Dynamics of a two-level system with priorities and application to an emergency call center

Vianney Boeuf

NNT : 2017SACLX120

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PARIS-SACLAY

préparée à

L'ÉCOLE POLYTECHNIQUE

ÉCOLE DOCTORALE N° 574
Mathématiques Hadamard (EDMH)

Spécialité de doctorat : Mathématiques

par

Vianney Bœuf

Dynamics of a two-level system with priorities and application to an emergency call center.

Thèse présentée et soutenue à Palaiseau, le 18 décembre 2017

Composition du jury :

Xavier Allamigeon	(Inria & CMAP, École polytechnique)	Co-directeur de thèse
Thomas Bonald	(Télécom ParisTech)	Rapporteur
Stéphane Gaubert	(Inria & CMAP, École polytechnique)	Directeur de thèse
Alessandro Giua	(Universités Aix-Marseille et Cagliari)	Rapporteur
Pierre L'Ecuyer	(Université de Montréal)	Rapporteur
Jean-Jacques Loiseau	(LS2N, université de Nantes)	Président
Philippe Robert	(INRIA de Paris)	Examinateur
Stéphane Raclot	(Préfecture de police de Paris)	Co-encadrant

« *Il n'y a que les routes qui sont belles...* »

À mes parents.

Pour l'esprit mathématique que vous m'avez légué ; pour mon goût mathématique que vous avez éveillé et fait s'épanouir.

Cette thèse vous doit beaucoup.

Remerciements

Je tiens à remercier en premier lieu mon directeur de thèse Stéphane Gaubert et mon co-directeur Xavier Allamigeon. Merci à tous deux pour m'avoir conduit sur ces belles routes scientifiques, et pour m'avoir beaucoup transmis. Merci à Stéphane pour ta science encyclopédique et ton fidèle soutien. Merci pour m'avoir accueilli jusqu'à chez toi lorsqu'il fallait traiter des urgences. J'associe naturellement Marianne à ces remerciements. Merci à Xavier pour la qualité de nos échanges, ta rigueur et la touche informatique dont cette thèse a beaucoup profité. Merci à vous deux pour tous ces repas si sympathiques, et la richesses de vos conversations.

Merci à Thomas Bonald, Alessandro Giua et Pierre L'Ecuyer, d'avoir bien voulu être rapporteurs de ma thèse. Merci pour leur lecture avisée de mon manuscrit. Je remercie aussi sincèrement l'ensemble des examinateurs d'avoir accepté, parfois en venant de très loin, de venir siéger à mon jury.

Cette thèse doit beaucoup à la Brigade des sapeurs-pompiers de Paris et à la Préfecture de police de Paris. Je remercie chaleureusement Stéphane Raclot, présent sur tous les sujets. Merci pour ton accueil à la Brigade. Merci pour ta capacité à interagir avec des scientifiques et à donner à tout cela un caractère opérationnel. Merci à Régis Reboul. Merci à tous les pompiers qui m'ont accueilli parmi eux, Frédéric Derreati, Pascal Dillenseger, David Vigier, Floriane Brill, Benjamin Fouilleul et l'ensemble de l'équipe SIOP et GGO. Merci au lieutenant-colonel Vilbé d'avoir particulièrement cherché à utiliser mon expertise (parfois sans doute surestimée !) avec une grande exigence, et pour votre soutien sur de nombreux sujets. Merci au BPO et notamment les lieutenants-colonels Deshayes et Cros. Merci aux capitaines Éric Faraon et Éric Gauyat. Merci à Philippe Boubel et François Botella. Merci à Carl de Barmon et Roman Lorencki. Merci aux médecins et aux juristes de la Brigade, avec qui j'ai toujours eu plaisir à échanger.

Mes remerciements s'étendent aux pompiers instructeurs qui m'ont formé aux premiers secours en équipe (et notamment le sergent-chef Lefèbvre), et aux pompiers du centre de secours de Montmartre qui m'ont accueilli pour des immersions en garde VSAV. Je salue aussi l'ensemble des camarades VSC de ma formation.

Les passages à l'équipe RAP ont toujours été des moments agréables et la personnalité et la disponibilité de Philippe Robert y sont pour beaucoup. Merci Philippe pour notre travail ensemble, merci pour m'avoir initié à la théorie de la mesure et à bien d'autres objets stochastiques, et pour les footings à 11h30 précises. Merci à Christine, Guilherme, Wen, Renaud, Sarah, Davit et les autres pour les agréables moments de détente, et pour ce séjour sympathique à Chicago.

Au CMAP, je veux remercier tout ceux à qui j'ai pu exposer des problèmes de mathématiques, et j'espère n'oublier personne en citant Marianne Akian, Vincent Bansaye, Yacine Chitour, Hadrien De March, Carl Graham, Benjamin Heymann, Ludovic Sacchelli, et, plus encore, Erwan Le Pennec, qui m'a apporté de précieuses informations sur le langage R et sur certaines méthodes statistiques. Je remercie aussi l'équipe administrative du CMAP et de l'INRIA, notamment Nassera Naar, Alexandra Noiret, Jessica Gameiro et Corinne Petitot. J'y associe mes remerciements à Sylvain Ferrand.

Je remercie Frédéric Meunier pour mon super stage de recherche opérationnelle sur le *train shunting*. Merci de m'avoir permis de croire que je pouvais non seulement chercher, mais aussi trouver. Merci aussi pour le TD d'optimisation aux Ponts, que j'ai eu grand plaisir à donner. Merci à Sébastien Blandin. Merci à Olivier Klopfenstein pour le projet à l'ENSTA.

Je remercie les chercheurs de l'ANR Democrite et notamment Emmanuel Lapébie. Je remercie aussi les huit élèves polytechniciens qui ont travaillé au cours de ces trois années sur la PFAU ou sur des aspects liés.

J'ose à peine citer au CMAP tous ceux avec qui j'ai eu plaisir à échanger, à passer des repas ou des pauses cafés. Je pense en premier lieu à mes co-bureau Etienne, Antoine, Gustaw, Aleksey, Florine, Jean-Bernard, Céline, Pierre. Merci particulièrement à mon « prédécesseur » Antoine et mon « successeur » Jean-Bernard : cela crée des liens particuliers ! Je salue aussi Joon, Perle, Hadrien, Aline, Romain, Antoine, Nikolas, Mateusz et encore de nombreux autres !

Mes camarades des autres labos, Lucile, Simon et Olga, méritent aussi de grands remerciements. Merci beaucoup à Lucile pour l'escalade.

Merci, dans un autre registre, à Monsieur Christian Makhmoufi pour avoir retrouvé mes affaires de thèse au fond d'un ruisseau à Beauvoir-en-Royans après un vol à la roulotte.

L'amitié se mesure entre autres par la capacité à reprendre avec passion la conversation après un intervalle parfois prolongé par la distance géographique, le travail ou les responsabilités diverses. Je ne prendrais évidemment pas le risque de citer tous les amis avec qui j'ai pu passer du temps pendant ces trois années, mais qu'ils soient ici chaleureusement remerciés. Je dois une dédicace toute particulière à mes quatre « chapeaux verts », Axel, Gaétan, Gaëtan et Matthieu qui savent toute l'affection que je leur porte. Pour Axel s'ajoutent des remerciements pour nos nombreuses discussions scientifiques et d'avenir. Je remercie aussi ici les Oyens pour Oya, et les Cramés, Matthieu, Étienne, Guillaume, Alexandre, Augustin, Maximilien, pour près de deux ans de coloc dans un super esprit.

Merci à ma famille, qui représente tant pour moi, et est un vrai ressourcement. Merci à Céline, Mathilde, Solène, à Éric et à mes nièces et neveu. Merci pour ces liens inconditionnels qui nous unissent. Mes remerciements du fond du cœur vont à mes parents pour tout ce que vous m'avez transmis et donné, depuis les premiers jours et bien sûr pendant ces trois dernières années. Cette thèse vous est dédiée.

Merci à ma deuxième famille depuis peu, et à mes nouveaux beaux-frères et belles-sœurs. Merci particulièrement à Père et Mère, pour m'avoir offert des conditions de travail épatantes dans des lieux magnifiques.

J'adresse enfin mes remerciements aux deux personnes qui comptent le plus dans ma vie : merci Sixtine d'avoir été un petit bébé si adorable et si sage pendant tes dix premiers mois. En plus de faire l'admiration de ton papa, tu as aussi grandement facilité l'avancement de ma thèse. Merci enfin, surtout, à mon épouse Isabelle. Merci pour tous tes sacrifices pendant la rédaction de ma thèse, tous ces jours de congés sans moi, ces soirées, ces weekends où tu devais, non seulement te passer de moi mais aussi t'occuper de Sixtine et de la maison. Cette fois-ci la thèse est bel et bien bouclée, quel soulagement ! Merci pour ton dévouement et ta sollicitude. Merci, plus fondamentalement, d'être une si belle raison d'être et d'aimer depuis plusieurs années.

Abstract

We analyze the dynamics of discrete event systems with synchronization and priorities, by means of Petri nets and queueing networks. We apply this to the performance evaluation of an emergency call center.

Our original motivation is practical. In 2016, a new emergency call center became operative in Paris area, handling emergency calls to police and firemen. The new organization includes a two-level call treatment. A first level of operators answers calls, identifies urgent calls and handles (numerous) non-urgent calls. Second level operators are specialists (policemen or firemen) and handle emergency demands. In this architecture, some calls are qualified as extremely urgent and receive a priority treatment. We are interested in the performance evaluation of bilevel systems corresponding to this general description.

We propose three different models addressing this kind of systems. The first two are timed Petri net models. We enrich the classical framework of free choice Petri nets by allowing conflict situations in which the routing is solved by priorities. The main difficulty in this situation is that the dynamics becomes non monotone.

In a first model, we consider discrete dynamics for this class of Petri nets. We prove that the counter variables of the Petri net are solutions of a piecewise linear system with delays. We investigate the stationary regimes of the dynamics, and characterize the affine ones as solutions of a piecewise linear system, which can be thought of as a system of rational equations over a tropical (min-plus) semifield of germs. Numerical experiments show that, however, convergence does not always holds towards these affine stationary regimes.

The second model is a infinitesimal version of the previous one. For the same class of Petri nets, we introduce a dynamics expressed by differential equations, so that the tokens and time events become continuous. For this differential system with discontinuous righthandside, we establish the existence and uniqueness of the solution. The benefit of this continuous model is that the discrete time pathologies disappear. We show however that the stationary regimes are the same as the stationary regimes of the discrete time dynamics. Numerical experiments tend to show that, in this setting, convergence effectively holds.

We also model the emergency call center described above as a queueing system, taking into account the randomness of the different call center variables. For this system, we prove that, under an appropriate scaling, the dynamics converges to a fluid limit which corresponds to the differential equations of our Petri net model. This provides support for the second model. Stochastic calculus for Poisson processes, generalized Skorokhod problems formulations and coupling arguments are the main tools used to establish this convergence.

Hence, our three models of an identical emergency call center yield the same schematic asymptotic behavior, expressed as a piecewise linear system of the parameters, and describing the different congestion phases of the system.

In a second part of this thesis, simulations are carried out and analyzed, taking into account the many details of our case study. The simulations confirm the schematic behavior described by our mathematical models. We also address the complex interactions coming from the heterogeneous nature of level 2.

Resumé

Nous analysons la dynamique de systèmes à événements discrets avec synchronisation et priorités, au moyen de réseaux de Petri et de réseaux de files d'attente. Nous appliquons cela à l'évaluation de performance d'un centre d'appels d'urgence.
Notre motivation est en premier lieu pratique. En 2016, un nouveau centre d'appels d'urgence a été mis en place pour l'agglomération parisienne, traitant les appels pour la police et les pompiers. La nouvelle organisation comporte deux niveaux de traitement. Un premier niveau d'opérateurs répond aux appels, identifie les appels urgents et traite les appels non urgents. Les opérateurs de second niveau sont spécialistes (policiers ou pompiers) et traitent les demandes d'intervention. De plus, certains appels sont identifiés comme très urgents et bénéficient d'un traitement prioritaire. Nous nous intéressons à l'évaluation de performance de divers systèmes correspondant à cette description générale.
Nous proposons trois modélisations différentes. Les deux premières sont des modèles de réseaux de Petri temporisés. Nous enrichissons le cadre classique des réseaux de Petri à choix libres en autorisant des situations de conflit où le routage est résolu par des priorités. La principale difficulté est alors que l'opérateur de la dynamique n'est plus monotone.
Dans un premier modèle, nous proposons une dynamique discrète pour cette classe de réseaux de Petri. Nous prouvons que les variables compteurs du réseau sont les solutions d'un système affine par morceaux avec retards. Nous étudions les régimes stationnaires de cette dynamique, et caractérisons les régimes affines comme solutions d'un système affine par morceaux, qui peut être vu comme un système d'équations rationnelles sur le semi-corps de germes tropical (min plus). Les applications numériques montrent cependant que la convergence ne se fait pas toujours vers ces régimes stationnaires affines.
Le second modèle est une version infinitésimale du précédent. Pour la même classe de réseaux de Petri, nous proposons une dynamique sous forme d'équations différentielles : les jetons et le temps deviennent continus. Pour ce système différentiel discontinu, nous établissons l'existence et l'unicité de la solution. L'avantage de cette modélisation continue est que les pathologies du temps discret disparaissent. Nous montrons cependant que les régimes stationaires sont les mêmes que ceux de la dynamique discrète. Les simulations numériques semblent montrer que la convergence s'obtient effectivement dans ce cas.
Nous modélisons aussi le centre d'appels d'urgence comme un réseau de files d'attente, prenant ainsi en compte le caractère aléatoire des différentes variables du centre d'appel. Pour ce système, nous prouvons que la dynamique, après une transformation d'échelle, converge vers une limite fluide, qui correspond au système d'équations différentielles de notre modèle de réseau de Petri. Cela conforte notre seconde modélisation. Les principaux outils de la preuve de convergence sont le calcul stochastique pour les processus de Poisson, des formulations en terme de problème de Skorokhod généralisé, ou encore des arguments de couplage.
Ainsi, nos trois modèles d'un même centre d'appels d'urgence définissent un même comportement asymptotique schématique, exprimé comme un système linéaire affine par morceaux, décrivant différentes phases de congestion du centre.
Dans une seconde partie de cette thèse, nous analysons des simulations poussées, prenant en compte les nombreux détails de notre étude de cas. Les simulations confirment le comportement schématique prédit par nos modèles mathématiques. Nous discutons aussi des interactions complexes provenant de la nature hétérogène du niveau 2.

Contents

Chapter 1

Introduction

1.1 A practical motivation: priority routing of calls in an emergency call center

The new organization of emergency call treatment in Paris area This thesis finds its source and impetus in a project led by Préfecture de police de Paris (PP) [1], in collaboration with the Brigade de sapeurs de pompiers de Paris (BSPP) [2].

Since early 2016, a new organization of the treatment of emergency calls became operative in Paris area. In a single call center, emergency calls to Police (17), Firemen (18) [3] and untyped emergency calls (112), are answered, and treated according to the type of emergency. The new architecture involves two levels of operators. At the first level, operators detect the type and urgency of the calls, handle (numerous) non urgent calls, and transfer urgent calls to second level operators, police or firemen. At the second level, operators handle the call request, and dispatch emergency means, if needed. The second level is split into two pools of operators, policemen and firemen. They have specific missions, and answer different types of calls. In contrast, the first level is common to police and firemen.

For the project leaders (see [dpdP16]), this new organization aims at improving emergency calls treatment, by identifying them faster and dedicating operators to them. Another objective is to decrease waiting times, by increasing the number of operators, and pooling part of police and firemen resources. Finally, gathering police and firemen operators in the same place is meant to bring a better coordination between both security forces for joint operations.

In order to improve the quality of the response for the most urgent calls, a key feature of the new organization is that, once they are identified as such, extremely urgent (EU) calls should always be in line with an operator. As a consequence, when a level 1 operator transfers a call, if the destination level is busy, the operator waits with the caller until a level 2 operator is available. Moreover, if several calls are waiting for the same destination level, EU calls have priority. We depict in Figure 1.1 the itinerary of an emergency call in the call center. Note that

1. Paris security authority
2. Paris Fire Brigade
3. also in charge of first aid

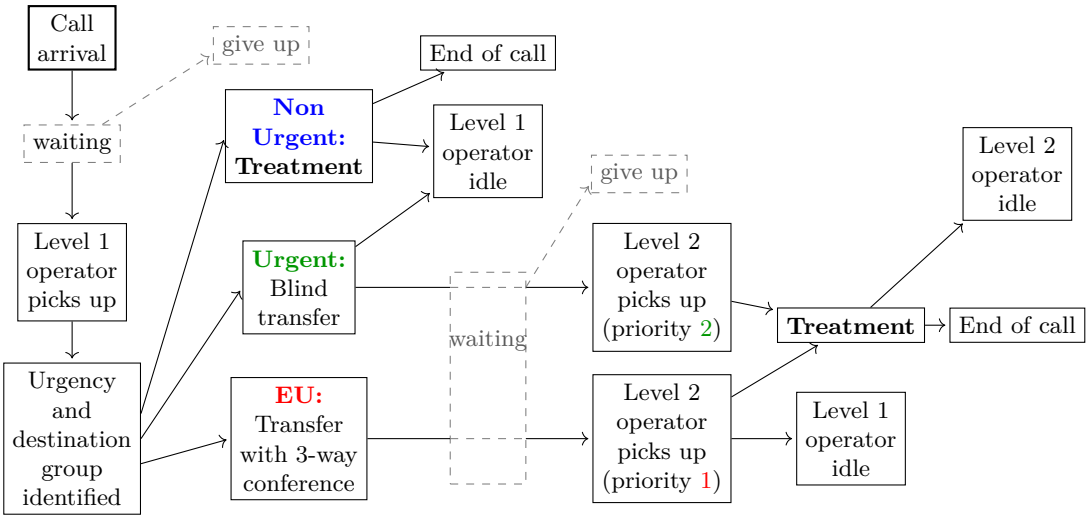


Figure 1.1 – Schematic flow chart of a call treatment in the two-level emergency call center.

this is a simplified model. For example, a few partner organizations have direct access to level 2 (*e.g.*, gas or electricity companies, public transportation operations centers).

This two-level architecture, together with the blocking of level 1 operators when a group of level 2 is busy, does not enter in the classical call center models, nor in the standard queueing network models. Specifically, these models would fail to account for the fact that the capacity of level 1 is diminished when a level 2 group is saturated. Moreover, we shall also not expect an exactly solvable Markov model: in these complex configurations, the invariant distribution cannot have a simple analytical expression.

This calls for exploring mathematical models in more details, in order to provide suitable formulæ for this system. From a user point of view, one would like to compute performance bounds, performance indicators, depending on the parameters of the system, as well as to deliver a general understanding on the different regimes and limit behaviors. In such situations, simple operating principles are as helpful as complex simulation-based tables and charts. Furthermore, a flexible tool is required, as the detailed architecture may vary in time and depending on the successive analyses and feedbacks.

This is the kind of results we endeavored to deliver to the project leaders and heads of the new emergency call center.

Petri net modeling Petri nets are a modeling language appropriate to account for concurrency and parallelism. A *Petri net* is a graph whose nodes are *transitions* and *places*, connected by directed arcs. Places hold *tokens*, which circulate from place to place, moved by *transition firings*. A transition can fire only if a token is available in each of the upstream places, and when a transition fires, it consumes (removes) one token in each upstream place, and produces (creates) one token in each downstream place. Therefore, a transition operates as a local synchronization and distribution module in a network. Tokens typically represent resources in a manufacturing process, or requests and servers in a communication network.

A Petri net modeling a simple call center is given in Figure 1.2.

Petri nets can also be given a timed semantics, in which case the circulation of tokens can encounter delays, or traveling times. In our simple call center, for example, we associate with each place a holding time: a token entering a place can leave it (by the firing of a downstream transition) only after having sojourned a given delay in this place.

Being able to model complex concurrency phenomena, together with a timed interpretation, Petri net is an appropriate, workable tool for modeling our two-level emergency call center.

On top of its flexibility, it is also a very convenient language for interacting with practitioners. Petri net's graphical representation, including token evolution by transition firings, makes it a directly intelligible language, and this facilitates the delicate process of modeling a real system.

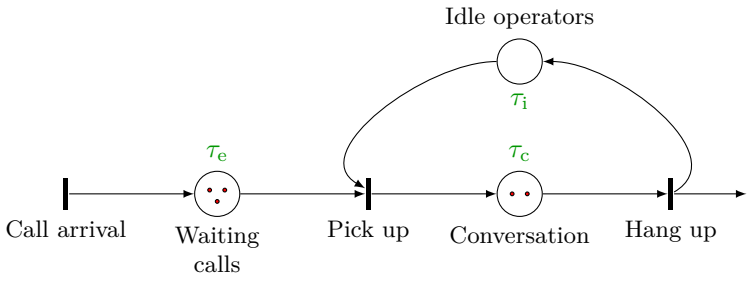


Figure 1.2 – Petri net of a simple call center. Transition are represented by thick black segments, and places by circles. Tokens are dark red dots in places. The arrival of calls is modeled by the left-most transition, and their release by the right-most transition. Here, there are two operators, both in conversation with a caller, and three other calls queueing. In green, we have represented the places' holding times*.

* Note that τ_e does not represent a token waiting time, but rather a fixed delay before entering in the queue, for example, an automatic welcome message. The waiting time if no operator is idle comes in addition to this holding time.

Priority Among the characteristics of the system described above, the priority allocation of level 2 operators to EU calls is the only (but crucial) non standard Petri net feature. In this thesis, we formalize priority routing of tokens in a timed Petri net.

Previously, Petri nets with priorities had already been studied in a non timed setting, which involves order relations between all transitions in the net. See for example [BK92]. In contrast, the priority rules that we study in our timed setting are local, restricted to *clusters* (group of connected upstream places and output transitions), because it takes a certain amount of time for a token to go from a cluster to another.

We found our inspiration in the anterior work of Farhi, Goursat and Quadrat [FGQ11], where such local priority routings are applied to a timed road traffic model.

Specific features of an emergency call center Emergency call centers differ from classical call centers in terms of objectives and characteristics.

Firstly, serving all callers, and serving them with minimum waiting times, is a much more involving imperative in an emergency call center, for which spared minutes and answered calls result in direct benefits in terms of lives, health and goods.

Secondly, an emergency call center must be designed, not only to face every-day situations, but also critical situations arising from expected or unexpected events (*e.g.*, storms, floods, terrorist attacks), in which the characteristics of demand (incoming calls) may be completely different. In such situations, one would like to design specific procedures to alert the people in charge, and to ensure that calls are served. This comprises resorting to reinforcements, and shifting in degraded modes.

In our work, we find simulations and formal analyses to be appropriate and complementary to account for such critical situations. While simulations allow one to focus on specific case studies, and to test *in silico* the performances of the planned organizations in the critical situations observed in the past, formal analyses provide information on the general behavior of the system, including at the limits. Besides, in our models, we will be particularly interested in stressed situations, in which the system is saturated in incoming calls.

1.2 A Petri net theory motivation: the "emergency physician" paradox

The paradox This apparent paradox was reported by Benchimol [Ben09], in the modeling of a hospital emergency department by the means of Petri nets. For the Petri net constructed in this work, some simulations were observed to yield a larger asymptotic throughput than what was expected from computation, by applying the formulæ of Cohen, Gaubert and Quadrat [CGQ95], allowing to compute the throughput of fluid approximations of Petri nets in which tokens are routed according to *preselection rules*. The author identified the resource which caused this

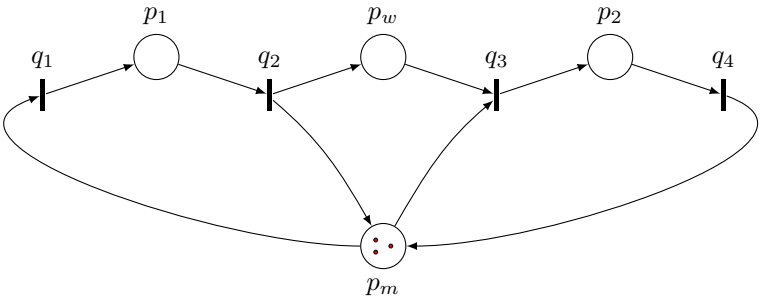


Figure 1.3 – The Petri net of the emergency physician paradox

throughput increase, and its associated subnet (a subset of places of transitions involving the resource). This is the Petri net depicted in Figure 1.3.

It models the medical consultations that a group of emergency physicians deliver to patients in a emergency department. After his or her arrival, and after a consultation by a nurse, a patient undergoes a first visit by a physician, place p_1. The physician usually asks for some complementary examinations in order to set the diagnosis (in fact, he or she always does in our modeling). Then, the patient goes through the series of exams without the presence of the physician, and, afterwards, returns in a consultation room where he or she waits for a second visit of a physician. Place p_w models both the series of examinations and the waiting of the second visit. Place p_2 models the second visit. The time for the complementary examinations is supposed to be fixed, equal to τ_w. Similarly, the time for a first (resp. second) consultation is τ_1 (resp. τ_2), and a physician who becomes available stays in place p_m a time τ_m before being dispatched to a patient.

The Petri net model is a very simple one. It is *consistent*, which means that firing once every transition yields an identical number of tokens in each place as before but it is not *conservative*, because the number of tokens in place p_w is unbounded: if doctors are always dispatched to first visits, the number of patients in place p_w never decreases and goes to infinity. Yet, it is not *free choice*, which means that the concurrency situations are not simple ones.

In the daily workflow, physicians ensure first visits as often as second visits. The transition throughputs are identical for every transition, and the greatest throughput is achieved if physicians always find a patient waiting when they become available, and it is

$$\rho^* = \frac{N_m}{\tau_1 + \tau_2 + 2\tau_m}\,, \tag{1.1}$$

with N_m being the number of physicians in the system. The throughput computed by Benchimol in his simulations was close to this throughput.

However, when modeling the routing of tokens in this Petri net by preselection rules, in which tokens are allocated to downstream transitions, regardless of which are fireable, the theoretical throughput can be much lower. Consider a situation in which the number of patients entering the system is saturated, physicians are dispatched, half the time to a first visit, half the time to a second visit, but, at time 0, the N_m physicians are all in place p_m. Then, at the beginning of the Petri net execution, half of the physicians are dispatched to second visits: no patient having being treated at this time, these physicians have to wait a time $\tau_1 + \tau_w$ before meeting their first patients, and this additional delay propagates during all the Petri net execution. Therefore, the throughput of this Petri net is then

$$\rho = \frac{N_m}{2\tau_1 + \tau_w + \tau_2 + 2\tau_m} < \rho^*\,,$$

and, as the delays τ_w and τ_1 are likely to be large, the throughput loss is sizable. This was the throughput computed by Benchimol.

The same phenomenon was identified by Gaujal and Giua [GG04b, Section 4], who underlined that, even if the routing proportions are those optimizing the throughput in a Petri net (sending the physicians half the time to first visits and half the time to second visits is the best strategy), this optimal throughput is not automatically obtained: the asymptotic throughput

still depends on the initial markings in the places. In the Emergency physician Petri net, assigning the N_m physicians to place p_1 at time 0 allows to reach the optimal throughput ρ^*, even with preselection routing.

Discussion This apparent paradox, while not dissimulating complex mathematical issues, still raises interesting modeling questions.

In a Petri net, we call *conflict* the situation in which one place has several output transitions. This term underlines the fact that a token entering a place can be fired by only one of the downstream transitions. In the modeling of a timed system, one would like to set a *routing rule*, which would solve the conflict for each token entering the place, that is, allocate the token to one of the place's output transitions.

Assigning tokens (or fractions of) to the output transitions according to fixed ratios is a convenient routing rule, which has led to powerful analytic results, allowing one to express the asymptotic throughput of the system as the solution of linear programming [CGQ95, GG04b]. In addition, it is also a good upper approximation of periodic routing, and of *Bernoulli routing* (assigning a token to output transitions according to fixed probabilities). It has also a strong relationship with the *race policy* routing. See [BGM06] for an analysis of all these routing rules.

However, it fails to model downstream-dependent behaviors, like the one we have in this model: in reality, in an emergency department, a physician who does not find any patient waiting for a second visit would not stay unoccupied until a second-visit patient arrival, but would take care of a first-visit patient. Thus, in terms of Petri net, the allocation of tokens to the output transitions of place p_m is conditioned by the availability of tokens in place p_w. This cannot be modeled by a pre-allocation routing scheme.

In this regard, the priority routing which is proposed in this work can be seen as an alternative, downstream-dependent routing procedure for tokens in Petri nets. Its analysis and the subsequent dynamical equations proposed in this dissertation could hence be useful to every Petri net user encountering priority or other downstream-dependent phenomena in the systems being modeled. Furthermore, as we will show, analytical formulæ and algorithms are still available to compute the corresponding throughputs.

One can retain from this example that the throughputs obtained with priority routing can be completely different from the throughputs computed in a preselection setting. Our interpretation is that this is because the monotonicity of the system is lost. This is the topic of the next section, which enters one step deeper in theoretical questions.

Still, we cannot leave this section without setting the reader's mind at rest: an alternative Petri net model for the emergency physician case shall be proposed later on, addressing the drawbacks of the current model. See Section 2.4.2. In other words, we solve the paradox, by replacing preselection rules by priority rules, and showing that the latter are still amenable to an algebraic analysis.

1.3 A dynamical systems motivation: beyond non-expansive operators

Let X be a vector space, and $T : X \to X$ an operator on X. The dynamics of a Petri net, as many other discrete event dynamics, can be modeled by a system of the form:

$$x(n) = T(x(n-1)) \quad \forall n \in \mathbb{N}, \qquad \text{with } x(0) = x_0 \in X\,.$$

For example, in this thesis, we will be interested in the counter variables of the different transitions of a Petri net. The $x(n)$ will then be variables in $\mathbb{R}^{\mathcal{Q}}$, and, for q a transition, $x_q(n)$ will be the number of firings of transition q up to date n included.

An important question in discrete event dynamical systems is to determine whether the sequence $x(n)/n = T^n(x_0)/n$ has a limit as n tends to ∞, and whether this limit depends on the initial condition (with T^n denoting the n-th iterate of T). The limit, if it exists, is commonly named the ***cycle time*** of T, denoted by $\chi(T)$, due to the following dual interpretation of x: n describes the n-th event of a discrete event process, and $x(n)$ is the date of this event. Despite this terminology, in the context of counter variables of Petri net, counting the number of firings of transitions up to a given date, such limit corresponds to an asymptotic throughput.

If T is monotone homogeneous, the answer is known.

The eigenvalue problem An operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is

- ***non-expansive*** for a given norm $\|\cdot\|$ of $\mathbb{R}^n$ if

$$\|T(x) - T(y)\| \leqslant \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n ,$$

- ***monotone*** if

$$x \leqslant y \implies T(x) \leqslant T(y), \quad \text{for all } x, y \in \mathbb{R}^n ,$$

  where $\leqslant$ is the usual partial order of $\mathbb{R}^n$,

- ***additively homogeneous*** if

$$T(\lambda + x) = \lambda + T(x), \quad \text{for all } x \in \mathbb{R}^n, \lambda \in \mathbb{R} ,$$

  where the addition of a scalar and a vector must be understood as an addition of this scalar to each coordinate of the vector.

These three properties are closely related. Crandall and Tartar [CT80] proved that an additively homogeneous map is non-expansive in the sup-norm if and only if it is monotone.

If T is non-expansive, then the limit of $T^n(x_0)/n$, if it exists, is independent on the initial condition.

Suppose now that T admits an *additive eigenvector*, that is, a vector ρ associated with a scalar u (*additive eigenvalue*) such that $T(\rho) = u + \rho$. Then, $T^n(\rho) = \rho + nu$, and therefore, all the coordinates of $x(n)/n$ have a common limit, independent of the initial conditions, equal to u. Therefore, an important question is to determine the existence of such generalized eigenvectors. If T is monotone (and if it respects a condition of connexity), a nonlinear equivalent of the Perron-Frobenius theorem allows to answer in the affirmative. See [GG04a]. More generally, nonlinear Perron-Frobenius theory provides a number of results allowing to characterize the cycle time of monotone non-expansive operators. We refer to Gaubert [Gau05], who surveys these results and their application to discrete event systems.

We do not detail these results, except for the following one, which applies to piecewise linear systems. For u, ρ in $\mathbb{R}^n$, the mapping $t \mapsto u + \rho t$ is an ***invariant half-line*** of T if $T(u + \rho t) = u + \rho(t + 1)$, for any $t > 0$. Kohlberg [Koh80] proves that, if T is a piecewise linear, non-expansive map, T has an invariant half-line, and, moreover, ρ is unique. A direct corollary is that the sequence $x(n)/n$ converges and has a limit independent of the initial conditions. This applies, in particular, if T is expressed as the minimum of a finite family of linear maps, $T(x) = \min_{i \in I} t_i(x)$.

Ergodic theory An important question in dynamical systems is to relate time averages with space averages: if the operator of a dynamics converges to a given orbit, is the asymptotic time average of the trace equal to the mean value of the orbit?

Birkhoff's ergodic theorem is central to this regard. We state it, following [Rob03, Chapter 10], in the case of endomorphisms of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathcal{F}$ a Borel σ-field. Recall that T is an endomorphism of $(\Omega, \mathcal{F}, \mathbb{P})$ if it is measurable, and if the probability measure $\mathbb{P}$ is invariant by T.

▶ **Theorem 1.1** (Birkhoff's ergodic theorem)**.** *If $T : \Omega \to \Omega$ is an endomorphism and f an integrable function, $\mathbb{P}$-almost surely,*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(T^k(\omega)) = \mathbb{E}(f \mid \mathcal{I})(\omega)$$

where $\mathcal{I}$ is the σ-field of the invariant measurable sets.

An operator T is ***ergodic*** if any invariant set by T has probability 0 or 1. If T is ergodic, a consequence of Theorem 1.1 is that, for any integrable function f, $\mathbb{P}$-almost surely,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(T^k(\omega)) = \mathbb{E}(f)$$

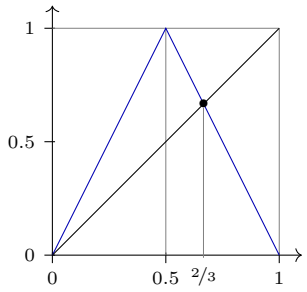


Figure 1.4 – The 'tent' map (in blue). 0 and $2/3$ are its two fixed points.

As a consequence, for an ergodic operator, the asymptotic throughput is almost surely constant, equal to a mean in space.

A reference on ergodic theory is the book of Cornfeld [CFS12].

Counters of the dynamics of Petri nets are determined by more complex operators, involving time shifts. However, it was proved that, with classical routing rules (pre-allocation, fluid approximation of Bernoulli routing), such operators are monotone, non-expansive, and, with some additional constraints, the dynamics usually converges towards an asymptotic value, possibly infinite, independent of the initial conditions. See Cohen, Gaubert and Quadrat [CGQ95] and Gaujal and Giua [GG04b]. In stochastic settings, ergodic theory, together with monotonicity properties, often helps to determine the convergence towards asymptotic throughputs. See for example Baccelli and Mairesse [BM98] and Gaujal, Haar and Mairesse [GHM03].

In contrast, when allowing priority routing in Petri nets, the operator of the dynamics becomes non monotone: a token entering later in the system can pass by a token which entered before. The classical results obtained on Petri nets with monotone operators do not apply in this situation, and we need to investigate the behavior of non monotone operators.

Non monotone operators The following example, developed by Farhi, Goursat and Quadrat [FGQ11], shows that the general non monotone case is more complicated, even for an additively homogeneous, ergodic operator.

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $T(x_1, x_2) = (x_2, 3x_2 - 2x_1 \wedge 2 + 2x_1 - x_2)$, where $\wedge$ stands for a min operator. The map T is 1-homogeneous, and therefore, its analysis can be reduced to analyzing the operator $\hat{T}$ in the projective space $\mathbb{R}^2/\mathbb{R}$ (with $\hat{y} = y_2 - y_1$):

$$\hat{T}(\hat{x}) = 2\hat{x} \wedge 2(1 - \hat{x})\,. \tag{1.2}$$

This is the (well-known) tent map, depicted in Figure 1.4.

It admits two fixed points, 0 and $2/3$, which are, therefore, eigenvalues of T. Moreover, $\hat{T}^k(x)$, with x any number with a finite binary development, reaches 0 after a finite number of iterates: thus, a dense set of initial conditions of $[0, 1]$ is such that $\hat{T}^k(x)$ converges towards 0.

However, neither 0 nor $2/3$ corresponds to a mean asymptotic value of $\hat{T}^k(x)/k$, independent of the initial conditions. In fact, the trajectory of $\hat{T}^k(x)$ is chaotic for any irrational number. In addition, for the Lebesgue measure, the unique invariant sets are sets of measure 0 or 1. Therefore, $\hat{T}$ is ergodic, and the asymptotic cycle time $\hat{T}^k(\omega)/k$ converges almost surely towards the mean value of $\hat{T}$, that is, $1/2$. We refer to Collet and Eckmann [CE09] for a more detailed analysis of the tent map dynamics.

