Maximum Weight Matching with Hysteresis in Overloaded Queues with Setups

Carri W. Chan
Division of Decision, Risk and Operations, Columbia Business School cwchan@columbia.edu, tel.: 212-854-1609, fax: 212-316-9180

Mor Armony
Stern School of Business, New York University marmony@stern.nyu.edu

Nicholas Bambos
Departments of Electrical Engineering and Management Science & Engineering, Stanford University bambos@stanford.edu

We consider a system of parallel queues where arriving service tasks are buffered, according to type. Available service resources are dynamically configured and allocated to the queues to process the tasks. At each point in time, a scheduler chooses a service configuration across the queues, in response to queue backlogs. Switching from one service configuration to another incurs a setup time, during which idling occurs and service bandwidth is lost. Such setup times are inherent in manufacturing and computer systems. Frequent switchings can significantly compromise the service capacity of such systems.

A Maximum Weight Matching (MWM) scheduler, which is known to maximize throughput in the absence of setups, can easily go unstable with setups, even under low load. To remedy this problem, we propose a new MWM-H scheduler which utilizes a controller introduced *Hysteresis* and achieves maximum throughput even with setups, without requiring knowledge of arrival rates and average traffic loads.

During prolonged traffic bursts, the queues may become overloaded and the issue becomes how to reasonably distribute the growing backlog under MWM-H. It is shown that by appropriately selecting the MWM-H parameters, one can control the backlog amongst the individual queues in order to achieve a desired balance.

Key words: Hysteresis, Maximum Weight Matching, Setup times

1. Introduction

Various services conform to the following operational paradigm: arriving service requests are placed in multiple queues and a scheduler, observing the queue backlogs, allocates service resources to the queues for executing the tasks. However, moving resources between queues requires setup times, during which processing capacity is lost. At the same time, traffic bursts may also drive the queues to (temporary) overload, which makes this loss of throughput even more critical. In this paper, we introduce a hysteresis mechanism for judiciously reconfiguring resources across the queues to mitigate lost bandwidth during setup times and maximize the throughput, while controlling queue overloads.

One particularly important example of such systems are modern information services that range from e-commerce (e.g. Amazon, Expedia) and e-trading (e.g. eTrade, Fidelity) to content delivery and media streaming (e.g. Netflix, YouTube). They operate using data centers accessed over networks. They may scale to tens of millions of users, who interact with Web servers , launching cascades of service (computation and communication) tasks. At various levels of modeling abstraction, tasks are buffered in (virtual) queues of various types awaiting service, while processing cycles and communication slots are dynamically allocated to process the tasks in each queue. A key complicating factor in these settings is that reconfiguration of processing communication resources requires some service idling time, for example, for accessing memory to store the previous context/state and retrieve new ones, update databases, disengage resources from previous task types and move them over to new ones, etc. During such setup times service bandwidth can be lost; therefore, resource switching and reconfiguration has to be exercised judiciously and as infrequently as possible.

Indeed, systems with setup times are quite common in computer, manufacturing, and service systems. In data centers, servers are often turned off or reallocated in order to save energy consumption (Gandhi et al. 2012, Gandhi and Harchol-Balter 2013, Gandhi et al. 2013). It can take minutes to turn servers back on, while each job often only requires a few seconds (DeCandia et al. 2007), making the decision of whether to turn servers on or off an important one. In manufacturing environments, it can take time to warm up machines as the product type on the line is changed. In service systems, bringing in additional staff or alternatively trained staff to address different customer classes can take time as the human servers may require transportation and psychological/cognitive adjustment time. For example, in call centers, handling in-bound versus out-bound calls requires a different skill set and mental aptitude. Similarly, a hospital physician may take time adjusting to treating existing ED patients, after treating a multi-trauma case. In this work, we consider how to allocate resources amongst queues in a system with such setup times.

Another complication is that prolonged bursts of service requests can temporarily overload the service resources and drive the queues through a temporary instability phase, until there is a chance to clear the backlogs and have overload subside. The issue then is how to distribute the resources during these (temporary) overload periods.

In this paper, we consider a system of parallel queues, where arriving service tasks/requests are queued up, according to type. In response to queue backlogs, a scheduler dynamically chooses an allocation of service resources to the queues for processing their tasks. There are setup times to switch from one service allocation/configuration to another, during which the queues idle and service bandwidth is lost. Hence, frequent reconfigurations are inefficient. In order to mitigate the effect of the setup times, we propose a scheduler which introduces a hysteresis into the system. During prolonged traffic bursts the scheduler tries to balance the backlogs of the overloaded queues, by achieving a desirable, chosen ratio between the rising backlogs. This is described in Section 2 in detail.

1.1. Related Research

In the absence of setup times for service reconfigurations, a well investigated class of scheduling policies is Maximum Weight Matching (MWM), which has been studied by several authors, including Tassiulas and Ephremides (1992), Stolyar (2004), Armony and Bambos (2003), Ross and Bambos (2009), etc. and more recently, in overloaded systems, by Shah and Wischik (2011) (discussed below).

This paper focuses on systems with setup times, which are inherent in various computer and manufacturing systems. In the presence of setup times, the MWM scheduler could actually drive the system unstable even under negligible traffic loads, due to frequent service reconfigurations inducing service idling. To remedy this situation we introduce a MWM scheduler with *Hysteresis* (MWM-H). Intuitively speaking, our scheme, with controller introduced hysteresis, purposely delays the switching from one service configuration to another until it “makes sure” that this is clearly needed. This is explained more precisely later.

Service systems with setup times have been studied by a number of authors with emphasis on stability and cost optimization. Bertsimas and Nino-Mora (1999a,b) studied queueing networks with multiple servers and Poisson job flows following Markovian routes through various queues, receiving exponential service at each one (general i.i.d service times in the case of a single server). When a server switches from serving one queue to another, it incurs a random (general i.i.d.) changeover (setup) time, whose distribution depends on the initial and final queues. Lan and Olsen (2006) studied a system of multiple queues and a single server, where there is a fixed cost and a random setup time when a server switches from serving one queue to another. There is also backlog cost and the overall scheduling objective is to minimize the long-run average cost. Takagi (1997) provides a nice review of the study of setup times in the context of polling systems.

Armony and Bambos (2003) have taken an “adaptive” batching approach to maximize throughput under setup times for the system considered here. Each batch of jobs is formed by grouping together the jobs arriving while the previous batch is processed; hence, the next batch is “adapted” to the previous one. Scheduling job service in batches lowers the frequency of service setups and leads to maximizing throughput, even when the average load is unknown (because of the adaptive nature of the process).

A different approach has been taken by Dai and Jennings (2004). They consider several servers, each serving a subset of buffers. When a server completes (non-preemptive) processing of a job, the latter can be forwarded to a buffer served by the same or another server. At each point in time, the problem is how to allocated each server to one of the buffers it serves, given that there is a setup time to switch it from one buffer to another. The approach proposed in Dai and Jennings (2004) is to have each server “stick” to the buffer currently being served and do a “service run” on several jobs in that buffer, before being allowed to switch to a different buffer and incur a setup time. How many jobs are included in each service run depends on the average traffic load of the system. The higher the load, the longer the service runs, so that the lost service bandwidth due to setup times is amortized across more jobs. This adds a small “virtual inflation” of service times, which gets smaller as the run length increases. The paper demonstrates that this policy achieves maximum throughput of the system, the same as if there were no setup times; the latter only induces higher backlogs, but does not compromise throughput.

There are two key differences between the policy in Dai and Jennings (2004) and ours: 1) While their policy sticks to the same service configuration (servers assigned to buffers) for a fixed number of jobs before considering a switch, our hysteresis-based policy examines the system state continuously and may initiate a switch at any time. 2) to determine this fixed length of the run, their policy requires the knowledge of traffic parameters such as the average load, while our hysteretic scheme *adapts* to the system state without requiring any such knowledge. Both of these qualities make our scheme robust to changes in traffic characteristics, and requires no forecast or learning of the relevant parameters.

Hysteresis typically refers to a relationship between inputs and outputs that have *memory*, which manifest itself through delayed impacts. Hysteresis has its roots in physics, where, for example, phenomenon due to magnetic fields introduce hysteresis. See Hassani et al. (2014) for a survey of these results. Hysteresis has found its way into other applications, such as image processing (Medina-Carnicer et al. 2010, Xu et al. 2011) and control (Chen et al. 2008, Morse et al. 1992). Hysteresis has also been considered in queueing contexts. Kelly (1986, 1991) found that the blocking probability in a loss network exhibits hysteretic behavior. Dshalalow (1998), Dikong and Dshalalow (1999) consider a two threshold hysteretic control for queues where the server comes and goes depending on the number in system. Lu and Serfozo (1984), Plum (1991) demonstrate that, for an M/M/1 queue whose arrival and service rates can be changed, a control policy with hysteresis is optimal when there are switching costs. In a multi-server setting, Ibe and Keilson (1995), Golubchik and Lui (1997) look at how hysteresis thresholds can be used to determine the number of servers being used. In all of these works, the hysteresis is given by fixed thresholds. In contrast, our hysteresis thresholds depend on the system state; in particular, the magnitude of the threshold increases with the number of jobs in system. In fact, because we examine heavily-loaded systems, fixed hysteresis thresholds quickly become ineffective; thus, we introduce a hysteresis function which specifies thresholds based on the system state. To the best of our knowledge, our work is the first to consider the use of a state-dependent hysteresis in control of queues.

Contrary to the aforementioned lines of research, our focus here is not simply on system stability under MWM with setups, but also on how to distribute the growing backlog in a desirable manner when the system is overloaded beyond stability. Given that MWM *with setups* can easily become unstable, even under low traffic load, we first need to prove that the proposed hysteresis scheme MWM-H does maximize the system throughput and achieves the same stability region, as with no setup times. In Section 2 below we describe the queueing system in detail and in Section 2.1 we define the MWM-H policy and show its essential stability properties.

In the absence of setup times, various queueing systems in overload have also been investigated. Egorova et al. (2007) studied bandwidth-sharing networks of work flows, utilizing α-fair schedules (instead of MWM) introduced by Mo and Walrand (2000) to allocate service bandwidth, and have characterized the backlog growth rates. Neely et al. (2008) consider networks of queues (in the context of communication networks), where the set of available service vectors changes over time.They consider distributed fair schedules under both network stability and overload, leveraging flow utilities (see also references therein). More recently, Shah and Wischik (2011) have studied the behavior of MWM in overload, considering a fluid model for a collection of queues, where work (fluid) on one queue may be forwarded after service to another queue (but can never come back to the original one) and there is no splitting and routing of the work to various queues. The network is analyzed under both MWM schedules and α-fair ones in overload and the backlog growth rates are characterized. In this paper, after showing that MWM in a system without setups and our proposed policy (MWM-H) in the presence of setups have identical fluid limits, we leverage the results of Shah and Wischik (2011) to establish our limiting result. Thus, all of our control results apply to both MWM and MWM-H. Note that these results pertaining to control via the MWM matrix were not considered in Shah and Wischik (2011). Perry and Whitt (2009, 2011) have also investigated the behavior of queueing systems in overload , albeit in a specialized network topology (an X network).

1.2. Contributions

In summary, the contributions of this paper are as follows. First, we address the behavior of the standard MWM scheme in the presence of setup times, which are common in many practical systems. To remedy the inherent instability of standard MWM with setups even under low load, we introduce a novel MWM scheme with *Hysteresis* (MWM-H). It is shown that MWM-H *with setups* achieves the same stability (and instability) region as the standard MWM with *no setups*. In contrast to previous literature on hysteresis in queues, the hysteresis in our proposed MWM-H is *state-dependent*, and does not use information on average traffic load. It should be noted that most previous schemes that handle control of queues with setups heavily rely on knowing that information.

Next, we focus on how MWM-H handles overloads, especially how to desirably distribute the backlog according to target proportions or various cost minimization objectives. We show that by controlling the weights of MWM-H one can achieve desirable target ratios for the backlog growth across the queues. Additionally, controlling the MWM-H weights, one can opt to maximize revenue (when job service generates revenue), minimize aggregate backlog, maximize aggregate service rate, etc.

2. The Queueing Model and the MWM-H Policy

We consider a queueing system of Q parallel queues, indexed by $q \in \mathcal{Q} = \{1, 2, \ldots, Q\}$. Time is slotted and indexed by $t \in \mathbb{Z}_0^+$. Jobs arrive to the system, adding service requirement[1] (work to be performed) to one or more queues. Let $t_j \in \mathbb{Z}_0^+$ be the arrival time of the j^{th} job to the system and $\sigma_j \in \mathbb{R}_0^Q$ its (vector) service requirement, where $(\sigma_j)_q \in \mathbb{R}_0$ is the workload added to queue $q \in \mathcal{Q}$ by that job. We assume that $0 \leq (\sigma_j)_q \leq \bar{\sigma}_q$ for each $q \in \mathcal{Q}$, where $\bar{\sigma}_q$ is an arbitrarily large fixed ceiling. At most one job arrives in each time-slot, adding some positive workload to at least one queue. Let

$$R(s,t) = \sum_j \sigma_j 1_{\{t_j \in (s,t]\}} \tag{2.1}$$

as the (vector) workload arriving in the system between time slot s and t, where $R_q(s,t)$ is the workload arriving to queue $q \in \mathcal{Q}$. We assume that the long-term average traffic load to each queue $q \in \mathcal{Q}$,

$$\lim_{t \to \infty} \frac{R_q(0,t)}{t} = \rho_q \in (0, \infty), \tag{2.2}$$

is well-defined, positive, and finite. The traffic load vector is $\rho = (\rho_1, \rho_2, \ldots, \rho_q, \ldots, \rho_Q)$.

In each time slot, a service vector[2] $S = (S_1, S_2, \ldots, S_q, \ldots, S_Q) \in \mathcal{S}$ can be used, chosen from a family $\mathcal{S} = \{S^1, S^2, \ldots, S^n, \ldots, S^N\}$ of N feasible ones. When $S = (S_1, S_2, \ldots, S_q, \ldots, S_Q) \in \mathcal{S}$ is used in a time slot, $S_q \in \mathbb{R}_0^+$ amount of workload is removed from queue $q \in \mathcal{Q}$, assuming there is enough workload in the queue (as explained momentarily). Let $S(t)$ denote the service vector used in time slot t. We define $S(t) = 0$ if *no* service vector is in use in time slot t.

Let $W_q(t)$ be the workload in queue $q \in \mathcal{Q}$ at time t, hence, corresponding workload vector is $W(t) = (W_1(t), W_2(t), \ldots, W_q(t), \ldots, W_Q(t)) \in \mathbb{R}_0^Q$. When the workload vector is $W(t)$ and the service vector $S(t)$ is used, the amount of work which is served and removed from queue q is simply

$$D_q(t) = \min\{S_q(t), W_q(t)\} \tag{2.3}$$

where the minimum accounts for the fact that work can only be serviced if it is already waiting in the queue. Hence, if $W_q(t) < S_q(t)$, there is some idle service provided by service vector $S(t)$ due to the lack of enough workload to be processed.

