Stochastic Models for Resource Allocation in Large Distributed Systems

Guilherme Raposo Thompson

Stochastic Models for Resource Allocation in Large Distributed Systems

THÈSE

présentée pour obtenir le titre de

DOCTEUR

de l'Université Pierre et Marie Curie, Paris VI

École doctorale : Sciences Mathématiques de Paris Centre, ED 386

Spécialité : Mathématiques Appliquées

par

Guilherme THOMPSON

Soutenue publiquement le 08 Décembre 2017 devant le jury composé de :

Directeur de thèse :		
Philippe ROBERT	Directeur de Recherche	INRIA Paris
Rapporteurs :		
Urtzi AYESTA	Directeur de Recherche	IRIT Toulouse
Rudesindo NUNEZ	Professeur	CWI Amsterdam
Examinateurs :		
Laurent DECREUSEFOND	Professeur	Telecom ParisTech
Fabrice GUILLEMIN	Ingénieur de Recherche	Orange Labs
Pierre SENS	Professeur	UPMC

Acknowledgments

First, I would like to thank my advisor Philippe Robert and my coauthors Christine Fricker and Fabrice Guillemin for all their guidance, for their time, dedication and patience. You have helped me to develop myself not only as a researcher but you have marked me also as a person, an imprint that will stay with me along my future journeys. There are no words that can express my gratitude towards your efforts.

I am grateful to both *rapporteurs*, Urtzi Ayesta and Sindo Nunez, who carefully reviewed my work and were very kind and attentive in their reports. Thank you also to the other members of the jury, Laurent Decreusefond and Pierre Sens, who took their time to read my work and to be present at my *soutenance*.

I would like to thank my team-mates at RAP, who made me enjoy my time at INRIA with very long and interesting talks during our breaks: Nicolas Broutin, Othmane Safsafi, Davit Martirosyan, Ravi Mazumdar, Nelly Maloisel and my math siblings, Renauld Dessalles, Sarah Eugène and Wen Sun. A thank you also to those who have been with me during countless hours in the *navette* and thank you to our coffee machine neighbours from Gallium. During my whole stay at INRIA, be it close to the woods of Rocquencourt or in the very core of Paris, I always felt well surrounded.

I thank my friends from CAp and UFRJ and especially Lucas De Tomaso, Isabela Kuschnir, Amannda Dacache, Isabelle Dutra, Rodrigo Fraga and O Grupo. The distance could not tear us apart. To Charline Hélène Dutarte, for being the hub of our friends in Paris and organising all our Christmas diners and *soirées de tarot*. A special mention to my teachers, the real and best kind of teachers who were not just dutifully standing in front of the black board but those who connected with me and influenced deeply the way I think up to today, Ilydio Pereira de Sá, Nilton Alves Jr., Luiz Antônio Meirelles, Maria Alice da Rocha and Samuel Jurkiewicz. Thank you also to my colleagues at Eleva who were very supportive for a smooth transition to my PhD journey, especially my mentor Lucas Vivone.

I am very grateful to my family in Brazil, who have been with me from the beginning to the end. Thank you for your support, for your deep unconditional love. My parents, Isabel and Celso, my big sister, Nathalia, my *cunhado* (in-law), Marcelo and Zandor, a very good boy. I would also like to thank all of those who cast me and today continue to live in my thoughts, especially my grandfather Aldahyr. Lastly, thank you, Julia for having borne with me all the troubles of a PhD as well as for being my source of inspiration, motivation

and endurance. Without you I would still be writing the introduction of this document... or I would be in Brazil even! My reviewer, translator, coach, girlfriend (transmuted into wife), my Love and much more. Thank you.

Abstract

This PhD thesis investigates four problems in the context of Large Distributed Systems. This work is motivated by the questions arising with the expansion of Cloud Computing and related technologies (Fog Computing, VNF, etc.). The present work investigates the efficiency of different resource allocation algorithms in this framework. The methods used involve a mathematical analysis of several stochastic models associated to these networks.

Chapter 1 provides an introduction to the subject in general, as well as a presentation of the main mathematical tools used throughout the subsequent chapters.

Chapter 2 presents a congestion control mechanism in Video on Demand services delivering files encoded in various resolutions. We propose a policy under which the server delivers the video only at minimal bit rate when the occupancy rate of the server is above a certain threshold. The performance of the system under this policy is then evaluated based on both the rejection and degradation rates.

Chapters 3, 4 and 5 explore problems related to cooperation schemes between data centres on the edge of the network. In the first setting, we analyse an offloading policy in the context of multi-resource cloud services. In second case, requests that arrive at a congested data centre are forwarded to a neighbouring data centre with some given probability. In the third case, requests blocked at one data centre are forwarded systematically to another where a trunk reservation policy is introduced such that a redirected request is accepted only if there are a certain minimum number of free servers at this data centre.

Résumé

Cette thèse étudie quatre problèmes dans le contexte des grands systèmes distribués. Ce travail est motivé par les questions soulevées par l'expansion du *Cloud Computing* et des technologies associées (*Fog Computing*, *VNF*, etc.). Le présent travail étudie l'efficacité de différents algorithmes d'allocation de ressources dans ce cadre. Les méthodes utilisées impliquent une analyse mathématique de plusieurs modèles stochastiques associés à ces réseaux.

Le Chapitre 1 fournit une introduction au sujet en général, ainsi qu'une présentation des principaux outils mathématiques utilisés dans les chapitres suivants.

Le Chapitre 2 présente un mécanisme de contrôle de congestion dans les services de *Video on Demand* fournissant des fichiers encodés dans diverses résolutions. On propose une politique selon laquelle le serveur ne livre la vidéo qu'à un débit minimal lorsque le taux d'occupation du serveur est supérieur à un certain seuil. La performance du système dans le cadre de cette politique est ensuite évaluée en fonction des taux de rejet et de dégradation.

Les Chapitres 3, 4 et 5 explorent les problèmes liés aux schémas de coopération entre centres de données situés à la périphérie du réseau. Dans le premier cas, on analyse une politique de déchargement dans le contexte des services de cloud multi-ressources. Dans le second cas, les demandes arrivant à un centre de données encombré sont transmises à un centre de données voisin avec une probabilité donnée. Dans le troisième cas, les requêtes bloquées dans un centre de données sont transmises systématiquement à une autre où une politique de réservation une politique de rÃĭservation (*a trunk*) est introduite tel qu'une requête redirigée est acceptée seulement s'il y a un certain nombre minimum de serveurs libres dans ce centre de données.

Contents

Chapter 1

Introduction

This introduction begins with a presentation of Cloud Computing which is the main object of study of this thesis in order to familiarise *non-expert* readers with a certain vocabulary and important key concepts. An overview of the issues and challenges related to Cloud Computing provides reasons for why conducting research in this field is so crucial. The focus of this thesis being the dynamic allocation of resources in the framework of stochastic networks, a selection of the most relevant literature of this topic is presented.

It follows an exposition of relatively simple stochastic queueing and network models in order to facilitate the reader's understanding of the more sophisticated models used in the thesis for the analysis of different aspects of the Cloud Computing environment. These simple models rely on similar assumptions and shall provide a good first understanding of basic model mechanisms.

In a next section, the main mathematical tools are described and their usage is illustrated through their application to the simple stochastic models introduced previously. A large part of this section concerns *Markov processes* and related tools necessary for its analysis, namely kernel problems and scaling techniques. This is followed by a quick introduction of the *stochastic averaging principle* and an explanation of how the cohabitation of different time scales determines the macroscopic evolution of large scale systems.

Finally, the content of the four chapters of this thesis is presented.

1.1 Cloud Computing

Cloud Computing is a service model, whose principle is the on-demand offer of access to shared computing resources. In the literature, the definition of this relatively recent concept is not yet pinned down. Several authors interpret the term in different ways (see [YBDS08, FZRL08, VRMCL08, Gee09, ASZ+10, ZCB10, MG11]). Among these authors, Foster *et al.* [FZRL08] propose one of the more broad and inclusive definitions

> [Cloud Computing is] 'A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualised, dynamically-scalable, managed computing power, stor-

> age, platforms, and services are delivered on demand to external customers over the Internet.'

From this extract, we retain the main concept of 'a shared large pool of digital resources delivered on demand'. From now on, we refer to this rather general definition whenever we use the term *Cloud*.

The idea of pooling resources together in order to mitigate idleness and benefit from economies of scale to reduce costs is not particularly new. Since the popularisation of computers and the access to Internet, many techniques have been explored to exploit the full potential of the processing power associated with these new technologies. We witnessed for example the development of *Grid Computing*, *Utility Computing*, *Service Computing* which are service concepts relatively similar to Cloud Computing. See Foster *et al.* [FZRL08] for a detailed discussion of the taxonomy of these technologies and Voas and Zhang [VZ09] for a critical discussion about the existence of Cloud Computing as a new computational paradigm.

Although these techniques have been in vogue for some while, none of them was as influential as *Cloud Computing*. Only as Internet access and computing resources have become cheaper, more powerful and ubiquitously available, users and companies felt confident enough to adopt the new service model. Cloud Computing has particularly been boosted by the interests of major players in the Internet market such as Amazon, with the *Amazon Web Services* (AWS) [Ama], Microsoft, with *Microsoft Azure* [Mic], and Google, with *Google Cloud Platform* (GCP) [Goo], leading to a genuine trend over the last decade which is currently still gaining momentum. See for example Amazon's portfolio of customer success stories which provides a glimpse of how much Cloud services are penetrating the Internet market ([Ama]).

Essential Characteristics

Despite the variety of definitions of Cloud Computing services (see for instance Armbrust *et al.* [ASZ+10] and Mell and Grance *et al.* [MG11]), there exists a set of criteria recurrently cited by the authors of the field which has to be fulfilled for a service to be considered a part of the Cloud Computing environment.

Elasticity is probably the most notorious feature of Cloud services. Resources can be allotted and released (almost) instantaneously. Users can ask for more resources when in need, as well as release superfluous resources. Elasticity can be understood as an enhancement on scalability, as it does not only include keeping up with the increasing loads of jobs but also takes into account the proper adjustment for decreases in the need for resources.

Another common aspect of Cloud services and a key factor for its commercial success is the billing system. The Cloud is based on Pay-as-you-go (PAYG) charging mechanisms, where customers chose a virtual machine based on its specifications and are billed as a function of these specifications and for the time they use the service. For example, Amazon's Elastic Compute Cloud (EC2) allows users to configure their virtual machines (VM) in terms

of memory RAM, CPU cores, storage space, connectivity bandwidth, Operational System (OS), etc. However, PAYG is not necessarily a characteristic of all Cloud services today. Cloud operators are currently developing middle and long term contracts, intended to smooth resource consumption, lock clients in and improve the dimensioning of their data centres. We also observe the emergence of *serverless computing* where Cloud containers hold only the necessary resources for the execution of a specific application, and services like Amazon Lambda, where customers do not have to reserve an entire VM any more but pay for the computing resources necessary for the execution of their tasks.

Availability is an essential condition for Cloud services because it is a precondition for people trusting in the new technology. This is why all players of the Cloud ecosystem are today committed to high availability, proposing elevated Service Level Agreements (SLA). See for instance the “SLA" section in [Ama], [Goo] and [Mic]. High availability is ensured through the offer of a reliable pool of resources and redundancy mechanisms.

Clouds can mainly be divided into either public or private Clouds, even though there also exist some hybrid forms or community Clouds which host private and public services in a single structure. The infrastructure of public Clouds is intended to serve many customers from different organisations. Customers should not be affected by the use of resources by others sharing the same physical machine (PM). Privates Clouds are designed to serve exclusively a single organisation. Today, many companies opt for this deployment level because data security issues and legislation constraints render the adoption of public services prohibitive.

Cloud Computing services can be organised into the following categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). This classification is usually referred to as *SIP model*. Certain subcategories have been proposed, such as Business Process as a Service (BPaaS) as a part of SaaS or Desktop as a Service (DaaS) as a category of PaaS. However, these subcategories fail to offer a real surplus over the original SIP model, especially in terms of mathematical modelling.

Infrastructure as a Service (IaaS) is a service model where providers offer on-demand computing resources to their customers. This is done through virtualisation, which is the emulation of a real physical system (with hardware specifications) in a virtualised manner. We say a virtual machine is instantiated in a server or data centre when the physical resources (Physical Machines (PM)) supporting the virtual machines are located there. Customers can create such virtual machines which use resources such as Random Access Memory (RAM) and Computing Processing Unit (CPU) cores. In this manner, multiple users can simultaneously allocate customised virtual machines and access them over the Internet or a local network. The provision of IaaS contributes to the reduction of equipment acquisition and maintenance costs for the users.

Platform as a Service (PaaS) allows the user to deploy onto the Cloud

infrastructure applications created using programming languages, libraries, services, and tools supported by the service provider. The user does not manage or control the underlying Cloud infrastructure (the network, servers, operating systems, etc.) but he has control over the deployed applications and possibly configuration settings for the application-hosting environment.

In Service as a Service (SaaS), the user has the possibility to run the provider's applications which are deployed onto a Cloud infrastructure. The user does not manage or control the underlying Cloud infrastructure, neither individual application capabilities, with the possible exception of limited user-specific application configuration settings.

In the hierarchy of the Cloud, IaaS is considered to be the core of Cloud services because both PaaS and SaaS are dependent on the infrastructure, without which they could not exist. Emphasising the importance of infrastructure, Zhang *et al.* [ZCB10] classify Cloud service providers into *infrastructure providers*, whose core business is the IaaS, and *service providers*, who propose services in the Cloud but rent resources from infrastructure providers. IaaS is also the most interesting aspect of Cloud Computing for mathematical studies as it is relatively simple to model. The requests executed in this level of Cloud require "raw" units of resources, such as CPU cores or GB of RAM, which need to be available lo host the incoming clients. In this context, the mathematical framework of Queueing Theory and also some tools from Operational Research are very promising for the analytical evaluation of the performance of Cloud systems.

Issues and challenges

Although Cloud Computing has been widely adopted by many major players in the telecommunication industry, the research in this field is still at an early stage. As for many other scientific advances, the technological possibilities develop much faster than our understanding of their functioning. Many key challenges, including automatic resource provisioning, power management and security management, are only starting to receive attention from the research community, while new challenges keep emerging from new applications and methods. See Zhang *et al.* [ZCB10] for more details.

The work in this thesis focuses more particularly on the challenges related to the current emergence of new decentralised Cloud architectures. Most of the commercial Clouds are today still implemented in large data centres and operated in a centralised fashion. In this set-up, all (or most of) the requests, originated near or far from the large data centre, is executed in this unit. This design has the advantage of economies of scale and high manageability but it comes at the price of high energy expenses (usually associated with the cooling down of the data centres), elevated up-front investments in infrastructure and increased latency delays. Long transmissions across the network are costly and often associated with security risks and cross-boarder juridical issues. Furthermore, the centralised system constitutes a potential bottleneck to the further development of Cloud Services, as its delivery channels are likely to get congested due to the continuous increase in the volume of Internet traffic.

Small-sized data centres that are better geographically distributed and

work in a decentralised manner can be more advantageous. They do not only require a less powerful and less expensive cooling system but, being located closer to the user, they also constitute a solution to the high transmission costs, saturation of delivery channels and latency times present in a centralised system. This is particularly important for response time-critical services such as content delivery and interactive gaming. In 2009, Valancius *et al.* [VLM+09] propose for example the usage of small (nano) data centres for hosting of Video on Demand services. We therefore witness a general movement of data centres towards the edge of the network away from centralised structures. This distribution of Cloud Computing resources at the edge of the network is known as Fog Computing (see [WRSvdM11, BMZA12, RBG12]) and allows for example to handle the enormous amount of data generated by devices located all over the network (i.e. Internet of Things (IoT), see Atzori *et al.* [AIM10]).

With these developments towards a more distributed Cloud architecture, the problematic of how to handle local demands relying on much smaller service units is currently gaining in importance. It is now possible that local data centres face scarcity of some resources which results in the rejection of user demands despite the total amount of resources in the system being sufficient to satisfy all user demands if the resources would be pooled in a centralised data centre. In order to reduce the occurrence of request blocking, Fog Computing data centres will have to collaborate, for example by offloading user demands to another data centres that does not face saturation. See Sharif *et al.* [SCAFV16] for an extensive discussion about the perspectives for decentralised service models for Cloud technologies.

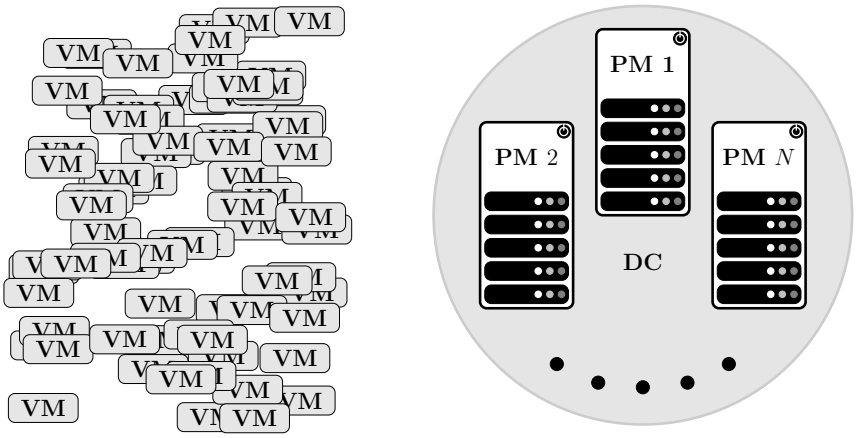
The development of new models is crucial for the study of the Cloud and future technologies building upon it. The computation associated with these models needs to be adapted to the increased volume and diversity of the Cloud Computing traffic. For instance, classical approaches for network optimisation, such as global load balancing strategies might be too slow depending on the magnitude of the system and necessitate revision. Another example, bandwidth sharing policies have to be redesigned in order to ensure a *fair* (equitable) division of network resources in the current context of an increase and diversification of customer demands. This thesis develops such models, taking into account the stochasticity of the arrival of user demands and duration of service times, which is absent in most of the research conducted so far.

Resource Allocation in the Cloud

The focus of this thesis is the dynamic allocation of resources in the framework of large stochastic networks. This part presents a selection of the most relevant literature on this topic.

In the context of Infrastructure as a Service, the Cloud service provider faces the challenge of allocating resources efficiently by assigning incoming requests of virtual machines to physical machines located in the Cloud data centre.

This issue of resource allocation in Cloud systems has been addressed by many authors with different perspectives, scopes and objectives. As Cloud



Flow of incoming virtual machines *Cloud provider data centre*

Figure 1.1 – Representation of a Cloud providers data centre

systems are very different one from another, research branches in many directions. See Jennings and Stadler [JS15] for a complete survey about resource management techniques for Cloud services and Masdari *et al.* [MNA16] for an overview of the techniques used in Cloud services for determining virtual machine placement.

Most of the literature focuses on Cloud Computing systems which rely on one type of resource exclusively (e.g. RAM memory, CPU cores, bandwidth, etc.). Queueing theory has been widely used for the study of such systems as it is particularly adapted to this context and provides a rich toolbox for system optimisation. One of the first works in this field is Yang *et al.* [YTDG09] who consider queues to evaluate metrics regarding the performance of Cloud services and Khazaei *et al.* [KMM12] study a simple Cloud system using $M/G/m/m+r$ queues. These papers as well as many other papers consider single resources systems. In the setting of Cloud Computing environments proposing multiple resources, the allocation of resources is more complex. The tools of queueing theory are less adapted and provide fewer information on the system under scrutiny. In the literature on multiple resources, the issue of resource allocation has classically been addressed through bin-packing and knapsack problems in the framework of stochastic optimisation [RT89, Ros95, GI99]. From a utility (or reward) perspective, these methods aim to determine which (possible) configuration would maximise some given utility function (over time). For example, consider a data centre which is equipped with C_i units of resources $i \in I$, and VM of type $j \in J$ requiring $A_{i,j}$ units of each resource. If a VM of type j is hosted in this system, the operator will be rewarded with a w_j bonus. If we denote x_j the number of VM of type j in the system, then the optimisation problem is simply defined by

$$\max_{x \in \mathcal{S}} \sum_{j \in J} w_j x_j \text{ with } \mathcal{S} = \left\{ x = (x_j) \in \mathbb{N}^{|J|} : \sum_{j \in J} A_{i,j} x_j \leq C_i \, \forall i \in I \right\}.$$

However, bin-packing and knapsack problems are NP-Complete [CKPT16] and

are therefore too slow to be used as a viable resource orchestration practice for large Cloud systems despite the development of many efficient heuristic methods in recent years. In addition, such methods allow only to consider the static characteristics of the systems with very limited applications in dynamical contexts.

Given these shortcomings in the multi-resource context, the recent literature focuses back on systems providing a single resource exclusively, or straightforward application of queueing theory resulting in quicker calculations and providing information on system dynamics. However, contrarily to the literature on queueing theory mentioned previously, this recent strand of literature considers more complex systems which are composed of several queues, modelling the rack of servers, servers farms or data centres. One approach in this framework of resource allocation of a single resource considering multiple queues is intelligent dispatching. Instead of running calculation power intense optimisation programs, the focus lies on the "sending" of a VM to a PM. One popular technique consists of dispatching virtual machines to the most busy physical machines in order to be able to generate as many idle PM as possible which can then be turned off (for power saving considerations). For example, Stolyar and Zhong [SZ13] introduce the algorithm *Greedy with sub-linear Safety Stocks (GSS)* with the objective of minimising the number of active servers in a multi-resource data centre in a heuristic manner, and Xiao *et al.* [XSC13] present a "skewness" metric to control for unevenness in the multidimensional resource utilisation for VM dispatching, aiming to minimise the number of servers in use. On the contrary, load balancing techniques have the aim to deploy the full capacity of the system, thus tempting to assign VM homogeneously across the pool of physical machines. In telecommunications theory, a common reference is the policy known as *join the shortest queue* (JSQ) which is, however, only effective in small-scale systems as the state of each physical machine must be known in order to dispatch VM accordingly. For instance, Maguluri *et al.* [MSY12] discuss the implementation of JSQ in a multi-resource Cloud service data centre and show that this policy is not suitable for large scale systems. Other techniques aim to improve global performance using only local information (from a small sample of data centres). The *power-of-choice* (or supermarket model) has been introduced by Mitzenmacher [Mit96] and Vvedenskaya *et al.* [VDK96]. In this approach, the dispatching program randomly compares the size of the queue in a limited amount of different physical machines (usually 2 or 3) and assigns the VM to the server with the shortest queue (the smallest workload), resulting in quick and efficient dispatching. This technique inspired the recent framework of *pull based policies* such as *Join-Idle-Queue* (JIQ), proposed by Lu *et al.* [LXK+11] where a list of idle physical machines is kept to which virtual machines are then allocated. See Stolyar [Sto15] and Foss and Stolyar [FS17] for a similar policy, the *PULL* algorithm, which assigns servers to customers instead of the common customer to server assignment.

In this thesis I consider resource allocation for *on-time* services, such as for example Microsoft Azure [Mic] systems, which necessitate instantiation of virtual machines as soon as they are assigned to a physical machine. If a system

does not dispose of a sufficient amount of resources (in one or several PM) to host an arriving VM, the request is rejected. Cloud service providers have an incentive to avoid high rejection rates which result in the loss of customers and eventually a decline in revenues. On-time systems have been studied by many authors. Recently, Xie *et al.* [XDLS15] and Mukhopadhyay *et al.* [MKMG15] use loss systems in the study of intra data centre VM allocation with power of choice mechanisms. In Xie *et al.*, the customers require different amounts of resources during their service. In Mukhopadhyay *et al.*, heterogeneous types of servers are considered, differentiated by their capacity (size). However, these models mostly are adapted to cases where system performance is determined exclusively by one resource (i.e. the system's bottleneck). In this thesis, similar loss systems are considered, but introducing in addition a generalisation to the multi-resource case (see Chapter 3 concerning cooperation schemes between multi-resource processing facilities).

In the literature, the allocation of resources has been focused on the *intra data centre schemes*, i.e. allocation of resources within a given data centre. However, as mentioned before the Cloud is evolving, there exists a need for collaboration in between small-size data centres in order to reduce the rejection of user demands in locally saturated data centres in the new Cloud architecture. For this reason, this thesis focuses on allocation of resources in between data centres (inter data centre schemes). For instance, consider a system with N servers distributed in 2 data centres – N_1 servers in data centre 1 and N_2 in data centre 2, with $N_1 + N_2 = N$. The system has the same capacity as

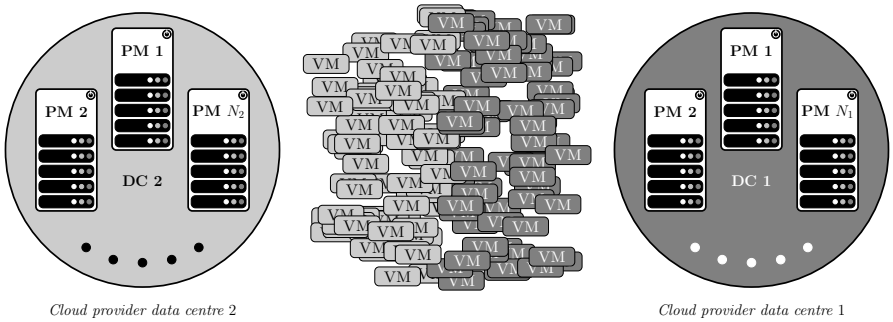


Figure 1.2 – Representation of a Cloud system with multiple data centres

its centralised version, but the customers in each processing facility may experience a very different quality of service due to traffic asymmetry. In the worse case, resource can be idle in one data centre while depleted in the other facility, causing "unnecessary" blocking of customers.

In Chapter 2 of this thesis, I investigate an allocation scheme for on demand video services where clients are served with the lowest service grade (level or resolution in terms of video quality) as soon as the link occupation is above a certain threshold, allowing the system to mitigate rejection of customers. Chapter 3, 4 and 5 focus on policies to enable the cooperation between decentralised facilities in a effort to improve the performance (particularly reducing the blocking rates of customers) which can find an application in the new decentralised Cloud architectures which are currently emerging. In Chapter 3,

I study an offloading scheme between two data centres in the multi-resource context. In the framework of resource-specialised virtual machines requiring different proportions of each resource, this policy aims to alleviate the local charge of a resource by forwarding the customers (VM) which require the most of the resource which is depleted locally. In Chapter 4, I consider the policy which forwards jobs from a data centre to another with some probability if the request cannot be served locally. And, in Chapter 5, a system similar to the one presented in Chapter 4 is considered under a another offloading policy: jobs are systematically forwarded from one data centre to the other, and are only accepted in the second data centre if there are a minimum amount of free servers. Notice that in this case, only one processing facility (data centre 1) is offloading customers to the other (data centre 2), which protects its original requests using a trunk reservation mechanism.

1.2 Stochastic Modelling of Services

This section introduces simple versions of the models used throughout this thesis for the analysis of Cloud Computing systems. The models are presented in the order of their complexity to acquaint the reader step by step to the classical tools of stochastic modelling.

Queueing Systems

A queueing system is the representation of a real system, where clients (or jobs) arrive, demand a service which takes some time, and then leave the system. A queue may be equipped with one or more servers and a buffer (or waiting) zone, such that clients who are not being served can wait for the starting of the service. If a queue has no buffer zone or if all of its servers and buffer are already used, all arriving customers are rejected until some space is freed. The duration of the time that clients are waiting is called waiting time. The duration of the time that the service is being executed is called service time.

The study of such queueing systems originated from the development of telecommunication technologies and gained in importance with the arrival of centralised computing units (main frame architecture) some years later. The main objective was to be able to calculate blocking probabilities of clients in the telephone network and waiting times of customers until access to the computing power in the main frame.

Erlang's Problem

In the beginning of the twentieth century, while working for the Copenhagen Telephone Company (CTC), the Danish engineer Agner K. Erlang (1878–1929) was confronted with the intriguing problem of dimensioning the company's telephone network.

Back then, a phone call was the realisation of a connection between a caller and receiver, using a circuit board on the links between these two interlocutors.

While the phone call was in progress, the circuit stayed occupied. If a new call was attempted when all the circuits were occupied, this call was rejected.

Local communities were connected by one board of circuits, see Newman [New10]. Erlang was responsible for determining the number of circuits to ensure a certain service level (or grade), given by the probability a client is rejected by the exhaustion of the circuits. In his efforts to engineer this system, Erlang published many papers with two being of particular importance for the development of this thesis.

His 1909 seminal paper "The Theory of Probabilities and Telephone Conversations" [Erl09] [1] showed that the number of calls coming in follows a *Poisson distribution*. This approximation allows to calculate system performance and is used until today in many areas due to its practicality.

In 1917, Erlang published "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges". Analysing teletraffic data, Erlang observed that the duration of the phone calls were exponentially distributed with mean μ^{-1}. Using this, the fact that incoming calls follow a Poisson distribution at rate $\lambda > 0$ and analysing the evolution of the number of used circuits, Erlang derived his famous formula for traffic design which we call today the Erlang-C formula.

$$B(\lambda, \mu, C) = \frac{(\lambda/\mu)^C/C!}{1 + (\lambda/\mu) + (\lambda/\mu)^2/2 + \cdots + (\lambda/\mu)^C/C!}. \tag{1.1}$$

This formula expresses the probability of an arriving customer to be blocked in a system composed of C circuits.

It is interesting to note that the theory of stochastic processes which would offer a formal support for the study of queueing systems was developed only after Erlang's death by Andrey Kolmogorov (1903–1987). This means that Erlang studied the telephone network only using classic probabilistic reasoning which is remarkable. Erlang was not only a pioneer in stochastic modelling of services, but he also laid the cornerstone of *Queueing Theory* and its branches. See Cohen and Boxma [CB85] and Kingman [Kin09] for more details.

Queues with one server

Queues are characterised by a number of servers, a waiting zone, arrival and departure processes of clients and a service discipline. The simplest queue is composed of a single server and an unlimited waiting zone. Clients are served one at a time and in their order of arrival which is why this service discipline is called First Come First Served (FCSF) or First In First Out (FIFO). Those who are not served directly upon their arrival start waiting in the queue. It is assumed that clients arrive according to a Poisson process with intensity λ. The time a client occupies the server (the service time) is exponentially distributed with mean $1/\mu$. The key characteristic of the Poisson arrival process and the exponentially distributed service times is that they

1. Both cited papers by Erlang, [Erl09] and [Erl17] were published in English in the biographical anthology "The life and works of A.K. Erlang" by Brockmeyer, Halstrøm and Jensen [BHJ48], which includes some works from the authors.

describe a system which is memoryless. In Kendall's notation [Ken53], this queue is therefore called a $M/M/1$ queue, with M standing for memorylessness and 1 for the number of servers.

The $M/M/1$ queue can also be analysed through an associated embedded discrete-time Markov chain. This Markov chain behaves like a random walk on $\mathbb{N}$ or, more illustratively, as a birth-death process, where the number of individuals (i) in a population is analogous to the number of clients in the queue. With probability $\lambda/(\lambda+\mu)$ the number of individuals (clients) transitions from i to $i+1$. A decrease of the queue (the population) from i to $i-1$ happens with probability $\mu/(\lambda+\mu)$, for $i \geq 1$. See Chapter III of Asmussen [Asm03, p. 75].

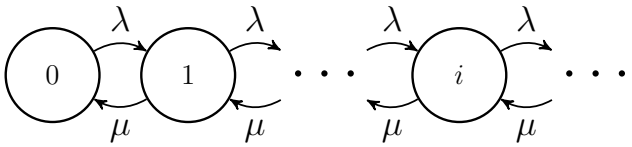


Figure 1.3 – Transition diagram of the $M/M/1$ Queue

Erlang's famous result (as introduced in the previous section) is based on his observation of the relationship between the probability of finding i clients present in the queue and the "speed" in the change of the number of clients queueing ($i \pm 1$). He supposed that if the system reaches its equilibrium, the "flows" $i \to i+1$ and $i+1 \to i$ (conversely to $i \to i-1$) should be of the same order. These "flows" are given by $\pi(i)q(i,i+1)$ and $\pi(i+1)q(i+1,i)$, where $\pi(i)$ denotes the probability of finding i customers in the system and $q(i,j)$ denotes the transitions rates from i to j. Thus, at equilibrium we have the following relation for $i \geq 1$

$$\pi(i)q(i,i+1) = \pi(i+1)q(i+1,i). \tag{1.2}$$

The number of customers in the system can be described by a Markov process whose transitions for the $M/M/1$ queue are $q(i,j)$ are given by

$$q(i,j) = \begin{cases} \lambda & \text{if } j = i+1 \\ \mu \mathbb{1}_{\{i>0\}} & \text{if } j = i-1. \end{cases}$$

Recursively, if $\lambda < \mu$, we have can solve difference equation for $i \in \mathbb{N}$, and obtain

$$\pi(i) = (1-\rho)\rho^i.$$

One sees that this relatively simple model already provides some crucial information on the behaviour and the performance of queueing systems. Among other performance metrics that can be derived we have for example the distribution of waiting times, the utilisation rate of the server or the mean length of a queue. The simplicity and yet versatility of Erlang's results has thus propelled the popularity of queueing models. In later work, Pollaczek (1892–1981) and Khinchine (1894–1959) derive further fundamental results for similar queueing

systems with Poisson arrivals and general service times, ($M/G/1$ queues in Kendall's notation) which led to an explosion of the application possibilities of queueing systems.

Queues with many servers

Other important mathematical objects used in this thesis are queues with many servers and no waiting zone. These queueing systems are used in the modelling of on-demand services where clients do not wait for getting served and are rejected if the capacity of the system is exhausted (in case all servers are occupied). These systems are the original object of study of the works of Erlang.

The system is equipped with C servers and can host up to C clients simultaneously, which is why it is called $M/M/C/C$ in Kendall's notation. Just as for the $M/M/1$ queue presented previously, customers in the $M/M/C/C$ arrive according to a Poisson process with rate λ and service times are again independently and exponentially distributed with finite mean $1/\mu$.

The number of customers in the system can be described by a Markov process taking values in $\mathcal{A} \overset{\text{def.}}{=} \{1, \dots, C\}$, whose transition rates are given by

$$q(i,j) = \begin{cases} \lambda \mathbb{1}_{\{j \leq C\}} & \text{if } j = i+1, \\ \mu j & \text{if } j = i-1. \end{cases}$$

Note that the $M/M/C/C$ queue is in fact a truncated version of the $M/M/\infty$ queue because its maximal capacity is fixed at C. If the original process takes values in $\mathbb{N}$ and has an invariant distribution $\pi = (\pi(i),\, i \in \mathcal{S})$ and the truncated process lives in $\mathcal{A}$ then, for any $i \in \mathcal{A}$, one has

$$\nu(i) = \pi(i) Z^{-1},$$

where Z is a normalisation constant given by $Z = \sum_{i \in \mathcal{A}} \pi(i)$. The invariant distribution of the $M/M/\infty$ can be simply obtained by recurrence using the balance equations (the same as Equation (1.2)), such that $\pi(i) \propto \rho^i / i!$, where $\rho = \lambda/\mu$. Therefore, the blocking probability in this system is simply given by Erlang's formula

$$\nu(C) = \left(\sum_{i=0}^{C} \frac{\rho^i}{i!} \right)^{-1} \frac{\rho^C}{C!} = B(\lambda, \mu, C).$$

Forty years later, Takács [Tak69] showed Erlang's formula is insensitive to the distribution of the service times, as long as the service time has a finite mean, i.e. blocking rate of customers on a $M/G/C/C$ queue is also given by Erlang's formula.

From here on, we look at extensions of Erlang's $M/M/C/C$ queue to more general systems, so called loss systems. These systems received their name because customers are either served directly upon their arrival or they are rejected, generating losses, in case all of the servers are occupied. Loss systems

are useful to model communication systems or, more generally, on-demand services where customer requests are usually served instantaneously until exhaustion of the system's resources and after which subsequent customer requests are rejected. The work in this thesis is mainly based on such loss systems.

For sake of illustration we consider the following example of a loss system. The system is equipped with $C_i \in \mathbb{N}^*$ units of each resource i, for $1 \leq i \leq I$. There exist $J \in \mathbb{N}^*$ types of jobs. Jobs of type $1 \leq j \leq J$ arrive in the system accordingly to a Poisson process with rate λ_j and stay in the system for a service time which is exponentially distributed with rate μ_j^{-1}, while holding $A_{i,j} \in \mathbb{N}$ units of each resource i. If we denote by $n(t) = (n_j(t), 1 \leq j \leq J) \in \mathbb{N}^J$ the vector describing the state of the system, where $n_j(t)$ is the number of clients then $(n(t)) = (n(t),\, t \geq 0)$ is a Markov process taking values in

$$\mathcal{S} = \left\{ n \in \mathbb{N}^J : A \cdot n \leq C \right\},$$

where $C = (C_i,\, 1 \leq i \leq I)$ and $A = (A_{i,j},\, 1 \leq i \leq I, 1 \leq j \leq J)$. Using the fact that the Poisson processes are all independent, one can write the detailed balance equations, for $n \in \mathbb{N}^J$,

$$\pi(n)\lambda_j = \pi(n + e_j)\mu_j(n_j + 1)\mathbb{1}_{\{n+e_j \in \mathcal{S}\}},$$

where $\pi = (\pi(n),\, n \in \mathcal{S})$ is the stationary distribution of the process $(n(t))$ and e_j is the j-th column of the canonical base of $\mathbb{N}^J$. Since this system is a truncated version of a system where $C_i = +\infty$, for $1 \leq i \leq I$, the following relation must hold

$$\pi(n_1, \ldots, n_J) \propto \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!},$$

where $\rho_j = \lambda_j/\mu_j$. Therefore, the invariant distribution is given by

$$\pi(n) = \frac{1}{Z} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \quad \text{with} \quad Z = \sum_{n \in \mathcal{S}} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!}. \tag{1.3}$$

Hence, the blocking probability of a job of type j is

$$\sum_{n \in \mathcal{B}_j} \pi(n) \quad \text{with} \quad \mathcal{B}_j = \{ n \in \mathcal{S} : n + e_j \notin \mathcal{S} \}.$$

Models similar to the current loss systems were developed as early as the beginning of the 20$^{\text{th}}$ century with the rise of the telecommunication technology (see Brockmeyer *et al.* [BHJ48], Frenkel [Fre74], Roberts [Rob81], Burman *et al.* [BLL84], Dziong and Roberts [DR87] and others). As the calculations involving the analysis of such systems are complicated (especially for obtaining the normalisation constant, see Louth *et al.* [LMK94]), much of the subsequent work has focused on expanding the mathematical toolbox (see or Choudhury *et al.* [CLW95] or Theberge *et al.* [TSM98]). It was particularly the publication of Kelly's papers [Kel86, Kel91] which through the introduction of a more simple and efficient framework for the study of the system's

equilibrium properties allowed the teletraffic community to gain a deeper understanding of the behaviour of loss systems. Recently, the framework of loss systems received renewed attention due to the rise of the Cloud Computing technology. The search for even more efficient calculation methods is subject of ongoing research (see Bonald and Virtamo [BV05], Jung *et al.* [JLS+08] and Tan *et al.* [TLX12a, TLX12b] for instance).

1.3 Mathematical Framework

Markovian Stochastic Processes

In this thesis, most of the queueing systems are analysed under Markovian assumptions, that is to say Poisson arrivals and Exponentially distributed service times. This is an important technical aspect of queueing modelling. In the case of $M/M/1$ queues, transitions from one state to another (increase or decrease of the number of customers queueing) happen with the same rate λ and μ independent of time. Due to this memorylessness, the system can be fully characterised by the number of clients in the queue at some time and the transition rates. The Markov processes in this document are Continuous-Time Markov Chains, usually analysed through the properties of its associated embedded Discrete-Time Markov Chain. The processes belong exclusively to the class of Markov jump processes. In these processes, the state might not change at all in a small time interval or, when a change does occur, the transition is discrete and of a relatively large magnitude (or "sharp") compared to the "smooth" changes such as observed in Brownian motion. We consider only irreducible and homogeneous (or at least partially homogeneous) Markov jump processes.

Kernel Methods

Let $(X(t))$ be a Markov process taking values on $\mathcal{S}$, whose $\mathcal{Q}$-matrix (or infinitesimal generator) is denote by $Q = (q(x, y),\, x, y \in \mathcal{S})$. When it exists, the stationary distribution of $(X(t))$ is denoted by $\pi = (\pi(x),\, x \in \mathcal{S})$ and the following relation holds for any function f with finite support on $\mathcal{S}$

$$\sum_{\substack{x,y\in\mathcal{S}\\ x\neq y}} \pi(x)q(x,y)(f(y)-f(x)) = 0.$$

In this thesis, the invariant distribution of some Markov processes are obtained from the relation of the associated generating function and its kernel, often making us resort to the toolbox of real and complex analysis i.e. the framework of Riemann-Hilbert problems. For example, if we have a process taking values in $\mathbb{N}$, the function $f_z(x) = z^x$, for $z \in D = \{z \in \mathbb{C} : |z| < 1\}$ plugged into the previous expression is associated with the generating function of the invariant distribution of such a process. Commonly, these objects are studied in the framework of functional analysis and are particularly recurrent in the field study of random walks. Consider the $M/M/1$ queue. If X is a random variable

with distribution π, the balancing equations of the embedded Markov process associated with the $M/M/1$ queue yield, for $\varphi(z) \overset{\text{def.}}{=} \mathbb{E}\left[z^X\right]$ and $|z| = D$,

$$\varphi(z)P(z) = \mu\pi(0)\left(z^{-1} - 1\right),$$

where $P(z) = \lambda(z-1) + \mu\left(z^{-1} - 1\right)$, whose roots are simply 1 and $\mu/\lambda \notin D$, considering $\lambda < \mu$. As the function $\varphi(z)$ is analytic for $z \in D$, the roots of $P(z)$ in D need to be zeros of the right hand side of the previous equation (in this example there are none). By definition, $\lim_{z\to 1}\varphi(z) = 1$, thus $\pi(0) = 1 - \rho$, where $\rho = \lambda/\mu$. Decomposing the previous expression in partial fractions, one has

$$\varphi(z) = \frac{\mu(1-\rho)\left(z^{-1}-1\right)}{P(z)} = \frac{1-\rho}{1-\rho z} = (1-\rho)\sum_{i=0}^{+\infty}\rho^i z^i$$

and deduces $\pi(i) = (1-\rho)\rho^i$, for $i \in \mathbb{N}$, which corresponds to the geometric distribution with parameter $0 < \rho < 1$. Similar methods from the framework of Complex Analysis are used throughout this thesis for the analysis of some processes that behave (at least locally) like random walks (see Cohen and Boxma [CB84], Chapter VIII of Asmussen [Asm03, p. 220], Fayolle *et al.* [FI79, FMM95, FIM17]). Furthermore, see the recent survey of El hady *et al.* [EhBN17] for a compilation of problems in queueing systems which are studied in the framework of functional analysis.

Ergodicity

In this document, the notion of *ergodicity* of Markov processes is very important as it plays a major role in the analyses of problems involving scaling techniques, particularly for the study of large loss networks in the framework proposed by Hunt and Kurtz [HK94]. The ergodicity of a Markov processes can be loosely understood as the equivalent concept to stability in dynamical systems, meaning that the system does not "explode" and that it has an unique invariant distribution.

In terms of queueing systems and networks, the easiest way to think of ergodicity is to consider that the queue lengths (or waiting times) do not go to infinity (or "explode" as mentioned before). For instance, in the $M/M/1$, the stability is given by the direct relation between arrival and service rates. The queue is stable if $\lambda < \mu$ or transient ("exploding") if $\lambda > \mu$ or null recurrent if $\lambda = \mu$. The last case is a critical case and those are not considered in this thesis.

As exposed in the next section, a frequent approach for the determination of the ergodicity conditions of Markov processes is the analysis of the fluid limits associate with the processes. A Markov process is ergodic if all of its fluid limits converge to zero in a finite time (see Corollary 9.8 of Robert [Rob03, p. 259]) and see the subsequent sections for a definition of fluid limits.

Scaling Techniques

Throughout this thesis, we have to deal with Markov processes whose invariant distribution may or may not exist or may be too complicated to be

calculated for obtaining practical information about the systems behaviour and performance. Therefore, we search for methods that facilitate the study of complex systems through simple models without neglecting however the main features of the system's functioning. For instance, in order to catch interesting aspects of some processes living in large state spaces, we may need to accelerate its time scale and shrink its state space, otherwise we cannot observe changes from one instant to another. The scaling considered in this thesis is of first order (functional law of large numbers). In the analysed cases, the renormalised version of a Markov process $(X(t))$ converges to a dynamic system $(x(t))$, which is the solution of some ordinary differential equation $\dot{x}(t) = f(x,t)$ related to the drifts of the original Markov process. Studying the equilibrium (and transient) properties of this dynamic system allows us to retrieve key information about the associated Markov process.

Fluid Limits A fluid limit is an asymptotic description of sample paths of a Markov process with a large initial state, after the renormalisation of space and time to capture first order phenomena in the evolution of the Markov process. The renormalisation consists in accelerating the time proportionally and contracting the space inversely proportionally to the size of the initial state of the process. Note that the renormalised process starts on the unit sphere of some normed vector space. We investigate the behaviour of the renormalised process when the norm of its initial state tends to infinity.

Fluid limits are a convenient tool for the study of Markov processes behaving like random walks (at least locally). Using the same framework deployed in Rybko and Stolyar [RS92], we are able to derive sufficient conditions for the process to be ergodic based on the contraction feature of its fluid limits. If the fluid limits of a Markov process are absorbed (reach and stay) at 0 in a deterministic finite time, then the process is ergodic. See Robert and Véber [RV15] for an interesting application of fluid limits in the analysis of processor sharing queueing systems.

Consider the stochastic process $(X(t))$ which takes values in the state space $\mathcal{S} \subseteq \mathbb{Z}^d$ with $X(0) = x$ and $d \in \mathbb{N}^*$. If the norm $\|x\| = |x_1| + \cdots + |x_d| = N$ then a fluid limit associated with the process $(X(t))$ is a stochastic process which is one of the possible limits of the process

$$\left(\overline{X}_N(t)\right) = \left(\frac{X\left(Nt\right)}{N}\right)$$

when N goes to infinity. Note that x/N is an element in the unit sphere of $\mathbb{R}^d$.

For a simple example, consider the $M/M/1$ queue with arrival rate λ and service rate μ. Using fluid limits, we are able to show in a straightforward manner that the stability condition for such a system is simply given by $\lambda < \mu$, or $\rho < 1$ for $\rho = \lambda/\mu$. Let us denote by $L(t)$ the number of customers in the queue at time $t \geq 0$. Therefore, $L(t)$ can be written as the solution of the following SDE

$$\mathrm{d}L(t) = \mathcal{N}_\lambda(\mathrm{d}t) - \mathcal{N}_\mu(\mathrm{d}t)\mathbb{1}_{\{L(t^-)\geq 0\}},$$

with $L(0) = N$. Consider now the renormalised process $\overline{L}_N(t) = L(Nt)/N$, for $t \geq 0$ and the hitting time

$$T_{0,N} = \inf \{t > 0 : L(t) = 0\}.$$

For $\tau_N = T_{0,N}/N$, the Strong Law of Large Numbers yields the identity $\lim_{N\to+\infty} \tau_N \overset{\text{dist.}}{=} 1/(\mu - \lambda)$. Thus, when N goes to infinity the following convergence holds for $t < \tau_N$

$$\overline{L}_N(t) = 1 + \frac{\mathcal{N}_\lambda]0, Nt] - \mathcal{N}_\mu]0, Nt]}{N} \overset{\text{dist.}}{=} 1 + (\lambda - \mu)t.$$

Using a simple coupling argument, we have that $\lim_{N\to+\infty} \overline{L}_N(t) = 0$, for $t > \tau_N$. Hence, the following convergence holds for $t \geq 0$

$$\lim_{N\to+\infty} \overline{L}_N(t) \overset{\text{dist.}}{=} (1 + (\lambda - \mu)t)^+.$$

were $(a)^+ = \max(0, a)$ for $a \in \mathbb{R}$. If $\lambda < \mu$, then the fluid limits of $(L(t))$ are absorbed at 0 and the queue is stable. Figure 1.4 illustrates the adherence of the fluid limits of a typical $M/M/1$ queue and the evolution of the number of customers in this system where $L(0) = N$. This example is particularly remarkable because it illustrates the linear relation between the initial number of customers N in a $M/M/1$ queue and the time it takes for this queue to become empty $T_{0,N}$.

