On the stability of a batch clearing system with Poisson arrivals and subadditive service times *

David Aldous † Masakiyo Miyazawa ‡ Tomasz Rolski §

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

$$\rho := \lim_{t\to\infty} t^{-1} Es(\text{all tasks arriving during time } [0,t]).$$

We show that $\rho < 1$, and 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

---

*First published in Journal of Applied Probability, **38.3** ©2001 by The Applied Probability Trust.

†Department of Statistics, University of California, Berkeley CA 94720-3860, U.S.A. aldous@stat.berkeley.edu. Research supported by N.S.F Grant DMS 9970901

‡Department of Information Sciences, Science University of Tokyo, Noda, Chiba 278-8510, Japan. miyazawa@is.noda.sut.ac.jp. Research is supported in part by JSPS under grant No. 13680532.

§Mathematical Institute, Wrocław University, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland. rolski@math.uni.wroc.pl. Research supported in part by KBN under grant 2 P03A 049 15 (1998–2001).

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:

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

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

- 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 multi-class 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 Lyapounov exponents: see e.g. Baccelli and his coworkers ([2] and references) regarding max-plus systems, and Bambos and Walrand [3] regarding parallel processing systems. But our use of subadditivity is less standard since it is used a weak *model hypothesis* in place of more structured ones.

Perhaps the most interesting questions about our model involve the server's choice of strategy, where one seeks to minimize some long-run average cost per unit time, and we outline some such questions in Section 5. But such long-run questions beg the more foundational 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

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

because under this condition we expect that (for large t_0) all arrivals in one 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 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 $Es^2(X_1) < \infty$ 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 (see Section 5 for elaboration) the $M/G/1$ queue, for which the process of arrivals during successive service periods is i.i.d. 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 $0 < T_1 < T_2 < \ldots$. The n-th customer has 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 $s : \{$ finite subsets of $\mathcal{X}\} \to [0, \infty)$ for which the key assumption is the subadditive property (1.1). We assume $s(\emptyset) = 0$ for the empty set $\emptyset$. It may be also natural to assume monotonicity:

$$\text{if } B_1 \subset B_2 \text{ then } s(B_1) \leq s(B_2) \tag{2.1}$$

and non-triviality

$$s(B) = 0 \text{ only if } B \text{ is empty.} \tag{2.2}$$

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 n'th service period $[\gamma_n, \eta_n)$ and the index J_n of the final customer in the n'th batch. For $n = 1, 2, \ldots$,

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

initialized by $\gamma_0 = \eta_0 = J_0 = 0$. Note we write $s(X_1, \ldots, X_j)$ in place 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

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

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

$$A'_n = \max(1, A_n). \tag{2.3}$$

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

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

$$p_{ij} = E\left(\frac{(\lambda s(X_1, \ldots, X_{i'}))^j}{j!} e^{-\lambda s(X_1, \ldots, X_{i'})}\right), \quad \text{where } i' = \max(1, i). \tag{2.4}$$

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

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

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

$$\begin{aligned} S_{n+1} &= s(X_{J_n+1}, \ldots, X_{J_{n+1}}) \\ &\stackrel{d}{=} s(\hat{X}_1, \ldots, \hat{X}_{A'_n}) \end{aligned} \tag{2.5}$$

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

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

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

$$E(S'_n - S_n) = \lambda^{-1} P(A_n = 0). \tag{2.6}$$

We next note some consequences of subadditivity. Write

$$\begin{aligned} Y_n &= s(X_1, \ldots, X_n) \\ f(n) &= EY_n. \end{aligned}$$

If $Es(X_1) < \infty$, then by subadditivity of $s(\cdot)$ we have $EY_n \leq nEY_1 < \infty$. 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 $EY_1 < \infty$ then*

$$\lim_{n \to \infty} \frac{EY_n}{n} = \beta, \tag{2.7}$$

for some $0 \leq \beta < \infty$.
(ii) *For each $k \geq 2$, if $EY_1^k < \infty$ then*

$$\lim_{n \to \infty} \frac{EY_n^k}{n^k} = \beta^k. \tag{2.8}$$

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

$$\lim_{n\to\infty} \frac{Y_n}{n} = \beta, \qquad \text{a.s.}, \tag{2.9}$$

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

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

where $1(\cdot)$ is the indicator function for the statement "$\cdot$". Using the bounded convergence theorem and (2.9), we have

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

On the other side, the subadditivity of Y_n implies

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

This together with the convexity of x^k yields

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

where the last equality holds because the $s(X_i)$'s are independent identically distributed. Since we can choose any $a > \beta$, take $a > EY_1$. Then the law of the large numbers implies that the last term of the above formula converges to 0. Thus we get (2.8). □

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

$$\rho = \lambda\beta \tag{2.10}$$

where λ is the Poisson arrival rate of customers and β is the mean congested service time per customer defined by (2.7).