Let $M_q(s,t)$ denote the amount of work that has been executed during time $[s,t)$ at queue $q \in \mathcal{Q}$, that is,

$$M_q(s,t) = \sum_{\tau=s}^{t-1} D_q(\tau) \tag{2.4}$$

[1] The terms service requirement and work, as well as workload and backlog, are used interchangeably.

[2] We use the terms service vector, service configuration and service mode interchangeably.

and $M(s,t)$ the corresponding vector across all queues. The workload vector then evolves as

$$W(t) = W(0) + R(0,t) - M(0,t). \tag{2.5}$$

There are service setup[3] (switching) times; that is, in order to switch from service vector S_i to service vector S_j a switching time of T time slots is incurred, during which no service is provided to the queues. At any time t, let $U(t)$ denote the remaining time until the new service vector can be activated and let $V(t) \in \mathcal{S}$ be the service vector the system is in the process of switching to. Thus, if $S(t) = V(t)$, then the service vector has already been switched; if $S(t) \neq V(t)$, then the system is currently switching to a new service vector. Suppose that, starting at time t, the service vector is to be changed from S_i to S_j. Then the system evolves as follows:

$$\begin{array}{rl} S(t-1) = S_i,\ S(t') = 0 & \text{for } t' \in \{t, t+1, t+2, ..., t+T-1\},\ \text{and } S(t+T) = S_j \\ V(t-1) = S_i,\ V(t') = S_j & \text{for } t' \in \{t, t+1, t+2, ..., t+T-1, t+T\} \\ U(t') = t+T-t' & \text{for } t' \in \{t, t+1, t+2, ..., t+T-1, t+T\} \end{array} \tag{2.6}$$

When a service vector switches from S_i to S_j is initiated, it has to complete before a different switch can commence. Finally, let $Y(s,t)$ denote the cumulative time the system spends idling in $[s, t-1)$ because of switching,

$$Y(s,t) = \sum_{\tau=s}^{t-1} 1_{\{U(\tau)>0\}} \tag{2.7}$$

that is, the amount of time during the period $[s,t)$ such that $S = 0$ due to the reconfiguring of service vectors.

2.1. The Maximum Weight Matching Policy with Hysteresis (MWM-H)

As mentioned before, it is known that - in the absence of setup times - the Maximum Weight Matching (MWM) scheduler maximizes the throughput (see, e.g. Armony and Bambos (2003)). MWM chooses a service vector

$$S^*(W) \in \mathcal{S}^*(W) = \arg\max_{S' \in \mathcal{S}} \langle W, \mathbf{\Delta} S' \rangle = \arg\max_{S' \in \mathcal{S}} \left\{ \sum_{q=1}^{Q} W_q \delta_q S'_q \right\} \tag{2.8}$$

when the workload is W, given any fixed positive-diagonal matrix $\mathbf{\Delta} = \text{diag}\{\delta_1, \delta_2, ...\delta_q, ...\delta_Q\}$ with $\delta_q > 0$ for all $q \in \mathcal{Q}$. Thus, when the backlog is W, the MWM scheduler chooses service vectors S' whose $\mathbf{\Delta} S'$ has maximal projection on W. Note that there may be more than one such service vector, hence, the set $\mathcal{S}^*(W)$ may have more than one element; in that case, one such service vector $S^*(W)$ is chosen arbitrarily amongst the ones available in $\mathcal{S}^*(W)$.

However, it is easy to see that, in the presence of setup times, the throughput of MWM can be driven to 0 by an adequately “adversarial” arrival pattern $\{(t_j, \sigma_j), \in \mathbb{Z}_0^+\}$, even when the its average traffic load ρ is very small. For example, a “bad” arrival pattern (see Example 1) could cause $\{W(t), t \in \mathbb{R}_0^+\}$ to constantly

[3] The terms setup time, setup, switching time and reconfiguration time are used interchangeably in general.

oscillate between backlog values requiring distinct service vectors under MWM. That would simply cause the system to "freeze" and spend all its time in setup mode, switching between these two service vectors. Thus, even with very low arrival traffic load the backlog would explode.

Example 1 *Consider a system with 2 service vectors: $S_1 = [5,0]^T, S_2 = [0,5]^T$. The switching time required to setup each service vectors is $T = 5$ time slots. Suppose the system load is $\rho = [2,2]^T$, which is certainly within the stability region (for example, the system is stable in the absence of setup times if at each time slot one randomly chooses S_1 or S_2 with equal probabilities). The initial state is $X(0) = [0,0]^T$ and $S(0) = S_2$.*

Consider a sample path where arrivals alternate between queue 1 and queue 2. In particular, the arrival pattern is:

$$\begin{bmatrix} 4\;0\;4\;0\;4\;0\;\ldots \\ 0\;4\;0\;4\;0\;4\;\ldots \end{bmatrix}$$

A quick numeric examination of this arrival pattern will demonstrate that the backlogs are growing without bound due to the time spent in switching mode, which renders the system idle for 5 time slots at a time.

$$W(t) = \begin{bmatrix} 0 & 0 & 4 & 4 & 8 & 8 & 12 & 12 & 11 & 6 & 10 & 10 & 14 & 14 & 18 & 18 & 22 & 22 & 26 & 26 & 30 & 30 & 34 & \ldots \\ 0 & 0 & 0 & 4 & 4 & 8 & 8 & 12 & 12 & 16 & 16 & 20 & 20 & 24 & 24 & 23 & 18 & 17 & 12 & 16 & 16 & 20 & 20 & \ldots \end{bmatrix}$$

Given this inherent limitation of MWM, we introduce a hysteresis mechanism which appropriately suppresses the frequency of service switchings/setups. The resulting algorithm, MWM with *hysteresis* (MWM-H), is shown to achieve maximal throughput under setup times. We start by defining a **Hysteresis Function** $h(W): \mathbb{R}_{0+}^{Q} \to \mathbb{R}_{0+}$ of the backlog W, with the following properties:

1. $h(W)$ is positive, monotonically increasing in each component $W_q, q \in \mathcal{Q}$, and
2. $h(W) \to \infty$ as $\|W\| \to \infty$ and

$$\lim_{\|W\| \to \infty} \frac{h(W)}{\|W\|} = 0, \tag{2.9}$$

uniformly on compact sets (u.o.c.), where $\|\cdot\|$ is a norm in $\mathbb{R}_{0+}^{Q}$.

Note that the latter property basically forces the hysteresis $h(W)$ to grow sub-linearly with respect to W. A simple example is $h(W) = \sqrt{\sum_{q \in \mathcal{Q}} W_q}$.

We can now define the scheduler **MWM *with Hysteresis* (MWM-H)**. Suppose that in time slot t the system is in operational mode (as opposed to in a setup mode) and is using the service vector $S(t)$. In the next time slot $t+1$, the workload becomes $W(t+1)$ and the system needs to decide whether to either 1) remain operational and keep using the same service vector $S(t+1) = S(t)$ or 2) initiate a switching to a new service vector $S' = S^*\Big(W(t+1)\Big)$, in which case it will have to halt service for the next T consecutive time slots for setup, hence, $S(\tau) = 0$ for $\tau \in \{t+1, t+2, ..., t+T\}$. Specifically, MWM-H scheduler is defined inductively as follows.

1. *When in time slot t the system is in operational mode with service vector $S(t)$, then:*

(a) *if*

$$\Big\langle W(t+1), \mathbf{\Delta} S^*\Big(W(t+1)\Big)\Big\rangle - \Big\langle W(t+1), \mathbf{\Delta} S(t)\Big\rangle \leq h\Big(W(t+1)\Big), \tag{2.10}$$

the service vector at time slot t remains the same $S(t+1) = S(t)$ and no switching is initiated;

(b) *else, if*

$$h\Big(W(t+1)\Big) < \Big\langle W(t+1), \mathbf{\Delta} S^*\Big(W(t+1)\Big)\Big\rangle - \Big\langle W(t+1), \mathbf{\Delta} S(t)\Big\rangle, \tag{2.11}$$

a switch to service vector $S^*\Big(W(t+1)\Big)$ is initiated and the system enters a setup mode for the next T time slots, that is, $S(\tau) = 0$ for $\tau \in \{t+1, t+2, ..., t+T\}$.

2. *When the system enters setup mode in time slot $t+1$ (as per step 1(b) above) it will halt service for T consecutive time slots (hence, $S(\tau) = 0$ for $\tau \in \{t+1, t+2, ..., t+T\}$) and will go back to operational mode again in time slot $t+T+1$ with service vector*

$$S(t+T+1) = S^*\Big(W(t+1)\Big) \tag{2.12}$$

and cycles back to step 1.

3. *By convention, we can initialize the system to starting from operational mode with $S(0) = S^*\Big(W(0)\Big)$, when the initial backlog is $W(0)$.*

Let us now make some observations on the MWM-H scheduler and briefly discuss the intuition behind it. Note first that the service $S(t)$ under MWM-H could be quite different than

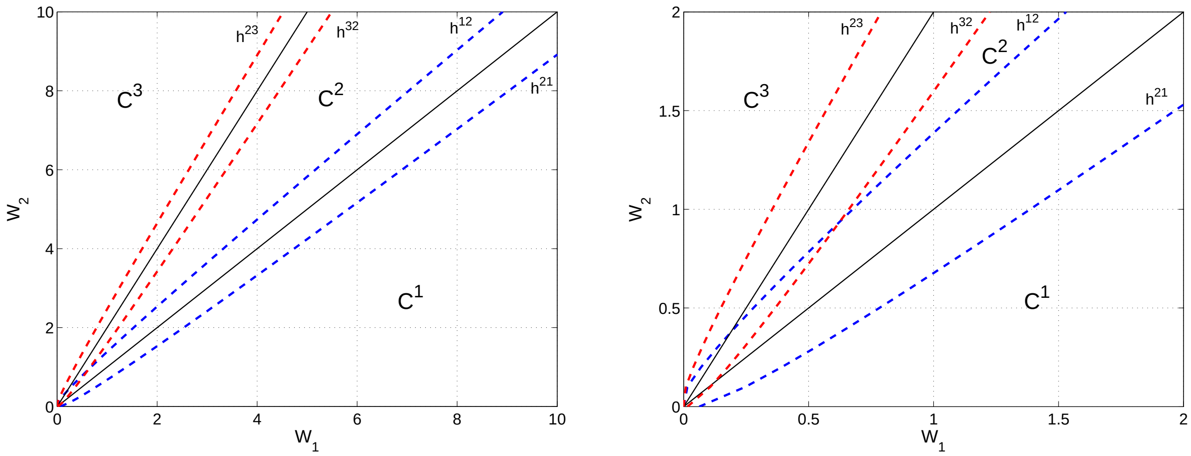
$$\text{any } S^*\Big(W(t)\Big) \in \mathcal{S}^*\Big(W(t)\Big) = \arg\max_{S' \in \mathcal{S}} \Big\langle W(t), \mathbf{\Delta} S'\Big\rangle, \tag{2.13}$$

which would have been the service vector under the original MWM (without hysteresis), as per (2.8). Next, note that the quantity

$$\Phi(t+1) = \Big\langle W(t+1), \mathbf{\Delta} S^*\Big(W(t+1)\Big)\Big\rangle - \Big\langle W(t+1), \mathbf{\Delta} S(t)\Big\rangle \tag{2.14}$$

(used in the definition of MWM-H) is non-negative[4] and measures how "suboptimal" it is to keep using $S(t)$ in time slot $t+1$ in MWM-H compared to $S^*\Big(W(t+1)\Big)$ that MWM would have used. If the sub-optimality gap $\Phi(t+1)$ is less than the threshold $h\Big(W(t+1)\Big)$, then MWM-H has no incentive to switch to $S^*\Big(W(t+1)\Big) \neq S(t)$ and halt service for T slots during setup; therefore, it is content with using the old suboptimal $S(t)$. But if the sub-optimality gap $\Phi(t+1)$ exceeds $h\Big(W(t+1)\Big)$, the MWM-H decides it is worth halting service for T time slots in order to switch to a better service vector. That is, as the backlog drifts away from using $S(t)$ efficiently, MWM-H does not respond immediately (as original MWM would

[4] Because $\Big\langle W(t+1), \mathbf{\Delta} S'\Big\rangle$ is maximized for $S' = S^*\Big(W(t+1)\Big)$ over the set of all service vectors $S' \in \mathcal{S}$



(a) Zoomed-out

(b) Zoomed-in

Figure 1 Standard MWM service cones C^1, C^2, C^3 and their boundaries (straight solid lines) for a simple system of 2 queues and 3 service vectors $S^1 = (4,0), S^2 = (3,1), S^3 = (1,2)$ and $\mathbf{\Delta} = \text{diag}(1,1)$ the identity matrix. The curved dashed lines are the hysteresis boundaries (around the corresponding cone boundaries) under MWM-H with setups, for hysteresis function $h(W) = \sqrt{W_1 + W_2}$. Consistent with (2.9), the hysteresis boundaries grow sub-linearly away from their corresponding cone boundaries as the workload increases.

have done), trying to avoid halting service for T setup slots. Instead, it sticks with $S(t)$ and waits until this drift goes beyond the tipping point $h\Big(W(t+1)\Big)$, before it launches the switching and trades T slots of service inactivity for getting better service at the end. Therefore, MWM-H is *hysteretic* in the sense that it waits to make sure it is worth switching, before launching the process and paying the price. This property suppresses the frequency of service switchings and gets MWM-H with setups to achieve maximal throughput (as shown below) as the raw MWM without setups, where the latter with setups could have collapsed to zero throughput.

An alternate geometric look at the MWM-H operation is demonstrated in Figure 1. Define in general the MWM *service cone* C^n as the set of backlogs W for which the MWM could use the service vector $S^n \in \mathcal{S} = \{S^1, S^2, ..., S^N\}$, that is,

$$C^n = \Big\{ W \in \mathbb{R}^Q_{0+} : S^n \in \mathcal{S}^*(W) = \arg\max_{S' \in \mathcal{S}} \langle W, \mathbf{\Delta} S' \rangle \Big\} \tag{2.15}$$

It is easy to see that these are (linear) cones, since scaling αW up or down by changing the scalar α does not change the service vector selection under MWM. Figure 1 shows (in straight solid lines) the MWM service cones C^1, C^2, C^3 for a system of 2 queues and 3 service vectors $S^1 = [4,0]^T, S^2 = [3,1]^T, S^3 = [1,2]^T$ and $\mathbf{\Delta} = \text{diag}(1,1)$ the identity matrix. The curved dashed lines define the hysteresis boundaries under MWM-H with setups for hysteresis function $h(W) = \sqrt{W_1 + W_2}$.

