Bandwidth Packing

E. G. Coffman, Jr. and A. L. Stolyar*
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
{egc,stolyar}@research.bell-labs.com

Abstract

We model a server that allocates varying amounts of bandwidth to "customers" during service. Customers could be computer jobs with demands for storage bandwidth or they could be calls with demands for transmission bandwidth on a network link. Service times are constants, each normalized to 1 time unit, and the system operates in discrete time, with packing (scheduling) decisions made only at integer times. Demands for bandwidths are for fractions of the total available and are limited to the discrete set $\{1/k, 2/k, \ldots, 1\}$ where k is a given parameter. More than one customer can be served at a time, but the total bandwidth allocated to the customers in service must be at most the total available. Customers arrive in k flows and join a queue. The jth flow has rate λ_j and contains just those customers with bandwidth demands j/k.

We study the performance of the two packing algorithms First Fit and Best Fit, both allocating bandwidth by a greedy rule, the first scanning the queue in arrival order and the second scanning the queue in decreasing order of bandwidth demand. We determine necessary and sufficient conditions for stability of the system under the two packing rules. The average total bandwidth demand of the arrivals in a time slot must be less than 1 for stability under any packing rule, i.e., the condition

$$\rho := \sum_i \lambda_i (i/k) < 1$$

must hold. We prove that if the arrival rates $\lambda_1, \ldots, \lambda_{k-1}$ are symmetric, i.e., $\lambda_i = \lambda_{k-i}$ for all $i, 1 \leq i \leq k-1$, then $\rho < 1$ is also sufficient for stability under both rules. Our Best Fit result strengthens an earlier result confined to Poisson flows and equal rates $\lambda_1 = \cdots = \lambda_{k-1}$, and does so using a far simper proof. Our First Fit result is completely new. The work here extends earlier results on bandwidth packing in multimedia communication systems, on storage allocation in computer systems, and on message transmission along slotted communication channels.

It is not surprising that $\rho < 1$ is sufficient under Best Fit, since in a congested system, Best Fit tends to serve two complementary (matched) customers in each time slot, with bandwidth demands being i/k and $(k-i)/k$ for some i, $1 \leq i \leq k-1$. It is not so obvious, however, that $\rho < 1$ is also sufficient under First Fit. Interestingly, when the system becomes congested, First Fit exhibits a "self-organizing" property whereby an ordering of the queue by time of arrival becomes approximately the same as an ordering by decreasing bandwidth demand.

---

* Part of the work of this author was completed while he was with AT&T Labs–Research, Murray Hill, NJ 07974 (now located in Florham Park, NJ 07932).

1 Introduction

We study a queueing model of storage and transmission bandwidth allocation in computer and communication systems. To define the model, we use the terminology of queueing systems; later, we will map this terminology into that of the applications. In our queueing model, customers are allocated available bandwidth according to their demands, each customer holding its allocation while it is being served, then releasing its allocation when it departs from the system. More than one customer can be served at a time, but the total bandwidth allocated to customers in service at any time must be less than the total available. Bandwidth demands are discretized and specified in fractions; for some given integer $k > 0$, a demand can be any multiple of $1/k$ up to $k/k = 1$. The discretization loses no generality in practice and expands the potential applications, as we shall see below.

Customers arrive in k flows to a single queue, the ith flow having rate λ_i and just those customers with bandwidth demands i/k. Each customer service time is a constant, which we take to be the unit of time. In addition, the system operates in discrete time; packing decisions are made, and customer services begin, only at integer times. Unit time intervals beginning at integer times will also be called *time slots*. Note that our model is a stochastic version of one dimensional bin packing [5], where a bin corresponds to the total bandwidth available over a time slot. We study the performance of both the First Fit and Best Fit packing rules. At service completions, both rules scan the queue and pack customer bandwidth demands by a greedy rule: the demand being considered is packed if and only if it is for at most the bandwidth still remaining from the demands already packed for customers to be served in the next time slot. The difference between the two policies is that First-Fit scans the queue in arrival order, while Best Fit scans the queue in decreasing order of bandwidth demand.

Our analysis addresses stability problems: determine necessary and sufficient conditions on the arrival rates λ_i such that the system is stable under First Fit and Best Fit, i.e., the underlying Markov queueing processes are ergodic. We prove that if the bandwidth-demand rates $\lambda_1, \ldots, \lambda_{k-1}$ are symmetric, i.e., $\lambda_i = \lambda_{k-i}$, $1 \le i \le k-1$, then under a very general class of arrival processes,

$$\rho := \sum_{i=1}^{k} \lambda_i (i/k) < 1$$

is necessary and sufficient for stability for both First Fit and Best Fit. Since ρ is the average total bandwidth demand in each time unit, $\rho < 1$ is clearly a necessary condition, so our proofs focus entirely on showing that the condition is sufficient. These proofs are trivially extended to any system with arrival rates majorized by the symmetric set $\{\lambda_i\}$ with $\rho < 1$, i.e., if the input rates of another system are λ_i' with $\lambda_i' \le \lambda_i = \lambda_{k-i}$, $1 \le i \le k-1$, and $\lambda_k' \le \lambda_k$, then stability obtains there as well.

In what follows we will take $\lambda_k = 0$. It would be trivial for us to generalize our results to the case $\lambda_k > 0$, but we have chosen not to do so as it creates a lack of symmetry and clutters the analysis.

Models similar to ours were studied in [3], where applications to multimedia communications were emphasized. The term 'bandwidth packing' was introduced in [3] as a name for the class of problems of interest here. In the applications, bandwidth on a network link is being allocated to several competing demands in varying amounts such as those needed for video, audio, and data transmission. It was proved in [3] that $\rho < 1$ was sufficient for the stability of Best Fit when the input flows were specialized to the Poisson process and when the demand distribution was uniform with all λ_i's equal. The stability question for First Fit was left as an open problem. The analysis in [3] applied the classical potential (Lyapunov) function approach. Our approach uses the relatively recent fluid-limit techniques described in Section 2. As a consequence, our proofs are more compact and more easily adapted to general arrival processes.

In another important application, the available bandwidth refers to the storage (memory) bandwidth of a multiprocessor system, a model studied in [9]. Customers are jobs using varying amounts of storage while running on a computer. The model in [9] differs from ours in requiring a strict FIFO service discipline. This is a substantial simplification of our model, but the analysis in [9] leads to further results, including formulas for invariant measures. Our work solves the stability problem in this application for much more efficient packing rules.

Our work also contributes new results to the analysis of an equivalent model of slotted communication systems [6]. In this new interpretation, customers are messages and bandwidth demands are message durations (fractions of a time slot); the available bandwidth in a time unit of our model becomes a unit-duration slot in which subsets of messages are packed and transmitted. With arriving messages modeled by a discretized Markov process, the analysis in [6] focuses on the Next Fit algorithm: When a message arrives and finds no messages waiting, it is packed (will be sent) in the next time slot. If a message arrives and finds other waiting messages, it is packed in the latest time slot already allocated at least one message, if it fits in the remaining unallocated time of that slot; otherwise, the message is packed in the next, as yet unused, time slot (and hence eventually transmitted one time unit later). Our analysis extends the earlier work to the much more efficient First Fit and Best Fit packing algorithms. For example, with equal arrival rates $\lambda_1 = \cdots = \lambda_{k-1}$, the message-rate capacity under Next Fit is only 3/2, whereas it is 2 under First Fit and Best Fit.

