Unison: Enabling Content Provider/ISP Collaboration using a vSwitch Abstraction

Yimeng Zhao, Ahmed Saeed, Mostafa Ammar, Ellen Zegura
Georgia Institute of Technology
{yzhao389, amsmti3, ammar, ewz}@cc.gatech.edu

***Abstract*—BGP was initially created assuming by default that all ASes are equal. Its policies and protocols, namely BGP, evolved to accommodate a hierarchical Internet, allowing an autonomous system more control over outgoing traffic than incoming traffic. However, the modern Internet is flat, making BGP asymmetrical. In particular, routing decisions are mostly in the hands of traffic sources (i.e., content providers). This leads to suboptimal routing decisions as traffic sources can only estimate route capacity at the destination (i.e., ISP). In this paper, we present the design of Unison, a system that allows an ISP to jointly optimize its intra-domain routes and inter-domain routes, in collaboration with content providers. Unison provides the ISP operator and the neighbors of the ISP with an abstraction ISP network in the form of a virtual switch. This abstraction allows the content providers to program the virtual switch with their requirements. It also allows the ISP to use that information to optimize the overall performance of its network. We show through extensive simulations that Unison can improve ISP throughput by up to 30% through cooperation with content providers. We also show that cooperation of content providers only improves performance, even for non-cooperating content providers (e.g., a single cooperating neighbour can improve ISP throughput by up to 6%).**

I. Introduction

BGP policies have evolved to accommodate a hierarchy of Autonomous Systems (ASes) within the Internet. Policies, agreed on through pair-wise contracts, determine the role of an AS within the hierarchy. Mechanisms developed for BGP allow for fine-grain control over how traffic exists an AS. Each AS determines the next hop according to the policies agreed on with that next hop. This approach to inter-AS routing is feasible to manage under a hierarchical structure in which roles are clear, and in turn who pays whom for carrying the traffic is well established.

The modern Internet is flat; source ASes (i.e., content providers) are connected directly to destination ASes (i.e., ISPs) [1]. Moreover, the source can be connected to the destination through other transit ASes. Source ASes have the flexibility to choose how to reach their destination but it is not easy for destination ASes to control inbound traffic [2]. *This makes BGP highly asymmetrical.* The asymmetry is exacerbated by advancements in Software Defined Interconnects that make both decision making as well as decision making frequency asymmetrical. For example, systems like Egde Fabric [3] and Espresso [4], employed by Facebook and Google, respectively, improve reaction time of content providers to congestion as well as load balancing among different inter-domain links. On the other hand, ISPs still have to rely on typical, ineffective, standard BGP tools that take tens of minutes to converge.

978-1-7281-2700-2/19/$31.00 2019 © IEEE

The asymmetry of Internet routing, along with the current flat topology of the network, leave routing decisions largely in the hand of content providers. This is not ideal for two main reasons. First, content providers can only make decisions based on their view of the network which is typically based on estimates of capacity from the ISP entry point to the user (e.g., relying on CDNs or Points of Presence physically closest to the user). This is especially problematic in the presence of congestion when an alternative entry point has to be selected that does not have to be close geographically to the user [5]. The content provider can only estimate the characteristics of the path inside the ISP carrying traffic from the alternative entry point to the end user, which can be erroneous [6]. However, the ISP has the ground truth, making ISP selection of alternative entry points more reliable.

Second, dynamic path selection by a content provider, independent from the ISP, complicates fault attribution. In particular, a user facing poor quality of experience of an online service will typically blame the ISP, despite that the problem can be caused by the content provider selecting a longer or more congested path. This is a known source of dispute between content providers and ISPs [7], [8], [9], [10], [11]. This problem can be alleviated with better coordination between ISPs and content providers. There has been attempts to allow such exchange of information through brokers or at Internet Exchange Points (IXPs) [12], [13], [14]. Broker-based solutions are not scalable as they represent a centralized Internet. IXP-based solutions (e.g., SDX [15]) provide a good first step, however, their impact is limited to entry points connected to a single IXP and do not specify how to operate at the full scale of an ISP network.

In this paper, we present the design of Unison, a system that allows an ISP to jointly optimize its intra-domain routes and inter-domain routes, in collaboration with content providers (§III). Unison's design is based on the argument that deciding which entry point traffic should take to reach a user is a decision that should be performed jointly by both the ISP and the content provider. Our work is motivated by two observations: 1) measurements of interconnect congestion show that while some entry points between a content provider and an ISP can be congested several other entry points are uncongested across

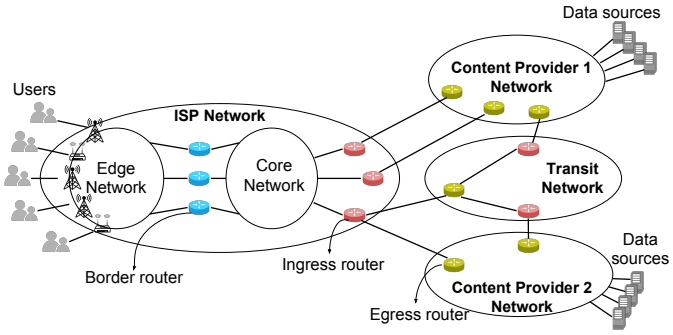


Fig. 1: Network context.

geographic regions (§II-C), and 2) the availability of software defined interconnect systems at content providers makes it feasible to coordinate between multiple networks and control inter-domain traffic.

The basic idea of Unison is to provide the ISP operator and the neighbors of the ISP with an abstraction of the ISP network in the form of a virtual switch. This abstraction allows the content providers to program the virtual switch with their requirements. It also allows the ISP to use that information to optimize the performance of its network. In addition, Unison allows the ISP to provide hints to its neighbors, suggesting alternative routes that can improve their performance. Unison leverages recent advancements in SDN. In particular, Unison makes use of SDN infrastructure at most modern ISPs [16], [17] as well as the programmable Interconnects at content providers [3], [4]. It also leverages SDX as a means to convert a vSwitch configuration into OpenFlow and BGP rules. This enables Unison to be a programmable platform that can be used for multiple Inter-domain routing applications (e.g. load balancing, or redirection through middleboxes).

We focus on the objective of maximizing throughput of content provider traffic going through the ISP. In particular, we are interested in the creation of a vSwitch abstraction from an ISP topology (§IV). Then, we investigate how this abstraction can be used to maximize the throughput of the ISP (§V). We formulate the problem as an integer program. Through that formulation, we investigate the the value of Unison in terms of improving ISP throughput in case of congestion. We also show the impact of non-cooperating content providers. Finally, we present a simple heuristic for selecting which content providers to approach for cooperation, if not all content providers can be approached. Our evaluation of Unison is conducted through simulations (§VI). We show that Unison can improve ISP throughput by up to 30% through cooperation with content providers. We also show that cooperation of content providers only improves performance, even for non-cooperating content providers (e.g., a single cooperating neighbour can improve ISP throughput by up to 6%).

II. Motivation and Background

A. Network Context

It has become increasingly important for content providers (CPs) to reach consumers with low latency. One way this has been achieved is through direct peering between CP and ISP networks. While this has helped, we believe it is necessary in today's demanding environment to also coordinate traffic routing across this peering connection. Recent work provides evidence that large CPs use peering links to carry majority of the traffic to access ISPs [3], making this coordination essential for the CP to achieve its reduced latency objective.

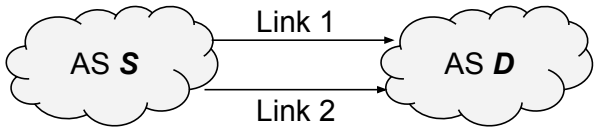


Fig. 2: Example of two ASes connected through two links (i.e., destination AS has two entry points).

