Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

NEW APPROACHES FOR INFERENCE OF UNOBSERVABLE QUEUES

Yun Bae KIM
Jinsoo PARK

Dept. of Systems Management Engineering
Sungkyunkwan University
Cheon-cheon-dong 300, Jang-an-gu
Suwon, KOREA

ABSTRACT

Many inference methods of queueing systems have been developed on the basis of Larson's QIE(queue inference engine) with the assumption of homogeneous Poisson arrivals. It inferred the queueing systems with starting and ending times of service. However, the arrival processes are becoming complex lately, so there are some limits to apply the method. Our study introduces new methods of queue inference which can find the internal behaviors of queueing systems with only external observations, arrival and departure time.

This study deals with general *GI*/*G*/*c* queueing systems.

(a) FCFC (first come first served)
(b) LCFS (last come first served)
(c) RSS (random selection for service)

The accurate inferences were obtained from FCFS and LCFS systems, and the approximate solutions from RSS systems.

1 INTRODUCTION

This paper suggests new methods that infer the queueing systems. One might still be able to observe, from outside the system, the arrival and departure times of customers. Given each customer's arrival and departure times, if one knows the number of servers, then one can exactly compute both time in queue and time in service of every customer irrelevant of service discipline, FCFS or LCFS. For the RSS discipline, one can approximately calculate them. If these same assumptions hold except that one does not know the number of servers, one can accurately estimate the number of servers in addition to the queueing and service times. Our only assumption is that the service times are independent.

This paper is in the tradition of queue inference studies. Much of that work derives from R. C. Larson's (1990) seminal paper on the Queue Inference Engine, which assumed that the only available data are the times at which service starts and stops. Our assumptions are complementary, since Larson assumes only the internal operation of the service facility is visible, while we assume system arrivals and departures are visible but both the queue and the internal operations of the service facility are invisible. Since Larson's original work, a number of related papers have appeared which improve computing times and/or generalize assumptions (Bertimas and Servi 1992, Daley and Servi 1992, Basawa et al. 1996, Bingham and Pitts 1999, Jang et al. 2001). There are also a number of other papers dealing with some aspects of queue inference (Masuda 1995, Dimitrijevic 1996, Pickands and Stine 1997, Toyozumi 1997, Mandelbaum and Zeltyn 1998, Jones 1999, Coolen 2003). To our knowledge, the work presented here is the first to draw exact inferences about queueing and service times from arrival and departure times when the number of servers is unknown. There is some previous work on estimating the number of servers in queueing systems.

The organization of this paper follows. In section 2, we develop inference for three service disciplines, FCFS, LCFS and RSS. At first we introduce the method for the case of known number of servers. We expand this method to the case of unknown number of servers. In section 3, we present simulations validating the inference methods. In section 4, we summarize our results.

2 METHODOLOGY

2.1 Purpose of Inference

The final goal is to find the waiting time and the service time of each customer. From this result we calculate the mean and variance of waiting time and service time. We use only arrival times and departure times of customers. Let us define some notations.

N : number of customers
A_i : arrival time of customer i

D_i : departure time of customer i
B_i : service starting time of customer i
S_i : service time of customer i
Q_i : waiting time of customer i
T_i : system sojourn time of customer i
c : number of servers

Observable terms are N, A_i's and D_i's. If we know the number of servers c, we can find the exact B_i's from the process of number of customers in system. We expand the method to the case of unknown number of servers. To identify the service stating times, we use the concept of congestion period below.

