Overlaying Delay-Tolerant Service using SDN

Patrick Maillé
Telecom Bretagne
Rennes, France
patrick.maille@telecom-bretagne.eu

Shyam Parekh
AT&T Labs
San Ramon, USA
shyam.parekh@att.com

Jean Walrand
University of California
Berkeley, USA
walrand@berkeley.edu

***Abstract*—Telecommunication networks are generally dimensioned to provide services with small delays and high throughput during peak-periods. Due to the sizable difference in the network utilization between the peak and off-peak periods as well as the requirements of robust performance in face of both traffic burstiness and various types of network failures, these networks are significantly over-dimensioned for the average network loads.**

In this paper, we propose to use this extra capacity for supporting a deferrable traffic class with some guarantees on its end-to-end delays. Using the Software-Defined Networking (SDN) capabilities for controlling the network ingress rates of the deferrable traffic class in real time, we ensure that such a service would remain transparent to existing delay-sensitive traffic. To estimate the available capacities for the deferrable service, we analyze large deviations for the proposed traffic model.

Starting from an initial network designed for delay-sensitive traffic, one can readily "overlay" a new network for the deferrable service at no extra cost. This overlaid network has the same topology as the original one, and its link capacities can be directly computed from the characteristics of the existing traffic, the original link capacities, and the end-to-end delay tolerances.

I. Introduction

Telecommunication networks have been witnessing an exponential increase in traffic volumes since the 1990s, driven in the last years by the widespread adoption of cloud services, the generalization of 4G mobile usage, and the user consumption changes from television to video streaming. This trend is very likely to continue with the advent of 5G and the higher definition of videos viewed online, putting again more pressure on the infrastructure owners to increase transmission capacities.

A comparable increase in demand is observed in electric power distribution networks. There, many solutions are envisioned to reduce infrastructure, production, and/or environment costs by smoothing out the (also highly variable) demand. Those solutions include *deferring* part of the demand in exchange for a lower unit price, and quantitative analyses show how much can be saved, for example when different types of demand have different deadlines [4], [5].

Telecommunication network features differ from electric distribution, including faster variations over time and the absence of alternative sources of supply. (Note that [4], [5] focus on the use of renewable energy, but assume that grid power is always available.) Nevertheless, we believe the idea of deferred traffic–treatable within a deadline with high probability–is worth investigating also there. A typical example is for video on demand: consumers could be asked to select a movie *in advance*, which would then be "pushed" through the network within a deadline, using only the capacity left unused by other flows instead of being downloaded or streamed as a delay-sensitive flow. Such a new service would then be transparent to existing traffic, and could help postpone capacity investments through a more efficient use of the existing infrastructure.

In this paper, we use large deviations analysis [16], [18] to estimate the amount of capacity that could be used by deferrable traffic. The idea is to control the probability that the average capacity available over some duration T is insufficient to carry some amount of deferrable traffic: given T and a target failure probability, we compute a corresponding capacity for deferrable traffic. While the analysis only provides results for the *rate* at which that probability decreases with T, simulation results show that ignoring smaller-order terms leads to very good estimates of the available capacity for deferrable traffic.

In terms of implementation, our approach relies on the current Software-Defined Networking (SDN) efforts [1], [11], in that it can leverage the use of logically centralized controllers, aware of the current network conditions, to inject deferrable traffic so as to remain transparent to delay-sensitive flows. Beyond the controller, some other management tools can also be applied, ranging from lower prioritization of deferrable traffic to more elaborate methods aimed at reducing the need for buffering in intermediate nodes, such as the Fastpass approach proposed in [15]. Since deferrable traffic will use the volatile resource left available by non-deferrable flows, we can also imagine that the routing applied to deferrable traffic be subject to rapid changes in order to optimize the instantaneous throughputs; this again implies the knowledge of the current network states, and the capacity to impact rapidly the behavior of routers through interfaces such as OpenFlow [13].

We are not the first ones to apply large deviations to analyze delays. In [17] the focus is on scheduling jobs in a multi-class queue so that out-of-time probabilities decrease at target rates for each class when the tolerated delay increases. Considering the network aspect, the large deviations of a network of G/G/1 (single-class) queues are analyzed in [3], also at the job (or packet) level. With regard to those references, our interpretation of delays here is for fluid-like models (continuous flows), not jobs. Additionally, our methodology focuses on estimating the throughput that can be offered to a low-priority service for a given tolerated delay, while the references focus on the performance for high-priority jobs: in [17] the objective is to minimize out-of-time probabilities, and [3] analyzes the waiting times and queue lengths (two notions we do not have in this paper for nondeferrable traffic given our "session" modeling for non-deferrable flows).
Large deviations are also applied, again in the power grid context, in [14] to control the risks of delaying some part of

ISBN 978-3-901882-83-8, Networking 2016 © 2016 IFIP

the energy demand from specific devices (pool pumps) with specific constraints (e.g., at least one cycle per device per 24 hours). In this paper we consider a steady-state setting of demand (e.g., during the peak hour) instead of relying on partially predictable (daily) cycles in demand, and we concentrate on providing some deferrable service at a constant *perceived* rate, the delay being a consequence of variations in the primary use of the network.

Our approach is close to *stochastic network calculus* [6], [7]; the main differences being twofold. First, most stochastic network calculus models consider random arrival flows served by a (non-random) network node, while here the service provided is the capacity left unused by nondeferrable traffic, hence randomness on the *service* side. Second, while the goal in network calculus is to provide conservative bounds (e.g., on usable capacity for delay to be below a threshold with high probability), we intend here to estimate the actual value, and for that we treat the large-deviation results (giving the speed for deviation probabilities) as "direct" estimates. Extensive simulations highlight the accuracy of this method. In stochastic network calculus, the closest notion to what we are investigating is that of *leftover capacity*, studied in [2], where the focus is still on obtaining bounds rather than on approaching the actual value. The contribution of this paper is then a method to estimate the usable capacity for given quality constraints given the characteristics of the nondeferrable traffic using it, and its extension in a very simple manner to the network case: it is indeed sufficient to apply the single-link method independently on each link of a network.

Our work is also related to the literature on delay-tolerant networks [8], [19], but the paradigm is sensibly different. Indeed, delay-tolerant networks are generally studied in a wireless context, the changes in connectivity coming from node mobility, hence a focus on routing [20] and buffering [12] strategies. In contrast, here the topology is assumed fixed and the instantaneous "connectivity" (the available capacity) results from demand variations over time of the non-deferrable service, which can be studied with a specific stochastic model. Studying that stochastic model to infer delay guarantees for the deferrable traffic is the main focus of this paper.