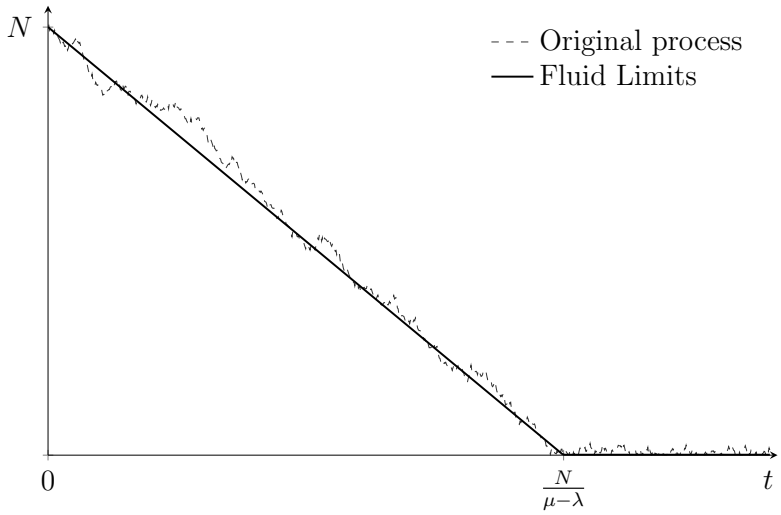


Figure 1.4 – Fluid limits of the $M/M/1$ queue

Kelly's Regime Another important scaling tecnique used in this thesis, it the scaling introduced by Kelly [Kel86, Kel91]. Despite the exactness and the revolutionary importance of Erlang's formula, Equation (1.1), and the product form expressions for loss networks, Equation (1.3), their practical use is limited to small capacity systems (see Ross and Tsang [RT90]). For example considering a single server with a installed capacity $C \in \mathbb{N}$ serving $J \in \mathbb{N}$ types

of customers. The computation used in Erlang's model grows very quickly with the capacity C as its related with its integer partition. Therefore, for systems disposing of large capacity, Erlang's approach becomes impractical.

Assume that the installed capacity in the system and the request arrival rate are sizeable i.e. of same order of magnitude. The server capacity is scaled up by a factor $N \in \mathbb{N}^*$ and the request arrival rate is scaled up accordingly, i.e.

$$\lambda_j \mapsto \lambda_j N \quad \text{and} \quad C \mapsto \lfloor cN \rfloor.$$

for $1 \leq j \leq J$. In the following, we consider two examples of the application of such technique to illustrate its power.

The M/M/N/N queue in Kelly's regime As a first illustration of the usage of Kelly's framework, consider a system equipped with N servers, where jobs arrive with rate λN (Poisson) and stay during a time exponentially distributed with mean $1/\mu$. Consider also that this system is under heavy traffic (or saturated), such that $\lambda > \mu$. Let $L_N(t)$ describe the number of jobs in the system at time $t \geq 0$ and $m_N(t) = N - L_N(t)$ the number of free servers at time t. We assume that the system is saturated at time $t = 0$, with $L_N(0) = N$. It is simple to see that the transition rates of process $(m_N(t))$, for $x \in \{0, \ldots, N\}$, are given by

$$\begin{cases} x \to x - 1 & N\lambda \mathbb{1}_{\{x>0\}} \\ x \to x + 1 & (N - x)\mu. \end{cases}$$

We note that this process lives in the time scale $t \mapsto tN$. We define the random walk $(\overline{m}(t))$ on the extended real line, such that, for $n \in \mathbb{N}$, the transitions $n \mapsto n + 1$ and $n \mapsto n - 1$ occur respectively at rates μ and λ (if $n > 0$). We denote by π the invariant distribution of $(\overline{m}(t))$, which is a geometric distribution with parameter μ/λ. Using some technical arguments, we can show that, when N goes to infinity, the process $(m_N(t/N))$ has the same invariant distribution π of $(\overline{m}(t))$, hence the probability of a customer being blocked is simply given by $\pi(0) = 1 - 1/\rho$, with $\rho = \lambda/\mu$.

Furthermore, it is simple to see that if $\rho < 1$, then after some (random) time the process $(m_N(t))/N$ converges to $1 - \rho$ when N goes to infinity. The system is therefore either underloaded ($\lambda < \mu$) and then, with probability 1, there is always free servers (no customer is lost with probability 1), or the system is overloaded ($\lambda > \mu$) and the blocking probability is simply given by $1 - \mu/\lambda$. This short example shows how Kelly's regime is a powerful tool for the characterisation of large scale systems. When N goes to infinity, the blocking probability in this system is simply $(1 - 1/\rho)^+$ for $\rho > 0$. Note that Kelly's method relies on capturing the deterministic behaviour of the system: a "flow" intensity $N\lambda$ of arriving jobs which are replaced by a flow of jobs that leave the system with rate $N\mu$. In this deterministic setting, the blocking rate in an interval $[0, t]$ is given by the ratio of the number of arriving customers, $N\lambda t$, minus the customers which are served directly, $N\mu t$, and the arriving customers, or equivalently, $(N\lambda t - N\mu t)/(N\lambda t) = 1 - \mu/\lambda$. Figure 1.5

illustrates the adherence of Erlang's formula, defined in Equation (1.1), and the limiting result as N gets larger.

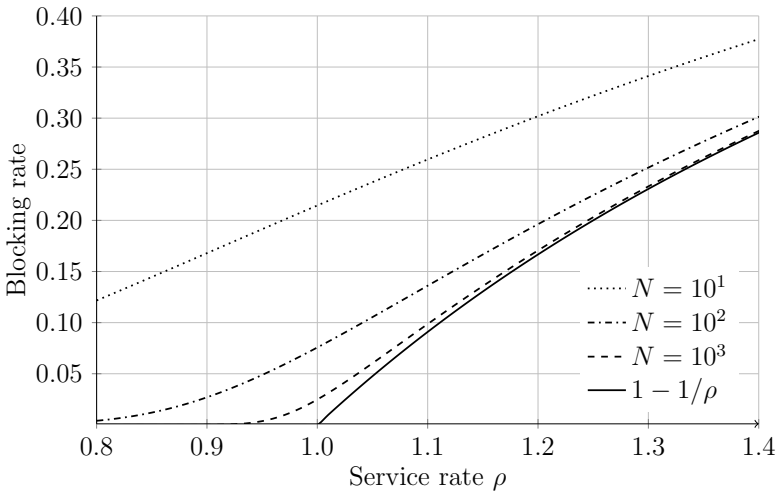


Figure 1.5 – Exact Erlang Formula and Large Scale Approximation

The M/M/N/N queue under heavy traffic and two types of customers We pass on to a second example illustrating Kelly's regime. Consider this time a system similar to the one presented before, equipped with $N \in \mathbb{N}$ servers and hosting two types of jobs $j = \{1, 2\}$. Jobs of type j arrive into the system with rate $\lambda_j N$ (Poisson) asking for a $A_j \in \mathbb{N}$ servers, where $A_1 = 1$, and $A_2 = a > 1$. If there are enough idle servers to fulfil the customer's requirements, the job is accepted and stays in the system during a certain service time which is exponentially distributed with mean $1/\mu_j$. Otherwise the job is rejected. We consider a saturated system, meaning that, under Kelly's regime [Kel86, Kel91], $\rho_1 + a\rho_2 > 1$, with $\rho_j = \lambda_j/\mu_j$ for $1 \leq j \leq 2$.

Using the same notation as before, let $L_{N,j}(t)$ describe the number of jobs of type j in the system and $m_N(t) = N - L_{N,1}(t) - aL_{N,2}(t)$ the number of free servers at time $t \geq 0$. The process $(\overline{L}_N(t)) = ((L_{N,1}(t)/N,\, L_{N,2}(t)/N))$ is given by, for $1 \leq j \leq 2$,

$$\overline{L}_{N,j}(t) = \overline{L}_{N,j}(0) + \overline{\mathcal{M}}_{N,j}(t) + \lambda_j \int_0^t \mathbb{1}_{\{m_N(s) \geq A_j\}}\, \mathrm{d}s - \mu_j \int_0^t \overline{L}_{N,j}(s)\, \mathrm{d}s,$$

where $(\overline{\mathcal{M}}_N(t)) = ((\overline{\mathcal{M}}_{N,1}(t),\, \overline{\mathcal{M}}_{N,2}(t)))$ is a Martingale, whose increasing predictable process at time t is given by, for $1 \leq j \leq 2$,

$$\langle \overline{\mathcal{M}}_{N,j} \rangle(t) = \frac{1}{N} \left(\lambda_j \int_0^t \mathbb{1}_{\{m_N(s) \geq A_j\}}\, \mathrm{d}s + \mu_j \int_0^t \overline{L}_{N,j}(s)\, \mathrm{d}s \right).$$

It can be shown that $(\overline{\mathcal{M}}_N(t))$ becomes negligible of order $1/\sqrt{N}$ as N goes to infinity.

The process $(\overline{L}_N(t))$ evolves in the $t \mapsto t$ time scale while the process $(m_N(t))$ evolves in the $t \mapsto Nt$ time scale. The system is governed by the

time-scale interaction between these two processes. In particular, the acceptance and rejection rates of jobs are determined by the equilibrium properties of the process $(m_N(t))$, which are influenced by the (renormalised) number of each type of customer. To study such a process, we introduce a random walk on $\mathbb{N}\cup\{+\infty\}$ as follows: for fixed $x \in \mathcal{D} = \{x \in \mathbb{R}_+^2 : x_1 + ax_2 \leq 1\}$, we define the process $(\overline{m}_x(t))$ on $\mathbb{N}\cup\{+\infty\}$, whose transitions occur from n to n' at rate $\mu_j x_j$ if $n' = n + A_j$, $\lambda_j \mathbb{1}_{\{n \geq A_j\}}$ if $n' = n - A_j$ and 0 otherwise. It is simple to determine that if

$$x \in \Delta = \{x \in \mathcal{D} : \mu_1 x_1 + a\mu_2 x_2 < \lambda_1 + a\lambda_2, x_1 + ax_2 = 1\}$$

then the Markov process $(\overline{m}_x(t))$ is ergodic on $\mathbb{N}$. In this case its unique invariant distribution is denoted by $\pi_x = (\pi_x(n), n \in \mathbb{N})$ (see for instance Bean *et al.* [BGZ95] and Fricker *et al.* [FRT01, FRT03]). For convenience, we extend the definition of π_x, such that if $x \in \mathcal{D} \setminus \Delta$, then the unique invariant distribution of $(\overline{m}_x(t))$ on $\mathbb{N}\cup\{+\infty\}$ is a Dirac mass at infinity $\delta_{+\infty}$.

Using the same method as Hunt and Kurtz [HK94], one gets the analogue of Theorem 3 of this reference. For any function f with finite support on $\mathbb{N}$ we have the following convergence

$$\lim_{N\to+\infty} \left(\frac{L_{N,1}(t)}{N}, \frac{L_{N,2}(t)}{N}, \int_0^t f(m_N(s))\,\mathrm{d}s \right) \overset{\text{dist.}}{=} \left(l_1(t), l_2(t), \int_0^t \int_{\mathbb{N}} f(u)\pi_{l(s)}(\mathrm{d}u)\,\mathrm{d}s \right),$$

where $l_1(t)$ and $l_2(t)$ satisfy the following differential equations

$$\dot{l}_j(t) = \begin{cases} -\mu_j l_j(t) + \lambda_j & \text{if } l_1 + al_2 < 1, \\ -\mu_j l_j(t) + \lambda_j \pi_{l(t)}([A_j, \ldots, +\infty[) & \text{if } l_1 + al_2 = 1, \end{cases} \tag{1.4}$$

with $\lim_{N\to+\infty} \overline{L}_{N,j}(0) = l_j(0)$ for $1 \leq j \leq 2$. Note that we have indeed a convergence in distribution since, for any $x \in \mathcal{D}$, $(\overline{m}_x(t))$ has exactly one invariant distribution. For this particular loss system under heavy traffic, there are no fixed points $x \in \mathbb{R}_+^2$ such that $x_1 + ax_2 < 1$ and $x_j = \rho_j$. Thus, if this system has an equilibrium point, it lays on the frontier $x_1 + ax_2 = 1$ — which sounds intuitive for a saturated system. Using the balance equations of $(\overline{m}_x(t))$, it is simple to see that if $x_j = \rho_j \beta^{A_j}$, for $1 \leq j \leq 2$, then π_x is a geometric distribution with parameter $0 < \beta < 1$. Finally, if β is the unique solution in the interval $(0, 1)$ of

$$\rho_1 \beta + a\rho_2 \beta^a = 1, \tag{1.5}$$

such that $x^\star = (\rho_1\beta, \rho_2\beta^a)$ satisfies Equation (1.4), then $x^\star$ is the unique Equilibrium point of $(l(t))$ (located in the frontier $x_1 + ax_2 = 1$). Equation (1.5) is referred as Kelly's Fixed Point Equation (see Kelly Kelly1991).

Coexistence of distinct Time Scales While investigating large circuit switching systems ([Kel86, Kel91]), Kelly observed that the macroscopic (and

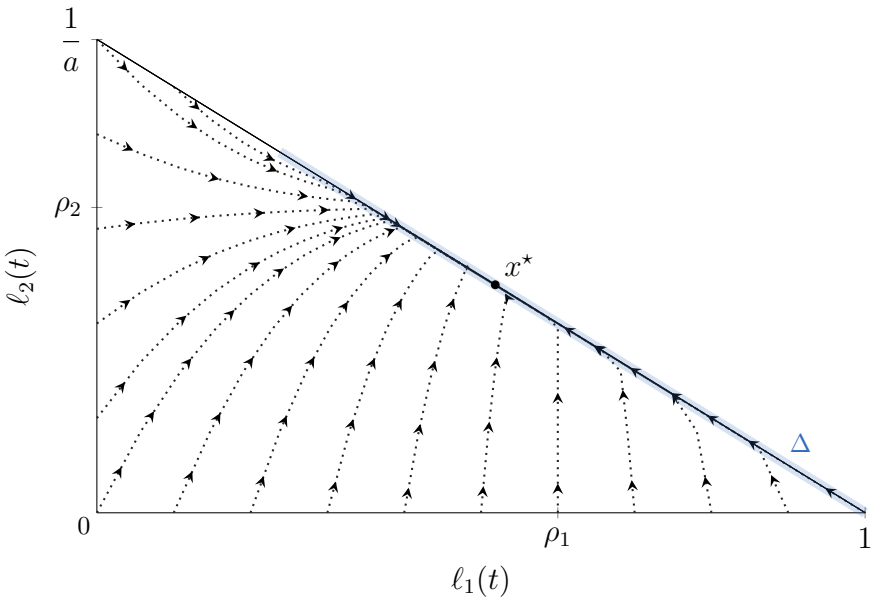


Figure 1.6 – Limiting sample paths of an overloaded loss system
The Equilibrium point $x^\star$ is determined in Equation (1.5). Note that if $l(t_0) \in \Delta$, for some $t_0 \geq 0$, then $l(t) \in \Delta$ for any $t > t_0$.

slow) behaviour of these systems is intrinsically related with the local equilibrium of the microscopic (and fast) process with the parameters of the fast process depending on the slow process. For the loss networks, the process $(\overline{L}_N(t))$ evolves in the $t \mapsto t$ time scale, since its transitions $+e_j$ occur at rate λ_j and $-e_j$ at rate $\mu_j \overline{L}_{N,j}(t)$, for $1 \leq j \leq 2$, whereas the process $(m_N(t))$ evolves in the much faster $t \mapsto Nt$ time scale, with transitions $-A_j$ and $+A_j$ occurring respectively at rates $\mu_j L_{N,j}$ and $N\lambda_j \mathbb{1}_{\{m \geq A_j\}}$, for $1 \leq j \leq 2$. When the scaling factor N is sufficient large, we observe the separation of these time scales which then coexist in a separate manner but still influence each other. The interplay of these processes is based on the Stochastic Averaging Principle (SAP). The dynamics of the limiting trajectories of the slow process are governed by the local equilibrium of the (infinitely) faster process, whose transitions depend on the slow process, seen as fixed by the fast process.

The averaging phenomenon was already established in the context of deterministic dynamical systems (see for instance Chapter 4 of Guckenheimer and Holmes [GH83, p. 168]). Hunt and Kurtz [HK94] further develop a convenient framework for the analysis of time-scale interaction in the framework of loss systems. However, the usage of this a framework is not straightforward in stochastic contexts due to the difficulty of determining the regularity of the properties of the invariant distributions of the fast processes.

1.4 Presentation of the following chapters

The concepts, techniques and models presented in this introduction are regularly used in the following chapters. The systems analysed in this thesis

are investigated under Markovian assumptions, i.e. Poison arrivals and exponential service times. The scaling techniques seen before are widely used in every chapter in the search for (good) approximations of the (macro- and microscopic) behaviour of these systems as well as the derivation of ergodicity conditions of some of the processes. The invariant distribution of the key-processes are often obtained using the framework of kernel problems and the toolbox of real and complex analysis. This thesis provides additions to the framework of the stochastic analysis of large scale systems through the exploration of Markovian models and limiting properties of large scale systems. The aim is to address the new perspectives and problems of policy design due to the changes in the architecture of Cloud Computing systems and the evolution of related technologies.

The four chapters of this thesis are somewhat related (at least in the methods deployed in their study). Chapter 2 is the most isolated from the others, since it focuses on a resource sharing policy for Video on Demand systems. The subsequent three chapters, focus on cooperative schemes between decentralised (smaller) data centres. In Chapter 3 an offloading policy is study in the context of multi-resource systems. In Chapter 4 and Chapter 5 two related offloading policies are investigated.

Allocation of resources with downgrading

In collaboration with: Christine Fricker, Fabrice Guillemin and Philippe Robert
Published on: Advances in Applied Probability. vol. 49 iss. 2
doi:10.1017/apr.2017.15

In the second chapter, we offer some insights about the introduction of (scaled) threshold in adaptation policies for resource sharing in Video on Demand systems. We consider a server with large capacity delivering video files encoded in various resolutions. We assume that the system is under saturation in the sense that the total demand exceeds the server capacity C. In such case, requests may be rejected. For the policies considered in this chapter, instead of rejecting a video request, it is downgraded. When the occupancy of the server is above some fixed threshold $C_0 < C$, the server delivers the video at a minimal bit rate. For these policies, request blocking is thus replaced with bit rate adaptation.

We assume that customers request video files encoded at various rates, say, A_j for $j = 1, \ldots, J$, with $1 = A_1 < A_2 < \cdots < A_J$. Jobs of class $j \in \{1, \ldots, J\}$ require bit rate A_j, arrive according to a Poisson process with rate λ_j and have an exponentially distributed transmission time with rate μ_j. We denote by $L_j(t)$ the number of jobs of type j in the system at time t, and $(L(t)) = ((L_j(t), 1 \leq j \leq J))$ the Markov process which describes the state of such a system. The invariant distribution of $(L(t))$ does not have an explicit expression. Note that this system is closely related with the loss systems introduced previously, thus, to study this allocation scheme, a scaling approach is used. It is assumed that the server capacity is very large, namely scaled up by a factor N. The bit rate adaptation threshold and the request

arrival rates are scaled up accordingly, i.e.

$$\begin{cases} \lambda_j \mapsto \lambda_j N & 1 \leq j \leq J, \\ C_0 \mapsto c_0 N & \text{and} \quad C \mapsto cN. \end{cases}$$

In the scaled version of the system, the process associated with the occupancy of the server (around the threshold) is denoted by $(m^N(t)) = (\sum_{j=1}^{J} A_j L_j^N(t) - C_0)$ and $L_j^N(t)$ denotes the number of jobs of type j in the system at time t. We show that, when the scaling factor N goes to infinity, the process $(m^N(t))$ converges to the random walk defined as follows: for λ, μ, $\ell \in \mathbb{R}_+^J$, let $(m_\ell(t))$ be the Markov jump process on $\mathbb{Z}$ whose non-negative elements of its $\mathcal{Q}$-matrix, $Q_\ell = (q_\ell(n, n'),\, n, n' \in \mathbb{Z})$, are given by

$$q_\ell(n, n') = \begin{cases} \Lambda & \text{if } n' = n + 1 \quad \text{and } n \geq 0, \\ \lambda_j & \text{if } n' = n + A_j \quad \text{and } n < 0, \\ \mu_j \ell_j & \text{if } n' = n - A_j, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_J$ and $1 = A_1 < A_2 < \cdots < A_J$. Note $n \geq 0$ represents the states where the link occupation rate is above the threshold c_0, and in this case, all the arriving requests are treated as class 1 and otherwise, if $n < 0$, the customers are threaded as they request.

Using some methods from real and complex analysis related to Wiener-Hopf factorisation, we are able to obtain the unique stationary invariant distribution of such a random walk when it exists. We can derive an asymptotic expression of the key performance measure of such a policy, since, in the limiting scenario, the stationary probability that a request is transmitted at requested bitrate is given by the probability that the process $(m^N(t))$ is on the negative half line $\mathbb{Z}_-^*$ and ℓ is the equilibrium point of this system.

Our main result shows that, if c_0 is conveniently chosen, such that

$$\frac{\Lambda}{\mu_1} < c_0 < \sum_{j=1}^{J} A_j \frac{\lambda_j}{\mu_j},$$

then, as N goes to infinity,

— the equilibrium probability of rejecting a job converges to 0, and

— the equilibrium probability of accepting a job at requested bitrate converges to

$$\pi^- \overset{\text{def.}}{=} (c_0 - \Lambda/\mu_1) \Bigg/ \left(\sum_{j=1}^{J} \frac{\lambda_j}{\mu_j} A_j - \Lambda/\mu_1 \right).$$

We compare this policy with the standard loss system (where no admission control is performed), and show that for systems where the access refusal is very expensive, our policy is an effective method for resource allocation between clients with very diversified demands on terms of servers. As we

show, under our scheme, the system can run without rejecting customers (with probability 1 when N goes to infinity) at the price of downgrading some jobs to the minimum available service quality.

In this chapter, one of the main technical difficulties is to prove the convergence of invariant distributions of the process $(m^N(t))$. We need to show simultaneously that the process $(L^N(t))/N$ converges in distribution to some dynamical system, which is governed by the equilibrium properties of the process $(m^N(t))$, while the invariant distribution of $(m^N(t))$ is influenced by $(L^N(t))/N$. However, the regularity properties of the process $(m^N(t))$ are not straight forward, since the invariant distribution of such a process depends on where the process $(L^N(t))/N$ is. Therefore, we need to show additionally that in the dynamical system's path (towards its equilibrium point), after some random time, the process $(m^N(t))$ remains in the region where it is ergodic.

Cooperation mechanisms in multi-resource services

In collaboration with: Christine Fricker, Fabrice Guillemin and Philippe Robert

In the third chapter, we investigate the deployment of a local state-aware policy for load balancing in the multi-resource framework. We consider a Cloud Computing environment where virtual machines (or jobs) are instantiated upon the allocation of `GB RAM` and `CPU core`, denoted resource $i \in \{1, 2\}$, respectively. The system is composed of two parallel data centres denoted $k \in \{1, 2\}$. Each data centre k is equipped with some large (but finite) amount $\mathcal{R}_k$ of `GB RAM` and $\mathcal{C}_k$ of `CPU cores`. To capture the particular aspect of resource specialised (unbalanced) virtual machines, we consider two types of jobs arriving to the system, denoted type $j \in \{1, 2\}$. Jobs of type 1 require $R > 1$ `GB RAM` and 1 `CPU core`, whereas jobs of type 2 require 1 `GB RAM` and $C > 1$ `CPU core`. Requests of type j arrive at data centre k according to a Poisson process with rate $\lambda_{j,k}$, for $1 \leq j, k \leq 2$. The duration of the service of type j jobs is exponentially distributed with mean $1/\mu_j$. We suppose that on the aggregate level the system is properly dimensioned, i.e.

$$\frac{R(\lambda_{1,1} + \lambda_{1,2})}{\mu_1} + \frac{\lambda_{2,1} + \lambda_{2,2}}{\mu_2} < \mathcal{R}_1 + \mathcal{R}_2 \quad \text{and}$$
$$\frac{\lambda_{1,1} + \lambda_{1,2}}{\mu_1} + \frac{C(\lambda_{2,1} + \lambda_{2,2})}{\mu_2} < \mathcal{C}_1 + \mathcal{C}_2,$$

but that locally each data centre has one resource which is saturated (with loss of generality, we consider resource p is saturated at data centre p, for $1 \leq p \leq 2$), meaning

$$\frac{\lambda_{1,1}}{\mu_1} R + \frac{\lambda_{2,1}}{\mu_2} > \mathcal{R}_1 \quad \text{and} \quad \frac{\lambda_{1,2}}{\mu_1} + \frac{\lambda_{2,2}}{\mu_2} C > \mathcal{C}_2.$$

Note that this case is the most interesting and challenging scenario to be investigated. The other configurations fall back into simpler models, which can be analysed using the classical toolbox of queueing theory. We propose

a threshold policy to handle local saturation and control the offloading of VM from one data centre to another. Thresholds are chosen to anticipate sufficiently in advance potential shortages of any resource in any data centre. The idea is to buffer some resources to avoid the complete jamming of a processing facility due to the lack of either resource. Jobs with the largest demand of the locally saturated resource are systematically forwarded to the other data centre when the threshold level is reached. Other requests are forwarded only if they cannot be accommodated locally. A threshold is set for the most demanded resource at each data centre, selected by the relation between its local load and capacity. The study is therefore restricted to the implementation of thresholds $0 < \delta_p < 1$ in data centre p to avoid exhaustion of resource p by offloading the jobs of type p, for $1 \leq p \leq 2$.

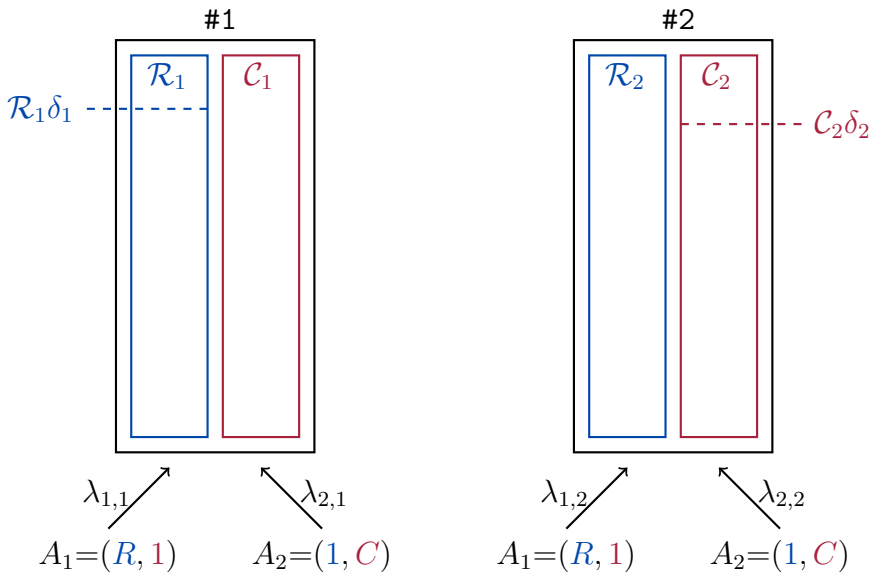


Figure 1.7 – System configuration

Jobs of type j arriving at data centre k with rate $\lambda_{j,k}$ requesting A_j units of the resources. Data centre k is equipped with a finite capacity or each resource, namely $\mathcal{R}_k$ `GB RAM` and $\mathcal{C}_k$ `CPU core`, for $1 \leq k \leq 2$. Clients of type p are forwarded to data centre $3-p$ when resource p is being used above the threshold $0 < \delta_p < 1$, for $1 \leq p \leq 2$.

Using Kelly's scaling, we are able to express the performance of the system in terms of the invariant distribution of the random walk on the plane $\mathbb{Z}^2$ in the Markovian environment defined as follows: for fixed $x = (x_{j,k},\, 1 \leq j,k \leq 2) \in \mathbb{R}_+^{2\times 2}$, let $(U_x(t)) = ((U_{x,h}(t), U_{x,v}(t)))$ be a semi-homogeneous random walk on $\mathbb{Z}^2$ whose transitions occur from $n = (n_1, n_2)$ to $n + b$ at rate

$$\begin{cases} \lambda_{1,1}\mathbb{1}_{\{n_1<0\}} & \text{if} \quad b=(R,0), \\ \lambda_{2,1}+\lambda_{2,2}\mathbb{1}_{\{n_2\geq 0\}} & \text{if} \quad b=(1,0), \\ \lambda_{1,2}+\lambda_{1,1}\mathbb{1}_{\{n_1\geq 0\}} & \text{if} \quad b=(0,1), \\ \lambda_{2,2}\mathbb{1}_{\{n_2<0\}} & \text{if} \quad b=(0,C) \end{cases} \quad \text{and} \quad \begin{cases} \mu_1 x_{1,1} & \text{if} \quad b=-(R,0), \\ \mu_2 x_{2,1} & \text{if} \quad b=-(1,0), \\ \mu_1 x_{1,2} & \text{if} \quad b=-(0,1), \\ \mu_2 x_{2,2} & \text{if} \quad b=-(0,C). \end{cases}$$

Note that $n_1 \geq 0$ represents the fact that the utilisation rate of resource 1 in data centre 1 is above the threshold δ_1 (the same for n_2 regarding resource 2 at data centre 2). We show that when N goes to infinity, the probability of a job being forwarded from one data centre to another given as a function of the stationary invariant distribution of this random walk i.e. the probability of this random walk to be in wither half plane determines the offloading rates between the servers. Using some fluid limits, we are able to derive the ergodicity conditions for such random walk and, resorting to the framework of kernel methods, we are able to obtain the probability which this random walk stays on each half-plane.

As main result, we show that when the scale of this system is large, we have that

— the probability of rejecting jobs in both data centres converges to 0;

— the probabilities of forwarding a type 2 job from data centre 1 and type 1 job from data centre 2 converge to 0;

— the probabilities of forwarding a job of type 1 from data centre 1 and type 2 from data centre 2 converge to

$$\gamma_1^\star(\delta) \stackrel{\text{def.}}{=} \frac{\mu_1}{(RC-1)\lambda_{1,1}}\left[\mathcal{R}_1\delta_1 C + \mathcal{C}_2\delta_2 - \frac{(\lambda_{1,1}+\lambda_{1,2})}{\mu_1} - \frac{C\,(\lambda_{2,1}+\lambda_{2,2})}{\mu_2}\right],$$

and

$$\gamma_2^\star(\delta) \stackrel{\text{def.}}{=} \frac{\mu_2}{(RC-1)\lambda_{2,2}}\left[\mathcal{R}_1\delta_1 + \mathcal{C}_2\delta_2 R - \frac{R\,(\lambda_{1,1}+\lambda_{1,2})}{\mu_1} - \frac{(\lambda_{2,1}+\lambda_{2,2})}{\mu_2}\right].$$

The multi-resource characteristic of this system brings some extra difficulties for its analysis. In this chapter, we had to resort to some heuristic reasoning (which are then formally proved later in the chapter) in the analysis of the threshold parametrisation. Also, we have to rely on the fluid limits analysis to derive the ergodicity conditions of the process $(U_x(t))$.

The results presented in this chapter are a first step towards the understanding of cooperation in multi-resource structures, still barely explored in literature – specially in the framework of stochastically modelling. The proposed policy work on a decentralised manner, a key aspect to be observed in the coming organisation of Cloud structures. We are able derive optimal threshold parameters, allowing the distributed Cloud Computing systems to meet the efficiency of a centralised system of equivalent total capacity serving the combined flows of requests.

Analysis of an offloading scheme for data centres in the framework of fog computing

In collaboration with: Christine Fricker, Fabrice Guillemin and Philippe Robert
Published on: ACM ToMPECS. vol. 1 iss. 4
doi:10.1145/2950047

In the fourth chapter, we analyse, in the context of fog computing, a simple case when data centres are installed at the edge of the network and assume

that if a request arrives at an data centre whose capacity is currently depleted, then it is forwarded to a neighbouring data centre with some probability (in a logical ring). We consider that data centres have a large number of servers and that traffic at some data centres is causing saturation. In this case the other data centres may help to cope with this saturation by accepting some of the rejected requests. Our aim is to qualitatively estimate the gain achieved via cooperation between neighbouring data centres. The analysis is focused on the behaviour of some pair of neighbouring data centres. See Figure 1.8.

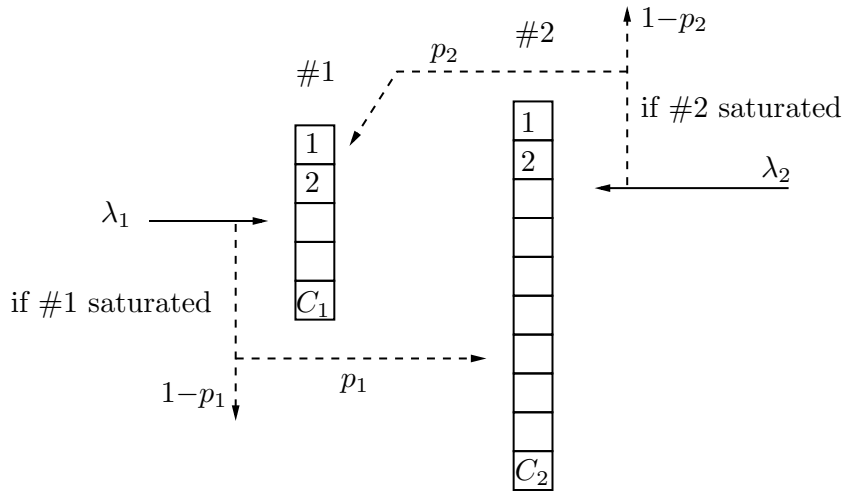


Figure 1.8 – System configuration ("probabilistic" Offloading)

Implementation of a probabilistic offloading scheme between two data centres in the framework of Fog Computing. In data centre p, the capacity is denoted by C_p and the arrival rates by λ_p, for $1 \leq p \leq 2$.

After proving some convergence results related to the scaling limits of loss systems for the process describing the number of free servers at both data centres, we show that the performance of the system can be expressed in terms of the invariant distribution of a reflected random walk on the positive quarter plane. For fixed $l = (l_1, l_2) \in \mathbb{R}_+^2$, we define the random walk $(\overline{m}_l(t))$ on $(\mathbb{N} \cup \{+\infty\})^2$ as follows: the transition from (m_1, m_2) to $(m_1 + a, m_2 + b)$ occurs at rate

$$\begin{cases} \mu_1 l_1 & \text{if } (a,b) = (1,0), \\ \mu_2 l_2 & \text{if } (a,b) = (0,1), \\ \lambda_1 + p_2 \lambda_2 \mathbb{1}_{\{m_2=0\}} & \text{if } (a,b) = (-1,0) \text{ and } m_1 > 0, \\ \lambda_2 + p_1 \lambda_1 \mathbb{1}_{\{m_1=0\}} & \text{if } (a,b) = (0,-1) \text{ and } m_2 > 0, \end{cases}$$

for $(m_1, m_2) \in (\mathbb{N} \cup \{+\infty\})^2$ with the convention that $+\infty \pm x = +\infty$ for $x \in \mathbb{N}$. We derive the ergodicity conditions for such a random walk using fluid limits. We are able obtain explicit formulas for the blocking rates of the system using and further developing existing results from the technical literature concerning random walks in the positive quarter plane, using some kernel methods. Numerical applications illustrate the power of the cooperation scheme, partic-

ularly focusing on the ability of such a policy to manage offloading systems, where one data centre is considerably larger than the other(s).

Analysis of a trunk reservation policy in the framework of Fog Computing

In collaboration with: Fabrice Guillemin
Preprint on arXiv: 1604.03823[GT16]

We analyse a system similar to the one in the previous chapter. The system is composed of two data centres with limited capacity in terms of servers. When one request for a single server is blocked at the first data centre, this request is forwarded to the second one. To protect the requests originally assigned to the second data centre, a trunk reservation policy is introduced (i.e., a redirected request is accepted only if there is a sufficient large number of free servers at the second data centre). After rescaling the

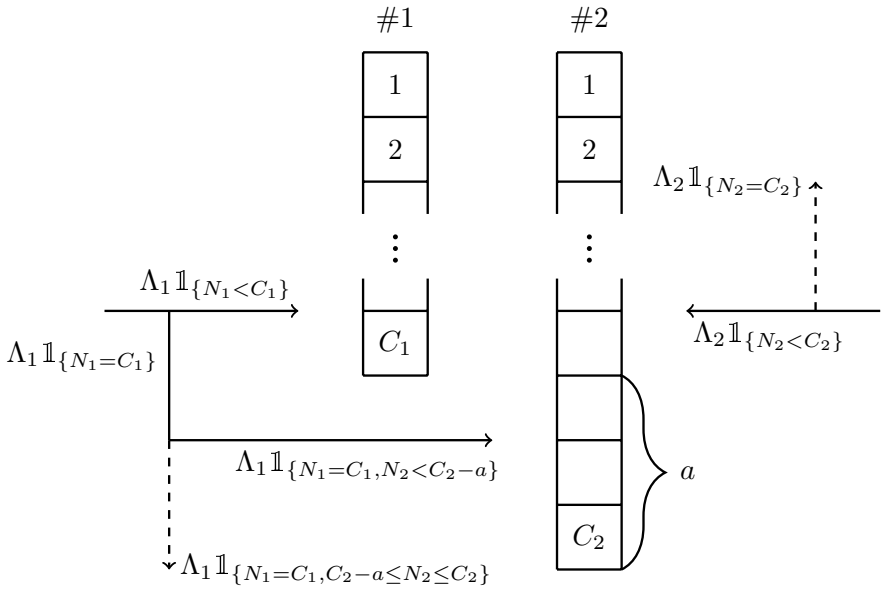


Figure 1.9 – System configuration ("truncated" offloading)

Implementation of a systematic offloading scheme between two data centres in the framework of Fog Computing, with trunk protection of a servers in the second data centre. In data centre p the number of requests hosted at p is denoted by N_p, the capacity by C_p and the arrival rates by Λ_p, for $1 \leq p \leq 2$.

system, assuming that there are many servers in both data centres and that request arrival rates are high, we show that the system's performance is related to a random walk on the quarter plane. For fixed $l = (l_1, l_2) \in \mathbb{R}_+^2$, we define the random walk $(\overline{m}_l(t))$ on $(\mathbb{N} \cup \{+\infty\})^2$ as follows: the transition from

(m_1, m_2) to $(m_1 + k, m_2 + \ell)$ occurs at rate

$$\begin{cases} \mu_1 l_1 & \text{if } (k,\ell) = (1,0), \\ \mu_2 l_2 & \text{if } (k,\ell) = (0,1), \\ \lambda_1 \mathbb{1}_{\{m_1>0\}} & \text{if } (k,\ell) = (-1,0), \\ \lambda_2 \mathbb{1}_{\{m_2>0\}} + \lambda_1 \mathbb{1}_{\{m_1=0,m_2>a\}} & \text{if } (k,\ell) = (0,-1), \end{cases}$$

where a is the threshold parameter. Note that this random walk has the particularity of having non-constant reflecting conditions on one boundary of the quarter plane. After some convergence results, we derive the ergodicity conditions for such a random walk. The main challenge of this chapter is the technical study of the kernel problem in the derivation of the invariant distribution of the aforementioned random walk. Contrary to usual reflected random walks, we have to determine three unknown functions, one polynomial and two infinite generating functions, in order to compute the stationary distribution of the random walk. We show that the coefficients of the polynomial are solutions to a linear system. After solving this linear system, we are able to compute the two other unknown functions. The blocking probabilities at both data centres are then derived.

Numerical experiments are performed to estimate the gain achieved by the trunk reservation policy, contrasting with the model of the previous chapter as well as the scenario where both data centres operate independently from each other, showing the trade-off between both policies.

Chapter 2

Allocation Schemes of Resources with Downgrading

2.1 Introduction

Video streaming applications have become over the past few years the dominant applications in the Internet and generate the prevalent part of traffic in today's IP networks; see for instance Guillemin *et al.* [GHM13] for an illustration of the application breakdown in a commercial IP backbone network. Video files are currently downloaded by customers from large data centres, like Google's data centres for YouTube files. In the future, it is very likely that video files will be delivered by smaller data centres located closer to end users, for instance cache servers disseminated in a national network. It is worth noting that as shown in Guillemin *et al.* [GKMS13], caching is a very efficient solution for YouTube traffic. While this solution can improve performances by reducing delays, the limited capacity of those servers in terms of bandwidth and computing can cause overload.

One possibility to reduce overload is to use bit rate adaptation. Video files can indeed be encoded at various bit rates (e.g., small and high definition video). If a node cannot serve a file at a high bit rate, then the video can be transmitted at a smaller rate. It is remarkable that video bit rate adaptation has become very popular in the past few years with the specification of MPEG-DASH standard where it is possible to downgrade the quality of a given transmission, see Schwarz *et al.* [SMW07], Vadlakonda *et al.* [VCH+10], Sieber *et al.* [SHZ+13] and Añorga *et al.* [AAS+15] (see also Fricker *et al.* [FGRT16b]). Adaptive streaming is also frequently used in mobile networks where bandwidth is highly varying. In this chapter, we investigate the effect of bit rate adaptation in a node under saturation.

Downgrading Policy

We assume that customers request video files encoded at various rates, say, A_j for $1 \leq j \leq J$, with $1 = A_1 < A_2 < \cdots < A_J$. Jobs of class $j \in \{1, \ldots, J\}$ require bit rate A_j. The total capacity of the communication link is C. If $\ell = (\ell_j)$ is the state of the network at some moment, with ℓ_j being the number

of class j jobs, the quantity $\langle A, \ell \rangle = A_1\ell_1 + \cdots + A_J\ell_J$ has to be less than C. The quantity $\langle A, \ell \rangle$ is defined as the *occupancy* of the link. The algorithm has a parameter $C_0 < C$ and works as follows: If there is an arrival of a job of class $1 \leq j_0 \leq J$, if

— $\langle A, \ell \rangle < C_0$ then the job is accepted;

— $C_0 \leq \langle A, \ell \rangle < C$ then the job is accepted but as a class 1 job, i.e. it has an allocated bit rate of $A_1 = 1$ and service rate μ_1;

— $\langle A, \ell \rangle = C$, the job is rejected.

For $1 \leq j \leq J$, jobs of class j arrive according to a Poisson process with rate λ_j and have an exponentially distributed transmission time with rate μ_j. Additionally, it is assumed that

$$\mu_1 \leq \min(\mu_j, 2 \leq j \leq J).$$

A Scaling Approach

To study this allocation scheme, a scaling approach is used. It is assumed that the server capacity is very large, namely scaled up by a factor N. The bit rate adaptation threshold and the request arrival rates are scaled up accordingly, i.e.

$$\begin{cases} \lambda_j \mapsto \lambda_j N, \quad 1 \leq j \leq J, \\ C_0 \mapsto c_0 N \text{ and } C \mapsto cN. \end{cases} \tag{2.1}$$

Performances of the algorithm. Our main result shows that, for the downgrading policy and if c_0 is chosen conveniently, then

1. the equilibrium probability of rejecting a job converges to 0 as N goes to infinity;
2. the equilibrium probability of accepting a job without downgrading it converges to

$$\pi^- \overset{\text{def.}}{=} \left(c_0\mu_1 - \sum_{j=1}^{J} \lambda_j \right) \Bigg/ \left(\mu_1 \sum_{j=1}^{J} \frac{\lambda_j}{\mu_j} A_j - \sum_{j=1}^{J} \lambda_j \right),$$

as N goes to infinity. See Theorem 2.2 and Corollary 2.1.

The above formula gives an explicit expression of the success rate of this allocation mechanism. The quantity $1 - \pi^-$, the probability of downgrading requests, can be seen as the "price" of the algorithm to avoid rejecting jobs.

The scaling (2.1) has been introduced by Kelly to study loss networks (see Kelly [Kel86, Kel91]). The transient behaviour of these networks under this scaling has been analysed by Hunt and Kurtz [HK94]. This last reference provides essentially a framework to establish convenient convergence theorems involving stochastic averaging principles. This line of research has been developed in the 1990's to study uncontrolled loss networks where a request is rejected as soon as its demand cannot be accepted.

When the demand can be adapted to the state of the network, for controlled loss networks, several (scarce) examples have been also analysed during that

period of time. One can mention Bean *et al.* [BGZ95, BGZ97], Zachary and Ziedins [ZZ02, ZZ11] for example. Our model can be seen as a "controlled" loss networks instead of a pure loss network. Controlled loss networks may have mechanisms such as trunk reservation or may allocate requests according to some complicated schemes depending on the state of the network. In our case, the capacity requirements of requests are modified when the network is in a "congested" state.

Contrary to classical uncontrolled loss networks, as it will be seen, the Markov process associated to the evolution of the vector of the number of jobs for each class is not reversible. Additionally, the invariant distribution of this process does not seem to have a closed form expression. Kelly's approach [Kel86] is based on an optimisation problem, it cannot be used in our case to get an asymptotic expression of some characteristics at equilibrium. For this reason, the equilibrium behaviour of these policies is investigated in a two step process:

1. Transient Analysis. We investigate the asymptotic behaviour of some characteristics of the process on a finite time interval when the scaling parameter N goes to infinity.
2. Equilibrium. The stability properties of the limiting process are analysed, we prove that the equilibrium of the system for a fixed N converges to the equilibrium of the limiting process.

For our model, the transient analysis involves the *explicit* representation of the invariant distribution of a specific class of Markov processes. It is obtained with complex analysis arguments. As it will be seen, this representation plays an important role in the analysis of the asymptotic behaviour at equilibrium.

It should be noted that related models have recently been introduced to investigate resource allocation in a cloud computing environment where the nodes receive requests of several types of resources. We believe that this domain will receive a renewed attention in the coming years. See Stolyar [Sto13, Sto15] and Fricker *et al.* [FGRT16a] for example. In some way one could say that the loss networks are back and this is also a motivation of this chapter to shed some light on the methods that can be used to study these systems.

Outline of the chapter

We consider a system in overload. Because of bit rate adaptation, requests may be downgraded but not systematically rejected as in a pure loss system. As it will be seen, the stability properties of this algorithm are linked to the behaviour of a Markov process associated to the occupation of the link. Under exponential assumptions for inter-arrival and service times, this process turns out to be, after rescaling by a large parameter N, a bilateral random walk instead of a reflected random walk as in the case of loss networks. Using complex analysis methods, an explicit expression of the invariant distribution of this random walk is obtained. With this result, the asymptotic expression of the probability that, at equilibrium, a job is transmitted at its requested rate (and therefore does not experience a bit rate adaptation) is derived.

This chapter is organised as follows: In Section 2.2, we present the model used to study the network under some saturation condition. Convergence results when the scaling factor N tends to infinity are proved in Section 2.3. The invariant distribution of a limiting process associated to the occupation of the link is computed in Section 2.4 by means of complex analysis techniques. Applications are discussed in Section 2.5.

2.2 Model description

One considers a service system where J classes of requests arrive at a server with bandwidth/capacity C. Requests of class j, $1 \leq j \leq J$, arrive according to a Poisson process $\mathcal{N}_{\lambda_j}$ with rate λ_j. A class j request has a bandwidth requirement of A_j units for a duration of time which is exponentially distributed with parameter μ_j. For the systems investigated in this chapter, there is no buffering, requests have to be processed at their arrival otherwise they are rejected. Without any flexibility on the resource allocation, this is a classical loss network with one link. See Kelly [Kel91] for example.

This chapter investigates allocation schemes which consist of reducing the bandwidth allocation of arriving requests to a minimal value when the link has a high level of congestion. In other words the service is downgraded for new requests arriving during a saturation phase. If the system is correctly designed, it will reduce significantly the fraction of rejected transmissions and, hopefully, few jobs will in fact experience downgrading.

2.2.1 Downgrading policy $\mathcal{D}(C_0)$

We introduce $C_0 < C$, the parameter C_0 will indicate the level of congestion of the link. It is assumed that the vector of integers $A = (A_j)$ is such that $A_1 = 1 < A_2 < \cdots < A_J$. The condition $A_1 = 1$ is used to simplify the presentation of the results and to avoid problems of irreducibility in particular but this is not essential.

If the network is in state $\ell = (\ell_j)$ and if the occupancy $\langle A, \ell \rangle$ is less than C_0, then any arriving request is accepted. If the occupancy is between C_0 and $C - 1$, it is accepted but with a minimal allocation, as a class 1 job. Finally it is rejected if the link is fully occupied, i.e. $\langle A, \ell \rangle = C$. It is assumed that $\mu_1 \leq \mu_j$, for $1 \leq j \leq J$, i.e. class 1 jobs are served with the smallest service rate.

Mathematically, the stochastic model is close to a loss network with the restriction that a job may change its requirements depending on the state of the network. This is a controlled loss network, see Zachary and Ziedins [ZZ11]. It does not seem that, like in uncontrolled loss networks, the associated Markov process giving the evolution of the vector ℓ has reversibility properties, or that its invariant distribution has a product form expression. Related schemes with product form are trunk reservation policies for which requests of a subset of classes are systematically rejected when the level of congestion of the link is above some threshold. See Bean *et al.* [BGZ95] and Zachary and Ziedins [ZZ02] for example. Concerning controlled loss networks,

mathematical results are more scarce. One can mention networks where jobs requiring congested links are redirected to less loaded links. Several mathematical approximations have been proposed to study these models. See the surveys Kelly [Kel91] and Zachary and Ziedins [ZZ11]. In our model, in the language of loss networks, the control is on the change of capacity requirements instead of a change of link.

