# Diversity Routing to Improve Delay-Jitter Tradeoff in Uncertain Network Environments

Arman Rezaee, *Student Member*, Vincent W.S. Chan, *Life Fellow IEEE*, *Fellow OSA*
Claude E. Shannon Communication and Network Group, RLE
Massachusetts Institute of Technology, Cambridge MA, USA
Emails: {armanr, chan}@mit.edu

***Abstract*—In this paper we propose a novel approach to deliver better delay-jitter performance in dynamic networks. Dynamic networks experience rapid and unpredictable fluctuations and hence, a certain amount of uncertainty about the delay-performance of various network elements is unavoidable. This uncertainty makes it difficult for network operators to guarantee a certain quality of service (in terms of delay and jitter) to users. The uncertainty about the state of the network is often overlooked to simplify problem formulation, but we capture it by modeling the delay on various links as general and potentially correlated random processes. Within this framework, a user will request a certain delay-jitter performance guarantee from the network. After verifying the feasibility of the request, the network will respond to the user by specifying a set of routes as well as the proportion of traffic which should be sent through each one to achieve the desired QoS. We propose to use *mean-variance analysis* as the basis for traffic distribution and route selection, and show that this technique can significantly reduce the end-to-end jitter because it accounts for the correlated nature of delay across different paths. The resulting traffic distribution is often non-uniform and the fractional flow on each path is the solution to a simple convex optimization problem. We conclude the paper by commenting on the potential application of this method to general transportation networks.**

## I. Introduction

Major networking companies are preparing for an unprecedented growth in adoption of time-sensitive and high bandwidth applications in the upcoming decade. According to Cisco, IP video traffic will be 82 percent of all IP traffic by 2021, up from 73 percent in 2016 [1], [2]. The same reports forecast live video to grow 15-fold while virtual reality (VR) and augmented reality (AR) traffic will increase 20-fold in the same period. The interactive nature of these applications requires low latency but more importantly they are extremely sensitive to "variation of delay", which is often referred to as ***jitter***. In packet switched networks, jitter is often defined as the standard deviation of packet delay, and we will use this definition of jitter in our analysis and exposition.

Other applications such as high frequency trading, and tele-surgery are also extremely sensitive to jitter. For example, high frequency traders would like to guarantee that their orders reach various exchanges at the same time ($<$1 ms), otherwise the execution of the first order at a given exchange may reveal their intent to other investors who can manipulate market prices by front-running the rest of their orders. Similarly, surgeons who want to conduct a remote operation on a patient (i.e. tele-surgery) expect a network that can deliver responsive and jitter-free haptic feedback.

The proliferation of these applications presents a chicken and egg problem for network engineers: on the one hand, these applications require low latency and jitter but at the same time the bursty and dynamic nature of their traffic introduces unpredictable delay and amplifies the jitter. The continuous variation and abrupt changes introduced by exogenous traffic will create temporary bottlenecks in the network which manifest themselves as jitter. Traditionally, packet buffers were deployed to combat the negative effects of jitter. When the instantaneous packet arrival rate at a queue exceeds its output rate, packets are stored in the buffer until they can be transmitted through outgoing ports. Not surprisingly, increasing the buffer size is not an attractive solution as it is costly and will increase the potential for increased delay and jitter. Other solutions include over-provisioning the network or providing dedicated paths/circuits to such applications, both of which are not economical.

Last but not least, we may think that jitter can be accounted for by a Network Management and Control (NMC) system that continuously monitors the state of the network to guarantee the desired QoS. Unfortunately, current NMC systems are much too slow to track the state of various network elements by the desired level of precision. Furthermore, tracking the state of dynamic networks of the future will be a rather costly undertaking and methods for optimal tracking of the network state are themselves the focus of current research projects [3]. Hence, any meaningful solution should strive to meet application demands despite the unavoidable uncertainty about the instantaneous state of the network

This brings us to the ultimate question: *can we accommodate these new applications with their stringent latency and jitter requirements despite our relative uncertainty about the state of the network and without massive over-provisioning?* This paper will demonstrate that in most cases we can answer this question with a (resounding) yes. The solution involves an innovative technique to distribute the traffic flow over multiple paths in such a way that guarantees lower end-to-end jitter despite the delay variations on individual links. This may seem counterintuitive at first, but as demonstrated in the following sections if we account for the correlated nature of delay across various paths we can trade slightly higher average delay for significantly reduced jitter.

This work is sponsored by NSF NeTS program, Grant No. #6936827 and the HKUST/MIT programs.

