Stationary workload of two non-work-conserving M/G/1 Preemptive LIFO Queues

Jacob Bergquist, Karl Sigman

IEOR 8100 Queueing Seminar, Spring 2020

Table of Contents

Introduction and Motivation

- We consider two variations of the M/G/1 preemptive LIFO queue which are non-work-conserving: customers do not retain their remaining service time when they are preempted.
  - Preemptive-repeat different (PRD): preempted customers are put back in the front of the line and assigned a new i.i.d. service time.
  - Preemptive-repeat identical (PRI): preempted customers are put back in the front of the line and retain their original service time.
- Studied recently in a paper due to Asmussen and Glynn via branching processes, Galton-Watson family trees, and stochastic fixed point equations.

Basic M/G/1 set-up

- Poisson point process of customer arrival times $\{t_n : n \geq 1\}$ at rate λ.
- Independent from the arrival process, each customer brings i.i.d non-negative service times $\{S_n : n \geq 1\}$ distributed according to some $G(x) = P(S \leq x)$, with $0 < E[S] = 1/\mu < \infty$.
- The workload process $\{V(t) : t \geq 0\}$ is defined as the sum of the service times of all customers in the queue plus the remaining service time of any customer in service.
- The times at which an arrival finds the system empty, $V(t_n-) = 0$, serve as regeneration points with i.i.d. cycles. So long as the cycle length distribution is proper and has finite first moment, we are ensured the existence of a limiting, stationary distribution of workload, which we denote by V.

Table of Contents

Warm-up: an always-stable PRD queue

Take the service time distribution $S = .99\delta_0 + .01\exp(.01)$, so the time-stationary remaining service time of a customer found in service is $\hat{S}_r = \exp(.01)$.

Then we will turn out to have $p_0 = \frac{1-2\lambda+\lambda(100)}{1-\lambda+\lambda(100)} = \frac{1+98\lambda}{1+99\lambda}$, so stability is ensured for all λ and as $\lambda \to \infty$, p_0 decreases monotonically to $98/99$.

PRD queue - a nice remark for the exponential service time case

Suppose we take exponential service times. Then in both the work-conserving PL and the non-work-conserving PRD disciplines, the distribution of the service times of the customers returned to the line are the same - exponential! Hence, for each fixed t, the distributions of workload $V(t)$ in the two models coincide, though the stochastic processes $\{V(t) : t \geq 0\}$ do not.

PRD model results

Lemma (PRD Idle Time)

For a stable M/G/1 PRD model,

$$p_0 \stackrel{d}{=} P(V=0) = \frac{2E[e^{-\lambda S}]-1}{E[e^{-\lambda S}]}$$

Proof idea: Note that $P(S < T) = E[e^{-\lambda S}]$. Let S_r denote time-stationary remaining service time of the customer in service. Let $\hat{S}_r = (S_r \mid S_r > 0)$. Then

$$\hat{S}_r \stackrel{d}{=} (S - T \mid S > T)$$

This allows us to compute $E[\hat{S}_r] = -1 + \frac{\rho}{1-E(e^{-\lambda S})}$.
Now apply the Rate Conservation Law to $\{V(t)\}$, and you too can solve for p_0.

PRD model results

Lemma (PRD Stability Condition)

An M/G/1 PRD queue is stable if and only if

$$E[e^{-\lambda S}] > 1/2$$

Alternative proof: Construct an equivalent FIFO M/G/1 queue and show stability. $\tau = \min\{n \geq 1 : S_n < T_n\}$ is (in distribution) the total number of times that a customer enters service. It is a geom($P(S < T)$) random variable. Thus we can define an effective service time

$$\mathbb{S} \stackrel{d}{=} (S|S < T) + \sum_{j=1}^{\tau-1}(T_j|T_j < S_j)$$

The stability condition for this queue is $\lambda E[\mathbb{S}] < 1$, and carrying out the requisite computation leads to the exact condition as above.

