Centrum voor Wiskunde en Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

PNA

Asymptotically optimal parallel resource assignment with interference

I.M. Verloop, R. Núñez-Queija

REPORT PNA-E0805 AUGUST 2008

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Asymptotically optimal parallel resource assignment with interference

ABSTRACT
Motivated by scheduling in multi-cell wireless networks and resource allocation in computer systems, we study a service facility with two types of users (or jobs) having heterogeneous size distributions. Our model may be viewed as a parallel two-server model, where either both job types can be served in parallel, each by a dedicated server, or both servers are simultaneously allocated to one type only (also known as "cycle stealing"). A special instance of the model is the coupled-processors model. In this paper, the aggregate service capacity is assumed to be largest when both job types are served in parallel, but giving preferential treatment to one of the job classes may be advantageous when aiming at minimization of the number of jobs, or when job types have different economic values, for example. The model finds applications in third generation wireless networks and in resource allocation of computer systems. For practical reasons, these application areas do not allow for centralized control. Still, knowledge of the theoretical achievable (centralized) optimum is extremely valuable to estimate the scope for improvement of the implemented decentralized control. We therefore set out to determine the optimal server allocation policies that in some appropriate sense minimize the total number of users in the system. At any given moment, the optimal resource allocation depends on the numbers of users present in each class. For some particular cases we can determine the optimal policy exactly, but in general this is not analytically feasible. Therefore, we study asymptotically optimal policies in the fluid limit which prove to be close to optimal. These policies can be characterized by either linear or exponential switching curves. We compare our results with existing approximations based on optimization in the heavy traffic regime, where threshold-based strategies and so-called Max-Weight policies are known to be asymptotically optimal. By simulations we show that our simple computable switching-curve strategies based on the fluid analysis in general perform well and that significant gains can be achieved compared to (i) the strategy that maximizes the aggregate service capacity at all times, (ii) the strategy that maximizes the job departure rate at all times, (iii) the threshold-based policies, and (iv) the Max-Weight policies.

2000 Mathematics Subject Classification: 60K25, 68M20, 90B22, 90B36, 93E20
Keywords and Phrases: parallel servers; resource sharing; dynamic control; optimal scheduling; switching curve; fluid control problem; fluid limit; asymptotic optimality

Asymptotically optimal parallel resource assignment with interference

Maaike Verloop* , Rudesindo Núñez-Queija*,‡

*CWI, The Netherlands
‡TNO Information and Communication Technology, The Netherlands

Abstract

Motivated by scheduling in multi-cell wireless networks and resource allocation in computer systems, we study a service facility with two types of users (or jobs) having heterogeneous size distributions. Our model may be viewed as a parallel two-server model, where either both job types can be served in parallel, each by a dedicated server, or both servers are simultaneously allocated to one type only (also known as "cycle stealing"). A special instance of the model is the coupled-processors model. In this paper, the aggregate service capacity is assumed to be largest when both job types are served in parallel, but giving preferential treatment to one of the job classes may be advantageous when aiming at minimization of the number of jobs, or when job types have different economic values, for example.

The model finds applications in third generation wireless networks and in resource allocation of computer systems. For practical reasons, these application areas do not allow for centralized control. Still, knowledge of the theoretical achievable (centralized) optimum is extremely valuable to estimate the scope for improvement of the implemented decentralized control. We therefore set out to determine the optimal server allocation policies that in some appropriate sense minimize the total number of users in the system. At any given moment, the optimal resource allocation depends on the numbers of users present in each class. For some particular cases we can determine the optimal policy exactly, but in general this is not analytically feasible. Therefore, we study asymptotically optimal policies in the fluid limit which prove to be close to optimal. These policies can be characterized by either linear or exponential switching curves. We compare our results with existing approximations based on optimization in the heavy traffic regime, where threshold-based strategies and so-called Max-Weight policies are known to be asymptotically optimal. By simulations we show that our simple computable switching-curve strategies based on the fluid analysis in general perform well and that significant gains can be achieved compared to (i) the strategy that maximizes the aggregate service capacity at all times, (ii) the strategy that maximizes the job departure rate at all times, (iii) the threshold-based policies, and (iv) the Max-Weight policies.

1 Introduction

In many practical applications where resources must be allocated among several contending users or tasks, the resource capacity itself may be affected by the scheduling strategy deployed. Our work is motivated by two specific application areas. In third generation wireless networks, neighboring base stations may interfere with each other when transmitting simultaneously. When one base station is not active, other base stations can work at higher rates, see for example [7]. For data applications, base stations may coordinate transmissions (i.e., transmit simultaneously or alternatingly) so as to optimize the use of the shared spectrum. A second motivating application is the scheduling of resources in computer systems (or Web servers) where jobs must be routed to one of several servers, see for example [22, 23]. There, the capacity depends on the allocation when servers are specialized for certain tasks.

Scheduling of resources with state-dependent capacities has attracted much attention in recent years. Most of the results concern stochastic stability properties of such systems. Due to the state-dependence of the resource capacity, even this most basic performance measure is a non-trivial task to determine. In [10] bounds for stability in a general class of systems with state-dependent capacity have been determined. In the specific context of wireless networking, stability of utility-based allocation strategies was shown to be intimately related with the shape of the feasible rate region [9], i.e., the set of simultaneously achievable rates by all users. With a convex rate region, the system is stabilized by any such allocation strategy, but this is not the case for non-convex rate regions. These results were later generalized to non-convex and time-varying rate regions in [18], showing the precise conditions for stability of utility-based strategies under quite general assumptions on the time-variations.

As may be expected from the complexity of determining stability, results on the flow level *performance* in terms of system delay or system occupancy are scarce. In particular, for parallel-server models where job types can be served in parallel, all by a dedicated server, or where several servers can be simultaneously allocated to one type only, most results in this direction concentrate on performance of a specific class of allocation strategies. For example, besides determining the stability conditions, [23] investigate the performance for threshold-based strategies. One main observation there is that finding reasonable values for the thresholds is not trivial since performance as well as stability can be quite sensitive to the threshold values. Approximations for mean response times are given in [22]. A general class of threshold-based priority policies for multi-class parallel-server networks is also proposed in [26]. For these strategies, the authors derive approximate formulas for the queue lengths and illustrate how these can be used to obtain reasonable threshold values. In [7] a parallel two-server model is analyzed under the policy that maximizes the aggregate service capacity at all times, and a diffusion approximation for the queue lengths is found for a specific heavy traffic setting.

Our goals here are to study the structural properties of optimal resource allocation strategies in a parallel-server model, and to determine computable approximations that are close to optimality. Our objective is to minimize (in some appropriate sense) the total number of users. A crucial observation when addressing optimality is that, in general, users will have class-specific sizes, so that few users of one class can typically add up to the same amount of work as many of another class. On one hand, it seems reasonable to maximize the departure rate of users/tasks, by serving the small users first. In the short run, this will keep the number of users/tasks in the system at a low level, thus shortening overall delays. On the other hand, it is also desirable to deploy the highest possible service capacity. That will minimize the volume of back-logged work and drain the system at maximum rate, thus ensuring maximum stability. In general, there can be a trade-off between these two objectives. The main challenge is then to weigh the trade-off of the two intrinsically different objectives and to find the optimal allocation policy.

Determining the exact optimal policy in a parallel-server model has so far proven analytically infeasible in literature. Most research on this area has focused on a heavily loaded systems under a (complete) resource pooling condition for which asymptotically optimal strategies are determined [1, 5, 6, 15, 16, 20, 25]. In [1, 15, 16] several kinds of discrete-review policies are proposed (at discrete points in time the system is reviewed, and decisions are based on the queue lengths at that moment) and are proved to be asymptotically optimal. In [20, 25] a generalized $c\mu$-rule (the Max-Weight policy is a special of this rule) is proposed which myopically maximizes the rate of decrease of certain instantaneous holding costs. This policy is robust in the sense that it only depends on the service rates and the cost function, and it is proved that in heavy traffic this policy asymptotically minimizes the cumulative costs over any finite interval. However, the cost function can not be chosen to represent the total number of users present in the system. In [5, 6], the authors prove that threshold-based strategies asymptotically minimize the scaled total number of users in a heavy-traffic setting. In general, the order of magnitude of the optimal thresholds as functions of the traffic load can be determined, but this does not give a recipe to choose good threshold values in under-loaded regime. In [26], the authors propose values for the threshold, which can be found by solving a minimization problem.

We consider a parallel two-server model with two traffic classes that can be served either in parallel or alternatingly. The highest service capacity is achieved when serving both classes in parallel, but with asymmetric job sizes the departure rate may be larger when serving one class only. This model is identical or similar to several of the aforementioned papers. Determining the mean number of jobs in closed form is not feasible: A special case of the model (known as the coupled-processors model in queueing literature) has been shown to be notoriously hard to analyze, requiring the solution of a Riemann-Hilbert boundary value problem [12]. For some special cases we can determine the optimal policy exactly, but this is not possible in general. In a similar setting, [4] state that switching-curve strategies are optimal (a proof will be included in a forthcoming paper by the authors of [4]). Numerical experiments included for illustration in the present paper indeed support this optimality. In order to find computable approximations for the optimal policies we then study the model in a fluid-limit regime for which we show that the optimal strategy is characterized by a linear switching curve. The optimal switching curves in the fluid regime can be used to determine asymptotically fluid optimal strategies for the stochastic model. These policies are characterized by either linear or exponential switching curves. Our analysis is inspired by that in [13, 14] where a multi-class tandem-network is studied. By simulations we compare these asymptotically fluid optimal switching-curve strategies with threshold-based policies [5, 6] and Max-Weight policies [20, 25] which are known to be optimal in heavy traffic. We show that the fluid based strategies give good performance in general and can achieve significant improvements over Max-Weight policies. Optimal threshold-based policies are equally competitive, while the choice of good parameters in practical settings is less involved for the fluid-based strategies.

It is worth noting that the optimal strategies studied in this paper require centralized control. Our aim here is to provide theoretical optimality bounds against which decentralized schemes, implemented in practice, can be compared. In the application area of bandwidth-sharing networks, it was found that certain distributed schemes may actually be close to the theoretical (centralized) optimum [27, 28, 29]. While computationally more involved, threshold-based strategies can be implemented in a distributed manner.

The paper is organized as follows. In Section 2 we describe the model and state some preliminary results. Section 3 contains our optimality results for the stochastic model. We first consider the case when a stochastic optimal strategy exists. Otherwise we resort to optimality in terms of mean numbers of users. The fluid analysis and the asymptotically fluid optimal policies are presented in Section 4. For comparison we briefly discuss the optimal strategies in heavy traffic using the results of [5, 6] and [20, 25] in Section 5. Numerical experiments and concluding remarks can be found in Sections 6 and 7.

2 Model description

We consider the following model. Class-i users, $i = 1, 2$, arrive according to independent Poisson processes with rate λ_i and have exponentially distributed service requirements with mean $1/\mu_i$, $i = 1, 2$. We assume throughout the paper that $\mu_1 \geq \mu_2$, that is the class-1 users are relatively small. Denote by $\rho_i = \frac{\lambda_i}{\mu_i}$. At any time, the server can serve one class only (either class 1 or class 2) with service rate 1, or serve classes 1 and 2 in parallel with service rates c_1 and c_2 respectively, $c_i \leq 1$, or take any convex combination of these three (in a time-sharing fashion). For a given policy π, denote by $s_i^\pi(t)$ the capacity given to class i at time t. The vector $s^\pi(t) = (s_1^\pi(t), s_2^\pi(t))$ lies in the convex hull of the set $\{(0, 0), (1, 0), (0, 1), (c_1, c_2)\}$, see Figure 1. Note that if $c_1 + c_2 > 1$, the total service rate is largest when both classes are served in

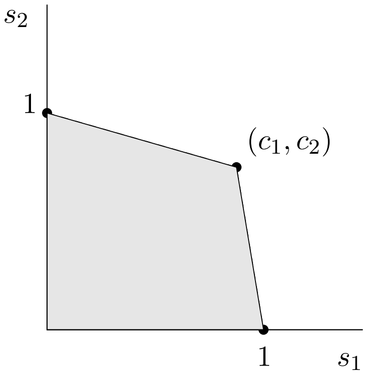

Figure 1: Rate region.

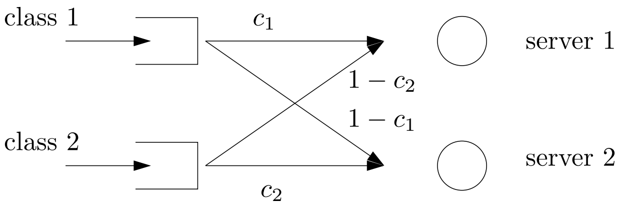

Figure 2: Parallel two-server model.

parallel. For application in wireless networks, this represents the joint capacity when both stations transmit in parallel, and in computer scheduling it corresponds to dedicated specialized servers.

With resource allocation in computer systems in mind, it is more natural to view the model as an equivalent parallel server model with two servers and two queues, as depicted in Figure 2. Server 1 can serve class 1 at rate c_1 or class 2 at rate $1 - c_2$. Similarly, server 2 can serve class 2 at rate c_2 or class 1 at rate $1 - c_1$. When the two servers are dedicated to their own classes, classes 1 and 2 are served in parallel at rates c_1 and c_2 respectively. When both servers serve class i, class i is served at rate 1. (In our setting, both servers can work together on one single task, thus achieving a service capacity of 1, even when there's only one task in the system.) Note that, although uncommon in this setting, it is no restriction to require that the service capacity obtained by combining the two servers equals 1 irrespective of the queue being served. In fact, this can be achieved for any parallel server model by normalizing the service requirements. [1]

In the remainder of the paper we will frequently use the terminology corresponding with the wireless network application and we will not specifically distinguish between two servers.

Let $S_i^\pi(t) := \int_0^t s_i^\pi(u)\mathrm{d}u$ denote the cumulative amount of capacity obtained by class i in the time interval $(0,t)$. Let $A_i(u,t)$ be the amount of class-i work that arrived in the time interval (u,t). Then, the workload in class i at time t can be written as

$$W_i^\pi(t) := W_i(0) + A_i(0,t) - S_i^\pi(t). \tag{1}$$

Denote by $N_i^\pi(t)$ the number of class-i users at time t, and let $N^\pi(t) = (N_1^\pi(t), N_2^\pi(t))$. We further define N_i^π and N^π as random variables with the corresponding steady-state distributions (when they exist).

At all times, one needs to decide how the service capacity should be allocated between the two classes. The objective of the paper is to identify service allocation policies that in some appropriate sense minimize the total number of users in the system. We only consider non-anticipating policies, i.e., policies that have no knowledge available of the remaining service requirements. Since we have exponentially distributed service requirements, the way the capacity is allocated within a class does not affect the stochastic behavior of the process. From now we therefore assume that within a class the First Come First Served discipline is applied. The set of non-anticipating policies is denoted by Π. We call a policy $\tilde{\pi}$ average optimal when $\tilde{\pi} = \mathrm{argmin}_{\pi\in\Pi}\mathbb{E}(N_1^\pi + N_2^\pi)$. A policy $\tilde{\pi}$ is stochastically optimal when $N_1^{\tilde{\pi}}(t) + N_2^{\tilde{\pi}}(t) \leq_{st} N_1^\pi(t) + N_2^\pi(t)$, for all $t \geq 0, \pi \in \Pi$, whenever $N^{\tilde{\pi}}(0) = N^\pi(0)$. By definition, $X \leq_{st} Y$ when $\mathbb{P}(X > s) \leq \mathbb{P}(Y > s)$ for all $s \geq 0$.

[1] One may think of μ_i to be the job completion rate of class i when served exclusively (with normalized service capacity 1). Then c_1 and c_2 may be adjusted so that $\mu_i c_i$ equals the job completion rate of class i when the two classes are served simultaneously.

Note that, if $c_1+c_2 \leq 1$, then serving either class 1 or class 2 exclusively will maximize the rate at which the total workload in the system decreases and since $\mu_i \geq c_1\mu_1 + c_2\mu_2$ it is readily proved that the optimal policy never serves both classes in parallel. The model becomes a multi-class server for which it is well known that when class-i users have exponentially distributed service requirements with mean $1/\mu_i$, the μ-rule, i.e. the policy that gives preemptive priority to the class with the highest departure rate μ_i, is stochastically optimal [24]. The rationale behind this rule is that it maximizes the departure rate at all times. These arguments imply the following lemma. (In fact, the result holds for any shape of the rate region where the points $(1, 0)$ and $(0, 1)$ are not dominated by any other point in the rate region.)

Lemma 2.1 *If $c_1 + c_2 \leq 1$, then the policy that gives preemptive priority to the class with the highest departure rate μ_i is stochastically optimal.*

In the remainder of the paper we will focus on the case $c_1 + c_2 > 1$.

