Robust Network Design for Stochastic Traffic Demands

Matthew Johnston, Hyang-Won Lee and Eytan Modiano
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: {mrj, hwlee, modiano}@mit.edu

Abstract—This paper addresses the problem of logical topology design for optical backbone networks subject to stochastic traffic demands. The network design problem is broken into three tasks: traffic routing, capacity allocation, and link placement. While the routing and capacity allocation subproblem can be formulated using convex optimization, the link placement component is prohibitive due to its nonlinearity. To address this issue, we develop a linear formulation for the routing and capacity allocation subproblem by extending tools from robust optimization to Gaussian random variables. We show that this linear formulation performs comparably to the optimal nonlinear formulation. Our formulation can then be used to solve the link-placement subproblem for stochastic traffic.

I. INTRODUCTION

This paper studies the problem of designing a robust network that can support stochastic traffic demands. The concept of an overlay network such as network virtualization and Virtual Private Networks (VPNs) has emerged as a promising solution for providing flexible Internet services. Such a network is constructed by purchasing a certain amount of bandwidth from the underlying infrastructure provider. Traffic engineering is used to route a set of demands from source to destination nodes to improve network performance [1]. Sufficient bandwidth must be purchased so that we can place logical links on the physical topology, assign capacity to these links, and route the traffic over the logical topology. It is especially important to provision sufficient but not excessive link capacities to support the demands. This problem has been extensively studied in the literature [2]–[6], with various design criteria such as minimizing the maximum congestion, total capacity, or the number of hops to route the demands. All these works assume that the traffic demand between every node pair is known a priori.

However, the design problem is typically solved during the network configuration stage, where the exact traffic information is unavailable. For this reason, traffic demands must be modeled as random quantities. Unlike the deterministic case, network links must be provisioned with sufficient capacity to support possible demand fluctuations. For models with bounded traffic variability, the most common approach is to allocate sufficient link capacity to support the worst-case set of demands, so that any demand fluctuations can be handled by the network [7], [8].

On the other hand, a demand is typically an aggregation of many independent traffic sources, and by the Central Limit Theorem, it can be approximated using a Gaussian random variable. In this case, allocating capacity to support the worst case set of demands yields an unacceptable over-provisioning of capacity, since Gaussian demands have unbounded support. Therefore, a different capacity provisioning approach is required for unbounded demands. The work of [9] addresses this issue by formulating an optimization problem that incorporates capacity underutilization into the objective as a penalty, to detract from over-provisioning.

In this paper, we take a different approach by explicitly taking into account the probability of traffic exceeding the provisioned capacity on each link. A similar approach was also considered in [10], where the overflow probability is explicitly derived for Gaussian distributed traffic and used in the optimization problem for routing and capacity allocation. While this formulation leads to a convex optimization problem that is efficiently solvable, its nonlinearity prevents extensions to discrete optimization problems such as link placement. For this reason [11] formulates a Linear Program (LP) by changing probabilistic constraints to linear deterministic ones. This approach suffers from an over-provisioning of capacity, since it inherently ignores the effect of statistical multiplexing of many demands. There are other approaches dealing with traffic variability [12], [13]; however, they ignore the distribution of the random demands, which also results in over-provisioning of capacity.

In this paper, we formulate and analyze the problem of topology design under Gaussian-distributed demands. We develop a linear formulation for routing traffic and allocating capacity to support random demands by extending the robust optimization techniques of [14] to Gaussian random variables. Our formulation exploits the effect of traffic aggregation to reduce the amount of allocated link capacity required, while at the same time yielding the computational advantages of a linear approach.

The rest of the paper is organized as follows. In Section II, we introduce the network model and motivate the development of a new approach to solving the routing and capacity alloc-
In Section III, we introduce robust optimization as a technique to develop a new formulation for the network design problem. In Section IV, we discuss how our formulation can be used to solve the link placement subproblem in parallel to the capacity allocation and routing problems, and we conclude in Section V.

