Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications xxx (xxxx) xxx

stochastic processes and their applications

www.elsevier.com/locate/spa

Heavy traffic limit for the workload plateau process in a tandem queue with identical service times

H. Christian Gromoll[a,*], Bryce Terwilliger[a], Bert Zwart[b]

[a] *Department of Mathematics, University of Virginia, Charlottesville, VA 22904, United States*
[b] *Centrum Wiskunde & Informatica, PO Box 94079, 1090 GB Amsterdam, Netherlands*

Received 27 September 2018; received in revised form 12 April 2019; accepted 13 May 2019
Available online xxxx

Abstract

We consider a two-node tandem queueing network in which the upstream queue has renewal arrivals with generally distributed service times, and each job reuses its upstream service requirement when moving to the downstream queue. Both servers employ the first-in-first-out policy. The reuse of service times creates strong dependence at the second queue, making its workload difficult to analyze. To investigate the evolution of workload in the second queue, we introduce and study a process M, called the plateau process, which encodes most of the information in the workload process. We focus on the case of infinite-variance service times and show that under appropriate scaling, workload in the first queue converges, and although the workload in the second queue does not converge, the plateau process does converge to a limit M^* that is a certain function of two independent Lévy processes. Using excursion theory, we derive some useful properties of M^* and compare a time changed version of it to a limit process derived in previous work.

MSC: 60K25; 90B22

Keywords: Tandem queue; Infinite variance; Process limit; Lévy process; Continuous mapping; Excursion theory

1. Introduction

The goal of this paper is to establish a stochastic process limit of a two-node tandem queueing network where the first queue is a $GI/GI/1$ queue (that is jobs have independent generally distributed service times and independent generally distributed interarrival times) but

* Corresponding author.
E-mail addresses: gromoll@virginia.edu (H.C. Gromoll), bat5ct@virginia.edu (B. Terwilliger), bert.zwart@cwi.nl (B. Zwart).

https://doi.org/10.1016/j.spa.2019.05.007

in contrast to most queueing models, customers reuse their specific service requirement when moving to the second queue. In other words, once a job's random service requirement has been generated at the first queue, it will also be that job's requirement at the second queue. Both servers process jobs in first-in-first-out order and have unlimited waiting space.

This structure induces a strong dependence between arrivals and services at the second queue, leading to unusual phenomena and making even simple performance measures such as the workload difficult to analyze.

To visualize the effect of identical service times, consider the workload in the second queue over a generic period during which both queues are busy. During a given interarrival time for the second queue, its workload will decrease by exactly the duration of this interarrival time (since we are assuming the second queue does not empty during this period). But this time equals the interdeparture time from the first queue, which equals the service time of the job about to transfer. Since this job reuses its service time at the second queue, this also equals the amount of work about to enter the second queue. So the workload in the second queue simply decreases by this job's service time and then increases by the same amount when the job transfers. The effect over a busy period of the second queue is a series of returns to the same level attained at the previous arrival time.

This continues until a job in service at the first queue is larger than any previous job in the first queue's busy period. The workload in the second queue will then empty and be zero for a while until the job transfers, at which time the workload will increase to a new level that is higher than the previous level, and resume a series of returns to this new level until the next record-setting job comes through.

Thus, during a busy period of the first queue, the workload in the second queue is characterized by oscillations below a series of increasing levels or plateaus. When the first queue experiences a period of idleness, this pattern in the second queue is interrupted and its workload can reset to a new starting height for the next series of plateaus.

The pattern of frequent returns to the same level can be seen in Fig. 1, where the workload in the second queue must hit zero before each level increase. When compared visually to the workload in the first queue, it is clear that the behavior is very different because the workload in the second queue has frequent consecutive local maxima of the same value, interspersed with occasional increases of that value.

Why study such a model? After all, most queueing models in the literature make the Jacksonian assumption that jobs generate new independent service times at each queue, and there are good reasons for this. For one, the independence assumption is crucial to the mathematical techniques most often employed, for example for deriving product-form descriptions of the steady state in Jackson networks, or for proving diffusion approximations in generalized Jackson networks. A second reason is that independence is a natural assumption in many applications. Consider an automobile assembly line for example, where it would make sense to assume that the time to attach the doors is independent of the time to apply rust protection.

On the other hand, if we consider a manual automobile washing operation, it seems natural that the main factor influencing service times is the soil level of the vehicle. A very dirty vehicle will tend to have a longer service time than others at the first washing station, but probably also at the wheel washing station and very likely as well at the interior cleaning station. That is, one would expect a vehicle's random service times at various stations to be correlated with its soil level and thus to each other.

Computer and telecommunications networks afford further examples in which jobs must pass through a series of processing queues (transmission, integrity check, decryption, format

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

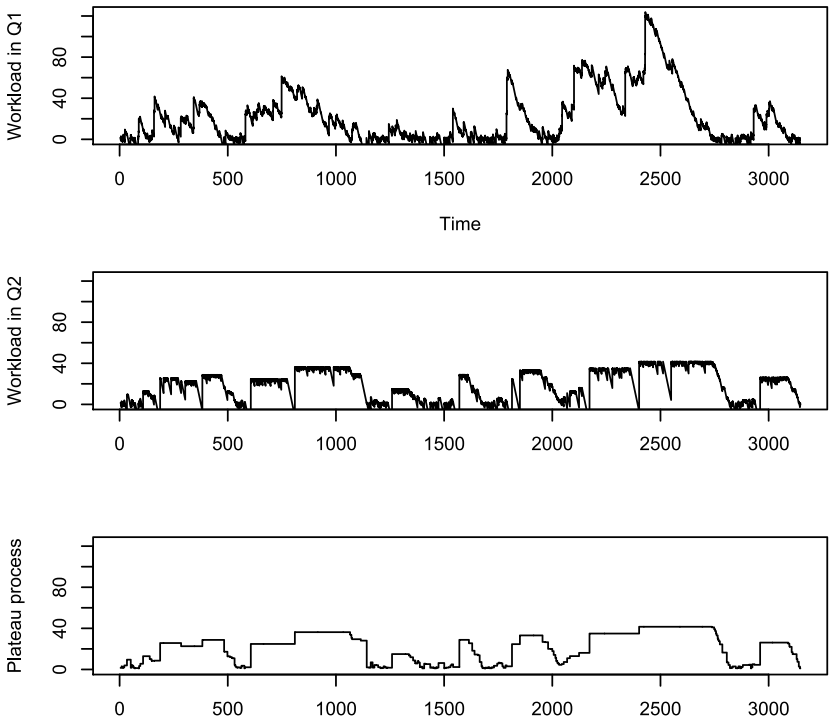

Fig. 1. The workload in both queues with identical service times in each. 1000 Poisson arrivals with parameter 1/3.1 service times are Pareto(1,3/2).

translation et al.), the random processing time of which will be correlated to the job's intrinsic size (file size). Indeed one can imagine many applications in which the successive service times of a given job are highly correlated due to some intrinsic property of the job, and this motivates consideration of models with correlated service times.

While we are not proposing that the model studied here, with just two nodes and identical service times is realistic for direct applications (such a model would allow for a more general network topology and arbitrarily correlated as opposed to *identical* service times), we view it as an archetype for more realistic models incorporating correlation. It is the simplest possible model in which the unusual effects of strong service time correlations are laid bare, and yet it already exhibits the serious difficulties in analyzing such effects. Our aim is to demonstrate some useful mathematical tools for dealing with such difficulties (adding to the small handful of results that exist for this model). We also speculate that some of the tools used here may be of use in analyzing non-queueing models incorporating similar correlation structures, such as models of world record evolution in improving populations as studied in [2].

The tandem model under consideration was first introduced for Poisson arrivals in the PhD thesis of O. Boxma [6] where a rather complete analysis of the invariant distribution was given, providing a rare example of a non-product form tandem queueing network for which an explicit analysis of the downstream queue is possible. For more on the broader theory of tandem queueing networks with independent service times, the reader is referred e.g. to [12,14,17] and references therein.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

The present model also shows unusual behavior in heavy traffic. In the finite variance case it is known [14,15] that the amount of work at the second node is of smaller order than the amount of work at the first node as the system load ρ (which is identical for both queues) tends to 1. For service times with bounded support, it is even shown in [7] that the expected value of the waiting time in the second queue is finite for $\rho = 1$. The intuition behind these results is that the amount of work in the first queue is driven by sums, and in the second queue is driven by maxima, suggesting that both queues should scale identically when the service times have infinite variance.

This behavior has recently been confirmed in our work [11], which is a prequel to the present study. In [11] we investigated the behavior of the workload of the second queue at embedded time points when the first queue empties. It was shown that this embedded Markov chain is sufficiently tractable, and analytic methods were used to investigate the process limit of this embedded Markov chain.

In the present paper we take up the task of analyzing the full workload process at the second node. We seek to prove a scaling limit theorem wherein the limit process is more tractable than the original workload process and can therefore serve as an approximation to it. One challenge is that the workload process does not converge in heavy traffic.

To see this intuitively, consider that on a space–time scale under which the successive plateaus of the process converge, there would be asymptotically infinitely many arrivals in between plateau increases. Under scaling, each such arrival causes a linear decrease at rate tending to infinity, followed by an upward jump the same size as the total decrease. The asymptotic result is oscillations below the level of the plateau that are too wild to converge in any of the Skorohod topologies.

We note that this type of behavior has been mentioned in Whitt's monograph [19], where new spaces (E and F) to potentially deal with such fluctuations have been suggested. Though an approach using this framework would be interesting, we take a different approach in the present paper which is more tailored to the specific model here. Notice that in Fig. 1, the silhouette of the workload in the second queue seems like it might converge in the usual J_1-topology under the same scaling as the workload in the first queue. Moreover, much of the information about the workload in the second queue is retained if we only keep track of these recurring levels or plateaus, so we do not lose much by working with just the silhouette. For example, if one is interested in the probability of the buffer at the second queue exceeding some critical threshold, the answer is the same for the silhouette. The silhouette also provides an upper bound for the actual workload at any time, provides information about how often the second queue is idle, etc.

In choosing to work with the silhouette, we eliminate the oscillating behavior that prevents us from working directly with the workload in the second queue, and gain the ability to prove a limit theorem.

This is the strategy we follow. We introduce and study a process M, called the plateau process, which encodes most of the information in the workload process. The plateau process is defined to be the workload in the second queue at the time of the most recent arrival. This definition eliminates the difficulty with scaling described above. We show that under an appropriate scaling, the plateau process converges to a limit M^* that is a certain function of two independent Lévy processes U^* and V^*.

More explicitly, it will be shown that the Nth job waits in the second queue for a period of time $F(U, V, 1)(N)$, where U and V are the arrival and service processes for the model, and

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

for two functions $x, y : [0, \infty) \to \mathbb{R}$,

$$F(x, y, c)(t) = \sup_{0\le s\le t}\left(y(s) - y(s-) + \sup_{0\le r\le s}\left(x(r) - y([r-c]^+)\right)\right) - \sup_{0\le s\le t}\left(x(s) - y([s-c]^+)\right).$$

Writing $R(t)$ for the number of jobs that have arrived to the second queue by time t, we will show that the plateau process can be written

$$M(t) = F(U, V, 1)(R(t)).$$

It is by no means obvious that the plateau process can be represented by this odd looking composition of functions, and it takes several steps in our proofs to establish it. Moreover, we show that the mapping F is continuous on a sufficient set in the Skorohod path space $\mathbb{D}$. Then for a sequence of models indexed by r, we have $M^r(t) = F(U^r, V^r, 1)(R^r(t))$, and letting $\check{M}^r(t) = \frac{1}{a_r}M^r(rt)$, we show that

$$\check{M}^r \Rightarrow M^*,$$

where $M^*(t) = F(U^* + \gamma\mu e, V^*, 0)(t/\mu)$; see Theorem 2.1 below.

