1ST QUARTER 2008, VOLUME 10, NO. 1

IEEE COMMUNICATIONS SURVEYS
The Electronic Magazine of Original Peer-Reviewed Survey Articles

www.comsoc.org/pubs/surveys

AN OVERVIEW OF ROUTING OPTIMIZATION FOR INTERNET TRAFFIC ENGINEERING

NING WANG, KIN HON HO, GEORGE PAVLOU, AND MICHAEL HOWARTH, UNIVERSITY OF SURREY

ABSTRACT

Traffic engineering is an important mechanism for Internet network providers seeking to optimize network performance and traffic delivery. Routing optimization plays a key role in traffic engineering, finding efficient routes so as to achieve the desired network performance. In this survey we review Internet traffic engineering from the perspective of routing optimization. A taxonomy of routing algorithms in the literature is provided, dating from the advent of the TE concept in the late 1990s. We classify the algorithms into multiple dimensions: unicast/multicast, intra-/inter-domain, IP-/MPLS-based and offline/online TE schemes. In addition, we investigate some important traffic engineering issues, including robustness, TE interactions, and interoperability with overlay selfish routing. In addition to a review of existing solutions, we also point out some challenges in TE operation and important issues that are worthy of investigation in future research activities.

The Internet is currently experiencing a transition from point-to-point best effort (BE) communications toward a multiservice network that supports many types of multimedia applications, with potentially high bandwidth demand. Thanks to the rapid development of communication network hardware, adding physical resources (fast-speed switching and routing elements, high-capacity network links, etc.) to the existing Internet has become relatively cheap in recent years. Typically, the advent of increasingly high-speed links has offered opportunities for IP network providers (INPs) to adopt a strategy of bandwidth overprovisioning in their networks. Nevertheless, this approach is currently only applicable to the core network, and the demand from sharply growing customer traffic over the global Internet still cannot be satisfied. The measurement results presented in [1] indicate that bottlenecks of the Internet backbone are not only located at interdomain links between autonomous systems (ASs), but also within individual domains. Given this information, it is essential for INPs to perform efficient resource optimization both intra- and interdomain so as to eliminate these bottlenecks. Internet traffic engineering (TE) is the process of performing this task. In [2] TE is defined as large-scale network engineering for dealing with IP network performance evaluation and optimization. A more straightforward explanation of TE is also given in [3]: "to put the traffic where the network bandwidth is available." Therefore, the nature of TE is effectively a routing optimization for enhancing network service capability without causing network congestion. In doing so, typical TE objectives include balancing the load distribution and minimizing bandwidth consumption in the network. Figure 1 illustrates this with a simple TE example. We assume that the bandwidth capacity of each link is 10 Mb/s, and there are three individual customer flows injected at node A, heading toward node C. If conventional shortest path routing is applied, all the customer flows are routed on the direct link A–C, thus causing the link utilization to be as high as 180 percent (6 × 3/10). On the other hand, if the three flows are routed through different paths, as shown in Fig. 1b, the total traffic within the network is evenly distributed without causing link congestion. As this example illustrates, routing optimization that uses alternative multiple paths other than conventional shortest-path-based approaches can be an effective means to improving the network service capability.

Two major issues that have recently received attention in TE approaches are quality of service (QoS) and resilience. First, many of the new multimedia applications not only have bandwidth requirements, but also require other QoS guarantees, such as end-to-end delay, jitter, or packet loss probability. These QoS requirements impose new challenges on INPs' TE in that the end-to-end QoS demands need to be satisfied through TE mechanisms. Second, given the fact that network node and link failure are still frequent events on the Internet [4], TE solutions have to consider how to minimize the impact of failures on network performance and resource utilization.

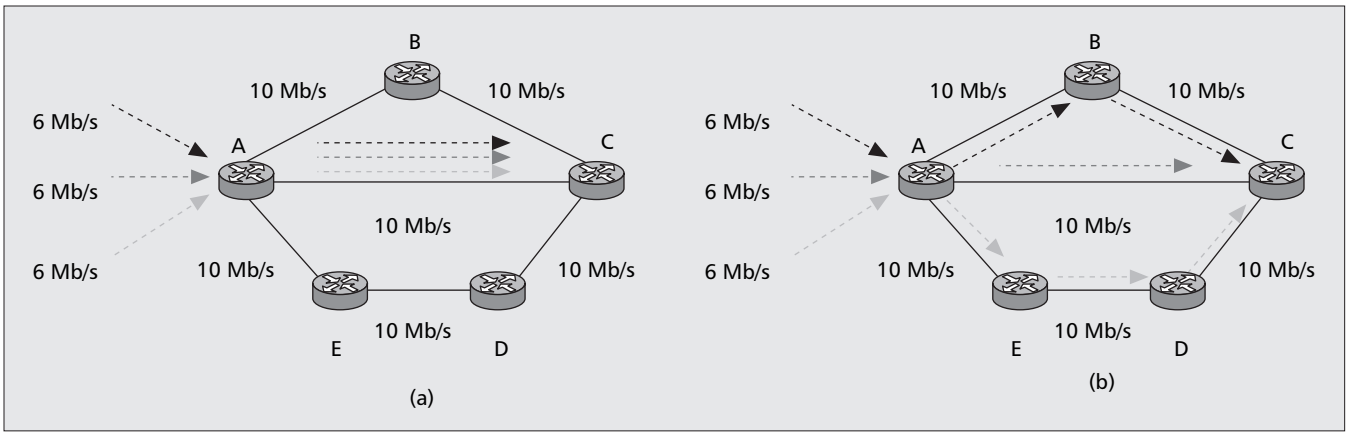


Figure 1. *A simple TE example: a) three traffic flows are routed over a common path, which causes overloading; b) traffic engineering directs the traffic flows onto different paths, thus achieving network load balancing.*

We discuss detailed robustness-aware TE solutions later.

Many papers have been published in the area of routing optimization. As a result, it is by no means an easy task to classify various TE solutions, and present a comprehensive and clear survey. In this article we classify these TE routing approaches according to four orthogonal criteria:

- From the aspect of traffic optimization scope, TE can be classified into intradomain TE and interdomain TE.
- From the aspect of routing enforcement mechanisms, TE can be classified into multiprotocol label switchinig (MPLS)-based TE and IP-based TE.
- From the aspect of availability of traffic demand or timescale of operations, TE can be classified into offline TE and online TE.
- From the aspect of traffic type, TE can be classified into unicast TE and multicast TE.

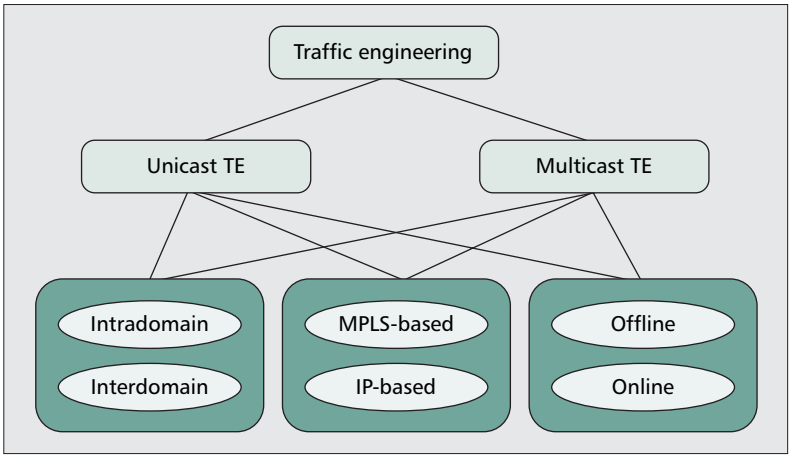


Figure 2. *Hierarchical classification of Internet traffic engineering.*

An overall taxonomy of Internet TE is presented in Fig. 2, and this article is organized following the structure of this diagram. The objective of this article is thus to provide a comprehensive survey on routing optimization for all the components in the TE hierarchy. The rest of the article is organized as follows. We specify the detailed characteristics of different types of TE according to Fig. 2. We introduce intradomain TE, which includes both MPLS- and IP-based routing optimization algorithms. Then we move on to interdomain TE, which we further divide into inbound and outbound TE. Multicast TE is presented, we discuss some important interactions between current TE approaches, and we conclude with a summary. It is worth mentioning that this survey does not claim to be exhaustive, although we attempt not to miss important work in the area.

Traffic Engineering Classifications

Intradomain TE vs. Interdomain TE

The task of intradomain TE is to optimize customer traffic routing between AS border routers (ASBRs) within a single domain. In comparison, interdomain TE deals with the problem of optimizing interdomain traffic traveling across multiple ASs. Interdomain TE mainly focuses on how to select ASBRs optimally as the ingress/egress points for interdomain traffic that travels across the local AS. That is to say, if the traffic has multiple potential ASBRs from which it can enter or leave the local domain, the problem of interdomain TE for an INP is: "which ASBR(s) should be used as the ingress/egress point(s) for routing the traffic through the local network so that the network resource utilization is optimized?" According to the control over how traffic enters/leaves the domain, interdomain TE can be further classified into inbound TE and outbound TE. Figure 3 presents a simple example to illustrate the difference between intra- and interdomain TE semantics, specifically using outbound TE as an example for interdomain TE. We assume that traffic destined to the remote prefix 20.20.20.0/24 (AS200) is injected into the local AS (AS100, 10.10.10.0/24) via ASBR 10.10.10.3, and both the internal peers 10.10.10.1 and 10.10.10.2 can provide a route to AS200 (i.e., both routers receive reachability information toward 20.20.20.0/24 through external Border Gateway Protocol [BGP] advertisements). In this scenario the decision to use ASBR 10.10.10.1 or 10.10.10.2 (or both for load balancing with interdomain multiple paths) as the egress point is the task of interdomain/outbound TE. Once the egress point has been selected, say ASBR 10.10.10.1, intradomain TE is then responsible for selecting the best intradomain path between each pair of ASBRs in the network. In this simple example, intradomain TE attempts to find an optimal internal path (or multiple paths if allowed) from ASBR 10.10.10.3 to ASBR 10.10.10.1 (e.g., path A or B or both) as well as an optimal path C from 10.10.10.3 to ASBR 10.10.10.2.

Despite their clear difference in definition, intra- and interdomain TE should not be considered independent of each other in practice, since the network configuration of one could potentially impact the other. Research has emerged

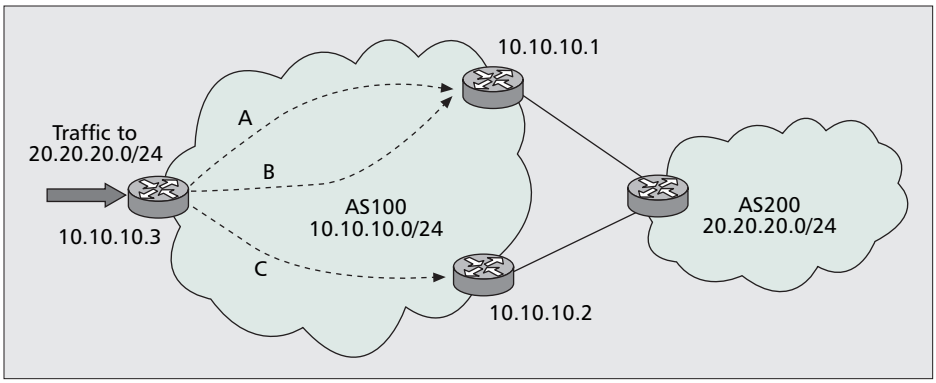


Figure 3. *Scope of traffic engineering. IntradomainTE considers optimized routing for each node pair within the network; for example, path A and/or B between 10.10.10.3 and 10.10.10.1. On the other hand, interdomain TE focuses on optimized ASBR selection; for example, the selection of egress point between 10.10.10.1 and 10.10.10.2 for the traffic destined to AS200.*

recently on the interaction between the two types of TE, and some results are presented in [5]. We provide more details on the interaction between intra- and interdomain TE later.

MPLS-Based TE vs. IP-Based TE

The concept of traffic engineering was first introduced in MPLS-based environments [6, 7]. By intelligently setting up dedicated label switched paths (LSPs) for delivering encapsulated IP packets, MPLS-based TE can provide an efficient paradigm for traffic optimization. The most distinct advantage of MPLS-based TE is its capability of explicit routing and arbitrary splitting of traffic, which is highly flexible for both routing and forwarding optimization purposes. However, since traffic trunks are delivered through dedicated LSPs, scalability and robustness become issues in MPLS-based TE. First, the total number of LSPs (assuming full mesh or equivalent) within a domain is $O(N^2)$ where N is the number of ASBRs. This means that the overhead of setting up LSPs can be very high in large-size networks. In addition, path protection mechanisms (e.g., using backup paths) are necessary in MPLS-based TE, as otherwise traffic cannot be automatically delivered through alternative paths in case of any link failure in active LSPs.

The first IP-based TE solution was proposed by Fortz *et al.* [8–10]. The basic idea of their approach is to set the link weights of interior gateway protocols (IGPs) according to the given network topology and traffic demand so as to control intradomain traffic and meet TE objectives. Unlike MPLS-based TE, which enables dedicated explicit routing for individual flows, such "fine-grained" path selection cannot be achieved in IP-based TE, as the changes of IGP link weight may affect the routing patterns of the entire set of traffic flows. More recently, schemes that manipulate BGP routing attributes, known as BGP tweaking [11], have also been proposed for interdomain TE. In this scenario optimized BGP routing is achieved through tuning of routing attributes on a per destination prefix basis. In comparison to the MPLS-based approach, these IP-based TE solutions lack flexibility in path selection, since explicit routing and uneven traffic splitting are not supported. However, the IP-based approach has better scalability and availability resilience than MPLS-based TE, because no overhead for dedicated LSPs is required, and also because traffic can be automatically delivered via alternative shortest paths in case of link failure without explicitly provisioning backup paths. However, given this type of auto-rerouting in the IP-based environment, link failures may introduce dramatic changes to traffic distribution (thus introducing new traffic congestion) even across multiple domains. For example, [12] indicates that link failures in IGP routing can increase the utilization of other links, as they have to carry the shifted traffic that originally traversed the broken shortest IGP path. In addition, in [13] the authors pointed out that in IGP/BGP routing an intradomain link failure may cause transit traffic to shift to alternative egress points due to a hot potato routing effect. This low TE robustness is in comparison to MPLS-based TE schemes, where a single link failure does not impact other primary LSPs unless they are using the faulty link. Table 1 summarizes the key differences between MPLS-based and IP-based TE.

Offline TE vs. Online TE

The third part of our taxonomy is to classify TE as offline and online. The principal difference between offline and online traffic engineering is the availability of a traffic matrix (TM) and timescale of traffic manipulation. The concept of a TM was originally associated with intradomain TE, where ingress/egress points of traffic are fixed. In this case the overall traffic demand on the network can be represented by a matrix TM, say, with each element $t(i, j)$ of the TM being the total bandwidth demand of all individual traffic flows (known as traffic trunk) from ingress node i to egress node j. Unlike intradomain TM, interdomain TM does not specify both ingress and egress points, as traffic travel across domains may enter/leave an AS through multiple border routers, which provides the opportunity for interdomain TE to select optimized ingress/egress points.

