Stochastic Traffic Engineering, with Applications to Network Revenue Management

Debasis Mitra Qiong Wang
Bell Laboratories, Lucent Technologies, Murray Hill NJ 07974
{mitra,chiwang}@lucent.com

Abstract—**We present a stochastic traffic engineering framework for optimizing bandwidth provisioning and path selection in networks. The objective is to maximize revenue from serving demands, which are uncertain and specified by probability distributions. We consider a two-tier market structure, where demands in the two markets are associated with different unit revenues and uncertainties. Based on mean-risk analysis, the optimization model enables a carrier to maximize mean revenue and contain the risk that the revenue falls below an acceptable level. Our framework is intended for off-line traffic engineering design, which takes a centralized view of network topology, link capacity, and demand. We obtain conditions under which the optimization problem is an instance of convex programming and therefore efficiently solvable. We derive properties of the optimal solution for the special case of Gaussian distributions of demands. We focus on the impact of demand variability on various aspects of traffic engineering, such as link utilization, routing, capacity provisioning, and total revenue.**

Method Keywords- **Mathematical Programming, Economics, Traffic Engineering, Demand Uncertainty, Risk**

I. Introduction

Traffic engineering (TE) is a mechanism for traffic and revenue management in networks [4]. Supported by emerging technologies, especially Multi-Path Label Switching (MPLS), TE performs provisioning and admission control functions to optimize network operators' objectives [3], [8], [7]. The TE mechanism takes two complementary forms, on-line and off-line [4], [23]. On-line TE is state-dependent and applies on a short time-scale. See [21] and [12] for works where the focus is on distributed traffic engineering and provisioning. Off-line TE applies on a longer time-scale. Instead of focusing on instantaneous network states and individual connections, the latter mechanism considers statistical behavior of traffic demands aggregated over all connections. Combining this demand information with a centralized view of network topology and link capacites, off-line TE selects the topology of routes and provisions resources on the selected routes for carrying the demands. These decisions are optimized globally for demands of various service types and origin-destination pairs [2], [16], [19], [22]. The solution of the off-line optimization has been proposed as a reference point for on-line operations. For example, in [22] capacity preallocated by the off-line TE process is used as the threshold for on-line admission control. Similarly, in [26] the off-line provisioning process is used to guide the real-time routing and admission control.

This paper focuses on the optimization of off-line traffic engineering. In previous work, the problem has been formulated as a deterministic multi-commodity flow (MCF) model, where demand of each service and node pair is given as a fixed quantity, such as expected value of forecasted traffic load [16], [22]. The goal is to find an appropriate amount of traffic to admit for each demand, and capacitated route(s) to carry the traffic, so that the carrier's objective, usually formulated as revenue earned by serving demands, is optimized subject to network capacity constraints.

There is much to be gained from a framework that takes into account uncertainty about demand and revenue. For instance, in the deterministic MCF model, revenue derived from carrying demand is assumed to increase linearly with the amount of capacity provisioned up to the point where all traffic demand is satisfied. This approach is incomplete in the presence of uncertainty in demand. When demand is random, if more capacity is provisioned then the probability that the incremental capacity will be fully utilized decreases. Consequently, the mean revenue should be a concave function of the allocated capacity. This non-linear effect is captured under the framework of probabilistic distributions of demands. Such distributional information is typically a byproduct of statistical procedures for forecasting network traffic from measurements [5], [6], [10], [24]. While this information is obtainable, it has not been used extensively in the past for the lack of a modeling framework.

In this paper, we develop a stochastic traffic engineering framework which uses probabilistic distributions of demands as inputs for off-line optimization. The model captures the aforementioned non-linear effect and also extends off-line traffic engineering in several interesting dimensions. First, the model enables us to assess the impact of statistical variability in demand. In our numerical studies, we find that, even when mean demands stay fixed, increasing demand variability significantly affects the optimal traffic engineered solution. We have observed lower link utilization and higher link shadow costs. Note that the optimal routing is based on link shadow costs [19], so variability also influences path selection.

Second, the use of distributional information of demands suggests extension of the objective function of traffic engineering beyond mean revenue. For example, the carrier may be interested not only in the mean revenue, but also the risk that the revenue will fall below an acceptable level. This risk is not calculable in the deterministic traffic engineering framework. It is, however, a major output of our stochastic traffic engineering framework. Finally, the framework allows the carrier to consider multiple objectives and their tradeoffs, e.g., maximizing the expected value and minimizing the risk of revenue shortfall. In fact, the optimization model developed in this paper is based on the mean-risk analysis approach, which was first developed in the finance community to address the needs of balancing growth and risk in resource allocation [15], [11], [13].

An important aspect of our model is the formulation of demand and revenue under a two-tier market structure. The first

The authors acknowledge K. G. Ramakrishnan for his contributions to the early development of this work, and Alan Weiss for the benefit of discussions.

tier is a wholesale market where bandwidth is sold as a commodity, and the second tier is a retail market where bandwidth is sold as a service. In the wholesale market demand is deterministic, so that there is no risk in revenue generation, but the price for carrying each unit of bandwidth is low. In the retail market, the carrier can charge a premium price, but the demand is stochastic, so that there exists the risk of revenue shortfall. By properly dimensioning bandwidth provisioned for each market, the carrier can maximize mean revenue at its acceptable risk level. Our numerical studies show that these decisions are affected by demand variability. As variability increases, more bandwidth is provisioned to satisfy stochastic retail demands and less for wholesale. Such a two-tier market structure is not uncommon in the telephone industry, but has not yet been fully developed in data transport. It is important to point out that in our modeling framework the existence of the wholesale market may be readily removed, in which case the stochastic demand in the retail market becomes the sole focus. On the other hand, we can further generalize the carrier's interaction with the wholesale market by assuming that it can both buy and sell (here we only allow selling) bandwidth in this market. We choose to defer this extension to a future publication in order to focus on the many aspects of the simpler model here.

The paper is organized as follows: In Section II we formulate the stochastic traffic engineering problem and present the optimization model. In Section III we discuss the complexity of obtaining the global optimal solution. In Section IV we derive properties of the optimal solution that provide important insights on bandwidth provisioning and routing. In Section V we give numerical results that show the impact of demand variability on network traffic engineering. We present our concluding remarks in Section VI.

II. Model

A. Conception

A.1 Demand for Bandwidth

A carrier derives its revenue by delivering traffic demand from one node to another for its clients. We assume that there are two markets for the carrier's transmission capability: a wholesale market where bandwidth is sold as a commodity; and a retail market where bandwidth is sold as a service.

In the wholesale market, carriers sell standardized "bandwidth pipes", e.g., DS3 circuits, that have little differentiation among the vendors. Naturally, only carriers that offers the lowest price per unit of bandwidth will be selected by buyers. Consequently, the competition in the wholesale market is best charaterized by a Cournot game, where every carrier charges the market prevailing price that is jointly determined by total capacities provided by all carriers. In this paper, we assume that there are many competing suppliers in the wholesale market, so that an individual carrier cannot dominate demands, and the prevailing market price cannot be influenced significantly by one carrier's tactical decision about how much capacity to bring to the wholesale market. The assumption implies that within the short time scale of our consideration, the carrier can consider as guaranteed the wholesale revenue, which equals the product of the market prevailing price per unit of bandwidth and the amount of bandwidth it chooses to sell.

In the retail market, rather than offering "raw bandwidth," carriers sell services, such as voice, data, and video, based on the delivery of bandwidth of course. Consequently, customers choose their preferred carrier not only on the basis of price, but also other factors, such as the richness of applications and content that are in the carrier's service package, as well as quality at the application level (e.g. clarity of voice or video image), security and reliability, and even brand name. These features differentiate carriers and make some more attractive than others to certain segments of customers. As differentiation renders more market power to a carrier over its targeted users, carriers are allowed to collect a higher price per unit of bandwidth from the retail market than from the wholesale market. However, the revenue also becomes more volatile in the retail market because, instead of serving a small portion of total industry demand, as is the case with the wholesale market, the carrier now serves a smaller group of users, whose demand is specific to its network service.

