**Submitted to *Operations Research***
**manuscript (Please, provide the manuscript number!)**



# Exploiting Hidden Convexity for Optimal Flow Control in Queueing Networks

Chaithanya Bandi
Kellogg School of Management, Northwestern University, Evanston, IL 60208, c-bandi@kellogg.northwestern.edu
Gar Goei Loke
Department of Mathematics, National University of Singapore, Singapore 119076, e0012863@u.nus.edu

Optimal flow control in queueing networks is a challenging problem occurring in many contexts, such as data centers, cloud computing, healthcare, revenue management, and distributed networks, etc. The traditional approach has been to adopt heuristic solutions or consider infinite-horizon fluid or diffusion approximations. Motivated by emerging techniques in Robust Optimization, we propose a framework, termed Pipeline Queues, which tracks the dynamics of a queue simultaneously in terms of its queue length and waiting time. We begin by showing that the dynamics of a traditional queueing system can be equivalently modeled using this approach. Our key contribution is the uncovering of the hidden convexity resulting from our modeling approach. This leads us to tractable optimization formulations for generic flow control problems of obtaining performance guarantees on average and quantiles of waiting time, under arbitrary arrival and service distributions with non-zero initial conditions. Our model is flexible enough to capture partial observability and uncertainty of the initial state, as well as various constraints on the control policy. We apply our approach to multiple examples from the literature and numerically illustrate their application. Finally, we implemented our model on a real dataset at a major hospital in India. Our proposed policies are near optimal and perform significantly better than present heuristics.

*Key words*: Optimal Control, Queueing networks, Delay constraints, Fluid models, Diffusion limits, Convex Optimization, Robust Optimization

## 1. Introduction

The origin of queueing theory dates back to the beginning of the $20^{\text{th}}$ century, with Erlang's fundamental paper on congestion in telephone traffic. In it, he laid the foundations for queueing theory in terms of the nature of assumptions and techniques of analysis that are being used to this day. In particular, the Poisson process, used by Erlang, played a privileged role in modeling the arrival and service processes of a queue resulting in tractable models of analysis and control.

However, loosening the assumptions to general distributions increases the difficulty of performing a near-exact analysis of the system considerably. In fact, the analysis of the $GI/GI/m$ queue with independent and generally distributed arrivals and services is, by and large, intractable.

This becomes even more challenging if one considers the problem of optimal control of queueing networks. This problem of control is crucial for various application settings that give rise to queueing networks. Examples of optimal control of these stochastic networks involve (1) optimal scheduling of incoming service requests, (2) optimal design of the queueing networks, (3) optimal routing of jobs within a queueing network, and finally (4) capacity pricing and allocation in a queueing network. These problems are common to settings ranging from cloud computing and data centers (Zhang et al. 2010), healthcare (Armony et al. 2015), and transportation (Larson et al. 1993).

### 1.1. Multi-class Queuing Networks and Patient Flow

In these settings, *multiclass queueing networks* are natural. These networks have multiple types of jobs which may differ in their arrival processes, service times, routes through the network, and cost per unit of holding time. A fundamental problem in these systems is that of *flow control* and *sequencing*. In particular, a *sequencing policy* determines at every point in time which type of jobs to serve at each server of the network.

#### Patient Flows in Hospitals

Take the example of patient flows. Patient flow is a central driver of a hospital's operational performance; it is tightly coupled with the overall quality and cost of healthcare (Niska et al. 2010, Armony et al. 2015). The control of patient flow is, thus, a major factor in improving hospital operations.

Patient dynamics are captured by a multiclass queueing system, with multiple servers (physicians), multiclasses of triage patients and multiclasses of in-patients. Patients within each class are served on a (FCFS) basis. A schematic of such a hosptial network is shown in Figure 1.
Triage patients arrive at the system exogenously and must be examined by a physician within a time deadline from their arrivals. After completing their first service, triage patients join the queue of in-patients, or exit the system. In-patients originate from either triage patients or from previous in-patient phases, wherein they require further treatment. While waiting, the in-patients incur queueing costs.

The control problem is challenging arising from two flow characteristics: deadlines and feedbacks. First, arriving patients must be served within time deadlines that are assigned after triage, based on clinical considerations (Mace and Mayer 2008, Farrohknia et al. 2011). Second, patient flows have a significant feedback component that must be accounted for: in-patients possibly return several times to physicians during their sojourn, before ultimately being released.

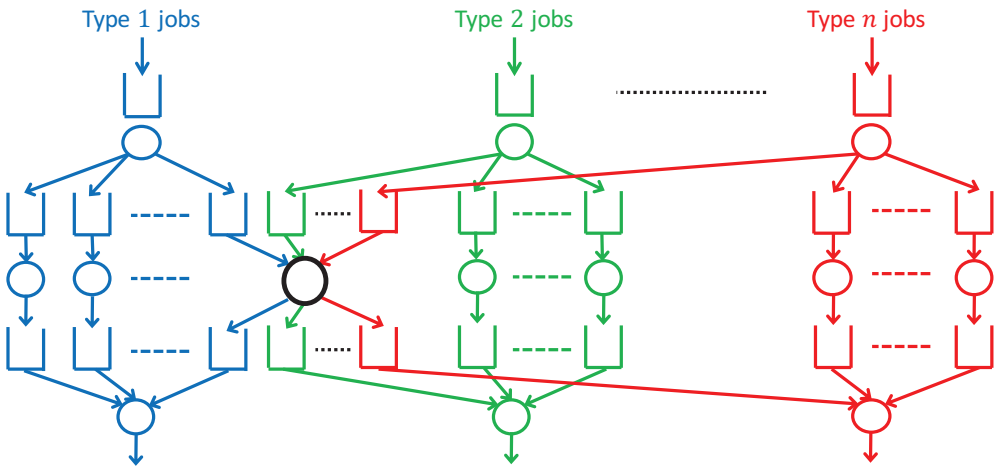



**Figure 1** Visualizing patient flows across different servers in a hospital.

### A Motivating Example of Transient Flow Control at PGIMER

These challenges are noticeably evident in the case of patient flow control at PGIMER Hospital in India. PGIMER is a premier teaching and research hospital. Last year, PGIMER examined around 2.5 million patients in its out-patient departments, with arrivals growing at a rate of 8.5% every year. As a result, on every working day, doctors examine close to 7,500 patients.

The sequence of steps that each patient follows is similar to that in hospitals in the US. Each arriving patient is first triaged to determine condition severity, and then (after some potential waiting) moves to the patient management phase before being discharged. The patient management phase is a series of diagnostic tests, that can depend sequentially on the results of previous tests. Sometimes, they may also be performed in parallel (blood and urine tests can be performed concurrently with a radiology exam, such as on a MRI machine). The discharge decision – admission or whether the patient can return home – cannot be made until all test results are received.

The main difference, however, is in the amount of data being collected and available. Each patient is assigned an Infrared tracker, that allows tracking of the patient's location at different points of time. This real time setting requires us to adaptively control the flow of patients in a *transient manner*, as opposed to the steady state considered in the literature.

## 1.2. Optimal Control Problems

At a broader level, we consider two types of flow routing control problems with **(P1)** *fixed capacity allocation* (sometimes termed *static capacity*); or **(P2)** *dynamic capacity allocation*, where capacity is a decision variable at each time step. These two types of problems are accompanied by one of two possible objectives:

**(a)** Obtaining guarantees on average waiting time $\overline{W}_n$ for each class $n$, in particular, satisfying $\overline{W}_n \leq w_n^{\max}$ for all classes, where $\{w_1^{\max}, \ldots, w_N^{\max}\}$ are given performance requirements; and
**(b)** Minimizing a separable convex function $\sum_{n=1}^N f_n(\overline{W}_n)$ of average waiting times $(\overline{W}_n)_{n=1}^N$ subject to constraints $\overline{W}_n \leq w_n^{\max}$ for all classes.

These optimal control problems **P1**–**P2** in multiclass queueing networks are in general dynamic and state-dependent, as the decision depends on load conditions not only at the server where it is to be made but also at other servers. Naturally, uncertainties regarding the arrival and service processes further complicate the problem. As a result, this problem is both theoretically and computationally hard to solve optimally, even for problems with a few number of servers and job types. It can be formulated as a stochastic dynamic programming problem but that does not lead to tractable approaches for large instances. Thus, a number of researchers have attempted to develop tractable approximations of the optimal policy.

In what follows, we organize our review of the relevant literature around the key approaches to these four control problems. For a more comprehensive review, we refer the readers to Chen and Yao (2013) and Srikant and Ying (2013).

**Brownian Models** The Brownian approach was first introduced by Harrison (1988) and further explored by other researchers (including Williams 1995, Bertsimas et al. 1994, Kumar and Kumar 1994). This approach approximates the queueing network in a heavy-traffic regime, that is, when the workload of the system reaches its capacity limit. In several instances, a policy can be constructed which is optimal in this limiting regime. Brownian models typically make use of the mean and variance of the associated stochastic processes in deriving a simpler control problem. However, except for problems that are essentially one-dimensional, this approach is itself intractable.

**Fluid Models** Fluid models are often tractable, but ignore the variance of the associated stochastic processes. They are deterministic, continuous approximations to stochastic, discrete networks. A major breakthrough was the theory developed by Dai (1995), who showed that the stability of the queueing network is implied by the stability of its associated fluid model (see also Dai 1995, Stolyar 1995). There is also a close connection between the control of processing networks and the optimal control of the corresponding fluid models. There are several examples where the solution of the fluid optimal control problem recovers significant information about the structure of an optimal policy in the original multiclass processing network (see, *e.g.* Avram et al. 1995). Several works have developed methods and guidelines for translating policies derived for the fluid optimal control problem into an implementable control policy for the stochastic, discrete network. Related work includes Meyn (1997), Bäuerle (2000), Maglaras (2000), Bäuerle (2002), Chen et al. (2004), Meyn (2005), Chen et al. (2006). A family of discrete review policies is also proposed by

Maglaras (2000) based on the BIGSTEP approach introduced by Harrison (1996). A comprehensive treatment of these models and policies can be found in Meyn (2008) which also provides specific guidelines on selecting safety stocks and hedging points based on the parameters of the stochastic system.

**Robust Queueing** This nascent literature stream deals with queuing systems under uncertainty in arrival and service times. Xie et al. (2011) use an approach based on the Stochastic Network Calculus framework to propose bounds on the delays in internet networks in transient regime. Bandi et al. (2015) model networks of single-class queues using a robust optimization approach via uncertainty sets and obtain bounds on the waiting times using a worst case analysis approach. These papers are mainly concerned with accurate and tractable performance analysis. Many of the modeling choices made in these papers are motivated by that goal. On the other hand, our goal in this paper is tractable formulations of the control problem.

### 1.3. Our Approach: Pipeline Queues

The above literature identifies the need to navigate between fidelity of the model to true dynamics and stochastic processes, and tractability. Despite extensive work, a tractable and practical approach accommodating all salient features of queueing and flow control problems continues to elude. While fluid models are tractable, they ignore the inherent uncertainties of the problem. The more natural approach to handle uncertainty via stochastic optimization, however, often leads to problems that are intractable to solve (Birge and Louveaux 1997). The alternative of using robust optimization, which can lead to tractable formulations (Bertsimas et al. 2011), however can produce conservative policies.

Instead, we propose an alternative framework to model queueing systems that leverages the tractability of convex optimization while remaining faithful to the probabilistic dynamics in a queueing network. We call this framework ***Pipeline Queues*** or ***'P-Queues'*** for short (Our model is motivated from Jaillet et al. 2018, where the talent *pipeline* of employees in a firm is considered. For their model, the authors view the system from the point-of-view of the amount of time each employee spent in the system. Hence, we use the term 'pipeline' to represent our model).

This framework views the dynamics of a queue by considering the present delay in the queue as a state variable. We show that it is able to equivalently model the dynamics of a traditional queuing model. In particular, our formulation is able to leverage *flow* variables and afford the flexibility of dealing with various priority rules (such as the FCFS discipline) that can be modeled as constraints on these flow variables. While such approaches may not be completely novel in the literature, our key result is the uncovering of the hidden convexity in our formulation, which we leverage to

present tractable optimization formulations for a large class of optimal flow control problems in a general multi-class queueing setting, while exhibiting enhanced computational performance. We are able to put our methodology into practice in a generalizable model of the PGIMER hospital.

### 1.4. Contributions and Structure of Paper

We next summarize our main contributions and discuss the structure of the paper. We propose a model, called a Pipeline Queue, or P-Queue, for short, and a reformulation technique that allows us to obtain tractable formulations for a large class of flow control problems such as problems **P1** and **P2**. More specifically, our contributions are:

1. ***Tractability*:** We present quasi-convex reformulations of the flow control optimization problems **P1** and **P2**. In particular, our formulations can be reduced to bisection search problems on an optimal parameter $k$, where each sub-problem is a quasi-convex feasibility problem (Theorem 2). Its complexity grows by a polynomial function of time horizon $T$, instead of exponentially (Theorem 3).

2. ***Modeling Power*:** Our approach allows for general control policies which can be time or state dependent, while remaining tractable. It is also set in the discrete time setting and can deal with non-zero initial conditions. The dynamics defined is shown to be one-to-one with the traditional queueing setting (Theorem 1). Using current delay $s$ as a dimension, we can model a broad library of random processes. This enables our model to address a large suite of queueing problems today (we illustrate this on Dai and Shi 2018, Dupuis and Ramanan 2000, and the case of PGIMER in §5). Simple adaptions can be made to ensure that the model continues to work even if the delay is not observed (§4.3).

3. ***Performance*:** In practice, the model yields sharper performance than its theoretical guarantees (§5.1). Using a real dataset, we verify that optimal policies obtained from our model are near optimal, with the benefit of improved computation times over common heuristics (§5.2).

In summary, our work contributes to the following literature streams. First, it contributes to the broader queuing literature by providing a procedure for optimizing delays in multi-class systems in a transient regime, that is tractable and accurate. Second, our work builds on and extends nascent robust queuing theory in a significant way by capturing multiple customer classes. This additional modeling component enables the incorporation of customer heterogeneity. Because, from a technical perspective, this relies on introducing customer allocation dynamics to servers, existing robust queuing theory tools are of little use. We show how, by capturing these dynamics via a novel convex optimization approach. Third, the work adds to the operations research literature that deals with patient flow control problem.

We introduce the motivation for P-Queues and its setting in §2. We also define its dynamics and prove its equivalence to traditional queues. The optimization model is set up in §3, and we illustrate how it is derived from Problems **P1** and **P2**. We prove its tractability, and illustrate its modelling power (§4). It is then applied to known problems and a numerical illustration appears in §5, along with performance results when tested against real data from PGIMER. We discuss some theoretical points in §6 and conclude in §7.

**Notation:** Given a directed graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, let $v,w\in\mathcal{V}$ be vertices. We use the notation $v\sim w$ to mean that the directed edge $e=(v,w)$ exists in the set of edges $\mathcal{E}$. Similarly, we use the notation $e\sim v$ (or $v\sim e$) to mean that the edge $e$ terminates at (or originates from) vertex $v$.

## 2. The Pipeline Queue Framework: Models and Formulations

In this section, we motivate our modeling framework by considering the case of a simple queueing system. Consider a discrete-time queuing system with infinite buffers, where customer arrivals and departures occur at discrete time epochs $t\in\{0,1,2,\dots\}$. At each epoch $t$, a total number of $D_t$ customers depart from the queue first, and then a total number of $A_t$ customers arrive. If there are enough servers, we admit all waiting customers (if any) and new arrivals into service; otherwise, we admit as many customers as possible, following the FCFS queueing discipline, until all servers are occupied and hold the remaining customers in the buffer. For the arrival process, we assume that $\{A_t, t=0,1\dots\}$ forms a sequence of independent but possibly time-nonhomogeneous random variables. For the departure process, we assume that each customer in service at the beginning of epoch $t$ (excluding new arrivals) has a probability $\mu\in(0,1)$ of departing in epoch $t$. It is equivalent to assuming that the "service time" in this discrete queue follows a *geometric* distribution with the success parameter being $\mu$, where $\mu$ can *arbitrarily* depend on the current state of system thus modeling arbitrary service time distributions; see Dai and Shi (2017) for a more rigorous proof on this equivalence using a coupling argument.

Our approach focuses on the customer count process and tracks $X_t$ which denotes the total number of customers in system at the beginning of epoch $t$, including both the customers in service and those waiting in the buffer. Under our arrival and departure assumptions, the customer count process $X=\{X_t : t=0,1,\dots\}$ is characterized by the following relationship:

$$X_{t+1}=X_t+A_t-D_t,\quad t=0,1,\dots. \tag{1}$$

Here, the total number of departures $D_t$ follows a binomial distribution with parameters $(Z_t,\mu)$, where $Z_t$ is the number of busy servers at the beginning of epoch $t$.

### Motivation of the Discrete Queue

The discrete queue is motivated by studying hospital inpatient flows (see Shi et al. 2016, Armony et al. 2015). The inpatient beds are modeled as the servers, and patients who need to be admitted to an inpatient bed are modeled as customers, for example, patients who have received treatment in the emergency department (ED) and wait to be hospitalized – commonly known as the ED boarding patients. The customer count $X_t$ corresponds to the the number of patients who are occupying an inpatient bed or waiting to be admitted at the start of day $t$. Naturally, $A_t$ and $D_t$ correspond to the total number of patient arrivals and discharges within day $t$, respectively.

Besides the applications in the healthcare setting, discrete queueing systems have been motivated from a variety of applications in the fields of telecommunication and computer systems, in which the time is usually divided into fixed-length time slots. For example, multi-server discrete queues with geometric service times are studied in the context of circuit-switched multiple-access communications channel (Bruneel and Wuyts 1994, Hluchyj and Karol 1988), data centers (Lin et al. 2013, Takine et al. 1994), and cloud computing (Calheiros et al. 2011, Buyya et al. 2009). Thus, the analysis and results we gain in this paper can potentially benefit a larger community.

## 2.1. Our Model

The dynamics in Equation (1) does not capture any control decisions. In this section, we present our *optimization-based model* that models the dynamics while naturally including controls. The key idea in our approach is to model, what we call, *present delay* $s$ as a dimension and track a new state variable $x^{t,s}$, representing the number of jobs in each node experiencing present delay $s$ at each epoch $t$.

### Model Primitives

Our model requires specifying four aspects – the structure of the queueing network, the variables associated with each feature of the network, the dynamics of the flows and decisions in this network, and the constraints to be applied on servers and flows.

Definition 1. A *P-Queue* is the tuple $(\mathcal{G},\mathcal{S},\mathcal{C})$, containing:

a) A connected directed graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ where the set of vertices $\mathcal{V}$ contains exactly one vertex $I$ with in-degree 0 and exactly one vertex $O$ with out-degree 0.

b) A subset of vertices $\mathcal{S}\subset\mathcal{V}$ with in-degree 1, adjacent to vertices not in $\mathcal{S}$, $w\sim v$ and $v\sim\mathcal{S}$ implies $w\notin\mathcal{S}$. We call all vertices in $\mathcal{S}$ *servers*, and all vertices in $\mathcal{Q}:=\mathcal{V}-\mathcal{S}-\{I,O\}$ *queues*.

c) A partitioning of edges, called the *control structure*, $\mathcal{E}=\mathcal{D}\cup\mathcal{P}\cup\mathcal{F}$, into decision edges $\mathcal{D}$, push edges $\mathcal{P}$ and flow edges $\mathcal{F}$, such that:

i. Pushes $\mathcal{P}$ do not contribute to the in-degree of queues $\mathcal{Q}$ nor out-degree of servers $\mathcal{S}$

ii. Decisions $\mathcal{D}$ do not contribute to the out-degree of queues $\mathcal{Q}$ nor in-degree of servers $\mathcal{S}$

iii. Flows $\mathcal{F}$ contribute at most 1 to in-degrees and out-degrees

Moreover, we say that a P-Queue is *proper* if the graph $\mathcal{G}|_{\mathcal{Q}}$ restricted to the queues has no edges.

REMARK 1. Notice that by condition (b), the graph $\mathcal{G}|_{\mathcal{S}}$ restricted to the servers also has no edges. In other words, if the P-Queue is proper, then $\mathcal{G}|_{\mathcal{S}\cup\mathcal{Q}}$ is a bipartite graph.

At first glance, this definition is common of queue networks, and it is so, for conditions (a) and (b). The key point of difference is that P-Queues deliberately differentiate servers from queues, in order to specify different flows and controls in (c).

Decision edges $\mathcal{D}$ enables the decision-maker to remove or re-route jobs from a server into a queue. Push edges $\mathcal{P}$ allow movement of jobs from queues into servers as capacity in the server is freed up. If a queue feeds into multiple servers, this becomes a routing decision. Flow edges $\mathcal{F}$ allows the movement of jobs that is exogenous to the decision-maker, and hence assumed to be stochastic and uncertain.

