Memorandum 1900 (June 2009). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Stability of Two Exponential Time-limited Polling Models

Roland de Haan, Richard J. Boucherie, Jan-Kees van Ommeren
University of Twente, Enschede, The Netherlands

Abstract

In this article, we consider the stability of two single-server polling models. More specifically, we will state and prove the stability conditions of single-server polling systems operating under the pure and exhaustive exponential time-limited service discipline. These conditions will be proven for the polling system operating under the periodic polling strategy and preemptive service. The stability proof of the pure time-limited discipline is straightforward as stability may be considered for each queue in isolation. The proof for the exhaustive time-limited discipline is more laborious. We follow the line of proof as introduced by Fricker and Jaïbi [1] for a large class of service disciplines. Unfortunately, the preemptive nature of the exhaustive time-limited discipline excludes it from this class and as a result substantial efforts are required to modify the proof as to allow for preemptive disciplines. Finally, the extension of the proofs to the Markovian polling strategy is discussed.

1 Introduction

In this article, we will state and prove the stability conditions for the pure and exhaustive service discipline in the context of polling models. The pure time-limited discipline states that the server visits a queue exactly for a random amount of time, while according to the exhaustive time-limited discipline the server already leaves a queue as soon as it becomes empty. Both service disciplines are preemptive disciplines meaning that any on-going service at the time limit will be interrupted and must be restarted at a next visit. The stability conditions prescribe limits on the amount of traffic that can be sustained by the system. Exceeding these limits leads to instable behavior, but operation of the system just below these limits will already lead to large buffers and long transfer delays.

Stability conditions for a large class of polling systems have been proven by Fricker and Jaïbi [1]. In particular, the authors considered periodic polling systems under non-preemptive and work-conserving service disciplines. The necessary and sufficient condition for stability reads:

$$\text{System is stable} \iff \rho + \max_{1\leq i\leq M}\left(\frac{\lambda_i}{\mathbb{E}[G_i^*]}\right)\cdot c_T < 1,$$

where ρ is the total offered load to the system, c_T is the mean total switch-over time during a cycle and $\mathbb{E}[G_i^*]$ denotes the mean number of served customers at Q_i during a cycle when Q_i is saturated. Saturation in this context means that there is an unlimited number of customers waiting to be served at a polling instant to Q_i. For the exhaustive and gated service discipline, this condition readily simplifies to:

$$\text{System is stable} \iff \rho < 1,$$

since the number of served customers may grow to infinity in these cases, i.e., $\mathbb{E}[G_i^*] = \infty$. Later, the same authors have also proven a similar stability condition for the Markovian polling strategy [2]. However, stability results for preemptive service disciplines, such as the pure and exhaustive time-limited disciplines, are not available in the literature.

To close this gap, we will prove stability conditions for both time-limited disciplines. This will be done for the periodic polling strategy and an exponential time limit. For the pure-time limited discipline, the stability question can be resolved by studying each queue in isolation. For the exhaustive time-limited discipline, stability must be considered for the system as a whole and we will prove the stability condition by adopting the line of proof of [1].

This article is organized as follows. The model is formally described in Sect. 2. Next, the stability conditions for the pure and the exhaustive time-limited disciplines are given in Sects. 3 and 4, respectively. We conclude the article in Sect. 5.

2 Model

Let us consider the basic polling system of M queues with Poisson arrivals and generally distributed service and switch-over times. The server visits the queues according to the periodic polling strategy. Without loss of generality (w.l.o.g.) we define a cycle as the time period between two consecutive polling instants at the 1st stage (or visit) of the cycle. A cycle consists of a stages and we denote by $t(j),\ j = 1, \ldots, a$, the queue served during stage j of the cycle. Further, the number of times Q_i is visited during a cycle is denoted by $a_i,\ i = 1, \ldots, M$, with $a_i \geq 1$ and $\sum_{i=1}^{M} a_i = a$.

The service discipline assumed in Sect. 3 is the pure time-limited discipline, whereas in Sect. 4 the exhaustive time-limited discipline is assumed. The time limit at Q_i for both disciplines is exponentially distributed with parameter ξ_i. Due to the time limit, service will be preempted at the timer expiration and in such case a service time will be redrawn from the original distribution at the start of the next visit; thus, we assume the so-called *preemptive-repeat with resampling* strategy.

3 Pure exponential time-limited discipline

For the pure time-limited discipline, the queues in the system are independent from a stability perspective as service capacity cannot be exchanged among the queues. The polling system is stable if and only if there exists a stationary regime in which each customer in the system can be served in a finite period of time. We will say that the system is stable if and only if all the queues in the system are stable.

A necessary and sufficient condition for the stability of a polling system with the server operating under the pure exponential time-limited discipline is given in the following theorem.

Theorem 1 (Pure exponential time-limited discipline)**.**

$$\textit{System is stable} \iff \rho_i < \kappa_i,\ \forall_{i \in \{1,\ldots,M\}},$$

where

$$\begin{aligned}\rho_i &= \lambda_i \cdot \frac{1-\tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)},\\ \kappa_i &= \frac{a_i/\xi_i}{\sum_{j=1}^{M} a_j/\xi_j + \sum_{k=1}^{a} c_{t(k),t(k+1)}},\end{aligned}$$

where $c_{t(k),t(k+1)}$ *is the mean switch-over time from the queue visited in stage* k *to the queue visited in stage* $k+1$.

Proof. It is well-known that for a single queue the nonsaturation condition is both a necessary and sufficient condition for stability, i.e.,

$$Q_i \text{ is stable } \iff \rho_i < \kappa_i,\ i = 1, \ldots, M,$$

where ρ_i is the mean effective amount of work arriving per time unit to Q_i and κ_i is the availability fraction of the server at Q_i.

Consider first the mean effective amount of work arriving per time unit to Q_i. This amount is determined by the total number of customers arriving per time unit λ_i and the mean effective amount of work each individual brings for the server, denoted by $\tilde{\sigma}_i$, as follows:

$$\rho_i = \lambda_i \cdot \tilde{\sigma}_i.$$

The quantity $\tilde{\sigma}_i$ is in fact the mean total time the server spends on serving a customer at Q_i including any interrupted services. Noting that the number of interruptions per customer is geometrically distributed, it can be found via simple calculus that:

$$\tilde{\sigma}_i = \frac{1-\tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)}.$$

The availability fraction of the server κ_i is fully specified by the mean visit times, the visit frequencies and the switch-over times between the queues. Notice that a complete cycle consists of a_i visits to Q_i, $i = 1, \ldots, M$, and the switch-over times between the queues. It then readily follows for the availability fraction of the server at Q_i:

$$\kappa_i = \frac{a_i/\xi_i}{\sum_{j=1}^{M} a_j/\xi_j + \sum_{k=1}^{a} c_{t(k),t(k+1)}}.$$

It is good to notice that the fraction κ_i is independent of the load at the queues. The observation that the system is stable if and only if all the queues in the system are stable completes the proof. □

4 Exhaustive exponential time-limited discipline

For the exhaustive time-limited discipline, service capacity can be exchanged between the queues. This suggests that stability must be considered for the system as a whole. However, as the visit time to each queue is bounded by the timer, the occupancy of individual queues also plays a role. The polling system is considered stable if there exists a stationary regime in which each customer in the system can be served in a finite period of time.