II. NETWORK MODEL AND MOTIVATION

Our goal is to support a set of demands \( \mathcal{D} \), where for \((s, d) \in \mathcal{D} \), \( \lambda_{sd} \) is the traffic demand from node \( s \) to node \( d \). The demand \( \lambda_{sd} \) represents the long-term average traffic flow between two nodes. As mentioned above, average traffic demands between node-pairs are often described using a Gaussian distribution [15]. We assume that all traffic demands are independent, and each \( \lambda_{sd} \) follows a Gaussian distribution with mean \( \mu_{sd} \) and variance \( \sigma_{sd}^2 \).

In our formulation, we allow a traffic demand to be split among various paths in the network. Let \( 0 \leq a_{ij} \leq 1 \) represent the fraction of the traffic flow from \( s \) to \( d \) traversing link \((i, j)\). Let \( C_{ij} \) represent the capacity allocated to link \((i, j)\). In this paper, we study the problem of finding a routing (determined by \( \{a_{ij}\} \)) and capacity allocation (determined by \( \{C_{ij}\} \)) that can support the demands in \( \mathcal{D} \) while minimizing the maximum link capacity. The chosen objective function has the effect of splitting the traffic over multiple paths, and thus balancing the load over the network.

Note that the routing and capacity allocation problem can be extended to include the link placement subproblem, by introducing a binary variable \( b_{ij} \), which takes on the value 1 if and only if a directed link is placed from node \( i \) to node \( j \). While this extension captures the network design problem, we will initially focus on the routing and capacity allocation problems. The network design problem will be discussed in Section IV. For now, we assume the network is described by a graph \( G(V, E) \) with a set of edges \( E = \{(i, j) | b_{ij} = 1\} \). Define the overflow probability of link \((i, j)\) to be the probability that the traffic on link \((i, j)\) exceeds its capacity, i.e., \( \Pr(\sum_{s,d} a_{ij}^{sd} \lambda_{sd} > C_{ij}) \). The routing and capacity allocation problem can be formulated as follows:

\[
\begin{align*}
\min_{C, a \geq 0} \quad & \max_{(i,j) \in E} C_{ij} \\
\text{subject to:} \quad & \Pr\left( \sum_{s,d} a_{ij}^{sd} \lambda_{sd} > C_{ij} \right) \leq \epsilon, \forall (i,j) \in E \quad (1) \\
& FC(s, d, a^{sd}), \forall (s, d) \in \mathcal{D}
\end{align*}
\]

The objective is to minimize the maximum link capacity. We note that the objective function can be linearized by replacing it with a new variable \( C_{max} \) and introducing new constraints \( C_{max} \geq C_{ij}, \forall (i,j) \), but for brevity, we will use the nonlinear form throughout the paper. The function \( FC(s, d, a^{sd}) \) represents a flow conservation constraint from \( s \) to \( d \) given as

\[
\sum_{j:(i,j) \in E} a_{ij}^{sd} - \sum_{j:(j,i) \in E} a_{ji}^{sd} = \begin{cases} 
1, & \text{if } s = i \\
-1, & \text{if } d = i, \forall i \in V \\
0, & \text{o.w.}
\end{cases} \quad (2)
\]

This constraint describes the routing of each demand.

The constraint in (1) requires sufficient capacity to be allocated to ensure the overflow probability at each link is no greater than a design parameter \( \epsilon \). Since the sum of Gaussian random variables is also Gaussian, this constraint can be rewritten as a convex, deterministic constraint as follows:

\[
C_{ij} \geq \sum_{sd} a_{ij}^{sd} \mu_{sd} + \Phi^{-1}(1 - \epsilon) \sqrt{\sum_{sd} (a_{ij}^{sd} \sigma_{sd})^2}, \quad (3)
\]

where \( \Phi(\cdot) \) is the CDF of a Gaussian random variable. While this optimization problem is solvable, the nonlinearity of (3) prevents the addition of integer constraints, as required by the network design problem, since the resulting formulation will be a nonlinear integer program and effective solution methodologies are yet unknown for general, non-linear, integer programs. Additionally, solving this convex optimization problem may be impractical for large networks.