Under MWM with no setups, when the backlog drifts, for example, from cone C^1 into cone C^2, the service vector immediately switches from S^1 to S^2. But under MWM with setups, if a “bad” pattern of arriving work causes the backlog to oscillate between the two sides of the C^1-C^2 boundary, then the system could lock in setup mode forever, transitioning back and forth between S^1 and S^2 repeatedly. That would halt service forever and cause the queues to explode even under very low traffic load.

In contrast, consider now MWM-H with setups and suppose the workload is in cone C^2 and S^2 is being used. As the workload drifts out of cone C^2 and crosses into cone C^1, MWM-H will keep using S^2 until the workload drifts beyond the hysteresis boundary h^{21}; only when that happens (if ever) will MWM-H initiate a switching to S^1 and halt service for T slots. In general, when operating with S^n and because of hysteresis, the backlog drifting out of cone C^n into cone C^m must *pass* the hysteresis boundary h^{nm} on the other side before triggering a switch to S^m and halting service in the next T slots for setup. Note that, consistent with (2.9), each hysteresis boundary h^{mn} grow sub-linearly away from the C^n- C^m cone boundary as the workload increases.

Remark 2.1 We note that the sublinearity and unboundedness of the hysteresis function, $h(W)$, are essential to the behavior of MWM-H. If the hysteresis grows too quickly with the workload, the system will never switch until the workload gets very large, at which point it is too late. Thus, even a stablizable load will result in unbounded backlogs. On the other hand, if the hysteresis function scales too small with the workload, the system will constantly switch, resulting in ‘freezing’, and instability.

Remark 2.2 As an alternative policy to MWM-H, one could also consider a policy where, once a service vector is being used, the system will not switch to a different service vector for a time that is greater than or equal to the value of a hysteresis evaluated as a function of the backlog. Once the system is ready to switch, the next service vector is selected according to MWM. In analyzing this alternative policy, we find that its performance is very similar to that of MWM-H, both analytically and numerically. Therefore, for the sake of brevity, we omit the details.

2.2. Overview of Main Results

It is known (see, e.g. Armony and Bambos (2003)) that when

$$\rho \in \mathcal{P} = \{\rho \in \mathbb{R}^Q_{0+} : \rho \leq \sum_{S \in \mathcal{S}} \phi_s S, \text{ for some } \phi_S \geq 0, S \in \mathcal{S} \text{ with } \sum_{S \in \mathcal{S}} \phi_s = 1\}$$

the standard MWM without setups is (weakly) stable (i.e. $\lim_{t\to\infty} \frac{W(t)}{t} = 0$). But, as discussed above, this is not the case in the presence of setups.

As shown below, given a fixed matrix $\mathbf{\Delta}$ and load ρ, the time-scaled backlog of the system operating under MWM-H with setups will converge to a limit vector $\eta_\rho(\mathbf{\Delta})$ which depends on $\mathbf{\Delta}$ and ρ, that is:

$$\lim_{t\to\infty}\frac{W(t)}{t}=\eta_\rho(\mathbf{\Delta}). \tag{2.16}$$

For all $\mathbf{\Delta}=\text{diag}\{\delta_1,\delta_2,...,\delta_Q\}\in\mathbb{R}^Q_{0+}$, let $\mathcal{K}_\rho$ be the range of the mapping $\eta_\rho(\mathbf{\Delta}):\ \mathbb{R}^Q_{0+}\to\mathbb{R}^Q_{0+}$, given fixed ρ.

In particular, there are two cases:

1. If $\rho\in\mathcal{P}$, then $\lim_{t\to\infty}\frac{W(t)}{t}=\eta_\rho(\mathbf{\Delta})=0$ for each $\mathbf{\Delta}=\text{diag}\{\delta_1,\delta_2,...,\delta_Q\}\in\mathbb{R}^Q_{0+}$. This establishes that MWM-H with setups achieves the same stability region and maximizes throughput just as the MWM without setups. This is in sharp contrast to MWM with setups which does not achieve maximal throughput and could actually have zero throughput under bad arrival patterns.

2. If $\rho\notin\mathcal{P}$, then $\lim_{t\to\infty}\frac{W(t)}{t}=\eta_\rho(\mathbf{\Delta})\neq 0$ for every $\mathbf{\Delta}=\text{diag}\{\delta_1,\delta_2,...,\delta_Q\}\in\mathbb{R}^Q_{0+}$. By choosing $\mathbf{\Delta}\in\mathbb{R}^Q_{0+}$ we can appropriately position $\eta_\rho(\mathbf{\Delta})\in\mathbb{R}^Q_{0+}$ in its range $\mathcal{K}_\rho$ so as to achieve a target proportion of overload $\frac{(\eta_\rho)_q(\mathbf{\Delta})}{\sum_{q\in\mathcal{Q}}(\eta_\rho)_q(\mathbf{\Delta})}$ for each queue $q\in\mathcal{Q}$ that the system manager considers desirable.

Controlling $\eta_\rho(\mathbf{\Delta})\in\mathcal{K}_\rho$ via $\mathbf{\Delta}\in\mathbb{R}^Q_{0+}$ for an overloaded MWM-H system with setups has some easy, yet interesting, implications. Some direct examples are given below:

(a) Note first that $\rho-\eta_\rho(\mathbf{\Delta})$ is the rate at which backlog is actually served by the overloaded system. Hence, manipulating $\eta_\rho(\mathbf{\Delta})$ via $\mathbf{\Delta}$ results in controlling the actual processing rate $\rho-\eta_\rho(\mathbf{\Delta})$, as desired.

(b) Another optimization consideration is to minimize the quadratic cost $\left\langle\left(\frac{W(t)}{t}\right),\mathbf{B}\left(\frac{W(t)}{t}\right)\right\rangle$ as $t\to\infty$ for the overloaded system under MWM-H with setups, where $\mathbf{B}$ is a positive diagonal matrix. That is, minimize the cost $\langle\eta_\rho(\mathbf{\Delta}),\mathbf{B}\eta_\rho(\mathbf{\Delta})\rangle$ over $\mathbf{\Delta}=\text{diag}\{\delta_1,\delta_2,...,\delta_Q\}\in\mathbb{R}^Q_{0+}$. As it turns out (see Section 4.1), choosing $\mathbf{\Delta}=\mathbf{B}$ is optimal in this case.

(c) Alternatively, one can consider revenue maximization. If serving a unit of backlog at queue q generates revenue m_q, then total revenue generation rate is simply $\sum_{q\in\mathcal{Q}}m_q\Big(\rho_q-(\eta_\rho)_q(\mathbf{\Delta})\Big)=\langle m,\rho-\eta_\rho(\mathbf{\Delta})\rangle=\langle m,\rho\rangle-\langle m,\eta_\rho(\mathbf{\Delta})\rangle$. In order to maximize the revenue generation rate we just need to minimize $\langle m,\eta_\rho(\mathbf{\Delta})\rangle$, that is, make $\eta_\rho(\mathbf{\Delta})$ as orthogonal to m as possible, within the bounds of $\mathcal{K}_\rho$.

(d) Another simple example relates to minimizing the total overload stress $\sum_{q\in\mathcal{Q}}\theta_q(\eta_\rho)_q(\mathbf{\Delta})=\langle\theta,\eta_\rho(\mathbf{\Delta})\rangle$, where θ_q is the stress incurred for a unit of overload at queue q. This can easily be done by making $\eta_\rho(\mathbf{\Delta})$ as orthogonal to θ as possible, within the bounds of $\mathcal{K}_\rho$.

(e) As a special case of the above, one may consider minimizing the aggregate backlog explosion rate $\sum_{q\in\mathcal{Q}}(\eta_\rho)_q(\mathbf{\Delta})=\langle 1,\eta_\rho(\mathbf{\Delta})\rangle$, can be done by making $\eta_\rho(\mathbf{\Delta})$ as orthogonal to 1 as possible, within the bounds of $\mathcal{K}_\rho$.

(f) Finally, one may wish to maintain fairness with respect to the rate of growth of the backlog for the various queues. For example, it is natural to want to minimize the maximum backlog with respect to all queues, which in the limit, corresponds to solving for $\min_{\eta \in \mathcal{K}_\rho} \max_{q \in \mathcal{Q}} (\eta_\rho)_q$.

3. Asymptotic Dynamics in Overload

We begin by building an understanding of the asymptotic dynamics of the scaled workload vector $W(t)/t$, as $t \to \infty$, given an MWM-H matrix, $\mathbf{\Delta}$, and hysteresis function, $h(\cdot)$. We start with a stability result for MWM-H and find that the stability region for MWM-H with setups is *identical* to the original stability region for MWM without setups. This justifies the need to use a hysteresis for throughput maximization to avoid a deadlock such as in Example 1. To do this we follow a similar argument to Armony (1999) and Shah and Wischik (2011) with important modifications that account for the setups and hysteresis. As our focus in this work is on overloaded systems, we simply state the stability result here. We note that the proof of this result comes as a simple corollary of Theorem 3.2.

Theorem 3.1 *If $\rho \in \mathcal{P}$, then $\lim_{t\to\infty} \frac{W_q(t)}{t} = 0, \forall q$.*

Note that the stability region is the same for the policy considered in Dai and Jennings (2004) (appropriately adjusted to our setting) even in the presence of setups. The key difference is that the implementation of their policy requires knowledge of the traffic load vector ρ, while our policy does not.

We now consider an overloaded system with ρ that is outside the stability region. From Armony and Bambos (2003), when $\rho \notin \mathcal{P}$ the workload explodes under a system with and without setup times: that is, $\|W(t)\| \to \infty$, as $t \to \infty$. We are interested in finding out how exactly this happens. Is there a finite limit for $\lim_{t\to\infty} \frac{W(t)}{t}$? If so, what is it and what does it depend on? This section is devoted to answering these questions. We note that in the process of establishing our results for a system in overload, we will also derive stability conditions for the MWM-H policy when there are setup times. Our key result in this section is that, given $\mathcal{S}$, ρ, $\mathbf{\Delta}$, and $h(\cdot)$, there is a **unique** limit $\eta_\rho(\mathbf{\Delta})$ of $\frac{W(t)}{t}$, as $t \to \infty$, that is independent of the hysteresis function h. For notational compactness, we will suppress the dependence of η on ρ and $\mathbf{\Delta}$ throughout this section.

Theorem 3.2 *Fix ρ, $\mathcal{S}$, $\mathbf{\Delta}$ and $h(\cdot)$. Then, there exists $\eta \in \mathbb{R}_+^Q$ such that*

$$\lim_{t\to\infty} \frac{W(t)}{t} = \eta. \tag{3.1}$$

Furthermore, η is defined as the unique solution to the following convex program:

$$\langle \eta, \mathbf{\Delta}\eta \rangle = \min_{\eta' \in \Psi(\rho,\mathcal{S})} \langle \eta', \mathbf{\Delta}\eta' \rangle \tag{3.2}$$

where

$$\Psi(\rho,\mathcal{S}) = \{\eta' : \eta' = (\rho - \psi)^+ \text{ with } \psi \in \mathcal{P}\} \tag{3.3}$$

and $\mathcal{P}$ is the stability region given by $\mathcal{S}$. Equivalently, η is the unique (fixed) point which satisfies $\eta = (\rho - \sum_{S\in\mathcal{S}} \alpha_S S)^+$ with $\sum_{S\in\mathcal{S}} \alpha_S = 1, \alpha_S \geq 0$ and

$$\alpha_S > 0 \;\; \Rightarrow \;\; S \in \mathcal{S}^*(\eta). \tag{3.4}$$

Corollary 3.1 *The limit η of $\frac{W(t)}{t}$ as $t{\to}\infty$ is independent of the hysteresis function $h(\cdot)$.*

To prove the theorem we use fluid model arguments. In the fluid model, work is infinitely divisible and arrives at a constant and deterministic rate. For any given $S \in \mathcal{S}$ work is also depleted at a constant and deterministic rate. In our proof we utilize results for the MWM policy from Armony (1999) which examines the stability region and Shah and Wischik (2011) which characterizes the limit for the corresponding fluid model; both of these works consider systems *without* switching times or hysteresis, so some key adjustments are required. The fluid model approach (Dai (1999)) may be outlined as follows:

Step 1. Postulate the fluid model equations.

Step 2. Establish Lipschitz continuity of the fluid model solutions.

Step 3. Establish that the fluid-scaled queueing process (time is scaled by r and space by $1/r$) is pre-compact as $r{\to}\infty$.

Step 4. Establish that every fluid limit of the fluid-scaled queueing process, as $r{\to}\infty$, satisfies the fluid model equations.

Step 5. Establish the desired property of fluid model solutions (in our case this involves establishing that Theorem 3.2 holds when $W(t)$ is replaced by the fluid content at time t).

Step 6. Use the above steps to establish that a similar properties hold for the original queueing system.

The key to our fluid proof is to establish that, under the MWM-H policy, the dynamics of the fluid model (Step 1) is identical to those of the fluid model for the MWM policy *without* switching times and no hysteresis, as long as the fluid state is non-zero. Once this is established, one can leverage Shah and Wischik (2011) to characterize the limit. This is in sharp contrast to the fluid model of Dai and Jennings (2004), where switching times appear in the fluid limit. All proofs of intermediate results can be found in the appendix.

Step 1. Fluid model equations

The fluid model is a formal construct, characterized by the set of solutions to the fluid model equations. In this step we postulate these fluid model equations (see (3.5)-(3.9)). In Step 4, it is established that every fluid limit is a solution to the fluid model equations. Throughout, an upper bar is used to denote all fluid related expressions. Notice that, as opposed to the queueing model, the fluid model is defined in continuous time.

$$\bar{R}(t) = \rho t, \quad t \geq 0. \tag{3.5}$$

$$\bar{W}(t) = \bar{W}(0) + \rho t - \sum_{S \in \mathcal{S}} S \cdot \bar{T}(S)(t) + \bar{L}(t), \quad t \geq 0, \tag{3.6}$$

$$\bar{L}_q(t) = - \inf_{0 \leq \tau \leq t} \left\{ \bar{W}_q(0) + \rho_q \tau - \sum_{S \in \mathcal{S}} S \cdot \bar{T}(S)(\tau) \right\}, \quad t \geq 0, \quad \bar{L}_q(0) = 0. \tag{3.7}$$

$$\sum_{S \in \mathcal{S}} \bar{T}(S)(t) + \bar{Y}(t) = t, \quad t \geq 0, \quad \bar{T}(S)(0) = 0, \ S \in \mathcal{S}, \quad \bar{Y}(0) = 0. \tag{3.8}$$

where the processes $\bar{T}(S), \ S \in \mathcal{S}$, $\bar{Y}$, and $\bar{L}$ are nondecreasing. $\bar{T}(S)(t)$ may be thought of as the cumulative time up to time t that the vector S has been used, and $\bar{Y}(t)$ is the total time up to time t that the system has been engaged in switching. We note that in the limit, the switching time practically disappears in the overloaded regime, as will be seen in the proof of Lemma 3.1. The process $\bar{L}(t)$ is the minimal process required to keep the fluid workload $\bar{W}(t)$ non-negative; thus, it captures the idling time. Finally, we add one more equation that is specific to the MWM-H policy. Recall that $\mathcal{S}^*(W) := \arg\max_{S \in \mathcal{S}} \langle S', \Delta W \rangle$, then we have that for all $t \geq 0$ which is a regular point[5], if $\bar{W}(t) \neq 0$ then

$$\sum_{S \in \mathcal{S}^*(\bar{W}(t))} \dot{\bar{T}}(S)(t) = 1, \tag{3.9}$$

Essentially, (3.9) implies that whenever the fluid workload vector is non-zero, the fluid system will not be in switching mode, and is processed by one of the "optimal" vectors according to the original MWM without hysteresis. In particular, the hysteresis and the switching do not play an active role in the fluid model. In fact, the hysteresis does not even show up in this scaling.