2.1 Stability

The stability condition depends on the policy being used. Maximum stability is obtained by the policy that serves classes 1 and 2 combined whenever possible, since this policy minimizes the total *workload* in the system at every moment in time ($c_1 + c_2 > 1$). Under this policy, the model becomes a coupled processors model for which the stability conditions are

$$\min(\frac{\rho_1}{c_1}, \frac{\rho_2}{c_2}) < 1 \quad \text{and} \tag{2}$$

$$\text{if} \quad \frac{\rho_i}{c_i} < 1 \quad \text{then} \quad \rho_j + \frac{\rho_i}{c_i}(1 - c_j) < 1, \; i \neq j, \tag{3}$$

as proved in [12]. Obviously, for the policy that serves classes 1 and 2 in parallel whenever possible, condition (2) is needed for stability, since otherwise the backlog in both classes will grow indefinitely. If $\frac{\rho_i}{c_i} < 1$, then class i will eventually be drained to zero. If the positive backlog of class j ($j \neq i$) persists for a long time (so that class i gets capacity c_i when present), class i will on average be non-empty for a fraction $\frac{\rho_i}{c_i}$ of the time, so that class j obtains (on average) a capacity of $\frac{\rho_i}{c_i} \cdot c_j + (1 - \frac{\rho_i}{c_i}) \cdot 1$. This being strictly larger than the offered load of class j, ρ_j (because of condition (3)), the class j must ultimately also empty.

Conditions (2) and (3) are necessary conditions to make the system stable for *any* policy. However, they do not guarantee stability and the exact (sufficient and necessary) stability conditions depend strongly on the used scheduling policy.

3 Optimality results

Motivated by the μ-rule, see the previous section, one might expect that such a rule is optimal in our model as well. Note that the departure rate corresponding to a certain allocation $s(t)$ is equal to $\mu_1 s_1(t) + \mu_2 s_2(t)$, hence the μ-rule would amount to choosing that $s(t)$ that maximizes this term. However, the total service capacity, $s_1(t) + s_2(t)$, depends on the chosen allocation as well. For example serving class i only, decreases the total amount of work at rate 1, while serving classes 1 and 2 in parallel implies a decrease of the workload at rate $c_1 + c_2 > 1$. In conclusion, the objective to maximize the departure rate may be conflictive with that of maximizing the total service capacity. The latter will minimize the total time needed to empty the system, which is optimal in the long run, while the former is better for the short run.

Recall that $\mu_1 \geq \mu_2$. If in addition $\mu_1 \leq \mu_1 c_1 + \mu_2 c_2$, then there is no trade-off and it is intuitively clear that the policy that always serves class 1 and 2 in parallel is optimal since this maximizes

both the speed at which the system works and the departure rate. In Section 3.1 we show that the above described policy is in fact stochastically optimal.
When $\mu_1 \geq \mu_1 c_1 + \mu_2 c_2$, we obtain the highest departure rate when we serve class 1 individually. So it may be better to sometimes serve class 1 individually even if that does not maximize the rate at which the total work in the system decreases. Hence as the number of users varies, the system will dynamically switch between different allocations. This setting is discussed in Section 3.2.

3.1 Stochastic optimality when $\mu_1 \leq \mu_1 c_1 + \mu_2 c_2$

In this section we will show that when $(\mu_2 \leq) \, \mu_1 \leq \mu_1 c_1 + \mu_2 c_2$, the policy that always serves both classes in parallel is stochastically optimal. This result can be proved using dynamic programming techniques. We choose a framework which is broader than strictly needed to prove the required stochastic optimality of the number of jobs (we only need a particular choice of the function $C(\cdot)$ below). It will be convenient to focus on the uniformized Markov chain. That is, transition epochs (possibly 'dummy' transitions that do not alter the system state) are generated by a Poisson process of uniform rate $\nu = \lambda_1 + \lambda_2 + (1 + c_1)\mu_1 + (1 + c_2)\mu_2$. Since ν is finite, we may assume $\nu = 1$ without loss of generality. We consider here the embedded discrete-time Markov chain and, for transparency of notation, again denote the state after k steps by $N_i(k), i = 1, 2$. Let $x = (x_1, x_2)$ where the variable x_i represents the number of class-i users. We define the functions $V_k(\cdot, \cdot)$, $k = 0, 1, \ldots$, as follows:

$$\begin{aligned}
V_0(x) &= C(x) \\
V_{k+1}(x) &= \lambda_1 V_k(x_1 + 1, x_2) + \lambda_2 V_k(x_1, x_2 + 1) \\
&\quad + \min\Big(\mu_1 V_k((x_1 - 1)^+, x_2) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2) V_k(x), \\
&\qquad\qquad \mu_2 V_k(x_1, (x_2 - 1)^+) + (\mu_1 + \mu_1 c_1 + \mu_2 c_2) V_k(x), \\
&\qquad\qquad \mu_1 c_1 V_k((x_1 - 1)^+, x_2) + \mu_2 c_2 V_k(x_1, (x_2 - 1)^+) + (\mu_1 + \mu_2) V_k(x)\Big)
\end{aligned} \tag{4}$$

for $x_1, x_2 \geq 0, k = 0, 1, \ldots,$ with $C(\cdot, \cdot)$ a cost function. The term $V_{k+1}(x)$ represents the minimum achievable expected costs after $k + 1$ steps, when the system starts in state x. If we choose as cost function $C(x) = \mathbf{1}_{(x_1 + x_2 > s)}$, then these costs are equal to $V_{k+1}(x) = \mathbb{P}(N_1(k + 1) + N_2(k + 1) > s | N(0) = x)$. Hence, if for this cost function we show that for every s and k we obtain the same minimizing action in (4) (the optimal action may depend on the state x), then the corresponding policy is stochastically optimal. In the next two lemmas we establish convenient properties of V_k, under certain conditions on the function $C(x)$.

Lemma 3.1 *If $C(x)$ is non-decreasing in x_1 and x_2, then $V_k(x)$ is non-decreasing in x_1 and x_2 for all k.*

Proof: The statement follows directly from the definition of V_k. □

Under certain conditions on $C(x)$, the minimizing action will be to always serve classes 1 and 2 in parallel, whenever possible. This is stated in Lemma 3.2 and the proof may be found in Appendix A.

Lemma 3.2 *If $c_1 + c_2 \geq 1$ and $W(x) = C(x)$ is non-decreasing in x_1 and x_2 and satisfies*

$$\begin{aligned}
&(\mu_1 + \mu_2) W(x) + \mu_1 c_1 W(x_1 - 1, x_2) + \mu_2 c_2 W(x_1, x_2 - 1) \\
&\leq \min(\mu_1 W(x_1 - 1, x_2) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2) W(x), \\
&\qquad \mu_2 W(x_1, x_2 - 1) + (\mu_1 + \mu_1 c_1 + \mu_2 c_2) W(x)),
\end{aligned} \tag{5}$$

for $x_1, x_2 > 0$, then $W = V_k$, $k \geq 0$, satisfies (5) as well.

We can now find the stochastically optimal policy when $\max(\mu_1, \mu_2) \leq \mu_1 c_1 + \mu_2 c_2$.

Proposition 3.3 *Assume $c_1 + c_2 \geq 1$. If $\max(\mu_1, \mu_2) \leq \mu_1 c_1 + \mu_2 c_2$, then it is stochastically optimal to serve both classes in parallel whenever possible.*

Proof: If $\max(\mu_1, \mu_2) \leq \mu_1 c_1 + \mu_2 c_2$, then the cost function $C(x_1, x_2) = \mathbf{1}_{(x_1+x_2>s)}, s \geq 0$, satisfies the conditions as given in Lemma 3.2. Hence, from Lemma 3.2 we obtain that serving both classes in parallel whenever possible is always the minimizing action in (4) and hence the corresponding policy is stochastically optimal. □

3.2 Characterization of the average-optimal policy when $\mu_1 > \mu_1 c_1 + \mu_2 c_2$

Now assume that $\mu_1 > \mu_1 c_1 + \mu_2 c_2$. A tradeoff must be made when users of both classes are present. On one hand serving only class 1 maximizes the departure rate since $\mu_1 > \mu_1 c_1 + \mu_2 c_2$. However, serving class 1 and 2 simultaneously maximizes the speed at which the total workload in the system decreases. Since a stochastically optimal policy may in general not exist, we focus on the average-optimal policy, i.e., the policy that minimizes $\mathbb{E}(N_1^\pi + N_2^\pi)$ over all policies $\pi \in \Pi$. From $c_1 + c_2 > 1$ and $\mu_1 > \mu_1 c_1 + \mu_2 c_2$ together we obtain $\mu_1 > \mu_2$. Hence maximizing the departure rate would imply that an optimal policy will never serve class 2 individually when class 1 is also present. At the same time, serving class 2 individually does not give the high service rate either. Therefore, this action will not be chosen by an optimal policy. This fact is proved in Proposition 3.5. First we will state a lemma that in fact holds for generally distributed service requirements and will be used later in the proof of Proposition 3.5. The proof may be found in Appendix B.

Lemma 3.4 *Assume we have generally distributed service requirements. Let $\tilde{\pi}$ be a policy that sometimes does serve class 2 individually while there are class-1 users present. Define policy π to be the policy that uses the same allocation as $\tilde{\pi}$ when possible, except when policy $\tilde{\pi}$ serves class 2 individually. In that case policy π serves classes 1 and 2 in parallel (if possible). Then the following sample-path inequalities hold:*

$$S_1^\pi(t) \geq S_1^{\tilde{\pi}}(t) \tag{6}$$

$$S_1^\pi(t) + S_2^\pi(t) \geq S_1^{\tilde{\pi}}(t) + S_2^{\tilde{\pi}}(t) \tag{7}$$

$$(1 - c_2)S_1^\pi(t) + c_1 S_2^\pi(t) \geq (1 - c_2)S_1^{\tilde{\pi}}(t) + c_1 S_2^{\tilde{\pi}}(t), \tag{8}$$

for all $t \geq 0$.

Proposition 3.5 *Assume $\mu_1 \geq \mu_2$ and $c_1 + c_2 > 1$. Then for any policy $\tilde{\pi}$ that serves class 2 individually when there is work of class 1 present, there exists a modified policy π that never serves class 2 individually and does not worse than $\tilde{\pi}$, i.e., $\mathbb{E}(N_1^\pi(t)+N_2^\pi(t)) \leq \mathbb{E}(N_1^{\tilde{\pi}}(t)+N_2^{\tilde{\pi}}(t))$, for all $t \geq 0$.*

Proof: Let $\tilde{\pi}$ be a policy that sometimes does serve class 2 individually while there are class-1 users present. Define policy π as in Lemma 3.4 and hence the sample-path inequalities (6) and (7) hold. Multiplying (6) by $\mu_1 - \mu_2 \geq 0$ and (7) by μ_2 and adding the two inequalities gives that $\mu_1 S_1^\pi(t) + \mu_2 S_2^\pi(t) \geq \mu_1 S_1^{\tilde{\pi}}(t) + \mu_2 S_2^{\tilde{\pi}}(t)$ and hence by (1) we obtain

$$\mu_1 W_1^\pi(t) + \mu_2 W_2^\pi(t) \leq \mu_1 W_1^{\tilde{\pi}}(t) + \mu_2 W_2^{\tilde{\pi}}(t) \tag{9}$$

for all t. Since we have exponentially distributed service requirements and we consider only non-anticipating policies, we have $\mathbb{E}(W_i^\pi(t)) = \frac{1}{\mu_i}\mathbb{E}(N_i^\pi(t))$. By taking expectations on both sides in (9), we obtain $\mathbb{E}(N_1^{\pi}(t) + N_2^{\pi}(t)) \leq \mathbb{E}(N_1^{\tilde{\pi}}(t) + N_2^{\tilde{\pi}}(t))$. Hence policy π is not worse than $\tilde{\pi}$ and policy π never serves class 2 individually when there is work of class 1 present. □

When seeking for an average-optimal policy, by Proposition 3.5 we only need to consider policies that never serve class-2 users individually when there are also class-1 users present. The decision between whether to serve class 1 individually or classes 1 and 2 in parallel depends on the number of class-1 and class-2 users present in the system. Intuitively, we expect that the optimal policy can be characterized by a switching curve, i.e. there exists a function h such that if $N_2 \geq h(N_1)$, then it is optimal to serve classes 1 and 2 in parallel, and otherwise it is optimal to serve only class 1. The authors in [4] state that for a model with slightly different behavior near the boundaries, the existence of such a switching curve can be proved using dynamic programming techniques. We expect that for our model, the existence of a switching curve can be proved using the same technique (see also [27] where this was done for a different model). However, dynamic programming techniques will not provide us with any information concerning the shape of the curve. Therefore, in the remainder of the paper we search for policies that are close to optimal by investigating two limiting regimes. In Section 4 this is done for a fluid scaled system and (asymptotically) optimal switching curve policies are found. In Section 5 a heavy traffic regime is considered.

4 Fluid analysis and asymptotic optimality

In this section we will consider the stochastic process under a fluid scaling and investigate close to optimal policies. In order to do so, it will be convenient to first study the related deterministic fluid control model. This will be done in Section 4.1 and for this relatively simple model we derive the optimal switching curve and the optimal trajectory. In Section 4.2 we then show that under certain switching curve policies in the stochastic process, the fluid scaled stochastic processes converge to the optimal trajectory as found for the deterministic fluid control model. This tells that these switching curve policies are asymptotically fluid optimal in the stochastic model.

4.1 Optimal policies for the fluid control model

In this section we consider the deterministic fluid control model, which arises from the original stochastic model by only taking into account the mean drifts of the stochastic model. Denote by $n_i(t)$ the amount of class-i fluid in the system at time t and let $n(t) = (n_1(t), n_2(t))$. The fluid processes $n_i(t)$ are described by the following differential equations:

$$\frac{\mathrm{d}n_i(t)}{\mathrm{d}t} = \lambda_i - u_i(t)\mu_i - u_c(t)\mu_i c_i, \;\; i = 1, 2, \tag{10}$$

$$n_i(t) \geq 0, \;\; i = 1, 2, \tag{11}$$

$$u_1(t) + u_2(t) + u_c(t) \leq 1 \tag{12}$$

$$u_j(t) \geq 0, \;\; j = 1, 2, c, \tag{13}$$

and

$$n(0) = n. \tag{14}$$

Note that when (2) and (3) are satisfied, the system can be drained in finite time and can be kept empty from that moment on.

A policy π for the fluid control model is described by the control functions $u_1^\pi(t),\ u_2^\pi(t)$ and $u_c^\pi(t)$. We are interested in the optimal fluid policy, which is defined as the policy that minimizes

$$\int_0^D (n_1^\pi(t) + n_2^\pi(t))\mathrm{d}t, \quad \text{with } n^\pi(t) \text{ satisfying (10)–(13),} \tag{15}$$

for every initial point $n(0) = n$. In this section we focus on finding the optimal policy for $D = \infty$. For $D = \infty$ we denote the optimal trajectory by $n^*(t)$ and the optimal control by $u_j^*(t), j = 1, 2, c$.
Before proceeding to find n^* and u_j^*, we will first state a lemma that will be useful later on. It states that $n^*(t)$, the optimal trajectory for the infinite horizon problem, is also optimal for the finite horizon problem when the horizon is large enough. The proof may be found in Appendix C.

Lemma 4.1 *There exists a function $D : \mathbb{R} \to \mathbb{R}$ such that,*

$$\min_{n(t)\ s.t.\ (10)-(14)} \int_0^{\tilde{D}} (n_1(t) + n_2(t))\mathrm{d}t = \int_0^{\tilde{D}} (n_1^*(t) + n_2^*(t))\mathrm{d}t,$$

for all $\tilde{D} \geq D(|n|)$ and with $n^*(t)$ the optimal solution of (15) for $D = \infty$ and initial state n.

For the stochastic model we know that it is never optimal to serve class 2 exclusively when also work of class 1 is present. In the fluid control model, it can be checked that this is true as well:

Observation 4.2 *If $n_1 > 0$, then $u_2^*(t) = 0$.*

The following lemma describes the existence of a switching curve in the fluid control model.

Lemma 4.3 *Assume $\mu_1 > c_1\mu_1 + c_2\mu_2$ and $c_1 + c_2 > 1$. Let $\tilde{n} \in \{n : n_1 > 0, n_2 \geq 0\}$ and $\hat{n} \in \{n : n_1 \geq 0, n_2 \geq 0\}$ and assume there exists a trajectory between these two points that does not coincide with the $n_1 = 0$ axis. Then among all trajectories that move from $\tilde{n}$ to $\hat{n}$ without coinciding with the $n_1 = 0$ axis, the path that first serves class 1 and at some point switches to serving both classes 1 and 2 simultaneously, minimizes the costs to go from $\tilde{n}$ to $\hat{n}$.*

Proof: Since $n_1 > 0$ on the whole trajectory, we know that no time is spent on serving class 2 individually. Denote by T_1 (T_c) the cumulative amount of time spend on serving class 1 (classes 1 and 2 simultaneously). The net change in the amount of fluid in the two classes can be written as

$$\begin{aligned} \hat{n}_1 - \tilde{n}_1 &= (\lambda_1 - \mu_1)T_1 + (\lambda_1 - c_1\mu_1)T_c \\ \hat{n}_2 - \tilde{n}_2 &= \lambda_2 T_1 + (\lambda_2 - c_2\mu_2)T_c. \end{aligned}$$

Under the necessary stability conditions (2) and (3) this has a unique solution for T_1 and T_c. Hence, all trajectories spend the same cumulative amount of time on serving both classes in parallel and serving class 1 individually.
The rate at which the total amount of fluid decreases when $n_1(t) > 0$ is given by $\frac{\mathrm{d}(n_1(t)+n_2(t))}{\mathrm{d}t} = \lambda_1 + \lambda_2 - u_1(t)\mu_1 - u_c(t)(\mu_1 c_1 + \mu_2 c_2)$. Since $\mu_1 > \mu_1 c_1 + \mu_2 c_2$, first serving only class 1 initially maximizes the rate at which the total amount of fluid decreases, and hence minimizes the costs.
□

For the fluid control model we can now determine the optimal switching curve. To do that, we will distinguish between whether $\rho_1 < c_1$ or $\rho_1 \geq c_1$.