Hence, for non monotone operators, the asymptotic cycle time can be different from the eigenvalues of the operator.

A motivation of our work is thus to go further in the analysis of such non monotone maps resulting from Petri net dynamics with priorities.

Note that it is an open problem to construct a Petri net with priorities whose dynamics would be reducible to the tent map.

1.4 Contributions

We analyze the dynamics of discrete event systems with synchronization and priorities, by means of Petri nets and queueing networks. We apply this to the performance evaluation of the bilevel emergency call center described in Section 1.1.

Timed Petri nets are a convenient tool to model discrete event systems with complex concurrency phenomena. Their performance is measured in terms of their counters variables, counting the number of firings of transitions up to the current date. For restricted classes of Petri nets, like event graphs, the dynamics of these counter variables are known to be expressed by *tropical* (min-plus) linear equations, see Baccelli, Cohen, Olsder and Quadrat [BCOQ92]. Cohen, Gaubert and Quadrat [CGQ95] characterized the dynamics of a larger class of Petri nets, Petri nets with *preselection routing*, as a combination of max-plus linear and classical linear equations. From these counters equations, one can compute the asymptotic throughputs of a *fluid approximation* of the Petri net, using the techniques mentioned in the above section (Section 1.3). Convergence towards stationary regimes, whose throughputs are computed as solutions of linear programs, was shown in [CGQ95] for Petri nets having a positive $\mathcal{Q}$*-invariant*, and in Gaujal and Giua [GG04b] in the general case.

Our approach in Chapter 3 builds on this series of results. A main novelty is that, while the previous models were limited to Petri nets with preselection routing, whose fluid approximation is in fact equivalent to the simpler class of choice-free Petri nets, we allow concurrency configurations, in which tokens are routed according to priority rules. We show that the dynamics of Petri nets with free choice and priority routing can be expressed by piecewise linear equations, leading to a rational tropical dynamics (3.3)–(3.7), thus generalizing the case of Petri nets with preselection routing. Moreover, we provide a complete proof of equivalence between the counters along execution traces of our Petri net, and the *càdlàg*, non-decreasing solutions of these piecewise linear equations (Theorem 3.1). We found our inspiration in the work of Farhi, Goursat and Quadrat [FGQ11], in which this modeling of priorities by rational tropical dynamics was applied to a timed Petri net describing a road traffic network.

Like in the case of Petri nets with preselection routing, this allows us to investigate the asymptotic regimes. For the fluid approximation of the dynamics, we show that the affine stationary regimes of our class of Petri nets are precisely the solutions of a set of lexicographic piecewise linear equations, which constitutes a rational system over a tropical semifield of germs of affine functions: see Theorem 3.6.

However, because of the priority rules, the operator of our dynamics becomes non monotone (in a Petri net, a token having priority can pass by a non priority token), contrary to the preselection routing case. This has two drawbacks. Firstly, one cannot apply classical iteration algorithms, inspired by value iteration in Markov decision processes, to compute these affine stationary solutions. Moreover, our applications show that several affine stationary regimes may exist, for a given set of parameters, depending on the initial conditions (and they do not form a convex set), so that one has to enumerate all the *policies* of the net in order to determine the solutions. Nevertheless, expressing this problem as a rational system over a tropical semifield shows that it reduces to solving a tropical polynomial system, so that the asymptotic throughputs can still be computed.

Secondly, more fundamentally, the asymptotic regimes of this dynamics are not always the expected affine regimes. Numerical experiments show that periodic behaviors can be reached asymptotically, leading to different asymptotic throughputs. Such phenomena are occasioned by arithmetical relationships between the holding times. Thus, it can be considered as a pathology originating from our discrete time modeling.

The model investigated in Chapter 4 is therefore a continuous time one, designed so as to avoid the pathologies of the discrete time one. In this chapter, we provide an alternative, infinitesimal version of the dynamics described above, for the same class of Petri nets. In this setting, the dynamics becomes a hybrid dynamics, expressed by a system of differential equations, with a discontinuous right-hand-side. We require the variables of the Petri net to be *forward Carathéodory solutions* of this hybrid system, that is, they are absolutely continuous solutions of the system in its integral form, and left-accumulation of switching times is forbidden (this corresponds to our solutions being càdlàg functions).

Differential dynamics of Petri nets were introduced by David and Alla [DA87], with the

same motivation as ours, that is, computing simple approximations of the behavior of a discrete Petri net. See also Vázquez *et al.* [VMJS13]. The main difference with our equations is that we model an infinitesimal equivalent of *holding durations*, while the original model of [DA87] rather considered *enabling durations*. Thus, the routing in this model was arbitrated by a race policy, while we can handle priority routing, in addition to preselection routing. Furthermore, this leads us to distinguish between tokens *under processing* and *idle* tokens in a place, and our dynamics become discontinuous, because the firing rate of a transition depends on the presence of idle tokens in its upstream place.

Our piecewise linear, piecewise continuous dynamics can be expressed as an infimum of linear dynamics, expressed in terms of policies of a Petri net. A policy associates with each transition a *bottleneck* upstream place, determining the flow of the transition when the policy reaches the infimum. One of our main results, Theorem 4.9, is that there exists a unique forward Carathéodory solution for this dynamics. It relies on the constructive proof that, at any time, there exists a policy determining a valid solution on a forward time interval, and that this policy is unique, except for the case that another policy yields the same dynamics.

Similarly to the discrete time case, we exhibit the affine stationary regimes of this hybrid dynamics, see Theorem 4.11. We show in Corollary 4.12 that they are the same as the ones computed in the discrete time case. Furthermore, numerical experiments tend to show that, unlike the discrete time model, stationary regimes are always reached by the Petri net dynamics, thus confirming the relevance of this model.

The idea of modeling a physical system by differential equations may seem remote from reality. Chapter 5 gives support to this modeling, by providing the proof that the dynamics of a stochastic, continuous-time network system, representing a bilevel call center, converges towards the same set of differential equations, up to an appropriate scaling of the system variables.

More precisely, the system considered in Chapter 5 is not a Petri net, but a queueing network, representing a bilevel emergency call center with only two classes of calls, urgent and non urgent. Level 1 operators answer all incoming calls, handle non urgent calls and transfer urgent calls to level 2 operators. If all level 2 operators are busy, then the level 1 operator waits with the urgent call, so that he is blocked until a level 2 operator becomes idle. The distribution of service times is exponential at each level, so that the corresponding stochastic process is Markovian.

For this model, under an appropriate scaling, we establish the convergence of the quantities of the system (fraction of blocked servers, fraction of free servers) towards asymptotic values expressed as a piecewise linear function of the parameters of the system. The proof of convergence is technical, because reflection conditions at boundaries (depending on the presence of blocked operators at level 1 or idle operators at level 2) makes the analysis more complex. We resort to a scaling analysis (this kind of analysis was applied by Kelly [Kel91] to loss networks), which allows us to separate two regimes, one in which a fraction of level 1 servers remains blocked after some fixed delay with probability one, and one in which a fraction of level 2 servers is idle after some fixed delay with probability one. Interestingly enough, we show that, at the scaling limit, after some fixed time, the dynamics is solution of an ordinary differential system, corresponding to the dynamics proposed in Chapter 4.

Finally, our three dynamical models, based on different semantics (see Table 1.1), applied to an identical physical system, lead to the same schematic asymptotic behavior, expressed as a piecewise linear system of the parameters, and describing the different congestion phases of the system.

Regarding our bilevel emergency call center, under a saturation hypothesis, our asymptotic analysis allows us to identify three different congestion phases, described in Figure 1.5, depending on the ratio between the number of operators of level 2, N_2, and the number of operators of level 1, N_1. Note that we suppose here that all level 2 operators have the same role. Going from right to left, when N_2/N_1 is large, level 2 is sufficiently sized, and all calls are handled at level 2. In an intermediate phase in which N_2/N_1 is between two critical rates, expressed in terms of the parameters of the system, level 2 is congested, so that some urgent calls are not answered, but the extremely urgent calls are protected by the priority rule. In the lower

	Discrete time	Continuous time
Discrete firings	Chapter 3 (fixed processing times)	Chapter 5 (stochastic processing times)
Infinitesimal firings		Chapter 4 (hybrid dynamical system)

Table 1.1 – Our three mathematical models in a nutshell.

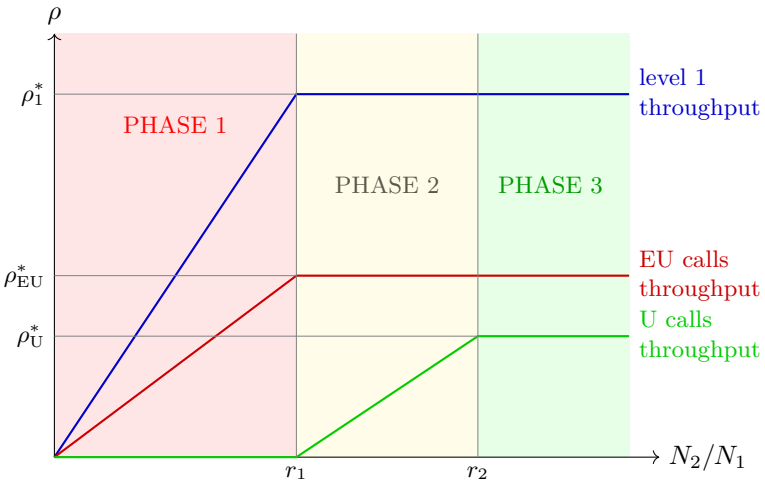


Figure 1.5 – The three phases of the bilevel emergency call center with an homogeneous level 2, depending on the ratio between the number of operators at level 2 (N_2) and the number of operators at level 1 (N_1).

phase when N_2/N_1 is small, level 2 is so congested that no urgent call is handled and that the treatment of extremely urgent calls is slowed down. Because of the three-way conversation between level 1 operators and level 2 operators, level 1 operators are also blocked, waiting for level 2, and the throughput of level 1 is diminished.

In the unique chapter of Part II (Chapter 6), we focus on our case study of the Parisian emergency call center, and apply the analytical methods of Part I to a bilevel call center in which the second level is composed of different groups of operators handling different kinds of calls (police and firemen, in our case study). We also use simulations, which help us to present our results with a more operational point of view, and which take into account a certain number of characteristics of our call center that were not incorporated in our simplified model of Part I.

Despite the operational advantages of this complex bilevel organization, a few situations are identified in which attention of practitioners is required. This is in particular the case of situations of "cross-congestion", in which one of the groups of level 2 is saturated, and slows down level 1, because of three-way conferences. In such situations, all other groups of level 2 are also slowed down, because level 1 becomes bottleneck. This is an unwelcome side effect of this organization, but simple procedures can help avoiding such situations.

We also propose a statistical analysis of our emergency call center data (without entering too much into details), and point out a few additional observations derived from our simulations, as for example, the unavoidable trade-off between operators activity and calls abandonment. Calls abandonment could not be modeled by our Petri net class of Part I.

Part I

A Dynamical Analysis

Chapter 2

Preliminaries on Petri nets

The objective of this chapter is to recall some basic definitions, notations and properties about Petri nets. In addition to the original model of Petri nets, we also develop some well-known extensions of this model that will be used or mentioned in this work. We try to propose a few relevant references for each of these extensions.

Going through all these notions and definitions (Sections 2.1, 2.2, 2.3), we arrive to the notion of *routing rules*, and introduce the one which is central to this thesis, *priority routing* (Section 2.4).

Finally, we propose three Petri net examples in Section 2.5. The first two ones will serve as applications of the results in Chapters 3 and 4. The third one is an equivalent of the queueing network examined in Chapter 5.

2.1 Untimed Petri nets

The modeling language of Petri nets was introduced in the beginning of the sixties by Petri [Pet62]. It was thought as a formal tool, aimed at modeling systems encountering concurrency and parallelism, and, therefore, convenient for practical applications such as manufacturing processes, communication networks, chemical reaction networks, and, more generally, processes where countable resources circulate between different places, encountering "joins" (rendez-vous, or synchronization), "forks" (splitting or branching) and "merges" (additions).

By the analysis of a Petri net, one would like to understand the behavior of a physical system, identify the critical paths of resources, the possible deadlocks and unexpected states, and provide guarantees of a "well-behaved" design.

We start by recalling the basic definitions of Petri nets.

2.1.1 General definitions of Petri nets

▶ **Definition 1.** A ***Petri net*** is a triple $(\mathcal{P}, \mathcal{Q}, F)$, consisting of a finite set $\mathcal{P}$, whose elements are called ***places***, a finite set $\mathcal{Q}$ whose elements are called ***transitions***, $\mathcal{P} \cap \mathcal{Q} = \emptyset$, and a mapping $F : (\mathcal{P} \times \mathcal{Q}) \cup (\mathcal{Q} \times \mathcal{P}) \to \mathbb{N}$ which indicates multiple directed ***arcs*** between places and transitions. $F(p, q)$ defines the number of arcs from place p to transition q, and $F(q, p)$ defines the number of arcs from q to p.

The mapping F of a Petri net is fully characterized by two $\mathcal{P} \times \mathcal{Q}$ matrices of natural integers, denoted by C_+ and C_-, such that the (p, q) entry of C_+ is $F(q, p)$, indicating the number of *forward* arcs from transition q to place p, and the (p, q) entry of C_- is the value $F(p, q)$, indicating the number of *backward* arcs, pointing to transition q, from place p. We call C_+ the ***forward matrix*** of the Petri net and C_- its ***backward matrix***. Note that the ordering of the columns and rows of C_+ and C_- entails a numbering of places and transitions: we usually note transitions $q_1, q_2, \ldots, q_{|\mathcal{Q}|}$, and places $p_1, p_2, \ldots, p_{|\mathcal{P}|}$, according to this ordering. Owing to the equivalence between F and the pair C_+, C_-, a Petri net can equivalently be defined by a tuple $(\mathcal{P}, \mathcal{Q}, C_+, C_-)$. Note that, in the Petri net literature, one often encounters the notations Post (or W^+) and Pre (or W^-) to designate, respectively, the forward matrix and the backward matrix. Equivalently, one also often consider $F(x, y)$ as the ***valuation*** of a single (x, y) arc.

For two elements $x, y \in \mathcal{P} \cup \mathcal{Q}$, we note $x \to y$ if $F(x, y) > 0$, and say that y is a forward neighbor of x, and x a backward neighbor of y. The set of backward neighbors of x is denoted $x^{\text{in}} := \{y \in \mathcal{P} \cup \mathcal{Q} \mid F(y, x) > 0\}$ and its set of forward neighbors is $x^{\text{out}} := \{y \in \mathcal{P} \cup \mathcal{Q} \mid F(x, y) > 0\}$. For a place p, sets p^{in} and p^{out} are subsets of $\mathcal{Q}$. The set p^{in} is called the set of ***input*** transitions of p, and p^{out} the set of ***output*** transitions of p. Similarly, for a transition q, sets q^{in} and q^{out} are subsets of $\mathcal{P}$. The set q^{in} is called the set of ***upstream*** places of q, and q^{out} is called the set of ***downstream*** places of q.

We call *self-loop* the situation where a pair (p, q) has at least one backward arc and one forward arc. A ***pure*** Petri net is a Petri net without self-loops, *i.e.*, where the existence of a (p, q) arc and of a (q, p) arc are mutually exclusive. This is equivalent to having $\min(C_+, C_-) = 0$ (by the minimum of two matrices or vectors, we mean the matrix or vector composed of entrywise minima). For pure Petri net, we define the ***place–transition incidence matrix***, or, for short, the ***incidence matrix*** $C := C_+ - C_-$. A matrix $C \in (\mathbb{N} \cup -\mathbb{N})^{\mathcal{P} \times \mathcal{Q}}$ uniquely determines the matrices C_+ and C_- of a pure Petri net by $C_+ = \min(C, 0)$ and $C_- = \min(-C, 0)$, so that a triple $(\mathcal{P}, \mathcal{Q}, C)$ defines a pure Petri net.

▶ **Definition 2.** A ***marking*** of a Petri net is a mapping $m : \mathcal{P} \to \mathbb{N}$. A place such that $m(p) > 0$ is called a ***marked place***. For such a place, we say that $m(p)$ designates the number of ***tokens*** of place p. We equivalently describe a marking as a column vector of $\mathbb{N}^{\mathcal{P}}$. A ***marked Petri net*** is a pair (N, m_0) where N is a Petri net, and m_0 a marking, called the ***initial marking*** of the marked Petri net.

By ***state*** of a Petri net N, we mean a given marking $m \in \mathbb{N}^{\mathcal{P}}$.

The following conventions hold for graphical representation of Petri nets: places are depicted by circles, transitions by rectangle or thick segment lines, and directed arcs link places to transitions, and vice-versa. Multiple arcs can equivalently be depicted by arcs with an integer valuation. Finally, tokens are represented by dots inside the circles of the places. See Figure 2.1.

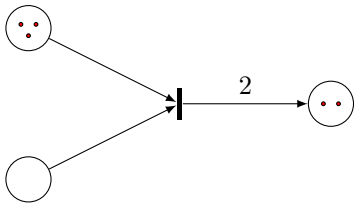


Figure 2.1 – A Petri net with three places and a transition. Despite the three tokens in one of its upstream places, the transition is not fireable, because its second upstream place is empty. A firing of the transition would produce two tokens in its downstream place.

A Petri net is ***connected*** if it is connected as a bipartite graph. A ***subnet*** of a Petri net is defined by $\mathcal{P}', \mathcal{Q}', C'$, where $\mathcal{P}' \subseteq \mathcal{P}$, $\mathcal{Q}'$ is the set of transitions of $\mathcal{Q}$ having at least one arc to or from $\mathcal{P}$, and $C' = C_{|\mathcal{P}',\mathcal{Q}'}$, that is, C restricted to its entries in $\mathcal{P}' \times \mathcal{Q}'$. A non connected Petri net can be described by the partition of its maximal connected subnets. The absence of arcs between two of these maximal connected subnets implies the absence of relationships between them. Therefore, in the following, without loss of generality, we only consider connected Petri nets.

2.1.2 Firing of transitions and Petri net dynamics

The dynamics of a marked Petri net describes the evolution of its marking under firings of transitions.

For a matrix A of dimensions $\mathcal{P} \times \mathcal{Q}$, the notation A_q with $q \in \mathcal{Q}$ stands for the q-th column of A. The notation $A_{p,q}$ with $p \in \mathcal{P}$ and $q \in \mathcal{Q}$ stands for the (p, q) entry of A.

▶ **Definition 3.** Let $N = (\mathcal{P}, \mathcal{Q}, C_+, C_-)$ be a Petri net and m_0 be a marking. We say that transition q is ***fireable*** with m_0 if $m_0 \geqslant (C_-)_q$ (the notation $(C_-)_q$ designates the q-th column of C_-, so this is a entrywise inequality on two vectors of $\mathbb{N}^{\mathcal{P}}$), that is, if for each upstream place p of q, the marking of p is larger than the number of arcs (p, q).

A transition q ***fires*** in state m to state m' if it is fireable with marking m, and if

$$m' = m - (C_-)_q + (C_+)_q \,. \tag{2.1}$$

We use the notation $m \xrightarrow{q} m'$.

The firing of a transition consists in decreasing the marking in upstream places and increasing the marking in downstream places, in quantities given by the backward and forward matrices. We say that transition q ***consumes*** $(C_-)_{p,q}$ tokens in each upstream place p and ***produces*** $(C_+)_{p',q}$ tokens in each downstream place p'.

We now define a firing sequence as a sequence of firings of transitions, such that the $(k+1)$-th transition is fireable for the marking reached after the firings of the k first transitions.

▶ **Definition 4.** A ***firing sequence*** σ for marked Petri net (N, m_0) is a word on transitions $\sigma \in \mathcal{Q}^*$, that satisfies the following, inductive, rules:

$$\begin{aligned} &m \xrightarrow{\varepsilon} m' && \text{if } m = m' \\ &m \xrightarrow{\sigma q} m' && \text{if } \exists m'' \in \mathbb{N}^{\mathcal{P}} : m \xrightarrow{\sigma} m'' \xrightarrow{q} m' \,, \end{aligned}$$

where σq is the word composed of the prefix σ and the letter q, and ε is the empty word.

We say that the firing sequence σ ***reaches*** marking m' from m if $m \xrightarrow{\sigma} m'$. We also speak of a firing sequence of a marked Petri net as an ***execution*** of the marked Petri net.

In discrete Petri nets, one is typically interested in describing the reachable markings of a Petri net, that is, the markings that can be reached by some sequence of transitions.

▶ **Definition 5.** Let (N, m_0) be a marked Petri net. A marking m is said to be ***reachable*** for (N, m_0) if there exists a firing sequence σ such that $m_0 \xrightarrow{\sigma} m$.

We call ***reachability set*** of (N, m_0) the set of reachable markings of (N, m_0).

If m is a reachable marking of (N, m_0) and σ a firing sequence from m_0 to m, we denote by $|\sigma|$ the vector of $\mathbb{N}^{\mathcal{Q}}$ counting the occurrences of every transition in σ, that is, $|\sigma|_q$ is the number of occurrences of q in σ. Vector $|\sigma|$ is called the ***occurrence count vector***, or *Parikh vector*, or *commutative image* of σ. The markings m_0 and m are related to $|\sigma|$ by the following result:

▶ **Lemma 2.1.** *Let (N, m_0) be a marked Petri net and let m be a reachable marking of m_0, with the associated firing sequence σ. We have:*

$$m = m_0 + C|\sigma| \,. \tag{2.2}$$

*This equation is called the **fundamental equation** of the Petri net.*

Given two markings m and m', the existence of an $x \in \mathbb{N}^{\mathcal{Q}}$ such that $m' = m + Cx$ is a necessary, but in general not sufficient condition for m' to be a reachable marking of m. Moreover, if m' is a reachable marking of m and if x satisfies the equation $m' = m + Cx$, there does not necessarily exist a firing sequence σ with $m \to^{\sigma} m'$, whose occurrence count vector would be x.

Proof. Equation (2.1) can be written $m' = m + Ce_q$, with e_q the vector of dimension $|\mathcal{Q}|$ such that $(e_q)_q = 1$ and $(e_q)_{q'} = 0$ for $q \neq q'$. One proves by induction on the length of σ that $m' = m + C(\sum_{q \in \mathcal{Q}} |\sigma|_q e_q)$. ◀

The following properties are related to the notion of reachable markings and firing sequences:

▶ **Definition 6** (Basic properties of a Petri net)**.** Let (N, m_0) be a marked Petri net.

- A transition q is ***dead*** if there does not exist a firing sequence σ such that σq is a firing sequence from m_0.
- A transition q is ***live*** (or *strongly live*) if for every reachable marking m, there exists a firing sequence σ such that σq is a firing sequence from m.
- A reachable marking m of (N, m_0) is a ***deadlock*** if no transition is fireable at m. If no reachable marking of (N, m_0) is a deadlock, then (N, m_0) is ***deadlock-free***.
- A place p is k-bounded if, for any reachable marking m, $m(p) \leqslant k$. It is ***bounded*** if there exists k such that it is k-bounded.
- If every transition of (N, m_0) is dead, we say that the Petri net is ***dead***. This is equivalent to saying that no transition is fireable for m_0.
- If every transition of (N, m_0) is live, we say that (N, m_0) is a ***live*** Petri net.
- If every place of (N, m_0) is bounded, we say that (N, m_0) is a ***bounded*** Petri net.

Characterizing the reachable states of a Petri net has been an important problem in Petri net theory for decades. It was observed that many other problems on Petri nets reduce to this reachability problem [Hac76]. We owe to Mayr [May84] and Kosaraju [Kos82] a major theorem in this respect: the reachability problem for Petri nets is ***decidable***, that is, that there exists an algorithm that answers if, for a marked Petri net (N, m) and a vector $m' \in \mathbb{N}^{\mathcal{P}}$, m' is reachable from m (and returns a firing sequence σ such that $m \to^{\sigma} m'$). The proof builds on the results of Karp and Miller on vector addition systems [KM69], which where shown to be equivalent to Petri nets. However, the reachability problem is EXPSPACE-hard [Lip76]. See also the recent algorithm of [FL15].

Reachability is just one of many useful Petri net properties. One would for example like to know if a Petri net is *bounded*, if some marking is a *home state*, if a marked Petri net is *live*, or *deadlock-free*, or *persistent*. One would also like to know if a given Petri net reachability set has some specific structure, for example, if it is a *language* or a *semilinear set*. We refer to the overview of Esparza and Nielsen [EN94] for decidability results (and definitions) of such properties, for various subclasses of Petri nets. Many of these problems reduce to the reachability problem.

We point out that these results build on the analysis of some structures associated with Petri nets, such as the *reachability graph* of a Petri net, its *coverability graph* (see definitions in [Mur89]), or its *occurrence net* (introduced in [BD90]).

2.1.3 Free choice Petri nets

In contrast with the difficulties of the analysis of general Petri nets, some subclasses of Petri nets were identified, for which most of the problems listed above find an easier answer (*e.g.*, polynomial-time algorithms may exist). Among these classes, free-choice nets are of particular interest.

▶ **Definition 7.** A Petri net $(\mathcal{P}, \mathcal{Q}, F)$ is said to be ***free-choice*** if, for any pair of transitions, either they have no common upstream places, or they have a single, identical upstream place with the same valuation (the same number of arcs):

$$\forall q, q' \in \mathcal{Q}\ ,\ q' \neq q\ ,\quad {q}^{\text{in}} \cap {q'}^{\text{in}} \neq \emptyset \implies \exists p \in \mathcal{P} \mid {q}^{\text{in}} = {q'}^{\text{in}} = \{p\} \text{ and } F(p,q) = F(p,q')$$

A marked Petri net (N, m_0) is said to be *free-choice* if the associated Petri net N is free-choice.

A Petri net $(\mathcal{P}, \mathcal{Q}, F)$ is said to be ***extended free-choice*** if, for any pair of transitions, either they have no common upstream places, or they have the same set of upstream places with the same valuations:

$$\forall q, q' \in \mathcal{Q}\ ,\ q' \neq q\ ,\quad {q}^{\text{in}} \cap {q'}^{\text{in}} \neq \emptyset \implies \forall p \in \mathcal{P}\ ,\quad F(p,q) = F(p,q')$$

The term "free-choice" is explained by the following property. In an (extended) free-choice net, for any place and marking, either the place's output transitions are all fireable, or none of them are fireable, so that, either one cannot fire an output transition, or one has the "choice" of which transition to fire. On the contrary, in general Petri nets, choosing which output transition to fire may not be "free", as some of the output transitions may not be fireable.

Note that many authors reserve the term "free-choice" to *plain* Petri nets, that is, Petri nets for which F takes values in $\{0, 1\}$ ([DE95], [BW13]). What we call here free-choice nets is called "equal-conflict" nets by other authors (see [TS96]). However, as we are here interested in generalizations of Petri nets, we choose to keep the term "free-choice" in a looser sense. Indeed, we aim at underlying that generalized free-choice nets still convey similar properties as free-choice, plain, discrete Petri nets.

For plain, extended free-choice nets, simple algebraic results allow to characterize liveness, boundedness, and other properties. For example, the Commoner/Hack Criterion characterizes liveness in a free-choice Petri net. The corresponding algorithm is, however, NP-complete. However, the import property of well-formedness of a Petri net (existence of an initial marking such that the marked Petri net is live and bounded) is equivalent to some simple algebraic properties of the incidence matrix of the Petri net, and of polynomial complexity.

The book by Desel and Esparza [DE95] is the classical reference on free-choice nets and extended free-choice nets. We also refer to the overview of [BW13], which includes further, more recent results.

Note also that the analysis is again simplified and more detailed if a Petri net belongs to one of the two well-known subclasses of free-choice nets, *state machines* (also called S-nets in [DE95], where S is the authors' notation for a set of places), and *marked graphs* (T-nets, where T similarly holds for a set of transitions).

2.1.4 Fractioned firings: a relaxation of Petri nets

In a Petri net, a transition firing consumes exactly $(C_-)_{p,q}$ tokens in each upstream place p and produces exactly $(C_-)_{p',q}$ tokens in each downstream place p'. In a ***Petri net with fractioned firings***, this rule on transition firings is relaxed in the following way: a firing is characterized by a transition q and a ***firing rate*** $\alpha \in \mathbb{Q}_{>0}$, where $\mathbb{Q}$ is the set of rational numbers, such that a firing of transition q with rate α consumes a rational number of tokens $\alpha(C_-)_{p,q}$ in each upstream place p and produces a rational number of tokens $\alpha(C_+)_{p',q}$ in each downstream place p'. Consequently, markings become elements of $\mathbb{Q}_{\geqslant 0}^{\mathcal{P}}$. As before, for a transition q to fire at rate α at state m, the marking in the upstream places before the firing is required to be larger than the token consumption, that is, $m \geqslant \alpha(C_-)_q$. If this condition is satisfied, we say that transition q is ***α-fireable*** at state m, and we say that is ***fireable*** at state m if there exists an $\alpha \in \mathbb{Q}_{>0}$ such that it is α-fireable (context should make it clear if a transition is *fireable* for the discrete setting or for the fractioned setting).

In this context, for a marked Petri net (N, m_0), a firing sequence σ is a sequence of pairs (q, α), representing the firing of transition q at rate α, and it is defined by induction in the following way: either it is the empty sequence, and the marking does not change, or it is of size $k \geqslant 1$, its prefix of length $k-1$ is a firing sequence, and from the marking m reached by the prefix, the last pair (q, α) is such that q is α-fireable. We use the notation $m \xrightarrow{\alpha q} m'$, and $m_0 \xrightarrow{\sigma} m'$. Associated to σ, we also define the counterpart of the occurrence count vector:

$$|\sigma| := \sum_{(q,\alpha_q)\in\sigma} \alpha_q e_q \,.$$

Now, $|\sigma|$ takes values in $\mathbb{Q}_{\geqslant 0}^{\mathcal{Q}}$. It satisfies a similar fundamental equation as the original one (2.2), but with rational numbers instead of integers: for any marking $m_0 \in \mathbb{Q}_{\geqslant 0}^{\mathcal{P}}$, and m such that $m_0 \to^{\sigma} m$ with a fractioned firings semantics,

$$m = m_0 + C|\sigma| \,. \tag{2.3}$$

The introduction of these Petri nets dates back to the seminal work of David and Alla [DA87]. It aimed at approximating the behavior of discrete Petri nets, especially in a context where a large number of tokens initially present in the system would lead to a blow up of the reachability set. Indeed, by analogy with the relaxation of integer programming in optimization, such a relaxation was expected to yield simplified behaviors and computations. Note that, because the rate-1 firing of a transition corresponds to its firing in the original integer definition, the reachability set of a marked Petri net with fractioned firings is a superset of its reachability set in the original setting. However, the gap between both models may be very large (examples exist of a discrete marked Petri net encountering a deadlock, while its counterpart with fractioned firings is unbounded).

Most basic properties of Petri nets with fractioned firings, such as the reachability of a given marking, are decidable and can be computed in polynomial time, which illustrates the tractability of the approximation. The work by Fraca and Haddad [FH15] contains the latest results on these issues. Other major results are exposed in [RTS99, JRS03, RHS10]. We note that, in these works, Petri nets with fractioned firings are named *continuous Petri nets* (and the firing rates are reals), but as this terminology is sometimes associated with time interpretations, we prefer using our maybe cumbersome designation to avoid confusion.

Note that, either in the setting of fractioned firings, or in the setting of integer firings, relaxing the incidence matrix entries to be rational numbers does not change the properties of the system: for such a net, the same behavior is obtained by a net in which the entries of the incidence matrix are scaled by an integer such that the new incidence matrix becomes integer-valued, and the initial marking is scaled by the same integer. Attributing rational valuations to arcs of a Petri net may be encountered, for example, when approximating stochastic routing in output of a conflict place (the arc valuations then represent the proportion of tokens routed along the arc).

2.1.5 External inputs and control

In the modeling of a physical system, it may be convenient to account for external arrivals of tokens, which are not related to the inner behavior of the system. This can be done by the means of ***input transitions***, which have no upstream places, but whose firings are an external input. Such models have a control flavor. Given information on the input firings, one would like to assert the stability of the system, the observability of some markings, or to compute some "output" quantities of the system. One may also want to control the system (for example, by enabling or disabling the firing of a transition), in order to ensure stability or to ensure some equations or inequalities on the place markings.

A Petri net with no external input can be named an ***autonomous Petri net***, or a ***closed Petri net*** (a generalization of "closed networks" in queueing theory) In the first part of this thesis, we do not consider control issues or external inputs, so that all our Petri nets are autonomous.

Nevertheless, we point out that Baccelli and Foss [BF95, BFG96] stated the following "saturation rule" theorem: stochastic networks which have the *monotone-homogeneous property* are stable under an arrival process of intensity λ if and only λ is strictly smaller than a constant

which is expressed in terms of the limit when $n \to \infty$ of the date of the last event of the system whose arrival process is a Dirac of size n at time 0. See also Bonald's PhD thesis [Bon99, Chapter 3]. This applies to stochastic, free-choice, consistent Petri nets with an external input. As an approximate rule, we can hence retain that the stability of an autonomous Petri net where some external input has been considered saturated implies the stability of the corresponding non autonomous Petri net.

2.2 Structural analysis of Petri nets

A number of key properties of Petri nets are satisfied by all Petri nets considered in this thesis (including their different extensions). They rely on the fundamental equation of the Petri net, already written in the discrete model (2.2), and in the fractioned firings model (2.3), and on the algebraic properties of its incidence matrix. We present them in this section.

We point out that the following results only hold for autonomous Petri nets, that is, Petri nets in which the firing of transitions is determined by the marking present in the net.

2.2.1 An algebraic characterization of reachability

Let us first define a dead transition for Petri nets with fractioned firings.

▶ **Definition 8.** Let (N, m_0) be a marked Petri net with fractioned firings. We say that transition q is ***dead*** if there does not exist a firing sequence and an $\alpha > 0$ such that (q, α) is an element of the firing sequence. We denote by $\mathcal{Q}^l(m_0)$ the set of transitions in $\mathcal{Q}$ which are not dead for marking m_0.

Of course, a transition which is dead for a marked Petri net with fractioned firings is also dead for the same marked Petri net restricted to discrete firings.

An immediate property is that, if a transition is dead for a given marked Petri net (N, m_0), then it is dead for any reachable marking of m_0. Algorithm 1 of [RTS99] computes the dead transitions of a Petri net (N, m_0) in quadratic time.

If marking m_0 is the initial marking of the Petri net, we also use the notation $\mathcal{Q}^l(0) = \mathcal{Q}^l(m_0)$.

▶ **Proposition 2.2.** *Let (N, m_0) be a marked Petri net (with discrete or fractioned firings), and C its place–transition incidence matrix. Any reachable marking of (N, m_0) belongs to the set $\mathbb{R}^{\mathcal{P}}_{\geqslant 0} \cap \left(m_0 + C\mathbb{R}^{\mathcal{Q}^l(0)}_{\geqslant 0}\right)$.*

Proof. This is a direct consequence of the fundamental equation (2.2) in the discrete case, and (2.3) in the fractioned case. ◀

Any reachable marking is hence in a polyhedron which is the intersection of two affine cones. We name it the ***reachability polyhedron***.

The reachability polyhedron provides information on the reachable markings of a given Petri net semantics. In the case of Petri nets with fractioned firings, it is a rather accurate approximation of the reachability set, as shown by the following proposition.

▶ **Theorem 2.3** ([JRS03])**.** *Let (N, m_0) be a marked Petri net with fractioned firings. The closure of the reachability set of (N, m_0) is equal to its reachability polyhedron.*

▶ **Remark 2.4.** The reachability set of a Petri net with fractioned firings is in general not closed, so that some markings of the reachability polyhedron do not belong to the reachability set. See the example of Recalde *et al.*, [RTS99, Figure 5].

▶ **Remark 2.5** (Pre-processing)**.** In a marked Petri net (N, m_0), dead transitions induce *dead places* of the Petri net, whose markings are and remain null. Therefore, the analysis of the Petri net reduces to the analysis of its non dead part, that is, one considers the Petri net in which the set of transitions is restricted to $Q^l(m_0)$, and in which the set of places is restricted to the ones having a non zero initial marking or a non dead input transition.

2.2.2 Structural properties in the reachability polyhedron

The following basic properties of Petri nets can be understood in a discrete firings setting or in a fractioned firings setting. The notion of reachable marking must be interpreted accordingly.

▶ **Definition 9.** Let N be a Petri net.

