Customizable network update planning in SDN☆

Shouxi Luo a,∗, Hongfang Yu b, Long Luo b, Lemin Li b

a Southwest Jiaotong University, Chengdu, 611756, PR China
b University of Electronic Science and Technology of China, Chengdu, 611731, PR China

ARTICLE INFO

Keywords:
SDN
Network update
Congestion-free
Rate-limiting

ABSTRACT

Updating network configurations responding to dynamic changes is an error-prone task in SDN. During the update process, in-flight packets might misuse different versions of rules, and “hot” links could be overloaded due to the unplanned update order. As for the problem of misusing rules, recently proposed suggestions like two-phase mechanism and Customizable Consistency Generator (CCG) have provided generic and customizable solutions. Yet, there does not exist an approach that is flexible to avoid the transient congestion on hot links respecting to diverse user requirements like guaranteeing update deadline, managing transient throughput loss, etc.; controllers are in urgent need of such a solution.

In this paper, we propose CUP, Customizable Update Planner, to seek the solution. Different from prior approaches that adopt fixed designs for a single purpose like optimizing the update speed (e.g., Dionysus) or avoiding congestions (e.g., zUpdate, SWAN), CUP introduces generic linear programming models to formulate user-specified requirements and the corresponding update planning problem. By solving these customized models, CUP is able to plan network updates according to a large fraction of user requirements, such as guaranteeing deadlines, prioritizing operation orders, managing throughput loss, etc., while avoiding transient congestion. We prototype CUP on Ryu and employ it to arrange updates for networks built upon Mininet. Results confirm the flexibility of CUP while indicating that it always obtains the “best” update plans following the user's wish.

1. Introduction

Reconfiguring forwarding rules in networks responding to dynamic demands such as periodical traffic optimization, unexpected failover, is always an error-prone task for operators (Luo et al., 2016; Raza et al., 2011; Liuet al., 2013; Jinel al., 2014; Luo et al., 2015a; Luo et al., 2015b; Reitblattel al., 2012; Luo et al., 2017). Recent trends toward Software Defined Networking (SDN) seem to provide a promising solution for network management—with a logical central controller, operators can directly operate the forwarding rules on all switches. Even so, the network is still an asynchronous system in essence. It is difficult to synchronize the changes to flows from different ingress switches. Therefore, when migrating a group of flows to their new paths, even if the network is safe both before and after the reconfiguration, some “hot” links could be overloaded during the update process in case new flows move in before those old ones move out (Liuet al., 2013; Jinel al., 2014; Luo et al., 2015a).

As an example, consider the toy case shown in Fig. 1. On executing WAN optimizations (Jinel al., 2014), the controller wants to update the network's configuration from Fig. 1a to b. For simplicity, we assume that the network uses tunnel-based routing and all necessary tunnels have already been established. If the controller carries out the update in one-shot, link S4-S3 or S1-S3 might be overloaded during the update, corresponding to the case that switch S4 happens to change F3 to its new path before S1 moving F1 away from link S4-S3, or vice versa. The congestion cannot be evaded by simply letting F1 and F3 be switched to their new paths at exactly the same time (Mizrahiet al., 2015)—because the incoming packets of F2 and F3, together with the in-flight packets of F1, could still congest S4-S3 until F1 drains; and so does S1-S3.

☆ The preliminary version of this paper titled “Arrange Your Network Updates as You Wish” is published in the IFIP Networking 2016 Conference (Luo et al., 2016). In this extended version, we mainly add the following work. We 1) give a proof of Theorem 1, 2) present more design rationales about why CUP adopts two-phase mechanism to achieve consistency, 3) show the detailed design of the LP-based heuristic algorithm employed by CUP, 4) add evaluations about the efficiency of CUP, and 5) discuss more about related work.

∗ Corresponding author.
E-mail addresses: sxluo@swjtu.edu.cn (S. Luo), yuhfluuestc.edu.cn (H. Yu), longvslong@gmail.com (L. Luo), lml@uestc.edu.cn (L. Li).
https://doi.org/10.1016/j.jnca.2019.05.007
Received 8 December 2018; Received in revised form 2 March 2019; Accepted 9 May 2019
Available online 14 May 2019
1084-8045/© 2019 Elsevier Ltd. All rights reserved.
Such a type of congestion disappears following the completion of update, but its destructibility lasts long—burst traffic leads to serious queuing delay, and even, packet drops, which will let involved TCPs’ windows collapse, or even worse, kill connections. These bad influences are not desirable, especially for real-time applications. Accordingly, carrying out network reconfigurations without introducing transient congestion is a fundamental function required by SDN controller.

Planning network updates to avoid transient congestion is not an easy task. Recent approaches like zUpdate (Liu et al., 2013) and SWAN (Hongel et al., 2013) try to solve the problem by introducing a sequence of intermediate configurations, among which, the update from a former stage to the latter must always be congestion-free. To ensure such a stage sequence exists, they require part of the link capacity to be left vacant, which results in a great waste of link capacities (Hongel et al., 2013; Zheng et al., 2015). Furthermore, the intermediate configurations they introduce will greatly complicate the update process, and might even disturb user’s QoS—e.g., an intermediate path might have a larger latency than both the initial and target ones. In contrast, Dionysus (Jinel et al., 2014) and ATOMIP (Luo et al., 2015a) address the challenges by scheduling updates in thoughtful orders without bringing in additional stages. For instance, by executing the update illustrated in Fig. 1 following the 3-round sequence of $[F4 \rightarrow F1 \rightarrow F3]$, no link will be overloaded and no extra paths are introduced. Order arrangement provides a more practical solution. However, it is not always the panacea because such a congestion-free operating sequence does not always exist.

Indeed, due to the various update scenarios and user demands that a controller would deal with, simply arranging the update operations, or introducing intermediate stages, is far from enough for a practical solution. We argue that, a practical planner should have these properties.

1) **Effective to handle deadlock and deadline.** First of all, the planner must be able to find feasible congestion-free solutions for any given task. On one hand, in some update scenarios, there does not exist a congestion-free sequence (Jinel et al., 2014; Liu et al., 2015a). For instance, in the case of Fig. 1, if the demand of either F1 or F3 increases to 6, it is impossible to migrate the network to its target routing state by arranging the execution order without overloading S1-S3 or S4-S3. This is a deadlock in update planning. On the other hand, even though congestion-free schedules are found, they may not meet the deadline requirements. This is because to remove overloads, the controller cannot switch flows belonging to round-$(i + 1)$ to their new paths until flows moved out from these paths in round-$i$ have exited. Suppose in-flight packets require about $r$ units of time to exit from a path on average; then, it would take about $k \cdot r$ for the entire network to perform a $k$-round update. Such an update delay/duration might be unacceptable for time-critical cases like failover routing (Liu et al., 2014). Therefore, on planning updates, the planner should have the ability to break deadlocks and guarantee deadlines.