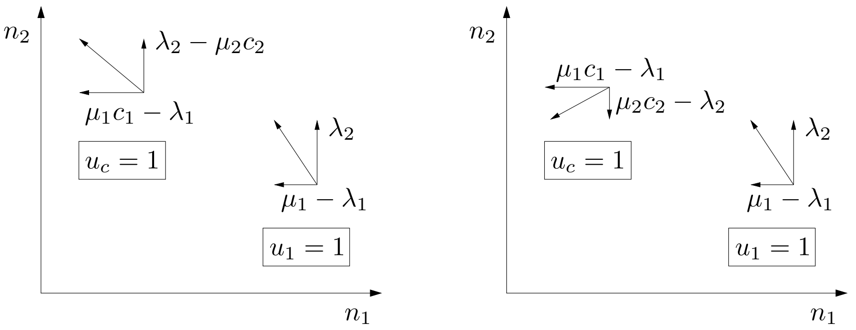

Figure 3: Drift vectors for $\rho_1 < c_1$ and $\rho_2 > c_2$ (left), and $\rho_1 < c_1$ and $\rho_2 < c_2$ (right), respectively.

4.1.1 Case $\rho_1 < c_1$

When $\rho_1 < c_1$, a necessary condition for stability is $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$ (see (2) and (3)). Depending on ρ_2 and c_2, the drifts are as in Figure 3.

In Proposition 4.4 we describe the optimal fluid policy which is characterized by a linear switching curve. In Figure 4 the optimal trajectory is shown. To state the proposition it is convenient to define

$$\alpha := \max\Big(0,\ \frac{c_2 - \rho_2}{c_1 - \rho_1} + \frac{c_1}{c_1 + c_2 - 1} \times \frac{1 - \rho_2 - \frac{\rho_1}{c_1}(1 - c_2)}{c_1 - \rho_1} \times \frac{\mu_1 - c_1\mu_1 - c_2\mu_2}{\mu_2}\Big).$$

Note that $\alpha > \frac{c_2 - \rho_2}{c_1 - \rho_1}$.

Proposition 4.4 *Let $\mu_1 > \mu_1 c_1 + \mu_2 c_2$ and $c_1 + c_2 > 1$. Assume $\rho_1 < c_1$ and $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$. The optimal policy in the fluid control model is*

- $u_1^* = 1$, *if $n_2 < \alpha \frac{\mu_2}{\mu_1} n_1$.*
- $u_c^* = 1$, *if $n_2 \geq \alpha \frac{\mu_2}{\mu_1} n_1$ and $n_1 > 0$.*
- $u_c^* = \frac{\rho_1}{c_1}$ *and $u_2^* = 1 - \frac{\rho_1}{c_1}$, if $n_1 = 0$.*

Proof: We first determine the optimal allocation for points with $n_1 = 0$. From the fluid dynamics (10)-(13), Observation 4.2 and noting that $\rho_1 < c_1$, we see that when $n_1(t) = 0$, then $\frac{\mathrm{d}n_1(t)}{\mathrm{d}t} = \lambda_1 - u_1^*(t)\mu_1 - u_c^*(t)\mu_1 c_1 \leq 0$, hence class 1 remains empty. So $\frac{\mathrm{d}n_1(t)}{\mathrm{d}t} = 0$, i.e., $\rho_1 - u_1^*(t) - u_c^*(t)c_1 = 0$. The optimal fluid policy will now maximize the departure rate of class 2. So maximize $u_2(t)\mu_2 + u_c(t)\mu_2 c_2$ given that $\rho_1 - u_1(t) - u_c(t)c_1 = 0$, $u_1(t) + u_2(t) + u_c(t) = 1$ and $u_j(t) \geq 0$. Solving this we obtain

$$u_c^* = \frac{\rho_1}{c_1},\ u_1^* = 0 \ \text{ and } \ u_2^* = 1 - \frac{\rho_1}{c_1}, \quad \text{when } \ n_1 = 0.$$

Now assume we start at time $t = 0$ in $n(0) = n \equiv (n_1, n_2)$ with $n_1 > 0$ and $n_2 \geq 0$. At some point the optimal trajectory will hit the n_1=0 axis for the first time. This point will be denoted by $d = (d_1, d_2)$, see Figure 4. By Lemma 4.3, the trajectory until d is known. Namely, first class 1 is served, and at some point the policy switches to serving both classes simultaneously. The point where this switch occurs, is denoted by $b = (b_1, b_2)$, the turning point, see Figure 4. We can calculate the costs corresponding to a certain turning point b. Let $T(x, y)$ be the time it takes to go from point x to y in the (n_1, n_2)-plane. We have $T(n, b) = \frac{n_1 - b_1}{\mu_1 - \lambda_1}$, $T(b, d) = \frac{b_1}{\mu_1 c_1 - \lambda_1}$, $T(d, 0)$

$= \frac{d_2}{u_2\mu_2+u_c\mu_2c_2-\lambda_2} = \frac{d_2}{\mu_2-\mu_2\frac{\rho_1}{c_1}(1-c_2)-\lambda_2}$, with $b_2 = n_2 + T(n,b)\lambda_2$ and $d_2 = b_2 + T(b,d)(\lambda_2 - \mu_2 c_2)$. Let $K_n(b_1) = \int_0^\infty (n_1(t) + n_2(t))\mathrm{d}t$ be the costs of the fluid trajectory going from n to the origin when the turning point is b. We have

$$K_n(b_1) = T(n,b)\Big(\frac{n_1+b_1}{2} + \frac{n_2+b_2}{2}\Big) + T(b,d)\Big(\frac{b_1}{2} + \frac{b_2+d_2}{2}\Big) + T(d,0)\frac{d_2}{2}. \qquad (16)$$

It can be checked that the function $K_n(b_1)$ is a quadratic function in b_1 and when minimizing the costs in (16), the optimal turning point b lies on the line $b_2 = \alpha\frac{\mu_2}{\mu_1}b_1$. Hence, if $n_2 < \alpha\frac{\mu_2}{\mu_1}n_1$, then $u_1^* = 1$, and if $n_2 \geq \alpha\frac{\mu_2}{\mu_1}n_1$ and $n_1 > 0$, then $u_c^* = 1$. □

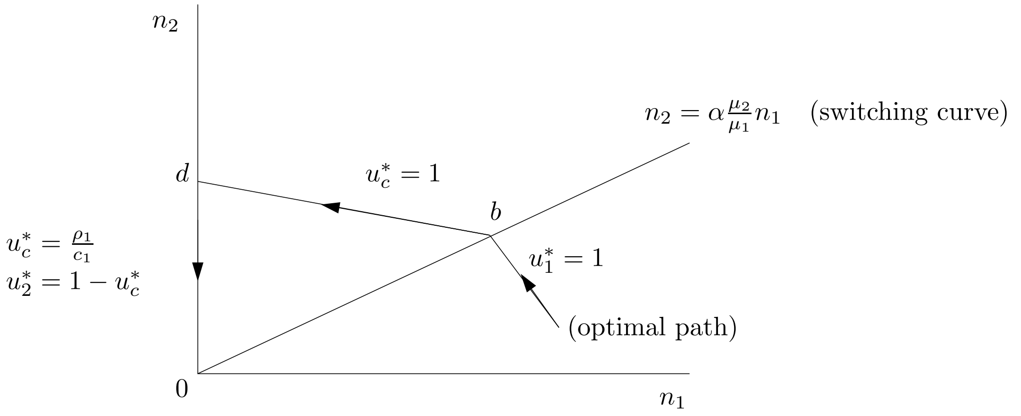
n_2
$n_2 = \alpha\frac{\mu_2}{\mu_1}n_1$ (switching curve)
d
$u_c^* = 1$
b
$u_c^* = \frac{\rho_1}{c_1}$
$u_2^* = 1 - u_c^*$
$u_1^* = 1$
(optimal path)
0
n_1

Figure 4: Optimal trajectory of the fluid control model when $\rho_1 < c_1$.

4.1.2 Case $\rho_1 \geq c_1$

When $\rho_1 \geq c_1$, the necessary stability condition is $\rho_2 < c_2$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1-c_1)$ (see (2) and (3)). Hence $\frac{\rho_2}{1-\rho_1} \leq \frac{c_2-\rho_2}{\rho_1-c_1}$ and the drifts are as in the left picture in Figure 5. When $\rho_1 \geq 1 - \frac{\rho_2}{c_2}(1-c_1)$, the system is unstable which corresponds to the right picture in Figure 5. The optimal fluid policy is described in the next proposition, and in Figure 6 the optimal trajectory is shown.

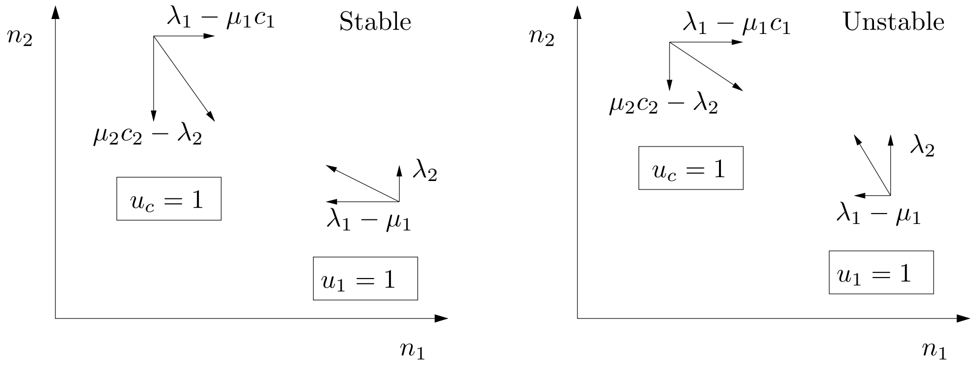

Figure 5: Vectors for $\rho_1 \geq c_1$ and $\rho_2 < c_2$. Left figure: $\rho_1 < 1 - \frac{\rho_2}{c_2}(1-c_1)$ and hence there are policies that give a stable system. Right figure: $\rho_1 > 1 - \frac{\rho_2}{c_2}(1-c_1)$ and hence unstable.

Proposition 4.5 *Let $\mu_1 > \mu_1 c_1 + \mu_2 c_2$ and $c_1 + c_2 > 1$. Assume $\rho_1 \geq c_1$, $\rho_2 < c_2$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1 - c_1)$. The optimal policy in the fluid control model is to give priority to class 1, i.e.,*

- $u_1^* = 1$ *if* $n_1 > 0$.
- $u_c^* = \frac{1-\rho_1}{1-c_1}$ *and* $u_1^* = \frac{\rho_1 - c_1}{1-c_1}$ *if* $n_1 = 0$.

Proof: By similar arguments as in the proof of Proposition 4.4, it can be shown that any turning point b with $b_1 > 0$ does worse than the policy that gives priority to class 1 until $n_1 = 0$. Hence $u_1^*(t) = 1$ when $n_1(t) > 0$.
Once $n_1(t) = 0$, we can conclude from the above that class 1 will remain empty since $\rho_1 < 1$. Hence we have $u_1^*(t) + u_c^*(t)c_1 = \rho_1$ when $n_1(t) = 0$. Now choose the allocations $u_j^*(t)$ such that the departure rate for class 2, $u_2(t)\mu_2 + u_c(t)\mu_2 c_2$, is maximized subject to $u_1(t) + u_c(t)c_1 = \rho_1$, $u_1(t) + u_2(t) + u_c(t) = 1$ and $u_j(t) \geq 0$. The solution to this is $u_2^*(t) = 0$, $u_1^*(t) = \frac{\rho_1 - c_1}{1-c_1}$ and $u_c^*(t) = \frac{1-\rho_1}{1-c_1}$ when $n_1(t) = 0$. □

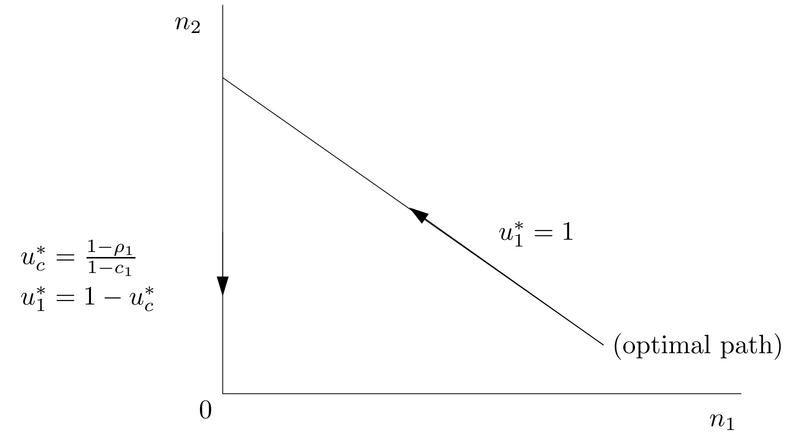

Figure 6: Optimal trajectory of the fluid control model when $\rho_1 \geq c_1$.

4.2 Asymptotic optimality

In this section we discuss the theoretical foundations that justify the use of the optimal policies from the fluid model as proxies of the optimal policies in the stochastic model. In particular, we prove that the stochastic processes of fluid scaled number of users under certain switching curve policies converge to the optimal fluid trajectories $n^*(t)$ as determined in Section 4.1. Using the latter we then show that these switching curve policies are asymptotically fluid optimal in the stochastic model. The analysis and terminology used in this section is motivated by [2, 13, 19, 21].
We consider a sequence of systems indexed by a superscript r. Let $N_i^r(t)$ be the number of class-i users at time t in the r-th system. The initial queue lenghts depend on r such that $N^r(0) = n^r$ with $\lim_{r\to\infty} \frac{n_i^r}{r} = n_i, i = 1, 2$. For the r-th system, during the time interval $(0, t)$, $T_0^r(t)$ is the cumulative amount of time that both queues are empty, $T_i^r(t)$ is the cumulative amount of time that was spent on serving class i only, $i = 1, 2$, and $T_c^r(t)$ the cumulative amount of time that was spent on serving classes 1 and 2 in parallel. Then $T_0^r(t) + T_1^r(t) + T_2^r(t) + T_c^r(t) = t$, and

$$N_i^r(t) = N_i^r(0) + E_i(t) - F_i(T_i^r(t)) - F_{c,i}(T_c^r(t)), \quad i = 1, 2, \tag{17}$$

with $E_i(t)$ a Poisson process with rate λ_i, $F_i(\cdot)$ a Poisson process with rate μ_i and $F_{c,i}(\cdot)$ a Poisson process with rate $c_i\mu_i$.
We will be interested in the processes under the fluid scaling, i.e., both time and space are scaled linearly:

$$\overline{N}_i^r(t) := \frac{N_i^r(rt)}{r} \quad \text{and} \quad \overline{T}_j^r(t) := \frac{T_j^r(rt)}{r}.$$

Limit points for $\overline{N}_i^r(t)$ and $\overline{T}_j^r(t)$ are described in the next lemma.

Lemma 4.6 *If* $\lim_{r\to\infty} \frac{n_i^r}{r} = n_i$, $i = 1, 2$, *then for almost all sample paths* ω *there exists a subsequence* r_k *such that*

$$\begin{aligned}
&\lim_{r_k\to\infty} \overline{T}_j^{r_k}(t) = \overline{T}_j(t), \quad j = 1, 2, c, \quad \textit{u.o.c.} \\
&\lim_{r_k\to\infty} \overline{N}_i^{r_k}(t) = \overline{N}_i(t) = n_i + \lambda_i t - \mu_i \overline{T}_i(t) - \mu_i c_i \overline{T}_c(t), \quad \textit{u.o.c.} \qquad (18)
\end{aligned}$$

where $\overline{N}_i(t) \geq 0$, $\overline{T}_j(0) = 0$, $\overline{T}_1(t) + \overline{T}_2(t) + \overline{T}_c(t) + \overline{T}_0(t) = t$ *and* $\overline{T}_j(t)$ *are non-decreasing and Lipschitz continuous functions.*

The notation u.o.c. stands for uniform convergence on compact sets. The processes $\overline{T}_j(t), j = 1, 2, c$, and $\overline{N}_i(t), i = 1, 2$, as obtained in Lemma 4.6 are by definition fluid limits for initial fluid level n. Note that the functions $\overline{T}_j(t)$, $j = 1, 2, c$, depend on the chosen policy in the stochastic model.

