Impatient customers in the *MAP/G/*1 queue

M.B. Combé

Department of Operations Research, Statistics, and System Theory

Report BS-R9413 April 1994

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

Copyright © Stichting Mathematisch Centrum
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

Impatient Customers in the $MAP/G/1$ queue

M.B. Combé
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

This paper considers a single server queueing model with customer impatience. This work extends results on the $M/G/1$ queue with impatient customers to the case of customers arriving according to a Markovian Arrival Process. We derive a transform for the virtual waiting time, from which expressions for customer waiting time and the probability of a premature departure of a customer are obtained.

AMS Subject Classification (1991): 60K25, 90B22
Keywords & Phrases: impatience, (batch) Markovian arrival process, waiting time
Note: The author was supported by NFI

1 INTRODUCTION

In queueing models of communication systems various forms of interaction between the arrival process of customers and the workload of the system may occur. Examples are queueing networks with limited buffer space and admission control, and queueing networks with customer impatience. A key reference is Baccelli et al.[1], which considers $GI/G/1$ queueing systems where customers at the moment of arrival decide whether to join the queue or not, their decision depending on the amount of work present at that time.

One can look at this situation from different perspectives; Baccelli et al.[1] describe the following situations:

-Customers are impatient, i.e., each customer is prepared to wait a limited period of time in the queue, and when the actual waiting time is larger, the customer's patience runs out and he leaves the system. Such a situation might for example occur in a telephone system, where customers disconnect if the waiting time is too large.

-The workload capacity of the system is limited, and customer access is regulated. For example, in a queueing system with finite buffer space a customer might be rejected when his service time would cause an overflow.

There are a number of variants of both examples. In the first example a customer might leave the system without joining the queue, or the decision whether to join the queue or not might be based on the sojourn time instead of the waiting time. In the second example, a customer causing an overflow might not be blocked completely; if at the moment of arrival

Report BS-R9413
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

the system is not completely loaded, a part of the customer's service request might fill up this space, the remainder of its service request being rejected.

In this paper we study the first example for the case of a Markovian Arrival Process (MAP) and exponentially(α) distributed patience. The latter means that when a customer enters the system and the total remaining work in the system is equal to x, the customer will eventually receive service with probability $e^{-\alpha x}$, and will eventually leave the queue without receiving service with probability $1 - e^{-\alpha x}$. The (FCFS) queueing model under consideration will be denoted by $MAP/G/1 + M$.

The motivation for the analysis of this queueing system is twofold. Firstly, it extends the results of Baccelli et al.[1] on the $M/G/1 + M$ queue for the more general MAP. The MAP allows a more accurate description of the arrival processes in modern queueing systems than the Poisson process. For example the Markov Modulated Poisson Process (MMPP), frequently used to model On/Off sources in a communication network, fits into the framework of the MAP. Secondly, any G(I) arrival process can be approximated arbitrarily close by a MAP, hence the analysis of the $MAP/G/1 + M$ queue can be used to provide (approximate) results for the more general $G(I)/G/1 + M$ queue. Note that no explixit expressions are known for the latter model, cf. Baccelli et al.[1], as their efforts on the $G(I)/G/1 + M$ did not result in explicit expressions for the waiting times of customers.

Overview of the paper.

In section 2, we first describe the $MAP/G/1 + M$ queue in more detail, then present an analysis of the workload process, resulting in the Laplace-Stieltjes Transform (LST) of its stationary distribution. In section 3 we consider the moments of the workload process, and also present the LST of the waiting time and the probability that a customer prematurely leaves the queue.

2 THE WORKLOAD PROCESS

In this section we present an analysis of the workload process in the $MAP/G/1 + M$ queue. First we describe the $MAP/G/1 + M$ queue in more detail.

Mathematical description of the $MAP/G/1 + M$ queue.

We consider a single server queue in which the arrival epochs of customers are directed by a continuous time Markov process, and in which for a customer at the moment of his arrival it is decided whether this customer will eventually receive service or not, this decision being based on the amount of work present in the system at that moment, i.e. the actual waiting time of the customer.