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 $Es(X_1) < \infty$, then $\{A_n; n \geq 0\}$ and $\{A'_n; n \geq 1\}$ and $\{S_n; n \geq 1\}$ and $\{S'_n; n \geq 1\}$ are positive recurrent.

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 $i \geq 1$

$$\begin{aligned} E(A_{n+1}|A_n = i) - i &= \lambda E(Y_i) - i \\ &= i\left(\lambda \frac{E(Y_i)}{i} - 1\right) \\ &\leq i(\rho + o(1) - 1) \qquad \text{by } (2.7, 2.10) \\ &\to -\infty \text{ as } i \to \infty, \end{aligned}$$

and thus the Lyapunov condition is satisfied. So $\{A_n\}$ is positive-recurrent and converges in distribution to a stationary batch size:

$$A_n \xrightarrow{d} A, \text{ say.}$$

From (2.3)

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

Also

$$\begin{aligned} S_{n+1} &\stackrel{d}{=} s(\hat{X}_1, \ldots, \hat{X}_{A'_n}) \qquad \text{by } (2.5) \\ &\xrightarrow{d} s(\hat{X}_1, \ldots, \hat{X}_{A'}) \\ &\stackrel{d}{=} S \text{ say,} \end{aligned}$$

where S is therefore the stationary service time. Similarly, using the argument above (2.6)

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

where in particular the limit satisfies

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

□

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

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

By positive-recurrence of $\{A_n\}$ (Lemma 3.1) we have $EN < \infty$. We need to know that the hypotheses of Lemma 3.1 imply $ES' < \infty$; this is part of Lemma 3.3, whose statement and proof we defer. Granted that $ES' < \infty$, we have shown that the mean duration $E(\tau - T_1)$ of a regenerative cycle is finite. So we can apply classical results on regenerative processes. 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 $\emptyset$. Let $\Xi(t)$ be the system state at time t. Then, the argument above is summarized by:

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

$$P(\Xi \in \cdot) = \frac{E \int_{T_1}^{\tau} 1(\Xi(t) \in \cdot)\, dt}{E(\tau - T_1)}. \tag{3.3}$$

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 $Es(X_1) < \infty$. Then $P(\Xi(t) \in \cdot) \to P(\Xi \in \cdot)$ 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 $i \geq 1$,*

$$\begin{aligned} &P(\Xi \in (du, \{C \in \mathcal{C}; \#C = i\}, \mathcal{B})) \\ &\quad = \frac{P(A' = i, s(\boldsymbol{X}(i)) \geq u, \boldsymbol{X}(i) \in \mathcal{C}, \boldsymbol{X}^*(\mathcal{P}(\lambda u)) \in \mathcal{B})\ du}{ES'}, \quad 0 < u < \infty, \end{aligned} \tag{3.4}$$

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

$$P(\Xi = \emptyset) = 1 - \frac{ES}{ES'}. \tag{3.5}$$

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 $ES' < \infty$. Thus we only need to show (3.4) and (3.5). From the mean cycle formula concerning the service starting instants, we have

$$P(\Xi \in \cdot) = \frac{1}{E(\gamma_2 - \gamma_1)} E\left(\int_{\gamma_1}^{\gamma_2} 1(\Xi(u) \in \cdot) du\right), \qquad t \geq 0. \tag{3.6}$$

Hence, from (3.6), we have, for $i \geq 1$,

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

since $E(\gamma_2 - \gamma_1) = E(S')$, where $S = s(\boldsymbol{X}(i))$. This is equivalent to (3.4). We finally get (3.5) from

$$\begin{aligned} P(\Xi = \emptyset) &= \frac{1}{ES'} E \int_0^{S'} 1(T_1 \geq u, \mathcal{P}(\lambda S) = 0)\ du \\ &= \frac{ET_1}{ES'} P(\mathcal{P}(\lambda S) = 0) \\ &= 1 - \frac{ES}{ES'}, \end{aligned}$$

where the last equality follows from $P(\mathcal{P}(\lambda S) = 0) = P(A = 0)$ and (2.6). □

As part of the proof of Theorem 3.1 we needed to know that $ES' < \infty$, which by (2.6) is equivalent to $ES < \infty$. 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) $Es^k(X_1) < \infty$
(ii) $EA^k < \infty$
(iii) $ES^k < \infty$.