Proof of Lemma 4.6: Making use of (17) and the fact that $\overline{T}_j^r(t), j = 1, 2, c$, is Lipschitz continuous with a constant less than or equal to 1, the proof follows similar to the proof in [11, Theorem 4.1], see also [5, Lemma 9.2]. □

As costs in the stochastic model we take $\mathbb{E}_n^\pi \int_0^D (N_1(t) + N_2(t))\mathrm{d}t$, with $\mathbb{E}_n^\pi$ the expectation when starting in $N(0) = n$ and using policy π. For the sequence of processes we consider, the costs $\mathbb{E}_{n^r}^\pi \int_0^D (N_1^r(t) + N_2^r(t))\mathrm{d}t$ will tend to infinity. In order to obtain a non-trivial limit we divide the costs by r^2 and consider a horizon that grows linearly in r. So we are interested in

$$\mathbb{E}_{n^r}^\pi \int_0^{r\cdot D} \frac{N_1^r(t) + N_2^r(t)}{r^2}\mathrm{d}t = \mathbb{E}_{n^r}^\pi \int_0^D \frac{N_1^r(rt) + N_2^r(rt)}{r}\mathrm{d}t = \mathbb{E}_{n^r}^\pi \int_0^D (\overline{N}_1^r(t) + \overline{N}_2^r(t))\mathrm{d}t. \quad (19)$$

Our goal is to find policies that minimize the costs (19) as $r \to \infty$. We have the following lower bound.

Lemma 4.7 *If* $\lim_{r\to\infty} \frac{n_i^r}{r} = n_i$, $i = 1, 2$, *then for any* π *we have*

$$\liminf_{r\to\infty} \mathbb{E}_{n^r}^\pi \Big(\int_0^D (\overline{N}_1^r(t) + \overline{N}_2^r(t))\mathrm{d}t \Big) \geq \int_0^\infty (n_1^*(t) + n_2^*(t))\mathrm{d}t, \quad \textit{with} \quad D > D(|n|)$$

and $n^*(t)$ *the optimal solution of (15) for initial state* n.

Proof: By applying Fatou's lemma twice, we obtain

$$\liminf_{r\to\infty} \mathbb{E}_{n^r}^\pi \Big(\int_0^D (\overline{N}_1^r(t) + \overline{N}_2^r(t))\mathrm{d}t \Big) \quad \geq \quad \mathbb{E}^\pi \Big(\int_0^D \liminf_{r\to\infty} \; (\overline{N}_1^r(t) + \overline{N}_2^r(t)) \; \mathrm{d}t \Big).$$

For almost all ω, we have that $\liminf_{r\to\infty} \; \overline{N}_1^r(t) + \overline{N}_2^r(t)$, is a fluid limit for initial fluid level

n (apply Lemma 4.6 to a subsequence that reaches the liminf), and is therefore an admissible trajectory for the fluid control problem (15).
When we choose $D \geq D(|n|)$, we know from Lemma 4.1 that n^* solves the minimization problem (15), and hence

$$\mathbb{E}^{\pi}\Big(\int_0^D \liminf_{r\to\infty}\ (\overline{N}_1^r(t) + \overline{N}_2^r(t))\ \mathrm{d}t\Big) \geq \int_0^\infty (n_1^*(t) + n_2^*(t))\mathrm{d}t.$$

This proves the lemma. □

We say that a policy is asymptotically fluid optimal when the lower bound is obtained, i.e., when the scaled costs under the policy converge to the costs of the optimal trajectory in the fluid model. In the remainder of this section we will characterize these asymptotically fluid optimal policies.

Definition 4.8 *A stationary policy π^* is called asymptotically fluid optimal if for any sequence n^r such that $\lim_{r\to\infty} \frac{n_i^r}{r} = n_i,\ i = 1,2,$ we have*

$$\lim_{r\to\infty} \mathbb{E}_{n^r}^{\pi^*}\Big(\int_0^D (\overline{N}_1^r(t) + \overline{N}_2^r(t))\mathrm{d}t\Big) = \int_0^\infty (n_1^*(t) + n_2^*(t))\mathrm{d}t, \quad \textit{with}\ \ D \geq D(n)$$

where $n^*(t)$ is the optimal solution of (15) for initial state n.

4.2.1 Case $\rho_1 < c_1$

In this section we consider the case $\rho_1 < c_1$. In Proposition 4.4 we found that the optimal switching curve for the fluid control problem was given by $h(n_1) = \alpha\frac{\mu_2}{\mu_1}n_1$. In the next proposition it is stated that the same switching curve provides a policy that is asymptotically fluid optimal for the original stochastic model.

Proposition 4.9 *Let $\mu_1 > \mu_1 c_1 + \mu_2 c_2$ and $c_1 + c_2 > 1$. If $\rho_1 < c_1$ and $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$, then the stationary policy π^* with switching curve $h(N_1) = \alpha\frac{\mu_2}{\mu_1}N_1$ is asymptotically fluid optimal.*

Proof: Any fluid limit of policy π^* satisfies the following. The functions $\overline{T}_j(t)$ are absolutely continuous (follows from Lipschitz continuity), hence are differentiable almost everywhere. It can be proved (see Appendix D) that for each regular point t the derivatives satisfy:

$$\frac{\mathrm{d}\overline{T}_1(t)}{\mathrm{d}t} = 1,\ \text{if}\ \ \overline{N}_2(t) < \alpha\frac{\mu_2}{\mu_1}\overline{N}_1(t), \tag{20}$$

$$\frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t} = 1,\ \text{if}\ \ \overline{N}_2(t) \geq \alpha\frac{\mu_2}{\mu_1}\overline{N}_1(t)\ \text{and}\ \overline{N}_1(t) > 0, \tag{21}$$

$$\frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t} = \frac{\rho_1}{c_1}\ \text{and}\ \frac{\mathrm{d}\overline{T}_2(t)}{\mathrm{d}t} = 1 - \frac{\rho_1}{c_1},\ \ \text{if}\ \ \overline{N}_1(t) = 0\ \text{and}\ \overline{N}_2(t) > 0. \tag{22}$$

and $\frac{\mathrm{d}\overline{T}_1(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_2(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_0(t)}{\mathrm{d}t} = 1.$
Taking $u_j(t) = \frac{\mathrm{d}\overline{T}_j(t)}{\mathrm{d}t}$, and from (20)–(22), we see that

$$\overline{N}(t) = n^*(t), \tag{23}$$

with n^* as defined in Proposition 4.4. From this it then follows that under policy π^* we have $\limsup_{r\to\infty} \mathbb{E}_{n^r}^{\pi^*}(\int_0^D (\overline{N}_1^r(t) + \overline{N}_2^r(t))\mathrm{d}t) \leq \int_0^\infty (n_1^*(t) + n_2^*(t))\mathrm{d}t$ (see Appendix D). Together with Lemma 4.7 we then obtain that the fluid scaled cost function under policy π^* converges to the costs of an optimal trajectory in the fluid model, and hence that π^* is asymptotically optimal.
□

4.2.2 Case $\rho_1 > c_1$

In this section we consider the case $\rho_1 > c_1$. In Proposition 4.5 we found that for the fluid control problem it is then optimal to give class 1 priority when $n_1 > 0$. A straight-forward translation of this policy to the original stochastic model would be to give preemptive priority to class-1 users. However, the stability conditions under this policy are $\rho_1 + \rho_2 < 1$, which are more strict than the maximum stability conditions as given in (2) and (3). Hence, a more precise interpretation of the fluid optimal policy is needed to avoid an unstable system.
Note that the optimal policy in the fluid control model does keep the system stable under (2) and (3) since on the vertical axis the fluid model partly serves class 1 individually and partly serves both classes in parallel. This suggest that for the stochastic model we should also sometimes serve classes 1 and 2 in parallel, especially when we come too close to the vertical axis. This implies that the switching curve for the original model lies close to the $N_1 = 0$ axis and is in fact non-observable in the fluid limit. In the next proposition we state that a policy with a switching curve of the shape $h(N_1) = e^{N_1/\gamma}$ is an asymptotically fluid optimal policy.

Proposition 4.10 *Let $\mu_1 > \mu_1 c_1 + \mu_2 c_2$ and $c_1 + c_2 > 1$. Assume $\rho_1 > c_1$, $\rho_2 < c_2$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1 - c_1)$. The stationary policy π^* with switching curve $h(N_1) = e^{N_1/\gamma}$ is asymptotically fluid optimal for $\gamma > 0$ large enough.*

Sketch of proof: Any fluid limit of policy π^* satisfies the following. The functions $\overline{T}_j(t)$ are absolutely continuous and using the same techniques as in [13, Section 7] it follows that for each regular point t the derivatives satisfy:

$$\frac{\mathrm{d}\overline{T}_1(t)}{\mathrm{d}t} = 1, \quad \text{if} \quad \overline{N}_1 > 0, \tag{24}$$

$$\frac{\mathrm{d}\overline{T}_1(t)}{\mathrm{d}t} = \frac{\rho_1 - c_1}{1 - c_1} \text{ and } \frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t} = \frac{1 - \rho_1}{1 - c_1}, \quad \text{if} \quad \overline{N}_1 = 0, \tag{25}$$

for γ large enough. Note that $\frac{\mathrm{d}\overline{T}_2(t)}{\mathrm{d}t} = 0$, which implies that the process does not stay long on the N_1=0 axis and therefore any capacity lost, when serving class 2 only, is negligible under fluid scaling.
Taking $u_j(t) = \frac{\mathrm{d}\overline{T}_j(t)}{\mathrm{d}t}$, and from (18), (24) and (25), we see that $\overline{N}(t) = n^*(t)$, with n^* as defined in Proposition 4.5. The remainder of the proof is similar to the proof of Proposition 4.9. □

4.3 Asymptotically optimal strategies with an exponential switching curve

When the asymptotically fluid optimal strategies have a linear switching curve, the slope of the curve has been exactly determined. For exponentially shaped optimal switching curves, the parameter γ in the exponent is yet unknown. In fact Proposition 4.10 proves that an exponential switching curve $h(N_1) = e^{N_1/\gamma}$ is asymptotically fluid optimal for any γ that is large enough. Therefore, it does not provide us with a good choice for γ. An asymptotically optimal policy satisfies $\mathbb{E}^{\pi}_{n^r}(\int_0^{r \cdot D}(N_1^r(t) + N_2^r(t))\mathrm{d}t) = r^2 \cdot \mathbb{E}^{\pi}_{n^r}(\int_0^D(\overline{N}_1^r(t) + \overline{N}_2^r(t))\mathrm{d}t) = r^2 \int_0^\infty (n_1^*(t) + n_2^*(t))\mathrm{d}t + o(r^2)$. Hence, a way to determine a good value for γ is by choosing the γ that minimizes the next order term. For the discrete time version of our model it is possible to find an estimate of this term under exponential switching curves, using the techniques of [14].
Consider a discrete-time system with Bernoulli arrivals. In an interval of length Δ, a class-i user arrives with probability $\lambda_i \Delta$, and it leaves the system with probability $(\mu_i s_i + \mu_i c_i s_c)\Delta$, $s_1 + s_2 + s_c \leq 1$. When $\Delta \to 0$, this approximates the continuous-time system with Poisson arrivals and exponential distributed service requirements. (The failure rate in the discrete model is $\mu_i s_i + \mu_i c_i s_c$ which is equal to the failure rate in the stochastic model.) For transparency of

notation, again denote the state at time k by $N_i(k)$, $i = 1, 2$, as we did previously in the continuous-time case.

Following the reasoning in [14] we consider a sequence of systems indexed by a superscript r. In the r-th system, we take as initial point $n^r = (\gamma \ln[rn_2], [rn_2])$ and as time horizon $r \cdot D$ for some fixed D with $n_2 > D$. Write $\mathbb{E}^{\pi}_{n^r}(\sum_{k=0}^{r \cdot D} N^r_1(k) + N^r_2(k)) = \sum_{k=1}^{4} V^{\pi}_k(n^r)$ with

$$\begin{aligned}
V^{\pi}_1(n^r) &= \sum_{k=0}^{r \cdot D} \mathbb{E}^{\pi}_{n^r}(N^r_1(k)), \\
V^{\pi}_2(n^r) &= \sum_{k=0}^{r \cdot D} (n^r_2 + k(\lambda_2 - \frac{1-\rho_1}{1-c_1}\mu_2 c_2)), \\
V^{\pi}_3(n^r) &= \sum_{k=0}^{r \cdot D} \frac{\mu_2 c_2}{\mu_1(1-c_1)}(n^r_1 - \mathbb{E}^{\pi}_{n^r}(N^r_1(k))), \\
V^{\pi}_4(n^r) &= \sum_{k=0}^{r \cdot D} \mu_2 \frac{c_1 + c_2 - 1}{1 - c_1} \mathbb{E}^{\pi}_{n^r}(v^r_k),
\end{aligned}$$

where $v^r_k = \sum_{m=0}^{k-1} \mathbf{1}_{(N^r_1(m)=0)}$ is the number of times the process serves class 2 individually, and π is a policy with switching curve $h(N_1) = e^{N_1/\gamma}$. Since $c_1 < \rho_1 < 1$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1 - c_1)$, we can use the large deviation results in [14] to show that

$$\begin{aligned}
V^{\pi}_1(n^r) &= D\gamma r \ln(r) + O(r), \\
V^{\pi}_2(n^r) &= r^2 \int_0^{\infty} (n^*_1(t) + n^*_2(t))\mathrm{d}t + O(r), \\
V^{\pi}_3(n^r) &= O(r) \\
V^{\pi}_4(n^r) &= \mu_2 \frac{c_1 + c_2 - 1}{1 - c_1} \cdot r^{2 - \beta(\Delta)\gamma + o(1)},
\end{aligned}$$

as $r \to \infty$, with $\beta(\Delta) = \ln(\frac{\rho_1}{c_1} \frac{1 - \mu_1 c_1 \Delta}{1 - \lambda_1 \Delta})$. As explained in [14], setting the value of γ larger than $1/\beta(\Delta)$ gives good second order asymptotics, and the second-order term is then given by $D\gamma r \ln r$.

Now letting $\Delta \to 0$, we have $\lim_{\Delta \to 0} \beta(\Delta) = \ln(\frac{\rho_1}{c_1})$. This suggest that in the continuous-time system we should choose a γ such that $\gamma > \frac{1}{\ln(\frac{\rho_1}{c_1})}$. The condition that γ should be large enough is natural. Setting the value of γ near 0 would almost everywhere give priority to class 1, a strategy that we know can be unstable.

5 Heavy traffic regime

Recall that our model may be viewed as a parallel two-server model. As mentioned in the introduction, policies that are in some sense asymptotically optimal in a heavy traffic setting with complete resource pooling have been investigated in among others [5, 6, 20, 25]. In fact, the results in the latter papers hold for even more general networks than the two-server model. In this section we will collect existing results in the literature that are specific for the parallel two-server model.

We know the system can be kept stable when (2) and (3) are satisfied. Equivalently, we may say that the system can be kept stable when the vector (λ_1, λ_2) lies in the interior of the stability set as depicted in Figure 7. The system is said to be in heavy traffic when the vector (λ_1, λ_2) lies on the northeast boundary of the stability set in Figure 7. In addition, the complete resource pooling condition is satisfied if the outer normal, η, to the stability set at that λ is unique up to scaling and all its coordinates are strictly positive. Hence, complete resource pooling is satisfied when (λ_1, λ_2) is such that $\lambda_i > 0$ for $i = 1, 2$ and $(\lambda_1, \lambda_2) \neq (\mu_1 c_1, \mu_2 c_2)$. This corresponds to one of the following two regions (see also Figure 7):

- Region A: $\rho_2 = 1 - \frac{\rho_1}{c_1}(1 - c_2)$ and $\rho_2 > c_2$, $c_1 > \rho_1$. The outer normal vector to a point in this region is $\eta = (\mu_2(1 - c_2), \mu_1 c_1)$.
- Region B: $\rho_1 = 1 - \frac{\rho_2}{c_2}(1 - c_1)$ and $\rho_1 > c_1$, $c_2 > \rho_2$. The outer normal vector to a point in this region is $\eta = (\mu_2 c_2, \mu_1(1 - c_1))$.

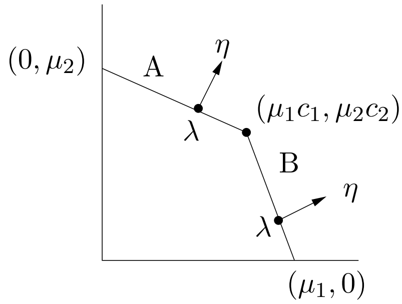

Figure 7: Stability set.

In Section 5.1 we briefly state the results from Bell and Williams [5, 6]. They prove that threshold-based policies asymptotically minimize the (scaled) total number of users in heavy traffic. In Section 5.2 we recall the definition of Max-Weight policies (or $Gc\mu$-rule) and describe the results concerning their behavior in heavy traffic as obtained by Stolyar and Mandelbaum [20, 25].

