Single server queue M/G/1

Samuli Aalto
Department of Communications and Networking
Aalto University

April 21, 2022

1 M/G/1 with a work-conserving service discipline

Consider an $M/G/1$ queue with arrival rate $\lambda > 0$ and mean service time $E[S] = 1/\mu < \infty$. Define $\rho := \lambda/\mu$ as the *load* of the system.

Customers arrive according to a *Poisson process*. Let α_i denote the arrival time of customer i, and $A(t)$ the number of arrivals until time t, with $A(0) = 0$. Thus, $A(t)$ is the counter process corresponding to the point process α_i.

Service times are assumed to be *IID* with a continuous distribution,

$$P\{S \le x\} = F(x) = \int_0^x f(y)\, dy.$$

Let S_i denote the service time of customer i.

We assume that the customers are served according to a *work-conserving* service discipline π. Let $B^\pi(t)$ denote the number of departures until time t and $X^\pi(t)$ the *queue length* (i.e., the number of customers in the system) at time t,

$$X^\pi(t) := A(t) - B^\pi(t).$$

1.1 Throughput

In the long run, the mean arrival rate converges to λ (by the Elementary Renewal Theorem),

$$\lim_{t\to\infty} E[A(t)]/t = \lambda.$$

Let $B^\pi(t)$ denote the number of departures until time t. In addition, let θ denote the long-run average departure rate,

$$\theta := \lim_{t\to\infty} E[B^\pi(t)]/t,$$

which is called the *throughput* of service discipline π.

Proposition 1

For any work-conserving service discipline π, the throughput is

$$\theta = \min\{\lambda, \mu\}.$$

Proof The throughput (of any discipline) is clearly upper-bounded by

$$\theta \le \min\{\lambda, \mu\}.$$

On the other hand, if $\theta < \min\{\lambda, \mu\}$, then $\theta < \lambda$ for sure. Thus, in the long run, customers accumulate in the system, and any work-conserving discipline urges the system to operate without breaks so that $\theta = \mu$, which is a contradiction. □

Corollary 1

(i) If $\rho \le 1$, then $\theta = \lambda$ for all work-conserving disciplines.

(ii) If $\rho \ge 1$, then $\theta = \mu$ for all work-conserving disciplines.

Note that the throughput is not only insensitive to the service discipline (as long as it is work-conserving) but also to the service time distribution (as long as the mean service time remains the same).

1.2 Stability

Service discipline π is *unstable* if, for all $n \geq 0$,

$$\lim_{t \to \infty} P\{X^{\pi}(t) \geq n\} = 1.$$

Service discipline π is *stable* if it is not unstable.

Proposition 2

All service disciplines are unstable if $\rho \geq 1$.

Proof Assume first that $\rho > 1$ so that $\lambda > \mu$. The throughput (of any discipline) is clearly upper-bounded by

$$\theta \leq \min\{\lambda, \mu\} < \lambda.$$

Thus, in the long run, customers accumulate in the system for sure so that the system is unstable.

For the case $\rho = 1$, the proof is much more delicate (and is therefore omitted in this course). □

Proposition 3

All work-conserving service disciplines are stable if $\rho < 1$.

Proof Consider what happens if the system is unstable. Then the number of users grows without limits (as a function of time t), implying that the probability that there is at least one customer approaches 1. Consequently, in the long run, the system with a work-conserving service discipline operates without breaks, i.e., $\theta = \mu$. Thus, $\lambda \geq \mu$ by Proposition 1, implying that $\rho \geq 1$. □

1.3 Unfinished work process

Let $U(t) \geq 0$ denote the *unfinished work* (or, workload) at time t,

$$U(t) = \sum_{i=1}^{A(t)} S_i - \int_0^t 1_{\{U(s)>0\}}\, ds. \tag{1}$$

Thus, process $U(t)$ jumps up by S_i when customer i arrives, and decreases at rate 1 when there is work (or, as well, customers) in the system. It follows that $U(t)$ is a continuous-state and continuous-time Markov process with state space $[0, \infty)$.