The process appearing in the limit is not Markovian, but a suitable time-change is shown to be. Our second result provides (for a subset of cases) a means of performing some calculations on the limit process M^*, by deriving an explicit formula for the one-dimensional distributions of a natural time change $\{M^*(\mu L^{-1}(v)), v \ge 0\}$ of the process. Here μ is a constant and L^{-1} is the inverse local time of a reflected version of the limiting service process V^*, which is an explicit α-stable Lévy process. These one-dimensional distributions are given for each $v \ge 0$ by the distribution functions

$$F_v(y) = \exp\left(-\int_y^{y+v}\frac{\kappa(q)}{q}dq\right), \qquad y \ge 0,$$

where κ is an explicit function; see Theorem 6.1 in Section 6. This second result also implies that the embedded Markov chain of the limit process coincides with the limit of the embedded Markov chains considered in [11].

The paper is organized as follows. We first carefully define the model and scaling, make mild asymptotic assumptions, and state our first result, Theorem 2.1.

The bulk of the paper, Sections 3 and 4, is then devoted to the proof, which is essentially an elaborate application of the continuous mapping theorem. This is a bit delicate because the relevant mapping F is not continuous everywhere. In particular, we first show in a series of steps that $M^r(t)$ can indeed be represented as the composition of functions $F(U^r, V^r, 1)(R^r(t))$ described above. Then a series of steps shows that F is continuous on a particular subset of $\mathbb{D} \times \mathbb{D} \times \mathbb{R}$ (see Lemma 4.5). Proving that for the limiting primitive processes U^* and V^*, the triple $(U^* + \gamma\mu e, V^*, 0)$ is almost surely in this set enables a final application of the continuous mapping theorem together with the random time change theorem.

Finally, in Section 6 we develop some ideas from excursion theory to analyze the limit process M^* for a subset of cases (when the interarrival times have finite second moment). After performing a time change using a local time derived from V^*, the process becomes Markov. We are then able to apply some excursion theory results to calculate one dimensional distributions and relate this process to the limit derived in [11].

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

1.1. Notation

The following notation will be used throughout. Let $\mathbb{N} = \{1, 2, \ldots\}$ and let $\mathbb{R}$ denote the real numbers. Let $\mathbb{R}_+ = [0, \infty)$. For $a, b \in \mathbb{R}$, write $a \vee b$ for the maximum, and $a \wedge b$ for the minimum, $[a]^+ = 0 \vee a$, $[a]^- = 0 \vee -a$, $\lfloor a \rfloor$ for the integer part of a. For $f : \mathbb{R}_+ \to \mathbb{R}$ let $f^{\uparrow}(t) = \sup_{0 \le s \le t} f(s)$.

Let $\mathbb{D} = \mathbb{D}([0, \infty), \mathbb{R})$ be the space of real valued, right-continuous functions on $[0, \infty)$ with finite left limits. We endow $\mathbb{D}$ with the Skorohod J_1-topology which makes $\mathbb{D}$ a Polish space [5]. For $T \ge 0$, let $\rho_T(x, y) = \sup_{s \in [0,T]} |x(s) - y(s)|$. Let $e \in \mathbb{D}$ be the identity function $e(t) = t$. For $x \in \mathbb{D}$, let $x(t-) = \lim_{s \uparrow t} x(s)$, and let $x^-(t) = x(t-)$ for $t > 0$ and $x^-(0) = x(0)$.

Following Ethier and Kurtz [9] let Λ' be the collection of strictly increasing functions mapping $\mathbb{R}_+$ onto $\mathbb{R}_+$. Let $\Lambda \subset \Lambda'$ be the set of Lipschitz continuous functions such that $\lambda \in \Lambda$ implies $\sup_{s > t \ge 0} \left| \log \frac{\lambda(s) - \lambda(t)}{s - t} \right| < \infty$.

We will often use [9] Proposition 3.5.3: let $\{x_n\} \subset \mathbb{D}$ and $x \in \mathbb{D}$. Then $x_n \xrightarrow{J_1} x$ if and only if for each $T > 0$ there exists $\{\lambda_n\} \subset \Lambda'$ (possibly depending on T) such that $\lim_{n \to \infty} \sup_{0 \le t \le T} |\lambda_n(t) - t| = 0$ and $\lim_{n \to \infty} \sup_{0 \le t \le T} |x_n(t) - x(\lambda_n(t))| = 0$.

We write $X \sim Y$ if X and Y are equal in distribution. Weak convergence of random elements will be denoted by $\Rightarrow$. We adopt the convention that a sum of the form $\sum_{i=n}^{m}$ with $n > m$, or a sum over an empty set of indices equals zero.

2. Tandem queue model and main result

In this section we give a precise description of the tandem queue, specify our assumptions, and state our main result.

2.1. Definition of the model

We formulate a model equivalent to the one in Boxma [8], although we allow for general renewal arrivals. The tandem queueing system consists of two queues Q1 and Q2 in series; both Q1 and Q2 are single-server queues with an unlimited buffer. Jobs enter the tandem system at Q1. After completion of service at Q1 a job immediately enters Q2, and when service at Q2, which is the exact same length as previously experienced in Q1, is completed it leaves the tandem system. Jobs are served individually and at both servers with the first in first out discipline. We assume the system is empty at time zero.

More precisely, at Q1 the *exogenous arrival process* $E(\cdot)$ is a renewal process. Jump times of this process correspond to times at which jobs enter the system. This renewal process is defined from a sequence of interarrival times $\{u_i\}_{i=1}^{\infty}$, where u_1 denotes the time at which the first job to arrive after time zero enters the system and u_i, $i \ge 2$, denotes the time between the arrival of the $(i-1)$st and the ith jobs to enter the system after time zero. Thus, $U_i = \sum_{j=1}^{i} u_j$ is the time at which the ith arrival enters the system, which is interpreted as zero if $i = 0$, and $E(t) = \sup\{i \ge 0 : U_i \le t\}$ is the number of exogenous arrivals by time t. We assume that the sequence $\{u_i\}_{i=1}^{\infty}$ is an independent and identically distributed sequence of nonnegative random variables with $\mathbb{E}[u_1] = \mu < \infty$.

At Q1, the service process, $\{V_i, i = 1, 2, \ldots\}$, is such that V_i records the total amount of service required from the server by the first i arrivals. More precisely, $\{v_i\}_{i=1}^{\infty}$ denotes an independent and identically distributed sequence of strictly positive random variables. We interpret v_i as the amount of processing time that the ith arrival requires from both servers.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

The v_i's are known as the *service times*. Then, $V_i = \sum_{j=1}^{i} v_j$, which is taken to be zero if $i = 0$. It is assumed that $\mathbb{E}[v_1] = \nu < \infty$.

For $t \geq 0$, let

$$I(t) = \sup_{s \leq t} \left[V_{E(s)} - s \right]^-$$

be the idle time, that is the cumulative amount of time that the first server has been idle up to time t. For $n \geq 0$, let

$$I_n = I(U_n).$$

Then I_n is the cumulative amount of time that the first server has been idle up to the arrival of the nth job in the first queue.

Let $W_i(t)$ denote the (immediate) workload at time t at Qi, $i = 1, 2$, which is the total amount of time that the server must work in order to satisfy the remaining service requirement of each job present in the system at time t, ignoring future arrivals. For $t \geq 0$ we define

$$W_1(t) = V_{E(t)} - t + I(t).$$

Let D_n be the *transfer time* of the nth job. So, the nth job exits Q1 and enters Q2 at time D_n. Let $d_1 = u_1 + v_1$ and $d_n = D_n - D_{n-1}$ for $n \geq 2$ be the *intertransfer time* between arrivals of the $n-1$st and nth job to the second queue. For $n \geq 0$ we have

$$D_n = V_n + I_n.$$

Let $R(t)$ denote the number of transfers to Q2 by time t. For $t \geq 0$ we have

$$R(t) = \sup\{n \geq 0 : D_n \leq t\}. \tag{1}$$

Let $J(t)$ denote the cumulative amount of time that the second server has been idle up to time t, and $W_2(t)$ as the workload in Q2 at time t. That is, for $t \geq 0$ let

$$\begin{aligned} J(t) &= \sup_{s \leq t} \left[V_{R(s)} - s \right]^-, \\ W_2(t) &= V_{R(t)} - t + J(t). \end{aligned}$$

If k is the index of the first job in a busy period of the first queue then $W_1(U_k) = v_k$. Similarly, $W_2(D_k) = v_k$ if the kth job arrives to the second queue at a time when the second queue is empty.

Finally, let M_n denote the workload in the second queue at the time of the arrival of the nth job to the second queue, which is just the sojourn time of the nth job in the second queue. Let $M(t)$ be the piecewise constant right continuous function that agrees with the work load in the second queue at each transfer time and whose discontinuities are contained in the transfer times. We call $M(t)$ the *plateau process*. For integers $n \geq 0$ and real numbers $t \geq 0$ we have

$$\begin{aligned} M_n &= W_2(D_n), \\ M(t) &= M_{R(t)}. \end{aligned} \tag{2}$$

Finally, we define for $t \geq 0$,

$$U(t) = U_{\lfloor t \rfloor} \qquad \text{and} \qquad V(t) = V_{\lfloor t \rfloor}. \tag{3}$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

2.2. Sequence of models, assumptions, and results

We now specify a sequence of tandem queueing models indexed by $r \in \mathbb{R}$, where r increases to ∞ through a sequence in $(0, \infty)$. Each model in the sequence is defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The rth model in the sequence is defined as in the previous section where we add a superscript r to each symbol. In particular, for $t \geq 0$ let $M^r(t)$ denote the plateau process in the rth system.

Then $\{v_i^r\}_{i=1}^{\infty}$ and $\{u_i^r\}_{i=1}^{\infty}$ are the service times and interarrival times to the first queue with positive, finite means $\mathbb{E}\left[v_i^r\right] = v^r$ and $\mathbb{E}\left[u_i^r\right] = \mu^r$ for each $i = 1, 2, \ldots$ independent of each other. Define the following scaled versions of processes in the rth model for a sequence of positive reals $a_r \to \infty$ and $t \geq 0$,

$$\begin{aligned} &\bar{U}^r(t) = r^{-1}U(rt) && \text{and} && \bar{V}^r(t) = r^{-1}V(rt) \\ &\check{U}^r(t) = a_r^{-1}\left(U(rt) - r\mu^r t\right) && \text{and} && \check{V}(t) = a_r^{-1}\left(V(rt) - rv^r t\right) \\ &&& \check{M}^r(t) = a_r^{-1}M^r(rt). \end{aligned} \tag{4}$$

Asymptotic assumptions. We make the following asymptotic assumptions, as $r \to \infty$, about our sequence of models. Assume there is a sequence $\{a_r\}$ such that $r/a_r \to \infty$, $\check{U}^r(1) \Rightarrow \mathcal{U}^*$, $\check{V}^r(1) \Rightarrow \mathcal{V}^*$ in $\mathbb{R}$. In this case $\mathcal{U}^*$ and $\mathcal{V}^*$ are centered infinitely divisible random variables; see Feller [10] XII.7. This holds for example if the service time and interarrival time distributions are regularly varying with parameter in (1,2), that is, have finite means and infinite variance. Then we have $U^r \Rightarrow U^*$ and $V^r \Rightarrow V^*$ in $\mathbb{D}$, where U^* and V^* are Lévy stable motions with $U^*(1) \sim \mathcal{U}^*$ and $V^*(1) \sim \mathcal{V}^*$; see [19] supplement 2.4.1. We further assume $\lim_{r\to\infty} \mu^r = \lim_{r\to\infty} v^r = \mu$ and the traffic intensity parameter for the rth system $\rho^r = \frac{\mu^r}{v^r}$ satisfies

$$\frac{r}{a_r}\left(1 - \rho^r\right) \to \gamma \in \mathbb{R}.$$

Definition 1. Define the mapping $F : \mathbb{D} \times \mathbb{D} \times \mathbb{R} \to \mathbb{D}$ by

$$F(x, y, c)(t) = \sup_{0\leq s\leq t}\left(y(s) - y(s-) + \sup_{0\leq r\leq s}\left(x(r) - y([r-c]^+)\right)\right) - \sup_{0\leq s\leq t}\left(x(s) - y([s-c]^+)\right)$$

The following is the main result of the paper.