5.1 Asymptotic optimality of threshold policies

Bell and Williams [5, 6] have investigated the parallel server model (with an arbitrary number of servers and classes) with i.i.d. interarrival times and service requirements. Their model is in fact a slight variation to the model we consider in this paper. First, in their model, once a server starts serving a user, this user has to obtain its full service from this server (non-preemption). Secondly, their model has slightly different behavior near the boundaries: when $N_i = 1$, their model works at rate $\mu_i c_i$ on class i, since a user cannot be split into two. In the model we consider, we can have a departure rate of μ_i. In heavy traffic, these states will be rarely visited, hence the heavy traffic results obtained by Bell and Williams continue to hold for the stochastic model we consider.

Bell and Williams [5, 6] consider a sequence of parameters indexed by r, μ_i^r and λ_i^r (denote $\rho_i^r = \frac{\lambda_i^r}{\mu_i^r}$), with $\lambda_i^r \to \lambda_i, \mu_i^r \to \mu_i$ such that $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$ correspond either to Region A or Region B. An additional condition involves the rate at which the system converges:

$$\lim_{r\to\infty} r\mu_1^r(\rho_1^r - \rho_1) = \theta_1, \qquad \lim_{r\to\infty} r\mu_2^r(\rho_2^r - \rho_2) = \theta_2.$$

Let $N_i^r(t)$ be the number of class-i users in the r-th system, and let $\hat{N}_i^r(t) = \frac{N_i^r(rt)}{\sqrt{r}}$ be the diffusion scaled number of class-i users in the r-th system. Define $\hat{J}^r(\pi) = \mathbb{E}^\pi(\int_0^\infty e^{-\xi t}(\hat{N}_1^r(t) + \hat{N}_2^r(t))\mathrm{d}t)$ where $\xi > 0$ is a constant. A sequence of policies $\tilde{\pi}^r$ is called asymptotically optimal when $\lim_{r\to\infty} \hat{J}^r(\tilde{\pi}^r) \leq \liminf_{r\to\infty} \hat{J}^r(\pi^r)$ for any sequence of policies π^r. For $\mu_1 \leq c_1\mu_1 + c_2\mu_2$ the optimal policy is to serve both classes in parallel whenever possible, which remains valid in heavy traffic. For $\mu_1 > c_1\mu_1 + c_2\mu_2$ the following result holds:

Proposition 5.1 ([6]) *Assume $\mu_1 > c_1\mu_1 + c_2\mu_2$, $c_1 + c_2 > 1$ and i.i.d. inter-arrival times and service requirements. Consider a heavy traffic setting with complete resource pooling.*

- *If (ρ_1, ρ_2) corresponds to Region A, then the policy that serves classes 1 and 2 in parallel whenever possible, is an asymptotically optimal policy (in heavy traffic).*
- *If (ρ_1, ρ_2) corresponds to Region B, then the sequence of threshold policies that gives priority to class 1 when $N_1 > T_1^r = c\log(r)$ (with c large enough), and that otherwise serves classes 1 and 2 in parallel is asymptotically optimal (in heavy traffic).*

In Section 6.2 we will evaluate the performance of threshold-based policies in the under-loaded case.

5.2 Max-Weight policies

In this section we will briefly state the results on Max-Weight policies cf. [20, 25]. For a parallel server system with K classes and L servers, the $Gc\mu$-rule (Max-Weight policy is a special case of this) is defined as follows. Server l serves class i that maximizes $\mu_{il}C_i'(N_i)$ among all classes. Here μ_{il} is the service rate of class-i users when served by server l, and the function $C_i(N_i)$ can be interpreted as the costs of having N_i users in class i. The function $C_i(N_i)$ needs to satisfy some conditions (as specified in [20]), in particular that the second derivative is strictly positive and continuous in $(0, \infty)$. Hence this excludes the function $C_i(N_i) = N_i$, which would be needed to minimize the queue lengths. We will focus on functions of the type $C_i(N_i) = \gamma_i N_i^{\beta+1}$ with $\beta > 0$, which corresponds to the Max-Weight policies. An important property of the Max-Weight policies is that they maintain a stable system under the necessary stability conditions [25].

Stolyar and Mandelbaum [20] consider i.i.d. inter-arrival times and service requirements. They take a sequence of systems indexed by r where the parameters λ_i^r may vary and where $\lambda^r \to \lambda$ with λ such that the system is in heavy traffic and the complete resource pooling condition is satisfied. In addition, $\lim_{r\to\infty} r(\eta \cdot \lambda^r - \eta \cdot \lambda) \to \theta$, with η the corresponding outer normal. It is then proved that in this heavy traffic setting with complete resource pooling, the Max-Weight policy minimizes (under diffusion scaling) both the queueing costs, $\sum_i \gamma_i N_i^{\beta+1}(t)$, and the "virtual" workload, $\sum_i \eta_i N_i(t)$, at all times. In addition, under the diffusion scaling, the vector $(\gamma_1 N_1(t)^\beta, \ldots, \gamma_K N_K(t)^\beta)$ is proportional to $(\eta_1, \ldots, \eta_K)$, hence the dimension of the queue-length processes decreases to one (state space collapse).

Note that the queueing costs cannot represent the total number of users at time t since $\beta > 0$. The Max-Weight policy does minimize the (diffusion-scaled) virtual workload $\sum_i \eta_i N_i(t)$. Hence, when trying to minimize the total number of users among the max-weight policies, it is best to set the parameters (γ_i's and β) such that $N_k(t)$ is as large as possible, where k is such that $\eta_k \geq \eta_i$ for all $i \neq k$. For this reason, in [20] it is suggested that in heavy traffic a good choice for the parameters is $\gamma_i = 1$, $i \neq k$ and $\gamma_k = \epsilon_k$, with $\epsilon_k > 0$ small, since the state space collapse result implies that then $N_k(t)$ will become relatively large compared to $N_i(t), i \neq k$.

For a two-class parallel server as we consider in this paper, the parameters μ_{il} are as follows: $\mu_{11} = \mu_1 c_1, \mu_{21} = \mu_2(1 - c_2), \mu_{12} = \mu_1(1 - c_1)$ and $\mu_{22} = \mu_2 c_2$, see also Figure 2. Assume $c_1 + c_2 \geq 1$ and $c_i \leq 1$. Then the corresponding Max-Weight policy for the two-parallel server model is described as follows:

- Serve class 1 if $N_2^\beta < \frac{\gamma_1(1-c_1)\mu_1}{\gamma_2 c_2 \mu_2} N_1^\beta$.
- Serve classes 1 and 2 in parallel if $\frac{\gamma_1(1-c_1)\mu_1}{\gamma_2 c_2 \mu_2} N_1^\beta \leq N_2^\beta < \frac{\gamma_1 c_1 \mu_1}{\gamma_2(1-c_2)\mu_2} N_1^\beta$.
- Serve class 2 if $\frac{\gamma_1 c_1 \mu_1}{\gamma_2(1-c_2)\mu_2} N_1^\beta \leq N_2^\beta$.

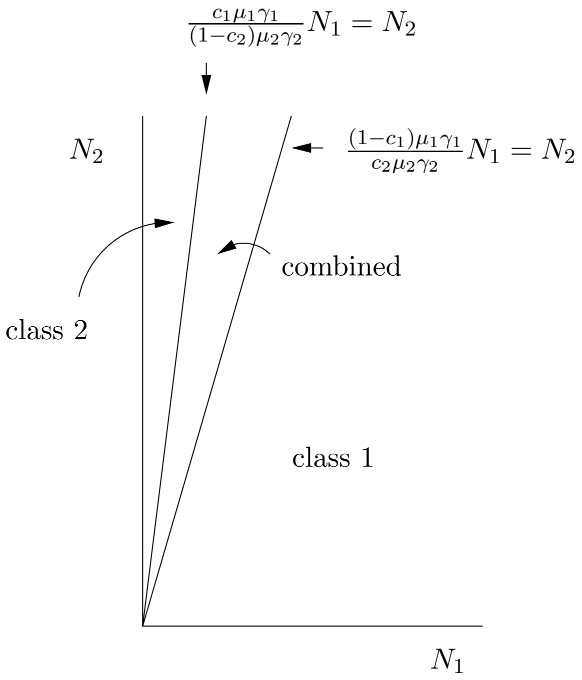
$\frac{c_1\mu_1\gamma_1}{(1-c_2)\mu_2\gamma_2}N_1 = N_2$
N_2
$\frac{(1-c_1)\mu_1\gamma_1}{c_2\mu_2\gamma_2}N_1 = N_2$
combined
class 2
class 1
N_1

Figure 8: The Max-Weight policy.

Hence the Max-Weight policy has two linear switching curves. In Figure 8 these switching curves are plotted for $\beta = 1$. From now on we fix $\beta = 1$. Note that in heavy traffic, the state space collapses to the line $N_2 = \frac{c_1\mu_1\gamma_1}{(1-c_2)\mu_2\gamma_2}N_1$ if we are in Region A and to the line $N_2 = \frac{(1-c_1)\mu_1\gamma_1}{c_2\mu_2\gamma_2}N_1$ if we are in Region B.

In Section 6.3 we will compare the performance of the Max-Weight policies with the optimal policy found numerically, and with the asymptotically fluid optimal policies as we proposed in this paper.

6 Numerical results

The average-optimal policy for the original stochastic model can be computed numerically by value iteration. Figure 9 illustrates for various scenarios that the optimal strategy is characterized by a switching curve. We note that finding these optimal curves numerically was extremely time-consuming. Figures 9 a) and b) consider the setting $\rho_1 < c_1$. We see that the switching curve is linear and coincides exactly with the asymptotically optimal switching curve $h(N_1) = \alpha\frac{\mu_2}{\mu_1}N_1$ from Proposition 4.9. Figure 9 c) corresponds to a scenario with $\rho_1 > c_1$ and illustrates that then the optimal strategy resembles an exponentially shaped curve, which coincides with Proposition 4.10. In the remainder of this section we will assess the gains that can be achieved by choosing the best switching-curve strategy.

6.1 Linear switching strategies for $\rho_1 < c_1$

We first focus on the case $\rho_1 < c_1$. In Figure 10 we plot the mean total number of users under policies with a linear switching curve $h(N_1) = dN_1$. On the horizontal axis we vary the value of d. Note that $d = 0$ corresponds to always serving both classes in parallel. When the slope grows large ($d \to \infty$) the policy gives higher priority to serving class 1 exclusively (whenever present). Note that strict priority for class 1 leads to instability if $\rho_1 + \rho_2 > 1$, which can be the case even if

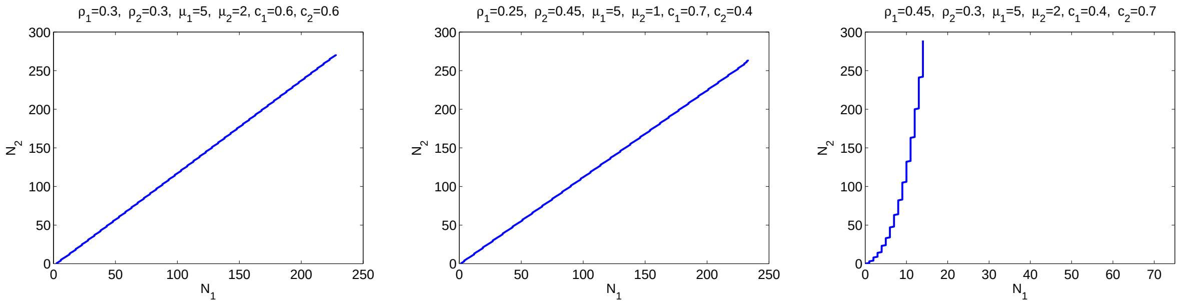

Figure 9: Optimal switching curve when a) $\rho_1 < c_1$, $\rho_2 < c_2$, b) $\rho_1 < c_1$ and $\rho_2 > c_2$ and c) $\rho_1 > c_1$ and $\rho_2 < c_2$.

the stability conditions (2) and (3) are met. The two graphs on the left in Figure 10 correspond to a medium-loaded system. There we also plot the optimal policy found numerically by value iteration. We observe that when the parameter d is chosen well, the linear switching curve policy coincides with the optimal policy. The two graphs on the right in Figure 10 represent a heavily-loaded system. We did not simulate the optimal policy for this parameter setting, since this is extremely time-consuming. Choosing d very large implies that the mean number of users will be large (since $\rho_1 + \rho_2 > 1$). It seems that a good choice for heavily-loaded systems is $d = 0$, i.e., always serve both classes in parallel. In a heavy-traffic setting with $\rho_1 < c_1$ (and necessarily $\rho_2 > c_2$ while $\rho_2 + \frac{\rho_1}{c_1}(1 - c_2) \to 1$) we see that the policy that always serves both classes in parallel is also the asymptotically optimal policy as found by both the fluid analysis (since then the slope is equal to 0) and the heavy traffic analysis. In Figure 11 we repeated the experiment for different parameter choices to illustrate that the relative differences between the optimal linear policy and the strategy that maximizes the service capacity at all times (slope $d = 0$) can be quite significant.

An important observation in Figure 10 and Figure 11 is that the asymptotically fluid optimal policy as found by the fluid analysis in Section 4 (denoted in the figures by “optimal slope fluid”) is always close to optimal and performs very well.

Remark 6.1 *If d tends to ∞, then the system behaves as a priority queue where class 1 is given preemptive priority. When $\rho_1 + \rho_2 < 1$, this policy is stable, and we indeed observe in the two graphs on the left in Figure 10 that the mean number of users will then converge to a constant. However, when $\rho_1 + \rho_2 > 1$, this policy is not stable, and $\mathbb{E}(N_1 + N_2)$ will grow infinitely large as $d \to \infty$. The two graphs on the right in Figure 10 suggest that the mean number of users grows linearly in d as $d \to \infty$. This can be intuitively understood as follows.*

Consider the policy with a linear switching curve $h(N_1) = dN_1$. Then, conditioned on $jd \leq N_2 < (j+1)d$, class 1 has as departure rate $\mu_1 c_1$ if $N_1 \leq j$, and μ_1 otherwise. The equilibrium distribution for such a process is $\pi_i(j) = \pi_0(j)(\frac{\rho_1}{c_1})^i$ if $i \leq j$ and $\pi_i(j) = \pi_0(j)(\frac{\rho_1}{c_1})^j \rho_1^{i-j}$ if $i > j$. If d is large, we assume that class 1 reaches equilibrium during the time that $jd \leq N_2 < (j+1)d$. Then the mean departure rate for class 2 is $\mu_2(j) := \mu_2 \pi_0(j) + \mu_2 c_2 \sum_{i=1}^{j} \pi_i(j)$ when $jd \leq N_2 < (j+1)d$. It can be checked that this is increasing in j, hence there exists a j^* such that $\mu_2(j^* - 1) < \lambda_2 \leq \mu_2(j^*)$ (for convenience we define $\mu_2(-1) = 0$). Note that $j^* > 0$, unless $\rho_1 + \rho_2 < 1$. Hence, if $jd \leq N_2 < (j+1)d$ with $j < j^*$, then the mean drift in class 2 is positive, and the probability that the increase in N_2 is $O(d)$ tends to 1 as $d \to \infty$. If $jd \leq N_2 < (j+1)d$ with $j \geq j^*$, then the mean drift in class 2 is negative. Hence, the probability that the decrement of N_2 is of order $O(d)$ tends to 1 as $d \to \infty$. It is therefore plausible that the process N_2/d will most of the time be around the level j^*.

