Technische Universität Berlin

# Forschungsberichte der Fakultät IV – Elektrotechnik und Informatik

## Content-aware Traffic Engineering


Benjamin Frank
Ingmar Poese
Georgios Smaragdakis
Steve Uhlig
Anja Feldmann



Bericht-Nr. 2012 – 03
ISSN 1436-9915

# Content-aware Traffic Engineering

Benjamin Frank
T-Labs/TU Berlin
bfrank@net.t-labs.tu-berlin.de

Ingmar Poese
T-Labs/TU Berlin
ingmar@net.t-labs.tu-berlin.de

Georgios Smaragdakis
T-Labs/TU Berlin
georgios@net.t-labs.tu-berlin.de

Steve Uhlig
Queen Mary University of London
steve@eecs.qmul.ac.uk

Anja Feldmann
T-Labs/TU Berlin
anja@net.t-labs.tu-berlin.de

## ABSTRACT

Today, a large fraction of Internet traffic is originated by Content Providers (CPs) such as content distribution networks and hyper-giants. To cope with the increasing demand for content, CPs deploy massively distributed infrastructures. This poses new challenges for CPs as they have to dynamically map end-users to appropriate servers, without being fully aware of network conditions within an ISP as well as the end-users network locations. Furthermore, ISPs struggle to cope with rapid traffic shifts caused by the dynamic server selection process of CPs.

In this paper, we argue that the challenges that CPs and ISPs face separately today can be turned into an opportunity. We show how they can jointly take advantage of the deployed distributed infrastructures to improve their operation and end-user performance. We propose *Content-aware Traffic Engineering* (CaTE), which dynamically adapts the traffic demand for content hosted on CPs by utilizing ISP network information and end-user location during the server selection process. As a result, CPs enhance their end-user to server mapping and improve end-user experience, thanks to the ability of network-informed server selection to circumvent network bottlenecks. In addition, ISPs gain the ability to partially influence the traffic demands in their networks. Our results with operational data show improvements in path length and delay between end-user and the assigned CP server, network wide traffic reduction of up to $15\%$, and a decrease in ISP link utilization of up to $40\%$ when applying CaTE to traffic delivered by a small number of major CPs.

## 1. INTRODUCTION

People value the Internet for the content it makes available [35]. For example, the demand for online entertainment and web browsing has exceeded 70% of the peak downstream traffic in the United States [34]. Recent traffic studies [27, 40, 52] show that a large fraction of Internet traffic is originated by a small number of Content Providers (CPs). Major CPs are highly popular rich media sites like YouTube and Netflix, One-Click Hosters (OCHs), e. g., RapidShare or MegaUpload, as well as Content Delivery Networks (CDN) such as Akamai or Limelight and hyper-giants, e. g., Google, Yahoo! or Microsoft. Gerber and Doverspike [27] report that a few CPs account for more than half of the traffic of a US-based Tier-1 carrier. Poese et al. [52] report a similar observation from the traffic of a European Tier-1 carrier. Labovitz et al. [40] infer that more than 10% of the total Internet inter-domain traffic originates from Google, and Akamai claims to deliver more than 20% of the total Web traffic in the Internet [50]. In North America, Netflix is responsible for around 30% of the traffic during peak hours [34] by offering a high definition video streaming service hosted on CDN infrastructures such as Limelight and the CDN operated by Level3.

To cope with the increasing demand for content, CPs deploy massively distributed server infrastructures [42] to replicate content and make it accessible from different locations in the Internet [62, 2]. For example, Akamai operates more than $60,000$ servers in more than $5,000$ locations across nearly $1,000$ networks [42, 50]. Google is reported to operate tens of data-centers and front-end server clusters worldwide [39, 61]. Microsoft has deployed its CDN infrastructure in 24 locations around the world [33]. Amazon maintains at least 5 large data-centers and caches in at least 21 locations around the world [55]. Limelight operates thousands of servers in more than 22 delivery centers and connects directly to more than 900 networks worldwide [49].

The growth of demand for content and the resulting deployment of content delivery infrastructures pose new challenges to CPs and to ISPs. For CPs, the cost of deploying and maintaining such a massive infrastructure has significantly increased during the last years [53] and the revenue from delivering traffic to end-users has decreased due to the intense competition. Furthermore, CPs struggle to engineer and manage their infrastructures, replicate content based on end-user demand, and assign users to appropriate servers.

The latter is challenging as end-user to server assignment is based on inaccurate end-user location information [47, 12], and inferring the network conditions within an ISP without direct information from the network is difficult. Moreover, due to highly distributed server deployment and adaptive server assignment, the traffic injected by CPs is volatile. For example, if one of its locations is overloaded, a CP will re-assign end-users to other locations, resulting in large traffic shifts in the ISP network within minutes. Current traffic engineering by ISP networks adapts the routing and operates on time scales of several hours, and is therefore too slow to react to rapid traffic changes caused by CPs.

The pressure for cost reduction and customer satisfaction that both CPs and ISPs are confronted with, coupled with the opportunity that distributed server infrastructures offer, motivate us to propose a new tool in the traffic engineering landscape. We introduce *Content-aware Traffic Engineering* (CaTE). CaTE leverages the location diversity offered by CPs and, through this, it allows to adapt to traffic demand shifts. In fact, CaTE relies on the observation that by selecting an appropriate server among those available to deliver the content, the path of the traffic in the network can be influenced in a desired way. Figure 1 illustrates the basic concept of CaTE. The content requested by the client is in principle available from three servers (A, B, and C) in the network. However, the client only connects to one of the network locations. Today, the decision of where the client will connect to is solely done by the CP and is partially based on measurements and/or inference of network information and end-user location. With CaTE the decision on end-user to server assignment can be done jointly between the CP and ISP.

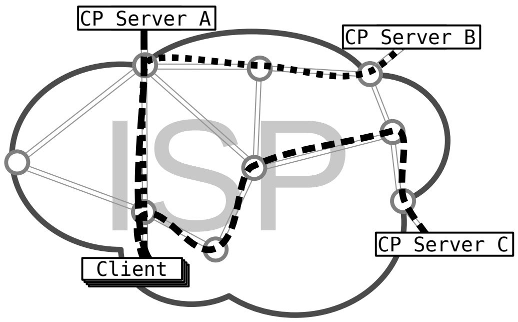


**Figure 1: By choosing a CP server for a client with the help of CaTE, traffic engineering goals and accurate end-user server assignment become possible.**

CaTE complements the existing traffic engineering ecosystem by focusing on traffic demands rather than routing, by combining (i) the knowledge of CPs about their location diversity and server load, with (ii) the ISPs detailed knowledge of the network conditions and end-user location. CaTE offers additional traffic engineering capabilities to ISPs to better manage the volatility of CP traffic. Also, thanks to the information about ISP networks, CPs gain the ability to better assign end-users to their servers and better amortize the cost of deploying and maintaining their infrastructure. Furthermore, the burden of measuring and inferring network topology and state is removed from the CPs. In short, all involved parties, including the end-users, benefit from CaTE, creating a win-win situation for everyone. Our contributions are as follows:

- We introduce the concept of CaTE.
- We present the design, incentives, and possible deployment schemes of systems to realize CaTE.
- We propose an online algorithm to map end-user requests to servers for CaTE and discuss its properties.
- We evaluate the performance of CaTE using real data from a European Tier-1 ISP. We show that CaTE can improve the assignment of end-users to servers for a number of metrics, namely, link utilization, path length and path delay. Our results show that the maximum link utilization can be reduced by half, especially during the peak hour, that the total traffic that flows in the network can be reduced by up to 15%, and the delay by 20% respectively when applying CaTE to a small number of major CPs. Similar results are obtained when evaluating CaTE on two other operational networks.

The remainder of this paper is structured as follows. In Section 2 we present the observations that motivate our work. In Section 3 we introduce our concept of CaTE and present the general architecture as well as possible deployment schemes. We formally define and model CaTE in Section 4. We propose algorithms to enable CaTE in Section 5. We evaluate the benefits of CaTE in Section 6 using data from operational networks with different metrics, including link utilization, path delay and length. We present related work in Section 7 and summarize in Section 8.

## 2. CHALLENGES AND OPPORTUNITIES IN CONTENT DISTRIBUTION

With the emergence of "hyper-giants" and other popular CPs, the traffic of the Internet has undergone drastic changes [40]. These changes stem from trends in business and organizational integration and consolidation. As a consequence, a small number of CPs are responsible for a large fraction of traffic [27, 52]. Content delivered by CPs, including highly popular rich media sites like Facebook and high definition video streaming such as Netflix or YouTube, is mostly carried over HTTP. Recent studies unveil that HTTP contributes more than 60% of Internet traffic [3, 17, 27, 34, 40, 46].

Moreover, CPs peer directly with a large number of ISPs and in many locations. For scalability reasons, most CPs make the content available from all their infrastructure locations [62]. The globally deployed infrastructures allow CPs to rapidly shift large amounts of traffic from one peering point to another. While the diverse footprint of CPs and the ability to shift traffic in short timescales poses new challenges to both CPs and ISPs, it also offers new opportunities for joint optimization of content delivery.

### 2.1 Challenges in Content Delivery

The scale and complexity of content delivery, especially from distributed infrastructures, brings multiple challenges to CPs. These challenges have a major impact on both the end-user performance and ISP operation.

**Content Delivery Cost.** CPs strive to minimize the overall cost of delivering huge amounts of content to end-users. To that end, their assignment strategy is mainly driven by economic aspects such as bandwidth or energy cost [53, 28]. While a CP will try to assign end-users in such a way that the server can deliver reasonable performance, this does not always result in end-users being assigned to the server able to deliver the best performance. Moreover, the intense competition in the content delivery market has led to diminishing returns of delivering traffic to end-users.

**End-user Mis-location.** End-user mapping requests received by the CP DNS servers originate from the DNS resolver of the end-user, not from the end-user itself. The assignment is therefore based on the assumption that end-users are close to their DNS resolvers. Recent studies have shown that in many cases this assumption does not hold [1, 47]. As a result, the end-user is mis-located and the server assignment is not optimal. As a response, DNS extensions have been proposed to include the end-user IP information [12].

**Network Bottlenecks.** Despite their efforts to discover the paths between the end-users and their servers to predict performances [32], CPs have limited information about the actual network conditions. Tracking the ever changing network conditions, i. e., through active measurements and end-user reports, incurs an extensive overhead for the CP without a guarantee of performance improvements for the end-user. Without sufficient information about the network paths between the CP servers and the end-user, an assignment performed by the CP can lead to additional load on existing network bottlenecks, or create new ones.

**End-user Performance.** Applications delivered by CPs often have requirements in terms of end-to-end delay [39]. Moreover, faster and more reliable content delivery results in higher revenues for e-commerce applications [50] as well as user engagement [15]. Despite the significant efforts of CPs, end-user mis-location and the limited view of network bottlenecks are major obstacles to improve end-user performance.

