

Link-State Routing With Hop-by-Hop Forwarding Can Achieve Optimal Traffic Engineering

Dahai Xu, *Member, IEEE*, Mung Chiang, *Senior Member, IEEE*, and Jennifer Rexford, *Senior Member, IEEE, Fellow, ACM*

Abstract—This paper settles an open question with a positive answer: Optimal traffic engineering (or optimal multicommodity flow) can be realized using just link-state routing protocols with hop-by-hop forwarding. Today's typical versions of these protocols, Open Shortest Path First (OSPF) and Intermediate System-Intermediate System (IS-IS), split traffic evenly over shortest paths based on link weights. However, optimizing the link weights for OSPF/IS-IS to the offered traffic is a well-known NP-hard problem, and even the best setting of the weights can deviate significantly from an optimal distribution of the traffic. In this paper, we propose a new link-state routing protocol, PEFT, that splits traffic over multiple paths with an exponential penalty on longer paths. Unlike its predecessor, DEFT, our new protocol provably achieves optimal traffic engineering while retaining the simplicity of hop-by-hop forwarding. The new protocol also leads to a significant reduction in the time needed to compute the best link weights. Both the protocol and the computational methods are developed in a conceptual framework, called Network Entropy Maximization, that is used to identify the traffic distributions that are not only optimal, but also realizable by link-state routing.

Index Terms—Interior gateway protocol, network entropy maximization, optimization, Open Shortest Path First (OSPF), routing, traffic engineering.

I. INTRODUCTION

DESIGNING a link-state routing protocol has three components. First is *weight computation*: The network-management system computes a set of link weights through a periodic and centralized optimization. The second is *traffic splitting*: Each router uses the link weights to decide traffic-splitting ratios among its outgoing links for every destination. The third is *packet forwarding*: Each router independently decides which outgoing link to forward a packet

Manuscript received April 12, 2010; revised January 04, 2011; accepted March 19, 2011; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor P. Van Mieghem. Date of publication April 07, 2011; date of current version December 16, 2011. This work was supported in part by DARPA W911NF-07-1-0057, ONR YIP N00014-07-1-0864, AFOSR FA9550-06-1-0297, and NSF CNS-0519880 and CNS 0720570. A preliminary short version of this paper was presented under the same title in the Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Phoenix, AZ, April 13–19, 2008.

D. Xu was with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA. He is now with AT&T Laboratories—Research, Florham Park, NJ 07932 USA (e-mail: dahaixu@research.att.com).

M. Chiang is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: chiangm@princeton.edu).

J. Rexford is with the Department of Computer Science, Princeton University, Princeton, NJ 08544 USA (e-mail: jrex@cs.princeton.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNET.2011.2134866

based only on its destination prefix in order to realize the desired traffic splitting. The popularity of link-state protocols can be attributed to their ease of management. In particular, each router's traffic-splitting decision is made autonomously based only on the link weights, without further assistance from the network-management system, and each packet's forwarding decision is made in a hop-by-hop fashion without end-to-end tunneling.

Such simplicity was thought to come at the expense of optimality. In a procedure known as traffic engineering (TE), network operators minimize a convex cost function of the link loads by tuning the link weights used by the routers. With Open Shortest Path First (OSPF) or Intermediate System-Intermediate System (IS-IS), the major variants of link-state protocols in use today, computing the right link weights is NP-hard, and even the best setting of the weights can deviate significantly from optimal TE [2], [32]. The following question remains open: Can a link-state protocol with hop-by-hop forwarding achieve optimal TE? This paper shows that the answer is in fact positive by developing a new link-state protocol, Penalizing Exponential Flow-splitting (PEFT), proving that it achieves optimal TE and demonstrating that link-weight computation for PEFT is highly efficient in theory and in practice.

In PEFT, packet forwarding is just the same as OSPF: destination-based and hop-by-hop. The key difference is in traffic splitting. OSPF splits traffic *evenly* among the shortest paths, and PEFT splits traffic along all paths, but penalizes longer paths (i.e., paths with larger sums of link weights) exponentially. While this is a difference in how link weights are *used* in the routers, it also mandates a change in how link weights are *computed* by the operator. It turns out that using link weights in the PEFT way enables optimal traffic engineering. Using the Abilene topology and traffic traces, we observe a 15% increase in the efficiency of capacity utilization by PEFT over OSPF. Furthermore, an exponential traffic-splitting penalty is the *only* penalty that can lead to this optimality result. The corresponding best link weights for PEFT can be efficiently computed: as efficiently as solving a linearly constrained concave maximization and much faster than the existing weight computation heuristics for OSPF.

Clearly, if the complexity of managing a routing protocol were not a concern, other approaches could be used to achieve optimal TE. One possibility is multicommodity-flow type of routing, where an optimal traffic distribution is realized by dividing an arbitrary fraction of traffic over many paths. This can be supported by the forwarding mechanism in multiprotocol label switching (MPLS) [3]. However, optimality then comes

TABLE I
COMPARISON OF VARIOUS TE SCHEMES (NEW CONTRIBUTIONS IN ITALICS)

	Commodity Routing	Link-State Routing	
		OSPF	PEFT
Traffic Splitting	Arbitrary	Even among shortest paths	Exponential
Scalability	Low	High	High
Optimal TE	Yes	No	<i>Yes</i>
Complexity Class	Convex Optimization	NP Hard	<i>Convex Optimization</i>

with a cost for establishing many end-to-end *tunnels* to forward packets. Second, other studies explored more flexible ways to split traffic over shortest paths [4]–[6], but these solutions do not enable routers to *independently* compute the flow-splitting ratios from the link weights. Instead, a central management system must compute and configure the traffic-splitting ratios and update them when the topology changes, sacrificing the main benefit of running a distributed link-state routing protocol in the first place. Clearly, there is a tension between optimal but complex routing or forwarding methods and the simple but to-date suboptimal link-state routing with hop-by-hop forwarding. Recent works [1], [7] attempted to attain optimality and simplicity simultaneously, but in contrast to this paper, they neither proved optimality for TE nor developed sufficiently fast methods for computing link weights. A summary is provided in Table I.

There are several new ideas in this paper that enable a proof of optimality and a much faster computation beyond, for example, the theory and algorithm in our own earlier Distributed Exponentially-weighted Flow splitTing (DEFT) [1] work. One of these ideas is to develop the traffic-splitting and weight-computation methods from the conceptual framework of network entropy maximization (NEM). As a proof technique and intermediate step of protocol development, we will construct an NEM optimization problem that is solved neither by the operator nor by the routers, but by us, the protocol developers. The optimality condition of NEM reveals the structure of hop-by-hop forwarding and is later used to guide both the router’s traffic splitting and the operator’s weight computation. In short, it turns out that a certain notion of entropy can precisely identify those optimal traffic distributions that can be realized by link-state protocols.

The general principle of entropy maximization has been used to solve other networking problems, e.g., [8]–[11]. This is the first work connecting entropy with IP routing. As we summarize later in Table V, our NEM framework for routing is different from and has interesting parallels to the recent work relating TCP congestion control to network utility maximization (NUM) [12]–[15]. Our work is not on solving the multicommodity flow problem approximately with distributed methods, such as [16] and [17].

The rest of this paper is organized as follows. Background on optimal traffic engineering is introduced in Section II. The theory of network entropy maximization in Section III leads to the routing protocol PEFT in Section IV and the associated link-weight computation algorithm in Section V. Extensive numerical experiments are then summarized in Section VI. The interesting and general framework of network entropy maximization is further discussed in Section VII. We conclude with further observations and extensions in Section VIII. In the

TABLE II
SUMMARY OF KEY NOTATION

Notation	Meaning
$D(s, t)$	Traffic demand from source s to destination t
$c_{u,v}$	Capacity of link (u, v)
$f_{u,v}$	Flow on link (u, v)
$\tilde{c}_{u,v}$	Necessary capacity of link (u, v)
$f_{u,v}^t$	Flow on link (u, v) destined to node t
f_u^t	Total incoming flow (destined to t) at u
$w_{u,v}$	Weight assigned to link (u, v)
w_{min}	Lower bound of all link weights
d_u^t	The shortest distance from node u to node t . $d_u^t = 0$
$h_{u,v}^t$	Gap of shortest distance, $h_{u,v}^t \triangleq d_v^t + w_{u,v} - d_u^t$
$\Gamma(h_{u,v}^t)$	Traffic splitting function

Appendix, we present more details about NEM and PEFT, as well as the key difference between PEFT and its predecessor, DEFT. The key notation used in this paper is shown in Table II.

II. BACKGROUND ON OPTIMAL TE

A. Definitions of Optimality

Consider a wireline network as a directed graph $G = (\mathbb{V}, \mathbb{E})$, where \mathbb{V} is the set of nodes (where $N = |\mathbb{V}|$), \mathbb{E} is the set of links (where $E = |\mathbb{E}|$), and link (u, v) has capacity $c_{u,v}$. The offered traffic is represented by a traffic matrix $D(s, t)$ for source–destination pairs indexed by (s, t) .

The load $f_{u,v}$ on each link (u, v) depends on how the network decides to route the traffic. An objective function enables quantitative comparisons between different routing solutions in terms of the load on the links. Traffic engineering usually considers a link-cost function $\Phi(f_{u,v}, c_{u,v})$ that is an increasing function of $f_{u,v}$.

For example, $\Phi(f_{u,v}, c_{u,v})$ can be the link utilization $f_{u,v}/c_{u,v}$, and the objective of traffic engineering can be to minimize $\max_{(u,v) \in \mathbb{E}} \Phi(f_{u,v}, c_{u,v})$.

As another example, let $\Phi(f_{u,v}, c_{u,v})$ be a piecewise-linear approximation of the M/M/1 delay formula [18], e.g.,

$$\Phi(f_{u,v}, c_{u,v}) = \begin{cases} f_{u,v}, & f_{u,v}/c_{u,v} \leq 1/3 \\ 3f_{u,v} - 2/3c_{u,v}, & 1/3 \leq f_{u,v}/c_{u,v} \leq 2/3 \\ 10f_{u,v} - 16/3c_{u,v}, & 2/3 \leq f_{u,v}/c_{u,v} \leq 9/10 \\ 70f_{u,v} - 178/3c_{u,v}, & 9/10 \leq f_{u,v}/c_{u,v} \leq 1 \\ 500f_{u,v} - 1468/3c_{u,v}, & 1 \leq f_{u,v}/c_{u,v} \leq 11/10 \\ 5000f_{u,v} - 16318/3c_{u,v}, & 11/10 \leq f_{u,v}/c_{u,v} \end{cases} \quad (1)$$

and the objective is to minimize $\sum_{(u,v)} \Phi(f_{u,v}, c_{u,v})$.

More generally, we use “ $\Phi(\{f_{u,v}, c_{u,v}\})$ ” to represent any increasing and convex objective function. The optimality of traffic engineering is with respect to this objective function.

