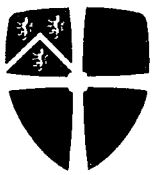
Stability Criteria for Controlled Queueing Networks

Lisa Johanna Müller

A Thesis presented for the degree of
Doctor of Philosophy



Statistics and Probability Group
Department of Mathematical Sciences
Durham University
England

November 2006

Stability Criteria for Controlled Queueing Networks

Lisa Johanna Müller

Submitted for the degree of Doctor of Philosophy

November 2006

Abstract

We give criteria for the stability of a very general queueing model under different levels of control. A complete classification of stability (or positive recurrence), transience and null-recurrence is presented for the two queue model. The stability and instability results are extended for models with $N \geq 3$ queues. We look at a broad class of models which can have the following features:

Customers arrive at one, several or all of the queues from the outside with exponential inter arrival times. We often have the case where a arrival stream can be routed so that under different routing schemes each queue can have external arrivals, i.e. we assume we have some control over the routing of the arrivals. We also consider models where the arrival streams are fixed.

We view the service in a more abstract way, in that we allow a number k of different service configurations. Under every such service configuration service is provided to some or all of the queues, length of service time can change from one service configuration to another and we can change from one configuration to another according to some control policy. The service times are assumed to be exponentially distributed.

The queueing models we consider are networks where, after completion at one queue, a customer might be fed back into another queue where it will be served another time often under with a different service time. These feedback probabilities change with the service configurations.

Our interest is in different types of control policies which allow us to change the

routing of arrivals and configurations of the service from time to time so that the controlled queue length process (which in most cases is Markov) is stable. The semi-martingale or Lyapunov function methods we use give necessary and sufficient conditions for the stability classification. We will look at some two queue models with different inter arrival and service times where the queueing process is still Markov.

Declaration

The work in this thesis is based on research carried out at the Statistics and Probability Group, Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Parts of Chapter 2 and parts of Chapter 4 are adapted from joint work with Iain MacPhee [26], [27], [28].

Acknowledgements

First I would like to thank me. Special thanks go to Iain MacPhee for being so patient with me, my work and working process, and also for the rather interesting conversations about everything else not related to mathematics. Thanks to Iain MacPhee and Misha Menshikov for proposing the problem and being constructively critical of my work. Special thanks also to Andrew Wade for being a great friend and having so many answers. Thanks to the Head of the Department Richard Ward and to Helene Rusby for the support and for making it possible for me to go to international conferences, which were a true enrichment. Frank Coolen and Pauline Coolen-Schrijner have been a great help and support since the first email I sent to the Department, my thanks to both of you and to everybody else in the group for making the postgraduate experience in the Statistics and Probability so good. Thanks to Jonty Rougier for being Jonty and for creating a sample path of the queue length process for those who see things differently, see Figure 2.13. I would also like to thank my parents who have supported my thoughout the three years.

Also (in random order) thanks to: Pamela for never failing to put a smile on my face (when I arrived in the Department early enough); to Alicia Huntriss and the other members of the Ustinov boat club for giving me the opportunity to get into rowing; to the people in my office Anna, Gina, Mark, David, Jennifer and Gavin, for sharing; to Lucy, Marek and the SCA for a change of perspective; to Eleanor and Bina for good times and cake; to Anton and Ian for truly messing up my English language; to Rachel and Fiona for being so helpful; to Sherburn Village Parish Council for providing such wonderful allotments; to all those I have not yet mentioned like Sandra, Holly, Keith, Owen, Norbert, Farid, James, James, Leo, Karen, Vanessa, Matt, Pascale, and my 🚲. And my thanks to Steve.

Contents

List of Figures

Chapter 1

Introduction

There are several things I would like to do in the introduction. First of all I would like to give a brief thought to what the title of the thesis means, which hopefully will give the reader an idea about the content of the thesis. Further I would like to give an introduction to some basic queueing models, then a short introduction to the methods for establishing stability that will be used later and a selective summary of some more queueing networks discussed in the literature - selective since the main focus here is on stability and there is a lot of literature that deals with rather different problems.

But first let us start with the title. The discrepancy, or not, between the linguistic definition and the mathematical meaning of a word can be rather confusing or enlightening (at least for me). Here I will state, and comment on, what I found in the Oxford Dictionary [33].

A queue is

> *"a line or sequence of people or vehicles awaiting their turn to be attended to or to proceed".*

Although we will not look at queues in terms of people or vehicles waiting, the main points we would like to consider are given in this definition - we shall keep these in mind when talking about queues in a mathematical sense. We have customers arriving from somewhere to be processed, the processing takes some time which can lead to waiting times for the customers and a queue starts forming.

We will not look at single queues but queueing networks, networks are *"a group*

or system of interconnected people or things" [33]. In terms of queues, networks arise if a customer has to go through several stages of service before it has received all the necessary attention and can be considered completed. One can think of this in terms of production of goods which require processing at several machines before being finished; or a visit to a doctor which starts at the reception, then one hopes to see the nurse or the doctor, and quite often has to visit the reception again before leaving.

A novelty of our approach is the way in which we control the queueing system. If we have control we have *"the power to influence or direct [...] the course of events"* [33]. We do not specify the control applied but try to keep it as general as possible, which means that a lot of control policies are included in our approach.

The most ambiguous word of those in the title is stability, at least in its mathematical interpretation. To say something stabilises or is stable can mean very different things depending on which branch of mathematics one considers. Here we are only interested in it in the sense of positive recurrence or ergodicity if the queue length process is a Markov chain (more details can be found in Section 1.2). In contrast to this linguistic meaning of the word stable is simply *"not likely to give way or overturn"* [33].

The word criteria is plural for criterion which is *"a principal or standard by which something maybe judged or decided"* [33], we will look for the standards needed in order to decide whether a queueing network is stable.

Let us summarise: we will try to control queueing networks in such a way that they are stable. We will give criteria (necessary and sufficient) to identify the situations when stability is present, possible or not at all achievable.

A PhD in brief or historical PhD briefing: Before we start fully with queues, stability and literature I would like to give a brief overview of how the work that is presented in this thesis came together and state which parts of it are published and where.

The original question was given by M. Menshikov as something like: *What if we*

have two queues and two servers and we can move the servers from one queue to another from time to time with the aim to make the system stable; when is it stable, and how can we classify when the system is stable or not?

The answer to this question is given with the convex hull of the system mean drifts given in Chapter 2 and some of the even earlier results for two queues are summarised in Section 2.4. Given the convex hull it was clear that introducing extra features and parameters to the model, such as Jackson feedback, would not change the stability classification. The idea of the control the we exercise over the system also evolved, starting with what we call pure policies to using (fully) randomised controls. It also became clear the some of the result would be easily extend-able to queueing models with more than 2 queues. At this state my supervisor Iain MacPhee and Misha Menshikov got a bit overexcited and Iain wrote the proof for Theorem 4.3.2 (the proof is here included in Appendix B) while I was away. We finished the first paper [27] and submitted it in January 2005.

I had submitted the work on two queues to a conference (the ASMDA (applied stochastic models and data analysis) held in May 2005 in Brest, France) earlier, so the two queue result can also be found in [26]. Since this conference paper won the IBM student price I was asked whether we have something to be submitted for a special conference issue of *Methodology and Computing in Applied Probability.* This became the second paper [28] about queueing networks with re-entrant lines, examples of this type of models can be found throughout the thesis. In the mean time I had worked on phase type distributions. The question was whether phase type services can be introduced into our model and whether the stability results would still hold in this case. Together with the phase type service I investigated the possibility to introduce discrete time Markovian arrival processes. The results of this can be found in Chapter 3. And in the last couple of month I became extremely interested in the queueing literature, which is why the fruits of this labour will be presented just after an introduction the Lyapunov function or semi-martingale methods.

Since ordering the content of the thesis chronologically in time does not make much sense the order of the content can be explained in the following way. There are

two "directions of travel", the macro direction, from start to finish, is the "queue-direction" and the micro direction, reappearing in some chapters, is the "stability-direction". The queue-direction goes from simple, low dimensional queueing models to more complex networks and higher dimensional processes; while the stability-direction starts with the most general stability result (under very general assumptions about the control) and is then narrowed down to the cases where we can talk about ergodicity or positive recurrence of a Markov chain (when the controls are stationary and Markov or even more restrictive).

A brief summary of content can be given as follows. We start by giving an introduction to queueing models, semi-martingale and Lyapunov function methods as well as giving stating some examples of queueing network discussed in the literature Sections 1.1 to 1.3. We then analyse a very general class of two queue models in detail under several different control policies in Chapter 2, including several examples where we also show how our results relate to existing models. In the proceeding Chapter 3 we consider some generalisations of the two queue model, such as service time of phase type and discrete time Markovian arrival processes. Chapter 4 has the N queue model as its main objective. We show which results from Chapter 2 extend readily into a N dimensional setting and where additional assumptions are needed in order to find similar results. In Chapter 4 we will take a closer look at some of the examples introduced in Section 1.3 such as the generalised Lu-Kumar network.

1.1 A Queueing Model

In this section we will we start by defining what exactly we mean when we say queue and what happens in such a queue. Additionally we will give some basic examples of queueing models and give some intuition about whether they are stable or not.

Generally queues form if customers or jobs need service or processing that takes some time and is only provided by a limited number of servers at stations, usually there is one server or one type of service offered per station (we will also consider stations with several queues and only one server). We say that there are *arrival to*

a queue at some rate. At the time of arrival of a customer three things can happen: the server is idle and the customer is served straight away, the server is occupied or there are customers already present in the queue, in which cases the new arrival joins the queue. We consider systems with so called *infinite buffers*, which means there is no limit to the number of customers that wait in the queue. We also assume that customers are served one at a time and there is a finite but non-zero *service or processing time* which has finite mean. A schematic picture of such a queue is given in Figure 1.1.

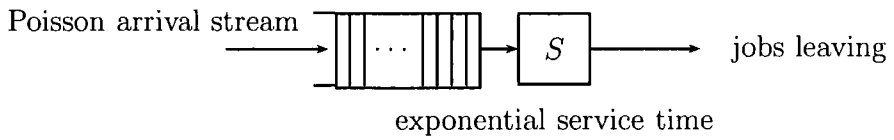


Figure 1.1: A $M/M/1$ queue with infinite buffer.

We will mainly consider queueing models where the arrival rate λ is Poisson and the service times of jobs are independently exponentially distributed with parameter μ. This means that the queue length process denoted by $\mathfrak{X}$ is a Markov chain. A consequence of these assumptions is that the rates of arrival and service time are directly comparable which leads to the following well known result.

Theorem 1.1.1 *Given a Poisson arrival rate λ and exponentially distributed service times with parameter μ, the queue length process $\mathfrak{X}$ is Markov chain and is*

(i) stable (ergodic/positive recurrent) if the workload is $\lambda/\mu < 1$,

(ii) unstable (transient) if $\lambda/\mu > 1$, or

(iii) neither (null-recurrent) if $\lambda/\mu = 1$.

These results about the *workload* or *traffic intensity* λ/μ go back to A.K. Erlang, thus the one server one Markovian queueing model is sometimes referred to as Erlang's model, but more popularly denoted by $M/M/1$ (M for Markovian arrivals and service times and 1 indicating the number of servers).

Intuitively (i) means that on average the time that it takes to serve a customer is shorter than the time interval between successive arrivals. So even if there exists a queue when the service starts it will disappear after some time and once the queue is empty new arrivals will be served without waiting. Part (ii) on the other hand implies that the number of customers waiting will keep increasing. We will show these basic results in the following Section when demonstrating the methods. The condition in (iii) implies that the all the jobs that enter will be served but the queue length might never be really reduced. More details about this and other queueing models can be found in text books such as Asmussen [1].

Once one knows that the system is stable the question is whether one can also find the stationary probability distribution for the underlying Markov chain, a question we will not study for our model, but point out for which models they exist or where attempts have been made.

The first variation of the $M/M/1$ is to have two $M/M/1$ queues running parallel to one another. The necessary and sufficient condition for the stability of such a system is that $\max_i \lambda_i/\mu_i < 1$, i.e. both λ_1/μ_1 and λ_2/μ_2 are smaller than 1. The idea is the same for N $M/M/1$ queues parallel (which we then denote by $M/M/N$) when there is no interaction between N queues arrivals and services.

Sometimes the queueing system denoted as $M/M/c$, is a system with only one Poisson arrival stream λ and c identical servers serving all jobs at rate μ. This means that if there are less than c jobs in the system some of the servers idle. The stability of a one arrival stream c server queue is still straight forwardly given by $\lambda/c\mu < 0$.

The stability of a model that incorporates a network structure, where customers that are processed at one queue can enter another queue and receive additional service there, was first analysed by Jackson [21], in a model now referred to as the Jackson Network.

Example 1.1.1 (Jackson Network) Consider the simplest Jackson Network which has two queues with independent Poisson rates λ_1 and λ_2 and independent, exponentially distributed service rates μ_1 and μ_2. On completion of service at queue 1

the customer either re-enters queue 1, joins the end of queue 2 or leaves the system with probabilities p_{11}, p_{12} or $p_{10} = 1 - (p_{11} + p_{12})$ respectively. There are similar routing probabilities for customers that have finished at queue 2. As long as $p_{i0} > 0$ for $i = 1$ or 2 there is a possibility for the system to be stable (i.e. some customers certainly leave the system so that the queue lengths can be reduced). From Jackson's [21] results, which hold for N queues as well, we know that the system is stable if $\tilde{\lambda}_i = \lambda_i + \mu_i p_{ii} + \mu_j p_{ji} < \mu_i$ for i and $j = 1, 2$ and $i \neq j$. Just like the conditions above this means that the total arrival rate, from the outside and from feed backs, is smaller than the service rate at each queue. In his paper Jackson [21] also gives the stationary probabilities for this Markov chain.

The following Example 1.1.2 investigates an additional idea that we will take even further in our model - the idea that one server can help the other if there are no customers waiting in his queue. In the previous examples, and for most of those that will follow, it is assumed that a server is allocated to a specific queue and cannot move to the other queues.

Example 1.1.2 (Modified Jackson Network) Foley and McDonald [14] consider a two queue Jackson network as in Example 1.1.1 given the addition that once queue 2 is empty server 2 can move to queue 1 and *help* serving customers there, thus increasing the service rate to $\mu_1 + \mu_2$. The authors [14] show that this increases the stability region, we will discuss this model later.

Queueing models and networks present in the literature are often motivated by some practical example which has been observed in reality, one of these is

Example 1.1.3 (Join the shortest queue or JSQ) Consider two queues with service times as in Example 1.1.1 but only one Poisson arrival stream with parameter λ. The customers that arrive in the system make a decision at the time of arrival to join the shorter one of the two queues - like a customer in a shop would do given the choice. If the two queues are the same length the customers are equally likely to join either of the two queues. This queue system is stable if $\lambda < \mu_1 + \mu_2$. Flatto and McKean [12] consider this model and the stationary probabilities where the two servers that have identical service time distribution. More details about the

stationary probabilities can also be found in Kurkova and Suhov [23]. Foley and McDonald [13] investigate the stability of the JSQ model for N queues.

Example 1.1.4 (Load-balanced network) Kurkova [22] considers a combination of the two models with two queues given in Examples 1.1.1 and 1.1.3, in that in addition to two dedicated arrival streams to the two queues there is another Poisson stream with parameter λ which is routed to the shorter of the two queues. Additionally some of the customers fed back into the queues join the shorter queue also. We will consider this model and a generalisation of it in more detail in Chapter 2.

In order to consider stability of more complex networks we will now take a closer look at stability criteria.

1.2 Stability criteria for countable Markov chains

In this section we would like to give a brief introduction to the mathematics we use to show stability or not of our queue length (or queueing) process. These methods are often referred to as semi-martingale or Lyapunov function methods (also called test functions see Meyn and Tweedie [29]). The main idea goes back to Foster's criterion [18] for ergodicity of Markov chains. A more general formulation of the theory gathered in the book by Fayolle, Malyshev, Menshikov [11]. Meyn and Tweedie [29] look at similar criteria when the Markov chain is not necessarily countable. Another book that summarises some of the earlier semi-martingale method results and has some more details about queueing models is Asmussen [1].

A continuous time queueing process is a combination of two random sequences: the inter arrival times of customers and their service times. The queue length process associated with the number of customers waiting in such a queueing system in a continuous time setting will be denoted by $\mathfrak{X} = \{\mathfrak{X}(t) : t \geq 0\}$. We will concentrate on the case where the inter arrival times are exponential with parameter λ and the service times are exponential with parameter μ.

The process we will consider mainly is the discrete time queue length process denoted by $\Xi = \{\xi(n) : n \geq 0\}$, where n indicated the discrete time units. Ξ is the

embedded continuous time queueing process $\mathfrak{X}$ observed at the times of events, the events being arrivals or departures of customers and a null event when the process remains in the same state, see Section 2.2 for more details. Ξ lives on the state space $\mathbf{Z}_0^N \equiv \{(x_1, \ldots, x_N) \in \mathbf{Z}^N : x_i \geq 0 \text{ for all } i = 1, \ldots, N\}$ and a vector in this state space is usually denoted by $\alpha = (x_1, \ldots, x_N) \in \mathbf{Z}_0^N$. We will assume Poisson arrival rates and exponentially distributed service times which means that Ξ will be Markov. Once we introduce control policies to the queueing system these also need to fulfil some requirements for the process to remain Markov, this is discussed in Section 2.2. What matters for now is that if we do not have a Markov chain we will talk about stability, but if we do have a Markov chain then stability and ergodicity (or positive recurrence) are the same. Unless specified otherwise the state space is $\mathbf{Z}_0^N$ where N is the number of queues in the model.

We will start with some very general stability results and then look at the stronger results we get when considering Markov chains. Towards the end of this section we will show more specifically how the stability of queueing systems can be analysed, introducing some notation and conditions.

Martingales: Since one of the methods presented here is the semi-martingale method, we give a very brief summary of martingales. Let $\{Z_n\}_{n \in \mathbf{Z}_+}$ be a sequence of real valued random variables with finite mean. We say $\{Z_n\}$ is *adapted* to an increasing family of σ-fields $\{\mathcal{F}_n\}_{n \in \mathbf{Z}_+}$ if Z_n is $\mathcal{F}_n$-measurable for each n. For $n \in \mathbf{Z}_+$ the sequence $\{Z_n\}$ is called

a *martingale* if $\mathbf{E}(Z_{n+1} \mid \mathcal{F}_n) = Z_n$ a.s.,

a *supermartingale* if $\mathbf{E}(Z_{n+1} \mid \mathcal{F}_n) \leq Z_n$ a.s., or

a *submartingale* if $\mathbf{E}(Z_{n+1} \mid \mathcal{F}_n) \geq Z_n$ a.s.

A lot of probability text books have information about martingales, for more details see for example Doob [9] or Meyn and Tweedie [29].

Stability: We consider stability as introduced in Chapter 2 of Fayolle, Malyshev, Menshikov [11], or loosely speaking in the sense that, *for a sequence of non-negative real valued random variables, the time to reach a finite ball $\mathcal{D}$ around the origin is*

finite, given that it started outside this ball. The three Theorems stated here are all taken from FMM [11] and the reader is referred to this source for the proofs.

Theorem 1.2.1 below gathers together three of semi-martingale results from Fayolle, Malyshev and Menshikov (FMM) [11] which we will use to help determine the stability or not of our queueing model.

Let $\{S_t, t \geq 0\}$ be a sequence of non-negative random variables with S_0 constant and S_t measurable with respect to the history $\mathcal{H}_t = \sigma(\xi(0), \eta_0, \ldots, \xi(t))$ for $t \geq 1$, where $\xi(t)$ are the realisations of the process and the η_t contain information about the way the system was run at time t under some control. Let $\{N_n, n \geq 0\}$ be an increasing sequence of stopping times adapted to $\{\mathcal{H}_n\}$ with $N_0 = 0$ and let $Y_0 = S_0$ and $Y_n = S_{N_n}$ for $n \geq 1$. Also, for constant $D > 0$, let $\tau = \min\{t \geq 1 : S_t \leq D\}$ and $\sigma = \min\{n \geq 1 : Y_n \leq D\}$. Finally let $\{X_{t\wedge\sigma}\}$ denote a sequence $\{X_t\}$ stopped at σ and I_E be the indicator function of an event E.

Theorem 1.2.1 *(i) If $S_0 > D$ and for some $\epsilon > 0$ and all $n \geq 0$*

$$\mathbf{E}\big(Y_{(n+1)\wedge\sigma} \mid \mathcal{H}_{N_n\wedge\sigma}\big) \leq Y_{n\wedge\sigma} - \epsilon\mathbf{E}\big(N_{(n+1)\wedge\sigma} - N_{n\wedge\sigma} \mid \mathcal{H}_{N_n\wedge\sigma}\big) \quad a.s.$$

then $\mathbf{E}(\tau) \leq S_0/\epsilon < \infty$.

(ii) If $S_0 > D$, the jumps $S_{n+1} - S_n$ are uniformly bounded below and there exists $\epsilon > 0$ and a positive constant b such that for every $n \geq 0$

$$\mathbf{E}\left(\big(S_{n+1} - S_n\big)I_{\{S_{n+1}-S_n<b\}} \mid \mathcal{H}_n\right) \geq \epsilon \quad a.s.$$

then $\mathbf{P}(\tau = \infty) > 0$ and for any $\delta_1 \in (0, \epsilon)$ there exist constants $C = C(S_0)$ and $\delta_2 > 0$ such that for any $n \geq 0$, $P(S_n < \delta_1 n) \leq Ce^{-\delta_2 n}$.

(i) is Theorem 2.1.2 while (ii) combines versions of 2.1.10 and 2.1.7 of FMM [11].

Setting up the process $Y_n = S_{N_n}$ in (i) guarantees the most general setting for stability. It means by observing the process S_n at certain stopping times N_n we get Y_n and if Y_n is a supermartingale this allows us to conclude that the time τ, the time it takes until S_n reaches the finite ball $\mathcal{D}$ around the origin, is finite and even has expectation bounded by S_0/ϵ. Note that if the first moment $\mathbf{E}(S_n \mid \mathcal{H}_n)$ is finite then the process remains bounded in mean. Part (ii) states that if S_n is a

submartingale there is a positive chance that the process will never reach $\mathcal{D}$, and that in fact the process goes to infinity at least linearly fast.

Positive recurrence and transience of Markov chains: Given an irreducible and aperiodic Markov chain on a countable state space, we can talk about the positive recurrence/ergodicity and transience of a Markov chain.

We say a Markov chain $\Xi = (\xi_0, \ldots, \xi_n \ldots)$ is positive recurrent if for all $\alpha \in \mathbf{Z}_0$ we have $\mathbf{P}(\xi_n = \alpha \text{ for infinitely many } n) = 1$ and the expected return time to α is $\mathbf{E}(\tau_\alpha) < \infty$.

One way to show the positive recurrence of a Markov chain was introduced by Foster [18]. The Theorems 1.2.2 and 1.2.3 below are Theorems 2.2.3 and 2.2.2 respectively from FMM [11].

Theorem 1.2.2 (Foster's criterion) *A Markov chain Ξ is positive recurrent (ergodic), if and only if there exists a positive function $f : \mathbf{Z}_0^N \to \mathbf{R}_+$, some $\epsilon > 0$ and a finite set $\mathcal{D} \subset \mathbf{Z}_0^N$ such that*

$$\mathbf{E}\left(f(\xi_{n+1}) - f(\xi_n) \mid \xi_n = \alpha_j\right) \leq -\epsilon, \text{ for } \alpha_j \notin \mathcal{D},$$

$$\mathbf{E}\left(f(\xi_{n+1}) \mid \xi_n = \alpha_i\right) < \infty, \text{ for } \alpha_i \in \mathcal{D}.$$

Here f is called a Lyapunov (or test) function and $f(\xi_n)$ is a supermartingale. The Lyapunov function projects the, often higher dimensional, process Ξ into a one dimensional space where one can see more easily whether it increases or decreases in expectation. Often the main problem when using this method is to find an appropriate Lyapunov function. Due to the nature of the processes and controls we consider most of the Lyapunov functions we will use are linear. Note that we can set $S_n = f(\xi_n)$ in Theorem 1.2.1(i) which, if the first moment $\mathbf{E}\left(f(\xi_{n+1}) \mid \xi_n = \alpha_i\right)$ is finitem would then also yield the stronger result of Theorem 1.2.2.

Theorem 1.2.3 (Transience) *A Markov chain Ξ is transient, if and only if there exists a positive function $f : \mathbf{Z}_0^N \to \mathbf{R}_+$, some $\epsilon > 0$ and a finite set $\mathcal{D} \subset \mathbf{Z}_0^N$ such that*

$$\mathbf{E}\left(f(\xi_{n+1}) - f(\xi_n) \mid \xi_n = \alpha_i\right) \leq 0, \text{ for } \alpha_i \notin \mathcal{D},$$

$$f(\alpha_k) < \inf_{\alpha_j \in \mathcal{D}} f(\alpha_j) \text{ for at least one } \alpha_k \notin \mathcal{D}.$$

Note: A Markov chain is either positive recurrent, transient or null-recurrent. Briefly null-recurrence means that the Markov chain is recurrent (we have a function $f : \mathbf{Z}_0^N \to \mathbf{R}_+$ and a finite set $\mathcal{D} \subset \mathbf{Z}_0^N$ such that $\mathbf{E}\left(f(\xi_{n+1}) - f(\xi_n) \mid \xi_n = \alpha\right) \leq 0$, for $\alpha \notin \mathcal{D}$), but not ergodic, we can find another positive function $g(\alpha)$ which is a submartingale for $\alpha \notin \mathcal{D}$, given some additional constraints, see for example Theorem 2.2.8 in FMM [11]. We will consider null-recurrent cases in more detail in Section 2.3.3.

After this short description of general stability criteria we will now return to the queueing models and introduce notions such as the mean drift to illustrate how the theorems above relate to the actual behaviour of the models we have already discussed in Section 1.1.

A discrete time queue length process: Consider the continuous time $M/M/1$ queue, the intensities (or rates) of the inter arrival and service times of the queue length process $\mathfrak{X}$ are given by λ and μ respectively. We would now like to consider the embedded discrete time Markov chain Ξ with transition probabilities $p_{\alpha\alpha+1}$ and $p_{\alpha\alpha-1}$, i.e. we observe the continuous process at the times of events such as arrivals and depatures. These probabilities are given by $p_{\alpha\alpha+1} = \frac{\lambda}{\lambda+\mu}$ and $p_{\alpha\alpha-1} = 1 - p_{\alpha\alpha+1} = \frac{\mu}{\lambda+\mu}$ (for all states α but $\alpha = 0$). We will introduce uniformising or embedding for more complicated models and the related processes in Section 2.2.

Note: The Lyapunov function methods also hold for continuous time processes, choosing the discrete version instead is mostly down to personal preference.

Given that we can observe the embedded discrete time queue length process Ξ on $\mathbf{Z}_0^N$ with states α and β we set the following two conditions for the transition probabilities of jumps $p_{\alpha\beta}$:

Condition B (Boundedness of jumps): $p_{\alpha\beta} = 0$ for $||\alpha - \beta|| > d > 0$,

where $||\alpha|| = \max_i |x_i|$ with $\alpha = (x_1, \ldots, x_N)$. In fact for all the models we consider the jumps are only on $\mathbf{Z}_0^N$ and most of the models have jumps bounded by $d = 1$ in both directions.

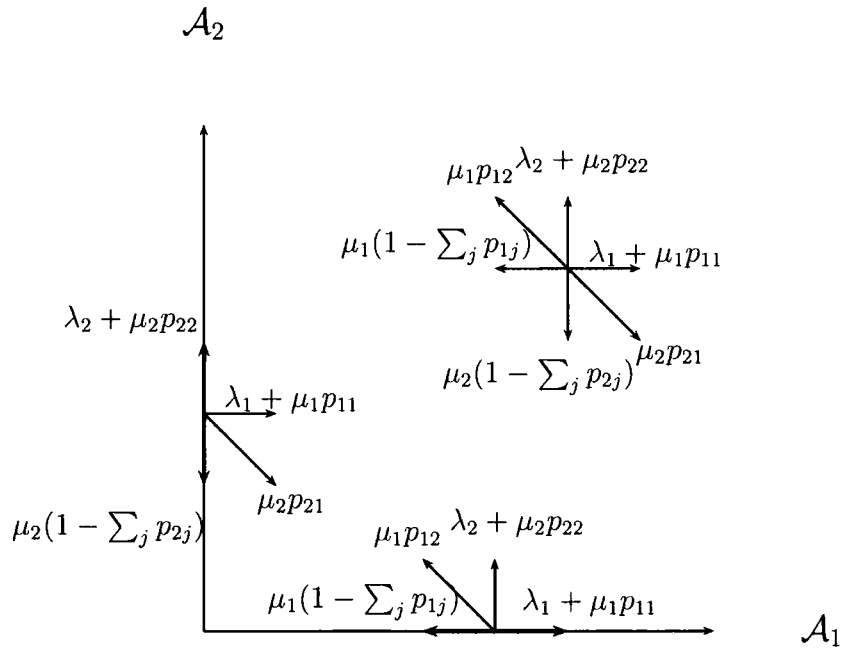


Figure 1.2: Rate of jumps for the two queue Jackson network.

Condition H (Homogeneity): There is a finite number of blocks $\mathcal{B}$ which partition the state space $\mathbf{Z}_0^N$ (see Section 2.2) and the jump distribution $p_{\alpha\beta}$ is the same for each α, β on a block $\mathcal{B}$, with $\alpha - \beta$ constant.

We introduce the notation $\mathcal{A}_i$ to denote the axis in i direction. Consider the two queues with two serves. We can observe three different jump distributions, the jumps in the interior when both servers are serving customers and the jumps on the boundary of the state space, the axes, when one of the servers is idling while waiting for a new customer to arrive. Figure 1.2 shows the three jump distributions in the case of the two queue Jackson network.

Mean drifts: Given the jump distribution we can now define the one step mean drift vectors for the discrete time process from $\alpha \in \mathbf{Z}_0^N$ as

$$M(\alpha) = (M_1(\alpha), \ldots, M_N(\alpha)) = \sum_{\beta} (\alpha - \beta) p_{\alpha\beta}.$$

Looking at the mean drifts of the queue length process Ξ we can immediately get some idea about whether Ξ is stable or not. Consider the $M/M/1$ queue again. The mean drift, while there are customers in the queue, is $M(\alpha) = \frac{1}{\lambda+\mu}(\lambda - \mu)$ and

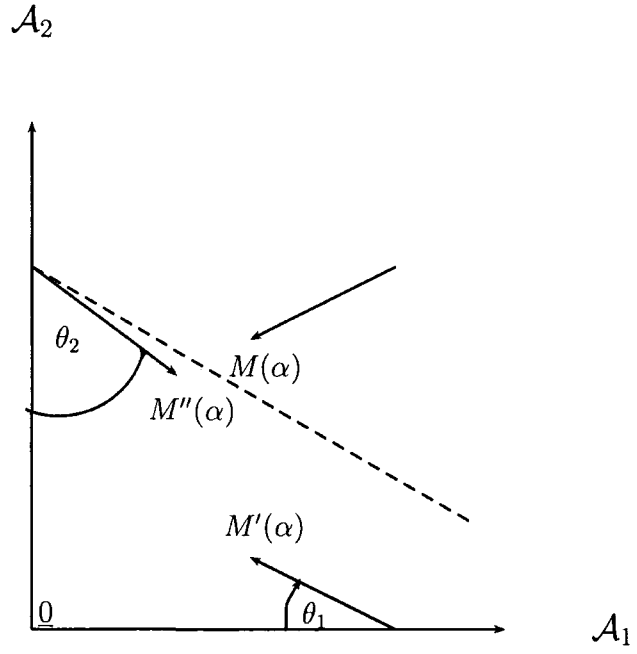


Figure 1.3: Possible mean drift vectors of a two queue model.

from Theorem 1.1.1 we know that the process is positive recurrent only if $\lambda/\mu < 1$ which is equivalent to a negative mean drift $M(\alpha) < 0$.

If we consider the two queue two server model again we get three different mean drift vectors for the three blocks:

$$\begin{aligned} M(\alpha) = (M_x(\alpha), M_y(\alpha)) &= \tfrac{1}{\lambda_1+\lambda_2+\mu_1+\mu_2}(\lambda_1 - \mu_1, \lambda_2 - \mu_2), \\ M'(\alpha) &= \tfrac{1}{\lambda_1+\lambda_2+\mu_1+\mu_2}(\lambda_1 - \mu_1, \lambda_2) \quad \text{and} \\ M''(\alpha) &= \tfrac{1}{\lambda_1+\lambda_2+\mu_1+\mu_2}(\lambda_1, \lambda_2 - \mu_2) \end{aligned}$$

for the interior, no customers in queue 2 and no customer in queue 1 respectively (the vectors M' and M'' are referred to as reflexion vectors).

Given set of parameters lets consider the three mean drift vectors to assess the stability of the process Ξ. Assume that the mean drifts in Figure 1.3 are the vectors we got from a given set of parameters in a two queue model. We can see that $\lambda_i < \mu_i$ for $i = 1, 2$ as $M(\alpha)$ is pictured with $M_x(\alpha), M_y(\alpha) < 0$.

Now consider the angles θ_1 and θ_2 that the reflexion vectors M' and M'' make with the respective axes. If the sum of these angles is $\theta_1 + \theta_2 < \pi/2$ and $M_x, M_y < 0$, then we can find a linear Lyapunov function (in Figure 1.3 the level curve of this

function is indicated by the dashed line) that shows that Ξ is positive recurrent. In fact if $M_x, M_y < 0$ FMM [11] Theorem 3.3.1(a) shows that the Markov chain is stable if and only if $M_x M_y' - M_y M_x' < 0$ and $M_y M_x'' - M_x M_y'' < 0$. If the sum is $\theta_1 + \theta_2 > \pi/2$ and/or M_x or $M_y \not< 0$,the positive recurrence of the underlying Markov chain depends on the angles of M with respect to the two reflexion vectors, see Theorem 1.2.4 below.

We can see now why the idea of Foley and McDonald [13], see Example 1.1.2, that one server helps the other can increase the range of parameters for which the system is stable; boosting the service at the non-empty queue 1 to $\mu_1 + \mu_2$ changes the reflexion vector M' so that the first component is now $\lambda_1 - (\mu_1 + \mu_2)$, this means the angle θ_1 is smaller and thus there are more cases where $\theta_1 + \theta_2 < \pi/2$.

Lyapunov functions can be interpreted in a very natural way. The linear function f can be chosen in such a way that all three mean drifts shown in Figure 1.3 point *inwards* (towards the origin) just like they do from the dashed line. So if we attach the function to our process we would expect that a decrease in the function value from time n to time $n+1$, which is exactly what Foster's criterion says. A complete classification of the stability or not of a Markov chain in $\mathbf{Z}_0^2$ with three homogeneous blocks (the interior and the two axis of $\mathbf{Z}_0^2$) is given in FMM [11] Chapter 3.3, Theorem 1.2.4 below is a version of Theorem 3.3.1.

Theorem 1.2.4 *Assume that condition* **C** *and* **H** *are satisfied.*

(1) If $M_x \geq 0$ and $M_y < 0$, then the Markov chain Ξ is

(a) positive recurrent if $M_x M_y' - M_y M_x' < 0$

(b) transient if $M_x M_y' - M_y M_x' > 0$.

(2) If $M_x < 0$ and $M_y \geq 0$, then the Markov chain Ξ is

(a) positive recurrent if $M_y M_x'' - M_x M_y'' < 0$

(b) transient if $M_y M_x'' - M_x M_y'' > 0$.

(3) If $M_x \geq 0$, $M_y \geq 0$ and $M_x + M_y > 0$ then the Markov chain Ξ is transient.

What we consider in detail in Chapter 2 are two queue models with more than three blocks of homogeneity. We also allow the model parameters such as the arrival rates, service times and feedback probabilities at each of the queues to not necessarily be the same in different blocks. The control of the system, depending on the queue lengths allows us to change the model parameters completely from time to time.

1.3 More queueing models

If we want to analyse more complex queueing models a simple comparison of arrival and service rates can sometimes only lead to a necessary stability condition but not a sufficient one. Often the problems that arise are related also to the way in which the process is run or managed with help of control policies and in some way feasible strategies. The queueing models we consider now are different to those we looked at earlier because there are different control policies one can apply. Our model, which will be introduced later, is different to the existing ones also because of our very general approach to control - we let the question about stability lead us towards a control policy and not examine the stability given a restrictive strategy.

First however we consider some queueing models discussed in the literature[1] We will focus on three points here: (1) polling systems, because the control of such systems is important and the N dimensional Markov chain associated with the queue lengths process is well studied because of its special characteristics; (2) multi-class queueing systems with re-entrant lines, because in some examples, as we show in Section 4.4, the control policies applied have led to problems with stability and the traffic intensity is not a sufficient stability condition any more; (3) maximum pressure or throughput approach, since this method to stabilise a queueing network is a very interesting approach.

[1] I do not intend to give a full review of the queueing literature, the focus is on those models which are interesting with respect to the question of stability, or their control features. For a more complete review of queueing literature please see Stidham [35].

Example 1.3.1 (Polling systems) These are queueing models with the following features: there are $i = 1, \ldots, N$ stations or nodes and each station has an independent Poisson arrival stream with rate λ_i but there is only one server that *polls* the N queues (i.e. asks the N queues whether they have customers). We assume that this server serves customers at queue i with exponentially distributed service times with mean $1/\mu_i$. There are various control policies or disciplines that are applied in order to run such systems. Generally we have two approaches to service: exhaustive (the server serves each queue that he visits until it is empty), or non-exhaustive (there is some rule by which the server stops serving a queue before it is empty); then there are models with finite or infinite buffers; and a choice on whether to do cyclic or non-cyclic polling. We will concentrate exhaustive service and infinite buffers and state some different ways to control the polling of the server.

Cyclic polling means that the server *visits the queues in a given order*, i.e. starting with queue 1 the server remains there until there is no customer left in the queue and then moves to queue 2, and so on until the server has finished the last customer at queue N when it starts again at queue 1. Takagi [36] gives an overview of the literature mainly on these cyclic polling models. For such an N-queue polling system there are $i = 1, \ldots, N$ mean drifts M^i. Each M^i has N components with $M_i^i = \lambda_i - \mu_i$ when the server serves queue i and $M_i^j = \lambda_j$ for $i \neq j$, the remaining $N - 1$ queue which do not receive service. The system is stable if and only if $\sum_i \lambda_i/\mu_i < 1$.