Theorem 2.1. *As $r \to \infty$,*

$$\check{M}^r \Rightarrow M^*,$$

where $M^*(t) = F(U^* + \gamma\mu e, V^*, 0)(t/\mu)$.

3. The plateau process as a function of U and V

In this section we derive various relationships between the stochastic processes comprising the tandem queueing model. These relationships hold for any of the r indexed models, so we suppress superscripts referring to a particular model in sequence.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

3.1. The idleness process for the first queue

This section is a prerequisite for understanding the arrival process in the second queue. If the cumulative idleness in the first queue is identically zero for all time, then the arrival process to the second queue is just a renewal process formed by the service times. Here we consider the cumulative idleness process in the first queue as a discrete time process. Consider the model defined in Section 2.1.

Lemma 3.1. *For each $n \geq 1$,*

$$I_n = u_1 + \max_{k=1}^{n} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right), \tag{5}$$

for $n = 1, 2, \ldots$

Proof. We proceed by induction. First observe that $\sum_{j=2}^{1}(u_j - v_{j-1}) = 0$, by convention, so

$$\max_{k=1}^{n} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right) \geq 0$$

for $n \geq 1$. $I_1 = u_1 + \max_{k=1}^{1} \sum_{j=1}^{k} (u_j - v_{j-1}) = u_1$. For $n = 2$,

$$I_2 = u_1 + [u_2 - v_1]^+ = u_1 + \max_{k=1}^{2} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right),$$

since there is no additional idleness if the second job arrives while the first job is in service. This is the base case for the induction.

For the inductive step, assume Eq. (5) holds for $n \geq 2$. There are two cases. In the first case the $(n+1)$st job arrives before the nth service is complete. In this case the first job in the current busy period had index $i \leq n$, arrived at time t_i, and the total amount of work that has arrived since t_i, $\sum_{k=i}^{n} v_k$ exceeds the amount of time $\sum_{k=i+1}^{n+1} u_k$ since t_i. That is,

$$\sum_{k=i+1}^{n+1} u_k - v_{k-1} < 0,$$

for some $i \leq n$. Thus

$$\max_{k=1}^{n+1} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right) = \max_{k=1}^{n} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right),$$

and the cumulative idle time has not increased

$$I_n = I_{n+1} = u_1 + \max_{k=1}^{n+1} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right).$$

In the second case, the $(n+1)$st job arrives after the nth service is complete, so the total idle time just before the arrival of the $n+1$ job is $u_1 + \sum_{k=2}^{n+1} u_k - v_{k-1}$. In this case, for any

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

job $i \le n$, the total amount of time $\sum_{k=i+1}^{n+1} u_k$ exceeds the total amount of work $\sum_{k=i}^{n} v_k$ since t_i. That is,

$$\sum_{k=i+1}^{n+1} u_k - v_{k-1} \ge 0.$$

Thus,

$$\left(\sum_{j=2}^{k}(u_j - v_{j-1})\right) \le \left(\sum_{j=2}^{n+1}(u_j - v_{j-1})\right)$$

for each $k = 2, \ldots, n+1$, and we have $\sum_{j=2}^{n+1} u_j - v_{j-1} = \max_{k=1}^{n+1}\left(\sum_{j=2}^{k}(u_j - v_{j-1})\right)$. ■

Note that the departure process of the first queue is equal to the arrival process $R(\cdot)$ of the second queue. Since the queueing discipline is FIFO, the number of jobs that have arrived to the second queue by time t is the greatest number N such that the total amount of time needed to complete the first N jobs, $\sum_{k=1}^{N} v_k$, is less than the amount of time spent working, t minus the cumulative idle time in the first queue.

3.2. Workload in the second queue

In this section we show how to write the plateau process $M(\cdot)$ as a function of the primitive arrival and service processes. The following formula relates sojourn times in the second queue to service times and idleness in the first queue. It comes from Lindley recursion [1] for a FIFO queue $W_2(D_{n+1}) = v_{n+1} + [W_2(D_n) - d_{n+1}]^+$, where no independence needs to be assumed about the intertransfer times d_k and service times v_k.

Lemma 3.2. *The sojourn time of the nth job in the second queue is*

$$M_n = \max_{k=1}^{n} \{v_k + I_k\} - I_n.$$

Proof. Note that the sojourn time of the nth job includes its service time. The second queue is initially empty and the service time of the nth job is the same in both queues. Clearly $I_1 = u_1$, since the first queue is empty until the arrival of the first job. So,

$$M_1 = v_1 = \max_{k=1}^{1} \{v_k + I_k\} - I_1.$$

The intertransfer time between the nth and $(n+1)$st job is $d_{n+1} = v_{n+1} + (I_{n+1} - I_n)$. Proceeding by induction, suppose $M_n = \max_{k=1}^{n} \{v_k + I_k\} - I_n$. Then, Lindley recursion gives

$$\begin{aligned}
M_{n+1} &= v_{n+1} + [M_n - v_{n+1} - (I_{n+1} - I_n)]^+ \\
&= v_{n+1} \vee (M_n - (I_{n+1} - I_n)) \\
&= v_{n+1} \vee \left(\max_{k=1}^{n} (v_k + I_k) - I_n - (I_{n+1} - I_n)\right) \\
&= \left[(v_{n+1} + I_{n+1}) \vee \max_{k=1}^{n} (v_k + I_k)\right] - I_{n+1} \\
&= \max_{k=1}^{n+1} (v_k + I_k) - I_{n+1}. \quad \blacksquare
\end{aligned}$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

Definition 2. Define the translation function $G : \mathbb{D} \times \mathbb{R} \to \mathbb{D}$ by

$$G(x, c)(t) = x([t - c]^+),$$

and define $H : \mathbb{D} \times \mathbb{D} \times \mathbb{R}_+ \to \mathbb{D}$ as the composition

$$H(x, y, c) = (x - G(y, c))^\uparrow .$$

More explicitly,

$$H(x, y, c)(t) = \sup_{0 \le s \le t} \left(x(s) - y([s - c]^+)\right).$$

We can write I_n in terms of V and U from (3).

Lemma 3.3. *For each $n \ge 1$,*

$$I_n = H(U, V, 1)(n),$$

Moreover H is constant on intervals of the form $[n, n + 1)$ where n is an integer, so for each integer n we have $H(U, V, n)(\lfloor t \rfloor) = H(U, V, n)(t)$ for all $t \ge 0$.

Proof. The processes V and U are constant between integers so H is constant on intervals of the form $[n, n + 1)$, where n is an integer. For an integer k, $v_k = V(k) - V(k-)$ and $u_k = U(k) - U(k-)$. By Lemma 3.1,

$$\begin{aligned}
I_n &= u_1 + \max_{k=1}^{n} \left(\sum_{j=2}^{k} (u_j - v_{j-1}) \right) \\
&= u_1 + \max_{k=1}^{n} \left(\sum_{j=2}^{k} u_j - \sum_{j=1}^{k-1} v_j \right) \\
&= \max_{k=1}^{n} \left(\sum_{j=1}^{k} u_j - \sum_{j=1}^{k-1} v_j \right) \\
&= \max_{k=1}^{n} (U(k) - V(k-1)) \\
&= \sup_{0 \le s \le n} \left(U(s) - V([s-1]^+)\right) \\
&= \sup_{0 \le s \le n} (U(s) - G(V, 1)(s)) \\
&= H(U, V, 1)(n). \quad \blacksquare
\end{aligned}$$

Now we can write R in terms of U and V.

Corollary 3.4.

$$R(t) = \max \{m \ge 0 : V(m) + H(U, V, 1)(m) \le t\}.$$

Proof. From Definition (1) we have $R(t) = \max\{N \ge 0 : \sum_{k=1}^{N} v_k + I_N \le t\}$. We have $\sum_{k=1}^{N} v_k = V(N)$ by Definition (3) and $I_N = H(U, V, 1)(N)$ by Lemma 3.3. $\blacksquare$

We can now write the plateau process in terms of the function F defined in Section 2.2. By Definitions 1 and 2,

$$F(x, y, c) = \left(y - y^- + H(x, y, c)\right)^\uparrow - H(x, y, c),$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

or more explicitly,

$$F(x, y, c)(t) = \sup_{0\le s\le t} (y(s) - y(s-) + H(x, y, c)(s)) - H(x, y, c)(t).$$

Lemma 3.5. *For all $t \ge 0$,*

$$M_{\lfloor t\rfloor} = F(U, V, 1)(t).$$

Proof. By Lemma 3.2

$$\begin{aligned} M_{\lfloor t\rfloor} &= \max_{k=1}^{\lfloor t\rfloor} (v_k + I_k) - I_{\lfloor t\rfloor} \\ &= \max_{k=1}^{\lfloor t\rfloor} (V(k) - V(k-) + I_k) - I_{\lfloor t\rfloor} \\ &= \max_{k=1}^{\lfloor t\rfloor} (V(k) - V(k-) + H(U, V, 1)(k)) - H(U, V, 1)(\lfloor t\rfloor) \end{aligned}$$

by Lemma 3.3. For a positive integer k we have $H(U, V, 1)(t)$ is constant for t in $[k, k+1)$ and $V(k) - V(k-) \ge V(t) - V(t-)$ for t in $[k, k+1)$. Thus, $V(t) - V(t-) + H(U, V, 1)(t)$ is maximized when t is an integer. Thus,

$$\begin{aligned} M_{\lfloor t\rfloor} &= \sup_{0\le s\le t} (V(s) - V(s-) + H(U, V, 1)(s)) - H(U, V, 1)(t) \\ &= F(U, V, 1)(t). \quad \blacksquare \end{aligned}$$

Finally we can express $M(\cdot)$ as function of U and V. By Definition (2), $M(t)$ is the composition $M_{(\cdot)}$ with the arrival process to the second queue. That is,

$$\begin{aligned} M(t) &= M_{R(t)} \\ &= F(U, V, 1)(\max\{m \ge 0 : V(m) + H(U, V, 1)(m) \le t\}). \end{aligned}$$

Notice that the plateau process is greater than or equal to the workload in the second queue at each time, that is $M(t) \ge W_2(t)$ for each $t \ge 0$.

4. Continuity properties of G, H, and F

Note that the function F is not continuous everywhere. For example, let $x_n = x = 1_{[1,\infty)} + 1_{[2,\infty)}$, let $y = 1_{[1,\infty)}$, and let $y_n = y(\cdot - 1/n)$ so that $(x_n, y_n, 0)$ clearly converges to $(x, y, 0)$ in $\mathbb{D} \times \mathbb{D} \times \mathbb{R}$. Then $F(x_n, y_n, 0) = y_n$ which converges in the Skorohod J_1-topology to y. But this does not equal $F(x, y, 0) = 1_{[1,2)}$, so F is not continuous at $(x, y, 0)$.