In some scenarios it is possible for an INP to forecast the traffic matrix before routing optimization is performed. Currently, there are two principal inputs from which traffic matrix can be forecasted: a service level specification (SLS) and monitoring/measurement (e.g., [14, 15]). An SLS is the detailed information on the agreement negotiated between customers and the INP. By aggregating the traffic predicted in SLSs with individual customers, the INP can estimate the overall bandwidth demand between each pair of ASBRs. In addition, the INP can also apply monitoring/measurement mechanisms at the network boundary for aiding traffic matrix estimation. Having obtained the traffic matrix for the specific network topology, an INP can perform offline TE (i.e., map optimally the whole traffic matrix onto the physical network). Figure 4 presents a basic diagram for the offline TE process. One important issue in offline TE is the average duration between two consecutive TE cycles, and this period is known as the resource provisioning cycle (RPC) [16]. In common practice, the RPC for offline TE is weekly or monthly, depending on various factors such as the frequency of establishing, modifying, and terminating SLSs with customers. The major weakness of offline TE is the lack of adaptive traffic manipulation according to traffic and network dynamics, such as traffic burst and network failures. These uncertainties may make offline TE less efficient as the actual traffic pattern might be different from what has been forecasted.

In some cases an INP might not be able to predict the overall TM in advance, and this requires that the INP perform online TE that does not require any knowledge about future traffic demands. In order to rapidly respond to dynamic traffic

	MPLS-based TE	IP-based TE
Routing mechanism	Explicit routing with packet encapsulation	Plain IGP/BGP-based routing
Routing optimization	Constraint-based routing (CBR)	IGP link weight adjustment BGP route attribute adjustment
Multipath forwarding	Arbitrary traffic splitting	Even traffic splitting only
Hardware requirement	MPLS capable routers required	Conventional IP routers
Route selection flexibility	More flexible — arbitrary path	Less flexible — shortest path only
Scalability (overhead in maintaining network state)	Less scalable	More scalable, with scalability of underlying routing protocol
Failure impact on traffic delivery (availability)	High (normally need backup paths in case of failures)	Low
Failure impact on TE performance	Low	High

■ Table 1. *MPLS/IP TE comparison.*

fluctuations, online TE is typically performed on a timescale of hours or even minutes. A practical concern for INPs to deploy online TE is how to make sure such a dynamic control system is self-converged without human intervention. Generally, the basic task of resource optimization is to optimally assign the new incoming traffic one by one so that the possibility of accommodating further incoming traffic without congestion can be maximized. Toward this end, online TE approaches should make sure that the traffic load is as evenly distributed as possible within the network, so random incoming traffic demand in the future can easily be satisfied. In some cases it is also possible to reroute existing flows in the network so as to reserve bandwidth for new and future incoming traffic. However, this rerouting should not involve a significant proportion of traffic flows in the network, as competing flows might interfere with each other and cause traffic instability and service disruption. In addition, due to the uncertainty of the traffic pattern (i.e., the lack of a global view on overall traffic conditions), online TE may have difficulties in handling future incoming traffic based on the current network state. To overcome these issues, a promising approach is to consider both offline and online TE together as complementary with each other. Specifically, offline TE provides guidelines to the behaviors of its online counterpart, which works as a more adaptive and local adjustment paradigm that tackles events that are not forecast by offline TE. This feature is addressed in more detail later.

Unicast TE vs. Multicast TE

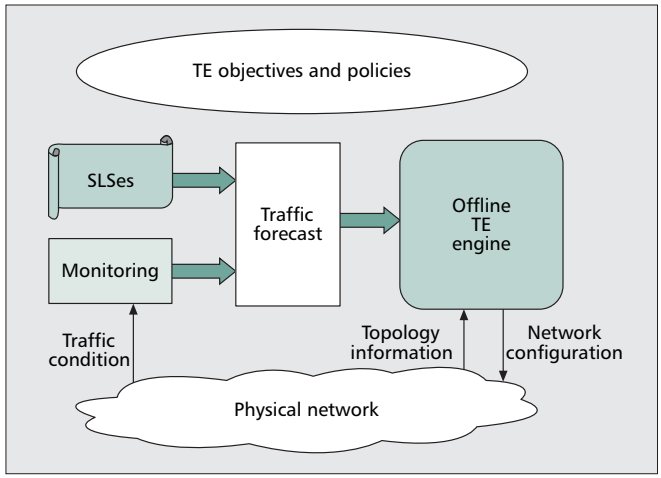
The Internet carries heterogeneous traffic, including both unicast/multicast traffic and various types of flows that use overlay routing techniques. In this article we survey not only unicast TE but also multicast TE, which is becoming important given recent progress in Internet multicast service development [17]. Compared to unicast TE, multicast TE is more complicated, since multicast routing is associated with point-to-multipoint tree construction. In the literature resource optimization in multicast TE is normally formulated as a Steiner tree related problem with the objective of minimizing bandwidth consumption. Although their TE problem formulations might be different, it should be noted that since IP unicast and multicast traffic can be simultaneously injected into the same physical network, TE for both types of traffic should not be done independently without an awareness of each other.

Intradomain Traffic Engineering

In this section we focus on routing optimization algorithms for intradomain TE. We first split intradomain TE into MPLS-based and IP-based subsections, and within each of them we discuss both offline and online TE.

Intradomain MPLS-based TE

MPLS is an Internet Engineering Task Force (IETF) standardized forwarding scheme. In MPLS traffic is sent along explicit LSPs. An LSP is the path between an ingress label switching router (LSR) and an egress LSR. At the boundary of an MPLS domain, LSRs classify IP packets into forwarding equivalence classes (FECs) and append different labels for packet forwarding within the MPLS domain. The Label Distribution Protocol (LDP) [18] is used to distribute label bindings during the setup of an LSP. MPLS is a powerful technology for Internet TE, as it allows traffic to be forwarded onto an arbitrary explicit route, which may not necessarily follow the shortest path computed by conventional IP routers. Typically, individual flows are aggregated by MPLS-based TE into traffic trunks identified by FECs, which are then carried on LSPs between ingress and egress routers. In this case the conventional shortest-path-based routing infrastructure (e.g.,



■ **Figure 4.** *Offline traffic engineering.*

Open Shortest Path First, OSPF) is overridden with MPLS explicit routing tunnels.

Offline Traffic Engineering — A generalized MPLS routing optimization can be formulated as a multicommodity flow problem [19], and can thus be solved using linear programming to yield an optimal solution for routing mechanisms that allow arbitrary traffic splitting. However, this approach is often regarded as impractical, especially in a large-sized network, since the number of LSPs required is potentially huge due to arbitrary traffic splitting. To obtain a more scalable TE solution, traffic splitting has to be limited in scope. An early MPLS-based TE approach used simple constraint-based routing (CBR) [20] without coordination between individual traffic trunks [21]. A typical CBR algorithm is as follows. Before setting up an LSP for a specific traffic trunk, all the infeasible network links (e.g., those with insufficient available bandwidth) are removed from the network topology. Shortest path routing (SPR) is then performed on the residual network graph, and the LSP is assigned to this shortest path. The algorithm repeats the above procedure until all the traffic trunks are assigned. This routing algorithm is known as Constrained Shortest Path First (CSPF). Other routing schemes have also been proposed to extend SPR, such as Widest Shortest Path (WSP) and Shortest Widest Path (SWP) [22, 23], both of which try to increase the available bandwidth at bottlenecks along the path. By applying WSP/SWP, not only has the underlying traffic a higher probability of finding a feasible path, but also network bottlenecks are avoided by "reserving" bandwidth resources for future demands, benefiting other traffic from this more sophisticated routing strategy.

In the literature many MPLS-based TE schemes have addressed the problem of minimizing the maximum utilization; this approach is often formulated as a linear or integer programming problem. In [24] TE is investigated using both single and multiple paths. The authors prove that TE using multiple paths (LSP bifurcation) and arbitrary traffic splitting is able to achieve optimal solutions using linear programming, while integer programming can be applied to MPLS-based TE without LSP bifurcations.

With the development of differentiated services (DiffServ), DiffServ-based MPLS-based TE has become a research area for supporting QoS differentiation. DiffServ-MPLS-based TE is now supported by both Cisco and Juniper routers, with CSPF being the fundamental routing algorithm. In addition, more sophisticated DiffServ aware/equivalent MPLS-based TE schemes have also been proposed in the literature [25–27]. The authors of [26] proposed a general framework for intradomain QoS provisioning through MPLS-based TE in DiffServ networks. From a routing optimization perspective, the TE objectives are to satisfy the QoS requirements of the traffic trunks and minimize the overall network cost (load). The cost function is formulated as a convex function of the traffic load on a per-QoS class basis, and the TE optimization task is formulated as a nonlinear programming problem. In order to find the optimal solution, the authors apply a general gradient projection method for calculating LSPs. The QoS metrics considered in this work include end-to-end delay and loss, both of which are transformed into unified hop-count-based constraints. In order to verify whether these QoS requirements are met during the optimization process, shortest path adaptations (e.g., kth shortest paths) are applied on a hop count basis. In [27], a differentiated TE (DTE) solution was proposed. To solve the path selection problem in DTE, the overall routing optimization is decomposed into two subproblems: the non-convex part of the optimization problem is solved by a simulated annealing technique, while the convex part is solved using the gradient projection method.

Apart from the pipe model, where LSPs are point-to-point (P2P), other papers have also proposed alternative models, such as the funnel model (multipoint-to-point, MP2P) [28–30] and the hose model (point-to-multipoint, P2MP) [31]. The advantage of these alternative models in LSP construction is to alleviate the scalability issues in LSP construction and maintenance. In order to reduce the total number of LSPs needed, the authors in [28] proposed a TE scheme using multiple MP2P LSPs. Specifically, the proposed approach consists of two distinct procedures: MP2P LSP construction and flow assignment. During the phase of LSP construction, a set of point-to-point paths is first selected between each ingress/egress pair with two constraints: the total hop counts of each path should not exceed the threshold that is the hops of the minimum hop count path plus a predefined number, and at least one path must be node-disjoint with the rest of the path set. If such a path set cannot be found, a path pair is selected comprising the minimum hop path and another disjoint path with a second minimum hop count. Thereafter, the MP2P LSP design applies binary integer programming on a per egress router basis, and merges the preselected point-to-point paths. In the flow assignment phase the task is to map the traffic trunks onto the constructed MP2P LSPs with the objective of minimizing the maximum load. In this work the design of MP2P LSPs has three distinct advantages: LSP scalability, load balancing, and resilience. In [29] MP2P LSPs are used for TE with deterministic end-to-end QoS guarantees. In addition, two admission control algorithms are introduced at the packet level, but routing optimization is not much addressed in this work. MP2P TE was also studied in [30], where the scalability issue in MPLS label space is investigated. The basic idea is similar to [28], which attempts to merge point-to-point paths into MP2P LSPs. However, this work assumes that the P2P paths are predefined, so the task is only to assign each of them to individual MP2P LSPs. From this point of view, routing and resource optimization are not the major concern in this work.

A summary of published offline MPLS-based TE work is presented in Table 2.

Online Traffic Engineering — Online MPLS-based TE can be classified into two categories: dynamically adjusting the traffic splitting ratio among preconstructed static LSPs [32, 33], and computing dynamic LSPs on the fly for each new traffic trunk demand. MATE [32] is a typical example of the first category, and its basic operation is to adaptively forward incoming traffic onto multiple preconstructed LSPs according to probing results from the network core. For this TE paradigm, routing optimization is not directly involved, as traffic and resource optimization are achieved through online forwarding adaptation. In the rest of this section we restrict our focus to the second category of online MPLS-based TE.

The CSPF, WSP, and SWP algorithms described earlier are the fundamental routing solutions that can be applied to online MPLS-based TE schemes. In DORA [34] the online TE solution contains two stages that maximize the ability of the network to accommodate future bandwidth-specified traffic demands. First, a parameter called path potential value (PPV) is computed for each link on a per ingress/egress node pair basis. The metric of PPV indicates the frequency with which each link has been used in the disjoint paths between ingress/egress node pairs. In the second stage network links without sufficient residual bandwidth are removed from the network graph, and then a combined weight is calculated for each remaining link based on the PPV value and the available bandwidth, with a tuning parameter known as bandwidth pro-

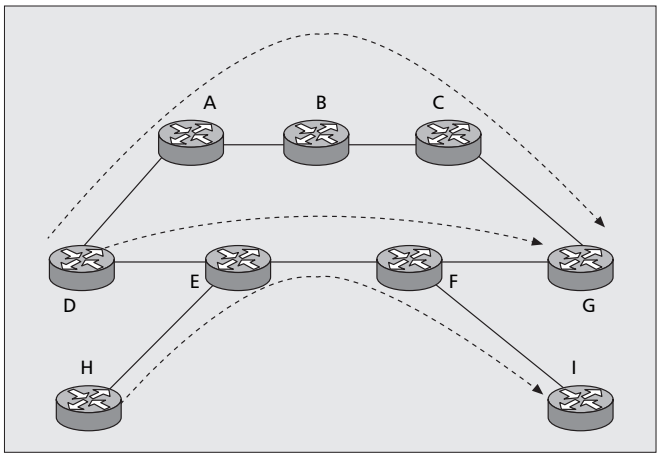
Reference	Optimization objectives/metrics	Optimization method	LSP type	Applicable environment
[24]	Minimize maximum utilization	Linear programming	P2P	Any
[26]	Minimize network cost with QoS constraints	Nonlinear programming (gradient projection)	P2P	DiffServ
[27]	Minimize network cost across multiple classes	Simulated annealing + gradient projection	P2P	DiffServ
[28]	Minimize the number of LSPs and hop counts	Heuristic + binary integer programming	MP2P	Any
[29]	Provide deterministic end-to-end QoS	Not available	MP2P	Any
[30]	Minimize the overhead in LSP labels	Not available	MP2P	Any
[31]	Minimize LSP bandwidth allocation	Not available	P2MP	Any

■ Table 2. *Offline MPLS-based solutions.*

portion (BWP) for handling the trade-off between the two metrics. Finally, a conventional Dijkstra's shortest path algorithm is applied based on the set of defined link weights.

One important issue often addressed in online MPLS-based TE schemes is the LSP interference problem [35–38]. The authors of [35, 36] noticed that by directly setting up LSPs (e.g., using CSPF) without considering the location of ingress/egress nodes for incoming traffic trunks, potential congestion is liable to take place at some critical links that multiple LSPs use. Competition by LSPs on the critical links that do not have sufficient available bandwidth for supporting all the LSP demands is known as LSP interference. Figure 5 provides a simple example of this. First, we assume an incoming traffic trunk from ingress node D to egress node G. If this is assigned the shortest-path-based LSP (D→E→F→G), future traffic trunks from H to I will be blocked if the residual link (E, F) cannot support both demands. In effect, we can find from the network topology that link (E, F) is critical to the traffic trunks from H to I in that any LSPs from H to I need to use that link. In this case a more intelligent strategy is to route the traffic trunk from D to G via an alternative longer path (D→A→B→C→G) and reserve the bandwidth on the critical link (E, F) for the future demand from the traffic trunk from H to I. From this example we can see that critical links are associated with the location of individual ingress/egress pairs. Hence, if the location of the ingress/egress nodes for traffic trunks is taken into consideration, the probability of LSP interference can be decreased if the LSP construction bypasses the critical links. Toward this end, the authors proposed the Minimum Interference Routing Algorithm (MIRA) to defer high loading on critical links. First, critical links associated with individual ingress/egress pairs are identified through calculating the maxflow value. Thereafter, an ingress/egress pair specific weight is created for each link, being an increasing function of its criticality. Finally, conventional shortest path algorithms are used according to the resulting link weights on top of the network graph containing only feasible links that can support the bandwidth demand of the incoming traffic trunk. The authors also implemented a software package called Routing and Traffic Engineering Server (RATES) [37], which is based on MIRA. In DAMOTE [38], decentralized agent for online MPLS-based TE, an algorithm for computing LSPs with minimization of a given objective function under bandwidth constraint is proposed. Examples of such objective functions are resource utilization, load balancing, and preemption-aware routing. DAMOTE computes in an efficient way that achieves near-optimal solutions.