### 2.2 Opportunities for CaTE

The idea behind CaTE is to provide solutions for the new challenges in content delivery. Indeed, ISPs are in a unique position, both in terms of knowledge as well as incentives, to improve content delivery. ISPs have the knowledge about the state of the underlying network topology and the status of individual links. This information not only helps CPs in their user-to-server mapping, but also reduces the need for CPs to perform large-scale active measurements and topology discovery [32]. It also enables CPs to better amortize their existing infrastructure, offer better quality of experience to their users, and postpone their infrastructure expansion.

The opportunity for ISPs to coordinate with CPs in their server selection is technically possible thanks to the decoupling of the server selection from the content delivery. In general, any end-user requesting content from a CP first does a mapping request, usually through the Domain Name System (DNS). During this request the CP needs to locate the network position of the end-user and assign a server capable of delivering the content, preferably close to the end-user. However, locating the user in a network and inferring the conditions of the path between the end-user and eligible CP servers is hard as the CP is missing network information. In contrast, ISPs have this information ready at their fingertips, but are currently missing a communication channel to inform the CPs. Furthermore, ISPs face the challenge of predicting the CP traffic, which is very difficult due to the lack of information on the mapping of end-users to server decided by CPs.

We propose to use CaTE during the server selection process of CPs. In today's CP deployment, the server selection is done directly between the end-user and the CP without the involvement of the ISP (see arrow A in Figure 2). Through CaTE, CPs are offered the opportunity to optimize their server selection beyond their current capabilities by communicating directly with the ISP (CP-ISP Communication, see Figure 2). Furthermore, ISPs gain the ability of adapting to the volatile traffic induced by content delivery, by being able to influence the choice of the CP. We believe that CaTE is a step forward in improving the end-user performance and enabling ISP and CP collaboration.

## 2.3 Incentives

The opportunities that CaTE enables for both CPs and ISPs require that both parties have incentives to work together. Furthermore, the growing awareness of end-users about CaTE's benefits will accelerate the penetration of CaTE in a highly commoditized content delivery market.

### 2.3.1 *Incentives for CPs*

The market of CPs requires them to enable new applications while reducing their operational cost, and to improve the end-user experience [50]. With CaTE improving the mapping of end-users to servers, CPs can expect improvements in the end-user experience, and thus, a competitive advantage. This is particularly important for CPs in light of the commoditization of the content delivery market and the choice that is offered to end-users, for example through meta-CDNs [15]. The improved mapping also yields better infrastructure amortization and thanks to CaTE, CPs will no longer have to perform and analyze voluminous measurements in order to infer the network conditions or end-user locations.

To stimulate the use of CaTE, ISPs can operate and provide CaTE as a free service to CPs or even offer discounts on peering or hosting prices, e. g., for early adopters and CPs that expose a higher server diversity while using CaTE. The loss of peering or hosting revenue is amortized with the benefits of a lowered network utilization, reduced investments in network capacity expansion and by taking back some control over the traffic within the network. Ma et al. [45] have developed a methodology to estimate the prices in such a cooperative scheme by utilizing the Shapley settlement mechanism. CaTE can also act as an enabler for CPs and ISPs to jointly launch new applications in a cost-effective way, for example traffic-intensive applications such as the delivery of high definition video on-demand, or real-time applications such as online games. In an ISP-CP collaborative scheme, CaTE can play the role of a recommendation system and is not intended to be applied unilaterally by the ISP.

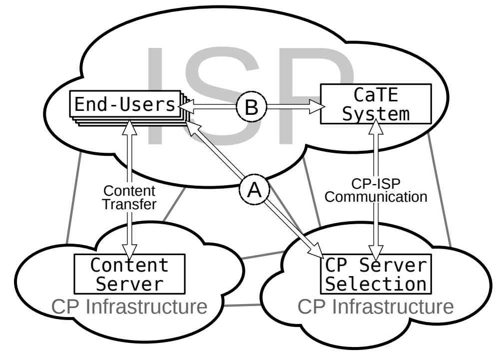


**Figure 2: CaTE deployment and interaction with CPs.**

### 2.3.2 *Incentives for ISPs*

ISPs are interested in reducing their operational and infrastructure upgrade costs, offering broadband services at competitive prices, and delivering the best end-user experience possible. Due to network congestion during the peak hour, ISPs in North America have recently revisited the flat pricing model and have announced data caps to broadband services. A better management of traffic in their network with CaTE can allow them to offer higher data caps or even alleviate the need to introduce them. From an ISP perspective, CaTE offers the possibility to do global traffic and peering management, through an improved awareness of the traffic across the whole network. For example, peering agreements with CPs can offer the use of CaTE in exchange for reduced costs to the CPs. This can be an incentive for CPs to peer with a CaTE-enable ISP and an additional revenue for an ISP, as such reduced prices can attract additional peering customers. An ISP can also offer CaTE to other ISPs it peers with, which makes sense especially in the case that the peering ISPs hosts content or also acts as CP. The interaction and federation of CPs run by ISPs can also be enabled through CaTE. There is high interest on the side of ISPs, as reflected by the creation of the IETF working group CDNi [44]. Furthermore, CaTE has the potential to reduce the significant overhead due to the handling of customer complaints that often do not stem from the operation of the ISP but the operation of CPs [8]. With CaTE, ISPs can identify and mitigate congestion, and react to short disturbances caused by an increased demand of content from CPs.

### 2.3.3 *Incentives for end-users*

CaTE offers a way to empower end-users to obtain the best possible quality of experience. As such, this creates an incentive for end-users to support the adoption of CaTE by both ISPs and CPs. For example, an ISP can offer more attractive products, i. e., higher bandwidth or lower prices, since it is able to better manage the traffic inside its network. Also, thanks to better traffic engineering, ISPs can increase data caps on their broadband offers, making the ISP more attractive to end-users. Moreover, CPs that utilize CaTE can offer better quality of experience to end-users. This can be done through premium services based on CaTE. For example, CPs delivering streaming services can offer higher quality videos to end-users thanks to better server assignment and network engineering. Also, applications running over the Internet can greatly benefit in their performance from CaTE (see Appendix B). This, in turn, gives end-users a good reason to choose CaTE enabled services.

# 3. CaTE APPROACH

The concept of CaTE relies on two key observations. First, a major fraction of the traffic in ISPs is delivered by massively distributed CP infrastructures. Therefore, the same content is often available at different network locations with different network paths to the end-user. Second, the server selection of CPs is decoupled from the content transfer. Thus, it is possible to augment the server selection strategy of CPs with detailed information from ISPs about the current network state, the status of links that are traversed and the precise network location of the end-user.

## 3.1 Concept of CaTE

CaTE relies on the fact that by selecting an appropriate server among those being able to satisfy a request, the flow of traffic within the network can be influenced. To illustrate the concept, we show in Figure 1 how, by selecting server A instead of B or C, a shorter path through the network is chosen. However, CPs have limited knowledge about the path characteristics inside a network. On the other hand, ISPs are aware of the state of their network, the location of their users, as well as the path conditions between end-users and servers. Given the large fraction of traffic that originates from CPs and their highly distributed infrastructure, CaTE can shift traffic among paths within a network and, through this, achieve traffic engineering goals for both CPs and the ISP.

## 3.2 CaTE Deployment Schemes

Our main architectural motivation is that the server selection is decoupled from the content transfer. In Figure 2 we provide a simplified version of how CPs handle content requests. Today, the server selection process of CPs works as follows. When an end-user wants to obtain a specific content, it first sends a request to the *CP server selection* of the CP (see Figure 2, *(A)*). Today, there are two prevalent techniques used to transfer this request: DNS queries and HTTP redirection. The *CP server selection* selects the content server based on the requested content, the objectives of the CP, its current view of the network, and its knowledge of the end-user network location. Finally, it returns the selected server IP, either through a DNS reply or a HTTP redirection, to the end-user, which in turn establishes a connection to the supplied server IP to download the content.

In order for CaTE to hook into the server selection of CPs, a new component inside the ISPs network is needed. In general, this component offers an interface between the CP and the ISP to get supplement information about the network position of end-users, path conditions between an end-user and eligible servers, etc. To this end, the system uses information readily available to an ISP, such as the actual network topology, routing information, end-user assignment databases, current network loads, etc. Today, systems capable of providing the interface between an ISP and a CP are for example the IETF ALTO service [4] or the Provider-aided Distance information System (PaDIS) [52]. In Figure 2 we outline the range of possible CaTE deployment schemes:

**1. CP contacts ISP:** The end-user contacts the CP server selection module via its DNS resolver *(A)* as it does today. When choosing the server for the end-user, the CP uses the *CP-ISP Communication* to retrieve information about the network status, topology, or a recommendation by the ISP based on the network conditions between the end-user and the candidate content servers. The advantage of the recommendation option is that no party reveals any sensitive operational information.

This can be implemented by including the client IP in the mapping request as proposed at the IETF dnsext working group [12] while using the IETF ALTO protocol or PaDIS by the CP to retrieve topology information, network status information, or server recommendation by the ISP.

**2. ISP contacts CP:** The end-user contacts CaTE directly *(B)* for the mapping. Then, CaTE uses the *CP-ISP Communication* to forward the request to the CP. The CP returns a list of potential servers and CaTE ranks them based on network characteristics and the current path conditions between end-user and server network location.

This can be implemented by utilizing the part of the DNS resolution process handled by CPs. When end-users query the ISP DNS resolver and, in turn, the CP DNS server, the CP returns all candidate content servers, which are re-ordered by the ISP DNS resolvers according to CaTE.

**3. ISP-based:** The end-user contacts CaTE directly *(B)* for the mapping. However, CaTE forwards the request through the *CP-ISP Communication* to the CP server selection, which returns the normal reply as it happens today. CaTE collects and aggregates the replies from the CP and overwrites the replies using the knowledge it has obtained from past results.

This can be implemented by using the DNS resolution process of CPs. When end-users query the ISP DNS resolver the ISP forwards the request. However, the answer from the CP is kept and aggregated as proposed by Poese et al. [52] and the DNS replies are overwritten as CaTE sees fit.

**4. User-based:** The end-user collects the potential content servers from the CP as well as the current network state from the ISP. By utilizing this information, it calculates the best server to connect to based on active end-to-end measurements or previously reported experience.

This can be achieved when both the CP and the ISP run the IETF ALTO service or PaDIS. In this case, the client downloads all the needed information and performs the server selection itself.

In the first three schemes CaTE can be incrementally deployed and interacts with the existing CP infrastructures while being transparent to the end-user. In the collaborative schemes 1 and 2, the final decision is made by the CPs to avoid any disturbance on their operation. The frequency of ranking exchanges as well as the granularity of end-user location identification is up to the administrator of the system. It is also possible to provide end-users the choice to opt-in or opt-out. CPs can also negotiate how many locations they make available to ISPs. Note, CPs can dynamically change the locations made available to the ISP depending on the utilization of each location. In the last deployment option, we describe how CaTE can also be deployed at the end-user, e. g., via the browser or home gateway, but the penetration will be slower as it requires the installation of software at the end-user.