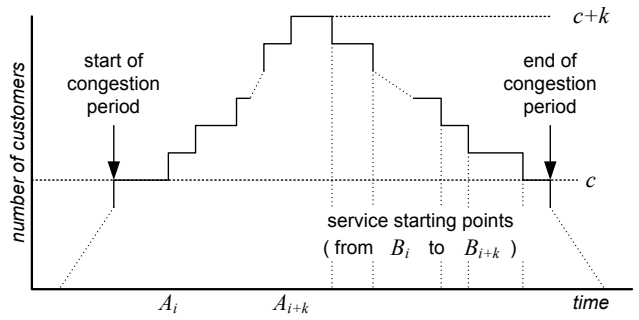

Figure 1: Congestion period and service starting times

Before congestion period, the service starting time of each customer is equal to arrival time. After that service starting time is determined by departure time. A departure means a service is over and a new service starts. Now, we have to allocate these service starting times to customers.

2.2 Allocating the Service Stating Times

2.2.1 FCFS Systems

In the FCFS system, we can directly determine the service starting time with arrival order. If we denote $D_{(i)}$ be the i-th order statistic of D_i's, the service starting times are determined by equation (1) in the congestion period

$$B_i = D_{(i-c)}. \tag{1}$$

Obviously, in the non-congestion period, service starting times are same with arrival times.

2.2.2 LCFS Systems

In the LCFS systems, more complex concept is required. In the congestion period, the service starting time of a customer who sees n customers is the first departure time that remains n customers. Figure 2 shows the mechanics of LCFS. If one (grayed customer in Figure 2) sees n customers upon arrival, c customers are being served and n-c customers are in the queue. As one joins the queue, the queue size becomes n-c+1, which is (a). There will be two situations follow. One is that an arrival is faster than all service completions; the other is that a service completion arises first. (b) and (c) in the Figure 2 represent the cases, respectively. The system eventually reaches (c) whether it takes the state of (b) or not. If the system reaches (c), the number of customers becomes n again and one starts his service at that time.

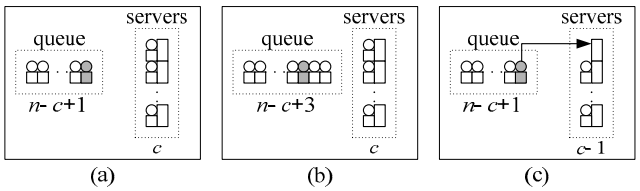

Figure 2: Service starting time of LSFS customer

To get the mathematical form of this method, let us define the next.

N_i^A : number of customers that i sees at his arrival

N_i^D : number of customers that i sees at his departure

$D^1(n,\ t)$: departure time when $N_i^D = n$ after time t

Now, we can get equation (2)

$$B_i = \begin{cases} A_i & (N_i^A < c) \\ D^1(N_i^A, A_i) & (o/w). \end{cases} \tag{2}$$

The above term of equation (2) is the case of non-congestion period and the below is congestion period.

2.2.3 RSS Systems

In the RSS systems, it is impossible to allocate the exact service starting times. We can only observe the characteristic of the queueing systems; probably customers leave the system in the sequential order of their service begun. From this fact, we can estimate the order of service approximately. The service order may follow with departure order. Therefore we can reconstruct the simulation with arrival and departure times. Figure 3 shows the simulation reconstruction algorithm. The action of "Get a customer from the queue who has the smallest departure time point" is processing in event 'Departure'.

From this algorithm, we can approximately allocate the service starting times. We can exactly calculate the mean queueing time and mean service time from this result. Besides, we can calculate approximate variance of them. If the number of servers is smaller, the variance is more close to real one. If the inherent variance of service time is small,

then we get the same result. Other than two cases of mentioned we need to develop a different method.

