Optimal scheduling algorithms for input-queued switches

Devavrat Shah
EECS & ESD, MIT
devavrat@mit.edu

Damon Wischik
Computer Science, UCL
D.Wischik@cs.ucl.ac.uk

***Abstract*— The input-queued switch architecture is widely used in Internet routers, due to its ability to run at very high line speeds. A central problem in designing an input-queued switch is choosing the scheduling algorithm, i.e. deciding which packets to transfer from ingress ports to egress ports in a given timeslot. Important metrics for evaluating a scheduling algorithm are its throughput and average delay.**

The well-studied 'Maximum-Weight' algorithm has been proved to have maximal throughput [1]; later work [2]–[4] found a wider class of algorithms which also have maximal throughput. The delay performance of these algorithms is less well understood.

In this paper, we present a new technique for analysing scheduling algorithms which can explain their delay performance. In particular, we are able to explain the empirical observations in [2] about the average delay in a parameterized class of algorithms akin to Maximum-Weight. We also propose an optimal scheduling algorithm. Our technique is based on critically-balanced fluid model equations.

I. Introduction

Switching is an integral function in a packet-switched data network. An Internet router has several input ports and several output ports. Its function is to receive packets at input ports, work out which output port to send them to, and then switch them to the correct output port. There are a variety of possible switch architectures; in this paper we are concerned with input-queued (IQ) switches, which work as follows:

A. Input-queued switch

Figure 1 illustrates a 3×3 IQ switch fabric. By '3×3' we mean it has 3 input ports and 3 output ports. (Not all ports need be used, so there is no loss in generality in assuming as many input as output ports.) Packets arriving at input i destined for output j are stored in Virtual Output Queue VOQ(i, j). In each timeslot, the switch fabric can transmit a number of packets from input ports to output ports, subject to the constraints:

i. each input can transmit at most one packet,
ii. each output can receive at most one packet.

Another way to express this is to say that, in each timeslot, the switch can choose a *matching* from inputs to outputs. For example, Figure 1 illustrates a matching in which one packet is transmitted from input port 1 to output port 3, and one from input port 2 to output port 1. The figure also shows a match from input port 3 to output port 2, but since VOQ$(3, 2)$ is empty no packet is transmitted.

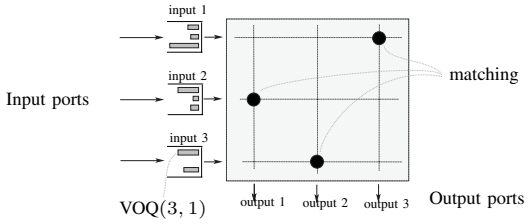

Fig. 1. An input-queued switch, and a matching of inputs to outputs.

The constraints (i)&(ii) mean that the buffer memory needs be accessed only twice per timeslot (once to write an incoming packet, once to read a packet for transmission). This low memory bandwidth means that IQ switches can operate at very high speeds. The constraint (ii) means that no buffers are required at the output ports. We have assumed here, and we will assume throughout, that all packets are of equal size, and that time is slotted so that at most one packet may arrive in any timeslot. In practice, packets are not all the same size, but they are broken up into equal-sized cells before being transmitted across the switch fabric.

B. Scheduling algorithms

The matching of inputs to outputs is chosen by a *scheduling algorithm*. It may take account of queue sizes, ages of packets, or quality-of-service constraints.

For the purposes of this paper, one scheduling algorithm is particularly interesting: the *Maximum-Weight Matching* (MWM) algorithm. In every timeslot, this algorithm chooses a matching as follows: Let Q_{ij} be the queue size at VOQ(i, j). Given a matching which matches input i to output $o(i)$, define the *weight* of that matching to be $\sum_i Q_{i\,o(i)}$. Among all possible matchings choose one with the greatest weight (breaking ties randomly). Another interesting algorithm is MWM-α, which among all matchings chooses one with the greatest $\sum_i Q^{\alpha}_{i\,o(i)}$, for a specified $\alpha > 0$. Thus MWM is the same as MWM-1.

There are two main metrics for evaluating scheduling algorithms: throughput, and delay. Roughly speaking, an algorithm is said to have *100% throughput* if it can carry as much traffic as an omniscient scheduling algorithm (i.e. one which knows all future packet arrivals). This is formalized in Section III. Delay performance is harder to define; we discuss it further below.

1-4244-0222-0/06/$20.00 (c)2006 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the Proceedings IEEE Infocom.

C. Previous work

The IQ switch architecture has been studied for more than a decade [5]–[8]. A good deal is now known about throughput. MWM has been shown to have 100% throughput, under a 'friendly' arrival distribution [1]. A generalization of this result in the context of radio-hop networks (under the same arrival distribution) was shown earlier [9]. These results have been generalized to arbitrary arrival distributions [10]. A class of algorithms akin to MWM have also been shown to have 100% throughput [2]–[4].

Less is known about the delay performance of scheduling algorithms. Bounds on delay have been derived for MWM and certain approximations to MWM, under 'friendly' arrival distributions [4], [11]. A systematic simulation study of the MWM-α algorithm led to the following conjecture [2]:

Conjecture 1 *The average delay of the MWM-α algorithm decreases as α decreases.*

The delay performance of a generalized switch (of which an IQ switch is a special case) under heavy traffic has been studied. In the special case where exactly one port of the switch is saturated, MWM has been shown to be optimal [12]. However, in this setting all the MWM-α algorithms have essentially the same performance, and so this theory does not help us resolve Conjecture 1.

Though MWM and related algorithms provide maximal throughput, they are too complex to be implementable in high-speed switches. This has motivated the design of simpler high-performance scheduling algorithms [6], [7], [13], [14]. In this paper we do not address complexity. Nonetheless we hope that the insights our analysis gives can be used to assist in the design of good implementable algorithms.

D. Contribution

The work in this paper is motivated by a desire to prove Conjecture 1. As in [12], we will look at systems which are heavily loaded. The formalism we will use is that of *heavy traffic theory*, a general body of theory which has been fully developed in the setting of queueing networks [15].

The outline of this paper is as follows. In Section II we give the *fluid model equations* for describing the behaviour of an IQ scheduling algorithm. In Section III we define throughput, and review the connection with fluid model equations.

The main contribution of the paper is in Sections IV & V. In Section IV we characterize the steady-state behaviour of the balanced fluid model equations, by studying combinatorial properties of switches and matchings; in Section V we describe the relationship between this steady-state behaviour and the performance of a heavily-loaded switch. The key discovery is that the $n \times n$ matrix of queue sizes Q_{ij} actually lives in a $2n-1$ dimensional subspace of $\mathbb{R}^{n^2}$ (called the invariant set), and the geometry of this subspace is determined by the scheduling algorithm.

Finally in Section VI we use the idea of the invariant set to resolve Conjecture 1, to conjecture an optimal algorithm, and to suggest future directions for work on scheduling algorithms.

E. Notation

We first specify our notation. Let $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ and $\mathbb{Z}_+ = \{i \in \mathbb{Z} : i \geq 0\}$. Let 1_X be the indicator function: $1_{\text{true}} = 1$ and $1_{\text{false}} = 0$.

Let $\mathbb{M}$ be the set of $n \times n$ real-valued matrices, and $\mathbb{M}_+$ the subset consisting of $\mathbb{R}_+$-valued matrices. Write matrices as $\mathbf{a} = [a_{ij}]$. Let $\mathbf{1} = [1]$. Let $\mathbb{S} \subset \mathbb{M}_+$ be the set of matrices whose row sums and column sums are all equal to 1, i.e. the set of doubly stochastic matrices. The set of doubly substochastic matrices is the subset of matrices in $\mathbb{M}_+$ whose row and column sums are all bounded above by 1. Let $\mathbb{P} \subset \mathbb{S}$ be the set of matrices $\boldsymbol{\pi}$ for which $\pi_{ij} \in \{0, 1\}$ for all i and j, i.e. the set of permutation matrices. For $\mathbf{a} \in \mathbb{M}$ write

$$a_{i\oplus} = \sum_j a_{ij}, \quad a_{\oplus j} = \sum_i a_{ij}, \quad a_{\oplus\oplus} = \sum_{i,j} a_{ij}.$$

When $\mathbf{a}$ is a matrix of queue sizes, we call these the workload at input port i, the workload at output port j, and the total workload respectively. Define the workload map

$$W(\mathbf{a}) = \left(a_{1\oplus}, \ldots, a_{n\oplus}; a_{\oplus 1}, \ldots, a_{\oplus n}; a_{\oplus\oplus}\right)$$

and let $\mathbb{W}$ be the set of all possible workload vectors, $\mathbb{W} = \{W(\mathbf{a}) : \mathbf{a} \in \mathbb{M}_+\}$. Write a workload $w \in \mathbb{W}$ as $w = (w_{1\cdot}, \ldots, w_{n\cdot}; w_{\cdot 1}, \ldots, w_{\cdot n}; w_{\cdot\cdot})$. Note that $\mathbb{W}$ has dimension $2n-1$, since $w_{\oplus\cdot} = w_{\cdot\oplus} = w_{\cdot\cdot}$.

For $\mathbf{a}$, $\mathbf{b} \in \mathbb{M}$ and $f : \mathbb{R} \to \mathbb{R}$ let

$$\mathbf{ab} = \left[a_{ij}b_{ij}\right] \in \mathbb{M}, \quad f(\mathbf{a}) = \left[f(a_{ij})\right] \in \mathbb{M}, \text{ and}$$
$$\mathbf{a} \cdot \mathbf{b} = \sum_{i,j} a_{ij}b_{ij} \in \mathbb{R}.$$

Let component-wise multiplication have precedence over $\cdot$, so that $\mathbf{a} \cdot \mathbf{bc} = \mathbf{a} \cdot (\mathbf{bc})$. The $\cdot$ operator is commutative, and distributive over addition so that $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$.