Steps 2. & 3. Establishing Lipschitz continuity and pre-compactness

This is analogous to Lemma B.2 and Proposition B.1, respectively, in Armony (1999) which analyzes the MWM policy with no switching times and no hysteresis. Details are omitted. We conclude that any fluid-scaled sequence of queueing processes converges almost surely (a.s.), u.o.c.

Step 4. Fluid limits satisfy the fluid model equations

Consider a fluid limit $\bar{R}(t) = \lim_{r \to \infty} \frac{R(0,rt)}{r}$, $\bar{W}(t) = \lim_{r \to \infty} \frac{W(rt)}{r}$, $\bar{Y}(t) = \lim_{r \to \infty} \frac{Y(0,rt)}{r}$, $\bar{T}(S)(t) = \lim_{r \to \infty} \frac{\int_0^{rt} 1_{\{S(\tau)=S\}} d\tau}{r}$, and $\bar{L}(t) = \lim_{r \to \infty} \frac{\sum_{S \in \mathcal{S}} S \int_0^{rt} 1_{\{S(\tau)=S\}} d\tau - M(0,rt)}{r}$. To establish that this fluid limit satisfies equations (3.5) through (3.8) is analogous to the arguments in Section B.2.1 in Armony (1999). In particular, it follows that every such limit is Lipschitz-continuous. Establishing that every such limit satisfies (3.9) requires more work due to the presence of hysteresis and switching times. To establish the above we need to show that if t is a regular point and if $\bar{W}(t) \neq 0$ then

[5] A regular point is a point where the fluid model is differentiable. Due to the absolute continuity of fluid model realizations (Dai (1999)) there are at most countably many non-regular points.

1. $\bar{T}(S)$ is non-increasing at this point, for all $S \notin \mathcal{S}^*(\bar{W}(t))$, and
2. $\bar{Y}$ is non-increasing at this point.

Propositions 3.1 and 3.2 do this.

Proposition 3.1 *Let t be a regular point, and suppose that $\bar{W}(t) \neq 0$. Then, $\bar{T}(S')$ is non-increasing at this point, for all $S' \notin \mathcal{S}^*(\bar{W}(t))$. In particular, $\dot{\bar{T}}(S')(t) = 0$ for all $S' \notin \mathcal{S}^*(\bar{W}(t))$.*

Proposition 3.2 *Let t be a regular point, and suppose that $\bar{W}(t) \neq 0$. Then, $\bar{Y}(t)$ is non-increasing at this point. In particular, $\dot{\bar{Y}}(t) = 0$.*

The proof of Proposition 3.2 relies on the following lemma that argues that the switching times are negligible in the limit.

Lemma 3.1 ***[Asymptotic negligibility of total switching times]*** *Let t and $\delta > 0$ be such that $\inf_{\tau \in [t-\delta, t+\delta]} \|W(r\tau)\| \to \infty$, as $r \to \infty$, then*

$$\lim_{r\to\infty} \frac{Y(r(t-\delta), r(t+\delta))}{r} = 0.$$

Step 5. Limit of the scaled fluid model

Shah and Wischik (2011) established that the fluid model, operating under MWM, satisfies $\lim_{t\to\infty} \bar{W}(t)/t = \eta$, where η is the unique solution of the convex program (3.2). Since we have established that the fluid model associated with the MWM-H policy is identical to the one associated with MWM (as long as $\bar{W}(t) \neq 0$), we have that $\lim_{t\to\infty} \frac{\bar{W}(t)}{t} = \eta$ under MWM-H as well. Note that this result is true regardless of whether the system is stable or not. Thus, MWM-H has the same stability region as MWM without setup times. Theorem 1 in Shah and Wischik (2011) also states that for the fluid model, if $\bar{W}(0) = 0$, then $\bar{W}(t) = \eta t$, for all $t \geq 0$. In particular, we have that if $\bar{W}(0) = 0$, then

$$\bar{W}(1) = \eta. \tag{3.10}$$

Step 6. Limit of the scaled original queueing model

Finally, we wish to show that, for the original queueing model,

$$\lim_{t\to\infty} \frac{W(t)}{t} = \eta, \tag{3.11}$$

where η is the unique solution of the convex program (3.2). But note that $\bar{W}(0) := \lim_{r\to\infty} W(r\cdot 0)/r = 0$. Therefore, by (3.10), we have that $\lim_{t\to\infty} W(t)/t = \bar{W}(1) = \eta$.

To complete the proof of Theorem 3.2 we need to show that the limit η in (3.2) is also the unique fixed point as defined in (3.4). The term "fixed-point" comes from the fact that in the fluid model, if $\frac{\bar{W}(t_0)}{t_0} = \eta$ then $\frac{\bar{W}(t)}{t} = \eta$ for all $t \geq t_0$.

Lemma 3.2 [Fixed point characterization of the limit η] *The limit η of $\frac{W(t)}{t}$ as $t\rightarrow\infty$ is a fixed point. That is,*

$$\eta = \lim_{t\rightarrow\infty} \frac{W(t)}{t} = (\rho - \sum_{S\in\mathcal{S}} \alpha_S S)^+ \tag{3.12}$$

with $\sum_{S\in\mathcal{S}} \alpha_S = 1, \alpha_S \geq 0$. Moreover,

$$\alpha_S > 0 \;\; \Rightarrow \;\; S \in \mathcal{S}^*(\eta). \tag{3.13}$$

We can further utilize the characterization of η as the solution to the convex program (3.2) and leverage convex optimization theory to conclude that there is only one fixed point. As we will see in Section 4, this characterization of η will be useful in the context of controlling the backlogs.

Corollary 3.2 *There exists exactly one fixed point, η:*

$$\begin{aligned} \eta &= [\rho - \sum_{S\in\mathcal{S}} \alpha_S S]^+ \\ 0 &\leq \alpha_S \\ 1 &= \sum_{S\in\mathcal{S}} \alpha_S \\ \alpha_S > 0 &\implies \langle \eta, \mathbf{\Delta} S\rangle \geq \langle \eta, \mathbf{\Delta} S'\rangle, \forall S' \in \mathcal{S}. \end{aligned} \tag{3.14}$$

To summarize, whenever the system is in overload the workload grows along a vector defined by η which is the solution of the convex program in (3.2). From Lemma 3.2, and recalling the geometric interpretation of the policy in Figure 1, we know that η is on the intersection of some set of cones. If there exists only one $\alpha_S > 0$, then $\eta = [\rho - S]^+$ and is in in the cone corresponding to S, C_S. If there are multiple $\alpha_S > 0$, then $\eta = [\rho - \sum_S \alpha_S S]^+$ is on the boundary of the set of cones C_S with $\alpha_S > 0$. In Section 4, we will discuss how we can use this observation to control the limit of the backlog process.

Note, now that we have established that the fluid limit of MWM and MWM-H ard identical, all subsequent results hold for both policies, as long as MWM is only used when there are no setups.

4. Control via the $\mathbf{\Delta}$ matrix

We have now seen how the asymptotic behavior of the workload vector, $W(t)$, behaves given service vectors $\mathcal{S}$ and MWM-H matrix $\mathbf{\Delta}$. In particular, when the system load is outside of the stability region, the workload will explode along a single direction. That is, $W(t) \approx \eta t$. During a long period of temporary stress, the queueing system is effectively *unstable* during this window and the valuable service resources become strained. Under the MWM-H scheduling policy, queues with exceptionally high load will starve resources from other, less stressed, queues. This begs the question of how to share resources in an optimal manner when the system is unstable. Our goal in this section is to discuss how to manipulate the MWM-H (and

MWM) matrix $\mathbf{\Delta}$ to achieve a desired limiting performance of the scaled workload vector. We start in Section 4.1 with a quadratic cost minimization objective where $\mathbf{\Delta}$ can be optimally selected in a simple way. With more general objective functions, one needs to first characterize the set $\mathcal{K}_\rho$ of feasible limits $\eta_\rho(\mathbf{\Delta})$ for all possible choices of the matrix $\mathbf{\Delta}$, and then see how to manipulate $\mathbf{\Delta}$ in order to ensure that the most desirable limit in this set is achieved. We pursue this in Section 4.2.

4.1. Cost Minimization

Consider the following quadratic cost as a function of the workload.

$$C(W) = \langle W, \mathbf{B}W \rangle ,$$

where $\mathbf{B}$ is a positive diagonal matrix. Consider now a policy that selects to use the service vector S at time t, where $S = \arg\max_{S\in\mathcal{S}} \left\langle S, \frac{\partial(C(W))}{\partial(W)} \right\rangle = \arg\max_{S\in\mathcal{S}} \langle S, \mathbf{B}W\rangle$. Whenever the system is not in the midst of switching, this policy selects the service vector S that myopically reduces the cost function by the maximum amount possible in the sense that it drains the largest amount of workload out of the system in the direction of the first derivative of the cost function. This is in fact the MWM-H policy with an MWM-H matrix $\mathbf{\Delta} = \mathbf{B}$. When introduced in the context of backlog cost minimization, the MWM-H policy is reminiscent of the generalized $c\mu$ policy that has been shown to be asymptotically optimal under various settings – without switching times (Van Mieghem (1995), Mandelbaum and Stolyar (2004), Gurvich and Whitt (2009), Armony and Ward (2012)) for separable convex cost functions.

Attempting to minimize the backlog cost is especially challenging in over-stressed systems; when the system is temporarily overloaded, the workload increases without bound, and so does the cost. Still, since backlog grows linearly in time, it is meaningful to attempt to minimize $C(W(t)/t) \equiv C(W(t))/t^2$ for large t, or, more formally, to minimize

$$\limsup_{t\to\infty} C\left(\frac{W(t)}{t}\right) = \limsup_{t\to\infty} \frac{\langle W(t), \mathbf{B}W(t)\rangle}{t^2}.$$

We call a policy *asymptotically optimal* if it obtains this minimum.

Theorem 4.1*Fix ρ, $\mathcal{S}$, and $\mathbf{B}$. Then, the policy MWM-H with MWM matrix $\mathbf{\Delta} = \mathbf{B}$ (and any hysteresis function h) is asymptotically optimal in the sense that it minimizes*

$$\limsup_{t\to\infty} C\left(\frac{W(t)}{t}\right) = \limsup_{t\to\infty} \frac{\langle W(t), \mathbf{B}W(t)\rangle}{t^2}.$$

The proof of Theorem 4.1 is immediate in light of Theorem 3.2, by first realizing that any sublimit of $W(t)/t$ must be in the set $\Psi(\rho, S)$ (3.3), and next seeing that minimizing the RHS of (3.2) is exactly like minimizing $\limsup_{t\to\infty} \frac{\langle W(t), \mathbf{B}W(t)\rangle}{t^2}$ with $\mathbf{B} = \mathbf{\Delta}$.

Remark: Skill-based routing. A special case of our queueing model is a parallel server queueing system with multiple customer classes, and multiple servers, each capable of serving a subset of the customer classes, with service rate that is both class and server dependent. Skill-based routing then refers to the dynamic assignment of servers to customer classes. Each such assignment corresponds to a service vector in which the q^{th} element is the sum of the service rates of all the servers working on customers of class q. In this setting, and without setups, the MWM has a simple analog which is implementable in settings where preemption is not allowed. Specifically, we consider the policy that upon service completion assigns the server j to the customer class whose queue is non-empty and which maximizes $(\mathbf{B}_{ii}X_i)\mu_{ij}$. This is precisely the policy proposed in Stolyar (2004), shown there to be asymptotically optimal in conventional heavy traffic, for a separable quadratic cost under a complete resource pooling condition. A similar analogy to the policy MWM-H can be established for skill-based routing with setups, by introducing a hysteresis.

4.2. Characterizing the set of feasible limits, $\mathcal{K}_\rho$, and the desired MWM-H matrix, $\mathbf{\Delta}$

Beyond quadratic cost minimization, we can choose the matrix $\mathbf{\Delta}$ in order to appropriately position $\eta_\rho(\mathbf{\Delta}) \in \mathcal{K}_\rho$ so as to achieve a target limit of the scaled workload that the system manager considers desirable. The desired limit might have to do with cost/utility optimization as outlined in Section 2.2. In general, and in contrast to the quadratic cost minimization case, both the matrix $\mathbf{\Delta}$ and the range $\mathcal{K}_\rho$ depend on the system parameters including the set $\mathcal{S}$ and the load vector ρ. We begin by characterizing the set $\mathcal{K}_\rho$ of feasible limits.

Recall that, from Lemma 3.2, the limit $\eta_\rho(\mathbf{\Delta})$ of the scaled workload, $W(t)/t$ is a fixed point, as defined in (3.4). In particular, η is on the intersection of the set of cones with $\alpha_S > 0$ in the definition of $\eta = (\rho - \sum_{S\in\mathcal{S}} \alpha_S S)^+$. If there is more than one $\alpha_S > 0$, this corresponds to a cone boundary. This boundary depends on the $\mathbf{\Delta}$ matrix used in the MWM-H scheduling policy.

Under certain necessary and sufficient conditions we can choose the matrix $\mathbf{\Delta}$ so as to arbitrarily place the cone boundary and thereby control the workload to explode along a desired vector defined by η. We leverage this observation in the following characterization of the set of feasible limits.

Proposition 4.1 *A vector $\eta \geq 0$ is in the set $\mathcal{K}_\rho$ if and only if the following conditions hold:*

1. $\eta = (\rho - \sum_{S\in\mathcal{S}} \alpha_S S)^+$, $\alpha_S \geq 0, \sum_{S\in\mathcal{S}} \alpha_S = 1$.
2. *There exists $v \geq 0$ such that $\alpha_S > 0$ implies $\langle v, S\rangle \geq \langle v, S'\rangle$ for all $S' \in \mathcal{S}$. Furthermore, $v_q = 0$ if and only if $\eta_q = 0$.*

Now that we have characterized the set $\mathcal{K}_\rho$, we can determine the $\mathbf{\Delta}$ matrix to achieve the desired limit, η.