Let us first consider the arrival process. In the MAP, the interarrival times of customers are directed by a continuous time Markov process $\{\mathbf{J}(t), t \geq 0\}$ on a finite state space E. A transition from a state i to a state j may induce the arrival of a customer. Let the sojourn time in state $i \in E$ be exponentially distributed with parameter $\lambda_i > 0$, and given that a transition takes place, let p_{ij} be the probability of a transition from i to a state $j \in E\backslash\{i\}$, $\sum_{j \in E\backslash\{i\}} p_{ij} = 1$. Defining $\lambda_{ij} = p_{ij}\lambda_i$, with $p_{ii} = -1$, then the generator of the Markov process $\{\mathbf{J}(t), t \geq 0\}$ is given by the matrix $D = (\lambda_{ij})$. Next, conditional on a transition from a state

i to a state j, the probability of an arrival is denoted by q_{ij}. With this notation, $p_{ij}q_{ij}\lambda_i$ is the transition rate from state i to state j inducing an arrival. Defining $q_{ii} = 0$, then the MAP is completely characterized by the matrices $D_0 = (\lambda_i p_{ij}(1 - q_{ij}))$ and $D_1 = (\lambda_i p_{ij} q_{ij})$.

The MAP is a special case of the Batch Markovian Arrival Process (BMAP), in which a transition from state i to state j generates a *batch arrival* of size k with probability q_{ij}^k, $k = 0, 1, \ldots$. The $BMAP/G/1$ queue is defined as the single server queue with the BMAP describing the arrival process of batch customers, where batches are served in FCFS order, and the service times of single customers are independent and identically distributed with a general distribution function $B_s(\cdot)$. The $BMAP/G/1$ has been studied in Lucantoni[6], in which one can find results for most of the performance measures of interest, such as the number of customers, waiting times, and the busy period.

Next we consider the impatience structure. If at the moment of an arrival of a customer the amount of work in the system is x, $(x \geq 0)$, then this customer will eventually receive service with probability $e^{-\alpha x}$, and will eventually leave the queue without receiving service with probability $1 - e^{-\alpha x}$. Hence the probability that a customer runs out of patience and leaves the queue before receiving service has an exponential distribution with parameter α.

We remark that for $\alpha \to \infty$ the model reduces to a $MAP/G/1/1$ loss system, and with $\alpha \to 0$ we return to the standard $MAP/G/1$ queue.

The workload process.

The workload of the system at time t is defined as the total remaining work of the customer presently in service plus the sum of the service times of the customers in the queue who will receive service. So, the service times of customers who will leave the system prematurely do not add to the amount of work in the system. Denote by $\{\mathbf{V}(t), t \geq 0\}$ the process describing the workload of the system, and let $\{\mathbf{J}(t), t \geq 0\}$ describe the state of the underlying Markov chain of the MAP. The finite state space of this Markov chain is denoted by E. The workload process is defined as the Markov process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$, which has state space $[0, \infty) \times E$.

Let $\{\mathbf{V}, \mathbf{J}\}$ be the random variable with distribution the stationary distribution of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$. In appendix A we show that a unique stationary distribution exists if $\alpha > 0$.

We start our analysis of the process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$ by writing down an equivalent of the Takács integro-differential equation.

First we describe the evolution of the workload process during a small period of time Δ.

Defining for $t \geq 0, x \geq 0, j \in E$: $V_j(t, x) = \Pr\{\mathbf{V}(t) \leq x, \mathbf{J}(t) = j\}$, then for $t > 0$, $x > 0$, and $\Delta > 0$ small,

$$\begin{aligned} V_j(t + \Delta, x) &= (1 - \lambda_j \Delta) V_j(t, x + \Delta) + \sum_{i \in E, i \neq j} \lambda_{ij}(1 - q_{ij}) \Delta V_i(t, x + \Delta) + \\ &\sum_{i \in E} \lambda_{ij} q_{ij} \Delta \int_{y=0}^{x} (e^{-\alpha y} B_s(x - y) + 1 - e^{-\alpha y}) dV_i(t, y) + o(\Delta). \end{aligned} \tag{2.1}$$

To avoid technical issues concerning continuity and differentiability of $V_j(t, x)$ we directly

turn to the LST of the amount of work. This LST is defined for $t > 0$ and $Re\,\omega \geq 0$ as

$$\Phi_j(t,\omega) = \mathrm{E}\{e^{-\mathbf{V}(t)\omega} I_{\{\mathbf{J}(t)=j\}}\} = \int_{x=0}^{\infty} e^{-\omega x} dV_j(t,x).$$