As a final application, one that takes us away from the bandwidth interpretation, we mention classical k-server queues. Our model generalizes these queues by allowing customers to require more than one server during their service. In the terminology of our model, a bandwidth demand of i/k is simply a request for i servers.

The next section formalizes our probability model and introduces the fluid-limit approach to our stability problems. Our main results appear in Sections 3 and 4 as theorems giving necessary and sufficient conditions for stability under the First Fit and Best Fit rules, respectively. In Section 5 we present a moment convergence result, which complements the stability results for both First Fit and Best Fit.

While our main results are in the stochastic analysis of algorithms, our methods also yield useful results in the asymptotic average-case analysis of algorithms. In the average-case (or fixed-input) model, a fixed number n of customers with i.i.d. bandwidth demands is given, and the objective is the large-n behavior of the expected total bandwidth wasted while serving the n demands. Section 6 applies the fluid-limit approach to the average-case model by giving a simple proof that, under First Fit and symmetric bandwidth-demand distributions, the expected total wasted bandwidth is $o(n)$, so the expected number of time slots needed to serve the n customers exceeds the expected sum of bandwidth demands ($n/2$) by a $o(n)$ term.

Section 7 concludes the paper with a discussion of open problems and the sensitivity of the analysis to various model assumptions.

2 Preliminaries

Under First Fit, a state of the queue is denoted by an element of the set of all finitely terminated sequences on $\{1, \ldots, k-1\}$. The length of the sequence is the queue length, and the ith element of the sequence gives the bandwidth demand of the customer that was ith to arrive among the customers currently waiting. Under Best Fit, the arrival order is not needed; the state just needs, for each $i = 1, \ldots, k-1$ the number of customers waiting with bandwidth demand i/k. Hereafter, a *type-i* demand is one for a fraction i/k of the bandwidth; type-i customers are those with type-i demands.

We assume that the aggregate arrival process of the $k-1$ customer types can be described by a finite number of independent, discrete-time regenerative processes with finite-mean regeneration cycles. Our proofs rely on two consequences of this assumption: The underlying queueing process, which we denote by $X = (X(t),\ t = 1, 2, \ldots)$, is a countable Markov chain, and the functional strong law of large numbers holds for the input process. To avoid trivial complications, we also assume that X is irreducible and aperiodic. These assumptions allow for virtually any process having a regenerative structure, e.g., discrete-time versions of Markov modulated Poisson processes, the processes generated by on-off sources, etc. However, to avoid complicated notation, in the rest of the paper we view the $k-1$ input flows as independent with the ith being an i.i.d. sequence of integer-valued random variables which give the numbers of type-i arrivals in $[t-1, t]$ and have a finite mean λ_i, the same for all $t = 1, 2, \ldots$. With this simplification, the underlying process X becomes the queue-content process.

In what follows, the norm $\|X(t)\|$ denotes the number of customers waiting at time t. Let $X^{(n)}$ denote a process X with an initial condition such that $\|X^{(n)}(0)\| = n$. In the analysis to follow, all variables associated with a process $X^{(n)}$ will be supplied with the upper index (n).

The following theorem is a corollary of a more general result of Malyshev and Menshikov [10].

Theorem 1 *Suppose there exists an integer $T > 0$ such that for any sequence of processes $X^{(n)}$, we have*

$$\lim_{n\to\infty} E[\frac{1}{n}\|X^{(n)}(nT)\|] = 1 - \epsilon \tag{1}$$

for some $\epsilon > 0$. Then X is ergodic.

It was shown by Rybko and Stolyar [11] that an ergodicity condition of the form (1) naturally leads to a fluid-limit approach to the stability problem of queueing systems. This approach was further developed by Dai [7], Chen [2], Stolyar [12], and Dai and Meyn [8]. As the form of (1) suggests, the approach studies a fluid process $x(t)$ obtained as a limit of the sequence of scaled processes $\frac{1}{n}X^{(n)}(nt), t \geq 0$; at the heart of the approach in its standard form is a proof that $x(t)$ starting from any initial state with norm $\|x(0)\| = 1$ reaches 0 in finite time T and stays there. (In most cases of interest, including ours, weaker conditions are sufficient, e.g., it is enough to verify that $\inf_{t\geq 0}\|x(t)\| < 1$, as shown in [12].) In our setting we need to define what the scaling $\frac{1}{n}X^{(n)}(nt)$ means. In order for this scaling to make sense, we will need an alternative definition of the queueing process.

To this end, we first adopt the convention $X(t) = X(\lfloor t \rfloor)$, $t \geq 0$, which allows us to view X as a continuous-time process defined for all $t \geq 0$, but with new arrivals and services still beginning only at integer times $t = 0, 1, 2, \ldots$. Next, we define the following random functions associated with the process $X^{(n)}(t)$: $F_i^{(n)}(t)$ is the total number of type-i customers that arrived by time $t \geq 0$, including the customers present at time 0; and $\hat{F}_i^{(n)}(t)$ is the number of type-i customers that were served by time $t \geq 0$. Obviously, $\hat{F}_i^{(n)}(0) = 0$ for all i. As in [11] and [12], we "encode" the initial state of the system; in particular, we extend the definition of $F_i^{(n)}(t)$ to the negative interval $t \in [-n, 0)$ by assuming that the customers present in the system in its initial state $X^{(n)}(0)$ arrived in the past at time instants $-(n-1), -(n-2), \ldots, 0$, exactly one customer at each time instant. In the case of First Fit, we require the order of these arrivals to be the same as in the state at time 0. By this convention $F_i^{(n)}(-n) = 0$ for all i and n, and $\sum_{i=1}^{k-1} F_i^{(n)}(0) = n$.

It is clear that the process $X^{(n)} = (X^{(n)}(t), t \geq 0)$ is a projection of the process $S^{(n)} = (F^{(n)}, \hat{F}^{(n)})$, where

$$F^{(n)} = (F_i^{(n)}(t), \quad t \geq -n, \quad i = 1, 2, \ldots, k-1)$$

and

$$\hat{F}^{(n)} = (\hat{F}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, k-1),$$

i.e., a sample path of $S^{(n)}$ uniquely defines a sample path of $X^{(n)}$.

Now consider the scaled process $s^{(n)} = (f^{(n)}, \hat{f}^{(n)})$, where

$$f^{(n)} = (f_i^{(n)}(t) = \frac{1}{n}F_i^{(n)}(nt), \quad t \geq -1, \quad i = 1, 2, \ldots, k-1)$$

and

$$\hat{f}^{(n)} = (\hat{f}_i^{(n)}(t) = \frac{1}{n}\hat{F}_i^{(n)}(nt), \quad t \geq 0, \quad i = 1, 2, \ldots, k-1)$$

The following lemma establishes convergence to a fluid process and is a variant of Theorem 4.1 in [7].