Corollary 4.1 *Let $\eta \in \mathcal{K}_\rho$ be the desired limit of the scaled workload vector, and let v be as in Proposition 4.1. Then the required $\mathbf{\Delta}$ such that $\eta = \eta_\rho(\mathbf{\Delta})$ satisfies:*

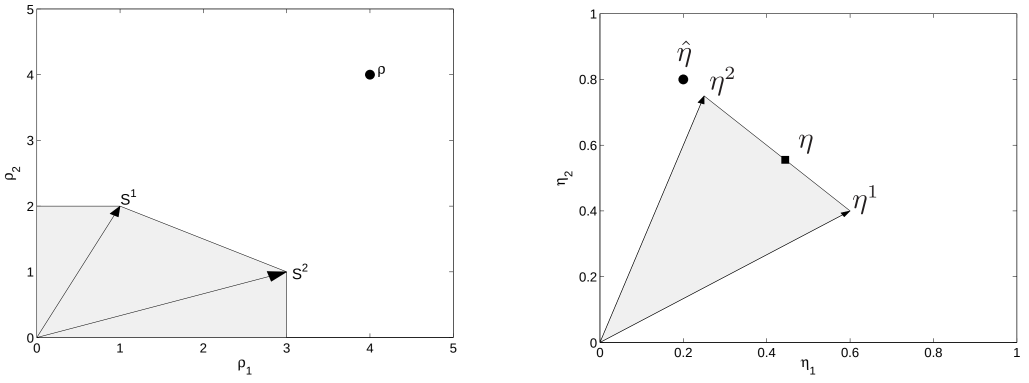
$$\mathbf{\Delta}_{qq} = \begin{cases} \frac{v_q}{\eta_q}, \; v_q > 0; \\ 1, \quad v_q = 0. \end{cases} \tag{4.1}$$

To verify that this $\mathbf{\Delta}$ is indeed the required MWM-H matrix, one can simply check that η is indeed a (unique) fixed point as is required by Theorem 3.2.

4.2.1. On the geometry of the set $\mathcal{K}_\rho$. To gain intuition about the geometry of $\mathcal{K}_\rho$, we now present an example with $N = 2$ service vectors and $Q = 2$ queues and specify the set of feasible limits given $\rho = [4,4]^T$, $S^1 = [1,2]^T$ and $S^2 = [3,1]^T$. It is easy to see that $v = [1,2]^T$ satisfies Condition 2 in Proposition 4.1 since $\langle v, S^1 \rangle = \langle v, S^2 \rangle$. Now, by Condition 1, any η that satisfies

$$\begin{aligned} \eta &= (\rho - \alpha S^1 - (1-\alpha)S^2)^+ \\ &= \alpha(\rho - S^1) + (1-\alpha)(\rho - S^2) \\ &= \alpha[1,3]^T + (1-\alpha)[3,2]^T, \; \text{ for some } \; \alpha \in [0,1] \end{aligned} \tag{4.2}$$

is a feasible limit of $W(t)/t$ and, hence, in the set $\mathcal{K}_\rho$. In Figure 4.2.1(a), we can see the stability region and ρ, which is outside of this region. The set $\mathcal{K}_\rho$ of feasible limits is the same as the set of vectors between the two vectors, $\eta^1 = (\rho - S^1)^+$ and $\eta^2 = (\rho - S^2)^+$ as seen in Figure 4.2.1(b). Thus $\eta \in \mathcal{K}_\rho$, but $\hat{\eta} \notin \mathcal{K}_\rho$. Geometrically, in this example, the set of feasible limits η may be obtained as the positive part of ρ minus the set of all convex combinations of S^1 and S^2.



(a) Stability Region

(b) Feasible Limits

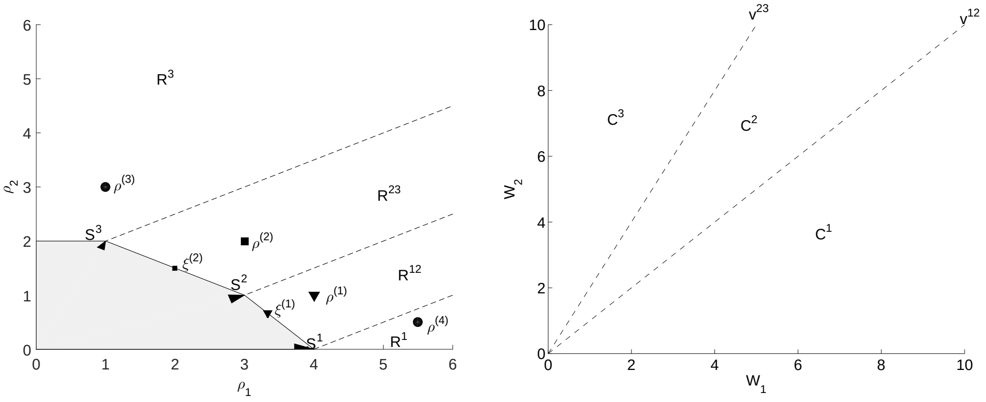
Figure 2 Feasible Limits for ρ: $N = 2$ service vectors and $Q = 2$ queues.

Relating the discussion back to the manipulation of the matrix $\mathbf{\Delta}$ in order to achieve a desired direction, we consider a specific example of asymptotically minimizing the aggregate scaled backlog (as discussed in example (e) in Section 2.2). The limit in $\mathcal{K}_\rho$ that achieves this minimum is $\eta = [1,3]^T$. By (4.1), this corresponds to selecting $\mathbf{\Delta}_{11} = \upsilon_1/\eta_1 = 1$ and $\mathbf{\Delta}_{22} = \upsilon_2/\eta_2 = 2/3$. In contrast, if one wishes to minimize the L_2 norm of the scaled backlog, the optimal choice is $\eta = [1.4, 2.8]^T$, which is achieved by selecting $\mathbf{\Delta}_{11} = \mathbf{\Delta}_{22} = 1$.

Now that we have fully characterized the region $\mathcal{K}_\rho$ and specified how to select $\mathbf{\Delta}$ in order to achieve any direction $\eta \in \mathcal{K}_\rho$, we note that this provides a framework in which to consider optimization problems over arbitrary cost (reward) functions $C(W)$ ($f(W)$). As long as the minimum (maximum) is given by a well-defined feasible direction as defined by $\mathcal{K}_\rho$ in Proposition 4.1, then by Corollary 4.1, we can determine $\mathbf{\Delta}$ necessary to achieve this minimum cost (maximum reward).

4.3. Robustness with Respect to ρ

Thus far, we have assumed that the load vector ρ is known. Under this assumption, we are able to select the necessary MWM-H matrix, $\mathbf{\Delta}$, to achieve any *normalized* limit (or direction) θ, where $\theta_q = \eta_q / \sum_k \eta_k$ for some $\eta \in \mathcal{K}_\rho$. Now we suppose that ρ is unknown or known with some error and examine whether we are still able to choose $\mathbf{\Delta}$ to achieve the desired normalized limit θ. Throughout this discussion, we will assume that $N > 1$; otherwise there is no control and $\eta = (\rho - S)^+$ for all ρ, irrespective of $\mathbf{\Delta}$.



(a) Stability Region

(b) Scheduling cones

Figure 3 **Stability and Cone regions for $N = 3$ service vectors and $Q = 2$ queues.**

Consider the following example with $N = 3$ service vectors and $Q = 2$ queues as depicted in Figure 3. Let $S^1 = [4,0]^T, S^2 = [3,1]^T, S^3 = [1,2]^T$. Additionally, suppose we desire $W(t)/t \rightarrow \eta$, with $\theta = [2/3, 1/3]^T$. We consider 4 different load vectors, which are outside of the stability region:

$$\rho^{(1)} = [4,1]^T, \rho^{(2)} = [3,2]^T, \rho^{(3)} = [1,3]^T, \rho^{(4)} = [5,.5]^T \tag{4.3}$$

When the system load $\rho = \rho^{(3)}$ or $\rho^{(4)}$, the conditions in Proposition 4.1 cannot be satisfied; indeed, starting from those load vectors and moving in the direction of θ, does not hit the stability region at a point that is a convex combination of the service vectors. Hence, the normalized limit, θ, is infeasible and there does not exist a MWM-H matrix to achieve it. With some algebra, we can see that the necessary MWM-H matrix to achieve the desired normalized limit depends on ρ:

$$\boldsymbol{\Delta}(\rho^{(1)}) := \boldsymbol{\Delta}^{(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \boldsymbol{\Delta}(\rho^{(2)}) = \boldsymbol{\Delta}^{(2)} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}. \tag{4.4}$$

From Theorem 3.2, the workload is given by $\eta = [\rho - \sum_m \alpha_m S_m]^+$. Hence, to achieve the desired normalized limit, the goal is to find a point on the boundary of the stability region such that subtracting that point from the system load, ρ, results in η, which we want to be a scaled version of the desired normalized limit θ. From Figure 3(a), we see that the point $\xi^{(1)}$ on the stability boundary which is given by the convex combination of S^1 and S^2 satisfies this constraint for $\rho^{(1)}$. The $\boldsymbol{\Delta}^{(1)}$ matrix places the boundary between the relevant cones along the direction of the desired limit. For $\rho^{(1)}$, the relevant cones are the ones corresponding to S^1 and S^2. This boundary vector can be moved to the direction θ by using $\boldsymbol{\Delta}^{(1)}$. Similarly, $\boldsymbol{\Delta}^{(2)}$ moves the boundary vector between cones 2 and 3 for $\rho^{(2)}$. This example shows that the boundary vector of interest and, subsequently, the necessary MWM-H matrix $\boldsymbol{\Delta}$ depends on ρ.

Despite the preceding example, it is possible to select $\boldsymbol{\Delta}$ without *precise* knowledge of ρ. This ability depends on the number of subsets of service vectors $\mathcal{S}$ of size greater than 1 which satisfy:

$$\langle v, S \rangle = \langle v, S' \rangle \geq \langle v, S'' \rangle, \forall S, S' \in \mathcal{S},\ S'' \in \mathcal{S} \setminus \mathcal{S} \tag{4.5}$$

for some $v \geq 0$, $v \neq \mathbf{0}$. Hence, v is a *boundary vector* as it is a vector on the boundary between the neighboring cones of $\mathcal{S}_M$, when the MWM matrix is the identity matrix. We refer to this boundary as a *relevant boundary*. In our example, there are 2 boundaries of interest: $v^{12} = [1, 1]^T$ and $v^{23} = [1, 2]^T$ (see Figure 3). The boundary which matters depends on both the system load, ρ, and the direction of the desired normalized limit, θ. For $\theta = [2/3, 1/3]^T$, if ρ is in the lower region, R^{12}, then the boundary vector of interest is v^{12}, between cones C^1 and C^2. If $\rho \in R^{23}$, then the boundary vector of interest is v^{23}, between cones C^2 and C^3. If $\rho \in R^1$ or $\rho \in R^3$, the desired limit is infeasible. These regions can be determined for each subset of service vectors by solving for the set of ρ which satisfy Condition 2 of Proposition 4.1. As long as we can determine in which *region* ρ resides, the MWM-H Matrix $\boldsymbol{\Delta}$ can be specified without exact knowledge of ρ. In particular, robustness of the choice of $\boldsymbol{\Delta}$ with respect to ρ holds to a certain extent.

5. Numerical Results

In this section, we present some numerical results to demonstrate the performance of MWM-H Scheduling. We examine how the backlogs grow and approach the established scaled-backlog limit, η. All of our results are asymptotic results with $t \to \infty$. We can see through some numerical simulations how large t must be in practice to approach our asymptotic results.

To start we look at a system with two ($Q = 2$) queues and two ($N = 2$) service vectors:

$$S^1 = [4,0]^T, S^2 = [3,1]^T$$

Our load vector is outside the stability region: $\rho = [4,1]^T \notin \mathcal{P}$. In each time slot, the number of jobs which arrive to queue 1 is uniformly distributed on $[0,8]$; for queue 2 it is uniformly distributed on $[0,2]$. We assume a setup time of 10 time slots and a hysteresis function: $h(W) = .25\sqrt{W_1 + W_2}$. Our goal is to minimize the following quadratic cost function $C(W/t) = \langle W/t, \mathbf{B}W/t \rangle$ for

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

From Theorem 4.1, we know that this implies one should use the MWM-H matrix $\mathbf{\Delta} = \mathbf{B}$. One can solve the convex program (3.2) to determine that $\alpha_1 = 1/3, \alpha_2 = 2/3$ so that $\eta = (\rho - \alpha_1 S^1 - \alpha_2 S^2)^+ = [2/3, 1/3]^T$.

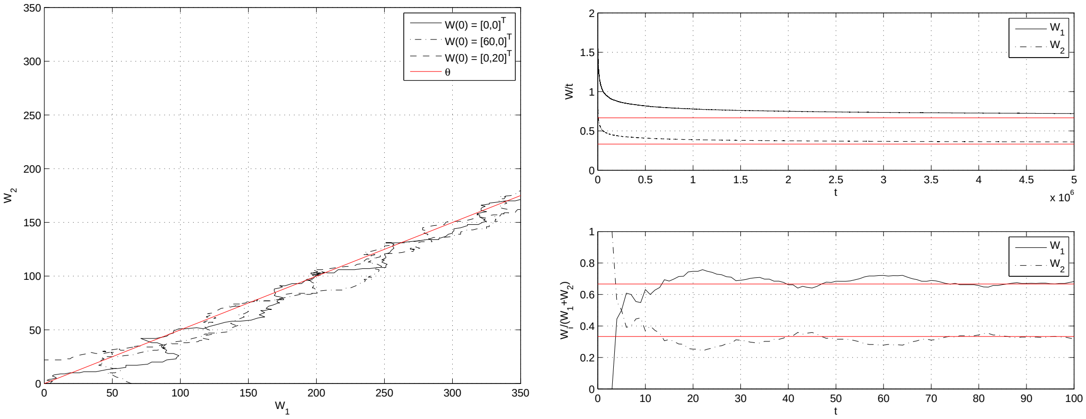
We consider how the workload vector grows for various initial conditions: $W(0) = [0,0]^T, W(0) = [60,0]^T, W(0) = [0,20]^T$. In Figure 4(a), we plot the trajectories of $W(t)$ for the different initial conditions, along with the line $W_1 = 2W_2$. We can see that all three trajectories converge to established vector $\eta_1 = 2/3, \eta_2 = 1/3$. In Figure 4(b), we see the scaled backlogs, $W_i(t)/t$, and the relative backlog, $W_i(t)/\sum_j W_j(t)$, converge starting from initial condition $W(0) = [0,0]$. We notice that it takes a *long* time for $W_i(t)/t$ to converge. This is due to the setup times and not the hysteresis. With a setup time of $T = 0$ and the same matrix $\mathbf{\Delta}$ and hysteresis function h, the scaled backlog converges to η within 200 time slots; with a setup time of $T = 10$, it takes nearly 5,000,000 time slots. That said, we see that the relative backlogs quickly achieve the desired direction.