The technical concerns are mainly caused by the distribution function of the service time, which might not be absolutely continuous, e.g. when the service time is deterministic. For details we refer to Hasofer[2]. However, in an analogous way to the derivation of similar equations by Loynes[5] and Takács[7, p. 52-53], we obtain from (2.1)

$$\begin{aligned}\frac{\partial}{\partial t}\Phi_j(t,\omega) &= \omega\Phi_j(t,\omega) - \omega V_j(t,0) - \lambda_j\Phi_j(t,\omega) + \sum_{i\in E, i\neq j} \lambda_{ij}\Phi_i(t,\omega) - \\ &\quad \sum_{i\in E} \lambda_{ij} q_{ij}(1-\beta(\omega))\Phi_i(t,\omega+\alpha), \qquad t>0, Re\,\omega \geq 0. \end{aligned} \tag{2.2}$$

$\beta(\cdot)$ in (2.2) is the LST of the service time distribution B_s.

We proceed by investigating the steady state of the workload process. However, the analysis can be extended for the transient behaviour.

Next we define the Laplace-Stieltjes Transform (LST) related to the process $\{\mathbf{V},\mathbf{J}\}$:
$\Phi_j(\omega) = \int_{x=0}^{\infty} e^{-\omega x} dV_j(x)$, $Re\,\omega \geq 0$.
Letting $t \to \infty$, and using the vector notation $\Phi(\omega) = (\Phi_j(\omega))$ together with the vector $V(0)$ of probabilities $\lim_{t\to\infty} V_j(t,0)$, we derive from (2.2)

$$\Phi(\omega)[\omega I + D] = \omega V(0) + (1-\beta(\omega))\Phi(\omega+\alpha)D_1, \quad Re\,\omega \geq 0. \tag{2.3}$$

In (2.3), I is the identity-matrix of dimension $|E|$, $D = D_0 + D_1$ is the generator of $\{\mathbf{J}(t), t \geq 0\}$.

We remark that for $\alpha = 0$ equation (2.3) gives the LST equation for the stationary workload process in an ordinary $MAP/G/1$ queue, provided that this workload process is ergodic. For $\alpha \to \infty$ the model can be interpreted as the $MAP/G/1/1$ loss model; in particular for $M = 1$, equation (2.3) reduces to the LST equation of the stationary workload distribution in an $M/G/1/1$ loss model.

We proceed with solving equation (2.3) by iteration.

First, D is a generator of a Markov process on a finite state space E, hence the matrix $[\omega I + D]$ is non-singular for $Re\,\omega \geq 0$, except for at most $|E|$ values of ω. Defining

$$\zeta_\omega = \{\omega | Re\,\omega \geq 0, \mathrm{Det}[\omega I + D] \neq 0\}, \quad (\text{note that } 0 \notin \zeta_\omega),$$

we rewrite (2.3) for $\omega \in \zeta_\omega$ into

$$\Phi(\omega) = V(0)\omega[\omega I + D]^{-1} + \Phi(\omega+\alpha)D_1[\omega I + D]^{-1}(1-\beta(\omega)). \tag{2.4}$$

We remark that $\omega[\omega I + D]^{-1}$ is a matrix of rational functions and that the set ζ_ω contains the poles of these functions.

Defining $A(\omega) = \omega[\omega I + D]^{-1}$, and $B(\omega) = D_1[\omega I + D]^{-1}(1 - \beta(\omega))$, equation (2.4) can be rewritten into

$$\Phi(\omega) = V(0)A(\omega) + \Phi(\omega + \alpha)B(\omega), \quad \omega \in \zeta_\omega. \tag{2.5}$$

Next define $\omega_k = \omega + k\alpha$, $k = 0, 1, \ldots$, and $\prod_{l=0}^{k} B(\omega_l) = B(\omega_k) \ldots B(\omega_1)B(\omega_0)$, with the convention that an empty product ($k < 0$) is equal to I.

Iterating (2.5) K times we derive

$$\Phi(\omega) = V(0) \sum_{k=0}^{K} A(\omega_k) \prod_{l=0}^{k-1} B(\omega_l) + \Phi(\omega_{K+1}) \prod_{l=0}^{K} B(\omega_l), \quad \omega \in \hat{\zeta}_\omega, \tag{2.6}$$

with $\hat{\zeta}_\omega = \{\omega \in \mathbb{C} | Re\ \omega \geq 0, \omega_l \in \zeta_\omega, l = 0, 1, \ldots\}$. Since $\alpha > 0$, this set $\hat{\zeta}_\omega$ is finite.