For any tolerable delay T, our method provides an estimate for the available capacities on a global network, obtained from a per-link analysis. The outcome of the analysis is a possibly simple exploitation of those unused resources in the near future, through the coordination possibilities offered by the SDN paradigm. Numerical examples show that even for networks optimized for delay-sensitive traffic, capacity utilization can be raised to 95% by adding deferrable traffic, while in current practice it is limited to at most 75-80%, and quite often in the vicinity of 50% due to the time-of-day and day-of-year traffic variations as well as inherent traffic burstiness and the provision of backup paths to be used in the event of failures. Hence, we think our proposition has the potential to enable new types of services without incurring any cost for additional capacity.

The remainder of the paper is organized as follows. The general model considered in the paper is presented in Section II, while Section III treats the special case of one communication link. A simple network case is detailed in Section IV, highlighting the key difficulties in the extension to more complex topologies, in particular insisting on the necessity of caching deferrable traffic in intermediate nodes. Section V summarizes the implications of our results for the "deferrable service network" that can be defined on an existing network, by explaining how to estimate the capacity of each link of this overlay network to satisfy delay constraints while remaining transparent to the non-deferrable traffic. Conclusions and directions for future work are given in Section VI.

II. General model

We consider a peak period during which the non-deferrable traffic is assumed in steady-state. We focus on links that carry the traffic of many users, such as backbone links (as schematized in Figure 1) and possibly backhaul links. Those links are now facing congestion issues because of the demand increase but also because of the increase in last-mile capacities. We nevertheless assume that access capacities are still the bottleneck for users most of the time, i.e., the network is designed so that users use all of their access capacities when active. We use the term *sessions* to refer to user flows, assumed with a constant throughput equal to their access link capacity. Thus, we consider that the network is dimensioned to offer a throughput limited only by the access rate, with a high probability. In that sense, we neglect sessions (flows) that are too short to reach the access transmission rate. A way to include those "mice" in our model is to average, for each given link, their throughputs and to subtract them from the link capacity. This corresponds to assuming that those sessions are such that their aggregate rate is approximately constant at the scale of the acceptable delays for deferrable services.

We also assume that all users have the same access rate, denoted by b. Therefore, when a number X of users have their sessions use a given link (considering fixed routing per flow), the used capacity is simply Xb. If the link has capacity $\underline{C}$, there consequently remains some bandwidth $\underline{C} - Xb$ that can be used for our new (deferrable) service. In the rest of the paper, b will be taken as the capacity unit.

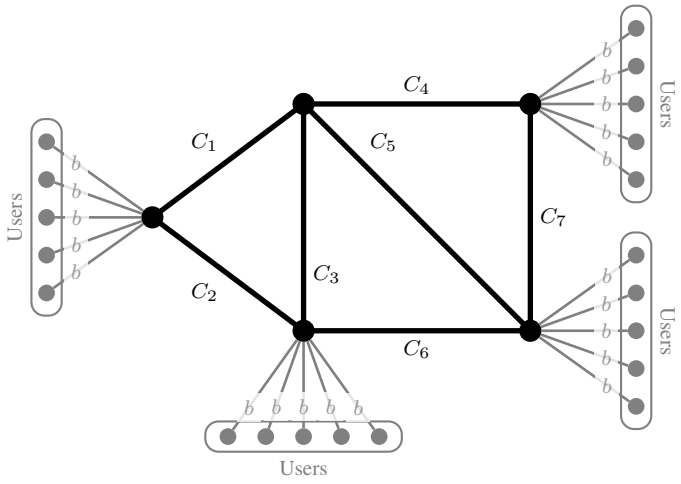

Figure 1. The type of backbone network considered (in black): individual rates are limited by user access rates. Grey parts schematize the access network, nodes in black can be entry points (for users and/or content providers) or simply intermediate nodes. All links are labeled by their capacity.

ISBN 978-3-901882-83-8, Networking 2016 © 2016 IFIP

We consider large numbers of users connected to each entry point, that behave independently. As a consequence, we assume that user sessions arrive according to independent Poisson processes. We moreover model session durations as exponentially distributed random variables with a common average duration denoted by $1/\mu$. Finally, we assume session routing is fixed, as least statistically: for each session route, arrivals follow a Poisson process. In practice, routes may adapt to network conditions, but since we consider networks that are dimensioned to keep saturation rare, we ignore that effect.

The question we now ask regards the use of the remaining capacity for deferrable service: given backbone link capacities, session arrival rates and average duration, we intend to offer a service based on that capacity, with looser delay constraints. More specifically, we want to choose a deadline T and offer a service for which delay is guaranteed to be below T with some high probability. In this paper we show how to compute the amount of such deferrable traffic that can be carried by the network, as a function of T, of the network capacities, and of the non-deferrable traffic characteristics.

We are aware that some of the assumptions we make are a considerable simplification of reality, but we believe the model we build on them provides useful insights regarding the potential offered by resources temporarily left idle by non-deferrable traffic.

III. The case of one link

In this section, we consider the case of a single (backbone) link, and detail the reasoning that will be applied in later sections to more complex topologies.

A. Setting and mathematical formulation

We denote the request arrival rate–assumed constant over the considered period–by λ (arrivals per time unit). Each session uses the same bandwidth b due to last-mile capacity limits, and goes through a single backbone link with capacity $\underline{C}$. As stated in Section II, the service duration of each request is assumed to follow an exponential distribution with parameter μ, so that if we assume that requests arriving while the link is full are rejected, the process describing the evolution of the number of active requests over time is an M/M/C/C queue [10], with $C := \lfloor \underline{C}/b \rfloor$.

Then the blocking probability for a non-deferrable request is simply given by the Erlang-B formula $B(\rho, C) = \frac{\rho^C/C!}{\sum_{k=0}^{C} \rho^k/k!}$ with $\rho := \lambda/\mu$. This formula can be used either to dimension the link (decide the value of C) for a given demand level ρ, or to decide how many users to route through this link (decide the value of λ).

In practice, the requests arriving while the link is fully used may not be rejected but rather be re-routed, or have to share the link capacity with existing sessions (although we can imagine an admission control scheme actually rejecting those requests). But we assume the decisions (on C or λ) are such that this occurs with small probability, so that the M/M/C/C model would still be a good approximation.

We consider providing deferrable service at an effective throughput represented by D, the equivalent number of access links with capacity of b bit/s each. For example, $D = 3$ corresponds to a effective throughput of $3b$ bit/s. The question is: depending on T and on the target probability of delay remaining below T, what value of D can the network handle? Equivalently, for given D and T, what is the probability that the amount DT of deferrable traffic is carried before the deadline T? The network controller will limit the amount of deferrable traffic to a value for which that probability is acceptably large, say 99%.