Online MPLS-based TE has also been studied in DiffServ environments for QoS support, a typical example being TEAM [39]. The Traffic Engineering Tool (TET) in the TEAM framework is responsible for LSP preemption and construction. First, for each incoming demand, three types of cost are considered in the cost function: bandwidth, switching, and signaling. The objective of LSP manipulation is to minimize the overall cost throughout the process, which can be achieved by a Markov-process-based decision. There are two distinct LSP operations in TEAM: LSP preemption and LSP routing. LSP preemption allows existing LSPs to be preempted by newly constructed LSPs with higher priority. To do this, each LSP is assigned a priority attribute, which is taken into account when there is competition for resources (i.e., interference). Thus, even if an LSP has already been assigned a path, it will be rerouted if it has a lower priority attribute than a new LSP that is competing for the shared network resources. In order to avoid frequent LSP switching and thus traffic instability, the proposed preemption policy includes the following three guidelines: preempt the LSP with the lowest priority attribute, preempt the fewest number of LSPs, and preempt the least amount of bandwidth while satisfying the traffic demand requirement. For LSP routing, the Stochastic Performance Comparison Routing Algorithm (SPeCRA) [40] is adopted in TEAM. SPeCRA behaves like a homogeneous Markov chain where the optimal routing scheme is a state of the chain that is visited at the steady state. Specifically, it attempts to select adaptively the best routing algorithm from a set of candidate schemes, each of which might be suitable for a specific type of traffic trunk. The same authors also pro-



■ **Figure 5.** *LSP interference.*

Reference	Optimization objectives/metrics	Major LSP computing method	Applicable environment
[34]	Maximize future traffic demands accommodation with bandwidth guarantees	Heuristic (CSPF based)	Any
[35, 36]	Minimize LSP interference so as to accommodate maximum future demands	Heuristic (CSPF based)	Any
[38]	Minimize path hop count and improve load balancing	Heuristic	Any
[39]	Minimize bandwidth, switching and signaling costs	The SPeCRA algorithm [40]	DiffServ
[41]	Optimize LSP priority, number of LSPs and preempted bandwidth	V-PREPT for LSP preemption	DiffServ
[42]	Minimize loss of traffic flow	Heuristic (kth shortest path based)	Any

■ Table 3. *Online MPLS-based TE solutions.*

posed a new DiffServ-based LSP preemption policy known as V-PREPT that attempts to avoid LSP rerouting [41]. Similar to the TEAM scheme, the optimization for LSP preemption considers multiple criteria, including LSP priority, the number of LSPs, and the preempted bandwidth. With V-PREPT, the trade-off between the three criteria can be adaptively tuned according to the policy adopted by the INP. Apart from the simple LSP preemption algorithm, an adaptive version of V-PREPT was also proposed for reducing the overhead (essentially in signaling) introduced by frequent events of LSP teardown and rerouting. The basic idea of the adaptation is to allow some LSPs with lower priority attributes to have their rate allocation reduced so as to accommodate more requests in the future. In this case Resource Reservation Protocol with TE (RSVP-TE) signaling is responsible for indicating the updated allocation of rate on the static LSP, while there is no extra signaling overhead in tearing down and setting up LSPs. In DiffServ-based networks, this adaptive V-PREPT scheme is useful in LSP operations for the assured forwarding (AF) per hop behavior (PHB). Given the common practice that the expedited forwarding (EF) PHB is normally used to support hard QoS guarantees, bandwidth allocation in AF can be more flexible and dynamic, and the proposed adaptive V-PREPT algorithm can be efficiently adopted for this class of PHBs.

Survivable online TE in MPLS networks has also been considered in [42]. Similar to MIRA, this scheme constructs LSPs dynamically by applying the shortest path algorithm to the dedicated link weight metric that reflects the specific TE requirement. This type of dynamic link metric is based on a Lost Flow in Link (LFL) function that is used to assign working routes with local restoration. In LFL the metric of a particular link reflects the change in the objective function if an incremental demand has been (re)routed through or even near that particular link.

A summary of the existing online MPLS-based TE approaches is shown in Table 3.

Intradomain IP-based Traffic Engineering

Theoretical Background — The advent of plain IP-based TE solutions has recently challenged MPLS-based approaches in that Internet traffic can also be effectively tuned through native hop-by-hop-based routing, without the associated complexity and cost of MPLS. In [43] the authors proved that any arbitrary set of loop-free routes can be resolved into shortest paths with respect to a set of positive link weights that can be calculated by solving the dual of a linear programming formulation. This implies theoretically that if a network is optimally engineered through a set of loop-free explicit LSPs, by setting appropriate OSPF/Intermediate System to Intermediate System (IS-IS) link weights, this set of LSPs can be transformed into shortest paths according to this set of link weights. As a result, plain IP routers can directly compute this set of paths by using Dijkstra's algorithm, and hence the associated LSPs are not necessary anymore. Take the small network in Fig. 6a as a simple example (with symmetric weight setting in both directions of each link): The explicit path set {a→c→b, b→c→d} are shortest paths if we assign the weight value of 3 to links (a, b) and (b, d), and set the weight of all the other links to 1. Nevertheless, there are two major issues that restrict the practical deployment of link weight-optimization-based TE. First, not any arbitrary set of paths can be represented into shortest paths according to a set of link weights. For example, if we add another explicit path d→b→c to the aforementioned path set, as shown in Fig. 6b, these three paths cannot be represented simultaneously as shortest paths with any set of link weights, as the two paths b→c→d and d→b→c form a path cycle. As a result, these three paths can be enforced with MPLS explicit routing, but not with IGP link weight setting. Second, the distinct advantage of MPLS-based TE is not only explicit routing, but also arbitrarily unequal splitting of traffic. In this case, even if a set of LSPs can be represented as shortest paths, it is still not possible to unequally split the traffic given the underlying OSPF/IS-IS routers. Evolving from [43], [44] presented further analysis on the relevant issues in shortest path representability. One important contribution from this work is how to prevent unintended paths from becoming shortest paths when setting specific link weights. The authors argue that the network could suffer from traffic suboptimality if some bad paths are included in the shortest path set configured to deliver customers' traffic.

ECMP-Based Link Weight Optimization — In the Equal Cost Multipaths (ECMP) mechanism, if there are multiple shortest paths with equal IGP link weights toward the same destination, traffic is evenly split onto the next hop routers on these paths. Normally, the forwarding behavior in ECMP is on a per flow basis rather than a per packet basis to avoid out-of-order packet arrival. This multipath approach was first adopted and analyzed in the Netscope TE tool [45].

Fortz and Thorup [8–10] claimed that by optimizing OSPF/IS-IS link weights for the purpose of load balancing, the network service capability can be improved by 50–110 percent in comparison to the conventional configuration of link weight setting using inverse proportional bandwidth capacity. The key idea of the proposed algorithm is to adjust the weight of a certain number of links that depart from one particular

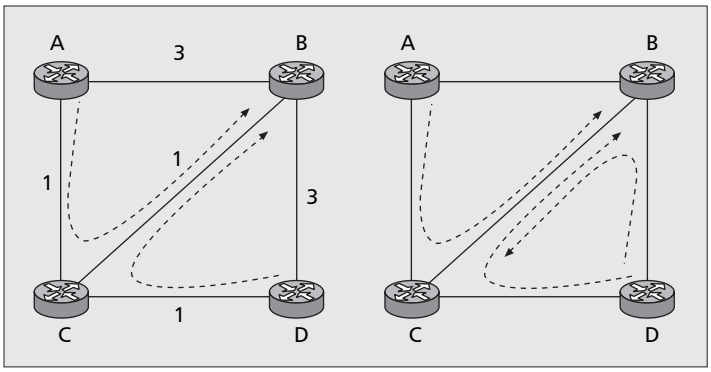


Figure 6. *Shortest path representation.*

node so new paths with equal cost are created from this node toward the destination. As a result, the traffic originally traveling through one single path can be evenly split into multiple paths with equal OSPF/IS-IS weights based on ECMP. In general, the authors proved that the optimal configuration of such link weights is NP-hard. Figure 7 provides a simple illustration of the basic idea of the algorithm. Consider destination node t and assume that part of traffic demand going to t travels through an intermediate node x. Fortz and Thorup's strategy is to split the flow to t going through x evenly along k links (x, x_i), $1 \le i \le k$, from x, if these links (x, x_i) belong to the shortest path from x to t. This type of "local adjustment" needs special attention, since shifting traffic might incur additional congestion to other links. In order to avoid this oscillation phenomenon, the authors apply sophisticated Tabu search for achieving the best load balancing performance.

Reference [46] also proposed a genetic algorithm (GA)-based approach for the same IP-based TE optimization problem, and the authors claimed that by properly tuning the GA parameters, the resulting performance is very close to that of [8–10]. Retvari *et al.* additionally raised some practical issues in OSPF traffic engineering, such as explicit knowledge of link capacity and reasonable range of OSPF link weight values [47]. Toward this end, the authors formulated the TE as the prime minimum cost maximum throughput problem, and the resulting link weight configuration provides a plausible basis to build a practical IP-based TE architecture.

Optimal routing often requires arbitrary traffic splitting. Instead of optimizing OSPF/IS-IS link weights, another TE approach for near-optimal network performance is to emulate uneven traffic splitting over ECMP paths at the edge or core routers. In [48] the authors proposed a scheme based on the manipulation of a subset of next hops for some routing prefixes; the scheme is capable of achieving near-optimal traffic distribution without any change of existing routing protocols and forwarding mechanisms. The basic idea is as follows. First, optimal link weights are calculated based on [43] through linear programming. Second, in order to deal with the requirement of arbitrary traffic splitting, the authors proposed activating only a subset of ECMP next hops for packet forwarding to the selected destination prefix so as to emulate unequal splitting of traffic in the MPLS-based solutions. Three different heuristic algorithms were studied for optimally configuring the next hop of unicast destination prefixes. This approach exhibits a typical strategy of making graceful trade-off between the performance and the overhead associated with the additional configuration needed.

Edge-Based Link Weight Setting — Wang *et al.* proposed in [49] a new OSPF traffic engineering approach without the necessity of ECMP splitting. Their approach is to divide the physical network into several logical routing planes, each being associated with a dedicated link weight configuration. There are two distinct procedures involved. First of all, the overall external traffic demands from all customers are partitioned properly into k traffic matrices only at the edge of the network, and each of the traffic matrices is identified by the type of service (ToS) or DiffServ code point (DSCP) in the IP header. Second, individual traffic matrices are independently routed over the k planes, each of which has its dedicated link weight configuration. The basic strategy of this approach is to emulate MPLS unequal splitting of flows by partitioning the overall traffic demand at the edge of the network so that traffic within different partitions is delivered through dedicated routing planes. To achieve the best overall traffic distribution, one of the most challenging tasks is to efficiently assign flows to the traffic matrices for different planes. Through simulations, the authors prove that a fairly small number of overlays ($k = 2$ or 4) can achieve near-optimal TE performance.

Table 4 presents a brief comparison of the IP-based TE approaches.

Online IP-Based Traffic Engineering — Unlike offline TE, which has been extensively studied, there are also few proposals for online or adaptive IP-based TE. Two online TE approaches are to change link weights on the fly and to make link weights sensitive to some loading or QoS parameters (e.g., to make the link weight a function of link utilization or delay). However, these approaches require the flooding of new link weights throughout the network, which can cause route instability and looping problems during the convergence process [50].

Another online TE approach is to dynamically adjust the traffic splitting ratio according to the network load. Adaptive multipath (AMP) [51] considers multiple nonequal cost paths and balances load by optimizing the traffic splitting ratios at each router. However, AMP only keeps network available information to a local scope rather than employing a global perspective of the network in each node.

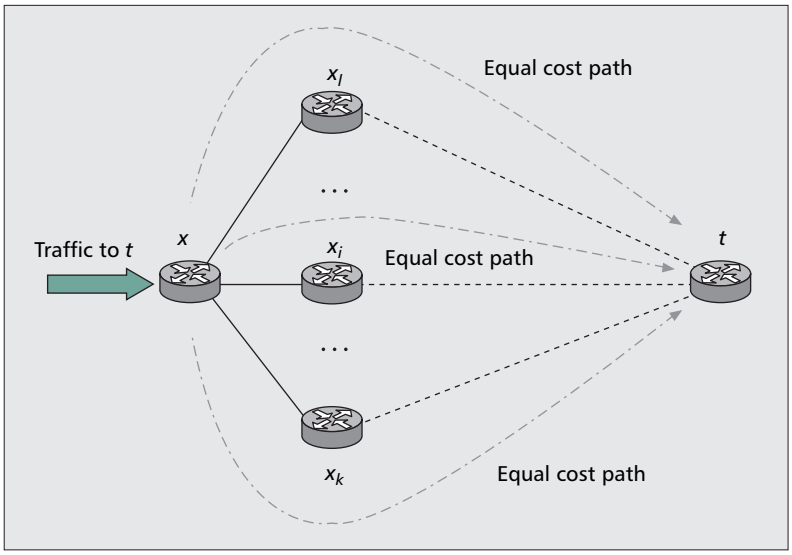


Figure 7. *Fortz and Thorup's link weight optimization algorithm.*

Reference	Feasibility	Traffic splitting	Protocol requirement	Configuration complexity	Performance
[43, 44]	Theoretical analysis only	Arbitrary splitting	—	—	Theoretically optimal
[8–10, 46]	Practical	ECMP	Plain IGP	Conventional IGP link weight setting	50–110% improvement
[48]	Practical	Selective ECMP	Plain IGP	Manual configuration of next hops for some prefixes	Near optimal
[49]	Practical	Traffic splitting at the network edge	ToS-aware routing with multi-RIB IGP	Configuration of multiple sets of link weights	Near optimal

■ Table 4. *IP-based TE solutions.*

Interdomain Traffic Engineering

In this section we introduce interdomain TE, an emerging topical research area that has evolved from its intradomain counterpart.

The Internet is a large decentralized internetwork composed of more than 26,000 ASs or domains by March 2007. From a business perspective, the relationship between any two domains can be classified into one of the following two types:

- Transit service (customer-provider relationship): This type of relationship exists commonly between low- and high-tier INP networks. Low-tier INPs (typically stub domains) purchase transit services from higher-tier INPs for Internet connectivity.
- Peering: This type of relationship exists commonly between neighboring INPs that are roughly equal in size and at the same tier. The INPs agree to simply exchange traffic without making any payment to each other.

We can also classify all the domains in the Internet into two categories: transit domains and stub domains. Transit domains offer transit services (i.e., interdomain traffic delivery across the Internet). Stub domains, on the other hand, are the leaf domains of the AS-level hierarchy. They only send or receive traffic, and do not provide transit services to any other AS. In general, the two types of domain have different interdomain TE objectives. The incentive for transit domains to perform interdomain TE is normally to optimize network resources so as to maximize their incoming revenue. On the other hand, stub domains compose more than 80 percent of ASs in the Internet, and most of them are multihomed. Hence, their principal interdomain issue is how to minimize the monetary expense of subscribing to Internet transit services from their INPs.