Note that $U(t)$ is, by definition, independent of the service discipline π (unlike $X^\pi(t)$). For work-conserving disciplines, the workload is equal to the time needed to empty the system unless there are no new arrivals. In particular, if the system is empty, the workload is naturally 0.

Proposition 4

The unfinished work process $U(t)$ is the same for all work-conserving service disciplines π. In addition,

$$\{X^\pi(t) = 0\} = \{U(t) = 0\}.$$

When the system is stable (i.e., $\rho < 1$), $U(t)$ has a *negative trend* whenever $U(t) > 0$. In other words, for any $x > 0$,

$$\lim_{h\to 0+} \frac{1}{h} E[U(t+h) - U(t) \mid U(t) = x] < 0.$$

The proof starts from the observation that the probability for an arrival in the interval $(t, t+h]$ is $\lambda h + o(h)$, where $\lim_{h\to 0+} o(h)/h = 0$.

1.4 Busy and idle periods

Consider a work-conserving service discipline. The server is *busy* whenever the system is non-empty, and *idle* otherwise. A *busy* [*idle*] *period* is an unbroken interval during which the server is busy [idle].

Assume that the system is empty in the beginning, i.e., $U(0) = 0$. The first idle period starts at time 0 lasting until the first arrival at time α_1,

$$I_1 := \inf\{t > 0 : U(t) > 0\} = \alpha_1.$$

The first busy period B_1 starts at time $I_1 = \alpha_1$, and it lasts until the system is empty again,

$$B_1 := \inf\{t > I_1 : U(t) = 0\} - I_1.$$

Let C_1 denote the first *busy cycle* defined by

$$C_1 := I_1 + B_1.$$

In addition, let N_1 denote the number of customers served in the first busy period,

$$N_1 := A(C_1).$$

Finally, let $\gamma_1 = 0$ denote the starting time of the first busy cycle.

The other idle periods I_n, busy periods B_n, and busy cycles C_n together with the number of customers N_n served in busy period n are defined recursively:

$$\begin{aligned}
\gamma_n &:= \gamma_{n-1} + C_{n-1}, \\
I_n &:= \inf\{t > \gamma_n : U(t) > 0\} - \gamma_n, \\
B_n &:= \inf\{t > \gamma_n + I_n : U(t) = 0\} - \gamma_n - I_n, \\
C_n &:= I_n + B_n, \\
N_n &:= A(\gamma_n + C_n) - A(\gamma_n).
\end{aligned}$$

Note that a new busy cycle starts whenever the system becomes empty. In addition, we have

$$B_n = \sum_{i=A(\gamma_n)+1}^{A(\gamma_n)+N_n} S_i, \tag{2}$$

and

$$B_n + I_{n+1} = \sum_{i=A(\gamma_n)+1}^{A(\gamma_n)+N_n} A_i, \tag{3}$$

where $A_i = \alpha_{i+1} - \alpha_i$ refers to the interarrival time between customers i and $i+1$.

Since all these random variables are determined from the arrival process $A(t)$ and the unfinished work process $U(t)$, they must be the same for all work-conserving disciplines.

Proposition 5

Idle periods I_n, busy periods B_n, busy cycles C_n, and the number of customers N_n served in a busy period are the same for all work-conserving service disciplines π.

Due to Poisson arrivals and IID service times, it is clear that what happens within a busy cycle is totally independent of the events of the other cycles.

Proposition 6

The busy cycles C_n constitute a renewal sequence *(γ_n). In addition, $X^\pi(t)$ and $U(t)$ are regenerative processes with respect to the renewal sequence (γ_n) for all work-conserving service disciplines π.*

Besides the busy cycles C_n, the idle periods I_n, the busy periods B_n, and the number of customers N_n served in a busy period are IID. Furthermore, for a stable system with $\rho < 1$, all the mean values are finite as shown below.