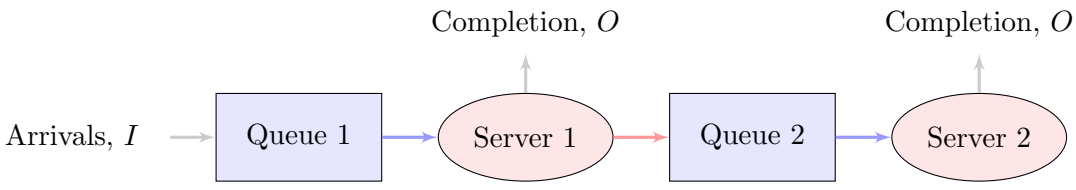
A proper network, where there are no queues feeding into queues, is required for tractability. As such, hereon, we only consider proper networks and omit the term 'proper'.

DEFINITION 2. A P-Queue is *simple* if there are no flows contributing to the out-degree of queues.

A simple P-Queue is one where there are no drop-outs from queues and queues only push units into servers. That said, it can be shown that every network with drop-outs can be re-formulated as a simple network (by adding an auxiliary server to represent the process of these drop-outs). As such, hereon, we shall also be concerned with only simple networks.

Figure 2 gives an example of a simple P-Queue representing a two-stage queue-server complex. Here, we have coloured the queues blue and the servers red. We have also coloured the push edges blue and the decision edges red, while leaving the flow edges grey. In the literature, finding an optimal policy here is often intractable. Instead, our model presents a solution for this problem.

**Figure 2** Simple example of a P-Queue



### Modeling Arrivals, Services and State of a Queue

We next describe how we model external arrivals, services at each server, and flows between the queueing nodes. The key idea of our approach is to model the state of all queues in the network using the variable: $x_v^{t,s}$, which represents the number of jobs at each node $v \in \mathcal{S} \cup \mathcal{Q}$, that have experienced delay $s$ at modeling time $t$.

1. Arrivals (Inflows): On flow edges $e \in \mathcal{F}$, let $\mathcal{I} := \{e \in \mathcal{F} : I \sim e\}$ and let $\Lambda_v^t$ be the real-valued random variable describing the stochastic arrival of jobs into vertex $v$ (where $e \sim v$ and $e \in \mathcal{I}$), with moment generating function $g_v^t$. This can model arbitrary arrivals to a queueing system.
2. Services (Outflows): For all other flow edges in $\mathcal{F} - \mathcal{I}$, define $f_v^{t,s}$ to represent the probability that a job in the vertex $v$ will leave along the edge at time $t$, after experiencing delay $s$. For convenience, let $d_v^{t,s} = 1 - f_v^{t,s}$ denote the probability of remaining in the vertex. Note that this allows us to model any service distribution using its hazard rate. For example, for exponential services at node $v$, $f_v^{t,s} = \mu_v$ for all $t, s$, where $\mu_v$ is the service rate at node $v$.
3. Controls on edges $e \in \mathcal{D} \cup \mathcal{P}$:
   - Push variables on $\mathcal{P}$: Let $p_v^{t,s}$ represent the number of jobs that are pushed into server $v$ from some queue $w : w \sim v$ at time $t$, after experiencing delay $s$ in queue $w$. (Note: There can only be at most one push edge incident to a server, so this is well-defined.)
   - Decision variables on $\mathcal{D}$: Let $\alpha_{v,w}^{t,s}$ represent the proportion of jobs in server $v$ that we route to queue $w : v \sim w$ at time $t$, after experiencing delay $s$ in server $v$.

### 2.2. Dynamics in P-Queues

The following assumptions are natural and applicable in many contexts.

Assumption 1 **(Independence)**. *Uncertainty on the flow of every job $i$ in the P-Queue is independent of any other job $j$ in the P-Queue.*

Assumption 2 **(Arrivals)**. *The support of all arrivals $\Lambda_v^t$ is bounded above by some constant $\Lambda$ for all $v : v \sim e, e \in \mathcal{I}$ and $t$.*

Assumption 1 legitimizes the use of Binomial distributions on the flows. Notice that this does not involve the arrivals. The arrival of jobs is allowed to be in bursts. Instead, arrivals are controlled via Assumption 2, which guarantees the finiteness of its moment generating function $g_v^t$. This is required later in order for our model to be well-defined.

The dynamics on each vertex depends on the kinds of edges incident to it.

1. If $v \in \mathcal{S}$ is a server, then its out-degree can possess both flows and decisions

$$x_v^{t,s} = \text{Bin}\left(x_v^{t-1,s-1}\left(1 - \sum_{v \sim w} \alpha_{v,w}^{t-1,s-1}\right), d_v^{t,s}\right), \qquad \forall t, \forall s \geq 1 \tag{2}$$

This simplifies if either are absent. At the in-edge, it can only maximally have one push edge, wherein $x_v^{t,0} = \sum_s p_v^{t,s}$, or one flow edge, wherein $x_v^{t,0} = \Lambda_v^t$, contributing to its in-degree.

2. If $v \in \mathcal{Q}$ is a queue, then only pushes are present in its out-degree

$$x_v^{t,s} = x_v^{t-1,s-1} - \sum_{v\sim w} p_w^{t,s} = \begin{cases} x_v^{t-s,0} - \sum_{v\sim w}\sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} & \text{if } 1 \le s < t \\ x_v^{0,s-t} - \sum_{v\sim w}\sum_{\tau=0}^{t-1} p_w^{t-\tau,s-\tau} & \text{if } s \ge t \end{cases} \tag{3}$$

It can also have decisions and at most one flow edge $e$ contributing to its in-degree. Then we can model, if $u \sim e$ and $u$ is a server,

$$x_v^{t,0} = \sum_s \text{Bin}\left( x_u^{t-1,s-1}\Big(1 - \sum_{u\sim y} \alpha_{u,y}^{t-1,s-1}\Big), f_u^{t,s} \right) + \sum_{\substack{w\sim v \\ w\neq u}} \sum_s x_w^{t-1,s-1} \alpha_{w,v}^{t-1,s-1}. \tag{4}$$

Else, if $u = I$, then we just replace the first term with $\Lambda_v^t$.

### 2.3. Equivalence to Standard Queueing Dynamics and Performance Metrics

Because the current delay $s$ at vertex $v$ is tracked, it is possible to compute time-based performance metrics of the P-Queue, despite being unable to track the exact delay for any single job.

**Proposition 1** *Let $v \in \mathcal{S} \cup \mathcal{Q}$. At time $t = T$, the (instantaneous) performance metrics include:*

*i. Queue length or service occupancy, $\sum_s x_v^{T,s}$; and,*

*ii. Total waiting or service time of present jobs in $v$, $\sum_s s x_v^{T,s}$.*

*Proof of Proposition 1.* (i) is evident from the definition of the $x_v^{T,s}$ variables . For (ii), again by definition, we have jobs amounting to $x_v^{T,s}$ have delay $s$ in vertex $v$. Hence, the total waiting time of these jobs so far is precisely $\sum_s s x_v^{T,s}$. □

THEOREM 1. *The total waiting and service times in a P-Queue equate to those for traditional queues. Specifically, let $\mathcal{A}$ be the set of jobs $i$ that have arrived since $t = 1$, $\mathcal{N}$ be the set of jobs $i$ that remain in the queue at time $t = T$ and $\mathcal{M} := \mathcal{A} \cap \mathcal{N}$ be the subset of jobs $i$ arriving after time $t = 0$ such that $i$ still remains in the queue at time $t = T$. Then,*

*i.*

$$\sum_{s=1}^{T-1} s x_v^{T,s} = \sum_{i\in\mathcal{M}} \tau_i, \tag{5}$$

*ii.*

$$\sum_s s x_v^{T,s} = \sum_{i\in\mathcal{N}} \tau_i, \tag{6}$$

*iii.*

$$\sum_{t=1}^{T}\sum_{s=0}^{t-1} x_v^{t,s} = \sum_{i\in\mathcal{A}} \tau_i, \tag{7}$$

*where, $\tau_i$ is the truncated waiting time of job $i$ up to time $T$.*

*Proof of Theorem 1.* For brevity, we suppress the index $v$. Let $\mathcal{A}^t$ be the collection of jobs $i$ arriving at time $t$ (defined for $t<0$ as well). Hence $\mathcal{A}=\bigcup_{t=1}^{T}\mathcal{A}^t$ and $|\mathcal{A}^t|=x^{t,0}$.

Define $\mathcal{M}^{t,s}:=\{i:\tau_i=s\}\cap\mathcal{A}^t$, allowing $t<0$. Note that for $s<T-t$, $x^{t+s,s}-x^{t+s+1,s+1}=|\mathcal{M}^{t,s}|:=m^{t,s}$ are precisely the jobs which arrived at time $t$ and left at time $t+s$. Also, $m^{t,T-t}:=|\mathcal{M}^{t,T-t}|=x^{T,T-t}$. Then, $\mathcal{M}=\bigcup_{t=1}^{T}\mathcal{M}^{t,T-t}$ and $\mathcal{N}=\bigcup_{t\leq T}\mathcal{M}^{t,T-t}$.

But every job in $\mathcal{M}^{t,s}$ has waiting time $s$! Thus, the right-hand side of (5) evaluates as

$$\sum_{i\in\mathcal{M}}\tau_i=\sum_{t=1}^{T}\sum_{i\in\mathcal{M}^{t,T-t}}\tau_i=\sum_{t=1}^{T}(T-t)m^{t,T-t}=\sum_{t=1}^{T}(T-t)x^{T,T-t}=\sum_{s=1}^{T-1}sx^{T,s}$$

Letting the limits of $s$ be unbounded recovers (6). Similarly, the right-hand side of (7) evaluates as

$$\begin{aligned}&\sum_{i\in\mathcal{A}}\tau_i=\sum_{t=1}^{T}\sum_{i\in\mathcal{A}^t}\tau_i=\sum_{t=1}^{T}\sum_{s=0}^{T-t}\sum_{i\in\mathcal{M}^{t,s}}\tau_i=\sum_{t=1}^{T}\sum_{s=0}^{T-t}sm^{t,s}\\&=\sum_{t=1}^{T}\Big((T-t)x^{T,T-t}+\sum_{s=0}^{T-t-1}s(x^{t+s,s}-x^{t+s+1,s+1})\Big)=\sum_{t=1}^{T}\sum_{s=0}^{T-t}x^{t+s,s}=\sum_{t=1}^{T}\sum_{s=0}^{t-1}x^{t,s}\end{aligned}$$

This completes the proof. □

Results (i) and (ii) effectively guarantee that the performance metrics in traditional queueing problems always have one-to-one correspondences to P-Queues. Result (iii) is reminiscent of the opening line in the proof of Little's Law. By Proposition 1, $\sum_{s=0}^{t-1}x_v^{t,s}$ is the queue length counting only jobs which arrived after time $t=0$. Hence, (7) is the statement that the sum of queue lengths after time $t=0$ is equal to the total waiting times of new arrivals. This result can be viewed as an adaptation Little's Law in the discrete setting where arrivals are general and the time horizon $T$ is finite (resulting in the truncation of waiting times at $T$).

### Modeling System Constraints

Based on this Theorem 1, we can model various system constraints as follows:

1. **Capacity Constaints**: Servers $v\in\mathcal{S}$ face capacity constraints, *i.e.* there is some capacity $\kappa_v$, which should never be exceeded, $\sum_s x_v^{t,s}\leq\kappa_v,\forall t$. Capacity constraints can be applied on individual servers or on groups of servers (*e.g.* for multi-class problems).
2. **Delay Constaints**: On queues $v\in\mathcal{Q}$, queueing constraints that can be written as linear combination of $x_v^{t,s}$ are permitted. This includes the average wait time constraint $\sum_s sx_v^{t,s}\leq W_v^t\sum_s x_v^{t,s},\forall t$. We will discuss more on the range of permitted queueing constraints later.
3. **Flow Constaints**: We may also want constraints on edges $e\in\mathcal{E}$, such as bounds on flows between edges (for example, a minimum number of jobs that must be cleared by the system in totality by some time $t$). Again, as long as these are linear combinations of the state, they are permitted.

***Summary:*** P-Queues are indeed very general and correspond one-to-one with traditional queues. While allowing us to model all the queueing dynamics, P-Queues model also enables calculating various performance metrics. We next demonstrate how this view of a queueing system leads to tractable formulations for solving the series of control problems **P1** and **P2** by choosing the decisions $\alpha_{v,w}^{t,s}$ and $p_v^{t,s}$, subject to queue performance guarantees.

## 3. General P-Queue Control Problem and Tractable Formulations

We next present our main results which show that our framework leads to tractable formulations for Problems **P1** and **P2** discussed in §1.2. We begin by considering Problem **P1a** to illustrate our approach.

### Illustration: Reformulation of Problem P1a

In Problem **P1a**, we seek an optimal flow control policy under capacity constraints to obtain guarantees on average waiting time $\overline{W}_n$ for each class $n$, while satisfying $\overline{W}_n \leq w_n^{\max}$ for all classes with high probability $(1-\epsilon)$, where $\{w_1^{\max}, \ldots, w_N^{\max}\}$ are given performance requirements. In particular, suppose the total capacity available is $K$ which needs to be distributed among nodes $v \in \mathcal{S}$. Suppose we assign capacity $\kappa_v$ to node $v$, then the following model the corresponding capacity constraints:

$$\sum_s x_v^{t,s} \leq \kappa_v \quad \forall v \in \mathcal{S}, \tag{8}$$

$$\sum_{v \in \mathcal{S}} \kappa_v \leq K. \tag{9}$$

The first constraint (8) represents the capacity constraint on the servers. Note that $x_v^{t,s}$ is a random variable; it is a mixture of the arrivals, service times and our decisions. Therefore, we will enforce this capacity constraint to be attained in probability. In particular, we relax the constraint to $\mathbb{P}\Big[\sum_s x_v^{t,s} \leq \kappa_v\Big] > 1-\epsilon$ for some small $\epsilon$.

Furthermore, in terms of the $x_v^{t,s}$ variables, mean waiting time is given by $\overline{W}_n = \sum_s s x_v^{t,s} / \sum_s x_v^{t,s}$, which allows us to reformulate the high probability waiting time constraint $\overline{W}_n \leq w_n^{\max}$ as

$$\mathbb{P}\left[\sum_s s x_v^{t,s} \leq w_n^{\max} \cdot \sum_s x_v^{t,s}\right] > 1-\epsilon_v^t. \tag{10}$$

Such reformulations are, in fact, possible for many traditional constraints on queue performance. In particular, they can be described in the following chance-constrained form:

$$\text{For a queue vertex } v \in \mathcal{Q},\ \mathbb{P}\left[\sum_s x_v^{t,s} a_v^s \leq b_v^t\right] > 1-\epsilon_v^t. \tag{11}$$

We have seen that this applies to average waiting time constraint as in Eq. (10). Similarly, a CVaR guarantee on queue wait time can be written as $\sum_{s \geq S} s x_v^{t,s} / \sum_{s \geq S} x_v^{t,s}$, which can be represented as (11).

To summarize, we can choose $\boldsymbol{a}$'s appropriately (they could be negative) to model wide variety of measures. This representation has two main advantages:

**(1)** Imposing bounds on some generic function $f(W_t)$ of service/wait times $W_t$, is possible through a discretization of $f(\cdot)$.This yields precisely the coefficients $a_v^s$, without additional conditions on $f$ – it need not be convex, or increasing even, as is often assumed in the literature.

**(2)** Enforcing particular queue dynamics is possible through such a formulation. For example, imposing $\sum_s s^2 x_v^{t,s} \leq P_v^t \sum_s x_v^{t,s}$ with high probability on queues will force a FCFS behaviour, for the quadratic term penalizes accumulating units with a long delay. A *first-come-last-serve* behaviour can be similarly achieved by reversing the inequalities.

### 3.1. Moving From Problems P1 and P2 to Pipeline Formulation

Obtaining tractable forms for (11) is often difficult. Instead, we consider the following convex outer approximation for some $k>0, \theta_v^t>0$:

$$k \log \mathbb{E}\left[\exp\left(\frac{\sum_s x_v^{t,s} a_v^s - b_v^t}{k\theta_v^t}\right)\right] \leq 0 \tag{12}$$

We next present one of our key results which relates the guarantees obtained from (12) to (11).

**Proposition 2** *Under (12), for any $\phi>0$,*

$$\mathbb{P}\left[\sum_s x_v^{t,s} a_v^s - b_v^t > \phi\right] < \exp\left(-\frac{\phi}{k\theta_v^t}\right). \tag{13}$$

*Proof of Proposition 2.* The result is a simple consequence of Markov's inequality. □

In Proposition 2, $\theta_v^t$ prescribes the severity of the constraint – the smaller the value of $\theta_v^t$, the more stringent the bound would be. This is referred to as the Satisficing constraint, which has appeared in this form involving the exponential disutility in the context of portfolio management (Brown and Sim 2009), vehicle routing (Jaillet et al. 2016), manpower planning (Jaillet et al. 2018) and with a different utility function in healthcare (Qi 2017).

The rationale for (12) is the following. For a given $k$, consider the feasibility problem:

$$\begin{array}{lll} \min & 0 & \\ \text{s.t.} & k\log\left(\mathbb{E}\exp\left(\sum_s x_v^{t,s} - \kappa_v/k\theta\right)\right) \leq 0 & \forall v \in \mathcal{S}, \forall t \\ & k\log\left(\mathbb{E}\exp\left(\sum_s x_v^{t,s} a_{v,j}^s - b_{v,j}^t/k\theta_{v,j}^t\right)\right) \leq 0 & \forall v \in \mathcal{Q}, \forall t, \forall j \end{array} \tag{14}$$

The index $j$ here labels as many queueing constraints as we may desire to impose on queue $v$.

By Proposition 2, any solution to (14) will be a solution to the following feasibility problem:

$$\begin{array}{lll} \min & 0 & \\ \text{s.t.} & \mathbb{P}\left[\sum_s x_v^{t,s} - \kappa_v > \phi\right] < \exp\left(-\frac{\phi}{k\theta}\right) & \forall v \in \mathcal{S}, \forall t, \forall \phi > 0 \\ & \mathbb{P}\left[\sum_s x_v^{t,s} a_{v,j}^s - b_{v,j}^t > \phi\right] < \exp\left(-\frac{\phi}{k\theta_{v,j}^t}\right) & \forall v \in \mathcal{Q}, \forall t, \forall j, \forall \phi > 0 \end{array} \tag{15}$$

As argued before, this is an approximation of Problem **P1a** where the approximation is controlled by the free parameter $k$. In particular, if $a_{v,j}^s$ and $b_{v,j}^t$ are integral, then

$$\begin{aligned} \mathbb{P}\left[\sum_s x_v^{t,s} a_{v,j}^s > b_{v,j}^t\right] &= \mathbb{P}\left[\sum_s x_v^{t,s} a_{v,j}^s - b_{v,j}^t \geq 1\right] = \lim_{\delta \searrow 0} \mathbb{P}\left[\sum_s x_v^{t,s} a_{v,j}^s - b_{v,j}^t > 1 - \delta\right] \\ &\leq \lim_{\delta \searrow 0} \exp\left(-\frac{1-\delta}{k\theta_{v,j}^t}\right) = \exp\left(-\frac{1}{k\theta_{v,j}^t}\right) \end{aligned}$$

If $\exp\left(-1/k\theta_{v,j}^t\right) \leq \epsilon_{v,j}^t$, then any solution to (15) will certainly be a solution to Problem **P1a**. Hence, we have proven the following specialized result in the case of integral coefficients:

**Proposition 3** *Suppose $a_{v,j}^s$ and $b_{v,j}^t$ are integers for all queues $v$, constraints $j$, delays $s$ and times $t$. If $k$ chosen such that $\exp\left(-1/k\theta_{v,j}^t\right) \leq \epsilon_{v,j}^t, \forall v, \forall j, \forall t$, then any solution to (14) is feasible for Problem **P1a**.*

To get the best approximation, we seek to minimize $\exp\left(-1/k\theta_{v,j}^t\right)$ using the free parameter $k$, which is referred to as the *risk parameter* in the literature. Given the monotonicity with respect to $k$, the goal of minimizing $\exp\left(-1/k\theta_{v,j}^t\right)$ is equivalent to minimizing $k$ itself. Even though we don't have these direct bounds in the general non-integer setting, a similar argument works here – minimizing $\exp\left(-\phi/k\theta_{v,j}^t\right)$ for all $\phi > 0$, which is equivalent to minimizing $k$, would yield good candidate solutions for Problem **P1a**.