We digress to observe that in the model here prices are not influenced by capacity provisioning decisions. It is possible to couple both through price-demand relationships, for instance, as in [19].

A.2 Admissible Route Sets

QoS and policy considerations are major constraints on provisioning decisions. The notion of *admissible route sets* allows these constraints to be taken into account in the optimization. Let $\mathcal{R}(v)$ denote the set of admissible routes for origin-destination pair v. Different admissible route sets for wholesale and retail services between the same node pair are allowed. For example, routes may be required to have lengths not exceeding specified thresholds, on account of propagation delay, and there may also be restrictions on the number of hops, since each hop is associated with addition switching node and consequent incremental delay. The admissibility of a route may also depend on policy, which might reflect diverse considerations, such as security, the capability of switching nodes in the routes to handle certain services, link capacity, etc. Generating the admissible route sets is a substantial task in itself. In this paper, as in [16], we consider that these sets are given.

A.3 Mean-Risk Analysis

In the presence of demand uncertainty, maximizing the mean revenue, which is implied in the earlier deterministic traffic engineering studies, may not be the best approach for a carrier. For example, suppose one TE design gives the carrier revenue of 10000 with probability 0.01, while another TE design gives, with certainty, revenue of 99. Then many carriers might choose the second design even though the mean revenue is lower. This preference for avoiding uncertainty in revenue is formally defined as *risk averseness* in decision science [11].

Mean-risk analysis, which has been widely applied to financial asset allocation, addresses the issue of risk averseness by offering a broader optimization objective. The approach starts with developing a risk index, which is a quantitative measure of the risk of revenue shortfall, based on the distributional information. It then maximizes the weighted combination of the mean

revenue and the risk index, i.e., $mean - \delta * (risk\ index)$, where $\delta \geq 0$ is a parameter. Different levels of risk averseness can be reflected by choosing different values for δ. A high δ value indicates larger willingness to sacrifice the mean revenue to avoid risk. On the other hand, by setting δ to 0, we can also include the case that the carrier maximizes the mean revenue only. Notice that, for any value of δ, the solution that maximizes the above combination is Pareto optimal in the sense that one cannot improve the mean revenue without increasing the risk, or reduce the risk without hurting the mean.

There are several candidates for the risk index, many of which may lead to inferior solutions that are stochastically dominated by other feasible alternatives. See [13] and [20] for definitions of stochastic dominance and their relevance to the mean-risk model. As discussed in [18], there is an important tradeoff between immunity from stochastic dominance of the optimal solution and tractability of the optimization problem in network revenue management. For example, variance is a widely used risk index. In our case, it facilitates the optimization, but the solutions obtained are susceptible to stochastic dominance. Based on these considerations, we find that the standard deviation of network revenue is an appropriate risk index for our optimization framework.

The reader may wish to take note of two popular approaches for handling risk, namely, the von Neuman - Morgenstern [25], [11] expected utility model and the mean-variance model due to Markowitz [15]. Also, robust optimization [17] is a technique related to the latter approach, which proactively addresses uncertainties in the model data of engineering systems.

B. Model Formulation

We formulate the network as a collection of nodes and links $(\mathcal{N}, \mathcal{L})$, where link $l \in \mathcal{L}$ has bandwidth C_l. Let $\mathcal{V} = \{(v_i, v_j) : v_i \in \mathcal{N}, v_j \in \mathcal{N}\}$ be the set of all node pairs. $\mathcal{V}_1 \subset \mathcal{V}$ is the collection of node pairs between which there are retail demands, and $\mathcal{V}_2 \subset \mathcal{V}$ is the collection of node pairs between which wholesale of bandwidth is admissible.

Retail demand between $v \in \mathcal{V}_1$ is characterized by a random distribution, with its Probability Density Function (PDF) denoted by $f_v(x)$, and Cumulative Distribution Function (CDF) denoted by $F_v(x)$. Let $d_v (v \in \mathcal{V}_1)$ be the amount of capacity provisioned to serve retail demand between v. The provisioned quantity, d_v, can be routed on one or more admissible routes. Denote the admissible route set for $v \in \mathcal{V}_1$ by $\mathcal{R}_1(v)$ and let $\xi_r (r \in \mathcal{R}_1(v))$ be the amount of capacity provisioned on route r. Then:

$$d_v = \sum_{r \in \mathcal{R}_1(v)} \xi_r. \tag{1}$$

Denote by T_v the random retail demand between node pair v. Then

$$x_v(d_v) = \min(T_v, d_v) \tag{2}$$

is the amount of retail demand that is actually carried between v. Note that the revenue earned by the carrier is based on the carried demand. Let $m_v(d_v)$ and $s_v^2(d_v)$ be the mean and variance of x_v, respectively. Then

$$m_v(d_v) = \int_0^{d_v} x f_v(x) dx + d_v \bar{F}_v(d_v) = \int_0^{d_v} \bar{F}_v(x) dx \tag{3}$$

and

$$\begin{aligned} s_v^2(d_v) &= \int_0^{d_v} x^2 f_v(x) dx + d_v^2 \bar{F}_v(d_v) - m_v^2(d_v) \\ &= 2 \int_0^{d_v} x \bar{F}_v(x) dx - m_v^2(d_v), \end{aligned} \tag{4}$$

where $\bar{F}_v(x) \equiv 1 - F_v(x)$. Notice that

$$\frac{\partial m_v}{\partial d_v} = \bar{F}_v(d_v) \geq 0,$$

$$\frac{\partial s_v^2}{\partial d_v} = 2[d_v - m_v(d_v)] \bar{F}_v(d_v) \geq 0,$$

i.e., both the mean and variance of carried demand increase with the amount of bandwidth provisioned. Their maximum values, denoted by $m_v(\infty)$ and $s_v^2(\infty)$, are the mean and variance of the demand between v, respectively. Let π_v be the revenue earned for each unit of retail traffic carried between v. The total revenue derived from serving retail demand between v is a random variable $\pi_v x_v(d_v)$, for which the mean is $\pi_v m_v(d_v)$ and the variance is $\pi_v^2 s_v^2(d_v)$.

Likewise, let y_v be the amount of bandwidth provisioned for wholesale between $v \in \mathcal{V}_2$,

$$y_v = \sum_{r \in \mathcal{R}_2(v)} \phi_r, \tag{5}$$

where $\mathcal{R}_2(v)$ is the admissible route set for $v \in \mathcal{V}_2$ and ϕ_r is the amount of provisioned bandwidth on route r to carry wholesale traffic. Suppose e_v is the unit wholesale price between node pair v. Then the wholesale revenue is $e_v y_v$.

The carrier's total revenue is the following function of decision variables d_v and y_v:

$$W = \sum_{v \in \mathcal{V}_1} \pi_v x_v + \sum_{v \in \mathcal{V}_2} e_v y_v. \tag{6}$$

We invoke the mean-risk framework, and assume that the carrier wants to balance maximization of the mean revenue, $E(W)$, with minimization of the risk of revenue shortfall, with the latter represented by the standard deviation, $\sqrt{Var(W)}$. Thus the objective function is formulated as:

$$\Theta = E(W) - \delta \sqrt{Var(W)} \tag{7}$$

where $\delta \geq 0$ is an input parameter that reflects the extent to which the carrier is willing to trade expected revenue with the risk of revenue loss.