In terms of *non-cyclic* polling we differentiate between deterministic and probabilistic polling (see Takagi [36] and references therein). A deterministic polling strategy can be a given, non-cyclic order (like $1 \to 3 \to 1 \to N \to 2 \ldots$ and so on) in which the server visits the queues.

An example for probabilistic polling is given by Borovkov and Schassberger [5], who consider a server that polls in a Markovian fashion which means that given the server serves station i it will serve station j next with some probability p_{ij}. Foss and Last [16] consider a polling model where the server's decision which queue to serve next depends on the configuration of the customers present in the system. Their polling system also has general service time and it takes the server a non-zero

walking time to move from one queue to another.

Another way to run a polling system is by what is called the *greedy* algorithm. There are several variations of this algorithm but the main idea about the greediness of the server works as follows: after completing the service of the last customer in queue i the server chooses queue j as the next queue to serve, given that queue j is the longest of all queues (or the longest within a neighbourhood of queue i) at this time instance. Foss and Last [17] analyse the stability of a greedy polling system with general service policies and non-zero walking time when the server changes from one queue to another. Their greedy service policy is a variation of the above in that at the instance of polling queue i at time n the server generates a pair of random variables (B_n, C_n) where $C_n < x_i$ (where x_i is the number of customers in queue i) and the server either serves B_n customers and leaves queue i or departs after there are only C_n jobs left in the queue, whatever event occurs first.

For all these different ways to run the polling systems (given that instantaneous switching time) the necessary and sufficient condition for stability can be summed up as *the total workload is smaller than one.* In terms of the mean drifts this means that all M^i have a negative component M_i^i such that $\sum_i \lambda_i/\mu_i < 1$. Another feature of the polling system which makes the issue of stability easier to evaluate is that the queue length process does not tend to "hang around" the boundaries of the state space - once a queue is served the server leaves and the queue will almost surely have new arrivals before the server starts serving it again. The next example we will consider showed that comparison of rates is not a sufficient condition under all reasonable appearing control strategies.

Example 1.3.2 (Queues in series, or re-entrant lines) The idea for this model is rather different from the one above. We assume there is one stream of Poisson arrivals of rate λ that arrive at a series of N servers or service stations. Once a customer has been served at the first server, with exponentially distributed service time at rate μ_1, it is routed to another server i where it is served at rate μ_i and so on. The customer may also be routed back to a server it has visited earlier, until it has completed its service at the N-th server say, after which the customer leaves the

system. The service times can be different at each queue and also depend upon the number of visits a customer has already made to a server. We assume that all customers follow the same route determined by a routing matrix R with entries $r_{ij} = 1$ if customers are routed from i to j and $r_{ij} = 0$ otherwise. The customers are served in the order of arrival at each queue (FIFO - first in first out). Unlike the previous example the workload conjecture does not hold any more, see for example Bramson [4], and one can ask whether this has anything to do with the way we route the customers through the system and how the customers are treated when they revisit a service station. Note that models with *service constituencies* as introduced by Dai [6] include this model.

A popular example of such a queueing system is called the Lu-Kumar model [24]. Lu and Kumar considered the simple routing (see Figure 1.4) which is motivated as a typical manufacturing process, where goods may require the attention at the same machine several times. The model has two stations with one server and two queues each. A customer (with Poisson arrival rate $\lambda = 1$) that enters the system queues at queue 1 at station 1 first, it is served by server 1 at exponentially distributed service times with rate μ_1. Then it joins queue 2 at station 2 and will be served by server 2 at rate μ_2, after completing service the customer remains at station 2 but is now waiting in queue 3 which is also attended by server 2 this time with rate μ_3. Before leaving the system the customer joins the last queue, queue 4, this time at station 1 where server 1 serves customers at rate μ_4. Given these parameters and that both servers are working we can have four mean drift vectors

$$\begin{aligned}
M^1(\alpha) &= \tfrac{1}{1+\sum_i \mu_i}(1-\mu_1,\ \mu_1-\mu_2,\ \mu_2,\ 0),\\
M^2(\alpha) &= \tfrac{1}{1+\sum_i \mu_i}(1-\mu_1,\ \mu_1,\ -\mu_3,\ \mu_3),\\
M^3(\alpha) &= \tfrac{1}{1+\sum_i \mu_i}(1,\ -\mu_2,\ \mu_2,\ -\mu_4),\\
M^4(\alpha) &= \tfrac{1}{1+\sum_i \mu_i}(1,\ 0,\ -\mu_3,\ \mu_3-\mu_4)
\end{aligned}$$

The obvious problem which arises here is that one has to decide how to split the attention of the two servers between the two queues that each of them has to serve. The usual control or strategy that is applied is called a priority scheduling. For example at station 1 serving queue 1 might have priority over serving queue 4, while are station 2 customers waiting in queue 2 has priority over those in queue 3.

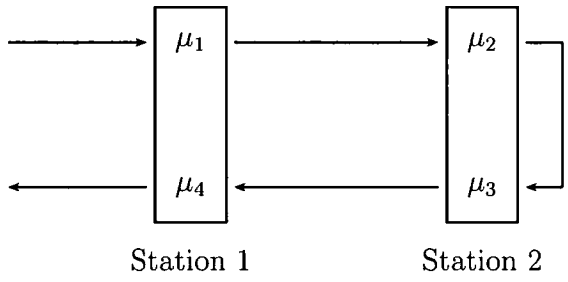


Figure 1.4: The Lu-Kumar network

We will now review some of the work that has been done on the Lu-Kumar network or variations of it. The Lu-Kumar network [24] has received a lot of attention. Bramson [4] considers a variation of this system where the customers revisit queue two over and over again. The phenomenon that occurs for the Lu-Kumar network, and is amplified in Bramson's [4] variation of the two station network, is the following: server 1 is left idling while server 2 serves the queues at his station, then once the jobs have received the last service at station 2, server 1 is over occupied at queue 4, which in turn means that it does not serve the jobs waiting at the first queue that would be fed into station two, so server 2 is forced to idle. What we observe here (that is very different to the polling system) is that the process Ξ "gets stuck" close to the boundary of the state space, which, as we will see in Chapter 4 this is manily down to poor choice of controls. The forced idling in this two station network was also observed by Dai [6], who considers the Harris ergodicity of a general class of multi-class queueing networks, via fluid limits, where each server serves a constituency of queues and jobs are routed from one queue to another.

Down and Meyn [10] use a fluid limit approach to find piecewise linear test functions (or Lyapunov functions) for re-entrant networks like the tandem queue (where all jobs from the first queue get fed into the second one). They also consider the Lu-Kumar and similar networks and come up with a buffer priority programming approach to find a range of service rates for which the system is stable. The stability or not of another variation of the two station re-entrant line network is considered by Dai et al. [7]. Niño-Mora and Glazebrook [32] consider what they call a generalised

Lu-Kumar network; it has additional arrivals from the outside at station two and the possibility for a proportion of the jobs to go from queue 1 to queue 4 (avoiding station two) and to leave the system after being served at queue 3. This model and the original Lu-Kumar network are considered in more detail in Chapter 4 where we show when there exists a randomised control policy that makes the system stable. The main idea in our approach is that after some or all queues are empty we wait for arrivals and change the servers around until all the queues populate again. Some two queue re-entrant lines are also discussed in Section 2.5.

The main results for our model in Chapter 4 can only deal with what is called *multi-class* queueing networks if the customers that require different service times at one station wait in different queues, we will discuss later in what way our approach is more general than the multi-class idea. We give some results under very general control policies (Theorem 4.2.1 and 4.2.2) which hold for a multi-class queueing model like the one in the following example.

Example 1.3.3 (Another multi-class queueing system) Foss and Chernova [15] study a multi-class model with N servers, general independent inter arrival times and service times and state dependent routing. They give stability criteria for join the shortest workload routing in cases where the service times are either server or class dependent and for the case of two queues where service times can be both server and class dependent. Towards the end of their paper they consider an interesting example with $N = 3$ queues and an arrival stream to each pair of queues where the service times depend not on the queue but on the routing decision i.e. for any cyclic permutation σ of 1, 2, 3 a job arriving for queues σ_1 and σ_2 has service time distribution F_l if routed to σ_1 and F_r if routed to σ_2.

The following class of models approaches the idea of stability from a different direction - that of maximum throughput or the pressure in a queueing network. The idea is that under some conditions there is a service allocation policy in the set of all feasible policies for which the throughput of customers is maximised.

Example 1.3.4 (The Generalised Constrained Queueing System) Tassiulas and Bhattacharya [37] considered this general network with service scheduling which

was published in 2000[2]. The network has $N+1$ nodes where the $i = 1, \ldots, N$ nodes are stations, or queues in our language, while the $N+1$-st node represents the outside world in that every customer that reaches this node is considered as having left the queueing system. There is an independent Poisson arrival stream for each node i, while the service time distribution is general and identical within each node i but different for different nodes. There are $k = 1, \ldots, K$ servers which can be allocated to the different nodes where they serve customers according to the service time distribution required for the node. After being served by server k at node i the customer might go to node j with probability $p(i,k,j)$. Tassiulas and Bhattacharya [37] define a $N \times K$ scheduling matrix $\mathcal{U} = \{u_{ik}\}$ where $u_{ik} = 1$ means that server k can serve customers at node i, $u_{ik} = 0$ otherwise. A service allocation is feasible if it guaranties that each server serves only one customer at a time and is still feasible if servers idle at empty queues.

Another more recent, in some ways more general but very similar queueing model with a different approach to control, is discussed below.

Example 1.3.5 (Stochastic Processing Networks) Stochastic processing networks were introduced by Harrison, see for example [20]. What we are interested in however is the paper by Dai and Lin [8] which looks at maximum pressure policies to control these type networks so that they are stable. Their model is similar but more general to Tassiulas and Bhattacharya [37] above. There are $N+1$ buffers or nodes, K processors and J what they call activities (this is most similar to what we will introduce as *management regimes* η in Chapter 2). Each activity j can process customers at a number of processors. They distinguish one input buffer 0 from the N service buffers and a number of the processors are only input processors. At each processor activities can be allocated, and the processing time of a customer depends on the number l of customers already processed under activity j. After completion of processing time the customer can be routed to another node. The activities are allocated in such a way that the customers are processed under maximum pressure,

[2]Unfortunately I had been unaware of this paper until Tassiulas sent me a copy in July 2006 after he had read our paper [27]

and an allocation is only feasible if no processor is idle, i.e. no activity is wasted at an empty buffer. Some of the more general features of Dai and Lin's [8] model are that it can handle activities which simultaneously process customers and that the processing time depends on the activity, not on the buffer or processor.

We will consider the model and the stability results which are given in Tassiulas and Bhattacharya [37] and compare them with our approach in Chapter 4. We will also compare our model and the assumptions made over control with Dai and Lin [8]. In particular we will look at customer types that require the attention of several servers simultaneously.

Chapter 2

The Two Queue Model

In this chapter we analyse a system which has two queues with arrival streams and servers that can be configured or managed in several ways. Our main aim is to identify conditions under which we can give a queue length dependent policy for choosing the system configuration that guarantees the stability of the queue length process. The queueing model in Section 2.1 is one example of a Markov chain and we use it to demonstrate how our results, which are true for all Markov chains, apply.

We start with the two queue model as we can get the most explicit results when the queue length process lives in a two dimensional state space. In terms of different queueing systems and their behaviour (like Jackson networks, join the shortest queue and so on) modelling two queues can often give a good idea about what will happen with more queues, while being easier to understand. The results given in this chapter are published in [26] and [27], only Section 2.3.3 is new. Those results that can be extended to more than two queues will be given in Chapter 4 .

As established earlier the most basic queueing model has arrivals and non-zero service times, which can lead to a queue forming. In this chapter we give a complete stability classification under control policies for a general two queue system with multiple service regimes, a dedicated traffic stream for each queue, and a further stream which can be routed to either queue with feedback of completed customers. The customers in one queue are of the same *class* in the sense that we do not keep track from where they joined the queue, but not all customers in one queue will necessarily receive the same service time because this and the feedback probabilities

depend upon the configuration of the servers. We will define the queue length process and the events that we observe. Several different levels of control of the service regimes are considered. We define the mean drifts of the queue length process. Given the mean drifts we identify four exclusive cases that the system can fall into. These cases are directly related to our main results about stability. We show these using the semi-martingale methods given in Fayolle, Malyshev and Menshikov [11] and described in Section 1.2. We start with stability criteria under very general control policies and also consider some null-recurrent cases. All cases on a very low level of control are discussed in Section 2.4 our results in this section generalise those of Kurkova [22]. We give examples throughout and concentrate on models with re-entrant lines, such as two queues in tandem in the last section of this chapter.

2.1 Model Parameters

In this Section we will introduce the parameters of the queueing model, such as arrival, service and feedback rates, introduce the assumptions that we make and give a basic example.

Arrival Rates: The queues have independent Poisson arrival streams with rates $\lambda_i \geq 0$, $i = 1, 2$. There is an independent Poisson arrival stream with rate $\lambda \geq 0$ of customers that can be sent to either queue, we will call this the *routable* arrival stream. We denote by $s = i$ with $i = 1, 2$ whether the jobs from the routable arrival stream are sent to queue 1 or 2 respectively.

Introducing the additional *routable* arrival stream means we can for example change the way in which we route the arrivals depending on the lengths of the two queues. A well known routing rule for queueing models with such an arrival stream is to *join the shortest queue* (JSQ).

Two queue models like *tandem queues* where $\lambda_2 = \lambda = 0$ and only $\lambda_1 > 0$ (i.e. only one queue has external arrivals) are considered separately in Section 2.5.

Service Times: We allow several ways to provide service to the customers waiting in the queues by introducing service configurations k. This means that under a service configuration $k = 1, \ldots, K$ a specific service time is offered at each queue, k

can include configurations under which no service is provided at one or both queues. We assume that there are distinct service configurations of distinct service rates, for example the configuration changes due end of service at an empty queue.

We assume all customers are served in the order in which they join a queue but that their service times depend upon their queue and the service scheme k in force while they are being served. Under server configuration k, at most one customer is in service at each non-empty queue and all customers in queue i have independent, exponentially distributed service times with rate μ_{ki}, $i = 1, 2$. The μ_{ki} may take any non-negative values so they may vary with k for any queue i, but we make the following

Assumption A1 (Efficient service): We allow only *efficient* service configurations, so whenever the queue i is empty we only permit configurations k where $\mu_{ki} = 0$.

On the other hand we do allow the use of configurations with $\mu_{ki} = 0$ when queue i is not empty.

No service is possible when the system is empty but as our main interest is in stability we are not concerned with what happens on any finite neighbourhood of the origin $\underline{0} = (0, 0)$.

Management Regimes: Given that the arrival stream λ can be routed to either queue by $s = 1, 2$ and that we can choose the service configuration k we define the finite collection $\mathcal{R}$ of overall *management regimes* whose members are the pairs $\eta = (k, s)$.

Our aim to control the system so that it is stable is directly related to these management regimes and the ability to change from one management regime η to another. For now we will make the following

Assumption A2 (Zero switchover times): We can instantaneously switch between different management regimes η at the instants just after changes to queue lengths.

Feedback Probabilities: In addition to the parameters above the system has Jackson-type feedback with probabilities that depend upon the current management regime η. Any job that completes service at queue i under regime η independently

enters queue i' with probability $p^{\eta}_{ii'}$, $i' = 1, 2$ or leaves the system with probability $p^{\eta}_{i0} \equiv 1 - (p^{\eta}_{i1} + p^{\eta}_{i2}) \geq 0$.

Example 2.1.1 (Two queues and two servers) In a simple case of the model described above we have two servers for the two queues. Each server S_i can be used to process customers at either queue (Q_i) which it does at rate μ_{ki}. The model and four different service configurations are illustrated in Figure 2.1. It implies that no server idles unless as long as there are customers in the system so there is no configuration with one server working at Q_i alone while Q_j is empty. Thus given our efficient service assumption A1, the regime where both servers are at Q_1 must be used when Q_2 is empty but might also be used when both queues contains customers.

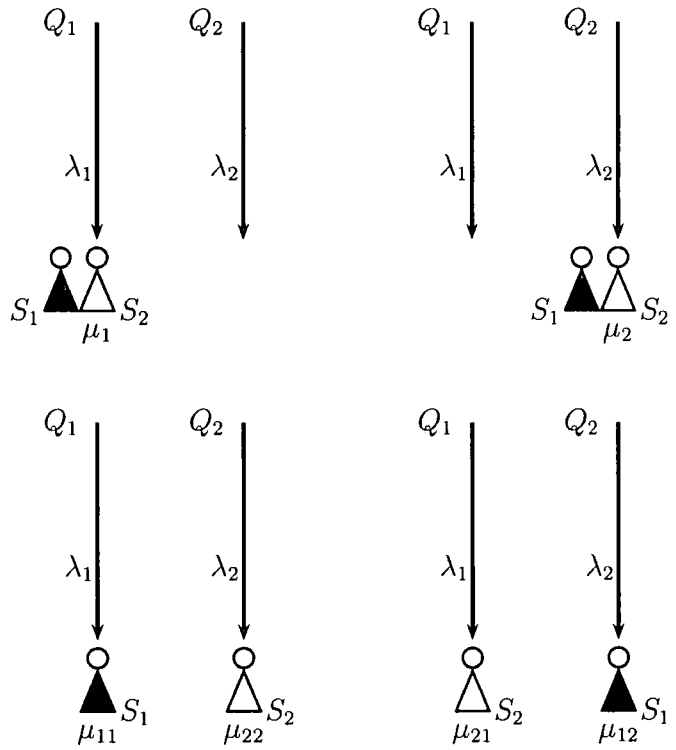


Figure 2.1: The four basic service regimes for two queues and two servers

The four service regimes are denoted by $ki = 12$, 21, 11 and 22 with service rate pairs (μ_{11}, μ_{22}), (μ_{21}, μ_{12}), $(\mu_1, 0)$ and $(0, \mu_2)$ respectively. Note that μ_1 is not

necessarily $\mu_{11}+\mu_{21}$ but can be bigger or smaller than the sum of the two service time rates. If we introduce a routable arrival stream which can be routed to either of the two queues we have eight management regimes $\eta = (k, s)$. Introducing feedback will not change the number of service regimes, but the feedback depends on the regime η in force. For example under $ki = 1\,1$ only customers at Q_1 receive service thus giving feedback probabilities p^{η}_{11} and p^{η}_{12} while customers leave with $p^{\eta}_{10} = 1 - (p^{\eta}_{11} + p^{\eta}_{12})$.

Other server configurations are possible. We could have more that two servers giving more than four service configurations k, or a server with no jobs may be switched to some background tasks and its service rate drops to $\mu_i = 0$, or it may be possible to boost the service rates at some cost or by contrast, the two servers may hinder each other when working at a single queue resulting in a service rate μ_1 which is less than μ_{11} or μ_{21}.

The question we consider is whether for such a model with a given set of parameters, the management regime can be changed from time to time to ensure that the queue lengths remain stable or whether the queue lengths must grow indefinitely regardless of how the system is managed. In the following section we define the queue length process and give details about the management and control of the system.

2.2 Definitions

This Section gives some of the important definitions we need in order to describe our results. First we define the discrete time queue length process by uniformising the continuous time process, then we define control and describe the classes of control policies that we wish to investigate. Given the control we can define the mean drifts of the queue length process and in the last part of this section we define different sets of the state space which will be used frequently throughout the thesis.

2.2.1 Uniformising the process

The Lyapunov function results sketched in Section 1.2 are described in terms of discrete time stochastic processes. Although the same results can be obtained for

continuous time it is convenient to study the discrete time process as *uniformising* has the positive side effect of simplifying the comparison of the behaviour of the queue length process under different management regimes η.

Consider the M/M/1 queue introduced in Section 1.1 which is equivalent to a birth and death process with constant birth and death rates. The birth rate is λ and death rate is μ. We also saw that the transition probabilities of the discrete time process are given by $p_{01} = 1$ and $p_{\alpha\alpha+1} = \frac{\lambda}{\lambda+\mu}$ as the probability of birth and $p_{\alpha\alpha-1} = 1 - p_{\alpha\alpha+1} = \frac{\mu}{\lambda+\mu}$ for all $\alpha \in \mathbf{Z}_+$.

The idea for uniformising our continuous time queueing process $\mathfrak{X}$ works in a rather similar way, only we now have more parameters due to two queues and k service configurations. More details about *uniformising* and the equivalence of the continuous and the discrete time process can be found in Serfozo [34], but here we give a brief outline.

Assume that the continuous time queueing process $\mathfrak{X}$ under a regime η has a generator matrix $\mathbf{A}^\eta = (a_{ij})^\eta$ with up, down and diagonal transitions at rates as given in Figure 2.2. The exponential rate for remaining in any given state is given by $(\lambda + \lambda_1 + \lambda_2 + \mu_{k1} + \mu_{k2})$. Using Serfozo's [34] approach we change the generator matrix $\mathbf{A}^\eta$ into a matrix of transition probabilities $\mathbb{A}^\eta = (p_{ij})^\eta$. We choose a constant

$$\rho \geq \max_k \{\lambda + \lambda_1 + \lambda_2 + \mu_{k1} + \mu_{k2}\} \tag{2.1}$$

and divide all off-diagonal elements a_{ij} for $i \neq j$ of $\mathbf{A}^\eta$ by ρ, so that $p_{ij} = a_{ij}/\rho$. For the diagonal elements a_{ii} we introduce what is called a *null* or *bell* event which has exponential inter-event times with rate $p_{ii} = \rho - (\lambda + \lambda_1 + \lambda_2 + \mu_{k1} + \mu_{k2})$ at any given queue lengths when regime $\eta = (k, s)$ is used. So the total event rate has the same value ρ in all states under all regimes which makes the process dynamically comparable under the different regimes. Observing the continuous time process when jumps occur gives the embedded Markov chain. This yields the matrix of transition probabilities $\mathbb{A}^\eta$ of the discrete time queue length process which is equivalent to the continuous time version.

From now on we consider the *uniformised discrete time process* Ξ on the state space $\mathbf{Z}_0^2 \equiv \{(x, y) \in \mathbf{Z}^2 : x \geq 0, y \geq 0\}$. The process is obtained by considering the

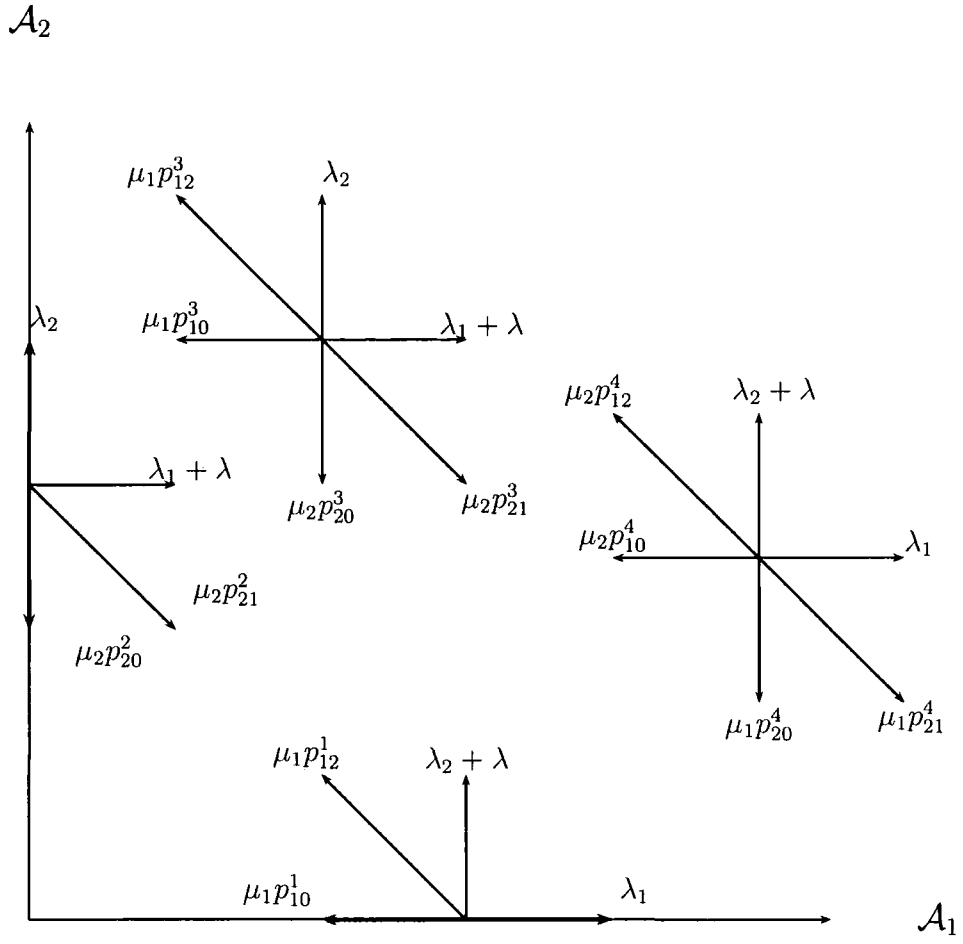


Figure 2.2: Typical jump rates in the two queue model.

queue lengths at bell events, arrival times of new jobs and at service completions and consequent re-entry to queues. The queue length process Ξ has jumps of the form $\pm e_i$ and $\pm(e_1 - e_2)$ where e_i denotes the unit vector in the i-th coordinate direction with $i = 1, 2$. We will use $\alpha = (x, y) \in \mathbf{Z}_0^2$ to denote a typical state vector for Ξ.

2.2.2 Control Policies

The control that can be apply over the queueing models arrival stream and service configurations is an important part of our stability results. We assume three different levels of control and give our stability results depending on these. As we have seen

in Section 1.3 not all sensibly or naturally seeming control policies lead to stability for all types of queueing models.

We define the policies by which the management regimes η at each state α are selected. A policy for controlling this discrete event system is a sequence $\Pi = \{\pi_n : n \geq 0\}$ of probabilities π_n from $\mathcal{H}_n$, the process history at time n, to $\mathcal{R}$, the set of regimes. This means for any history $\alpha_0, \eta_0, \ldots, \alpha_{n-1}, \eta_{n-1}, \alpha_n$ the next action is selected according to the distribution $\pi_n(\alpha_0, \eta_0, \ldots, \alpha_n, \cdot)$. This definition includes non-stationary, non-Markov randomised policies though they offer no performance benefits when applied to stationary Markov processes, see Blackwell [2]. The results for these very general controls are given in Theorems 2.3.1 and 2.3.2.

Let $\xi_i(n)$ denote the length of queue i at time n and $\xi(n) = (\xi_1(n), \xi_2(n))$. A policy Π along with an initial distribution for the queues determines a stochastic process $(\Xi, \Pi) = \{(\xi(n), \eta_n) : n \geq 0\}$ which will only be Markov when $\pi_n(\alpha_0, \eta_0, \ldots, \alpha_n, \cdot)$ is Markov and depends only on α_n and not on the whole history $\mathcal{H}_n$.

We are interested in stationary Markovian control policies which choose the same distribution over large *blocks* of the state space, this is interesting because it leads to a finite number of blocks with a homogeneous jump distribution in each block. We define these blocks formally in Section 2.2.4 for now denote them by $\mathcal{B}$.

We consider two of these *block policies*. Under this type of policy the state space $\mathbf{Z}_0^2$ is partitioned into a small number of disjoint blocks and we run the system according to a distribution $\pi(\alpha)$ in each block.

The first type of block policy is what we call a *block randomised policy*, denoted by Π^r, where the distribution $\pi^r(\alpha)$ is the same at every state α in the block $\mathcal{B}$. We make a decision on which η to use according to a randomised rule which is the same for all states in a block.

The second type is what we call *block pure policy* denoted by Π^p. This is a deterministic rule and can be seen as a degenerate case of the randomised version because the distribution $\pi^p(\alpha)$ is constant for every $\alpha \in \mathcal{B}$, i.e. we run the system under the same regime η in the whole block.

Note: Block pure or randomised policies are stationary and Markov which means that the process (Ξ, Π^r) is Markov because we assume that the arrivals are Poisson

and the services times are exponential for now (see Chapter 3 for different arrival and service distributions where the queueing process is still Markov).

Given the control policy we introduce an assumption about the routing of arrivals and change of service regimes. Let $\mathcal{A}_1 \equiv \{(x,0) : x > 0\}$ and $\mathcal{A}_2 \equiv \{(0,y) : y > 0\}$ denote the boundary or axes of the state space $\mathbf{Z}_0^2$.

Assumption A3 (Boundary Reflexion Condition): If $\lambda_i > 0$ for $i = 1, 2$ or a routable arrival stream $\lambda > 0$ exists that can be routed to queue $i = 1, 2$ by changing the management regime η then the boundary reflexion condition applies, i.e. we can jump back into the interior of the state space by simply directing the arrivals to the empty queue.

The assumption below is needed when we would like to analyse two queue models where only one of the queues has an external Poisson arrival stream of rate λ and the other queue gets only feedback. This makes it impossible to send arrivals to the second queue by changing the routing scheme s, i.e. we cannot direct the process away from the boundary of the state space by directing the arrivals only. To compensate for this we introduce the *boundary sojourn condition* below.

Assumption A4 (Boundary Sojourn Condition): For each α on the boudray of the state space $\mathbf{Z}_0^2$ let $\tau \equiv \min\{n \geq 1 : \xi_0 = \alpha \in \mathcal{A}_1 \cup \mathcal{A}_2, \xi_n \in \mathbf{Z}_+^2\}$ denote the length of the boundary sojourn. We assume there exists a constant $v > 0$ such that for any $\alpha \in \mathcal{A}_1 \cup \mathcal{A}_2$ there is a policy Π_α such that

$$\mathbf{E}(\tau \mid \xi_0 = \alpha, \Pi_\alpha) < v \tag{2.2}$$

This condition means that we assume that we can change service configurations and regimes $\eta = (k, s)$, where the arrival routing s is fixed, so that customers are fed into empty queues through feedback in order to drive away from the boundary as quickly as possible.

2.2.3 Mean drifts

The results about stability are based on the mean drifts of the process. The basic idea for this is described in Section 1.2; alternatively to FMM [11] we will look at

the mean drifts under the different management regimes η, given a control policy Π.

The process (Ξ, Π) has bounded jumps, specifically $\pm e_i$ with $i = 1, 2$ and $\pm(e_2 - e_1)$ and so all moments of its jump distributions exist under any policy Π, but our results can be stated in terms of their first moments. For each regime η and process history $\mathcal{H}_n$ let

$$M^\eta = \mathbf{E}(\xi(n+1) - \xi(n) \mid \mathcal{H}_n, \pi_n = \eta) \tag{2.3}$$

denote the *mean drift* vector for any period when the policy selects regime η. We have, for $k = 1, \ldots, K$ at states $\alpha \in \mathbf{Z}_+^2 \equiv \{(x, y) \in \mathbf{Z}^2 : x > 0, y > 0\}$

$$\begin{aligned} M^\eta &= (M_1^\eta, M_2^\eta) \\ &= \begin{cases} \frac{1}{\rho}\left(\lambda + \lambda_1 + \mu_{k2}p_{21}^\eta - \mu_{k1}p_{10}^\eta,\ \lambda_2 + \mu_{k1}p_{12}^\eta - \mu_{k2}p_{20}^\eta\right), & \eta = (k, 1) \\ \frac{1}{\rho}\left(\lambda_1 + \mu_{k2}p_{21}^\eta - \mu_{k1}p_{10}^\eta,\ \lambda + \lambda_2 + \mu_{k1}p_{12}^\eta - \mu_{k2}p_{20}^\eta\right), & \eta = (k, 2) \end{cases} \end{aligned} \tag{2.4}$$

Recalling the efficient service assumption A1, when queue i is empty the policy selects a regime η from among those with $\mu_{ki} = 0$. This ensures that equation (2.4) is also correct for histories leading to states $\alpha \in \mathcal{A}_1$ and $\alpha \in \mathcal{A}_2$ for such service regimes k.

Now consider any policy Π allowing randomisation. The mean drift of our process Ξ under Π when the current state is $\alpha \in \mathbf{Z}_0^2$ is a 2-dimensional vector M^Π lying in the convex set

$$\mathcal{M} = \left\{ \sum_\eta p_\eta M^\eta : p_\eta \in [0, 1] \text{ and } \sum_\eta p_\eta = 1 \right\} \tag{2.5}$$

the *convex hull* of the regime mean drifts M^η. The extreme points of $\mathcal{M}$ are a subset of the regime mean drifts M^η. When three or more of the M^η are distinct, or not parallel, it may happen that the two-dimensional interior,

$$\mathrm{Int}_2(\mathcal{M}) \equiv \{z \in \mathcal{M} : B(z, \epsilon) \subset \mathcal{M} \text{ for some } \epsilon > 0\}, \tag{2.6}$$

is non-empty, here $z \in \mathbf{R}_+^2$, $B(z, \epsilon) = \{z' \in \mathbf{R}^2 : |z - z'| < \epsilon\}$. Using randomised policies (stationary Markovian or not) implies that we can *create* a vector M^Π, so that there exists a Π for which for example $M_i^\Pi < 0$ for $i = 1,2$. When using the deterministic *block pure* policy only *pure* regimes M^η can be used to run the

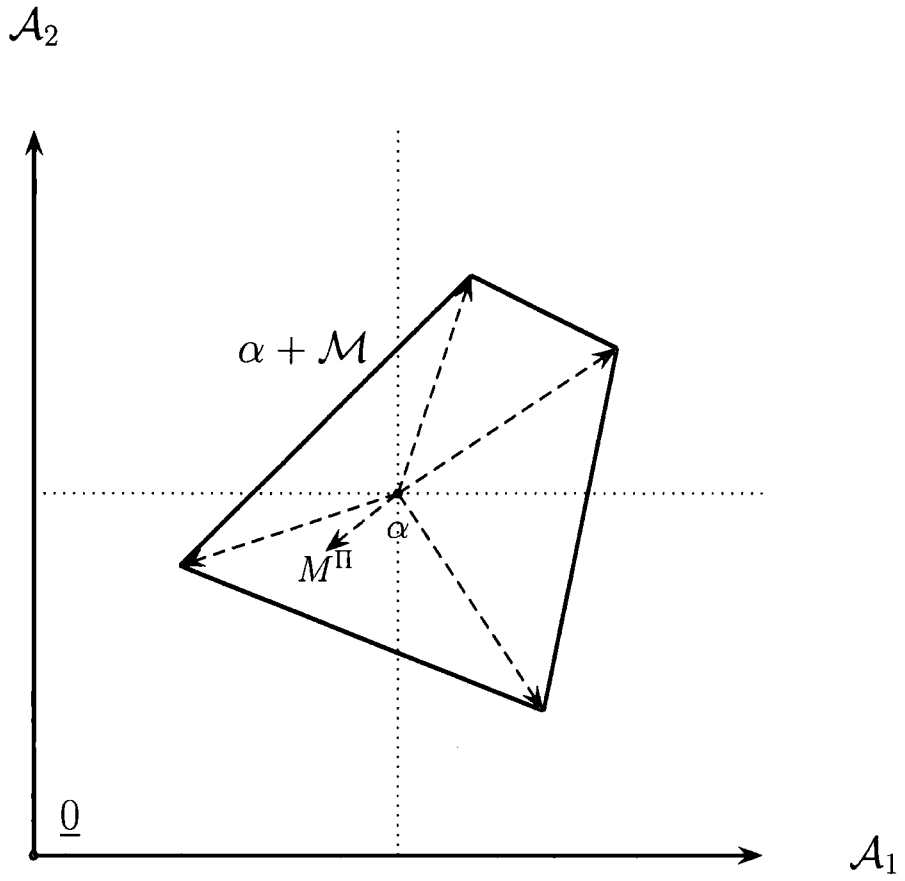


Figure 2.3: $\alpha + \mathcal{M}$ with non-empty 2 dimensional interior and an example for M^{Π}.

queueing system and we might not always have η such that $M_i^{\eta} < 0$ for $i = 1,2$. Figure 2.3 depicts a non-empty interior $\text{Int}_2(\mathcal{M}) \neq \emptyset$ with an example vector M^{Π}.

2.2.4 Cone shaped blocks in $\mathbf{Z}_0^2$

As stated earlier we are interested in policies which apply the same rule over a block of the state space $\mathbf{Z}_0^2$, specifically we will look at cone shaped blocks. For these we need some notation.

Summarising some of the earlier notation we have:

* the state space is $\mathbf{Z}_0^2$
* the interior of the state space is denoted by $\mathbf{Z}_+^2$, some times used as a block together with,

* the positive parts of the axes, $\mathcal{A}_1 \equiv \{(x,0) : x > 0\}$ and $\mathcal{A}_1 \equiv \{(0,y) : y > 0\}$ which are considered blocks.

Note that there is also the origin of the state space $\underline{0}$ which is not included in the three blocks given above, this is because Lyapunov function methods allow us to not really care about what happens to the process close to the origin.

$\mathbf{Z}_0^2$ is divided into cones in the following way. Let e_i denote the unit vector in the axis $\mathcal{A}_i$ direction and for non-zero $z = (z_1, z_2) \in \mathbf{R}_+^2$ let $|z|$ denote the length of z and $\arg_u(z)$ the argument relative to non-zero vector $u \in \mathbf{R}^2$. $\arg_u(z)$ gives the angle anticlockwise from u to z. For any non-zero u, $v \in \mathbf{R}^2$ let $\ell(u) = \{z \in \mathbf{R}^2 : z = tu, t > 0\}$ denote the half-line in the direction u and

$$\mathcal{C}(u,v) \equiv \left\{z \in \mathbf{R}^2 : |z| > 0,\ 0 < \arg_u(z) < \arg_u(v)\right\} \tag{2.7}$$

the cone swept anticlockwise from direction u to direction v. The closure of such a cone will be denoted $\bar{\mathcal{C}}(u,v)$.

An example of $\mathbf{Z}_0^2$ divided into five blocks given by the two axes $\mathcal{A}_i$ and three cones is shown in Figure 2.4. Here $\mathcal{C}_1 = \mathcal{C}(e_1, d_1) = \{z \in \mathbf{R}^2 : |z| > 0,\ 0 < \arg_1(e_1) < \arg_1(d_1)\}$, $\mathcal{C}_2 = \bar{\mathcal{C}}(d_1, d_2)$ and $\mathcal{C}_3 = \mathcal{C}(d_2, e_2)$.