Indeed, this leads to the following optimization problem, which we call the P-Queue Model. Suppose the model is ran over time horizon $t \in \{1, \ldots, T\}$ and delay space $s \in \{0, \ldots, M\}$, where it is assumed that $T < M$.

$$\begin{array}{lll} \min & k & \\ \text{s.t.} & k \log\left(\mathbb{E}\exp\left(\sum_s x_v^{t,s} - \kappa_v / k\theta\right)\right) \leq 0 & \forall v \in \mathcal{S}, \forall t \\ & k \log\left(\mathbb{E}\exp\left(\sum_s x_v^{t,s} a_{v,j}^s - b_{v,j}^t / k\theta_{v,j}^t\right)\right) \leq 0 & \forall v \in \mathcal{Q}, \forall t, \forall j \\ & k \log\left(\mathbb{E}\exp\left(\sum_{\substack{w \sim e \sim v \\ e \in \mathcal{P}}} p_v^{t,s} - q_w^{t-1,s-1} / k\theta\right)\right) \leq 0 & \forall w \in \mathcal{Q}, \forall t, \forall s \end{array} \tag{16}$$

We have added the last constraint to represent the condition that we cannot push more than there are jobs in the queue, $\sum_{\substack{w\sim e\sim v\\ e\in\mathcal{P}}} p_v^{t,s} \le q_w^{t-1,s-1}$. We call this the *push constraint*. We would want to set $\theta \ll \theta_{v,j}^t$, so that the model will treat the capacity and push constraints as tight constraints, that should never be violated under any circumstances.

### 3.2. Quasi-convexity of P-Queues Optimization Problem (16)

This formulation has three strengths. First, it is tractable, which we shall immediately illustrate in the following theorem. Second, it accepts general policies. We describe more on this in §4. Finally, it has stronger performance in practice than these guarantees provided by Proposition 2. We shall see this in §5.

THEOREM 2. *Let $\mathcal{PQ}=(\mathcal{G},\mathcal{S},\mathcal{C})$ be a P-Queue. Under Assumptions 1 and 2, suppose that*

*i. $\mathcal{PQ}$ is proper and simple, and,*

*ii. The initial state of $\mathcal{PQ}$, given by $x_v^{0,s}$, is known for all $v\in\mathcal{S}\cup\mathcal{Q}$.*

*Then for fixed $k$, the feasibility problem of seeking a solution in the constraint set of the P-Queue Model (16), has a quasi-convex reformulation depending only on the decisions $\alpha_v^{t,s}$, pushes $p_v^{t,s}$ and auxiliary variables. In particular, it can be tractably solved by bisection search on $k$.*

Note that Theorem 2's assumptions are not restrictive. In particular, assumptions 1 and 2 are common in the literature. Condition (i) is unbinding – a proper P-Queue has an equivalent simple P-Queue. The most stringent assumption is perhaps (ii) knowledge of the initial delay, which can be either costly to obtain (*e.g.* in distributed processing networks), or simply not observed. This discussion is deferred to §4.3.

While we have illustrated the reformulation for Problem **P1a**, the other problems are simple adaptations of this. For Problem **P2a**, we can obtain a tractable formulation that yields an upper bound by exploiting Jensen's inequality:

$$f_n(\overline{W}_n) := f_n\left(\frac{\sum_s s x_v^{t,s}}{\sum_s x_v^{t,s}}\right) \le \frac{\sum_s f_n(s) x_v^{t,s}}{\sum_s x_v^{t,s}}. \tag{17}$$

The right-hand objective is of the form of our P-Queue Model and is tractable using Theorem 2.

Problems **P1b** and **P2b** are similar, except where dynamic capacity allocation is allowed. In the P-Queue model, if one changes $\kappa_v$ to $\kappa_v^t$, and allows $\kappa_v^t$ to be a decision variable, then the reformulation remains quasi-convex. This is because $\kappa_v$ only appears as a linear term in the reformulation in Theorem 2 (refer to Lemmas in Appendix A). The tractability of this hierarchy of problems is summarized in Table 1 below.

**Table 1** Tractability of Delay and Capacity Control Problems

| Capacity | a) Obeys delay constraints | b) Minimizes separable convex fn |
|---|---|---|
| **P1**: **Static** | Tractable ✓ | Upper bounded |
| **P2**: **Dynamic** | Tractable ✓ | Upper bounded |

### 3.3. Proof of Theorem 2

As the proof is long and technical, we only broadly describe the key steps and components of the proof without belaboring into the details. The full proof is adjourned to Appendix A.

Our goal is to provide a quasi-convex reformulation on every constraint of the form $\sum_s x_v^{t,s} a_v^s \leq b_v^t$ expressed in the style of (12). We shall do this for two classes of vertices, servers $\mathcal{S}$ and queues $\mathcal{Q}$ separately. Since $\mathcal{PQ}$ is proper and simple, all vertices must be in one of the forms with dynamics as described in §2.2 in equations (2-4). In all subsequent discussions, let the moment generating functions of $\Lambda_v^t$ be denoted by $g_v^t(z) = \mathbb{E}[e^{z\Lambda_v^t}]$. Moreover, the operator $k \log \mathbb{E} \exp(\cdot)$ is additive under Assumption 1 of independence, hence it suffices to consider the contributions of each edge separately. These results are captured in the following series of Lemmas.

1. For server vertices $\mathcal{S}$
   a. Out-degree is 1 corresponding to a flow edge: Lemma 3
   b. Out-degree consists of both flows and decisions (Note that if there are no flow edges, then one can set the probability of remaining in the vertex as 1): Lemma 4
2. For queue vertices $\mathcal{Q}$, the out-degree can only be pushes since $\mathcal{PQ}$ is simple.
   a. In-degree consists of only arrivals: Lemma 5
   b. In-degree consists of only flow edges from servers: Lemma 6
   c. In-degree consists of only decisions: Lemma 7
3. For the push constraint: Lemma 8

The proof for these Lemmas follow a common strategy. First, decisions $\alpha^{t,s}$ are reformulated as a ratio $\beta^{t+1,s+1}/\beta^{t,s}$, with some care taken to ensure conservation of flow between the vertices. Second, the expectation on the Binomial-distributed $x_v^{t,s}$ is evaluated repeatedly in time. Then, we exploit the degree of freedom in the definition of $\beta$ to pick suitable boundary terms $\beta^{0,s}$ and $\beta^{t,0}$ that lead to nice formulations. Finally, we check that for a given $k$, the reformulations are quasi-convex in the decisions $\beta_v^{t,s}$, pushes $p_v^{t,s}$ and all subsequent auxiliary variables. As such, a bisection search algorithm on $k$ for the P-Queue Model (16) would only require solving each of the quasi-convex feasibility problems posed by these reformulations.

To glimpse the inner workings of the Lemmas, we present here Lemma 4 as an illustration.

**Lemma 4**. Let $v \in \mathcal{S}$ be a server. Suppose the out-degree of $v$ consists of a flow edge with $d_v^{t,s} = 1 - f_v^{t,s}$ representing the probability of remaining in the vertex and decisions $\alpha_{v,w}^{t,s}$ representing the proportion of existing units to be redirected to vertex $w$, then the capacity constraint

$$k \log \mathbb{E} \exp\left( \left( \sum_s x_v^{t,s} - \kappa_v \right) / k\theta \right) \leq 0 \tag{18}$$

has the reformulation

$$\begin{aligned}
& A(p_v, \xi_v^{\cdot,1}; k) + k \sum_{s=t}^{\text{last}} \xi_v^{1,s-t+1} \leq \frac{\kappa_v}{\theta} \\
& \beta_v^{t,s} \log(1 - d_v^{t,s} + d_v^{t,s} e^{1/k\theta}) \leq \xi_v^{t,s} \\
& \beta_v^{t-\tau,s-\tau} \log(1 - d_v^{t-\tau,s-\tau} + d_v^{t-\tau,s-\tau} e^{\xi_v^{t-\tau+1,s-\tau+1} / \beta_v^{t-\tau,s-\tau}}) \leq \xi_v^{t-\tau,s-\tau} \qquad \begin{matrix} \tau = 1 \ldots, t-1 \\ s > \tau \end{matrix}
\end{aligned} \tag{19}$$

where

$$A(p_v, \xi_v^{\cdot,1}; k) = \begin{cases} \frac{1}{\theta} \sum_s p_v^{t,s} + k \sum_{s=1}^{t-1} \xi_v^{t-s+1,1} & \text{if } v \text{ has only a push in-edge} \\ k \log g_v^t(1/k\theta) + k \sum_{s=1}^{t-1} \log g_v^t(\xi_v^{t-s+1,1}/k) & \text{if } v \text{ has only an arrival in-edge} \end{cases} \tag{20}$$

*Proof.* Notice that in the dynamics defined for such a server $v \in \mathcal{S}$ in (2), the actual distribution of units out of the server via the individual $\alpha_{v,w}^{t,s}$'s is immaterial in whether or not the capacity is going to be exceeded or not, insofar as their sum is controlled. As such, let us use the reformulation $\frac{\beta_v^{t,s}}{\beta_v^{t-1,s-1}} := 1 - \sum_{v \sim w} \alpha_{v,w}^{t-1,s-1}$, and then later define in Lemma 7, a suitable reformulation that preserves the conservation of flow. Also notice that this definition affords us some degrees of freedom to define $\beta_v^{0,s}$ and $\beta_v^{t,0}$ as suitable.

Under this regime, the dynamics (2) is reformulated as $x_v^{t,s} = \text{Bin}\left( x_v^{t-1,s-1} \frac{\beta_v^{t,s}}{\beta_v^{t-1,s-1}}, d_v^{t,s} \right)$. As such, by additivity and independence, (18) may be reformulated as

$$k \log \mathbb{E} \exp\left( x_v^{t,0} / k\theta \right) + k \sum_{s=1}^{\text{last}} \log \mathbb{E} \exp\left( x_v^{t,s} / k\theta \right) \leq \frac{\kappa_v}{\theta} \tag{21}$$

We now perform a series of repeated expansions of the expectation. The larger principle behind why this expansion works is described as the property of 'Pipeline Invariance' in Jaillet et al. (2018). Now the left-hand side of this expression is given as

$$\begin{aligned}
& k \log \mathbb{E} \exp\left( x_v^{t,0} / k\theta \right) + k \sum_{s=1}^{\text{last}} \log \mathbb{E} \exp\left( x_v^{t-1,s-1} \frac{\beta_v^{t,s}}{\beta_v^{t-1,s-1}} \log(1 - d_v^{t,s} + d_v^{t,s} e^{1/k\theta}) \right) \\
= & k \log \mathbb{E} \exp\left( x_v^{t,0} \frac{1}{k\theta} \right) + k \log \mathbb{E} \exp\left( x_v^{t-1,0} \frac{\xi_v^{t,1}}{\beta_v^{t-1,0}} \right) + k \sum_{s=2}^{\text{last}} \log \mathbb{E} \exp\left( x_v^{t-1,s-1} \frac{\xi_v^{t,s}}{\beta_v^{t-1,s-1}} \right) \\
= & k \log \mathbb{E} \exp\left( x_v^{t,0} \frac{1}{k\theta} \right) + k \log \mathbb{E} \exp\left( x_v^{t-1,0} \frac{\xi_v^{t,1}}{\beta_v^{t-1,0}} \right)
\end{aligned}$$

$$
\begin{aligned}
&+k\sum_{s=2}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-2,s-2}\frac{\beta_v^{t-1,s-1}}{\beta_v^{t-2,s-2}}\log(1-d_v^{t-1,s-1}+d_v^{t-1,s-1}e^{\xi_v^{t,s}/\beta_v^{t-1,s-1}})\right)\\
&=k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right)+\sum_{\tau=1}^{2}k\log\mathbb{E}\exp\left(x_v^{t-\tau,0}\frac{\xi_v^{t-\tau+1,1}}{\beta_v^{t-\tau,0}}\right)+k\sum_{s=3}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-2,s-2}\frac{\xi_v^{t-1,s-1}}{\beta_v^{t-2,s-2}}\right)\\
&=\dots\\
&=k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right)+\sum_{\tau=1}^{l}k\log\mathbb{E}\exp\left(x_v^{t-\tau,0}\frac{\xi_v^{t-\tau+1,1}}{\beta_v^{t-\tau,0}}\right)+k\sum_{s=l+1}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-l,s-l}\frac{\xi_v^{t-l+1,s-l+1}}{\beta_v^{t-l,s-l}}\right)\\
&=\dots\\
&=k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right)+\sum_{\tau=1}^{t}k\log\mathbb{E}\exp\left(x_v^{t-\tau,0}\frac{\xi_v^{t-\tau+1,1}}{\beta_v^{t-\tau,0}}\right)+k\sum_{s=t+1}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{0,s-t}\frac{\xi_v^{1,s-t+1}}{\beta_v^{0,s-t}}\right)\\
&=k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right)+\sum_{s=1}^{t-1}k\log\mathbb{E}\exp\left(x_v^{t-s,0}\frac{\xi_v^{t-s+1,1}}{\beta_v^{t-s,0}}\right)+k\sum_{s=t}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{0,s-t}\frac{\xi_v^{1,s-t+1}}{\beta_v^{0,s-t}}\right) \qquad (22)
\end{aligned}
$$

where in each iterate, we define the auxiliary variables $\xi$ to serve as epigraphs, as follows

$$
\begin{aligned}
&\beta_v^{t,s}\log(1-d_v^{t,s}+d_v^{t,s}e^{1/k\theta})\le\xi_v^{t,s}\\
&\beta_v^{t-\tau,s-\tau}\log(1-d_v^{t-\tau,s-\tau}+d_v^{t-\tau,s-\tau}e^{\xi_v^{t-\tau+1,s-\tau+1}/\beta_v^{t-\tau,s-\tau}})\le\xi_v^{t-\tau,s-\tau} \qquad \begin{matrix}\tau=1\dots,t-1\\ s>\tau\end{matrix} \qquad (23)
\end{aligned}
$$

Because the original constraint is convex, by the convexity of the operator $k\log\mathbb{E}\exp(\cdot)$ and moreover, each of these epigraphs is jointly convex in $\beta$ and $\xi$ because they are just perspectives of a convex function, this expansion is legitimate.

Now, the last term is known (for it contains either the initial data or decision variables). Hence, we utilize our degree of freedom to define $\beta_v^{0,s}:=x_v^{0,s}$ for all $s$, to obtain the simple expression $k\sum_{s=t}^{\text{last}}\xi_v^{1,s-t+1}$ for the last term.

To complete the computation, we evaluate (22). If indeed, the in-edge to server $v$ is a push edge, then $x_v^{t,0}=\sum_s p_v^{t,s}$ and again, the first two terms will be known. Similarly utilizing our degree of freedom, we can let $\beta_v^{t,0}:=x_v^{t,0}=\sum_s p_v^{t,s}$ for all $t$, to get $\frac{1}{\theta}\sum_s p_v^{t,s}$ for the first term and $k\sum_s\xi_v^{t-s+1,1}$ for the second term.

If instead, the in-edge is an arrival, then we continue the evaluation to obtain $k\log g_v^t(1/k\theta)$ for the first term and $k\sum_{s=1}^{t-1}\log g_v^{t-s}(\xi_v^{t-s+1,1}/\beta_v^{t-s,0})$ for the second. In this case, we can pick $\beta_v^{t,0}\equiv k$ for all $t$ to preserve the overall quasi-convexity of the problem. See Remark 4 for alternative choices of boundary terms. □

REMARK 2. i. Theorem 2 requires $\log g_v^t(\cdot)$ convex. But this is guaranteed by Hölder's inequality. Often, $\log g_v^t$ may possess a nicer form (see Remark 4 and Appendix B).

ii. Demarcating decision edges to servers and push edges to queues turns out to be important in the proof. This legitimizes the choice of $\beta_v^{t,0}:=x_v^{t,0}=\sum_s p_v^{t,s}$, because we have required that the inflow into the servers are not random variables, but rather decision variables.

Observe from the proof that the reformulations reduce to a series of perspective cone constraints that are shifts and dilations of the function $h(\cdot) = \log\left(1 - d + d\exp(\cdot)\right)$ for some $d \in [0,1]$. The functions $h(\cdot)$ are not just convex, but asymptotically linear, hence easily tractable even in practice.

**Corollary 1** *If each server has at most 1 queue node contributing to its out-degree, then the P-Queue Model (16) has a convex reformulation. Furthermore, if the graph $\mathcal{G}|_{\mathcal{S}\cup\mathcal{Q}}$ contains no cycles and arrivals form a Poisson distribution, then the reformulation reduces to a linear one.*

*Proof of Corollary 1.* This is a consequence of the proof of Theorem 2. □

Corollary 1 identifies the case where convex, as opposed to quasi-convex, reformulations are obtained. These circumstances are not uncommon – all the examples raised in §5 later save for the case of PGIMER will lead to convex reformulations. Moreover, suppose there is a server $s$ which routes to more than 1 queue $q_i$. If one is happy to accept a similar P-Queue with an artificial queue $\bar{q}$ and servers $\bar{s}_i$, where $s \sim \bar{q}$ and $\bar{q} \sim \bar{s}_i, \forall i$ and $\bar{s}_i \sim q_i, \forall i$, then this satisfies the conditions in Corollary 1 and hence is convex.

### 3.4. Tractability of the P-Queue Model

Tractability covers two aspects. First, specific features, most ostensibly, linearity of constraints within operator $k\log\mathbb{E}\exp(\cdot)$, guarantee quasi-convex reformulations in Theorem 2. This exploits the fact that the uncertainty is Binomial-distributed. Tractability is lost if functions of the uncertainty are present, except in cases involving only the decision variables, *e.g.* in priority queues – the constraint $\left(\sum_s q_i^{t,s}\right)\left(\sum_s p_j^{t,s}\right) \leq 0$ where Queue $i$ has a higher priority than Queue $j$, preserves the tractability of Theorem 2.

Second, tractability depends on the dimensions of the model, in particular, the time horizon $T$ and the delay space $M$. This is given by the following corollary.

THEOREM 3. *For fixed $k$, the P-Queue Model (16) can be reformulated as a collection of no more than $O\left(|\mathcal{S}||\mathcal{Q}|MT^3\right)$ quasi-convex constraints.*

*Proof of Theorem 3.* This is a direct consequence of the proof of Lemmas 3-8. □

Note that, as long as the time horizon $T$ is not large, the model remains tractable, even if the state space, governed by $M$, can be very large. This contrasts against other multi-period formulations where $T$ is in the exponent of the complexity. Moreover, in our model, full generality is assumed in the decisions $\alpha$ and $p$! Additionally, it is also interesting to note that the actual support of $x_v^{t,s}$ has no ramifications on tractability, unlike in the Markov setting.

The problem setting usually dictates the value of $M$, *e.g.* slots within a week in the case of patient scheduling. Tractability is less dependent on $M$, except where $M$ is very large, say heavy-tailed

service times. One approach is to truncate the distribution at the first service time $M$ where the probability of exceeding service time $M$ is below some bound. This grants new perspective on why heavy-tailed distributions lead to intractable problems.

On the other hand, $T$ is the look ahead horizon for which performance guarantees are required. It depends on the flexibility of decisions $\alpha$ and the 'memory' of the dynamics. In practice, the model is churned in a rolling horizon fashion and the dependence on $T$ is thus de-emphasized. Numerically, we have been able to solve for $T \approx 120$, on a 64 GB Linux machine within 2 hrs. Note that for 'light-tailed' queues, where the queuing system has a low relaxation time, the 'memory' of the system is also low leading to small-sized optimization problems.

In Table 2, we present computational times for different values of $T$ and $M$. We generate random instances of our problem with different parameters, and calculate the average computational time.

**Table 2** Average computation times (in minutes)

| Horizon | Maximum Delay, $M$ | | | | | |
|---|---|---|---|---|---|---|
| $T$ | 10 | 20 | 50 | 100 | 200 | 500 |
| 10 | 1.94 | 2.10 | 2.19 | 2.29 | 3.08 | 3.56 |
| 20 | 1.97 | 2.04 | 2.30 | 3.23 | 3.4 | 3.73 |
| 50 | 2.28 | 2.17 | 3.06 | 3.70 | 4.49 | 4.81 |
| 100 | 1.72 | 3.37 | 4.09 | 5.07 | 8.2 | 11.35 |
| 200 | 2.94 | 5.22 | 8.53 | 15.39 | 19.23 | 42.64 |
| 500 | 3.38 | 9.81 | 13.28 | 18.11 | 55.43 | 87.32 |

## 4. Modeling Flexibility of our Framework

In Section 3, we have demonstrated the computational tractability of our approach in modeling various types of queueing conrol problems. In this section, we discuss the other main advantage of our approach – modeling flexibility. In particular, we discuss (a) modeling different types of service distributions, (b) modeling arbitrary control policies, and (c) modeling partial observability or partial information.

### 4.1. Modeling General Service Distributions

Recall that, in our P-Queue model, we represent the service process by describing the dynamics of the flow edges using the $f_v^{t,s}$ variables. In particular, $f_v^{t,s}$ represents the probability that a job in the vertex $v$ will leave at time $t$, after experiencing delay $s$. That is, the $f$ variables are related to the hazard rate of service distribution.

Suppose the distribution at node $v$ at time $t$ is given by PDF $g_v^t(\cdot)$ and CDF $G_v^t(\cdot)$, the hazard rate function and consequently $f_v^{t,s}$ is given by

$$f_v^{t,s} = \frac{g_v^t(s)}{1 - G_v^t(s)}, \quad s > 0, \tag{24}$$

For example, for exponential services at node $v$ at time $t$ with rate $\mu_v^t$, the $f$ variables are given by $f_v^{t,s} = \mu_v^t$ for all $t, s$. Note that, in this manner, we can allow for time varying distributions as well.

