

HEAVY-TRAFFIC APPROXIMATIONS FOR A LAYERED NETWORK WITH LIMITED RESOURCES

BY

ANGELOS AVEKLOURIS (EINDHOVEN), MARIA VLASIOU (EINDHOVEN),
JIHENG ZHANG (HONG KONG), AND BERT ZWART (EINDHOVEN)

The paper is dedicated to Professor Tomasz Rolski. The authors would like to wish Professor Tomasz Rolski many more happy, productive and inspirational years. We are grateful for his guidance, friendship, and insights throughout our careers. The legacy he has built has had a profound impact on our academic lives. It has enriched our collaborations, formed friendships, and made us broader as researchers and human beings.

Abstract. Motivated by a web-server model, we present a queueing network consisting of two layers. The first layer incorporates the arrival of customers at a network of two single-server nodes. We assume that the interarrival and the service times have general distributions. Customers are served according to their arrival order at each node and after finishing their service they can re-enter at nodes several times for another service. At the second layer, active servers act as jobs that are served by a single server working at speed one in a processor-sharing fashion. We further assume that the degree of resource sharing is limited by choice, leading to a limited processor-sharing discipline. Our main result is a diffusion approximation for the process describing the number of customers in the system. Assuming a single bottleneck node and studying the system as it approaches heavy traffic, we prove a state-space collapse property.

2010 AMS Mathematics Subject Classification: Primary: 60K25, 90B15; Secondary: 68M20.

Key words and phrases: Layered queueing network, limited processor sharing, fluid model, diffusion approximation, heavy traffic.

1. INTRODUCTION

We consider a network with a two-layered architecture. The first layer models the processing of customers by a network of two nodes. Each node can have multiple (but finitely many) servers. Customers are served according to their order of arrival and after finishing their service, they can re-enter at nodes several times for

new services. The servers of the first layer act as jobs in the second layer, where they are simultaneously served by a common server working at speed one according to processor sharing with rates depending on the number of customers in the system. Our goal is to derive an explicit approximation of the process describing the number of customers in the system.

We analyze the system as it approaches heavy traffic. Under the assumption that there is a single bottleneck, we derive explicit results for the joint distribution of the number of customers in the system by proving a diffusion limit theorem. To achieve this, we look at the system in the second layer. In this way, we can aggregate the whole system since the total workload of the system (including the future workload due to customers re-entering the queues) acts as if were that of a single server queue with two independent renewal inputs.

To derive our diffusion limit theorem, we carry out a program inspired by the work of Bramson [1] and Williams [14], which consists of two main steps. First, we consider a critical fluid model, which can be thought of as a formal law of large numbers approximation under appropriate scaling. We identify the invariant states for the critical fluid model and we study the convergence to equilibrium of critical fluid model solutions as time goes to infinity. Our analysis has some similarities with the head-of-the-line processor-sharing discipline as studied in [1], but there are differences. In particular, as the degree of resource sharing at each node is finite in our case, we need to define appropriate spatial regions in which the fluid model solutions have qualitatively different behavior. Our main result is to show that a solution of the fluid model converges to equilibrium uniformly (in terms of the initial condition) on compact sets. To achieve this, we perform a time change that facilitates our analysis.

The second main step is to show a state-space collapse property for the joint queue length vector process in heavy traffic. For an appropriately defined sequence of stochastic processes, we show that the difference between this vector and an appropriate deterministic mapping of the one-dimensional total workload process vanishes. The latter process is shown to converge to a one-dimensional reflected Brownian motion.

Our work can be seen as a partial network extension of the limited processor-sharing queue of which fluid, diffusion, and steady-state heavy-traffic limit theorems have been derived [18]–[20]. In our model, we assume that the inter-arrival and the service times have general distributions, but we consider that only one customer at each node can receive service at any time. In case that the inter-arrival and the service times are exponential, the service discipline at each station becomes irrelevant. An extension in the direction of general service times, using processor sharing at each node would require measure valued processes and is beyond the scope of the present paper. A mostly heuristic description of the results in this paper has appeared in [11]. In the classical applied probability literature, a version of our model has been investigated in a steady-state setting using boundary value

techniques [6]; the solution in that paper may be used for numerical purposes, and is complementary to our heavy-traffic limit, which yields explicit formulae, both for time-dependent as well as steady-state results.

In addition, our work is a contribution to the performance analysis of layered queueing networks. These are queueing networks where some entities in the system have a dual role (e.g., servers become customers to a higher layer). In such systems, the dynamics in layers are correlated and the service speeds vary over time. Layered queueing networks can be characterized by separate layers (see [8] and [15]) or simultaneous layers. In the first case, customers receive service with some delay. An application where layered networks with separate layers appear is the manufacturing systems, e.g., [3] and [4]. On the other hand, in layered networks with simultaneous layers, customers receive service from the different layers simultaneously. Layered networks with simultaneous layers have applications in communications networks. An application example where layered networks with simultaneous layers (such as our model) appear naturally are web-based multi-tiered system architectures. In such environments, different applications compete for access to shared infrastructure resources, both at the software level (e.g., mutex and database locks, thread-pools) and at the hardware level (e.g., bandwidth, processing power, disk access). For background, see [9] and [10].

The paper is organized as follows. We provide a detailed model description in Section 2 and introduce the systems dynamics. In Section 3, we derive the fluid model and analyze it under the assumption of a single bottleneck and heavy traffic in the network. As we see, the assumption of the single bottleneck allows us to prove a state-space collapse (SSC) property. Then, we show that a fluid model solution converges to equilibrium uniformly on compact sets. The main result of this paper is contained in Section 4. Namely, we provide a diffusion limit theorem for the joint customer population process for this two-layered queueing network. First, we prove that the diffusion-scaled total workload process converges in distribution to a reflected Brownian motion. This result together with results in [1] lead to the main theorem.

2. MODEL

We assume a network with two layers. In layer 1, there are two single-server nodes indexed by i . Customers arrive at node $i \in \{1, 2\}$ randomly one by one and have a random service requirement. A customer completing service at node i may be routed at node l , $l \in \{1, 2\}$, for another service. It is assumed that customers are served according to their arrival order at each node; i.e., first in, first out. Only the first customer at each node can receive service at any time; i.e., the network is a Head of the Line network (HL).

In layer 2, there is a single server working at speed one. The servers of layer 1 are served by this single server simultaneously and at a rate which depends on

the number of customers in the system. The model is illustrated in Figure 1. In Section 2.1, we give a formal description of the model, and in Section 2.2, we introduce the dynamics describing the model. In the sequel, we use the subscript i to refer to processes or quantities pertaining to each node, and, by convention, we omit the subscript to denote the two-dimensional vector of these processes or quantities.

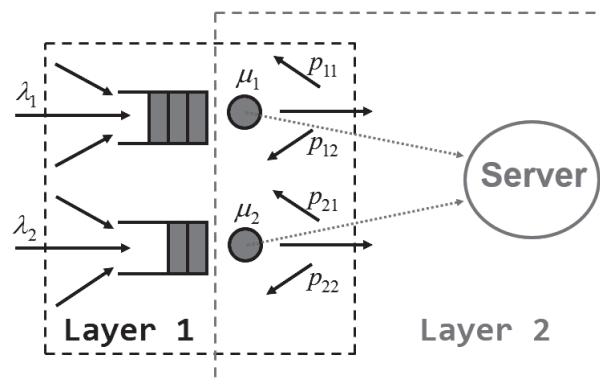


FIGURE 1. A two-layered network with two single-server nodes and routing.

2.1. Preliminaries and model description. In this section, we give a formal model description. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For $T > 0$, let $\mathcal{D}[0, T]^2$ be the Skorokhod space; i.e., the space of two-dimensional real-valued functions on $[0, T]$ that are right continuous with left limits endowed with the J_1 -topology (as all candidate limit objects we consider are continuous, we actually only need to work with the uniform topology); cf. [2]. Denote by $\mathcal{B}(\mathcal{D}[0, T]^2)$ the Borel σ -algebra of $\mathcal{D}[0, T]^2$. All the processes are defined from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathcal{B}(\mathcal{D}[0, T]^2), \mathcal{D}[0, T]^2)$. For a process $X(t)$, let $\|X(t)\|_T = \sup_{0 \leq t \leq T} |X(t)|$ denote the uniform norm, where $|X(\cdot)| = \max_i |X_i(\cdot)|$. We adopt the convention that all mentioned vectors are two-dimensional columns, and use a^T to denote the transpose of a vector or a matrix a . We use A^{-1} to denote the inverse of a square matrix A , A^k its k -th power, and $\|A\|$ the maximum element of A . Furthermore, I represents the identity matrix, and e and e_0 are the vectors consisting of 1's and 0's, respectively, the dimensions of which are clear from the context. Also, e_i is the vector whose i -th element is 1 and the rest are all 0. Last, for a real number x , its integer part is represented by $[x]$.

We start by describing the first layer. Let $u_i(j)$ for $j = 2, 3, \dots$ be the time between the $(j-1)$ -st and j -th external arrival at node i , and $u_i(1) > 0$ be the residual arrival time of the first customer entering at node i after time 0. We assume that the sequence $\{u_i(j)\}$ for $i = 1, 2$ and $j = 2, 3, \dots$ is a sequence of positive i.i.d. random variables with mean $1/\lambda_i$, $\lambda_i > 0$, and that $u_i(1)$ is independent of

this sequence but sampled from an arbitrary distribution with the same mean. For $i = 1, 2$, define the *cumulative arrival time process* $U_i(\cdot)$ as follows: $U_i(0) := 0$ and $U_i(m) := \sum_{j=1}^m u_i(j)$ for $m \in \mathbb{N}$. The number of the external arrivals at node i until time $t > 0$ is given by *the external arrival process*

$$E_i(t) := \max\{m \geq 0 : U_i(m) \leq t\}.$$

In order to be able to define the total workload process in the system, including future service requirements due to routing, we need to introduce a sequence of random variables for any customer j . For any fixed time $t \geq 0$ and $j > 1$, let $v_{ii}^{(1)}(j)$ be the immediate service requirement of the j -th customer (external or routed) at node i . Also, we define $v_{il}^{(k+1)}(j)$ to be the service requirement of the j -th customer (external or routed) at node i at the k -th future time he visits node l for $i, l \in \{1, 2\}$ and $j, k \in \mathbb{N}$. The sequence $\{v_{il}^{(k)}(j)\}$, indexed by j , is a sequence of i.i.d. random variables for any fixed i, l, k and for $j > 1$, and has mean $\beta_l := 1/\mu_l$, $\mu_l > 0$. The random variable $v_{ii}^{(1)}(1)$ denotes the residual service time for the first customer being served at node i at time 0; it is independent of the sequence $\{v_{il}^{(k)}(j)\}$ but sampled from an arbitrary distribution with the same mean. In addition, we assume that all the above-mentioned random variables have finite second moments (more precisely, we need a Lindeberg-type condition to hold to make sure that the exogenous input processes satisfy a functional central limit theorem; see Section 4 for more details). We define the *cumulative service time process* as $V_i(0) := 0$ and, for $m \in \mathbb{N}$,

$$V_i(m) := \sum_{j=1}^m v_{ii}^{(1)}(j), \quad i = 1, 2,$$

and the counting process

$$(2.1) \quad S_i(t) := \max\{m \geq 0 : V_i(m) \leq t\}.$$

We shall use the random variables $\{v_{il}^{(k+1)}(j)\}$ to count the future workload in the system at time t . For a fixed $t > 0$, $v_{il}^{(k+1)}(j)$ represents the k -th future service requirement of the j -th customer waiting to being served at node i and routed at node l . This event will occur after time t and after the completion of his service at node i .

Customers can move between queues according to Markovian routing. To describe the routing process, we define the following quantities. Let P be the (square) routing matrix of dimension two. It is assumed that it is substochastic with a spectral radius less than one; i.e., its largest eigenvalue is less than one. In other words, the network is open, so the following relations hold:

$$(I - P^T)^{-1} = I + \sum_{k=1}^{\infty} (P^T)^k \quad \text{and} \quad \lim_{k \rightarrow \infty} (P^T)^k = 0.$$

For any customer j at node i (external or routed), we define the random variables $\varphi_{il}^{(k)}(j) = 1$ if the j -th departing customer from node i is routed to node l in k steps. The probability of this event is given by

$$\mathbb{P}(\varphi_{il}^{(k)}(j) = 1) = p_{il}^{(k)},$$

where $p_{il}^{(k)}$ denotes the (i, l) -th element of the matrix P^k . For $i = 1, 2$, we define the two-dimensional random vector

$$\varphi_i^{(k)}(j) := (\varphi_{i1}^{(k)}(j), \varphi_{i2}^{(k)}(j))^T.$$

Note that $\varphi_i^{(k)}(j)$ can take values in the set $\{e_0, e_1, e_2\}$, where $\varphi_i^{(k)}(j) = e_0$ means that the j -th customer leaves the system. Let $\bar{p}_i^{(k)}$ be the i -th column of the matrix $(P^T)^k$. The expectation and the covariance matrix of $\varphi_i^{(k)}(j)$, for $i = 1, 2$, are given by

$$\mathbb{E}(\varphi_i^{(k)}(j)) = \bar{p}_i^{(k)} \quad \text{and} \quad \mathbb{C}(\varphi_i^{(k)}(j)) = \begin{bmatrix} p_{i1}^{(k)}(1 - p_{i2}^{(k)}) & -p_{i1}^{(k)}p_{i2}^{(k)} \\ -p_{i1}^{(k)}p_{i2}^{(k)} & p_{i2}^{(k)}(1 - p_{i1}^{(k)}) \end{bmatrix}.$$

Now, we can define the routing process, which counts the number of customers who are routed from node l to node i , as

$$\Phi_{li}(m) := \sum_{j=1}^m \varphi_{li}^{(1)}(j), \quad i, l = 1, 2, \quad m \in \mathbb{N}.$$

The total arrival rate at node i , γ_i , is given by the solution of the following traffic equations:

$$\gamma_i = \lambda_i + \sum_{l=1}^2 p_{li}^{(1)} \gamma_l, \quad i = 1, 2.$$

In vector form, this can be written as

$$(2.2) \quad \gamma = (I - P^T)^{-1} \lambda.$$

It is shown in [2], Theorem 7.3, that under the assumptions described above, (2.2) has a unique solution $\gamma = (\gamma_1, \gamma_2)^T$. The traffic intensity of node i is $\rho_i := \gamma_i / \mu_i$.

