Queueing Syst (2012) 71:97–136
DOI 10.1007/s11134-012-9306-x

Log-weight scheduling in switched networks

Devavrat Shah · Damon Wischik

Received: 11 August 2011 / Revised: 13 March 2012 / Published online: 17 May 2012

Abstract We consider *switched* queueing networks in which there are constraints on which queues may be served simultaneously. The scheduling policy for such a network specifies which queues to serve at any point in time. We introduce and study a variant of the popular maximum weight or backpressure policy which chooses the collection of queues to serve that has maximum weight. Unlike the maximum weight policies studied in the literature, the weight of a queue depends on logarithm of its queue-size in this paper. For any multihop switched network operating under such maximum log-weighted policy, we establish that the network Markov process is positive recurrent as long as it is underloaded. As the main result of this paper, a meaningful fluid model is established as the formal functional law of large numbers approximation. The fluid model is shown to be work-conserving. That is, work (or total queue-size) is nonincreasing as long as the network is underloaded or critically loaded. We identify invariant states or fixed points of the fluid model. When underloaded, null state is the unique invariant state. For a critically loaded fluid model, the space of invariant states is characterized as the solution space of an optimization problem whose objective is lexicographic ordering of total queue-size and the negative entropy of the queue state. An important contribution of this work is in overcoming the challenge presented by the log-weight function in establishing meaningful fluid model. Specifically, the known approaches in the literature primarily relied on the "scale invariance" property of the weight function that log-function does not possess.

D. Shah (✉)
Dept of EECS, MIT, Cambridge, MA 02139, USA
e-mail: devavrat@mit.edu

D. Wischik
Dept of Computer Science, UCL, Gower St, London WC1E 6BT, UK
e-mail: d.wischik@cs.ucl.ac.uk

Keywords Scheduling · Maximum weight · Fluid models · Log-weight · Work-conservation

Mathematics Subject Classification 60K25 · 60K30 · 90B36

1 Introduction

The scheduling problem is ubiquitous. The stochastic processing network model introduced by Harrison [10] has been quite effective in modeling a large class of such problems arising in communications, computer networks, operations research, transportation networks, etc. The switched network model (cf. Shah and Wischik [17]) considered in this paper is a restriction of the stochastic processing network model. Despite the restriction, it faithfully models a variety of application scenarios including core router's operation through input-queued switch or wireless access network. Further, it does seem to bring out the quintessential challenges involved in performance analysis of scheduling policies. Therefore, developing methods for performance analysis of switched networks are likely to lead to better understanding of general stochastic processing networks.

1.1 Switched network model

Consider a collection of N queues. Let time be discrete, indexed by $\tau \in \{0, 1, \ldots\}$. Let $Q_n(\tau)$ be the size of queue n at the beginning of timeslot τ, and write $\mathbf{Q}(\tau)$ for the vector $[Q_n(\tau)]_{1\le n\le N}$. Let $\mathbf{Q}(0)$ be the prespecified vector of initial queue sizes.

In each timeslot, each queue is offered either unit amount of service or no service as per *scheduling constraint* described below. If the queue is empty and it is offered unit amount of service, then we say that queue has idled by unit amount. Once work is served, it gets routed to another queue or leaves the network. New work may arrive in each timeslot; let each of the N queues have a dedicated exogenous arrival process.

The scheduling constraint is described by a finite set of *feasible schedules* $\mathcal{S} \subset \{0, 1\}^N$. In every timeslot, a *schedule* $\boldsymbol{\pi} \in \mathcal{S}$ is chosen; queue n is offered an amount of service π_n times the duration of the timeslot. We shall assume $\mathcal{S}$ to be *monotone*. That is,

$$\text{if } \boldsymbol{\pi} \in \mathcal{S},\ \boldsymbol{\rho} \in \{0,1\}^N \quad \text{and} \quad \boldsymbol{\rho} \le \boldsymbol{\pi} \quad \text{then } \boldsymbol{\rho} \in \mathcal{S}.$$

Further, we shall assume that $\mathbf{e}^n \in \mathcal{S}$ for all $1 \le n \le N$ where $\mathbf{e}^n$ is the schedule that only serves queue n. Let $S_{\boldsymbol{\pi}}(\tau)$ be the total length of time up to the beginning of timeslot τ in which schedule $\boldsymbol{\pi}$ has been chosen, and let $S_{\boldsymbol{\pi}}(0) = 0$. Let $Z_n(\tau)$ be the total amount of idling at queue n up to the beginning of timeslot τ, and let $Z_n(0) = 0$.

Let $A_n(\tau)$ be the total amount of work arriving to queue n up to the beginning of timeslot τ, and $A_n(0) = 0$. We will take $\mathbf{A}(\cdot)$ to be a random process satisfying the following properties: (1) The components $A_n(\cdot)$ of $\mathbf{A}(\cdot)$ are independent across n; and (2) $A_n(\cdot)$ is a Bernoulli process with mean λ_n. That is,

$$\lambda_n = \mathbb{P}\big(A_n(\tau) - A_n(\tau-1) = 1\big) = \lim_{\tau\to\infty} \frac{1}{\tau} A_n(\tau) \tag{1}$$

where the above limit exists almost surely for each queue due to strong law of large numbers for Bernoulli process. In summary, each arriving customer or job or packet has unit requirement.

Work served from a queue can get routed to another queue in the network or leave the network. We shall assume deterministic, unicast and acyclic routing. Let $R = [R_{mn}]$ represent the $N \times N$ routing matrix with $R_{mn} = 1$ if work departing from queue m joins queue n and 0 otherwise due to deterministic routing assumption. Unicast routing implies $\sum_n R_{mn} \leq 1$ for all m. Define $\vec{R} = (I - R^{\mathrm{T}})^{-1}$. Acyclic routing assumption implies that $\vec{R}$ is well-defined as $(I - R^{\mathrm{T}})^{-1} = I + R^{\mathrm{T}} + (R^{\mathrm{T}})^2 + \cdots$. In this setup, all the components of $\vec{R}$ are $\{0, 1\}$. And $\vec{R}_{mn} = 1$ if work departing from queue n eventually goes through queue m.

We will use the convention that $\mathbf{Q}(\tau)$ is the vector of queue sizes at the beginning of timeslot τ, and then the schedule for timeslot τ is chosen and service happens, and then arrivals for timeslot τ happen. Thus, with $\boldsymbol{\Sigma}(\tau) = \sum_{\boldsymbol{\pi} \in \mathcal{S}} \boldsymbol{\pi} S_{\boldsymbol{\pi}}(\tau)$,

$$Q_n(\tau) = Q_n(0) + A_n(\tau) - \big(\Sigma_n(\tau) - Z_n(\tau)\big) + \sum_m R_{mn}\big(\Sigma_m(\tau) - Z_m(\tau)\big). \quad (2)$$

Equivalently,

$$\mathbf{Q}(\tau) = \mathbf{Q}(0) + \mathbf{A}(\tau) - \big(I - R^{\mathrm{T}}\big)\big[\boldsymbol{\Sigma}(\tau) - \mathbf{Z}(\tau)\big].$$

Here,

$$Z_n(\tau) - Z_n(\tau - 1) = \max\big(0, \Sigma_n(\tau) - \Sigma_n(\tau - 1) - Q_n(\tau - 1)\big). \quad (3)$$

1.2 Maximum weight scheduling

The operational problem of interest is to decide which queues to schedule for service among all possible allowable options. This is done by a scheduling policy. A class of myopic scheduling policies known as the Maximum Weight (MW) have been of interest since their introduction by Tassiulas and Ephremides [19]. The basic version of the policy works as follows. Define weight of a queue n in timeslot τ as $Q_n(\tau) - Q_m(\tau)$ if $R_{nm} = 1$ for some m and $Q_n(\tau)$ if $R_{nm} = 0$ for all m. The weight of a schedule $\boldsymbol{\pi} \in \mathcal{S}$ is the summation of the weights of queues that are served. Then the MW policy chooses for timeslot τ a schedule with the largest weight (breaking ties as per a predetermined fixed order of schedules in $\mathcal{S}$). That is, in timeslot τ, a schedule $\boldsymbol{\pi}$ is chosen so that

$$\begin{aligned}\boldsymbol{\pi} \cdot (I - R)\mathbf{Q}(\tau) &= \max_{\boldsymbol{\rho} \in \mathcal{S}} \boldsymbol{\rho} \cdot (I - R)\mathbf{Q}(\tau) \\ &= \max_{\boldsymbol{\rho} \in \mathcal{S}} \sum_n \rho_n \Big(Q_n(\tau) - \sum_m R_{nm} Q_m(\tau) \Big) \quad (4)\end{aligned}$$

with notation $\mathbf{a} \cdot \mathbf{b} = \sum_n a_n b_n$. The above policy is also referred to as the *back-pressure*, since the weight of queue n is determined by the difference of its own queue-size and the next-hop queue-size.

This policy can be naturally generalized to choose a schedule which maximizes $\boldsymbol{\pi} \cdot (I - R)\mathbf{Q}(\tau)^{\alpha}$, where the exponent is taken componentwise for some $\alpha > 0$; call this the MW-α policy. Or more generally, MW-f policy that chooses a schedule $\boldsymbol{\pi}$ in each timeslot τ so that

$$\boldsymbol{\pi} \cdot (I - R) f\big(\mathbf{Q}(\tau)\big) = \max_{\boldsymbol{\rho} \in \mathcal{S}} \boldsymbol{\rho} \cdot (I - R) f\big(\mathbf{Q}(\tau)\big) \tag{5}$$

for some function $f : \mathbb{R}_+ \to \mathbb{R}_+$; call this the MW-$f$ policy. Here and throughout, $f(\mathbf{Q}(\tau)) = [f(Q_n(\tau))]$.

In the literature (cf. [17]), the choice of weight function is restricted to the class of *scale-invariant* functions (see (10) for precise definition). In this paper, interest is in the family of MW-f policies where the weight function f is weighted logarithm of queue-size, which is not scale-invariant. Specifically, given a constant weight vector $\mathbf{w} = [w_n]$ with $w_n > 0$ for all n and constant $G \geq 1$, define weight of queue n at timeslot τ as

$$\mathsf{LOG}_n\big(Q_n(\tau)\big) = w_n \log\big(w_n Q_n(\tau) + G\big) - w_n \log G. \tag{6}$$

The policy of interest, called MWL chooses schedule $\boldsymbol{\pi}$ in timeslot τ so that

$$\boldsymbol{\pi} \cdot (I - R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big) = \max_{\boldsymbol{\rho} \in \mathcal{S}} \boldsymbol{\rho} \cdot (I - R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big) \tag{7}$$

with notation $\mathsf{LOG}(\mathbf{Q}(\tau)) = [\mathsf{LOG}_n(Q_n(\tau))]$. Again, ties are broken uniformly at random independent of everything else.

We shall consider a sequence of systems, indexed by r, to establish a fluid model as a formal approximation and subsequently study its properties. Specifically, we shall establish two main properties of the fluid model: one, work-conservation property (see Definition 2, Sect. 3), and two, characterization of invariant manifold under critical loading (see Sect. 4). The work-conservation property is establish in generality, i.e., for any value of G (including $G = 1$) in (6). However, to characterize the invariant manifold, we shall assume that the G in the rth system in (6), denoted by G_r, scales so that as $r \to \infty$, $G_r \to \infty$ (equivalently $G_r = \omega(1)$) but not too fast, i.e., $G_r / \log r \to 0$ (equivalently $G_r = o(\log r)$). The requirement of $G_r = o(\log r)$ is to make sure that the policy behavior in the fluid model is independent of the choice of such constant; the requirement of $G_r = \omega(1)$; however, in Sect. 4 is primarily technical and induced by the limitation of current proof method.

An implication of the *monotonicity* property of $\mathcal{S}$ is as follows: if $Q_n(\tau) = 0$ then its weight, $Q_n(\tau) - \sum_m R_{nm} Q_m(\tau) \leq 0$. Therefore, there is a schedule $\boldsymbol{\pi} \in \mathcal{S}$ with maximum weight so that $\pi_n = 0$. We shall assume that such a schedule is chosen. Therefore, under the MWL policy, we shall impose

$$Z_n(\tau) = 0, \quad \text{for all } \tau, n. \tag{8}$$

Equivalently, the resulting queueing dynamics under the MWL policy is

$$\begin{aligned} Q_n(\tau) &= Q_n(0) + A_n(\tau) - \Sigma_n(\tau) + \sum_m R_{mn} \Sigma_m(\tau) \\ \Sigma_n(\tau) - \Sigma_n(\tau - 1) &= 0 \quad \text{if } Q_n(\tau - 1) = 0. \end{aligned} \tag{9}$$

1.3 Notations

Bold letters will be reserved for vectors in $\mathbb{R}^N$. Let $\mathbf{0}$ be the vector of all 0s, and $\mathbf{1}$ be the vector of all 1s. Let $1_{\{\cdot\}}$ be the indicator function, $1_{\text{true}} = 1$ and $1_{\text{false}} = 0$. Let $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$ and $[x]^+ = x \vee 0$. For vectors $\mathbf{x}$, $\mathbf{y}$, notation $\mathbf{x} \wedge \mathbf{y}$, $\mathbf{x} \vee \mathbf{y}$ and $[\mathbf{x}]^+$ means application of operators $\wedge$, $\vee$ and $[\cdot]^+$ componentwise. Use $|\mathbf{x}| = \max_i |x_i|$ as a norm for vectors $\mathbf{x}$. Use notation $\mathbf{x}^{\max} = \max_i x_i$ and $\mathbf{x}^{\min} = \min_i x_i$. For vectors $\mathbf{u}$ and $\mathbf{v}$ and functions $f : \mathbb{R} \to \mathbb{R}$, let

$$\mathbf{u} \cdot \mathbf{v} = \sum_{n=1}^{N} u_n v_n, \quad \text{and} \quad f(\mathbf{u}) = \left[f(u_n)\right]_{1 \le n \le N}.$$

Let $\mathbb{N}$ be the set of natural numbers $\{1, 2, \ldots\}$, let $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, let $\mathbb{R}$ be the set of real numbers, and let $\mathbb{R}_+ = \{x \in \mathbb{R} : x \ge 0\}$.

1.4 Related work

The primary performance goals of a scheduling policy are stability and small queue-sizes. A queueing network is called *stable* if the underlying network Markov process is positive (Harris) recurrent as long as the network is underloaded. Thus, a stable policy leads to efficient utilization of network resources in the long term. The short term efficiency of the network is captured by small queue-sizes—to begin with, on average and more generally, with respect to higher moments.

In a single server system, like a $G/G/1$ queue, work-conservation property leads to optimal performance in terms of minimizing unserved work in the system in a strong, path-wise, sense. Such optimality properties make work-conservation property highly desirable. In a system where many queues are served by a single server, again work-conservation can be achieved as long as any queue can be scheduled to serve at any time. In such systems, work conservation property has a strong implication: as long as there is work in the system, it is served at the maximal possible rate.

Now in a generic constrained multiserver queueing network, such as that considered in this paper, a policy with such a work-conserving property is unlikely to exist. Therefore, over the past few decades, attempts have been made to search for policies that have asymptotic work-conservation property in multiserver, constrained queueing networks such as the stochastic processing networks.

As the first step toward this, Harrison [9] proposed a parametric policy called BIGSTEP to achieve work-conservation under heavy traffic or diffusion approximation by utilizing explicit knowledge of arrival process statistics like rate vector $\boldsymbol{\lambda}$. This establishes the existence of a policy that is indeed work-conserving with respect to the diffusion approximation. However, it is a parametric policy and requires explicit knowledge of the arrival process statistics. It would be more desirable to have such an asymptotic work-conserving policy that is myopic like the MW policy both from a designer's viewpoint and for the sake of elegance.

Toward this, work by Stolyar [18] and Lin and Dai [14] establish work-conservation property of myopic MW-1 policy for generalized switched network and stochastic processing network respectively with respect to heavy traffic approximation. Their results are applicable to scenarios when essentially one resource is critically loaded. This condition is known as the *complete resource pooling*. In a nutshell, these results say that when exactly one resource is critically loaded; the MW-1 policy organically identifies it and does not idle on it or serves it in a work-conserving manner. These results do not extend beyond the *complete resource pooling* condition or in the presence of multiple critically loaded resources.

To overcome this limitation of complete resource pooling, Shah and Wischik [17] studied the performance of MW-α policy for $\alpha > 0$, under heavy traffic or diffusion scaling. They find that under the MW-α policy, the critically loaded fluid model (which is used to establish state space collapse result) is approximate work-conserving (see (40) for precise definition) even in the presence of multiple critically loaded resources—the approximation error goes to 0 as $\alpha \to 0^+$ suggesting that the limiting policy, say MW-0^+, is work-conserving with respect to the critically loaded fluid model.

Few remarks are in order. First, the work-conservation property of critical fluid model is weaker than work-conservation of the heavy traffic approximation. However, it is an important step toward it. Second, the limiting policy MW-0^+ was conjectured by Shah and Wischik [17] to be work-conserving with respect to critical fluid model. However, analyzing it seems quite challenging. Finally, the logarithm weight policy, the MWL, considered in this paper is closely related to the conjectured MW-0^+ policy [17].

1.5 Contribution

As an important contribution of this work, we establish work-conservation property (see Definition 2) of the MWL policy with respect to the critically loaded fluid model. More generally, we establish that the work (total queue-size) is strictly decreasing under the fluid model as long as there is some work in the network and the network is underloaded.

We start by establishing stability property of switched network operating under the MWL policy. Next, we identify fluid model of the switched network operating under the MWL policy and establish it as the formal functional law of large numbers approximation. The work-conservation property shows up explicitly in this fluid model (Eq. (36) in Sect. 3.2).

In all the prior work, such as [1, 8, 14, 16–18], that establishes fluid model as a formal approximation of network operating under maximum weight scheduling policy, the choice of weight function f in the MW-f policy always satisfies the following assumption: f is differentiable and strictly increasing with $f(0) = 0$; for any $\mathbf{q} \in \mathbb{R}_+^N$ and $\boldsymbol{\pi} \in \mathcal{S}$, with $m(\mathbf{q}) = \max_{\boldsymbol{\rho} \in \mathcal{S}} \boldsymbol{\rho} \cdot f(\mathbf{q})$,

$$\boldsymbol{\pi} \cdot f(\mathbf{q}) = m(\mathbf{q}) \quad \Rightarrow \quad \boldsymbol{\pi} \cdot f(\kappa \mathbf{q}) = m(\kappa \mathbf{q}), \quad \text{for all } \kappa > 0. \tag{10}$$

This "scale invariance" property (10) has been essential in obtaining useful fluid model and subsequently to study heavy traffic scaled network. As a consequence of the scale invariance, the fluid analog of queue-size vector evolves under the same, maximum weight, policy. That is, the “fluid control” remains the same as “discrete control.” This makes analysis relatively easier.

The logarithm weight function as defined in (6) does not satisfy (10), and hence it is not clear what is the right “fluid control” that such policy will induce. An important contribution of this work is to overcome this challenge and obtain useful fluid model as a formal approximation.