ASSUMPTION 2.1 : For the sake of simplicity we assume in our analysis of $\{\mathbf{V}, \mathbf{J}\}$ that the eigenvalues of D are distinct. The analysis can be extended to the case of non-simple eigenvalues. Moreover, we claim without proof that a $MAP/G/1 + M$ queue for which this assumption does not hold, can be approximated arbitrarily close by a $MAP/G/1 + M$ queue for which D does satisfy this condition.

In proving the convergence of (2.6) for $K \to \infty$ we use the following lemma

LEMMA 2.1 :

(*i*) As $k \to \infty$, $A(\omega_k) \to I$, for $Re\ \omega \geq 0$.

(*ii*) For $K = 0, 1, \ldots$, $|| \prod_{l=0}^{K} B(\omega_l)|| \leq c_\omega \tau^K$, for $\omega \in \hat{\zeta}_\omega$, with $\tau < 1$, $c_\omega < \infty$, and where $||C|| = |E| \max_{i,j} |C_{ij}|$ is a matrix norm of $C \in \mathbb{C}^{|E| \mathrm{x} |E|}$.

PROOF:

(*i*) Let η_j, $j = 1, \ldots, |E|$, be the eigenvalues of D. Since D is a generator matrix of a non-degenerate Markov process, there exists a constant $\xi < \infty$ such that $|\eta_j| < \xi$. According to the spectral theorem for matrices (cf. Lancaster & Tismenetsky[3] p. 314), for $\omega \in \zeta_\omega$ there exist matrices $Z_j, j = 1, \ldots, |E|$ such that $[\omega I + D]^{-1} = \sum_{j=1}^{|E|} \frac{1}{\omega + \eta_j} Z_j$. Since $\alpha > 0$ and $\xi < \infty$, there exists a $K_0 > 0$ such that $\omega_k \in \zeta_\omega$ for $k > K_0$. Moreover, $\frac{\eta_j}{\omega_k} \to 0$, as $k \to \infty$. It readily follows that $||A(\omega_k) - I|| \to 0$. Applying proposition 1 on page 361 of Lancaster & Tismenetsky[3] finishes the proof of (i).

(ii) Next, define $\theta = \max_{i,j \in E} D_1(i,j)$. Then $||D_1|| = |E|\theta \leq |E| \max_{i \in E} \lambda_i$. From (i) also follows $[\omega_K I + D]^{-1} \to 0$ when $K \to \infty$, hence there exists a K_1 such that for all $k \geq K_1$, $||\sum_{j=1}^{|E|} \frac{1}{\omega_k + \eta_j} Z_j|| \leq \frac{\tau}{2\theta|E|}$ with $\tau < 1$. Since $||\cdot||$ is a matrix norm, $||C_1 C_2|| \leq ||C_1||.||C_2||$, for $C_1, C_2 \in \mathbb{C}^{|E| \times |E|}$. It follows that $||B(\omega_k)|| \leq \tau$, for $k \geq K_1$. Hence there exists a $c_\omega < \infty$, such that $||\prod_{l=0}^{K} B(\omega_l)|| \leq c_\omega \tau^{K+1}$ for all $k = 0, 1, \ldots$. □

As a corollary of lemma 2.1 we obtain

THEOREM 2.1 :

$$\Phi(\omega) = V(0) \sum_{k=0}^{\infty} \left(A(\omega_k) \prod_{l=0}^{k-1} B(\omega_l) \right), \quad \omega \in \hat{\zeta}_\omega, \quad \Phi(0) = \pi. \tag{2.7}$$

□

In (2.7), π is the stationary probability (row) vector of $\{\mathbf{J}(t), t \geq 0\}$, fulfilling $\pi D = 0$ and $\pi e = 1$, with e the $|E|$ dimensional unit (column) vector.

Next we derive $V(0) = (V_j(0))$, with $V_j(0)$ the steady state joint probability that the server is idle and $\mathbf{J}$ is in state j.