- Let m_0 be an initial marking of N. A place p is ***bounded*** if there exists $K{>}0$ such that, for any reachable marking m of m_0, $m_p \leqslant K$.
- A place p is ***structurally bounded*** if, for any initial marking m_0, the place is bounded (there exists K depending only on C and m_0 such that, for any reachable marking m, $m_p \leqslant K$).
- N is ***structurally bounded*** if every place of $\mathcal{P}$ is structurally bounded.
- Let m_0 be an initial marking of N. We say that a marked Petri net (N, m_0) is ***partially consistent*** if m_0 is a reachable marking of m_0, associated with a non empty firing sequence.
- If, in this firing sequence, every transition is fired at least once, then the marked Petri net is said to be ***consistent***.
- N is ***structurally consistent*** if there exists an initial marking m_0 such that the marked net (N, m_0) is consistent.

We have the following well-known algebraic characterizations of boundedness and consistency in a Petri net with fractioned firings:

▶ **Proposition 2.6.** *Let N be a Petri net.*

(i) If there exists $y \in \mathbb{R}^{\mathcal{P}}_{\geqslant 0}$ such that $y_p > 0$ and $y^{\mathsf{T}}C \leqslant 0$, then place p is structurally bounded.

(ii) If there exists $y \in \mathbb{R}^{\mathcal{P}}, y > 0$ such that $y^{\mathsf{T}}C \leqslant 0$, then the Petri net is structurally bounded.

(iii) If there exists $y \in \mathbb{R}^{\mathcal{P}}_{\geqslant 0}$ such that $y^{\mathsf{T}}C \lneq 0$, then the Petri net is not consistent.

Actually, the converse holds for each statement. However, from Section 2.4 on, we consider restricted Petri net semantics, such that the set of reachable markings is a strict subset of the set of reachable markings. Therefore, we are only interested in the if-part of these statements, which holds for any such semantics.

Proof. (i) For such y and p, we have for every reachable marking m of (N, m_0) the identity $y^{\mathsf{T}}m_0 = y^{\mathsf{T}}m + y^{\mathsf{T}}Cx$, so that $y^{\mathsf{T}}m \leqslant y^{\mathsf{T}}m_0$. Therefore, $m_p \leqslant (y^{\mathsf{T}}m_0)/y_p$.

(ii) If y is positive, then for any place p, (i) holds.

(iii) If the net is consistent, then m_0 is a reachable marking of m_0 with a firing sequence such that every transition is fired at least once. By the fundamental equation, this implies that there exists an $x > 0$ such that $m_0 = m_0 + Cx$, so that $Cx = 0$. By Stiemke's theorem (which is a variant of Farka's lemma), exactly one of the two alternatives is true: there exists y such that $y^{\mathsf{T}}C \lneq 0$, and there exists $x > 0$ such that $Cx = 0$. ◀

Note that the above proposition is also true for the discrete Petri net model of Section 2.1, with integer-valued vectors y. This is because Farka's lemma and comparisons to zero also hold with integer-valued vectors. Moreover, for such Petri nets, the converse of (i), (ii), (iii) is also true. See Murata [Mur89, Section VIII].

From Proposition 2.6, we deduce that a well-chosen vector of $\mathbb{R}^{\mathcal{P}}_{\geqslant 0}$ can provide a simple Lyapunov function for the dynamics of our Petri net:

▶ **Corollary 2.7.** *Let (N, m_0) be a marked Petri net. Let $y \in \mathbb{R}^{\mathcal{P}}_{\geqslant 0}$ be such that $y^{\mathsf{T}}C \leqslant 0$, and $(y^{\mathsf{T}}C)_q < 0$ for some transition q. If m is a reachable marking, associated with $x \geqslant 0$, and if $x_q > 0$, then $y^{\mathsf{T}}m < y^{\mathsf{T}}m_0$.*

In fact, from a geometric point of view, a $y \neq 0$ such that $y^{\mathsf{T}}C \leqslant 0$ is the normal of a supporting hyperplane $y^{\mathsf{T}}m = y^{\mathsf{T}}m_0$ of the reachability polyhedron of a Petri net. If, moreover, $y > 0$, then the intersection of the half-space $y^{\mathsf{T}}m \leqslant y^{\mathsf{T}}m_0$ with $\mathbb{R}^{\mathcal{P}}_{\geqslant 0}$ is bounded, (one can say that the supporting hyperplane "separates the reachability polyhedron from infinity"). Note also that, if there exists a $y \neq 0$ such that $y^{\mathsf{T}}C = 0$, then the reachability polyhedron is included in the supporting hyperplane. See two possible configurations in Figure 2.2.

A vector satisfying these properties has an important role in algebraic analysis of Petri nets.

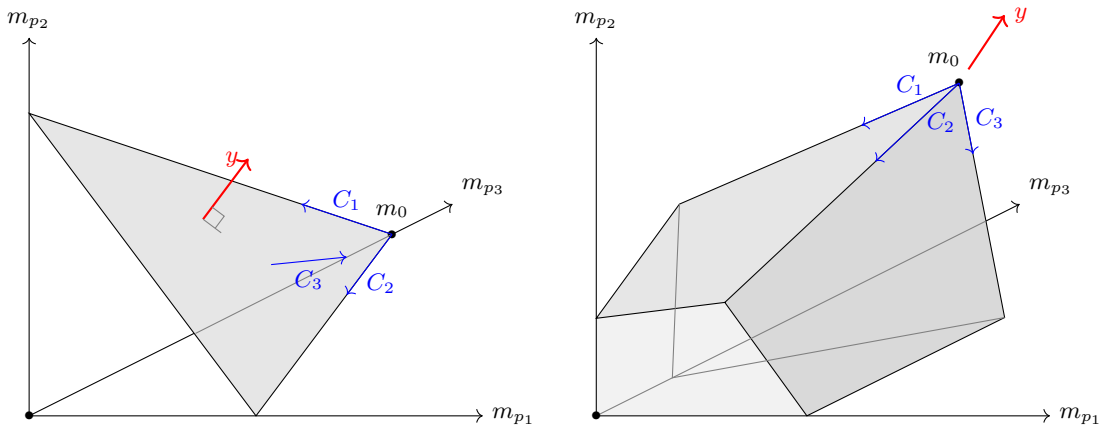


Figure 2.2 – Two bounded reachability polyhedra associated with two different marked Petri nets (with three places, so that the reachability polyhedra are subsets of $\mathbb{R}^3$). Left: there is a linear relationship between the transition vectors. The system is conservative (the vector y is such that the quantity $y^T m$ is invariant). Right: any transition firing is such that $y^T m$ decreases.

▶ **Definition 10.** Let N be a Petri net.

- A $\mathcal{P}$***-invariant*** is a vector $y \in \mathbb{R}^{\mathcal{P}}$, $y \neq 0$, such that $y^{\mathsf{T}} C = 0$.
- A $\mathcal{Q}$***-invariant*** is a vector $x \in \mathbb{R}^{\mathcal{Q}}$, $x \neq 0$, such that $Cx = 0$.

The term "invariant" comes from the fact that, if y is a $\mathcal{P}$-invariant of N, then the quantity $y^{\mathsf{T}} m$ is invariant among the set of reachable markings of (N, m_0).

Note that in the definition of a $\mathcal{P}$-invariant of a discrete Petri net, the vector is restricted to be integer-valued. However, any $\mathcal{P}$-invariant of a Petri net with C having integer entries is proportional to a rational-valued $\mathcal{P}$-invariant, and hence to an integer-valued $\mathcal{P}$-invariant.

The structural consistency of a Petri net is equivalent to the existence of a positive $\mathcal{Q}$-invariant:

▶ **Proposition 2.8.** *A Petri net N is structurally consistent if and only if it admits a positive $\mathcal{Q}$-invariant.*

Proof. The "only if" part is a direct consequence of the definition of a consistent net. For the "if" part, it suffices to remark that, with a positive initial marking $m_0 > 0$, every transition is active, so that $\mathcal{Q}^l(m_0) = \mathcal{Q}$. For x a positive $\mathcal{Q}$-invariant, and $\alpha > 0$ small enough, we have $m_0 = m_0 + \alpha C x$, and any sequence of firings of the different $(\alpha x_q, q)$ is fireable, so that m_0 is reachable from m_0 and the consistency holds. ◀

The notion of conservative Petri nets is associated with the $\mathcal{P}$-invariants of the net:

▶ **Definition 11.** A Petri net N is ***conservative*** if it admits a positive $\mathcal{P}$-invariant.

2.2.3 Minimal-support invariants of a Petri net

If x is a vector indexed by a set E, its ***support*** is defined by $[x] := \{e \in E \mid x_e \neq 0\}$. A ***minimal-support $\mathcal{P}$-invariant*** of N is a nonnegative P-invariant of N whose support is minimal for the inclusion. Of course, if y is a minimal-support $\mathcal{P}$-invariant of N, so is any λy, with $\lambda > 0$. It is a well-known fact that minimal-support $\mathcal{P}$-invariants generate all nonnegative $\mathcal{P}$-invariants of N:

▶ **Theorem 2.9** (Theorem 2.43 of [DE95])**.** *Every nonnegative $\mathcal{P}$-invariant is the sum of minimal-support $\mathcal{P}$-invariants.*

In fact, from a linear algebra point of view, this amounts to saying that the cone of the nonnegative vectors of $\ker(C^{\mathsf{T}})$ the kernel (or null space) of matrix C^{T} is generated by its vectors of minimal support (thus, these vectors are the *extreme rays* of the cone).

Minimal-support $\mathcal{P}$-invariants are of special interest, as they allow to compute lower bounds on the throughput of the system in a timed semantics. See for example Proposition 4.13. Their enumeration is however complicated: it reduces to computing the set of extreme rays of a cone defined by a set of linear inequalities.

▶ **Lemma 2.10** (Counting the minimal-support $\mathcal{P}$-invariants)**.** *The number of minimal-support nonnegative $\mathcal{P}$-invariants of a Petri net with incident matrix C is upper bounded by the quantity*

$$U(n,k) = \binom{n - \lfloor (n-k)/2 \rfloor}{k+1} + \binom{n - \lceil (n-k)/2 \rceil}{k+1}, \tag{2.4}$$

where $n = |\mathcal{P}|$ and $k = \mathrm{rank}(C)$.

The notation $\lfloor x \rfloor$ for $x \in \mathbb{R}$ stands for the largest integer k s.t. $k \leqslant x$. The notation $\lceil x \rceil$ stands for the smallest k s.t. $k \geqslant x$.

Proof. Let C be the incidence matrix of the Petri net, and $A = C^{\mathsf{T}}$. It will be convenient to define $n := |\mathcal{P}|$. We want to compute a minimal family of nonnegative elements of $\ker(A) \cap \mathbb{R}^n$ generating all the elements of this set by positive linear combinations. We first remark that this set is a polyhedral cone described by inequalities, and that this problem amounts to computing the extreme rays of the cone.

Without loss of generality, we can suppose that A it is full rank, that is, its number of rows is equal to its rank k. Indeed, if it is not, one can select k independent rows in A, $R^1(A), \ldots, R^k(A)$ such that $Ay = 0 \Leftrightarrow (R^1(A), \ldots, R^k(A))^{\mathsf{T}} y = 0$.

If $k = n$, then the kernel of A is reduced to $\{0\}$, and it has no positive invariant. Let us assume $k < n$. We re-order A in such a way that the k first columns of A form an independent family of vectors, that is, $A = (A_1, A_2)$ with A_1 nonsingular. We can similarly re-order y in two parts $y^{\mathsf{T}} = (y_1^{\mathsf{T}}, y_2^{\mathsf{T}})$. It is equivalent that $Ay = 0$ and that $y_1 = -A_1^{-1} A_2 y_2$. Therefore, the set of nonnegative elements of the kernel of A has the following equivalent representation in the smaller vectorial space $\mathbb{R}^{n-k}$.

$$\{x \in \mathbb{R}^{n-k} \mid x \geqslant 0, A_1^{-1} A_2 x \leqslant 0\}.$$

Moreover, computing the extreme rays of this cone amounts to computing the vertices of the polytope generated by its intersection with the hyperplane given by the equality $\sum_i x_i = 1$:

$$\{x \in \mathbb{R}^{n-k} \mid x \geqslant 0, A_1^{-1} A_2 x \leqslant 0, \sum_{i=1}^{n-k} x_i = 1\},$$

which itself admits an equivalent representation in the smaller vectorial space $\mathbb{R}^{n-k-1}$ by replacing the last coordinate by its expression $x_{n-k} = 1 - \sum_{i=1}^{n-k-1} x_i$:

$$\{x \in \mathbb{R}^{n-k-1} \mid x \geqslant 0, A_1^{-1} A_2 x \leqslant 0, 1 - \sum_{i=1}^{n-k-1} x_i \geqslant 0\}.$$

The number of vertices of this polytope equals the number of extreme rays of the cone $\ker(A) \cap \mathbb{R}^n_{\geqslant 0}$. An upper bound to this number is given by McMullen's Upper Bound Theorem, see [McM70]. Note that this polytope has at most $(n-k-1) + k + 1 = n$ facets, in dimension $n - k - 1$. Relation (2.4) follows from this upper bound, where we have used the identities $\lceil (n-k-1)/2 \rceil = \lfloor (n-k)/2 \rfloor$ and $\lceil (n-k-1)/2 \rceil + 1 = \lfloor (n-k)/2 \rfloor$. ◀

To the best of our knowledge, the best upper bound which was known until now, given by Silva *et al.*, (see for example [SCC92]), was

$$\binom{|\mathcal{P}|}{\lfloor |\mathcal{P}|/2 \rfloor}. \tag{2.5}$$

It relies on the fact that the support of two minimal-support nonnegative invariants are not comparable for the inclusion relation, so that an upper bound is the maximal length of an antichain in the graph of the inclusion relations for a set of size $|\mathcal{P}|$. This, however, does

not take advantage of the informations on the rank of the incidence matrix, and yields looser bounds. As a simple example, for a matrix C with 10 rows and rank 7, the upper bound given by (2.5) is 252, while the upper bound given by (2.4) is 18. In fact, the largest bounds in (2.4) are obtained when $k \sim n/3$. Yet, for a matrix C with 12 rows and rank 4, the upper bound given by (2.5) is 924, while we obtain 112 with (2.4). Note also that McMullen proved his upper bound to be tight for general polytopes.

Algorithms to enumerate the minimal-support $\mathcal{P}$-invariants of a Petri net can be derived from the polytope representation given in the proof of Lemma 2.10. Known algorithms are the reverse-search algorithm of Avis and Fukuda, see [Avi00], or the double-description method of Motzkin ([MRTT53]), for which we refer to Fukuda and Prodon [FP96].

Computing minimal-support invariants of Petri nets has historically relied on "Farkas' algorithm", see [Tre88] and references therein. The work of Colom and Silva [CS91] is, to our knowledge, the only one to identify the relationship with extreme rays of a cone and with Motzkin's method. This reference includes a comprehensive comparative study of algorithms for computing the minimal invariants.

2.3 Timed semantics of Petri nets

In this thesis, the circulation of tokens in a Petri net is restricted to follow some time constraints. Introducing time in Petri nets fundamentally changes the nature of analyses carried out. Instead of analyzing properties of the reachable markings, one focuses on performance issues, such as lower and upper bound for event times or route durations, or asymptotic throughputs.

Still, the structural and untimed analysis of Petri nets is essential for their timed analysis.

2.3.1 Place and transition durations in Petri nets

In a (classical) Petri net, a given execution of a Petri net can be characterized by a sequence of markings $(m(k))_{k\in K}$, where $K = \{0, 1, \ldots\}$, possibly infinite, where exactly one transition firing occurs between steps k and $k+1$. This is what we call a ***discrete-step*** execution. Introducing time in Petri nets amounts to considering executions of the form $(m(t))_{t\in\mathbb{R}_{\geqslant 0}}$. We remark that this includes situations in which the marking execution $(m(t))$ is fully characterized by its value at discrete time steps. In such situations, time can be modeled as an integer variable $t \in \mathbb{N}$.

The relationship and consistency between a timed execution of a Petri net and the previous untimed semantics is given by the following property:

$$\text{for any } s, t \in \mathbb{R}_{\geqslant 0},\ s < t,\ m(t) \text{ is a reachable marking of } m(s)\,. \tag{2.6}$$

Of course, depending on the model, "reachable" can also be interpreted as for a fractioned firing execution. In other words, a timed execution of a Petri net should follow the same kind of trajectories than an untimed execution. One does simply measure the durations of the steps of this untimed execution.

Typically, in a time semantics, a token is required to remain in a place at least a certain amount of time (named "holding duration"), or the firing of a transition is required to last a certain amount of time (named "firing duration"). A more complex setting is the following, which introduces a timed concurrency between transitions: when a transition becomes fireable, this triggers a clock, and the transition fires if it remains fireable until the clock reaches a certain duration ("enabling duration". The terminology comes from the expression *enabled transition*, which is a synonym for fireable transition). Whatever the type of duration implemented, the specification of the durations may also vary: a duration can be given as constant for each place or transition, or varies in a fixed interval, or may be given by a sequence specifying the duration of each successive event (one duration for each new token entering the place, or for each firing of a transition).

Durations can also be specified as random variables, having known stochastic distributions: it is usually required that the corresponding random variables are independent, and that they are identically distributed for each place or transition. In this case, we speak of a ***stochastic timed Petri net***. Note that the term "stochastic Petri net" was reserved for memoryless time distributions (exponential or geometric) in the initial work of Molloy [Mol82].

Because the markings take integer values (or rational values), a time execution $(m(t))$ is a piecewise constant function of time. It will be typically required to be a *càdlàg* function, that is, a function which is right continuous and has a left limit at any time. If $(t_n)_{n\in\mathbb{N}}$ is the increasing sequence of times where a jump of some component of $(m(t))$ occurs, then, the sequence $(m(t_n))_{n\in\mathbb{N}}$ should be required to be such that, for any n, $m(t_{n+1})$ is a reachable marking of $m(t_n)$. However, at a given jumping time t_n, several events may happen. Therefore, a timed semantic should specify how such events interact at a given time. For example, if two transitions become fireable at time t_n, the firing of one can prevent the firing of the other one. Besides, a timed semantics could lead to a *Zeno behavior*, that is, an infinite sequence of jumping times $(t_n)_{n\in\mathbb{N}}$ that has a finite limit. A Zeno behavior of another, more trivial kind occurs if tokens encounter holding (or firing) durations of value 0 in a cycle: this would lead to infinite cycling at a fixed time. Durations of value 0 along a cycle of places and transitions are, in general, forbidden (or the probability of this event should be 0).

In timed semantics in which durations are independent, memoryless random variables, and token routing is deterministic or Bernoulli, it can be shown that the Petri net has the Markov property. Moreover, with probability one, at most one transition fires at any time. We speak of a ***Markovian Petri net***. The processes of the marking states $(m(t))_{t\geqslant 0}$ are then a continuous-time Markov chain. This yields instructive results on the marking executions.

These different notions of time in Petri nets have led to a large variety of models, although a number of them were shown to have the same modeling power. Merlin and Farber's model [MF76] is recognized as the first model extending Petri nets by a time notion. A detailed review and analysis on the role of time in (non stochastic) Petri nets is the one of Bowden [Bow00]. See further references therein. A reference on stochastic timed extensions of Petri nets is the book of Marsan *et al.* [MBC+94].

Petri nets with a time semantics raise completely different issues and analyses than the original model. In a timed system, one would like to compute quantities such as waiting delays, reaching times and path journey delays, or throughputs at transitions, or to provide lower or upper bounds to these quantities. The set of states that can be reached in a timed semantics may be smaller than the reachability set of the corresponding untimed Petri net, because some sequences of firings would not be compatible with the given time constraints.

2.3.2 Differential Petri nets

Consider a Petri net with non integer firings of transitions and a timed semantics. If, in a marking execution, the marking jumps are of weak amplitude and are separated by small time intervals, it would look natural to consider the limit of such marking when both this jump amplitude and time interval length go tend to zero. This implies that the markings become a continuous function of time. In some cases, it can be shown that the markings are solutions of an ordinary differential system (or of a differential system with discontinuous righthandside). We use the loose term ***differential Petri nets*** to designate such systems.

The first introduction of differential Petri nets was proposed by David and Alla [DA87] in the very same work that introduced Petri nets with fractioned firings. Indeed, solutions of differential equations appear as the natural timed semantics for Petri nets with fractioned firings. The two authors soon proposed two different firing semantics, a "constant speed continuous Petri net" and a "variable speed continuous Petri net". The quality of both approximations for various classes of Petri nets was an important question in the years 1990-2000. It is addressed for example in [MRS06] for the class of live, consistent, conservative, connected Petri nets.

It was underlined later on that such models could also be understood as a *fluid approximation* (and sometimes as a *fluid limit*) of Petri nets with stochastic times. To our knowledge, the first reference pointing out the relationship between fluid limits of queueing networks and "fluidization" of Petri nets is [BGM98]. Since then, it became a known fact in this track of research, see for example [RS00] and [VMJS13].

Observe that a differential Petri net can be obtained as the limit of a Petri net with a discrete-event semantics. It is not clear however if, at this limit, the continuous time execution still follows the important property (2.6). We did not find any proof of this nature in the literature, in a differential setting.

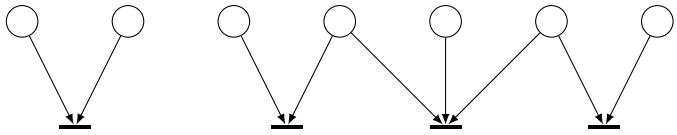


Figure 2.3 – Two clusters and their upstream places. Left: a trivial cluster, reduced to a single transition. Right: a more elaborate cluster.

2.4 Routing rules in Petri nets

In general, for a given marking in a Petri net, several transitions are fireable, but the firing of one of them usually makes some others not fireable afterwards: a Petri net has in general not the *persistency* property. Moreover, one firing of transition may yield significant changes on the reachability set, for example, it may lead to a deadlock in the discrete setting, or some transitions may become dead in the fractioned firings setting. In fact, if the reachability sets of markings of a Petri net are forward invariants, they still encounter branching at some states. The analysis and problems we described in Section 2.1 and 2.2 mostly consisted in characterizing the behavior of the Petri net under arbitrary firing sequences of transitions, regardless of the branchings.

Introducing time in Petri net, as is developed in Section 2.3, diminishes the number of potential branchings, because the different durations of paths in the Petri nets forbids some markings that were reachable in the untimed setting.

In this section, we consider another category of restrictions on the firing sequences of transitions that eliminate the potential branching of the reachability set induced by the choice between several transitions fireable in output of a place. We call these restrictions ***routing rules***, or ***routing policy***.

2.4.1 Introducing routing rules

A routing rule is meant to resolve conflicts occurring in clusters:

▶ **Definition 12.** Let N be a Petri net.

- We say that two transitions q and q' are ***in structural conflict*** if they have common upstream places, that is, if $q^{\text{in}} \cap {q'}^{\text{in}} \neq \emptyset$.
- Let us denote $q \sim_c q'$ if q and q' are in structural conflict. It is a reflexive and symmetric relation. Its transitive closure induces a partition of $\mathcal{Q}$ in equivalence classes. The equivalence class of transition q is denoted $\text{cl}[q]$, and it is named the ***cluster*** of q.

An example of clusters is given in Figure 2.3.

We extend the notion of set of backward or forward neighbors of a node, introduced in Section 2.1.1, to a set of node, that is, if $X \subseteq \mathcal{P} \cup \mathcal{Q}$, $X^{\text{in}} = \cup_{x \in X} x^{\text{in}}$ and $X^{\text{out}} = \cup_{x \in X} x^{\text{out}}$. In this way, we can denote the upstream set of a cluster of a transition q as $\text{cl}[q]^{\text{in}}$. The partition of $\mathcal{Q}$ into clusters induces a partition of $\mathcal{P}$ into upstream sets of clusters.

Routing rules require us to distinguish between *enabled* transitions and *fireable* transitions. In a context with routing rules, we say that a transition is ***enabled*** if it is fireable for the corresponding untimed Petri net (without routing rules). A transition is ***fireable*** if it is enabled and if, moreover, its firing respects the routing rules given in the Petri net.

We say that a transition firing ***disables*** a transition q if q was enabled before this firing and is not enabled afterwards.

▶ **Lemma 2.11.** *Let q, $q' \in \mathcal{Q}$. If $\text{cl}[q] \neq \text{cl}[q']$, and if q and q' are enabled, then the firing of q does not disable q'.*

Proof. After the firing of q, the marking strictly decreases only in places of q^{in}, which is a subset of $\text{cl}[q]^{\text{in}}$. As $\text{cl}[q]^{\text{in}} \cap \text{cl}[q']^{\text{in}} = \emptyset$, the marking of the upstream places of q' does not decrease, so that q' remains enabled. ◀

▶ **Definition 13.** Let (N, m_0) be a marked Petri net. If, with marking m_0, transitions q and q' are enabled, but if the firing of one disables the other, we say that there is a ***conflict*** between q and q'.

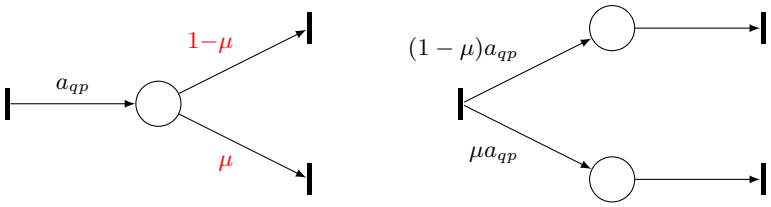


Figure 2.4 – Fluid approximation of a Bernoulli routing. Left: the tokens are routed with probability μ to the lower transition, and with probability $1-\mu$ to the upper transition. Right: the place is duplicated, and only one transition is available in output of each of both places. The valuations of the input arcs of the places are multiplied by a coefficient μ (or $1-\mu$).

Remark that a conflict situation may be asymmetric, that is, the firing of q may disable q', while the firing of q' does not disable q.

A *routing rule* resolves a conflict, by choosing which of the transitions to fire in a situation of conflict. Setting routing rules for all the clusters of a Petri net yields a *single-valued* dynamics of the Petri net: for a given realization of the random quantities, there is a unique possible state evolution in the net.

The following routing rules are standard in Petri net semantics (see for example [BGM06]):

Pre-allocation routing A token entering in a place is allocated to an output transition, regardless of which transitions are enabled or disabled (if the token does not enable the selected output transition, it waits until the transition becomes fireable). One associates with each place p a function $A_p : \mathbb{N} \to p^{\text{out}}$ determining the output transition of the n-th token entering the place. This function can also depend on which input transition brought the token in the place, or on the date of entrance of the token, in the case of a timed Petri net.

Bernoulli routing In this case, the output transition of a token in a place is a random variable. For different tokens or different places, these random variables are independent. They are also identically distributed for each place, with a fixed probability associated with each output transition.

Race policy This kind of routing holds for timed Petri nets with *enabling durations*, that is, a transition can fire only after it remains enabled during a given time. In this case, if, moreover, one requires that a transition fires as soon as possible, there is a *race* between the different output transitions of a place: the first transition that becomes fireable fires, and therefore, disables transitions with which it is in conflict. One speaks of a *race policy* in the case when the enabling time is given by a Poisson distribution. The parameter of this distribution depends only on the transition.

We call ***choice-free*** the Petri nets with no structural conflict, that is, each place has a single output transition. In these nets, routing is always trivial. Choice-free Petri nets are analyzed in Teruel and Silva [TCS97]. They are a subclass of free-choice Petri nets.

▶ **Remark 2.12** (Fluid approximation of Bernoulli routing). The following approximation is standard for systems with routing involving probability distributions: instead of routing a token with probability μ to transition q_1 and $1-\mu$ to transition q_2, send a fraction μ of the token to transition q_1 and $1-\mu$ to transition q_2. This is what we call a ***fluid approximation***. The marking becomes a continuous value, and fractions of tokens circulate in the Petri net.

This also corresponds to transforming a cluster with structural conflict into a choice-free configuration, see Figure 2.4.

If we apply this routing to a whole Petri net, we transform a Petri net into a choice-free Petri net, in which all the routing is trivial. Gaujal and Giua [GG04b] show that the analysis of the corresponding *stationary routing continuous Petri net* is much easier. In particular, the asymptotic throughputs can be computed by linear programming (for a time semantics with fixed holding times in each place).

2.4.2 Priority routing

The notion of priority routing that we consider in this dissertation is restricted to clusters. In a cluster with priority routing, a situation of conflict between two transitions is always arbitrated in favor of the same transition: this transition has priority over the other one.

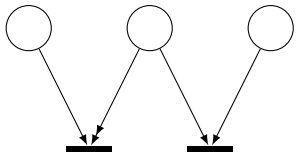


Figure 2.5 – A simple cluster with a priority routing.

More formally, in a cluster cl[q] of size c, the transitions are ranked according to a total ordering, $q_1 < q_2 < \cdots < q_c$. The smallest transition q_1 has the highest priority (we need it to appear before the other ones in our order), and the largest transition q_c has the lowest priority. A ***priority routing*** of a token entering in an upstream place of this cluster corresponds to assigning this token (after a potential holding time) to the transition with highest priority among the transitions it enables. In other words, a transition in a cluster is fireable only if it is enabled, and if no transition of the cluster with higher priority is enabled.

Graphically, the priority ranking of transition in a cluster is (partially) represented by multiple arrows on place-transition arcs: the more chevrons there are on a place-transition arc, the higher the priority of the transition is, among the output transitions of the upstream place.

In the following chapters, our analysis of priority routing will focus on elementary situations, in which only two transitions are in a structural conflict, like the one in Figure 2.5. More complicated cluster configurations can be solved with the same kind of analyses, up to considering more elaborate equations.

▶ **Remark 2.13** (Modeling power)**.** Authorizing multiple priorities in a cluster has the modeling power of a Turing machine. The proof consists in transforming a Petri net with priorities in a Petri net with *inhibitor arcs*, see the construction of Chiola, Donatelli, and Franceschinis [CDF91], which applies to our local priority rule. A place–transition inhibitor arc does not send tokens to the downstream transition, but disables it as soon as its upstream place is non empty. It was shown by Peterson [Pet81] that Petri nets with inhibitor arcs have the computation power of a Turing machine.

Solving the emergency physician paradox We propose a construction, involving Bernoulli routing and priority routing, modeling the emergency physician routing behavior in a Petri net.

Suppose that we have the configuration in Figure 2.6(a): transitions q_1 and q_2 are in structural conflict, because they share the same upstream place p_0. We want to model the following routing rule: if a new token in place p_0 enables both transitions q_1 and q_2, then it is routed according to a random draw (assigning probability $1/2$ to each choice). Otherwise, the token is routed towards the transition it enables, if any.

Our construction in Figure 2.6(b), which involves Bernoulli routing downstream place p_0, and priority routing for cluster $\{q_1, q_2, q_1', q_2'\}$, corresponds to this routing rule. Indeed, if a token enters place p_0, it is allocated to transition q_a with probability $1/2$, and to transition q_b otherwise. As transitions q_1 and q_2 have priority, if both p_1 and p_2 are non empty, then the token that entered place p_0 enters place p_3 with probability $1/2$ and place p_4 with probability $1/2$. Otherwise, suppose for example that place p_2 is empty. Then, the token is routed towards place p_3, either by the firing of transition q_1, or by transition q_1', depending on the place towards which it was routed, p_a or p_b. In any case, if exactly one of the places p_1 or p_2 is empty, the token does not stay waiting, but is routed towards the other part of the network.

The methods developed in Chapters 3 and 4 allow one to compute the stationary regimes of Petri nets with priority routing. One can show that the expected asymptotic throughput (1.1) is reached for the Petri net of Figure 1.3 transformed with this construction.

2.5 Three Petri net examples

The two Petri nets with priorities introduced in this chapter serve as running examples for Chapter 3 and Chapter 4. The first one is a simplified model of the emergency call center mentioned in the introduction. The second one is a well-known system in the literature on Petri nets.

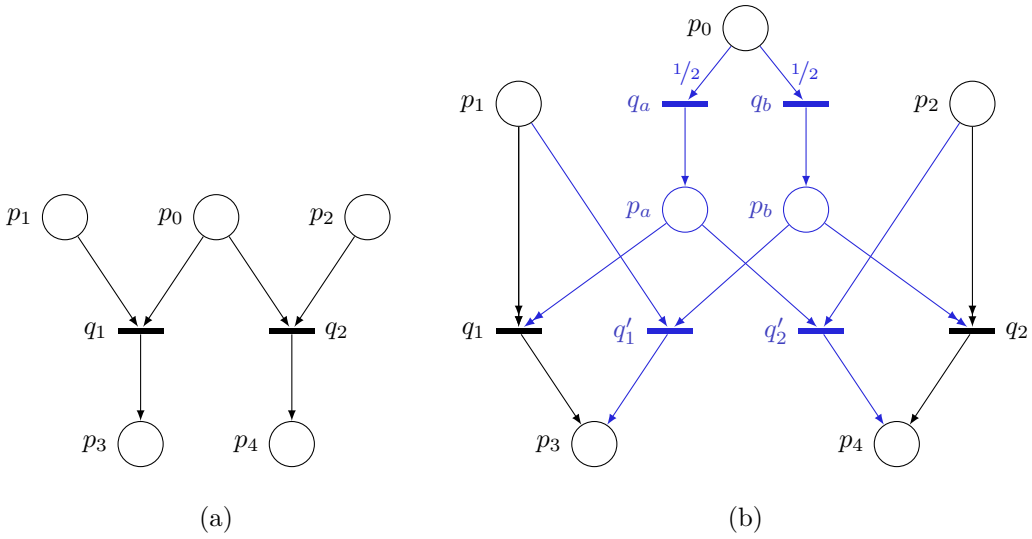


Figure 2.6 – A cluster configuration (a) transformed in (b) to account for the "emergency physician" routing rule. Transitions q_1 and q_2 have priority over transitions q'_1 and q'_2. In the emergency physician Petri net of Figure 1.3, place p_0 would correspond to place p_m.

The third Petri net presented in this section is a simplification of the first one, used in Chapter 5.

2.5.1 A simplified Petri net model of a two-level emergency call center

In this section, we describe a call center answering emergency calls according to the two level instruction procedure developed in Section 1.1. In the new organization planned by PP together with BSPP [RR15], the emergency calls to the police (number 17), to the firemen (18), and untyped emergency calls (European number 112) are dealt with according to a unified procedure, allowing a strong coordination. Another important feature of this organization is that it involves a two level treatment. The present example is designed, so as to analyzing the key features of this new architecture, the priority routing of extremely urgent calls, and the blocking of level-1 operators when level-2 operators are busy, while keeping the model as simple as possible. Hence, we discuss a simplified model, for academic purposes.

The first level operators filter the calls and assign them to three categories: extremely urgent (potentially life threatening situation), urgent (needing further instruction), and non urgent (e.g., call for advice). Non-urgent calls are dealt with entirely by level 1 operators. Extremely urgent and urgent calls are passed to level 2 operators. An advantage of this procedure lies in robustness considerations. In case of events generating bulk calls, the access to level 2 experts is protected by the filtering of level 1. This allows for better guarantees of service for the extremely urgent calls. Every call qualified as extremely urgent generates a 3-way conversation: the level 1 operator stays in line with the calling person when the call is passed to the level 2 operator. Proper dimensioning of resources is needed to make sure that the synchronizations between level 1 and level 2 operators created by these 3-way conversations do not create bottlenecks. We focus on the case where the system is saturated, that is, there is an infinite queue of calls that have to be handled. We want to evaluate the performance of the system, *i.e.* the throughput of treatment of calls by the operators.

The call center is modeled by the timed Petri net of Figure 2.7. We use the convention that all transitions can be fired instantaneously. Holding times are attached to places.