2.2.2 Scaling Regime

The invariant distribution being, in general, not known, a scaling approach is used. The network is investigated under Kelly's regime, i.e. under heavy traffic regime with a scaling factor N. It has been introduced in Kelly [Kel86] to study the equilibrium of uncontrolled networks. The arrival rates are scaled by N: λ_j is replaced by $\lambda_j N$ as well as the capacity C by C^N and the threshold C_0 by C_0^N which are such that

$$C^N = cN + o(N) \text{ and } C_0^N = c_0 N + o(N),$$

for $0 < c_0 < c$.

Definition 2.1. *For $1 \leq j \leq J$ and $t \geq 0$, $L_j^N(t)$ denotes the number of class j jobs at time t in this system and $L^N(t) = (L_j^N(t), 1 \leq j \leq J)$.*

It will be assumed that the system is overloaded when the jobs have their initial bandwidth requirements

$$\langle A, \rho \rangle > c \text{ and } \frac{\Lambda}{\mu_1} < c, \tag{R}$$

with $\Lambda = \lambda_1 + \cdots + \lambda_J$ and $\rho_j = \lambda_j/\mu_j$, $1 \leq j \leq J$. The first condition gives that, without any change on the bandwidth requirement of jobs, the system will reject jobs. The second condition implies that the network could accommodate all jobs without losses (with high probability) if all of them would require the reduced bit rate $A_1 = 1$ and service rate μ_1.

It should be noted that, from the point of view of the design of algorithms, the constant c_0 has to be defined. If one takes $c_0 \in (\Lambda/\mu_1, c)$ then,

$$\langle A, \rho \rangle > c_0, \tag{R_1}$$

$$\frac{\Lambda}{\mu_1} < c_0. \tag{R_2}$$

hold.

If $\langle A, \rho \rangle < c$, it is not difficult to see that the system is equivalent to a classical underloaded loss network with one link and multiple classes of jobs. There is, of course, no need to use downgrading policies since the system can accommodate incoming requests without any loss when N is large. See Kelly [Kel91] or Section 6.7 of Robert [Rob03, p. 164] for example.

2.3 Scaling Results

In this section, we prove convergence results when the scaling parameter N goes to infinity. These results are obtained by studying the asymptotic behaviour of the occupation of the link around C_0^N,

$$m^N(t) \stackrel{\text{def.}}{=} \left\langle A, L^N(t) \right\rangle - C_0^N. \tag{2.2}$$

In the context of loss networks, the analogue of such quantity is the number of empty places. The following proposition shows that, for the downgrading policy, the boundary C^N does not play a role after some time if Condition (R_2) holds.

Proposition 2.1. *Under Condition (R_2) and if the initial state is such that*

$$\lim_{N\to+\infty} \left(\frac{L_j^N(0)}{N} \right) = \ell(0) = (\ell_{j,0}) \in \mathcal{S} \stackrel{\text{def.}}{=} \{x \in \mathbb{R}_+^J : \langle A, x \rangle < c\},$$

then, for $\varepsilon > 0$, there exists $t_\varepsilon \geq 0$ such that, for $T > t_\varepsilon$,

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{t_\varepsilon \leq t \leq T} \left\langle A, L^N(t) \right\rangle < (c_0 + \varepsilon)N \right) = 1.$$

Proof. Define

$$\left(\widetilde{L}_j^N(t) \right) \stackrel{\text{def.}}{=} \left(D_1^N(t) + X^N(t), D_2^N(t), \ldots, D_J^N(t) \right),$$

where $(X^N(t))$ is the process of the number of jobs of an independent $M/M/\infty$ queue with $X^N(0) = 0$, service rate μ_1 and arrival rate $\Lambda = \lambda_1 + \cdots + \lambda_J$ and, for $1 \leq j \leq J$,

$$D_j^N(t) \stackrel{\text{def.}}{=} \sum_{k=1}^{L_j^N(0)} \mathbb{1}_{\left\{E_{\mu_j,k} > t\right\}},$$

where $(E_{\mu_j,k})$ is a sequence of i.i.d. exponentially distributed random variables with rate μ_j.The quantity $D_j^N(t)$ is the number of initial class j jobs still present at time t. Using Theorem 6.13 of Robert [Rob03], one gets the convergence in distribution

$$\lim_{N\to+\infty} \left(\frac{X^N(t)}{N} \right) = \frac{\Lambda}{\mu_1} \left(1 - e^{-\mu_1 t} \right),$$

and, consequently,

$$\lim_{N\to+\infty} \left(\frac{1}{N} \left\langle A, \widetilde{L}^N(t) \right\rangle \right) = \left(\frac{\Lambda}{\mu_1} \left(1 - e^{-\mu_1 t} \right) + \sum_{j=1}^{J} A_j \ell_{j,0} e^{-\mu_j t} \right). \tag{2.3}$$

Since $\mu_1 \leq \mu_j$ for $1 \leq j \leq J$,

$$\frac{\Lambda}{\mu_1}\left(1-e^{-\mu_1 t}\right)+\sum_{j=1}^{J} A_j \ell_{j,0} e^{-\mu_j t} \leq \frac{\Lambda}{\mu_1}\left(1-e^{-\mu_1 t}\right)+e^{-\mu_1 t}\langle A, \ell(0)\rangle \leq \max(c_0, \langle A, \ell(0)\rangle),$$

by Condition (R_2). Note that the asymptotic occupancy, when N is large, remains below the initial occupancy.

If $0 < \varepsilon N < C^N - C_0^N$ and $L^N(0) \in \mathbb{N}^J$ such that $C_0^N + \varepsilon N < \left\langle A, L^N(0)\right\rangle < C^N$, let

$$\tau_N \stackrel{\text{def.}}{=} \inf\left\{t > 0 : \left\langle A, L^N(t)\right\rangle \leq C_0^N + \frac{\varepsilon}{2}N\right\}$$

then, on the event $\{\tau_N > T\}$, the downgrading policy gives that the identity in distribution

$$\left(\left(L_j^N(t)\right), 0 \leq t \leq T\right) \stackrel{\text{dist.}}{=} \left(\left(\widetilde{L}_j^N(t)\right), 0 \leq t \leq T\right) \tag{2.4}$$

holds. Condition (R_2) gives the existence of t_ε such that

$$\frac{\Lambda}{\mu_1}\left(1-e^{-\mu_1 t_\varepsilon}\right)+\sum_{j=1}^{J} A_j \ell_{j,0} e^{-\mu_j t_\varepsilon} = c_0 + \frac{\varepsilon}{2}.$$

Convergence (2.3) shows that the sequence (τ_N) converges in distribution to t_ε.

Note that, if $S \in (t_\varepsilon, T)$, as long as the process $(\left\langle A, L^N(t)\right\rangle)$ stays above C_0^N on $I = [t_\varepsilon, S)$, a relation similar to (2.4) holds. Using again Convergence (2.3), one gets that, as N goes to infinity, the process $(\left\langle A, L^N(t)\right\rangle / N)$ remains below $c_0 + \varepsilon$ with probability close to 1 on I. The proposition is proved. □

We are now investigating the asymptotic behaviour of the process $(m^N(t))$ defined by Relation (2.2). The variable indicates if the network is operating in saturation at time t, $m^N(t) \geq 0$, or not, $m^N(t) < 0$. In pure loss networks, when N is large, up to a change of time scale, the analogue of this process, the process of the number of empty places converges to a reflected random walk on $\mathbb{N}$. In our case, as it will be seen, the corresponding process is in fact a random walk on $\mathbb{Z}$.

Definition 2.2. *For $\ell = (\ell_j) \in \mathcal{S}$, let $(m_\ell(t))$ be the Markov process on $\mathbb{Z}$ whose $\mathcal{Q}$-matrix, Q_ℓ, is defined by, for $x \in \mathbb{Z}$ and $1 \leq j \leq J$,*

$$\begin{cases} Q_\ell(x, x - A_j) = \mu_j \ell_j, \\ Q_\ell(x, x + A_j) = \lambda_j, \ \textit{if } x < 0, \\ Q_\ell(x, x + 1) = \Lambda, \ \textit{if } x \geq 0, \end{cases} \tag{2.5}$$

with $\Lambda \stackrel{\text{def.}}{=} \lambda_1 + \lambda_2 + \cdots + \lambda_J$.

The following proposition summarises the stability properties of the Markov process $(m_\ell(t))$.

Proposition 2.2. *If $\ell = (\ell_j) \in \mathcal{S}$, then the Markov process $(m_\ell(t))$ is ergodic if $\ell \in \Delta_0$ with*

$$\Delta_0 \stackrel{\text{def.}}{=} \left\{ x \in \mathcal{S} : \langle A, x\rangle = c_0, \sum_{j=1}^{J} (\lambda_j - \mu_j x_j) A_j > 0 \text{ and } \Lambda < \sum_{j=1}^{J} \mu_j x_j A_j \right\} \tag{2.6}$$

π_ℓ denotes the corresponding invariant distribution.

Proof. The Markov process $(m_\ell(t))$ on $\mathbb{Z}$ behaves like a random walk on each of the two half-lines $\mathbb{N}$ and $\mathbb{Z}_-^*$. Definition (2.6) implies that if $\ell \in \Delta_0$, then the drift of the random walk is positive when in $\mathbb{Z}_-^*$ and negative when in $\mathbb{N}$. This property implies the ergodicity of the Markov process using the Lyapounov function $F(x) = |x|$, for example. See Corollary 8.7 of Robert [Rob03] for example. □

One now extends the expression π_ℓ for the values $\ell \in \mathcal{S} \setminus \Delta_0$. This will be helpful to describe the asymptotic dynamic of the system. See Theorem 2.1 further.

Definition 2.3. *One denotes $\pi_\ell = \delta_{-\infty}$, the Dirac measure at $-\infty$ when $\ell \in \Delta_-$, with*

$$\Delta_- \stackrel{\text{def.}}{=} \left\{ x \in \mathcal{S} : \langle A, x\rangle = c_0, \sum_{j=1}^{J} (\lambda_j - \mu_j x_j) A_j \le 0 \right\} \cup \left\{ x \in \mathcal{S} : \langle A, x\rangle < c_0 \right\}$$

and $\pi_\ell = \delta_{+\infty}$ if $\ell \in \Delta_+$, with

$$\Delta_+ \stackrel{\text{def.}}{=} \left\{ x \in \mathcal{S} : \langle A, x\rangle = c_0, \sum_{j=1}^{J} \mu_j x_j A_j \le \Lambda \right\} \cup \left\{ x \in \mathcal{S} : \langle A, x\rangle > c_0 \right\}.$$

Stochastic Evolution Equations

For $\xi > 0$, denote by $\mathcal{N}_\xi(\mathrm{d}t)$ a Poisson process on $\mathbb{R}_+$ with rate ξ and $(\mathcal{N}_\xi^i(\mathrm{d}t))$ an i.i.d. sequence of such processes. All Poisson processes are assumed to be independent. Classically, the process $(L^N(t))$ can be seen as the unique solution to the following stochastic differential equations (SDE),

$$\begin{cases} \mathrm{d}L_1^N(t) = -\displaystyle\sum_{k=1}^{+\infty} \mathcal{N}_{\mu_1}^k(\mathrm{d}t) \mathbb{1}_{\{k \le L_1^N(t-)\}} + \mathbb{1}_{\{m^N(t-) < C^N - C_0^N\}} \mathcal{N}_{\lambda_1 N}(\mathrm{d}t) \\ \qquad\qquad + \displaystyle\sum_{j=2}^{J} \mathbb{1}_{\{0 \le m^N(t-) < C^N - C_0^N\}} \mathcal{N}_{\lambda_j N}(\mathrm{d}t), \\ \mathrm{d}L_j^N(t) = -\displaystyle\sum_{k=1}^{+\infty} \mathcal{N}_{\mu_j}^k(\mathrm{d}t) \mathbb{1}_{\{k \le L_j^N(t-)\}} + \mathbb{1}_{\{m^N(t-) < 0\}} \mathcal{N}_{\lambda_j N}(\mathrm{d}t), \end{cases} \tag{2.7}$$

for $2 \leq j \leq J$, with initial condition $(L_j^N(0)) \in \mathbb{N}^J$ such that

$$\left\langle A, L^N(0) \right\rangle \leq C^N.$$

Theorem 2.1 (Limiting Dynamical System)**.** *Under Condition (R_2), if the initial conditions are such that $m^N(0) = m \in \mathbb{Z}$ and*

$$\lim_{N \to +\infty} \left(\frac{L_j^N(0)}{N} \right) = (\ell_j(0)) \in \mathcal{S},$$

then there exists continuous process $(\ell(t)) = (\ell_j(t))$ such that the convergence in distribution

$$\lim_{N \to +\infty} \left(\left(\frac{L_j^N(t)}{N} \right), \int_0^t f\left(m^N(u) \right) \mathrm{d}u \right) \overset{\text{dist.}}{=} \left((\ell_j(t)), \int_0^t \int_{\mathbb{Z}} f(x) \pi_{\ell(u)}(\mathrm{d}x)\, \mathrm{d}u \right) \tag{2.8}$$

holds for any function f with finite support on $\mathbb{Z}$. Furthermore, there exists $t_0 > 0$ such that $(\ell(t), t \geq t_0)$ satisfies the differential equations

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \ell_1(t) = -\mu_1 \ell_1(t) + \lambda_1 + \pi_{\ell(t)}(\mathbb{N}) \displaystyle\sum_{k=2}^{J} \lambda_k, \\ \dfrac{\mathrm{d}}{\mathrm{d}t} \ell_j(t) = -\mu_j \ell_j(t) + \lambda_j \pi_{\ell(t)}(\mathbb{Z}_-^*), \quad 2 \leq j \leq J, \end{cases} \tag{2.9}$$

where π_ℓ, for $\ell \in \mathcal{S}$, is the distribution of Proposition 2.2 and Definition 2.3.

It should be noted that, since the convergence holds for the convergence in distribution of processes, the limit $(\ell(t))$ is a priori a *random* process.

Proof. Using the same method as Hunt and Kurtz [HK94], one gets the analogue of Theorem 3 of this reference. Fix $\varepsilon > 0$ such that $c_0 + \varepsilon < c$, from Proposition 2.1, one gets that the existence of t_0 such that

$$\lim_{N \to +\infty} \mathbb{P}\left(\sup_{t_0 \leq t \leq T} \left\langle A, L^N(t) \right\rangle < (c_0 + \varepsilon) N \right) = 1,$$

which implies that the boundary condition $m^N(t) < C^N - C_0^N$ in the evolution equations (2.7) can be removed. Consequently, only the boundary condition of $(m^N(t))$ at 0 plays a role which gives Relation (2.9) as in Hunt and Kurtz [HK94]. Note that, contrary to the general situation described in this reference, we have indeed a convergence in distribution because, for any $\ell \in \mathcal{S}$, $(m_\ell(t))$ has exactly one invariant distribution (which may be a Dirac mass at infinity) by Proposition 2.2. See Conjecture 5 of Hunt and Kurtz [HK94]. □

The following proposition gives a characterisation of the equilibrium point of the dynamical system $(\ell(t))$.

Proposition 2.3 (Fixed Point)**.** *Under Conditions* (R_1) *and* (R_2)*, there exists a unique equilibrium point* $\ell^* \in \Delta_0$ *of the process* $(\ell_j(t))$ *defined by Equation* (2.8) *given by*

$$\begin{cases} \ell_1^* = c_0 - \pi^-(\rho_2 A_2 + \cdots + \rho_J A_J), \\ \ell_j^* = \rho_j \pi^-, \quad 2 \le j \le J, \end{cases} \tag{2.10}$$

where

$$\pi^- \stackrel{\text{def.}}{=} \frac{c_0 - \Lambda/\mu_1}{\langle A, \rho \rangle - \Lambda/\mu_1}, \tag{2.11}$$

with $\Lambda = \lambda_1 + \cdots + \lambda_J$*. The process* $(m_{\ell^*}(t))$ *is ergodic in this case.*

Proof. Assume that there exists an equilibrium point $\ell^* = (\ell_j^*)$ of $(\ell_j(t))$ defined by Equation (2.8), it is also an equilibrium point of the dynamical system defined by Equation (2.9), then

$$\begin{cases} \mu_1 \ell_1^* = \lambda_1 + (\lambda_2 + \cdots + \lambda_J)(1 - \pi^-), \\ \mu_j \ell_j^* = \lambda_j \pi^-, \quad 2 \le j \le J, \end{cases} \tag{2.12}$$

with $\pi^- = \pi_{\ell^*}(\mathbb{Z}_-^*)$. One gets

$$\sum_{j=1}^{J} \lambda_j = \sum_{j=1}^{J} \mu_j \ell_j^* < \sum_{j=1}^{J} \mu_j \ell_j^* A_j = \pi^- \sum_{j=1}^{J} \lambda_j A_j + (1 - \pi^-) \sum_{j=1}^{J} \lambda_j < \sum_{j=1}^{J} \lambda_j \tag{2.13}$$

We now show that the vector ℓ^* is on the boundary, i.e.

$$\sum_{j=1}^{J} A_j \ell_j^* = c_0. \tag{2.14}$$

If we assume that

$$\lim_{N \to +\infty} \left(\frac{L_j^N(0)}{N} \right) = (\ell_j^*),$$

from Theorem 2.1 and the definition of $(m^N(t))$, we know that, for the convergence of processes, the following relation holds

$$\lim_{N \to +\infty} \left(\frac{m^N(t)}{N} \right) = (\kappa_0), \text{ with } \kappa_0 \stackrel{\text{def.}}{=} \sum_{j=1}^{J} A_j \ell_j^* - c_0.$$

For $N_0 \in \mathbb{N}$, $\varepsilon > 0$ and $N \ge N_0$,

$$\int_0^1 \mathbb{1}_{\{|m^N(u)| \ge \varepsilon N\}} \, \mathrm{d}u \le \int_0^1 \mathbb{1}_{\{|m^N(u)| \ge \varepsilon N_0\}} \, \mathrm{d}u.$$

Using again Theorem 2.1 and the fact that ℓ^* is an equilibrium point of the dynamical system, we have, for the convergence in distribution

$$\lim_{N \to +\infty} \int_0^1 \mathbb{1}_{\{|m^N(u)| \le \varepsilon N_0\}} \, \mathrm{d}u = \pi_{\ell^*}([-\varepsilon N_0, \varepsilon N_0]).$$

The left-hand side of the above expression can be arbitrarily close to 1 when N_0 is large. By convergence of the sequence $(m^N(t)/N)$ to (κ_0), one gets that, for the convergence in distribution, the relation

$$\lim_{N\to+\infty} \int_0^1 \mathbb{1}_{\{|m^N(u)|/N\geq\varepsilon\}}\, \mathrm{d}u = 0$$

holds for $\varepsilon > 0$, which implies that $\kappa_0 = 0$. Thus Relation (2.14) holds. Finally, Relations (2.12) and (2.14) give Relation (2.10). One concludes therefore that $\ell^* \in \Delta_0$, the associated process $(m_{\ell^*}(t))$ is necessarily ergodic by Proposition 2.2 and Relations (2.13).

To prove that the ℓ^* defined by Relations (2.10) and (2.11) is indeed an equilibrium point of the dynamical system defined by Equation (2.9), one has to show that the right-hand side of Equation (2.11) is indeed equal to $\pi_{\ell^*}(\mathbb{Z}_-^*)$. This is proved in Proposition 2.5 of Section 2.4. □

Convergence of Invariant Distributions

In this section our main result establishes the convergence of the invariant distribution of the process $(m^N(t))$ as N gets large. This will give in particular the convergence with respect to N of the probability of not downgrading a request at equilibrium.

Lemma 2.1. *If the process $(\widetilde{L}_j^N(t))$ is the process $(L_j^N(t))$ at equilibrium then, for any $\varepsilon > 0$ and $T > 0$,*

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leq t\leq T}\sup_{2\leq j\leq J} \frac{\widetilde{L}_j^N(t)}{N} \leq \rho_j + \varepsilon\right) = 1.$$

Proof. Let $(L_j^N(t))$ be the process with initial state empty, then one can easily construct a coupling such that the relation

$$L_j^N(t) \leq \widetilde{Q}_j^N(t), \quad t \geq 0, \quad 2 \leq j \leq J,$$

holds almost surely, where $(Q_j^N(t))$ is the $M/M/\infty$ queue associated to class j requests. One deduces that,

$$\widetilde{L}_j^N(0) \leq_{\mathrm{st}} \widetilde{Q}_j^N(0)$$

where $\widetilde{Q}_j^N(0)$ is a Poisson random variable with parameter $\rho_j N$ and $\leq_{\mathrm{st}}$ is the stochastic ordering of random variables. One can therefore construct another coupling such that

$$\widetilde{L}_j^N(t) \leq \widetilde{Q}_j^N(t), \quad t \geq 0, \quad 2 \leq j \leq J,$$

where $(\widetilde{Q}_j^N(t))$ is a stationary version of the $M/M/\infty$ queue associated to class j requests. The lemma is then a consequence of the following convergence in distribution of processes,

$$\lim_{N\to+\infty} \left(\frac{\widetilde{Q}_j^N(t)}{N}\right) = (\rho_j)$$

for $2 \leq j \leq J$, see Theorem 6.13 of Robert [Rob03, pp. 159] for example. □

Definition 2.4. *Let $(y(t))$ be the dynamical system on $\mathcal{S}$ satisfying*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} y_1(t) = -\mu_1 y_1(t) + \lambda_1 + \left(\displaystyle\sum_{k=2}^{J} \lambda_k\right) \frac{1}{\Lambda_A} \sum_{k=1}^{J} A_k \left(\lambda_k - \mu_k y_k(t)\right), \\ \dfrac{\mathrm{d}}{\mathrm{d}t} y_j(t) = -\mu_j y_j(t) + \lambda_j \dfrac{1}{\Lambda_A} \displaystyle\sum_{k=1}^{J} \left(A_k \mu_k y_k(t) - \lambda_k\right), \quad 2 \le j \le J, \end{cases} \tag{2.15}$$

with

$$\Lambda_A \stackrel{\text{def.}}{=} \sum_{k=1}^{J} \lambda_k (A_k - 1).$$

Lemma 2.2. *If $y(0) \in \Delta_0$ and if there exists an instant $T > 0$ such that $y(t) \in \Delta_0$ for $t \in [0, T]$ then $(y(t))$ and $(\ell(t)$ coincide on the time interval $[0, T]$, where $(\ell(t))$ is the solution of Equations (2.9) with $\ell(0) = y(0)$.*

Proof. The proposition is a simple consequence of the representation (2.9) of the differential equations defining the dynamical system $(\ell(t))$ and of the explicit expression of the quantity $\pi_\ell(\mathbb{Z}_-^*)$ given by Relation (2.22) when $\ell \in \Delta_0$, see Relation (2.6). □

The next proposition investigates the stability Properties of $(y(t))$.

Proposition 2.4. *Let H_0 be the hyperplane*

$$H_0 \stackrel{\text{def.}}{=} \{z \in \mathcal{S} \colon \langle A, z \rangle = c_0\}\}$$

if $y(0) \in H_0$ then $y(t) \in H_0$ for all $t \ge 0$ and $(y(t))$ is converging exponentially fast to ℓ^* defined in Proposition 2.3.

Proof. It is easily checked that

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle A, y(t) \rangle = 0,$$

so that if $y(0) \in H_0$, then the function $t \mapsto \langle A, y(t) \rangle$ is constant and equal to c_0, hence $y(t) \in H_0$ for all $t \ge 0$.

For $2 \le j \le J$,

$$\frac{\mathrm{d}}{\mathrm{d}t} y_j(t) = \lambda_j b_0 - \mu_j y_j(t) + \lambda_j \sum_{k=2}^{J} b_k y_k(t),$$

with

$$b_0 \stackrel{\text{def.}}{=} \frac{\mu_1 c_0 - \Lambda}{\Lambda_A} \text{ and } b_j = \frac{A_j(\mu_j - \mu_1)}{\Lambda_A}.$$

In matrix form, if $z(t) = (y_2(t), \ldots, y_J(t))$, it can be expressed as

$$\frac{\mathrm{d}}{\mathrm{d}t} z(t) = e_b + B z(t), \tag{2.16}$$

with $e_b = b_0(\lambda_2, \dots, \lambda_J) \in \mathbb{R}^{J-1}$ and $B = (B_{jk}, 2 \le j, k \le J)$ with

$$B_{jk} \overset{\text{def.}}{=} \lambda_j b_k - \mu_j \mathbb{1}_{\{k=j\}}.$$

If $v = (v_2, \dots, v_J)$ is an eigenvector for the eigenvalue x of B, then

$$(x + \mu_j)v_j = \lambda_j \sum_{k=2}^{J} b_k v_k, \quad 2 \le j \le J,$$

hence, x is an eigenvalue if and only if it is a solution of the equation

$$F(x) \overset{\text{def.}}{=} \sum_{j=2}^{J} \frac{b_j \lambda_j}{x + \mu_j} = 1.$$

If L is the number of distinct values of μ_j, $2 \le j \le J$, such that $\mu_j \neq \mu_1$, then the above equation shows that an eigenvalue is a zero of a polynomial of degree at most L. Using Conditions (R), it is easy to check that the relation $F(0) < 1$ holds. In particular 0 is not an eigenvalue and, consequently B is invertible. Due to the poles of F at the $-\mu_j$, $2 \le j \le J$ and the relations $F(0) < 1$ and $\mu_j \ge \mu_1$ for $2 \le j \le J$, one has already L negative solutions of the equation $F(x) = 1$. All eigenvalues of B are thus negative, consequently, $\exp(tB)$ converges to 0. (See Corollary 2 of Chapter 25 of Arnol'd [Arn92, p. 223] for example.)

Equation (2.16) can be solved as

$$z(t) = e^{tB}\left(z(0) + B^{-1}e_b\right) - B^{-1}e_b.$$

Therefore the function $(z(t))$ has a limit at infinity given by $-B^{-1}e_b$ which is clearly $(\ell_j^*, 2 \le j \le J)$. The proposition is proved. □

One can now prove the main result of this section.

Theorem 2.2. *If ℓ^* is the quantity defined in Proposition 2.3, then the equilibrium distribution of $(m^N(t))$ converges to π_{ℓ^*} when N goes to infinity.*

Proof. Recall that $m^N(t) = \left\langle A, L^N(t) \right\rangle - C_0^N$ and let Π^N be the invariant distribution of $(L^N(t))$. It is assumed that the distribution of $L^N(0)$ is Π^N for the rest of the proof. In particular $(m^N(t))$ is a stationary process.

One first proves that $(L^N(0)/N)$ converges in distribution to ℓ^*. The boundary condition $\left\langle A, L^N(0) \right\rangle \le C^N$ gives that the sequence of random variables $(L^N(0)/N)$ is tight. If $(L^{N_k}(0)/N_k)$ is a convergent subsequence to some random variable ℓ^∞, by Theorem 2.1, one gets that, for the convergence in distribution, the relation

$$\lim_{k \to +\infty} \left(\left(\frac{L^{N_k}(t)}{N_k} \right) \right) = (\ell(t))$$

holds, where $(\ell(t))$ is a solution of Equation (2.9) with initial point at $\ell(0) = \ell^\infty$. Note that $(\ell(t))$ is a stationary process, its distribution is invariant under any time shift.

By Lemma 2.1 one has that the relation $\ell_j(t) \leq \rho_j$, for $2 \leq j \leq J$, holds almost surely on any finite time interval and, by Proposition 2.1, $\langle A, \ell(t)\rangle \leq c_0$ also holds almost surely on finite time intervals.

Assume that $\langle A, \ell(0)\rangle < c_0$ holds. The ODEs defining the limiting dynamical system are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_j(t) = -\mu_j \ell_j(t) + \lambda_j, \quad 1 \leq j \leq J,$$

as long as the condition $\langle A, \ell(t)\rangle < c_0$ holds, hence on the corresponding time interval, one has

$$\ell_j(t) = \rho_j + (\ell_j(0) - \rho_j)e^{-\mu_j t}, \quad 1 \leq j \leq J,$$

so that

$$\langle A, \ell(t)\rangle = \langle A, \rho\rangle + \sum_{j=1}^{J} A_j \left(\ell_j(0) - \rho_j\right) e^{-\mu_j t}.$$

Since $\langle A, \rho\rangle > c_0$, there exists some $t_1 > 0$ such that $\langle A, \ell(t_1)\rangle = c_0$.

Hence, by stationarity in distribution of $(\ell(t))$, one can shift time at t_0 and assume that $\langle A, \ell(0)\rangle = c_0$. On this event

$$\sum_{j=1}^{J} \mu_j \ell_j(0) A_j \geq \mu_1 \sum_{j=1}^{J} \ell_j(0) A_j = \mu_1 c_0 > \Lambda = \sum_{j=1}^{J} \lambda_j. \tag{2.17}$$

Similarly, since $\ell_j(0) \leq \rho_j$ for all $2 \leq j \leq J$,

$$\begin{aligned}
&\sum_{j=1}^{J} A_j(\lambda_j - \mu_j \ell_j(0)) = \lambda_1 - \mu_1 c_0 + \mu_1 \sum_{j=2}^{J} A_j \ell_j(0) + \sum_{j=2}^{J} A_j(\lambda_j - \mu_j \ell_j(0)) \\
&= -\mu_1 c_0 + \sum_{j=1}^{J} A_j(\lambda_j + (\mu_1 - \mu_j)\ell_j(0)) \geq -\mu_1 c_0 + \sum_{j=1}^{J} A_j(\lambda_j + (\mu_1 - \mu_j)\rho_j) \\
&= -\mu_1 c_0 + \sum_{j=1}^{J} A_j \lambda_j \frac{\mu_1}{\mu_j} = \mu_1 \left(\langle A, \rho\rangle - c_0\right) > 0,
\end{aligned} \tag{2.18}$$

and the last quantity is independent of $\ell(0)$. Relations (2.17) and (2.18) show that $\ell(0) \in \Delta_0$ and, by Equations (2.9) and (2.15), they also hold for t in a small neighbourhood I of 0 independent of $\ell(0)$ so that $\ell(t) \in \Delta_0$ for $t \in I$. Consequently, the dynamical system $(\ell(t))$ never leaves Δ_0. Lemma 2.2 shows that the two dynamical systems $(\ell(t))$ and $(y(t))$ (with $y(0) = \ell(0)$) coincide. Hence, on the one hand $(\ell(t))$ is a stationary process and, on the other hand, it is a dynamical system converging to ℓ^*, one deduces that it is constant and equal to ℓ^*. We have thus proved that the sequence $(L^N(0)/N)$ converges in distribution to ℓ^*.

Using again Theorem 2.1, one gets that, for the convergence in distribution,

$$\lim_{N \to +\infty} \int_0^1 f(m^N(u))\, \mathrm{d}u \overset{\text{dist.}}{=} \int_{\mathbb{Z}} f(x) \pi_{\ell^*}(\mathrm{d}x)$$

holds for any function f with finite support on $\mathbb{Z}$. Using the stationarity of $(m^N(t))$ and Lebesgue's Theorem, one obtains

$$\lim_{N\to+\infty} \mathbb{E}\left(f(m^N(0))\right) = \int_{\mathbb{Z}} f(x)\pi_{\ell^*}(\mathrm{d}x).$$

The theorem is proved. □

Since a job arriving at time t is not downgraded if $m^N(t) < 0$, one obtains the following corollary.

Corollary 2.1. *As N goes to infinity, the probability that, at equilibrium, a job is not downgraded in this allocation scheme is converging to π^- defined in Proposition 2.10,*

$$\pi^- = \frac{c_0 - \Lambda/\mu_1}{\langle A, \rho\rangle - \Lambda/\mu_1}.$$

2.4 Invariant Distribution

We assume in this section that $\ell \in \Delta_0$, as defined in Proposition 2.2, so that $(m_\ell(t))$ is an ergodic Markov process. The goal of this section is to derive an explicit expression of the invariant distribution π_ℓ on $\mathbb{Z}$ of $(m_\ell(t))$. At the same time, Proposition 2.5 below gives the required argument to complete the proof of Proposition 2.3 on the characterisation of the fixed point of the dynamical system.

2.4.1 Functional Equation

In the following we denote by Y_ℓ a random variable with distribution $\pi_\ell = (\pi_\ell(n), n \in \mathbb{Z})$.

For $r > 0$, we will use the notation

$$D(r) \overset{\text{def.}}{=} \{z \in \mathbb{C}, |z| < r\}, \; D^c(r) \overset{\text{def.}}{=} \{z \in \mathbb{C}, |z| > r\} \text{ and}$$
$$\gamma(r) \overset{\text{def.}}{=} \{z \in \mathbb{C}, |z| = r\}.$$

For sake of simplicity, we will use $D \overset{\text{def.}}{=} D(1)$ and $D^c \overset{\text{def.}}{=} D^c(1)$.

Lemma 2.3. *With the notation*

$$\varphi_+(z) \overset{\text{def.}}{=} \mathbb{E}\left(z^{Y_\ell}\mathbb{1}_{\{Y_\ell\geq 0\}}\right), \quad \varphi_-(z) \overset{\text{def.}}{=} \mathbb{E}\left(z^{Y_\ell}\mathbb{1}_{\{Y_\ell< 0\}}\right),$$

the random variable Y_ℓ is such that

$$P_1(z)\varphi_+(z) = P_2(z)\varphi_-(z) \tag{2.19}$$

where P_1 and P_2 are polynomials defined by

$$\begin{cases} P_1(z) \overset{\text{def.}}{=} \displaystyle\sum_{j=1}^{J}\left[(\lambda_j + \mu_j\ell_j)z^{A_J} - \lambda_j z^{A_J+1} - \mu_j\ell_j z^{A_J-A_j}\right], \\ P_2(z) \overset{\text{def.}}{=} \displaystyle\sum_{j=1}^{J}\left[\lambda_j z^{A_J+A_j} + \mu_j\ell_j z^{A_J-A_j} - (\lambda_j + \mu_j\ell_j)z^{A_J}\right]. \end{cases} \tag{2.20}$$

Proof. For $z \in \gamma(1)$ define $f_z : \mathbb{Z} \mapsto \mathbb{C}$ such that $f_z(x) = z^x$, for $x \in \mathbb{Z}$. Equilibrium equations for $(m_\ell(t))$ give the identity

$$\sum_{\substack{x,y\in\mathbb{Z}\\ x\neq y}} \pi_\ell(x) Q_\ell(x,y)(f_z(y) - f_z(x)) = 0,$$

where Q_ℓ is the $\mathcal{Q}$-matrix of $(m_\ell(t))$ given by Equation (2.5). After some simple reordering, one gets the relation

$$\mathbb{E}\left(z^{Y_\ell}\mathbb{1}_{\{Y_l\geq 0\}}\right)\sum_{j=1}^{J}\left(\lambda_j(1-z)+\mu_j\ell_j\left(1-z^{-A_j}\right)\right) = \\ -\mathbb{E}\left(z^{Y_l}\mathbb{1}_{\{Y_l<0\}}\right)\sum_{j=1}^{J}\left(\lambda_j\left(1-z^{A_j}\right)+\mu_j\ell_j\left(1-z^{-A_j}\right)\right). \quad (2.21)$$

Using the definition of $\varphi_+(z)$ and $\varphi_-(z)$, Equation (2.21) can be rewritten as Equation (2.19). □

Proposition 2.5. *If $\ell \in \Delta_0$ then*

$$\pi_\ell(\mathbb{Z}_-^*) = \frac{\sum_{j=1}^J (A_j\mu_j\ell_j - \lambda_j)}{\sum_{j=1}^J \lambda_j(A_j-1)}. \quad (2.22)$$

In particular if $\ell^* \in \mathcal{S}$ is given by Relation (2.10) then

$$\pi_{\ell^*}(\mathbb{Z}_-^*) = \frac{c_0 - \Lambda/\mu_1}{\langle A,\rho\rangle - \Lambda/\mu_1}.$$

Note that the right-hand side of the last relation is precisely π^- of Relation (2.11) which is the result necessary to complete the proof of Proposition 2.3.

Proof. With the same notations as before, from Relation (2.19),

$$\frac{\varphi_-(z)}{\varphi_+(z)} = \frac{P_1(z)}{P_2(z)}$$

holds for $z \in \mathbb{C}$, with $z \in \gamma(1)$. By definition of $\varphi_-(z)$ and $\varphi_+(z)$,

$$\lim_{z\to 1}\varphi_-(z) = \pi_\ell(\mathbb{Z}_-^*) \text{ and } \lim_{z\to 1}\varphi_+(z) = 1 - \pi_\ell(\mathbb{Z}_-^*).$$

Since 1 is a zero of P_1 and P_2, this gives the relation

$$\frac{\pi_\ell(\mathbb{Z}_-^*)}{1-\pi_\ell(\mathbb{Z}_-^*)} = \frac{P_1'(1)}{P_2'(1)} = \frac{\sum_{j=1}^J (A_j\mu_j\ell_j - \lambda_j)}{\sum_{j=1}^J A_j(\lambda_j - \mu_j\ell_j)}.$$

Using the expression of (ℓ_j^*), with some algebra, one gets

$$\pi_{\ell^*}(\mathbb{Z}_-^*) = \left(c_0 - \sum_{j=1}^J \frac{\lambda_j}{\mu_1}\right) \Big/ \left(\sum_{j=1}^J \rho_j A_j - \sum_{j=1}^J \frac{\lambda_j}{\mu_1}\right) = \pi^-.$$

The proposition is proved. □

Relation (2.19) is valid on the unit circle, however the function φ_+ (resp. φ_-) is defined on D (resp. D^c). This can then be expressed as a Wiener-Hopf factorisation problem analogous to the one used in the analysis of reflected random walks on $\mathbb{N}$. This is used in the analysis of the $GI/GI/1$ queue (see Chapter VIII of Asmussen [Asm03] or Chapter 3 of Robert [Rob03] for example). In a functional context, this is a special case of a Riemann's problem (see Gakhov [Gak90]). In our case, this is a random walk on $\mathbb{Z}$, with a drift depending on the half-space where it is located. The first (resp. second) condition in the definition of the set Δ_0 in Definition (2.6) implies that the drift of the random walk on $\mathbb{Z}_-^*$ (resp. in $\mathbb{N}$) is positive (resp. negative).

The first step in the analysis of Equation (2.19) is to determine the locations of the zeros of P_1 and P_2. This is the purpose of the following lemma.

Lemma 2.4. *(Location of the Zeros of P_1 and P_2) Let ℓ be in Δ_0.*

(i) *Polynomial P_2 has exactly two positive real roots 1 and $z_2 \in]0,1[$. There are $A_J - 1$ roots in $D(z_2)$ and $A_J - 1$ roots whose modulus are strictly greater than 1.*

(ii) *Polynomial P_1 has exactly two positive real roots 1 and $z_1 > 1$. The $A_J - 1$ remaining roots have a modulus strictly smaller than 1.*

Proof. One first notes that P_2 is a polynomial with the same form as the f defined by Equation (13) in Bean *et al.* [BGZ95] (with $e_j = A_j$, $\kappa_j = \lambda_j$ and $\hat{e} = A_J$). The roots of Q are exactly the roots of f. Lemma 2.2 of Bean *et al.* [BGZ95] gives assertion *(i)* of our lemma.

The proof of assertion *(ii)* uses an adaptation of the argument for the proof of Lemma 2.2 of Bean *et al.* [BGZ95]. Define the function $f(z) \overset{\text{def.}}{=} z^{-A_J} P_1(z)$. Recall that P_1 is a polynomial with degree $A_J + 1$. There are exactly two real positive roots for P_1. Indeed, $f(1) = 0$ and it is easily checked that f is strictly concave with

$$f'(1) = \sum_{j=1}^{J} (-\lambda_j + A_j \mu_j \ell_j) > 0,$$

since $\ell \in \Delta_0$, by the second condition in Definition (2.6). Hence P_1 has a real zero z_1 greater than 1.

Let $r \in (1, z_1)$ be fixed, note that $P_1(r) > 0$. Define

$$f_1(z) \overset{\text{def.}}{=} K z^{A_J}, \text{ with } K \overset{\text{def.}}{=} \sum_{j=1}^{J} (\lambda_j + A_j \mu_j \ell_j),$$

$$f_2(z) \overset{\text{def.}}{=} \sum_{j=1}^{J} \left(\lambda_j z^{A_J+1} + \mu_j \ell_j z^{A_J - A_j} \right),$$

so that $P_1 = f_1 - f_2$.

Fix some $z \in \gamma(r)$. Expressing these functions in terms of real and imaginary parts,

$$z^{A_J} = \alpha_1 + i\beta_1 \text{ and } f_2(z) = \alpha_2 + i\beta_2,$$

one gets

$$\left|f_1(z) - f_2(z) - bz^{A_J}\right|^2 = |K(\alpha_1 + i\beta_1) - b(\alpha_1 + i\beta_1) - (\alpha_2 + i\beta_2)|^2$$
$$= (K\alpha_1 - \alpha_2)^2 + (K\beta_1 - \beta_2)^2 + H = |f_1(z) - f_2(z)|^2 + H, \quad (2.23)$$

with

$$H \stackrel{\text{def.}}{=} (b\alpha_1)^2 - 2b\alpha_1(K\alpha_1 - \alpha_2) + (b\beta_1)^2 - 2b\beta_1(K\beta_1 - \beta_2)$$
$$= b(b - 2K)(\alpha_1^2 + \beta_1^2) + 2b(\alpha_1\alpha_2 + \beta_1\beta_2).$$

Cauchy-Schwarz's Inequality gives the relation

$$\alpha_1\alpha_2 + \beta_1\beta_2 \leq \frac{1}{K}\,|f_2(z)|\,|f_1(z)| \leq \frac{1}{K} f_2(r) f_1(r),$$

since $|f_i(z)| \leq f_i(|z|)$ for $i = 1, 2$. Thus,

$$\frac{H}{b} = (b - 2K)(\alpha_1^2 + \beta_1^2) + 2(\alpha_1\alpha_2 + \beta_1\beta_2) \leq (b - 2K)\frac{f_1(r)^2}{K^2} + 2f_2(r)\frac{f_1(r)}{K}$$
$$= \frac{f_1(r)}{K^2}\left((b - 2K)f_1(r) + 2Kf_2(r)\right) = \frac{f_1(r)}{K^2}(bf_1(r) - 2KP_1(r)).$$

Since $P_1(r) > 0$, b can be chosen so that $bf_1(r) < 2KP_1(r)$. From the above relation and Equation (2.23), one gets that for $z \in \gamma(r)$, the relation

$$\left|f_1(z) - f_2(z) - bz^{A_J}\right| < |f_1(z) - f_2(z)|$$

holds. By Rouché's theorem, one obtains that, for any $r \in (1, z_1)$, P_1 has exactly A_J roots in $D(r)$. One concludes that P_1 has exactly A_J roots in $\overline{D}$. It is easily checked that if $z \in \gamma(1)$ and $z \notin \mathbb{R}$ then the real part of $P_1(z)$ is positive, hence z cannot be a root of the polynomial P_1. Consequently, P_1 has exactly $A_J - 1$ roots in D. The lemma is proved. □

Definition 2.5. *For $U \in \{P_1, P_2\}$, denote by $\mathcal{Z}_U$ the set of the zeros of U different from* 1.

Define

$$\Phi(z) \stackrel{\text{def.}}{=} \begin{cases} -\varphi_+(z){\lambda_J}^{-1}(z - z_1) \displaystyle\prod_{q \in \mathcal{Z}_{P_2} \cap D^c} (z - q)^{-1}, & z \in D \\ \\ \varphi_-(z)\Lambda^{-1} \displaystyle\prod_{q \in \mathcal{Z}_{P_2} \cap D} (z - q) \prod_{p \in \mathcal{Z}_{P_1} \cap D} (z - p)^{-1}, & z \in D^c \end{cases}$$

with $\Lambda = \lambda_1 + \cdots + \lambda_J$ and the same notations as before. By definition, function Φ is holomorphic in D and D^c and, from Relation (2.19), is continuous on $\gamma(1)$. The analytic continuation theorem, Theorem 16.8 of Rudin [Rud87] for example, gives that Φ is holomorphic on $\mathbb{C}$. For $z \in D^c$,

$$|\varphi_-(z)| \leq \mathbb{E}\left(\mathbb{1}_{\{Y_\ell < 0\}}\,|z|^{Y_\ell}\right) \leq \frac{1}{|z|},$$

since the cardinality of $\mathcal{Z}_{P_1} \cap D$ (resp. $\mathcal{Z}_{P_2} \cap D$) is $A_J - 1$ (resp. A_J), the holomorphic function Φ is therefore bounded on $\mathbb{C}$. By Liouville's theorem, Φ is constant, equal to $\kappa \in \mathbb{C}$. Therefore

$$\begin{cases} \varphi_+(z) = -\kappa\lambda_J(z - z_1)^{-1} \prod_{q \in \mathcal{Z}_{P_2}\ capD^c} (z - q), & z \in D, \\ \varphi_-(z) = \kappa\Lambda \prod_{q \in \mathcal{Z}_{P_2} \cap D} (z - q)^{-1} \prod_{p \in \mathcal{Z}_{P_1} \cap D} (z - p), & z \in D^c. \end{cases} \tag{2.24}$$

Recall that $\varphi(z) = \varphi_+(z) + \varphi_-(z) = \mathbb{E}\left(z^{Y_\ell}\right)$ is a generating function, in particular $\varphi(1) = 1$. Plugging the previous expressions for φ_+ and φ_- in $\varphi_+(1) + \varphi_-(1) = 1$, one gets the relation

$$1 = -\kappa \prod_{q \in \mathcal{Z}_{P_2} \cap D} (1 - q)^{-1} \frac{1}{1 - z_1} \left(P_1'(1) + P_2'(1)\right),$$

hence, using equation (2.20),

$$\kappa = \frac{z_1 - 1}{\Lambda_A} \prod_{q \in \mathcal{Z}_{P_2} \cap D} (1 - q),$$

where Λ_A is introduced in Definition 2.4. Note that κ is positive. We can now state the main result of this section.

Proposition 2.6 (Invariant Measure)**.** *If $\ell \in \Delta_0$ defined by Relation (2.6), then the invariant measure π_ℓ can be expressed, for $n \in \mathbb{Z}$, as*

$$\pi_\ell(n) = \begin{cases} -\kappa \sum_{q \in \mathcal{Z}_{P_2} \cap D} \dfrac{P_1(q) q^{-n-1}}{(q - z_1)(q - 1) R_D'(q)}, & n < 0, \\ \kappa\left(\alpha_n + \dfrac{P_2(z_1) z_1^{-n-1}}{(z_1 - 1) R_D(z_1)}\right), & 0 \le n < A_J - 1, \\ \kappa \dfrac{P_2(z_1) z_1^{-n-1}}{(z_1 - 1) R_D(z_1)}, & n \ge A_J - 1, \end{cases}$$

where z_1 is defined in Lemma 2.4, and P_1 and P_2 by Relation (2.20),

$$R_D(z) = \prod_{q \in \mathcal{Z}_{P_2} \cap D} (z - q), \quad \kappa = \frac{(z_1 - 1) R_D(1)}{\Lambda_A},$$

for $0 \le n < A_J - 1$, α_n is the coefficient of degree n of the polynomial

$$-\frac{1}{z - z_1}\left(\frac{P_2(z)}{(z - 1) R_D(z)} - \frac{P_2(z_1)}{(z_1 - 1) R_D(z_1)}\right).$$

Proof. Note that, for $z \in \mathbb{C}$,

$$\prod_{p \in \mathcal{Z}_{P_1} \cap D} (z - p) = -\frac{1}{\Lambda} \frac{P_1(z)}{(z - z_1)(z - 1)}.$$

For $z \in D^c$,

$$\varphi_-(z) = \kappa\Lambda \prod_{q\in\mathcal{Z}_{P_2}\cap D} (z-q)^{-1} \prod_{p\in\mathcal{Z}_{P_1}\cap D} (z-p).$$

Since $|\mathcal{Z}_{P_1} \cap D| = A_J - 1 < A_J = |\mathcal{Z}_{P_2} \cap D|$ by Lemma 2.4, φ_- has the following partial fraction decomposition

$$\begin{aligned}\varphi_-(z) &= -\kappa \sum_{q\in\mathcal{Z}_{P_2}\cap D} \frac{P_1(q)}{(q-z_1)(q-1)R'_D(q)} \frac{1}{z-q} \\ &= \sum_{i=0}^{\infty} -\kappa \sum_{q\in\mathcal{Z}_{P_2}\cap D} \frac{P_1(q)q^i}{(q-z_1)(q-1)R'_D(q)} \frac{1}{z^{i+1}}.\end{aligned}$$

Denote

$$R_{D^c}(z) = \prod_{q\in\mathcal{Z}_{P_2}\cap D^c} (z-q) = \frac{P_2(z)}{\lambda_J(z-1)R_D(z)},$$

then

$$\varphi_+(z) = -\kappa\lambda_J \frac{R_{D^c}(z)}{z-z_1} = \kappa\Big(-\lambda_J \frac{R_{D^c}(z) - R_{D^c}(z_1)}{z-z_1} + \frac{P_2(z_1)}{(1-z_1)R_D(z_1)} \frac{1}{z-z_1}\Big).$$

One concludes using the expression of κ obtained before. □

2.4.2 Some Moments of (π_{ℓ^*})

Using the probability generating function $\varphi(z)$ of π_{ℓ^*} from Equation (2.24), one can derive an explicit expression of the mean, the variance and the skewness of such distribution. The skewness of a random variable X is a measure of the asymmetry of the distribution of X,

$$\mathbb{S}(X) \overset{\text{def.}}{=} \mathbb{E}([X - \mathbb{E}(X)]^3).$$

See Doane and Seward [DS11] for example.

