Queueing with Future Information †

Joel Spencer
Courant Institute of Mathematical Sciences, NYU
New York, NY 10003
spencer@courant.nyu.edu

Madhu Sudan
Microsoft Research New England
Cambridge, MA 02139
madhu@mit.edu

Kuang Xu
Laboratory for Information and Decision Systems, MIT
Cambridge, MA 02139
kuangxu@mit.edu

ABSTRACT

We study an admissions control problem, where a queue with service rate $1-p$ receives incoming jobs at rate $\lambda \in (1-p, 1)$, and the decision maker is allowed to redirect away jobs up to a rate of p, with the objective of minimizing the time-average queue length.

We show that the amount of *information about the future* has a significant impact on system performance, in the heavy-traffic regime. When the future is unknown, the optimal average queue length diverges at rate $\sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}$, as $\lambda \to 1$. In sharp contrast, when all future arrival and service times are revealed beforehand, the optimal average queue length converges to a finite constant, $(1-p)/p$, as $\lambda \to 1$. We further show that the finite limit of $(1-p)/p$ can be achieved using only a *finite* lookahead window starting from the current time frame, whose length scales as $\mathcal{O}\left(\log \frac{1}{1-\lambda}\right)$, as $\lambda \to 1$. This leads to the conjecture of an interesting duality between queuing delay and the amount of information about the future.

1. INTRODUCTION

"Variable, but Predictable". The notion of *queues* have been used extensively as a powerful abstraction in studying dynamic resource allocation systems. Two important ingredients often make the design and analysis of a queueing system difficult: the demands and resources can be both *variable* and *unpredictable*. *Variability* refers to the fact that the arrivals of demands or the availability of resources can be highly volatile and non-uniformly distributed across the time horizon. *Unpredictability* means that such non-uniformity "tomorrow" is unknown to the decision maker "today", and she is obliged to make allocation decisions only based on the state of the system at the moment, and some statistical estimates of the future.

While the world will remain volatile as we know it, in many cases, the amount of unpredictability about the future may be reduced thanks to *forecasting* technologies and the increasing accessibility of data: future demands remain *exogenous* and variable, yet the decision maker is revealed with (some of) their realizations.

Is there significant performance gain to be harnessed by "looking into the future"? We provide a largely affirmative answer to this question, in the context of an admissions control problem.

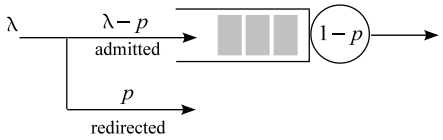

Figure 1: An illustration of the admissions control problem, with a constraint on the a rate of redirection.

2. PROBLEM FORMULATION

We consider a simple admissions control problem with a *single-server queue* running in continuous time. The system is depicted in Figure 1, and characterized by three parameters: λ, p, and w:

1. Jobs arrives to the system according to a Poisson process of rate λ, with $\lambda \in (0,1)$. The server operates at a rate of $1-p$, with $p \in (0,1)$. [1]

2. When $\lambda > 1-p$, the system is unstable, and hence some form of *admissions control* is necessary. The decision maker is allowed to decide whether an arriving job is admitted to the queue, or redirected away, subject to the constraint that the time-average rate of redirection *does not exceed* p. The goal is to minimize the resulted time-average queue length. [2]

3. The decision maker has access to *information about the future*, which takes the form of a *lookahead window* of length $w \in \mathbb{R}_+$. In particular, at any time t, the times of arrivals and service availability within the interval $[t, t+w]$ are revealed to the decision maker. We will consider the following cases of w.

† This document is an extended abstract for [1]. A manuscript is available at `http://arxiv.org/abs/1211.0618`.

[1] The server is modeled by a Poisson process of "service tokens", with rate $1-p$. If a service token is generated at time t, and if the queue is non-empty, then one job "consumes" the service token and departs from the system; otherwise, the service token is "wasted". Note that this implies the jobs are identified only by their arrival times, as opposed to job sizes. See [1] for a more detailed discussion on this modeling assumption.

[2] By Little's Law, the average queue length is essentially the same as average delay for the admitted jobs, up to a constant factor.