Fortunately, for any update, by limiting the rates of some flows at their senders or traffic shapers, controllers can always obtain a congestion-free update sequence that involves fewer rounds and satisfies the deadline requirements. Indeed, there is a trade-off between the time an update takes, and the throughput the network has to drop (induced by congestion or rate-limiting). For example, one can carry out the update request demonstrated in Fig. 1 within 2 rounds by limiting the rate of either F2 or F4 to 0 (e.g., when F2’s rate is limited to 0, $[F3 \rightarrow F1, F4]$ is congestion-free), or even perform the update within 1 round by limiting the rates of both F2 and F4 to 0. This example gives us a valuable insight: the planner should have the ability to trade throughput loss for update speed.

2) **Expressive to deal with user-specified requirements.** As infrastructure, today’s network is shared by numerous customers while simultaneously carrying various kinds of traffic. To be a universal tool for controller, the update planner should be extensible and easy to adapt to user-specified requirements (aka intents). As an example, consider the case of removing transient congestion for the update illustrated in Fig. 1 again. Provided the reconfiguration is time-sensitive and required to complete within 1 round, the controller has to reduce some flow rates to avoid congestion. Suppose this is an instance of inter-datacenter traffic optimization (Hongel et al., 2013), in which both F1 and F3 are interactive traffic while F2 and F4 are background traffic, and the operator prefers to minimize the amount of interactive traffic disturbed by the update. In such a scenario, the planner should temporarily reduce the rates of F2 and F4 to 0 to execute the update, i.e., limit the rates of $(F1, F2, F3, F4)$ to $([5, 0, 5, 0])$. On the contrary, if F2 and F4, instead of F1 and F3, are interactive, the result would be $([1, 4, 1, 4])$. As another example, if all flows share the same class and a fairness alike policy is expected (Lamel et al., 2012), the planner should set their rates to $([\frac{5}{7}, \frac{4}{7}, \frac{2}{7}, \frac{1}{7}])$, with the target of letting the decrease of throughput be fairly shared in proportion.

Indeed, due to network’s diversity, such a special constraint of rate-limiting is only the tip of an iceberg. In practice, there are plenty more kinds of user-specified demands (about the update execution time or throughput loss) that a controller would deal with. It follows that, on planning rate-limiting schemes, the planner should be flexible enough to suit various update scenarios, as well as user-specified demands.

3) **Efficient to scale up.** Last but not least, to be practical, the planner must be time-efficient to find feasible solutions for update requests in time. In consideration of that the size of today’s network might be really huge (e.g., Datacenter or backbone), the planner needs to easily scale up.

As the first step, this paper proposes CUP, Customizable Update Planner, to help controller deal with various updating requirements.

---

**Table 1**

<table>
<thead>
<tr>
<th>#Proposal</th>
<th>Introduce intermediate status?</th>
<th>Effectiveness</th>
<th>Expressiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Handle deadlock</td>
<td>Deal with deadline</td>
</tr>
<tr>
<td>zUpdate (Liu et al., 2013)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>SWAN (Hongel et al., 2013)</td>
<td>Yes</td>
<td>Yes</td>
<td>Single deadline for all</td>
</tr>
<tr>
<td>GI (Zheng et al., 2015)</td>
<td>Yes</td>
<td>Yes</td>
<td>Single deadline for all</td>
</tr>
<tr>
<td>Dionysus (Jinel et al., 2014)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ATOMIP (Luo et al., 2015a)</td>
<td>No</td>
<td>Yes</td>
<td>Single deadline for all</td>
</tr>
<tr>
<td>CUP</td>
<td>No</td>
<td>Yes</td>
<td>Per-flow deadline</td>
</tr>
</tbody>
</table>

---

**Fig. 1** A network update example. Each link has 10 units of capacity and flows are labeled with their sizes. If the controller carries out the update in one-shot, link S1-S3 or S4-S3 will be overloaded during the update.
CUP suggests adopting generic methods such as two-phase mechanism [7, 6] to enforce rule consistency, and focuses on eliminating the transient congestion during updates. Distinguished from existing solutions proposed for fixed targets, CUP is effective and expressive to deal with deadlock, deadline, prioritization, and many other user-specified requirements as Table 1 summaries (Note that, proposals focusing on enforcing rule consistency are not listed, e.g., CCG (Zhou et al., 2015).). We analyze various demands and realize that, besides consistency, what users/operators concern about the implementation of an update, no matter how complex it is, generally involves two types of fundamental issues—i) when a flow could take advantage of its new path(s) and ii) how its throughput would be impacted during the update process.

At a high-level, CUP provides an expressive user-friendly language, with which, customers and operators can describe their own requirements easily and explicitly. When the network is to be updated, CUP maps these high-level requirements into the essence (involved) flows, and translates them into low-level linear constrains. At its core, CUP builds a couple of generic linear programming models to formulate the update request while capturing constrains from users. Via solving these customized models, CUP obtains a congestion-free update execution plan that explicitly follows the user’s wish.

Roughly, CUP’s model involves two parts, Order Scheduler and Rate Manager, which respectively answer the two basic problems mentioned above. On planning an update, Order Scheduler first determines the operation order respecting to time-related requirements. If congestion-free sequences are found, Order Scheduler outputs the one involving the minimum rounds; otherwise, it chooses the sequence causing least overload on links. For the overloaded traffic, Rate Manager then figures out the optimal rate-limiting scheme that is able to erase the congestion while satisfying all throughput-related requirements. As the core of both Order Scheduler and Rate Manager is to solve a single Linear Program (LP), with high performance LP solvers, CUP obtains solutions within polynomial time and is able to scale up.

We prototype CUP upon Ryu (An open-source SDN control) and use it to plan updates for networks conducted by Mininet (Händel et al., 2012). Results show that CUP is quite flexible to exactly meet user-specified requirements, while effective to outperform existing approaches.

In summary, we make three contributions in this paper.

- **Abstraction**: We show how to express various user-specified updating requirements with a high-level language, and show how to dynamically translate them into low-level linear constraints (Section 2).
- **Model**: We propose generic linear programming models to formulate and solve the customized update planning problem, with which, controllers obtain the “best” update plan explicitly following user’s wish (Section 3 and 4).
- **Evaluation**: We show that our CUP tool is flexible, effective, and efficient to make update plans for “real” networks built by Mininet (Section 5).

### 2. Flexible CUP

In CUP, network users as well as operators describe their desired properties about the update with the high-level CUP language; they can change the clauses at any time. On planning a network update, at the first step, CUP “compiles” the user’s codes to figure out their exact “meaning” in this instance. After that, CUP employs back-end solvers, Order Scheduler and Rate Manager, to find the update processing plan that exactly follows the user’s wish. Basically, the entire workflow of how CUP produces is as Fig. 2 shows.

In the following, we present the high-level language in Section 2.1 and show the compilation process in Section 2.2. After that, we introduce how CUP solves the planning problems in Section 3 and discuss how it handles multi-tenants and concurrent update requests in Section 4.