The $MAP/G/1 + M$ is ergodic for $\alpha > 0$ (cf. appendix A), hence equation (2.3) has a unique solution, as given by (2.7). Moreover, $V(0)$ is the only vector that fits (2.7) for all $\omega \in \mathbb{C}, Re\ \omega \geq 0$. The latter remains valid if we post-multiply both sides of (2.3) with the non-singular matrix $R = (r_j)$, r_j being the right (column) eigenvector of D, associated with eigenvalue η_j, $j = 1, \ldots, |E|$. In particular, $r_1 = e$ is the eigenvector for the eigenvalue $\eta_1 = 0$. The matrix R is non-singular because all eigenvalues of D are distinct (cf. Lancaster & Tismenetsky[3] pg. 153). Dividing by ω, observing that the limit $\omega \downarrow 0$ exists, results in the following set of equations

$$\Phi(\omega) \left(1 + \frac{\eta_j}{\omega}\right) r_j = V(0) r_j + \frac{1 - \beta(\omega)}{\omega} \Phi(\omega + \alpha) D_1 r_j, \quad j = 1, \ldots, |E|, \ Re\ \omega \geq 0. \tag{2.8}$$

Subsequently, we fixate $\omega = -\eta_j$ in the $j-th$ equation of (2.8). Rewriting the right hand side of (2.7) as $V(0)C(\omega)$, and defining constants $\beta_j = \lim_{\omega \downarrow -\eta_j} \frac{1-\beta(\omega)}{-\omega}$, then (2.8) becomes

$$V(0) r_j = I_{\{1=j\}} + [V(0) C(\alpha - \eta_j) \beta_j D_1] r_j, \quad j = 1, \ldots, |E|, \tag{2.9}$$

with $I_{\{\cdot\}}$ the indicator function.

The set of equations (2.9) has a unique positive solution. This follows from a probabilistic argument. Since the workload process is ergodic, $\Phi(\omega)$ is analytic for $\omega \in \{Re\ \omega \geq 0\}$.

Moreover, $\Phi(\omega)$ is unique. Since the solution of (2.9) establishes an analytic solution of the functional equation (2.3), the uniqueness of $\Phi(\omega)$ forces $V(0)$ to be the only solution of (2.9).

The vector $V(0)$ is obtained from (2.9) by post-multiplying the $j-th$ equation of (2.9) with l_j, the left (row) eigenvector of D, associated with r_j ($l_i r_j = I_{\{i=j\}}, i,j = 1,\ldots,|E|$), and applying $\sum_{j=1}^{|E|} r_j l_j = I$. Note that $l_1 = \pi$, since π is the unique solution of $xD = 0, xe = 1$.

Finally, we have obtained

THEOREM 2.2 :

$$V(0) = \pi \left[I - \sum_{j=1}^{|E|} C(\alpha - \eta_j)\beta_j D_1 r_j l_j \right]^{-1}. \tag{2.10}$$

□

REMARK 2.1 : For $\omega = -\eta_j$, $j = 2,\ldots,|E|$, $\Phi(\omega)$ can be derived by taking limits in (2.7), then the remaining $\omega \notin \hat{\zeta}_\omega$ can be obtained using relation (2.3). Since $\Phi(\omega)$ is analytic for $\omega \in \{Re\,\omega \geq 0\}$, these limits exist. □

REMARK 2.2 : For the analysis it is necessary that $\alpha \in \hat{\zeta}_\omega$. We assume that this is the case; by slightly altering the parameters we can approximate any $MAP/G/1+M$ queue arbitrarily closely by a $MAP/G/1+M$ queue for which this assumption does hold. □

REMARK 2.3 : We conclude this section by reflecting on other MAP generalisations of queueing models with impatient customers. In Baccelli et al.[1] the only other model that could be solved completely was the $M/M/1+D$ queue, i.e. the $M/M/1$ queue with deterministically distributed patience. Unfortunately its MAP equivalent results in a differential equation which seems hard to solve. The general $M/G/1+G$ queue was not explicitly solvable, so one could not expect more for the general $MAP/G/1+G$ queue. □

3 DERIVED RESULTS

Moments of $\{\mathbf{V},\mathbf{J}\}$.

The moments of $\{\mathbf{V},\mathbf{J}\}$ can be obtained from (2.3). These are not obtainable by differentiating in equation (2.7) because $0 \notin \hat{\zeta}_\omega$. Below we present the first moment, higher moments follow analogously.