The overall optimization model is as follows:

$$\begin{aligned} &\max \Theta(d_v, y_v, \xi_r, \phi_r) = E(W) - \delta \sqrt{Var(W)} \\ &= \sum_{v \in \mathcal{V}_1} \pi_v m_v(d_v) + \sum_{v \in \mathcal{V}_2} e_v y_v - \delta \sqrt{\sum_{v \in \mathcal{V}_1} \pi_v^2 s_v^2(d_v)} \end{aligned} \tag{8}$$

subject to:

$$\sum_{r \in \mathcal{R}_1(v)} \xi_r = d_v \ (v \in \mathcal{V}_1) \qquad \sum_{r \in \mathcal{R}_2(v)} \phi_r = y_v \ (v \in \mathcal{V}_2) \tag{9}$$

$$\sum_{r\in\mathcal{R}_1(v):l\in r}\xi_r+\sum_{r\in\mathcal{R}_2(v):l\in r}\phi_r\le C_l \qquad l\in\mathcal{L} \quad (10)$$

$$0<\bar{d}_v\le d_v\ (v\in\mathcal{V}_1), \quad (11)$$

$$0\le\xi_r\ (r\in\mathcal{R}_1(v)), \quad (12)$$

$$0\le y_v\ (v\in\mathcal{V}_2),\quad 0\le\phi_r\ (r\in\mathcal{R}_2(v)) \quad (13)$$

Here, we comment on parameter $\bar{d}_v$ in (11), which is defined as the minimum amount of bandwidth that must be provisioned for retail demand between v. The value of $\bar{d}_v$ is determined by the grade of service that the carrier offers to its retail customers. For example, the requirement that the blocking rate of a retail demand be kept below some threshold can be translated into the condition that d_v be greater than a suitably chosen value of $\bar{d}_v$. Notice that with fixed link capacities, it may be infeasible to satisfy the minimum amount requirement of some demands. Should this situation arise, the carrier has to supplement its installed capacities by buying bandwidth from other carriers. As discussed in Section I, our model can be extended to cover this scenario. Let p_l be the price of buying unit bandwidth on link l, and b_l be the amount to buy, which is a decision variable. We can integrate buying decisions into the model by generalizing the revenue function in (6) to be

$$\hat{W}=W-\sum_{l\in\mathcal{L}}p_l b_l,$$

and replacing W with $\hat{W}$ in the objective function (7). Constraints in (10) should also be changed to

$$\sum_{r\in\mathcal{R}_1(v):l\in r}\xi_r+\sum_{r\in\mathcal{R}_2(v)}\phi_r\le C_l+b_l \qquad l\in\mathcal{L} \quad (14)$$

The generalized model makes it always feasible to provide minimum quantity guarantee for retail demands and its output recommends the optimal amount of bandwidth to buy on each link. Nevertheless, the extended model that incorporates the carrier's buying decisions is not the focus of this paper, as has been mentioned in Section I. Consequently, we will analyze the original model in (8)-(13), and assume that the guaranteed grade of service is such that a feasible solution exists for the existing link capacities.

III. Solution Strategy

Given that all constraints of the above model are linear, the difficulty of finding the global optimum depends on the shape of the objective function Θ. If Θ is concave in all nonlinear variables $d_v(v\in\mathcal{V}_1)$, the model falls into the class of concave maximization problem. In this case, the global optimum can be found efficiently with existing standard algorithms, which facilitates the implementation of our framework.

In general, Θ is not concave everywhere if $\delta>0$, as can be verified by considering a restricted case with only one nonlinear variable. Despite this inconvenience, in this section, we show that in many circumstances, the model can still be solved as a concave maximization problem. Our approach is based on two theorems developed in III-A and the subsequent analysis in III-B.

A. The Shape of the Objective Function

In the following, we use m_v and s_v^2 to represent $m_v(d_v)$ and $s_v^2(d_v)$, as defined in equations (3) and (4), respectively. We also omit the argument d_v in the distribution functions, and use F_v, $\bar{F}_v$, and f_v to represent $F_v(d_v)$, $\bar{F}_v(d_v)$, and $f_v(d_v)$, respectively. Lemma 1 is the basis for the proofs of Theorems 1 and 2.

Lemma 1 For any $v\in\mathcal{V}_1$ and $d_v\ge 0$,

$$F_v s_v^2\ge(d_v-m_v)^2\bar{F}_v. \quad (15)$$

Proof: Because $f_v\ge 0$, from equations (3) and (4):

$$\begin{aligned}\frac{\partial(F_v s_v^2)}{\partial d_v} &= 2F_v\bar{F}_v(d_v-m_v)+f_v s_v^2\\ &\ge 2F_v\bar{F}_v(d_v-m_v)-f_v(d_v-m_v)^2\\ &= \frac{\partial[(d_v-m_v)^2\bar{F}_v]}{\partial d_v}.\end{aligned} \quad (16)$$

Since $F_v s_v^2=(d_v-m_v)^2\bar{F}_v=0$ at $d_v=0$,

$$F_v s_v^2\ge(d_v-m_v)^2\bar{F}_v \quad \text{for all} \quad d_v\ge 0. \ \square \quad (17)$$

We show in Theorem 1 that Θ is unimodal in every d_v.

Theorem 1 For $v\in\mathcal{V}_1$, given fixed values for $d_{v'}(v'\ne v)$,

$$\frac{\partial\Theta}{\partial d_v}=\begin{cases}\ge 0 & \text{if } 0\le d_v\le\hat{d}_v\\ <0 & \text{if } d_v>\hat{d}_v\end{cases}, \quad (18)$$

$$\hat{d}_v=\sup\{d_v:\ \frac{(d_v-m_v)^2}{s_v^2+\Psi_v}\le\frac{1}{\delta^2}\}. \quad (19)$$

where $\Psi_v=\sum_{v'\ne v}(\pi_{v'}/\pi_v)^2 s_{v'}^2$, and $\hat{d}_v$ is the unique solution to (19), i.e., $\hat{d}_v$ is calculated by solving the equation obtained by replacing $\le$ by $=$.

Proof: Let $S^2=Var(W)=\sum_{v\in\mathcal{V}_1}\pi_v^2 s_v^2$,

$$\begin{aligned}\frac{\partial\Theta}{\partial d_v} &= \pi_v\frac{\partial m_v}{\partial d_v}-\delta\frac{\partial S}{\partial d_v}=\pi_v\bar{F}_v-\delta\pi_v\bar{F}_v\frac{(d_v-m_v)}{S/\pi_v}\\ &= \pi_v\bar{F}_v[1-\delta\frac{(d_v-m_v)}{\sqrt{s_v^2+\Psi_v}}].\end{aligned} \quad (20)$$

Notice that by Lemma 1,

$$\partial[\frac{(d_v-m_v)}{\sqrt{s_v^2+\Psi_v}}]/\partial d_v=\frac{F_v(s_v^2+\Psi_v)-(d_v-m_v)^2\bar{F}_v}{\sqrt{(s_v^2+\Psi_v)^3}}\ge 0. \quad (21)$$

Therefore $(d_v-m_v)/\sqrt{s_v^2+\Psi_v}]$ monotonically increases from 0 to $+\infty$ as d_v goes from 0 to $+\infty$. It follows that $\hat{d}_v$ as defined by (19) is unique, and $\partial\Theta/\partial d_v\ >,=,$ or <0, depending on whether $d_v<,=,$ or $>\hat{d}_v$. $\square$

Holding (19) at equality and applying the Implicit Function Theorem,

$$\frac{\partial\hat{d}_v}{\partial\Psi_v}=\frac{d_v-m_v}{F_v(s_v^2+\Psi_v)-(d_v-m_v)^2\bar{F}_v}\ge 0, \quad (22)$$

indicating that $\hat{d}_v$ increases with Ψ_v. This result will be used in III-B.

Clearly, any maximum point of Θ can be reached only in areas where $d_v \le \hat{d}_v$ for all $v \in \mathcal{V}_1$. Theorem 2 shows that Θ is concave in this region. Before presenting the theorem, we first specify the second-order derivatives,

$$\frac{\partial^2\Theta}{\partial d_v^2} = -\pi_v f_v[1-\delta\frac{(d_v-m_v)}{S/\pi_v}] \\ -\delta\frac{\pi_v\bar{F}_v}{S/\pi_v}[F_v-(\frac{(d_v-m_v)}{S/\pi_v})^2\bar{F}_v] \;\; v\in\mathcal{V}_1, \tag{23}$$

and for $v \neq v'$,

$$\frac{\partial^2\Theta}{\partial d_v \partial d_{v'}} = \frac{\delta}{S^3}\pi_v^2\pi_{v'}^2(d_v-m_v)\bar{F}_v(d_{v'}-m_{v'})\bar{F}_{v'} \tag{24}$$

Theorem 2 : Let $H(\Theta)$ be the Hessian matrix of Θ.
If

$$\frac{\partial\Theta}{\partial d_v} = \pi_v\bar{F}_v[1-\delta\frac{(d_v-m_v)}{S/\pi_v}] \ge 0 \text{ for all } v\in\mathcal{V}_1 \tag{25}$$

then $H(\Theta)$ is negative semi-definite.
Proof:

$$H(\Theta) = \begin{pmatrix} H_1 & 0 \\ 0 & 0 \end{pmatrix} \tag{26}$$

where H_1 is a square matrix of dimension $n = |\mathcal{V}_1|$, and its entries are $\partial^2\Theta/\partial d_{v_i}\partial d_{v_j}(v_i, v_j \in \mathcal{V}_1)$. Other elements that take the zero value correspond to second-order derivatives with respect to linear variables in the objective function. $H(\Theta)$ is negative semi-definite if and only if H_1 is negative semi-definite.