Let us give the interpretation in terms of places and transitions. The number of operators of level 1 and 2 is equal to N_1 and N_2, respectively. The marking in places p_1 and p_2, respectively, represents the number of idle operators of level 1 or 2 at a given time. In particular, the number of tokens initially available in places p_1 and p_2 is N_1 and N_2. The initial marking of other places is zero. A firing of transition q_1 represents the beginning of a treatment of an incoming emergency call by a level 1 operator. The arc from place p_1 to transition q_1 indicates that every call requires one level 1 operator. The routing from transition q_1 to transitions q_2, q_3, q_4

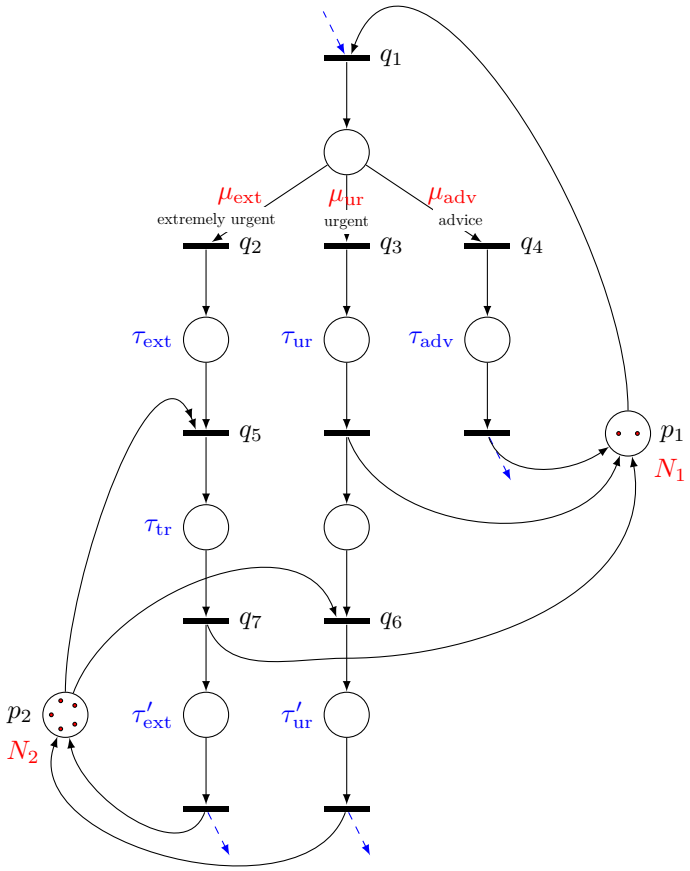


Figure 2.7 – Simplified Petri net model of the Parisian 17-18-112 emergency call center. Blue arrows do not belong to the Petri net and symbolize the entrance and exit of calls in the system.

represents the qualification of a call as extremely urgent, urgent, or non urgent (advice). The proportions of these calls are denoted by μ_{ext}, μ_{ur}, and μ_{adv}, respectively, so that $\mu_{\text{ext}} + \mu_{\text{ur}} + \mu_{\text{adv}} = 1$. The proportions are known from historical data. The instruction of the call at level 1 is assumed to take a deterministic time τ_{ext}, τ_{ur}, or τ_{adv}, respectively, depending on the type of call.

After the treatment of a non urgent or urgent call at level 1, the level 1 operator is made immediately available to handle a new call. This is represented by the arcs leading to place p_1 from the transitions located below the places with holding times τ_{ur} and τ_{adv}. Before an idle operator of level 2 is assigned to the treatment of an urgent call, which is represented by the firing of transition q_6, the call is stocked in the place located above q_6. In contrast, the sequel of the processing of an extremely urgent call (transition q_5) requires the availability of a level 2 operator (incoming arc $p_2 \to q_5$) in order to initiate a 3-way conversation. The level 1 operator is released only after a time τ_{tr} corresponding to the duration of this conversation. This is represented by the arc $q_7 \to p_1$. The double arrow depicted on the arc $p_2 \to q_5$ means that level 2 operators are assigned to the treatment of extremely urgent calls (if any) in priority. The holding times τ'_{ext} and τ'_{ur} represent the time needed by a level 2 operator to complete the instruction of extremely urgent and urgent calls respectively.

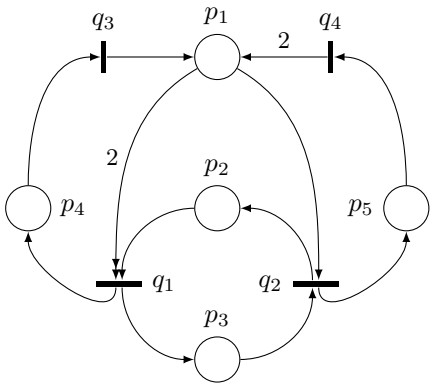


Figure 2.8 – The SR Petri net. Place p_1 is subject to priority.

2.5.2 The Petri net of Silva and Recalde (2002)

The first occurrence of this Petri net we are aware of is in Silva and Recalde [SR02, Fig.14]. Therefore, in the following, we refer to it as the SR Petri net. This Petri net example appeared many times since then, first as an example in which the throughput of the continuous model is lower than the throughput of the discrete model, and more recently as an example in which the asymptotic throughputs, expressed as functions of the firing rates or of the initial markings, encounter discontinuities and non monotonicities [JRS05, Mey12, NGRTS16]. It is of interest to notice that these phenomena appear, even in a consistent, conservative Petri net (but not free-choice).

In this Petri net, place p_1 is subject to a conflict between downstream transitions q_1 and q_2. In all the references cited above, the downstream routing at place p_1 encounters race policy. Here, we implement priority routing, and assign priority to transition q_1.

The incidence matrix of the Petri net is given by

$$C = \begin{pmatrix} -2 & -1 & 1 & 2 \\ -1 & 1 & & \\ 1 & -1 & & \\ 1 & & -1 & \\ & 1 & & -1 \end{pmatrix}.$$

It admits one $\mathcal{Q}$-invariant and two $\mathcal{P}$-invariants:

$$x = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}^{\mathsf{T}}$$
$$y_1 = \begin{pmatrix} 1 & 0 & 1 & 1 & 2 \end{pmatrix}^{\mathsf{T}}$$
$$y_2 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \end{pmatrix}^{\mathsf{T}}$$

These two $\mathcal{P}$-invariants are also the two minimal $\mathcal{P}$-invariants of the net. Their sum $y_1 + y_2$ is a positive invariant, so that the system is conservative: the weighted sum of markings of the net is a constant.

2.5.3 The equivalent Petri net model of Chapter 5

In the stochastic model of Chapter 5, we, again, simplify our emergency call center model, by considering a unique class of urgent calls, all being kept in line by level 1 operators until a level 2 operator handles them.

This corresponds to the Petri net of Figure 2.9.

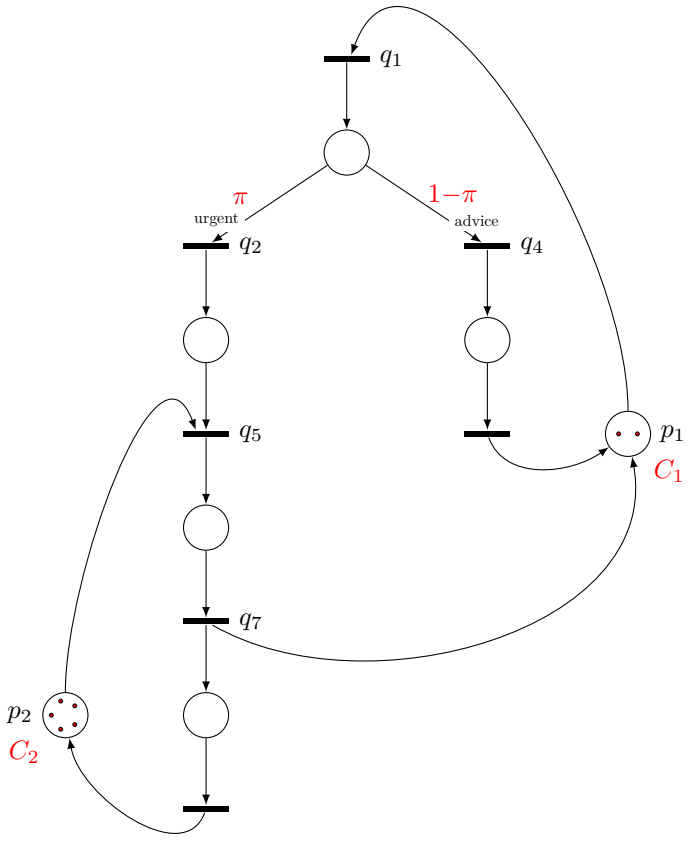


Figure 2.9 – The two-class model of Chapter 5. This corresponds to the model of Figure 2.7, in which the sequence of places indicating the circulation of urgent calls has been removed. The places corresponding to the extremely urgent calls circulation now correspond to all urgent calls, which are all kept in line between level 1 and level 2.

Chapter 3

Discrete dynamics and fluid approximation for Petri nets with priorities

The results of this chapter, without the proofs, are published in the proceedings of the conference FORMATS [ABG15]. On top of the proofs of our theorems, we also established and proved the converse of one of the main theorem: it was already stated that the counter variables of the execution of a Petri net with free choice and priority routing are solutions of a piecewise linear dynamical system with delays. We also establish that a càdlàg, nondecreasing solution of such dynamical system is also the counter variable of an execution of the Petri net. We have also added to this chapter the analysis of a second example, and the corresponding numerical application.

3.1 Introduction

In the present chapter, we consider a first, natural modeling of our emergency call center as a Petri net with discrete firings and a discrete time evolution. In our model, places are given a

constant holding time. Free choice conflicts are solved by a stationary probability distribution (with fixed probability, a token is routed towards a given downstream transition), and non free choice situations are handled with priority rules.

This leads us to analyze the class of Petri nets in which the places can be partitioned in two categories: the routing in certain places is subject to priority rules, whereas the routing at the other places is free choice. We present an algebraic approach which allows to analyzing the performance of such timed systems.

Counter variables determine the number of firings of the different transitions as a function of time. A main result of this chapter shows that, for the earliest firing rule, the counter variables are the solutions of a piecewise linear dynamical system, and that, moreover, any nondecreasing solution of this dynamical system corresponds to a correct execution of the Petri net (Section 3.2). Then, we introduce a fluid approximation in which the counter variables are real valued, instead of integer valued. Our second main result shows that in the fluid model, the affine stationary regimes are precisely the solutions of a set of lexicographic piecewise linear equations, which constitutes a polynomial system over a tropical (min-plus) semifield of germs (Section 3.3). The latter is a modification of the ordinary tropical semifield. In essence, this result shows that computing affine stationary regimes reduces to solving tropical polynomial systems.

Solving tropical polynomial systems is one of the most basic problems of tropical geometry. The latter provides insights on the nature of solutions, as well as algorithmic tools. In particular, the tropical approach allows one to determine the different congestion phases of the system.

We apply this approach to the case study of PP and BSPP. For the simplified model of the emergency call center, introduced in Section 2.5.1, we solve the associated system of tropical polynomial equations and arrive at an explicit computation of the different congestion phases, depending on the ratio N_2/N_1 of the numbers of operators of level 2 and 1 (Section 3.4). Our analytical results are obtained only for the approximate fluid model. However, they are confirmed by simulations in which the original semantics of the Petri nets (with integer firings) is respected (Section 3.6).

However, computations on this example and a second one also show that affine stationary regimes are not the only possible asymptotic regimes of the fluid approximation of the dynamics. Periodic behaviors may also be reached, yielding a different throughput (Section 3.6.2).

Related work.

Our approach finds its origin in the maxplus modeling of timed discrete event systems, introduced by Cohen, Quadrat and Viot and further developed by Baccelli and Olsder, see [BCOQ92, HOvdW06] for background. The idea of using counter variables already appeared in their work. However, the classical results only apply to restricted classes of Petri nets, like event graphs, or event graphs with weights as, for instance, in recent work by Cottenceau, Hardouin and Boimond [CHB14]. The modeling of more general Petri nets by a combination of min-plus linear constraints and classical linear constraints was proposed by Cohen, Gaubert and Quadrat [CGQ98, CGQ95] and Libeaut and Loiseau (see [Lib96]). The question of analyzing the behavior of the dynamical systems arising in this way was stated in a compendium of open problems in control theory [Plu99]. For this model, Gaujal and Giua [GG04b] proved the asymptotic throughput to be solution of a linear program. A key discrepancy with the previously developed min-plus algebraic models lies in the *semantics* of the Petri nets. The model of [CGQ98, CGQ95, GG04b] requires the routing to be based on open loop *preselection* policies of tokens at places, and it does not allow for priority rules. This is remedied in the present work: we show that priority rules can be written in a piecewise linear way, leading to a rational tropical dynamics. However, as is shown in numerical experiments, the stationary throughput computed in the case with priority routing does not always correspond to the asymptotic throughput of the dynamics, contrarily to the case with preselection policies [GG04b].

Our approach is inspired by a work of Farhi, Goursat and Quadrat [FGQ11], who developed a min-plus model for a road traffic network. The idea of modeling priorities by rational min-plus dynamics first appeared there. By comparison, one aspect of novelty of the present approach consists in showing that this idea applies to a large class of Petri nets, mixing free choice and priority routing, so that its scope is not limited to a special class of road traffic models. Moreover, we provide a complete proof that these Petri nets follow the rational trop-

ical dynamics, based on a precise analysis of the counter variables along an execution trace. Finally, the approach of [FGQ11] was developed in the discrete time case. A novelty of the present work consists in the treatment of the continuous time. This requires the introduction of a symbolic perturbation technique, working with semifield of germs. This technique was used in [GG98] for algorithmic purposes. It has been recently applied by Allamigeon, Fahrenberg, Gaubert, Katz, and Legay to the analysis of timed systems [AFG+14].

The analysis of timed Petri nets is a major question, which has been extensively studied. We refer to [BD91, AN01, GRR04, JJMS11] for a non-exhaustive account on the topic, and to [BV06, LRST09, BJS09] for examples of tools implementing these techniques. An important effort has been devoted to the comparison of timed Petri nets with timed automata in terms of expressivity, see for instance [BCH+05, Srb08]. The approaches developed in the aforementioned works aim at checking whether a given specification is satisfied (for instance, reachability, or more generally, a property expressed in a certain temporal logic), or at determining whether two Petri nets are equivalent in the sense of bisimulation. Hence, the emphasis is on issues different from the present ones: we focus on the performance analysis of timed Petri nets, by determining the asymptotic throughputs of transitions.

3.2 Piecewise linear dynamics of timed Petri nets with free choice and priority routing

3.2.1 Timed Petri nets: notation and semantics

A ***timed Petri net*** consists of a set $\mathcal{P}$ of places and a set $\mathcal{Q}$ of transitions, in which each place $p \in \mathcal{P}$ is equipped with a holding time $\tau_p \in \mathbb{R}_{>0}$ as well as an initial marking $M_p \in \mathbb{N}$. Given a place $p \in \mathcal{P}$, we respectively denote by p^{in} and p^{out} the sets of input and output transitions. Similarly, for all $q \in \mathcal{Q}$, the sets of upstream and downstream places are denoted by q^{in} and q^{out} respectively.

The semantics of the timed Petri net which we use in this chapter is based on the fact that every token entering a place $p \in \mathcal{P}$ must stay at least τ_p time units in place p before becoming available for a firing of a downstream transition. More formally, a state of the semantics of the Petri net specifies, for each place $p \in \mathcal{P}$, the set of tokens located at place p, together with the age of these tokens since they have entered place p. In a given state σ, the Petri net can evolve into a new state σ' in two different ways:

(i) either a transition $q \in \mathcal{Q}$ is fired, which we denote $\sigma \xrightarrow{q} \sigma'$. This occurs when every upstream place p contains a token whose age is greater than or equal to τ_p. The transition is supposed to be instantaneous. A token enters in each downstream place, and its age is set to 0;

(ii) or all the tokens remain at their original places, and their ages are incremented by the same amount of time $d \in \mathbb{R}_{\geqslant 0}$. This is denoted $\sigma \xrightarrow{d} \sigma'$.

In the initial state σ^0, all the tokens of the initial marking are supposed to have an "infinite" age, so that they are available for firings of downstream transitions from the beginning of the execution of the Petri net. The set of relations of the form $\xrightarrow{q}$ and $\xrightarrow{d}$ constitutes a timed transition system which, together with the initial state σ^0, fully describe the semantics of the Petri net. Note that in this semantics, transitions can be fired simultaneously. In particular, a given transition can be fired several times at the same moment. Recall that every holding time τ_p is positive, so that we cannot have any Zeno behavior.

In this setting, we can write any execution trace of the Petri net as a sequence of transitions of the form:

$$\sigma^0 \xrightarrow{d^0} \xrightarrow{q_1^0} \xrightarrow{q_2^0} \dots \xrightarrow{q_{n^0}^0} \sigma^1 \xrightarrow{d^1} \xrightarrow{q_1^1} \xrightarrow{q_2^1} \dots \xrightarrow{q_{n^1}^1} \sigma^2 \xrightarrow{d^2} \dots \tag{3.1}$$

where $d^0 \geqslant 0$ and $d^1, d^2, \dots > 0$. In other words, we consider traces in which we remove all the time-elapsing transitions of duration 0, except the first one, and in which time-elapsing transitions are separated by groups of firing transitions occurring simultaneously. We say that a transition q is *fired at the instant* t if there is a transition $\xrightarrow{q}$ in the trace such that the sum of the durations of the transitions of the form $\xrightarrow{d}$ which occur before in the trace is equal to t. The *state of the Petri net at the instant* t refers to the state of the Petri net appearing in the trace (3.1) after all transitions have been fired at the instant t.

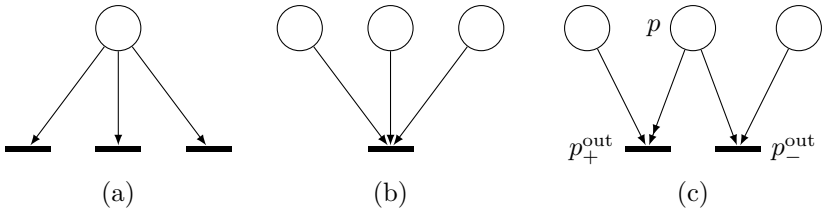


Figure 3.1 – Conflict, synchronization and priority configurations.

In the rest of the chapter, we stick to a stronger variant of the semantics, referred to as ***earliest behavior*** semantics, in which every transition q is fired at the earliest moment possible. More formally, this means that in any state σ arising during the execution, a place p is allowed to contain a token of age (strictly) greater than τ_p only if no downstream transition can be fired (*i.e.* no transition $\xrightarrow{q}$ with $q \in p^{\text{out}}$ can be applied to σ). The motivation to study the earliest behavior semantics originates from our interest for emergency call centers, in which all calls are supposed to be handled as soon as possible.

3.2.2 Timed Petri nets with free choice and priority routing

In this chapter, we consider timed Petri nets in which places are free choice, or subject to priorities. This class of nets includes our model of emergency call center. Recall that a place $p \in \mathcal{P}$ is said to be ***free choice*** if either $|p^{\text{out}}| = 1$, or all the downstream transitions $q \in p^{\text{out}}$ satisfy $q^{\text{in}} = \{p\}$. The main property of such a place is the following: if one of the downstream transitions is activated (*i.e.* it can be potentially fired), then the other downstream transitions are also activated. A place is ***subject to priority*** if the available tokens in this place are routed to downstream transitions according to a certain priority rule. We denote by $\mathcal{P}_{\mathsf{priority}}$ the set of such places. We assume that no transition has more than one upstream place subject to priority, that is, for any transition q, the set $q^{\text{in}} \cap \mathcal{P}_{\mathsf{priority}}$ has at most one element. This allows to avoid inconsistency between priority rules (e.g. two priority places acting on the same transitions in a contradictory way). For the sake of simplicity, we also assume in the following that every $p \in \mathcal{P}_{\mathsf{priority}}$ has precisely two downstream transitions, which we respectively denote by p_+^{out} and p_-^{out}. Then, if both transitions are activated, the tokens available in place p are assigned to p_+^{out} as a priority. Equivalently, in the execution trace of the Petri net, we have $\sigma \to^{p_-^{\text{out}}} \sigma'$ only if the transition $\to^{p_+^{\text{out}}}$ cannot be applied to the state σ. We remark that it is possible to handle multiple priority levels, up to making the presentation of the subsequent results more complicated.

To summarize, there are three possible place/transition patterns which can occur in the timed Petri nets that we consider, see Figure 3.1. The first two involve only free choice places, and are referred to as ***conflict*** and ***synchronization*** patterns respectively. We denote by $\mathcal{P}_{\mathsf{conflict}}$ the set of free choice places that have at least two output transitions, and by $\mathcal{Q}_{\mathsf{sync}}$ the set of transitions such that every upstream place p satisfies $|p^{\text{out}}| = 1$. By definition, we have $\mathcal{P}_{\mathsf{conflict}} \cap (\mathcal{Q}_{\mathsf{sync}})^{\text{in}} = \emptyset$. The third configuration in Figure 3.1 depicts a place p subject to priority. In order to distinguish p_+^{out} and p_-^{out}, we depict the arc leading to the transition p_+^{out} by a double arrow. By assumption, the places $r \neq p$ located upstream p_+^{out} and p_-^{out} are non-priority, so that they are free-choice and have only one output transition, as depicted in Figure 3.1(c).

3.2.3 Piecewise linear representation by counter variables

Since we are interested in estimating the throughput of transitions in a Petri net, we associate with any transition $q \in \mathcal{Q}$ a counter variable z_q from $\mathbb{R}$ to $\mathbb{N}$ such that $z_q(t)$ represents the number of firings of transition q that occurred up to time t included. Similarly, given a place $p \in \mathcal{P}$, we denote $x_p(t)$ the number of tokens that have entered place p up to time t included. Note that the tokens initially present in place p are counted. More formally, $x_p(t)$ is given by the sum of the initial marking M_p and of the numbers of firings of transitions $q \in p^{\text{in}}$ which occurred before the instant t (included). We extend the counter variables x_p and z_q to $\mathbb{R}_{<0}$ by

setting:

$$x_p(t) = M_p\,, \quad z_q(t) = 0\,, \quad \text{for all } t < 0\ . \tag{3.2}$$

By construction, the functions x_p and z_q are non-decreasing. Besides, since they count tokens up to time t *included*, they are *càdlàg* functions, which means that they are right continuous and have left limits at any point. Given a càdlàg function f, we denote by $f(t^-)$ the left limit at the point t.

The goal of this section is to describe the dynamics of timed Petri nets with free choice and priority routing by means of a set of piecewise linear equality constraints over the counter variables. We provide an informal presentation of these constraints. First observe that we necessarily have:

$$\forall p \in \mathcal{P}\,, \quad x_p(t) = M_p + \sum_{q \in p^{\mathrm{in}}} z_q(t)\,, \tag{3.3}$$

as the initial marking M_p is counted in $x_p(t)$, and any token entering place p up to instant t must have been fired from an upstream transition $q \in p^{\mathrm{in}}$ before. In a similar way, if $p \in \mathcal{P}_{\mathsf{conflict}}$, the total number of times the downstream transitions have been fired up to instant t is necessarily equal to the number of tokens which entered place p before time $t - \tau_p$ (included). This is due to the fact that if a token enters p at the instant s, then it is consumed *exactly* at the instant $s + \tau_p$ (by definition of the earliest behavior semantics). This yields the identity:

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \quad \sum_{q \in p^{\mathrm{out}}} z_q(t) = x_p(t - \tau_p)\,. \tag{3.4}$$

Now consider a transition $q \in \mathcal{Q}_{\mathsf{sync}}$. The number of times this transition is fired at the instant t is given by $z_q(t) - z_q(t^-)$. In each upstream place $p \in q^{\mathrm{in}}$, the number of tokens which are available for firing q is equal to $x_p(t - \tau_p) - z_q(t^-)$. Indeed, since place p does not have any other output transition, the total number of tokens which have left place q until the instant t equals $z_q(t^-)$. By definition of the earliest behavior semantics, the number of firings of q at the instant t must be exactly equal to the minimum number of tokens available in places $p \in q^{\mathrm{in}}$. If we denote $\min(x, y)$ by $x \wedge y$, we consequently get:

$$\forall q \in \mathcal{Q}_{\mathsf{sync}}\,, \quad z_q(t) = \bigwedge_{p \in q^{\mathrm{in}}} x_p(t - \tau_p)\,. \tag{3.5}$$

Finally, let us take a place $p \in \mathcal{P}_{\mathsf{priority}}$. Since the transition p_+^{out} has priority over p_-^{out}, the quantity $z_{p_+^{\mathrm{out}}}(t) - z_{p_+^{\mathrm{out}}}(t^-)$ must be equal to the minimal number of tokens available in the upstream places, including p. For every place $r \in (p_+^{\mathrm{out}})^{\mathrm{in}}$ distinct from p, the number of available tokens is given by $x_r(t - \tau_r) - z_{p_+^{\mathrm{out}}}(t^-)$ (recall that p_+^{out} is the only downstream transition of r). In contrast, the number of tokens available for firing in place p is equal to $x_p(t - \tau_p) - (z_{p_+^{\mathrm{out}}}(t^-) + z_{p_-^{\mathrm{out}}}(t^-))$. We deduce that we have:

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,, \quad z_{p_+^{\mathrm{out}}}(t) = \big(x_p(t - \tau_p) - z_{p_-^{\mathrm{out}}}(t^-)\big) \wedge \bigwedge_{\substack{r \in (p_+^{\mathrm{out}})^{\mathrm{in}} \\ r \neq p}} x_r(t - \tau_r)\,. \tag{3.6}$$

The number of tokens from place p which are available for the transition p_-^{out} after the firings of p_+^{out} is given by $x_p(t-\tau_p) - (z_{p_+^{\mathrm{out}}}(t^-) + z_{p_-^{\mathrm{out}}}(t^-)) - (z_{p_+^{\mathrm{out}}}(t) - z_{p_+^{\mathrm{out}}}(t^-))$. Hence, we obtain:

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,, \quad z_{p_-^{\mathrm{out}}}(t) = \big(x_p(t - \tau_p) - z_{p_+^{\mathrm{out}}}(t)\big) \wedge \bigwedge_{\substack{r \in (p_-^{\mathrm{out}})^{\mathrm{in}} \\ r \neq p}} x_r(t - \tau_r)\,. \tag{3.7}$$

We summarize the previous discussion by the following result:

▶ **Theorem 3.1.** *Given any execution trace of a timed Petri net with free choice and priority routing, the counter variables x_p ($p \in \mathcal{P}$) and z_q ($q \in \mathcal{Q}$) satisfy the constraints (3.3)–(3.7) for all $t \geqslant 0$, together with the initial conditions (3.2).*

Conversely, any nondecreasing, càdlàg solution of (3.3)–(3.7), with initial conditions (3.2), consists in counter variables of an execution trace of a timed Petri net with free choice and priority routing.

We prove the first statement in Section 3.2.5, and its converse in 3.2.6. Notice that, if we do not restrict to the earliest behavior semantics, the constraints (3.4)–(3.7) are relaxed to inequalities.

So far, we have described the dynamics of timed Petri nets in the continuous time setting. However, since the Petri net of our case study is a model of a real system which is implemented in silico, we need to investigate the dynamics in discrete time as well. In more details, assuming that all the quantities τ_p are multiple of an elementary time step $\delta > 0$, the discrete-time version of the semantics of the Petri net restricts the transitions $\xrightarrow{d}$ to the case where d is a multiple of δ. In this case, on top of being càdlàg, the functions x_p and z_q are constant on any interval of the form $[k\delta, (k+1)\delta)$ for all $k \in \mathbb{N}$. Then, we can verify that the following result holds:

▶ **Proposition 3.2.** *In the discrete time semantics, the counter variables x_p and z_q satisfy the constraints (3.3)–(3.7) for all $t \geqslant 0$, independently of the choice of the elementary time step δ.*

In other words, the dynamics in continuous-time is a valid representation of the dynamics in discrete time which allows to abstract from the discretization time step. We also note that we can refine the constraint given in (3.6) by replacing the left limit $z_{p^{\mathrm{out}}_-}(t^-)$ by an explicit value:

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,, \quad z_{p^{\mathrm{out}}_+}(t) = \begin{cases} \big(x_p(t-\tau_p) - z_{p^{\mathrm{out}}_-}(t-\delta)\big) \\ \qquad \wedge \bigwedge_{r \in (p^{\mathrm{out}}_+)^{\mathrm{in}}\,,\, r \neq p} x_r(t-\tau_r) & \text{if } t \in \delta\mathbb{N}\,, \\ \big(x_p(t-\tau_p) - z_{p^{\mathrm{out}}_-}(t)\big) \\ \qquad \wedge \bigwedge_{r \in (p^{\mathrm{out}}_+)^{\mathrm{in}}\,,\, r \neq p} x_r(t-\tau_r) & \text{otherwise.} \end{cases} \tag{3.8}$$

(Here and below, we denote by $\delta\mathbb{N}$ the set $\{0, \delta, 2\delta, \ldots\}$.) The system formed by the constraints (3.3)–(3.5), (3.7), (3.8) is referred to as the *δ-discretization of the Petri net dynamics*.

The only source of non-determinism in the model that we consider is the routing policy in the conflict pattern (Figure 3.1(a)). In the sequel, we assume that the tokens are assigned according to a stationary probability distribution. Given a free choice place $p \in \mathcal{P}_{\mathsf{conflict}}$, we denote by π_{qp} the probability that an available token is assigned to the transition $q \in p^{\mathrm{out}}$. In the following, we consider a ***fluid approximation of the dynamics*** of the system, in which the x_p and z_q are non-decreasing càdlàg functions from $\mathbb{R}$ to itself, and the routing policy degenerates in sharing the tokens in fractions π_{qp}. Equivalently, the fluid dynamics is defined by the constraints (3.3)–(3.7) and the following additional constraints:

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \forall q \in p^{\mathrm{out}}\,, \quad z_q(t) = \pi_{qp} x_p(t-\tau_p)\,. \tag{3.9}$$

Note that the latter equation is still valid in the context of discrete time. By extension, the system formed by the constraints (3.3)–(3.5), (3.7)–(3.9) is referred to as the *δ-discretization of the fluid dynamics*.

▶ **Remark 3.3** (The case with valuations)**.** We have considered plain Petri nets, that is, Petri nets whose arcs all have valuation 1. The case with valuations yields more complicated equations, which we do not detail here. The difficulty lies in the fact that transitions can only fire a given integer number of tokens from upstream places.

In the case of the fluid approximation, however, we allow transitions to fire fractions of tokens, so that the dynamics is much simpler. We postpone to the next chapter the extension of the fluid approximation of the dynamics to the case with valuations: see Equations (4.21).

3.2.4 Application to our Petri net model of emergency call center

We illustrate Theorem 3.1 on the Petri net of Figure 2.7. We point out that in Figure 2.7, we have omitted to specify the holding time of some places. By default, this holding time is set to a certain $\tau_\varepsilon > 0$, and is meant to be negligible w.r.t. the other holding times.

For simplicity, we omit the counter variables of the places distinct from p_1 and p_2. Indeed, each of theses places p has a unique input transition q, and its initial marking is 0. Therefore, by definition, we have $x_p(t) = z_q(t)$ for all t, which means that x_p can be trivially substituted in the constraints. Similarly, we omit the transitions which lead to places p_1 and p_2, as their counter variables correspond the counter variables of some transitions located upstream and

shifted by the holding time of the place in between. Finally, we denote by z_i the counter variables of transitions q_i, and by x_i the counter variables of places p_i. We can verify that the fluid dynamics is then given by the following constraints:

$$\begin{aligned}
z_1(t) &= x_1(t-\tau_\varepsilon)\\
z_2(t) &= \mu_{\text{ext}} z_1(t-\tau_\varepsilon)\\
z_3(t) &= \mu_{\text{ur}} z_1(t-\tau_\varepsilon)\\
z_4(t) &= \mu_{\text{adv}} z_1(t-\tau_\varepsilon)\\
z_5(t) &= (x_2(t-\tau_\varepsilon) - z_6(t^-)) \wedge z_2(t-\tau_{\text{ext}})\\
z_6(t) &= (x_2(t-\tau_\varepsilon) - z_5(t)) \wedge z_3(t-\tau_{\text{ur}}-\tau_\varepsilon)\\
z_7(t) &= z_5(t-\tau_{\text{tr}})\\
x_1(t) &= N_1 + z_7(t) + z_3(t-\tau_{\text{ur}}) + z_4(t-\tau_{\text{adv}})\\
x_2(t) &= N_2 + z_7(t-\tau'_{\text{ext}}) + z_6(t-\tau'_{\text{ur}})
\end{aligned}$$

They can be simplified into the following system:

$$\begin{aligned}
z_1(t) &= N_1 + z_5(t-\tau_{\text{tr}}) + \mu_{\text{ur}} z_1(t-\tau_{\text{ur}}-2\tau_\varepsilon) + \mu_{\text{adv}} z_1(t-\tau_{\text{adv}}-2\tau_\varepsilon)\\
z_5(t) &= \big(N_2 + z_5(t-\tau_{\text{tr}}-\tau'_{\text{ext}}-\tau_\varepsilon) + z_6(t-\tau'_{\text{ur}}-\tau_\varepsilon) - z_6(t^-)\big)\\
&\qquad \wedge \mu_{\text{ext}} z_1(t-\tau_{\text{ext}}-\tau_\varepsilon)\\
z_6(t) &= \big(N_2 + z_5(t-\tau_{\text{tr}}-\tau'_{\text{ext}}-\tau_\varepsilon) + z_6(t-\tau'_{\text{ur}}-\tau_\varepsilon) - z_5(t)\big)\\
&\qquad \wedge \mu_{\text{ur}} z_1(t-\tau_{\text{ur}}-\tau_\varepsilon)
\end{aligned} \tag{3.10}$$

which involve the counter variables z_1, z_5 and z_6 only. These variables correspond to the key characteristics of the system. They respectively represent the number of calls handled at level 1, and the number of extremely urgent and urgent calls handled at level 2, up to time t. All the other counter variables can be straightforwardly obtained from z_1, z_5 and z_6.

For the sake of readability, we slightly modify the original holding times τ_{ext}, τ_{ur}, ... to incorporate the effect of τ_ε. In more details, we substitute τ_{ext}, τ_{ur}, τ_{adv}, τ'_{ext} and τ'_{ur} by $\tau_{\text{ext}}-\tau_\varepsilon$, $\tau_{\text{ur}}-2\tau_\varepsilon$, $\tau_{\text{adv}}-2\tau_\varepsilon$, $\tau'_{\text{ext}}-\tau_\varepsilon$ and $\tau'_{\text{ur}}-\tau_\varepsilon$ respectively. Then, System (3.10) simply reads as:

$$\begin{aligned}
z_1(t) &= N_1 + z_5(t-\tau_{\text{tr}}) + \mu_{\text{ur}} z_1(t-\tau_{\text{ur}}) + \mu_{\text{adv}} z_1(t-\tau_{\text{adv}})\\
z_5(t) &= \big(N_2 + z_5(t-\tau_{\text{tr}}-\tau'_{\text{ext}}) + z_6(t-\tau'_{\text{ur}}) - z_6(t^-)\big) \wedge \mu_{\text{ext}} z_1(t-\tau_{\text{ext}})\\
z_6(t) &= \big(N_2 + z_5(t-\tau_{\text{tr}}-\tau'_{\text{ext}}) + z_6(t-\tau'_{\text{ur}}) - z_5(t)\big) \wedge \mu_{\text{ur}} z_1(t-\tau_{\text{ur}})
\end{aligned} \tag{3.11}$$

This is the system which we consider in the rest of the chapter.

3.2.5 Proof of the first part of Theorem 3.1

▶ **Lemma 3.4.** *Suppose that all the holding times τ_p ($p \in \mathcal{P}$) are positive, and consider the subpart of the execution trace formed by the transitions fired at the instant t, i.e.:*

$$\ldots \xrightarrow{d} \sigma^{t^-} \xrightarrow{q_1} \xrightarrow{q_2} \ldots \xrightarrow{q_n} \sigma^t \xrightarrow{d'} \tag{3.12}$$

(with $d > 0$ unless $t = 0$, and $d' > 0$). Then the following two properties hold:

(i) for all $p \in \mathcal{P}_{\textsf{priority}}$, no transition $\xrightarrow{p^{out}_-}$ can occur before a transition $\xrightarrow{p^{out}_+}$ in (3.12);

(ii) any pair of consecutive transitions $\xrightarrow{q_i}\xrightarrow{q_{i+1}}$ can be switched in (3.12) without changing the states occurring after, provided that (q_i, q_{i+1}) is not equal to (p^{out}_+, p^{out}_-) for some $p \in \mathcal{P}_{\textsf{priority}}$.

Proof. (i) Suppose that a transition $\xrightarrow{p^{\text{out}}_-}$ occurs before $\xrightarrow{p^{\text{out}}_+}$ in (3.12), *i.e.* we have a subsequence of the form $\sigma \xrightarrow{p^{\text{out}}_-} \sigma' \ldots \xrightarrow{p^{\text{out}}_+}$. As all the holding times are positive, all the tokens consumed by p^{out}_+ are already present in the state σ. In other words, the transition $\sigma \xrightarrow{p^{\text{out}}_+} \ldots$ is valid in the semantics. This contradicts the priority rule.