This novel solution is inspired by Harry Markowitz's Nobel prize winning work on *portfolio selection* [4]. His work has been instrumental in construction of investment portfolios that exhibit a pre-determined risk-return behavior. We do not seek credit for any of the mathematical formulations and/or developments of this subject which have been exhaustively studied in economics literature. On the other hand, we are unaware of other works that apply these ideas to communication networks and specifically questions of delay and jitter. We refer interested readers to [5] and [6] for a short history on the development of Modern Portfolio Theory, as well as the mathematical derivations and consequences of the theory. The namesake "diversity routing" has been chosen to draw parallels to diversification of financial investments.

The rest of the paper is organized as follows: Section II introduces the general model under which diversity routing is considered. Section III casts the optimal allocation of traffic as a convex quadratic optimization problem and describes the solution space. Section IV discusses the theoretical limits of diversification. Sections V incorporates additional cost criteria into the optimization framework. Section VI extends the results to general transportation networks. Discussion of our contributions as well as future works is given in Section VII. Concluding remarks are provided in Section VIII.

## II. General Model

Consider a network management and control system that monitors the state of the network at all layers, reconfigures network resources when necessary, and provides data and instructions to applications upon request. More specifically, when an application requires network resources, it will contact the NMC system and specify its requirements, including delay, jitter, and bandwidth. It is preferable for the application to specify a few possible variations of its desirable requirements, each corresponding to a different QoS and/or QoE level. The NMC system will in turn evaluate the feasibility of the requests, and respond by specifying the routes that should be used to achieve the highest possible QoS and/or QoE levels. If the network, in its current state, is unable to satisfy the application's demand, the NMC system would either reconfigure the network to meet the requirements or reject the request. Figure 1 illustrates of such an interaction. The following discussion exemplifies this process in the context of routing with delay and jitter requirements.

Let us suppose that an origin-destination (OD) pair is connected via $n$ paths $P_1, \ldots, P_n$. Based on the information available to the NMC system, packets transmitted on path $P_i$ will experience a delay $d_i$, where $d_i$ is a random variable with known mean and variance,

$$\mu_i = \mathbb{E}[d_i], \qquad \sigma_i^2 = \mathrm{Var}[d_i]$$

recall that jitter is defined as the standard deviation of delay, and thus $\sigma_i$ denotes the jitter on path $P_i$. In general, the delay incurred on these paths are not independent.[1] In what

[1]This fact is contrary to the assumption used in many queuing theory texts, known as the Kleinrock Independence Approximation [7].

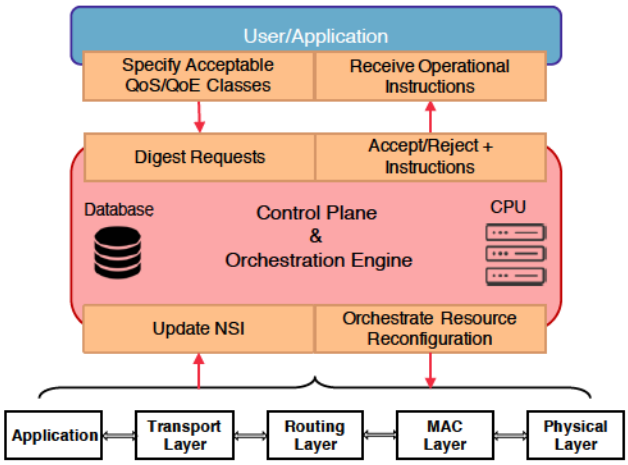


Figure 1: Roles and responsibilities of the NMC system: 1) Determine feasibility of application requests, 2) Instruct the application of operational requirements, 3) Orchestrate reconfiguration of network resources.

follows, we use $\sigma_{i,j}$ to denote the covariance between delays on paths $P_i$ and $P_j$, i.e. $\sigma_{i,j} = \mathrm{Cov}(d_i, d_j)$. Thus, we have a covariance matrix $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_n^2 \end{pmatrix}$$

Figure 2 illustrates the relationship between a few arbitrarily chosen paths and their respective mean delay and jitter. Clearly, applications that utilize this network for data transport are affected by the delay performance of its individual paths. But, can the network as a whole provide delay characteristics which outperform the convex hull created by individual path characteristics? The following section demonstrates how diversity routing enables applications to meet delay/jitter requirements that exceed the performance of individual paths as well as the convex hull of their performance.

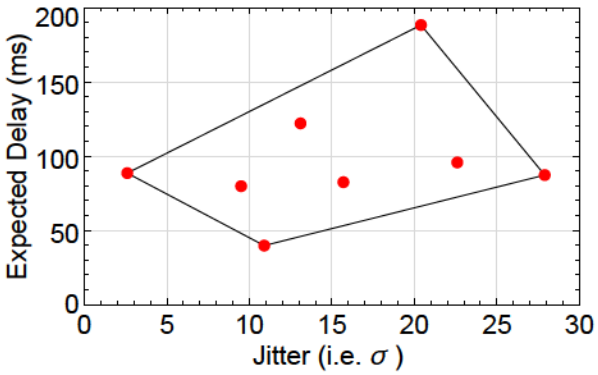


Figure 2: Mean delay vs. jitter for a few paths, and the resulting convex hull.