Given work-conservation property of fluid model, it is natural to wonder the validity of work-conservation of MWL policy under the heavy traffic approximation. The method proposed by Bramson [6] and Williams [20] to establish heavy traffic approximation involves characterization of the invariant manifold of critical fluid model as the key initial step. In this paper, we identify invariant manifold for critically loaded fluid model as the solution space of a *two stage* optimization problem (see Sect. 4.3). It is worth taking note of the fact that the form of optimization problem is unusual compared to what is observed in literature, e.g., [13, 17].

We note that the characterization of invariant manifold is restricted to single-hop network and requires certain additional assumptions that are stated in Sect. 4.1. However, we believe that such characterization must hold more generally. Finally, inspired by the form of optimization problem arising in the characterization of invariant manifold, we conjecture that the MWL policy is work-conserving in heavy traffic approximation as long as appropriate weight vector, $\mathbf{w} > 0$, is used (Condition (71)).

1.6 An illustration of results: input-queued switch

Here, we illustrate the main results of this paper about MWL policy in the context of an instance of single-hop switched network, the input-queued switch. The input-queued switch architecture is commercially popular for performing task of switching packets in an internet router. Figure 1 illustrates an input-queued switch with three input ports and three output ports. Packets arriving at input i destined for output j are stored at input port i, in queue $Q_{i,j}$, thus there are $N = 9$ queues in total. (When we discuss the specific example of an input-queued switch, it is most natural to use double indexing, e.g., $Q_{3,2}$, whereas when we give general results about switched networks we will use single indexing, e.g., Q_n for $1 \le n \le N$.)

The switch operates in discrete time. In each timeslot, the switch fabric can transmit a number of packets from input ports to output ports, subject to the two constraints that each input can transmit at most one packet and that each output can receive at most one packet. In other words, in each timeslot, the switch can choose a *matching* from inputs to outputs as its schedule. The matching or schedule $\boldsymbol{\pi} \in \{0,1\}^{3\times 3}$ is given by $\pi_{i,j} = 1$ if input port i is matched to output port j in a given timeslot, and $\pi_{i,j} = 0$ otherwise. Thus,

$$\mathcal{S} = \left\{\boldsymbol{\pi} \in \{0,1\}^{3\times 3} : \sum_k \pi_{i,k} \le 1,\ \sum_k \pi_{k,j} \le 1,\ \forall 1 \le i, j \le 3\right\}.$$

It can be checked that an arrival rate vector $\boldsymbol{\lambda} \in [0,1]^{3\times 3}$ is supportable if

$$\sum_k \lambda_{i,k} \le 1, \qquad \sum_k \lambda_{k,j} \le 1, \quad \forall 1 \le i, j \le 3.$$

Define

$$\mathsf{L}(\boldsymbol{\lambda}) = \max_{1 \le i,j \le 3} \left(\sum_{k=1}^{3} \lambda_{i,k}, \sum_{k=1}^{3} \lambda_{k,j} \right).$$

Now consider a switch operating under the MWL policy with $\mathbf{w} = \mathbf{1}$, the vector of all 1s. The work-conservation property that is established in Sect. 3 implies[1] that

$$\sum_{i,j=1}^{3} \frac{dq_{i,j}(t)}{dt} \le -\frac{1 - \mathsf{L}(\boldsymbol{\lambda})}{3}, \quad \text{if } \sum_{i,j=1}^{3} q_{i,j}(t) > 0.$$

In above, $q_{i,j}(\cdot)$ represents fluid model analog for queue (i, j). It says that the summation of the queue-sizes, the total work, is nonincreasing and it decreases at a rate that depends on the loading. Even when switch is critically loaded, i.e., $\mathsf{L}(\boldsymbol{\lambda}) = 1$, the net queue-size is nonincreasing. And if all ports are critically loaded, i.e., $\sum_k \lambda_{i,k} = 1$ for all $1 \le i \le 3$, then net queue-size cannot decrease either. That is, in such a setting, $\sum_{i,j} q_{i,j}(\cdot)$ remains unchanged.

In this critically loaded setup, that is $\boldsymbol{\lambda}_{i,\cdot} = \boldsymbol{\lambda}_{\cdot,j} = 1$ for all $1 \le i, j \le 3$, where $\boldsymbol{\lambda}_{i,\cdot} = \sum_k \lambda_{i,k}$, $\boldsymbol{\lambda}_{\cdot,j} = \sum_k \lambda_{k,j}$, a lot more can be said about the invariant states of the fluid model. To begin with, such a critically loaded switch corresponds to having 6 servers critically loaded: each input-port i, $1 \le i \le 3$ (similarly output-port j, $1 \le j \le 3$), receives data at net-rate $\boldsymbol{\lambda}_{i,\cdot}$ and can serve at most at rate 1. Each of these six servers, three input-ports, and three output-ports, are *virtual resources* of the switch. And they are all critically loaded under the above described setting. For generic switched network, virtual resources and critically loading is defined in Sect. 4.2.

Under the above setting, Theorem 3 implies that $\mathbf{q} \in \mathbb{R}_+^{3\times 3}$ is a fixed or invariant state of the fluid model under critical loading if it solves the following optimization problem:

$$\begin{aligned}
&\text{minimize} && \left(\sum_{i,j} y_{i,j}, \sum_{i,j} y_{i,j} \log y_{i,j} \right) \\
&\text{over} && \mathbf{y} \in \mathbb{R}_+^N \\
&\text{such that} && \sum_{k=1}^{3} y_{i,k} \ge \sum_{k=1}^{3} q_{i,k}, \qquad \sum_{k=1}^{3} y_{k,j} \ge \sum_{k=1}^{3} q_{k,j}, \quad \forall 1 \le i, j \le 3.
\end{aligned}$$

In the above optimization problem, the minimization is with respect to the lexicographic ordering of two objectives: first minimize $\sum_{i,j} y_{i,j}$ and then minimize $\sum_{i,j} y_{i,j} \log y_{i,j}$. To understand solution of this optimization problem, consider $\mathbf{q}$ such that $\mathbf{q}_{i,\cdot}, \mathbf{q}_{\cdot,j} > 0$ for all $1 \le i, j \le 3$. Then the minimization of the first objective suggests that the solution of optimization problem, say $\mathbf{q}^*$, must preserve the "row-sums" and "column-sums," i.e., $\sum_{k=1}^{3} q^*_{i,k} = \sum_{k=1}^{3} q_{i,k}$ and $\sum_{k=1}^{3} q^*_{k,j} = \sum_{k=1}^{3} q_{k,j}$

[1] Tighter analysis for the specific instance of input-queued switch leads to denominator 3 in place of 81 as per the general result.

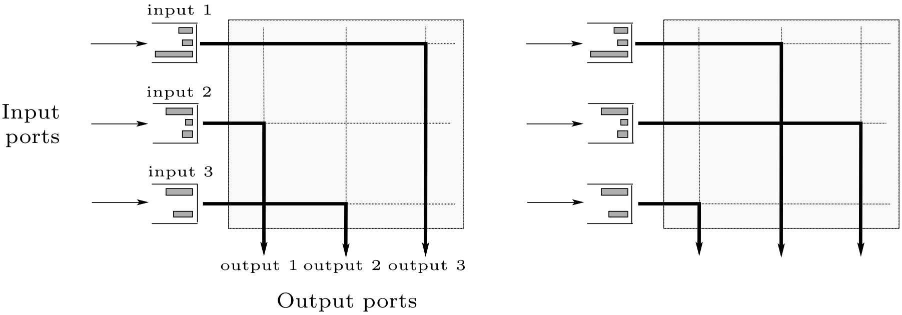

Fig. 1 An input-queued switch, and two example matching of inputs to outputs

for all $1 \le i, j \le 3$. Subject to this, the minimization of strictly convex objective $\sum_{i,j} y_{i,j} \log y_{i,j}$ leads to the form of $\mathbf{q}^*$ such that

$$q^*_{i,j} = \frac{\mathbf{q}_{i,\cdot}\mathbf{q}_{\cdot,j}}{\mathbf{q}_{\cdot,\cdot}}, \quad \text{for all } 1 \le i, j \le 3, \tag{11}$$

where $\mathbf{q}_{i,\cdot} = \sum_{k=1}^3 q_{i,k}$, $\mathbf{q}_{\cdot,j} = \sum_{k=1}^3 q_{k,j}$ and $\mathbf{q}_{\cdot,\cdot} = \sum_{i,j} q_{i,j}$. That is, effectively the fixed or invariant point has extremely simple rank-1 matrix structure subject to the work-conservation property. Such form of invariant or fixed point is highly desirable to establish heavy traffic optimality property of scheduling policy as discussed in work [17].

1.7 Organization

Section 2 establishes positive recurrence of the network operating under the MWL policy when it is underloaded. Section 3 presents fluid model and establishes it as formal functional law of large numbers approximation. The fluid model is established to possess work-conservation property when it is underloaded and critically loaded. Section 4 studies properties of critically loaded fluid model. Specifically, the invariant manifold under critically loaded fluid model is characterized. This characterization suggests the form of state-space collapse. The results of Sect. 4 apply to single-hop network while results of Sects. 2 and 3 apply in the general form.

2 Stability

This section establishes stability property of the network operating under the MWL policy, i.e., the network Markov process is positive recurrent as long as the network is underloaded. We start by formally defining notion of load of the system. This will allow one to formally distinguish underloaded and critically loaded scenarios.

2.1 Admissible arrival rates

In each timeslot, a schedule $\boldsymbol{\pi} \in \mathcal{S}$ must be chosen. Let Σ be the convex hull of $\mathcal{S}$,

$$\Sigma = \left\{ \sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} \boldsymbol{\pi} : \sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} = 1, \text{ and } \alpha_{\boldsymbol{\pi}} \geq 0 \text{ for all } \boldsymbol{\pi} \right\}. \tag{12}$$

Given an arrival rate vector $\boldsymbol{\lambda}$, the effective load induced on queues is given by $\vec{\boldsymbol{\lambda}} = \vec{R}\boldsymbol{\lambda}$. We say that an arrival rate vector $\boldsymbol{\lambda}$ is *admissible* if $\vec{\boldsymbol{\lambda}} \in \Lambda$ where

$$\Lambda = \left\{ \boldsymbol{\mu} \in \mathbb{R}_+^N : \boldsymbol{\mu} \leq \boldsymbol{\sigma} \text{ componentwise, for some } \boldsymbol{\sigma} \in \Sigma \right\}. \tag{13}$$

Intuitively, this means that there is some combination of feasible schedules which permits all incoming work to be served. Also define

$$\Lambda^\circ = \left\{ \boldsymbol{\rho} \in \Lambda : \boldsymbol{\rho} \leq \sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} \boldsymbol{\pi}, \text{ where } \sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} < 1 \text{ and } \alpha_{\boldsymbol{\pi}} \geq 0 \text{ for all } \boldsymbol{\pi} \right\},$$

$$\partial\Lambda = \Lambda \setminus \Lambda^\circ.$$

Say that $\boldsymbol{\lambda}$ is *strictly admissible* or the system is *underloaded* if $\vec{\boldsymbol{\lambda}} \in \Lambda^\circ$, and that $\boldsymbol{\lambda}$ is *critical* or the system is *critically loaded* if $\vec{\boldsymbol{\lambda}} \in \partial\Lambda$. A useful corresponding optimization problem is PRIMAL($\vec{\boldsymbol{\lambda}}$):

$$\begin{array}{lll} \text{minimize} & \displaystyle\sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} & \\ \text{over} & \alpha_{\boldsymbol{\pi}} \in \mathbb{R}_+ \quad \text{for all } \boldsymbol{\pi} \in \mathcal{S} & \\ \text{such that} & \vec{\boldsymbol{\lambda}} \leq \displaystyle\sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} \boldsymbol{\pi} \quad \text{componentwise} & \end{array}$$

This problem asks whether it is possible to find a combination of schedules which can serve arrival rates $\boldsymbol{\lambda}$; clearly $\boldsymbol{\lambda}$ is admissible if and only if the solution to the problem is ≤ 1; it is strictly admissible if the solution to the problem is < 1 and it is critically loaded if $= 1$. We shall define the load of $\boldsymbol{\lambda}$, denoted by $\mathsf{L}(\boldsymbol{\lambda})$, as the solution of the optimization problem PRIMAL($\vec{\boldsymbol{\lambda}}$).

2.2 Positive recurrence

We establish the stability property of the MWL policy.

Theorem 1 *Consider a switched network operating under the* MWL *policy with* $\mathbf{w} > 0$, $G \geq 1$. *For any strictly admissible* $\boldsymbol{\lambda}$, *i.e.*, $\mathsf{L}(\boldsymbol{\lambda}) < 1$, $\mathbf{Q}(\cdot)$ *is a positive recurrent Markov chain.*

Proof First observe that $\mathbf{Q}(\cdot)$ is a discrete time, countable state space Markov chain under the MWL policy. This is because of the Bernoulli arrival process, the myopic nature of the MWL policy and $\mathcal{S} \subset \{0,1\}^N$. The monotonicity property of $\mathcal{S}$ and Bernoulli nature of arrival process makes $\mathbf{Q}(\cdot)$ irreducible; aperiodicity follows because there is a positive probability of remaining at state $\mathbf{0}$. By Lyapunov and Foster's criteria, to establish positive recurrence of $\mathbf{Q}(\cdot)$, it is sufficient to find an appropriate Lyapunov function with negative drift, e.g., see [15]. Consider a candidate Lyapunov function

$$L(\mathbf{Q}) = \sum_n (w_n Q_n + G)\log(w_n Q_n + G) - w_n Q_n \log G - (w_n Q_n + G). \quad (14)$$

The choice of $L(\cdot)$ is made so that

$$\frac{\partial L(\mathbf{Q})}{\partial Q_n} = w_n \log(w_n Q_n + G) - w_n \log G = \mathsf{LOG}_n(Q_n).$$

Subsequently, $L(\cdot)$ is a strictly convex function over $\mathbb{R}_+^N$. Therefore,

$$\begin{aligned} L\big(\mathbf{Q}(\tau+1)\big) - L\big(\mathbf{Q}(\tau)\big) &\le \Delta L\big(\mathbf{Q}(\tau+1)\big)\cdot\big(\mathbf{Q}(\tau+1)-\mathbf{Q}(\tau)\big) \\ &= \sum_n \mathsf{LOG}_n\big(Q_n(\tau+1)\big)\big(Q_n(\tau+1)-Q_n(\tau)\big). \quad (15) \end{aligned}$$

Now $|Q_n(\tau+1) - Q_n(\tau)| \le 1$ since arrival process is Bernoulli and $\mathcal{S} \subset \{0,1\}^N$. LOG_n is a concave strictly increasing function and $\frac{d\mathsf{LOG}_n(x)}{dx} \le w_n/G \le \mathbf{w}^{\max}/G$ for all $x \ge 0$, where recall $\mathbf{w}^{\max} = \max_n w_n$. Therefore, it follows that

$$-\mathbf{w}^{\max}/G \le \mathsf{LOG}_n\big(Q_n(\tau+1)\big) - \mathsf{LOG}_n\big(Q_n(\tau)\big) \le \mathbf{w}^{\max}/G. \quad (16)$$

Using (16) in (15),

$$\begin{aligned} &L\big(\mathbf{Q}(\tau+1)\big) - L\big(\mathbf{Q}(\tau)\big) \\ &\quad \le N\mathbf{w}^{\max}/G + \sum_n \mathsf{LOG}_n\big(Q_n(\tau)\big)\big(Q_n(\tau+1)-Q_n(\tau)\big). \quad (17) \end{aligned}$$

Using (9) in (17) and using $\boldsymbol{\sigma}(\tau) = \boldsymbol{\Sigma}(\tau+1) - \boldsymbol{\Sigma}(\tau)$,

$$\begin{aligned} &\mathbb{E}\big[L\big(\mathbf{Q}(\tau+1)\big) - L\big(\mathbf{Q}(\tau)\big)\big|\mathbf{Q}(\tau)\big] \\ &\quad \le N\mathbf{w}^{\max}/G + \sum_n \mathsf{LOG}_n\big(Q_n(\tau)\big)\mathbb{E}\big[Q_n(\tau+1)-Q_n(\tau)\big|\mathbf{Q}(\tau)\big] \\ &\quad = N\mathbf{w}^{\max}/G + \sum_n \mathsf{LOG}_n\big(Q_n(\tau)\big)\bigg(\lambda_n - \sigma_n(\tau) + \sum_m R_{mn}\sigma_m(\tau)\bigg) \\ &\quad = N\mathbf{w}^{\max}/G + \sum_n \mathsf{LOG}_n\big(Q_n(\tau)\big)\lambda_n \\ &\qquad - \sum_n \sigma_n(\tau)\bigg[\mathsf{LOG}_n\big(Q_n(\tau)\big) - \sum_m R_{nm}\mathsf{LOG}_m(Q_m(\tau))\bigg] \\ &\quad = N\mathbf{w}^{\max}/G + \mathsf{LOG}\big(\mathbf{Q}(\tau)\big)\cdot\boldsymbol{\lambda} - \boldsymbol{\sigma}(\tau)\cdot(I-R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big). \quad (18) \end{aligned}$$

Since $\mathsf{L}(\boldsymbol{\lambda}) < 1$, i.e., value of optimization problem PRIMAL$(\vec{\boldsymbol{\lambda}})$ is < 1, and using *monotonicity* of $\mathcal{S}$, we have

$$\vec{\boldsymbol{\lambda}} = \sum_{\boldsymbol{\pi} \in \mathcal{S}} \alpha_{\boldsymbol{\pi}} \boldsymbol{\pi} \quad \text{so that} \quad \alpha_{\boldsymbol{\pi}} \geq 0, \ \sum_{\boldsymbol{\pi}} \alpha_{\boldsymbol{\pi}} = \mathsf{L}(\boldsymbol{\lambda}) < 1.$$

That is,

$$\boldsymbol{\lambda} = \sum_{\boldsymbol{\pi}} \alpha_{\boldsymbol{\pi}} \big(I - R^{\mathrm{T}}\big) \boldsymbol{\pi}.$$

Therefore, we obtain

$$\begin{aligned}
&\mathbb{E}\big[L\big(\mathbf{Q}(\tau+1)\big) - L\big(\mathbf{Q}(\tau)\big) \big| \mathbf{Q}(\tau)\big] \\
&\quad \leq N\mathbf{w}^{\max}/G + \sum_{\boldsymbol{\pi}} \alpha_{\boldsymbol{\pi}} \boldsymbol{\pi} \cdot (I - R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big) \\
&\qquad - \boldsymbol{\sigma}(\tau) \cdot (I - R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big) \\
&\quad \leq N\mathbf{w}^{\max}/G + \Big(\sum_{\boldsymbol{\pi}} \alpha_{\boldsymbol{\pi}} - 1\Big) \boldsymbol{\sigma}(\tau) \cdot (I - R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big),
\end{aligned} \tag{19}$$

where the last inequality follows since MWL policy chooses $\boldsymbol{\sigma}(\tau)$ that maximizes $\boldsymbol{\rho} \cdot (I - R)\mathsf{LOG}(\mathbf{Q}(\tau))$ over all $\boldsymbol{\rho} \in \mathcal{S}$. To complete the proof, we claim that

$$\begin{aligned}
\boldsymbol{\sigma}(\tau) \cdot (I - R)\mathsf{LOG}\big(\mathbf{Q}(\tau)\big) &\geq \frac{1}{N} \max_n \mathsf{LOG}_n\big(Q_n(\tau)\big) \\
&\geq \frac{1}{N^2} \sum_n \mathsf{LOG}_n\big(Q_n(\tau)\big).
\end{aligned} \tag{20}$$