To address the question, let us consider a deferrable bit that enters the network at time t. If the network provides a first-come-first-served service for deferrable demand, that bit will be served after all the deferrable bits that arrived in the time interval $[t-T, t)$, since the value of D is chosen so that the delay does not exceed T. The probability that our considered bit can be served before $t+T$ at least equals the probability that the capacity left unused by the non-deferrable traffic during $[t, t+T]$ exceeds DT, i.e.,

$$\mathbb{P}\left(\int_t^{t+T} (C - X_\tau)\mathrm{d}\tau \geq DT\right) = 1 - \mathbb{P}\left(\frac{1}{T}\int_t^{t+T} X_\tau \mathrm{d}\tau > C - D\right), \quad (1)$$

where X_τ is the number of active users of the non-deferrable service at time τ.

The situation is illustrated in Figure 2 for a given realization of non-deferrable traffic: the network can offer an equivalent throughput D to a delay-T deferrable traffic if the average idle capacity over a duration T exceeds D with a sufficiently high probability.

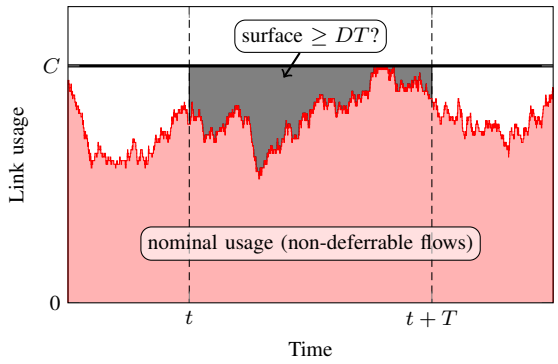

Figure 2. A trajectory for nominal (non-deferrable) usage, and the corresponding instantaneous available capacity (C=65, λ =50, μ=1, T=5).

In the following, we will therefore look for the relation between $A > 0$, T, and the "failure" probability

$$P_{A,T} := \mathbb{P}\left(\frac{1}{T}\int_0^T X_\tau \mathrm{d}\tau > A\right) \quad (2)$$

where (X_τ) is a continuous-time Markov chain corresponding to the number of clients in an M/M/C/C queue with offered load ρ, and X_0 is assumed to be distributed according to the stationary distribution of X, i.e., $\mathbb{P}(X_0 = x_0) = \frac{\rho^{x_0}/x_0!}{\sum_{k=0}^{C} \rho^k/k!}$ for $x_0 = 0, 1, \ldots, C$. We call $P_{A,T}$ the failure probability, since for $A = C - D$ it gives the probability that the average available capacity for deferrable traffic over T is below D, as indicated in (1).

B. Large deviations analysis

For delay durations that are large (e.g., with respect to the mean session duration $1/\mu$), the probability $P_{A,T}$ in (2) can be studied using large deviations [16], [18], and should then verify

$$P_{A,T} = e^{-TI(A)+o(T)} \tag{3}$$

where

$$I(A) := \sup_{\theta\in\mathbb{R}} \left[\theta A - \Lambda_\theta\right], \tag{4}$$

with Λ_θ the principal eigenvalue (eigenvalue with largest real part) of the matrix $Q+\theta V$, V a diagonal matrix with $V(i,i) = i$ for $i = 0, ..., C$ (assuming matrix indices start at 0), and Q the infinitesimal generator matrix for the process X. The function $I(\cdot)$ is called the large deviations *rate function*, and is continuous and convex.

Figure 3 displays examples of the objective function in (4), and of the large deviation rate $I(A)$ when A varies. Note that

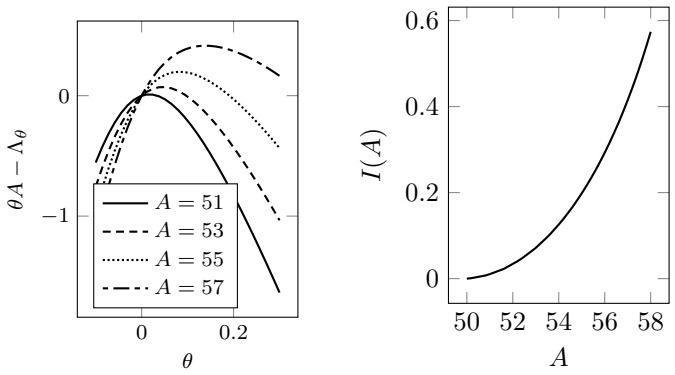

Figure 3. Some values of $\theta A - \Lambda_\theta$ *(left)*, and corresponding large deviation rates *(right)* for the average occupancy during T, when $\lambda = 50$, $\mu = 1$, $C = 65$ (blocking rate for non-deferrable demand: 0.0064).

the large deviation analysis only provides the *rate* $I(A) > 0$ at which the probability tends to 0 as T increases. In this paper we nevertheless intend to ignore the $o(T)$ in (3), or more precisely to ignore its variations with T, and directly use $Ke^{-TI(A)}$ (for an appropriate constant K) as an approximation for the probability of the average occupancy over a period T to exceed A. This will allow us to look for combinations of T and A such that $P_{A,T}$ is small enough. We choose the value of the constant K such that the formula gives a correct response when T tends to 0, hence we will consider that

$$P_{A,T} \approx P_{A,0}e^{-TI(A)}, \tag{5}$$

with $P_{A,0}$ approximating the probability that the instantaneous bandwidth used by non-deferrable traffic exceeds A. For our session model this probability is simply

$$\frac{1}{\sum_{k=0}^{C}\rho^k/k!}\sum_{i=\lceil A\rceil}^{C}\frac{\rho^i}{i!}, \tag{6}$$

which is not continuous in A. For later convenience we will preferably use for $P_{A,0}$ an approximation that is *continuous* in A, for example by taking $P_{A,0}$ as in (6) for integer values of A and piecewise linear between (our choice for the curves plotted in this paper), keeping the difference very small.

The approximation (5) gives us a relationship between T and A: the minimum T such that we can offer some capacity $C - A$ to deferrable service with "failure" probability below ϵ would be

$$T \approx \frac{\log P_{A,0} - \log\epsilon}{I(A)}.$$

Inverting that function in A, the difference $C-A$ is the amount of capacity that can be offered for deferrable service with probability $1-\epsilon$ within delay T during the peak hour, which we denote by $D(T,\epsilon)$:

$$D(T,\epsilon) \approx C - \inf\left\{A : \frac{\log(P_{A,0}/\epsilon)}{I(A)} < T\right\}. \tag{7}$$

As expected, that capacity increases with the guaranteed delay: the rate $I(A)$ increases with A (see Figure 3) while $P_{A,0}$ decreases from (6), hence the inf in (7) decreases with T. Moreover, since $I(A)$ and $P_{A,0}$ vary continuously with A, the right-hand side of (7) is continuous in T, as the inf describes the inverse of the continuous and strictly decreasing function $A \mapsto \frac{\log(P_{A,0}/\epsilon)}{I(A)}$.