At this point, we can already observe that there is a “gap” between the objective of TE and the mechanism of link-state routing. Optimality is defined directly in terms of the traffic flows, whereas link-state protocols represent the paths *indirectly* in terms of link weights. Bridging this gap is one of the challenges that have prevented researchers from achieving optimal traffic engineering using link-state routing thus far.

B. Optimal TE Via Multicommodity Flow

Consider the following convex optimization problem: minimizing the TE cost function over flow conservation and link capacity constraints.

COMMODITY:

$$\min \Phi(\{f_{u,v}, c_{u,v}\}) \quad (2a)$$

$$\text{s.t.} \quad \sum_{v:(s,v) \in \mathbb{E}} f_{s,v}^t - \sum_{u:(u,s) \in \mathbb{E}} f_{u,s}^t = D(s,t) \quad \forall s \neq t \quad (2b)$$

$$f_{u,v} \triangleq \sum_{t \in \mathbb{V}} f_{u,v}^t \leq c_{u,v} \quad \forall (u,v) \quad (2c)$$

$$\text{vars.} \quad f_{u,v}^t, f_{u,v} \geq 0. \quad (2d)$$

This multicommodity problem¹ can be readily solved efficiently, where the flow destined to a single destination is treated as a commodity, and $f_{u,v}^t$ is the amount of flow on link (u,v) destined to node t .²

The resulting solution, however, may not be realizable through link-state routing and hop-by-hop forwarding. Indeed, for a network with N nodes and E links, the multicommodity-flow solution may require up to $O(N^2E)$ tunnels, i.e., explicit routing (see Appendix-E), making it difficult to scale. In contrast, link-state routing is much simpler, requiring only $O(E)$ parameters (i.e., one per link).

Furthermore, while it is true that, from the solution of the COMMODITY problem, a set of link weights can be computed such that all the commodity flow will be forwarded along the shortest paths [4], [5], the flow-splitting ratios among these shortest paths are *not* related to the *link weights*, forcing the operator to specify up to $O(NE)$ *additional* parameters (one parameter on each link for each destination) as the flow-splitting ratios for all the routers.

Henceforth, we use the following phrases: optimal traffic engineering, optimal multicommodity flow (2) and optimal distribution of traffic, interchangeably. We formally define the problem addressed in this paper.

Optimal Traffic Engineering With Link-State Routing: In a network $G = (\mathbb{V}, \mathbb{E})$ using a link-state routing protocol with destination-based hop-by-hop forwarding, each router is aware of the weight of each link. Based on the E link weights, each router independently computes the flow-splitting ratios across its outgoing links. Is there such a protocol, with efficient computation of the link weights, that can achieve the optimal distribution of traffic as defined in (2)?

The rest of this paper shows that optimal traffic engineering can, in fact, be achieved using only E link weights.

¹We first remark that solving this COMMODITY problem is only an intermediate step in the proof. The actual PEFT protocol in Section IV will not be implementing a multicommodity-flow-based routing with end-to-end tunneling. Another clarifying remark is that while we will later show that PEFT link-weight computation is as easy as solving a convex optimization. However, that optimization is not this well-known COMMODITY problem.

²If the objective $\Phi(\{f_{u,v}, c_{u,v}\})$ is not a strictly increasing function of link flow $f_{u,v}$ (like minimizing the maximum link utilization), the optimal solution of COMMODITY problem (2) may contain flow cycles. To prevent bandwidth waste, we can eliminate flow cycles in the optimal routing with a $O(E \log N)$ -time algorithm for each commodity [19].

III. THEORETICAL FOUNDATION: NEM

In this section, we present the theory of realizing optimal TE with link-state protocols. We first compute the minimal load that each link must carry to achieve optimal traffic distribution, then examine all the traffic-splitting choices subject to necessary (minimal) link capacities. It turns out that the traffic-splitting configuration that is realizable with hop-by-hop forwarding can be picked out by maximizing a weighted sum of the entropies of traffic-splitting vectors. In addition, the corresponding link weights can be found efficiently by solving the new optimization problem using the gradient descent algorithm. It is important to realize that the proposed NEM framework developed in this section is used to *design* the protocol. The NEM problem itself is *not* solved by the operator or routers—it is constructed as a proof technique and an intermediate step toward the results in Sections IV and V.

A. Necessary Capacity

Given the traffic matrix and the objective function, the solution to the COMMODITY problem (2) provides the optimal distribution of traffic. We represent the resulting flow on each link (u,v) as the *necessary capacity* $\check{c}_{u,v} \triangleq f_{u,v}$ (or \check{c} as a vector). The necessary capacity is a minimal³ set of link capacities to realize optimal traffic engineering.

There could be numerous ways of traffic splitting that realize optimal TE. If we replace link capacity $c_{u,v}$ in COMMODITY (2) with the necessary capacity $\check{c}_{u,v}$,⁴ we are free to impose another objective function to pick out a particular optimal solution to the original problem. A key challenge here is to design a new objective function, purely for the purpose of protocol development, such that the resulting routing of flow can be realized *distributively with link-state routing protocols and hop-by-hop forwarding*.

B. Network Entropy Maximization

Denote $P_{s,t}$ as the set of paths from s to t (repeated nodes are allowed) and $x_{s,t}^i$ as the probability (fraction) of forwarding a packet of demand $D(s,t)$ to the i th path ($P_{s,t}^i$). Obviously, $\sum_i x_{s,t}^i = 1$. If we require the probabilities of using two paths to be the same as long as they are of the same length (see Appendix-B for details), to be realized with hop-by-hop forwarding, the values of $x_{s,t}^i$ should satisfy

$$\frac{x_{s,t}^i}{x_{s,t}^j} = \frac{g\left(\sum_{(u,v) \in \mathbb{E}} K_{P_{s,t}^i}^{(u,v)} w_{u,v}\right)}{g\left(\sum_{(u,v) \in \mathbb{E}} K_{P_{s,t}^j}^{(u,v)} w_{u,v}\right)}. \quad (3)$$

where $w_{u,v}$ is the weight assigned to link (u,v) , $K_{P_{s,t}^i}^{(u,v)}$ is the number of times $P_{s,t}^i$ passes through link (u,v) ($P_{s,t}^i$ can contain cycles), and $g(\cdot)$ is a known function for all the routers. We find

³But may not be the minimum capacity. \bar{c} is minimal if $\nexists \bar{c}' : \bar{c}' \neq \bar{c} \wedge \bar{c}' \leq \bar{c}$, whereas \check{c} is the minimum if $\forall \check{c}' : \check{c} \leq \check{c}'$.

⁴The link cost is still defined in terms of the original link capacity, i.e., link utilization or cost will not be changed due to the use of necessary capacity.

that the set of values of $x_{s,t}^i$ satisfying (3) maximizes a “network entropy” defined as follows. Consider the entropy function $z(x_{s,t}^i) = -x_{s,t}^i \log x_{s,t}^i$ for source–destination pair (s, t) . The weighted sum, $\sum_{s,t} \left(D(s, t) \sum_{i \in P_{s,t}} z(x_{s,t}^i) \right)$, is defined as the network entropy.⁵

Now we define the NEM problem under the necessary capacity constraints as follows:

$$\max \sum_{s,t} \left(D(s, t) \sum_{i \in P_{s,t}} z(x_{s,t}^i) \right) \quad (4a)$$

$$\text{s.t.} \sum_{s,t,i} D(s, t) K_{P_{s,t}^i}^{(u,v)} x_{s,t}^i \leq \tilde{c}_{u,v} \quad \forall (u, v) \quad (4b)$$

$$\sum_i x_{s,t}^i = 1 \quad \forall s, t \quad (4c)$$

$$\text{vars.} \quad x_{s,t}^i \geq 0. \quad (4d)$$

From the optimal solution of the COMMODITY problem, we know the feasibility set of NEM is nonempty. For a concave maximization over a nonempty, compact constraint set, there exist globally optimal solutions to NEM.

C. Solve NEM by Dual Decomposition

We will connect the characterization of optimal solutions to NEM that are realizable with hop-by-hop forwarding to exponential penalty. Toward that end and to provide a foundation for link weight computation in Section V, we first investigate the Lagrange dual problem of NEM and a gradient-based solution.

Denote dual variables for constraints (4b) as $\lambda_{u,v}$ for link (u, v) (or λ as a vector). We first write the Lagrangian $L(\mathbf{x}, \lambda)$ associated with the NEM problem

$$L(\mathbf{x}, \lambda) = \sum_{s,t} \left(D(s, t) \sum_{i \in P_{s,t}} z(x_{s,t}^i) \right) - \sum_{(u,v) \in E} \lambda_{u,v} \left(\sum_{s,t,i} D(s, t) K_{P_{s,t}^i}^{(u,v)} x_{s,t}^i - \tilde{c}_{u,v} \right). \quad (5)$$

The Lagrange dual function is

$$Q(\lambda) = \max_{\substack{\mathbf{1} \succeq \mathbf{x} \succeq \mathbf{0} \\ \|\mathbf{x}_{s,t}\| = 1}} L(\mathbf{x}, \lambda) \quad (6)$$

where $\mathbf{0}$ and $\mathbf{1}$ are the vectors whose elements are all zeros and ones, respectively, and $\mathbf{x}_{s,t}$ is the vector of $x_{s,t}^i$.

The dual problem is formulated as

$$\begin{aligned} \min \quad & Q(\lambda) \\ \text{s.t.} \quad & \lambda \succeq \mathbf{0}. \end{aligned} \quad (7)$$

To solve the dual problem, we first consider problem (6). The maximization of the Lagrangian over \mathbf{x} can be solved as TRAFFIC-DISTRIBUTION problem (8).

⁵The physical interpretation of entropy for IP routing and the uniqueness of choosing the entropy function to pick out the right flow distributions are presented in Appendix-C and Appendix-B, respectively.

TRAFFIC-DISTRIBUTION:

$$\max \sum_{(u,v) \in E} \lambda_{u,v} \tilde{c}_{u,v} + \sum_{s,t} \left(D(s, t) \sum_{i \in P_{s,t}} z(x_{s,t}^i) \right)$$

$$- \sum_{(u,v) \in E} \lambda_{u,v} \left(\sum_{s,t,i} D(s, t) K_{P_{s,t}^i}^{(u,v)} x_{s,t}^i \right) \quad (8a)$$

$$\text{s.t.} \quad \sum_i x_{s,t}^i = 1 \quad \forall s, t. \quad (8b)$$

Then, the dual problem (7) can be solved by using the gradient descent algorithm as follows for iterations indexed by q :

$$\begin{aligned} & \lambda_{u,v}(q+1) \\ &= \left[\lambda_{u,v}(q) - \alpha(q) \left(\tilde{c}_{u,v} - \sum_{s,t,i} D(s, t) K_{P_{s,t}^i}^{(u,v)} x_{s,t}^i(q) \right) \right]^+ \\ &= [\lambda_{u,v}(q) - \alpha(q) (\tilde{c}_{u,v} - f_{u,v}(q))]^+ \quad \forall (u, v) \in E \end{aligned} \quad (9)$$

where $\alpha(q) > 0$ is the step-size, $x_{s,t}^i(q)$ are solutions of the TRAFFIC-DISTRIBUTION problem (8) for a given $\lambda(q)$, and $f_{u,v}(q)$ is the total flow on link (u, v) .