To see this, observe that by monotonicity property of $\mathcal{S}$, schedule $\mathbf{e}^n$ that only schedules nth queue belongs to $\mathcal{S}$. Consider a queue, say m, such that $Q_m(\tau) = \max_n Q_n(\tau)$. Either $R_{mn} = 0$ for all n or there exists $1 \leq l \leq N - 1$ and $m_1, \dots, m_l$ so that $R_{mm_1} = R_{m_1m_2} = \cdots = R_{m_{l-1}m_l} = 1$ and $R_{m_l n} = 0$ for all n. In the former case, schedule $\mathbf{e}^m$ has weight $\mathsf{LOG}_m(Q_m(\tau))$ and hence the weight of schedule with maximum weight must be at least $\max_n \mathsf{LOG}_m(Q_n(\tau))/N$. In the latter case, consider schedules $\mathbf{e}^m, \mathbf{e}^{m_1}, \dots, \mathbf{e}^{m_l}$. Summation of their weights is $\mathsf{LOG}_m(Q_m(\tau))$. Therefore, the average weight of these, at most N, schedules is at least $\max_n \mathsf{LOG}_n(Q_n(\tau))/N$. And it provides a lower bound on the weight of maximum weighted schedule. To complete the proof, from (19) and (20), we obtain

$$\begin{aligned}
&\mathbb{E}\big[L\big(\mathbf{Q}(\tau+1)\big) - L\big(\mathbf{Q}(\tau)\big) \big| \mathbf{Q}(\tau)\big] \\
&\quad \leq \frac{N\mathbf{w}^{\max}}{G} - \frac{(1 - \mathsf{L}(\boldsymbol{\lambda}))}{N^2} \Big(\sum_n \mathsf{LOG}_n\big(Q_n(\tau)\big)\Big).
\end{aligned} \tag{21}$$

Therefore, for $(\sum_n \mathsf{LOG}_n(Q_n(\tau))) \geq \frac{2N^3\mathbf{w}^{\max}}{G(1-\mathsf{L}(\boldsymbol{\lambda}))}$, the RHS of (21) is at most $-N\mathbf{w}^{\max}/G$. That is, Lyapunov function $L(\cdot)$ has strictly negative drift outside set $\mathcal{B}$ defined as

$$\mathcal{B} = \Big\{\mathbf{Q} = [Q_n] : \sum_n \mathsf{LOG}_n(Q_n) \leq \frac{2N^3\mathbf{w}^{\max}}{G(1 - \mathsf{L}(\boldsymbol{\lambda}))}\Big\}.$$

The set $\mathcal{B}$ is clearly bounded since $L(\mathbf{Q}) \to \infty$ as $\|\mathbf{Q}\| \to \infty$. And recall that $\mathbf{Q}(\cdot)$ is irreducible, aperiodic discrete time Markov chain with countable state space. Therefore, we have shown that $\mathbf{Q}(\cdot)$ along with $L(\cdot)$ satisfies Lyapunov–Foster's criterion for establishing positive recurrence. In summary, we have established that the network Markov chain $\mathbf{Q}(\cdot)$ is indeed positive recurrent when $\mathsf{L}(\boldsymbol{\lambda}) < 1$. □

2.3 Concentration of $L(\cdot)$

We state an exponential tail bound on the deviation of the Lyapunov function $L(\cdot)$. This will be useful in establishing fluid model as an approximation of the associated stochastic performance processes as well as in characterizing invariant manifold for critically loaded fluid model.

Lemma 1 *Given $\tau \geq 0$ and $\mathbf{Q}(\tau)$, consider any $\tilde{\tau} > \tau$. Then the following holds with probability at least $1-\delta$:*

$$\begin{aligned} &L\big(\mathbf{Q}(\tilde{\tau})\big) - L\big(\mathbf{Q}(\tau)\big) \\ &\quad \leq \frac{N\mathbf{w}^{\max}(\tilde{\tau}-\tau)}{G} - \frac{(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2}\sum_{s=\tau}^{\tilde{\tau}-1}\Big(\sum_n \mathsf{LOG}_n\big(Q_n(s)\big)\Big) \\ &\qquad + 2N\mathbf{w}^{\max}\left(\frac{1}{G} + \log\left(\frac{\mathbf{w}^{\max}(\mathbf{Q}(\tau)\cdot\mathbf{1}+\tilde{\tau}-\tau+1)}{G}+1\right)\right) \\ &\qquad \times \sqrt{2(\tilde{\tau}-\tau)\log\left(\frac{1}{\delta}\right)}. \end{aligned} \tag{22}$$

Proof Throughout, assume that $\mathbf{Q}(\tau)$ is given and fixed. Define

$$\begin{aligned} X(0) &= L\big(\mathbf{Q}(\tau)\big) \\ X(s) &= L\big(\mathbf{Q}(\tau+s)\big) - \frac{N\mathbf{w}^{\max}s}{G} + \frac{(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2}\sum_{\ell=0}^{s-1}\Big(\sum_n \mathsf{LOG}_n\big(Q_n(\tau+\ell)\big)\Big), \\ &\quad \text{for } s \geq 1. \end{aligned} \tag{23}$$

Let $\mathcal{F}_s$, $s \geq 0$ be the smallest sigma algebra containing information about $(\mathbf{Q}(\tau), \mathbf{Q}(\tau+1), \ldots, \mathbf{Q}(\tau+s))$. Then $X(s), s \geq 0$ is measurable with respect to $\mathcal{F}_s$. From (21), it follows that $X(s)$ is a super-Martingale. This is because for $s \geq 0$,

$$\begin{aligned} \mathbb{E}\big[X(s+1)-X(s)\big|\mathcal{F}_s\big] &= \mathbb{E}\big[L\big(\mathbf{Q}(\tau+s+1)\big)-L\big(\mathbf{Q}(\tau+s)\big)\big|\mathcal{F}_s\big] - \frac{N\mathbf{w}^{\max}}{G} \\ &\quad + \frac{(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2}\Big(\sum_n \mathsf{LOG}_n\big(Q_n(\tau+s)\big)\Big) \\ &\leq 0, \quad \text{using (21)}. \end{aligned} \tag{24}$$

Now since at most one arrival happens per timeslot per queue,

$$Q_n(\tau+s) \leq Q_n(\tau)+s, \quad \text{for all } n. \tag{25}$$

And, hence

$$\mathsf{LOG}_n\big(Q_n(\tau+s)\big) \le w_n \log\left(\frac{w_n(Q_n(\tau)+s)}{G}+1\right).$$

Therefore, using arguments similar to those in (15)–(17), it follows that

$$\begin{aligned}&\big|L\big(\mathbf{Q}(\tau+s+1)\big)-L\big(\mathbf{Q}(\tau+s)\big)\big|\\&\quad\le \frac{N\mathbf{w}^{\max}}{G}+N\mathbf{w}^{\max}\log\left(\frac{\mathbf{w}^{\max}(\mathbf{Q}(\tau)\cdot\mathbf{1}+s+1)}{G}+1\right).\end{aligned}\tag{26}$$

And, therefore, for any $0 \le s \le \tilde{\tau}-\tau$,

$$\begin{aligned}\big|X(s+1)-X(s)\big| &\le \frac{2N\mathbf{w}^{\max}}{G}+2N\mathbf{w}^{\max}\log\left(\frac{\mathbf{w}^{\max}(\mathbf{Q}(\tau)\cdot\mathbf{1}+s+1)}{G}+1\right)\\&\le \frac{2N\mathbf{w}^{\max}}{G}+2N\mathbf{w}^{\max}\log\left(\frac{\mathbf{w}^{\max}(\mathbf{Q}(\tau)\cdot\mathbf{1}+\tilde{\tau}-\tau+1)}{G}+1\right)\\&\overset{\triangle}{=} \Delta(\tau,\tilde{\tau}).\end{aligned}\tag{27}$$

We recall inequality for super-Martingales with bounded increments by Azuma [2] and Hoeffding [12]. □

Proposition 1 *Let $\{Y_k\}_{k\ge 0}$ be super-Martingale. Let $\{C_k\}_{k\ge 0}$ be nonnegative constants so that with probability* 1, $|Y_{k+1}-Y_k| \le C_k$ *for* $k \ge 0$. *Then for any* $\gamma > 0$ *and* $m \ge 1$,

$$\mathbb{P}(Y_m - Y_0 > \gamma) \le \exp\left(-\frac{\gamma^2}{2\sum_{k=0}^{m-1} C_k^2}\right).$$

From Proposition 1, (24) and (27) it follows that for any $\gamma > 0$,

$$\mathbb{P}\big(X(\tilde{\tau}-\tau)-X(0)\ge\gamma\big) \le \exp\left(-\frac{\gamma^2}{2(\tilde{\tau}-\tau)\Delta(\tau,\tilde{\tau})^2}\right).\tag{28}$$

That is, for any $\delta \in (0,1)$,

$$\mathbb{P}\Big(X(\tilde{\tau}-\tau)-X(0)\ge \Delta(\tau,\tilde{\tau})\sqrt{2\log(1/\delta)(\tilde{\tau}-\tau)}\Big) \le \delta.\tag{29}$$

That is, with probability at least $1-\delta$,

$$\begin{aligned}L\big(\mathbf{Q}(\tilde{\tau})\big)-L\big(\mathbf{Q}(\tau)\big) &\le \frac{N\mathbf{w}^{\max}(\tilde{\tau}-\tau)}{G}-\frac{(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2}\sum_{s=\tau}^{\tilde{\tau}-1}\left(\sum_n \mathsf{LOG}_n\big(Q_n(s)\big)\right)\\&\quad+\Delta(\tau,\tilde{\tau})\sqrt{2\log(1/\delta)(\tilde{\tau}-\tau)}.\end{aligned}$$

This completes the proof of Lemma 1.

3 Fluid model

This section introduces fluid model for switched network operating under the MWL policy with $\boldsymbol{\lambda} \in \Lambda$. The fluid model is established as the formal functional law of large numbers approximation of the network. An important feature of the fluid model is that it is work conserving for any $\boldsymbol{\lambda} \in \Lambda$ (specifically see (36)).

3.1 Technical preliminaries

We begin with necessary technicalities. For a given fixed $T > 0$, let $C(T)$ be the set of continuous functions $[0, T] \to \mathbb{R}^I$ for some $I \in \mathbb{N}$, where $\mathbb{R}^I$ is equipped with the norm $|x| = \max_i |x_i|$. Here, we want $I = 3N + |\mathcal{S}|$. Equip $C(T)$ with the norm

$$\|x\| = \sup_{t\in[0,T]} \bigl|x(t)\bigr|.$$

Let $d(\cdot, \cdot)$ be the metric induced by this norm on $C(T)$, i.e.,

$$d(x, y) = \|x - y\| \quad \text{for all } x, y \in C(T).$$

For $E \subset C(T)$ and $x \in C(T)$, define

$$d(x, E) = \inf\bigl\{d(x, y) : y \in E\bigr\}.$$

Define the *modulus of continuity* by

$$\mathrm{mc}_\delta(x) = \sup_{|s-t|<\delta} \bigl|x(s) - x(t)\bigr|$$

where $s, t \in [0, T]$. Since $[0, T]$ is compact, each $x \in C(T)$ is uniformly continuous, therefore, $\mathrm{mc}_\delta(x) \to 0$ as $\delta \to 0$.

3.1.1 Cluster points

We shall build on the methodology of Bramson [6] that utilizes notion of *cluster points* that we shall introduce next. In what follows, we shall use C in place of $C(T)$ when T is clear from the context. Here, interest is in convergence in space $(C, \|\cdot\|)$ endowed by metric $\|\cdot\|$. The appropriate concept is *cluster points*. Consider any metric space E with metric d and a sequence $(E_1, E_2, \ldots)$ of subsets of E. Say that $x \in E$ is a *cluster point* of the sequence if $\liminf_{r\to\infty} d(x, E_r) = 0$ where $d(x, E_r) = \inf\{d(x, y) : y \in E_r\}$.

Proposition 1 (Cluster points in C)[2] *Given $K > 0$, $A > 0$ and a sequence $B_r \to 0$, let*

$$K_r = \bigl\{x \in C : \bigl|x(0)\bigr| \le K \textit{ and } \mathrm{mc}_\delta(x) \le A\delta + B_r \textit{ for all } \delta > 0\bigr\}$$

and consider a sequence $(E_1, E_2, \ldots)$ of subsets of C for which $E_r \subset K_r$. Then $\sup_{y\in E_r} d(y, \mathrm{CP}) \to 0$ as $r \to \infty$, where CP is the set of cluster points of $(E_1, E_2, \ldots)$.

3.2 Fluid model solution

We introduce the fluid model for network operating under MWL policy. Let time be measured by $t \in \mathbb{R}_+$. Let $\mathbf{q}$, $\mathbf{a}$ and $\mathbf{z}$ all be functions $\mathbb{R}_+ \to \mathbb{R}^N_+$, and let $s = (s_\pi)_{\pi\in\mathcal{S}}$

[2] Taken from Bramson [6, Proposition 4.1].

be a collection of functions $\mathbb{R}_+ \to \mathbb{R}_+$. The vector of queue sizes at time t is $\mathbf{q}(t)$, the cumulative arrivals up to time t is $\mathbf{a}(t)$, the cumulative idleness up to time t is $\mathbf{z}(t)$, and $s_{\boldsymbol{\pi}}(t)$ is the total amount of time spent on schedule $\boldsymbol{\pi}$ up to time t.

We say that the process $x(\cdot) = (\mathbf{q}(\cdot), \mathbf{a}(\cdot), \mathbf{z}(\cdot), s(\cdot))$ satisfies the fluid model for the MWL scheduling policy if

$$\mathbf{a}(t) = \boldsymbol{\lambda} t \tag{30}$$

$$\mathbf{q}(t) = \mathbf{q}(0) + \mathbf{a}(t) - \big(I - R^{\mathrm{T}}\big) \sum_{\pi} s_{\boldsymbol{\pi}}(t) \boldsymbol{\pi} + \mathbf{z}(t) \tag{31}$$

$$\sum_{\pi \in \mathcal{S}} s_{\boldsymbol{\pi}}(t) = t \tag{32}$$

each $s_{\boldsymbol{\pi}}(\cdot)$ and $z_n(\cdot)$ is increasing (not necessarily strictly increasing) (33)

all the components of $x(\cdot)$ are absolutely continuous (34)

for almost all t, all n, $\quad \dot{z}_n(t) = 0$ if $q_n(t) > 0$ (35)

And the additional fluid model equations that capture the work-conservation property of MWL policy are (with definition $\mathbf{w}^{\min} = \min_n w_n$)

for almost all t, (36)

$$\text{if } \mathbf{q}(t) \neq \mathbf{0} \text{ then } \sum_n \dot{q}_n(t) w_n \leq -\frac{(1 - \mathsf{L}(\boldsymbol{\lambda}))\mathbf{w}^{\min}}{N^2}$$

for all t, $\quad \mathbf{z}(t) = \mathbf{0}$ (37)

We define notion of fluid model solution.

Definition 1 (Fluid Model Solution) Given $T > 0$, we call $x(\cdot) \in C(T)$ a fluid model solution for a switched network operating under MWL policy if $x(\cdot)$ satisfies Eqs. (30)–(37).

We shall use notation FMS to denote the set of all fluid model solutions for a switched network operating under the MWL policy.

Definition 2 (Work-Conservation) Given vector $\mathbf{w} \in \mathbb{R}_+^N$ with $\mathbf{w} > 0$ component-wise, a policy is called $\mathbf{w}$ work-conserving if the corresponding fluid model satisfies the following: there exists constant $C = C(\mathbf{w}, N, \mathcal{S}) > 0$ such that for any $\boldsymbol{\lambda} \in \Lambda$ and starting state $\mathbf{q}_0$, all fluid model solutions $\mathbf{q}(\cdot)$ satisfy

$$\sum_n w_n \frac{dq_n(t)}{dt} \leq -C\big(1 - \mathsf{L}(\boldsymbol{\lambda})\big), \quad \text{for almost all } t > 0, \tag{38}$$

as long as $\mathbf{q}(t) \neq \mathbf{0}$.

In the above definition, by corresponding fluid model, we mean one that can be established to be a formal approximation. As per the above definition of work-conservation, when $\boldsymbol{\lambda}$ is critical, i.e., $\mathsf{L}(\boldsymbol{\lambda}) = 1$, we have that for any $t > 0$,

$$\sum_n w_n q_n(t) \leq \sum_n w_n q_n(0). \tag{39}$$

Given $\beta \geq 1$, we call a policy β-approximate $\mathbf{w}$ work-conserving if for any $t > 0$

$$\sum_n w_n q_n(t) \leq \beta \left(\sum_n w_n q_n(0) \right), \tag{40}$$

under any critical loading, i.e., $\mathsf{L}(\boldsymbol{\lambda}) = 1$.

3.3 Fluid model as a formal approximation

We state how the fluid model formally approximates the original system here. To that end, we introduce a fluid model scaling followed by statement of the main result.

3.3.1 Fluid scaling

Consider a sequence of systems of the type described in Sect. 1.1, indexed by $r \in \mathbb{N}$. Write $X^r(\tau) = (\mathbf{Q}^r(\tau), \mathbf{A}^r(\tau), \mathbf{Z}^r(\tau), S^r(\tau))$, $\tau \in \mathbb{Z}_+$, for the rth system. The scaling parameter will be denoted by r. Define the scaled system $x^r(t) = (\mathbf{q}^r(t), \mathbf{a}^r(t), \mathbf{z}^r(t), s^r(t))$ for $t \in \mathbb{R}_+$ by

$$\begin{aligned} \mathbf{q}^r(t) &= \mathbf{Q}^r(rt)/r, & \mathbf{a}^r(t) &= \mathbf{A}^r(rt)/r, \\ \mathbf{z}^r(t) &= \mathbf{Z}^r(rt)/r, & s^r_{\boldsymbol{\pi}}(t) &= S^r_{\boldsymbol{\pi}}(rt)/r \end{aligned}$$

after extending the domain of $X^r(\cdot)$ to $\mathbb{R}_+$ by linear interpolation in each interval $(\tau - 1, \tau)$. Then each sample-path of the fluid-scaled systems $x^r(\cdot)$ over $[0, T]$ must lie in $C(T)$ with $I = 3N + |\mathcal{S}|$.

3.3.2 Main result

We now state the result formally for the multihop switched network.