Proposition 2.7. *If Y_{ℓ^*} is a random variable with distribution π_{ℓ^*} then*

$$\begin{aligned}\mathbb{E}(Y_{\ell^*}) &= A_J + \frac{\theta_2}{2\theta_1} - S(1), \\ \mathbb{V}(Y_{\ell^*}) &= \frac{\theta_2 + 2\theta_3}{6\theta_1} - \left(\frac{\theta_2}{2\theta_1}\right)^2 - \left(S(1) + S'(1)\right), \\ \mathbb{S}(Y_{\ell^*}) &= \frac{\theta_2^3}{4\theta_1^3} + \theta_2\frac{\theta_2 - 2\theta_3}{4\theta_1^2} + \frac{\theta_4 - \theta_3}{4\theta_1} - \left(S(1) + 3S'(1) + S''(1)\right),\end{aligned}$$

where, for $i \geq 1$,

$$\theta_i \overset{\text{def.}}{=} \sum_{j=2}^{J} \lambda_j A_j^{i-1}(A_j - 1),$$

and

$$S(z) \stackrel{\text{def.}}{=} \frac{1}{z - z_1} + \sum_{q \in \mathcal{Z}_{P_2} \cap D} \frac{1}{z - q},$$

with $R_D(z)$ defined in Proposition 2.6.

The proof is straightforward, modulo some tedious calculations of the successive derivatives of $\varphi(z)$ evaluated at 1. Figure 2.1 shows that the distribution of Y_{ℓ^*} is significantly asymmetrical. For this example $\mathbb{E}(Y_{\ell^*}) = 8.04819$, $\mathbb{V}(Y_{\ell^*}) = 77.2284$ and $\mathbb{S}(Y_{\ell^*}) = 0.967069$.

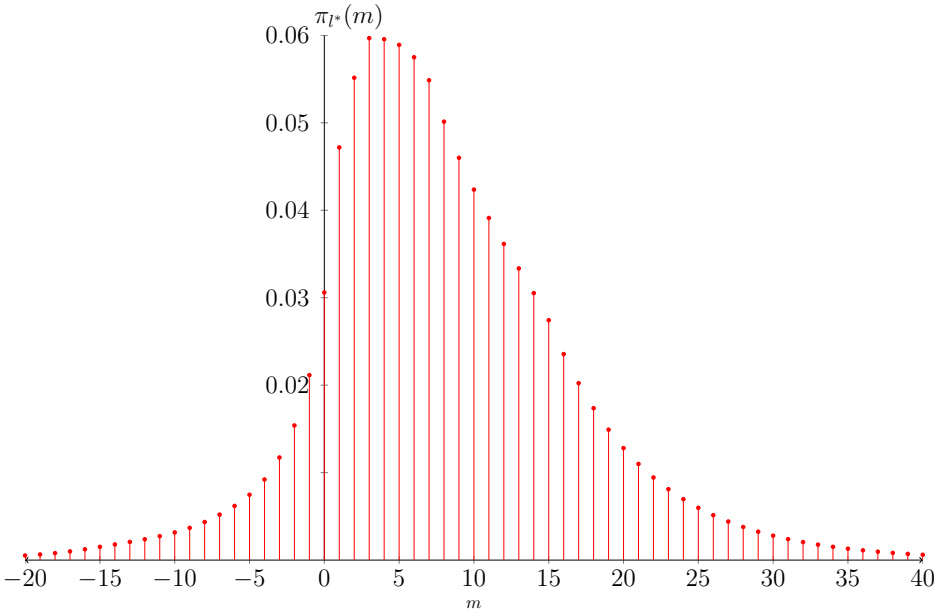


Figure 2.1 – The histogram of Y_{ℓ^*} with the parameters $J = 5$, $A = (1, 2, 4, 8, 16)$, $\lambda = (0.25, 0.2, 0.15, 0.1, 0.05)$, $\mu = (1, 1, 1, 1, 1)$ and $c_0 = 0.97$.

2.5 Applications

Comparison with a Pure Loss System

In this case, a request which cannot be accommodated is rejected right away. Recall that, with probability 1, our algorithm does not reject any request. The purpose of this section is to discuss the price of such a policy. Intuitively, at equilibrium the probability W_L of accepting a job at requested capacity in a pure loss system is greater that the corresponding quantity W_D for the downgrading algorithm. See Proposition 2.8 below. A further question is to assess the impact of such policy, i.e. the order of magnitude of the difference $W_L - W_D$.

Under the same assumptions about the arrivals and under the condition

$$\langle A, \rho \rangle > c \quad \text{and} \quad \frac{\Lambda}{\mu_1} < c,$$

with $\Lambda = \lambda_1 + \cdots + \lambda_J$, then, as N gets large, the equilibrium probability that a request of class $1 \leq j \leq J$ is accepted in the pure loss system is converging to β^{A_j}, where $\beta \in (0,1)$ is the unique solution of the equation

$$\sum_{j=1}^{J} A_j \rho_j \beta^{A_j} = c. \tag{2.25}$$

see Kelly [Kel86, Kel91]. Consequently, the asymptotic load of accepted requests is given by

$$W_L \overset{\text{def.}}{=} \frac{1}{\Lambda} \sum_{j=1}^{J} \rho_j \beta^{A_j}.$$

Under the downgrading policy, the equilibrium probability that a job is accepted without degradation is given by π^-, the asymptotic load of requests accepted without degradation is

$$W_D \overset{\text{def.}}{=} \frac{1}{\Lambda} \sum_{j=1}^{J} \rho_j \frac{c_0 - \Lambda/\mu_1}{\langle A, \rho \rangle - \Lambda/\mu_1},$$

for $c_0 \in (\Lambda/\mu_1, c)$. Note that, when the service rates are constant equal to 1, then W_L (resp. W_D) is the asymptotic throughput of accepted requests (resp. of non-degraded requests).

The following proposition establishes the intuitive property that a pure loss system has better performances in terms of acceptance.

Proposition 2.8. *For $c_0 \in (\Lambda/\mu_1, c)$, the relation $W_D \leq W_L$ holds.*

Proof. The representation of these quantities gives that the relation to prove is equivalent to the inequality

$$\sum_{j=1}^{J} \rho_j \beta^{A_j} \left(\sum_{j=1}^{J} \rho_j A_j - \sum_{j=1}^{J} \frac{\lambda_j}{\mu_1} \right) - \sum_{j=1}^{J} \rho_j \left(c_0 - \sum_{j=1}^{J} \frac{\lambda_j}{\mu_1} \right) \geq 0.$$

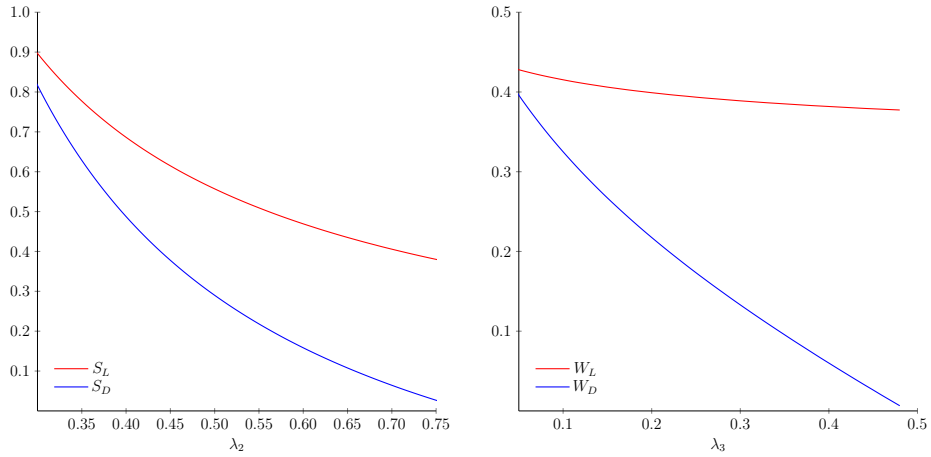
Using the fact that $c_0 < c$ and Equation (2.5), it is enough to show that the quantity

$$\Delta \overset{\text{def.}}{=} \sum_{j=1}^{J} \rho_j \beta^{A_j} \left(\sum_{i=1}^{J} \rho_j A_i - \sum_{i=1}^{J} \frac{\lambda_i}{\mu_1} \right) - \sum_{j=1}^{J} \rho_j \left(\sum_{i=1}^{J} A_i \rho_j \beta^{A_i} - \sum_{i=1}^{J} \frac{\lambda_i}{\mu_1} \right)$$

is positive. But this is clear since

$$\begin{aligned} \Delta &= \sum_{1 \leq i,j \leq J} \rho_i \rho_j \left(A_j \left(\beta^{A_i} - \beta^{A_j} \right) \right) + \sum_{1 \leq i,j \leq J} \rho_j \frac{\lambda_i}{\mu_1} \left(1 - \beta^{A_j} \right) \\ &= \sum_{1 \leq i < j \leq J} \rho_i \rho_j \left((A_j - A_i) \left(\beta^{A_i} - \beta^{A_j} \right) \right) + \sum_{1 \leq i,j \leq J} \rho_j \frac{\lambda_i}{\mu_1} \left(1 - \beta^{A_j} \right) \end{aligned}$$

and the terms of both series of the right hand side of this relation are non-negative due to the fact that $0 < \beta < 1$. □



(a) $J = 2, A_2 = 3, \lambda_1 = 0.2$
(R) conditions: $0.2633 < \lambda_2 < 0.7$

(b) $J = 3, A_2 = 2, A_3 = 3, \lambda_1 = \lambda_2 = 0.2$
(R) conditions: $0.03 < \lambda_3 < 0.49$

Figure 2.2 – Comparison between policies

Numerical experiments have been done to estimate the difference $W_L - W_D$, see Figure 2.2. The general conclusion is that, at moderate load under Condition (R), the downgrading algorithm performs quite well with only a small fraction of downgraded jobs. As it can be seen this is no longer valid for high load where, as expected, most of requests are downgraded but nobody is lost.

Application to Video Transmission

We consider now a link with large bandwidth, 10.0 Gbps, in charge of video streaming. Requests that cannot be immediately served are lost. Video transmission is offered in two standard qualities, namely, *Low Quality* (LQ) and *High Quality* (HQ). From Añorga *et al.* [AAS+15], the bandwidth requirement for YouTube's videos at 240p is 1485 Kbps, and for 720p it is 2737.27 Kbps.

Using the values above, after renormalisation, one takes $A_1 = 1$, $N = C^N = 7061$ and $A_2 = 2$, $c = 1$. Jobs arrive at rate λ_2 in this system asking for HQ transmission, but clients accept to watch the video in LQ. In particular $\lambda_1 = 0$. Service times are assumed to be the same for both qualities and taken as the unity, $\mu_1 = \mu_2 = 1$. Condition (R) is satisfied when

$$0.5 < \lambda_2 < 1.$$

We define $C_0 = \alpha C$, with $0 < \alpha < 1$. The quantity α_ε is defined as the largest value of α such that the loss probability of a job is less than $\varepsilon > 0$. With the notations of Section 2.4, we write

$$\alpha_\varepsilon \overset{\text{def.}}{=} \sup \left\{ \alpha \in (0,1) : \mathbb{P}\left(Y_{\ell^*} + C_0 > C \right) < \varepsilon \right\}.$$

Note that this is an approximation, since the variable Y_{ℓ^*} corresponds to the case when the scaling parameter N goes to infinity. Using the explicit expression of the distribution of Y_{ℓ^*} of Proposition 2.6, Figure 2.3 plots the threshold

α_ε that ensures a loss rate less than ε as a function of ε, for several values of λ_2. In the numerical example, taking $C_0 = 0.98C$ is sufficient to get a loss probability less than 10^{-7}.

Now let π_ε^- be the value of π^- defined by Corollary 2.1 for $C_0 = \alpha_\varepsilon C$. Recall that π_ε^- is the asymptotic equilibrium probability that a job is not downgraded is given by Relation (2.11),

$$\pi_\varepsilon^- = \frac{\alpha_\varepsilon}{\lambda_2} - 1.$$

For comparison, β is defined as the corresponding acceptance probability when no control is used in the system. We show in Figure 2.4 the relation between these quantities and the workload λ_2, for fixed loss rates of 10^{-3}, 10^{-6} and 10^{-9}. We have $\beta = 1 - 1/(2\lambda_2)$, see Proposition 6.19 of Robert [Rob03, p. 169]. The difference $\beta - \pi^-$ can be seen as the fraction of jobs which are downgraded for our policy but lost in the uncontrolled policy. Intuitively it can be seen as the price of not rejecting any job. Notice also that the curves plotting π_ε^- for $\varepsilon = 10^{-3}$, 10^{-6}, 10^{-9} are close and that β is larger than π^-. One remarks nevertheless that, for high loads, the system cannot hold these demands, because our policy is no longer effective.

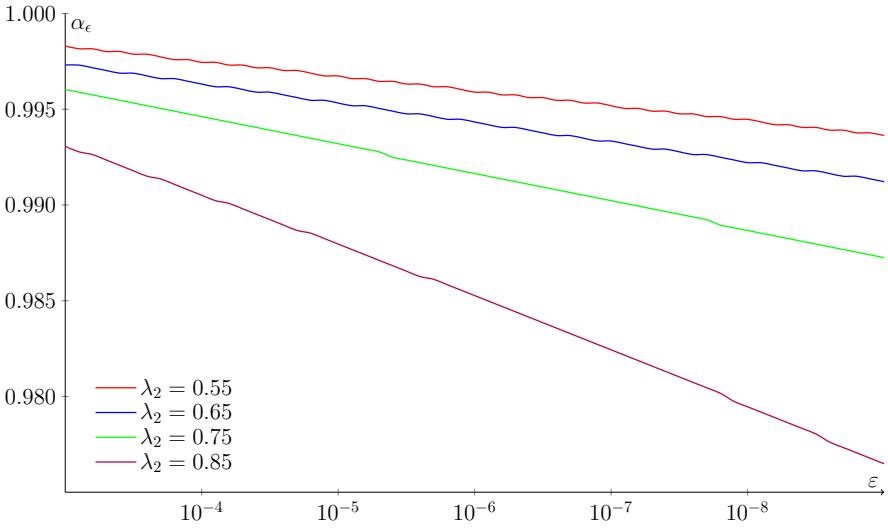


Figure 2.3 – Maximal threshold for a loss probability equal to ε.

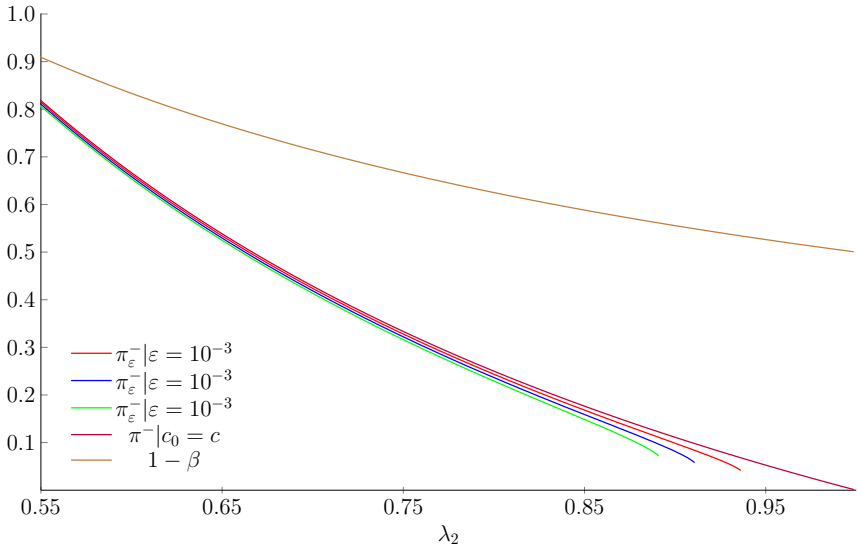


Figure 2.4 – Fraction of jobs served at requested bit-rate at equilibrium for the downgrading policy compared to the fraction of lost jobs in a pure loss system.

Chapter 3

Cooperative Schemes in the framework of multi-resource Cloud Computing

3.1 Introduction

Over the last years, Cloud Computing systems have become increasingly decentralised as processing facilities (data centres) have been moved closer to the final user in an effort to reduced the volume of data sent across networks. On average, these decentralised facilities dispose of a lower capacity of resources than centralised ones and may also face different flows of service requests. This results in a system where certain facilities may be overloaded whereas the resources of others can be underexploited.

This problem still gains in complexity, considering that the instantiation of a common virtual machine necessitates at least two different kinds of resources, namely RAM memory and CPU cores. The demand and supply of these resources might be imbalanced in a decentralised architecture. Not only can users request different amounts of each of the resources but the installed capacity of each resource may change from one processing facility to another. Due to the stochastic nature of service request arrivals and service times, mismatches between demand and supply may occur. Certain facilities might run out of one of the resources whereas the others resource have high amount of idle units, causing rejection of user demands. This is particularly critical in a system which disposes of a sufficiently large capacity on an aggregate level (summing the capacities of all the decentralised processing facilities) to serve the entire traffic of user demands. There thus exist potential gains from establishing cooperation schemes between data centres in order to handle the local demands while maintaining service standards.

It should be noted that in the framework of random access networks, allocation and dimensioning of resources have been widely investigated, beginning with the work of Erlang more than one century ago. However, from *circuit-switch* networks to Cloud Computing, most of the studies in this field focus on models and policies for single resource services.

One of the first among the few papers dealing with multi-resource alloc-

ation is Ghodsi *et al.* [GZH+11, GSZS12]. The authors propose the concept of *Dominant Resource Fairness* (DRF) allocations to share resources in an equitable (*fair*) manner between a fixed number of users. The authors develop the DRF by abstracting from the *weighted Max-Min fairness* to a multi-resource context. In Ghodsi *et al.* [GZH+11], the authors show the DRF has desirable properties for sharing mechanisms such as Sharing Incentive, Strategy-proofness, Envy-freeness, Pareto efficiency, Single Resource Fairness and Bottleneck Fairness in the static environment. In Ghodsi *et al.* [GSZS12] the authors define the *Dominant Resource Fair Queueing* (DRFQ), whose goal is to promote DRF allocations in time-evolving processing systems.

However, these *fair* allocations may lack applicability in many Cloud services where jobs are neither divisible nor preemptive nor postponable, contrary to what is assumed in the previously mentioned literature. For example, in many Cloud services such as Microsoft Azure [Mic], jobs are not divisible. A virtual machine may actually fail to be instantiated because there are not enough resources in the data centre. As this is more likely when the data centre works close to its full capacity, exploiting at full capacity is often avoided in practice. In such systems, every data centre usually has enough idle resources to accommodate an incoming request or the request will not be hosted locally. If offloading is not possible, these systems rather fit into the framework of *loss networks* presented in the seminal paper by Kelly [Kel91].

In the context of cooperation in single resource systems, we refer to the papers by Fricker *et al.* [FGRT16a] and Guillemin and Thompson [GT16]. They advocate the introduction of cooperative schemes in between processing facilities to circumvent load asymmetry arising from the decentralisation of Cloud services whose performance is determined solely by one resource. For the design of cooperation schemes in the multi-resource framework, the load of jobs arriving in each data centre has to be considered, as well as the charge each resource is subjected to at the local and at the global level. This issue is particularly relevant for services whose virtual machines are resource-specialised, necessitating very different ratios of each resource. See for instance the predefined virtual machine types from Google Cloud Platform [Goo]. In the worst case scenario, the traffic asymmetry may generate a situation where each data centre is running low on some specific resource while it is available abundantly in the other processing units.

To mitigate the negative effects on system performance in such a case, we propose a cooperation scheme where jobs are forwarded to another processing facility in order alleviate the local charge of the depleted resource. We propose the introduction of a threshold on each depleted resource (on local level) which, if surpassed, triggers the offloading of the jobs with the highest requirements of the scarce resource.

Our model and assumptions

The system we consider is composed of two data centres, each equipped with a finite capacity of `GB RAM` and `CPU cores`. There are two virtual machine configurations: type 1 occupies a big chunk of `GB RAM` and only 1 `CPU core` whereas type 2 uses only 1 `GB RAM` and requires a large amount of `CPU cores`.

Virtual machines of both types arrive at each data centre at different rates. The average service time only depends on the type of the job (virtual machine) but not on the hosting data centre. All of the necessary resources for the execution of a job should be available upon its instantiation. If the demanded resources are not free, the arriving virtual machine is either forwarded to another data centre or rejected. The resources are released after service completion.

Considering the large scale of data centres deployed in Cloud services, the service capability of the system is defined by the relation between the load of requests and the installed capacity. The present paper focuses on the cases where the capacity of the system on the aggregate level is large enough to accommodate the combined flow of jobs, meaning the system is *globally underloaded*. This setting is particularly interesting because it implies that the system could accommodate (with high probability) all jobs without losses on the aggregated level (see Kelly [Kel91] and Section 6.7 of Robert [Rob03, p. 164]). The aim of the presented policy is to enable the decentralised system to achieve the efficiency of a centralised architecture with the same capacity.

We focus on the case where both data centres are facing *local saturation* because each lacks one of the resources. Systematically, clients arriving at some data centre will be rejected due to the unavailability of the same resource which is clearly the saturated on. In this scenario, forwarding jobs after exhaustion of one of the resources is no longer effective, since jobs require both resources for the instantiation of their virtual machine. Although the probability of rejection of a job ought to be smaller than in isolated data centres, significant losses are still observed.

To bypass this, a set of thresholds is introduced to trigger the offloading of the jobs whenever a data centre faces a demand such that its resources might be depleted. The idea is to buffer some resources to avoid the complete jamming of a processing facility due to the lack of either resource. A threshold is set up for the most demanded resource at each data centre, selected by the relation between its local load and capacity. The study is therefore restricted to the implementation of thresholds $0 < \delta_p < 1$ in data centre p to avoid exhaustion of resource p by offloading the jobs of type p, for $1 \leq p \leq 2$. Thus, an arriving job of type p at data centre p is served locally only if resource p utilisation factor is bellow δ_p, otherwise it is forwarded to data centre $3 - p$, for $1 \leq p \leq 2$.

It is shown that the rejection rates of requests become negligible (with high probability for a large system) by triggering early enough the previously defined offloading mechanism. In fact, different thresholds can be set up for each resource, type of job and data centre to improve systems performance. However, if the thresholds are poorly dimensioned, communication channels can be jammed or resources which were *a priori* well dimensioned for the local charge can be saturated by the large volume of offloaded jobs.

Admission Control Policy. It is assumed that jobs of type j arrive at data centre k according to a Poisson process with rate $\lambda_{j,k}$, for $1 \leq j, k \leq 2$. Jobs of type j hold $A_{i,j}$ units of resource i during their service time which is exponentially distributed with mean $1/\mu_j$, and data centre k is equipped

with $G_{i,k}$ units of resource i, for $1 \leq i, j, k \leq 2$. Throughout this paper, the following notation is kept

$$A \overset{\text{def.}}{=} \begin{pmatrix} R & 1 \\ 1 & C \end{pmatrix} \quad \text{and} \quad G \overset{\text{def.}}{=} \begin{pmatrix} \mathcal{R}_1 & \mathcal{R}_2 \\ \mathcal{C}_1 & \mathcal{C}_2 \end{pmatrix}.$$

Let $\ell = (\ell_{j,k},\, 1 \leq j, k \leq 2)$ represent the number of jobs of type j at data centre k at a certain moment, for $1 \leq j, k \leq 2$. The admission and offloading mechanism rules that, in the event of an arrival of a job of type 1 at data centre 1, if

— $R\ell_{1,1} + \ell_{2,1} < \mathcal{R}_1\delta_1$ and $(\ell_{1,1} + 1) + C\ell_{2,1} \leq \mathcal{C}_1$ then

the request is hosted in data centre 1, otherwise the request is forwarded to the other processing facility. Once forwarded to data centre 2, if

— $R\,(\ell_{1,2} + 1) + \ell_{2,2} \leq \mathcal{R}_2$ and $(\ell_{1,2} + 1) + C\ell_{2,2} \leq \mathcal{C}_2$ then

the request is hosted in data centre 2, otherwise the request is rejected.

In the event of an arrival of type 2 at data centre 1, if

— $R\ell_{1,1} + (\ell_{2,1} + 1) \leq \mathcal{R}_1$ and $\ell_{1,1} + C\,(\ell_{2,1} + 1) \leq \mathcal{C}_1$ then

the request is hosted in data centre 1, otherwise the request is forwarded to data centre 2. Once in the other processing facility, if

— $R\ell_{1,2} + (\ell_{2,2} + 1) \leq \mathcal{R}_2$ and $\ell_{1,2} + C\,(\ell_{2,2} + 1) \leq \mathcal{C}_2$ then

the request is hosted in data centre 2, otherwise the request is rejected.

At data centre 2 the mirror conditions rule the admission and offloading of jobs.

Because of the exponential assumptions for inter-arrival and service times, the stochastic vector describing the number of jobs of type j at data centre k is a Markov process in a finite state space. The invariant distribution of this Markov process does likely not have a closed expression, so a scaling approach is used in order to obtain qualitative and quantitative results information about the system, considering its large scale.

Scaling Approach. To study the offloading policy, a scaling approach is used such as in Kelly [Kel91, Kel86] in the analysis of *loss networks*. We assume that the loads and installed capacities of resources are of the same order, meaning both are scalable by a positive factor N. Namely,

$$\begin{cases} G_{i,k} \mapsto G_{i,k}N & \text{for } 1 \leq i, k \leq 2, \\ \lambda_{j,k} \mapsto \lambda_{j,k}N & \text{for } 1 \leq j, k \leq 2. \end{cases}$$

Asymptotic results will be derived for the rescaled system, when N goes to infinity. At the equilibrium and if the thresholds are well chosen,

— the probability of rejecting jobs in both data centres converges to 0;

— the probabilities of forwarding a type 2 job from data centre 1 and type 1 job from data centre 2 converge to 0;

— the probabilities of forwarding a job of type 1 from data centre 1 and type 2 from data centre 2 can be explicitly derived.

Using some approximations (which are rigorously verified later), we deduce that the thresholds δ_1 and δ_2 are well chosen if

$$\begin{cases} \dfrac{1}{\mathcal{R}_1}\left[\dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}R + \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2} - \mathcal{R}_2\right] < \delta_1 < 1, \\ \dfrac{1}{\mathcal{C}_2}\left[\dfrac{\lambda_{2,2}+\lambda_{2,1}}{\mu_2}C + \dfrac{\lambda_{1,2}+\lambda_{1,1}}{\mu_1} - \mathcal{C}_1\right] < \delta_2 < 1; \end{cases}$$

Moreover, using the same program as in Hunt and Kurtz [HK94] we are able to study the process of the renormalised number of customers in the system trough a dynamical system, when N goes to infinity. We show show that this dynamical system converges indeed for a fixed point which is the same obtained by the approximations previously established.

Outline of the paper

The paper is organised as follows: In Section 3.2, we describe the model, the offloading policy and the stochastic processes characterising the functioning of the model. In Section 3.3, we introduce the scaled version of the system and analyse system performance under Kelly's regime. We derive approximations of the main metrics of the scaled version of the system in steady state, such as the average occupation of each data centre and the offloading rates. We finish this section by defining the optimal operation range of the threshold parameters. In Section 3.4, we study the transient behaviour of the system and the dynamical system associated with the normalised number of virtual machines in each data centre.

3.2 Model description and notation

We consider a Cloud Computing environment where virtual machines (or jobs) are instantiated upon the allocation of `GB RAM` and `CPU core`, denoted resource $i \in \{1, 2\}$, respectively. To capture the particular aspect of resource specialised (unbalanced) virtual machines, we consider two types of jobs arriving to the system, denoted type $j \in \{1, 2\}$. Jobs of type 1 require $R > 1$ `GB RAM` and 1 `CPU core`, whereas jobs of type 2 require 1 `GB RAM` and $C > 1$ `CPU core`. We write the matrix A, analogue to the routing matrix in Kelly [Kel91] for loss networks with fixed routing, as

$$A \stackrel{\text{def.}}{=} (A_{i,j},\, 1 \le i, j \le 2) = \begin{pmatrix} R & 1 \\ 1 & C \end{pmatrix}.$$

We denote by A_j the column vector $(A_{i,j},\, 1 \le i \le 2)$ which represents the requirements of jobs of type j, for $1 \le j \le 2$. Throughout this paper, vectorial and matrix equalities as well as inequalities are defined element wise, hence for $u, v \in \mathbb{R}^d$, $u \# v$ if and only if $u_1 \# v_1, u_2 \# v_2, \ldots, u_d \# v_d$, for $\# \in \{>, \ge, =, \le, <\}$.

The system is composed of two parallel data centres denoted $k \in \{1, 2\}$, respectively. Each data centre k is equipped with some large (but finite)

amount $\mathcal{R}_k$ of `GB RAM` and $\mathcal{C}_k$ of `CPU cores`. We write the local capacity matrix G, as

$$G \stackrel{\text{def.}}{=} (G_{i,k},\, 1 \leq i,k \leq 2) = \begin{pmatrix} \mathcal{R}_1 & \mathcal{R}_2 \\ \mathcal{C}_1 & \mathcal{C}_2 \end{pmatrix},$$

where $G_{i,k}$ is the capacity of resource i at data centre k. Requests of type j arrive at data centre k according to a Poisson process $\mathcal{N}_{\lambda_{j,k}}$ with rate $\lambda_{j,k}$, for $1 \leq j,k \leq 2$. If accepted, a job of type j holds the required units of each resource in the hosting data centre during its service time, which is exponentially distributed with mean μ_j^{-1}. After service completion, the job leaves the system and releases the occupied units of both resources. It is assumed that the service rates can be different, conditioned that $\mu_p/\mu_{3-p} < RC$, for $1 \leq p \leq 2$.

See Figure 3.1 for an illustration of the arrival processes in the system.

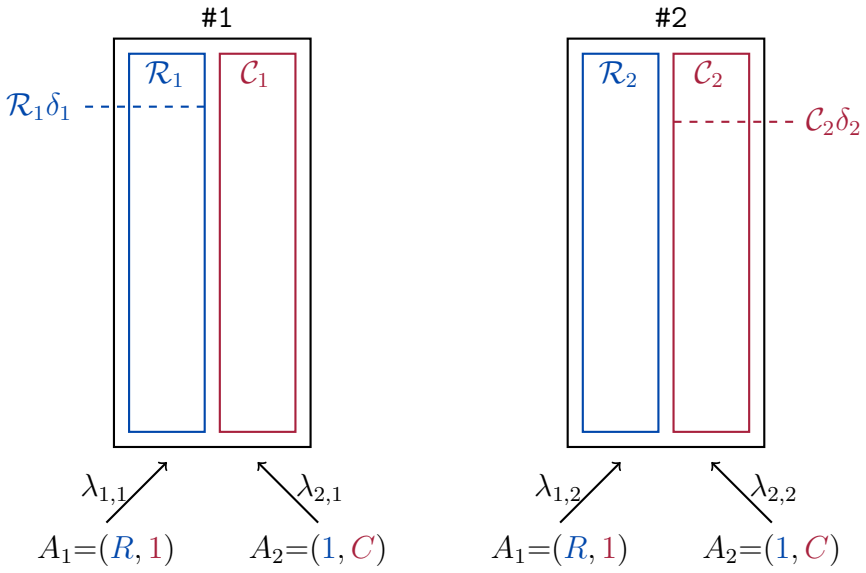


Figure 3.1 – System configuration

Jobs of type j arriving at data centre k with rate $\lambda_{j,k}$ requesting A_j units of the resources.

Load and Capacity

In order to understand the behaviour of the system, we need to examine the load its components are subject to and the ability of each resource to meet user demand. As we consider multiple resources and multiple data centres, we have to abstract the notions of service rate and, hence, of saturation from the ones used in the classical framework of queuing systems. This leads us to consider separately the local and global properties of the system for each of the resources.

The *load and capacity* characterisation is particularly relevant when we consider the large scale of the system, further explored in Section 3.3. These characteristics reflect the ability of the system to handle properly the current

level of access and elucidates whether the fraction of jobs which will eventually be rejected is negligible or not.

The *local load* of a resource i in data centre k is given by

$$H_{i,k} \overset{\text{def.}}{=} A_{i,1}\frac{\lambda_{1,k}}{\mu_1} + A_{i,2}\frac{\lambda_{2,k}}{\mu_2}, \text{ for } 1 \leq i,k \leq 2$$

This measure contrasts with the local capacity of each resource. In large scale networks, resource i is said to be saturated at data centre k if $H_{i,k} > G_{i,k}$, see Kelly [Kel91]. In the context of cooperation between data centres, we need to keep in mind the global aspects of the system. The system is said to be *globally underloaded* if and only if

$$H_{i,1} + H_{i,2} < G_{i,1} + G_{i,2}, \text{ for } 1 \leq i \leq 2.$$

These relations indicate the capacity of the system to meet user demand if all resources would be aggregated in one large structure to serve the combined flows of requests.In conclusion, the scenario we are studying is characterised by local saturation of one resource of such a globally underloaded system

$$H_{p,p} > G_{p,p} \quad \text{and} \quad H_{p,p} + H_{q,p} < G_{p,p} + G_{q,p}$$

for $q = 3 - p$ and $1 \leq p \leq 2$.

Anticipated Offloading

To avoid local mismatching between the load and capacity of a resource, we propose to anticipate the offloading of some jobs. The general idea is to locally mitigate the saturation of a resource by forwarding the jobs with the largest demand for the concerned resource to another data centre much before actual depletion occurs.

Algorithm definition.

> If the fraction of `GB RAM` being used at data centre 1 is above some threshold $\delta_1 \in (0,1)$, then jobs of type 1 are systematically forwarded to data centre 2 (`GB RAM` being in high demand at data centre 1). If data centre 2 cannot host the job, then the request is then rejected. The symmetric admission scheme is applied to jobs of type 2 at data centre 2 according to some threshold $\delta_2 \in (0,1)$ on the fraction of `CPU core` being used (CPU constituting the resource that is most demanded at data centre 2).
>
> Jobs of type 2 in data centre 1 and of type 1 in data centre 2 are forwarded if the original data centre cannot meet their demands upon their arrival.

To understand the effects of the deployment of such a policy, we need to look at the amount of jobs of each type instantiated in each data centre over time. Considering Poisson arrivals and exponentially distributed service times, this quantity is a Markov process described in the following. Using these notations, the control policy is formally defined in Algorithm 1.

Markovian process

Let us denote by $L_{j,k}(t)$ the number of jobs of type j instantiated at data centre k at time t, and

$$L(t) \overset{\text{def.}}{=} (L_{j,k}(t),\, 1 \leq j,k \leq 2)\,.$$

Because of the Poisson arrival and exponential service time assumptions, the process $(L(t)) = (L(t),\, t \geq 0)$ is a Markov process on

$$\mathcal{S} \overset{\text{def.}}{=} \left\{\ell \in \mathbb{N}^{2\times 2} :\, A\ell \leq G\right\}.$$

The positive elements of its $\mathcal{Q}$-matrix, $Q = (q(\ell,\ell'),\, (\ell,\ell') \in \mathcal{S}^2)$, are given by

$$q(\ell,\ell') = \begin{cases} \lambda_{j,k}\mathbb{1}_{\mathcal{A}_{j,k}} + \lambda_{j,3-k}\mathbb{1}_{\mathcal{A}^{\mathrm{c}}_{j,3-k}\cap\mathcal{B}_{j,k}} & \text{if} \quad \ell' = \ell + e_{j,k} \\ \mu_j \ell_{j,k}, & \text{if} \quad \ell' = \ell - e_{j,k}, \end{cases},$$

for $\ell,\ell' \in \mathcal{S}$ and $1 \leq j,k \leq 2$, with $e_{j,k} = \left(\mathbb{1}_{\{j=p,k=q\}},\, 1 \leq p,q \leq 2\right)$,

$$\mathcal{A}_{j,k} = \begin{cases} \{v_k < 0, u_k \geq A_j\} & \text{if} \quad j = k, \\ \{u_k \geq A_j\} & \text{if} \quad j \neq k, \end{cases} \quad \text{and} \quad \mathcal{B}_{j,k} = \{u_k \geq A_j\}$$

where $v_1 \overset{\text{def.}}{=} R\ell_{1,1} + \ell_{2,1} - \lfloor \mathcal{R}_1\delta_1 \rfloor$, $v_2 \overset{\text{def.}}{=} \ell_{1,2} + C\ell_{2,2} - \lfloor \mathcal{C}_2\delta_2 \rfloor$ and

$$u_k \overset{\text{def.}}{=} (G_{1,k} - R\ell_{1,k} - \ell_{2,k},\, G_{2,k} - \ell_{1,k} - C\ell_{2,k})\,, \text{ for } 1 \leq k \leq 2.$$

The invariant distribution of this process π being in general not known, we use a scaling approach. See Kelly [Kel86, Kel91], Hunt and Kurtz [HK94], Bean *et al.* [BGZ95, BGZ97] and Fricker *et al.* [FRT01]. For applications of such a framework in contexts similar to the one presented in this work, see also Fricker *et al.* [FGRT16a] and Guillemin and Thompson [GT16].

3.3 Scaling Results

Kelly's Regime

In this section, the system is investigated under Kellys regime which has been introduced to study the equilibrium properties of uncontrolled loss networks (see Kelly [Kel86, Kel91]). We assume that the capacity of the system resources and the arrival rates of jobs are scalable by a positive factor N. Namely, the capacities of resource i at data centre k is replaced by $G^N_{i,k}$, with

$$\lim_{N\to+\infty} \frac{G^N_{i,k}}{N} = G_{i,k}, \text{ for } 1 \leq i,k \leq 2,$$

and the arrival rate of jobs of type j at data centre k by $\lambda^N_{j,k}$, with

$$\lim_{N\to+\infty} \frac{\lambda^N_{j,k}}{N} = \lambda_{j,k}, \text{ for } 1 \leq j,k \leq 2.$$

To indicate the dependency of the number of virtual machines (or jobs) of each type in this system on the parameter N, we denote by $L^N_{j,k}(t)$ the number of jobs of type j instantiated at data centre k at time t and

$$L^N(t) \stackrel{\text{def.}}{=} \left(L^N_{j,k}(t),\, 1 \leq j,k \leq 2\right).$$

Also, let

$$m^N(t) \stackrel{\text{def.}}{=} (m_{i,k}(t),\, 1 \leq i,k \leq 2) \stackrel{\text{def.}}{=} G^N - AL^N(t)$$

represent the amount of free slots of resource i in data centre k at time t (in data centre k the vector $m_k(t) = (m_{1,k}(t), m_{2,k}(t))$ denotes the amount of idle units of both resources) and

$$Y^N(t) \stackrel{\text{def.}}{=} \left(Y^N_p(t),\, 1 \leq p \leq 2\right) \stackrel{\text{def.}}{=} \left(RL^N_{1,1}(t) + L^N_{2,1}(t) - \lfloor N\mathcal{R}_1\delta_1 \rfloor,\, L^N_{1,2}(t) + CL^N_{2,2}(t) - \lfloor N\mathcal{C}_2\delta_2 \rfloor\right) \quad (3.1)$$

denote the offload triggering process of jobs of type p at data centre p. If the utilisation of resource p is above the threshold δ_p at time t then $Y^N_p(t) \geq 0$, otherwise, if the amount of resource p being used is below the threshold δ_p then $Y^N_p(t) < 0$.

Load and Capacity under Kelly's Regime

As said in Section 3.2, we focus our efforts on the case where each data centre is running low on one resource and has idle units of the other. Following the definition introduced before, we consider a system which is *globally underloaded* but where each data centre faces local saturation of one of the resources. Formally, the system is *globally underloaded* if

$$\begin{cases} \dfrac{R}{\mu_1}\left(\lambda_{1,1} + \lambda_{1,2}\right) + \dfrac{1}{\mu_2}\left(\lambda_{2,1} + \lambda_{2,2}\right) < \mathcal{R}_1 + \mathcal{R}_2, \\ \dfrac{1}{\mu_1}\left(\lambda_{1,1} + \lambda_{1,2}\right) + \dfrac{C}{\mu_2}\left(\lambda_{2,1} + \lambda_{2,2}\right) < \mathcal{C}_1 + \mathcal{C}_2. \end{cases}$$

Locally, in data centre 1, resource 1 is exhausted and resource 2 is not if

$$\frac{\lambda_{1,1}}{\mu_1}R + \frac{\lambda_{2,1}}{\mu_2} > \mathcal{R}_1 \quad \text{and} \quad \frac{\lambda_{1,1}}{\mu_1} + \frac{\lambda_{2,1}}{\mu_2}C < \mathcal{C}_1. \quad (3.2)$$

In data centre 2, resource 2 is exhausted and resource 1 is not if

$$\frac{\lambda_{1,2}}{\mu_1}R + \frac{\lambda_{2,2}}{\mu_2} < \mathcal{R}_2 \quad \text{and} \quad \frac{\lambda_{1,2}}{\mu_1} + \frac{\lambda_{2,2}}{\mu_2}C > \mathcal{C}_2. \quad (3.3)$$

Also, no resource is exclusively the bottleneck of any data centre k if and only if

$$\mathcal{C}_k < \mathcal{R}_k C \quad \text{and} \quad \mathcal{R}_k < R\mathcal{C}_k,$$

for $1 \leq k \leq 2$. This condition implies that there exists, in the limiting case when N goes to infinity, some configuration of the available types of virtual machines which could exhaust both resources at the same time in both data centres. Otherwise, if this condition is violated at any data centre, then its performance is ruled solely by one of the resources. This assumption arises from a particular application of *Farkas' Lemma*, see Roos [Roo09].

Approximations from Kelly's Regime

The following results are intended to provide a convenient framework for the analysis of the operational boundaries of thresholds $\delta = (\delta_1, \delta_2)$, considering in particular the limiting case when the scaling factor N goes to infinity. We assess the effects of the choice of different thresholds δ on some key quantities related to the operation of this policy such as the probability of forwarding jobs from one data centre to another or the probability of a job getting blocked at a given data centre. Based on these results, we obtain some upper and lower boundaries for the different thresholds δ.

We begin by considering how the system would behave in the steady state. If the renormalised process describing the number of jobs in the system is at equilibrium in the fixed point $x(\delta) = (x_{j,k}(\delta),\, 1 \leq j,k \leq 2)$ then, the flow conservation equations, which rule the proportion of jobs that are either accepted, offloaded or rejected, imply that for $1 \leq j,k \leq 2$ the following relation holds

$$\mu_j x_{j,k}(\delta) = \lambda_{j,k}(1 - \gamma_{j,k}(\delta)) + (1 - \beta_{j,k}(\delta))\lambda_{j,3-k}\gamma_{j,3-k}(\delta), \tag{3.4}$$

where $\beta_{j,k}(\delta)$ and $\gamma_{j,k}(\delta)$ are respectively the rejection and forwarding probabilities for jobs of type j at data centre k as a function of δ. If there exists a set of thresholds $\delta = (\delta_1, \delta_2)$ such that both resources in both data centres are underloaded after the deployment of this scheme, then, with probability equals to 1, no job is rejected ($\beta_{j,k}(\delta) = 0$, for $1 \leq j,k \leq 2$) and only jobs of type 1 are forwarded from data centre 1 and jobs of type 2 from data centre 2 ($\gamma_{p,3-p}(\delta) = 0$, for $1 \leq p \leq 2$). Therefore, if these parameters exist, the following relations must hold

$$\begin{cases} \dfrac{R}{\mu_1}\lambda_{1,1}\left(1 - \gamma_1(\delta)\right) + \dfrac{1}{\mu_2}\left(\lambda_{2,1} + \lambda_{2,2}\gamma_2(\delta)\right) < \mathcal{R}_1, \\ \dfrac{1}{\mu_1}\left(\lambda_{2,1} + \lambda_{1,1}\gamma_1(\delta)\right) + \dfrac{C}{\mu_2}\lambda_{2,2}\left(1 - \gamma_2(\delta)\right) < \mathcal{C}_2, \end{cases} \tag{3.5}$$

and

$$\begin{cases} \dfrac{1}{\mu_1}\lambda_{1,1}\left(1 - \gamma_1(\delta)\right) + \dfrac{C}{\mu_2}\left(\lambda_{2,1} + \lambda_{2,2}\gamma_2(\delta)\right) < \mathcal{C}_1, \\ \dfrac{R}{\mu_1}\left(\lambda_{2,1} + \lambda_{1,1}\gamma_1(\delta)\right) + \dfrac{1}{\mu_2}\lambda_{2,2}\left(1 - \gamma_2(\delta)\right) < \mathcal{R}_2, \end{cases} \tag{3.6}$$

where $\gamma_p(\delta) = \gamma_{p,p}(\delta)$, for $1 \leq p \leq 2$. Equation (3.4) is rewritten as

$$\begin{cases} \mu_p x_{p,p}(\delta) = \lambda_{p,p}(1 - \gamma_p(\delta)), \\ \mu_q x_{q,p}(\delta) = \lambda_{q,p} + \lambda_{q,q}\gamma_q(\delta), \end{cases} \tag{3.7}$$

for $q = 3 - p$ and $1 \leq p \leq 2$.

For practical reasons, we consider the case where neither data centre 1 has a sufficient large capacity of resource 2 to host the combined flow of jobs of type 2 nor data centre 2 disposes of enough units of resource 1 to host the combined flows of jobs of type 1, which is true if

$$\mathcal{C}_1 < \frac{\lambda_{2,1} + \lambda_{2,2}}{\mu_2}C \quad \text{and} \quad \mathcal{R}_2 < \frac{\lambda_{1,2} + \lambda_{1,1}}{\mu_1}R. \tag{3.8}$$

Equation (3.8) combined with the Equations (3.2) and (3.3) implies that $0 < \gamma_p(\delta) < 1$, for $1 \leq p \leq 2$. Moreover $\gamma_p(\delta)$ represents the probability that the p coordinate of the process $(Y^N(t)) = (Y^N(t), t \geq 0)$ is positive, meaning resource p is used above the threshold δ_p. If the system is stable, the equilibrium point $x(\delta)$ must be such that the process $(Y^N(t))$ stays around 0 (in both coordinates) - meaning the process is neither the whole time positive nor negative. Thus, supposing that in the steady state resource p is used "around" the threshold δ_p, for $1 \leq p \leq 2$, we obtain two additional relations

$$\begin{cases} Rx_{1,1}(\delta) + x_{2,1}(\delta) = \mathcal{R}_1\delta_1, \\ x_{1,2}(\delta) + Cx_{2,2}(\delta) = \mathcal{C}_2\delta_2. \end{cases} \tag{3.9}$$

Equations (3.7) and (3.9) describe a linear system, whose unique solution is $x^\star(\delta) = (x^\star_{j,k}(\delta),\, 1 \leq j,k \leq 2)$ and $\gamma^\star(\delta) = \left(\gamma^\star_p(\delta), 1 \leq p \leq 2\right)$, given by

$$\begin{cases} x^\star_{1,1}(\delta) &= \mathcal{R}_1\delta_1 C + \mathcal{C}_2\delta_2 - \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1} - \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2}C, \\ x^\star_{2,1}(\delta) &= -\mathcal{C}_2\delta_2 R - \mathcal{R}_1\delta_1 + \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}R + \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2}RC, \\ x^\star_{1,2}(\delta) &= -\mathcal{R}_1\delta_1 C - \mathcal{C}_2\delta_2 + \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2}C + \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}RC, \\ x^\star_{2,2}(\delta) &= \mathcal{C}_2\delta_2 R + \mathcal{R}_1\delta_1 - \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2} - \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}R \end{cases} \tag{3.10}$$

and

$$\begin{cases} \gamma^\star_1(\delta) = \dfrac{\mu_1}{(RC-1)\lambda_{1,1}}\left[\mathcal{R}_1\delta_1 C + \mathcal{C}_2\delta_2 - \dfrac{(\lambda_{1,1}+\lambda_{1,2})}{\mu_1} - \dfrac{C\,(\lambda_{2,1}+\lambda_{2,2})}{\mu_2}\right], \\ \gamma^\star_2(\delta) = \dfrac{\mu_2}{(RC-1)\lambda_{2,2}}\left[\mathcal{R}_1\delta_1 + \mathcal{C}_2\delta_2 R - \dfrac{R\,(\lambda_{1,1}+\lambda_{1,2})}{\mu_1} - \dfrac{(\lambda_{2,1}+\lambda_{2,2})}{\mu_2}\right]. \end{cases}$$

We are now able to express the main result of this paper. We present an operational bandwidth for the thresholds δ_1 and δ_2 which ensure that both resources run beneath the total capacity of the system and that rejections and unnecessary transfers of jobs from one data centre to another are avoided as much as possible. The upper and lower boundaries for the thresholds δ_1 and δ_2 such that Inequalities (3.5) and (3.6) hold are defined in the following proposition.