After this dual decomposition, the following result can be proven with standard convergence analysis for gradient algorithms [20].

Lemma 1: By solving the TRAFFIC-DISTRIBUTION problem for the NEM problem and the dual variable update (9), $\lambda(q)$ converge to the optimal dual solutions λ^* , and the corresponding primal variables \mathbf{x}^* are the globally optimal primal solutions of (4).

Proof: See Appendix-D. ■

D. Solve TRAFFIC-DISTRIBUTION Problem

Note that, the TRAFFIC-DISTRIBUTION problem is also separable, i.e., the traffic splitting for each demand across its paths is independent of the others since they are not coupled together with link capacity constraint (4b). Therefore, we can solve a subproblem (10) for each demand $D(s, t)$ separately.

DEMAND-DISTRIBUTION for $D(s, t)$:

$$\begin{aligned} \max \quad & D(s, t) \sum_{i \in P_{s,t}} z(x_{s,t}^i) \\ & - \sum_{(u,v) \in E} \lambda_{u,v} \left(\sum_i D(s, t) K_{P_{s,t}^i}^{(u,v)} x_{s,t}^i \right) \end{aligned} \quad (10a)$$

$$\text{s.t.} \quad \sum_i x_{s,t}^i = 1. \quad (10b)$$

We write the Lagrangian associated with the DEMAND-DISTRIBUTION subproblem as

$$\begin{aligned} L^r(\mathbf{x}_{s,t}, \mu_{s,t}) &= \left(D(s, t) \sum_{i \in P_{s,t}} z(x_{s,t}^i) \right) - \mu_{s,t} \left(\sum_i x_{s,t}^i - 1 \right) \\ & - \sum_{(u,v) \in E} \lambda_{u,v} \left(\sum_i D(s, t) K_{P_{s,t}^i}^{(u,v)} x_{s,t}^i \right) \end{aligned} \quad (11)$$

where $\mu_{s,t}$ is the Lagrangian variable associated with (10b).

According to Karush–Kuhn–Tucker (KKT) conditions⁶ [21], at the optimal solution of the DEMAND-DISTRIBUTION sub-problem, we have

$$z'(x_{s,t}^{i*}) - \sum_{(u,v)} K_{P_{s,t}^i}^{(u,v)} \lambda_{u,v} - \frac{\mu_{s,t}^*}{D(s,t)} = 0. \quad (12)$$

For the entropy function, $z(x) = -x \log x$, $z'(x) = -1 - \log x$, we have

$$x_{s,t}^{i*} = e^{-\left(\sum_{(u,v)} K_{P_{s,t}^i}^{(u,v)} \lambda_{u,v} + \frac{\mu_{s,t}^*}{D(s,t)} + 1\right)} \quad (13)$$

where $x_{s,t}^{i*}$, $\mu_{s,t}^*$ are the values of the $x_{s,t}^i$, $\mu_{s,t}$, respectively, at the optimal solution.

Then, for two paths i, j from s to t , we have

$$\frac{x_{s,t}^{i*}}{x_{s,t}^{j*}} = \frac{e^{-\left(\sum_{(u,v)} K_{P_{s,t}^i}^{(u,v)} \lambda_{u,v}\right)}}{e^{-\left(\sum_{(u,v)} K_{P_{s,t}^j}^{(u,v)} \lambda_{u,v}\right)}}. \quad (14)$$

We pause to examine the engineering implications of (14). If we use $\lambda_{u,v}$ as the weight $w_{u,v}$ for link (u, v) , the probability of using path $P_{s,t}^i$ is inversely proportional to the exponential value of its path length. It is important to observe at this point that since (14) has no factor of $\mu_{s,t}^*$, an intermediate router can ignore the source of the packet when making forwarding decisions. Equally importantly, from (9), in iteration q , the procedure for updating link weights does not need the values of $x_{s,t}^i(q)$. Instead, the procedure just needs $f_{u,v}(q)$, the aggregated bandwidth usage. We will show how to calculate $f_{u,v}(q)$ efficiently in Section V-B.

Now, combining the optimality results in Section II-B and Lemma 1 with the existence of (14), we have the following.

Theorem 1: Optimal traffic engineering (i.e., the optimal multicommodity flow) for a given traffic matrix can be realized with link weights using exponential flow splitting (14).

IV. NEW LINK-STATE ROUTING PROTOCOL: PEFT

In this section, we translate the theoretical results in Section III into a new link-state routing protocol run by routers. Each router makes an *independent* decision on how to forward traffic to a destination (i.e., flow-splitting ratios) among its outgoing links, using *only* the link weights. We first present PEFT from (14) and summarize the notation of the traffic-splitting function [1] for calculating flow-splitting ratios. Then, we show an efficient way to calculate the traffic-splitting function for the flow with PEFT routing, which can be approximated to further simplify the computation of traffic-splitting ratios in practice.

A. PEFT

Based on (14), we propose a new link-state routing protocol, called PEFT. The fraction of the traffic (from u to t) distributed across the i th path (or probability of forwarding a packet), $x_{u,t}^i$,

⁶KKT is a necessary condition, but NEM must have a global optimal solution. Thus, we must have one set of $x_{s,t}^{i*}$, $\mu_{s,t}^*$ for (12).

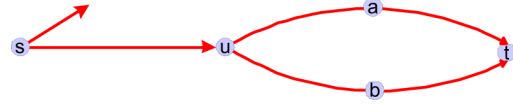


Fig. 1. Realize a PEFT flow using hop-by-hop forwarding.

is inversely proportional to the exponential value of its path length $p_{u,t}^i \triangleq \sum_{(u',v) \in E} K_{P_{u,t}^i}^{(u',v)} w_{u',v}$

$$\text{PEFT: } x_{u,t}^i = \frac{e^{-p_{u,t}^i}}{\sum_j e^{-p_{u,t}^j}}. \quad (15)$$

Theorem 1 in Section III shows PEFT can achieve optimal TE. A PEFT flow can be realized with hop-by-hop forwarding. For the sample network in Fig. 1, for the two paths from s to t ($s \rightarrow u \rightarrow a \rightarrow t$ and $s \rightarrow u \rightarrow b \rightarrow t$) and two paths from u to t , the flows on them for PEFT (15) satisfy

$$f_{s \rightarrow u \rightarrow a \rightarrow t} : f_{s \rightarrow u \rightarrow b \rightarrow t} = f_{u \rightarrow a \rightarrow t} : f_{u \rightarrow b \rightarrow t} \quad (16)$$

Therefore, router u can treat the packets from different sources (e.g., s or u) equally by forwarding them among the outgoing links with precalculated splitting ratios. Formally, we have the following.

Proposition 1: The PEFT flow for a set of link weights can be realized with hop-by-hop forwarding.

Proof: For the traffic from s to t , assume $P_i(s, u, t)$ is the set of all the paths (having flow from s to t) that share i , a sub-path (segment) from s to u , and $P(u, t)$ is the set of all paths having flow from u to t . From PEFT (15), the traffic-splitting ratio of the flows on $P_i(s, u, t)$ is equal to that of $P(u, t)$. The equality holds for every set of $P_i(s, u, t)$ for a PEFT flow. Thus, the flow can be realized with hop-by-hop forwarding. ■

As a link-state routing protocol, we need to define the traffic-splitting function for PEFT as follows.

B. Review: Traffic-Splitting Function

The notation of traffic-splitting (allocation) function was introduced in [1] to succinctly describe link-state routing protocols. In a directed graph, each unidirectional link (u, v) has a single, configurable weight $w_{u,v}$. Based on a complete view of the topology and link weights, a router can compute the shortest distance d_u^t from any node u to node t ; $d_v^t + w_{u,v}$ represents the distance from u to t when routed through neighboring node v . *Shortest-distance gap* $h_{u,v}^t$ is defined as $d_v^t + w_{u,v} - d_u^t$, which is always greater than or equal to 0. Then, (u, v) lies on a shortest path to t if and only if $h_{u,v}^t = 0$. Traffic-splitting function ($\Gamma(h_{u,v}^t)$) indicates the relative amount of traffic destined to t that node u will forward via outgoing link (u, v) .⁷ Let f_u^t denote the total incoming flow (destined to t) at node u (including the

⁷For example, the traffic-splitting function for even splitting across shortest paths (e.g., OSPF) is

$$\Gamma_O(h_{u,v}^t) = \begin{cases} 1, & \text{if } h_{u,v}^t = 0 \\ 0, & \text{if } h_{u,v}^t > 0. \end{cases}$$

passing-through flow and self-originated flow). The total outgoing flow of traffic (destined to t) traversing link (u, v) , $f_{u,v}^t$, can be computed as follows:

$$f_{u,v}^t = f_u^t \frac{\Gamma(h_{u,v}^t)}{\sum_{(u,j) \in E} \Gamma(h_{u,j}^t)}. \quad (17)$$

Consistent with hop-by-hop forwarding, u splits the traffic over the outgoing links without regard to the source node or the incoming link from which the traffic arrived.

C. Exact Traffic-Splitting Function for PEFT

The traffic-splitting function for PEFT can be calculated by each node autonomously and in polynomial time. From the definition of PEFT (15), more traffic should be sent along an outgoing link used by more paths, and the paths should be treated differently based on their path lengths. To compute the traffic splitting on each outgoing link, we first define a positive real number Υ_u^t , possibly interpretable as the “equivalent number” of shortest paths from node u to destination t , and let $\Upsilon_t^t \triangleq 1$.

