Queueing Systems Vol. 100
Special Issue: 100 views on queues

Statistical Inference in Queueing Networks with Probing Information

Nelson Antunes · Gonçalo Jacinto · António Pacheco

February 25, 2022

1 Introduction

Statistical inference of the model parameters of a queueing system from data is an essential part of the study of the system. The data collection procedure specifies the type of observations from the system which are available. The estimation of the parameters can follow standard statistical procedures if we can collect all the necessary information from the system (e.g. arrival times and service times of customers) [**?**]. This explains the initial limited literature in statistical inference of queueing systems compared with performance analysis studies. For a survey on estimation in queueing systems see [**?**].

The explosion of data in the operation of large computer networks and in the management of human service systems is creating opportunities for new inference problems in queueing theory. Reflecting this new setting, there is interest in inference using partial information of the system motivated by active measurements of Internet traffic. In this context to which queueing models are applied, probe packets (say, special customers) are sent to the network and their delays are observed through the route path. The goal is to infer the statistical characteristics of the stream of packets flowing through the network. The majority of the networking literature is based on heuristic methods or assumes non-intrusive probes. Note that probes are packets, which have a minimum size and therefore perturbate the system. The analysis of a queueing system with additional input (probes) is a difficult problem [**?**]. A small number of works have provided rigorous results for classical queueing systems (e.g. M/M/1, M/G/1, G/M/1), see e.g. [**?**,**?**,**?**]. The idea consists in sending a stream of probes, according to a point process, with a given size distribution. From the arrival and departure times of the probes, the intensity

N. Antunes
CEMAT, University of Lisbon, and University of Algarve, Portugal. E-mail: nantunes@ualg.pt

G. Jacinto
CIMA/IIFA and ECT/DMat, University of Évora, Portugal. E-mail: gjcj@uevora.pt

A. Pacheco
CEMAT and IST, University of Lisbon, Portugal. E-mail: apacheco@math.tecnico.ulisboa.pt

rate and the size distribution of the original packet stream are inferred. There have been few analytical extensions with more than one queue. For instance, two $M/G/1$ queues in tandem is investigated in [**?**]. A Kelly network is used to estimate the residual capacity in each queue with probing based on the end-to-end delay of the probes, notably in [**?**]. However, in contrast to these networks, packets have constant size when they progress through a network and so service times in different queues of the network are correlated. Additionally, probes need to be sent according to a Poisson process with exponential sizes. Other probing strategies have been considered based on the arrival and departure times of probe-pairs at the different queues [**?**].

2 Problem statement

We first describe a general system without the probes. Consider a network with n queues $Q_1,\dots,Q_n$ in tandem and $n+1$ arrivals of packets streams. Each Q_i is a single-server with processing speed C_i, FIFO discipline and infinite capacity. The stream 0 enters Q_1 according to an arrival point process of intensity λ_0 and passes through all queues; stream $i=1,\dots,n$ enters Q_i according to an arrival point process of intensity λ_i and after service leaves the system. The packet sizes of stream i are i.i.d. and follow a general distribution G_i. The arrival processes of the streams to the network and their size distributions are independent and their statistical characteristics are unknown.

Probes enter the network at selected time instants with specific size and follow the path $Q_1,\dots,Q_n$. The arrival times and sizes of the probes are defined by the probing strategy and therefore known. The addition of probes should be constrainted to creating a small perturbation in the original streams. It is assumed that the system with the probes is stable. Additionally, the probe information available are the instants at which the stream probes leave Q_i and enter Q_{i+1}, $i=1,\dots,n-1$, and also when they depart from Q_n. *The problem is to define a probing strategy and an inference method to estimate the arrival point process and the distribution size of the different streams.*

From the probing information the following quantities can be observed. Let A^i_j, S^i_j, D^i_j, $j\geq 1$, denote the arrival time, service time, and departure time of the jth probe entering Q_i, respectively. If two consecutive probes, say jth and $(j+1)$th probes, share the same busy period in Q_1, then the corresponding output separation time observed between the probes, will be equal to S^1_{j+1} in case there is no stream traffic between them. Otherwise, the output separation time will contain the workload of the streams 0 and 1 that arrives to the queue during the interval (A^1_j,A^1_{j+1}). Thus, the workload in Q_1 between jth and $(j+1)$th probes is given by $W^1_j = D^1_{j+1} - D^1_j - S^1_{j+1}$.

If the arrival processes of the original streams are Poisson then the probing strategy in [**?**] can be used to estimate the arrival rate $\lambda_0+\lambda_1$ to Q_1 through $E[I]=E[e^{-(\lambda_0+\lambda_1)T}]$, where I is a random variable which is equal to 1 if no workload arrives between two probes and 0 otherwise, and T is the inter-arrival time between the probes. The moments of packet sizes could also be estimated assuming that streams 0 and 1 have the same distribution (see [**?**], Eqs. (7)-(10)). In this work, the probe sizes are general and their arrival times to Q_1 follow a renewal point process. The rate λ_0 could also be estimated in a cumbersome way through the comparison of the workloads W_j^1 and W_j^2. To extend this probing strategy to estimate the streams characteristics in the other queues would be very inefficient. For instance, for stream 2 we will need to observe consecutive probe pairs $(j, j+1)$ in $Q_1$with $W_j^1 = S_{j+1}^1$. In this case the workload between the probes W_j^2 in Q_2 will be only from stream 2, and if they share the same busy period in Q_2, the same approach described above can be used to estimate the stream characteristics. However, the number of probes under these two conditions will be reduced as they progress through the network and at the expense of a prolonged observation time. Alternatively, probes could arrive in batch to increase the chances to be in the same busy period at a queue, but this would increase the perturbation in the system. On the other hand, the estimation procedure uses the method of moments characterized by its simplicity and resulting in estimators which are consistent but biased. Also other desired properties of the estimators (e.g., asymptotic normality, efficiency) are difficult to establish or do not exist.

3 Discusssion

The inference of the arrival processes and size distributions of different streams in tandem queues with probes is a difficult and open problem. A probing strategy framework allowing to estimate the stream characteristics in the different queues should be designed. It is possible to consider several probing phases. Other route paths of the streams can also be considered. In order to avoid the problems associated with the method of moments, other estimation procedures should be investigated based on the method of maximum likelihood, Expectation-Maximization algorithm or Bayes method. Addressing non-Poisson arrival processes for packet streams and non-parametric size distributions are also challenging problems. Good properties of the estimators, such as efficiency and asymptotic normality, are desired. Another direction is the study of optimal probing strategies with minimum variance of the estimators. Finally, the consideration of piecewise-stationary arrival processes for the streams is also a possible direction [**?**].