Proposition 7

Assume that $\rho < 1$. Then for all work-conserving service disciplines:

(i) Idle periods I_n are IID with mean $E[I] = 1/\lambda$.

(ii) Busy periods B_n are IID with mean $E[B] = \frac{E[S]}{1-\rho}$.

(iii) Busy cycles C_n are IID with mean $E[C] = \frac{1/\lambda}{1-\rho}$.

(iv) The number of customers N_n served in a busy period are IID with mean $E[N] = \frac{1}{1-\rho}$.

Proof (i) is clear due to the memoryless property of the exponential distribution. So we have to prove (ii), (iii), and (iv).

It follows from (2) that

$$B_1 = \sum_{i=1}^{N_1} S_i.$$

On the other hand, it is easy to see that N_1 is a stopping time of sequence (A_n, S_n). Thus, by Wald's equation, we conclude that

$$E[B] = E[B_1] = E[N_1]E[S] = E[N]E[S]. \tag{4}$$

In addition, by definition,

$$E[C] = E[C_1] = E[I_1] + E[B_1] = E[I] + E[B]. \tag{5}$$

Finally, it follows from (3) that

$$B_1 + I_2 = \sum_{i=1}^{N_1} A_i.$$

Thus, again by Wald's equation, we have

$$E[C] = E[B] + E[I] = E[B_1] + E[I_2] = E[N_1]E[A] = E[N]E[I]. \tag{6}$$

The result follows now by solving the three unknowns, $E[B]$, $E[C]$, and $E[N]$, from the three equations (4), (5), and (6). $\square$

Note that the mean values $E[I]$, $E[B]$, $E[C]$, and $E[N]$ are not only insensitive to the service discipline (as long as it is work-conserving) but also to the service time distribution (as long as the mean service time remains the same).

By Propositions 6 and 7, the *steady-state variables* X^π and U,

$$\begin{aligned} P\{X^\pi \le x\} &:= \lim_{t\to\infty} P\{X^\pi(t) \le x\} = \frac{E[\int_0^C 1_{\{X^\pi(t)\le x\}}\, dt]}{E[C]}, \\ P\{U \le x\} &:= \lim_{t\to\infty} P\{U(t) \le x\} = \frac{E[\int_0^C 1_{\{U(t)\le x\}}\, dt]}{E[C]}, \end{aligned}$$

are well defined whenever the system is stable, i.e., $\rho < 1$.

Proposition 8

Assume that $\rho < 1$. Then for all work-conserving service disciplines:

$$\begin{aligned} P\{X^\pi = 0\} &= P\{U = 0\} = 1 - \rho, \\ P\{X^\pi > 0\} &= P\{U > 0\} = \rho. \end{aligned}$$

Proof By Proposition 7 (iii), we have

$$\begin{aligned} P\{X^\pi \le x\} &= \frac{E[\int_0^C 1_{\{X^\pi(t)\le x\}}\, dt]}{E[C]} = \lambda(1-\rho)E[\int_0^C 1_{\{X^\pi(t)\le x\}}\, dt], \\ P\{U \le x\} &= \frac{E[\int_0^C 1_{\{U(t)\le x\}}\, dt]}{E[C]} = \lambda(1-\rho)E[\int_0^C 1_{\{U(t)\le x\}}\, dt]. \end{aligned}$$

In particular, for $x = 0$, we thus have

$$P\{X^\pi = 0\} = P\{U = 0\} = \lambda(1-\rho)E[I] = 1 - \rho,$$

which completes the proof. $\square$

Note that the probabilities $P\{X^\pi = 0\}$ and $P\{X^\pi > 0\}$ (and, as well as, $P\{U = 0\}$ and $P\{U > 0\}$) are not only insensitive to the service discipline (as long as it is work-conserving) but also to the service time distribution (as long as the mean service time remains the same).

In the long run, the system with a work-conserving service discipline is non-empty with probability ρ. Therefore, load ρ is also often called the *utilization* (factor) of the system.