Lemma 1 *The following statements hold with probability* 1*. For any sequence of processes* $X^{(n)}$*, there exists a subsequence* $X^{(m)}$*,* $\{m\} \subseteq \{n\}$*, such that for each* i*,* $1 \le i \le k-1$*,*

$$(f_i^{(m)}(t), t \ge -1) \to (f_i(t), t \ge -1) \quad u.o.c. \tag{2}$$

$$(\hat{f}_i^{(m)}(t), t \ge 0) \to (\hat{f}_i(t), t \ge 0) \quad u.o.c. \tag{3}$$

where the functions f_i*,* $\hat{f}_i$*, are non-negative non-decreasing Lipschitz-continuous in the given time intervals, and u.o.c. means that convergence is uniform on compact sets as* $n \to \infty$*. The limiting set of functions*

$$s = (f, \hat{f}) = \{(f_i(t), t \ge -1), (\hat{f}_i(t), t \ge 0), i = 1, 2, \ldots, k-1\}$$

also satisfies

$$\sum_{i=1}^{k-1} f_i(0) = 1 \tag{4}$$

and for all i*,* $1 \le i \le k-1$*,*

$$f_i(-1) = 0, \quad \hat{f}_i(0) = 0, \tag{5}$$

$$f_i(t) - f_i(0) = \lambda_i t, \quad t \ge 0, \tag{6}$$

$$\hat{f}_i(t) \le f_i(t), \quad t \ge 0\ . \tag{7}$$

Finally, for any $0 \le t_1 \le t_2$*,*

$$\sum_{i=1}^{k-1} \frac{i}{k}(\hat{f}_i(t_2) - \hat{f}_i(t_1)) \le t_2 - t_1\ . \tag{8}$$

Proof. It follows from the strong law of large numbers that, with probability 1 for every i,

$$(f_i^{(n)}(t) - f_i^{(n)}(0),\ t \ge 0) \to (\lambda_i t, \quad t \ge 0) \quad u.o.c.$$

Also, for every n and i, the functions $(f_i^{(n)}(t), -1 \le t \le 0)$ and $(\hat{f}_i^{(n)}(t), t \ge 0)$ can increase by at most $k(1/n)$ in any interval of length $1/n$. Together with the above limit, this implies (2) and (3). We get (6) as a byproduct. Equations (4) and (5) follow from the construction representing the initial state. Equation (7) follows immediately from definitions, and the conservation law (8) from the trivial observation that the total bandwidth of customers served in one time slot is limited to 1. ∎

3 First Fit

To prove that $\rho < 1$ is sufficient for stability under First Fit, we need two lemmas.

Lemma 2 *For any fixed $T_1 > 1$, the following holds with probability 1. A limiting set of functions $s = (f, \hat{f})$ defined in Lemma 1 has the following additional property: For all i, $1 \leq i \leq k-1$,*

$$\hat{f}_i(T_1) > f_i(0). \tag{9}$$

Proof. Consider the sequence of sample paths of the scaled process $s^{(n)}$ converging to the set of functions s defined in Lemma 1. For any fixed $\epsilon > 0$ and $\delta > 0$, we have for all sufficiently large n,

$$\sum_{i=1}^{k-1} f_i^{(n)}(\epsilon) < 1 + (\sum_{i=1}^{k-1} \lambda_i)(1+\delta)\epsilon.$$

Since as long as the queue is non-empty at least one customer is served in every time slot, we conclude that

$$\hat{f}_i^{(n)}(\epsilon + 1 + (\sum_{i=1}^{k-1} \lambda_i)(1+\delta)\epsilon) \geq f_i^{(n)}(\epsilon),$$

which implies (9) via a simple passage to the limit $n \to \infty$, and by the fact that ϵ and δ can be arbitrarily small. ∎

For a set s of functions as defined in Lemma 1, let us define

$$\tau_i(t) = \inf\{\xi \geq -1 |\ f_i(\xi) > \hat{f}_i(t)\} \tag{10}$$

and let

$$\tau(t) = \min_{1 \leq i \leq k-1} \tau_i(t). \tag{11}$$

The proof of the First Fit stability result centers on the analysis of the times $\tau_i(t)$ because of their useful properties, one being a simple relation to the fluid limit of the queue-length processes $\|X^{(n)}(t)\|$. For this reason, we give below an informal interpretation of these times defined in terms of the unscaled processes $S^{(n)} = \{(F_i^{(n)}, \hat{F}_i^{(n)})\}$. According to (10), $\tau_i(t)$ is the earliest time by which the number of type-i arrivals exceeds the number of type-i departures by time t. Under suitable conditions to be covered in the lemma below, $\tau_i(t)$ can be expressed as the inverse of f_i evaluated at $\hat{f}_i(t)$, i.e.,

$$\tau_i(t) = f_i^{-1}(\hat{f}_i(t)), \tag{12}$$

as illustrated in Figure 1. We remark that $\tau_i(t)$ need not be a smooth function. For example, the initial state can be contrived so that $f_i(t)$ is flat in some subinterval of $[-1, 0]$, thus creating a discontinuity in $\tau_i(t)$. On the other hand, as proved later, $\tau_i(t)$ has to be Lipschitz-continuous in the interval $1 < t < \infty$. Under First Fit, the queue of type-i customers at

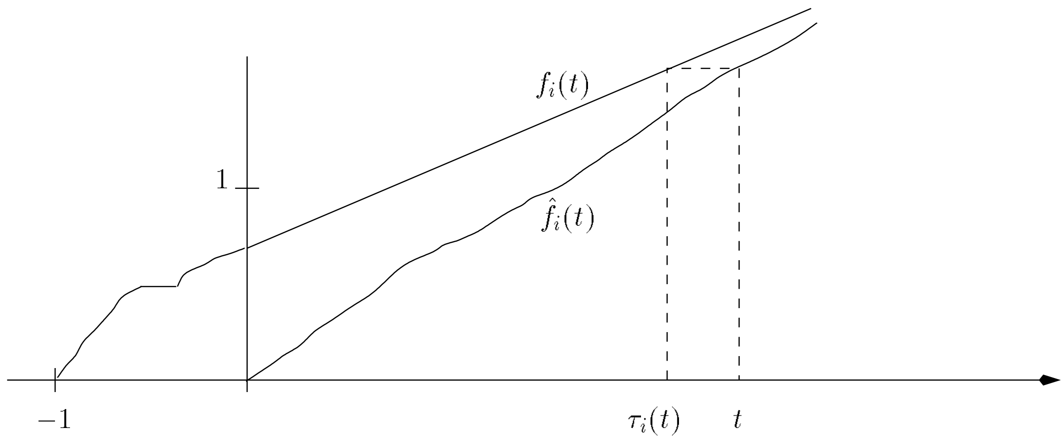

Figure 1: The functions $\tau_i(t)$, $\hat{f}_i(t)$, and $f_i(t)$ with $f_i(t) = \lambda_i t + f_i(0)$ for $t \geq 0$.

time t consists of just those type-i customers that arrived during $[\tau_i(t), t]$. Then $\tau_i(t) = t$ is a type-i empty-queue condition. Recalling our discussion of Theorem 1, we want to show that $\tau(t)$ tends towards t, i.e., $\tau'(t) > 1$, and absorbs in the empty-queue condition $\tau(t) = t$. These and related properties of the $\tau_i(t)$ are formalized in the following result.

