THE TELEOLOGY OF SCHEDULING ALGORITHMS FOR SWITCHED NETWORKS UNDER LIGHT LOAD, CRITICAL LOAD, AND OVERLOAD

BY DEVAVRAT SHAH* AND DAMON WISCHIK†

MIT and UCL

We consider a queueing network in which there are constraints on which queues may be served simultaneously; such networks may be used to model input-queued switches and wireless networks. The scheduling algorithm for such a network specifies which queues to serve at any point in time. We analyse a family of scheduling algorithms, related to the maximum-weight algorithm of Tassiulas and Ephremides [32]. We establish a fluid limit and analyse it under light load, critical load and overload. In all three regimes we characterize the invariant states as solutions to a network-wide optimization problem, and show that the system converges to an invariant state whatever the initial condition. Under light load, we recover known stability results. Under critical load, we obtain multiplicative state space collapse; related results for a bandwidth-sharing model of the Internet were obtained by Kelly and Williams [17]. For overload, the characterization of invariant states and the convergence result are both new. The light-load and critical-load results apply to single-hop and multihop networks; the overload results only apply to single-hop networks. We use these results to demonstrate a scheduling algorithm whose fluid model is near-optimal under all three loading regimes. Based on this, we conjecture an optimal algorithm.

1. Introduction.

1.1. Explanation of ‘switched network’. Consider a collection of queues operating in discrete time. In each timeslot, queues are offered service according to a service schedule chosen from a specified finite set. For example, in a three-queue system, the set of allowed schedules might consist of “Serve 3 units of work each from queues A & B” and “Serve 1 unit of work each from queues A & C, and 2 units from queue B”. New work may arrive in each timeslot; let each queue have a dedicated exogenous arrival process, with specified mean arrival rates. Once work is served,

*Supported by NSF CNS-0546590
†Supported by a Royal Society university research fellowship. Collaboration supported by the British Council Researcher Exchange program.
AMS 2000 subject classifications: 60K25, 60K30, 90B36
Keywords and phrases: switch, input-queued switch, scheduling, maximum weight matching, fluid models, heavy traffic
it leaves the network. Switched networks are closely related to what Harrison [12] calls ‘stochastic processing networks’.

The switched network model has been used to describe a wireless network in which radio interference limits the amount of service that can be given to each host. It has been used to describe an input-queued switch, the device at the heart of high-end Internet routers, whose underlying silicon architecture imposes constraints on which traffic streams can be transmitted simultaneously. It can be extended to describe the overall operation of congestion control in the Internet, whereby TCP assigns transmission rates to flows, subject to network constraints about the capacity on shared links. It might even describe road junctions, where the schedule of which lanes can move is controlled by traffic lights.

1.2. Outline of paper. We give the name scheduling algorithm to the procedure whereby a schedule is chosen for each timeslot. In this paper we study how the scheduling algorithm affects the behaviour of the network; we do this by finding an optimization problem that characterizes the behaviour. Teleology is a philosophical term meaning ‘the study of the end or purpose for which a thing is done, especially as related to the evidence of design in natural phenomena’; we use the term to refer to an algorithm’s implicit optimization problem.

We shall discuss the behaviour of the network in three regimes: light load, critical load, and overload. The mathematical tools we will use operate at three different scales of description: at the ‘microscopic’ level we work with a probabilistic characterization of the detailed packet-level behaviour of the system, and relate it to fluid models (i.e. dynamical system models); at the ‘macroscopic’ level we analyse the dynamics of these fluid models; and at the ‘teleological’ level we characterize the steady-state behaviour of the fluid models by means of optimization problems. The sections of the main body of this paper are each flagged as microscopic, macroscopic or teleological. Teleology can be read on its own. Macroscopic work relies on teleology, and microscopic work relies on both macroscopic and teleology. The sections are as follows:

§3 (teleology). Optimization problems for describing the capacity of a switched network, used to characterize the difference between light load, critical load, and overload

§4 (macroscopic). The fluid model, and some remarks about light load

§5 (microscopic). Derivation of the fluid equations from a detailed probabilistic model

§6–7 (teleology & macroscopic). An emergent optimization problem called the ‘lifting map’, and how it relates to fixed points of the critically-loaded fluid model

§8 (microscopic). A limit theorem describing the system in critical load, via the lifting map
§ 9 (teleology). How to use the lifting map to infer performance in critical load
§ 10 (teleology & macroscopic). An optimization problem that emerges from the
overloaded fluid model
§ 11 (microscopic, macroscopic & teleology). Rerun of the analysis of critical load,
this time for a multihop network

In the remainder of this section we will explain precisely our network model, and
explain in detail the input-queued switch model. In Section 2 we summarize previous
work in this area, and explain our contribution. The main body of the paper follows.
Then in Section 12 we summarize our findings about the performance of scheduling
algorithms under light load, critical load, and overload, and use these to conjecture
an optimal algorithm. We conclude in Section 13 by outlining some problems for
further study, and suggesting other systems where switched network models might
be useful.

1.3. First some notation. Let $\mathbf{1}_{\{\cdot\}}$ be the indicator function, $\mathbf{1}_{\text{true}} = 1$ and $\mathbf{1}_{\text{false}} = 0$. Let $x \land y = \min(x, y)$ and $x \lor y = \max(x, y)$ and $[x]^+ = x \lor 0$. When $x$ is a vector,
the maximum is taken componentwise. We will reserve bold letters for vectors in
$\mathbb{R}^N$, where $N$ is the number of queues, for example $\mathbf{x} = [x_n]_{1 \leq n \leq N}$. Let $\mathbf{0}$ be
the vector of all 0s, and $\mathbf{1}$ be the vector of all 1s. Use the norm $|\mathbf{x}| = \max_n |x_n|$. For
vectors $\mathbf{u}$ and $\mathbf{v}$ and functions $f : \mathbb{R} \to \mathbb{R}$, let
$$
\mathbf{u} \cdot \mathbf{v} = \sum_{n=1}^{N} u_n v_n, \quad \mathbf{u} \mathbf{v} = [u_n v_n]_{1 \leq n \leq N}, \quad \text{and} \quad f(\mathbf{u}) = [f(u_n)]_{1 \leq n \leq N}
$$
and let matrix multiplication take precedence over dot product so that
$$
\mathbf{u} \cdot A \mathbf{v} = \sum_{n=1}^{N} u_n \left( \sum_{m=1}^{N} A_{nm} v_m \right).
$$
Let $A^T$ be the transpose of matrix $A$. 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 $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$.

1.4. Switched network model. Consider a collection of $N$ queues. Let time be
discrete: timeslot $\tau \in \{0, 1, \ldots \}$ runs from time $\tau$ to $\tau + 1$. Let $Q_n(\tau)$ be the amount
of work in queue $n \in \{1, \ldots, N\}$ at the beginning of timeslot $\tau$, and write $\mathbf{Q}(\tau)$ for
the vector $[Q_n(\tau)]_{1 \leq n \leq N}$. Let $\mathbf{Q}(0)$ be the prespecified vector of initial queue sizes.

In each timeslot, each queue is offered service subject to a scheduling constraint
described below. If the amount of service $\pi_n$ offered to a queue is greater than the
current queue size $Q_n(\tau)$ then we say that the queue has idled by $\pi_n - Q_n(\tau)$,
otherwise it does not idle. Once work is served it 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 \( S \subset \mathbb{R}_+^N \). In every timeslot a schedule \( \pi \in S \) is chosen; queue \( n \) is offered an amount of service \( \pi_n \).

Let \( S_\pi(\tau) \) be the total number of timeslots up to the beginning of timeslot \( \tau \) in which schedule \( \pi \) has been chosen, and let \( S_\pi(0) = 0 \). Let \( Z_n(\tau) \) be the total amount of idling at queue \( n \) up to the beginning of timeslot \( \tau \), 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 \( \tau \), and \( A_n(0) = 0 \). We will use the convention that \( Q(\tau) \) is the vector of queue sizes at the beginning of timeslot \( \tau \), and then the schedule for timeslot \( \tau \) is chosen and service happens, and then arrivals for timeslot \( \tau \) happen. Thus, with \( \Sigma(\tau) = \sum_{\pi \in S} \pi S_\pi(\tau) \),

\[
Q_n(\tau) = Q_n(0) + A_n(\tau) - \Sigma_n(\tau) + Z_n(\tau) \\
Z_n(\tau + 1) - Z_n(\tau) = \left[ \Sigma_n(\tau + 1) - \Sigma_n(\tau) - Q_n(\tau) \right]^+.
\]

We will take \( A(\cdot) \) to be a random process with stationary increments and a well-defined mean arrival rate vector \( \lambda \), i.e. we assume \( \lim_{\tau \to \infty} A_n(\tau) / \tau \) exists almost surely and is deterministic for every queue \( 1 \leq n \leq N \), and we define

\[
\lambda_n = \lim_{\tau \to \infty} \frac{1}{\tau} A_n(\tau).
\]

We assume for convenience that \( A(\cdot) \) has bounded increments, i.e. there exists some \( A^{\text{max}} \in \mathbb{R}_+ \) such that

\[
A_n(\tau + 1) - A_n(\tau) \in [0, A^{\text{max}}] \quad \text{for all} \quad n, \tau.
\]

We will also need to assume that the arrival process is close to its mean, in the sense that there is a non-increasing function \( R : \mathbb{R}_+ \to \mathbb{R}_+ \) such that

\[
P \left( \sup_{\tau \leq r} |A(\tau) - \lambda \tau| < \varepsilon r \right) = 1 - o(R(r)) \quad \text{as} \quad r \to \infty, \quad \text{for all} \quad \varepsilon > 0.
\]

The \( R \) function controls the chance of large fluctuations in the arrival process, and it thereby controls the chance of large fluctuations in all other parts of the system. For the fluid limit in Section 5 all we need is \( R(r) = O(1) \); for the critical-load limit in Section 8 we will want \( R(r) = O(1/r) \). Assumption (5) holds for a wide range of input processes with i.i.d. increments, e.g. when the arrivals to queue \( n \) form a Bernoulli process independent of all other queues; it holds with \( R(r) = 1 \) by the weak law of large numbers, or with \( R(r) = 1/r \) by a simple application of Chebyshev’s inequality\(^1\).

\(^1\)cf. Bramson [4], Proposition 4.3
1.5. Motivating example: input-queued switch. An Internet router has several input ports and output ports. A data transmission cable is attached to each of these ports. Packets arrive at the input ports. The function of the router is to work out which output port each packet should go to, and to transfer packets to the correct output ports. This last function is called switching. There are a number of possible switch architectures; we will consider the commercially popular input-queued switch architecture.

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’s 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 $1 \leq n \leq N$.)

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, at each timeslot the switch can choose a matching from inputs to outputs. The schedule $\pi \in \mathbb{R}_{+}^{3 \times 3}$ 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. Clearly $\pi$ is a permutation matrix, and the set $\mathcal{S}$ of feasible schedules is the set of $3 \times 3$ permutation matrices.

Figure 1 shows two possible matchings. In the left hand figure, the matching allows a packet to be transmitted from input port 3 to output port 2, but since $Q_{3,2}$ is empty no packet is actually transmitted. The specific matching of inputs to outputs in each timeslot is chosen by the scheduling algorithm. It may take account of queue sizes, ages of packets, or other quality-of-service constraints. The goal of the scheduling algorithm is to make good use of the switch’s capacity.

Some notes on modelling. (1) We have illustrated a switch with as many inputs...
as outputs. It may be that some of these do not actually carry any traffic; thus there is no loss of generality in assuming as many inputs as outputs. (2) In the Internet, packets may have different sizes. Before the packet reaches the switch fabric, it will be fragmented into a collection of smaller packets (called cells) of fixed size. (3) There will typically be a block of memory at each input port for the queues, and one packet’s worth of memory at each output port to hold the packet as it is serialized onto the outgoing cable. Memory access speeds are a limiting factor, and indeed the time it takes to read or write a packet from memory is what determines the length of a timeslot. There are switches which perform several matchings per timeslot—but then the timeslots need to last several times longer, to give time for the extra reads and writes.

2. Previous work, and our contribution.

2.1. Scheduling algorithms. The scheduling algorithm that has received the most attention is Maximum-Weight scheduling, which we will call MW, introduced by Tassiulas and Ephremides [32]. It works as follows. Let $Q(\tau)$ be the vector of queue sizes at the beginning of timeslot $\tau$. Define the weight of a schedule $\pi \in S$ to be $\pi \cdot Q(\tau)$. The algorithm then chooses for timeslot $\tau$ a scheduling with the greatest weight (breaking ties arbitrarily). This algorithm can be generalized to choose a schedule which maximizes $\pi \cdot Q(\tau)^\alpha$, where the exponent is taken componentwise for some $\alpha > 0$; call this the MW-$\alpha$ algorithm. In this paper, we will study the MW-$\alpha$ algorithm in detail. More generally, one could choose a schedule $\pi$ such that

\begin{equation}
\pi \cdot f(Q(\tau)) = \max_{\rho \in S} \rho \cdot f(Q(\tau))
\end{equation}

for some function $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$; call this the MW-$f$ algorithm. (The function $f$ needs to satisfy an assumption spelled out in Section 4.)

The MW-$f$ algorithm is myopic, i.e. it chooses a schedule based only on the current queue sizes and doesn’t need to try to learn traffic parameters etc. An important reason for the popularity of the MW algorithm is that MW-$f$ is the only class of myopic scheduling algorithms known to have the largest possible stability region, for a large class of constrained scheduling problems.

In the setting of bandwidth sharing in the Internet, described in detail in Section 13, the name rate allocation algorithm is used instead of scheduling algorithm; the set of possible rates/schedules is infinite, and has special structure. The algorithms that have received the most attention are from the family of $\alpha$-fair rate allocations, described by Mo and Walrand [25]. This family of algorithms arises from models for congestion control, and is not directly related to the MW-$f$ family.
2.2. Analysis. To date, there have been four main modes of analysis applied to scheduling algorithms for switched networks.

The first mode of analysis is to formulate a Markov model, and to calculate the stability region, namely the set of arrival rates for which an algorithm is stable. This is what we mean by the ‘light load regime’. Tassiulas and Ephremides [32] showed that the stability region for MW is optimal, in that it contains the stability region for any other algorithm. Their work was in the context of a multihop wireless network. McKeown et al. [23] independently proposed the algorithm, and proved the same result for an input-queued switch. These are remarkable finding given that the algorithm uses no information other than current queue lengths. Bonald and Massoulie [3] and de Veciana et al. [7] have studied the stability region for the switched-network model of bandwidth sharing in the Internet.

The second method of analysis is also concerned with finding the stability region, but uses fluid models instead of Markov techniques. There is a large body of systematic work on fluid modeling; see for example the lecture notes of Dai [6]. Using fluid models, one obtains results for arbitrary arrival processes, assuming only a strong law of large numbers. In this setting Dai and Prabhakar [5] showed that the stability region for MW is as large as possible, in an input-queued switch, and Keslassy and McKeown [18] and Shah [27] independently extended this to apply to the MW-α algorithm; Andrews et al. [1] proved the same result for a wireless network. Eryilmaz and Srikant [9] analyse a fluid model for MW combined with rate allocation, for a wireless network.

Stability analysis leaves some questions unanswered. The MW-α family of algorithms, as α varies in (0, ∞), all have the optimal stability region—but Keslassy and McKeown [18] found from extensive simulations of an input-queued switch that the average queueing delay is very different. They conjecture:

**Conjecture 2.1** For an input-queued switch running the MW-α algorithm, The average queueing delay decreases as α decreases.

The third mode of analysis attempts to understand average delay by looking at queue size using Markov techniques, e.g. the moment bound based on the Foster–Lyapunov criterion [24]. Marsan et al. [22] used such criterion to bound the average queue size for MW, in an input-queued switch; this was extended by Shah and Kopikare [28] to algorithms which approximate MW. We are not aware of any work of this sort which addresses the above conjecture.

The fourth mode of analysis looks at queue size in the critically loaded regime, via a heavy traffic limit. This regime was first described by Kingman [19]; since then a substantial body of theory has developed, and modern treatments can be found in [4, 11, 33, 34]. A characteristic phenomenon of heavily-loaded queueing systems is that the timescale over which queue sizes change is much longer than the timescale
over which scheduling decisions are made: from the perspective of the queue size process, scheduling decisions are near-instantaneous.

Harrison [12] introduced a general model of a ‘stochastic processing network’, building on the work of Laws [20, 21]; he studies the controllability of stochastic processing networks with near-instantaneous scheduling timescale, and he describes optimization problems that are closely related to those we study in Section 3. Our switched network is a single ‘resource’ in his terminology; he interprets the work in the queues to be ‘material’, and the schedules to be ‘processing activities’. This leads to an interesting difference between his work and ours: his model does not permit idling. Equivalently, it requires the set $S$ to be monotone, by which we mean that if $\pi \in S$ is some schedule and $\rho \in \mathbb{R}^N_+$ is another schedule for which $\rho_n \in \{0, \pi_n\}$ for all $n$, then $\rho \in S$. Tassiulas and Ephremides [32] had the same requirement. This requirement plays a significant role when we come to multihop networks in Section 11, though we do not understand why.

Stolyar has studied MW-\(\alpha\) for a generalized switch model in the heavy traffic regime, and obtained a complete characterization of the limiting queue size process, under a condition known as ‘complete resource pooling’. This condition effectively requires that a clever scheduling algorithm be able to balance work between all the heavily loaded queues. Stolyar [31] showed in a remarkable paper that the limiting queue size lives in a one-dimensional state space. Operationally, this means that all one needs to keep track of is the one-dimensional total amount of work in the system (called the workload), and at any point in time one can assume that the individual queues have all been balanced.

We have previously reported on some results for a critical-load limit in an input-queued switch [29].

Kelly and Williams [17] have studied a switched-network model for bandwidth sharing, in the critically-loaded regime. They obtain fluid limits, study the behaviour of fluid limits in a critically loaded network, and conjecture a state space collapse result. Many of their results are parallel to our own, and we footnote where there are links. Kang et al. [15] prove a diffusion limit for this model.

Egorova et al. [8] have studied the bandwidth-sharing model in overload. They identify a solution to the fluid model, and conjecture that all fluid limits should converge to this solution. This is the only study of the performance of overloaded networks of which we are aware.

More generally, there are good pragmatic reasons for studying overload, as observed by Harrison and Zeevi [14]. They point out that staffing levels in a call center are typically set in advance, and that over a given period the system may be either overloaded or underloaded; it is therefore prudent to set staffing levels so that the cost is not too great in either case. Likewise a web server or an Internet switch will alternate between periods of underload and overload, hence it is useful to under-
stand the behaviour of overloaded switches. Harrison and Zeevi studied the policy problem of how to set staffing levels given uncertainty in the arrival rate, and assumed that once the arrival rate is realized then the call center is run optimally; by contrast we will take the arrival rate as given, and compare the performance of different scheduling algorithms.

2.3. Our contribution. This paper was motivated by a desire to prove the conjecture of Keslassy and McKeown about the performance of MW-\(\alpha\) as \(\alpha\) decreases. It turns out that the results of Stolyar are not sufficient, since under his ‘complete resource pooling’ condition, MW-\(\alpha\) has exactly the same performance whatever the value of \(\alpha\). For example, in an input-queued switch the algorithm simply distributes the workload equally between the critically-loaded queues, regardless of \(\alpha\).

In this paper, we remove Stolyar’s ‘complete resource pooling’ condition, and provide a critical-load analysis of a switched network running a generalized version of MW-\(\alpha\) (Theorem 8.2). We extend this analysis to a multihop switched network (Theorem 11.9.)

We find that the performance of MW-\(\alpha\) does indeed depend on \(\alpha\). In technical terms, we prove multiplicative state space collapse. We follow the method laid out by Bramson [4]; we prove a fluid limit (Theorem 5.1), and characterize fixed points of the fluid limit by means of a lifting map. (The complement of Bramson’s work is by Williams [34], and consists in proving a diffusion limit as a consequence of multiplicative state space collapse; Stolyar [31] proved this limit for a switched network under complete resource pooling, Kang and Williams [16] are currently working on the general result, and Kang et al. [15] have obtained the diffusion limit for bandwidth-sharing networks.) Our conclusions about MW-\(\alpha\) are a by-product of a teleological characterization of the system:

When the system is critically loaded, we find that the generalized version of MW-\(\alpha\) implicitly solves an optimization problem, which we call ALGD (Theorem 7.3). We call this the teleology of the algorithm.

The objective of ALGD is the Lyapunov function introduced by Tassiulas and Ephremides [32], and its constraints are closely linked to the canonical representation of workload identified by Harrison [12]. The constraints arise from a fundamental static-planning optimization problem, which finds the set of workload constraints that any scheduling algorithm must satisfy; we call this problem DUAL. Stolyar’s condition requires that DUAL have only one biting constraint.

Whereas in heavy traffic models of other systems [4, 11, 31, 34] the map from workloads to queue sizes is linear, we find instead that it is nonlinear; to be precise it yields the queue size vector that optimizes ALGD. Kelly and Williams [17] found a similar result, for a bandwidth-sharing model of congestion control. This nonlinearity
seems to capture some fundamental constraint imposed by switched networks.

Furthermore, we provide an overload analysis of a switched network running a generalized version of MW-$\alpha$. We find again that the system implicitly solves an optimization problem, which we call ALGD$^\dagger$, whose objective is the same as ALGD but whose constraints are slightly different (Theorem 10.2).

We find that there are situations where the throughput of MW decreases as the switch becomes more overloaded; we are not aware of any earlier work that has looked at overloaded switches or discovered this paradoxical behaviour. To obtain the result, we study a time-scaled version of the fluid model (what Egorova et al. [8] call a 'linear solution'), and prove that it converges to the solution of ALGD$^\dagger$. We again observe that the performance of MW-$\alpha$ depends on $\alpha$.

We characterize optimal scheduling algorithms (Theorems 12.2, 12.3 & 12.6).

The title of this paper mentions ‘optimal scheduling algorithms’. In one sense, every algorithm we consider is optimal, in that it solves a certain optimization problem! While there is considerable understanding to be gained from identifying the teleology of a scheduling algorithm, it is also useful to know how close we are to optimal performance in a conventional sense. In Section 12 we prove several results concerning optimality at the fluid scale, for critical load and overload: we find lower bounds on how far a scheduling algorithm can minimize total queue size, and we show that the total queue size for a switch running MW-$\alpha$ becomes arbitrarily close to the lower bound as $\alpha \to 0$. This suggests that at critical load the average queueing delay is near-optimal (assuming some fluid-scale version of Little’s Law applies); it also shows that at overload the average throughput is near-optimal. These results apply to a fully-loaded input-queued switch, supporting Conjecture 2.1, but they do not all extend to general switched networks. We also propose an optimal scheduling algorithm, the formal limit of MW-$\alpha$ as $\alpha \to 0$, though we have not been able to analyse it.

3. Critical workloads and virtual resources (teleology).

3.1. Admissible arrival rates. At each timeslot, a schedule $\pi \in S$ must be chosen. Let $\Sigma$ be the convex hull of $S$,

$$\Sigma = \left\{ \sum_{\pi \in S} \alpha_{\pi} \pi : \sum_{\pi \in S} \alpha_{\pi} = 1, \text{ and } \alpha_{\pi} \geq 0 \text{ for all } \pi \right\}. \tag{7}$$

We say that an arrival rate vector $\lambda$ is admissible if $\lambda \in \Lambda$ where

$$\Lambda = \left\{ \lambda \in \mathbb{R}^N_+ : \lambda \leq \sigma \text{ componentwise, for some } \sigma \in \Sigma \right\}. \tag{8}$$
Intuitively, this means that there is some combination of feasible schedules which permits all incoming work to be served. Also define

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

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

Say that $\lambda$ is strictly admissible if $\lambda \in \Lambda^\circ$, and that $\lambda$ is critical if $\lambda \in \partial \Lambda$.