#### 2.1. High-level language

CUP language (Fig. 3) provides end-users and operators with an easy way to specify their requirements on configuring the network. A CUP policy is a collection of rules, in which, each term specifies a specific requirement, of either the activation time of new paths or the degradation of throughputs, for a (group of) flow(s). CUP uses a regular expression on the match fields of flow header to define the involved traffic. For instance, * defines all traffic passing through the network; dstTCP = 80 defines all web access traffic; srcIP = 10.0.0.1/24 ∧ dstIP = 20.0.0.11 defines those flows from subnet 10.0.0.1/24 to destination 20.0.0.11; and srcIP = 10.0.0.2 ∨ dstIP = 10.0.0.4 defines the traffic from 10.0.0.2, or to 10.0.0.4.

For the update of a collection of flows specified by m, there are two basic types of indicators that customers and operators might concern: i) how long it would wait before taking advantage of the new path(s), and ii) how its throughput (i.e., rate) would be limited to avoid transient congestion. CUP uses T(m) and R(m) to denote, respectively. Using their relation expressions, these two basic elements can generate other complicated requirements. For instance, T(m1) ≤ T(m2) says, flows matched with m1 should be switched into the new paths “no later than” those matched with m2, while T(m2) ≤ val indicates the waiting time before m2 switched should be “no larger than” val. Similarly, R(m1) ≤ R(m2) implies the effective bandwidth of m1 during the update should “no less than” that of m2, while R(m1) ≥ amap means the user would like the effective bandwidth of m1 be maximized.

CUP language is simple yet expressive for many requirements. As examples, revisit the toy update cases of Fig. 1. With CUP language, network operators can formulate their own requirements precisely and concisely as the instances in Table 2 illustrate.

#### 2.2. Dynamic translator

High-level CUP policies are compiled into low-level restrictions, which tell the planner how to process each flow’s reconfiguration is
in line with user requirements. To achieve this, the most challenging task is to figure out the probable time cost of migrating a flow. CUP employs the approach of pre-installing new rules then triggering two-phase reconfigurations to address the problem. In this part, we first present how to make the estimation of reconfiguration’s time cost possible in Section 2.2.1, then introduce the way of binding high-level requirements with flows and translating them into low-level linear constraints in Section 2.2.2.

### 2.2.1. Estimating time cost of traffic migration

As Section 1 and Fig. 1 have shown, to not overload any link during the update, the controller has to wait the flow that is moving out from a link exits, before moving other flows in. Thus, the time cost of migrating a flow to its new path(s) mainly involves two parts of \(i)\) waiting the moving-out traffic exits (if any); and \(ii)\) installing rules to shift the flow to its new path(s).

As for the first part of draining time, we can simply use the well-known One-Way Delay (OWD) as an approximation, which can be estimated at end hosts (Gurewitz et al., 2006; Pathakel al., 2008), or at edge switches in OpenFlow-enabled networks. CUP suggests adopting two-phase update mechanism to guarantee strong rule consistency (refer to Appendix A for the discussion). On carrying out an \(N\)-rounds flow migration, at the first step, CUP pre-installs the new configurations and sets rate-limits. Supposing the time of installing/modifying a rule from the controller is \(\epsilon\), the total time cost of this step is \(\epsilon\) because all rule installations (for both new paths and rate-limits) can perform in parallel. Thus, the rest operations for each round are to \(i)\) wait a draining time then \(ii)\) touch some flows' ingress switches to activate their new paths.

Provided the largest OWD in network is \(\tau\), we get the point that flows migrated in the \(k\)th round would take advantage of their new paths at time \(k \times \tau + (k + 1) \times \epsilon\). Consequently, if a flow's deadline requirement on the update process is \(w\), we know that the controller should make sure it get migrated no later than round \(\lfloor \frac{w}{\tau + \epsilon} \rfloor\).

In practice, the time cost of modifying a rule on physical switches is usually inconsistent (Hanel et al., 2015; Kuzniaret al., 2015; Jinelal et al., 2013; Jinel et al., 2014). Yet, recent studies have shown its long-tailed characteristic (Jinel et al., 2014). That is to say, simply choosing the 95th percentile value (or other thresholds) as the estimated time is reasonable in most cases. Moreover, since OpenFlow-style control is still in its early stages, most switch software and SDKs are not optimized for dynamic table programming yet (Jinel et al., 2013).

Some effects have been put on improving this and we argue that future switches will be more stable and fast for table changes (Hanel et al., 2015; Bifulco and Matsiuk, 2015; Chen and Benson, 2017).

As yet, we have found a way to estimate the time cost of migrating a flow based on network's maximum OWD and ingress’s rule modification delay. In real networks, both types of delays can be measured by the controller. With this information, CUP is able to translate the absolute deadline requirements into round requirements. For simplicity, hereafter, all deadline requirements we discuss in this paper are in the form of round number.

### 2.2.2. Mapping requirements to each flow

Now, we show how CUP maps user requirements into each flow. The basic notations that CUP’s model uses are summarized in Table 3.

#### Lexical analysis and preprocessing. CUP first parses user-specified policies to get the semantics. Obviously, there are four types of constraints on flow predicates, indicating the absolute update deadline (i.e., \(T(m) \leq w\)), the relative update order in “no-later-than” form (i.e., \(T(m_2) \leq T(m_1)\)), relative rate-limiting setting in “no-less-than” form (i.e., \(R(m_2) \geq R(m_1)\)), and the expected targets that should be optimized (e.g., \(R(m) \geq \text{amap}\)). Without loss of generality, we let \(M_{\text{update}}\) be the set of predicates holding the relation of \(T(m) \leq w\), and \(M_R\) be the set of predicates holding \(R(m) \geq \text{amap}\). As well, we further use \(M_P\) and \(M_R\) to denote the set of predicate pairs (e.g., \((m_2, m_1)\)) that have the relation of \(T(m_2) \leq T(m_1)\) and \(R(m_2) \geq R(m_1)\), respectively.

As discussed above, for a deadline requirement on flows specified by predicate \(m\), CUP can transfer it into a round number requirement with Equation (1), where \(\hat{\tau}\) is the network’s measured maximum OWD and \(\hat{\epsilon}\) is the measured 95th rule modification delay.