(ii) Consider a pair of consecutive transitions

$$\sigma^i \xrightarrow{q_i} \xrightarrow{q_{i+1}} \sigma^{i+2}$$

such that $(q_i, q_{i+1}) \neq (p_+^{\text{out}}, p_-^{\text{out}})$. As discussed in the previous case, since the holding times are positive, the transition q_{i+1} does not consume tokens produced by the transition q_i. Besides, there is no priority rule between q_i and q_{i+1}. Therefore, q_{i+1} can be fired before q_i, and the sequence $\sigma^i \xrightarrow{q_{i+1}} \xrightarrow{q_i}$ leads to the same state σ^{i+2}. It follows that all the subsequent states remain identical. ◀

Let us now prove Theorem 3.1. It is useful to extract the part of the execution trace leading to the state σ^t of the Petri net at the instant t. It has one of the following two forms:

$$\ldots \xrightarrow{d} \sigma^{t^-} \xrightarrow{q_1} \xrightarrow{q_2} \ldots \xrightarrow{q_n} \sigma^t \xrightarrow{d'} \tag{3.13}$$

or

$$\ldots \xrightarrow{d} \xrightarrow{q_1} \xrightarrow{q_2} \ldots \xrightarrow{q_n} \xrightarrow{d''} \sigma^t \xrightarrow{d'} \tag{3.14}$$

depending on whether some transitions $q \in \mathcal{Q}$ are fired at the instant t or not. In both cases, $d', d'' > 0$ and $d > 0$ unless $t = 0$, and the durations of the time-elapsing transitions occurring before the state σ^t in the trace sum up to t.

In this context, $z_q(t)$ counts the number of transitions occurring before σ^t in the trace, while $x_p(t)$ is given by the sum of M_p and the number of transitions $q \in p^{\text{in}}$ occurring before σ^t. The constraint (3.3) is therefore trivially satisfied by definition of $x_p(t)$ and the $z_q(t)$.

Consider $p \in \mathcal{P}_{\mathsf{conflict}}$. Since we use the earliest behavior semantics and p is free choice, any token of the initial marking is consumed at the instant 0, and any token brought by an upstream transition $q' \in p^{\text{in}}$ at the instant $s \geqslant 0$ is consumed at the instant $s + \tau_p$. As a consequence, we can build a bijection which maps each initial token with the transition $\xrightarrow{q}$ which consumes it at the instant 0, and any transition $\xrightarrow{q'}$ occurring at the instant $s - \tau_p$ with the transition $\xrightarrow{q}$ which consumes at the instant s the token brought by q' to place p. We deduce that

$$M_p + \sum_{q' \in p^{\text{in}}} z_{q'}(t - \tau_p) = \sum_{q \in p^{\text{out}}} z_q(t)\,.$$

Using the constraint (3.3), this yields to $x_p(t - \tau_p) = \sum_{q \in p^{\text{out}}} z_q(t)$.

Now, let us take $q \in \mathcal{Q}_{\mathsf{sync}}$. Consider $p \in q^{\text{in}}$. Recall that, by definition of $\mathcal{Q}_{\mathsf{sync}}$, q is the only downstream transition of place p. Therefore, every transition $\xrightarrow{q}$ arising at the instant s consumes a token from place p. This token is either a initial token from M_p, or a token brought by a transition $q' \in p^{\text{in}}$ fired before the instant $s - \tau_p$ (included). Therefore, we have $z_q(t) \leqslant x_p(t - \tau_p)$. In fact, $x_p(t - \tau_p) - z_q(t)$ is equal to the number of tokens with age greater than or equal to τ_p located in place p in the state σ^t. At the instant $t + \varepsilon$ with $0 < \varepsilon < d'$, the age of these tokens will be strictly greater than τ_p. Therefore, if $x_p(t - \tau_p) - z_q(t) > 0$ for all $p \in q^{\text{in}}$, the transition q can be fired at the instant $t + \varepsilon$. But this is impossible in the earliest behavior semantics, since the places $p \in q^{\text{in}}$ are not allowed to contain tokens with age strictly greater than τ_p while their downstream transition q can be fired. We deduce that $x_p(t - \tau_p) = z_q(t)$ for some $p \in q^{\text{in}}$. This proves (3.5).

Finally, consider $p \in \mathcal{P}_{\mathsf{priority}}$. Using similar arguments as the ones used in the previous case, we can show that $z_{p_+^{\text{out}}}(t) \leqslant x_r(t - \tau_r)$ for all $r \in (p_+^{\text{out}})^{\text{in}}$, $r \neq p$, and $z_{p_-^{\text{out}}}(t) \leqslant x_r(t - \tau_r)$ for all $r \in (p_-^{\text{out}})^{\text{in}}$, $r \neq p$. Besides, we have $z_{p_+^{\text{out}}}(t) + z_{p_-^{\text{out}}}(t) \leqslant x_p(t - \tau_p)$, since every firing of the transition p_+^{out} or p_-^{out} at the instant s consumes a token of M_p or a token brought by an upstream transition of p before the instant $s - \tau_p$. In consequence, as the function $z_{p_-^{\text{out}}}$ is non-decreasing, we obtain:

$$z_{p_+^{\text{out}}}(t) + z_{p_-^{\text{out}}}(t^-) \leqslant x_p(t - \tau_p)\,.$$

In order to prove that (3.6) is satisfied, we distinguish two cases depending on the form of the trace:

(i) if the trace is of the form (3.13), then, by Lemma 3.4, we can rewrite the subpart of the trace as follows:

$$\ldots \xrightarrow{d} \sigma^{t^-} \xrightarrow{q'_1} \xrightarrow{q'_2} \ldots \xrightarrow{q'_k} \underbrace{\xrightarrow{p_+^{\text{out}}} \ldots \xrightarrow{p_+^{\text{out}}}}_{k_+ \text{ times}} \sigma \underbrace{\xrightarrow{p_-^{\text{out}}} \ldots \xrightarrow{p_-^{\text{out}}}}_{k_- \text{ times}} \sigma^t \xrightarrow{d'}$$

where $k_+, k_- \geqslant 0$, and where p_+^{out} and p_-^{out} do not appear in the q'_i. Then, the quantity $x_p(t-\tau_p) - z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t^-)$ corresponds to the number of tokens with age greater than or equal to τ_p in the intermediary state σ. If it is positive, and if $x_r(t-\tau_r) - z_{p_+^{\text{out}}}(t) > 0$ for all $r \in (p_+^{\text{out}})^{\text{in}}$ such that $r \neq p$, then the transition p_+^{out} can be fired right after the state σ. This contradicts the priority rule if $k_- > 0$. If $k_- = 0$, we can fire p_+^{out} at the instant $t+\varepsilon$ $(0 < \varepsilon < d')$, which contradicts the definition of the earliest behavior semantics (all the upstream place of p_+^{out} contains a token older than allowed).

(ii) if the trace is of the form (3.14), then $z_{p_-^{\text{out}}}(t^-) = z_{p_-^{\text{out}}}(t)$. In this case, the quantity $x_p(t-\tau_p) - z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t^-)$ represents the number of tokens with age greater than or equal to τ_p at place p in the state σ^t. If $x_p(t-\tau_p) - z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t^-) > 0$ and $x_r(t-\tau_r) - z_{p_+^{\text{out}}}(t) > 0$ for all $r \in (p_+^{\text{out}})^{\text{in}}$ such that $r \neq p$, then the transition p_+^{out} can be fired at the instant $t+\varepsilon$ with $0 < \varepsilon < d'$. This is again a contradiction with the earliest behavior semantics.

In both cases, we have $x_p(t-\tau_p) - z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t^-) = 0$ or $x_r(t-\tau_r) - z_{p_+^{\text{out}}}(t) = 0$ for some $r \in (p_+^{\text{out}})^{\text{in}}$ such that $r \neq p$. We deduce that the constraint (3.6) holds.

Now assume that $x_p(t-\tau_p) - z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t) > 0$ and $x_r(t-\tau_r) - z_{p_-^{\text{out}}}(t) > 0$ for all $r \in (p_-^{\text{out}})^{\text{in}}$ such that $r \neq p$. These quantities correspond to the number of tokens in places p and r with age greater than or equal to τ_p and τ_r respectively, in the state σ^t. Thus, the transition p_-^{out} is activated at the instant $t+\varepsilon$ for all $\varepsilon > 0$ sufficiently small. Note that $x_p(t-\tau_p) - z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t^-) > 0$ as $z_{p_-^{\text{out}}}(t^-) < z_{p_-^{\text{out}}}(t)$. Thus, there exists a place $r' \in (p_+^{\text{out}})^{\text{in}}$ with $r \neq p$, such that $x_{r'}(t-\tau_{r'}) = z_{p_+^{\text{out}}}(t)$. In other words, place r' does not contain any token with age greater than or equal to $\tau_{r'}$. Given $\varepsilon > 0$ sufficiently small, this is still true at the instant $t+\varepsilon$, so that the transition p_+^{out} cannot be fired at the instant $t+\varepsilon$. Therefore, we are allowed to fire the transition p_-^{out} at the instant $t+\varepsilon$, which is a contradiction with the definition of the earliest behavior semantics. As a result, (3.7) is satisfied. ◀

3.2.6 Proof of the second part of Theorem 3.1

Consider a timed Petri net with free choice and priority routing, and suppose that we are given functions $x_p : \mathbb{R}_{\geqslant 0} \to \mathbb{N}$ for $p \in \mathcal{P}$ and $z_q : \mathbb{R}_{\geqslant 0} \to \mathbb{N}$ for $q \in \mathcal{Q}$ satisfying equations (3.3)–(3.7) for all $t \geqslant 0$, together with the initial conditions (3.2). Let us further assume that the x_p and z_q are càdlàg, nondecreasing functions.

We first prove that the functions x_p and z_q are well defined on $\mathbb{R}_{\geqslant 0}$ and admit a finite number of discontinuities on any interval $[0,T]$, $T > 0$, which implies that the union of the times of discontinuities of those functions is at most countable.

▶ **Lemma 3.5.** *(i) If a function x_p or z_q admits an infinite number of discontinuities on a time interval, then it goes to infinity on this interval.*

(ii) If the functions x_p, for $p \in \mathcal{P}$, and z_q, for $q \in \mathcal{Q}$, are defined on $[0,T]$, then they are defined on $[0, T+\tau^m]$, where $\tau^m = \min_{p\in\mathcal{P}} \tau_p$.

(iii) The functions (x_p) and (z_q) are defined on $\mathbb{R}_{\geqslant 0}$ and admit a finite number of discontinuities on any interval $[0,T]$, $T > 0$.

Proof. (i) The functions being nondecreasing and having integer values, each point of discontinuity is associated with a jump of size $\geqslant 1$.

(ii) For any transition q, at any time, $z_q(t)$ is upper bounded by one of the $x_p(t-\tau_p)$, with $p \in \mathcal{P}$. Besides, the $x_p(t)$ are expressed in terms of the $z_q(t)$.

(iii) This follows directly from (i) and (ii). ◀

We now prove that the x_p and z_q are the counter variables of an execution trace of the given timed Petri net with free choice and priority routing.

Let us first construct the corresponding execution trace, denoted σ. The z_q being càdlàg, they are piecewise constant, and encounter a positive jump at any point of discontinuity. Let $t_1, t_2 \dots$ be the (possibly infinite) increasing sequence of times of jumps of all the z_q, $q \in \mathcal{Q}$. We define the execution trace inductively. It will be convenient to note $t_0 = -\infty$ and σ^{t_0} the initial state preceding the beginning of the execution trace. Let $i \geqslant 1$ and t_i be the i-th jump time ; we suppose that we have constructed an execution trace up to time t_{i-1} included. Let $\sigma^{t_{i-1}}$ be the associated state. Let us order the transitions in $\mathcal{Q}$ such that the transitions being the non priority transitions of some place subject to priority are greater than all the other transitions, that is, we have $q^{(1)} \prec q^{(2)} \prec \cdots \prec q^{(|\mathcal{Q}|)}$ and, for any pair q, q', if $\exists p$ s.t. $q = p_-^{\text{out}}$ and $\nexists p$ s.t. $q' = p_-^{\text{out}}$, then $q \succ q'$. For any transition q, we define $J_q := z_q(t_i) - z_q(t_i^-)$. The functions z_q being càdlàg and nondecreasing, $J_q \in \mathbb{N}$. The execution trace between $\sigma^{t_{i-1}}$ and σ^{t_i} is constructed as follows:

$$\sigma^{t_{i-1}} \xrightarrow{t_i - t_{i-1}} \sigma^{t_{i-1}^-} \underbrace{\xrightarrow{q^{(1)}} \cdots \xrightarrow{q^{(1)}}}_{J_{q^{(1)}} \text{ times}} \underbrace{\xrightarrow{q^{(2)}} \cdots \xrightarrow{q^{(2)}}}_{J_{q^{(2)}} \text{ times}} \cdots \underbrace{\xrightarrow{q^{(|\mathcal{Q}|)}} \cdots \xrightarrow{q^{(|\mathcal{Q}|)}}}_{J_{q^{(|\mathcal{Q}|)}} \text{ times}} \sigma^{t_i} . \tag{3.15}$$

If $i = 1$, we replace the duration transition $\xrightarrow{t_i - t_{i-1}}$ by $\xrightarrow{t_i}$.

We prove by induction that this execution trace is a sound execution trace of the Petri net: that is, at any state of σ, a transition firing occurs only if every upstream place p contains a token of age $\geqslant \tau_p$, the execution sticks to the earliest behavior semantics, and the priority rule is respected, that is, if the non priority transition p_-^{out} of some place p fires, then at the state before the firing, the priority transition p_+^{out} is not activated. We prove also that, for this execution of the Petri net, at any time t, for any q, $z_q(t)$ counts the number of firings of transition q up to time t included, and for any p, $x_p(t)$ counts the number of tokens entered in place p up to time t included, plus the initial marking. We proceed by induction on the sequence $(t_0, t_1, t_2, \dots)$.

For $t \in]-\infty, 0[$, by the initial conditions (3.2), $z_q(t) = 0$ and $x_p(t) = M_p$ for any q and p. Moreover, the execution of σ does not start before time 0. Therefore, the execution is correct (nothing can happen in a Petri net before time 0), and the z_q and x_p have the correct meaning.

Now, suppose that the induction holds up to time t_{i-1} included, that is, the execution trace defined by (3.15) is correct up to state $\sigma^{t_{i-1}}$, and z_q and x_p are the counters of this Petri net execution up to time t_{i-1} included.

The time t_i is defined as the first time of discontinuity of some $z_q(t)$ after t_{i-1}. If the sequence $t_0, t_1, t_2, \dots$ is finite, and if t_{i-1} is its last element, we also define $t_i = +\infty$. In this case $\sigma^{t_i^-}$ shall stand for the state of the execution trace obtained after $\sigma^{t_{i-1}}$ and an infinite duration transition $\rightarrow^{+\infty}$.

We prove that the execution trace is correct up to $\sigma^{t_i^-}$, that is, we prove that no transition could be fired in the Petri net in the time interval $]t_{i-1}, t_i[$. Suppose that this is not true, and let $t_f \in]t_{i-1}, t_i[$ be the first instant when a transition should fire after t_{i-1} in a correct Petri net execution, and q be such transition. Let p be a limiting place of q at time t_{i-1}. As the number of tokens of age $\geqslant \tau_p$ in p at time t_{i-1} is null, there is necessarily a token entering place p in the time interval $]t_{i-1} - \tau_p, t_f - \tau_p]$, and we define t_e as being the date of its entrance. If $t_e > t_{i-1}$, then this would mean that one of the upstream transitions of p fires at time t_e, $t_e \in]t_{i-1}, t_f - \tau_p]$. This contradicts the minimality of t_f. On the other hand, if $t_e \leqslant t_{i-1}$, this token entrance should be associated with an increase of $x_p(t_e)$, and therefore, with an increase of $z_q(t_e + \tau_p)$. But $t_e + \tau_p \leqslant t_f < t_i$, and $t_e + \tau_p > t_{i-1}$, so that it would contradict the minimality of t_i. Finally, this proves that no transition can fire before time t_i in the Petri net execution, and therefore, that σ is correct up to state $\sigma^{t_i^-}$.

In addition, the functions z_q have no discontinuities in the interval $]t_i, t_{i-1}[$, so that $z_q(t_{i-1}) = z_q(t) = z_q(t_i^-)$, for any t in the interval. Therefore, the z_q have the correct meaning (no transition fires in the time interval). For the x_p, the result follows from the relation (3.3), and from the induction hypothesis.

Now, let us prove the correctness of the transition firings described by σ at time t_i. We distinguish the following cases:

- Let $p \in \mathcal{P}_{\mathsf{conflict}}$. The number of tokens of age greater than or equal to τ_p in place p in state $\sigma^{t_i^-}$ is given by $x_p(t_i - \tau_p) - \sum_{q \in p^{\text{out}}} z_q(t_i^-)$, by the induction hypothesis. A correct execution at time t_i should be that these tokens (if any) are fired by the output transitions of p, whatever the distribution and sequencing of these transition firings (otherwise, just after t_i,

there are tokens of age greater than τ_p in place p, while any downstream transition of p can be fired, which would contradict the earliest behavior semantics). Therefore, the execution is correct if and only if

$$\sum_q z_q(t_i) - z_q(t_i^-) = x_p(t_i - \tau_p) - \sum_{q \in p^{\text{out}}} z_q(t_i^-)\,,$$

the quantity $z_q(t_i) - z_q(t_i^-)$ being equal to the number of firings of q at time t_i in the execution trace (3.15). By (3.4) for $t = t_i$, the above equality holds, and therefore, the execution is correct for this part of the Petri net.

• Let $q \in \mathcal{Q}_{\mathsf{sync}}$. At state $\sigma^{t_i^-}$, the number of tokens of age greater than or equal to τ_p in each upstream place p of q is given by $x_p(t_i - \tau_p) - z_q(t_i^-)$, by the induction hypothesis. Because of the earliest behavior semantics, the number of firings of transition q at time t_i is given by the minimal number of tokens of sufficient age in its upstream places, that is,

$$\bigwedge_{p \in q^{\text{in}}} \left(x_p(t_i - \tau_p) - z_q(t_i^-)\right)\,.$$

By (3.5), this is also equal to $z_q(t_i) - z_q(t_i^-)$. This is precisely the number of occurrences of $\rightarrow^q$ in σ at time t_i.

• Let $p \in \mathcal{P}_{\mathsf{priority}}$, and p_+^{out} its priority output transition. At state $\sigma^{t_i^-}$, the number of tokens of age greater than or equal to τ_p in an upstream place $r \in (p_+^{\text{out}})^{\text{in}}$, $r \neq p$, is $x_r(t_i - \tau_r) - z_{p_+^{\text{out}}}(t_i^-)$, by correctness of the counters for t up to t_i (not included). In place p, this number is $x_p(t - \tau_p) - z_{p_+^{\text{out}}}(t_i^-) - z_{p_-^{\text{out}}}(t_i^-)$. Therefore, under our earliest behavior semantics and priority routing, the number of tokens that should be fired by priority transition p_+^{out} is

$$\left(x_p(t_i - \tau_p) - z_{p_+^{\text{out}}}(t_i^-) - z_{p_-^{\text{out}}}(t_i^-)\right) \wedge \bigwedge_{r \in (p_+^{\text{out}})^{\text{in}}, r \neq p} \left(x_r(t_i - \tau_r) - z_q(t_i^-)\right)\,.$$

By (3.5), this is also equal to $z_{p_+^{\text{out}}}(t_i) - z_{p_+^{\text{out}}}(t_i^-)$. This is precisely the number of occurrence of $\rightarrow^{p_+^{\text{out}}}$ in σ at time t_i. In addition, these transitions appear before the transitions of type $\xrightarrow{p_-^{\text{out}}}$ in σ.

Now, consider p_-^{out} the non priority transition of p, and let us consider the number of tokens in place p at time t_i, just after the firings of transition p_+^{out} (if any). In this state, the number of tokens of age greater than or equal to τ_p is number of tokens in this place in state $\sigma^{t_i^-}$, which equals $x_p(t - \tau_p) - z_{p_+^{\text{out}}}(t_i^-) - z_{p_-^{\text{out}}}(t_i^-)$ by the induction hypothesis, minus the number of firings of transition p_+^{out} at this time, which we just proved to be correctly equal to $z_{p_+^{\text{out}}}(t_i) - z_{p_+^{\text{out}}}(t_i^-)$. The other places upstream p_-^{out} have no other output transition, and therefore, the number of tokens of age greater than or equal to τ_p is $x_r(t_i - \tau_r) - z_{p_-^{\text{out}}}(t_i^-)$. Finally, the number of tokens that transition p_-^{out} can fire (and, because of the earliest behavior semantics, has to fire) is

$$(x_p(t_i - \tau_p) - z_{p_+^{\text{out}}}(t_i) - z_{p_-^{\text{out}}}(t_i^-)) \wedge \bigwedge_{r \in (p_-^{\text{out}})^{\text{in}}, r \neq p} \left(x_r(t_i - \tau_r) - z_{p_-^{\text{out}}}(t_i^-)\right)\,.$$

This is precisely $z_{p_-^{\text{out}}}(t_i) - z_{p_-^{\text{out}}}(t_i^-)$ by (3.7), so that the number of firings of p_-^{out} at time t_i equals the number of occurrences of $\rightarrow^{p_-^{\text{out}}}$ at time t_i.

The sets $(\mathcal{P}_{\mathsf{conflict}})^{\text{out}}$, $(\mathcal{P}_{\mathsf{priority}})^{\text{out}}$ and $\mathcal{Q}_{\mathsf{sync}}$ describe the whole $\mathcal{Q}$. As a conclusion, all the transitions appearing between $\sigma^{t_i^-}$ and σ^{t_i} correspond to appropriate firings of a Petri net execution, in a correct order, and none is missing. Moreover, for any transition q, the difference $z_q(t_i) - z_q(t_i^-)$ correctly counts the number of transitions occurring at time t_i, and consequently, by (3.3), for any place p, $x_p(t_i) - x_p(t_i^-)$ correctly counts the number of tokens entering in place p at time t_i. This completes the proof. ◀

3.3 Computing stationary regimes

We investigate the stationary regimes of the fluid dynamics associated with Petri nets with free choice and priority routing. More specifically, our goal is to characterize the non-decreasing

càdlàg solutions x_p and z_q of the dynamics which behave ultimately as affine functions $t \mapsto u+\rho t$ ($u \in \mathbb{R}$ and $\rho \in \mathbb{R}_{\geqslant 0}$). By *ultimately*, we mean that the property holds for t large enough. In this case, the scalar ρ corresponds to the asymptotic throughput of the associated place or transition. However, if the functions x_p and z_q are continuous, and a fortiori if they are affine, their values at points t and t^- coincide, and then, the effect of the priority rule on the dynamics vanishes (see Equation (3.6)). Hence, looking for ultimately affine solutions of the continuous time equations might look as an ill-posed problem, if one interprets it in a naive way. In contrast, looking for the ultimately affine solutions of the δ-discretization of the fluid dynamics is a perfectly well-posed problem. In other words, we aim at determining the solutions x_p and z_q of the discrete dynamics which coincide with affine functions at points $k\delta$ for all sufficiently large $k \in \mathbb{N}$. These solutions are referred to as the ***stationary solutions*** of the dynamics. As we shall prove in Theorem 3.6, the characterization of these solutions does not depend on the value of δ, leading to a proper definition of ultimately affine solutions of the continuous time dynamics.

In order to determine the stationary regimes, we use the notion of germs of affine functions. We introduce an equivalence relation $\sim$ over functions from $\mathbb{R}$ to itself, defined by $f \sim g$ if $f(t)$ and $g(t)$ are equal for all $t \in \delta\mathbb{N}$ sufficiently large. A ***germ of function*** (at point infinity) is an equivalence class of functions with respect to the relation $\sim$. For brevity, we refer to the germs of affine functions as ***affine germs***, and we denote by (ρ, u) the germ of the function $t \mapsto u + \rho t$. In this setting, our goal is to determine the affine germs of the counter variables of the Petri net in the stationary regimes.

Given two functions f and g of affine germs (ρ, u) and (ρ', u') respectively, it is easy to show that $f(t) \leqslant g(t)$ for all sufficiently large $t \in \delta\mathbb{N}$ if, and only if, the couple (ρ, u) is smaller than or equal to (ρ', u') in the lexicographic order. Moreover, the affine germ of the function $f + g$ is simply given by the germ $(\rho + \rho', u + u')$, which we denote by $(\rho, u) + (\rho', u')$ by abuse of notation. As a consequence, affine germs provide an ordered group. Let us add to this group a greatest element $\top$, with the convention that $\top + (\rho, u) = (\rho, u) + \top = \top$. Then, we obtain the *tropical (min-plus) semiring of affine germs* $(\mathbb{G}, \wedge, +)$, where $\mathbb{G}$ is defined as $\{\top\} \cup \mathbb{R}^2$, and for all $x, y \in \mathbb{G}$, $x \wedge y$ stands for the minimum of x and y in lexicographic order (extended to $\top$). Since in $\mathbb{G}$, the addition plays the role of the multiplicative law, the additive inversion defined by $-(\rho, u) := (-\rho, -u)$ corresponds to a division over $\mathbb{G}$. This makes $\mathbb{G}$ a semifield, *i.e.*, in loose terms, a structure similar to a field, except that the additive law has no inverse. Finally, we can define the multiplication by a scalar $\lambda \in \mathbb{R}$ by $\lambda(\rho, u) := (\lambda\rho, \lambda u)$. When $\lambda \in \mathbb{N}$, this can be understood as an exponentiation operation in $\mathbb{G}$.

Instantiating the functions x_p and z_q by affine asymptotics $t \mapsto u_p + t\rho_p$ and $t \mapsto u_q + t\rho_q$ in the δ-discretization of the fluid dynamics leads to the following counterparts of the constraints (3.3), (3.5), (3.7) and (3.9), the variables being now elements of the semifield $\mathbb{G}$ of germs:

$$\forall p \in \mathcal{P}\,, \qquad (\rho_p, u_p) = (0, M_p) + \sum_{q \in p^{\mathrm{in}}} (\rho_q, u_q) \tag{3.16a}$$

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \forall q \in p^{\mathrm{out}}\,, \qquad (\rho_q, u_q) = \pi_{qp}(\rho_p, u_p - \rho_p\tau_p) \tag{3.16b}$$

$$\forall q \in \mathcal{Q}_{\mathsf{sync}}\,, \qquad (\rho_q, u_q) = \bigwedge_{p \in q^{\mathrm{in}}} (\rho_p, u_p - \rho_p\tau_p) \tag{3.16c}$$

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,, \qquad (\rho_{p_-^{\mathrm{out}}}, u_{p_-^{\mathrm{out}}}) = (\rho_p - \rho_{p_+^{\mathrm{out}}}, u_p - \rho_p\tau_p - u_{p_+^{\mathrm{out}}}) \wedge \bigwedge_{r \in (p_-^{\mathrm{out}})^{\mathrm{in}}\,,\, r \neq p} (\rho_r, u_r - \rho_r\tau_r) \tag{3.16d}$$

Given $p \in \mathcal{P}_{\mathsf{priority}}$, the transposition of (3.6) (or equivalently (3.8)) to germs is more elaborate due to the occurrence of the left limit $x_{p_-^{\mathrm{out}}}(t^-)$. We obtain:

$$(\rho_{p_+^{\mathrm{out}}}, u_{p_+^{\mathrm{out}}}) = \begin{cases} (\rho_p - \rho_{p_-^{\mathrm{out}}}, u_p - \rho_p\tau_p - u_{p_-^{\mathrm{out}}}) \\ \quad \wedge \displaystyle\bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}}\,,\, r \neq p} (\rho_r, u_r - \rho_r\tau_r) & \text{if } \rho_{p_-^{\mathrm{out}}} = 0\,, \\ \displaystyle\bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}}\,,\, r \neq p} (\rho_r, u_r - \rho_r\tau_r) & \text{otherwise.} \end{cases} \tag{3.16e}$$

The correctness of these constraints is stated in the following result:

▶ **Theorem 3.6.** *The affine germs of the stationary solutions of the δ-discretization of the fluid dynamics are precisely the solutions of System (3.16) such that $\rho_p, \rho_q \geqslant 0$ ($p \in \mathcal{P}$, $q \in \mathcal{Q}$).*

Proof. We denote the lexicographic order over $\mathbb{R}^2$ by $\preccurlyeq$, and we use the notation $x \prec y$ when $x \preccurlyeq y$ and $x \neq y$.

We first remark that given two functions f and g of affine germs (ρ, u) and (ρ', u'), the function $t \mapsto f(t) \wedge g(t)$ belongs to the affine germ given by the minimum $(\rho, u) \wedge (\rho', u')$ taken in the lexicographic order. Besides, if $\tau \in \delta\mathbb{N}$, the function $t \mapsto f(t-\tau)$ belongs to the affine germ $(\rho, u - \tau\rho)$. Finally, for all $\lambda \in \mathbb{R}$, the affine germ of the map $t \mapsto \lambda f(t)$ is equal to $\lambda(\rho, u) = (\lambda\rho, \lambda u)$.

Now, suppose that x_p and z_q are stationary solutions of the δ-discretization of the fluid dynamics, and let (ρ_p, u_p) and (ρ_q, u_q) be the respective germs, for $p \in \mathcal{P}$ and $q \in \mathcal{Q}$. Since the functions x_p and z_q satisfy the constraints (3.3), (3.5), (3.7), (3.9) for all $t \in \delta\mathbb{N}$, we deduce from the previous properties that the constraints (3.16a)–(3.16d) are satisfied. Besides, given $p \in \mathcal{P}_{\mathsf{priority}}$, (3.8) ensures that for all $t \in \delta\mathbb{N}$, we have:

$$z_{p_+^{\mathrm{out}}}(t) = \big(x_p(t-\tau_p) - z_{p_-^{\mathrm{out}}}(t-\delta)\big) \wedge \bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}},\, r \neq p} x_r(t-\tau_r)\,.$$

Consequently, we obtain

$$(\rho_{p_+^{\mathrm{out}}}, u_{p_+^{\mathrm{out}}}) = (\rho_p - \rho_{p_-^{\mathrm{out}}}, u_p - \rho_p\tau_p - u_{p_-^{\mathrm{out}}} + \rho_{p_-^{\mathrm{out}}}\delta) \wedge \bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}},\, r \neq p} (\rho_r, u_r - \rho_r\tau_r)\,. \quad (3.17)$$

If $\rho_{p_-^{\mathrm{out}}} = 0$, this amounts to the constraint given in (3.16e). Now consider the case where $\rho_{p_-^{\mathrm{out}}} > 0$. Using (3.16d), we know that

$$(\rho_{p_-^{\mathrm{out}}}, u_{p_-^{\mathrm{out}}}) \preccurlyeq (\rho_p - \rho_{p_+^{\mathrm{out}}}, u_p - \rho_p\tau_p - u_{p_+^{\mathrm{out}}})\,,$$

and thus

$$(\rho_{p_+^{\mathrm{out}}}, u_{p_+^{\mathrm{out}}}) \preccurlyeq (\rho_p - \rho_{p_-^{\mathrm{out}}}, u_p - \rho_p\tau_p - u_{p_-^{\mathrm{out}}})\,,$$

Since $\delta > 0$, it follows that

$$(\rho_{p_+^{\mathrm{out}}}, u_{p_+^{\mathrm{out}}}) \prec (\rho_p - \rho_{p_-^{\mathrm{out}}}, u_p - \rho_p\tau_p - u_{p_-^{\mathrm{out}}} + \rho_{p_-^{\mathrm{out}}}\delta)\,.$$

We conclude that the constraint (3.17) is equivalent to (3.16e) when $\rho_{p_-^{\mathrm{out}}} > 0$.

Conversely, let (ρ_p, u_p) and (ρ_q, u_q) be solutions of System (3.16). We define x_p and z_q as the functions given by $x_p(t) = u_p + \rho_p k\delta$ and $z_q(t) = u_q + \rho_q k\delta$ for all $t \in [k\delta, (k+1)\delta)$ and $k \in \mathbb{N}$. The constraints (3.16a)–(3.16d) ensure that (3.3), (3.5), (3.7), (3.9) hold for all $t \in \delta\mathbb{N}$. Since all the holding times τ_p belong to $\delta\mathbb{N}$ and the functions x_p and z_q are constant on the intervals of the form $[k\delta, (k+1)\delta)$, we deduce that these constraints (3.3), (3.5), (3.7), (3.9) actually hold for all t in such intervals, and so for all $t \geqslant 0$. Moreover, as previously shown, the constraint given in (3.16e) is equivalent to (3.17), since (3.16d) is satisfied. This proves that the constraint (3.8) holds for all $t \in \delta\mathbb{N}$. It remains to show that the latter constraint is satisfied when $t \in (k\delta, (k+1)\delta)$. First observe that $z_{p_+^{\mathrm{out}}}(k\delta) \leqslant \bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}},\, r \neq p} x_r(k\delta - \tau_r)$ ensures that $z_{p_+^{\mathrm{out}}}(t) \leqslant \bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}},\, r \neq p} x_r(t-\tau_r)$. Besides, by (3.7), we know that $z_{p_+^{\mathrm{out}}}(t) \leqslant x_p(t-\tau_p) - z_{p_-^{\mathrm{out}}}(t)$. We now distinguish two cases:

(i) if we have $z_{p_+^{\mathrm{out}}}(k\delta) = \bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}},\, r \neq p} x_r(k\delta - \tau_r)$, then straightforwardly, $z_{p_+^{\mathrm{out}}}(t) = \bigwedge_{r \in (p_+^{\mathrm{out}})^{\mathrm{in}},\, r \neq p} x_r(t-\tau_r)$.

(ii) if $z_{p_+^{\mathrm{out}}}(k\delta) = x_p(k\delta - \tau_p) - z_{p_-^{\mathrm{out}}}((k-1)\delta)$, we obtain:

$$z_{p_+^{\mathrm{out}}}(k\delta) \geqslant x_p(k\delta - \tau_p) - z_{p_-^{\mathrm{out}}}(k\delta) \geqslant z_{p_+^{\mathrm{out}}}(k\delta)\,,$$

where the first inequality comes from the fact that $z_{p_-^{\mathrm{out}}}$ is non-decreasing, and the second inequality from (3.7). We deduce that $z_{p_+^{\mathrm{out}}}(k\delta) = x_p(k\delta - \tau_p) - z_{p_-^{\mathrm{out}}}(k\delta)$. Hence, we get $z_{p_+^{\mathrm{out}}}(t) = x_p(t-\tau_p) - z_{p_-^{\mathrm{out}}}(t)$.

As a consequence, in both cases, we have proved that $z_{p_+^{\text{out}}}(t)$ is the minimum between $z_{p_+^{\text{out}}}(t) - z_{p_-^{\text{out}}}(t)$ and $\bigwedge_{r \in (p_+^{\text{out}})^{\text{in}}, r \neq p} x_r(t - \tau_r)$. This shows that (3.8) holds for all $t \geqslant 0$. ◀

Since the expressions at the right hand side of the constraints of System (3.16) involve minima of linear terms, these expressions can be interpreted as fractional functions over the tropical semifield $\mathbb{G}$. In this way, System (3.16) can be thought of as a set of tropical polynomial constraints (or more precisely, rational constraints).

The solutions of tropical polynomial systems is a topic of current interest, owing to its relations with fundamental algorithmic issues concerning classical polynomial system solving over the reals. Here, we describe a simple method to solve System (3.16), which is akin to *policy search* in stochastic control. Observe that System (3.16) corresponds to a fixed point equation $(\rho, u) = f(\rho, u)$, where the function f can be expressed as the infimum $\bigwedge_\pi f^\pi$ of finitely many linear (affine) maps f^π. In more details, every function f^π is obtained by selecting one term for each minimum operation $\bigwedge$ occurring in the constraints (for instance, in (3.16c), we select one term $(\rho_p, u_p - \rho_p \tau_p)$ with $p \in q^{\text{in}}$). For every selection π, we can solve the associated linear system $(\rho, u) = f^\pi(\rho, u)$, and under some structural assumptions on the Petri net, the solution (ρ^π, u^π) is unique [1]. If $f^\pi(\rho^\pi, u^\pi) = f(\rho^\pi, u^\pi)$, *i.e.* in every constraint, the term we selected is smaller than or equal to the other terms appearing in the minimum, then (ρ^π, u^π) forms a solution of System (3.16) associated with the selection π. Otherwise, the selection π does not lead to any solution. Iterating this technique over the set of selections provides all the solutions of System (3.16). Every iteration can be done in polynomial time. However, since there is an exponential number of possible selections, the overall time complexity of the method is exponential in the size of the Petri net.

Section 3.5 provides an example in which several policies provide several valid, but different solutions.