Another dimension for categorizing interdomain TE is inbound and outbound TE, which focus respectively on how to control interdomain traffic entering or leaving a domain. A domain may only require either inbound or outbound TE, or both, according to its business objectives. For example, a domain that contains popular content providers generates a large amount of traffic that needs to be sent out of the network efficiently, and thus outbound TE is needed. On the other hand, domains that have a large number of multimedia application receivers (e.g., Internet TV/MP3 subscribers) are typically traffic consumers. They therefore need to perform inbound TE in order to control traffic injected into their networks. Finally, since transit domains normally exchange Internet traffic between each other, both inbound and outbound TE may be required.

In the rest of this interdomain TE section we first give a brief introduction to the de facto interdomain routing protocol, BGP-4 [52], which can be used to perform interdomain TE by appropriately adjusting route attributes. Then some general guidelines for interdomain TE are presented. We then describe relevant TE work, classifying it into inbound and outbound TE. Finally, we discuss advanced interdomain TE paradigms such as cooperative TE between adjacent domains.

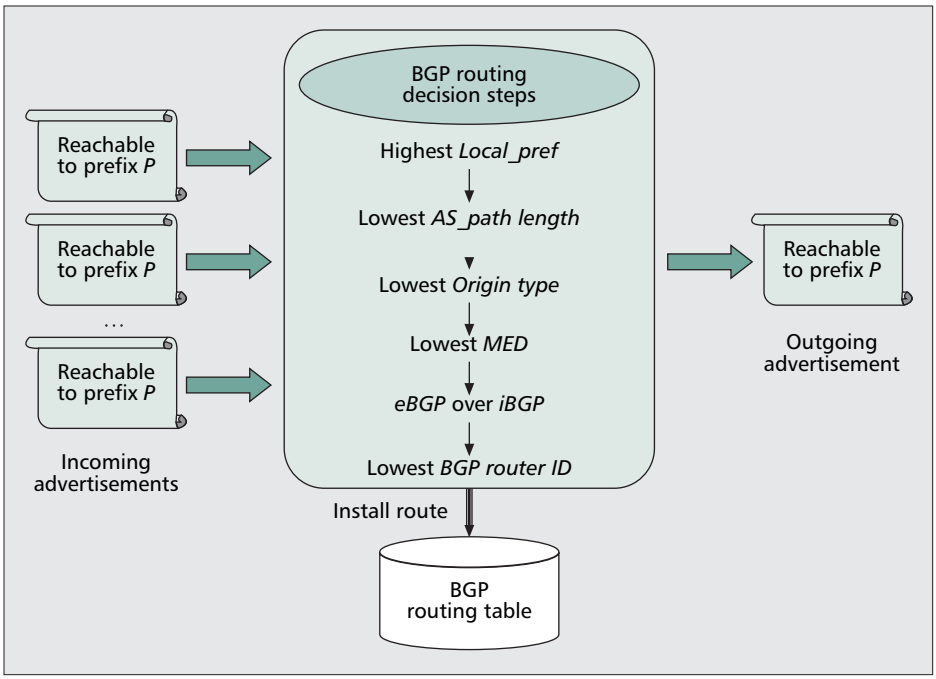
BGP Overview

BGP is the de facto interdomain routing protocol used to exchange routing information for the Internet. ASs are interconnected via dedicated interdomain links or Internet exchange points (IXPs). Border routers from different ASs exchange routing reachability advertisements through external BGP (eBGP) sessions, and these advertisements are also propagated to all other BGP speakers within the AS through internal BGP (iBGP) sessions. BGP enables import/export policies that enable INPs to control interdomain routes rather than always using the shortest AS paths. In the case where a BGP speaker receives multiple route advertisements for the same destination prefix, it selects only one of them as the best path according to the prioritized path selection process (Fig. 8), using the attributes associated with each route advertisement as the selection criterion. More specifically, if multiple BGP routes are received with the same value of the attribute in a higher priority, tie breaking is applied through comparing the attribute in the next priority, as the arrows indicate in Fig. 8. The best path is then installed in the IP routing table and exported to other peers.

As described above, interdomain TE can be classified into inbound/outbound TE, and an INP can configure BGP attributes to help achieve its TE objectives (Tables 5 and 7). From Fig. 8, it is obvious that only one single path should be selected for a particular destination prefix, because the final step of tie breaking is based on the unique IP address of the next hop of BGP peer. Some vendors have also implemented the BGP multipath functionality. In Cisco's BGP implementation, if the INP chooses to enable BGP multiple paths, the tie breaking criteria in steps 6–7 in the above process are overridden [53], which means that multiple (up to six) interdomain routes can be installed simultaneously into the BGP routing table for the same destination prefix. Similar to the intradomain scenario, this BGP multipath functionality provides flexible mechanisms for the INP to perform load balancing for transit traffic traveling through the network.

Interdomain TE Guidelines

Interdomain TE is performed by taking into account the routing information advertised by adjacent domains. We note that the change of TE configuration in one domain might affect the routing decisions of other ASs nearby, and this can propagate in a cascaded fashion. This often introduces route instability problems across the whole Internet, where a single



■ **Figure 8.** *BGP path selection process. The attributes used in BGP path selection are shown in the middle box.*

change of interdomain path may take up to several minutes to converge [54]. As a result, domains may be unable to predict whether their interdomain TE solutions can produce the target performance. Thus, interdomain TE should take into consideration how to preserve its predictability as well as stability so as to ensure stable traffic distribution and fast routing convergence [55]. For this purpose, recent research has proposed several guidelines for interdomain TE. We summarize the guidelines proposed in [54, 56] as follows:

- Achieving predictable traffic flow changes: The objective is to minimize the frequency with which upstream domains need to switch their outgoing traffic to different domains by changing the local BGP configuration. This adversely affects the traffic volume entering their networks.
- Limiting the influence of neighboring domains: The objective is to minimize the impact on routing decisions of neighboring domains. These routing decisions may contain inconsistent route advertisements from adjacent domains, which reduce the operator's control capability over traffic flows.
- Reducing the overhead of routing changes: If the traffic has to be separately engineered for all address prefixes in the Internet, the configuration overhead is too high to be realistic. To reduce this overhead, the number of destination prefixes to be considered should be limited through efficient address aggregation. In effect, it is suggested that INPs need only engineer the traffic toward a small number of popular destination prefixes that account for a large portion of Internet traffic [56]. This TE strategy allows INPs to efficiently control a large portion of traffic in the Internet by considering only a small number of prefixes.
- Customer routes preferred: Reference [54] has shown that Internet stability can be achieved by imposing a set of policies on individual domains. Thus, global coordination among all domains across the Internet is not necessary. The guidelines proposed in [54] ensure stable TE with fast convergence by favoring routing via customer domains over peer and provider domains. If customer domains are not directly available, routing via peer domains is preferred over provider domains.

Outbound Traffic Engineering

Outbound TE Mechanisms — A number of mechanisms are currently known for outbound TE, as shown in Table 5:

- Setting local preference (Local_pref): The local preference attribute has the highest priority in the BGP route selection process. The value assigned to this attribute indicates the preference on one border router to other candidates as the best egress point. Take Fig. 3 as an example. If the local preference value for the prefix 20.20.20.0/24 on the border router 10.10.10.1 is higher than that on 10.10.10.2, the traffic destined for AS 200 will use 10.10.10.1 as the egress point in AS 100.
- Hot potato routing: If multiple routes exist with equal value of BGP route attributes up to step 5 of the BGP route selection process shown in Fig. 8, the route with the lowest IGP weight from the ingress to the egress point is selected. This scenario is known as *hot potato* or *early-exit routing*, which is often adopted by large INPs. The objective of hot potato routing is to send the traffic to downstream domains across the core network as quickly as possible. By manipulating IGP link weights, an

Mechanism	Description	Implementation techniques	Applicable environment
BGP local preference (*Local_pref*)	To select the egress router directly by setting the highest BGP local preference value	BGP	Stub/transit domains
Hot potato routing	To select the egress router with the lowest IGP weight	BGP/IGP	Usually transit domains
Explicit routing (MPLS)	To select the egress router by establishing explicit paths across domains	RSVP-TE, BGP/IGP-TE, PCE	Stub/transit domains

■ Table 5. *Mechanisms for outbound inter-domain TE.*

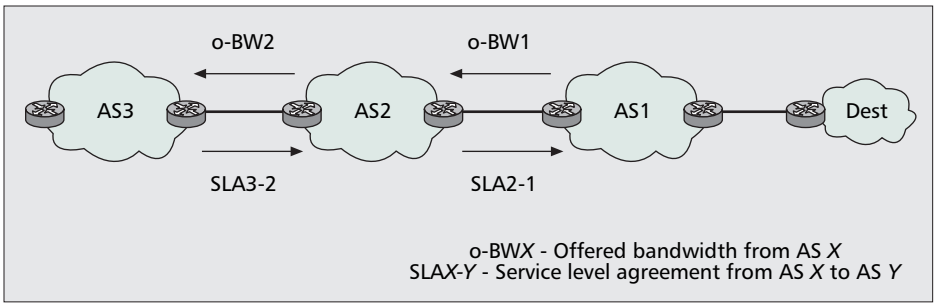


Figure 9. *Cascaded model for end-to-end bandwidth guarantee.*

INP is able to influence egress router selections within the local domain. In Fig. 3 we now assume that all the route attributes are "equally good" (Fig. 8, steps 1 to 5) for both 10.10.10.1 and 10.10.10.2. If the IGP weight of shortest path A (between 10.10.10.3 and 10.10.10.1) is lower than that of shortest path C (between 10.10.10.3 and 10.10.10.2), 10.10.10.1 is selected as the egress point according to hot potato routing.

- Explicit routing (interdomain MPLS): Interdomain MPLS enables a domain to enforce traffic to be delivered on the explicit paths to the destination across downstream domains. Thus, domains may establish explicit paths through their desired egress points to the downstream domains and destinations. Currently, mechanisms supporting interdomain MPLS have been proposed and implemented such as path computation element (PCE) [57].

Offline Outbound Traffic Engineering — We initially consider offline outbound TE in stub domains. The authors in [58] proposed offline optimization algorithms to distribute the traffic of a multihomed stub domain among multiple downstream INPs. The TE objective is to optimize both monetary expenses and network performance (measured by average latency). The authors found that the optimization of expenses and performance are often in conflict. In order to cope with this, they consider an approach that tackles expense and performance optimization separately and sequentially. First of all, the optimization of monetary expense is performed. This is based on the business operation viewpoint that minimizing the overall expense has higher priority than optimizing the network resource utilization in stub domains. Based on a percentile-based charging model, the objective of the optimization is to determine the amount of traffic to be sent to each of the downstream INPs so that the total charge is minimized. The performance optimization is then applied to assign the traffic to the downstream INPs. As a result, the total latency is minimized within the constraint of the computed expense. Instead of tackling the expense and performance optimization in a lexicological importance order, the authors in [59, 60] proposed a multi-objective evolutionary algorithm to solve a similar optimization problem. The aim is to find a compromising solution that is good with respect to all the optimization objectives. As with [58], the metric to be minimized is the charge incurred by the downstream INP, whereas the performance to be optimized is the load balancing across the interdomain links. In addition to these two objectives, the authors also consider how to minimize the iBGP communication overhead in order to enforce the TE decisions. The authors in [61] introduced an INP subscription problem of subscribing to a set of downstream INPs so as to minimize the cost in payment. The INP subscription problem is different from the above mentioned expense optimization in that the latter assumes that the INP subscription decision has already been made; thus, traffic can only be assigned to the subscribed downstream INPs. However, in order to further minimize the monetary expense, a domain may have the freedom to select the optimal set of downstream INPs from all the available candidates and then assign traffic to this set of INPs. The INP subscription problem is based on a percentile-based charging model and is solved through dynamic programming. The authors in [62] addressed a similar INP subscription problem on top of a total-volume-based charging model. Their work goes one step further: the chosen downstream INPs also need to provide end-to-end bandwidth guarantees toward the destination domains. The problem is solved by a GA-based approach.

We now describe a number of schemes that focus on transit domain TE issues. The BGP TE approach proposed by Bressoud *et al.* [63] was the first piece of work dealing specifically with outbound interdomain TE for transit domains. The objective of the TE problem is to determine an optimal set of egress points for the advertisement of destination prefixes so as to minimize the traffic cost (i.e., bandwidth consumption) while satisfying the bandwidth capacity constraints of the interdomain links. The outbound interdomain TE problem is further subdivided into two parts: single egress selection (SES) and multiple egress selection (MES). SES ensures that one and only one egress point is selected for each destination prefix, whereas MES allows multiple egress points. Two heuristic algorithms, combining the approximation algorithm proposed for the generalized assignment problem (GAP) with a simple greedy heuristic, were proposed to solve the SES and MES problems. Finally, the authors in [64] proposed an open source tool, called Tweak-it, for outbound interdomain TE in large transit domains. The authors in [65] extended outbound interdomain TE to support end-to-end bandwidth guarantees across transit domains. Their work is based on the MESCAL cascaded model that allows negotiations between adjacent domains and achieves bandwidth guarantee by establishing INP-level service level agreements (SLAs) [66]. As Fig. 9 shows, each domain offers its upstream neighbor (through provider SLAs) a guaranteed bandwidth (o-BW) toward each destination aggregate prefix (Dest). Each SLA is associated with the amount of available bandwidth that is guaranteed from the offering downstream domains to the destination domains. In order to provide end-to-end bandwidth guarantees for the traffic, the outbound interdomain TE problem has been extended for finding not only an optimal egress point that maintains the capacity constraints of interdomain links and SLAs, but also the paths within the network to satisfy the traffic demand requirement. In [65] the TE objectives are to minimize the total bandwidth consumption and balance the load over intra- and interdomain links. The authors in [67] proposed an interdomain TE system for provisioning end-to-end delay guarantees in addition to meeting bandwidth requirements.

Online Outbound Traffic Engineering — In the literature, online outbound TE schemes have only focused on stub domains. They can be classified into the following two types:

- Proactive: These TE solutions rely on traffic predictors to forecast traffic for a short time interval (e.g., minutes) and then run a lightweight TE algorithm in a quasi-offline manner to produce solutions in a short timescale.
- Reactive: These TE solutions are adaptive and dynamic to incoming traffic demand without traffic prediction beforehand.

Reference	Optimization objectives/metrics	TE semantics	Implementation techniques	Applicable environment
[58]	Minimize overall expenses and end-to-end latency	Offline/online	Not specified	Stub
[59, 60]	Minimize overall expenses, improve inter-domain load balancing and minimize BGP communication overhead	Offline	Local_pref	Stub
[61]	Minimize overall expenses	Offline	Not specified	Stub
[62]	Minimize overall expenses and provide end-to-end bandwidth guarantee	Offline	Not specified	Stub
[63]	Minimize network cost (e.g., bandwidth consumption)	Offline	Local_pref, AS path	Transit
[65]	Minimize network cost and provide end-to-end bandwidth guarantee	Offline	Not specified	Transit
[69]	Minimize overall expenses, improve inter-domain load balancing and minimize iBGP communication overhead	Online	Local_pref	Stub
[70]	Turn-around delay	Online	Not specified	Stub
[71]	Round Trip Time (RTT)	Online	Local_pref	Stub

■ Table 6. *Outbound traffic engineering approaches.*

In [58] the authors proposed proactive online algorithms for multihomed domains to select appropriate INPs for outbound traffic. The objective is to minimize first the total expense and then the end-to-end latency. The approach to short-term traffic forecast is based on the exponentially weighted moving average (EWMA) method. In this scenario traffic prediction is performed through detecting traffic changes based on a sequence of independent preceding observations. The proposed online TE algorithm is a greedy heuristic based on traffic sorting, which has also been used for solving the bin-packing problem [68]. Another proactive online TE approach was addressed in [69]. The authors designed a systematic BGP-based outbound TE technique for stub domains over the timescale of minutes. Apart from the TE objectives considered in [58, 69] also investigated how to minimize the overhead of the associated iBGP message advertisements. A quasi offline multi-objective evaluation algorithm was proposed to solve the online outbound TE problem.