If the region $(j^* + 1)d \leq N_2$ is not reached (which is not a strong assumption, since this region

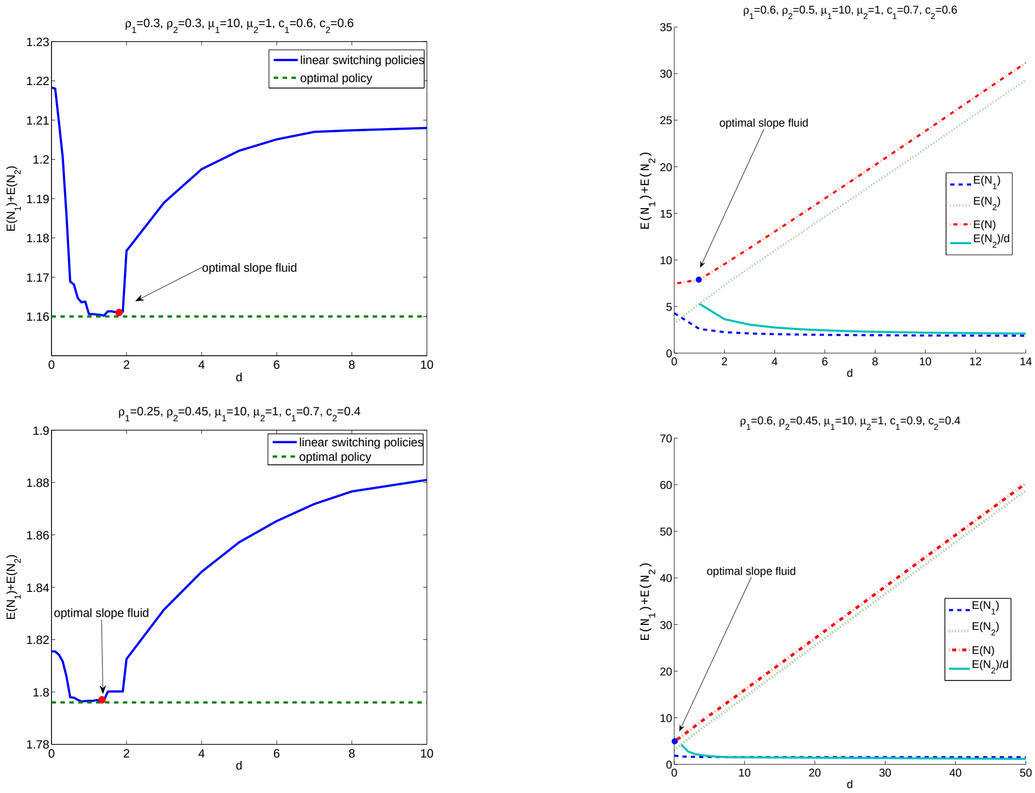

Figure 10: Mean total number of users for policies with a linear switching curve. The marker indicates the optimal slope for the fluid approximation. The two graphs on the top row correspond to cases with $\rho_1 < c_1$ and $\rho_2 < c_2$. The lower graphs have $\rho_1 < c_1$ and $\rho_2 > c_2$.

will be rarely visited as $d \to \infty$), then the number of class-1 users can be upper bounded by the number of class-1 users in a system with departure rates $\mu_1 c_1$ if $N_1 \leq j^*$ and μ_1 otherwise. Since j^* does not depend on d, the upper bound for the number of class-1 users does not scale with d. For the parameters used in the graph on the top right in Figure 10, the j^* is equal to 2. We observe in the figure that $\mathbb{E}(N_2)/d$ indeed converges to $j^* = 2$ and that $\mathbb{E}(N_1)$ does not scale with d. For the parameters that belong to the graph on the bottom right in Figure 10, the j^* is equal to 1. Then as well, we observe in the figure that $\mathbb{E}(N_2)/d$ indeed converges to $j^* = 1$ and that $\mathbb{E}(N_1)$ does not scale with d.

6.2 Exponentially shaped switching strategies for $\rho_1 > c_1$

In Figure 12 we consider several parameter settings with $\rho_1 > c_1$, and plot the total mean number of users under policies with switching curves of the shape $h(N_1) = e^{N_1/\gamma}$. On the horizontal axis we vary the value of γ. Note that when γ grows large, this converges to the policy that always serves both classes in parallel. We observe that the best choice for the parameter γ, delivers the same performance as the optimal policy. As shown in Proposition 4.10, exponential switching curves are asymptotically optimal. The large deviation analysis further suggests that $\gamma > \frac{1}{\ln(\rho_1/c_1)}$ is a good choice, see Section 4.3. For the choices of parameters in the pictures of Figure 12, γ should be larger than 8.5. From Figure 12 we observe that in fact the better choices for the parameter γ are smaller than 8.5. Still, the large deviations result gives a safe estimate (the policy is stable) with better performance than the capacity-maximizing strategy (serving both classes in parallel when possible).

In a heavy traffic regime with $\rho_1 > c_1$, a threshold policy is asymptotically optimal. That

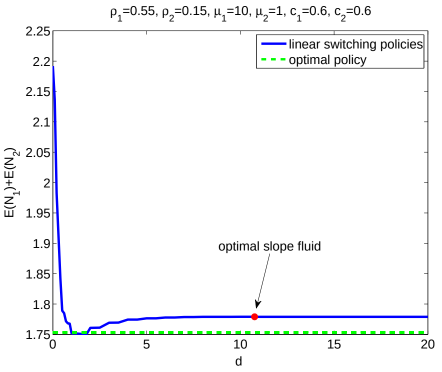

Figure 11: Mean total number of users for policies with a linear switching curve.

is, both classes should be served in parallel whenever the number of class-1 users is below the threshold. When the threshold grows large, this coincides with the policy that always serves both classes in parallel. In Figure 12 we vary the value of this threshold and plot the mean total number of users. For certain small values of the threshold, this policy performs rather well. However, when the threshold is chosen too small, the performance of the system can degrade considerably. For a system with large loads ($\rho_1 + \rho_2 > 1$), this policy is in fact unstable. In Figure 12 c) we already see that the total number of users doubles when the threshold is set equal to 1. In [26] the authors propose estimates for the value of the threshold.

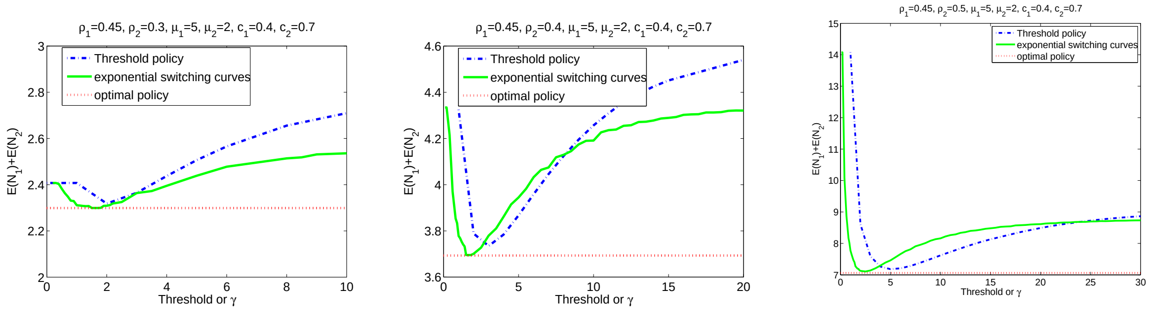

Figure 12: Mean total number of users when $\rho_1 > c_1$ and $\rho_2 < c_2$ for policies with exponential switching curves and for threshold policies for different values of ρ_2.

6.3 Comparison with Max-Weight policies for moderate loads

As stated in Section 5.2, the Max-Weight policies will be close to optimal in a heavy-traffic setting. In this section we investigate the performance of the Max-Weight policies in a mildly loaded system and compare this to the performance of the asymptotically fluid optimal policies as found in this paper. We need to distinguish between whether $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$ or $\mu_1 c_1 + \mu_2 c_2 < \mu_1$. We will see that in both cases the fluid optimal policies can outperform the Max-Weight policies. In addition, we will see that it is not clear which choice of the parameters for the Max-Weight policies gives a good performance for a normally-loaded system.

Case $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$

Assume $\mu_1 > \mu_2$ and $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$. Hence also $\mu_1 c_1 + \mu_2 c_2 \geq \mu_2$. From Section 3.1 we know that the policy which serves classes 1 and 2 in parallel, stochastically minimizes the total number of users present in the system. The Max-Weight policy does not do this: in certain states it still serves class-2 users individually, see Figure 8. According to the reasoning in Section 5.2, in a heavy traffic setting, the parameters of the Max-Weight policy, γ_1 and γ_2, should be set depending on whether the arrival rate vector is closer to Region A or Region B, as depicted in Figure 7. In Region A, the outer normal vector is given by $\eta = (\mu_2(1 - c_2), \mu_1 c_1)$, so that $\eta_1 < \eta_2$. Hence, choose $\gamma_1 = 1$ and $\gamma_2 = \epsilon_2$, with ϵ_2 small. In Region B the outer normal vector is given by $\eta = (\mu_2 c_2, \mu_1(1 - c_1))$, so $\eta_1 > \eta_2$. Hence, choose $\gamma_1 = \epsilon_1$ and $\gamma_2 = 1$, with ϵ_1 small. In Figures 13 and 14 the mean number of users under the Max-Weight policy are plotted for $\gamma_i = 1$ and $\gamma_j = \epsilon_j$, $i \neq j$. The ϵ_j is varied on the horizontal axis. The optimal policy which serves both classes in parallel is plotted as well. We observe that when $\epsilon_j \downarrow 0$, the performance significantly degrades. This comes from the fact that when ϵ_j tends to zero, both switching curves collapse to the same axis, which implies that one class is almost always given full priority, which is far from optimal. A better choice for the parameters seems to be $\gamma_1 = \gamma_2 = 1$. In Figure 13 this gives the same performance as the optimal policy. This comes from the fact that for these parameters the Max-Weight policy will take the optimal action, i.e., serve classes 1 and 2 parallel, whenever $\frac{N_1}{9} \leq N_2 < 9N_1$, which will be the case most of the time. Figure 14 shows that $\gamma_1 = \gamma_2 = 1$ is not always a good choice, since then the optimal action is only taken in states with $\frac{4N_1}{5} \leq N_2 < \frac{6N_1}{5}$. Hence, often either class 1 or class 2 is served individually, which causes the Max-Weight policy with $\gamma_1 = \gamma_2 = 1$ to be approximately 15% worse than the optimal policy.

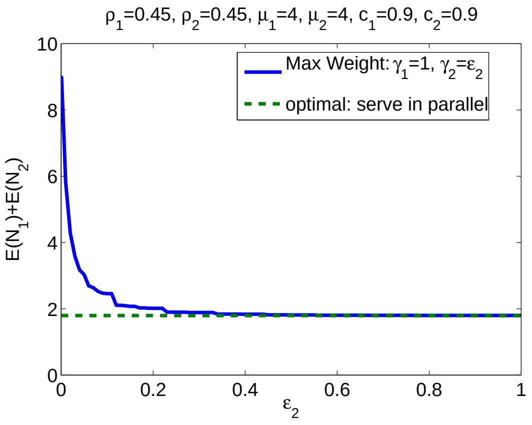

Figure 13: Mean total number of users under the Max-Weight policy, with $\mu_1 c_1 + \mu_2 c_2 > \mu_1$.

Case $\mu_1 c_1 + \mu_2 c_2 < \mu_1$

When $\mu_1 > \mu_2$ and $\mu_1 c_1 + \mu_2 c_2 < \mu_1$, the asymptotically fluid optimal policy we proposed in this paper is described by a switching curve $h(N_1)$ (either linear or exponential), where class 1 is served in states below the switching curve, and classes 1 and 2 are served in parallel in states above the switching curve. When $\mu_1 > \mu_2$ and $\mu_1 c_1 + \mu_2 c_2 < \mu_1$, we have $\eta_1 < \eta_2$, both in Region A and in B of Figure 7. Hence, we consider Max-Weight policies with $\beta = 1, \gamma_1 = 1$ and $\gamma_2 = \epsilon_2$.

In Figures 15 and 16, we compare the performance of Max-Weight policies with the minimum

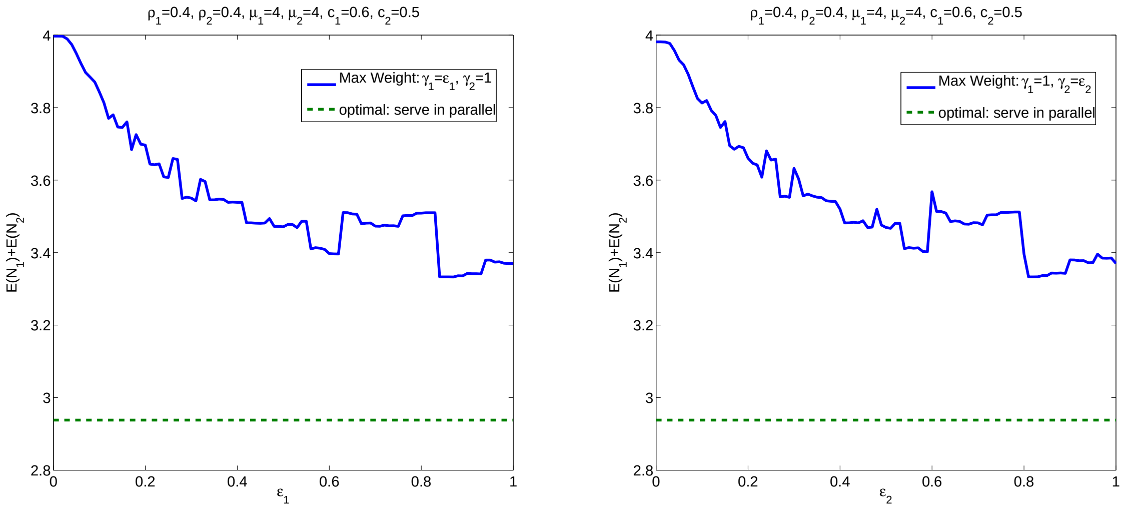

Figure 14: Mean total number of users under the Max-Weight policy, with $\mu_1 c_1 + \mu_2 c_2 > \mu_1$ and a) $\gamma_1 = \epsilon_1$, $\gamma_2 = 1$, and b) $\gamma_1 = 1$, $\gamma_2 = \epsilon_2$.

total mean number of users we found numerically under the best linear or exponential switching curve.

For the parameters as in Figure 15 a), the fluid approximation suggests that if $N_2 \leq 1.8N_1$, then serve class 1, and otherwise serve both classes in parallel. The Max-Weight policy will serve class 1 most of the time, since that is the prescribed action in states such that $N_2 \leq 6\frac{2}{3\epsilon_2}N_1$. From the figure, we see that this is only 5% worse than the optimal policy.

For the parameters as in Figure 15 b), the fluid approximation serves always classes 1 and 2 in parallel. The Max-Weight policy however, serves class 1 individually as soon as $N_2 \leq \frac{12}{10\epsilon_2}N_1$. These states will be visited more often when $\epsilon_2 \downarrow 0$. In the figure, the performance degrades from 15% worse ($\epsilon_2 = 1$), to 30% worse ($\epsilon_2 \downarrow 0$), compared to the optimal linear policy.

In Figure 16, the parameters are such that an exponential switching curve is fluid asymptotically optimal. When $\mu_1 = 10$, the Max-Weight policy is about 15% worse. When $\mu_1 = 2$, it is close to optimal when $\epsilon_2 = 1$, but the performance degrades when $\epsilon_2 \downarrow 0$.

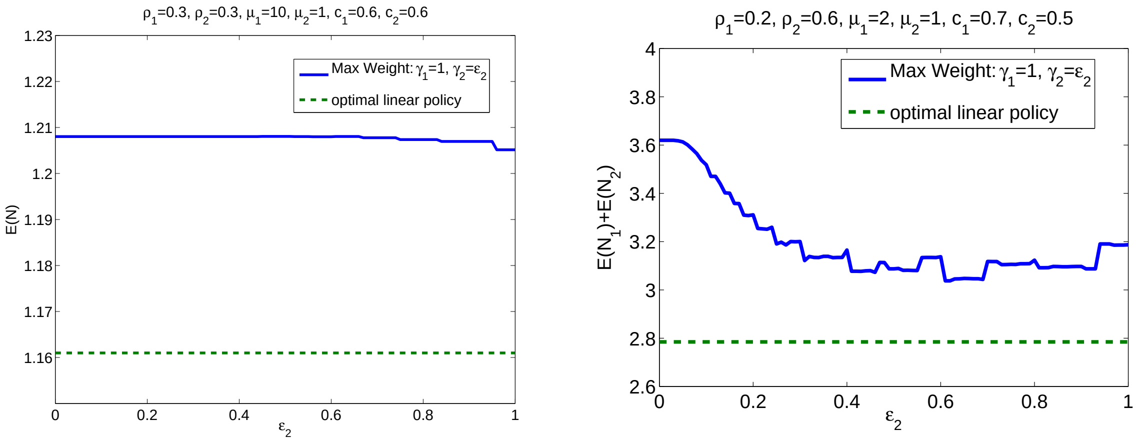

Figure 15: Mean total number of users under the Max-Weight policy and under the optimal linear switching curve, with $\mu_1 c_1 + \mu_2 c_2 < \mu_1$ and $\rho_1 < c_1$.

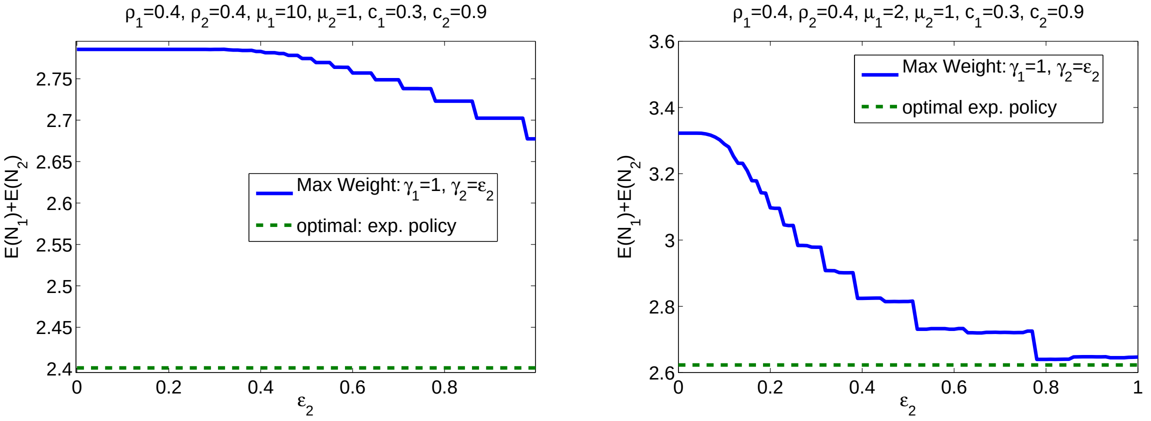

Figure 16: Mean total number of users under the Max-Weight policy and under the optimal exponential switching curve, with $\mu_1 c_1 + \mu_2 c_2 < \mu_1$ and $\rho_1 > c_1$.