### 4.2. Modeling General Control Policies

Our approach can also be used to model different types of control policies, and different types of performance objectives.

Recall that our key control variable $\alpha_{v,w}^{t,s}$ represents the proportion of jobs in server $v$, which have spent $s$ amount of time in server $v$ by modeling time $t$, that we route to queue $w : v \sim w$ at time $t$. By imposing constraints on $\alpha_{v,w}^{t,s}$, we can model the following control policies:

1. Stationary policies, which do not depend on time $t$ can be modeled by imposing the following linear constraint: $\alpha_{v,w}^{t,s} = \alpha_{v,w}^{t',s} \quad \forall t, t'$.
2. State independent policies, which do not depend on state $s$ can be modeled by imposing the following linear constraint: $\alpha_{v,w}^{t,s} = \alpha_{v,w}^{t,s'} \quad \forall s, s'$.
3. Static Priority policies, which do not depend on state $s$ and time $t$, can be modeled by imposing the following linear constraint: $\alpha_{v,w}^{t,s} = \alpha_{v,w}^{t',s'} \quad \forall s, s'$.

Additional constraints on $p_v^{t,s}$ and initial data $x_v^{0,s}$ would be required to make sense in these cases.

### 4.3. Modeling Partial Observability using a Robust P-Queue Model

We identified earlier that full observability, including the present delay of all jobs, may potentially be a restrictive condition. In reality and practice, it is not common for this history to be tracked. What is available is information about the queue length, or present server usage rates. In this case, define $\tilde{x}_v^0 := \sum_s x_v^{0,s}$ as the observed sum of units in vertex $v$ at time $t = 0$. Instead, $x_v^{0,s}$ is assumed to be not observed. Then the Robust P-Queue Model is the following:

$$
\begin{array}{lll}
\min & k & \\
\text{s.t.} & k \log\left(\mathbb{E}\exp\left(\sum_s x_v^{t,s} - \kappa_v / k\theta\right)\right) \leq 0 & \forall v \in \mathcal{S}, \forall t \\
& k \log\left(\mathbb{E}\exp\left(\sum_s x_v^{t,s} a_{v,j}^s - b_{v,j}^t / k\theta_{v,j}^t\right)\right) \leq 0 & \forall v \in \mathcal{Q}, \forall t, \forall j \\
& k \log\left(\mathbb{E}\exp\left(\sum_{\substack{w \sim e \sim v \\ e \in \mathcal{P}}} p_v^{t,s} - q_w^{t-1,s-1} / k\theta\right)\right) \leq 0 & \forall w \in \mathcal{Q}, \forall t, \forall s \\
& \forall x_v^{0,s} \in \left\{ x_v^{0,s} \geq 0 : \sum_s x_v^{0,s} = \tilde{x}_v^0 \right\} & \forall v \in \mathcal{S} \cup \mathcal{Q}
\end{array}
\tag{25}
$$

This model turns out to be tractable. Indeed, a closer inspection of the reformulation of Theorem 2 reveals that the terms $x_v^{0,s}$ only appear in linear terms of the form $\sum_s x_v^{0,s} u_v^s$ for some known coefficients $u_v^s$ that do not involve $k$ (see Lemma 3-8 in Appendix A for more details). Hence, the robust counterpart involves changing $\sum_s x_v^{0,s} u_v^s$ into $\tilde{x}_v^0 \cdot \max_s \{u_v^s\}$ and Theorem 2 remains tractable.

**Corollary 2** *Let $\mathcal{PQ} = (\mathcal{G}, \mathcal{S}, \mathcal{C})$ be a proper and simple P-Queue. Then under Assumptions 1 and 2, for fixed $k$, the feasibility problem of seeking a solution in the constraint set of the Robust P-Queue Model (25), has a quasi-convex reformulation depending only on the decisions $\alpha_v^{t,s}$, pushes $p_v^{t,s}$ and auxiliary variables. In particular, it can be tractably solved by bisection search on $k$.*

Note that this instance illustrates a basic form of partial observability, however, it addresses cases where the policies themselves depend on the queue length. Using our approach, we can use other uncertainty sets describing the extent of observability but is beyond the scope of current discussion.

## 5. Numerical Illustration of Model Performance

In this section, we apply our approach to classic problems in literature and subsequently present results from its implementation in a problem of patient-flow control at a major hospital in India.

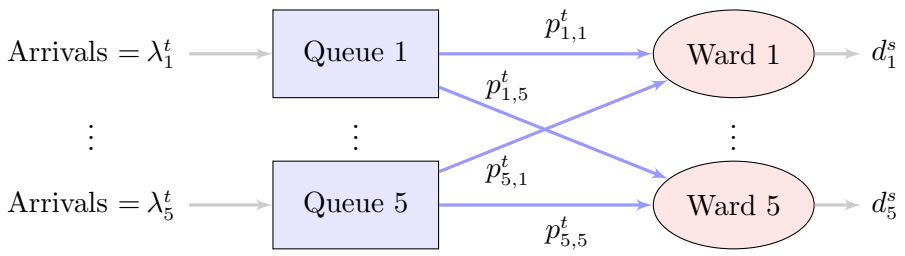
### 5.1. Performance Evaluation in Simulations

We begin by presenting results from computational experiments on some of the networks considered in the literature with different types of features such as presence of multiple classes and feedback.

#### Multi-class Overflow Management

Dai and Shi (2018) describe a model with five classes of patients and wards associated to each class of patients. Until beds stipulated for their class frees up, patients remain in the queue. The planner has the option of warding the patient of class $i$ in another class $j$, though at some cost $c_{i,j}$. There is also cost to holding patients in the queue. This problem can be formulated as a P-Queue, as illustrated in Figure 3, with dynamics in the fashion of §2.2. The costs to holding patients can be formulated as Satisficing queue constraints, while the costs to overflowing patients becomes the linear constraint $\sum_{i,j}\sum_{s} c_{i,j}^s p_{i,j}^{t,s} \leq C$. Notice the flexibility in the definition of these costs – overflow costs can depend on present delay $s$. Moreover, because the only decisions here are push decisions, the P-Queue model (16) simplifies into a convex problem; linear if arrivals are Poisson.

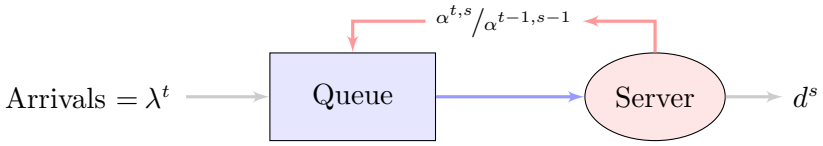
**Figure 3** P-Queue model of Dai and Shi (2018)'s inpatient overflow model



### Feedback Queueing Networks

We consider a simple feedback queueing system (Coffman and Kleinrock 1968, Dupuis and Ramanan 2000), where the manager has the flexibility to force jobs that experienced delays of more than $\bar{S}$ in the server to rejoin the queue. This can be expressed as the P-Queue in Figure 4. Again, this will have the usual dynamics in the P-Queue, except with the condition that $\alpha^{t,s} = \alpha^{t-1,s-1}, \forall s \leq \bar{S}$. As expected, this too leads to a convex reformulation of Theorem 2.

**Figure 4** P-Queue model of the restarting queue




Here, we provide a numerical illustration on the feedback model in Figure 4. The time horizon is set at $T = 10$. The queues and servers admit jobs that remain for at most $s = 19$; at $s = 20$, jobs are assumed to leave the queues/servers. The server has a capacity of $\kappa_i = 500$ and are initiated at 95% capacity as in Figure 5. Queues, too, are initiated from a non-zero state. Arrivals are Poisson-distributed but time non-homogeneous, averaging around 85. The service rate is also given in Figure 5. The performance targets for the queue are set as follows: Average waiting time, $W = 3.75$; CVaR waiting time, $S = 4$ and $W^S = 6$; and FCFS-enforcing constriant, $P = 20$. A bound $\alpha^{t,s}/\alpha^{t-1,s-1} \geq 0.5$ was artificially set to prevent the model from clearing servers.

For wait time constraints, $\theta$ was set at 1, while for the prioritizing constraint $\theta = 5$, and for capacity and push constraints, $\theta = 0.01$. This ensured they were not violated without compromising performance which occurs if $\theta$ was too low. Two techniques are described in Jaillet et al. (2018) to compute the perspective cone constraints – a cutting plane algorithm and a second-order cone approximation. Here, we adopt the former.

The optimal risk level was $k^* = 10.34$. In other words, there is at most a $1/e$ chance that any of the queue constraint at any one time $t$ is violated by a magnitude of $k^*$. Table 3 describes the optimal push policies $p_i^{t,s}$ from Queues $i$ into Servers $i$. The push policies make sense – since the target average wait time was 3.75, the model preferentially pushes jobs with present delay of around 7 to 9. Moreover, the FCFS discipline is apparent. Table 4 also gives the optimal feedback policy $\alpha^{t,s}/\alpha^{t-1,s-1}$ for jobs removed from the Server to rejoin the Queue.

To verify the performance of the optimal policies, we ran 1,000 simulations, where job completion and arrivals were simulated to their assumed distributions. For each simulation, deviations from the queueing constraints and the capacity were computed – here, a positive score indicates constraint

Table 3: Optimal Push Policy from Queue to Server, $p_i^{t,s}$

| **Time, t** | **Experienced delay, s** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 39 | 131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 53 | 30 | 111 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 120 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 73 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 144 | 0 | 0 | 104 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 119 | 150 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 72 | 99 | 0 | 0 | 0 | 0 | 46 | 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 71 | 36 | 0 | 24 | 101 | 31 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 4: Optimal Transfer Policy from Server back into Queue, $\alpha^{t,s}/\alpha^{t-1,s-1}$

| **Time, t** | **Experienced delay, s** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** |
| 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.51 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 0.5 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 0.5 | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 0.5 | 0.94 | 0.95 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.97 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 | 0.5 | 0.83 | 0.76 | 0.56 | 0.5 | 0.5 | 0.5 | 0.55 | 0.75 | 0.87 | 0.94 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 1 | 1 | 1 | 1 |
| 8 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 1 | 1 | 1 |
| 9 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 1 | 1 |
| 10 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 1 |

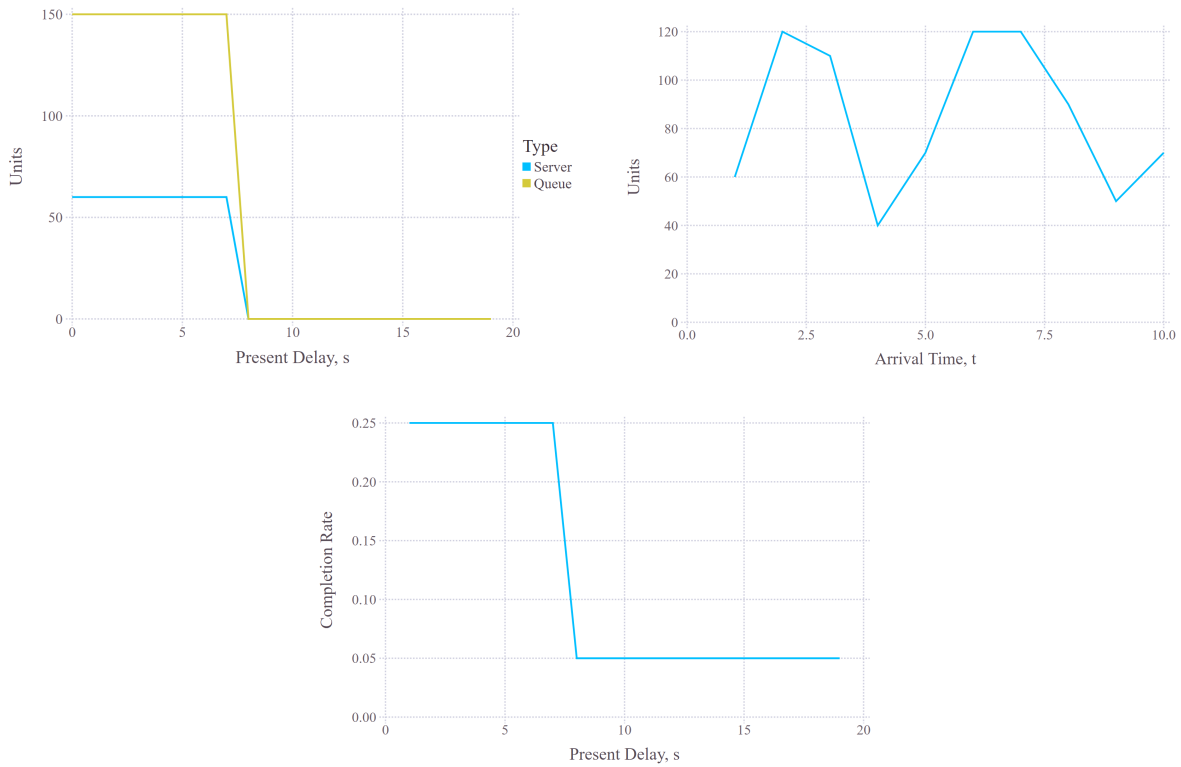



**Figure 5** Model Inputs: Initial State (left), Arrivals (right) and Service Distribution (bottom)

violation by that magnitude, and vice versa. Figure 6 plots these violations. Capacity violations have not occurred, which is expected since $\theta$ was small – capacity violation runs a probability of no more than $e^{-1/0.01/k^*} \approx 6.3 \times 10^{-5}$. That is $< 1$ of the $1,000$ simulations.

For the wait time and CVaR constraints, simulations that exceeded $k^*$ were colored red. To give a sense of the magnitude of $k^* = 10.34$, since the average wait time was set at 3.75, exceeding by $k^*$ equates to around 41 jobs experiencing additional delay of 0.25. Since the model was initiated with $1,200$ jobs in the queue, this translates to an estimated proportion of 3.4% of jobs overstaying by 0.25. Nonetheless, the instances of constraint violation were significantly lower than the guarantee. Table 5 counts this actual proportion. The performance of the model is read follows: There is a probability of around 3.3% that 3.4% of jobs in the queue will experience additional delay of 0.25, at any time $t$. For a model projecting $T = 10$ into the future, this performance is very reasonable.

**Table 5** Proportion of Simulations where Deviations Violated Bound $k^*$

| | Wait Time | CVaR |
|---|---|---|
| Actual proportion | 3.3% | 25.0% |
| Guarantee | 36.8% | 36.8% |

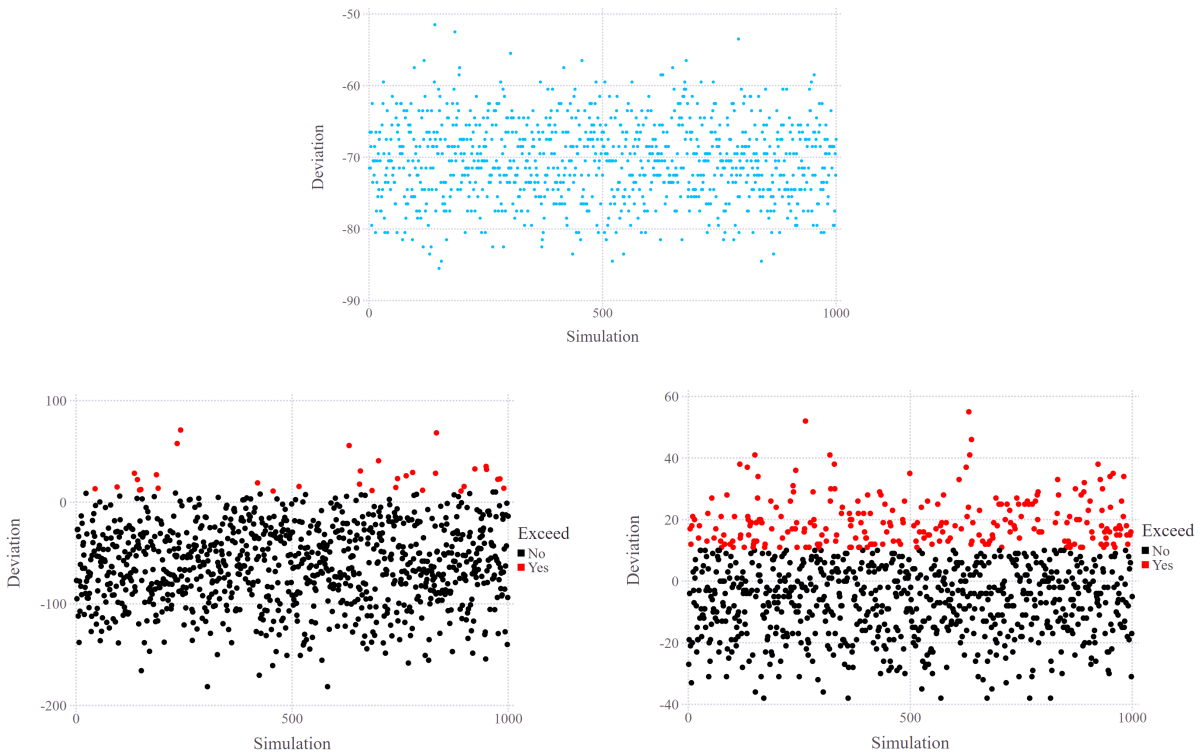



**Figure 6** Deviations from Server Capacity Limit (above), Average Waiting Time Constraint (left) and CVaR Constraint (right)

In this last segment, we examine computation time and complexity. The model was solved using Gurobi in Julia on an Intel® i7-6650U dual-core processor. Table 6 illustrates the computational time (in seconds) and the number of linear constraints in the last cutting plane iterate for different time horizons $T$. Figure 7 provides log-log plots of these variables.

**Table 6** Complexity of the Program

| **Time horizon,** $T$ | 7 | 8 | 9 | 10 | 12 | 15 |
|---|---|---|---|---|---|---|
| **Computational time (s)** | 19 | 66 | 234 | 724 | 3936 ($\approx$ 66 mins) | 13801 ($\approx$ 230 mins) |
| **Optimal risk,** $k^*$ | 0.88 | 2.12 | 4.61 | 10.47 | 52.33 | 170.76 |
| **# Constraints** | 14,352 | 23,084 | 39,468 | 60,971 | 104,385 | 162,867 |

As expected, the number of linear constraints grew by around $O(T^3)$, while actual computation time grew by around $O(T^9)$. This difference is attributed to the re-solving of sub-problems with ever growing number of constraints. This could be averted if the second-order cone method for solving the perspective cones (as described in Jaillet et al. 2018) is adopted.

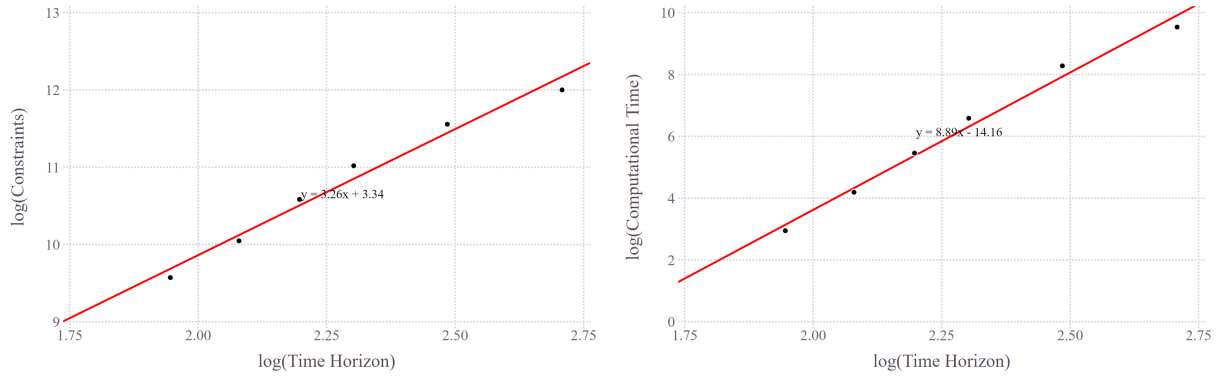



**Figure 7** Log-log Plots of Constraints (left) and Computational Time (right) against Time Horizon, $T$

## 5.2. Implementing our approach at PGIMER

The Postgraduate Institute of Medical Education and Research (PGIMER) is a premier medical and research institution in Chandigarh, India. With a capacity of 1950 beds serving over 2.5 million patients annually, PGIMER operates one of the biggest hospitals in India. PGIMER also operates a free tele-medicine hotline where people talk directly to physicians, nurses and volnteer medical students who offer medical assistance. In our collaboration with PGIMER, we were provided access to the complete dataset of their calls. This data set consists of 493,827 total number of patient calls over 12 months of PGIMER from September 1, 2017, until August 31, 2018. It is based on information collected solely from phone calls made between the patients and PGIMER. Out of the data set, 263,004 calls were identified with distinct patients. The data set does include recurring patients, some making as high as 101 calls. The composition of medical providers varies with time (during the day, week, or year). In our data, there are 61 medical providers who belong in one of four teams (main physicians, nurses, volunteers, and local language-speaking staff), representing a total of 33 distinct specializations.