PRD model results

Theorem (PRD Stationary Workload)

For the stable M/G/1 PRD model, let Q be a geometric random variable with mass at 0 and parameter $p_0 = \frac{2E[e^{-\lambda S}]-1}{E[e^{-\lambda S}]}$. *Then we have*

$$(V|V>0) \stackrel{d}{=} \hat{S}_r + \sum_{j=1}^{Q} S_j$$

and thus

$$F_V = p_0\delta_0 + (1-p_0)F_{\hat{V}}$$

Stationary workload proof idea

Proof idea: The same classical approach applied by Fakinos in 81 to the work-conserving model works here. Let N denote the time-stationary number-in-system. Then $P(N \geq 1) = 1 - p_0$. Now the event $\{N \geq 2\}$ means that the arriving customer C_0 finds a customer C_{-k} in system, and customer C_{-k} found a customer as well. But these events are independent because of the PL discipline, and hence $P(N \geq 2) = P(N \geq 1)^2 = (1 - p_0)^2$.

PRD heavy-traffic limit

Theorem (PRD Heavy-traffic limit)

Let $\lambda_2 > \lambda$ be the solution to $E[e^{-\lambda_2 S}] = 1/2$. Then as $\lambda \uparrow \lambda_2$,

$$p_0 V \longrightarrow exp(\mu),$$

where $\mu^{-1} = E(S)$.

Proof sketch: We write

$$p_0(\lambda)\hat{V}(\lambda) \stackrel{d}{=} p_0(\lambda)S_r(\lambda) + p_0(\lambda)\sum_{i=1}^{N(\lambda)} S_i$$

First term goes to 0 by Markov's, $p_0(\lambda)N(\lambda) \longrightarrow \text{exp}(1)$ is standard, and WLLN gives $1/N(\lambda)\sum_{i=1}^{N(\lambda)} S_i \xrightarrow{p} E[S]$.

PRD average sojourn time

Using Little's law, we now have all the ingredients to immediately arrive at the average sojourn time. It aligns with that found in Asmussen and Glynn's work, modulo a small error.

$$W = \frac{1}{\lambda}L = \frac{1}{\lambda}E[\text{geom}(\frac{2E[e^{-\lambda S}]-1}{E[e^{-\lambda S}]})]$$

$$= \frac{1}{\lambda}\left[-1 + \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S})-1}\right]$$

Average number of customers served in a busy period

In all preemptive LIFO models, the distribution of sojourn time is identical to the distribution of the length of a busy period, so the expression for the average sojourn time is exactly the expected length of a busy period. Since by PASTA p_0 is the long-run proportion of customers who find the system empty, we get by regenerative process theory that

$$p_0 = \frac{1}{E[N_B]}$$

where N_B is the number of customers served during a busy period. Hence we can compute

$$E[N_B] = \frac{E[e^{-\lambda S}]}{2E[e^{-\lambda S}] - 1}$$

Table of Contents

PRI Model

Here we can assume WLOG $P(S > 0) = 1$: customers with service times of 0 are never seen, they pass right through and do not impact the model.
Thus if $P(S = 0) > 0$, we replace λ with $\lambda P(S > 0)$ and G with the conditional distribution of $(S|S > 0)$, reducing us to the case where as assume S is almost surely positive.

PRI Model

The analysis here is more difficult: though $\hat{S}_r$ is of the form $(S - T | T < S)$, the S here has a biased distribution. The more times it has been preempted, the more biased it becomes).
We will let $\hat{B}$ denote the stationary age of the service found when preempted, and summing this with $\hat{S}_r$ will give the whole service time of the customer found in service when preempted.

$$\hat{S} = \hat{B} + \hat{S}_r$$

In stationarity, it is this $\hat{S}$ which is placed back into the queue when a customer is preempted.

Applying the RCL to PRI

When we apply the RCL to the workload process, we immediately arrive at the equation

$$1 - p_0 = \rho + \lambda(1 - p_0)E[\hat{B}]$$

And thus to solve for p_0, it remains only to solve for $E[\hat{B}]$.