#### Basic network model

We assume that the network, \(G\), is hosting a set of flows \(F\) with links \(E\). The rate of flow \(f \in F\) is denoted by \(\tau_f\) while the capacity of link \(e \in E\) is denoted by \(c_e\). By letting \(\tau_{f,e}\) be the traffic load of flow \(f\) on link \(e\), the network’s state can be formulated as \(S = \{\tau_{f,e}\} \forall (f \in F, e \in E)\). Then, a network update is to change its state from \(S\) to \(S’ = \{\tau’_{f,e}\} \forall (f \in F, e \in E)\) by rerouting some flows,

#### Table 2

<table>
<thead>
<tr>
<th>#</th>
<th>Update scenarios</th>
<th>Policy expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Minimize transient congestion without deadline requirements on the update process.</td>
<td>((R^* \geq \text{amap}))</td>
</tr>
<tr>
<td>2</td>
<td>Let interactive flows (F_2) and (F_s) take advantage of new paths no later than 1 unit time, while minimizing the impacts on their throughputs (e.g., inter-DC WAN optimization [17, 10]).</td>
<td>((T(m_2) \vee m_1) \leq 1; R(m_2) \geq amap))</td>
</tr>
<tr>
<td>3</td>
<td>Execute all flow migrations no later than 1 unit time, and let the throughput loss be shared in proportion since they are in the same class.</td>
<td>((T^* \leq 1; R(m_s) \geq amap; R(m_3) \geq amap; R(m_4) \geq amap))</td>
</tr>
</tbody>
</table>

#### Table 3

The key notations of the network model.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M^\text{update})</td>
<td>the set of predicates ((m)) holding (T(m) \leq w)</td>
</tr>
<tr>
<td>(M_R)</td>
<td>the set of predicates ((m)) holding (R(m) \geq \text{amap})</td>
</tr>
<tr>
<td>(M_P)</td>
<td>the set of ((m_2, m_1)) pairs holding (T(m_2) \leq T(m_1))</td>
</tr>
<tr>
<td>(M_R)</td>
<td>the set of ((m_2, m_1)) pairs holding (R(m_2) \geq R(m_1))</td>
</tr>
<tr>
<td>(R^\text{due})</td>
<td>the round deadline for flow matching with (m)</td>
</tr>
<tr>
<td>(f \in F)</td>
<td>the set of all current flows in the network</td>
</tr>
<tr>
<td>(F(m))</td>
<td>the set of all flows matching with predicate (m)</td>
</tr>
<tr>
<td>(\tau_f)</td>
<td>the demand of flow (f)</td>
</tr>
<tr>
<td>(\tau_{f,e})</td>
<td>the rate-limit setting of flow (f) during the update</td>
</tr>
<tr>
<td>(\rho_e)</td>
<td>the rate-limit setting for all flows matching with (m)</td>
</tr>
<tr>
<td>(e \in E)</td>
<td>the set of all (directed) links in the network</td>
</tr>
<tr>
<td>(c_e)</td>
<td>the capacity of link (e)</td>
</tr>
<tr>
<td>(t_j)</td>
<td>the load of (f) on link (e) after the update</td>
</tr>
<tr>
<td>(t_{f,e})</td>
<td>the load of flow (f) on link (e) after the update</td>
</tr>
<tr>
<td>(p^\text{old})</td>
<td>the set of flows that will not be updated/migrated</td>
</tr>
<tr>
<td>(p^\text{new})</td>
<td>the set of flows that will be updated/migrated</td>
</tr>
<tr>
<td>(F^\text{old}(m))</td>
<td>the set of to-be-updated flows matching with (m)</td>
</tr>
<tr>
<td>(F^\text{new}(m))</td>
<td>the set of flows that will be updated/migrated</td>
</tr>
<tr>
<td>(\Psi(f, e) \in F^\text{old}(m))</td>
<td>should be updated no later than (f)</td>
</tr>
<tr>
<td>(\gamma_{f,e})</td>
<td>whether (f) has been updated in round (k)</td>
</tr>
<tr>
<td>(\gamma_{f,e}^\text{new})</td>
<td>the (maximum) load of (f) on (e) in round (k)</td>
</tr>
</tbody>
</table>

\[ \hat{\tau}_{m} = \min \left( \frac{\text{val} - \hat{\tau}}{\hat{\tau} + \epsilon}, 1 \right) \]
or changing their traffic split ratios in the case of multi-path routing. For the update of $S \rightarrow S'$, let $F^U$ be the set of updated flows and $F^R$ be the set of unmodified flows. Obviously, there must be $F^U \cap F^R = \emptyset$, $F^U \cup F^R = F$, and $\eta_{f} = \tau_{f} \cdot e$ for all $f \in F^U, e \in E$. We assume that the update of flow $f$ is required to be finished within $N_{f}^{due}$ rounds, and use binary variable $y_{f,k} \leq k \leq N_{f}^{due}$ to indicate whether $f \in F^U$ has been migrated/updated in the $k$-th round. By defining $y_{f,0} = 0$ for convenience, we get the constraints as Equations (2) and (3) show.

\[ \forall k, f \in F^U : y_{f,k} \in \{0, 1\} \]  
\[ \forall f \in F^U : 0 = y_{f,0} \leq y_{f,1} \leq \cdots \leq y_{f,N_{f}^{due}} = 1 \]

Besides, we let $r_f$ denote the proportion of rate-limiting that flow $f$ would be set to during the update. Then, after rate-limiting is enabled, the total load of $f$ would be reduced to $r_f \cdot \tau_f$, and the subpart on $e$ before and after the update would also decrease to $\eta_{f,e} \cdot r_f + \tau_{f} \cdot r_f$, respectively.

\[ \forall f : 0 \leq r_f \leq 1 \]  

**Embedding user-specified requirements.** In networks, flows are defined by predicate strings of the packet header fields. By checking whether a flow’s predicate intersects with the user-specified predicate, CUP figures out which flows are involved with that rule. For rule predicate $\mathcal{P}$, whether a flow’s predicate intersect with the user-specified predicate, defined by predicate strings of the packet header fields. By checking Equation (7) shows.

\[ \forall (f, e) \in \mathcal{F_P}, f \in F^U(m), f \in F^U(m) \]  

As for the “no-later-than” order requirements, $T(m_j) \leq T(m_j)$, if two to-be-updated flows, $f_i$ and $f_j$, happen to hold the relations of $f_i \in F^U(m_i)$ and $f_j \in F^U(m_j)$, it means they have order dependencies on the update active time, namely, $y_{f_i,k} \geq y_{f_j,k}$ for all feasible $k$. Let $FP_T$ be the set of such order-dependent flow pairs; CUP can easily get it by computing the relevant.

\[ FP_T = \{(f_i, f_j) : \exists (m_x, m_y) \in MP_T; f_i \in F^U(m_i); f_j \in F^U(m_j)\} \]  

Now, CUP deals with rate/throughput related requirements. Same to the case of time-related predicates, the predicate $m$ in a rate-specified rule also might match with multiple flows at the same time. We denote the collection of involved flows as $P(m)$ and regard them as a “virtual” aggregated flow. For this “virtual” flow, we further use $r^m_e$ to present what its rate-limit would be during the update process. Then the two types of throughput requirements could be formulated as Equations (8) and (9) show, in which $r^m_e$ is defined by Equation (10) and amap is the index/variable that should be optimized.

\[ \forall (m_i, m_j) \in MP_k : r^m_e \geq r^m_j \]  

\[ \forall m_i \in M_{R}^{amap} : r^m_e \geq amap \]  

\[ r^m_e = \sum_{f \in P(m)} r_f \cdot t_f \sum_{f \in P(m)} t_f \]  

So far, CUP has translated all user-specified requirements into low-level flow-based constraints, which are all linear.

3. **Efficient solver**

To handle various updates, CUP needs a generic yet efficient solver. However, the design is not easy since planning updates is computationally intractable in ordinary sense—even answering the question of whether there exists a congestion-free solution for a given update is NP-hard as Theorem 1 says.