7 Conclusion and future work

We have studied optimal policies for systems that have capacity gains when serving users in parallel. Fluid limit analysis shows that the asymptotically optimal strategies are characterized by either linear or exponentially shaped switching curves. The results yield directly usable estimates for good strategies in the stochastic setting, comparing favorably to generally more involved parameter choices in under-loaded regimes for threshold-based strategies and Max-Weight policies (that are asymptotically optimal under heavy-traffic conditions). Note that in [26] the authors provide a method to calculate values for the threshold-based policy.

Several extensions to this work are of interest. For example, it is interesting to investigate how our results change if the capacity is also favorably affected by the numbers of users within each class. For example, in wireless networks the aggregate transmission rate increases with the number of users, due to opportunistic scheduling that deploys multiuser diversity [17].

An intermediate step that is of interest in its own would be to consider our current model with several possible service rate vectors when serving classes in parallel. For example, if in addition to the service rates c_1 and c_2 we can choose d_1 and d_2 which are not in the convex hull depicted in Figure 1.

A third direction of interest is to study our model with more than two classes. This could also serve as an intermediate step towards the first extension mentioned above, which is presumably more difficult to handle. These issues will be addressed in on-going and future research.

Acknowledgment

The authors are grateful to Sem Borst for valuable discussions and comments.

References

[1] Ata, B., Kumar, S. (2005). Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. Annals of Applied Probability **15**, 331–391.

[2] Bäuerle, N. (2000). Asymptotic optimality of tracking policies in stochastic networks. Annals of Applied Probability **10**, 1065–1083.

[3] Bäuerle, N. (2002). Optimal control of queueing networks: an approach via fluid models. Advances in Applied Probability **34**, 313–328.

[4] Bayati, M., Sharma, M., Squillante, M.S. (2006). Optimal scheduling in a multiserver stochastic network. *ACM SIGMETRICS/Performance Evaluation Review* **34**, 45–47.

[5] Bell, S. L., Williams, R. J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Annals of Applied Probability* **11**, 608–649.

[6] Bell, S. L., Williams, R. J. (2005). Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: asymptotic optimality of a threshold policy. *Electronic J. of Probability* **10**, 1044–1115.

[7] Bhardwaj, S., Williams, R. J., Acampora, A.S. (2007). On the performance of a two user MIMO downlink system in heavy traffic. *IEEE Transactions on Information Theory* **53**, 1851–1859.

[8] Billingsley, P. (1999). *Convergence of Probability Measures.* Wiley, New York.

[9] Bonald, T., Proutière, A. (2006). Flow-level stability of utility-based allocations for nonconvex rate regions. In: *Proc. CISS 2006 Conference on Information Sciences and Systems (Princeton University).*

[10] Borst, S.C., Leskela, L., Jonckheere, M. Stability of parallel queueing systems with coupled rates. *Journal of Discrete Events and Dynamic Systems*, to appear.

[11] Dai, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* **5**, 49–77.

[12] Fayolle, G., Iasnogorodski, R. (1979). Two coupled processors: the reduction to a Riemann-Hilbert problem. *Z. Wahr. verw. Geb.* **47**, 325–351.

[13] Gajrat, A., Hordijk, A. (2000). Fluid approximation of a controlled multiclass tandem network. *Queueing Systems* **35**, 349–380.

[14] Gajrat, A., Hordijk, A., Ridder, A. (2003). Large-deviations analysis of the fluid approximation for a controllable tandem queue. *Annals of Applied Probability* **13**, 1423–1448.

[15] Harrison, J.M. (1998). Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of Applied Probability* **8**, 822–848.

[16] Harrison, J.M., López, M.J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing systems* **33**, 339–368.

[17] Liu, X., Chong, E., Shroff, N. (2003). A framework for opportunistic scheduling in wireless networks. *Computer Networks* **41**, 451–474.

[18] Liu, J., Proutière, A., Yi, Y., Chiang, M., Poor, H.V. (2007). Flow-level stability of data networks with non-convex and time-varying rate regions. *ACM SIGMETRICS Performance Evaluation Review* **35**, 239–250.

[19] Maglaras, C. (2000). Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Annals of Applied Probability* **10**, 897–929.

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

[21] Meyn, S.P. (1997). Stability and optimization of queueing networks and their fluid models. In: *Mathematics of stochastic manufacturing systems. Lectures in Applied Mathematics* **33**, 175–199.

[22] Osogami, T., Harchol-Balter, M., Scheller-Wolf, A. (2005). Analysis of cycle stealing with switching times and thresholds. *Performance Evaluation* **61**, 347–369.

[23] Osogami, T., Harchol-Balter, M., Scheller-Wolf, A., Zhang, L. (2004). Exploring Threshold-based Policies for Load Sharing. *Forty-second Annual Allerton Conference on Communication, Control, and Computing*, 1012–1021.

[24] Righter, R., Shanthikumar, J.G. (1989). Scheduling multiclass single-server queueing systems to stochastically maximize the number of successful departures. *Prob. Eng. Inf. Sc.* **3**, 323–333.

[25] Stolyar, A.L. (2004). MaxWeight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Annals of Applied Probability* **14**, 1–53.

[26] Squillante, M.S., Xia, C.H., Yao, D.D., Zhang, L. (2001). Threshold-based priority policies for parallel-server systems with affinity scheduling. *Proc. of the IEEE American Control Conference* **4**, 2992–2999.

[27] Verloop, I.M., Borst, S.C., Núñez-Queija, R. (2006). Delay optimization in bandwidth-sharing networks. In: *Proc. CISS 2006 Conference on Information Sciences and Systems (Princeton University).*

[28] Verloop, I.M., Borst, S.C. (2007). Heavy-traffic delay minimization in bandwidth-sharing networks. In: *Proc. IEEE Infocom 2007.*

[29] Verloop, I.M., Núñez-Queija, R. (2007). Assessing the efficiency of resource allocations in bandwidth-sharing networks. *CWI-Report PNA-E0702.*

Appendix A: Proof of Lemma 3.2

The proof is by induction on the time index k. For $k = 0$ the statement holds. Assume it holds for $W = V_k$. We show that it holds for $W = V_{k+1}$ as well.
When $x_1, x_2 > 1$, by induction the optimal action in the minimization terms in the functions $V_{k+1}(x_1, x_2), V_{k+1}(x_1 - 1, x_2)$ and $V_{k+1}(x_1, x_2 - 1)$ is to serve classes 1 and 2 in parallel. Using this fact, the proof follows easily.
We are left with the cases $x_1 = 1$ or $x_2 = 1$. First assume $x_1 > 0$ and $x_2 = 1$. We will show that $(\mu_1 + \mu_2)V_{k+1}(x_1, 1) + \mu_1 c_1 V_{k+1}(x_1 - 1, 1) + \mu_2 c_2 V_{k+1}(x_1, 0) \le \mu_1 V_{k+1}(x_1, 1) + \mu_2 V_{k+1}(x_1, 0) + (\mu_1 c_1 + \mu_2 c_2)V_{k+1}(x_1, 1)$ is indeed satisfied.
We can write

$$\begin{aligned}
&\mu_2 V_{k+1}(x_1, 1) + \mu_1 c_1 V_{k+1}(x_1 - 1, 1) + \mu_2 c_2 V_{k+1}(x_1, 0)\\
&\le \mu_2[\lambda_1 V_k(x_1 + 1, 1) + \lambda_2 V_k(x_1, 2) + \mu_1 V_k(x_1, 1) + \mu_2 V_k(x_1, 0) + (\mu_1 c_1 + \mu_2 c_2)V_k(x_1, 1)]\\
&\quad + \mu_1 c_1[\lambda_1 V_k(x_1, 1) + \lambda_2 V_k(x_1 - 1, 2) + \mu_1 V_k(x_1 - 1, 1) + \mu_2 V_k(x_1 - 1, 0)\\
&\qquad\qquad + (\mu_1 c_1 + \mu_2 c_2)V_k(x_1 - 1, 1)]\\
&\quad + \mu_2 c_2[\lambda_1 V_k(x_1 + 1, 0) + \lambda_2 V_k(x_1, 1) + \mu_1 V_k(x_1 - 1, 0) + \mu_2 V_k(x_1, 0) + (\mu_1 c_1 + \mu_2 c_2)V_k(x_1, 0)]\\
&= \lambda_1[\mu_2 V_k(x_1 + 1, 1) + \mu_1 c_1 V_k(x_1, 1) + \mu_2 c_2 V_k(x_1 + 1, 0)]\\
&\quad + \lambda_2[\mu_2 V_k(x_1, 2) + \mu_1 c_1 V_k(x_1 - 1, 2) + \mu_2 c_2 V_k(x_1, 1)]\\
&\quad + \mu_1[\mu_2 V_k(x_1, 1) + \mu_1 c_1 V_k(x_1 - 1, 1) + \mu_2 c_2 V_k(x_1, 0)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[\mu_2 V_k(x_1, 1) + \mu_1 c_1 V_k(x_1 - 1, 1) + \mu_2 c_2 V_k(x_1, 0)]\\
&\quad + \mu_2[\mu_2 V_k(x_1, 0) + \mu_2 c_2 V_k(x_1, 0)]\\
&\quad + \mu_1\mu_2[(c_1 + c_2)V_k(x_1 - 1, 0) - c_2 V_k(x_1, 0)]\\
&\le \lambda_1[\mu_2 V_k(x_1 + 1, 0) + (\mu_1 c_1 + \mu_2 c_2)V_k(x_1 + 1, 1)]\\
&\quad + \lambda_2[\mu_2 V_k(x_1, 1) + (\mu_1 c_1 + \mu_2 c_2)V_k(x_1, 2)]\\
&\quad + \mu_1[\mu_2 V_k(x_1, 0) + (\mu_1 c_1 + \mu_2 c_2)V_k(x_1, 1)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[\mu_2 V_k(x_1, 1) + \mu_1 c_1 V_k(x_1 - 1, 1) + \mu_2 c_2 V_k(x_1, 0)]\\
&\quad + \mu_2[\mu_2 V_k(x_1, 0) + \mu_2 c_2 V_k(x_1, 0)]\\
&\quad + \mu_1\mu_2[(c_1 - 1)V_k(x_1, 0) + V_k(x_1 - 1, 0)]\\
&= \mu_2[\lambda_1 V_k(x_1 + 1, 0) + \lambda_2 V_k(x_1, 1) + \mu_1 V_k(x_1 - 1, 0) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2)V_k(x_1, 0)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[\lambda_1 V_k(x_1 + 1, 1) + \lambda_2 V_k(x_1, 2) + (\mu_1 + \mu_2)V_k(x_1, 1)\\
&\qquad\qquad\qquad + \mu_1 c_1 V_k(x_1 - 1, 1) + \mu_2 c_2 V_k(x_1, 0)]\\
&= \mu_2 V_{k+1}(x_1, 0) + (\mu_1 c_1 + \mu_2 c_2)V_{k+1}(x_1, 1),
\end{aligned}$$

which was to be proved. In the second inequality we used that V_k is increasing in x_1, $c_1 + c_2 > 1$ and that (5) holds by induction for V_k.
The remaining cases can be checked in a similar fashion. □

Appendix B: Proof of Lemma 3.4:

Relation (6) can be easily shown. Assume at time t we have $S_1^{\pi}(t) = S_1^{\tilde{\pi}}(t)$. By (1) and since $W_i(0)$ and $A_i(0, t)$ are independent of the policy, we have $W_1^{\pi}(t) = W_1^{\tilde{\pi}}(t)$. By construction of policy π we obtain that $s_1^{\pi}(t^+) \ge s_1^{\tilde{\pi}}(t^+)$. Hence (6) holds for all $t \ge 0$.
Let time t be the first time instant that either (7) or (8) holds with equality and is violated at time t^+.

First assume equation (7) is the first equation that fails to hold, i.e., $S_1^{\pi}(t)+S_2^{\pi}(t) = S_1^{\tilde{\pi}}(t)+S_2^{\tilde{\pi}}(t)$,

$s_1^{\pi}(t^+) + s_2^{\pi}(t^+) < s_1^{\tilde{\pi}}(t^+) + s_2^{\tilde{\pi}}(t^+)$ and by (1) also $W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$. We will show that

$$W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t) \quad \text{implies that} \quad W_i^{\pi}(t) = W_i^{\tilde{\pi}}(t), \;\; i = 1, 2, \tag{26}$$

and hence $N^{\pi}(t) = N^{\tilde{\pi}}(t)$. By construction of policy π this means that $s_1^{\pi}(t^+) + s_2^{\pi}(t^+) \geq s_1^{\tilde{\pi}}(t^+) + s_2^{\tilde{\pi}}(t^+)$ and hence we reach a contradiction.
So let us prove (26). Without loss of generality we assume that $W_1^{\pi}(t)+W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t)+W_2^{\tilde{\pi}}(t)$ and $W_1^{\pi}(t^-)+W_2^{\pi}(t^-) < W_1^{\tilde{\pi}}(t^-)+W_2^{\tilde{\pi}}(t^-)$. This implies that there is an interval $[u, t]$ in which policy $\tilde{\pi}$ has more work in the system compared to policy π, but at time t it has made up for the lost capacity in one of the following ways:

(i) In the interval $[u, t]$ policy $\tilde{\pi}$ serves both classes combined, while policy π serves class 2 with service rate 1. Hence $W_1^{\pi}(v) = 0$ and $W_2^{\pi}(v) > 0$, for all $v \in [u, t]$.

(ii) In the interval $[u, t]$ policy $\tilde{\pi}$ serves both classes combined, while policy π serves class 1 with service rate 1. Hence $W_2^{\pi}(v) = 0$ and $W_1^{\pi}(v) > 0$, for all $v \in [u, t]$.

(iii) In the interval $[u, t]$ policy $\tilde{\pi}$ serves either class 1, class 2 or both classes simultaneously, while policy π has an empty system in the time interval.

In case (i) the total amount of additional capacity that policy $\tilde{\pi}$ gets compared with policy π in the interval $[u, t]$ is equal to the difference in the total workload at time u, so $M(u, t)(c_1+c_2-1) = W_1^{\tilde{\pi}}(u) + W_2^{\tilde{\pi}}(u) - W_2^{\pi}(u)$, with $M(u, t)$ the cumulative amount of time that both classes are served combined under policy $\tilde{\pi}$ in the time interval $[u, t]$. Together with (1) and (8), we obtain that $W_1^{\tilde{\pi}}(u) \leq \frac{c_1}{c_1+c_2-1}(W_1^{\tilde{\pi}}(u) + W_2^{\tilde{\pi}}(u) - W_2^{\pi}(u)) = c_1 M(u, t)$. Also $S_1^{\tilde{\pi}}(t) - S_1^{\tilde{\pi}}(u) \geq c_1 M(u, t)$ and $A_1(u, t) = 0$ (since π serves class 2 and $W_1^{\pi}(v) = 0$ for all $v \in [u, t]$). Together this gives $W_1^{\tilde{\pi}}(t) = W_1^{\tilde{\pi}}(u) + A_1(u, t) - (S_1^{\tilde{\pi}}(t) - S_1^{\tilde{\pi}}(u)) \leq 0$. Hence $W_1^{\tilde{\pi}}(t) = 0$, but we also had $W_1^{\pi}(t) = 0$. It now immediately follows that $W_2^{\tilde{\pi}}(t) = W_2^{\pi}(t)$.
In case (ii) we have that $W_1^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$. By (1) and (6) we have $W_1^{\pi}(t) \leq W_1^{\tilde{\pi}}(t)$. Hence $W_2^{\tilde{\pi}}(t) = 0$ $(= W_2^{\pi}(t))$ and $W_1^{\tilde{\pi}}(t) = W_1^{\pi}(t)$, $i = 1, 2$.
In case (iii) we have that $W_i^{\pi}(t) = 0$ for $i = 1, 2$. Since at time t also $W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$, we obtain that $W_i^{\tilde{\pi}}(t) = 0$, $i = 1, 2$, as well.
Hence we have shown that (26) holds for (i), (ii) and (iii).