To alleviate this problem, the constraint in (3) can be replaced by the following linear constraint:

\[
C_{ij} \geq \sum_{sd} a_{ij}^{sd} (\mu_{sd} + \Phi^{-1}(1 - \epsilon) \sigma_{sd}). \quad (4)
\]

This constraint, which allocates dedicated capacity for each flow sufficient to satisfy (1) without statistical multiplexing gains, is wasteful in terms of the maximum capacity on any link in the network. To understand the difference between constraints (3) and (4), we study the following example. Consider the network shown in Figure 1. Traffic demands exist from \( s_1 \) to \( d_1 \) and from \( s_2 \) to \( d_2 \), and are independently and identically distributed according to \( N(\mu, \sigma^2) \). Due to symmetry, both demands have the same fraction of traffic (\( \alpha \)) routed on the shared link. Clearly, the optimal solution allocates equal capacity on all links. Using the linear capacity constraint in (4), we can analytically show that the optimal routing is \( \alpha = \frac{1}{3} \). This routing is optimal for any deterministic traffic. Alternatively, for the non-linear constraint in (3) it can be shown that the optimal routing will send more traffic through the middle link (i.e., larger value of alpha), to take advantage of statistical multiplexing gain from combining multiple traffic flows, so that the resulting variance is smaller than the sum of individual variances. Consequently, the required capacity is \( O(\sqrt{n}) \) less than that with the linear constraint, which simply adds up indi-
individual variances, resulting in an over-provisioning of capacity. In the next section, we use robust optimization to develop a linear formulation exploiting the statistical multiplexing gain, so that it can be applied to network design problems without significant capacity over-provisioning.

III. ROBUST OPTIMIZATION APPROACH

Robust optimization is a technique for solving optimization problems with uncertain parameters. In [14], the authors introduce a novel robust optimization approach based on the idea that it is unlikely for all of the random parameters to simultaneously take values far from their means. Therefore, resources can be allocated to protect against scenarios where only a fraction of the parameters exceed their mean, and the probability of constraint violation can still be kept small. This technique, however, applies only to bounded uncertainty, and thus cannot be directly applied to our problem. In this section, we extend the robust optimization approach of [14] to Gaussian random variables by considering a truncated Gaussian random variable with a large truncation point.

A. Robust Formulation for Truncated Gaussian Demands

Initially, we assume that demands are drawn from a truncated Gaussian distribution, i.e., the traffic $\lambda_B^{sd}$ between demand pair $(s, d)$ satisfies $\lambda_B^{sd} \sim \mathcal{N}(z| z \leq \mu_{sd} + k\sigma_{sd}; \mu_{sd}, \sigma_{sd}^2)$ for some constant $k$, and has a PDF

$$f_{\lambda_B^{sd}}(z| \mu_{sd}, \sigma_{sd}^2) = \frac{1}{\sigma_{sd} \Phi(k)} \phi\left( \frac{z - \mu_{sd}}\sigma_{sd} \right), \quad z < \mu_{sd} + k\sigma_{sd}$$

(6)

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and CDF of a standard normal random variable respectively. We now apply the approach of [14] to our problem.

For each link $(i, j)$, a parameter $\Gamma_{ij}$ is introduced taking a value between 0 and $n_{ij}$, where $n_{ij}$ is the number of demands traversing link $(i, j)$. Capacity is assigned to protect against any scenario where $\Gamma_{ij}$ of the demands exceed their means. The other demands are assumed not to exceed their mean, and are therefore allocated a capacity equal to their mean. Using this idea, we formulate a new capacity constraint to replace those in (3) and (4). Let $S_d \subseteq D$ be a subset of demands of size $|S_d|$, and let $t_{ij} \in D \setminus S_d$ be another demand. On each link $(i, j)$, sufficient capacity is allocated to support the mean $a_{ij}^{sd} \mu_{sd}$ for each demand, as well as the $|S_d|$ largest values of $a_{ij}^{sd} \kappa_{sd}$ to guarantee support for the $|\Gamma_{ij}|$ demands exceeding their mean. In the case where $\Gamma_{ij}$ is not an integer, the fraction $(\Gamma_{ij} - |\Gamma_{ij}|)$ of $a_{ij}^{sd} \kappa_{sd}$ is also supported. The constraint is presented below.