We consider a network similar to the schematic in Figure 1. Multiple CPs are connected to an ISP, either directly at their points of presence, or through transit autonomous systems. We focus on the prevalent scenario where access ISPs connect directly with CPs. The ISP network is composed of *Ingress Routers* that receive traffic intended for users. Traffic is routed through the ISP's *Core Network* to *Edge Networks* that users connect to directly (e.g., cellular edge). The Core Network and Edge Network are connected through *Border Routers*. These Border Routers deliver traffic to a large number of users. In this paper, we are concerned with the problem of routing data from Ingress Routers to Border Routers, as multiple such routes can exist [4]. However, we assume that once traffic reaches a Border Router, its path to the user is deterministic.

We assume that some or all participating networks rely on programmable infrastructure to determine and configure routes. These assumptions are increasingly becoming the reality in the modern Internet as announced by ISPs [16], [17] and CPs [3], [4]. We note that CP networks without programmable infrastructure are able to handle routing suggestions from the ISP using existing APIs, motivated by the promise of higher throughput. Moreover, Unison does not require all CPs to cooperate with the ISP. Our results show that Unison can remain beneficial for the majority of CPs, even if only a subset of them cooperate.

We do not make any assumptions about data placement, as none is needed for our context. This is because our interest is in cases where network congestion, rather than physical distance, is the main bottleneck. Although ISP-CDN collaboration allows for strategic placement of data and routing optimization, which improves data delivery [18], [19], [20], [21], we are interested in reducing congestion where some links between the CP and the ISP are congested. Circumventing this congestion requires using a different entry point, that can be in a different physical location. This scenario is typical as we show later.

B. The need for Content Provider-ISP Cooperation

Internet Routing Asymmetry: Current Internet routing is asymmetric because it gives traffic sources much more control over route selection compared to traffic destinations. This

asymmetry is necessary to ensure traffic is routable in case of conflicting preferences. For instance, consider the case in Figure 2. Suppose the source AS prefers to send traffic over Link 1. An irresolvable conflict would arise if the destination AS prefers to receive the traffic over Link 2.

Current BGP mechanisms such as path prepending and selective announcement are very limited in terms of their expression of preference. In particular, an ISP can stop announcing certain prefixes through certain entry points, which is an extreme approach and typically not preferred for redundancy. The other available approach is path prepending which does not provide clear preference between paths and does not necessarily differentiate between CPs. Furthermore, these approaches rely on BGP convergence which is known to be slow, especially compared to Software Defined Interconnects. The asymmetry problem can be mitigated through the use of BGP communities that depend on cooperation between peering partners, but BGP comminities tend to leak critical information such as network topology hence is not an ideal solution [22]. We consider our solution as an argument against using BGP communities.

Determining Best Path to End Users: Typically, CPs try to route traffic to end users through the geographically closest point of presence (i.e., entry point to the ISP). However, if that entry point is congested, CPs can only guess which alternative entry point to use. CPs do not have visibility into the ISP's network. This means that by selecting another entry point, CPs cannot guarantee enough capacity from that entry point to the end user. Selecting the best entry point can only be achieved if the CPs cooperate with the ISP.

Attribution of Bad QoE: When end users face bad quality of experience (QoE), it is natural for users to blame the ISP [11]. Blaming the ISP implies that the ISP did not allocate enough capacity for traffic to reach the user. However, this does not necessarily have to be the case. It can be that the entry point used by the CP is congested due to large traffic volume from that CP. It can also be due to the CP choosing an entry point that does not have the proper capacity in its connection to the targeted users, while other entry points have that needed capacity. It can also be the case that the CP's network is congested. This attribution is very hard to achieve accurately and can be costly to the ISP if the CP unilaterally moves traffic between entry points. For example, this unilateral behavior can force the ISP to upgrade and increase the capacity of parts of its network while the same outcome could have been achieved by simply asking the CP to use a different entry point.

C. Interdomain Congestion across ISP Entry Points

Our main hypothesis in developing Unison is that when one point of entry to an ISP is congested, several other entry points are not congested. This hypothesis is critical as it implies the existence of the option to move traffic from the congested entry point to another. Unison allows this decision to be made by the ISP rather than the ISP's neighbouring AS because the ISP knows the best, or second best, entry point to reach its customers. We validate our hypothesis by examining interdomain congestion data between a single ISP and three CPs over a period of two years [23]. The data provides measurements of latency over multiple links connecting the ASes (i.e., egress-ingress router links in Figure 1). Links are grouped based on location where every location captures a single point of entry in our analysis. Each data point represent the congestion status inferred from the latency over a period of 24 hours. The congestion measurement method is based on an intuition that if the latency to the far end of the link is elevated but that to the near end is not, then it is highly likely that the interdomain link is congested [24].

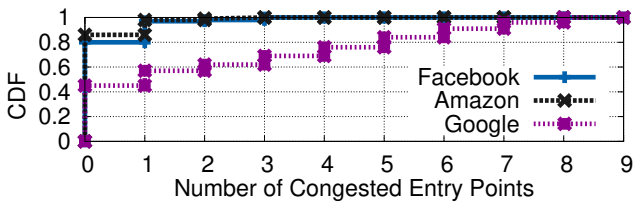


Fig. 3: CDF showing likelihood of congestion at one or more entry points.

Figure 3 shows the CDF of simultaneously congested entry points between Comcast and three CPs: Facebook, Google, and Amazon. Each content provider AS is connected to the ISP through at least twelve points of entry. The results validate our hypothesis. It is not uncommon to have multiple links congested at the same time while there are still links that have available capacity. In particular, we find that in worst cases of congestion there are 20.1%, 14.6%, 55.3% of the links are congested at a certain data point for Facebook, Amazon, and Google respectively. This means that there are at least seven alternative entry points available even in the worst cases of congestion. The objective of Unison is to allow the ISP to help the CPs choose between the entry points.

III. Unison Overview

Unison allows an ISP to expose a programmable interface to other autonomous systems connected to it. Unison limits the amount of information exchanged by only providing the vSwitch abstraction which does not reveal information about exact ISP topology, but provides some hints about capacity in exchange for improving performance. In particular, an AS can specify its preferred routing policies, which without Unison it would enforce regardless of the state of the ISP. Furthermore, the ISP can take into account the preferences of ASes connected to it, as well as its own capacity, to send hints back to neighbouring ASes suggesting better routes if any. It is left up to the neighbouring ASes to use these hints, thus preserving the distributed nature of the current Internet. Unison performs these functions by configuring ISP inter-domain and intra-domain routing simultaneously. Inter-domain routes are configured based on the state of the ISP as well as the routing preferences of its neighbouring ASes by providing neighbouring ASes with hints on entry points for aggregates of traffic that would optimize network performance. Intra-domain

routing is optimized by configuring capacity within the ISP to accommodate demand from peer ASes. We leave further anonymization of the hints to future research.

The insight we build on is that, given proper controller infrastructure, BGP routers can be programmed dynamically based on centrally made decisions to control inter-domain routing at scale. This was demonstrated by software defined Internet routing systems developed and deployed by CPs such as Espresso by Google [4] and Edge Fabric by Facebook [3]. Unison also builds on systems that allow the realization of a single policy from preferences set by multiple autonomous systems developed for Internet Exchange Points (e.g., SDX [15]). Unison is developed for an ISP setting which requires interaction with peer ASes, as well as consolidating ISP objectives and peer ASes objectives. The design of Unison has two major components:

1) Overlay software defined control over BGP infrastructure, à la Espresso, designed to control the Interconnect.
2) Cross-controller coordination and consolidation, à la SDX, designed to handle coordination between the ISP and ASes connected to it in order to reach a feasible resource allocation.