We would like to acknowledge that in current systems, the network is often unaware of the user's specific quality of service requirements (except for some limited SDN services). Our proposal requires a deliberate negotiation between the user and the NMC system to convey such requirements in order to achieve better efficiency and performance. We would like to be clear that our decision to take this (unconventional) approach is a conscious trade-off.

## III. Optimal Traffic Allocation

### A. Formulation

Consider a routing algorithm that assigns a fraction $f_i$ of the total flow to path $P_i$. Let us use $\mathbf{F}$, and $\mu$ to denote the vector of fractions, and vector of mean delays respectively,

$$\mathbf{F} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Note that each vector $\mathbf{F}$ corresponds to a unique *Traffic Allocation*. The mean and variance of delay for a given traffic allocation can be computed as

$$\begin{aligned} \mathbb{E}\left[d_{TA}\right] &= \mathbb{E}\left[\sum_{j=1}^{n} f_j d_j\right] = \sum_{j=1}^{n} f_j \mu_j = \mathbf{F}^T \mu \\ \mathrm{Var}\left[d_{TA}\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n} f_i f_j \mathrm{Cov}(d_i, d_j) = \mathbf{F}^T \mathbf{\Sigma} \mathbf{F} \end{aligned}$$

Given the aforementioned quantities, how can we define the "optimal" traffic allocation? An optimal allocation may refer to one that minimizes the expected delay or jitter, or a combination of them. Since expectation is a linear operation, the expected delay of the allocation is simply the weighted linear combination of individual mean delays and is thus minimized if the entire traffic is allocated to the path with the lowest expected delay. On the other hand, variance of delay is a quadratic function of the traffic allocation $\mathbf{F}$, and the allocation that minimizes jitter depends on the covariance matrix $\mathbf{\Sigma}$. One natural way to incorporate both criteria into an optimization framework is to find the minimum-jitter allocation that achieves a pre-specified expected delay, $\mu^*$. Noting that jitter is minimized when $\mathrm{Var}\left[d_{TA}\right]$ is minimized, the optimization can be written as,

$$\begin{aligned} \underset{\mathbf{F}}{\text{minimize}} \quad & \mathbf{F}^T \mathbf{\Sigma} \mathbf{F} \\ \text{subject to} \quad & \mathbf{e}^T \mathbf{F} = 1 \\ & \mathbf{F}^T \mu = \mu^* \\ & 0 \le f_i \le 1, \; i = 1, \ldots, n. \end{aligned}$$

where $\mathbf{e}$ denotes a vector of all ones, and the constraint $\mathbf{e}^T\mathbf{F} = 1$ ensures that fractional flows sum-up to one.

Algorithmically speaking, the application specifies a pair of numbers $(\mu^*, \sigma^*)$ to the NMC system, representing the maximum acceptable average delay and jitter respectively. The NMC system evaluates the aforementioned optimization to determine the feasibility of the request. If a feasible traffic allocation exists, the application's request is accepted and appropriate routing information is provided by recommending a specific allocation. If the request is infeasible, the NMC system will either reject the application's request or will reconfigure the network in such a way to accommodate the original request. Network reconfiguration tactics are outside the scope of this paper and is left as future work.

### B. Solution

The investigation of minimum variance allocations for a desired average performance, as expressed above, was originally proposed by Harry Markowitz in the context of portfolio theory and allocation of financial assets [4]. In that context, he expressed the goal of rational investors as hoping to allocate/invest their assets in such a way to achieve the lowest risk (lowest standard deviation) for a desired expected return on investments. In a similar manner, we wish to identify a traffic allocation that achieves the lowest jitter for a desired expected delay.

Before getting into the details of the optimization, let us consider the simplest possible setting whereby an OD pair is connected via exactly two paths. Let us denote the expected delay and jitter characteristics of each path as a point on the Cartesian plane, and suppose that the red dots in Figure 3 represent the characteristics of these two paths. The first path, $P_1$, has a mean delay of 150 ms and jitter of 15 ms while the second path, $P_2$, has a mean delay of 50 ms but a jitter of 20 ms. In this case our traffic allocation vector is $\mathbf{F} = (f_1, f_2)^T$, where $f_2 = 1 - f_1$. Hence, we can determine performance of all traffic allocations by sweeping the $f_1$ parameter between 0 and 1. Each line in Figure 3 traces the set of achievable mean delay and jitter combinations for a specific correlation coefficient. Note that by transmitting through both paths we can obtain an overall jitter that is significantly lower than that afforded by either of the individual paths. In particular, if the delay on the two paths are perfectly anti-correlated, i.e. $\rho = -1$, there exists a traffic allocation that can achieve a jitter-free performance. For our example, this jitter-free point occurs when traffic allocation fractions are chosen to be $(f_1, f_2) = (0.43, 0.57)$.