In terms of the current operations, we observed that 79.74% of all calls made were queued either due to abandonment from the patient's end or overcapacity on PGIMER's end. After waiting for 9.44 minutes on average, 39.60% of patients in the queue abandoned before reaching a medical provider. Those who waited were then re-routed to main physicians (43%), nurses (36%), volunteers (6%), or local language-speaker (15%). The main physician queue saw a mean waiting time before service of 35.87 minutes, while the service itself took on average 14.25 minutes. On the other hand, the mean waiting and service times for the nurse queue were the shortest (11.91 minutes and 2.78 minutes, respectively). We summarize the statistics in Table 7.

**Table 7** Current Operational Statistics in PGIMER

| **Time (mins)** | Main Physician | Nurses | Volunteer | Local language |
|---|---|---|---|---|
| Mean Waiting Time | 35.87 | 3.17 | 16.95 | 37.75 |
| Mean Service Time | 14.25 | 3.09 | 23.14 | 12.37 |
| Mean Post-queue Time | 0.01 | 11.91 | 0.003 | 0.00 |
| Mean Wrap-up Time | 7.41 | 2.78 | 6.62 | 7.36 |

### PGIMER Current Operation

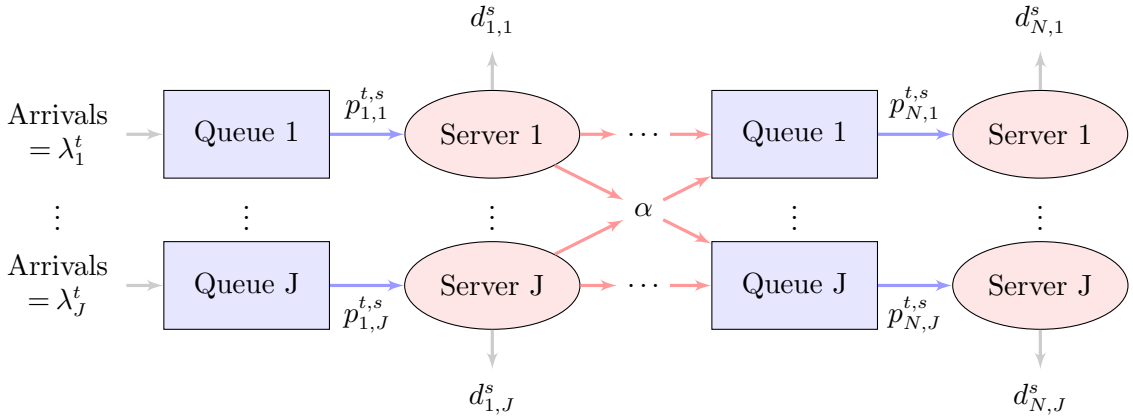
In the current operational structure of PGIMER, each call is assigned to a medical provider who will then be responsible for the call. This operational structure provides many advantages to both the main physicians and the nurses/volunteers by centralizing the communication between the patients and PGIMER. However, looking closely at the data provided, we observe that peak loads for different specializations occur at different points of times. Consequently, this variance is transferred to the loads of PGIMER physicians, hence creating a workload imbalance. The current operational structure does not inherently allow flexibility regarding assigning patients based on the load that a main physician is facing in a given period of time. In particular, we observed that there exists at least a single week where a PGIMER physician had to deal with 23 patients more than other PGIMER physicians.

### Modeling PGIMER as a P-Queue

As calls naturally undergo multiple stages of interaction and treatment where at each stage, they have some chance of being resolved, it is conceivable that routing calls between medical providers, as opposed to the assigning a dedicated one, would likely lead to overall reductions in waiting times. This can be modeled as a patient-flow control problem using a P-Queue, where at each stage, calls might be handled by one of the $J = 4$ possible types of medical providers. Figure 8 provides a schematic for this. In this subsection, we illustrate the results obtained under such an approach, where the aim is to control the flow of calls between stages, $\boldsymbol{\alpha}$, in order to minimize the total average waiting time across the queueing network. As discussed in Theorem 2, this problem has a quasi-convex reformulation.

We calibrate our models using actual data obtained from PGIMER, and compare our approach with various approaches proposed in the literature. In particular, we calculate the arrival and service rates at various nodes in the queueing network using historical data. We chose a value of $\theta = 0.05$ and obtained an optimal value of $k^* = 12.1$. We then compare our approach with the stochastic optimal obtained by solving the optimal control problem exactly using stochastic dynamic programming. We also report the waiting times obtained by using other approximations such as using the optimal static priority policy (see Andersson et al. 2001) and using a randomized

**Figure 8** P-Queue model of PGIMER inpatient flow




policy (see Huang et al. 2015). Specifically, we consider 10 of the 36 queues in the hospital queueing network and compute the performance of our policy and other policies using simulation using the calibrated distributions. The results are presented in Table 8.

**Table 8** Detailed results of simulations: Average queue lengths with their 95% confidence intervals.

| **S.No** | P-Queue approach | Stochastic Optimal | Static priority approximation | Randomized policy |
|---|---|---|---|---|
| 1 | 20.34 ± 0.24 | 20.12 ± 0.27 | 23.44 ± 0.31 | 25.74 ± 0.30 |
| 2 | 28.15 ± 0.28 | 27.86 ± 0.22 | 29.34 ± 0.21 | 30.24 ± 0.27 |
| 3 | 41.52 ± 0.34 | 38.76 ± 0.37 | 48.18 ± 0.36 | 43.14 ± 0.36 |
| 4 | 35.17± 0.27 | 35.14 ± 0.28 | 38.49 ± 0.31 | 36.26 ± 0.33 |
| 5 | 22.14 ± 0.21 | 20.46 ± 0.20 | 26.33 ± 0.21 | 24.68 ± 0.29 |
| 6 | 22.45 ± 0.22 | 22.23 ± 0.23 | 24.63 ± 0.28 | 27.52 ± 0.29 |
| 7 | 29.67 ± 0.18 | 28.25 ± 0.20 | 31.73 ± 0.23 | 32.04 ± 0.24 |
| 8 | 32.25 ± 0.29 | 31.64 ± 0.31 | 37.19 ± 0.33 | 36.44 ± 0.37 |
| 9 | 31.56± 0.23 | 31.04 ± 0.24 | 34.89 ± 0.29 | 37.62 ± 0.30 |
| 10 | 24.29 ± 0.19 | 22.16 ± 0.20 | 23.67 ± 0.21 | 25.79 ± 0.22 |

We observe that our approach achieves waiting times that are within 5% of the optimal waiting times in all the cases. In particular, for majority of the cases, we obtain the optimal solution within 1%. The other policies can obtain solutions that can be up to 15% away from the optimal solution. This demonstrates that our approach, while being tractable, also allows us to accurately approximate the true optimal control policy.

## 6. Discussions

The key idea in this paper is the tractability obtained through the introduction of state variable $s$, tracking the present delay in the queue. The central step in this procedure (say in the proof of Theorem 2) lies in the preservation of the $\exp(\cdot)$ form under expectations. In Jaillet et al. (2018),

they described this as 'pipeline invariance' and referred to $\exp(\cdot)$ as the 'preservation function'. The exponential dis-utility in the context of P-Queues is, in a broad sense, a geometry which is preserved under expectations. This can be viewed as a generalization of Lyapunov functions – if there exists a geometry preserving the joint space of decisions and uncertainty, then the problem becomes tractable. Nonetheless, as also remarked in Jaillet et al. (2018), it is not known if this technique can be extended to other dis-utilities, nor is it known if the technique works for more general dynamics.

Here, we also make a short comment about the connection to Robust Queueing. At first glance, the P-Queue model appears to be built on different primitives as Robust Queueing – in Robust Queueing, the distribution of the uncertainty is assumed to be unknown; in P-Queues, survival probabilities and hence service time distributions are known. However, constraint (12) has another interpretation – its dual is related to the Kullback-Leibler (KL) divergence. Hence, (12) can be interpreted as the expected delays under the uncertainty set of model rates deviating from the true rates about a ball of some KL divergence radius (see Jaillet et al. 2018, for details). In this lens, P-Queues are 'distributionally robust' to the estimation of service times.

**Duality and Lagrange Multipliers as Prices**. In the last discussion point, we mention a few comments about the duality of the P-Queue model. In typical convex optimization problems, the cost/price of the constraints can be understood in terms of the lagrange multipliers or the dual variables. Such insights can be drawn of the P-Queue Model.

This is made possible by the nature of the Binomial distribution, which leads to the P-Queue Model involving perspective cone constraints of the form $\beta \log(1 - d + d\exp(\xi_+/\beta)) \leq \xi_-$. Furthermore, the following well-known result enables the dual to be easily obtained:

LEMMA 1. *Let $\mathcal{K}_j$, $j \in \mathcal{J}$ be a collection of proper convex cones and $\mathcal{K} = \bigcap_{j\in\mathcal{J}} \mathcal{K}_j$. Then*

$$\mathcal{K}^* = \bigplus_{j\in\mathcal{J}} \mathcal{K}_j^* := \left\{ \sum_{j\in\mathcal{J}} k_j^* \;\middle|\; k_j^* \in \mathcal{K}_j^* \right\}. \tag{26}$$

The proof is illustrated in Appendix B. Since the P-Queue model can be written as a series of perspective cones, Lemma 1 reduces the dual to a sum of dual variables per cone. Moreover, if Slater's condition is satisfied, there would be no optimality gap between the primal and the dual formulations. But this is satisfied in the cases where the service distribution $d_v^{t,s} \in (0,1)$ is non-trivial and the risk parameter $k$ is non-zero and bounded. Moreover, Jaillet et al. (2018) gives a further decomposition of the perspective cone into exponential cones:

LEMMA 2. *(Jaillet et al. 2018) The constraint* $\beta \log(1-d+d\exp(\xi_+/\beta)) \leq \xi_-$ *has reformulation*

$$\begin{aligned} &(1-d)y_1 + dy_2 \leq \beta \\ &\beta \exp(-\xi_-/\beta) \leq y_1 \\ &\beta \exp(\xi_+ - \xi_-/\beta) \leq y_2 \end{aligned} \tag{27}$$

*Proof of Lemma 2.* The constraint is easily equivalent to $(1-d)\exp(-\xi_-/\beta) + d\exp(\xi_+ - \xi_-/\beta) \leq 1$. Since the expressions $\beta\exp(-\xi_-/\beta)$ and $\beta\exp(\xi_+ - \xi_-/\beta)$ are quasi-convex, it suffices to take their epigraphs, which leads to (27). □

For an illustration of duality applied on a stylized example, we refer the reader to Appendix B.

## 7. Conclusions

To summarize, in this paper, we demonstrate the power of considering hybrid state variables which lead to tractable formulations. As we discussed in the paper, many flow control problems reduce to quasi-convex optimization problems, which in many practical cases, are in fact, convex. In future work, we seek to explore the power of this viewpoint in obtaining structural results on the optimal control policies for various problems. Furthermore, note that the optimal control problems considered in this paper have structured convex representations, in particular, perspective cones. Because of this, it is possible to exploit these formulations to understand the associated dual problems to understand the 'price' of various capacity constraints and service level constraints. We hope to explore this in future follow-up works.

## Online Supplementary Materials

Appendices and the online version of this paper may be found on https://papers.ssrn.com/abstract=3190874.

## References

Andersson, Björn, Sanjoy Baruah, Jan Jonsson. 2001. Static-priority scheduling on multiprocessors. *Real-Time Systems Symposium, 2001.(RTSS 2001). Proceedings. 22nd IEEE*. IEEE, 193–202.

Armony, M., S. Israelit, A. Mandelbaum, Y N Marmor, Y. Tseytlin, G B Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.

Avram, F., D. Bertsimas, M. Ricard. 1995. An optimal control approach to optimization of multi-class queueing networks. F. Kelly, R. Williams, eds., *Volume 71 of IMA volumes in Mathematics and its Applications*. Springer-Verlag, New York, 199–234.

Bandi, C., D. Bertsimas, N. Youssef. 2015. Robust queueing theory. *Operations Research* **63**(3) 676–700.

Bäuerle, N. 2000. Asymptotic optimality of tracking policies in stochastic networks. *Annals of Applied Probability* 1065–1083.

Bäuerle, N. 2002. Optimal control of queueing networks: an approach via fluid models. *Advances in Applied Probability* **34**(2) 313–328.

Bertsimas, D., D. B. Brown, C. Caramanis. 2011. Theory and applications of robust optimization. *SIAM Review* **53**(3) 464–501.

Bertsimas, D., I. C. Paschalidis, J. N. Tsitsiklis. 1994. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *The Annals of Applied Probability* 43–75.

Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer-Verlag.

Brown, DB., M. Sim. 2009. Satisficing measures for analysis of risky positions. *Management Science* **55**(1) 71–84.

Bruneel, H., I. Wuyts. 1994. Analysis of discrete-time multiserver queueing models with constant service times. *Operations Research Letters* **15**(5) 231–236.

Buyya, R., R. Ranjan, R. N. Calheiros. 2009. Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities. *High Performance Computing & Simulation, 2009. HPCS'09. International Conference on*. IEEE, 1–11.

Calheiros, R. N., R. Ranjan, A. Beloglazov, C. A. F. De Rose, R. Buyya. 2011. Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience* **41**(1) 23–50.

Chen, H., D. D. Yao. 2013. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, vol. 46. Springer Science & Business Media.

Chen, M., I.-K. Cho, S. P. Meyn. 2006. Reliability by design in distributed power transmission networks. *Automatica* **42**(8) 1267–1281.

Chen, M., R. Dubrawski, S. P. Meyn. 2004. Management of demand-driven production systems. *IEEE Transactions on Automatic Control* **49**(5) 686–698.

Coffman, E. G., L. Kleinrock. 1968. Feedback queueing models for time-shared systems. *Journal of the ACM (JACM)* **15**(4) 549–576.

Dai, J. G. 1995. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* **5** 49–77.

Dai, J. G., P. Shi. 2017. A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* **65**(2) 514–536.

Dai, J. G., P. Shi. 2018. Inpatient overflow: An approximate dynamic programming approach. *MSOM* **Forthcoming**.

Dupuis, P., K. Ramanan. 2000. A multiclass feedback queueing network with a regular skorokhod problem. *Queueing Systems* **36**(4) 327–349.

Farrohknia, N., M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, K. E. Göransson. 2011. Emergency department triage scales and their components: a systematic review of the scientific evidence. *Scandinavian journal of trauma, resuscitation and emergency medicine* **19**(1) 42.

Harrison, J. M. 1988. Brownian models of queueing networks with heterogeneous customer populations. *Stochastic differential systems, stochastic control theory and applications*. Springer, 147–186.

Harrison, J. M. 1996. The BIGSTEP approach to flow management in stochastic processing networks. *Stochastic Networks: Theory and Applications* **4** 147–186.

Hluchyj, M. G., M. J. Karol. 1988. Queueing in high-performance packet switching. *IEEE Journal on selected Areas in Communications* **6**(9) 1587–1597.

Huang, Junfei, Boaz Carmeli, Avishai Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.

Jaillet, P., G.G. Loke, M. Sim. 2018. Risk-based manpower planning: A tractable multi-period model. Submitted. Extracted online from https://papers.ssrn.com/abstract=3168168.

Jaillet, P., J. Qi, M. Sim. 2016. Routing optimization under uncertainty. *Operations Research* **64**(1) 186–200.

Kumar, S., P. R Kumar. 1994. Performance bounds for queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* **39**(8) 1600–1611.

Larson, R. C., M. F. Cahn, M. C. Shell. 1993. Improving the New York City arrest-to-arraignment system. *Interfaces* **23** 76–96.

Lin, M., A. Wierman, L. L. H. Andrew, E. Thereska. 2013. Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Transactions on Networking* **21**(5) 1378–1391.

Mace, S. E., T. A. Mayer. 2008. Triage. *Pediatric emergency medicine*. Elsevier, 1087–1096.

Maglaras, C. 2000. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Annals of Applied Probability* 897–929.

Meyn, S. 1997. Stability and optimization of queueing networks and their fluid models. *Lectures in applied mathematics-American Mathematical Society* **33** 175–200.

Meyn, S. P. 2005. Workload models for stochastic networks: Value functions and performance evaluation. *IEEE Transactions on Automatic Control* **50**(8) 1106–1122.

Meyn, S. P. 2008. *Control techniques for complex networks*. Cambridge University Press.

Niska, R., F. Bhuiya, J. Xu, et al. 2010. National hospital ambulatory medical care survey: 2007 emergency department summary. *Natl Health Stat Report* **26**(26) 1–31.

Qi, J. 2017. Mitigating delays and unfairness in appointment systems. *Management Science* **63**(2) 566–583.

Shi, P., M. Chou, J. D. Dai, D. Ding, J. Sim. 2016. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* **62**(1) 1–28.

Srikant, R., L. Ying. 2013. *Communication networks: An optimization, control, and stochastic networks perspective*. Cambridge University Press.

Stolyar, A. L. 1995. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes and Related Fields* **1**(4) 491–512.

Takine, T., B. Sengupta, T. Hasegawa. 1994. An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. *IEEE Transactions on Communications* **42**(234) 1837–1845.

Williams, R. J. 1995. Semimartingale reflecting Brownian motions in the orthant. *IMA Volumes in Mathematics and its Applications* **71** 125–125.

Xie, J., Y. Jiang, M. Xie. 2011. A temporal approach to stochastic network calculus. Working paper available online at http://arxiv.org/abs/1112.2822.

Zhang, Q., L. Cheng, R. Boutaba. 2010. Cloud computing: State-of-the-art and research challenges. *Journal of internet services and applications* **1**(1) 7–18.

## A. Lemmas Used in Theorem 2

Here, we prove each of the lemmas in Theorem 2, as deferred.

LEMMA 3. *Let $v \in \mathcal{S}$ be a server. Suppose the out-degree of $v$ is 1, corresponding to a flow edge with $d_v^{t,s} = 1 - f_v^{t,s}$ representing the probability of remaining in the vertex, then the capacity constraint*

$$k \log \mathbb{E} \exp\left( \left( \sum_s x_v^{t,s} - \kappa_v \right) / k\theta \right) \leq 0 \tag{28}$$

*has the reformulation*

$$\begin{cases} k \sum_{s=0}^{t-1} \tilde{d}^{t,s} \sum_{s'} p_v^{t-s,s'} + k \sum_{s=t}^{last} x_v^{0,s-t} \hat{d}^{t,s} \leq \frac{\kappa_v}{\theta} & \text{if } v \text{ has only a push in-edge} \\ k \sum_{s=0}^{t-1} \log g_v^{t-s}(\tilde{d}^{t,s}) + k \sum_{s=t}^{last} x_v^{0,s-t} \hat{d}^{t,s} \leq \frac{\kappa_v}{\theta} & \text{if } v \text{ has only an arrival in-edge} \end{cases} \tag{29}$$

*Proof of Lemma 3.* In the absence of decisions, the dynamics equation (2) simplifies to

$$x_v^{t,s} = \text{Bin}(x_v^{t-1,s-1}, d_v^{t,s}) = \begin{cases} \text{Bin}\left( x_v^{t-s,0}, \prod_{\tau=0}^{s-1} d_v^{t-\tau,s-\tau} \right) & \text{if } s < t \\ \text{Bin}\left( x_v^{0,s-t}, \prod_{\tau=0}^{t-1} d_v^{t-\tau,s-\tau} \right) & \text{if } s \geq t \end{cases} \tag{30}$$

Additivity of the operator $k \log \mathbb{E} \exp(\cdot)$ under independence assures that (28) simplifies to

$$\sum_{s=0}^{t-1} k \log \mathbb{E} \exp\left( x_v^{t,s} / k\theta \right) + \sum_{s=t}^{\text{last}} k \log \mathbb{E} \exp\left( x_v^{t,s} / k\theta \right) \leq \frac{\kappa_v}{\theta} \tag{31}$$

The left-hand side simplifies further into

$$\sum_{s=0}^{t-1} k \log \mathbb{E} \exp\left( x_v^{t-s,0} \tilde{d}^{t,s} \right) + \sum_{s=t}^{\text{last}} k \log \mathbb{E} \exp\left( x_v^{0,s-t} \hat{d}^{t,s} \right) \tag{32}$$

where $\tilde{d}^{t,s} = \log\left( 1 - \prod_{\tau=0}^{s-1} d_v^{t-\tau,s-\tau} + e^{1/k\theta} \prod_{\tau=0}^{s-1} d_v^{t-\tau,s-\tau} \right)$ for $s \geq 1$, $\tilde{d}^{t,0} = 1$ and $\hat{d}^{t,s} = \log\Big( 1 - \prod_{\tau=0}^{t-1} d_v^{t-\tau,s-\tau} + e^{1/k\theta} \prod_{\tau=0}^{t-1} d_v^{t-\tau,s-\tau} \Big)$ are all constants given fixed $k$. The second term is, by assumption, known data. For the first term, there are two cases. If the in-edges into $v$ consists of only push edges, then $x_v^{t,0} = \sum_s p_v^{t,s}$, and the first term will also only consist of known values (decisions, in this case). As such, we recover the expression

$$k \sum_{s=0}^{t-1} \tilde{d}^{t,s} \sum_{s'} p_v^{t-s,s'} + k \sum_{s=t}^{\text{last}} x_v^{0,s-t} \hat{d}^{t,s} \leq \frac{\kappa_v}{\theta}. \tag{33}$$