3.4 Application: the emergency call center PN

We now apply the results of Section 3.3 to determine the stationary regimes of the fluid dynamics associated with our timed Petri net model of emergency call center. As in Section 3.2.4, we consider the subsystem reduced to the variables z_1, z_5 and z_6. The corresponding system of constraints over the germ variables (u_1, ρ_1), (u_5, ρ_5) and (u_6, ρ_6) is given by:

$$(\rho_1, u_1) = \big(\rho_5 + \mu_{\text{ur}}\rho_1 + \mu_{\text{adv}}\rho_1, \quad N_1 + (u_5 - \rho_5\tau_{\text{tr}}) + \mu_{\text{ur}}(u_1 - \rho_1\tau_{\text{ur}}) + \mu_{\text{adv}}(u_1 - \rho_1\tau_{\text{adv}})\big) \tag{3.18a}$$

$$(\rho_5, u_5) = \begin{cases} \big(\rho_5, N_2 + u_5 - \rho_5(\tau_{\text{tr}} + \tau'_{\text{ext}})\big) \wedge \mu_{\text{ext}}(\rho_1, u_1 - \rho_1\tau_{\text{ext}}) & \text{if } \rho_6 = 0 \\ \mu_{\text{ext}}(\rho_1, u_1 - \rho_1\tau_{\text{ext}}) & \text{if } \rho_6 > 0 \end{cases} \tag{3.18b}$$

$$(\rho_6, u_6) = \big(\rho_6, N_2 - \rho_5(\tau_{\text{tr}} + \tau'_{\text{ext}}) + (u_6 - \rho_6\tau'_{\text{ur}})\big) \wedge \mu_{\text{ur}}(\rho_1, u_1 - \rho_1\tau_{\text{ur}}) \tag{3.18c}$$

To solve this system, it is convenient to introduce the following quantity

$$\bar{\tau} := \mu_{\text{ext}}(\tau_{\text{ext}} + \tau_{\text{tr}}) + \mu_{\text{ur}}\tau_{\text{ur}} + \mu_{\text{adv}}\tau_{\text{adv}} \,,$$

which represents the average time of treatment of a call at level 1 of the model. Note that we exclude the trivial case where $\rho_1 = 0$ (and subsequently $\rho_5 = \rho_6 = 0$), since it cannot occur unless the quantity N_1 is null.

The ρ-part of (3.18a) and (3.18c) show that

$$\rho_5 = \mu_{\text{ext}}\rho_1 \,, \qquad 0 \leqslant \rho_6 \leqslant \mu_{\text{ur}}\rho_1 \,.$$

We start by considering the case where $\rho_6 = 0$. Since $\rho_1 > 0$, the minimum in (3.18c) is necessarily attained by the left term. From this, we deduce

$$\rho_1 = \frac{N_2}{\mu_{\text{ext}}(\tau_{\text{tr}} + \tau'_{\text{ext}})} \,.$$

1. More on this in Chapter 4, in which an equivalent formula is used to compute stationary affine solutions.

Table 3.1 – The normalized throughputs ρ_1, ρ_5 and ρ_6 as piecewise linear functions of N_2/N_1.

	$0 \leqslant N_2/N_1 \leqslant r_1$	$r_1 \leqslant N_2/N_1 \leqslant r_2$	$r_2 \leqslant N_2/N_1$
ρ_1/ρ^*	$\dfrac{\bar\tau}{\mu_{\text{ext}}(\tau_{\text{tr}}+\tau'_{\text{ext}})}\dfrac{N_2}{N_1}$	1	1
ρ_5/ρ^*	$\dfrac{\bar\tau}{\tau_{\text{tr}}+\tau'_{\text{ext}}}\dfrac{N_2}{N_1}$	μ_{ext}	μ_{ext}
ρ_6/ρ^*	0	$\dfrac{\bar\tau}{\tau'_{\text{ur}}}\Big(\dfrac{N_2}{N_1}-r_1\Big)$	μ_{ur}

As $(\rho_5, u_5) \leqslant \mu_{\text{ext}}(\rho_1, u_1 - \rho_1\tau_{\text{ext}})$ (by (3.18b)) and $\rho_5 = \mu_{\text{ext}}\rho_1$, the inequality $u_5 \leqslant \mu_{\text{ext}}(u_1 - \rho_1\tau_{\text{ext}})$ holds. Using the u-part of (3.18a), we can show that this amounts to the inequality

$$\frac{N_2}{N_1} \leqslant r_1 := \frac{\mu_{\text{ext}}(\tau_{\text{tr}}+\tau'_{\text{ext}})}{\bar\tau}\,.$$

We now assume that $\rho_6 > 0$. The fact that $u_5 = \mu_{\text{ext}}(u_1 - \rho_1\tau_{\text{ext}})$ (by (3.18b)) leads to the identity

$$\rho_1 = \frac{N_1}{\bar\tau}\,.$$

It remains to distinguish the subcases corresponding to the minimum in (3.18c).

- Suppose that the minimum is attained by the left term. We deduce that:

$$\rho_6 = \frac{N_2}{\tau'_{\text{ur}}} - \frac{N_1}{\bar\tau}\frac{\mu_{\text{ext}}(\tau_{\text{tr}}+\tau'_{\text{ext}})}{\tau'_{\text{ur}}} = \frac{N_2 - N_1 r_1}{\tau'_{\text{ur}}}\,.$$

Since $0 < \rho_6 \leqslant \mu_{\text{ur}}\rho_1$, we also derive:

$$r_1 < \frac{N_2}{N_1} \leqslant r_2 := \frac{\mu_{\text{ext}}(\tau_{\text{tr}}+\tau'_{\text{ext}}) + \mu_{\text{ur}}\tau'_{\text{ur}}}{\bar\tau}\,.$$

- If the minimum is reached by the right term, then we have $\rho_6 = \mu_{\text{ur}}\rho_1$, or equivalently $\rho_6 = \mu_{\text{ur}}\frac{N_1}{\bar\tau}$. Moreover, we necessarily have $u_6 \leqslant N_2 - \rho_5(\tau_{\text{tr}}+\tau'_{\text{ext}}) + (u_6 - \rho_6\tau'_{\text{ur}})$, which provides $\frac{N_2}{N_1} \geqslant r_2$. Note that the latter inequality is strict as soon as the minimum in (3.18c) is attained by the right term only.

To summarize, we report the possible values of the throughputs ρ_1, ρ_5 and ρ_6 in Table 3.1 in the stationary regimes. We normalize these values by a quantity ρ^* which corresponds to the throughput (of transition q_1) in an "ideal" call center which involves as many level 2 operators as necessary, *i.e.* $N_2 = +\infty$. Then, the throughput ρ^* is given by $N_1/\bar\tau$, where $\bar\tau := \mu_{\text{ext}}(\tau_{\text{ext}}+\tau_{\text{tr}}) + \mu_{\text{ur}}\tau_{\text{ur}} + \mu_{\text{adv}}\tau_{\text{adv}}$ represents the average time of treatment at level 1.

As shown in Table 3.1, the ratios ρ_1/ρ^*, ρ_5/ρ^* and ρ_6/ρ^* are piecewise linear functions of the ratio N_2/N_1. The non-differentiability points are given by:

$$r_1 := \frac{\mu_{\text{ext}}(\tau_{\text{tr}}+\tau'_{\text{ext}})}{\bar\tau} \qquad r_2 := \frac{\mu_{\text{ext}}(\tau_{\text{tr}}+\tau'_{\text{ext}}) + \mu_{\text{ur}}\tau'_{\text{ur}}}{\bar\tau}\,.$$

They separate three phases:

(i) when N_2/N_1 is strictly smaller than r_1, the number of level 2 operators is so small that some extremely urgent calls cannot be handled, and no urgent call is handled. This is why the throughput of the latter calls at level 2 is null. Also, level 1 operators are slowed down by the congestion of level 2, since, in the treatment of an extremely urgent call, a level 1 operator cannot be released until the call is handled by a level 2 operator.

(ii) when N_2/N_1 is between r_1 and r_2, there are enough level 2 operators to handle all the extremely urgent calls, which is why the throughput ρ_5 is equal to ρ_1 multiplied by the proportion μ_{ext} of extremely urgent calls. As a consequence, level 2 is no longer slowing down level 1 (the throughput ρ_1 reaches its maximal value ρ^*). However, the throughput of urgent calls at level 2 is still limited because N_2 is not sufficiently large.

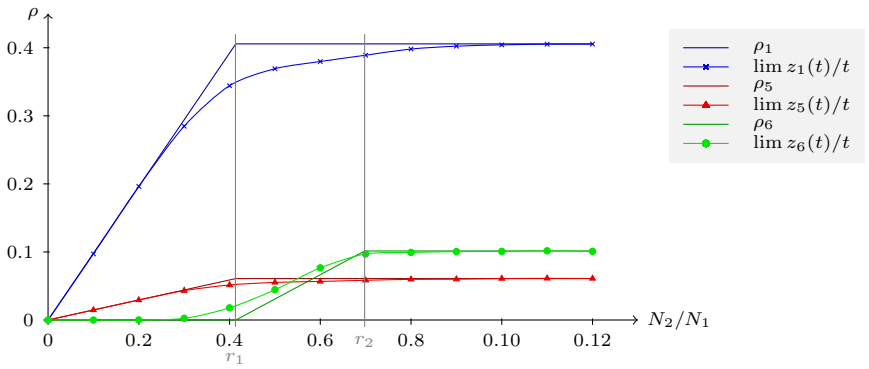


Figure 3.2 – Comparison of the throughputs of the non-fluid simulations with the theoretical throughputs (fluid model). The three phases are identified by two vertical lines.

(iii) if N_2/N_1 is larger than r_2, the three throughputs reach their maximal values. This means that level 2 is sufficiently well-staffed w.r.t. level 1.

This analysis provides a qualitative method to determine an optimal dimensioning of the system in stationary regimes. Given a fixed N_1, the number N_2 of level 2 operators should be taken to be the minimal integer such that $N_2/N_1 \geqslant r_2$. This ensures that the level 2 properly handles the calls transmitted by the level 1 (all calls are treated). Then, N_1 should be the minimal integer such that $\rho_1 = N_1/\bar{\tau}$ dominates the arrival rate of calls.

3.5 Application: the SR Petri net

The Petri net of Silva and Recalde (2002) is introduced in Section 2.5.2. Note that this Petri net has no conflict configuration (in the sense of Figure 3.1(a)), and therefore, no randomized feature.

We extend the equations of the fluid dynamics (3.3)–(3.7) and (3.9) to Petri net with valuations in a straightforward way. Note that this leads to an additional fluid approximation. Indeed, in a discrete setting, a place-transition arc with valuation 2 implies that a firing is possible only when two tokens are available in the corresponding place. In contrast, in the counter equations of the fluid dynamics, the transition is allowed to fire even if there are less than two tokens in the place. Of course, the proportions of token consumed in upstream places are preserved.

The fluid approximation of the discrete dynamics is given by the following system:

$$\begin{aligned} Z_1(t) &= \tfrac{1}{2}\left(N_1 + Z_3(t-\tau_1) + 2Z_4(t-\tau_1) - Z_2(t^-)\right) \wedge N_2 + Z_2(t-\tau_2) \\ Z_2(t) &= \left(N_1 + Z_3(t-\tau_1) + 2Z_4(t-\tau_1) - 2Z_1(t)\right) \wedge N_3 + Z_1(t-\tau_3) \\ Z_3(t) &= N_4 + Z_1(t-\tau_4) \\ Z_4(t) &= N_5 + Z_2(t-\tau_5)\,, \end{aligned} \tag{3.19}$$

which can be simplified in a system in Z_1 and Z_2 only:

$$\begin{aligned} Z_1(t) &= \tfrac{1}{2}\left(N_1 + N_4 + 2N_5 + Z_1(t-\tau_1-\tau_4) + 2Z_2(t-\tau_1-\tau_5) - Z_2(t^-)\right) \wedge N_2 + Z_2(t-\tau_2) \\ Z_2(t) &= \left(N_1 + N_4 + 2N_5 + Z_1(t-\tau_1-\tau_4) + 2Z_2(t-\tau_1-\tau_5) - 2Z_1(t)\right) \wedge N_3 + Z_1(t-\tau_3)\,. \end{aligned}$$

The system of equations over the germ variables (ρ, u) is given by:

$$\begin{aligned} (\rho_1, u_1) &= \begin{cases} \left(\tfrac{1}{2}(\rho_1+\rho_2), \tfrac{1}{2}(N_1 + N_4 + 2N_5 + u_1 + u_2 - \rho_1(\tau_1+\tau_4) - 2\rho_2(\tau_1+\tau_5)\right) \wedge \\ \qquad\qquad (\rho_2, N_2 + u_2 - \rho_2\tau_2) \quad \text{if } \rho_2 = 0 \\ (\rho_2, N_2 + u_2 - \rho_2\tau_2) \quad \text{if } \rho_2 > 0 \end{cases} \\ (\rho_2, u_2) &= (2\rho_2 - \rho_1, N_1 + N_4 + 2N_5 + 2u_2 - u_1 - \rho_1(\tau_1+\tau_4) - 2\rho_2(\tau_1+\tau_5)) \wedge \\ &\qquad\qquad (\rho_1, N_3 + u_1 - \rho_1\tau_3)\,. \end{aligned} \tag{3.20}$$

	$y_1^\mathsf{T}N^0 \leqslant y_2^\mathsf{T}N^0$		$y_1^\mathsf{T}N^0 \geqslant y_2^\mathsf{T}N^0$	
	$\pi_1\tau \leqslant \pi_2\tau$	$\pi_1\tau \geqslant \pi_2\tau$	$\pi_1\tau \leqslant \pi_2\tau$	$\pi_1\tau \geqslant \pi_2\tau$
$\frac{y_1^\mathsf{T}N^0}{\pi_1\tau} \leqslant \frac{y_2^\mathsf{T}N^0}{\pi_2\tau}$	0		infeasible	$\frac{(y_1-y_2)^\mathsf{T}N^0}{(\pi_1-\pi_2)\tau}$
$\frac{y_1^\mathsf{T}N^0}{\pi_1\tau} \geqslant \frac{y_2^\mathsf{T}N^0}{\pi_2\tau}$	0 or $\frac{y_2^\mathsf{T}N^0}{\pi_2\tau}$ or $\frac{(y_1-y_2)^\mathsf{T}N^0}{(\pi_1-\pi_2)\tau}$	infeasible	$\frac{y_2^\mathsf{T}N^0}{\pi_2\tau}$	

Table 3.2 – Throughput of the SR Petri net, depending on τ and N^0.

Solving the system in the ρ coordinate yields $\rho_1 = \rho_2$, and equality in each side of the minimum. Therefore, denoting ρ this common value, we have the following system in u:

$$u_1 = \begin{cases} \frac{1}{2}\left(N_1 + N_4 + 2N_5 + u_1 + u_2\right) \wedge N_2 + u_2 & \text{if } \rho = 0 \\ N_2 + u_2 - \rho\tau_2 & \text{if } \rho > 0 \end{cases} \tag{3.21a}$$

$$u_2 = N_1 + N_4 + 2N_5 + 2u_2 - u_1 - \rho(3\tau_1 + \tau_4 + 2\tau_5) \wedge N_3 + u_1 - \rho\tau_3\,. \tag{3.21b}$$

If $\boldsymbol{\rho = 0}$, then (3.21a) yields

$$u_1 - u_2 = N_2 \wedge N_1 + N_4 + 2N_5$$

and (3.21b) yields

$$\begin{cases} u_2 - u_1 = N_3 \\ u_1 - u_2 \leqslant N_1 + N_4 + 2N_5 \end{cases} \quad \text{or} \quad \begin{cases} u_1 - u_2 = N_1 + N_4 + 2N_5 \\ u_2 - u_1 \leqslant N_3 \end{cases}\,.$$

Therefore, either $u_2 - u_1 = N_1 + N_4 + 2N_5 \leqslant N_2$ and $N_1 + N_4 + 2N_5 \geqslant -N_3$. Or $u_2 - u_1 = -N_2 = N_3 \leqslant N_1 + N_4 + 2N_5$, but then $N_2 = N_3 = 0$. The condition is hence $\rho = 0 \Rightarrow N_1 - N_2 + N_4 + 2N_5 \leqslant 0$ or $N_2 = N_3 = 0$.

If $\boldsymbol{\rho > 0}$, $u_2 - u_1 = -N_2 + \rho\tau_2$ and, either $u_1 - u_2 = N_1 + N_4 + 2N_5 - \rho(3\tau_1 + \tau_4 + 2\tau_5) \geqslant -N_3 + \rho\tau_3$, or $u_1 - u_2 = -N_3 + \rho\tau_3 \leqslant N_1 + N_4 + 2N_5 - \rho(3\tau_1 + \tau_4 + 2\tau_5)$.
In any case, we have

$$\rho \leqslant \frac{N_1 + N_3 + N_4 + 2N_5}{3\tau_1 + \tau_3 + \tau_4 + 2\tau_5}\,.$$

In the first case, if $(3\tau_1 - \tau_2 + \tau_4 + 2\tau_5) \neq 0$ this yields

$$\rho = \frac{N_1 - N_2 + N_4 + 2N_5}{3\tau_1 - \tau_2 + \tau_4 + 2\tau_5} \leqslant \frac{N_2 + N_3}{\tau_2 + \tau_3}\,.$$

Of course, this is possible only if $(3\tau_1 - \tau_2 + \tau_4 + 2\tau_5)(N_1 - N_2 + N_4 + 2N_5) \geqslant 0$.
In the second case, we have

$$\rho = \frac{N_2 + N_3}{\tau_2 + \tau_3} \text{ and } \rho(3\tau_1 - \tau_2 + \tau_4 + 2\tau_5) \leqslant N_1 - N_2 + N_4 + 2N_5$$

Finally, the different possibilities are listed in the Table 3.2, depending on the sign of different parameters of the system (and we easily verify that they all satisfy (3.20)). We have denoted by $y_1 = (1,0,1,1,3)^\mathsf{T}$, $y_2 = (0,1,1,0,0)^\mathsf{T}$ the two nonnegative minimal $\mathcal{P}$-invariants of the Petri net, and $\pi_1 = (3,0,1,1,2)$, $\pi_2 = (0,1,1,0,0)$ the corresponding row vectors by which the holding times are multiplied.

The case $\boldsymbol{\tau_2}$ small In the case $\pi_1\tau \geqslant \pi_2\tau$, the expression of the throughput is very similar to the expression of the throughput ρ_6 of the Petri net analyzed in Section 3.4. Let us denote $r := \pi_1\tau/\pi_2\tau$ and $\rho^* = y_2^\mathsf{T}N^0/\pi_2\tau$, we have the following values of ρ, as a nondecreasing, continuous, piecewise linear function of $y_1^\mathsf{T}N^0/y_2^\mathsf{T}N^0$.

	$0 \leqslant y_1^\mathsf{T}N^0/y_2^\mathsf{T}N^0 \leqslant 1$	$1 \leqslant y_1^\mathsf{T}N^0/y_2^\mathsf{T}N^0 \leqslant r$	$r \leqslant y_1^\mathsf{T}N^0/y_2^\mathsf{T}N^0$
ρ/ρ^*	0	$\frac{1}{r-1}\left(\frac{y_1^\mathsf{T}N^0}{y_2^\mathsf{T}N^0} - 1\right)$	1

The case τ_2 large On the other hand, if $r < 1$ (τ_2 large), the throughput in the case $r < y_1^\mathsf{T} N^0 / y_2^\mathsf{T} N^0 < 1$ is not fully determined, but can take three different values, for initial markings belonging to the same affine space $m_0' = m_0 + Cx$, $x \in \mathbb{R}$.

Incidentally, this proves that the operator of the dynamics is *not* non-expansive. Indeed, observe that the operator of the dynamics is homogeneous – firing once every transition does not change the place markings, the net is consistent, – so that, if the operator was non-expansive, the asymptotic throughput would be unique for a given set of parameters. See Akian and Gaubert [AG03], and references in Section 1.3.

3.6 Numerical experiments

We first compare the analytical results of Section 3.4 with the asymptotic throughputs of the non-fluid settings.

We then simulate the fluid approximation of the dynamics, and compare the asymptotic throughputs with the affine stationary throughputs computed in the previous section, for the two applications of Section 3.4 and Section 3.5.

3.6.1 Asymptotic behavior of the non-fluid dynamics for the emergency call center PN

We have implemented the δ-discretization of the non-fluid dynamics (Equations (3.3)–(3.5), (3.7) and (3.8)) since this setting is the closest to reality. Recall that, in this case, tokens are routed towards transitions q_2, q_3 and q_4 randomly according to a constant probability distribution. We assume that holding times are given by integer numbers of seconds, so that we take $\delta = 1\,\mathrm{s}$. In this way, we compute the quantities $z_1(t)$, $z_5(t)$ and $z_6(t)$ by induction on $t \in \mathbb{N}$ using the equations describing the dynamics. In the simulations, we choose holding times and probabilities which are representative of the urgency of calls.

Figure 3.2 compares the limits when $t \to +\infty$ of the throughputs $z_1(t)/t$, $z_5(t)/t$, $z_6(t)/t$ of the "real" system, with the throughputs ρ_1, ρ_5 and ρ_6 of the stationary solutions which have been determined in Section 3.4. The latter are simply computed using the analytical formulæ of Table 3.1. We estimate the limits of the throughput $z_i(t)/t$ by evaluating the latter quantity for $t = 10^6\,\mathrm{s}$. As shown in Figure 3.2, these estimations confirm the existence of three phases, as described in the previous section. The convergence of $z_i(t)/t$ towards the throughputs ρ_i is mostly reached in the two extreme phases. In the intermediate phase, the difference between the limit of $z_i(t)/t$ and the throughput ρ_i is more important. This originates from the stochastic nature of the routing, which causes more variations in the realization of the minima in the $z_i(t)$: the throughput of q_6 increases and the throughputs of q_1 and q_5 decrease.

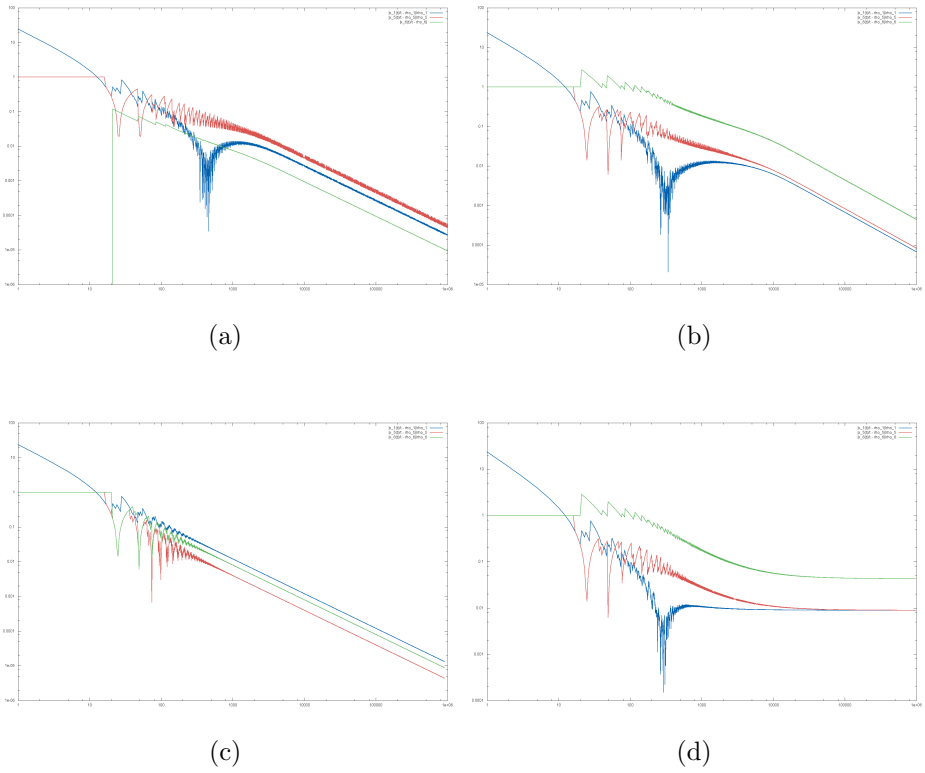
3.6.2 Asymptotic behavior of the fluid dynamics

We also simulate the discrete-time fluid dymamics (using Equations (3.3)–(3.5) and (3.7)–(3.9)) of our two example Petri nets.

The Emergency call center Petri net We implement the dynamics (3.11). All simulations have been computed with exact rationals in $\mathbb{Q}$, using the GMP library [Gt16].

In most cases, we observe that the corresponding asymptotic throughputs converge to the throughputs of the stationary solutions. This is illustrated in Figures 3.3(a), 3.3(b) and 3.3(c), which are obtained using the same set of holding times, and by varying the ratio N_2/N_1 (lower, intermediate and upper phase respectively).

However, there are also cases in which the convergence does not hold. In the experiments we have made, this happens only in the lower phase and in the intermediate phase, that is, when $N_2/N_1 < r_2$, and when $\tau_{\mathrm{tr}} + \tau'_{\mathrm{ext}}$ and τ'_{ur} are in an arithmetical relationship. This is illustrated in Figure 3.3(d), in which we have increased τ'_{ext} by one unit of time in comparison to Figure 3.3(b). Such cases suggest the existence of other kinds of stationary regimes of the dynamics, in which the system oscillates between different phases. An interpretation lies in the fact that, if cycle times are not coprime in the system, phenomena of synchronization may lead to recurrent slow-down of extreme urgent calls by urgent calls in the two lower phases, which could lower the throughput of the system.



(a) (b)

(c) (d)

Figure 3.3 – Comparison of the fluid dymamics with the stationary regimes. Error ratios $|z_i(t)/t - \rho_i|/\rho_i$ are plotted in log-log scale, respectively in blue, red and green when $i = 1, 5, 6$.

In other words, this would correspond to stationary regimes for which the stationary counters would be the sum of an affine function $t \mapsto \rho t$ and a periodic function, taking values in $[0, \rho[$. For such a function, the minima in (3.16) are reached periodically by different terms.

We show in Figure 3.4 the asymptotic throughputs computed in such situation where arithmetical relationships occur between the time parameters. Interestingly enough, the throughputs appear to be (more complex) piecewise linear functions of the variables.

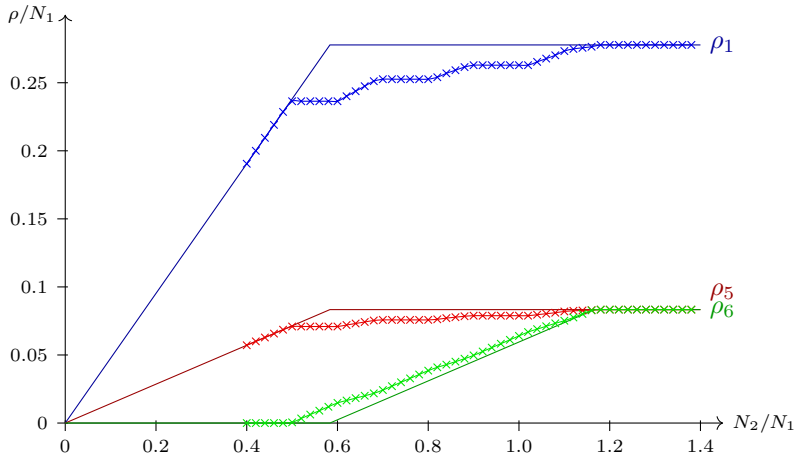
▶ **Remark 3.7** (Analogy with Markov decision processes)**.** It is pointed out in [CGQ95] that, in the homogeneous, free-choice case, the evolution equation of a counter of a Petri net corresponds to the value function of a semi-Markov decision process. It is proven for such problems that history dependent or randomized policies do not provide a better throughput than Markovian deterministic policies, see [Put94, Sections 5.5, 7.1].

The present Petri net is also homogeneous, but encounters priorities, so that the probability matrices of the corresponding semi Markov decision processes (whose value iteration would correspond to the counter equations) would have negative entries. It can be considered as a semi Markov decision process with negative probabilities. For such models, the optimality of the Markovian deterministic policies do not hold. In our case, periodic policies achieve smaller throughputs than policies assigning a fixed decision to each state.

The SR Petri net The fluid dynamics is given by System (3.19). We recall that this Petri net is consistent and conservative, with a unique nonnegative $\mathcal{Q}$-invariant. These properties are often associated with simple dynamical behaviors.

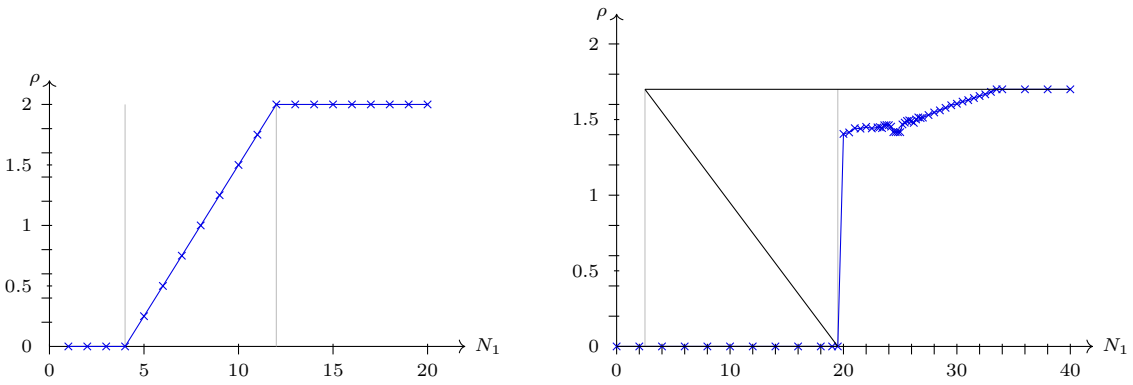
However, even in this situation, we observe that the asymptotic throughputs computed by simulation can differ from the stationary affine throughputs. In our experiments, this happen only in the case when τ_2 is large. See Figure 3.5.

We also note that, in the case τ_2 large, any of the three valuations that are valid for the throughput in the intermediate phase ($\pi_1\tau/\pi_2\tau < y_1^{\mathsf{T}} N^0/y_2^{\mathsf{T}} N^0 < 1$) can be reached by simulations. For example the stationary regime itself is solution of the counter equations.



The parameters are $N_1 = 200$, N_2 varying, $(\tau_{\text{ext}}, \tau_{\text{ur}}, \tau_{\text{adv}}, \tau_{\text{tr}}, \tau'_{\text{ext}}, \tau'_{\text{ur}}) = (4, 3, 3, 1, 6, 7)$, and $(\mu_{\text{ext}}, \mu_{\text{ur}}, \mu_{\text{adv}}) = (0.3, 0.3, 0.4)$.

Figure 3.4 – Petri net asymptotic throughputs in a case with $\tau_{\text{tr}} + \tau'_{\text{ext}} = \tau'_{\text{ur}}$. The plain lines represent the theoretical throughputs, and the marked data represents asymptotic throughputs computed by simulation.



Left: $N^0 = (*, 4, 4, 0, 0)$, $\tau = (1, 2, 2, 1, 1)$. Right: $N^0 = (*, 24, 10, 2.5, 2)$, $\tau = (1, 16, 4, 1, 1)$.

Figure 3.5 – Asymptotic throughputs ($t = 10000$) of the SR Petri net, in blue, and theoretical throughputs, in black, for N_1 varying. The vertical gray lines separate the different phases. Left: τ_2 small. The asymptotic throughputs are identical to the theoretical ones. Right: τ_2 large. The asymptotic throughputs differ from the theoretical ones when N_1 enters the third phase ($y_1^{\mathsf{T}} N^0 > y_2^{\mathsf{T}} N^0$). Observe also the – unexplained – throughput decrease when $N_1 \in [24, 26]$.

3.7 Concluding remarks

The dynamics of Petri nets with free choice and priority routing are characterized by delay equations on the counters of the Petri net. We prove an equivalence between this dynamics and executions of the Petri net. From these delay equations, the affine stationary regimes can then be computed for the fluid approximation of the dynamics, by means of tropical geometry. This allows us to characterize the asymptotic behavior of the Petri net, by identifying different regimes, depending on the parameters of the system. Numerical experiments indicate that these theoretical results are representative of the real dynamics.

However, our experiments also exhibit pathological behaviors of the fluid approximation of our dynamics. While our fluid approximation was designed so as to provide schematic behaviors of the Petri net executions, simulations show that the asymptotic behavior is not always the affine stationary regime that was expected, and that, for some parameters, this can yield different throughputs. These phenomena seem to be related with the discrete time semantics of our Petri net: arithmetical relationships between holding times in different places affect the long-term behavior of the Petri net. This motivates the continuous time modeling we study in the forthcoming chapter.

We also underline here that the discrete dynamics and its fluid counterpart we proposed in this chapter would deserve further analysis. For example, on top of the affine regimes, it remains to characterize periodic regimes, which were shown to modify the asymptotic regimes of the net. More generally, we would like to exhibit conditions of convergence of the fluid dynamics to a stationary regime. Future works should also focus on analyzing the treatment times of the system, on top of the throughputs.

Finally, we point out that the enumeration algorithm that we quickly sketched in Section 3.3 should be further detailed and analyzed. This would allow to implement an analysis tool determining automatically the stationary regimes of a timed Petri net given in input.

Chapter 4

Hybrid dynamics of Petri nets with priorities

This chapter originated from a joint work with Xavier Allamigeon and Stéphane Gaubert, and was first presented to the conference Valuetools [ABG16], and then published in the Performance Evaluation Journal [ABG17]. The current chapter comprises the content of this journal article, and some original content: a whole section comprising the result of well-posedness of our hybrid dynamics (Section 4.3), an additional proposition on the stationary solutions of the dynamics (Proposition 4.13) and the application of the dynamics to the Petri net example of Silva and Recalde (Section 4.5.2).

4.1 Introduction

In this chapter, we analyze an alternative dynamics modeling of Petri nets with priorities, based on systems of differential equations. This continuous-time model aims at avoiding the

discrete-time pathologies observed in the model of Chapter 3, and at obtaining a sound fluid approximation of our emergency call center system. Whereas classical (discrete) Petri nets belong to the class of discrete event dynamic systems, the circulation of tokens in hybrid Petri nets is a continuous phenomenon: tokens are assumed to be fluid, *i.e.*, a transition can fire an infinitesimal quantity of tokens. In this way, the continuous dynamics can be represented by a system of differential equations or differential inclusions in which, however, the right-hand-side may be discontinuous (piecewise continuous).

Contributions The general organization of this chapter has some similarities with the organization of the previous one: we first propose a dynamics for a class of Petri nets comprising our model of an emergency call center (Section 4.2), and then, compute the stationary regimes of this dynamics (Section 4.4). Numerical experiments complete the chapter (Section 4.5).

However, our dynamics here is quite different, as we describe the markings of the Petri net as solutions of a differential (hybrid) system. Our system builds on the differential equations of continuous Petri nets proposed by David and Alla [DA87], and studied in many subsequent works. This differential dynamics models nets in which routing is arbitrated according to a fluid equivalent of the race policy. The main novelty of our approach is that it handles a class of Petri nets in which tokens can be routed according to priority rules (Section 4.2). For this purpose, we found it convenient to attach time to places instead of transitions. This also leads us to distinguish tokens being processed in a place, and tokens available for the firing of a downstream transition.

A consequence of this modeling is that the operator of the dynamics is only piecewise continuous: the flow of a transition encounters discontinuities, as its value depends on whether its different upstream places have a positive amount of tokens available for firing or not. A first main result of this chapter shows that, for the class of Petri nets with free choice and priority routing, the dynamics is well-posed, that is, that there exists a unique *forward Carathéodory solution* (a solution of the system in its integral form, that does not encounter a left-accumulation of switches) (Section 4.3). For this result, we had to express the continuous dynamics in terms of *policies*. A policy is a map associating with each transition one of its upstream places. In this way, the dynamics of the Petri net can be written as an infimum of the dynamics of subnets induced by the different policies. The policies reaching the infimum indicate the places which are bottleneck in the Petri net. We prove that, at any time, one can find a policy which determines the execution of the dynamics on some time interval ahead. On this interval, the dynamics reduces to a linear system.

We characterize the stationary solutions in terms of the policies of the Petri net. This allows us to set up a correspondence between the (ultimately affine) stationary solutions of the discrete dynamics that were described in Chapter 3 and the stationary solutions of the continuous dynamics (Section 4.4). We also relate the continuous stationary solutions to the initial marking of the Petri net or invariants of the Petri net, for some restrictive assumptions.

We finally provide some numerical simulations of the continuous dynamics, for our model of a two-level emergency call center, and for the SR Petri net. On both Petri nets, numerical experiments illustrate the convergence of the trajectory towards the stationary solution. This exhibits an advantage of this hybrid continuous setting in comparison to the discrete one, in which, for certain values of the parameters, the asymptotic throughputs computed by simulations differ from the stationary solutions (Section 4.5.1).

Related work Analysis of differential equations for Petri nets dates back to the work of David and Alla [DA87] in the end of the 80s. They were first seen as a relaxation of the integer firings of a Petri net with discrete firing sequences, the parallel being drawn with the continuous relaxation of integer programming. There is in general no guarantee on the quality of this relaxation to compute quantities such as the asymptotic throughputs of the transitions. However, the comparison between continuous nets and their discrete counterparts has led to fruitful results for some classes of Petri nets. The quality of these approximations for various classes of Petri nets was an important question in the years 1990-2000. It is addressed for example in [MRS06] for the class of live, consistent, conservative, connected Petri nets. A recent introduction to continuous models can be found in [VMJS13], while a more extensive reference is [DA10].