For a PEFT flow, we have

$$\begin{aligned} \Upsilon_u^t &\triangleq \sum_{i \in P_{u,t}} e^{-(p_{u,t}^i - d_u^t)} \quad (18a) \\ &= \sum_{(u,v) \in E} \left(\sum_{j \in P_{v,t}} e^{-(p_{v,t}^j + w_{u,v} - d_u^t - d_v^t + d_v^t)} \right) \\ &= \sum_{(u,v) \in E} \left(e^{-(d_v^t + w_{u,v} - d_u^t)} \sum_{j \in P_{v,t}} e^{-(p_{v,t}^j - d_v^t)} \right) \\ &= \sum_{(u,v) \in E} \left(e^{-h_{u,v}^t} \Upsilon_v^t \right). \quad (18b) \end{aligned}$$

The recursive relationship represented in (18b)⁸ can be used in the following way: $e^{-h_{u,v}^t} \Upsilon_v^t$ is an “equivalent number” of shortest paths from u to t for those paths passing through link (u, v) and the router should distribute the traffic from u on link (u, v) in proportion to $e^{-h_{u,v}^t} \Upsilon_v^t$. Then, we have an exact traffic-splitting function⁹ for PEFT at link (u, v)

$$\Gamma_{PX}(h_{u,v}^t) = \Upsilon_u^t e^{-h_{u,v}^t}. \quad (19)$$

To enable hop-by-hop forwarding, each router needs to independently calculate $\Gamma_{PX}(h_{u,v}^t)$ for all node pairs. Then, each router first computes the all-pairs shortest paths, using, e.g., the Floyd–Warshall algorithm with time complexity $O(N^3)$ [22] and calculates the values of $e^{-h_{u,v}^t}$. Then, for each destination t , to compute the values of Υ_u^t , each router needs to solve

⁸Allowing for paths with cycles is required for the recursive derivation of (18b) (i.e., from $\sum_{j \in P_{v,t}} e^{-(p_{v,t}^j - d_v^t)}$ to Υ_v^t). Consider a simple example with two unidirectional links between u and v [i.e., (u, v) and (v, u)], and $P_{u,t}^i$ and $P_{v,t}^i$ are the sets of the paths to t from u and v , respectively. Then, the concatenation of link (u, v) and $P_{v,t}^i$, which may create paths with cycle, is a subset of $P_{u,t}^i$. Similarly, the concatenation of link (v, u) and $P_{u,t}^i$ is a subset of $P_{v,t}^i$. However, if optimal TE is acyclic, only cycle-free paths will be used because longer paths are exponentially penalized.

⁹ P in the subscript emphasizes that the calculation of traffic splitting considers the paths toward destination, and X denotes exactness.

N linear (18b), which requires $O(N^3)$ time [22]. Thus, the total complexity is $O(N^4)$.

D. Detour: Traffic-Splitting Function for “Downward PEFT”

To prevent cycles in link-state routing, packets are usually forwarded along a “downward path” where the next hop is closer to destination. This inspires the following *Downward PEFT*, whose traffic-splitting function is $\Gamma_{PD}(h_{u,v}^t)$:¹⁰

$$\Gamma_{PD}(h_{u,v}^t) = \begin{cases} \Upsilon_v^t e^{-h_{u,v}^t}, & \text{if } d_u^t > d_v^t \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

$\Gamma_{PD}(h_{u,v}^t)$ can approximate $\Gamma_{PX}(h_{u,v}^t)$ and further simplify the computation of Υ_u^t and traffic splitting as discussed below and utilized in Section V-C.

We consider each destination t independently. After temporarily removing link (u, v) where $d_u^t \leq d_v^t$ since there is no flow on it, we get an acyclic network and do topological sorting on the remaining network. Proceeding through the nodes u in *increasing* topological order (starting with destination t), we compute the value of Υ_u^t using (18b). For each destination, topology sorting requires $O(N + E)$ time, and summarizing the Υ_u^t across the outgoing links requires $O(N + E)$ time. Thus, the total time complexity to calculate Υ_u^t is $O(N^3 + N(N + E)) = O(N^3)$.

In general, “*Downward PEFT*” does not provably achieve optimal TE, in contrast to PEFT, although it comes extremely close to optimal TE in practice, with the associated link weight computation even faster than that for PEFT. In the case where the lower bound of all link weights, w_{\min} , is large enough, the downward PEFT is same as PEFT.¹¹

E. Discussion

In the control plane, PEFT does not change the routing-protocol messages that are sent between the routers (an important consideration for practical use), but does change the computation done locally on each router based on the weights.

In the data plane, routers today implement hash-based splitting over multiple outgoing links, typically with an even (1 out of n) splitting ratio. PEFT requires flexible splitting over multiple outgoing links, thus we need to store the splitting percentages—whereas for $1/n$ spitting, the splitting ratio is implicitly even. It requires a little extra storage and processing, not enough to become a new bottleneck, when packets arrive to direct packets to the appropriate outgoing links.

An optimal distribution of traffic could have flow cycles if the objective $\Phi(\{f_{u,v}, c_{u,v}\})$ is not a strictly increasing function of link flow $f_{u,v}$. Both cyclic or acyclic optimal traffic distributions can be realized with Exact PEFT. For a cyclic optimal traffic distribution, Exact PEFT may result in cycles in link-state routing. For an acyclic optimal traffic distribution (or with flow cycles removed as in [19]), the flow on the cyclic paths in Exact PEFT solution should be sufficiently close to 0. Downward PEFT is

¹⁰ D in the subscript emphasizes “downward.”

¹¹For link (u, v) , if the shortest distance to t of u is $d_u^t \leq d_v^t$, then $h_{u,v}^t = d_v^t + w_{u,v} - d_u^t \geq w_{u,v}$ and $\Gamma_{PX}(h_{u,v}^t) \leq \Upsilon_v^t e^{-w_{u,v}}$, and the flow destined to t on (u, v) is close to 0 if $w_{u,v}$ is large enough, e.g., $e^{-10} \approx 0.005\%$. Therefore, most flow in PEFT always makes forward progress toward the destination, i.e., from router u with larger d_u^t to router v with smaller d_v^t .

Algorithm 1: Optimize Over Link Weights

```

1: Compute necessary capacities  $\tilde{c}$  by solving (2)
2:  $\mathbf{w} \leftarrow$  Any set of link weights
3:  $\mathbf{f} \leftarrow$  Traffic_Distribution( $\mathbf{w}$ )
4: while  $\mathbf{f} \neq \tilde{c}$  do
5:  $\mathbf{w} \leftarrow$  Link_Weight_Update( $\mathbf{f}$ )
6:  $\mathbf{f} \leftarrow$  Traffic_Distribution( $\mathbf{w}$ )
7: end while
8: Return  $\mathbf{w}$  /*final link weights*/

```

a faster but approximate solution to realize an acyclic optimal traffic distribution.

V. LINK-WEIGHT COMPUTATION FOR PEFT

Section IV described how routers split traffic under PEFT. A new way to use link weights also means the network operator needs a new way to compute, centrally and offline, the optimal link weights. It turns out that the NP-hard problem of link-weight computation in OSPF can be turned into a convex optimization when link weights are used by PEFT. To do that, we will convert the iterative method of solving the NEM problem in Section III into a simple and efficient algorithm. We first present an algorithm that iteratively chooses a tentative set of link weights and evaluates the corresponding traffic distribution by simulating the PEFT traffic splitting run by the routers. From Theorem 1, the algorithm is guaranteed to converge to a set of link weights, which realizes optimal TE with PEFT. To further speed up the calculation, the traffic distribution with PEFT for each iteration can be *approximated* with downward PEFT. The simulation in Section VI shows that such an approximation is very close to optimal and provides substantial speedup in practice.

A. Algorithm Framework for Optimizing Link Weights

The iterative algorithm consists of two main parts:

- 1) computing the optimal traffic distribution (necessary capacities) for a given traffic matrix by solving the COMMODITY problem (2);
- 2) computing the link weights that would achieve the optimal traffic distribution.

The second step uses the optimal traffic distribution found in the first step as input and need not consider the objective function ($\Phi(\{f_{u,v}, c_{u,v}\})$) any further. Starting with an initial setting of link weights, the algorithm (see Algorithm 1) repeatedly updates the link weights until the load on each link is the same as the necessary capacity. Each setting of the link weights corresponds to a particular way of splitting the traffic over a set of paths. The Traffic_Distribution procedure computes the resulting link loads $f_{u,v}$ based on the traffic matrix. Then, the Link_Weight_Update procedure (see Algorithm 2) increases the weight of each link (u, v) linearly if the traffic exceeds the necessary capacity, or decreases it otherwise. The parameter α is a positive step-size, which can be constant or dynamically adjusted; we find that setting α to the reciprocal of the maximum

Algorithm 2: Link_Weight_Update(\mathbf{f})

```

1: for each link  $(u, v)$  do
2:  $w_{u,v} \leftarrow w_{u,v} - \alpha(\tilde{c}_{u,v} - f_{u,v})$ 
3: end for
4: Return new link weights  $\mathbf{w}$ 

```

Algorithm 3: Traffic_Distribution \mathbf{w} with $\Gamma_{PX}(\cdot)$

```

1: For link weights  $\mathbf{w}$ , construct all-pairs shortest paths (e.g.,
   with Floyd–Warshall algorithm) and compute  $\Gamma_{PX}(h_{u,v}^t)$ 
2: For each  $t$ , compute  $f_u^t$  by solving linear (21)
3:  $f_{u,v}^t \leftarrow f_u^t \frac{\Gamma_{PX}(h_{u,v}^t)}{\sum_{(u,j) \in E} \Gamma_{PX}(h_{u,j}^t)}$ 
4:  $f_{u,v} \leftarrow \sum_{t \in \mathcal{V}} f_{u,v}^t$ 
5: Return  $\mathbf{f}$  /*set of  $f_{u,v}$ , total flow on each link*/

```

necessary link capacity $\left(\frac{1}{\max \tilde{c}_{u,v}}\right)$ performs well in practice. Algorithm 1 is guaranteed to converge to the global optimal solution as stated in Lemma 1.

In terms of computational complexity, we know that COMMODITY can be solved efficiently. The complexity of Algorithm 2 is $O(E)$. The remaining question is how to solve the subproblem Traffic_Distribution(\mathbf{w}) efficiently.

B. Compute Traffic Distribution With PEFT

To compute the traffic distribution for PEFT, we should first compute the shortest paths between each pair of nodes and all the values $\Gamma_{PX}(h_{u,v}^t)$ as in Section IV-C, which is shown as the first step of Algorithm 3. Computing the resulting distribution of traffic is complicated by the fact that $\Gamma_{PX}(\cdot)$ may direct traffic “backwards” to a node that is farther away from the destination. To capture these effects, recall that f_u^t is the total incoming flow at node u (including traffic originating at u as well as any traffic arriving from other nodes) that is destined to node t . In particular, the traffic $D(s, t)$ that enters the network at node s and leaves at node t satisfies the following linear equation:

$$f_s^t - \sum_{x:(x,s) \in E} f_x^t \left(\frac{\Gamma_{PX}(h_{x,s}^t)}{\sum_{(x,j) \in E} \Gamma_{PX}(h_{x,j}^t)} \right) = D(s, t). \quad (21)$$

That is, the traffic $D(s, t)$ entering the network at node s matches the total incoming flow f_s^t at node s (destined to node t), excluding the traffic entering s from other nodes. The transit flow is captured as a sum over all incoming links from neighboring nodes x , which split their incoming traffic f_x^t over their links based on the traffic-splitting function.

Algorithm 3 computes the traffic distribution by solving the system of linear (21) and computing the resulting flow on each link (u, v) . The N linear (21) for each t typically require $O(N^3)$ time [22] to solve. Thus, the total complexity is $O(N^4)$.

C. Approximate Traffic Distribution With “Downward PEFT”

If optimal traffic distribution is cycle-free, we can further reduce the computational overhead in link-weight computation.