Proposition 3.1 (Operational Parameters of δ)**.** *If $\bar{\delta}_p < \delta_p < 1$, for $1 \leq p \leq 2$, then Conditions (3.5) and (3.6) hold, where*

$$\begin{cases} \bar{\delta}_1 \overset{\text{def.}}{=} \dfrac{1}{\mathcal{R}_1}\left(\dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}R + \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2} - \mathcal{R}_2\right), \\ \bar{\delta}_2 \overset{\text{def.}}{=} \dfrac{1}{\mathcal{C}_2}\left(\dfrac{\lambda_{1,2}+\lambda_{1,1}}{\mu_1} + \dfrac{\lambda_{2,2}+\lambda_{2,1}}{\mu_2}C - \mathcal{C}_1\right), \end{cases}$$

such that $x^\star(\delta) \in \mathcal{D} \overset{\text{def.}}{=} \left\{x \in \mathbb{R}_+^{2\times 2} : Ax < G\right\}$ and $\gamma^\star(\delta) \in (0,1)^2$.

Proof. Combining Equation (3.10) and Inequality (3.6), we define the real-valued function $\varphi : (0,1)^2 \to \mathbb{R}_+^2$, $\varphi(\delta) = (\varphi_1(\delta), \varphi_2(\delta))$, such that

$$\begin{cases} \varphi_1(\delta) \stackrel{\text{def.}}{=} \mathcal{C}_1 - \dfrac{\lambda_{1,1}\left(1-\gamma_1^\star(\delta)\right)}{\mu_1} - \dfrac{\lambda_{2,1}+\lambda_{2,2}\gamma_2^\star(\delta)}{\mu_2} C, \\ \varphi_2(\delta) \stackrel{\text{def.}}{=} \mathcal{R}_2 - \dfrac{\lambda_{2,1}+\lambda_{1,1}\gamma_1^\star(\delta)}{\mu_1} R - \dfrac{\lambda_{2,2}\left(1-\gamma_2^\star(\delta)\right)}{\mu_2}. \end{cases}$$

It is simple to see that $\bar{\delta} = (\bar{\delta}_1, \bar{\delta}_2)$ is the unique element of the set $\varphi^{-1}(0)$ and is given by

$$\begin{cases} \bar{\delta}_1 = \dfrac{1}{\mathcal{R}_1}\left[\dfrac{R}{\mu_1}\left(\lambda_{1,1}+\lambda_{1,2}\right) + \dfrac{1}{\mu_2}\left(\lambda_{2,1}+\lambda_{2,2}\right) - \mathcal{R}_2\right], \\ \bar{\delta}_2 = \dfrac{1}{\mathcal{C}_2}\left[\dfrac{C}{\mu_2}\left(\lambda_{2,2}+\lambda_{2,1}\right) + \dfrac{1}{\mu_1}\left(\lambda_{1,2}+\lambda_{1,1}\right) - \mathcal{C}_1\right]. \end{cases}$$

Hence, if $\delta > \bar{\delta}$ then Inequality (3.6) holds, since the gradient $\nabla_\delta \varphi(\delta)$ is non-negative. Similarly, if $\delta < (1,1)$ then Inequality (3.5) holds. Therefore, if $\bar{\delta} < \delta < (1,1)$ then $Ax^\star(\delta) < G$.

In addition, all the elements of the gradient $\nabla_\delta \gamma(\delta)$ are negative. Thus, the single elements in $\gamma^{-1}(0)$ and $\gamma^{-1}(1)$ define respectively an upper and a lower boundary for δ with

$$\gamma^{-1}(0) = \left\{ \left(\frac{\frac{\lambda_{1,1}}{\mu_1}R + \frac{\lambda_{2,1}}{\mu_2}}{\mathcal{R}_1} \;,\; \frac{\frac{\lambda_{1,2}}{\mu_1} + \frac{\lambda_{2,2}}{\mu_2}C}{\mathcal{C}_2} \right) \right\}$$

and

$$\gamma^{-1}(1) = \left\{ \left(\frac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2 \mathcal{R}_1}, \; \frac{\lambda_{1,2}+\lambda_{1,1}}{\mu_1 \mathcal{C}_2} \right) \right\}.$$

If Equations (3.2) and (3.3) hold then both elements of $\gamma^{-1}(0)$ are greater than 1, and if Equation (3.8) holds then $\gamma^{-1}(1) < \bar{\delta}$. Therefore, if $\bar{\delta} < \delta < (1,1)$ then $0 < \gamma_1^\star(\delta), \gamma_2^\star(\delta) < 1$. □

It is worth noting that the lower boundary $\bar{\delta}$ is obtained by assessing which proportion of jobs can be offloaded from one data centre to another without jamming the *a priori* underloaded resources in the hosting data centre. With this in mind, we are able to show that the point $x^\star(\delta)$ (and the associated offloading rates $\gamma^\star(\delta)$) is indeed the equilibrium point of the dynamical system where the probability that each coordinate of $Y^N(t)$ is positive is $\gamma^\star(\delta)$. This point captures the first order behaviour of the system and the probability that (at equilibrium) a job is transferred from one data centre to another.

The limiting process

We are now investigating the asymptotic behaviour of the process $(Y^N(t))$ defined by Relation (3.1). Recall that the resource p at data centre p is operating above the threshold level δ_p at time t if $Y_p^N(t) \geq 0$ or bellow the threshold level if $Y_p^N(t) < 0$, for $1 \leq p \leq 2$.

If N is large, up to the level of a change in time scale, the processes $(Y^N(t))$ and $(L^N(t)/N)$ coexist in different time scales. The fast process $(Y^N(t))$ sees the show process $(L^N(t)/N)$ as fixed while the slow process sees the fast process at equilibrium. Using the same framework of Hunt and Kurtz [HK94] we have that the process $(Y^N(t/N))$ converges in distribution to a random walk $U_{x(t)}(t)$ on the extended integers line, where $x(t)$ is the limit of $(L^N(t)/N)$, seen as constant by $(Y^N(t/N))$. In the following, we introduce the random walk $U_x(t)$.

Definition 3.1. *For fixed* $x = (x_{j,k},\, 1 \leq j,k \leq 2) \in \mathbb{R}_+^{2\times 2}$, *let*

$$(U_x(t)) = ((U_{x,h}(t),\, U_{x,v}(t)),\, t \geq 0)$$

be the Markov process on $\mathbb{Z}^2$ *with the positive elements of its* $\mathcal{Q}$*-matrix,*

$$Q_U = (q_x(n,n'),\, n,n' \in \mathbb{Z}^2),$$

given by

$$q_x(n,n+b) = \begin{cases} \lambda_{1,1}\mathbb{1}_{\{n_1<0\}} & \text{if} \quad b=(R,0),\\ \lambda_{2,1}+\lambda_{2,2}\mathbb{1}_{\{n_2\geq 0\}} & \text{if} \quad b=(1,0),\\ \lambda_{1,2}+\lambda_{1,1}\mathbb{1}_{\{n_1\geq 0\}} & \text{if} \quad b=(0,1),\\ \lambda_{2,2}\mathbb{1}_{\{n_2<0\}} & \text{if} \quad b=(0,C),\\ \mu_1 x_{1,1} & \text{if} \quad b=-(R,0),\\ \mu_2 x_{2,1} & \text{if} \quad b=-(1,0),\\ \mu_1 x_{1,2} & \text{if} \quad b=-(0,1),\\ \mu_2 x_{2,2} & \text{if} \quad b=-(0,C). \end{cases}$$

In the following, we define the regions Δ_0 where the process $U_x(t)$ is ergodic.

Definition 3.2. *The region* $\Delta_0 \subset \mathbb{R}_+^{2\times 2}$ *is given by*

$$\Delta_0 \overset{\text{def.}}{=} \Delta_A \cap (\Delta_{B.1} \cup \Delta_{B.2} \cup \Delta_{B.3}),$$

such that

— $x \in \Delta_A$ *if and only if*

$$\begin{cases} R\lambda_{1,1}+\lambda_{2,1} > R\mu_1 x_{1,1}+\mu_2 x_{2,1},\\ \lambda_{1,2}+C\lambda_{2,2} > \mu_1 x_{1,2}+C\mu_2 x_{2,2}. \end{cases}$$

— $x \in \Delta_{B,1}$ *if and only if*

$$\begin{cases} \lambda_{2,1}+\lambda_{2,2} < R\mu_1 x_{1,1}+\mu_2 x_{2,1},\\ \lambda_{1,2}+\lambda_{1,1} < \mu_1 x_{1,2}+C\mu_2 x_{2,2}. \end{cases}$$

— $x \in \Delta_{B,2}$ *if and only if*

$$\begin{cases} \lambda_{2,1}+\lambda_{2,2} < R\mu_1 x_{1,1}+\mu_2 x_{2,1},\\ \lambda_{1,2}+\lambda_{1,1} > \mu_1 x_{1,2}+C\mu_2 x_{2,2},\\ \lambda_{1,2}+\lambda_{1,1}+\dfrac{\lambda_{2,1}+\lambda_{2,2}}{R} < \mu_1 x_{1,2}+C\mu_2 x_{2,2}+\mu_1 x_{1,1}+\dfrac{\mu_2 x_{2,1}}{R}. \end{cases}$$

— $x \in \Delta_{B,3}$ *if and only if*

$$\begin{cases} \lambda_{2,1} + \lambda_{2,2} + \dfrac{\lambda_{1,2} + \lambda_{1,1}}{C} < R\mu_1 x_{1,1} + \mu_2 x_{2,1} + \dfrac{\mu_1 x_{1,2}}{C} + \mu_2 x_{2,2}, \\ \lambda_{2,1} + \lambda_{2,2} > R\mu_1 x_{1,1} + \mu_2 x_{2,1}, \\ \lambda_{1,2} + \lambda_{1,1} < \mu_1 x_{1,2} + C\mu_2 x_{2,2}. \end{cases}$$

In the following proposition, we show that if $x \in \Delta_0$ then the process $(U_x(t))$ is ergodic, using the framework in Kingman [Kin61], Rybko and Stolyar [RS92] and Malyshev [Mal93]. This result builds upon the the convergence of the *fluid limits* associated with the process $(U_x(t))$ to zero in a finite time.

Fluid Limits Consider the stochastic process $(X(t))$ which takes values in $\mathbb{Z}^d$ with $X(0) = x$. If $\|x\| \stackrel{\text{def.}}{=} |x_1| + \cdots + |x_d| = N$ then a fluid limit associated with $(X(t))$ is a stochastic process and one of the limits of

$$\left(\bar{X}^N(t)\right) = \left(\frac{X\left(Nt\right)}{N}\right)$$

when N goes to infinity. Note that $\bar{X}^N(0)$ is an element in the unit sphere of $\mathbb{R}^d$. Particularly, fluid limits are a convenient tool for the study of Markov processes behaving like random walks (at least locally). If the fluid limits of a Markov process converge to zero in a deterministic finite time then the process is ergodic. See Corollary 9.8 of Robert [Rob03, p. 259].

Proposition 3.2. *The Markov process* $(U_x(t))$ *is ergodic if* $x \in \Delta_0$.

Proof. The ergodicity conditions for this Markov process are derived by determining under which conditions the associated *fluid limits* converge to zero after some finite time. Kingman [Kin61] takes a similar approach in the analysis of reflected random walks in $\mathbb{N}^2$, observing the fact that far away from the axes the (first order) evolution of these processes is solely defined by their mean drifts.

Let $\varphi : \mathbb{Z}^2 \to \mathbb{R}^2$ be the drift at any point $n \in \mathbb{Z}^2$, given by

$$\varphi(n) \stackrel{\text{def.}}{=} \sum_{\substack{n \in \mathbb{Z}^2 \\ n \neq n'}} q_x(n, n')(n' - n).$$

The drift in the interior of each quadrant p is the constant vector

$$\theta_p = (\theta_{p,h}, \theta_{p,v}) = \varphi(n),$$

for any n in the quadrant p, for $1 \leq p \leq 4$, see Figure 3.2.

In the following, we prove that if $x \in \Delta_A \cap \Delta_{B,2}$ then any fluid limit of such process starting from some initial state $u \in \mathbb{N}^2$ reaches the ordinate axis in finite time and then evolves along this axis towards the origin of the plane. Once the fluid limit reaches abscissa axis (at the origin $\mathbb{R}^2$), it stays.

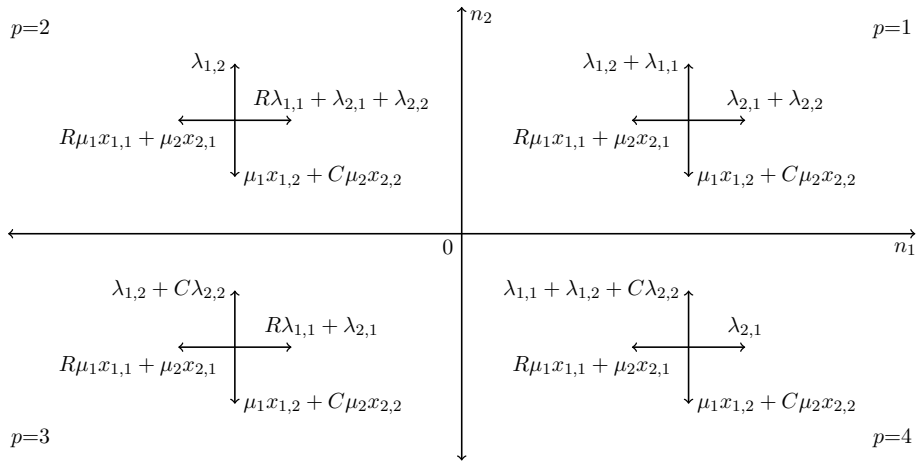


Figure 3.2 – Drifts on the quarter planes of $\mathbb{Z}^2$

The fluid limit in the interior of $\mathbb{N}^2$ We start by showing that from any initial state $u \in \mathbb{N}^2$, the process $(U_x(t))$ reaches either axis in some finite time.

For $N \geq 1$, we define the process

$$\left(U_x^N(t)\right) \overset{\text{def.}}{=} \left(\left(U_{x,h}^N(t),\, U_{x,v}^N(t)\right),\, t \geq 0\right)$$

with the same $\mathcal{Q}$-matrix as the process $(U_x(t))$ and initial value

$$u = (\lfloor N\alpha \rfloor, \lfloor N - N\alpha \rfloor),$$

such that $0 < \alpha < 1$ and $\lim_{N\to+\infty} u/N = (\alpha, 1-\alpha)$.

In the interior of the positive quarter plane the components of the Markov $(U_x^N(t))$ behave like independent random walks on $\mathbb{N}$. Therefore, the renormalised version of the process converges as follows

$$\lim_{N\to+\infty} \left(\frac{U_x^N(Nt)}{N},\, t < \tau\right) \overset{\text{dist.}}{=} \left((\alpha,\, 1-\alpha) + \theta_1 t,\, t < \tau\right),$$

with

$$\theta_1 = (\lambda_{2,1} + \lambda_{2,2} - R\mu_1 x_{1,1} - \mu_2 x_{2,1},\, \lambda_{1,2} + \lambda_{1,1} - \mu_1 x_{1,2} - C\mu_2 x_{2,2})$$

and

$$\tau = \alpha / \left(-\theta_{1,h}\right)^+ \wedge (1-\alpha) / \left(-\theta_{1,v}\right)^+ \leq +\infty.$$

Thus, if $x \in \Delta_{B,2}$ then $\theta_{1,h}$ is negative and $\theta_{1,v}$ is positive, so

$$\tau = \frac{\alpha}{R\mu_1 x_{1,1} + \mu_2 x_{2,1} - \lambda_{2,1} - \lambda_{2,2}} < +\infty.$$

Therefore, if $x \in \Delta_{B,2}$ then any fluid limit in the interior of the positive quadrant reaches the ordinate axis after some finite time τ, or, equivalently, the time that is necessary for the horizontal component of the rescaled process to reach zero converges almost surely to $\tau < +\infty$.

The fluid limit along the ordinate axis Due to the strong Markov property, the analysis is reduced to the case where the initial state of the process $(U_x^N(t))$ satisfies $\lim_{N\to+\infty} u/N = (0,1)$. At the fluid time scale (Nt), from the perspective of the large component (in this case, the vertical) the other component (the horizontal) is at equilibrium. Moreover, at u, this point "far away" from the abscissa axis, the discontinuities involving this axis do not interfere in the evolution of the process $(U_{x,v}^N(t))$.

Using the results in Propositions 3.5 and 3.6 (of the Appendix), we derive the equilibrium properties of the horizontal component of this process. Particularly, the drift of the vertical component of this process is dependent on the location of its horizontal component, such that

$$\theta_{1,v} = \lambda_{1,2} + \lambda_{1,1} - \mu_1 x_{1,2} - C\mu_2 x_{2,2} \quad \text{or} \quad \theta_{2,v} = \lambda_{1,2} - \mu_1 x_{1,2} - C\mu_2 x_{2,2},$$

depending on whether the horizontal component is positive or negative. Thus, the local equilibrium of the horizontal component determines the mean vertical drift of the (large) vertical component.

Indeed, far from the abscissa axis and along the ordinate axis, $(U_{x,h}^N(t))$ behaves like the random walk $(W(t))$ on $\mathbb{Z}$, described in Definition 3.4, with $P = 2$, $B_1 = 1$, $B_2 = R$, $\kappa_1 = \lambda_{2,1} + \lambda_{2,2}$, $\kappa_2 = \lambda_{1,1}$, $\eta_1 = \mu_2 x_{2,1}$, $\eta_2 = \mu_1 x_{1,1}$, $\rho_1 = \lambda_{2,1} + \lambda_{2,2}$ and $\rho_2 = 0$. Proposition 3.5 shows that if $x \in \Delta_A \cap \Delta_{B,2}$, then the associated process $(W(t))$ is ergodic. In this case, $(U_{x,v}^N(t))$ behaves like a random walk in the Markovian Environment (see Section 9.6 of Robert [Rob03, p. 271]). From this reference, we derive the following convergence

$$\lim_{N\to+\infty} \left(\frac{U_x^N(t)}{N}, t < \tau' \right) \overset{\text{dist.}}{=} \left(\left(0,\, 1 + \bar{\theta}_v t \right), t < \tau' \right)$$

where

$$\begin{aligned} \bar{\theta}_v &\overset{\text{def.}}{=} \theta_{1,v}\nu(\mathbb{N}) + \theta_{2,v}\nu(\mathbb{Z}_-^*) \\ &= \frac{\lambda_{2,1} + \lambda_{2,2} - R\mu_1 x_{1,1} - \mu_2 x_{2,1}}{R} - \lambda_{1,2} - \lambda_{1,1} + \mu_1 x_{1,2} + C\mu_2 x_{2,2}, \end{aligned}$$

with $\nu(\mathbb{N})$ and $\nu(\mathbb{Z}_-^*)$ given by Proposition 3.6, and $\tau' = 1/\left(-\bar{\theta}_v\right)^+$. Finally, if $x \in \Delta_{B,2}$ then $\bar{\theta}_v < 0$ and $\tau' < +\infty$.

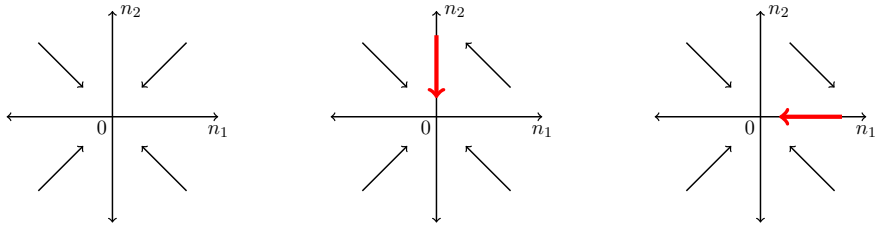
In the origin, if $x \in \Delta_A \cap \Delta_{B,2}$, then $\theta_{2,h}, \theta_{3,h}, \theta_{3,v}, \theta_{4,v} > 0$ and $\theta_{2,v}, \theta_{4,h} < 0$ such that the fluid limit does not scape from any direction. This concludes the case where $x \in \Delta_A \cap \Delta_{B,2}$ and $u \in \mathbb{N}^2$.

Other cases For the cases where $u \in \mathbb{N}^2$, the previous arguments can be easily adapted for the cases where $x \in \Delta_0 \setminus \Delta_{B,2}$. If $x \in \Delta_A \cap \Delta_{B,3}$ then fluid limits reach the abscissa axis instead, evolving along this axis until the origin of the plane. Or, if $x \in \Delta_A \cap \Delta_{B,1}$ then the fluid limits reach some of the axis and then goes along until reaching the origin of the plane.

With the same sort of arguments used in the case $u \in \mathbb{N}^2$, the direction of the constant drift θ_p for $2 \le p \le 4$ is enough to extend the results for any

$u \in \mathbb{Z}^2 \setminus \mathbb{N}^2$ and $x \in \Delta_0$. The dynamics of the fluid limits are the same as before, such that the fluid limits reach some axis and evolve along it until reaching the origin of the plane where it stays. It is worth mentioning that, if $u \in \mathbb{N} \times \mathbb{Z}_-^*$ and $x \in \Delta_A \cap \Delta_{B,2}$ then fluid limits may reach and cross the abscissa axis, going from $\mathbb{N} \times \mathbb{Z}_-^*$ to $\mathbb{N}^2$. However, the strong Markov property yields that this scenario is the same presented before for $u \in N^2$. The symmetric case may happen if $u \in \mathbb{Z}_-^* \times \mathbb{N}$ and $x \in \Delta_A \cap \Delta_{B,3}$, with same consequences.

Figure 3.3 synthesises directions of the fluid limits drifts, for $x \in \Delta_0$. Note that the mean drift in the axis of the positive quarter plane ($\bar{\theta}_v$ in the case above) are highlighted in red only if the relevant drifts in their neighbours do not have the same direction. The proposition is proved.



(a) if $x \in \Delta_A \cap \Delta_{B,1}$ (b) if $x \in \Delta_A \cap \Delta_{B,2}$ (c) if $x \in \Delta_A \cap \Delta_{B,3}$

Figure 3.3 – Resulting drifts of the fluids limits associated with $(U_x(t))$.

□

Finally, the probabilities of each component of this process being either negative or positive can be derived. This measure is related to the performance of the system, since it indicates the probability of a job to be hosted locally or to be offloaded. We use a similar approach as in Proposition 3.6.

Proposition 3.3 (Invariant measure)**.** *For fixed $x \in \Delta_0$, given by Definition 3.2, the stationary probability that either component of the process $(U_x(t))$ is negative is*

$$\pi_x(\mathbb{Z}_-^* \times \mathbb{Z}) = \frac{RC\mu_1 x_{1,1} + C\left(\mu_2 x_{2,2} + \mu_2 x_{2,1} - \lambda_{2,2} - \lambda_{2,1}\right) + \mu_1 x_{1,2} - \lambda_{1,2} - \lambda_{1,1}}{\lambda_{1,1}(RC - 1)}$$

and

$$\pi_x(\mathbb{Z} \times \mathbb{Z}_-^*) = \frac{RC\mu_1 x_{2,2} + R\left(\mu_1 x_{1,1} + \mu_1 x_{1,2} - \lambda_{1,1} - \lambda_{1,2}\right) + \mu_2 x_{2,1} - \lambda_{2,1} - \lambda_{2,2}}{\lambda_{2,2}(RC - 1)}.$$

Proof. Let $x \in \Delta_0$,X be fixed such that the process $(U_x(t))$ is ergodic with invariant probability measure $\pi_x = (\pi_x(n), n \in \mathbb{Z}^2)$ and X be a random variable with finite support on $\mathbb{Z}^2$ and distribution π_x. Also let $f_u : \mathbb{Z}^2 \to \mathbb{C}$ be a complex-valued function, such that $u \overset{\text{def.}}{=} (u_1, u_2)$ and $f_u(n) = u_1^{n_1} u_2^{n_2}$.

Writing the balance equation of the process $(U_x(t))$, we obtain the following relation

$$\sum_{\substack{n,n'\in\mathbb{Z}^2\\ n\neq n'}} \pi_x(n)q_x(n,n')(f_u(n')-f_u(n))=0. \tag{3.11}$$

Let $u\in\partial D^2(1)\overset{\text{def.}}{=}\{u\in\mathbb{C}^2:|u_1|=|u_2|=1\}$ and, for $\mathcal{Z}\subseteq\mathbb{Z}^2$,

$$\mathbb{E}\left(f_u(X)\mathbb{1}_{\{X\in\mathcal{Z}\}}\right)=\sum_{n\in\mathcal{Z}}\pi_x(n)f_u(n).$$

Rearranging the terms of the Equation (3.11), for $u\in\partial D^2(1)$, it yields

$$\begin{aligned}\mathbb{E}\left(f_u(X)\right)K_1(u)=&\\ &\mathbb{E}\left(f_u(X)\mathbb{1}_{\left\{X\in\mathbb{Z}_-^*\times\mathbb{Z}\right\}}\right)K_2(u)+\mathbb{E}\left(f_u(X)\mathbb{1}_{\left\{X\in\mathbb{Z}\times\mathbb{Z}_-^*\right\}}\right)K_3(u),\end{aligned}$$

with

$$\begin{aligned}K_1(u)&\overset{\text{def.}}{=}\mu_1x_{1,1}\left(1-u_1^{-R}\right)+\mu_2x_{2,1}\left(1-u_1^{-1}\right)+\left(\lambda_{2,1}+\lambda_{2,2}\right)\left(1-u_1\right)\\ &\quad+\mu_2x_{2,2}\left(1-u_2^{-C}\right)+\mu_1x_{1,2}\left(1-u_2^{-1}\right)+\left(\lambda_{1,2}+\lambda_{1,1}\right)\left(1-u_2\right)\\ K_2(u)&\overset{\text{def.}}{=}\lambda_{1,1}\left(u_1^R-u_2\right)\\ K_3(u)&\overset{\text{def.}}{=}\lambda_{2,2}\left(u_2^C-u_1\right).\end{aligned}$$

By definition, for $\mathcal{Z}\subseteq\mathbb{Z}^2$,

$$\lim_{u\to(1,1)}\mathbb{E}\left(f_u(X)\mathbb{1}_{\{X\in\mathcal{Z}\}}\right)=\pi_x(\mathcal{Z}),$$

and as the point $(1,1)$ is a simple root $K_1(u)$, $K_2(u)$ and $K_3(u)$, we have

$$\begin{aligned}\pi_x\left(\mathbb{Z}_-^*\times\mathbb{Z}\right)=\lim_{u_1\to1}\frac{K_1\left(u_1,\sqrt[C]{u_1}\right)}{K_2\left(u_1,\sqrt[C]{u_1}\right)}=&\\ \frac{RC\mu_1x_{1,1}+C\left(\mu_2x_{2,2}+\mu_2x_{2,1}-\lambda_{2,2}-\lambda_{2,1}\right)+\mu_1x_{1,2}-\lambda_{1,2}-\lambda_{1,1}}{\lambda_{1,1}(RC-1)}&.\end{aligned}$$

With a similar method one obtains $\pi_x\left(\mathbb{Z}\times\mathbb{Z}_-^*\right)$, taking the limit when $u_2\to1$ and $u_1=\sqrt[R]{u_2}$. The proposition is proved. □

The following corollary makes explicit that the fixed point obtained using the heuristic approach is indeed the equilibrium point of the stochastic system.

Corollary 3.1. *If $x=x^\star$, given by Equation (3.10) then*

$$\pi_{x^\star}(\mathbb{N}\times\mathbb{Z})=\gamma_1^\star(\delta)\quad\textit{and}\quad\pi_{x^\star}(\mathbb{Z}\times\mathbb{N})=\gamma_2^\star(\delta).$$

3.4 Time Evolution

In this section, we study the transient behaviour the system. Using a similar approach to Hunt and Kurtz [HK94], we are able to prove that the dynamical system associated with the Markovian process $(L^N(t))$ has only one stable point, which is absorbent and is the same as $x^\star$, given by Equation (3.10) while the probabilities that of each component of $(Y^N(t))$ is either negative or positive converges to $\gamma^\star$.

Stochastic Evolution Equations

For $\xi > 0$, let $\mathcal{N}_\xi$ denote a Poisson process on $\mathbb{R}_+$ with rate ξ and $(\mathcal{N}_\xi^\ell)_\ell$ an i.i.d. sequence of such processes. All Poisson processes are assumed to be independent. Classically, the process $(L^N(t))$ can be seen as the unique solution to the following stochastic differential equations (SDE),

$$\begin{cases} \mathrm{d}L^N_{p,p}(t) &= \mathcal{N}_{N\lambda_{p,p}}(\mathrm{d}t)\mathbb{1}_{\left\{Y_p^N(t^-)<0, m^N_{q,p}(t^-)\geq 1\right\}} - \sum_{\ell=1}^{+\infty} \mathcal{N}^\ell_{\mu_p}(\mathrm{d}t)\mathbb{1}_{\left\{\ell \leq L^N_{p,p}(t^-)\right\}} \\ &\quad + \mathcal{N}_{N\lambda_{p,q}}(\mathrm{d}t)\left(1 - \mathbb{1}_{\left\{m_q^N(t^-)\geq A_p\right\}}\right)\mathbb{1}_{\left\{m_p^N(t^-)\geq A_p\right\}} \\ \mathrm{d}L^N_{q,p}(t) &= \mathcal{N}_{N\lambda_{q,p}}(\mathrm{d}t)\mathbb{1}_{\left\{m_p^N(t^-)\geq A_q\right\}} \sum_{\ell=1}^{+\infty} \mathcal{N}^\ell_{\mu_q}(\mathrm{d}t)\mathbb{1}_{\left\{\ell \leq L^N_{q,p}(t^-)\right\}} \\ &\quad + \mathcal{N}_{N\lambda_{q,q}}(\mathrm{d}t)\left(1 - \mathbb{1}_{\left\{Y_q^N(t^-)<0, m^N_{p,q}(t^-)\geq 1\right\}}\right)\mathbb{1}_{\left\{m_p^N(t^-)\geq A_q\right\}} \end{cases} \tag{3.12}$$

for $1 \leq p \leq 2$ and $q = 3 - p$, with initial condition $L^N(0) \in \mathbb{N}^{2\times 2}$ such that $AL^N(0) \leq G^N$.

We start the analysis showing that the equivalent dynamical system does not reach the boundaries of the convex hull where it lives if its initial point observes some conditions. Therefore, in that case, the process $(m^N(t))$ does not influence the evolution of process $(L^N(t))$ since, with probability 1, this process lives inside the convex hull without ever touching its boundaries. In the following, we introduce the dynamical system $z(t)$ and the region Δ_F, where it takes values.

Definition 3.3. *If $x \in \Delta_F(\delta)$ then*

$$\begin{cases} x_{1,1}R + x_{2,1} = \mathcal{R}_1\delta_1, \\ x_{1,1} + x_{2,1}C < \mathcal{C}_1, \end{cases} \quad \textit{and} \quad \begin{cases} x_{1,2}R + x_{2,1} < \mathcal{R}_2, \\ x_{1,2} + x_{2,2}C = \mathcal{C}_2\delta_2. \end{cases}$$

Lemma 3.1 (Dynamical System Convergence)**.** *If*

$$(z(t)) \overset{\text{def.}}{=} \left((z_{i,k}(t), 1 \leq i,k \leq 2), t \geq 0\right)$$

is governed by the following dynamical system

$$\begin{cases} \dot{z}_{p,p}(t) = -\mu_1 z_{p,p}(t) + \lambda_{p,p}(1 - \psi_p(z(t))), \\ \dot{z}_{q,p}(t) = -\mu_q z_{q,p}(t) + \lambda_{q,p} + \lambda_{q,q}\psi_q(z(t)), \end{cases} \tag{3.13}$$

for $q = 3 - p$ and $1 \leq p \leq 2$ with $z(0) \in \Delta_F(\delta)$,

$$\psi_1(z(t)) = \frac{1}{\lambda_{1,1}(RC-1)}\left[RC\left(\lambda_{1,1} - \mu_1 z_{1,1}(t)\right) + \left(\lambda_{1,2} - \mu_1 z_{1,2}(t)\right) + C\left(\lambda_{2,2} - \mu_2 z_{2,2}(t) + \lambda_{2,1} - \mu_2 z_{2,1}(t)\right)\right]$$

and

$$\psi_2(z(t)) = \frac{1}{\lambda_{2,2}(RC-1)}\left[RC\left(\lambda_{2,2} - \mu_2 z_{2,2}(t)\right) + \left(\lambda_{2,1} - \mu_2 z_{2,1}(t)\right) + R\left(\lambda_{1,1} - \mu_1 z_{1,1}(t) + \lambda_{1,2} - \mu_1 z_{1,2}(t)\right)\right]$$

then the dynamical system converges exponentially fast to the stable fixed point fixed point $z^\star/(RC-1)$, where $z^\star = (z^\star_{i,k}, 1 \leq i,k \leq 2)$ and

$$\begin{cases} z^\star_{1,1} &= \mathcal{R}_1\delta_1 C + \mathcal{C}_2\delta_2 - \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1} - \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2}C, \\ z^\star_{2,1} &= -\mathcal{C}_2\delta_2 R - \mathcal{R}_1\delta_1 + \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}R + \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2}RC, \\ z^\star_{1,2} &= -\mathcal{R}_1\delta_1 C - \mathcal{C}_2\delta_2 + \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2}C + \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}RC, \\ z^\star_{2,2} &= \mathcal{C}_2\delta_2 R + \mathcal{R}_1\delta_1 - \dfrac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2} - \dfrac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1}R. \end{cases} \tag{3.14}$$

Also, for $t \geq 0$, the dynamical system $z(t)$ at t stays in region $\Delta_F(\delta)$.

Proof. The mapping $t \mapsto z(t)$ has the following integral form

$$z(t) = z(0) + \int_0^t \dot{z}(\mathrm{d}s)$$

which admits the following explicit solution

$$z(t) = \begin{cases} z_{1,1}(t) &= z^\star_{1,1}(RC-1)^{-1} + \alpha_1 e^{-\mu_1 t} + C\beta_1 e^{-\mu_2 t}, \\ z_{2,1}(t) &= z^\star_{2,1}(RC-1)^{-1} - R\alpha_1 e^{-\mu_1 t} - RC\beta_1 e^{-\mu_2 t}, \\ z_{1,2}(t) &= z^\star_{1,2}(RC-1)^{-1} - RC\alpha_2 e^{-\mu_1 t} - C\beta_2 e^{-\mu_2 t}, \\ z_{1,2}(t) &= z^\star_{2,2}(RC-1)^{-1} + R\alpha_2 e^{-\mu_1 t} + \beta_2 e^{-\mu_2 t}, \end{cases}$$

where

$$\alpha_1 = \frac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1} - z_{1,1}(0) - z_{1,2}(0), \quad \beta_1 = \frac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2} - z_{2,2}(0) - z_{2,1}(0),$$
$$\alpha_2 = \frac{\lambda_{1,1}+\lambda_{1,2}}{\mu_1} - z_{1,1}(0) - z_{1,2}(0), \quad \beta_2 = \frac{\lambda_{2,1}+\lambda_{2,2}}{\mu_2} - z_{2,2}(0) - z_{2,1}(0).$$

So, for $t \geq 0$, the following relations hold

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big[Rz_{1,1}(t) + z_{2,1}(t)\Big] = \frac{\mathrm{d}}{\mathrm{d}t}\Big[z_{1,2}(t) + Cz_{2,2}(t)\Big] = 0.$$

Finally, we can rewrite the dynamical system $t \mapsto z(t)$ by using the autonomous system $\omega(t) = (\omega_1(t), \omega_2(t))$, such that

$$\begin{cases} Rz_{1,1}(t) + z_{2,1}(t) = \mathcal{R}_1\delta_1, \\ z_{1,2}(t) + Cz_{2,2}(t) = \mathcal{C}_2\delta_2 \end{cases} \text{ and } \begin{cases} z_{1,1}(t) + Cz_{2,1}(t) = z^\star_{1,1} + Cz^\star_{2,1} + \omega_1(t) \\ Rz_{1,2}(t) + z_{2,2}(t) = Rz^\star_{1,2} + z^\star_{2,2} + \omega_2(t), \end{cases}$$

where

$$\begin{cases} \omega_1(0) = z_{1,1}(0) - z^\star_{1,1} + C(z_{2,1}(0) - z^\star_{2,1}) \\ \omega_2(0) = R(z_{1,2}(0) - z^\star_{1,2}) + z_{2,2}(0) - z^\star_{2,2}, \end{cases}$$

and

$$\dot{\omega}(t) = F\omega(t),$$

with

$$F = \frac{1}{RC-1}\begin{pmatrix} \mu_1 - RC\mu_2 & C(\mu_2 - \mu_1) \\ R(\mu_1 - \mu_2) & \mu_2 - RC\mu_1 \end{pmatrix}.$$

It is straightforward to see that the eigenvalues of F are negative since it is assumed that $\mu_p/\mu_{3-p} < RC$, for $1 \le p \le 2$. Therefore, one concludes that if $z(0) \in \Delta_F(\delta)$ then $z(t) \in \Delta_F(\delta)$ for $t > 0$. □

Proposition 3.4 (Limiting Dynamical System)**.** *Under Conditions* (3.5) *and* (3.6) *and for* $\bar{\delta} < \delta < (1,1)$*, where* $\bar{\delta}$ *is defined by Proposition 3.1, if the initial state* $L^N(0)$ *is such that*

$$\begin{cases} RL^N_{1,1}(0) + L^N_{2,1}(0) = \mathcal{R}_1\delta_1 + o(N), \\ L^N_{1,2}(0) + CL^N_{2,2}(0) = \mathcal{C}_2\delta_2 + o(N) \end{cases} \text{ and } \begin{cases} L^N_{1,1}(0) + CL^N_{2,1}(0) < \mathcal{C}_1 + o(N), \\ RL^N_{1,2}(0) + L^N_{2,2}(0) < \mathcal{R}_2 + o(N), \end{cases}$$

such that

$$\lim_{N\to+\infty} \frac{L^N(0)}{N} = l(0) \in \Delta_F(\delta)$$

then there exists a continuous process

$$(l(t)) = ((l_{j,k}(t),\, 1 \le j,k \le 2),\, t \ge 0),$$

such that the convergence below

$$\lim_{N\to+\infty} \left(\frac{L^N(t)}{N}, \int_0^t f\left(Y^N(u)\right) \mathrm{d}u \right) \overset{\text{dist.}}{=} \left(l(t), \int_0^t \int_{\mathbb{Z}^2} f(x)\, \pi_{l(u)}(\mathrm{d}x)\, \mathrm{d}u \right)$$

holds for any function f *with finite support on* $\mathbb{Z}^2$ *and* $l(t)$ *satisfies*

$$\begin{cases} \dot{l}_{p,p}(t) = -\mu_p l_{p,p}(t) + \lambda_{p,p}(1 - \pi_{l(t)}(\mathcal{A}_p)), \\ \dot{l}_{q,p}(t) = -\mu_q l_{q,p}(t) + \lambda_{q,p} + \lambda_{q,q}\pi_{l(t)}(\mathcal{A}_q), \end{cases}$$

with $\pi_{l(t)}$ *given in Proposition 3.3,* $\mathcal{A}_p = \{(y_1, y_2) \in \mathbb{Z}^2 : y_p \in \mathbb{N}\}$..

Proof. Just as it is common in classical stochastic calculus, $(L^N(t)/N)$ can be written as the renormalised integral solution of Equation (3.12). However, by Lemma 3.1 we show that an equivalent dynamical system $(z(t))$, starting in $\Delta_F(\delta)$ remains $\Delta_F(\delta)$ and, consequentially, does not reach the boundaries of the closed hull where it is defined, with probability 1. Therefore, translating to the equation above, it implies that the fluid limits live sufficiently far from the borders that enclose the set

$$\mathcal{D} \overset{\text{def.}}{=} \left\{ x \in \mathbb{R}_+^{2\times 2} : Ax < G \right\}$$

such that the free slots process $(m^N(t))$ does not influence the evolution of process $(L^N(t))$. Thus, the solution of the SDE (3.12) is simply written as, for $1 \leq p \leq 2$ and $q = 3 - p$,

$$\begin{cases} \dfrac{L^N_{p,p}(t)}{N} = \dfrac{L^N_{p,p}(0)}{N} + \overline{\mathcal{M}}^N_{p,p}(t) + \lambda_{p,p} \displaystyle\int_0^t \mathbb{1}_{\left\{Y^N_p(s)<0\right\}} \,\mathrm{d}s - \mu_p \int_0^t \frac{L^N_{p,p}(s)}{N} \,\mathrm{d}s, \\ \dfrac{L^N_{q,p}(t)}{N} = \dfrac{L^N_{q,p}(0)}{N} + \overline{\mathcal{M}}^N_{q,p}(t) + \lambda_{q,p} t - \mu_q \displaystyle\int_0^t \frac{L^N_{q,p}(s)}{N} \,\mathrm{d}s \\ \qquad\qquad\qquad\qquad\qquad\qquad + \lambda_{q,q} \displaystyle\int_0^t \mathbb{1}_{\left\{Y^N_q(s)\geq 0\right\}} \,\mathrm{d}s \end{cases}$$

where $\overline{\mathcal{M}}^N_{p,p}(t)$ and $\overline{\mathcal{M}}^N_{q,p}(t)$ are martingales, whose increasing predictable processes at time t are given by

$$\begin{cases} \left\langle \overline{\mathcal{M}}^N_{p,p} \right\rangle (t) = \dfrac{1}{N} \left((\lambda_{p,p} \displaystyle\int_0^t \mathbb{1}_{\left\{Y^N_p(s)<0\right\}} \,\mathrm{d}s + \mu_p \int_0^t \frac{L^N_{p,p}(s)}{N} \,\mathrm{d}s \right) \\ \left\langle \overline{\mathcal{M}}^N_{q,p} \right\rangle (t) = \dfrac{1}{N} \left(\lambda_{q,p} t + \lambda_{q,q} \displaystyle\int_0^t \mathbb{1}_{\left\{Y^N_q(s)\geq 0\right\}} \,\mathrm{d}s + \mu_q \int_0^t \frac{L^N_{q,p}(s)}{N} \,\mathrm{d}s \right). \end{cases}$$

Doob's inequality yields the martingale $\overline{\mathcal{M}}^N(t) = \left(\overline{\mathcal{M}}^N_{j,k}(t),\, 1 \leq j,k \leq 2\right)$ becomes negligible of order $1/\sqrt{N}$ as N goes to infinity. For $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{0\leq s\leq t} \sup_{1\leq j,k\leq 2} \left| \overline{\mathcal{M}}^N_{j,k}(s) \right| \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \mathbb{E}\left(\sup_{1\leq j,k\leq 2} \left(\overline{\mathcal{M}}^N_{j,k}(t) \right)^2 \right) \xrightarrow[N\to+\infty]{} 0.$$

Moreover, using the same method as Hunt and Kurtz [HK94], we obtain a result analogous to the Theorem 3 of this reference. As the process $(Y^N(t))$ is ergodic, the following convergence holds

$$\lim_{N\to+\infty} \left(\frac{L^N(t)}{N}, \int_0^t \mathbb{1}_{\left\{Y^N_1(s)<0\right\}} \,\mathrm{d}s, \int_0^t \mathbb{1}_{\left\{Y^N_2(s)<0\right\}} \,\mathrm{d}s \right) \overset{\text{dist.}}{=} \left(l(t), \int_0^t \pi_{l(s)} \left(\mathbb{Z}^*_- \times \mathbb{Z}\right) \mathrm{d}s, \int_0^t \pi_{l(s)} \left(\mathbb{Z} \times \mathbb{Z}^*_-\right) \mathrm{d}s \right).$$

From Proposition 3.3, we have that

$$\lim_{N\to+\infty} \left(\pi_{l(t)}(\mathbb{N}\times\mathbb{Z}),\, \pi_{l(t)}(\mathbb{Z}\times\mathbb{N}) \right) = \left(\psi_1(l(t)),\, \psi_2(l(t)) \right).$$

which implies that $\dot{l}(t) = \dot{z}(t)$. Hence, if $l(0) = z(0)$ then $l(t) = z(t)$ for $t > 0$. Therefore, the convergence holds, such that $\lim_{t\to+\infty} l(t) = z^\star = x^\star$. The proposition is proved. □

Corollary 3.2. *The equilibrium point of the Dynamical System $z(t)$ described in Equation (3.13) is given by $z^\star$ the solution of Equation (3.14). Moreover,*

$$(z^\star, \psi(z^\star)) = (x^\star, \gamma^\star)\,.$$

3.5 Conclusion

We have proposed in this paper an analytical model to study a simple offloading strategy for data centres in the framework of multi-resource Cloud Computing, under asymmetric service loads such each data centre is running low on one of the resources. The strategy considered consists of forwarding jobs with the highest demand for the resource locally depleted, if this resource is being used above a fixed threshold. The key finding is that the proposed strategy can significantly improve the performance of both data centres. A next step consists of studying the implementation of a dynamic mechanism of the parametrisation of these thresholds.

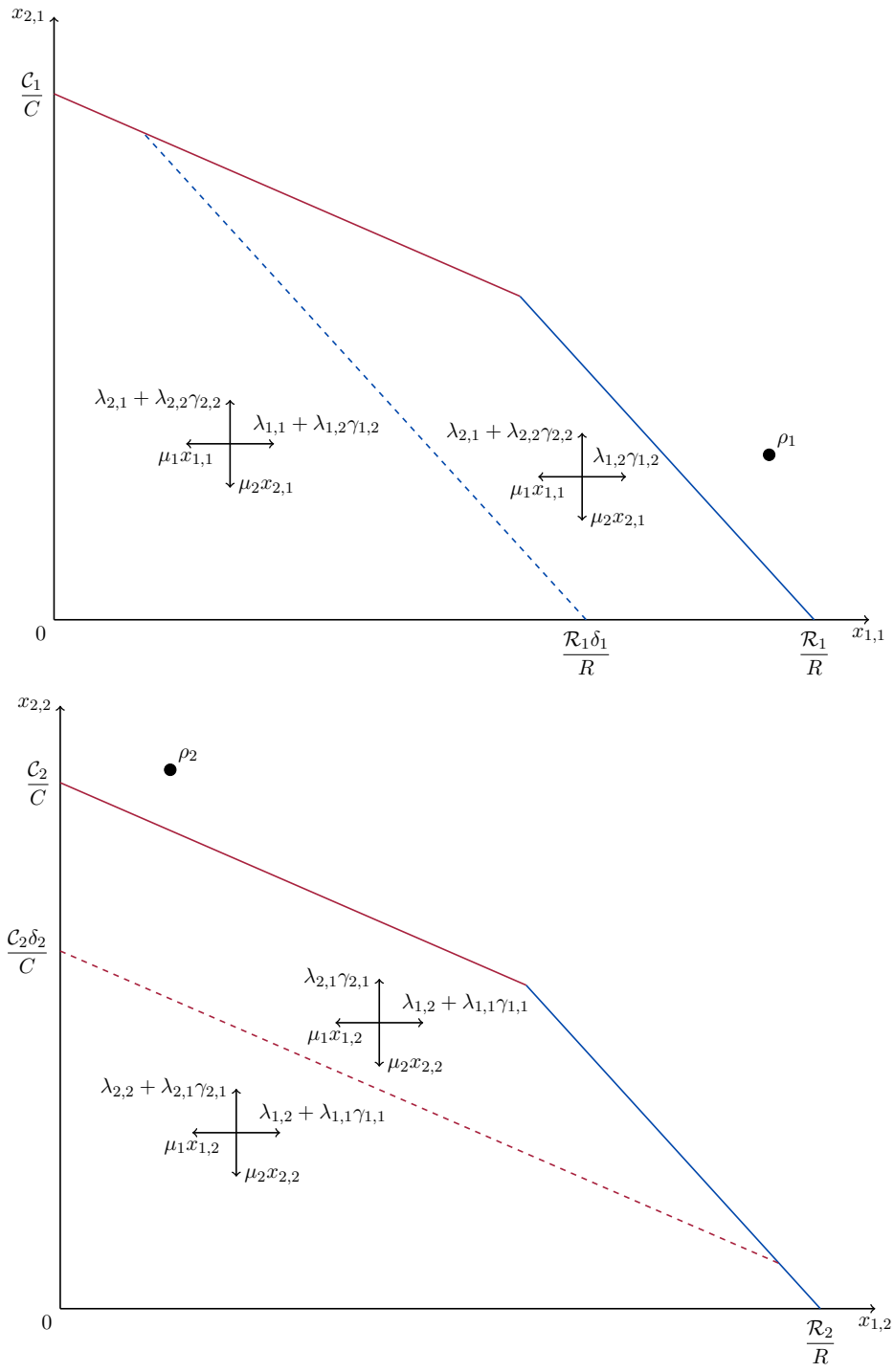


Figure 3.4 – Transitions of threshold forwarding policy

Algorithm 1 Jobs placement at data centre k

Require: $m_k = (m_{1,k}, m_{2,k})$ ▷ Free resources vector
Require: Y_k ▷ Offloading trigger
Require: $j \in \{1, 2\}, A_j = (A_{1,j}, A_{2,j})$ ▷ Type of request
Require: $p \in \{k, k'\}$ ▷ Origin of the request
Ensure: $q \in \{k, k', 0\}$ ▷ Placement, offloading or rejection

$q \leftarrow k$ ▷ Initial Placement
if $p = k$ **then** ▷ Local Job treatment
  if $j = j^\star(k)$ **then** ▷ $j^\star(k) = k$: jamming type of Job
    if $Y_k > 0$ **then** ▷ Is occupation threshold reached?
      $q \leftarrow k'$
    else
      for $i \leftarrow 1, 2$ **do** ▷ **?** $m_k \geq A_j$
        if $m_{i,k} < A_{i,j}$ **then**
          $q \leftarrow k'$ ▷ Offloading
          break
        end if
      end for
    end if
  else
    for $i \leftarrow 1, 2$ **do**
      if $m_{i,k} < A_{i,j}$ **then**
        $q \leftarrow k'$ ▷ Offloading
        break
      end if
    end for
  end if
else ▷ Alien Job treatment
  for $i \leftarrow 1, 2$ **do**
    if $m_{i,k} < A_{i,j}$ **then**
      $q \leftarrow 0$ ▷ Rejection
      break
    end if
  end for
end if
return q ▷ Returns: Placement, offloading or rejection

Appendix: limiting Markov processes

A random walk on $\mathbb{Z}$

We define a random walk on $\mathbb{Z}$ where the transition rates depend whether the state is in either half lines $\mathbb{Z}_-^*$ or $\mathbb{N}$. The goal of this analysis is to determine the stationary probability of this random walk to be in either half-line. For that the characteristic function of the invariant distribution of such process is decomposed in two (analytical) parts using some complex analysis arguments (similar to the Wiener-Hopf factorisation). This method is often used in the analysis of reflected random walks on $\mathbb{N}$, such as $GI/GI/1$ queue, see Chapter VIII of Asmussen [Asm03, p. 220] and Section 2.2 of Robert [Rob03, p. 33] for instance. See also Fricker *et al.* [FGRT17] for a similar (and extended) application of such techniques.