2 M/G/1-FIFO

In this section, we assume that the service discipline is FIFO. So we focus on the performance analysis of an $M/G/1$-FIFO queue. By utilizing the theory of *regenerative processes*, we derive the so called *Pollaczek-Khinchin mean value formulas* for the steady-state variables (queue length, waiting time, and delay).

The main conclusion is that, for a fixed mean service time $E[S]$, the mean steady-state waiting time, sojourn time, and queue length are all increasing functions of the coefficient of variation of the service time distribution. In other words, FIFO gives the best performance for deterministic service times, while the performance is getting worse and worse as the variability in the service time distribution increases (as long as the arrival rate λ and the mean service time $E[S]$ remain the same).

2.1 Waiting times W_i

Since the customers are served in their arrival order, the *sojourn time* T_i of customer i consists of two phases, the *waiting time* W_i and the *service*

time S_i,

$$T_i = W_i + S_i. \tag{7}$$

However, if the system is empty when customer i arrives, the waiting time is zero ($W_i = 0$) and the sojourn time equals the service time ($T_i = S_i$).

Let Y_i^w denote the number of waiting customers that the arriving customer i sees,

$$Y_i^w = \max\{X(\alpha_i-) - 1, 0\}, \tag{8}$$

where $X(t-) := \lim_{h\to 0} X(t-h)$ is the left limit of the queue length process. In addition, let $R(t)$ denote the *remaining service time* of the customer in service at time t (if any). If the system is empty (i.e., $X(t) = 0$), we define $R(t) = 0$. Because of the FIFO discipline, the waiting time of customer i satisfies clearly

$$W_i = \sum_{j=1}^{Y_i^w} S_{i-j} + R(\alpha_i-) = U(\alpha_i-), \tag{9}$$

which is just the left limit of the unfinished work process at the arrival time.

Recall that the busy cycles C_n constitute a renewal sequence (γ_n), and all the queueing related processes (including $R(t)$) are regenerative with respect to the renewal sequence (γ_n). In addition, the mean cycle $E[C]$ is finite for $\rho < 1$. Thus, the steady-state variable R,

$$P\{R \le x\} := \lim_{t\to\infty} P\{R(t) \le x\} = \frac{E[\int_0^C 1_{\{R(t)\le x\}}\, dt]}{E[C]},$$

is well defined whenever $\rho < 1$.

Proposition 9 *Consider the M/G/1-FIFO queue with $\rho < 1$. The mean steady-state remaining service time is*

$$E[R] = \frac{\lambda}{2} E[S^2].$$

Proof Since $R(t)$ is regenerative in each busy cycle, we have

$$E[R] = \frac{E[\int_0^C R(t)\,dt]}{E[C]}.$$

Let N denote the number of customers served in a busy cycle, which is a stopping time of sequence (A_n, S_n). Recall from Proposition 7 that

$$E[N] = \frac{1}{1-\rho}.$$

Now the key observation is

$$\int_0^C R(t)\,dt = \sum_{i=1}^{N} \frac{1}{2} S_i^2,$$

implying, by Wald's equation, that

$$E[\int_0^C R(t)\,dt] = \frac{1}{2} E[N] E[S^2] = \frac{E[S^2]}{2(1-\rho)}.$$

Recall further from Proposition 7 that

$$E[C] = \frac{1/\lambda}{1-\rho}.$$

Thus,

$$E[R] = \frac{E[\int_0^C R(t)\,dt]}{E[C]} = \frac{\lambda}{2} E[S^2],$$

which completes the proof. □

2.2 Pollaczek-Khinchin mean value formulas

Let us now first derive the mean steady-state waiting time $E[W]$,

$$E[W] = \lim_{i \to \infty} E[W_i],$$

and then utilize the result when determining the mean steady-state unfinished work $E[U]$, sojourn time $E[T]$, and queue length $E[X]$.

