J. Appl. Prob. **38,** 621–634 (2001)
Printed in Israel

ON THE STABILITY OF A BATCH CLEARING SYSTEM WITH POISSON ARRIVALS AND SUBADDITIVE SERVICE TIMES

DAVID ALDOUS,* *University of California, Berkeley*
MASAKIYO MIYAZAWA,** *Science University of Tokyo*
TOMASZ ROLSKI,*** *Wrocław University*

Abstract

We study a service system in which, in each service period, the server performs the current set B of tasks as a batch, taking time $s(B)$, where the function $s(\cdot)$ is subadditive. A natural definition of 'traffic intensity under congestion' in this setting is $\rho := \lim_{t\to\infty} t^{-1}\mathrm{E}s(\text{all tasks arriving during time } [0,t])$. We show that $\rho < 1$ and a finite mean of individual service times are necessary and sufficient to imply stability of the system. A key observation is that the numbers of arrivals during successive service periods form a Markov chain $\{A_n\}$, enabling us to apply classical regenerative techniques and to express the stationary distribution of the process in terms of the stationary distribution of $\{A_n\}$.

Keywords: Gated service discipline; job scheduling; queueing; regenerative process; stability; stochastic scheduling; subadditive

AMS 2000 Subject Classification: Primary 60K25; 90B36

1. Introduction

In a general model of a batch service system, tasks are presented to a *server* at random times. On completing a service, the server examines the set A of tasks to be done, and chooses (according to some strategy) a subset $B \subseteq A$ as the next batch of tasks to be accomplished. In many contexts, the service time $s(B)$ to accomplish task set B (for simplicity we assume service times are deterministic) will be a *subadditive* function of task sets:

$$s(B_1 \cup B_2) \le s(B_1) + s(B_2). \tag{1.1}$$

In particular, subadditivity is pervasive when a server must physically move (combining two trips into one trip saves time and distance) or where there is some start-up time for each new batch (so combining two batches eliminates one start-up time). For instance, consider the following examples:

Received 20 June 2000; revision received 4 April 2001.
* Postal address: Department of Statistics, University of California, Berkeley, CA 94720-3860, USA.
Research supported by NSF grant DMS 9970901.
** Postal address: Department of Information Sciences, Science University of Tokyo, Noda, Chiba 278-8510, Japan.
Research supported in part by JSPS under grant No. 13680532.
*** Postal address: Mathematical Institute, Wrocław University, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland.
Email address: rolski@math.uni.wroc.pl
Research supported in part by KBN under grant 2 P03A 049 15 (1998–2001).

- A retail store's delivery van. A 'task' is to deliver a package to a house.
- Thin client computing, that is, replacing a PC and purchased software applications by a cheaper device which downloads rented software from the Internet as needed. So a 'task' involves start-up time spent downloading some set of software.

Realistic modeling of any particular example will involve more specific structure (e.g. specific forms of $s(\cdot)$, capacity constraints). But can we say anything interesting when we assume only subadditivity for $s(\cdot)$? This mathematically natural question has apparently not been studied before, so we make a modest start here. We take a model (stated more precisely in Section 2) which is simple in other respects:

- single server;
- deterministic service times;
- Poisson arrivals (with general type-space).

In contrast to classical multiclass queueing theory which envisages a small number of customer classes (see e.g. [14, Chapter 10]), we envisage every task being different, that is, the type of each arrival may be chosen from some diffuse distribution.

Subadditivity as a *proof technique* is pervasive throughout modern applied probability. In the queueing context it has been used to study stability and Lyapunov exponents: see e.g. the work of Baccelli and his coworkers ([2] and references) regarding max-plus systems, and [3] regarding parallel processing systems. Our use of subadditivity is less standard since a weak *model hypothesis* is used in place of more structured ones.

Perhaps the most interesting questions about our model involve the server's choice of strategy, where we seek to minimize some long-run average cost per unit time; we outline some such questions in Section 5. Such long-run questions beg the more fundamental question of when the system is *stable*. In this paper we study the simple strategy in which the server adopts the entire set of waiting tasks as the next batch. This can be called a batch *clearing* system; or in the terminology of polling service systems a *gated* service discipline. Intuitively, stability should be closely related to the condition

$$\rho := \lim_{t\to\infty} t^{-1}\mathrm{E}s(\text{all tasks arriving during time } [0,t]) < 1 \tag{1.2}$$

because under this condition we expect that (for large t_0) all arrivals in an interval of duration t_0 can typically be served in the next interval of duration t_0, so that waiting times should not grow much beyond t_0. Our main result, Theorem 3.1, shows that the condition (1.2), together with finite expectation of the time $s(X_1)$ to serve a *single* customer, establishes stability (i.e. convergence to stationarity) of the queueing system as a whole, and hence of the usual characteristics such as service time, waiting time and queue length. These conditions are also necessary. Moreover, Corollary 4.2 shows that if $\mathrm{E}s^2(X_1) < \infty$, then the stationary waiting time or queue length have finite mean. So if there is a bounded waiting-cost function then (cf. Corollary 4.1) the asymptotic waiting cost per unit time is finite.

The case where $s(\cdot)$ is *additive* is essentially the $M/G/1$ queue, for which the process of arrivals during successive service periods is i.i.d. (see Section 5 for elaboration). Keys to our analysis are the observations (Lemmas 2.1 and 3.3) that in the subadditive case the process of arrivals during successive service periods is *Markov* and the mean service time of a batch is *finite*. In Section 3 we combine this with the observation that the process regenerates when empty, and deduce the convergence theorem.

Our model could be restated in the general framework of *state-dependent service* models. Such queueing models (in particular, polling systems [12]) have been widely studied. The regenerative technique is the standard way to prove stability—see e.g. [13] for a recent account of its queueing uses and [11] for the case of clearing systems—but it seems simpler to give direct regenerative proofs of our results than to adapt some other general set-up.