A necessary and sufficient condition for the stability of a polling system with the server operating under the exhaustive exponential time-limited discipline is given in the following theorem.

Theorem 2 (Exhaustive exponential time-limited discipline)**.**

$$\textit{System is stable} \iff \rho + \max_{1\le i\le M}\left(\frac{\lambda_i}{\mathbb{E}[G_i^{*-}]}\right)\cdot c_T < 1,$$

where

$$\mathbb{E}[G_i^{*-}] = \frac{a_i\cdot \tilde{X}_i(\xi_i)}{1-\tilde{X}_i(\xi_i)},$$

denotes the mean number of served customers at Q_i during a cycle when Q_i is saturated and c_T is the mean total switch-over time during a cycle.

We will prove the theorem adopting the approach of Fricker and Jaïbi [1]. To this end, we will often stick to their notation whenever it does not lead to ambiguity. We should emphasize that the authors of [1] considered only work-conserving service disciplines. The exhaustive time-limited (E-TL) discipline allows for preemption of service and thus is definitely not work conserving.

The organization of the proof is as follows. In Sects. 4.1 and 4.2, we state several preliminary and monotonicity results which are analogous to the results in [1]. Essentially, we need to introduce notation to account for the preemption of the service, but the line of proof remains similar. Hence, the lemmas and theorems corresponding directly to the ones provided in [1] will be given without proof. In Sect. 4.3, we present several novel results for the visit time to the queues during a cycle and also give the proofs. These results are then incorporated in the final necessary and sufficiency proofs of [1], so that these account for preemptive service.

4.1 Preliminaries and stochastic monotonicity

The general service disciplines for which stability is proven in [1] should satisfy four properties. Property 1 and 3 refer to the independence of the service discipline on the history of the service process and on the independence of the customer selection. These properties are readily seen to be satisfied for the E-TL discipline. Property 2 deals with the work conservation and is not satisfied for this discipline since work is created due to preemptions. However, during the course of a visit the server is always working and does not idle. Finally, Property 4 is the so-called stochastic monotonicity property and is defined as follows [1]: "*As the queue size grows, the number of customers served during one stage (visit) grows stochastically, but such that the number of customers left at the end of the stage (visit) grows stochastically as well.*" This latter property plays a crucial role in the proof.

Let us w.l.o.g. consider an arbitrary queue in the polling system. First, we define an independent and identically distributed (i.i.d.) sequence $(\sigma^m)_{m=1,2,\ldots}$, as the modified service times of a customer, with mean σ, and being distributed as $\min(X,V)$, where X is distributed according to the original service time distribution, and V is exponentially distributed and independent of X. That is, the modified service times can be seen as the duration of a service attempt (which can either be successful or interrupted). For non-preemptive service disciplines, the number of customers taken into service is equal to the number of customers served during a visit. However, this is not always true for the preemptive discipline that we consider here. Hence, we will define also the following quantities for a visit with x customers present at the start (i.e., $t=0$):

- $f^+(x)$: the number of customers that is taken into service during the visit;

- $f^-(x)$: the number of customers that is actually served during the visit;
- $v(x)$: the duration of the visit;
- $\phi(x)$: the number of customers at the end of the visit.

The quantities $f^+(x)$ and $f^-(x)$ are given by:

$$\begin{aligned} f^+(x) &= \min(N^0(x), N^*), \\ f^-(x) &= \min(N^0(x), N^* - 1), \end{aligned} \tag{1}$$

where $N^0(x)$ refers to the number of served customers during a visit which started with x customers and which ends due to the queue becoming empty, and N^* refers to the number of customers taken into service during a visit which ends due to the expiration of the time limit. Notice that, due to the exponential visit times, N^* is in fact a geometrically distributed random variable independent of x.

Let us denote by $N(a, b]$ the number of arrivals to the queue during the interval $(a, b]$. Thus, we may write the following relations between $f^+(x)$, $f^-(x)$, $v(x)$ and $\phi(x)$:

$$v(x) = \sum_{m=1}^{f^+(x)} \sigma^m, \tag{2}$$

$$\phi(x) = x - f^-(x) + N(0, v(x)], \tag{3}$$

with $f^+(0) = f^-(0) = v(0) = 0$. It is good to notice that σ^m in Eq. (2) refers to the duration of an arbitrary service attempt rather than an original service time of an arbitrary customer.

Let us next recall the definitions of $\leq$-monotonicity and $\leq_d$-monotonicity as given in [1]:

Definition 1. *($\leq$-monotonicity)*
A real function h defined on $\mathbb{R}^n$ is called $\leq$-monotone when:

$$x \leq y \Rightarrow h(x) \leq h(y).$$

Definition 2. *($\leq_d$-monotonicity)*
Two (cumulative) distributions functions P_1 and P_2 on $\mathbb{R}^n$ satisfy $P_1 \leq_d P_2$ when:

$$\int h \; dP_1 \leq \int h \; dP_2,$$

for any $\leq$-monotone function h such that the integrals are well defined.
Two random vectors X_1 and X_2 satisfy $X_1 \leq_d X_2$ if their distributions satisfy $P_1 \leq_d P_2$.

Hence, the monotonicity property for the E-TL discipline is that $(f^+(x), f^-(x), \phi(x))$ is $\leq_d$-monotone in x. It follows immediately from Eq. (2) that $\leq_d$-monotonicity of $f^+(x)$ implies $\leq_d$-monotonicity of $v(x)$, and that $\leq_d$-monotonicity of $f^-(x)$ does not imply that of $\phi(x)$.

Next, we embed the queue into the polling system. Let the nth visit to the queue start at stopping time T_n (with respect to the complete history of the system) with N_n customers waiting. Define the following quantities:

- F_n^+: the number of customers that is taken into service during visit n;

- F_n^-: the number of customers that is actually served during visit n;
- V_n: the duration of visit n;
- Φ_n: the number of customers at the end of visit n.