We find that despite progress made in such systems from the perspective of CPs and Internet exchange points, the ISP perspective poses a set of new challenges and constraints. For the rest of this section, we elaborate on these challenges as well as give an overview of Unison.

A. Unison Design Goals

An access ISP can be connected with multiple CPs at potentially multiple *ingress* points for each CP. Our goal is to provide a way for ISPs to control the network taking into account considerations from all CPs and users in addition to its own network and business considerations. Although Unison can be used to achieve a wide range of objectives, we focus on the simple and natural objective of maximizing ISP network throughput subject to weighted differential treatment of different CPs. Hence, all CPs observe a less congested ISP network, which is the main goal of CP-based solutions [4], [3]. Moreover, the ISP achieves higher utilization of its network in addition to achieving its business obligations by providing paying CPs more bandwidth. This approach is challenging and has to be handled under a very strict set of constraints:

- Benefits both CPs and ISPs. ISPs have to balance many CPs. Unison should improve ISP throughput while ensuring weighted differential treatment of CPs. Our system should also improve throughput of CPs within an ISP.
- Does not require CPs to cooperate. The benefits of Unison should be achieved even if only a subset of CPs connecting to an ISP agree to participate, without penalizing non-participating ones.
- Limits information disclosure: ISPs will not be willing to disclose detailed information such as network topology, traffic load, and customer information, often considered proprietary by ISPs, to third parties such as CPs.

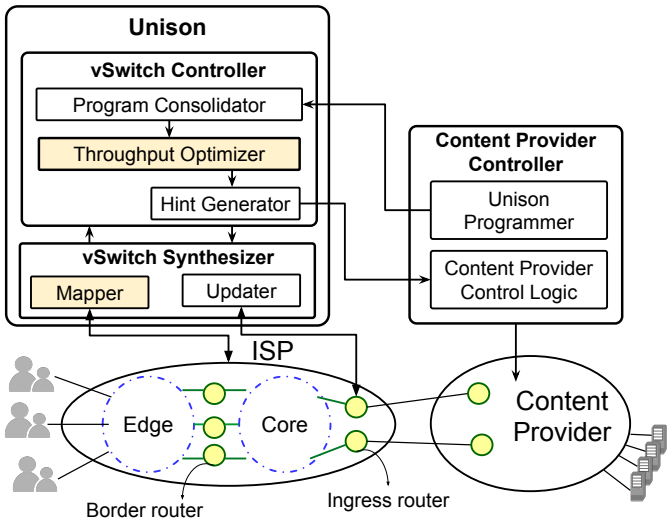


Fig. 4: Overview of Unison architecture.

B. Architecture Overview and Operation

Figure 4 shows the details of connectivity between an ISP and a Content Provider (CP). Note that we focus on the prevalent scenario where access ISPs connect directly with CPs. However, our technique will work as long as we can construct a traffic matrix that can be translated into a virtual switch. Our approach works on a (src ip prefix, dst ip prefix) granularity. Hence, it should tailor different negotiations to different clients of the transit network. Unison operates in the control plane of the ISP and provides an interface to CPs. Unison has two main components: (i) a vSwitch Synthesizer that creates and maintains the mapping between the vSwitch abstraction and actual network equipment at the ISP, and (ii) a vSwitch Controller that combines programs from CPs as well as the ISP to generate vSwitch configurations. Unison also requires minor changes in the controller of the CP network to specify its policies as well as receive and take into account hints from the ISP.

vSwitch Synthesizer: The main function of this component is to convert the complex topology of an ISP's network to a simple vSwitch with well-defined input and output ports. It also realizes high level programs of the vSwitch into actual route configurations in network elements. These two functions are the responsibility of the Mapper and Updater modules, respectively. The vSwitch Synthesizer is inspired by recent work in programmable inter-domain controllers introduced by content providers [4], [3]. In particular, these recent advancements show that a central controller can make routing decisions that reconfigure BGP routers either on per packet basis [4] or per point-of-presence basis [3].

vSwitch Controller: This component is responsible for programming the vSwitch as well as providing hints to neighbour ASes. A vSwitch program is created by combining programs from different neighbours of the ISP as well as the objective from ISP. Each program from each neighbour AS specifies its traffic demand as well as its routing preferences. Programs are combined in the *Program Consolidator*. The combined

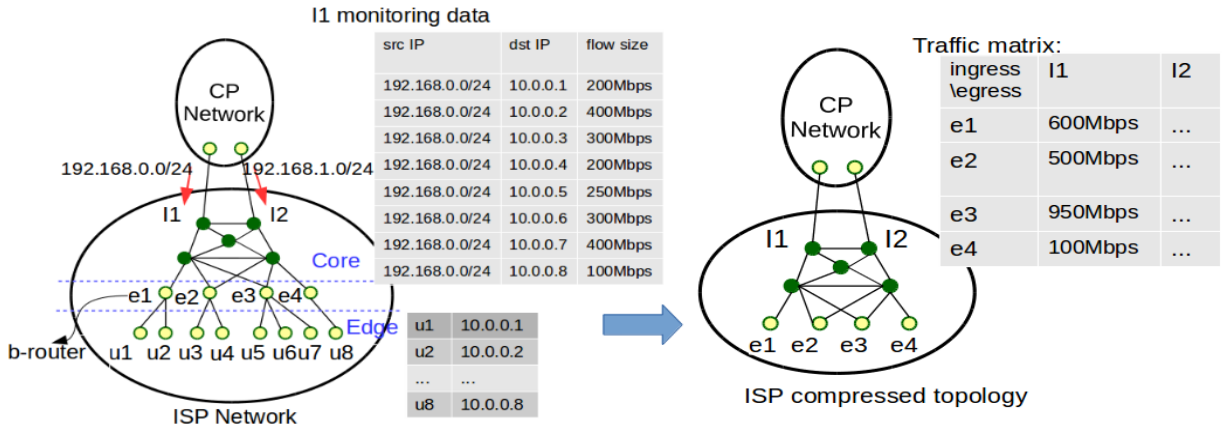
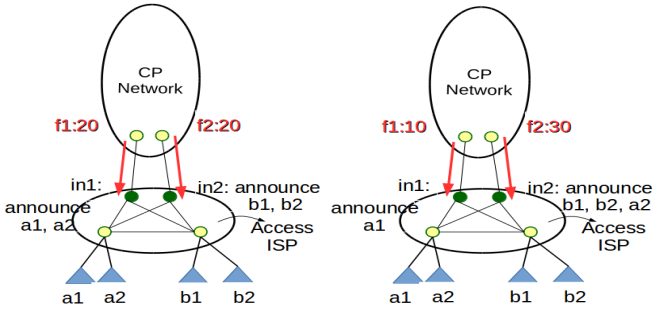


Fig. 5: Traffic matrix determination in the Mapping module



(a) before update (b) after update

Fig. 6: Influencing inbound traffic through hints.

program is fed to the *Throughput Optimizer* which generates network configuration as well as hints to the neighbours of the ISP. This component is inspired by recent advancements in interconnect abstractions (e.g., SDX [15]). We leverage these advancements to allow the neighbours of an ISP to indicate to the ISP how they prefer their work to be routed. Similar to SDX virtual switch abstraction, our proposed approach allows for dynamic allocation of IPs within the virtual switch, allowing for capturing of the complex dynamic topology.

This architecture captures a generic Unison that can be programmed to perform a wide variety of functions, depending on the programs provided by the CPs. However, in this paper we focus on the case of maximizing ISP throughput where CP programs only provide demand as a function of the input and output ports of the vSwitch. We note that challenges in building components such as the Program Consolidator and the Updater have been addressed in SDX. In particular, SDX combines and joins policies from multiple participating ASes to program a virtual switch abstraction of an IXP. Then, it converts such combined program into BGP and OpenFlow rules. In the paper, we focus on the two components highlighted in Figure 4: the Mapper and the Throughput Optimizer.