We now consider an extension of the previous example with $N = 3$ service vectors

$$S^1 = [4,0]^T, S^2 = [3,1]^T, S^3 = [0,2.2]^T$$

Again, we assume the setup time is 10 time slots and a hysteresis function: $h(W) = .25\sqrt{W_1 + W_2}$. Instead of cost minimization, our goal is to control the backlog to grow along $\eta = [2/3, 1/3]^T$. Specifically, we want:

$$\lim_{t \to \infty} \frac{W(t)}{t} = \eta = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}.$$



(a) Different initial conditions

(b) Initial condition $W(0) = [0, 0]$

Figure 4 Dynamics for $N = 2$ service vectors and $Q = 2$ queues.

As we saw in the previous example, when $\rho = [4, 1]^T$, this occurs for

$$\mathbf{\Delta} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

However, we note that the inclusion of the third service vector alters the set $\mathcal{K}_\rho$. In fact, there are now two boundary vectors of interest when deciding the appropriate $\mathbf{\Delta}$ matrix as specified by Proposition 4.1: the one between C^1 and C^2 as well as the one between C^2 and C^3. The boundary which matters depends on ρ. Specifically, consider two load vectors: $\rho = [4, 1]^T$, as before, and $\rho' = [2, 2]^T$. As is indicated by Figure 5, $\eta = [2/3, 1/3]^T$ is indeed a feasible limit for both ρ and ρ', the vector $\xi = \rho - \eta$ is a convex combination of S^1 and S^2, while $\xi' = \rho' - \eta$ is a convex combination of S^2 and S^3. In particular, to achieve the limit η, when the load vector is ρ we need to manipulate $\mathbf{\Delta}$ so as to align the boundary between C^1 and C^2 with η. As established above, in this case, $\mathbf{\Delta} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$; In contrast, when the load vector is ρ' the boundary that needs to be aligned with η is the one between C^2 and C^3. The appropriate MWM-H matrix in this case is $\mathbf{\Delta}' = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$.

We can see in Figure 6 how the asymptotic dynamics of the queues depend on ρ and $\mathbf{\Delta}$. η is shown in red.

In the next experiment, we consider a system with $Q = 3$ queues and $N = 3$ service vectors.

$$S^1 = [5, 0, 0]^T, S^2 = [0, 5, 0]^T, S^3 = [0, 0, 5]^T$$

and MWM-H matrix:

$$\mathbf{\Delta} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix} \tag{5.1}$$

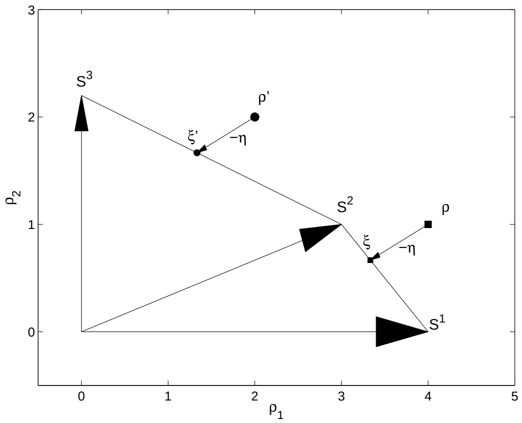


Figure 5 Geometry of 2-queue queueing system with 3 service vectors.

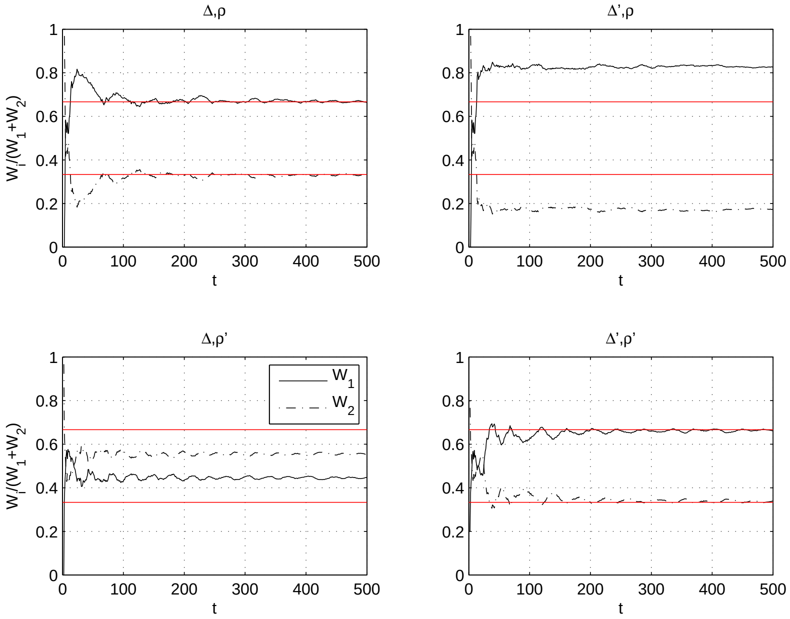
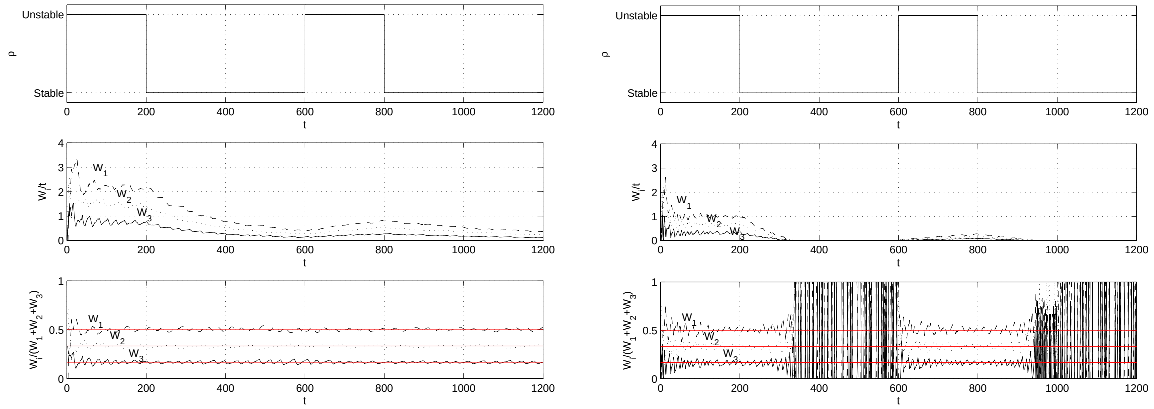


Figure 6 Dynamics of 2-queue queueing system with 3 service vectors.

We consider a hysteresis function: $h(W) = 10\sqrt{W_1 + W_2 + W_3}$ and vary the setup times. The system load oscillates between being stable and unstable. Hence, there are *temporary periods of overload*. $\rho^{\text{stable}} = [1, 0, 1]^T$, $\rho^{\text{unstable}} = [3, 2, 2]^T$. The system spends 200 time slots in the stable mode–$\rho = \rho^{\text{stable}}$–then switches to spend 400 time slots in the unstable mode–$\rho = \rho^{\text{unstable}}$. Arrivals to queue q in each time slot are uniformly distributed between $[0, 2\rho_q]$. When the system is in the stable mode, MWM-H should stabilize the workload. When it is in the unstable mode, MWM-H should converge to a single direction, η. By solving the convex program (3.2), we find that $\eta = [1, 2/3, 1/3]^T$.

Figure 7(a) and 7(b) plot the scaled workload, $W(t)/t$, and relative workloads, $W_i(t)/\sum_i W_j(t)$, under this unstable/stable system when the setup time is 5 and 0 time slots, respectively. We can see that for the first unstable period ($t \in [0, 200]$), the direction η (plotted in red) is quickly achieved. That said, the setup time makes the rate of convergence of $W(t)/t \rightarrow \eta$ much slower. During the stable period ($t \in [200, 600]$), we can see that the length of time spent stable is too short for the system with setups to stabilize, though it is clear the scaled workload is going to zero. Without setups, the system is stabilized, though it takes nearly 150 time slots to do so. In the next unstable period, the scaled backlogs ($W_i(t)/t$) do not appear to stabilize within the 200 epoch period for the system with and without setup times. This is because we are scaling by the *total* time, not just the time from when we enter the period of instability. Hence, it may actually take a very long time before the scaled backlogs converge. On the other hand, the relative workload ($W_i(t)/\sum_j W_j(t)$) quickly aligns with the direction of η. During stable periods this relative backlog is not very informative since all the scaled backlogs will go to zero.



(a) 5 slot setup time: $T = 5$

(b) No setup time: $T = 0$

Figure 7 Dynamics of 3-queue queueing system oscillating between stable and unstable modes.

5.1. Impact of the Hysteresis Function

In order to explore the impact of the hysteresis function on system performance, we are considering a family of related functions. In particular, we consider a hysteresis function of the form:

$$h^{\alpha}(W) = \left(\sum_q W_q \right)^{\alpha}$$

for $\alpha \in (0, 1)$. Our goal is to examine how α, which varies the growth rate of the hysteresis function, impacts the system dynamics. Small α means the hysteresis function will grow slowly, presumably resulting in more switching, while large α will result in less switching, but also less time when the employed service vector is aligned with the ‘optimal’ one that MWM without hysteresis would use.

For this numeric exploration, we consider a similar setup to the one explored in Figure 4 with

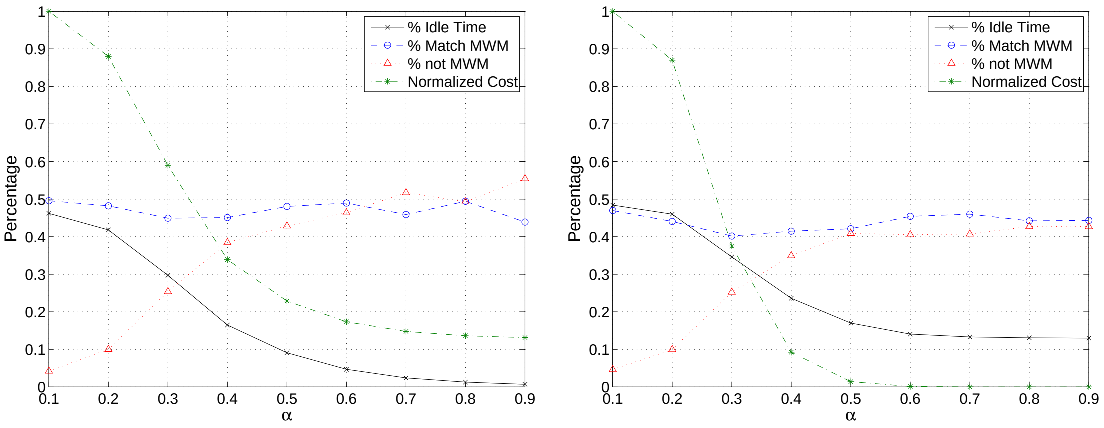
$$S^1 = [4,0]^T, S^2 = [3,1]^T$$

. Again, we assume the setup time is 10 time slots and a cost function of $C(W/t) = \langle W/t, \mathbf{B}W/t \rangle$ with

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

We consider an unstable load vector, $\rho_{\text{unstable}} = [4,1]^T$, as well as a stable load vector $\rho_{\text{stable}} = [3,.5]^T$.

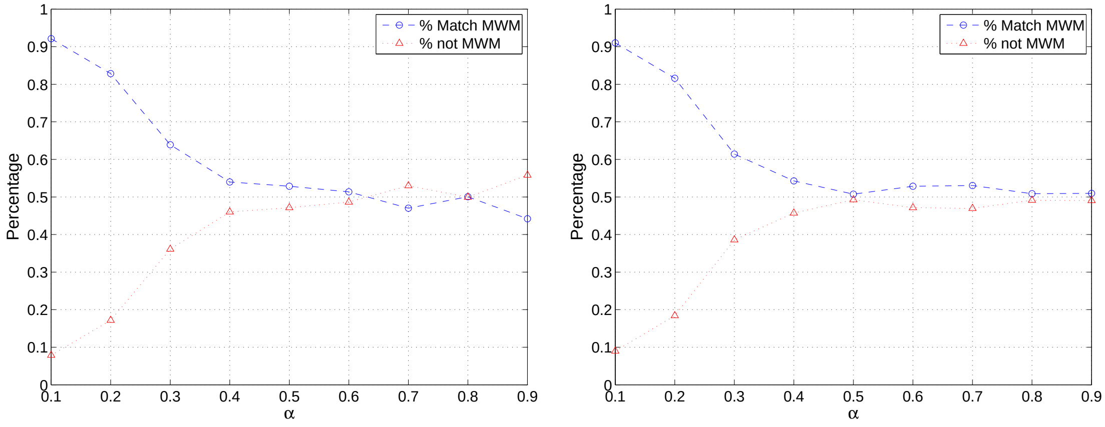
We initialize backlog to $W(0) = [0,0]^T$ and run the simulation for $T = 10,000$ time slots. Figure 8 depicts the percentage of time MWM-H spends idling due the switching time, the percentage of time MWM-H is using the same service vector MWM would be using in a given time slot, and the percentage of time MWM-H is using a different service vector than MWM due to the hysteresis. Finally, we see the normalized cost $\langle W^\alpha/t, \mathbf{B}W^\alpha/t \rangle \,/\, \max_\alpha \langle W^\alpha/t, \mathbf{B}W^\alpha/t \rangle$, where W^α is the backlog at $T = 10,000$ when hysteresis function h^α is used. As expected, the time spent idling due to switching is highest for small α. Because the service vectors are being changed so frequently, this results in the highest cost. As α increases, the time spent switching decreases, but then the time spent using an alternative service vector increases.



(a) Unstable ρ (b) Stable ρ

Figure 8 System performance for different hysteresis functions, $h^\alpha(w) = \left(\sum_q W_q\right)^\alpha$**.**

Figure 9 considers only times when the system is actually working, i.e. not during the switching times. We can see that as α increases, the growth rate of the hysteresis increases, meaning there are larger delays before changing service vectors and a larger discrepancy between MWM and MWM-H. Note that the performance of $h(W) = \log\left(\sum_q W_q\right)$ with respect to matching MWM is practically the same as that of $h^\alpha(W)$ for $\alpha = .25$.