3.2. Optimization problems. First consider the optimization problem PRIMAL($\lambda$):

\[
\begin{align*}
\text{minimize} & \quad \sum_{\pi \in S} \alpha_\pi \\
\text{over} & \quad \alpha_\pi \in \mathbb{R}_+ \text{ for all } \pi \in S \\
\text{such that} & \quad \lambda \leq \sum_{\pi \in S} \alpha_\pi \pi \text{ componentwise}
\end{align*}
\]

This problem asks whether it is possible to find a combination of schedules which can serve arrival rates $\lambda$; clearly $\lambda$ is admissible if and only if the solution to the primal is $\leq 1$. Now consider its dual problem DUAL($\lambda$):

\[
\begin{align*}
\text{maximize} & \quad \xi \cdot \lambda \\
\text{over} & \quad \xi \in \mathbb{R}_+^N \\
\text{such that} & \quad \max_{\pi \in S} \xi \cdot \pi \leq 1
\end{align*}
\]

The solution is clearly attained when the constraint is tight. Given a queue size vector $Q$ and any dual-feasible $\xi$ satisfying the constraint with equality, call $\xi \cdot Q$ the workload at the virtual resource $\xi$. The virtual resource specifies a combination of several actual resources (namely the queues themselves). The long-run rate at which work arrives at the virtual resource is $\xi \cdot \lambda$, and the maximum rate at which it can be served is 1. If $\xi \cdot \lambda > 1$ then the workload must increase to infinity, and if $\xi \cdot \lambda \leq 1$ then there is a chance that the workload might be stable. When there are several virtual resources all with $\xi \cdot \lambda = 1$ then it is far from clear that one can keep all those workloads stable at the same time. This is in fact possible (at least at the fluid level) by using a well-designed scheduling algorithm, as shown in Section 4.3.

Laws [20, 21] introduced primal and dual problems of this general sort for describing multihop queueing networks with routing choice. Harrison [12] extends the problems to a more general setting which encompasses switched networks. Our switched
network is a single ‘resource’ in his terminology, and our set of schedules plus all subschedules is his set of ‘activities’. (A vector $\rho \in \mathbb{R}_+^N$ is a subschedule of $\pi \in \mathcal{S}$ if for each queue $n$ either $\rho_n = \pi_n$ or $\rho_n = 0$.) Our dual constraint maps exactly to his dual constraint (2.4).

3.3. Critical workloads. Both problems are soluble so, by strong duality, the solutions to both problems are equal. Clearly the solution to the optimization problems is $\leq 1$ for any $\lambda \in \Lambda$. For $\lambda \in \Lambda^\circ$ it is $< 1$, and for $\lambda \in \partial \Lambda$ it is $= 1$. In this paper we will be especially interested in situations where the solution is $1$, i.e. where the system needs to work non-stop to keep the virtual resources from becoming overloaded. When this is so, we call the solutions to the dual problem the critically-loaded virtual resources.

Let $\mathcal{S}^* = \mathcal{S}^*(\lambda)$ be the set of all critically loaded virtual resources that are extreme points of the feasible region. Call these the principal critically-loaded virtual resources. Note that the feasible region is a polytope, therefore $\mathcal{S}^*$ is finite; and that the feasible region is convex, therefore any critically-loaded virtual resource $\zeta$ can be written

$$\zeta = \sum_{\xi \in \mathcal{S}^*} x_\xi \xi \text{ with } \sum x_\xi = 1 \text{ and all } x_\xi \geq 0.$$  \hspace{1cm} (9)

The critical workloads have a useful property. Suppose $\lambda \in \Lambda$, so $\lambda \leq \sigma$ for some $\sigma \in \Sigma$, as per the definition of $\Lambda$. Then

$$\xi_n > 0 \text{ for some critically-loaded } \xi \implies \lambda_n = \sigma_n$$  \hspace{1cm} (10)

In words, if queue $n$ is critical, then it is not possible to reduce it without increasing some other queue. To see this, pick some critically-loaded $\xi$ with $\xi_n > 0$. Then $\xi \cdot \sigma \geq \xi \cdot \lambda$ since $\sigma \geq \lambda$. Also $\xi \cdot \lambda = 1$ since $\xi$ is critical, and $\xi \cdot \sigma \leq 1$ since $\xi$ is feasible for $\text{DUAL}(\sigma)$, and $\text{PRIMAL}(\sigma) \leq 1$. Therefore there is equality, therefore $\lambda_n = \sigma_n$.

3.4. A simple example. Consider a system with $N = 2$ queues, $A$ and $B$. Suppose the set $\mathcal{S}$ of possible schedules consists of “serve three packets from queue $A$” (schedule 1) and “serve one packet each from $A$ and $B$” (schedule 2). Let $\lambda_A$ and $\lambda_B$ be the arrival rates at the two queues, measured in packets per second.

PRIMAL description. We will first derive $\Lambda$ from PRIMAL. Schedule 2 is the only action which serves queue $B$, so we need to perform schedule 2 at least $\lambda_B$ times per second. There’s no point performing schedule 2 any more than this. This allows for serving $\lambda_B$ packets per second from queue $A$, so we additionally need to perform schedule 1 at a rate of $[\lambda_A - \lambda_B]^+ / 3$ times per second. If we’re only allowed to choose
one schedule per second, we require $\lambda_B \leq 1$ and $(\lambda_A - \lambda_B)^+ / 3 + \lambda_B \leq 1$; equivalently we can replace the second inequality by $\lambda_A / 3 + 2\lambda_B / 3 \leq 1$. The two inequalities define $\Lambda$.

**DUAL description**. Define a virtual queue $W$ as follows. Every time a packet arrives to queue $A$ put $\xi_A \geq 0$ tokens into $W$; every time a packet arrives to queue $B$ put $\xi_B \geq 0$ tokens into $W$. The most tokens that schedule 1 can remove from $W$ is $3\xi_A$, and the most tokens that schedule 2 can remove from $W$ is $\xi_A + \xi_B$. We may as well normalize $(\xi_A, \xi_B)$ so that the largest of these is 1. The normalized vectors $(\xi_A, \xi_B)$ constitute the set of virtual resources. The extremes of this set are

$$(\xi_A, \xi_B) \in \{(0, 0), (0, 1), (1/3, 2/3), (1/3, 0)\}.$$ 

In order that the total rate at which tokens arrive should be less than the maximum rate at which we can remove tokens, we need

$$\lambda_A \xi_A + \lambda_B \xi_B \leq 1.$$ 

When a virtual resource $(\xi_A, \xi_B)$ is such that (11) is tight, it is said to be a critically loaded virtual resource. If $(\xi_A, \xi_B)$ is an extreme point of the feasible set and it is critically loaded, it is included in $S^*$. In this example it may consist of neither, one or both of both of $(0, 1)$ and $(1/3, 2/3)$, depending on $\lambda_A$ and $\lambda_B$.

Note that when $(\xi_A, \xi_B) = (0, 1)$, inequality (11) turns into the first defining constraint on $\Lambda$, and when $(\xi_A, \xi_B) = (1/3, 2/3)$ it turns into the second.

### 3.5. Example: input-queued switch.

Consider a switch with $M$ input ports and $M$ output ports. Let $\lambda_{i,j}$ be the arrival rate at the queue at input port $i$ of packets destined for output port $j$, $\lambda \in \mathbb{R}^{M \times M}$. The set $S$ consists of all $M \times M$ permutation matrices. The Birkhoff–von-Neumann decomposition result says that any doubly substochastic matrix is less than or equal to a convex combination of permutation matrices, which gives us

$$\Lambda = \left\{ \lambda \in [0, 1]^{M \times M} : \sum_{j=1}^{M} \lambda_{i,j} \leq 1 \text{ and } \sum_{i=1}^{M} \lambda_{i,j} \leq 1 \text{ for all } i, j \right\}.$$ 

It is easy to check that

$$\partial \Lambda = \left\{ \lambda \in \Lambda : \sum_{j=1}^{M} \lambda_{i,j} = 1 \text{ or } \sum_{i=1}^{M} \lambda_{i,j} = 1 \text{ for at least one } i \text{ or } j \right\}.$$ 

---

2cf. Laws [20], Example 4.4.3
We propose the following set $S^*$ of principal critically-loaded virtual resources. This set is obtained from the row and column indicators $r_\hat{i}$ and $c_\hat{j}$, defined by $(r_\hat{i})_{i,j} = 1_{i=\hat{i}}$ and $(c_\hat{j})_{i,j} = 1_{j=\hat{j}}$. We also need

$$\mathcal{N} = \{ n \in \{0,1\}^{M \times M} : n_{i,j} = 1 \text{ if } \lambda_{i,j} > 0 \}$$

Then

$$S^* = \{ r_\hat{i}n \text{ for } n \in \mathcal{N} \text{ and } \hat{i} \text{ such that } \sum_j \lambda_{i,j} = 1 \} \cup \{ c_\hat{j}n \text{ for } n \in \mathcal{N} \text{ and } \hat{j} \text{ such that } \sum_i \lambda_{i,j} = 1 \}$$

The virtual resource $r_1$, for example, corresponds to the constraint that at most one packet can be served from input port 1 in any timeslot, therefore the total arrival rate at input port 1 must be $\leq 1$. If say $\lambda_{1,3} = 0$ then the total arrival rate to the remaining $M-1$ queues at input port 1 must also be $\leq 1$, and this corresponds to the virtual resource $r_1n$ for $n_{i,j} = 1_{i>1 \text{ or } j\neq3}$.

It is easy to see that every $\xi \in S^*$ is a critically-loaded virtual resource, and it is not hard to check that they are all extreme as well. To show (9) requires some more work.

First, we remark upon a dual to the Birkhoff–von-Neumann decomposition. Let

$$\mathcal{D} = \{ r_\hat{i} \text{ for all } \hat{i} \} \cup \{ c_\hat{j} \text{ for all } \hat{j} \}.$$ 

Then, given any vector $\zeta \in \mathbb{R}^{M \times M}$ for which $\max_{\pi \in \mathcal{S}} \zeta \cdot \pi \leq 1$, we can find some $\zeta'$ which is a convex combination of elements of $\mathcal{D}$ such that $\zeta \leq \zeta'$ componentwise. This is because $\text{DUAL}(\zeta) \leq 1$ when taken with respect to the schedule set $\mathcal{D}$, by the condition on $\zeta$; and $\zeta'$ is then obtained from $\text{PRIMAL}(\zeta)$.

Now suppose that $\zeta$ is any critically-loaded virtual resource for $\text{DUAL}(\lambda)$. We need to show that (9) holds. First, use the dual decomposition above to write

$$\zeta = \sum_i x_i r_i + \sum_j y_j c_j - z.$$ 

Note that $r_1 \cdot \lambda \leq 1$ with equality only if $r_1 \in S^*$, and similarly for $c_j$. Since $\zeta$ is assumed to be critically loaded, $\zeta \cdot \lambda = 1$; it must therefore be that the coefficients $x_i$ and $y_j$ are 0 unless the corresponding virtual resource is in $S^*$, and also that $z_{i,j} > 0$ only when $\lambda_{i,j} = 0$.

To recap, we have found

$$\zeta = \sum_{\xi \in \mathcal{S}^*} a_{\xi} \xi - z.$$
where $\sum a_\xi = 1$ and $a_\xi \geq 0$, and $z_{i,j} > 0$ only when $\lambda_{i,j} = 0$. It remains to dispose of $z$. Suppose $z_{k,l} > 0$ for some $k, l$, and define $n^{k,l}_{i,j}$ by $n^{k,l}_{i,j} = 1_{i \neq k \text{ or } j \neq l}$; note that $n^{k,l} \in \mathcal{N}$ by the condition on $z$. Also note that $\zeta \in \mathbb{R}^{M \times M}_+$, and so $\sum a_\xi \xi^{k,l} \geq z^{k,l}$. Now we can rewrite

$$\zeta = \frac{z^{k,l}}{\sum a_\xi \xi^{k,l}} \sum a_\xi \xi + \left(1 - \frac{z^{k,l}}{\sum a_\xi \xi^{k,l}}\right) \sum a_\xi \xi n^{k,l} - zn^{k,l}.$$ 

Continuing in this way we can remove all non-zero elements of $z$, until we are left with an expression of the form (9).

4. The fluid model (macroscopic).

4.1. Purpose. Fluid models describe the dynamics of the system at the ‘rate’ level rather than the packet level. Fluid models are the cornerstone for proving stability in light load, for proving state space collapse at critical load, and for characterizing the behaviour in overload.

To prove stability, and to characterize the behaviour in overload, one needs to establish a relationship between the fluid model and a rescaled version of the original stochastic system (Section 5), and to establish certain simple properties of the fluid model (Section 4.3 and 10).

To obtain a critical-load limit, one needs to establish a relationship between the fluid model and a differently rescaled version of the original stochastic system (Section 8 as well as Section 5), and to establish other more involved properties of the fluid model (Section 7).

4.2. Specification of the fluid model. In this section we will specify the fluid model, and give a heuristic explanation for the equations.

Let time be measured by $t \in \mathbb{R}_+$. Let $q$, $a$ and $z$ all be functions $\mathbb{R}_+ \rightarrow \mathbb{R}_+^{N}$, and let $s = (s_\pi)_{\pi \in \mathcal{S}}$ be a collection of functions $\mathbb{R}_+ \rightarrow \mathbb{R}$. The vector of queue sizes at time $t$ is $q(t)$, the cumulative arrivals up to time $t$ is $a(t)$, the cumulative idleness up to time $t$ is $z(t)$, and $s_\pi(t)$ is the total amount of time spent on schedule $\pi$ up to time $t$.

We say that the process $x(\cdot) = (q(\cdot), a(\cdot), z(\cdot), s(\cdot))$ satisfies the fluid model for...
the MW-\( f \) scheduling algorithm if

\begin{align*}
\text{(12)} & \quad a(t) = \lambda t \\
\text{(13)} & \quad q(t) = q(0) + a(t) - \sum_{\pi} s_{\pi}(t) \pi + z(t) \\
\text{(14)} & \quad \sum_{\pi \in S} s_{\pi}(t) = t \\
\text{(15)} & \quad \text{each } s_{\pi}(\cdot) \text{ and } z_n(\cdot) \text{ is increasing (not necessarily strictly increasing)} \\
\text{(16)} & \quad \text{all the components of } x(\cdot) \text{ are absolutely continuous—indeed they are Lipschitz} \\
\text{(17)} & \quad \text{for almost all } t, \text{ all } n, \quad \dot{z}_n(t) = 0 \text{ if } q_n(t) > 0 \\
\text{(18)} & \quad \text{for almost all } t, \text{ all } \pi \in S, \quad \dot{s}_{\pi}(t) = 0 \text{ if } \pi \cdot f(q(t)) < \max_{\rho \in S} \rho \cdot f(q(t))
\end{align*}

We will only consider weight functions \( f \) which satisfy

**Assumption 4.1** Assume \( f \) is differentiable and strictly increasing with \( f(0) = 0 \). Assume also that for any \( q \in \mathbb{R}^N_+ \) and \( \pi \in S \), with \( m(q) = \max_{\rho \in S} \rho \cdot f(q) \),

\[ \pi \cdot f(q) = m(q) \quad \implies \quad \pi \cdot f(\kappa q) = m(\kappa q) \text{ for all } \kappa \in \mathbb{R}_+. \]

The first two equations are the continuous analogues of (3) and (1). Equation (14) says that the scheduling algorithm must choose some schedule at every timestep. The requirement of absolute continuity implies that all components of \( x(\cdot) \) are differentiable almost everywhere; any equations we write which involve derivatives are meant to apply to such \( t \). Equation (17) is the continuous analogue of (2). Equations (12) to (17) apply to any scheduling algorithm; whereas the final equation applies only to MW-\( f \), and is the continuous analogue of (6). The assumption about the weight function \( f \) says that the scheduling algorithm has the same behaviour if the queues are rescaled, and is needed in the proof that relates the original stochastic system to the fluid-scaled system. It is satisfied by \( f(x) = x^\alpha, \alpha > 0 \), but it is not satisfied for example for an input-queued switch with \( f(x) = \log(1 + x) \).

4.3. **Fluid stability.** A fluid model is said to be *stable* if there is some draining time \( H \in \mathbb{R}_+ \) such that every fluid model solution with bounded initial queue size \(|q(0)| \leq 1 \) ends up with \( q(t) = 0 \) for all \( t \geq H \). It is said to be *weakly stable* if every fluid model with empty initial queues \( q(0) = 0 \) remains at \( q(t) = 0 \) for all \( t \geq 0 \).

Define the function \( L : \mathbb{R}^N_+ \to \mathbb{R}_+ \) by

\[ L(q) = F(q) \cdot 1 \]
where \( F(x) = \int_0^x f(y) \, dy \), and \( F(q) = [F(q_n)]_{1 \leq n \leq N} \) as per the notation in Section 1.3. The next lemma shows that \( L \) is a Lyapunov function, and it can be used to prove weak stability.

**Lemma 4.2** Suppose the system is running MW-f. Then every fluid model solution satisfies

\[ \dot{L}(q(t)) = \lambda \cdot f(q(t)) - \max_{\pi \in S} \pi \cdot f(q(t)). \]

Furthermore if \( \lambda \in \Lambda \) then

\[ \lambda \cdot f(q) - \max_{\pi \in S} \pi \cdot f(q) \leq 0 \quad \text{for all } q \in \mathbb{R}_+^N \]

hence \( \dot{L}(q(t)) \leq 0 \). Furthermore, if \( \lambda \in \Lambda^\circ \) then either \( \dot{L}(q(t)) < 0 \) or \( q(t) = 0 \).

**Proof.**

\[
\frac{d}{dt} L(q(t)) = \dot{q}(t) \cdot f(q(t)) = \left( \lambda - \sum_{\pi \in S} \dot{s}_\pi(t) \pi + \dot{z}(t) \right) \cdot f(q(t)) \quad \text{by differentiating (13)}
\]

\[
= \left( \lambda - \sum_{\pi} \dot{s}_\pi(t) \pi \right) \cdot f(q(t)) \quad \text{by (17), using } f(0) = 0
\]

\[
= \lambda \cdot f(q(t)) - \max_{\rho} \rho \cdot f(q(t)) \sum_{\pi} \dot{s}_\pi(t) \quad \text{by (18)}
\]

\[
= \lambda \cdot f(q(t)) - \max_{\rho} \rho \cdot f(q(t)) \quad \text{by (14)}.
\]

When \( \lambda \in \Lambda \), we can write \( \lambda \leq \sigma \) componentwise for some \( \sigma = \sum_{\pi} \alpha_\pi \pi \) with \( \alpha_\pi \geq 0 \) and \( \sum \alpha_\pi \leq 1 \); if \( \lambda \in \Lambda^\circ \) then \( \sum \alpha_\pi < 1 \). Hence for any \( q \in \mathbb{R}_+^N \),

\[
\lambda \cdot f(q) - \max_{\rho} \rho \cdot f(q) = \sum_{\pi} \alpha_\pi \pi \cdot f(q) - \max_{\rho} \rho \cdot f(q)
\]

\[
\leq \left( \sum_{\pi} \alpha_\pi - 1 \right) \max_{\rho} \rho \cdot f(q).
\]

Clearly the maximum is > 0 unless \( q = 0 \), hence the result follows. \( \Box \)

When we use this result, we almost always implicitly pair it with a standard fact which is worth stating here: if \( f : \mathbb{R}_+ \to \mathbb{R} \) is an absolutely continuous function, and \( f(t) \leq 0 \) at almost all \( t \), then \( f(t) \leq f(0) \) for all \( t \geq 0 \). This fact, together with the fact that \( L(q) = 0 \iff q = 0 \), implies that the fluid model for MW-f is weakly stable for \( \lambda \in \Lambda \).
Indeed, the fluid model for MW-$f$ is stable for $\lambda \in \Lambda^\circ$. To see this it helps to use an alternative condition for characterizing stability, due to Stolyar [30], Theorem 6.1: the fluid model is stable iff all fluid model solutions with $1 \cdot q(0) = 1$ satisfy $\inf_{t \geq 0} 1 \cdot q(t) < 1$. We prove that switched networks satisfy Stolyar’s condition, by contradiction. Suppose there exists a fluid model solution with $1 \cdot q(t) \geq 1$ for all $t$. Then, for all $t$,

$$
\dot{L}(q(t)) \leq \left( \sum_{\pi} \alpha_{\pi} - 1 \right) \max_{\rho} \rho \cdot f(q) \quad \text{by (20)}
$$

$$
\leq \left( \sum_{\pi} \alpha_{\pi} - 1 \right) S^{\min} f(1/N) \quad \text{where } S^{\min} = \min_{\rho \in S} \min_{n, \rho_n > 0} \rho_n
$$

$$
< -\varepsilon \quad \text{for some } \varepsilon > 0.
$$

This is inconsistent with $L(\cdot) \geq 0$, hence the supposition is false, hence Stolyar’s condition is satisfied, hence the fluid model is stable.

4.4. Consequences of fluid stability. Dai [6], Section 2.6 and Stolyar [30] describe the relationship between stability of the fluid model and stability of the original (unscaled) stochastic process. For example, for a wide class of queueing systems that may be represented by Markov chains, stability of the fluid model implies positive Harris recurrence of the Markov chain.

In our case, suppose the fluid model is stable, and let $H$ be the draining time. By Theorem 5.1 below, assuming that (3)–(5) hold with $R(r) = O(1)$, $Q(rH)/r$ converges in distribution to 0; furthermore the assumption of bounded arrivals (4) implies that the collection $\{Q(rH)/r : r \in \mathbb{N}\}$ is uniformly integrable; hence $Q(rH)/r$ converges to 0 in expectation. Now suppose that the arrival process has stationary independent increments, so that $Q(\cdot)$ is a Markov chain in $\mathbb{R}_+^N$. We can then apply [10, Theorem 1], which is based on an idea due to Tweedie, to deduce that $Q(\cdot)$ is positive recurrent.

5. Derivation of fluid model (microscopic). The development in this section follows closely the pattern of Bramson [4]. To help the reader familiar with that work, we footnote our results with references to the corresponding results of Bramson.

5.1. Fluid model scaling. Consider a sequence of systems of the type described in Section 1.4, indexed by $r \in \mathbb{N}$. Write $X^r(\tau) = (Q^r(\tau), A^r(\tau), Z^r(\tau), S^r(\tau)), \tau \in \mathbb{Z}_+$, for the $r$th system. Let $r' : \mathbb{N} \to \mathbb{R}$ be a function with $r'(r) \to \infty$ as $r \to \infty$, and write $r'$ for $r'(r)$ when the context indicates which $r$ is meant. Define the scaled
system \( x^r(t) = (q^r(t), a^r(t), z^r(t), s^r(t)) \) for \( t \in \mathbb{R}_+ \) by

\[
\begin{align*}
q^r(t) &= Q^r(r't)/r' \\
a^r(t) &= A^r(r't)/r' \\
z^r(t) &= Z^r(r't)/r' \\
s^r(t) &= S^r(r't)/r'
\end{align*}
\]

after extending the domain of \( X^r(\cdot) \) to \( \mathbb{R}_+ \) by linear interpolation in each interval \((\tau - 1, \tau)\).

For some purposes, e.g. analysing stability as described in Section 4.3 or analysing overloaded systems as described in Section 10, it is sufficient to drop the dependence of \( X^r(\cdot) \) on \( r \), and let \( r'(r) = r \), giving \( a^r(t) = A(rt)/r \) etc. But when we come to study the critically-loaded switch in Section 8, we might need to consider a sequence of queueing systems where the \( r \)th system has arrival process \( A^r(\cdot) \), and let

\[
\begin{align*}
a^1(t) &= A^1(t) \\
a^2(t) &= A^2(2t)/2 \\
a^3(t) &= A^2(2 + 2t)/2 \\
a^4(t) &= A^3(3t)/3 \\
a^5(t) &= A^3(3 + 3t)/3 \\
a^6(t) &= A^3(6 + 3t)/3 \\
a^7(t) &= A^4(4t)/4 \\& \ldots
\end{align*}
\]

Beware the fact that the critical-load index \( r \) is not the same as the fluid index \( r' \)!

The exact rescaling is given in Section 8.4.

We now state a general theorem, which covers both cases.

5.2. **Main result.** 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 \( A^r(\cdot) \) satisfies assumptions (3)–(5). We will also make the following uniformity assumptions:

Let the arrival rate vector for the \( r \)th system be \( \lambda^r \). Assume that

\begin{equation}
\lim_{r \to \infty} \lambda^r = \lambda \quad \text{for some} \quad \lambda \in \mathbb{R}_+^N.
\end{equation}

Assume that the increments are uniformly bounded by \( A^\text{max} \), which implies

\begin{equation}
a^r_n(t) - a^r_n(s) \in \left[0, A^\text{max} + (t - s)A^\text{max}\right] \quad \text{for any} \quad s \leq t.
\end{equation}

Assume that there is a function \( R : \mathbb{R}_+ \to \mathbb{R}_+ \) which gives a uniform bound on the chance of large fluctuations in the arrival processes, in that

\begin{equation}
P\left( \sup_{t \in [0, T]} \left| a^r(t) - \lambda^r t \right| < \varepsilon \right) = 1 - o(R(r')) \quad \text{for all} \quad \varepsilon > 0.
\end{equation}

Assume that the initial queue sizes are bounded,

\begin{equation}
|q^r(0)| < K \quad \text{for some} \quad K \in \mathbb{R}_+.
\end{equation}
We may optionally assume that the initial size is non-random and that it converges,
\[(25)\]
\[q^r(0) \to q_0 \text{ for some } q_0 \in \mathbb{R}_+^N.\]

Note that in the simple case where \(X^r\) does not depend on \(r\) and \(r'(r) = r\), these conditions all follow automatically from assumptions (3)–(5).