Lemma 3 *Let $T_1 > 1$ be fixed. There exist fixed constants T_2 and T with $T_1 \leq T_2 \leq T < \infty$ such that with probability 1, a limiting set of functions $s = (f, \hat{f})$ defined in Lemma 1 has the following additional properties:*

(i) We have

$$\tau_i(T_1) > 0 \quad \text{for all } i,\ 1 \leq i \leq k-1. \tag{13}$$

(ii) In the interval $t \geq T_1$, every function $\tau_i(t)$, $1 \leq i \leq k-1$, is non-decreasing Lipschitz-continuous, and therefore so is $\tau(t)$.

(iii) At any regular point $t \geq T_1$, i.e., a point where all the derivatives of each of the functions $f_i, \hat{f}_i, \tau_i$, and τ exist for all i, $1 \leq i \leq k-1$, we have

$$\tau_i'(t) = \hat{f}_i'(t)/\lambda_i \tag{14}$$

$$\tau(t) < t \ \Rightarrow\ \tau'(t) \geq 1/\sum_{i=1}^{k-1} \lambda_i \tag{15}$$

$$(\tau_i(t) < \tau_j(t)\ \wedge\ i < j)\ \Rightarrow\ \tau_j'(t) = 0. \tag{16}$$

(iv) For all $t \geq T_2$,

$$i < j\ \Rightarrow\ \tau_i(t) \geq \tau_j(t). \tag{17}$$

(v) If the input flows are symmetric, i.e., if $\lambda_i = \lambda_{k-i}$ for every i, then we have at any regular point $t \geq T_2$,

$$\tau(t) < t \;\Rightarrow\; \tau'(t) \geq 1/\rho > 1 \tag{18}$$

(vi) For symmetric input flows, for all $t \geq T$,

$$\tau(t) = t, \tag{19}$$

which is equivalent to the assertion that, for all $t \geq T$,

$$\hat{f}_i(t) = f_i(t) \quad \textit{for all } i, 1 \leq i \leq k-1. \tag{20}$$

Proof Property (i) follows from Lemmas 1 and 2. When $t \geq T_1 > 1$, the effects of the initial state have dissipated and we know that τ_i is positive (by property (i)), that $f_i(t) = f_i(0) + \lambda_i t$, and that $\hat{f}_i(t)$ is nondecreasing Lipschitz-continuous (by Lemma 1). It follows easily that $\tau_i(t)$ is nondecreasing Lipschitz-continuous for $t \geq T_1$, which proves property (ii).

For property (iii), we first differentiate (12) at regular points t and substitute $f_i'(t) = \lambda_i$; this gives (14). To prove (15), define

$$M(t) := \{i \mid \tau_i(t) = \tau(t)\},$$

so that, since t is a regular point, we can write for all $i \in M(t)$,

$$\tau_i(t) = \tau(t), \quad \tau_i'(t) = \tau'(t), \quad \hat{f}_i'(t)/\lambda_i = \tau'(t). \tag{21}$$

For the unscaled sample path, it is easy to see that for any sufficiently small $\epsilon > 0$ and all sufficiently large n, at least one customer of a type $i \in M(t)$ will be served in each time slot of the time interval $[nt, n(t+\epsilon)]$. This means that $\sum_{i \in M(t)} \hat{f}_i'(t) \geq 1$, which together with (21) implies $\tau'(t) \geq 1/\sum_{i=1}^{k-1} \lambda_i$, thus proving (15).

To prove (16) and complete the proof of property (iii), consider an unscaled sample path at time nt. If $\tau_i(t) < \tau_j(t)$, then for small $\delta > 0$ and all sufficiently large n, there are at least $\lambda_i[\tau_j(t) - \tau_i(t)](1-\delta)n$ type-i customers in the queue ahead of any type-j customer. This means that, if $i < j$, there exists a small $\epsilon > 0$ such that no type-j customers will be served in the interval $[nt, n(t+\epsilon)]$. This in turn implies that $\hat{f}_j^{(n)}(t+\epsilon) = \hat{f}_j^{(n)}(t)$ for all n large enough, and therefore that (16) holds.

Property (iv) follows from (15) and (16) in property (iii). The constant T_2 can be chosen to be

$$T_2 = T_1 + \frac{\max_i \tau_i(T_1) - \tau(T_1)}{1/\sum_{i=1}^{k-1} \lambda_i}.$$

To prove property (v), we note first that, by property (iv), the set $M(t)$ for $t \geq T_2$ has the form $M(t) = \{k-1, k-2, \ldots, r\}$ with $r \leq k-1$. Then we can rewrite (21) as

$$\tau(t) = \tau_{k-1}(t) = \ldots = \tau_r(t) < \tau_{r-1}(t) \tag{22}$$

$$\tau'(t) = \tau'_{k-1}(t) = \ldots = \tau'_r(t) > 0 \tag{23}$$

$$\hat{f}'_{k-1}(t)/\lambda_{k-1} = \cdots = \hat{f}'_r(t)/\lambda_r = \tau'(t), \tag{24}$$

for some $r \leq k-1$. Here, we need to show that, if $\tau(t) < t$, then

$$\tau'(t) \geq 1/\rho \tag{25}$$

Let us assume that k is odd; the proof of (25) for k even is very similar and left to the reader. First, we make an observation similar to the one we made in the proof of (16). Consider the unscaled sample path at time nt. Equation (22) implies that, for any small $\delta > 0$ and all sufficiently large n, there are at least $\lambda_i(\tau_{r-1}(t) - \tau(t))(1-\delta)n$ customers of each type $i = k-1, k-2, \ldots, r$ in the queue ahead of any customer of type $j = r-1, \ldots, 1$. This means that there exists a small $\epsilon > 0$ such that in the interval $[nt, n(t+\epsilon)]$ the customers of types $r-1, \ldots, 1$ have lower priority than customers of types $k-1, k-2, \ldots, r$. More precisely, no customer of type $j < r$ will be packed in a time slot as long as a customer of type $i \geq r$ can be packed into that time slot instead. Therefore, as far as the behavior of the functions $\hat{f}_i^{(n)}, i \geq r$, in the interval $[t, t+\epsilon]$ is concerned, we can ignore the customers of types $j < r$.

In the remainder of the proof of property (v), we fix ϵ with the above observation in mind; we confine ourselves to the behavior of scaled processes in the interval $[t, t+\epsilon]$, and the behavior of the corresponding unscaled processes in the interval $[nt, n(t+\epsilon)]$, with n sufficiently large.