Knowing the angle between the axes $\mathcal{A}_i$ and the mean drift vectors M^η is important when assessing the behaviour of the system. Therefore it will be convenient to define two special versions of the argument, one relative to each axis $\mathcal{A}_i$. Let $R : \mathbf{R}^2 \to \mathbf{R}^2$ be reflection in the line $z_1 = z_2$ i.e. $R(z_1, z_2) = (z_2, z_1)$ and define

$$\arg_1(z) = \arg_{e_1}(z)\ , \qquad \arg_2(z) = \arg_1\big(R(z)\big) \tag{2.8}$$

so $\arg_2(z)$ is the angle measured clockwise from $\mathcal{A}_2$ to z. These are used to define the two angles of the reflexion vectors M' and M'' which are the two mean drift vectors, under regimes η, such that $\mu_{k1} = 0$ and $\mu_{k2} = 0$ respectively. If there is more than one regime η which we can use on an axis under the efficiency assumption A1 the η which maximises the angles $\psi_1 = \arg_1(M')$ and $\psi_2 = \arg_2(M'')$ relative to the two axes is chosen, the angles are also depicted in Figure 2.4.

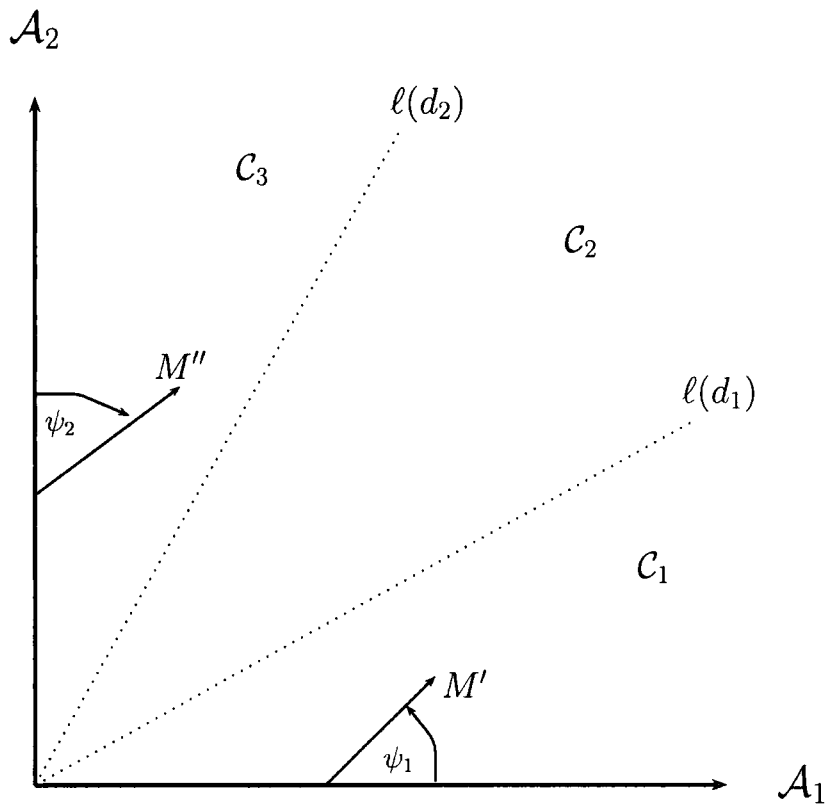


Figure 2.4: An example of a finite number of blocks, with three cones and the two axis giving five blocks.

2.3 Classification and Results

The classification of our queueing model using the mean drifts M^η is the basis for the results. It is based on the position and shape of the convex hull $\alpha + \mathcal{M}$ with respect to the origin $\underline{0} = (0,0)$, and whether $\alpha + \mathcal{M}$ can be separated from the origin by a line (or hyperplane) through α or not. The classification leads directly to the Theorems about stability or ergodicity and transience, under different control policies Π given later on in this Section.

For any set of parameters λ_i, μ_{ki} and p_{ij}^η the mean drifts M^η and thus $\mathcal{M}$ of the process (Ξ, Π) fall into one the following four exclusive cases:

C1 $(0,0) = \underline{0} \notin \mathcal{M}$ and there exists a state $\alpha \in \mathbf{Z}_0^2$ and a hyperplane

$$L_v(\alpha) \equiv \{\beta \in \mathbf{R}^2 : v^T(\beta - \alpha) = 0\} \tag{2.9}$$

with normal vector v through α separating $\alpha + \mathcal{M}$ from the origin $\underline{0}$. If there exists one such $\alpha \in \mathbf{Z}_0^2$ then there is an infinite cone of such α.

C2 $\underline{0} \notin \mathcal{M}$ and there exists no $\alpha \in \mathbf{Z}_0^2$ for which there exists a hyperplane $L_v(\alpha)$ through α which separates $\alpha + \mathcal{M}$ from $\underline{0}$.

C3 $\mathrm{Int}_2(\mathcal{M})$ is non-empty, $\underline{0} \in \mathcal{M}$ and there exists no $\alpha \in \mathbf{Z}_0^2$, $v \in \mathbf{R}^2$ such that the line $L_v(\alpha)$ separates $\alpha + \mathrm{Int}_2(\mathcal{M})$ from the origin.

C4 $\underline{0}$ is a boundary point of $\mathcal{M}$ and either $\mathrm{Int}_2(\mathcal{M}) = \emptyset$ or the tangent line to $\alpha + \mathcal{M}$ through α separates the origin $\underline{0}$ from $\alpha + \mathrm{Int}_2(\mathcal{M})$ for each α in a cone within $\mathbf{Z}_0^2$.

See Figure 2.5 for examples of C1-C4 (for further reference the classification is also given in Appendix A).

The results are separated by the type of control policy Π that is applied. We start by giving sufficient conditions for instability or stability respectively of the system under fully randomised controls in cases C1 and C2 respectively. Next we show that in case C3 there is always a block pure policy that makes (Ξ, Π^p) ergodic and we also show that randomisation allows the use of fewer blocks. Finally we will consider the null recurrent case C4.

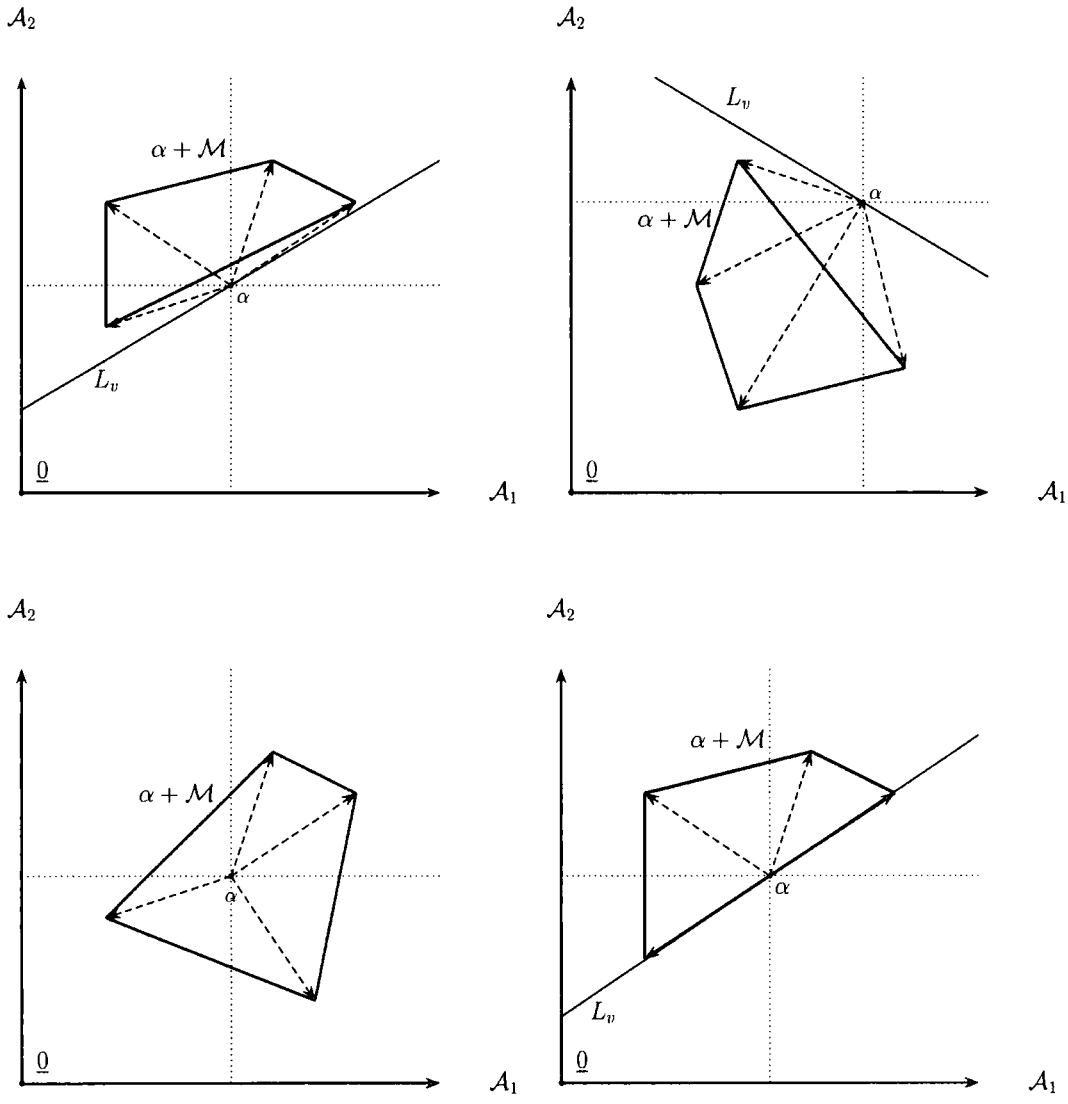


Figure 2.5: From top left: C1, C2, and below C3, C4.

We will explore lower levels of control relating to the results in Kurkova [22] separately in Section 2.4.

2.3.1 Fully randomised controls

The following two results apply when even the most general policy Π, non-stationary and non-Markov, is used to control the queueing system. The results imply that,

given our assumptions about the queueing process, in cases C1 and C2 the control policy used does not affect the stability or otherwise of the process.

Theorem 2.3.1 *If* $\underline{0} \notin \mathcal{M}$ *and there exists an* $\alpha \in \mathbf{Z}_0^2$ *and* $v \in \mathbf{R}^2$ *such that the line* $L_v(\alpha)$*, see (2.9), separates* $\alpha + \mathcal{M}$ *from the origin* $\underline{0}$ *then the process* (Ξ, Π) *is unstable, in the sense that the total number of queued jobs almost surely goes to* ∞ *linearly in time for any policy* Π.

The conditions of the theorem can be pictured in an alternative way. Specifically there exists a state $\alpha \in \mathbf{Z}_+^2$ such that the line segment from $\underline{0}$ to α does not intersect $\alpha + \mathcal{M}$. It follows that if there is any such pair α, v then there is an infinite cone of points α' such that $L_v(\alpha')$ separates $\underline{0}$ and $\alpha' + \mathcal{M}$.

Proof: Under the conditions of Theorem 2.3.1 there exists $\alpha \in \mathbf{Z}_+^2$ and $v = (v_1, v_2) \in \mathbf{R}^2 \setminus \bar{C}(-e_1, -e_2)$ (i.e. v has at least one positive component) such that $L_v(\alpha)$ separates $\alpha + \mathcal{M}$ from $\underline{0}$ and hence $v^{\mathsf{T}} M^\eta > 0$ for every regime η. As the number of regimes η is finite, we can choose $\varepsilon = \frac{1}{2} \min_\eta v^{\mathsf{T}} M^\eta > 0$. In order to apply Theorem 1.2.1 to show that the process is unstable we first need to define a sequence of non-negative random variables $\{S_n\}$. In this case $S_n = v_1 \xi_1(n) + v_2 \xi_2(n)$ for $n = 0, 1, \ldots$ satisfies

$$\mathbf{E}(S_{n+1} - S_n \mid \mathcal{H}_n, \pi_n = \eta) = v^{\mathsf{T}} M^\eta > \varepsilon$$

whatever policy Π is used. It then follows from part (ii) of Theorem 1.2.1 that

$$\mathbf{P}(S_n < \delta_1 n \mid \Pi) < Ce^{-\delta_2 n} \text{ for all } n \geq 0$$

so by Borel-Cantelli these events almost surely occur only finitely often. This means the number of jobs waiting in the queues goes to infinity linearly in time due to the Poisson arrival stream. In addition we have $P(\tau = \infty) > 0$ so with positive probability the process makes no visits to $\{\alpha \in \mathbf{Z}_+^2 : v^{\mathsf{T}} \alpha < D\}$ which for large D contains the region of $\mathbf{Z}_+^2$ around the origin $\underline{0}$. □

Next we present the results for case C2, here the process is always stable no matter what policy Π is used.

Theorem 2.3.2 *If* $\underline{0} \notin \mathcal{M}$ *and there is no* $\alpha \in \mathbf{Z}_+^2$, $v \in \mathbf{R}^2$ *such that* $L_v(\alpha)$ *separates* $\alpha + \mathcal{M}$ *from* $\underline{0}$ *then* (Ξ, Π) *is stable, in the sense that the total number of queued jobs remains bounded in mean, under every policy* Π.

The alternative description of the conditions here is that for every $\alpha \in \mathbf{Z}_+^2$ the line segment joining $\underline{0}$ to α intersects $\alpha + \mathcal{M}$. From this it follows there is some $v \in \mathbf{R}_+^2$ such that $\underline{0}$ and $\alpha + \mathcal{M}$ are in the same half space created by $L_v(\alpha)$.

Proof: Under the conditions of Theorem 2.3.2 we can find a $v \in \mathbf{R}_+^2$ such that $v^\mathsf{T} M^\eta < 0$ for every regime η. As before we choose $S_n = v_1\xi_1(n) + v_2\xi_2(n)$ for $n = 0$, $1, \ldots$ which now satisfies

$$\mathbf{E}(S_{n+1} - S_n \mid \mathcal{H}_n, \pi_n = \eta) = v^\mathsf{T} M^\eta < -\varepsilon$$

for some $\varepsilon > 0$. Applying part (i) of Theorem 1.2.1 we see that $\mathbf{E}(\tau \mid \Pi, S_0 > D) \leq S_0/\varepsilon < \infty$. Thus from any finite state α the process reaches $\{\alpha \in \mathbf{Z}_+^2 : v^\mathsf{T}\alpha < D\}$ in finite time and the process must remain finite almost surely. □

Note: these two results have immediate extensions to models with $i \geq 3$ queues this is given in Chapter 4.

2.3.2 Block controls

In case C3 it does make a difference which policy is used for running the system. In fact we can show that block pure policies Π^p with at most a handful of blocks are adequate to ensure stability of the process. Under policies of this type the process (Ξ, Π^p) is Markov so we can now talk about ergodicity and transience.

Corollary 2.3.3 *Under block randomised or block pure policy the processes* (Ξ, Π^r) *and* (Ξ, Π^p) *are Markov, so that under the conditions of Theorem 2.3.2 these processes are ergodic if the first moment* $\mathbf{E}(\xi_n \mid \mathcal{H}_n m \pi_n)$ *is finite.*

The main result for two queues is given by the following

Theorem 2.3.4 *If* $\underline{0} \in Int_2(\mathcal{M})$ *then there is a block pure policy* Π^p *with at most five blocks such that the Markov chain* (Ξ, Π^p) *is ergodic.*

This implies that when $\underline{0}$ is inside $\mathcal{M}$ we can always find a policy which is stationary, Markov and uses only five blocks so that the process is ergodic. Theorems 2.3.2 and 2.3.4 imply the following result.

Corollary 2.3.5 *If $\underline{0}$ is a boundary point of $\mathcal{M}$, $Int_2(\mathcal{M})$ is non-empty and there exists no $\alpha \in \mathbf{Z}_0^2$, $v \in \mathbf{R}^2$ such that $L_v(\alpha)$ separates $\alpha + Int_2(\mathcal{M})$ from $\underline{0}$ then there is a policy Π^p with at most three blocks such that (Ξ, Π^p) is ergodic.*

In Theorem 2.3.4 the number of blocks required to achieve ergodicity can be reduced if block randomised policies Π^r are used. This is due to the fact that under a randomised policy and the conditions of case C3 we can choose a mean drift with $M_i^{\Pi^r} < 0$ for $i = 1, 2$.

Corollary 2.3.6 *If $\underline{0} \in Int_2(\mathcal{M})$ and a block randomised policy Π^r is used then at most four blocks are necessary to ensure that (Ξ, Π^r) is ergodic.*

Example 2.3.1 (Fixed servers and JSQ) Foley and MacDonald [13] carry out the large deviations analysis of a model with N queues which has fixed servers, no feedback and is strictly join the shortest queue (JSQ). For $N = 2$ queues their are three Poisson arrival streams, two dedicated ones with parameters λ_1 and λ_2 and a routable one with rate λ. The service times at queue 1 and 2 are exponentially distributed with parameters μ_1 and μ_2. We can see that the boundary reflexion condition (A3) applies. The stability criterion for $N = 2$ queues given in Foley and MacDonald [13] is that $\rho_{\max} \leq 1$ where

$$\rho_{\max} = \max\{\lambda_1/\mu_1,\ \lambda_2/\mu_2,\ (\lambda + \lambda_1 + \lambda_2)/(\mu_1 + \mu_2)\ \}.$$

For the policy which sends the routable stream λ to the shortest queue our model has four regimes depending upon where the routable traffic is sent and which queues have customers. We have four blocks the two axes $\mathcal{A}_i$ and two cones $\mathcal{C}_1 = \mathcal{C}(e_1, d) \cap \ell(d)$ and $\mathcal{C}_2 = \mathcal{C}(d, e_2)$ where the slope of $\ell(d)$ is $d' = 1$; join the shortest queue means that the two cones split $\mathbf{Z}_+^2$ in half and in this example we also have symmetry of the jump distributions. The four mean drift vectors are

$$M^1 = \tfrac{1}{\rho}(\lambda_1 - \mu_1, \lambda + \lambda_2 - \mu_2) \text{ and } M^2 = \tfrac{1}{\rho}(\lambda + \lambda_1 - \mu_1, \lambda_2 - \mu_2)$$

when both queues are non-empty (on cones $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively) and

$$M' = \tfrac{1}{\rho}(\lambda_1 - \mu_1, \lambda + \lambda_2) \text{ and } M'' = \tfrac{1}{\rho}(\lambda + \lambda_1, \lambda_2 - \mu_2)$$

on $\mathcal{A}_1$ and $\mathcal{A}_2$ where ρ is defined in (2.1). The condition $\rho_{\max} < 1$ guarantees one of the cases in Theorem 2.4.1(i) (see Section 2.4) holds so the system is ergodic. Similarly $\rho_{\max} > 1$ or $\rho_{\max} = 1$ leads to cases in Theorem 2.4.1(ii) or (iii) respectively. We see that our conditions are consistent with those of Foley and McDonald for $N = 2$.

Our results also apply where these regimes may be chosen by policies different to JSQ. As $\rho(M^1 - M^2) = (-\lambda, \lambda) \perp (1, 1)$ the line segment joining these two drift vectors has the form $z_1 + z_2 = (\lambda + \lambda_1 - \mu_1 + \lambda_2 - \mu_2)/\rho$ which can only intersect $\mathbf{R}^2_-$ when $\rho_{\max} < 1$. In this case Theorem 2.3.4 guarantees a routing scheme that makes the system stable.

Proofs: We will establish Theorem 2.3.4 by using linear and *smoothed* piecewise linear Lyapunov functions and appropriate waiting times N_i, $i = 1, 2$ when Ξ visits the axes $\mathcal{A}_i$. We set up some additional preliminary results before the proofs of our main results.

The lemma below about second vector fields states a special case we need of general results from Fayolle, Malyshev and Menshikov [11].

Let X be a Markov chain on $\mathbf{Z} \times \mathbf{Z}_0$ with typical state $\alpha = (x, y)$ and transition probabilities $P_{\alpha\beta} = p(\beta - \alpha)$ for α, $\beta \in \mathbf{Z} \times \mathbf{Z}_0$ with $y \geq 1$ while $P_{\alpha\beta} = q(\beta - \alpha)$ when $y = 0$ for some appropriate distributions p, q. Suppose there exists some $b > 0$ such that $p(\beta - \alpha) = 0$ and $q(\beta - \alpha) = 0$ whenever $\|\beta - \alpha\| > b$. Let $M^1 = E(X(n+1) - X(n) \mid X(n) = \alpha)$ for α with $y \geq 1$ and let $M' = E(X(n+1) - X(n) \mid X(n) = \alpha)$ when $y = 0$. Let $\varphi = 2\pi - \arg_1(M^1)$, $\psi = \arg_1(M')$ and let $\mathcal{A} = \{\alpha \in \mathbf{Z} \times \mathbf{Z}_0 : y = 0\}$.

Lemma 2.3.7 *For any given $w \in \mathbf{R}^2$ let $\varepsilon = |w^\top M^1|$. There exist constants $\gamma = \gamma(w)$ and $\delta \in (0, \varepsilon)$ such that*

(i) if $M_y^1 < 0$, $w_1 > 0$ *then* $\mathbf{E}(w^\mathsf{T}X(n+\gamma) - w^\mathsf{T}X(n) \mid X(n) \in \mathcal{A}) < -\gamma\delta$ $(> \gamma\delta)$ *according as to* $\varphi + \psi > \pi$ $(< \pi)$;

(ii) if $M_y^1 \geq 0$ *then* $E(w^\mathsf{T}X(n+\gamma) - w^\mathsf{T}X(n) \mid X(n) \in \mathcal{A}) < -\gamma\delta$ $(> \gamma\delta)$ *according as to* $w^\mathsf{T}M^1 < 0$ (> 0) .

Proof of Lemma 2.3.7: the proof is routine as the projection of X onto the y dimension is a one dimensional Markov chain which is ergodic in case (i) and transient or null recurrent in case (ii). In the language of FMM [11], in case (i) $\mathcal{A}$ is an ergodic face with the second vector field (here a scalar) ingoing when $\varphi + \psi > \pi$ and otherwise outgoing. In case (ii) face $\mathcal{A}$ is transient so the jump distribution there has no major influence on the long term behaviour of X. □

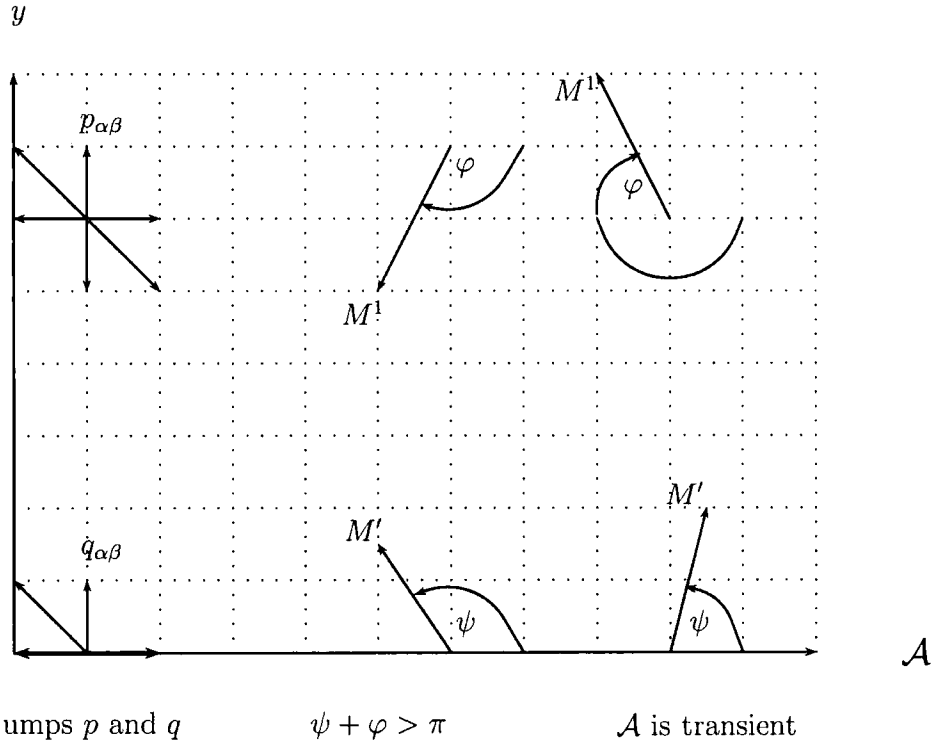


Figure 2.6: Jumps and possible mean drifts of the Markov chain X on $\mathbf{Z} \times \mathbf{Z}_0$.

We will use Lemma 3.3.4 from [11] to extend our semi-martingale results to the smoothed functions so briefly state it here for our process Ξ. Let f be a real function defined on $\mathbf{R}^2$, b a bound on the length of Ξ's jumps and Λ the space of linear functions on $\mathbf{R}^2$.

Lemma 2.3.8 (local linearity) *For $\alpha \in \mathbf{Z}_0^2$, if $f(\alpha + M^\eta) < f(\alpha) - 5\epsilon$ for some $\epsilon > 0$ and*

$$\inf_{\lambda \in \Lambda} \sup_{z \in B(\alpha,b)} |f(z) - \lambda(z)| < \epsilon \tag{2.10}$$

then $\mathbf{E}\big(f(\xi(n+1)) - f(\xi(n)) \mid \xi(n) = \alpha, \pi_n = \eta\big) < -\epsilon$.

Smoothing The smoothing we use is elementary. For any $d \in \mathbf{R}_+^2$, $u, v \in \mathbf{R}^2$ such that $0 < u^\mathsf{T} d = v^\mathsf{T} d$ and $u_2/u_1 < d_2/d_1 < v_2/v_1$, the continuous piecewise-linear function

$$f(z) = \begin{cases} u^\mathsf{T} z\ , & z \in \mathcal{C}(e_1, d) \\ v^\mathsf{T} z\ , & z \in \mathcal{C}(d, e_2) \end{cases}$$

defined on cones $\mathcal{C}(e_1, d)$ and $\mathcal{C}(d, e_2)$ in $\mathbf{R}_+^2$ with common boundary line $\ell = \{z \in \mathbf{R}_+^2 : z = td, t > 0\}$ can be modified outside a ball around the origin to give a smoothed version $\tilde{f}$. To do this we redefine f on a strip parallel to and containing $\ell(d)$ so that its contours there are circular arcs as follows (see also Figure 2.7). For

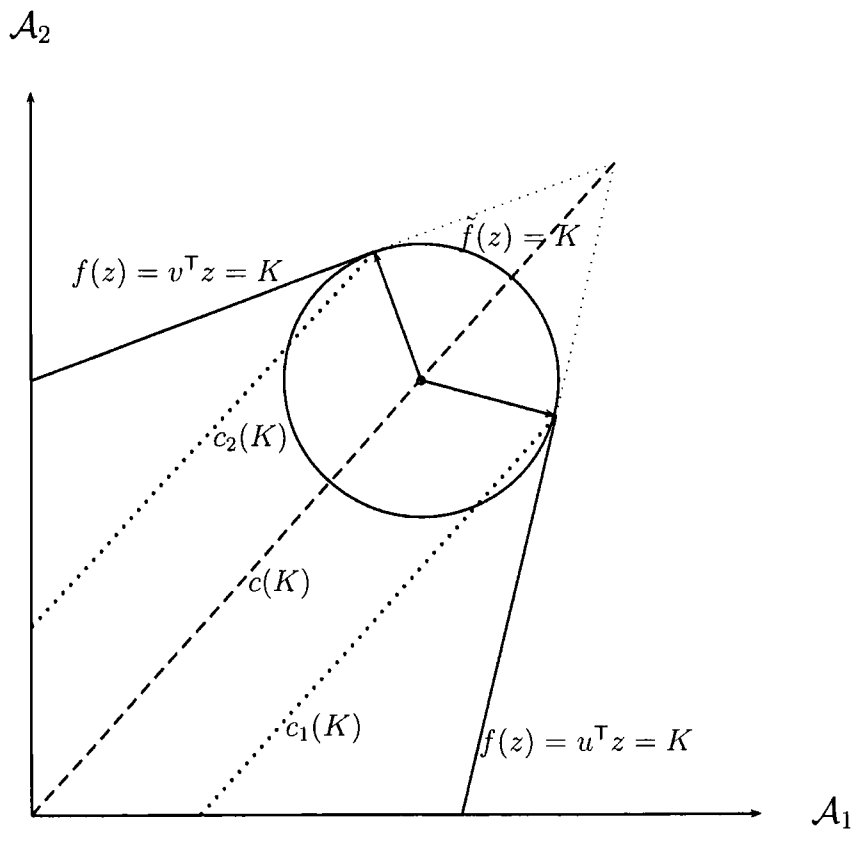


Figure 2.7: An example of a smoothed piecewise linear function.

any given $r > 0$ and for all $K \geq r\max(|d|, |u|, |v|)$ join the two linear parts of the contour of f with value K by a circular arc of radius r centred on the point $c(K) = \{z \in \mathbf{R}^2_+ : v^\mathsf{T} z = K - r|v|\} \cap \{z \in \mathbf{R}^2_+ : u^\mathsf{T} z = K - r|u|\} \in \mathbf{R}^2_+$ and extending from $c_1(K) = c(K) + ru/|u|$ to $c_2(K) = c(K) + rv/|v|$ and set $\tilde{f}(z) = K$ on this arc. The centres $c(K)$ lie on a straight line parallel to d as do the $c_1(K)$ and $c_2(K)$ and we define $\tilde{f}(z) = f(z)$ outside the strip bounded by these lines. For any z in this strip we note that $\tilde{f}(z + \gamma d) = \tilde{f}(z) + \gamma u^\mathsf{T} d$ for scalar γ. By choosing r sufficiently large we can (i) ensure that the distance from $\ell(d)$ to the edge of the strip is further than the longest possible jump; (ii) use the tangent to the contour at point α as our linear function λ in which case we may take $\epsilon \approx b^2(u^\mathsf{T} d)/2r$ in condition (2.10), where b is an upper bound on jump length.

For some additional but useful notation denote by $\mathcal{M}\{1, \ldots k\} = \text{conv}\{v_1, \ldots, v_k\}$ the convex hull of the vectors v_i, we will also use this notation to describe the convex hull of for example two mean drifts M' and M^2 by $\mathcal{M}\{', 2\}$ and so on.

Dichotomy The final piece of our analysis is a dichotomy which we use to split the proofs of the remaining theorems into cases.

Lemma 2.3.9 *Let v_1, v_2, ..., v_k be distinct vectors in $\mathbf{R}^2$ and $\mathcal{M}\{1, \ldots k\} = conv\{v_1, \ldots, v_k\}$, the convex hull of the v_i. Suppose that $B(\underline{0}, \epsilon) \subset \mathcal{M}\{1, \ldots k\}$ (so the origin is inside the convex hull, not on the boundary) for some $\epsilon > 0$ and there exists $u \in \mathbf{R}^2_+$ such that $u^\mathsf{T} v_i > 0$ for $i = 1, 2$. For definiteness suppose that $0 < \arg_{v_1}(v_2) < \pi$ so that $\mathcal{C}(v_1, v_2) \cap \mathcal{C}(-e_1, -e_2) = \emptyset$ (so v_1, v_2 are in the upper half plane of $L_u(\underline{0})$ and v_1 is clockwise of v_2). Then one of the following holds:*

D1: $v_i \in \mathcal{C}(-v_1, -v_2)$ for some $i \leq k$

D2: D1 does not hold but there exist $w \in \mathbf{R}^2$, $v_m \in \mathcal{C}(v_2, -v_1), v_n \in \mathcal{C}(-v_2, v_1)$ such that $w^\mathsf{T} v_i > 0$, $i = 1, 2$ $w^\mathsf{T} v_m < 0$, $w^\mathsf{T} v_n < 0$

Proof of Lemma 2.3.9: Suppose $D1$ does not hold. If there is no $v_i \in \mathcal{C}(v_2, -v_1)$ then, as $v_i \notin \mathcal{C}(-v_1, -v_2)$ for all i it follows that there are no v_i in the half plane $\mathcal{C}(v_2, -v_2)$ which contradicts $B(\underline{0}, \epsilon) \subset \mathcal{M}\{1, \ldots k\}$. It follows there exist $v_i \in \mathcal{C}(v_2, -v_1)$ and by a similar argument some other $v_i \in \mathcal{C}(-v_2, v_1)$.

Let v_m denote the $v_i \in \mathcal{C}(v_2, -v_1)$ with maximal value of $\arg_{v_2}(v_i)$ and v_n the $v_i \in \mathcal{C}(-v_2, v_1)$ with minimal value of $\arg_{-v_2}(v_i)$. Now consider the unit vector u_m perpendicular to v_m such that $u_m^{\mathsf{T}} v_i > 0$ for $i = 1, 2$ which exists as $\mathcal{C}(v_1, v_2) \subset \mathcal{C}(-v_m, v_m)$. It is necessary that $u_m^{\mathsf{T}} v_n < 0$ for otherwise there are no $v_i \in \mathcal{C}(v_m, -v_m)$. Replicating this argument there exists u_n with $u_n^{\mathsf{T}} v_i > 0$ for $i = 1, 2$ and $u_n^{\mathsf{T}} v_m < 0$ and we complete the proof by setting $w = u_m + u_n$. □

Proof of Theorem 2.3.4 We wish to define a pure block policy Π^p to ensure the ergodicity of our Markov queueing process (Ξ, Π^p). We start by selecting regimes η' and η'' on blocks $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. Their mean drifts M' and M'' make angles $\psi_1 = \arg_1(M')$ and $\psi_2 = \arg_2(M'')$ with their axial directions. If there is a choice of regimes in either of these blocks we choose the regimes that make ψ_1 and ψ_2 as large as possible. The result depends on the sum of the angles, the smaller the sum of the angles the more work is need to show the positive recurrence on the Markov chain. Below we state how to split the state space into blocks and which regimes to use in each of these blocks depending on the angles and the dichotomy stated above.

If $\psi_1 + \psi_2 > 3\pi/2$ ergodicity can be achieved only using regimes η' and η'' as follows. Choose $d \in \mathbf{R}_+^2$ and use regime η' on $\bar{\mathcal{C}}(e_1, d)$ and η'' on $\mathcal{C}(d, e_2) \cup \mathcal{A}_2$. As $\psi_1 + \psi_2 > 3\pi/2$ we can choose $w \in \mathbf{R}_+^2$ such that $-\varepsilon = \max\{w^{\mathsf{T}} M',\ w^{\mathsf{T}} M''\} < 0$. Now consider the process $S_n = w^{\mathsf{T}} \xi(n)$. At all states $\alpha \in \mathbf{Z}_+^2$ we have

$$\mathbf{E}(w^{\mathsf{T}} \xi(n+1) - w^{\mathsf{T}} \xi(n) \mid \Pi^p, \xi(n) = \alpha) \leq -\varepsilon$$

and by Theorem 1.2.1(i), with $N_{n+1} - N_n = 1$ for all n, it follows that $\tau = \min\{n \geq 1 : S_n \leq D\}$ satisfies $E(\tau) \leq S_0/\varepsilon$ from any initial value $S_0 > D$. As $w \in \mathbf{R}_+^2$, $\{\alpha \in \mathbf{Z}_0^2 : w^{\mathsf{T}} \alpha \leq D\}$ is a finite triangle around $\underline{0}$ and the ergodicity of (Ξ, Π^p) is assured.

When $\psi_1 + \psi_2 \leq 3\pi/2$ it is necessary to use further regimes to achieve ergodicity. We will denote these by η_i, $i = 1, 2$ or 3 as necessary and their corresponding drift vectors by M^i. We split the proof into three sub-cases according to the value of $\psi_1 + \psi_2$ and the cases of the dichotomy established in Lemma 2.3.9.

Suppose that $3\pi/2 \geq \psi_1 + \psi_2 > \pi/2$ and apply Lemma 2.3.9 with $v_2 = M'$ and $v_1 = M''$. If D1 holds then there exists η^1 such that $\psi_i + \varphi_i > \pi$, where $\varphi_i = 2\pi - \arg_i(M^1)$, for $i = 1, 2$ (Case I); otherwise D2 applies and there exist η_1, η_2 and $w \in \mathbf{R}^2$ such that $M^1 \in \mathcal{C}(M', -M'')$, $M^2 \in \mathcal{C}(-M', M'')$, $w^{\mathsf{T}}M^i < 0$ for $i = 1, 2$ and also $w^{\mathsf{T}}M' > 0$ and $w^{\mathsf{T}}M'' > 0$ (Case II).

This leaves the possibility that $\psi_1 + \psi_2 \leq \pi/2$ in which case $\psi_i + \varphi_i \leq \pi$ (again $\varphi_i = 2\pi - \arg_i(M^\eta)$) for at least one of $i = 1, 2$ for every regime η. This time employ Lemma 2.3.9 with $v_1 = M'$ and $v_2 = M''$. If D1 applies then there exists η_1 such that $\psi_1 + \varphi_2 > \pi$ and $\psi_2 + \varphi_1 > \pi$ (Case III); otherwise D2 applies and we are again in Case II (with M' and M'' swapped). It remains to describe the pure block policies required and construct the super-martingales $\{S_n\}$.

Case I: Our policy Π^p uses three blocks, $\mathcal{A}_1$ where regime η' is used, $\mathcal{A}_2$ with regime η'' and $\mathbf{Z}_0^2$ where we use regime η^1 such that $\psi_i + \varphi_i > \pi$ for $i = 1, 2$. Pick $\varepsilon > 0$. As $M^1 \notin \mathbf{R}_+^2$ it has at least one negative component and further we can choose a $w \in \mathbf{R}_+^2$ such that

$$w^{\mathsf{T}}M^1 = \mathbf{E}(w^{\mathsf{T}}\xi(t+1) - w^{\mathsf{T}}\xi(t) \mid \Pi^p,\ \xi(t) \in \mathbf{Z}_+^2) < -\varepsilon$$

Applying Lemma 2.3.7 to (Ξ, Π^p) with each $\mathcal{A}_i$ as $\mathcal{A}$ (using part (i) if $M_i^1 < 0$ and (ii) if $M_i^1 \geq 0$) we see there are constants γ_1, γ_2 and $\delta \in (0, \varepsilon)$ such that

$$\mathbf{E}(w^{\mathsf{T}}\xi(t+\gamma_i) - w^{\mathsf{T}}\xi(t) \mid \Pi^p,\ \xi(t) = \alpha \in \mathcal{A}_i) < -\gamma_i\delta\ , \quad i = 1,\ 2$$

for α outside some finite ball where the other axis can be reached in γ_i steps. Choose $D > 0$ large enough that $\{w^{\mathsf{T}}\alpha \leq D\}$ contains such a ball. Next use the γ_i to define a sequence of random times N_n by $N_0 = 0$ and for $n = 0, 1, \ldots$

$$N_{n+1} - N_n = \begin{cases} 1, & \xi(N_n) \in \mathbf{Z}_+^2 \\ \gamma_1, & \xi(N_n) \in \mathcal{A}_1 \\ \gamma_2, & \xi(N_n) \in \mathcal{A}_2 \end{cases}$$

Now set $S_n = w^{\mathsf{T}}\xi(n)$ for $n \geq 0$, $Y_n = S_{N_n}$ and define hitting times τ, σ as for Theorem 1.2.1. We have shown that

$$\mathbf{E}\big(Y_{(n+1)\wedge\sigma} - Y_{n\wedge\sigma} \mid \Pi^p, Y_{n\wedge\sigma}\big) < -\delta\mathbf{E}\big(N_{(n+1)\wedge\sigma} - N_{n\wedge\sigma} \mid Y_{n\wedge\sigma}\big)$$

so by Theorem 1.2.1(i), $E(\tau) \leq S_0/\delta$ from any initial value $S_0 > D$. As $w \in \mathbf{R}^2_+$, the ergodicity of (Ξ, Π^p) is assured as before.