Assumptions Our goal is to study the dynamics of $x^r(t)$, for t in a fixed interval $[0, T]$, as $r \to \infty$. We will assume that for every r the arrival process $\mathbf{A}^r(\cdot)$ is Bernoulli with rate vector $\boldsymbol{\lambda}^r$. Further,

$$\lim_{r\to\infty} \boldsymbol{\lambda}^r = \boldsymbol{\lambda} \quad \text{for some } \boldsymbol{\lambda} \in \Lambda. \tag{41}$$

We assume that the initial queue sizes are uniformly bounded (need not be non-random/deterministic). That is, for all r

$$\left| \mathbf{q}^r(0) \right| \leq K \quad \text{for some } K \in \mathbb{R}_+. \tag{42}$$

The rth system operates under scheduling policy MWL that utilizes weight function LOG_n for queue n, where

$$\mathsf{LOG}_n(y) = w_n \log(w_n y + G_r) - w_n \log G_r. \tag{43}$$

The constant $G_r \geq 1$ and in general it may depend on r. However, we shall always assume that $G_r = o(\log r)$. Note that the results of this section *do not* require G_r to be changing with r; $G_r = 1$ for all r is a perfectly valid choice of constants. An event of interest is

$$\mathrm{ARR}^r = \Big\{ \sup_{t\in[0,T]} \big| \mathbf{a}^r(t) - \boldsymbol{\lambda}^r t \big| < \varepsilon(r) \Big\},$$
$$\text{where } \varepsilon(r) = K_0 \sqrt{\frac{\log r}{r}} = \Theta\bigg(\sqrt{\frac{1}{r}\log r}\bigg)$$

for appropriate choice of large enough constant K_0. Given that the arrival process to each queue n is an independent Bernoulli process with rate λ_n, an immediate application of the Azuma and Hoeffding's inequality (cf. Proposition 1) or Chernoff bound implies that for K_0 large enough (depending on T, N),

$$\mathbb{P}\big(\mathrm{ARR}^r\big) \geq 1 - \frac{1}{r^2}. \tag{44}$$

Another event of interest will be

$$\mathrm{CON}^r = \Big\{ \max_{0\leq\tau<\tilde{\tau}\leq rT} \varGamma(\tau,\tilde{\tau}) \leq 0 \Big\},$$

where $\varGamma(\tau,\tilde{\tau})$ is defined as

$$\begin{aligned}\varGamma(\tau,\tilde{\tau}) = {} & L\big(\mathbf{Q}(\tilde{\tau})\big) - L\big(\mathbf{Q}(\tau)\big) - \frac{N\mathbf{w}^{\max}(\tilde{\tau}-\tau)}{G} \\ & + \frac{(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2} \sum_{s=\tau}^{\tilde{\tau}-1} \Big(\sum_n \mathsf{LOG}_n\big(Q_n(s)\big) \Big) \\ & - 8N\mathbf{w}^{\max}\bigg(\frac{1}{G} + \log\bigg(\frac{\mathbf{w}^{\max}(\mathbf{Q}(\tau)\cdot\mathbf{1}+\tilde{\tau}-\tau+1)}{G} + 1\bigg)\bigg) \\ & \times \sqrt{(\tilde{\tau}-\tau)\log r}.\end{aligned} \tag{45}$$

By Lemma 1, using the fact that total number of possible pairs $(\tau,\tilde{\tau})$ such that $0 \leq \tau < \tilde{\tau} \leq rT$ are $O(r^2)$ and union bound, it follows that

$$\mathbb{P}\big(\mathrm{CON}^r\big) \geq 1 - \frac{1}{r^2}, \quad \text{for all large enough } r. \tag{46}$$

Notice that the definition of event CON^r (equivalently, $\varGamma(\tau,\tilde{\tau})$) may seem ad hoc. But in fact, it is precisely the event that is implied by the concentration inequality established in Lemma 1.

Formal statement The following result establishes fluid model as a formal approximation of the fluid scaled network operating under the MWL policy. It should be noted that the fluid model defined by (30)–(35) is applicable to *any* measurable scheduling policy. In addition, under the MWL policy with monotonicity of scheduling set $\mathcal{S}$, (36)–(37) are satisfied.

Theorem 2 *Given fixed $T > 0$, let* FMS *be the set of all $x(\cdot) \in C(T)$ satisfying fluid model equations, namely*

- *Equations* (30)–(35) *and* (36)–(37)
- $|\mathbf{q}(0)| \le K$.

And for any $\delta > 0$, *define* FMS_δ *to be the* δ*-fattening of* FMS *as*

$$\mathrm{FMS}_\delta = \Big\{ x \in C(T) : \sup_{t\in[0,T]} \big|x(t) - y(t)\big| < \delta \textit{ for some } y \in \mathrm{FMS} \Big\}.$$

Let assumptions (41)–(42) *be satisfied. Then under the* MWL *policy*

$$\mathbb{P}\big(x^r(\cdot) \in \mathrm{FMS}_\delta\big) \to 1 \quad \textit{as } r \to \infty.$$

Corollary 1 *In addition to the setup of Theorem* 2, *suppose* $\mathbf{q}^r(0) \to \mathbf{q}_0$ *as* $r \to \infty$ *with* $\mathbf{q}^r(0), \mathbf{q}_0$ *nonrandom. With* FMS *as defined in Theorem* 2, *let* $\mathrm{FMS}(\mathbf{q}_0) \subset \mathrm{FMS}$ *be such that in addition* $\mathbf{q}(0) = \mathbf{q}_0$. *Then*

$$\mathbb{P}\big(x^r(\cdot) \in \mathrm{FMS}_\delta(\mathbf{q}_0)\big) \to 1 \quad \textit{as } r \to \infty.$$

3.4 Justifying the fluid model: proof of Theorem 2

The proof of Theorem 2 will build on methodology of Bramson [6] that utilizes notion of *cluster points* introduced earlier. Rather than following [6]'s use of cluster points, we might instead have used the approach based on weak convergence such as that used by Dai [7] or Kelly and Williams [13]. The general line of argument would be (i) the sequence of measures of $x^r(\cdot)$ is tight; (ii) by Prohorov's theorem Billingsley [4, Theorem 5.1], it is relatively compact, so there exists a weakly convergent subsequence; (iii) by the Skorohod representation theorem Billingsley [4, Theorem 6.7], we can express this weak convergence as pathwise convergence; (iv) pathwise limits must satisfy the fluid model equations. Lemma 2 below does the job of (i), Lemma 3 does the job of (iv), and Proposition 1 does the rest of the work—notice how similar it is to the characterization of compact sets in $(C, \|\cdot\|)$ from Billingsley [4, Theorem 7.2]. The benefit of the cluster-point technique is that it gives tighter control of the probability of rare events. The results of this paper are established with the hope that they may be useful in future for establishing other results requiring such tighter control over rare events. For example, the multiplicative state space collapse result (cf. [17]).

3.4.1 Proof of Theorem 2

The proof strategy is to show that under well-behaved arrival process, the entire process x^r is well-behaved in that it is close to a cluster point; then to show that all cluster points of this sequence are fluid model solutions, i.e., satisfy the fluid model equations.

System behavior under good arrivals Let $E_r = \{x^r(\omega) : \omega \in \mathrm{ARR}^r \cap \mathrm{CON}^r\}$, i.e., the set of all possible system trajectories induced under well-behaved sample-paths of arrival process (over $[0, T]$). Lemma 2 below shows that $E_r \subset K_r$, K_r defined in Proposition 1, for appropriate constants K, A and B_r. Therefore,

$$\sup_{\omega\in E_r} d\big(x^r(\omega), \mathrm{CP}\big) \to 0 \quad \text{as } r \to \infty$$

where CP is the set of cluster points of the sequence of events, E_r. Lemma 3 below shows that by our choice of E_r, all cluster points satisfy the fluid model equations, therefore,

$$\sup_{\omega \in E_r} d\big(x^r(\omega), \mathrm{FMS}\big) \to 0 \quad \text{as } r \to \infty.$$

Finally, from (44) and (46), it follows that

$$\mathbb{P}(E_r) \geq 1 - 2/r^2 \to 1, \quad \text{as } r \to \infty.$$

This establishes Theorem 2. □

Lemma 2 (Tightness of fluid scaling) *For every r, with K_r as defined in Proposition 1 and E_r as defined above in the proof of Theorem 2, $E_r \subset K_r$. The constants used to define K_r are K as given in (42), and A and B_r from (49) below.*

Proof We will prove that $x^r \in E_r$ satisfies the two defining conditions of K_r.

For the condition about initial state, note from the definition of the model in Sect. 1.1 that the only nonzero component of $x^r(0)$ is $\mathbf{q}^r(0)$, and that $|\mathbf{q}^r(0)| \leq K$ by Assumption 42.

For condition about the modulus of continuity, consider any $0 \leq s < t \leq T$ with $t - s < \delta$. Write $\lceil t \rceil$ or $\lfloor t \rfloor$ for t rounded up or down to the nearest integral timeslot. We will now look at each component of x^r in turn.

For arrivals,

$$\begin{aligned} \big|\mathbf{a}^r(t) - \mathbf{a}^r(s)\big| &\leq \big|\mathbf{a}^r(t) - \boldsymbol{\lambda}^r t\big| + \big|\mathbf{a}^r(s) - \boldsymbol{\lambda}^r s\big| + \big|\boldsymbol{\lambda}^r(t-s)\big| \\ &\leq 2\varepsilon(r) + \big|\boldsymbol{\lambda}^r\big|\delta \quad \text{from definition of } E_r \\ &\leq 2\varepsilon(r) + A^{\max}\delta, \end{aligned}$$

where the last inequality uses bound

$$\big|\boldsymbol{\lambda}^r\big| \leq A^{\max} \quad \text{for all } r, \tag{47}$$

which follows from Assumption (41).

For idling, consider the following. The maximum amount of service that can be offered to any queue per unit time is unit since $\mathcal{S} \subset \{0,1\}^N$. Then, based on (3), for each n

$$\big|z_n^r(t) - z_n^r(s)\big| < \delta + 2/r.$$

For each $\boldsymbol{\pi}$, since $S_{\boldsymbol{\pi}}(\cdot)$ is increasing and since a schedule must be chosen not more than once every timeslot,

$$\big|s_{\boldsymbol{\pi}}^r(t) - s_{\boldsymbol{\pi}}^r(s)\big| \leq \frac{1}{r}\big(S_{\boldsymbol{\pi}}^r\big(\lceil rt \rceil\big) - S_{\boldsymbol{\pi}}^r\big(\lfloor rs \rfloor\big)\big) < \delta + 2/r.$$

For queue size, note that (2) carries through to the fluid model scaling, i.e.,

$$\mathbf{q}^r(t) = \mathbf{q}^r(0) + \mathbf{a}^r(t) - \big(I - R^{\mathrm{T}}\big)\sum_{\boldsymbol{\pi}} s_{\boldsymbol{\pi}}^r(t)\boldsymbol{\pi} + \mathbf{z}^r(t),$$

thus

$$\begin{aligned}\left|q_n^r(t)-q_n^r(s)\right| &\le \left|a_n^r(t)-a_n^r(s)\right| \\ &\quad + \sum_{\boldsymbol{\pi}} \left|\left[\left(I-R^T\right)\boldsymbol{\pi}\right]_n\right| \left|s_{\boldsymbol{\pi}}^r(t)-s_{\boldsymbol{\pi}}^r(s)\right| + \left|z_n^r(t)-z_n^r(s)\right| \\ &< \delta\left(A^{\max}+|\mathcal{S}|N+1\right)+\left(2|\mathcal{S}|N+2\right)/r+2\varepsilon(r).\end{aligned}$$

Putting all these together,

$$w_\delta\left(x^r\right) < \delta A + B_r, \tag{48}$$

where constants A and B_r are

$$\begin{aligned} A &= \left(2NA^{\max}+2N+\left(N^2+1\right)|\mathcal{S}|\right), \\ B_r &= \left(4N+2\left(N^2+1\right)|\mathcal{S}|\right)/r+4N\varepsilon(r). \end{aligned} \tag{49}$$

□

Lemma 3 (Dynamics at cluster points) *Let x be a cluster point of the sequence $E_r = \{x^r(\omega) : \omega \in \text{ARR}^r \cap \text{CON}^r\}$. Then $x \in$ FMS.*

Proof By definition of cluster point we can find a subsequence r_k and a collection $x^{r_k} \in E_{r_k}$ such that $x^{r_k} \to x$. For simplicity, we shall drop the index k hence forth, i.e., $x^r \to x$ with $x^r \in E_r$. We now use this in proving that x satisfies all the fluid model equations: (30)–(37).

Proof of (30) Observe that

$$\sup_{t\in[0,T]}\left|\mathbf{a}(t)-\boldsymbol{\lambda}t\right| \le \sup_{t\in[0,T]}\left|\mathbf{a}(t)-\mathbf{a}^r(t)\right| + \sup_{t\in[0,T]}\left|\mathbf{a}^r(t)-\boldsymbol{\lambda}^r t\right| + T\left|\boldsymbol{\lambda}^r-\boldsymbol{\lambda}\right|.$$

Each term converges to 0 as $r \to \infty$: the first because $x^r \to x$, the second because $x^r \in E_r$ so a^r is consistent with the event ARR^r, the third by (41). Since the left-hand side does not depend on r, it must be that $\mathbf{a}(t) = \boldsymbol{\lambda}t$.

Proof of (31)–(33) The discrete (unscaled) system satisfies these properties, therefore, the scaled systems x^r do, also. Taking the limit yields the fluid equations.

Proof of (34) In Eq. (49), we found constants A and B_r such that

$$\left|x^r(t)-x^r(s)\right| \le A|t-s|+B_r,$$

with $B_r \to 0$ as $r \to \infty$. Taking the limit as $r \to \infty$, we find that $|x(t)-x(s)| \le A|t-s|$, i.e., x is (globally) Lipschitz continuous (of order 1). And, this immediately implies that x is absolutely continuous.

Proof of (35) Since x is absolutely continuous, each component is too, which means that z_n is differentiable for almost all t. Pick some such t, and suppose that $q_n(t) > 0$. Consider some small interval $I = [t, t+\delta]$ about t. Since q_n is continuous, we can choose δ sufficiently small such that $\inf_{s\in I} q_n(s) > 0$. Since $\|\mathbf{q}^r(\cdot)-\mathbf{q}(\cdot)\| \to 0$, we can find $\alpha > 0$ such that $\inf_{s\in I} q_n^r(s) > \alpha$ for all r sufficiently large. In the unscaled version of the process, this means $\inf_{s\in I} Q_n^r(rs) > r\alpha$. By (3), there is too much work in the queue over this entire interval for there to be any idling, so after rescaling we find $z_n^r(t+\delta/2) = z_n^r(t)$. (The switch from δ to $\delta/2$ sidesteps any discretization problems.) Therefore, the same holds for z_n in the limit. We assumed z_n to be differentiable at t; the derivative must be 0.

Proof of (37) The monotonicity property of $\mathcal{S}$ immediately implies that the discrete (unscaled) system satisfies (8). Therefore the scaled systems x^r do, also. Taking the limit yields the fluid equation.

Initial queue size Clearly, $|\mathbf{q}(0)| \le K$ since $|\mathbf{q}^r(0)| \le K$.

Proof of (36) As discussed earlier, $x(\cdot)$ is absolutely continuous. Therefore, $\mathbf{q}(\cdot)$ is differentiable for almost all $t \in [0, T]$. Consider any such t. We wish to establish the validity of (36) for such t. We shall prove it by contradiction.

To that end, if $\mathbf{q}(t) = \mathbf{0}$ then there is nothing to prove. Therefore, let $\mathbf{q}(t) \neq \mathbf{0}$ and assume on the contrary that $\sum_n w_n \dot{q}_n(t) > -\mathbf{w}^{\min}(1 - \mathsf{L}(\boldsymbol{\lambda}))/N^2$. Note that, $\mathsf{L}(\boldsymbol{\lambda}) \le 1$ for $\boldsymbol{\lambda} \in \Lambda$. Therefore, when $\mathsf{L}(\boldsymbol{\lambda}) = 1$ the above inequality becomes $\sum_n w_n \dot{q}_n(t) > 0$. Therefore, for any small enough $\varepsilon, \varepsilon_1 > 0$, we have

$$\sum_n w_n \big(q_n(t+\varepsilon) - q_n(t)\big) > -\varepsilon \mathbf{w}^{\min}\big(1 - \mathsf{L}(\boldsymbol{\lambda})\big)/N^2 + 2\varepsilon\varepsilon_1.$$

Since $\|\mathbf{q}^r(\cdot) - \mathbf{q}(\cdot)\| \to 0$, for all r large enough,

$$\sum_n w_n \big(q_n^r(t+\varepsilon) - q_n^r(t)\big) > -\varepsilon \mathbf{w}^{\min}\big(1 - \mathsf{L}(\boldsymbol{\lambda})\big)/N^2 + \varepsilon\varepsilon_1.$$

That is,

$$\big(\mathbf{q}^r(t+\varepsilon) - \mathbf{q}^r(t)\big) \cdot \mathbf{w} > -\varepsilon \mathbf{w}^{\min}\big(1 - \mathsf{L}(\boldsymbol{\lambda})\big)/N^2 + \varepsilon\varepsilon_1. \tag{50}$$

Since $\mathbf{q}(t) \neq \mathbf{0}$, $\|\mathbf{q}^r(\cdot) - \mathbf{q}(\cdot)\| \to 0$ and $\mathbf{q}(\cdot)$ being Lipschitz continuous, it follows that by choice of $\varepsilon > 0$ small enough, there exists $\delta > 0$ so that for all $s \in [t, t+\varepsilon]$ and r large enough,

$$\mathbf{q}^r(s) \cdot \mathbf{w} \ge \delta. \tag{51}$$

Therefore, for any $s \in [t, t+\varepsilon]$,

$$\begin{aligned}\sum_n \mathsf{LOG}_n\big(r q_n^r(s)\big) &= \sum_n w_n \log\left(\frac{r w_n q_n^r(s)}{G} + 1\right) \\ &\ge \mathbf{w}^{\min} \log\left(\max_n \frac{r w_n q_n^r(s)}{G} + 1\right) \\ &\ge \mathbf{w}^{\min} \log\left(\frac{r\delta}{NG} + 1\right) \\ &\ge \mathbf{w}^{\min} \log r + K(\delta),\end{aligned} \tag{52}$$

where $K(\delta) = \log\delta - \log N - \log G$ is a finite real valued constant that does not scale with r but depends on N, G and δ. We would like to use (50) and (52) to argue that they will lead to violation of event CON^r, and thus reach desired contradiction. For this, we shall use relation between $L(\mathbf{Q})$ and the LHS of (50) that is stated below in the Proposition 2.

Proposition 2 *Consider* $\mathbf{x}\in\mathbb{R}_+^N$ *with* $|\mathbf{x}|\le B$ *for some constant* B. *Then with respect to large* r,

$$\left|L(r\mathbf{x})-\left[r\log\left(\frac{r}{eG_r}\right)\right]\mathbf{x}\cdot\mathbf{w}-r\,\mathsf{entr}(\mathbf{x}\circ\mathbf{w})\right|=O\big(\log^2 r\big),$$

where $\mathsf{entr}(\mathbf{y})=\sum_n y_n\log y_n$ *and* $\mathbf{x}\circ\mathbf{w}$ *denotes component-wise multiplication, i.e.,* $\mathbf{x}\circ\mathbf{w}=[x_nw_n]$. *The constant in* $O(\cdot)$ *term depends on* N *and* B.