IV. Unison vSwitch Mapper

The function of the Mapper is to convert the complicated topology of the ISP into a vSwitch. In particular, the Mapper aims at identifying the input and output ports of the vSwitch. This is particularly challenging when taking scalability into account. In particular, a vSwitch defined by individual ports on individual routers in the ISP topology will lead to an intractably large vSwitch. To handle the TE problem that considers millions of egress flows destined to hundreds thousands of external IP prefixes, recent work [25] proposes a hierarchical framework for ISP network. The framework divides a global optimization problem into sub-problems, each of which is assigned to a child worker so the computation can be accelerated through parallelism. Our mapper, provides an alternative for ISPs who are not capable or not willing to build such a hierarchy framework. To insure scalability, the ISP simplifies its network by representing it as a traffic matrix where each element is an aggregate flow. An aggregate flow is defined by an entry point to the ISP from a specific CP to a group of users. The entry point for an aggregate flow pair is easy to define, and is fixed. The function of the Mapper is mostly concerned with grouping users which are typically represented by an IP-prefix. We take a greedy approach, starting to de-aggregate flows from the entry points with a predefined value for the maximum number of flows we can handle. We consider boundary routers the same as the entry points so we have one aggregate flow for each entry point. Then we look at the routers that are directly connected with the boundary routers and consider them as new boundary router. We keep doing this until the number of aggregated flows exceeds the threshold. To that end, we divide the ISP's network into "core" and an "edge" networks. Boundary Routers (*b-routers*) separate the ISP's core from the edge. Unison is concerned with routing CP data within the core network only, representing the core network by vSwitch. End users are aggregated such that data flow to and from the users is routed to a single b-router. The division of core from edge network is determined by the ISP. The closer the b-routers are to the users, the more effective Unison will be in controlling individual user performance but the large the scale of the problem Unison solves.

Realization of the aggregate flow can be achieved through existing tunneling techniques, such as MPLS, GRE, and VPNs,

or emerging SDN approaches based on flow space allocation [26]. An example of the mapping function is shown Figure 5. In the figure $I1$ and $I2$ are ISP ingress routers connected directly with the CP network. $e1$ to $e4$ are b-routers connecting the core and edge networks. Traffic demands from each ingress router to each user are shown in the left side of the figure. The aggregate flows are shown in one column of the traffic matrix resulting from the mapping process.

For Unison efficiency, the traffic matrix, which is the output of the mapper, has to be relatively fixed. This means that traffic from a specific CP to a group of users has to go through the same entry point. This is achieved by the Hint Generator which communicates to CPs to fix traffic going to a specific user IP prefix to a specific entry point. This feature is already supported by SDX for inbound traffic engineering. For ISPs that are not connected with SDX, we discuss other alternatives. Figure 6 demonstrates an examples of such process, which announces nonoverlaping prefixes to different interconnection links, to inform CPs of the suggested inter-domain traffic metrics. Some configuration is required between CP and ISP (e.g., disabling route damping). Figure 6a shows the situation before the update takes place. There are four flows, each with a size of 10 units, destined to $a1$, $a2$, $b1$, $b2$ respectively. Ingress router in_1 announces IP address of $a1$ and $a2$, and ingress router in_2 announces IP address of $b1$ and $b2$. To shift traffic from the left peering link to the right peering link, the ISP could announce IP address $a2$ at ingress router in_2 instead of at in_1. Although this approach is easy to deploy, it reduces the network resilience and may lead to routing table explosion. An alternative approach is to use AS path prepending or MEDs. For the example shown in Figure 6, the ISP could announce IP address of a2 at both ingress router in_1 and in_2 but with a shorter AS path or a smaller MEDs value in the announcement from in_2. To generate the router-level BGP configuration from high-level BGP policies, ISPs may use an BGP synthesizer [27], which takes as input the routing policies and generates Quagga router configurations.

One possible concern of dynamically changing BGP entries is that unstable routes may cause unexpected interactions among multiple nodes in a large network [28], [29]. One possible solution is to use Root Cause Notification (RCN), shown in recent work [29] to effectively eliminate false suppression and undesirable timer interactions. Further, although our design expects the ISP to trigger the monitoring and optimizing periodically (e.g., every few minutes), the ISPs are not obliged to change the routing every time when the Optimizer module generates a new inter-domain routing. The ISP may change the inter-domain routing only when the new routing can significantly improve the throughput.

V. Unison Throughput Optimizer

Unison provides the neighbor ASes of an ISP with a virtual switch abstraction connecting the neighbour AS to customers of the ISP. This abstraction allows the ISP as well as its neighbours to program the virtual switch to implement different inter-domain applications. We focus on the application of maximizing the total throughput of the ISP (i.e., the number of bits per second delivered from the ISP entry points to the ISP customers). With arguments still raging aroung Net Neutrality in the US [30], we propose a framework that can be tuned to provide a neutral or biased ISP. In particular, we look at throughput optimization with weighted fairness constraints, assigning different weights in the lack of net neutrality regulations and equal weights otherwise.

The optimization problem that runs at the Optimizer is critical to the performance of the system. The Optimizer module takes as input elements a network topology (ingress routers, core network and b-routers as shown in the previous section), a traffic matrix that represents the demand for each CP from each ingress location to/from each of the b-routers, the pre-defined CPs' behavior (i.e., participating or non-participating) stated in the agreement between CPs and ISPs, and the link capacity constraints. These constraints are translated into decision variables for an optimization solver [31], [32] . Table I summarizes our notation.

A. Traffic Matrix

For the *intra-domain traffic matrix*, we define $intraTM_{i,i'}$ as the volume of traffic that enters the ISP network at ingress point i and exits at b-router point i'. We use this intra-domain traffic matrix for traffic from non-participating CPs. For participating CPs, an *inter-domain traffic matrix* is constructed by summing the intra-domain traffic matrices for each b-router point. For a network with two ingress points i and j, and two b-router points i' and j', given the intra-domain matrix $intraTM_{i,i'}$, $intraTM_{i,j'}$, $intraTM_{j,i'}$, $intraTM_{j,j'}$, the elements of the inter-domain matrices are constructed as $interTM_{k,i'} = intraTM_{i,i'} + intraTM_{j,i'}$ and $interTM_{k,j'} = intraTM_{i,j'} + intraTM_{j,j'}$.

Non-cooperating Neighbours: Unison can also handle cases when neighbouring ASes do not provide their traffic demand or accept the hints provided by the Hint Generator. In particular, the traffic matrix can be inferred through monitoring. Recent work [33] shows that it is possible to monitor network traffic for any prefix within an ISP network within milliseconds. The proposed approach does not require modification on current vendor hardware and is easy to deploy. Furthermore, to fix the aggregate flow pairs, the ISP can leveraging the existing BGP mechanisms like selective announcements or prepending. We show the impact of non-cooperating neighbours on performance in Section VI.

B. Feasibility and Weighted Fairness

We model the ISP network (ingress routers, core network and b-routers) as a directed graph $G = (V, E)$, where V is the set of routers and E is the set of links that connect the routers. Assume there are CPs competing for resources, each CP requesting resources for n_i aggregate flows (e.g., in Figure 5 the CP is requesting resources for 8 aggregate flows). We define I as the set of CPs and M as the set of aggregate flows among all CPs. We use $d_{i,m}$ to represent the bandwidth demand from the i^{th} CP for aggregate flow m and use $r_{i,\,m}$ to express the rate allocated to aggregate flow m for the i^{th} CP.