For reactive TE paradigms, the first work on quantifying the benefits of dynamic route selection with multihoming was proposed in [70]. The multihomed domain under consideration may subscribe to multiple downstream INPs, and it also measures the end-to-end path performance (turnaround delay) through each downstream INP toward the destination. Based on the performance obtained from measurement, the domain dynamically switches traffic to the INP that has the best instant performance. Compared to random selection of INPs, the measurement-based multihoming approach can achieve a 40 percent performance improvement in terms of the average turnaround delay. Based on this approach, the authors in [71] proposed a round-trip time (RTT) measurement approach for outbound route selection. The proposed approach is scalable and does not require RTT measurements via each INP to individual large numbers of destinations.

To summarize the outbound traffic engineering schemes in this section, we list and compare in Table 6 the major characteristics of the solutions presented in this subsection.

Inbound Traffic Engineering

Inbound TE Mechanisms — In this section we first provide an overview of available mechanisms for inbound TE. As with outbound TE, although there are various candidate implementation mechanisms, inbound TE routing optimization algorithms have only used a few of them (e.g., AS path prepending.) Nevertheless, we list all of the potential mechanisms in Table 7 based on which inbound TE can be performed.

- Selective advertisement. In this approach routes toward a destination prefix are only advertised through a set of chosen ingress links. We take Fig. 10 as an example. If AS300 would like to receive traffic from AS400 via ASBR 30.30.0.1 heading toward AS301, it chooses not to advertise the route to AS301 through ASBR 30.30.0.2. However, the shortcoming of this approach is that if the chosen ingress point fails, no alternative routes can be used as backup.
- More specific advertisement. In this approach, if multiple routes exist toward the same destination, the one with the longest matching prefix will be selected. In Fig. 10 we assume AS300 advertises to AS400 the reachability of destination prefix 30.30.0.0/16 on 30.30.0.1, and its subprefix 30.30.30.0/24 on 30.30.0.2. As a result, the traffic toward any destination in "nested" AS301 will not use 30.30.0.1, as the other ingress router has a route with a more specific prefix. Compared to selective advertisement, this type of ingress point selection is more robust in case of link failure. If the interdomain link attached to 30.30.0.2 breaks, the traffic toward AS301 can still be routed via 30.30.0.1 using the route with a more coarse-grained prefix. It is worth mentioning that advertising too many specific prefixes may cause the scalability problem in terms of increase in BGP routing tables, which is the main reason this approach is not commonly considered for interdomain TE.
- AS path prepending. A route advertisement is made less attractive to upstream domains by adding several instances of AS number to the AS path attribute to inflate the AS

Mechanism	Description	Implementation techniques	Applicability environment
Selective advertisement	Advertise a route only at the set of ingress points that is expected to receive traffic	BGP	Stub/transit
More specific advertisement	Advertise routes with more specific prefixes to suppress the coarse-grained ones	BGP	Stub/transit
AS path prepending	Inflate the length of the AS path attribute to reduce the attractiveness of the route	BGP	Stub/transit
Lowest MED value	Advertise preferred routes with the lowest value of MED	BGP	Stub/transit
Community attribute	Suggest to adjacent domains how to manipulate the advertised routes	BGP	Stub/transit
Network address translation	Modify the packet headers by assigning the desired ingress point as the source of packets	NAT	Usually stub
BGP overlay	Direct communication between any two domains bypassing BGP	User specified	Stub/transit

Table 7. *Mechanisms for inbound interdomain TE.*

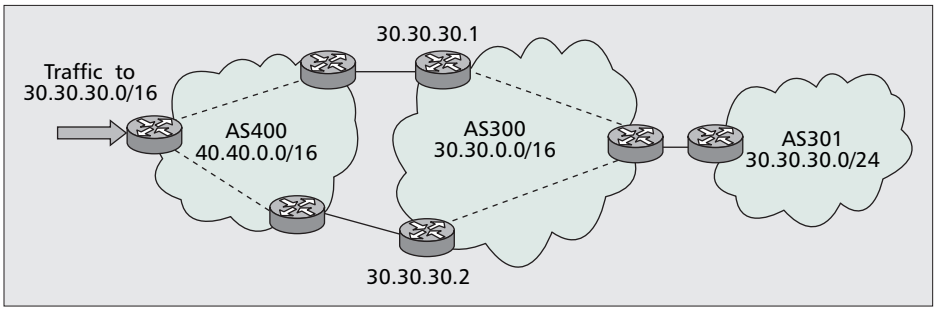


Figure 10. *Inbound traffic engineering examples.*

path length of that route. In Fig. 10, if AS300 would like to receive traffic from AS400 toward AS301 via ingress point 30.30.0.1, it may prepend its own AS number in the advertisement on 30.30.0.2 such that the overall AS path via this ASBR is made "longer" than via 30.30.0.1. It should be noted that this is only possible if AS400 does not apply the Local_pref metric to select the preferred route. Related work on and performance evaluation of AS path prepending can be found in [72–74].

- Setting Multi-exit Discriminator (MED) value. This applies only if two adjacent ASs have two or more direct connections between them, and both ASs agree to implement MED. In these circumstances a domain may select its preferred ingress router by assigning a lower MED value. Consider the example of Fig. 10; if AS300 would like to receive traffic from AS400 via 30.30.0.1, it may advertise a BGP route with a lower MED value through this router than the one on 30.30.0.2. The prerequisite for using the MED metric for ingress point selection is that all the route attributes with higher BGP route selection priority for the two routes should be set equal (e.g., the Local_pref metric set internally by AS 400 and the AS path length via the two border routers).
- Community attribute. In this approach a route can be advertised associated with the community attribute that instructs upstream domains how to manipulate this route with certain actions. For example, AS path prepending can be included in the community attribute to instruct upstream domains to perform AS path prepending before sending route advertisements to their specific upstream domains [75, 76].
- NAT address translation. This approach manipulates Network Address Translation (NAT) tables [77, 78]. The NAT rules associate destination prefixes with the best ingress point such that the source address in packets for the destination is translated to the address of the chosen ingress point.
- BGP overlay. An overlay policy control architecture (OPCA) has been proposed to separate the policy from routing so that a faster channel can be used to handle routing policy changes [79]. OPCA consists of several major components including policy agent and database, measurement infrastructure, message propagation, and so on. The aims of OPCA are to solve the BGP convergence problem by improving route failover time and to balance the inbound traffic load for multihomed domains.

Offline Inbound Traffic Engineering — In [80] the authors addressed an offline inbound interdomain TE problem by optimizing AS path prepending for stub domains. The problem is called constrained optimal prepending (COP). The objective of COP is to determine the minimum number of prepended ASs for each prefix advertised through each ingress link such that the load constraint on each ingress link is satisfied. An essential assumption in this work is that the inbound route selection at the local domain is not affected by the setting of the Local_pref attributes in its upstream domains. This is because, if Local_pref is used, the upstream domains may send the traffic through another path toward the local domain using different ingress links. As a result, this makes the effect of AS path prepending hard to predict. An Optimal Padding Vector (OPV) heuristic algorithm is proposed for solving the COP problem. The basic idea of the OPV algorithm is first to identify the most overloaded ingress link at each time, and then to increase the AS path length by one of all customer prefixes to be advertised through the

Reference	Optimization objectives/metrics	TE semantics	Implementation techniques	Application Environment
[72]	Minimize link congestion and foresee performance impact	Online	AS path prepending	Stub
[73]	Improve load balancing	Online	AS path prepending	Stub/Transit
[78]	Reduce traffic request response time	Online	NAT	Stub
[80]	Minimize the number of prepending with the bandwidth constraint of ingress links	Offline	AS path prepending	Stub

■ Table 8. *Inbound traffic engineering solutions.*

ingress link. The algorithm iterates until the traffic load received by each ingress link satisfies its maximum load constraint.

Online Inbound Traffic Engineering — In [72] the authors proposed a systematic and automated procedure named AutoPrepend to control inbound traffic using AS path prepending. The basic operation of AutoPrepend is to artificially inflate the length of the AS path attribute in order to divert traffic onto different ingress links until the outcome network performance meets the TE goals. AutoPrepend is composed of four components:

- Passive measurement: To identify a set of top senders responsible for most of the inbound traffic.
- Active measurement: To send ICMP echo requests to the set of top senders and record the ingress links that receive the ICMP replies. A virtual beacon prefix with inflated AS path length on one of the ingress links is sent to the set of top senders. The ingress links where the top senders respond to the beacon prefix are examined.
- Traffic prediction: Based on passive and active measurement, to predict the changes in the traffic volume on each ingress link when AS path length increases. This is accomplished by comparing the measurements from the ICMP requests and the beacon prefixes described above.
- AS path update: To check if the predicted outcome satisfies the traffic engineering goals. If so, enforce the change by advertising the prefixes with the chosen AS path length.

The authors in [73] proposed a greedy AS path prepending heuristic algorithm to apply the above mentioned algorithm to the most heavily (or least) loaded ingress link and then virtually inflate (or decrease) the AS path length of the routes through the link by one until the TE goals are met.

In [78] the authors proposed the use of the NAT-based approach to control inbound traffic through the best ingress point. The instantaneous performance of the connected ingress points is continuously measured through active or passive measurement methods. The ingress link that gives the best performance is then selected for a given transfer.

A summary of the existing inbound TE work is presented in Table 8.

Compared to the outbound scenario, BGP-based inbound TE is more difficult for INPs to put into practice. This is generally because the BGP routing attribute used for outbound TE (Local_pref) can always "suppress" those attributes used for inbound TE (e.g., AS path and MED). Let's take Fig. 10 as an example again. If AS300 decides to receive traffic from AS400 via 30.30.0.1 through either AS path prepending or setting appropriate MED values, AS400 can still force the traffic to be injected into the downstream domain through 30.30.0.2 by setting higher Local_pref on this border router. Given the situation that individual INPs may have different or even conflicting routing policies, it is not surprising that this happens from time to time in the Internet. To solve this problem, cooperative TE between adjacent domains have been proposed, which are described in the following section.

Cooperative Interdomain Traffic Engineering

Since most domains in the Internet are self-governed entities and are effectively in competition with each other for customers, it is natural that they perform interdomain TE individually without considering their neighbors. However, recent research has found that when adjacent domains perform their interdomain TE selfishly, not only is the global network performance not optimized, but also the interdomain TE strategies of each domain may adversely affect each other [81]. In this case routing instability may occur, as domains need to change their path selection strategies whenever the TE decisions of their adjacent domains change. Such instability is primarily due to interdomain TE policy conflicts between domains. A desirable way to achieve overall good TE performance is to encourage INPs to negotiate with each other in order to obtain a compromising solution that benefits them all. This is known as cooperative-based TE [82].

Cooperative-based TE relies on the negotiation between two adjacent domains to achieve an agreement on how traffic is routed between their networks. The TE objectives of the adjacent domains should be jointly considered in order to achieve a "win-win" agreement that is satisfied by participating domains. Such an agreement can be determined through intelligent optimization methods, taking into consideration the topologies, TE objectives, and traffic matrices of the two domains.

Compared to the existing effort on independent outbound and inbound TE, a very limited number of papers have investigated routing optimization using cooperative TE. In [83] the authors formulated an optimal peering problem for two domains that have agreed to establish peering relationships. The problem is to determine how many peering points are needed and how they are located such that the total cost of peering is minimized without compromising interdomain service quality. With the peering point fixed, traffic is routed through the agreed ingress and egress points. A similar optimal peering problem has also been formulated in [84]. Most recently, cooperative interdomain TE schemes have also been addressed using game theory and nonlinear programming (specifically Nash bargaining and dual decomposition techniques) [85].

Apart from the optimal peering problem, the authors in [86] proposed using IP tunneling to establish explicit paths between source and destination domains through the ingress links that are chosen to receive traffic. This approach is assumed valid in the environment where all network domains are cooperative. In addition, the authors in [87] proposed an

algorithm for optimal route control among a group of cooperative multihomed stub domains in order to reach a global TE solution that avoids oscillation caused by any conflict on TE objectives between domains.

Multicast Traffic Engineering

The problem of how to optimally engineer multicast traffic is far less well understood than unicast TE. A common objective of multicast TE is to minimize the total amount of bandwidth to be consumed. This objective is also known as bandwidth conservation, where conventional shortest-path-based routing paradigms are normally not optimal solutions. In the literature bandwidth conservation in multicast routing is formulated as the directed Steiner tree problem [88], which has been proved to be NP-hard. It is worth mentioning that the task of multicast TE is not necessarily identical to the classic Steiner tree problem. For example, apart from bandwidth conservation, there are also some other TE objectives such as load balancing and maximizing throughput.

MPLS-based Multicast Traffic Engineering

The most straightforward approach to MPLS-based multicast TE is to set up P2MP LSPs, and this is where Steiner tree algorithms play a role. Before considering individual multicast TE schemes, we first investigate how to aggregate multicast traffic from different groups, which is an important procedure prior to LSP computation. This issue was first addressed in [89], and a scheme known as Aggregate Multicast was proposed. In this scheme multiple multicast groups are forced to share one single P2MP LSP, even if the egress router set of these groups does not completely overlap. At the expense of some extra bandwidth consumption, this approach is able to significantly reduce the total number of LSPs needed, thus improving scalability.

In [90] the authors proposed the Edge Router Multicasting (ERM) scheme for setting up P2MP LSPs only at the boundary of an MPLS domain. In ERM multicast traffic aggregation in LSPs is confined to the network edge; thus, the task is reduced to unicast TE within the domain. The authors studied two types of ERM: the first scheme is based on modifications to the existing multicast protocols, while the second approach applies a Steiner-tree-based routing heuristic at edge routers.

Apart from an offline approach, online multicast traffic engineering has also been investigated, where future multicast sessions are not known a priori. In [88] Kodialam *et al.* extended their MPLS-based online unicast TE scheme [36] to a multicast semantic. The basic objective is to accommodate as many multicast routing requests as possible without knowing about any incoming traffic in advance. The authors proposed a directed Steiner-tree-based online multicast routing algorithm for computing dynamic multicast trees with minimum bandwidth interference between individual sessions.

IP-based Multicast Traffic Engineering

Despite theeir flexibility, explicit-routing-based TE approaches suffer from the complexity and cost associated with MPLS deployment. This problem becomes more serious in supporting multicast services, as P2MP (other than point-to-point) LSPs need to be maintained throughout the network. Compared to the unicast scenario, another difficulty in MPLS multicast TE is how to aggregate multicast flows, because different multicast sessions tend to have different egress routers attached to group members. As described above, this problem was addressed in the Aggregate Multicast scheme [89], but the associated scalability issue is still left open for further investigation. Naturally, one might wonder if it is also possible to engineer multicast traffic without MPLS enforcement (e.g., by using plain IP based paradigms). The answer is yes, but the number of relevant publications has been very small. The reason for this situation can be summarized as follows. First, Protocol Independent Multicast - Sparse Mode (PIM-SM) [91] uses the underlying IP unicast routing table for the construction of multicast trees, and hence it is difficult to decouple multicast TE from its unicast counterpart. Second, the enforcement of Steiner trees can be achieved through packet encapsulation and explicit routing mechanisms such as MPLS tunneling. However, this approach lacks support from hop-by-hop protocols, due to reverse path forwarding (RPF) in the IP multicast routing protocol family. In PIM-SM, if multicast packets are not received on the shortest path through which unicast traffic is delivered back to the source, they are discarded so as to avoid traffic loops. Given the difference between the shortest path tree used by PIM-SM and the optimized minimum hop Steiner tree, engineered multicast traffic for bandwidth optimization through Steiner tree heuristics could result in RPF check failure.