**Theorem 5.1**  
Make the above assumptions (3)–(5) and (21)–(24). Let \(FMS\) be the set of all processes \(x(t)\) over \(t \in [0, T]\) which satisfy the appropriate fluid model equations, namely
- equations (12)–(17), for any scheduling algorithm,
- equation (18) in addition if the network is running MW-f and Condition 4.1 holds,
- \(q(0) = q_0\) in addition, if (25) holds.

Let \(FMS_\varepsilon\) be the \(\varepsilon\)-fattening
\[FMS_\varepsilon = \left\{ x : \sup_{t \in [0,T]} |x(t) - y(t)| < \varepsilon \text{ for some } y \in FMS \right\}.
\]

Then for any \(\varepsilon > 0\), \(P(x^r(\cdot) \in FMS_\varepsilon) = 1 - o(R(r'))\) as \(r \to \infty\).

The proof of Theorem 5.1 is in Section 5.4, following some technical preliminaries.

5.3. Technical preliminaries. Let \(C\) 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 \(|x| = \max_i |x_i|\). For our work here we want \(I = 3N + |\mathcal{S}|\); then each of the fluid-scaled systems \(x^r(\cdot)\), and any process \(x(\cdot)\) which satisfies the fluid model, must lie in \(C\). Equip \(C\) with the norm
\[\|x\| = \sup_{t \in [0,T]} |x(t)|.\]

Define the modulus of continuity by
\[mc_\delta(x) = \sup_{|s-t|<\delta} |x(s) - x(t)|\]

where \(s, t \in [0, T]\). Since \([0, T]\) is compact, each \(x \in C\) is uniformly continuous, therefore \(mc_\delta(x) \to 0\) as \(\delta \to 0\).

We are interested in convergence in \((C, \| \cdot \|)\). The appropriate concept is cluster points. Consider any metric space \(E\) 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 \(\liminf_{r \to \infty} d(x, E_r) = 0\).

\(^3\)cf. Bramson [4], Proposition 6.1
Proposition 5.2 (Cluster points in $C$)  $^4$ Given $K > 0$, $A > 0$ and a sequence $B_r \to 0$, let

$$K_r = \{ x \in C : |x(0)| \leq K \text{ and } mc_\delta(x) \leq A\delta + B_r \text{ for all } \delta > 0 \}$$

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

Rather than following Bramson’s use of cluster points, we might instead have used weak convergence. The general line of argument would be (i) the sequence of measures of $x^r(\cdot)$ is tight; (ii) by Prohorov’s theorem [2, Theorem 5.1] it is relatively compact, so there exists a weakly convergent subsequence; (iii) by the Skorohod representation theorem [2, Theorem 6.7], we can express this weak convergence as pathwise convergence; (iv) pathwise limits must satisfy the fluid model equations. Lemma 5.3 below does the job of (i), Lemma 5.4 does the job of (iv), and Proposition 5.2 does the rest of the work—notice how similar it is to the characterization of compact sets in $(C, \| \cdot \|)$ from [2, Theorem 7.2]. The benefit of the cluster-point technique is that it gives tighter control of the probability of rare events, and this tighter control is needed for proving state space collapse in Section 8.

5.4. Proof of main result. The proof strategy is to choose a sequence of events which have high probability and on which the arrival process for the $r$th system is well-behaved; then to show that this implies 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 satisfy the fluid model equations.

Choice of ‘good arrivals’ events. Write $\text{ARR}^r_\varepsilon$ for the event

$$\text{ARR}^r_\varepsilon = \left\{ \sup_{t \in [0,T]} |a^r(t) - \lambda^r t| < \varepsilon \right\}.$$ 

By (23), $P(\text{ARR}^r_\varepsilon) = 1 - o(R(r'))$ for all $\varepsilon > 0$, therefore it is possible to choose a sequence $\varepsilon(r) \to 0$ such that

$$P(\text{ARR}^r_{\varepsilon(r)}) = 1 - o(R(r')).$$ 

We will use these $\text{ARR}^r_{\varepsilon(r)}$ for the good events. Note: a suitable sequence $\varepsilon(r)$ can be obtained by setting $\varepsilon_i = 1/i$, $r_0 = 1$, and

$$r_i = \inf \left\{ s > r_{i-1} : \sup_{r \geq s} \frac{P(a^r \notin \text{ARR}_{\varepsilon_i})}{R(r')} < \varepsilon_i \right\},$$

then defining $\varepsilon(r) = \varepsilon_{\max\{i : r_i \leq r\}}$.

$^4$taken from Bramson [4], Proposition 4.1
System behaviour under good arrivals. Let \( E_r = \{ x^r(\omega) : \omega \in \text{ARR}^r_{\varepsilon(r)} \} \), i.e. the set of all possible paths for the entire system for any well-behaved arrival process. Lemma 5.3 below shows that every possible path \( x^r \) lies in the set \( K_r \) defined in Proposition 5.2, for appropriate constants \( K \), \( A \) and \( B_r \), so certainly \( E_r \subset K_r \) as required by that proposition. Therefore

\[
\sup_{\omega \in \text{ARR}^r_{\varepsilon(r)}} d(x^r(\omega), \text{CP}) \to 0
\]

where CP is the set of cluster points of the \( E_r \). Lemma 5.4 below shows that by our choice of \( E_r \), all cluster points satisfy the fluid model equations, therefore

\[
\sup_{\omega \in \text{ARR}^r_{\varepsilon(r)}} d(x^r(\omega), \text{FMS}) \to 0.
\]

Wrapping up. All that remains is to interpret this last convergence. Pick any \( \delta > 0 \). The equation says that \( x^r \in \text{FMS}_\delta \) (for \( r \) sufficiently large) in the event that \( \text{ARR}^r_{\varepsilon(r)} \) occurs. But by (26), this has probability \( 1 - o(R(r')) \). Therefore \( \mathbb{P}(x^r \in \text{FMS}_\delta) \geq 1 - o(R(r')) \). This establishes the desired result. \( \square \)

**Lemma 5.3 (Tightness of fluid scaling)\(^5\)** For every \( r \), with \( K_r \) as defined in Proposition 5.2, \( x^r \in K_r \). The constants used to define \( K_r \) are \( K \) as given in (24), and \( A \) and \( B_r \) from (27) below.

**Proof.** We will prove that \( x^r \) satisfies the two defining conditions of \( K_r \).

For the condition about initial state, note from the definition of the model in Section 1.4 that the only non-zero component of \( x^r(0) \) is \( q^r(0) \), and that \( |q^r(0)| \leq K \) by assumption (24).

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, for every queue \( n \),

\[
|a^r_n(t) - a^r_n(s)| \leq \frac{1}{r'} \left( A^r_n([r't]) - A^r_n([r's]) \right)
\]

since \( A_n(\cdot) \) is increasing

\[
\leq \frac{A^\text{max}}{r'} \left( r'(t - s) + 2 \right)
\]

by the bound (22) on arrivals

\[
< \delta A^\text{max} + 2A^\text{max}/r'.
\]

For idling, let \( S^\text{max} = \max_{\pi \in S} \max_n \pi_n \). This is the maximum amount of service that can be offered to any queue per unit time, and it must be finite since \( |S| \) is finite. Then, by a similar argument based on (2),

\[
|z^r_n(t) - z^r_n(s)| < \delta S^\text{max} + 2S^\text{max}/r'.
\]

\(^5\)cf. Bramson [4], Proposition 5.2
For service, since $S_{\pi}(\cdot)$ is increasing and since a schedule must be chosen not more than once every timeslot,

$$|s_{\pi}^r(t) - s_{\pi}^r(s)| \leq \frac{1}{r'} \left( S_{\pi}^r([r't]) - S_{\pi}^r([r's]) \right) < \delta + 2/r'$$

For queue size, note that (1) carries through to the fluid model scaling, i.e.

$$q^r(t) = q^r(0) + a^r(t) - \sum_\pi s_{\pi}^r(t)\pi + z^r(t),$$

thus

$$|q_n^r(t) - q_n^r(s)| \leq |a_n^r(t) - a_n^r(s)| + \sum_\pi |s_{\pi}^r(t) - s_{\pi}^r(s)| + |z_n^r(t) - z_n^r(s)|$$

$$< \left( \delta + 2/r' \right) \left( A^{\text{max}} + |S|S^{\text{max}} + S^{\text{max}} \right).$$

Putting all these together,

(27)

$$w_\delta(x^r) < \left( \delta + 2/r' \right) \left( 2NA^{\text{max}} + 2NS^{\text{max}} + N|S|S^{\text{max}} + |S| \right).$$

Lemma 5.4 (Dynamics at cluster points) \(^6\) Let $x$ be a cluster point of the sequence $E_r = \{x^r(\omega) : \omega \in \text{ARR}^r_{\varepsilon(r)} \}$. Then $x \in \text{FMS}$.

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 $x^{r_k} \to x$. We now use this in proving all the fluid model equations. It turns out that the only equation which relies on the choice of $E_r$ is (12); all the other equations hold even if we let $E_r$ consist of all possible paths $x^r(\omega)$. We will omit the subscript $k$ in the rest of the proof.

Proof of (12). Observe that

$$\sup_{t \in [0,T]} |a(t) - \lambda t| \leq \sup_{t \in [0,T]} |a(t) - a^r(t)| + \sup_{t \in [0,T]} |a^r(t) - \lambda^r t| + T|\lambda^r - \lambda|.$$

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 $\text{ARR}^r_{\varepsilon(r)}$, the third by (21). Since the left hand side does not depend on $r$, it must be that $a(t) = \lambda t$.

Proof of (13)–(15). The discrete (unscaled) system satisfies these properties, therefore the scaled systems $x^r$ do too. Taking the limit yields the fluid equations.

\(^6\)cf. Bramson [4], Proposition 6.2
Proof of (16). In equation (27) we found constants $A$ and $B$ such that
\[ |x'(t) - x'(s)| \leq A|t-s| + B/r'. \]
Taking the limit as $r \to \infty$, we find that $|x(t) - x(s)| \leq A|t-s|$, i.e. $x$ is Lipschitz continuous. This implies that $x$ is absolutely continuous.

Proof of (17). 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 $\delta$ sufficiently small that $\inf_{s \in I} q_n(s) > 0$. Since $\|q^* - q\| \to 0$, we can find $\alpha > 0$ such that $\inf_{s \in I} q_n(s) > \alpha$ for all $r$ sufficiently large. In the unscaled version of the process, this means $\inf_{s \in I} Q_n^I(r') > r'\alpha$. By (2), there is too much work in the queue over this entire interval for there to be any idling, so after rescaling we find $z_n^I(t + \delta/2) = z_n^I(t)$. (The switch from $\delta$ 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 (18). Pick a $t$ at which $s_\pi$ is differentiable, and suppose that $\pi \cdot f(q(t)) < \max_{\rho \in \mathcal{S}} \rho \cdot f(q(t))$. As above, pick some small interval $I = [t, t + \delta]$ and $r$ sufficiently large that
\[ \pi \cdot f(q^T(s)) < \max_{\rho \in \mathcal{S}} \rho \cdot f(q^T(s)) \quad \text{for } s \in I. \]
Writing this in terms of the unscaled system and applying Condition 4.1,
\[ \pi \cdot f(Q^I(r's)) < \max_{\rho \in \mathcal{S}} \rho \cdot f(Q^I(r's)) \quad \text{for } s \in I. \]
The MW-$f$ algorithm ensures by (6) that $\pi$ will not be chosen throughout this entire interval, so after rescaling we find $s_\pi^I(t + \delta/2) - s_\pi^I(t) \leq r_{\pi}^{\text{max}}/r$, and taking the limit gives $s_\pi(t + \delta/2) = s_\pi(t)$. We assumed $s_\pi$ to be differentiable at $t$; the derivative must be 0.

Initial queue size. If (25) holds then certainly $q^I(0) \to q_0$. \qed

6. The lifting map (teleology). This section describes two optimization problems which capture the operation of the MW-$f$ scheduling algorithm. The link between the problems and the fluid model will be explained in Section 7.

One problem is expressed in the language of DUAL and is useful for reasoning about what the algorithm achieves. The other is expressed in the language of PRIMAL and is useful for reasoning about the operation of the algorithm.

In Section 6.2 we use the problems to define the lifting map, and to prove two properties which will be needed when we analyse state space collapse in Section 8.

Throughout this section, let the arrival rate vector be $\lambda \in \Lambda$, and let $q \in \mathbb{R}_+^N$ be a vector of queue sizes. Let $\mathcal{S}^\pi(\lambda)$ be the set of principal critically loaded virtual
resources, defined in Section 3.3, let it consist of \( V \) elements, and define the workload map \( W: \mathbb{R}_+^N \to \mathbb{R}_+^V \) by \( W(\mathbf{q}) = [w_\xi | \xi \in S^r(\lambda)] \), where \( w_\xi = \xi \cdot \mathbf{q} \). Let \( L(\mathbf{q}) \) be the Lyapunov function for the MW-\( f \) algorithm, defined by (19), and assume that \( f \) satisfies Condition 4.1.

There are many similarities between our results and those of Kelly and Williams [17] for bandwidth-sharing models of congestion control, and we will signpost the links with footnotes.

6.1. Optimization problems. Consider first the optimization problem ALGD(\( W^\lambda(\mathbf{q}) \)):

\[
\begin{array}{ll}
\text{minimize} & L(\mathbf{r}) \\
\text{over} & \mathbf{r} \in \mathbb{R}_+^N \\
\text{such that} & \xi \cdot \mathbf{r} \geq w_\xi \quad \text{for all} \quad \xi \in S^r(\lambda)
\end{array}
\]

Note that \( \mathbf{r} \) is feasible for ALGD(\( W^\lambda(\mathbf{q}) \)) if and only if \( \zeta \cdot \mathbf{r} \geq \zeta \cdot \mathbf{q} \) for all critically loaded virtual resources (not just the principal ones) by (9). In words, the optimization problem says “Rearrange work between the queues so as to make the Lyapunov function as small as possible, except that the critical workloads must all be conserved.”

Now consider the optimization problem ALGP(\( \mathbf{q} \)):

\[
\begin{array}{ll}
\text{minimize} & L(\mathbf{r}) \\
\text{over} & \mathbf{r} \in \mathbb{R}_+^N \\
\text{such that} & \mathbf{r} \geq \mathbf{q} + t(\lambda - \sigma) \quad \text{for some} \quad t \geq 0, \sigma \in \Sigma
\end{array}
\]

Note that the optimum must have \( \mathbf{r} = [\mathbf{q} + t(\lambda - \sigma)]^+ \), since \( L(\cdot) \) is strictly increasing componentwise. In words, this optimization problem says “Rearrange work between the queues so as to make the Lyapunov function as small as possible, by running some schedule or mix of schedules \( \sigma \) for as long as it takes.”

We chose the names ALGP and ALGD because the constraints in the problems are suggestive of PRIMAL and DUAL, not because the two problems are dual to one another. Rather, the following lemma shows they are equivalent because they have the same feasible sets. We tend to use ALGP for proofs in which we construct a feasible state, and ALGD for proofs in which we argue about all possible states.

**Lemma 6.1** \(^7\) Let \( \lambda \in \Lambda \) and \( \mathbf{q}, \mathbf{r} \in \mathbb{R}_+^N \). Then the following are equivalent:

\(^7\)This was suggested to us by C. Moallemi. Also cf. Harrison [12], Proposition 3 for the equivalent result for monotone scheduling sets.
i. $r \geq q + t(\lambda - \sigma)$ for some $t \geq 0$, $\sigma \in \Sigma$.

ii. $\xi \cdot r \geq \xi \cdot q$ for all $\xi \in S^*(\lambda)$

Proof that (i) $\implies$ (ii). Pick any critically loaded virtual resource $\xi$. By definition $\xi \geq 0$, so

$$\xi \cdot r \geq \xi \cdot q + t(\xi \cdot \lambda - \xi \cdot \sigma)$$

assuming $r$ satisfies (i)

$$= \xi \cdot q + t(1 - \xi \cdot \sigma)$$

since $\xi$ is critical for $\text{DUAL}(\lambda)$

$$\geq \xi \cdot q + t(1 - \text{DUAL}(\sigma))$$

since $\xi$ is feasible for $\text{DUAL}(\sigma)$

$$= \xi \cdot q + t(1 - \text{PRIMAL}(\sigma))$$

by strong duality

$$\geq \xi \cdot q + t(1 - 1)$$

since $\sigma \in \Sigma \subset \Lambda$, so $\text{PRIMAL}(\sigma) \leq 1$

$$= \xi \cdot q$$

therefore $r$ satisfies (ii).

Proof that (i) $\iff$ (ii). Let $r$ satisfy (ii), and let $\sigma' = \lambda - (r - q)/t$ for some sufficiently large $t \in \mathbb{R}_+$. We shortly show that the solution to $\text{DUAL}(\sigma')$ is $\leq 1$. By strong duality the solution to $\text{PRIMAL}(\sigma')$ is likewise $\leq 1$, and so by definition of $\text{PRIMAL}$ we can find some $\sigma \in \Sigma$ such that $\sigma' \leq \sigma$ componentwise. Then

$$r = q + t(\lambda - \sigma') \geq q + t(\lambda - \sigma)$$

i.e. $r$ satisfies (i).

It remains to show that the solution to the $\text{DUAL}(\sigma')$ optimization problem is $\leq 1$, i.e. that $\zeta \cdot \sigma' \leq 1$ for all virtual resources $\zeta$. Without loss of generality, assume that $\zeta$ is an extreme feasible point. Either $\zeta$ is a critically-loaded virtual resource, in which case

$$\zeta \cdot \sigma' = \zeta \cdot \lambda - \zeta \cdot (r - q)/t$$

$$= 1 - \zeta \cdot (r - q)/t$$

since $\zeta$ is critically loaded

$$\leq 1$$

by assumption (ii),

or $\zeta$ is not critically loaded in which case

$$\zeta \cdot \sigma' < 1 - \zeta \cdot (r - q)/t$$

and this is $< 1$ for $t$ sufficiently large. Either way, $\zeta \cdot \sigma' \leq 1$. Therefore the solution to $\text{DUAL}(\sigma')$ is $\leq 1$.

6.2. The lifting map. We assumed $f$ is strictly increasing, hence $F$ is strictly convex, hence $L$ is strictly convex. Also, the feasible set for ALGD is non-empty and convex. Therefore the problem has a unique solution.
Definition 6.2 (Lifting map \( \Delta \)) Given a weight function \( f \) satisfying Condition 4.1, and a vector of arrival rates \( \lambda \in \Lambda \), and the set of \( V \) principal critically-loaded virtual resources, define the lifting map \( \Delta f, \lambda : \mathbb{R}^{V}_+ \rightarrow \mathbb{R}^N_+ \) which maps \( w \) to the unique solution to ALGD(\( w \)).

We now state two lemmas which will be needed when we prove state space collapse in Section 8. We will drop the superscripts on \( W^\lambda \) and \( \Delta f, \lambda \) when the context makes them clear, and we drop the outer brackets in \( \Delta(W(q)) \).

Lemma 6.3 (Continuity of \( \Delta \)) \(^8\) If \( \lambda \in \Lambda \), the lifting map is continuous.

Proof. If \( S^*(\lambda) \) is empty, then the lifting map is trivial and the result is trivial. In what follows, we shall assume that \( S^*(\lambda) \) is non-empty, and we will abbreviate it to \( S^* \). Furthermore note that, by definition of \( S^* \), for every \( \xi \in S^* \) we know \( \xi \cdot \lambda = 1 \) and hence there is some queue \( n \) such that \( \xi_n > 0 \).

Pick any sequence \( w^k \rightarrow w \in \mathbb{R}^V_+ \), and let \( r^k = \Delta(w^k) \) and \( r = \Delta(w) \). We want to prove that \( r^k \rightarrow r \). We shall first prove that there is a compact set \( [0, h]^N \) such that \( r^k \in [0, h]^N \) for all \( k \). We shall then prove that any convergent subsequence of \( r^k \) converges to \( r \); this establishes continuity of \( \Delta \).

First, compactness. A suitable value for \( h \) is

$$ h = \max \max \sup_{\xi \in S^*} \xi_n > 0 \quad k \quad \xi_n $$

Note than the maximums are over a non-empty set, as noted at the beginning of the proof. Note also that \( h \) is finite because \( w \) is finite. Now, suppose that \( r^k \not\in [0, h]^N \) for some \( k \), i.e. that there is some queue \( n \) for which \( r^k_n > h \), and let \( r' = r^k \) in each component except for \( r'_n = h \). We claim that \( r' \) satisfies the constraints of ALGD(\( w^k \)). To see this, pick any \( \xi \in S^* \); either \( \xi_n = 0 \) in which case \( \xi r' = \xi r^k \geq w^k_\xi \), or \( \xi_n > 0 \) in which case \( \xi \cdot r^k \geq \xi_n h \geq w^k_\xi \) by construction of \( h \). Applying this repeatedly, if \( r^k \not\in [0, h]^N \) then we can reduce it to a queue size vector in \( [0, h]^N \), thereby improving on \( L(r^k) \), yet still meeting the constraints of ALGD(\( w^k \)); this contradicts the optimality of \( r^k \). Hence \( r^k \in [0, h]^N \).

Next, convergence on subsequences. With a slight abuse of notation, let \( \Delta(w^k) = r^k \rightarrow s \) be a convergent subsequence, and recall that \( \Delta(w) = r \) and \( w^k \rightarrow w \). By continuity of the constraints of ALGD, \( s \) is feasible for ALGD(\( w \)); we shall next show that \( L(s) \leq L(r) \). Since \( r \) is the unique optimum, it must be that \( s = r \).

It remains to show that \( L(s) \leq L(r) \). Consider the sequence \( r + \varepsilon^k 1 \) as candidate solutions to ALGD(\( w^k \)) where

$$ \varepsilon^k = \max_{\xi \in S^*} \frac{w^k_\xi - w_\xi}{\xi \cdot 1}. $$

\(^8\)cf. Kelly and Williams [17], Lemma A.3 for the special case where \( f(\kappa x) = \kappa^\alpha x \).
This choice ensures that the candidates are feasible, since
\[ \xi \cdot (r + \varepsilon^k 1) = \xi \cdot r + \varepsilon^k \xi \cdot 1 \geq \xi \cdot r + w^k - w \xi \geq w^k. \]
Since the candidates are feasible solutions to ALGD(\(w^k\)), and \(r^k\) is an optimal solution, it must be that
\[ L(r^k) \leq L(r + \varepsilon^k 1). \]
Taking the limit as \(k \to \infty\), and noting that \(L\) is continuous and \(\varepsilon^k \to 0\), we find
\[ L(s) \leq L(r) \]
as required. This completes the proof. \(\square\)

**Lemma 6.4 (Scale-invariance of \(\Delta\))** \(^9\) Let \(\lambda \in \Lambda\) and \(q \in \mathbb{R}^N_+\). Then \(\Delta W(\kappa q) = \kappa \Delta W(q)\) for all \(\kappa > 0\).

**Proof.** This result is about the scale-invariance of \(\Delta\), and it hinges on Condition 4.1 about the scale-invariance of \(f\). Since that condition was expressed in terms of schedules \(\pi \in S\), we will prove the lemma using the ALGP characterization of \(\Delta\). Before delving into the two scales, we will establish three preliminary properties of \(\Delta\). Preliminary 1 is used to prove 2, and 2 & 3 are used in the main proof.

**Preliminary 1.** If \(q = \Delta(w)\) for some \(w \in \mathbb{R}^V_+\) then
\[ (28) \quad \lambda \cdot f(q) = \max_{\pi \in S} \pi \cdot f(q) \]
To see this, suppose \(\pi \in S\) has maximal weight and consider \(r = [q + t(\lambda - \pi)]^+\).
This is feasible for ALGD by Lemma 6.1. Now (using the fact that \(f(0) = 0\)),
\[ \frac{d}{dt} L\left([q + t(\lambda - \pi)]^+\right) \bigg|_{t=0} = (\lambda - \pi) \cdot f(q) \]
and since \(q\) is optimal for ALGD it must be that \(\lambda \cdot f(q) \geq \pi \cdot f(q)\). On the other hand, \(\lambda \in \Lambda\) so \(\lambda \leq \sigma\) for some \(\sigma \in \Sigma\), hence \(\lambda \cdot f(q) \leq \sigma \cdot f(q) \leq \pi \cdot f(q)\). Hence the result follows.