Let $p = (k-1)/2$, and note that the symmetry condition $\lambda_i = \lambda_{k-i},\ 1 \leq i \leq p$, implies

$$\begin{aligned} \rho := \sum_{i=1}^{k-1} \lambda_i \frac{i}{k} &= \sum_{i=1}^{p} \left[\lambda_i \frac{i}{k} + \lambda_{k-i} \frac{k-i}{k} \right] \\ &= \sum_{i=1}^{p} \lambda_i = \sum_{i=p+1}^{k-1} \lambda_i. \end{aligned} \tag{26}$$

If $r \geq p+1$, then

$$\tau'(t) = \frac{1}{\sum_{i=r}^{k-1} \lambda_i} \geq \frac{1}{\sum_{i=p+1}^{k-1} \lambda_i} = \frac{1}{\rho}. \tag{27}$$

To see this, note that exactly one customer of some type $i \geq r$ will be served in each time slot. For, since $2r \geq k+1$, two such customers have demands exceeding the total available bandwidth. This immediately implies that

$$\sum_{i=r}^{k-1} \hat{f}'_i(t) = 1$$

which means that $\tau'(t) \sum_{i=r}^{k-1} \lambda_i = 1$, and hence that (27) holds.

To finish the proof of property (v), it remains to dispose of the case $r \leq p$. We will show that $\tau'(t) = 1/\rho$. First, we observe that $\tau'(t) \leq 1/\rho$. This is because, by an argument similar to the one used for the case $r \geq p+1$ above, we have

$$\sum_{i=p+1}^{k-1} \hat{f}_i'(t) \leq 1$$

and so

$$\tau'(t) \leq \frac{1}{\sum_{i=p+1}^{k-1} \lambda_i} = \frac{1}{\rho}.$$

Now assume that strict inequality holds,

$$\tau'(t) < 1/\rho. \tag{28}$$

We will prove that this implies that $\hat{f}_r'(t) \geq 1/\rho$, which is a contradiction, since $\hat{f}_r'(t) = \tau'(t)$ by definition of r.

If (28) were to hold, then for any $\delta > 0$, any sufficiently small η, $0 < \eta < \epsilon$, (with η depending on δ), and all sufficiently large n (depending on δ and η), the following three observations would hold for an unscaled sample path in the interval $[nt, n(t+\eta)]$:

(a) The number of time slots not serving any customers of types $k-1, \ldots, k-r+1$, which do not fit together with type-r customers, is at least

$$\left[1 - (\tau'(t) + \delta) \sum_{i=k-r+1}^{k-1} \lambda_i\right] \eta n.$$

(b) Consider a time slot described in (a). If this slot does not serve any customers of types $p, p-1, \ldots, r+1$, then it must serve at least one type-r customer.

(c) The total number of served customers of types $i = p, \ldots, r+1$ does not exceed

$$[\tau'(t) + \delta] \left(\sum_{i=r+1}^{p} \lambda_i\right) \eta n.$$

Since the number of slots occupied by customers of types $p, p-1, \ldots, r+1$ is at most the number of such customers, observations (a)-(c) imply that the limiting type-r service rate has the lower bound

$$\hat{f}_r'(t) \geq \left[1 - (\tau'(t) + \delta) \sum_{i=k-r+1}^{k-1} \lambda_i\right] - [\tau'(t) + \delta] \sum_{i=r+1}^{p} \lambda_i$$

Since $\delta > 0$ can be arbitrarily small, we get

$$\hat{f}_r'(t) \geq 1 - \tau'(t) \left[\sum_{i=k-r+1}^{k-1} \lambda_i + \sum_{i=r+1}^{p} \lambda_i\right].$$

By the symmetry condition, $\sum_{i=k-r+1}^{k-1} \lambda_i = \sum_{i=1}^{r-1} \lambda_i$, so

$$\begin{aligned} \hat{f}_r'(t) &\geq 1 - \tau'(t)\sum_{i=1}^{p} \lambda_i + \tau'(t)\lambda_r \\ &= 1 - \tau'(t)\rho + \tau'(t)\lambda_r \\ &> \tau'(t)\lambda_r, \end{aligned}$$

the desired contradiction. Thus, (28) can not hold, we can conclude that $\tau'(t) = 1/\rho$, and property (v) is proved.

It follows from property (v) that

$$\inf\{t|\ \tau(t) = t\} \leq T_2 + \frac{T_2 - \tau(T_2)}{1/\rho - 1} \leq T_2 + \frac{T_2}{1/\rho - 1}$$

Let us choose the constant T to be

$$T = T_2 + \frac{T_2}{1/\rho - 1}.$$

Since we know that $\frac{d}{dt}(t - \tau(t)) \leq 1 - 1/\rho < 0$ at any regular point $t \geq T_2$ such that $t - \tau(t) > 0$, we conclude that $\tau(t) = t$ for all $t \geq T$. This proves property (vi) and hence the lemma. ∎

Theorem 2 *Suppose the input flow intensities are symmetric, and $\rho < 1$. Then under First Fit X is ergodic.*

Proof The proof is a slight modification of the proof of Theorem 4.2 in [7]. In particular, Lemmas 1 and 3 imply that there exists a $T > 0$, which can be chosen to be an integer, such that for any sequence of processes $\{X^{(n)}\}$ we have

$$\lim_{n\to\infty} \frac{1}{n}\|X^{(n)}(nT)\| = \lim_{n\to\infty} \sum_{i=1}^{k-1}(f_i^{(n)}(T) - \hat{f}_i^{(n)}(T)) = 0, \tag{29}$$

with probability 1. The uniform integrability of the sequence $\{X^{(n)}\}$ can be proved in ways similar to those in [11] and [7]. Uniform integrability and the convergence in (29) imply that

$$\lim_{n\to\infty} E[\frac{1}{n}\|X^{(n)}(nT)\|] = 0.$$

Then the condition in (1) of Theorem 1 holds, and we are done. ∎

We can also make strong statements about convergence properties and the existence of moments. These apply to Best Fit as well, so we defer these results to Section 5.

4 Best Fit Discipline

In this section, we prove that the analog of Theorem 2 for Best Fit also holds. We use the same general fluid-limit approach, but the arguments will be simpler. We again need Lemma 1, but we will create a new version of Lemma 3, a simpler version in that there will be no need to deal with the times $\tau_i(t)$ or the encoding of the initial state; instead of analyzing $t-\tau_i(t)$, we will analyze the difference $q_i(t) := f_i(t) - \hat{f}_i(t)$, proving that it reaches zero in finite time and stays there.

Theorem 3 *Suppose the input flow intensities are symmetric, and $\rho < 1$. Then under Best Fit X is ergodic.*

Proof: We need only prove Lemma 4 below; with Lemma 4 replacing Lemmas 2 and 3, the proof of Theorem 2 will apply to Best Fit.