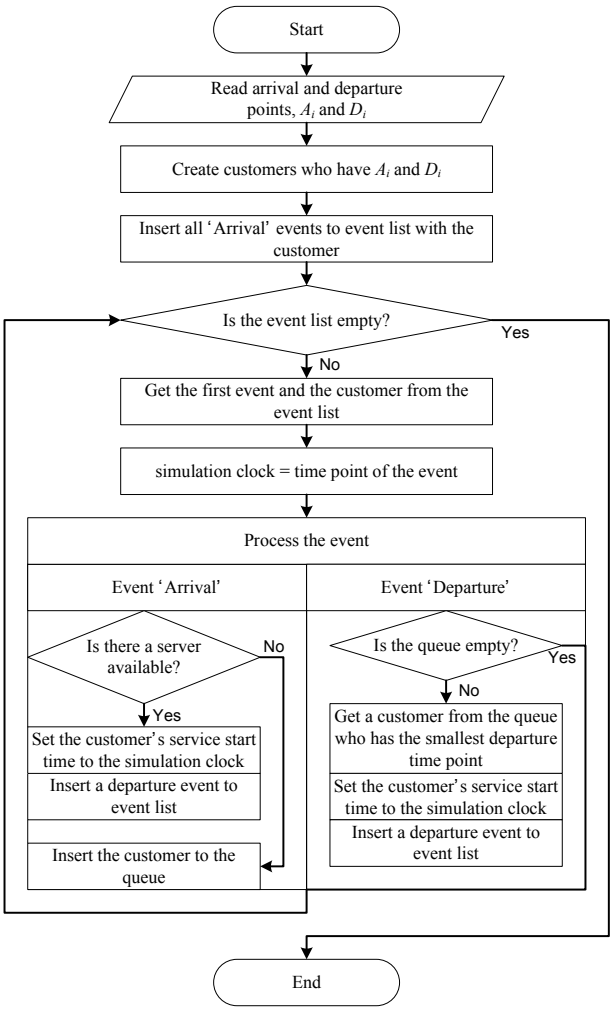

Figure 3: Simulation reconstruction algorithm

2.2.4 Expanded Method for RSS System

If the number of servers is large or the real variance of service times is large, the estimated variance of service times has large error. In these cases, we use the independence of queueing time and service time to estimate the variance. Theoretically, if these times are independent, the covariance of them is zero. We are expecting a successful allocation of the service starting times that result in zero covariance. As a result, we construct an optimization problem with covariance of them as objective function.

To make the mathematical form of optimization problem, we denote $C(i, r)$ as the optional r-th departure time in the congestion period that contains customer i. Now we have the optimization model.

$$
\begin{aligned}
\textit{minimize} \quad & \left| \sum_{i=i}^{N} \left(\hat{Q}_i - \overline{\hat{Q}} \right) \left(\hat{S}_i - \overline{\hat{S}} \right) \right| \\
\textit{s.t.} \quad & \hat{B}_i = A_i \qquad (N_i^A < c) \\
& \hat{B}_i = C(i,r) \qquad (o/w) \\
& A_i \le \hat{B}_i < D_i \\
& \hat{Q}_i = \hat{B}_i - A_i \\
& \hat{S}_i = D_i - \hat{B}_i
\end{aligned}
$$

$\overline{\hat{Q}}$ and $\overline{\hat{S}}$ imply the sample means of $\hat{Q}_i$ and $\hat{S}_i$. N_i^A is that of section 2.2.2.

2.3 The Cases of Unknown Number of Servers

In the case of unknown number of servers, we cannot allocate the service starting times directly. The service starting times depend on the number of servers. For any randomly selected value of c, we can get only estimated service starting times. After we have estimated service starting times, we can calculate the service times. Using properties of service times, we can find the real number of servers. Actually, if the number of servers is estimated wrong, the estimated service times can be represented as $\hat{S}_i = S_i + \Delta_i$, where $\hat{S}_i$ is the estimated service time and S_i is the real service time. If we know the real number of servers c, we can get the real service time S_i; but if we estimate the number of servers wrong, we get the estimated service time as $\hat{S}_i$. We can show that S_i and Δ_i are independent. The proof of independency is provided in Park (2007). For any cases, the optimization problem successfully finds the real number of servers, c, which minimizes the sample variance of service times.