In this section we identify a subset of the domain of F that almost surely contains the limits of the processes we are interested in and on which F is indeed continuous. This result is obtained by treating F as a composition of continuous functions. The strategy of proof is similar to showing addition is continuous on a large subset of $\mathbb{D} \times \mathbb{D}$ (see e.g. [18]).

Lemma 4.1. *For any $x \in \mathbb{D}$, G is continuous at $(x, 0)$ in the product topology on $\mathbb{D} \times \mathbb{R}$.*

Proof. Let c_n be a sequence in $\mathbb{R}$ with $c_n \to 0$, and let $x_n \to x$ in $\mathbb{D}$. Then for each $T > 0$ there exists $\{\lambda_n\} \subset \Lambda$ such that $\sup_{0\le t\le T} |\lambda_n(t) - t| \to 0$ as $n \to \infty$ and $\sup_{0\le t\le T} |x_n(t) - x(\lambda_n(t))| \to 0$ as $n \to \infty$.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

For each $n = 1, 2, \ldots$ define

$$\tilde{\lambda}_n(t) = \begin{cases} \lambda_n(t - c_n), & \text{if } t \geq 2|c_n|, \\ \lambda_n\left(\left(1 - \frac{\text{sgn}(c_n)}{2}\right)t\right), & \text{if } t < 2|c_n|, \end{cases}$$

where $\text{sgn}(c_n) = -1$ if $c_n < 0$, $\text{sgn}(c_n) = 1$ if $c_n > 0$, and $\text{sgn}(c_n) = 0$ if $c_n = 0$.

We have $\{\tilde{\lambda}_n\} \subset \Lambda$ because each $\tilde{\lambda}_n$ is the composition of two functions in Λ. Now,

$$\begin{aligned}
&\sup_{0\leq t\leq T}\left|\tilde{\lambda}_n(t) - t\right| = \left(\sup_{0\leq t<2|c_n|}\left|\tilde{\lambda}_n(t) - t\right|\right) \vee \left(\sup_{2|c_n|\leq t\leq T}\left|\tilde{\lambda}_n(t) - t\right|\right) \\
&= \left(\sup_{0\leq t<2|c_n|}\left|\lambda_n\left(\left(1 - \frac{\text{sgn}(c_n)}{2}\right)t\right) - t\right|\right) \vee \left(\sup_{2|c_n|\leq t\leq T}|\lambda_n(t - c_n) - t|\right) \\
&\leq \left(\sup_{0\leq t<2|c_n|}\left|\lambda_n\left(\left(1 - \frac{\text{sgn}(c_n)}{2}\right)t\right) - \left(1 - \frac{\text{sgn}(c_n)}{2}\right)t\right|\right. \\
&\left. + \sup_{0\leq t\leq 2|c_n|}\left|\left(1 - \frac{\text{sgn}(c_n)}{2}\right)t - t\right|\right) \vee \left(\sup_{2|c_n|\leq t\leq T}|\lambda_n(t - c_n) - (t - c_n)| + |c_n|\right).
\end{aligned}$$

When $0 \leq t < 2|c_n|$ we have $0 \leq \left(1 - \frac{\text{sgn}(c_n)}{2}\right)t \leq 3|c_n|$, so

$$\begin{aligned}
\sup_{0\leq t\leq T}\left|\tilde{\lambda}_n(t) - t\right| &\leq \left(\sup_{0\leq t<3|c_n|}|\lambda_n(t) - t| + 3|c_n|\right) \\
&\quad \vee \left(\sup_{2|c_n|-c_n\leq t\leq T-c_n}|\lambda_n(t) - t| + |c_n|\right) \\
&\leq \sup_{0\leq t\leq T}|\lambda_n(t) - t| + 3|c_n|,
\end{aligned}$$

so $\sup_{0\leq t\leq T}\left|\tilde{\lambda}_n(t) - t\right| \to 0$ as $n \to \infty$.

Now, it suffices to show $\sup_{0\leq t\leq T}\left|G(x_n, c_n)(t) - G(x, 0)(\tilde{\lambda}_n(t))\right| \to 0$ by [9] Proposition 3.5.3. We have

$$\begin{aligned}
&\sup_{2|c_n|\leq t\leq T}\left|G(x_n, c_n)(t) - G(x, 0)(\tilde{\lambda}_n(t))\right| \\
&\quad = \sup_{2|c_n|\leq t\leq T}\left|x_n([t - c_n]^+) - x(\tilde{\lambda}_n(t))\right| \\
&\quad = \sup_{2|c_n|\leq t\leq T}|x_n(t - c_n) - x(\lambda_n(t - c_n))| \\
&\quad = \sup_{2|c_n|-c_n\leq t\leq T-c_n}|x_n(t) - x(\lambda_n(t))| \to 0
\end{aligned} \tag{6}$$

So it suffices to show $\sup_{0\leq t<2|c_n|}\left|G(x_n, c_n)(t) - G(x, 0)(\tilde{\lambda}_n(t))\right| \to 0$.

Fix $\epsilon > 0$ and let $\eta > 0$ such that $\sup_{0\leq t\leq\eta}|x(0) - x(t)| < \epsilon$ by right continuity of x at zero. Now, for n so large that $|c_n| < \min(T/3, \eta/6)$, $\sup_{0\leq t\leq T}|\lambda_n(t) - t| < \epsilon \wedge \eta/2$, and $\sup_{0\leq t\leq T}|x_n(t) - x(\lambda_n(t))| < \epsilon$ consider the $c_n < 0$, $c_n > 0$, and $c_n = 0$ cases.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

If $c_n < 0$,

$$\begin{aligned}
&sup_{0\le t<2|c_n|}\left|G(x_n,c_n)(t)-G(x,0)(\tilde{\lambda}_n(t))\right| \\
&\quad=\sup_{0\le t<2|c_n|}\left|x_n([t-c_n]^+)-x(\tilde{\lambda}_n(t))\right| \\
&\quad=\sup_{0\le t<-2c_n}|x_n(t-c_n)-x(\lambda_n(3t/2))| \\
&\quad\le\sup_{0\le t<-2c_n}|x_n(t-c_n)-x(\lambda_n(t-c_n))|+|x(\lambda_n(t-c_n))-x(\lambda_n(3t/2))| \\
&\quad\le\sup_{0\le t\le T}|x_n(t)-x(\lambda_n(t))|+\sup_{0\le t<-2c_n}|x(\lambda_n(t-c_n))-x(\lambda_n(3t/2))| \\
&\quad\le\sup_{0\le t\le T}|x_n(t)-x(\lambda_n(t))|+\sup_{0\le t<-2c_n}|x(\lambda_n(t-c_n))|+\sup_{0\le t<-2c_n}|x(\lambda_n(3t/2))|\,.
\end{aligned}$$

We have $(t-c_n)\vee(3t/2)\le -3c_n$ for $0\le t<-2c_n$, and so

$$\lambda_n(t-c_n)\vee\lambda_n(3t/2)\le\lambda_n(-3c_n)\le -3c_n+\eta/2\le\eta.$$

Thus,

$$\begin{aligned}
&\sup_{0\le t<2|c_n|}\left|G(x_n,c_n)(t)-G(x,0)(\tilde{\lambda}_n(t))\right| \\
&\quad\le\epsilon+\sup_{0\le t<-2c_n}|x(\lambda_n(t-c_n))|+\sup_{0\le t<-2c_n}|x(\lambda_n(3t/2))| \\
&\quad\le\epsilon+\sup_{0\le t\le\eta}|x(t)|+\sup_{0\le t\le\eta}|x(t)|\le 3\epsilon
\end{aligned}$$

If $c_n > 0$,

$$\begin{aligned}
&\sup_{0\le t<2|c_n|}\left|G(x_n,c_n)(t)-G(x,0)(\tilde{\lambda}_n(t))\right| \\
&\quad=\sup_{0\le t<2c_n}\left|x_n([t-c_n]^+)-x(\tilde{\lambda}_n(t))\right| \\
&\quad=\sup_{0\le t<2c_n}\left|x_n([t-c_n]^+)-x(\lambda_n(t/2))\right| \\
&\quad\le\sup_{0\le t<c_n}|x_n(0)-x(\lambda_n(t/2))|\vee\sup_{c_n\le t<2c_n}|x_n(t-c_n)-x(\lambda_n(t/2))|\,. \qquad (7)
\end{aligned}$$

For the first term,

$$\begin{aligned}
\sup_{0\le t\le c_n}|x_n(0)-x(\lambda_n(t/2))| &\le\sup_{0\le t<c_n}|x_n(0)-x(0)|+|x(0)-x(\lambda_n(t/2))| \\
&=|x_n(0)-x(\lambda_n(0))|+\sup_{0\le t<c_n}|x(0)-x(\lambda_n(t/2))| \\
&\le\sup_{0\le t\le T}|x_n(t)-x(\lambda_n(t))|+\sup_{0\le t\le\eta}|x(0)-x(t)|\le 2\epsilon,
\end{aligned}$$

since $\lambda_n(t/2)\le\lambda_n(c_n/2)\le c_n/2+\eta/2\le\eta$ for $0\le t\le c_n$. For the second term,

$$\begin{aligned}
\sup_{c_n\le t<2c_n}|x_n(t-c_n)-x(\lambda_n(t/2))| &=\sup_{0\le t<c_n}\left|x_n(t)-x\left(\lambda_n\left(\frac{t+c_n}{2}\right)\right)\right| \\
&\le\sup_{0\le t<c_n}|x_n(t)-x(\lambda_n(t))|+\left|x(\lambda_n(t))-x\left(\lambda_n\left(\frac{t+c_n}{2}\right)\right)\right| \\
&\le\epsilon+\sup_{0\le t<c_n}\left|x(\lambda_n(t))-x(0)+x(0)-x\left(\lambda_n\left(\frac{t+c_n}{2}\right)\right)\right|
\end{aligned}$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

$$\leq \epsilon + \sup_{0\leq t<c_n} |x(\lambda_n(t)) - x(0)| + \sup_{0\leq t<c_n} \left| x(0) - x\left(\lambda_n\left(\frac{t+c_n}{2}\right)\right)\right|$$
$$\leq \epsilon + 2 \sup_{0\leq t<\eta} |x(0) - x(t)| \leq 3\epsilon,$$

since $\lambda_n(t) \vee \lambda_n(\frac{t+c_n}{2}) \leq \lambda_n(c_n) \leq c_n + \eta/2 \leq \eta$ for $0 \leq t \leq c_n$.

If $c_n = 0$ then $\tilde{\lambda}_n = \lambda_n$ so $G(x_n, c_n)(t) - G(x, 0)(\tilde{\lambda}_n(t)) = x_n(t) - x(\lambda_n(t))$, which converges to zero uniformly by assumption.

So in all three cases we have

$$\sup_{0\leq t<2|c_n|} \left| G(x_n, c_n)(t) - G(x, 0)(\tilde{\lambda}_n(t))\right| \leq 3\epsilon.$$

Together with (6) and since ϵ was arbitrary, we have

$$\sup_{0\leq t\leq T} \left| G(x_n, c_n)(t) - G(x, 0)(\tilde{\lambda}_n(t))\right| \to 0$$

as $n \to \infty$.