The random walk is properly defined as follows.

Definition 3.4. *For fixed $P \in \mathbb{N}^*$, let $(W(t)) \overset{\text{def.}}{=} (W(t),\, t \geq 0)$ be the Markov process on $\mathbb{Z}$ whose non-negative elements of its $\mathcal{Q}$-matrix,*

$$Q_W = (q(n, n'),\, n, n' \in \mathbb{Z})\,,$$

are given by

$$q(n,n') = \begin{cases} \kappa_p & \text{if} \quad n' = n + B_p \quad \text{and} \quad n < 0, \\ \rho_p & \text{if} \quad n' = n + B_p \quad \text{and} \quad n \geq 0, \\ \eta_p & \text{if} \quad n' = n - B_p, \\ 0 & \text{otherwise}, \end{cases}$$

with $1 = B_1 < \cdots < B_P \in \mathbb{N}$ and $\kappa_1 + \rho_1 + \eta_1 > 0$.

In the following proposition, we present the stability properties of the Markov process $(W(t))$.

Proposition 3.5. *The Markov process $(W(t))$ is ergodic if*

$$\sum_{p=1}^{P} \rho_p B_p < \sum_{p=1}^{P} \eta_p B_p < \sum_{p=1}^{P} \kappa_p B_p. \tag{3.15}$$

When it exists, $\nu = (\nu(n),\, n \in \mathbb{Z})$ denotes the corresponding invariant distribution.

Proof. The Markov process $(W(t))$ on $\mathbb{Z}$ behaves like a random walk on each of the half-lines $\mathbb{N}$ and $\mathbb{Z}_-^*$. At some point $n \in \mathbb{Z}$, the drift of the random walk is positive if $n \in \mathbb{Z}_-^*$ and negative if $n \in \mathbb{N}$. This property implies the ergodicity of the Markov process by using the Lyapounov function $F(x) = |x|$, for example. See Corollary 8.7 of Robert [Rob03, p. 214]. □

In the following proposition, we derive the probability that, at equilibrium, the Markov process $(W(t))$ is either in the negative or positive half-line.

Proposition 3.6. *If Equation (3.15) holds, then $(W(t))$ is ergodic and at equilibrium the probability that the process is either in the negative or positive half-line is given by*

$$\nu(\mathbb{Z}_-^*) = \frac{\sum_{p=1}^P B_p\left(\eta_p - \rho_p\right)}{\sum_{p=1}^P B_p\left(\kappa_p - \rho_p\right)} \quad \textit{and} \quad \nu(\mathbb{N}) = \frac{\sum_{p=1}^P B_p\left(\kappa_p - \eta_p\right)}{\sum_{p=1}^P B_p\left(\kappa_p - \rho_p\right)}. \tag{3.16}$$

Proof. Let X a random variable with finite support on $\mathbb{Z}$ and distribution ν and, for $u \in \mathbb{C} \setminus \mathbb{R}_-$ and $n \in \mathbb{Z}$, $f_u : \mathbb{Z} \mapsto \mathbb{C}$ such that $f_u(n) = u^n$.

For any function f with finite support on $\mathbb{Z}$, the equilibrium equations for the Markov process $(W(t))$ yield the following relation

$$\nu(Q_W(f)) = \sum_{\substack{n,n' \in \mathbb{Z} \\ n \neq n'}} \nu(n) q(n,n')(f(n') - f(n)) = 0, \tag{3.17}$$

where Q_W is the $\mathcal{Q}$-matrix of $(W(t))$. Replacing f by f_u in Equation (3.17), for $u \in \partial D(1) = \{u \in \mathbb{C} : |u| = 1\}$, one gets the relation

$$\mathbb{E}\left(f_u(X)\mathbb{1}_{\{X \geq 0\}}\right) P_1(u) = -\mathbb{E}\left(f_u(X)\mathbb{1}_{\{X<0\}}\right) P_2(u),$$

where

$$\begin{cases} P_1(u) \overset{\text{def.}}{=} u^{B_P} \displaystyle\sum_{p=1}^P \left[\left(u^{B_p} - 1\right) \rho_p + \eta_p \left(u^{-B_p} - 1\right)\right], \\ P_2(u) \overset{\text{def.}}{=} u^{B_P} \displaystyle\sum_{p=1}^P \left[\left(u^{B_p} - 1\right) \kappa_p + \eta_p \left(u^{-B_p} - 1\right)\right]. \end{cases}$$

By definition,

$$\lim_{u \to 1} \mathbb{E}\left(f_u(X)\mathbb{1}_{\{X \geq 0\}}\right) = \nu(\mathbb{N}) \quad \text{and} \quad \lim_{u \to 1} \mathbb{E}\left(f_u(X)\mathbb{1}_{\{X<0\}}\right) = \nu(\mathbb{Z}_-^*).$$

Using Equation (3.15), the 1 is a simple root of $P_1(u)$ and $P_2(u)$ then one on obtains the following relation

$$\frac{\nu(\mathbb{Z}_-^*)}{\nu(\mathbb{N})} = \lim_{u \to 1} -\frac{P_2'(u)}{P_1'(u)} = -\frac{\sum_{p=1}^P B_p\left(\rho_p - \eta_p\right)}{\sum_{p=1}^P B_p\left(\kappa_p - \eta_p\right)}.$$

Hence, using the fact that $\nu(\mathbb{Z}_-^*) + \nu(\mathbb{N}) = 1$, one gets Equation (3.16). □

Chapter 4

Analysis of an offloading scheme for data centres in the framework of Fog Computing

4.1 Introduction

Cloud computing has become one of the major stakes in the development of information technology by offering the possibility of reserving computing resources online. Commercial offers already exist for customers (residential or business) relying on big data centres like Amazon or Azure [Mic] for example. This kind of technology is also relevant for network operators in the framework of network function virtualisation, where network functions can be instantiated on data centres instead of dedicated hardware. In this context, there is currently a clear trend to distribute data centres. For network operators, it is possible to instantiate at the edge of the network functions which were so far centralised in servers (e.g., mobile core functions). Furthermore, by allocating resources closer to end users, it is expected to offer better quality of experience. Distributing cloud computing resources at the edge of the network is known as fog computing (see [WRSvdM11, BMZA12, RBG12]).

Data centres involved in fog computing have a smaller capacity than those in the case of cloud computing and therefore more subject to congestion. Hence, to reduce the probability of request blocking, fog computing data centres have to collaborate. For instance, when one request cannot be accommodated by one of them, it may be forwarded to another one.

A typical example of such a situation is when data centres are located on a logical ring at the edge of the network. See Figure 4.1. A request arriving in an overloaded data centre with index i, may be forwarded to a neighbouring data centre $i-1$ or $i+1$ with some probability. Hence, if the traffic to a data centre is causing saturation, the other data centres may help alleviate this saturation regime. The aim of this chapter is of investigating the impact of such a cooperative scheme. In practice, the network could be backed up by a central (bigger) data centre at the core of the network but at a price in terms of latency. We will not consider this additional feature here.

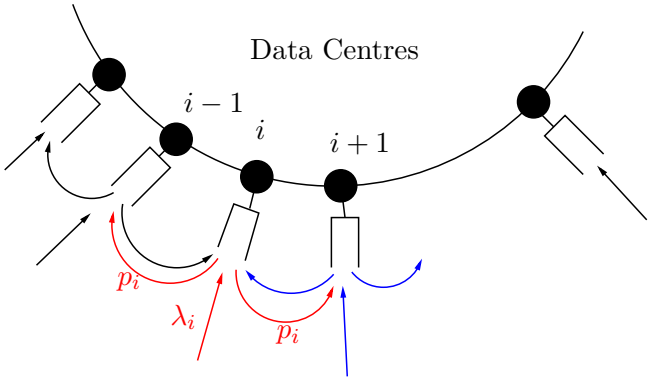


Figure 4.1 – A Fog Computing Architecture

Collaboration of Two Data Centres

Our aim here is to qualitatively estimate the gain achieved by the collaboration of data centres at the edge of the network. The main part of our analysis will concern the impact of the collaboration of two data centres. It is shown in Section 4.2.4 that the analysis applies also to more general architectures of fog computing, as in Figure 4.1, provided they are not congested.

For $i \in \{1, 2\}$, the external arrival process of requests to facility/data centre $\#i$, referred to as class i requests, is Poisson with parameter λ_i. If one of the C_i servers is idle upon arrival, then the request is processed by this data centre. Otherwise, if the data centre is saturated, i.e., all the C_i servers are busy, then with probability p_i the request is forwarded to the other data centre if it is not saturated too, otherwise with probability $1 - p_i$ the request is rejected. A request allocated at data centre $\#i$ is processed at rate μ_i.

Considering the number of requests processed at both data centres, this scheme can be clearly represented by a two dimensional Markov process on $\{0, \ldots, C_1\} \times \{0, \ldots, C_2\}$. This Markov process, related to loss networks, is not reversible in general and its invariant distribution does not have a product form expression. Even if a numerical analysis of the equilibrium equations is always possible, it is very likely that it will not give precise qualitative and quantitative results concerning the impact of rerouting parameters p_1 and p_2 of the offloading scheme. Our goal is of giving *explicit* closed form expressions of the equilibrium probability that a request is rejected, see Theorem 4.3 which is our main result in this domain.

To overcome the difficulty of not having an explicit expression of the equilibrium, we have chosen to study a scaled version of this network. The input rates λ_1, λ_2 and the capacities C_1, C_2 are assumed to be proportional to a large parameter N which goes to infinity. This scaling has been introduced by Kelly in the context of loss networks (see [Kel91]). As it will be seen, there is a relation between the parameters (see Condition (E) below), which implies that both data centres can be saturated with positive probability. We will focus mainly on this case which is, in our view, the most interesting situation to assess the benefit of offloading mechanisms in a congested environment. Otherwise, the situation is much simpler. One of the data centres will be

underloaded, so that the rejection rate at equilibrium will converge to 0 as N gets large, in particular external arrivals to this data centre and the rerouted jobs from the other data centre will be accepted with probability 1 in the limit. See Proposition 4.1 and Theorem 4.1.

In this limiting regime we prove convergence results for the process describing the number of free servers at both data centres in the same way as in Hunt and Kurtz [HK94] for loss networks. We show that the invariant distribution of a random walk in the quarter plane is playing a key role in the asymptotic behaviour of the loss probabilities at equilibrium. The derivation of the equilibrium is based on the analysis of random walks on $\mathbb{N}^2$ by [FI79]. Taking advantage of the specific characteristics of the random walks considered, we are able to get an explicit expression of the generating function of their invariant distributions in terms of elliptic integrals instead of contour integrals in the complex plane as in [FI79]. See Theorem 4.3. With these results we can then assess quantitatively the interest of this load balancing mechanism by comparing the respective loss probabilities of the two streams of requests.

The organisation of this chapter is as follows: in Section 4.2 the stochastic model is introduced and the limit results for the scaling regime are obtained. A family of random walks is shown to play a central role. Section 4.3 establishes the functional relation satisfied by the generating function of the invariant measure of one of these random walks. Section 4.4 gives an explicit representation of this generating function in Theorem 4.3 and therefore of the performance metrics of the load balancing mechanism. Section 4.5 presents some numerical examples of these results. Concluding remarks are presented in Section 4.6.

4.2 Model description

4.2.1 Model

We consider in this chapter two processing facilities in parallel. The first one is equipped with C_1 servers and serve requests (for computing resources) arriving according to a Poisson process with rate λ_1; each request requires an exponentially distributed service time with mean $1/\mu_1$ (a request if accepted occupies a single server). Similarly, the second processing facility is equipped with C_2 servers and serves service requests arriving according to a Poisson process with rate λ_2; service times are exponentially distributed with mean $1/\mu_2$.

To reduce the blocking probability, we assume that requests arriving at a service facility with no available servers are forwarded to the other one with a given probability. More precisely, if a request arrives at service facility #1 with no available servers, the request is forwarded to the other service facility with probability p_1. Similarly, a request arriving at facility #2 with no available servers is forwarded to the other facility with probability p_2. See Figure 4.2.

Let $L_1(t)$ and $L_2(t)$ denote the number of occupied servers in facilities #1 and #2 at time t, respectively. Owing to the Poisson and exponential service time assumptions, $(L(t)) = ((L_1(t), L_2(t)))$ is a Markov process with values in

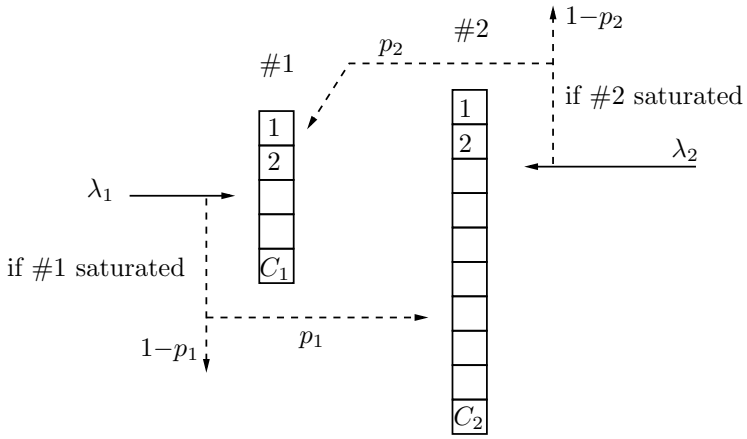


Figure 4.2 – Load Balancing between Two Data Centres

the set $\{0,\dots,C_1\}\times\{0,\dots,C_2\}$, and transitions from (ℓ_1,ℓ_2) to (ℓ_1+i,ℓ_2+j) occurring at rate

$$\begin{cases} (\lambda_1 + p_2\lambda_2 \mathbb{1}_{\{\ell_2=C_2\}})\cdot \mathbb{1}_{\{\ell_1<C_1\}} & \text{if } (i,j)=(1,0),\\ (p_1\lambda_1 \mathbb{1}_{\{\ell_1=C_1\}} + \lambda_2)\cdot \mathbb{1}_{\{\ell_2<C_2\}} & \text{if } (i,j)=(0,1),\\ \mu_1\ell_1 & \text{if } (i,j)=(-1,0),\\ \mu_2\ell_2 & \text{if } (i,j)=(0,-1) \end{cases}$$

and 0 otherwise.

The equilibrium characteristics of this Markov process on a finite state space, like loss probabilities, do not seem to have closed form expressions in general. A scaling approach is used in the following to get some insight on the performance of such a strategy. We first introduce a random walk in $\mathbb{N}^2$.

4.2.2 A random walk in the extended positive quadrant

We now consider the following random walk in the extended positive quadrant.

Definition 4.1. *For fixed $l=(l_1,l_2)\in\mathbb{R}_+^2$, one defines the random walk $(\overline{m}_l(t))$ on $(\mathbb{N}\cup\{+\infty\})^2$ as follows: the transition from (m_1,m_2) to (m_1+a,m_2+b) occurs at rate*

$$\begin{cases} \mu_1 l_1 & \text{if } (a,b)=(1,0),\\ \mu_2 l_2 & \text{if } (a,b)=(0,1),\\ \lambda_1 + p_2\lambda_2\mathbb{1}_{\{m_2=0\}} & \text{if } (a,b)=(-1,0) \text{ and } m_1>0,\\ \lambda_2 + p_1\lambda_1\mathbb{1}_{\{m_1=0\}} & \text{if } (a,b)=(0,-1) \text{ and } m_2>0, \end{cases}$$

for $(m_1,m_2)\in(\mathbb{N}\cup\{+\infty\})^2$ with the convention that $+\infty\pm x=+\infty$ for $x\in\mathbb{N}$ (see Figure 4.3).

In particular $(+\infty,+\infty)$ is an absorbing point for the process $(\overline{m}_l(t))$. The random walk $(\overline{m}_l(t))$ is a special case of the Markov process investigated in [FI79].

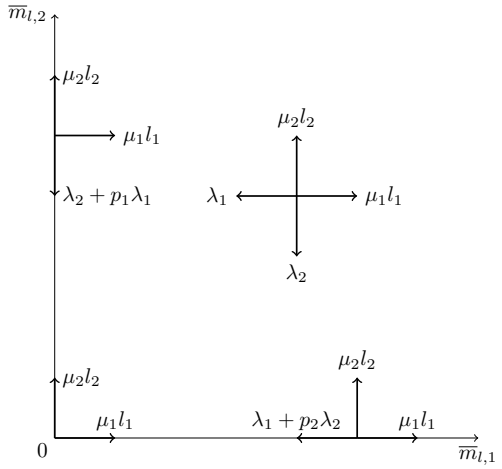


Figure 4.3 – Transitions for $\bar{m}_l(t)$.

The following result summarises the stability properties of this random walk. Critical cases are omitted.

Proposition 4.1. *For $l = (l_1, l_2) \in \mathbb{R}_+^2$,*

(i) If one of the conditions

$$a)\ \lambda_2 < \mu_2 l_2 \text{ and } \lambda_1 p_1 + \lambda_2 > \mu_1 l_1 p_1 + \mu_2 l_2$$
$$b)\ \lambda_1 < \mu_1 l_1 \text{ and } \lambda_1 + \lambda_2 p_2 > \mu_1 l_1 + \mu_2 l_2 p_2$$
$$c)\ \lambda_1 > \mu_1 l_1 \text{ and } \lambda_2 > \mu_2 l_2$$

holds then the Markov process $(\overline{m}_l(t))$ is ergodic on $\mathbb{N}^2$. In this case the unique invariant distribution on $\mathbb{N}^2$ is denoted by π_l.

(ii) If $\lambda_1 < \mu_1 l_1$ and $\lambda_2 < \mu_2 l_2$, the unique invariant distribution of $(\overline{m}_l(t))$ on the extended state space $(\mathbb{N} \cup \{+\infty\})^2$ is the Dirac measure $\delta_{(\infty,\infty)}$.

(iii) If $\lambda_1 > \mu_1 l_1$, $\lambda_2 < \mu_2 l_2$ and $\lambda_1 p_1 + \lambda_2 < p_1\mu_1 l_1 + \mu_2 l_2$, the unique invariant distribution of $(\overline{m}_l(t))$ on $(\mathbb{N} \cup \{+\infty\})^2$ is $G_{\delta_1} \otimes \delta_\infty$, where G_{δ_1} is the geometric distribution with parameter $\delta_1 = \mu_1 l_1/\lambda_1$.

(iv) If $\lambda_2 > \mu_2 l_2$, $\lambda_1 < \mu_1 l_1$ and $\lambda_2 p_2 + \lambda_1 < p_2\mu_2 l_2 + \mu_1 l_1$, the unique invariant distribution of $(\overline{m}_l(t))$ on $(\mathbb{N} \cup \{+\infty\})^2$ is $\delta_\infty \otimes G_{\delta_2}$, where G_{δ_2} is the geometric distribution with parameter $\delta_2 = \mu_2 l_2/\lambda_2$.

Proof. Due to Fayolle and Iasnogorodski [FI79] see also Proposition 9.15 of Robert [Rob03, p. 276], $(\overline{m}_l(t))$ is ergodic if and only if one of the conditions of (i) holds. As long as it does not hit 0, the first (resp. second) coordinate of $(\overline{m}_l(t))$ behaves as an $M/M/1$ queue with arrival rate $\mu_1 l_1$ (resp. $\mu_2 l_2$) and service rate λ_1 (resp. λ_2). Under the conditions of (ii), each of these $M/M/1$ queues is transient, in particular starting from 1, it has a positive probability of not returning to 0. This implies that after some random time, the process $(\overline{m}_l(t))$ stays in the interior of the quadrant $\mathbb{N}^2$ and therefore behaves as a couple of independent transient $M/M/1$ queues. Consequently, both coordinates of $(m_l(t))$ are converging in distribution to $+\infty$. Similarly, for

(*iii*) and (*iv*), the process $(\overline{m}_l(t))$ can be coupled to two queues, the first one, an $M/M/1$ queue which is transient and the second one, an ergodic $M/M/1$ queue, with an invariant distribution which is geometrically distributed. □

4.2.3 Heavy Traffic Scaling Regime

We investigate now the case when some of the parameters of the processing facilities are scaled up by a factor $N \in \mathbb{N}$. The arrival rates are given by $\lambda_1 N$ and $\lambda_2 N$ with $\lambda_1 > 0$ and $\lambda_2 > 0$. Similarly the capacities are given by $C_1^N = Nc_1$ and $C_2^N = Nc_2$ for some positive constants c_1 and c_2. To indicate the dependence of the numbers of idle servers upon N, an upper index N is added to the stochastic processes. A similar approach has been used in Alanyali and Hajek [AH97] to study a load balancing scheme in an Erlang system.

We will consider the process

$$(m^N(t)) \stackrel{\text{def.}}{=} \left(C_1^N - L_1^N(t), C_2^N - L_2^N(t)\right)$$

describing the number of idle servers in both processing facilities. As it will be seen, the random walks $(\overline{m}_l(t))$, $l \in \mathbb{R}_+^2$, play an important role in the asymptotic behaviour of $(m^N(t))$ as N goes to infinity.

Theorem 4.1. *If one of the following conditions*

$$\begin{cases}\lambda_2 < \mu_2 c_2,\\ \lambda_1 p_1 + \lambda_2 > \mu_1 c_1 p_1 + \mu_2 c_2,\end{cases} \begin{cases}\lambda_1 < \mu_1 c_1,\\ \lambda_1 + \lambda_2 p_2 > \mu_1 c_1 + \mu_2 c_2 p_2,\end{cases} \text{ or } \begin{cases}\lambda_1 > \mu_1 c_1,\\ \lambda_2 > \mu_2 c_2\end{cases}$$

holds, and if the initial conditions are such that $m^N(0) = m \in \mathbb{N}^2$ and

$$\lim_{N\to+\infty} \left(\frac{L_1^N(0)}{N}, \frac{L_2^N(0)}{N}\right) = c = (c_1, c_2)$$

then, for the convergence in distribution,

$$\lim_{N\to+\infty} \left(\frac{L_1^N(t)}{N}, \frac{L_2^N(t)}{N}, \int_0^t f(m^N(u))\ \mathrm{d}u\right) \stackrel{\text{dist.}}{=} \left(c_1, c_2, t\int_{\mathbb{N}^2} f(x)\pi_c(\mathrm{d}x)\right)$$

for any function f with finite support on $\mathbb{N}^2$, π_c is the invariant distribution of the process $(\overline{m}_c(t))$ defined previously.

Proof. Using the same method as in [HK94], one gets an analogous result to Theorem 3 of this reference. For the convergence in distribution of processes, the relation

$$\lim_{N\to+\infty} \left(\frac{L_1^N(t)}{N}, \frac{L_2^N(t)}{N}, \int_0^t f(m^N(u))\ \mathrm{d}u\right) \stackrel{\text{dist.}}{=} \left(l_1(t), l_2(t), \int_0^t\int_{\mathbb{N}^2} f(x)\pi_{l(u)}(\mathrm{d}x)\,\mathrm{d}u\right) \quad (4.2)$$

holds, where $(l(t)) = ((l_1(t), l_2(t)))$ satisfying the following integral equations

$$\begin{cases} l_1(t) = c_1 + \displaystyle\int_0^t \Big(\lambda_1 \pi_{l(u)}(\mathcal{A}_1) + p_2\lambda_2 \pi_{l(u)}(\mathcal{A}_1 \cap \mathcal{A}_2^c) - \mu_1 l_1(u)\Big)\, \mathrm{d}u \\ l_2(t) = c_2 + \displaystyle\int_0^t \Big(\lambda_2 \pi_{l(u)}(\mathcal{A}_2) + p_1\lambda_1 \pi_{l(u)}(\mathcal{A}_2 \cap \mathcal{A}_1^c) - \mu_2 l_2(u)\Big)\, \mathrm{d}u, \end{cases}$$

for $i \in \{1, 2\}$, $\mathcal{A}_i = \{m \in \mathbb{N}^2 : m_i \neq 0\}$ and, for $l \in \mathbb{R}_+^2$, π_l is the *unique* invariant distribution of $(\overline{m}_l(t))$.

Let us assume without loss of generality that, under condition (E), for example the first condition of (E) is satisfied. It will be assumed throughout the chapter. It is not difficult to construct a coupling so $L_1^N(t) \geq Q_1^N(t)$ holds almost surely for all $t \geq 0$, where $(Q_1^N(t))$ is the number of jobs of an $M/M/C_1^N/C_1^N$ queue with arrival rate $\lambda_1 N$ and service rate μ_1. Since $\lambda_1 > \mu_1 c_1$, a classical result, see Section 6.7 of [Rob03, p. 164] for example, gives the convergence in distribution

$$\lim_{N\to+\infty} \left(\frac{L_1^N(t)}{N}\right) = (c_1),$$

in particular, $(l_1(t))$ is a constant equal to c_1.

If, for $i \in \{1, 2\}$, $\mathcal{N}_{\lambda_i N}$ is the Poisson process of arrivals at facility $\#i$, using the same coupling as before one gets that the number $U_2^N(t)$, arrivals at facility $\#2$ up to time t, satisfies, for $0 \leq s \leq t$, $U_2^N(t) - U_2^N(s) \geq \underline{U}^N(t) - \underline{U}^N(s)$,

$$\underline{U}^N(t) \overset{\text{def.}}{=} \mathcal{N}_{\lambda_2 N}[0, t] + \int_0^t \mathbb{1}_{\{Q_1^N(u-)=C_1^N, B_1(u-)=1\}} \mathcal{N}_{\lambda_1 N}(\mathrm{d}u),$$

where $(B_1(u), u \geq 0)$ is a family of independent Bernoulli random variables with parameter p_1. The lower bound includes the direct arrivals $\mathcal{N}_{\lambda_2 N}[0, t]$ to facility $\#2$ and the rejected jobs from $\#1$. One gets that

$$L_2^N(t) \geq Q_2^N(t),$$

where $(Q_2^N(t))$ is a $G/M/C_2^N/C_2^N$ queue with the arrival process $(\underline{U}^N(t))$.

The ergodic theorem gives that, almost surely

$$\lim_{N\to+\infty} \frac{\underline{U}^N(t)}{N} = t\left(\lambda_2 + \lambda_1 p_1 \left(1 - \frac{\mu_1 c_1}{\lambda_1}\right)\right) > \mu_2 c_2 t,$$

by condition (E). Using this relation, one can show that, for the convergence in distribution, the relation

$$\lim_{N\to+\infty} \left(\frac{Q_2^N(t)}{N}\right) \overset{\text{dist.}}{=} (c_2)$$

holds. In particular $(l_2(t))$ is constant and equal to c_2. Therefore, almost surely, $(l(t)) = (c)$ holds, hence $\pi_{l(t)} = \pi_c$. Relation (4.2) shows that the theorem is proven. □

The following proposition states that the performances of the load balancing mechanism can be expressed with the invariant distribution π_c.

Proposition 4.2. *Under Condition (E), as N goes to infinity, the probability that at equilibrium a job of class $i \in \{1,2\}$ is rejected converges to*

$$\beta_i = \pi_c\left(m \in \mathbb{N}^2 : m_i = 0\right)(1-p_i) + p_i \pi_c\,(0,0)\,,$$

where π_c is the invariant distribution of $(\overline{m}_c(t))$.

Proof. Assume that $(L_1^N(t), L_2^N(t))$ is at equilibrium, the number of class 1 jobs rejected between 0 and t is given by

$$R_1^N(t) \stackrel{\text{def.}}{=} \int_0^t \mathbb{1}_{\{m_1^N(u-)=0, B_1(u-)=0\}} \mathcal{N}_{\lambda_1 N}(\mathrm{d}u) + \int_0^t \mathbb{1}_{\{m_1^N(u-)=0, m_2^N(u-)=0, B_1(u-)=1\}} \mathcal{N}_{\lambda_1 N}(\mathrm{d}u).$$

The probability of rejecting a class 1 job is hence given by

$$\mathbb{P}\left(m_1^N(0)=0, B_1(0)=0\right) + \mathbb{P}\left(m_1^N(0)=0, B_1(0)=1, m_2^N(0)=0\right) = \frac{\mathbb{E}(R_1^N(1))}{\lambda_1 N}.$$

Using the martingales associated with Poisson processes, one gets

$$\frac{\mathbb{E}(R_1^N(1))}{\lambda_1 N} = (1-p_1)\mathbb{E}\left(\int_0^1 \mathbb{1}_{\{m_1^N(u-)=0\}}\,\mathrm{d}u\right) + p_1 \mathbb{E}\left(\int_0^1 \mathbb{1}_{\{m_1^N(u-)=0, m_2^N(u-)=0\}}\,\mathrm{d}u\right),$$

one concludes with the convergence of the previous theorem. □

When condition (E) is not satisfied, one can obtain an analogous result. Its (elementary) proof is skipped. The results on the asymptotic blocking probability of jobs are summarised in the following proposition, where (A), (B_1) and (B_2) are exclusive.

Proposition 4.3. *Let*

$$(A) \stackrel{\text{def.}}{=} \begin{cases} \lambda_1 < \mu_1 c_1, \\ \lambda_2 < \mu_2 c_2, \end{cases} \quad (B_1) \stackrel{\text{def.}}{=} \begin{cases} \lambda_2 > \mu_2 c_2, \\ \lambda_1 + \lambda_2 p_2 < \mu_1 c_1 + \mu_2 c_2 p_2 \end{cases} \quad \textit{and}$$

$$(B_2) \stackrel{\text{def.}}{=} \begin{cases} \lambda_1 > \mu_1 c_1, \\ \lambda_2 + \lambda_1 p_1 < c_2\mu_2 + \mu_1 c_1 p_1, \end{cases}$$

then, at equilibrium, the loss probability of a job of class $i \in \{1,2\}$ is converging to β_i as N goes to infinity, where

$$\beta_i \stackrel{\text{def.}}{=} \begin{cases} \pi_c\left(m \in \mathbb{N}^2 : m_i = 0\right)(1-p_i) + p_i \pi_c(0,0) & \textit{if (E) holds,} \\ 0 & \textit{if } (A) \textit{ or } (B_i) \textit{ holds,} \\ (1-p_i)\,(1-\mu_i c_i/\lambda_i) & \textit{if } (B_{3-i}) \textit{ holds.} \end{cases}$$

4.2.4 An Extension to Multiple Data Centres

In this section, it is assumed that there are J data centres, for $1 \leq j \leq J$, the $j-$th data centre has $c_j N$ servers and the external arrivals to it are Poisson with parameter $\lambda_j N$ and services are exponentially distributed with parameter μ_j. If an external request at data centre j finds all $c_j N$ servers occupied, it is re-routed to data centre $j+1$ (with $J+1=1$) or to data centre $j-1$ (with $0=J$) with probability p_j, otherwise it is rejected. In particular a job is rerouted with probability $2p_j$ in the case of congestion. See Figure 4.1.

For $1 \leq j \leq J$, one defines the random walk $(\overline{m}_c^j(t))$ on $\mathbb{N}^2$ as follows: the transition from $(m,n) \in \mathbb{N}^2$ to $(m+a, n+b)$ occurs at rate

$$\begin{cases} \mu_j c_j & \text{if } (a,b) = (1,0), \\ \mu_{j+1} c_{j+1} & \text{if } (a,b) = (0,1), \\ \lambda_j + p_{j+1}\lambda_{j+1}\mathbb{1}_{\{n=0\}} & \text{if } (a,b) = (-1,0) \text{ and } m > 0, \\ \lambda_{j+1} + p_j \lambda_j \mathbb{1}_{\{m=0\}} & \text{if } (a,b) = (0,-1) \text{ and } n > 0. \end{cases}$$

If one of the conditions

$$\begin{cases} \lambda_j > \mu_j c_j, \\ \lambda_{j+1} > \mu_{j+1} c_{j+1}, \end{cases} \begin{cases} \lambda_{j+1} < \mu_{j+1} c_{j+1} \\ \lambda_j + \lambda_{j+1} p_{j+1} > \mu_j c_j + p_{j+1}\mu_{j+1} c_{j+1}, \end{cases}$$
$$\text{or } \begin{cases} \lambda_j < \mu_j c_j, \\ \lambda_{j+1} + \lambda_j p_j > \mu_{j+1} c_{j+1} + p_j \mu_j c_j, \end{cases}$$

holds, one gets that the associated Markov process is ergodic by Proposition 4.1, one denotes by π_c^j its invariant probability distribution. As before, one takes the following convention for the indices, $J+1=1$ and $1-1=J$.

We now give a version of the previous proposition in this context.

Proposition 4.4. *At equilibrium, the loss probability of a job of class $j \in \{1,\dots,J\}$ is converging to β_j as N goes to infinity in the following cases,*

1. *No Congestion.*
   If $\lambda_j < \mu_j c_j$ for all $j \in \{1,\dots,J\}$ then $\beta_j \equiv 0$.
2. *One saturated node.*
   If, for some $j_0 \in \{1,\dots,J\}$, $\lambda_{j_0} > \mu_{j_0} c_{j_0}$ and if $\lambda_j < \mu_j c_j$ holds for all $j \neq j_0$ and

   $$\begin{cases} \lambda_{j_0+1} + \lambda_{j_0} p_{j_0} < \mu_{j_0+1} c_{j_0+1} + p_{j_0}\mu_{j_0} c_{j_0} \\ \lambda_{j_0-1} + \lambda_{j_0} p_{j_0} < \mu_{j_0-1} c_{j_0-1} + p_{j_0}\mu_{j_0} c_{j_0}, \end{cases}$$

   then $\beta_j = 0$ if $j \neq j_0$ and $\beta_{j_0} = (1-2p_{j_0})(1-\mu_{j_0} c_{j_0}/\lambda_{j_0})$
3. *Two saturated neighbouring nodes.*
   If, for some $j_0 \in \{1,\dots,J\}$, one of the conditions (E_{j_0}) holds and $\lambda_j < \mu_j c_j$ holds for all $j \neq j_0, j_0+1$ and

   $$\begin{cases} \lambda_{j_0+2} + \lambda_{j_0+1} p_{j_0+1} \pi_c^{j_0}(\mathbb{N}\times\{0\}) < \mu_{j_0+2} c_{j_0+2} \\ \lambda_{j_0-1} + \lambda_{j_0} p_{j_0} \pi_c^{j_0}(\{0\}\times\mathbb{N}) < \mu_{j_0-1} c_{j_0-1} \end{cases} \tag{4.4}$$

then

$$\begin{cases} \beta_{j_0} = \pi_c^{j_0}\left(\{0\}\times\mathbb{N}\right)(1-2p_{j_0}) + p_{j_0}\pi_c^{j_0}(0,0) \\ \beta_{j_0+1} = \pi_c^{j_0}\left(\mathbb{N}\times\{0\}\right)(1-2p_{j_0+1}) + p_{j_0+1}\pi_c^{j_0}(0,0). \end{cases}$$

The proof is similar to the proof of the previous proposition and is therefore omitted. Note that Condition (3) implies that the nodes with index $j_0 - 1$ and $j_0 + 2$ are underloaded, so that only nodes with index j_0 and $j_0 + 1$ are congested. This result covers partially the set of various possibilities but, as long as only two neighbouring nodes are congested, it can be extended quite easily to the case where only pairs of nodes are congested.

When there are at least three neighbouring congested nodes, this method does not apply. It occurs when one of the conditions (E_{j_0}) holds and one of the conditions of (3) is not satisfied. One has to consider the invariant distributions of a three dimensional random walk in $\mathbb{N}^3$ for which there are scarce results. Nevertheless this situation should be, in practice, unlikely if the fog computing architecture is conveniently designed so that a local congestion can be solved using the neighbouring resources.

This proposition shows that the evaluation of the performances of the offloading algorithm can be expressed in terms of the invariant distributions of the random walks $(\overline{m}_c(t))$ introduced in Definition 4.1. The rest of the chapter is devoted to the analysis of these invariant distributions when they exist. In particular, we will derive an explicit expression of the blocking probabilities β_i at facility $\#i$.

4.3 Characteristics of the limiting random walk

4.3.1 Fundamental equations

Throughout this section, we assume that the first condition of (E) holds. Let $m_{c,1}$ and $m_{c,2}$ denote the abscissa and the ordinate of the random walk $(\overline{m}_c(t))$ in the stationary regime. Under stability condition (E), it is shown in [FI79] that the generating function of the stationary numbers $m_{c,1}$ and $m_{c,2}$, defined by

$$P(x,y) \stackrel{\text{def.}}{=} \mathbb{E}(x^{m_{c,1}} y^{m_{c,2}})$$

for complex x and y such that $|x| \leq 1$ and $absy \leq 1$, satisfies the functional equation

$$h_1(x,y)P(x,y) = h_2(x,y)P(x,0) + h_3(x,y)P(0,y) + h_4(x,y)\pi_c(0,0),$$

with $\pi_c(0,0)$ standing for $P(0,0)$ and

$$\begin{aligned} h_1(x,y) &\stackrel{\text{def.}}{=} -\mu_1 c_1 x^2 y - \mu_2 c_2 x y^2 + (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)xy - \lambda_1 y - \lambda_2 x, \\ h_2(x,y) &\stackrel{\text{def.}}{=} \lambda_2\left((1-p_2)xy - x + p_2 y\right), \\ h_3(x,y) &\stackrel{\text{def.}}{=} \lambda_1\left((1-p_1)xy - y + p_1 x\right), \\ h_4(x,y) &\stackrel{\text{def.}}{=} (\lambda_1 p_1 + \lambda_2 p_2)xy - p_2\lambda_2 y - p_1\lambda_1 x. \end{aligned}$$

It is worth noting that

$$\lambda_1 p_1(\lambda_1 + \lambda_2 p_2)h_2(x,y) + \lambda_2 p_2(\lambda_2 + \lambda_1 p_1)h_3(x,y) = \lambda_1\lambda_2(1 - p_1p_2)h_4(x,y). \quad (4.5)$$

In [FI79, FIM17], it is shown how to compute the unknown functions using the zeros of the kernel $h_1(x,y)$ and the results on Riemann-Hilbert problems. In the following we briefly describe how to achieve this goal. For the system under consideration, let us recall that the performance of the system is characterised by the blocking probabilities of the two classes of customers. For customers arriving at facility #1, the blocking probability is given by

$$\beta_1 \overset{\text{def.}}{=} P(0,1)(1-p_1) + p_1\pi_c(0,0) \quad (4.6)$$

and that for customers arriving at the second facility by

$$\beta_2 \overset{\text{def.}}{=} P(1,0)(1-p_2) + p_2\pi_c(0,0). \quad (4.7)$$

Using the normalising condition $P(1,1) = 1$, we can easily show that

$$\lambda_1 + \lambda_2 p_2 P(1,0) - \mu_1 c_1 = \lambda_1 P(0,1) + \lambda_2 p_2 \pi_c(0,0)$$

and

$$\lambda_2 + \lambda_1 p_1 P(0,1) - \mu_2 c_2 = \lambda_2 P(1,0) + \lambda_1 p_1 \pi_c(0,0).$$

We then deduce that

$$P(0,1) = \frac{\lambda_1 - \mu_1 c_1 + p_2(\lambda_2 - \mu_2 c_2) - p_2(\lambda_2 + \lambda_1 p_1)\pi_c(0,0)}{(1-p_1p_2)\lambda_1} \quad (4.8)$$

and

$$P(1,0) = \frac{\lambda_2 - \mu_2 c_2 + p_1(\lambda_1 - \mu_1 c_1) - p_1(\lambda_1 + \lambda_2 p_2)\pi_c(0,0)}{(1-p_1p_2)\lambda_2}. \quad (4.9)$$

The above relations show that the blocking probabilities β_1 and β_2 can be estimated as soon as the quantity $\pi_c(0,0)$ is known.

4.3.2 Zero pairs of the kernel

The kernel $h_1(x,y)$ has already been studied in Fayolle and Iasnogorodski [FI79] in the framework of coupled servers. For fixed y, the kernel $h_1(x,y)$ has two roots $X_0(y)$ and $X_1(y)$. Using the common definition of the square root such that $\sqrt{a} > 0$ for $a > 0$, the solution which is null at the origin and denoted by $X_0(y)$, is defined and analytic in $\mathbb{C} \setminus ([y_1,y_2] \cup [y_3,y_4])$ where the real numbers y_1, y_2, y_3 and y_4 are such that $0 < y_1 < y_2 < 1 < y_3 < y_4$. The other solution $X_1(y)$ is meromorphic in $\mathbb{C} \setminus ([y_1,y_2] \cup [y_3,y_4])$ with a pole at 0. The function $X_0(y)$ is precisely defined by

$$X_0(y) \overset{\text{def.}}{=} \frac{-(\mu_2 c_2 y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)y + \lambda_2) + \sigma_1(y)}{2\mu_1 c_1 y}$$

with

$$\Delta_1(y) \overset{\text{def.}}{=} (\mu_2 c_2 y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)y + \lambda_2)^2 - 4\mu_1 c_1 \lambda_1 y^2,$$

where $\sigma_1(y)$ is the analytic continuation in $\mathbb{C}\setminus([y_1, y_2]\cup[y_3, y_4])$ of the function $\sqrt{\Delta_1(y)}$ defined in the neighbourhood of 0. The other solution is defined as $X_1(y) \overset{\text{def.}}{=} \lambda_1/(\mu_1 c_1 X_0(y))$.

When y crosses the segment $[y_1, y_2]$, $X_0(y)$ and $X_1(y)$ describe the circle C_{r_1} with centre 0 and radius $r_1 \overset{\text{def.}}{=} \sqrt{\lambda_1/(\mu_1 c_1)} > 1$, since from the first of condition of (E), we have $\lambda_1 > \mu_1 c_1$.

Similarly, for fixed x, the kernel $h_1(x, y)$ has two roots $Y_0(x)$ and $Y_1(x)$. The root $Y_0(x)$, which is null at the origin, is analytic in $\mathbb{C}\setminus([x_1, x_2]\cup[x_3, x_4])$, where the real numbers x_1, x_2, x_3 and x_4 read $0 < x_1 < x_2 < 1 < x_3 < x_4$, and is given by

$$Y_0(x) \overset{\text{def.}}{=} \frac{-(\mu_1 c_1 x^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)x + \lambda_1) + \sigma_2(x)}{2\mu_2 c_2 x}$$

with

$$\Delta_2(x) \overset{\text{def.}}{=} (\mu_1 c_1 x^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)x + \lambda_1)^2 - 4\mu_2 c_2 \lambda_2 x^2,$$

where $\sigma_2(x)$ is the analytic continuation in $\mathbb{C}\setminus([x_1, x_2]\cup[x_3, x_4])$ of the function $\sqrt{\Delta_2(x)}$ defined in the neighbourhood of 0. The other root is defined as $Y_1(x) \overset{\text{def.}}{=} \lambda_2/(\mu_2 c_2 Y_0(x))$ and is meromorphic in $\mathbb{C} \setminus ([x_1, x_2] \cup [x_3, x_4])$ with a pole at the origin.

When x crosses the segment $[x_1, x_2]$, $Y_0(y)$ and $Y_1(y)$ describe the circle C_{r_2} with centre 0 and radius $r_2 \overset{\text{def.}}{=} \sqrt{\lambda_2/(\mu_2 c_2)}$.

4.4 Boundary value problems

4.4.1 Problem formulation

In Fayolle *et al.* [FIM17], it is proven that the functions $P(x, 0)$ and $P(0, y)$ can be extended as meromorphic functions in $\mathbb{C} \setminus [x_3, x_4]$ and $\mathbb{C} \setminus [y_3, y_4]$, respectively. Using the fact that $X_0(y)$ and $X_1(y)$ are on circle C_{r_1} for $y \in [y_1, y_2]$, we easily deduce that the function $P(x, 0)$, analytic in D_{r_1} (the disk with centre 0 and radius r_1), is such that for $x \in C_{r_1}$

$$\Re\left(i\frac{h_2(x, Y_0(x))}{h_3(x, Y_0(x))}P(x, 0)\right) = \Im\left(\frac{h_4(x, Y_0(x))}{h_3(x, Y_0(x))}\pi_c(0, 0)\right) \tag{4.10}$$

where $Y_0(x) \in [y_1, y_2]$.

Similarly, the function $P(0, y)$ is analytic in D_{r_2}, which is the disk with centre 0 and radius r_2, and for $x \in C_{r_2}$, we have

$$\Re\left(i\frac{h_3(X_0(y), y)}{h_2(X_0(y), y)}P(0, y)\right) = \Im\left(\frac{h_4(X_0(y), y)}{h_2(X_0(y), y)}\pi_c(0, 0)\right). \tag{4.11}$$

Using Equation (4.5), we have

$$\Im\left(\frac{h_4(x,y)}{h_2(x,y)}\right) = -\frac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\Re\left(i\frac{h_3(x,y)}{h_2(x,y)}\right).$$

Equation (4.11) can then be rewritten as

$$\Re\left(i\frac{h_3(X_0(y),y)}{h_2(X_0(y),y)}\left(P(0,y)+\frac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\pi_c(0,0)\right)\right)=0. \tag{4.12}$$

Similarly, Equation (4.10) can be rewritten as

$$\Re\left(i\frac{h_2(x,Y_0(x))}{h_3(x,Y_0(x))}\left(P(x,0)+\frac{p_1(\lambda_1+\lambda_2 p_2)}{\lambda_2(1-p_1p_2)}\pi_c(0,0)\right)\right)=0. \tag{4.13}$$

Problem (4.13) corresponds to Problem (7.6) in [FI79] for which $i_1 = i_2 = i_3 = 0$ in the notation of that chapter. The ratio $h_2(x, Y_0(x))/h_3(x, Y_0(x))$ corresponds to the function $J(x)$ in [FI79].

In the following, we focus on Riemann-Hilbert problem (4.12). The analysis of problem (4.13) is completely symmetrical. Moreover, to compute the blocking probabilities β_1 and β_2, we only need to compute the quantity $\pi_c(0,0)$.

4.4.2 Problem resolution

The function $P(0,y)$ is analytic in the open disk D_{r_2}. Using the reflection principle [Car50], the function

$$y \mapsto \overline{P\left(0, r_2^2/\overline{y}\right)}$$

is analytic on the outside of the closed disk $\overline{D_{r_2}}$. It is then easily checked that if we define

$$F_Y(y) \stackrel{\text{def.}}{=} \begin{cases} P(0,y)+\dfrac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\pi_c(0,0), & y\in D_{r_2},\\ \overline{P\left(0,r_2^2/\overline{y}\right)}+\dfrac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\pi_c(0,0), & y\in\mathbb{C}\setminus\overline{D_{r_2}}, \end{cases}$$

the function $F_Y(y)$ is sectionally analytic with respect to the circle C_{r_2}, the quantity $F_Y(y)$ tends to $\pi_c(0,0)(\lambda_1+\lambda_2 p_2)/(\lambda_1(1-p_1p_2))$ when y goes to infinity, and for $y \in C_{r_2}$

$$F_Y^i(y) = \alpha_Y(y) F_Y^e(y), \tag{4.14}$$

where $F_Y^i(y)$ (resp. $F_Y^e(y)$) is the interior (resp. exterior) limit of the function $F_Y(y)$ at the circle C_{r_2}, and the function $\alpha_Y(y)$ is defined on C_{r_2} by

$$\alpha_Y(y) \stackrel{\text{def.}}{=} \frac{\overline{a_Y(y)}}{a_Y(y)} \tag{4.15}$$

with

$$a_Y(y) \stackrel{\text{def.}}{=} \frac{h_3(X_0(y),y)}{h_2(X_0(y),y)}.$$

The solutions to Riemann-Hilbert problems of form (4.14) are given in [DL90]. We first have to determine the index of the problem defined as

$$\kappa_Y \overset{\text{def.}}{=} \frac{1}{2\pi}\text{var}_{y\in Cr_2} \arg \alpha_Y(y).$$

Theorem 7.2 of Fayolle and Iasnogorodski [FI79] it is shown that the stability condition (E) is equivalent to $\kappa_Y = 0$.