# 4. MODELLING CaTE

Next, we formalize CaTE and discuss how it relates to traditional traffic engineering and multipath routing.

## 4.1 Traffic Engineering

We model the network as a directed graph $G(V, E)$ where $V$ is the set of nodes and $E$ is the set of links. An origin-destination (OD) flow $f_{od}$ consists of all traffic entering the network at a given point $o \in V$ (origin) and exiting the network at some point $d \in V$ (destination). The traffic on a link is the superposition of all OD flows that traverse the link.

The relationship between link and OD flow traffic is expressed by the routing matrix $A$. The matrix $A$ has size $|E| \times |V|^2$. Each element of matrix $A$ has a boolean value. $A_{ml} = 1$ if OD flow $m$ traverses link $l$, and 0 otherwise. The routing matrix $A$ can be derived from routing protocols, e.g., OSPF, ISIS, BGP. Typically, $A$ is very sparse since each OD flow traverses only a very small number of links. Let $\mathbf{y}$ be a vector of size $|E|$ with traffic counts on links and $\mathbf{x}$ a vector of size $|V|^2$ with traffic counts in OD flows, then $\mathbf{y}=A\mathbf{x}$. Note, $\mathbf{x}$ is the vector representation of the traffic matrix.

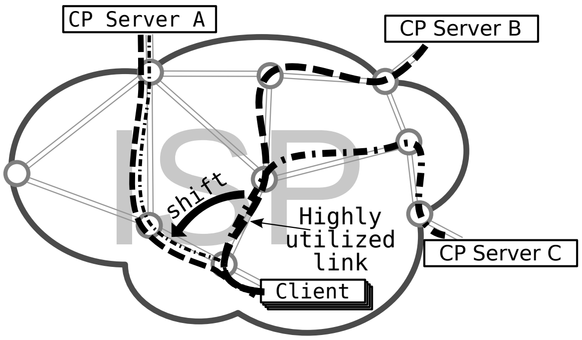


**Figure 3: Content-aware Traffic Engineering Process**

**Traditional Traffic Engineering:** In its broadest sense, traffic engineering encompasses the application of technology and scientific principles to the measurement, characterization, modeling, and control of Internet traffic [5]. Traditionally, traffic engineering reduces to controlling and optimizing the routing function and to steering traffic through the network in the most effective way. Translated into the above matrix form, traffic engineering is the process of adjusting $A$, given the OD flows **x**, so as to influence the link traffic **y** in a desirable way, as coined in [41]. The above definition assumes that the OD flow vector **x** is known. For instance, direct observations can be obtained, e. g., with Netflow data [9, 19].

**Terminology:** We denote as *flow* an OD flow between two routers in the network. We call a flow *splittable* if arbitrarily small pieces of the flow can be assigned to other flows. This is not to be confused with end-to-end sessions, i. e., TCP connections, which are *un-splittable*. The assumption that flows are splittable is reasonable, as the percentage of traffic of a single end-to-end session is small compared to that of a flow between routers. Let $C$ be the set of nominal capacities of the links in the network $G$. We denote as *link utilization* the fraction of the link capacity that is used by flows. We denote as *flow utilization* the maximum link utilization among all links that a flow traverses. We introduce the terms of *traffic consumer* and *traffic producer* which refer to the aggregated demand of users attached to a router, and the CPs that are responsible for the traffic respectively. Throughout this paper, we refer to the different alternatives from which content can be supplied by a given CP as *network locations* that host servers.

## 4.2 Definition of CaTE

We revisit traffic engineering by focusing on the traffic demands rather than changing the routing.

**Definition 1: Content-aware Traffic Engineering(CaTE)** is the process of adjusting the traffic demand vector **x**, given a routing matrix $A$, so as to change the link traffic **y**.

Not all the traffic can be adjusted arbitrarily. Only traffic for which location diversity is available can be adjusted by CaTE. Therefore, $\mathbf{x}=\mathbf{x}_r+\mathbf{x}_s$ where $\mathbf{x}_r$ denotes the content demands that can be adjusted and $\mathbf{x}_s$ denotes the content demands that can not be adjusted as there is only a single location in the network where the content can be downloaded from. The amount of traffic that can be adjusted depends on the diversity of locations from which the content can be obtained. We can rewrite the relation between traffic counts on links and traffic counts in flows as follows: $\mathbf{y}=A(\mathbf{x}_s+\mathbf{x}_r)$. CaTE adjusts the traffic on each link of the network by adjusting the content demands $\mathbf{x}_r$: $\mathbf{y}_r=A\mathbf{x}_r$. Applying CaTE means adjusting the content demand to satisfy a traffic engineering goal.

**Definition 2: Optimal Traffic Matrix** is the new traffic matrix, $\mathbf{x}^*$, after applying CaTE, given a network topology $G$, a routing matrix $A$ and an initial traffic matrix **x**.

Figure 3 illustrates the CaTE process. A content consumer requests content that three different servers can deliver. Let us assume that, without CaTE, the CP redirects the clients to servers B and C. Unfortunately, the resulting traffic crosses a highly-utilized link. With CaTE, content can also be downloaded from server A, thus, the traffic within the network is better balanced as the highly utilized link is circumvented.

Minimizing the maximum utilization across all links in a network is a popular traffic engineering goal [24, 25, 42]. It potentially improves the quality of experience and postpones the need for capacity increase. CaTE mitigates bottlenecks and minimizes the maximum link utilization by re-assigning parts of the traffic traversing heavily loaded paths. Thus it redirects traffic to other, less utilized paths. As we will elaborate in Section 6, different metrics such as path length or network delay can also be used in CaTE.

## 4.3 CaTE and Traditional TE

CaTE is complementary to routing-based traffic engineering as it does not modify the routing. Routing-based traffic engineering adjusts routing weights to adapt to traffic matrix changes. To avoid micro-loops during IGP convergence [26], it is common practice to only adjust a small number of routing weights [25]. To limit the number of changes in routing weights, routing-based traffic engineering relies on traffic matrices computed over long time periods and offline estimation of the routing weights. Therefore, routing-based traffic engineering operates on time scales of hours, which can be too slow to react to rapid change of traffic demands. CaTE complements routing-based traffic engineering and can influence flows at shorter time scales by assigning clients to servers on a per request basis. Thus, CaTE influences the traffic within a network online in a fine-grained fashion.

## 4.4 CaTE and Multipath Routing

Multipath routing helps end-hosts to increase and control their upload capacity [37]. It can be used to minimize transit costs [28]. Multipath also enables ASes to dynamically distribute the load inside networks in the presence of volatile and hard to predict traffic demand changes [19, 16, 58, 21]. This is a significant advantage, as routing-based traffic engineering can be too slow to react to phenomena such as flash crowds. Multipath takes advantage of the diversity of paths to better distribute traffic.

CaTE also leverages the path diversity, and can be advantageously combined with multipath to further improve traffic engineering and end-user performance. One of the advantages of CaTE is its limited investments in hardware deployed within an ISP. It can be realized with no change to routers, contrary to some of the previous multipath proposals [58, 16, 21]. The overhead of CaTE is also limited as no state about individual TCP connections needs to be maintained, contrary to multipath [58, 16, 21]. In contrast to [16, 58], CaTE is not restricted to MPLS-like solutions and is easily deployable in todays networks.

## 4.5 CaTE and Oscillations

Theoretical results [23, 22] have shown that load balancing algorithms can take advantage of multipath while provably avoiding traffic oscillations. In addition, their convergence is fast. Building on these theoretical results, Fischer et al. proposed REPLEX [21], a dynamic traffic engineering algorithm that exploits the fact that there are multiple paths to a destination. It dynamically changes the traffic load routed on each path. Extensive simulations show that REPLEX leads to fast convergence, without oscillations, even when there is lag between consecutive updates about the state of

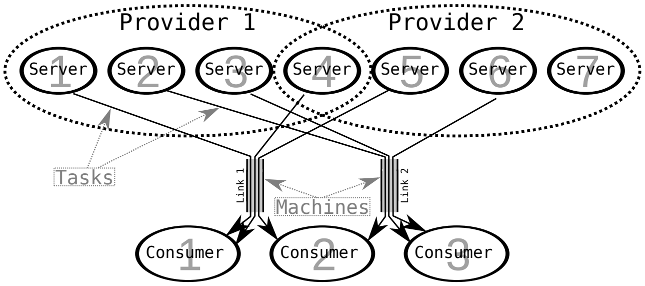


**Figure 4: CaTE and Restricted Machine Load Balancing.**

the network. CaTE is derived from the same principles and thus inherits all the above-mentioned desired properties.

## 5. CaTE ALGORITHMS

In this section we propose algorithms to realize CaTE, in the context of an ISP. A key observation is that CaTE can be reduced to the restricted machine load balancing problem [7] for which optimal online algorithms are available. The benefit of the CaTE online algorithm can be estimated either by reporting results from field tests within an ISP or by using trace-driven simulations. Typically, in operational networks only aggregated monitoring data is available. To estimate the benefit that CaTE offers to an ISP, we present offline algorithms that uses traffic demands and server diversity over time extracted from those statistics as input.

### 5.1 Connection to Restricted Machine Load Balancing

Given a set of CPs and their network location diversity, we consider the problem of re-assigning the flows that correspond to demands of content consumers to the CPs in such a way that a specific traffic engineering goal is achieved. Given that sub-flows between end-systems and content provider servers can be re-distributed only to a subset of the network paths, we show that the solution of the optimal traffic matrix problem corresponds to solving the *restricted machine load balancing problem* [7]. In the restricted machine load balancing problem, a sequence of tasks is arriving, where each task can be executed by a subset of all the available machines. The goal is to assign each task upon arrival to one of the machines that can execute it so that the total load is minimized. Note, contrary to the case of multipath where paths between only one source-destination pair are utilized, CaTE can utilize any eligible path between any candidate source and destination of traffic.

For ease of presentation let us assume that the traffic engineering goal is to minimize the maximum link utilization in the network [24, 25]. Let us consider three consumers where each one wants to download one unit of content from two different content providers, see Figure 4. Given that different servers can deliver the content on behalf of the two providers, the problem consists in assigning consumers to servers in such a way that their demands are satisfied while minimizing the maximum link utilization in the network. Thus, the problem is the restricted machine load balancing one where tasks are the demands satisfied by the servers and machines are the bottleneck links that are traversed when a path, out of all eligible server-consumer paths, is selected. Figure 4 shows one of the possible solutions to this problem, where consumer 1 is assigned to servers 1 and 4, consumer 2 to servers 5 and 2, and consumer 3 to servers 3 and 6. Note that the machine load refers to the utilization of the bottleneck links of eligible paths, denoted as link 1 and 2.