Variable	Description
$G(V, E)$	network with V routers and E links
c_e	capacity of edge e in E
I	a set of CPs
M	a set of aggregate flows
J	a set of paths
T	a set of aggregate flows that cannot receive a higher allocated bandwidth
$t_{i,m}$	allocated bandwidth to aggregate flow m for CP i in T
$b_{j,e}$	is edge e contained in path j; binary
$d_{i,m}$	bandwidth demand from CP i on aggregate flow m
$r_{i,m,j}$	bandwidth allocated to flows from CP i on aggregate flow m over path j
$r_{i,m}$	bandwidth allocated to flows from CP i on aggregate flow m
n_i	number of aggregate flows for CP i
$w_{i,m}$	weight of aggregate flow m for CP i
w_i	weight of CP i
$CPSat_i$	satisfaction of CP i
b_l	lower bound of allocated bandwidth
b_h	upper bound of allocated bandwidth

TABLE I: List of Notation

We use $r_{i,m,\,j}$ to express the rate allocated to flows from i^{th} CP on aggregate flow m over path j. The ISP takes as input the CP's bandwidth requests for aggregate flows (i.e., $d_{i,m}$) as well as the topology capacity, and generates allocations for each aggregate flow (i.e., $r_{i,m}$).

Feasibility: An allocation policy is feasible if no link capacity is exceeded. The upper limit of a feasible solution can be found by solving the following optimization:

$$\begin{aligned} \text{maximize} \quad & \sum_i \sum_m \sum_j r_{i,m,j} \\ \text{subject to} \quad & \sum_j r_{i,m,j} \le d_{i,m}, \forall m \in M, \forall i \in I \\ & \sum_i \sum_m \sum_j r_{i,m,j} \times b_{j,e} \le c_e, \forall e \in E \\ & b_{j,e} \in \{0,1\}, r_{i,m,j} \succeq 0, \forall j \in J \end{aligned} \tag{1}$$

$b_{j,e}$ is the binary variable on whether path j contains edge e and c_e is capacity of edge e. There is no fairness constraint on this optimization, so the result may assign high bandwidth to aggregate flows from to a few CPs and completely starve the others in an effort to maximize the total throughput.

Weighted Fairness In addition to being feasible, a bandwidth allocation policy should also be fair. Fair bandwidth allocation to flows has been extensively studied in the past [34]. Demirci et al. [35] studied how to extend these fairness definitions to multiple overlay networks instantiated on one substrate. Kleinberg et al. [36] take routing into consideration and prove the problem is NP-hard. In this section, we present a definition for fair allocation among multiple CPs.

We define a weighted fairness index (WFI) to evaluate the fairness of a bandwidth allocation policy in a multi-CP-setting. We define normalized weight of aggregate flows as follows:

$$w_{i,m} = \frac{d_{i,m}}{\frac{\sum_i \sum_m d_{i,m}}{\sum_i n_i}} \tag{2}$$

The weight of a aggregate flow is proportional to its demand and is normalized by the average aggregate flow demand for all CPs. This insures that bandwidth allocations are positively correlated with aggregate flow demands. As will be described in the next section, we use these weights to insure the weighted fairness of the allocation algorithm. The weight of a CP is defined as the sum of the CP aggregate flow weights: $w_i = \sum_m w_{i,m}$. We define CP satisfaction (*CPSat*) in the same way as the network satisfaction metric (*NetSat*) in [35] with $CPSat_i$ denoting the satisfaction of CP i. The *CPSat* describes how close the CP aggregate flow bandwidth allocation in the presence of other CPs is to the allocation it would receive had it been without competition. WFI is defined as the weighted standard deviation of the CP satisfaction metrics across all CPs sharing the resources of the ISP.

Algorithm 1 WBA: Weighted Bandwidth Allocation

Input: Traffic metrics $d_{i,m}$**, a set of paths** $b_{j,e}$ in
Output: Allocated rate $t_{i,m}$ out

1: $w_{i,m} \leftarrow \frac{\sum_i n_i \times d_{i,m}}{\sum_i \sum_m d_{i,m}}$, $k \leftarrow \lceil \log_a [\frac{\max_{d_{i,m} \times w_{i,m}}}{u}] \rceil$
2: $T \leftarrow \emptyset$
3: **for** $n = 1...k$ **do**
4: **for** $r_{i,m} \in BMCF(a^{n-1}u, a^n u)$ **do**
5: **if** $(i,m) \notin T$ and $r_{i,m} \le \min(d_{i,m}, a^n u \times w_{i,m})$ **then**
6: $T \leftarrow T + (i,m)$, $t_{i,m} \leftarrow r_{i,m}$
7: **end if**
8: **end for**
9: **end for**
10: **return** $t_{i,m} : (i,m) \in T$

C. Bandwidth Allocation Algorithm

The brute force method to find the optimal routing is to (i) enumerate all paths between every ingress node and b-router node pair, and then (ii) apply max-min fair bandwidth allocation algorithm to all possible path selections to find the optimal selection that achieves the highest total rate. To make the computation faster, we limit the possible paths to k shortest paths instead of enumerating all paths between ingress and b-router node pair. To further reduce the computation time, the path generation process is performed offline. We expect valid paths to change infrequently.

The max-min fairness bandwidth allocation algorithm computes the allocation for each flow iteratively: maximizing the minimal flow rate, freezing the minimal flows and then repeating the steps for the second minimal flow. The computation quickly becomes infeasible as the number and size of a network grows. Inspired by SWAN approximate max-min fairness heuristic [37], we use the Weighted Bandwidth Allocation (WBA) algorithm shown in Algorithm 1. The algorithm achieves weighted fairness between CPs by solving an optimization problem which we call Bounded MCF (BMCF) in k steps. In every iteration, BMCF solves a multi-commodity problem (MCF) problem that aims at maximizing $\sum_i \sum_m \sum_j r_{i,m,j}$, which is similar with optimization problem (1). The difference is that BMCF tries to achieve weighted

fairness among CPs, so in each iteration it puts a lower bound and upper bound on rate allocated to each aggregate flow:

$$b_l w_{i,m} \leq \sum_j r_{i,m,j} \leq \min(d_{i,m}, b_h w_{i,m}), \forall (i,m) \notin T \quad (3)$$

$b_l = a^{n-1}u$ and $b_h = a^n u$, which is passed by WBA in step n (line 4). Aggregate flows with lower demands have smaller weights, ending with fewer allocated rates. If an aggregate flow is allocated with its full demand or it cannot receive a higher allocation because of the link capacity constraints, the aggregate flow is frozen and is removed from the next round of computation. If every aggregate flow has the same demand, this allocation is identical to max-min fair allocation. Note that any changes in the traffic matrix require rerunning the optimization problems. This overhead can be mitigated by only recalculating routes for the affected parts of the network. We leave such enhancements for future work.

VI. Evaluation

In this section, we focus our evaluation efforts on exploring how useful Unison can be under the limitations discussed in §III-A. We show that Unison can provide improved throughput and differential treatment between CPs while not harming the performance of any CPs, even with limited number of cooperating CPs. We also evaluate the impact of various parameters and settings on the system's performance. To evaluate the performance of our design, we implement a proof-of-concept Optimizer that calls the CPLEX solver through its python API to solve the optimization problem described earlier. We conduct simulations to study the performance of our algorithm in realistic settings.

A. Experimental Setup

Topologies: We conduct experiments with a setting of one ISP and twenty CPs. Each CP is connected with the ISP in multiple interconnection nodes (i.e, ingress nodes) and the number of inter-domain links ranges from 1 to 5. Our simulation uses a variety of topologies from topology zoo [38] for the ISP and CP network.

Traffic demand: We assume that there is one flow from each CP source node to each ISP egress node. We consider all nodes excluding egress points in CP topology as source nodes and all nodes excluding ingress points in ISP topology as egress nodes. We simulate the traffic demand using a gravity model [39], which predicts that the traffic demand of a CP is proportional to the corresponding node population.