So we have $G(x_n, c_n) \to G(x, 0)$ on $\mathbb{D}$. ■

For $x \in \mathbb{D}$, let $\mathrm{Disc}(x)$ denote the set of discontinuities of x.

Lemma 4.2. *H is continuous at $(x, y, 0)$ for all $x, y \in \mathbb{D}$ such that*

$$Disc(x) \cap Disc(y) = \varnothing.$$

Proof. Let $c_n \in \mathbb{R}$ with $c_n \to 0$ and let x_n and y_n be in $\mathbb{D}$ such that $x_n \to x$ and $y_n \to y$ and fix a time $T > 0$. Let $z_n = y_n - x_n$ and $z = y - x$. Since $\mathrm{Disc}(x) \cap \mathrm{Disc}(-y) = \varnothing$, [18] Theorem 4.1 tells us that there exists $\{\lambda_n\} \subset \Lambda'$ such that $\rho_T(\lambda_n, e) \to 0$ and $\rho_T(z_n, z \circ \lambda_n) \to 0$. Since G is continuous at $(z, 0)$ by Lemma 4.1, and $(z_n, c_n) \to (z, 0)$ we have $\{\tilde{\lambda}_n\} \subset \Lambda'$ such that $\rho_T(\tilde{\lambda}_n, e) \to 0$ and $\rho_T(G(z_n, c_n), z \circ \tilde{\lambda}_n) \to 0$. In fact, we may construct $\tilde{\lambda}_n$ as in the proof of 4.1. Since $x \mapsto x^{\uparrow}$ is continuous on $\mathbb{D}$ and $(x)^{\uparrow} \circ \tilde{\lambda} = (x \circ \tilde{\lambda})^{\uparrow}$, we have $\rho_T(H(x_n, y_n, c_n), H(x, y, 0) \circ \tilde{\lambda}_n) \to 0$. Since T was arbitrary we have H is continuous $(x, y, 0)$. ■

Lemma 4.3. *For all $x, y \in \mathbb{D}$,*

$$Disc(H(x, y, 0)) \subset \{t : y(t) - y(t-) > 0\} \cup \{t : x(t) - x(t-) < 0\}.$$

In particular, if $\{t : x(t) - x(t-) < 0\} = \varnothing$, then

$$Disc(H(x, y, 0)) \subset Disc(y).$$

Proof. $\mathrm{Disc}(H(x, y, 0)) = \{t : H(x, y, 0)(t) - H(x, y, 0)(t-) \neq 0\} = \{t : H(x, y, 0)(t) - H(x, y, 0)(t-) > 0\}$ since $H(x, y, 0)$ is nondecreasing. Thus,

$$\mathrm{Disc}(H(x, y, 0)) \subset \{t : (y - x)(t) - (y - x)(t-) > 0\}$$
$$\subset \{t : y(t) - y(t-) > 0\} \cup \{t : x(t) - x(t-) < 0\}.$$ ■

Lemma 4.4. *Let λ_n and γ_n be strictly increasing homeomorphisms from $[0, T]$ onto $[0, T]$ and $x_n, x \in \mathbb{D}$ such that for some finite collection $\{t_j\}_{j=0}^{N} \subset [0, T]$ with*

(i) $0 = t_0 < t_1 < \cdots < t_N = T$ we have $\lambda_n^{-1}(t_j) = \gamma_n^{-1}(t_j)$ for each $j = 0, 1, 2, \ldots, N$,

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

(ii) $\rho_T(x_n, x \circ \lambda_n) < \epsilon$, *and*

(iii) $w(x, [t_{j-1}, t_j)) = \sup\left(|x(t) - x(s)| : t, s \in [t_{j-1}, t_j)\right) < \epsilon$ *for each* $j = 1, 2, \ldots, N$,

then

$$\rho_T(x_n, x \circ \gamma_n) < 3\epsilon.$$

Proof. Let $r_j = \gamma_n^{-1}(t_j) = \lambda_n^{-1}(t_j)$ for $j = 0, 1, \ldots, N$, so that $\cup_{j=1}^N \left[r_{j-1}, r_j\right) = \cup_{j=1}^N \left[t_{j-1}, t_j\right) = [0, T)$. Then

$$\begin{aligned}
\rho_T(x_n, x \circ \gamma_n) &= \sup_{0 \le t \le T} |x_n(t) - x(\gamma_n(t))| \\
&= \max_{k=1}^N \sup_{r_{j-1} \le t < r_j} |x_n(t) - x(\gamma_n(t))| \vee |x_n(T) - x(T)| \\
&= \max_{k=1}^N \sup_{t_{j-1} \le t < t_j} \left|x_n(\gamma_n^{-1}(t)) - x(t)\right| \vee |x_n(T) - x(T)| \\
&= \max_{k=1}^N \sup_{t_{j-1} \le t < t_j} \left|x_n(\gamma_n^{-1}(t)) - x(t_{j-1}) + x(t_{j-1}) - x(t)\right| \\
&\qquad \vee |x_n(T) - x(T)|,
\end{aligned}$$

and so

$$\begin{aligned}
\rho_T(x_n, x \circ \gamma_n) &\le \max_{k=1}^N \left(\sup_{t_{j-1} \le t < t_j} \left|x_n(\gamma_n^{-1}(t)) - x(t_{j-1})\right| + w(x, [t_{j-1}, t_j)) \right) \\
&\qquad \vee |x_n(T) - x(T)| \\
&\le \max_{k=1}^N \left(\sup_{r_{j-1} \le t < r_j} \left|x_n(t) - x(\lambda_n(r_{j-1}))\right| + \epsilon \right) \\
&\qquad \vee |x_n(T) - x(T)| \\
&\le \max_{k=1}^N \left(\sup_{r_{j-1} \le t < r_j} |x_n(t) - x(\lambda_n(t))| \right. \\
&\qquad \left. + \left|x(\lambda_n(t)) - x(\lambda_n(r_{j-1}))\right| + \epsilon \right) \vee |x_n(T) - x(T)| \\
&\le \max_{k=1}^N \left(\sup_{r_{j-1} \le t < r_j} |x_n(t) - x(\lambda_n(t))| + w(x, [t_{j-1}, t_j)) + \epsilon \right) \\
&\qquad \vee |x_n(T) - x(T)| \\
&\le \max_{k=1}^N \left(\sup_{r_{j-1} \le t < r_j} |x_n(t) - x(\lambda_n(t))| + 2\epsilon \right) \vee |x_n(T) - x(T)| \\
&\le \rho_T(x_n, x \circ \lambda_n) + 2\epsilon \\
&\le 3\epsilon. \quad \blacksquare
\end{aligned}$$

Finally, we prove that F is continuous on a relevant set.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

Lemma 4.5. *F is continuous at $(x, y, 0)$ in the product topology on $\mathbb{D} \times \mathbb{D} \times \mathbb{R}$, for all x and $y \in \mathbb{D}$ with $Disc(x) \cap Disc(y) = \varnothing$ and*

$$\{t : y(t) - y(t-) < 0\} = \varnothing.$$

Proof. Let $T > 0$, let ρ_T be the uniform metric on function from $[0, T]$ to $\mathbb{R}$, and fix $\epsilon > 0$. Apply Lemma 1 on page 110 of [5] to construct finite subsets $A_1 = \{t'_j\}$ and $A_2 = \{s_j\}$ of $[0, T]$ such that $0 = t'_0 < \cdots < t'_k = T$, $0 = s_0 < \cdots < s_m = T$, $w(y; [t'_{j-1}, t'_j)) = \sup\{|y(s) - y(t)| : s, t \in [t'_{j-1}, t'_j)\} < \epsilon$ and $w(H(x, y, 0); [s_{j-1}, s_j)) < \epsilon$ for all j. Since $\text{Disc}(y) \cap \text{Disc}(H(x, y, 0)) \subset \text{Disc}(x) \cap \text{Disc}(y) = \varnothing$, the two sets A_1 and A_2 can be chosen so that $A_1 \cap A_2 = \{0, T\}$. Note that $w(y; [t_{j-1}, t_j)) < \epsilon$ and $w(H(x, y, 0); [t_{j-1}, t_j)) < \epsilon$ for $\{t_j\} = A_1 \cup A_2$. Let 2δ be the distance between the closest two points in $A_1 \cup A_2$. Choose n_0 and homeomorphisms λ_n and μ_n in Λ so that

(i) $\rho_T(y_n, y \circ \lambda_n) < (\delta \wedge \epsilon)$,
(ii) $\rho_T(\lambda_n, e) < (\delta \wedge \epsilon)$,
(iii) $\rho_T(H(x_n, y_n, c_n), H(x, y, 0) \circ \mu_n) < (\delta \wedge \epsilon)$, and
(iv) $\rho_T(\mu_n, e) < (\delta \wedge \epsilon)$

for $n \geq n_0$. Thus for $n \geq n_0$

$$\lambda_n^{-1}(A_1) \cap \mu_n^{-1}(A_2) = \{0, T\}$$

and $\{r_j\} = \lambda_n^{-1}(A_1) \cup \mu_n^{-1}(A_2)$ has corresponding points in the same order as $\{t_j\} = A_1 \cup A_2$. Let γ_n be homeomorphisms of $[0, T]$ defined by

$$\gamma_n(r_j) = t_j$$

for corresponding points $r_j \in \lambda_n^{-1}(A_1) \cup \mu_n^{-1}(A_2)$ and $t_j \in A_1 \cup A_2$ and by linear interpolation elsewhere.

Note that for each $r_j \in \lambda_n^{-1}(A_1) \cup \mu_n^{-1}(A_2)$ either

$$\lambda_n(r_j) = t_j \qquad \text{or} \qquad \mu_n(r_j) = t_j.$$

Since $t \mapsto |\gamma_n(t) - t|$ is continuous the maximum is attained at some critical point (exposed point) r_j, so $\rho_T(\gamma_n, e) < \rho_T(\lambda_n, e) \vee \rho_T(\mu_n, e) < \epsilon$. Now,

$$\begin{aligned}
&\rho_T(F(x_n, y_n, c_n), F(x, y, 0) \circ \gamma_n) \\
&\quad \leq \rho_T\left(\left(y_n - y_n^- + H(x_n, y_n, c_n)\right)^\uparrow, \left(\left(y - y^- + H(x, y, 0)\right)^\uparrow\right) \circ \gamma_n\right) \\
&\qquad + \rho_T\left(H(x_n, y_n, c_n), (H(x, y, 0)) \circ \gamma_n\right).
\end{aligned}$$

For the first term we have

$$\begin{aligned}
&\rho_T\left(\left(y_n - y_n^- + H(x_n, y_n, c_n)\right)^\uparrow, \left(\left(y - y^- + H(x, y, 0)\right)^\uparrow\right) \circ \gamma_n\right) \\
&\leq \rho_T\left(y_n - y_n^- + H(x_n, y_n, c_n), \left(y - y^- + H(x, y, 0)\right) \circ \gamma_n\right),
\end{aligned} \tag{8}$$

and

$$\begin{aligned}
&\rho_T\left(y_n - y_n^- + H(x_n, y_n, c_n), \left(y - y^- + H(x, y, 0)\right) \circ \gamma_n\right) \\
&\quad \leq \rho_T\left(y_n, y \circ \gamma_n\right) + \rho_T\left(y_n^-, y^- \circ \gamma_n\right) + \rho_T\left(H(x_n, y_n, c_n), H(x, y, 0) \circ \gamma_n\right).
\end{aligned} \tag{9}$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