Case II: now policy Π^p needs four blocks, the $\mathcal{A}_i$ as before and now, for any $d \in \mathbf{R}^2_+$, the cones $\mathcal{C}(e_1, d) \cup \ell(d)$ on which regime η^1 is used and $\mathcal{C}(d, e_2)$ where η^2 is used. We know that $M^1 \in \mathcal{C}(-M', M'')$, $M^2 \in \mathcal{C}(M', -M'')$ and there exists $w \in \mathbf{R}^2$ such that $w^\mathsf{T} M^i < 0$ for $i = 1, 2$. Hence there exists $\varepsilon > 0$ such that

$$\mathbf{E}\big(w^\mathsf{T}\xi(t+1) - w^\mathsf{T}\xi(t) \mid \Pi^p,\ \xi(t) = \alpha\big) \leq -\varepsilon\ , \quad \alpha \in \mathbf{Z}^2_+$$

Next consider axis $\mathcal{A}_1$ with reflexion vector M' in relation to M^2, since $M^2 \in \mathcal{C}(M', -M'')$ we will check whether the axis is a transient or an ergodic face according to Lemma 2.3.7. If $M^2_y < 0$ we know that $\psi_1 + (2\pi - \arg_1(M^2)) > \pi$ and we can choose $w_1 > 0$. Next we apply Lemma 2.3.7, part (i) or (ii) depending upon the sign of M^2_y, to obtain constants γ_1 and $\delta \in (0, \varepsilon)$ such that

$$\mathbf{E}\big(w^\mathsf{T}\xi(t+\gamma_1) - w^\mathsf{T}\xi(t) \mid \Pi^p,\ \xi(t) = \alpha \in \mathcal{A}_1\big) \leq -\gamma_1\delta$$

as long as α is outside a finite ball around $\underline{0}$ where Ξ can reach $\ell(d)$ in γ_1 steps from α. Repeat this argument for block $\mathcal{A}_2$. Finally define the times N_n as in Case I and proceeding exactly as before, the ergodicity of (Ξ, Π^p) is established in Case II.

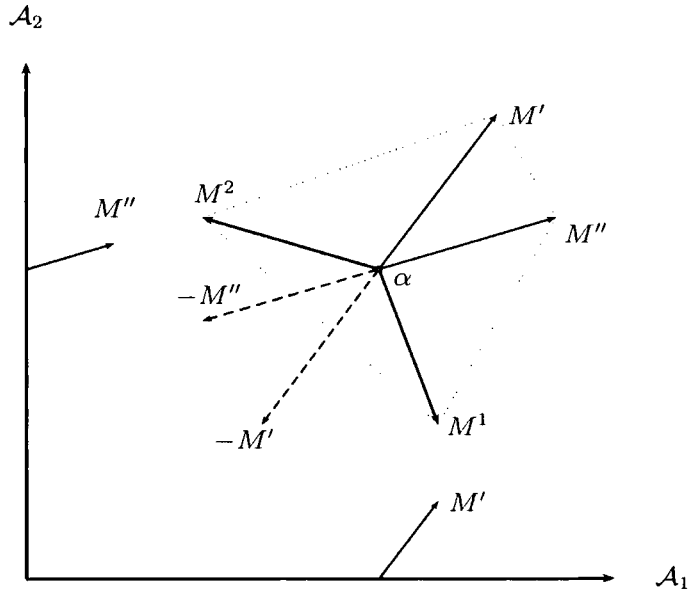


Figure 2.8: Possible mean drift vectors for Case II.

Case III: this time Π^p has five blocks and we use a smoothed piecewise linear Lyapunov function. For any d_1, $d_2 \in \mathbf{R}^2_+$ with d_2 strictly anticlockwise of d_1 define blocks $\mathcal{A}_1$, $\mathcal{C}_1 = \mathcal{C}(e_1, d_1) \cup \ell(d_1)$, $\mathcal{C}_2 = \mathcal{C}(d_1, d_2) \cup \ell(d_2)$, $\mathcal{C}_3 = \mathcal{C}(d_2, e_2)$ and $\mathcal{A}_2$. Π^p uses regimes η' on blocks $\mathcal{A}_1$ and $\mathcal{C}_3$ and η'' on $\mathcal{A}_2$ and $\mathcal{C}_1$. In addition we know there is a regime η^1 such that $\psi_1 + \varphi_2 > \pi$ and $\psi_2 + \varphi_1 > \pi$ where $\varphi_i = 2\pi - \arg_i(M^1)$.

Choose $\varepsilon > 0$. As $M^1 \in \mathcal{C}(-M', -M'')$ there exist v, $w \in \mathbf{R}^2$ such that $v_1 > 0$, $v^\mathsf{T} M'' < -\varepsilon$, $v^\mathsf{T} M^1 < -\varepsilon$ and $w_2 > 0$, $w^\mathsf{T} M' < -\varepsilon$, $w^\mathsf{T} M^1 < -\varepsilon$. To define our Lyapunov function let $d = -M^1$ and use it to define cones $\mathcal{D}^- = \mathcal{A}_1 \cup \mathcal{C}(e_1, d) \cup \ell(d)$, $\mathcal{D}^+ = \mathcal{C}(d, e_2) \cup \mathcal{A}_2$. Now let

$$f(z) = \begin{cases} v^\mathsf{T} z, & z \in \mathcal{D}^- \\ w^\mathsf{T} z, & z \in \mathcal{D}^+ \end{cases}$$

scaling v say, so that $v^\mathsf{T} d = w^\mathsf{T} d$ to make f continuous (we know both $v^\mathsf{T} d > 0$ and $w^\mathsf{T} d > 0$ so this does not change this sign of v_1).

By construction we know that for $\alpha \in \mathbf{Z}^2_+$ but not too near $\ell(d)$ we have

$$\mathbf{E}\big(f(\xi(t+1)) - f(\xi(t)) \mid \Pi^p,\ \xi(t) = \alpha\big) < -\varepsilon$$

but we do not know this expectation if a single jump from $\alpha \in \mathcal{D}^-$ can reach $\beta \in \mathcal{D}^+$ or vice versa. To deal with this we use the smoothing described earlier in this section

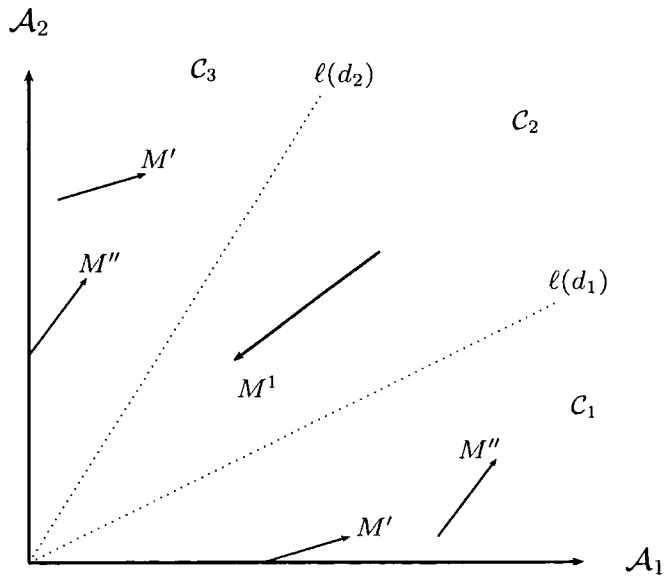


Figure 2.9: Regimes and mean drifts for Case III.

to produce a function $\tilde{f}$ which coincides with f except on a corridor containing $\ell(d)$ where its level curves are circular arcs of sufficiently large radius r.

Because of the choice of $d = -M^1$ we calculate that $\tilde{f}(\alpha+M^1) = \tilde{f}(\alpha)+w^{\mathsf{T}}M^1 < \tilde{f}(\alpha) - \varepsilon$ for every $\alpha \in \mathbf{Z}_+^2$ so by Lemma 2.3.8

$$\mathbf{E}\big(\tilde{f}(\xi(t+1)) - \tilde{f}(\xi(t)) \mid \Pi^p,\ \xi(t) \in \mathbf{Z}_+^2\big) < -\varepsilon/5.$$

Finally as $M_2'' > 0$ and $w^{\mathsf{T}}M < -\varepsilon$ it follows from Lemma 2.3.7(ii) that for $\alpha \in \mathcal{A}_1$ there exist constants γ_1 and $\delta \in (0, \varepsilon)$ such that

$$\mathbf{E}\big(w^{\mathsf{T}}\xi(t+\gamma_1) - w^{\mathsf{T}}\xi(t) \mid \Pi^p,\ \xi(t) = \alpha \in \mathcal{A}_1\big) \leq -\gamma_1\delta$$

as long as α is outside a finite ball around $\underline{0}$ where Ξ can reach $\ell(d_1)$ in γ_1 steps from α. Repeat this argument for block $\mathcal{A}_2$. Finally define the times N_n as in Case I and proceeding exactly as before, the ergodicity of (Ξ, Π^p) is established in all cases. This completes the proof of Theorem 2.3.4. □

Proof of Corollary 2.3.5 In this case $\psi_1 + \psi_2 \geq 3\pi/2$ so at the worst we can employ the argument of Case I of the proof of Theorem 2.3.4. □

Proof of Corollary 2.3.6 If $\psi_1 + \psi_2 > \pi/2$ then we use an appropriate mixture of regimes to obtain a drift vector $M^\Pi \in \mathcal{C}(-M'', -M')$ and we can use the argument of Case I of Theorem 2.3.4. If on the other hand $\psi_1 + \psi_2 \leq \pi/2$ then we use a policy Π that creates the conditions of Case II of Theorem 2.3.4. □

2.3.3 Null-recurrence of the Markov chain (Ξ, Π)

In order to complete the stability classification of the queue length process (Ξ, Π) we would like to evaluate the case C4. Under the conditions of case C4 the queueing system can be controlled so that (Ξ, Π) is null-recurrent (i.e. it cannot be positive recurrent and there are many more ways to control the system so that it is transient). We can formulate the following

Theorem 2.3.10 *If $\underline{0}$ is a boundary point of $\mathcal{M}$ and either $Int_2(\mathcal{M}) = \emptyset$ or the tangent line $L_v(\alpha)$ to $\alpha + \mathcal{M}$ through α separates the origin $\underline{0}$ from $\alpha + Int_2(\mathcal{M})$ for*

each α in $\mathbf{Z}_0^2$ then the Markov chain (Ξ, Π) can, "at best", be controlled so that it is null-recurrent.

Note, "at best" in the sense that any control that does not make the Markov chain null-recurrent will make it transient.

Proof: In order to establish the proof of Theorem 2.3.10 we first give a slightly modified version of Theorem 2.2.8, p. 31 in Fayolle, Malyshev, Menshikov [11]

Theorem 2.3.11 *For an irreducible Markov chain Ξ on $\mathbf{Z}_0^2$ to be null-recurrent it suffices that there exist two functions $f, \phi : \mathbf{Z}_0^2 \to \mathbf{R}_+$, $\alpha \in \mathbf{Z}_+^2$ and a finite set $\mathcal{D} \subset \mathbf{Z}_0^2$ so that the following conditions hold:*

(1) $f(\alpha) \geq 0$, $\phi(\alpha) \geq 0$ *for all* $\alpha \in \mathbf{Z}_+^2$.

(2) For some $\gamma > 0$ *and* $0 < \beta \leq 2$ *we have* $f(\alpha) \leq \gamma[\phi(\alpha)]^\beta$, *for all* $\alpha \in \mathbf{Z}_+^2$.

(3) $\lim_{x_i \to \infty} \phi(x_i) = \infty$ *for* $i = 1$, *2 and* $\sup_{\alpha \notin \mathcal{D}} f(\alpha) > \sup_{\alpha \in \mathcal{D}} f(\alpha)$.

(4) (a) $\mathbf{E}(f(\xi_{n+1}) - f(\xi_n) \mid \xi_n = \alpha) \geq 0$ *for all* $\alpha \notin \mathcal{D}$;

(b) $\mathbf{E}(\phi(\xi_{n+1}) - \phi(\xi_n) \mid \xi_n = \alpha) \leq 0$ *for all* $\alpha \notin \mathcal{D}$;

(c) $\sup_{\alpha \in \mathbf{Z}_+^2} \mathbf{E}(|\phi(\xi_{n+1}) - \phi(\xi_n)|^\beta \mid \xi_n = \alpha) = C < \infty$.

This means that we have to find two positive functions f and ϕ, so that $f(\alpha)$ is a submartingale while $\phi(\alpha)$ is a supermartingale for $\alpha \notin \mathcal{D}$. Or we can find one function for which the equality in (a) and (b) holds.

We will concentrate on one case of the null-recurrent cases only and first show how the process can be controlled so that it is null-recurrent and then give an example why it is transient if it is controlled any other way.

Again it is necessary to split the proof into cases depending on the reflexion vectors and their angles with the axes. Since we know that there exists a tangent line with $v \notin \mathbf{R}_+^2$ such that $v^\mathsf{T} M^\eta \geq 0$ for all η we can exclude the case when $\psi_1 + \psi_2 > 3\pi/2$. We distinguish the two cases for the angles, either $\psi_1 + \psi_2 = 3\pi/2$ or $\psi_1 + \psi_2 < 3\pi/2$, where the latter would need to be split into yet more cases.

If $\psi_1 + \psi_2 = 3\pi/2$ then we need only two blocks to run the queueing system under a block pure policy Π^p (a different policy would not change anything about

the stability). The two blocks are as follows, we choose $d \in \mathbf{R}_+^2$ and use regime η' on $\bar{\mathcal{C}}(e_1, d)$ and η'' on $\mathcal{C}(d, e_2) \cup \mathcal{A}_2$. Since $\psi_1 + \psi_2 = 3\pi/2$ we can choose $v \in \mathbf{R}_+^2$. Now consider the function $f(\xi) = \phi(\xi) = v^\mathsf{T}\xi(n)$. At all states $\alpha \in \mathbf{Z}_0^2$ we have

$$\mathbf{E}(v^\mathsf{T}\xi(n+1) - v^\mathsf{T}\xi(n) \mid \Pi^p, \xi(n) = \alpha) = 0.$$

We can see that $f(\alpha)$ and $\phi(\alpha)$ fulfil all conditions of Theorem 2.3.11 thus (Ξ, Π^p) is null recurrent.

Note that by choosing any other regime η on one of the two block for which $\psi_1 + \psi_2 \neq 3\pi/2$ under the conditions of Theorem 2.3.11 would make the process (Ξ, Π) transient. Using $f(\alpha)$ as above the transience can be established as in the proof of Theorem 2.3.1.

For $\psi_1 + \psi_2 < 3\pi/2$ we consider the following case. We can find a $v \in \mathbf{R}_+^2$ such that $v^\mathsf{T}M^\eta > 0$ for η', η''. In which case we need only four blocks to run the queueing system under a block pure policy Π^p. The blocks are $\mathcal{A}_1$ where we run η' and $\mathcal{A}_2$ with η'', we choose $d \in \mathbf{R}_+^2$ so that η^1 is run on $\mathcal{C}(e_1, d)$ and η^2 on $\mathcal{C}(d, e_2)$. The regimes η^1 and η^2 have mean drifts M^1 and M^2 which lie on the boundary of the convex hull $\partial\mathcal{M}$, i.e. they both lie on the tangent line with normal vector $v \in \mathbf{R}_+^2$ so that $v^\mathsf{T}M^1 = v^\mathsf{T}M^2 = 0$ and $M_x^1, M_y^2 < 0$. Consider first the linear function $f(\xi) = v^\mathsf{T}\xi(n)$. At all states $\alpha \in \mathbf{Z}_0^2$ we have

$$\mathbf{E}(v^\mathsf{T}\xi(n+1) - v^\mathsf{T}\xi(n) \mid \Pi^p, \xi(n) = \alpha) \geq 0$$

so $f(\alpha)$ is a positive submartingale, i.e. fulfils conditions (1) and (4)a of Theorem 2.3.10.

Construction the function ϕ to be a positive supermartingale is a little bit more involved and based on the smoothing and local linearity, see Lemma 2.3.8 in the proof of Theorem 2.3.4. We will briefly describe how such a function can be constructed.

The idea is to produce a function which is linear over most of $\mathbf{Z}_+^2$ but to *glue* a quadratic onto the end of the linear function, so that the reflexion vectors M' and M'' with $\psi_1 + \psi_2 < 3\pi/2$ point *inwards* from the locally linear function ϕ. For a suitably large radius r and along the line segment $l_K = \{\alpha \in \mathbf{R}_+^2 : w^\mathsf{T}\alpha = K\}$ select $p_1(K)$ so that l_K is tangential to a circle of radius r that *sits* on the axes $\mathcal{A}_1$. This

means the centre of the circle is given by (x_K, r) and $p_1(K) = (x_K, r) + r\frac{w}{|w|}$. Since we know that (x_K, r) sits on a line segment given by $\{\alpha \in \mathbf{R}_+^2 : w^\mathsf{T}\alpha = K - r|w|\}$ we get $x_K = \frac{1}{w_1}(K - r(|w| + w_2))$. The construction is also depicted in Figure 2.10. We can repeat the same idea for the other axes $\mathcal{A}_2$. The function ϕ is l_K between the point $p_1(K)$ and $p_2(K)$, and a part of a circle of radius r between $p_i(K)$ and the respective axes $\mathcal{A}_i$. We can observe that $f(\alpha) \leq \phi(\alpha)$ for all α outside some set

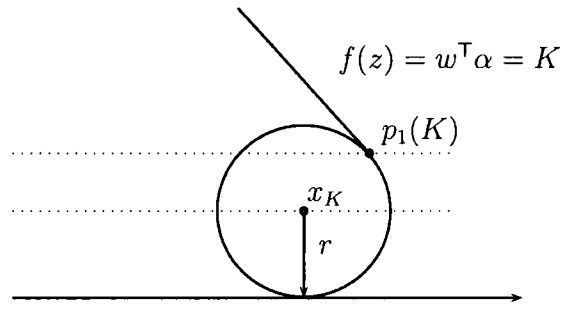


Figure 2.10: Constructing the locally linear function ϕ.

$D = \{\alpha \in \mathbf{Z}_0^2 : x + y \leq 4r\}$.

There are several ways to make the system unstable, for example one could reduce the number of blocks to three run M'' on $\mathcal{C}(d, e_2) \cup \mathcal{A}_2$.

There are the cases where we cannot find a $w \in \mathbf{R}_+^2 \cup \mathbf{R}_-^2$ such that $v^\mathsf{T}M^\eta > 0$ but we will not consider them here.

2.4 Low levels of control

In this section we will explore the ergodicity, transience and null-recurrence of a Markov chain with set blocks in $\mathbf{Z}_0^2$, so that the control that can be applied is limited. The results of Fayolle, Malyshev and Menshikov [11] can also be used to classify the process for any control policy that is block homogeneous for any small number of blocks. It soon becomes evident to anybody who attempts this that there are many ways for the process to remain stable and many more for it to be transient. To illustrate this we now spell out the possible behaviour of the queueing system with four blocks, specifically the axes $\mathcal{A}_1$, $\mathcal{A}_2$ and two cones, $\mathcal{C}_1 = \mathcal{C}(e_1, d) \cup \ell(d)$ and $\mathcal{C}_2 = \mathcal{C}(d, e_2)$ (see (2.7) for this notation), that partition $\mathbf{Z}_0^2$. The two cones are

not assumed to be symmetric i.e. the vector $d \in \mathbf{R}_+^2$ need not be parallel to (1, 1). Some of the cases given here with d equal to (1, 1) were analysed in Kurkova [22], see Example 2.4.1.

We assume that in each of the blocks $\mathcal{A}_i$ and $\mathcal{C}_i$, $i = 1, 2$ a single management regime is used (different blocks may have a common regime) with mean drift vectors M^1, M^2 in blocks $\mathcal{C}_1$, $\mathcal{C}_2$ respectively and M', M'' in blocks $\mathcal{A}_1$, $\mathcal{A}_2$ respectively.

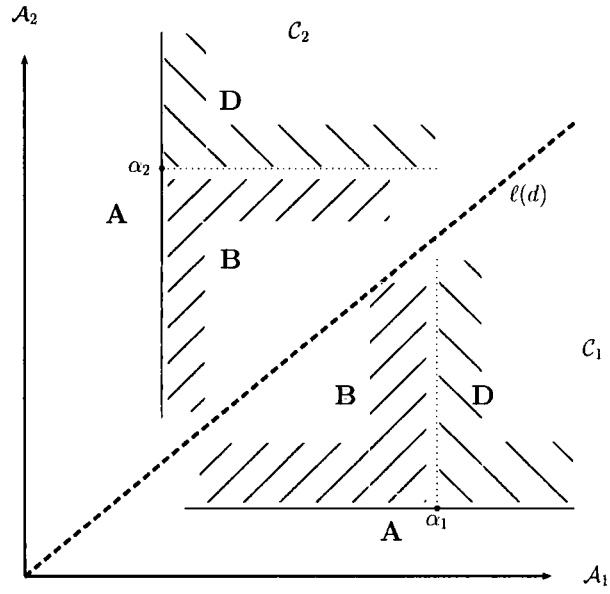


Figure 2.11: Graphical explanation of the labels.

We first label the M^i according to the angles φ_i they make relative to the axes $\mathcal{A}_i$, $i = 1, 2$. For each M^i angle $\varphi_i = 0$ is in the direction of $\mathcal{A}_i$ and φ_1 increases clockwise while φ_2 increases anticlockwise i.e. $\varphi_i = 2\pi - \arg_i(M^i)$. We label the directions of the M^i as case **A** when $0 < \varphi_i < \pi$, **B** when $\pi \leq \varphi_i \leq \frac{3\pi}{2}$ and **D** when $\frac{3\pi}{2} < \varphi_i \leq 2\pi$. The various cases of this model are labelled with *label of* M^1*/label of* M^2 so a label **B/A** means M^1 has a positive y and a negative x component and M^2 has x component negative with y of either sign. Figure 2.11 illustrates this labelling scheme for the directions of the M^i from origins α_i.

From the results FMM [11] on the random walk in the positive quadrant, we have (as also given in Lemma 2.3.7):

(i) if a drift M^i has an **A** label then axis $\mathcal{A}_i$ is an *ergodic* face;

(ii) if (i) is true the face $\mathcal{A}_i$ will be *outgoing*, *ingoing* or *neutral* according to the sign of the *second vector field* (which is scalar in this case);

(iii) if M^i has a **B** or **D** label then face $\mathcal{A}_i$ is transient and has no second vector field.

In this two dimensional case the sign of the second vector field depends only upon the angles of M' and M^1 for $\mathcal{A}_1$, M'' and M^2 for $\mathcal{A}_2$. We repeat the angles $\psi_1 = \arg_1(M')$ and $\psi_2 = \arg_2(M'')$ that M' and M'' make relative to axes $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively, so $\psi_i = 0$ is in the $\mathcal{A}_i$ direction and ψ_1 increases anticlockwise while ψ_2 increases clockwise.

The labelling below is a direct result of Lemma 2.3.7, so that following the sign of the second vector field, we modify the labels for M^i, $i = 1, 2$ to

$$\mathbf{A}^+ : \varphi_i + \psi_i < \pi, \quad \mathbf{A}^- : \varphi_i + \psi_i > \pi, \quad \mathbf{A}^0 : \varphi_i + \psi_i = \pi. \tag{2.11}$$

Using this labelling system we can identify 25 different cases to deal with. It turns out that in many of the cases we get the same result for all choices of the two cones i.e. all slopes $d' \equiv d_2/d_1 \in (0, \infty)$ of the line $\ell(d)$ separating them. Theorem 2.4.1 classifies these invariant cases.

Theorem 2.4.1 *The system is*

(1) ergodic in cases $\mathbf{A}^-/\mathbf{A}^- \cup \mathbf{B}$, $\mathbf{B}/\mathbf{A}^-$, $\mathbf{B}/\mathbf{B}$ *with* $\left|\frac{M_1^1}{M_2^1}\right| > \left|\frac{M_1^2}{M_2^2}\right|$

(2) transient in cases $\mathbf{A}^+/\mathbf{A}^+$, $\mathbf{A}^+/\mathbf{A}^- \cup \mathbf{B} \cup \mathbf{D}$, $\mathbf{A}^- \cup \mathbf{B} \cup \mathbf{D}/\mathbf{A}^+$, $\mathbf{B}/\mathbf{B}$ *with* $\left|\frac{M_1^1}{M_2^1}\right| < \left|\frac{M_1^2}{M_2^2}\right|$, $\mathbf{D}/\mathbf{B}$, $\mathbf{B}/\mathbf{D}$, $\mathbf{D}/\mathbf{D}$;

(3) null recurrent in cases $\mathbf{A}^0/\mathbf{A}^0 \cup \mathbf{A}^+ \cup \mathbf{B}$, $\mathbf{A}^+ \cup \mathbf{B}/\mathbf{A}^0$, $\mathbf{B}/\mathbf{B}$ *with* $\left|\frac{M_1^1}{M_2^1}\right| = \left|\frac{M_1^2}{M_2^2}\right|$.

For systems with no control over the service regimes there still may be some control over the routable traffic stream. The next theorem shows that there are sets of parameters such that a change to the slope of the switching line $\ell(d)$ can change Ξ from a transient to an ergodic process. We describe in detail only the case $\mathbf{D}/\mathbf{A}^0 \cup \mathbf{A}^-$, depicted in Fig. 2.12, as case $\mathbf{A}^0 \cup \mathbf{A}^-/\mathbf{D}$ is very similar. The relative slopes of M^1, $\ell(d)$ and M^2 are crucial so we label two key conditions:

E1: $M_2^1 < d'M_1^1$ (so $\ell(d)$ is steeper than M^1); E1′: $M_2^1 > d'M_1^1$;

E2: $-M_2^2 \leq d'(-M_1^2)$ (includes cases with $M_2^2 \geq 0$ and implies $-M^2$ is not steeper than $\ell(d)$).

Theorem 2.4.2 *In case* $\mathbf{D}/\mathbf{A}^0 \cup \mathbf{A}^-$ *the ergodicity or non-ergodicity of the Markov chain* Ξ *also depends on the slope* $d' > 0$ *of the line* $\ell(d)$ *separating* $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *as follows:*

(a) if E1 holds then Ξ *is transient,*

(b) if E1′ holds then Ξ*'s excursions into* $\mathcal{C}_1$ *have finite mean time and* Ξ *is*

(i) ergodic if E2 holds and M^2 *is* $\mathbf{A}^-$ *or if E2 does not hold and* $(-M_2^2)M_1^1 < M_2^1(-M_1^2)$ *(so* M^1 *is steeper than* $-M^2$*);*

(ii) null recurrent if E2 holds and M^2 *is* $\mathbf{A}^0$ *or if E2 does not hold and* $(-M_2^2)M_1^1 = M_2^1(-M_1^2)$*;*

(iii) transient if E2 does not hold and $(-M_2^2)M_1^1 > M_2^1(-M_1^2)$.

The case $\mathbf{A}^0 \cup \mathbf{A}^-/\mathbf{D}$ *is simply the reflection of the above in the line* $\ell(1,1)$.

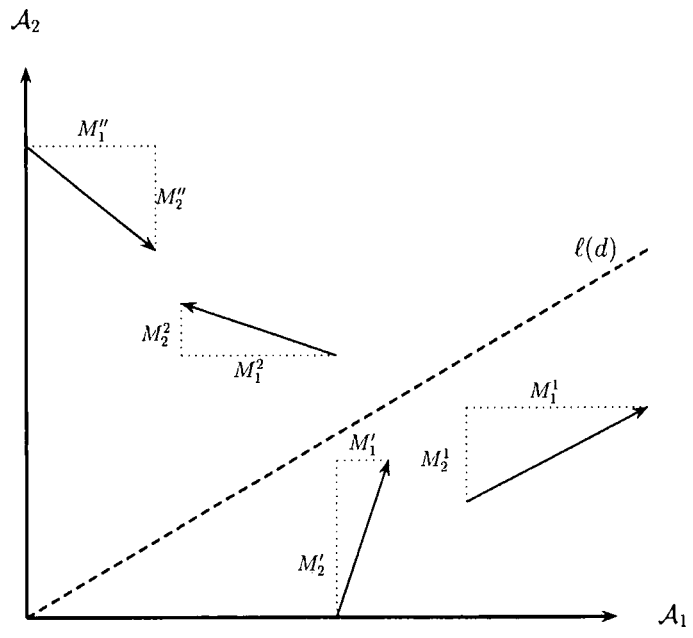


Figure 2.12: Example of case $\mathbf{D}/\mathbf{A}^-$ where $\ell(d)$ is important.

Note: this theorem says nothing about the cases where M^1 is parallel to $\ell(d)$ but in practice this will not be a major problem if the slope of the line $\ell(d)$ is under user control.

Example 2.4.1 (Load-balanced network) Kurkova [22] establishes stability conditions for a two queue model with Poisson arrival streams at rates λ_1 and λ_2, two fixed servers with rate 1 exponentially distributed service times. There is also an arrival stream λ which is routed to the shortest queue and some of the fed back jobs also join the shortest queue. Again we see that the boundary reflexion condition can be applied. The feedback from queue i to queue 1 and 2 is given by $p_i = p_{i,1} + p_{i,2} + p_{i,\{1,2\}}$, where $p_{i,\{1,2\}}$ indicated the proportion of the customers that have completed service at queue i and join the shorter of the two queues. Customers leave this system with probability $p_{i0} = 1 - p_i$. Our Theorems 2.4.1 and 2.4.2 above hold for a in many ways more general model and our results coincide with those of Kurkova. The only slight difference is that Kurkova [22] assumes a jump distribution on $\ell(d)$ different to those on $\mathcal{C}_i$.

Assume we have the following five blocks, the axes $\mathcal{A}_i$, two cones $\mathcal{C}_i$ with $i = 1$, 2 and $\ell(d)$ with slope $d' = 1$ as the last block. Since the service rates are all 1 the mean drifts are given by

$$\begin{aligned}
M' &= \frac{1}{\rho}\left(\lambda_1 - p_{10},\ \lambda + \lambda_2 + p_{12} + p_{1,\{2\}}\right) \\
M'' &= \frac{1}{\rho}\left(\lambda + \lambda_1 + p_{21} + p_{2,\{1\}},\ \lambda_2 - p_{20}\right) \\
M^1 &= \frac{1}{\rho}\left(\lambda_1 + p_{21} - p_{10},\ \lambda + \lambda_2 + p_{12} + p_{1,\{2\}} - p_{20}\right) \\
M^2 &= \frac{1}{\rho}\left(\lambda + \lambda_1 + p_{21} + p_{2,\{1\}} - p_{10},\ \lambda_2 + p_{12} - p_{20}\right) \\
M^\ell &= \frac{1}{2}\left(M_1^1 + M_1^2,\ M_2^1 + M_2^2\right).
\end{aligned}$$

We can see that the additional mean drift along the $\ell(d)$ does not make the problem any more complicated. Consider the four mean drifts M^i, M' and M'' in the positive recurrent cases given in Theorem 2.4.1; due to the nature of the JSQ routing Kurkova's model is stable in the cases $\mathbf{A}^-/\mathbf{A}^-$, $\mathbf{A}^-/\mathbf{B}$ and $\mathbf{B}/\mathbf{A}^-$, i.e. all the cases

where the process hardly visits $\ell(d)$ or *drives through* it. If we are in case **B/B** then M^ℓ has more impact as the process Ξ spends most of its time *driving into* $\ell(d)$. Therefore we need the following there exists a $v \in \mathbf{R}^2_+$ such that $v^\mathsf{T} M^i < 0$ for $i = 1$, 2, ℓ.

Example 2.4.2 (Move-able servers but no routable arrivals) In the modified Jackson feedback model described by Foley and MacDonald [14] there are two queues and server 2 helps the other if queue 2 is empty. This is a special case of our model but their main interest is in asymptotic estimates of the equilibrium distribution. Their stability result, Proposition 1, is quite old and appears as case (i) of Theorem 3.3.1(b) of [11]. Our Theorem 2.4.1 above discusses stability conditions for more complex models with two queues.

To illustrate the typical behaviour of the process Ξ under some of the cases some sample path simulations are included in Figure 2.13.

The proofs of Theorems 2.4.1 and 2.4.2 are very much based on the previous proofs. The positive recurrent cases are special cases of Theorems 2.3.2 and 2.3.4, transience can be established via Theorem 2.3.1 and the null-recurrent cases are partly covered by Theorem 2.3.10.

Proof of Theorem 2.4.1(i) & (ii) In part (i) we can use a linear Lyapunov function. If the mean drift in either $\mathcal{C}_i$ is labelled $\mathbf{A}^-$ then we need to use the waiting times given in Lemma 2.3.7 whenever the process visits the $\mathcal{A}_i$. The argument follows that for Theorem 2.3.2.

For part (ii) we can again use linear Lyapunov functions for all pairs of labels. For example, with pair $\mathbf{A}^-/\mathbf{A}^+$ we use function $f(z) = w^\mathsf{T} z$ where $w_1 < 0$ and $w_2 > 0$, again employing the waiting times given in Lemma 2.3.7 when the process visits the $\mathcal{A}_i$. The other cases all work similarly and the argument follows that for Theorem 2.3.1.

A note on part (iii) The null recurrence of the cases listed in part (iii) cannot be established using purely linear Lyapunov functions. The problem is that we have

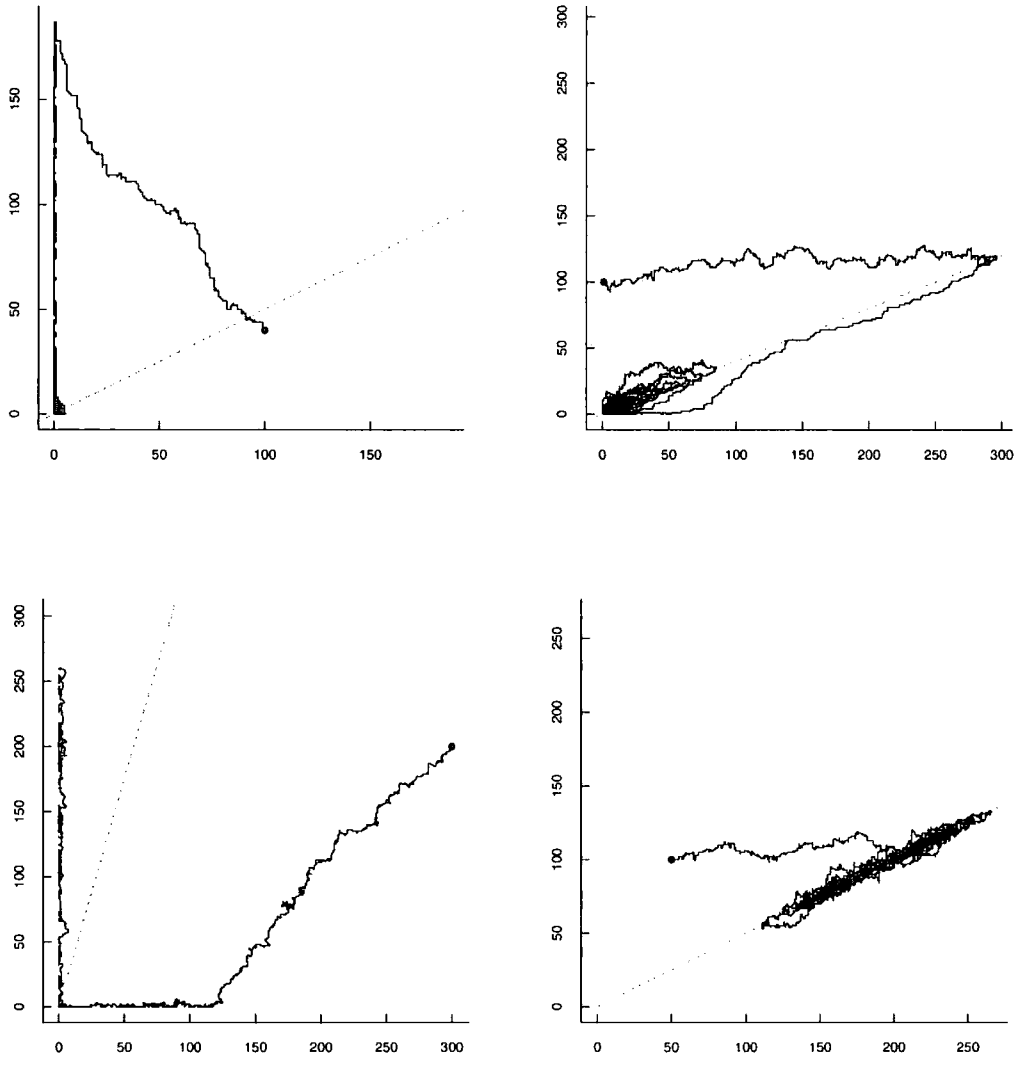


Figure 2.13: Simulations of typical paths for cases, from top left: $\mathbf{B}/\mathbf{A}^-$, $\mathbf{A}^-/\mathbf{D}$, $\mathbf{A}^-/\mathbf{A}^+$, $\mathbf{B}/\mathbf{B}$ with $\left|\frac{M_x^1}{M_y^1}\right| = \left|\frac{M_x^2}{M_y^2}\right|$ cases.

The dot indicates the starting point and the dashed line is $\ell(\cdot)$. One can see that the process in case $\mathbf{B}/\mathbf{A}^-$ does not look positive recurrent until it hits $\mathcal{A}_2$; for $\mathbf{D}/\mathbf{A}^-$ the process changes it behaviour when hitting $\ell(\cdot)$; the process in case $\mathbf{A}^-/\mathbf{A}^+$ shows its transience only after hitting the $\mathcal{A}_2$; and the null-recurrent case $\mathbf{B}/\mathbf{B}$ hugs $\ell(\cdot)$ but never reaches the origin.

to prove the non-ergodicity (case 4(a) in Theorem 2.3.11) as well as the recurrence (see 4(b)) of the Markov chain - usually one of the two is reasonably easy while the other one is difficult to establish.