Let us introduce the tuple (f^+, f^-, v, ϕ) which represents the service discipline. It can readily be argued (cf. [1, p.215]) that for each n:

$$(F_n^+, F_n^-, V_n, \Phi_n) =_d (f^+(N_n), f^-(N_n), v(N_n), \phi(N_n)).$$

Along the single-queue equations, Eqs. (2) and (3), we find that for any n, V_n and Φ_n are related to F_n^+ and F_n^- as follows.

$$\begin{aligned} V_n &= \sum_{i=D_n+1}^{D_n+F_n^+} \sigma^i, \\ \Phi_n &= N_n - F_n^- + N(T_n, T_n + V_n], \end{aligned}$$

where D_n denotes the number of service attempts performed up to T_n. Since $(F_n^+, F_n^-, V_n, \Phi_n)$ is independent of future service attempt durations, i.e.,
$(\sigma^i)_{i>D_n+F_n^+}$, and the future customer arrival process, i.e., $N(T_n+V_n, T_n+V_n+\cdot]$, we may apply Wald's equation and obtain:

$$\mathbb{E}[V_n] = \mathbb{E}[F_n^+] \cdot \sigma, \tag{4}$$

$$\mathbb{E}[N(T_n, T_n + V_n]] = \mathbb{E}[F_n^+] \cdot \lambda \cdot \sigma. \tag{5}$$

Notice that the expectations in (4) and (5) are finite, since the visit duration is always bounded by the exponential timer.

Let F^{*+} (F^{*-}) be the number of customers that are taken into service (served) during a visit if there are infinitely many customers waiting in the queue, and V^* the duration of such a visit, i.e.,

$$\begin{aligned} 0 < \lim_{x\to\infty} \mathbb{E}[f^+(x)] = \mathbb{E}[F^{*+}] < \infty, \\ 0 < \lim_{x\to\infty} \mathbb{E}[f^-(x)] = \mathbb{E}[F^{*-}] < \infty, \end{aligned}$$

and also,

$$\lim_{x\to\infty} \mathbb{E}[v(x)] = \mathbb{E}[V^*] = \mathbb{E}[F^{*+}] \cdot \sigma < \infty.$$

Next, we present a lemma which will be needed in the final part of the proof. This lemma substitutes in fact Lemma 1 of [1].

Lemma 1. *Let $(N_n)_n$ be a sequence of random variables converging in distribution to a, possibly degenerate, integer-valued random variable N. Let (f^+, f^-, v, ϕ) be induced by the E-TL service discipline and be independent of $(N, (N_n)_n)$, i.e., N^* is independent of $(N, (N_n)_n)$ (see, Eq.(1)). The sequence $(N_n, f^+(N_n), f^-(N_n), v(N_n)$,
$\phi(N_n))_n$ converges in distribution to $(N, f^+(N), f^-(N), v(N), \phi(N))$, and
(i) when $\mathbb{E}[F^{*-}] < \infty$, and if N has a defective distribution, then so is the limiting distribution of $N_n - f^-(N_n)$;*

(ii) when $\mathbb{E}[F^{*-}] < \infty$, $\mathbb{E}[F^-(N)] < \mathbb{E}[F^{*-}]$ *if and only if there exists a* $y < \infty$ *such that* $\mathbb{P}(N \leq y) > 0$ *and* $\mathbb{E}[f^-(y)] < \mathbb{E}[F^{*-}]$.
In both cases, if $(N_n)_n$ *is* $\leq_d$*-monotone,* $\lim_{n\to\infty} \mathbb{E}[F^-(N_n)] = \mathbb{E}[f^-(N)]$ *and* $\lim_{n\to\infty} \mathbb{E}[v(N_n)] = \mathbb{E}[v(N)]$.

Proof. The proof is immediate from the proof of Lemma 1 in [1]. □

Remark 1 (Number of customers taken into service)**.** *We have defined Lemma 1 in terms of the number of customers served. Analogously, this lemma can be defined for the number of customers taken into service.*

Finally, the following lemma is essential for the remainder of the proof and refers to a specific property of the service discipline addressed in the beginning of this section.

Lemma 2. *The E-TL discipline satisfies the stochastic monotonicity property.*

Proof. The proof follows by sample-path arguments and is immediate from the proof of Lemma 2 in [1]. □

4.2 Monotonicity

The stochastic monotonicity property plays a key role in the stability proof. Therefore, we will state several monotonicity results [1] which are valid for service disciplines satisfying this property.

To this end, we describe the system by the queue lengths at the polling instants and define $M(t)$ as follows:

$$M(t) = (N_1(t), \ldots, N_M(t)),\ t \geq 0.$$

Recall that $t(i)$ denotes the queue served at visit i of a cycle. We denote by visit (n, i) the ith visit in the nth cycle and let visit $(1, 1)$ start at time $t = 0$. Let $T_{n,i}$ denote the time of the polling instant of visit (n, i), so that we have:

$$0 = T_{1,1} \leq T_{1,2} \leq \ldots \leq T_{1,a} \leq T_{2,1} \leq \ldots\ .$$

For convenience, we write $M_{n,i}$ for $M(T_{n,i})$ and $N_{n,i}$ for $N_{t(i)}(T_{n,i})$. Hence, we can describe the Markovian behavior of the system as follows.

Proposition 1. *(Prop. 1 of [1]) The sequence* $(M_{n,i})_{n,i}$ *is a Markov chain. For each* i *fixed in* $\{1, \ldots, a\}$*, the Markov chain* $(M_{n,i})_n$ *is homogeneous, aperiodic and irreducible on (a subset of)* $\mathbb{N}^M$.

Proof. See [1]. □

Let us define by π_i the transition operator at visit $i,\ 1 \leq i \leq a$, of the Markov chain $(M_{n,i})_{n,i}$ as follows:

$$\pi_i h(\mathbf{m}) = \mathbb{E}[h(M_{n,i+1}) | M_{n,i} = \mathbf{m}],$$

for any $\mathbf{m} = (m_1, \ldots, m_M)$ and any real function h defined on $\mathbb{N}^M$ for which the expectation exists. Besides, we let $\tilde{\pi}$ be the transition operator of the Markov chain $(M_{n,i})_n$. An operator π is said to be $\leq_d$-monotone if for all distributions $P_1 \leq_d P_2$, $\pi P_1 \leq_d \pi P_2$. This holds if πh is $\leq$-monotone when h is.