Proposition 10

Consider the M/G/1-FIFO queue with $\rho < 1$. The mean steady-state waiting time is

$$E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{10}$$

Proof From (9) we deduce that

$$\begin{aligned} E[W_i] &= E[E[\sum_{j=1}^{Y_i^w} S_{i-j}|Y_i^w]] + E[R(\alpha_i-)] \\ &= E[Y_i^w E[S]] + E[R(\alpha_i-)] \\ &= E[Y_i^w]E[S] + E[R(\alpha_i-)], \end{aligned}$$

where Y_i^w refers to the number of waiting customers and $R(\alpha_i-)$ to the remaining service time seen by arriving customer i. Note that equation

$$E[\sum_{j=1}^{Y_i^w} S_{i-j}|Y_i^w] = Y_i^w E[S]$$

is justified by the fact that Y_i^w is independent of service times $S_{i-1}, \ldots, S_{i-Y_i^w}$. Due to PASTA, the steady-state variables $\lim_{i\to\infty} Y_i^w$ and $\lim_{i\to\infty} R(\alpha_i-)$ are distributed as steady-state variables $X^w := \lim_{t\to\infty} \max\{X(t)-1, 0\}$ and $R := \lim_{t\to\infty} R(t)$. Thus,

$$E[W] = E[X^w]E[S] + E[R].$$

By further taking into account Little's formula $E[X^w] = \lambda E[W]$, we get

$$E[W] = \lambda E[W]E[S] + E[R],$$

implying, by Proposition 9, that

$$E[W] = \frac{E[R]}{1-\rho} = \frac{\lambda E[S^2]}{2(1-\rho)},$$

which completes the proof. □

Equation (10) is known as the *Pollaczek-Khinchin mean value formula* for the steady-state waiting time.

The mean steady state waiting time $E[W]$ equals the mean steady-state unfinished work $E[U]$. This is explained by the PASTA property of the Poisson arrival process as follows: For the FIFO discipline, $V(t) := U(t-)$ can be interpreted as the *virtual waiting time process*, i.e., $V(t)$ is the time that the arriving customer has to wait if it happens to arrive at time t. Due to PASTA, arriving customers see the system in equilibrium, implying that $E[W] = E[V] = E[U]$, where the last equation follows from a continuity argument. Thus, we have the following Pollaczek-Khinchin mean value formula for the *steady-state unfinished work*:

$$E[U] = E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{11}$$

The corresponding Pollaczek-Khinchin mean value formula for the *steady-state sojourn time* is clearly

$$E[T] = E[S] + E[W] = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{12}$$

By applying Little's formula $E[X] = \lambda E[T]$, we also get the Pollaczek-Khinchin mean value formula for the *steady-state queue length*:

$$E[X] = \lambda E[T] = \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}. \tag{13}$$

Note that, when the mean service time $E[S]$ is kept fixed, the mean steady-state waiting time, sojourn time, and queue length are increasing functions of the coefficient of variation of the service time distribution,

$$C[S] := \sqrt{\frac{E[S^2]}{E[S]^2} - 1}.$$

Therefore, any Erlang service time distribution ($C[S] < 1$) performs better than the exponential distribution ($C[S] = 1$), while hyperexponential and Pareto distributions ($C[S] > 1$) result in a worse mean performance in the $M/G/1$-FIFO queue.

In addition, we notice that the optimal mean performance for the FIFO discipline is clearly achieved with deterministic service times (with $C[S] = 0$).

On the other hand, a service time distribution for which $E[S^2] = \infty$ (such as Pareto(b, β) with $\beta \leq 2$) results in an infinite mean sojourn time, waiting time, and queue length for any load $\rho > 0$.

3 M/G/1-PS

In this section, we assume that the service discipline is PS. So we focus on the performance analysis of an $M/G/1$-PS queue. By applying the so called *phase method*, we derive the steady-state queue length distribution, from which we also get the mean values for the steady-state variables (queue length, sojourn time).

The main conclusion is that, for a fixed mean service time $E[S]$, the queue length distribution is *insensitive* to the shape of the service time distribution. Thus, PS gives the same performance for any service time distribution (as long as the arrival rate λ and the mean service time $E[S]$ remain the same). Thus, the results are the same as in an $M/M/1$-PS queue.