2. The model and first lemmas

We restate the model more carefully using the language of queueing theory. Consider a single-server queue with Poisson arrivals at rate λ. Customers are numbered as $1, 2, 3, \ldots$ according to their arrival times $0 < T_1 < T_2 < \cdots$. The nth customer has a task of type X_n, where $\{X_n; n = 1, 2, 3, \ldots\}$ are i.i.d. random variables (with some distribution Θ on some type-space $\mathcal{X}$, the details being irrelevant for our purposes). Service time is specified by a measurable function $s : \{\text{finite subsets of } \mathcal{X}\} \to [0, \infty)$ for which the key assumption is the subadditive property (1.1). We assume $s(\varnothing) = 0$ for the empty set $\varnothing$. It may also be natural to assume monotonicity:

$$\text{if } B_1 \subset B_2 \quad \text{then } s(B_1) \le s(B_2),$$

and nontriviality:

$$s(B) = 0 \quad \text{only if } B \text{ is empty.}$$

However, we shall not use these assumptions throughout the paper. Sets like B are really *multisets*; we won't labor the distinction. The verbal description of the batch clearing system translates into the following inductive description of the nth service period $[\gamma_n, \eta_n)$ and the index J_n of the final customer in the nth batch. For $n = 1, 2, \ldots,$

$$\begin{aligned}
\gamma_n &= \max(\eta_{n-1}, T_{J_{n-1}+1}),\\
J_n &= \max\{j : T_j \le \gamma_n\},\\
\eta_n &= \gamma_n + s(X_{J_{n-1}+1}, \ldots, X_{J_n}),
\end{aligned}$$

initialized by $\gamma_0 = \eta_0 = J_0 = 0$. Note that we write $s(X_1, \ldots, X_j)$ instead of $s(\{X_1, \ldots, X_j\})$. Consider

$$\begin{aligned}
A_n &= \text{number of arrivals during } n\text{th service period}\\
&= \max\{j : T_j < \eta_n\} - J_n,
\end{aligned}$$

setting $A_0 = 0$. This is almost the same as

$$\begin{aligned}
A'_n &= J_{n+1} - J_n\\
&= \text{size of } (n+1)\text{th batch served.}
\end{aligned}$$

The difference is that $A_n = 0$ implies that $A'_n = 1$; in other words,

$$A'_n = \max(1, A_n). \tag{2.1}$$

A key observation is that $\{A_n\}$ is Markov. This is intuitively clear: the number of arrivals during the $(n+1)$th service depends only on the duration of the $(n+1)$th service, which depends only on the number and types of arrivals during the nth service, but the types are independent of the number. We write the argument more carefully below.

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

Lemma 2.1. *The sequence* $\{A_n; n \geq 0\}$ *is the discrete-time Markov chain on states* $\{0, 1, 2, \dots\}$ *with* $A_0 = 0$ *and transition probabilities*

$$p_{ij} = \mathrm{E}\left(\frac{(\lambda s(X_1, \dots, X_{i'}))^j}{j!} \, \mathrm{e}^{-\lambda s(X_1, \dots, X_{i'})} \right), \quad \textit{where } i' = \max(1, i). \tag{2.2}$$

Hence the Markov chain $\{A_n\}$ *is irreducible and aperiodic.*

Proof. Write $\mathcal{G}_n = \sigma(J_1, \dots, J_{n+1}; X_1, \dots, X_{J_{n+1}})$ for the information known at the start of the $(n+1)$th service. So A_n and the duration $\eta_{n+1} - \gamma_{n+1}$ are $\mathcal{G}_n$-measurable. The conditional distribution of A_{n+1} given $\mathcal{G}_n$ is Poisson with mean $\eta_{n+1} - \gamma_{n+1} = s(X_{J_n+1}, \dots, X_{J_{n+1}})$. Write $\mathcal{F}_n = \sigma(\mathcal{G}_{n-1}, J_n + 1, \dots, J_{n+1})$, so that A_n is $\mathcal{F}_n$-measurable. Conditional on $\mathcal{F}_n$, $(X_{J_n+1}, \dots, X_{J_{n+1}})$ is distributed as $(\hat{X}_1, \dots, \hat{X}_{A'_n})$, where the $\hat{X}_i$ are independent copies of the X_i. So the conditional distribution of A_{n+1} given $\sigma(A_1, \dots, A_n) \subseteq \mathcal{F}_n$ is the Poisson mixture specified by the random parameter $s(\hat{X}_1, \dots, \hat{X}_{A'_n})$. This establishes the Markov property and the formula for transition probabilities.

Lemma 2.1 immediately implies that $\{A'_n, n \geq 0\}$ is also Markov. For $n = 1, 2, 3, \dots$ write S_n for the service time of the nth batch. So, as in the proof above,

$$\begin{aligned} S_{n+1} &= s(X_{J_n+1}, \dots, X_{J_{n+1}}) \\ &\stackrel{\mathrm{D}}{=} s(\hat{X}_1, \dots, \hat{X}_{A'_n}) \end{aligned} \tag{2.3}$$

and $(S_n, n \geq 1)$ is also a Markov chain. Related to S_n is

$$\begin{aligned} S'_n &= \text{time between start of } n\text{th and } (n+1)\text{th services} \\ &= \gamma_{n+1} - \gamma_n. \end{aligned}$$

Here $S'_n - S_n = 0$ unless $A_n = 0$, in which case $S'_n - S_n$ has exponential(λ) distribution, i.e., with mean $1/\lambda$. In particular

$$\mathrm{E}(S'_n - S_n) = \lambda^{-1} \mathrm{P}(A_n = 0). \tag{2.4}$$

We next note some consequences of subadditivity. Write

$$\begin{aligned} Y_n &= s(X_1, \dots, X_n), \\ f(n) &= \mathrm{E}Y_n. \end{aligned}$$

If $\mathrm{E}s(X_1) < \infty$, then by subadditivity of $s(\cdot)$ we have $\mathrm{E}Y_n \leq n\mathrm{E}Y_1 < \infty$. So $f(n)$ is finite-valued and subadditive:

$$f(n_1 + n_2) \leq f(n_1) + f(n_2).$$

Lemma 2.2. (i) *If* $\mathrm{E}Y_1 < \infty$, *then*

$$\lim_{n \to \infty} \frac{\mathrm{E}Y_n}{n} = \beta, \tag{2.5}$$

for some $0 \leq \beta < \infty$.