Lemma 3. *(Lemma 3 of [1]) For all i, π_i and $\tilde{\pi}_i$ are $\leq_d$-monotone.*

Proof. See [1]. □

Let us next define the following quantities:

- $F^+_{n,i}$: the number of customers taken into service during visit (n,i);
- $F^-_{n,i}$: the number of customers served during visit (n,i);
- $V_{n,i}$: the duration of visit (n,i).

An immediate consequence of Lemma 3 is the monotonicity property of the state process.

Proposition 2. *Suppose $M_{1,1} = (0,\ldots,0)$. Then, for each i, $(M_{n,i})_n$ and $(F^+_{n,i}, F^-_{n,i}, V_{n,i})$ are $\leq_d$-monotone.*

Proof. The proof is immediate from the proof of Proposition 2 in [1]. □

Next, we turn to dominance relations between polling systems. In particular, we compare systems with a different number of saturated queues. Here, saturation means that at a polling instant of a queue there is an infinite number of customers waiting. The saturation of a queue implies that the server serves the queue up to the time limit and then leaves. From the viewpoint of the other queues in the system, such a visit to a saturated queue is merely an additional switch-over time. Let $\mathcal{S}$ be the initial polling system with queues $1,\ldots,M$. For $e \in \{0,\ldots,M\}$, we define the subsystem $\mathcal{S}^e$ as the polling system consisting of the queues $1,\ldots,e$, resulting from the saturation of the queues $e+1,\ldots,M$, and served according to the same periodic schedule as the original system. We emphasize that if $t(i) > e$ then in $\mathcal{S}^e$ no queue is served but the server becomes unavailable for a duration of $V^*_{t(i)}$, which is defined as the stationary duration of a visit to queue $t(i)$ with an infinite number of customers waiting at the start of the visit. Let us define σ_j as the mean duration of a service attempt at Q_j, i.e., $\sigma_j := \mathbb{E}[min(X_j, V_j)]$, where X_j refers to the original service time of a customer at Q_j and V_j to the visit time of the server to Q_j. Further, denote by $\mathbb{E}[G^{*+}_j]$ the expected number of customers taken into service at Q_j during a cycle when Q_j is saturated. Then, the mean total switch-over time in the subsystem c^e_T can be written as:

$$c^e_T = c_T + \sum_{j=e+1}^{M} \sigma_j \mathbb{E}[G^{*+}_j] = c_T + \sum_{j=e+1}^{M} a_j/\xi_j.$$

The state space of the subsystem $\mathcal{S}^e$ is given by the sequence $M^e_{n,i} = (N^e_1(T^e_{n,i}),$ $\ldots, N^e_e(T^e_{n,i}))$ at the polling instants $T^e_{n,i}$. For each visit i, $(M^e_{n,i})_n$ is a Markov chain and is $\leq_d$-monotone if the initial state is the empty state. The subsystem $\mathcal{S}^e$ is similar to the original system $\mathcal{S}$ in the sense that all previous results apply to it. Let denote by $M^{g|e}$ the e first components of a vector M^g having $g > e$ components. Then, the subsystems $\mathcal{S}^e$ satisfy the following dominance property.

Lemma 4. *(Lemma 4 of [1]) For $e < g$ both in $\{0,\ldots,M\}$, $\mathcal{S}^e$ dominates $\mathcal{S}^g$ in the sense that if $M^{g|e}_{1,1} \leq_d M^e_{1,1}$ then $M^{g|e}_{n,i} \leq_d M^e_{n,i}$ for all (n,i).*

Proof. See [1]. □

4.3 Stability proof

The polling system is said to be stable if:

- there exists a proper stationary joint-distribution for the queue lengths at the polling instants at stage k, for all $k = 1, \ldots, a$;
- the stationary cycle time is finite.

4.3.1 Proof: Sufficient condition

We assume w.l.o.g. that the system is empty at time 0 as the stationary distribution of the Markov chain does not depend on the initial distribution. For convenience, let us introduce several definitions for the number of customers at a specific queue, viz.,

- H_k^- : number of customers actually served at Q_k during a visit to Q_k;
- H_k^+ : number of customers taken into service at Q_k during a visit to Q_k;
- H_k^{*-} : number of customers actually served at Q_k during a visit when Q_k is saturated;
- H_k^{*+} : number of customers taken into service at Q_k during a visit when Q_k is saturated.

Notice that these definitions resemble the definitions of F_n^-, F_n^+, F_n^{*-} and F_n^{*+}. However, the latter quantities refer to the number of customers at the nth visit rather than to the number at a specific queue.

W.l.o.g. we consider the cycle from $T_{n,1}$ to $T_{n+1,1}$. Then, we may similarly define the counterparts G_k^-, G_k^+, G_k^{*-} and G_k^{*+} which count the same quantities but over a complete cycle. Hence, we may then also write:

$$\begin{aligned}
\mathbb{E}[G_k^-] &:= \mathbb{E}[H_{k,1}^-] + \cdots + \mathbb{E}[H_{k,a_k}^-], \\
\mathbb{E}[G_k^+] &:= \mathbb{E}[H_{k,1}^+] + \cdots + \mathbb{E}[H_{k,a_k}^+], \\
\mathbb{E}[G_k^{*-}] &:= \mathbb{E}[H_{k,1}^{*-}] + \cdots + \mathbb{E}[H_{k,a_k}^{*-}], \\
\mathbb{E}[G_k^{*+}] &:= \mathbb{E}[H_{k,1}^{*+}] + \cdots + \mathbb{E}[H_{k,a_k}^{*+}],
\end{aligned}$$

where $\mathbb{E}[H_{k,i}^-]$ is the mean number of customers served at Q_k during the ith visit to Q_i in a cycle, and $\mathbb{E}[H_{k,i}^+]$, $\mathbb{E}[H_{k,i}^{*-}]$, and $\mathbb{E}[H_{k,i}^{*+}]$ are defined similarly. Besides, we define $\tilde{\sigma}_k$ as the mean effective service time of a customer at Q_k. The effective service time refers to the total time spent by the server on serving a customer (including interrupted service attempts, but excluding the periods that the server is not present at the queue) and it is in fact a geometric sum of service attempt durations. Thus, using Wald's equation, we may write for its mean:

$$\tilde{\sigma}_k = \mathbb{E}\left[\sum_{n=1}^{N} \min(X_{k,n}, V_{k,n})\right] = \sigma_k / \tilde{X}_k(\xi_k), \tag{6}$$

where N is geometrically distributed with success probability $p = \tilde{X}_k(\xi_k)$, and $\{X_{k,n}\}_{n\geq 1}$ and $\{V_{k,n}\}_{n\geq 1}$ are two independent families of independent random variables distributed as X_k and V_k, respectively.