To be consistent with our terminology, we define the *restricted flow load balancing problem*. Let $J$ be the set of the consumers in the network, $K$ be the set of content producers, and $I$ be the set of servers for a given content provider, i.e., the set of locations where a request can be satisfied. Note, this set is offered by the CP in order to satisfy its own objectives and can change over time. We denote as $M_{jk}$ the set of flows that can deliver content for a given content producer $k$ to consumer $j$.

**Definition 3: Restricted Flow Load Balancing Problem** is the problem of finding a feasible assignment of flows such that a traffic engineering goal is achieved, given a set of sub-flows $\{f_{ijk}\}$ from all eligible servers $i \in I$ of a given content provider $k \in K$ to a consumer $j \in J$, and a set of eligible residual flows $f_{ij}^{-k}$, $i \in M_{jk}$ (after removing the traffic of the above mentioned sub-flows).

Despite some similarities, the nature of our problem differs from the multi-commodity flow and bin packing. In the multi-commodity flow problem [6], the demand between source and destination pairs is given while in our problem the assignment of demands is part of the solution. In the bin packing problem [11], the objective is to minimize the number of bins, i.e., number of flows in our setting, even if this means deviating from the given traffic engineering goal. Note, in the restricted flow load balancing problem any eligible path from a candidate source to the destination can be used, contrary to the multipath problem where only equal-cost paths can be used.

### 5.2 Online Algorithm and Competitiveness

We next turn to the design of online algorithms. It has been shown that in the online restricted machine load balancing problem, the greedy algorithm that schedules a permanent task to an eligible processor having the least load is exactly optimal [7], i. e., it is the best that can be found, achieving a competitive ratio of $\lceil \log_2 n \rceil + 1$, where $n$ is the number of machines. If tasks are splittable then the greedy algorithm is 1-competitive, i. e., it yields the same performance as an offline optimal algorithm. The greedy algorithm is an online one, thus it converges to the optimal solution immediately without oscillations.

In the restricted flow load balancing problem, the set $M_{jk}$ can be obtained from the set of candidate servers that can deliver content when utilizing CaTE as described in Section 3.2. The online assignment of users to servers per request, which minimizes the overall load, leads to an optimal assignment of sessions within subflows. In our case, flows are splittable since the content corresponding to each content request is negligible compared to the overall traffic traversing a link. Note, the end-to-end TCP connections are not splittable. Thus, the following online algorithm is optimal:

**Algorithm 1. Online Greedy Server Selection.** Upon the arrival of a content user request, assign the user to the server that can deliver the content, out of all the servers offered by the CP, such that the traffic engineering goal is achieved.

### 5.3 Estimating the Benefit of CaTE with Passive Measurements

Before applying CaTE in real operational networks, it is important to understand the potential benefits that it can bring in a given context. For example, the operator of an ISP network would like to know in advance what are the gains when applying CaTE, as well as being able to answer what-if scenarios, when applying CaTE to traffic delivered by different CPs. Operators of CPs would also like to quantify the benefits by participating in CaTE before collaborating with an ISP. In most operational networks, aggregated statistics and passive measurements are collected to support operational decisions. Therefore, we provide a framework that allows a simulation-driven evaluation of CaTE. To that end, we present

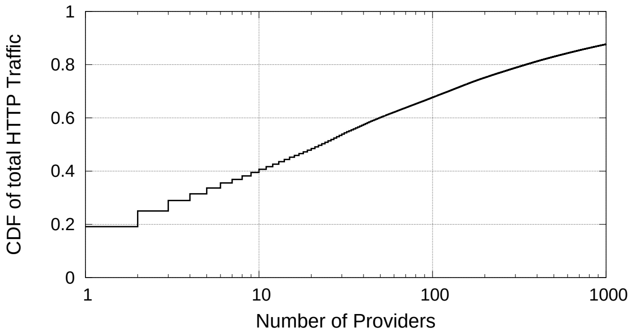


**Figure 5: CDF of traffic volume of CPs in ISP1.**

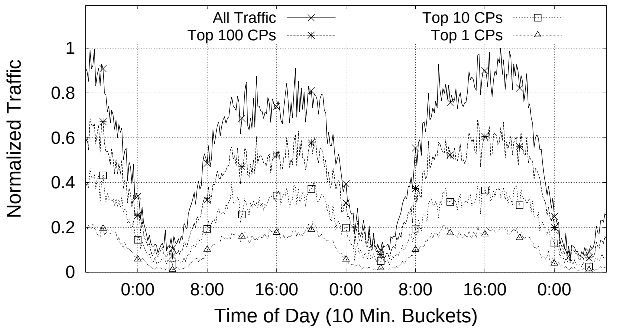


**Figure 6: Normalized traffic for top CPs by volume in ISP1.**

offline algorithms that can take as input passive measurements and evaluate the potential gain when applying CaTE in different scenarios in Appendix A. We propose a linear programming formulation as well as greedy approximation algorithms to speed-up the process of estimating the gain when using CaTE.

# 6. EVALUATION OF CaTE

In this section, we quantify the potential of CaTE with different traffic engineering goals in mind. We evaluate CaTE with operational data from three different networks. For the first network, we rely on content demands built from observed traffic of a European Tier-1 ISP. The other two networks, namely AT&T and Abilene, allow us to evaluate the impact of the ISP topology structure.

## 6.1 Experimental Setting

To evaluate CaTE, an understanding of the studied ISP network is necessary, including its topological properties and their implications on the flow of traffic. Indeed, the topological properties of the ISP network influence the availability of disjoint paths, which are key to benefit from the load-balancing ability of CaTE. Because CaTE influences traffic aggregates inside the ISP network at the granularity of requests directed to CPs, fine-grained traffic statistics are necessary. Traffic counts per-OD flow, often used in the literature, are too coarse an input for CaTE.

### *6.1.1 Data from a Large European ISP*

To build fine-grained traffic demands, we rely on anonymized packet-level traces of residential DSL connections from a large European Tier-1 ISP, henceforth called *ISP1*. For ISP1, we have the complete annotated router-level topology including the router locations as well as all public and private peerings. ISP1 contains more than $650$ routers and $30$ peering points all over the world.

We collect a $10$ days long trace starting on May 7, 2010. Our monitor, using Endace monitoring cards [10], allows us to observe the traffic of more than $20,000$ DSL lines to the Internet. We capture HTTP and DNS traffic using the Bro intrusion detection system [51]. We observe 720 million DNS messages as well as more than 1 billion HTTP requests involving about 1.4 million unique hostnames, representing more than 35 TBytes of data. With regards to the application mix, more than 65% of the traffic volume is due to HTTP. Other popular applications that contribute to the overall traffic volume are NNTP, BitTorrent, and eDonkey.

A large fraction of the traffic in the Internet is due to large CPs, including CDNs, hyper-giants, and OCHs, as reported in earlier studies [27, 40, 52]. In Figure 5, we plot the cumulative fraction of HTTP traffic volume as a function of the CPs that originate the traffic. We define a CP as a organizational unit where all servers from the distributed infrastructure serve the same content, such as Akamai or Google. We rank the CPs by decreasing traffic volume observed in our trace. Note that the x-axis uses a logarithmic scale. The top 10 CPs are responsible for around 40% of the HTTP traffic volume and the top 100 CPs for close to 70% of the HTTP traffic volume. The marginal increase of traffic is diminishing when increasing the number of CPs. This shows that collaborating directly with a small number of large CPs, can yield significant savings.

In Figure 6 we plot the traffic of the top 1, 10, 100 CPs by volume as well as the total traffic over time normalized to the peak traffic in our dataset. For illustrative purposes, we show the evolution across the first $60$ hours of our trace. A strong diurnal pattern of traffic activity is observed. We again observe that a small number of CPs are responsible for about half of the traffic. Similar observations are made for the rest of the trace.

### *6.1.2 Understanding the Location Diversity of CPs*

To achieve traffic engineering goals, it is crucial to also understand the location diversity of the top CPs, as CaTE relies on the fact that the same content is available at multiple locations. Traffic originated from multiple network locations by a given CP is seen by CaTE as a single atomic traffic aggregate to be engineered. Furthermore, as routing in the Internet works per prefix, we assume that the granularity of subnets is the finest at which CaTE should engineer the traffic demand. Thus, we differentiate candidate locations of CPs by their subnets and quantify the location diversity of CPs through the number of subnets from which content can be obtained.

We examine the amount of location diversity offered by CPs based on traces from ISP1. To identify the subnets of individual CPs, we rely on a similar methodology to the one from Poese et al. [52]. Our granularity is comparable to their "infrastructure redirection aggregation". Figure 7 shows the cumulative fraction of HTTP traffic as a function of the number of subnets (logarithmic scale) from which a given content can be obtained, over the entire $10$ days of the trace. We observe that more than $50\%$ of the HTTP traffic can be delivered from at least $8$ different subnets, and more than $60\%$ of the HTTP traffic from more than $3$ locations. These results confirm the observations made in [52].

### *6.1.3 Dynamics in Location Diversity*

So far the location diversity of CPs has been evaluated irrespective of time. To complement the finding, we turn our attention to the location diversity exposed by CPs at small time-scales, i. e., in the order of minutes. To this end, we split the original trace into $10$ minutes bins. Figure 8 shows the evolution of the number of exposed subnets of five of the top 10 CPs by volume. Note that the diversity exposed by some CPs exhibits explicit time of day patterns, while others do not. This can be due to the structural setup or the type of content served by the CP. The exposed location diver-

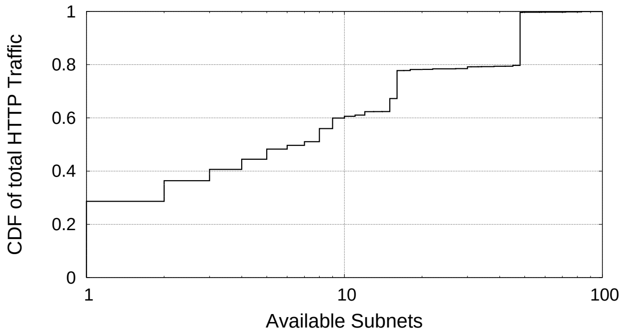


**Figure 7: Subnet diversity from which content is available.**

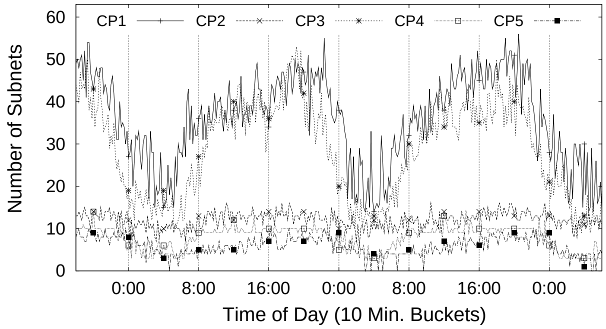


**Figure 8: Evolution over time of number of subnets for selected CPs in the top 10 CPs.**

sity patterns, i. e., flat or diurnal, are representative for all CPs with a major traffic share in our trace. We conclude that a significant location diversity is exposed by popular CPs at any point in time, and is quite extensive during the peak hour.