(ii) *For each* $k \geq 2$, *if* $\mathrm{E}Y_1^k < \infty$, *then*

$$\lim_{n \to \infty} \frac{\mathrm{E}Y_n^k}{n^k} = \beta^k. \tag{2.6}$$

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

Proof. Part (i) is a classical consequence of deterministic subadditivity (see e.g. Theorem 6.6.1(a) of [5]). For (ii), we first note that Y_n is subadditive, since $s(\cdot)$ is so. Kingman's subadditive ergodic theorem (see e.g. Theorem 6.6.1 of [5]) implies that

$$\lim_{n\to\infty} \frac{Y_n}{n} = \beta, \quad \text{a.s.}, \tag{2.7}$$

for the β defined by (i). Fix $a > \beta$ and write Y_n^k as

$$Y_n^k = Y_n^k \mathbf{1}(Y_n < na) + Y_n^k \mathbf{1}(Y_n \geq na),$$

where $\mathbf{1}(\cdot)$ is the indicator function. Using the bounded convergence theorem and (2.7), we have

$$\lim_{n\to\infty} \frac{\mathrm{E} Y_n^k \mathbf{1}(Y_n < na)}{n^k} = \beta^k.$$

On the other hand, the subadditivity of Y_n implies that

$$Y_n \leq \sum_{i=1}^n s(X_i).$$

This together with the convexity of x^k yields

$$\begin{aligned}
\mathrm{E}\left(\frac{Y_n^k}{n^k}\mathbf{1}(Y_n \geq na)\right) &\leq \mathrm{E}\left(\left(\frac{1}{n}\sum_{i=1}^n s(X_i)\right)^k; \sum_{i=1}^n s(X_i) \geq na\right) \\
&\leq \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^n s^k(X_i); \sum_{i=1}^n s(X_i) \geq na\right) \\
&= \mathrm{E}\left(s^k(X_1); \sum_{i=1}^n s(X_i) \geq na\right),
\end{aligned}$$

where the last equality holds because the $s(X_i)$s are i.i.d. Since we can choose any $a > \beta$, take $a > \mathrm{E}Y_1$. Then the law of large numbers implies that the last term of the above formula converges to 0. Thus we get (2.6).

It is straightforward to check that the congested traffic intensity ρ defined at (1.2) satisfies

$$\rho = \lambda\beta, \tag{2.8}$$

where λ is the Poisson arrival rate of customers and β is the mean congested service time per customer defined by (2.5).

3. The convergence theorem

In the following lemma, by positive-recurrence of a Markov chain we mean stability, that is, the existence of a limiting stationary distribution.

Lemma 3.1. *If $\rho < 1$ and $\mathrm{E}s(X_1) < \infty$, then $\{A_n; n \geq 0\}$, $\{A'_n; n \geq 1\}$, $\{S_n; n \geq 1\}$ and $\{S'_n; n \geq 1\}$ are positive-recurrent.*

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

Proof. Consider $\{A_n\}$, which is irreducible and aperiodic. We use Foster's theorem with test function $h(i) = i$ for a discrete-time Markov chain (see e.g. Theorem 5.1.1 in [4]). For $i \geq 1$

$$\begin{aligned} \mathrm{E}(A_{n+1} \mid A_n = i) - i &= \lambda \mathrm{E}(Y_i) - i \\ &= i\left(\lambda \frac{\mathrm{E}(Y_i)}{i} - 1\right) \\ &\leq i(\rho + o(1) - 1) \quad \text{(by (2.5), (2.8))} \\ &\to -\infty \quad \text{as } i \to \infty. \end{aligned}$$

The Lyapunov condition is thus satisfied, and so $\{A_n\}$ is positive-recurrent and converges in distribution to a stationary batch size:

$$A_n \xrightarrow{\mathrm{D}} A, \text{ say.}$$

From (2.1),

$$A'_n = \max(1, A_n) \xrightarrow{\mathrm{D}} \max(1, A) = A', \quad \text{say.}$$

Also

$$\begin{aligned} S_{n+1} &\stackrel{\mathrm{D}}{=} s(\hat{X}_1, \ldots, \hat{X}_{A'_n}) \quad \text{(by (2.3))} \\ &\xrightarrow{\mathrm{D}} s(\hat{X}_1, \ldots, \hat{X}_{A'}) \\ &\stackrel{\mathrm{D}}{=} S, \quad \text{say,} \end{aligned}$$

where S is therefore the stationary service time. Similarly, using the argument preceding (2.4),

$$(S_n, S'_n) \xrightarrow{\mathrm{D}} (S, S'),$$

where in particular the limit satisfies

$$\mathrm{E}(S' - S) = \lambda^{-1}\mathrm{P}(A = 0). \tag{3.1}$$

An obvious feature of our batch clearing system is that it regenerates each time an arriving customer finds an empty queue. The first such time is the first arrival time T_1. The next regeneration time τ is given by

$$\tau - T_1 = \sum_{n=1}^{N} S'_n,$$

where $N = \min\{n \geq 1 : A_n = 0\}$. By the regenerative cycle formula,

$$\mathrm{E}\sum_{n=1}^{N} S'_n = (\mathrm{E}N)(\mathrm{E}S'). \tag{3.2}$$

By positive-recurrence of $\{A_n\}$ (Lemma 3.1) we have $\mathrm{E}N < \infty$. We need to know that the hypotheses of Lemma 3.1 imply that $\mathrm{E}S' < \infty$; this is part of Lemma 3.3, whose statement and proof we defer. Granted that $\mathrm{E}S' < \infty$, we have shown that the mean duration $\mathrm{E}(\tau - T_1)$ of a regenerative cycle is finite. So we can apply classical results on regenerative processes.