Let us denote by $\mathbb{E}[V_k^c]$ the mean total visit time to Q_k during a cycle, i.e.,

$$\mathbb{E}[V_k^c] = \mathbb{E}[V_{k,1}] + \cdots + \mathbb{E}[V_{k,a_k}],$$

where $\mathbb{E}[V_{k,j}]$ stands for the mean visit time during the jth visit to Q_k of a cycle. Then, we are ready to present the following lemma:

Lemma 5.

$$\mathbb{E}[V_k^c] = \mathbb{E}[G_k^-] \cdot \tilde{\sigma}_k, \ k = 1, \ldots, M. \tag{7}$$

The proof of the lemma will be given below. However, we will derive several intermediate results first.

Clearly, when Q_k is saturated, there is always exactly one interrupted service. Thus, we have the following property:

Property 1.

$$H_k^{*+} = H_k^{*-} + 1, \ k = 1, \ldots, M,$$

and since $\mathbb{E}[H_k^{*+}] < \infty$ *also:*

$$\mathbb{E}[H_k^{*+}] = \mathbb{E}[H_k^{*-}] + 1, \ k = 1, \ldots, M.$$

Besides, there is a less obvious relation between the quantities H_k^+, H_k^-, H_k^{*+} and H_k^{*-}. However, before we get to this relation, we give a lemma and present some useful properties for H_k^+ and H_k^-.

Lemma 6. *Let H be a geometrically distributed random variable and let W be a non-negative discrete random variable independent of H. Then, the following assertion holds:*

$$\mathbb{E}[H \mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[W \mathbf{1}_{\{W < H\}}] = \mathbb{E}[H] \cdot \mathbb{E}[\mathbf{1}_{\{W \geq H\}}].$$

Proof.

$$\begin{aligned} \mathbb{E}[H] &= \mathbb{E}[H\mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[H\mathbf{1}_{\{W < H\}}] \\ &= \mathbb{E}[H\mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[W\mathbf{1}_{\{W < H\}}] + \mathbb{E}[(H - W)\mathbf{1}_{\{W < H\}}]. \end{aligned}$$

Next, we may use the fact that H is a geometric and thus memoryless random variable, i.e., $H - W|_{H>W} =_d H$, so that:

$$\mathbb{E}[H] = \mathbb{E}[H\mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[W\mathbf{1}_{\{W < H\}}] + \mathbb{E}[H] \cdot \mathbb{E}[\mathbf{1}_{\{W < H\}}].$$

This completes the proof. □

Denote by N_k^0 the number of customers served until Q_k would become empty for the first time if there were no timer. The following properties are readily verified:

Property 2.

$$\begin{aligned} H_k^+ &= \min(N_k^0, H_k^{*+}) = N_k^0 \mathbf{1}_{\{N_k^0 \le H_k^{*+}\}} + H_k^{*+} \mathbf{1}_{\{N_k^0 > H_k^{*+}\}}, \\ H_k^- &= \min(N_k^0, H_k^{*-}) = N_k^0 \mathbf{1}_{\{N_k^0 \le H_k^{*-}\}} + H_k^{*-} \mathbf{1}_{\{N_k^0 > H_k^{*-}\}}. \end{aligned}$$

These properties imply that if the server leaves Q_k because it is empty, then $H_k^+ = H_k^-$ and $H_k^+ = H_k^- + 1$, otherwise.

The following lemma demonstrates that the ratio of mean number of served customers and mean number of customers taken into service is equal both for a saturated and a non-saturated queue.

Lemma 7.

$$\frac{\mathbb{E}[H_k^{*-}]}{\mathbb{E}[H_k^{*+}]} = \frac{\mathbb{E}[H_k^-]}{\mathbb{E}[H_k^+]}, \ k = 1, \ldots, M.$$

Proof. Note that H_k^{*+} is a geometrically distributed random variable (with success probability $p = 1 - \tilde{X}_k(\xi_k)$, since an interruption is seen as a success). Then,

$$\begin{aligned} \mathbb{E}[H_k^+] \cdot \mathbb{E}[H_k^{*-}] &= \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}]\right) \left(\mathbb{E}[H_k^{*+}] - 1\right) \\ &= \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}]\right) \cdot \mathbb{E}[H_k^{*+}] \\ &\quad - \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}]\right) \\ &= \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}]\right) \cdot \mathbb{E}[H_k^{*+}] \\ &\quad - \mathbb{E}[\mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}] \cdot \mathbb{E}[H_k^{*+}] \\ &= \mathbb{E}[H_k^{*+}] \cdot \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[(H_k^{*+} - 1) \mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}]\right), \end{aligned}$$

where for the third equality sign we used Lemma 6. Finally, observe that $\{N_k^0 < H_k^{*+}\} = \{N_k^0 \le H_k^{*-}\}$ (since all variables are discrete), so that we may write:

$$\begin{aligned} &\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[(H_k^{*+} - 1) \mathbf{1}_{\{N_k^0 \ge H_k^{*+}\}}] \\ &= \mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 \le H_k^{*-}\}}] + \mathbb{E}[(H_k^{*+} - 1) \mathbf{1}_{\{N_k^0 > H_k^{*-}\}}] \\ &= \mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 \le H_k^{*-}\}}] + \mathbb{E}[H_k^{*-} \mathbf{1}_{\{N_k^0 > H_k^{*-}\}}] \\ &= \mathbb{E}[H_k^-], \end{aligned}$$

where we used Proposition 2 in the final step. □

Remark 2 (Independence of N_k^0)**.** *It is important to notice that the equivalence of the ratios does not depend on the distribution of N_k^0. In particular, we have made no assumptions whatsoever on the number of customers present at the start of a visit in the unsaturated case. So, the time of the previous polling instant of the queue does not impact the ratio of the unsaturated case (while the ratio of the saturated case is obviously fixed).*

Recall that $\mathbb{E}[V_k^*]$ denotes the mean visit time of the server to Q_k when Q_k is saturated. This quantity satisfies the following relation.