II. Switch dynamics

In this section we describe fluid model equations, a natural and intuitive way to describe the dynamics of a switch. They are used rigorously in two ways, in the following sections:

i. to show that a scheduling algorithm has 100% throughput (Section III)
ii. to derive state space collapse (Section IV–V)

We will not give rigorous derivations of the fluid model equations. This can be found in [10], [16]. Instead we will give motivation.

We will first describe the general setup in Section II-A and the fluid scaling in Section II-B. Then we give the fluid model equations in two parts: the algorithm-independent equations in Section II-C, and the algorithm-dependent equations in Section II-D. Any solution to all these equations is called a *fluid model solution*.

A. Queueing model

Let timeslots be indexed by $\tau \in \mathbb{Z}_+$, starting at $\tau = 0$. Let $\mathbf{Q}(\tau) = [Q_{ij}(\tau)] \in \mathbb{M}_+$ denote the matrix of the queue sizes at the end of timeslot τ. Since work arrives in discrete packets, $Q_{ij}(\tau) \in \mathbb{Z}_+$ for all τ. We are interested in describing the dynamics of $\mathbf{Q}(\cdot)$, which depend on the initial conditions, the arrival process and the scheduling algorithm.

First, the initial condition. For simplicity we make the following assumption:

Assumption 2 *We assume that the switch starts empty at time* 0*, i.e.*

$$\mathbf{Q}(0) = \mathbf{0}. \tag{1}$$

Next, the dynamics. Let $\mathbf{A}(\tau)$ be the cumulative arrival process up to timeslot τ, i.e. $A_{ij}(\tau)$ is the number of packets that have arrived at input i destined for output j in the time interval $[0, \tau]$, with $\mathbf{A}(0) = \mathbf{0}$. The arrivals in timeslot τ are thus $\mathbf{A}(\tau) - \mathbf{A}(\tau - 1)$. Similarly, let $\mathbf{D}(\tau)$ be the cumulative departure process from the virtual output queues. Then

$$\mathbf{Q}(\tau) = \mathbf{Q}(0) + \mathbf{A}(\tau) - \mathbf{D}(\tau) = \mathbf{A}(\tau) - \mathbf{D}(\tau) \tag{2}$$

(where the last equality uses Assumption 2). Now for the scheduling algorithm. Let $S_{\boldsymbol{\pi}}(\tau)$ be the cumulative number of timeslots that the scheduling algorithm has devoted to matching $\boldsymbol{\pi} \in \mathbb{P}$ in the time interval $[0, \tau]$, with $S_{\boldsymbol{\pi}}(0) = 0$ for all $\boldsymbol{\pi}$. We will use the convention that departures in timeslot τ happen at the beginning of the timeslot, and that arrivals happen at the end, so that

$$\begin{aligned} D_{ij}(\tau) - D_{ij}(\tau - 1) = \\ \sum_{\boldsymbol{\pi} \in \mathbb{P}} \pi_{ij} \big[S_{\boldsymbol{\pi}}(\tau) - S_{\boldsymbol{\pi}}(\tau - 1) \big] 1_{Q_{ij}(\tau-1)>0} \end{aligned} \tag{3}$$

In each time slot, exactly one matching is chosen. Thus

$$\sum_{\boldsymbol{\pi} \in \mathbb{P}} S_{\boldsymbol{\pi}}(\tau) = \tau. \tag{4}$$

The service process is completely described by $S(\cdot) = \{S_{\boldsymbol{\pi}}(\cdot), \boldsymbol{\pi} \in \mathbb{P}\}$, and the tuple

$$\mathcal{X}(\cdot) = \big(\mathbf{Q}(\cdot), \mathbf{A}(\cdot), \mathbf{D}(\cdot), S(\cdot) \big)$$

completely describes the dynamics of the switch.

B. Fluid scaling

The fluid model equations describe the switch at the 'rate' level, rather than the packet level. Instead of looking at $\mathcal{X}(\tau)$, we now look at the limit

$$x(t) = \lim_{r \to \infty} \frac{1}{r} \mathcal{X}(rt), \quad t \in \mathbb{R}_+ \tag{5}$$

where for $t \notin \mathbb{Z}_+$

$$\mathcal{X}(t) = (1 - t - \lfloor t \rfloor) \mathcal{X}(\lfloor t \rfloor) + (t - \lfloor t \rfloor) \mathcal{X}(\lfloor t \rfloor + 1).$$

Let $x(t) = \big(\mathbf{q}(t), \mathbf{a}(t), \mathbf{d}(t), s(t) \big)$. We make the following assumption about $x(\cdot)$:

Assumption 3 *Assume that the limit $x(\cdot)$ exists, and is absolutely continuous.*

An absolutely continuous function is differentiable almost everywhere. Write $\dot{\xi}(t)$ for the derivative of $\xi(t)$ at t, and take any equations involving $\dot{\xi}(t)$ to hold only where $\xi(t)$ is differentiable.

Under mild probabilistic assumptions on the arrival process, this assumption holds almost surely—see [10], [16]—though the limit $x(\cdot)$ is in general random. We will also make an assumption about the form of the limiting arrival process $\mathbf{a}(\cdot)$, the first part of which also holds under very mild probabilistic assumptions.

Assumption 4 *Assume that $\mathbf{a}(t) = \boldsymbol{\lambda} t$ for some $\boldsymbol{\lambda} \in \mathbb{M}_+$, called the matrix of* mean arrival rates*. Suppose that $\boldsymbol{\lambda}$ is doubly substochastic, i.e.*

$$\lambda_{i\oplus} \leq 1 \text{ for all } i, \quad \lambda_{\oplus j} \leq 1 \text{ for all } j.$$

The assumption that $\lambda_{i\oplus} \leq 1$ reflects the constraint that at most one packet can be transmitted from input port i per timeslot; the assumption that $\lambda_{\oplus j} \leq 1$ reflects the constraint that at most one packet can be received by output port j per timeslot. If these constraints are not met then the queue sizes will blow up. We say that $\boldsymbol{\lambda}$ is *admissible* if it is strictly doubly substochastic, i.e.

$$\lambda_{i\oplus} < 1 \text{ for all } i, \quad \lambda_{\oplus j} < 1 \text{ for all } j. \tag{6}$$

If $\lambda_{i\oplus} = 1$ we say that input port i is *critically loaded*; similarly for output ports.

C. Algorithm-independent dynamics

The fluid model equations corresponding to Assumption 4 and equations (2)–(4) are

$$\mathbf{a}(t) = \boldsymbol{\lambda} t \tag{7}$$

$$\mathbf{q}(t) = \mathbf{a}(t) - \mathbf{d}(t) \tag{8}$$

$$\dot{d}_{ij}(t) = \sum_{\boldsymbol{\pi} \in \mathbb{P}} \pi_{ij} \dot{s}_{\boldsymbol{\pi}}(t) 1_{q_{ij}(t)>0}. \tag{9}$$

$$\sum_{\boldsymbol{\pi} \in \mathbb{P}} s_{\boldsymbol{\pi}}(t) = t \tag{10}$$

We have remarked that Assumption 3 can be proved under mild probabilistic assumptions on the arrival process; it can also be proved that the limit $x(\cdot)$ almost surely satisfies these equations [10], [16].

Some further notation will be helpful for the rest of this paper. The matrix of instantaneous service rates $\boldsymbol{\sigma}(t)$ is

$$\boldsymbol{\sigma}(t) = \sum_{\boldsymbol{\pi} \in \mathbb{P}} \boldsymbol{\pi} \dot{s}_{\boldsymbol{\pi}}(t).$$

Then equations (7)–(9) can be rewritten as

$$\dot{q}_{ij}(t) = \begin{cases} \lambda_{ij} - \sigma_{ij}(t) & \text{if } q_{ij} > 0 \\ \big(\lambda_{ij} - \sigma_{ij}(t) \big)^+ & \text{otherwise} \end{cases}$$

We will write this in the following compact form:

$$\dot{\mathbf{q}}(t) = \big(\boldsymbol{\lambda} - \boldsymbol{\sigma}(t) \big)^{+[\mathbf{q}(t)=0]}. \tag{11}$$

D. Algorithm-dependent dynamics

The scheduling algorithm decides which matching to use in each timeslot, i.e. it specifies $\{S_{\boldsymbol{\pi}}(\cdot),\ \boldsymbol{\pi} \in \mathbb{P}\}$. We will now describe some different scheduling algorithms, and associated fluid model equations: the basic Maximum-Weight matching algorithm MWM (II-D.1); and Generalized Maximum-Weight matching MWMf, which includes MWM-α as a subclass (II-D.2).

As with the algorithm-independent fluid model equations, it can be shown (under mild probabilistic assumptions on the arrival process) that the limit process $x(\cdot)$ satisfies the following fluid model equations almost surely [10], [16].

1) MWM: At time τ, MWM chooses a matching $\boldsymbol{\pi}^* \in \mathbb{P}$ such that

$$\boldsymbol{\pi}^* = \operatorname*{argmax}_{\boldsymbol{\pi}\in\mathbb{P}} \boldsymbol{\pi} \cdot \mathbf{Q}(\tau - 1).$$

If there are several optimal matchings, $\boldsymbol{\pi}^*$ is chosen randomly among them. (Recall that under our convention departures occur at the beginning of a timeslor, and so the matching at time τ depends on the queue sizes at $\tau - 1$.) Equivalently,

$$S_{\boldsymbol{\pi}}(\tau) = S_{\boldsymbol{\pi}}(\tau - 1) \quad \text{if} \quad \boldsymbol{\pi} \cdot \mathbf{Q}(\tau - 1) < \max_{\boldsymbol{\rho}\in\mathbb{P}} \boldsymbol{\rho} \cdot \mathbf{Q}(\tau - 1).$$

The corresponding fluid model equation is

$$\dot{s}_{\boldsymbol{\pi}}(t) = 0 \quad \text{if} \quad \boldsymbol{\pi} \cdot \mathbf{q}(t) < \max_{\boldsymbol{\rho}\in\mathbb{P}} \boldsymbol{\rho} \cdot \mathbf{q}(t) \tag{12}$$

2) MWMf: MWMf is a generalization of MWM. Let f be some function $\mathbb{R}_+ \to \mathbb{R}_+$. Then MWMf chooses a matching $\boldsymbol{\pi}^*$ in timeslot τ such that

$$\boldsymbol{\pi}^* \cdot f\big(\mathbf{Q}(\tau - 1)\big) = \max_{\boldsymbol{\pi}\in\mathbb{P}} \boldsymbol{\pi} \cdot f\big(\mathbf{Q}(\tau - 1)\big).$$

The fluid equation analogous to (12) is

$$\dot{s}_{\boldsymbol{\pi}}(t) = 0 \quad \text{if} \quad \boldsymbol{\pi} \cdot f(\mathbf{q}(t)) < \max_{\boldsymbol{\rho}\in\mathbb{P}} \boldsymbol{\rho} \cdot f(\mathbf{q}(t)). \tag{13}$$

The special case of $f(x) = x^\alpha$, $\alpha > 0$, is called MWM-α.