Now, we describe the service discipline at the second layer. Here, there is a single server. The servers of layer 1 become jobs at layer 2 in the sense they are served by the server of layer 2 simultaneously and at a rate that depends on the number of customers in layer 1 (at any time). The rate that each node receives is given by the service allocation function $R(\cdot) : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+^2$, with $R(\cdot) := (R_1(\cdot), R_2(\cdot))^T$ and, for $i = 1, 2$,

$$(2.3) \quad R_i(q) := \begin{cases} \min\{q_i, K_i\} / \sum_{j=1}^2 \min\{q_j, K_j\} & \text{if } q_i \neq 0, \\ 0 & \text{if } q_i = 0. \end{cases}$$

The quantity q_i represents the number of customers at node i . The two-dimensional vector $K = (K_1, K_2)^T$ is constant, and we call it *the degree of resource sharing*. We assume that it is always finite and the user can choose it as a parameter of the system. Observe that K in this service allocation function guarantees a minimum service rate for each customer in the system. Also, note that the above function is Lipschitz continuous for $q \neq (0, 0)$.

We make the additional assumption that there exists a unique bottleneck in our system, which without loss of generality we let it be node 1. The definition of bottleneck is the following.

DEFINITION 2.1 (Bottleneck). Node i is a *bottleneck* if $i = \arg \min_j \mu_j K_j / \gamma_j$ for $i = 1, 2$.

By the previous definition, a straightforward inequality follows:

$$(2.4) \quad \frac{\rho_1}{K_1} > \frac{\rho_2}{K_2}.$$

Observe that, if $K_1 = K_2$, an intuitive explanation of the above definition is that the average occupancy of the server at node 1 is strictly greater than that of the server at node 2. In case of multiserver nodes, where K_i represents the number of servers at node i , the fraction ρ_i / K_i is the average occupancy of a server at node i .

2.2. System dynamics. In this section, we introduce the dynamics that describe our model. We denote by $Q_i(t)$ the number of customers at node i at time t . This is given by

$$(2.5) \quad Q_i(t) = Q_i(0) + E_i(t) + \sum_{l=1}^2 \Phi_{li} \left(S_l(T_l(t)) \right) - S_i(T_i(t)),$$

where $Q_i(0)$ denotes the number of customers initially at node i . We define the *cumulative service time* of the server at node i as

$$(2.6) \quad T_i(t) = \int_0^t R_i(Q(s)) ds.$$

This quantity can be viewed as the effort that the server of node i has put in processing customers during $[0, t]$. Note that as the allocation function might be less than one, the above process is not necessarily equal to the amount of time that the server at node i is busy during $[0, t]$. In case the other node is empty during $[0, t]$, (2.6) coincides with the busy time at node i . Recall that $E_i(t)$ is the number of external arrivals at node i up to time t . Observe that $S_i(T_i(t))$, which is a composition of the renewal process (2.1) and the process $T_i(t)$, represents the number of departures at node i until time t . Furthermore, the *total arrival process* is given by

$$(2.7) \quad A_i(t) = E_i(t) + \sum_{l=1}^2 \Phi_{li} \left(S_l(T_l(t)) \right).$$

The amount of time that both servers at the nodes are idle during $[0, t]$ is given by the one-dimensional process

$$(2.8) \quad Y_{L_2}(t) = t - \sum_{i=1}^2 T_i(t).$$

Alternatively, we can see this quantity as the idle time of the server in layer 2 during $[0, t]$. Further, the *immediate workload* at node i at time t is defined as

$$(2.9) \quad W_i(t) = V_i(Q_i(0) + A_i(t)) - T_i(t).$$

Observe that $W_i(t)$ is nonnegative for any $t \geq 0$. Last, due to the work-conserving property in layer 2, the following relation holds:

$$(2.10) \quad Y_{L_2}(t) \text{ increases} \Rightarrow W_1(t) + W_2(t) = 0, \quad t \geq 0, \quad i = 1, 2.$$

Recall that when we omit the subscript i , we refer to the two-dimensional column vector of the corresponding process/quantity; for example, $A(\cdot) = (A_1(\cdot), A_2(\cdot))^T$ and $W(\cdot) = (W_1(\cdot), W_2(\cdot))^T$. All the essential information of the evolution of the system is contained in the following six-tuple:

$$\mathfrak{X}(\cdot) := (A(\cdot), S(\cdot), Q(\cdot), T(\cdot), Y_{L_2}(\cdot), W(\cdot)).$$

In addition, the total (immediate and future) workload of the system plays a key role in our analysis. First, we define the remaining service requirement of the j -th customer waiting to be served at node $i = 1, 2$ as

$$(2.11) \quad s_i(j) := v_{ii}^{(1)}(j) + s'_i(j)$$

where

$$s'_i(j) := \sum_{l=1}^2 \sum_{k=1}^{\infty} \varphi_{il}^{(k)}(j) v_{il}^{(k+1)}(j)$$

is the future service requirement of the above-mentioned customer. Observe that, for an external arrival, $s_i(j)$ is the total service requirement. The first and the second moments of (2.11) are given (in the vector form) by

$$(2.12) \quad \tau := \mathbb{E}(s(j)) = (I - P)^{-1} \beta$$

and

$$(2.13) \quad \tau^{(2)} := \mathbb{E}(s^2(j)) = (I - P)^{-1} \left(\mathbb{E}(v^2(j)) + 2\beta(P\tau) \right).$$

Now, we can define the (one-dimensional) total workload of the system as

$$(2.14) \quad W_{\text{Tot}}(t) := \sum_{i=1}^2 W_i(t) + \sum_{i=1}^2 \sum_{j=S_i(T_i(t))+1}^{Q_i(0)+A_i(t)} s'_i(j).$$

In case that $S_i(T_i(t)) = Q_i(0) + A_i(t)$, i.e., there are no customers at node i , we understand the second sum of the last equation as zero. Obviously, the total workload is not a Markov process as it is dependent on future service requirements. In Section 4.1, we shall see that under an appropriate scaling (i.e., the diffusion scaling) the dependence of the total workload on the future vanishes.

Last, as our network is HL, only one customer can be in service at node i at any time. This property gives an upper and a lower bound for the cumulative service time (2.6) at node i , which is given in [1], Inequality 2.13, namely,

$$(2.15) \quad V_i\left(S_i(T_i(t))\right) \leq T_i(t) < V_i\left(S_i(T_i(t)) + 1\right).$$

We have so far defined the system dynamics for the above-mentioned two-layered network and stated all the assumptions we need for our analysis. We are now ready to study the fluid model of this network, which is the first essential step to show an SSC property.

3. FLUID ANALYSIS

In this section, we study a critical fluid model, which is a deterministic model and can be thought of as a formal law of large numbers approximation under appropriate scaling. We shall give a rigorous proof of the last statement in the next section.

The main goal is to prove uniform convergence (with respect to the initial condition) on compact sets for the fluid model under the critical loading assumption; i.e., the traffic intensity of the network is one. First, we find the invariant points (or equilibrium states) and define an appropriate lifting map which describes these points. Then, we define a time-changed version of the original fluid model and we show that it is enough to prove the convergence for the time-changed function. As the time-changed function is given by a piecewise linear ODE, we are able to find the solution and to show the convergence. Because the degree of resource sharing (the vector K) is finite, we need to separate the state space in suitable regions and to distinguish cases depending on initial conditions.

3.1. Definition and invariant points. The traffic intensity of the network is given by $\rho := \beta^T \gamma = \sum_{i=1}^2 \rho_i$. We make the critical loading assumption, i.e.,

$$(3.1) \quad \rho = \rho_1 + \rho_2 = 1.$$

To derive the fluid model equations, we replace any random quantity in (2.5)–(2.10) with its mean. The fluid model equations are given by

$$(3.2) \quad \bar{Q}_i(t) = \bar{Q}_i(0) + \lambda_i t + \sum_{l=1}^2 p_{li} \mu_l \bar{T}_l(t) - \mu_i \bar{T}_i(t),$$

$$(3.3) \quad \bar{T}_i(t) = \int_0^t R_i(\bar{Q}(s)) ds,$$

$$(3.4) \quad \sum_{i=1}^2 \bar{T}_i(t) + \bar{Y}_{L_2}(t) = t,$$

$$(3.5) \quad \bar{W}_i(t) = \beta_i(\lambda_i t + \sum_{l=1}^2 p_{li} \mu_l \bar{T}_l(t) + \bar{Q}_i(0)) - \bar{T}_i(t),$$

$$(3.6) \quad \bar{Y}_{L_2}(t) \text{ increases} \Rightarrow \bar{W}_1(t) + \bar{W}_2(t) = 0, \quad t \geq 0, \quad i = 1, 2.$$

We can show that the immediate workload in the fluid model can be written as $\bar{W}_i(t) = \beta_i \bar{Q}_i(t)$.

DEFINITION 3.1 (Fluid model). We say that a two-dimensional vector $\bar{Q}(\cdot)$ with nonnegative components is a solution of the *fluid model* if it is continuous and satisfies the equations (3.2)–(3.6) for $t \in [0, \delta]$, and $Q(t) = 0$ for $t \geq \delta$, with $\delta = \inf\{t : Q(t) = 0\}$.

We define an auxiliary quantity which can be interpreted as the total workload in the fluid model. It is defined by the function $\bar{Q}(\cdot)$ as follows:

$$\begin{aligned} \bar{W}_{\text{Tot}}(t) &= \beta^T \bar{Q}(t) + \sum_{k=1}^{\infty} \sum_{i=1}^2 \beta^T \bar{p}_i^{(k)} \bar{Q}_i(t) = \beta^T \bar{Q}(t) + \sum_{k=1}^{\infty} \beta^T (P^T)^k \bar{Q}(t) \\ &= \beta^T (I - P^T)^{-1} \bar{Q}(t) = \tau^T \bar{Q}(t). \end{aligned}$$

A useful result in our analysis is that the fluid total workload in the system remains constant under the critical loading assumption.

PROPOSITION 3.1. *For any fluid model solution $\bar{Q}(\cdot)$, we have*

$$\bar{W}_{\text{Tot}}(t) = \beta^T (I - P^T)^{-1} \bar{Q}(t) = \bar{W}_{\text{Tot}}(0).$$

Proof. If $\bar{Q}(0) = 0$, then $\bar{W}_{\text{Tot}}(t) \equiv 0$ for $t \geq 0$. Let $\bar{Q}(0) > 0$. Assume now $\bar{Q}(0) = 0$. By definition, $\bar{Q}(t)$ is continuous, so let t be such that $\bar{Q}(t)$ is positive in a neighborhood of t . Calculating the derivative of the total workload at time t , we derive

$$\begin{aligned} \bar{W}'_{\text{Tot}}(t) &= \beta^T (I - P^T)^{-1} \bar{Q}'(t) = \beta^T (I - P^T)^{-1} (\lambda - (I - P^T)(\mu \circ R(\bar{Q}(t)))) \\ &= \beta^T \gamma - \beta^T (\mu \circ R(\bar{Q}(t))) = \beta^T \gamma - \sum_{i=1}^2 R_i(\bar{Q}(t)) = 0. \end{aligned}$$

The last equation holds due to (3.1) and the property of the service allocation function; i.e. $\sum_{i=1}^2 R_i(\bar{Q}(t)) = 1$. It follows that $\bar{W}_{\text{Tot}}(t) = \bar{W}_{\text{Tot}}(0)$ for $t \geq 0$. Thus, $\bar{W}_{\text{Tot}}(t)$ is constant on $[0, \delta]$. Combining this with the continuity of $\bar{Q}(t)$, we see that, necessarily, $\delta = \infty$. Thus, the result extends to all positive t . ■

In the following lemma, we show that there exists a solution of the fluid model equations for all non-zero initial states and it is unique.

LEMMA 3.1 (Existence and uniqueness). *For any $\bar{Q}(0) \in \mathbb{R}_+^2 \setminus \{0\}$ there exists a unique solution of the fluid model equations.*

PROOF. Let $\bar{Q}(0) > 0$. A by-product of the previous lemma is $\delta = \infty$. Thus, we can discard the origin and define the function $\Psi(\cdot) : \mathbb{R}_+^2 \setminus \{0\} \rightarrow \mathbb{R}_+^2$ as

$$(3.7) \quad \Psi(\cdot) := \lambda - (I - P^T)(\mu \circ R(\cdot)),$$

where $\mu \circ R(\cdot)$ indicates the Hadamard product; $\mu \circ R(\cdot) = (\mu_1 R_1(\cdot), \mu_2 R_2(\cdot))^T$. This function is Lipschitz continuous because such is $R(\cdot)$. Now, note that (3.2) can be written as

$$(3.8) \quad \bar{Q}'(t) = \Psi(\bar{Q}(t)), \quad t \geq 0,$$

where the prime denotes the derivative with respect to time. The existence and uniqueness follow directly by [12], Section 10, Theorem IV. ■

Now, we characterize the invariant points $x \in \mathbb{R}_+^2$ of the fluid model. Before we state our result, we first proceed in an informal manner. Equate the total rate into node i with the total rate out of node i . That is,

$$(3.9) \quad \gamma_i = \mu_i R_i(x).$$

Thus, for the points on the invariant manifold (i.e. the set of the invariant points), we have $\rho_i = R_i(x)$. Using the definition of a bottleneck and keeping in mind that we assume node 1 to be the bottleneck, we now describe the invariant points. We know by (2.4) that $\rho_1/K_1 > \rho_2/K_2$, which yields

$$\frac{K_1}{K_2} < \frac{\rho_1}{\rho_2} = \frac{R_1(x)}{R_2(x)}.$$

Thus, by combining the last inequality and the definition of the service allocation function (2.3), we see that the following inequality holds:

$$\min\{x_2, K_2\} < \frac{K_2}{K_1} \min\{x_1, K_1\} \leq K_2.$$

The last inequality implies that for all invariant points $x = (x_1, x_2)$ of the fluid model, we have $x_2 < K_2$. Thus, solving the equation (3.9) for x_2 now yields $x_2 = \frac{\rho_2}{\rho_1} \min\{x_1, K_1\}$. The invariant manifold is thus given by

$$(3.10) \quad \mathcal{I} = \left\{ x \in \mathbb{R}_+^2 : x_2 = \frac{\mu_1 \gamma_2}{\gamma_1 \mu_2} \min\{x_1, K_1\} \right\}.$$

We make the previous arguments rigorous by showing that a sufficient and necessary condition of the fluid queue length to remain constant in time is the initial state lies on the invariant manifold.

