_i \cdot (\alpha \cdot \beta + 2 \cdot \alpha - a_i) \leq 2 \cdot (\alpha \cdot \beta - 2 \cdot \alpha - 2),$$

then for each $i \leq \beta$: $b_i = \alpha$ and for each $i \leq \alpha$: $a_i = 2$ (we have expected number of nodes in variable and clause gadgets).