3.1 Exponential service times

We start with the M/M/1-PS queue, i.e., we assume first that the service times S_i are independent and follow the Exp(μ) distribution with mean $1/\mu$.

Consequently, the queue length process $X(t)$ is an irreducible Markov birth-death process with state space $\mathcal{S} = \{0, 1, \ldots\}$ and the following (positive) state transition rates for any $n \in \mathcal{S}$:

$$\begin{aligned} q(n, n+1) &= \lambda, \\ q(n+1, n) &= (n+1)\,\frac{\mu}{n+1} \;=\; \mu. \end{aligned}$$

Proposition 11

Consider the M/M/1-PS queue with $\rho < 1$. The steady-state queue length distribution is

$$P\{X = n\} = (1-\rho)\rho^n, \quad n = 0, 1, \ldots$$

Proof Denote $\pi_n := (1-\rho)\rho^n$. Since $\rho < 1$, the normalization condition (N) is clearly satisfied:

$$\sum_{n=0}^{\infty} \pi_n = (1-\rho) \sum_{n=0}^{\infty} \rho^n = \frac{1-\rho}{1-\rho} = 1.$$

It remains to prove that the *detailed balance equations* (DBE) are also satisfied for all n:

$$\pi_n \lambda = \pi_{n+1} \mu.$$

Now

$$\pi_n \lambda = (1-\rho)\rho^n \lambda = (1-\rho)\rho^n \rho\mu = (1-\rho)\rho^{n+1}\mu = \pi_{n+1}\mu,$$

which completes the proof. □

In addition to the PS discipline, the result is, in fact, valid for any *work-conserving* queueing discipline of an M/M/1 queue.

3.2 Erlangian service times

Next we consider the M/E_K/1-PS queue. So we assume here that the service times S_i are independent and follow the Erl$(K, K\mu)$ distribution with mean

$1/\mu$. We further assume that $K \geq 2$ so that the service time consists of multiple sequential *phases* $k = 1, \ldots, K$.

In this case, the queue length process $X(t)$ is *not* a Markov process. To get a Markovian description of the system, we have to additionally keep track of the current phases of the customers. Let $N(t) = (N_k(t); k = 1, \ldots, K)$, where $N_k(t)$ refers to the total number of customers in phase k at time t. The state space of process $N(t)$ is clearly

$$\mathcal{S} = \{n = (n_1, \ldots, n_K); n_k \in \{0, 1, \ldots\}\}.$$

In addition, let e_k denote the unit vector to direction k in this space, $e_k = (n_1, \ldots, n_K)$ with $n_k = 1$ and $n_j = 0$ for $j \neq k$.

Due to the PS service discipline, $N(t)$ is an irreducible Markov process with the following (positive) state transition rates for any $n \in \mathcal{S}$ and $k \in \{1, \ldots, K-1\}$:

$$\begin{aligned} q(n, n + e_1) &= \lambda, \\ q(n + e_k, n + e_{k+1}) &= K\mu \frac{n_k + 1}{n_1 + \ldots + n_K + 1}, \\ q(n + e_K, n) &= K\mu \frac{n_K + 1}{n_1 + \ldots + n_K + 1}. \end{aligned}$$

Proposition 12

Consider the $M/E_K/1$-PS queue with $\rho < 1$. The steady-state distribution of the Markov process $N(t)$ is

$$P\{N = n\} = (1 - \rho)(\rho/K)^{n_1 + \ldots + n_K} \frac{(n_1 + \ldots + n_K)!}{n_1! \ldots n_K!}, \quad n \in \mathcal{S}.$$

Proof Denote

$$\pi(n) := (1 - \rho)(\rho/K)^{n_1 + \ldots + n_K} \frac{(n_1 + \ldots + n_K)!}{n_1! \ldots n_K!}.$$