(a) Unstable ρ

(b) Stable ρ

Figure 9 **System performance for different hysteresis functions,** $h^{\alpha}(w) = \left(\sum_q W_q\right)^{\alpha}$**.**

From Figure 8, we can see that larger α results in lower costs. Even though larger α also increases the mismatch with MWM (Figure 9), because the amount of idling time also decreases, the total amount of work done increases with α, resulting in lower costs. It also appears that the differences in the performance of MWM-H become very small for $\alpha \geq .5$. This suggests that it may be reasonable to use MWM-H with any $\alpha \geq .5$.

6. Conclusions and Discussion

In many real world systems, traffic load is unpredictable and often bursty in nature. In any finite window of time, the system may enter a period of temporary instability where the rate of incoming jobs is larger than the rate at which jobs can be serviced. Additionally, many systems require a setup time when service configurations are changed. During this time, no jobs can be serviced, creating additional stress to an already overloaded system.

Our focus in this work on the *instability* region is different than traditional queueing. While it is certainly desirable to operate systems within the stability region, there are many real world scenarios where this may not be possible. Input traffic may surge due to unplanned circumstances. Service resources may be reduced due to unavoidable accidents or catastrophes. During these periods of temporary instability it is often necessary to allocate limited resources in a desirable manner. Once the system exits the window of stress, it will be stabilizable and the natural goals of throughput maximization and cost minimization can be restored.

Our second main focus in this work is considering how to serve queues under the paradigm of *switching times*. We restrict attention to MaxWeight Matching scheduling policies because they are simple to implement and behave well during stable periods when there are no switching times. In particular, MaxWeight

Matching policies guarantee finite backlogs when the system load is within the stability region. To account for the setup times, we introduce a new class of policies: MaxWeight Matching with *Hysteresis*. We find that for appropriately defined hysteresis functions, MWM-H 1) has the same stability region of MWM without setup times and 2) whenever the system is overloaded, the time-scaled backlog approaches a straight line as the time window during which the system is overloaded increases. This straight line can be characterized as a fixed point, or equivalently, as the solution to a simple convex program. As such, it is straightforward to identify this line, as a function of the system parameters, the load vector ρ, and the MaxWeight matrix $\boldsymbol{\Delta}$. Interestingly, this line does not depend on the hysteresis function. Moreover, we are able to adjust the $\boldsymbol{\Delta}$ matrix to achieve various control objectives, such as minimizing quadratic costs.

The proposed MWM-H policy addresses an inherent problem with MWM: frequent switching. Even in the absence of switching times, one could utilize the MWM-H control to mitigate switching, e.g. if there are costs associated with switching service configurations. More generally, there are other policies which can be used in the presence of switching times (e.g. Dai and Jennings (2004), Armony and Bambos (2003)). That said, MWM is a policy which has received substantial attention in the literature and the proposed MWM-H policy inherits many nice properties of MWM, while addressing the problems which arise with setups.

It should be noted that the complexity of the MWM-H policy proposed in this paper is at least as large as that of the original MWM policy without Hysteresis. In particular, one must first solve the optimization problem to identify the MWM service vector and then evaluate the hysteresis function to determine whether a switch should be initiated. One way to reduce the complexity of evaluating MWM is utilizing *Local MWM*, which only considers the nearest neighbor service vectors as alternatives to switch to. Ross and Bambos (2009) demonstrates that such local optimization results in a stabilizing policy (in a setting without setups) that mimics the standard MWM with some delay. Such local optimization could also be applied to the MWM-H policy.

This work can be extended in various directions. First, one might consider whether introducing hysteresis to other policies, such as Projective Cone Scheduling from Ross and Bambos (2009), would also ensure identical stability regions in the presence of setup times. Next, it may be possible to extend this work to networks of parallel queueing systems, by relying on results from Shah and Wischik (2011). Finally, while we have established convergence of the backlog vector under very mild traffic conditions, if more restrictive assumptions are made (such as Markovian queues) one might be able to obtain results on the rate of convergence as well.

References

Armony, M. 1999. Queueing networks with interacting service resources. Ph.D. thesis, Stanford University.

Armony, M., N. Bambos. 2003. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Systems* **44** 209–252.

Armony, M., A. Ward. 2012. Blind fair routing in large-scale service systems. Working paper.

Bertsimas, D., J. Nino-Mora. 1999a. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part i, the single-station case. *Mathematics of Operations Research* **24** 306–330.

Bertsimas, D., J. Nino-Mora. 1999b. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part ii, the multi-station case. *Mathematics of Operations Research* **24** 331–361.

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Chen, X., C.-Y. Su, T. Fukuda. 2008. Adaptive control for the systems preceded by hysteresis. *Automatic Control, IEEE Transactions on* **53**(4) 1019–1025.

Dai, J. G. 1999. Stability of fluid and stochastic processing networks. *Mathematical Physics and Stochastics Miscellanea Publication* **9**.

Dai, J. G., O. B. Jennings. 2004. Stabilizing queueing networks with setups. *Mathematics of Operations Research* **29** 891–922.

DeCandia, G., D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels. 2007. Dynamo: Amazons highly available key-value store. *Proceedings of twenty-first ACM SIGOPS Symposium on Operating Systems Principles*. 205220.

Dikong, E., J.H. Dshalalow. 1999. Bulk input queues with hysteretic control. *Queueing Systems* **32**(4).

Dshalalow, J.H. 1998. Queues with hysteretic control by vacation and post-vacation periods. *Queueing Systems* **29**(2-4) 231–268.

Egorova, R., S. Borst, B. Zwart. 2007. Bandwidth-sharing networks in overload. *Performance Evaluation* **64** 978–993.

Gandhi, A., S. Doroudi, M. Harchol-Balter, A. Scheller-Wolf. 2013. Exact analysis of the M/M/K/Setup class of markov chains via recursive renewal reward. *SIGMETRICS Perform. Eval. Rev.* **41**(1) 153–166.

Gandhi, A., M. Harchol-Balter. 2013. M/G/k with staggered setup. *Operations Research Letters* **41**(4) 317 – 320.

Gandhi, A., M. Harchol-Balter, R. Raghunathan, M.A. Kozuch. 2012. Autoscale: Dynamic, robust capacity management for multi-tier data centers. *ACM Trans. Comput. Syst.* **30**(4) 14:1–14:26.

Golubchik, L., J.C.S. Lui. 1997. Bounding of performance measures for a threshold-based queueing system with hysteresis. *SIGMETRICS Perform. Eval. Rev.* **25**(1) 147–157.

Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management* **11**(2) 237–253.

Hassani, V., T. Tjahjowidodo, T.N. Do. 2014. A survey on hysteresis modeling, identification and control. *Mechanical Systems and Signal Processing* **49**(12) 209 – 233.

Ibe, O. C., J. Keilson. 1995. Multi-server threshold queues with hysteresis. *Performance Evaluation* **21**(3) 185 – 213.

Kelly, F. P. 1986. Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability* **18** 473–505.

Kelly, F. P. 1991. Loss networks. *Annals of Applied Probability* **1** 319–378.

Lan, W-M, T. Lennon Olsen. 2006. Multiproduct systems with both setup times and costs: Fluid bounds and schedules. *Operations Research* **54** 505–522.

Lu, F. V., R. F. Serfozo. 1984. M/M/1 queueing decision processes with monotone hysteretic optimal policies. *Operations Research* **32**(5) 1116–1132.

Mandelbaum, A., S. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855.

Medina-Carnicer, R., F. J. Madrid-Cuevas, R. Muñoz Salinas, A. Carmona-Poyato. 2010. Solving the process of hysteresis without determining the optimal thresholds. *Pattern Recogn.* **43**(4) 1224–1232.

Mo, J., J. Walrand. 2000. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking* **8** 556–567.

Morse, A.S., D.Q. Mayne, G.C. Goodwin. 1992. Applications of hysteresis switching in parameter adaptive control. *Automatic Control, IEEE Transactions on* **37**(9) 1343–1354.

Neely, M.J., E. Modiano, C-P. Li. 2008. Fairness and optimal stochastic control for heterogeneous networks. *IEEE/ACM Transactions on Networking* **16** 396–409.

Perry, O., W. Whitt. 2009. Responding to unexpected overloads in large-scale service systems. *Management Science* **55** 1353–1367.

Perry, O., W. Whitt. 2011. A fluid approximation for service systems responding to unexpected overloads. *Operations Research* **59(5)** 1159–1170.

Plum, H.J. 1991. Optimal monotone hysteretic markov policies in an m/m/1 queueing model with switching costs and finite time horizon. *Zeitschrift fr Operations Research* **35**(5).

Ross, K., N. Bambos. 2009. Projective cone scheduling (PCS) algorithms for packet switches of maximal throughput. *IEEE/ACM Transactions on Networking* **17**(3) 976–989.

Shah, D., D. Wischik. 2011. Fluid models of congestion collapse in overloaded switched networks. *Queueing Syst. Theory Appl.* **69** 121–143.

Stolyar, S. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* **14**(1) 1–53.

Takagi, H. 1997. Queueing analysis of polling models. *Frontiers in Queueing: Models and Applications in Science and Engineering (Dshalalow, Ed.)* **Chapter 5** 119146.

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

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* **5**(3) 809–833.

Xu, Qian, C. Chakrabarti, L.J. Karam. 2011. A distributed canny edge detector and its implementation on fpga. *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE*. 500–505.

Appendix

A. Proofs of results in Section 3

PROOF OF PROPOSITION 3.1: Let $\{r_k\}$ be a subsequence of $\{r\}$ such that $\frac{W(r_k\cdot)}{r_k}$ converges a.s., u.o.c. to $\bar{W}(\cdot)$[6]. Consider a regular point t with $\bar{W}(t) \neq 0$. For convenience of notation we remove k from the notation and assume, without loss of generality, that $\lim_{r\to\infty} \frac{W(r\cdot)}{r} = \bar{W}(\cdot)$. We wish to show that there exists $\delta > 0$, such that for all r large enough $S(r\tau) \in \mathcal{S}^*(\bar{W}(t))$ for all $\tau \in [t-\delta, t+\delta]$.

Let $\delta > 0$ be such that $\inf_{\tau\in[t-\delta,t+\delta]} \left|\bar{W}(\tau)\right| > 0$ and $\mathcal{S}^*(\bar{W}(\tau)) \subseteq \mathcal{S}^*(\bar{W}(t))$, for all $\tau \in [t-\delta, t+\delta]$. The existence of such δ follows from the continuity of the limit $\bar{W}$. We now argue that for all r large enough we have that if $S \in \mathcal{S}^*(\bar{W}(\tau))$ and $S' \notin \mathcal{S}^*(\bar{W}(\tau))$ then

$$S' \notin \mathcal{S}^*(W(r\tau)), \tag{A1}$$

and

$$\langle S, \Delta W(r\tau)\rangle > \langle S', \Delta W(r\tau)\rangle + h(W(r\tau)), \tag{A2}$$

for all $\tau \in [t-\delta, t+\delta]$. To establish (A1) note that $\mathcal{S}^*(W(r\tau)) = \mathcal{S}^*\left(\frac{W(r\tau)}{r}\right)$, and the result follows from the the fact that $\bar{W}(\cdot) = \lim_{r\to\infty} \frac{W(r\cdot)}{r}$, a.s., u.o.c. Similarly, (A2) follows after dividing all expressions by r, taking the limit, and recalling that h is sublinear, that is, $\lim_{\|w\|\to\infty} \frac{h(w)}{\|w\|} = 0$, u.o.c., so that $\lim_{r\to\infty} \frac{h(W(r\tau))}{\|W(r\tau)\|} \frac{\|W(r\tau)\|}{r} = 0$, u.o.c. ■

PROOF OF PROPOSITION 3.2: Let $\{r_k\}$ be a subsequence of $\{r\}$ such that $\frac{W(r_k\cdot)}{r_k}$ converges a.s., u.o.c. to $\bar{W}(\cdot)$, and consider a regular point t with $\bar{W}(t) \neq 0$. For convenience of notation we remove k from the notation and assume, without loss of generality, that $\lim_{r\to\infty} \frac{W(r\cdot)}{r} = \bar{W}(\cdot)$. Let $\delta > 0$ be such that $\inf_{\tau\in[t-\delta,t+\delta]} \left|\bar{W}(\tau)\right| > 0$. We wish to show that for all r large enough the system engages in switching for a negligible amount of time in $[t-\delta, t+\delta]$. More precisely, we wish to show that

$$\lim_{r\to\infty} \frac{Y(r(t-\delta), r(t+\delta))}{r} = 0. \tag{A3}$$

To establish (A3) we first prove Lemma 3.1. The proof of the proposition then follows from the fact that $\inf_{\tau\in[t-\delta,t+\delta]} \|W(r\tau)\| \to \infty$, as $r \to \infty$, by the definition of δ. ■

[6] Recall that the queueing model is in discrete time. With a slight abuse of notation we let $W(t) := W(\lfloor t \rfloor)$. We use a similar notational convention with other quantities that are associated with the queueing model.