Due to the bound on initial queue-size $|\mathbf{q}^r(0)|\le K$ and Lipschitz continuity of $\mathbf{q}^r(\cdot)$, it follows that $|\mathbf{q}^r(s)|\le B$ for any $s\in[0,T]$ for an appropriately defined constant B, dependent on T, K. Therefore, by an application of Proposition 2, it follows that

$$\begin{aligned}&L\big(r\mathbf{q}^r(t+\varepsilon)\big)-L\big(r\mathbf{q}^r(t)\big)\\&\quad=r\log\left(\frac{r}{eG_r}\right)\big(\mathbf{q}^r(t+\varepsilon)-\mathbf{q}^r(t)\big)\cdot\mathbf{w}\\&\qquad+r\big(\mathsf{entr}\big(\mathbf{q}^r(t+\varepsilon)\circ\mathbf{w}\big)-\mathsf{entr}\big(\mathbf{q}^r(t)\circ\mathbf{w}\big)\big)+O\big(\log^2 r\big)\\&\quad=r\log r\big(\mathbf{q}^r(t+\varepsilon)-\mathbf{q}^r(t)\big)\cdot\mathbf{w}+O(r),\end{aligned}\tag{53}$$

because $\mathsf{entr}(\cdot)$ is a bounded continuous function on $\{\mathbf{x}\in\mathbb{R}_+^N:|\mathbf{x}|\le B\}$ (with bound dependent on N, B). From (50) and (53), it follows that for r large enough

$$L\big(r\mathbf{q}^r(t+\varepsilon)\big)-L\big(r\mathbf{q}^r(t)\big)\ge -r\log r\,\frac{\varepsilon\mathbf{w}^{\min}(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2}+r\log r\,\frac{\varepsilon\varepsilon_1}{2}.\tag{54}$$

We wish to show that (54) violates event CON^r. To see this, use $\tilde{\tau}=rt+r\varepsilon$ and $\tau=rt$ in (45), to obtain

$$\begin{aligned}&L\big(r\mathbf{q}^r(t+\varepsilon)\big)-L\big(r\mathbf{q}^r(t)\big)\\&\quad\le O(r)-\frac{(1-\mathsf{L}(\boldsymbol{\lambda}^r))}{N^2}\left(\sum_{rs=rt}^{rt+r\varepsilon-1}\sum_n\mathsf{LOG}_n\big(rq_n^r(s)\big)\right)+O\big(\sqrt{r\log^3 r}\big)\\&\quad\overset{(a)}{\le}-\frac{1-\mathsf{L}(\boldsymbol{\lambda}^r)}{N^2}\big(r\varepsilon\mathbf{w}^{\min}\log r+r\varepsilon K(\delta)\big)+O(r)\\&\quad\overset{(b)}{\le}-r\log r\,\frac{\varepsilon\mathbf{w}^{\min}(1-\mathsf{L}(\boldsymbol{\lambda}))}{N^2}+r\log r\,\frac{\varepsilon\varepsilon_1}{4}+O(r),\end{aligned}\tag{55}$$

where (a) follows from (52) and (b) holds for large enough r since $\boldsymbol{\lambda}^r\to\boldsymbol{\lambda}$ as $r\to\infty$ and $\mathsf{L}(\cdot)$ is a continuous function. Thus, we have that if CON^r is satisfied then (55) holds, which is contradicted by (54) for r large enough. This completes the proof of (36) and subsequently proof of Lemma 3. □

Proof of Corollary 1 Essentially, we need to show that each cluster point satisfies additional equation $\mathbf{q}(0)=\mathbf{q}_0$. That follows trivially, since for all $\omega\in\mathrm{ARR}^r$ for which $\mathbf{q}^r(0)$ does not converge to $\mathbf{q}_0$ has zero probability. □

Proof of Proposition 2 Define $F_n:\mathbb{R}_+\to\mathbb{R}_+$ as

$$F_n(x)=(w_nx+G_r)\log(w_nx+G_r)-w_nx\log G_r-(w_nx+G_r)\,.$$

And hence $L(r\mathbf{x}) = \sum_n F_n(rx_n)$. Now for any $x \geq 0$,

$$\begin{aligned}
F_n(rx) &= (rxw_n + G_r)\log(w_n rx + G_r) - rxw_n \log G_r - (rxw_n + G_r) \\
&= (rxw_n + G_r)\big[\log r + \log(w_n x + G_r/r)\big] - rxw_n \log eG_r - G_r \\
&= rxw_n[\log r - \log eG_r] + rxw_n \log\left(w_n x + \frac{G_r}{r}\right) \\
&\quad + G_r\big[\log(rxw_n + G_r) - 1\big] \\
&= rxw_n \log\left(\frac{r}{eG_r}\right) + rxw_n \log\left(w_n x + \frac{G_r}{r}\right) + O\big(\log^2 r\big), \qquad (56)
\end{aligned}$$

where the last equality follows from the fact that $G_r = o(\log r)$, $\|\mathbf{x}\|_\infty \leq B$, and hence $|\log(rxw_n + G_r)| = O(\log r)$. To complete the proof, we would like to establish that

$$rxw_n \log\left(w_n x + \frac{G_r}{r}\right) = rxw_n \log w_n x + O\big(\log^2 r\big). \qquad (57)$$

To that end, consider two cases: (i) $x \leq 1/r^2$, (ii) $x > 1/r^2$. In case (i) when $x \leq 1/r^2$, it follows that

$$\left|rxw_n \log\left(w_n x + \frac{G_r}{r}\right)\right| \leq O(\log r), \qquad (58)$$

$$|rxw_n \log w_n x| \leq \left|\frac{w_n}{r} \log\left(\frac{w_n}{r^2}\right)\right| = O(\log r). \qquad (59)$$

Therefore, (57) follows immediately when $x \leq 1/r^2$. For case (ii), when $x > 1/r^2$, consider first-order Taylor's expansion of function $f(z) = \log z$ around $z = w_n x$ to obtain

$$\log\left(w_n x + \frac{G_r}{r}\right) = \log(w_n x) + \frac{G_r}{r}\frac{1}{\theta},$$

where $\theta \in [w_n x, w_n x + G_r/r]$. Therefore, it follows that

$$\left|rw_n x \log\left(w_n x + \frac{G_r}{r}\right) - rw_n x \log(w_n x)\right| \leq G_r = O(\log r). \qquad (60)$$

In summary, for both cases we have established (57). Therefore, we obtain

$$\left|F_n(rx) - rxw_n \log\left(\frac{r}{eG_r}\right) + rxw_n \log(w_n x)\right| = O\big(\log^2 r\big). \qquad (61)$$

Therefore,

$$L(r\mathbf{x}) = r\log\left(\frac{r}{eG_r}\right)\left(\sum_n w_n x_n\right) + r\left(\sum_n w_n x_n \log(w_n x_n)\right) + O\big(\log^2 r\big). (62)$$

This completes the proof of Proposition 2. □

4 Critical fluid model: fixed points

This section characterizes invariant manifold or equivalently the space of fixed points of fluid model solutions. Unlike the previous section, results in this section are limited to single hop network setting. If $\boldsymbol{\lambda}$ is such that $\mathsf{L}(\boldsymbol{\lambda}) < 1$, then from (36) it follows immediately that $\mathbf{q}(t) = \mathbf{0}$ is the only fixed point. When $\mathsf{L}(\boldsymbol{\lambda}) = 1$, nontrivial fixed points do exist. We shall characterize the corresponding space of fixed points for critically loaded fluid model solutions. It is worth remarking here that such characterization of fixed points for critical fluid model solutions is an essential step toward establishing multiplicative state space collapse in the method developed by Bramson [6] and subsequently utilized by Shah and Wischik [17] to study the class of MW policies for the switched network with weight function satisfying scale invariance property. As mentioned earlier, the MWL policy of interest does not have this property. Results of this section, therefore, will serve as an important step toward establishing multiplicative state space collapse of switched network operating under the MWL policy.

4.1 Assumptions

We shall restrict our attention in this section to single-hop network, i.e., $R = 0$. Primary motivation for studying fixed points for critical fluid models is multiplicative state space collapse (cf. see [6, 17]). Therefore, we shall consider the standard *heavy traffic scaling*. Specifically, as before consider a sequence of systems indexed by $r \in \mathbb{N}$. The stochastic model of the rth system obeys all the assumptions stated in Sect. 1.1 and this sequence of systems satisfy assumptions stated in Sect. 3.3.2. In addition, we shall impose additional constraints on G_r that

$$\lim_{r\to\infty}\inf G_r \to \infty \quad \text{and} \quad \lim_{r\to\infty}\sup \frac{G_r}{\log r} = 0. \tag{63}$$

As mentioned earlier in the paper, the condition $G_r = o(\log r)$ with $G_r \to \infty$ is to make sure that the policy behavior at fluid scale is not affected by the choice of constant G_r and only the fluid queue-state appears in the critical fluid model as well as characterization of the invariant points. However, $G_r = \omega(1)$ is imposed by the proof method of this paper and not clear if essential.

The arrival rate $\boldsymbol{\lambda}^r$ of the rth system is such that

$$\boldsymbol{\lambda}^r = \boldsymbol{\lambda} - \frac{1}{r}\Omega, \tag{64}$$

for all large enough r, for some vector $\Omega \in \mathbb{R}^N_+$ and $\boldsymbol{\lambda}$ such that $\mathsf{L}(\boldsymbol{\lambda}) = 1$. We shall assume that $\boldsymbol{\lambda} > 0$ component-wise. This is because, if $\lambda_n = 0$, then we shall ignore such a queue from consideration.

4.2 Preliminaries

Since the network is single-hop, i.e., $R = 0$, $\vec{\boldsymbol{\lambda}} = \boldsymbol{\lambda}$. Recall from Sect. 2.1, that for a critical $\boldsymbol{\lambda}$, i.e., $\mathsf{L}(\boldsymbol{\lambda}) = 1$, the cost of the optimal solution of problem PRIMAL($\boldsymbol{\lambda}$) is 1. The dual of PRIMAL($\boldsymbol{\lambda}$), denoted by DUAL($\boldsymbol{\lambda}$), is as follows:

$$\boxed{\begin{array}{ll} \text{maximize} & \boldsymbol{\xi}\cdot\boldsymbol{\lambda} \\ \text{over} & \boldsymbol{\xi}\in\mathbb{R}_+^N \\ \text{such that} & \max_{\boldsymbol{\pi}\in\mathcal{S}}\boldsymbol{\xi}\cdot\boldsymbol{\pi}\le 1 \end{array}}$$

The solution is clearly attained when the constraint is tight. Define the dual feasible $\boldsymbol{\xi}$s as virtual resources, i.e.,

$$\mathsf{VR}=\left\{\boldsymbol{\xi}\in\mathbb{R}_+^N:\max_{\boldsymbol{\pi}\in\mathcal{S}}\boldsymbol{\xi}\cdot\boldsymbol{\pi}\le 1\right\}.$$

Given a queue size vector $\mathbf{Q}$ and $\boldsymbol{\xi}\in\mathsf{VR}$, define $\boldsymbol{\xi}\cdot\mathbf{Q}$ as the *workload* at the *virtual resource* $\boldsymbol{\xi}$. In this section, our interest is in critical $\boldsymbol{\lambda}$. That is, $\boldsymbol{\lambda}$ such that $\text{PRIMAL}(\boldsymbol{\lambda})=1$. By strong duality we have $\text{PRIMAL}(\boldsymbol{\lambda})=\text{DUAL}(\boldsymbol{\lambda})=1$. Define critical virtual resources as $\mathsf{CVR}(\boldsymbol{\lambda})$ where

$$\mathsf{CVR}(\boldsymbol{\lambda})=\{\boldsymbol{\xi}\in\mathsf{VR}:\boldsymbol{\xi}\cdot\boldsymbol{\lambda}=1\}.$$

It can be checked that $\mathsf{CVR}(\boldsymbol{\lambda})$ is non-empty and finite dimensional bounded polytope. Therefore, it has finitely many extreme points. Let $\mathcal{S}^*=\mathcal{S}^*(\boldsymbol{\lambda})$ be the set of extreme points of $\mathsf{CVR}(\boldsymbol{\lambda})$. We shall denote $\boldsymbol{\xi}\in\mathcal{S}^*$ by *principal* critically-loaded virtual resources. Subsequently, any $\boldsymbol{\zeta}\in\mathsf{CVR}$ can be expressed as

$$\boldsymbol{\zeta}=\sum_{\boldsymbol{\xi}\in\mathcal{S}^*}x_{\boldsymbol{\xi}}\boldsymbol{\xi}\quad\text{with }\sum x_{\boldsymbol{\xi}}=1\text{ and all }x_{\boldsymbol{\xi}}\ge 0. \tag{65}$$

The following is a useful proposition.

Proposition 3 *Given critically loaded* $\boldsymbol{\lambda}$, *let* $\mathbf{q}$, $\tilde{\mathbf{q}}\in\mathbb{R}_+^N$ *be such that*

$$\tilde{\mathbf{q}}\cdot\boldsymbol{\xi}\ge\mathbf{q}\cdot\boldsymbol{\xi},\quad\forall\boldsymbol{\xi}\in\mathcal{S}^*(\boldsymbol{\lambda}).$$

Then there exists $U\ge 0$ *and* $\boldsymbol{\sigma}\in\Sigma$ *so that*

$$\tilde{\mathbf{q}}=\mathbf{q}+U(\boldsymbol{\lambda}-\boldsymbol{\sigma}).$$

Further, if $|\mathbf{q}|,|\tilde{\mathbf{q}}|\le B$ *for some constant* B, *then* U *is uniformly bounded.*

Proof Consider the space of all virtual resources VR. It is a finite dimensional polytope. Since $\mathcal{S}\subset\{0,1\}^N$, $\mathcal{S}$ is monotone and $\mathbf{e}^n\in\mathcal{S}$ for all $1\le n\le N$, it follows that for each $\boldsymbol{\xi}\in\mathsf{VR}$, $\xi_n\le 1$ for all n. Further VR has finitely many extreme points. Let they be denoted by EVR. Given critically loaded $\boldsymbol{\lambda}$, as explained earlier, there is a subset of EVR, denoted by $\mathcal{S}^*(\boldsymbol{\lambda})$ so that for each $\boldsymbol{\xi}\in\mathcal{S}^*(\boldsymbol{\lambda})$ we have $\boldsymbol{\xi}\cdot\boldsymbol{\lambda}=1$. For all $\boldsymbol{\xi}\in\mathsf{EVR}\backslash\mathcal{S}^*(\boldsymbol{\lambda})$, $\boldsymbol{\xi}\cdot\boldsymbol{\lambda}<1$. Let,

$$\varepsilon=1-\max_{\boldsymbol{\xi}\in\mathsf{EVR}\backslash\mathcal{S}^*(\boldsymbol{\lambda})}\boldsymbol{\xi}\cdot\boldsymbol{\lambda}. \tag{66}$$

And define

$$\boldsymbol{\sigma}=\boldsymbol{\lambda}-\frac{1}{U}(\tilde{\mathbf{q}}-\mathbf{q}), \tag{67}$$

where

$$U = \frac{\max(\widetilde{\mathbf{q}}^{\max}, \mathbf{q}^{\max})}{\min(\varepsilon, \boldsymbol{\lambda}^{\min})}. \tag{68}$$

Here, U is well defined since $\boldsymbol{\lambda} > 0$ component-wise, i.e., $\boldsymbol{\lambda}^{\min} > 0$. We claim that $\boldsymbol{\sigma} \in \Sigma$. To start with, note that $\boldsymbol{\sigma} \geq 0$ due to choice of U. Next, we shall establish that $\boldsymbol{\xi} \cdot \boldsymbol{\sigma} \leq 1$ for all $\boldsymbol{\xi} \in \mathsf{VR}$. This will, immediately imply that DUAL$(\boldsymbol{\sigma}) \leq 1$. Therefore, PRIMAL$(\boldsymbol{\sigma}) \leq 1$, i.e., $\boldsymbol{\sigma} \in \Sigma$ as desired. To complete the proof, note that it is sufficient to consider $\boldsymbol{\xi} \in \mathsf{EVR}$. For $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$,

$$\begin{aligned} \boldsymbol{\xi} \cdot \boldsymbol{\sigma} &= \boldsymbol{\xi} \cdot \boldsymbol{\lambda} - \frac{1}{U}(\boldsymbol{\xi} \cdot \widetilde{\mathbf{q}} - \boldsymbol{\xi} \cdot \mathbf{q}) \\ &= 1 - (\boldsymbol{\xi} \cdot \widetilde{\mathbf{q}} - \boldsymbol{\xi} \cdot \mathbf{q}) \\ &\leq 1, \end{aligned} \tag{69}$$

where we have used hypothesis of Proposition that $\boldsymbol{\xi} \cdot \widetilde{\mathbf{q}} \geq \boldsymbol{\xi} \cdot \mathbf{q}$ for all $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$. For $\boldsymbol{\xi} \in \mathsf{EVR} \backslash \mathcal{S}^*(\boldsymbol{\lambda})$, due to choice of U, it can be checked that $\boldsymbol{\xi} \cdot \boldsymbol{\sigma} \leq 1$. This completes the proof. Note that the choice of U is uniform when $|\widetilde{\mathbf{q}}|$, $|\mathbf{q}|$ are bounded by a constant B. □

4.3 A useful optimization

Define a function $\ell : \mathbb{R}_+^N \to \mathbb{R}^2$ as

$$\ell(\mathbf{y}) = \big(\mathbf{y} \cdot \mathbf{w}, \mathsf{entr}(\mathbf{y} \circ \mathbf{w})\big),$$

where recall that $\mathsf{entr}(\mathbf{z}) = \sum_n z_n \log z_n$ and $\mathbf{y} \circ \mathbf{w}$ represents component-wise multiplication $[y_n w_n]$. The function $\ell(\mathbf{y})$ is to be interpreted as assigning a real valued tuple to each N dimensional vector $\mathbf{y} \in \mathbb{R}_+^N$ with strict lexicographic order on these resulting tuples. That is, $\ell(\mathbf{y}) < \ell(\mathbf{y}')$ if and only if either $\mathbf{y} \cdot \mathbf{w} < \mathbf{y}' \cdot \mathbf{w}$ or $\mathbf{y} \cdot \mathbf{w} = \mathbf{y}' \cdot \mathbf{w}$, $\mathsf{entr}(\mathbf{y} \circ \mathbf{w}) < \mathsf{entr}(\mathbf{y}' \circ \mathbf{w})$.

Given $\mathbf{q} \in \mathbb{R}_+^N$, define optimization problem $\mathsf{opt}(\mathbf{q})$ as follows:

minimize	$\ell(\mathbf{y})$
over	$\mathbf{y} \in \mathbb{R}_+^N$
such that	$\boldsymbol{\xi} \cdot \mathbf{y} \geq \boldsymbol{\xi} \cdot \mathbf{q}$ for all $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$

4.4 Fluid model: fixed points

Theorem 2 implies that under the above stated assumptions, the fluid scaled system is well approximated by fluid model solutions, denoted by FMS, that were defined earlier for given $\boldsymbol{\lambda}$. That is, fluid limit points of the fluid scaled system satisfy the fluid model Eqs. (30)–(37). Here, we study additional properties of these fluid limit points. Specifically, we characterize fixed or invariant points for the limiting dynamics of the fluid scaled system.