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

To this end, we describe the state of the batch clearing system as follows. Write the state as $\xi = (u, C, B)$, where

u is the time since the starting instant of the latest service;

C is the set of types of customers being served;

B is the set of types of customers waiting.

Note that C is empty only when the system state is empty. For convenience, this empty state is denoted by $\varnothing$. Let $\Xi(t)$ be the system state at time t. The argument above is then summarized by the following lemma.

Lemma 3.2. *Under the assumptions of Lemma 3.1, the process $\Xi(t)$ has the stationary distribution given by*

$$\mathrm{P}(\Xi \in \cdot) = \frac{\mathrm{E}\int_{T_1}^{\tau} \mathbf{1}(\Xi(t) \in \cdot)\,\mathrm{d}t}{\mathrm{E}(\tau - T_1)}.$$

To state a more helpful expression for the stationary distribution, we introduce the following notation. Write $\#B$ for the size of set B. Write $\boldsymbol{X}(i)$ or $\boldsymbol{X}^*(i)$ for a random set distributed as $\{X_1, \ldots, X_i\}$. Write $\mathcal{P}(t)$ for a Poisson process with rate 1. We also write $\mathcal{C}$ and $\mathcal{B}$ for measurable subsets of the second and third components of the system state, respectively.

Theorem 3.1. *Suppose $\rho < 1$ and $\mathrm{E}s(X_1) < \infty$. Then $\mathrm{P}(\Xi(t) \in \cdot) \to \mathrm{P}(\Xi \in \cdot)$ as t goes to infinity, where $\to$ means the setwise convergence for all measurable subsets. The limit distribution is as follows. For each pair $\mathcal{C}$, $\mathcal{B}$ of measurable subsets and each integer $i \geq 1$,*

$$\begin{aligned}&\mathrm{P}(\Xi \in (\mathrm{d}u, \{C \in \mathcal{C}; \#C = i\}, \mathcal{B}))\\ &\quad = \frac{\mathrm{P}(A' = i, s(\boldsymbol{X}(i)) \geq u, \boldsymbol{X}(i) \in \mathcal{C}, \boldsymbol{X}^*(\mathcal{P}(\lambda u)) \in \mathcal{B})\,\mathrm{d}u}{\mathrm{E}S'}, \qquad 0 < u < \infty, \end{aligned} \tag{3.3}$$

where the random quantities A', $\boldsymbol{X}(i)$, $\boldsymbol{X}^*(i)$, $\mathcal{P}(\cdot)$ in the numerator are independent. Moreover

$$\mathrm{P}(\Xi = \varnothing) = 1 - \frac{\mathrm{E}S}{\mathrm{E}S'}. \tag{3.4}$$

Proof. The setwise convergence is immediate from the following observations. A regenerative process converges to its stationary version in the sense of the total variation as the time goes to infinity if the regeneration cycle has a finite expectation (see, e.g., Section III.18 of [7]), and we have verified this condition, provided $\mathrm{E}S' < \infty$. Thus, we only need to show (3.3) and (3.4). From the mean cycle formula concerning the service starting instants, we have

$$\mathrm{P}(\Xi \in \cdot) = \frac{1}{\mathrm{E}(\gamma_2 - \gamma_1)}\mathrm{E}\left(\int_{\gamma_1}^{\gamma_2} \mathbf{1}(\Xi(u) \in \cdot)\,\mathrm{d}u\right), \qquad t \geq 0. \tag{3.5}$$

Hence, from (3.5), we have, for $i \geq 1$,

$$\begin{aligned}&\mathrm{P}(\Xi \in ((0, t], \{C \in \mathcal{C}; \#C = i\}, \mathcal{B}))\\ &\quad = \frac{1}{\mathrm{E}S'}\mathrm{E}\int_0^{S'} \mathbf{1}(u \leq t \leq S, A' = i, \boldsymbol{X}(i) \in \mathcal{C}, \boldsymbol{X}(\mathcal{P}(\lambda u)) \in \mathcal{B})\,\mathrm{d}u\\ &\quad = \frac{1}{\mathrm{E}S'}\int_0^{t} \mathrm{P}(S \geq u, A' = i, \boldsymbol{X}(i) \in \mathcal{C}, \boldsymbol{X}(\mathcal{P}(\lambda u)) \in \mathcal{B})\,\mathrm{d}u,\end{aligned}$$

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

since $\mathrm{E}(\gamma_2 - \gamma_1) = \mathrm{E}(S')$, where $S = s(\boldsymbol{X}(i))$. This is equivalent to (3.3). We finally get (3.4) from

$$\begin{aligned}\mathrm{P}(\Xi = \varnothing) &= \frac{1}{\mathrm{E}S'}\mathrm{E}\int_0^{S'} \mathbf{1}(T_1 \geq u, \mathcal{P}(\lambda S) = 0)\,\mathrm{d}u \\ &= \frac{\mathrm{E}T_1}{\mathrm{E}S'}\mathrm{P}(\mathcal{P}(\lambda S) = 0) \\ &= 1 - \frac{\mathrm{E}S}{\mathrm{E}S'},\end{aligned}$$

where the last equality follows from $\mathrm{P}(\mathcal{P}(\lambda S) = 0) = \mathrm{P}(A = 0)$ and (2.4).