An example is displayed in Figure 4, together with simulation results to illustrate that the large deviations theory very accurately predicts the throughput that can be offered to deferrable service as a function of the delay T. More evi-

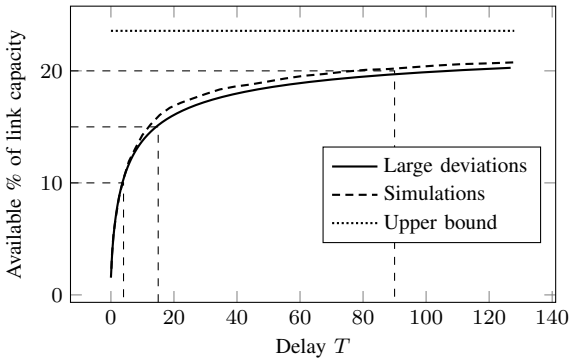

Figure 4. Available capacity for deferrable demand (proportion of the link capacity), with out-of-time probability less than 0.01 (C=65, λ =50, μ=1, blocking probability=0.00645, non-deferrable usage=76%)

dence of this accuracy is given in Figure 5, suggesting that the large deviations approach slightly underestimates the available capacity, with a relative error below 5% (for reasonable link loads) that decreases when the link capacity increases.

Figure 4 shows the case of a link for which the offered non-deferrable traffic (in number of sessions) is 50, but that is dimensioned to $C = 65$ to keep a blocking rate below 0.8%. This results in only 76% of the link capacity being used on average, hence some margin (up to 24% of the link capacity) to offer deferrable service. Both simulation and large-deviation results indicate that we could use 15% of the link capacity–thus reaching 91% link utilization–by proposing a service with delay below $15/\mu$ and a 99% guarantee.

More stringent delay constraints could be preferred: with the same 99% guarantee but for the delay $4/\mu$ we can use 10% of the total link capacity, thus reaching a 86% usage for that link. Alternatively, for very large delays (around $90/\mu$) the link usage rate can get as high as 96%.

ISBN 978-3-901882-83-8, Networking 2016 © 2016 IFIP

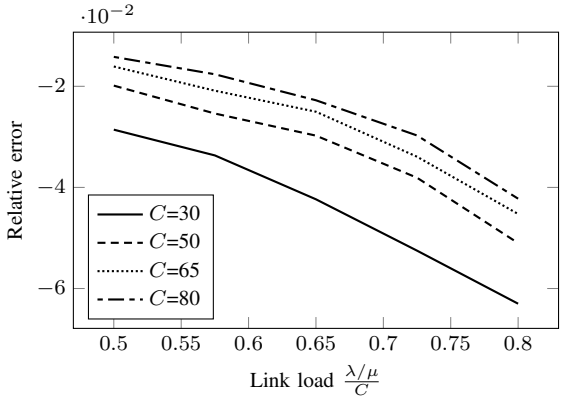

Figure 5. Relative error of (7) to predict deferrable supply with respect to simulations for delay $T = 10$, with out-of-time probability less than 0.01, μ=1.

C. Offering different delay guarantees

Before extending our results to the network case, let us briefly evoke the possibility of proposing simultaneously different "deferrable traffic" offers, with distinct delays and likely with different prices. This provides the network manager with even more flexibility, to segment demand and reach a higher social welfare (and/or higher revenues).

A way to provide that service in practice is to use priorities, traffic with tighter delay constraints having higher priority (and of course, non-deferrable traffic having the highest priority). For the example of Figure 4, as much as 10% of the link capacity can be sold for a "$4/\mu$-delay" service. If that amount is sold, then the network manager can still devote an additional 5% of the link capacity to a "$15/\mu$-delay" service, and even another additional 5% to a "$90/\mu$-delay" service.

This option, and in particular the revenue-maximization possibilities it offers, are not developed in this paper: in the following sections we still consider a unique delay T. But the same simple reasoning as done here is applicable to our next results as well.

IV. A simple network case

In this section we consider the simplest generalization of our results, to a 2-link network topology. We explain how the large-deviation results obtained for one link can be applied for multiple-link transfers, insisting on the importance of caching data in intermediate nodes.

A. Model

Let us consider the simple network topology depicted in Figure 6, with three nodes, two links, and three types of flow: we denote by $X_i(t)$ the number of ongoing non-deferrable sessions using link i only for $i = 1, 2$ at time t, and by $X(t)$ the number of ongoing non-deferrable sessions using both links. Mirroring the previous section, we denote by λ, λ_1 and λ_2 the arrival rates for sessions using both links, link 1 only, and link 2 only, respectively. We assume as before that all sessions use the same bandwidth b, so that the capacity C_i of link i can be expressed as the maximum number of sessions that can simultaneously use that link. Recall that we assume sessions for the three types of flows have the same duration distribution (namely, exponential with parameter μ).

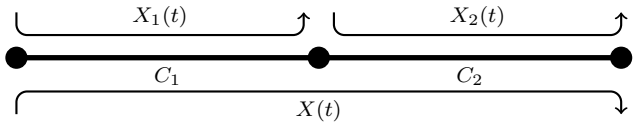

Figure 6. A simple network topology with three types of non-deferrable sessions (represented by arrows). Arc are labelled with their capacity.

B. Available capacity on one link

Under our assumptions, the number of sessions using a given link i is not exactly an M/M/C_i/C_i queue since some requests of two-link connections can be blocked because of the other link. Hence treating $X(t) + X_i(t)$ as an M/M/C_i/C_i queue will be over-pessimistic, but will yield a lower bound of what can be offered as deferrable traffic on that link. Additionally, we can expect this lower bound to be close to the actual value when the blocking probability of non-deferrable flows is low, which is the case in properly dimensioned systems.

Therefore we will take, as an estimate of the available capacity on each link, the result obtained from the analysis in Section III, taking for the arrival rate the sum of the arrival rates of all paths using that link (hence, for our 2-link example, taking $\bar{\lambda}_i = \lambda + \lambda_i$ as the arrival rate on link i, $i = 1, 2$). Figure 7 provides an illustration, where for each link we observe results similar to the one-link case: the large deviation approach provides a very accurate estimation of the available capacity on each link. The gap is a bit larger than in Figure 4, though, especially for link 1, because of the blocking of some two-link sessions due to saturation of link 2 (link 2 has a larger blocking rate than link 1), an aspect neglected in our large deviation approach as explained above.