Algorithm 4: Traffic_Distribution(\mathbf{w}) with $\Gamma_{PD}(\cdot)$

-
- 1: For link weights \mathbf{w} , construct all-pairs shortest paths and compute $\Gamma_{PD}(h_{u,v}^t)$
 - 2: **for each** destination t **do**
 - 3: Temporarily remove link (u, v) where $d_u^t \leq d_v^t$
 - 4: Do topological sorting on the remaining network
 - 5: **for each** source $s \neq t$ in the decreasing topological order **do**
 - 6: $f_s^t \leftarrow D(s, t) + \sum_{x:(x,s) \in E} f_{x,s}^t$
 - 7: $f_{s,v}^t \leftarrow f_s^t \frac{\Gamma_{PD}(h_{s,v}^t)}{\sum_{(s,j) \in E} \Gamma_{PD}(h_{s,j}^t)}$
 - 8: **end for**
 - 9: **end for**
 - 10: $f_{u,v} \leftarrow \sum_{t \in V} f_{u,v}^t$
 - 11: Return \mathbf{f} /*set of $f_{u,v}$ */
-

Note that, if the optimal traffic distribution is acyclic, in the last iteration in Algorithm 1, the flow cycle will be negligible. In addition, the accurate solution for each intermediate iteration is not necessary in practice, we can approximate PEFT ($\Gamma_{PX}(\cdot)$) with Downward PEFT ($\Gamma_{PD}(\cdot)$) to forward traffic only on “downward” paths, and the traffic distribution for each intermediate iteration can be computed using a combinatorial algorithm, which is significantly faster than solving linear (21).

As in Section V-B, we first compute the shortest paths between all pairs of nodes, as well as the values of $\Gamma_{PD}(h_{u,v}^t)$, as shown in the first step of Algorithm 4. The following procedure is very similar to, but subtly different from, that for calculating $\Gamma_{PD}(h_{u,v}^t)$. We consider each destination t independently since the flow to each destination can be computed without regard to the other destinations. After temporarily removing link (u, v) where $d_u^t \leq d_v^t$ since there is no flow on it, we get an acyclic network and do topological sorting on the remaining network. The computation starts at the node without any incoming link in the acyclic network since this node would never carry any traffic to t that originates at other nodes. Proceeding through the nodes s in *decreasing* topological order, we compute the total incoming flow at node s (destined to t) as the sum of the flow originating at s [i.e., $D(s, t)$] and the flow arriving from neighboring nodes x ($f_{x,s}^t$). Then, we use the total incoming flow at s to compute the flow of traffic toward t on each of its outgoing links (s, v) using the traffic-splitting function $\Gamma_{PD}(\cdot)$.

In Algorithm 4, computing the all-pairs shortest paths with the Floyd–Warshall algorithm has time complexity $O(N^3)$ [22]. For each destination, topology sorting requires $O(N + E)$ time, and summarizing the incoming flow and splitting across the outgoing links requires $O(N + E)$ time. Thus, the total time complexity to run Algorithm 4 in each iteration of Algorithm 1 is $O(N^3 + N(N + E)) = O(N^3)$.

Finally, the total running time for Algorithm 1 depends on the time required to solve (2) and the total number of iterations required for Algorithms 2 and 4. Interestingly, although the original NEM problem involves an infinite number of variables, the complexity of Algorithm 1 is still comparable to solving a convex optimization with polynomial number of variables [like

the COMMODITY problem (2)] using the gradient descent algorithm since we do not need to solve NEM directly.¹² However, in the terminology of complexity theory, link-weight computation for PEFT is not yet proven to be polynomial-time, although in the special case of single destination, we can compute PEFT in polynomial time as shown in Proposition 2.

Proposition 2: Downward PEFT can achieve acyclic optimal traffic engineering with a single destination in polynomial time. See Appendix-F for proof.

VI. PERFORMANCE EVALUATION

How well can the new routing protocol PEFT perform, and how fast can the new link weight computation be? PEFT has been already proven to achieve optimal TE in Section III, with a complexity of link-weight computation similar to that of solving convex optimization (with a polynomial number of variables). In this section, we numerically demonstrate that its approximate version, Downward PEFT, can make convergence very fast in practice while coming extremely close to TE optimality.

A. Simulation Environment

We consider two network objective functions ($\Phi(\{f_{u,v}, c_{u,v}\})$): maximum link utilization and total link cost (1) (as used in operator’s TE formulation). For benchmarking, the optimal values of both objectives are computed by solving linear program (2) with CPLEX 9.1 [23] via AMPL [24].

To compare to OSPF, we use the state-of-the-art local-search method in [2]. We adopt TOTEM 1.1 [25], which follows the same approach as [2] and has similar quality of the results.¹³ We use the same parameter setting for local search as in [2], [18], where the link weights are restricted as integers from 1 to 20 since a larger weight range would slow down the searching [18], initial link weights are chosen randomly, and the best result is collected after 5000 iterations.

Note that here we do not evaluate and compare some previous works using noneven splitting over shortest paths [4], [5] since these solutions do not enable routers to *independently* compute the flow-splitting ratios from link weights.

To determine link weights under PEFT, we run Algorithm 1 with up to 5000 iterations of computing the traffic distribution and updating link weights. Abusing terminology a little, in this section we use the term PEFT to denote the traffic engineering with Algorithm 1 (including two sub-Algorithms 2 and 4).

We run the simulation for a real backbone network and several synthetic networks. The properties of the networks used are summarized in Table IV, which will be presented in Subsection VI-E. First is the Abilene network (Fig. 2) [26], which has 11 nodes and 28 directional links with 10-Gb/s capacity. The traffic demands are extracted from the sampled Netflow data on November 15, 2005. To simulate networks

¹²We do not need to write down the NEM problem explicitly or obtain the optimal value for each variable. Instead, we just search for E dual variables (link weights) that can enable optimal solution of NEM problem. Each step in the proposed gradient descent algorithm has polynomial-time complexity in terms of the number of nodes and edges.

¹³Proprietary enhancements can bring in factors of improvement, but as we will see, PEFT’s advantage on computational speed is orders of magnitude.

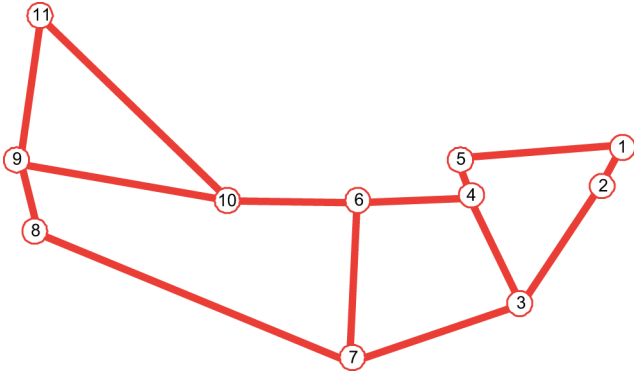


Fig. 2. Abilene network.

TABLE III
MAXIMUM LINK UTILIZATION OF OPTIMAL TRAFFIC ENGINEERING, PEFT,
AND LOCAL SEARCH OSPF FOR LIGHT-LOADING NETWORKS

Net. ID	Optimal TE	PEFT	OSPF
abilene	33.9%	33.9%	39.8%
hier50a	56.4%	56.5%	58.6%
hier50b	44.7%	45.0%	59.2%
rand50	60.6%	60.6%	60.6%
rand50a	60.8%	60.8%	64.7%
rand100	55.0%	55.0%	71.5%

with different congestion levels, we create different test cases by uniformly decreasing the link capacity until the maximal link utilization reaches 100% with optimal TE.

We also test the algorithms on the same topologies and traffic matrices as those in [2]. The two-level hierarchical networks were generated using GT-ITM, which consists of two kinds of links: local access links with 200-unit capacity and long-distance links with 1000-unit capacity. In the random topologies, the probability of having a link between two nodes is a constant parameter, and all link capacities are 1000 units. In these test cases, for each network, traffic demands are uniformly increased to simulate different congestion levels.

B. Minimization of Maximum Link Utilization

Since we create different levels of congestion for the same network by uniformly decreasing link capacities or uniformly increasing traffic demands, we just need to compute the maximum link utilization (MLU) for one test case in each network because MLU is proportional to the ratio of total demand over total capacity. In addition to MLU, we are particularly interested in the metric “efficiency of capacity utilization,” η , which is defined as the following ratio: the percentage of the traffic demand satisfied when the MLU reaches 100% under a traffic engineering scheme over that in the optimal traffic engineering. The improvement in η is referred to as the “Internet capacity increase” in [2].

For any test case of a network, if MLU of optimal TE, OSPF, and PEFT are ξ , ξ_O , and ξ_D , respectively, then $\eta_O = \frac{\xi}{\xi_O}$ and $\eta_D = \frac{\xi}{\xi_D}$. Thus, PEFT can increase Internet capacity over OSPF by $\eta_D - \eta_O$. Table III shows the maximum link utilizations of optimal traffic engineering, PEFT, and Local Search OSPF for the test case with the lightest loading of each network. Fig. 3 illustrates the efficiency of capacity utilization of the three schemes. They show that PEFT is very close to optimal traffic

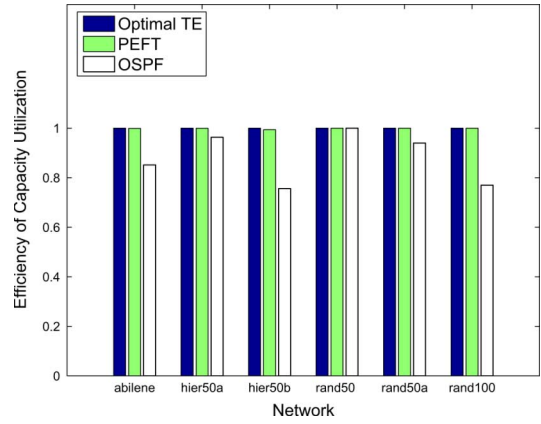


Fig. 3. Efficiency of capacity utilization of optimal traffic engineering, PEFT and Local Search OSPF.