**Theorem 1.** Determining whether there is a congestion-free update order scheduling that meets user-specified deadline is NP-Hard in ordinary sense.

**Proof.** The proof is quite similar to that of Theorem 2 in (Jin et al., 2014). Consider a network in which a set of integer traffic demands travel through link $e_1$ or link $e_2$, alternatively, and the capacity of both links is $c$. Initially, flows in group $G_A$ go through $e_1$, while flows in group $G_B$ go through $e_2$. Their total load are $c_A$ and $c_B$, respectively, where $c_A \leq \frac{c}{2}$ and $c_B = c$. Suppose the update is to swap their routes. Obviously, the fastest updating plan that might be congestion-free is a 3-round solution: 1) migrate a part of $G_A$ with the total load of $c - c_A$ from $e_1$ to $e_2$; 2) migrate $G_A$ from $e_1$ to $e_2$; and finally 3) migrate the rest of $G_B$ from $e_2$ to $e_1$ (with load $c_B - c_A$). However, to figure out whether this 3-round congestion-free solution exists, we have to solve the subset sum problem of finding a subset flows from $G_B$ sum to $c - c_A$, which is known as NP-complete.

Corresponding to the fact that planning an update involves two parts of i) finding an execution order and ii) computing the relevant rate-limiting scheme, CUP heuristically decouples the original problem into two parts as Fig. 2 shows. On planning a group of flow migrations, the Order Scheduler module first determines which round each flow should be moved in, based on user-specified time-related requirements. If there exists congestion-free sequences, Order Scheduler outputs the one with the minimum rounds; otherwise, it suggests the sequence causing smallest traffic overloads. Then, for the congested traffic, Rate Manager further finds the optimal rate-limiting scheme that makes the update free of congestion, respecting to throughput/time-related rules.

3.1. **Order Scheduler**

As Section 3.1.1 will show, the problem of scheduling flow migration order to reduce congestion can be formulated as a Mixed Integer Linear Program (MIP). Then, for the schedule of a small number of flows, it is possible to obtain the optimal order by directly solving this MIP with efficient solvers. However, as finding the optimization scheduling order is theoretically NP-hard, the computation process becomes quite time-consuming when the network scales up. To find scheduling orders quickly, we further relax the original MIP into a Linear Program (LP), and develop an efficient heuristic solution based on the relaxed LP’s outputs as Section 3.1.2 illustrates.

In practice, a simple way to achieve both efficiency and effectiveness on ordering scheduler is to employ a “dual-core” design. For each planning request, CUP can perform the MIP solving and heuristic computation, simultaneously. If MIP completes within a certain time (e.g., 1 s), CUP gets the optimal results; otherwise, CUP chooses the heuristic result and stops the task of MIP solving.

3.1.1. **The MIP model**

The first step of planning update to prevent transmit congestions is to evaluate what link loads would be during the update procedure. For flow $f \in F^U$, we let $\gamma_{f,e,k}$ indicate its maximum possible load on link $e$ when preforming the reconfiguration of round $k$. Then, the maximum (possible) load on link $e$ in this round is $\sum_{f \in F^U} \gamma_{f,e,k} = \sum_{f \in F^U} \gamma_{f,e,k}$. 

\[ \sum_{f \in F^U} \gamma_{f,e,k} \]
The calculation of $t_{f,e,k}$ for round $k$ has two formulations depending on $f$'s update senses as Equation (11) shows. In both formulations, it is certain that $f$s load on link $e$ equals $t_{f,e}$ if $f$ has not been migrated yet, i.e., $y_{f,k-1} = y_{f,k} = 0$, or equals $t'_{f,e}$ if its migration has completed, i.e., $y_{f,k-1} = y_{f,k} = 1$. The difference exists in the case when $f$ happens to be migrated in round $k$, i.e., $y_{f,k-1} = 0$ and $y_{f,k} = 1$, and the link is used by both $f$'s old path(s) and new path(s).

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.

In datacenter networks, the multiple paths between two end-hosts usually share the same hops and packets traveling through them are likely to experience the similar delay (Lii et al., 2013). Accordingly, usually share the same hops and packets traveling through them are used by both $f$.
The procedure of fixing the "no-later-than" relationship on DAG is similar to that of the topological sorting algorithm described by Kahn (1962). At each turn, LPHA picks a node with no incoming arrows (i.e., no-later-than requirements) from the DAG. After computing a round number that satisfies the requirements of all flows belonging to node 10, it computes the set round number of these flows. After the rounds of flows in this node are established, LPHA immediately updates the maximum possible round number for the un-scheduled flows having "no-later-than" requirements on this node. This guarantees all "no-later-than" requirements always being kept.

Finally, LPHA makes a rearrangement to remove the unused round numbers, and get the final schedule—π(\(X_j\)) (Line 19).

### 3.2. Rate Manager

Once the update order is determined, CUP gets the value of \(\{t_{f, e, k} \mid v(f, e, k)\}\). The next issue is to find a rate-limiting scheme avoiding congestion respecting to user's requirements. As defined in Section 2.2.2, \(r_f\) is the ratio that flow \(f\) should decrease to for removing transient congestions; then, the straightforward solution to obtain the optimal rate-limiting scheme for user-specified requirements is to solve the corresponding LP shown in Fig. 7.

\[
\forall e, k > 0 : \sum_{f \in F^e} r_f \cdot t_{f, e} + \sum_{f \in F^e} r_f \cdot t_{f, e, k} \leq c_e \tag{13}
\]

Note that, when no \(amap\)-based rule is specified, CUP adopts \(R(+) \geq amap\) by default, which results in minimizing the total throughput loss. In some cases, there might be multiple rate-setting schemes that obtain the same optimal \(amap\). CUP adds a tail of \(\rho \times \min t_{f, e}\) (\(\rho\) is a small positive constant) into the objective to gain the one limiting less flows.

### 3.3. Tricks for scalability

So far, we have built a generic solver made up of Order Scheduler and Rate Manager for CUP. Obviously, the core computation in both Order Scheduler and Rate Manager is solving LPs, which can be efficiently done within polynomial time by leveraging fast solvers like CPLEX and MOSEK. Consequently, the entire solver is a polynomial time approach as well. Furthermore, there are several simple yet efficacious designs that CUP can employ to simplify the model and accelerate the computation.

For instance, it is easy and possible to remove these "free" variables from the models to accelerate the speed of solvers. If a link would never be overloaded during the update, CUP can exclude its related constraints from the model safely. We call such links non-critical, and they can be determined by Equation (14) easily. Correspondingly, if a flow only encounters with non-critical links, there is no need to limit its rate. CUP can remove its constraints from the Rate Manager model. As well, if a to-be-updated flow is non-critical and does not have “no-later-than” relation with other flows, it can be migrated directly in the first round without planning computations.

\[
E_{non-crit} = \{\forall e \mid \sum_{f \in F^e} t_{f, e} + \sum_{f \in F^e} \max_{v(k)} t_{f, e, k} \leq c_e\} \tag{14}
\]

Moreover, for an update that involves a huge number of critical flows and links, an intuitive heuristic to control the model scale is to i) split it into multiple tiny scheduling tasks, ii) solve them respectively, and iii) merge these results to get the final one. However, both the split and merge process are non-trivial tasks, as flows to be updated might be binded with user-specified "no-later-than" and "no-larger-than" requirements. We leave the detailed designs as our future work.