The relationship between this dynamics and stochastic Markovian dynamics was underlined later on, see Recalde and Silva [RS00]. Indeed, the continuous dynamics of Petri nets can be seen as a fluid limit of a queueing system in which each place consists of a queue with an infinite number of servers having i.i.d. exponential service rates. For this fluid limit, the number of tokens in each place is scaled by a factor proportional to N, and the continuous marking is the limit $M_p(N)/N$. The analysis of stochastic systems by the means of a fluid limit leading to differential equations is a classical method in stochastic calculus, see Darling and Norris [DN08] for some applications. However, in the context of Petri nets, differential equations generally come with no result of convergence. Chapter 5 aims at providing such convergence results. We remark that, in this regard, our model can also be seen as the fluid limit of the Queueing Petri nets of Bause [Bau93].

The "continuization" of our dynamics draws inspiration from the original continuous model where time is attached to transitions. In particular, the situation in which the routing of a token at a given place is influenced by the firing times of the output transitions through a race policy has received much attention, see [VMJS13]. Here, we address the situation in which the routing is specified by priority or preselection rules which are independent of the processing rates. To do so, it is convenient to attach times to places, instead of attaching firing rates to transitions. We point out in Remark 4.4 that our model can be reduced to a variant of the standard continuous model [VMJS13] in which we allow immediate transitions and require non-trivial routings to occur only at these transitions. A benefit of our presentation is to allow a more transparent comparison between the continuous model and the discrete time piecewise affine models studied in [CGQ95, GG04b], and in Chapter 3.

In order to evaluate the long-term performance of Petri nets, one has to characterize the stationary or steady states of the Petri nets dynamics. Cohen, Gaubert and Quadrat [CGQ95] introduced an approximation of a discrete Petri net by a fluid, piecewise affine dynamics with finite delays, and showed that the limit throughput does exist for a class of consistent and free choice Petri nets. In the more recent work of Gaujal and Giua [GG04b], the result is extended to larger classes of Petri nets, and the stationary throughputs are computed as the solutions of a linear program. The results obtained using this fluid approximation hardly apply to the discrete model, up to a remarkable exception identified by Bouillard, Gaujal and Mairesse [BGM06] (bounded Petri nets under total allocation). This reference illustrates the many difficulties that arise from the discrete setting (*e.g.*, some firing sequences may lead to a deadlock).

In the hybrid dynamics setting, with time attached to transitions, Recalde and Silva [RS00] showed that the steady states of free choice Petri nets as well as upper bounds of the throughputs in larger classes of Petri nets can be determined by linear programming. However, in general, the asymptotic throughputs are non-monotone with respect to the initial marking or the firing rates of the transitions [MRS06]. An example of oscillations in infinite time around a steady state is also given in [MRTRS08].

Contrarily to the original differential model, the dynamics of our model belongs to the category of differential equations with a discontinuous right-hand-side. In such situation, existence, uniqueness and stability of solutions require a specific analysis. An acknowledged reference is the monograph of Filippov [Fil88]. We also refer to the tutorial of Cortés [Cor08], and to [HcVdSS02] in the specific cases of hybrid systems and piecewise linear systems.

The use of the term "policy" refers to the theory of Markov decision processes, owing to the analogy between the discrete time dynamics and the value function of a semi-Markovian decision process. Note that in the context of hybrid or continuous Petri nets, policies are also known as "configurations", see [MRS06] for an example.

4.2 Hybrid dynamics of Petri nets with time attached to places

4.2.1 General notation

A Petri net consists of a set $\mathcal{P}$ of places, a set $\mathcal{Q}$ of transitions and a set of arcs $\mathcal{E} \subset (\mathcal{P} \times \mathcal{Q}) \cup (\mathcal{Q} \times \mathcal{P})$. Every arc is given a valuation in $\mathbb{N}$. Each place $p \in \mathcal{P}$ is given an initial marking $M_p^0 \in \mathbb{N}$, which represents the number of tokens initially present in the place.

We denote by a^+_{qp} the valuation of the arc from transition q to place p, with the convention that $a^+_{qp} = 0$ if there is no such arc. Similarly, we denote by a^-_{qp} the valuation of the arc from place p to transition q, with the same convention. We set $a_{qp} := a^+_{qp} - a^-_{qp}$. The place-transition incidence matrix C of the Petri net is the $\mathcal{P} \times \mathcal{Q}$ matrix defined by $C := (a_{qp})_{p\in\mathcal{P}, q\in\mathcal{Q}}$. We also denote by C^+ (resp. C^-) the $\mathcal{P} \times \mathcal{Q}$ matrix with entry a^+_{qp} (resp. a^-_{qp}), so that $C = C^+ - C^-$. We limit our attention to *pure* Petri nets, *i.e.*, Petri nets with no self-loop: for every pair (q, p), at least one of a^+_{qp} and a^-_{qp} is zero.

We denote by q^{in} the set of upstream places of transition q and by q^{out} the set of downstream places of transition q. Similarly, we use the notation p^{in} and p^{out} to refer to the sets of input and output transitions of a place p.

4.2.2 General dynamics

We now equip the Petri net with a continuous semantics. Given a transition q, we associate a flow $f_q(t)$ which represents the instantaneous firing rate of transition q at time t. We also associate with each place p a ***marking*** $M_p(t)$, which is a continuous real valued function of the time t. In the case of *discrete* timed Petri nets, one typically requires that every token stays a minimum time in the place, — at this stage, the token may be considered as *under processing* — before becoming available for the firing of output transitions. To capture this property in the continuous setting, we assume that the marking $M_p(t)$ can be decomposed as $M_p(t) = m_p(t) + w_p(t)$, where $m_p(t)$ is the quantity of tokens under processing and $w_p(t)$ is the quantity of tokens waiting to contribute to the firing of an output transition.

We associate with each place p a time constant $\tau_p > 0$. Each token entering in a place is processed with the rate $1/\tau_p$. This leads to the following differential equation:

$$\dot{m}_p(t) = \sum_{q\in p^{\text{in}}} a^+_{qp} f_q(t) - \frac{m_p(t)}{\tau_p}\,. \tag{4.1}$$

The evolution of the number of tokens waiting in place p is described by the relation:

$$\dot{w}_p(t) = \frac{m_p(t)}{\tau_p} - \sum_{q\in p^{\text{out}}} a^-_{qp} f_q(t)\,. \tag{4.2}$$

Moreover, for all transitions q, we require that

$$\min_{p\in q^{\text{in}},\, w_p(t)=0} \Big(\frac{m_p(t)}{\tau_p} - \sum_{q'\in p^{\text{out}}} a^-_{q'p} f_{q'}(t) \Big) = 0\,. \tag{4.3}$$

In particular, this implies that at least one place $p \in q^{\text{in}}$ verifies $w_p(t) = 0$. In this case, (4.3) means that each of the upstream places p that has a zero quantity of waiting tokens ($w_p(t) = 0$) must satisfy $\dot{w}_p(t) \geqslant 0$, and that at least one of these places satisfies $\dot{w}_p(t) = 0$. In other words, there is at least one *bottleneck* upstream place p of q, which has no waiting tokens and whose outgoing flow $\sum_{q'\in p^{\text{out}}} a^-_{q'p} f_{q'}(t)$ coincides with its processing flow $m_p(t)/\tau_p$.

The relation provided in (4.3) can be simplified in the case of free choice conflict and synchronization patterns. In more detail, if q has a unique upstream place p, and this place is free choice (conflict pattern), then (4.3) reduces to:

$$\frac{m_p(t)}{\tau_p} - \sum_{q'\in p^{\text{out}}} a^-_{q'p} f_{q'}(t) = 0\,. \tag{4.4}$$

Now, if q has several upstream places, which are all free choice (synchronization pattern), then (4.3) reads as:

$$f_q(t) = \min_{p\in q^{\text{in}},\, w_p(t)=0} \frac{m_p(t)}{a^-_{qp}\tau_p}\,. \tag{4.5}$$

This equation also holds if $|q^{\text{in}}| = 1$ and if the upstream place of q has a single output transition.

We respectively denote by $m(t)$, $w(t)$ and $f(t)$ the vectors of entries $m_p(t)$, $w_p(t)$ and $f_q(t)$.

Albeit the dynamics that we presented so far is piecewise affine, a trajectory $t \mapsto (m(t), w(t), f(t))$ may be discontinuous. Indeed, in (4.5), the set of the places over which

the minimum is taken may change over time. If at time t, there is a new place $p \in q^{\text{in}}$ such that $w_p(t)$ cancels, and if the quantity $m_p(t)/(a^-_{qp}\tau_p)$ is sufficiently small, then the minimum in (4.5) (and subsequently the flow $f_q(t)$) discontinuously jumps to the latter value. In the case of differential equations with discontinuous right-hand-side, various notions of solutions of the dynamics exist. In the following, any trajectory considered is supposed be a *forward Carathéodory solution* of the dynamics. We report to Section 4.3.3 the definition of this category of solutions of hybrid systems.

Initial conditions of the dynamics are specified by a pair $(m(t_i), w(t_i))$ such that the minimum in (4.3) makes sense, *i.e.*, at least one $w_p(t_i)$ is equal to 0 for each set of places q^{in}. One can easily show that if the set $\{p \in q^{\text{in}} \colon w_p(t) = 0\}$ is nonempty for all transitions $q \in \mathcal{Q}$ at time $t = t_i$, then it remains nonempty at any time $t \geqslant t_i$.

▶ **Remark 4.1.** Suppose that, at time 0, for a transition q, in every upstream place $p \in q^{\text{in}}$, a given marking is instantaneously available for firing, that is, for all $p \in q^{\text{in}}$, $w_p(0) > 0$. This is an ill-posed initial condition in our setting. In fact, transition q would be able to fire a batch of tokens in instantaneous time, that is, $f_q(0) = \min_{p \in q^{\text{in}}}(1/a^-_{qp})w_p(0)\delta_0$, where δ_0 represents a Dirac at time 0, so that, for any downstream place of q, $m_p(0^+) = m_p(0) + \min_{p \in q^{\text{in}}}(1/a^-_{qp})w_p(0)$.

From an ill-posed initial condition, a simple procedure can hence build a well-posed initial condition: it suffices to fire for each transition the batches of tokens available for firing, until one of the upstream w_p reaches 0, and to increase the downstream markings $m_p(0)$ according to these firings.

The dynamics (4.1)–(4.3) may admit different trajectories for a given initial condition. These correspond to different routings of tokens in places with several output transitions. However, each of these trajectories satisfies the conservation law:

$$\dot{m}(t) + \dot{w}(t) = Cf(t)\,. \tag{4.6}$$

As usual, matrix C is the place–transition incidence matrix of the Petri net.

Recall that a ***P-invariant*** of the Petri net refers to a solution $y \neq 0$ of the system $y^{\mathsf{T}}C = 0$. In the discrete setting, a P-invariant corresponds to a weighting of places that is constant for any reachable marking, meaning that the quantity $y^{\mathsf{T}}M$ is preserved under any firing of transition. An analogous statement holds in the hybrid continuous setting:

▶ **Proposition 4.2.** *Given a P-invariant y of the Petri net, the quantity $y^{\mathsf{T}}(m(t) + w(t))_{p\in\mathcal{P}}$ is independent of t.*

In particular, if the entries of y are all positive, then the Petri net is bounded, i.e., each function $t \mapsto M_p(t)$ is bounded.

Proof. The proof consists in multiplying both sides of (4.6) by the row vector y^{T}. It follows that the derivative of $y^{\mathsf{T}}(m(t) + w(t))_{p\in\mathcal{P}}$ is zero. If the entries of y are all positive, then $\big(y^{\mathsf{T}}(m(0) + w(0))_{p\in\mathcal{P}}\big)/y_p$ is an upper bound to the marking of any place p. ◀

The following proposition collects several homogeneity properties of the continuous dynamics:

▶ **Proposition 4.3.** *Let $(m(t), w(t), f(t))$ be a trajectory solution of the dynamics (4.1)–(4.3), with the initial markings $(m_p(0))_{p\in\mathcal{P}}$, and the holding times $(\tau_p)_{p\in\mathcal{P}}$ and let $\alpha \in \mathbb{R}_{>0}$, then:*

(i) $(\alpha m(t), \alpha w(t), \alpha f(t))$ is a trajectory solution of the dynamics, associated with the initial markings $(\alpha m_p(0))_{p\in\mathcal{P}}$.

(ii) $(m(t/\alpha), w(t/\alpha), (1/\alpha)f(t/\alpha))$ is a trajectory solution of the dynamics, associated with the holding times $(\alpha\tau_p)_{p\in\mathcal{P}}$ and the same initial conditions.

(iii) let x be a vector of the kernel of C, and $D = \operatorname{diag}(\tau)$ be the $\mathcal{P} \times \mathcal{P}$ diagonal matrix such that $D_{pp} = \tau_p$, then $(m(t) + \alpha DC^+x, w(t), f(t) + \alpha x)$ is a trajectory solution of the dynamics, associated with the initial markings $(m(0) + \alpha DC^+x)$.

Proof. The first two statements derive easily from the homogeneity properties of Equations (4.1)–(4.3). For the third statement, one can note that adding αx_q to each $f_q(t)$ and adding $\alpha \sum_{q\in p^{\text{in}}} a^+_{qp}x_q$ to each $m_p(t)/\tau_p$ in (4.1)–(4.3) does not change the right hand sides of (4.1) and (4.2), or the expression within the minimum in (4.3). For (4.2) and (4.3), this is due to the fact that $(C^+ - C^-)x = Cx = 0$. ◀

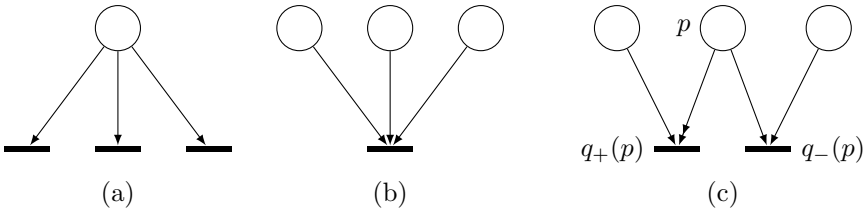


Figure 4.1 – Conflict, synchronization and priority patterns

4.2.3 Hybrid dynamics of Petri nets with free choice and priority routing

Like in the previous chapter, we consider the class of Petri nets in which places are either free choice or subject to priority. Recall that a place $p \in \mathcal{P}$ is said to be ***free choice*** if either all the output transitions $q \in p^{\text{out}}$ satisfy $q^{\text{in}} = \{p\}$ (***conflict***, see Figure 4.1(a)), or $|p^{\text{out}}| = 1$ (***synchronization***, see Figure 4.1(b)). A place is ***subject to priority*** if its tokens are routed to output transitions according to a priority rule. We refer to Figure 4.1(c) for an illustration. For the sake of simplicity, we assume that each place subject to priority has exactly two output transitions, and that any transition has at most one upstream place subject to priority. Given a place p subject to priority, we denote by $q^+(p)$ and $q^-(p)$ its two output transitions, with the convention that $q^+(p)$ has priority over $q^-(p)$. For the sake of readability, we use the notation q^+ and q^- when the place p is clear from context.

The set of transitions such that every upstream place p satisfies $|p^{\text{out}}| = 1$ is referred to as $\mathcal{Q}_{\mathsf{sync}}$ and the set of free choice places that have at least two output transitions is referred to as $\mathcal{P}_{\mathsf{conflict}}$. We denote by $\mathcal{P}_{\mathsf{priority}}$ the set of places subject to priority. The sets $(\mathcal{P}_{\mathsf{conflict}})^{\text{out}}$, $\mathcal{Q}_{\mathsf{sync}}$ and $(\mathcal{P}_{\mathsf{priority}})^{\text{out}}$ form a partition of $\mathcal{Q}$. Figure 4.1 hence summarizes the three possible place/transition patterns that can occur in this class of Petri nets.

We now complete the description of the continuous dynamics (4.1)–(4.3) by additional equations which arise from the specification of routing rules. As we will show in Section 4.3, these routing rules yield a unique solution of the dynamics.

Such rules occur in the following two situations:

Conflict. Given $p \in \mathcal{P}_{\mathsf{conflict}}$, we suppose that tokens are routed according to a stationary distribution specified by weights $\mu_{qp} > 0$ associated with each output transition q. Therefore,

$$\forall p \in \mathcal{P}_{\mathsf{conflict}},\ \forall q \in p^{\text{out}}, \quad a^-_{qp} f_q(t) = \mu_{qp} \frac{m_p(t)}{\tau_p}\,. \tag{4.7}$$

Priority. Let $p \in \mathcal{P}_{\mathsf{priority}}$, and q_+ and q_- be the two output transitions, as illustrated in Figure 4.1(c). In order to specify that the flow is routed in priority to transition q_+, we require that:

$$f_{q_+}(t) = \min_{r \in q_+^{\text{in}},\, w_r(t)=0} \frac{m_r(t)}{a^-_{q_+ r}\tau_r}\,, \tag{4.8}$$

$$f_{q_-}(t) = \begin{cases} \min_{r \in q_-^{\text{in}} \setminus \{p\}, w_r(t)=0} \dfrac{m_r(t)}{a^-_{q_- r}\tau_r} & \text{if } w_p(t) \neq 0\,, \\ \min\Big(\dfrac{m_p(t)}{a^-_{qp}\tau_p} - \dfrac{a^-_{q_+ p}}{a^-_{qp}} f_{q_+}(t)\,, \min_{r \in q_-^{\text{in}} \setminus \{p\}, w_r(t)=0} \dfrac{m_r(t)}{a^-_{q_- r}\tau_r} \Big) & \text{if } w_p(t) = 0\,. \end{cases} \tag{4.9}$$

The expression of $f_{q_-}(t)$ in (4.9), when $w_p = 0$, indicates that only the outgoing flow from p that is not already consumed by the priority transition q_+ is available to q_-. Note that p is uniquely determined by q_- or q_+. See also Figure 4.2.

The first two properties of homogeneity in Proposition 4.3 are still satisfied by the dynamics extended by the routing rules (4.7)–(4.9). However, a trajectory translated by a $\mathcal{Q}$-invariant according to the equations given in the third item of this proposition is a solution of the general dynamics (4.1)–(4.3) but may not respect the priority routing rules (4.8)–(4.9).

▶ **Remark 4.4.** We already mentioned in the introduction that our model differs from the standard continuous Petri net model in which transitions are equipped with firing rates, in the

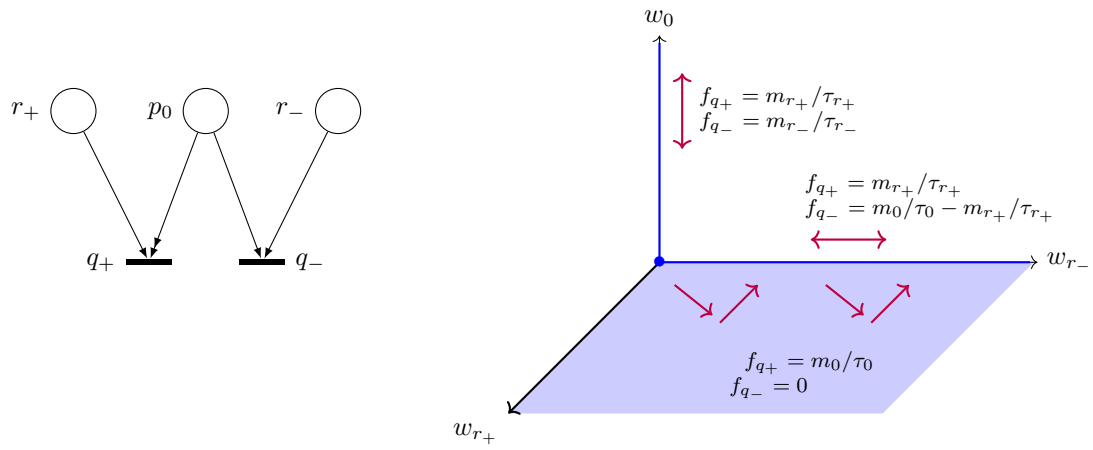


Figure 4.2 – Priority configuration (left), and states reachable by the quantities w_0, w_{r_+} and w_{r_-} corresponding to this configuration (right). The blue color characterizes reachable regions. The purple arrows depict the achievable moves. Here, all the non-zero valuations a^-_{qp} are set to 1.

sense that in the latter model, the flows of the output transitions of a given place are pairwise independent. To overcome this limitation, *immediate transitions* have been introduced [RMS06]. These transitions come with the specification of routing rules, for instance, in the case of conflict pattern. In this way, our model could be reduced to a classical continuous model enriched with immediate transitions. In this reduction, we require timed transitions to have exactly one upstream place and one downstream place, so that all the routing is determined by immediate transitions, which inherit the equations defined in our place-timed dynamics.

Simply put, our model is the continuous analogue of discrete Petri nets equipped with "holding durations", in which tokens are frozen during processing, whereas the usual continuous Petri net model can be seen as the continuous analogues of Petri nets with "enabling durations", in which transitions preempt tokens. We refer to Bowden [Bow00] for a discussion on the different interpretations of time in Petri nets.

4.3 Well-posedness of the dynamics

The general dynamics proposed in (4.1)–(4.3) for continuous Petri nets with holding durations belongs to the category of differential inclusions. Moreover, if at least one transition has several upstream places, the operator of the dynamics is discontinuous, because of the condition $w_p(t) = 0$ in the minimum operator in (4.3), see Section 4.2.2.

In such settings, existence and uniqueness of a solution of the differential system cannot be taken for granted, and a more detailed analysis is required. This is the purpose of this section. We prove that, for the class of Petri nets with free choice and priority routing, there exists a unique *forward Carathéodory solution* of the system on $\mathbb{R}_{\geqslant 0}$.

The technical analysis developed here relies on the notion of *policies* of a Petri net. Fixing a policy allows one to solve the dynamics on a region where it is linear. We shall see in Section 4.4 that policies also arise in the characterization of stationary solutions.

4.3.1 Policies and bottleneck places

Even if our continuous dynamics holds for more general classes of Petri nets, we focus in the remaining of this section on autonomous Petri nets (without external inputs), and require that each transition has at least one upstream place.

We observe that the dynamics of Petri nets with free choice and priority routing (4.1)–(4.2), (4.5) and (4.7)–(4.9) is linear on each region where the arguments of the minimum operators do not change. More precisely, at any time t, for any transition $q \in \mathcal{Q}$, there exists a place $p \in q^{\text{in}}$ such that $w_p(t) = 0$ and, either p is the unique upstream place of q, or p realizes the minimum

in the expression (4.5), (4.8) or (4.9) of $f_q(t)$. Place p is then referred to as a ***bottleneck place*** of transition q at time t.

We define a ***policy*** π as a function from $\mathcal{Q}$ to $\mathcal{P}$, which maps any transition q to one of its upstream places $\pi(q) \in q^{\text{in}}$. A policy is meant to indicate the bottleneck place of each transition q. We denote by S_π the ***selection matrix*** associated with π, that is, the $\mathcal{Q} \times \mathcal{P}$ matrix such that $(S_\pi)_{qp} = 1$ if $p = \pi(q)$, and 0 otherwise. In particular, $(S_\pi)_{qp} = 1$ implies that $a_{qp} < 0$.

Note that, if p realizes the minimum in one of the equations (4.5), (4.8) or (4.9) for some transition, then p also realizes the minimum in (4.3). The converse is not true if places are subject to priority. For p denoting a priority place and q_+ its priority output transition, if p realizes the minimum in (4.3) for transition q_+, then, p does not necessarily realize the minimum in (4.8). In other words, our definition of a bottleneck place is dependent on the routing rules of the net.

We point out that notions comparable to policies are used in [MRS06] (and subsequent works) in the context of continuous Petri nets with time attached to transitions.

Our term "policy" comes from the analogy with Markov decision processes, in which the value iteration formula corresponds to choosing the policy (the set of actions) realizing the minimum in an expression similar to the ones we handle in the following.

Hybrid dynamics of Petri nets in terms of policies Let us first propose a matrix representation of the initial conditions of our dynamics. Let $w_0, m_0 \in \mathbb{R}^{\mathcal{P}}_{\geqslant 0}$:

$$m(0) = m_0\,, \quad w(0) = w_0\,, \quad \text{and there exists } \pi \text{ s.t. } S_\pi w_0 = 0\,. \tag{4.10}$$

The latter condition ensures that, for any transition q, there exists an upstream place p of q such that $(w_0)_p = 0$.

We recall the notation $D := \operatorname{diag}(\tau)$, the $\mathcal{P}\times\mathcal{P}$ diagonal matrix such that $D_{pp} = \tau_p$. The use of policies allows us to introduce the following equivalent representation of the general dynamics of continuous Petri nets (4.1)–(4.3).

$$\dot{m}(t) = C^+ f(t) - D^{-1} m(t) \tag{4.11}$$

$$\dot{w}(t) = D^{-1} m(t) - C^- f(t) \tag{4.12}$$

and

$$\min_{\pi \text{ s.t. } S_\pi w(t)=0} \left(S_\pi D^{-1} m(t) - S_\pi C^- f(t)\right) = 0 \tag{4.13}$$

The minimum in (4.13) is taken over the different policies of the net, and must be understood as an infimum for the partial order over $\mathbb{R}^{\mathcal{Q}}$ induced by $\leqslant$. However, at any time, there is at least one policy attaining the infimum: it suffices to build this policy componentwise, by associating with each transition q a place that attains the minimum for row q of (4.13).

Policies for Petri nets with free choice and priority routing In the following, we propose an expression of our routing rules (4.7)–(4.9) in terms of policies of our Petri net. Naturally, this expression shall have some similarities with (4.13). In this latter equation, the quantity $(S_\pi C^- f(t))_q$ corresponds to the output flow of the upstream place of transition q selected by π. In fact, our routing rules relate an upstream place with its output flow more precisely, depending on the policy π: this leads us to replace matrix $S_\pi C^-$ by a new downstream incidence matrix C^-_π, whose coefficients depend on the policy π. We shall see that this matrix is nonsingular, which means that the flow can be fully characterized by the knowledge of the bottleneck places of each transition.

Now, Equations (4.5) and (4.7)–(4.9) imply, at any time t,

$$\min_{\pi \text{ s.t. } S_\pi w(t)=0} \left(S_\pi D^{-1} m(t) - C^-_\pi f(t)\right) = 0\,, \tag{4.14}$$

where, for any policy π, the matrix C^-_π is such that, for $f \in \mathbb{R}^{\mathcal{Q}}$:

$$\forall q \in \mathcal{Q}\,, (C^-_\pi f)_q = \begin{cases} a^-_{q\pi(q)} f_q & \text{if } \pi(q) \notin (\mathcal{P}_{\mathsf{conflict}} \cup \mathcal{P}_{\mathsf{priority}})\,, \\ \dfrac{a^-_{q\pi(q)}}{\mu_{q\pi(q)}} f_q & \text{if } \pi(q) \in \mathcal{P}_{\mathsf{conflict}}\,, \\ a^-_{q_+\pi(q_+)} f_{q_+} & \text{if } \pi(q) \in \mathcal{P}_{\mathsf{priority}} \text{ and } q = q_+ \text{ is its priority transition}\,, \\ a^-_{q_-\pi(q_-)} f_{q_-} + a^-_{q_+\pi(q_+)} f_{q_+} & \text{if } \pi(q) \in \mathcal{P}_{\mathsf{priority}} \text{ and } q = q_- \text{ is its non priority transition}\,. \end{cases}$$

One can easily check that, by construction of C^-_π, the routing equations (4.5) and (4.7)–(4.9) are equivalent to (4.14).

Similarly to the case of the general dynamics (4.13), the minimum in (4.14) is reached at any time, since it suffices to choose a policy which associates with each transition one of its bottleneck places.

Consequently, at any time t, there exists a policy δ_t such that,

$$S_{\delta_t} w(t) = 0\,, \tag{4.15a}$$

$$S_{\delta_t} D^{-1} m(t) = C^-_{\delta_t} f(t)\,, \tag{4.15b}$$

and for any policy π,

$$S_\pi w(t) = 0 \Rightarrow S_\pi D^{-1} m(t) \geqslant C^-_\pi f(t)\,. \tag{4.15c}$$

A policy δ which satisfies (4.15) on some time interval for a trajectory $(m(t), w(t), f(t))$ solution of the dynamics is said to be a ***valid policy*** on this time interval for this trajectory.

Matrix C^-_π is nonsingular, which implies that the flow at a given time is uniquely determined by the marking of the different bottleneck places of the Petri net. Indeed, all diagonal entries of C^-_π are positive. Moreover, if we order the transitions of the Petri net such that the transitions which have lower priority for some upstream place subject to priority are the largest for this order [1], then the matrix becomes lower triangular.

It will be useful to provide a description of C^-_π and of its inverse by blocks. We define by $\mathcal{Q}_-$ the set of transitions being lower priority transitions of some place subject to priority, and $\mathcal{Q}_{0+} := \mathcal{Q} \setminus \mathcal{Q}_-$ the set of the remaining transitions. Note that, in our class of Petri nets, transitions being assigned a high priority necessarily belong to $\mathcal{Q}_{0+}$. The order that we introduced just before is such that the transitions of $\mathcal{Q}_{0+}$ are before the transitions of $\mathcal{Q}_-$. Matrix C^-_π admits the following block decomposition, according to the partition $\mathcal{Q} = \mathcal{Q}_{0+} \cup \mathcal{Q}_-$:

$$C^-_\pi = \begin{pmatrix} A^\pi_+ & 0 \\ A^\pi_{-+} & A^\pi_- \end{pmatrix}\,,$$

where A^π_+ and A^π_- are diagonal matrices whose diagonal entries are all positive. Both matrices are hence nonsingular, and, for $x, y \in \mathbb{R}^{\mathcal{Q}}$, with notation $x = (x_{0+}, x_-)$ and $y = (y_{0+}, y_-)$,

$$C^-_\pi x = y \quad \Leftrightarrow \quad \begin{cases} x_{0+} = (A^\pi_+)^{-1} y_{0+}\,, \\ x_- = (A^\pi_-)^{-1} \left(y_- - A^\pi_{-+} x_{0+}\right) \\ \qquad = (A^\pi_-)^{-1} \left(y_- - A^\pi_{-+} (A^\pi_+)^{-1} y_{0+}\right)\,. \end{cases}$$

As a consequence, the inverse of C^-_π is given by

$$(C^-_\pi)^{-1} = \begin{pmatrix} (A^\pi_+)^{-1} & 0 \\ -(A^\pi_-)^{-1} A^\pi_{-+} (A^\pi_+)^{-1} & (A^\pi_-)^{-1} \end{pmatrix}\,.$$

Computing the flow at a given time Given an initial condition m_0, w_0, computing the flow at time 0 is straightforward using Equations (4.5) and (4.7)–(4.9). Remark that this computation almost corresponds to computing a minimum, transition by transition. However, one must ensure that the flow in transitions of type q_+ is computed before the flow of transitions of type q_-.

1. Note that this ordering is valid, because in our class of Petri net, each transition has at most one usptream place subject to priority.

Now, starting from Relation (4.14), if all the matrices C_π^- were diagonal with positive diagonal entries, then $f(t)$ would have a simple expression as an infimum:

$$f(t) = \min_{\pi \text{ s.t. } S_\pi w(t)=0} \left((C_\pi^-)^{-1} S_\pi D^{-1} m(t)\right) ,$$

and a valid policy at time t would be a policy attaining the minimum. This actually holds when the Petri net is free choice. Moreover, we remark that, if two policies attain the minimum in the latter equation, then any mixing of these policies (choosing arbitrarily the bottleneck place of one or the other policy) also attain the minimum.

When there is a place subject to priority, this simple expression is no more valid. However, we invite the reader to think of the resolution of (4.14) as a componentwise computation of an infimum, in which one considers elements of $\mathcal{Q}_{0+}$ before elements of $\mathcal{Q}_-$. A formal meaning shall be given in Section 4.7 to this informal discussion. We will be interested in computing, not only $f(t)$, but also its successive derivatives, using this componentwise computation with an appropriate ordering of the transitions.

Indeed, our purpose is to determine if, starting from a given time t_0, there exists one or several policies that remain valid on some time interval $[t_0, t_0 + \varepsilon]$, with $\varepsilon{>}0$. If there exists a unique policy attaining the minimum in (4.14) at time t_0, then, by continuity of the different terms of this minimum, this policy is valid on a time interval. However, if there are several policies attaining the minimum, then, determining which of these policies (if any) is valid just after time t_0 requires comparing the derivatives of the different candidate flows, and, if there are still several candidate policies, again, the successive derivatives.

This is the subject of the next section.

4.3.2 Working out a valid policy in a forward time interval

We consider the system (4.11), (4.12), (4.14), together with initial conditions (4.10). Let us assume that π is a valid policy for the dynamics during a time interval $[0, T]$, that is, for any transition q, place $\pi(q)$ is bottleneck for q on $[0, T[$. Then, (4.15a) and (4.15b) hold for π on $[0, T[$.

If we multiply (4.11) by S_π, and replace the term $S_\pi D^{-1} m(t)$ by its expression given in (4.15b), we obtain

$$S_\pi \dot{m}(t) = (S_\pi C^+ - C_\pi^-) f(t) . \tag{4.16}$$

Let D_π be the $\mathcal{Q}{\times}\mathcal{Q}$ diagonal matrix such that $(D_\pi)_{qq} = \tau_{\pi(q)}$, that is, $D_\pi := S_\pi D {S_\pi}^{\mathsf{T}}$. Equation (4.15b) then writes

$$S_\pi m(t) = D_\pi C_\pi^- f(t) . \tag{4.17}$$

By differentiating this expression, we can replace the left-hand-side of (4.16), and get

$$\dot{f}(t) = (C_\pi^-)^{-1} D_\pi^{-1} \left(S_\pi C^+ - C_\pi^-\right) f(t) . \tag{4.18}$$

In addition, by (4.11) and (4.15b), we also have

$$\dot{m}(t) = (C^+ (C_\pi^-)^{-1} S_\pi - I) D^{-1} m(t) .$$

Finally, if π is a valid policy during a time interval $[0, T]$, then, for $t \in [0, T]$,

$$\begin{aligned} m(t) &= e^{A_\pi t} m_0 \\ f(t) &= e^{B_\pi t} (C_\pi^-)^{-1} S_\pi D^{-1} m_0 , \end{aligned}$$

with $A_\pi := (C^+ (C_\pi^-)^{-1} S_\pi - I) D^{-1}$ and $B_\pi := (C_\pi^-)^{-1} D_\pi^{-1} (S_\pi C^+ - C_\pi^-)$.

Now, for any policy π, we define the candidate flow f_π starting from $t = 0$ associated with π as the flow given by this expression: $f_\pi : t \mapsto e^{B_\pi t} (C_\pi^-)^{-1} S_\pi D^{-1} m_0$, and the candidate marking $m_\pi : t \mapsto e^{A_\pi t} m_0$.

The following proposition is crucial for establishing the well-posedness of the differential system. It is a characterization of smooth continuation of a piecewise linear system in one of the different linear modes by a lexicographic relation on the successive derivatives of the operator

of the dynamics. A similar characterization was established in [IvdS00] in a more general case. For the sake of readability, the lexicographical ordering relation $((x_1)_q, (x_2)_q, \dots) \overset{\text{lex.}}{\leqslant} ((y_1)_q, (y_2)_q, \dots)$, where $x_1, x_2, \dots$ and $y_1, y_2, \dots$ are vectors of $\mathbb{R}^{\mathcal{Q}}$, is abridged in the following way: $(x_1, x_2, \dots)_q \overset{\text{lex.}}{\leqslant} (y_1, y_2, \dots)_q$.

▶ **Proposition 4.5.** *Let δ be a policy, and m_δ and f_δ its associated candidate flow and candidate marking. The two following properties are equivalent:*

- *There exists a trajectory $(m(t), w(t), f(t))$ solution of (4.10), (4.11), (4.12), and (4.14), and $\varepsilon > 0$ such that δ is a valid policy on $[0, \varepsilon]$,*
- *The equality $S_\delta w_0 = 0$ holds, and, for any policy π,*

$$\forall q \in \mathcal{Q}, \quad (S_\delta w_0, C_\pi^- f_\delta^{(1)}(0), C_\pi^- f_\delta^{(2)}(0), \dots)_q \overset{\text{lex.}}{\leqslant} (S_\pi w_0, S_\pi D^{-1} m_\delta^{(1)}(0), S_\pi D^{-1} m_\delta^{(2)}(0), \dots)_q . \quad (4.19)$$

Observe that, if δ is such that the second property holds, then (4.19) implies in particular that δ realizes the minimum in (4.14) at time 0, since $m_\delta^{(1)}(0) = m_0$ and $S_\pi D^{-1} m_0 = C_\pi^- f_\pi^{(1)}(0)$.