2.3.1 FCFS systems

In the FCFS systems, if we unite both case of congestion and non-congestion period, we get next term.

$$B_i = \max\left\{A_i, D_{(i-c)}\right\} \tag{3}$$

From equation (3) we can get the next optimization form.

$$
\begin{aligned}
\textit{minimize} \quad & \sum_{i=1}^{N} \left(\hat{S}_i - \overline{S} \right)^2 \\
\textit{s.t.} \quad & \hat{B}_i = A_i, \qquad i \le \hat{c} \\
& \hat{B}_i = \max\{A_i, D_{(i-\hat{c})}\}, \qquad i > \hat{c} \\
& \hat{S}_i = D_i - \hat{B}_i \\
& \hat{S}_i > 0
\end{aligned}
$$

Hat notations mean the estimated values and barred notations imply their sample means. Grid search over the integers provides the estimated number of servers.

2.3.2 LCFS systems

We can construct the optimization model with same objective function and equation (2).

$$\begin{aligned} &\textit{minimize} && \sum_{i=1}^{N}\left(\hat{S}_i-\overline{S}\right)^2 \\ &\textit{s.t.} && \hat{B}_i=\begin{cases} A_i & (N_i^A<\hat{c}) \\ D^1(N_i^A,A_i) & (o/w) \end{cases} \\ & && \hat{S}_i=D_i-\hat{B}_i \\ & && \hat{S}_i>0 \end{aligned}$$

2.3.3 RSS Systems

In the RSS System, we cannot allocate the service starting times exactly. So we use the result from simulation reconstruction for constraint term. Let us denote $SR_{\hat{c}}(i)$ as service starting time of customer i from simulation reconstruction under number of servers is $\hat{c}$. Now we get

$$\begin{aligned} &\textit{minimize} && \sum_{i=1}^{N}\left(\hat{S}_i-\overline{S}\right)^2 \\ &\textit{s.t.} && \hat{B}_i=SR_{\hat{c}}(i) \\ & && \hat{S}_i=D_i-\hat{B}_i \\ & && \hat{S}_i>0\,. \end{aligned}$$

This problem provides the real number of servers too. We will confirm this fact at Section 3.

3 SIMULATIONS AND INFERENCES

3.1 Simulation Models

We simulated general *GI*/*G*/*c* systems which have general inter-arrival time distribution, general service time distribution and the number of servers is *c*. We used deterministic, exponential and normal distribution as inter-arrival and service time distribution, and tried 1, 3 and 7 as the number of servers. We also varied the server utilization 0.8, 0.9 and 1.0. Three different service disciplines (FCFS, LCFS, and RSS) are tested. Therefore the number of combinations we simulated is 243(3 arrival types×3 service types×3 number of servers×3 server utilizations×3 service disciplines). Table 1 shows the mean service time as system parameters. We fix the arrival rate to 1 and change the mean service time by server utilization ρ.

Table 1: Mean service times

c	mean service time		
	ρ=0.8	ρ=0.9	ρ=0.1
1	0.8	0.9	1
3	2.4	2.7	3
7	5.6	6.3	7

We collected 1,000 arrival and departure times from each simulation to infer the system of interest. We also collected 1,000 service starting times from each simulation for true values. We only used the arrival and departure times to find service starting times.

3.2 Inference Results

We inferred the service starting times for the situation of known and unknown number of servers. We calculated mean queueing time, mean service time, variance of queueing times and variance of service times from inferences.

3.2.1 Case of Known Number of Servers

In this case, we get exact solution for FCFS and LCFS systems. So we omit the results of these service disciplines. However, in the RSS systems, we get approximate solutions. Table 2 shows some results of RSS systems.

The optimization model in section 2.2.4 provided the results. $Var(Q)$ means true variance of queueing times and $Var(\hat{Q})$ is estimated variance of queueing times. R.E. means relative error and the values are calculated down to four decimal places. In the left most column of Table 2, we use a normal distribution for *GI*(or *G*). All average values are same to actual values.