PROPOSITION 3.2. *Let $\bar{Q}(t)$ be a solution of (3.8). Then, $\bar{Q}(t) = \bar{Q}(0)$ for all $t \geq 0$ if and only if $\bar{Q}(0) \in \mathcal{I}$.*

PROOF. The definition of the invariant manifold of an ODE is the set of all initial states such that the function remains constant; i.e., $\{\bar{Q}(0) \in \mathbb{R}_+^2 : \bar{Q}(t) = \bar{Q}(0), t \geq 0\}$. Suppose now that $\bar{Q}(0) \in \mathcal{I}$. In this case, by the definition of an invariant point, $\bar{Q}(t)$ should be constant. So, $\bar{Q}(t) = \bar{Q}(0) \in \mathcal{I}$.

Now, supposing that $\bar{Q}(t) = \bar{Q}(0) \in \mathbb{R}_+^2$ for all $t \geq 0$, it follows from the previous discussion that $\bar{Q}(0) \in \mathcal{I}$. ■

Having found the invariant (or equilibrium) points of the fluid model, we now turn to its stability property, namely, the convergence of the solutions of fluid model equations to the invariant manifold as time goes to infinity.

3.2. Convergence to the invariant manifold for the fluid model. Let x^* be the critical point in the invariant manifold where $x_1^* = K_1$, which means that $x_2^* = \frac{\mu_1 \gamma_2}{\gamma_1 \mu_2} K_1$. For this point, we define the critical workload as (cf. (3.1))

$$(3.11) \quad w^* := \beta^T (I - P^T)^{-1} x^* = \tau^T x^*.$$

In order to prove an SSC property based on the critical workload level w^* , we define a lifting map, $\Delta : \mathbb{R}_+ \rightarrow \mathbb{R}_+^2$, as follows:

$$(3.12) \quad \begin{aligned} \Delta_1(w) &:= \frac{\min\{w, w^*\}}{w^*} K_1 + \frac{\max\{w - w^*, 0\}}{\tau_1}, \\ \Delta_2(w) &:= \frac{\min\{w, w^*\}}{w^*} \frac{\rho_2 K_1}{\rho_1}. \end{aligned}$$

Note that the lifting map is Lipschitz continuous with constant

$$C_1 = \max \left\{ 2 \frac{K_1}{w^*} + \mu_1, 2C_2 \right\}, \quad \text{where } C_2 = \frac{\mu_1 K_1}{\lambda_1 w^*} \max_i \frac{\lambda_i}{\mu_i}.$$

In the sequel, we show that the fluid model solution converges to the invariant manifold as t goes to infinity.

THEOREM 3.1 (Convergence to the invariant manifold for the fluid model). *If $\bar{Q}(0) = (\bar{Q}_1(0), \bar{Q}_2(0)) \in [0, M]^2$ for some $M > 0$, then for any $\epsilon > 0$, there exists a $t_0 \geq 0$ (independent of M) such that*

$$(3.13) \quad \sup_{\bar{Q}(0) \in [0, M]^2} |\bar{Q}(t) - \Delta \bar{W}_{\text{Tot}}(0)| \leq \epsilon$$

for $t > t_0$, and $\Delta \bar{W}_{\text{Tot}}(0)$ is an invariant state.

P r o o f. Here, we give a sketch of the proof. The complete proof is extended in the rest of this section. The first step is to define a function $y(\cdot)$ and a function $G(\cdot)$, and to show that $y(\cdot)$ can be interpreted as a time-change of $\bar{Q}(t)$, namely $\bar{Q}(t) = y(G(t))$. Then, we show that the convergence of the time-changed version implies the convergence of the original function $\bar{Q}(\cdot)$. To this end, let $\Xi = \{\xi_{ij}\}$, $i, j \in \{1, 2\}$, be the matrix

$$\Xi = \begin{bmatrix} \lambda_1 + \mu_1 p_{11} - \mu_1 & \lambda_1 + \mu_2 p_{21} \\ \lambda_2 + \mu_1 p_{12} & \lambda_2 + \mu_2 p_{22} - \mu_2 \end{bmatrix}.$$

Define a function $y(\cdot) : [0, \infty) \rightarrow [0, \infty)^2$ such that $y_i(0) = \bar{Q}_i(0)$ and

$$(3.14) \quad y'(t) = \Xi(\min\{y_1(t), K_1\}, \min\{y_2(t), K_2\})^T.$$

We shall show that the above-defined function can be interpreted as a time change of $\bar{Q}(t)$. Let $G(\cdot) : [0, \infty) \rightarrow [0, \infty)$ be the solution of the equation

$$G'(t) = \frac{1}{\sum_{i=1}^2 \min\{y_i(G(t)), K_i\}}.$$

Note that $G(\cdot)$ is continuous and that

$$G(t) = \int_0^t \frac{1}{\sum_{i=1}^2 \min\{y_i(G(s)), K_i\}} ds \geq \frac{1}{\sum_{i=1}^2 K_i} t.$$

This means that the function $G(\cdot)$ is strictly increasing and unbounded in time, which implies that $G(\cdot)$ is also invertible. The original function $\bar{Q}(t)$ can be interpreted as $\bar{Q}_i(t) = y_i(G(t))$ for $t > 0$. To see this,

$$\begin{aligned} \bar{Q}'(t) &= y'(G(t))G'(t) = \frac{\Xi(\min\{y_1(G(t)), K_1\}, \min\{y_2(G(t)), K_2\})^T}{\sum_{i=1}^2 \min\{y_i(G(t)), K_i\}} \\ &= \Xi(R_1(\bar{Q}(t)), R_2(\bar{Q}(t)))^T = \Psi(\bar{Q}(t)), \end{aligned}$$

where the function $\Psi(\cdot)$ is defined in (3.7).

The idea now is to prove (3.13) by showing that $y(t)$ converges as $t \rightarrow \infty$. We will formally do so in the remainder of Section 3. Assuming that $y(t)$ converges, we now show that $\bar{Q}(t)$ converges. To see this, if for any $\epsilon > 0$ there exists t_0 (independent of M) such that

$$(3.15) \quad \sup_{y(0) \in [0, M]^2} |y(t) - \Delta \bar{W}_{\text{Tot}}(0)| \leq \epsilon, \quad t > t_0,$$

then

$$\begin{aligned} \sup_{\bar{Q}(0) \in [0, M]^2} |\bar{Q}(t) - \Delta \bar{W}_{\text{Tot}}(0)| &= \sup_{y(0) \in [0, M]^2} |y(G(t)) - \Delta \bar{W}_{\text{Tot}}(0)| \\ &= \sup_{y(0) \in [0, M]^2} |y(u) - \Delta \bar{W}_{\text{Tot}}(0)|, \end{aligned}$$

where $u = G(t)$. For $u > G^{-1}(t_0)$, the last term becomes smaller than ϵ . ■

The remainder of the current section is devoted to the proof of (3.15). To do it, we first define appropriate spatial regions in which the fluid model solutions have qualitatively different behavior and we solve (3.14) in these regions. This is done in Section 3.3. In Section 3.4, these solutions are used to show that both $y_1(t)$ and $y_2(t)$ are monotone in t . A crucial observation is that we only need to look at $y_2(t)$ as $\tau_1 y_1(t) + \tau_2 y_2(t) = \bar{W}_{\text{Tot}}(0)$, by Proposition 3.1. This paves the way for a global convergence analysis of $y(t)$, also establishing the desired uniformity. This is done in Section 3.5.

3.3. Explicit local solutions of time-changed ODE. Because we assume that the degree of resource sharing K is finite, the form of (3.14) depends on the value of $y(t)$. For this reason, we need to define the following regions. For $x = (x_1, x_2) \in \mathbb{R}_+^2$, we define

$$\begin{aligned} \Pi_1 &= \{x \in \mathbb{R}_+^2 : x_1 \leq K_1, x_2 \geq K_2\}, \quad \Pi_2 = \{x \in \mathbb{R}_+^2 : x_1 \geq K_1, x_2 \geq K_2\}, \\ \Pi_3 &= \{x \in \mathbb{R}_+^2 : x_1 \geq K_1, x_2 \leq K_2\}, \quad \Pi_4 = \{x \in \mathbb{R}_+^2 : x_1 \leq K_1, x_2 \leq K_2\}. \end{aligned}$$

The following picture (Figure 2) shows these regions and the invariant manifold as defined in (3.10).

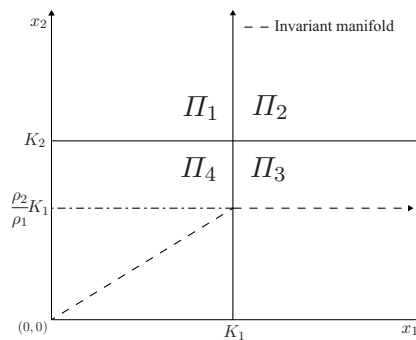


FIGURE 2. The regions Π_i and the invariant manifold.

Let $y(0) = (y_1(0), y_2(0))^T \in \mathbb{R}_+^2$. We solve the time-changed ODE, which is given by (3.14), in regions Π_3 and Π_4 (considering these two regions only is sufficient for our purposes). It is useful to observe the relations between the coefficients

of the matrix Ξ in (3.14), which will be used later. We know by the definition of the total arrival rate that

$$(3.16) \quad \rho_2(1 - p_{22}) = \frac{\lambda_2}{\mu_2} + \frac{p_{12}\gamma_1}{\mu_2}.$$

The constant ξ_{22} can be expressed as

$$(3.17) \quad \xi_{22} = \lambda_2 + \mu_2(p_{22} - 1) \stackrel{(3.16)}{=} \lambda_2 - \frac{1}{\rho_2}(\lambda_2 + p_{12}\gamma_1) = -\frac{\rho_1}{\rho_2}\xi_{21} < 0.$$

In a similar way, we can obtain

$$(3.18) \quad \xi_{11} = -\frac{\rho_2}{\rho_1}\xi_{12} < 0.$$

By definition (2.12), τ_2 can be written as $\tau_2(1 - p_{22}) = \beta_2 + p_{21}\tau_1$. Also, by (2.2) and (2.12), we have $\tau^T\lambda = \beta^T(I - P^T)^{-1}\lambda = \beta^T\gamma = 1$. Combining the previous two equations, we get

$$\xi_{12} = -\frac{\tau_2}{\tau_1}\xi_{22}.$$

Last, by (3.17) and (3.18) we have

$$\xi_{21} = -\frac{\tau_1}{\tau_2}\xi_{11}.$$

3.3.1. Solution in region Π_3 . Assuming that $y(s)$ is in region Π_3 for $s \in [0, t]$, we can directly solve the second equation of the system (3.14) since it is independent of $y_1(t)$:

$$y_2'(t) = \xi_{21}K_1 + \xi_{22}y_2(t).$$

Then, by using (3.17), the solution is given by

$$y_2(t) = \left(y_2(0) - \frac{\rho_2}{\rho_1}K_1 \right) \exp\{\xi_{22}t\} + \frac{\rho_2}{\rho_1}K_1.$$

Now, we can easily obtain the solution of the first equation of the system (3.14). By the relations between the coefficients of (3.14), this solution is given by

$$(3.19) \quad y_1(t) = -\frac{\tau_2}{\tau_1} \left(y_2(0) - \frac{\rho_2}{\rho_1}K_1 \right) \exp\{\xi_{22}t\} + \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1.$$

3.3.2. Solution in region Π_4 . Assuming that $y(s)$ is in region Π_4 for $s \in [0, t]$, we can write the system given by (3.14) as

$$(3.20) \quad y'(t) = \Xi y(t)^T.$$

The eigenvalues of Ξ are $\alpha_1 = 0$ and $\alpha_2 = \xi_{11} + \xi_{22} = \lambda_1 + \mu_1(p_{11} - 1) + \lambda_2 + \mu_2(p_{22} - 1) < 0$. Let V_1, V_2 be the corresponding eigenvectors; i.e.,

$$V_1 = \begin{pmatrix} 1 \\ -\frac{\xi_{11}}{\xi_{12}} \end{pmatrix} \quad \text{and} \quad V_2 = \begin{pmatrix} 1 \\ \frac{\alpha_2 - \xi_{11}}{\xi_{12}} \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{\xi_{21}}{\xi_{11}} \end{pmatrix}.$$

Using the relations between the coefficients of the matrix Ξ , we obtain the solution of (3.20) in the form

$$(3.21) \quad \begin{aligned} y_1(t) &= c_2 + c'_2 \exp\{\alpha_2 t\}, \\ y_2(t) &= c_2 \frac{\rho_2}{\rho_1} - c'_2 \frac{\tau_1}{\tau_2} \exp\{\alpha_2 t\}, \end{aligned}$$

with

$$c_2 = \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 \quad \text{and} \quad c'_2 = - \left[y_2(0) - \frac{\rho_2}{\rho_1} y_1(0) \right] \left[\frac{\tau_2 \rho_1}{\tau_1 \rho_1 + \tau_2 \rho_2} \right].$$

Having found the solution of (3.14) in each of the regions Π_3 and Π_4 , we observe that the two-dimensional equation can be reduced to a one-dimensional equation since $\bar{W}_{\text{Tot}}(0) = \tau_1 y_1(t) + \tau_2 y_2(t)$ for $t \geq 0$. Now, it is enough to show the convergence of this equation. To see this, define $x(\cdot)$ as

$$(3.22) \quad x(t) := y_2(t) = \frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 y_1(t)}{\tau_2},$$

and its derivative as

$$(3.23) \quad x'(t) = - \frac{\tau_1 y_1'(t)}{\tau_2}.$$

Using the observation that we can reduce the dimension by one and the solutions to system (3.14), we show that the fluid model solutions converge to an equilibrium state uniformly for all initial states within a compact set. First, we find the sign of the derivative of the above reduced equation. Then, as the limiting point depends on the sign of the quantity $\bar{W}_{\text{Tot}}(0) - w^*$, we have to distinguish between the following three cases: the total workload in the fluid model is (i) greater than, (ii) less than or (iii) equal to the critical workload.