### 6.1.4 *Content Demand Generation*

The location diversity is not a mere observation about CPs deployment. It requires to revisit the mapping between a given content demand and the realized traffic matrix. Given the location diversity for content, multiple traffic matrices can be realized from a given content demand. The standard view of the OD flows therefore provides an incomplete picture of the options available for CaTE.

As an input for CaTE, we introduce an abstraction of the demand that reflects the available location diversity. We rely on the notion of *potential vectors*, that were denoted as $x_r$ in Section 4.2. To generate the potential vector for a given CP, the amount of traffic this CP originates as well as the potential ingress points need to be known. Combining all potential vectors and $x_s$, we synthesize a network-wide content demand matrix for each time bin, by scaling the traffic demand to match the network utilization of ISP1. For our evaluation, we use the series of content demand matrices over a period of $10$ days. The content demands are based exclusively on the HTTP traffic of our trace.

## 6.2 CaTE in ISP1

To quantify the benefits of CaTE, we first consider one of the most popular traffic engineering goals, namely minimizing the maximum utilization of the links in the network [24, 25]. The rationale is that by minimizing the maximum link utilization, network bottlenecks are reduced, in turn limiting queueing delays, improving the quality of experience and postponing the need for increased network capacity.

With CaTE, an ISP can collaborate with any CP. It is up to the

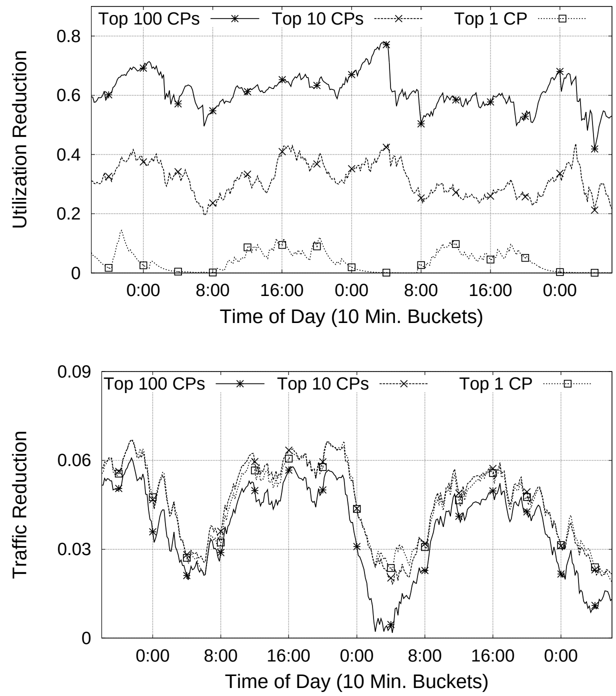


**Figure 9: Maximum link utilization reduction (top) and total traffic reduction (bottom) with CaTE for the top CPs.**

ISP to select the set of CPs that are the most important to establish collaboration with. Since a significant fraction of the traffic originates from a small number of CPs, we consider the most popular CPs by volume to evaluate CaTE. In the following, we perform a sensitivity study where we quantify the benefits of CaTE when restricting its use to the top 1, 10 and 100 CPs by volume. All other traffic remains unaffected by CaTE. For all experiments, we use the Algorithm 2 from Appendix A.2.

**Effect on Maximum Link Utilization**. Figure 9 (top) shows the reduction of the maximum link utilization over a period of $2$ days when considering the top 1, 10 and 100 CPs. Once again, we normalized the absolute link utilization by the maximal one. The largest gain in maximum link utilization reduction is up to $15\%$, $40\%$ and $70\%$ respectively. We observe large fluctuations of the gains which are due to variations in traffic (see Figure 7) and location diversity (Figure 8) throughout the day. The largest gains are obtained during peak times, when there is more traffic and the highest location diversity is available. This is also when congestion is at its peak and CaTE is most needed. Our results show that CaTE is able to react to diurnal changes in traffic volume and utilizes the available location diversity.

**Effect on Network-wide Traffic**. Although optimizing for link utilization, CaTE reduces the overall traffic that flows through the network, see Figure 9 (bottom). This is due to CaTE choosing the shortest path when multiple ones with the same utilization are available, thus, as a side effect, content is fetched from closer locations and therefore traverses less links. With CaTE, the gains in overall traffic reduction are up to $6\%$ and follows a clear diurnal pattern. It is worth noticing that just with the top 10 CPs, the total traffic reduction is very close to the one when considering the top 100 CPs, indicating that CaTE only needs to be implemented with the major players. Also, an ISP that is able to reduce the overall traffic inside its network is more competitive as it can serve more end-users with the same infrastructure, delay additional investments in

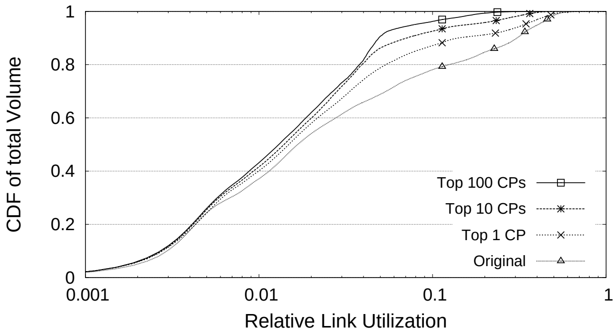


**Figure 10: Improvements in link utilization with CaTE.**

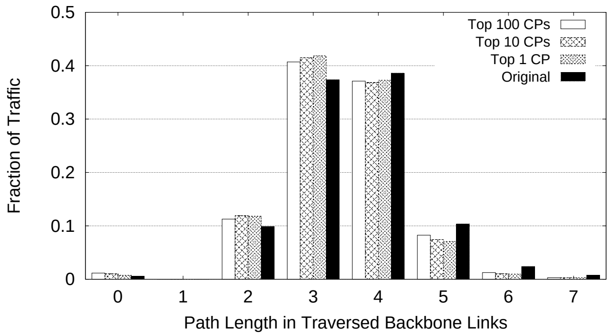


**Figure 11: Backbone path length count with CaTE.**

capacity upgrades and improve end-user satisfaction.

**Effect on Distribution of Link Utilization**. Reducing the maximum link utilization shifts traffic away from congested links. However, it should not be done at the expense of creating congestion on other highly utilized links. In Figure 10 we plot the CDF of traffic volume in ISP1 across all link utilizations, normalized by the maximum one when considering sets of the top CPs by volume. The results show that CaTE shifts the traffic away from highly utilized links to low utilized ones.

**Effect on Traffic Path Length**. Our results in Figure 9 (bottom) show a reduction in the overall traffic in ISP1, which can be attributed to an overall reduction of the path length. Path length reduction is an important metric for ISPs for the dimensioning of the network as well as the reduction of operational costs. To quantify this reduction in terms of the path length inside ISP1, Figure 11 shows the relative traffic across different path lengths inside the network. CaTE redirects the traffic towards paths with the same or even shorter length than the ones used without CaTE, only in the rare case where a longer paths yields a lower utilization, CaTE can choose a longer one. Note that there is no traffic for backbone path length equal to 1 due to the network design of ISP1. We conclude that applying CaTE to a small number of CPs yields major improvements in terms of path length.

**Effect on Path Delay**. Although the objective of minimizing maximum link utilization is not directly related to the reduction of path delay, the achieved reduction in path length directly affects the path delay. Figure 12 shows the accumulated path delay for the traffic that flows within ISP1, when applying CaTE. The reported numbers for the backbone path delay are relatively modest compared to the values for the access part of the network [46]. However, improving the access delay requires significant investments as it can be done mostly through changes in the access technology, e. g., from copper to fiber. When considering the end-to-end delay, the delay of the path outside the ISP's network also needs to be considered. As content infrastructures are located close to peering points [40, 39, 2], e. g., IXPs or private peerings, the delays are expected to be relatively small, especially for popular CPs. Estimating the impact of CaTE on the end-to-end performance for every application is very challenging, due to the many factors that influence flow performance, especially network bottlenecks outside the considered ISP. In Appendix B we show the results from active measurements conducted in the case of traffic-heavy applications, confirming the significant improvements in end-to-end delay as well as download time that can be achieved thanks to CaTE.

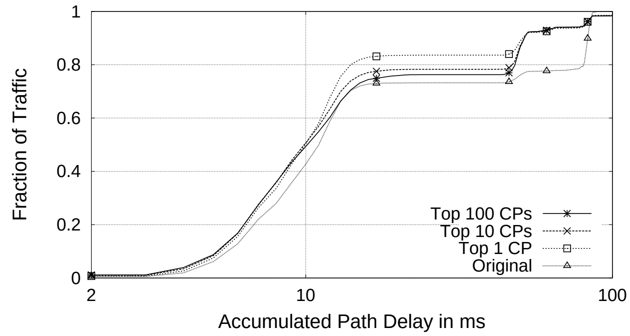


**Figure 12: Improvement in path delay (in ms) with CaTE.**

**Summary**. Our evaluation shows that CaTE yields encouraging results, even when only a few large CPs are collaborating with an ISP. In fact, even metrics that are not directly related to the optimization function of CaTE are improved. Besides significant improvements for the operation of ISP networks, the end-users are expected to also benefit from these gains. This can be attributed to the decrease of delay as well as the reduced link utilization.

## 6.3 CaTE with other Network Metrics

So far we have evaluated CaTE with one traffic engineering objective, namely, the minimization of maximum link utilization. CaTE allows ISPs and CPs to to optimize for other network metrics such as path length or path delay. To this end, we quantify the effects of CaTE when using path length and delay and compare it with the results presented in Section 6.2. We focus on the top 10 CPs as our results show that most of the benefits from CaTE can be achieved with this rather low number of CPs. Similar observations are made when applying CaTE to the top 1 and 100 CPs.

In Figure 13 (top) we plot the total traffic reduction when applying CaTE to the top 10 CPs with different optimization goals. The first observation is that when the network metric is path length, the total traffic reduction is the highest, with up to 15%. The total traffic reduction when optimizing for path length are close to the one achieved when the metric is delay. Optimizing for other metrics provides the expected result: the optimized metric is significantly improved, but at the cost of not optimizing other metrics as much. For example, optimizing for link utilization diminishes the benefits from path length (Figure 14 top) and vice-versa (Figure 13 bottom). Still, significant improvements can be achieved even when optimizing for another network metric and we encountered no case of significant deterioration in on of the network metrics throughout our experiments, see Figure 13 and Figure 14.