In this paper we will only consider functions f which satisfy

Assumption 5 *Assume f is differentiable and strictly increasing with $f(0) = 0$. Assume also that for any $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n) \in \mathbb{R}_+^n$*

$$\sum_i f(x_i) > \sum_i f(y_i) \Rightarrow \sum_i f(rx_i) > \sum_i f(ry_i) \quad \forall r > 0.$$

This is needed to ensure that the fluid limit exists [16, Section 4.2.2]. Clearly $f(x) = x^\alpha$ satisfies this assumption.

III. Throughput analysis via fluid model

A very powerful use of fluid models is in analysing throughput [10]. The general idea is this:

i. Take the arrival process $\mathbf{A}$ to be stochastic, and make some mild assumptions on its distribution.
ii. Prove that any solution $x(\cdot)$ of the fluid model equations, with initial queue size $\mathbf{q}(0) = \mathbf{0}$, satisfies $\mathbf{q}(t) = \mathbf{0}$ for almost all $t \geq 0$. This is called *weak stability*. ('Almost all' is with respect to the Lebesgue measure.)
iii. It can be shown that (i)&(ii) together imply that $\mathbf{Q}(t) = O(t)$ whenever $\boldsymbol{\lambda}$ is admissible. This is known as *rate-stability* or *having 100% throughput*.

We will now demonstrate step (ii) for MWMf. The remaining steps are fleshed out in [16]. The conclusion is that MWMf has 100% throughput.

Theorem 1 *Under any arrival process satisfying Assumption 4, with admissible rate-matrix $\boldsymbol{\lambda}$, the switch operating under the MWMf algorithm is rate-stable.*

Proof. Assume $\boldsymbol{\lambda}$ is admissible. Let f be the weight function for the MWMf algorithm. Define

$$L(\mathbf{q}) = F(\mathbf{q}) \cdot \mathbf{1}, \quad \text{where } F(x) = \int_0^x f(y)\, dy.$$

It can be shown that L is a Lyapunov function, i.e. that for any fluid model solution, at any t such that $\mathbf{q}(t) \neq \mathbf{0}$,

$$\frac{d}{dt} L\big(\mathbf{q}(t)\big) < 0.$$

The proof is very similar to the proof of Theorem 5(i), so we omit it.

It is shown in [10] that for any absolutely continuous function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(0) = 0$ and $df(t)/dt \leq 0$ almost everywhere where $f(t) > 0$, $f(t) = 0$ for almost all $t \geq 0$.

This lets us conclude that $L(\mathbf{q}(t)) = 0$ almost everywhere.□

IV. Equilibrium Analysis of Fluid Model

The analysis of stability in the previous section assumes that the arrival rate matrix $\boldsymbol{\lambda}$ is admissible, i.e. that $\lambda_{i\oplus} < 1$ and $\lambda_{\oplus j} < 1$ for every i and j. By studying the behaviour of the fluid model equations for admissible $\boldsymbol{\lambda}$ we showed that MWMf has 100% throughput.

In this section we will use the fluid model equations in a different way. We will analyse their behaviour when the switch is critically loaded, i.e. when $\lambda_{i\oplus} = 1$ for some i, and/or $\lambda_{\oplus j} = 1$ for some j. In particular, we will characterize the *invariant states*.

Definition 6 (Invariant State) *We say that a state $\mathbf{q} \in \mathbb{M}_+$ is* invariant *(for a switch with given arrival rate matrix and scheduling algorithm) if any solution to the fluid model equations for that switch has queue size $\mathbf{r}(\cdot)$ which satisfies*

$$\mathbf{r}(t) = \mathbf{q} \quad \Longrightarrow \quad \mathbf{r}(s) = \mathbf{q} \text{ for all } s \geq t.$$

In Theorem 1 we showed that for MWMf with admissible $\boldsymbol{\lambda}$ the state $\mathbf{0}$ is invariant. In this section we will prove some useful results about matchings (Section IV-A), characterize invariant states of MWMf as the solutions to a certain optimization problem (Section IV-B), and find the time taken to converge to an invariant state (Section IV-C).

A. Preliminary results about matchings

The well-known Birkhoff–von Neumann theorem states that the set of all doubly stochastic matrices $\mathbb{S}$ is convex, and the set of its extreme points is $\mathbb{P}$. Thus any $\mathbf{a} \in \mathbb{S}$ can be written

$$\mathbf{a} = \sum_{\boldsymbol{\pi} \in \mathbb{P}} \gamma_{\boldsymbol{\pi}} \boldsymbol{\pi}, \quad \sum_{\boldsymbol{\pi}} \gamma_{\boldsymbol{\pi}} = 1, \quad \gamma_{\boldsymbol{\pi}} \geq 0 \text{ for all } \boldsymbol{\pi}.$$

Many of our results concern maximum weight matchings. Given $\mathbf{a} \in \mathbb{M}_+$, let $m(\mathbf{a})$ be the weight of a maximum weight matching $m(\mathbf{a}) = \operatorname{argmax}_{\boldsymbol{\pi}} \boldsymbol{\pi} \cdot \mathbf{a}$, let $\mathcal{M}(\mathbf{a})$ be the set of maximum weight matchings $\mathcal{M}(\mathbf{a}) = \{\boldsymbol{\pi} : \boldsymbol{\pi} \cdot \mathbf{a} = m(\mathbf{a})\}$, and let $M(\mathbf{a})$ be the matrix which indicates which entries are involved in a maximum weight matching:

$$M(\mathbf{a})_{ij} = \begin{cases} 1 & \text{if } \pi_{ij} = 1 \text{ for some } \boldsymbol{\pi} \in \mathcal{M}(\mathbf{a}) \\ 0 & \text{otherwise} \end{cases}$$

The set $\mathcal{M}(\mathbf{a})$ exhibits an important closure property:

Lemma 2 *Let $\boldsymbol{\pi} \in \mathbb{P}$, and suppose $M(\mathbf{a})_{ij} = 1$ whenever $\pi_{ij} = 1$. Then $\boldsymbol{\pi} \in \mathcal{M}(\mathbf{a})$.*

Proof. Define the matrix $\mathbf{b}$ by

$$\mathbf{b} = \sum_{\boldsymbol{\rho} \in \mathcal{M}(\mathbf{a})} \boldsymbol{\rho}. \tag{14}$$

It is easy to see that $\mathbf{b} - \boldsymbol{\pi}$ has non-negative entries, and that its row and column sums are all equal, so by the Birkhoff–von Neuman decomposition

$$\mathbf{b} = \boldsymbol{\pi} + \sum_{\boldsymbol{\rho} \in \mathbb{P}} \gamma_{\boldsymbol{\rho}} \boldsymbol{\rho} \tag{15}$$

where each $\gamma_{\boldsymbol{\rho}} \geq 0$ and $\sum \gamma_{\boldsymbol{\rho}} = |\mathcal{M}(\mathbf{a}) - 1|$.

Let $m = m(\mathbf{a})$ be the weight of a maximum weight matching. By (14), $\mathbf{a} \cdot \mathbf{b} = |\mathcal{M}(\mathbf{a})| m$. On the other hand, by maximality and (15), it must be that $\mathbf{a} \cdot \boldsymbol{\pi} \leq m$ and $\mathbf{a} \cdot (\mathbf{b} - \pi) \leq |\mathcal{M}(q) - 1| m$. If either of these inequalities were strict we would obtain $|\mathcal{M}(q)| m < |\mathcal{M}(q)| m$, a contradiction. Hence $\mathbf{a} \cdot \boldsymbol{\pi} = m$, and so $\boldsymbol{\pi} \in \mathcal{M}(\mathbf{a})$. □

Let $\boldsymbol{\lambda} \in \mathbb{M}_+$ be doubly sub-stochastic. It can be augmented to form a doubly stochastic matrix $\boldsymbol{\lambda} + \boldsymbol{\delta}$, where the matrix $\boldsymbol{\delta}$ satisfies

$$\delta_{ij} > 0 \quad \text{if} \quad \lambda_{i\oplus} < 1 \text{ and } \lambda_{\oplus j} < 1.$$

We will say that such a $\boldsymbol{\delta}$ is *complementary* to $\boldsymbol{\lambda}$, and that $\boldsymbol{\lambda} + \boldsymbol{\delta}$ is an *augmentation* of $\boldsymbol{\lambda}$. (One way to obtain such a $\boldsymbol{\delta}$ is to start with $\delta_{ij} = \varepsilon$ for the entries specified above, where $\varepsilon = n^{-1} \min_i (1 - \lambda_{i\oplus}) \wedge \min_j (1 - \lambda_{\oplus j})$, and then to add the 'deficit' amount according to the transport algorithm.)

The next lemma gives a useful description of which switch states may be reached from other switch states.