The emergence of the MGF

Lemma (Distribution of the number of preemptions)

For a fixed service time S, let τ denote the number of times it enters service before completion. Conditional on S, τ is a geometric random variable with success probability $e^{-\lambda S}$, and hence $E[\tau] = E[e^{\lambda S}]$

Proof sketch: Let $\{T_n : n \geq 1\}$ denote i.i.d. exponential random variables at rate λ. Then we have

$$P(\tau = 1) = P(S \leq T_1) = E[e^{-\lambda S}]$$

$$P(\tau = n) = P(S > T_1, \ldots, S > T_{n-1}, S \leq T_n)$$

Hence, conditioning on S furnishes

$$P(\tau = n|S) = (1 - e^{-\lambda S})^{n-1} e^{-\lambda S}$$

Finding $\hat{B}$

We can take each customer alone with their independent service time S and their own i.i.d. sequence of interarrival times $\{T_n\}$ and sum up their ages when preempted as if they form one regenerative cycle which ends at interarrival time $K = \tau - 1$. We can thus write, for any non-negative measurable function f,

$$E[f(\hat{B})] = \frac{E[\sum_{j=1}^{K} f(T_j)]}{E[K]}$$

Finding $\hat{B}$

K is not a stopping time - but $\tau = K + 1$ is. Thus we can compute the numerator by using Wald's and subtracting the last piece

$$E[\sum_{j=1}^{K} f(T_j)] = E[\tau]E[f(T)] - E[f(T_\tau)]$$

Letting $f(b) = b$,

$$E[\hat{B}] = \frac{\frac{1}{\lambda}E[e^{\lambda S}] - E[T_\tau]}{E[e^{\lambda S}] - 1}$$

And after computing $E[T_\tau] = \frac{1}{\mu} + \frac{1}{\lambda}$, we can arrive at the stability condition.

PRI model results

Lemma (PRI Stability Condition)

The PRI model is stable if and only if $E[e^{\lambda S}] < 2$.

Lemma (PRI Idle Time)

For a stable M/G/1 PRI model,

$$p_0 = 2 - E[e^{\lambda S}]$$

Since the expected number of times a customer is preempted is $E[e^{\lambda S}] - 1$, we can interpret the stability condition as: the PRI model is stable if and only if the expected number of times a customer is preempted is strictly less than 1.

CDF of $\hat{B}$

Lemma

For the M/G/1 PRI model, the CDF of $\hat{B}$ is as follows:

$$F_{\hat{B}}(x) = \frac{E[e^{\lambda S}] - e^{-\lambda x} E[e^{\lambda S}; S > x] - G(x)}{E[e^{\lambda S}] - 1}$$

and $\hat{B}$ is in fact always a continuous random variable, with density

$$f_{\hat{B}}(x) = \frac{\lambda e^{-\lambda x} E[e^{\lambda S}; S > x]}{E[e^{\lambda S}] - 1}$$

CDF of $\hat{S}$

Lemma

For the M/G/1 PRI model, the CDF of $\hat{S}$ is as follows:

$$F_{\hat{S}}(x) = \frac{E[e^{\lambda S}; S \leq x] - G(x)}{E[e^{\lambda S}] - 1}$$

and if S has a density, then so does $\hat{S}$ and it is given by

$$f_{\hat{S}}(x) = \frac{g(x)e^{\lambda x} - 1}{E[e^{\lambda S}] - 1}$$

CDF of $\hat{S}_r$

Lemma

The cumulative distribution function (cdf) of $\hat{S}_r$, $F_{\hat{S}_r}(x) = P(\hat{S}_r \le x)$ is given by

$$F_{\hat{S}_r}(x) = \frac{e^{\lambda x}\bar{G}(x) + E(e^{\lambda S}; S \le x) - 1}{E(e^{\lambda S}) - 1}, x \ge 0.$$

$\hat{S}_r$ always has a density (it is alway a continuous r.v.) and it is given by

$$f_{\hat{S}_r}(x) = \frac{\lambda e^{\lambda x}\bar{G}(x)}{E(e^{\lambda S}) - 1}, x \ge 0.$$