Lemma 8.

$$\mathbb{E}[V_k^*] = \mathbb{E}[H_k^{*+}] \cdot \sigma_k \;=\; \mathbb{E}[H_k^{*-}] \cdot \tilde{\sigma}_k,\; k = 1, \ldots, M.$$

Proof. Consider the saturated case. We write: $V_k^* = \sum_{j=1}^{H_k^{*+}} Y_{k,j}$, where $Y_{k,j},\; j = 1, 2, \ldots,$ are i.i.d. random variables distributed as $min(X_k, V_k)$ and with mean σ_k. Let further $V_{k,j}$ and $X_{k,j},\; j = 1, 2, \ldots,$ be i.i.d. random variables distributed as the generic random variables X_k and V_k, with X_k and V_k independent. Notice that $H_k^{*+} = \min\{j : X_{k,j} > V_{k,j}\}$. Therefore, H_k^{*+} is a stopping time for $Y_{k,j},\; j = 1, 2, \ldots$, so that we may apply Wald's equation yielding: $\mathbb{E}[V_k^*] = \mathbb{E}[H_k^{*+}] \cdot \sigma_k$. Next, consider a period Z comprising a single visit of length V_k^* to Q_k extended with the (service) time needed to complete the service of the customer that was interrupted at the end of the visit. That is, Z is the time needed to complete all (residual) services that were started during V_k^* (in particular, we include a possible residual service time, which is in fact distributed as an effective service time due to the geometric nature of the effective service time). Thus, $\mathbb{E}[Z] = \mathbb{E}[H_k^{*+}] \cdot \tilde{\sigma}_k$, but also $\mathbb{E}[Z] = \mathbb{E}[V_k^*] + \tilde{\sigma}_k$, since there is always an interrupted service with a mean residual service time identical to the original mean effective service time. Hence, it follows that: $\mathbb{E}[V_k^*] = (\mathbb{E}[H_k^{*+}] - 1) \cdot \tilde{\sigma}_k = \mathbb{E}[H_k^{*-}] \cdot \tilde{\sigma}_k$. □

Combining Lemma 8 with Eq. (6) gives after some manipulations:

Corollary 1.

$$\mathbb{E}[G_k^{*-}] = a_k \cdot \mathbb{E}[H_k^{*-}] = \frac{a_k \cdot \tilde{X}_k(\xi_k)}{1 - \tilde{X}_k(\xi_k)}.$$

Proof. (Proof of Lemma 5) It is readily seen that to prove Eq. (7) it is sufficient to show:

$$\mathbb{E}[V_{k,j}] = \mathbb{E}[H_{k,j}^{-}] \cdot \tilde{\sigma}_k,\; j = 1, \ldots, a_k.$$

W.l.o.g. we consider the first visit to Q_k in a cycle and leave out the subscript 1. Thus, we need to prove the following:

$$\mathbb{E}[V_k] = \mathbb{E}[H_k^{-}] \cdot \tilde{\sigma}_k.$$

Analogously to the proof of Lemma 8, we write $V_k = \sum_{j=1}^{H_k^{+}} X_{k,j}$ for the unsaturated case. By arguing that H_k^{+} is a stopping time for the sequence $\{X_{k,j}\}_j$, it immediately follows via Wald that: $\mathbb{E}[V_k] = \mathbb{E}[H_k^{+}] \cdot \sigma_k$. The proof is then completed by appealing to Lemma 7 and Lemma 8. □

Let us define $\hat{\rho}_k,\; k = 1, \ldots, M$ as follows:

$$\hat{\rho}_k := \sum_{j=1}^{k} \rho_j = \sum_{j=1}^{k} \lambda_j \tilde{\sigma}_j.$$

Next, we define a stability condition for the complete system and for the subsystems $\mathcal{S}^e$ with $e \in \{0, \ldots, M\}$:

Definition 3. *(Condition $\mathcal{C}^M$)*

$$\mathcal{C}^M : \hat{\rho}_M + \max_{1\leq j\leq M}(\lambda_j/\mathbb{E}[G_j^{*-}])c_T < 1.$$

Definition 4. *(Condition $\mathcal{C}^e$)*

$$\mathcal{C}^e : \hat{\rho}_e + \max_{1\leq j\leq e}(\lambda_j/\mathbb{E}[G_j^{*-}])c_T^e < 1.$$

We number the queues according to the ratio $\lambda_j/\mathbb{E}[G_j^{*-}]$ in non-decreasing order. Hence, we have that:

$$\mathcal{C}^e : \hat{\rho}_e + (\lambda_e/\mathbb{E}[G_e^{*-}])c_T^e < 1.$$

Further, we note that it can verified by simple calculations that $\mathcal{C}^{e+1}$ implies $\mathcal{C}^e$.

We are now ready to present the following lemma (cf. Lemma 6 of [1]) which forms a crucial link in the proof:

Lemma 9. *If condition $\mathcal{C}^e$ holds, then*

$$\mathbb{E}[G_k^{e-}] < \mathbb{E}[G_k^{*-}],\ 1 \leq k \leq e.$$

Proof. Let us consider the mean duration of a cycle. W.l.o.g. we say that the nth cycle starts at time $T_{n,1}$ and ends at time $T_{n+1,1}$. A cycle consists of the visits to the queues and the switch-over times, so that we may write (cf. Lemma 5):

$$\mathbb{E}[T_{n+1,1} - T_{n,1}] = \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T,\ n = 0, 1, \dots\ .$$

Hence, for the change in number of customers at Q_k during this cycle, we readily have for $k = 1, \dots, M,\ n = 0, 1, \dots$:

$$\mathbb{E}[N_k(T_{n+1,1}) - N_k(T_{n,1})] = \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T \right) - \mathbb{E}[G_{n,k}^-].$$

Suppose w.l.o.g. that the system is empty at time 0. Using the $\leq_d$-monotonicity for each given visit, it follows that the expectations of the queue lengths at the polling times are non-decreasing, i.e.,

$$\mathbb{E}[N_k(T_{n+1,1}) - N_k(T_{n,1})] \geq 0,\ k = 1, \dots, M,\ n = 0, 1, \dots,$$

which provides us immediately with the following system of equations:

$$\mathbb{E}[G_{n,k}^-] \leq \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T \right),\ k = 1, \dots, M,\ n = 0, 1, \dots\ . \tag{8}$$

Observe that $\mathbb{E}[G_{n,k}^-]$ and $\mathbb{E}[G_{n,k}^+]$ are non-decreasing in n and are bounded from above by $\mathbb{E}[G_k^{*+}] < \infty$. Thus, we may write for the following limits, $k = 1, \ldots, M$:

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}[G_{n,k}^-] &= \mathbb{E}[G_k^-], \\ \lim_{n\to\infty} \mathbb{E}[G_{n,k}^+] &= \mathbb{E}[G_k^+]. \end{aligned}$$

Let us consider next Eq. (8) for $k = 1$ and let $n \to \infty$:

$$\mathbb{E}[G_1^-] \leq \lambda_1 \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right).$$