Lemma 3 *Let $\boldsymbol{\lambda}$ be doubly substochastic. Let $\mathbf{q}$, $\mathbf{r} \in \mathbb{M}_+$ be such that $q_{ij} = r_{ij} = 0$ whenever $\lambda_{ij} = 0$. Suppose that*

$$r_{i\oplus} \geq q_{i\oplus} \text{ if } \lambda_{i\oplus} = 1, \quad \text{for all } i,$$
$$r_{\oplus j} \geq q_{\oplus j} \text{ if } \lambda_{\oplus j} = 1, \quad \text{for all } j.$$

Then there exists a doubly stochastic matrix $\boldsymbol{\sigma} \in \mathbb{S}$, a positive matrix $\boldsymbol{\varepsilon} \in \mathbb{M}_+$, and a duration $t > 0$ such that

$$\mathbf{r} = \mathbf{q} + t(\boldsymbol{\lambda} - \boldsymbol{\sigma}) + \boldsymbol{\varepsilon}. \tag{16}$$

Suppose that in addition

$$r_{i\oplus} \geq q_{i\oplus} \text{ for all } i \quad \text{and} \quad r_{\oplus j} \geq q_{\oplus j} \text{ for all } j.$$

Then for any augmentation $\boldsymbol{\lambda}^+$ of $\boldsymbol{\lambda}$ there exist $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}$ and t as above such that

$$\mathbf{r} = \mathbf{q} + t(\boldsymbol{\lambda}^+ - \boldsymbol{\sigma}) + \boldsymbol{\varepsilon}. \tag{17}$$

Proof. Let $\boldsymbol{\rho} = \boldsymbol{\lambda} - \delta(\mathbf{r} - \mathbf{q})$ for sufficiently small $\delta > 0$. We will show that $\boldsymbol{\rho}$ is a doubly sub-stochastic matrix with non-negative entries.

First we show that all entries of $\boldsymbol{\rho}$ are non-negative, that is, $\rho_{ij} \geq 0$. Now, if $\lambda_{ij} > 0$, then by choosing δ small enough, ρ_{ij} can be made positive; else if $\lambda_{ij} = 0$ then trivially by constraints on $\mathbf{q}$ and $\mathbf{r}$ we obtain $\rho_{ij} = 0$. Thus, $\boldsymbol{\rho} \in \mathbb{M}_+$.

Next, we show that $\boldsymbol{\rho}$ is doubly substochastic, that is, $\rho_{i\oplus} \leq 1$ and $\rho_{\oplus j} \leq 1$ for all i and j. Consider $\rho_{i\oplus}$: either $\lambda_{i\oplus} < 1$, in which case $\rho_{i\oplus} < 1$ for sufficiently small δ; or $\lambda_{i\oplus} = 1$, in which case $r_{i\oplus} \geq q_{i\oplus}$ and $\rho_{i\oplus} \leq 1$. Similarly, $\rho_{\oplus j} \leq 1$ for all j.

Thus $\boldsymbol{\rho}$ is doubly substochastic non-negative matrix. Hence there exists an augmentation of $\boldsymbol{\rho}$, i.e. there exists a doubly stochastic matrix $\boldsymbol{\sigma}$ for which $\boldsymbol{\rho} \leq \boldsymbol{\sigma}$ componentwise. Then

$$\mathbf{q} + \delta^{-1}(\boldsymbol{\lambda} - \boldsymbol{\sigma}) \leq \mathbf{q} + \delta^{-1}(\boldsymbol{\lambda} - \boldsymbol{\rho}) = \mathbf{r}.$$

This proves (16).

The proof of (17) is similar, with $\boldsymbol{\rho} = \boldsymbol{\lambda}^+ - \delta(\mathbf{r} - \mathbf{q})$. It makes use of the fact that $\lambda^+_{ij} = 0$ implies $\lambda_{ij} = 0$. □

B. Invariant states of MWMf

In this section we study the invariant states of fluid model solutions of MWMf, in a critically loaded switch. We exhibit a Lyapunov function for the system state, and we characterize invariant states as the solution to an optimization problem whose objective is the Lyapunov function.

Let f be the weight function, which we take to satisfy Assumption 5. Recall the Lyapunov function $L(\mathbf{q}) = F(\mathbf{q}) \cdot \mathbf{1}$ where $F(x) = \int_0^x f(y)\, dy$. Let $\boldsymbol{\lambda}$ be the doubly stochastic matrix of mean arrival rates. For a workload vector $w \in \mathbb{W}$, define the convex optimization problem MWMf-CP(w) to be

$$\begin{aligned} \text{minimize} \quad & L(\mathbf{q}) \quad \text{over } \mathbf{q} \in \mathbb{M}_+ \\ \text{such that} \quad & q_{i\oplus} \geq w_{i\cdot} \text{ if } \lambda_{i\oplus} = 1 \\ & q_{\oplus j} \geq w_{\cdot j} \text{ if } \lambda_{\oplus j} = 1 \\ & q_{ij} = 0 \text{ if } \lambda_{ij} = 0 \end{aligned}$$

Note that we may as well take the optimum over $\{\mathbf{q} : q_{ij} \leq w_{\cdot\cdot} \; \forall i, j\}$, which is a bounded set. Note also that the objective function is strictly convex, since f is a strictly increasing function on $\mathbb{R}_+$. Thus the optimization problem has a unique solution. Accordingly we define

Definition 7 (Lifting Map) *The lifting map $\Delta : \mathbb{W} \to \mathbb{M}_+$ maps w to the unique solution of optimization problem MWMf-CP(w).*

Lemma 4 *The lifting map is continuous.*

Proof. Let $w^n \to w$ in $\mathbb{W}$. Let $\mathbf{q}^n = \Delta w^n$ and $\mathbf{q} = \Delta w$. As we noted above, $\mathbf{q}^n$ lies in the bounded set $\{\mathbf{r} : r_{ij} \leq w^n_{::}\ \forall i, j\}$. Thus there is a convergent subsequence $\mathbf{q}^{m(n)} \to \mathbf{q}^*$. By continuity of the constraints in MWMf-CP(w^n), $\mathbf{q}^*$ satisfies the constraints in MWMf-CP(w). By optimality of $\mathbf{q}$ for MWMf-CP(w), it must be that $L(\mathbf{q}^*) \geq L(\mathbf{q})$. We will now show that $L(\mathbf{q}^*) \leq L(\mathbf{q})$; then $\mathbf{q}^* = \mathbf{q}$ by uniqueness of the optimum, and hence Δ is continuous.

Let $\varepsilon^m = (\max_i w^m_{i\cdot} - w_{i\cdot}) \vee (\max_j w^m_{\cdot j} - w_{\cdot j})$. Since $w^m \to w$, $\varepsilon^m \to 0$. Now consider $\mathbf{q} + \varepsilon^m \mathbf{1}$ as a candidate solution to MWMf-CP(w^m). By choice of ε^m it is a feasible solution. By optimality of $\mathbf{q}^m$,

$$L(\mathbf{q}^m) \leq L(\mathbf{q} + \varepsilon^m \mathbf{1}).$$

Since L is continuous and $\mathbf{q}^m \to \mathbf{q}^*$, we find $L(\mathbf{q}^*) \leq L(\mathbf{q})$. This completes the proof. □

The following two theorems give two equivalent characterizations of invariant states, one of them in terms of Δ. Recall that $W(\mathbf{q})$ gives the vector of workloads for $\mathbf{q}$.

Theorem 5 *For a switch operating under the MWMf algorithm,*

i. For any fluid model solution $\mathbf{r}(\cdot)$, $dL(\mathbf{r}(t))/dt \leq 0$;
ii. $\mathbf{q}$ is an invariant state $\Leftrightarrow \mathbf{q} = \Delta W(\mathbf{q})$;
iii. $\mathbf{q}$ is an invariant state $\Leftrightarrow dL(\mathbf{r}(t))/dt = 0$ for any fluid model solution $\mathbf{r}(\cdot)$ starting at $\mathbf{r}(0) = \mathbf{q}$.

Proof. Proof of (i). Recall that $\mathbf{r}(t)$ is absolutely continuous (Assumption 3), and note that $L(\cdot)$ is continuous; thus the derivative of $L(\mathbf{r}(t))$ exists for almost all t. At such points, the fluid model equations tell us

$$\begin{aligned}
\frac{d}{dt}L(\mathbf{r}(t)) &= f(\mathbf{r}(t)) \cdot (\boldsymbol{\lambda} - \boldsymbol{\sigma}(t))^{+[\mathbf{r}(t)=0]} \\
&= f(\mathbf{r}(t)) \cdot (\boldsymbol{\lambda} - \boldsymbol{\sigma}(t)) \quad \text{since } f(0) = 0 \\
&\leq f(\mathbf{r}(t)) \cdot (\boldsymbol{\lambda}^+ - \boldsymbol{\sigma}(t)) \quad \text{since } \boldsymbol{\lambda} \leq \boldsymbol{\lambda}^+ \qquad (18) \\
&= f(\mathbf{r}(t)) \cdot \boldsymbol{\lambda}^+ - m(f(\mathbf{r}(t))) \quad \text{by (13)} \\
&= \sum_{\boldsymbol{\pi} \in \mathbb{P}} \gamma_\pi f(\mathbf{r}(t)) \cdot \boldsymbol{\pi} - m(f(\mathbf{r}(t))) \text{ decomposing } \boldsymbol{\lambda}^+ \\
&\leq m(f(\mathbf{r}(t))) - m(f(\mathbf{r}(t))) \qquad (19) \\
&= 0.
\end{aligned}$$

Proof of (ii,⇐). Let $w = W(\mathbf{q})$, and suppose that $\mathbf{q}$ solves MWMf-CP(w). Let $\mathbf{r}(t)$ be any fluid model solution with $\mathbf{r}(0) = \mathbf{q}$. Now $dL(\mathbf{r}(t))/dt \leq 0$ by (i). We will shortly show that $\mathbf{r}(t)$ is a feasible solution to MWMf-CP(w) for all t. Then $dL(\mathbf{r}(t))/dt = 0$ by optimality of $\mathbf{q}$, and each $\mathbf{r}(t)$ is also an optimal solution. But since the optimum is unique, it must be that $\mathbf{r}(t) = \mathbf{q}$ for all t, i.e. $\mathbf{q}$ is invariant.