Multiplying both sides of (2.3) with the unit vector e, using $De = 0$, dividing by ω, and taking derivatives with respect to ω, we obtain

$$\Phi'(\omega)e = \left(\frac{1-\beta(\omega)}{\omega}\right)' \Phi(\omega+\alpha)D_1 e + \left(\frac{1-\beta(\omega)}{\omega}\right) \Phi'(\omega+\alpha)D_1 e. \tag{3.1}$$

Letting $\omega \downarrow 0$ we find the mean total workload of the system

$$\Phi'(0)e = -\frac{\beta^{(2)}}{2}\Phi(\alpha)D_1 e + \beta\Phi'(\alpha)D_1 e, \tag{3.2}$$

in which β and $\beta^{(2)}$ respectively are the first and second moment of the service time distribution. $\Phi(\alpha)$ and $\Phi'(\alpha)$ can be derived from (2.7). By taking derivatives directly in equation (2.3), we also find an expression for $\Phi'(0)D$. Post-multiplying in (3.2) with π, and noting that $[e\pi + D]$ is non-singular, we finally obtain the mean workload vector

$$\mathbf{EV} = -\Phi'(0) = [\pi - V(0) - \beta\Phi(\alpha)D_1][e\pi + D]^{-1} + \left[\frac{\beta^{(2)}}{2}\Phi(\alpha) - \beta\Phi'(\alpha)\right] D_1 e\pi. \qquad (3.3)$$

REMARK 3.1 : As stated in the beginning of this section, for $\alpha \downarrow 0$, the $MAP/G/1+M$ queue reduces to the ordinary $MAP/G/1$ queue, although the $MAP/G/1+M$ queue is always stable for $\alpha > 0$ while the corresponding $MAP/G/1$ might not be. The stability condition for the ordinary $MAP/G/1$ queue is $\pi D_1 e\beta < 1$. Provided that this condition is satisfied, then, for $\alpha \downarrow 0$, the equation (2.3) for the LST's of the workload process holds for the $MAP/G/1$ queue. Moreover, the expressions (3.2) and (3.3) are also consistent with the ordinary $MAP/G/1$ queue when $\alpha \downarrow 0$. However, they do not yield the moments for the ordinary $MAP/G/1$ queue, since in the right hand sides of (3.2) and (3.3) the term $\Phi'(0)D_1$ appears for $\alpha \downarrow 0$. The moments of the workload vector can be obtained as follows: rewrite (2.3) for $\alpha \downarrow 0$ as $\Phi(\omega)[\omega I + D_0 + \beta(\omega)D_1] = \omega V(0)$, postmultiply with the eigenvector $r_1(\omega)$ which belongs to the smallest eigenvalue $\eta(\omega)$ of $[\omega I + D_0 + \beta(\omega)D_1]$. Taking derivatives with respect to ω and letting $\omega \downarrow 0$ gives an expression for $E\mathbf{V}$, containing the unknown vector $V(0)$. This vector can be obtained in the same way as for the $MAP/G/1+M$ queue. A second method to derive this vector is presented in Lucantoni [6], it involves the iteration of the matrix equivalent of the $M/G/1$ busy period equation. □

The waiting time.
$\Phi(\omega)$ is the LST vector of the virtual waiting time and the probability that $\mathbf{J}$ is in state j. The time between transitions in the Markov process is exponentially distributed, hence we can use the PASTA property to obtain the LST vector of the waiting time $\mathbf{W}$ and the probability that a customer prematurely leaves the system. Denote by $q_j = \sum_{i \in E} p_{ji} q_{ji}$ the probability that an arrival occurs given a transition out of state j, $j \in E$. Then, with $q = (q_j)$,

$$\begin{aligned} \Pr\{\mathbf{W} \le x\} &= \sum_{j \in E} \Pr\{\mathbf{V} \le x, \mathbf{J} = j | \text{arrival of a customer}\} \\ &= \sum_{j \in E} \frac{\Pr\{\mathbf{V} \le x, \mathbf{J} = j, \text{arrival of a customer}\}}{\Pr\{\text{arrival of a customer}\}} = \frac{V(x)q}{\pi q}, \qquad x \ge 0. \end{aligned}$$

From this we find $W(\omega)$, the LST of the waiting time of an arbitrary customer in steady state

$$W(\omega) = \frac{\Phi(\omega)q}{\pi q}, \quad Re\,\omega \ge 0. \qquad (3.4)$$

The probability that a customer leaves the system prematurely.
Analogous to the derivation of the LST for the waiting time we find P_α, the probability of a premature departure

$$P_\alpha = 1 - \int_{x=0}^{\infty} e^{-\alpha x} d\frac{V(x)q}{\pi q} = 1 - \frac{\Phi(\alpha)q}{\pi q} = 1 - W(\alpha). \tag{3.5}$$

ACKNOWLEDGEMENT: The author is grateful to Professors O.J. Boxma and J.W. Cohen for several useful discussions.