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

$$E\mathcal{P}(x)(\mathcal{P}(x) - 1) \dots (\mathcal{P}(x) - k + 1) = x^k. \tag{3.7}$$

Let $b > 0$, and recall $Y_i = s(\mathbf{X}(i))$. From (2.4)

$$E(A_{n+1}^k \wedge b) = P(A_n = 0) E(\mathcal{P}^k(\lambda Y_1) \wedge b) + \sum_{j=1}^{\infty} P(A_n = j) E(\mathcal{P}^k(\lambda Y_j) \wedge b), \tag{3.8}$$

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

$$E(A^k \wedge b) \quad = \quad P(A=0)E(\mathcal{P}^k(\lambda Y_1) \wedge b) + \sum_{j=1}^{\infty} P(A=j)E(\mathcal{P}^k(\lambda Y_j) \wedge b). \qquad (3.9)$$

Letting $b \to \infty$ shows

$$E(A^k \wedge b) \geq P(A=0)E\mathcal{P}^k(\lambda Y_1) \geq \lambda^k EY_1^k.$$

So (ii) implies $EY_1^k < \infty$, which is assertion (i). Conversely, suppose $EY_1^k < \infty$. From (3.7), for every $0 < \delta < 1$ we can find d such that

$$E(\mathcal{P}^k(x) \wedge b) \leq (E\mathcal{P}^k(x)) \wedge b \leq (1+\delta)(x^k \wedge b) + d. \qquad (3.10)$$

This implies

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

Substituting into (3.9),

$$E(A^k \wedge b) \leq P(A=0)((1+\delta)\lambda^k EY_1^k + d) + \sum_{j=1}^{\infty}((1+\delta)\lambda^k(EY_j^k \wedge b) + d)P(A=j).$$

Using either (2.7) for $k=1$ or (2.8) for $k \geq 2$, we have for all ϵ and suitable chosen d'

$$\lambda^k EY_j^k \leq \lambda^k(m^k + \epsilon)j^k + d'.$$

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

$$E(A^k \wedge b) \leq gE(A^k \wedge b) + h$$

for $n \geq 1$ and therefore

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

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

$$ES^k = \sum_{j=1}^{\infty} EY_j^k P(A=j) \leq \sum_{j=1}^{\infty}((m^k+\epsilon)j^k + d')j^k P(A=j),$$

and the corresponding lower bound with $-\epsilon$. $\square$

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

Proposition 3.1 *If $\Xi(t)$ converges in distribution then $\rho < 1$ and $Es(X_1) < \infty$.*

PROOF. It is not difficult to see that convergence of $\Xi(t)$ to some limit Ξ implies $P(\Xi = \emptyset) > 0$. Indeed, for each state ξ, there is a non-random time $t_0 \geq 0$ (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 $P(\Xi = \emptyset) > 0$. Hence the inter-regeneration time $\tau - T_1$ has finite mean. Since $\tau - T_1$ is stochastically larger than $s(X_1)$, we deduce $Es(X_1) < \infty$. Moreover in the notation of (3.2)

$$(EN)(ES') < \infty$$

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

$$ES = Es(\mathbf{X}(A')) \leq ES' < \infty.$$

In the subadditive limit (2.7) $\beta = \inf_n \frac{Es(\mathbf{X}(n))}{n}$ and so

$$Es(\mathbf{X}(n)) \geq n\beta. \tag{3.11}$$

So $\beta EA' \leq Es(\mathbf{X}(A')) < \infty$, implying

$$\limsup_n EA_n \leq EA \leq EA' < \infty. \tag{3.12}$$

But as at (3.8),

$$\begin{aligned} EA_{n+1} &= P(A_n = 0)\lambda Es(X_1) + \sum_{j=1}^{\infty} \lambda Es(\mathbf{X}(A_n))1(A_n = j) \\ &\geq P(A_n = 0)\lambda\beta + \lambda\beta \sum_{j=1}^{\infty} EA_n 1(A_n = j) \text{ by (3.11)} \\ &= \rho P(A_n = 0) + \rho EA_n. \end{aligned}$$