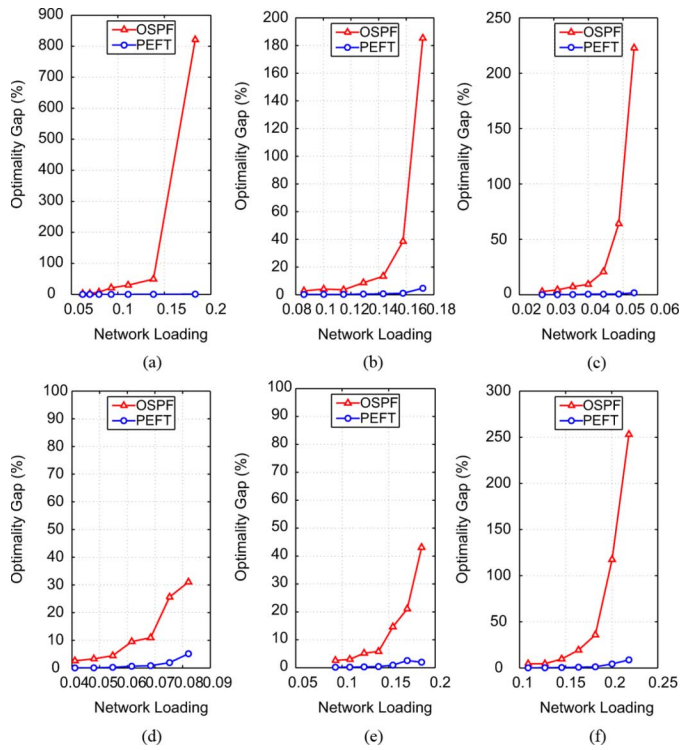


Fig. 4. Comparison of PEFT and Local Search OSPF in terms of optimality gap on minimizing total link cost. (a) Abilene network. (b) Rand100 network. (c) hier50b network. (d) hier50a network. (e) Rand50 network. (f) Rand50a network.

engineering in minimizing MLU and increases Internet capacity over OSPF by 15% for the Abilene network and 24% for the hier50b network, respectively.

C. Minimization of Total Link Cost

We also employ the cost function (1) as in [2]. The comparison is based on the optimality gap, in terms of the total link cost, compared against the value achieved by the optimal traffic engineering. Typical results for different topologies with various traffic matrices are shown in Fig. 4, where the network loading is the ratio of total demand over total capacity. From the results, we observe that the gap between OSPF and the optimal traffic engineering can be very significant (up to 821%) for the most

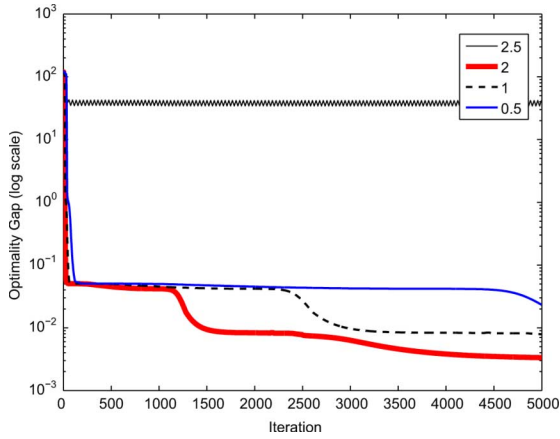


Fig. 5. Evolution of optimality gap of PEFT with different step-sizes.

congested case of the Abilene network. In contrast, PEFT can achieve almost the same performance as the optimal traffic engineering in terms of total link cost. Note that, within those figures, the maximum optimality gap of PEFT is only up to 8.8% in Fig. 4(b), which can be further reduced to 1.5% with a larger step-size and more iterations (which is feasible as the algorithm runs very quickly, to be shown in Section VI-E).

D. Convergence Behavior

Fig. 5 shows the optimality gap in terms of total cost achieved by PEFT, using different step-sizes, within the first 5000 iterations for the Abilene network with the least link capacities. It provides convergence behavior typically observed. The legends show the ratio of the step-size over the default setting. It demonstrates that the algorithm developed in Section V for the PEFT protocol converges very fast even with the default setting, and reduces the gap to 5% after 100 iterations and 1% after 3000 iterations. In addition, increasing step-size a little will speed up the convergence and as expected; too large a step-size (e.g., 2.5 in the above example) would cause oscillation. Notice that there is a wide range of step-sizes that can make convergence very fast. An even faster solution with Newton's method can be found in [27].

E. Running Time Requirement

Besides the convergence behavior, the actual running time is also an important evaluation criteria. The tests for PEFT and local search OSPF were performed under the time-sharing servers of Redhat Enterprise Linux 4 with Intel Pentium IV processors at 2.8 ~ 3.2 GHz. Note that the running time for local search OSPF is sensitive to the traffic matrix since a near-optimal solution can be reached very quickly for light traffic matrices. Therefore, we show the range of their average running times per iteration for qualitative reference.

Fig. 6 shows the optimality gap (on a log scale) achieved by local search OSPF and PEFT within the first 500 iterations for a typical scenario [Fig. 4(c)]. It demonstrates that Algorithm 1 for PEFT converges much faster than local search for OSPF. Table IV shows the average running time per iteration for different networks. We observe that our algorithm is very

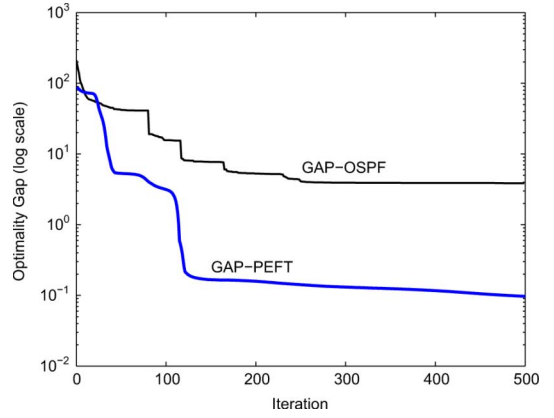


Fig. 6. Comparison of the drop in optimality gap between Local Search OSPF and PEFT in a two-level topology with 50 nodes and 212 links.

TABLE IV
AVERAGE RUNNING TIME PER ITERATION REQUIRED BY PEFT AND LOCAL SEARCH OSPF TO ATTAIN THE PERFORMANCE IN FIG. 4

Net. ID	Topology	Node #	Link #	Time per Iteration (second)	
				PEFT	OSPF
abilene	Backbone	11	28	0.002	6.0~13.9
hier50a	2-level	50	148	0.006	6.0~13.9
hier50b	2-level	50	212	0.007	6.4~17.4
rand50	Random	50	228	0.007	3.2~9.0
rand50a	Random	50	245	0.007	6.1~14.1
rand100	Random	100	403	0.042	39.5~105.1

fast, requiring at most 2 min even for the largest network (with 100 nodes) tested, while the OSPF local search needs tens of hours on the same computer. On average, the algorithm developed in this paper to find link weights for PEFT routing is 2000 times faster than local search algorithms for OSPF routing.

VII. NEM: A FRAMEWORK FOR LINK-STATE ROUTING

In this section, we highlight the conceptual framework of NEM and the differences between NEM and NUM.

As explained in Section III, NEM is developed in this paper as a unifying mathematical model that enables the discovery and development of new link-state routing protocol PEFT. Although NEM is solved by neither routers nor operators, its solution leads to both the development of PEFT traffic splitting and link-weight computation algorithms. More discussions on the intuitions behind NEM can be found in Appendix-C.

On the other hand, TCP congestion control protocols have been studied extensively since 1998 as solutions to another family of optimization models called NUM. The notion of network utility was first advocated in [28] in 1995 for bandwidth allocation among elastic demands on *source rates*. The NUM problem (22) was first introduced for TCP congestion control (e.g., [12]–[15]). Consider a communication network with L logical links, each with a fixed capacity of c_l b/s and S sources (i.e., end-users), each transmitting at a source rate of x_s b/s. Each source s emits one flow, using a fixed set $L(s)$ of links in its path, and has an increasing (and often concave) function $U_s(x_s)$ called utility function. Each link l is shared by a set $S(l)$ of sources. NUM, in its basic version, is the following problem of maximizing the network utility $\sum_s U_s(x_s)$, over the source rates \mathbf{x} , subject to linear flow

TABLE V
NUM FOR TCP AND NEM FOR IP: MAIN DIFFERENCES

Property	Congestion Control (TCP)	Traffic Engineering (IP)
Traffic type	Elastic	Inelastic
Flow distribution	Fixed	Variable
Participants	End user and router	Operator and router
Timescale	Seconds	Hours
Optimization Model	Network Utility Maximization	<i>Network Entropy Maximization</i>
Lagrange multipliers	Congestion price	<i>Link weight</i>
Reverse engineering	Tahoe, Reno, Vegas, etc.	<i>Even splitting in OSPF</i>
Forward engineering	FAST TCP, etc.	<i>PEFT</i>

constraints $\sum_{s \in \mathcal{S}(l)} x_s \leq c_l$ for all links l (note that routing is fixed in NUM formulation):

$$\begin{aligned}
 & \text{maximize} && \sum_s U_s(x_s) \\
 & \text{subject to} && \sum_{s \in \mathcal{S}(l)} x_s \leq c_l \quad \forall l \\
 & \text{variables} && \mathbf{x} \succeq 0.
 \end{aligned} \tag{22}$$

There is a useful economics interpretation of the dual-based distributed algorithm for NUM, in which the Lagrange dual variables can be interpreted as shadow prices for resource allocation, and end-users and the network maximize their net utilities and net revenue, respectively. Much reverse-engineering of existing TCP variants and forward-engineering of new congestion control protocols have been developed with the NUM model as a starting point.

The NEM problem proposed in this paper is *not* a special case of NUM since entropy is not an increasing function and the design freedom in NEM is routing rather than rate control. Instead, there is a useful and interesting *parallel* between the framework of NEM proposed this paper, for link-state routing protocols in the IP layer, and that of NUM matured over the last decade, for end-to-end congestion control protocols in the TCP layer. The comparison between the two frameworks is shown in Table V, where results from this paper are highlighted in italics.

VIII. CONCLUDING REMARKS

Commodity-flow-based routing protocols are optimal for any convex objective in Internet TE, but introduce much configuration complexity. Link-state routing is simple, but prior work suggests it does not achieve optimal TE. This paper proves that optimal traffic engineering, in fact, *can* be achieved by link-state routing with hop-by-hop forwarding, and the right link weights *can* be computed efficiently, as long as flow splitting on non-shortest paths is allowed but properly penalized. In the Appendix, we also show uniqueness of the exponential penalty in achieving optimal TE and discuss interpretations of NEM from the viewpoints of statistical physics and combinatorics.

Before concluding this paper, we would like to highlight that optimization is used in three different ways in this paper. First and obviously, it is used when developing algorithms to solve the link-weight computation problem for PEFT.

In a more interesting way, the level of difficulty of optimizing link weights for OSPF is used as a hint that perhaps we need to

revisit the standard assumption on how link weights should be used. In this approach of “Design For Optimizability,” sometimes a restrictive assumption in the protocol can be perturbed at low “cost” and yet turn a very hard network-management problem into an efficiently solvable one. In this case, better (and indeed the best) TE and faster weight computation are simultaneously achieved.

In yet another way, optimization in the form of NEM is introduced as a conceptual framework to develop routing protocols. The NEM framework for distributed routing also leads to several interesting future directions, including extensions to robust TE and to the interactions between congestion control at sources with link-state routing in the network.