As part of the proof of Theorem 3.1 we needed to know that $\mathrm{E}S' < \infty$, which by (2.4) is equivalent to $\mathrm{E}S < \infty$. This follows from (3.1) and the $k = 1$ case of the next lemma.

Lemma 3.3. *Suppose $\rho < 1$. For each positive integer k, the following are equivalent:*

(i) $\mathrm{E}s^k(X_1) < \infty$;

(ii) $\mathrm{E}A^k < \infty$;

(iii) $\mathrm{E}S^k < \infty$.

Proof. Recall [5, p. 19] that the factorial moments of $\mathcal{P}(x)$ are

$$\mathrm{E}\mathcal{P}(x)(\mathcal{P}(x) - 1)\cdots(\mathcal{P}(x) - k + 1) = x^k. \tag{3.6}$$

Let $b > 0$, and recall that $Y_i = s(\boldsymbol{X}(i))$. From (2.2)

$$\mathrm{E}(A_{n+1}^k \wedge b) = \mathrm{P}(A_n = 0)\mathrm{E}(\mathcal{P}^k(\lambda Y_1) \wedge b) + \sum_{j=1}^{\infty}\mathrm{P}(A_n = j)\mathrm{E}(\mathcal{P}^k(\lambda Y_j) \wedge b), \tag{3.7}$$

with $\mathcal{P}(\cdot)$ independent of $\{Y_n\}$. Since $A_n \xrightarrow{\mathrm{D}} A$, letting $n \to \infty$ in the formula above yields

$$\mathrm{E}(A^k \wedge b) = \mathrm{P}(A = 0)\mathrm{E}(\mathcal{P}^k(\lambda Y_1) \wedge b) + \sum_{j=1}^{\infty}\mathrm{P}(A = j)\mathrm{E}(\mathcal{P}^k(\lambda Y_j) \wedge b). \tag{3.8}$$

Letting $b \to \infty$ shows that

$$\mathrm{E}(A^k \wedge b) \geq \mathrm{P}(A = 0)\mathrm{E}\mathcal{P}^k(\lambda Y_1) \geq \lambda^k \mathrm{E}Y_1^k.$$

So (ii) implies that $\mathrm{E}Y_1^k < \infty$, which is assertion (i). Conversely, suppose $\mathrm{E}Y_1^k < \infty$. From (3.6), for every $0 < \delta < 1$ we can find a d such that

$$\mathrm{E}(\mathcal{P}^k(x) \wedge b) \leq (\mathrm{E}\mathcal{P}^k(x)) \wedge b \leq (1 + \delta)(x^k \wedge b) + d.$$

This implies that

$$\mathrm{E}(\mathcal{P}(\lambda Y_j^k) \wedge b) \leq (1 + \delta)(\lambda^k \mathrm{E}Y_j^k \wedge b) + d < \infty.$$

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

Substituting into (3.8),

$$\mathrm{E}(A^k \wedge b) \le \mathrm{P}(A=0)((1+\delta)\lambda^k \mathrm{E}Y_1^k + d) + \sum_{j=1}^{\infty}((1+\delta)\lambda^k(\mathrm{E}Y_j^k \wedge b) + d)\mathrm{P}(A=j).$$

Using either (2.5) for $k=1$ or (2.6) for $k \ge 2$, we have for all ε and suitably chosen d'

$$\lambda^k \mathrm{E}Y_j^k \le \lambda^k (m^k + \varepsilon) j^k + d'.$$

Hence we conclude that there exist $0 < g < 1$ and $h > 0$ such that

$$\mathrm{E}(A^k \wedge b) \le g\mathrm{E}(A^k \wedge b) + h$$

for $n \ge 1$ and therefore

$$\mathrm{E}(A^k \wedge b) < \frac{h}{1-g} < \infty.$$

Letting $b \to \infty$ implies that $\mathrm{E}A^k < \infty$, which is (ii). The equivalence of (iii) follows from the fact that

$$\mathrm{E}S^k = \sum_{j=1}^{\infty} \mathrm{E}Y_j^k \mathrm{P}(A=j) \le \sum_{j=1}^{\infty}((m^k+\varepsilon)j^k + d')j^k \mathrm{P}(A=j),$$

and the corresponding lower bound with $-\varepsilon$.

For completeness, let us prove *necessity* in Theorem 3.1.

Proposition 3.1. *If $\Xi(t)$ converges in distribution, then $\rho < 1$ and $\mathrm{E}s(X_1) < \infty$.*

Proof. It is not difficult to see that convergence of $\Xi(t)$ to some limit Ξ implies that $\mathrm{P}(\Xi = \varnothing) > 0$. Indeed, for each state ξ, there is a nonrandom time $t_0 \ge 0$ (the time to serve all customers present) such that, for the process started at state ξ, $\Xi(t)$ attains the empty state with positive probability for each $t > t_0$. This remains true if the initial state is random. So any stationary distribution Ξ for the process must have $\mathrm{P}(\Xi = \varnothing) > 0$. Hence the inter-regeneration time $\tau - T_1$ has finite mean. Since $\tau - T_1$ is stochastically larger than $s(X_1)$, we deduce that $\mathrm{E}s(X_1) < \infty$. Moreover, in the notation of (3.2),

$$(\mathrm{E}N)(\mathrm{E}S') < \infty,$$

so that $\{A_n\}$ is positive-recurrent, implying that $A_n \xrightarrow{\mathrm{D}} A$ for some A, and

$$\mathrm{E}S = \mathrm{E}s(\boldsymbol{X}(A')) \le \mathrm{E}S' < \infty.$$

In the subadditive limit (2.5), $\beta = \inf_n(\mathrm{E}s(\boldsymbol{X}(n)))/n$, and so

$$\mathrm{E}s(\boldsymbol{X}(n)) \ge n\beta. \tag{3.9}$$

So $\beta \mathrm{E}A' \le \mathrm{E}s(\boldsymbol{X}(A')) < \infty$, implying that

$$\limsup_n \mathrm{E}A_n \le \mathrm{E}A \le \mathrm{E}A' < \infty. \tag{3.10}$$