Now assume (8) is the first equation that fails to hold, i.e., $(1 - c_2)S_1^{\pi}(t) + c_1 S_2^{\pi}(t) = (1 - c_2)S_1^{\tilde{\pi}}(t) + c_1 S_2^{\tilde{\pi}}(t)$, $(1 - c_2)s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) < (1 - c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+)$ and by (1) also $(1 - c_2)W_1^{\pi}(t) + c_1 W_2^{\pi}(t) = (1 - c_2)W_1^{\tilde{\pi}}(t) + c_1 W_2^{\tilde{\pi}}(t)$. We have the following three possibilities:

- When $s_1^{\pi}(t^+) = 0$, i.e. π serves class 2 individually, or both classes in parallel at time t^+, we have $(1-c_2)s_1^{\pi}(t^+)+c_1 s_2^{\pi}(t^+) = c_1$. Since $c_1+c_2 > 1$ we have that $c_1 \geq (1-c_2)s_1^{\tilde{\pi}}(t^+)+c_1 s_2^{\tilde{\pi}}(t^+)$, and hence we obtain a contradiction.

- When $s_1^{\pi}(t^+) > 0$, i.e. π serves class 1 individually at time t^+, then $W_1^{\pi}(t) > 0$. By definition $s_1^{\pi}(t^+) \geq s_1^{\tilde{\pi}}(t^+)$. In order for $(1-c_2)s_1^{\pi}(t^+)+c_1 s_2^{\pi}(t^+) < (1-c_2)s_1^{\tilde{\pi}}(t^+)+c_1 s_2^{\tilde{\pi}}(t^+)$, that is $(1-c_2)(s_1^{\pi}(t^+)-s_1^{\tilde{\pi}}(t^+)) < c_1(s_2^{\tilde{\pi}}(t^+)-s_2^{\pi}(t^+))$, we then need $s_2^{\tilde{\pi}}(t^+) > s_2^{\pi}(t^+)$. Hence $W_2^{\tilde{\pi}}(t) > 0$. But from $W_1^{\pi}(t) \leq W_1^{\tilde{\pi}}(t)$ and $(1 - c_2)W_1^{\pi}(t) + c_1 W_2^{\pi}(t) = (1 - c_2)W_1^{\tilde{\pi}}(t) + c_1 W_2^{\tilde{\pi}}(t)$, it follows that $W_2^{\pi}(t) \geq W_2^{\tilde{\pi}}(t)$. So under policy π there is work of both classes present. Since policy $\tilde{\pi}$ serves class 2, policy π serves by definition classes 1 and 2 in parallel. We again have a contradiction.

- When π has an empty system at time t^+, then $0 = (1 - c_2)W_1^{\tilde{\pi}}(t) + c_1 W_2^{\tilde{\pi}}(t)$. Hence also $\tilde{\pi}$ has an empty system and $(1 - c_2)c_2 s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) = (1 - c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+) = 0$.

In all three cases we reach a contradiction and this concludes the proof. □

Appendix C: Proof of Lemma 4.1

Denote by $n^D(t)$ the minimizing trajectory of $\min_{n(t) \text{ s.t. } (10)-(14)} \int_0^D (n_1(t) + n_2(t))\mathrm{d}t$. First assume that there exists a function $D(\cdot)$ such that

$$|n^D(t)| = 0, \qquad \text{for all } t \geq D(|n|). \tag{27}$$

Then we have that $\min_{n(t) \text{ s.t. } (10)-(14)} \int_0^D (n_1(t)+n_2(t))\mathrm{d}t = \int_0^D (n_1^D(t)+n_2^D(t))\mathrm{d}t = \int_0^\infty (n_1^D(t)+n_2^D(t))\mathrm{d}t$ for all $D \geq D(|n|)$. In addition $\min_{n(t) \text{ s.t. } (10)-(14)} \int_0^\infty (n_1(t) + n_2(t))\mathrm{d}t = \int_0^\infty (n_1^*(t) + n_2^*(t))\mathrm{d}t \geq \int_0^D (n_1^*(t) + n_2^*(t))\mathrm{d}t$. Together this proves the result.

In the remainder of the proof we will show that there exists a $D(|n|)$ such that (27) holds. The proof uses the arguments from [19, Proposition 6.1].

Denote by π^p the policy that always serves classes 1 and 2 in parallel whenever possible. Let $n^p(t)$ be the trajectory that corresponds to policy π^p. Under the stability conditions we know that $n^p(t)$ hits zero after a finite time and then remains empty. Denote by $T^p(n, 0)$ the time it takes for the system to become empty when starting in state n. Note that the depletion time scales as follows: $T^p(\zeta^{m-1} \cdot n, 0) = \zeta^{m-1} \cdot T^p(n, 0)$. For the case $\rho_1 < c_1$, this can be easily seen from the expressions in the proof of Proposition 4.4 (take $b = n$). For $\rho_1 \geq c_1$ this can be checked similarly.

Now consider as initial point $n(0) = n^{(m-1)}$ such that $|n^{(m-1)}| = |n| \cdot \zeta^{m-1}$, with $0 < \zeta < 1$. Then we have for all D:

$$\begin{aligned}
\int_0^D (n_1^D(t) + n_2^D(t))\mathrm{d}t &= \min_{\substack{n(t) \text{ s.t. } (10)-(13) \\ n(0)=n^{(m-1)}}} \int_0^D (n_1(t) + n_2(t))\mathrm{d}t \\
&\leq \int_0^D (n_1^p(t) + n_2^p(t))\mathrm{d}t \\
&\leq |n| \cdot \zeta^{m-1} \cdot T^p(n^{(m-1)}, 0) \\
&\leq |n| \cdot \zeta^{2(m-1)} \cdot \sup_{l:|l|=|n|} T^p(l, 0).
\end{aligned}$$

Hence, for all $D \geq \zeta^{m-1} \frac{\sup_{l:|l|=|n|} T^p(l,0)}{\zeta}$ it holds that

$$\begin{aligned}
\min_{t \leq D} \{n_1^D(t) + n_2^D(t) | n^D(0) = n^{(m-1)}\} &\leq \frac{|n| \cdot \zeta^{2(m-1)} \cdot \sup_{l:|l|=|n|} T^p(l, 0)}{D} \\
&\leq |n| \cdot \zeta^m.
\end{aligned} \tag{28}$$

Define

$$\tau_m = \min\{t > 0 : n_1^D(t) + n_2^D(t) < |n| \cdot \zeta^m, \ n_1^D(0) + n_2^D(0) = |n| \cdot \zeta^{m-1}\}, \ m = 1, 2, \ldots$$

From (28) it follows that $\tau_m \leq \zeta^{m-1} \cdot \frac{\sup_{l:|l|=|n|} T^p(l,0)}{\zeta}$. Hence,

$$\sum_{m=1}^\infty \tau_m \leq \frac{\sup_{l:|l|=|n|} T^p(l, 0)}{\zeta(1 - \zeta)} =: D(|n|)$$

and $D(|n|) < \infty$. Continuity now gives that $|n^D(\sum_{m=1}^\infty \tau_m)| = 0$. Further note that the optimal trajectory will remain empty from then on. Hence, given initial state n, $|n^D(t)| = 0$ for all $t \geq D(|n|)$, i.e., (27) holds. □

Appendix D: Proposition 4.9

We will first prove equations (20)–(22).

Proof of (20)–(22):

Fix a sample path ω such that there is a subsequence r_k with $\lim_{r_k\to\infty}\overline{N}_i^{r_k}(t)=\overline{N}_i(t)$ u.o.c. and $\lim_{r_k\to\infty}\overline{T}_j^{r_k}(t))=\overline{T}_j(t)$ u.o.c.. Here $\overline{N}_i(t)$ is given by (18). Further, let $t>0$ be a regular point for $\overline{T}_j(t)$ for all $j=1,2,c$.
First assume $\overline{N}_2(t)<\alpha\frac{\mu_2}{\mu_1}\overline{N}_1(t)$. Then there is an $\epsilon>0$ such that $\overline{N}_2(s)<\alpha\frac{\mu_2}{\mu_1}\overline{N}_1(s)$ for $s\in[t-\epsilon,t+\epsilon]$. By the uniform convergence of $\overline{N}_i^{r_k}(t)$ to $\overline{N}_i(t)$, $i=1,2$, for r_k large enough we have $N_2(r_ks)<\alpha\frac{\mu_1}{\mu_2}N_1(r_ks)$ for $s\in[t-\epsilon,t+\epsilon]$. Hence, under policy π^*, in the interval $[r_k(t-\epsilon),r_k(t+\epsilon)]$ class 1 is served and we obtain $\overline{T}_1^{r_k}(t+\epsilon)-\overline{T}_1^{r_k}(t-\epsilon)=2\epsilon$. Letting $\epsilon\downarrow 0$ we obtain $\frac{\mathrm{d}\overline{T}_1(t)}{\mathrm{d}t}=1$.
Now assume $\overline{N}_2(t)>\alpha\frac{\mu_2}{\mu_1}\overline{N}_1(t)$ and $\overline{N}_1(t)>0$. Then there is an ϵ such that $N_2(r_ks)>\alpha\frac{\mu_2}{\mu_1}N_1(r_ks)$ and $N_1(r_ks)>0$ for $s\in[t-\epsilon,t+\epsilon]$ and r_k large enough. Under policy π^*, in this interval both classes are served in parallel, hence $\frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t}=1$.
Assume $\overline{N}_2(t)=\alpha\frac{\mu_2}{\mu_1}\overline{N}_1(t)$ and $\overline{N}_1(t)>0$. There is an ϵ such that $N_1(r_ks)>0$ for $s\in(t-\epsilon,t+\epsilon)$ and r_k large enough. In this interval, class 1 is always served, so $\frac{\mathrm{d}\overline{T}_1(s)}{\mathrm{d}s}+\frac{\mathrm{d}\overline{T}_c(s)}{\mathrm{d}s}=1$, for any regular point $s\in(t-\epsilon,t+\epsilon)$. Together with (18), $\rho_1<c_1$, $\alpha\geq 0$ and $\alpha>\frac{c_2-\rho_2}{c_1-\rho_1}$ we obtain $\alpha\frac{\mu_2}{\mu_1}\frac{\mathrm{d}\overline{N}_1(s)}{\mathrm{d}s}-\frac{\mathrm{d}\overline{N}_2(s)}{\mathrm{d}s}=\mu_2(\alpha(\rho_1-\frac{\mathrm{d}\overline{T}_1(s)}{\mathrm{d}s}-c_1\frac{\mathrm{d}\overline{T}_c(s)}{\mathrm{d}s})-(\rho_2-c_2\frac{\mathrm{d}\overline{T}_c(s)}{\mathrm{d}s}))<0$, for s a regular point, $s\in(t-\epsilon,t+\epsilon)$. This means that if $\overline{N}$ lies below the switching curve, then it moves towards the switching curve and when $\overline{N}$ lies on or above the switching curve, it will move upwards, away from the switching curve. Since we are in a state on the switching curve, there is an $\epsilon>0$ such that $\overline{N}_2(s)<\alpha\frac{\mu_2}{\mu_1}\overline{N}_1(s)$ for $s\in(t-\epsilon,t)$ and $\overline{N}_2(s)>\alpha\frac{\mu_2}{\mu_1}\overline{N}_1(s)$ for $s\in(t,t+\epsilon)$. Hence, the derivative from the left is $\frac{\mathrm{d}\overline{T}_1(t^-)}{\mathrm{d}t}=1$, and the derivative from the right is $\frac{\mathrm{d}\overline{T}_c(t^+)}{\mathrm{d}t}=1$. Hence, the point t itself is not a regular point.
Finally assume $\overline{N}_1(t)=0$ and $\overline{N}_2(t)>0$. Then there is an $\epsilon>0$ such that $\overline{N}_2(s)>\alpha\frac{\mu_1}{\mu_2}\overline{N}_1(s)$ for $s\in[t-\epsilon,t+\epsilon]$ and hence $N_2(r_ks)>\alpha\frac{\mu_1}{\mu_2}N_1(r_ks)$ for $s\in[t-\epsilon,t+\epsilon]$ and r_k large enough. Recall that t is a regular point, so $\frac{\mathrm{d}\overline{T}_1(t)}{\mathrm{d}t}=0$ and from (18) we then have

$$\frac{\mathrm{d}\overline{N}_1(t)}{\mathrm{d}t}=\lambda_1-\mu_1c_1\frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t}. \tag{29}$$

Note that if $\overline{N}_1(t^+)>0$, then $\frac{\mathrm{d}\overline{T}_c(t^+)}{\mathrm{d}t}=1$. Since $\rho_1<c_1$, we obtain from (29) that $\frac{\mathrm{d}\overline{N}_1(t)}{\mathrm{d}t}=0$. Hence $\frac{\mathrm{d}\overline{T}_c(t)}{\mathrm{d}t}=\frac{\rho_1}{c_1}$.
Therefore, (20)–(22) are satisfied for each fluid limit $\overline{T}(t)$. □

Proof of Proposition 4.9:

For any sequence n^r such that $\lim_{r\to\infty}\frac{n^r}{r}=n$, we have

$$\begin{aligned}\limsup_{r\to\infty}\mathbb{E}^{\pi}_{n^r}(\int_0^D\overline{N}_1^r(t)+\overline{N}_2^r(t)\mathrm{d}t) &= \limsup_{r\to\infty}\int_0^D\mathbb{E}^{\pi}_{n^r}(\overline{N}_1^r(t)+\overline{N}_2^r(t))\mathrm{d}t\\ &\leq \int_0^D\limsup_{r\to\infty}\mathbb{E}^{\pi}_{n^r}(\overline{N}_1^r(t)+\overline{N}_2^r(t))\mathrm{d}t. \end{aligned}\tag{30}$$

The inequality follows from Fatou's lemma, which applies since $\mathbb{E}^{\pi}_{n^r}(\overline{N}_1^r(t)+\overline{N}_2^r(t))\leq(\lambda_1+\lambda_2)t$. In Section 4.1 we determined the optimal trajectory for the fluid control model, $n^*(t)$. It remains to show that under the stationary policy π^* with switching curve $h(N_1) = \alpha\frac{\mu_2}{\mu_1}N_1$ we have that $\text{limsup}_{r\to\infty}\mathbb{E}^{\pi^*}_{n^r}(\overline{N}^r_1(t) + \overline{N}^r_2(t)) = n^*_1(t) + n^*_2(t)$.

Fix the sample path ω. For a $t > 0$, let the subsequence r_k be such that $\text{limsup}_{r\to\infty}\overline{N}^r_i(t) = \lim_{k\to\infty}\overline{N}^{r_k}_i(t)$. From Lemma 4.6 it follows that almost surely there exists a subsequence r_{k_l} of r_k such that $\lim_{l\to\infty}\overline{N}^{r_{k_l}}(t) = \overline{N}(t)$. Since every fluid limit $\overline{N}(t)$ must coincide with the optimal fluid control solution $n^*(t)$ (see (23)) we obtain $\limsup_{r\to\infty}\overline{N}^r_i(t) = n^*_i(t)$ pointwise almost surely. The same holds for the $\liminf$, and hence we obtain that $\lim_{r\to\infty}\overline{N}^r_i(t) = n^*_i(t)$ pointwise almost surely.

We also know that the function $\overline{N}^r_1(t)+\overline{N}^r_2(t)$ is uniform integrable. This follows from the same argument as in the proof of [11, Lemma 4.5]. Here we state it briefly. Note that $\overline{N}^r_1(t)+\overline{N}^r_2(t) \leq \frac{n^r_1+n^r_2}{r} + \frac{E_1(rt)+E_2(rt)}{r}$, with $E_i(\cdot)$ a Poisson process with rate λ_i. Since $\lim_{r\to\infty}\frac{E_1(rt)+E_2(rt)}{r} = (\lambda_1 + \lambda_2)t$ almost surely (see Lemma 4.6) and $\mathbb{E}(\frac{E_1(rt)+E_2(rt)}{r}) = (\lambda_1 + \lambda_2)t$, we obtain from [8, Theorem 3.6] that $\frac{E_1(rt)+E_2(rt)}{r}$ is uniform integrable. Hence by definition of uniform integrability it is immediate that also the function $\overline{N}^r_1(t) + \overline{N}^r_2(t)$ is uniform integrable.

Since the function $\overline{N}^r_1(t)+\overline{N}^r_2(t)$ is uniform integrable and converges point-wise to $n^*_1(t)+n^*_2(t)$ a.s., we can interchange the limit and the expectation (see [8, Theorem 3.5]). We obtain

$$\begin{aligned}\int_0^D \limsup_{r\to\infty}\mathbb{E}^{\pi^*}_{n^r}(\overline{N}^r_1(t) + \overline{N}^r_2(t))\mathrm{d}t &= \int_0^D \mathbb{E}^{\pi^*}\Big(\lim_{r\to\infty}\overline{N}^r_1(t) + \overline{N}^r_2(t)\Big)\mathrm{d}t \\ &= \int_0^\infty (n^*_1(t) + n^*_2(t))\mathrm{d}t.\end{aligned}$$

Together with (30) we obtain $\limsup_{r\to\infty}\mathbb{E}^{\pi}_{n^r}(\int_0^D(\overline{N}^r_1(t) + \overline{N}^r_2(t))\mathrm{d}t) \leq \int_0^\infty(n^*_1(t) + n^*_2(t))\mathrm{d}t$. □