APPENDIX

In this Appendix, we present more details about NEM and PEFT. Appendix-A explains the differences between PEFT and DEFT [1]. Appendix-B proves the uniqueness of choosing the entropy function to pick out the right flow distributions realizable with link-state routing. Appendix-C introduces a physical interpretation of entropy for IP routing. Appendix-D proves Lemma 1 on the convergence of solving the NEM problem with the gradient descent algorithm. Appendix-E introduces how to realize the multicommodity-flow solution with up to $O(N^2E)$ tunnels, which also can be used as an initialization for the NEM problem (4). Appendix-F proves Proposition 2 and shows a polynomial-time algorithm of setting optimal link weights for PEFT in a single-destination network.

A. Differences Between PEFT and DEFT

Here, we explain several points of potential confusion between PEFT in this paper and DEFT in [1]. Link-state routing protocols can be categorized as link-based and path-based in terms of flow splitting. Their difference is illustrated in Fig. 7, with a network that only has traffic demand from s to t . Assume the weights of the links are shown in Fig. 7(a). Obviously, the shortest distance from s to t is 2 units, and both nodes t and u are on the shortest paths from s to t . In a link-based splitting scheme (e.g., OSPF, Fong [7], and DEFT [1]), node s evenly splits traffic across its *two* outgoing links (s, t) and (s, u) as shown in Fig. 7(b), whereas in a path-based splitting scheme, e.g., PEFT, there are *three* equal-length paths from (s, t) and s evenly splits traffic across them as shown in Fig. 7(c). Note that the path-based model does not imply explicit routing to set up tunnels for all the possible paths. Instead, each node just needs to compute and stores the aggregated flow-splitting ratio across

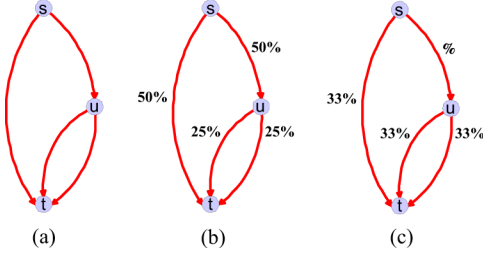


Fig. 7. Difference in traffic splittings for link-based and path-based link-state routing protocol. (a) Link weights. (b) Link-based splitting. (c) Path-based splitting.

its outgoing links, like 66% on link (s, u) for the sample network in Fig. 7(c). Therefore, path-based splitting schemes can still be realized with hop-by-hop forwarding.

The key differences between PEFT and DEFT are summarized as follows.

- 1) DEFT is a link-based flow splitting, while PEFT is a path-based flow splitting.
- 2) The core algorithms for setting link weights are completely different. Reference [1] introduces a nonconvex, nonsmooth optimization for DEFT and a two-stage iterative solution method, while the theory for PEFT is NEM. The two-stage method for DEFT is *much slower* than the algorithms developed for PEFT in this paper.
- 3) Reference [1] *numerically* shows DEFT can realize *near-optimal* TE in terms of a particular objective (total link cost), while this paper *proves* that PEFT can realize *optimal* TE with any convex objective function.

B. Uniqueness of Exponential Penalty

Can optimal traffic engineering be achieved by other penalty functions on longer paths? Here, we demonstrate that exponential penalty is the only way of realizing optimal traffic distribution with path-based link-state routing.

As in (12), we use $\lambda_{u,v}$ as weight for link (u, v) , denote $p \triangleq K_{P_{s,t}^{(u,v)}} \lambda_{u,v}$ as the length of the i th path, define $\frac{\mu_{s,t}^*}{D(s,t)}$ as q , and simplify $x_{s,t}^{i*}$ as x , then we have

$$z'(x) - p - q = 0 \quad (27)$$

then

$$z(x) = (p + q)x + C_1 \quad (28)$$

where C_1 is a constant and

$$p + q = \frac{z(x) - C_1}{x} \triangleq \psi(x). \quad (29)$$

Assume $\psi(x)$ is reversible, then we have

$$x = \psi^{-1}(p + q). \quad (30)$$

We also denote $x = \varphi(p, q)$. Note that, for path-based link-state routing, for two paths of the same demand $D(s, t)$, the ratio of

the traffic over them should depend only on their path lengths. For a path of length p and a shortest path of length p_0 , we have

$$\begin{aligned} \frac{\varphi(p, q)}{\varphi(p_0, q)} &= f_1(p, p_0) \\ \Rightarrow \log \varphi(p, q) - \log \varphi(p_0, q) &= \log f_1(p, p_0) \\ \Rightarrow \frac{d \log \varphi(p, q)}{dq} - \frac{d \log \varphi(p_0, q)}{dq} &= 0 \\ \Rightarrow \int_{q_0}^q \frac{d \log \varphi(p, q)}{dq} dq &= \int_{q_0}^q \frac{d \log \varphi(p_0, q)}{dq} dq \\ \Rightarrow \log \varphi(p, q)|_{q_0}^q &= \log \varphi(p_0, q)|_{q_0}^q \\ \Rightarrow \varphi(p, q) &= \frac{\varphi(p, q_0)\varphi(p_0, q)}{\varphi(p_0, q_0)} \quad (31) \end{aligned}$$

where p_0, q_0 are constants.

Therefore, we can define two functions $f(p) \geq 0$ and $g(q) \geq 0$, such that

$$x = f(p)g(q) \quad (32)$$

where

$$\frac{dx}{dp} = f'(p)g(q) \quad \frac{dx}{dq} = f(p)g'(q). \quad (33)$$

From (30), $\frac{dx}{dp} = \frac{dx}{dq}$, thus

$$f'(p)g(q) = f(p)g'(q) \Rightarrow \frac{f'(p)}{f(p)} = \frac{g'(q)}{g(q)}. \quad (34)$$

Since $\frac{f'(p)}{f(p)}$ is a function of p and $\frac{g'(q)}{g(q)}$ is a function of q , thus

$$\frac{f'(p)}{f(p)} = \frac{g'(q)}{g(q)} = C. \quad (35)$$

where $C < 0$ since $f'(p) \leq 0$ assuming we send more traffic on a shorter path.

Therefore, $f(p) = Ae^{Cp}$ and $g(q) = Be^{Cq}$, $x = AB e^{C(p+q)}$. Then, $z(x) = \frac{x \log \frac{x}{AB}}{C} + C_1 = \frac{x \log x}{C} - \frac{\log(AB)}{C}x + C_1$. Consider the objective function (4a) and constraint (4c) of the NEM problem and ignore the exact values of the constant parameters A, B, C , and C_1 . It is now clear that we can choose $z(x) = -x \log x$ as the objective function and there is no other format of $z(x)$ resulting in a flow that can be realized by link-state routing.

C. Entropy Maximization and Most Likely Flow Configuration

There are several intriguing relationships between the framework of network entropy maximization for link-state routing and statistical physics. We speculate about some of the thought-provoking connections in this Appendix.

In classical statistical mechanics, many microscopic behaviors aggregate into macroscopic states, and an isolated thermodynamic system will eventually reach an equilibrium macroscopic state that is the most likely one. Interestingly, entropy maximization for traffic engineering can be motivated by an argument of “most likely flow configuration,” shown as follows.

Consider a network with only one source–destination pair (s, t) and P uncapacitated paths between them. If there

are T packets to be transmitted from s to t , let $T_i \geq 0$ be the number of packets on path i , with $\sum_i T_i = T$. Each set of such $\{T_i\}$, which can be represented as a vector, is referred to as a *macroscopic state*. In contrast, each collection of routing decisions for individual packets represents a *microscopic state*. There are a total of P^T possible microscopic states. The number of microscopic states consistent with a given macroscopic state can be viewed as a measure of likelihood of that macroscopic state.

The number of microscopic states corresponding to the macroscopic state $\{T_i\}$ is $K = \frac{T!}{\prod_i T_i!}$. We want to search for the macroscopic state with the largest number of K , i.e., $\max K$, or, equivalently, $\max \log K = \max \log \frac{T!}{\prod_i T_i!}$. For a large system asymptote, T and T_i are large numbers. Hence, using Stirling's approximation, $n! \approx n^n e^{-n}$, we have $\log K \approx \log(e^{-T} T^T) - \sum_i \log(e^{-T_i} T_i^{T_i}) = -T \sum_i \frac{T_i}{T} \log \frac{T_i}{T}$.

This shows that the system equilibrium is the flow configuration that maximizes the entropy, $-\sum_i T x_i \log x_i$, where $x_i = \frac{T_i}{T}$ is the fraction of flow on path i .

The optimality result of PEFT through NEM suggests an intriguing connection between the *principle of entropy maximization* and that of *shortest description length* since maximizing entropy picks out those traffic distribution that can be realized by the simplest set of routing configuration parameters: one weight per link to be used independently by each router.

D. Proof of Lemma 1

Proof: Since strong duality holds for problem (4) and its Lagrange dual problem (7), we solve the dual problem through gradient method and recover the primal optimizers from the dual optimizers. By Danskin's Theorem [20]

$$\frac{\partial Q(\boldsymbol{\lambda}(q))}{\partial \lambda_{u,v}(q)} = \tilde{c}_{u,v} - \sum_{s,t,i} D(s,t) K_{P_{s,t}^{(u,v)}}^i x_{s,t}^i(q) \quad \forall (u,v) \in \mathbb{E}.$$

Hence, the algorithm in (9) is a gradient descent algorithm for dual problem (7). Since the dual objective function $Q(\boldsymbol{\lambda})$ is a convex function, there exists a step-size $\alpha(q)$ that guarantees $\boldsymbol{\lambda}(q)$ to converge to the optimal dual solutions $\boldsymbol{\lambda}^*$ [20]. Also, if $\nabla Q(\boldsymbol{\lambda})$ satisfies a Lipschitz continuity condition, i.e., there exists a constant $H > 0$ such that

$$\|\nabla Q(\boldsymbol{\lambda}_1) - \nabla Q(\boldsymbol{\lambda}_2)\| \leq H \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \quad \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \geq \mathbf{0}$$

then $\boldsymbol{\lambda}(q)$ converges to the optimal dual solution $\boldsymbol{\lambda}^*$ with a sufficiently small constant step-size $\alpha(q) = \alpha, 0 < \alpha < 2/H$ [20]. The Lipschitz continuity condition is satisfied if the curvatures of the entropy functions are bounded away from zero; see [29] for further details. Furthermore, since problem (4) is a strictly convex optimization problem and TRAFFIC-DISTRIBUTION problems (8) have unique solutions, \boldsymbol{x}^* are the globally optimal primal solutions of (4) [30]. ■

E. Tunnel-Based Routing to Realize Optimal TE

A tunnel-based routing can be derived from the optimal solution of the COMMODITY problem (2) based on dual decomposition. The approach follows the same way as the flow decomposition technique in [31]. We rephrase the approach and illus-

trate its complexity. The flow destined to the same destination is treated as a commodity. In the optimal solution of (2), there are up to N acyclic commodity flows, where N is the node number. The paths with flow can be determined for each commodity independently. For commodity t , starting with any source s , temporarily remove all the links without flow to t (i.e., $f_{u,v}^t = 0$). In the remaining network, choose any path from s to t , and let (u', v') be the link with the least $f_{u,v}^t$ along the path, then deduct $f_{u',v'}^t$ from demand $D(s, t)$ and flow $f_{u,v}^t$ for all the links along the path. Remove link (u', v') from further consideration. Repeat the above procedure until the paths for $D(s, t)$ have been determined. For each demand $D(s, t)$, there are at most E paths with flow since at least one link is removed during each step. Therefore, the total number of paths for N commodities (and $O(N^2)$ source/destination pair) is $O(N^2 E)$. Hence, the above procedure finishes within polynomial time.