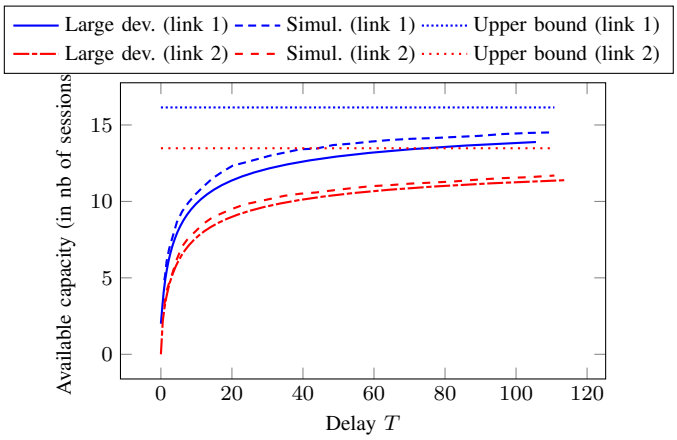

Figure 7. Available capacity for deferrable demand, with out-of-time probability less than 0.01 ($C_1 = 58, C_2 = 60$, $\lambda = 27$, $\lambda_1 = 15$, $\lambda_2 = 20$, μ=1, blocking probabilities$\approx$(0.0035,0.01), non-deferrable usage$\approx$ $(72\%, 78\%)$).

C. Available capacity on a path

Now consider the "long" path in Figure 6. To quantify the amount of bandwidth that could be offered on that path for

deferrable service, we distinguish two cases, according to the possibility or not of caching (storing) data in the middle node.

1) Without caching at the middle node: When no data caching is possible at the middle node, deferrable traffic on the path should be controlled by the source, to send data only when there is capacity available on the whole path, i.e., at an instantaneous rate equal to the minimum of the available rates on the traversed links as proposed in [15]. We leverage here the fact that an SDN architecture can be aware of the usage of each link, and use that knowledge to control the sending rate of each deferrable traffic source. Hence for our two-link path, we are looking for the maximum capacity D such that, assuming the system in stationary regime at time 0,

$$\mathbb{P}\left(\frac{1}{T}\int_{t=0}^{T}\min(C_1-X_1-X\ ,\ C_2-X_2-X)\mathrm{d}t<D\right)\leq\epsilon, \quad (8)$$

where we omit the dependence on t of X_1, X_2, and X: at time t, $X_i(t)$ is the number of non-deferrable sessions using link i only, and $X(t)$ the number of non-deferrable sessions on the two-link path. The left-hand side of (8) being continuous in D, we actually have equality in (8) for the optimal D, that we denote by D_{path}.

We can now state a result lower-bounding D_{path} to the value obtained when no non-deferrable traffic uses the two-link path.

Proposition 1: The available capacity D_{path} on the two-link path is lower-bounded by the one obtained when only one-link sessions arrive, with arrival rate $\bar{\lambda}_i = \lambda_i + \lambda$ on link $i = 1, 2$.

The proof is provided in Appendix A.

Proposition 1 is illustrated by simulations in Figure 8, where we plot the available transmission rates on the two-link path when arrival rates of non-deferrable sessions on link-1, link-2, and the two-link paths are respectively $\lambda_1 + \beta\lambda$, $\lambda_2 + \beta\lambda$, and $(1-\beta)\lambda$, for β varying in $[0, 1]$. As stated in the proposition,

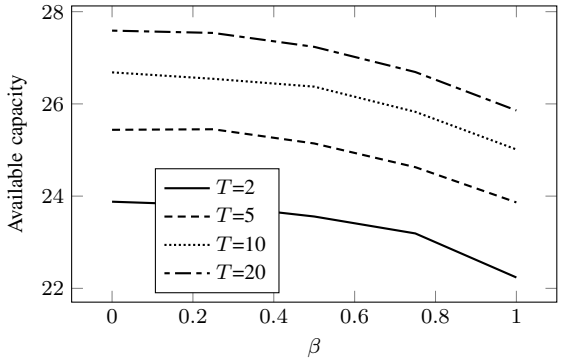

Figure 8. Available capacity for deferrable demand, with out-of-time probability less than 0.01 when $\lambda_1 = 10 + 10\beta$, $\lambda_2 = 15 + 10\beta$, $\lambda = 10(1-\beta)$, $\mu = 1$, C_1=40, C_2=45 (simulation results).

the available rates are the lowest when $\beta = 1$.

Unfortunately, we do not have a large-deviation derivation for that case for a general delay T. However we think that a networked version of the deferrable service *should* involve caching in intermediate nodes to reach a significant use of network links. As an illustration, consider a chain of M links behaving as independent and identical M/M/C/C queues with offered traffic ρ on each link. Then, without caching, the steady-state probability that at least some capacity D is available on the whole path equals $U(D)^M$, with

$$U(D) = \sum_{i=0}^{C-\lceil D\rceil} \frac{\rho^i/i!}{\sum_{k=0}^{C}\rho^k/k!} < 1,$$

and therefore decreases exponentially in M. When $T \to \infty$, the maximum capacity that could be offered on the M-link path equals the average minimum available capacity among the M links, that can be computed as $\sum_{D=1}^{C} U(D)^M$. For $T < \infty$ we can of course offer even less.

Figure 9 plots this upper bound for some example values, showing that the available bandwidth for the deferrable service decreases very fast with M, hence a very limited service offer even for numbers of hops around 5, a reasonable value [9]. Therefore, we think a network application of the deferrable

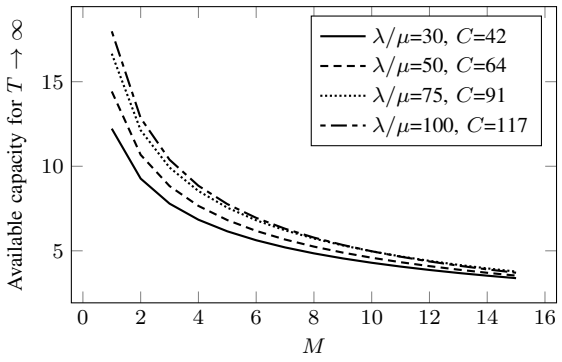

Figure 9. Available capacity on an M-link path (multiple of the session rate b) without caching, when all links behave as independent M/M/C/C queues. Link capacity C is optimized to maintain blocking rate below 0.01.

service is worth considering only when caching is available at intermediate nodes. This is also illustrated later, in Figure 10.

2) With caching at the middle node: With the possibility of caching data in the middle node, the deferrable service does not need to limit instantaneous data rates to the minimum of the instantaneous available rates on the path links: data can be sent on a per-link basis, just being constrained by the currently used link instantaneous available capacity, and is then possibly cached at the next hop. The capacity of interest then becomes the minimum (over links) *average* (over time) available capacity on a period of length T. Mathematically, while without caching we were looking for D_{path} such that

$$\mathbb{P}\left(\frac{1}{T}\int_{t=0}^{T}\min(C_1-X_1-X\ ,\ C_2-X_2-X)\mathrm{d}t<D_{\text{path}}\right)=\epsilon,$$

with caching we are looking for $D^{\text{c}}_{\text{path}}$ such that

$$\mathbb{P}\left(\frac{1}{T}\min\left\{\int_{t=0}^{T}C_1-X_1-X\mathrm{d}t,\ \int_{t=0}^{T}C_2-X_2-X\mathrm{d}t\right\}<D^{\text{c}}_{\text{path}}\right)=\epsilon, \quad (9)$$

which will give larger available capacities, i.e., $D^{\text{c}}_{\text{path}} \geq D_{\text{path}}$, since the minimum of averages is larger than the average of minimums.