To obtain explicit expressions, let us first study the function $\alpha_Y(y)$, which can be expressed as follows.

Lemma 4.1. *The function $\alpha_Y(y)$ defined for $y \in C_{r_2}$ by Equation (4.15) can be extended as a meromorphic function in $\mathbb{C} \setminus ([y_1, y_2] \cup [y_3, y_4])$ by setting*

$$\alpha_Y(y) \overset{\text{def.}}{=} \frac{\lambda_2(1-p_1p_2)X_0(y) + yR_Y(X_0(y))}{y(\mu_2c_2(1-p_1p_2)yX_0(y) + R_Y(X_0(y)))}, \tag{4.16}$$

where

$$\begin{aligned} R_Y(x) \overset{\text{def.}}{=}\ & p_1\mu_1c_1(1-p_2)x^2 \\ & + (p_1p_2(\mu_1c_1+\mu_2c_2) - (1-p_2)(\lambda_2+\lambda_1p_1))x - p_2(\lambda_2+\lambda_1p_1). \end{aligned} \tag{4.17}$$

Proof. We have for (x, y) such that $y \in C_{r_2}$ and $x = X_0(y)$

$$\begin{aligned} h_3(x,\bar{y})h_2(x,y) = \lambda_1\lambda_2 \big(&((1-p_1)x-1)((1-p_2)x+p_2)y\bar{y} \\ &-((1-p_1)x-1)x\bar{y} + p_1x((1-p_2)x+p_2)y - p_1x^2\big) \end{aligned}$$

Using the fact that $y\bar{y} = \lambda_2/(\mu_2c_2)$ and $h_1(x, y) = 0$, we deduce that

$$h_3(x,\bar{y})h_2(x,y) = -\frac{\lambda_1\lambda_2(x-1)}{\mu_2c_2y}\left(\lambda_2(1-p_1p_2)x + yR_Y(x)\right),$$

where $R_Y(x)$ is defined by Equation (4.17), and the result follows. □

Since the index of the Riemann-Hilbert (4.14) is null, the solution is as follows.

Lemma 4.2. *The solution to the Riemann-Hilbert problem (4.14) exists and is unique and given for $y \in D_{r_2}$ by*

$$F_Y(y) \overset{\text{def.}}{=} \frac{\lambda_1+\lambda_2p_2}{\lambda_1(1-p_1p_2)}\pi_c(0,0)\varphi_Y(y), \tag{4.18}$$

where

$$\varphi_Y(y) \overset{\text{def.}}{=} \exp\left(\frac{y}{\pi}\int_{x_1}^{x_2}\frac{(\mu_1c_1x^2-\lambda_1)\Theta_Y(x)}{xh_1(x,y)}\,\mathrm{d}x\right) \tag{4.19}$$

and

$$\begin{aligned} &\Theta_Y(x) \overset{\text{def.}}{=} \\ &\text{ArcTan}\left(\frac{(1-p_1p_2)\sqrt{-\Delta_2(x)}}{(1-p_1p_2)(\mu_1c_1x^2-(\lambda_1+\lambda_2+\mu_1c_1+\mu_2c_2)x+\lambda_1)-2R_Y(x)}\right). \end{aligned} \tag{4.20}$$

Proof. Since the index of the Riemann-Hilbert (4.14) is null, the solution reads

$$F_Y(y) = \phi_Y(y) \exp\left(\frac{1}{2i\pi}\int_{C_{r_2}} \frac{\log \alpha_Y(z)}{z-y}\,\mathrm{d}z\right)$$

where the function $\alpha_Y(y)$ is defined by Equation (4.16) and $\phi_Y(y)$ is a polynomial (see [DL90]). Since we know that

$$F_Y(y) \to \pi_c(0,0)\frac{\lambda_1+\lambda_2 p_2}{\lambda_1(1-p_1p_2)}$$

as $|y| \to \infty$, then

$$\phi_Y(y) = \frac{\lambda_1+\lambda_2 p_2}{\lambda_1(1-p_1p_2)}\pi_c(0,0).$$

Let for $y \in C_{r_2}$ and $y = Y_0(x+i0)$ for $x \in [x_1, x_2]$

$$\Theta_Y(x) = \arg\left(-\mu_2 c_2(1-p_1p_2)Y_0(x+0i)x - R_Y(x)\right)$$

Using the expression of $Y_0(x)$, Equation (4.20) follows. It is clear that

$$\log \alpha_Y(Y_0(x+0i)) = -2i\Theta_Y(x).$$

Since $Y_0(x+0i) = \overline{Y_0(x-0i)}$, we have

$$\begin{aligned}\frac{1}{2i\pi}\int_{C_{r_2}} \frac{\log \alpha_Y(z)}{z-y}\,\mathrm{d}z &= \frac{1}{\pi}\int_{x_1}^{x_2} \Im\left(\frac{\log \alpha_Y(Y_0(x+0i))}{Y_0(x+0i)-y}\frac{\mathrm{d}Y_0}{\mathrm{d}x}(x+0i)\right)\mathrm{d}x \\ &= \frac{1}{\pi}\int_{x_1}^{x_2} \Im\left(\frac{-2i}{Y_0(x+0i)-y}\frac{\mathrm{d}Y_0}{\mathrm{d}x}(x+0i)\right)\Theta_Y(x)\,\mathrm{d}x.\end{aligned}$$

It is easily checked from the equation $h_1(x, Y_0(x)) = 0$ that

$$\frac{\mathrm{d}Y_0}{\mathrm{d}x} = \frac{-2\mu_1c_1xY_0(x) - \mu_2c_2Y_0(x)^2 + (\lambda_1+\lambda_2+\mu_1c_1+\mu_2c_2)Y_0(x) - \lambda_2}{\mu_1c_1x^2 + 2\mu_2c_2xY_0(x) - (\lambda_1+\lambda_2+\mu_1c_1+\mu_2c_2)x + \lambda_1}.$$

For $x \in [x_1, x_2]$, we have

$$\mu_1c_1x^2 + 2\mu_2c_2xY_0(x+0i) - (\lambda_1+\lambda_2+\mu_1c_1+\mu_2c_2)x + \lambda_1 = -i\sqrt{-\Delta_2(x)}$$

Using once again $h_1(x, Y_0(x+0i)) = 0$, we obtain for $x \in [x_1, x_2]$

$$\frac{\mathrm{d}Y_0}{\mathrm{d}x}(x+0i) = \frac{(\lambda_1-\mu_1c_1x^2)Y_0(x+0i)}{-ix\sqrt{-\Delta_2(x)}}$$

and then for real y

$$\Im\left(\frac{-2i}{Y_0(x+0i)-y}\frac{\mathrm{d}Y_0}{\mathrm{d}x}(x+0i)\right) = \frac{(\mu_1c_1x^2-\lambda_1)y}{xh_1(x,y)}.$$

It follows that for real y

$$\frac{1}{2i\pi}\int_{C_{r_2}} \frac{\log \alpha_Y(z)}{z-y}\,\mathrm{d}z = \frac{y}{\pi}\int_{x_1}^{x_2} \frac{(\mu_1c_1x^2-\lambda_1)\Theta_Y(x)}{xh_1(x,y)}\,\mathrm{d}x$$

It is easily checked that the function on the right hand side of the above equation can analytically be continued in the disk D_{r_2} and the result follows. Hence for $y \in D_{r_2}$, the first part of Equation (4.18) follows. When y is not in the closed disk $\overline{D_{r_2}}$, similar arguments can be used to derive the second part of Equation (4.18). □

In view of the above lemma, we can state the main result of this section.

Theorem 4.2. *The function $P(0,y)$ can be defined as a meromorphic function in $\mathbb{C} \setminus [y_3, y_4]$ by setting*

$$P(0,y) \stackrel{\text{def.}}{=} \begin{cases} \dfrac{\lambda_1 + \lambda_2 p_2}{\lambda_1(1-p_1p_2)}\pi_c(0,0)\varphi_Y(y) - \dfrac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\pi_c(0,0), \ y \in D_{r_2}, \\ \dfrac{\lambda_1 + \lambda_2 p_2}{\lambda_1(1-p_1p_2)}\pi_c(0,0)\alpha_Y(y)\varphi_Y(y) - \dfrac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\pi_c(0,0), \ y \in \mathbb{C} \setminus \overline{D_{r_2}}, \end{cases} \tag{4.21}$$

where $\varphi_Y(y)$ is defined by Equation (4.19).

Proof. Since the solution to the Riemann-Hilbert problem (4.12) is unique, the function $P(0,y)$ coincides with the function

$$F_Y(y) + \frac{p_2(\lambda_2+\lambda_1 p_1)}{\lambda_1(1-p_1p_2)}\pi_c(0,0)$$

in D_{r_2}. We can extend this function as follows. Noting that the function $\log \alpha_Y(y)$ is analytic in a neighbourhood of the circle C_{r_2}, the function

$$y \mapsto \exp\left(\frac{1}{2i\pi}\int_{C_{r_2}} \frac{\log \alpha_Y(z)}{z-y}\,\mathrm{d}z\right)$$

defined for $y \in D_{r_2}$ can be continued as a meromorphic function in $\mathbb{C} \setminus [x_3, x_4]$ considering the function defined for $y \in \mathbb{C} \setminus \overline{D_{r_2}}$, by

$$\alpha_Y(y)\exp\left(\frac{1}{2i\pi}\int_{C_{r_2}} \frac{\log \alpha_Y(z)}{z-y}\,\mathrm{d}z\right) = \alpha_Y(y)\exp\left(\frac{y}{\pi}\int_{x_1}^{x_2} \frac{(\mu_1 c_1 x^2 - \lambda_1)\Theta_Y(x)}{x h_1(x,y)}\,\mathrm{d}x\right),$$

where the last equality is obtained using the same arguments as above (consider first real y and then extend the function by analytic continuation). □

For the system under consideration, let us recall that the performance of the system is characterised by the blocking probabilities of the two classes of customers. The following theorem summarises the main results of the chapter for Condition (E). Proposition 4.3 covers the other cases.

Theorem 4.3. *Under Condition (E), as N goes to infinity, the probability that at equilibrium a job of facility #i, $i \in \{1,2\}$ is rejected converges to β_i with*

$$\beta_1 = (1-p_1)\frac{\lambda_1 - \mu_1 c_1 + p_2(\lambda_2 - \mu_2 c_2) - p_2(\lambda_2 + \lambda_1 p_1)\pi_c(0,0)}{\lambda_1(1-p_1p_2)} + p_1\pi_c(0,0)$$
$$\beta_2 = (1-p_2)\frac{\lambda_2 - \mu_2 c_2 + p_1(\lambda_1 - \mu_1 c_1) - p_1(\lambda_1 + \lambda_2 p_2)\pi_c(0,0)}{\lambda_2(1-p_1p_2)} + p_2\pi_c(0,0)$$

and the quantity $\pi_c(0,0)$ is given by

$$\pi_c(0,0) = \begin{cases} \dfrac{\lambda_1 + \lambda_2 p_2 - \mu_1 c_1 - \mu_2 c_2 p_2}{(\lambda_1 + \lambda_2 p_2)\varphi_Y(1)} & \text{if } \lambda_2 > \mu_2 c_2, \\ \\ \dfrac{\lambda_2 + \lambda_1 p_1 - \mu_1 c_1 p_1 - \mu_2 c_2}{p_1(\lambda_1 + \lambda_2 p_2)\varphi_Y(1)} & \text{if } \lambda_2 < \mu_2 c_2, \end{cases} \tag{4.22}$$

where $\varphi_Y(y)$ is defined by Equation (4.19).

Proof. In the case $\lambda_2 > \mu_2 c_2$, the result easily follows using Equation (4.21) for $y = 1$ and Equation (4.8).

In the case $\lambda_2 < \mu_2 c_2$ (and then $\lambda_1 > \mu_1 c_1$ by Condition (E)), we have $X_0(1) = 1$ and then, by Relation (4.16), one gets the expression for $\alpha_Y(1)$,

$$\alpha_Y(1) = p_1 \frac{\lambda_1 + \lambda_2 p_2 - \mu_1 c_1 - \mu_2 c_2 p_2}{\lambda_2 + \lambda_1 p_1 - \mu_1 c_1 p_1 - \mu_2 c_2}.$$

Equation (4.22) then easily follows. The formulas for the blocking probabilities are obtained using Relations (4.6) and (4.7) for β_1 and β_2 and the expressions (4.8) and (4.9) for $P(0,1)$ and $P(1,0)$. □

To conclude this section, it is worth noting that the computation of the function $\varphi_Y(y)$ in the quantity $\pi_c(0,0)$ involves elliptic integrals. In addition, a similar result holds for the function $P(x,0)$. This enables us to completely compute the generating function $P(x,y)$.

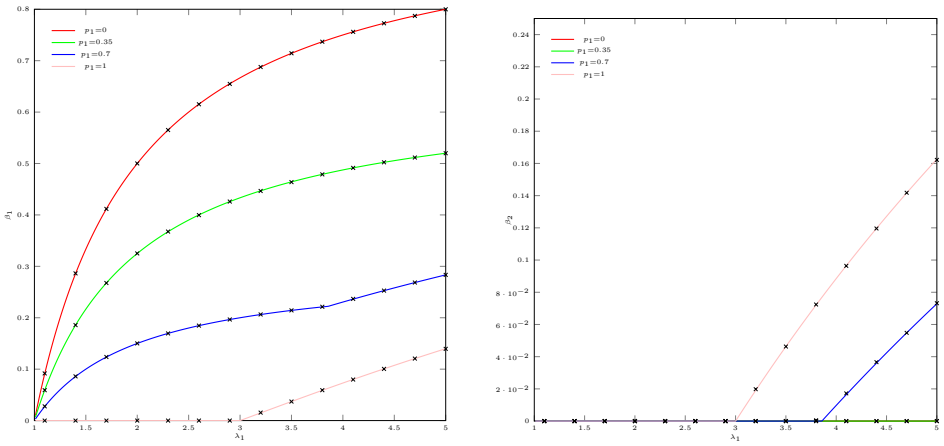
4.5 Numerical results: Offloading small data centres

In this section, we illustrate the results obtained in the previous sections (in particular Theorem 4.3) in order to estimate the gain achieved by the offloading scheme. We assume that the service rate at both facilities is the same and taken equal to unity ($\mu_1 = \mu_2 = 1$). Assume in addition that the first data centre has a capacity much smaller than the second one, e.g., $c_1 = 1$ and $c_2 = 10$. We consider the case when all the requests blocked at the first data centre are forwarded to the second one ($p_1 = 1$) and none blocked at the second data centre is forwarded to the first one ($p_2 = 0$).

In Figures 4.4 and 4.5, when the arrival rate λ_1 at the first data centre increases, the loss rate β_1 goes from 0 if (A) or (B_1) holds to a positive value if (B_2) or (E) holds. For example in Figure 4.4, for $p_1 = 1$, we can see the

transition from (B_2) to (E) when $\lambda_1 = 3$, and for $p_1 = 0.7$, the transition from (B_2) to (E) when $\lambda_1 = 1 + 2/0.7 \simeq 3.85$. We can checked that β_1 is a continuous and not differentiable function of λ_1 at $1 + 2/0.7$. If $p_1 = 0.35$ or $p_1 = 0$, (E) holds for the range of values $[1, 5]$ considered here for λ_1. In Figure 4.4, $\lambda_2 = 12$ thus (E) holds for $\lambda_1 \in [1, 5]$, as $\lambda_1 > \mu_1 c_1$ and $\lambda_2 > \mu_2 c_2$.

In conclusion, Figures 4.4 and 4.5 show that the offloading mechanism improves a lot the loss rate β_1 of the requests of class 1 and does not significantly deteriorate the corresponding performances at facility #2 in the case of systematic rerouting ($p_1 = 1$), even when this data centre is already significantly loaded as in Figure 4.5 (B). This means that offloading small date centres with a big back-up data centre is a good strategy to reduce blocking in fog computing.



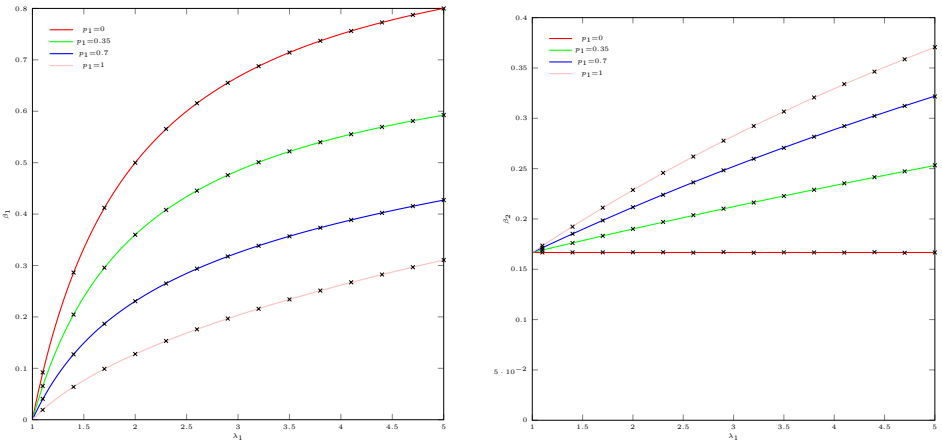
(a) Loss probability of class 1 (b) Loss probability of class 2

Figure 4.4 – Loss probabilities as a function of λ_1 with $\lambda_2 = 8$, $c_1 = 1$, $c_2 = 10$, $\mu_1 = 1$, $\mu_2 = 1$, $p_2 = 0$. The crosses represent simulation points while solid curves are plotted from analytical results.

Figures 4.6 illustrate the impact of the choice of p_1 when facility #2 is almost overloaded, $\lambda_2 = 9.9$ so that $\lambda_2 < c_2\mu_2$, and with a high load $\lambda_2 = 11.1$. As it can be seen, even when $p_1 = 1$, the performances of class 2 requests are not really impacted by the offloading scheme, whereas the loss rate of class 1 is significantly changed. This confirms the benefit of the offloading strategy.

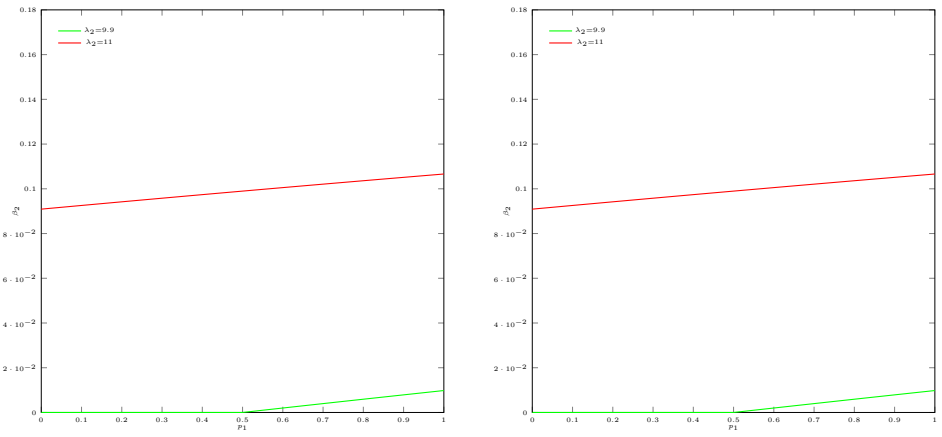
4.6 Conclusion

We have proposed in this chapter an analytical model to study a simple offloading strategy for data centres in the framework of fog computing under heavy loads. The strategy considered consists of forwarding with a certain probability requests blocked at a small data centre to a big back-up data centre. The model considered could also be used to study the offload of requests blocked at the big data centre onto a small data centre but this case has not been considered in the numerical applications. The key finding is that the proposed strategy can significantly improves blocking at a small data centre



(a) Loss probability of class 1 (b) Loss probability of class 2

Figure 4.5 – Loss probabilities as a function of λ_1 with $\lambda_2 = 12$, $c_1 = 1$, $c_2 = 10$, $\mu_1 = 1$, $\mu_2 = 1$, $p_2 = 0$. The crosses represent simulation points while solid curves are plotted from analytical results.



(a) Loss probability of class 1 (b) Loss probability of class 2

Figure 4.6 – Loss probabilities as a function of p_1 with $c_1 = 1$, $c_2 = 10$, $\lambda_1 = 1.2$, $\mu_1 = 1$, $\mu_2 = 1$, $p_2 = 0$.

without affecting too much blocking at the big data centre.

Chapter 5

Analysis of a trunk reservation policy in the framework of Fog Computing

5.1 Introduction

Fog computing [WRSvdM11, BMZA12, RBG12] is considered by many actors of the telecommunication ecosystem as a major breakthrough in the design of networks for both network operators and content providers. For the former, deploying storage and computing data centres at the edge of the network enables them to reduce the load within the network and on critical links such as peering links. In addition, network operators can take benefit of these data centres to dynamically instantiate virtualised network functions. Content providers can take benefit of distributed storage and computing data centres to optimise service platform placement and thus to improve the quality experienced by end users.

Fog computing relies on distributed data centres, which are much smaller than big centralised data centres generally used in cloud computing. Because of potential resource limitation, user requests may be blocked if resources are exhausted at a (small) data centre. This is a key difference with cloud computing where resources are often considered as infinite. In this chapter, we consider the case of a computing resource service where users request servers available at a data centre. If no server is available, then a user request may be blocked.

To reduce the blocking probability, it may be suitable that data centres collaborate. This is certainly a key issue in fog computing, which makes the design of networks and fog computing very different from that of cloud computing. Along this line of investigations, an offloading scheme has been investigated in [FGRT16a], where requests blocked at a data centre are forwarded to another one with a given probability. In this chapter, we investigate the case when a blocked request is systematically forwarded to another data centre but to protect those requests which are originally assigned to that data centre, a redirected request is accepted only if there is a sufficient large number of idle servers. In the framework of telephone networks, this policy is known as *trunk*

reservation [Ros95].

In the following, we consider the case of two data centres, where the trunk reservation policy is applied in one server only; the analysis of the case when the policy is applied in both data centres is a straightforward extension of the case considered but involves much more computations. We further simplify the system by reasonably assuming that both data centres have a large number of servers. From a theoretical point of view, this leads us to rescale the system and to consider limiting processes. The eventual goal of the present analysis is to estimate the gain achieved by the trunk reservation policy.

Considering the number of free servers in both data centres, we are led after rescaling to analyse a random walk in the quarter plane. This kind of process has been extensively studied in the technical literature (see for instance the book by Fayolle *et al.* [FIM17]). For the random walk appearing in this chapter, even if the kernel is similar to that analysed in [FI79] (and in Fricker *et al.* [FGRT16a]), the key difference is that the reflecting conditions on the boundaries of the quarter plane are not constant. More precisely, the reflecting coefficients in the negative vertical direction along the y-axis take three different values depending on a given threshold (namely, the trunk reservation threshold).

This simple difference with usual random walks in the quarter plane makes the analysis much more challenging. Contrary to the usual case which consists of determining two unknown functions, we have in the present case to decompose one unknown function into two pieces (one polynomial and one infinite generating function) and thus to determine three unknown functions. We show that the coefficients of the unknown polynomial can be computed by solving a linear system. Once this polynomial is determined, the two other functions can be derived. This eventually allows us to compute the blocking probabilities at the two data centres and to estimate the efficiency of the trunk reservation policy in the framework of fog computing.

This chapter is organised as follows: In Section 5.2, we introduce the notation and we show convergence results for the rescaled system. We analyse in Section 5.3, the limiting random walk, in particular its kernel. The associated boundary value problems are formulated and solved in Section 5.4. Finally, some numerical results are discussed in Section 5.5.

5.2 Model description

5.2.1 Notation

We consider in this chapter two data centres in parallel. The first one is equipped with C_1 servers and serves customers arriving according to a Poisson process with rate Λ_1 and requesting exponentially distributed service times with mean $1/\mu_1$ (a customer if accepted occupies a single server of the data centre); the number of busy servers in this first data centre is denoted by $N_1(t)$ at time t. Similarly, the second data centre is equipped with C_2 servers and accommodate service requests arriving according to a Poisson process with rate Λ_2 and service demands exponentially distributed with mean $1/\mu_2$; the

number of requests in progress is denoted by $N_2(t)$ at time t. Note that the system being with finite capacity is always stable.

To reduce the blocking probability at the first service data centre without exhausting the resources of the second one, we assume that when data centre 1 is full and there are at least a servers available at data centre 2, then requests arriving at data centre 1 are served at data centre 2. Figure 5.1 shows how both data centres deal with this cooperative scheme. Dashed lines represent the flows of blocked requests.

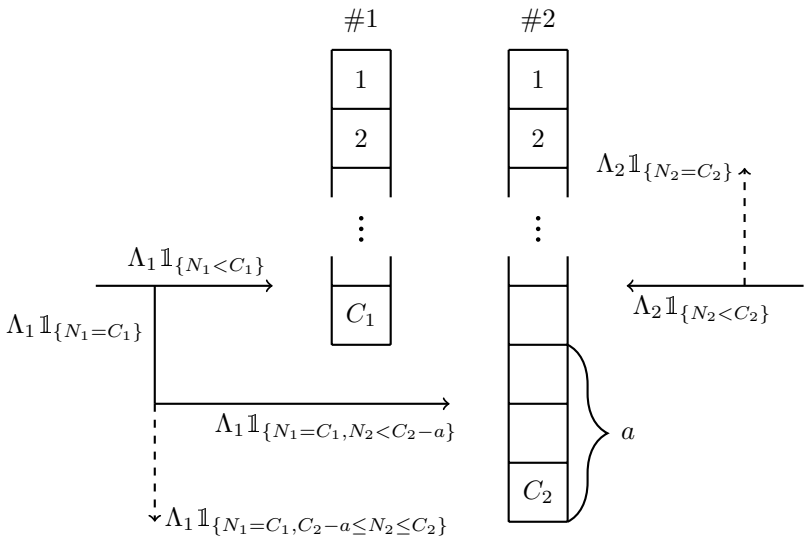


Figure 5.1 – Policy implementation scheme

Owing to the Poisson arrival and exponential service time assumptions, the process $(N(t)) = ((N_1(t), N_2(t)),\, t \geq 0)$ is a Markov chain, which takes values in the set $\{0, \ldots, C_1\} \times \{0, \ldots, C_2\}$. The transition rates of the Markov chain $(N(t))$ are given by

$$q(N, N + (k, \ell)) = \begin{cases} \Lambda_1 \mathbb{1}_{\{N_1<C_1\}} & \text{if } (k, \ell) = (1, 0) \\ \Lambda_2 \mathbb{1}_{\{N_2<C_2\}} + \Lambda_1 \mathbb{1}_{\{N_1=C_1, N_2<C_2-a\}} & \text{if } (k, \ell) = (0, 1) \\ \mu_1 N_1 & \text{if } (k, \ell) = (-1, 0) \\ \mu_2 N_2 & \text{if } (k, \ell) = (0, -1). \end{cases}$$

In the following, we consider the process

$$(m(t)) \overset{\text{def.}}{=} ((C_1 - N_1(t), C_2 - N_2(t)),\, t \geq 0),$$

describing the number of idle servers in both data centres. In the next section, we investigate the case when the arrival rates at data centres are scaled up by a factor $\nu > 0$.

5.2.2 Rescaled system

Let us assume that the arrival rates Λ_1 and Λ_2 are scaled up by a factor ν, i.e., $\Lambda_1 = \nu\lambda_1$ and $\Lambda_2 = \nu\lambda_2$ for some factor ν and real $\lambda_1 > 0$ and

$\lambda_2 > 0$. We further assume that the capacities C_1 and C_2 scale with ν, namely $C_1 = \nu c_1$ and $C_2 = \nu c_2$ for some positive constants c_1 and c_2. To indicate the dependence of the numbers of occupied and idle servers upon ν, we write $N_i^{[\nu]}$ and $m_i^{[\nu]}$ instead of N_i and m_i to denote respectively the number of occupied and idle servers in data centres i for $i = 1, 2$.

With the above hypotheses, we are led to consider $\left(m^{[\nu]}(t)\right)$ as a random walk. For this purpose, let us introduce the random walk $(n(t)) = ((n_1(t), n_2(t)),\, t \geq 0)$ in the positive quadrant with transition rates for $n \in \mathbb{N}_*^2$

$$r((n_1, n_2), (n_1 + k, n_2 + \ell)) = \begin{cases} \lambda_1 & \text{if } (k, \ell) = (-1, 0) \\ \lambda_2 & \text{if } (k, \ell) = (0, -1) \\ \mu_1 c_1 & \text{if } (k, \ell) = (1, 0) \\ \mu_2 c_2 & \text{if } (k, \ell) = (0, 1) \end{cases}$$

and the reflecting conditions for $n_1 > 0$ and $n_2 = 0$

$$r((n_1, 0), (n_1 + k, \ell)) = \begin{cases} \lambda_1 \mathbb{1}_{\{n_1 > 0\}} & \text{if } (k, \ell) = (-1, 0) \\ \mu_1 c_1 & \text{if } (k, \ell) = (1, 0) \\ \mu_2 c_2 & \text{if } (k, \ell) = (0, 1), \end{cases}$$

for $n_1 = 0$ and $n_2 > 0$

$$r((0, n_2), (k, n_2 + \ell)) = \begin{cases} \lambda_2 + \lambda_1 \mathbb{1}_{\{n_2 > a\}} & \text{if } (k, \ell) = (-1, 0) \\ \mu_1 c_1 & \text{if } (k, \ell) = (1, 0) \\ \mu_2 c_2 & \text{if } (k, \ell) = (0, 1), \end{cases}$$

and for $n = 0$

$$r(0, (k, \ell)) = \begin{cases} \mu_1 c_1 & \text{if } (k, \ell) = (1, 0) \\ \mu_2 c_2 & \text{if } (k, \ell) = (0, 1), \end{cases}$$

where a is the threshold introduced in the previous section.

Proposition 5.1. *If $m^{[\nu]}(0) = (k, \ell) \in \mathbb{N}^2$ is fixed then, for the convergence in distribution,*

$$\lim_{\nu \to +\infty} \left(m^{[\nu]}(t/\nu),\, t \geq 0\right) = (n(t),\, t \geq 0).$$

Proof. If f is a function on $\mathbb{N}^2$ with finite support then classical stochastic

calculus gives the relation

$$
\begin{aligned}
f\left(m^{[\nu]}(t/\nu)\right) &= f(k,\ell) + M^{[\nu]}(t/\nu) \qquad (5.1)\\
&+ \int_0^t \mu_1 \frac{\nu c_1 - m_1^{[\nu]}(s/\nu)}{\nu} \left[f\left(m^{[\nu]}(s/\nu) + e_1\right) - f\left(m^{[\nu]}(s/\nu)\right)\right] \mathrm{d}s\\
&+ \int_0^t \mu_2 \frac{\nu c_2 - m_2^{[\nu]}(s/\nu)}{\nu} \left[f\left(m^{[\nu]}(s/\nu) + e_2\right) - f\left(m^{[\nu]}(s/\nu)\right)\right] \mathrm{d}s\\
&+ \int_0^t \lambda_1 \mathbb{1}_{\left\{m_1^{[\nu]}(s/\nu)>0\right\}} \left[f\left(m^{[\nu]}(s/\nu) - e_1\right) - f\left(m^{[\nu]}(s/\nu)\right)\right] \mathrm{d}s\\
&+ \int_0^t \lambda_2 \mathbb{1}_{\left\{m_2^{[\nu]}(s/\nu)>0\right\}} \left[f\left(m^{[\nu]}(s/\nu) - e_2\right) - f\left(m^{[\nu]}(s/\nu)\right)\right] \mathrm{d}s\\
&+ \int_0^t \lambda_1 \mathbb{1}_{\left\{m_1^{[\nu]}(s/\nu)=0, m_2^{[\nu]}(s/\nu)>a\right\}} \left[f\left(m^{[\nu]}(s/\nu) - e_2\right) - f\left(m^{[\nu]}(s/\nu)\right)\right] \mathrm{d}s
\end{aligned}
$$

where $M^{[\nu]}(t) = \left(M_1^{[\nu]}(t), M_2^{[\nu]}(t)\right)$ is a martingale, $e_1 = (1,0)$ and $e_2 = (0,1)$.

For $i = 1, 2$, since the process $\left(N_i^{[\nu]}(s/\nu)\right)$ is stochastically bounded by a Poisson process with rate $\mu_i c_i$, hence, for the convergence of processes,

$$\lim_{\nu\to+\infty} \left(\frac{\nu c_i - m_i^{[\nu]}(s/\nu)}{\nu}\right) = (c_i).$$

Using by Theorem 4.5 of Jacod and Shiryaev [JS03, p. 320], one gets that the sequence of processes $\left(m^{[\nu]}(t/\nu), t \geq 0\right)$ is tight. If $(z(t))$ is the limit of some convergent subsequence $\left(m^{[\nu_k]}(t/\nu_k)\right)$, then, if $R = \left(r(a,b),\ a,b \in \mathbb{N}^2\right)$ is the $\mathcal{Q}$-matrix of $(n(t))$, Relation (5.1) gives that

$$\left(f(z(t)) - f(k,\ell) - \int_0^t R \cdot f(z(u))\,\mathrm{d}u\right)$$

is a martingale. Some of the technical details are skipped, see Hunt and Kurtz [HK94] for a similar context for example. This shows that $(z(t))$ is the Markov process with transition matrix R, hence $(z(t))$ has the same distribution as $(n(t))$. See Section IV-20 of Rogers and Williams [RW00, p. 30]. Since this is the only possible limit, the convergence in distribution is proved. □

Using the results of [FMM95], we have the following stability condition.

Lemma 5.1. *A stationary regime exists for the random walk $(n(t))$ if and only if*

$$\lambda_1 > \mu_1 c_1 \text{ and } \mu_1 c_1 + \mu_2 c_2 < \lambda_1 + \lambda_2. \qquad (5.2)$$

To conclude this section, let us note that the limiting random walk $(n(t))$ describes the number of customers in two $M/M/1$ queues in tandem. The arrival rate at the first queue is $\mu_1 c_1$ and the service rate is λ_1; this queue is independent of the second one and is stable under Condition (5.2). The arrival rate at the second queue is $\mu_2 c_2$ and the service rate is λ_2 plus λ_1 when the first queue is empty and there are more than a customers in the second queue. This last term introduces a coupling between the two queues. See Figure 5.2 for an illustration.

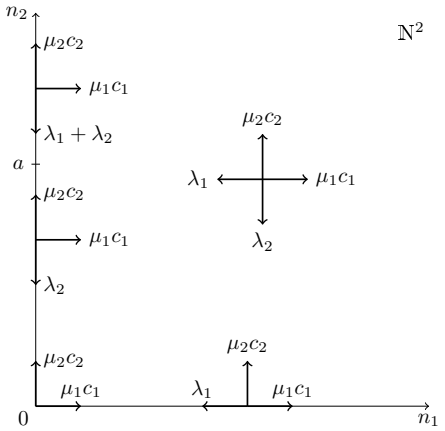


Figure 5.2 – Transitions for $(n(t))$.

5.3 Analysis of the limiting random walk

We assume in this section that the random walk is ergodic. In other words, Condition (5.2) is satisfied.

5.3.1 Functional equation

Let $p(n_1, n_2)$ denote the stationary probability of being in state $(n_1, n_2) \in \mathbb{N}^2$ in the stationary regime. It is easily checked that the balance equation reads for $n_1, n_2 \geq 0$

$$\begin{aligned}\Big(\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2 - \lambda_2 \mathbb{1}_{\{n_2=0\}} - \lambda_1 \mathbb{1}_{\{n_1=0, 0\leq n_2\leq a\}}\Big)\, p(n_1, n_2) \\ = \mu_1 c_1 p(n_1 - 1, n_2) + \mu_2 c_2 p(n_1, n_2 - 1) + \lambda_1 p(n_1 + 1, n_2) \\ + \lambda_2 p(n_1, n_2 + 1) + \lambda_1 p(0, n_2 + 1) \mathbb{1}_{\{n_1=0, n_2\geq a\}}. \quad (5.3)\end{aligned}$$

Define for x and y in the unit disk $D \overset{\text{def.}}{=} \{z \in \mathbb{C} : |z| < 1\}$ the generating functions

$$\begin{aligned}P(x, y) &\overset{\text{def.}}{=} \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(n_1, n_2) x^{n_1} y^{n_2}, \\ P_1(x) &\overset{\text{def.}}{=} \sum_{n_1=0}^{\infty} p(n_1, 0) x^{n_1}, \text{ and} \\ P_2(y) &\overset{\text{def.}}{=} \sum_{n_2=0}^{\infty} p(0, n_2) y^{n_2}.\end{aligned}$$

By definition, the function $P(x, y)$ is analytic in $D \times D$ and the functions $P_1(x)$ and $P_2(y)$ are analytic in D.

Multiplying Equation (5.3) by the term $x^{n_1} y^{n_2}$ and summing for n_1 and n_2 ranging from zero to infinity, we obtain the functional equation

$$K(x, y) P(x, y) = \lambda_2 x(1 - y) P_1(x) + \lambda_1 (y - x) P_2(y) + \lambda_1 x(1 - y) P_2^-(y), \quad (5.4)$$

where the kernel $K(x,y)$ is defined by

$$K(x,y) \overset{\text{def.}}{=} \mu_1 c_1 x^2 y + \mu_2 c_2 x y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2) xy + \lambda_1 y + \lambda_2 x \tag{5.5}$$

and the polynomial $P_2^-(y)$ by

$$P_2^-(y) \overset{\text{def.}}{=} \sum_{n_2=0}^{a} p(0,n_2) y^{n_2}.$$

From the functional equation (5.4), let us note that the generating function of the number of customers in the first queue is given by

$$P(x,1) = \frac{\lambda_1}{\lambda_1 - \mu_1 c_1 x} P_2(1),$$

from which we deduce that $P_2(1) = \mathbb{P}(n_1 = 0) = 1 - \mu_1 c_1/\lambda_1$. This is obvious since the first queue is an $M/M/1$ queue with input rate $\mu_1 c_1$ and service rate λ_1, independent of the second queue. See Section 6.7 of Robert [Rob03, p. 164] for a complete proof.

In addition, we have

$$P(1,y) = \frac{\lambda_2 P_1(1) - \lambda_1 (P_2(y) - P_2^-(y))}{\lambda_2 - \mu_2 c_2 y}.$$

The normalising condition $P(1,1) = 1$ yields

$$\lambda_1 \mathbb{P}(n_1 = 0, n_2 \leq a) + \lambda_2 \mathbb{P}(n_2 = 0) = \lambda_1 + \lambda_2 - \mu_1 c_1 - \mu_2 c_2. \tag{5.6}$$

The quantity $B_1 = \mathbb{P}(N_1 = 0, N_2 \leq a)$ is the blocking probability of customers originally trying to access the first data centre and $B_2 = \mathbb{P}(n_2 = 0)$ is the blocking probability of customers trying to reach the second data centre. The above relation is the global rate conservation equation of the system. The performance of the system is actually characterised by the blocking probabilities, given by the generating functions as

$$B \overset{\text{def.}}{=} \left(P_2^-(1), P_1(1) \right).$$

In the following, we intend to compute the unknown generating functions $P_1(x)$, $P_2(y)$ and $P_2^-(y)$.

5.3.2 Zero pairs of the kernel

The kernel $K(x,y)$ has already been studied in Fayolle and Iasnogorodski [FI79] in the framework of coupled data centres. For fixed y, the kernel $K(x,y)$ defined by Equation (5.5) has two roots $X_0(y)$ and $X_1(y)$. Using the usual definition of the square root such that $\sqrt{a} > 0$ for $a > 0$, the solution which is null at the origin and denoted by $X_0(y)$, is defined and analytic in $\mathbb{C} \setminus ([y_1, y_2] \cup [y_3, y_4])$ where the real numbers y_1, y_2, y_3 and y_4 are such that $0 < y_1 < y_2 < 1 < y_3 < y_4$. The other solution $X_1(y)$ is meromorphic in

$\mathbb{C} \setminus ([y_1, y_2] \cup [y_3, y_4])$ with a pole at 0. The function $X_0(y)$ is precisely defined by

$$X_0(y) \stackrel{\text{def.}}{=} \frac{-(\mu_2 c_2 y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)y + \lambda_2) + \sigma_1(y)}{2\mu_1 c_1 y},$$

where $\sigma_1(y)$ is the analytic extension in $\mathbb{C} \setminus ([y_1, y_2] \cup [y_3, y_4])$ of the function defined in the neighbourhood of 0 as $\sqrt{\Delta_1(y)}$ with

$$\Delta_1(y) \stackrel{\text{def.}}{=} (\mu_2 c_2 y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)y + \lambda_2)^2 - 4\mu_1 c_1 \lambda_1 y^2.$$

The other solution $X_1(y) \stackrel{\text{def.}}{=} \lambda_1/(\mu_1 c_1 X_0(y))$.

When y crosses the segment $[y_1, y_2]$, $X_0(y)$ and $X_1(y)$ describe the circle C_{r_1} with centre 0 and radius $r_1 = \sqrt{\lambda_1/(\mu_1 c_1)} > 1$.

Similarly, for fixed x, the kernel $K(x, y)$ has two roots $Y_0(x)$ and $Y_1(x)$. The root $Y_0(x)$, which is null at the origin, is analytic in $\mathbb{C} \setminus ([x_1, x_2] \cup [x_3, x_4])$ where the real numbers x_1, x_2, x_3 and x_4 are such that with $0 < x_1 < x_2 < 1 < x_3 < x_4$ and is given by

$$Y_0(x) \stackrel{\text{def.}}{=} \frac{-(\mu_1 c_1 x^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)x + \lambda_1) + \sigma_2(x)}{2\mu_2 c_2 x}$$

where $\sigma_2(x)$ is the analytic extension in of the function defined in the neighbourhood of 0 as $\sqrt{\Delta_2(x)}$ with

$$\Delta_2(x) \stackrel{\text{def.}}{=} (\mu_1 c_1 x^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)x + \lambda_1)^2 - 4\mu_2 c_2 \lambda_2 x^2.$$

The other root $Y_1(x) \stackrel{\text{def.}}{=} \lambda_2/(\mu_2 c_2 Y_0(x))$ and is meromorphic in $\mathbb{C} \setminus ([x_1, x_2] \cup [x_3, x_4])$ with a pole at the origin.

When x crosses the segment $[x_1, x_2]$, $Y_0(y)$ and $Y_1(y)$ describe the circle C_{r_2} with centre 0 and radius $r_2 = \sqrt{\lambda_2/(\mu_2 c_2)}$.

5.4 Boundary value problems

To solve the boundary value problems encountered in the following, let us recall that if we search for a function $P(z)$ analytic in the disk $D_r \stackrel{\text{def.}}{=} \{z \in \mathbb{C} : |z| < r\}$ for some $r > 0$, such that for $z \in C_r \stackrel{\text{def.}}{=} \{z \in \mathbb{C} : |z| = r\}$, $P(z)$ satisfies

$$\Re(ig(z)P(z)) = \Re(ih(z))$$

for some functions $g(z)$ and $h(z)$ analytic in a neighbourhood of C_r and $g(z)$ does not cancel in this neighbourhood, then the function

$$\tilde{P}(z) \stackrel{\text{def.}}{=} \begin{cases} P(z) & z \in D_r \\ \overline{P\,(1/z)} & z \in \mathbb{C} \setminus \overline{D_r} \end{cases}$$

is solution to the following Riemann-Hilbert problem: The function $\tilde{P}(z)$ is sectionally analytic with respect to the circle C_r and verifies for $z \in C_r$

$$g(z)\tilde{P}^i(z) - \overline{g(z)}\tilde{P}^e(z) = 2i\Im(h(z)),$$

where $\tilde{P}^i(z)$ (resp. $\tilde{P}^e(z)$) is the interior (resp. exterior) limit of the function $\tilde{P}(z)$ at the circle C_r, and $\overline{D_r} \stackrel{\text{def.}}{=} D_r \cup C_r$.

From [DL90], the solution to this Riemann-Hilbert problem when it exists is given for $z \in \mathbb{C} \setminus C_r$ by

$$\tilde{P}(z) \stackrel{\text{def.}}{=} \frac{\phi(z)}{\pi} \int_{C_r} \frac{\Im(h(\xi))}{g(\xi)\phi^i(\xi)} \frac{\mathrm{d}\xi}{\xi - z} + \phi(z)\mathcal{P}(z),$$

where $\mathcal{P}(z)$ is a polynomial and $\phi^i(z)$ (resp. $\phi^e(z)$) is the interior (resp. exterior) limit at the circle C_r of the solution $\phi(z)$ to the following homogeneous Riemann-Hilbert problem: The function $\phi(z)$ is sectionally analytic with respect to the circle C_r and, for $z \in C_r$,

$$\phi^i(z) \stackrel{\text{def.}}{=} \alpha(z)\phi^e(z),$$

where

$$\alpha(z) \stackrel{\text{def.}}{=} \frac{\overline{g(z)}}{g(z)}.$$

The function $\phi(z)$ is given by

$$\phi(z) \stackrel{\text{def.}}{=} \begin{cases} \exp\left(\frac{1}{2\pi i} \int_{C_r} \log\left(\frac{\alpha(\xi)}{\xi^\kappa}\right) \frac{\mathrm{d}\xi}{\xi - z}\right) & z \in D_r \\ \frac{1}{z^\kappa} \exp\left(\frac{1}{2\pi i} \int_{C_r} \log\left(\frac{\alpha(\xi)}{\xi^\kappa}\right) \frac{\mathrm{d}\xi}{\xi - z}\right) & z \in \mathbb{C} \setminus \overline{D_r}, \end{cases}$$

where κ is the index of the Riemann-Hilbert problem defined by

$$\kappa \stackrel{\text{def.}}{=} \frac{1}{2\pi} \operatorname{var}_{z \in C_r} \arg \alpha(z).$$

The existence and the uniqueness of the solution depends on the value of the index κ. When $\kappa = 0$, the solution exists and is unique with $\mathcal{P}(z) = P(0)$.

5.4.1 Function $P_1(x)$

Let us first establish a relation between the functions $P_1(x)$ and $P_2^-(y)$. For this purpose, let us define for $x \in C_{r_1}$

$$\alpha_1(x) \stackrel{\text{def.}}{=} \frac{\bar{x}(Y_0(x) - x)}{x(Y_0(x) - \bar{x})} = \frac{\lambda_1(Y_0(x) - x)}{x(\mu_1 c_1 x Y_0(x) - \lambda_1)} \tag{5.7}$$

and let us consider the following homogeneous Riemann-Hilbert problem: The function $\phi_1(x)$ is sectionally analytic with respect to the circle C_{r_1}, and for some $x \in C_{r_1}$

$$\phi_1^i(x) = \alpha_1(x)\phi_1^e(x), \tag{5.8}$$

where $\phi_1^i(x)$ (resp. $\phi_1^e(x)$) is the interior (resp. exterior) limit at the circle C_{r_1}.

Lemma 5.2. *The function $\phi_1(x)$ is given for some $x \in D_{r_1}$, by*

$$\phi_1(x) \stackrel{\text{def.}}{=} \exp\left(\frac{x}{\pi}\int_{y_1}^{y_2} \frac{(\lambda_2 - \mu_2 c_2 y^2)\,\Theta_1(y)}{yK(x,y)}\,\mathrm{d}y\right), \tag{5.9}$$

where

$$\Theta_1(y) \stackrel{\text{def.}}{=} \operatorname{ArcTan}\left(\frac{\sqrt{-\Delta_1(y)}}{\mu_2 c_2 y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)y + 2\lambda_1 + \lambda_2}\right). \tag{5.10}$$

Proof. The index of the Riemann-Hilbert problem (5.8) is given by

$$\kappa_1 \stackrel{\text{def.}}{=} \frac{1}{2\pi}\operatorname{var}_{x\in C_{r_1}} \arg\ \alpha_1(x)$$

where $\alpha_1(x)$ is defined by Equation (5.7). For $x \in C_{r_1}$, we have

$$\Re\left(\bar{x}Y_0(x) - \frac{\lambda_1}{\mu_1 c_1}\right) < 0,$$

since $Y_0(x) \in [y_1, y_2]$ and $\Re(\bar{x}) \le r_1$. Hence, $\kappa_1 = 0$.