Looking the cases $\mathbf{A}^0/\mathbf{A}^0 \cup \mathbf{A}^+ \cup \mathbf{B}$, $\mathbf{A}^+ \cup \mathbf{B}/\mathbf{A}^0$, $\mathbf{B}/\mathbf{B}$ with $\left|\frac{M_x^1}{M_y^1}\right| = \left|\frac{M_x^2}{M_y^2}\right|$ it becomes clear that they all behave rather differently and each case might have to be split again into separate cases. We will consider one case briefly in more detail, the case $\mathbf{B}/\mathbf{B}$ with $\left|\frac{M_x^1}{M_y^1}\right| = \left|\frac{M_x^2}{M_y^2}\right|$. There exists a w with $w_i > 0$ for the line $L_w(\alpha)$ containing to $M^i(\alpha)$ for $i = 1, 2$. We can see that

$$\mathbf{E}(g(\xi_{n+1}) - g(\xi_n) \mid \Pi^p \xi_n = \alpha) = 0 \text{ for } \alpha \in \mathbf{Z}_+^2 \tag{2.12}$$

where $g(\alpha) = w_1 x + w_2 y$. Including the reflexion vectors M' and M'' we can distinguish (1) $\psi_1 + \psi_2 > 3\pi/2$, (2) $\psi_1 + \psi_2 = 3\pi/2$, or (3) $\psi_1 + \psi_2 < 3\pi/2$.

Cases (2) and (3) are covered in the proof of Theorem 2.3.10. The case (1) is special in that we only have null-recurrency because of the low levels of control (remembering that $\psi_1 + \psi_2 > 3\pi/2$ in the proof of Theorem 2.3.4 we enough for positive recurrence with two blocks). Here however (1) implies that g in (2.12) is a supermartingale on $\mathbf{Z}_0^2$ since $g(M'), g(M'') < 0$ giving the recurrence of the process together with (2.12). The non-ergodicity can be established using a locally linear function, but we will not do this here.

Sketch of proof of Theorem 2.4.2 The argument uses all the same ideas as the earlier proofs. We discuss only the case $\mathbf{D}/\mathbf{A}^-$ as $\mathbf{A}^-/\mathbf{D}$ is so similar and the null recurrent cases are rather more delicate. When conditions E1 ($\ell(d)$ is steeper than M^1) or E2 ($-M^2$ is no steeper than $\ell(d)$) hold there exists $w \in \mathbf{R}^2$ with $w_1 > 0$ such that $w^\mathsf{T} M^1 = -w^\mathsf{T} M^2 = 1$ and we will use the process $w^\mathsf{T}\xi(t)$ to generate our semi-martingales.

(a) For $\xi(0) = \alpha \in \mathcal{C}_1$ if E1 holds then, by suitable application of Theorem 1.2.1(ii), we can show there is positive probability that Ξ never exits $\mathcal{C}_1$ and conditioned on this event, $w^\mathsf{T}\xi(t) \to \infty$ almost surely as $t \to \infty$ and so Ξ is transient.

(b) If E1 does not hold then Ξ surely exits $\mathcal{C}_1$ and if E2 holds we can define a sequence of times N_n that enable us to use Theorem 1.2.1(i). For $\alpha \in \mathcal{C}_1$ we define a state dependent time $T(\alpha)$ as follows. Let $d^\perp$ denote the unit vector perpendicular

to d on the side of cone $\mathcal{C}_2$, the function $h(\alpha)$ be the perpendicular distance from α to $\ell(d)$ and $m(\alpha) = h(\alpha)/M^2 \cdot d^{\perp} > 0$. For suitably large m_0 and small $\delta > 0$ let $T(\alpha) = \max\{m_0, \lceil 3m(\alpha)^{1+\delta} \rceil\}$. Using Law of Large Numbers type estimates we can show that

$$\mathbf{E}\big(w^{\mathsf{T}}\xi(t+T(\alpha)) - w^{\mathsf{T}}\xi(t) \mid \xi(t) = \alpha \in \mathcal{C}_1\big) < -\frac{1}{4}T(\alpha)$$

so given $\xi(N_n) = \alpha \in \mathcal{C}_1$ we define $N_{n+1} = N_n + T(\alpha)$. For $\alpha \in \mathcal{C}_2$ we proceed exactly as in Case I of the proof of Theorem 2.3.4 to define a supermartingale $Y_n = S_{N_n} = w^{\mathsf{T}}\xi(N_n)$ on all of $\mathbf{Z}_+^2$ outside some finite ball around $\underline{0}$ and ergodicity follows from Theorem 1.2.1(i).

If neither E1 or E2 hold then the LLN type estimate fails because the line $\ell(d)$ becomes an ergodic face and the behaviour of Ξ is determined by the relative sizes of $\arg_1(M^1)$ and $\arg_1(-M^2)$ just as in Lemma 2.3.7(i).

2.5 Examples

Networks with or without re-entrant lines seem to become very interesting only when there are more than two queues (see for example the generalised Lu-Kumar network in Section 4.4). Nevertheless looking at the two dimensional example can help to understand why some of the intuition fails once we remove the possibility to have external arrivals to all queues. The material presented here on re-entrant lines is published in [28].

The three examples discussed here are the *one server re-entrant line*, *two queues in tandem* and *two queues with feedback*. For the first two of these models we have a Poisson arrival stream λ which is sent to queue i only and never to queue j thus making it impossible to send arrivals to queue j by changing the routing scheme s, thus we need to apply the *boundary sojourn condition* (2.2).

Example 2.5.1 (One server re-entrant line) The smallest networks with a re-entrant line is a one single server station with two queues and one external arrival stream of rate λ as depicted in Figure 2.14. The external arrivals enter queue 1

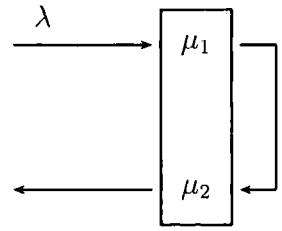


Figure 2.14: One server re-entrant station

where the exponentially distributed service time has parameter μ_1 when the server is there. After a customer has received service at queue 1 it enters queue 2 to receive further service, this time at rate μ_2. The system has two service regimes with mean drift vectors $M^1 = \frac{1}{\rho}(\lambda - \mu_1, \mu_1)$ and $M^2 = \frac{1}{\rho}(\lambda, -\mu_2)$. $\mathcal{M}$ satisfies the conditions of Theorem 2 precisely when the well-known condition $\lambda/\mu_1 + \lambda/\mu_2 < 1$ holds and in this case the system is stable no matter what non-idling control policy is used. If $\lambda/\mu_1 + \lambda/\mu_2 > 1$ then $\mathcal{M}$ satisfies the conditions of Theorem 2.3.1 and the system is always unstable. Further Theorem 2.4.1(iii) states that the system is null recurrent if $\lambda/\mu_1 + \lambda/\mu_2 = 1$.

Example 2.5.2 (Two queues in tandem) Consider a network with two single server stations where each job visits servers 1 and 2 in that order. This model, with general arrival and service time distributions, appears as an example in Down and Meyn [10] to demonstrate how stability conditions for it can be established using piecewise linear test functions. Figure 2.15 with $\lambda_2 = 0$ and $p^\eta_{12} = p^\eta_{20} = 1$ depicts this network. We have a Poisson arrival stream at rate λ_1 and service rates μ_i for $i = 1, 2$. This two dimensional system has regimes $\eta = \{1, 2, 3\}$ corresponding to service at queue 1, 2 or both queues respectively with mean drift vectors M^i where

$$M^1 = \frac{1}{\rho}(\lambda_1 - \mu_1,\ \mu_1), \qquad M^2 = \frac{1}{\rho}(\lambda_1,\ -\mu_2), \qquad M^3 = \frac{1}{\rho}(\lambda_1 - \mu_1,\ \mu_1 - \mu_2).$$

$\mathcal{M}$ satisfies the conditions of Theorem 2.3.2 if $\lambda_1/\mu_1 + \lambda_1/\mu_2 < 1$ and in this case the process is stable under any non-idling control. If, on the other hand, $\lambda_1/\mu_1 > 1$ or $\lambda_1/\mu_2 > 1$ then $\mathcal{M}$ satisfies the conditions of Theorem 2.3.1 and the system is sure to be unstable.

If $\lambda_1/\mu_1+\lambda_1/\mu_2 > 1$ but $\lambda_1/\mu_1 < 1$ and $\lambda_1/\mu_2 < 1$ then $\mathcal{M}$ satisfies the conditions of case C3 under any non-idling strategy. If the boundary sojourn condition (2.2) is satisfied we can apply Theorem 2.3.4 to deduce that a policy which keeps the system stable exists. We now establish the boundary sojourn condition. From the state $\underline{0}$ (with both queues empty) the jumps e_1, e_1, $e_2 - e_1$ lead to the state $(1,1) \in \mathbf{Z}_+^2$ and has positive probability under any policy that ensures service at queue 1. Hence there exists a constant $\hat{p} > 0$ such that for each $\alpha \in \partial\mathbf{Z}_0^N$ there is a policy Π_α such that the sojourn time τ satisfies

$$\mathbf{P}(\tau > t+3 \mid \tau > t, \xi(0) = \alpha, \Pi_\alpha) < 1 - \hat{p}$$

It follows from this that starting from any boundary point α τ is stochastically smaller than a random variable Z with geometrically bounded tails and finite mean ζ say and so condition (2.2) is satisfied. This shows that our results are consistent with the known properties of this system.

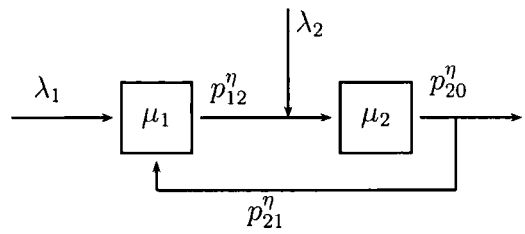


Figure 2.15: Two queues in tandem

Example 2.5.3 (Two queues with feedback) This is a variation of the tandem queue above with the following additions. We introduce a second arrival stream with rate λ_2 to queue 2 and allow completed jobs to feed back into the system. Jobs leaving queue $i = 1, 2$ independently enter queue $j = 1, 2$ with probability p_{ij}^η and leave the system with $p_{i0}^\eta = 1 - (p_{ii}^\eta + p_{ij}^\eta) \geq 0$ (see Figure 2.15). The mean drifts under the three service regimes are $\rho M^1 = (\lambda_1 + p_{11}\mu_1 - \mu_1, \lambda_2 + p_{12}\mu_1)$, $\rho M^2 = (\lambda_1 + p_{21}\mu_2, \lambda_2 + p_{22}\mu_2 - \mu_2)$ and $\rho M^3 = (\lambda_1 + \sum_j p_{j1}\mu_j - \mu_1, \lambda_2 + \sum_j p_{j2}\mu_j - \mu_2)$. This example satisfies the boundary reflection condition and so is covered by the complete stability classification of two queue models in Theorems 2.4.1 and 2.4.2.

Chapter 3

Extensions of the two queue model

The results stated in the previous chapter rely on the fact that we have a Markov chain Ξ as the queue length process on $\mathbf{Z}_0^2$ and that the mean drifts M^η under a management regime η can be expressed as vectors in $\mathbf{Z}_0^2$ in order to describe the convex hull $\mathcal{M}$, which is used to establish our results under different control policies. In this chapter we will drop some of the restrictive assumptions on the model parameters which leads to a state space different from $\mathbf{Z}_0^2$ that the Markov chain lives in. We will investigate where and how our results from Chapter 2 could apply.

The first assumption we drop is that of exponentially distributed service times. In practice it is often desirable to have a more general distribution for the service times, for modelling purposes we would however still like to be able to model the queueing process as a Markov chain. In order to do this we introduce phase type service. A detailed introduction to phase type distributions can be found in Neuts' book [30]. The advantage of assuming phase type service is that we can keep the Markovian setting while being able to approximate any distribution on $(0, \infty)$ using phase type distributions, see Asmussen [1]. The challenge is to describe the state space of the queueing process, which now also involves stages of service, then to find an appropriate way to describe the mean drifts M^η to get the convex hull $\mathcal{M}$ and thus being able to apply the earlier results.

Similarly restrictive to assuming exponentially distributed service times is to

consider exponential inter arrival times. The question is whether we can relax the arrival process assumption be introducing a Markovian arrival process (short denoted by MAP). The advantage being that again the Markovian property of the queueing process is preserved by considering the stages of arrival as part of the process. An introduction to the Markovian and other arrival processes can be found in Neuts [30].

There are two ideas we would like to introduce in order to help us with the analysis of the two variations of our model, one is the idea of induced Markov chains and the other is the joint transition or generator matrix of two finite state Markov chain.

The induced Markov chain

We will start with a low dimensional example for an induced Markov chain, called a *Markov chain in a half-strip.* The result in Fayolle, Malyshev, Menshikov [11] Section 3.1, pp. 33, says that if a discrete time Markov chain Ξ, on a state space $\mathbf{Z}_0 \times \{1, \ldots, n\}$ with states (x, i) is homogeneous for all but finitely many x and its jumps are bounded, the ergodicity or transience can be determined by checking the sign of the mean drift of the one dimensional Markov chain in x direction. The homogeneity assumption is that the jumps of the Markov chain Ξ on the half-strip are given by p_{ij}^k, i.e. a transition from (x, i) to $(x+k, j)$ does not depend on the state x (except for a finite number of x). The induced Markov chain Ξ is the finite state Markov chain with states $i \in \mathcal{B} = \{1, \ldots, n\}$. We assume that the Markov chain on the finite state space $\mathcal{B}$ is irreducible so there is only one essential class. The transitions of the induced chain from state i to j are denoted by $q_{ij} = \sum_k p_{ij}^k$ and each state i we get $\pi(i)$ where π is the stationary distribution of the Markov chain with jumps on $\mathcal{B}$ and a mean jump from the point (x, i) given by $M(i) = \sum_{j,k} k p_{ij}^k$. The mean drift on the half-strip is then given by

$$M = \sum_i \pi(i) M(i) \quad \text{for } i \in \mathcal{B}. \tag{3.1}$$

It follows that if $M(i) < \infty$ for all i then the Markov chain Ξ is ergodic (FMM [11], Theorem 3.1.2).

FMM [11], Chapter 4 has an even more detailed introduction to induced Markov chains and the idea of second vector fields for Markov chains on $\mathbf{Z}_+^N$. One of the changes to the *half-strip* model we are interested in is where several finite state spaces are attached to the countable state space, such that the Markov chain lives on $\mathbf{Z}_0 \times \{0, \ldots, c\} \times \{1, \ldots, m\}$ which is some kind of *bundle* of $\mathbf{Z}_0$ (this corresponds to the state space of the two queues in tandem model with a single Markovian arrival process and a finite buffer at the second queue, see Section 3.2). The other application we are interested in is a Markov chain on $\mathbf{Z}_0^2 \times \{1, \ldots, s_1\} \times \{1, \ldots, s_2\}$, which corresponds to the state space of a two queue model with phase type service at both queues, see Section 3.1.

In order to work with these additions to the half-strip model idea in the proceeding sections we will first introduce some notation concerning the stationary distribution of two Markov chains which are *joint*.

Two independent Markov chain's jumps combined

Assume we have two finite state, irreducible continuous time Markov chains $\mathfrak{V} = \{V(t) : t \geq 0\}$ and $\mathfrak{W} = \{W(t) : t \geq 0\}$ with states $i_1 \in S_1 = \{1, \ldots, s_1\}$ and $i_2 \in S_2 = \{1, \ldots, s_2\}$ respectively. The generator matrices are given by

$$\mathbf{L} = \begin{pmatrix} -l_1 & l_{12} & \ldots & l_{1s_1} \\ l_{21} & -l_2 & \ldots & l_{2s_1} \\ \vdots & & \ddots & \vdots \\ l_{s_1 1} & l_{s_1 2} & \ldots & -l_{s_1} \end{pmatrix} \quad \text{and} \quad \mathbf{K} = \begin{pmatrix} -k_1 & k_{12} & \ldots & k_{1s_2} \\ k_{21} & -k_2 & \ldots & k_{2s_2} \\ \vdots & & \ddots & \vdots \\ k_{s_2 1} & k_{s_2 2} & \ldots & -k_{s_2} \end{pmatrix} \tag{3.2}$$

By definition we know that the stationary distributions of both Markov chains exist.

We would now like to consider the joint Markov chain $\mathfrak{Z} = \{Z(t) : t \geq 0\}$ of the two processes $\mathfrak{V}$ and $\mathfrak{W}$. The joint process we are interested in has non-zero transitions of the form

$$(i_1, i_2) \to (j_1, i_2) \text{ with rate } l_{i_1 j_1} \text{ and}$$
$$(i_1, i_2) \to (i_1, j_2) \text{ with rate } k_{i_2 j_2}$$

where $\to$ denotes that there is a jump from one state (i_1, i_2) in $\mathcal{L}^2$ to another. The process stays in state (i_1, i_2) for an exponentially distributed time with parameter

$l_{i_1} + k_{i_2}$. This means the joint process $\mathfrak{Z}$ has jumps on a finite two dimensional lattice $\mathcal{L}^2 = \{(i_1, i_2) : i_1 = 1, \ldots, s_1 \text{ and } i_2 = 1, \ldots, s_2\}$. There are no diagonal jumps on $\mathcal{L}^2$.

Denote by $l_{i_1 j_1}\mathbf{I}$ the $s_2 \times s_2$ matrix with zero entries apart from $l_{i_1 j_1}$ on the diagonal. The generator matrix for the joint service processes $\mathfrak{Z}$ is then given by the $(s_1 s_2) \times (s_1 s_2)$ matrix

$$\begin{aligned} \mathbf{M} &= \mathbf{L} \circledast \mathbf{K} \\ &= \begin{pmatrix} -l_1\mathbf{I} + \mathbf{K} & l_{12}\mathbf{I} & l_{13}\mathbf{I} & \ldots & l_{1s_1}\mathbf{I} \\ l_{21}\mathbf{I} & -l_2\mathbf{I} + \mathbf{K} & l_{23}\mathbf{I} & \ldots & l_{2s_1}\mathbf{I} \\ l_{31}\mathbf{I} & l_{32}\mathbf{I} & -l_3\mathbf{I} + \mathbf{K} & \ldots & l_{3s_1}\mathbf{I} \\ \vdots & \vdots & & \ddots & \vdots \\ l_{s_1 1}\mathbf{I} & l_{s_1 2}\mathbf{I} & l_{s_1 3}\mathbf{I} & \ldots & -l_{s_1}\mathbf{I} + \mathbf{K} \end{pmatrix} \end{aligned} \tag{3.3}$$

the symbol $\circledast$ is introduced to denote this type of matrix multiplication[1]. Following Serfozo we can uniformise the Markov chain $\mathfrak{Z}$ governed by $\mathbf{M}$ by choosing a constant

$$\rho^* \geq \max_{i_1, i_2}\{l_{i_1} + k_{i_2}\} \tag{3.4}$$

so ρ^* is at least as big as the maximum diagonal element of $\mathbf{M}$. Using (3.4) we can define the discrete time Markov chain $Z = \{z(n) : n \geq 0\}$ of the joint process with the matrix $\mathbb{M}$ of transition probabilities given by

$$(\mathbb{M})_{(i_1 i_2),(j_1 j_2)} = \begin{cases} (\rho^*)^{-1} l_{i_1 j_1}, & (i_1 i_2) \to (j_1 i_2) \\ (\rho^*)^{-1} k_{i_2 j_2}, & (i_1 i_2) \to (i_1 j_2) \\ \rho^* - (l_{i_1} + k_{i_2}), & (i_1 i_2) \to (i_1 i_2) \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

Note that we have set up the continuous time processes $\mathfrak{Z}$ first in order to find the right *bell/null* event rate or *clock* speed ρ^* at which to compare the two processes. The product $\circledast$ works also for discrete time processes and their transition matrices.

[1]Although this seems to be so straight forward that it surely exists in some text book I have been unable to find it.

We also define the two discrete time versions of $\mathfrak{V}$ and $\mathfrak{W}$ as the Markov chains $V = \{v(n) : n \geq 0\}$ and $W = \{w(n) : n \geq 0\}$ respectively. The transition probability matrices $\mathbb{L}$ and $\mathbb{K}$ with ρ^* as in (3.4) are given by

$$(\mathbb{L})_{i_1,j_1} = \begin{cases} (\rho^*)^{-1} l_{i_1 j_1}, & (i_1) \to (j_1) \\ \rho^* - l_{i_1}, & (i_1) \to (i_1) \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (\mathbb{K})_{i_2,j_2} = \begin{cases} (\rho^*)^{-1} k_{i_2 j_2}, & (i_2) \to (j_2) \\ \rho^* - k_{i_2}, & (i_2) \to (i_2) \\ 0, & \text{otherwise} \end{cases}$$

Given all these Markov chains and their transition matrices we can note the following

Lemma 3.0.1 *Given the Markov chains V and W above with stationary distributions $\nu\mathbb{L} = \nu$ and $\theta\mathbb{K} = \theta$ respectively and $\mathbb{M} = \mathbb{L} \circledast \mathbb{K}$ we have*

$$\pi(i_1, i_2) = \nu(i_1)\theta(i_2)$$

where $\pi\mathbb{M} = \pi$ is the stationary distribution of Z.

Proof: Let $\bar{\pi}(i_1, i_2) = \nu(i_1)\theta(i_2)$ for all i_1, i_2. The rest follows from the uniqueness of the stationary distribution and

$$\begin{aligned} (\bar{\pi}\mathbb{M})_{\imath_1\imath_2} &= \textstyle\sum_{j_1}\sum_{j_2} \pi(i_1, i_2)(\mathbb{M})_{(\imath_1,\imath_2)\to(i_1,i_2)} \\ &= \pi(\imath_1, \imath_2)(\rho^* - (l_{\imath_1} + k_{\imath_2})) + \textstyle\sum_{i_1 \neq j_1} \pi(i_1, \imath_2)(\frac{1}{\rho^*} l_{i_1 j_1}) + \sum_{i_2 \neq j_2} \pi(\imath_1, j_2)(\frac{1}{\rho^*} k_{i_2 j_2}) \\ &= \nu(\imath_1)\theta(\imath_2)(\rho^* - (l_{\imath_1} + k_{\imath_2})) + \nu(\imath_1)\theta(\imath_2)(\tfrac{1}{\rho^*} l_{\imath_1}) + \nu(\imath_1)\theta(\imath_2)(\tfrac{1}{\rho^*} k_{\imath_2}) \\ &= \nu(\imath_1)\theta(\imath_2) \end{aligned}$$

where $l_{i_1} = \sum_{i_1 \neq j_1} l_{i_1 j_1}$. □

3.1 Service time of phase type

In this section we will consider the phase type distribution as the distribution of the service time in a two queue model. The class of phase type distributions is a highly versatile class of probability distributions - it is the distribution of the time until absorption in a finite Markov chain. One distribution in this class, introduced as a generalisation of the exponential distribution, is called the Erlang distribution or also the method of stages. The idea is that the customer goes through m stages of

exponentially distributed service time with the same rate μ. This can be generalised by assuming different rates at each stage. Another distribution that falls into the class of phase type distributions is the hyperexponential distribution, or exponential distributions in parallel channels; here one of m different exponentially distributed service times is chosen at random by the customer.

The general information about the phase type distributions is taken from, and more details and applications can be found in, Neuts [30], Chapter 2 and Asmussen [1] Chapter III, Part 6.

The class of phase type distributions dense in the set of all probability distributions on $(0, \infty)$ (see Asmussen [1] Chapter III, Section 6), which means it is applicable for a much broader range of service time distributions than the exponential distribution. Additionally the use of phase type distributions allows us to analyse the queueing process as a Markov chain as the phase type distribution preserves the Markov property, if the process is observed at the stages of service as well as the changes in queue length.

The two queue model that we will consider is a bit simpler than that in Chapter 2 in that we have two dedicated Poisson arrival streams λ_1 and λ_2 only and no feedback of customers that have completed service is allowed. We keep the general notion for service regimes with η. The simplifications are introduced to avoid over complicated notation, the results apply for the general model in Chapter 2 as well.

We will start with a short definition of the phase type distribution we would like to use and then describe what happens for a two queue model. Note that we will use the continuous time setting for the phase type distribution first and change to a discrete time queueing process when the continuous time process is completely set up.

The **phase type distribution** as the distribution of service times implies the following: every customer that receives service starts at some state, according to some initial distribution $\underline{\beta}$, of a finite state continuous time Markov chain $\mathfrak{Y} = \{Y(t) : t \geq 0\}$. This Markov chain has $i = 1, \ldots, s$ transient states and one absorbing state $s + 1$; upon arrival at state $s + 1$ the customer has completed service. There is only one customer per queue in service at any given time n. $\mathfrak{Y}$ has a transition matrix of the

form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{T} & \underline{t} \\ \underline{0} & 0 \end{pmatrix}$$

where $\mathbf{T}$ is a $s \times s$ matrix giving the transitions t_{ij} between the transient states $i, j = 1, \ldots, s$ and $\underline{t} = (t_{1s+1}, \ldots, t_{ss+1})^{\mathsf{T}}$ is the vector that gives the transitions to the absorbing state $s+1$. For our purpose of modelling service times in a queueing model we make the following additional assumptions about the phase type distribution:

(a) all jumps from states $i : i \neq 1$ to state 1 have transition $t_{i1} = 0$ and also $t_{1s+1} = 0$;

(b) the initial distribution $(\underline{\beta}(x), \beta_{s+1}(x))$ of any phase type distribution given here is

$$\underline{\beta}(x) = (\beta_1(x), 0, \ldots, 0) \text{ where } \beta_1(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \end{cases},$$

$$\beta_{s+1}(x) = \begin{cases} 0, & \text{if } x > 0 \\ 1, & \text{if } x = 0 \end{cases}$$

and x the number of customers waiting in the queue.

We say the phase type distribution, as given above, has representation $(\mathbf{T}, \underline{\beta}(x))$, is of order s and the first moment is given by

$$\mu = -\underline{\beta}(x)\mathbf{T}^{-1}\underline{e} \tag{3.6}$$

where $\underline{e} = (1, \ldots, 1)^{\mathsf{T}}$.

We will now briefly define the **renewal representation** of the phase type distribution. The idea is that as long as the queue length is positive the next customer starts service when the previous one leaves, in other words to restart the Markov chain $\mathfrak{Y}$, in our case from state 1, over and over again until the queue is empty. This is done using the so called *renewal* phase type distribution (or irreducible representation) which is given by

$$\mathbf{Q}^* = \mathbf{T} + \mathbf{T}^0\mathbf{A}^0 \tag{3.7}$$

where $\mathbf{T}^0 = (\underline{t}, \underline{t}, \ldots, \underline{t})$ and $\mathbf{A}^0 = (1 - \beta_{s+1}(x))^{-1}\text{diag}(\beta_1(x), 0, \ldots, 0)$ are $s \times s$ matrices, where diag indicates the matrix with all zero's except of the diagonal elements here given as $(\beta_1(x), 0, \ldots, 0)$. We will use the matrices $\mathbf{T}$ and $\mathbf{T}^0\mathbf{A}^0$ when setting up the queueing process Ξ.

Phase type service for the two queue model

For the two queue model as described above, under a service regime η which provides service to both queues, denote by $\mathbf{Q}_1^\eta$ on $S_1 = \{1, \ldots, s_1 + 1\}$ and $\mathbf{Q}_2^\eta$ on $S_2 = \{1, \ldots, s_2 + 1\}$ the transition matrices of the two Markov processes $\mathfrak{V}^\eta = \{V(t) : t \geq 0\}$ and $\mathfrak{W}^\eta = \{W(t) : t \geq 0\}$ for the service at queue 1 and 2 respectively. Note that the state space S_l of the phase type service at queue l can also change from one regime η to another but we will refrain from introducing further sub- or superscripts here. We denote by $(x, y) \in \mathbf{R}_0^2$ the number of customers in queue 1 and queue 2 respectively. We have two transition matrices of size $s_1 \times s_1$ and $s_2 \times s_2$ for the transient states only given by

$$\mathbf{D}^\eta = \begin{pmatrix} -d_1^\eta & d_{12}^\eta & \ldots & d_{1s_1}^\eta \\ 0 & -d_2^\eta & \ldots & d_{2s_1}^\eta \\ \vdots & & \ddots & \vdots \\ 0 & d_{s_12}^\eta & \ldots & -d_{s_1}^\eta \end{pmatrix} \quad \text{and} \quad \mathbf{T}^\eta = \begin{pmatrix} -t_1^\eta & t_{12}^\eta & \ldots & t_{1s_2}^\eta \\ 0 & -t_2^\eta & \ldots & t_{2s_2}^\eta \\ \vdots & & \ddots & \vdots \\ 0 & t_{s_22}^\eta & \ldots & -t_{s_2}^\eta \end{pmatrix} \tag{3.8}$$

where $t_{i_1}^\eta = t_{i_1 i_1}^\eta$ and $d_{i_2}^\eta = d_{i_2 i_2}^\eta$.

The joint Markov jump process $\mathfrak{Z}^\eta = \{Z(t) : t \geq 0\}$ of two service processes together jumps on a finite two dimensional lattice $\mathcal{L}_\eta^2 = \{(i_1, i_2) : i_1 = 1, \ldots, s_1 + 1 \text{ and } i_2 = 1, \ldots, s_2 + 1\}$ and the following non-zero transitions rates for all $j_1, j_2 \neq 1$

$$\begin{aligned} &(i_1, i_2) \rightarrow (j_1, i_2) \text{ with rate } d_{i_1 j_1}^\eta \text{ and} \\ &(i_1, i_2) \rightarrow (i_1, j_2) \text{ with rate } t_{i_2 j_2}^\eta \end{aligned} \tag{3.9}$$

where $\rightarrow$ denotes the transitions between states in $\mathcal{L}_\eta^2$. The time that the process remains in state (i_1, i_2) is exponentially distributed with parameter $d_{i_1}^\eta + t_{i_2}^\eta$.

In order to define the queueing process as a whole we have to set up some notation first. Denote by $d^{\eta}_{i_1 j_1}\mathbf{I}$ the $s_2 \times s_2$ matrix with zero entries apart from $d^{\eta}_{i_1 j_1}$ along the diagonal. The transition matrix (again excluding the absorbing states) for the joint service processes $\mathfrak{Z}^{\eta}$ is then given by the $(s_1 s_2) \times (s_1 s_2)$ matrix

$$\mathbf{Q}^{\eta} = \mathbf{D}^{\eta} \circledast \mathbf{T}^{\eta} = \begin{pmatrix} -d^{\eta}_1\,\mathbf{I} + \mathbf{T}^{\eta} & d^{\eta}_{12}\mathbf{I} & d^{\eta}_{13}\mathbf{I} & \dots & d^{\eta}_{1s_1}\mathbf{I} \\ 0 & -d^{\eta}_2\,\mathbf{I} + \mathbf{T}^{\eta} & d^{\eta}_{23}\mathbf{I} & \dots & d^{\eta}_{2s_1}\mathbf{I} \\ 0 & d^{\eta}_{32}\mathbf{I} & -d^{\eta}_3\,\mathbf{I} + \mathbf{T}^{\eta} & \dots & d^{\eta}_{3s_1}\mathbf{I} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & d^{\eta}_{s_1 2}\mathbf{I} & d^{\eta}_{s_1 3}\mathbf{I} & \dots & -d^{\eta}_{s_1}\,\mathbf{I} + \mathbf{T}^{\eta} \end{pmatrix} \tag{3.10}$$

as in (3.3).

In the following we will not use the superscript η every time any more, everything to do with service can however change with a change in service regime η. We define another two matrices of size $(s_1 s_2) \times (s_1 s_2)$ with $\mathbf{D}^0$ and $\mathbf{T}^0$, $\mathbf{A}^0_1$ and $\mathbf{A}^0_2$ as defined in (3.7) for queue 1 and queue 2 respectively.

$$\mathbf{D1} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ d_{2(s_1+1)}\mathbf{I} & 0 & \dots & 0 \\ d_{3(s_1+1)}\mathbf{I} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ d_{s_1(s_1+1)}\mathbf{I} & 0 & \dots & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{T2} = \begin{pmatrix} \mathbf{T}^0\mathbf{A}^0_2 & 0 & \dots & 0 \\ 0 & \mathbf{T}^0\mathbf{A}^0_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{T}^0\mathbf{A}^0_2 \end{pmatrix} \tag{3.11}$$

where the matrix $\mathbf{D1}^{\eta}$ gives the transition rates for customers leaving after service at queue 1 under regime η, and $\mathbf{T2}^{\eta}$ gives those for customers leaving from queue 2.

We set $\lambda = \lambda_1 + \lambda_2$ and $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_{12}$ denote identity matrices of size $(s_1 \times s_1)$, $(s_2 \times s_2)$ and $(s_1 s_2 \times s_1 s_2)$ respectively. The transitions to and from the boundary states are denoted by the following:

(1) If there are no jobs in queue 2, $(x, 0)$ with $x > 0$ then

$$\mathbf{DA} = \mathbf{D}^0\mathbf{A}^0_1 \otimes \mathbf{A}^0_1 \quad , \quad \mathbf{L}_1 = \lambda_1\mathbf{I}_1 \otimes \mathbf{A}^0_1 \quad \text{and} \quad \mathbf{LD} = (-\lambda_1\mathbf{I}_1 + \mathbf{D}) \otimes \mathbf{A}^0_1$$

where $\otimes$ denotes the tensor or Kronecker product of the two matrices. The three

matrices are of size $(s_1 s_2 \times s_1 s_2)$. The matrix $\mathbf{D}^0\mathbf{A}_1^0$ corresponds to the departing jobs, $\lambda_1\mathbf{I}_1$ to the arrivals and $-\lambda_1\mathbf{I}_1 + \mathbf{D}$ to remaining in service at queue 1. After each entry of these $(s_1 \times s_1)$ matrices there are $s_2 - 1$ rows and columns of zeros (this is achieved by the tensor product with $\mathbf{A}_1^0$), which makes them the desired size of $(s_1 s_2 \times s_1 s_2)$.

(2) For the case where $(0, y)$ with $y > 0$ we get

$$\mathbf{TA} = \begin{pmatrix} \mathbf{T}^0\mathbf{A}_2^0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix}, \quad \mathbf{L}_2 = \begin{pmatrix} \lambda_2\mathbf{I}_2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} \quad \text{and}$$

$$\mathbf{LT} = \begin{pmatrix} -\lambda_2\mathbf{I}_2 + \mathbf{T} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix}$$

as the corresponding matrices for queue 2 which are of size $(s_1 s_2 \times s_1 s_2)$ also. This time we add $s_1 - 1$ times $(s_2 \times s_2)$ matrices with all zero entries to the first $(s_2 \times s_2)$ non-zero matrix.

For the whole queueing process we now have two two-dimensional state spaces, $(i_1, i_2) \in \mathcal{L}_\eta^2$ and the states $(x, y) \in \mathbf{R}_0^2$. If we look only at the changes in queue length, then these happen at the following transition rates:

$$\begin{aligned} (x, i_1, y, i_2) &\to (x+1, i_1, y, i_2) \text{ with rate } \lambda_1 \\ (x, i_1, y, i_2) &\to (x, i_1, y+1, i_2) \text{ with rate } \lambda_2 \\ (x, i_1, y, i_2) &\to (x-1, i_1, y, i_2) \text{ with rate } \underline{d} \\ (x, i_1, y, i_2) &\to (x, i_1, y-1, i_2) \text{ with rate } \underline{t} \end{aligned} \tag{3.12}$$

where $\underline{d}$ and $\underline{t}$ are the vectors of transition from states i_1 and i_2 to the absorbing states $s_1 + 1$ and $s_2 + 1$ respectively.

Given all these transitions we can now set up the actual transition matrix $\mathbf{P}$ for the $N = 2$ queue process with phase type service. We will write out four matrices

depending on whether we are in the interior, between interior and boundary or on the boundary of the queue lengths state space $\mathbf{R}_0^2$.

Between the boundary states we have

$$\mathbf{P}_{00} = \begin{pmatrix} -\lambda\mathbf{I}_{12} & \lambda_1\mathbf{I}_{12} & 0 & 0 & \dots & \lambda_2\mathbf{I}_{12} & 0 & \dots \\ \mathbf{DA} & \mathbf{LD} & \mathbf{L}_1 & 0 & \dots & 0 & 0 & \dots \\ 0 & \mathbf{DA} & \mathbf{LD} & \mathbf{L}_1 & \dots & 0 & 0 & \dots \\ \vdots & & \vdots & & \vdots & & \vdots & \\ \mathbf{TA} & 0 & 0 & 0 & \dots & \mathbf{LT} & \mathbf{L}_2 & \dots \\ 0 & 0 & 0 & 0 & \dots & \mathbf{TA} & \mathbf{LT} & \dots \\ \vdots & & \vdots & & \vdots & & \vdots & \end{pmatrix}$$

Form boundary states to interior states

$$\mathbf{P}_{01} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots \\ \mathbf{L}_2 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \\ 0 & \mathbf{L}_2 & \dots & 0 & 0 & \dots & 0 & 0 & \dots \\ \vdots & & \vdots & & \vdots & & \vdots & & \\ \mathbf{L}_1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots \\ 0 & 0 & \dots & \mathbf{L}_1 & 0 & \dots & 0 & 0 & \dots \\ \vdots & & \vdots & & \vdots & & \vdots & & \end{pmatrix}$$

From interior states to boundary states

$$\mathbf{P}_{10} = \begin{pmatrix} \mathbf{T2} & 0 & 0 & \dots & \mathbf{D1} & 0 & \dots & \\ 0 & 0 & \mathbf{T2} & 0 & \dots & 0 & 0 & \dots \\ \vdots & \vdots & & \vdots & & \vdots & & \\ 0 & 0 & 0 & 0 & \dots & 0 & \mathbf{D1} & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots \\ \vdots & \vdots & & \vdots & & \vdots & & \end{pmatrix}$$

Between interior states

$$\mathbf{P}_{11} = \begin{pmatrix} -\lambda\mathbf{I}_{12}+\mathbf{Q} & \lambda_1\mathbf{I}_{12} & \dots & \lambda_2\mathbf{I}_{12} & 0 & \dots & 0 & 0 & \dots \\ \mathbf{D1} & -\lambda\mathbf{I}_{12}+\mathbf{Q} & \dots & 0 & \lambda_1\mathbf{I}_{12} & \dots & 0 & 0 & \dots \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & \\ \mathbf{T2} & 0 & \dots & -\lambda\mathbf{I}_{12}+\mathbf{Q} & \lambda_1\mathbf{I}_{12} & \dots & 0 & 0 & \dots \\ 0 & \mathbf{T2} & \dots & \mathbf{D1} & -\lambda\mathbf{I}_{12}+\mathbf{Q} & \dots & 0 & \lambda_2\mathbf{I}_{12} & \dots \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & \end{pmatrix}$$

Given these we can see that the transition matrix of the Markovian queueing process $\mathfrak{X}$ under η on $\mathcal{L}_\eta^2 \times \mathbf{R}_0^2$ has a transition matrix of the form

$$\mathbf{P}^\eta = \begin{pmatrix} \mathbf{P}_{00} & \mathbf{P}_{01} \\ \mathbf{P}_{10} & \mathbf{P}_{11} \end{pmatrix}.$$

This generator matrix depends on the regime η in force in the sense that all matrices $\mathbf{D}^\eta$ and $\mathbf{T}^\eta$ will change with η.