Table 2: Variance of queueing times for the RSS systems

System	ρ	$Var(Q)$	$Var(\hat{Q})$	R.E.
M/*M*/*7*	0.8	128.74	131.10	0.0183
	0.9	101.52	115.16	0.1344
	1.0	423.27	439.36	0.0380
D/*M*/*7*	0.8	5.60	6.64	0.1873
	0.9	124.64	133.04	0.0674
	1.0	81.87	89.37	0.0916
GI/*M*/*7*	0.8	0.77	0.81	0.0521
	0.9	5.69	6.11	0.0730
	1.0	16.19	18.00	0.1122

The small gaps between $Var(Q)$ and $Var(\hat{Q})$ are due to the statistical noise coming from sampling. The last column of the Table 2 (R.E.) can be ignored if we repeat the replications in simulation experiments.

3.2.2 Case of Unknown Number of Servers

In this section, we deal with the case of unknown number of servers. We want to show the uni-modality of variance of service times. That is, we want to show that the optimization models in section 2.3 work correctly.

Figure 4 is the results of some FCFS systems. In these systems, the number of servers is 7 and server utilization is 0.8. The rest of the systems which have higher utilization, another number of servers, or different arrival and service distributions show similar shape on this plot. The horizontal axis means the number of servers and the vertical axis implies the variance of service times according to the number of servers. The uncharted area represents infeasible solution area.

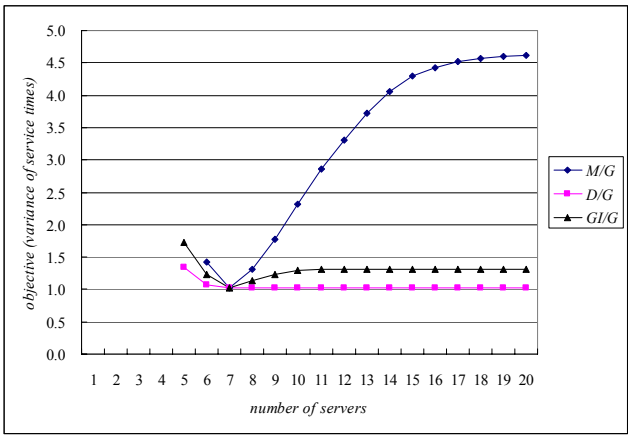

Figure 4: Example results of FCFS systems

See Figure 5 to obtain the case of LCFS discipline. All of the system condition is same to FCFS previous results (Figure 4). With same condition, Figure 6 shows that case of RSS works.

The fact that we can realize is the convergence of objective, the variance of service times. If we overestimate the number of servers larger and larger, the inference understands that all the customers start their service as they arrive. Therefore, the service times converge to system sojourn times. This fact provides the upper limit of the number of servers. Needless to say, the rest of those systems show similar results.