Apply equations (23) and (24), and note that
$S^2 = \pi_v^2 s_v^2 + \sum_{v'\neq v}\pi_{v'}^2 s_{v'}^2$ for any $v \in \mathcal{V}_1$, we have

$$H_1 = -H_{1a} - \frac{\delta}{S^3}H_{1b} - \frac{\delta}{S^3}H_{1c}, \tag{27}$$

where elements of matrix H_{1a}, H_{1b}, and H_{1c}, denoted correspondingly as $h_a(i,j)$, $h_b(i,j)$, and $h_c(i,j)$. H_{1a} and H_{1b} are diagonal matrices with

$$h_a(i,i) = \pi_{v_i}f_{v_i}[1-\delta\frac{(d_{v_i}-m_{v_i})}{S/\pi_{v_i}}], \text{ and} \\ h_b(i,i) = \pi_{v_i}^4\bar{F}_{v_i}[s_{v_i}^2F_{v_i} - (d_{v_i}-m_{v_i})^2\bar{F}_{v_i}]. \tag{28}$$

$$h_c(i,i) = \pi_{v_i}^2\bar{F}_{v_i}F_{v_i}(\sum_{v\neq v_i}\pi_v^2 s_v^2), \text{ and for } i\neq j, \\ h_c(i,j) = -\pi_{v_i}^2\pi_{v_j}^2(d_{v_i}-m_{v_i})\bar{F}_{v_i}(d_{v_j}-m_{v_j})\bar{F}_{v_j}. \tag{29}$$

H_{1a} is positive semi-definite by assumption, and H_{1b} is positive semi-definite by Lemma 1. To prove that $H(\Theta)$ is negative semi-definite, it suffices to show that H_{1c} is positive semi-definite, i.e., for any real vector $\vec{X}$ of dimension $n = |\mathcal{V}_1|$,

$$\vec{X}^T H_{1c}\vec{X} = \sum_{v\in\mathcal{V}_1} x_v^2\pi_v^2\bar{F}_vF_v(\sum_{v'\neq v}\pi_{v'}^2 s_{v'}^2) \\ -\sum_{v'\neq v} x_v x_{v'}\pi_v^2\pi_{v'}^2(d_v-m_v)\bar{F}_v(d_{v'}-m_{v'})\bar{F}_{v'} \ge 0. \tag{30}$$

This is true because by Lemma 1,

$$\sqrt{\bar{F}_vF_v\bar{F}_{v'}F_{v'}s_v^2s_{v'}^2} \ge (d_v-m_v)\bar{F}_v(d_{v'}-m_v)\bar{F}_{v'}, \tag{31}$$

and by the A-G Mean Inequality($a^2+b^2 \ge 2ab$),

$$\sum_v x_v^2\pi_v^2\bar{F}_vF_v(\sum_{v'\neq v}\pi_{v'}^2 s_{v'}^2) \\ \ge \sum_{v'\neq v}\pi_v^2\pi_{v'}^2 x_v x_{v'}\sqrt{\bar{F}_vF_v\bar{F}_{v'}F_{v'}s_v^2s_{v'}^2}. \;\; \square \tag{32}$$

B. Discussions on Solution Procedures

The two theorems in the last section define a set:

$$\Omega \equiv \{d_v(v\in\mathcal{V}_1) : \bar{d}_v \le d_v \le \hat{d}_v\}, \tag{33}$$

over which Θ is concave. The set Ω also contains all local maximum point(s) of Θ. The model is a concave maximization problem if Ω, or its intersection with the feasible region, is a convex set. In this case, we can find the global optimum efficiently.

The set Ω may or may not be convex, depending on demand distribution functions. In case Ω is not convex, we can still solve the model as a concave maximization problem under certain conditions. Consider $\hat{d}_v(v\in\mathcal{V}_1)$ in Theorem 1, which are obtained by solving:

$$\frac{[\hat{d}_v - m_v(\hat{d}_v)]^2}{s_v^2(\hat{d}_v)+\Psi_v} = \frac{1}{\delta^2}, \text{ where } \Psi_v = \sum_{v'\neq v}(\frac{\pi_{v'}}{\pi_v})^2 s_{v'}^2(d_{v'}). \tag{34}$$

Since $\bar{d}_v > 0$ is the minimum amount of guaranteed bandwidth for v, and $s_{v'}^2(d_{v'})$ increases with $d_{v'}$,

$$\bar{\Psi}_v = \sum_{v'\neq v}(\frac{\pi_{v'}}{\pi_v})^2 s_{v'}^2(\bar{d}_{v'})$$

is the lower bound of Ψ_v. Suppose that for each v, we solve (34) by using $\Psi_v = \bar{\Psi}_v$ and denote the solution by $\hat{d}_v^e$, i.e.,

$$\frac{[\hat{d}_v^e - m_v(\hat{d}_v^e)]^2}{s_v^2(\hat{d}_v^e)+\bar{\Psi}_v} = \frac{1}{\delta^2}. \tag{35}$$

Then $\hat{d}_v^e \le \hat{d}_v$ because as mentioned in III-A, $\hat{d}_v$ that solves (34) increases with Ψ_v. Consequently, the polyhedra

$$\Omega^e \equiv \{d_v(v\in\mathcal{V}_1) : \bar{d}_v \le d_v \le \hat{d}_v^e\}$$

is a subset of Ω. If Ω^e is large enough to contain the global optimum, the model can be solved by exercising a concave maximization algorithm on the polyhedra.

We now discuss conditions for the global optimum to be "trapped" inside Ω^e. Define

$$k_v = \frac{\bar{\Psi}_v}{s_{v,\infty}^2},$$

where $1/k_v$ is the ratio of the maximum variance of the revenue from node pair v to the minimum variance of the revenue from all other node pairs. Because $s_{v,\infty}^2 \ge s^2(\hat{d}_v^e)$,

$$\frac{1}{\delta^2} = \frac{[\hat{d}_v^e - m_v(\hat{d}_v^e)]^2}{s_v^2(\hat{d}_v^e)+\bar{\Psi}_v} \le \frac{[\hat{d}_v^e - m_v(\hat{d}_v^e)]^2}{(k+1)s_v^2(\hat{d}_v^e)}.$$

By Lemma 1,

$$\frac{1}{\delta^2} = \frac{[\hat{d}_v^e - m_v(\hat{d}_v^e)]^2}{s_v^2(\hat{d}_v^e) + \bar{\Psi}_v} \leq \frac{F_v(\hat{d}_v^e)}{(k+1)\bar{F}_v(\hat{d}_v^e)},$$

so

$$F_v(\hat{d}_v^e) \geq \frac{k_v + 1}{k_v + 1 + \delta^2}.$$

If $k_v \gg \delta^2$ for each v, then $F_v(\hat{d}_v^e) \approx 1$, and it becomes certain that Ω^e contains the global optimum. For example, when $k_v = 20$, $\delta = 1$, then $F_v(\hat{d}_v^e) \geq 0.95$. This means that within Ω^e, the amount of bandwidth provisioned to serve a demand can vary from the minimum amount required to 95% quantile of the demand distribution. For most networks, a solution that maximizes revenue will fall into that region. Especially when wholesale is allowed and the wholesale price is more than 5% of retail price, exceeding the expected incremental revenue that the carrier can get by provisioning bandwidth beyond the 95% quantile. Therefore, the model can be optimized by performing concave maximization over Ω^e.