Since γ_n is strictly increasing,

$$\begin{aligned}\rho_T\left(y_n^-, y^- \circ \gamma_n\right) &= \sup_{0\le t\le T}\left|\lim_{s\nearrow t} y_n(s) - \lim_{r\nearrow \gamma_n(t)} y(r)\right| \\ &= \sup_{0\le t\le T}\left|\lim_{s\nearrow t} y_n(s) - \lim_{r\nearrow t} y(\gamma_n(r))\right|,\end{aligned}$$

and so

$$\rho_T\left(y_n^-, y^- \circ \gamma_n\right) \le \sup_{0\le t\le T}\left|y_n(t) - y(\gamma_n(t))\right|,$$

since the left limit of y_n and $y\circ\gamma_n$ exists at each t. Therefore,

$$\rho_T\left(y_n^-, y^- \circ \gamma_n\right) \le \rho_T\left(y_n, y\circ\gamma_n\right). \tag{10}$$

Combining (8)–(10) we have,

$$\begin{aligned}&\rho_T(F(x_n, y_n, c_n), F(x, y, 0)\circ\gamma_n) \\ &\le \rho_T\left(\left(y_n - y_n^- + H(x_n, y_n, c_n)\right)^\uparrow, \left(\left(y - y^- + H(x, y, 0)\right)^\uparrow\right)\circ\gamma_n\right) \\ &\quad + \rho_T\left(H(x_n, y_n, c_n), H(x, y, 0)\circ\gamma_n\right) \\ &\le 2\rho_T\left(y_n, y\circ\gamma_n\right) + 2\rho_T\left(H(x_n, y_n, c_n), H(x, y, 0)\circ\gamma_n\right) \\ &\le 12\epsilon,\end{aligned}$$

by Lemma 4.4. ■

5. Scaling limit of the plateau process

In this section we prove several results concerning the sequence of models, and then combine these to prove Theorem 2.1. We begin by showing that the function H scales nicely when no centering is required.

Lemma 5.1. *For positive constants a_n and n,*

$$a_n^{-1}H(x, y, c)(nt) = H(x^n, y^n, c/n)(t),$$

for all $t \ge 0$, where $x^n(t) = a_n^{-1}x(nt)$ and $y^n(t) = a_n^{-1}y(nt)$.

Proof. By definition,

$$\begin{aligned}a_n^{-1}H(x, y, c)(nt) &= a_n^{-1}\sup_{0\le s\le nt}\left(x(s) - y([s-c]^+)\right) \\ &= \sup_{0\le s\le t}\left(a_n^{-1}x(ns) - a_n^{-1}y([ns-c]^+)\right) \\ &= \sup_{0\le s\le t}\left(a_n^{-1}x(ns) - a_n^{-1}y(n\,[s-c/n]^+)\right) \\ &= \sup_{0\le s\le t}\left(x^n(s) - y^n([s-c/n]^+)\right) \\ &= H(x^n, y^n, c/n)(t) \quad ■\end{aligned}$$

Lemma 5.2. *The set $\mathscr{K} = \{x\in\mathbb{D} : x(t) - x(t-) \ge 0$ for each $t\in(0,\infty)\}$ is closed in $\mathbb{D}$.*

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

Proof. Let $\{x_n\}$ be a sequence in $\mathscr{K}$ such that $x_n \to x$. Fix $t_0 \in (0,\infty)$ with $x(t_0)-x(t_0-) \neq 0$. There exists $t_n \to t_0$ with $x_n(t_n) - x_n(t_n-) \to x(t_0) - x(t_0-)$ by [13] proposition VI.2.1. We have $x_n(t_n) - x_n(t_n-) \geq 0$ for each n since $x_n \in \mathscr{K}$, so $x(t_0) - x(t_0-) \geq 0$ and we must have $x \in \mathscr{K}$. ■

The next Lemma establishes a joint convergence involving the primitive input processes. Recall that $\check{U}^r \Rightarrow U^*$ and $\check{V}^r \Rightarrow V^*$ in $\mathbb{D}$.

Lemma 5.3. *For any sequence of real numbers $c_r \to c$,*

$$(\check{U}^r + c_r e, \check{V}^r, 1/r) \Rightarrow (U^* + ce, V^*, 0),$$

in the product topology on $\mathbb{D} \times \mathbb{D} \times \mathbb{R}$. Moreover,

$$Disc(U^* + ce) \cap Disc(V^*) = \varnothing \text{ a.s.}$$

and $\{t : V^*(t) - V^*(t-) < 0\} = \varnothing$ a.s.

Proof. Since ce is continuous, $\check{U}^r \Rightarrow U^*$, and $c_r e \Rightarrow ce$ we have $\check{U}^r + c_r e \Rightarrow U^* + ce$ by [18]. We have joint convergence $(\check{U}^r + c_r e, \check{V}^r) \Rightarrow (U^* + ce, V^*)$ since $\check{V}^r$ is independent of $\check{U}^r$ and therefore $\check{U}^r + c_r e$ is independent of $\check{V}^r$ because c_r is constant in ω, [19] Theorem 11.4.4, moreover U^* is independent of V^*. Since $1/r$ is constant in ω we have $1/r \to 0$ in probability so [5] Theorem 4.4 gives joint convergence

$$(\check{V}^r + c_r e, \check{U}^r, 1/n) \Rightarrow (U^* + ce, V^*, 0).$$

V^* is a stable Lévy motion by 2.4.1 of the online supplement to [19]. So V^* has no fixed discontinuities: $\mathbb{P}\{U^*(t) = U^*(t-)\} = 1$ for all $t \in (0,\infty)$. By [18] Lemma 4.3, gives $\mathbb{P}\{\text{Disc}(U^*) \cap \text{Disc}(V^*) = \varnothing\} = 1$ and since ce is continuous we have

$$\mathbb{P}\left\{\text{Disc}(U^* + ce) \cap \text{Disc}(V^*) = \varnothing\right\} = 1.$$

Finally, $\mathbb{P}\left\{\check{V}^r \in \mathscr{K}\right\} = 1$, $\check{V}^r \Rightarrow V^*$, and $\mathscr{K}$ is closed by Lemma 5.2, so the Portmanteau theorem gives

$$\mathbb{P}\left\{V^* \in \mathscr{K}\right\} \geq \limsup_{n\to\infty} \mathbb{P}\left\{\check{V}^r \in \mathscr{K}\right\} = 1. \quad ■$$

For each $r > 0$ and $t \geq 0$ define $\bar{D}^r(t) = \frac{1}{r} D^r(rt)$. Using Corollary 3.4 under this fluid scaling, we have for all $t \geq 0$,

$$\bar{R}^r(t) = \frac{1}{r} R(rt).$$

We will need the fluid limit of $\bar{D}^r(\cdot)$.

Lemma 5.4. *As $r \to \infty$,*

$$\bar{R}^r \Rightarrow e/\mu$$

Proof. $\check{U}^r(1) \Rightarrow U^*(1)$ implies $\frac{r}{a_r}\left(\bar{U}^r(1) - \mu^r\right) \Rightarrow U^*(1)$, but $r/a_r \to \infty$ implies $\bar{U}^r(1) - \mu_r \Rightarrow 0$. Since $\mu^r \to \mu$ we have $\bar{U}^r(1) \Rightarrow \mu$. By Theorem 2.4.1 of the internet

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

supplement to [19], we have $\bar{U}^r \Rightarrow \mu e$ in $\mathbb{D}$. Similarly, $\bar{V}^r \Rightarrow \mu e$ in $\mathbb{D}$. Now compute

$$\begin{aligned}
\bar{R}^r(t) &= \frac{1}{r}\sup\left\{m \geq 0 : V^r(m) + H(U^r, V^r, 1)(m) \leq rt\right\} \\
&= \sup\left\{x/r \geq 0 : V^r(x) + H(U^r, V^r, 1)(x) \leq rt\right\} \\
&= \sup\left\{x/r \geq 0 : \frac{V^r(x)}{r} + \frac{1}{r}H(U^r, V^r, 1)(x) \leq t\right\} \\
&= \sup\left\{y \geq 0 : \frac{V^r(ry)}{r} + \frac{1}{r}H(U^r, V^r, 1)(ry) \leq t\right\} \\
&= \sup\left\{y \geq 0 : \bar{V}^r(y) + H(\bar{U}^r, \bar{V}^r, 1/r)(y) \leq t\right\},
\end{aligned}$$

by Lemma 5.1. We have $(\bar{U}^r, \bar{V}^r, 1/r) \Rightarrow (\mu e, \mu e, 0)$ in $\mathbb{D}$ since the processes are independent. The function H is continuous at $(\mu_u e, \mu_v e, 0)$, and addition is continuous at continuous elements of $\mathbb{D}$, so

$$\bar{V}^r + H(\bar{U}^r, \bar{V}^r, 1/r) \Rightarrow \mu e$$

in $\mathbb{D}$. The result follows because μe is in the set of continuity for the function $x \mapsto \sup\{y \geq 0 : x(y) \leq t\}$ by Corollary 13.6.4 in [19]. ■

We now prove the main result.

Proof of Theorem 2.1. By Lemma 3.5

$$M(t) = F(U^r, V^r, 1)(R(t)).$$

Under fluid scaling $\bar{R}^r \Rightarrow e/\mu$ by 5.4. We first consider the scaling limit for F, before composing with R.

$$\begin{aligned}
a_r^{-1}F(U^r, V^r, 1)(rt) &= a_r^{-1}\sup_{0\leq s\leq rt}\left(V^r(s) - V^r(s-) + H(U^r, V^r, 1)(s)\right) \\
&\quad - a_r^{-1}H(U^r, V^r, 1)(rt) \\
&= \sup_{0\leq s\leq rt}\left(a_r^{-1}V^r(s) - a_r^{-1}V^r(s-) + a_r^{-1}H(U^r, V^r, 1)(s)\right) \\
&\quad - a_r^{-1}H(U^r, V^r, 1)(rt) \\
&= \sup_{0\leq s\leq t}\left(a_r^{-1}V^r(rs) - a_r^{-1}V^r(rs-) + a_r^{-1}H(U^r, V^r, 1)(rs)\right) \\
&\quad - a_r^{-1}H(U^r, V^r, 1)(rt).
\end{aligned}$$

$t \mapsto rv^r t$ is continuous so $rv^r(rs) - rv^r(rs-) = 0$ and

$$\begin{aligned}
a_r^{-1}F(U^r, V^r, 1)(rt) = \sup_{0\leq s\leq t}&\left(\check{V}^r(s) - \check{V}^r(s-) + a_r^{-1}H(U^r, V^r, 1)(rs)\right) \\
&- a_r^{-1}H(U^r, V^r, 1)(rt).
\end{aligned} \tag{11}$$

Now, we address the idleness part of (11) that occurs twice.

$$\begin{aligned}
&a_r^{-1} H(U^r, V^r, 1)(rt) \\
&\quad = a_r^{-1} \sup_{0\le s\le rt} \left(U^r(s) - V^r([s-1]^+)\right) \\
&\quad = \sup_{0\le s\le t} \left(a_r^{-1} U^r(rs) - a_r^{-1} V^r(r[s-1/r]^+)\right) \\
&\quad = \sup_{0\le s\le t} \Bigg(a_r^{-1}\left(U^r(rs) - r\mu^r s\right) + a_r^{-1} r\mu^r s \\
&\qquad\quad - a_r^{-1}\left(V^r(r[s-1/r]^+) - r\nu^r[s-1/r]^+\right) - a_r^{-1} r\nu^r[s-1/r]^+ \Bigg) \\
&\quad = \sup_{0\le s\le t} \left(\check{U}^r(s) + a_r^{-1} r\mu^r s - \check{V}^r([s-1/r]^+) - a_r^{-1} r\nu^r[s-1/r]^+ \right) \\
&\quad = \sup_{0\le s\le t} \Bigg(\check{U}^r(s) + a_r^{-1} r(\mu^r - \nu^r)s + a_r^{-1} r\nu^r(s - [s-1/r]^+) \\
&\qquad\quad - \check{V}^r([s-1/r]^+) \Bigg).
\end{aligned}$$