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

But, as in (3.7),

$$\begin{aligned}
\mathrm{E}A_{n+1} &= \mathrm{P}(A_n = 0)\lambda \mathrm{E}s(X_1) + \sum_{j=1}^{\infty} \lambda \mathrm{E}s(\boldsymbol{X}(A_n))\mathbf{1}(A_n = j) \\
&\geq \mathrm{P}(A_n = 0)\lambda\beta + \lambda\beta \sum_{j=1}^{\infty} \mathrm{E}A_n \mathbf{1}(A_n = j) \quad \text{(by (3.9))} \\
&= \rho \mathrm{P}(A_n = 0) + \rho \mathrm{E}A_n.
\end{aligned}$$

If $\rho \geq 1$, summing over all $n \geq 1$ yields

$$\begin{aligned}
\liminf_{n\to\infty} \mathrm{E}A_n &\geq \rho \sum_{n=1}^{\infty} \mathrm{P}(A_n = 0) \\
&= \infty
\end{aligned}$$

because $\mathrm{P}(A_n = 0) \to \mathrm{P}(A = 0) > 0$. But this contradicts (3.10), so we must instead have $\rho < 1$.

For the discrete-time chain $\{A_n\}$ the situation is more complicated, since $\rho < 1$ may not be necessary for $\{A_n\}$ to be positive-recurrent (see [10] for the additive case). We state here a partial result.

Proposition 3.2. *The chain* $\{A_n\}$ *is transient if there exist a* $\theta_0 \in (0, 1]$ *and an* $\varepsilon > 0$ *such that*

$$\limsup_{n\to\infty} \mathrm{E}(\mathrm{e}^{-\theta_0 Y_n + \theta_0(1+\varepsilon)n}) \leq 1. \tag{3.11}$$

The proof is given in Appendix A. It is easy to see that (3.11) implies that $\rho > 1$. For some $s(\cdot)$, in particular if $s(\cdot)$ is additive, (3.11) is equivalent to $\rho > 1$, provided $\mathrm{E}s(X_1) < \infty$.

4. Characteristics of the stationary distribution

Assume now that $\rho < 1$ and $\mathrm{E}s(X_1) < \infty$, so we are in the setting of Theorem 3.1. Let us elaborate the model by introducing a *waiting cost* function $c : \mathcal{X} \to (0, \infty)$, where $c(x)$ is interpreted as a waiting cost per unit time for a type-x customer, incurred from arrival until service is complete. For a set B write $c(B) = \sum_{x\in B} c(x)$. So the instantaneous cost rate associated with a state ξ is

$$\hat{c}(\xi) = \begin{cases} 0 & \text{if } \xi = \varnothing, \\ c(C) + c(B) & \text{if } \xi = (u, C, B). \end{cases}$$

In the setting of Theorem 3.1 the system has a long run average waiting cost per unit time given by

$$\bar{c} = \mathrm{E}\hat{c}(\Xi).$$

Corollary 4.1. *We have*

$$\bar{c} = \frac{(\lambda/2)(\mathrm{E}S^2)(\mathrm{E}c(X_1)) + \mathrm{E}[c(\boldsymbol{X}(A'))s(\boldsymbol{X}(A'))]}{\mathrm{E}S'}.$$

In particular, if $\mathrm{E}s^2(X_1) < \infty$ *and* $\mathrm{E}c^2(X_1) < \infty$, *then* $\bar{c} < \infty$.

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

Proof. The formula can be established by integrating over the distribution (3.3) of Ξ. More intuitively, consider a typical S'-interval. The first term in the numerator is the mean total cost over the interval associated with new customers arriving during the interval, while the second term is the cost associated with the customers being served. Because S'-intervals occur at rate $1/\mathrm{E}S'$, a regenerative argument rederives the formula.

If $\mathrm{E}s^2(X_1) < \infty$, then Lemma 3.3 implies that $S \stackrel{\mathrm{D}}{=} s(\boldsymbol{X}(A'))$ has finite second moment; similarly, if $\mathrm{E}c^2(X_1) < \infty$, then $c(\boldsymbol{X}(A'))$ has finite second moment; and so when both conditions hold we have $\bar{c} < \infty$.

A natural characteristic of the batch clearing system is the queue length process $\{L(t)\}$. Of course, Theorem 3.1 implies that as $t \to \infty$ this characteristic converges in distribution to the stationary distribution L, and we can write expressions in the spirit of (3.3) for their distributions. Note that the special case $c(\cdot) \equiv 1$ of Corollary 4.1 gives the stationary mean queue length, which is related to the mean stationary sojourn time of a customer by Little's law. Thus, writing W for the stationary sojourn time, we have the next corollary.

Corollary 4.2. *We have*

$$\mathrm{E}L = \lambda \mathrm{E}W = \frac{(\lambda/2)\mathrm{E}S^2 + \mathrm{E}[A's(\boldsymbol{X}(A'))]}{\mathrm{E}S'}.$$

In particular, Lemma 3.3 implies that $\mathrm{E}L$ *(or* $\mathrm{E}W$*) is finite if and only if* $\mathrm{E}s^2(X_1) < \infty$.

As in classical queueing theory, we expect that the kth moments of L and W are finite if and only if $\mathrm{E}s^{k+1}(X_1) < \infty$, and this can be verified in our model (see Appendix B for their verifications).