The discrete time process

Given the rates of jumps between the stages of service and changes in queue length in $\mathbf{P}^\eta$ we can define the discrete time queueing process Ξ by choosing a constant

$$\rho \geq \max_{\eta, i_1, i_2} \left\{ d_{i_1}^\eta + t_{i_2}^\eta + \lambda \right\}, \tag{3.13}$$

as in (3.4) ρ is bigger than the maximum diagonal element of all η generator matrices $\mathbf{P}^\eta$. The uniformised discrete time process Ξ on $\mathcal{L}_\eta^2 \times \mathbf{Z}_0^2$ has *null* events at exponential rate $\rho - (d_{i_1}^\eta + t_{i_2}^\eta + \lambda)$ under η when both queues receive service; under η when only queue 1 or 2 receive service the rate are given by $\rho - (d_{i_1}^\eta + \lambda)$ or $\rho - (t_{i_2}^\eta + \lambda)$ respectively. Ξ is the continuous time queueing process $\mathfrak{X}$ observed at the following events: a customer changes stage during service (at either of the two queues), a new arrival occurs, a customer leaves the system and at all null events.

We can picture the state space of Ξ in the following way. Changes to queue length happen on the positive lattice $\mathbf{Z}_0^2$ with states $\alpha = (x, y)$ indicating the number of customers waiting at each queue. There are $s_1 s_2$ layers of $\mathbf{Z}_0^2$ and each of these layers corresponds to a joint state of service (i_1, i_2).

We will now define the **mean drifts** with respect to the queue length i.e. in direction of $\alpha = (x, y) \in \mathbf{Z}_0^2$. As long as there are customers in queue $l = 1, 2$ we know that the service process has a renewal phase type distribution or irreducible representation (see $\mathbf{Q}_l^*$ in (3.7)). This means both service processes have a stationary distribution which we denote by ν and θ for queue 1 and 2 respectively. From Lemma 3.0.1 we know that we also have $\pi(i_1, i_2) = \nu(i_1)\theta(i_2)$ where π is the stationary distribution of $\mathbb{Q}^*$ - the irreducible representation of the discrete time equivalent of (3.7) with uniformising constant as in (3.13). Note that these distributions and matrices again change with η. With $\alpha, \beta \in \mathbf{Z}_0^2$ we also define

$$M(i_1, i_2) = \sum_{\beta} q_{\alpha\beta}$$

the mean jump from point (α, i_1, i_2) where $q_{\alpha\beta}$ gives the jumps that correspond to arrival $\rho^{-1}\lambda_l$ with $l = 1, 2$ or to departures from the queues $\rho^{-1}\underline{d}$, $\rho^{-1}\underline{t}$ with ρ as in (3.13).

The mean drift is given by

$$M^{\eta} = \sum_{i_1, i_2} \pi(i_1, i_2) M(i_1, i_2)$$

under regime η when both queues receive service and for regimes η' and η'' on axis $\mathcal{A}_1$ and $\mathcal{A}_2$ of $\mathbf{Z}_0^2$ by

$$M^{\eta'} = \sum_{i_1, i_2} \nu(i_1) M(i_1, i_2) \quad \text{and} \quad M^{\eta''} = \sum_{i_1, i_2} \theta(i_2) M(i_1, i_2)$$

respectively.

The mean drifts form a convex hull as in Chapter 2 (2.5) and the discrete time process (Ξ, Π) is Markov if Π is a stationary, Markovian control policy. Thus the results of Theorems 2.3.1, 2.3.2 and 2.3.4 could be applied if we sample the process at fewer time points so that the service process is in some sense averaged out.

3.2 Markovian arrival process

The continuous time Markovian arrival processes denoted by $\mathfrak{A}$, similarly to the phase type service above, involves a finite state Markov chain. The idea is that

specific transition of the Markov chain $\mathfrak{A}$ correspond to a certain number b of arrivals, so if the Markov chain makes a transition from a state i to j there are b arrivals. We will first describe the Markovian arrival process for two queues and see what we can say about this and later consider a specific example of two queues in tandem with blocking at the second queue which was analysed by van Houdt and Alfa [38].

Although the Markovian arrival process can deal with batch arrivals where batches are of size b, see Lucantoni [25], we will concentrate on the Markovian arrival process with $b = 0$ or $b = 1$ as discussed by van Houdt and Alfa [38] or Neuts and Alfa [31].

We use a Markovian arrival processes as introduced in van Houdt and Alfa [38], though they use the discrete time Markovian arrival process (D-MAP) while we will start in a continuous time setting. The matrix of transition rates $\mathbf{B}$ of the Markovian arrival process $\mathfrak{B}$ is composed as follows. There are two $m \times m$ matrices B_0 and B_1 such that $\mathbf{B} = B_0 + B_1$ is a generator matrix, i.e. $\sum_j (B)_{ij} = 0$ where $(B)_{ij}$ are the elements of $\mathbf{B}$. We assume that $\mathfrak{B}$ is irreducible and aperiodic. The elements of the matrix $\mathbf{B}$ denoted by $(B_b)_{ij}$ do not only bear information about the rate of transitions from states i to j but also of making a transition from state i to j with b arrivals where $b = 0$ or 1. The Markovian process described by $\mathbf{B}$ has a stationary distribution which satisfies $\delta\mathbf{B} = 0$ and $\delta\underline{e} = 1$, where $\underline{e}$ is a m column vector with all ones. Given δ the arrival rate is given by $\Lambda = \delta B_1 \underline{e}$, see Lucantoni [25].

A variation of this process is given in Neuts and Alfa [31]. In their model arrivals of two types are generated by the discrete time Markovian arrival process. We can use this idea to generate two types of arrivals. In this case we have four matrices $A_{0,0}$, $A_{0,1}$, $A_{1,0}$ and $A_{1,1}$ of size m so that $\mathbf{A} = A_{0,0} + A_{0,1} + A_{1,0} + A_{1,1}$ is the generator matrix of the Markov chain $\mathfrak{A}$ which is irreducible and aperiodic. Similar to above we have $b, d \in \{0, 1\}$ and elements $(A_{b,d})_{i,j}$ that give information on the number b of arrivals to queue 1 and the number d of arrivals to queue 2 with any transition from state i to state j of the Markov chain. The stationary distribution $\theta\mathbf{A} = 0$ is again used to determine the two types of arrival rates

$$\Lambda_1 = \theta(A_{1,0} + A_{1,1})\underline{e} \quad \text{and} \quad \Lambda_2 = \theta(A_{0,1} + A_{1,1})\underline{e}.$$

This idea can of cause be taken further, say we would like a third type of arrivals which could than be routed to either queue 1 or 2 depending on our routing decision, then use a formulation like the following: $A_{0,0,0}$, $A_{1,0,0}$, $A_{0,1,0}$, $A_{0,0,1}$, $A_{1,1,0}$, $A_{0,1,1}$, $A_{1,0,1}$ and $A_{1,1,1}$ are eight matrices giving the transition rate with arrivals of type 1, 2 or 3. For simplicity we will refrain from doing this and consider a two queue system with a Markovian arrival process $\mathfrak{A}$ as above and arrival rates Λ_1 and Λ_2 to queues 1 and 2 respectively. We assume for now that the customers waiting in queue l have exponentially distributed service times at rate μ_{kl} under regime $\eta = (k, s)$ (with s fixed) at queue $l = 1, 2$. After receiving service at queue l a customer might be fed back into queue l' with probability $p_{ll'}$ or leave with probability $p_{l0} = 1 - \sum_{l'} p_{ll'}$.

As before we would like to consider the discrete time queueing process Ξ on $\mathbf{Z}_0^2 \times \{1, \ldots, m\}$. The non-zero jumps have rates

$$
\begin{aligned}
(x, y, i) &\to (x+1, y, j) \text{ with rate } (A_{1,0})_{ij} \\
(x, y, i) &\to (x, y+1, j) \text{ with rate } (A_{0,1})_{ij} \\
(x, y, i) &\to (x+1, y+1, j) \text{ with rate } (A_{1,1})_{ij} \\
(x, y, i) &\to (x, y, j) \text{ with rate } (A_{0,0})_{i,j} \\
(x, y, i) &\to (x+1, y, i) \text{ with rate } \sum_l \mu_l p_{l1} \\
(x, y, i) &\to (x, y+1, i) \text{ with rate } \sum_l \mu_l p_{l2} \\
(x, y, i) &\to (x-1, y, i) \text{ with rate } \mu_{k1} p_{10} \\
(x, y, i) &\to (x, y-1, i) \text{ with rate } \mu_{k2} p_{20}
\end{aligned}
$$

thus we can find a $\rho \geq \max_{i,k}(\mu_{k1} + \mu_{k2} + (A_{b,d})_i)$ where $(A_{b,d})_i = \sum_{i \neq j}(A_{b,d})_{i,j}$ to uniformise the process $\mathfrak{A}$ giving the equivalent discrete time process Ξ by observing the process at all stages of arrival, null events and departures or subsequent reentries.

We will look at the discrete time queueing process Ξ in the following way. For each state i of the arrival process $\mathfrak{A}$ governed by $\mathbf{A}$ there is a $\mathbf{Z}_0^2$ plane so that the state space of the queueing process is $\mathbf{Z}_0^2 \times \{1, \ldots, m\}$ as depicted in Figure

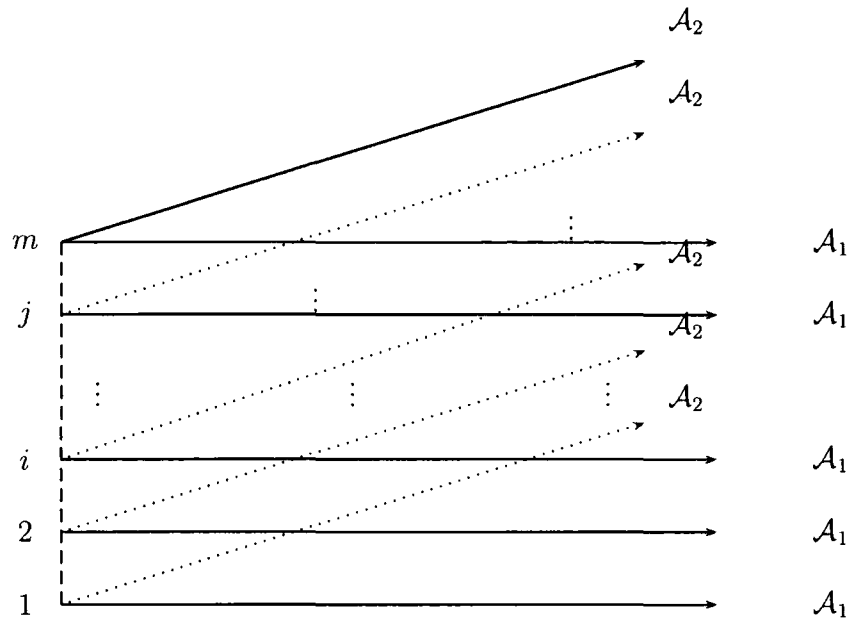


Figure 3.1: State space $\mathbf{Z}_0^2 \times \{1, \ldots, m\}$ of the queueing process with discrete time Markovian arrivals.

3.1. To define the mean drifts with respect to the queue length in direction of $\alpha = (x, y) \in \mathbf{Z}_0^2$ we consider the stationary distribution of the arrival process $\mathfrak{A}$ given by θ. Unfortunately we have no control over the arrival process and it is not yet clear to me how to get nice results for this model, I am however working on it.

Two queues in tandem with blocking We have seen an example of two queues in tandem in Section 2.5. Here we will discuss a variation similar to the model analysed in van Houdt and Alfa [38]. The first queue has a Markovian arrival process $\mathfrak{B}$ (as above) with stationary distribution δ and arrival rate Λ. Customers are served in the order in which they join the queue, we assume for now that the service time is exponentially distributed with rate μ_1 at queue 1 and rate μ_2 at queue 2. The buffer at queue 2 is finite, so that at any time only c customers can wait at queue 2. So whenever there are c customers waiting, a customer that has completed service at queue 1 has to wait there, preventing any other customers being served by server 1. This reduces the state space that the queue length process lives on to $\mathbf{Z}_0 \times \{0, \ldots, c\}$, where queue 2 has only $c+1$ states and the queue length changes of queue 1 happen on $\mathbf{Z}_0$, see Figure 3.2 for an example.

The continuous time queueing process $\mathfrak{X}$ with states (x, κ, i) (corresponding to number of customers in queue 1, queue 2 and stages of arrival respectively) has non-zero jumps at rates

$$(x, \kappa, i) \to (x+1, \kappa, i) \text{ with rate } (B_1)_{ij}$$
$$(x, \kappa, i) \to (x, \kappa, j) \text{ with rate } (B_0)_{i,j}$$
$$(x, \kappa, i) \to (x-1, \kappa+1, i) \text{ with rate } \mu_1$$
$$(x, \kappa, i) \to (x, \kappa-1, i) \text{ with rate } \mu_2$$

We choose $\rho \geq \max_i(\mu_1 + \mu_2 + (B_b)_i)$ where $(B_b)_i = \sum_{i \neq j}(B_b)_{i,j}$ and get the discrete time queueing process Ξ by observing $\mathfrak{X}$ at the times of events as before.

The discrete time queueing process Ξ lives on $\mathbf{Z}_0 \times \{0, \ldots, c\} \times \{1, \ldots, m\}$. The stability or not of a process on such a state space follows from the result about the half-strip model in Fayolle, Malyshev and Menshikov [11] given earlier.

For the model with two queues in tandem, Markovian arrival process and blocking in the second queue we have states $(x, \kappa, i) \in \mathbf{Z}_0 \times \{0, \ldots, c\} \times \{1, \ldots, m\}$, so it is not a *half-strip* as such but more a *bundle* which can be treated in a very similar fashion. By definition the Markov chain on $\kappa = \{0, \ldots, c\}$ is irreducible and aperi-

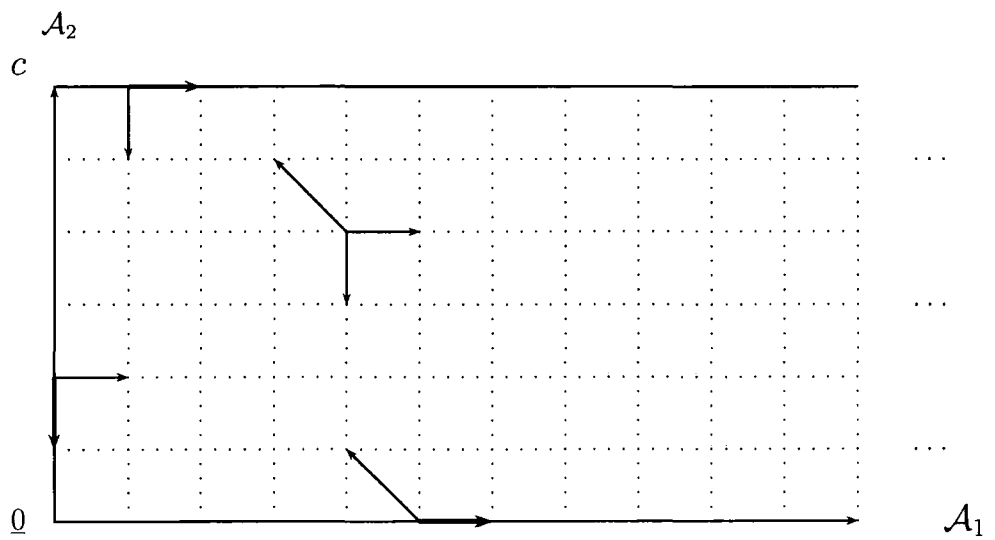


Figure 3.2: Queue length process state space $\mathbf{Z}_0 \times \{0, \ldots, c\}$ with jumps, for two queues in tandem with blocking at the second queue.

odic and has a stationary distribution which we denote by ν. Given Lemma 3.0.1 we know that the stationary distribution π of Ξ fulfils $\pi(\kappa, i) = \nu(\kappa)\delta(i)$. We also have

$$M(\kappa, i) = \Lambda_i - \mu_1 \mathbf{1}_{\kappa<c}$$

where Λ_i denotes the arrival rate at a given state i and $\mathbf{1}$ is the indicator function. We can see that

$$M = \sum_{\kappa,i} \nu(\kappa)\delta(i)\Lambda_i - \mu_1 \mathbf{1}_{\kappa<c} c = \sum_i \delta(i)(\Lambda_i - \mu_1(1 - \nu(c)).$$

The process is stable if $M < 0$ which corresponds to the results of van Houdt and Alfa [38].

Note the introducing the phase type service that van Houdt and Alfa [38] assume for the service at both queues would add an extra finite two dimensional state space $\mathcal{L}^2$, we will refrain from introducing this extra complication here.

Chapter 4

Stability criteria for $N \geq 3$ queues

Given our results for two queues it is natural to ask which of these are still valid if we have three or more queues, when the queue length process lives in $\mathbf{Z}_0^N$ with $N \geq 3$. In this chapter we will look at the discrete time queue length process Ξ on $\mathbf{Z}_0^N$ for N queue models and state stability criteria under different control policies or levels of control. We will also look at some examples such as the (generalised) Lu-Kumar network as introduced in Niño-Mora and Glazebrook [32], the generalised constraint queueing system in Tassiulas and Bhattacharya [37] and networks with customers that require simultaneous service at several service stations. The results given in Theorem 4.2.1 to Corollary 4.3.3 and the proofs are published in [27], the results on the Lu-Kumar network are published in [28].

There are some aspects of positive recurrence which are much more challenging in $N \geq 3$ dimensions than in two, we try to give the reader some idea why. In Section 2.4 we introduced the notion of second vector fields. In $\mathbf{Z}_0^2$ these are scalars and their signs, which determine whether the second vector field is ingoing or outgoing are easily obtained by comparing angles of the mean drifts as in Lemma 2.3.7. This rather nice idea does not work that simply for $N \geq 3$. Imagine one of the three axes in $\mathbf{Z}_0^3$, say $\mathcal{A}_1 \equiv \{\alpha = (x_1, x_2, x_3) \in \mathbf{Z}_0^3 : x_1 > 0, x_2 = 0, x_3 = 0\}$. On this axis we have a reflexion M' with at most one negative component - the x_1 component if the service at queue 1 is faster that the rate of input. We define the two mean drift vectors $M^{2'}$ and $M^{3'}$ as the reflexion from the two two dimensional planes $\mathbf{Z}^2_{(x_2=0)} \equiv \{\alpha \in \mathbf{Z}_0^3 : x_1 > 0, x_2 = 0, x_3 > 0\}$ and $\mathbf{Z}^2_{(x_3=0)} \equiv \{\alpha \in \mathbf{Z}_0^3 : x_1 > 0, x_2 > 0, x_3 = 0\}$

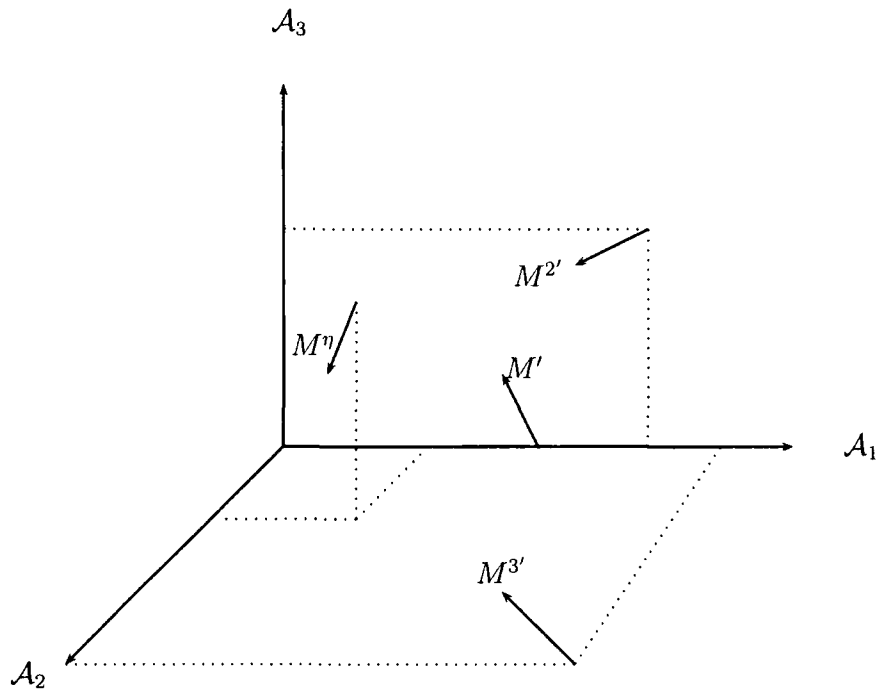


Figure 4.1: Mean drifts around $\mathcal{A}_1$ in $\mathbf{Z}_0^3$.

respectively. And finally there is one mean drift vector M^{η} under regime η on states $\alpha \in \mathbf{Z}_+^3$. See Figure 4.1 for an example of these mean drifts. Around the axis $\mathcal{A}_1$ all four jump distributions and not only the mean drift vectors are important to determine the second vector field; similarly if we introduce blocks and change between regimes we will have to consider second vector fields at the boundaries of these blocks.

We have seen some results for induced Markov chains in Chapter 3 when parts of the state space are finite. The results for second vector fields in higher dimensions with countable state space are given in FMM [11] Chapter 4, we will give the basic idea of this method here. If the mean drift in the interior M^{η} has $M_i^{\eta} < 0$ for $i = 1$, 2, 3 then we know that for example $\mathbf{Z}^2_{(x_3=0)}$ is an ergodic face (in the terminology of FMM [11]). We can calculate the stationary distribution of the birth and death process associated with the jumps in x_3 direction, we could have $p_{\alpha\alpha+1} = \rho^{-1}\lambda_3$, $p_{\alpha\alpha-1} = \rho^{-1}\mu_{k3}$, $p_{\alpha\alpha} = \rho - (\lambda_3 + \mu_{k3})$ and $p_{01} = \rho^1\lambda_3$ (given a queueing system with dedicated arrival streams only and no feedback and service configuration k, with ρ as the uniformising constant). Lets denote the stationary distribution of this one

dimensional process by π_0. If we are far away from the boundaries $\mathcal{A}_1$ and $\mathcal{A}_2$ on the plane $\mathbf{Z}^2_{(x_3=0)}$ we know that the mean drifts that influences the second vector field are M^η and $M^{3'}$. We can now calculate $M_{x_3=0} = (1-\pi_0)M^\eta + \pi_0 M^{3'}$ which is a drift vector with zero x_3 component, if $M_{x_3=0} < 0$ then we say the second vector field is ingoing. In order to avoid long calculations of second vector fields on all ergodic faces for different management regimes we will show here how, given some control, we can drive the process away from the boundaries so that the faces are transient.

We start be defining the model parameters of N queues and repeat some assumption made in Chapter 2. We will define the discrete time queue length process and state the mean drift vectors under η. We will see that the results for the clean cut cases such as C1 and C2 are straight forward, while C3 requires controls that keep the process away from the boundary; we will consider block randomised and block pure policies in this case.

4.1 Model parameters

Queueing models with $N \geq 3$ queues require some modifications of the parameters introduced in Section 2.1 we will state these here.

Arrival streams Let $\mathcal{P}_N$ denote the collection of non-empty subsets, Q, of the queues. Each such Q has an independent Poisson arrival stream with rate λ_Q and upon arrival each job is routed to a queue $i \in Q$. Any rule for doing this is a *routing scheme*

$$s : \mathcal{P}_N \to \{1, \ldots, N\} \text{ such that } s(Q) = i \in Q \text{ for all } Q \in \mathcal{P}_N.$$

We distinguish two cases: (1) we permit $\lambda_Q = 0$ for some sets Q but assume that no queue has zero arrival rate under all routing schemes, so that $\sum_{Q:s(Q)=i} \lambda_Q > 0$. (2) we will look at networks with re-entrant lines where $\lambda_Q > 0$ for at least one $i \in Q$ and $\sum_{Q:s(Q)=i} \lambda_Q = 0$ is allowed.

Service times We assume all jobs have service times that depend upon their queue and the service scheme in force while they are being served. Specifically,

under server configuration k, at most one job is in service at each non-empty queue and all jobs in queue i have independent, exponentially distributed service times with parameter μ_{ki}, $i = 1, 2, \ldots, N$.

The μ_{ki} may take any non-negative values so they may vary with k for any queue i but we allow only *efficient* server configurations (Assumption A1: whenever the queues in some set Q are all empty we only permit configurations k where $\mu_{ki} = 0$ for each $i \in Q$). We do allow the use of configurations with $\mu_{ki} = 0$ at states where queue i is not empty.

Regimes and switching As before the set $\mathcal{R}$ of overall *management regimes* is a finite collection of pairs $\eta = (k, s)$ and Assumption A2 (zero switchover time when changing from one η to another) applies.

Feedback For $N \geq 3$ any job that completes service at queue i under regime η independently enters queue j with probability p_{ij}^{η}, $j = 1, \ldots, N$ or leaves the system with probability $p_{i0}^{\eta} \equiv 1 - \sum_{j=1}^{N} p_{ij}^{\eta} \geq 0$.

Uniformising We *uniformise* the continuous time jump process, following Serfozo [34], by choosing a constant

$$\rho \geq \max_{k} \{ \sum_{Q} \lambda_Q + \sum_{i} \mu_{ki} \} \tag{4.1}$$

and introducing a *null* or *bell* event which has exponential inter-event times with rate $\rho - (\sum_Q \lambda_Q + \sum_i \mu_{ki})$ at any given queue lengths when regime $\eta = (k, s)$ is used.

We consider the uniformised discrete time process Ξ on state space $\mathbf{Z}_0^N \equiv \{(x_1, \ldots, x_N) \in \mathbf{Z}^N : x_i \geq 0,\ i = 1, \ldots, N\}$, obtained by observing the queue lengths at all null events, arrival times of new jobs, at service completions and consequent re-entry to queues. The jumps are of the form $\pm e_i$ and $e_i - e_j$ where e_i denotes the unit vector in i^{th} coordinate direction, so they are bounded in L_∞-norm, by 1. We will use $\alpha = (x_1, \ldots, x_N) \in \mathbf{Z}_0^N$ to denote a typical state vector for Ξ.

Control We define a policy for controlling this discrete event system as a sequence $\Pi = \{\pi_n : n \geq 0\}$ of probability distributions π_n, as in Section 2.2.2. We consider non-stationary, non-Markov *fully randomised* policies in Theorems 4.2.1 and 4.2.2. Then *stationary randomised* policies are considered in Theorem 4.3.2 and Corollary

4.3.3 where we investigates the use of *block randomised* policies Π^r. Some special cases when *block pure* policies Π^b can be used for N queue models are given in Section 4.5.

The blocks we consider are denoted by $\mathcal{B}$, one can still think of them as higher dimensional cones, but we will also consider different ways to define blocks. The basic blocks in three dimensions are given in the example below.

Example 4.1.1 (Block controls for $N = 3$) Consider a system with three queues. It has at least seven blocks of interest, these are linked to the collection of seven non-empty subsets $Q \in \mathcal{P}_3$ $(\{1\},\ldots,\{1,2\},\ldots\{2,3\},)$. Let $Q' = \{1,2,3\} \setminus Q$ and let $\mathcal{B}_Q = \{\alpha = (x_1,x_2,x_3) \in \mathbf{Z}_0^3 : x_i > 0, i \in Q; x_i = 0, i \in Q'\}$ so $\mathcal{B}_Q$ contains those states where precisely the queues in set Q have customers. All states except (0, 0, 0) lie in exactly one of these blocks. We may choose to partition these further but dealing with these blocks for now, the list of regimes η usable on block $\mathcal{B}_Q$ may be constrained by our efficient server assumption A1. At any visit to $\alpha \in \mathcal{B}_Q$: any block randomised policy chooses at random a regime from this list using the same distribution every time; a block pure policy chooses the same regime η every time.

With respect to the routable or fixed arrival rate we repeat the two assumptions made earlier

Assumption A3 (Boundary Reflexion Condition): If $\lambda_Q = 0$ for some sets Q but no queue has zero arrival rate under all routing schemes s, so that $\sum_{Q:s(Q)=i} \lambda_Q > 0$, then the process (Ξ, Π) can be reflected off the boundary simply by changing the routing scheme s.

We also assume that if we cannot route arrivals to all queues that assumption A4 given in Section 2.2 holds.

Assumption A4 (Boundary Sojourn Condition): If for for some non-empty subsets of queue, $Q \in \mathcal{P}_N$ we have

$$\sum_{Q:s(Q)=i} \lambda_Q = 0$$

then we require the following condition. For each $\alpha \in \partial\mathbf{Z}_0^N$ let $\tau \equiv \min\{n \geq 1 : \xi(0) = \alpha, \xi(n) \in \mathbf{Z}_+^N\}$ denote the length of the boundary sojourn. We assume there

exists a constant $v > 0$ such that for any $\alpha \in \partial \mathbf{Z}_0^N$ there is a policy Π_α such that

$$\mathbf{E}(\tau \mid \xi(0) = \alpha, \Pi_\alpha) < v \tag{4.2}$$

Mean drifts For routing policies s and service regimes $k = 1, \ldots, K$ at states $\alpha \in \mathbf{Z}_+^N \equiv \{(x_1, \ldots, x_N) \in \mathbf{Z}^N : x_i > 0, i = 1, \ldots, N\}$ we have

$$M_i^\eta = \rho^{-1} \left(\sum_{S:s(S)=i} \lambda_S + \sum_{j=1}^{N} \mu_{kj} p_{ji}^\eta - \mu_{ki} \right), \quad \eta = (k, s) \tag{4.3}$$

for queues $i = 1, 2, \ldots, N$ with ρ as defined in (4.1). Under the efficient service assumption (A1) equation (4.3) is also correct for histories leading to states $\alpha \in \partial \mathbf{Z}_0^N \equiv \mathbf{Z}_0^N \setminus \mathbf{Z}_+^N$ i.e. where at least one queue is empty.

The convex hull of the regime mean drifts is given by

$$\mathcal{M} = \left\{ \sum_\eta p_\eta M^\eta \in \mathbf{R}^N : p_\eta \in [0, 1] \text{ and } \sum_\eta p_\eta = 1 \right\} \tag{4.4}$$

and, if it is non-empty, its N-dimensional interior is given as

$$\mathrm{Int}_N(\mathcal{M}) \equiv \{\alpha \in \mathcal{M} : B(\alpha, \epsilon) \subset \mathcal{M} \text{ for some } \epsilon > 0\}.$$

Given all the details above we can now state the results for the $N \geq 3$ queueing model, using the classification of the convex hull with respect to the origin, see Appendix A. As before we have for each v the hyperplane

$$L_v(\alpha) \equiv \{\beta \in \mathbf{R}^N : v^T(\beta - \alpha) = 0\} \tag{4.5}$$

through α with normal vector v, which either separates $\alpha + \mathcal{M}$ from the origin $\underline{0}$ or not.

4.2 Fully randomised controls

The following two results apply when even the most general policy Π is used to control the queueing system. They imply that in cases C1 and C2 the control policy used does not affect the stability or otherwise of the process.

Theorem 4.2.1 *If* $\underline{0} \notin \mathcal{M}$ *and there exists an* $\alpha \in \mathbf{Z}_+^N$ *and* $v \in \mathbf{R}^N$ *such that the hyperplane* $L_v(\alpha)$ *separates* $\alpha + \mathcal{M}$ *from the origin* $\underline{0}$ *then the process* (Ξ, Π) *is unstable for any policy* Π*, in the sense that the total number of queued jobs almost surely goes to* ∞ *linearly in time.*

Proof: Under the conditions of Theorem 4.2.1 there exists $\alpha \in \mathbf{Z}_0^N$ and $w \in \mathbf{R}^N$ with at least one positive component, such that $L_w(\alpha)$ separates $\alpha + \mathcal{M}$ from $\underline{0}$ and $w^{\mathsf{T}} M^\eta > 0$ for every regime η. As the number of regimes is finite, $\varepsilon = 1/2 \min_\eta w^{\mathsf{T}} M^\eta > 0$. In this case $S_n = w^{\mathsf{T}} \xi(n)$ for $n = 0, 1, \ldots$ satisfies

$$\mathbf{E}(S_{n+1} - S_n \mid \mathcal{H}_n, \pi_n = \eta) = w^{\mathsf{T}} M^\eta > \varepsilon$$

whatever policy Π is used. It follows from part (ii) of Theorem 1.2.1 that there exists a $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$\mathbf{P}(S_n < \delta_1 n \mid \mathcal{H}_0, \Pi) < Ce^{-\delta_2 n} \text{ for all } n \geq 0$$

so by Borel-Cantelli these events almost surely occur only finitely often i.e. the number of customers waiting in the queues goes to infinity at least linearly in time. In addition we have $\mathbf{P}(\tau = \infty) > 0$ so with positive probability the process makes no visits to $\{\alpha \in \mathbf{Z}_0^N : v^{\mathsf{T}} \alpha < D\}$ which for large D contains the region of $\mathbf{Z}_0^N$ around the origin $\underline{0}$. □

Theorem 4.2.2 *If* $\underline{0} \notin \mathcal{M}$ *and there is no* $\alpha \in \mathbf{Z}_+^N$*,* $v \in \mathbf{R}^N$ *such that* $L_v(\alpha)$ *separates* $\alpha + \mathcal{M}$ *from* $\underline{0}$ *then* (Ξ, Π) *is stable under every policy* Π*, in the sense that the mean time to reach a bounded set around* $\underline{0}$ *is finite from any state* α*.*

Proof: Under the conditions of Theorem 4.2.2 we can find $w \in \mathbf{R}_+^N$ such that $w^{\mathsf{T}} M^\eta < 0$ for every regime η. In this case $S_n = w^{\mathsf{T}} \xi(n)$ for $n = 0, 1, \ldots$ satisfies

$$\mathbf{E}(S_{n+1} - S_n \mid \mathcal{H}_n, \pi_n = \eta) = w^{\mathsf{T}} M^\eta < -\varepsilon$$

for some $\varepsilon > 0$ and for every η. Applying part (i) of Theorem 1.2.1 we see that $E(\tau \mid \Pi, S_0 > D) \leq S_0/\varepsilon < \infty$. Thus from any finite state the process reaches $\{\alpha \in \mathbf{Z}_0^N : w^{\mathsf{T}} \alpha < D\}$ in finite time almost surely. □

Example 4.2.1 (Fixed servers and no routable arrivals) Consider a system with fixed servers (with service rate that drops to 0 when their queues are empty), Poisson arrivals with rate vector λ, exponential service times with rate vector μ and Markov feedback according to substochastic matrix P. We can see that the boundary reflexion condition applies. Jackson's result [21] for such models says that the system is stable if $\nu < \mu$, where $\nu = (I - P)^{-1}\lambda$, but transient if $\nu_i > \mu_i$ for any i. In our notation, for the regime η where all servers have jobs to process and so $\rho M^\eta = \lambda - (I - P)\mu$ from (4.3), these conditions correspond to stability if $[(I - P)^{-1}M^\eta]_i < 0$ for each server i and transience if $[(I - P)^{-1}M^\eta]_i > 0$ for any i. Our results do apply to this system if we assume we can shut down servers at any time and our results are consistent with what is known as we now show.

The possible regimes are as follows. For each subset $Q \in \mathcal{P}_N$ of the queues our model has a regime η_Q in which all servers outside Q are idle; feedback uses probabilities from P in all regimes. The mean drift M^Q of Ξ under η_Q satisfies

$$\rho M^Q = \lambda - (I - P)\mu^Q, \quad \text{where } \mu_j^Q = \begin{cases} 0, & \text{for } j \notin Q, \\ \mu_j & \text{for } j \in Q \end{cases}$$

As $\rho(I - P)^{-1}M^Q = (I - P)^{-1}\lambda - \mu^Q$, where $(I - P)^{-1}$ is non-negative matrix and hence $(I - P)^{-1}\lambda > 0$ it follows that if $[(I - P)^{-1}M^Q]_i < 0$ (i.e. $(I - P)^{-1}\lambda_i < \mu_i^Q$) for each i then $\underline{0} \in \mathrm{Int}_N\big((I - P)^{-1}\mathcal{M}\big)$. As $(I - P)^{-1}$ is invertible this implies $\underline{0} \in \mathrm{Int}_N\mathcal{M}$ so the conditions for Theorem 4.3.2 apply and the system with multiple regimes can be controlled to ensure it is ergodic.

If on the other hand $[(I - P)^{-1}M^\eta]_i > 0$ for some queue i then there exists a vector $w > \underline{0}$ such that $w^\mathsf{T}(I - P)^{-1}M^Q > 0$. In this case

$$\rho w^\mathsf{T}(I - P)^{-1}M^Q = w^\mathsf{T}(I - P)^{-1}\lambda - w^\mathsf{T}\mu^Q > \rho w^\mathsf{T}(I - P)^{-1}M^\eta > 0$$

and hence $w^\mathsf{T}(I - P)^{-1}M > 0$ for all $M \in \mathcal{M}$. As $v^\mathsf{T} = w^\mathsf{T}(I - P)^{-1} > \underline{0}$ we have $v^\mathsf{T}M > 0$ for all $M \in \mathcal{M}$ and hence the conditions of Theorem 4.2.1 hold and the system is transient under any control scheme. This equivalence of conditions is natural as switching regimes here can only make sets of servers idle whether or not they have jobs to process.

For $N = 2$ queues we see directly that if either $M_i^\eta > 0$ then one of the conditions of Theorem 2.4.1(ii) holds and the system is transient while if both $M_i^\eta < 0$ we are in case $\mathbf{A}^-/\mathbf{A}^-$ and the system is stable. The criteria for transience differ from those in the previous paragraphs if the servers can assist each other. Our results are of course for Poisson arrivals and exponential service while for example, Baccelli and Foss [3] establish essentially the same stability criteria for a model with an ergodic arrival process.