Lemma 4 *Suppose that the input flows are symmetric. Then there exist constants $0 = T_k < T_{k-1} < \ldots < T_1 = T < \infty$ such that, with probability 1, a limiting set of functions s defined in Lemma 1 has the following additional property for every $i = 1, 2, \ldots, k-1$: At any regular point $t \geq T_{i+1}$,*

$$\hat{f}_i(t) < f_i(t) \;\Rightarrow\; \hat{f}'_i(t) \geq \lambda_i + (1-\rho), \tag{30}$$

and for any $t \geq T_i$,

$$\hat{f}_i(t) = f_i(t) \quad \textit{and therefore} \quad \hat{f}'_i(t) = \lambda_i. \tag{31}$$

Thus, for all $t \geq T$, and for all i, $1 \leq i \leq k-1$,

$$\hat{f}_i(t) = f_i(t) \tag{32}$$

Proof All the conventions introduced in the proof of Lemma 3 are still in force. Thus, $s^{(n)}, n = 1, 2, \ldots$ is the sequence of sample paths of the scaled process $s^{(n)}$ which converges to s. And when we refer to the unscaled sample path, we mean the corresponding sample path of the process $S^{(n)}$. We consider only the case when k is odd; the proof for even k is analogous.) Define $p := (k-1)/2$, as before, and recall that $q_i(t) := f_i(t) - \hat{f}_i(t)$, so that (30) and (31) become

$$q_i(t) > 0 \Rightarrow q'_i(t) \leq -(1-\rho) \tag{33}$$

and

$$q_i(t) = 0 \text{ for every } t \geq T_i. \tag{34}$$

We need a couple of key observations, the first following from the fact that, the higher the bandwidth demand, the higher the packing priority under Best Fit.

(a) For any fixed r, the service of customers of types $k-1, k-2, \ldots, r$ is completely unaffected by the service of customers of types $j < r$.

(b) The condition $q_i(t) > 0$ for the limiting set of functions implies that, for a sufficiently small, fixed $\epsilon > 0$, and all n sufficiently large, the corresponding unscaled sample path is such that in the interval $[tn, (t+\epsilon)n]$:

(b_1) there are always type-i customers available for service;

(b_2) if $i \leq p$, then every time slot serving a type-$(k-i)$ customer must serve a type-i customer; every time slot not serving a type-i customer, or a type-$(k-i)$ or larger customer must serve one or more customers of types $p, p-1, \ldots, i+1$.

The proof is by induction on i decreasing from $k-1$ to 1. If $i = k-1$, it follows easily from observations (a) and (b_1) that at any regular point $t \geq 0 = T_k$, the condition $q_{k-1}(t) > 0$ implies that $\hat{f}'_{k-1}(t) = 1 \geq \lambda_{k-1} + (1-\rho)$, and hence that (33) holds.

Notice that $q_{k-1}(0) \leq 1$, so if we choose

$$T_{k-1} = T_k + 1/(1-\rho), \tag{35}$$

then (34) follows from (33). This establishes the basis of the induction.

For the induction step, suppose (33) and (34) hold for $i = k-1, k-2, \ldots, r+1$. We will now prove that (33) and (34) also hold for $i = r$.

Consider a regular point $t \geq T_{r+1}$. If $r \geq p+1$, then the condition $q_r(t) > 0$ must imply

$$\hat{f}'_r(t) = 1 - \sum_{i=r+1}^{k-1} \lambda_i > \lambda_r + (1-\rho), \tag{36}$$

which gives $q'_r \leq -(1-\rho)$. To see this note that, in an unscaled sample path, one and only one customer of types $k-1, k-2, \ldots, r$ can be served in a time slot. Thus, (36) follows from observations (a) and (b_1) and the inductive hypothesis, which asserts that the customers of each of the types $i = k-1, \ldots, r+1$ are served at exactly the corresponding rate λ_i for all $t > T_i$. (We omit routine ϵ, δ-technicalities similar to those used in the proof of Lemma 3.)

If $r \leq p$, then by applying observations (a) and (b_2), we get

$$\hat{f}'_r(t) \geq 1 - \sum_{i=k-r+1}^{k-1} \lambda_i - \sum_{i=r+1}^{p} \lambda_i$$

$$= 1 - \sum_{i=1}^{r-1} \lambda_i - \sum_{i=r+1}^{p} \lambda_i - \lambda_r + \lambda_r = \lambda_r + (1-\rho)$$

so $q'_r \leq -(1-\rho)$ again holds.

Now if we observe that $q_r(T_{r+1}) \leq 1 + \lambda_r T_{r+1}$, and set, in analogy with (35),

$$T_r = (1 + \lambda_r T_{r+1})/(1-\rho),$$

then we see that (34) follows from (33). The inductive step and hence the proof of Lemma 4 and Theorem 3 is complete. ∎

5 Moment Convergence

It is shown in [8] that condition (1) implies not only stability, but also very strong moment-existence and convergence properties. For example, Theorem 4 below follows directly from (1) and Theorem 6.2 in [8] (which can easily be adjusted for our discrete-time case).

Theorem 4 *Suppose the λ_i are symmetric, $\rho < 1$ holds, and the input processes are i.i.d. sequences with finite $(p+1)$-st moments ($p \geq 1$ is an integer). Let $X(\infty)$ have the stationary distribution of the Markov chain X under either First Fit or Best Fit. Then*

$$E\|X(\infty)\|^p < \infty$$

and for any initial state $X(0)$,

$$\lim_{t\to\infty} E\|X(t)\|^p = E\|X(\infty)\|^p$$

6 Connection to Average-Case Analysis

Our First Fit stability analysis is closely related to the average-case analysis of First Fit bin packing under discrete distributions [4]. We can recast the average-case model into our setting as follows. Suppose the initial state consists of a queue of n customers with customer types being a sequence of independent samples from a given distribution on $\{1,\ldots,k-1\}$; and assume there are no new arrivals. A formula for the expected total wasted bandwidth in the packing of n customers is the objective of the average-case analysis. The problem is difficult, so virtually all of the results to-date describe large-n asymptotic behavior. In particular, it has been shown that for certain customer-type distributions, the expected total wasted bandwidth is $o(n)$. This section demonstrates how the First Fit average-case result can easily be obtained from the properties of the fluid limit of a stochastic system like the one considered in previous sections. Further results of this type are discussed in the next section.

Consider an infinite sequence of i.i.d. customer types $\xi_1, \xi_2, \ldots$, taking values in the set $\{1,\ldots,k-1\}$, with the distribution $\{\lambda_i, i=1,\ldots,k-1\}$, $\sum_i \lambda_i = 1$. The reason for adopting our arrival rate notation for the customer-type distribution in an average-case model will be clear when the theorem below links up the average-case and stochastic analysis. Define a sequence of systems (just like those analyzed in previous sections) indexed by $n=1,2,\ldots$. The n^{th} system has an initial state consisting of customers of types $\xi_1,\ldots,\xi_n$ waiting in queue in the order listed. Suppose there are no new arrivals after time 0. Note that in this setting the initial state is random.

For the n^{th} system, define $U^{(n)}$ to be the time slot in which the last customer of the initial state is served under First Fit, and define

$$W^{(n)} = U^{(n)} - \sum_{i=1}^{n} \xi_i/k \ ;$$

$W^{(n)}$ is the total bandwidth (or server capacity) wasted by the First Fit packing process.