$$C_{ij} \geq \sum_{sd \in D} \mu_{sd} a_{ij}^{sd} + \sum_{s, \{i, j\} \subseteq D, S_d \subseteq D, |S_d| = |\Gamma_{ij}|, t_{ij} \in D \setminus S_d} \max_{S_d} \left\{ \sum_{(s, d) \in S_d} k\sigma_{sd} a_{ij}^{sd} + (\Gamma_{ij} - |\Gamma_{ij}|) k\sigma_{sd} a_{ij}^{sd} \right\}.$$  

(7)

The first term in the right hand side is the mean of the traversing demands, and the second term is the sum of $\Gamma_{ij}$ largest possible fluctuations from the mean. This formulation exploits the statistical multiplexing effect discussed above by assuming that the demands are unlikely to simultaneously exceed all of the provisioned link capacity. Note that the solution is feasible for any realization of random demands such that no more than $\Gamma_{ij}$ demands exceed their mean. Hence, the robustness of the formulation is determined by the value of $\Gamma_{ij}$. For example, when $\Gamma_{ij} = n_{ij}$, the solution yields the most conservative capacity allocation, and all the demands are supported with probability 1. In contrast, for small $\Gamma_{ij}$, a small amount of extra capacity is allocated at the expense of higher overflow probability.

The link overflow probability can be bounded as a function of $\Gamma_{ij}$, and the value of $\Gamma_{ij}$ that satisfies the probabilistic constraint in (1) can be characterized. [14] provides an upper bound on that probability for general symmetric, bounded random variables, and we modify this bound for truncated Gaussians to relate $\epsilon$ to $\Gamma_{ij}$ for each link. This bound is shown in the following theorem.

**Theorem 1:** Let $\lambda_B^{sd}$ be the traffic from source $s$ to destination $d$. Further, let $\lambda_B^{sd}$ be a continuous random variable with density $\mathcal{N}(\lambda_B^{sd}| \lambda_B^{sd} \leq \mu_{sd} + k\sigma_{sd}; \mu_{sd}, \sigma_{sd}^2)$. Let $0 \leq a_{ij}^{sd} \leq 1$ and $C_{ij} \geq 0$ satisfy (7). Let $n_{ij}$ be the number of demands routed over link $(i, j)$. Then, the probability that link $(i, j)$ overflows its allocated capacity is bounded by

$$\Pr \left[ \sum_{sd} \lambda_B^{sd} a_{ij}^{sd} > C_{ij} \right] \leq \exp \left( -\frac{\Gamma_{ij}^2 k^2}{2n_{ij}} \right).$$

(8)

The proof is omitted for brevity, but involves an application of the Chernoff Bound. Given a desired probability $\epsilon$ of overflow on link $(i, j)$, the value of the parameter $\Gamma_{ij}$ required on that link can be computed by

$$\Gamma_{ij} = \sqrt{-2n_{ij} \log \epsilon k-1}.$$  

(9)

Hence, with the value of $\Gamma_{ij}$ in (9), the capacity constraint (7) satisfies the overflow probability requirement in (1). Although the formulation obtained by replacing (1) with (7) is nonlinear, it can be reformulated as an equivalent LP by following the procedure used in [14] (see [16] for details). Note that the constraint (7) is valid only for truncated Gaussian demands, and in Section III-B, we extend this approach to general Gaussian demands.

Note also that capacity is assigned to a link as a linear function of $\Gamma_{ij}$, and grows with $\sqrt{n_{ij}}$. This is similar to the optimal convex capacity constraint in (3). Therefore, our formulation results in routings in which traffic is multiplexed to reduce capacity provisioning.