3.4. Local analysis: Establishing monotonicity. By (3.22), it is clear that we need to study only the behavior of $y_1(t)$. In the sequel, we find the sign of (3.23) in each region Π_i for $i = 1, \dots, 4$.

In region Π_1 , we know that $y_1(t) \leq K_1$. By (3.14) and (3.18), we have

$$\begin{aligned} x'(t) &= - \frac{\tau_1 y_1'(t)}{\tau_2} = - \frac{\tau_1}{\tau_2} (\xi_{11} y_1(t) + \xi_{12} K_2) = - \frac{\tau_1}{\tau_2} \xi_{12} \left(- \frac{\rho_2}{\rho_1} y_1(t) + K_2 \right) \\ &\leq - \frac{\tau_1}{\tau_2} \xi_{12} \left(- \frac{\rho_2}{\rho_1} K_1 + K_2 \right) = -\epsilon_1 < 0, \end{aligned}$$

where $\epsilon_1 = \frac{\tau_1}{\tau_2} \xi_{12} \left(-\frac{\rho_1}{\rho_2} K_1 + K_2 \right)$, which is strictly positive by (2.4) and the fact that $\xi_{12} > 0$. Therefore, we conclude that the derivative of $x(t)$ is strictly negative in region Π_1 .

In region Π_2 , by (3.14) and (3.18), we have

$$x'(t) = -\frac{\tau_1}{\tau_2} (\xi_{11} K_1 + \xi_{12} K_2) = -\frac{\tau_1 \xi_{12}}{\tau_2} \left(-\frac{\rho_2}{\rho_1} K_1 + K_2 \right) = -\epsilon_1 < 0.$$

Thus, the derivative of $x(t)$ is strictly negative for all $y_1(0)$ in regions Π_1 and Π_2 . In other words, the trajectory of $x(t)$ leaves regions Π_1 and Π_2 after a finite time. Now, we move to the regions where the invariant points lie, i.e., Π_3 and Π_4 .

In region Π_3 , by (3.11) and (3.19), we obtain

$$\begin{aligned} x'(t) &= \frac{\tau_1}{\tau_2} \xi_{22} \left(\frac{\bar{W}_{\text{Tot}}(0)}{\tau_1} - y_1(0) - \frac{\tau_2 \rho_2}{\tau_1 \rho_1} K_1 \right) \exp\{\xi_{22} t\} \\ &= \frac{\tau_1}{\tau_2} \xi_{22} \left(\frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 - y_1(0) \right) \exp\{\xi_{22} t\}. \end{aligned}$$

We saw in (3.17) that $\xi_{22} < 0$. We therefore infer that, in Π_3 ,

$$(3.24) \quad x'(t) = \begin{cases} > 0 & \text{if } y_1(0) > \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1, \\ < 0 & \text{if } y_1(0) < \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1, \end{cases}$$

and

$$x'(t) = 0 \quad \text{if } y_1(0) = \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1.$$

In region Π_4 , by (3.11) and (3.21), we have

$$x'(t) = -\frac{\tau_1 y_1'(t)}{\tau_2} = -\frac{\tau_1}{\tau_2} c_4' \alpha_2 \exp\{\alpha_2 t\},$$

where

$$\begin{aligned} c_4' &= -\left[\frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1}{\tau_2} y_1(0) - \frac{\rho_2}{\rho_1} y_1(0) \right] \left[\frac{\tau_2 \rho_1}{\tau_1 \rho_1 + \tau_2 \rho_2} \right] \\ &= -\left[\frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 \rho_1 + \tau_2 \rho_2}{\tau_2 \rho_1} y_1(0) \right] \left[\frac{\tau_2 \rho_1}{\tau_1 \rho_1 + \tau_2 \rho_2} \right] \\ &= -\left[\bar{W}_{\text{Tot}}(0) - \frac{w^*}{K_1} y_1(0) \right] \left[\frac{\rho_1}{\tau_1 \rho_1 + \tau_2 \rho_2} \right]. \end{aligned}$$

Recall that from Section 3.3.2 it follows that $\alpha_2 < 0$, and thus, in Π_4 ,

$$(3.25) \quad x'(t) = \begin{cases} > 0 & \text{if } y_1(0) > \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1, \\ < 0 & \text{if } y_1(0) < \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1, \end{cases}$$

and

$$x'(t) = 0 \quad \text{if } y_1(0) = \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1.$$

Combining (3.24), (3.25), and keeping in mind that $\Delta_1(w) = \frac{\min\{w, w^*\}}{w^*} K_1 + \frac{\max\{w - w^*, 0\}}{\tau_1}$, we see that, in $\Pi_3 \cup \Pi_4$,

$$x'(t) = \begin{cases} > 0 & \text{if } y_1(0) > \Delta_1(\bar{W}_{\text{Tot}}(0)), \\ < 0 & \text{if } y_1(0) < \Delta_1(\bar{W}_{\text{Tot}}(0)), \end{cases}$$

and $x'(t) = 0$ if $y_1(0) = \Delta_1(\bar{W}_{\text{Tot}}(0))$.

3.5. Global analysis: Convergence to invariant manifold. We are now ready to connect all pieces. From the previous section we know that $x(t) = y_2(t)$ must be smaller than K_2 after a finite time, thus exiting regions Π_1 and Π_2 . Therefore, we can focus on the remaining two regions. In order to do so, we need to consider whether $y_1(t)$ will eventually be larger than, smaller than, or equal to K_1 . This leads to three cases, treated separately in the remainder of this section.

Case 1: $\bar{W}_{\text{Tot}}(0) - w^* > 0$. In this case the invariant point (limiting point) lies in region Π_3 . If $y_1(0) \in \Pi_3$, then we know that $y_1(0) \geq K_1$. By (3.19), we have

$$\begin{aligned} y_1(t) &= -\frac{\tau_2}{\tau_1} \left(\frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 y_1(0)}{\tau_2} - \frac{\rho_2}{\rho_1} K_1 \right) \exp\{\xi_{22}t\} + \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 \\ &\geq \left(-\frac{\bar{W}_{\text{Tot}}(0)}{\tau_1} + K_1 + \frac{\tau_2 \rho_2}{\tau_1 \rho_1} K_1 \right) \exp\{\xi_{22}t\} + \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 \\ &\geq \left(-\frac{\bar{W}_{\text{Tot}}(0)}{\tau_1} + K_1 + \frac{w^*}{\tau_1} - K_1 \right) \exp\{\xi_{22}t\} + \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 \\ &\geq \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} (1 - \exp\{\xi_{22}t\}) + K_1 \geq K_1. \end{aligned}$$

If $y_1(0) \in \Pi_4$, then $y_1(0) \leq K_1$. Also, by the assumption that $\bar{W}_{\text{Tot}}(0) - w^* > 0$ and the definition of the lifting map (3.12), we infer that $y_1(0) \leq K_1 < \Delta_1 \bar{W}_{\text{Tot}}(0)$. This implies that $x(0) \geq \bar{W}_{\text{Tot}}(0)/\tau_2 - \tau_1 K_1/\tau_2$ and $x(t)$ is strictly decreasing. In the sequel, we show that there exists a time t^* such that $y_1(t^*) = K_1$. This means that the function $x(t)$ lies in region Π_3 after that time. It is enough to prove that the equation $y_1(t) = K_1$ has a positive solution. Note that, by (3.21), we have

$$y_1(t) = \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 - \left[\bar{W}_{\text{Tot}}(0) - \frac{w^*}{K_1} y_1(0) \right] \left[\frac{\rho_1}{\tau_1 \rho_1 + \tau_2 \rho_2} \right] \exp\{\alpha_2 t\}.$$

Now, setting $y_1(t) = K_1$, we see that the previous equation becomes

$$(3.26) \quad \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 - K_1 = \left[\bar{W}_{\text{Tot}}(0) - \frac{w^*}{K_1} y_1(0) \right] \left[\frac{\rho_1}{\tau_1 \rho_1 + \tau_2 \rho_2} \right] \exp\{\alpha_2 t\}.$$

We argue that (3.26) is satisfied by a $t \geq 0$ as follows. Observe that by (3.11) the quantity $\rho_1/(\tau_1\rho_1 + \tau_2\rho_2)$ is equal to K_1/w^* , and by the assumption that $y_1(0) \in \Pi_4$, we have $y_1(0) \leq K_1$. Now, we can obtain the inequality

$$\left[\bar{W}_{\text{Tot}}(0) - \frac{w^*}{K_1} y_1(0) \right] \left[\frac{\rho_1}{\tau_1\rho_1 + \tau_2\rho_2} \right] \geq \left[\frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 - K_1 \right],$$

which proves the statement. Note that if the total workload is equal to the critical workload, then (3.26) does not have a positive solution since (3.26) would imply that $\exp\{\alpha_2 t\} = 0$. It would only be satisfied by $t = 0$. This means that if $y_1(0) \in \Pi_4$ and $\bar{W}_{\text{Tot}}(0) - w^* = 0$, then the function $x(t)$ remains in region Π_4 for ever. Since in this first case $\bar{W}_{\text{Tot}}(0) - w^* > 0$, we have $\frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 - K_1 > 0$. Thus, by combining the fact that $y_1(0) \leq K_1$ and the previous display, we have shown that $\bar{W}_{\text{Tot}}(0) - \frac{w^*}{K_1} y_1(0) > \bar{W}_{\text{Tot}}(0) - w^* > 0$. Combining these arguments leads to the conclusion that the equation $y_1(t^*) = K_1$ has a (unique) positive solution, say t^* . Therefore, $y(t) \in \Pi_3$ for $t > t^*$. In other words, it is enough to prove that, for $\bar{W}_{\text{Tot}}(0) - w^* > 0$, the function $x(t)$ converges to a point in region Π_3 .

We now show that it converges to the invariant manifold in region Π_3 . By (3.12), we have $\Delta_2 \bar{W}_{\text{Tot}}(0) = \frac{\rho_2}{\rho_1} K_1$. By (3.19) and (3.22),

$$|x(t) - \Delta_2 \bar{W}_{\text{Tot}}(0)| \leq \left| \frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 y_1(0)}{\tau_2} - \frac{\rho_2}{\rho_1} K_1 \right| \exp\{\xi_{22} t\},$$

and recall that $\xi_{22} < 0$. Also, for any closed and bounded interval of the form $[0, M]$ for $M > 0$, the quantity $\left| \frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 y_1(0)}{\tau_2} - \frac{\rho_2}{\rho_1} K_1 \right|$ is uniformly bounded by $\sup_{y_1(0) \in [0, M]} \left| \frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 y_1(0)}{\tau_2} - \frac{\rho_2}{\rho_1} K_1 \right|$. That is, the convergence is uniform for any initial state in a compact set.

Case 2: $\bar{W}_{\text{Tot}}(0) - w^* < 0$. Adapting the previous case, we first show that if $\bar{W}_{\text{Tot}}(0) - w^* < 0$ and $y_1(0) \in \Pi_4$, then the function $x(t)$ remains for ever in region Π_4 . To see this, by (3.21) we have

$$y_1(t) = \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 - \left[\bar{W}_{\text{Tot}}(0) - \frac{w^*}{K_1} y_1(0) \right] \frac{K_1}{w^*} \exp\{\alpha_2 t\}.$$

Observing that $y_1(0) \leq K_1$ in Π_4 , we derive the following inequality:

$$y_1(t) \leq \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1 + \left(1 - \frac{\bar{W}_{\text{Tot}}(0)}{w^*} \right) K_1 \exp\{\alpha_2 t\}.$$

Note that the term $\left(1 - \frac{\bar{W}_{\text{Tot}}(0)}{w^*} \right)$ is positive by the assumption $\bar{W}_{\text{Tot}}(0) - w^* < 0$ and the inequality $\exp\{\alpha_2 t\} \leq 1$. Combining these three facts, we have $y_1(t) \leq K_1$.

Now, we show that if the process starts in region Π_3 and $\bar{W}_{\text{Tot}}(0) - w^* < 0$, then there exists a finite time t^{**} such that $x(t) \in \Pi_4$ after that time. Again, here

we prove that the equation $y_1(t^*) = K_1$ has a positive solution. Then, the result follows by observing that $y_1(0) \geq K_1 > \Delta_1 \bar{W}_{\text{Tot}}(0)$, and in that case $x(t)$ is an increasing function. By (3.19), we have

$$\begin{aligned} y_1(t) &= -\frac{\tau_2}{\tau_1} \left(\frac{\bar{W}_{\text{Tot}}(0)}{\tau_2} - \frac{\tau_1 y_1(0)}{\tau_2} - \frac{\rho_2}{\rho_1} K_1 \right) \exp\{\xi_{22}t\} + \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 \\ &= -\left(\frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 - y_1(0) \right) \exp\{\xi_{22}t\} + \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1. \end{aligned}$$

Setting $y_1(t) = K_1$, we obtain

$$(3.27) \quad \frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} = \left(\frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 - y_1(0) \right) \exp\{\xi_{22}t\}.$$

To show that the previous equation has a positive solution, it is enough to prove that

$$\frac{\tau_1}{\bar{W}_{\text{Tot}}(0) - w^*} \left(\frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 - y_1(0) \right) \geq 1.$$

Recall that $y_1(0) \geq K_1$ and that $\bar{W}_{\text{Tot}}(0) - w^* < 0$. We can now derive the previous inequality by observing that

$$\frac{\tau_1}{\bar{W}_{\text{Tot}}(0) - w^*} \left(\frac{\bar{W}_{\text{Tot}}(0) - w^*}{\tau_1} + K_1 - K_1 \right) \geq 1.$$

Analogously with the previous case, we note that if the total workload is equal to the critical workload, (3.27) does not have a positive solution. This means that if $y_1(0) \in \Pi_3$ and $\bar{W}_{\text{Tot}}(0) - w^* = 0$, then the function $x(t)$ remains in region Π_3 for ever.