This can readily be rewritten to:

$$\mathbb{E}[G_1^-] \cdot (1 - \hat{\rho}_1) \leq \lambda_1 \cdot \left(\sum_{j=2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right).$$

Applying a triangularization procedure (see Appendix 6), we may obtain for $1 \leq k \leq M$:

$$\mathbb{E}[G_k^-] \cdot (1 - \hat{\rho}_k) \leq \lambda_k \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right). \tag{9}$$

The latter result does also hold if we consider the system with queues $e + 1$ up to M being saturated. From the point of view of the first e queues only the return time of the server will change while the behavior of the server during a visit remains identical. Denoting the quantities in this modified system by adding the superscript e, we may write:

$$\mathbb{E}[G_k^{e-}] \leq \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^{e-}] + c_T^e \right), \ k = 1, \ldots, e,$$

where $c_T^e = c_T + \sum_{j=e+1}^{M} \mathbb{E}[V_j^*]$. Since, $\mathbb{E}[G_j^{e-}] \leq \mathbb{E}[G_j^{*-}], \ j = 1, \ldots, M$, we obtain:

$$\begin{aligned} \mathbb{E}[G_k^{e-}] &\leq \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^{*-}] + c_T \right) \\ &= \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot c_T^k, \ k = 1, \ldots, e. \end{aligned}$$

On the other hand, the condition $\mathcal{C}^k, \ k = 1, \ldots, e$, which is implied by $\mathcal{C}^e$, reads:

$$\mathcal{C}^k : \hat{\rho}_k + \max_{1 \leq j \leq k} (\lambda_j / \mathbb{E}[G_j^{*-}]) \cdot c_T^k < 1.$$

Under the assumption that the ratios $\lambda_j / \mathbb{E}[G_j^{*-}]$ are ordered non-decreasingly, it is readily found that $\mathcal{C}^k$ implies:

$$\mathbb{E}[G_k^{*-}] > \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot c_T^k,$$

which completes the proof. □

The remainder of the proof is along the lines of [1]. Recall that we want to show here that if condition $\mathcal{C}^M$ is satisfied, then the system is stable. An equivalent definition of stability (see [1]) is that there exists a proper stationary joint queue-length distribution at the polling instants such that the expectation of the stationary cycle time is finite. A sufficient condition for the stationary distribution to exist is that the multi-dimensional Markov chain $(M^e_{n,1})$ is ergodic. The ergodicity of this chain $(M^e_{n,1})$ is equivalent to the existence of $\mathbf{m}^e$ such that the limit

$$\lim_{n\to\infty} \mathbb{P}(M^e_{n,1} \leq \mathbf{m}^e) \geq 1 - \sum_{k=1}^{e} \lim_{n\to\infty} \mathbb{P}(N_k(T^e_{n,1}) \geq \mathbf{m}_k), \tag{10}$$

is strictly positive. We note that the system will become empty once the chain enters some state $\leq \mathbf{m}^e$ with a strictly positive probability (due to no arrivals for a specific time). Since the positiveness of the limit implies that you return infinitely often to some state $\leq \mathbf{m}^e$, it follows that with probability one you will reach the empty state in a finite amount of time; in other words, it excludes transient or null-recurrent behaviour of the chain. Thus, to have ergodicity, we need the sum on the right-hand side of Eq. (10) to be strictly smaller than one. This can be established if for one k the limiting distribution $(N_k(T^e_{n,1}))_n$ is not concentrated at infinity, i.e., $\mathbb{P}(N_k < \infty) > 0$ and all other limiting distributions are proper, i.e., $\mathbb{P}(N_k < \infty) = 1$.

We will prove this by induction starting with the subsystem $\mathcal{S}^0$. This system $\mathcal{S}^0$ is readily seen to be stable. Next, we suppose $\mathcal{S}^{e-1}$ is stable, and consider $\mathcal{S}^e$, $1 \leq e \leq M$. We note since $\mathcal{S}^{e-1}$ is stable, the Markov chain $(M^{e-1}_{n,1})$ is ergodic and in particular $(N_k(T^{e-1}_{n,1}))_n$, $1 \leq k \leq e-1$ has a proper distribution. Also, $(M^{e-1}_{n,i})_n$, $i = 1, ..., a$ has a proper limiting distribution and by Lemma 4, $M^{e|e-1}_{n,i} \leq_d M^{e-1}_{n,i}$ for all n. Thus, $(M^{e|e-1}_{n,i})_n$ has a proper limiting distribution. Moreover, from Lemma 5, we have that $\mathbb{E}[G^{e-}_e] < \mathbb{E}[G^{*-}_e]$. Hence, there exists a visit r such that $\lim_{n\to\infty} \mathbb{E}[F^{e-}_{n,r}] < \mathbb{E}[F^{*-}_r]$. Then, by Lemma 1-ii there exists a y such that $\lim_{n\to\infty} \mathbb{P}(N^e_{n,r} \leq y) > 0$, i.e., the limiting distribution of the last component $N^e_{n,r} = N^e_e(T_{n,r})$ of $M^e_{n,r}$ is not concentrated at infinity. Thus, the chain $(M^e_{n,r})_n$ is ergodic. The observation that the expectation of the cycle time is finite completes the proof.

□

4.3.2 Proof: Necessary condition

Suppose the polling system $\mathcal{S}$ is stable. Let us define F^-_{n,k_l} as the mean number of customers served during the k_l-th stage of the nth cycle, where the k_l-th stage corresponds to exactly the lth visit to Q_k in the cycle. We let for each visit i the initial distribution of $(M_{n,i})_n$ be its stationary distribution. Since $\mathcal{S}$ is stable, these chains are stationary with positive-recurrent states. As a result, $\mathbb{P}(N_k(T_{n,i}) = 0) > 0$ for all k and (n, i). Further, as the expected cycle time is finite, $\mathbb{E}[G^-_k] = \sum_{l=1}^{a_k} \mathbb{E}[F^-_{n,k_l}]$ does not depend on n and is finite for all k. It follows by Lemma 1 that $\mathbb{E}[G^-_k] < \mathbb{E}[G^{*-}_k]$ for $1 \leq k \leq M$ and in particular that $\mathbb{E}[G^-_M] < \mathbb{E}[G^{*-}_M]$.