Given $T > 0$ let FMS$'$ be the set of all $x(\cdot) = (\mathbf{q}(\cdot), \mathbf{q}(\cdot), \mathbf{z}(\cdot), s(\cdot)) \in C(T)$ such that they satisfy

- Equations (30)–(35) and (36)–(37)
- $|\mathbf{q}(0)| \le K$
- and, for any regular point $t \in [0, T]$,

$$\dot{\mathbf{q}}(t) = \mathbf{0} \quad \text{iff} \quad \mathbf{q}(t) \text{ solves } \mathsf{opt}\big(\mathbf{q}(t)\big). \tag{70}$$

As before, define FMS'_δ to be the δ-fattening of FMS' as

$$\mathrm{FMS}'_\delta = \Big\{x \in C(T) : \sup_{t\in[0,T]} \big|x(t) - y(t)\big| < \delta \text{ for some } y \in \mathrm{FMS}'\Big\}.$$

Theorem 3 *Given fixed $T > 0$ and $\boldsymbol{\lambda}$ with $\mathsf{L}(\boldsymbol{\lambda}) = 1$, let the* MWL *policy utilize weights* $\mathbf{w}$ *such that*

$$c\mathbf{w} \in \mathsf{CVR}(\boldsymbol{\lambda}), \quad \textit{for some } c > 0. \tag{71}$$

Then under the above stated assumptions,

$$\mathbb{P}\big(x^r(\cdot) \in \mathrm{FMS}'_\delta\big) \to 1 \quad \textit{as } r \to \infty.$$

4.5 Proof of Theorem 3

We shall essentially build on proof of Theorem 2. Recall that the proof strategy for Theorem 2 involved showing that under well-behaved arrival process, the entire process x^r is well-behaved in that it is close to a cluster point; then to show that all cluster points of this sequence are fluid model solutions, i.e., satisfy the fluid model equations. Here, we shall use the same definition for well-behaved arrival process. To this end, $E_r = \{x^r(\omega) : \omega \in \mathrm{ARR}^r \cap \mathrm{CON}^r\}$, i.e., the set of all possible paths for the entire system for any well-behaved arrival process (over time interval $[0, T]$). Here we shall use definition of CON^r with $G = G_r$; as per assumption in Theorem 3 $\liminf_{r\to\infty} G_r = \infty$, $\limsup_{r\to\infty} G_r/\log r = 0$. Recall that, in the definition of ARR^r, $\varepsilon(r) = \Theta\big(\sqrt{\frac{1}{r}\log r}\big)$. An important consequence of initial condition $|\mathbf{q}^r(0)| \le K$, event ARR^r and $\boldsymbol{\lambda}^r \to \boldsymbol{\lambda}$ is that for given fixed T, we have that for any $t \in [0, T]$ and all r

$$\big|\mathbf{q}^r(t)\big| \le B, \tag{72}$$

for some constant B (dependent on $N, T, K, \boldsymbol{\lambda}$). That is $\mathbf{q}^r(\cdot)$ is always in a bounded set $[0, B]^N$ over $[0, T]$.

By similar application of Lemma 2 as in proof of Theorem 2, it follows that

$$\sup_{\omega\in E_r} d\big(x^r(\omega), \mathrm{CP}\big) \to 0 \quad \text{as } r \to \infty$$

where CP is the set of cluster points of the E_r. Lemma 3 shows that by our choice of E_r, all cluster points satisfy the basic fluid model equations: (30)–(37). In what follows (Sect. 4.6), we shall establish that state $\mathbf{q}$ is a fixed point for cluster points of E_r if and only if $\mathbf{q}$ solves optimization problem $\mathsf{opt}(\mathbf{q})$. Then it follows that

$$\sup_{\omega\in E_r} d\big(x^r(\omega), \mathrm{FMS}'\big) \to 0 \quad \text{as } r \to \infty.$$

Finally, from (44) and (46), it follows that

$$\mathbb{P}(E_r) \ge 1 - 2/r^2 \to 1, \quad \text{as } r \to \infty.$$

This establishes Theorem 3. □

4.6 Fixed-point characterization

By the definition of a cluster point, we can find a subsequence r_k and a collection $x^{r_k} \in E_{r_k}$ such that $x^{r_k} \to x$. For simplicity, we shall drop the index k hence forth, i.e., $x^r \to x$ with $x^r \in E_r$. We have that x satisfies all the basic fluid model equations: (30)–(37). We wish to establish that x satisfies the fixed-point equation (70). Equivalently, we wish to show that with $\mathbf{q}(0) = \mathbf{q}$, $\mathbf{q}(t) = \mathbf{q}$ for all $t > 0$ if and only if $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$. Note that $\mathbf{q} = \mathbf{0}$ is a fixed point from work-conservation property and it solves $\mathsf{opt}(\mathbf{0})$. Therefore, we need to establish this property for the case when $\mathbf{q} \neq \mathbf{0}$.

In what follows, we shall establish this by studying evolution of the unscaled system in detail. Naturally, the remainder of the proof will be divided into two parts with the first part establishing implication that if $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$ then it is a fixed point; the second part establishing the reverse implication.

4.6.1 $\mathbf{q}$ *solves* $\mathsf{opt}(\mathbf{q}) \Rightarrow \mathbf{q}$ *is a fixed point*

We shall establish that if $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$, then it is a fixed point, i.e., for such a $\mathbf{q}$ if $\mathbf{q}(0) = \mathbf{q}$ then $\mathbf{q}(t) = \mathbf{q}$ for all $t > 0$. As mentioned earlier, we shall do this by analyzing evolution of the unscaled system. To this end, let the rth system have scaled initial state $\mathbf{q}^r(0)$ such that $\mathbf{q}^r(0) \to \mathbf{q}$ as $r \to \infty$. That is, $\mathbf{q}^r(0) = \mathbf{q}(0) + \boldsymbol{\varepsilon}_1(r)$ with $|\boldsymbol{\varepsilon}_1(r)| \to 0$ as $r \to \infty$. It will be sufficient to show that $|\mathbf{q}^r(t) - \mathbf{q}^r(0)| \to 0$ as $r \to \infty$.

With the above goal in sight, let us start by defining useful optimization problems. Recall the Lyapunov function

$$L^r(\mathbf{Y}) = \sum_n (w_n Q_n + G_r)\log(w_n Q_n + G_r) - w_n Q_n \log G_r - (w_n Q_n + G_r),$$

where explicit use of r in $L^r(\cdot)$ is to note that $G = G_r$ varies with r. Define optimization problem $\mathsf{OPT}^r(\mathbf{Q})$ as follows:

$$\boxed{\begin{array}{ll} \text{minimize} & L^r(\mathbf{Y}) \\ \text{over} & \mathbf{Y} \in \mathbb{R}_+^N \\ \text{such that} & \boldsymbol{\xi} \cdot \mathbf{Y} \ge \boldsymbol{\xi} \cdot \mathbf{Q} \quad \text{for all } \boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda}) \end{array}}$$

The Lagrangian form, denoted as $\mathsf{Lagr}^r(\mathbf{Q})$, of $\mathsf{OPT}^r(\mathbf{Q})$ is as follows:

$$\boxed{\begin{array}{ll} \text{minimize} & \mathrm{La}^r(\mathbf{Y}, \boldsymbol{\phi}; \mathbf{Q}) \stackrel{\triangle}{=} L^r(\mathbf{Y}) - \displaystyle\sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi_{\boldsymbol{\xi}}(\boldsymbol{\xi} \cdot \mathbf{Y} - \boldsymbol{\xi} \cdot \mathbf{Q}) \\ \text{over} & \mathbf{Y} \in \mathbb{R}_+^N \\ \text{suchthat} & \phi_{\boldsymbol{\xi}} \ge 0 \quad \text{for all } \boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda}) \end{array}}$$

For any $\mathbf{Q} \in \mathbb{R}_+^N$, the optimization problem $\mathsf{OPT}^r(\mathbf{Q})$ has a convex objective with linear constraints. Since $\mathbf{Q}$ itself is a feasible solution, the domain can be restricted to a bounded convex set. Over this restricted set, the objective function is strictly convex, and hence achieves a unique minimum, say $\hat{\mathbf{Q}}$. The optimization problem $\mathsf{OPT}^r(\mathbf{Q})$ satisfies the Slater's condition, and hence by strong duality (cf. see [3, 5]), the following holds: there exists a choice of Lagrangian dual variables $\boldsymbol{\phi}(\mathbf{Q})$, so that $(\hat{\mathbf{Q}}, \boldsymbol{\phi}(\mathbf{Q}))$ is a solution of $\mathsf{Lagr}^r(\mathbf{Q})$ and

$$\mathrm{La}^r\big(\hat{\mathbf{Q}}, \boldsymbol{\phi}(\mathbf{Q}); \mathbf{Q}\big) = L^r(\hat{\mathbf{Q}}). \tag{73}$$

And, for any $\mathbf{Y} \in \mathbb{R}_+^N$ and choice of nonnegative dual variables $\boldsymbol{\phi}$,

$$\mathrm{La}^r(\mathbf{Y}, \boldsymbol{\phi}; \mathbf{Q}) \geq L^r(\mathbf{Q}). \tag{74}$$

Consider objective $\mathrm{La}^r(\mathbf{Y}, \boldsymbol{\phi}; \mathbf{Q})$ of $\mathsf{Lagr}^r(\mathbf{Q})$. For any n,

$$\begin{aligned}\frac{\partial\,\mathrm{La}^r(\mathbf{Y}, \boldsymbol{\phi}; \mathbf{Q})}{\partial Y_n} &= \frac{\partial L^r(\mathbf{Y})}{\partial Y_n} - \sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi_{\boldsymbol{\xi}} \left(\sum_m \xi_m \frac{\partial Y_m}{\partial Y_n}\right) \\ &= \mathsf{LOG}_n^r(Y_n) - \sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi_{\boldsymbol{\xi}} \xi_n,\end{aligned} \tag{75}$$

where LOG_n^r is the same as LOG_n with $G = G_r$, that is,

$$\mathsf{LOG}_n^r(x) = w_n \log(w_n Q_n + G_r) - w_n \log G_r.$$

Since $\hat{\mathbf{Q}}$ and $\boldsymbol{\phi}(\mathbf{Q})$ is an optimal solution of $\mathsf{Lagr}^r(\mathbf{Q})$, either of the following holds: (i) $\hat{Q}_n = 0$ and $\frac{\partial\,\mathrm{La}^r(\mathbf{Y}, \boldsymbol{\phi}(\mathbf{Q}); \mathbf{Q})}{\partial Y_n}|_{Y=\hat{\mathbf{Q}}} > 0$, or (ii) $\frac{\partial\,\mathrm{La}^r(\mathbf{Y}, \boldsymbol{\phi}(\mathbf{Q}); \mathbf{Q})}{\partial Y_n}|_{Y=\hat{\mathbf{Q}}} = 0$. Since $\mathsf{LOG}_n^r(0) = 0$ for all n, r, and dual variables $\boldsymbol{\phi}$ are always nonnegative, from (75) it follows that (i) is not possible. That is, the optimal solution $(\hat{\mathbf{Q}}, \boldsymbol{\phi}(\mathbf{Q}))$ must satisfy the following: for all n,

$$\mathsf{LOG}_n^r(\hat{Q}_n) - \sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi_{\boldsymbol{\xi}}(\mathbf{Q}) \xi_n = 0. \tag{76}$$

Now we shall embark on proving $|\mathbf{q}^r(t) - \mathbf{q}^r(0)| \to 0$ as $r \to \infty$ assuming $\mathbf{q}^r(0) \approx \mathbf{q}$ with $\mathbf{q}$ being a solution of $\mathsf{opt}(\mathbf{q})$. To this end, let $r\hat{\mathbf{q}}^r(0)$ be the solution of $\mathsf{OPT}(r\mathbf{q}^r(0))$. From the above discussion, there exists a choice of nonnegative valued $\boldsymbol{\phi}^r \overset{\triangle}{=} \boldsymbol{\phi}(r\mathbf{q}^r(0))$ such that

$$\begin{aligned}\mathrm{La}^r\big(r\hat{\mathbf{q}}^r(0), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)\big) &= L^r\big(r\hat{\mathbf{q}}^r(0)\big), \\ \mathrm{La}^r\big(r\mathbf{y}, \boldsymbol{\phi}^r; r\mathbf{q}^r(0)\big) &\geq L^r\big(r\hat{\mathbf{q}}^r(0)\big), \quad \text{for any } \mathbf{y} \in \mathbb{R}_+^N.\end{aligned} \tag{77}$$

Next we state two important lemmas. They assume that $x^r(\cdot) \in E_r$. Their proofs are provided later.

Lemma 4 *For any $t \in [0, T]$,*

$$\big|\mathrm{La}^r\big(r\mathbf{q}^r(t), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)\big) - \mathrm{La}^r\big(r\hat{\mathbf{q}}^r(0), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)\big)\big| = o(r).$$

Lemma 5 *For any $t \in [0, T]$,*

$$\big|\mathbf{q}^r(t) - \hat{\mathbf{q}}^r(0)\big| = O\left(\sqrt{\frac{|\mathrm{La}^r(r\mathbf{q}^r(t), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)) - \mathrm{La}^r(r\hat{\mathbf{q}}^r(0), \boldsymbol{\phi}^r; r\mathbf{q}^r(0))|}{r}}\right).$$

As per Lemmas 4 and 5, it follows that $|\mathbf{q}^r(t) - \hat{\mathbf{q}}^r(0)| \to 0$ as $r \to \infty$ for all t. Therefore, it must be that $|\mathbf{q}^r(t) - \mathbf{q}^r(0)| \to 0$ as $r \to \infty$. This is because $\|\mathbf{q}^r(\cdot) - \mathbf{q}(\cdot)\| \to 0$ as $r \to \infty$ and $\mathbf{q}(\cdot)$ is Lipschitz continuous. Therefore, it follows that $|\hat{\mathbf{q}}^r(0) - \mathbf{q}| \to 0$ as $r \to \infty$ and $\mathbf{q}(t) = \mathbf{q}(0) = \mathbf{q}$. This completes the proof of $\mathbf{q}$ being a fixed point.

Proof of Lemma 4 Consider the following:

$$\begin{aligned}
&\mathrm{La}^r\big(r\mathbf{q}^r(t), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)\big) \\
&\quad = L^r\big(r\mathbf{q}^r(t)\big) - \sum_{\boldsymbol{\xi}\in\mathcal{S}^*} \phi^r_{\boldsymbol{\xi}}\big(\boldsymbol{\xi}\cdot r\mathbf{q}^r(t) - \boldsymbol{\xi}\cdot r\mathbf{q}^r(0)\big) \\
&\quad = L^r\big(r\hat{\mathbf{q}}^r(0)\big) + \big[L^r\big(r\mathbf{q}^r(0)\big) - L^r\big(r\hat{\mathbf{q}}^r(0)\big)\big] \\
&\qquad + \big[L^r\big(r\mathbf{q}^r(t)\big) - L^r\big(r\mathbf{q}^r(0)\big)\big] - r\sum_{\boldsymbol{\xi}\in\mathcal{S}^*} \phi^r_{\boldsymbol{\xi}}\boldsymbol{\xi}\cdot\big(\mathbf{q}^r(t) - \mathbf{q}^r(0)\big). \qquad (78)
\end{aligned}$$

Define,

$$\begin{aligned}
\delta_1(r) &\triangleq L^r\big(r\mathbf{q}^r(0)\big) - L^r\big(r\hat{\mathbf{q}}^r(0)\big), \\
\delta_2(r) &\triangleq L^r\big(r\mathbf{q}^r(t)\big) - L^r\big(r\mathbf{q}^r(0)\big), \\
\delta_3(r) &\triangleq \sum_{\boldsymbol{\xi}\in\mathcal{S}^*} \phi^r_{\boldsymbol{\xi}}\boldsymbol{\xi}\cdot\big(\mathbf{q}^r(t) - \mathbf{q}^r(0)\big).
\end{aligned}$$

Given that $\mathrm{La}^r(r\hat{\mathbf{q}}^r(0), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)) = L^r(r\hat{\mathbf{q}}^r(0))$, $\mathrm{La}^r(r\mathbf{q}^r(t), \boldsymbol{\phi}^r; r\mathbf{q}^r(0)) \geq \mathrm{La}^r(r\hat{\mathbf{q}}^r(0), \boldsymbol{\phi}^r; r\mathbf{q}^r(0))$ and (78), it is sufficient to prove the following:

1. $\delta_1(r) \leq o(r)$, i.e., $\limsup_{r\to\infty} \delta_1(r)/r \leq 0$,
2. $\delta_2(r) \leq o(r)$, i.e., $\limsup_{r\to\infty} \delta_2(r)/r \leq 0$, and
3. $\delta_3(r) \geq -o(1)$, i.e., $\liminf_{r\to\infty} \delta_3(r) \geq 0$.

1. *Proof of* $\delta_1(r) \leq o(r)$. Recall that $r\hat{\mathbf{q}}^r(0)$ solves $\mathsf{OPT}^r(r\mathbf{q}^r(0))$. Therefore, by feasibility conditions

$$\hat{\mathbf{q}}^r(0)\cdot\boldsymbol{\xi} \geq \mathbf{q}^r(0)\cdot\boldsymbol{\xi}, \quad \forall\boldsymbol{\xi}\in\mathcal{S}^*(\boldsymbol{\lambda}), \qquad (79)$$

and since $r\mathbf{q}^r(0)$ is a feasible solution,

$$L^r\big(r\hat{\mathbf{q}}^r(0)\big) \leq L^r\big(r\mathbf{q}^r(0)\big). \qquad (80)$$

By assumption in statement of Theorem 3, we have $c\mathbf{w} \in \mathsf{CVR}(\boldsymbol{\lambda})$ for some $c > 0$. Therefore, (79) implies that

$$c\hat{\mathbf{q}}^r(0)\cdot\mathbf{w} \geq c\mathbf{q}^r(0)\cdot\mathbf{w}, \quad \text{equivalently } \hat{\mathbf{q}}^r(0)\cdot\mathbf{w} \geq \mathbf{q}^r(0)\cdot\mathbf{w}. \qquad (81)$$

Let $\theta_1(r) = (\hat{\mathbf{q}}^r(0) - \mathbf{q}^r(0))\cdot\mathbf{w}$ then $\theta_1(r) \geq 0$. Now $\delta_1(r) = L^r(r\mathbf{q}^r(0)) - L^r(r\hat{\mathbf{q}}^r(0)) \geq 0$ and we need to upper bound it. By Proposition 2,

$$\begin{aligned}\delta_1(r) &= L^r\big(r\mathbf{q}^r(0)\big) - L^r\big(r\hat{\mathbf{q}}^r(0)\big)\\ &= r\log\left(\frac{r}{eG_r}\right)\big(\mathbf{q}^r(0) - \hat{\mathbf{q}}^r(0)\big)\cdot\mathbf{w} + r\big(\mathsf{entr}\big(\mathbf{q}^r(0)\circ\mathbf{w}\big)\\ &\quad - \mathsf{entr}\big(\hat{\mathbf{q}}^r(0)\big)\circ\mathbf{w}\big) + O\big(\log^2 r\big)\\ &= -r\log\left(\frac{r}{eG_r}\right)\theta_1(r) + r\big(\mathsf{entr}\big(\mathbf{q}^r(0)\circ\mathbf{w}\big) - \mathsf{entr}\big(\hat{\mathbf{q}}^r(0)\big)\circ\mathbf{w}\big)\\ &\quad + O\big(\log^2 r\big).\end{aligned} \tag{82}$$

We need the RHS of (82) to be $o(r)$. First some observations about $\theta_1(r)$. As stated earlier, $\theta_1(r)\geq 0$. Suppose $\limsup_r \theta_1(r) > 0$. Then $\liminf_r \delta_1(r) = -\infty$. This is because $\mathbf{q}^r(0), \hat{\mathbf{q}}^r(0)$ are in a bounded set, $\mathsf{entr}(\cdot)$ is bounded continuous function and (82). However, this is a contradiction to $\delta_1(r)\geq 0$ as established earlier. Therefore, $\limsup_r \theta_1(r) = 0$, i.e.

$$\theta_1(r) = o(1). \tag{83}$$

In fact, the same argument will imply that

$$\big|\theta_1(r)\big| = O(1/\log r). \tag{84}$$

Subsequently, to conclude $\delta_1(r)\leq o(r)$ as desired, it is sufficient to establish that

$$\mathsf{entr}\big(\hat{\mathbf{q}}^r(0)\circ\mathbf{w}\big) \geq \mathsf{entr}\big(\mathbf{q}^r(0)\circ\mathbf{w}\big) - o(1). \tag{85}$$

Towards this, recall that $\boldsymbol{\varepsilon}_1(r) = \mathbf{q}^r(0) - \mathbf{q}(0)$ with $\|\boldsymbol{\varepsilon}_1(r)\| = o(1)$. Again, since $\mathsf{entr}(\cdot)$ is a bounded continuous function on bounded sets of type $[0, B]^N$, it follows that

$$\mathsf{entr}\big(\mathbf{q}^r(0)\circ\mathbf{w}\big) = \mathsf{entr}\big(\mathbf{q}(0)\circ\mathbf{w}\big) + o(1). \tag{86}$$

Now let $\boldsymbol{\varepsilon}_2(r) = \hat{\mathbf{q}}^r(0) - \mathbf{q}^r(0)$. Either $|\boldsymbol{\varepsilon}_2(r)| = o(1)$, i.e., $\limsup_r |\boldsymbol{\varepsilon}_2(r)| = 0$, or $|\boldsymbol{\varepsilon}_2(r)| = \Omega(1)$, i.e., $\liminf_r |\boldsymbol{\varepsilon}_2(r)| > 0$. When $|\boldsymbol{\varepsilon}_2(r)| = o(1)$, $\mathsf{entr}(\hat{\mathbf{q}}^r(0)\circ\mathbf{w}) = \mathsf{entr}(\mathbf{q}^r(0)\circ\mathbf{w}) + o(1)$ follows from uniform continuity of $\mathsf{entr}(\cdot)$ on bounded set, and hence we obtain desired (85). Therefore, the situation to worry is the one when $|\boldsymbol{\varepsilon}_2(r)| = \Omega(1)$.