For the convergence to the invariant manifold, it follows that $\Delta_2 \bar{W}_{\text{Tot}}(0) = \frac{\rho_2}{\rho_1} \frac{\bar{W}_{\text{Tot}}(0)}{w^*} K_1$ by (3.12). Moreover, (3.21) and (3.22) lead to

$$|x(t) - \Delta_2 \bar{W}_{\text{Tot}}(0)| \leq \left| \frac{\tau_1}{\tau_2} \left(\frac{\bar{W}_{\text{Tot}}(0) w^*}{w^* K_1} - y_1(0) \right) \right| \exp\{\alpha_2 t\},$$

and recall that $\alpha_2 < 0$. The uniform convergence in a compact set for any initial state follows for the same reason as in the previous case.

C a s e 3: $\bar{W}_{\text{Tot}}(0) - w^* = 0$. In this case, the convergence follows from the comments we made in the previous two cases. If $\bar{W}_{\text{Tot}}(0) - w^* = 0$, then the function $x(\cdot)$ always stays in the region where $y_1(0)$ lies (see comments after (3.26) and (3.27)). As we see, the function converges in regions Π_3 and Π_4 .

This concludes the proof of Theorem 3.1, which will be applied to prove a diffusion theorem for the queue length process in the next section.

4. DIFFUSION APPROXIMATIONS

The main objective in this section is to show a state-space collapse (SSC) property for the diffusion queue length process. This yields a diffusion limit theorem for the diffusion-scaled process. To do it, we follow the strategy set up in [1]. Let us consider a family of single-server systems indexed by $n \in \mathbb{N}$, where n tends to infinity, with the same basic structure as that of the network described in Section 2. To indicate the position in the sequence of networks, a superscript n will be appended to the network parameters and processes. Diffusion (or central limit theorem) scaling is indicated by placing a hat over a process. Thus, the well-known diffusion scaling is given by $\widehat{\mathfrak{X}}^n(\cdot) = \frac{1}{n} \mathfrak{X}^n(n^2 \cdot)$. Let $\rho^n = \sum_{i=1}^2 \rho_i^n = (\gamma^n)^T \beta$. We set $\gamma^n = (I - P^T)^{-1} \lambda^n$, $\lambda^n = \lambda(1 - \frac{\theta}{n})$, $\mu^n \equiv \mu$, $P^n \equiv P$, and $K_i^n = nK_i$, where θ is a positive real number. Thus, we have $\rho^n = 1 - \frac{\theta}{n}$. It is clear that under the critical loading assumption, $\lambda^n \rightarrow \lambda$ and $n(1 - \rho^n) \rightarrow \theta$ as $n \rightarrow \infty$. These are our heavy-traffic assumptions. Furthermore, we assume that $\widehat{Q}^n(0) = \frac{1}{n} Q^n(0) \rightarrow \widehat{Q}(0)$, where $\widehat{Q}(0)$ is a positive constant. The service allocation function for the n -th model is given by

$$R_i^n(q) = \begin{cases} \min_i \{q_i, nK_i\} / \sum_{j=1}^I \min\{q_j, nK_j\} & \text{if } q_i \neq 0, \\ 0 & \text{if } q_i = 0, \end{cases}$$

where we observe that $R_i^1(\cdot) = R_i(\cdot)$. Recall that $R(\cdot)$ is a Lipschitz-continuous function on $\mathbb{R}_+^2 \setminus \{0\}$ and note that the following scaling property holds: $R^n(n \cdot) = R(\cdot)$. In the sequel, we state the technical assumptions that allow us to apply the functional central limit theorem and Bramson's weak law estimates. We assume that, for $i = 1, 2$,

$$\frac{u_i^n(1)}{n} \rightarrow 0 \quad \text{and} \quad \frac{v_{ii}^{(1),n}(1)}{n} \rightarrow 0,$$

in probability as $n \rightarrow \infty$. In addition, we assume that there exists a function $\eta(\cdot)$, with zero limit at infinity, such that, for $j > 1$, $i, l \in \{1, 2\}$ and $k \in \mathbb{N}$,

$$\mathbb{E}(u_i^n(j)^2 \mathbb{1}_{\{u_i^n(j) > a\}}) \leq \eta(a) \quad \text{and} \quad \mathbb{E}(v_{il}^{(k),n}(j)^2 \mathbb{1}_{\{v_{il}^{(k),n}(j) > a\}}) \leq \eta(a).$$

More details about these assumptions can be found in [1] and [17]. In the rest of this section, we assume that the previous assumptions are satisfied without referring to them again. The main result in this section is the following

THEOREM 4.1. *Assume that the diffusion-scaled initial state converges in distribution as $n \rightarrow \infty$, i.e. $\widehat{Q}^n(0) \xrightarrow{d} \Delta \widehat{W}_{\text{Tot}}(0)$, where “ \xrightarrow{d} ” denotes convergence in distribution. Then, the diffusion-scaled stochastic process converges in distribution as $n \rightarrow \infty$, i.e.*

$$\widehat{Q}^n(\cdot) \xrightarrow{d} \Delta \widehat{W}_{\text{Tot}}(\cdot),$$

where $\widehat{W}_{\text{Tot}}(t)$ is a one-dimensional Brownian motion with drift $-\theta$ and variance $\sigma^2 = \sum_{i=1}^2 \lambda_i \sigma_{s_i}^2 + \tau_i^2 \lambda_i c_{u_i}^2$, where $\sigma_{s_i}^2$ denotes the variance of $s_i(j)$, and c_{u_i} denotes the coefficient of variation of $u_i(j)$.

The proof of this theorem is given at the end of this section. This theorem can be applied to develop heavy-traffic approximations for the joint queue length process, as they are both a piecewise linear function of a one-dimensional RBM, of which the time-dependent distribution can be expressed in closed form (in terms of the Gaussian cdf and pdf); cf. [2]. To see this, note that the functions Δ_1 and Δ_2 are invertible with inverses

$$\Delta_1^{-1}(w) = \begin{cases} ww^*/K_1 & \text{if } w < w^*, \\ \tau_1(w - K_1) + w^* & \text{if } w \geq w^*, \end{cases}$$

$$\Delta_2^{-1}(w) = \begin{cases} \rho_1 ww^*/(\rho_2 K_1) & \text{if } w < w^*, \\ \infty & \text{if } w \geq w^*. \end{cases}$$

We know that $\widehat{W}_{\text{Tot}}(t)$ is an RBM($-\theta, \sigma^2$). Let $\widehat{Q}(\cdot)$ be the diffusion limit. We have, for $x, y \geq 0$,

$$\begin{aligned} \mathbb{P}(\widehat{Q}_1(t) > x, \widehat{Q}_2(t) > y) &= \mathbb{P}(\Delta_1 \widehat{W}_{\text{Tot}}(t) > x, \Delta_2 \widehat{W}_{\text{Tot}}(t) > y) \\ &= \mathbb{P}(\widehat{W}_{\text{Tot}}(t) > \Delta_1^{-1}(x), \widehat{W}_{\text{Tot}}(t) > \Delta_2^{-1}(y)) \\ &= \mathbb{P}(\widehat{W}_{\text{Tot}}(t) > z), \end{aligned}$$

where $z = \max\{\Delta_1^{-1}(x), \Delta_2^{-1}(y)\}$. The last expression can be written in terms of the Gaussian cdf and pdf; cf. [2]. Also, using a similar coupling argument as in [20], it can be shown that one can interchange the steady-state and heavy-traffic limits in this case. For space considerations we will leave this as detail to the reader.

The rest of this section is devoted to a proof of Theorem 4.1. It is organized as follows.

1. We first prove a heavy-traffic limit theorem for the total workload process.
2. After that, we define a family of shifted fluid-scaled processes in Section 4.2 and we show that they are stochastically bounded in Section 4.3.
3. In Section 4.4, we establish some technical auxiliary estimates and tightness of these families. Moreover, we establish that limit points of these fluid-scaled processes, which are called *fluid limits*, are in fact fluid model solutions as defined in Section 3. The development in this section is very similar to those in Bramson [1] and is therefore kept concise.
4. In Section 4.5, we establish a similar tightness property for a family of shifted fluid-scaled workload processes.
5. The proof is then completed by showing a state-space collapse result in Section 4.6.

4.1. Convergence of the total workload.

LEMMA 4.1. *Under the critical loading assumption, the diffusion-scaled total workload, $\widehat{W}_{\text{Tot}}^n(t) = \frac{1}{n}\widehat{W}_{\text{Tot}}^n(n^2t)$, converges in distribution to an RBM($-\theta, \sigma^2$).*

Proof. By (2.11) for $i, l = 1, 2$ and $i \neq l$, we know that the total service requirement of the j -th external customer (including customers who are already in queue i at time zero) who enters queue i is given by

$$s_i(j) = v_{ii}^{(1)}(j) + \sum_{k=1}^{\infty} \varphi_{ii}^{(k)}(j)v_{ii}^{(k+1)}(j) + \varphi_{il}^{(k)}(j)v_{il}^{(k+1)}(j).$$

We define the following process:

$$(4.1) \quad W_G(t) := \sum_{i=1}^2 \sum_{j=1}^{Q_i(0)+E_i(t)} s_i(j) - \int_0^t \mathbb{1}_{\{W_G(s)>0\}} ds.$$

We recall that $E_i(\cdot)$ denotes the external arrivals at queue i , and, by construction of the model, $\{s_i(j)\}_{j=1}^{\infty}$ is a sequence of positive i.i.d. random variables for $i = 1, 2$. The process given in (4.1) represents the workload of a single queue with input given by two independent renewal processes that have independent service requirements of each other. The busy time of this system is $\int_0^t \mathbb{1}_{\{W_G(s)>0\}} ds = T_1(t) + T_2(t)$, and it represents the busy time of server in layer 2. Note that the busy time is zero if and only if both queues are empty. The diffusion-scaled process (after subtracting and adding the means of the random quantities in (4.1)) is given by

$$\begin{aligned} \widehat{W}_G^n(t) &= \sum_{i=1}^2 \frac{1}{n} \left(\sum_{j=1}^{n^2(\bar{Q}_i^n(0)+\bar{E}_i^n(t))} s_i^n(j) - \tau_i n^2 (\bar{Q}_i^n(0) + \bar{E}_i^n(t)) \right) \\ &\quad + \tau_i (\widehat{E}_i^n(t) - \lambda_i^n nt) + \tau_i \bar{Q}_i^n(0) + \tau_i \lambda_i^n nt - nt + \widehat{Y}^n(t), \end{aligned}$$

where $\widehat{Y}^n(t) = \int_0^t \mathbb{1}_{\{\widehat{W}_G^n(s)=0\}} ds$. By the time change theorem in [13] and the functional central limit theorem (see also [2], Theorem 6.8), we have $\widehat{W}_G^n(\cdot) \xrightarrow{d} \widehat{W}_G(\cdot)$ as $n \rightarrow \infty$. Furthermore, the limit can be described as

$$\begin{aligned} \widehat{W}_G(t) &= \sum_{i=1}^2 \tau_i \bar{Q}_i(0) - \theta t + \sqrt{\lambda_1 \sigma_{s_1}^2} \mathcal{W}_1(t) + \sqrt{\lambda_2 \sigma_{s_2}^2} \mathcal{W}_2(t) \\ &\quad + \tau_1 \sqrt{\lambda_1 c_{u_1}^2} \mathcal{W}_3(t) + \tau_2 \sqrt{\lambda_2 c_{u_2}^2} \mathcal{W}_4(t) + \widehat{Y}(t), \end{aligned}$$

where $\mathcal{W}_i(t)$, $1 \leq i \leq 4$, are independent one-dimensional standard Brownian motions, and $\widehat{Y}(t)$ can be increased only if $\widehat{W}_G(t) = 0$. Thus, the process $\widehat{W}_G(t)$ satisfies a one-dimensional Skorokhod problem. That is, $\widehat{W}_G(t)$ is a reflected Brownian

motion starting at point $\sum_{i=1}^2 \tau_i \bar{Q}_i(0)$ with drift $-\theta$ and variance σ^2 which is given by $\sigma^2 = \sum_{i=1}^2 \lambda_i \sigma_{s_i}^2 + \tau_i^2 \lambda_i c_{u_i}^2$, where $\sigma_{s_i}^2$ denotes the variance of $s_i(j)$, and c_{u_i} denotes the coefficient of variation of $u_i(j)$. The second moment of the random variables $s_i(j)$ is given by (2.13). In case of Poisson external arrivals, this result is reduced to the well-known heavy-traffic limit (see, e.g., [7], Theorem 2.3).

Now, we shall prove that

$$(4.2) \quad W_{\text{Tot}}(t) = W_G(t), \quad t \geq 0.$$

We do it by showing that we can change the label of how we count the service requirements of the customers in the system. Counting the total service requirements of the external arrivals until time t is the same as counting the immediate and remaining service requirements of the total arrivals in the system until time t . Recall that $s_i(j) = v_{ii}^{(1)}(j) + s'_i(j)$ for $i = 1, 2$, where $s'_i(j)$ are the future service requirements. If $t = 0$, then we have nothing to prove as $E_i(0) = A_i(0) = 0$. If $t > 0$ and $\Phi_{li}(S_l(T_l(t))) = 0$ for $l, i \in \{1, 2\}$, then (4.2) holds as all the departures until time t leave the system, and so $s'_i(j) = 0$ for $j = 1, \dots, S_i(T_i(t))$ and $E_i(t) = A_i(t)$.