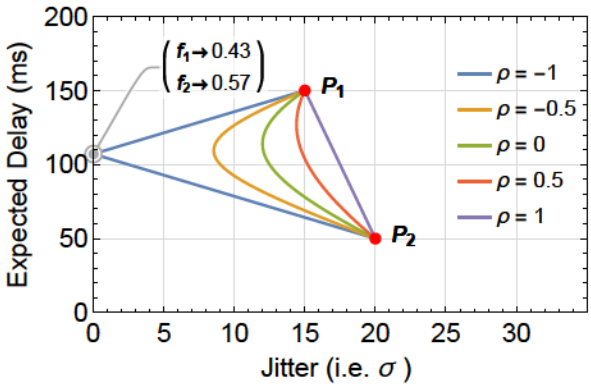


Figure 3: Achievable delay-jitter combinations using diversity routing for various correlation coefficients.

We should also note that a particular jitter requirement can be satisfied at two different mean delays (corresponding to two different traffic allocations). In the absence of additional selection criteria, we should always choose the traffic allocation that has a smaller mean delay. In other words, we will always be interested in the bottom portion of the delay-jitter traces. The set of traffic allocations that constitute the bottom portion of the curves will be referred to as the set of *Efficient Allocations*, following a similar naming convention in [4].

Fortunately, the same basic behavior is observed when the number of paths increases, as shown in the following simulated scenario. Suppose that the NMC system has observed the instantaneous delay corresponding to 9 paths that connect a particular origin-destination pair over a long period of time. A possible realization of the observed instantaneous delays is depicted in Figure 4. The delay traces correspond to a set of correlated random processes whose mean and standard deviation (i.e. jitter) correspond to the grid of black dots shown in Figure 6.[2]

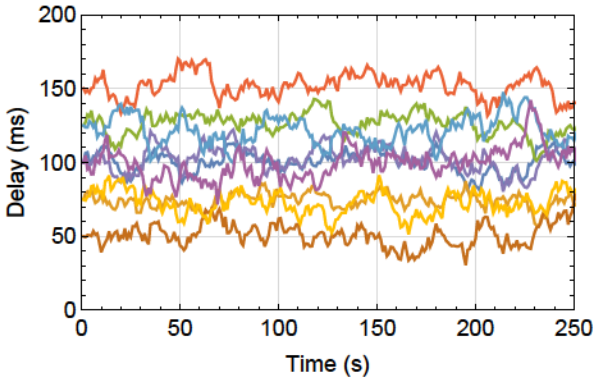


Figure 4: Simulated delay of 9 paths connecting an origin-destination pair.

The NMC system can compute the correlation matrix (or equivalently the covariance matrix) corresponding to these observations, as depicted in Figure 5. We can then numerically solve the following convex optimization problem, for all feasible values of $\mu^*$. Feasible values of $\mu^*$ are those that fall between the minimum mean-delay and the maximum mean-delay of the 9 paths, i.e. between 50 ms and 150 ms,

$$\begin{aligned} \underset{\mathbf{F}}{\text{minimize}} \quad & \mathbf{F}^T \mathbf{\Sigma} \mathbf{F} \\ \text{subject to} \quad & \mathbf{e}^T \mathbf{F} = 1 \\ & \mathbf{F}^T \mu = \mu^* \\ & 0 \leq f_i \leq 1, \; i = 1, \ldots, n. \end{aligned}$$

Figure 6 depicts the solution of the aforementioned optimization for all feasible expected delays. The nine black dots on the right-hand side of the figure represent the mean delay and jitter of each of the individual paths. Each blue point on the left represents a specific traffic allocation vector,

[2]More specifically, these sample paths were drawn from a set of correlated Ornstein-Uhlenbeck processes with pre-specified mean delay and jitter and are good candidates for our demonstration purposes.

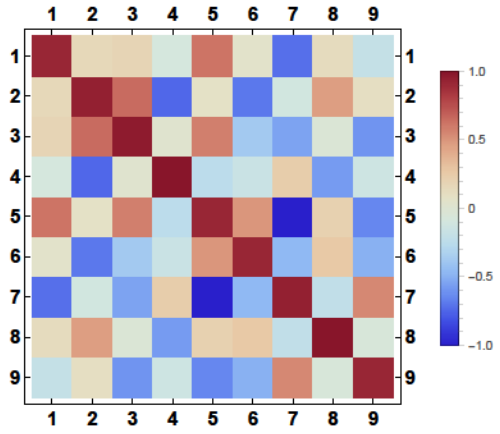


Figure 5: Correlation matrix corresponding to the delay observations of the 9 paths shown in Figure 4.

and the resulting mean delay and jitter. Note that the blue points present a significantly reduced jitter in comparison to the original paths. Finally, the red dot corresponds to the "minimum-jitter allocation". The network cannot support an application that requires more stringent jitter than that afforded by the allocation corresponding to the red dot. As shown in the figure, the minimum-jitter allocation sends most of its traffic through paths $P_5$ and $P_7$ as denoted by $f_5 = 0.51, f_7 = 0.45$.