Toward this, define $\theta_2(r) = (\hat{\mathbf{q}}^r(0) - \mathbf{q}(0))\cdot\mathbf{w}$. And let $\mathcal{O}(\boldsymbol{\varepsilon},\theta)$ be the value of optimization problem

$$\begin{aligned}&\text{minimize} \quad \mathsf{entr}(\mathbf{x}\circ\mathbf{w}) \quad \text{over } \mathbf{x}\in\mathbb{R}^N_+\\ &\text{subject to} \quad \mathbf{x}\cdot\boldsymbol{\xi} \geq \mathbf{q}(0)\cdot\boldsymbol{\xi} + \boldsymbol{\varepsilon}\cdot\boldsymbol{\xi}, \quad \forall\boldsymbol{\xi}\in\mathcal{S}^*\\ &\qquad\qquad\quad \big|\mathbf{x}\cdot\mathbf{w} - \mathbf{q}(0)\cdot\mathbf{w}\big| \leq |\boldsymbol{\varepsilon}\cdot\mathbf{w}| + |\theta|.\end{aligned} \tag{87}$$

By definition, $r\hat{\mathbf{q}}^r(0)$ solves $\mathsf{OPT}(r\mathbf{q}^r(0))$. Now $\mathbf{q}^r(0) = \mathbf{q}(0) + \boldsymbol{\varepsilon}_1(r)$, $\theta_2(r) = (\hat{\mathbf{q}}^r(0) - \mathbf{q}(0))\cdot\mathbf{w}$, it follows that $\hat{\mathbf{q}}^r(0)$ is feasible for optimization (87) with $(\boldsymbol{\varepsilon},\theta) = (\boldsymbol{\varepsilon}_1(r), \theta_2(r))$. Therefore, it follows that

$$\mathsf{entr}\big(\hat{\mathbf{q}}^r(0)\circ\mathbf{w}\big) \geq \mathcal{O}\big(\boldsymbol{\varepsilon}_1(r), \theta_2(r)\big). \tag{88}$$

Now

$$\begin{aligned}\theta_2(r) &= \big(\hat{\mathbf{q}}^r(0)-\mathbf{q}(0)\big)\cdot\mathbf{w}\\ &= \big(\hat{\mathbf{q}}^r(0)-\mathbf{q}^r(0)+\mathbf{q}^r(0)-\mathbf{q}(0)\big)\cdot\mathbf{w}\\ &= \theta_1(r)+\theta_1'(r),\end{aligned}$$

where $\theta_1'(r)=(\mathbf{q}^r(0)-\mathbf{q}(0))\cdot\mathbf{w}=\boldsymbol{\varepsilon}_1(r)\cdot\mathbf{w}$, which is $o(1)$ since $|\boldsymbol{\varepsilon}_1(r)|=o(1)$. From (83), $\theta_1(r)=o(1)$, and hence

$$\theta_2(r)=o(1). \tag{89}$$

Finally, since $\mathbf{q}(0)$ solves $\mathsf{opt}(\mathbf{q}(0))$, we have that

$$\mathsf{entr}\big(\mathbf{q}(0)\big)=\mathcal{O}(\mathbf{0},0), \tag{90}$$

where $\mathbf{0}$ is the vector of all 0s. Finally, we claim that since $|\boldsymbol{\varepsilon}_1(r)|\to 0$ and $\theta_2(r)\to 0$ as $r\to\infty$ (i.e., both are $o(1)$ terms),

$$\lim_{r\to\infty}\big|\mathcal{O}(\mathbf{0},0)-\mathcal{O}\big(\boldsymbol{\varepsilon}_1(r),\theta_2(r)\big)\big|=0, \tag{91}$$

or equivalently, $\mathcal{O}(\boldsymbol{\varepsilon}_1(r),\theta_2(r))=\mathcal{O}(\mathbf{0},0)+o(1)$. Therefore, (85) follows from (86) and (88). Thus, establishing $\delta_1(r)\le o(r)$. Now we justify (91). Note that optimization problem (87) has strictly convex objective with linear constraints. The region of optimization can be easily restricted to a bounded set. Therefore, it achieves minimum and this optimal solution is unique. Let $\mathbf{x}(r)$ denote this solution for optimization problem with $\boldsymbol{\varepsilon}_1(r)$, $\theta_2(r)$; $\mathsf{entr}(\mathbf{x}(r))=\mathcal{O}(\boldsymbol{\varepsilon}_1(r),\theta_2(r))$ as defined earlier. As stated in (90), $\mathbf{q}(0)=\mathbf{q}$ is the solution to optimization problem with $\mathbf{0},0$ since it solves $\mathsf{opt}(\mathbf{q})$; $\mathsf{entr}(\mathbf{q})=\mathcal{O}(\mathbf{0},0)$. Now consider $\mathbf{q}(0)+\boldsymbol{\varepsilon}_1(r)=\mathbf{q}^r(0)$. Then by definition it satisfies constraints of optimization problem (87) with $\boldsymbol{\varepsilon}_1(r),\theta_2(r)$. That is,

$$\mathsf{entr}\big(\mathbf{q}^r(0)\big)\ge\mathsf{entr}\big(\mathbf{x}(r)\big). \tag{92}$$

By definition $\mathbf{q}^r(0)\to\mathbf{q}$, and hence $\mathsf{entr}(\mathbf{q}^r(0))\to\mathsf{entr}(\mathbf{q})$. It can be argued that $\mathbf{x}(r)$ is always in a bounded, compact set. Hence, $\mathbf{x}(r)$ has limit points which all by the fact that $|\boldsymbol{\varepsilon}_1(r)|\to 0$ and $\theta_2(r)\to 0$, satisfy feasibility conditions of optimization problem (87) for $\mathbf{0}$ and 0. Therefore, from the above discussion, it follows that

$$\begin{aligned}\lim_{r\to\infty}\mathsf{entr}\big(\mathbf{x}(r)\big)&=\mathsf{entr}(\mathbf{q})\\ &=\lim_{r\to\infty}\mathsf{entr}\big(\mathbf{q}^r(0)\big).\end{aligned} \tag{93}$$

That is, $\mathcal{O}(\boldsymbol{\varepsilon}_1(r),\theta_2(r))\to\mathcal{O}(\mathbf{0},0)$ as $r\to\infty$. This complete the justification of (91) and hence the proof of $\delta_1(r)\le o(r)$.

2. *Proof of* $\delta_2(r)\le o(r)$. Since $\mathrm{CON}_r\subset E_r$, from (45) it follows that for any $x^r\in E_r$,

$$\begin{aligned}L^r\big(r\mathbf{q}^r(t)\big)-L^r\big(r\mathbf{q}^r(0)\big)&\le\frac{Nrt}{G_r}+\frac{8N\mathbf{w}^{\max}}{G_r}\\ &\quad+8N\mathbf{w}^{\max}\sqrt{rt\log r}\,\log\left(\frac{\mathbf{w}^{\max}(r\mathbf{q}^r(0)\cdot\mathbf{1}+rt+1)}{G_r}+1\right)\\ &\le\frac{Nrt}{G_r}+O\Big(\sqrt{r\log^3 r}\Big)\\ &\le o(r),\end{aligned} \tag{94}$$

for $t \in [0, T]$. This establishes that $\delta_2(r) \leq o(r)$.

3. *Proof of* $\delta_3(r) \geq -o(1)$. We wish to show that

$$\sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi^r_{\boldsymbol{\xi}} \bigl(\mathbf{q}^r(t) \cdot \boldsymbol{\xi} - \mathbf{q}^r(0) \cdot \boldsymbol{\xi}\bigr) \geq -o(1).$$

By assumption $x^r(\cdot) \in E_r$ and $E_r \subset \mathrm{ARR}^r$. That is, under this event the following component-wise inequality holds: for any $t \in [0, T]$

$$\mathbf{a}^r(t) - \boldsymbol{\lambda}^r t \geq -\varepsilon(r)\mathbf{1}, \tag{95}$$

with $\varepsilon(r) = \Theta(\sqrt{\frac{1}{r}\log r})$. The dynamics of $\mathbf{q}^r(\cdot)$ implies that

$$\mathbf{q}^r(t) = \mathbf{q}^r(0) + \mathbf{a}^r(t) - \sum_{\boldsymbol{\pi}} s^r_{\boldsymbol{\pi}}(t)\boldsymbol{\pi} + \mathbf{z}^r(t).$$

Therefore,

$$\begin{aligned}\mathbf{q}^r(t) &= \mathbf{q}^r(0) + \mathbf{a}^r(t) - \sum_{\boldsymbol{\pi}} s^r_{\boldsymbol{\pi}}(t)\boldsymbol{\pi} + \mathbf{z}^r(t)\\ &\geq \mathbf{q}^r(0) + \mathbf{a}^r(t) - \sum_{\boldsymbol{\pi}} s^r_{\boldsymbol{\pi}}(t)\boldsymbol{\pi}.\end{aligned}$$

Hence, for any $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$

$$\boldsymbol{\xi} \cdot \mathbf{q}^r(t) \geq \boldsymbol{\xi} \cdot \mathbf{q}^r(0) + \boldsymbol{\xi} \cdot \mathbf{a}^r(t) - \sum_{\boldsymbol{\pi}} s^r_{\boldsymbol{\pi}}(t)\boldsymbol{\xi} \cdot \boldsymbol{\pi}. \tag{96}$$

From (95), (96), fact that $\boldsymbol{\xi} \cdot \boldsymbol{\pi} \leq 1$ for all $\boldsymbol{\pi} \in \mathcal{S}$, we have

$$\boldsymbol{\xi} \cdot \mathbf{q}^r(t) - \boldsymbol{\xi} \cdot \mathbf{q}^r(0) \geq \bigl(\boldsymbol{\xi} \cdot \boldsymbol{\lambda}^r - 1\bigr)t - N\varepsilon(r)\boldsymbol{\xi} \cdot \mathbf{1}. \tag{97}$$

Recall that $\boldsymbol{\lambda}^r = \boldsymbol{\lambda} - \Gamma/r$. Since $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$, or equivalently $\boldsymbol{\xi} \cdot \boldsymbol{\lambda} = 1$,

$$\boldsymbol{\xi} \cdot \boldsymbol{\lambda}^r \geq 1 - \frac{1}{r}\boldsymbol{\xi} \cdot \Gamma. \tag{98}$$

Since $\mathcal{S}^*(\boldsymbol{\lambda}) \subset \mathbb{R}^N_+$ is a finite set and $\varepsilon(r) = \Theta(\sqrt{\frac{1}{r}\log r})$, it follows that for some constant K_1 (dependent on $\mathcal{S}^*(\boldsymbol{\lambda})$, N),

$$\boldsymbol{\xi} \cdot \mathbf{q}^r(t) - \boldsymbol{\xi} \cdot \mathbf{q}^r(0) \geq -K_1\sqrt{\frac{\log r}{r}}. \tag{99}$$

Next, we shall establish that for any $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$, $|\boldsymbol{\phi}^r| = O(\log r)$. This along with (99) will suffice to conclude that $\delta_3(r) \geq -o(1)$. For each $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$, there must exist n such that $\xi_n > 0$, or else $\boldsymbol{\xi} = \mathbf{0}$ which is of no interest (we will not have such a $\boldsymbol{\xi}$ in consideration from the beginning). By (76), we have

$$\mathsf{LOG}^r_n\bigl(r\hat{q}^r_n(0)\bigr) = \sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi^r_{\boldsymbol{\xi}}\xi_n. \tag{100}$$

Now $\hat{\mathbf{q}}^r(0)$ is bounded (in terms of N, $\mathbf{q}(0) = \mathbf{q}$), it follows that the LHS of (100) is $O(\log r)$. Since $\mathcal{S}^*(\boldsymbol{\lambda})$ is finite, we conclude that $\phi^r_{\boldsymbol{\xi}} = O(\log r)$. Since this is true for any $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$, we conclude that $|\boldsymbol{\phi}^r| = O(\log r)$. From this and (99), it follows that

$$\delta_3(r) = \sum_{\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})} \phi^r_{\boldsymbol{\xi}} \big(\mathbf{q}^r(t) \cdot \boldsymbol{\xi} - \mathbf{q}^r(0) \cdot \boldsymbol{\xi} \big)$$
$$\geq -K_2 \sqrt{\frac{\log^3 r}{r}}, \tag{101}$$

for some constant K_2 (dependent on $\mathbf{q}$, N). That is, $\delta_3(r) \geq -o(1)$ as desired. This completes the proof of all three claims and that of Lemma 4. □

Proof of Lemma 5 Given $\boldsymbol{\phi}^r$ and $\mathbf{q}^r(0)$, define function $M : \mathbb{R}^N_+ \to \mathbb{R}$ as

$$M(\mathbf{Y}) = \mathrm{La}^r \big(\mathbf{Y}, \boldsymbol{\phi}^r; r\mathbf{q}^r(0) \big). \tag{102}$$

By second-order Taylor's expansion, it follows that for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^N_+$,

$$M(\mathbf{Y}) = M(\mathbf{X}) + \nabla M(\mathbf{X})(\mathbf{Y} - \mathbf{X}) + \frac{1}{2}(\mathbf{Y} - \mathbf{X})^{\mathrm{T}} \nabla^2 M(\mathbf{Z})(\mathbf{Y} - \mathbf{X}), \tag{103}$$

where $\mathbf{Z} = \alpha \mathbf{X} + (1 - \alpha)\mathbf{Y}$ for some $\alpha \in [0, 1]$. Our interest is in choice of $\mathbf{X} = r\hat{\mathbf{q}}^r(0)$ and $\mathbf{Y} = r\mathbf{q}^r(t)$. From (76),

$$\nabla M \big(r\hat{\mathbf{q}}^r(0) \big) = \mathbf{0}. \tag{104}$$

And from the form of La^r, it follows that $\nabla^2 M(\mathbf{Z})$ is an $N \times N$ diagonal matrix with nth entry of diagonal as

$$\nabla^2 M(\mathbf{Z})_{nn} = \frac{w_n^2}{w_n Z_n + G_r}. \tag{105}$$

From (104) and (105), it follows that for some $r\mathbf{z} = \alpha \hat{\mathbf{q}}^r(0) + (1 - \alpha)\mathbf{q}^r(t)$, $\alpha \in [0, 1]$,

$$M\big(r\mathbf{q}^r(t) \big) = M\big(r\hat{\mathbf{q}}^r(0) \big) + \frac{1}{2} \sum_n \frac{r(q^r_n(t) - \hat{q}^r_n(0))^2 w_n^2}{w_n z_n + \frac{G_r}{r}}$$
$$\geq M\big(r\hat{\mathbf{q}}^r(0) \big) + K_3 r \big\| \mathbf{q}^r(t) - \hat{\mathbf{q}}^r(0) \big\|_2^2, \tag{106}$$

where K_3 is a constant that depends on $\mathbf{q}(0)$, N and time interval length T. That is,

$$\big\| \mathbf{q}^r(t) - \hat{\mathbf{q}}^r(0) \big\|_2^2 \leq \frac{|M(r\mathbf{q}^r(t)) - M(r\hat{\mathbf{q}}^r(0))|}{K_3 r}. \tag{107}$$

This completes the proof of Lemma 5. □

4.6.2 $\mathbf{q}$ *is a fixed point* $\Rightarrow$ $\mathbf{q}$ *solves* $\mathsf{opt}(\mathbf{q})$

Here, we wish to establish the other side of Theorem 3: if $\mathbf{q}$ be such that starting with $\mathbf{q}(0) = \mathbf{q}$, we have $\mathbf{q}(t) = \mathbf{q}$ for all $t \in [0, T]$, then $\mathbf{q}$ must solve $\mathsf{opt}(\mathbf{q})$. As before, we have $\|\mathbf{q}^r(\cdot) - \mathbf{q}(\cdot)\| \to 0$. That is $\mathbf{q}^r(t) \to \mathbf{q}$ for all $t \in [0, T]$ as $r \to \infty$ since $\mathbf{q}(t) = \mathbf{q}$ for all $t \in [0, T]$. To establish $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$, we shall execute the following three steps:

1. For any finite t, under event E_r,

$$\big| L^r\big(r\mathbf{q}^r(t) \big) - L^r\big(r\mathbf{q}^r(0) \big) \big| = o(r).$$

2. For any $t > 0$,

$$\left|\sum_{\tau=0}^{rt-1} \mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\mathbf{Q}^r(\tau+1)-\mathbf{Q}^r(\tau)\big)\right| = o(r),$$

where $\mathsf{LOG}^r(\mathbf{X}) = \sum_n \mathsf{LOG}_n^r(X_n)$.