## 6.4 CaTE in AT&T and Abilene

To quantify the potential benefits of CaTE in networks with different topological structures than ISP1, we repeat our experiments for two other ISPs: AT&T and Abilene.

**AT&T** is one of the largest commercial networks. We use the

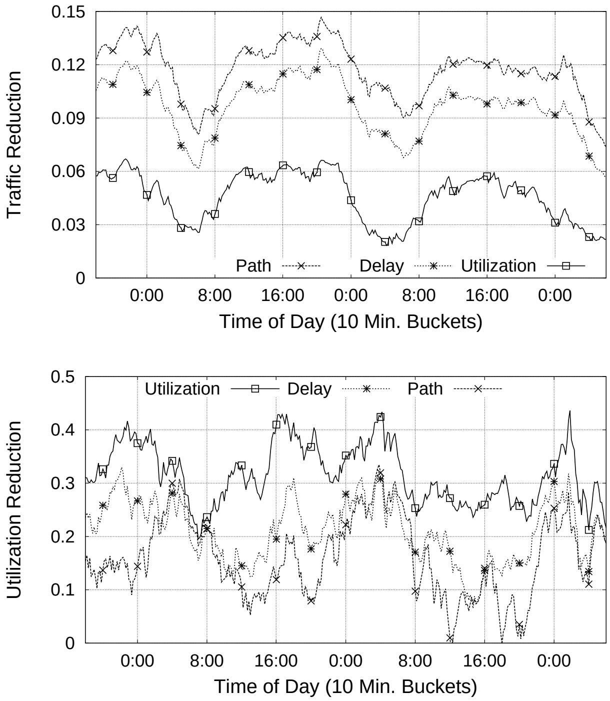


**Figure 13: Total traffic (top) and maximum link utilization (bottom) reduction with CaTE and different network metrics.**

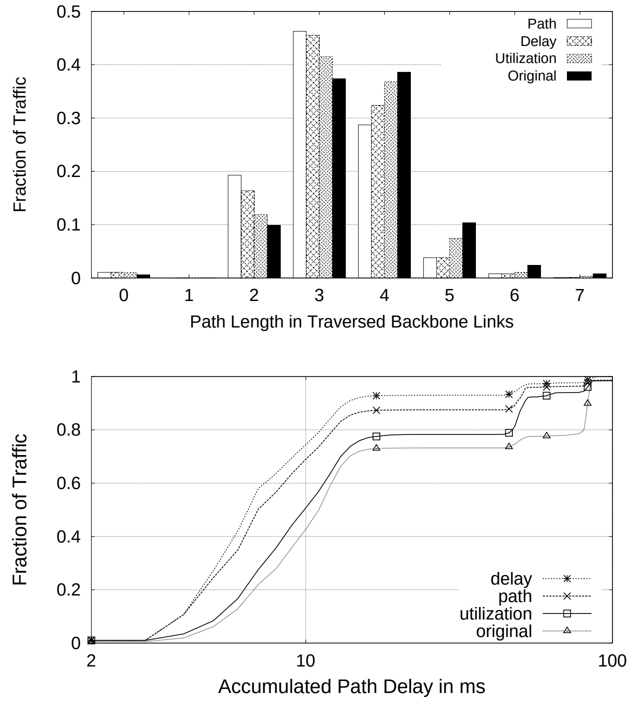


**Figure 14: Backbone path length (top) and accumulated path delay (bottom) with CaTE and different network metrics.**

topology for the US backbone of AT&T as measured by the Rocketfuel project [60, 57]. Given that no publicly available traffic demands exist for AT&T, we rely on the gravity model [54] to generate several traffic demand matrices as in ISP1.

**Abilene** is the academic network in the US. We use the Abilene topology and traffic demands covering a 6 month period that are both publicly available.[1]

The topology of both networks differ significantly from the one of ISP1. In AT&T, many smaller nodes within a geographical area are aggregated into a larger one. Abilene has few but large and well connected nodes with a high degree of peerings. For the application mix we rely on recent measurements in AT&T [27] and for server diversity we rely on measurements of users in these networks [2].

Figure 15 shows the cumulative fraction of normalized link utilizations for AT&T and Abilene with different optimization goals. As already done in ISP, only the Top 10 CPs are considered for CaTE, while all other traffic stays unaffected. For AT&T the benefit for the maximum link utilization is about $36\%$ when the network is optimized for minimizing the maximum link utilization, while the median reduction in terms of network-wide traffic is about $3.7\%$. When other optimizations are used, the benefits of CaTE regarding the link utilization minimization are approximately $12\%$ for path length and delay. However, when looking at the median traffic reduction of these metrics, the traffic is reduced by $5.4\%$ when path length is used, while delay achieves a reduction of $5\%$. In the Abilene network benefits of CaTE are more significant: $45\%$ reduction in the maximum link utilization and $18\%$ for network-wide traffic when CaTE optimizes for link utilization. When targeting the other two metrics, i. e., path length and delay, the results show that CaTE does not reduce the maximum link utilization. In fact, the maximum link utilizations stays constant. This is due to the structure of the network and the fact that the content is available closer, but at the cost of keeping the high utilization on some of the links. However, when looking at the median traffic reduction, both metrics manage to reduce the traffic by over $24\%$. These results show that CaTE is capable of targeting different optimization goals in different network structures and is able to optimize for different metrics.

It is worth noting that for AT&T $40\%$ of the links have a normalized link utilization less than $10\%$ while the remaining link utilizations are distributed almost linear. This distribution fits the structural observations made for the AT&T network: many links from smaller nodes are aggregated into larger ones. This also explains why the benefits for AT&T are smaller, since such a structure reduces the path diversity. Turning our attention to Abilene, we attribute the higher reduction of maximum link utilization and network-wide traffic to the non-hierarchical structure of the network and a higher ratio of peering locations. Applying CaTE to both AT&T and Abilene networks where the network metric is delay or path length shows similar behavior of CaTE as it does in ISP1.

## 6.5 CaTE and Popular Applications

Today, the launch of new content hosted on CPs such as high definition video or others that share flash-crowd characteristics, is not done in coordination with ISPs. This is challenging to ISPs that have to deal with rapid shifts of traffic volume as currently deployed traffic engineering tools are too slow to react to rapid demand changes. Furthermore, the end-user experience for popular applications is far from optimal as application designers have limited means to optimize the end-to-end delivery of content [39]. Both ISPs and applications would benefit from the improved traffic engineering capabilities of CaTE. We believe that CaTE can act as an enabler for ISP-application collaboration.

For example, Netflix, a very popular application that delivers high quality videos to end-users, relies on commercial CDNs such

[1] http://userweb.cs.utexas.edu/~yzhang/research/AbileneTM/

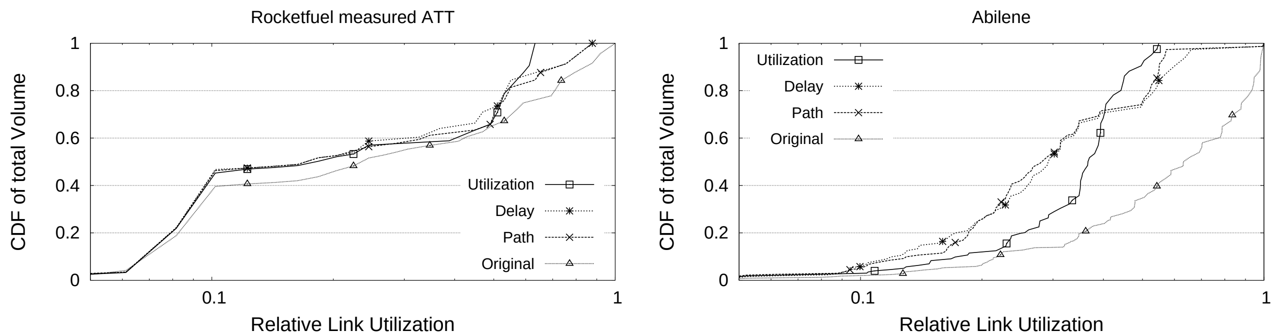


**Figure 15: Link utilization improvements after applying when using CaTE in AT&T and Abilene.**

as Level3 and Limelight to improve the content delivery. Today, Netflix is only available in North and Latin America. However, Netflix has announced that it will be launching its services in Europe early 2012. To quantify the effect of Netflix coming to Europe, we use our simulation to estimate the effect on ISP1. We run a series of experiments, assuming that the traffic of the CPs hosting Netflix will increase 20-fold. Our results show that with CaTE, the total HTTP traffic volume is reduced by up to 8% and the utilization of the most utilized link by 60%. More detailed results can be found in Appendix C.

## 7. RELATED WORK

To meet the requirements of mission critical applications with stringent Service Level Agreements (SLAs), today's ISPs rely on traffic engineering [5] to better control the flow of IP packets inside their network. Several techniques have been proposed in the literature, some require tuning of the IP routing protocols used inside the ISP network [24, 25, 64], while others rely on multipath [19, 16, 58, 21, 28, 65]. Changing routing weights can lead to oscillations [30] and is applied on time scales of hours. Multipath enables ISPs to dynamically distribute the traffic load within the network in the presence of volatile and hard to predict traffic demand changes [19, 16, 58, 21], even at very small time scales, but requires additional configuration and management or router support. CaTE is complementary to both routing-based traffic engineering and multipath enabled networks.

Traffic engineering relies on the availability of information about the traffic demands, which can be obtained either by direct observations [19, 31, 20, 63] or through inference [48, 66, 67, 59, 18]. CaTE relies on the network location diversity exposed by current hosting and content delivery infrastructures [2].

Game-theoretic results [36, 14, 45] show that the collaboration between CPs and ISPs can lead to a win-win situation. Recent studies also show that content location diversity has significant implications on traffic engineering within an ISP [56]. To our knowledge, CaTE is the first system that is proposed to leverage the benefits of a direct collaboration between CPs and ISPs.

## 8. SUMMARY

Today, a large fraction of Internet traffic is due to a few content providers that rely on highly distributed infrastructures [40, 42, 2]. These distributed infrastructures expose a significant location diversity, which opens new opportunities to improve end-user performance, help CPs to better locale end-user and circumvent network bottlenecks, and enables new traffic engineering capabilities. We introduce the concept of *content-aware traffic engineering* (CaTE), that leverages this location diversity to engineer the traffic through careful selection of the locations from which content is obtained. We propose deployment schemes of CaTE based on an online algorithm. The algorithm is stable and incurs no oscillations in link utilizations. Furthermore, CaTE works on time scales ranging between the TCP control loop and traditional traffic engineering, and therefore advantageously complements existing traffic engineering techniques.

We evaluate some of the potential benefits of CaTE on multiple operational networks using an offline derivative of the online algorithm. Our results show that CaTE provides benefits to CPs, ISPs and end-users, by reducing the maximum link utilization, the path length and the delay inside an ISP network, as well as enabling improved end-user to CP server assignment.