More specifically, we claim that with caching, the available capacity on the two-link path is very close to the minimum of

the available capacities along the path, computed separately with the common delay target T.

The reasoning is as follows: the principle of large deviations not only gives the rate at which the probability of "exceptionally large average occupancy" on each link decreases with the considered duration T, but also indicates *how* such large occupancies can be attained. Specifically, only the most likely behaviors leading to such large average occupancies should be considered.

In an M/M/C/C queue, one can show that due to the convexity of the rate function, the most likely trajectories yielding to a given high average occupancy are those with a (almost) constant occupancy, equal to that average. The intuition is that trajectories going below that level must also have periods with even higher occupancy (to reach the same average value), which have a high "likelihood cost" since the likelihood of having an extra client (i.e., an arrival rather than a departure) decreases with the occupancy.

Going back to our two-link path, the most likely way to have "exceptionally bad" performance on the path is to have only one link with "exceptionally large average occupancy", more specifically, the one for which such occupancy is the most likely. But when the target probability of those exceptional events is ϵ, this is precisely the link i with the smallest available capacity D_i computed from (7) for link i. Then, the most likely behavior for the other link is to have more than D_i available.

This reasoning leads to the simple method below.

Method 1: To estimate the available capacity on the two-link path with caching, take the minimum available capacity of both links, computed independently from (7), with a session arrival rate $\bar{\lambda}_i = \lambda + \lambda_i$ on link $i = 1, 2$.

Figure 10 displays an example for a symmetric ($C_1 = C_2$ and $\lambda_1 = \lambda_2$) and "pessimistic" case ($\lambda = 0$), showing that our large-deviation results applied separately to each link still capture very accurately the variations of what can be offered end-to-end with the tolerable delay. Figure 10 also shows the importance of caching: without caching, Method 1 does not apply and the available capacity on the path is significantly below the available capacity on each link. Note that for that case the upper bound as T increases is consistent with the observations in Figure 9 for $M = 2$.

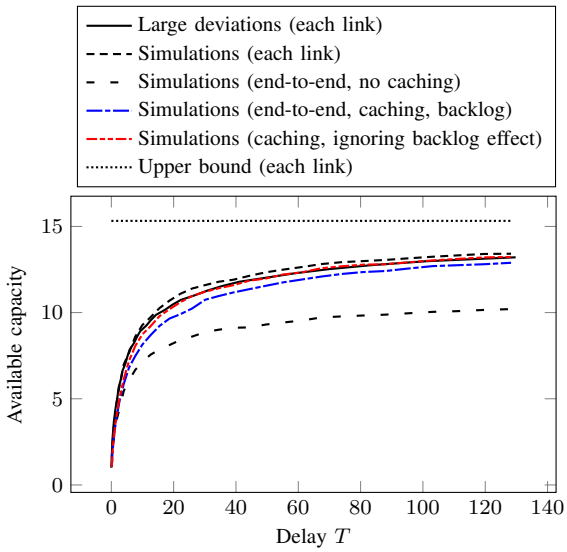

Figure 10. Available capacity for deferrable demand, with out-of-time probability less than 0.01. (Parameters: C_1=C_2=65, λ=0, λ_1=λ_2=50, μ=1, blocking probabilities$\approx$ 0.0064, non-deferrable usage$\approx$ 76%)

A direct consequence of Method 1 is that the use of the available "delay-T" capacity on the links can be on any path, leading to a straightforward method to check feasibility of a deferrable-traffic matrix:

Method 2: To check whether a deferrable traffic throughput profile R_1, R_2, R (on the link-1, the link-2, and the two-link paths respectively) can be served with the delay guarantee T and the out-of-time probability ϵ, verify that the link capacity constraints $R_1 + R \leq D_1$ and $R_2 + R \leq D_2$ are satisfied, with $D_i, i = 1, 2$, obtained as in Method 1.

3) Possible loss of efficiency: The expression (9) actually forgets a part of the problem, by just focusing on the average bandwidth available along the path: there may indeed be cases when some bandwidth is available on link 2 *before* the equivalent amount is available on link 1. In that case, even if link 2 is the bottleneck in the sense of Method 1, not all the capacity of link 2 can be used, hence some possible loss with respect to the proposition due to this "backlog effect", as simulation results show in Figure 10.

However, we think this effect should be minor in practice because of the pipelining that occurs: recall that we have been pessimistic in Section III by considering that no deferrable data received for treatment in the interval $[t-T, t]$ was treated in that interval. This is how we reached (1), and which is simulated in Figure 10. Additionally, Figure 10 considers a worst-case situation, where both links have the same available capacity: this maximizes the likelihood of data being backlogged by link 1 among situations where link 2 is the bottleneck. Even with those two pessimistic assumptions the effect is not so salient, we expect it to be even less important in practice and therefore ignore it in the remainder of this paper.

V. Building a "delay-T network"

In this section, we propose to extend the results of the two previous sections over a whole network. More specifically, we suggest that the manager of an existing network decide on a delay T for the deferrable service, and we provide a methodology to estimate the capacities that could be offered with that delay constraint. We first extend Method 1 to claim that an analysis on a per-link basis is sufficient: the delay guarantee on each link will still be satisfied end-to-end. Hence we can just represent a "delay-T network" as a network with the same topology as the original one, with some "delay-T capacity" on each link.

Note that while our "delay-T network" comes at no costs in terms of transmission capacities, there may be some storage costs at the network nodes to provide caching as described in the previous section. We expect the storage amounts to remain small because of pipelining of data treatment, but quantifying the amount of storage space needed is of interest and should be studied in future work. Here, we assume that storage is cheap and focus on transmission capacities.

A. Lower-bounds: assuming independence among links

In the rest of this section we assume sufficient caching is available within the network. As in Method 1, we target delay guarantees on an end-to-end basis. The reasoning is exactly the same: given a path, considering all traversed links as independent (with arrival rates equal to the sum of the arrival rates of all flows using that link), should leave less capacity than the initial setting. We will use the lower bound obtained this way as an estimate of the available capacities on links.