4.3 Block randomised controls

In case C3 it does make a difference which policy is used for running the system. In fact we can show that block randomised policies Π^r with a finite number of blocks are adequate to ensure stability of the process. Under policies of this type the process (Ξ, Π^r) is Markov so we can now talk about ergodicity and transience.

Corollary 4.3.1 *Under the conditions of Theorem 4.2.2 if the policy* Π *is Markov then* (Ξ, Π) *is an ergodic Markov chain.*

Proof: If Π is Markov then (Ξ, Π) is an aperiodic Markov chain. By construction (Ξ, Π) is irreducible and so it is ergodic by Foster's criterion, see e.g. Theorem 2.2.3 of [11]. □

Note that under the conditions of Theorem 4.2.1 when Π is Markov, (Ξ, Π) is transient. For block randomised policies we get the following result:

Theorem 4.3.2 *If* $\underline{0} \in Int_N(\mathcal{M})$ *then there is a block randomised policy* Π^r *with a finite number of blocks such that the Markov chain* (Ξ, Π^r) *is ergodic.*

Using randomised policies means that we can choose any vector M^Π starting in α that lies in $\mathcal{M}(\alpha)$ as the mean drift under which we run the system. M^Π is then a mixture of the *pure* mean drifts M^η. The condition $\underline{0} \in \mathrm{Int}_N(\mathcal{M})$ is more restrictive than case C3. Let $\partial\mathcal{M} = \mathcal{M} \setminus \mathrm{Int}_N(\mathcal{M})$, then we can extend our results to the following

Corollary 4.3.3 *The result of Theorem 4.3.2 still holds under the conditions* $\underline{0} \in \partial\mathcal{M}$, $Int_N(\mathcal{M}) \neq \emptyset$ *and there exist no* $\alpha \in \mathbf{Z}_+^N$, $v \in \mathbf{R}^N$ *such that the hyperplane* $L_v(\alpha)$ *separates* $\underline{0}$ *from* $\alpha + Int_N(\mathcal{M})$.

These results resolve the stability problem when block randomised policies can be used. The proof of Theorem 4.3.2 and Corollary 4.3.3 can be found in the Appendix B and in [27]. We do however give the idea of the proof here.

Idea of the proof of Theorem 4.3.2: Given that $\underline{0} \in \mathrm{Int}_N(\mathcal{M})$ is true we know that there is a ball around $\underline{0}$ (denoted by $B(\underline{0}, \Delta)$) with distance $\Delta > 0$ to the origin which is also inside $\mathcal{M}$. Since we can use block randomised policies Π^r we are free to choose a mean drift vector M^{Π^r} that points into any direction from the origin that we like (at least within the ball $B(\underline{0}, \Delta)$ as we know for a fact that it is part of $\mathcal{M}$).

The only problem that can arise in this case is that the boundary reflexion vectors are not very good; remember that given our efficient service assumption A1 we have to run the system under the appropriate regimes until we are in the interior $\mathbf{Z}_+^N$ (see the case $\psi_1 + \psi_2 \leq \pi/2$ in the proof of Theorem 2.3.4 for a $N = 2$ example of this). To avoid any problems that might arise due to the reflexion vectors we use a similar idea for N queues as we did for two. We construct a large block $\mathcal{B}^a$ in the interior $\mathbf{Z}_+^N$ in which we run a mean drift M^a which has all negative components and we make sure that this block has a large enough distance from the boundary $\partial\mathbf{Z}_0^N$ so that we are unlikely to hit it. Around $\mathcal{B}^a$ we have a safety zone where we use mean drifts M^i such that the process drives away from what ever boundary it is closest to and into the interior block. The M^i are chosen is such a way that $\mathcal{M}(M^a, M^i)$ is in case C2 for all i, which ensures that there is no problem on the boundaries between the interior blocks. On the boundary of the state space we use the fact that we can route arrivals so that we can populate the empty queues, i.e. we use the regime which guarantees that the *boundary reflexion condition* holds. Thus we can make the boundary faces transient and use the Lyapunov function $f(\alpha) = \sum_1^N \alpha_i$, given we can estimate appropriate sequence of random times N_n, makes the process $Y_n = f(\xi(N_n))$ a supermartingale on $\mathbf{Z}_0^N$.

4.4 The generalised Lu-Kumar network

In this section we look at a queueing network with re-entrant lines, the Lu-Kumar network as introduced in [24] and the generalised Lu-Kumar network discussed in [32]. This two station, two server network with four queues and variations of it have received a lot of attention as we highlighted in Section 1.3. Here we demonstrate how a randomised control policy can be applied so that the network is stable.

The Lu-Kumar network has a re-entrant structure i.e. there is a non-empty subset $i \in Q$ of queues with $\sum_{Q:s(Q)=i} \lambda_Q = 0$ which means we require the *boundary sojourn condition* 4.2 to be true. Given this we will consider the (generalised) Lu-Kumar model which falls into the group of multi-class queueing networks, see Dai [6] for an early paper about the stability of such networks. The multi-class networks considered in [6] and here are those where jobs that require different service times (i.e. are of different classes) queue in separate queues. Our model is in some ways more general than the multi-class network, as the idea of service regimes η allows a more general treatment in service of jobs. Jobs that wait at queue i can receive different service depending on the length of queue i. The way our model is set up we do not care whether these changes is due to actual servers coming or going to queue i (where service rate possibly add up) or some much more abstract notion is applicable by which the service speed is just changed with the regimes η.

We will describe one example of a multi-class queueing network given by Niño-Mora and Glazebrook [32]; differences in notation to the original paper are due to the attempt of keeping our notation.

The model considered is an open multi-class queueing network with N queues, where each queue contains a single class of customers, and K single server stations. Each station $\sigma \in \mathcal{S} = \{1, \ldots, S\}$ provides service for a constituency $C_\sigma \subseteq \mathcal{N} \equiv \{1, \ldots, N\}$ with $C_1, \ldots, C_S$ forming a partition of $\mathcal{N}$. Let $\sigma = \sigma(i)$ denote the station containing queue i. Customers of class i arrive at queue i as a Poisson stream with rate λ_i independently of other arrivals and of server location, note that $\lambda_i = 0$ is possible for some classes i. Class i jobs have exponential service times with parameter μ_i but only receive service when server $\sigma(i)$ is at queue i. Once its service

is completed an i-customer is routed back into the network as a j-customer with probability p_{ij} or leaves the network with probability $p_{i0} = 1 - \sum_{j \in \mathcal{N}} p_{ij}$. There are several results in the literature stated in terms of the *total arrival rate* of customers at queue i which is defined by the traffic equation $\tilde{\lambda}_i = \lambda_i + \sum_{j \in \mathcal{N}} \tilde{\lambda}_j p_{ji}$ for $i \in \mathcal{N}$ but we do not make any particular use of this.

To simplify comparison of the queueing process given in Niño-Mora and Glazebrook [32] with our model (in Section 4.1 and [27]) we translate their model into our language. In our model the service time and routing probabilities depend upon the current management regime $\eta = (k, s)$. For the model in [32] this corresponds to allocating server σ to queues i in constituency C_σ which gives the service configurations k, but the routing scheme for arrivals, s, is fixed. Under management regime η we change the service times as follow

$$\mu_{ki} = \begin{cases} \mu_i, & \text{if queue } i \text{ receives service under } \eta = (k, s) \\ 0, & \text{otherwise} \end{cases}$$

with $\mu_{ki} = 0$ for the special cases where there are no customers in C_σ and server σ is idle.

Given this translation we now look at the generalised Lu-Kumar network in more detail.

NIÑO-MORA AND GLAZEBROOK NETWORK This network was introduced in [32]. Like the standard Lu-Kumar network this one has two service stations S_1 and S_2 with one server each. There are two external arrival streams with rates $\lambda_1 = 1$ and $\lambda_2 = 1 - q$ where $0 \leq q \leq 1$. The service rate at queue i is μ_i, $i = 1, \ldots, 4$. The routing probabilities are $p_{12} = q$, $p_{14} = 1 - q$, $p_{23} = 1$, $p_{34} = q$, $p_{30} = 1 - q$ and $p_{40} = 1$ which gives a total arrival rate of 1 at each queue, see Figure 4.2. The case $q = 1$ corresponds to the original Lu-Kumar network in [24].

We assume throughout that the servers can switch between the two queues at their station at any given queue length without time delays.

This generalised Lu-Kumar model has eight service regimes: $A = 12$, $B = 13$, $C = 42$ and $D = 43$ (service provided at two queues); 1, 2, 3 and 4 under which only one queue gets served. For convenience we introduce an idling regime 0 where

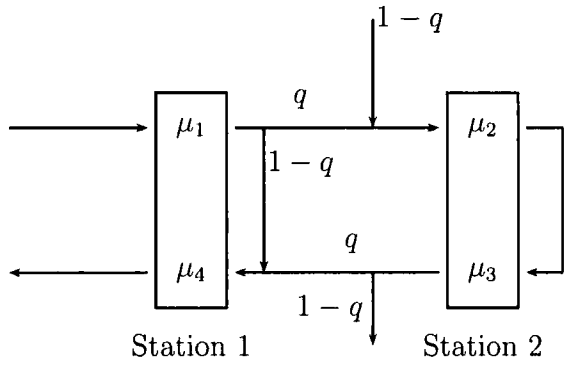


Figure 4.2: The generalised Lu-Kumar network

no queue is served. The mean drifts are

$$
\begin{aligned}
\rho^* M^A &= (\ 1-\mu_1,\ 1-q+q\mu_1-\mu_2,\ \mu_2,\ (1-q)\mu_1\ ) \\
\rho^* M^B &= (\ 1-\mu_1,\ 1-q+q\mu_1,\ -\mu_3,\ (1-q)\mu_1+q\mu_3\ ) \\
\rho^* M^C &= (\ 1,\ 1-q-\mu_2,\ \mu_2,\ -\mu_4\ ) \\
\rho^* M^D &= (\ 1,\ 1-q,\ -\mu_3,\ q\mu_3-\mu_4\ ) \\
\rho^* M^1 &= (\ 1-\mu_1,\ 1-q+q\mu_1,\ 0,\ (1-q)\mu_1\ ) \\
\rho^* M^2 &= (\ 1,\ 1-q-\mu_2,\ \mu_2,\ 0\ ) \\
\rho^* M^3 &= (\ 1,\ 1-q,\ -\mu_3,\ q\mu_3\ ) \\
\rho^* M^4 &= (\ 1,\ 1-q,\ 0,\ -\mu_4\ ) \\
\rho^* M^0 &= (\ 1,\ 1-q,\ 0,\ 0\ )
\end{aligned}
$$

where ρ^* denotes the uniformising constant in (4.1).

It is necessary to show that this model satisfies our boundary sojourn condition. Consider any $\alpha \in \partial\mathbf{Z}_0^N$. For the model with $q < 1$ the sequence of jumps e_1, e_1, e_2, e_2 (in any order) followed by $e_4 - e_1$, $e_3 - e_2$ (either way around) leads from α to $\alpha + e_1 + e_2 + e_3 + e_4 \in \mathbf{Z}_+^4$ and has strictly positive probability when the servers are at queues 1 and 2. As in the *two queues in tandem* example (Section 2.5) this implies that every sojourn time τ is stochastically smaller than a random variable Z with geometrically bounded tails and finite mean ζ say and so condition (4.2) is satisfied.

For the case $q = 1$ we can ensure that the sequence of jumps e_1 (four times), $e_2 - e_1$ (three times), $e_3 - e_2$ (twice) and finally $e_4 - e_3$ has positive probability and then, as above, we can show that condition (4.2) is satisfied.

We now explain how our classification relates to known results. Let $\mathbf{M}$ be the matrix with rows which are the vectors M^η for the various regimes η. Let $S_9 \equiv \{p \in [0,1]^9 : \sum_\eta p_\eta \leq 1\}$ and let $\|\cdot\|_2$ denote the Euclidean distance. If $\underline{0} \in \mathcal{M}$ then there exists a $p = (p_A, \ldots, p_D, p_1, \ldots, p_4, p_0) \in S_9$ such that $\|p\mathbf{M}\|_2 = 0$. We have

$$\begin{aligned}\|p\mathbf{M}\|_2^2 &= \big(1-\mu_1(p_A+p_B+p_1)\big)^2 + \big(1-q+q\mu_1(p_A+p_B+p_1)-\mu_2(p_A+p_C+p_2)\big)^2 \\ &+ \big(\mu_2(p_A+p_C+p_2)-\mu_3(p_B+p_D+p_3)\big)^2 \\ &+ \big((1-q)\mu_1(p_A+p_B+p_1)+q\mu_3(p_B+p_D+p_3)-\mu_4(p_C+p_D+p_4)\big)^2\end{aligned}$$

To obtain $\|p\mathbf{M}\|_2 = 0$ we need $p_A+p_B+p_1 = \frac{1}{\mu_1}$, $p_A+p_C+p_2 = \frac{1}{\mu_2}$, $p_B+p_D+p_3 = \frac{1}{\mu_3}$ and $p_C+p_D+p_4 = \frac{1}{\mu_4}$ which imply

$$\rho_1 \equiv \frac{1}{\mu_1}+\frac{1}{\mu_4} = p_A+p_B+p_C+p_D+p_1+p_4 \leq 1$$

and

$$\rho_2 \equiv \frac{1}{\mu_2}+\frac{1}{\mu_3} = p_A+p_B+p_C+p_D+p_2+p_3 \leq 1$$

This means that if $\rho_i > 1$ for either $i = 1$ or 2 then $\underline{0} \notin \mathcal{M}$ and we will show below that the system is in case C1.

Suppose now that the system is in case C2 or C3. Then there exists a vector $v \in \mathcal{M}$ such that $v_i < 0$ for $i = 1, 2, 3, 4$. For any $v \in \mathcal{M}$ there exists $p \in S_9$ such that $pM = \sum_\eta p_\eta M^\eta = v$. As each $v_i < 0$ we get the following set of inequalities:

(i) $\sum_\eta p_\eta < \mu_1(p_A+p_B+p_1)$

(ii) $(1-q)\sum_\eta p_\eta + q\mu_1(p_A+p_B+p_1) < \mu_2(p_A+p_C+p_2)$

(iii) $\mu_2(p_A+p_C+p_2) < \mu_3(p_B+p_D+p_3)$

(iv) $(1-q)\mu_1(p_A+p_B+p_1)+q\mu_3(p_B+p_D+p_3) < \mu_4(p_C+p_D+p_4)$

From these inequalities it readily follows that

$$\rho_1 < \frac{p_A+p_B+p_C+p_D+p_1+p_4}{\sum_\eta p_\eta} \leq 1 \quad \text{and} \quad \rho_2 < \frac{p_A+p_B+p_C+p_D+p_2+p_3}{\sum_\eta p_\eta} \leq 1$$

These are the well known necessary conditions for stability for this system.

Next we assume these necessary conditions hold and establish that either case C2 or C3 holds so that by Theorems 4.2.2 and 4.3.2 the process (Ξ, Π^r) can be made ergodic by an appropriate control policy.

Suppose that $\rho_1 < \rho_2 < 1$. The vector

$$p = \left(\frac{1}{\rho_2\mu_1\mu_2}, \frac{1}{\rho_2\mu_1\mu_3}, \frac{1}{\rho_2\mu_4\mu_2}, \frac{1}{\rho_2\mu_4\mu_3}, 0, \frac{\rho_2-\rho_1}{\rho_2\mu_2}, \frac{\rho_2-\rho_1}{\rho_2\mu_3}, 0, 1-\rho_2 \right)$$

has $\sum_\eta p_\eta = 1$ with seven of the $p_\eta \in (0,1)$ and also $\sum_\eta p_\eta M^\eta = \underline{0}$. Further the 4×4 matrix which has rows M^A, M^B, M^C, M^3 has determinant $(1-\rho_2)\prod_1^4 \mu_i > 0$ so these four vectors are linearly independent and hence $\underline{0} \in \mathrm{Int}_4(\mathcal{M})$ so there is $v \in \mathcal{M}$ with $v_i < 0$. The case $\rho_2 < \rho_1 < 1$ is very similar – this time use

$$p = \left(\frac{1}{\rho_1\mu_1\mu_2}, \frac{1}{\rho_1\mu_1\mu_3}, \frac{1}{\rho_1\mu_4\mu_2}, \frac{1}{\rho_1\mu_4\mu_3}, \frac{\rho_1-\rho_2}{\rho_1\mu_1}, 0, 0, \frac{\rho_1-\rho_2}{\rho_1\mu_4}, 1-\rho_1 \right)$$

When $\rho_1 = \rho_2 < 1$ use the convex combination

$$p = \left(\frac{1}{\mu_1\mu_2}, \frac{1}{\mu_1\mu_3}, \frac{1}{\mu_4\mu_2}, \frac{1}{\mu_4\mu_3}, \frac{1-\rho_1}{\mu_1}, \frac{1-\rho_1}{\mu_2}, \frac{1-\rho_1}{\mu_3}, \frac{1-\rho_1}{\mu_4}, (1-\rho_1)^2 \right)$$

Now consider the transient case: if either $\rho_1 > 1$ or $\rho_2 > 1$ the system must be in case C1 since $\underline{0} \notin \mathcal{M}$ and $\mathcal{M} \cap \mathbf{R}^4_- = \emptyset$. Thus by Theorem 4.2.1 the system is unstable under any possible control policy. In some cases it is easy to find the normal vector w of a separating hyperplane L_w (4.5) such that $w^\mathsf{T} M^\eta > 0$ for all $M^\eta \in \mathcal{M}$.

Consider the following case, suppose $\rho_1 > 1$, $\rho_2 < 1$. If $q = 0$ (i.e. two independent tandem queues) we can use normal vector $w_T = (1, 0, 0, a)$; for $q = 1$ (the Lu-Kumar network) we we can use $w_L = (1, a, a, a)$, where $a = \frac{1}{\mu_4+\epsilon}$ where $\epsilon > 0$ is such that $a > 1 - \frac{1}{\mu_1}$. Given these $w^\mathsf{T} M^\eta > 0$ for all $M^\eta \in \mathcal{M}$ is true.

Finally there are the cases where $\rho_i \leq 1$ for $i = 1, 2$ with one or both $\rho_i = 1$. These are in case C4 which we have discussed for $N = 2$ cases only in Section 2.3.3.

Our results also apply to a generalised Niño-Mora and Glazebrook model as depicted in Figure 4.2 which has Poisson arrival streams with *arbitrary* rates and with feedback of completed jobs.

4.5 Block pure controls

Using block pure policies Π^p in for models with $N \geq 3$ queues makes the description of how the system is run more complicated. Remember the two queue model under

block policies, we have already seen that when block randomised policy are used only a maximum of four blocks is required while it is 5 for block pure policies. In FMM [11] Section 3.3 one can find the 15 possible cases for the three essential blocks of $\mathbf{Z}_0^2$, $\mathbf{Z}_+^2$, $\mathcal{A}_1$ and $\mathcal{A}_2$. When establishing the stability of the queueing network under the conditions of Theorem 2.3.4 we allow the number of blocks to change depending on the mean drifts and single out four general cases. For lower levels of control in Section 2.4, with four given blocks the number of all possible cases is 27 and this is when we have only two queues.

From Example 4.1.1 we can see that there are seven essential blocks in $\mathbf{Z}_0^3$ (three axis, three two dimensional faces and the interior). Assuming that we would like to run the system with these seven blocks only and given that we know that the mean drift M^η that is run in the interior has components $i = 1, 2, 3$ with $M_i^\eta < 0$ we still need to check the interior mean drift against all the boundary ones. Loosely speaking we can have the following: all second vector fields are ergodic and ingoing, one of the 6 vector fields is ergodic but outgoing, there are 15 combinations in which two boundary faces can be outgoing, and so on. In the latter cases we will need more than seven blocks and think in detail about how to run the system for each case. Since we could see already how many cases there can be for $\mathbf{Z}_0^3$ we will try to avoid this possible but not practical splitting into cases for $N \geq 3$ and rather highlight some more general straight forward cases where block pure policies can be used.

Consider case C3, where $\underline{0} \in \mathrm{Int}_N(\mathcal{M})$. The convex hull might fall into C3 and the following situation is given: there is one mean drift vector M^a which has components $M_i^a < 0$ for all $i = 1, \ldots, N$ while the convex hull of the remaining mean drifts $\mathcal{M} \setminus M^a$ falls into case C1 (i.e. is separable from the origin). From the proof of Theorem 4.3.2 we know that having only one mean drift vector with all negative components is not sufficient to establish the positive recurrence as the boundary behaviour might make the process unstable.

We can guarantee stability if there are mean drifts M^a and M^i as given in the proof of Theorem 4.3.2 (see Appendix B). However there is another condition which together with a change in control strategy can leads to a stable system.

As we have seen in Example 4.2.1 and for the Lu-Kumar network our Section 4.4 definition of regimes includes the following: For each subset $Q \in \mathcal{P}_N$ of the queues our model has a regime η_Q in which all servers outside Q are idle. The mean drifts are denoted as M^Q of Ξ under η_Q and the service rate is given by

$$\mu_j^Q = \begin{cases} 0, & \text{for } j \notin Q, \\ \mu_j & \text{for } j \in Q \end{cases}.$$

This is a very natural assumption to make and, as we have seen in previous examples, it is the case for lot of models present in the literature (see for example the Jackson and Lu-Kumar network). If we know that such regimes η_Q exist then a case like case CIII in the proof of Theorem 2.3.4 cannot happen and neither can the situation given above where $\mathcal{M} \setminus M^a$ falls into case C1.

We can identify the two following sub cases of C3 which enable us to find block pure policies so that the N queue network is stable

Ci There exists at least one regime η_a with mean drift M^a such that $M_i^a < 0$ for all $i = 1, \ldots, N$ when all $x_i > 0$.

Cii There exists a v as in (4.5) with all $v_i > 0$ so that for any queue i

$$\exists\, \eta \in \mathcal{R} \equiv \{\eta : v^T M^\eta < 0, e_i^\mathsf{T} M^\eta < 0\} \tag{4.6}$$

Given the service rate as above, no feedback or routable arrivals, the first case Ci corresponds to the stable standard N queues in parallel denoted by $M/M/N$, with stability condition $\max_i \lambda_i/\mu_i < 1$ - the lengths of all queues is reduced while there are jobs present and the servers just idle when there are no jobs in the queues. If we introduce feedback it corresponds to Jackson's network [21]. We will not consider Ci in any more detail but concentrate on Cii.

The strategies we introduce here is to show stability in case Cii is what we call a *cyclic* or *greedy strategy*. We can formulate Cii as follows:

Cii For a system with N queues there are L regimes η_l with mean drifts M^l, $l = 1, 2, \ldots, L$ and $L \leq N$ such that Cii is true and

$$\mathcal{M}^L \equiv \left\{ \sum_{l=1}^{L} p_l M^l \in \mathbf{R}^N \ : \ p_l \in [0,1] \text{ and } \sum_l p_l = 1 \right\},$$

the convex hull of the M^l, $l = 1, 2, \ldots, L$ only, falls into case C2 (given in Appendix A). The M^l are such that we can reduce every queue i with at least one mean drift M^l. The extreme of this case is given by $L = N$ which means that there are N mean drifts and each of them has only one negative component $M_i^l < 0$ while all the other components M_j^l for $j \neq i$ are positive.

Considering the extreme case its similarity to a polling system becomes clear. Given that $L = N$ the model could be controlled in such a way that the queue length process Ξ on $\mathbf{Z}_0^N$ looks like a polling system where queue i is emptied first and then the policy changes to a regime under which queue j can be emptied, and so on. See Example 1.3.1 for a brief description of polling systems.

Different to polling systems, under our model assumptions service can be provided to more than one queue and under all regime η_l which satisfies Cii service might even be provided to all N queues but only queues $i \in Q \in \mathcal{P}_N$ are reduced in mean.

Cyclic strategy For Cii if $L \leq N$ consider the discrete time queue length process Ξ in state $\alpha \in \mathbf{Z}_+^N$. Assume that the system is run under a policy Π^p so that the mean drift M^1 which is in force at time n reduces the length of the set of queues Q_1. Given that started in a state $\alpha \in \mathcal{B}_a$ (an interior block defined in Appendix B) at time n we know that after a finite amount of time n_1 the process will first hit a state $\alpha \in \partial_{Q_1} \mathbf{Z}_0^N$ (the set of the boundary states where queues $i \in Q_1$ are empty). Once Ξ is in such a boundary state following the cyclic strategy the policy choose the mean drift from all M^l which reduces queues Q_l as follows.

Choosing a feasible regime Given that the process Ξ is in state $\xi(n) = \alpha \in \partial_{Q_l} \mathbf{Z}_0^N$ having run the system under regime η_l using the cyclic strategy, the feasible regime η_m is given by

$$m = \min_{\ell=\{1,2,\ldots L\}} \{m = l + \ell : \xi(n) = \alpha \notin \partial_{Q_m} \mathbf{Z}_0^N\}.$$

This means, if we check all regimes η_{l+1}, η_{l+2} ... then $\eta_{m=l+\ell}$ is the first regime (following the cyclic order) under which we reduce queues i so that all $x_i > 0$ and

$\xi(n) = \alpha = (x_1, \ldots, x_N) \in \partial_{Q_l} \mathbf{Z}_0^N$ for $i \in Q_m$. We assume that selecting a feasible regime can be done instantaneously.

Corollary 4.5.1 *Under Cii given a cyclic strategy which chooses feasible regimes η the Markov chain (Ξ, Π^p) is positive recurrent.*

Proof: Since the condition Cii requires that there exists a v as in (4.5) with all $v_i > 0$ so that for any queue i $\exists$ $\eta \in \mathcal{R} \equiv \{\eta : v^\top M^\eta < 0\}$ with $M_i^\eta < 0$ we know from the proof of Theorem 4.2.2 combined with Corollary 4.3.1 that setting $S_n = v^\top \xi(n)$ in Theorem 1.2.1 the process in stable for all $\alpha \in \mathbf{Z}_+^N$.

On the boundary we know that once we hit a state $\alpha \in \partial_{Q_l} \mathbf{Z}_0^N$ under η_l we next change to regime η_{l+1}. This means that if the process Ξ is in $\xi(n) = \alpha \in \partial_{Q_l} \mathbf{Z}_0^N$ at time n we run the system under a new mean drift M^{l+1} which reduces queues $i \in Q_{l+1}$ if $Q_l \cap Q_{l+1} = \emptyset$, then we have

$$\mathbf{E}(S_{n+1} - S_n \mid \mathcal{H}_n, \xi(n) = \alpha) = w^\top M^\eta < -\varepsilon \text{ for } \alpha \in \mathbf{Z}_+^N \cap \partial_{Q_l} \mathbf{Z}_0^N$$

If $Q_l \cap Q_{l+1} \neq \emptyset$ we can have the following two possibilities (a) the boundary state is $\alpha \notin \partial_{Q_{l+1}} \mathbf{Z}_0^N$, i.e. it is a state of the boundary of the set Q_l but not Q_{l+1} in which case we run M^{l+1} until we hit $\partial_{Q_{l+1}} \mathbf{Z}_0^N$ and start again; or (b) $\xi(n) = \alpha \in \partial_{Q_l} \mathbf{Z}_0^N \cap \partial_{Q_{l+1}} \mathbf{Z}_0^N$ in which case the regime chose is η_m such that the feasibility condition above is fulfilled.

Since we have assumed that switching regimes and choosing a feasible regime η_m does not take any time we know that $\mathbf{E}(S_{n+1} - S_n \mid \mathcal{H}_n, \xi(n) = \alpha) = w^\top M^\eta < -\varepsilon$ is true for all $\alpha \in \mathbf{Z}_0^N$. □

Greedy strategy This strategy works as follows: when Ξ is in a boundary state $\xi(n) = \alpha \in \partial_{Q_l} \mathbf{Z}_0^N$ after being run under regime η_b applying the greedy strategy mean we choose regime η_c with mean drift M^c, such that

(1) M^c reduces the set of queues Q_c which contains the longest queue i, with $i = \max\{j : x_j = \max_{\ell=1,\ldots N}(x_\ell)\}$ given that $\xi(n) = \alpha \in \partial_{Q_l} \mathbf{Z}_0^N$, and

(2) M^c is feasible if $\xi(n) = \alpha \notin \partial_{Q_m} \mathbf{Z}_0^N$. If M^c is not feasible, and there is no other regime η_d which reduces queue i and is feasible, then the greedy strategy

selects a regime η_e which reduces (in mean) the next longest queue j for which $\xi(n) = \alpha \notin \partial_{Q_e} \mathbf{Z}_0^N$.

Corollary 4.5.2 *Under Cii and the greedy strategy which selects feasible regimes η the Markov chain (Ξ, Π^p) is positive recurrent.*

Proof: Follows directly from the proof of Corollary 4.5.1. If we hit the boundary $\xi(n) = \alpha \in \partial_{Q_b} \mathbf{Z}_0^N$ under regime η_b given the greedy strategy regime η_c is selected which reduce $i \in Q_c$ where i is the longest queue at time n. As before we check if $Q_b \cap Q_c = \emptyset$ and run η_b if this is the case. Otherwise ($Q_b \cap Q_c \neq \emptyset$) there are two possibilities: (a) there is another η_d so that CII is true which reduces the longest queue i; or (b) we choose a regime η_e under which the second longest queue j is reduced. □

Note: Given our earlier assumption about the service configurations under η_Q we know that if regime η_b is not feasible after Ξ hits a state $\alpha \in \partial_{Q_a} \mathbf{Z}_0^N$ because $Q_a \cap Q_b = \{j\}$, then there is a regime η_c for which all components of M^{η_c} but j are identical to those of M^{η_b} and component j has in M^{η_c} has $\mu_j = 0$.

4.6 Further Examples

In this section we would like to discuss some more examples of queueing networks. The first one we look at is an example of a three queue system where one queue requires simultaneous processing by the two only service stations in the system. We show how our assumptions about management regimes can straightforwardly deal with models of this type. We also show how the three queue example can be controlled so that it is stable under a block pure strategy/policy introduced in the previous section. As another example for a model run under a block pure strategies can be found in a paper by Tassiulas and Bhattacharya [37]. We will state their model and results briefly and compare them to ours using an example.

4.6.1 Customers with simultaneous service requirements

Networks with customers which have simultaneous service (or station, or processor) requirements can be found for example in Dai and Lin [8] and Hansen, Reynolds and Zachary [19]. We will only consider one example to demonstrate how our methods can be applied. The example is the following three queue model. There are three Poisson arrival streams with rate λ_i $i = 1, 2, 3$ and two stations where customers are served with exponentially distributed service times with parameter μ_{ki} where k is the service configuration under regime $\eta = (k, s)$ and i indicates which queue i is served. Customers in queue 1 are served at station 1 and customer from queue 2 are served at station 2, while serving customers that are waiting in queue 3 requires the attention of both stations S_1 and S_2, see Figure 4.3. This means $\mu_{ki} = 0$ for $i = 1$ and 2 if $\mu_{k3} > 0$ and $\mu_{ki} > 0$ for $i = 1$ and/or 2 if $\mu_{k3} = 0$, leaving out the case where no queue receives service.

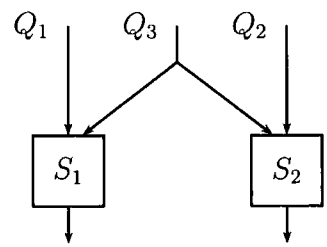


Figure 4.3: Three queue model where queue Q_3 requires the attention of both service stations S_1 and S_2.

We have regimes η_{12} with mean drifts

$$M^{12} = \rho^{-1}(\lambda_1 - \mu_{k1}, \lambda_2 - \mu_{k2}, \lambda_3)$$

when queues 1 and 2 receive service, regimes η_3 with mean drifts

$$M^3 = \rho^{-1}(\lambda_1, \lambda_2, \lambda_3 - \mu_{k3})$$

if queue 3 is served and regimes η_j with

$$M^1 = \rho^{-1}(\lambda_1 - \mu_{k1}, \lambda_2, \lambda_3) \text{ and } M^2 = \rho^{-1}(\lambda_1, \lambda_2 - \mu_2, \lambda_3)$$

when either queue 1 or queue 2 are served and $\rho \geq \max_\eta(\sum_i \lambda_i + \mu_{k1} + \mu_{k2}, \sum_i \lambda_i + \mu_{k3})$.

Let us consider the possible ways in which this three queue model can be stable. The convex hull $\mathcal{M}$ which is formed by the mean drifts above would have to be either in case C2 or C3. If C3 is true it is obvious that for this model to be stable condition Cii has to be true (as Ci is impossible). We will assume that we are in case Cii (a model with $\mathcal{M}$ that falls into C2 can be controlled to be stable in exactly the same way).

We can have the following situations: (i) there is at least one regime η_{12} and one regime η_3 such that there exists a $v \in \mathbf{R}^3_+$ for which $v^\top M^{12} < 0$ and $v^\top M^3 < 0$, (ii) there are three regimes η_i so that for $v \in \mathbf{R}^3_+$ we have $v^\top M^i < 0$ for $i = 1, 2, 3$, or (iii) both. In any of these cases we can apply the cyclic or the greedy strategy and run the system using *pure* (or non-randomised) regime mean drifts.

Given that we are in a state $\alpha \in \mathcal{B}_a \subset \mathbf{Z}^3_+$ away from the boundary and that the system is run under a regime η_3 we know that M^3 only reduces the length of queue 3 in mean while queue 1 and 2 have only arrivals. After some finite time we will hit the boundary $\partial_{Q_3}\mathbf{Z}^3_0$. We can see that $Q_3 \cap Q_{12} = Q_3 \cap Q_1 \cap Q_2 = \emptyset$ and if we have a regime η_{12} that falls into Cii we can run the system under this regime until Ξ hits $\xi(n) = \alpha \in \partial_{Q_{12}}\mathbf{Z}^3_0$ at which point it is feasible to select regime η_3.

In general for networks with customers which have simultaneous service requirements with $i = 1, \ldots, N$ queues and $j = 1, \ldots, J$ service stations, we know that if customers in queues $Q \in \mathcal{P}_N$ require the attention of the server at stations $S \in \mathcal{P}_J$ (where $\mathcal{P}_J$ is the collection of non-empty subsets S of stations j) then

$$\mu^Q_{ki} = \begin{cases} 0, & \text{if queue } i \notin Q \text{ but requires station(s) } S \text{ under } k \\ \mu_{ki}, & \text{if queue } i \in Q \text{ is served at station(s) } S \text{ under } k \end{cases}$$

Note that under our assumptions about service configurations k and regimes η it is also possible to model a queueing system where the simultaneous service requirements change from one regime to another.

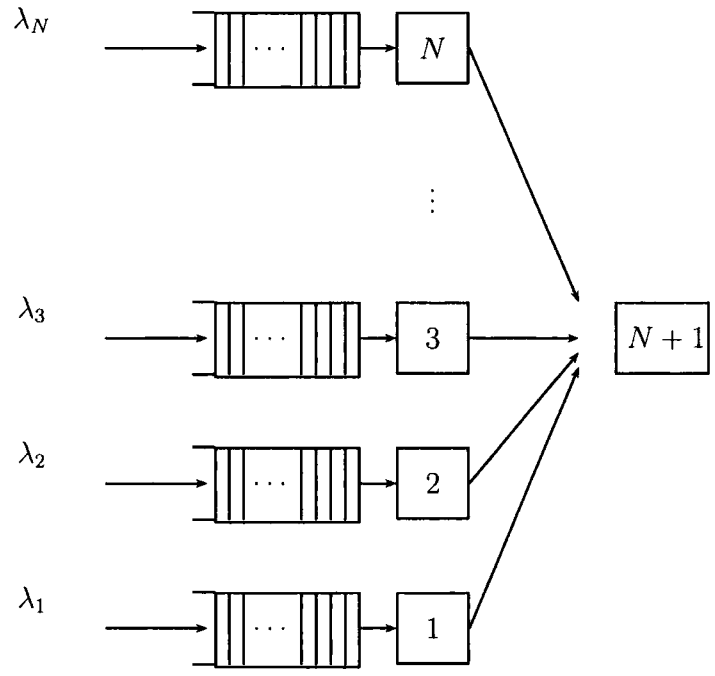


Figure 4.4: The $N+1$ nodes of the GCQS without feedback.

4.6.2 The generalised constrained queueing system (GCQS)

We will now discuss the model and results in Tassiulas and Bhattacharya [37], see also Example 1.3.4. Note that the notation used for describing the model is the same as in [37] and we state whenever any of our results and notations are used. The network has $N+1$ nodes where the $i = 1, \ldots, N$ nodes are stations or queues in our language, while the $N+1$-st node represents the outside world, customers that reach $N+1$ have left the system, these nodes are depicted in Figure 4.4. There is an independent Poisson arrival stream with parameter λ_i for each node i. The service time distribution is general and identical within each node i with mean β_i. There are $k = 1, \ldots, K$ servers which are allocated to nodes i where they serve at the required rate. After being served by server k at node i the customer might go to node j with probability $p(i,k,j)$. There is a $N \times K$ schedule matrix $\mathcal{U} = \{u_{ik}\}$ where $u_{ik} = 1$ means that server k can serve customers at node i, $u_{ik} = 0$ otherwise. Tassiulas and Bhattacharya's [37] aim is to find a *server allocation policy* that stabilises the network under the condition that only feasible schedules can be applied. A schedule is feasible if no more than K serves are needed, the set of feasible schedules includes schedules under which servers idle at empty queue.