(a) $w = 0$, the *online problem*, where no future information is available.

(b) $w = \infty$, the *offline problem*, where entire the future has been revealed.

(c) $0 < w < \infty$, where future is revealed only up to a finite lookahead window.

One can also think of our problem as that of *resource allocation*, where a decision maker tries to match incoming demands with two types of processing resources: a *slow local resource* that corresponds to the server, and a *fast external resource* that can process any job redirected to it almost instantaneously. Both types of resources are *constrained*, in the sense that their capacities ($1-p$ and p, respectively) cannot not change over time, by physical or contractual predispositions. The processing time of a job at the fast resource is *negligible compared to that at the slow resource*, as long as the rate of redirection to the fast resource stays below p in the long run. Under this interpretation, minimizing the average delay across *all* jobs is equivalent to minimizing the average delay across just the *admitted* jobs, since the jobs redirected to the fast resource can be thought of being processed immediately and experiencing no delay at all.[3]

3. SUMMARY OF MAIN RESULTS

Our main contribution is to demonstrate that the performance of a redirection policy is highly sensitive to the amount of future information available, measured by the value of w. In particular, when sufficient amount of *future lookahead* is available, the queueing delay is *infinitely* smaller than what is achievable with an optimal online policy, in the heavy-traffic regime of $\lambda \to 1$. [4]

Throughout, we fix $p \in (0,1)$.

1. For **online policies** ($w = 0$), we show the optimal time-average queue length, C_0^{opt}, approaches infinity in the heavy-traffic regime, at the rate[5]

$$C_0^{opt} \sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}, \quad \text{as } \lambda \to 1, \tag{1}$$

and this scaling is achieved by a threshold policy.

2. In sharp contrast, the optimal average queue length among **offline policies** ($w = \infty$), C_∞^{opt}, converges to a *constant*,

$$C_\infty^{opt} \to \frac{1-p}{p}, \quad \text{as } \lambda \to 1, \tag{2}$$

and this limit is achieved by an explicit (simple) policy [1]. Figure 2 illustrates this difference in delay performance for a particular value of p, and Figure 3 demonstrates the differences via an example sample path.

3. Finally, we show that the optimal policy for the offline problem can be modified, so that the *same* optimal heavy-traffic limit of $\frac{1-p}{p}$ is achieved even with a **finite lookahead window**, $w(\lambda)$, where

$$w(\lambda) = \mathcal{O}\left(\log \frac{1}{1-\lambda}\right), \quad \text{as } \lambda \to 1. \tag{3}$$

This is of practical important, because in any realistic application only a finite amount of future information can be obtained.

REMARK 1. *As a by-product of Eq.* (2)*, observe that the heavy-traffic limit scales, in* p*, as*

$$\lim_{\lambda \to 1} C_\infty^{opt} \sim \frac{1}{p}, \quad \text{as } p \to 0. \tag{4}$$

This is consistent with an intuitive notion of "flexibility": delay should degenerate as the system's ability to redirect away jobs diminishes.

On the methodological end, we use a sample-path-based framework to analyze the performance of the offline and finite-lookahead policies, borrowing tools from renewal theory and the theory of random walks. We believe that our techniques could be extended to incorporate more general arrival and service processes, diffusion approximations, as well as observational noises.

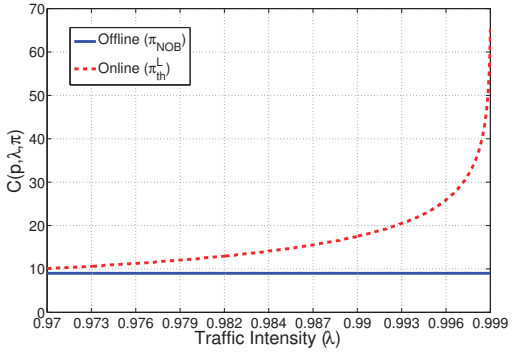

Figure 2: Comparison of optimal heavy-traffic delay scalings between online and offline policies, with $p = 0.1$ and $\lambda \to 1$. The value $C(p,\lambda,\pi)$ is the resulting average queue length as a function of p, λ, and a policy π.