F. Polynomial-Time Algorithm of Link Weight Setting for Single-Destination Network

For a single-destination (sink) network, the link weights to realize acyclic optimal TE with PEFT can be found in polynomial time. The method is much faster than solving the NEM problem with the gradient descent algorithm. We have the following lemma first.

Lemma 2: "Downward PEFT" can realize any acyclic flow for a single destination in polynomial time.

Proof: The links without flow can be assigned infinitely large weights and excluded from further processing. Denote $f_u^t = \sum_{(u,v) \in \mathbb{E}} f_{u,v}^t$, where $f_{u,v}^t$ is the amount of flow on link (u, v) . The nodes are processed in their reverse topological order in the acyclic flow, where the first node is the destination t , with $\Upsilon_t^t = 1$ (Section IV-C). When node u is processed, from (17), (18b), and (19), we have

$$f_{u,v}^t = f_u^t \frac{e^{-h_{u,v}^t} \Upsilon_v^t}{\Upsilon_u^t} \quad (36)$$

and

$$h_{u,v}^t = -\log \frac{f_{u,v}^t \Upsilon_u^t}{f_u^t \Upsilon_v^t} \geq 0 \quad (37)$$

then

$$\Upsilon_u^t \leq \frac{f_u^t \Upsilon_v^t}{f_{u,v}^t}. \quad (38)$$

We can set $\Upsilon_u^t = \min_{(u,v) \in \mathbb{E}} \frac{f_u^t \Upsilon_v^t}{f_{u,v}^t}$ since at least one link (u, v_0) is on the shortest path from u to t , i.e., $h_{u,v_0}^t = 0$. Then, we set the weight for link (u, v_0) as w_{\min} and the shortest distance from node u to t , $d_u^t = w_{\min} + d_{v_0}^t$. Then, the weight of link (u, v) is $-\log \frac{f_{u,v}^t \Upsilon_u^t}{f_u^t \Upsilon_v^t} + d_u^t - d_v^t$ from (37). It is easy to verify that the above link weighting satisfies the definition of downward PEFT (20)¹⁴ and the time complexity is $O(N + E)$.

Proof of Proposition 2:

Proof: An obvious conclusion from Lemma 2 if optimal TE is cycle-free. ■

¹⁴All d_v^t have been determined since the nodes are processed in the reverse topological order and $d_t^t \equiv 0$.

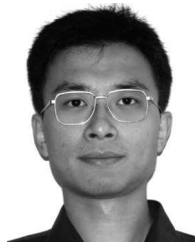
ACKNOWLEDGMENT

The authors appreciate the helpful discussions with D. Applegate, B. Fortz, J. He, J. Huang, D., Johnson, H. Karloff, Y. Li, J. Liu, M. Prytz, A. Tang, M. Thorup, J. Yu and J. Zhang.

REFERENCES

- [1] D. Xu, M. Chiang, and J. Rexford, "DEFT: Distributed exponentially-weighted flow splitting," in *Proc. IEEE INFOCOM*, Anchorage, AK, May 2007, pp. 71–79.
- [2] B. Fortz and M. Thorup, "Increasing Internet capacity using local search," *Comput. Optimiz. Appl.*, vol. 29, no. 1, pp. 13–48, 2004.
- [3] D. Awduche, "MPLS and traffic engineering in IP networks," *IEEE Commun. Mag.*, vol. 37, no. 12, pp. 42–47, Dec. 1999.
- [4] Z. Wang, Y. Wang, and L. Zhang, "Internet traffic engineering without full mesh overlaying," in *Proc. IEEE INFOCOM*, Anchorage, AK, 2001, vol. 1, pp. 565–571.
- [5] A. Sridharan, R. Guérin, and C. Diot, "Achieving near-optimal traffic engineering solutions for current OSPF/IS-IS networks," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 234–247, Apr. 2005.
- [6] S. Srivastava, G. Agrawal, M. Pioro, and D. Medhi, "Determining link weight system under various objectives for OSPF networks using a Lagrangian relaxation-based approach," *IEEE Trans on Network & Service Management*, vol. 2, no. 1, pp. 9–18, Nov. 2005.
- [7] J. H. Fong, A. C. Gilbert, S. Kannan, and M. J. Strauss, "Better alternatives to OSPF routing," *Algorithmica*, vol. 43, no. 1-2, pp. 113–131, 2005.
- [8] W. R. Blunden, *Introduction to Traffic Science*. London, U.K.: Printershall, 1967.
- [9] J. A. Tomlin and S. G. Tomlin, "Traffic distribution and entropy," *Nature*, vol. 220, pp. 974–976, 1968.
- [10] J. A. Tomlin, "A new paradigm for ranking pages on the world wide web," in *Proc. 12th WWW*, New York, 2003, pp. 350–355.
- [11] A. K. Agrawal, D. Mohan, and R. S. Singh, "Traffic planning in a constrained network using entropy maximisation approach," *J. Inst. Eng., India. Civil Eng. Div.*, vol. 85, pp. 236–240, 2005.
- [12] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.
- [13] H. Yäiche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 667–678, Oct. 2000.
- [14] S. H. Low, "A duality model of TCP and queue management algorithms," *IEEE/ACM Trans. Netw.*, vol. 11, no. 4, pp. 525–536, Aug. 2003.
- [15] R. Srikant, *The Mathematics of Internet Congestion Control (Systems and Control: Foundations and Applications)*. New York: Springer Verlag, 2004.
- [16] N. Garg and J. Könemann, "Faster and simpler algorithms for multicommodity flow and other fractional packing problems," *SIAM J. Comput.*, vol. 37, no. 2, pp. 630–652, 2007.
- [17] B. Awerbuch and R. Khandekar, "Distributed network monitoring and multicommodity flows: A primal-dual approach," in *Proc. 26th Annu. ACM PODC*, New York, 2007, pp. 284–291.
- [18] B. Fortz and M. Thorup, "Internet traffic engineering by optimizing OSPF weights," in *Proc. IEEE INFOCOM*, Tel-Aviv, Israel, 2000, pp. 519–528.
- [19] D. D. Sleator and R. E. Tarjan, "A data structure for dynamic trees," *J. Comput. Syst. Sci.*, vol. 26, no. 3, pp. 362–391, 1983.
- [20] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.
- [21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [22] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*. Cambridge, MA: MIT Press, 1990.
- [23] ILOG CPLEX optimizer IBM, Armonk, NY [Online]. Available: <http://www.ilog.com/products/cplex/>
- [24] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*. Danvers, MA: Boyd & Fraser, 1993.
- [25] "TOTEM," Univ. Catholique de Louvain, Louvain-la-Neuve, Belgium [Online]. Available: <http://totem.info.ucl.ac.be>
- [26] "Abilene backbone network," Internet2, Ann Arbor, MI [Online]. Available: <http://abilene.internet2.edu/>,

- [27] D. Xu, "Optimal traffic engineering via Newton's method," in *Proc. CISS*, Princeton, NJ, Mar. 2008, pp. 46–51.
- [28] S. Shenker, "Fundamental design issues for the future Internet," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1176–1188, Sep. 1995.
- [29] S. H. Low and D. E. Lapsley, "Optimization flow control—I: Basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, Dec. 1999.
- [30] M. Minoux, *Mathematical Programming: Theory and Algorithms*. New York: Wiley, 1986.
- [31] D. Mitra and K. G. Ramakrishnan, "A case study of multiservice multipriority traffic engineering design for data networks," in *Proc. IEEE GLOBECOM*, Rio de Janeiro, Brazil, Dec. 1999, pp. 1077–1083.
- [32] D. Lorenz, A. Orda, D. Raz, and Y. Shavitt, "How good can IP routing be?," DIMACS Rep. 2001-17, May 2001.



Dahai Xu (S'01–M'05) received the Ph.D. degree in computer science from the State University of New York at Buffalo in 2005.

He is currently a Research Staff Member with AT&T Laboratories—Research, Florham Park, NJ. After receiving the Ph.D. degree, he spent two years as a Post-Doctoral Research Associate with Princeton University, Princeton, NJ. His research interests include Internet design, control, and management; algorithm design and fast implementation; large-scale nonlinear network optimization; and

secure communication in wireless ad hoc networks.



Mung Chiang (S'00–M'03–SM'08) received the B.S. degree (Honors) in electrical engineering and mathematics and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1999, 2000, and 2003, respectively.

He is an Associate Professor of electrical engineering and an affiliated faculty member of Applied and Computational Mathematics and of Computer Science with Princeton University, Princeton, NJ. He was an Assistant Professor with Princeton University from 2003 to 2008. He has four U.S.

patents issued. His research areas include optimization, distributed control and stochastic analysis of communication networks, with applications to the Internet, wireless networks, broadband access networks, content distribution, and network economics. He founded the Princeton EDGE Lab in 2009 (<http://scenic.princeton.edu>).

Dr. Chiang co-chaired the 38th Conference on Information Sciences and Systems, the 9th IEEE WiOpt Conference. He received the following awards: the Presidential Early Career Award for Scientists and Engineers in 2008 from the White House, the TR35 Young Innovator Award in 2007 from Technology Review, the Young Investigator Award in 2007 from the Office of Naval Research (ONR), the Young Researcher Award Runner-up 2004–2007 from the Mathematical Programming Society, the CAREER Award in 2005 from the National Science Foundation (NSF), as well as Frontiers of Engineering Symposium participant in 2008 from the National Academy of Engineering (NAE) and an Engineering Teaching Commendation in 2007 from Princeton University. He was a Princeton University Howard B. Wentz Junior Faculty and a Hertz Foundation Fellow. His paper awards include the ISI citation Fast Breaking Paper in Computer Science and the IEEE GLOBECOM Best Paper Award three times.



Jennifer Rexford (S'89–M'96–SM'01) received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 1991, and the M.S.E. and Ph.D. degrees in computer science and electrical engineering from the University of Michigan, Ann Arbor, in 1993 and 1996, respectively.

She is a Professor with the Computer Science Department, Princeton University. From 1996 to 2004, she was a member of the Network Management and Performance Department, AT&T Laboratories—Research, Florham Park, NJ. She is coauthor of the book *Web Protocols and Practice* (Addison-Wesley, 2001).

Prof. Rexford served as the Chair of ACM SIGCOMM from 2003 to 2007. She was the 2004 winner of the ACM's Grace Murray Hopper Award for an outstanding young computer professional.