For the condition $k_v \gg \delta^2$ to hold, either δ is small (in an extreme case when $\delta = 0$, $\hat{d}_v^e = \infty$), or k_v is large. The latter corresponds to situations where the contribution of each node pair to the variance of total revenue is insignificant. This happens when the network has many node pairs, and the total revenue is not dominated by the revenue from an individul pair, which is the case that we consider in subsequent analysis.

If the above condition does not apply, the model becomes a global optimization problem, which no general algorithm can solve efficiently. Development of specialized procedures is under consideration.

IV. Analysis

A. Necessary Condition and Its Implications

In this section, we apply a Lagrangian method and discuss the implications for the optimal solution based on the first-order necessary condition.

The Lagrangian takes the following form:

$$\begin{aligned}\Lambda &= \sum_{v\in\mathcal{V}_1} \pi_v m_v + \sum_{v\in\mathcal{V}_2} e_v y_v - \delta\sqrt{\sum_{v\in\mathcal{V}_1} \pi_v^2 s_v^2} \\ &+ \sum_{v\in\mathcal{V}_1} \chi_v^1 \Big(\sum_{r\in\mathcal{R}_1(v)} \xi_r - d_v\Big) + \sum_{v\in\mathcal{V}_2} \chi_v^2 \Big(\sum_{r\in\mathcal{R}_2(v)} \phi_r - y_v\Big) \\ &+ \sum_{l\in\mathcal{L}} \lambda_l \Big(c_l - \sum_{r\in\mathcal{R}_1(v): l\in r} \xi_r - \sum_{r\in\mathcal{R}_2(v): l\in r} \phi_r\Big) \qquad (36)\end{aligned}$$

It follows that the first-order necessary conditions are (using results in (20)):

$$\left.\begin{aligned} &\partial\Lambda/\partial d_v = \pi_v \bar{F}_v[1 - \delta\tfrac{\pi_v(d_v - m_v)}{S}] - \chi_v^1 \leq 0 \\ &d_v(\partial\Lambda/\partial d_v) = 0,\ \ d_v \geq 0,\ \ \chi_v^1 \geq 0\ \ (v \in \mathcal{V}_1) \end{aligned}\right\} \qquad (37)$$

$$\left.\begin{aligned} &\partial\Lambda/\partial y_v = e_v - \chi_v^2 \leq 0 \\ &y_v(\partial\Lambda/\partial y_v) = 0,\ y_v \geq 0,\ \ \chi_v^2 \geq 0\ \ (v \in \mathcal{V}_2) \end{aligned}\right\} \qquad (38)$$

$$\left.\begin{aligned} &\partial\Lambda/\partial \xi_r = \chi_v^1 - \textstyle\sum_{l:l\in r} \lambda_l \leq 0 \\ &\xi_r(\partial\Lambda/\partial\xi_r) = 0,\ \xi_r \geq 0,\ \lambda_l \leq 0\ (r \in \mathcal{R}_1(v)) \end{aligned}\right\} \qquad (39)$$

$$\left.\begin{aligned} &\partial\Lambda/\partial \phi_r = \chi_v^2 - \textstyle\sum_{l:l\in r} \lambda_l \leq 0 \\ &\phi_r(\partial\Lambda/\partial\phi_r) = 0,\ \phi_r \geq 0,\ \lambda_l \geq 0\ (r \in \mathcal{R}_2(v)) \end{aligned}\right\} \qquad (40)$$

As in [19], λ_l is interpreted as the link shadow cost, which reflects the marginal value of capacity on link l. It is a critical quantity that unifies routing and bandwidth provisioning decisions. Specifically,

1. By (39), for any route $r_0 \in \mathcal{R}_1(v)$, $\xi_{r_0} > 0$ only when

$$\sum_{l\in r_0} \lambda_l = \min_{r\in\mathcal{R}_1(v)} \sum_{l\in r} \lambda_l,$$

suggesting that traffic for retail demand is carried solely on the *minimum cost* path(s) of all the admissible routes, where path costs are obtained by summing link costs, and link costs are given by λ_l. This is also true for routing of wholesale traffic, as implied by (40).

2. We define $\chi_v^2 = \min_{r\in\mathcal{R}_2(v)} \sum_{l\in r} \lambda_l$ as the opportunity cost of carrying wholesale traffic between $v \in \mathcal{V}_2$. By (38), e_v is the lower bound of χ_v^2 and $y_v > 0$ only when $e_v = \chi_v^2$. This means in order to provide bandwidth for wholesale between v, the opportunity cost has to be at its lower bound, which requires low shadow costs of links on v's admissible routes.

3. Similarly, we define $\chi_v^1 = \min_{r\in\mathcal{R}_1(v)} \sum_{l\in r} \lambda_l$ as the opportunity cost of carrying retail demand between $v \in \mathcal{V}_1$. By (37), the optimal quantity to be provisioned is determined at the point where the marginal increase of mean revenue, $\pi_v \bar{F}_v$, compensated by the marginal change of risk, $\delta\pi_v(d_v - m_v)/\sqrt{\sum_{v\in\mathcal{V}_1} \pi_v^2 s_v^2}$, equals the opportunity cost, χ_v^1.

B. Truncated Gaussian Distribution

When demand between a node pair comes from many independent individual sources, the total demand can be approximated by the Gaussian distribution. We will make this assumption in what follows. Of course, the distribution needs to be restricted to nonnegative values, and the PDF should also be normalized properly so that the total probability over the restricted sample space is unity. As a result, we will consider the *Truncated Gaussian Distribution* characterized by the following PDF function:

$$f_v(x) = \frac{1}{\sqrt{2\pi}\sigma_v G_v} e^{-(x-\mu_v)^2/2\sigma_v^2} \quad x \geq 0 \qquad (41)$$

where the normalizing parameter is:

$$G_v = \frac{Erfc(-\tau_v)}{2} \quad \text{and} \quad \tau_v = \frac{\mu_v}{\sqrt{2}\sigma_v}. \qquad (42)$$

As before, d_v is the amount of bandwidth provisioned to carry retail demand between v. Then the mean and standard deviation of carried demand can be derived as

$$\begin{aligned} m_v(d_v) &= \frac{1}{\sqrt{2\pi}\sigma_v G_v} \int_0^\infty \min(x, d_v) e^{-(x-\mu_v)^2/2\sigma_v^2} dx \\ &= \mu_v + \gamma(\tilde{d}_v), \qquad (43)\end{aligned}$$

$$s_v^2(d_v) = \frac{\int_0^\infty \min(x^2, d_v^2) e^{-(x-\mu_v)^2/2\sigma_v^2} dx}{\sqrt{2\pi}\sigma_v G_v} - m_v^2$$
$$= \sigma_v^2[1 - \frac{Erfc(\tilde{d}_v)}{2G_v} - \gamma^2(\tilde{d}_v) + \sqrt{2}\tilde{d}_v\gamma(\tilde{d}_v)$$
$$- \frac{\tilde{d}_v + \tilde{\mu}_v}{\sqrt{\pi}G_v} e^{-\tilde{\mu}_v^2}], \quad (44)$$

where $\tilde{d}_v = (d_v - \mu_v)/\sqrt{2}\sigma_v$, $\tilde{\mu}_v = \mu_v/\sqrt{2}\sigma_v$, and $\gamma(\tilde{d}_v) = [e^{-\tilde{\mu}_v^2} - e^{-\tilde{d}_v^2} + \sqrt{\pi}\tilde{d}_v Erfc(\tilde{d}_v)]/\sqrt{2\pi}G_v$.

The following inequalities provide important insights of the optimal solution.