3.1 Delay-Information Duality

Eq. (3) says that one can attain the same heavy-traffic delay performance as the the optimal offline algorithm, if the size of the lookahead window scales as $\mathcal{O}(\log \frac{1}{1-\lambda})$. Is this the minimum amount of future information necessary to achieve the same (or comparable) heavy-traffic delay limit as the optimal offline policy? We conjecture that this is the case, in the sense that thee exists a matching lower-bound, as follows.

CONJECTURE 1. *Fix* $p \in (0,1)$*. If* $w(\lambda) = o\left(\log \frac{1}{1-\lambda}\right)$ *as* $\lambda \to 1$*, then under any feasible policy, the time-average queue length diverges to infinity as* $\lambda \to 1$*.*

If the conjecture is proven, it would imply a *sharp transition* in the system's heavy-traffic delay scaling behavior, around the critical "threshold" of $w(\lambda) = \Theta\left(\log \frac{1}{1-\lambda}\right)$. It would also imply the existence of a symmetric dual relationship between *future information* and *queueing delay*: $\Theta\left(\log \frac{1}{1-\lambda}\right)$ amount of information is required to achieve a

[3]See [1] for a more detailed discussion on a connection of our model to resource pooling.

[4]Formal statements for the results can be found in [1].

[5]We write $f(x) \sim g(x)$ if $\lim_{x \to 1} \frac{f(x)}{g(x)} = 1$.

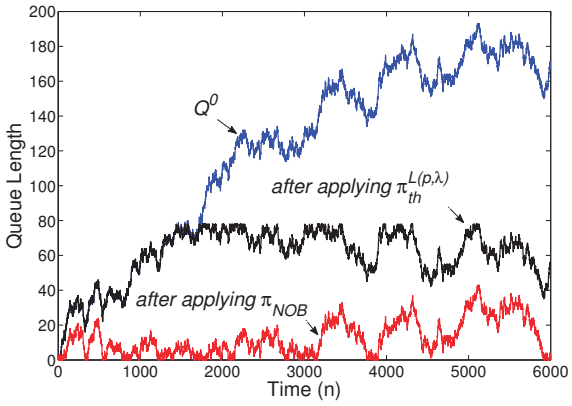

Figure 3: Example sample paths of the initial queue length process (i.e., without any deletion), Q_0, and those obtained after applying the optimal online ($\pi_{th}^{L(p,\lambda)}$) and offline (π_{NOB}) policies. ($p = 0.05$ and $\lambda = 0.999$)

finite delay limit, and one has to suffer $\Theta\left(\log \frac{1}{1-\lambda}\right)$ in delay, if only finite amount of future information is available.

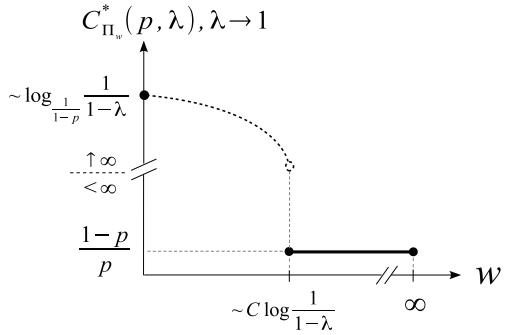

Figure 4: "Delay v.s Information." Best achievable heavy traffic delay scaling, $C^*_{\pi_w}$, as a function of the size of the lookahead window, w. Results presented in this paper are illustrated in the solid lines and circles, and the gray dotted line depicts our conjecture of the unknown regime of $0 < w(\lambda) \lesssim \log\left(\frac{1}{1-\lambda}\right)$.

Figure 4 summarizes our main results from the angle of the delay-information duality. The dotted line segment marks the unknown regime, and the sharp transition at its right end point reflects the view of Conjecture 1.

4. ACKNOWLEDGMENT

This research was partially supported by a summer research internship at Microsoft Research New England, NSF grants CMMI-0856063 and CMMI-1234062, and an MIT-Xerox Fellowship.

5. REFERENCES

[1] J. Spencer, M. Sudan and K. Xu, "Queueing with future information," under submission. arXiv:1211.0618.