### 4. Discussion

In this section, we give brief discussions on several concerns with CUP, i.e., i) whether there exists solutions for any given CUP policy, ii) could CUP help the controller arrange updates from multiple tenants, and iii) how CUP could handle multiple concurrent update requests.

#### 4.1. Solvability

As Fig. 3 shows, the requirements on both the updating time and flow rates that CUP is capable of describing are non-strict partial orders by design. Accordingly, constraints that users specify their updates to comply with would never conflict. For instance, if flow \(f_i\) is required to be updated no later than \(f_j\) by one rule while the converse requirement is specified by another, these two flows should be updated within the same round. Similarly, for two flows, provided that rate-related requirements ask the degree of one flow’s rate limiting to be no-more-than that of the other, these two flows could share the same rate-limiting settings. Therefore, the models involved by Order Scheduler and Rate Manager are always solvable. Actually, for any update request, it is obvious that moving all flows in one round while limiting flow rates zeros always yields a feasible planning for any given update request.

#### 4.2. Multi-tenant

In practice, a network might be shared by multiple tenants (or virtual operators) simultaneously (Sherwood et al., 2010). The requirements specified by a tenant should only impact its own updates and own traffic. In such cases, CUP would look into the tenant information when embedding policies. As for CUP’s solver, Order Scheduler is able to handle this directly because there is no difference on the subproblem of order scheduling; however, Rate Manager needs a modification as the rate management problem is a multi-objective optimization problem now—\(\max(\text{amap}_1, \text{amap}_2, \ldots, \text{amap}_n)\). Multi-objective optimization has been studied for very long time and there are so many solutions, such as scalarization, no-preference methods, priori methods, etc (Wikipedia and Multi-objective, 2015). In this paper, CUP simply adopts the approach of linearly scalarizing (Wikipedia and Multi-objective, 2015) the multiple objectives into the single objective of \(\max \sum v_w \cdot \text{amap}_w\), where \(w_0 \geq 0\) stands for the weight of the \(i\)th tenant. By simply pursuing this scalarized objective, CUP supports multi-tenant updates. We note that there is room to improve and CUP is flexible to be upgraded.

#### 4.3. Concurrent updates

In general, a “fat” update request involving many flow migrations would be planned to execute in more than one round. As the network configuration is volatile, new update request is likely to occur before the current “fat” one completes. This should be handled appropriately and immediately as some new flow migration requests might have urgent deadline requirements. CUP adopts the generic two-phase mechanism (Luo et al., 2015b) to implement the reconfiguration of each round, which naturally supports update streams. Accordingly, CUP can immediately deal with a new request by just regarding it together with these unperformed rounds as a fresh request; rule consistency is always guaranteed.
Unfortunately, this approach makes rule management installs a group of exact-match rules specifying the tunnel for consistent enforcement (Reitblatt et al., 2012; Luo et al., 2015b).

to reconfigure its tunnel weights at the ingress, so that each update is sub-tunnels respecting to tunnel weights. Then, updating a flow is only path routing, in which ingress switches split and assign a flow to its 
of backbone traffic in WAN, and use VLAN tags to implement tunnel-sends UDP packets with each other in steady rates to simulate the case default routes and tunnel rules via OpenFlow 1.3. We let end-hosts 

realizations of their types. In normal, forwarding functional rules like tunnels and 
amap (Reitblatt et al., 2012; Luo et al., 2015b).

To carry out weighted traffic splitting on Open vSwitch, the controller installs a group of exact-match rules specifying the tunnel for each microflow. 1 Unfortunately, this approach makes rule management on ingress complex as the update of a single flow might trigger the modification of a collection of microflow rules. We address the problem by using the Multiple Flow Table mechanism provided by OpenFlow switches (supports start from OpenFlow 1.1). Basically, rules in an ingress switch are either stored in Table 0 or Table 1 depending on their types. In normal, forwarding functional rules like tunnels and default routes reside in Table 1, and these microflow rules that realize traffic splits and tunnel selections, together with a lower priority all- whose action is “goto Table 1”, reside in Table 0. When a flow’s splitting weights are to be updated, the controller first installs microflow rules that implement the new weights in Table 1, then installs a high-priority wildcard rule with action “goto Table 1” into the first table to “guide” involved packets to the new weights. After that, the controller silently modifies the actions of those unmatched microflow rules in Table 0 following the new weights, then deletes the previously installed wildcard rule and microflow rules. Following this, we make rules easy to manage and guarantee the consistency property during weight recon-

1 In tests, the traffic from a host to another is equally dispersed over 20 UDP flows, and its ingress switch holds a corresponding number of microflow rules for traffic splitting. Thus, the accuracy of traffic-splitting is 0.05.

5. Evaluation

In this section, we implement a simplified CUP based on Ryu, and conduct virtual networks with Mininet to test CUP. Our results indicate that CUP is flexible enough to handle user-specified time- and throughput-requirements. Moreover, CUP is very effective. On each type of requirement, CUP always significantly outperforms the variant of Dionysus which is modified to handle that requirement type.

5.1. Implementation

We prototype CUP upon Ryu 3.26, and employ it to plan traffic migrations for toy virtual networks on Mininet 2.2 (Handigolle et al., 2012).

Network setup. When switches start up, the controller installs default routes and tunnel rules via OpenFlow 1.3. We let end-hosts send UDP packets with each other in steady rates to simulate the case of backbone traffic in WAN, and use VLAN tags to implement tunnel-based forwarding for them. We assume that the network adopts multi-path routing, in which ingress switches split and assign a flow to its sub-tunnels respecting to tunnel weights. Then, updating a flow is only to reconfigure its tunnel weights at the ingress, so that each update is consistent in essence (Reitblatt et al., 2012; Luo et al., 2015b).

To evaluate how transient congestion caused by unplanned updates would influence the traffic, we first conduct experiments for the toy update cases shown in Fig. 1. Note that all virtual hosts and switches in Mininet use the shared CPU and bandwidth resources for simulation (Handigolle et al., 2012). To avoid resource competition between them and to highlight the results, we set link bandwidth to 5 Mbps with 100 ms delay, and let port buffer size be large enough to hold all overloaded traffic. Accordingly, in the case of no congestion, the transmission delay of all old paths is about 200 ms, same to the network’s maximum OWD, and that of the new paths is about 100 ms.