It remains to show that $\mathbf{r}(t)$ is feasible for all t. According to the fluid equations,

$$\dot{\mathbf{r}}(t) = \big(\boldsymbol{\lambda} - \boldsymbol{\sigma}(t)\big)^{+[\mathbf{r}(t)=0]}$$

If $\lambda_{i\oplus} = 1$ then

$$\dot{r}_{i\oplus}(t) \geq \lambda_{i\oplus} - \sigma_{i\oplus}(t) = 0$$

and so $r_{i\oplus}(t) \geq r_{i\oplus}(0) = w_{i\cdot}$. Similarly for $r_{\oplus j}(t)$. Also, if $\lambda_{ij} = 0$ then

$$\dot{r}_{ij}(t) \leq 0$$

and by assumption $r_{ij} = 0$; thus $r_{ij}(t) = 0$. Therefore $\mathbf{r}(t)$ is a feasible solution to MWMf-CP(w) for all $t \geq 0$.

Proof of (ii,⇒). Let $\mathbf{q}$ be an invariant state. Consider any fluid model solution $\mathbf{r}(\cdot)$ with $\mathbf{r}(0) = \mathbf{q}$. Since $\mathbf{q}$ is invariant, $dL(\mathbf{r}(t))/dt = 0$. Hence (18) and (19) must be equalities, which implies

$$f(\mathbf{q}) \cdot \boldsymbol{\lambda} = m\big(f(\mathbf{q})\big). \qquad (20)$$

Now let $w = W(\mathbf{q})$ and let $\mathbf{r}$ be any feasible solution to MWMf-CP(w), and suppose $\mathbf{r} \neq \mathbf{q}$. By Lemma 3, we can write

$$\mathbf{r} = \mathbf{r}' + \boldsymbol{\varepsilon} \quad \text{where} \quad \mathbf{r}' = \mathbf{q} + t(\boldsymbol{\lambda} - \boldsymbol{\sigma})$$

for some doubly-stochastic $\boldsymbol{\sigma}$, some $t > 0$, and some $\boldsymbol{\varepsilon} \geq 0$ componentwise; and either $\boldsymbol{\lambda} \neq \boldsymbol{\sigma}$ or $\boldsymbol{\varepsilon} > 0$ in some component. Now consider the family of states

$$\mathbf{s}(u) = \mathbf{q} + u(\boldsymbol{\lambda} - \boldsymbol{\sigma}), \quad u \in [0, t]$$

giving $\mathbf{s}(0) = \mathbf{q}$ and $\mathbf{s}(t) = \mathbf{r}'$. It is the case that

$$\begin{aligned}
\frac{d}{du}L(\mathbf{s}(u))\Big|_{u=0} &= f(\mathbf{q}) \cdot (\boldsymbol{\lambda} - \boldsymbol{\sigma}) \\
&= f(\mathbf{q}) \cdot \boldsymbol{\lambda} - f(\mathbf{q}) \cdot \boldsymbol{\sigma} \\
&= m(f(\mathbf{q})) - f(\mathbf{q}) \cdot \boldsymbol{\sigma} \quad \text{by (20)} \\
&\geq m(f(\mathbf{q})) - m(f(\mathbf{q})) \quad \text{decomposing } \boldsymbol{\sigma} \\
&= 0.
\end{aligned}$$

Now $L(\mathbf{s}(u))$ is strictly convex as a function of u; thus if $\boldsymbol{\lambda} \neq \boldsymbol{\sigma}$ then $L(\mathbf{r}') = L(\mathbf{s}(t)) > L(\mathbf{s}(0)) = L(\mathbf{q})$, and since L is increasing, $L(\mathbf{r}) = L(\mathbf{r}' + \boldsymbol{\varepsilon}) > L(\mathbf{q})$. Otherwise $\boldsymbol{\lambda} = \boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon} > 0$ in some component, so again $L(\mathbf{r}) = L(\mathbf{r}' + \boldsymbol{\varepsilon}) > L(\mathbf{q})$. Either way, we have shown that if $\mathbf{r} \neq \mathbf{q}$ then $m(f(\mathbf{r})) > m(f(\mathbf{q}))$, i.e. that $\mathbf{q}$ solves MWMf-CP($W(\mathbf{q})$).

Proof of (iii,⇒). If $\mathbf{q}$ is an invariant state then any fluid model solution $\mathbf{r}(\cdot)$ starting at $\mathbf{r}(0) = \mathbf{q}$ satisfies $\dot{\mathbf{r}}(t) = 0$; hence $dL(\mathbf{r}(t))/dt = 0$.

Proof of (iii,⇐). If $dL(\mathbf{r}(t))/dt = 0$ then (20) holds and as argued in (ii,⇒) $\mathbf{q}$ solves MWMf-CP($\mathbf{q}$). By (ii,⇐) $\mathbf{q}$ is an invariant state. □

Next we present an alternative characterization of invariant states.

Definition 8 (MWMf-endstate) *A state $\mathbf{q}$ is an MWMf-endstate if*

i. $M(f(\mathbf{q}))_{ij} = 1$ if $\lambda_{ij} > 0$,
ii. $M(f(\mathbf{q}))_{ij} = 1$ if both $\lambda_{i\oplus} < 1$ and $\lambda_{\oplus j} < 1$,

iii. $q_{ij} = 0$ *if both* $\lambda_{i\oplus} < 1$ *and* $\lambda_{\oplus j} < 1$.

Theorem 6 *A state* $\mathbf{q}$ *is an MWMf-endstate if and only if it is an invariant state.*

Proof. From Theorem 5, $\mathbf{q}$ is invariant if and only if $dL(\mathbf{r}(t))/dt = 0$ for any fluid model solution $\mathbf{r}(\cdot)$ with $\mathbf{r}(0) = \mathbf{q}$. Hence, from (18) and (19), $\mathbf{q}$ is invariant if and only if $f(\mathbf{q}) \cdot \boldsymbol{\lambda} = m(f(\mathbf{q}))$. So we will now prove that

$$\mathbf{q} \text{ an MWMf-endstate} \;\Leftrightarrow\; f(\mathbf{q}) \cdot \boldsymbol{\lambda} = m(f(\mathbf{q})). \qquad (21)$$

Proof of (21,⇒). First write

$$f(\mathbf{q}) \cdot \boldsymbol{\lambda} = f(\mathbf{q}) \cdot \boldsymbol{\lambda}^+ - f(\mathbf{q}) \cdot \boldsymbol{\delta}$$

where $\boldsymbol{\lambda}^+$ is an augmentation of $\boldsymbol{\lambda}$. Since $\boldsymbol{\delta}$ is a complementary matrix, $\delta_{ij} > 0$ only if $\lambda_{i\oplus} \vee \lambda_{\oplus j} < 1$; by property (iii) of an MWMf-endstate we see that $f(\mathbf{q}) \cdot \boldsymbol{\delta} = 0$. So it remains to prove that $f(\mathbf{q}) \cdot \boldsymbol{\lambda}^+ = m(f(\mathbf{q}))$.

Since $\boldsymbol{\lambda}^+$ is doubly stochastic it has a decomposition $\boldsymbol{\lambda}^+ = \sum \gamma_{\boldsymbol{\pi}} \boldsymbol{\pi}$ over $\boldsymbol{\pi} \in \mathbb{P}$, with $\gamma_{\boldsymbol{\pi}} \geq 0$ and $\sum \gamma_{\boldsymbol{\pi}} = 1$. Suppose that $\gamma_{\boldsymbol{\pi}} > 0$ for some $\boldsymbol{\pi}$. Then, whenever $\pi_{ij} > 0$, $\lambda^+_{ij} > 0$. There are then two possibilities:

i. either $\lambda^+_{ij} = \lambda_{ij}$, in which case $M(f(\mathbf{q}))_{ij} = 1$ by property (i) of an MWMf-endstate;
ii. or $\lambda^+_{ij} > \lambda_{ij}$, in which case $\delta_{ij} > 0$ and so $M(f(\mathbf{q}))_{ij} = 1$ by property (ii) of an MWMf-endstate.

Either way, $M(f(\mathbf{q}))_{ij} = 1$. By Lemma 2, $\boldsymbol{\pi}$ is a maximum weight matching, i.e. $f(\mathbf{q}) \cdot \boldsymbol{\pi} = m(f(\mathbf{q}))$. Therefore

$$f(\mathbf{q}) \cdot \boldsymbol{\lambda}^+ = \sum_{\boldsymbol{\pi} \in \mathbb{P}} \gamma_{\boldsymbol{\pi}} f(\mathbf{q}) \cdot \boldsymbol{\pi} = m(f(q)).$$

Thus, if q is an MWMf-endstate then $f(q) \cdot \lambda = m(f(q))$.

Proof of (21,⇐). If $\mathbf{q}$ is not an MWMf-endstate then at least one of the three properties of an MWMf-endstate does not hold.

i. If property (i) fails, then $M(f(\mathbf{q}))_{ij} = 0$ and $\lambda_{ij} > 0$ for some i, j. Thus $\lambda^+_{ij} > 0$, and so in the decomposition $\boldsymbol{\lambda}^+ = \sum \gamma_{\boldsymbol{\pi}} \boldsymbol{\pi}$ there must be some $\boldsymbol{\pi} \notin \mathcal{M}(f(\mathbf{q}))$ with $\gamma_{\boldsymbol{\pi}} > 0$. Since this $\boldsymbol{\pi}$ is not a maximum weight matching, $f(q) \cdot \boldsymbol{\pi} < m(f(\mathbf{q}))$ and so
$$f(\mathbf{q}) \cdot \boldsymbol{\lambda} \leq f(\mathbf{q}) \cdot \boldsymbol{\lambda}^+ < m(f(\mathbf{q})).$$
ii. If property (ii) of MWMf-endstate fails, then $M(f(q))_{ij} = 0$ and $\delta_{ij} > 0$ for some i, j. Thus, $\lambda^+_{ij} > 0$ with the same consequences as above.
iii. If property (iii) of MWMf-endstate fails, then $q_{ij} > 0$ and $\delta_{ij} > 0$ for some i, j. Thus $f(\mathbf{q}) \cdot \boldsymbol{\delta} > 0$. Also, by decomposing $\boldsymbol{\lambda}^+$ into permutations, $f(\mathbf{q}) \cdot \boldsymbol{\lambda}^+ \leq m(f(\mathbf{q}))$. Hence
$$f(\mathbf{q}) \cdot \lambda \leq m(f(\mathbf{q})) - f(\mathbf{q}) \cdot \boldsymbol{\delta} < m(f(\mathbf{q})).$$

Thus, if $\mathbf{q}$ is not an MWMf-endstate then $f(\mathbf{q}) \cdot \boldsymbol{\lambda} < m(f(\mathbf{q}))$. □

C. Time to convergence

The last result of this section concerns the speed of convergence. Intuitively, if any fluid model solution converges quickly to an invariant state, then the switch spends most of its time in or close to an invariant state. This will be made rigorous in Section V; for now we simply prove the lemma.