Joint CDF of $\hat{B}$, $\hat{S}_r$

Lemma

For the M/G/1 PRI model,

$$P(\hat{B} > x,\ \hat{S}_r > y) =$$

$$\frac{e^{-\lambda x} E(e^{\lambda S}; S > x + y) - e^{\lambda y} P(S > x + y)}{E(e^{\lambda S}) - 1},\ x \geq 0, y \geq 0$$

Thus, if G has a density g, then the joint density of $(\hat{B}, \hat{S}_r)$ exists and is given by

$$f_{(\hat{B},\hat{S}_r)}(x, y) = \frac{\lambda e^{\lambda y} g(x + y)}{E(e^{\lambda S}) - 1},\ x \geq 0, y \geq 0.$$

Relationship between $\hat{S}_r$ and $\hat{B}$

We can immediately deduce that
$P(\hat{S}_r > x) = e^{\lambda x} P(\hat{B} > x)$, from which we can conclude

- $\hat{S}_r$ is stochastically larger than $\hat{B}$.
- If S has unbounded support, then

$$\frac{P(\hat{S}_r > x)}{P(\hat{B} > x)} \to \infty$$

as $x \to \infty$; the tail of $\hat{S}_r$ is heavier than the tail of $\hat{B}$.

PRI Stationary Workload

Theorem (PRI Stationary Workload)

For the stable $M/G/1$ PRI model, with $\hat{Q}$ geometric with mass at 0 and parameter $p_0 = 2 - E[e^{\lambda S}]$,

$$(V|V>0) \overset{d}{=} \hat{S}_r + \sum_{j=1}^{\hat{Q}} \hat{S}_j$$

and thus

$$F_V = p_0 \delta_0 + (1-p_0) F_{\hat{V}}$$

We can now for example compute

$$E(V) = (1-p_0)\Big[E(\hat{S}_r) + E(\hat{Q})E(\hat{S})\Big]$$

$$= (E(e^{\lambda S})-1)\Big[\frac{E(Se^{\lambda S})}{E(e^{\lambda S})-1} - \frac{1}{\lambda} + \Big[\frac{1}{2-E(e^{\lambda S})} - 1\Big]\Big[\frac{E(Se^{\lambda S}) - \frac{1}{\mu}}{E(e^{\lambda S})-1}\Big]\Big].$$

PRI heavy-traffic limit

Theorem (PRI heavy-traffic limit)

Suppose that there exists a $\lambda_2 > \lambda$ *such that as* $\lambda \uparrow \lambda_2$, $E(e^{\lambda S}) \to E(e^{\lambda_2 S}) = 2$, *and that* $E(Se^{\lambda_2 S}) < \infty$. *Then as* $\lambda \uparrow \lambda_2$

$$p_0 V \longrightarrow exp(\alpha),$$

where $\alpha^{-1} = E(\hat{S}(\lambda_2)) = E(Se^{\lambda_2 S}) - 1/\mu$.

Proof is more delicate than before: the random variables in the geometric sum now themselves depend upon λ, as opposed to the PRD case when they are simply i.i.d. copies of S.

References I

Asmussen, S. & Glynn, P.W. (2017). On preemptive-repeat LIFO queues. *Queueing Systems*, **87**, 1-22.

Asmussen, S. & Glynn, P.W. (2007). *Stochastic Simulation*, **87**, 1-22.

Fakinos, D. (1981). The G/G/1 Queueing System with a Particular Queue Discipline. *J.R. Statist. Soc.*, **B43**, 190-196.

Fakinos, D. (1986). On the Single-Server Queue with the Preemptive-Resume Last-Come- First-Served Queue Discipline. *Journal of Applied Probability*, **23**, 243-248.

Sigman, K. (1995). *Stationary Marked Point Processes: An Intuitive Approach*. Chapman and Hall. New York.

Wolff, R.W. (1989). *Stochastic Modeling And The Theory Of Queues*. Prentice Hall. New Jersey.

Woods, F.S. (1926). *Advanced Calculus: A Course Arranged with Special Reference to the Needs of Students of Applied Mathematics*. Springer-Verlag. New York. 148-151.