Then, as in the previous section, we exploit the properties of large deviations as depending only on the most likely trajectories, to claim it is sufficient to consider the minimum available capacity (for the chosen delay T and guarantee level ϵ) among the path links. Indeed, again the most likely way to get bad average performance over a sufficiently long period T is through the "weakest" link in the path, i.e., the most saturated. And for that path, the most likely trajectory leaving a capacity D on average is one leaving a (almost) constant capacity D; for the other links the most likely behavior would not be far away from the average, hence leaving at least D except for very short durations (managed through caching, and only slightly affecting the delay for deferrable service).

Method 3: Assume that there are sufficiently large caching capacities in intermediate nodes in the network, and consider a single path on that network. Then, to estimate the available capacity on any path, take the minimum available capacity of the links on that path, computed independently from (7), with an arrival rate equal to the sum of all arrival rates for sessions using that link.

B. How much capacity to offer?

Treating all links as independent has the advantage of removing complex delay constraints due to multi-link paths: to get some delay-T capacity D over a path, one just needs to ensure to get delay-T capacity on each link over that path, i.e., exactly as in the initial network for non-deferrable traffic. We therefore have the counterpart of Method 2:

Method 4: To check whether a deferrable-traffic throughput profile (on all possible paths on the network) can be served with the delay guarantee T and the out-of-time probability ϵ, verify that the link delay-T capacity constraints on all links are satisfied, where those capacities are obtained independently on each link from (7), taking for the arrival rate the sum of all arrival rates for non-deferrable flows using that link.

A simple way of representing the delay-T service is therefore to keep the network topology, and display the available capacity on each link for the chosen delay T. An example is provided in Figure 11, for two different values of the delay.

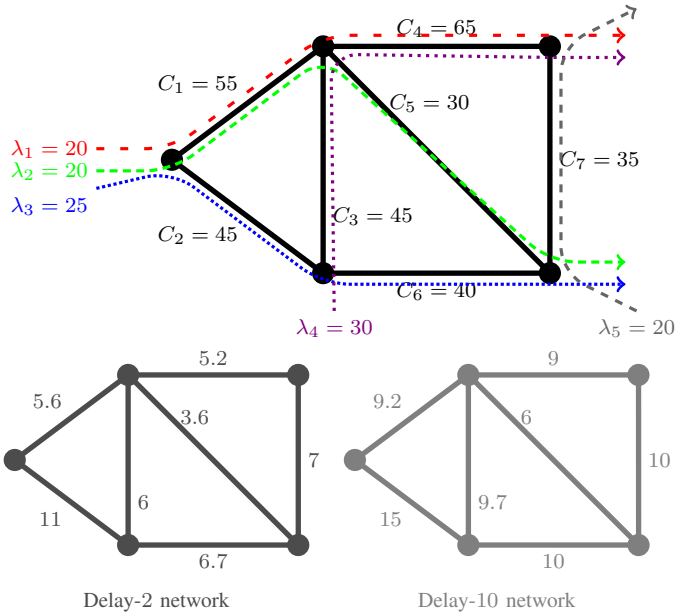

Figure 11. An example of network topology with existing non-deferrable demands *(top)*, and the associated delay-T network for $T = 2$ and $T = 10$, when $\epsilon = 0.01$ and $\mu = 1$. Arcs are labeled with their capacities; all link blocking rates are below 1%.

VI. Conclusion and Perspectives

Telecommunication networks are over-dimensioned with respect to the average traffic they carry, because of traffic demand variations. In this paper, we propose to leverage these extra capacities to provide a new service, using only the resources left available by the non-deferrable traffic. We show that we can still provide guarantees for the delay experienced by such traffic, and provide a methodology based on large deviations analysis to estimate the capacities of the corresponding deferrable-service network, a "new" network that does not imply any capacity expansion costs but possibly some in-network storage costs. Even if a new external service is not offered, the ideas discussed in this paper can be used for internal purposes by a large operator. For example, large operators periodically have to perform some synchronization or backup of large distributed databases, which is very bandwidth-consuming. Although we suspect they already perform those operations using low-priority traffic, our results help understand the type of delays that could be guaranteed, or reciprocally the maximum loads of such low-priority traffic that could be supported while keeping delays reasonable.

Possible future work includes a quantitative study of the amount of in-network storage needed to make the most of such a system: we have assumed that there is sufficient caching space in the network, and would be able to estimate the associated costs to gain even more insight regarding the realizability of our proposition. Another interesting extension is to consider heterogeneous access rates among users: our model (with equal access capacities for all users) provides useful insights, but given the diversity of available access technologies it would be more realistic to study different types of sessions, with different capacities and probably different duration distributions. A first step could be to assume an access rate that depends only on the entry node: in Figure 1 we would have a common b_j for all users accessing the network through entry point j. The case of routing (for non-deferrable traffic) that would depend on the current network conditions is also worth considering: we have ignored it here for simplicity, so that we have Poisson arrivals for each type of route, but in practice we may have spill-over sessions on secondary routes.

APPENDIX A
PROOF OF PROPOSITION 1

Proof: Let us consider the process $(X_1 + X, X_2 + X)$, and consider a slightly different Markov process $(\tilde{X}_1, \tilde{X}_2)$ such that arrivals of one-link sessions are unchanged but arrivals of 2-link sessions are "duplicated" into a one-link session on each link. Mathematically, for arrivals we have transitions
- $(\tilde{X}_1, \tilde{X}_2) \to (\min(C_1, \tilde{X}_1 + 1), X_2)$ with rate λ_1,
- $(\tilde{X}_1, \tilde{X}_2) \to (X_1, \min(C_2, \tilde{X}_2 + 1))$ with rate λ_2,
- $(\tilde{X}_1, \tilde{X}_2) \to (\min(C_1, \tilde{X}_1{+}1), \min(C_2, \tilde{X}_2{+}1))$ with rate λ.

In terms of departures, all sessions of $(\tilde{X}_1, \tilde{X}_2)$ leave after independent exponentially distributed times with parameter μ (i.e., the "duplicated" sessions are then independent).

Then $\min(C_1 - \tilde{X}_1, C_2 - \tilde{X}_2)$ is stochastically smaller than $\min(C_1 - (X_1{+}X), C_2 - (X_2{+}X))$, since the differences are:
i) in the original case more sessions are blocked: when a link is saturated and a 2-link session arrives, the state is unchanged while in the new case there is a new session on one link.
ii) in the original case, two-link sessions leave after an exponentially distributed time with parameter μ, freeing one "server" (the space for one session) simultaneously on both links. In contrast, in the new case the duplicated sessions leave one by one, each one with an exponentially distributed time with parameter μ.
Hence there tends to be more active sessions in the new case than in the original one, thus less space for deferrable flows.