Link capacity: In our simulation, inter-domain link capacities are drawn from distribution of congested interconnections in recent work [3]. We generate the inter-domain link capacity by multiplying the traffic demand with the fraction of congestion shown in [3]. For the intra-domain link capacities, we assume that all links in the ISP have the same capacity and the link weights are assumed to be one. The value of this capacity is calculated through the following steps. First, we compute the routing with the default routing (i.e, OSPF) for the ISP network and identify the link with the most

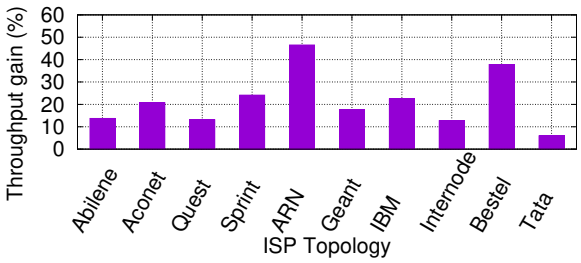


Fig. 7: Throughput gain created by Unison compared to the baseline over different ISP topologies.

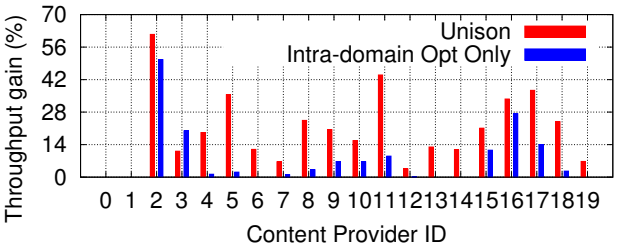


Fig. 8: Throughput gains comparing optimizing intra-domain only and Unison.

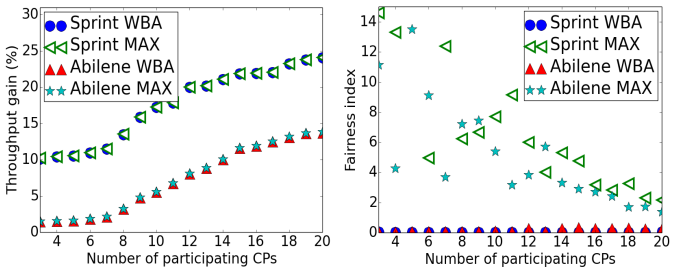
traffic demand. Then we compute a capacity by multiplying a congestion parameter with the demand carried in the most heavily loaded link. The goal of this approach is guarantee that a few links are congested. We also experimented with other link capacity distributions (i.e., uniform random distribution) and we observe that the results remain qualitatively similar.

Baseline: We use early-exit policy for the default routing. The chosen interconnection is the one that is the closest to the source. We assume that flows belong to the same aggregate flow will be routed in the same way and will not be split between multiple paths.

B. The Value of Unison

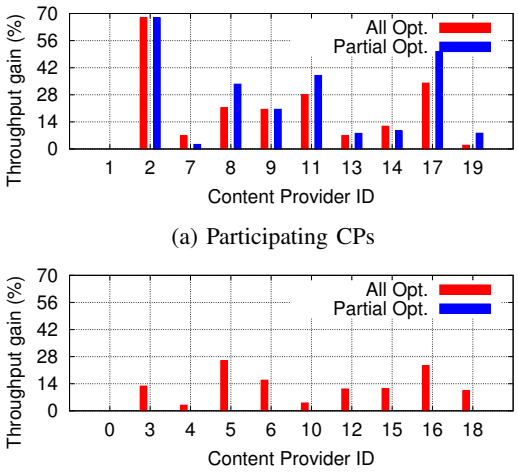
Value to ISP: We compare Unison to the baseline in terms of the amount of traffic they can deliver from CPs to end users. Figure 7 shows the throughput gain of Unison. We assume all CPs are participating, i.e., agree to use the new inter-domain routing as suggested by the ISP. It is clear that Unison's approach to jointly optimize inter-intra-domain routing improves ISP throughput. We also note that topologies with smaller average node degree improve more with Unison. Compared to complex topologies, simple topologies have less candidates paths between each ingress and egress node pair and it is more likely that a few links are heavily used by a large portion of paths. Therefore, changing the inter-domain routing is effective at diminishing unbalanced link usage.

Value to CPs: To better understand the value of Unison, we look at how increase in ISP throughput is viewed from the CP. Figure 8 shows the performance gain in percentage. We compare Unison to a baseline that attempts to optimize network utilization through optimizing OSPF parameters only (i.e., Optimizing intra-domain only). We observe that most CPs achieve a much higher throughput gain when the ISP relies on Unison compared to only optimizing intra-domain routing.



(a) Throughput gain (b) Weighted Fairness Index

Fig. 9: Comparison between WBA and MFC algorithm showing minor throughput impact with weighted fairness.



(a) Participating CPs

(b) Non-participating CPs

Fig. 10: OptAll v.s. only for OptPartial showing that non-participating CPs enjoy a free ride of increased throughput in OptAll as participating CPs while only participating CPs achieve increased throughput in OptPartial.

This shows the value in the joint optimization of inter- and intra-domain routing even from the perspective of CPs.

Value of jointly optimizing for weighted fairness and throughput: To show the value of our proposed algorithm WBA, we compare it to the multi-commodity flow (MCF) algorithm. We compare the two algorithms on Sprint and Abilene topology. Each ISP is connected to 20 CPs and we change the number of participating CPs from 3 to 20. Figures 9a and 9b show the total allocated bandwidth and the weighted fairness index (WFI) respectively. For both of the topologies, the WFI for the bandwidth allocation generated by the WBA algorithm is lower (better) than that of the MCF algorithm. The throughput gain for both algorithms almost match with MCF performing negligibly better in some cases. We also observe that the WFI achieved by the MCF algorithm shows a decreasing trend as the number of participating CPs increases. The main reason is that the MCF algorithm does not enforce any constraints on the bandwidth assigned to each aggregate flow and participating CPs have advantages over non-participating CPs by adjusting the inter-domain routing. As the number of participating CPs increases, the effect of favorable treatment on a few CPs starts to diminish.

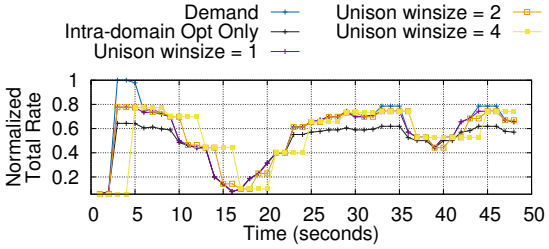


Fig. 11: Impact of window size using in Unison on total ISP throughput under dynamic traffic demand.

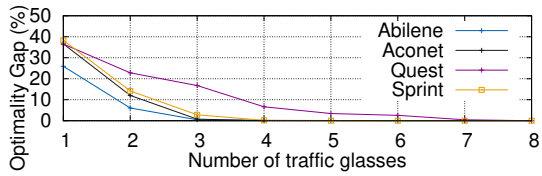


Fig. 12: Optimality gap as function of aggregate flow.

C. Impact of CP participation

In the previous experiments, our algorithm optimizes the inter-domain routing only for participating CPs and optimizes the intra-domain traffic for both participating and non-participating CPs. We call this approach Optimizing for All (OptAll). An alternative approach is to not change any routing, including intra-domain routing and the allocated rate, for non-participating CPs while optimizing both intra- and inter-domain routing for participating CPs. We now compare the performance of optimizing for all (OptAll) against optimizing only for participating CPs (OptPartial). We conduct four sets of experiments with different participating CPs and non-participating CPs selection, but due to space limits we only show results (Figure 10) for the experiment with 10 participating CPs and 10 non-participating CPs. Our result shows that OptAll achieves a slightly higher ISP throughput gain than OptPartial does. Compared with OptAll, OptPartial achieves similar or higher throughput gain for participating CPs. Non-participating CPs enjoy a free ride of increased throughput in OptAll as participating CPs. This effect may decrease CP's motivation to participate if there are changes to have throughput gain without participating.