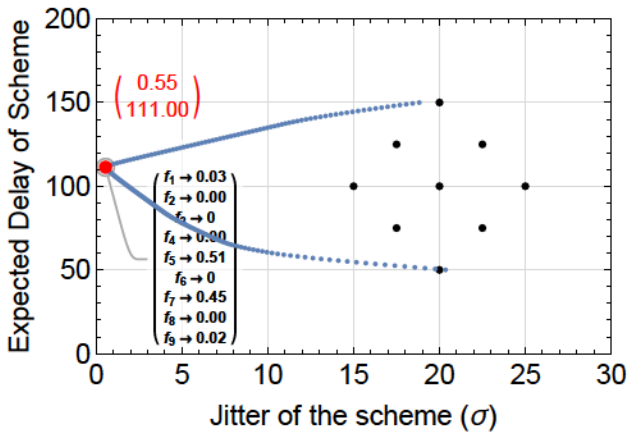


Figure 6: Optimal delay-jitter combinations and the characteristics of the original paths.

To get some insight about this, we draw your attention to the significant negative correlation between paths $P_5$ and $P_7$ as shown in Figure 5. Not surprisingly, the overall jitter is reduced when the flow is split amongst negatively correlated paths. Negative correlations can arise in many situations in real networks. One such instance is exemplified by Autonomous Systems (AS) that compete for traffic share by advertising different costs to a given destination. When an AS advertises a cheaper route, it will attract traffic from other AS's, resulting in negatively correlated delay on the respective paths. Another example is caused by the cyclical nature of traffic demand which corresponds to the time of day, hence, geographical areas that are offset by certain time

differences will exhibit negatively correlated behaviors. We shall wrap up this section by reiterating that the lower half of the plot corresponds to *Efficient Allocations*.

## IV. Limits to Diversification

The successful examples of the last section may lead us to wonder whether we could completely eliminate jitter by using additional paths. As we will see, there are limits to the diversification effect afforded to us by the presence of additional paths. The discussions of this section will closely follow the developments of similar material in section 7.3 of [8] which showed the limits of diversification in context of Modern Portfolio Theory.

Recall the expression for the variance of a given traffic allocation, and rewrite it as

$$\begin{aligned}\text{Var}\left[d_{TA}\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n} f_i f_j \text{Cov}(d_i, d_j) \\ &= \sum_{i=1}^{n} f_i^2\sigma_i^2 + 2\sum_{i<j} f_i f_j \sigma_{i,j}\end{aligned}$$

Clearly, variance of delay in individual paths contributes $n$ terms to the sum while the covariances contribute approximately $n^2$ terms to the sum. This simple observation signifies the importance and contribution of covariances/correlations between various paths which can easily outweigh the jitter of individual paths! It can be shown that the contribution of the variances can be eliminated through the introduction of additional paths, but the covariances will dominate and constitute the bulk of the remaining jitter. The following example is often used to convey the aforementioned idea. Let us consider the case of equal-splitting of the traffic amongst all $n$ paths, i.e. $f_i = 1/n$. Then:

$$\begin{aligned}\text{Var}\left[d_{TA}\right] &= \sum_{i=1}^{n} f_i^2\sigma_i^2 + 2\sum_{i<j} f_i f_j \sigma_{i,j} \\ &= \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma_i^2 + 2\sum_{i<j} \left(\frac{1}{n}\right)^2 \sigma_{i,j} \\ &= \frac{\text{(Avg. Var)}}{n} + \left(1 - \frac{1}{n}\right)\text{(Avg. Covar)}\end{aligned}$$

interestingly, as $n \to \infty$, the first term (i.e. the contribution of variances) will become 0, and the only remaining factor is the average covariance of delay. In other words, diversity routing reduces jitter by incorporating paths whose average delay covariance is negligible. The aforementioned analysis is often the basis of the common practice that dictates "diversification reduces risk" in financial literature. We conclude this section by presenting the following simple bound on the minimum achievable jitter,

$$\frac{1}{\sqrt{\mathbf{e}^T\Sigma^{-1}\mathbf{e}}} \quad \leq \quad \text{Minimum Jitter}$$

This bound is derived in Appendix A and can be used by the NMC system to reject application requests that are in conflict with this bound.