For the general case, we first assume that all customers are routed only one time until the time t . The right-hand side in (4.1) can be written as

$$(4.3) \quad \sum_{i=1}^2 \frac{Q_i(0)+E_i(t)}{\sum_{j=1}^{S_i(T_i(t))} v_{ii}^{(1)}(j) + \sum_{j=S_i(\mu_i T_i(t))+1}^{Q_i(0)+E_i(t)} s'_i(j) + \sum_{j=1}^{S_i(T_i(t))} s'_i(j).$$

In order to separate the customers who depart from node i (if they are routed or leave the system), we define for $i, l \in \{1, 2\}$ and $i \neq l$ the following set: $\mathcal{A}^i = \{j : 1 \leq j \leq S_i(T_i(t))\}$, which includes all the customers who depart from node i . Customers can be routed at the same node, and we denote this set by $\mathcal{A}_1^i = \{j \in \mathcal{A}^i : \varphi_{ii}(j) = 1\}$, or to the other node, and we denote this set by $\mathcal{A}_2^i = \{j \in \mathcal{A}^i : \varphi_{il}(j) = 1\}$. Last, we define the set $\mathcal{A}_3^i = \mathcal{A}^i \setminus (\mathcal{A}_1^i \cup \mathcal{A}_2^i)$, which represents all customers who leave the system. If $j \in \mathcal{A}_1^i$, then there exist natural numbers k_j^i such that $S_i(T_i(t)) \leq k_j^i \leq A_i(t) - 1$ and

$$s'_i(j) = v_{ii}^{(1)}(k_j^i + 1) + s'_i(k_j^i + 1).$$

Similarly, if $j \in \mathcal{A}_2^i$, then for $i \neq l$ there exist h_j^i such that $S_l(T_l(t)) \leq h_j^i \leq A_l(t) - 1$ and

$$s'_i(j) = v_{il}^{(1)}(h_j^i + 1) + s'_i(h_j^i + 1).$$

Last, $j \in \mathcal{A}_3^i$ means that the j -th customer leaves the system after his first service, and so $s'_i(j) = 0$. The quantities k_j^i and h_j^i denote the number of customers in

node $i = 1, 2$ (including that one in service) who the j -th customer meets after his departure from node $l = 1, 2$. Therefore, (4.3) can be written as

$$\sum_{i=1}^2 \sum_{j=1}^{Q_i(0)+A_i(t)} v_{ii}^{(1)}(j) + \sum_{j=S_i(T_i(t))+1}^{Q_i(0)+A_i(t)} s'_i(j).$$

Now, let $k_0^i(j)$ be the number of routes at node i , for $1 \leq j \leq Q_i(0) + E_i(t)$, until time t . Also, we define the number $k_0^i = \max_j k_0^i(j)$, and the number of maximum routes for any external arrival in the system until time t , $k_0 = \max_{i=1,2} \{k_0^i\}$. Observe that, for any $t \geq 0$, $k_0 \in \mathbb{N}$ and it is finite because we assume Markov routing. Take the partition $[0, t] = \bigcup_{m=0}^{k_0} [t_m, t_{m+1}]$, where $t_0 = 0$ and $t_{k_0+1} = t$, so that in each interval $[t_m, t_{m+1}]$ any customer in the system can be routed only one time and so that in the interval $[t_{k_0}, t]$ there is no routing. Now, we can write the right-hand side of (4.1) as

$$\sum_{i=1}^2 \sum_{j=1}^{Q_i(0)+E_i(t)} v_{ii}^{(1)}(j) + \sum_{j=S_i(T_i(t_1))+1}^{Q_i(0)+E_i(t)} s'_i(j) + \sum_{j=1}^{S_i(T_i(t_1))} s'_i(j).$$

Applying the previous idea where customers are routed only one time per interval, we see that the above quantity can be written as

$$\sum_{i=1}^2 \sum_{j=1}^{Q_i(0)+E_i(t)+\sum_{l=1}^2 \Phi_{li}(S_l(T_l(t_1)))} v_{ii}^{(1)}(j) + \sum_{j=S_i(T_i(t_1))+1}^{Q_i(0)+E_i(t)+\sum_{l=1}^2 \Phi_{li}(S_l(T_l(t_1)))} s'_i(j).$$

Split again the last term of the previous quantity until time t_2 , and apply the previous idea when customers are routed only one time to obtain

$$\sum_{i=1}^2 \sum_{j=1}^{Q_i(0)+E_i(t)+\sum_{l=1}^2 \Phi_{li}(S_l(T_l(t_2)))} v_{ii}^{(1)}(j) + \sum_{j=S_i(T_i(t_2))+1}^{Q_i(0)+E_i(t)+\sum_{l=1}^2 \Phi_{li}(S_l(T_l(t_2)))} s'_i(j).$$

Adapting the previous steps until time t_{k_0} , and recalling that the total arrival process is given by (2.7), we derive (4.2). ■

In the sequel, this result plays a key role. The next step is to define the so-called *shifted fluid-scaled processes* and to show that they are stochastically bounded.

4.2. Shifted fluid-scaled processes. We introduce the shifted fluid scaling, which is an extension of the classical fluid scaling. Let $T > 0$ and $m \leq nT$. We define

$$\bar{A}_i^{n,m}(t) = \frac{1}{n} (A_i^n(nm + nt) - A_i^n(nm)),$$

and the analogous scaling for the processes $T(\cdot)$, $Y_{L_2}(\cdot)$, and $E(\cdot)$. For the departure process, the cumulative service time process, and the routing process we have

$$\bar{S}_i^{n,m}(t) = \frac{1}{n} \left(S_i^n(nt + T_i^n(nm)) - S_i^n(T_i^n(nm)) \right),$$

$$\bar{V}_i^{n,m}(k) = \frac{1}{n} \left[V_i^n(nk + S_i^n(T_i^n(nm))) - T_i^n(nm) \right],$$

$$\bar{\Phi}_{li}^{n,m}(k) = \frac{1}{n} \Phi_{li}^n(nk + S_l^n(T_l^n(nm))) - \frac{1}{n} \Phi_{li}^n(S_l^n(T_l^n(nm))).$$

Last, the queue length process is scaled as follows: $\bar{Q}_i^{n,m}(t) = \frac{1}{n} Q_i^n(nm + nt)$, and analogously for the scaling of the immediate and total workload. The system dynamics (2.5)–(2.10) under the shifted fluid scaling become

$$(4.4) \quad \begin{aligned} \bar{Q}_i^{n,m}(t) &= \bar{Q}_i^{n,m}(0) + \bar{E}_i^{n,m}(t) \\ &\quad + \sum_{l=1}^2 \bar{\Phi}_{li}^{n,m} \left(\bar{S}_l^{n,m}(\bar{T}_l^{n,m}(t)) \right) - \bar{S}_i^{n,m}(\bar{T}_i^{n,m}(t)), \end{aligned}$$

$$(4.5) \quad \bar{T}_i^{n,m}(t) = \int_0^t R_i^n(n\bar{Q}^{n,m}(s)) ds = \int_0^t R_i(\bar{Q}^{n,m}(s)) ds,$$

$$(4.6) \quad \begin{aligned} \bar{W}_i^{n,m}(t) &= \bar{V}_i^{n,m} \left[\bar{Q}_i^{n,m}(0) + \bar{E}_i^{n,m}(t) \right. \\ &\quad \left. + \sum_{l=1}^2 \bar{\Phi}_{li}^{n,m} \left(\bar{S}_l^{n,m}(\bar{T}_l^{n,m}(t)) \right) \right] - \bar{T}_i^{n,m}(t), \end{aligned}$$

$$(4.7) \quad \sum_{i=1}^2 \bar{T}_i^{n,m}(t) + \bar{Y}_{L_2}^{n,m}(t) = t,$$

$$\bar{Y}_{L_2}^{n,m}(t) \text{ increases} \Rightarrow \bar{W}_1^{n,m}(t) + \bar{W}_2^{n,m}(t) = 0, \quad t \geq 0, \quad i = 1, 2.$$

In the sequel, we shall be referring to *shifted fluid scaling*, and *shifted fluid process* as *shifted scaling*, and *shifted process*, for simplicity. The main step of the SSC property is to show that the shifted process can be approximated by a solution of the fluid model. This is done in Section 4.4. We first need to prove that the shifted workload and shifted queue length are bounded at zero, which we do in the following section. Using these bounds and some properties of the cumulative service time (2.6), we can apply the results of [1], Sections 4 and 5.

4.3. Bounding the shifted processes. First, we find the relation between the diffusion scaling and the shifted scaling. Although this relation is easily obtained and is already known in the literature (e.g., [19]), we provide it here for completeness. Fix $L > 1$ and define the shifted fluid processes on $[0, L]$. The interval $[0, n^2T]$ can be covered by $[nt] + 1$ overlapping intervals as follows. For $t \in [0, n^2T]$, there exist $s \in [0, L]$ and $m \in \{0, 1, \dots, [nt]\}$ such that

$$n^2t = nm + ns.$$

We can write the relation between the diffusion scaling and the shifted scaling as follows. For $s \leq L$,

$$(4.8) \quad \bar{Q}^{n,m}(s) = \widehat{Q}^n\left(\frac{nm + ns}{n^2}\right),$$

$$(4.9) \quad \bar{W}_{\text{Tot}}^{n,m}(s) = \widehat{W}_{\text{Tot}}^n\left(\frac{nm + ns}{n^2}\right).$$

By Lemma 4.1 and (4.9), it follows that for any $\epsilon > 0$ there exists a constant $B_1 > 0$ such that

$$(4.10) \quad \liminf_{n \rightarrow \infty} \mathbb{P}\left(\max_{m \leq nT} \bar{W}_{\text{Tot}}^{n,m}(0) \leq B_1\right) \geq 1 - \epsilon.$$

We denote the event $\{\omega \in \Omega : \max_{m \leq nT} \bar{W}_{\text{Tot}}^{n,m}(0) \leq B_1\}$ by $\mathcal{G}_1^n(B_1)$. Using now (4.10), it can be shown that the shifted queue length process is stochastically bounded at zero.

LEMMA 4.2. *Let $T > 0$. For any $\epsilon > 0$ there exists a constant $B_2 > 0$ such that*

$$(4.11) \quad \liminf_{n \rightarrow \infty} \mathbb{P}\left(\max_{m \leq nT} \bar{Q}_i^{n,m}(0) \leq B_2\right) \geq 1 - \epsilon, \quad i = 1, 2.$$

We denote the event $\{\omega \in \Omega : \max_{m \leq nT} \bar{Q}_i^{n,m}(0) \leq B_2\}$ by $\mathcal{G}_2^n(B_2)$.

Proof. We prove the result by deriving a contradiction, so suppose that (4.11) does not hold. Thus, assume there exists at least one i such that $\bar{Q}_i^{n,m}(0)$ is not stochastically bounded. In other words, there exists a $\delta > 0$ such that, for any $B > 0$,

$$(4.12) \quad \liminf_{n \rightarrow \infty} \mathbb{P}\left(\max_{m \leq nT} Q_i^n(nm) > Bn\right) > \delta.$$

Suppose that m_n is such that it optimizes the quantity $\max_{m \leq nT} Q_i^n(nm)$. We can choose in (4.10), $\epsilon = \frac{\delta}{3}$ and a (large enough) constant B_1 . Also, we choose a

constant B such that $B > 2\frac{B_1}{\beta_i}$. By the definition of the total workload (2.14) and (2.15), we have

$$\begin{aligned} \max_{m \leq nT} W_{\text{Tot}}^n(nm) &\geq \sum_{j=1}^{S_i^n(T_i^n(nm_n))+1} v_{ii}^{(1)}(j) - T_i^n(nm_n) \\ &+ \sum_{j=S_i^n(T_i^n(nm_n))+2}^{S_i^n(T_i^n(nm_n))+Q_i^n(nm_n)} v_{ii}^{(1)}(j) \geq \sum_{j=S_i^n(T_i^n(nm_n))+2}^{S_i^n(T_i^n(nm_n))+Q_i^n(nm_n)} v_{ii}^{(1)}(j). \end{aligned}$$

We know that $v_{ii}^{(1)}(j)$ are i.i.d. with mean β_i . Also, in the previous summation we have $j > S_i^n(T_i^n(nm_n))$, which means that $v_{ii}^{(1)}(j)$ are independent of the process $S_i^n(T_i^n(nm_n))$. Define the following event:

$$(4.13) \quad \mathcal{G}^n = \left\{ \omega \in \Omega : \left| \frac{1}{Bn} \sum_{j=S_i^n(T_i^n(nm_n))+2}^{S_i^n(T_i^n(nm_n))+Bn} v_{ii}^{(1)}(j) - \beta_i \right| < \frac{\beta_i}{2} \right\}.$$

By the weak law of large numbers (which we can apply due to the independence of the $v_{ii}^{(1)}(j)$ and $S_i^n(T_i^n(nm_n))$), we have, for large n , $\mathbb{P}(\mathcal{G}^n) \geq 1 - \frac{\delta}{3}$. In the sequel, we assume that $\omega \in \mathcal{G}^n \cap \mathcal{G}_1^n(B_1) \cap (\mathcal{G}_2^n(B))^c$, and observe that $\mathbb{P}(\mathcal{G}^n \cap \mathcal{G}_1^n(B_1) \cap (\mathcal{G}_2^n(B))^c) \geq \frac{\delta}{3}$. Applying (4.12) and dividing by n , we derive

$$B_1 \geq \max_{m \leq nT} \bar{W}_{\text{Tot}}^{n,m}(0) \geq \frac{B}{Bn} \sum_{j=S_i^n(T_i^n(nm_n))+2}^{S_i^n(T_i^n(nm_n))+Bn} v_{ii}^{(1)}(j).$$

By (4.13) and the last inequality, we obtain for sufficiently large n the relation $B_1 \geq B(\beta_i - \frac{\beta_i}{2}) > 2\frac{B_1}{2}$. This yields a contradiction. ■

Having proved that the shifted processes are bounded, we can show that the shifted processes can be approximated by a solution of the fluid model. This is the topic of the next section, in which we use a very similar approach to that in Bramson [1], Sections 4 and 5.

4.4. Uniform fluid approximation. By Proposition 5.1 in [1], we have, for any $\epsilon > 0$,

$$(4.14) \quad \mathbb{P}\left(\max_{m < nT} \|\bar{E}^{n,m}(\cdot) - \lambda^n \cdot\|_L > \epsilon \right) \leq \epsilon.$$

Also, by Proposition 5.2 in [1] it is known that the shifted arrival process is almost Lipschitz continuous, which means that, for some $N_1 > 0$,

$$\mathbb{P}\left(\sup_{t_1, t_2 \in [0, L]} |\bar{E}^{n,m}(t_2) - \bar{E}^{n,m}(t_1)| > N_1 |t_2 - t_1| + \epsilon \text{ for some } m < nT \right) \leq \epsilon.$$

Furthermore, using the definition of the cumulative service time (2.6), the property $\sum_{i=1}^2 R_i^n(q) = 1$, and the observation that $Y_{L_2}^n(\cdot)$ and $T^n(\cdot)$ are increasing functions in time, we conclude that the shifted process, $T^{n,m}(\cdot)$, and the shifted idle time are Lipschitz continuous with constant equal to one.