Theorem 3 Let $F_v(x)$ be the CDF of the distribution specified by (41), Then

$$a) \quad \frac{\partial(m_v/d_v)}{\partial d_v} \leq 0 \quad (45)$$

$$b) \quad \frac{\partial m_v}{\partial \sigma_v} \leq 0 \text{ if } d_v \leq 2\mu_v \quad (46)$$

$$c) \quad \frac{\partial \bar{F}_v}{\partial \sigma_v} \geq 0 \text{ if } d_v \geq (1 + \frac{e^{-\mu_v^2/2\sigma_v^2}}{G_v})\mu_v \quad (47)$$

Proof: a) Since $\bar{F}_v(x)$ decreases in x, from (3),

$$\frac{\partial(m_v/d_v)}{\partial d_v} = \frac{d_v\bar{F}_v - \int_0^{d_v} \bar{F}_v(x)dx}{d_v^2} \leq 0$$

b) From (3) and (41)

$$m_v = \frac{1}{\sqrt{2\pi}G_v}[\int_{-\mu_v/\sigma_v}^{(d_v-\mu_v)/\sigma_v} (\sigma_v t + \mu_v)e^{-t^2/2}dt$$
$$+ d_v \int_{(d_v-\mu_v)/\sigma_v}^{+\infty} e^{-t^2/2}dt] \quad (48)$$

$$\frac{\partial m_v}{\partial \sigma_v} = \frac{e^{-\mu_v^2/2\sigma_v^2}}{\sqrt{2\pi}G_v}[1 - e^{\frac{d_v(2\mu_v - d_v)}{2\sigma_v^2}}] - \frac{m_v}{G_v}\frac{\partial G_v}{\partial \sigma_v} \quad (49)$$

$$\frac{\partial G_v}{\partial \sigma_v} = \frac{\mu_v}{\sqrt{2\pi}\sigma_v^2} e^{-\mu_v^2/2\sigma_v^2} \geq 0, \text{ so} \quad (50)$$

$$\frac{\partial m_v}{\partial \sigma_v} \leq 0 \text{ if } d_v \leq 2\mu_v.$$

c) From (41)

$$\bar{F}_v = \frac{1}{\sqrt{2\pi}G_v} \int_{(d_v-\mu_v)/\sigma_v}^{\infty} e^{-t^2/2}dt \quad (51)$$

$$\frac{\partial \bar{F}_v}{\partial \sigma_v} = \frac{(d_v - \mu_v)}{\sqrt{2\pi}G_v\sigma_v^2} e^{-(d_v-\mu_v)^2/2\sigma_v^2}$$
$$- \frac{1}{\sqrt{2\pi}G_v^2}\frac{\partial G_v}{\partial \sigma_v} \int_{(d_v-\mu_v)/\sigma_v}^{+\infty} e^{-t^2/2}dt \quad (52)$$

Because $e^{-x^2/2} > \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}dt$ for all $x \geq 0$,

$$\frac{\partial \bar{F}_v}{\partial \sigma_v} \geq \frac{e^{-(d_v-\mu_v)^2/2\sigma^2}}{G_v}[\frac{(d_v - \mu_v)}{\sqrt{2\pi}\sigma_v^2} - \frac{\partial G_v}{\partial \sigma_v}/G_v], \quad (53)$$

By (50), the right hand side equals

$$\frac{e^{-(d_v-\mu_v)^2/2\sigma^2}}{\sqrt{2\pi}\sigma_v^2 G_v}[d_v - (1 + \frac{e^{-\mu_v^2/2\sigma_v^2}}{G_v})\mu_v], \quad (54)$$

which leads to the conclusion.□

Theorem 3 has the following implications:

Equation (45) shows that the ratio of the mean carried traffic to the provisioned capacity decreases when more capacity is provisioned to carry the demand. Notice that this trend of declining return from bandwidth provisioning is not specific to the truncated Gaussian distribution assumption, as can be seen from the proof.

The mean carried traffic also decreases when the standard deviation of the demand distribution increases, provided that the provisioned capacity does not exceed the critical value in (46). Higher variability will reduce the mean carried traffic, because the risk that the demand falls below the mean value becomes dominant. When the demand distribution is close to the Gaussian distribution, then the threshold is unlikely to be violated unless the network is extremely under-loaded. For example, in the numerical study of the next section, we let $\mu_v \geq 2.75\sigma_v$ to minimize the impact of truncation. In this case, exceeding the threshold value requires provisioning bandwidth up to the 99.5% quantile of the demand distribution, which is not likely in data networks.

The last inequality (47) shows the influence of demand variability on bandwidth provisioning. To understand its implication, we refer to the necessary condition in (37) and let $\delta = 0$, so

$$\pi_v\bar{F}_v = \chi_1^v.$$

To maintain the equality when $\partial\bar{F}_v/\partial\sigma_v \geq 0$ and σ_v increases, one has to increase d_v (notice that $\partial\bar{F}_v/\partial d_v \leq 0$) to neutralize the impact on $\pi_v\bar{F}_v$ by σ_v increase, or raise χ_1^v, which makes the bandwidth between v more expensive. Of course, the result is premised on the condition in (47), which is usually satisfied in networks that are not over-loaded. For example, as we set $\mu_v \geq 2.75\sigma_v$, the impact of $e^{-\mu_v^2/2\sigma_v^2}/G_v$ can be ignored. So the condition is met when the amount of bandwidth provisioned to serve a demand is more than the mean of its distribution in the optimal solution.

V. Numerical Studies

In this section, we study the effects of uncertainty in demand on traffic engineering through numerical examples. We first describe the network topology and base case scenario in section V-A. In V-B, we discuss the impact of demand variability on stochastic traffic engineering design.

A. Framework and Base Case

We consider a sample network which has 12 nodes and 14 bidirectional links. The network topology, as well as indices of nodes and links are shown in Figure 1. We assume that retail demands are symmetric in both directions for each node pair, and subject to the Truncated Gaussian distribution which has been discussed in the previous section. We keep $\mu_v \geq 2.75\sigma_v$ for all $v \in \mathcal{V}_1$. In this case, μ_v and σ_v approximately equal the mean (m_v) and standard deviation (s_v) of the demand distribution, and

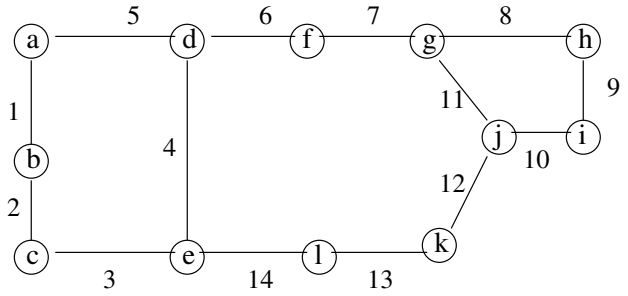

Fig. 1. Network Topology

are used to approximate these parameters in the following analysis.

We assume demand distributions for all origin-destination pairs have the same mean, i.e., $\mu_v = \bar{\mu}$ for all $v \in \mathcal{V}_1$. Let h_v be the minimum number of hops between node pair $v \in \mathcal{V}_1$. Then $\sum_{v\in\mathcal{V}_1} \mu_v h_v = \bar{\mu} \sum_{v\in\mathcal{V}_1} h_v$ is an estimate of average bandwidth demand for the retail service. We define the ratio of this quantity to the total installed network capacity to be the network *load factor*, denoted by ρ, i.e.,

$$\rho = \frac{\bar{\mu} \sum_{v\in\mathcal{V}_1} h_v}{\sum_{l\in\mathcal{L}} C_l}. \tag{55}$$

Note that for a given network topology and link capacities, $\bar{\mu}$ and ρ are uniquely determined by each other. In the following discussions, we will vary ρ.

The variability of retail demand is characterized by the coefficient of variation, defined by

$$CV_v = \frac{\sigma_v}{\mu_v}. \tag{56}$$

Below, for a given μ_v, we vary CV_v, and thus the standard deviation σ_v.

We make the following assumptions to create a base case scenario.