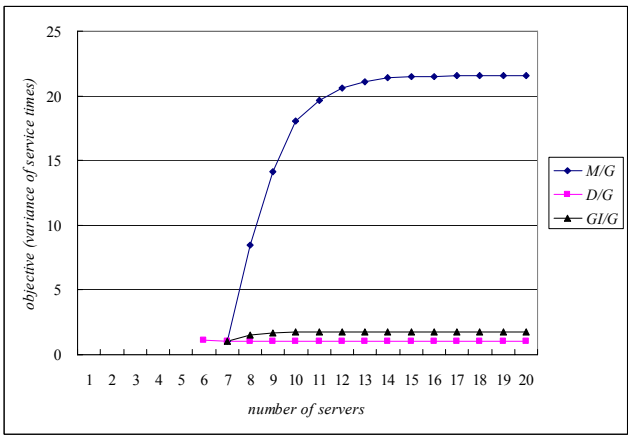

Figure 5: Example results of LCFS systems

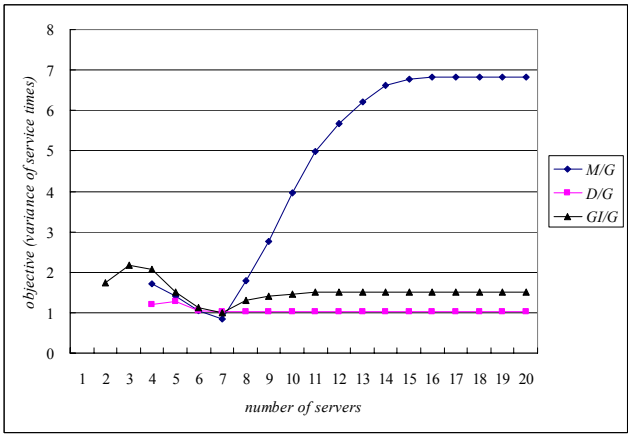

Figure 6: Example results of RSS systems

4 SUMMARY

We inferred the queueing systems using arrival and departure points. The only assumption of our research is independence of service times. We find internal behaviors of FCFS, LCFS and RSS systems. At first, we proposed inference methods when the number of servers is known. And we expended to the case of unknown number of servers. We inferred the accurate solution from FCFS and LCFS systems, and approximate solution from RSS systems.

REFERENCES

Basawa, I., U. Bhat, and R. Lund. 1996. Maximum likelihood estimation for single server queues from waiting time data. *Queueing Systems* 24: 155-167.

Bertimas, D. and L. Servi. 1992. Deducing queueing from transactional data: The queue inference engine, revisited. *Operations Research* 40: 217-228.

Bingham, N. and S. Pitts. 1999. Non-parametric estimation for the M/G/infinity queue. *Annals of the Institute of Statistical Mathematics* 51: 71-97.

Coolen, F., P. Coolen-Schrijner. 2003. A nonparametric predictive method for queues. *European Journal of Operational Research* 145, 425–442.

Daley, D. and L. Servi. 1992. Exploiting Markov-chains to infer queue length from transactional data. *Journal of Applied Probability* 29, 713-732.

Dimitrijevic, D. 1996. Inferring most likely queue length from transactional data. *Operations Research Letters* 19: 191-199.

Jang, J., J. Suh and C. Liu. 2001. A new procedure to estimate waiting time in GI/G/2 systems by server observation. *Computers and Operations Research* 28: 597-611.

Jones, L. 1999. Inferring balking behavior from transactional data. *Operations Research* 47: 778-784.

Larson, R. 1990. The queue inference engine: Deducing queue statistics from transactional data *Management Science* 36: 586-601.

Mandelbaum, A. and S. Zeltyn. 1998. Estimating characteristics of queueing networks using transactional data. *Queueing Systems* 29: 75-127.

Masuda, Y. 1995. Exploiting partial information in queueing-systems *Operations Research* 43: 530-538.

Park, J. 2007. A Study of Inference Methods for Queueing Systems with Arrival and Departure Time Data. unpublished Ph. D. thesis. Sungkyunkwan University, Suwon, Korea.

Pickands, J. and R. Stine. 1997. Estimation for the M/G/infinite queue with incomplete information. *Biometrika* 84: 295-308.

Toyoizumi, H. 1997. Sengupta's invariant relationship and its application to waiting time inference. *Journal of Applied Probability* 34: 795-799.

AUTHOR BIOGRAPHIES

YUN BAE KIM is a Professor at the Sungkyunkwan university, Suwon, Korea. He received master degree from University of Florida and Ph.D degree from Rensselaer Polytechnic Institute. His current research interests are simulation methodology, agent based simulation and simulation based acquisition.

JINSOO PARK holds a postdoctoral position in the Department of Systems Management Engineering at the Sungkyunkwan university. He received a master's degree in industrial engineering and a Ph. D. in the same subject from the Sungkyunkwan university, Seoul, Korea. His current research interests are in queue inference, analysis of queueing systems and agent based modeling and simulation.