PROOF OF LEMMA 3.1: For every fixed r, let $t-\delta \leq \tau_{r,1} < \tau_{r,2} < ... < \tau_{r,m_r} < t+\delta$ be the sequence of points where the system initiates switching at time $r\tau_{r,k}$, $1 \leq k \leq m_r$. Note that $\tau_{r,1} = t-\delta$ if the system is in the midst of switching at time $r(t-\delta)$. By definition,

$$Y(r(t-\delta), r(t+\delta)) \leq T(m_r+1). \tag{A4}$$

At the same time, for $m_r > 1$:

$$2\delta \geq \sum_{k=1}^{m_r-1} (\tau_{r,k+1} - \tau_{r,k}).$$

We wish to evaluate the term $\tau_{r,k+1} - \tau_{r,k}$. Let $S^{r,k} = H(r\tau_{r,k})$ be the service vector which is being changed to at time $r\tau_{n,k}$. By definition we have that:

$$\begin{aligned} \left\langle W(r\tau_{n,k}), \mathbf{\Delta} S^{r,k} \right\rangle &\geq h(W(r\tau_{n,k})) + \left\langle W(r\tau_{n,k}), \mathbf{\Delta} S^{r,k+1} \right\rangle \\ \left\langle W(r\tau_{n,k+1}), \mathbf{\Delta} S^{r,k+1} \right\rangle &\geq h(W(r\tau_{n,k+1})) + \left\langle W(r\tau_{n,k+1}), \mathbf{\Delta} S^{r,k} \right\rangle \end{aligned}$$

With a little algebra, we can see that

$$\left\langle W(r\tau_{n,k+1}) - W(r\tau_{n,k}), \mathbf{\Delta} S^{r,k+1} \right\rangle - \left\langle W(r\tau_{n,k+1}) - W(r\tau_{n,k}), \mathbf{\Delta} S^{r,k} \right\rangle \geq h(W(r\tau_{n,k+1})) + h(W(r\tau_{n,k})) \tag{A5}$$

We consider the expression on the left hand side:

$$\begin{aligned} &\left\langle W(r\tau_{n,k+1}) - W(r\tau_{n,k}), \mathbf{\Delta}(S^{r,k+1} - S^{r,k}) \right\rangle \\ &= \left\langle \left(R(r\tau_{n,k+1}, r\tau_{n,k} - 1) - \sum_{\tau = r\tau_{n,k}}^{r\tau_{n,k+1}-1} D(\tau) \right), \mathbf{\Delta}(S^{r,k+1} - S^{r,k}) \right\rangle \\ &= \left\langle R(r\tau_{n,k+1}, r\tau_{n,k} - 1) - \sum_{\tau = r\tau_{n,k}}^{r\tau_{n,k+1}-1} D(\tau), \mathbf{\Delta}[(S^{r,k+1} - S^{r,k})^+ + (S^{r,k+1} - S^{r,k})^-] \right\rangle \\ &\leq \left\langle R(r\tau_{n,k+1}, r\tau_{n,k} - 1), \mathbf{\Delta}(S^{r,k+1} - S^{r,k})^+ \right\rangle - \left\langle \sum_{\tau = r\tau_{n,k}}^{r\tau_{n,k+1}-1} D(\tau), \mathbf{\Delta}(S^{r,k+1} - S^{r,k})^- \right\rangle \\ &\leq rQ \times (\max_q \bar{\sigma}_q + \max_{S \in \mathcal{S}} \max_q S_q)(\tau_{n,k+1} - \tau_{n,k}) \max_q \mathbf{\Delta}_q \max_{S \in \mathcal{S}} \max_q S_q \\ &= rK(\tau_{n,k+1} - \tau_{n,k}) \end{aligned}$$

Where $K = Q \times (\max_q \bar{\sigma}_q + \max_{S \in \mathcal{S}} \max_q S_q) \max_q \mathbf{\Delta}_q \max_{S \in \mathcal{S}} \max_q S_q < \infty$ is a well-defined positive finite constant. Combining this with (A5), we have:

$$r(\tau_{n,k+1} - \tau_{n,k}) \geq \frac{2}{K} \inf_{\tau_{n,k} \leq \tau < \tau_{n,k+1}} h(W(r\tau)) \tag{A6}$$

Thus,

$$2r\delta \geq (m_n - 1) \frac{2}{K} \inf_{t-\delta \leq \tau < t+\delta} h(W(r\tau)). \tag{A7}$$

Combining this result with our initial observation in (A4):

$$\frac{Y(r(t-\delta),r(t+\delta))}{2r\delta} \leq \frac{T(m_r+1)}{(m_r-1)\frac{2}{K}\inf_{t-\delta\leq\tau<t+\delta}h(W(r\tau))} \leq \frac{\frac{3TK}{2}}{\inf_{t-\delta\leq\tau<t+\delta}h(W(r\tau))} \tag{A8}$$

for all r such that $m_r > 1$. By assumption, we have that $\inf_{\tau\in[t-\delta,t+\delta]}\|W(r\tau)\|\to\infty$ as $r\to\infty$, and hence $\inf_{t-\delta\leq\tau<t+\delta}h(W(r\tau))\to\infty$ as $r\to\infty$.

Alternatively, with $m_r \leq 1$, there is at most T time spent switching in the interval:

$$\frac{Y(r(t-\delta),r(t+\delta))}{2r\delta} \leq \frac{T}{2r\delta}\to 0, \text{ as } r\to\infty. \tag{A9}$$

■

Proof of Lemma 3.2: The proof is an immediate consequence of the fluid model equations (3.5)-(3.9). In particular, assuming that $\bar{W}(0)=0$, we have that $\bar{W}(t)=\eta t$ and every point t is a regular point. The proof follows by setting for all $S\in\mathcal{S}$, $\alpha_S := \bar{T}(S)(1) = \bar{T}(S)(t)/t, \ \forall t.$ ■

Proof of Corollary 3.2: Now, we demonstrate that any fixed point is a solution to the convex program (3.2). Since the solution to (3.2) is unique, the result will follow. Consider the KKT conditions for optimality. Our goal is to show that if $\eta = (\rho - \sum_{S\in\mathcal{S}}\alpha_S S)^+$, $\alpha_S \geq 0$, $\sum_{S\in\mathcal{S}}\alpha_S = 1$ is such that for all S with $\alpha_S > 0$, we have that $S\in\mathcal{S}^*(\eta)$, then it is a solution to the convex program (3.2). The KKT conditions are necessary and sufficient for optimality if the objective function is differentiable and Slater's constraint is satisfied (Boyd and Vandenberghe (2004)). Both are easily verifiable in our case.

To examine the KKT conditions, we first rewrite the convex program in (3.2) as an equivalent convex program:

$$\begin{aligned} \min_{\psi,\alpha} \quad & \langle \rho-\psi, \mathbf{\Delta}(\rho-\psi)\rangle \\ s.t. \quad & \psi \leq \sum_{S\in\mathcal{S}}\alpha_S S \\ & 0\leq\psi\leq\rho \\ & \alpha_S \geq 0, \forall S\in\mathcal{S} \\ & \sum_{S\in\mathcal{S}}\alpha_S = 1 \end{aligned} \tag{A10}$$

The KKT conditions for optimality say that all the constraints must be satisfied and:

$$\begin{aligned} \nabla\langle\rho-\psi,\mathbf{\Delta}(\rho-\psi)\rangle + \nabla\lambda'(\psi-\sum_{S\in\mathcal{S}}\alpha_S S) + \nabla\lambda''(\psi-\rho) - \nabla\lambda\alpha + \nabla v(\sum_{S\in\mathcal{S}}\alpha_S - 1) &= 0 \\ \lambda'_q(\psi_q - \sum_{S\in\mathcal{S}}\alpha_S S_q) &= 0 \\ \lambda''_q(\psi_q-\rho_q) &= 0 \\ \lambda_S\alpha_S &= 0 \\ \lambda,\lambda',\lambda'',v &\geq 0 \end{aligned} \tag{A11}$$

We look at the first condition:

$$\begin{aligned}\nabla_{\psi}:&\ 2\boldsymbol{\Delta}(\rho-\psi)=\lambda'+\lambda''\\ \nabla_{\alpha_S}:&\sum_q \lambda'_q S_q = v-\lambda_S\\ \Rightarrow&\ \ 2\langle(\rho-\psi),\boldsymbol{\Delta}S\rangle = v-\lambda_S+\langle\lambda'',S\rangle\end{aligned} \tag{A12}$$

Now we show that for any fixed point, there exists $\lambda,\lambda',\lambda'',v$ which satisfy the KKT condition (A11). To do this, suppose we are given a fixed point $\eta=\rho-\psi=(\rho-\sum_{S\in\mathcal{S}}\alpha_S S)^+$ with

$$\alpha_S>0 \implies \langle\rho-\psi,\boldsymbol{\Delta}S\rangle \geq \langle\rho-\psi,\boldsymbol{\Delta}S'\rangle\,,\ \ \forall S'\in\mathcal{S}. \tag{A13}$$

We will construct non-negative Lagrange multipliers to satisfy the KKT conditions.

Consider q such that $\rho_q>\psi_q$. In order to satisfy the third constraint in (A11), $\lambda''_q=0$. Now if $\rho_q\leq r_q$, the first constraint in (A12) implies that $\lambda'_q+\lambda''_q\leq 0$, which means that $\lambda'_q=\lambda''_q=0$. To ensure that the Lagrange multipliers are non-negative, we must have that $\lambda''_q=0$ for all q. Subsequently:

$$\lambda'=2\boldsymbol{\Delta}(\rho-\psi) \tag{A14}$$

Hence, λ'_q is non-zero if and only if $(\rho-\psi)_q>0$. Considering the second constraint in (A11), this would also require that $\psi_q=\sum_{S\in\mathcal{S}}\alpha_S S_q$, which is certainly feasible.

Consider $\alpha_S>0$. To satisfy the fourth constraint in (A11), $\lambda_S=0$. Now to satisfy the third constraint in (A12):

$$0\leq 2\langle\rho-\psi,\boldsymbol{\Delta}S\rangle = v, \forall S \text{ such that } \alpha_S>0 \tag{A15}$$

which also satisfies the non-negativity of v. Note that because $\lambda'\geq 0$ and $S\geq 0$ the above expression for v is necessarily non-negative. Consider $\alpha_{S'}=0$ and $\alpha_S>0$. By the assumption that $\rho-\psi$ is a fixed point:

$$\begin{aligned}\langle\rho-\psi,\boldsymbol{\Delta}S'\rangle &\leq \langle\rho-\psi,\boldsymbol{\Delta}S\rangle\\ &\implies \frac{v-\lambda_{S'}}{2}\leq\frac{v}{2}\\ &\implies \lambda_{S'}\geq 0\end{aligned} \tag{A16}$$

Hence, we can satisfy the KKT conditions in (A11) with non-negative $\lambda,\lambda',\lambda'',v$. Since any fixed point satisfies the necessary and sufficient KKT conditions, all fixed points are solutions to the convex program. There is only one solution and so there is only one fixed point. This concludes the proof. ■

B. Proofs of Results in Section 4

We begin with an auxiliary result, that we will utilize next.

Lemma B.1 *In Q-dimensions, consider any M service vectors $S^{i_1}, S^{i_2}, \ldots, S^{i_M}$. Suppose there exists a diagonal positive definite matrix $\hat{\mathbf{\Delta}}$ and non-negative Q-dimensional vector $v \geq 0$ such that v is a* boundary vector *of the M cones, i.e. for each $k \in [1, M]$ and for all j:*

$$\left\langle v, \hat{\mathbf{\Delta}} S^{i_k} \right\rangle \geq \left\langle v, \hat{\mathbf{\Delta}} S^{j} \right\rangle \tag{B1}$$

Then, for any non-negative vector η such that $\eta_q = 0$ if and only if $v_q = 0$, there exists a diagonal positive definite matrix $\mathbf{\Delta}$ such that for each $k \in [1, M]$ and all j:

$$\left\langle \eta, \mathbf{\Delta} S^{i_k} \right\rangle \geq \left\langle \eta, \mathbf{\Delta} S^{j} \right\rangle \tag{B2}$$

i.e. the boundary between the M cones can be placed arbitrarily in $\mathbb{R}_+^Q$. This matrix is specified as:

$$\mathbf{\Delta}_{qq} = \begin{cases} \frac{\hat{\mathbf{\Delta}}_{qq} v_q}{\eta_q}, & v_q > 0; \\ 1, & v_q = 0. \end{cases} \tag{B3}$$

PROOF: We show this via construction. Recall that $\hat{\mathbf{\Delta}}$ is diagonal: $\left\langle v, \hat{\mathbf{\Delta}} S \right\rangle = \sum_q v_q \hat{\mathbf{\Delta}}_{qq} S_q$. For $v_q > 0$, $\mathbf{\Delta}_{qq} = \frac{\hat{\mathbf{\Delta}}_{qq} v_q}{\eta_q} \geq 0$, where the positivity comes from the fact that each element is positive. If $v_q = 0$, $\mathbf{\Delta}_{qq} = 1$ or some other arbitrary positive value.

Now, for any i_j ($j \in [1, m]$) and k the following holds:

$$\begin{aligned} \left\langle v, \hat{\mathbf{\Delta}} S^{i_j} \right\rangle \geq \left\langle v, \hat{\mathbf{\Delta}} S^{k} \right\rangle &\Rightarrow \sum_q \hat{\mathbf{\Delta}}_{qq} v_q (S^{i_j})_q \geq \sum_q \hat{\mathbf{\Delta}}_{qq} v_q (S^{k})_q \\ &\Rightarrow \sum_q v_q \hat{\mathbf{\Delta}}_{qq} (S^{i_j})_q - \sum_q \eta_q \mathbf{\Delta}_{qq} (S^{i_j})_q - \sum_q \eta_q \mathbf{\Delta}_{qq} (S^{k})_q \\ &\quad \geq \sum_q v_q \hat{\mathbf{\Delta}}_{qq} (S^{k})_q - \sum_q \eta_q \mathbf{\Delta}_{qq} (S^{i_j})_q - \sum_q \eta_q \mathbf{\Delta}_{qq} (S^{k})_q \\ &\Rightarrow \sum_q [(S^{i_j})_q (\hat{\mathbf{\Delta}}_{qq} v_q - \mathbf{\Delta}_{qq} \eta_q) - (S^{k})_q \mathbf{\Delta}_{qq} \eta_q)] \\ &\quad \geq \sum_q [(S^{k})_q (\hat{\mathbf{\Delta}}_{qq} v_q - \mathbf{\Delta}_{qq} \eta_q) - (S^{i_j})_q \mathbf{\Delta}_{qq} \eta_q)] \\ &\Rightarrow -\sum_q (S^{k})_q \mathbf{\Delta}_{qq} \eta_q \geq -\sum_q (S^{i_j})_q \mathbf{\Delta}_{qq} \eta_q \Rightarrow \left\langle \eta, \mathbf{\Delta} S^{i_j} \right\rangle \geq \left\langle \eta, \mathbf{\Delta} S^{k} \right\rangle \end{aligned} \tag{B4}$$

■

PROOF OF PROPOSITION 4.1: Assume that Conditions 1 and 2 are satisfied. We show this implies there exists a $\mathbf{\Delta}$ such that $\lim_{t \to \infty} \frac{W(t)}{t} = \eta$. We first consider Condition 2. This says that $v \geq 0$ is the boundary of cones C^S where $\alpha_S > 0$ defined by $\hat{\mathbf{\Delta}} = \mathbf{I}$. By Lemma B.1, we can construct a $\mathbf{\Delta}$ such that for all

$\alpha_S > 0$:$\langle \eta, \mathbf{\Delta} S \rangle \geq \langle \eta, \mathbf{\Delta} S' \rangle\ ,\ \ \forall S' \in \mathcal{S}$. Now, in conjunction with Condition 1, we have that $\eta \in \Psi(\rho, \mathcal{S})$ is a fixed point. By Theorem 3.2, we have that $\lim_{t\to\infty} W(t)/t = \eta$.

Now suppose there exists a $\mathbf{\Delta}$ such that $\lim_{t\to\infty} W(t)/t = \eta$. By Theorem 3.2, η is a (the only) fixed point and, hence, satisfies Condition 1. Now, we show that Condition 2 is satisfied by constructing the necessary $v \geq 0$. Similar to the construction of $\mathbf{\Delta}$ in the proof of Lemma B.1 we can determine v–the boundary induced when $\hat{\mathbf{\Delta}} = \mathbf{I}$. That is $v_q = \mathbf{\Delta}_{qq}\eta_q$ which equals 0 if and only if $\eta_q = 0$. This v_q satisfies Condition 2. ■