1. All links have 150 units of installed capacity, except links $6, 7, 12, 13$ and 14, which have 200 units. The latter links are given higher installed capacities since they are likely to carry more traffic on account of their central locations.
2. Retail demand and opportunities of bandwidth wholesale are ubiquitous in the network, i.e., $\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}$. The load factor of the network is $\rho = 0.65$. All retail demands have the same coefficient of variation, which is set at $CV_v = 0.1$. Consequently, $\mu_v = 8.7$, and $\sigma_v = 0.87$ $(v \in \mathcal{V}_1)$.
3. The unit price for carrying retail demand (π_v) is proportional to the distance between the originating and terminating nodes, where the distance for node pair v is measured by h_v, i.e., $\pi_v = \kappa h_v$ with $\kappa = 50$. The unit wholesale price (e_v) is 10% of the unit retail price for the same node pair, i.e., $e_v = 0.1\pi_v$
4. The risk parameter is set at $\delta = 0.5$.
5. A path between node pair v is an admissible route for both retail and wholesale traffic if and only if the number of links on this path does not exceed $h_v + 2$, i.e., the minimum number of hops plus 2.

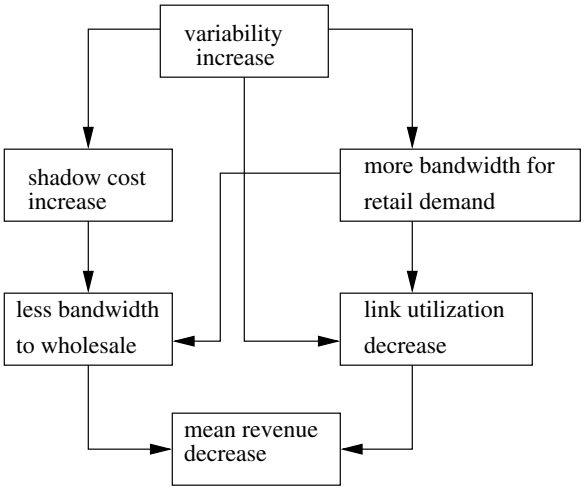

Fig. 2. Result Summary

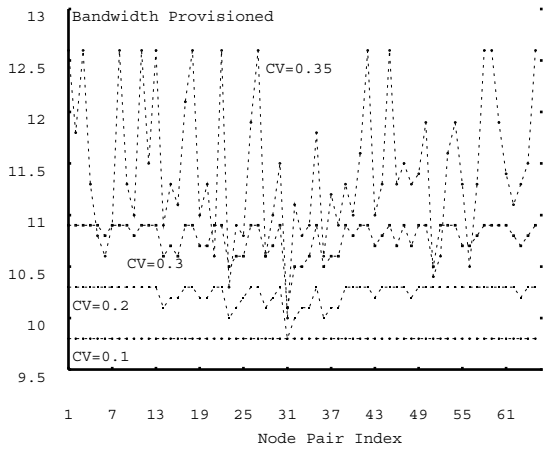

Fig. 3. Bandwidth Provisioned for Retail Increases with Variability

B. Implications of Demand Variability

We start with the base case, and increase demand variability by increasing the coefficient of variation from 0.1 to $0.2, 0.3$, and 0.35. All other parameters are kept unchanged from the base case. Figure 2 summarizes various important implications of increasing demand variability on the optimal design. Detailed explanations follow.

As implied in (47) of Theorem 3 in IV-B, when the variability for retail demand between $v \in \mathcal{V}_1$ increases, the optimal solution provisions more bandwidth, i.e., d_v is larger, which may or may not be accompanied by an increase of the opportunity cost of the minimum cost route(s). Since the opportunity cost of a route is the sum of shadow costs of links on that route, higher route costs imply that the shadow costs of some links are also higher. In our example, we observe that in general, both provisioned bandwidth for retail demand and link shadow costs increase with demand variability, as shown in Figures 3 and 4. Notice that both increases are notably non-uniform across links and node pairs, and depend on the locations of links and nodes. Shadow costs increase faster on "busy" links, i.e., links that are

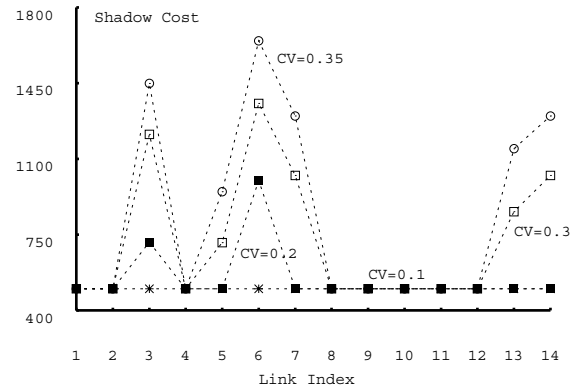

Fig. 4. Shadow Costs Increase with Variability

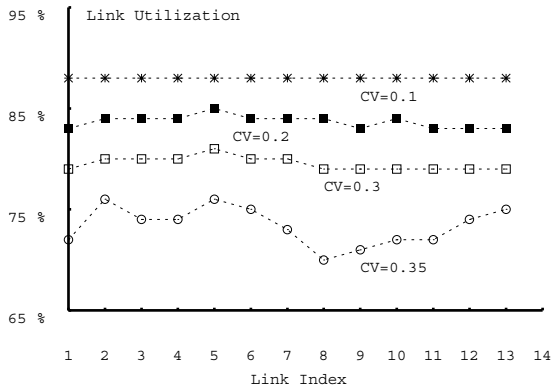

Fig. 6. Link Utilization Decreases with Variability

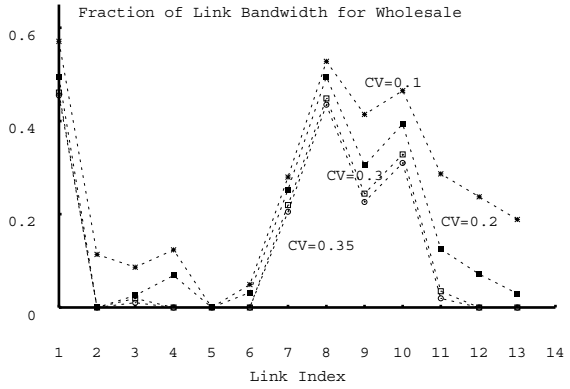

Fig. 5. Bandwidth for Wholesale Decreases with Variability

on admissible routes of a larger number of node pairs. For node pairs that are connected by "busy" links, the increase in provisioned bandwidth is relatively small, and in some cases, even negative. The situation is reversed for links that carry traffic for a smaller number of demands. Shadow costs stay the same on these links, and we observe that there are large increases in the amounts of provisioned bandwidth for retail demands that only pass through these links.

For networks with fixed capacity, increasing bandwidth provisioned to carry retail demand reduces the amount of bandwidth for wholesale. This can seen from Figure 5, which shows the percentage of bandwidth set aside for wholesale on each link. The figure can also be explained by using the shadow cost argument. As indicated by the necessary condition (38), it is optimal to wholesale bandwidth between node pair $v \in \mathcal{V}_2$ only when the minimum route cost χ_v^2 stays at its lower bound. As demand variability increases, link shadow costs increase, which raises the minimum route costs above their lower bounds for certain node pairs. As a result, the amount of bandwidth set aside for wholesale between these pairs decreases.

Another implication of increasing bandwidth variability is that it reduces the *carried load*. The carried load is given by the ratio of the expected amount of carried demand to the amount of provisioned bandwidth. For the fixed amount of provisioned bandwidth for a given node pair, this quantity decreases as the variability increases, as indicated by (46) of Theorem 3. Moreover, since the optimal amount of bandwidth provisioned to carry retail demand generally increases with variability, the carried load decreases even more according to (45) of Theorem 3. For any link, define the link utilization rate as the ratio of the mean total retail demand carried on the link to the amount of bandwidth provisioned for retail service. The aggregated effect of decreasing normalized carried load of node pairs is reflected by the decline of link utilization, as shown in Figure 6. Notice that utilization of all links decreases as demand variability increases and the change is also not uniform.

The above shows that when demand variability increases, bandwidth provisioned to carry retail demands is used less efficiently, and bandwidth provisioned for wholesale is reduced. Therefore, it should be no surprise that the expected revenue decreases with demand variability. In Figure 7, we vary network load from 0.30 to 0.80 in 0.15 increments, and plot the expected revenues as functions of the coefficient of variation. At all load levels, the expected revenue decreases monotonically as demand variability increases. Furthermore, we also plot the change of standard deviation of total revenue in Figure 8. Clearly, when coefficient of variation increases, the standard deviation of total revenue increases. Both figures show that demand variation is detrimental to revenue, reducing the mean revenue and increasing the risk of revenue shortfall.