Finally, consider another process $(\bar{X}_1, \bar{X}_2)$, that only differs from $(\tilde{X}_1, \tilde{X}_2)$ in that the "duplicated" sessions now arrive independently (hence we have independent arrivals on each link according to two independent Poisson processes with rate λ for those specific sessions). In summary, $\bar{X}_1$ and $\bar{X}_2$ are simply two *independent* processes, each $\bar{X}_i$ $(i = 1, 2)$ corresponding to an M/M/C_i/C_i queue with arrival rate $\bar{\lambda}_i$ and service rate μ.
Now remark that in both cases, for any fixed $i \in \{1, 2\}$ the "marginal" processes $\bar{X}_i$ and $\tilde{X}_i$ both correspond to an M/M/C_i/C_i with the same arrival rate $\bar{\lambda}_i$ and service rate μ, hence are stochastically equivalent. But because of some joint arrivals (the duplicated ones) in the case of $(\tilde{X}_1, \tilde{X}_2)$, the processes $\tilde{X}_1$ and $\tilde{X}_2$ are positively correlated.
It results that $(C_1 - \bar{X}_1, C_2 - \bar{X}_2)$ and $(C_1 - \tilde{X}_1, C_2 - \tilde{X}_2)$ also have marginal processes that are stochastically equivalent, but $C_1 - \tilde{X}_1$ and $C_2 - \tilde{X}_2$ are positively correlated while $C_1 - \bar{X}_1$ and $C_2 - \bar{X}_2$ are independent.

Let us now define, for $i = 1, 2$, $p_i(\delta) := \mathbb{P}(C_i - \bar{X}_i < \delta)$. Then for any $\delta > 0$:
- because of the independence between $\bar{X}_1$ and $\bar{X}_2$ we have $\mathbb{P}\left(\min(C_1 - \bar{X}_1, C_2 - \bar{X}_2) < \delta\right) = p_1(\delta) + p_2(\delta) - p_1(\delta)p_2(\delta)$;
- now since $\tilde{X}_i$ is stochastically equivalent to $\bar{X}_i$ for $i = 1, 2$, we have at each instant

$$\mathbb{P}\left(\min(C_1 - \tilde{X}_1, C_2 - \bar{X}_2) < \delta\right) = p_1(\delta) + p_2(\delta) - \mathbb{P}\left(\{C_1 - \tilde{X}_1 < \delta\} \cap \{C_2 - \tilde{X}_2 < \delta\}\right).$$

But because of the positive correlation between $C_1 - \tilde{X}_1$ and $C_2 - \tilde{X}_2$, the probability that *both* $C_1 - \tilde{X}_1$ and $C_2 - \tilde{X}_2$ exceed δ is larger than if those processes were independent:

$$\mathbb{P}\left(\{C_1 - \tilde{X}_1 < \delta\} \cap \{C_2 - \tilde{X}_2 < \delta\}\right) \geq p_1(\delta)p_2(\delta).$$

Hence $\min(C_1 - \bar{X}_1, C_2 - \bar{X}_2)$ is stochastically smaller than $\min(C_1 - \tilde{X}_1, C_2 - \tilde{X}_2)$, which yields

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{T}\int_{t=0}^{T} \min(C_1 - \bar{X}_1(t)\ ,\ C_2 - \bar{X}_2(t))\mathrm{d}t < D_{\text{path}}\right) \\
&\geq \mathbb{P}\left(\frac{1}{T}\int_{t=0}^{T} \min(C_1 - \tilde{X}_1(t)\ ,\ C_2 - \tilde{X}_2(t))\mathrm{d}t < D_{\text{path}}\right) \\
&\geq \mathbb{P}\left(\frac{1}{T}\int_{t=0}^{T} \min(C_1 - X_1 - X, C_2 - X_2 - X)\mathrm{d}t < D_{\text{path}}\right) = \epsilon,
\end{aligned}$$

thus we cannot offer more than D_{path} to the system with arrival rates $\lambda_1 = \bar{\lambda}_1, \lambda_2 = \bar{\lambda}_2, \lambda = 0$. Hence the proposition. ∎

REFERENCES

[1] S. Agarwal, M. Kodialam, and T. V. Lakshman. Traffic engineering in software defined networks. In *Proc. of IEEE INFOCOM*, 2013.

[2] K. Angrishi. An end-to-end stochastic network calculus with effective bandwidth and effective capacity. *Computer Networks*, 57(1):78–84, 2013.

[3] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis. On the large deviations behavior of acyclic networks of G/G/1 queues. *The Annals of Applied Probability*, 8(4):1027–1069, 1998.

[4] E. Bitar and S. Low. Deadline differentiated pricing of deferrable electric power service. In *Proc. of IEEE CDC*, 2012.

[5] E. Bitar and Y. Xu. Deadline differentiated pricing of delay-tolerant demand. http://arxiv.org/abs/1407.1601, 2015.

[6] C.-S. Chang. *Performance Guarantees in Communication Networks*. Springer, 2000.

[7] F. Ciucu, A. Burchard, and J. Liebeherr. Scaling properties of statistical end-to-end bounds in the network calculus. *IEEE/ACM Trans. Networking*, 14(6):2300–2312, 2006.

[8] K. Fall. A delay-tolerant network architecture for challenged internets. In *Proc. of ACM SIGCOMM*, 2003.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proc. of ACM SIGCOMM*, 1999.

[10] Bolch. G., S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Wiley, 2006.

[11] H. Kim and N. Feamster. Improving network management with software defined networking. *IEEE Comm. Mag.*, 51(2):114–119, 2013.

[12] A. Krifa, C. Barakat, and T. Spyropoulos. Optimal buffer management policies for delay tolerant networks. In *Proc. of IEEE SECON*, 2008.

[13] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner. Openflow: Enabling innovation in campus networks. *ACM SIGCOMM Comp. Comm. Rev.*, 38(2):69–74, 2008.

[14] S. Meyn, P. Barooah, A. Bušić, and J. Ehren. Ancillary service to the grid from deferrable loads: the case for intelligent pool pumps in Florida. In *Proc. of IEEE CDC*, 2013.

[15] J. Perry, A. Ousterhout, H. Balakrishnan, D. Shah, and H. Fugal. Fastpass: A centralized "zero-queue" datacenter network. In *Proc. of ACM SIGCOMM*, 2014.

[16] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis*. Chapman & Hall, 1995.

[17] A. L. Stolyar and K. Ramanan. Largest weighted delay first scheduling: Large deviations and optimality. *The Annals of Applied Probability*, 11(1):1–48, 2001.

[18] S. R. S. Varadhan. Large deviations. *The Annals of Probability*, 36(2):397–419, 2008.

[19] A. V. Vasilakos, Y. Zhang, and T. Spyropoulos. *Delay Tolerant Networks: Protocols and Applications*. CRC Press, 2011.

[20] Z. Zhang. Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges. *IEEE Communications Surveys & Tutorials*, 8(1):24–37, 2006.