B. Robust Formulation for Gaussian Demands

So far we have assumed that demands obey a truncated Gaussian distribution. In this section, we extend the robust optimization formulation and associated probability bounds on constraint violation developed in the previous section to unbounded Gaussian demands.

Recall that the demands $\lambda_B^{sd}$ are upper bounded by $\mu_{sd} + k\sigma_{sd}$. In the capacity constraint in (7), $\Gamma_{ij}$ corresponds to
the number of demands that require additional capacity, and a capacity of \( k \sigma_{sd} a_{ij}^{sd} \) must be allocated to account for the worst case realization for each of those demands. It is easy to see that if \( \{X_k, k \geq 0\} \) is an independent sequence of continuous random variables, with \( X_k \) distributed according to a truncated gaussian described by (6), then the sequence of \( X_k \) converges in distribution to a Gaussian random variable with mean \( \mu_{sd} \) and variance \( \sigma_{sd}^2 \).

In equation (9), \( k \) and \( \Gamma_{ij} \) are inversely proportional, implying that as \( k \) increases, fewer demands are allocated extra capacity, but more capacity is allocated to those demands. Since Theorem 1 holds for all finite values of \( k \), \( \Gamma_{ij} \) eventually satisfies \( 0 \leq \Gamma_{ij} < 1 \) as \( k \) becomes large, so that \( S_{ij} \) in (7) is the empty set and the maximization is taken over one \( s-d \) pair representing \( t_{ij} \). Consequently, equation (7) can be rewritten as

\[
C_{ij} \geq \sum_{(s,d) \in D} a_{ij}^{sd} \mu_{sd} + \Gamma_{ij} k \max_{(s,d) \in D} \{ \sigma_{sd} a_{ij}^{sd} \}
\]

\[
= \sum_{(s,d) \in D} a_{ij}^{sd} \mu_{sd} + \sqrt{-2 n_{ij} \log \epsilon} \max_{(s,d) \in D} \{ \sigma_{sd} a_{ij}^{sd} \}, \quad (10)
\]

where we used the relationship (9). Using this constraint, the routing and capacity allocation problem can be formulated as follows:

\[
\min_{C, \alpha \geq 0} \max_{(i,j) \in E} C_{ij}
\]

s.t.:

\[
C_{ij} \geq \sum_{(k,l) \in D} a_{ij}^{kl} \mu_{kl} + \alpha \sqrt{\Gamma_{ij} \sigma_{sd} a_{ij}^{sd}}, \forall (s,d), \forall (i,j)
\]

\[
FC(s, d, a^{sd}), \forall (s,d) \in D
\]

(11)

where \( \alpha = \sqrt{-2 \log \epsilon} \). Note that \( n_{ij} \) is a variable determined by \( a_{ij}^{sd} \)'s, and therefore (11) is a nonlinear formulation. It can be reformulated into an mixed integer linear program (MILP), but due to space limitations, we omit the formulation (See [16] for more details). We can compare the optimal convex nonlinear formulation (3), its linearization (4), and our robust formulation (11). We use the following algorithm to solve our formulation:

- Every \( n_{ij} \) is initialized to 1.
- The optimization problem in (11) is solved, and the number of demands traversing each link is counted.
- If more than \( n_{ij} \) demands are routed on link \((i,j)\), then \( n_{ij} \) is increased by 1.
- The algorithm repeats until each link \((i,j)\) has \( n_{ij} \) higher than the number of demands routed on it.
- Link capacities are computed after termination by plugging the obtained solution \( a_{ij}^{sd} \)'s into (3).

This iterative scheme is called the Increasing Cost Algorithm (ICA), and is guaranteed to converge in at most \( N^2 \) iterations, where \( N \) is the number of links. The value of \( n_{ij} \) can increase during each iteration, but never decrease. Note that since some values of \( n_{ij} \) can become large and cannot be lowered, ICA may find a suboptimal solution to the problem in (11) since it can end up searching a restricted space that does not contain an optimal solution. However, each iteration only requires solving an LP, and thus the entire algorithm can be run in polynomial time.