3. Use 1 and 2 to establish that $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$.

Proof of Step 1 By assumption in the statement of Theorem 3 that $c\mathbf{w} \in \mathcal{S}^*(\boldsymbol{\lambda})$ for some $c > 0$ and (99), it follows that

$$\mathbf{q}^r(t)\cdot\mathbf{w} \ge \mathbf{q}^r(0)\cdot\mathbf{w} - K_6\sqrt{\frac{\log r}{r}}, \tag{108}$$

for some constant K_6. Therefore, using Proposition 2, (108), and fact that $\mathsf{entr}(\mathbf{q}^r(t)) - \mathsf{entr}(\mathbf{q}^r(0)) = o(1)$ since $\mathbf{q}^r(t), \mathbf{q}^r(0) = \mathbf{q}(0) + o(1)$, and hence $\mathbf{q}^r(\cdot)$ is in a bounded set, it follows that

$$\begin{aligned} L^r\big(r\mathbf{q}^r(t)\big) - L^r\big(r\mathbf{q}^r(0)\big) &= r\log\left(\frac{r}{eG_r}\right)\big(\mathbf{q}^r(t)-\mathbf{q}^r(0)\big)\cdot\mathbf{w} \\ &\quad + r\big(\mathsf{entr}\big(\mathbf{q}^r(t)\big) - \mathsf{entr}\big(\mathbf{q}^r(0)\big)\big) + O\big(\log^2 r\big) \\ &\ge -O\left(\sqrt{r\log^3 r}\right) + o(r) \\ &\ge -o(r). \end{aligned} \tag{109}$$

Then from (94) along with (109), Step 1 follows.

Proof of Step 2 Recall that the Lyapunov function for the rth system, $L^r(\mathbf{Q}) = \sum_n F_n^r(Q_n)$ with

$$F_n^r(x) = (w_n x + G_r)\log(w_n x + G_r) - w_n x \log G_r - (w_n x + G_r).$$

And, $dF_n^r(x)/dx = \mathsf{LOG}_n^r(x)$. By convexity of F_n^r, it follows that for any $x, y \ge 0$,

$$\mathsf{LOG}_n^r(y)(x-y) \le F_n^r(x) - F_n^r(y) \le \mathsf{LOG}_n^r(x)(x-y). \tag{110}$$

Using (110), one obtains for any $\tau \ge 0$,

$$\begin{aligned} &\sum_n \mathsf{LOG}_n^r\big(Q_n^r(\tau)\big)\big(Q_n^r(\tau+1) - Q_n^r(\tau)\big) \\ &\quad \le L^r\big(\mathbf{Q}^r(\tau+1)\big) - L^r\big(\mathbf{Q}^r(\tau)\big) \\ &\quad \le \sum_n \mathsf{LOG}_n^r\big(Q_n^r(\tau+1)\big)\big(Q_n^r(\tau+1) - Q_n^r(\tau)\big) \\ &\quad \le \sum_n \mathsf{LOG}_n^r\big(Q_n^r(\tau)\big)\big(Q_n^r(\tau+1) - Q_n^r(\tau)\big) + \frac{N\mathbf{w}^{\max}}{G_r}. \end{aligned} \tag{111}$$

Therefore, for any $t > 0$,

$$\left|\left[\sum_{\tau=0}^{rt-1} L^r\big(\mathbf{Q}^r(\tau+1)\big) - L^r\big(\mathbf{Q}^r(\tau)\big)\right] - \left[\sum_{\tau=0}^{rt-1} \mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\mathbf{Q}^r(\tau+1)-\mathbf{Q}^r(\tau)\big)\right]\right| = O\left(\frac{r}{G_r}\right) = o(r). \tag{112}$$

Using Step 1 and (112), it follows that

$$\left|\sum_{\tau=0}^{rt-1} \mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\mathbf{Q}^r(\tau+1)-\mathbf{Q}^r(\tau)\big)\right| = o(r). \tag{113}$$

This completes the proof of Step 2.

4.6.3 Proof of Step 3

Using Step 2, we wish to conclude that $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$. Suppose, to the contrary that $\mathbf{q}$ does not solve $\mathsf{opt}(\mathbf{q})$ and let $\mathbf{q}' \neq \mathbf{q}$ is its solution. Since $c\mathbf{w} \in \mathcal{S}^*(\boldsymbol{\lambda})$ for some $c > 0$ and $\boldsymbol{\xi}\cdot\mathbf{q}' \geq \boldsymbol{\xi}\cdot\mathbf{q}$ for all $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$, it must be that $\mathbf{q}'\cdot\mathbf{w} = \mathbf{q}\cdot\mathbf{w}$. Since $\mathbf{q}' \neq \mathbf{q}$, we must have $\mathsf{entr}(\mathbf{q}') < \mathsf{entr}(\mathbf{q})$. Therefore, by Proposition 2,

$$L^r(r\mathbf{q}) \geq L^r\big(r\mathbf{q}'\big) + K_7 r, \tag{114}$$

for some constant K_7. We shall use (114) to first show that there exists a feasible solution $\widetilde{\mathbf{Q}}^r = r\widetilde{\mathbf{q}}^r$ of $\mathsf{OPT}(r\mathbf{q}^r(0)) = \mathsf{OPT}(\mathbf{Q}^r(0))$ such

$$L^r\big(\mathbf{Q}^r(0)\big) \geq L^r\big(\widetilde{\mathbf{Q}}^r\big) + K_8 r. \tag{115}$$

for some constant K_8. Next, we shall use (115) and property of $\mathsf{OPT}(\cdot)$, to arrive at a contradiction using Step 2 to conclude that indeed $\mathbf{q}$ solves $\mathsf{opt}(\mathbf{q})$.

Toward that, we start by establishing (115). Define $\widetilde{\mathbf{q}}^r = \mathbf{q}' + (\mathbf{q}^r(0) - \mathbf{q})$. Clearly, $\widetilde{\mathbf{q}}^r = \mathbf{q}' + o(1)$ since $\mathbf{q}' \in \mathbb{R}_+^N$ and $\mathbf{q}^r(0) = \mathbf{q}(0) + o(1) = \mathbf{q} + o(1)$. We claim that for r large enough $\widetilde{\mathbf{q}}^r \in \mathbb{R}_+^N$. To see this, observe that the function $x\log x$ has derivative $-\infty$ at $x = 0$. Therefore, the minimum of $\mathsf{entr}(\mathbf{x}) = \sum_n x_n \log x_n$, over subset of $\mathbb{R}_+^N$ with a feasible point that has all strictly positive component, must have all of its components strictly positive. Indeed, $\mathbf{q}'$ is solution to such an optimization problem for any $\mathbf{q} \neq \mathbf{0}$. That is, $\mathbf{q}' > 0$ component-wise, and hence the nonnegativity of $\widetilde{\mathbf{q}}^r$ follows since $\widetilde{\mathbf{q}}^r = \mathbf{q}' + o(1)$.

Now for all $\boldsymbol{\xi} \in \mathcal{S}^*(\boldsymbol{\lambda})$,

$$\widetilde{\mathbf{q}}^r\cdot\boldsymbol{\xi} = \big(\mathbf{q}' - \mathbf{q}\big)\cdot\boldsymbol{\xi} + \mathbf{q}^r(0)\cdot\boldsymbol{\xi} \geq \mathbf{q}^r(0)\cdot\boldsymbol{\xi},$$

since $\mathbf{q}'$ is feasible for $\mathsf{opt}(\mathbf{q})$. Thus, it follows that $r\widetilde{\mathbf{q}}^r$ is feasible for $\mathsf{OPT}^r(r\mathbf{q}^r(0)) = \mathsf{OPT}^r(\mathbf{Q}^r(0))$. Next, we compare $L^r(\widetilde{\mathbf{Q}}^r)$ and $L^r(\mathbf{Q}^r(0))$ where $\widetilde{\mathbf{Q}}^r = r\widetilde{\mathbf{q}}^r$ and $\mathbf{Q}^r(0) = r\mathbf{q}^r(0)$.

As noted earlier, $\mathbf{q}'\cdot\mathbf{w} = \mathbf{q}\cdot\mathbf{w} = \mathbf{q}(0)\cdot\mathbf{w}$. Therefore,

$$\widetilde{\mathbf{q}}^r\cdot\mathbf{w} = \big(\mathbf{q}' - \mathbf{q}(0)\big)\cdot\mathbf{w} + \mathbf{q}^r(0)\cdot\mathbf{w} = \mathbf{q}^r(0)\cdot\mathbf{w}.$$

Clearly, $\widetilde{\mathbf{q}}^r = \mathbf{q}' + o(1)$, $\mathbf{q}^r(0) = \mathbf{q}(0) + o(1)$. Further, we have $\mathsf{entr}(\mathbf{q}') < \mathsf{entr}(\mathbf{q}) = \mathsf{entr}(\mathbf{q}(0))$. Therefore, by uniform continuity of $\mathsf{entr}(\cdot)$ over a compact set it follows that $\mathsf{entr}(\widetilde{\mathbf{q}}^r) < \mathsf{entr}(\mathbf{q}^r(0))$ for all r large enough. Therefore, from Proposition 2

$$\begin{aligned} L^r\big(r\mathbf{q}^r(0)\big) - L^r\big(r\widetilde{\mathbf{q}}^r\big) &= r\log r\big(\big(\mathbf{q}^r(0) - \widetilde{\mathbf{q}}^r\big)\cdot\mathbf{w}\big) + r\big(\mathsf{entr}\big(\mathbf{q}^r(0)\big) - \mathsf{entr}\big(\widetilde{\mathbf{q}}^r\big)\big) \\ &\quad + O\big(\log^2 r\big) \\ &\geq K_9 r, \end{aligned} \tag{116}$$

for some constant K_9. Thus, $\widetilde{\mathbf{Q}}^r = r\widetilde{\mathbf{q}}^r$ is a feasible solution of $\mathsf{OPT}(\mathbf{Q}^r(0))$ with property (116). Next, we use this to obtain contradiction with Step 2. Now using Proposition 3, it follows that there exists finite U and $\boldsymbol{\sigma} \in \Sigma$ such that

$$\widetilde{\mathbf{q}}^r = \mathbf{q}^r(0) + U(\boldsymbol{\lambda} - \boldsymbol{\sigma}) \quad \Leftrightarrow \quad \widetilde{\mathbf{Q}}^r = \mathbf{Q}^r(0) + rU(\boldsymbol{\lambda} - \boldsymbol{\sigma}). \tag{117}$$

Using convexity of L^r, we obtain

$$\begin{aligned} L^r\big(\widetilde{\mathbf{Q}}^r\big) - L^r\big(\mathbf{Q}^r(0)\big) &\geq \nabla L^r\big(\mathbf{Q}^r(0)\big)\cdot\big(\widetilde{\mathbf{Q}}^r - \mathbf{Q}^r(0)\big) \\ &= rU\big(\mathsf{LOG}^r\big(\mathbf{Q}^r(0)\big)\cdot(\boldsymbol{\lambda} - \boldsymbol{\sigma})\big). \end{aligned}$$

Using this and (116), we obtain that

$$\mathsf{LOG}^r\big(\mathbf{Q}^r(0)\big)\cdot(\boldsymbol{\lambda} - \boldsymbol{\sigma}) \leq -K_{10}, \tag{118}$$

for some constant $K_{10} > 0$. Let $\boldsymbol{\sigma}^r(0)$ be schedule chosen by the MWL policy at timeslot 0 for the rth system. Since $\boldsymbol{\sigma}^r(0)$ is suppose to be the maximum weight schedule, we have that

$$\mathsf{LOG}^r\big(\mathbf{Q}^r(0)\big)\cdot\big(\boldsymbol{\lambda} - \boldsymbol{\sigma}^r(0)\big) \leq -K_{10} < 0. \tag{119}$$

The basic premise has been that $\mathbf{q}$ is a fixed point. That is, $\mathbf{q}(t) = \mathbf{q}$ if $\mathbf{q}(0) = \mathbf{q}$. That is, $\mathbf{q}^r(t) = \mathbf{q}(t) + o(1) = \mathbf{q} + o(1)$ for any t (with error term $o(1)$ being uniformly applicable to all t in a given bounded time interval). Therefore, exactly the same argument as that used to obtain (119) will lead to the following: for given $t > 0$, for any $0 \leq \tau \leq rt$ with $\tau \in \mathbb{Z}_+$, we have that

$$\mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\boldsymbol{\lambda} - \boldsymbol{\sigma}^r(\tau)\big) \leq -K_{10} < 0. \tag{120}$$

The uniform choice of K_{10} is possible since choice of U, as per Proposition 3, in (117) can be bounded uniformly given that $\mathbf{q}'$ always comes from a bounded set (depending on initial $\mathbf{q}$, T, N) and uniform convergence of $\mathbf{q}^r(\cdot) \to \mathbf{q}(\cdot)$ (over a given bounded time interval). From this, $\boldsymbol{\lambda}^r = \boldsymbol{\lambda} - \Gamma/r$ and $\mathsf{LOG}^r(\mathbf{Q}^r(\cdot)) = O(\log r)$, it follows that

$$\sum_{\tau=0}^{rt-1} \mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\boldsymbol{\lambda} - \boldsymbol{\sigma}^r(\tau)\big) \leq -K_{10}tr. \tag{121}$$

Now conclusion (121) is derived under event E_r, which holds with probability $1 - 1/r^2$. The term on the left in the above equation is at most $O(r\log r)$ due to Lipschitz property of queue-size of $[0, rt]$. Therefore, it follows that

$$\mathbb{E}\left[\sum_{\tau=0}^{rt-1} \mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\boldsymbol{\lambda} - \boldsymbol{\sigma}^r(\tau)\big)\right] \leq -K_{11}r, \tag{122}$$

for some constant $K_{11} > 0$.

Next, we shall use Step 2 to argue that (122) can not hold. This will be the contradiction to our assumption that $\mathbf{q}$ does not solve $\mathsf{opt}(\mathbf{q})$. Toward that, using arguments similar to those used in derivation of (18) from (17), we obtain that

$$\begin{aligned}&\mathbb{E}\left[\sum_{\tau=0}^{rt-1}\mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\mathbf{Q}^r(\tau+1)-\mathbf{Q}^r(\tau)\big)\right]\\&\quad=\mathbb{E}\left[\sum_{\tau=0}^{rt-1}\mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\boldsymbol{\lambda}^r-\boldsymbol{\sigma}^r(\tau)\big)\right].\end{aligned}\tag{123}$$

Now Step 2 suggests that under event E_r,

$$\left|\sum_{\tau=0}^{rt-1}\mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\mathbf{Q}^r(\tau+1)-\mathbf{Q}^r(\tau)\big)\right|=o(r).$$

Since event E_r holds with probability at least $1-1/r^2$ and since the term on the left in the above equation is at most $O(r\log r)$ due to Lipschitz property of queue-size of $[0,rt]$, it follows that

$$\left|\mathbb{E}\left[\sum_{\tau=0}^{rt-1}\mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\mathbf{Q}^r(\tau+1)-\mathbf{Q}^r(\tau)\big)\right]\right|=o(r).\tag{124}$$

From (123) and (124), it follows that

$$\left|\mathbb{E}\left[\sum_{\tau=0}^{rt-1}\mathsf{LOG}^r\big(\mathbf{Q}^r(\tau)\big)\cdot\big(\boldsymbol{\lambda}^r-\boldsymbol{\sigma}^r(\tau)\big)\right]\right|=o(r).\tag{125}$$

Indeed, (122) and (125) contradict our assumption that $\mathbf{q}$ does not solve $\mathsf{opt}(\mathbf{q})$. This concludes the proof of the Theorem 3.

5 Discussion

In this paper we introduced and studied the MWL policy, a variant of the well-studied maximum weight policy of Tassiulas and Ephremides. The MWL policy utilizes logarithm of queue-sizes as weight to select a maximum weight schedule. The primary motivation to study this policy is the authors' belief that it exhibits work-conservation property under the heavy traffic approximation. While this paper stops short of establishing this property, the work-conservation property is established for its fluid model. As a step toward establishing heavy traffic approximation, we have identified the invariant manifold of the critical fluid model (under additional restrictions). The invariant manifold corresponds to solution space of a two-stage optimization problem. The form of this optimization problem hints toward the work-conservation property being true, at least with certain restrictions, under heavy traffic approximation.

The MWL policy, due to lack of scale invariance of logarithm weight function, presents a serious challenge to the existing approaches for identifying a meaningful fluid model and subsequently in establishing a heavy traffic approximation. Results of this paper overcome this challenge for the specific problem at hand. Going forward, it would be worth developing a general method to deal with such scenarios. Obtaining the heavy traffic approximation and subsequently establishing work-conservation property of a switched network operating under the MWL policy remain outstanding problems.

Acknowledgements DS is supported by NSF CAREER CNS-0546590 and NSF Collaborative project on Flow Level Models. DJW is supported by a Royal Society university research fellowship. We are grateful for further support from the British Council Researcher Exchange program, and the Newton Institute programme on Stochastic Processes in Communication Sciences.

References

1. Andrews, M., Kumaran, K., Ramanan, K., Stolyar, S., Vijayakumar, R., Whiting, P.: Scheduling in a queueing system with asynchronously varying service rates. Probab. Eng. Inf. Sci. (2001)
2. Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Math. J. **19**, 357–367 (1967)
3. Bertsekas, D., Nedic, A., Ozdaglar, A.: Convex Analysis and Optimization. Athena Scientific, Belmont (2003)
4. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley, New York (1999)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
6. Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. Queueing Syst. **30**, 89–148 (1998)
7. Dai, J.G.: Stability of fluid and stochastic processing networks. MaPhySto Lecture Notes (1999). http://www.maphysto.dk/cgi-bin/gp.cgi?publ=70
8. Dai, J.G., Prabhakar, B.: The throughput of switches with and without speed-up. In: Proceedings of IEEE Infocom, pp. 556–564 (2000)
9. Harrison, J.M.: The bigstep approach to flow management in stochastic processing networks. In: Stochastic Networks: Theory and Applications, p. 57–90 (1996)
10. Harrison, J.M.: Brownian models of open processing networks: canonical representation of workload. Ann. Appl. Probab. **10**, 75–103 (2000) Also see [11], http://projecteuclid.org/euclid.aoap/1019737665
11. Harrison, J.M.: Correction to Harrison, J. M. (2000). Brownian models of open processing networks: canonical representation of workload. Annals Applied Probab. **10**, 75–103 (2000). Ann. Appl. Probab. **13**, 390–393 (2003)
12. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
13. Kelly, F.P., Williams, R.J.: Fluid model for a network operating under a fair bandwidth-sharing policy. Ann. Appl. Probab. **14**, 1055–1083 (2004)
14. Lin, W., Dai, J.G.: Maximum pressure policies in stochastic processing networks. Oper. Res. **53**, 197–218 (2005)
15. Meyn, S., Tweedie, R.: Markov Chains and Stochastic Stability. Springer, New York (1993)
16. Shah, D., Wischik, D.J.: Optimal scheduling algorithms for input-queued switches. In: Proceedings of IEEE Infocom (2006)
17. Shah, D., Wischik, D.J.: Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse. Ann. Appl. Probab. **22**(1), 70–127 (2012)
18. Stolyar, A.L.: MaxWeight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. Ann. Appl. Probab. **14**(1), 1–53 (2004)
19. Tassiulas, L., Ephremides, A.: Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. IEEE Trans. Autom. Control **37**, 1936–1948 (1992)
20. Williams, R.J.: Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. Queueing Syst. **30**, 27–88 (1998)