D. Impact of environmental parameters

Impact of inaccurate prediction of dynamic traffic: In our design, the Estimation module collects the current traffic demand and uses it to estimate demand for a future window. This estimation is the main source of error. This error in demand prediction can cause under- or over-provisioning of bandwidth to some aggregate flows. The value of the error is a function of the estimation window size. To measure the impact of errors in estimating the traffic demand, we conduct an experiment with dynamic traffic. The traffic data is drawn from recent measurement study on YouTube network traffic at a campus network [40]. Figure 11 compares Unison with different window sizes to the baseline routing scheme. Unison adapts to changes when a small window size is used leading to better throughput than the baseline. When a large window size

is used, Unison causes under-and over-provisioning frequently, which makes the default routing more preferable. Note that the window size depends on the frequency monitoring, mapping, optimization, and update can be run.

Impact of the traffic granularity: Traffic granularity refers to the level of traffic aggregation. In our baseline experiment, we consider aggregating all traffic entering at an ISP ingress router and routed to a b- router as an aggregate flow, and flows in the same aggregate flow are not splittable. We expect increasing the number aggregate flows per each ingress and b-router pair should increase the total rate until the highest rate has been achieved. Fig 12 shows the gap between the achieved total rate and the optimal allocation. Unsurprisingly, the computation time increases linearly as the number of aggregate flows increases. However, this is not a big concern. In most cases, we find that the throughput gain reaches its largest value with aggregate flows numbers as low as 5.

VII. Related Work

Centralized approach There are an increasing number of proposals that suggest an inter-domain routing broker to provide end-to-end guaranteed paths [12], [13], [14]. In some of these proposals, ISPs provide QoS-enabled pathlets [12], which are stitched together by a centralized mediator called a service broker. The design requires that users submit their requirements and service providers submit their topology information to a service broker, who chooses the proper path in each domain and stitches the paths together to form an end-to-end path based on a global view of all participating networks. While a centralized network provisioning approach may optimize the inter-domain routing in an efficient way, the system is difficult to scale. Other proposals show that Internet exchange points (IXPs), the physical locations where multiple networks connect to exchange traffic, provide an ideal location to improve the existing routing system [15], [41]. Those approaches build on recent technology trends of Software Defined Networking (SDN) to utilize traffic-management capabilities and explore various use cases ranging from inbound route selection to application-specific peering. In the SDX approach [15], participants exchange BGP update messages with the IXP route server, and the SDN controller combines the SDN policy with the BGP routing information to compute forwarding table entries in the IXP fabric. However, such proposals do not provide control over all possible ingress paths to an ISP as IXPs do not represent all possible connection points between between ISPs and CPs [3].

Negotiation-based approach A few research studies have explored the benefits of allowing neighboring domains to collaboratively manage traffic [42], [43]. In these inter-domain architectures, neighboring ISPs exchange information about their traffic volume and preferred routes, and participate in negotiations until they reach mutually acceptable routes. Mahajan et al. [42] propose a negotiation-based routing framework where neighboring ISPs exchange their preference for inter-domain paths. Shrimali et al. [43] use the idea of multi-criteria optimization and Nash bargaining to approach the inter-domain routing problem. However, realizing ISP collaboration in practice is not straight-forward. The negotiation-based approach requires clean-slate architectures and protocols, which suffer from deployment challenges. Similar to the centralized approach, having to disclose sensitive information such as network structures and link capacities may prevent ISPs from participating.

Distributed Routing Systems: Large CPs have already taken initiatives to improve inter-domain routing aimed at delivering high-volume traffic while improving user-perceived performance [3], [4]. To tackle the limitations of BGP, Facebook designed Edge Fabric, a system for optimizing routing at the edge [4]. Edge Fabric monitors capacities and demand for outgoing traffic, and enforces better route selection by overriding the router's normal BGP selection for outbound traffic in Points of Presence (PoPs). Google takes a similar approach, designing an edge architecture that delivers high-demand traffic with low latency [3]. While Facebook only optimizes routing in PoPs, Google's architecture has a global traffic engineering system that enables application-aware routing at Internet scale. Both systems use their already deployed SDN infrastructure to dynamically change BGP entries.

ISP-CDN Collaboration: In order to improve the content delivery efficiency, the collaboration of ISP and CDN has been proposed. Jiang et al. [18] focus on the joint optimization of TE in ISP and server selection in CDN. Poese et al. [19] shows that server selection alone, without TE in ISP, is sufficient enough to improve the content delivery. Another line of work focus on mapping clients' request to the closest CDN clusters using DNS-based approach or SDN-based approach [20], [21]. Our work focuses on reacting to variability in network condition, which is different from previous ISP-CDN collaboration that focus on content placement and routing.

VIII. Conclusion

In this paper we propose a framework to be deployed in an access ISP network for joint inter-intra-domain routing. We consider practical deployment issues and evaluate different design choices. We develop a resource allocation strategy that can be deployed by ISPs that maximizes the allocation to the CPs within the ISP capacity constraints while insuring fairness among CP allocations. Our evaluation shows that such framework is beneficial to both CPs and ISPs, improving total throughput of CPs within an ISP and improving ISP throughput. We also show that the benefits of Unison can be achieved even if only a subset of CPs connecting to an ISP agree to participate. Future research is needed to understand the economic model to establish such collaborative relationship between content providers and ISPs. Recent work [44], [45] shows that SDN has the potential of reducing operational expense (OPEX) and capital expense (CAPEX), especially the network operation cost, but the analysis should be extended to consider the cost and benefit of collaboration.

IX. Acknowledgement

This work was funded in part by the National Science Foundation grant NETS 1816331.

REFERENCES

[1] A. Dhamdhere and C. Dovrolis, "The Internet is Flat: Modeling the Transition from a Transit Hierarchy to a Peering Mesh," in *Proc. ACM CoNEXT*, 2010.

[2] B. Quoitin, C. Pelsser, L. Swinnen, O. Bonaventure, and S. Uhlig, "Interdomain Traffic Engineering with BGP," *IEEE Communications magazine*, 2003.

[3] B. Schlinker, H. Kim, T. Cui, E. Katz-Bassett, H. V. Madhyastha, I. Cunha, J. Quinn, S. Hasan, P. Lapukhov, and H. Zeng, "Engineering egress with edge fabric: Steering oceans of content to the world," in *Proc. ACM SIGCOMM*, 2017.

[4] K.-K. Yap, M. Motiwala, J. Rahe, S. Padgett, M. Holliman, G. Baldus, M. Hines, T. Kim, A. Narayanan, A. Jain, V. Lin, C. Rice, B. Rogan, A. Singh, B. Tanaka, M. Verma, P. Sood, M. Tariq, M. Tierney, D. Trumic, V. Valancius, C. Ying, M. Kallahalla, B. Koley, and A. Vahdat, "Taking the edge off with espresso: Scale, reliability and programmability for global internet peering," in *Proc. ACM SIGCOMM*, 2017.

[5] R. Torres, A. Finamore, J. R. Kim, M. Mellia, M. M. Munafo, and S. Rao, "Dissecting video server selection strategies in the Youtube CDN," in *Proc. of IEEE ICDCS*, 2011.

[6] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang, "Unreeling Netflix: Understanding and improving multi-CDN movie delivery," in *Proc. of IEEE INFOCOM*, 2012.