Fig. 8a shows the transmission delay of packets in each flow when the controller sends the “activate the new path” commands for {F1, F3, F4} in One Shot at the 0.4 s. About 150 ms later, receivers get packets through the new paths. Obviously, the latency of packets in incoming packets of F1 together with the in-flight packets of F3 and F4 overload Link S1-S3, while F1’s in-flight packets together with the newly incoming packets of F2 and F3 overload Link S4-S3. In practice, the activation time of new rule might be distinct on switches; transient congestion happens once a flow moves in the hot link before the old in-flight packets exit. And these overloaded packets in high speed networks can be really huge, which would slowly eat up switch buffers and result in heavy packet loss (Jinel et al., 2014). As a comparison, Fig. 8b shows the case of migrating flows in order of [F4 → F1 → F3], which is the result planned by both Dionysus and CUP under the policy of (R(∗) ≥ amap). In this case, the controller
triggers flow migrations round by round, and waits the maximum OWD time (200 ms) between them. Following the plan, the update process takes about 600 ms to complete, but avoids all transient congestion.

Then, we look into the case of planning updates with time- and throughput-requirements. Provided the update request appear at the 0.4 s, and the operator wants all flows to take advantage of their new paths no later than 300 ms; that is to say, all flow migrations must be carried out within one round, and rate-limits are needed to avoid congestion. Fig. 9 and Fig. 10 show the results planned by CUP under user-specified policies \((T^+ \leq 1; R(m_{F2} \lor m_{F4}) \geq \text{amap})\) and \((T^+ \leq 1; R(m_{F2} \lor m_{F4}) \geq \text{amap})\), respectively. In the case of Fig. 9, all flows share the same importance and the operator prefers the total throughput be reduced as less as possible. With the objective function shown in Fig. 7, CUP’s Rate Manager lets the throughput loss be shared by all flows in proportion as Fig. 9b shows, where \(\frac{200}{150} = \frac{4}{3}\) stands for the flow rates observed by the sender or receiver—about \(\{0.5 \times \frac{2}{3}, 0.5 \times \frac{4}{3}, 0.5 \times \frac{4}{3}\}\). Different from Figs. 9 and 10 demonstrates the results show that F1 and F4 are background traffic while F2 and F4 are interactive whose throughput should be keep as much as possible. As the results show, CUP finds the update plan exactly following the operator’s wish. In contrast, Dionysus will handle the requirements in a rough way—completely kill F1 and F2 to avoid congestion.

5.3. CUP flexibility

To investigate the flexibility of CUP, we further employ it to plan updates for a small WAN topology (Jin et al., 2014; Zheng et al., 2015), which involves 8 nodes and 14 links as Fig. 11 illustrates. In this case, each link is assumed to have the capacity of 10 Mbps and delay of 200 ms. We consider the case of WAN optimization, where ingress switches split the traffic to a destination among its 4-shortest paths to pursue load balancing. Because of lacking real traffic matrices, we assume that all the possible paths of a source-destination share the equal weight initially, and use gravity model (Luo et al., 2015a) to synthesize the current traffic demands, which make the maximum link load be 99% in the old configuration. Then, the update scenario is to reconfigure traffic split weights to the new one that reduces the maximum link load to the minimized value, 78%. The longest path(s) in tests involves 4 links; accordingly, the network’s maximum OWD is 800 ms. For each link \(e\), we consider it as unchanged independently for flow \(f\) if \(f\) has more than one path going through \(e\) and these paths hold distinct lengths (i.e., delays).

\(^{2}\) It takes about 200 ms to pre-install new rules and wait rate-limits coming into force; then less than 100 ms is left for performing the updates.

When no update deadline is required, CUP finds a congestion-free plan involving 5 rounds without limiting flow rates, while Dionysus obtains a 6-round plan that achieves the same goal. Then, we artificially add deadline requirements to all flows and compute the proportion of network throughput that CUP, as well as Dionysus, has to abandon for congestion freedom. Numerical results indicate that CUP outperforms Dionysus about 3 × on reducing the impact of network throughput as Fig. 12 shows. CUP is excellent because its Rate Manager always obtains the optimal rate-limiting scheme respecting to user’s requirements. On the contrary, Dionysus just randomly kills some flows to move on. In addition, Dionysus would never touch the rate of the un-updated flows. But in some cases, slowing down some of them really helps.

We also study the cases that some traffic is background and the operator wishes interactive traffic be less impacted during the update. To this end, we assume that a certain percentage of traffic between each source-destination pair is background, then calculate how many round CUP, as well as Dionysus, would need to perform congestion-free reconfiguration without reducing the throughput of interactive traffic. Fig. 13 demonstrates the results. It implies that, with the proportion of background traffic increasing, the round number required by CUP rapidly decreases. And after the background traffic accounts for half of the traffic, CUP always performs congestion-free updates in one round without reducing the rates of interactive flows. In contrast, Dionysus cannot achieve this because of its unawareness of user-specified requirements. If we pre-limit the rates of background traffic to zeros, Dionysus then obtains small update rounds as CUP does. However, similar to the cases shown in Fig. 12, such a solution is far from good because too many flows are killed unnecessarily.

5.4. CUP efficiency

To understand the efficiency of CUP, we examine the time that CUP solvers spend on constructing as well as solving linear models for update scheduling. Recall that, CUP adopts the heuristic design of i) first deciding the updating order ii) then computing the optimal rate limits.
Accordingly, the total time that CUP needs for scheduling an update, is the sum of these took by Order Scheduler and Rate Manager. Fig. 14 illustrates the detailed results of the aforementioned updating cases of rebalancing traffic for inter-DC WAN topology (Fig. 12), in which each measured time is the mean value of 20 trials, carried out by Mosek on a PC running 64-bit Ubuntu 14.04 server with 8G RAM and a single Intel ES-1620 v2 CPU.

As Fig. 14a shows, given an update request, it would take non-trivial time, ranging from tens to hundreds of milliseconds, for CUP to build linear models for the corresponding scheduling problem. Basically, this is due to the fact that our initial CUP implementation acts as a first step for functional verification and it employs the Mosek Fusion API (ApS, 2016) for fast prototyping. With the high-level programming abstraction provided by Fusion, we could focus explicitly on modeling oriented aspects rather than reformulating problems into a single matrix and a few vectors, which is a time-consuming and error-prone process. Fusion APIs make the life much easier; but it simultaneously introduces computational overheads compared to using the low-level C APIs (ApS, 2016). Thus, for production code, building CUP solvers with these low-level APIs would be a better choice. Also, we observe that, the time for model building increases with Round, the maximum allowed round number. This is reasonable since more variables and constraints CUP solvers (i.e., these shown in Figs. 5, Figure 6, and Fig. 7) would get involved in, when a larger Round is set.

Besides model building, we further count the time that CUP takes on solving each of them. As Fig. 14b summarized, for each update request, the MIP-based Order Scheduler (Fig. 5) would need notable yet unpredictable solving time, up to tens of seconds, while its LP-based heuristic, LPHA, always yields a feasible order scheduling within several milliseconds. We also notice that, when a loose deadline is set (e.g., when Round ≥ 5), the MIP-based Rate Manager would be more efficient than its colleague that uses LPHA as the core. Mainly, this is because when Round ≥ 5, MIP-based Order Manager has found congestion-free schedulings in which no rate limit is needed thus it is easy for Rate Manager to find a feasible solution in that case. As Section 3.1 suggests, in practice, Order Scheduler could achieve both efficiency and effectiveness via a “dual-core” design.