In the future, we envision CaTE as an enabler for coordinated and Internet-wide traffic engineering. Meanwhile, CaTE creates incentives for both ISPs and CPs to interlock their traffic engineering planes through the mutual benefits it brings. As further work, we want to deploy a prototype implementation of CaTE and evaluate it through a direct collaboration between a CP and an ISP.

## 9. REFERENCES


[1] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig. Comparing DNS Resolvers in the Wild. In *ACM IMC*, 2010.

[2] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig. Web Content Cartography. In *ACM IMC*, 2011.

[3] B. Ager, F. Schneider, J. Kim, and A. Feldmann. Revisiting Cacheability in Times of User Generated Content. In *IEEE Global Internet*, 2010.

[4] R. Alimi, R. Penno, and Y. Yang. ALTO Protocol. draft-ietf-alto-protocol-08, May 2011.

[5] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and principles of internet traffic engineering. RFC3272.

[6] B. Awerbuch and F. Leighton. Multicommodity Flows: A Survey of Recent Research. In *ISAAC*, 1993.

[7] Y. Azar, J. Naor, and R. Rom. The competitiveness of on-line assignments. *J. Algorithms*, 1995.

[8] B. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina. Controlling Data in the Cloud: Outsourcing Computation withour Outsourcing Control. 2009.

[9] Cisco. NetFlow services and applications. White paper: http://www.cisco.com/warp/public/732/netflow.

[10] J. Cleary, S. Donnelly, I. Graham, A. McGregor, and M. Pearson. Design Principles for Accurate Passive Measurement. In *PAM*, 2000.

[11] Jr. E. Coffman, M. Garey, and D. Johnson. Approximation Algorithms for Bin Packing: A Survey. In *Approximation algorithms for NP-hard problems*, 1997.

[12] C. Contavalli, W. van der Gaast, S. Leach, and D. Rodden. Client IP Information in DNS Requests. IETF draft, work in progress, draft-vandergaast-edns-suspend-ip-01.txt, 2011.

[13] A. Czumaj, C. Riley, and C. Scheideler. Perfectly Balanced Allocation. In *RANDOM-APPROX*, 2003.

[14] D. DiPalantino and R. Johari. Traffic Engineering versus Content Distribution: A Game-theoretic Perspective. In *IEEE INFOCOM*, 2009.

[15] F. Dobrian, A. Awan, I. Stoica, V. Sekar, A. Ganjam, D. Joseph, J. Zhan, and H. Zhang. Understanding the Impact of Video Quality on User Engagement. In *ACM SIGCOMM*, 2011.

[16] A. Elwalid, C. Jin, S. Low, and I. Widjaja. MATE : MPLS adaptive traffic engineering. In *IEEE INFOCOM*, 2001.

[17] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware forward caching. In *WWW Conf.*, 2009.

[18] V. Erramilli, M. Crovella, and N. Taft. An Independent-connection Model for Traffic Matrices. In *ACM IMC*, 2006.

[19] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving Traffic Demands for Operational IP Networks: Methodology and Experience. *IEEE/ACM Trans. Netw.*, 2001.

[20] A. Feldmann, N. Kammenhuber, O. Maennel, B. Maggs, R. De Prisco, and R. Sundaram. A Methodology for Estimating Interdomain Web Traffic Demand. In *ACM IMC*, 2004.

[21] S. Fischer, N. Kammenhuber, and A. Feldmann. REPLEX: Dynamic Traffic Engineering based on Wardrop Routing Policies. In *ACM CoNEXT*, 2006.

[22] S. Fischer, H. Räcke, and B. Vöcking. Fast Convergence to Wardrop Equilibria by Adaptive Sampling Methods. In *ACM STOC*, 2006.

[23] S. Fischer and B. Vöcking. Adaptive Routing with Stale Information. In *ACM PODC*, 2005.

[24] B. Fortz and M. Thorup. Internet Traffic Engineering by Optimizing OSPF Weights. In *IEEE INFOCOM*, 2000.

[25] B. Fortz and M. Thorup. Optimizing OSPF/IS-IS Weights in a Changing World. *IEEE J. Sel. Areas in Commun.*, 2002.

[26] P. Francois and O. Bonaventure. Avoiding transient loops during the convergence of link-state routing protocols. *IEEE/ACM Trans. Netw.*, 2007.

[27] A. Gerber and R. Doverspike. Traffic Types and Growth in Backbone Networks. In *OFC/NFOEC*, 2011.

[28] D. Goldenberg, L. Qiuy, H. Xie, Y. Yang, and Y. Zhang. Optimizing Cost and Performance for Multihoming. In *ACM SIGCOMM*, 2004.

[29] R. Graham. Bounds on Multiprocessing Timing Anomalies. *SIAM J. Applied Math.*, 1969.

[30] T. Griffin and G. Wilfong. On the Correctness of iBGP Configuration. In *ACM SIGCOMM*, 2002.

[31] A. Gunnar, M. Johansson, and T. Telkamp. Traffic Matrix Estimation on a Large IP Backbone: A Comparison on Real Data. In *ACM IMC*, 2004.

[32] Akamai Inc. SureRoute. www.akamai.com/dl/feature_sheets/fs_edgesuite_sureroute.pdf.

[33] Microsoft Inc. Windows azure. http://www.microsoft.com/windowsazure/cdn.

[34] Sandvine Inc. Global broadband phenomena. Research Report http://www.sandvine.com/news/global_broadband_trends.asp, 2011.

[35] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard. Networking Named Content. In *ACM CoNEXT*, 2009.

[36] W. Jiang, R. Zhang-Shen, J. Rexford, and M. Chiang. Cooperative Content Distribution and Traffic Engineering in an ISP Network. In *ACM SIGMETRICS*, 2009.

[37] P. Key, L. Massoulié, and D. Towsley. Path Selection and Multipath Congestion Control. *Commun. of the ACM*, 2011.

[38] L. Khachiyan. A Polynomial Time Algorithm for Linear Programming. *Dokl. Akad. Nauk SSSR*, 1979.

[39] R. Krishnan, H. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao. Moving Beyond End-to-end Path Information to Optimize CDN Performance. In *ACM IMC*, 2009.

[40] C. Labovitz, S. Lekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In *ACM SIGCOMM*, 2010.

[41] A. Lakhina, K. Papagiannaki, Mark. Crovella, C. Diot, E. Kolaczyk, and N. Taft. Structural Analysis of Network Traffic Flows. In *ACM SIGMETRICS*, 2004.

[42] T. Leighton. Improving Performance on the Internet. *Commun. of the ACM*, 2009.

[43] J. Lenstra, D. Shmoys, and É. Tardos. Approximation Algorithms for Scheduling Unrelated Parallel Machines. *Math. Program.*, 1990.

[44] K. Ma. Content Distribution Network Interconnection (CDNI) Metadata Interface. draft-ma-cdni-metadata-00, April 2011.

[45] R. T. B. Ma, D. .M. Chiu, J. C. S. Lui, V. Misra, and D. Rubenstein. On Cooperative Settlement Between Content, Transit, and Eyeball Internet Service Providers. *IEEE/ACM Trans. Netw.*, 2011.

[46] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *ACM IMC*, 2009.

[47] Z. Mao, C. Cranor, F. Douglis, M. Rabinovich, O. Spatscheck, and J. Wang. A Precise and Efficient Evaluation of the Proximity Between Web Clients and Their Local DNS Servers. In *Usenix ATC*, 2002.

[48] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic Matrix Estimation: Existing Techniques and New Directions. In *ACM SIGCOMM*, 2002.

[49] Limelight Networks. http://www.limelightnetworks.com/platform/cdn.

[50] E. Nygren, R. K. Sitaraman, and J. Sun. The Akamai Network: A Platform for High-performance Internet Applications. *SIGOPS Oper. Syst. Rev.*, 44:2–19, August 2010.

[51] V. Paxson. Bro: A System for Detecting Network Intruders in Real-Time. *Computer Networks*, 1999.

[52] I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann. Improving Content Delivery using Provider-Aided Distance Information. In *ACM IMC*, 2010.

[53] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the Electric Bill for Internet-scale Systems. In *ACM SIGCOMM*, 2009.

[54] M. Roughan. Simplifying the Synthesis of Internet Traffic Matrices. *ACM Comp. Comm. Rev.*, 2005.

[55] AMAZON Web Services. http://aws.amazon.com.

[56] A. Sharma, A. Mishra, V. Kumar, and A. Venkataramani. Beyond MLU: An application-centric comparison of traffic engineering schemes. In *IEEE INFOCOM*, 2011.

[57] R. Sherwood, A. Bender, and N. Spring. DisCarte: A Disjunctive Internet Cartographer. In *ACM SIGCOMM*, 2008.

[58] S.Kandula, D. Katabi, B. Davie, and A. Charny. Walking the Tightrope: Responsive Yet Stable Traffic Engineering. In *ACM SIGCOMM*, 2005.

[59] A. Soule, A. Nucci, R. Cruz, E. Leonardi, and N. Taft. How to Identify and Estimate the Largest Traffic Matrix Elements in a Dynamic Environment. In *ACM SIGMETRICS*, 2004.

[60] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson. Measuring ISP Topologies with Rocketfuel. *IEEE/ACM Trans. Netw.*, 2004.

[61] M. Tariq, A. Zeitoun, V. Valancius, N. Feamster, and M. Ammar. Answering What-if Deployment and Configuration Questions with Wise. In *ACM SIGCOMM*, 2009.

[62] S. Triukose, Z. Al-Qudah, and M. Rabinovich. Content Delivery Networks: Protection or Threat? In *ESORICS*, 2009.

[63] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon. Providing Public Intradomain Traffic Matrices to the Research Community. *ACM Comp. Comm. Rev.*, 2006.

[64] Y. Wang, Z. Wang, and L. Zhang. Internet Traffic Engineering Without Full Mesh Overlaying. In *IEEE INFOCOM*, 2001.

[65] M. Zhang, C. Yi, B. Liu, and B. Zhang. GreenTE: Power-aware Traffic Engineering. In *ICNP*, 2010.

[66] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast Accurate Computation of Large-scale IP Traffic Matrices from Link Loads. In *ACM SIGMETRICS*, 2003.

[67] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An Information-theoretic Approach to Traffic Matrix Estimation. In *ACM SIGCOMM*, 2003.

# APPENDIX

# A. ESTIMATING THE BENEFITS OF CaTE WITH PASSIVE MEASUREMENTS

We answer the question of the potential benefit CaTE can offer to CPs, ISPs, and end-users. The online algorithm requires deployment of CaTE inside an operational network. An alternative is to rely on a simulation-driven evaluation of CaTE. For this, we design offline algorithms that take as input passive measurements and estimate the gain when applying CaTE under different scenarios. We first propose a linear programming formulation and then we present greedy algorithms to speed-up the process of estimating the benefits of CaTE.