Modelling the departures from the network through node $N+1$ is an essential part of Tassiulas and Bhattacharya's [37] analysis. Let ϕ_{ij} denote the flow from node i to node j, then the set of flows which satisfy what they call the *flow conservation equations* is given by

$$\mathcal{F} = \left\{ \phi \in \mathbf{R}^{N \times N+1} \ : \ \lambda_i \sum_{j=1}^{N} \phi_{ji} = \sum_{j=1}^{N+1} \phi_{ij}, \ i = 1, \ldots, N \right\}.$$

So $\mathcal{F}$ contains those flows for which the number of customers that are served at node i is equal to the number of customers that reach node $N+1$ for all $i = 1, \ldots, N$. Tassiulas and Bhattacharya [37] also construct a convex hull. The components of their convex hull $\mathcal{S}$ are matrices S^l with elements

$$S^l_{ij} = \frac{1}{\beta} \sum_{k=1}^{K} u^l_{ik} p(i,k,j), \quad i \leq N, \quad j \leq N+1$$

which gives the maximum flow matrices under a possible, feasible server schedule $u^l \in \mathcal{U}$. The convex hull is gives the flow achieved by mixing all schedules $u^l \in \mathcal{U}$ is

$$\mathcal{S} = \left\{ \sum_{l=1}^{L} c_l S^l \ : \ \sum_{l=1}^{L} c_l \leq 1, \ c_l \geq 0 \right\}.$$

Given these two sets $\mathcal{F}$ and $\mathcal{S}$ Tassiulas and Bhattacharya [37] state their first necessary result, saying that it is possible to allocate the servers so that the system is stable if $\mathcal{F} \cap \mathcal{S} \neq \emptyset$ (see Theorem 1 in [37]). For the necessary and sufficient stability results they construct a policy π^* which is rather similar to the *greedy strategy* described above. The policy π^* runs a server allocation schedule u^* for some time so that the longest queue in the network is reduced. τ_n denotes the sequence of times when the queue length's are checked and a new decision about the schedule u^* is made. Given this policy π^* Tassiulas and Bhattacharya [37] conclude that if $\mathcal{F} \cap \mathcal{S}^0 \neq \emptyset$ (where $\mathcal{S}^0$ is the interior of the convex hull) then π^* stabilises the GCQS (see Theorem 2 in [37]). Observing the process at times $\tan_n$ guarantees that the conditions of Theorem 1.2.1(i) hold.

We would now like to demonstrate how our convex hull $\mathcal{M}$ relates to the set of flows and the convex hull of Tassiulas and Bhattacharya [37] by means of a three queue version of the *parallel processing* example. The parallel processing network has

customers that are split into component i where each component requires a server and all components finish service at the same time. For the particular example we would like to consider this works as follows. There are $i = 1, 2, 3$ nodes, a Poisson arrival stream at rate λ_i and at node i the customer is split into i components which are processed by i servers simultaneously at the identical mean service rate β_i (i.e. all i components are finished at the same time and enter the next queue as one customer). The number of servers in the parallel processing network is $K = 6$, so that there can be i servers for each node i. There are seven feasible service schedules $u \in \mathcal{U}$, one where all nodes $i = 1, 2, 3$ receive service, three where two queues are served (queues 12, 13 and 23) and three where only one queue is served. After completing service at queue i a customer is fed into queue j with probability p_{ij} while we assume that $p_{ii} = 0$ for all i. The probability for customers leaving the system is given by $p_{i0} = 1 - \sum_j p_{ij}$.

The total traffic at node i is given by $\tilde{\lambda} = (I - P)^{-1}\lambda$ (as in Section 4.4) which in this example gives the unique solution to the flow equation in [37]. The convex hull $\mathcal{S}$ has seven matrices S^l with elements $S^l_{ij} = \frac{1}{\beta_i} p_{ij} \sum_k u^l_{ik}$ with $i \leq 3$ and $j \leq 3 + 1$ (for the fourth node that is used to model the exiting jobs). The necessary condition $\mathcal{F} \cap \mathcal{S} \neq \emptyset$ reduces to

$$\begin{pmatrix} \tilde{\lambda}_1 \\ \tilde{\lambda}_2 \\ \tilde{\lambda}_3 \end{pmatrix} \in \text{convex hull} \left[\begin{pmatrix} \frac{1}{\beta_1} \sum_k u_{1k} \\ \frac{1}{\beta_2} \sum_k u_{2k} \\ \frac{1}{\beta_3} \sum_k u_{3k} \end{pmatrix} : u \in \mathcal{U} \right]$$

We can adapt this example quite easily, the only major change is that we assume exponentially distributed service times. The arrival rates are given by λ_i for each i and we have exponentially distributed service times at rate μ_{ki}. There are seven service regimes $\eta = (k, s)$ as mentioned above. Under our modelling assumptions a customer at queue i is not split into components but just served under a specific regime. This works since a service configuration under our assumptions could mean there are in fact i servers working parallel but we do not care as all i components finish service at the same time and are fed into other queues as one customer and we look at the changes to queue length only. We get the following mean drift

components:

$$M_i^{\eta} = \frac{1}{\sum_i(\lambda_i + \mu_{ki})}(\lambda_i + \sum_j(\mu_{kj}p_{ij}) - \mu_{ki}), \text{ for } i \neq j.$$

Our stability criteria for this type of model are given in Example 4.2.1 of the Jackson Network.

Using the example above we can see that the necessary and sufficient stability criterion in [37] given by $\mathcal{F} \cap \mathcal{S}^0 \neq \emptyset$, is only true if $[(I - P)^{-1}\lambda]_i < \mu_i$ for all i, which corresponds exactly to our results for Example 4.2.1 in Section 4.2. Thus the result that $\mathcal{F} \cap \mathcal{S}^0 \neq \emptyset$ corresponds to our the convex hull $\mathcal{M}$ intersecting with the negative quadrant.

Chapter 5

Discussion

A very general queueing network that can be controlled at several levels has been introduced. The aim was to determine criteria for which control policies can be found, these control the routable arrivals or inputs, the service configurations and with it the feedback probabilities, so that the process (Ξ, Π), which is Markov if Π is stationary and Markov, is stable; or when the process is unstable, no matter what controls are applied.

The novelty of our approach is that we look at the mean drifts for each possible system configuration or under each possible management and then ask which control to apply so that the network can be stable. As we have demonstrated in Chapter 2 and more so in Chapter 4 this leads to straight forward conclusions about the stability of the queueing network as we do not limit our search by making prior assumptions about the control.

We do admit that limiting our analysis to queue length processes Ξ which are Markov is restrictive. Chapter 3 however gives an idea of how many different service time and arrival distributions are possible while still keeping the Markovian assumption. The main problem with Markovian arrival processes and phase type service is the formulation of the state space that the queueing process lives on and produce mean drift vectors that give information about the queue length changes.

One queueing model in the literature which is similar in terms of control but more general with respect to the inter arrival and service time distributions is the stochastic processing model given in Dai and Lin [8]. Their model has $N+1$ buffers or

nodes, K processors and J what they call activities. These activities are comparable to our management regimes η in that the service times of customers are determined by the activity which is active at a processor and not the processor. Each activity j can process customers at a number of processors. They distinguish one input buffer 0 from the N service buffers and a number of the processors are only input processor, there are also input activities. Under our assumptions we can model a queueing network which has only one external arrival stream to one queue (see Section 4.4) but the stream needs to be Poisson and the arrivals to all other queues in the network would be determined through feedback probabilities and by the exponentially distributed service times at this queue.

In Dai and Lin's model activities can be allocated to each processor, and the processing time of a customer depends on the number of customer already processed under activity j (this generalistaion is not possible for our model as we do not remember the number of customers which are served under a regime and assume exponentially distributed service times). After completion of processing time the customer can be routed to another node. The activities are allocated in such a way that the customers are processed under maximum pressure, an allocation is only feasible if no processor is idle, i.e. no activity is wasted at an empty buffer.

This assumption about the feasibility of the activities is in some ways rather restrictive and leads to models such as the one in Figure 5.1 (from Section 6 in [8]) being unstable. This is because the activity (activity 6) under which all buffers are served (indicated by the solid line) can be unfeasible most of the time, the result being that queue 3 is growing infinitely.

Due to the way in which we model our service regimes a problem like this (given our more restrictive assumptions about service and inter arrival times) would not occur. As Dai and Lin [8] point out, if one would drop the feasibility assumption activity 6 could be run all the time, leaving the servers or parts of the processor to idle when there are no jobs in the queues, the system would be stable.

Dai and Lin [8] assume for their first result, Theorem 2, that processors can be split, i.e. two activities that need the same processor can split the attention of this processor according to some function. This is equivalent to our assumption about

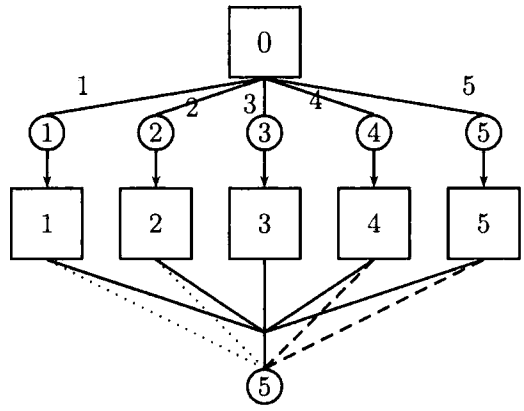


Figure 5.1: Dai and Lin's [8] counter example under the feasibility assumption. The boxes indicate buffers and the circles processors, buffer 0 and processors 1-5 are for the input, so are the activities 1 to 5. There are three service activities: activity 6 (indicated by the solid lines) under which all buffers are served, activity 7 (dotted) and activity 8 (dashed).

randomised controls where pure regimes η are mixed according to some randomised rule Π^r. Dai and Lin also find that it requires additional assumptions to run the queueing system if the processor cannot be split. The methods that Dai and Lin use to show the stability of their model are rather different from our which is why we will not go into further details with the comparison of the results here. It seems however that the approaches to control and the model (except for the Markovian assumption) are very similar and lead to similar results.

Another assumption made by Dai and Lin [8] which we would like to comment on is that of *preemptive service*. This idea has not been mentioned at all in our model. Recall the assumption about switching between different management regimes η we made in the beginning; we assume that switching takes no time is possible just after the event of service completion. This means that if a job completes service at queue i we can change service regime while all the other $N-1$ jobs that could possibly be in service at this time are preempted and are served again under a different service time whenever service is next provided to their queue. For the queue length process we consider this means that we have a jump of the form $-e_i$ or $\pm(e_i - e_j)$ (leaving

or re-entering) after service completion, we do not remember *half served* jobs as is the case in Dai and Lin [8]. It would however be possible in our model to include a short waiting time just before switching in which all unfinished services can be completed. Note that we would need to make sure that under any given control policy there are only a finite number of regime switches.

Another extension of our model which is possible but not considered here is to introduce batch arrival and batch service, both are possible if the batch size is finite since the Lyapunov function results apply as long as the jumps are bounded.

Finally I would like to say that I think that the model is sufficiently general even under the Markov assumption and the results, or at least the main idea which is captured in the convex hull, are easy to understand. It also shows that it is better to start asking for stability in general and then find suitable control policies than doing it the other way round as it is always easier to loose generality.

Bibliography

[1] S. Asmussen, *Applied Probability and Queues*, John Wiley & Sons Ltd., (1987)

[2] D. Blackwell, Discounted dynamic programming, *Annals of Mathematical Statistics* 36 (1965) 226 – 235.

[3] F. Baccelli, S. Foss, Ergodicity of Jackson-type queueing networks. *Queueing Systems*, 17 (1994), 5 – 72.

[4] M. Bramson, Instability of FIFO queueing networks, *Annals of Applied Probability* 4, No. 2 (1994) 414 – 431.

[5] A.A. Borovkov, R. Schassberger, Ergodicity of a polling network, *Stochastic Processes and their Applications* 50 (1994) 253 – 262.

[6] J.G. Dai, On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models, *Annals of Applied Probability* 5, No. 1 (1995) 49 – 77.

[7] J.G. Dai, J.J. Hasenbein, J.H. Vande Vate, Stability and Instability of a two-station queueing network, *Annals of Applied Probability* 14, No. 1 (2004) 326 – 377.

[8] J.G. Dai, W. Lin, Maximum Pressure Policies in Stochstic Processing Networks, *Operations Research* 53, No. 2 (2005) 197 – 218.

[9] J.L. Doob, *Stochastic Processe*, John Wiley and Sons, New York (1953).

[10] D. Down, S.P. Meyn, Piecewise Linear Test Functions for Stability and Instability of Queueing Systems. *Queueing Systems* 27 (1997) 205–226.

[11] G. Fayolle, V. A. Malyshev, M. V. Menshikov, *Topics in the Constructive Theory of countable Markov Chains*, Cambridge University Press (1995).

[12] L. Flatto, H.P. McKean Two queues in parallel, *Communications in Pure and Applied Mathematics* XXX (1977) 255 – 263.

[13] R.D. Foley, D.R. McDonald, Join the shortest queue: stability and exact asymptotics. *Annals of Applied Probability* 11 (3) (2001) 569–607.

[14] R.D. Foley, D.R. McDonald Large deviations of modified Jackson networks: stability and rough asymptotics, *Annals of Applied Probability* 15 (1B)(2005) 519 – 541.

[15] S. Foss, N. Chernova, On the stability of partially accessible multi-station queue with state-dependent routing. *Queueing Systems* 29 (1998) 55–73.

[16] S. Foss, G. Last, Stability of polling systems with exhaustive service policies and state-dependent routing, *Annals of Applied Probability* 6, No. 1 (1996) 116 – 137.

[17] S. Foss, G. Last, On the stability of a greedy polling systems with general service policies, *Probability in the Engineering and Informational Sciences* 12, No. 1 (1998) 46 – 68.

[18] F.G. Foster, On stochastic matrices associated with certain queueing processes, *Annals of Mathematical Statistics* 24 (1953) 355 – 360.

[19] J. Hansen, C. Reynolds, S. Zachary, Stability of processor sharing networks with simultaneous resource requirements, *arXiv:math.PR/0605477 v1* May 2006.

[20] J.M. Harrison, Brownian Models of Open Processing Networks: Canonical Representation of Worklaod, *Annals of Applied Probability* 10, No. 1 (2000) 75 – 103.

[21] J.R. Jackson, Networks of waiting lines. *Operations Research* 5 (1957), 518–521.

[22] I.A. Kurkova, A Load-Balanced Network with two Servers. *Queueing Systems* 37 (2001) 379 – 389.

[23] I.A. Kurkova, Y.M. Suhov, Malyshev's theory and JS-queues. Asymptotics of stationary probabilities. *Annals of Applied probability* 13 No. 4 (2003) 1313 – 1354.

[24] S.H. Lu, P.R. Kumar, Distributed scheduling based on due dates and buffer priorities, *IEEE Transactions on Automatic Control* 36 No. 12 (1991) 1406 – 1416.

[25] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, *Stochastic Models* 7 No. 1 (1991) 1 – 46.

[26] I.M. MacPhee, L.J. Müller, Stability criteria for controlled two-queue systems, Conference proceedings of the ASMDA 2005 in Brest, France.

[27] I.M. MacPhee, L.J. Müller, Stability criteria for controlled queueing systems, *Queueing Systems* 52 (2006) 215 – 229.

[28] I.M. MacPhee, L.J. Müller, Stability criteria for multi-class queueing networks with re-entrant lines, accepted for *Methodology and Computing in Applied Probability.*

[29] S.P. Meyn, R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer (1995).

[30] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, Dover Publications, Inc. New York (1981).

[31] M.F. Neuts, A.S. Alfa, Pair formation in a Markovian arrival process with two event labels, *Journal of Applied Probability* 41 (2004) 1124 – 1137.

[32] J. Niño-Mora, K.D. Glazebrook, Assessing an Intuitive Condition for Stability Under a Range of Traffic Conditions via a Generalised Lu-Kumar Network, *Journal of Applied Probabilty* 37 (2000), 890 – 899.

[33] The New Oxford Dictionary of English, Oxford University Press, 2001.

[34] R.F. Serfozo, An Equivalence Between Continuous and Discrete Time Markov Decision Processes, *Operations Research* 27 (1979), 616 – 620.

[35] S. Stidham Jr.,Analysis, design and control of queueing systems, 50th anniversary issue of Operations Research *Operations Research* 50, No. 1 (2002) 197 – 216.

[36] H. Takagi, Queueing Analysis of Polling Models *ACM Computing Surveys* 20, No. 1 (1988), 5 – 20.

[37] L. Tassiulas, P.B. Bhattacharya, Allocation of Independent Recources for Maximum Throughput *Stochastic Models* 16 (2000), 27 – 48.

[38] B. Van Houdt, A.S. Alfa, Response Time in a Tandem Queue with Blocking, Markovian Arrivals and Phase-Type Service, *Operations Research Letters* 33 (2005) 373 – 381.

Appendix A

Process classification Pull-out

Component i of the regime mean drifts

$$M_i^\eta = \rho^{-1}\left(\sum_{S:s(S)=i}\lambda_S + \sum_{j=1}^{N}\mu_{kj}p_{ji}^\eta - \mu_{ki}\right), \quad \eta = (k,s) \tag{A.0.1}$$

for queues $i = 1, 2, \ldots, N$ with ρ defined as $\rho \geq \max_k\{\sum_S \lambda_S + \sum_i \mu_{ki}\}$.

The convex hull of the regime mean drifts

$$\mathcal{M} = \left\{\sum_\eta p_\eta M^\eta \in \mathbf{R}^N : p_\eta \in [0,1] \text{ and } \sum_\eta p_\eta = 1\right\} \tag{A.0.2}$$

and it interior $\mathrm{Int}_N(\mathcal{M}) \equiv \{z \in \mathcal{M} : B(z,\epsilon) \subset \mathcal{M} \text{ for some } \epsilon > 0\}$, where $B(z,\epsilon) = \{z' \in \mathbf{R}^N : |z - z'| < \epsilon\}$ for $z \in \mathbf{R}^N$.

Any set of parameters for the process (Ξ, Π) falls into one the following four exclusive cases:

C1 $(0,\ldots,0) = \underline{0} \notin \mathcal{M}$ and there exists a state $\alpha \in \mathbf{Z}_+^N$ and a hyperplane

$$L_v(\alpha) \equiv \{z \in \mathbf{R}^N : v^T(z - \alpha) = 0\} \tag{A.0.3}$$

through α with normal vector v separating $\alpha + \mathcal{M}$ from the origin $\underline{0}$. If there exists one such $\alpha \in \mathbf{Z}_+^N$ then there is an infinite cone of such α.

C2 $\underline{0} \notin \mathcal{M}$ and there is no $\alpha \in \mathbf{Z}_+^N$ for which a hyperplane $L_v(\alpha)$ exists separating $\alpha + \mathcal{M}$ from $\underline{0}$.

C3 $\mathrm{Int}_N(\mathcal{M})$ is non-empty, $\underline{0} \in \mathcal{M}$ and there exists no $\alpha \in \mathbf{Z}_+^N$, $v \in \mathbf{R}^N$ such that a hyperplane $L_v(\alpha)$ separates $\alpha + \mathrm{Int}_N(\mathcal{M})$ from the origin.

C4 $\underline{0}$ is a boundary point of $\mathcal{M}$ and either $\mathrm{Int}_N(\mathcal{M}) = \emptyset$ or a supporting hyperplane to $\alpha + \mathcal{M}$ at α separates the origin from $\alpha + \mathrm{Int}_N(\mathcal{M})$ for each α in some cone within $\mathbf{Z}_+^N$.

These cases are depicted for $N = 2$ in Figure A.1.

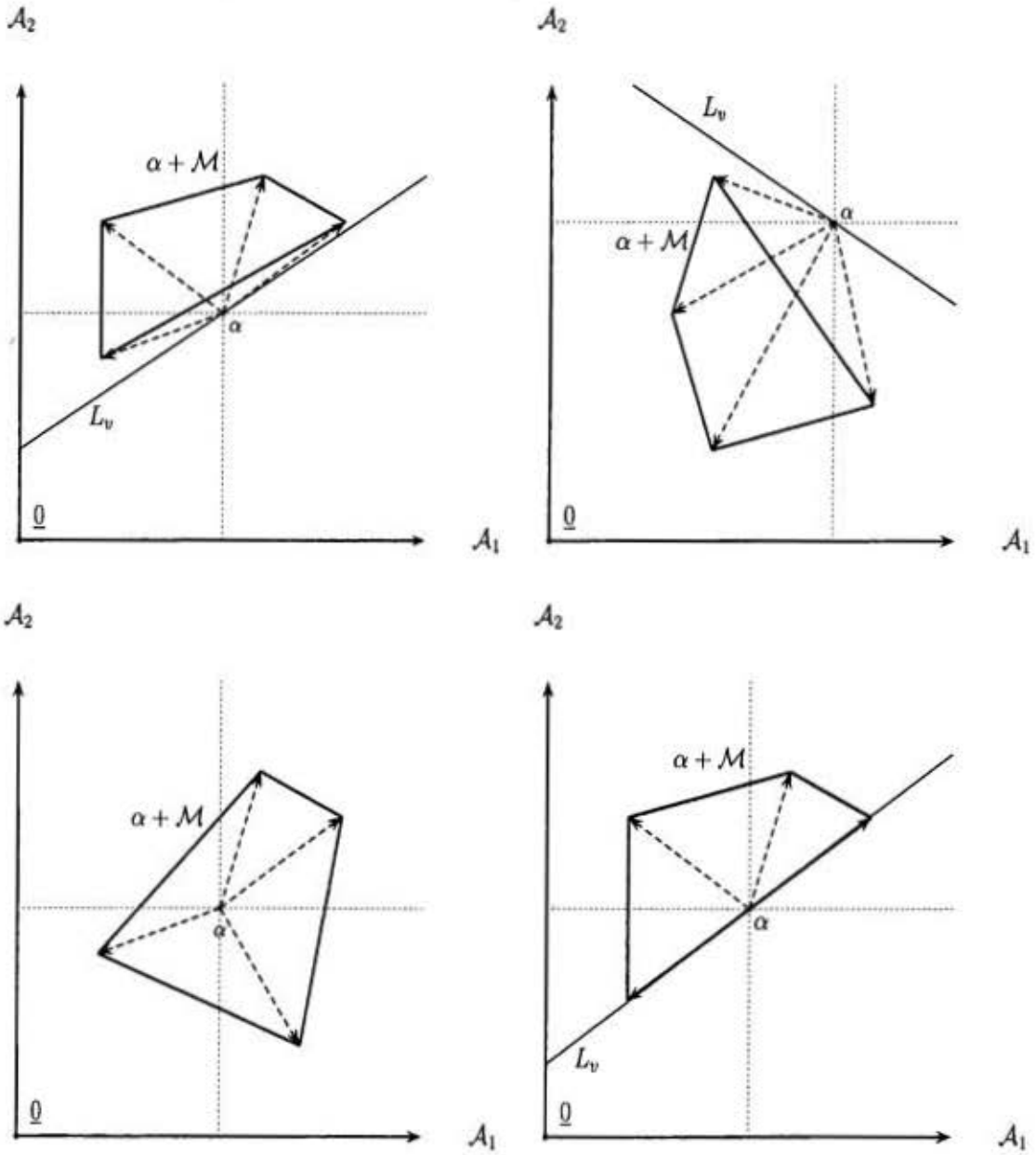


Figure A.1: From top left: C1, C2, and below C3, C4 for $N = 2$.

Appendix B

Proof of Theorem 4.3.2 and Corollary 4.3.3

The proof of Theorem 4.3.2 given here is a slightly extended version of the proof given in our first paper MacPhee and Müller [27] and was written by Iain MacPhee.

We will establish Theorem 4.3.2 with a linear Lyapunov function and appropriate waiting times N_n, $n = 1, 2 \ldots$ whenever Ξ visits states $\alpha \in \partial \mathbf{Z}_0^N$ where queues are empty. We start with a calculation that is needed in the proof of Theorem 4.3.2. The idea is the same as for two queues, since we would like to avoid the boundary we choose the control policies so that the boundary is transient, i.e. the number of visits to states $\alpha \in \partial \mathbf{Z}_0^N$ is finite and in fact decays exponentially fast, while there is some positive probability that the time it takes to reach the the boundary from any state α is infinite.

Lemma B.0.1 *Consider a process X on state space $\mathbf{Z}_0^k$ with bounded jumps that is adapted to some filtration $\mathcal{F} = \{\mathcal{F}_n\}$ and satisfies*

$$\mathbf{E}(X_j(n+1) - X_j(n) \mid \mathcal{F}_n) \geq \varepsilon > 0, \quad j = 1, 2, \ldots, k$$

for all n (for $X(n) = \alpha \in \partial \mathbf{Z}_0^k$ we assume that $X(n+1) = \alpha + \sum_{j \in J(\alpha)} e_j$ where $J(\alpha) = \{j : \alpha_j = 0\}$ i.e. X jumps back to $\mathbf{Z}_+^k$ as quickly as possible). There exist constants C and $\delta_2 > 0$ such that for any $\alpha \in \mathbf{Z}_0^k$

$$\mathbf{P}(X(n) \in \partial \mathbf{Z}_0^k \mid X(0) = \alpha) \leq Ce^{-n\delta_2} \qquad \text{for all } n \geq 1 \,.$$

Further there exists $p(\varepsilon) > 0$ *such that* $\tau_X = \min\{n \geq 1 : X(n) \in \partial \mathbf{Z}_0^k\}$ *satisfies*

$$\mathbf{P}(\tau_X = \infty \mid X(0) = \alpha) \geq p(\varepsilon) \qquad \textit{for any } \alpha \in \mathbf{Z}_+^k \, .$$

Proof We can apply Theorem 1.2.1(ii) to each component X_j of X. We find that for any $\alpha \in \mathbf{Z}_0^k$ and for all $n \geq 1$

$$\begin{aligned} \mathbf{P}(X(n) \in \partial \mathbf{Z}_0^k \mid X(0) = \alpha) &\leq \sum_{j=1}^{k} \mathbf{P}(X_j(n) = 0 \mid X(0) = \alpha) \\ &\leq \sum_{j=1}^{k} \mathbf{P}(X_j(n) < \delta_1 n \mid X(0) = \alpha) \leq kC' e^{-n\delta_2} \end{aligned}$$

for some constants $C' > 0$, $\delta_1 \in (0, \varepsilon)$, $\delta_2 > 0$ which is independent of α. Using these inequalities the proof of Theorem 1.2.1(ii) now guarantees the existence of $p(\varepsilon) > 0$ such that $\mathbf{P}(\tau_X = \infty \mid X(0) = \alpha) \geq p(\varepsilon)$ as required. □

Proof of Theorem 4.3.2 The content of Theorems 2.4.1 and 2.4.2 show that it is not sufficient to ensure that Ξ has negative drift in all components on $\mathbf{Z}_+^N$ as this may lead to difficulties on boundary faces. The scheme of the proof is to define a policy Π^r with blocks and drifts so that Ξ is pushed away from the boundary of its state space but decreases in the interior so the jump distributions on its boundary faces have no major influence on its long term behaviour.

We use as our Lyapunov function $f(\alpha) = \sum_1^N \alpha_i$, the total number of jobs in the system, and we study the process $Y_n = f(\xi(N_n)) = \sum_1^N \xi_i(N_n)$ where the N_n are a strictly increasing sequence of random times where $|N_{n+1} - N_n|$ is bounded for all $n \geq 0$ which we must find.

We start by defining the blocks on which our randomised policy will be based. Our assumptions on routing ensure that Ξ is irreducible under all allowed policies so we do not need to worry about a finite set of states near $\underline{0}$ when establishing ergodicity. For some constant b to be determined so it is sufficiently large, let $\mathcal{B}_a = \{\alpha \in \mathbf{Z}_+^N : \alpha_i \geq b, i = 1, \ldots, N\}$ and let

$$\mathcal{B}_i = \{\alpha \in \mathbf{Z}_+^N : i = \min(m : z_m = \max_j\{\alpha_j\}), z_j < b \text{ for some } j\}$$

for $j = 1, 2, \ldots, N$. In addition to these interior sets we define boundary sets

$$E_Q = \{\alpha \in \mathbf{Z}_0^N : \alpha_i > 0 \text{ for } i \in Q, \alpha_i = 0 \text{ for } i \in Q'\}$$

for each $Q \in \mathcal{P}_N$, where $Q' = \{1, \ldots, N\} \setminus Q$. Our blocks are the sets $\mathcal{B}_i$ and E_Q.

If $\underline{0} \in \mathrm{Int}_N(\mathcal{M})$ there exists $\Delta > 0$ such that $\mathcal{M} \supset B(\underline{0}, \Delta)$ (the convex hull contains the origin with a little ball around it, or $\underline{0}$ is not on the boundary of $\mathcal{M}$). We choose randomised policies π_i^r on the blocks $\mathcal{B}_i$ with mean drifts M^i so that for $i = a$, the expected mean drift M^a has components $M_j^a = -\Delta$, $j = 1, \ldots, N$ while for $i = 1, \ldots, N$

$$M_j^i = \begin{cases} -\Delta, & j = i \\ \Delta/N, & j \neq i \end{cases}$$

This ensures that the mean drift in the interior block $\mathcal{B}_a$ is negative, while on the blocks $\mathcal{B}_i$ for $i = 1,, \ldots, N$ which are close to the boundary the mean drifts drive the process towards $\mathcal{B}_a$.

On the boundary $\partial \mathbf{Z}_+^N$ a policy Π_Q^r on blocks E_Q is such that

$$\mathbf{P}(\xi(t+1) = \alpha + e_i \mid \xi(t) = \alpha \in E_Q \cap \mathbf{Z}_0^k, \Pi_Q^r) > 0 \quad \text{for all } i \in Q' \tag{B.0.1}$$

i.e. we route some arrivals to empty queues, again aiding our aim to drive away from the boundary. By construction $S_n \equiv \sum_{j=1}^N \xi_j(n)$ satisfies

$$\mathbf{E}(S_{n+1} - S_n \mid \xi(n) = \alpha \in \mathbf{Z}_+^N) = \begin{cases} -N\Delta\,, & \alpha \in \mathcal{B}_a \\ -\Delta/N\,, & \alpha \in \mathcal{B}_i,\ i \geq 1 \end{cases}$$

so this process is a strong supermartingale on $\mathbf{Z}_+^N$ and we can set $N_{n+1} = N_n + 1$ whenever $\xi(N_n) \in \mathbf{Z}_+^N$. It remains to define N_{n+1} when $\xi(N_n) \in \partial \mathbf{Z}_0^N$ so that we can apply Theorem 1.2.1(ii).

Consider the evolution of Ξ started from $\hat{\alpha} \in E_Q \cap \{\alpha \in \mathbf{Z}_0^N : \alpha_i \geq b\}$ for some $i \in Q$ over a period of $t_b < b/3$ steps (b is used in the definition of $\mathcal{B}_a$ above). It spends some time in $\partial \mathbf{Z}_0^N$, then enters $\mathbf{Z}_+^N$ for some number of steps and may return to $\partial \mathbf{Z}_0^N$ before time t_b, repeating this pattern. Let W_b denote the number of distinct sojourns in $\mathbf{Z}_+^N$ by time t_b and let σ_n, τ_n denote the lengths of the successive sojourns in $\partial \mathbf{Z}_0^N$ and $\mathbf{Z}_+^N$, $n = 1, 2, \ldots$

The boundary reflection assumption (B.0.1) on (Ξ, Π) implies there is some $p_B > 0$ such that under Π^r, independently of $\alpha \in \partial \mathbf{Z}_0^N$

$$\mathbf{P}(\xi(N) \in \mathbf{Z}_+^k \mid \xi(0) = \alpha \in \partial \mathbf{Z}_0^N) \geq p_B \quad \Longrightarrow \quad \mathbf{P}(\sigma_n > t + N \mid \sigma > t) \leq 1 - p_B$$

i.e. each sojourn of (Ξ, Π^r) in $\partial \mathbf{Z}_0^N$ is stochastically smaller than some random variable V with geometrically bounded tails and mean v say. Further, from our assumption of bounded jumps, there exists a constant κ, independent of α, such that

$$\mathbf{E}(S_1 - S_0 \mid \xi(0) = \alpha) \leq \kappa \text{ for all } \alpha \in \partial \mathbf{Z}_0^k .$$

so the expected change to S_n during each sojourn in $\partial \mathbf{Z}_0^N$ is bounded by κv.

Now we estimate W_b, the number of sojourns in $\partial \mathbf{Z}_0^N$. For any j such that $\hat{\alpha}_j \geq b/3$ we know that component ξ_j cannot reach 0 by time t_b as its jumps are bounded below by -1. This applies particularly to ξ_i as by assumption $\hat{\alpha}_i \geq b$. For $j \in J(\hat{\alpha}) \equiv \{j : \hat{\alpha}_j < b/3\}$ we observe that $\xi_j(n) < \xi_i(n)$ at all times $n \leq t_b$ so Ξ cannot reach $\mathcal{B}_j$ before time t_b. Hence, for these components,

$$\mathbf{E}\big(\xi_j(n+1) - \xi_j(n) \mid \xi(n) \in \mathbf{Z}_+^N\big) = \Delta/N \quad \text{for any } n \leq t_b$$

As the projection of Ξ onto components ξ_j, $j \in J(\hat{\alpha})$ satisfies the conditions of Lemma B.0.1 and the other components cannot reach 0 by time t_b it follows that the lengths τ_n of excursions in $\mathbf{Z}_+^N$ satisfy

$$\mathbf{P}\Big(\tau_n \geq t_b - \sum_{j=1}^{n-1} \tau_j \mid \xi(0) = \hat{\alpha}\Big) \geq \mathbf{P}(\tau_n \geq t_b \mid \xi(0) = \hat{\alpha}) \geq p(\Delta/N)$$

for each $n \leq W_b$ and this estimate is uniform in b. Hence the number of separate sojourns in $\mathbf{Z}_+^N$ by time t_b, W_+ say, is bounded by a geometric random variable with parameter $p(\Delta/N)$ with mean $(1 - p(\Delta/N))^{-1}$. As $W_b \leq W_+ + 1$, there is some finite $w \geq \mathbf{E}(W_b)$ for all b.

Let $I_n \equiv I_{\{\xi(n) \in \mathbf{Z}_+^k\}}$ for $n = 1, 2, \ldots$ and $\bar{I}_n = 1 - I_n$. We have shown that the total time, $\sum_{n=0}^{t_b} \bar{I}_n$, spent by Ξ in $\partial \mathbf{Z}_0^N$ up to time t_b is bounded by $\sum_{n=1}^{W_b} V_n$, a variable with geometrically bounded tails and further

$$\mathbf{E}\Big(\sum_{n=0}^{t_b} \bar{I}_n \mid \xi(0) = \hat{\alpha}\Big) \leq vw$$

We see now that for $n \leq t_b$ and $\hat{\alpha} \in E_Q \cap \{\alpha \in \mathbf{Z}_0^N : \alpha_i \geq b\}$ for some $i \in Q$

$$\begin{aligned}
\mathbf{E}(S_n - S_0 \mid \xi(0) = \hat{\alpha}) &= \sum_{j=0}^{n-1} \mathbf{E}(S_{j+1} - S_j \mid \xi(0) = \hat{\alpha}) \\
&= \sum_{j=0}^{n-1} \mathbf{E}\big(S_{j+1} - S_j \mid I_j\big)\mathbf{E}(I_j \mid \xi(0) = \hat{\alpha}) + \sum_{j=0}^{n-1} \mathbf{E}\big(S_{j+1} - S_j \mid \bar{I}_j\big)\mathbf{E}(\bar{I}_j \mid \xi(0) = \hat{\alpha}) \\
&= -n\Delta/N + \sum_{j=0}^{n-1} \big\{\mathbf{E}\big(S_{j+1} - S_j \mid \bar{I}_j\big) + \Delta/N\big\}\mathbf{E}(\bar{I}_j \mid \xi(0) = \hat{\alpha}) \\
&\leq -n\Delta/N + (\kappa + \Delta/N) \sum_{j=0}^{n-1} \mathbf{E}(\bar{I}_j \mid \xi(0) = \hat{\alpha})
\end{aligned}$$

and hence

$$|\mathbf{E}(S_{t_b} - S_0 \mid \xi(0) = \hat{\alpha}) + t_b\Delta/N| \leq (\kappa + \Delta/N)vw$$

By choosing t_b (and hence b) large enough we can be sure that

$$\mathbf{E}(S_{t_b} - S_0 \mid \xi(0) = \hat{\alpha}) \leq -t_b\Delta/2N < 0$$

Choosing suitably large b and t_b we set our stopping times

$$N_{n+1} = N_n + t_b \quad \text{when } \xi(N_n) \in \partial\mathbf{Z}_0^N \text{ with } \xi_i(N_n) \geq b \text{ for some } i$$

and by Theorem 1.2.1(i) with $D = Nb$ we see that the hitting time to $\{\alpha \in \mathbf{Z}_0^N : \sum_i \alpha_i \leq Nb\}$ is finite and our process (Ξ, Π^r) is ergodic. □

Proof of Corollary 4.3.3 We do not have complete freedom to choose directions M^i now but we can still use the proof of Theorem 4.3.2.

If $\underline{0} \in \mathcal{M} \setminus \mathrm{Int}_N(\mathcal{M})$ but there is no $\alpha \in \mathbf{Z}_+^N$ with a supporting hyperplane $L_v(\alpha)$ that separates $\alpha + \mathrm{Int}_N(\mathcal{M})$ from $\underline{0}$ then any supporting hyperplane $L_v(\alpha)$ with $v^{\mathsf{T}}\beta < 0$ for $\beta \in \mathrm{Int}_N(\mathcal{M})$ has $v_i \geq 0$ for all i. In addition, the line segment joining $\underline{0}$ and α must intersect $\alpha + \mathrm{Int}_N(\mathcal{M})$ for all $\alpha \in \mathbf{Z}_+^N$ and so, for small enough $\delta > 0$ we have $-\delta\alpha \in \mathrm{Int}_N(\mathcal{M})$ for all $\alpha \in \mathbf{Z}_+^N$. Hence there exists $\Delta > 0$ such that $\mathrm{Int}_N(\mathcal{M}) \supset \{z \in \mathbf{R}_-^N : |z| < \Delta\}$ i.e. $\mathrm{Int}_N(\mathcal{M})$ contains a small ball intersected with the strictly negative orthant. Specifically this means there is a randomised strategy π_a with mean drift vector M^a such that $M_i^a < 0$ for each i.

The regimes available at states of the form $\alpha = \alpha_i e_i \in \partial\mathbf{Z}_0^N$ all have zero service rate at all queues $j \neq i$. As we can route arrivals to any queues we know for each

$i = 1, 2, \ldots, N$ there is a randomised strategy π_i with mean drift vector M^i such that $M^i_j > 0$ for all $j \neq i$. As $v \neq 0$ and $v^\mathsf{T} M^i \leq 0$ for each i it follows that $M^i_i < 0$ and $v_i > 0$ for each i.

If at all $\alpha \in \partial \mathbf{Z}^N_0$ there are regimes η available with $v^\mathsf{T} M^\eta < 0$ for some v then by choosing regimes appropriately the weighted total queue process $v^\mathsf{T}\xi(t) = \sum_1^N v_i \xi_i(t)$ can be made a strong supermartingale bounded below by 0 and ergodicity of Ξ follows as in Theorem 4.2.2 as all $v_i > 0$.

If there are boundary faces where the only available regimes η satisfy $v^\mathsf{T} M^\eta = 0$ then we can apply the argument of Theorem 4.3.2 to the process $v^\mathsf{T}\xi(t)$ using the strategies with drifts M^0, M^1, ..., M^N found above. □