We end this section by noting that many of above results are implications of Theorem 3, which is specialized to the case of normal loading of the network. Different results may emerge for extreme values of the load factor. For instance, if the network is extremely heavily loaded, then increasing variability may reduce the optimal amount of bandwidth provisioned to retail demand. Also for instance, if the network is extremely lightly loaded, then bandwidth utilization can increase with demand variability. Nevertheless, we consider these scenarios are unlikely to happen in a properly dimensioned network, and exclude them from our discussions.

VI. CONCLUSION

We have presented and analyzed a stochastic traffic engineering framework for off-line planning of bandwidth provisioning

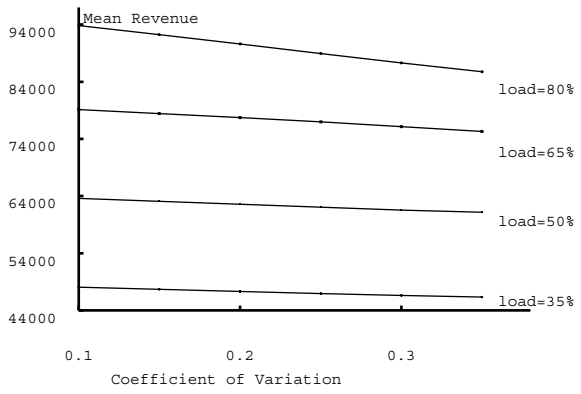

Fig. 7. Mean Revenue Decreases with Variability

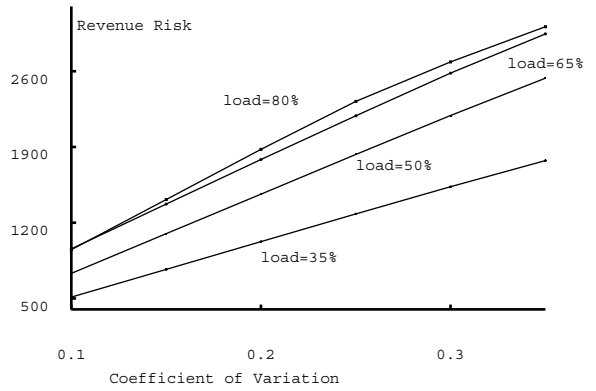

Fig. 8. Revenue Risk Increases with Variability

and routing. The framework is based on an optimization model that uses probability distributions of demands as inputs and maximizes the weighted combination of the mean revenue and the risk of revenue shortfall. We discuss properties of the objective function, and strategies of solving the model as a concave maximization problem. We consider a two-tier market structure for demand and revenue. Link shadow costs, which are outputs of our model, serve as the basis for the optimal bandwidth provisioning and routing in both markets. In our numerical studies, we analyze the impacts of demand variability on various aspects of traffic engineering design. We observe significant changes in shadow costs, link utilization, bandwidth provisioning and routing with demand variability, and explain their causes and implications.

Our analysis can be extended in several directions. Different carriers, or the same carrier in different situations, may have different tolerance to the risk of revenue shortall. Therefore, it is useful to understand how the risk parameter in our model, δ, affects the outcome of traffic engineering design. It is also important to study the influence of the grade of service parameter $(\bar{d}_v, v \in \mathcal{V}_1)$, on the optimal solution. Another interesting topic is to compare the difference between the stochastic traffic engineering model with a deterministic approach with additional compensation mechanisms [16], and examine the effectiveness of latter under different circumstances. These issues are currently being investigated, and will be presented in future publications.

REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows.* Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.

[2] D. Applegate and M. Thorup, Load Optimal MPLS Routing with N+M Labels, *preprint*

[3] P. Aukia *et al*., RATES: A Server for MPLS Traffic Engineering, *IEEE Network*, **14**, pp. 34-41, March/April 2000.

[4] D. O. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, Overview and Principles of Internet Traffic Engineering *RFC 3272, IETF*, May 2002.

[5] J. Cao, D. Davis, S. V. Wiel, and B. Yu, Time-Varying Network Tomography, *Journal of the American Statistical Association*, **95**, pp. 1063-1075.

[6] J. Cao, S. V. Wiel, B. Yu, and Z. Zhu, A Scalable Method for Estimating Network Traffic Matrices, *Bell Labs Technical Memorandum*, 2000.

[7] A. Elwalid, C. Jin, S. Low, and I. Widjaja, Mate: MPLS Traffic Engineering, *Proceedings of IEEE INFOCOM2001*, pp. 1300-1309, 2001.

[8] A. Feldmann *et al*., NetScope: Traffic Engineering for IP Networks, *IEEE Network*, **14**, pp. 11-19, March/April 2001.

[9] Fortz, B. and M. Thorup, Internet Traffic Engineering by Optimizing OSPF Weights, *Proceedings of IEEE INFOCOM2000*, pp. 519-528, 2000.

[10] M. Grossglauser and J. Rexford, Passive Traffic Measurement for IP Operations *INFORMS Telecom Meeting 2002*, Ft. Lauderdale, FL, March, 2002.

[11] J. Hirshleifer and J. G. Riley, *The Analytics of Uncertainty and Information*, Cambridge University Press, Cambridge, England, 1992.

[12] Lagoa, C. and H. Che, Decentralized Optimal Traffic Engineering in the Internet, *Computer Communications Review*, vol. 30, No. 5, pp. 39-47, October 2000.

[13] H. Levy, *Stochastic Dominance: Investment Decision Making Under Uncertainty*, Kluwer Academic Publishers, MA, 1998.

[14] D. Luenberger, *Introuduction to Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, Reading, MA, 1973.

[15] H. Markowitz, *Mean Variance Analysis in Portfolio Choice and Capital Markets*, Basil Blackwell, New York, 1987.

[16] D. Mitra and K. G. Ramakrishnan, A Case Study of Multiservice Multipriority Traffic Engineering Design for Data Networks, *Proceedings of IEEE GLOBECOM 99*, pp. 1077-1083, December 1999.

[17] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, Robust Optimization of Large-Scale Systems, *Operations Research*, vol. 43, No. 2, pp. 264-281, March-April 1995.

[18] D. Mitra and Q. Wang, Network Revenue Management in the Presence of Demand Uncertainty, *submitted for publication*, December 2002.

[19] Mitra, D., Ramakrishnan, K. G., and Q. Wang, Combined Economic Modeling and Traffic Engineering: Joint Optimization of Pricing and Routing in Multi-Service Networks, *Traffic Engineering in the Internet Era: Proceedings of 17th International Traffic Congress-ITC17*, eds. J.M. de Souza, N.L.S. da Fonseca and E.A. de Souza e Silva, Elsevier, pp. 73-85, 2001.

[20] W. Ogryczak and A. Ruszczynski, Dual stochastic dominance and related mean–risk models, *SIAM Journal on Optimization*, vol. 13 pp. 60–78, 2002.

[21] N. Semret, R. R.-F. Liao, A. T. Campell, and A. A. Lazar, Peering and Provisioning of Differentiated Internet Services, *Proceedings of IEEE INFOCOM 2000*, pp. 100-107, 2000.

[22] S. Suri, M. Waldvogel, and P. R. Warkhede, Profile-Based Routing: A New Framework for MPLS Traffic Engineering, *QofIS'01*, Coimbra, Portugal, 2001.

[23] P. Trimintzios *et al*. A Management and Control Architecture for Providing IP Differentiated Services in MPLS-Based Networks, *IEEE Communications*, vol. 39, no. 5, pp. 80-87, May 2001.

[24] Y. Vardi, Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data, *Journal of the American Statistical Association, Theory and Method*, **91**, pp. 365-377.

[25] J. von Neuman and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, New Jersey, 1953.

[26] I. Widjaja, I. Saniee, A. Elwalid, and D. Mitra, Online Traffic Engineering with Design-based Routing, *ITC Specialist Workshop*, Wurzburg, July 2002.