## V. Generalized Cost Function

It should come as no surprise that mean delay and jitter are not the sole criteria for path selection in diversity routing; but how can we incorporate additional cost criteria into the model? The most natural way of adding cost criteria is to realize that transmission over different paths may have different "costs". One reason for this may be the heterogeneity of the underlying physical layer. For example, a given path may be a fiber optic while another one is a satellite link. Even in homogeneous networks where transmission over all links have the same cost, we can associate a cost with the "length" of a given path. Clearly, a path consisting of 3 links will have 3 times the cost as a path with 1 segment. Last but not least, we should recognize that real networks (e.g. the US fiber backbone) are often a collection of independently owned and operated subnets (often referred to as Autonomous Systems). Hence, we should expect various vendors to charge different amounts for using their systems. Either way, we can easily associate and incorporate the respective costs with each path.

Let us use $\mathbf{C}$ to denote a cost vector, whose $i^{\text{th}}$ element $c_i$ denotes the cost per unit flow over path $P_i$. We can then rewrite our optimization problem as,

$$\begin{aligned}\underset{\mathbf{F}}{\text{minimize}} \quad & \mathbf{C}^T\mathbf{F} + \mathbf{F}^T\mathbf{\Sigma}\mathbf{F} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{F} = 1 \\ & \mathbf{F}^T\mu = \mu^* \\ & 0 \leq f_i \leq 1,\ i = 1, \ldots, n.\end{aligned}$$

One way to interpret this new formulation is to think of a service provider that has to balance two competing goals. The first goal is to reduce the transportation cost as captured by $\mathbf{C}^T\mathbf{F}$ and the second goal is to reduce the potential loss of revenue associated with delivering lower QoS. This loss of revenue may reflect the immediate drop in customer satisfaction or the eventual customer defection caused by subpar QoS. In effect, we have taken variance of delay, $\mathbf{F}^T\mathbf{\Sigma}\mathbf{F}$, as a stand-in for this loss of revenue, reflecting our preference for lower jitter. It is clear that the formulation could be further generalized by using a convex function of $\mathbf{F}^T\mathbf{\Sigma}\mathbf{F}$ as the second term of the objective function, but we shall sacrifice that generality in favor of simplicity, for now.

Note that the aforementioned formulation is still a convex quadratic optimization whose solution is obtained as easily as before, and hence we will refrain from additional discussion of the solution space except for the following example. Let us revisit the routing example used in section III-B and incorporate a specific cost vector $\mathbf{C}$ as shown in Figure 7. We have assigned the paths to 3 different cost groups, (150, 100, and 50). While the numbers were chosen arbitrarily, it is reasonable to expect an increasing trend as we move towards the bottom-left corner of the plot, because this direction corresponds to paths that exhibit lower delay and lower jitter.

Once again, the blue dots constitute an "optimal" traffic allocations for each value of expected delay, with the caveat that the minimum-cost allocation is no longer on the leftmost point on the plot.

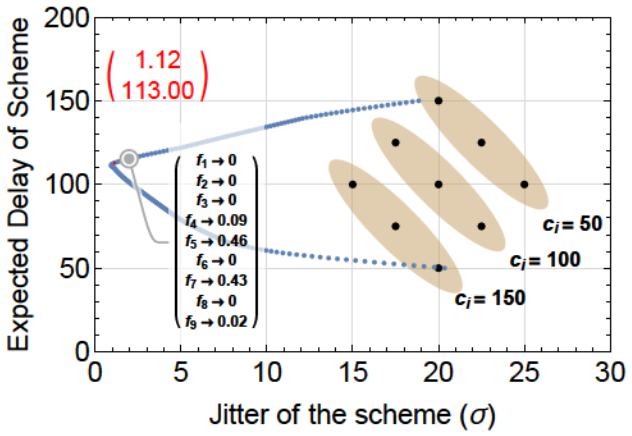


Figure 7: Optimal delay-jitter combinations, with indifference curves that correspond to path costs of 150, 100, and 50.

## VI. Generalization to Transportation Networks

Our analysis has so far focused on the importance of diversity routing in communication networks. Fortunately, our proposed mechanism can be used to reduce uncertainty in delivery time of goods over general transportation networks.