Otherwise, there is an arrival into $v$ given by $x_v^{t,0} = \Lambda_v^t$, with moment generating function $g_v^t$. Hence, we instead recover the expression

$$k \sum_{s=0}^{t-1} \log g_v^{t-s}(\tilde{d}^{t,s}) + k \sum_{s=t}^{\text{last}} x_v^{0,s-t} \hat{d}^{t,s} \leq \frac{\kappa_v}{\theta}. \tag{34}$$

□

REMARK 3. Notice that apart from $k$, there are no decisions variables in the formulation for arrivals. As such, feasibility of the model immediately necessitates

$$k \leq \frac{\kappa_v}{\theta}\left(\sum_{s=0}^{t-1}\log g_v^{t-s}(\tilde{d}^{t,s}) + \sum_{s=t}^{\text{last}} x_v^{0,s-t}\hat{d}^{t,s}\right)^{-1} \tag{35}$$

LEMMA 4. *Let $v \in \mathcal{S}$ be a server. Suppose the out-degree of $v$ consists of a flow edge with $d_v^{t,s} = 1 - f_v^{t,s}$ representing the probability of remaining in the vertex and decisions $\alpha_{v,w}^{t,s}$ representing the proportion of existing units to be redirected to vertex $w$, then the capacity constraint*

$$k\log\mathbb{E}\exp\left(\left(\textstyle\sum_s x_v^{t,s} - \kappa_v\right)/k\theta\right) \leq 0 \tag{36}$$

*has the reformulation*

$$\begin{aligned}
&A(p_v,\xi_v^{\cdot,1};k) + k\sum_{s=t}^{last}\xi_v^{1,s-t+1} \leq \frac{\kappa_v}{\theta}\\
&\beta_v^{t,s}\log(1-d_v^{t,s}+d_v^{t,s}e^{1/k\theta}) \leq \xi_v^{t,s}\\
&\beta_v^{t-\tau,s-\tau}\log(1-d_v^{t-\tau,s-\tau}+d_v^{t-\tau,s-\tau}e^{\xi_v^{t-\tau+1,s-\tau+1}/\beta_v^{t-\tau,s-\tau}}) \leq \xi_v^{t-\tau,s-\tau} \qquad \begin{matrix}\tau = 1\ldots,t-1\\ s>\tau\end{matrix}
\end{aligned} \tag{37}$$

*where*

$$A(p_v,\xi_v^{\cdot,1};k) = \begin{cases} \frac{1}{\theta}\sum_s p_v^{t,s} + k\sum_{s=1}^{t-1}\xi_v^{t-s+1,1} & \text{if } v \text{ has only a push in-edge}\\ k\log g_v^t(1/k\theta) + k\sum_{s=1}^{t-1}\log g_v^t(\xi_v^{t-s+1,1}/k) & \text{if } v \text{ has only an arrival in-edge}\end{cases} \tag{38}$$

*Proof of Lemma 4.* Notice that in the dynamics defined for such a server $v \in \mathcal{S}$ in (2), the actual distribution of units out of the server via the individual $\alpha_{v,w}^{t,s}$'s is immaterial in whether or not the capacity is going to be exceeded or not, insofar as their sum is controlled. As such, let us use the reformulation $\frac{\beta_v^{t,s}}{\beta_v^{t-1,s-1}} := 1 - \sum_{v\sim w}\alpha_{v,w}^{t-1,s-1}$, and then later define in Lemma 7, a suitable reformulation that preserves the conservation of flow. Also notice that this definition affords us some degrees of freedom to define $\beta_v^{0,s}$ and $\beta_v^{t,0}$ as suitable.

Under this regime, the dynamics (2) is reformulated as $x_v^{t,s} = \text{Bin}\left(x_v^{t-1,s-1}\frac{\beta_v^{t,s}}{\beta_v^{t-1,s-1}}, d_v^{t,s}\right)$. As such, by additivity and independence, (36) may be reformulated as

$$k\log\mathbb{E}\exp\left(x_v^{t,0}/k\theta\right) + k\sum_{s=1}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t,s}/k\theta\right) \leq \frac{\kappa_v}{\theta} \tag{39}$$

We now perform a series of repeated expansions of the expectation. The larger principle behind why this expansion works is described as the property of 'Pipeline Invariance' in Jaillet et al. (2018). Now the left-hand side of this expression is given as

$$k\log\mathbb{E}\exp\left(x_v^{t,0}/k\theta\right) + k\sum_{s=1}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-1,s-1}\frac{\beta_v^{t,s}}{\beta_v^{t-1,s-1}}\log(1-d_v^{t,s}+d_v^{t,s}e^{1/k\theta})\right)$$

$$
\begin{aligned}
&= k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right) + k\log\mathbb{E}\exp\left(x_v^{t-1,0}\frac{\xi_v^{t,1}}{\beta_v^{t-1,0}}\right) + k\sum_{s=2}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-1,s-1}\frac{\xi_v^{t,s}}{\beta_v^{t-1,s-1}}\right)\\
&= k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right) + k\log\mathbb{E}\exp\left(x_v^{t-1,0}\frac{\xi_v^{t,1}}{\beta_v^{t-1,0}}\right)\\
&\quad + k\sum_{s=2}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-2,s-2}\frac{\beta_v^{t-1,s-1}}{\beta_v^{t-2,s-2}}\log(1-d_v^{t-1,s-1}+d_v^{t-1,s-1}e^{\xi_v^{t,s}/\beta_v^{t-1,s-1}})\right)\\
&= k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right) + \sum_{\tau=1}^{2} k\log\mathbb{E}\exp\left(x_v^{t-\tau,0}\frac{\xi_v^{t-\tau+1,1}}{\beta_v^{t-\tau,0}}\right) + k\sum_{s=3}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-2,s-2}\frac{\xi_v^{t-1,s-1}}{\beta_v^{t-2,s-2}}\right)\\
&= \dots\\
&= k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right) + \sum_{\tau=1}^{l} k\log\mathbb{E}\exp\left(x_v^{t-\tau,0}\frac{\xi_v^{t-\tau+1,1}}{\beta_v^{t-\tau,0}}\right) + k\sum_{s=l+1}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{t-l,s-l}\frac{\xi_v^{t-l+1,s-l+1}}{\beta_v^{t-l,s-l}}\right)\\
&= \dots\\
&= k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right) + \sum_{\tau=1}^{t} k\log\mathbb{E}\exp\left(x_v^{t-\tau,0}\frac{\xi_v^{t-\tau+1,1}}{\beta_v^{t-\tau,0}}\right) + k\sum_{s=t+1}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{0,s-t}\frac{\xi_v^{1,s-t+1}}{\beta_v^{0,s-t}}\right)\\
&= k\log\mathbb{E}\exp\left(x_v^{t,0}\frac{1}{k\theta}\right) + \sum_{s=1}^{t-1} k\log\mathbb{E}\exp\left(x_v^{t-s,0}\frac{\xi_v^{t-s+1,1}}{\beta_v^{t-s,0}}\right) + k\sum_{s=t}^{\text{last}}\log\mathbb{E}\exp\left(x_v^{0,s-t}\frac{\xi_v^{1,s-t+1}}{\beta_v^{0,s-t}}\right) \qquad (40)
\end{aligned}
$$

where in each iterate, we define the auxiliary variables $\xi$ to serve as epigraphs, as follows

$$
\begin{aligned}
&\beta_v^{t,s}\log(1-d_v^{t,s}+d_v^{t,s}e^{1/k\theta}) \le \xi_v^{t,s}\\
&\beta_v^{t-\tau,s-\tau}\log(1-d_v^{t-\tau,s-\tau}+d_v^{t-\tau,s-\tau}e^{\xi_v^{t-\tau+1,s-\tau+1}/\beta_v^{t-\tau,s-\tau}}) \le \xi_v^{t-\tau,s-\tau} \qquad \begin{matrix}\tau=1\dots,t-1\\ s>\tau\end{matrix} \qquad (41)
\end{aligned}
$$

Because the original constraint is convex, by the convexity of the operator $k\log\mathbb{E}\exp(\cdot)$ and moreover, each of these epigraphs is jointly convex in $\beta$ and $\xi$ because they are just perspectives of a convex function, this expansion is legitimate.

Now, the last term is known (for it contains either the initial data or decision variables). Hence, we utilize our degree of freedom to define $\beta_v^{0,s} := x_v^{0,s}$ for all $s$, to obtain the simple expression $k\sum_{s=t}^{\text{last}}\xi_v^{1,s-t+1}$ for the last term.

To complete the computation, we evaluate (40). If indeed, the in-edge to server $v$ is a push edge, then $x_v^{t,0} = \sum_s p_v^{t,s}$ and again, the first two terms will be known. Similarly utilizing our degree of freedom, we can let $\beta_v^{t,0} := x_v^{t,0} = \sum_s p_v^{t,s}$ for all $t$, to get $\frac{1}{\theta}\sum_s p_v^{t,s}$ for the first term and $k\sum_s \xi_v^{t-s+1,1}$ for the second term.

If instead, the in-edge is an arrival, then we continue the evaluation to obtain $k\log g_v^t(1/k\theta)$ for the first term and $k\sum_{s=1}^{t-1}\log g_v^{t-s}(\xi_v^{t-s+1,1}/\beta_v^{t-s,0})$ for the second. In this case, we can pick $\beta_v^{t,0}\equiv k$ for all $t$ to preserve the overall quasi-convexity of the problem. □

Remark 4. Often, for a fixed server $v$, the arrivals $\Lambda_v^t$ belong to some family where the moment generating functions are parametrized by some parameter $\lambda^t$ over time $t$, that is, $g_v^t(\cdot) = g_v(\cdot;\lambda^t)$,

*e.g.* Poisson arrivals. If indeed further that $\log g_v(z;\lambda^t) = \lambda^t h(z)$ for some convex $h$, then it may make more sense to define $\beta_v^{t,0} = \lambda^t$. In this way, one can create a further auxiliary variable where $\lambda^{t-s} h(\xi_v^{t-s+1,1}/\lambda^{t-s}) \leq \eta_v^{t-s}$ and the sum simply reduces to $k\sum_{s=1}^{t-1} \eta_v^{t-s}$. The benefit of this approach is that $\lambda^t$ has the same role as $\alpha$ in the constraint here, so in the case where $\lambda^t$ is a decision variable, the problem remains tractable.

LEMMA 5. *Let $v \in \mathcal{Q}$ be a queue. Suppose the in-degree of $v$ is 1, corresponding to an arrival $\Lambda_v^t$, with moment generating function $g_v^t$, then the capacity constraint*

$$k \log \mathbb{E} \exp\left( \left( \sum_s x_v^{t,s} a_v^s - b_v^t \right) / k\theta_v^t \right) \leq 0 \tag{42}$$

*has the reformulation*

$$\begin{aligned} k\sum_{s=0}^{t-1} \log g_v^t(a_v^s/k\theta_v^t) + \frac{1}{\theta_v^t}\sum_{s=t}^{last} x_v^{0,s-t} a_v^s \\ - \frac{1}{\theta_v^t}\sum_{s=0}^{t-1} a_v^s \sum_{v\sim w}\sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} - \frac{1}{\theta_v^t}\sum_{s=t}^{last} a_v^s \sum_{v\sim w}\sum_{\tau=0}^{t-1} p_w^{t-\tau,s-\tau} \leq \frac{b_v^t}{\theta_v^t} \end{aligned} \tag{43}$$

*Proof of Lemma 5.* Starting from the dynamics (3) and using additivity under independence as before, we have that (42) has reformulation

$$\begin{aligned} \sum_{s=0}^{t-1} k \log \mathbb{E} \exp\left( x_v^{t-s,0} a_v^s / k\theta_v^t \right) + \sum_{s=t}^{\text{last}} k \log \mathbb{E} \exp\left( x_v^{0,s-t} a_v^s / k\theta_v^t \right) \\ - k\sum_{s=0}^{t-1} \log \mathbb{E} \exp\left( a_v^s \sum_{v\sim w}\sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} / k\theta_v^t \right) - k\sum_{s=t}^{\text{last}} \log \mathbb{E} \exp\left( a_v^s \sum_{v\sim w}\sum_{\tau=0}^{t-1} p_w^{t-\tau,s-\tau} / k\theta_v^t \right) \leq \frac{b_v^t}{\theta_v^t} \end{aligned} \tag{44}$$

The last three terms are known quantities. Using the moment generating function expression for arrivals at $x_v^{t-s,0}$, we recover the desired expression. □

LEMMA 6. *Let $v \in \mathcal{Q}$ be a queue. Suppose the in-degree of $v$ consists of only a flow from server $u \in \mathbb{S}$, then the capacity constraint*

$$k \log \mathbb{E} \exp\left( \left( \sum_s x_v^{t,s} a_v^s - b_v^t \right) / k\theta_v^t \right) \leq 0 \tag{45}$$

*has the reformulation*

$$\begin{aligned} \frac{1}{\theta_v^t}\sum_{s=t}^{last} x_v^{0,s-t} a_v^s - \frac{1}{\theta_v^t}\sum_{s=0}^{t-1} a_v^s \sum_{v\sim w}\sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} - \frac{1}{\theta_v^t}\sum_{s=t}^{last} a_v^s \sum_{v\sim w}\sum_{\tau=0}^{t-1} p_w^{t-\tau,s-\tau} \\ + k\sum_{s=1}^{t}\sum_{\tau=s}^{last} \zeta_u^{s,\tau,s} + A_u(p_u, \zeta_u; k) \leq \frac{b_v^t}{\theta_v^t} \end{aligned}$$

$$\begin{aligned}
&\beta_u^{s,\tau}\log\left(1-f_u^{s,\tau}+f_u^{s,\tau}e^{a_v^{t-s}/k\theta_v^t}\right)\le\zeta_u^{s,\tau,1} && s=1\dots,t,\ \tau\ge 1\\
&\beta_u^{s-l,\tau-l}\log(1-f_u^{s-l,\tau-l}+f_u^{s-l,\tau-l}e^{\zeta_u^{s,\tau,l}/\beta_u^{s-l,\tau-l}})\le\zeta_u^{s,\tau,l+1} && s=2\dots,t,\ \tau\ge 2,\ l\le\min\{s-1,\tau-1\}
\end{aligned} \tag{46}$$

*where*

$$A_u(p_u,\zeta_u;k)=\begin{cases}k\displaystyle\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t}\zeta_u^{s,\tau,\tau} & \text{if } u \text{ has only a push in-edge}\\ k\displaystyle\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t}\log g_u^{s-\tau}(\zeta_u^{s,\tau,\tau}/k) & \text{if } u \text{ has only an arrival in-edge}\end{cases}$$

*Proof of Lemma 6.* As with the proof of Lemma 5, we arrive at equation (44), where the last three terms are known quantities. It leaves to evaluate the unwieldy term

$$\sum_{s=0}^{t-1}k\log\mathbb{E}\exp\left(x_v^{t-s,0}a_v^s/k\theta_v^t\right)=\sum_{s=1}^{t}k\log\mathbb{E}\exp\left(x_v^{s,0}a_v^{t-s}/k\theta_v^t\right) \tag{47}$$

Contiguous with the definition of $\frac{\beta_u^{t,s}}{\beta_u^{t-1,s-1}}:=1-\sum_{u\sim y}\alpha_{u,y}^{t-1,s-1}$ in Lemma 4, we have that the dynamics in (4) becomes $x_v^{t,0}=\sum_{s=1}^{\text{last}}\chi_v^{t-1,s}$, where

$$\chi_v^{t-1,s}\sim\text{Bin}\left(x_u^{t-1,s-1}\frac{\beta_u^{t,s}}{\beta_u^{t-1,s-1}},f_u^{t,s}\right)$$

In this case, (47) can be evaluated in a fashion similar to the proof in Lemma 4.

$$\begin{aligned}
&k\sum_{s=1}^{t}\log\mathbb{E}\exp\left(x_v^{s,0}a_v^{t-s}/k\theta_v^t\right)\\
&=k\sum_{s=1}^{t}\sum_{\tau=1}^{\text{last}}\log\mathbb{E}\exp\left(\chi_v^{s-1,\tau}a_v^{t-s}/k\theta_v^t\right)\\
&=k\sum_{s=1}^{t}\sum_{\tau=1}^{\text{last}}\log\mathbb{E}\exp\left(x_u^{s-1,\tau-1}\frac{\beta_u^{s,\tau}}{\beta_u^{s-1,\tau-1}}\log(1-f_v^{s,\tau}+f_v^{s,\tau}e^{a_v^{t-s}/k\theta_v^t})\right)\\
&=k\sum_{s=1}^{t}\sum_{\tau=1}^{\text{last}}\log\mathbb{E}\exp\left(x_u^{s-1,\tau-1}\frac{\zeta_u^{s,\tau,1}}{\beta_u^{s-1,\tau-1}}\right)\\
&=k\sum_{\tau=1}^{\text{last}}\log\mathbb{E}\exp\left(x_u^{0,\tau-1}\frac{\zeta_u^{1,\tau,1}}{\beta_u^{0,\tau-1}}\right)+k\sum_{s=2}^{t}\log\mathbb{E}\exp\left(x_u^{s-1,0}\frac{\zeta_u^{s,1,1}}{\beta_u^{s-1,0}}\right)\\
&\quad+k\sum_{s=2}^{t}\sum_{\tau=2}^{\text{last}}\log\mathbb{E}\exp\left(x_u^{s-2,\tau-2}\frac{\beta_u^{s-1,\tau-1}}{\beta_u^{s-2,\tau-2}}\log\left(1-f_v^{s-1,\tau-1}+f_v^{s-1,\tau-1}e^{\zeta_u^{s,\tau,1}/\beta_u^{s-1,\tau-1}}\right)\right)\\
&=k\sum_{s=1}^{2}\sum_{\tau=s}^{\text{last}}\log\mathbb{E}\exp\left(x_u^{0,\tau-s}\frac{\zeta_u^{s,\tau,s}}{\beta_u^{0,\tau-s}}\right)+k\sum_{\tau=1}^{2}\sum_{s=\tau+1}^{t}\log\mathbb{E}\exp\left(x_u^{s-\tau,0}\frac{\zeta_u^{s,\tau,\tau}}{\beta_u^{s-\tau,0}}\right)\\
&\quad+k\sum_{s=3}^{t}\sum_{\tau=3}^{\text{last}}\log\mathbb{E}\exp\left(x_u^{s-2,\tau-2}\frac{\zeta_u^{s,\tau,2}}{\beta_u^{s-2,\tau-2}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \cdots \\
&= k\sum_{s=1}^{l}\sum_{\tau=s}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{0,\tau-s}\frac{\zeta_u^{s,\tau,s}}{\beta_u^{0,\tau-s}}\right) + k\sum_{\tau=1}^{l}\sum_{s=\tau+1}^{t} \log \mathbb{E}\exp\left(x_u^{s-\tau,0}\frac{\zeta_u^{s,\tau,\tau}}{\beta_u^{s-\tau,0}}\right) \\
&\quad + k\sum_{s=l+1}^{t}\sum_{\tau=l+1}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{s-l,\tau-l}\frac{\zeta_u^{s,\tau,l}}{\beta_u^{s-l,\tau-l}}\right) \\
&= \cdots \\
&= k\sum_{s=1}^{t-1}\sum_{\tau=s}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{0,\tau-s}\frac{\zeta_u^{s,\tau,s}}{\beta_u^{0,\tau-s}}\right) + k\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t} \log \mathbb{E}\exp\left(x_u^{s-\tau,0}\frac{\zeta_u^{s,\tau,\tau}}{\beta_u^{s-\tau,0}}\right) \\
&\quad + \sum_{\tau=t}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{1,\tau-t+1}\frac{\zeta_u^{t,\tau,t-1}}{\beta_u^{1,\tau-t+1}}\right) \\
&= k\sum_{s=1}^{t-1}\sum_{\tau=s}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{0,\tau-s}\frac{\zeta_u^{s,\tau,s}}{\beta_u^{0,\tau-s}}\right) + k\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t} \log \mathbb{E}\exp\left(x_u^{s-\tau,0}\frac{\zeta_u^{s,\tau,\tau}}{\beta_u^{s-\tau,0}}\right) \\
&\quad + \sum_{\tau=t}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{0,\tau-t}\frac{\zeta_u^{t,\tau,t}}{\beta_u^{0,\tau-t}}\right) \\
&= k\sum_{s=1}^{t}\sum_{\tau=s}^{\text{last}} \log \mathbb{E}\exp\left(x_u^{0,\tau-s}\frac{\zeta_u^{s,\tau,s}}{\beta_u^{0,\tau-s}}\right) + k\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t} \log \mathbb{E}\exp\left(x_u^{s-\tau,0}\frac{\zeta_u^{s,\tau,\tau}}{\beta_u^{s-\tau,0}}\right) \qquad (48)
\end{aligned}
$$

with accompanying epigraphs

$$
\begin{aligned}
&\beta_u^{s,\tau}\log\left(1-f_u^{s,\tau}+f_u^{s,\tau}e^{a_v^{t-s}/k\theta_v^t}\right) \le \zeta_u^{s,\tau,1} && s=1\ldots,t,\ \tau\ge 1 \\
&\beta_u^{s-l,\tau-l}\log(1-f_u^{s-l,\tau-l}+f_u^{s-l,\tau-l}e^{\zeta_u^{s,\tau,l}/\beta_u^{s-l,\tau-l}}) \le \zeta_u^{s,\tau,l+1} && s=2\ldots,t,\ \tau\ge 2,\ l\le\min\{s-1,\tau-1\} \qquad (49)
\end{aligned}
$$

The first summation in (48), aligned with the proof in Lemma 4, the definition of $\beta_u^{0,s}=x_u^{0,s}$ will reduce it to $k\sum_{s=1}^{t}\sum_{\tau=s}^{\text{last}}\zeta_u^{s,\tau,s}$. For the second sum, again, depending on the nature of the in-edges of the server $u$, we either recover $k\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t}\zeta_u^{s,\tau,\tau}$, with the contiguous definition of $\beta_u^{t,0}:=x_u^{t,0}=\sum_s p_u^{t,s}$ in the case of a push edge, or $k\sum_{\tau=1}^{t-1}\sum_{s=\tau+1}^{t}\log g_u^{s-\tau}(\zeta_u^{s,\tau,\tau}/k)$, again aligned with the definition of $\beta_u^{t,0}\equiv k$ in the case of an arrival. This completes the proof. □

Remark 5. Lemma 6 omits the case when the dispatching server $u$ does not have any decisions. This is easy to fix as the case simplifies further.