On the other hand, it can readily be seen that:

$$N_k(T_{2,1}) - N_k(T_{1,1}) = N_k(T_{1,1}, T_{2,1}] - \sum_{l=1}^{a_k} F^-_{1,k_l}.$$

Hence, we can bound $N_k(T_{2,1}) - N_k(T_{1,1})$ as follows:

$$-\sum_{l=1}^{a_k} F^-_{1,k_l} \leq N_k(T_{2,1}) - N_k(T_{1,1}) \leq N_k(T_{1,1}, T_{2,1}].$$

Both the lower and upper bound have finite expectation, such that for all k (see [1, Lemma 7]):

$$\mathbb{E}[N_k(T_{2,1}) - N_k(T_{1,1})] = 0,$$

and in general for $n \geq 1$:

$$\mathbb{E}[N_k(T_{n+1,1}) - N_k(T_{n,1})] = 0.$$

This leads to (cf. Eq. (8)) the following system of equalities:

$$\mathbb{E}[G^-_k] = \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G^-_j] + c_T \right), \ 1 \leq k \leq M,$$

and along the lines of deriving Eq. (9), we obtain:

$$\mathbb{E}[G^-_k] \cdot (1 - \hat{\rho}_k) = \lambda_k \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G^-_j] + c_T \right), \ 1 \leq k \leq M.$$

Specifically, for $k = M$ this implies:

$$\mathbb{E}[G^-_M] \cdot (1 - \hat{\rho}_M) = \lambda_M \cdot S.$$

Together with the observation above, $\mathbb{E}[G^-_M] < \mathbb{E}[G^{*-}_M]$, it follows that condition $\mathcal{C}^M$ holds.

5 Concluding remarks

We have proven the stability conditions for two polling models with time-limited service and periodic polling. The proof for the pure time-limited discipline is straightforward, since the queues can in fact be decoupled and thus studied in isolation. For the proof of the exhaustive time-limited discipline, we have relied largely on the rigorous stability proof of Fricker and Jaïbi [1] for a class of service disciplines. Unfortunately, this class covers only non-preemptive and work-conserving service disciplines. Though, the main ideas of their proof could still be used to prove stability here.

A logical next step would be to extend the results to Markovian polling of the server. The fixed cycle of periodic polling will then become a random cycle. One would typically consider per-queue cycles, i.e., a cycle starting and ending at consecutive polling instants of a specific queue. Consequently, the number of visits to the other queues during a cycle are random variables. For the pure time-limited discipline, such an extension can readily be incorporated by appropriately adjusting the availability fraction κ_i. For the exhaustive time-limited, it might require some more work to prove this extension. However, we strongly believe this could also be done using similar techniques as the ones presented in this article.

References

[1] C. Fricker and M. R. Jaïbi, "Monotonicity and stability of periodic polling systems," *Queueing Systems*, vol. 15(1-4), pp. 211–238, 1994.

[2] ——, "Stability of a polling model with a Markovian scheme," INRIA report 2278, 1994.

6 Triangularization

Let us explain below the triangularization method that we apply. We depart from the following set of equalities:

$$\mathbb{E}[G_k^-] \leq \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right), \ k = 1, \ldots, M. \tag{11}$$

Rearranging the equality for $k = 1$, we obtain:

$$(1 - \hat{\rho}_1) \cdot \mathbb{E}[G_1^-] \leq \lambda_1 \cdot \left(\sum_{j=2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right). \tag{12}$$

Next, we will show that also for $2 \leq k \leq M$ we may write:

$$(1 - \hat{\rho}_k) \cdot \mathbb{E}[G_k^-] \leq \lambda_k \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right).$$

This will be done by proving the following inequalities by induction.

$$\sum_{j=1}^{k} \tilde{\sigma}_j \mathbb{E}[G_j^-] \leq \frac{\hat{\rho}_k}{1 - \hat{\rho}_k} \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right), \ k = 1, \ldots, M, \tag{13}$$

$$(1 - \hat{\rho}_{k+1}) \cdot \mathbb{E}[G_{k+1}^-] \leq \lambda_{k+1} \cdot \left(\sum_{j=k+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right), \ k = 1, \ldots, M-1. \tag{14}$$

First, notice that for $k = 1$ Eq. (13) has been shown above, while Eq. (14) reads as follows:

$$(1 - \hat{\rho}_2) \cdot \mathbb{E}[G_2^-] \leq \lambda_2 \cdot \left(\sum_{j=3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right).$$

This inequality can be proven from Eq. (11) and taking $k = 2$. First, we take all terms $\mathbb{E}[G_2^-]$ to the left-hand side, second we apply Eq. (12), and finally some simple manipulations provide us with the desired result. Next, we show that once these inequalities hold for l these also hold for

$l+1$. First, consider Eq. (13) for $l+1$:

$$\begin{aligned}
\sum_{j=1}^{l+1} \tilde{\sigma}_j \mathbb{E}[G_j^-] &= \sum_{j=1}^{l} \tilde{\sigma}_j \mathbb{E}[G_j^-] + \tilde{\sigma}_{l+1} \mathbb{E}[G_{l+1}^-] \\
&\leq \frac{\hat{\rho}_l}{1-\hat{\rho}_l} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \frac{1}{1-\hat{\rho}_l} \cdot \tilde{\sigma}_{l+1} \mathbb{E}[G_{l+1}^-] \\
&\leq \left(\frac{\hat{\rho}_l}{1-\hat{\rho}_l} + \frac{\rho_{l+1}}{(1-\hat{\rho}_l)(1-\hat{\rho}_{l+1})} \right) \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&= \frac{\hat{\rho}_{l+1}}{1-\hat{\rho}_{l+1}} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right).
\end{aligned}$$

Second, we have to prove Eq. (14) for $l+1$, i.e.,

$$(1-\hat{\rho}_{l+2}) \cdot \mathbb{E}[G_{l+2}^-] \leq \lambda_{l+2} \cdot \left(\sum_{j=l+3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right). \tag{15}$$

To this end, we depart from Eq. (11) for $l+2$:

$$\begin{aligned}
\mathbb{E}[G_{l+2}^-] &\leq \lambda_{l+2} \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&= \lambda_{l+2} \cdot \sum_{j=1}^{l+1} \tilde{\sigma}_j \mathbb{E}[G_j^-] + \lambda_{l+2} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&\leq \lambda_{l+2} \cdot \frac{\hat{\rho}_{l+1}}{1-\hat{\rho}_{l+1}} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&\quad + \lambda_{l+2} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&= \frac{\lambda_{l+2}}{1-\hat{\rho}_{l+1}} \cdot \left(\sum_{j=l+3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \frac{\lambda_{l+2}}{1-\hat{\rho}_{l+1}} \cdot \tilde{\sigma}_{l+2} \mathbb{E}[G_{l+2}^-].
\end{aligned}$$

Hence, moving all the terms $\mathbb{E}[G_{l+2}^-]$ to the left-hand side and performing some rearrangements yields Eq. (15).