First some definitions. Let $\mathcal{D} = \{\mathbf{q} \in \mathbb{M}_+ : L(\mathbf{q}) \leq L(\mathbf{1})\}$. Note that since L is a Lyapunov function, if $\mathbf{q}(0) \in \mathcal{D}$ then $\mathbf{q}(t) \in \mathcal{D}$ for all $t \geq 0$. Clearly $\mathcal{D}$ is closed and bounded, and hence compact. Now let

$$\mathcal{I} = \big\{\mathbf{q} \in \mathcal{D} : \Delta W(\mathbf{q}) = \mathbf{q}\big\}$$
$$\mathcal{I}_\delta = \big\{\mathbf{r} \in \mathcal{D} : \|\mathbf{r} - \mathbf{q}\| < \delta \text{ for some } \mathbf{q} \in \mathcal{I}\big\}.$$

Since Δ and W are continuous, $\mathcal{I}$ is closed; clearly $\mathcal{I}_\delta$ is open.

We saw in the proof of Theorem 5 that for any fluid model solution

$$\frac{d}{dt} L(\mathbf{q}(t)) = g(\mathbf{q}(t)) \quad \text{where} \quad g(\mathbf{q}) = f(\mathbf{q}) \cdot \lambda - m(f(\mathbf{q})).$$

Since g is continuous, it attains its supremum inside the closed and bounded set $\mathcal{D} \cap \mathcal{I}^{\mathrm{c}}_\delta$; from Theorem 5 this supremum is strictly negative. Let η_δ be the value of the supremum. Finally we can state the result:

Lemma 7 *Given* ε*, and any fluid model solution* $\mathbf{q}(\cdot)$ *with* $\mathbf{q}(0) \in \mathcal{D}$*, let*

$$T_\varepsilon = \inf\big\{t \geq 0 : \|\mathbf{q}(t) - \Delta W(\mathbf{q}(t))\| \leq \varepsilon\big\}.$$

Then there exists some $\delta > 0$ *which does not depend on* $\mathbf{q}(\cdot)$ *such that*

$$T_\varepsilon \leq \frac{L(\mathbf{1})}{|\eta_\delta|}. \qquad (22)$$

Proof. First we will argue that if $\mathbf{q} \in \mathcal{I}_\delta$ then $\|\mathbf{q} - \Delta W(\mathbf{q})\| < \varepsilon$, for δ sufficiently small. Suppose that $\mathbf{q} \in \mathcal{I}_\delta \subset \mathcal{D}$; then $\|\mathbf{q} - \mathbf{r}\| < \delta$ for some $\mathbf{r} \in \mathcal{D}$ such that $\mathbf{r} = \Delta W(\mathbf{r})$. The map $\Delta W(\cdot)$ is continuous, hence it is uniformly continuous on the closed and bounded set $\mathcal{D}$. Hence for any ε there exists a $\delta > 0$ such that

$$\|\mathbf{q} - \mathbf{r}\| < \delta \;\Rightarrow\; \|\Delta W(\mathbf{q}) - \Delta W(\mathbf{r})\| < \varepsilon/2.$$

Hence

$$\begin{aligned}\|\mathbf{q} - \Delta W(\mathbf{q})\| &\leq \|\mathbf{q} - \mathbf{r}\| + \|\mathbf{r} - \Delta W(\mathbf{r})\| + \|\Delta W(\mathbf{r}) - \Delta W(\mathbf{q})\| \\ &\leq \delta + \varepsilon/2.\end{aligned}$$

Simply choose $\delta < \varepsilon/2$; then $\|\mathbf{q} - \Delta W(\mathbf{q})\| < \varepsilon$.

Now all we need to do is to bound the time it takes $\mathbf{q}(\cdot)$ to reach $\mathcal{I}_\delta$. If $\mathbf{q}(0) \in \mathcal{I}_\delta$ then (22) holds trivially. If not, then until $\mathbf{q}(t) \in \mathcal{I}_\delta$,

$$\frac{d}{dt} L(\mathbf{q}(t)) = g(\mathbf{q}(t)) \leq \eta_\delta < 0.$$

Since $L(\mathbf{q}(0)) \leq L(\mathbf{1})$ and $L(\mathbf{q}(t)) \geq 0$, we obtain (22). □

V. Heavy Traffic and State Space Collapse

In Section III we described how the fluid model equations can be used to reason about the throughput of a switch. In this section we will describe what the balanced fluid model equations can tell us about the behaviour of the switch. The conclusion is that the switch spends most of its time at or near an invariant state.

To make this statement precise, we need to introduce the *heavy traffic* limiting regime. Consider a sequence of MWMf switches indexed by r, satisfying Assumptions (2)–(5), where the matrix of mean arrival rates for the rth system is

$$\boldsymbol{\lambda}^r = \boldsymbol{\lambda} - \frac{1}{r}\boldsymbol{\phi}$$

where $\boldsymbol{\phi}$ is a fixed constant matrix in $\mathbb{M}_+$. Assume further that $\boldsymbol{\lambda}$ is such that one or more of the input and/or output ports is critically loaded. Let $\mathcal{X}^r(\cdot)$ be the tuple describing the dynamics of the rth system. In the heavy traffic scaling, we are interested in

$$\hat{x}^r(t) = \frac{1}{r}\mathcal{X}^r(r^2 t).$$

By contrast, in the fluid scaling (Section II-B) we considered $x^r(t) = r^{-1}\mathcal{X}^r(rt)$.

The main result, which is proved in [16, Theorem 16], is the following. It holds under mild probabilistic assumptions on the arrival process. The proof is along the lines of [15].

Theorem 8 *For any finite $T \geq 0$, where $\hat{q}^r(t)$ is the first component of $\hat{x}^r(t)$,*

$$\frac{\left|\hat{q}^r(\cdot) - \Delta W(\hat{q}^r(\cdot))\right|_T}{\left|\hat{q}^r(\cdot)\right|_T \vee 1} \to 0 \quad \textit{as } r \to \infty$$

in probability.

Here, $|\cdot|_T$ is the supremum norm of a function defined on $[0, T]$.

This is called *weak state space collapse*, weak because we do not have control of $|\hat{q}^r(\cdot)|_T$.

Figure 2 illustrates state space collapse for a 3×3 switch running MWM. Each cell shows one of the nine queues; it plots the queue size as a function of time. The horizontal axis runs for 5000 time steps; the vertical axis runs from 0 to 65 packets. Arrivals are Bernoulli, with arrival rate matrix $\boldsymbol{\lambda}$ chosen so that all ports are nearly critically loaded:

$$\boldsymbol{\lambda} = \begin{pmatrix} 0.143 & 0.435 & 0.417 \\ 0.435 & 0.002 & 0.558 \\ 0.417 & 0.558 & 0.020 \end{pmatrix}$$

The lifted queue sizes $\Delta W(Q(t))$ match very closely the actual queue sizes, except for Q_{22} where the arrival rate is so slow that the queue 'can't keep up'. If we had run the simulation for longer, the match would be closer.

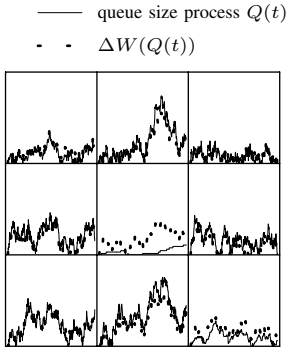

Fig. 2. Evolution of queue sizes in a 3×3 switch running MWM. The actual queue sizes $Q(t)$ are very close to the projected queue sizes $\Delta W(Q(t))$.

VI. Inferring performance via state space collapse

The results of the last section show that, when a $n \times n$ switch running MWMf is heavily loaded, the n^2 queue sizes are essentially determined by the $(2n-1)$-dimensional workload vector, via the lifting map. In other words, to understand the behaviour of the switching algorithm it is sufficient to reason about what it does to the workload vector. Since the workload vector is $(2n-1)$-dimensional rather than n^2-dimensional it can sometimes be simpler to reason about the workload vector.

In fact, state space collapse tells us something more. The queue size matrix $\mathbf{q}$ is usually very close to $\Delta W(\mathbf{q})$ (it is exactly equal to $\Delta W(\mathbf{q})$ in the heavy traffic limit). Therefore the queue size matrix is restricted to the set $\mathcal{I} = \{\mathbf{q} \in \mathbb{M}_+ : \mathbf{q} = \Delta W(\mathbf{q})\}$. Call this the *invariant set*.

Of course, we don't need to keep track of the queue size matrix; it is sufficient to keep track of the workload vector. The workload vector always lies in the set $\mathcal{W} = \{W(\mathbf{q}) : \mathbf{q} \in \mathcal{I}\}$. Call this the *collapsed invariant set*.

The workload vector therefore roams inside $\mathcal{W}$. It cannot leave $\mathcal{W}$; whenever it hits a boundary the scheduling algorithm chooses matchings so that the workload vector remains inside $\mathcal{W}$. Typically this is achieved by idling on some port. For MWM-α, which we study in depth here, $\Delta(w) > \mathbf{0}$ componentwise for w in the interior of $\mathcal{W}$, which means that no queue is empty, and so any matching that the scheduler chooses is work-conserving.