5. Discussion

The requirement that service times be deterministic is in fact no restriction. Random service times could be represented as $s(X_1, \ldots, X_i, U_i)$, where as before the Xs are i.i.d. with some distribution Θ on some type-space $\mathcal{X}$, and now the U_i are independent $U(0, 1)$. Subadditivity is now defined via the usual stochastic partial order on probability measures on $[0, \infty)$. But an exercise in measure theory (which we leave to the reader) shows that, given any such $s(\cdot)$, we can find an enlarged type-space $\hat{\mathcal{X}}$ and i.i.d. $\hat{\mathcal{X}}$-valued random variables $\hat{X}_i$ and a subadditive function $\hat{s}(\cdot)$ such that

$$s(X_1, \ldots, X_i, U_i) \stackrel{\mathrm{D}}{=} \hat{s}(\hat{X}_1, \ldots, \hat{X}_i), \qquad i = 1, 2, \ldots .$$

We take X_n as the type of a customer. It is also natural to consider it as the original service time of the customer. In this case the type-space is $(0, \infty)$ (note that this identification cannot be made in the general subadditive case). A typical service function $s(\cdot)$ of this case is a linear function, i.e. for some nonnegative constant $a > 0$,

$$s(X_1, \ldots, X_i) = a + X_1 + \cdots + X_i.$$

In particular, when $s(\cdot)$ is *additive* (i.e. $a = 0$), our model is essentially the $M/G/1$ queue as mentioned in Section 1. More precisely, consider the usual Galton–Watson branching process associated with the $M/G/1$ queue, in which arrivals during one customer's service are the children of that customer. Then a batch service interval in our additive model corresponds to

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

the time taken to serve all members of one generation in the $M/G/1$ queue. And the server's busy periods are identical in the two processes. See [10], [12] for related work.

As also mentioned in Section 1, perhaps the most interesting questions about the model involve the server's choice of strategy. Consider the setting of Corollary 4.1 where the 'clearing' algorithm CLEAR has some mean cost per unit time $\bar{c}(\text{CLEAR}) < \infty$. There will be some optimal strategy OPT (depending on $c(\cdot)$, $s(\cdot)$, λ and the type-distribution Θ) such that $\bar{c}(\text{OPT})$ is the minimal cost over all strategies. It is not hard to give an example to show that $\bar{c}(\text{CLEAR})/\bar{c}(\text{OPT})$ is not bounded by any absolute constant (there is an example with two types of customer and $c(\cdot) \equiv 1$), so that the 'clearing' strategy may be inefficient. Calculating the exact optimal strategy at any level of generality seems hopeless. But in the spirit of the *competitive analysis of algorithms* [6] we can ask if there exists any simple-to-describe strategy STRAT such that

$$\bar{c}(\text{STRAT})/\bar{c}(\text{OPT}) \text{ is bounded by some constant.} \tag{5.1}$$

In a first draft of this paper we conjectured that a greedy algorithm

$$\text{choose as the next batch the subset } B \text{ of current tasks that maximizes } \sum_{x \in B} \frac{c(x)}{s(B)}$$

might satisfy (5.1), but John Tsitsiklis (private communication) gave an example where this greedy strategy is not even stable.

Appendix A.

We give here the proof of Proposition 3.2. We first state the following lemma.

Lemma A.1. *Let*

$$\phi_n(t) = -\log \mathrm{E}(\mathrm{e}^{-tY_n}), \qquad n = 1, 2, \ldots .$$

Then

(i) *for each* $t \geq 0$ *the sequence* $\{\phi_n(t), n = 1, 2, \ldots\}$ *is subadditive;*

(ii) *the function* $\phi_n(t)$ *is concave;*

(iii) *the limit*

$$\phi(t) = \inf_{n \geq 1} \frac{\phi_n(t)}{n} = \lim_{n \to \infty} \frac{\phi_n(t)}{n}$$

exists and is increasing and concave.

Proof. Since $s(\cdot)$ is subadditive we have

$$\mathrm{e}^{-ts(X_1, \ldots, X_{n+n'})} \geq \mathrm{e}^{-ts(X_1, \ldots, X_n)}\, \mathrm{e}^{-ts(X_{n+1}, \ldots, X_{n+n'})}.$$

Hence

$$\mathrm{E}(\mathrm{e}^{-tY_{n+n'}}) \geq \mathrm{E}(\mathrm{e}^{-tY_n})\mathrm{E}(\mathrm{e}^{-tY_{n'}}).$$

Taking the logarithms of both sides and multiplying by -1 shows the subadditivity. Laplace transforms are logarithmically convex. Hence $\phi_n(t)/n$ is concave and $\phi(t)$ is concave as the infimum of concave functions.

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

Proof of Proposition 3.2. Without loss of generality we assume that $\lambda = 1$. We use Theorem 8.4.2 of [8] (or Theorem 3.7 of [4, Chapter 5] but the test function must be negative) with $h(n) = 1 - \theta^n$, $F = \{0, 1, \ldots, n_0\}$, where $0 < \theta < 1$ and n_0 are suitably chosen. We have

$$\sum_{k=0}^{\infty} p_{nk} h(k) = \sum_{k=0}^{\infty} (1 - \theta^k) \mathrm{E}\left(\frac{Y_n^k}{k!} \, \mathrm{e}^{-Y_n} \right) = 1 - \mathrm{E}(\mathrm{e}^{-(1-\theta)Y_n})$$

and therefore we want

$$1 - \mathrm{E}(\mathrm{e}^{-(1-\theta)Y_n}) > 1 - \theta^n, \qquad n > n_0, \tag{A.1}$$

which implies that A_n is transient. We show now how to choose θ and n_0. From (iii) of Lemma A.1, the limit $\phi(t)$ of $\phi_n(t)/n$ exists. From (3.11), there exists i_0 such that

$$-\frac{1}{n} \log \mathrm{E}(\mathrm{e}^{-\theta_0 Y_n}) \geq \theta_0 (1 + \varepsilon), \qquad \forall n \geq i_0.$$

Hence, writing $\theta_0 = 1 - \theta_1$, we have

$$\phi(1 - \theta_1) = \lim_{n \to \infty} -\frac{1}{n} \log \mathrm{E}(\mathrm{e}^{-\theta_0 Y_n}) > \theta_0 = 1 - \theta_1.$$

Since $\phi(1 - \theta)$ is concave, this implies that the left-hand derivative of $\phi(1 - \theta)$ at $\theta = 1$ is less than -1. Hence the concavity of $-\log\theta$ together with its derivative at $\theta = 1$ yields that there exists a positive $\theta_2 < \theta_1$ such that

$$\phi(1 - \theta_2) > -\log\theta_2.$$

Then, for $n_0 = 0$,

$$\frac{1}{n} \phi_n(1 - \theta_2) \geq \phi(1 - \theta_2) > -\log\theta_2, \qquad n > n_0,$$

which yields (A.1) for $\theta = \theta_2$.