Theorem 5 *Under First Fit and a symmetric distribution* $\{\lambda_i\}$

$$\lambda_i = \lambda_{k-1}, \quad 1 \le i \le k-1,$$

the following holds with probability 1*:*

$$\lim_{n\to\infty} W^{(n)}/n = 0 \ , \tag{37}$$

$$\lim_{n\to\infty} U^{(n)}/n = 1/2 \ . \tag{38}$$

Remark. Since the random variables $W^{(n)}/n$ and $U^{(n)}/n$ are bounded above uniformly in n, the probability 1 convergence implies convergence of mean values.

Proof First, it is clear that the properties (37) and (38) are equivalent, since

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{k} = \frac{1}{2},$$

with probability 1. Also, $W^{(n)} \ge 0$ obviously holds, and therefore

$$\liminf_{n\to\infty} U^{(n)}/n \ge 1/2,$$

so it will suffice to show that

$$\limsup_{n\to\infty} U^{(n)}/n \le 1/2 \ . \tag{39}$$

For every index n, consider a modified system in which new arrivals after time 0 *do* occur; the input (say Poisson) flow of type i arrivals has intensity λ_i. By the definition of First Fit, such a modification can not change the random variables $W^{(n)}$ and $U^{(n)}$, because it has no effect on the service of initial customers. The primary reason for considering the modified system is to comply with the formulations of the results in previous sections, to ease the 'reuse' of those results.

We observe that our sequence of processes, with index n, satisfies the conditions of Lemmas 1 and 3, except for the fact that the initial state is now random. But since the initial states are drawn from a sequence of i.i.d. random variables, the functional strong law of large numbers applies to the sequence of initial states. We can conclude that: *Except for property (13), Lemma 1 and Lemma 3 are valid for our sequence of modified processes. Moreover, they are valid with* $T_1 = T_2 = 0$ *and a limiting set of functions (a fluid process)* s *such that*

$$f_i(t) = \lambda_i(t-(-1)), \quad -1 \le t \le 0, \quad \forall i \tag{40}$$

Indeed, it follows that

$$f_i(t) = \lambda_i(t-(-1)), \quad t \ge -1, \quad \forall i \tag{41}$$

and hence that $\tau_i(0) = -1$ for any i. The only property of the constant T_1 required in earlier proofs was that each function $f_i(\cdot)$ be strictly linear with slope λ_i in the interval $[\tau_i(T_1), \infty)$; with $T_1 = 0$, we still have this property. The only property of the constant T_2 required in the earlier proofs was that $T_2 \geq T_1$ and $\tau_{k-1}(T_2) \leq \ldots \leq \tau_1(T_2)$. Again, with $T_2 = 0$, we still have this property.

Applying the results of Lemma 1 and Lemma 3, we get $\tau'(t) \geq 1/\rho = 2$ at any regular point $t \geq 0$. In fact, from the conservation law (8), we see that equality must hold: $\tau'(t) = 1/\rho = 2$, at any regular point $t \geq 0$. Then the following must also hold:

$$\tau(t) = \tau_{k-1}(t) = \ldots = \tau_1(t) = -1 + 2t, \quad t \geq 0, \tag{42}$$

because an inequality $\tau(t) < \tau_i(t)$ for any fixed t and i would contradict property (8).

We are now in position to prove (39). Let us fix a small $\epsilon > 0$. It follows from (42) that any limiting set of functions s is such that, for all i,

$$\hat{f}_i((1-\epsilon)/2) = \lambda_i(1-\epsilon) < \lambda_i = f_i(0).$$

This means that, with probability 1 for any i, the sequence of scaled processes $\hat{f}_i^{(n)}(\cdot)$ is such that for all n, except perhaps for values in some finite set,

$$\lambda_i(1-2\epsilon) \leq \hat{f}_i^{(n)}((1-\epsilon)/2) < f_i^{(n)}(0) < \lambda_i(1+\epsilon)\ .$$

This in turn means that, in the unscaled systems:

(a) in the first $\lfloor (n/2)(1-\epsilon) \rfloor$ time slots, no server bandwidth was wasted and only initial customers were served;

(b) $U^{(n)} \leq (n/2)(1-\epsilon) + nk \cdot 3\epsilon$.

Therefore, with probability 1, $\limsup_{n\to\infty} U^{(n)}/n \leq (1-\epsilon)/2 + 3k\epsilon$.

Since $\epsilon > 0$ can be chosen arbitrarily small, we get (39), which concludes the proof. ∎

7 Discussion

Our proofs of Theorems 2 and 3 verify that, for arrival processes within the broad framework given in Section 2, the stability of the system under consideration depends essentially on the input flow intensities and is insensitive to the precise probabilistic structure of the input flows.

Other special cases of interest to which the result of Theorem 2 is easily extended, are sets of divisible bandwidth demands. If h/k and j/k, $j > h$ are any two demands in a divisible subset of $\{1/k, 2/k, \ldots, 1\}$, then h divides j and j in turn divides k. The special case

$\{1/2^a, 1/2^{a-1}, \ldots, 1/2, 1\}$ for some positive integer a is of interest in computer applications. We leave to the interested reader an easy adaptation of the fluid approach to a proof that $\rho < 1$ is sufficient for stability under First Fit and Best Fit *independent* of the relative sizes of the arrival rates.

In heavy congestion, First Fit typically matches demands i/k with their complements $(k - i)/k$, thus wasting no bandwidth and allowing $\rho < 1$ to be sufficient as well as necessary for stability. For this matching to occur, the queue must reorder itself dynamically into decreasing order by size. In an attempt to find other examples where First Fit has a similar self-organizing property, consider any distribution that yields perfect packings respecting the arrival rates, i.e., examples for which there are integers $n, n_1, \ldots, n_{k-1}$ such that, for some ρ, $0 < \rho < 1$, we have $\lambda_i = \rho n_i/n$, $1 \leq i \leq k-1$, and we can pack n_1 type-1 customers, n_2 type-2 customers, …,and n_{k-1} type-$(k-1)$ customers into n time units with no available bandwidth left over. It is easy to define an algorithm that suitably restricts the demand-type configurations allowed in each time unit so as to guarantee that $\rho < 1$ is sufficient for stability. (What makes this algorithmic technique impractical, of course, is that it requires advance knowledge of the arrival rates.) Typically, however, it is not possible to make the same stability claim about First Fit. As a simple example, take $k = 7$ and let the only nonzero arrival rates be $\lambda_2 = 2\lambda_3$. A greedy algorithm that, whenever possible, packs 2 type-2 customers and 1 type-3 customer in a time slot will be stable as long as $\rho < 1$. However, it can be shown that, during periods of congestion under First Fit, a positive fraction of the time slots will be packed with 2 type-3 customers or 3 type-2 customers, wasting 1/7 of the bandwidth in each case. A formal proof that First Fit is unstable for values of ρ in an interval $[1-\delta, 1]$, $\delta > 0$ is sketched in the appendix.