REFERENCES

[1] Baccelli, F., Boyer, P., Hebuterne, G. (1984). Single-server queues with impatient customers. *Adv. Appl. Prob.* **16**, 887-905.

[2] Hasofer, A.M. (1963). On the integrability, continuity and differentiability of a family of functions introduced by L. Takács. *Ann. Math. Statist.* **34**, 1045-1049.

[3] Lancaster, P., Tismenetsky, M. (1985). *The Theory of Matrices* (Academic Press, Orlando).

[4] Laslett, G.M., Pollard, D.M., Tweedie R.L. (1978). Techniques for establishing ergodic and recurrence properties of continuous-valued Markov chains. *Naval Res. Logist. Quart.* **25**, 455-472.

[5] Loynes, R.M. (1962). A continuous-time treatment of certain queues and infinite dams. *J. Austr. Math. Soc.* **2**, 484-498.

[6] Lucantoni, D.M. (1991). New results on the single server queue with a batch Markovian arrival process. *Commun. Statist.-Stochastic Models* **7**, 1-46.

[7] Takács, L. (1962). *Introduction to the Theory of Queues.* (Oxford University Press, New York).

APPENDIX

A ERGODICITY OF THE $MAP/G/1+M$ QUEUE.

In this appendix we prove that the workload process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$ of the $MAP/G/1+M$ queue is ergodic for $\alpha > 0$.

Intuitively this is clear, due to the impatience of the customers: as the amount of work in the system increases, the number of customers joining the queue gets smaller and smaller. Eventually, the work that joins the queue per unit of time is less than 1, hence there exists an $x_0 \geq 0$ such that the drift of the Markov process is towards the origin for all states $x \in [x_0, \infty) \times E$, ($[0, \infty) \times E$ is the state space of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$). The basic idea of the proof follows this intuition. If one can show the existence of a positive recurrent subset $A \subset [0, \infty) \times E$ then, under certain conditions, this is sufficient for the existence of a unique stationary distribution of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$.

The proof follows the scheme of Laslett et al.[4], to which we refer for details and technical issues.

First we mention a number of concepts that are used in Laslett et al.[4].

ϕ-irreducibility.
A Markov chain $\{X_n\}$ with state space χ is called *ϕ-irreducible* if there exists a nonzero measure ϕ on F (F is the σ-algebra of Borel subsets of χ), such that for any $x \in \chi$ and $A \in F$ with $\phi(A) > 0$, there exists an n for which $P^n(x, A) > 0$. $P^n(x, A)$ denotes the n-step transition probability of $\{X_n\}$.

Ergodicity.
If $\{X_n\}$ is ϕ-irreducible then there exists *at most one* stationary distribution, $\{X_n\}$ is *ergodic* if it has a *unique* stationary distribution. If $\{X_n\}$ is ergodic the n-step probability transitions converge to this stationary distribution in the strong Cesaro sense.

Hitting times.
Hitting times T_A are defined as $T_A = \inf\{n > 0 | X_n \in A\}$.

Test set.
Assume $\{X_n\}$ is ϕ-irreducible. Then, if $\sup_{x \in A} \mathrm{E}[T_A | X_0 = x] < \infty$ is a sufficient condition for $\{X_n\}$ to be ergodic, $A \in F$ is called a *test set.*

The main theorem for testing this condition for a set $A \in F$ is (cf. Theorem 2.1 of [4])

THEOREM A.1 : Let g be a nonnegative function on χ. If for some $\epsilon > 0$ and $A \in F$

$$\mathrm{E}[g(X_1) | X_0 = x] \leq g(x) - \epsilon, \text{ for } x \in A^c,$$

then

$$\sup_{x \in A} \mathrm{E}[T_A | X_0 = x] < \infty.$$

□

The scheme in [4] for proving that a Markov chain $\{X_n\}$ is ergodic consists of three steps:

1. Prove that $\{X_n\}$ is ϕ-irreducible.
2. Identify possible test sets.
3. Apply theorem A.1 to one of these sets.

For the workload process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$ of the $MAP/G/1{+}M$ queue, we prove ergodicity of the embedded Markov chain $\{(\mathbf{V}_n, \mathbf{J}_n), n = 0, 1, \ldots\} = \{(\mathbf{V}(t_n), \mathbf{J}(t_n)), n = 0, 1, \ldots\}$, where t_n is the time just before the $n-th$ transition in $\{\mathbf{J}(t), t \geq 0\}$. Applying the conditional PASTA property, we find that the stationary distribution of the embedded Markov chain equals the stationary distribution of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$.