Lemma 7. *Let $v\in\mathcal{Q}$ be a queue. Suppose the in-degree of $v$ consists of only decision in-edges, then the capacity constraint*

$$k\log\mathbb{E}\exp\left(\left(\sum_s x_v^{t,s}a_v^s-b_v^t\right)/k\theta_v^t\right)\le 0 \qquad (50)$$

*has the reformulation*

$$\frac{1}{\theta_v^t}\sum_{s=t}^{last}x_v^{0,s-t}a_v^s-\frac{1}{\theta_v^t}\sum_{s=0}^{t-1}a_v^s\sum_{v\sim w}\sum_{\tau=0}^{s-1}p_w^{t-\tau,s-\tau}-\frac{1}{\theta_v^t}\sum_{s=t}^{last}a_v^s\sum_{v\sim w}\sum_{\tau=0}^{t-1}p_w^{t-\tau,s-\tau}$$

$$+\sum_{\substack{w\sim v\\ w\neq u}}\left[\frac{1}{\theta_v^t}\sum_{\tau=1}^{last}a_v^{t-1}(\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau})+k\sum_{s=2}^{t}\sum_{\tau=s}^{last}\phi_w^{s-1,\tau-1,s-1}+B_w(p_w,\phi_w;k)\right]\leq\frac{b_v^t}{\theta_v^t}$$

$$\begin{aligned}
&\beta_w^{s-1,\tau-1}\log\left(1-d_w^{s-1,\tau-1}+d_w^{s-1,\tau-1}e^{\frac{\beta_{w,i_w-1}^{s,\tau}-\beta_{w,i_w}^{s,\tau}}{\beta_w^{s-1,\tau-1}}\frac{a_v^{t-s}}{k\theta_v^t}}\right)\leq\phi_w^{s-1,\tau-1,1} && \begin{array}{l}s=2\ldots,t\\ \tau\geq 2\end{array}\\
&\beta_w^{s-l,\tau-l}\log(1-d_w^{s-l,\tau-l}+d_w^{s-l,\tau-l}e^{\phi_w^{s-1,\tau-1,l-1}/\beta_w^{s-l,\tau-l}})\leq\phi_u^{s-1,\tau-1,l} && \begin{array}{l}s=3\ldots,t\\ \tau\geq 3\\ l=2,\ldots,\min\{s-1,\tau-1\}\end{array}
\end{aligned}\tag{51}$$

*where*

$$B_w(\zeta_w;k,p_w)=\begin{cases}\frac{1}{\theta_j^t}\sum_{s=2}^{t}a_j^{t-s}\ (\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1})\\ \qquad +k\sum_{\tau=2}^{t-1}\sum_{s=\tau+1}^{t}\phi_w^{s-1,\tau-1,\tau-1} & \text{if } w \text{ has only a push in-edge}\\ k\sum_{s=2}^{t}\log\ g_w^{s-1}\left(a_j^{t-s}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{k^2\theta_j^t}\right)\\ \qquad +k\sum_{\tau=2}^{t-1}\sum_{s=\tau+1}^{t}\log g_w^{s-\tau}\left(\frac{\phi_w^{s-1,\tau-1,\tau-1}}{k}\right) & \text{if } w \text{ has only an arrival in-edge}\end{cases}$$

*Proof of Lemma 7.* Like Lemma 6, we have to evaluate the term (47). In this case, however, we need to examine the definition of $\beta$ carefully so as to maintain conservation of flow.

Consider a server vertex $w$ dispatching units to $Y$ different queues $y_i$, $i=1,\ldots,Y$. At time $t$, the total proportion of units that will be dispatched amongst those that have spent $s-1$ amount of time in $w$ will be $1-\frac{\beta_w^{t,s}}{\beta_w^{t-1,s-1}}$. Notice in fact that we could write, by means of a telescopic series

$$\frac{\beta_w^{t-1,s-1}-\beta_w^{t,s}}{\beta_w^{t-1,s-1}}=\sum_{i=1}^{Y}\frac{\beta_{w,i-1}^{t,s}-\beta_{w,i}^{t,s}}{\beta_w^{t-1,s-1}}\tag{52}$$

where $\beta_{w,0}^{t,s}=\beta_w^{t-1,s-1}$. Then we can define the proportion of units to be dispatched to queue $y_i$ to be $\alpha_{w,y_i}^{t-1,s-1}:=\frac{\beta_{w,i-1}^{t,s}-\beta_{w,i}^{t,s}}{\beta_u^{t-1,s-1}}$. Thus, if for server $w$, the queue $v$ is the $i_w$-th queue, then the incoming dynamics for $v$ can be stated as

$$x_v^{t,0}=\sum_{\substack{w\sim v\\ w\neq u}}\sum_{s=1}^{\text{last}}x_w^{t-1,s-1}\frac{\beta_{w,i_w-1}^{t,s}-\beta_{w,i_w}^{t,s}}{\beta_w^{t-1,s-1}}\tag{53}$$

Hence the term in question (47) evaluates as follows:

$$\begin{aligned}
&k\sum_{s=1}^{t}\log\mathbb{E}\exp\left(x_v^{s,0}a_v^{t-s}/k\theta_v^t\right)\\
=&k\sum_{s=1}^{t}\sum_{\substack{w\sim v\\ w\neq u}}\sum_{\tau=1}^{\text{last}}\log\mathbb{E}\exp\left(x_w^{s-1,\tau-1}\frac{\beta_{w,i_w-1}^{s,\tau}-\beta_{w,i_w}^{s,\tau}}{\beta_w^{s-1,\tau-1}}\frac{a_v^{t-s}}{k\theta_v^t}\right)
\end{aligned}\tag{54}$$

It is understood that the sum will be over all $w : w \sim v, w \neq u$. Hence, for brevity, let us omit that summation over all subsequent analysis.

$$
\begin{aligned}
& k\sum_{s=1}^{t}\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-1,\tau-1}\frac{\beta_{w,i_w-1}^{s,\tau}-\beta_{w,i_w}^{s,\tau}}{\beta_w^{s-1,\tau-1}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
=& k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t} \log \mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
& + k\sum_{s=2}^{t}\sum_{\tau=2}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-1,\tau-1}\frac{\beta_{w,i_w-1}^{s,\tau}-\beta_{w,i_w}^{s,\tau}}{\beta_w^{s-1,\tau-1}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
=& k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t} \log \mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
& + k\sum_{s=2}^{t}\sum_{\tau=2}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-2,\tau-2}\frac{\beta_w^{s-1,\tau-1}}{\beta_w^{s-2,\tau-2}}\log\left(1-d_w^{s-1,\tau-1}+d_w^{s-1,\tau-1}\exp\left(\frac{\beta_{w,i_w-1}^{s,\tau}-\beta_{w,i_w}^{s,\tau}}{\beta_w^{s-1,\tau-1}}\frac{a_v^{t-s}}{k\theta_v^t}\right)\right)\right) \\
=& k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t} \log \mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
& + k\sum_{s=2}^{t}\sum_{\tau=2}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-2,\tau-2}\frac{\phi_w^{s-1,\tau-1,1}}{\beta_w^{s-2,\tau-2}}\right) \\
=& k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t} \log \mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
& + k\sum_{\tau=2}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-2}\frac{\phi_w^{1,\tau-1,1}}{\beta_w^{0,\tau-2}}\right) + k\sum_{s=3}^{t} \log \mathbb{E}\exp\left(x_w^{s-2,0}\frac{\phi_w^{s-1,1,1}}{\beta_w^{s-2,0}}\right) \\
& + k\sum_{s=3}^{t}\sum_{\tau=3}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-2,\tau-2}\frac{\phi_w^{s-1,\tau-1,1}}{\beta_w^{s-2,\tau-2}}\right) \\
=& k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t} \log \mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
& + k\sum_{s=2}^{3}\sum_{\tau=s}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-s}\frac{\phi_w^{s-1,\tau-1,s-1}}{\beta_w^{0,\tau-s}}\right) + k\sum_{\tau=2}^{3}\sum_{s=\tau+1}^{t} \log \mathbb{E}\exp\left(x_w^{s-\tau,0}\frac{\phi_w^{s-1,\tau-1,\tau-1}}{\beta_w^{s-\tau,0}}\right) \\
& + k\sum_{s=4}^{t}\sum_{\tau=4}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-3,\tau-3}\frac{\phi_w^{s-1,\tau-1,2}}{\beta_w^{s-2,\tau-2}}\right) \\
=& \cdots \\
=& k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t} \log \mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
& + k\sum_{s=2}^{l}\sum_{\tau=s}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-s}\frac{\phi_w^{s-1,\tau-1,s-1}}{\beta_w^{0,\tau-s}}\right) + k\sum_{\tau=2}^{l}\sum_{s=\tau+1}^{t} \log \mathbb{E}\exp\left(x_w^{s-\tau,0}\frac{\phi_w^{s-1,\tau-1,\tau-1}}{\beta_w^{s-\tau,0}}\right) \\
& + k\sum_{s=l+1}^{t}\sum_{\tau=l+1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{s-l,\tau-l}\frac{\phi_w^{s-1,\tau-1,l-1}}{\beta_w^{s-l,\tau-l}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \cdots \\
&= k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t}\log\mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
&\quad + k\sum_{s=2}^{t-1}\sum_{\tau=s}^{\text{last}}\log\mathbb{E}\exp\left(x_w^{0,\tau-s}\frac{\phi_w^{s-1,\tau-1,s-1}}{\beta_w^{0,\tau-s}}\right) + k\sum_{\tau=2}^{t-1}\sum_{s=\tau+1}^{t}\log\mathbb{E}\exp\left(x_w^{s-\tau,0}\frac{\phi_w^{s-1,\tau-1,\tau-1}}{\beta_w^{s-\tau,0}}\right) \\
&\quad + k\sum_{\tau=t}^{\text{last}}\log\mathbb{E}\exp\left(x_w^{1,\tau-t+1}\frac{\phi_w^{t-1,\tau-1,t-2}}{\beta_w^{1,\tau-t+1}}\right) \\
&= k\sum_{\tau=1}^{\text{last}} \log \mathbb{E}\exp\left(x_w^{0,\tau-1}\frac{\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}}{\beta_w^{0,\tau-1}}\frac{a_v^{t-1}}{k\theta_v^t}\right) + k\sum_{s=2}^{t}\log\mathbb{E}\exp\left(x_w^{s-1,0}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{\beta_w^{s-1,0}}\frac{a_v^{t-s}}{k\theta_v^t}\right) \\
&\quad + k\sum_{s=2}^{t}\sum_{\tau=s}^{\text{last}}\log\mathbb{E}\exp\left(x_w^{0,\tau-s}\frac{\phi_w^{s-1,\tau-1,s-1}}{\beta_w^{0,\tau-s}}\right) + k\sum_{\tau=2}^{t-1}\sum_{s=\tau+1}^{t}\log\mathbb{E}\exp\left(x_w^{s-\tau,0}\frac{\phi_w^{s-1,\tau-1,\tau-1}}{\beta_w^{s-\tau,0}}\right)
\end{aligned}
$$

This time, the expression is not convex, but it is still quasi-convex, which enables us to utilize the same epigraph approach to create auxiliary variables $\phi_w$, as follows:

$$
\begin{aligned}
&\beta_w^{s-1,\tau-1}\log\left(1-d_w^{s-1,\tau-1}+d_w^{s-1,\tau-1}e^{\frac{\beta_{w,i_w-1}^{s,\tau}-\beta_{w,i_w}^{s,\tau}}{\beta_w^{s-1,\tau-1}}\frac{a_v^{t-s}}{k\theta_v^t}}\right)\le\phi_w^{s-1,\tau-1,1} && \begin{array}{l}s=2\ldots,t\\ \tau\ge 2\end{array}\\
&\beta_w^{s-l,\tau-l}\log(1-d_w^{s-l,\tau-l}+d_w^{s-l,\tau-l}e^{\phi_w^{s-1,\tau-1,l-1}/\beta_w^{s-l,\tau-l}})\le\phi_u^{s-1,\tau-1,l} && \begin{array}{l}s=3\ldots,t\\ \tau\ge 3\\ l=2,\ldots,\min\{s-1,\tau-1\}\end{array}
\end{aligned}
\tag{55}
$$

Again, depending on the nature of the in-edges of $w$, we can end up with either the simple looking

$$
\begin{aligned}
\frac{1}{\theta_v^t}\sum_{\tau=1}^{\text{last}}a_v^{t-1}(\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}) + \frac{1}{\theta_v^t}\sum_{s=2}^{t}a_v^{t-s}(\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}) \\
+k\sum_{s=2}^{t}\sum_{\tau=s}^{\text{last}}\phi_w^{s-1,\tau-1,s-1}+k\sum_{\tau=2}^{t-1}\sum_{s=\tau+1}^{t}\phi_w^{s-1,\tau-1,\tau-1}
\end{aligned}
$$

for the case when $w$ receives pushes or the more daunting

$$
\begin{aligned}
\frac{1}{\theta_v^t}\sum_{\tau=1}^{\text{last}}a_v^{t-1}(\beta_{w,i_w-1}^{1,\tau}-\beta_{w,i_w}^{1,\tau}) + k\sum_{s=2}^{t}\log g_w^{s-1}\left(a_v^{t-s}\frac{\beta_{w,i_w-1}^{s,1}-\beta_{w,i_w}^{s,1}}{k^2\theta_v^t}\right) \\
+k\sum_{s=2}^{t}\sum_{\tau=s}^{\text{last}}\phi_w^{s-1,\tau-1,s-1}+k\sum_{\tau=2}^{t-1}\sum_{s=\tau+1}^{t}\log g_w^{s-\tau}\left(\frac{\phi_w^{s-1,\tau-1,\tau-1}}{k}\right)
\end{aligned}
$$

in the presence of arrivals. □

LEMMA 8. *Let $v\in\mathcal{Q}$ be a queue. Then $k\log\left(\mathbb{E}\exp\left(\sum_{v\sim w}p_w^{t,s}-x_v^{t-1,s-1}/k\theta\right)\right)\le 0$ has reformulation*

*a. If $t\le s$, then*

$$\sum_{v\sim w}\sum_{\tau=0}^{t-1}p_w^{t-\tau,s-\tau}\le x_v^{0,s-t}. \tag{56}$$

*b. If $t>s$, then*

*i. If the in-degree of $v$ is 1, corresponding to an arrival $\Lambda_v^t$, with moment generating function $g_v^t$, then*

$$k \log g_v^{t-s}(-1/k\theta) + \frac{1}{\theta} \sum_{v \sim w} \sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} \leq 0 \tag{57}$$

*ii. If the in-degree of $v$ consists of only a flow from server $u \in \mathbb{S}$, then*

$$\frac{1}{\theta} \sum_{v \sim w} \sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} + \bar{A}_u(p_u, \eta_u; k) + k \sum_{s'=t-s}^{last} \eta_u^{1,s'-t+s+1} \leq 0 \tag{58}$$

$$\begin{aligned}
&\beta_u^{t-s,s'} \log\left(1 - f_u^{t-s,s'} + f_u^{t-s,s'} e^{-1/k\theta}\right) \leq \eta_u^{t-s,s'} && s' \geq 1 \\
&\beta_u^{t-s-\tau,s'-\tau} \log(1 - f_u^{t-s-\tau,s'-\tau} + f_u^{t-s-\tau,s'-\tau} e^{\frac{\eta_u^{t-s-\tau+1,s'-\tau+1}}{\beta_u^{t-s-\tau,s'-\tau}}}) && \leq \eta_u^{t-s-\tau,s'-\tau} \\
& && s' \geq 2 \\
& && \tau \leq \min\{s'-1, t-s-1\}
\end{aligned} \tag{59}$$

*where*

$$\bar{A}_u(p_u, \eta_u; k) = \begin{cases} k \sum_{\tau=1}^{t-s-1} \eta_u^{t-s-\tau+1,1} & \text{if } u \text{ has only a push in-edge} \\ k \sum_{\tau=1}^{t-s-1} \log g_u^{t-s-\tau}(\eta_u^{t-s-\tau+1,1}/k) & \text{if } u \text{ has only an arrival in-edge} \end{cases} \tag{60}$$

*iii. If the in-degree of $v$ consists of only decision in-edges. Then*

$$\frac{1}{\theta} \sum_{v \sim w} \sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} + \sum_{\substack{y \sim v \\ y \neq u}} \left[ \bar{B}_y(p_y, \psi_y; k) + k \sum_{s'=t-s}^{last} \psi_y^{1,s'-t+s+1} \right] \leq 0 \tag{61}$$

$$\begin{aligned}
&\beta_y^{t-s-1,s'-1} \log\left(1 - f_y^{t-s-1,s'-1} + f_y^{t-s-1,s'-1} e^{-\frac{1}{k\theta} \frac{\beta_{y,i_y-1}^{t,s} - \beta_{y,i_y}^{t,s}}{\beta_y^{t-1,s-1}}}\right) && \leq \psi_y^{t-s-1,s'-1} \\
& && s' \geq 2 \\
&\beta_y^{t-s-\tau,s'-\tau} \log(1 - f_y^{t-s-\tau,s'-\tau} + f_y^{t-s-\tau,s'-\tau} e^{\frac{\psi_y^{t-s-\tau+1,s'-\tau+1}}{\beta_y^{t-s-\tau,s'-\tau}}}) && \leq \eta_y^{t-s-\tau,s'-\tau} \\
& && s' \geq 3 \\
& && \tau = 2, \ldots, \min\{s'-1, t-s-1\}
\end{aligned} \tag{62}$$

*where*

$$\bar{B}_y(p_y, \psi_y; k) = \begin{cases} -\frac{1}{\theta}(\beta_{y,i_y-1}^{t-s,1} - \beta_{y,i_y}^{t-s,1}) + k \sum_{\tau=2}^{t-s-1} \psi_y^{t-s-\tau+1,1} & \text{if } y \text{ has only a push in-edge} \\ k \log g_y^{t-s-1}\left( - \left(\beta_{y,i_y-1}^{t-s,1} - \beta_{y,i_y}^{t-s,1}\right)/\theta \right) & \\ \quad + k \sum_{\tau=2}^{t-s-1} \log g_y^{t-s-\tau}(\psi_y^{t-s-\tau+1,1}/k) & \text{if } y \text{ has only an arrival in-edge} \end{cases} \tag{63}$$

*Proof of Lemma 8.* Notice that $\sum_{v\sim w} p_w^{t,s} - x_v^{t-1,s-1} = -x_v^{t,s}$. Moreover,

$$x_v^{t,s} = \begin{cases} x_v^{0,s-t} - \sum_{v\sim w}\sum_{\tau=0}^{t-1} p_w^{t-\tau,s-\tau} & \text{if } s \geq t \\ x_v^{t-s,0} - \sum_{v\sim w}\sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} & \text{if } s < t \end{cases} \tag{64}$$

Hence, (a) is clear from here. For (b), the constraint reduces to $k\log\mathbb{E}\exp(-x_v^{t-s,0}/k\theta) + \frac{1}{\theta}\sum_{v\sim w}\sum_{\tau=0}^{s-1} p_w^{t-\tau,s-\tau} \leq 0$. In the case b(i), the first term simply evaluates to the moment generating function of the arrival. In the case of b(ii) and b(iii), we follow the same derivation as in Lemmas 6 and 7 respectively:

- If there is only a flow from server $u \in \mathcal{S}$ where $x_v^{t,0} = \sum_{s=1}^{\text{last}} \chi_v^{t-1,s}$, and $\chi_v^{t-1,s} \sim \text{Bin}\left(x_u^{t-1,s-1}\frac{\beta_u^{t,s}}{\beta_u^{t-1,s-1}}, f_u^{t,s}\right)$, we obtain

$$\bar{A}_u(p_u,\eta_u;k) + k\sum_{s'=t-s}^{\text{last}} \eta_u^{1,s'-t+s+1} \leq 0 \tag{65}$$

$$\begin{aligned} &\beta_u^{t-s,s'}\log\left(1 - f_u^{t-s,s'} + f_u^{t-s,s'}e^{-1/k\theta}\right) \leq \eta_u^{t-s,s'} && s' \geq 1 \\ &\beta_u^{t-s-\tau,s'-\tau}\log(1 - f_u^{t-s-\tau,s'-\tau} + f_u^{t-s-\tau,s'-\tau}e^{\frac{\eta_u^{t-s-\tau+1,s'-\tau+1}}{\beta_u^{t-s-\tau,s'-\tau}}}) && \leq \eta_u^{t-s-\tau,s'-\tau} \\ & && s' \geq 2 \\ & && \tau \leq \min\{s'-1, t-s-1\} \end{aligned} \tag{66}$$

where

$$\bar{A}_u(p_u,\eta_u;k) = \begin{cases} k\sum_{\tau=1}^{t-s-1} \eta_u^{t-s-\tau+1,1} & \text{if } u \text{ has only a push in-edge} \\ k\sum_{\tau=1}^{t-s-1} \log g_u^{t-s-\tau}(\eta_u^{t-s-\tau+1,1}/k) & \text{if } u \text{ has only an arrival in-edge} \end{cases} \tag{67}$$

- The other case is the decision in-edges. Again, we adopt dynamics as in Lemma 7, $x_v^{t,0} = \sum_{\substack{y\sim v \\ y\neq u}}\sum_{s=1}^{\text{last}} x_y^{t-1,s-1}\frac{\beta_{y,i_y-1}^{t,s} - \beta_{y,i_y}^{t,s}}{\beta_y^{t-1,s-1}}$. Mirroring the same computation, we obtain the reformulation

$$\sum_{\substack{y\sim v \\ y\neq u}}\left[\bar{B}_y(p_y,\psi_y;k) + k\sum_{s'=t-s}^{\text{last}} \psi_y^{1,s'-t+s+1}\right] \leq 0 \tag{68}$$

$$\begin{aligned} &\beta_y^{t-s-1,s'-1}\log\left(1 - f_y^{t-s-1,s'-1} + f_y^{t-s-1,s'-1}e^{-\frac{1}{k\theta}\frac{\beta_{y,i_y-1}^{t,s} - \beta_{y,i_y}^{t,s}}{\beta_y^{t-1,s-1}}}\right) && \leq \psi_y^{t-s-1,s'-1} \\ & && s' \geq 2 \\ &\beta_y^{t-s-\tau,s'-\tau}\log(1 - f_y^{t-s-\tau,s'-\tau} + f_y^{t-s-\tau,s'-\tau}e^{\frac{\psi_y^{t-s-\tau+1,s'-\tau+1}}{\beta_y^{t-s-\tau,s'-\tau}}}) && \leq \eta_y^{t-s-\tau,s'-\tau} \\ & && s' \geq 3 \\ & && \tau = 2,\ldots,\min\{s'-1, t-s-1\} \end{aligned} \tag{69}$$

where

$$\bar{B}_y(p_y,\psi_y;k)=\begin{cases}-\frac{1}{\theta}(\beta_{y,i_y-1}^{t-s,1}-\beta_{y,i_y}^{t-s,1})+k\sum_{\tau=2}^{t-s-1}\psi_y^{t-s-\tau+1,1} & \text{if } y \text{ has only a push in-edge}\\ k\log g_y^{t-s-1}\Big(-\big(\beta_{y,i_y-1}^{t-s,1}-\beta_{y,i_y}^{t-s,1}\big)/\theta\Big) & \\ \quad +k\sum_{\tau=2}^{t-s-1}\log g_y^{t-s-\tau}\big(\psi_y^{t-s-\tau+1,1}/k\big) & \text{if } y \text{ has only an arrival in-edge}\end{cases} \tag{70}$$

This completes the proof. □

## B. Duality of the P-Queue Model

We first prove Lemma 1. There is an alternate proof via proving $\left(\underset{j\in\mathcal{J}}{+}\mathcal{K}_j\right)^*=\bigcap_{j\in\mathcal{J}}\mathcal{K}_j^*$ and then using duality, but the following is more instructive as to why convexity is required.

*Proof of Lemma 1.* We first show the inclusion $\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*\subseteq\left(\bigcap_{j\in\mathcal{J}}\mathcal{K}_j\right)^*$. Let $\boldsymbol{x}^*\in\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$, then we may write $x^*=\sum_{j\in\mathcal{J}}\boldsymbol{x}^{(j)*}$ for some $\boldsymbol{x}^{(j)*}\in\mathcal{K}_j^*$. In other words, $(\boldsymbol{x}^{(j)*})^T\boldsymbol{x}\geq 0$ for all $x\in\mathcal{K}_j$. But this holds for all $j$, hence in particular, $(\boldsymbol{x}^{(j)*})^T\boldsymbol{x}\geq 0$ for all $x\in\bigcap_{j\in\mathcal{J}}\mathcal{K}_j$ for all $j$. Thus, $(\boldsymbol{x}^*)^T\boldsymbol{x}=\sum_{j\in\mathcal{J}}(\boldsymbol{x}^{(j)*})^T\boldsymbol{x}\geq 0$ for all $x\in\bigcap_{j\in\mathcal{J}}\mathcal{K}_j$. Thus, $\boldsymbol{x}^*\in\left(\bigcap_{j\in\mathcal{J}}\mathcal{K}_j\right)^*$.

For the reverse inclusion, let $\boldsymbol{x}^*\in\left(\bigcap_{j\in\mathcal{J}}\mathcal{K}_j\right)^*$. Suppose $\boldsymbol{x}^*\notin\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$. Since $\mathcal{K}_j^*$ are dual cones, hence they are closed and convex. Thus, so is $\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$. Now, we invoke separating hyperplane theorem – there exists some $\boldsymbol{y}\neq\mathbf{0}$ and real number $r$ such that $\boldsymbol{y}^T\boldsymbol{x}^*<r$ but $\boldsymbol{y}^T\boldsymbol{z}^*\geq r$ for all $\boldsymbol{z}^*\in\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$. First, notice that $r\geq 0$. Indeed, if there exists some $\boldsymbol{z}^*\in\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$ such that $\boldsymbol{y}^T\boldsymbol{z}^*=\beta<0$, then $\boldsymbol{y}^T(\alpha\boldsymbol{z}^*)$ cannot be bounded from below. Moreover, $0\in\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$, so $r\leq 0$. Hence, $r=0$.

Thus, $\boldsymbol{y}^T\boldsymbol{x}^*<0$ but $\boldsymbol{y}^T\boldsymbol{z}^*\geq 0$ for all $\boldsymbol{z}^*\in\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*$. Since $\underset{j\in\mathcal{J}}{+}\mathcal{K}_j^*\supseteq\mathcal{K}_j^*$ for all $j$, hence $\boldsymbol{y}^T\boldsymbol{z}^*\geq 0$ for all $\boldsymbol{z}^*\in\mathcal{K}_j^*$, that is, $\boldsymbol{y}\in(\mathcal{K}_j^*)^*=\mathcal{K}_j$ for all $j$. Thus, $\boldsymbol{y}\in\bigcap_{j\in\mathcal{J}}\mathcal{K}_j$. But $\boldsymbol{y}^T\boldsymbol{x}^*<0$ contradicts $\boldsymbol{x}^*\in\left(\bigcap_{j\in\mathcal{J}}\mathcal{K}_j\right)^*$. This completes the proof. □

Lemmas 2 and 1 are sufficient for writing out the dual to the P-Queue Model (16). The expressions however can be messy. For simplicity, let us illustrate it on an extremely simple example.

Suppose we only have a single server with a time non-homogeneous Poisson-distributed arrival rate $\lambda^t$. At each time, we have the decision to retain $\beta^{t,s}/\beta^{t-1,s-1}$ of jobs experiencing delay $s=0,\ldots,M$ in the server (kicking the rest out from the server permanently). The jobs retained have a $1-d^s$ chance of completion. As such, the system has dynamics $x^{t,s}=\text{Bin}\left(x^{t-1,s-1}\frac{\beta^{t,s}}{\beta^{t-1,s-1}},d^s\right)$ for $s\geq 1$ and $x^{t,0}=\text{Pois}(\lambda^t)$.

Only at time $t=T$, will the system be evaluated (it may violate constraints prior to time $T$). The server must not have exceeded capacity $\kappa$, that is, $\sum_{s=0}^{M}x^{T,s}\leq\kappa$ and the total jobs exited under

policy $\beta$ cannot exceed some cost threshold $\sum_{t=1}^{T}\sum_{s=1}^{M} x^{t-1,s-1}\left(1-\frac{\beta^{t,s}}{\beta^{t-1,s-1}}\right)c_s \le C$. The latter can be approximated (again for simplicity of illustration) to $\sum_{t=1}^{T}\sum_{s=1}^{M}(\beta^{t-1,s-1}-\beta^{t,s})c_s \le C$, or equivalently,

$$\sum_{s=1}^{M}\beta^{0,s-1}c^s+\sum_{t=2}^{T}\beta^{t-1,0}c^1+\sum_{t=1}^{T-1}\sum_{s=1}^{M-1}\beta^{t,s}(c_{s+1}-c_s)-\sum_{t=1}^{T}\beta^{t,M}c_M-\sum_{s=1}^{M-1}\beta^{T,s}c^s\le C.$$

We consider the capacity constraint in the P-Queue form $k\log\mathbb{E}\exp(\sum_{s=0}^{\text{last}} x^{T,s}/k\theta)\le \kappa/\theta$. Hence, by Theorem 2, given fixed $k$, the feasibility sub-problem is given by

$$\begin{aligned}
&k\lambda^T(e^{1/k\theta}-1)+k\sum_{s=1}^{T-1}\lambda^{T-s}(e^{\xi^{T-s+1,1}/\lambda^{T-s}}-1)+k\sum_{s=T}^{\text{last}}\xi^{1,s-T+1}\le\frac{\kappa}{\theta}\\
&\beta^{T,s}D^s\le\xi^{T,s} && s=1,\dots,M\\
&\beta^{T-\tau,s-\tau}\log(1-d^{s-\tau}+d^{s-\tau}e^{\xi^{T-\tau+1,s-\tau+1}/\beta^{T-\tau,s-\tau}})\le\xi^{T-\tau,s-\tau} && \tau=1,\dots,T-1\\ &&& s>\tau\\
&\sum_{s=1}^{M}x^{0,s-1}c^s+\sum_{t=2}^{T}\lambda^{t-1}c^1+\sum_{t=1}^{T-1}\sum_{s=1}^{M-1}\beta^{t,s}(c_{s+1}-c_s)-\sum_{t=1}^{T}\beta^{t,M}c_M-\sum_{s=1}^{M-1}\beta^{T,s}c^s\le C
\end{aligned}$$

where $\beta^{t,0}$ was set at $\lambda^t$, $\beta^{0,s}$ at $x^{0,s}$, and $D^s=\log(1-d^s+d^se^{1/k\theta})$. Under Lemma 2, this becomes

$$\begin{aligned}
&k\lambda^T(e^{1/k\theta}-1)+k\sum_{t=1}^{T-1}(Y^t-\lambda^t)+k\sum_{s=T}^{\text{last}}\xi^{1,s-T+1}\le\frac{\kappa}{\theta}\\
&\lambda^t e^{\xi^{t+1,1}/\lambda^t}\le Y^t && t=1,\dots,T-1\\
&\beta^{T,s}D^s\le\xi^{T,s} && s=1,\dots,M\\
&\left.\begin{aligned}
&(1-d^{s-\tau})y_1^{T-\tau,s-\tau}+d^{s-\tau}y_2^{T-\tau,s-\tau}\le\beta^{T-\tau,s-\tau}\\
&\beta^{T-\tau,s-\tau}\exp\left(\tfrac{-\xi^{T-\tau,s-\tau}}{\beta^{T-\tau,s-\tau}}\right)\le y_1^{T-\tau,s-\tau}\\
&\beta^{T-\tau,s-\tau}\exp\left(\tfrac{\xi^{T-\tau+1,s-\tau+1}-\xi^{T-\tau,s-\tau}}{\beta^{T-\tau,s-\tau}}\right)\le y_2^{T-\tau,s-\tau}
\end{aligned}\right\} && \begin{aligned}&\tau=1,\dots,T-1\\&s>\tau\end{aligned}\\
&\sum_{s=1}^{M}x^{0,s-1}c^s+\sum_{t=2}^{T}\lambda^{t-1}c^1+\sum_{t=1}^{T-1}\sum_{s=1}^{M-1}\beta^{t,s}(c_{s+1}-c_s)-\sum_{t=1}^{T}\beta^{t,M}c_M-\sum_{s=1}^{M-1}\beta^{T,s}c^s\le C
\end{aligned}$$

Notice in fact here that even if $\lambda$ were a decision variable, this problem is still convex and tractable in $\lambda$! For more details, refer to Remark 4. Hence, for simplicity, let $\lambda$ enter the decision variables with some overall constraint $\sum_{t=1}^{T}\lambda^t\ge\Lambda$. Then we can write this as a conic linear program:

$$\begin{aligned}
\min\quad & k\lambda^T(1-e^{1/k\theta})+k\sum_{t=1}^{T-1}(\lambda^t-Y^t)-k\sum_{s=T}^{\text{last}}\xi^{1,s-T+1} && && (71)\\
&\beta^{T,s}D^s\le\xi^{T,s} && s=1,\dots,M\\
&(1-d^{s-\tau})y_1^{T-\tau,s-\tau}+d^{s-\tau}y_2^{T-\tau,s-\tau}\le\beta^{T-\tau,s-\tau} && \begin{aligned}&\tau=1,\dots,T-1\\&s>\tau\end{aligned}\\
&\sum_{s=1}^{M}x^{0,s-1}c^s+\sum_{t=2}^{T}\lambda^{t-1}c^1+\sum_{t=1}^{T-1}\sum_{s=1}^{M-1}\beta^{t,s}(c_{s+1}-c_s)-\sum_{t=1}^{T}\beta^{t,M}c_M-\sum_{s=1}^{M-1}\beta^{T,s}c^s\le C\\
&\sum_{t=1}^{T}-\lambda^t\le-\Lambda
\end{aligned}$$

$$\begin{aligned}
&\beta^{\tau,s} \le \beta^{\tau-1,s-1} && \tau = 1,\dots,T-1\\
&(\beta^{t,s},\xi^{t,s},y_1^{t,s},y_2^{t,s},\lambda^t,Y^s) \in \mathcal{K} && s = 1,\dots,M
\end{aligned}$$

with feasibility if the optimal value is greater than $-\kappa/\theta$ and where the cone $\mathcal{K} = \left(\bigcap_{t=1}^{T-1} \mathcal{K}_{\lambda,t}\right) \cap \left(\bigcap_{\tau=1}^{T-1}\bigcap_{s=\tau+1}^{M} \mathcal{K}_{y,\tau,s}\right) \cap \left(\bigcap_{\tau=1}^{T-1}\bigcap_{s=\tau+1}^{M} \bar{\mathcal{K}}_{y,\tau,s}\right)$ is an intersection of three types of exponential cones

$$\begin{aligned}
\mathcal{K}_{\lambda,t} &= \mathrm{cl}\left\{\lambda^t e^{\xi^{t+1,1}/\lambda^t} \le Y^t\right\}\\
\mathcal{K}_{y,\tau,s} &= \mathrm{cl}\left\{\beta^{T-\tau,s-\tau}\exp\left(\frac{-\xi^{T-\tau,s-\tau}}{\beta^{T-\tau,s-\tau}}\right) \le y_1^{T-\tau,s-\tau}\right\}\\
\bar{\mathcal{K}}_{y,\tau,s} &= \mathrm{cl}\left\{\beta^{T-\tau,s-\tau}\exp\left(\frac{\xi^{T-\tau+1,s-\tau+1}-\xi^{T-\tau,s-\tau}}{\beta^{T-\tau,s-\tau}}\right) \le y_2^{T-\tau,s-\tau}\right\}.
\end{aligned}$$

These cones have duals where all other variables are 0:

$$\begin{aligned}
\mathcal{K}^*_{\lambda,t} &= \mathrm{cl}\left\{-\xi_*^{t+1,1}e^{\xi_*^{t+1,1}-\lambda_*^t/-\xi_*^{t+1,1}} \le Y_*^t\right\}\\
\mathcal{K}^*_{y,\tau,s} &= \mathrm{cl}\left\{\xi_*^{T-\tau,s-\tau}\exp\left(\frac{-\xi_*^{T-\tau,s-\tau}-\beta_*^{T-\tau,s-\tau}}{\xi_*^{T-\tau,s-\tau}}\right) \le y_{1,*}^{T-\tau,s-\tau}\right\}\\
\bar{\mathcal{K}}^*_{y,\tau,s} &= \mathrm{cl}\left\{\xi_*^{T-\tau,s-\tau}\exp\left(\frac{\beta_*^{T-\tau,s-\tau}+\xi_*^{T-\tau,s-\tau}}{\xi_*^{T-\tau,s-\tau}}\right) \le -y_{2,*}^{T-\tau,s-\tau},\ \xi_*^{T-\tau+1,s-\tau+1}+\xi_*^{T-\tau,s-\tau}=0\right\}
\end{aligned}$$

Hence (71) has dual formulation:

$$\begin{aligned}
\max\quad & \left(C-\sum_{s=1}^{M} x^{0,s-1}c^s\right)\gamma - \Lambda\mu + \sum_{s=1}^{M} x^{0,s-1}\epsilon^{1,s} && && (72)\\
& (1-d^{s-\tau})\delta^{\tau,s} + \boldsymbol{\omega}_{y_1} = 0 && \tau = 1\dots,T-1;\ s>\tau\\
& d^{s-\tau}\delta^{\tau,s} + \boldsymbol{\omega}_{y_2} = 0 && \tau = 1\dots,T-1;\ s>\tau\\
& c^1\gamma - \mu - \epsilon^{t+1,1} + \boldsymbol{\omega}_{\lambda^t} = k && t = 1,\dots,T-1\\
& -\mu + \boldsymbol{\omega}_{\lambda^T} = k\left(1-e^{1/k\theta}\right) &&\\
& \boldsymbol{\omega}_Y = -k && t = 1,\dots,T-1\\
& D^s\alpha^s - c^s\gamma + \epsilon^{T,s} + \boldsymbol{\omega}_{\beta^{T,s}} = 0 && s = 1,\dots,M\\
& -\delta^{\tau,s} + (c^{s+1}-c^s)\gamma + \epsilon^{t,s} - \epsilon^{t+1,s+1} + \boldsymbol{\omega}_{\beta^{t,s}} = 0 && t = 1,\dots,T-1;\ s = 1,\dots,M-T+t\\
& -\delta^{t,M} - c^M\gamma + \epsilon^{t,M} + \boldsymbol{\omega}_{\beta^{t,M}} = 0 && t = 1,\dots,T-1\\
& -\alpha^s + \boldsymbol{\omega}_{\xi^{\tau,s}} = 0 && s = 1,\dots,M\\
& \boldsymbol{\omega}_{\xi^{t,s}} = 0 && t = 2,\dots,T-1;\ s = 1,\dots,M-T+t\\
& \boldsymbol{\omega}_{\xi^{1,s}} = -k && s = 1,\dots,M-T+1\\
& \boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\epsilon},\gamma,\mu \le 0 &&
\end{aligned}$$

$$\begin{bmatrix} \boldsymbol{\omega}_{\boldsymbol{\beta}} \\ \boldsymbol{\omega}_{\boldsymbol{\xi}} \\ \boldsymbol{\omega}_{\boldsymbol{y}_1} \\ \boldsymbol{\omega}_{\boldsymbol{y}_2} \\ \boldsymbol{\omega}_{\boldsymbol{\lambda}} \\ \boldsymbol{\omega}_{\boldsymbol{Y}} \end{bmatrix} \in \mathcal{K}^* = \left( \bigplus_{t=1}^{T-1} \mathcal{K}^*_{\lambda,t} \right) + \left( \bigplus_{\tau=1}^{T-1} \bigplus_{s=\tau+1}^{M} \mathcal{K}^*_{y,\tau,s} \right) + \left( \bigplus_{\tau=1}^{T-1} \bigplus_{s=\tau+1}^{M} \bar{\mathcal{K}}^*_{y,\tau,s} \right).$$