Examples like those above suggest that the class of bandwidth-demand distributions for which $\rho < 1$ is sufficient for stability under First Fit or Best Fit is likely to be relatively small. More definitive statements of this kind present interesting directions for further research.

We established in Section 6 that the large-n packing process in the average-case model and the queueing process under heavy congestion in the stochastic model are described by essentially the same fluid process. This connection makes the following conjecture quite plausible. Consider an average-case model with a fixed customer-type distribution $\{b_i\}$, and a family of stochastic models with this same customer-type distribution; then for each model in the family $\lambda_i = b_i\lambda$, where the total arrival rate $\lambda = \sum_i \lambda_i$ indexes the family of stochastic models. We conjecture that, for the First Fit algorithm, the expected total wasted bandwidth is $o(n)$ in the average-case model if and only if $\rho < 1$ is sufficient for stability in the family of stochastic models. Expected total wasted bandwidth is known to be $O(\sqrt{nk})$ for First Fit packing [4] and customer types independently and uniformly distributed on $\{1, \ldots, k-1\}$. This result and our Theorem 5 on the more general symmetric distributions lends support for the conjecture.

Further support is provided by recent results of Albers and Mitzenmacher [1] who showed that, in the average-case model, the expected wasted bandwidth under First Fit is $O(1)$ when $b_1 = \cdots = b_{k-2} = 1/(k-2)$. The fluid approach shows easily that First Fit is stable in the corresponding stochastic model with intensities $\lambda_1 = \cdots = \lambda_{k-2}$ and $\lambda_{k-1} = 0$, and with $\rho < 1$. In fact, the proof of Theorem 2 is easily generalized to prove the same result for any set of intensities satisfying: (i) λ_1 is arbitrary, (ii) for some given integer m, $2 \le m < k/2$, we have the symmetry $\lambda_i = \lambda_{k-i}$ for all i, $m \le i \le k-m$, and (iii) all other intensities are 0.

Acknowledgments We gratefully acknowledge Anja Feldmann and Nabil Kahale for having helped to initiate the research culminating in this paper.

Appendix

Proposition 1 *Consider First Fit bandwidth packing with $k = 7$ and Poisson input flows of only types 2 and 3 customers, with intensities satisfying $\lambda_2 = 2\lambda_3 > 0$. There exists a $\delta > 0$ such that the system with $\rho > 1 - \delta$ is unstable.*

Proof sketch We consider first a simpler, 'saturated' version of the system in which new arrivals are generated if and only if there is room for more customers in the current time slot. In other words, when packing the current time slot, if the entire queue has been scanned and 2/7 of the time slot is still empty, the queue is immediately extended by new arrivals, with the numbers of new customers of types 2 and 3 being Poisson distributed with means λ_2 and λ_3. These arrivals are immediately available for further packing. The process is repeated as necessary until the slot is at most 1/7 empty. The queue state just after each integer time is a discrete-time countable Markov chain. Its ergodicity is easily verified. It is also easy to see that the average per-slot rates μ_2 and μ_3 at which customers of types 2 and 3 are served are such that $\mu_3 = \mu_2/2 < 1$. Also, the Markov chain can be viewed as a regenerative process with an empty queue being a regeneration state. All moments of the regeneration cycle are finite.

Let us now return to our original system with λ_2 and λ_3 such that $\mu_3 < \lambda_3 = \lambda_2/2 < 1$, which means $\mu_3 < \rho < 1$. This system is unstable. To prove this, we consider an initial state formed by arrivals without service (packing is suspended) for M time slots, with M large. When the packing starts, the packing process is indistinguishable from the packing process in the saturated system, until the time slot when the last initial customer is reached (i.e., scanned for the first time). It will take approximately αM time slots, with $\alpha = \lambda_3/\mu_3 > 1$, for the packing process to reach the last initial customer. By that time, the queue is longer than the initial queue; it is extended by new arrivals during approximately αM time slots. The packing process will then take approximately $\alpha^2 M$ slots to reach the end of that queue. This continues, with the maximum queue length growing without bound. Using routine large-deviation estimates, it is a simple matter to convert the above observations into a rigorous argument that, with positive probability, the queue length tends to infinity (see, for example, the instability example in [11]).

References

[1] S. Albers and M. Mitzenmacher. Average-Case Analysis of First Fit and Random Fit Bin Packing. *Proc. Ninth Ann. ACM-SIAM Symp. Disc. Alg.*, (1998), pp. 290-299.

[2] H. Chen. Fluid Approximations and Stability of Multiclass Queueing Networks: Work-conserving Disciplines. *Annals of Applied Probability*, Vol. 5, (1995), pp. 637-665.

[3] E. G. Coffman, Jr., A. Feldmann, N. Kahale, B. Poonen. Computing Call Admission Capacities in Linear Networks. Bell Labs, Lucent Technologies, Murray Hill, NJ 07974 (1997), submitted for publication.

[4] E. G. Coffman, Jr., D. S. Johnson, P. W. Shor, and R. R. Weber. Bin packing with discrete item sizes, Part II: Tight bounds on First Fit. *Random Structures and Algorithms*, 10:69–101, 1997.

[5] E. G. Coffman, Jr., M. R. Garey, and D. S. Johnson. Approximation Algorithms for Bin Packing: A Survey. in *Approximation Algorithms for NP-Complete Problems*, D. S. Hochbaum (ed.). PWS Publishing Co. (1995), pp. 46-93.

[6] E. G. Coffman, Jr., S. Halfin, A. Jean-Marie, and Ph. Robert. Stochastic Analysis of a Slotted, FIFO Communication Channel. *IEEE Trans. Inf. Th.*, **39**(1993), pp. 1555–1566.

[7] J. G. Dai. On the Positive Harris Recurrence for Open Multiclass Queueing Networks: A Unified Approach Via Fluid Limit Models. *Annals of Applied Probability*, Vol. 5, (1995), pp. 49-77.

[8] J. G. Dai and S. P. Meyn. Stability and Convergence of Moments for Open Multiclass Queueing Networks Via Fluid Limit Models. *IEEE Transactions on Automatic Control*, Vol. 40, (1995), pp. 1889-1904.

[9] C. Kipnis and Ph. Robert. A Dynamic Storage Process. *Stoch. Proc. and Their Applic.,* **34**(1990), pp. 155–169.

[10] V.A. Malyshev and M.V. Menshikov. Ergodicity, Continuity, and Analyticity of Countable Markov Chains. *Transactions of Moscow Mathematical Society*, Vol. 39, (1979), pp. 3-48.

[11] A.N. Rybko and A.L. Stolyar. Ergodicity of Stochastic Processes Describing the Operation of Open Queueing Networks *Problems of Information Transmission*, Vol. 28, (1992), pp. 199-220.

[12] A.L. Stolyar. On the Stability of Multiclass Queueing Networks: A Relaxed Sufficient Condition via Limiting Fluid Processes *Markov Processes and Related Fields*, 1(4), 1995, pp.491-512.