[7] R. Andrews and S. Higginbotham, "YouTube sucks on French ISP Free, and French regulators want to know why," 2013, (Date last accessed 1-April-2019). [Online]. Available: https://gigaom.com/2013/01/02/youtube-sucks-on-french-isp-free-french-regulators-want-to-know-why/

[8] J. Brodkin, "Time Warner, net neutrality foes cry foul over Netflix Super HD demands," 2013, (Date last accessed 1-April-2019). [Online]. Available: https://arstechnica.com/information-technology/2013/01/timewarner-net-neutrality-foes-cry-foul-netflix-requirements-for-super-hd/

[9] ——, "Why YouTube buffers: The secret deals that makeand breakonline video," 2013, (Date last accessed 1-April-2019). [Online]. Available: https://arstechnica.com/information-technology/2013/07/why-youtube-buffers-the-secret-deals-that-make-and-break-online-video/

[10] S. Buckley, "France Telecom and Google entangled in peering fight," 2013, (Date last accessed 1-April-2019). [Online]. Available: https://www.fiercetelecom.com/telecom/france-telecom-and-google-entangled-peering-fight

[11] C. Dovrolis, "The evolution and economics of internet interconnections," *Submitted to Federal Communications Commission*, 2015.

[12] P. Godfrey, I. Ganichev, S. Shenker, and I. Stoica, "Pathlet Routing," *Proc. of ACM SIGCOMM*, 2009.

[13] H. Esquivel, C. Muthukrishnan, F. Niu, S. Chawla, and A. Akella, "RouteBazaar: An Economic Framework for Flexible Routing," *Technical Report TR1654, Department of Computer Sciences*, 2009.

[14] V. Kotronis, R. Klöti, M. Rost, P. Georgopoulos, B. Ager, S. Schmid, and X. Dimitropoulos, "Stitching Inter-domain Paths over IXPs," in *Proc. ACM SOSR*, 2016.

[15] A. Gupta, L. Vanbever, M. Shahbaz, S. P. Donovan, B. Schlinker, N. Feamster, J. Rexford, S. Shenker, R. Clark, and E. Katz-Bassett, "SDX: A Software Defined Internet Exchange," in *Proc. ACM SIGCOMM*, 2014.

[16] S. Tse and G. Choudhury, "Real-time traffic management in at&t's sdn-enabled core ip/optical network," *Optical Fiber Communication Conference*, 2018.

[17] M. Birk, G. Choudhury, B. Cortez, A. Goddard, N. Padi, A. Raghuram, K. Tse, S. Tse, A. Wallace, and K. Xi, "Evolving to an sdn-enabled isp backbone: key technologies and applications," *IEEE Communications Magazine*, vol. 54, no. 10, pp. 129–135, 2016.

[18] W. Jiang, R. Zhang-Shen, J. Rexford, and M. Chiang, "Cooperative content distribution and traffic engineering in an isp network," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 1. ACM, 2009, pp. 239–250.

[19] I. Poese, B. Frank, G. Smaragdakis, S. Uhlig, A. Feldmann, and B. Maggs, "Enabling content-aware traffic engineering," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 5, pp. 21–28, 2012.

[20] I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann, "Improving content delivery using provider-aided distance information," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 22–34.

[21] M. Wichtlhuber, J. Kessler, S. Bücker, I. Poese, J. Blendin, C. Koch, and D. Hausheer, "Soda: Enabling cdn-isp collaboration with software defined anycast," in *2017 IFIP Networking Conference (IFIP Networking) and Workshops*. IEEE, 2017, pp. 1–9.

[22] B. Donnet and O. Bonaventure, "On bgp communities," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 55–59, 2008.

[23] "MANIC: Measurement and ANalysis of Internet Congestion," (Date last accessed 18-August-2019). [Online]. Available: https://manic.caida.org

[24] A. Dhamdhere, D. D. Clark, A. Gamero-Garrido, M. Luckie, R. K. P. Mok, G. Akiwate, K. Gogia, V. Bajpai, A. C. Snoeren, and K. Claffy, "Inferring persistent interdomain congestion," in *Proc. ACM SIGCOMM*, 2018.

[25] M. Moradi, Y. Zhang, Z. M. Mao, and R. Manghirmalani, "Dragon: Scalable, flexible, and efficient traffic engineering in software defined isp networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 12, pp. 2744–2756, 2018.

[26] A. Al-Shabibi, M. De Leenheer, M. Gerola, A. Koshibe, G. Parulkar, E. Salvadori, and B. Snow, "OpenVirteX: Make your Virtual SDNs Programmable," in *Proc. ACM SIGCOMM HotSDN Workshop*, 2014.

[27] R. Beckett, R. Mahajan, T. Millstein, J. Padhye, and D. Walker, "Don't Mind the Gap: Bridging Network-wide Objectives and Device-level Configurations," in *Proc. ACM SIGCOMM*, 2016.

[28] Z. M. Mao, R. Govindan, G. Varghese, and R. H. Katz, "Route Flap Damping Exacerbates Internet Routing Convergence," in *ACM SIGCOMM Computer Communication Review*, 2002.

[29] B. Zhang, D. Pei, D. Massey, and L. Zhang, "Timer Interaction in Route Flap Damping," in *Proc. IEEE ICDCS*, 2005.

[30] A. A. Gilroy, "The net neutrality debate: Access to broadband networks," *Congressional Research Service, Washington, DC*, 2019.

[31] "Using the CPLEX Callable Library and CPLEX Mixed Integer Library," *CPLEX Optimization, Incline Village*, 1993.

[32] "The gurobi optimizer," (Date last accessed 15-May-2018). [Online]. Available: http://www.gurobi.com/

[33] O. Tilmans, T. Bühler, S. Vissicchio, and L. Vanbever, "Mille-Feuille: Putting ISP traffic under the Scalpel," in *Proc. ACM HotNets*, 2016.

[34] J. Jaffe, "Bottleneck Flow Control," *IEEE Transactions on Communications*, 1981.

[35] M. Demirci and M. Ammar, "Fair Allocation of Substrate Resources among Multiple Overlay Networks," in *Proc. IEEE MASCOTS*, 2010.

[36] J. Kleinberg, Y. Rabani, and É. Tardos, "Fairness in routing and load balancing," in *Proc. of IEEE FOCS*, 1999.

[37] C.-Y. Hong, S. Kandula, R. Mahajan, M. Zhang, V. Gill, M. Nanduri, and R. Wattenhofer, "Achieving High Utilization with Software-Driven WAN," in *Proc. ACM SIGCOMM*, 2013.

[38] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet Topology Zoo," *IEEE Journal on Selected Areas in Communications*, 2011.

[39] M. Roughan, M. Thorup, and Y. Zhang, "Traffic Engineering with Estimated Traffic Matrices," in *Proc. ACM IMC*, 2003.

[40] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of Youtube Network Traffic at a Campus Network–Measurements, Models, and Implications," *Computer networks*, 2009.

[41] A. Gupta, R. MacDavid, R. Birkner, M. Canini, N. Feamster, J. Rexford, and L. Vanbever, "An Industrial-Scale Software Defined Internet Exchange Point," in *Proc. USENIX NSDI*, 2016.

[42] R. Mahajan, D. Wetherall, and T. Anderson, "Negotiation-based routing between neighboring ISPs," in *Proc. USENIX NSDI*, 2005.

[43] G. Shrimali, A. Akella, and A. Mutapcic, "Cooperative Interdomain Traffic Engineering Using Nash Bargaining and Decomposition," *IEEE/ACM Transactions on Networking (TON)*, 2010.

[44] J. G. Herrera and J. F. Botero, "Resource allocation in nfv: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, 2016.

[45] E. Hernandez-Valencia, S. Izzo, and B. Polonsky, "How will nfv/sdn transform service provider opex?" *IEEE Network*, vol. 29, no. 3, pp. 60–67, 2015.