A natural goal is to choose a scheduling algorithm which makes $\mathcal{W}$ as large as possible, so that idling is avoided as much as possible. In the rest of this section we compare scheduling algorithms by reasoning about the geometry of $\mathcal{W}$. The geometry can be complicated, and so in many cases we still only have partial results.

A more rigorous line of argument would be to calculate the stationary distribution of the workload process, or use large deviations theory to estimate the probability of large workloads, under reasonable probabilistic assumptions on the arrival process. This stationary distribution would depend on the scheduling algorithm both through the geometry of $\mathcal{W}$ and

through its behaviour at the boundary of $\mathcal{W}$. This approach has been developed for standard queueing networks [17]; it seems very challenging, and we leave it as a topic for further research.

A. Example: state space collapse of 2×2 MWM-α

Consider a 2×2 switch running MWM-α, and suppose that the arrival rate matrix $\boldsymbol{\lambda}$ is $> \mathbf{0}$ componentwise. To find $\mathcal{I}$, use the characterization of MWM-endstates which says that all matchings must have the same (maximum) weight, i.e.

$$\mathcal{I} = \left\{ \mathbf{q} \in \mathbb{M}_+ : q_{11}^\alpha + q_{22}^\alpha = q_{12}^\alpha + q_{21}^\alpha \right\}.$$

To find $\mathcal{W}$: Suppose $w = W(\mathbf{q})$ for some $\mathbf{q} \in \mathcal{I}$. Solving the equations which correspond to these two conditions (i.e. $q_{11} + q_{12} = w_{1\cdot}$ etc., $q_{11}^\alpha + q_{22}^\alpha = q_{12}^\alpha + q_{21}^\alpha$) we can find $\mathbf{q}$. To be concrete, q_{ij} solves

$$q_{ij}^\alpha + (w_{\cdot\cdot} - w_{i\cdot} - w_{\cdot j})^\alpha = (w_{i\cdot} - q_{ij})^\alpha + (w_{\cdot j} - q_{ij})^\alpha.$$

This is soluble (for $q_{ij} \geq 0$) if and only if

$$w_{\cdot\cdot} \leq w_{i\cdot} + w_{\cdot j} + \left(w_{i\cdot}^\alpha + w_{\cdot j}^\alpha \right)^{1/\alpha}. \tag{23}$$

In other words, $w \in \mathcal{W}$ if and only if (23) is satisfied for all i and j.

Note that the boundaries of $\mathcal{W}$ correspond to regions where (23) is tight, i.e. where $q_{ij} = 0$ for some i and j. In the interior of $\mathcal{W}$, $\mathbf{q} > \mathbf{0}$ componentwise, and so there cannot be any idling. At the boundary of $\mathcal{W}$, some queues are empty and so the scheduling algorithm may choose a matching which results in wasted service.

Now, it is a standard inequality that for any a,$b \in \mathbb{R}_+$, and any $0 < \alpha < \beta$,

$$(a^\alpha + b^\alpha)^{1/\alpha} \geq (a^\beta + b^\beta)^{1/\beta}.$$

Applying this inequality to (23), we see that the collapsed invariant set $\mathcal{W}$ is decreasing as α increases. Note that $\mathcal{W}$ becomes arbitrarily small as α increases, which indicates that delays get arbitrarily bad.

We conjecture that for an $n \times n$ switch running MWM-α, the set $\mathcal{W}$ is decreasing in α (though the above proof only works for $n = 2$). If this is so, we have an explanation for Conjecture 1: when α is larger, the set $\mathcal{W}$ is smaller, so the workload process hits the boundaries more often, so there is more wasted service, so the average queue sizes are larger, so the average delay is bigger.

B. Limiting state space collapse for MWM-α as $\alpha \to 0$

Write $\mathcal{W}(\alpha)$ for the collapsed invariant set for MWM-α. We have just seen that $\mathcal{W}(\alpha)$ is increasing as $\alpha \to 0$ for a 2×2 switch. Does it tend to a limit? Clearly, the collapsed invariant set $\mathcal{W}$ for *any* scheduling algorithm is a subset of

$$\mathcal{W}^{\max} = \mathbb{W} = \left\{ w : w_{\oplus\cdot} = w_{\cdot\oplus} = w_{\cdot\cdot} \right\}$$

i.e. the set of workload vectors where the row and column sums add up to give the same total workload; so $\mathcal{W}(\alpha)$ is certainly constrained. For an $n \times n$ switch, we do not know if $\mathcal{W}(\alpha)$ is increasing as $\alpha \to 0$, but we do have the following partial result:

Lemma 9 *Suppose that the arrival rate matrix $\boldsymbol{\lambda}$ is > 0 componentwise. For any w in the interior of $\mathcal{W}^{max}$, $w \in \mathcal{W}(\alpha)$ for α sufficiently small.*

Proof. Suppose not, i.e. suppose there exist arbitrarily small α with $w \notin \mathcal{W}(\alpha)$. Write Δ^α for the lifting map for MWM-α. Consider a sequence of $\mathbf{q}(\alpha) = \Delta^\alpha(w)$ taken along $\alpha \to 0$ such that $w \notin \mathcal{W}(\alpha)$. We will prove that $\mathbf{q}(\alpha)$ is not an MWMf-endstate, which means that $\mathbf{q}(\alpha) \neq \Delta^\alpha(\mathbf{q}(\alpha))$. But if $\mathbf{q}(\alpha) = \Delta^\alpha(w)$ then from the definition of the lifting map $\mathbf{q}(\alpha) = \Delta^\alpha(\mathbf{q}(\alpha))$, a contradiction. This contradiction falsifies our supposition about w.

We next note some properties of $\mathbf{q} = \Delta(w)$, for any fixed α with $w \notin \mathcal{W}(\alpha)$. Since $w \notin \mathcal{W}(\alpha)$, $W(\mathbf{q}) \neq w$. By the definition of the lifting map (and in particular the requirement of feasibility for the optimization problem), $W(\mathbf{q}) \geq w$. This means that there is some i or j such that $q_{i\oplus} > w_{i\cdot}$ or $q_{\oplus j} > w_{\cdot j}$. 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. Now, $q_{ij} = 0$, since if $q_{ij} > 0$ we could reduce q_{ij} and still have a feasible solution to the optimization problem but with smaller $L(\mathbf{q})$. There must also be a $q_{i'j} > 0$ since if $q_{i'j} = 0$ for all i' then $w_{\cdot j} = 0$, which by assumption is not the case. Similarly there must be some $q_{ij'} > 0$. Furthermore we can bound these away from zero: $q_{i'j} > w_{\cdot j}/n$ and similarly for $q_{ij'}$.

Now return to the sequence $\mathbf{q}(\alpha)$. For each α along this sequence, we can find indices $i(\alpha)$, $j(\alpha)$ etc. as above. Some set of indices (i, j, i', j') must be repeated infinitely often (since there are only finitely many choices). Consider the subsequence of α for which $i(\alpha) = i$ etc. The subsequence $\mathbf{q}(\alpha)$ is bounded (by the remark before the definition of lifting map), and so it has a convergent subsequence. Let $\mathbf{q}^*$ be the limit of the convergent subsequence. By our choice of subsequence, $q_{ij}^* = 0$ and $q_{i'j}^* \geq w_{\cdot j}/n$ and $q_{ij'}^* \geq w_{i\cdot}/n$.

Finally we can return to matchings. By our assumption that $\boldsymbol{\lambda} > 0$ componentwise, using the characterization of MWMf-endstates, all matchings are maximum-weight matchings for $\mathbf{q}(\alpha)$. Let $\boldsymbol{\pi}$ be any matching with $\pi_{ij} = \pi_{i'j'} = 1$, and let $\boldsymbol{\rho}$ be like $\boldsymbol{\pi}$ but with $\rho_{ij'} = \rho_{i'j} = 1$ and $\rho_{ij} = \rho_{i'j'} = 0$. Thus $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ differ simply by a transposition. The difference in weight is

$$\boldsymbol{\rho} \cdot \mathbf{q}(\alpha) - \boldsymbol{\pi} \cdot \mathbf{q}(\alpha) = q_{i'j}(\alpha)^\alpha + q_{ij'}(\alpha)^\alpha - q_{i'j'}(\alpha)^\alpha. \tag{24}$$

Recall that along the subsequence we have chosen $q_{i'j}(\alpha) \to q_{i'j}^*$ etc. Since these limits exist, the limit of (24) as $\alpha \to 0$ is strictly positive. This means that not all matchings have the same weight. Therefore $\mathbf{q}(\alpha)$ is not an MWMf-endstate, for α sufficiently small. □

C. Non-idling in the interior of $\mathcal{W}$

As we discussed earlier, for w in the interior of $\mathcal{W}$ all the VOQs are non-empty and so the switch is work-conserving. In the notation of the last section,

Lemma 10 *Suppose $\boldsymbol{\lambda} > \mathbf{0}$ componentwise. For w in the interior of $\mathcal{W}(\alpha)$, $\Delta^\alpha(w) > \mathbf{0}$ componentwise.*

Proof. Let $\mathbf{q} = \Delta^\alpha(w)$. Consider any 2×2 submatrix

$$\begin{pmatrix} q_{ij} & q_{ij'} \\ q_{i'j} & q_{i'j'} \end{pmatrix}$$

Since $\mathbf{q}$ is an MWM-endstate, and $\boldsymbol{\lambda} > 0$ componentwise, both matchings of this 2×2 submatrix have the same weight.

Suppose $q_{ij} = 0$. Since w is in the interior of $\mathcal{W}(\alpha) \subset \mathcal{W}^{\max}$, $w > 0$ componentwise, which means we can choose i' and j' such that $q_{ij'} > 0$ and $q_{i'j} > 0$. Thus we can choose this 2×2 submatrix so that its row and column workloads are strictly positive.