**Preliminary 2.** Suppose that \(r = \Delta W(q)\). From the structure of ALGP, \(r = [q + t(\lambda - \sigma)]^+\) for some \(t \geq 0\) and \(\sigma \in \Sigma\). Then either \(t = 0\) or
\[ (29) \quad \sigma \cdot f(r) = \max_{\pi \in \mathcal{S}} \pi \cdot f(r). \]

\(^9\)cf. Kelly and Williams [17], Lemma A.2 for the special case where \(f(\kappa x) = \kappa^\alpha f(x)\).
This is because $t$ is an optimal choice, so either $t$ is constrained to be 0 or
\[
\frac{d}{du} L \left( [q + u(\lambda - \sigma)]^+ \right) \bigg|_{u=t} = (\lambda - \sigma) \cdot f(r) = 0.
\]
In this second case, $\lambda \cdot f(r) = \max_{\pi} \pi \cdot f(r)$ by (28) so the same is true for $\sigma$.

Preliminary 3. Suppose that $r = \Delta W(q)$. From ALGP, we can write it as $r = [q + t(\lambda - \sigma)]^+$ for some $\sigma \in \Sigma$. In fact, for any $T \geq t$ we can write it as (30)
\[
r = [q + T(\lambda - \rho)]^+ \quad \text{for some } \rho \in \Sigma.
\]
To see this, recall that PRIMAL$(\lambda) \leq 1$, so we pick some $\bar{\lambda} \in \Sigma$ such that $\lambda \leq \bar{\lambda}$, whence
\[
r \geq [q + t(\lambda - \sigma) + (T-t)(\lambda - \bar{\lambda})]^+
= [q + T(\lambda - \rho)]^+ \quad \text{where } \rho = \frac{t}{T} \sigma + \frac{T-t}{T} \bar{\lambda} \in \Sigma.
\]
This last expression is certainly feasible for ALGP$(q)$. Since $r$ is optimal for ALGP$(q)$, and the objective function is increasing pointwise, $r = [q + T(\lambda - \rho)]^+$ as claimed.

Main proof. Let $r = \Delta W(q)$ and $\kappa r' = \Delta W(\kappa q)$. We know that $\kappa r$ is feasible for ALGD$(W(\kappa q))$ because the constraints are linear; we will now show that $L(\kappa r) \leq L(\kappa r')$; hence $\kappa r$ is also optimal for ALGD$(W(\kappa q))$. By uniqueness of the optimum, $\kappa r = \kappa r'$ as required.

It remains to prove that $L(\kappa r) \leq L(\kappa r')$. Since $r$ solves ALGP$(q)$ and $\kappa r'$ solves ALGP$(\kappa q)$, we can write
\[
r = [q + t(\lambda - \sigma)]^+, \quad \kappa r' = [\kappa q + \kappa t' (\lambda - \sigma')]^+
\]
for $t, t' \in \mathbb{R}_+$ and $\sigma, \sigma' \in \Sigma$. Indeed, for $T > \max(t, t')$ we can use (30) to write
\[
r = q + T (\lambda - \rho + z) \quad \text{for } \rho \in \Sigma, z \in \mathbb{R}_N^+, \text{ where } z_n = 0 \text{ if } r_n > 0
r' = q + T (\lambda - \rho' + z') \quad \text{for } \rho' \in \Sigma, z' \in \mathbb{R}_N^+, \text{ where } z'_n = 0 \text{ if } r'_n > 0.
\]
Now consider the value of $L(\cdot)$ along the trajectory from $\kappa r$ to $\kappa r'$. Along this trajectory,
\[
\frac{d}{du} L \left( \kappa r + (r' - r)u/T \right) \bigg|_{u=0} = (r' - r) \cdot f(\kappa r)/T
\geq (\rho - \rho' - z + z') \cdot f(\kappa r)
\geq (\rho - \rho' - z) \cdot f(\kappa r) \quad \text{since } z' \geq 0
\geq \rho \cdot f(\kappa r) \quad \text{since } z_n = 0 \text{ if } r_n > 0
\geq \rho \cdot f(\kappa r) = \max_{\pi \in \mathcal{S}} \pi \cdot f(\kappa r) \quad \text{for any } \rho \in \Sigma
= 0.
\]
The final equality is because $\rho f(r) = \max_\pi \pi f(r)$ by (29), so $\rho f(\kappa r) = \max_\pi \pi f(\kappa r)$ by Condition 4.1. Since $L(\cdot)$ is convex, it follows that $L(\kappa r') \geq L(\kappa r)$. This completes the proof. \hfill \qedsymbol

7. The critical fluid model (macroscopic). We now justify our claim that the optimization problems ALGD and ALGP capture the behaviour of the MW-$f$ scheduling algorithm. Consider a critically loaded system, i.e. $\lambda \in \partial \Lambda$, running MW-$f$. Theorem 7.3 shows that the fluid model heads towards a solution of ALGD/ALGP, and Lemma 7.4 says how long it takes to get there.

7.1. A basic feasibility result. The first hint that the optimization problems are related to the fluid model is the fact that feasibility is preserved along fluid model trajectories:

**Lemma 7.1 (Feasibility of fluid model solutions)** Suppose $\lambda \in \Lambda$. Consider any fluid model solution, for any scheduling algorithm, with initial queue size $q(0)$. Then $q(t)$ is feasible for $ALGD(W(q(0)))$ for all $t \geq 0$.

**Proof.** Let $\xi$ be any critically loaded virtual resource. By (13),

$$\xi \cdot q(t) = \xi \cdot q(0) + t \left( \xi \cdot \lambda - \xi \cdot \sigma(t) \right) + \xi \cdot z(t) \quad \text{where } \sigma(t) = \sum \pi s_\pi(t)/t$$

$$\geq \xi \cdot q(0) + t \left( \xi \cdot \lambda - \xi \cdot \sigma(t) \right) \quad \text{since } z \geq 0$$

$$= \xi \cdot q(0) + t \left( 1 - \xi \cdot \sigma(t) \right) \quad \text{since } \xi \text{ is critical for } \lambda$$

$$\geq \xi \cdot q(0) + t \left( 1 - \text{DUAL}(\sigma(t)) \right) \quad \text{since } \xi \text{ is feasible for DUAL}(\sigma(t))$$

$$= \xi \cdot q(0) + t \left( 1 - \text{PRIMAL}(\sigma(t)) \right) \quad \text{by strong duality}$$

$$\geq \xi \cdot q(0) \quad \text{since } \sigma(t) \in \Sigma \text{ so PRIMAL}(\sigma(t)) \leq 1.$$ 

This holds for all $\xi \in S^*(\lambda)$, therefore $q(t)$ is feasible. \hfill \qedsymbol

7.2. Fixed points of the fluid model.

**Definition 7.2 (Fixed Point)** Say that $q \in \mathbb{R}_+^N$ is a fixed point of the fluid model if all fluid model solutions starting at $q(0) = q$ satisfy $q(t) = q$ for all $t$.

**Theorem 7.3 (Characterization of fixed points of MW-$f$)** Consider a system running the MW-$f$ scheduling algorithm, with arrival rate vector $\lambda \in \Lambda$. Then, for $q \in \mathbb{R}_+^N$, the following are equivalent:

i. $q = \Delta W(q)$

---

10cf. Kelly and Williams [17], Theorems 5.1 & 5.3
ii. \( q \) is a fixed point

iii. there exists a fluid model solution with \( q(t) = q \) for all \( t \)

iv. \( \lambda \cdot f(q) = \max_{\pi \in \mathcal{S}} \pi \cdot f(q) \)

v. either \( q = 0 \), or \( \zeta(q) = f(q)/\max_{\pi} \pi \cdot f(q) \) is a critically-loaded virtual resource

Proof that \( (i) \implies (ii) \). Suppose that \( q = \Delta W(q) \), i.e. that \( q \) is optimal for ALGP\((q)\), and consider any fluid model solution which starts with \( q(0) = q \). On one hand, Lemma 4.2 says that \( L(q(t)) \leq L(q(0)) \). On the other hand, Lemma 7.1 says that \( q(t) \) is feasible for ALGP\((q)\). Since ALGP\((q)\) has a unique solution, it must be that \( q(t) = q \).

Proof that \( (ii) \implies (iii) \). It is easy to find a fluid model solution which starts at \( q(0) = q \): a limit of the microscopic model from Theorem 5.1 will do. By (ii), the queue size vector is constant.

Proof that \( (iii) \implies (iv) \). Suppose there is a fluid model solution with \( q(t) = q \). Since \( q(\cdot) \) is constant, \( \dot{L}(q(t)) = 0 \). Lemma 4.2 says that \( \dot{L}(q(t)) \leq 0 \), so the inequality in the proof must be tight for all \( t \), i.e.

\[
\lambda \cdot f(q) = \max_{\pi \in \mathcal{S}} \pi \cdot f(q)
\]

Proof that \( (iv) \implies (v) \). If \( q \neq 0 \) then certainly \( \max_{\pi} \pi \cdot \zeta(q) = 1 \), so \( \zeta(q) \) is a virtual resource. The condition of criticality, that \( \zeta(q) \cdot \lambda = 1 \), is a straightforward rewriting of (iv).

Proof that \( (v) \implies (i) \). If \( q = 0 \) the result is trivial. Otherwise, let \( r = \Delta W(q) \), i.e. let \( r \) solve ALGD\((W(q))\). Therefore it solves ALGP\((q)\), so \( r = [q + t(\lambda - \sigma)]^+ \) for some \( t \geq 0 \) and \( \sigma \in \Sigma \). Consider the value of \( L(\cdot) \) along the trajectory from \( q \) to \( r \):

\[
\frac{d}{du} L([q + (\lambda - \sigma)u]^+) \bigg|_{u=0} = (\lambda - \sigma) \cdot f(q) = (\lambda \cdot \zeta - \sigma \cdot \zeta) \max_{\pi} \pi \cdot f(q) \quad \text{writing} \quad \zeta \quad \text{for} \quad \zeta(q) \\
= (1 - \sigma \cdot \zeta) \max_{\pi} \pi \cdot f(q) \quad \text{since} \quad \zeta \quad \text{is optimal for DUAL(}\lambda\text{)} \\
\geq (1 - \text{DUAL(}\sigma\text{)}) \max_{\pi} \pi \cdot f(q) \quad \text{since} \quad \zeta \quad \text{is feasible for DUAL(}\sigma\text{)} \\
\geq 0 \quad \text{since} \quad \sigma \in \Sigma \quad \text{so} \quad \text{DUAL(}\sigma\text{)} \leq 1.
\]

By convexity, \( L(r) \geq L(q) \); but \( q \) is obviously feasible for ALGD\((W(q))\); so it must be optimal. By uniqueness of the solution, \( q = \Delta W(q) \).

7.3. Time to converge. The following lemma bounds how long it takes for a fluid model solution to get close to a fixed point. Use the norm \( |q| = \max_n |q_n| \). Given
\( \varepsilon > 0 \), let \( \mathcal{J}_\varepsilon \) be the set of queue size vectors which are within \( \varepsilon \) of their lifted versions,
\[
\mathcal{J}_\varepsilon = \{ \mathbf{q} \in \mathbb{R}^N_+ : |\mathbf{q} - \Delta W(\mathbf{q})| < \varepsilon \}.
\]
Also, for any fluid model solution with queue size process \( \mathbf{q}(\cdot) \), define \( h_\varepsilon(\mathbf{q}(\cdot)) \) to be the ‘sticking time’ for this set,
\[
h_\varepsilon(\mathbf{q}(\cdot)) = \inf \{ t \geq 0 : \mathbf{q}(s) \in \mathcal{J}_\varepsilon \text{ for all } s \geq t \}.
\]

Lemma 7.4 11 Fix an arrival rate vector \( \lambda \in \Lambda \), and a weight function \( f \) satisfying Condition 4.1. For any \( \varepsilon > 0 \) there exists an \( h_\varepsilon \) such that \( h_\varepsilon(\mathbf{q}(\cdot)) \leq h_\varepsilon \) for every MW-\( f \) fluid model solution which starts with \( \mathbf{q}(0) \in [0, 1]^N \).

Proof. Some definitions. Let
\[
\mathcal{D} = \{ \mathbf{q} \in \mathbb{R}^N_+ : L(\mathbf{q}) \leq L(1) \} \quad \text{for } L(\cdot) \text{ as in Lemma 4.2}
\]
\[
\mathcal{I} = \{ \mathbf{q} \in \mathcal{D} : \Delta W(\mathbf{q}) = \mathbf{q} \}
\]
\[
\mathcal{I}_\delta = \{ \mathbf{q} \in \mathcal{D} : |\mathbf{q} - \mathbf{r}| < \delta \text{ for some } \mathbf{r} \in \mathcal{I} \}
\]
\[
\mathcal{K}_\delta = \{ \mathbf{q} \in \mathcal{D} : L(\mathbf{q}) - L(\Delta W(\mathbf{q})) < \inf_{\mathbf{r} \in \mathcal{D} \setminus \mathcal{I}_\delta} (L(\mathbf{r}) - L(\Delta W(\mathbf{r}))) \}
\]
We will argue that the function \( K(\mathbf{q}) = L(\mathbf{q}) - L(\Delta W(\mathbf{q})) \) is decreasing along fluid model trajectories, so once you hit \( \mathcal{K}_\delta \) you stay there. We will then argue that \( \mathcal{I} \subset \mathcal{K}_\delta \subset \mathcal{I}_\delta \subset \mathcal{J}_\varepsilon \) for sufficiently small \( \delta \). Finally, we will bound the time it takes to hit \( \mathcal{K}_\delta \).

\( K \) is decreasing. Lemma 4.2 says that for any fluid model solution, \( L(\mathbf{q}(\cdot)) \) is decreasing. From Lemma 7.1, the feasible set for ALGP(\( \mathbf{q}(u) \)) is a subset of the feasible set for ALGP(\( \mathbf{q}(t) \)) for any \( u \geq t \geq 0 \), hence \( \Delta W(\mathbf{q}(u)) \geq \Delta W(\mathbf{q}(t)) \), i.e. \( \Delta W(\mathbf{q}(\cdot)) \) is increasing. Therefore \( K \) is decreasing (not necessarily strictly).

\( \mathcal{I} \subset \mathcal{K}_\delta \subset \mathcal{I}_\delta \subset \mathcal{J}_\varepsilon \). To show \( \mathcal{I} \subset \mathcal{K}_\delta \): The map \( \Delta(\cdot) \) is continuous by Lemma 6.3, and \( W(\cdot) \) and \( L(\cdot) \) are clearly continuous, so \( K(\cdot) \) is continuous; also the set \( \mathcal{D} \) is closed and bounded, and \( \mathcal{I}_\delta \) is open, so \( \mathcal{D} \setminus \mathcal{I}_\delta \) is closed and bounded; so the infimum in the definition of \( \mathcal{K}_\delta \) is attained at some \( \mathbf{r} \in \mathcal{D} \setminus \mathcal{I}_\delta \). Now, \( K(\mathbf{q}) > 0 \) for \( \mathbf{q} \in \mathcal{D} \setminus \mathcal{I} \), so \( K(\mathbf{r}) > 0 \). Yet \( K(\mathbf{q}) = 0 \) for \( \mathbf{q} \in \mathcal{I} \). Thus \( \mathcal{I} \subset \mathcal{K}_\delta \).

It is clear by construction that \( \mathcal{K}_\delta \subset \mathcal{I}_\delta \).

To show \( \mathcal{I}_\delta \subset \mathcal{J}_\varepsilon \): The map \( \Delta W(\cdot) \) is continuous, hence it is uniformly continuous on the closed and bounded set \( \mathcal{D} \), so for any \( \varepsilon > 0 \) there exists a \( \delta > 0 \) such that
\[
|\mathbf{q} - \mathbf{r}| < \delta \implies |\Delta W(\mathbf{q}) - \Delta W(\mathbf{r})| < \varepsilon / 2 \quad \text{for } \mathbf{q}, \mathbf{r} \in \mathcal{D}.
\]

11 cf. Kelly and Williams [17], Theorem 5.2, Lemma 6.3
If \( q \in I_{\delta} \) then it is within \( \delta \) of some \( r \in I \), hence
\[
|q - \Delta W(q)| \leq |q - r| + |r - \Delta W(r)| + |\Delta W(r) - \Delta W(q)| \\
< \delta + \epsilon/2 \\
< \epsilon \quad \text{for } \delta \text{ sufficiently small.}
\]

**Time to hit \( K_{\delta} \).** Consider first the rate of change of \( K(\cdot) \) while the process is in \( D \setminus K_{\delta} \):
\[
\dot{K}(q(t)) \leq \dot{L}(q(t)) = \lambda \cdot f(q(t)) - \max_{\pi \in S} \pi \cdot f(q(t)) \\
\leq \sup_{r \in D \setminus K_{\delta}} \left[ \lambda \cdot f(r) - \max_{\pi \in S} \pi \cdot f(r) \right] \\
\leq 0 \quad \text{by Lemma 4.2.}
\]

The supremum in (32) is of a continuous function of \( r \), taken over a closed and bounded set, hence the supremum is attained at some \( \hat{r} \in D \setminus K_{\delta} \). If the supremum were equal to 0 then \( \lambda \cdot f(\hat{r}) = \max_{\pi \in S} \pi \cdot f(\hat{r}) \) so \( \hat{r} \in I \) by Theorem 7.3; but \( \hat{r} \in D \setminus K_{\delta} \) and we just proved that \( I \subset K_{\delta} \); hence the supremum is some \( -\eta_{\delta} < 0 \).

Now consider any fluid model solution starting at \( q(0) \in [0,1]^N \). If \( q(0) \in K_{\delta} \) then \( h_\epsilon(q(\cdot)) = 0 \) so the theorem holds trivially. If not, then \( q(0) \leq 1 \) componentwise, so \( L(q(0)) \leq L(1) \), so \( q(0) \in D \); also \( L(q(t)) \) is decreasing so \( q(t) \in D \) for all \( t \geq 0 \). Now, \( \dot{K}(q(t)) \leq -\eta_{\delta} \) all the time that \( q(t) \in D \setminus K_{\delta} \), and this can’t go on for longer than \( H_\epsilon = \dot{K}(q(0))/\eta_{\delta} \leq L(1)/\eta_{\delta} \).

This proves that the process approaches \( I \) asymptotically. In all the examples we’ve looked at, the process actually hits \( I \) in bounded time; but we’ve not been able to prove this is generally true.

8. Critical load (microscopic). This section is devoted to proving multiplicative state space collapse under a critical load limit. Essentially, the results says that when the system is critically loaded (i.e. \( \lambda \in \partial \Lambda \)) then the queue size vector is constrained to lie in or close to the invariant set
\[
I = \{ q \in \mathbb{R}_+^N : q = \Delta^\lambda(W^\lambda(q)) \}.
\]

The development in this section follows closely the pattern of Bramson [4]. To help the reader familiar with that work, we footnote our results with references to the corresponding results of Bramson.

8.1. CL scaling. Consider a sequence of systems of the type described in Section 1.4, indexed by \( r \in \mathbb{N} \). Write \( X^r(\tau) = (Q^r(\tau), A^r(\tau), Z^r(\tau), S^r(\tau)) \), \( \tau \in \mathbb{Z}_+ \), for the
rth system. Define the scaled system \( \hat{x}^r(t) = (\hat{q}^r(t), \hat{a}^r(t), \hat{z}^r(t), \hat{s}^r(t)) \) for \( t \in \mathbb{R} \) by

\[
\begin{align*}
\hat{q}^r(t) &= Q^r(r^2 t)/r \\
\hat{a}^r(t) &= A^r(r^2 t)/r \\
\hat{z}^r(t) &= Z^r(r^2 t)/r \\
\hat{s}^r(t) &= S^r(\pi t)/r
\end{align*}
\]

after extending the domain of \( X^r(\cdot) \) to \( \mathbb{R}_+ \) by linear interpolation in each interval \((\tau - 1, \tau)\). Note the difference between this scaling and the fluid model scaling in Section 5.1. Whereas the fluid limit makes sense for a single rescaled arrival process \( a^r(t) = A(\tau t)/r \), the critical-load limit requires us to consider a sequence of arrival processes.

8.2. Main result. Our goal is to study the dynamics of \( \hat{x}^r(t) \), for \( t \) in a fixed interval \([0, T]\), as \( r \to \infty \). We will assume that for every \( r \) the arrival process \( A^r \) satisfies assumptions (3)–(5). We will also make the following uniformity assumptions:

Let the arrival rate vector for the \( r \)th system be \( \lambda^r \). Assume that

\[
\lambda^r \to \lambda \quad \text{for some } \lambda \in \Lambda
\]

Note: the result is most interesting when \( \lambda \in \partial \Lambda \), e.g. in the widely studied heavy traffic scaling, in which \( \lambda^r = \lambda - \Gamma/r \) for some \( \Gamma \geq 0 \) and \( \lambda \in \partial \Lambda \); but it still holds when \( \lambda \in \Lambda^0 \).

Assume that the arrival processes all have stationary increments, and that the bound on the increments given by (4) is uniform in \( r \). Assume that the limit (5) is uniform in \( r \), in the sense that for all \( \varepsilon > 0 \) and \( T > 0 \) there is a non-increasing function \( R : \mathbb{R}_+ \to \mathbb{R}_+ \) and a value \( r_0 \in \mathbb{N} \) such that

\[
P\left( \sup_{t \in [0, T]} \left| \frac{\Delta^r(tz)}{z} - \lambda^r t \right| < \varepsilon \right) = 1 - o(R(z)) \quad \text{as } z \to \infty, \text{ uniformly in } r \geq r_0.
\]

Also assume that the initial queue size distribution is non-random and satisfies

\[
\lim_{r \to \infty} \hat{q}^r(0) = \hat{q}_0 \quad \text{for some } \hat{q}_0 \in \mathcal{I}.
\]

Consider the MW-\( \cdot \)f scheduling algorithm for some \( f \) satisfying Condition 4.1, and write \( \Delta W \) for the combined workload and lifting map \( q \mapsto \Delta f_\lambda(W^\lambda(q)) \). These are the properties of \( \Delta W \) we’ll need, and they all hold for any \( \lambda \in \Lambda \):

**Condition 8.1**

i. \( \Delta W \) is continuous (Lemma 6.3)

ii. If \( q = \Delta W(q) \) then \( \Delta W(\kappa q) = \kappa \Delta W(q) \) for all \( \kappa > 0 \) (Lemma 6.4)

iii. For any \( \varepsilon > 0 \) there exists some \( H_\varepsilon \) such that, for every fluid model solution with bounded initial value, \( |q(t) - \Delta W(q(t))| < \varepsilon \) for all \( t \geq H_\varepsilon \) (Lemma 7.4)
iv. \( q \) is a fixed point \( \iff q = \Delta W(q) \) (Theorem 7.3)

Theorem 8.2 12 Fix \( T > 0 \), and write \( \|x(\cdot)\| \) for \( \sup_{0 \leq t \leq T} |x(t)| \). Under the above assumptions, for any \( \varepsilon > 0 \),

\[
\mathbb{P}\left( \left\| \hat{q}^r(\cdot) - \Delta W(\hat{q}^r(\cdot)) \right\| < \varepsilon \right) = 1 - o(rR(r)) \quad \text{as} \quad r \to \infty.
\]

This is only interesting when the arrival process assumption (34) is satisfied with \( R(r) = O(1/r) \), in which case the theorem yields convergence in probability.