PROPOSITION 4.1. *Let $\epsilon > 0$. Then, for an appropriate large n and for $i = 1, 2$, we have*

$$(4.15) \quad \mathbb{P}\left(\max_{m < nT} \|\bar{S}_i^{n,m}(\bar{T}_i^{n,m}(\cdot)) - \mu_i \bar{T}_i^{n,m}(\cdot)\|_L > \epsilon\right) < \epsilon,$$

$$(4.16) \quad \mathbb{P}\left(\max_{m < nT} \left\| \sum_{l=1}^2 \bar{\Phi}_{li}^{n,m}(\bar{S}_l^{n,m}(\bar{T}_l^{n,m}(\cdot))) - \sum_{l=1}^2 \mu_l p_{li} \bar{T}_l^{n,m}(\cdot) \right\|_L > \epsilon\right) < \epsilon,$$

$$(4.17) \quad \mathbb{P}\left(\max_{m < nT} \left\| \bar{V}_i^{n,m} \left(\bar{Q}_i^{n,m}(0) + \bar{E}_i^{n,m}(\cdot) + \sum_{l=1}^2 \bar{\Phi}_{li}^{n,m}(\bar{S}_l^{n,m}(\bar{T}_l^{n,m}(\cdot))) \right) - \beta_i (\bar{Q}_i^{n,m}(0) + (\lambda_i^n \cdot) + \sum_{l=1}^2 p_{li} \bar{T}_l^{n,m}(\cdot)) \right\|_L > \epsilon\right) < \epsilon.$$

Proof. It is shown in [1] that, for a renewal process $S(\cdot)$,

$$\mathbb{P}\left(\sup_{m \leq nT} \sup_{t \leq L} \left| \frac{1}{n} (S_i^n((nm + nt)) - S_i^n(nm)) - \mu t \right| \geq \epsilon\right) < \epsilon,$$

which is equivalent to (the process can start anywhere in the interval $[0, n^2T]$)

$$(4.18) \quad \mathbb{P}\left(\sup_{u \in [0, n^2T]} \sup_{t \leq L} \left| \frac{1}{n} (S_i^n((u + nt)) - S_i^n(u)) - \mu_i t \right| \geq \epsilon\right) < \epsilon.$$

Let $t' = \bar{T}_i^{n,m}(t) = \frac{1}{n}(T_i^n(nm + nt) - T_i^n(nm)) \in [0, L]$ and $u = T_i^n(nm) \leq n^2T$ for $m \leq nT$. By (4.18) we obtain

$$\mathbb{P}\left(\sup_{u \in [0, n^2T]} \sup_{t' \leq L} \left| \frac{1}{n} (S_i^n((u + nt')) - S_i^n(u)) - \mu_i t' \right| \geq \epsilon\right) < \epsilon$$

for each $i = 1, 2$. Then, (4.15) follows.

By the Lipschitz continuity of the departure process in Proposition 4.2 below, which we can prove only using (4.15), we know that $\|S_i(T_i(\cdot))\|_L \leq N_2 L n$. Using Proposition 4.2 of [1], we derive

$$\mathbb{P}\left(\max_{m < nT} \left\| \sum_{l=1}^2 \bar{\Phi}_{li}^{n,0}(\bar{S}_l^{n,0}(\bar{T}_l^{n,0}(\cdot))) - \sum_{l=1}^2 p_{li} \bar{S}_l^{n,0}(\bar{T}_l^{n,0}(\cdot)) \right\|_L > N_2 L \epsilon\right) < \frac{\epsilon}{n}.$$

Furthermore, using the conclusion of the proof of Proposition 5.19 in [1], we obtain

$$\mathbb{P}\left(\max_{m < nT} \left\| \sum_{l=1}^2 \bar{\Phi}_{li}^{n,m} \left(\bar{S}_l^{n,m}(\bar{T}_l^{n,m}(\cdot)) \right) - \sum_{l=1}^2 p_{li} \bar{S}_l^{n,m}(\bar{T}_l^{n,m}(\cdot)) \right\|_L > N_2 L \epsilon \right) < \frac{\epsilon}{n}.$$

Applying (4.15) to the last inequality, we obtain (4.16).

To prove (4.17), we know by Lemma 4.2 that, for some $B_2 > 0$, $|Q^n(0)| \leq B_2 n$. Using (4.14), (4.17), and applying Proposition 4.2 of [1], we obtain the result. ■

In the following proposition, we show that all the shifted processes are almost Lipschitz continuous.

PROPOSITION 4.2. *Let $\bar{X}^{n,m}(\cdot)$ be any of the processes $\bar{S}^{n,m}(\cdot)$, $\bar{Q}^{n,m}(\cdot)$ and $\bar{W}^{n,m}(\cdot)$. Then, for large n , for $\epsilon > 0$ and some $N > 0$, we have*

(4.19)

$$\mathbb{P}\left(\sup_{t_1, t_2 \in [0, L]} |\bar{X}^{n,m}(t_2) - \bar{X}^{n,m}(t_1)| > N|t_2 - t_1| + \epsilon \text{ for some } m < nT\right) \leq \epsilon.$$

Proof. For the departure process and by using (4.15), we have for $i = 1, 2$ the relations

$$\begin{aligned} |\bar{S}_i^{n,m}(T_i^{n,m}(t_2)) - \bar{S}_i^{n,m}(\bar{T}_i^{n,m}(t_1))| &\leq |\bar{S}_i^{n,m}(\bar{T}_i^{n,m}(t_2)) - \mu_i \bar{T}_i^{n,m}(t_2)| \\ &\quad + |\bar{S}_i^{n,m}(\bar{T}_i^{n,m}(t_1)) - \mu_i \bar{T}_i^{n,m}(t_1)| \\ &\quad + |\mu_i \bar{T}_i^{n,m}(t_2) - \mu_i \bar{T}_i^{n,m}(t_1)| \\ &\leq N_2 |t_2 - t_1| + 2\epsilon, \end{aligned}$$

where $N_2 = \max_i \mu_i$. Using (4.14), (4.17), and the Lipschitz continuity of the cumulative service time (2.6), $T(\cdot)$, it is easy to show that the shifted total arrival process, $\bar{A}^{n,m}(\cdot)$, is almost Lipschitz continuous with $N_3 = N_1 + \|P\|N_2$, where $N_1 = \max_i \lambda_i$.

By combining the almost Lipschitz continuity for the shifted arrival and shifted departure process, $\bar{A}^{n,m}(\cdot)$, $\bar{S}^{n,m}(T(\cdot))$, the result for the shifted queue length process, $\bar{Q}^{n,m}(\cdot)$, follows with the constant $N_4 = N_3 + N_2$. Using the same idea and (4.17), we obtain the same result for the shifted immediate workload process, $\bar{W}^{n,m}(\cdot)$, with $N_5 = \frac{N_1}{\min_i \mu_i} + \|P\|$. ■

REMARK 4.1. *Adapting the techniques in [1], we can replace ϵ in the propositions above by $\epsilon(n)$ such that $\epsilon(n) \rightarrow 0$. Let $\mathcal{G}_i^n \subseteq \Omega$, $1 \leq i \leq 5$, be the “good events” such that the complements of inequalities (4.14)–(4.17) and (4.19) hold if we replace ϵ by $\epsilon(n)$. Also, let $\mathcal{G}_1^n(B)$ and $\mathcal{G}_2^n(B)$ be as in (4.10) and (4.11) with $B = \max\{B_1, B_2\}$. Denote by $\mathcal{G}_0^n(B)$ the intersection of the previous events. Because $\omega \in \mathcal{G}_0^n(B)$, we know that $|\tilde{\mathbf{x}}^{n,m}(t_2) - \tilde{\mathbf{x}}^{n,m}(t_1)| \leq N|t_2 - t_1| + \epsilon$. Also,*

by Lemmas 4.1, 4.2, and by the definition of the shifted processes (4.4)–(4.7), we have $|\tilde{\mathfrak{X}}^{n,m}(0)| \leq B$ for some positive constant B . In addition, if we replace the bound in Inequality 4.6 of [1] by a general real number, we can again show that the set of Lipschitz functions with this property is compact; see [16], Lemma 6.3.

By Remark 4.1, all the requirements in Section 4.1 of [1] hold. Thus, we can find a Lipschitz-continuous function $\tilde{\mathfrak{X}}(\cdot)$ such that, for $\epsilon(n) \rightarrow 0$,

$$(4.20) \quad \|\tilde{\mathfrak{X}}^{n,m}(\cdot, \omega) - \tilde{\mathfrak{X}}(\cdot)\|_L \leq \epsilon(n), \quad \forall \omega \in \mathcal{G}_0^n(B), \forall m \leq nT.$$

PROPOSITION 4.3. *The function $\tilde{\mathfrak{X}}(\cdot)$ is a solution of the fluid model equations on $[0, L]$.*

Proof. We shall show that the function $\tilde{\mathfrak{X}}(\cdot)$ verifies the fluid model equations (3.2)–(3.6). To do this, let $\delta > 0$. As $\epsilon(n) \rightarrow 0$, we can find a large n such that $\epsilon(n) < \delta$. It is known by (4.20) that, for large n ,

$$\|\tilde{\mathfrak{X}}^{n,m}(\cdot) - \tilde{\mathfrak{X}}(\cdot)\|_L < \delta.$$

Thus, from the heavy-traffic assumption we conclude that $|\lambda^n - \lambda| < \delta$. Using the above inequalities, (4.14), Proposition 4.1, and the triangle inequality, it can be proved in the same way as in [1], Proposition 6.2, that all the functions $\tilde{\mathfrak{X}}(\cdot)$, except for $\tilde{T}(\cdot)$, verify the fluid model equations. To prove that $\tilde{T}(\cdot)$ satisfies (3.3), we need to use the following two properties of the service allocation function: (i) $R^n(nq) = R(q)$ and (ii) $R(\cdot)$ is a Lipschitz-continuous function on $\mathbb{R}_+^2 \setminus \{0\}$; i.e., there exists a constant C such that for $q_1, q_2 \in \mathbb{R}_+^2 \setminus \{0\}$

$$|R(q_2) - R(q_1)| \leq C|q_2 - q_1|.$$

Now, using (4.5) and the above properties of the service allocation function, we can show that $\tilde{T}(\cdot)$ satisfies (3.3), and thus is a solution of the fluid model:

$$\begin{aligned} |\tilde{T}(t) - \int_0^t R(\tilde{Q}(s))ds| &= |\tilde{T}(t) - \bar{T}^{n,m}(t) + \int_0^t R(\bar{Q}^{n,m}(s))ds - \int_0^t R(\tilde{Q}(s))ds| \\ &\leq |\tilde{T}(t) - \bar{T}^{n,m}(t)| + \int_0^t |R(\bar{Q}^{n,m}(s)) - R(\tilde{Q}(s))|ds \\ &\leq \delta + \int_0^t C|\bar{Q}^{n,m}(s) - \tilde{Q}(s)| \leq \delta + \int_0^t C\delta \leq \delta + CL\delta. \quad \blacksquare \end{aligned}$$

4.5. The scaled shifted total workload process. In this section, we see that we can approximate the scaled shifted total workload process by a solution of the fluid model. We begin with a preliminary result.

PROPOSITION 4.4. For appropriately large $n \in \mathbb{N}$ and $\epsilon > 0$, we have

$$(4.21) \quad \mathbb{P} \left(\max_{m < nT} \left\| \sum_{i=1}^2 \frac{1}{n} \sum_{j=Z_i^{n,m}(\cdot)}^{B_i^{n,m}(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T (P^T)^k \bar{Q}^{n,m}(\cdot) \right\|_L > \epsilon \right) < \epsilon,$$

where

$$\begin{aligned} Z_i^{n,m}(\cdot) &= n\bar{S}_i^{n,m}(\bar{T}_i^{n,m}(\cdot)) + S_i^n(T_i^n(nm)) + 1, \\ B_i^{n,m}(\cdot) &= Q_i^n(0) + n\bar{A}_i^{n,m}(\cdot) + A_i^n(nm). \end{aligned}$$

Proof. The random variables $s'_i(j)$ depend on n , but to keep the notation simple we omit the index n . Note that (4.21) can be written as

$$\mathbb{P} \left(\max_{m < nT} \left\| \sum_{i=1}^2 \frac{1}{n} \sum_{j=Z_i^{n,m}(\cdot)}^{B_i^{n,m}(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T \bar{p}_i^{(k)} \bar{Q}_i^{n,m}(\cdot) \right\|_L > \epsilon \right) < \epsilon,$$

where $\beta^T \bar{p}_i^{(k)} = \beta_1 p_{i1}^{(k)} + \beta_2 p_{i2}^{(k)}$. We have

$$\begin{aligned} &\mathbb{P} \left(\max_{m < nT} \left\| \sum_{i=1}^2 \frac{1}{n} \sum_{j=Z_i^{n,m}(\cdot)}^{B_i^{n,m}(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T \bar{p}_i^{(k)} \bar{Q}_i^{n,m}(\cdot) \right\|_L > \epsilon \right) \\ &\leq \sum_{i=1}^2 \mathbb{P} \left(\max_{m < nT} \left\| \frac{1}{n} \sum_{j=Z_i^{n,m}(\cdot)}^{B_i^{n,m}(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T \bar{p}_i^{(k)} \bar{Q}_i^{n,m}(\cdot) \right\|_L > \frac{\epsilon}{2} \right), \end{aligned}$$

so it is enough to show that the last term is sufficiently small for $i = 1, 2$. First, we shall prove it for $m = 0$. We know by Proposition 4.2 that the shifted queue length process is Lipschitz continuous and that the shifted queue length process at zero is stochastically bounded; i.e., for $t \leq L$, $|Q_i^n(t)| \leq (N_4 + B_2)Ln$. By [1], Proposition 4.2, we derive

$$\begin{aligned} &\mathbb{P} \left(\left\| \sum_{j=S_i^n(T_i^n(\cdot))}^{Q_i^n(0)+A_i^n(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T \bar{p}_i^{(k)} Q_i^n(\cdot) \right\|_{N_4Ln} > \frac{\epsilon(N_4 + B_2)Ln}{2} \right) \\ &\leq \frac{\epsilon}{2(N_4 + B_2)Ln}, \end{aligned}$$

which leads to

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{j=Z_i^{n,0}(\cdot)}^{B_i^{n,0}(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T \bar{p}_i^{(k)} \bar{Q}_i^{n,0}(\cdot) \right\|_L > \frac{\epsilon(N_4 + B_2)L}{2} \right) \leq \frac{\epsilon}{2(N_4 + B_2)n},$$

where $B_i^{n,0}(\cdot) - Z_i^{n,0}(\cdot) = \bar{Q}_i^{n,0}(\cdot)$. Multiplying the error bounds by the number of processes $[nT] + 1$ and choosing a suitable ϵ , we derive

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{j=Z_i^{n,0}(\cdot)}^{B_i^{n,m}(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T \bar{p}_i^{(k)} \bar{Q}_i^{n,m}(\cdot)\right\|_L > \frac{\epsilon}{2}\right) \leq \frac{\epsilon}{2}. \blacksquare$$

Adapting Remark 4.1, we can replace ϵ in the proposition above by $\epsilon(n)$ such that $\epsilon(n) \rightarrow 0$ as $n \rightarrow \infty$. This will be done in the next result, where we combine all technical estimates so far to construct a “good” event.