The solution to the Riemann-Hilbert problem is then given for $x \in D_{r_1}$ by

$$\phi_1(x) = \exp\left(\frac{1}{2\pi i}\int_{C_{r_1}} \frac{\log\alpha_1(z)}{z-x}\,\mathrm{d}z\right).$$

For $x \in C_{r_1}$, simple manipulations show that for $x = X_0(y + 0i)$ for $y \in [y_1, y_2]$

$$\alpha_1(x) = e^{-2i\Theta_1(y)},$$

where $\Theta_1(y)$ is defined by Equation (5.10), since

$$x = \frac{-(\mu_2 c_2 y^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)y + \lambda_2) - i\sqrt{\Delta_1(y)}}{2\mu_1 c_1 y}.$$

It follows that for $x \in D_{r_1}$

$$\begin{aligned}\frac{1}{2\pi i}\int_{C_{r_1}} \frac{\log\alpha_1(z)}{z-x}\,\mathrm{d}z = \\ \frac{-1}{\pi}\int_{y_1}^{y_2} \Theta_1(y)\left(\frac{\frac{\mathrm{d}X_0(y+0i)}{\mathrm{d}y}}{X_0(y+0i)-x} + \frac{\frac{\mathrm{d}X_0(y-0i)}{\mathrm{d}y}}{X_0(y-0i)-x}\right)\mathrm{d}y = \\ \frac{1}{\pi}\int_{y_1}^{y_2} \frac{x(\lambda_2 - \mu_2 c_2 y^2)\Theta_1(y)}{yK(x,y)}\,\mathrm{d}y,\end{aligned}$$

where we have used the fact that

$$\frac{\mathrm{d}X_0(y+0i)}{\mathrm{d}y} = \frac{X_0(y+0i)\left(-\mu_2 c_2 y + \frac{\lambda_2}{y}\right)}{-i\sqrt{-\Delta_1(y)}}$$

and

$$(X_0(y+0i) - x)(X_0(y-0i) - x) = \frac{K(x,y)}{\mu_1 c_1 y}.$$

Equation (5.9) then follows. □

Corollary 5.1. *The interior limit of the function $\phi_1(x)$ at the circle C_{r_1} is given for $x = X_0(y + 0i)$ with $y \in [y_1, y_2]$ by*

$$\phi_1^i(x) = \exp\left(-i\Theta_1(y) + \Phi_1(y)\right),$$

where

$$\Phi_1(y) \overset{\text{def.}}{=} \frac{1}{\pi} \oint_{y_1}^{y_2} \frac{y(\lambda_2 - \mu_2 c_2 \xi^2)\Theta_1(\xi)}{\xi(\mu_2 c_2 y\xi - \lambda_2)} \frac{\mathrm{d}\xi}{\xi - y}, \tag{5.11}$$

and the symbol $\oint$ denotes the Cauchy integral [DL90].

Using Plemelj's formula, we have for $x = X_0(y + 0i)$ with $y \in [y_1, y_2]$

$$\begin{aligned}\phi_1^i(x) &= \exp\left(\frac{\log \alpha_1(x)}{2} + \frac{1}{2\pi i}\int_{C_{r_1}} \frac{\log \alpha_1(z)}{z - x}\,\mathrm{d}z\right) \\ &= \exp\left(-i\Theta_1(y) + \frac{1}{\pi}\oint_{y_1}^{y_2} \frac{x(\lambda_2 - \mu_2 c_2 y^2)\Theta_1(\xi)}{yK(x,y)}\,\mathrm{d}y\right).\end{aligned}$$

Also, using the fact that for $x' = \frac{\lambda_1}{\mu_1 c_1 x}$,

$$\frac{x'}{K(x', y)} = \frac{x}{K(x, y)},$$

we deduce that for $x = X_0(y \pm 0i)$ the Cauchy integral

$$\frac{1}{\pi}\oint_{y_1}^{y_2} \frac{x(\mu_2 c_2 \xi^2 - \lambda_2)\Theta_1(\xi)}{\xi K(x, \xi)}\,\mathrm{d}\xi$$

is real and equal to

$$\Phi_1(y) = \frac{1}{\pi}\oint_{y_1}^{y_2} \frac{y(\lambda_2 - \mu_2 c_2 \xi^2)\Theta_1(\xi)}{\xi(\xi - y)(\mu_2 c_2 y\xi - \lambda_2)}\,\mathrm{d}\xi.$$

It follows that for $x = X_0(y \pm 0i)$

$$\phi_1^i(x) = \exp\left(\Phi_1(y) - i\Theta_1(y)\right).$$

The results above allow us to establish a relation between functions $P_1(x)$ and $P_2^-(y)$.

Proposition 5.2. *The functions $P_1(x)$ and $P_2^-(y)$ are such that for $x \in D_{r_1}$*

$$\begin{aligned}P_1(x) &= \phi_1(x)P(0,0) \\ &\quad + \frac{x\lambda_1\phi_1(x)}{\lambda_2\pi}\int_{y_1}^{y_2} \frac{(\lambda_2 - \mu_2 c_2 y^2)P_2^-(y)}{yK(x,y)} \sin(\Theta_1(y))e^{-\Phi_1(y)}\,\mathrm{d}y,\end{aligned} \tag{5.12}$$

where the functions $\phi_1(x)$ and $\Phi_1(y)$ are defined by Equations (5.9) and (5.11), respectively.

Proof. For $x = X_0(y)$, we have from Equation (5.4)

$$\frac{\lambda_2 X_0(y)}{y - X_0(y)} P_1(x) = -\frac{X_0(y)\lambda_1 P_2^-(y)}{y - X_0(y)} - \frac{\lambda_1 P_2(y)}{1-y},$$

The above equation implies that for some $y \in [y_1, y_2]$, and hence $x \in C_{r_1}$,

$$\Re\left(i\lambda_2 \frac{x}{Y_0(x) - x} P_1(x)\right) = \Re\left(i \frac{x}{x - Y_0(x)} \lambda_1 P_2^-(Y_0(x))\right), \tag{5.13}$$

where we have used the fact that if $x = X_0(y + 0i)$ with $y \in [y_1, y_2]$ then $y = Y_0(x)$.

Using the results recalled in Section 5.4.1, the solution to the Riemann-Hilbert problem (5.13) then reads for $x \in D_{r_1}$

$$P_1(x) = \frac{\phi_1(x)}{\lambda_2 \pi} \int_{C_{r_1}} \frac{\Im(h_1(z))}{g_1(z)\phi_1^i(z)} \frac{\mathrm{d}z}{z - x} + \phi_1(x) P(0,0),$$

where

$$\begin{aligned} h_1(x) &\overset{\text{def.}}{=} \frac{x}{x - Y_0(x)} \lambda_1 P_2^-(Y_0(x)), \\ g_1(x) &\overset{\text{def.}}{=} \frac{x}{Y_0(x) - x}, \end{aligned}$$

and the function $\phi_1(x)$ is solution to the homogeneous Riemann-Hilbert problem (5.8).

Using the fact that $h_1(z) = -g_1(z)\lambda_1 P_2^-(Y_0(z))$, we deduce for $z = X_0(y + 0i)$

$$\frac{\Im(h_1(z))}{g_1(z)\phi_1^i(z)} = -\sin(\Theta_1(y)) e^{-\Phi_1(y)} \lambda_1 P_2^-(y).$$

Hence,

$$\frac{1}{\lambda_2 \pi} \int_{C_{r_1}} \frac{\Im(h_1(z))}{g_1(z)\phi_1^i(z)} \frac{\mathrm{d}z}{z - x} = \frac{x\lambda_1}{\lambda_2 \pi} \int_{y_1}^{y_2} \frac{(\lambda_2 - \mu_2 c_2 y^2) P_2^-(y)}{y K(x,y)} \sin(\Theta_1(y)) e^{-\Phi_1(y)} \mathrm{d}y$$

and Equation (5.12) follows. □

5.4.2 Function $P_2(y)$

We now establish a relation between the function $P_2(y)$, the function $P_1(x)$ and the polynomial $P_2^-(y)$. We use the same technique as for function $P_1(x)$.

Proposition 5.3. *The function $P_2(y)$ is related to polynomial $P_2^-(y)$ as*

$$\begin{aligned} P_2(y) = \frac{y\lambda_2}{2\pi\lambda_1} \int_{x_1}^{x_2} \frac{P_1(x)(\lambda_1 - \mu_1 c_1 x^2)\sqrt{-\Delta_2(x)}}{x(\lambda_1 + \lambda_2 - (\mu_1 c_1 + \mu_2 c_2)x) K(x,y)} \mathrm{d}x \\ + \frac{y}{2\pi i} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z)^2 P_2^-(z)}{z^2 (z - X_0(z)) K(X_0(z), y)} \mathrm{d}z + P(0,0). \end{aligned} \tag{5.14}$$

Proof. From Equation (5.4), we have for $y = Y_0(x)$

$$\lambda_2 \frac{1 - Y_0(x)}{Y_0(x) - x} x P_1(x) + \lambda_1 P_2(Y_0(x)) + \lambda_2 \frac{1 - Y_0(x)}{Y_0(x) - x} x P_2^-(Y_0(x)) = 0.$$

When $x \in [x_1, x_2]$, $Y_0(x) \in C_{r_2}$ and it then follows that the function $P_2(y)$ satisfies for $y \in C_{r_2}$

$$\Re\left(i\lambda_1 P_2(y)\right) = \Re\left(i h_2(y)\right),$$

where

$$h_2(y) \stackrel{\text{def.}}{=} \lambda_2 \frac{(y-1)X_0(y)}{y - X_0(x)} P_1(X_0(y)) + \lambda_1 \frac{(y-1)X_0(y)}{y - X_0(y)} P_2^-(y).$$

Using the results recalled in Section 5.4.1, the function $P_2(y)$ is given by

$$P_2(y) = \frac{1}{\lambda_1 \pi} \int_{C_{r_2}} \Im(h_2(z)) \frac{\mathrm{d}z}{z - y} + P(0,0).$$

(Note that we have in the present case to deal with a Dirichlet problem.)

Simple computations show that for $y = Y_0(x + 0i)$

$$\Im\left(\lambda_2 \frac{(y-1)X_0(y)}{y - X_0(x)} P_1(X_0(y))\right) = -\frac{\lambda_2 P_1(x)}{2(\lambda_1 + \lambda_2 - (\mu_1 c_1 + \mu_2 c_2)x)} \sqrt{-\Delta_2(x)}.$$

In addition, using the fact that the polynomial $P_2^-(y)$ is with real coefficients, we have

$$\Im\left(\lambda_1 \frac{y-1}{y - X_0(y)} X_0(y) P_2^-(y)\right) = \frac{\lambda_1}{2\pi i}\left(\frac{y-1}{y - X_0(y)} X_0(y) P_2^-(y) - \frac{\bar{y}-1}{\bar{y} - X_0(y)} X_0(y) P_2^-(\bar{y})\right).$$

Since

$$\frac{\mathrm{d}Y_0(x + 0i)}{\mathrm{d}x} = \frac{Y_0(x + 0i)\left(\frac{\lambda_1}{x} - \mu_1 c_1 x\right)}{-i\sqrt{-\Delta_2(x)}},$$

we have

$$\int_{C_{r_2}} \Im\left(\lambda_2 \frac{(y-1)X_0(y)}{y - X_0(x)} P_1(X_0(y))\right) \frac{\mathrm{d}z}{z - y} = \int_{x_1}^{x_2} \frac{\lambda_2 y P_1(x)(\lambda_1 - \mu_1 c_1 x^2)\sqrt{-\Delta_2(x)}}{2x(\lambda_1 + \lambda_2 - (\mu_1 c_1 + \mu_2 c_2)x)K(x,y)} \mathrm{d}x.$$

Moreover,

$$\begin{aligned}
&\frac{1}{\pi} \int_{C_{r_2}} \Im\left(\lambda_1 \frac{z-1}{z - X_0(z)} X_0(z) P_2^-(z)\right) \frac{\mathrm{d}z}{z - y} \\
&\quad = \frac{y\lambda_1}{2\pi i} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z) P_2^-(z)}{z(z - X_0(z))(z - y)(\lambda_2 - \mu_2 c_2 y z)} \mathrm{d}z \\
&\quad = \frac{y\lambda_1}{2\pi i} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z)^2 P_2^-(z)}{z^2(z - X_0(z))K(X_0(z), y)} \mathrm{d}z.
\end{aligned}$$

Assembling the two above relations, Equation (5.14) follows. □

5.4.3 Determination of the polynomial $P_2^-(y)$

We use the two previous results to establish a linear system satisfied by the coefficients of the polynomial $P_2^-(y)$. Let us first introduce some notation

Set

$$x(\theta) \overset{\text{def.}}{=} \frac{\lambda_1+\lambda_2+\mu_1c_1+\mu_2c_2-2\sqrt{\mu_2c_2\lambda_2}\cos\theta-\sqrt{\delta_1(\theta)}}{2\mu_1c_1}$$

with

$$\delta_1(\theta) \overset{\text{def.}}{=} (\lambda_1+\lambda_2+\mu_2c_2+\mu_1c_1-2\sqrt{\mu_2c_2\lambda_2}\cos\theta)^2-4\mu_1c_1\lambda_1,$$

so that

$$\cos\theta=-\frac{\mu_1c_1x(\theta)^2-(\lambda_1+\lambda_2+\mu_1c_1+\mu_2c_2)x(\theta)+\lambda_1}{2\sqrt{\mu_2c_2\lambda_2}x(\theta)}.$$

Moreover, define the coefficients for $0\le n,k\le a$

$$\alpha_{n,k}^{(1)} \overset{\text{def.}}{=} \frac{2\mu_2c_2}{\pi^2}\int_0^\pi \frac{x(\theta)\phi_1(x(\theta))}{\lambda_1+\lambda_2-(\mu_1c_1+\mu_2c_2)x(\theta)}J_k(\theta)\sin((n+1)\theta)\sin\theta\mathrm{d}\theta, \tag{5.15}$$

$$\alpha_{n,k}^{(2)} \overset{\text{def.}}{=} \frac{2r_2^{k-1}\mu_2c_2}{\pi}\int_0^\pi x(\theta)j_k(\theta)\sin((n+1)\theta)\mathrm{d}\theta, \tag{5.16}$$

$$\beta_n \overset{\text{def.}}{=} \frac{2\mu_2c_2\lambda_2}{\lambda_1\pi}\int_0^\pi \frac{\phi_1(x(\theta))x(\theta)}{\lambda_1+\lambda_2-(\mu_1c_1+\mu_2c_2)x(\theta)}\sin((n+1)\theta)\sin\theta\mathrm{d}\theta, \tag{5.17}$$

where

$$J_k(\theta) \overset{\text{def.}}{=} \int_{y_1}^{y_2}\frac{(\lambda_2-\mu_2c_2y^2)y^{k-1}}{\mu_2c_2y^2-2y\sqrt{\mu_2c_2\lambda_2}\cos\theta+\lambda_2}\sin(\Theta_1(y))e^{-\Phi_1(y)}\mathrm{d}y,$$

$$j_k(\theta) \overset{\text{def.}}{=} \frac{(r_2^2+x(\theta))\sin(k\theta)-x(\theta)r_2\sin((k+1)\theta)-r_2\sin((k-1)\theta)}{(1-x(\theta))(\lambda_1+\lambda_2-(\mu_1c_1+\mu_2c_2)x(\theta))}.$$

Theorem 5.1. *The probabilities $p(0,n)$ for $n=1,\dots,a$ can be expressed as a function of $p(0,0)$ by solving the linear system: for $0\le n\le a-1$*

$$r_2^np(0,n+1)=\beta_np(0,0)+\sum_{k=0}^{a}\alpha_{n,k}p(0,k) \tag{5.18}$$

where β_n is given by Equation (5.17) and $\alpha_{n,k}=\alpha_{n,k}^{(1)}+\alpha_{n,k}^{(2)}$ with $\alpha_{n,k}^{(1)}$ and $\alpha_{n,k}^{(2)}$ being defined by Equation (5.15) and (5.16), respectively. The probability $p(0,0)$ is determined by plugging the solution the linear system into equation (5.14), and such that it satisfies the normalisation condition (5.6).

Proof. Using the fact that

$$\frac{1}{K(x,y)}=\frac{1}{\lambda_2x}\sum_{n=0}^{\infty}U_n\left(\cos\theta_2(x)\right)\left(\frac{y}{r_2}\right)^n,$$

where $U_n(x)$ is the n-th Chebyshev polynomials of the second kind [AS72], $\theta_2(x) \in [0, \pi]$ is such that for $x \in [x_1, x_2]$

$$\cos\theta_2(x) = -\frac{\mu_1 c_1 x^2 - (\lambda_1 + \lambda_2 + \mu_1 c_1 + \mu_2 c_2)x + \lambda_1}{2x\sqrt{\mu_2 c_2 \lambda_2}},$$
$$\sin\theta_2(x) = \frac{\sqrt{-\Delta_2(x)}}{2x\sqrt{\mu_2 c_2 \lambda_2}}$$

and then $Y_0(x + 0i) = r_2 e^{-i\theta_2(x)}$, we deduce from Equation (5.14) that

$$r_2^n p(0, n+1) = + \frac{1}{2\pi i \lambda_2} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z) P_2^-(z)}{z^2(z - X_0(z))} U_n\left(\cos\theta_2(X_0(z))\right) \mathrm{d}z + \frac{1}{2\pi\lambda_1} \int_{x_1}^{x_2} \frac{P_1(x)(\lambda_1 - \mu_1 c_1 x^2)\sqrt{-\Delta_2(x)}}{x^2(\lambda_1 + \lambda_2 - (\mu_1 c_1 + \mu_2 c_2)x)} U_n\left(\cos\theta_2(x)\right) \mathrm{d}x.$$

We have

$$\frac{1}{2\pi i \lambda_2} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z) P_2^-(z)}{z^2(z - X_0(z))} U_n\left(\cos\theta_2(X_0(z))\right) \mathrm{d}z = \sum_{k=0}^{a} p(0,k) \frac{1}{2\pi i \lambda_2} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z) z^k}{z^2(z - X_0(z))} U_n\left(\cos\theta_2(X_0(z))\right) \mathrm{d}z.$$

For the integrals appearing in the above equation, we note that

$$\frac{1}{2\pi i \lambda_2} \int_{C_{r_2}} \frac{(z-1)(\lambda_2 - \mu_2 c_2 z^2) X_0(z) z^k}{z^2(z - X_0(z))} U_n\left(\cos\theta_2(X_0(z))\right) \mathrm{d}z = - \frac{1}{\pi r_2} \int_{C_{r_2}} \frac{(z(\theta) - 1) X_0(z(\theta))}{z(\theta) - X_0(z(\theta))} z(\theta)^{k-1} \sin((n+1)\theta) \mathrm{d}\theta,$$

where $z(\theta) \stackrel{\text{def.}}{=} r_2 e^{i\theta}$ and where we have used the fact that

$$U_n(\cos\theta) = \frac{\sin((n+1)\theta)}{\sin\theta}.$$

Simple computations show that the integral in the right hand side of the above equation is equal to $\alpha_{n,k}^{(2)}$ defined by Equation (5.16).

Using the fact that

$$2\sqrt{\mu_2 c_2 \lambda_2} \sin\theta_2(x) \frac{\mathrm{d}\theta_2(x)}{\mathrm{d}x} = \frac{\mu_1 c_1 x^2 - \lambda_1}{x^2}$$

we easily obtain that

$$\frac{1}{2\pi\lambda_1} \int_{x_1}^{x_2} \frac{P_1(x)(\lambda_1 - \mu_1 c_1 x^2)\sqrt{-\Delta_2(x)}}{x^2(\lambda_1 + \lambda_2 - (\mu_1 c_1 + \mu_2 c_2)x)} U_n\left(\cos\theta_2(x)\right) \mathrm{d}x = \frac{2\mu_2 c_2 \lambda_2}{\pi\lambda_1} \int_0^{\pi} \frac{P_1(x(\theta)) x(\theta)}{\lambda_1 + \lambda_2 - (\mu_1 c_1 + \mu_2 c_2) x(\theta)} \sin((n+1)\theta) \sin\theta \mathrm{d}\theta$$

Simple manipulations show that the above integral can be expressed as

$$\beta_n p(0,0) + \sum_{k=0}^{a} \alpha_{n,k}^{(1)} p(0,k)$$

with $\alpha_{n,k}^{(1)}$ and β_n defined by Equations (5.15) and (5.17), respectively. We finally obtain the linear system (5.18). □

In the next section, we show how the above theorem can be used to carry out numerical experiments.

5.5 Numerical experiments

We report in this section some numerical results in order to illustrate the computations performed in the previous sections. We analyse the implementation of a systematic offloading scheme in the framework of fog computing. Data centre 1 forwards all the requests which cannot be hosted locally to data centre 2 which, in order to protect its own requests, only accepts offloaded requests if there are a sufficient large number of free cores. We assess the potential gains on such policy using the blocking probabilities of customers at each data centre. In the following, B_i denotes the blocking rate of customers assigned to data centre i, for $1 \leq i \leq 2$.

In all cases described below, it is considered that data centre 1 is saturated by its demands, i.e., $\lambda_1 > \mu_1 c_1$, and data centre 2 is saturated by the combined flows of requests after this policy is implemented, i.e., $\lambda_2 + \lambda_1 - \mu_1 c_1 > \mu_2 c_2$ (Condition (5.2)). The analysis is decomposed into the following cases:

i. If $a \in \mathbb{N}^*$, the loss probabilities are calculated using Theorem 5.1. For this purpose, we first numerically compute the coefficients $\alpha_{n,k}^{(1)}$ and $\alpha_{n,k}^{(2)}$ for $0 \leq n, k \leq a$ and β_n for $0 \leq n \leq a$ by taking $p(0,0) = 1$ via Equations (5.15), (5.16), and (5.17), respectively. We thus obtain the unnormalised probabilities $p(0,k)$ and then polynomial $P_2^-(y)$ up to the normalising coefficient $p(0,0) = P(0,0)$. This determines the function $P_1(x)$ up to the proportional coefficient $p(0,0)$ by using Equation (5.12) and subsequently the function $P_2(x)$ up to the coefficient $p(0,0)$. This coefficient is eventually computed by using the normalising condition (5.6). This allows us to numerically compute $P_2^-(1)$ and $P_1(1)$, equal to B_1 and B_2, respectively.

ii. When a is very large the requests arriving at the second data centre are not influenced by those offloaded from first data centre. In this case, the blocking probabilities, denoted by $B^\infty = (B_i^\infty,\ 1 \leq i \leq 2)$, are simply given by

$$\begin{cases} B_1^\infty \overset{\text{def.}}{=} \dfrac{\lambda_1 - \mu_1 c_1 - (\mu_2 c_2 - \lambda_2)^+}{\lambda_1} \\ B_2^\infty \overset{\text{def.}}{=} \left(\dfrac{\lambda_2 - \mu_2 c_2}{\lambda_2}\right)^+, \end{cases} \tag{5.19}$$

where we have used the global rate conservation law given by Equation (5.6).

iii. If $a = 0$, there is no protection for the requests which originally arrive at data centre 2, i.e., every request rejected at data centre 1 is systematically forwarded to data centre 2 and accepted if there are idle cores in this data centre. This scenario is a particular case of the offloading scheme studied in Fricker *et al.* [FGRT16a], where two parallel data centres forward demands they cannot hold to each other with a given probability. Namely, in the notation used in that paper, we have $p_1 = 1$ and $p_2 = 0$, where p_i is probability that data centre i forwards a request to data centre $3 - i$, for $1 \leq i \leq 2$. Using Theorem 3 from this reference, we are able to calculate the blocking probabilities, denoted by $B^0 = (B_i^0,\, 1 \leq i \leq 2)$, such that

$$\begin{cases} B_1^0 \overset{\text{def.}}{=} \dfrac{\lambda_1 - \mu_1 c_1 - (\mu_2 c_2 - \lambda_2)^+}{\lambda_1 \phi_2(1)}, \\ B_2^0 \overset{\text{def.}}{=} \dfrac{\lambda_1 (1 - B_1^0) - \mu_1 c_1 + \lambda_2 - \mu_2 c_2}{\lambda_2}, \end{cases} \tag{5.20}$$

where

$$\phi_2(y) \overset{\text{def.}}{=} \exp\left(\frac{y}{\pi} \int_{x_1}^{x_2} \frac{\lambda_1 - \mu_1 c_1 x^2}{x K(x,y)} \Theta_2(x) \mathrm{d}x \right)$$

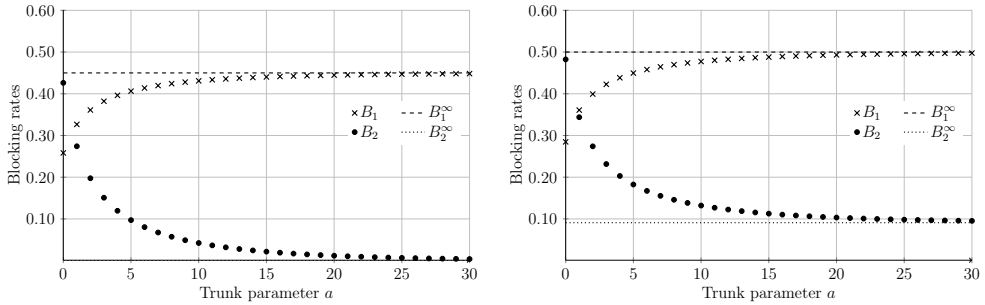
and

$$\Theta_2(x) \overset{\text{def.}}{=} \operatorname{ArcTan}\left(\frac{\sqrt{-\Delta_2(x)}}{(x+1)(\lambda_1 - x\mu_1 c_1) + x(\lambda_2 - \mu_2 c_2)} \right).$$

To evaluate the impact of the choice of the parameter a on the loss probabilities at both data centres, we compare the blocking rates in this system with the intuitive boundaries B^0 and B^∞ defined previously in Equations (5.19) and (5.20), respectively. For this purpose, we consider the case where requests are accommodated by a edge data centre possibly offloaded to the other data centre; the trunk reservation policy is implemented in the second data centre.

We first consider the case when both data centres have the same capacity, say, $c = (1.0, 1.0)$ and the edge data centre is saturated while the other is underloaded but the global system is overloaded. To illustrate this case, we have chosen the parameters $\lambda = (2.0, 0.9)$ and $\mu = (1.0, 1.0)$. Blocking rates are depicted in Figure 5.3a. We see that small values of a (say, $a = 1, 2, 3$) can help reduce the blocking rate of its original requests while moderately increasing that of the offloaded ones. Choosing large values of a does not improve blocking rates. The same conclusion holds when both data centres are overloaded as shown by Figure 5.3b.

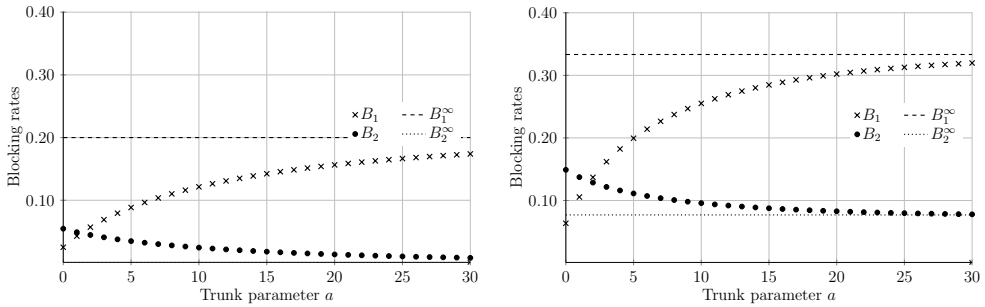
To further investigate the performance of the trunk reservation policy, we have considered asymmetric data centres, the data centre 2 is larger than the data centre 1. We have specifically chosen $c = (1., 5.)$ and still with the same mean service rates $\mu = (1., 1.)$. In the first case, $\lambda = (1.5, 4.8)$ so that the second data centre is underloaded. Taking moderate values of a (say, $a = 3, \ldots, 10$) can decrease the blocking rate at data centre 1 without



(a) $\lambda_1 = 2.0, \lambda_2 = 0.9$. (b) $\lambda_1 = 2.0, \lambda_2 = 1.1$.

Figure 5.3 – Blocking probabilities in both data centres with same size as a function of a, with $c = (1.0, 1.0)$ and $\mu = (1.0, 1.0)$

impacting too much the performance of data centre 2, as shown in Figure 5.4a. Smaller values of a (say, $a = 1, \ldots, 5$) are better when the second data centre is overloaded as shown in Figure 5.4b.



(a) $\lambda_1 = 1.5, \lambda_2 = 4.8$. (b) $\lambda_1 = 1.5, \lambda_2 = 6.0$.

Figure 5.4 – Blocking probabilities in both data centres with different sizes as a function of a, with $c = (1.0, 5.0)$ and $\mu = (1.0, 1.0)$

5.6 Conclusion

We have considered in this paper a trunk reservation policy implemented in a data centre accommodating external requests but also some requests from another data centre when saturated. This model reasonably illustrates what happens in a distributed data centre system. We consider in this paper a system under saturation to stress the performance of the trunk reservation policy.

The offload mechanism introduces a coupling between both data centres. By rescaling the system under an assumption of congestion, we are led to study a random walk in the quarter plane, which has the property of having non constant reflecting conditions on one axis. We have developed an original method of determining the unknown functions appearing when computing the generating function of the joint numbers of active jobs in both data centres and subsequently the blocking rates of requests. Numerical results show that

the trunk reservation policy proves efficient for reducing the blocking rates at the edge data centre while limiting the impact of the offload mechanism on the rejection of requests at the second data centre. An algorithm for dynamically adjusting the trunk reservation parameter a could be envisaged in practice by using real time measurement information.

Bibliography

[AAS+15] J. Añorga, S. Arrizabalaga, B. Sedano, M. Alonso-Arce, and J. Mendizabal. YouTube's DASHimplementation analysis. In *Proceedings of the 19th International Conference on Communications (part of CSCC'15)*, volume 50 of *Recent Advances in Electrical Engineering Series*, pages 61–66, Zakynthos, Ionian Islands, Greece, 07 2015. 31, 53

[AH97] M. Alanyali and B. Hajek. Analysis of simple algorithms for dynamic load balancing. *Mathematics of Operations Research*, 22(4):840–871, 1997. 90

[AIM10] L. Atzori, A. Iera, and G. Morabito. *Computer Networks*, 54(15):2787–2805, 10 2010. 5

[Ama] Amazon.com. Amazon Web Services. http://aws.amazon.com/. 2, 3

[Arn92] V. I. Arnol'd. *Ordinary differential equations*. Springer-Verlag, Berlin, Germany, third edition, 1992. Translated from the third Russian edition by Roger Cooke. 43

[AS72] M. Abramowitz and I. Stegun. *Handbook of mathematical functions*. National Bureau of Standards, Applied Mathematics Series 55, 1972. 119

[Asm03] S. Asmussen. *Applied Probability and Queues*, volume 51 of *Applications of Mathematics*. Springer-Verlag, New York, USA, second edition, 2003. 11, 15, 47, 82

[ASZ+10] M. Armbrust, I. Stoica, M. Zaharia, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, and A. Rabkin. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 04 2010. 1, 2

[BGZ95] N. G. Bean, R. J. Gibbens, and S. Zachary. Asymptotic analysis of single resource loss systems in heavy traffic, with applications to integrated networks. *Advances in Applied Probability*, 27(1):273–292, 03 1995. 20, 33, 34, 47, 64

[BGZ97] N. G. Bean, R. J. Gibbens, and S. Zachary. Dynamic and equilibrium behavior of controlled loss networks. *The Annals of Applied Probability*, 7(4):873–885, 11 1997. 33, 64

[BHJ48] E. Brockmeyer, H. L. Halstrøm, and A. Jensen. *The life and works of A. K. Erlang*. Transactions of the Danish Academy

of Technical Sciences (Akademiet for de Tekniske Videnskaber, Denmark). Akademiet for de Tekniske Videnskaber, first edition, 1948. 10, 13

[BLL84] D. Y. Burman, J. P. Lehoczky, and Y. Lim. Insensitivity of blocking probabilities in a circuit-switching network. *Journal of Applied Probability*, 21(4):850–859, 12 1984. 13

[BMZA12] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. Fog computing and its role in the Internet of Things. In *Proceedings of the First Edition of the MCCWorkshop on Mobile Cloud Computing*, MCC '12, pages 13–16, New York, NY, USA, 08 2012. ACM. 5, 85, 105

[BV05] T. Bonald and J. Virtamo. A recursive formula for multirate systems with elastic traffic. *IEEE Communications Letters*, 9(8):753–755, 08 2005. 14

[Car50] H. Cartan. *Elementary theory of one or several complex variables*. Dover Publications, 1950. 97

[CB84] J. W. Cohen and O. J. Boxma. *Boundary value problems in queueing system analysis*. Mathematics Studies, North Holland, 1984. 15

[CB85] J. W. Cohen and O. J. Boxma. A survey of the evolution of queueing theory. *Statistica Neerlandica*, 39(2):143–158, 06 1985. 10

[CKPT16] H. I. Christensen, A. Khan, S. Pokutta, and P. Tetali. Multidimensional bin packing and other related problems: A survey, 2016. 6

[CLW95] G. L. Choudhury, K. K. Leung, and W. Whitt. An algorithm to compute blocking probabilities in multi-rate multiclass multi-resource loss models. *Advances in Applied Probability*, 27(4):1104–âĂŞ1143, 12 1995. 13

[DL90] R. Dautray and J. L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 4 Integral Equations and Numerical Methods*, volume 4. Springer Berlin Heidelberg, Berlin, Germany, 1990. 98, 99, 113, 115

[DR87] Z. Dziong and J. W. Roberts. Congestion probabilities in a circuit-switched integrated services network. *Performance Evaluation*, 7(4):267–284, 11 1987. 13

[DS11] D. P. Doane and L. E. Seward. Measuring skewness: A forgotten statistic? *Journal of Statistics Education*, 19(2):1–18, 2011. 50

[EhBN17] E. El-hady, J. Brzdęk, and H. Nassar. On the structure and solutions of functional equations arising from queueing models. *Aequationes mathematicae*, 91(3):445–477, 06 2017. 15

[Erl09] A. K. Erlang. The theory of probabilities and telephone conversations (sandsynlighedsregning og telefonsamtaler). *Nyt Tidsskrift for Matematik*, 20:33–39, 1909. 10

[Erl17] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineers' Journal*, 10:189–197, 1917. 10

[FGRT16a] C. Fricker, F. Guillemin, Ph. Robert, and G. Thompson. Analysis of an offloading scheme for data centers in the framework of fog computing. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (ToMPECS)*, 1(4):16:1–16:18, 09 2016. 33, 58, 64, 105, 106, 121

[FGRT16b] C. Fricker, F. Guillemin, Ph. Robert, and G. Thompson. Analysis of downgrading for resource allocation. *ACM SIGMETRICS Performance Evaluation Review*, 44(2):24–26, 09 2016. 31

[FGRT17] C. Fricker, F. Guillemin, Ph. Robert, and G. Thompson. Allocation schemes of resources with downgrading. *Advances in Applied Probability*, 49(2):629–651, 04 2017. 82

[FI79] G. Fayolle and R. Iasnogorodski. The two coupled server: the reduction to a Riemann-Hilbert problem. *Zur Wahrscheinlichkeitstheorie und verwandte Gebiete*, B47(3):325–351, 01 1979. 15, 87, 88, 89, 94, 95, 97, 98, 106, 111

[FIM17] G. Fayolle, R. Iasnogorodski, and V. A. Malyshev. *Random Walks in the Quarter-Plane. Algebraic Methods, Boundary Value Problems and Applications*, volume 40 of *Applications of Mathematics*. Springer-Verlag, second edition, 2017. 15, 95, 96, 106

[FMM95] G. Fayolle, V. A. Malyshev, and M. V. Menshikov. *Topics in the constructive theory of countable Markov chains*. Cambridge University Press, Cambridge, EN, UK, 1995. 15, 109

[Fre74] G. Frenkel. The grade of service in multiple-access satellite communications systems with demand assignments. *IEEE Transactions on Communications*, 22(10):1681–1685, 10 1974. 13

[FRT01] C. Fricker, Ph. Robert, and D Tibi. On the fluid limits of some loss networks. Research Report RR-4171, INRIA, 04 2001. 20, 64

[FRT03] C. Fricker, Ph. Robert, and D. Tibi. A degenerate central limit theorem for single resource loss systems. *The Annals of Applied Probability*, 13(2):561–575, 05 2003. 20

[FS17] S. Foss and A. L. Stolyar. Large-scale join-idle-queue system with general service times. Preprint, 02 2017. 7

[FZRL08] I. Foster, Y. Zhao, I. Raicu, and S. Lu. Cloud computing and grid computing 360-degree compared. In *Grid Computing Environments Workshop, 2008. GCE'08*, pages 1–10. IEEE, 2008. 1, 2

[Gak90] F. D. Gakhov. *Boundary value problems*. Dover Publications Inc., New York, USA, 1990. Translated from the Russian, Reprint of the 1966 translation. 47

[Gee09] J. Geelan. Twenty-one experts define cloud computing. http://cloudcomputing.sys-con.com/node/612375/, 01 2009. 1

[GH83] J. Guckenheimer and P. J. Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42 of *Applied Mathematical Sciences*. Springer-Verlag, New York, USA, first edition, 1983. 21

[GHM13] F. Guillemin, T. Houdoin, and S Moteau. Volatility of YouTubecontent in Orangenetworks and consequences. In *2013 IEEEInternational Conference on Communications (ICC)*, pages 2381–2385. IEEE, 06 2013. 31

[GI99] A. Goel and P. Indyk. Stochastic load balancing and related problems. In *40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039)*, pages 579–586, 1999. 6

[GKMS13] F. Guillemin, B. Kauffmann, S. Moteau, and A Simonian. Experimental analysis of caching efficiency for YouTubetraffic in an ISPnetwork. In *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, pages 1–9. IEEE, 09 2013. 31

[Goo] Google Inc. Google Cloud Platform. http://cloud.google.com/. 2, 3, 58

[GSZS12] A. Ghodsi, V. Sekar, M. Zaharia, and I. Stoica. Multi-resource fair queueing for packet processing. In *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '12, pages 1–12, New York, NY, USA, 2012. ACM. 58

[GT16] F. Guillemin and G. Thompson. Analysis of a trunk reservation policy in the framework of fog computing. Preprint, 04 2016. 28, 58, 64

[GZH+11] A. Ghodsi, M. Zaharia, B Hindman, A. Konwinski, S. Shenker, and I Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In *Proceedings of the 8th USENIX-Conference on Networked Systems Design and Implementation*, NSDI'11, pages 323–336, Berkeley, CA, USA, 03 2011. USENIX Association. 58

[HK94] P. J. Hunt and T. G. Kurtz. Large loss networks. *Stochastic Processes and their Applications*, 53(2):363–378, 10 1994. 15, 20, 21, 32, 39, 61, 64, 69, 75, 78, 87, 90, 109

[JLS+08] K. Jung, Y. Lu, D. Shah, M. Sharma, and M. S. Squillante. Revisiting stochastic loss networks: Structures and algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 36(1):407–418, 06 2008. 14

[JS03] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg, second edition, 2003. 109

[JS15] B. Jennings and R. Stadler. Resource management in clouds: Survey and research challenges. *Journal of Network and Systems Management*, 23(3):567–619, 07 2015. 6

[Kel86] F. P. Kelly. Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability*, 18(2):473–505, 06 1986. 13, 17, 19, 20, 32, 33, 35, 52, 60, 64

[Kel91] F. P. Kelly. Loss networks. *The Annals of Applied Probability*, 1(3):319–378, 08 1991. 13, 17, 19, 20, 32, 34, 35, 52, 58, 59, 60, 61, 63, 64, 86

[Ken53] D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, 09 1953. 11

[Kin61] J. F. C. Kingman. The ergodic behaviour of random walks. *Biometrika*, 48(3/4):391–396, 12 1961. 70

[Kin09] J. F. C. Kingman. The first erlang century — and the next. *Queueing Systems: Theory and Applications*, 63(1):3–12, 11 2009. 10

[KMM12] H. Khazaei, J. Misic, and V. B. Misic. Performance analysis of cloud computing centers using $M/G/m/m+r$ queuing systems. *IEEE Transactions on Parallel and Distributed Systems*, 23(5):936–943, 05 2012. 6

[LMK94] G. Louth, M. Mitzenmacher, and F. P. Kelly. Computational complexity of loss networks. *Theoretical Computer Science*, 125(1):45–59, 1994. 13

[LXK+11] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg. Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68(11):1056–1071, 11 2011. 7

[Mal93] V. A. Malyshev. Networks and dynamical systems. *Advances in Applied Probability*, 25(1):140–175, 03 1993. 70

[MG11] P. Mell and T. Grance. The NIST definition of cloud computing recommendations of the national institute of standards and technology. *Nist Special Publication*, 145:7, 09 2011. 1, 2

[Mic] Microsoft. Azure. http://azure.microsoft.com/. 2, 3, 7, 58, 85

[Mit96] M. Mitzenmacher. Load balancing and density dependent jump Markov processes. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 213–222, 10 1996. 7

[MKMG15] A. Mukhopadhyay, A. Karthik, R. R. Mazumdar, and F. Guillemin. Mean field and propagation of chaos in multi-class heterogeneous loss models. *Performance Evaluation*, 91:117–131, 2015. Special Issue: Performance 2015. 8

[MNA16] M. Masdari, S. S. Nabavi, and V. Ahmadi. An overview of virtual machine placement schemes in cloud computing. *Journal of Network and Computer Applications*, 66:106–127, 2016. 6

[MSY12] S. T. Maguluri, R. Srikant, and L. Ying. Stochastic models of load balancing and scheduling in cloud computing clusters. In *2012 Proceedings IEEE INFOCOM*, pages 702–710. IEEE, 03 2012. 7

[New10] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, first edition, 2010. 10

[RBG12] A. Rai, R. Bhagwan, and S. Guha. Generalized resource allocation for the cloud. In *Proceedings of the Third ACM Symposium on Cloud Computing*, SoCC '12, pages 15:1–15:12, New York, NY, USA, 2012. ACM. 5, 85, 105

[Rob81] J. W. Roberts. A service system with heterogeneous user requirements - application to multi-service telecommunications systems. In G. Pujolle, editor, *Performance of Data Communication Systems and their Applications*, pages 423–431. North-Holland Publishing Company, 1981. 13

[Rob03] Ph. Robert. *Stochastic Networks and Queues*, volume 52 of *Applications of Mathematics*. Springer-Verlag, New York, USA, first edition, 2003. 15, 35, 36, 38, 41, 47, 54, 59, 70, 72, 82, 89, 91, 111

[Roo09] K. Roos. *Farkas lemma*, pages 995–998. Springer US, Boston, MA, USA, 2009. 65

[Ros95] K. W. Ross. *Multiservice loss models for broadband telecommunication networks*. Springer, 1995. 6, 106

[RS92] A. N. Rybko and A. L. Stolyar. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii*, 28(3):3–26, 07 1992. Translated from Russian. 16, 70

[RT89] K. W. Ross and D. H. K. Tsang. The stochastic knapsack problem. *IEEE Transactions on Communications*, 37(7):740–747, 07 1989. 6

[RT90] K. W. Ross and D. H. K. Tsang. Teletraffic engineering for product-form circuit-switched networks. *Advances in Applied Probability*, 22(3):657–675, 09 1990. 17

[Rud87] W. Rudin. *Real and complex analysis*. Mathematics Series. McGraw-Hill Book Co., New York, USA, third edition, 1987. 48

[RV15] Ph. Robert and A. Véber. A stochastic analysis of resource sharing with logarithmic weights. *The Annals of Applied Probability*, 25(5):2626–2670, 10 2015. 16

[RW00] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge University Press, second edition, 2000. 109

[SCAFV16] L. Sharifi, L. Cerdà-Alabern, F. Freitag, and L. Veiga. Energy efficient cloud service provisioning: Keeping data center gran-

ularity in perspective. *Journal of Grid Computing*, 14(2):299–325, 06 2016. 5

[SHZ+13] C. Sieber, T. Hoßfeld, T. Zinner, Ph. Tran-Gia, and C. Timmerer. Implementation and user-centric comparison of a novel adaptation logic for DASHwith SVC. In *IM'13*, pages 1318–1323. IEEE, 2013. 31

[SMW07] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H.264/AVCstandard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, 09 2007. 31

[Sto13] A. L. Stolyar. An infinite server system with general packing constraints. *Operations Research*, 61(5):1200–1217, 07 2013. 33

[Sto15] A. L. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems: Theory and Applications*, 80(4):341–361, 06 2015. 7, 33

[SZ13] A. L. Stolyar and Y. Zhong. A large-scale service system with packing constraints: Minimizing the number of occupied servers. *ACM SIGMETRICS Performance Evaluation Review*, 41(1):41–52, 06 2013. 7

[Tak69] L. Takács. On Erlang's formula. *The Annals of Mathematical Statistics*, 40(1):71–78, 02 1969. 12

[TLX12a] Y. Tan, Y. Lu, and C. H. Xia. Provisioning for large scale cloud computing services. *ACM SIGMETRICS Performance Evaluation Review*, 40(1):407–408, 06 2012. 14

[TLX12b] Y. Tan, Y. Lu, and C. H. Xia. Provisioning for large scale loss network systems with applications in cloud computing. *ACM SIGMETRICS Performance Evaluation Review*, 40(3):83–85, 01 2012. 14

[TSM98] F. Theberge, A. Simonian, and R. R. Mazumdar. Upper bounds for blocking probabilities in large multi-rate loss networks. *Telecommunication Systems*, 9(1):23–39, 1998. 13

[VCH+10] S. Vadlakonda, A. Chotai, B.D. Ha, A. Asthana, and S. Shaffer. System and method for dynamically upgrading / downgrading a conference session, 04 2010. 31

[VDK96] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996. Translated from Russian. 7

[VLM+09] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez. Greening the internet with nano data centers. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, CoNEXT '09, pages 37–48. ACM, ACM, 2009. 5

[VRMCL08] L. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner. A break in the clouds: Towards a cloud definition. *Computer Communication Review*, 39(1):50–55, 12 2008. 1

[VZ09] J. Voas and J. Zhang. Cloud Computing: New wine or just a new bottle? *IT Professional*, 11(2):15–17, 03 2009. 2

[WRSvdM11] T. Wood, K. K. Ramakrishnan, P. Shenoy, and J. van der Merwe. CloudNet: Dynamic pooling of cloud resources by live wan migration of virtual machines. *IEEE/ACM Transactions on Networking*, 46(7):121–132, 03 2011. 5, 85, 105

[XDLS15] Q. Xie, X. Dong, Y. Lu, and R. Srikant. Power of d choices for large-scale bin packing: A loss model. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):321–334, 06 2015. 8

[XSC13] Z. Xiao, W. Song, and Q. Chen. Dynamic resource allocation using virtual machines for cloud computing environment. *Parallel and Distributed Systems, IEEE Transactions on*, 24(6):1107–1117, 06 2013. 7

[YBDS08] L. Youseff, M. Butrico, and D. Da Silva. Toward a unified ontology of cloud computing. In *Grid Computing Environments Workshop, 2008. GCE'08*, pages 1–10. IEEE, 11 2008. 1

[YTDG09] B. Yang, F. Tan, Y.-S. Dai, and S. Guo. *Performance Evaluation of Cloud Service Considering Fault Recovery*, pages 571–576. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 6

[ZCB10] Q. Zhang, L. Cheng, and R. Boutaba. Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1):7–18, 05 2010. 1, 4

[ZZ02] S. Zachary and I. Ziedins. A refinement of the Hunt-Kurtz theory of large loss networks, with an application to virtual partitioning. *The Annals of Applied Probability*, 12(1):1–22, 02 2002. 33, 34

[ZZ11] S. Zachary and I. Ziedins. *Loss Networks*, volume 154 of *International Series in Operations Research & Management Science*, pages 701–728. Springer US, 2011. 33, 34, 35

Abstract

This PhD thesis investigates four problems in the context of Large Distributed Systems. This work is motivated by the questions arising with the expansion of Cloud Computing and related technologies (Fog Computing, VNF, etc.). The present work investigates the efficiency of different resource allocation algorithms in this framework. The methods used involve a mathematical analysis of several stochastic models associated to these networks.

Chapter 1 provides an introduction to the subject in general, as well as a presentation of the main mathematical tools used throughout the subsequent chapters.

Chapter 2 presents a congestion control mechanism in Video on Demand services delivering files encoded in various resolutions. We propose a policy under which the server delivers the video only at minimal bit rate when the occupancy rate of the server is above a certain threshold. The performance of the system under this policy is then evaluated based on both the rejection and degradation rates.

Chapters 3, 4 and 5 explore problems related to cooperation schemes between data centres on the edge of the network. In the first setting, we analyse an offloading policy in the context of multi-resource cloud services. In second case, requests that arrive at a congested data centre are forwarded to a neighbouring data centre with some given probability. In the third case, requests blocked at one data centre are forwarded systematically to another where a trunk reservation policy is introduced such that a redirected request is accepted only if there are a certain minimum number of free servers at this data centre.