## A.1 Linear Programming Formulation

To estimate the potential improvement of CaTE we formulate the Restricted Flow Load Balancing problem (see Section 5.1) as a Linear Program (LP) with restrictions on the variable values. Variables $f_{ijk}$ correspond to flows that can be influenced. Setting $f_{ijk} = 0$ indicates that consumer $j$ cannot download the content from server $i$ of a content provider $k$. For each consumer $j$ we require that its demand $d_{jk}$ for content provider $k$ is satisfied, i. e., we require $\sum_{i \in M_{jk}} f_{ijk} = d_{jk}$. The utilization on a flow $f_{ij}$ is expressed as $f_{ij} = \sum_k f_{ijk}$.

We use the objective function to encode the traffic engineering goal. For ease of presentation we use as objective function the minimization of the maximum link utilization. Let $T_e$ be the set of flows $f_{ij}$ that traverse a link $e \in E$. The link utilization of a link $e \in E$ is expressed as $L_e = \sum_{T_e} f_{ij}$. Let variable $L$ correspond to the maximum link utilization. We use the inequality $\sum_{T_e} f_{ij} \leq L$ for all links. This results in the following LP problem:

$$
\begin{aligned}
&\texttt{min}\, L \\
&\sum_i f_{ijk} = d_{jk}, && \forall\, j \in J,\, k \in K \\
&\sum_{T_e} f_{ijk} \leq L, && \forall\, j \in J,\, i \in I,\, k \in K,\, e \in E \\
&0 \leq f_{ijk} \leq d_{jk}, && \forall\, j \in J,\, i \in M_{jk},\, k \in K \\
&f_{ijk} = 0, && \forall\, j \in J,\, i \notin M_{jk},\, k \in K
\end{aligned}
$$

The solution of the above LP provides a fractional assignment of flows under the assumption that flows are splittable and thus can be solved in polynomial time [38]. The solution is the optimal flow assignment, $f^*_{ijk}$, that corresponds to the optimal traffic matrix $\mathbf{x}^*$. If flows are not splittable, or the sub-flows are discretized, then the integer programming formulation has to be solved. In this case the Restricted Flow Load Balancing problem is NP-hard and a polynomial time rounding algorithm that approximates the assignment within a factor of 2 exists [43].

## A.2 Approximation Algorithms

Since it is a common practice for operators to study multiple scenarios to quantify the effect of changes in traffic matrices over periods that spans multiple weeks or months, solutions based on LP may be too slow. It might be also too slow to estimate the gain of CaTE when applying it to an arbitrary combination of CPs. To that end, we turn our attention to the design of fast approximation algorithms. Simple greedy algorithms for load balancing problems [29] are among the best known. Accordingly, we propose a greedy algorithm for our problem which starts with the largest flow first.

**Algorithm 2: Iterative Greedy-Sort-Flow.**

**INPUT:** $I$, $J$, $K$, $\{f_{ijk}\}$, $\{M_{jk}\}$, $A$.
**OUTPUT:** $\{f^*_{ijk}\}$.

**Initialization:**
1. Sort $k \in K$ by decreasing volume: $\sum_i \sum_j f_{ijk}$.
2. Sort $j \in J$ by decreasing volume: $\sum_i f_{ijk}$ for all $k \in K$.

**Iteration:**
Until no sub-flow is re-assigned or the maximum number of iterations has been reached.
▷ Pick unprocessed $k \in K$ in descending order.
▷ Pick unprocessed $j \in J$ in descending order.
▷ Re-assign $f_{ijk}$ in $f_{ij}^{-k}$, $i \in M_{jk}$ s.t. the engineering goal is achieved.

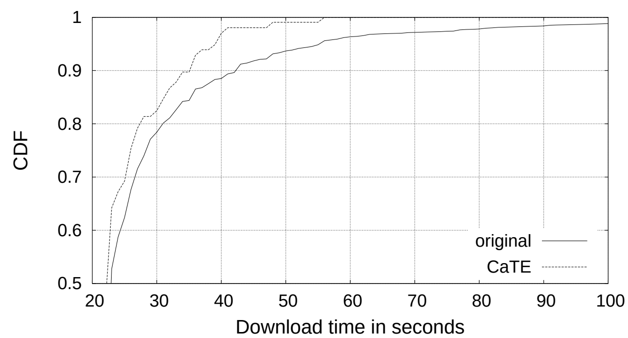


**Figure 16: Distribution of download times of a CP.**

**Algorithm 3: Greedy-Sort-Flow.** Sort sub-flows in decreasing order based on volume and re-assign them in this order to any other eligible flow which, after assigning the sub-flow $f_{ijk}$, will yield the desire traffic engineering goal.

Assignment in sorted order has been shown to significantly improve the approximation ratio and the convergence speed [13, 29]. Recent studies [27, 40, 52] show that a small number of content providers are responsible for a large fraction of the traffic. Therefore it is expected that the algorithm yields results close to the optimal ones. To further improve the accuracy of the proposed approximation algorithm, we design an *iterative* version of the algorithm, presented in Algorithm 2, that converges to the optimal solution. Indeed, a small number of iterations, typically one, suffice to provide a stable assignment of flows.

As we elaborate in Section 6, we performed a number of simulations using real operational traces, and different sets of CPs. Our evaluation show that the performance of the iterative greedy algorithm presented in Algorithm 2 yields results very close to this obtained with LP, but in significantly shorter time.

# B. ACTIVE MEASUREMENTS IN ISP1

The CaTE evaluation in Section 6.2 does not allow us to argue about end-user performance, as it is based on simulations. To this end, we complement our previous network-wide simulations with active measurements. Over a period of one week, we repeatedly downloaded a 60MB object from one of the major CPs. This CP is an OCH distributed across 12 locations. The downloads were performed every two hours, from each of the 12 locations. Additionally, mapping requests were issued every 200ms to find out the dynamics in the server assignment of this CP. Figure 16 shows the

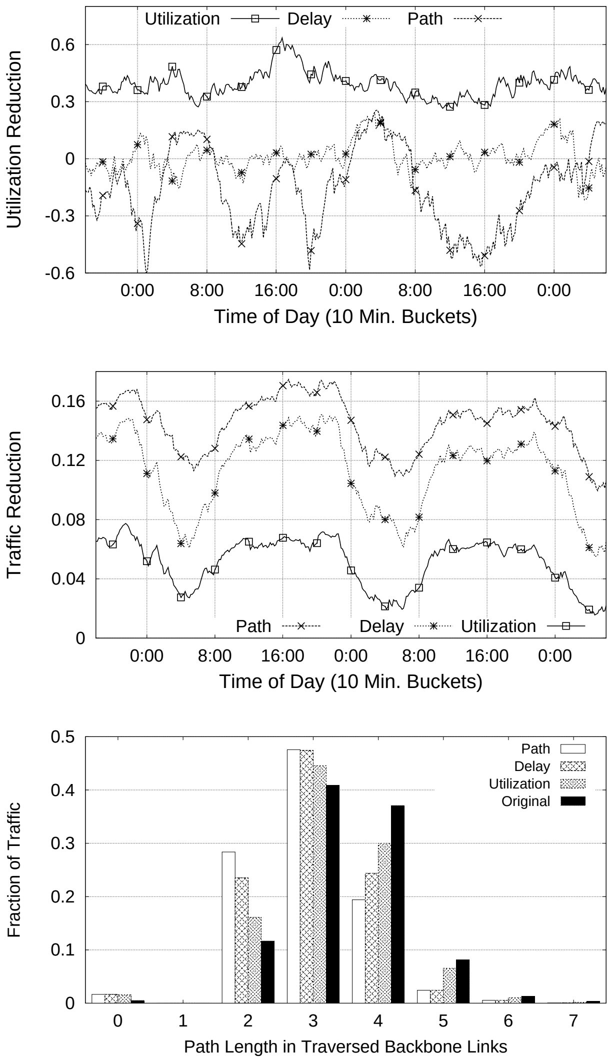


**Figure 17: Projection of reduction in link utilization (top), reduction in overall network traffic (middle) and fraction of volume by path length (bottom) if Netflix is launched in ISP1.**

distribution of total download times when the CP assigns end-users to its servers ("original") and compares it to the download time that would be observed if CaTE had been used. We observe that more than 50% of the downloads do not show a significant difference. This happens when congestion is low, e. g., during non-peak hours. For 20% of the downloads, we observe a significant difference in the download times, mainly during peak hours. This confirms our observation that CaTE is most beneficial during peak hours.

## C. CASE STUDY: NETFLIX IN ISP1

Netflix, a very popular application that delivers high quality videos to end-users, relies on commercial CDNs such as Level3 and Limelight to improve the content delivery. Today, Netflix is available in North and Latin America, and is announced to arrive in the UK soon. Recent studies show that Netflix is responsible for more than 30% of the peak downstream traffic in large ISPs [34]. Consider the scenario where Netflix is launching its service in the large European ISP1 we described in Section 6.1. If the launch happens overnight, ISP1 would have to deal with a huge amount of highly variable traffic, which would have significant implications on the operation of ISP1. With CaTE, the traffic of Netflix can be spread across the ingress points of ISP1. This will limit the negative consequences imposed by additional traffic for the CP delivering Netflix as well as for ISP1 and thus avoids a deteriorated end-user experience.

To quantify the effect of Netflix being deployed in Europe, we simulate the launch of Netflix in ISP1, assuming that the CP currently hosting Netflix increases its traffic 20-fold, while keeping the distribution of the requests. Next, we generate a new set of traffic demands for CaTE accordingly. We consider the the top 10 CPs by volume for CaTE, and show the benefits when optimizing for different metrics.

Our results show that with CaTE, the utilization of the most utilized link can be reduced by up to 60% (see top of Figure 17), the total HTTP traffic volume can be reduced by 15% (see middle of Figure 17) and traffic can be shifted towards shorter paths inside the network of ISP1 (bottom of Figure 17). However, when considering all metrics, we observe that not all metrics can be optimized to their full extend at the same time. For example, a reduction of traffic in the order of 15% would actually increase the utilization on the highest loaded link by 60%. This indicates that the optimization function employed by CaTE needs to be carefully chosen to target the most important metrics when deploying CaTE inside a network. Nonetheless, if minimizing the maximum link utilization is chosen as the optimization function for CaTE, benefits in all metrics can be observed.

Internet applications such as Netflix are in a position to negotiate how they should be deployed in order to improve end-user experience and not disturb the operation of ISPs. CaTE can be used to identify the best peering points between the CPs that deliver Netflix traffic and the ISPs that receive its traffic. In addition, ISPs might offer better peering prices if the CPs hosting Netflix are willing to provide a higher diversity in the locations from which the traffic can be obtained. This would lead to a win-win situation where Netflix can offer better service to its users, the CPs achieve reduced pricing on their peering agreements, and ISPs can compensate the reduced peering revenue through more efficient operations.