From our calculations for the 2×2 switch with MWM-α, we know that the optimal configuration of queue sizes (i.e. the configuration that minimizes $L(\mathbf{q})$) has each of these four queues non-empty. This contradicts our assumption that $q_{ij} = 0$. We conclude that all queues are non-empty. □

D. An optimal scheduling algorithm

We saw in the previous section that the collapsed invariant set $\mathcal{W}$ for MWM-α becomes as large as it can be as $\alpha \to 0$ (and is smaller than it need be when α is large—hence MWM is not optimal). We have explained why it is desirable to make $\mathcal{W}$ as large as possible: it leads to less wasted service. It is natural to wonder if there is a single scheduling algorithm which achieves the maximum possible $\mathcal{W}$, without having to take a limit.

A sensible guess would be to take the formal limit of MWM-α as $\alpha \to 0$, in the following sense. MWM-α chooses a matching $\boldsymbol{\pi}$ which maximizes $\sum_{i,j} \pi_{ij} q_{ij}^\alpha$. Now, as $\alpha \to 0$,

$$x^\alpha \approx \begin{cases} 1 + \alpha \log x & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

So the weight of matching $\boldsymbol{\pi}$ is roughly

$$\sum_{i,j} \pi_{ij} 1_{q_{ij}>0} + \alpha \sum_{i,j:q_{ij}>0} \pi_{ij} \log q_{ij}.$$

This suggests the formal limit algorithm:

Definition 9 (MWM-0$^+$ algorithm) *At each timeslot, consider all matchings which have maximal size, i.e. all matchings $\boldsymbol{\pi}$ such that $\sum_{i,j} \pi_{ij} 1_{q_{ij}>0}$ is maximal. Among these choose one which has maximum weight, with weight function* log. *Break ties arbitrarily.*

Interestingly, this is very similar to the Longest Port First algorithm proposed by [18].

We have not been able to obtain useful fluid model equations for MWM-0$^+$, and so we have not been able to find the collapsed invariant set. The difficulty is the discontinuity at $q_{ij} = 0$. We leave this as a topic for further work.

E. Example: MWMw

As a further illustration of the general technique we have described, we now consider a different scheduling algorithm, weighted maximum weight matching, or MWMw. Think of it as a way of understanding the impact of giving priority to some virtual output queue.

The algorithm (which was pointed out to us by Dan C. O'Neill) is as follows. Let $\mathbf{w} \in \mathbb{M}_+$ be a weight matrix with $\mathbf{w} > \mathbf{0}$ componentwise. Suppose the queue size matrix is $\mathbf{Q}$. Then MWMw chooses a matching $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi} \cdot \mathbf{wQ} = \operatorname*{argmax}_{\boldsymbol{\rho} \in \mathbb{P}} \boldsymbol{\rho} \cdot \mathbf{wQ}.$$

The natural fluid model equation is

$$\dot{s}_{\boldsymbol{\pi}}(t) = 0 \quad \text{if} \quad \boldsymbol{\pi} \cdot \mathbf{wq}(t) < \max_{\boldsymbol{\rho} \in \mathbb{P}} \boldsymbol{\rho} \cdot \mathbf{wq}(t).$$

The next step is to find the invariant states. Using similar arguments to those in Section IV, a suitable Lyapunov function is $L^w(\mathbf{q}) = \mathbf{wq}^2 \cdot \mathbf{1}$, the lifting map Δ is like that for MWMf but with this new Lyapunov function L^w, and $\mathbf{q}$ is an invariant state if and only if all matchings $\boldsymbol{\pi}$ yield the same value of $\mathbf{wq} \cdot \boldsymbol{\pi}$.

To begin to understand the implications for performance, consider the simple case of a 2×2 switch where $w_{ij} = 1$ except for $w_{11} = \omega$. As ω increases, greater priority is given to queue q_{11}. How does this impact the configuration of queue sizes?

First calculate $\mathcal{I}$. Suppose $\mathbf{q} \in \mathcal{I}$, i.e. that $\mathbf{q}$ is invariant, so that both matchings have the same $\mathbf{w}$-weighted weight, i.e. $\omega q_{11} + q_{22} = q_{12} + q_{21}$. Let $w = W(\mathbf{q})$, i.e. $q_{1\oplus} = w_{1\cdot}$ etc. Solving these together,

$$q_{11} = \frac{2(w_{1\cdot} + w_{\cdot 1}) - w_{\cdot\cdot}}{3 + \omega}$$
$$q_{12} = w_{1\cdot} - q_{11} \quad \text{etc.}$$

The space $\mathcal{W}$ is the space of all w such that the above is a proper solution, i.e. such that $\mathbf{q} \geq \mathbf{0}$. The space $\mathcal{I}$ is the space of all proper solutions $\mathbf{q}$.

As ω increases, q_{11} decreases (for fixed w). This is as we expect: the higher the priority given to q_{11}, the smaller the queue size. The space $\mathcal{W}$ also changes, although not in a simple monotonic way. As $\omega \to \infty$, the space $\mathcal{W}$ converges to

$$\mathcal{W} = \left\{ w \in \mathcal{W}^{\max} : w_{\cdot\cdot}/2 \leq w_{1\cdot} + w_{\cdot 1} \leq w_{\cdot\cdot} \right\}$$

Since $\mathcal{W}$ converges to a non-empty space, we know that giving absolute priority to q_{11} will not make the other queue sizes arbitrarily big. (Compare this to MWM-α, which gives arbitrarily bad delays as $\alpha \to \infty$.)

VII. Conclusion

In this paper we have described a new technique for analysing scheduling algorithms for input-queued switches. The technique consists in writing down fluid model equations, then analysing the invariant states of those equations. Previous work [10] has used fluid model equations to analyse the

throughput attainable by scheduling algorithms; our technique lets us reason about the queue size distribution.

The basic idea is that, when one or more ports of the switch is heavily loaded, the switch spends most of its time in or near the invariant states. (This can be made formal in the heavy traffic limit). By analysing the geometry of the set of invariant states, we can make inferences about the performance of scheduling algorithms. This has allowed us to explain the conjecture raised in [2] about delay performance of MWM-α as $\alpha \to 0$. It has also led us to conjecture an optimal scheduling algorithm.

We believe that this technique is quite general. In particular, we believe that it can be extended to a large class of scheduling problems where 'MWM-type' algorithms have 100% throughput, e.g. radio-hop networks [9].

There are several lines of further work which would be useful to develop the technique. We have not yet been able to find useful fluid model equations for scheduling algorithms with discontinuities, such as our conjectured optimal algorithm. It would be desirable to be able to calculate the stationary distribution of queue size in the heavy traffic limit, along the lines of [17], in order to make more rigorous our analysis of performance. Finally, it would be fascinating to apply the technique to a wide range of other interesting scheduling algorithms.

VIII. Acknowledgements

We would like to thank Mike Harrison, Frank Kelly, Balaji Prabhakar and Ruth Williams for helpful discussions. DJW is supported by a Royal Society university research fellowship.

References

[1] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proceedings of IEEE Infocom*, 1996, pp. 296–302.

[2] I. Keslassy and N. McKeown, "Analysis of scheduling algorithms that provide 100% throughput in input-queued switches," in *Proceedings of Allerton Conference on Communication, Control and Computing*, 2001.

[3] D. Shah, "Stable algorithms for input queued switches," in *Proceedings of Allerton Conference on Communication, Control and Computing*, 2001. [Online]. Available: http://www.stanford.edu/~devavrat/ilqf.ps

[4] D. Shah and M. Kopikare, "Delay bounds for the approximate Maximum Weight matching algorithm for input queued switches," in *Proceedings of IEEE Infocom*, 2002.

[5] M. Karol, M. Hluchyj, and S. Morgan, "Input versus output queueing on a space division packet switch," *IEEE Transactions on Communications*, vol. 35, no. 12, pp. 1347–1356, 1987.

[6] Y. Tamir and H. Chi, "Symmetric crossbar arbiters for vlsi communication switches," *IEEE Transaction on Parallel and Distributed Systems*, vol. 4, no. 1, pp. 13–27, 1993.

[7] T. Anderson, S. Owicki, J. Saxe, and C. Thacker, "High speed switch scheduling for local area networks," *ACM Transactions on Computer Systems*, vol. 11, pp. 319–351, 1993.

[8] M. Karol, K. Eng, and H. Obara, "Improving the performance of input-queued atm packet switch," in *IEEE INFOCOM*, 1992, pp. 110–115.

[9] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, pp. 1936–1948, 1992.

[10] J. Dai and B. Prabhakar, "The throughput of switches with and without speed-up," in *Proceedings of IEEE Infocom*, 2000, pp. 556–564.

[11] M. A. Marsan, P. Giaccone, E. Leonardi, and F. Neri, "On the stability of local scheduling policies in networks of packet switches with input queues," *IEEE Journal on Selected Areas in Communications "High-performance electronic switches/routers for high-speed internet"*, vol. 21, no. 4, pp. 642–655, 2003.

[12] A. L. Stolyar, "MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, 2004.

[13] N. McKeown, "iSLIP: a scheduling algorithm for input-queued switches," *IEEE Transaction on Networking*, vol. 7, no. 2, pp. 188–201, 1999.

[14] P. Giaccone, B. Prabhakar, and D. Shah, "Randomized scheduling algorithms for high-aggregate bandwidth switches," *IEEE Journal on Selected Areas in Communications High-performance electronic switches/routers for high-speed internet*, vol. 21, no. 4, pp. 546–559, 2003.

[15] M. Bramson, "State space collapse with application to heavy traffic limits for multiclass queueing networks," *Queueing Systems*, vol. 30, pp. 89–148, 1998.

[16] D. Shah, "Randomization and heavy traffic theory: New approaches to the design and analysis of switch algorithms," Ph.D. dissertation, Computer Science, Stanford University, October, 2004.

[17] R. Williams, "Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse," *Queueing Systems*, vol. 30, pp. 27–88, 1998.

[18] A. Mekkittikul and N. McKeown, "A practical scheduling algorithm to achieve 100% throughput in input-queued switches," in *IEEE INFOCOM*, 1998, pp. 792–799.