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

$$\begin{aligned} \liminf_{n\to\infty} EA_n &\geq \rho \sum_{n=1}^{\infty} P(A_n = 0) \\ &= \infty \text{ because } P(A_n = 0) \to P(A = 0) > 0. \end{aligned}$$

But this contradicts (3.12), so we must have instead $\rho < 1$. □

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 *$\{A_n\}$ is transient if there exist a $\theta_0 \in (0,1]$ and an $\epsilon > 0$ such that*

$$\limsup_{n\to\infty} E\left(e^{-\theta_0 Y_n + \theta_0(1+\epsilon)n}\right) \leq 1. \tag{3.13}$$

The proof is given in Appendix A. It is easy to see that (3.13) implies $\rho > 1$. For some $s(\cdot)$, in particular if $s(\cdot)$ is additive, (3.13) is equivalent to $\rho > 1$, provided $Es(X_1) < \infty$.

4. Characteristics of the stationary distribution

Assume now that $\rho < 1$ and $Es(X_1) < \infty$, 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

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

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

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

Corollary 4.1

$$\bar{c} = \frac{\frac{\lambda}{2}(ES^2)(Ec(X_1)) \ + \ E\left[c(\mathbf{X}(A'))s(\mathbf{X}(A'))\right]}{ES'}.$$

In particular, if $Es^2(X_1) < \infty$ and $Ec^2(X_1) < \infty$ then $\bar{c} < \infty$.

Proof. The formula can be established by integrating over the distribution (3.4) 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 $1/ES'$, a regenerative argument rederives the formula.

If $Es^2(X_1) < \infty$ then Lemma 3.3 implies that $S \stackrel{d}{=} s(\mathbf{X}(A'))$ has finite second moment; similarly if $Ec^2(X_1) < \infty$ then $c(\mathbf{X}(A'))$ 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 one can write expressions in the spirit of (3.4) 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

Corollary 4.2

$$EL = \lambda EW = \frac{\frac{\lambda}{2}ES^2 \ + \ E\left[A's(\mathbf{X}(A'))\right]}{ES'}.$$

In particular, Lemma 3.3 implies EL (or EW) is finite if and only if $Es^2(X_1) < \infty$.

As in classical queueing theory, one expects that k'th moments of L and W are finite if and only if $Es^{k+1}(X_1) < \infty$, and this can be verified in our model (see Appendix B for their verifications).

5. Discussion

1. 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 X's 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

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

2. We take X_n as a type of a customer. It is also natural to consider it as an original service time of the customer. In this case the type-space to be $(0, \infty)$ (note 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 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.

3. 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 $c(\cdot), s(\cdot), \lambda$ 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 *competitive analysis of algorithms* [6] one 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

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

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

Acknowledgments

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

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, *Annals of Applied Probability* 10, 779-827.

[3] Bambos, N. and Walrand, J. (1991). On the stability and performance of parallel processing systems, *J. ACM* 38, 429-452. Springer, Berlin.

[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 & Brooks, Pacific Grove.

[6] Fiat, A. and Woeginger, G. J., *Eds.* (1998). *Online Algorithms.* Springer, Berlin. Lecture Notes in Computer Science, 1442.

[7] Lindvall, T. (1992). *Lectures on the Coupling Method.* 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.*

[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.

Appendix A

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

Lemma A.1 *Let*

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

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 following limit exists*

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

which is increasing and concave.

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

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

Hence

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

Taking the logarithms of the 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. □

Proof of Proposition 3.2. Without loss of generality we assume $\lambda = 1$. We use Theorem 8.4.2 from Meyn and Tweedie [8] (or Theorem 3.7 from Chapter 5 from Brémaud

[4] 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 suitable chosen. We have

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

and therefore we want

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

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 (3.13), there exists i_0 such that

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

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

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

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]),

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

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

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

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

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

In particular, for $k = 0$, $\lambda = \nu EA'$. Since

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

we have $\nu = 1/ES'$ as is 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, $EL^k < \infty$ if and only if $EA^{k+1} < \infty$, which is equivalent to $Es^{k+1}(X_1) < \infty$ by Lemma 3.3.

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

$$W = W_Q + S,$$

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 the finite k-th moment if only if $ES^{k+1} < \infty$. Hence, using the inequality,

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

we have that $EW^k < \infty$ if and only if $ES^{k+1} < \infty$.