8.3. An illustration. Here is a simulation experiment which illustrates Theorem 8.2. Consider an input-queued switch with three inputs and three outputs, as described in Section 1.5. We simulated the switch running the plain MW scheduling algorithm, for Bernoulli i.i.d. arrivals with arrival rate matrix

\[
\lambda = \begin{pmatrix}
0.1413 & 0.4626 & 0.3910 \\
0.4626 & 0.0055 & 0.5268 \\
0.3910 & 0.5268 & 0.0771
\end{pmatrix}
\]

where \( \lambda_{i,j} \) is the arrival rate at input port \( i \) of packets destined for output port \( j \). As explained in Section 3.5, the principal critically-loaded virtual resources are those input ports and output ports for which the total arrival rate is 1. In this case, the arrival rates at each input and output are close enough to 1 that we will treat them all as critically loaded. That is, we consider the workload vector

\[
W(Q) = (r_1 \cdot Q, r_2 \cdot Q, r_3 \cdot Q; c_1 \cdot Q, c_2 \cdot Q, c_3 \cdot Q).
\]

Figure 2 shows the evolution of \( Q(\cdot) \). The cell in row \( i \) and column \( j \) plots \( Q_{i,j}(\tau) \) as a function of time \( \tau \). The horizontal axis runs for 5000 timesteps, and the vertical axis runs from 0 to 75 packets. The plot shows \( Q(\cdot) \) and superimposes \( \Delta W(Q(\cdot)) \). It can be seen that the two match very closely, except for \( Q_{2,2} \) where the arrival rate is so low that the queue ‘can’t keep up’, and for a small knock-on effect on the other queues in the same row or column as \( Q_{2,2} \). If we had run the simulation for longer and rescaled the plots, the match would look closer.

In effect, the 9-dimensional queue-size matrix \( Q \) can be described using only the 6-dimensional workload vector \( W(Q) \), which itself is a function of the 5-dimensional vector \( (r_1 \cdot Q, r_2 \cdot Q; c_1 \cdot Q, c_2 \cdot Q; 1 \cdot Q) \).

12cf. Bramson [4], Theorem 1
8.4. **Proof of main result.** The idea of the proof is to split the process into \( \lfloor rT \rfloor + 1 \) pieces starting at 0, \( r \), 2\( r \), ..., and to look at each piece under the fluid scaling. Each fluid-scaled piece is close to a fluid model solution by Theorem 5.1. Figure 3 illustrates the idea. At every point \( t \in [0, T] \) we will approximate \( \hat{q}(t) \) by one of these fluid model solutions. If the fluid model solution has run for long enough before we observe it at \( t \), then it is in or close to the invariant set, by Condition 8.1(iii). Therefore the original process is close to the invariant set.

Define the fluid-scaled parts of the original process \( x^{r,m}(\cdot) \) for \( 0 \leq m \leq \lfloor rT \rfloor \) by

\[
\begin{align*}
z_{r,m} &= |Q_r^{r}(rm)| \lor r \\
q^{r,m}(u) &= Q_r^{r}(rm + z_{r,m}u) / z_{r,m}
\end{align*}
\]
\[ a^{r,m}(u) = \frac{(A^r(rm + z_{r,m}u) - A^r(rm))}{z_{r,m}} \]
\[ z^{r,m}(u) = \frac{(Z^r(rm + z_{r,m}u) - Z^r(rm))}{z_{r,m}} \]
\[ s^{r,m}(u) = \frac{(S^r(rm + z_{r,m}u) - S^r(rm))}{z_{r,m}}. \]

We will use \( u \) to denote a time on the fluid scale, as opposed to \( t \) which is a time on the CL scale, i.e. \( r^2t \) on the timescale of the original process. For any CL time \( t \), we can choose \( m \) and \( u \) such that \( r^2t = rm + z_{r,m}u \).

The proof of Theorem 8.2 starts with two lemmas. The first lemma is about how to choose \( m \) and \( u \) sensibly, with \( u \) large enough that the fluid-scaled parts have nearly reached a fixed point by the time we observe them, and small enough that we only need to run them for over a finite timescale. The second lemma says that for well-behaved sample points \( \omega \) the quantity we are estimating in the statement of Theorem 8.2 can be bounded above (pathwise, not just in probability). Finally, the main proof says that these well-behaved sample points have high probability, and establishes the result.

In the two lemmas that follow, let \( \varepsilon > 0 \), let \( H_\varepsilon \) be as in Condition 8.1(iii), and let \( T^\text{fluid}_\varepsilon = 2(A^{\max} \lor 1)(H_\varepsilon + 1) \).

**Lemma 8.3 (Choice of \( m \) and \( u \))** \(^{13}\) Define \( m^* = m^*(r,t) \) and \( u^* = u^*(r,t) \) by

\[
m^* = \min \left\{ m \in \mathbb{Z}_+ : rm \leq r^2t \leq rm + T^\text{fluid}_\varepsilon z_{r,m} \right\}, \quad u^* = \frac{r^2t - rm^*}{z_{r,m^*}}.
\]

This is a sound definition (i.e. the set for \( m^* \) is non-empty). Also, either \( m^* = 0 \) and \( 0 \leq u^* \leq T^\text{fluid}_\varepsilon \), or \( m^* > 0 \) and \( H_\varepsilon < u^* \leq T^\text{fluid}_\varepsilon \).

**Proof.** The set for \( m^* \) is non-empty because \( z_{r,m} \geq r \) and \( T^\text{fluid}_\varepsilon \geq 1 \). The upper bound for \( u^* \) in either case is trivial. To prove the lower bound for \( u^* \) when \( m^* > 0 \), note that \( m^* - 1 \) is not suitable and so

\[
u^* = \frac{r^2t - rm^*}{z_{r,m^*}} > \frac{T^\text{fluid}_\varepsilon z_{r,m^* - 1} - r}{z_{r,m^*}} \quad \text{since} \quad r^2t > r(m^* - 1) + T^\text{fluid}_\varepsilon z_{r,m^* - 1}
\]

To bound \( u^* \), observe

\[
z_{r,m} = |Q(rm)| \lor r \leq |Q(r(m - 1)) + rA^{\max}| \lor r \quad \text{since arrival rate} \leq A^{\max}
\]
\[
\leq |Q(r(m - 1))| \lor r + r(A^{\max} \lor 1)
\leq 2(A^{\max} \lor 1)z_{r,m - 1} \quad \text{since} \quad z_{r,m - 1} \geq r
\]

\(^{13}\)cf. Bramson [4], Lemmas 6.2 and 6.3
and substituting this back into the earlier bound for $u^*$

$$u^* > \frac{T_{\varepsilon}^{\text{fluid}}}{2(A_{\max} \lor 1)} - \frac{r}{z_{r,m^*}} \geq \frac{T_{\varepsilon}^{\text{fluid}}}{2(A_{\max} \lor 1)} - 1 = H_{\varepsilon} \text{ by choice of } T_{\varepsilon}^{\text{fluid}}.$$  

**Lemma 8.4 (Pathwise state space collapse)**  

Pick $t \in [0, T]$, and $r \in \mathbb{N}$, and $m^*$ and $u^*$ as in Lemma 8.3. Let $\text{FMS}_\varepsilon$ be the $\varepsilon$-fattening of the set of fluid model solutions over the interval $[0, T_{\varepsilon}^{\text{fluid}}]$, as in Theorem 5.1, and let $\text{FMS}_\varepsilon(q_0)$ be the $\varepsilon$-fattening of the set of those fluid model solutions which start at $q(0) = q_0$. If $x^{r,m^*}(\cdot) \in \text{FMS}_\varepsilon$ and $x^{r,0}(\cdot) \in \text{FMS}_\varepsilon(q_0)$ where $q_0 = \hat{q}_0/(|\hat{q}_0| \lor 1)$, then

$$\begin{align*}
|\hat{q}^*(t) - \Delta W(q^*(t))| &\leq \frac{|\hat{q}^*(t) - \Delta W(q^*(t))|}{z_{r,m^*}/r} \leq 2\varepsilon + mc_\varepsilon(\Delta W) \tag{36}
\end{align*}$$

where $mc_\varepsilon(\Delta W)$ is the modulus of continuity of the map $q \mapsto \Delta W(q)$ over $q \in [0, q_{\max}]^N$ and $q_{\max} = 1 + \varepsilon + A_{\max}T_{\varepsilon}^{\text{fluid}}$.

**Proof.** The first inequality is trivially true because $z_{r,m}/r \leq \sup_t |\hat{q}^*(t)| \lor 1$ for all $m$, from the definition of $z_{r,m}$.

Unwrapping the CL scaling and wrapping it up again in the fluid model scaling, the middle term in the claimed inequality is

$$MT = |q^{r,m^*}(u^*) - \Delta W(q^{r,m^*}(u^*)|.$$ 

Since $x^{r,m^*} \in \text{FMS}_\varepsilon$ we can choose some $x \in \text{FMS}$ such that $\|x^{r,m} - x\| < \varepsilon$. Writing $q$ for the queue component of $x$,

$$MT \leq |q^{r,m^*}(u^*) - q(u^*)| + |q(u^*) - \Delta W(q(u^*))| + |\Delta W(q(u^*)) - \Delta W(q^{r,m^*}(u^*))|$$

$$= (37a) + (37b) + (37c) \text{ respectively.}$$

We can bound each term as follows:

- (37a) is $< \varepsilon$ since $\|x^{r,m^*} - x\| < \varepsilon$ by choice of $x$.
- (37b) is $\leq \varepsilon$ when $m^* > 0$: We know from Lemma 8.3 that $u^* > H_{\varepsilon}$; also $q^{r,m^*}(0) \in [0,1]^N$ by virtue of the scaling term $z_{r,m^*}$. The requirements of Condition 8.1(iii) are met, and we obtain the inequality.
- (37b) is $= 0$ when $m^* = 0$: In this case, we can by the assumption of the lemma choose $x \in \text{FMS}$ with $q(0) = q_0$. By (35), $\hat{q}_0 \in \mathcal{I}$, that is $\hat{q}_0 = \Delta W(q_0)$, therefore by Condition 8.1(ii) $q_0 \in \mathcal{I}$, therefore by Condition 8.1(iv) the fluid model solution $q(\cdot)$ stays constant at $q_0$ and so (37b) = 0.

\footnote{cf. Bramson [4], Corollary 6.2}
(37c) is \(\leq mc_\varepsilon(\Delta W)\): We know \(q^{r,m*}(0) \in [0,1]^N\) by virtue of the scaling term \(z_{r,m*}\); also the maximum arrival rate is \(A^{\text{max}}\); thus \(q^{r,m*}(u^*) \in [0, q^{\text{max}} - \varepsilon]^N\). Also, \(\|q^{r,m*}(\cdot) - q(\cdot)\| < \varepsilon\), which implies \(q(u^*) \in [0, q^{\text{max}}]^N\). The inequality follows from the definition of the modulus of continuity. \(\square\)

**Proof of Theorem 8.2.** Fix \(\varepsilon > 0\) and \(T > 0\), and consider the fluid-scaled versions \(x^{r,m}(\cdot)\) for \(r \in \mathbb{N}\) and \(0 \leq m \leq \lfloor rT \rfloor\). Recall that \(x^{r,m}\) is the fluid-scaled version of \(X^r\), time-shifted by \(rm\) and scaled in space and time by \(z_{r,m}\). We want to apply Lemma 8.4, but we first need to show that there is a high probability that the fluid-scaled versions lie in \(\text{FMS}_\varepsilon\), by appealing to Theorem 5.1.

**Applying fluid limit to \(x^{r,m}\).** Consider the sequence consisting of \(x^{1,0}, \ldots, x^{1,\lfloor T \rfloor}\) followed by \(x^{2,0}, \ldots, x^{2,\lfloor 2T \rfloor}\) and so on. We shall show that over the interval \([0, T_{\text{fluid}}]\) this sequence satisfies the assumptions of the fluid limit theorem, Theorem 5.1, namely (3)–(5) and (21)–(24). A note of caution: the role of \(r\) in the fluid limit is here played by the pair \((r, m)\). The scaling factor \(r'\) in the fluid limit is here given by \(z_{r,m}\); note that \(z_{r,m} \to \infty\) as required for the fluid limit scaling.

First, equations (3)–(5) apply straightforwardly, because we assumed that for every \(r\) the arrival process \(A^r\) satisfies those equations; equation (5) holds for the time-shifted version because \(A^r\) is assumed to have stationary increments. Equation (21) holds because of (33). Equation (22) holds because we assumed that increments in the arrival processes are uniformly bounded. For equation (23), note that by the assumption of stationarity

\[
\mathbb{P} \left( \sup_{t \in [0, T_{\text{fluid}}]} \left| a^{r,m}(t) - \lambda^r t \right| < \varepsilon \right) = \mathbb{P} \left( \sup_{t \in [0, T_{\text{fluid}}]} \left| \frac{A^r(z_{r,m}t)}{z_{r,m}} - \lambda^r t \right| < \varepsilon \right);
\]

and recall assumption (34) which says there is an \(r_0\) such that

\[
\mathbb{P} \left( \sup_{t \in [0, T_{\text{fluid}}]} \left| \frac{A^r(zt)}{z} - \lambda^r t \right| < \varepsilon \right) = 1 - o(R(z)) \quad \text{uniformly in } r \geq r_0;
\]

putting these together with the fact that \(z_{r,m} \to \infty\) as \(r \to \infty\) we conclude

\[
\mathbb{P} \left( \sup_{t \in [0, T_{\text{fluid}}]} \left| a^{r,m}(t) - \lambda^r t \right| < \varepsilon \right) = 1 - o(R(z_{r,m}))
\]
as required for (23). Finally, (24) holds because

\[
|q^{r,m}(0)| = \frac{|Q^r(rm)|}{z_{r,m}} = \frac{|Q^r(rm)|}{|Q^r(rm)| \lor r} \leq 1.
\]

We have shown that the sequence of \(x^{r,m}\) satisfies all the assumptions of Theorem 5.1. We conclude that

\[
\mathbb{P}(x^{r,m}(\cdot) \in \text{FMS}_\varepsilon) = 1 - o(R(z_{r,m})).
\]
Furthermore, since we assumed that \( R \) is non-increasing, the probability is in fact \( 1 - o(R(r)) \). Applying the union bound to \( 0 \leq m \leq \lfloor rT \rfloor \),

\[
P(x^{r,m} \in \text{FMS}_\varepsilon \text{ for all } 0 \leq m \leq \lfloor rT \rfloor) = 1 - o(rR(r)).
\]

\( (38) \)

 Applying fluid limit to \( x^{r,0} \). All the previous checks still apply. Furthermore, equation (25) also applies, because

\[
q^{r,0}(0) = \frac{Q^{r}(0)}{\|Q^{r}(0)\|} \rightarrow q_0
\]

where \( q_0 \) is defined as in Lemma 8.4. Applying Theorem 5.1, we conclude that

\[
P(x^{r,0} \in \text{FMS}_\varepsilon(q_0)) = 1 - o(R(z_{r,0})).
\]

Since \( R \) is non-increasing and \( z_{r,0} \geq r \), this probability is \( 1 - o(R(r)) \).

Wrapping up. Let \( G_r \) denote the ‘good’ event in which both the events in (38) and (39) hold. Putting the two probabilities together, \( P(G_r) = 1 - o(R(r)) \). Furthermore, when \( G_r \) is true, the conditions of Lemma 8.4 are satisfied for every \( t \in [0,T] \), which implies

\[
\|\hat{q}(\cdot) - \Delta W(\hat{q}(\cdot))\| \leq 2\varepsilon + mc_\varepsilon(\Delta W).
\]

Finally, we can shrink the error term \( 2\varepsilon + mc_\varepsilon(\Delta W) \) by making \( \varepsilon \) smaller. (This is because \( \Delta W \) is continuous, therefore it is uniformly continuous on compact sets, therefore \( mc_\varepsilon(\Delta W) \rightarrow 0 \) as \( \varepsilon \rightarrow 0 \).) By shrinking it appropriately, we obtain Theorem 8.2.

\[\square\]

9. Inferring performance from state space collapse (teleology). In this section we will explain the implications of state space collapse for performance. We will do this in the context of an input-queued switch running MW-\( \alpha \).

The results in Section 7 and 8 show that the vector of queue lengths \( q(\cdot) \) spends ‘most of its time’ in states where \( q \approx \Delta W(q) \), in other words that the queue size is restricted to a neighbourhood of the invariant set

\[
\mathcal{I} = \{ q \in \mathbb{R}_+^N : q = \Delta W(q) \}.
\]

We don’t need to keep track of the vector of queue lengths \( q(\cdot) \); it is sufficient to keep track of the workload vector \( W(q(\cdot)) \) which lies in the collapsed invariant set

\[
\mathcal{W} = \{ W(q) : q \in \mathcal{I} \}.
\]
The workload vector roams inside \( \mathcal{W} \). It cannot leave \( \mathcal{W} \); whenever it hits a boundary the scheduling algorithm chooses schedules to keep it inside \( \mathcal{W} \). Typically this is achieved by idling at some queue (we will prove so for the input-queued switch).

A natural goal is to choose a scheduling algorithm which makes \( \mathcal{I} \) as large as possible, or equivalently makes \( \mathcal{W} \) as large as possible, in the hope that this will reduce idling. Motivated by this, we will compare scheduling algorithms by reasoning about the geometry of \( \mathcal{W} \): loosely speaking, the larger \( \mathcal{W} \), the better should be the performance of the algorithm. Typically \( \mathcal{W} \) is a lower-dimensional space than \( \mathcal{I} \), and it is easier to reason about. The lifting map tells us the disposition of queue sizes, and in particular it tells us which queues are empty and in danger of idling (that is, about the reflection angles of the queue-length process). This geometric approach does not however tell us about the interplay between size of \( \mathcal{W} \) and reflection angles, that is, about how much time the system is likely to spend in any particular part of \( \mathcal{W} \).

A similar relation between state space geometry and algorithmic performance has been observed by Kelly and Williams [17] in the context of bandwidth sharing.

9.1. Example: 2 \times 2 \ switch running MW-\( \alpha \). Consider a 2 \times 2 switch running MW-\( \alpha \), with arrival rate vector \( \lambda \in \partial \Lambda \). Suppose that \( \lambda = (\lambda_{1,1}, 1 - \lambda_{1,1}, 1 - \lambda_{1,1}, \lambda_{1,1}) \) for some \( \lambda_{1,1} \in (0, 1) \). This is a maximally loaded switch, i.e. all input ports and output ports are critically loaded. As described in Section 3.5, the critical workloads are

\[
\mathcal{W}(q) = (w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2}) = (r_{1,1} \cdot q, r_{2,1} \cdot q, c_{1,1} \cdot q, c_{2,1} \cdot q).
\]

There is a straightforward bijection between \( \mathcal{W}(q) \) and

\[
\hat{\mathcal{W}}(q) = (w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2}) = (r_{1,1} \cdot q, c_{1,1} \cdot q, 1 \cdot q)
\]

which we prefer because it has no redundancy. Let \( \hat{\mathcal{I}} \) be the image of \( \hat{\mathcal{W}}(q) \) for \( q \in \mathcal{I} \).

To find \( \mathcal{I} \), we use a characterization from Theorem 7.3, which says that for \( q \in \mathbb{R}_+^{2\times 2} \), \( q \in \mathcal{I} \) if and only if \( \lambda \cdot q^\alpha = \max_{\pi, \pi} \pi \cdot q^\alpha \), which becomes

\[
\lambda_{1,1}(q_{1,1}^\alpha + q_{2,2}^\alpha) + (1 - \lambda_{1,1})(q_{1,2}^\alpha + q_{2,1}^\alpha) = (q_{1,1}^\alpha + q_{2,2}^\alpha) \land (q_{1,2}^\alpha + q_{2,1}^\alpha)
\]

Since \( \lambda_{1,1} \in (0, 1) \),

\[
\mathcal{I} = \left\{ q \in \mathbb{R}_+^{2\times 2} : q_{1,1}^\alpha + q_{2,2}^\alpha = q_{1,2}^\alpha + q_{2,1}^\alpha \right\}.
\]
Since \( q_{1,2} = w_1 - q_{1,1} \) etc., we can rewrite the equation as
\[
q_{i,j}^\alpha + (w_\cdot - w_i - w_j + q_{i,j})^\alpha = (w_i - q_{i,j})^\alpha + (w_j - q_{i,j})^\alpha \quad \text{for each } i, j \in \{1, 2\}.
\]
The set \( \hat{\mathcal{W}} \) is the set of \( \hat{w} \) which arise from \( q \geq 0 \), namely those \( \hat{w} \) which satisfy
\[
\text{(40)} \quad w_\cdot \leq w_i + w_j + (w_i^\alpha + w_j^\alpha)^{1/\alpha} \quad \text{for each } i, j \in \{1, 2\}.
\]
In the interior of \( \hat{\mathcal{W}} \), i.e. when \( w > 0 \) componentwise, it is easy to see that \( q > 0 \) componentwise; thus there cannot be any idling. At the boundary of \( \hat{\mathcal{W}} \), some queues are empty and so the scheduling algorithm may choose a matching which results in wasted service. The boundary where the only tight constraint is \( w_\cdot = w_i + w_j + (w_i^\alpha + w_j^\alpha)^{1/\alpha} \) has \( q_{i,j} = 0 \) and the other three queues non-empty, therefore idling can only occur on \( q_{i,j} \). If there is some amount \( \delta \) of idling here, it has the impact of increasing \( \hat{\mathcal{W}}(q) = (r_1 \cdot q, c_1 \cdot q, 1 \cdot q) \) by \( \delta \) in the direction \((1_i = 1, 1_j = 1, 1)\). The four direction vectors for idling (for \( i, j \in \{1, 2\} \)) are referred to as the reflection angles at the four surfaces defined by (40).

Now, it is a standard inequality that for any \( x, y \in \mathbb{R}_+ \), and any \( 0 < \alpha < \beta \),
\[
(x^\alpha + y^\alpha)^{1/\alpha} \geq (x^\beta + y^\beta)^{1/\beta}.
\]
Applying this inequality to (40), it follows that the collapsed invariant set \( \hat{\mathcal{W}} \) becomes larger as \( \alpha \) decreases. As we have discussed, this hints at improved performance.

10. The overloaded fluid model (macroscopic & teleology). We now study the behaviour of the MW-\( f \) scheduling algorithm when the system is overloaded, i.e. \( \lambda \not\in \Lambda \). Obviously, when the system is overloaded, some queues will grow unboundedly. We will characterize how fast the queues grow, by means of an optimization problem which is closely related to the lifting map. First, an example which illustrates that MW can be ‘misled’ in overload.

10.1. Example: an overloaded input-queued switch. Consider a \( 2 \times 2 \) input-queued switch, with two possible arrival rate matrices
\[
\lambda^{\text{critical}} = \begin{pmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{pmatrix} \quad \text{or} \quad \lambda^{\text{overload}} = \begin{pmatrix} 0.3 & 1.0 \\ 0.7 & 0.3 \end{pmatrix}
\]
where the two possible service actions are
\[
\pi^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \pi^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.
\]
In the critically loaded case, the system can only be stabilized by \( \sigma^{\text{crit}} = 0.3\pi^1 + 0.7\pi^2 \), and it is straightforward using the fluid model equations to check that MW
will eventually achieve this service rate, for any initial queue size. However, in the overloaded case, starting from $q(0) = 0$, MW will achieve

$$q(t) = \begin{pmatrix} 0.1t \\ 0.2t \\ 0 \\ 0.1t \end{pmatrix}$$

by serving at rate $0.2\pi + 0.8\pi^2$, which means that queue $q_{2,1}$ is idling, and the total departure rate is 1.9. A different scheduling algorithm might have chosen to serve at rate $\sigma^{\text{crit}}$, which would result in a total departure rate of 2.

Now let’s generalize: suppose this switch is running MW-$\alpha$. According to Theorem 10.2 below, $q(t)/t$ converges as $t \to \infty$ to the (unique) solution to the following optimization problem:

$$\text{minimize } \sum_{i,j} r_{i,j}^{1+\alpha}$$

over $r \in \mathbb{R}^{2\times2}_+$

such that $r_{1,1} + r_{1,2} \geq 0.3$ and $r_{1,2} + r_{2,2} \geq 0.3$.

It is easy to check that the solution is

$$\hat{q} = \frac{0.3}{1 + 2^{-1/\alpha}} \begin{pmatrix} 2^{-1/\alpha} & 1 \\ 0 & 2^{-1\alpha} \end{pmatrix}$$

and the total departure rate is $2 - 0.3 \times 2^{-1/\alpha}/(1 + 2^{-1/\alpha})$. Observe that as $\alpha \to 0$, the overall throughput approaches 2, which is the largest possible. We will argue in Section 12 that a similar result holds for any overloaded switched network.