Since

$$a_r^{-1} r\nu^r(s - [s-1/r]^+) = a_r^{-1} r\nu^r(1/r \wedge s) = a_r^{-1}\nu^r(1 \wedge rs),$$

we have

$$\begin{aligned}
&a_r^{-1} H(U^r, V^r, 1)(rt) \\
&\quad = H(\check{U}^r + a_r^{-1} r(\mu^r - \nu^r)e + a_r^{-1}\nu^r(1 \wedge re), \check{V}^r, 1/r)(t).
\end{aligned}$$

Putting this expression back into (11),

$$\begin{aligned}
a_r^{-1} F(U^r, V^r, 1)(rt) &= \sup_{0\le s\le t} \Big[\check{V}^r(s) - \check{V}^r(s-) \\
&\quad + H(\check{U}^r + a_r^{-1} r(\mu^r - \nu^r)e + a_r^{-1}\nu^r(1 \wedge re), \check{V}^r, 1/r)(s) \Big] \\
&\quad - H(\check{U}^r + a_r^{-1} r(\mu^r - \nu^r)e + a_r^{-1}\nu^r(1 \wedge re), \check{V}^r, 1/r)(t) \\
&= F(\check{U}^r + a_r^{-1} r(\mu^r - \nu^r)e + a_r^{-1}\nu^r(1 \wedge re), \check{V}^r, 1/r)(t).
\end{aligned}$$

By Lemma 5.3 we have $(U^* + \gamma\mu e, V^*, 0)$ satisfies the continuity criterion of Lemma 4.5. By the continuous mapping theorem

$$F(\check{U}^r + a_r^{-1} r(\mu^r - \nu^r)e + a_r^{-1}\nu^r(1 \wedge re), \check{V}^r, 1/r) \Rightarrow F(U^* + \gamma\mu e, V^*, 0).$$

Finally, the scaled plateau process is a composition of F with R,

$$a_r^{-1} F(U^r, V^r, 1)(R(rt)) = a_r^{-1} F(U^r, V^r, 1)(r\bar{R}^r(t)).$$

Composition is continuous on $(\mathbb{D}\times C_0)$ by [18] Theorem 3.1, where $C_0 \subset \mathbb{D}$ denotes the strictly increasing, continuous functions. So the continuous mapping theorem yields

$$a_r^{-1} M^r(r\cdot) = \check{M}^r \Rightarrow M^* = F(U^* + \gamma\mu e, V^*, 0)(\cdot/\mu). \quad \blacksquare$$

6. Analysis of the limit process

In this section we derive, for certain cases, some properties of the stochastic process M^* that appears as the scaling limit of the plateau process. We focus on cases where the interarrival

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

time distribution has finite variance (but the service time distribution still has infinite variance), leading to a trivial limit for the arrival process $U^*(t) \equiv 0$ and a non-trivial α-stable process V^* for the limit of the service process. By Theorem 2.1, the limit of the plateau process is then

$$M^*(t) = F(\gamma\mu e, V^*, 0)(t/\mu), \qquad t \geq 0.$$

Although this process is not Markov, a suitable time change of it is and has one-dimensional distributions that can be derived explicitly. The time change is simply an inverse local time of the reflected (at zero) version of the process $V^*(t) - \gamma\mu t$, $t \geq 0$. More explicitly, letting $X(t) = V^*(t) - \gamma\mu t$ and $\underline{X}(t) = \inf_{0\leq s\leq t} X(s)$, the process $L(t) = -\underline{X}(t)$ is the local time at zero for the reflected process $Y(t) = X(t) - \underline{X}(t)$ associated with X. We use its right-continuous inverse L^{-1} to define the time-changed version

$$Z(v) = M^*(\mu L^{-1}(v)), \qquad v \geq 0,$$

of our limit process. The one-dimensional distributions of Z are given by the following.

Theorem 6.1. *If the limiting arrival process is identically zero, then for each $v \geq 0$, the distribution function F_v of $Z(v)$ is given by*

$$F_v(y) = \exp\left(-\int_y^{y+v} \frac{\kappa(q)}{q} dq\right), \qquad y \geq 0, \tag{12}$$

where $\kappa(q)/q = \phi_q(c_\alpha q^{-\alpha}/\alpha)$, ϕ_q is the right-inverse of

$$s \mapsto s + s^\alpha + c_\alpha \int_q^\infty (1 - e^{-sx}) x^{-\alpha-1} dx,$$

and c_α is an explicit constant (see below).

Not only does this result provide some means to perform calculations on the process Z (and thus on the process M^*), but it also allows us to relate Theorem 2.1 to the results obtained in [11]. In particular, comparing with Theorem 2.2 in [11], we see that the above one-dimensional distributions of our time changed limit process $Z(v) = M^*(\mu L^{-1}(v))$ are precisely the limiting laws of the one-dimensional distributions of the process studied in [11] (a discrete-time Markov chain embedded in the plateau process), in which the analogous time change was performed on the original (prelimit) process before scaling and taking the limit.

The remainder of this section provides the proof.

6.1. Proof of Theorem 6.1

Since $U^* \equiv 0$, we are using the function

$$F(\gamma\mu e, y, 0)(t/\mu) = \sup_{0\leq s\leq t/\mu} [y(s) - y(s-) + \sup_{0\leq r\leq s} [\gamma\mu r - y(r)]] - \sup_{0\leq s\leq t/\mu} [\gamma\mu s - y(s)],$$

where y is replaced by the α-stable process V^*. Using the definition of $X(t)$ and $\underline{X}(t)$, this expression reduces to

$$\begin{aligned} F(\gamma\mu e, V^*, 0)(t/\mu) &= \sup_{0\leq s\leq t/\mu} [X(s) - X(s-) - \inf_{0\leq r\leq s} X(r)] + \inf_{0\leq s\leq t/\mu} X(s) \\ &= \sup_{0\leq s\leq t/\mu} [X(s) - X(s-) + \underline{X}(t/\mu) - \underline{X}(s)]. \end{aligned} \tag{13}$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

Recall that $L(t) = -\underline{X}(t)$ is the local time at zero for the reflected process $Y(t) = X(t) - \underline{X}(t)$ associated with X. For $v \geq 0$ define

$$Z(v) = \sup_{0 \leq s \leq L^{-1}(v)} [X(s) - X(s-) - (v - L(s))],$$

where L^{-1} denotes the right-continuous inverse. Then from (13) we see that for times t such that $L(t/\mu) = v$, $M^*(t) = Z(v)$. Put another way, we have for all $t \geq 0$ that $M^*(\mu L^{-1}(L(t/\mu))) = Z(L(t/\mu))$. That is, the process $Z(v) = M^*(\mu L^{-1}(v))$ is a certain time-changed (and embedded) version of the process M^*, evaluated at times (scaled by μ) when the local time of Y has attained the level v. We now examine the one-dimensional distributions of the process Z.

For each $v \geq 0$ we will derive the distribution function $F_v(y)$, $y \geq 0$, of $Z(v)$ using some calculations from excursion theory (note that $Z(v)$ is a nonnegative random variable). As in Chapter 4 of Bertoin [4], define $N = ((v, \varepsilon(v)), v \geq 0)$ as the Poisson point process of excursions away from 0 for the reflected process Y. That is, $(v, \varepsilon(v))$ takes values in $[0, \infty) \times \mathscr{E}$, where $\mathscr{E}$ is the space of excursions from zero, and $\varepsilon(v)$ corresponds to the excursion of Y beginning when its local time has attained level v. Let ℓ denote Lebesgue measure and denote by n the excursion measure of Y, which is the sigma-finite measure on $\mathscr{E}$ such that $\ell \times n$ is the intensity on $[0, \infty) \times \mathscr{E}$ of the Poisson random measure N.

Defining $\Delta(v) = \Delta(\varepsilon(v))$ to be the largest jump made during the excursion $\varepsilon(v)$ (which we set to be 0 if there is no excursion at v), we see that

$$Z(v) = \sup_{0 \leq u \leq v} [\Delta(u) - (v - u)]. \tag{14}$$

Since $N' = \left(\sum_v \delta_{(v, \Delta(v))}, v \geq 0\right)$ is a Poisson point process on $[0, \infty) \times [0, \infty)$, the process Z is Markov. Note that for any $w \in [0, v]$,

$$\begin{aligned} Z(v) &= \max\{\sup_{0 \leq u \leq w} [\Delta(u) - (w - u)] - (v - w), \sup_{w \leq u \leq v} [\Delta(u) - (v - u)]\} \\ &= \max\{Z(w) - (v - w), \sup_{w \leq u \leq v} [\Delta(u) - (v - u)]\} \\ &\sim \max\{Z(w) - (v - w), \sup_{u \in [0, v-w]} [\Delta(u) - u]\}. \end{aligned}$$

In particular, taking $w = 0$, we obtain

$$Z(v) \sim \sup_{0 \leq u \leq v} [\Delta(u) - u]. \tag{15}$$

Define $A = A_{v,y} = \{(u, \varepsilon) \in [0, \infty) \times \mathscr{E} : u \in [0, v], \Delta(\varepsilon) > y + u\}$. Then using standard results (e.g. Section 0.5 of Bertoin [4]), we see that for $y > 0$,

$$P(Z(v) > y) = P(N(A) \geq 1). \tag{16}$$

The random variable $N(A)$ is Poisson with mean

$$\lambda(v, y) = (\ell \times n)(A) = \int_0^v n(\Delta(\varepsilon) > y + u) du = \int_y^{y+v} n(\Delta(\varepsilon) > q) dq. \tag{17}$$

So the distribution function of $Z(v)$ is $F_v(y) = \exp(-\lambda(v, y))$, $y > 0$, which is explicit as long as we can derive an expression for $n(\Delta(\varepsilon) > q)$ for each $q > 0$.

To this end, fix $q > 0$. The idea is to compare the set of excursions with a jump bigger than q to the set of excursions of a modified process, whose lifetimes are longer than the exponential waiting time until the first q-jump of the original process. The modified process $\tilde{Y}$ is obtained

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

from Y by thinning all jumps of size greater than q, yielding a Lévy process for which the Lévy measure is now restricted to $[0, q]$, so that we may apply a formula of Baurdoux [3] for excursion lifetimes.

In more detail, write $X = \tilde{X} + J_q$, where J_q is a pure jump process independent of $\tilde{X}$ with all jumps greater than q, and $\tilde{X}$ almost surely has all jumps bounded by q. Define the modified process $\tilde{Y}(t) = \tilde{X}(t) - \underline{\tilde{X}}(t)$ and let $\tilde{n}$ denote the excursion measure on $\mathscr{E}$ of the process $\tilde{Y}$.