In addition, let

$$\mathcal{S}_m = \{n \in \mathcal{S} : n_1 + \ldots + n_K = m\}.$$

Note that, for all $n \in \mathcal{S}_m$ and $k = 1, \ldots, K$,

$$\pi(n + e_k) = \pi(n)(\rho/K)\,\frac{m+1}{n_k+1}. \tag{14}$$

Since $\rho < 1$ and

$$\sum_{n \in \mathcal{S}_m} \frac{(n_1 + \ldots + n_K)!}{n_1! \ldots n_K!}(1/K)^m = ((1/K) + \ldots + (1/K))^m = 1,$$

the normalization condition (N) is satisfied:

$$\begin{aligned}
\sum_{n \in \mathcal{S}} \pi(n) &= (1-\rho) \sum_{m=0}^{\infty} \sum_{n \in \mathcal{S}_m} \frac{(n_1 + \ldots + n_K)!}{n_1! \ldots n_K!}(\rho/K)^m \\
&= (1-\rho) \sum_{m=0}^{\infty} \rho^m \\
&= \frac{1-\rho}{1-\rho} = 1.
\end{aligned}$$

It remains to prove that the *global balance equations* (GBE) are also satisfied for all n:

$$\sum_{n' \neq n} \pi(n) q(n, n') = \sum_{n' \neq n} \pi(n') q(n', n). \tag{15}$$

Let $n \in \mathcal{S}$, and denote $m = n_1 + \ldots + n_K$. Note first that

$$\begin{aligned}
\pi(n + e_1) q(n + e_1, n + e_2) &\overset{(14)}{=} \pi(n)(\rho/K)\,\frac{m+1}{n_1+1} \cdot K\mu\,\frac{n_1+1}{m+1} \\
&= \pi(n) \cdot \lambda \\
&= \pi(n) q(n, n + e_1).
\end{aligned} \tag{16}$$

Consider then any $k = 2, \ldots, K-1$. Now

$$\begin{aligned}
\pi(n + e_k) q(n + e_k, n + e_{k+1}) &\overset{(14)}{=} \pi(n)(\rho/K)\,\frac{m+1}{n_k+1} \cdot K\mu\,\frac{n_k+1}{m+1} \\
&= \pi(n)(\rho/K)\,\frac{m+1}{n_{k-1}+1} \cdot K\mu\,\frac{n_{k-1}+1}{m+1} \\
&\overset{(14)}{=} \pi(n + e_{k-1}) q(n + e_{k-1}, n + e_k).
\end{aligned} \tag{17}$$

For $k = K$, we have

$$\begin{aligned}
\pi(n + e_K)q(n + e_K, n) &\overset{(14)}{=} \pi(n)(\rho/K)\,\frac{m+1}{n_K+1} \cdot K\mu\,\frac{n_K+1}{m+1} \\
&= \pi(n)(\rho/K)\,\frac{m+1}{n_{K-1}+1} \cdot K\mu\,\frac{n_{K-1}+1}{m+1} \\
&\overset{(14)}{=} \pi(n + e_{K-1})q(n + e_{K-1}, n + e_K). \qquad (18)
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\pi(n)q(n, n + e_1) &\overset{(14)}{=} \pi(n + e_K)(K/\rho)\,\frac{n_K+1}{m+1} \cdot \lambda \\
&= \pi(n + e_K)K\mu\,\frac{n_K+1}{m+1} \\
&= \pi(n + e_K)q(n + e_K, n). \qquad (19)
\end{aligned}$$

Thus, equations (16), (17), (18), and (19), which together are called *station balance equations* (SBE), are true for any $n \in \mathcal{S}$. Note that each of these equations (except (19)) correspond to transitions into a state and out of that state generated by a customer entering a phase and leaving that phase, respectively! Equation (19) corresponds to one customer entering the whole system and another leaving it. The global balance equations (15) follow from these station balance equations in a straightforward way by summing up the related SBE's. □