10.2. Two optimization problems. Since the switched network is overloaded, i.e. $\lambda \notin \Lambda$, the solution to DUAL($\lambda$) is $> 1$. Let $S^\dagger = S^\dagger(\lambda)$ be the set of extreme feasible solutions $\xi$ such that $\xi \cdot \lambda > 1$; call these the principal overloaded virtual resources. Now consider the optimization problem ALGP$^\dagger$:

$$\text{minimize } L(r)$$

over $r \in \mathbb{R}^N_+$

such that $r \geq \lambda - \sigma$ for some $\sigma \in \Sigma$,

and also the optimization problem ALGD$^\dagger$ in which the last constraint is replaced by

$$r \cdot \xi \geq \lambda \cdot \xi - 1 \text{ for all } \xi \in S^\dagger(\lambda).$$

The objective function $L(\cdot)$ is strictly convex, and the feasible set for ALGD$^\dagger$ is non-empty and convex, therefore the problem has a unique solution. Furthermore, the two problems are identical, by the following lemma.
Lemma 10.1 \( r \) is feasible for ALGD\(^\dagger \) \( \iff \) \( r \) is feasible for ALGP\(^\dagger \).

Proof of \( \iff \). Let \( r \) be feasible for ALGP\(^\dagger \). Then \( r \cdot \xi \geq \lambda \cdot \xi - \sigma \cdot \xi \geq \lambda \cdot \xi - 1 \) for any virtual resource \( \xi \), where the last inequality is because \( \xi \) satisfies the constraint in DUAL(\( \lambda \)).

Proof of \( \Rightarrow \). Assume that \( r \) is feasible for ALGD\(^\dagger \). Let \( \sigma' = \lambda - r \). We will shortly show that the solution to DUAL(\( \sigma' \)) is \( \leq 1 \). This implies that the solution to PRIMAL(\( \sigma' \)) is \( \leq 1 \), which means that \( \sigma' \leq \sigma \) for some \( \sigma \in \Sigma \), and so \( r \geq \lambda - \sigma \) i.e. \( r \) is feasible for ALGP\(^\dagger \).

It remains to show that the solution to DUAL(\( \sigma' \)) is \( \leq 1 \), i.e. \( \zeta \cdot \sigma' \leq 1 \) for all virtual resources \( \zeta \). Without loss of generality, assume that \( \zeta \) is an extreme feasible point. Either \( \zeta \) is an overloaded virtual resource, \( \zeta \in \mathcal{S}^\dagger \), in which case
\[
\zeta \cdot \sigma' = \zeta \cdot \lambda - \zeta \cdot r \leq 1 \quad \text{since } r \text{ is feasible for ALGD}^\dagger,
\]
or \( \zeta \) is not overloaded in which case
\[
\zeta \cdot \sigma' = \zeta \cdot \lambda - \zeta \cdot r \leq \zeta \cdot \lambda \quad \text{since } r \geq 0
\leq 1 \quad \text{since } \zeta \text{ is not overloaded}.
\]
Either way, \( \zeta \cdot \sigma' \leq 1 \). This is true for all virtual resources \( \zeta \), hence DUAL(\( \sigma' \)) \( \leq 1 \).

10.3. Dynamics of the overloaded system.

Theorem 10.2 (Convergence of scaled queue size) Consider a system running the MW-f scheduling algorithm with arrival rate vector \( \lambda \). Let \( q(t)/t \) be any fluid model solution. Then \( q(t)/t \) converges to the unique solution to ALGP\(^\dagger \).

Proof. Let \( \tilde{q} \) be the solution to ALGP\(^\dagger \); by continuity of the constraint there exists some \( \tilde{\sigma} \in \Sigma \) such that \( \tilde{q} \geq \lambda - \tilde{\sigma} \). Now observe that, at instants where the fluid model solution is differentiable,
\[
\frac{d}{dt} L\left( \frac{q(t)}{t} \right) = f\left( \frac{q(t)}{t} \right) \cdot \left( \tilde{q}(t) - \frac{q(t)}{t} \right)
= f\left( \frac{q(t)}{t} \right) \cdot \left( \lambda - \sigma(t) - \frac{q(t)}{t} \right) \quad \text{where } \sigma(t) = \sum_{\pi} \hat{\lambda}_{\pi}(t) \pi
\leq f\left( \frac{q(t)}{t} \right) \cdot \left( \lambda - \tilde{\sigma} - \frac{q(t)}{t} \right) \quad \text{using (18) and Condition 4.1}
\leq f\left( \frac{q(t)}{t} \right) \cdot \left( \tilde{q} - \frac{q(t)}{t} \right) \quad \text{since } \lambda - \tilde{\sigma} \leq \tilde{q}
\leq L(\tilde{q}) - L\left( \frac{q(t)}{t} \right) \quad \text{since } L(\cdot) \text{ is convex}
\leq 0 \quad \text{since } \tilde{q} \text{ is optimal and } q(t)/t \text{ is feasible for ALGP}^\dagger.
\]
This shows that \( L(q(t)/t) \) is decreasing. To show that \( q(t)/t \to \hat{q} \) it will be useful to work with a transform of \( L \): define \( g(u) = L(q(t)/t) \) where \( t = e^u \). Like \( L \) this function is differentiable almost everywhere, and

\[
g'(u) = t \frac{d}{dt} L\left(\frac{q(t)}{t}\right) \leq 0.
\]

We will prove that \( q(t)/t \to \hat{q} \), arguing by contradiction.

Suppose that \( q(t)/t \neq \hat{q} \). Then there exists some \( \varepsilon > 0 \) such that \( |q(t)/t - \hat{q}| > \varepsilon \) for arbitrarily large \( t \). We show below (under Duration of excursions) that on each such occasion, there is some \( h > 0 \) such that \( |q(s)/s - \hat{q}| \geq \varepsilon/2 \) for at least the interval \( s \in [t, (1+h)t] \), i.e. for the interval \( u \in [\log t, \log(1+h) + \log t] \). Furthermore, we can find some \( \eta > 0 \) such that for every such interval

\[
L(\hat{q}) - L\left(\frac{q(s)}{s}\right) \leq -\eta.
\]

Therefore, by (41) and (42) it follows that \( g'(u) \leq -\eta \).

To see (43), note that the continuous function \( L(\cdot) \) must attain its infimum on the compact set

\[\{ r : |r| \leq |\lambda| \text{ and } |r - \hat{q}| \geq \varepsilon/2 \}\]

and that \( q(s)/s \) lies in this set for \( s \in [t, (1+h)t] \).

We have shown that unless \( q(t)/t \to \hat{q} \), there exist arbitrarily large \( t \) such that \( g'(u) \leq -\eta \) for \( u \in [\log t, \log(1+h) + \log t] \). Furthermore \( g'(u) \leq 0 \). Therefore \( g(u) \to -\infty \), and this in turn implies \( L(q(t)/t) \to -\infty \). But this is impossible, by definition of \( L \). Therefore \( q(t)/t \to \hat{q} \).

**Duration of excursions.** It remains to find a lower bound for the duration of excursions away from \( \hat{q} \). Suppose that \( |q(t)/t - \hat{q}| > \varepsilon \) for some \( t \). Suppose \( s > t \) is such that \( |q(s)/s - \hat{q}| < \varepsilon/2 \). We can find a lower bound for \( s - t \) as follows.

\[
\left|\frac{q(t)}{t} - \frac{q(s)}{s}\right| = s^{-1}\left|\frac{s}{t}q(t) - q(s)\right|
\leq s^{-1}\left|\frac{s}{t}q(t) - q(t)\right| + s^{-1} |q(t) - q(s)|
= s^{-1} |q(t)| \left(\frac{s}{t} - 1\right) + s^{-1} |q(s) - q(t)|
= \frac{s - t}{s} \left(\frac{|q(t)|}{t} + \frac{|q(s) - q(t)|}{s - t}\right)
\leq \frac{s - t}{s} (c_1 + c_2)
\]

where \( c_1 \geq 0 \) is the Lipschitz constant for \( q(t) \) from (16), and \( c_2 = |\lambda| \geq 0 \) is a bound on the arrival rate. (Without loss of generality, assume \( c_1 + c_2 > 0 \).) On the
other hand, by the triangle inequality,

\[ \left| \frac{q(t)}{t} - \frac{q(s)}{s} \right| \geq \left| \frac{q(t)}{t} - \hat{q} \right| - \left| \frac{q(s)}{s} - \hat{q} \right| > \frac{\varepsilon}{2}. \]

Combining these two bounds and solving for \( s - t \), we find

\[ s - t > t \frac{\varepsilon/2}{c_1 + c_2}. \]

This gives us \( h > 0 \) as desired, and completes the proof. \( \square \)

11. A multihop network (microscopic, macroscopic & teleology). We now study a multihop network. The techniques in this section show nothing substantially new—the multihop case is mostly a simple reworking of the single hop case. This section may therefore be seen as a quick review of the train of argument of the preceding sections, for a slightly more involved setup. But there is one important technical difference: we will make further assumptions on the arrival process and on the set \( S \) of possible actions (Condition 11.2 below), so that it is always possible to serve the maximum weight schedule without idling. This condition is needed for our proofs, and it is assumed in the work of Tassiulas and Ephremides [32] and Harrison [12] and all the other work we are aware of on scheduling in multihop networks; but we believe this is limitation of the proof techniques and that the assumption is not in fact necessary.

The rest of this section is in three parts. In Section 11.1 we define the microscopic network model, explain the admissible traffic region, state the fluid model equations, define a Lyapunov function, and define a lifting map via two optimization problems. In Section 11.2 we state the main results: that the fluid model equations are appropriate, that the Lyapunov is indeed a Lyapunov function, that the fixed points of the fluid model correspond to solutions to the optimization problems, and that multiplicative state space collapse holds. In Section 11.3 we give sketch proofs.

11.1. Definitions.

Network model. Consider a collection of \( N \) queues as in Section 1.4, but with the difference that once work is served from a queue it may join another queue. Let \( R \in \{0, 1\}^{N \times N} \) be the routing matrix, \( R_{mn} = 1 \) if work served from queue \( m \) is sent to queue \( n \), and \( R_{mn} = 0 \) otherwise; if \( R_{mn} = 0 \) for all \( n \) then work served from queue \( m \) departs the system. For each \( m \) we require \( R_{mn} = 1 \) for at most one \( n \). We will assume that routing is acyclic, i.e. that work served from some queue \( n \) never returns to queue \( n \). Also, let \( \hat{R} = (I - R^T)^{-1} \); by considering the expansion \( \hat{R} = I + R^T + (R^T)^2 + \cdots \) it is clear that \( R_{mn} = 1 \) if work injected at queue \( n \) eventually passes through \( m \), and 0 otherwise. If \( q \) is a vector of queue sizes we will
write \( \mathbf{q} \) for \( \mathbf{Rq} \); \( q_n \) is the amount of work at queue \( n \), and \( \mathbf{q}_n \) is the amount of work for queue \( n \). (Tassiulas and Ephremides [32] described a network model with routing choice, whereas we have restricted ourselves to deterministic routing for the sake of simplicity.)

We will use the convention that \( \mathbf{Q}(\tau) \) is the vector of queue sizes at the beginning of timeslot \( \tau \), then the schedule for timeslot \( \tau \) is chosen and work departs queues according to the schedule, then external arrivals for timeslot \( \tau \) happen and so do arrivals due to internal routing. The queueing dynamics are changed from (1) to (44)

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

**Scheduling algorithm.** We will be interested in the multihop version of the MW-\( f \) scheduling algorithm, known as *backpressure*. This algorithm chooses a schedule \( \pi \) for timeslot \( \tau \) such that

\[
\pi \cdot (I - R)f(\mathbf{Q}(\tau)) = \max_{\rho \in \mathcal{S}} \rho \cdot (I - R)f(\mathbf{Q}(\tau)).
\]

Recall that matrix multiplication takes precedence over the \( \cdot \) operator, so the left hand side is \( \pi \cdot \{(I - R)f(\mathbf{q}(\tau))\} \). Note that

\[
[(I - R)f(\mathbf{q})]_n = f(Q_n) - f([R\mathbf{q}]_n)
\]

where \([R\mathbf{q}]_n\) is the queue size at the first queue downstream from \( n \) (or 0 if there is no queue downstream). The difference \([ (I - R)f(\mathbf{q})]_n \) is interpreted as the pressure to send work from queue \( n \) to the queue downstream of \( n \); if the downstream queue has more work in it than the upstream queue then there is no pressure to send work downstream.

We will assume that \( f \) satisfies a scale-invariance property, the multihop equivalent of Condition 4.1:

**Assumption 11.1** Assume \( f \) is differentiable and strictly increasing with \( f(0) = 0 \). Assume also that for any \( \mathbf{q} \in \mathbb{R}_+^N \) and \( \pi \in \mathcal{S} \), with \( m(\mathbf{q}) = \max_{\rho \in \mathcal{S}} \rho \cdot (I - R)f(\mathbf{q}) \),

\[
\pi \cdot (I - R)f(\mathbf{q}) = m(\mathbf{q}) \implies \pi \cdot (I - R)f(\kappa \mathbf{q}) = m(\kappa \mathbf{q}) \text{ for all } \kappa \in \mathbb{R}_+.
\]

We will make some further assumptions about the chosen schedule: we will require that it should not incur any idling, hence

\[
\mathbf{Z}(\tau) = 0 \quad \text{for all } \tau,
\]

and we will also need the chosen schedule to always have positive weight,

\[
\max_{\pi \in \mathcal{S}} \pi \cdot (I - R)f(\mathbf{Q}) > 0 \quad \text{for all } \mathbf{Q} \neq \mathbf{0}, \mathbf{Q} \in \mathbb{R}_+^N.
\]
In order to meet these two requirements, we will make some extra assumptions about the arrival process and the set $S$ of possible actions.

**Assumption 11.2 (Monotonicity)** Assume that

i. if $\pi \in S$ is an allowed service action, and $\rho \in \mathbb{R}^N_+$ is some other service action with $\rho_n \in \{0, \pi_n\}$ for all $n$, then $\rho \in S$

ii. $A_n(\tau + 1) - A_n(\tau) \in \{0, 1, 2, \ldots\}$ for all $n$ and $\tau$

iii. $\pi_n \in \{0, 1\}$ for all $n$ and $\pi \in S$

It is easy to see that these three assumptions are sufficient to permit (45). We say that the set $S$ is *monotone* if it satisfies the first of the three; this is sufficient to permit (46). Whenever we refer to a network running the MW-$f$ backpressure algorithm, we mean that Conditions 11.1 and 11.2 are to be assumed.

**Admissible region.** Let $S$ be the set of feasible schedules, and assume it is monotone. Let $\lambda$ be the vector of external arrival rates. The total arrival rate of work for each queue, taking account of routing, is $\bar{X} = R\lambda$, and so we define the admissible region to be

$$\Lambda = \{ \lambda \in \mathbb{R}^N_+ : \bar{X} \leq \sigma \text{ componentwise, for some } \sigma \in \Sigma \}$$

where $\Sigma$ is the convex hull of $S$ as in Section 3.1. We say $\lambda$ is admissible if and only if PRIMAL($\bar{X}$) $\leq 1$. When it is equal to 1, define the critically-loaded virtual resources to be the solutions to DUAL($\bar{X}$).

Given that $S$ is monotone, it is not hard to show that

$$\text{if } \sigma \in \Sigma \text{ and } 0 \leq \sigma' \leq \sigma \text{ then } \sigma' \in \Sigma.$$  

This is how we will most commonly use monotonicity in the sketch proofs in Section 11.3. Applying it to the definition of the admissible region, we see

$$\Lambda = \{ \lambda \in \mathbb{R}^N_+ : \bar{X} = \sigma \text{ for some } \sigma \in \Sigma \}.$$

**Fluid model equations.** The fluid model equations are mostly unchanged from (12)–(18), but for the sake of clarity we will write them all out again. The algorithm-independent equations are

(48) \[ a(t) = \lambda t \]

(49) \[ q(t) = q(0) + a(t) - (I - R^T) \sum_\pi s_\pi(t) \pi + z(t) \]

(50) \[ \sum_{\pi \in S} s_\pi(t) = t \]

(51) each $s_\pi(\cdot)$ and $z_n(\cdot)$ is increasing (not necessarily strictly increasing)

(52) all the components of $x(\cdot)$ are Lipschitz

(53) for almost all $t$, all $n$, $\dot{z}_n(t) = 0$ if $q_n(t) > 0$
and the additional fluid equations for the non-idling backpressure algorithm are

\begin{equation}
\dot{s}_\pi(t) = 0 \text{ if } \pi \cdot (I - R)f(q(t)) < \max_{\rho \in S} \rho \cdot (I - R)f(q(t)) \tag{54}
\end{equation}

\begin{equation}
\text{for almost all } t, \text{ all } \pi \in S, \tag{55}
\end{equation}

\begin{equation*}
z(t) = 0
\end{equation*}

**Lyapunov function.** We will use exactly the same Lyapunov function as before, \(L(q) = 1 \cdot F(q)\).

**Lifting map.** We redefine the optimization problems ALGD and ALGP as follows. Let \(S^*(\bar{X})\) be the set of principal critically loaded virtual resources, as defined in Section 3. The optimization problem ALGD(\(q\)) for \(q \in \mathbb{R}^N_+\) is

\[
\begin{align*}
\text{minimize} & \quad L(r) \\
\text{over} & \quad r \in \mathbb{R}^N_+ \\
\text{such that} & \quad \xi \cdot \bar{r} \geq \xi \cdot \bar{q} \quad \text{for all } \xi \in S^*(\bar{X}) \\
& \quad \text{and } \bar{r}_n \leq \bar{q}_n \quad \text{for all } n \text{ such that } \bar{x}_n = 0.
\end{align*}
\]

Remembering that \(\bar{r}_n = [\bar{R}r]_n\) is the amount of work that will have to pass through queue \(n\), this constraint may be thought of as saying “regardless of where in the upstream path the work is held, it’s still got to get through the bottleneck”. The optimization problem ALGP(\(q\)) is like ALGD but with the first line of the constraint replaced by

\[
\bar{r} \geq \bar{q} + t(\bar{X} - \sigma) \quad \text{for some } t \geq 0, \sigma \in \Sigma.
\]

The Lyapunov function \(L\) is strictly convex as before, and the feasible set for ALGD is non-empty and convex, so the first problem has a unique solution. We define the lifting map \(\Delta W\) to map \(q\) to the unique solution of ALGD(\(q\)). It turns out that both problems have the same feasible set. When \(S\) is monotone, Lemma 11.6 below shows that the common solution satisfies

\[
\bar{r} = \bar{q} + t(\bar{X} - \sigma) \quad \text{for some } t \geq 0, \sigma \in \Sigma.
\]

Note that the map \(\Delta W\) is a function \(\mathbb{R}^N_+ \to \mathbb{R}^N_+\), whereas in the single-hop case we defined two separate functions, \(W\) which maps a queue size vector to a workload vector, and \(\Delta\) which maps a workload vector to a queue size vector. The difference here is that the problem ALGD has an extra constraint relating to queue sizes at queues with no arrivals, and to express this constraint we need to know not just \(w\) but also \(q\). We speculate that one might be able to find a more appropriate Lyapunov
function, under which the extra constraint would be redundant. Alternatively, we might reasonably restrict our attention throughout this section to the space of queue size vectors $q \in \mathbb{R}_+^N$ for which $q_n = 0$ wherever $\lambda_n = 0$; this restriction allows us to write $\Delta W$ as the composition of two separate maps $\Delta$ and $W$.

11.2. Results.

**Theorem 11.3 (Fluid model equations)*** Let $x^r(t)$ be a sequence of fluid-scale queueing systems, as in Section 5.1, and make the same assumptions (21)–(24) about the arrival process as for Theorem 5.1. Let FMS be the set of all processes $x(t)$ over $t \in [0, T]$ which satisfy the appropriate fluid model equations, namely

- equations (48)–(53), for any scheduling algorithm,
- equations (54) and (55) in addition if the network is running the MW-f backpressure algorithm described in Section 11.1,
- $q(0) = q_0$ in addition, if (25) holds.

Let $\text{FMS}_\varepsilon$ be the $\varepsilon$-fattening of FMS as defined in Theorem 5.1. Then for any $\varepsilon > 0$, $\mathbb{P}(x^r(\cdot) \in \text{FMS}_\varepsilon) = 1 - o(R(r'))$ as $r \to \infty$.

**Lemma 11.4 (Lyapunov stability)** Suppose the network is running the MW-f backpressure algorithm. Then every fluid model solution satisfies

$$\dot{L}(q(t)) = \lambda \cdot f(q(t)) - \max_{\pi \in \Pi} \pi \cdot (I - R)f(q(t)).$$

Furthermore, if $\lambda \in \Lambda$ then

$$\lambda \cdot f(q) - \max_{\pi \in \Pi} \pi \cdot (I - R)f(q) \leq 0 \quad \text{for all } q \in \mathbb{R}_+^N,$$

hence $\dot{L}(q(t)) \leq 0$. Furthermore, if $\lambda \in \Lambda^\circ$ then either $\dot{L}(q(t)) < 0$ or $q(t) = 0$.

**Lemma 11.5 (Feasibility)** Suppose that $\lambda \in \Lambda$. Then the optimization problems $\text{ALGD}(q)$ and $\text{ALGP}(q)$ have the same feasible set, for any $q \in \mathbb{R}_+^N$. Furthermore, for any scheduling algorithm, every fluid model solution with initial queue size $q(0)$ has the property that $q(t)$ is feasible for $\text{ALGP}(q(0))$ for all $t \geq 0$.

**Lemma 11.6 (Monotonicity)** Suppose that $\lambda \in \Lambda$. For any $q \in \mathbb{R}_+^N$, $\Delta W(q)$ can be written

$$\Delta W(q) = q + t(\lambda - (I - R^T)\sigma) \quad \text{for some } t \geq 0, \sigma \in \Sigma.$$

**Lemma 11.7 (Fixed points)** Suppose $\lambda \in \Lambda$ and the network is running the MW-f backpressure algorithm. Then $q = \Delta W(q) \iff q$ is a fixed point of the fluid model.
Lemma 11.8 (Properties of the lifting map) Suppose $\lambda \in \Lambda$ and the network is running the MW-f backpressure algorithm.

i. $\Delta W : \mathbb{R}_+^N \to \mathbb{R}_+^N$ is a continuous function;

ii. if $q = \Delta W(q)$ then $\Delta W(\kappa q) = \kappa \Delta W(q)$ for all $\kappa > 0$;

iii. for any $\varepsilon > 0$ there exists some $H_\varepsilon > 0$ such that, for every fluid model solution with bounded initial value, $|q(t) - \Delta W(q(t))| < \varepsilon$ for all $t \geq H_\varepsilon$.

Theorem 11.9 (Multiplicative state space collapse) Theorem 8.2 holds for the multihop network running the MW-f backpressure algorithm.

11.3. Proofs. Here are the proofs. Mostly they are the same as for the single-hop case, and we leave the details to the appendix. But where the proofs involve multihop or monotonicity in an interesting way, we will spell them out in this section.

Proof of Theorem 11.3. This is much the same as the proof of Theorem 5.1; details in the appendix.

Proof of Lemma 11.4.

\[
\frac{d}{dt}L(q(t)) = \dot{q}(t) \cdot f(q(t))
\]

\[
= \left(\lambda - (I - R^T) \sum_\pi s_\pi(t) \pi + \dot{z}(t)\right) \cdot f(q(t)) \quad \text{by differentiating (49)}
\]

\[
= \left(\lambda - (I - R^T) \sum_\pi s_\pi(t) \pi\right) \cdot f(q(t)) \quad \text{by (55), no idling}
\]

\[
= \lambda \cdot f(q(t)) - \sum_\pi s_\pi(t) \pi \cdot (I - R)f(q(t)) \quad \text{by rearranging terms}
\]

\[
= \lambda \cdot f(q(t)) - \max_\rho \rho \cdot (I - R)f(q(t)) \quad \text{by (54)}
\]

\[
= \lambda \cdot f(q(t)) - \max_\rho \rho \cdot (I - R)f(q(t)) \quad \text{by (50)}.
\]

When $\lambda \in \Lambda$ we can by definition write $\bar{R}\lambda \leq \sigma$ for some $\sigma \in \Sigma$. Because $S$ is monotone we can by (47) strengthen this to $\bar{R}\lambda = \sigma$ for some $\sigma \in \Sigma$, hence $\lambda = (I - R^T)\sigma = \sum_\pi \alpha_\pi (I - R^T)\pi$ with $\alpha_\pi \geq 0$ and $\sum \alpha_\pi = 1$. Hence, for any $q \in \mathbb{R}_+^N$,

\[
\lambda \cdot f(q) - \max_\rho \rho \cdot (I - R)f(q) = \sum_\pi \alpha_\pi \pi \cdot (I - R)f(q) - \max_\rho \rho \cdot (I - R)f(q)
\]

\[
\leq \left(\sum \alpha_\pi - 1\right) \max_\rho \rho \cdot (I - R)f(q).
\]

When $\lambda \in \Lambda^\circ$ the same holds but with $\sum \alpha_\pi < 1$. In both cases, either $q = 0$ or $\max_\pi \pi \cdot (I - R)f(q) > 0$ by the assumption (46) that $S$ is monotone. Hence the result follows. \qed
Proof of Lemma 11.5. Applying Lemma 6.1 to $\vec{q} = \vec{R}q$, $\vec{r} = \vec{R}r$ and $\vec{\lambda} = \vec{R}\lambda$, the feasible sets of ALGD(q) and ALGP(q) are the same. The proof of the second claim is much like that of Lemma 7.1; details in the appendix.