6. Related work

Managing network updates in SDN is a hot topic. We mainly revisit the most related work here and refer the interested readers to (Foerster et al., 2018) for a comprehensive survey.

As in-flight packets might be handled by a mix of different versions of rules during the update, several approaches are proposed to provide strong consistent properties such that no packet or flow misuse rules (Reitblatt et al., 2012; Katta et al., 2013; Luo et al., 2015b), or weaker yet specific properties such as loop freedom (Mahajan and Wattenhofer, 2013; Zhou et al., 2015; Ludwig et al., 2015), and waypoint invariant (Ludwig et al., 2014; Zhou et al., 2015). While orthogonal to them, CUP focuses on another problem of managing transient congestion during the reconfiguration process, and directly employs generic two-phase approach (Luo et al., 2015b) to guarantee strong rule consistency for each step/round.

Schedulers like zUpdate (Liu et al., 2013), SWAN (Hong et al., 2013), and GI (Zheng et al., 2015) attempt to avoid transient congestion by introducing a sequence of intermediate traffic distributions (i.e., configurations), following which, the transition might be congestion-free. These introduced intermediate configurations greatly complicate the update procedure, and make the network error-prone (Miserez et al., 2015). Even worse, these intermediate configurations might hurt user’s QoS because of their paths might have unsatisfied delays and jitters. Moreover, for some updates, there does exist congestion-free transition plans. To avoid this, a portion (10%–50% (Hong et al., 2013)) has to be left vacant, which leads a great waste of link capacities (GI (Zheng et al., 2015) chooses to bear the transient congestion instead of reserving vacant bandwidth). Differently, Dionysus (Jin et al., 2014) and ATOMIP (Luo et al., 2015b) try to handle transient congestion by scheduling update operations according to a dynamic-determined or pre-designed order, which might avoid the problem of intermediate configurations. Even though, they only maintain a pre-defined specific objective by design—either towards fast speed, or congestion freedom. Accordingly, they cannot deal with various update scenario properly. By comparison, CUP formulates the update planning problem with generic models. Via binding models with user-specific constrains and objective functions, CUP adapts to a large fraction of scenarios easily.
Like CUP, many other researchers also realize the advantage of customizable network update planning and propose attractive proposals, recently. For example, Atoman enables operators to manually choose both the optimization objective and constraints for network update scheduling (Luo et al., 2019). However, Atoman would not limit flow rates; thus, it could not guarantee congestion freedom, in case the network is heavily loaded and there does not exist a congestion-free order scheduling. The work of Hermes does employ customizable rate limiting schemes to avoid transient congestion; but it focuses on the target of maximizing the sum of service utilities during the update, which is different from what CUP pursues (Zheng et al., 2018).

7. Conclusion

As transient congestions are prone to occur during SDN updates, controllers are in urgent need of a planner to handle the trouble. We argue that planning the reconfiguration process respecting to specified requirements is an import issue. In this paper, we have analyzed the desired properties of such planners and proposed a case design—CUP. CUP translates high-level user-specific requirements into linear constraints and formulates the planning problem as generic linear programs. By solving customized LPs, CUP is flexible to obtain “best” plans for a large fraction of updates.

Acknowledgements

We would like to thank the anonymous reviewers and editors for their useful feedback. This work is supported in part by the Fundamental Research Funds for the Central Universities (2682019CX61).

Appendix A. Why CUP adopts two-phase for rule consistency

As is known, when reconfiguring the network, in-flight packets might misuse different versions of rules (Reitblatt et al., 2012; Luo et al., 2015b; Mahajan and Wattenhofer, 2013; Ludwigel et al., 2014; Zhouel et al., 2015) and the solution is to either perform rule changes following a well-designed order (Mahajan and Wattenhofer, 2013; Ludwigel et al., 2014; Zhouel et al., 2015), or use version tags to avoid the mix use (Reitblatt et al., 2012; Luo et al., 2015b). Order arrangement does not require extra rules spaces, however, it has two fatal flaws. First, it is not universal and only works in specified cases (Mahajan and Wattenhofer, 2013; Ludwigel et al., 2014). Second and crucially, it is time-costly since it has to change rules one-by-one; this results in big update durations (Luo et al., 2015b). For example, provided the longest path in the dependency tree of rule changing order is L, and the average time for changing a rule from the controller is τ, it would take about L × τ to go through the entire process. In contrast, the version-based two-phase mechanism is generic and fast. If new rules already exists, the controller only needs to modify the ingress to switch a flow to the new path(s)—the time cost is τ; even though new rules are absent, the controller can let all new rules ready within another τ because these rule installations can be executed concurrently—the total time cost is 2τ, still greatly smaller than L × τ.

The possible price of version-based methods is rule-space overheads—switches have to hold two version of rules temporarily. For this problem, recent study has shown that, with the help of wildcard in match fields, switches only needs to store two versions for rules that are being modified (Luo et al., 2015b); this greatly reduce the overheads. Moreover, after an update procedure completes, all old rules can be removed immediately. Thus, we argue that rule-space overheads are not serious in many cases. Even in the case that the rule-space is the bottleneck, the controller can still pre-split updates to reduce the demands of extra rule-space (Katta et al., 2013), or employs rule aggregation techniques to get more useful rules space (Luo et al., 2014, 2015c).

References


Hongfang Yu is a Professor at University of Electronic Science and Technology of China. She received the BS degree in electrical engineering in 1996 from XiDian University, and the MS and PhD degrees in communication and in formation engineering in 1999 and 2006 from the University of Electronic Science and Technology of China, respectively. From 2009 to 2010, she was a visiting scholar at the Department of Computer Science and Engineering, University at Buffalo (SUNY). Her research interests include SDN/NFV, data center network, network for AI system and network security.

Long Luo is currently working toward the PhD degree from the University of Electronic Science and Technology of China (UESTC). She received the BS degree in communication engineering from Xian University of Technology in July 2012 and MS degree in communication engineering from the UESTC in July 2015. Her research interests include software-defined networks, traffic engineering and data-driven network.

Lemin Li received his B.S. degree in Electrical Engineering from Jiaotong University, Shanghai, in 1952. Then, he was with the Dept. of Electrical Communications at Jiaotong University until 1956. Since 1956 he has been with Chengdu Institute of Radio Engineering (now the UESTC). During Aug. 1980 to Aug. 1982, he was a Visiting Scholar in the Dept. of Electrical Engineering and Computer Science at the University of California at San Diego. Currently, his research interests are in the area of communication networks.

Shouxi Luo received his B.S. degree in Communication Engineering and Ph.D. degree in Communication and Information System from University of Electronic Science and Technology of China in 2011 and 2016, respectively. From Oct. 2015 to Sep. 2016, he was an Academic Guest at the Dept. of Information Technology and Electrical Engineering of ETH Zurich, Switzerland. His research interests include data center networks and software-defined networks.


Shouxi Luo

Hongfang Yu

Long Luo

Lemin Li

115