Appendix B.

We first consider a distributional relationship between L and A. To this end, we apply the rate conservation law to the process $U(t) \equiv z^{L^*(t)}$ with $0 \leq z \leq 1$, assuming $\{L^*(t)\}$ to be a stationary version of the queue length process $\{L(t)\}$ (see e.g. [9] for the rate conservation law). Since $U(t)$ has jumps at arrival instants as well as service completion instants, we have, using PASTA (see e.g. [14]),

$$\lambda \mathrm{E}(z^L - z^{(L+1)}) + \nu \mathrm{E}(z^{(A' + \mathcal{P}(\lambda Y_{A'}))} - z^{\mathcal{P}(\lambda Y_{A'})}) = 0,$$

where ν is the mean departure rate of the batches. This yields

$$\lambda \mathrm{E}(z^L (z - 1)) = \nu \mathrm{E}(z^{\mathcal{P}(\lambda Y_{A'})} (z^{A'} - 1)). \tag{B.1}$$

Let $B(j, k)$ be $j!/(j - k)!$ for $j \geq k$ and 0 otherwise. For $z < 1$, differentiate both sides of (B.1) $k + 1$ times. Then, letting z go to 1 yields

$$\lambda (k + 1) \mathrm{E}B(L, k) = \nu \sum_{\ell=0}^{k} \binom{k+1}{\ell} \mathrm{E}B(\mathcal{P}(\lambda Y_{A'}), \ell) B(A', k + 1 - \ell). \tag{B.2}$$

https://doi.org/10.1239/jap/1005091027 Published online by Cambridge University Press

In particular, for $k = 0$, $\lambda = \nu \mathrm{E}A'$. Since

$$\begin{aligned}\mathrm{E}A' &= \mathrm{P}(A = 0) + \mathrm{E}(A) \\ &= \lambda \left(\mathrm{P}(A = 0)\frac{1}{\lambda} + \mathrm{E}(Y_{A'}) \right) = \lambda \mathrm{E}S',\end{aligned}$$

we have $\nu = 1/\mathrm{E}S'$ as expected. For $k = 1$, (B.2) obviously leads to Corollary 4.2. From (B.2), it is also not hard to see that, for any positive integer k, $\mathrm{E}L^k < \infty$ if and only if $\mathrm{E}A^{k+1} < \infty$, which is equivalent to $\mathrm{E}s^{k+1}(X_1) < \infty$ by Lemma 3.3.

For the stationary sojourn time W for a customer, we decompose it as

$$W = W_{\mathrm{Q}} + S,$$

where W_{Q} is the waiting time before service. Obviously we can use PASTA for W_{Q}, so it has the same distribution as the remaining service time of a batch at an arbitrary point in time. From Theorem 3.1, it is easy to see that the latter has finite kth moment if and only if $\mathrm{E}S^{k+1} < \infty$. Hence, using the inequality

$$x^k \le (x + y)^k \le 2^{k-1}(x^k + y^k), \qquad x, y \ge 0,$$

we have that $\mathrm{E}W^k < \infty$ if and only if $\mathrm{E}S^{k+1} < \infty$.

Acknowledgements

The paper was written in part while the third author was staying at the Department of Mathematical and Computing Sciences of Tokyo Institute of Technology. TR wishes to thank the Department and Professor Yukio Takahashi for their hospitality.

References

[1] Baccelli, F. and Brémaud, P. (1994). *Elements of Queueing Theory*. Springer, Berlin.
[2] Baccelli, F. and Hong, D. (2000). Analytic expansions of max-plus Lyapunov exponents. *Ann. Appl. Prob.* **10,** 779–827.
[3] Bambos, N. and Walrand, J. (1991). On the stability and performance of parallel processing systems. *J. Assoc. Comput. Mach.* **38,** 429–452.
[4] Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York.
[5] Durrett, R. (1991). *Probability: Theory and Examples*. Wadsworth and Brooks, Pacific Grove, CA.
[6] Fiat, A. and Woeginger, G. J. (eds) (1998). *Online Algorithms* (Lecture Notes Comput. Sci. **1442**). Springer, Berlin.
[7] Lindvall, T. (1992). *Lectures on the Coupling Method*. John Wiley, New York.
[8] Meyn, S. P. and Tweedie, R. D. (1993). *Markov Chains and Stochastic Stability*. Springer, London.
[9] Miyazawa, M. (1994). Rate conservation laws: a survey. *Queueing Systems* **15,** 1–58.
[10] Pakes, A. G. (1971). A branching process with a state dependent immigration component. *Adv. Appl. Prob.* **3,** 301–314.
[11] Perry, D. and Stadje, W. (2001). A stochastic clearing model with a Brownian and a compound Poisson component. Unpublished manuscript.
[12] Resing, J. A. C. (1993). Polling systems and branching processes. *Queueing Systems* **13,** 409–426.
[13] Sharma, V. (1998). Some limit theorems for regenerative queues. *Queueing Systems* **30,** 341–363.
[14] Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Prentice Hall, New Jersey.