Next we perform the three steps of the scheme.

1. We choose for ϕ an arbitrary nonzero measure on F, such that ϕ has an atom at $(0, i) \in [0, \infty) \times E$, where i is an arbitrary element of E. Since the Markov chain is irreducible

it follows that $(0,i)$ can be reached from any subset $A \subset [0,\infty) \times E$ in at most $|E|$ steps. Hence, $\{(\mathbf{V}_n, \mathbf{J}_n), n = 0,1,\ldots\}$ is ϕ-irreducible.

2. According to Theorem 3.2 of Laslett et al.[4], any set A containing $\{(0,i)\}$ is a test set if for some integer N and some $\delta > 0$, $\max_{n \le N} P^n(y,(0,i)) \ge \delta$, for all $y \in A$. With $N = |E|$ this condition holds when A is of the form $[0,x_0) \times E$, with $x_0 > 0$.

3. We apply theorem A.1, using the function $g((x,i)) = x$, $(x,i) \in [0,\infty) \times E$. t_0 is defined as the time just before the first transition in $\{\mathbf{J}(t), t \ge 0\}$. Denote by (x_0,i) the state of $(\mathbf{V}_0,\mathbf{J}_0)$ and $\tau_{(x_0,i)}$ the amount of work joining the queue at time t_0. Define by $\sigma_{(x_0,i)} = t_1 - t_0$ the time until the first jump after t_0 in $\{\mathbf{J}(t), t \ge 0\}$, conditioned on the state of $(\mathbf{V}_0,\mathbf{J}_0)$. Then

$$\mathrm{E}[g(\mathbf{V}_1,\mathbf{J}_1) - g(\mathbf{V}_0,\mathbf{J}_0) | (\mathbf{V}_0,\mathbf{J}_0) = (x_0,i)] = \mathrm{E}[\max[x_0 + \tau_{(x_0,i)} - \sigma_{(x_0,i)}, 0] - x_0] =$$

$$\mathrm{E}[\max[\tau_{(x_0,i)} - \sigma_{(x_0,i)}, -x_0]] = \int_{u=-\infty}^{-x_0} (-x_0) dH_{(x_0,i)}(u) + \int_{u=-x_0}^{\infty} u dH_{(x_0,i)}(u),$$

with $H_{(x_0,i)} = \tau_{(x_0,i)} - \sigma_{(x_0,i)}$, and $H_{(x_0,i)}(u) = \Pr\{H_{(x_0,i)} \le u\}$, $u \in \mathbb{R}$.

It is readily verified that $\mathrm{E}H_{(x_0,i)}$ is a continuous non-increasing function of x_0 when $\alpha > 0$. Finally, $\lim_{x_0 \to \infty} \mathrm{E}H_{(x_0,i)} \le -\frac{1}{\theta} < 0$, with $\theta = \max_{j \in E} \lambda_j$.

It follows that for $\alpha > 0$, $\exists x_0 \ge 0$ such that for all $i \in E$, $\mathrm{E}[g(\mathbf{V}_1,\mathbf{J}_1) - g(\mathbf{V}_0,\mathbf{J}_0) | (\mathbf{V}_0,\mathbf{J}_0) = (x_0,i)] < 0$, hence $\sup_{(x,i) \in A} \mathrm{E}[T_A | (\mathbf{V}_0,\mathbf{J}_0) = (x,i)] < \infty$, for $A = [0,x_0) \times E$ (we remark that in general x_0 depends on α). This completes the last step of our scheme.

Remark A.1: From equation (2.2) the ergodicity of the workload process can also be obtained via a more classical analysis method. This method involves the (double) LST of $V_j(t,x)$ with respect to t and x. The ergodicity of the workload process is proved by first showing that if $\alpha > 0$ the limit $\lim_{t \to \infty} V(t,0).e = V(0).e$ exist and that $V(0).e > 0$. From renewal theory then follows that the set of states $\{V_j(0), j \in E\}$ is positive recurrent, which completes the proof. □