Note that the queue length process $X(t)$ satisfies

$$X(t) = N_1(t) + \ldots + N_K(t).$$

As an immediate consequence of the previous result, we get the steady-state queue length distribution, which (surprisingly) appears to be the same as for the M/M/1-PS queue:

$$P\{X = m\} \;=\; \sum_{n \in \mathcal{S}_m} P\{N = n\}$$

$$
\begin{aligned}
&= (1-\rho) \sum_{n \in \mathcal{S}_m} \frac{(n_1 + \ldots + n_K)!}{n_1! \ldots n_K!} (\rho/K)^m \\
&= (1-\rho)\rho^m.
\end{aligned}
$$

Corollary 2

Consider the M/E_K/1-PS queue with $\rho < 1$. The steady-state queue length distribution is

$$P\{X = n\} = (1-\rho)\rho^n, \quad n = 0, 1, \ldots$$

3.3 General service times

Here we finally consider the general M/G/1-PS queue. So we assume that the service times S_i are IID with mean $1/\mu$. As in the previous section, the queue length process $X(t)$ is neither a Markov process in this case. Below we describe the *phase method* that can be used to derive the steady-state queue length distribution.

Any service time distribution can be approximated (with an arbitrary precision) by a *phase-type* distribution, represented by the absorption time in an absorbing Markov process with a finite state space $1, \ldots, K$. To get a Markovian description of the approximating system, we have to keep track of the current phases of the customers. Let $N(t) = (N_k(t); k = 1, \ldots, K)$, where $N_k(t)$ refers to the total number of customers in phase k at time t. The state space of process $N(t)$ is clearly

$$\mathcal{S} = \{n = (n_1, \ldots, n_K); n_k \in \{0, 1, \ldots\}\}.$$

It is possible to find the steady-state distribution of the Markov process $N(t)$ by verifying the *station balance equations*, which are generated by jumps of single customers from one phase to another (including the arrivals and the

departures) in a similar manner as in the case of M/E_K/1-PS queues. Since the queue length process $X(t)$ satisfies

$$X(t) = N_1(t) + \ldots + N_K(t),$$

one is able to derive the steady-state queue length distribution, which (surprisingly) appears to be the same as for the M/M/1-PS queue.

Proposition 13

Consider the M/G/1-PS queue with $\rho < 1$. The steady-state queue length is geometrically distributed with point probabilities

$$P\{X = n\} = (1-\rho)\rho^n, \quad n = 0, 1, \ldots, \tag{20}$$

and mean value

$$E[X] = \frac{\rho}{1-\rho}. \tag{21}$$

Proof This can be proved by the phase method utilizing the station balance equations, see, e.g., [1, Sect. 3.3]. □

Since the queue length distribution remains the same for any service time distribution with the same mean $1/\mu$, the steady-state queue length distribution of the M/G/1-PS queue is said to be *insensitive* to the service time distribution.

Little's formula, $E[X] = \lambda E[T]$, gives the following corollary.

Corollary 3

Consider the M/G/1-PS queue with $\rho < 1$. The mean steady-state sojourn time is

$$E[T] = \frac{E[S]}{1-\rho}. \tag{22}$$

Interestingly, the mean sojourn time $E[T]$ in the M/G/1-PS queue equals the mean busy period $E[B]$. We note also that it is only the mean delay which is insensitive to the service time distribution, but not the whole delay distribution.

4 Performance comparison between FIFO and PS

Below we summarize the performance comparison between FIFO and PS disciplines.

Corollary 4

Consider an M/G/1 queue with $\rho < 1$. Then

$$E[X^{\mathrm{FIFO}}] \leq E[X^{\mathrm{PS}}] \iff E[T^{\mathrm{FIFO}}] \leq E[T^{\mathrm{PS}}] \iff C[S] \leq 1,$$

where $C[S] := D[S]/E[S]$ refers to the coefficient of variation of the service time distribution.

References

[1] F.P. Kelly, *Reversibility and Stochastic Networks*, Wiley, 1979.

[2] L. Kleinrock, *Queueing Systems, Vol. I: Theory*, Wiley, 1975.