The authors in [92] first stated that the theorem proved in [43] can also be applied to P2MP routes. This implies that a set of loop-free Steiner trees can also be represented theoretically in shortest path trees with a proper set of link weights. Thus, it is also possible to engineer multicast trees into Steiner trees for bandwidth conservation purposes without IP layer RPF check failure. However, the authors did not propose how to achieve this type of tree representation in their work. To fill this gap, the authors of [93] proposed a GA-based approach to optimize PIM-SM multicast trees with bandwidth constraint by setting properly the underlying IGP link weights. The objective is to achieve bandwidth conservation and load balancing through tuning the link weight of multitopology-enabled IGP (MT-IGP) protocols such as M-ISIS [94] and MT-OSPF [95]. The most distinct advantage of these two protocols is that they allow multiple sets of link weights for the same physical topology, with each corresponding to a specific type of traffic. In this scenario multicast TE can be effectively decoupled from its unicast counterpart given the underlying MPLS-free environment. Figure 11 illustrates a simple example of how to conserve bandwidth in multicast routing by configuring optimized M-ISIS/MT-OSPF link weights. In this example the single multicast source is node A, and nodes E, F, and G are multicast group members. By conventional hop count shortest-path-based PIM-SM routing, the bandwidth consumption is 6 units, with 1 unit consumed on each on-tree link. However, with proper link weight setting for MT-IGP, the optimal multicast tree for the same group is in effect a Steiner tree in terms of hop counts, with only 4 units of bandwidth being consumed (Fig. 11b). In general, the practical approach is to optimize multiple multicast trees with only a set of MT-IGP link weights.

Some Traffic Engineering Considerations

In this section we discuss some important issues that need to be considered in routing optimization for advanced TE, specifically TE robustness, TE interactions, and interoperability between TE and overlay selfish routing.

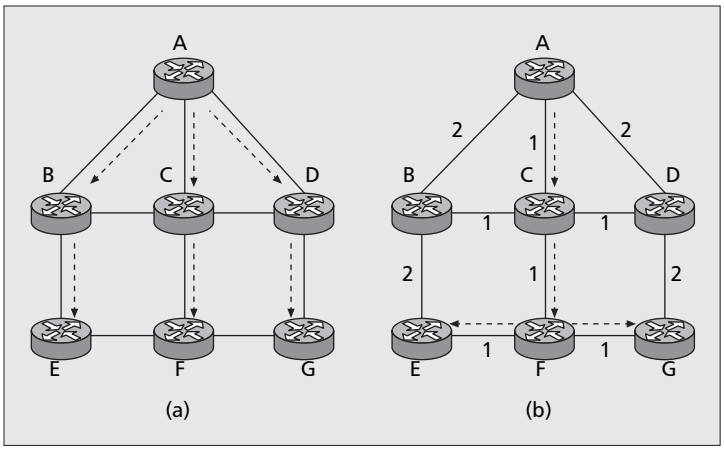


Figure 11. *Steiner tree with IGP link weight optimization.*

TE Robustness

Most of the offline TE solutions described in this article are based on the assumption that TMs are accurate and the network is operating under normal conditions. However, to derive accurate TMs is far from trivial due to the dynamic nature of Internet traffic. Moreover, failures, in particular logical ones, often occur in core networks. As a result, traffic fluctuation and network failure may cause TE performance to be unpredictable, and thus make network management more complicated. Hence, it is necessary to make TE more robust in order to maintain the expected performance when any of those situations take place. Apart from achieving the expected performance, another advantage of this robust approach is that only one relatively stable network configuration is needed without frequent changes in response to the occurrence of any unexpected situation.

In the literature robust TE has considered two issues: link failure and traffic demand uncertainty. The idea of the robust TE approach is first to model these issues as separate scenarios. For example, each link failure or TM represents a distinct scenario. Thereafter, a single TE configuration is produced that performs well in any given scenario.

As for the case of intradomain link failure, which has been found to be common and transient, [4, 12, 96–99] proposed IGP link weight setting algorithms to achieve the desired performance at any single link failure scenario. However, the computational complexity of algorithms increases significantly as the number of links in the network gets larger. In order to reduce such complexity, [12] further suggested performing robust TE optimization only on the critical links that have a significant impact on overall network performance. Recently, multitopology IGP link weight setting for robust intradomain TE has also been proposed [99]. The idea is that traffic can be shifted to alternative IGP topologies (hence alternative IGP paths) in order to retain load balancing once link failures are detected. For MPLS, the authors of [100] considered combined working and backup LSP optimization for all traffic demands. Specifically, a proactive ingress-to-egress restoration scheme with resource reservation was studied. The objective is to maximize the network's ability to carry future demands. Through this MPLS-based TE, the traffic carried over the network is fully restorable against all single event failures. Given that interdomain peering link failures are as common and transient as intradomain link failures, the authors of [101] proposed a local search heuristic to obtain an outbound interdomain TE solution that is robust to any interdomain link failure. Their objective is to minimize interdomain link utilization under both normal state (no failure) and failure state with any single interdomain link failure.

Traffic engineering in the case of multiple TM scenarios for the purpose of handling traffic demand uncertainty is relatively new. For intradomain TE, Applegate and Cohen [102] found that it is possible to obtain a robust routing configuration that guarantees nearly optimal utilization with fairly limited knowledge of the applicable TMs. Similar work with link failure consideration was also proposed by the same authors [103]. Based on their work, the authors in [104] proposed algorithms to solve the robust intradomain TE problem. Instead of using distinct TM scenarios, Mitra and Wang [105] proposed a stochastic optimization approach which assumes that the traffic demands are given probability distributions. Apart from being used for TM uncertainty, the robust TE approach can be used to obtain a high chance of performing well for multiple TMs, each of which represents traffic demands in a distinct period (e.g., days and evenings). This can be achieved through a set of OSPF link weight settings with the changing of a few link weights for different time periods [9]. This approach reduces the complexities in network management, as network operators do not need to change link weights on a regular basis. The COPE MPLS-based TE approach [106] was proposed to optimize for the expected TM scenarios while providing a worst case performance guarantee for unexpected ones, including those caused by link failures and traffic spikes. On the other hand, for interdomain TE, the authors in [107] proposed an outbound TE approach based on scenario-based robust optimization, taking as input a set of interdomain TMs. The objective of their work is to obtain an outbound TE solution that achieves good maximum interdomain link utilization while minimizing the performance gap between the achieved solution and the optimal solution for any given interdomain TM.

The ultimate objective of using robust TE approaches is to make network design and provisioning more predictable. This topic has been further receiving attention in designing a predictable Internet backbone network using novel approaches. Zhang and McKeown [108] proposed using Valiant load balancing over a fully connected logical mesh for backbone network design. The aim of this approach is to achieve predictable and guaranteed performance, even when TMs change, and links and routers fail. Kodialam *et al.* [109] proposed a simple static routing scheme that is robust to extreme traffic fluctuations without requiring significant network overprovisioning.

TE Interactions

Earlier we classified traffic engineering into a set of categories. In this section we discuss TE interactions within each category from the viewpoint of routing optimization.

Intra-/Interdomain TE Interaction — Much research has been conducted on intradomain and interdomain TE, respectively, but how they work together as an integrated TE paradigm has not been well addressed. Recently, some publications have indicated that the interaction between intra- and interdomain TE significantly impacts overall performance [110]. First, any change of BGP ingress/egress point for traffic across a domain influences the intradomain TM and leads to significant impact on the effectiveness of intradomain TE [110]. Hence, a more appropriate TE strategy is to take intradomain conditions into consideration when performing interdomain TE. For example, when selecting an egress point for any traffic trunk with bandwidth requirements, a prerequisite is to guarantee that at least one feasible intradomain path with sufficient network resources exists between the ingress-

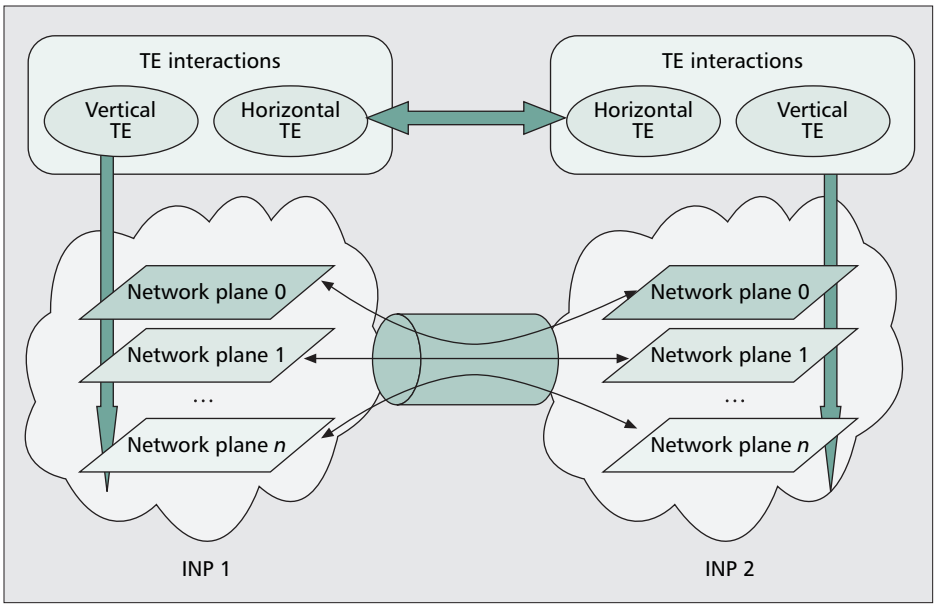


Figure 12. *Horizontal/vertical TE interactions.*

egress pair. In [111] the authors proposed a joint optimization approach of intra- and interdomain TE that is solved by a local search heuristic algorithm. Their results show that performing intra- and interdomain TE simultaneously can maximize the network's capability to accommodate future traffic demands better than a sequential or nested approach that performs both types of TE separately.

The configuration of intradomain TE can, however, also impact interdomain path selection. A typical example is HPR, often used by large INPs [5]. According to the BGP route selection policy, if multiple routes toward the same destination prefix are received through the same type of e/iBGP advertisement with identical values of Local_pref, origin type, AS path length, and MED, the route having the lowest intradomain IGP link weight is selected. Today, many INPs adopt HPR, which allows IGP link weights to influence egress router selection. By doing so, they hope that the traffic can be delivered out of the local domain using the least number of hops (assuming each IGP link weight to be 1), which indicates that the least bandwidth resources are consumed. However, HPR also potentially leaves the interdomain traffic instability problem in time of link failure. We reuse Fig. 3 as an example. Assume that the INP of AS100 applies HPR for traffic delivery toward AS200 via egress node 10.10.10.1 according to its TE requirement. To achieve this, the configured IGP link weight for the shortest path between 10.10.10.3 and 10.10.10.1 (i.e., path A) should be lower than its counterpart between 10.10.10.3 and 10.10.10.2 (path C). Under this configuration, in case of a link failure on path A, the whole traffic trunk toward AS200 will shift automatically to use 10.10.10.2 as the egress point in AS100 if the IGP weight of the newly formed shortest path between 10.10.10.3 and 10.10.10.1 (e.g., path B) is larger than that of path C. In this scenario, not only does traffic routing within the network become unstable, but also the original TE objectives may be violated. With this example, we can see that intradomain TE might also interact with interdomain path selection. By showing the above examples, we indicate the importance of the intra-/interdomain TE interaction, and we believe that further investigation in this area is worthwhile for more effective and robust TE.

MPLS/IP-based TE Interaction — We showed earlier the distinct advantages and disadvantages of using IP/MPLS-based TE schemes. Recently, some proposals have been made to integrate IP and MPLS technologies to provide a hybrid TE solution. In [112] the authors suggested the option of using LSPs only to reroute the traffic trunks that potentially contribute to network congestion, while the rest of the traffic is routed through plain IGP. In this case the overhead introduced from LSP states can be reduced significantly at the expense of reasonably less flexibility in path selection. In the offline scenario, how to set up LSPs and configure IGP link weights so as to achieve overall network optimality is the key objective of the hybrid TE approach. If the IGP link weight is properly calculated, the number of LSPs needed for explicit routing to eliminate congestion can be reduced. In addition, hybrid online TE with both IGP and MPLS has also been investigated in [113–115]. These works aim at efficient allocation of unpredictable incoming traffic trunks onto different routing planes. In both cases the interaction between IP-based and MPLS-based TE on top of the same physical network is of significant importance, as there is a typical trade-off between performance and scalability that should be taken into consideration by INPs.

Offline/Online TE Interaction — Despite the fundamental difference between offline/online TE described earlier, it is still possible, and even desirable in some circumstances, to combine them for more sophisticated TE optimization. Although TMs can sometimes be obtained in advance (e.g., through SLSs) to provide the possibility of offline TE, it is not always the case that the overall traffic demands can be accurately predicted. In this case static configuration according to the result from offline TE may not be able to handle unexpected traffic dynamics within each resource provisioning cycle. To compensate for this inefficiency, online TE can be used to dynamically adjust traffic trunks according to the instant network condition obtained from real-time monitoring mechanisms. On the other hand, online TE should not completely discard the original configuration from offline TE, as significant traffic flapping and oscillation might be incurred, introducing network instability. In effect, a desired strategy to handle the relationship between offline and online TE is to allow offline TE to provide proper guidelines and restrictions to the online TE component so that dynamic routing adjustment can be applied in a controlled manner. A typical example is the TEQUILA [16] architecture, where the offline network dimensioning (ND) functional block provides directives and nonspecific "hard" values so as to leave space for unpredictable traffic fluctuations that will be handled by the dynamic route/resource management (DRtM, DRsM) functional blocks. In addition, design-based routing has been proposed in [116] to use offline TE results to guide online traffic routing. During the offline network provisioning phase, the INP may configure multiple routes toward a remote destination prefix, while BGP speakers can split traffic dynamically onto different next hop peers based on the advertised interdomain link bandwidth through eBGP [117].