PROPOSITION 4.5. *Let $\epsilon > 0$ and $\mathcal{G}_0^n \subseteq \Omega$ be as in Remark 4.1. Let $\mathcal{G}_6^n \subseteq \Omega$ be the event such that the complement of (4.21) holds if we replace ϵ by $\epsilon(n)$. Define the event $\mathcal{G}^n(B) = \mathcal{G}_0^n(B) \cap \mathcal{G}_6^n$. Then*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{G}^n(B)) \geq 1 - \epsilon.$$

Proof. Note that $\mathcal{G}^n(B) = \bigcap_{i=1}^6 \mathcal{G}_i^n \cap \mathcal{G}_1^n(B) \cap \mathcal{G}_2^n(B)$. We denote by $(\mathcal{G}_i)^c$ the complement of the event \mathcal{G}_i . For $1 \leq i \leq 6$, it follows by construction that $\mathbb{P}((\mathcal{G}_i^n)^c) \leq \epsilon(n)$. Also, by (4.10) and (4.11) we can choose a constant B such that, for $i = 1, 2$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{G}_i^n(B)) \geq 1 - \frac{\epsilon}{2}.$$

Combining the above inequalities, we obtain, as $n \rightarrow \infty$,

$$\begin{aligned} \mathbb{P}(\mathcal{G}^n(B)) &= \mathbb{P}\left(\left(\bigcup_{i=1}^6 (\mathcal{G}_i^n)^c \cup (\mathcal{G}_1^n(B))^c \cup (\mathcal{G}_2^n(B))^c\right)^c\right) \\ &= 1 - \mathbb{P}\left(\bigcup_{i=1}^6 (\mathcal{G}_i^n)^c \cup (\mathcal{G}_1^n(B))^c \cup (\mathcal{G}_2^n(B))^c\right) \\ &\geq 1 - 6\epsilon(n) - \epsilon \rightarrow 1 - \epsilon. \blacksquare \end{aligned}$$

In the sequel, we assume that $\omega \in \mathcal{G}^n(B)$. In other words, Proposition 4.5 allows us to use a sample-path approach. As a final step towards proving the state-space collapse property, we use this approach to show that there exists a fluid approximation for the total workload of the system.

PROPOSITION 4.6. *For all $\omega \in \mathcal{G}^n(B)$ there exists a solution of the fluid model equations, $\widetilde{W}_{\text{Tot}}(\cdot)$, such that for $m \leq nT$*

$$\|\widetilde{W}_{\text{Tot}}^{n,m}(\cdot, \omega) - \widetilde{W}_{\text{Tot}}(\cdot)\|_L \leq \epsilon(n).$$

Proof. Take $\epsilon > 0$ and let $\omega \in \mathcal{G}^n(B)$ as defined in Proposition 4.5. Assume that the functions $\widetilde{Q}(t)$ and $\widetilde{W}(t)$ satisfy (4.20). Define the function

$$\widetilde{W}_{\text{Tot}}(t) = \beta^T (I - P^T)^{-1} \widetilde{Q}(t),$$

which is a solution of the fluid model equations because so is $\tilde{Q}(t)$. Omitting again the index n in the quantity $s'_i(j)$, by the definition of the total workload, we obtain

$$\bar{W}_{\text{Tot}}^{n,m}(t) = \sum_{i=1}^2 \bar{W}_i^{n,m}(t) + \sum_{i=1}^2 \frac{1}{n} \sum_{j=Z_i^{n,m}(t)}^{B_i^{n,m}(t)} s'_i(j),$$

where the quantities $Z_i^{n,m}(t)$ and $B_i^{n,m}(t)$ are defined in Proposition 4.4. Using the triangle inequality, we thus have

$$\begin{aligned} \|\bar{W}_{\text{Tot}}^{n,m}(\cdot) - \widetilde{W}_{\text{Tot}}(\cdot)\|_L &\leq \left\| \sum_{i=1}^2 \frac{1}{n} \sum_{j=Z_i(\cdot)}^{B_i(\cdot)} s'_i(j) - \sum_{k=1}^{\infty} \beta^T (P^T)^k \bar{Q}^{n,m}(\cdot) \right\|_L \\ &+ \left\| \sum_{k=1}^{\infty} \beta^T (P^T)^k (\bar{Q}^{n,m}(\cdot) - \tilde{Q}(\cdot)) \right\|_L + \sum_{i=1}^2 \|\bar{W}_i^{n,m}(\cdot) - \widetilde{W}_i(\cdot)\|_L. \end{aligned}$$

By (4.20) and (4.21), $\|\bar{W}_{\text{Tot}}^{n,m}(\cdot) - \widetilde{W}_{\text{Tot}}(\cdot)\|_L \leq (2 + 1 + \max_i (\tau_i - \beta_i)) \epsilon(n)$. ■

4.6. State-space collapse. Now, we can state and prove the SSC property for the diffusion queue length process.

THEOREM 4.2 (SSC). *Assume that*

$$(4.22) \quad |\widehat{Q}^n(0) - \Delta \widehat{W}_{\text{Tot}}^n(0)| \rightarrow 0 \text{ in probability as } n \rightarrow \infty.$$

Then, for any $T > 0$,

$$(4.23) \quad \|\widehat{Q}^n(\cdot) - \Delta \widehat{W}_{\text{Tot}}^n(\cdot)\|_T \rightarrow 0 \text{ in probability as } n \rightarrow \infty.$$

Proof. Take $\epsilon > 0$ and let $\omega \in \mathcal{G}^n(B)$ as defined in Proposition 4.5. By Theorem 3.1, we know that there exists a constant L^* such that, for $t \geq L^*$,

$$(4.24) \quad |\tilde{Q}(t) - \Delta \widetilde{W}_{\text{Tot}}(t)| \leq \epsilon.$$

Fix $L > L^* + 1$. It is known that

$$[0, n^2 T] \subseteq [0, nL^*] \cup_{m=0}^{[nT]} [n(m + L^*), n(m + L)].$$

So, it suffices to show that

$$(4.25) \quad \max_{m \leq nT} \sup_{t \in [L^*, L]} |\bar{Q}^{n,m}(t) - \Delta \bar{W}_{\text{Tot}}^{n,m}(t)| < \epsilon$$

and

$$(4.26) \quad \sup_{t \in [0, L^*]} |\bar{Q}^{n,0}(t) - \Delta \bar{W}_{\text{Tot}}^{n,0}(t)| < \epsilon.$$

Then, by using (4.8) and (4.9), we derive (4.23). To prove (4.25), we know that by (4.20) and Proposition 4.6, for $t \leq L$,

$$(4.27) \quad |\bar{Q}^{n,m}(t) - \tilde{Q}(t)| < \epsilon$$

and

$$(4.28) \quad |\bar{W}_{\text{Tot}}^{n,m}(t) - \tilde{W}_{\text{Tot}}(t)| < \epsilon.$$

Recall that the lifting map is Lipschitz continuous with constant C_1 . Combining this with (4.24), (4.27), and (4.28), we get (4.25).

To prove (4.26), we infer by (4.20) that, for $t \leq L$,

$$|\bar{Q}^{n,0}(t) - \tilde{Q}(t)| < \epsilon \quad \text{and} \quad |\bar{W}_{\text{Tot}}^{n,0}(t) - \tilde{W}_{\text{Tot}}(t)| < \epsilon.$$

Also, by the assumption (4.22) we obtain $|\bar{Q}^{n,0}(0) - \Delta \bar{W}_{\text{Tot}}^{n,0}(0)| < \epsilon$. By the last three inequalities and Proposition 3.2, we can apply Lemma 6.1 of [1] and derive $|\tilde{Q}(t) - \Delta \tilde{W}_{\text{Tot}}(t)| < \epsilon$ for $0 \leq t \leq L^*$. In a similar way as before, we get the inequality (4.26). ■

Now, we are ready to prove Theorem 4.1, which is a result of Lemma 4.1, Theorem 4.2, and the continuous mapping theorem.

Proof of Theorem 4.1. By Lemma 4.1, we have $\widehat{W}^n(\cdot) \xrightarrow{d} \widehat{W}_{\text{Tot}}(\cdot)$. Also, the lifting map Δ is continuous. Applying the continuous mapping theorem (Theorem 1.2 in [5]), we have $\widehat{W}^n(\cdot) \xrightarrow{d} \Delta \widehat{W}_{\text{Tot}}(\cdot)$. Now, the result follows by our Theorem 4.2 and Lemma 1.3 in [5]. ■

Acknowledgments. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing, Berkeley. The research of Angelos Avelouris is funded by a TOP grant of the Netherlands Organization for Scientific Research (NWO) through project 613.001.301. The research of Maria Vlasiou and Jiheng Zhang is partly supported by two grants from the ‘Joint Research Scheme’ program, sponsored by NWO and the Research Grants Council of Hong Kong (RGC) through projects 649.000.005 and DHK007/ 11T, respectively. The research of Bert Zwart is partly supported by an NWO VICI grant.

REFERENCES

- [1] M. Bramson, *State space collapse with application to heavy traffic limits for multiclass queueing networks*, Queueing Syst. 30 (1–2) (1998), pp. 89–140.
- [2] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, Springer, New York 2001.
- [3] J. L. Dorsman, O. J. Boxma, and M. Vlasiou, *Marginal queue length approximations for a two-layered network with correlated queues*, Queueing Syst. 75 (1) (2013), pp. 29–63.
- [4] J. L. Dorsman, M. Vlasiou, and B. Zwart, *Heavy-traffic asymptotics for networks of parallel queues with Markov-modulated service speeds*, Queueing Syst. 79 (3–4) (2015), pp. 293–319.

- [5] R. Durrett, *Stochastic Calculus: a Practical Introduction*, CRC Press, Boca Raton 1996.
- [6] G. Fayolle, P. J. B. King, and I. Mitrani, *The solution of certain two-dimensional Markov models*, Adv. in Appl. Probab. 14 (2) (1982), pp. 295–308.
- [7] H. C. Gromoll, *Diffusion approximation for a processor sharing queue in heavy traffic*, Ann. Appl. Probab. 14 (2) (2004), pp. 555–611.
- [8] J. A. Rolia and K. C. Sevcik, *The method of layers*, IEEE Trans. Softw. Eng. 21 (8) (1995), pp. 689–700.
- [9] R. D. van der Mei, R. Hariharan, and P. Reeser, *Web server performance modeling*, Telecommunication Systems 16 (3–4) (2001), pp. 361–378.
- [10] W. van der Weij, S. Bhulai, and R. van der Mei, *Dynamic thread assignment in web server performance optimization*, Perform. Eval. 66 (6) (2009), pp. 301–310.
- [11] M. Vlasiou, J. Zhang, B. Zwart, and R. D. van der Mei, *Separation of timescales in a two-layered network*, in: *Proceedings of the 24th International Teletraffic Congress, Kraków 2012*.
- [12] W. Walter, *Ordinary Differential Equations*, Springer, New York 1998.
- [13] W. Whitt, *Some useful functions for functional limit theorems*, Math. Oper. Res. 5 (1) (1980), pp. 67–85.
- [14] R. J. Williams, *Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse*, Queueing Syst. 30 (1–2) (1998), pp. 27–88.
- [15] M. Woodside, J. E. Neilson, D. C. Petriu, and S. Majumdar, *The stochastic rendezvous network model for performance of synchronous client-server-like distributed software*, IEEE Trans. Comput. 44 (1) (1995), pp. 20–34.
- [16] H.-Q. Ye, J. Ou, and X.-M. Yuan, *Stability of data networks: Stationary and bursty models*, Oper. Res. 53 (1) (2005), pp. 107–125.
- [17] H.-Q. Ye and D. D. Yao, *A stochastic network under proportional fair resource control: Diffusion limit with multiple bottlenecks*, Oper. Res. 60 (3) (2012), pp. 716–738.
- [18] J. Zhang, J. Dai, and B. Zwart, *Law of large number limits of limited processor-sharing queues*, Math. Oper. Res. 34 (4) (2009), pp. 937–970.
- [19] J. Zhang, J. Dai, and B. Zwart, *Diffusion limits of limited processor sharing queues*, Ann. Appl. Probab. 21 (2) (2011), pp. 745–799.
- [20] J. Zhang and B. Zwart, *Steady state approximations of limited processor sharing queues in heavy traffic*, Queueing Syst. 60 (3–4) (2008), pp. 227–246.

Angelos Aveklouris
 Department of Mathematics and
 Computer Science
 Eindhoven University of Technology
 P.O. Box 513, 5600 MB Eindhoven
 The Netherlands
E-mail: a.aveklouris@tue.nl

Maria Vlasiou
 Department of Mathematics and
 Computer Science
 Eindhoven University of Technology
 P.O. Box 513, 5600 MB Eindhoven
 The Netherlands
E-mail: m.vlasiou@tue.nl

Jiheng Zhang
 Department of Industrial Engineering and
 Logistics Management
 The Hong Kong University of Science and
 Technology
 Hong Kong S.A.R., China
E-mail: jiheng@ust.hk

Bert Zwart
 Department of Mathematics and
 Computer Science
 Eindhoven University of Technology
 Centrum Wiskunde en Informatica
 P.O. Box 513, 5600 MB Eindhoven
 The Netherlands
E-mail: Bert.Zwart@cwi.nl

*Received on 7.1.2017;
 revised version on 26.5.2017*