Proof of Lemma 11.6. We will choose $\sigma$ simply by multiplying each side of the desired equation by $\vec{R}$:

$$\vec{r} = \vec{q} + t(\vec{\lambda} - \sigma) \text{ where } r = \Delta W(q)$$

or, rearranging,

$$\sigma = \vec{\lambda} - (\vec{r} - \vec{q})/t.$$

We will show that $0 \leq \sigma \leq \rho$ for some $\rho \in \Sigma$, hence by the monotonicity equation (47) $\sigma \in \Sigma$.

First, we show $0 \leq \sigma$. If $\vec{\lambda}_n > 0$ this can be achieved by choosing $t$ sufficiently large. If $\vec{\lambda}_n = 0$ then by the second constraint of ALGD(q) we know that $\vec{r}_n \leq \vec{q}_n$ so $\sigma_n \geq 0$.

Second, we show $\sigma \cdot \xi \leq 1$ for all virtual resources $\xi$. Either $\xi \cdot \vec{\lambda} = 1$, in which case $\vec{r} \cdot \xi \geq \vec{q} \cdot \xi$. Or $\xi \cdot \vec{\lambda} < 1$, in which case we simply need to choose $t$ sufficiently large. Either way, $\sigma \cdot \xi \leq 1$ for all virtual resources $\xi$, hence DUAL(\sigma) \leq 1, hence PRIMAL(\sigma) \leq 1, hence $\sigma \leq \rho$ for some $\rho \in \Sigma$ by the definition of PRIMAL.

Proof of Lemma 11.7. The corresponding result for a single-hop network, Theorem 7.3, expands the result into five equivalent statements, and the same is helpful here. We claim the following are equivalent:

   i. $q = \Delta W(q)$
   ii. $q$ is a fixed point
   iii. there exists a fluid model solution with $q(t) = q$ for all $t$
   iv. $\lambda \cdot f(q) = \max_{\pi \in S} \pi \cdot (I - R)f(q)$
   v. either $q = 0$, or $\zeta(q) = (I - R)f(q)/\max_{\pi \in S} \pi \cdot (I - R)f(q)$ is a critically-loaded virtual resource.

The only substantive difference between this and the proof of Theorem 7.3 is in going from (v) to (i)—here we have to use the representation of $\Delta W(q)$ from Lemma 11.6, whereas Theorem 7.3 used a different representation. Full details are in the appendix.

Proof of Lemma 11.8. The proofs are much the same as for a single-hop network.

   (i) is like Lemma 6.3 $\rightarrow$ Lemma A.3
   (ii) is like Lemma 6.4 $\rightarrow$ Lemma A.4
   (iii) is like Lemma 7.4 $\rightarrow$ Lemma A.5.

The proof of (i) needs some trivial modifications to take account of the network
structure. The proof of (ii) is based on the representation of the lifting map for a monotonic scheduling algorithm given in Lemma 11.6 rather than that used in Lemma 6.4. The proof of (iii) uses the multihop Lyapunov result from Lemma 11.4 rather than the single-hop result used in Lemma 7.4. There is no novelty in these proofs; they are detailed in the appendix.

12. An optimal algorithm? (macroscopic & teleology). A motivation for our work was Conjecture 2.1, which claims that the delay performance of an input-queued switch running MW-\(\alpha\) will improve as \(\alpha \to 0\). In this section, we give three different pieces of circumstantial evidence for the conjecture.

The first piece of evidence is based on critical fluid models, and is a proof that the total workload for MW-\(\alpha\) is no more than \(N^\alpha/(1+\alpha)\) times that for any other algorithm. Thus as \(\alpha \to 0\), MW-\(\alpha\) approaches optimal. The proof works for any single-hop switched network of the general sort described in this paper, under a ‘complete loading’ condition on the arrival rate vector \(\lambda\).

The second piece of evidence is based on state space collapse, and is a proof that as \(\alpha \to 0\) the collapsed state space \(W\) converges to the maximal possible collapsed state space for any other algorithm, and furthermore that MW-\(\alpha\) does not permit any idling in the interior of \(W\). Thus as \(\alpha \to 0\), the performance of MW-\(\alpha\) should approach optimal. This proof only applies to input-queued switches, and assumes that \(\lambda > 0\) componentwise, and assumes a stricter ‘complete loading’ condition.

The third piece of evidence is based on fluid models for overloaded switched networks, and is a proof that the long-run difference between net arrivals and net departures for MW-\(\alpha\) is no more than \(N^\alpha/(1-\alpha)\) times that for any other algorithm. Thus as \(\alpha \to 0\), MW-\(\alpha\) approaches the optimal net departure rate. The proof works for any overloaded single-hop network of the general sort described in this paper.

Given these three pieces of evidence, it is tempting to speculate about a formal limit of MW-\(\alpha\) as \(\alpha \to 0\). Since MW-\(\alpha\) chooses a schedule \(\pi\) to maximize \(\pi \cdot q^\alpha\), and since

\[
\alpha \approx \begin{cases} 
1 + \alpha \log x & \text{if } x > 0 \\
0 & \text{if } x = 0 
\end{cases}
\]

we make the following conjecture:

**Conjecture 12.1** Define the MW-0 scheduling algorithm as follows: at each timeslot it looks at all maximum-size schedules (i.e. those \(\pi \in S\) for which \(\sum_n \pi_n 1_{\eta_n > 0}\) is maximal), and among these picks one which has maximal log-weight (i.e. for which \(\sum_{n:q_n > 0} \pi_n \log q_n\) is maximal), breaking ties arbitrarily. We conjecture that this algorithm is stable for \(\lambda \in \Lambda^\circ\), and that it minimizes the total amount of idling in both the fluid limit and the critical-load limit for \(\lambda \in \partial\Lambda\), for any single-hop switched network.
Since our first two pieces of circumstantial evidence require certain assumptions about $\lambda$ and $S$, it may be that this conjecture only holds for those cases.\footnote{Both Stolyar and Srikant have recently found single-hop switched networks that show our conjecture to be false (private communication).}

Incidentally, McKeown et al. [23] showed that maximum-size matching (without use of weights to break ties) is not stable for certain $\lambda \in \Lambda^\circ$, for an input-queued switch.

12.1. Critical fluid model of MW-$\alpha$. Consider a single-hop switched network of the general sort described in this paper.

**Theorem 12.2** Let $\lambda \in \Lambda$. For any fluid model solution for the MW-$\alpha$ algorithm, $1 \cdot q(t) \leq N^{\alpha/(1+\alpha)} 1 \cdot q(0)$. If there is a critically-loaded virtual resource which assigns equal weight to each queue (i.e. if the virtual resource $1/\max_\pi 1\pi$ is critically loaded) then $1 \cdot q(t) \geq 1 \cdot q(0)$ for every scheduling algorithm.

**Proof.** The claim about MW-$\alpha$ uses the standard result that for any $x \in \mathbb{R}_+^N$ and $\beta > 1$

$$\frac{1}{N^{1-1/\beta}} \sum_n x_n \leq (\sum_n x_n^\beta)^{1/\beta} \leq \sum_n x_n. \tag{56}$$

Then using $L(\cdot)$ as in Lemma 4.2,

$$1 \cdot q(t) \leq N^{1-1/(1+\alpha)} (\sum_n q_n(t)^{(1+\alpha)})^{1/(1+\alpha)} \quad \text{by first standard inequality}$$

$$= N^{\alpha/(1+\alpha)} L(q(t))^{1/(1+\alpha)} \quad \text{by definition of } L(\cdot)$$

$$\leq N^{\alpha/(1+\alpha)} L(q(0))^{1/(1+\alpha)} \quad \text{since } \dot{L}(q(t)) \leq 0 \text{ by Lemma 4.2}$$

$$\leq N^{\alpha/(1+\alpha)} 1 \cdot q(0) \quad \text{by second standard inequality.}$$

The claim about every scheduling algorithm is a simple consequence of Lemma 7.1. $\square$

In an input-queued switch, the requirement that there be a critically-loaded virtual resource which assigns equal weight to all queues is equivalent to the requirement that either there is some set of critically loaded input ports and $\lambda_{i,j} = 0$ for all input ports $i$ which are not critical; or that there is some set of critically loaded output ports and $\lambda_{i,j} = 0$ for all output ports $j$ which are not critical.

12.2. State space collapse of MW-$\alpha$ for an input-queued switch. Consider an input-queued switch running MW-$\alpha$, and write $\Delta^\alpha$ for its lifting map. Suppose the arrival rate matrix is $\lambda \in \partial \Lambda$, and that $\lambda > 0$ componentwise, and that every input
port and every output port is critically loaded, i.e.
\[ \sum_j \lambda_{ij} = 1 \quad \text{and} \quad \sum_i \lambda_{ij} = 1 \quad \text{for every } i, j. \]

As discussed in Section 3.5, the critical workloads are
\[ W(q) = (w_1, \ldots, w_M; w_1, \ldots, w_M) = (r_1 \cdot q, \ldots, r_M \cdot q; c_1 \cdot q, \ldots, c_M \cdot q). \]

There is a straightforward bijection between \( W(q) \) and \( \hat{W}(q) = (w_2, \ldots, w_M; w_2, \ldots, w_M; w_1, \ldots, w_M). \)

Write \( \hat{W}(\alpha) \) for the image of \( \hat{W}(q) \) over \( q \in \mathcal{I} \). Observe that \( \hat{W}(q) \) is constrained to lie in the set
\[ \hat{W}^{\text{max}} = \{ \hat{w} \in \mathbb{R}^{2M-1}_+ : w_2 + \cdots + w_M \leq w_1 \text{ and } w_2 + \cdots + w_M \leq w_M \}. \]

For the 2 \( \times \) 2 case, we proved that \( \hat{W}(\alpha) \) is increasing as \( \alpha \to 0 \), and also that all the queues of \( \Delta^{\alpha}(\hat{w}) \) are nonempty (which means that the switch cannot idle) when \( \hat{w} \) is in the interior of \( \hat{W} \). We have so far been unable to prove this for the \( M \times M \) case, but we can prove

**Theorem 12.3** Make the above assumptions about the arrival rate vector \( \lambda \). Then for any \( \alpha > 0 \), all the queues of \( \Delta^{\alpha}(\hat{w}) \) are nonempty for \( \hat{w} \) in the interior of \( \hat{W}(\alpha) \). Also, for any \( \hat{w} \) in the interior of \( \hat{W}^{\text{max}} \), \( \hat{w} \in \hat{W}(\alpha) \) for sufficiently small \( \alpha \).

The proof makes use of the two following lemmas. The first is a general closure property of permutation matrices, and the second is a property of MW-\( f \) matchings in input-queued switches.

**Lemma 12.4** Let \( x \in \mathbb{R}^{M \times M}_+ \), and let \( A_{i,j} = 1 \) if there is some matching \( \rho \in \mathcal{S} \) whose weight \( \rho \cdot x \) is maximal and for which \( \rho_{i,j} = 1 \); and \( A_{i,j} = 0 \) otherwise. Then, for any matching \( \pi \) with \( \pi_{i,j} = 1 \implies A_{i,j} = 1 \), \( \pi \) is itself a maximum weight matching.

**Proof of Lemma 12.4.** Let \( a \) be the weight of the maximum weight matching, let \( \mathcal{A} = \{ \rho : \rho \cdot x = a \} \), and let \( \sigma = \sum_{\rho \in \mathcal{A}} \rho \). The matrix \( \sigma - \pi \) is non-negative because \( \sigma \geq A \geq \pi \); also all its row sums and column sums are equal to \( |\mathcal{A}| - 1 \); therefore by the Birkhoff–von-Neumann decomposition
\[ \sigma = \pi + \sum_{\rho \in \mathcal{A}} \alpha_{\rho} \rho \quad (57) \]

where each \( \alpha_{\rho} \geq 0 \) and \( \sum_{\rho} \alpha_{\rho} = |\mathcal{A}| - 1 \). Now, \( \sigma \cdot x = |\mathcal{A}| m \) by construction of \( \sigma \), and since \( a \) is the maximum weight it must be that all the matchings in (57) have maximum weight. \( \square \)
Lemma 12.5  For any $\lambda \in \partial \Lambda$ and any weight function $f$ satisfying Condition 4.1, if $q \in R_{+}^{M \times M}$ is a fixed point then all queues with $\lambda_{i,j} > 0$ are involved in some maximum weight matching.

Proof of Lemma 12.5. Pick some $\sigma = \sum \alpha_{\pi} \pi \in \Sigma$ such that $\lambda \leq \sigma$. By Theorem 7.3, $\lambda \cdot f(q) = \max_{\pi} \pi \cdot f(q)$, and this must also be the value of $\sigma \cdot f(q)$, i.e. all the matchings $\pi$ with $\alpha_{\pi} > 0$ are maximum weight. If $\lambda_{i,j} > 0$ then at least one of these matchings has $\pi_{i,j} = 1$.

Proof of first claim of Theorem 12.3. Suppose that $\hat{w}$ is in the interior of $W(\alpha)$, and let $q = \Delta^{\alpha}(\hat{w})$. Since $\hat{w}$ is in the interior of $W(\alpha)$, it is in the interior of $W^{\max}$ and so every workload of $W(q)$ is strictly positive.

Suppose that some $q_{i,j} = 0$. Then there must also be a $q_{i',j'} > 0$ since if $q_{i',j} = 0$ for all $i'$ then $w_{j} = 0$, which by assumption is not the case. Similarly there must be some $q_{i,j'} > 0$. Indeed we can bound these away from zero: $q_{i',j} > w_{j}/M$ and similarly for $q_{i,j'}$.

Since $q = \Delta^{\alpha}(\hat{w})$ and $\hat{w} \in W(\alpha)$, it is the case that $q = \Delta^{\alpha}W(\alpha)$. Therefore by Theorem 7.3 $q$ is a fixed point. By assumption $\lambda > 0$ componentwise, so by Lemma 12.5 every queue is involved in some maximum weight matching. Therefore by Lemma 12.4 every matching is a maximum weight matching.

Now consider the submatrix

\[
\begin{pmatrix}
q_{i,j} & q_{i,j'} \\
q_{i',j} & q_{i',j'}
\end{pmatrix}
\]

Let $\pi$ be any matching which includes the leading diagonal of this submatrix (i.e. with $\pi_{i,j} = \pi_{i',j'} = 1$), and let $\rho$ be like $\pi$ but including instead the off diagonal of the submatrix (i.e. $\pi$ and $\rho$ differ simply by a transposition). By our earlier remark, each of these two matchings has the same weight. From our direct calculations for the $2 \times 2$ switch with MW-$\alpha$, we know that the configuration of these four queues which makes both $\pi$ and $\rho$ have the same weight has each of these four queues non-empty. This contradicts our assumption that $q_{i,j} = 0$.

We conclude that all queues are non-empty.

Proof of second claim of Theorem 12.3. First note some properties of $\Delta^{f}(\cdot)$, for any MW-$f$ algorithm. Suppose that $w \not\in W$, and let $q = \Delta^{f}(w)$. If it were the case that $W(q) = w$ then $q \in I$ and so $w \in W$; by assumption this is not the case so $W(q) \neq w$. Indeed by the definition of the lifting map, $W(q) \geq w$. Therefore there is some $i$ or $j$ such that $r_{i}q > w_{i}$ or $c_{j}q > w_{j}$. Indeed, there must be both such an $i$ and a $j$, since otherwise the sum of row workloads and column workloads would not be equal. Therefore $q_{i,j} = 0$, since if $q_{i,j} > 0$ we could reduce $q_{i,j}$ and still have a feasible solution to the optimization problem but with smaller $L(q)$.
Now we can proceed to prove the claim. Let $\hat{w}$ be in the interior of $\hat{W}^{\max}$. Suppose the claim is not true, i.e. suppose there exists a sequence $\alpha \to 0$ with $w \not\in W(\alpha)$. Let $q(\alpha) = \Delta^\alpha(w)$. For each $\alpha$ along the sequence, we can find indices $i(\alpha), j(\alpha)$ etc. as in the proof of the first claim. Some set of indices $(i, j, i', j')$ must be repeated infinitely often (since there are only finitely many choices). Consider the subsequence of $\alpha$ for which $i(\alpha) = i$ etc. The subsequence $q(\alpha)$ is bounded (by the same construction as in the proof of Lemma 6.3), and so it has a convergent subsequence. Let $q^*$ be the limit of the convergent subsequence. By our choice of subsequence, $q^*_{i,j} = 0$ and $q^*_{i',j'} \geq w_i/M$ and $q^*_{i,j} \geq w_j/M$.

Note that $q$ is optimal for ALGD($\hat{w}$), therefore it must also be optimal for ALGD($W(q)$) which has a smaller feasible region. Therefore $q = \Delta W(q)$, so $q$ is a fixed point by Theorem 7.3. Consider the same submatrix as in the first part of the proof, and the same two matchings $\pi$ and $\rho$; since $q$ is a fixed point these two matchings must have the same weight.

We can write out explicitly the difference in weight between the two matchings:

$$
\rho \cdot q(\alpha) - \pi \cdot q(\alpha) = q^{\alpha}_{i',j'} + q^{\alpha}_{i,j'} - q^{\alpha}_{i',j'} - q^{\alpha}_{i,j}.
$$

Recall that along the subsequence we have chosen, $q^{\alpha}_{i',j'} \to q^{*}_{i',j'}$ etc. Since these limits exist, (59) must lie in an open neighbourhood of $[1, 2]$ for sufficiently small $\alpha$. This contradicts the finding that all matchings must have the same weight. Therefore we have contradicted our original assumption, that there are arbitrarily small $\alpha > 0$ with $w \not\in W(\alpha)$.

12.3. Overloaded fluid model of MW-$\alpha$. Consider a single-hop switched network of the general sort described in this paper, and suppose it is overloaded.

**Theorem 12.6** Let $\lambda \notin \Lambda$. There is some $q^{\min} = q^{\min}(\lambda) > 0$ such that for any scheduling algorithm, all fluid model solutions satisfy

$$
q^{\min} \leq 1 \cdot q(t)/t \quad \text{for all } t \geq 0.
$$

Furthermore, for the MW-$\alpha$ scheduling algorithm,

$$
\lim_{t \to \infty} 1 \cdot q(t)/t \leq N^{\alpha/(1+\alpha)}q^{\min}.
$$

Note that the departures up to time $t$ are defined as $d(t) = q(0) + a(t) - q(t)$, and $a(t) = \lambda t$ by (12), thus the net departure rate over the interval $[0, t]$ is

$$
1 \cdot d(t)/t = 1 \cdot q(0)/t + \lambda - 1 \cdot q(t)/t.
$$

Hence in order to maximize the net departure rate we should minimize $1 \cdot q(t)/t$. The theorem tells us that MW-$\alpha$ has near-optimal long-run departure rate, for $\alpha \approx 0$. 

imsart-aap ver. 2007/04/13 file: netsched.tex date: April 27, 2009
Proof. We needn’t be coy; we can identify $q_{\text{min}}$ as the solution to

$$\text{minimize } 1 \cdot r$$
$$\text{over } r \in \mathbb{R}_+^N$$
$$\text{such that } \xi \cdot r \geq \xi \cdot \lambda - 1 \text{ for all } \xi \in S^\dag(\lambda)$$

where $S^\dag$ is the set of principal overloaded virtual resources defined in Section 10.2.

Proof of first claim. From the fluid model equation (13),

$$q(t)/t = q(0)/t + a(t)/t - \sum_{\pi} s_{\pi}(t) \pi / t + z(t)/t$$
$$\geq \lambda - \sum_{\pi} s_{\pi}(t) \pi / t.$$ 

Therefore, for any $\xi \in S^\dag$,

$$\xi \cdot q(t)/t \geq \xi \cdot \lambda - t^{-1} \sum_{\pi} s_{\pi}(t) \xi \cdot \pi$$
$$\geq \xi \cdot \lambda - 1 \text{ since } \xi \cdot \pi \leq 1 \text{ for any virtual resource } \xi.$$

We have shown that $q(t)/t$ is feasible for the optimization problem that defines $q_{\text{min}}$, hence $1 \cdot q(t)/t \geq q_{\text{min}}$ as required.

Proof of second claim. Theorem 10.2 says that for MW-$\alpha$, every fluid model solution has the property that $q(t)/t$ converges as $t \to \infty$, and the limit is the (unique) solution $\hat{q}$ to the optimization problem ALGD$^\dag$, namely to

$$\text{minimize } 1 \cdot r^{1+\alpha}$$
$$\text{over } r \in \mathbb{R}_+^N$$
$$\text{such that } \xi \cdot r \geq \xi \cdot \lambda - 1 \text{ for all } \xi \in S^\dag(\lambda).$$

We will again need the standard inequality (56). It gives us

$$1 \cdot \hat{q} \leq N^{\alpha/(1+\alpha)}(1 \cdot \hat{q}^{1+\alpha})^{1/(1+\alpha)} \text{ using LHS of (56) with } \beta = 1 + \alpha$$
$$\leq N^{\alpha/(1+\alpha)}(1 \cdot r^{1+\alpha})^{1/(1+\alpha)} \text{ for any } r \text{ feasible for ALGD}^\dag$$
$$\leq N^{\alpha/(1+\alpha)}1 \cdot r \text{ using RHS of (56) with } \beta = 1 + \alpha.$$ 

Now, taking the minimum over all $r$ feasible for ALGD$^\dag$,

$$1 \cdot \hat{q} \leq N^{\alpha/(1+\alpha)} q_{\text{min}}$$

as required. \qed
13. Conclusion. We have considered a class of maximum-weight scheduling algorithms for a general model of a switched network. This model includes the popular input-queued switch, as well as a wireless model described below. We produced a fluid model, and analysed its performance when the network is lightly loaded, critically loaded and overloaded. Under all three loadings, we identified the space of invariant states and convergence to this space. Under light load, these results recover known stability results; for critical load, we obtain multiplicative state space collapse; and for overload, we obtain a new limit result.

The noteworthy feature of these networks is that invariant states are characterized through an optimization problem, in which constraints are imposed by loading and resources. In the context of bandwidth sharing in the Internet, similar results have been observed by Kelly and Williams [17] for critical load, and by Egorova et al. [8] for overload.

We used these techniques to analyse the performance of the MW-$\alpha$ algorithm for a general switched network. We analysed the critical fluid model, and found that it gets arbitrarily close to optimal as $\alpha \to 0$, under a condition on the traffic pattern. For the case of the input-queued switch, we analysed the geometry of the set of fixed points, and found that as $\alpha \to 0$ this geometry indicates there is less and less idling. These results provide support for the experimental conjecture of Keslassy and McKeown, Conjecture 2.1. Furthermore, our analysis of the overloaded switched network shows that as $\alpha \to 0$ the net throughput gets arbitrarily close to optimal. (We also observed that for the plain MW algorithm, net throughput might actually decrease as load increases.) We conjectured an optimal algorithm, the formal limit of MW-$\alpha$ as $\alpha \to 0$.

13.1. Questions for future work.

Question one. We conjectured an optimal algorithm, but we were not able to obtain a fluid model for it, therefore we were not able to prove anything about its state space collapse. The fluid models we used cannot distinguish between queues which are very-nearly-empty and queues which are actually empty, whereas our conjectured algorithm does distinguish. Overcoming this limitation is likely to advance our understanding of how small one can keep a queue without it idling.

Question two. We proved multiplicative state space collapse. We have not proved full state space collapse, nor any sort of diffusion limit, as in Williams [34]. Such results are being worked on by Kang and Williams [16]. It may also be possible to obtain moderate deviations limits for the probability of large fluctuations in queue size.

Question three. We have described the geometry of the invariant state space, and the reflection angles at the boundaries, but we have not been able to analyse the...
two together. A full state space collapse result, which characterizes the invariant distribution of the diffusion limit, may provide the answer, but it is likely to be very challenging to obtain any sort of explicit answer. If we could, it is likely to lead to a better understanding of the performance of MW-α.