The Laplace exponent of the Lévy process X is $\Psi(s) = s + s^\alpha$, and the corresponding Lévy measure $\nu(dx) = c_\alpha x^{-\alpha-1}dx$, for a strictly positive constant c_α (an expression is given in Exercise 1.4 of [16]). So the Lévy measures of $\tilde{X}$ and J_q are ν restricted to $[0, q]$ and (q, ∞) respectively. The Lévy exponent of $\tilde{X}(t)$ can be written as

$$\tilde{\Psi}_q(s) = s + s^\alpha + c_\alpha \int_q^\infty (1 - e^{-sx})x^{-\alpha-1}dx. \tag{18}$$

Define

$$\mathrm{e}_q = \inf\{t \geq 0 : \ Y(t) - Y(t-) > q\}$$

as the waiting time until the first jump of Y of size greater than q. Then $\mathrm{e}_q = \inf\{t \geq 0 : J_q(t) > J_q(t-)\}$, and since J_q is independent of $\tilde{X}$, the random variable e_q is exponential with rate $\beta_q = \nu(q, \infty) = c_\alpha q^{-\alpha}/\alpha$ and is independent of $\tilde{X}$.

Lemma 6.2. *For each $q > 0$,*

$$n(\Delta(\varepsilon) > q) = \int_{\mathscr{E}} (1 - e^{-\beta_q|\varepsilon|})d\tilde{n}(\varepsilon), \tag{19}$$

where $|\varepsilon|$ denotes the lifetime of an excursion $\varepsilon \in \mathscr{E}$.

Proof. We show that both expressions are equal to $1/E[L(\mathrm{e}_q)]$. Beginning with the left side, multiply and divide by $E[L(\mathrm{e}_q)]$ to obtain

$$\begin{aligned} n(\Delta(\varepsilon) > q) &= \int_{\mathscr{E}} 1_{\{\Delta(\varepsilon)>q\}} dn(\varepsilon) \\ &= \frac{1}{E[L(\mathrm{e}_q)]} E\left[\int_0^\infty \int_{\mathscr{E}} 1_{\{\Delta(\varepsilon)>q\}} 1_{\{s\leq \mathrm{e}_q\}} dn(\varepsilon)dL(s)\right]. \end{aligned}$$

We show the second expectation on the right equals one. Since the function $G(s, \omega, \varepsilon) = 1_{\{\Delta(\varepsilon)>q\}} 1_{\{s\leq \mathrm{e}_q(\omega)\}}$ on $[0, \infty) \times \Omega \times \mathscr{E}$ is measurable and almost surely left-continuous in s, the compensation formula in excursion theory (see Corollary 11 in Section IV.4 of [4]) yields

$$E\left[\int_0^\infty \int_{\mathscr{E}} 1_{\{\Delta(\varepsilon)>q\}} 1_{\{s\leq \mathrm{e}_q\}} dn(\varepsilon)dL(s)\right] = E\left[\sum_g 1_{\{\Delta(\varepsilon_g)>q\}} 1_{\{g\leq \mathrm{e}_q\}}\right],$$

where for each sample path, the sum is over the left endpoints of all excursion intervals (g, d) and ε_g is the excursion of Y beginning at time g. But since e_q falls during the first excursion with a jump greater than q, the sum equals one almost surely.

Turning to the right side of (19), we again multiply and divide, noting that $E[L(\mathrm{e}_q)] = E[\tilde{L}(\mathrm{e}_q)]$, where $\tilde{L}(t) = -\underline{\tilde{X}}(t)$ is the local time for $\tilde{Y}$, because the sample paths of Y and $\tilde{Y}$ are identical up to time e_q. This gives

$$\int_{\mathscr{E}} (1 - e^{-\beta_q|\varepsilon|})d\tilde{n}(\varepsilon) = \frac{1}{E[L(\mathrm{e}_q)]} E\left[\int_0^\infty \int_{\mathscr{E}} (1 - e^{-\beta_q|\varepsilon|}) 1_{\{s\leq \mathrm{e}_q\}} d\tilde{n}(\varepsilon)d\tilde{L}(s)\right],$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

and we must show the second expectation on the right equals one. Using the compensation formula,

$$E\left[\int_0^\infty \int_{\mathscr{E}} (1-e^{-\beta_q|\varepsilon|})1_{\{s\le \mathrm{e}_q\}} d\tilde{n}(\varepsilon) d\tilde{L}(s)\right] = E\left[\sum_g (1-e^{-\beta_q|\varepsilon_g|})1_{\{g\le \mathrm{e}_q\}}\right],$$

where this time the sum is over all excursion intervals (g,d) of $\tilde{Y}$ and ε_g are the corresponding excursions. Since e_q is independent of $\tilde{Y}$, the expectation on the right can be computed as an iterated integral over $\mathbb{D}\times[0,\infty)$ with respect to the product law $P_{\tilde{Y}}\times P_e$ of the random pair $(\tilde{Y},\mathrm{e}_q)$. This yields

$$\begin{aligned} E_{\tilde{Y}}E_e\left[\sum_g (1-e^{-\beta_q|\varepsilon_g|})1_{\{g\le \mathrm{e}_q\}}\right] &= E_{\tilde{Y}}\left[\sum_g (1-e^{-\beta_q|\varepsilon_g|})P_e(g\le \mathrm{e}_q)\right] \\ &= E_{\tilde{Y}}\left[\sum_g P_e(|\varepsilon_g|>\mathrm{e}_q)P_e(\mathrm{e}_q>g)\right]. \end{aligned}$$

Note that for each excursion interval (g,d), the lifetime $|\varepsilon_g| = d-g$. So by the memoryless property of the exponential and since the excursion intervals are disjoint, the right side above is equal to

$$\begin{aligned} E_{\tilde{Y}}\left[\sum_g P_e(\mathrm{e}_q < d \,|\, \mathrm{e}_q > g)P_e(\mathrm{e}_q>g)\right] &= E_{\tilde{Y}}\left[\sum_g P_e\left(\mathrm{e}_q\in(g,d)\right)\right] \\ &= E_{\tilde{Y}}\left[P_e\left(\mathrm{e}_q\in[0,\infty)\setminus\overline{\mathscr{Z}}\right)\right], \end{aligned}$$

where $\overline{\mathscr{Z}}$ denotes the closure of the zero set of $\tilde{Y}$. Since $X = V^* - \gamma\mu t$ is not a monotone or pure jump process, this set has Lebesgue measure zero and the right side above equals one. ■

Since the Lévy measure of $\tilde{X}$ has bounded support, we can apply Equation (3.3) of [3] to the right side of (19), which in the notation of [3] would be written "$\tilde{n}(|\varepsilon|>\mathrm{e}_q)$". Let P_x denote the law of $\tilde{X}+x$ and $\tau_q^x = \inf\{t\ge 0 : \tilde{X}(t)+x=0\}$ be the hitting time of zero. Then (19) combined with [3] Equation (3.3) in our setting (in particular $h(x)$ there is simply x here) yields

$$n(\Delta(\varepsilon)>q) = \lim_{x\downarrow 0}\frac{P_x(\tau_q^x>\mathrm{e}_q)}{x} = \lim_{x\downarrow 0}\frac{1-E_x[e^{-\beta_q\tau_q^x}]}{x}.$$

Observe that

$$E_x[e^{-\beta_q\tau_q^x}] = e^{-x\phi_q(\beta_q)},$$

with ϕ_q the right inverse of $\tilde{\Psi}_q$.

Thus, we obtain

$$n(\Delta(\varepsilon)>q) = \phi_q(\beta_q) = \phi_q(c_\alpha q^{-\alpha}/\alpha) =: h(q). \tag{20}$$

Rewrite the last expression using (18) to get

$$c_\alpha q^{-\alpha}/\alpha = \tilde{\Psi}_q(h(q)) = h(q) + h(q)^\alpha + c_\alpha\int_q^\infty (1-e^{-h(q)x})x^{-\alpha-1}dx.$$

This can be simplified to

$$h(q)+h(q)^\alpha = c_\alpha\int_q^\infty e^{-h(q)x}x^{-\alpha-1}dx.$$

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.

Defining $\kappa(q) = h(q)q$, performing a change of variables $t = x/q$, and letting T_α be a Pareto distributed random variable with index α we obtain

$$q^{\alpha-1}\kappa(q) + \kappa(q)^\alpha = \frac{c_\alpha}{\alpha} E[e^{-\kappa(q)T_\alpha}]. \tag{21}$$

This equation can be transformed into Equation (7) in [11] for $\kappa(y)$ (using $\lambda = 1$ and $\gamma = -\Gamma(1-\nu)$ there), and so we see by Lemma 3.9 of [11] that (21) has a unique solution $\kappa(q)$, which by Lemma 3.11 in [11] is a continuous, bounded, regularly varying function of q with index $1-\alpha$. Combining (17) with (20) establishes (12) and proves Theorem 6.1.

Acknowledgment

BZ is supported by NWO grant #639.013.433.

References

[1] S. Asmussen, Applied Probability and Queues, John Wiley and Sons, 1987.
[2] R. Ballerini, S.I. Resnick, Records in the presence of a linear trend, Adv. Appl. Probab. 19 (4) (1987) 801–828, http://dx.doi.org/10.2307/1427103.
[3] E.J. Baurdoux, Some excursion calculations for reflected Lévy processes, Lat. Am. J. Probab. Math. Stat. 6 (2009) 149–162.
[4] J. Bertoin, Lévy Processes, Cambridge University Press, 1996.
[5] P. Billingsley, Convergence of Probability Measures, Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics, Wiley, 1968, URL https://books.google.com/books?id=O9oQAQAAIAAJ.
[6] O.J. Boxma, Analysis of Models for Tandem Queues (Ph.D. thesis), University of Utrecht, Utrecht, 1977.
[7] O.J. Boxma, On the longest service time in a busy period of the M\G\1 queue, Stochastic Process. Appl. 8 (1) (1978) 93–100.
[8] O.J. Boxma, On a tandem queueing model with identical service times at both counters, i, Adv. Appl. Probab. (1979) 616–643.
[9] S.N. Ethier, T.G. Kurtz, Markov Processes: Characterization and Convergence, 282, John Wiley & Sons, 2009.
[10] W. Feller, An Introduction to Probability and Its Applications, Vol. ii, Wiley, New York, 1971.
[11] H.C. Gromoll, B. Terwilliger, B. Zwart, Heavy traffic limit for a tandem queue with identical service times, Queueing Syst. 89 (2018) 213–241, http://dx.doi.org/10.1007/s11134-017-9560-z.
[12] J.M. Harrison, The diffusion approximation for tandem queues in heavy traffic, Adv. Appl. Probab. 10 (1978) 886–905.
[13] J. Jacod, A.N. Shiryaev, Limit Theorems for Stochastic Processes, in: A Series of Comprehensive Studies in Mathematics, vol. 288, Springer-Verlag, 1987.
[14] F.I. Karpelevich, A.Y. Kreĭnin, Heavy Traffic Limits for Multiphase Queues, in: Translations of Mathematical Monographs, 137, American Mathematical Society, Providence, RI, 1994, p. xii+143, Translated from the Russian manuscript by Kreĭnin and A. Vainstein.
[15] F.I. Karpelevitch, A.Y. Kreinin, Asymptotic analysis of queueing systems with identical service, J. Appl. Probab. 33 (1) (1996) 267–281.
[16] A. Kyprianou, Fluctuations of Lévy Processes with Applications, Springer, 2014.
[17] M.I. Reiman, L.M. Wein, Heavy traffic analysis of polling systems in tandem, Oper. Res. 47 (1999) 524–534.
[18] W. Whitt, Some useful functions for functional limit theorems, Math. Oper. Res. 5 (1) (1980) 67–85.
[19] W. Whitt, Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and Their Application to Queues, Springer Science & Business Media, 2002.

Please cite this article as: H.C. Gromoll, B. Terwilliger and B. Zwart, Heavy traffic limit for the workload plateau process in a tandem queue with identical service times, Stochastic Processes and their Applications (2019), https://doi.org/10.1016/j.spa.2019.05.007.