Multiplane TE Interaction — Finally, if we regard intra-/interdomain TE interaction (including interdomain TE itself)

as a type of horizontal TE semantic between adjacent domains, the terminology of vertical TE can be borrowed as the concept of network resource optimization across multiple network planes within a domain (Fig. 12). Currently, there are two major scenarios of TE with multiple network planes: routing incongruence between different traffic types (e.g., IPv4/IPv6, unicast/multicast) and different QoS requirements (e.g., DiffServ TE). Recently, with the advent of multitopology-aware routing protocols such as MT-OSPF, M-ISIS, and the multiprotocol extension to BGP (MP-BGP [118]), together with DiffServ-MPLS-based solutions, vertical TE for multiple traffic types and QoS/TE requirements becomes a feasible option. However, even if these multiplane routing protocols offer high flexibility in path selection, TE in the management plane concerning overall resource optimization is still indispensable, as all types of traffic are mapped onto the same physical network infrastructure. In this case TE for individual network planes needs to be coordinated so as to achieve "vertical" optimization across all planes. Taking unicast/multicast TE as an example, the MT-IGP link weights can be assigned for unicast traffic and multicast traffic independently, aiming at different TE objectives (e.g., load balancing for unicast traffic and bandwidth conservation for multicast traffic). However, the calculation of link weights for the two planes should not proceed independently, as both unicast and multicast traffic are projected onto the same network resources. This means that the link weight setting for the two planes should concern overall TE optimization as well as the objectives in individual planes. It is also worth mentioning that multiplane routing protocols are not absolutely necessary for routing of different traffic types. In fact, all types of traffic can be routed through a single plane with conventional OSPF/ISIS and BGP. In this scenario configuration of the unique set of link weight and BGP path selection should include all TE objectives. Since multiplane routing protocols have not been widely deployed in the Internet, it would be interesting to investigate the relevant performance against the scalability in the routing information base (RIB) needed to store the routing information for multiple planes, compared to the conventional single plane routing semantics.

Traffic Engineering vs. Overlay Selfish Routing

In some circumstances there are conflicts between TE objectives and end-to-end QoS demands from individual customers in which TE cannot satisfy the QoS requirements. In this case overlay selfish routing is a flexible mechanism for end users to bypass TE constraints. A distinct characteristic of overlay routing is that path selection is performed by end hosts running applications according to their QoS requirements, and the underlying IP routing infrastructure is not aware of any overlay traffic.[1] In this sense overlay routing is also known as selfish routing, as it does not consider the optimization for any other traffic within the network [119]. As has been mentioned, TE aims at overall optimization of network performance by controlling traffic across the network. With the introduction of overlay routing, TE becomes less efficient because the routing of overlay traffic is outside the control of the INP. This problem has been identified recently, and several research papers have addressed the interaction between TE and overlay routing. In [119] the authors applied game theory to analyze the behavior of overlay routing and IP/MPLS-based TE, taking end-to-end delay as a typical QoS metric. The result of their work showed that through dedicated overlay routing, near-optimal traffic delay can be achieved provided that the network layer routing of other traffic is static. However, network congestion still occurs at some hot spots within the network, because the overall traffic distribution cannot be fully managed by TE. Furthermore, the performance of IP-based TE with overlay traffic coexistence was found to be very poor, while the situation can be improved using MPLS-based TE with explicit routing and uneven splitting functionality. Other research work, such as [120], also indicated the same conclusion based on both theoretical and experimental analysis. As a conclusion, the more traffic in the network that is outside the management scope of the INP, the poorer the TE performance results. This indicates that excessive overlay traffic brings significant negative impacts to effective TE.

[1] *This flexible functionality of overlay routing is very similar to MPLS explicit routing. The key difference is that overlay routing is always performed by end users for their own QoS benefits, while MPLS explicit routing is normally adopted by INPs for TE purposes.*

Summary

Today, Internet traffic engineering techniques have been largely confined to theoretical analysis, and most of them have not been applied in real operational networks. As far as intradomain traffic engineering is concerned, it has been receiving less and less attention due to the trend of bandwidth overprovisioning at core networks. Nevertheless, as network congestion may still occur due to the significant changes in traffic load distribution caused by network element failures and traffic spikes, making TE robust to failures and traffic demand uncertainty is a topic worthy of investigation even within a single domain. It is not difficult to imagine that this resilience issue also needs to be considered for supporting QoS (e.g., edge-to-edge delay), specifically, how to compute optimized backup paths in order to support the original QoS requirements while at the same time meeting the TE objectives.

Compared to its intradomain counterpart, interdomain TE is not yet fully understood, let alone ready for practical deployment. We argue that there are two interrelated issues that plague the practicality of interdomain TE. First of all, the original design of BGP did not consider how individual routing policies can be systematically used to optimize interdomain traffic. Although various BGP-based TE techniques have been proposed in recent years, routing issues of stability and divergence have generally been ignored, which are vital to consider in practice. It is worth mentioning that these are not problems from the individual TE schemes that have been proposed, but are inherently associated with the policy-based BGP routing infrastructure. It is still unknown how many of the existing interdomain TE approaches can be practically configurable when stability issues are taken into account. Nevertheless, recent BGP-based TE with stability consideration has already made efforts in this direction.

Another issue that hinders the practicality of interdomain TE is the noncooperative or even conflicting routing strategies of individual, especially adjacent, INPs. Strictly speaking, the concept of interdomain TE can be split into two categories: engineering interdomain (transit) traffic within one single domain and a more "genuine" interdomain TE that considers engineering traffic across multiple cooperative domains. Most interdomain TE proposals belong to the first category, as the TE objectives are purely for the benefit of the local domain without considering whether the local BGP routing decision will introduce negative impacts on its neighbors or even the global Internet. In effect, inconsistency in interdomain routing policies between neighboring domains may cause routing

anomalies such as the "bad gadget" effect [121]. Hence, a much more challenging task is to consider the second scenario, which requires cooperation between participating domains in order to achieve mutual benefits. Unfortunately, given the current situation where individual INPs are generally in competitive rather than cooperative relationships, it is hard to tell whether such a "genuine" interdomain TE, especially at the global Internet scale, is realistic or just Utopian. A new trend in interdomain TE is to use alternative routing mechanisms to avoid the inefficiency of the BGP routing paradigm, such as interdomain MPLS-based path computation services. By the time this article is written, IETF activities in this area are only in the architectural design phase, and there have been very few operational experiences thus far. Nevertheless, it is envisaged that PCE-based interdomain TE is able to provide promising solutions, not only for traffic optimizations across domains, but also for more advanced services such as end-to-end QoS and resilience. Apart from the problem of how to compute optimized interdomain paths, one important issue to be considered in deploying PCE-based TE infrastructure is scalability, mainly in how interdomain traffic can be efficiently aggregated in order to avoid deploying massive LSPs across multiple domains.

Finally, with the Internet becoming a multiservice platform, Internet TE needs to take into account different types of traffic (e.g., unicast vs. multicast, IPv4 vs. IPv6) as well as heterogeneous service requirements. Although the idea of using multitopology routing or combined routing techniques such as IP+MPLS may provide some potential solutions, the actual management of heterogeneous traffic on top of the common physical routing infrastructure in order to achieve both operational objectives and service requirements still needs further research effort, typically on the vertical interactions across different traffic types and traffic with heterogeneous service requirements. Again, how to achieve this goal in the interdomain environment in order to enable a genuine multiservice Internet is a much more challenging task for the future.

REFERENCES

[1] N. Hu *et al.*, "Locating Internet Bottlenecks: Algorithms, Measurements and Implications," *Proc. ACM SIGCOMM*, 2004, pp. 41–54.
[2] D. Awduche *et al.*, "Overview and Principles of Internet Traffic Engineering," IETF RFC 3272, May 2002.
[3] Y. Lee *et al.*, "Traffic Engineering in Next-Generation Optical Networks," *IEEE Commun. Surveys & Tutorials*, vol. 6, no. 3, 2004, pp. 16–33.
[4] G. Iannaccone *et al.*, "Analysis of Link Failures in an IP Backbone," *Proc. ACM IMW*, 2002, pp. 237–42.
[5] R. Teixeira *et al.*, "Network Sensitivity to Hot-Potato Disruptions," *Proc. ACM SIGCOMM*, 2004, pp. 231–44.
[6] D. Awduche et al, "Requirements on Traffic Engineering over MPLS," RFC 2702, June 1999.
[7] D. Awduche *et al.*, "MPLS and Traffic Engineering in IP Networks," *IEEE Commun. Mag.*, vol. 37, no. 12, Dec. 1999, pp. 42–47.
[8] B. Fortz *et al.*, "Internet Traffic Engineering by Optimising OSPF Weights," *Proc. IEEE INFOCOM*, 2000, pp. 519–28.
[9] B. Fortz *et al.*, "Optimizing OSPF/IS-IS Weights in a Changing World," *IEEE JSAC*, vol. 20, no. 4, May 2000, pp. 756–67.
[10] B. Fortz *et al.*, "Traffic Engineering with Traditional IP Routing Protocols," *IEEE Commun. Mag.*, vol. 40, no. 10, Oct. 2002, pp. 118–24.
[11] B. Quoitin *et al.*, "Interdomain Traffic Engineering with BGP," *IEEE Commun. Mag.*, vol. 41, no. 5, May 2003, pp. 122–28.
[12] B. Fortz *et al.*, "Robust Optimization of OSPF/IS-IS Weights," *Proc. INOC 2003*, pp. 225–30.
[13] R. Teixeira *et al.*, "Dynamics of Hot-Potato Routing in IP Networks," *Proc. ACM SIGMETRICS 2004*, pp. 307–19.
[14] Cisco IOS Netflow, http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html
[15] A. Asgari *et al.*, "Scalable Monitoring Support for Resource Management and Service Assurance," *IEEE Network*, vol. 18, no. 6, Nov./Dec. 2004, pp. 6–18.
[16] P. Trimintzios *et al.*, "A Management and Contral Architecture for Providing IP Differentiated Services in MPLS-Based Networks," *IEEE Commun. Mag.*, vol. 39, no. 5, May 2001, pp. 80–88.
[17] Multicast Deployment Status, http://multicast.internet2.edu/wg-multicast-status.shtml
[18] L. Andersson *et al.*, "LDP Specification," IETF RFC 3036, Jan. 2001.
[19] D. Mitra and K. G. Ramakrishnan, "A Case Study of Multiservice, Multipriority Traffic Engineering Design for Data Networks," *Proc. IEEE GLOBECOM*, 1999, pp. 1077–83.
[20] O. Younis *et al*, "Constraint-Based Routing in the Internet: Basic Principles and Recent Research," *IEEE Commun. Surveys & Tutorials*, 3rd qtr., 2003, pp. 2–13.
[21] X. Xiao *et al.*, "Traffic Engineering with MPLS in the Internet," *IEEE Network*, vol. 14, no. 12, Mar./Apr. 2000, pp. 28–33.
[22] Z. Wang *et al.*, "Quality of Service Routing for Supporting Multimedia Applications," *IEEE JSAC*, vol. 14, no. 7, Sept. 1996, pp. 1228–34.
[23] R. Guerin, *et al.*, "QoS Routing Mechanisms and OSPF Extensions," *Proc. IEEE GLOBECOM* 1997, pp. 1903–08.
[24] Y. Wang *et al.*, "Explicit Routing Algorithms for Internet Traffic Engineering," *Proc. IEEE ICCCN*, 1999, pp. 582–88.
[25] F. Le Faucheur *et al.*, "Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering," IETF RFC 3564, July 2003.
[26] P. Trimintzios *et al.*, "Quality of Service Provisioning through Traffic Engineering with Applicability to IP Based Production Networks," *Comp. Commun.*, vol. 26, no. 8, May 2003, pp. 845–60.
[27] V. Tabatabaee *et al.*, "Differentiated Traffic Engineering for QoS Provisioning," *Proc. IEEE INFOCOM*, 2005, pp. 2349–59.
[28] H. Saito *et al.*, "Traffic Engineering Using Multiple Multipoint-to-point LSPs," *Proc. IEEE INFOCOM*, 2000, pp. 894–901.
[29] G. Urvoy-Keller *et al.*, "Traffic Engineering in a Multipoint-to-Point Network," *IEEE JSAC*, vol. 20, no. 4, May 2002, pp. 834–49.
[30] S. Bhatnagar *et al.*, "Creating Multipoint-to-Point LSPs for Traffic Engineering," *IEEE Commun. Mag.*, vol. 43, no. 1, Jan. 2005, pp. 95–100.
[31] P. Trimintzios *et al.*, "Engineering the Multi-Service Internet: MPLS and IP-Based Techniques," *Proc. IEEE ICT*, 2001, pp. 129–34.
[32] A. Elwalid *et al.*, "MATE: MPLS Adaptive Traffic Engineering," *Proc. IEEE INFOCOM*, 2001, pp. 1300–09.
[33] S. Kandula *et al.*, "Walking the Tightrope: Responsive Yet Stable Traffic Engineering," *ACM SGICOMM Comp. Commun. Review*, vol. 35, no. 4, Oct. 2005, pp. 253–64.
[34] R. Boutaba *et al.*, "DORA: Efficient Routing for MPLS Traffic Engineering," *J. Network and Sys. Mgmt.*, vol. 10, no. 3, Sept. 2002, pp. 309–25.
[35] K. Kar *et al.*, "Minimum Interference Routing of Bandwidth Guaranteed Tunnels with MPLS Traffic Engineering Applications," *IEEE JSAC*, vol. 18, no. 12, Dec. 2000, pp. 2566–79.
[36] K. Kodialam *et al.*, "Minimum Interference Routing of Applications to MPLS Traffic Engineering," *Proc. IEEE INFOCOM*, 2000, pp. 884–93.
[37] P. Aukia *et al.*, "RATES: A Server for MPLS Traffic Engineering," *IEEE Network*, vol. 14, no. 2, Mar./Apr. 2000, pp. 34–41.
[38] F. Blanchy, L. Melon, and G. Leduc, "A Preemption-Aware On-line Routing Algorithm for MPLS Networks," *Telecommun. Sys.*, vol. 24, no. 2–4, Oct. 2003, pp. 187–206.
[39] C. Scoglio *et al.*, "TEAM: A Traffic Engineering Automated Manager for DiffServ Based MPLS Networks," *IEEE Commun. Mag.*, vol. 42, no. 10, Oct. 2004, pp. 134–45.
[40] J. C. de Oliveira *et al.*, "SPeCRA: A Stochastic Performance Comparison Routing Algorithm for LSP setup in MPLS Networks," *Proc. IEEE GLOBECOM*, 2002, pp. 2190–94.
[41] J. C. de Oliveira *et al.*, "New Preemption Policies for DiffServ Aware Traffic Engineering to Minimize Rerouting in MPLS Networks," *IEEE/ACM Trans. Networking*, vol. 12, no. 4, Aug. 2004, pp. 733–45.
[42] K. Walkowiak, "Survivable Online Routing for MPLS Traffic Engineering," *Proc. QoFIS*, 2004, pp. 288–97.
[43] Y. Wang *et al.*, "Internet Traffic Engineering without Full Mesh Overlaying," *Proc. IEEE INFOCOM*, 2001, pp. 565–71.