We compare the performance of the optimal nonlinear formulation in (1), linear formulation with constraint (4), and ICA. As in the ICA, the capacity of linear formulation is recomputed using (3). Consider a network in Figure 2, where demands with \( \mu = 100 \) and \( \sigma = 35 \) are randomly placed. Note that this network allows for ample sharing of resources, allowing for a clear comparison of the different formulations.

Fig. 3 compares the performance of three different approaches. Fig. 3(a) shows that the maximum link capacity for the optimal solution offers a 15\% savings over that required by the linear approach. Our ICA approach allocates the same capacity as in the optimal approach. Fig. 3(b) plots the running time of each algorithm. The linear approach instantly finds a
solution, whereas the running time of the nonlinear approach dramatically increases as the number of demands increases. Note that our algorithm finds a solution much faster than the nonlinear approach. This shows that even though convex optimization problems can be efficiently solved in general, it is desirable to develop a formulation with better scaling properties for problems over large networks.

IV. ROBUST TOPOLOGY DESIGN

The formulation in the previous section can be used to develop network design algorithms that explicitly solve the link-placement sub-problem. This is necessary because the optimal link placement for stochastic traffic may differ from that for deterministic traffic. As an example, consider the following three-node network. Suppose we have i.i.d. traffic \( \mathcal{N}(10, \sigma) \) from node 1 to node 2 and node 1 to node 3. We want to design a network and route the traffic over such a network such that max link capacity is minimized, and the link placement is restricted to a limit of four links, assuming no parallel links. If traffic is deterministic \( (\sigma = 0) \), the optimal link placement only requires a link from 1 to 2 and 1 to 3, while the other two links can be placed arbitrarily as in Fig. 4(a). With this link placement, the traffic can be sent on the one hop path to its destination, with a maximum link capacity of \( C_{\text{max}} = 10 \).

Now consider the stochastic traffic case, by letting \( \sigma = 1 \). The optimal topology is shown in Figure 4(b). On this topology, half of each demand can be sent on each link, and a lower \( C_{\text{max}} \) is achievable than on the network in Figure 4(a). To be precise, the solution to (1) is \( C_{\text{max}} = 10 + \frac{\Phi^{-1}(1-\epsilon)}{\sigma^2} \). If the first topology is used to route the stochastic traffic instead, it would be sent directly to the destination, and a larger capacity of \( C_{\text{max}} = 10 + \Phi^{-1}(1-\epsilon) \) would be required.

Consequently, design algorithms for deterministic traffic are often insufficient to find optimal topologies for stochastic traffic. Instead, we propose that our robust optimization formulation be used as a capacity constraint alongside link placement and routing constraints, to create topologies which lend themselves to high amounts of resource sharing. Integer constraints can be added to the formulation of (11) to account for the link placement subproblem. As mentioned in the previous section, the capacity constraint (11) can be linearized, and thus the problem of topology design, routing and capacity allocation can be formulated a MILP (See [16] for details).

V. CONCLUSION

Modern backbone networks must be designed to be robust to random traffic fluctuations and errors in traffic estimation. Previous attempts to solve the routing and capacity allocation problems have either resulted in formulations which are too conservative in terms of allocated capacity, or formulations which are non-linear. By extending results from robust optimization, we have developed a formulation to allocate capacity in a less conservative fashion, while computing near optimal routes for traffic flows. The resulting routing for stochastic traffic takes advantage of multiplexing opportunities in the network, such that shared capacity is efficiently utilized. We provide an iterative scheme to solve our robust optimization problem, and show that our formulation finds near-optimal routings and capacity allocations, while remaining computationally tractable.

While our formulation allows an extension to the link-placement problem, the resulting ILP formulation is still computationally intractable. Our analysis shows that networks supporting random traffic should be designed in a manner to allow link capacity to be shared by many traffic flows. Therefore, it would be interesting to develop algorithms and heuristics to solve the link placement subproblem exploiting these strategies.

REFERENCES