Question four. Algorithms based on MW are computationally difficult to implement, because there are so many comparisons to be made. Can the analysis techniques described in this paper help us to design simpler algorithms which do not sacrifice too much performance? Also, there may be policy reasons for giving higher priority to certain flows; how do such design decisions impact on overall performance?

13.2. Applications where switched network models might be useful.

Wireless networks. Consider several wifi networks close to each other and sharing the same frequency. If two devices close to each other transmit at the same time, then there is interference and the data may be lost; whereas two devices far from each other may successfully transmit at the same time. A popular way to model this sort of interference is to draw a graph with a node for each device and an edge between two nodes if they can interfere with each other; in other words a transmission from a node is successful only if none of its neighbours in the network graph is transmitting at the same time. (This is called the independent set model for interference.) The model of Stolyar [31] allows for more detailed specification of the degree of interference between transmissions, as well as a random component to the set of schedules which might reflect atmospheric conditions.

Job-level model for congestion control. Roberts and Massoulie [26] proposed a model for bandwidth sharing between flows in an Internet-like network, running a particular class of congestion control algorithm. Consider a network consisting of J links, 1,...,J, where link j has capacity Cj. A route r can be considered to be a subset of {1,...,J}, specifying which links are used in transmitting data along that route. Let Qr be the number of active flows transmitting data on route r. We can think of a congestion control scheme like TCP as a function which maps the vector of Qr to an allocation of rates for each of the active flows. Given an allocation of rates, and assuming exponential flow sizes, we can compute the rate at which Qr decreases. In other words, the congestion-control scheme is an algorithm for specifying a schedule.

Kelly and Williams [17] and Egorova et al. [8] have modelled this as a switched network, and obtained results that are closely related to ours. Additionally, Kang et al. [15] have proved a full diffusion limit. There are of course differences in the model: we have worked with a slotted-time model whereas the bandwidth-sharing model uses continuous time; we assumed a finite set of possible schedules, whereas the bandwidth-sharing model considers a particular scheduling algorithm which chooses
a schedule from a space which is uncountable (but simply parameterized); we assumed that service is non-random, whereas the bandwidth-sharing model takes it to be random.

Road roundabouts. Figure 4 shows a four-way road junction with a roundabout. There are twelve traffic flows in total, one from each road to each other road (assuming no one is using the roundabout to do a U-turn). According to UK traffic regulations, you may only enter the roundabout if there is no traffic coming from your right; this limits which of the twelve possible traffic flows may move simultaneously. Figure 4 shows two possibilities. This might perhaps be modelled as a switched network, though a proper ‘microscopic’ model should probably not use a slotted-time model, but should instead take account of the time it takes a car to cross the roundabout.

Acknowledgements. We would like to thank Mike Harrison, Frank Kelly, Balaji Prabhakar and Ruth Williams for helpful discussions. We thank Ciamac Moallemi for suggesting the Lemma 6.1.

References.


APPENDIX: MULTIHOP PROOFS

Restated theorem (11.3 Fluid model equations) Let \( x(t) \) be as in Theorem 5.1, and make the same assumptions (21)–(24). Let FMS be the set of all processes \( x(t) \) over \( t \in [0,T] \) which satisfy the appropriate fluid model equations, namely

- equations (48)–(53), for any scheduling algorithm,
- equations (54) and (55) in addition if the network is running the MW-f backpressure algorithm described in Section 11.1,
- \( q(0) = q_0 \) in addition, if (25) holds.

Let \( FMS_\epsilon \) be the \( \epsilon \)-fattening of FMS as defined in Theorem 5.1. Then for any \( \epsilon > 0 \), \( \mathbb{P}(x^r(\cdot) \in FMS_\epsilon) = 1 - o(R(r')) \) as \( r \to \infty \).

Proof. The “proof of main result”, Section 5.4, applies verbatim, except that the two lemmas need to be replaced.

Lemma 5.3 (Tightness of fluid scaling) \( \rightarrow \) Lemma A.1
Lemma 5.4 (Dynamics at cluster points) \( \rightarrow \) Lemma A.2

Lemma A.1 (Tightness of fluid scaling) For every \( r \), with \( K_r \) as defined in Proposition 5.2, \( x^r \in K_r \). The constants used to define \( K_r \) are \( K \) as given in (24), and \( A \) and \( B_r \) from (60) below.

Proof. We prove that \( x^r \) satisfies the two defining conditions of \( K_r \). The condition about initial state is as in the single-hop case. The bounds on the arrival process, the idleness, and service allocation, are as in the single-hop case: for any \( 0 \leq s < t \leq T \) with \( t - s < \delta \),

\[
\begin{align*}
|a^r_n(t) - a^r_n(s)| &< \delta A^{\text{max}} + 2A^{\text{max}}/r', \\
|z^r_n(t) - z^r_n(s)| &< \delta S^{\text{max}} + 2S^{\text{max}}/r', \\
|s^r_\pi(t) - s^r_\pi(s)| &< \delta + 2/r'.
\end{align*}
\]

The bound on queue size is a little different. Note that (44) carries through to the fluid model scaling, i.e.

\[
q^r(t) = q^r(0) + a^r(t) - (I - R^T) \sum_\pi s^r_\pi(t)\pi + z^r(t),
\]

thus

\[
|q^r_n(t) - q^r_n(s)| \leq |a^r_n(t) - a^r_n(s)| + \sum_\pi |(I - R^T)\pi_n| \left| s^r_\pi(t) - s^r_\pi(s) \right| + |z^r_n(t) - z^r_n(s)|
< \left( \delta + 2/r' \right) \left( A^{\text{max}} + |S|NS^{\text{max}}S^{\text{max}} + S^{\text{max}} \right).
\]
Putting all these together,
\begin{equation}
\omega(x^r) < \left( \delta + 2/r' \right) \left( 2NA_{\max} + 2NS_{\max} + N^2 |S|(S_{\max})^2 + |S| \right). \tag{60}
\end{equation}

**Lemma A.2 (Dynamics at cluster points)** Let \( x \) be a cluster point of the sequence \( E_r = \{ x^r(\omega) : \omega \in \text{ARR}^r_{(r)} \} \). Then \( x \in \text{FMS} \).

**Proof.** Equations (48)–(53) all work exactly as in the single-hop case, as does the final equation \( q(0) = q_0 \). The only equations that need further thought are the MW-f equations (54) and (55).

**Proof of (54).** Pick a \( t \) at which \( s_\pi \) is differentiable, and suppose that \( \pi \cdot (I - R)f(q(t)) < \max_{\rho \text{ } R^{-1}} -\rho(I - R)f(q(t)) \). As in Lemma 5.4, proof of (18), it must be that there is some small interval \( I = [r't, r't + r'd] \) such that matching \( \pi \) is not chosen for any \( \tau \in I \), therefore \( \dot{s}_\pi(t) = 0 \).

**Proof of (55).** The discrete (unscaled) system satisfies (45), therefore the scaled systems \( x^r \) do too. Taking the limit yields the fluid equation. \( \square \)

**Restated lemma (11.5 Feasibility)** Suppose that \( \lambda \in \Lambda \). Then the optimization problems \( \text{ALGD}(q) \) and \( \text{ALGP}(q) \) have the same feasible set, for any \( q \in \mathbb{R}_+^N \).

Furthermore, for any scheduling algorithm, every fluid model solution with initial queue size \( q(0) \) has the property that \( q(t) \) is feasible for \( \text{ALGP}(q(0)) \) for all \( t \geq 0 \).

**Proof.** Applying Lemma 6.1 to \( \bar{q} = \bar{R}q, \bar{f} = \bar{R}r \) and \( \bar{\lambda} = \bar{R}\lambda \), the feasible sets of \( \text{ALGD}(W(q)) \) and \( \text{ALGP}(q) \) are the same.

The proof of the second claim is much that of 7.1. By (49),
\begin{align*}
q(t) &= q(0) + t(\lambda - (I - R^T)\sigma(t)) + z(t) \quad \text{where } \sigma(t) = \sum_{\pi} \pi s_\pi(t)/t \\
&\geq q(0) + t(\lambda - (I - R^T)\sigma(t)) \quad \text{since } z \geq 0.
\end{align*}

Next, multiply each side by \( \bar{R} \) and then by \( \xi \). Note that \( \bar{R} \) is a routing matrix and \( \xi \) is a virtual resource, and both consist of nonnegative entries, so the inequality is preserved. We find
\begin{align*}
\xi \cdot \bar{q}(t) &\geq \xi \cdot \bar{q}(0) + t\left( \xi \cdot \bar{X} - \xi \cdot \sigma(t) \right) \quad \text{since } \bar{R} = (I - R^T)^{-1} \\
&= \xi \cdot \bar{q}(0) + t\left( 1 - \xi \cdot \sigma(t) \right) \quad \text{since } \xi \text{ is critical for } \bar{X} \\
&\geq \xi \cdot \bar{q}(0) + t\left( 1 - \text{DUAL}(\sigma(t)) \right) \quad \text{since } \xi \text{ is feasible for } \text{DUAL}(\sigma(t)) \\
&= \xi \cdot \bar{q}(0) + t\left( 1 - \text{PRIMAL}(\sigma(t)) \right) \quad \text{by strong duality} \\
&\geq \xi \cdot \bar{q}(0) \quad \text{since } \sigma(t) \in \Sigma \text{ so } \text{PRIMAL}(\sigma(t)) \leq 1.
\end{align*}
Finally, suppose that $\bar{x}_n = 0$ for some queue $n$. Take (49) and multiply each side by $\bar{R}$ to get

$$\bar{q}(t) = \bar{q}(0) + \bar{\lambda}t - \sum_{\pi} \bar{\pi}\bar{s}(t) + \bar{R}\bar{z}(t)$$

$$= \bar{q}(0) + \bar{\lambda}t - \sum_{\pi} \bar{\pi}\bar{s}(t) \quad \text{by (55)}$$

$$\implies \bar{q}_n(t) \leq \bar{q}_n(0) \quad \text{since } \bar{x}_n = 0.$$ 

We have shown that $\bar{\xi} \cdot \bar{q}(t) \geq \bar{\xi} \cdot \bar{q}(0)$ for all $\bar{\xi} \in \mathcal{S}^*(\bar{X})$, and that $\bar{q}_n(t) \leq \bar{q}_n(0)$ whenever $\bar{x}_n = 0$. Therefore $\bar{q}(t)$ is feasible. \hfill $\square$

**Restated lemma (11.7 Fixed points)** Suppose $\bar{\lambda} \in \Lambda$ and the network is running the MW-f backpressure algorithm. Then $\bar{q} = \Delta W(\bar{q}) \iff \bar{q}$ is a fixed point of the fluid model.

**Proof.** The corresponding result for a single-hop network, Theorem 7.3, expands the result into five equivalent statements, and the same is helpful here. We claim the following are equivalent:

i. $\bar{q} = \Delta W(\bar{q})$

ii. $\bar{q}$ is a fixed point

iii. there exists a fluid model solution with $\bar{q}(t) = \bar{q}$ for all $t$

iv. $\bar{\lambda} \cdot f(\bar{q}) = \max_{\pi \in S} \pi \cdot (I - R)f(\bar{q})$

v. either $\bar{q} = 0$, or $\zeta(\bar{q}) = (I - R)f(\bar{q})/\max_{\pi} \pi \cdot (I - R)f(\bar{q})$ is a critically-loaded virtual resource.

**Proof that** $(i) \implies (ii) \implies (iii) \implies (iv) \implies (v)$. These are very similar to the proofs for Theorem 7.3. We just need to appeal to Lemma 11.4 rather than 4.2 for Lyapunov stability, and to Lemma 11.5 rather than 7.1 for the fact that $\bar{q}(t)$ remains feasible.

**Proof that** $(v) \implies (i)$. If $\bar{q} = 0$ the result is trivial. Otherwise, let $\bar{r} = \Delta W(\bar{q})$. By Lemma 11.6, $\bar{r} = \bar{q} + t(\bar{\lambda} - (I - R^T)\bar{\sigma})$ for some $t \geq 0$ and $\bar{\sigma} \in \Sigma$. Consider the value of $L(\cdot)$ along the trajectory from $\bar{q}$ to $\bar{r}$:

$$\frac{d}{du} L(\bar{r} + t(\bar{\lambda} - (I - R^T)\bar{\sigma})\bar{u}) \bigg|_{u=0} = (\bar{\lambda} - (I - R^T)\bar{\sigma}) \cdot f(\bar{q})$$

$$= (\bar{R}\bar{\lambda} - \bar{\sigma}) \cdot (I - R)f(\bar{q}) \quad \text{since } \bar{R}(I - R)^T = I$$

$$= (\bar{R}\bar{\lambda} - \bar{\sigma}) \cdot (\zeta(\bar{q}) \max_{\pi} \pi \cdot (I - R)f(\bar{q})) \quad \text{substituting in } \zeta$$

$$= (\bar{\lambda} \cdot \zeta - \bar{\sigma} \cdot \zeta) \max_{\pi} \pi \cdot (I - R)f(\bar{q}) \quad \text{writing } \zeta \text{ for } \zeta(\bar{q})$$

$$= (1 - \bar{\sigma} \cdot \zeta) \max_{\pi} \pi \cdot (I - R)f(\bar{q}) \quad \text{since } \zeta \text{ is optimal for DUAL}(\bar{X})$$

$$\geq (1 - \text{DUAL}(\bar{\sigma})) \max_{\pi} \pi \cdot (I - R)f(\bar{q}) \quad \text{since } \zeta \text{ is feasible for DUAL}(\bar{\sigma})$$
Lemma A.3 (Continuity of lifting map) Suppose $\lambda \in \Lambda$ and the network is running the MW-f backpressure algorithm. Then $\Delta W : \mathbb{R}_+^N \rightarrow \mathbb{R}_+^N$ is a continuous function.

Proof. The proof is very similar to that of Lemma 6.3, but the formulae all have to be adjusted to deal with multihop so we will write it all out again.

If $S^*(\tilde{\lambda})$ is empty, then the lifting map is trivial and the result is trivial. In what follows, we shall assume that $S^*(\tilde{\lambda})$ is non-empty, and we will abbreviate it to $S^*$. Furthermore note that, by definition of $S^*$, for every $\xi \in S^*$ we know $\xi \cdot \tilde{\lambda} = 1$ and hence there is some queue $n$ such that $\xi_n > 0$ and $\tilde{\lambda}_n > 0$.

Pick any sequence $q_k^k \rightarrow q$, and let $r^k = \Delta W(q_k^k)$ and $r = \Delta W(q)$. We want to prove that $r^k \rightarrow r$. We shall first prove that there is a compact set $[0, h]^N$ such that $r^k \in [0, h]^N$ for all $k$. We shall then prove that any convergent subsequence of $r^k$ converges to $r$; this establishes continuity of $\Delta W$.

First, compactness. A suitable value for $h$ is

$$h = \max \max \sup_{\xi \in S^*} \frac{\xi \cdot \bar{q}^k}{\xi_n}.$$

Note that the maximums are over a non-empty set, as noted at the beginning of the proof. Note also that $h$ is finite because $q$ is finite. Now, suppose that $r^k \notin [0, h]^N$ for some $k$, i.e. that there is some queue $n$ for which $r^k_n > h$, and let $r' = r^k$ in every coordinate except for $r'_n = h$. We claim that $r'$ satisfies the two constraints of $\text{ALGD}(q^k)$. To see that it satisfies the second constraint, note that $r' \leq r^k$ and hence if $\bar{\lambda}_n = 0$ then $r'_n \leq r^k_n \leq q_n$. To see that it satisfies the first constraint, pick any $\xi \in S^*$. Either $\xi_m = 0$ for all queues $m$ that are downstream of $n$, i.e. for which $\bar{R}_{mn} = 1$; if this is so then

$$\xi \cdot r' = \xi \cdot r^k + \xi \cdot (\tilde{r} - r^k) = \xi \cdot r^k + \sum_{l} (r'_l - r^k_l) \sum_{m} \xi_m \bar{R}_{ml} = \xi \cdot r^k.$$

Or $\xi_m > 0$ for some queue $m$ that is downstream of $n$; if this is so then

$$\xi \cdot r' \geq \xi_m r'_m \geq \xi_m h \geq \xi \cdot \bar{q}^k$$

by construction of $h$.

Applying this repeatedly, if $r^k \notin [0, h]^N$ then we can reduce it to a queue size vector in $[0, h]^N$, thereby improving on $L(r^k)$, yet still meeting the constraints of $\text{ALGD}(q^k)$; this contradicts the optimality of $r^k$. Hence $r^k \in [0, h]^N$.
Next, convergence on subsequences. With a slight abuse of notation, let \( \Delta W(q^k) = r^k \to s \) be a convergent subsequence, and recall that \( \Delta W(q) = r \) and \( q^k \to q \). By continuity of the constraints of ALGD, \( s \) is feasible for ALGD\((q)\); we shall next show that \( L(s) \leq L(r) \). Since \( r \) is the unique optimum, it must be that \( s = r \).

It remains to show that \( L(s) \leq L(r) \). We will construct a sequence \( r^k - \delta^k + \varepsilon^k \) of candidate solutions to ALGD\((q^k)\), choosing \( \delta^k \geq 0 \) and \( \varepsilon^k \) to ensure that the candidate solutions are feasible. Specifically, we define

\[
\delta^k_n = \begin{cases} 
0 & \text{if } \lambda_n > 0 \\
(q_n - q^k_n)^+ & \text{if } \lambda_n = 0 
\end{cases}
\]

and \( P_n = 1 \) if \( \lambda_n > 0 \) and \( 0 \) otherwise, and

\[
\varepsilon^k = \max_{\xi \in S^*} \frac{(\xi \cdot q^k - \xi \cdot q)}{\xi \cdot P}.
\]

We will first deal with the feasibility constraint that kicks in when \( \lambda_n = 0 \). Note that this implies \( \lambda_m = 0 \) for all queues \( m \) that are upstream of \( n \), since \( \lambda_n = \sum_m \bar{R}_{nm} \lambda_m \), and hence that \( \lambda_m = 0 \) for all upstream queues. Using this we find

\[
[\bar{R}(r^k - \delta^k + \varepsilon^k \cdot P)]_n = \sum_m \bar{R}_{nm} [r^k - \delta^k + \varepsilon^k \cdot P]_m
\]

\[
= \sum_m \bar{R}_{nm} (r_m - (\bar{q}_m - \bar{q}^k_m)^+) \quad \text{since } \lambda_m = 0 \text{ when } \bar{R}_{nm} = 1
\]

\[
\leq \sum_m \bar{R}_{nm} r_m - (\bar{q}_n - \bar{q}^k_n)^+
\]

\[
= \bar{r}_n - (\bar{q}_n - \bar{q}^k_n)^+
\]

\[
\leq \bar{q}_n - (\bar{q}_n - \bar{q}^k_n)^+ \quad \text{since } r \text{ is feasible for ALGD}\(q)\)
\]

\[
= \min(q_n, \bar{q}^k_n) \leq \bar{q}^k_n.
\]

Hence \( r^k - \delta^k + \varepsilon^k \) satisfies the second feasibility constraint of ALGD\((q^k)\). For the other feasibility constraint of ALGD\((q^k)\), pick any \( \xi \in S^* \). Then

\[
\xi \cdot \bar{R}(r - \delta^k + \varepsilon^k \cdot P) = \xi \cdot (r - \bar{r}^k) + \varepsilon^k \xi \cdot \bar{P}
\]

\[
\geq (\xi \cdot r - \xi \cdot \bar{q}^k) + (\xi \cdot q^k - \xi \cdot \bar{q})^+ + \xi \cdot \bar{q}^k \quad \text{by construction of } \varepsilon^k
\]

\[
\geq (\xi \cdot q - (\xi \cdot q^k - \xi \cdot \bar{q})^+) \quad \text{since } \bar{r} \text{ is feasible for ALGD}\(q)\)
\]

\[
= \max(\xi \cdot q, \xi \cdot q^k) \geq \xi \cdot q^k.
\]

Since the candidates are feasible solutions to ALGD\((q^k)\), and \( r^k \) is an optimal solution, it must be that

\[
L(r^k) \leq L(r) - \delta^k + \varepsilon^k \cdot P).
\]
Taking the limit as $k \to \infty$, and noting that $L$ is continuous and $\delta^k \to 0$ and $\varepsilon^k \to 0$, we find

$$L(s) \leq L(r)$$

as required. This completes the proof.

Lemma A.4 (Scale-invariance of the lifting map) Suppose $\lambda \in \Lambda$ and the network is running the MW-f backpressure algorithm. If $q$ is a fixed point then $\kappa q$ is also a fixed point for all $\kappa > 0$.

This proof uses a similar method to the proof of the single-hop version, Lemma 6.4. It is much shorter, because the monotonicity assumption gives us a more powerful representation of the lifting map, Lemma 11.6. Also, this version makes a weaker claim, namely that the lifting map is scale-invariant at fixed points, whereas the single-hop version shows that the lifting map is invariant everywhere.

Proof. Suppose that $q$ is a fixed point i.e. $q = \Delta W(q)$, and let $\kappa r = \Delta W(\kappa q)$. Clearly $\kappa q$ is feasible for ALGD($\kappa q$); we shall show that $L(\kappa r) \geq L(\kappa q)$, whence $\kappa q = \kappa r$ by uniqueness of the optimum.

It remains to prove that $L(\kappa r) \geq L(\kappa q)$. By the representation lemma, Lemma 11.6, we can write $\kappa r$ as

$$\kappa r = \kappa q + t(\lambda - (I - R^T)\sigma)$$

for some $t \geq 0$ and some $\sigma \in \Sigma$. Now consider the value of $L$ along a straight-line trajectory from $\kappa q$ to $\kappa r$:

$$\frac{d}{du} L(\kappa q + (\lambda - (I - R^T)\sigma)u)|_{u=0} = (\lambda - (I - R^T)\sigma) \cdot f(\kappa q) = \lambda \cdot f(\kappa q) - \sigma \cdot (I - R) f(\kappa q) \geq \lambda \cdot f(\kappa q) - \max_{\rho \in S} \rho \cdot (I - R) f(\kappa q) \quad \text{for any } \sigma \in \Sigma$$

$$= 0.$$

The final equality is because

$$\lambda \cdot f(q) = \hat{X} \cdot (I - R) f(q) = \max_{\pi \in S} \pi \cdot (I - R) f(q)$$

by part (iv) in the proof of Lemma 11.7, and so

$$\lambda \cdot f(\kappa q) = \hat{X} \cdot (I - R) f(\kappa q) = \max_{\pi \in S} \pi \cdot (I - R) f(\kappa q)$$

using Condition 11.1 and also the monotonicity property (47) that tells us $\hat{X} \in \Sigma$.\qed
Lemma A.5 (Convergence time of the lifting map) Suppose $\lambda \in \Lambda$ and the network is running the MW-f backpressure algorithm. For any $\varepsilon > 0$ there exists some $H_\varepsilon > 0$ such that, for all fluid model solutions with bounded initial value, $|q(t) - \Delta W(q(t))| < \varepsilon$ for all $t \geq H_\varepsilon$.

Proof. The proof of Lemma 7.4 goes through almost verbatim. The only changes are that we need to use the multihop Lyapunov result Lemma 11.4 rather than the single-hop result Lemma 4.2, and we need to use the multihop fixed point result in part (iv) of Lemma 11.7 rather than the single-hop result Theorem 7.3. Specifically, replace the penultimate paragraph by the following.

Time to hit $K_\delta$. Consider first the rate of change of $K(\cdot)$ while the process is in $D \setminus K_\delta$:

$$\dot{K}(q(t)) \leq \dot{L}(q(t)) = \lambda \cdot f(q(t)) - \max_{\pi \in S} \pi \cdot (I - R)f(q(t))$$

$$\leq \sup_{r \in D \setminus K_\delta} \left[ \lambda \cdot f(r) - \max_{\pi \in S} \pi \cdot (I - R)f(r) \right]$$

$$\leq 0 \quad \text{by Lemma 11.4.}$$

This supremum is of a continuous function of $r$, taken over a closed and bounded set, hence the supremum is attained at some $\hat{r} \in D \setminus K_\delta$. If the supremum were equal to 0 then $\lambda \cdot f(\hat{r}) = \max_{\pi} \pi \cdot (I - R)f(\hat{r})$ so $\hat{r} \in I$ by Lemma 11.7; but $\hat{r} \in D \setminus K_\delta$ and we just proved that $I \subset K_\delta$; hence the supremum is some $-\eta_\delta < 0$.

Restated theorem (11.9 Multiplicative state space collapse) Theorem 8.2 holds for the multihop network running the MW-f backpressure algorithm.

Proof. The proof of Theorem 8.2 applies verbatim.