[44] G. Retvari *et al.*, "On the Representability of Arbitrary Path Sets as Shortest Paths: Theory, Algorithms and Complexity," *Proc. IFIP NETWORKING*, 2004, pp. 1180–91.
[45] A. Feldmann *et al.*, "NetScope: Traffic Engineering for IP Networks," *IEEE Network*, vol. 14, no. 2, Mar./Apr. 2000, pp. 11–19.
[46] M. Ericsson *et al.*, "A Genetic Algorithm for the Weight Setting Problem in OSPF Routing," *J. Combinatorial Optimization*, vol. 6, no. 3, Sept. 2002, pp. 299–333.
[47] G. Retvari *et al.*, "Practical OSPF Traffic Engineering," *IEEE Commun. Letters*, vol. 8, no. 11, Nov. 2004, pp.689–91.
[48] A. Sridharan *et al.*, "Achieving Near-Optimal Traffic Engineering Solutions for Current OSPF/IS-IS Networks," *IEEE/ACM Trans. Networking*, vol. 13, no. 2, Apr. 2005, pp. 234–47.
[49] J. Wang *et al.*, "Edge Based Traffic Engineering for OSPF Networks," *Comp. Networks*, vol. 48, no. 4, July 2005, pp. 605–25.
[50] C. Labovitz *et al.*, "Internet Routing Instability," *IEEE/ACM Trans. Networking*, vol. 6, no. 5, Oct. 1998, pp. 515–28.
[51] I. Gojmerac, T. Ziegler, F. Ricciato and P. Reichl, "Adaptive Multipath Routing for Dynamic Traffic Engineering," *Proc. IEEE GLOBECOM*, 2003, pp. 3058–62.
[52] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," IETF RFC 1771, Mar. 1995.
[53] Cisco Systems, "BGP Multipath Load Sharing for Both eBGP and iBGP in an MPLS-VPN," 2005.
[54] L. Gao and J. Rexford, "Stable Internet Routing without Global Coordination," *IEEE/ACM Trans. Networking*, vol. 9, no. 6, Dec. 2001, pp.681–92.
[55] Y. R. Yang *et al.*, "On Route Selection for Interdomain Traffic Engineering," *IEEE Network*, vol. 19, no. 6, Nov./Dec. 2005, pp 20–27.
[56] N. Feamster *et al.*, "Guidelines for Interdomain Traffic Engineering," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 33, no. 5, Oct. 2003, pp. 19–30.
[57] A. Farrel *et al.*, "A Path Computation Element (PCE)-Based Architecture," IETF RFC 4655, Aug. 2006.
[58] D. Goldenberg *et al.*, "Optimizing Cost and Performance for Multihoming," *Proc. ACM SIGCOMM* 2004, pp. 79–92.
[59] S. Uhlig *et al.*, "Interdomain Traffic Engineering with Minimal BGP Configurations," *Proc. 18th Int'l. Teletraffic Cong.,* 2003.
[60] S. Uhlig, "A Multiple-Objectives Evolutionary Perspective to Interdomain Traffic Engineering in the Internet," *Int'l. J. Computational Intelligence and Apps.*, vol. 5, no. 2, June 2005, pp. 215–30.
[61] H. Wang, "Optimal ISP Subscription for Internet Multihoming: Algorithm Design and Implication Analysis," *Proc. IEEE INFOCOM*, 2005, pp. 2360–71.
[62] K. Ho *et al.*, "An Incentive-based Quality of Service Aware Algorithm for Offline Inter-AS Traffic Engineering," *Proc. IEEE IPOM*, 2004, pp. 34–40.
[63] T.C. Bressoud *et al.*, "Optimal Configuration for BGP Route Selection," *Proc. IEEE INFOCOM* 2003, pp. 916–26.
[64] S. Uhlig and B. Quoitin, "Tweak-it: BGP-Based Interdomain Traffic Engineering for Transit ASs," *Proc. Next Gen. Internet Networks*, 2005, pp. 75–82.
[65] K. Ho *et al.*, "Multi-Objective Egress Router Selection Policies for Inter-domain Traffic with Bandwidth Guarantees," *Proc. IFIP Networking*, 2004, pp. 271–83.
[66] M. Howarth *et al.*, "Provisioning for Inter-domain Quality of Service: the MESCAL Approach," *IEEE Commun. Mag.*, vol. 43, no. 6, June 2005, pp. 129–37.
[67] M. Howarth *et al.*, "End-to-end Quality of Service Provisioning Through Inter-provider Traffic Engineering," *Comp. Commun.*, vol. 29, no. 6, Mar. 2006, pp. 683–02.
[68] M.R. Gary and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, 1979.
[69] S. Uhlig and O. Bonaventure, "Designing BGP-based Outbound Traffic Engineering Techniques for Stub ASs," *ACM SIGCOMM Comp. Commun.* Rev., vol. 34, no. 5, Oct. 2004, pp. 89–106.
[70] A. Akella *et al.*, "A Measurement-Based Analysis of Multihoming," *Proc. ACM SIGCOMM* 2003, pp. 353–64.
[71] S. Lee *et al.*, "Exploiting AS Hierarchy for Scalable Route Selection in Multi-Homed Stub Networks," *Proc. ACM IMC*, 2004, pp. 294–99.
[72] R.K.C. Chang and M. Lo, "Inbound Traffic Engineering for Multihomed ASs Using AS Path Prepending," *IEEE Network*, vol. 19, no. 2, Mar./Apr. 2005, pp. 18–25.
[73] H. Wang *et al.*, "Characterizing the Performance and stability Issues of the AS Path Prepending Method: Taxonomy, Measurement Study and Analysis," *Proc. ACM SIGCOMM Asia Wksp.*, 2005.
[74] B. Quoitin *et al.*, "A Performance Evaluation of BGP-based Traffic Engineering," *Int'l. J. Network Mgmt.*, vol. 15, no. 3, May/June 2005, pp. 177–91.
[75] S.R. Sangli *et al.*, "BGP Extended Communities Attribute," Internet draft, draft-ietf-idr-bgp-ext-communities-08.txt, Feb. 2005.
[76] B. Quoitin *et al.*, "Interdomain Traffic Engineering with Redistributed Communities," *Comp. Commun.*, vol. 27, no. 4, Oct. 2003, pp. 355–63.
[77] S. Kalyanaraman, "Load Balancing in BGP Environments using Online Simulation and Dynamic NAT," presented at the Internet Statistic and Metrics Analysis Wksp. 2001.
[78] A. Akella *et al.*, "Multihoming Performance Benefits: An Experimental Evaluation of Practical Enterprise Strategies," *Proc. USENIX Annual Tech. Conf.*, 2004.
[79] S. Agarwal *et al.*, "OPCA: Robust Interdomain Policy Routing and Traffic Control," *Proc. IEEE OPENARCH*, 2003, pp. 55–64.
[80] R. Gao *et al.*, "Interdomain Ingress Traffic Engineering through Optimized AS-Path Prepending," *Proc. IFIP Networking*, 2005, pp. 647–58.
[81] R. Mahajan *et al.*, "Towards Coordinated Interdomain Traffic Engineering," *Proc. ACM HotNets-III Wksp.*, 2004.
[82] R. Mahajan *et al.*, "Negotiation-based Routing Between Neighboring Domains," *Proc. ACM/USENIX Networked Sys. Design and Implementation*, 2005.
[83] D. Awduche *et al.*, "An Approach to Optimal Peering Between Autonomous Systems in the Internet," *Proc. IEEE ICCCN*, 1998, pp. 346–51.
[84] R. Johari and J.N. Tsitsiklis, "Routing and Peering in a Competitive Internet," tech. rep. P-2570, MIT Lab. for Info. and Decision Sys., Jan. 2003.
[85] G. Shrimali et al, "Cooperative Interdomain Traffic Engineering using Nash Bargaining and Decomposition," *Proc. IEEE INFOCOM*, 2007.
[86] B. Quoitin and O. Bonaventure, "A Cooperative Approach to Inter-domain Traffic Engineering," *Proc. NGI Networks*, 2005, pp. 450–57.
[87] Y. Liu and N. Reddy, "Multihoming Route Control among a Group of Multihomed Stub Networks," *Comp. Commun.*, vol. 30, no. 17, 2007, pp. 3335–45.
[88] M. Kodialam *et al.*, "Online Multicast Routing with Bandwidth Guarantees: A New Approach Using Multicast Network Flow," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, Aug. 2003, pp. 676–86.
[89] A. Fei at al, "Aggregated Multicast with Inter-Group Tree Sharing," *Proc. Int'l. Wksp. Networked Group Commun.*, 2001, pp. 172–88.
[90] B. Yang *et al.*, "Multicasting in MPLS Domains," *Comp. Commun.*, vol. 27, no. 2, Feb. 2004, pp. 162–70.
[91] B. Fenner *et al.*, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)," IETF RFC 4601, Aug. 2006.
[92] Y. D. Meisel *et al.*, "Multicast Routing with Traffic Engineering: a Multi-Objective Optimization Scheme and a Polynomial Shortest Path Tree Algorithm with Load Balancing," *Proc. CCIO*, 2004.
[93] N. Wang *et al.*, "Traffic Engineered Multicast Content Delivery without MPLS Overlay," *IEEE Trans. Multimedia*, vol. 9, no. 3, Apr. 2007, pp. 619–28.
[94] T. Przygienda *et al.*, "M-ISIS: Multi Topology (MT) Routing in IS-IS," RFC 5120, Feb. 2008.
[95] P. Psenak *et al.*, "Multi-Topology (MT) Routing in OSPF," RFC 4915, June 2007.
[96] A. Nucci *et al.*, "IGP Link Weight Assignment for Operational Tier-1 Backbones," *IEEE/ACM Trans. Networking*, vol. 15, no. 1, pp. 789–802.c
[97] A. Sridharan *et al.*, "Making IGP Routing Robust to Link Failures," *Proc. IFIP Networking*, 2005, pp. 634–46.
[98] D. Yuan, "A Bi-Criteria Optimization Approach for Robust OSPF Routing," *Proc. IEEE IPOM*, 2003, pp. 91–97.
[99] A. Kvalbein *et al.*, "Post Failure Routing Performance with Multiple Routing Configurations," *Proc. IEEE INFOCOM*, 2007.

[100] E. Karasan *et al.*, "Robust Path Design Algorithms for Traffic Engineering with Restoration in MPLS Networks," *IEICE Trans. Commun.*, vol. E86-b, no. 5, pp 1632–40.
[101] M. Amin *et al.*, "Making Outbound Route Selection Robust to Egress Point Failure," *Proc. IFIP Networking*, 2006, pp. 233–46.
[102] D. Applegate and E. Cohen, "Making Intra-Domain Routing Robust to Changing and Uncertain Traffic Demands: Understand Fundamental Tradeoffs," *Proc. ACM SIGCOMM*, 2003, pp. 313–24.
[103] D. Applegate *et al.*, "Coping with Network Failures: Routing Strategies for Optimal Demand Oblivious Restoration," *Proc. ACM SIGMETRICS*, 2004, pp. 270–81.
[104] C. Zhang *et al.*, "On Optimal Routing with Multiple Traffic Matrices," *Proc. IEEE INFOCOM*, 2005, pp. 607–18.
[105] D. Mitra and Q. Wang, "Stochastic Traffic Engineering for Demand Uncertainty and Risk-Aware Network Revenue Management," *IEEE/ACM Trans. Networking*, vol. 13, no. 2, Apr. 2005, pp. 221–33.
[106] H. Wang et al, "COPE: Traffic Engineering in Dynamic Networks," *Proc. ACM SIGCOMM*, 2006.
[107] K.H. Ho *et al.*, "A Robustness Approach to Inter-AS Outbound Traffic Engineering," *Proc. IEEE ICC*, 2006, pp. 560–65.
[108] R. Zhang and N. Mckeown, "Designing a Predictable Internet Backbone Network," *Proc. ACM HotNets*, 2004.
[109] M. Kodialam, T.V. Lakshmann and S. Sengupta, "Efficient and Robust Routing of Highly Variable Traffic," *Proc. ACM HotNets*, 2004.
[110] S. Agarwal *et al.*, "The Impact of BGP Dynamics on Intra-domain Traffic," *ACM SIGMETRICS Perf. Eval. Rev.*, vol. 32, no. 1, June 2004, pp. 319–30.
[111] K. H. Ho *et al.*, "Joint Optimization of Intra- and Inter-Autonomous System Traffic Engineering," *Proc. IEEE/IFIP NOMS*, 2006, pp. 248–59.
[112] J. Boyle *et al.*, "Applicability Statement for Traffic Engineering with MPLS," IETF RFC 3346, Aug. 2002.
[113] H. Pham *et al.*, "Hybrid Routing for Scalable IP/MPLS Traffic Engineering," *Proc. IEEE ICC*, 2003, pp. 332–337.
[114] A. Bagula, "Hybrid Routing in Next Generation IP Networks," *Comp. Commun.*, vol. 29, no. 7, Apr. 2006, pp. 879–92.
[115] A. Bagula, "Hybrid IGP + MPLS Routing in Next Generation IP Networks: An Online Traffic Engineering Model," *Proc. QoSIP*, 2005, pp. 325–38.
[116] A. Elwalid, "Routing and Protection in GMPLS Networks: From Shortest Paths to Optimized Designs," *J. Lightwave Tech.*, vol. 21, no. 11, Nov. 2003, pp. 2828–38.
[117] Cisco white paper, "BGP Bandwidth Link," http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122newft/122t/122t2/ftbgplb.htm
[118] T. Bates *et al.*, "Multiprotocol Extensions for BGP-4," IETF RFC 4760, Jan. 2007.
[119] L. Qiu *et al.*, "On Selfish Routing in Internet-like Environments," *IEEE/ACM Trans. Networking*, vol. 14, no. 4, Aug. 2006, pp. 725–38.
[120] Y. Liu *et al.*, "On the Interaction between Overlay Routing and Traffic Engineering," *Proc. IEEE INFOCOM*, 2005, pp. 2543–53.
[121] T. Griffin *et al.*, "An Analysis of BGP Convergence Properties," *Proc. ACM SIGCOMM*, 1999, pp. 277–88.

BIOGRAPHIES

NING WANG [M'01] (n.wang@surrey.ac.uk) received a B.Eng. degree (Honors) in computing from Changchun University of Science and Technology, China, in 1996, an M.Eng. degree in electronic engineering from Nanyang Technological University, Singapore, in 2000, and a Ph.D. degree in electronic engineering from the University of Surrey, Guildford, United Kingdom, in 2004. He is currently a postdoctoral research fellow in the Center for Communication Systems Research, University of Surrey. His major research interests include traffic engineering and network optimization algorithms, Internet QoS provisioning, multicast communication, and overlay networks.

KIN-HON HO (k.ho@surrey.ac.uk) is a research fellow at the Center for Communication Systems Research, University of Surrey, United Kingdom. He received his B.Sc. (Honors) in computer studies from the City University of Hong Kong, his M.Sc. (Eng) in data communications from the University of Sheffield, United Kingdom, and his Ph.D. in electronic engineering from the University of Surrey. His research interests include Internet traffic engineering, QoS, and network planning and optimization.

GEORGE PAVLOU [M'95] (g.pavlau@surrey.ac.uk) received a Diploma degree in electrical and mechanical engineering from the National Technical University of Athens, Greece, and M.Sc. and Ph.D. degrees in computer science from University College London, United Kingdom. He is currently a professor of communication and information systems at the Centre of Communication Systems Research (CCSR), Department of Electronic Engineering, University of Surrey, where he leads the activities of the Networks Research Group. He was previously a senior research fellow and lecturer in the Department of Computer Science, University College London, where he led research activities on network and service management. His research interests focus on network management, networking, and service engineering. He has been a Chartered Engineer and member of the Technical Chamber of Greece since 1984. He is on the Editorial Board of *IEEE Transactions on Network and Service Management*, *IEEE Communication Surveys and Tutorials*, and the *Journal of Network and Systems Management*. He is Network and Service Management Series Editor of *IEEE Communications Magazine*.

MICHAEL HOWARTH (m.howarth@surrey.ac.uk) is a lecturer in networking at CCSR, University of Surrey. He holds a Bachelor's degree in engineering science and a D.Phil. in electrical engineering, both from Oxford University, and an M.Sc. in telecommunications from the University of Surrey. Prior to joining Surrey he worked for several networking and IT consultancies. His research interests include traffic engineering, QoS, security systems, protocol design, and optimization of satellite communications. He is a Chartered Electrical Engineer and member of the U.K. IET.