As an example, consider a retail store in Boston that wants to receive a steady supply of a given product from New York City. Commonly, retail stores contract a logistics and transportation company to transport the products from NYC to Boston. If the logistics company uses one mode of transportation (*e.g.* trucks), the exact delivery time can be impacted by the often unpredictable road conditions. On the other hand, the uncertainty in the delivery time of the products can be minimized, if multiple modes of transportation (such as air, sea, railroad, etc.) are used. It is important to recognize that this improvement is due to the fact that different modes of transportation are affected by different factors, and thus conditions that impact one mode of transportation are often different from those that impact another. For example, note that an accident on a road is unlikely to be related to conditions of shipping lanes. By accounting for the correlation of delay on various modes of transportation (or various roads), the logistics company can deliver the goods on a more regular basis (*i.e.* lower jitter). Note that regular and steady delivery of goods can be immensely important to the retail store as well, because it will eliminate the cost of excessive local storage and warehousing for the retail company.

A similar argument can be used to reduce the uncertainty in other aspects of supply-chain management. For example, a company can order raw materials from multiple suppliers in such a way as to reduce the uncertainty in their arrival rate, where the fraction ordered from each supplier is computed according to our formulation. We believe that our proposed method can be used to systematically achieve high-level managerial goals which are often referred to as "just-in-time manufacturing" or "lean manufacturing".

## VII. Discussion and Future Work

In this section we discuss some of the overarching principles which should be considered with regards to the adoption of our diversity routing mechanism. Recall that our treatment of diversity routing started with a network management and control system that has visibility to all layers. This included the ability to monitor the state and performance characteristics of various elements, as well as orchestration and resource reconfiguration capabilities. Resource reconfiguration may include tasks such as addition or removal of wavelengths on a particular fiber connection, which is currently carried out by human operators. More importantly, the NMC system will interact with applications to identify appropriate routes that can deliver a desired level of service. This challenges the conventional wisdom that networks should avoid any coordination or interaction with applications.[3] This long-held strategy has forced a whole host of responsibilities to the end-user terminal. For example, rate control, congestion control and backoff algorithms are largely delegated to the communication end-point and the rapid growth of internet access is often attributed to this choice.

We challenge this paradigm by promoting a user-centric view that expects the network to do its best to deliver the desired quality of service to the user. Of course, this approach comes at the expense of additional complexity to the network, but we believe that this added complexity can be justified when it enables the rapid adoption of next-generation applications. Simply put, current networking practices may impede the arrival of new applications that will constitute the next wave of innovation. Of course, introduction of additional complexity will have diminishing return and thus the appropriate level of complexity should be investigated.

On a related note, we should point-out that we have not addressed the security issues that arise when the network interface is opened to various applications. Not surprisingly, malicious applications may leverage this ability to manipulate and/or attack the network by making requests which can result in misallocation of resource and ultimately resource exhaustion within the network. This topic is of immense importance and will be the focus of future investigations.

We should emphasize that diversity routing or multiple-path routing is not an entirely new idea. As far back as 1998, the Internet Engineering Task Force was considering the use of multiple paths to achieve QoS-based routing [9]. Singh *et al.* provide a detailed survey of various such routing schemes in [10], and we shall address a few major differences in our approach vs. the prior art. Most prior work concentrate on throughput maximization as their central objective and not surprisingly using *all* available paths is the simplest way to achieve this goal. Furthermore, most of their analyses considers a static network as opposed to a truly dynamic network. Note that from an optimal routing perspective, static vs. dynamic is simply a matter of the precision by which network state (*e.g.* congestion/load) is known. Clearly,

[3]Often phrased as "dumb networks are the smart choice".

optimal decisions can be made if the precise network state is known at all times. The overarching assumption in the prior work is that the network state is either fixed or varies slowly enough to ensure that underlying routing algorithms have precise and consistent view of the network. As a result, their formulation does not account for the unavoidable uncertainty in the state of the network and overlooks the fact that routing decisions should be made despite this uncertainty. In our approach, the uncertainty in delay characteristic of a link/path is captured by variance of delay on each path, denoted by $\sigma_i^2$, which can be computed from the historical behavior of a given link. Another unique feature of our development is that we account for and utilize the correlation between various links to achieve higher quality of service. This is in contrast to traditional approaches that disregard the presence of correlated behavior and often assume independence to achieve/design simpler operating paradigms.

Additionally, most authors employ a narrow network-centric approach in their formulation. These approaches lead network architects to attach undue value to goals that are certainly reasonable but secondary in nature. For example, multiple-path routing is often used to achieve load balancing and avoid undesirable/intolerable oscillatory behavior in the network. But load balancing should be a byproduct of clever network design and operation and not its primary purpose. Our formulation uses a combination of factors, such as delay and jitter as the primary design parameters and achieves a certain level of load-balancing as a byproduct of our solution.

Last but not least, we should mention that communication networks can suffer from out of order packet delivery associated with multiple-path routing. It suffices to say that this issue can be handled separately and in fact error correcting codes can be utilized to address the effects of out of order packet delivery.

## VIII. Conclusion

In this paper, we introduced a new mechanism for efficient allocation of traffic across a diversified set of paths. This allocation allows the network to deliver customizable quality of service to different users and reduces the need for buffers at various network elements. Our work focused on the tradeoff between mean delay and jitter as the main contributors to QoS. An important feature of this approach is its ability to achieve the desired QoS despite the relative uncertainty about the state of the network. Noting that the introduction of demanding (and data hungry) applications often outpace that of network upgrades, we have argued that our innovative solution can accelerate the adoption of these applications without the need for immediate capital expenditure. We concluded our remarks by extending our findings to general transportation networks and argued that this approach can significantly improve the supply chain predictability and reduce the need for storage facilities.

## Appendix

By relaxing the positivity constraints on $f_i$'s, we obtain an analytical solution to the relaxed optimization problem via a Lagrange multiplier. Let us write the Lagrangian as

$$\mathcal{L}(\mathbf{F},\ell) \;=\; \frac{1}{2}\;\mathbf{F}^T\mathbf{\Sigma}\mathbf{F} + \ell(1-\mathbf{e}^T\mathbf{F})$$

which can be solved as the solution to $\frac{\partial\mathcal{L}}{\partial f_i} = \frac{\partial\mathcal{L}}{\partial \ell} = 0$. Where

$$\frac{\partial\mathcal{L}}{\partial f_i} = f_i\sigma_i^2 + \sum_{j\neq i} f_j\sigma_{i,j} - \ell = 0$$

rewriting the solution as a matrix gives us $\mathbf{\Sigma F} = \ell\mathbf{e}$ or equivalently, $\mathbf{F} = \ell\Sigma^{-1}\mathbf{e}$. Noting that $\mathbf{e}^T\mathbf{F} = 1$, we get

$$\begin{aligned}\mathbf{e}^T\mathbf{F} &= \mathbf{e}^T\left(\ell\Sigma^{-1}\mathbf{e}\right) = \ell\mathbf{e}^T\Sigma^{-1}\mathbf{e} = 1\\ \ell &= \frac{1}{\mathbf{e}^T\Sigma^{-1}\mathbf{e}}\end{aligned}$$

which gives us the following allocation

$$\mathbf{F} \;=\; \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}^T\Sigma^{-1}\mathbf{e}}$$

Let us use $U$ to denote this unconstrained traffic allocation. Then we have the following mean and variance for the delay:

$$\begin{aligned}\mathbb{E}[d_U] &= \mathbf{F}^T\mu = \left(\ell\Sigma^{-1}\mathbf{e}\right)^T\mu\\ &= \ell\mathbf{e}^T\left(\Sigma^{-1}\right)^T\mu = \ell\mathbf{e}^T\Sigma^{-1}\mu = \frac{\mathbf{e}^T\Sigma^{-1}\mu}{\mathbf{e}^T\Sigma^{-1}\mathbf{e}}\\ \mathrm{Var}[d_U] &= \mathbf{F}^T\mathbf{\Sigma}\mathbf{F} = \mathbf{F}^T\ell\mathbf{e} = \ell\mathbf{e}^T\mathbf{F} = \ell\end{aligned}$$

Recall that we ignored the positivity constraints on $f_i$, and hence the aforementioned variance, is a lower bound to the achievable minimum variance. If we use minVar to denote the minimum achievable variance for operationally feasible traffic allocations we have

$$\frac{1}{\mathbf{e}^T\Sigma^{-1}\mathbf{e}} \;\leq\; \mathrm{Var}[d_{\mathrm{minVar}}]$$

## References


[1] "The zettabyte era: Trends and analysis," June 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

[2] "Cisco vni: Forecast and methodology, 2016-2021," June 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[3] A. Rezaee and V. Chan, "Cognitive network management and control with significantly reduced state sensing," in *Global Telecommunications Conference (GLOBECOM 2018)*. IEEE, 2018, pp. 1–7.

[4] H. Markowitz, "Portfolio selection," *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.

[5] H. M. Markowitz, "The early history of portfolio theory: 1600–1960," *Financial Analysts Journal*, vol. 55, no. 4, pp. 5–16, 1999.

[6] E. Elton, M. Gruber, S. Brown, and W. Goetzmann, *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, 2009. [Online]. Available: https://books.google.com/books?id=aOtcTEQ3DAUC

[7] L. Kleinrock, *Queueing Systems Volume I: Theory*. New York: John Wiley & Sons, 1975, vol. 1.

[8] R. A. Brealey, S. C. Myers, F. Allen, and P. Mohanty, *Principles of corporate finance*. Tata McGraw-Hill Education, 2012.

[9] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, "RFC2386: A Framework for QoS-Based Routing," *Network Working Group*, 1998.

[10] S. K. Singh, T. Das, and A. Jukan, "A survey on internet multipath routing and provisioning," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2157–2175, Fourthquarter 2015.