Proof. $\Rightarrow$ Let $\varepsilon{>}0$. We suppose that δ is a valid policy on $[0, \varepsilon]$ for $m(t), w(t), f(t)$. This implies that $S_\delta w(t) = 0$, $f(t) = f_\delta(t)$ and $m(t) = m_\delta(t)$ on this time interval. Let π be another policy, and let q be a transition. If $(S_\pi w_0)_q > 0$, then Relation (4.19) holds for π and q. Now, suppose $(S_\pi w_0)_q = 0$. Suppose also, for the contradiction, that there exists an index $j{\geqslant}0$ such that, for $i{<}j$, $(C_\pi^- f_\delta^{(i)}(0))_q = (S_\pi D^{-1} m_\delta^{(i)}(0))_q$, and $(C_\pi^- f_\delta^{(j)}(0))_q > (S_\pi D^{-1} m_\delta^{(j)}(0))_q$. Then, on an interval $]0, \eta[$, with $\eta > 0$, $(C_\pi^- f_\delta(0) - S_\pi D^{-1} m_\delta(0))_q$ is positive. Note that this implies that $\pi(q) \neq \delta(q)$, because of (4.15c). In particular, $\pi(q)$ cannot be in $\mathcal{P}_{\mathsf{conflict}}$. We show that $w_{\pi(q)}$ would become negative. Remember that $w_{\pi(q)}(0) = 0$. In the cases when $q \in \mathcal{Q}_{\mathsf{sync}}$, or when q has one upstream place subject to priority, but this place is not $\pi(q)$, or when $\pi(q)$ is subject to priority, and q is its non priority transition, then this simply comes from the fact that $(C_\pi^- f_\delta(0))_q = a^-_{q\pi(q)}(f_\delta)_q(0) = (S_\pi C^- f_\delta(0))_q$. Consequently, by (4.12) multiplied by S_π, the first nonzero derivative of w is negative, and so $w_{\pi(q)}$ becomes negative, which is a contradiction.

It remains to handle the case when q is the priority transition of place $\pi(q)$. We note q_- the non priority transition of $\pi(q)$. In this case, $(C_\pi^- f_\delta(0))_q = a^-_{q\pi(q)}(f_\delta)_q(0) \leqslant a^-_{q\pi(q)}(f_\delta)_q(0) + a^-_{q_-\pi(q)}(f_\delta)_{q_-}(0) \leqslant (S_\pi C^- f_\delta(0))_q$, so that, again, by (4.12) multiplied by S_π, the first nonzero derivative of w is negative, and so $w_{\pi(q)}$ becomes negative. Contradiction.

$\Leftarrow$ We prove that, at time 0 and for some time afterwards, Equations (4.11), (4.12), (4.15) hold for the candidate marking and flow $m_\delta(t)$, $f_\delta(t)$, and for $w(t)$ solution of (4.12) with $m = m_\delta$, $f = f_\delta$ and initial condition w_0. Note that Relations (4.12), (4.11) and (4.15b) automatically hold, by assumption, and by construction of f_δ and m_δ. The fact that $S_\delta w(t)$ remains 0 is a consequence of (4.13), in which δ realizes the minimum: when multiplying (4.12) by S_δ, the right-hand-side cancels, and we get $S_\delta \dot{w}(t) = 0$. This proves (4.15a).

It remains to prove that (4.15c) holds on some time interval.

Let π be a policy, and q be a transition. If $(S_\pi w_0)_q > 0$, then, by continuity of w, there exists a time interval $[0, \varepsilon_q]$ such that the inequality remains true. If $(S_\pi w_0)_q = 0$, then, (4.19) implies that the first nonzero derivative of $(S_\pi D^{-1} m_\delta(t) - C_\pi^- f_\delta(t))_q$ is positive, or that all derivatives are null. In the first situation, just after time 0, $(S_\pi D^{-1} m_\delta(t) - C_\pi^- f_\delta(t))_q$ becomes positive, and remains positive for some time interval $]0, \varepsilon_q]$, so that inequality in (4.15c) holds on this time interval. Otherwise, all the derivatives are zero. The difference $(S_\pi D^{-1} m_\delta(t) - C_\pi^- f_\delta(t))_q$ being a linear transformation of two functions solutions of linear differential systems, the fact that its derivatives are all zero at some time implies that the function itself is zero, and (4.15c) also holds for row q, for any ε_q. Finally, by taking $\varepsilon = \min_{q \in \mathcal{Q}} \varepsilon_q$, the result is proven. ◀

Proposition 4.5 provides a characterization of a valid policy in terms of Relation (4.19), expressed in terms of the sequence of derivatives $(f_\delta, f_\delta^{(2)}, \dots)$. One would like to use this relation to prove the existence and uniqueness (in some sense) of a valid policy on some interval $[0, \varepsilon]$. However, Relation (4.19) has several drawbacks.

Observe first that the formula of (4.19) is not symmetric in policies δ, π, so that one has to prove that the relation is however antisymmetric: if Relation (4.19) holds for the pair of policies (δ, π) and for the pair of policies (π, δ), then, it is an equality.

Furthermore, some policies are not comparable for Relation (4.19), because different coordinates q may lead to opposite comparisons. However, if a pair of policies is not comparable, it will be shown that a third policy is comparable and smaller than both policies for this relation.

The proposition below follows from such manipulations of Relation (4.19):

▶ **Proposition 4.6.** *There exists a policy δ such that $S_\delta w_0 = 0$ and, for any policy π, Relation (4.19) holds. Moreover, if there exists two such policies δ_1 and δ_2, then, $f_{\delta_1} = f_{\delta_2}$, and Relation (4.19) with $\delta = \delta_1$ and $\pi = \delta_2$ is an equality, and similarly, it is an equality for $\delta = \delta_2$ and $\pi = \delta_1$.*

The proof of Proposition 4.6 is demanding and requires some technical developments. We report it to the end of this chapter, see Section 4.7. We remark here that it can be largely simplified if the Petri net is free-choice (without places subject to priority). In this situation, the matrices C_π^- are all diagonal, so that, for a vector $x \in \mathbb{R}^{\mathcal{P}}_{\geqslant 0}$, comparing $C_\pi^-(C_\sigma^-)^{-1}S_\sigma x$ and $S_\pi x$ is equivalent to comparing $(C_\sigma^-)^{-1}S_\sigma x$ and $(C_\pi^-)^{-1}S_\pi x$. The policies that realize the minimum of the set $\{(C_\pi^-)^{-1}S_\pi D^{-1}m_0\}$ attain this minimum componentwise (and can be computed componentwise), and if there are several, the same componentwise selection can be achieved successively for the sequence of derivatives.

4.3.3 Smooth continuation and well-posedness

A direct corollary of Proposition 4.6 is the following:

▶ **Corollary 4.7** (Smooth continuation)**.** *At any time t, there exists a unique smooth continuation of the dynamics, i.e., there exists a trajectory solution of the dynamics, $\varepsilon > 0$, and a policy δ such that δ is valid on $[t, t+\varepsilon]$, and any other policy π valid on an interval of the form $[t, t+\varepsilon']$ yields the same trajectory.*

Let δ be a valid policy of a trajectory solution of the dynamics during some time interval. If there exists a time at which the policy is no more valid, then, the right bound of the maximal time interval on which δ is valid is called a ***switching time***. At this switching time, the trajectory is not smooth.

By the latter corollary, one can construct a solution of the dynamics starting from time 0, such that there exists a sequence of switching times $t_1, t_2, \dots$ and a sequence of policies $\delta(1), \delta(2), \dots$ with, for any i, $\delta(i)$ is a valid policy on $[t_{i-1}, t_i[$, $\delta(i) \neq \delta(i-1)$, and, for $t \geqslant 0$ and for n such that $t_n \leqslant t < t_{n+1}$,

$$m(t) = e^{A_{\delta(n)}(t-t_n)} \dots e^{A_{\delta(2)}(t_2-t_1)} e^{A_{\delta(1)}t_1} m_0 \,, \tag{4.20}$$

is a solution of the system.

Such a solution is bounded on any finite time interval:

▶ **Corollary 4.8.** *A solution of the kind (4.20) does not goes to infinity in finite time.*

Proof. Let $a > 0$ be a common upper bound of all the matrices A_π, for an operator norm. Then, for any π, $\|e^{A_\pi t}\| \leqslant e^{at}$. Consequently, by (4.20), $\|m(t)\| \leqslant e^{at}\|m(0)\|$. ◀

A ***forward Carathéodory solution***[2] of a differential system $\dot{x}(t) = F(x(t))$, $x(0) = x_0$, with F a discontinuous vector field, is an absolutely continuous function $x(t)$ on an interval $[0, T[$ satisfying $x(t) = \int_0^t F(x(s))\,\mathrm{d}s$ for $t \in [0, T[$, and, such that, moreover, there is no left-accumulation point of event times on $[0, T[$, that is, no left-accumulation point of switching times of $F(x(t))$.

Note that the notion of forward Carathéodory solution of a differential system with discontinuous right-hand-side excludes sliding modes (which appear in Filippov solutions) and points of left-accumulation (which can exist in classical Carathéodory solutions). This seems the right notion of a solution of a differential system involving a timed system, where we expect functions describing the behavior of the system to be càdlàg. An example of a piecewise linear system having an infinity of Carathédory solutions and a point of left-accumulation can be found in [Fil88, p.116], see also [HcVdSS02].

2. See definition in [IvdS00], where the notion appears under the vocable *extended Carathéodory solution*. Our terminology is borrowed from [Tc11], see comments therein.

▶ **Theorem 4.9.** *The dynamical system (4.11), (4.12), (4.15), together with initial conditions (4.10), admits a unique forward Carathéodory solution on* $\mathbb{R}_{\geqslant 0}$.

Proof. This is a direct consequence of Corollary 4.7, by Lemma 2.1 of [IvdS00], which establishes the equivalence between the smooth continuation property and forward Carathéodory solutions.

Note that Lemma 2.1 of [IvdS00] is stated in the bimodal case, for which the vector space is divided in two half-spaces on which the operator of the dynamics is continuous. This lemma has a natural extension in the multi-modal case (number of maximal regions of continuity larger than 2). This is the version used in Section 6 of [IvdS00] (see, in particular, proof of Theorem 6.1). ◀

▶ **Remark 4.10** (Zeno behavior)**.** A forward Carathéodory solution forbids points of left-accumulation. However, we do not exclude that right-accumulation points exist. In the case when there is such a right-accumulation, *i.e.*, a sequence of switching times $t_1 < t_2 < \cdots \to t_\infty$, with finite t_∞, the place markings m, w still have a limit at t_∞, and from this time on, the execution continues.

This, however, raises tractability issues when implementing the Petri net, so that one would like to determine in which conditions a Petri net execution can encounter such right-accumulation points. This is an open question for executions of continuous Petri nets with holding durations [3].

A simple case is when there is only one transition having two upstream places (all the other transitions having a unique upstream place). In this case, the system is a bimodal linear system, and it can be proven that its Carathéodory solution is also a Filippov solution, and that Zeno behavior cannot happen. See [Tc11].

4.4 Stationary solutions

In this section, we prove that the stationary solutions of the continuous and discrete dynamics of a timed Petri net with free-choice and priority routing are the same. To do so, we first recall in Section 4.4.1 the formulation of the discrete dynamics and the associated stationary solutions given in [ABG15].

4.4.1 Stationary solutions of the discrete dynamics

The discrete dynamics of Petri nets with free choice and priority is expressed in terms of ***counter variables*** associated with transitions and places. Given a transition q, the counter variable $z_q : \mathbb{R}_{\geqslant 0} \to \mathbb{N}$ denotes the number of firings of q that occurred up to time t included. Similarly, the counter variable of place p is a function $x_p : \mathbb{R}_{\geqslant 0} \to \mathbb{N}$ which represents the number of tokens that have visited place p up to time t included (taking into account the initial marking). On top of being non-decreasing, the counter variables are ***càdlàg*** functions, which means that they are right continuous and have a left limit at any time.

In this setting, the parameter τ_p associated with the place p represents a minimal holding time. It is shown in [ABG15] that, if tokens are supposed to be fired as early as possible, the counter variables satisfy the following equations (we generalize the equations to the case with valuations):

$$\forall p \in \mathcal{P}\,, \quad x_p(t) = M_p^0 + \sum_{q \in p^{\text{in}}} a_{qp}^+ z_q(t)\,, \tag{4.21a}$$

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \quad \sum_{q \in p^{\text{out}}} a_{qp}^- z_q(t) = x_p(t - \tau_p)\,, \tag{4.21b}$$

$$\forall q \in \mathcal{Q}_{\mathsf{sync}}\,, \quad z_q(t) = \min_{p \in q^{\text{in}}} x_p(t - \tau_p)/a_{qp}^-\,, \tag{4.21c}$$

3. Note that Zeno behaviors appear in very simple physical systems ; a well-known example is that of the bouncing times of a bouncing ball.

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,,$$

$$z_{q_+}(t) = \min\Bigg(\Big(\frac{1}{a^-_{q_+p}} x_p(t-\tau_p) - \frac{a^-_{q_-p}}{a^-_{q_+p}} \lim_{s\uparrow t} z_{q_-}(s)\Big), \min_{r\in q_+^{\mathrm{in}}, r\neq p} \frac{1}{a^-_{q_+r}} x_r(t-\tau_r)\Bigg)\,, \tag{4.21d}$$

$$z_{q_-}(t) = \min\Bigg(\Big(\frac{1}{a^-_{q_-p}} x_p(t-\tau_p) - \frac{a^-_{q_+p}}{a^-_{q_-p}} z_{q_+}(t)\Big), \min_{r\in q_-^{\mathrm{in}}, r\neq p} \frac{1}{a^-_{q_-r}} x_r(t-\tau_r)\Bigg)\,, \tag{4.21e}$$

where q_+ (q_-) is the priority (non priority) output transition of $p \in \mathcal{P}_{\mathsf{priority}}$.

Note that if all the holding times τ_p are integer multiples of a fixed time δ, the left limit $\lim_{s\uparrow t} z_{q_-}(s)$ in (4.21d) can be replaced by $z_{q_-}(t-\delta)$. This is helpful in particular to simulate these equations.

In the setting of [ABG15], all conflicts are solved by a stationary distribution routing. The equivalent of the routing rule introduced to solve conflicts in the continuous setting is obtained here by allowing the tokens to be shared in fractions, so that the counter functions take real values. This corresponds to a ***fluid approximation*** of the discrete dynamics. In this setting, for each $p \in \mathcal{P}_{\mathsf{conflict}}$ and $q \in p^{\mathrm{out}}$, we fix $\mu_{qp} > 0$, giving the proportion of the tokens routed from p to q. We have:

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \forall q \in p^{\mathrm{out}}\,, \quad z_q(t) = \frac{\mu_{qp}}{a^-_{qp}} x_p(t-\tau_p)\ . \tag{4.22}$$

The stationary solutions of the discrete dynamics are defined as functions x_p and z_q satisfying the relations (4.21)–(4.22) and which ultimately behave as affine functions, *i.e.*, $x_p(t) = u_p + t\rho_p$ and $z_q(t) = u_q + t\rho_q$ for all t large enough. In this case, ρ_p (resp. ρ_q) represents the asymptotic throughput of place p (resp. transition q). We have shown in [ABG15, Theorem 3] that these stationary solutions are precisely given by the following system (we generalize the equations to the case with valuations):

$$\forall p \in \mathcal{P}\,, \qquad \rho_p = \sum_{q\in p^{\mathrm{in}}} a^+_{qp}\rho_q\,, \tag{4.23a}$$

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \forall q \in p^{\mathrm{out}}\,, \qquad \rho_q = \mu_{qp}\rho_p / a^-_{qp}\,, \tag{4.23b}$$

$$\forall q \in \mathcal{Q}_{\mathsf{sync}}\,, \qquad \rho_q = \min_{p\in q^{\mathrm{in}}} \rho_p / a^-_{qp}\,, \tag{4.23c}$$

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,, \qquad \rho_{q_+} = \min_{r\in q_+^{\mathrm{in}}} \rho_r / a^-_{q_+r}\,, \tag{4.23d}$$

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,, \quad \rho_{q_-} = \min\Big((\rho_p - a^-_{q_+p}\rho_{q_+})/a^-_{q_-p}, \min_{r\in q_-^{\mathrm{in}}\setminus\{p\}} \rho_r / a^-_{q_-r}\Big)\,, \tag{4.23e}$$

$$\forall p \in \mathcal{P}\,, \quad u_p = M^0_p + \sum_{q\in p^{\mathrm{in}}} a^+_{qp} u_q\,, \tag{4.24a}$$

$$\forall p \in \mathcal{P}_{\mathsf{conflict}}\,, \forall q \in p^{\mathrm{out}},\ u_q = (\mu_{qp}/a^-_{qp})(u_p - \rho_p\tau_p)\,, \tag{4.24b}$$

$$\forall q \in \mathcal{Q}_{\mathsf{sync}}\,, \quad u_q = \min_{p\in q^{\mathrm{in}}, \rho_q = \rho_p} (u_p - \rho_p\tau_p)/a^-_{qp}\,, \tag{4.24c}$$

$$\forall p \in \mathcal{P}_{\mathsf{priority}}\,,$$

$$u_{q_+} = \begin{cases} \min\Big((u_p - \rho_p\tau_p - a^-_{q_-p}u_{q_-})/a^-_{q_+p}, & \text{if } \rho_{q_-} = 0\,, \\ \qquad \min_{r\in q_+^{\mathrm{in}}\setminus\{p\},\, \rho_{q_+}=\rho_r} (u_r - \rho_r\tau_r)/a^-_{q_+r}\Big) & \\ \min_{r\in q_+^{\mathrm{in}}\setminus\{p\},\, \rho_{q_+}=\rho_r} (u_r - \rho_r\tau_r)/a^-_{q_+r} & \text{otherwise,} \end{cases} \tag{4.24d}$$

$$u_{q_-} = \begin{cases} \min\Big((u_p - \rho_p\tau_p - a^-_{q_+p}u_{q_+})/a^-_{q_-p}, & \text{if } \rho_{q_-} + \rho_{q_+} = \rho_p\,, \\ \qquad \min_{r\in q_-^{\mathrm{in}}\setminus\{p\},\, \rho_{q_-}=\rho_r} (u_r - \rho_r\tau_r)/a^-_{q_-r}\Big) & \\ \min_{r\in q_-^{\mathrm{in}}\setminus\{p\},\, \rho_{q_-}=\rho_r} (u_r - \rho_r\tau_r)/a^-_{q_-r} & \text{otherwise.} \end{cases} \tag{4.24e}$$

The above equations are expressed in a more compact form in [ABG15], using a semiring of germs of affine functions, which encodes lexicographic minimization operations.

4.4.2 Stationary solutions of the continuous time dynamics

In the continuous setting, we define a ***stationary solution*** as a solution (m, w, f) of the continuous dynamics such that for any place, m_p is constant and w_p is affine ($\dot{w}_p$ is constant). The following theorem provides a characterization of the stationary solutions.

▶ **Theorem 4.11.** *A triple (m, w, f) of vectors of resp. $|\mathcal{P}|$, $|\mathcal{P}|$ and $|\mathcal{Q}|$ functions from $\mathbb{R}_{\geqslant 0}$ to $\mathbb{R}_{\geqslant 0}$, with all the m_p constant and all the w_p affine, is a stationary solution of the continuous dynamics if and only if the following conditions hold:*

$$D^{-1}m = C^+ f\,, \tag{4.25a}$$
$$\dot{w} = D^{-1}m - C^- f\,, \tag{4.25b}$$
$$Cf \geqslant 0\,, \tag{4.25c}$$

and there exists a policy δ, such that

$$\forall t\,, \quad S_\delta w(t) = 0\,, \tag{4.25d}$$
$$\left(S_\delta C^+ - C^-_\delta\right) f = 0\,. \tag{4.25e}$$

Note that the existence of an $f \gneq 0$ that satisfies (4.25c) provides a simple algebraic necessary condition to the existence of a stationary flow in a Petri net. This corresponds to the net being *partially repetitive* (see [Mur89] for a definition).

Proof. Equations (4.25a) and (4.25b) are derived from (4.11) and (4.12), with $\dot{m} = 0$ for a stationary solution.

In a stationary solution, for any place p, $\dot{w}_p$ is constant, so that one cannot have $\dot{w}_p < 0$, otherwise this would yield $\lim_{t\to\infty} w_p(t) = -\infty$. Therefore, by (4.25b), $D^{-1}m \geqslant C^- f$, and by (4.25a), we can replace $D^{-1}m$ by $C^+ f$, and get (4.25c).

As the $\dot{w}$ are constant, if, for some place p and at some time $t_0 > 0$, $w_p(t_0) = 0$, then $\dot{w}_p = 0$, (otherwise it would contradict $w_p(t) \geqslant 0$ for $0 \leqslant t < t_0$ or for $t > t_0$). Hence, the set of places p such that $w_p(t) = 0$ is independent of time for $t > 0$.

Moreover, the m_p are constant, so that, if a policy π is valid at some time, then it is valid at any time. This means that, if (m, w, f) is a solution of the continuous dynamics, then there exists a policy δ such that:

$$\forall t\,, \quad S_\delta w(t) = 0\,,$$
$$C^-_\delta f = S_\delta D^{-1} m(t)\,.$$

Now, by (4.25a) again, we can replace $D^{-1}m$ by $C^+ f$ in the above equation, and we get Equations (4.25d) and (4.25e).

Conversely, suppose that a triple of functions (m, w, f) satisfies the conditions of the theorem, with policy δ. We prove that the the relations (4.11), (4.12) and (4.15), describing the dynamics are satisfied. First, (4.11) and (4.12) are derived from (4.25a) and (4.25b), with $\dot{m}_p = 0$. Equations (4.25d) is (4.15a). We also note that, in Equations (4.25c) and (4.25e), replacing the term $C^+ f$ by m/τ (by (4.25a)) leads to the following equations:

$$C^- f \leqslant D^{-1} m\,, \tag{4.26}$$
$$f = (C^-_\delta)^{-1} S_\delta D^{-1} m\,. \tag{4.27}$$

Equation (4.27) implies (4.15b).

It remains to prove (4.15c). We prove this inequality row by row. Let q be a transition. We distinguish the following cases:

- if $q \in \mathcal{Q}_{\mathsf{sync}}$, then $(C^- f)_{\pi(q)} = (C^-_\pi f)_q$ for any π (for any choice of an upstream place of q) so that (4.15c) follows from (4.26).
- if q has a unique upstream place p, with $p \in \mathcal{P}_{\mathsf{conflict}}$, then for any π, $\pi(q) = p_\delta(q)$ so that (4.15c) follows from (4.27).

- assume now that q_+ is the priority transition of a place p subject to priority. Then, by (4.26), $m_p/\tau_p \geqslant a^-_{q_+p} f_{q_+} + a^-_{q_-p} f_{q_-} \geqslant a^-_{q_+p} f_{q_+}$ and for $r \in q_+^{\text{in}} \setminus \{p\}$, $m_r/\tau_r \geqslant a^-_{q_+r} f_{q_+}$. Finally, for any $r \in q_+^{\text{in}}$, $a^-_{q_+r} f_{q_+} \leqslant m_r/\tau_r$. This proves (4.15c).
- let q_- be the non priority transition of a place p subject to priority. Then $(C^- f)_{\pi(q_-)} = (C^-_\pi f)_{q_-}$ for any policy π, so that (4.15c) follows from (4.26). ◀

As a consequence of Theorem 4.11, we obtain a correspondence between the stationary solutions of the continuous dynamics and the stationary solutions of the discrete dynamics. In order to highlight the parallel between the discrete and the continuous setting, we denote by f_p the processing flow m_p/τ_p for every place p.

▶ **Corollary 4.12.** *(i) Suppose (m, w, f) defines a stationary solution of the continuous dynamics. Then, for the initial marking $M^0_p = m_p$, setting $\rho := f$, $u_p := M^0_p$, and $u_q := 0$ yields a stationary solution of the discrete dynamics.*

(ii) Conversely, suppose (ρ, u) is a stationary solution of the discrete dynamics. Then, defining $f := \rho$, setting $m_p := \rho_p \tau_p$ for every place p, and defining w according to (4.25b) and (4.25d) yields a stationary solution of the continuous dynamics.

Proof. Both statements are straightforward. We point out that (4.23a) reads $\rho_p = C^+ \rho_q$ and that (4.23b)–(4.23e) are equivalent to $\rho_q = \min_\pi (C^-_\pi)^{-1} S_\pi \rho_p$. The same relationship between the f_q and the f_p was established in the proof of Theorem 4.11. ◀

4.4.3 Properties of the stationary solutions

The characterization (4.25) leads to simple algebraic facts on the affine stationary solutions of Petri nets with free-choice and priority. For example, a stationary flow always exists, because the vector $0^{\mathcal{Q}}$ is such that $C0 \geqslant 0$ (but the associated marking may not exist).

If the Petri net is conservative, *i.e.*, there exists a positive y such that $y^\mathsf{T} C = 0$, then any $x \geqslant 0$ such that $Cx \geqslant 0$ necessarily satisfies $Cx = 0$, because $y^\mathsf{T} Cx = 0$ and $y > 0$. Consequently, if the Petri net is conservative, the possible stationary flows are the nonnegative vectors of the kernel of C (the nonnegative $\mathcal{Q}$-invariants of the Petri net). If 0 is the only annuller of C, this means that a stationary solution is such that all the markings m_p are zero. In such stationary solution, the transitions do not fire anymore and hence no tokens are under processing, even if some tokens may remain idle in some places.

Furthermore, we come across an equivalent of the classical Little's law applied to Petri nets, and the corresponding classical upper bound on the Petri net flow:

▶ **Proposition 4.13.** *Suppose that the stationary flow f is known to be proportional to a given vector u, $f = \lambda u$. (For example, the Petri net is conservative, and it admits a unique $\mathcal{Q}$-invariant, up to a multiplicative constant).*

(i) Let y be a nonnegative $\mathcal{P}$-invariant of the Petri net. Then,

$$\lambda (y^\mathsf{T} D C^+ u) \leqslant y^\mathsf{T} M^0,$$

where M^0 is the vector of the initial markings, $M^0_p = m_p(0) + w_p(0)$.

(ii) If $\mathcal{Y}$ is a family of minimal nonnegative $\mathcal{P}$-invariants of the Petri net,

$$\lambda \leqslant \min_{y \in \mathcal{Y}} \frac{y^\mathsf{T} M^0}{y^\mathsf{T} D C^+ u} \tag{4.28}$$

Proof. (i) This follows from the token conservation $y^\mathsf{T}(m_\infty + w_\infty) = y^\mathsf{T} M^0$, for a stationary solution with markings m_∞, w_∞, and from Equation (4.25a).

(ii) Straightforward from the previous item. ◀

Determining under which conditions equality holds in (4.28) is a well-known question in the Petri net literature. See for example [CCS91]. See Section 2.2.3 on computing minimal $\mathcal{P}$-invariants of a Petri net.

Regarding this important issue of relating the stationary flow to the initial marking, very little can be obtained on top of the relations given by the invariants of the Petri nets, as most results in this direction are limited to nets without priorities, and rely on monotonicity

properties of the dynamics. The next theorem identifies, however, a somehow special situation in which such a relation persists even in the presence of priority. This applies in particular to the Petri net of the next section.

▶ **Theorem 4.14.** *If a trajectory of the continuous Petri net converges towards a stationary solution $(m^\infty, w^\infty, f^\infty)$, if for this trajectory, there exists a policy π valid on $\mathbb{R}_{\geqslant 0}$, and if 0 is a semi-simple eigenvalue of $(S_\pi C^+ - C_\pi^-)$ associated with this policy, then f^∞ is uniquely determined by the initial marking.*

(Recall that the eigenvalue λ of a matrix B is said to be *semi-simple* if the dimension of its eigenspace is equal to its algebraic multiplicity, that is, to the multiplicity of λ as the root of the characteristic polynomial of B. In particular, if 0 is a semi-simple eigenvalue of B, then the kernel of B and its range space are complementary subspaces.)

Proof. Under the conditions of the theorem, there exists a policy π such that, for any t,

$$S_\pi \dot{m}(t) = (S_\pi C^+ - C_\pi^-) f(t)\,, \tag{4.29}$$

$$S_\pi m(t) = D_\pi C_\pi^- f(t)\,, \tag{4.30}$$

see Relations (4.16) and (4.17).

Since 0 is a semi-simple eigenvalue of $(S_\pi C^+ - C_\pi^-)$, the same property holds for the matrix $(S_\pi C^+ - C_\pi^-)(D_\pi C_\pi^-)^{-1} = (S_\pi C^+ (C_\pi^-)^{-1} - I) D_\pi^{-1}$. Therefore, the kernel of this matrix and its range space are complementary subspaces. We denote by Q the projection onto the former along the latter.

By (4.29), we obtain that $QS_\pi \dot{m}(t) = Q(S_\pi C^+ - C_\pi^-) f(t) = 0$, so that $QS_\pi m(t)$ is independent of time, and

$$QS_\pi m(0) = QS_\pi m_\infty = QD_\pi C_\pi^- f_\infty\,.$$

Moreover, as $(m^\infty, w^\infty, f^\infty)$ is a stationary solution of the continuous dynamics, Equation (4.25e) holds and $D_\pi C_\pi^- f_\infty$ belongs to the kernel of $(S_\pi C^+ (C_\pi^-)^{-1} - I) D_\pi^{-1}$. Therefore,

$$f_\infty = (C_\pi^-)^{-1} D_\pi^{-1} Q S_\pi m(0)\,. \qquad \blacktriangleleft$$

4.5 Numerical experiments

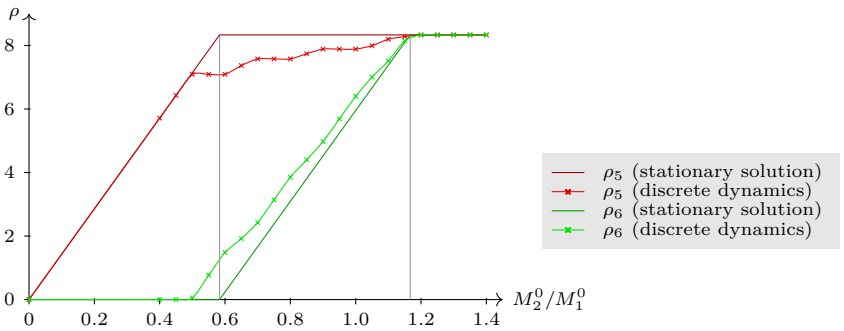
The dynamics expressed by (4.1)–(4.2), (4.5) and (4.7)–(4.9) belongs to the class of hybrid automata [Hen00], which can handle piecewise linear but discontinuous dynamics like ours. We simulate our dynamics with the tool SpaceEx [FLGD+11], which is a verification platform for hybrid systems. The particularity of SpaceEx is that it computes a sound over-approximation of the trajectories.

4.5.1 The two-level emergency call center Petri net

In this section, we illustrate our results on our running model of an emergency call center with two treatment levels, introduced in Section 2.5.1, see Figure 2.7. Every arc has a valuation equal to one. The initial marking $M_1^0 = N_1$ (resp. $M_2^0 = N_2$) of place p_1 (p_2) denotes the available number of operators of level 1 (level 2) in the call center.

It was observed in Chapter 3 that the discrete dynamics has a pathological feature: when certain arithmetic relations between the time delays are satisfied, the discrete time trajectory may not converge to a stationary solution, and its asymptotic throughput may differ from the throughput of the stationary solution. It follows from our correspondence result (Corollary 4.12) that the continuous dynamics has the same stationary solutions as the corresponding discrete dynamics. We shall observe that, in this continuous setting, the trajectory converges towards a stationary solution, so that the former pathology vanishes.

We recall in Figure 4.3 the throughputs of transitions q_5 and q_6 (see Figure 2.7), obtained for the discrete dynamics, compared with the throughputs of the stationary solutions, computed by System (4.25). This is the same computation that was led in Section 3.6.2, and we keep the same parameters.



The initial markings of the places different from p_1 and p_2 are null. The holding times are $\tau_{\text{ext}} = 4$, $\tau_{\text{ur}} = 3$, $\tau_{\text{adv}} = 3$, $\tau_{\text{tr}} = 1$, $\tau'_{\text{ext}} = 6$, $\tau'_{\text{ur}} = 7$, and 0.01 for the remaining places.

Figure 4.3 – Emergency call center Petri net. Comparison of the throughputs of the discrete dynamics simulations with the theoretical throughputs (fluid model).

M_2^0/M_1^0	0.2	0.4	0.6	0.8	1.0	1.2
ρ_5	2.857	5.714	8.333	8.333	8.333	8.333
f_5^{up}	2.865	5.716	8.334	8.338	8.339	8.340
f_5^{down}	2.849	5.707	8.333	8.328	8.328	8.327
ρ_6	0	0	0.238	3.095	5.952	8.333
f_6^{up}	< 0.001	< 0.001	0.239	3.107	5.968	8.340
f_6^{down}	0	0	0.237	3.083	5.936	8.327

Table 4.1 – Lower and upper bounds of the throughputs of the continuous dynamics computed by SpaceEx, and comparison with the stationary throughputs

At the scale of Figure 4.3, the lower and upper bounds to the values of the throughputs, computed by SpaceEx, coincide with the shape of the stationary throughputs curve. Table 4.1 compares the numerical values of these lower and upper bounds with the stationary throughputs for a few values of M_2^0/M_1^0. We observe that the over-approximation computed by SpaceEx provides an accurate estimate of the stationary throughput computed via System (4.25). This tends to show that the continuous dynamics converges towards the stationary throughputs, unlike the discrete dynamics.

4.5.2 The SR Petri net

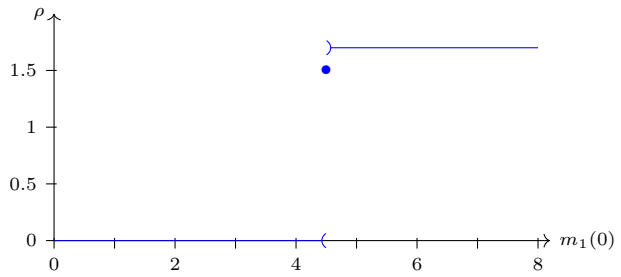
We consider the Petri net of Silva and Recalde (2002), introduced in Section 2.5.2. As for the previous Petri net, our experiments show that the throughput always reaches one of the stationary states computed in Section 3.5.

This contrasts with our observations in the case of the fluid approximation of the discrete dynamics, see Section 3.6.2.

In Figure 4.4, we show simulation results in the case when several asymptotic throughputs are possible, that is, when the system parameters satisfy $\pi_1\tau \leqslant \pi_2\tau$ and $\pi_1\tau/\pi_2\tau \leqslant y_1^{\mathsf{T}}N^0/y_2^{\mathsf{T}}N^0 \leqslant 1$. We vary the initial markings $m_1(0)$, $m_2(0)$, $m_3(0)$, while keeping the Petri net invariants $y_1^{\mathsf{T}}N^0$, $y_2^{\mathsf{T}}N^0$ constant.

We observe that, when $m_1(0)$ increases, the different possible asymptotic throughputs are reached. In our experiments, the throughput $(y_1 - y_2)^{\mathsf{T}}N^0/(\pi_1 - \pi_2)\tau$ is reached only when the initial markings are the markings of the corresponding stationary solution. Therefore, this seems to be an unstable asymptotic state.

▶ **Remark 4.15** (Discontinuities and non monotonicities)**.** For the model of continuous Petri nets with race policy routing ([DA87, VMJS13]), discontinuities and non monotonicities of throughputs in terms of the parameters of the system were observed and analyzed in a series of works [JRS05, Mey12, NGRTS16]. Such behaviors require special attention, because, in the



The parameters are those of Figure 3.5 Right. $\tau = (1, 16, 4, 1, 1)$, $m_4(0) = 2.5$, $m_5(0) = 1$, $m_3(0) = 14.5 - m_1(0)$ and $m_2(0) + m_3(0) = 34$. All the w_p are initially null.

Figure 4.4 – SR Petri net. Asymptotic throughputs with $m_1(0)$, $m_2(0)$, $m_3(0)$ varying, while keeping the invariants constant (with τ_2 large, and in the intermediate phase). The three possible throughputs are reached when the parameters vary.

modeling of real systems, a slight perturbation on one transition's firing time would lead to a completely different throughput, or increasing one transition's firing rate would decrease the global throughput of the Petri net (for example, increasing the failure rate of an equipment in a manufacturing system decreases the production throughput).

In comparison, in the same SR Petri net, but with a priority routing and our differential semantics, we also observe that, in the intermediate case, the throughput is decreasing in the marking N_2, (not that it is nondecreasing in the normalized parameter $y_1^\mathsf{T} N^0 / y_2^\mathsf{T} N^0$). It is also increasing in the holding time τ_2. Moreover, in the situation $\pi_1\tau \leqslant \pi_2\tau$, we can observe discontinuities in the values of the throughput. Take the case $\pi_1\tau = \pi_2\tau$ for example: if $y_1^\mathsf{T} N^0 < y_2^\mathsf{T} N^0$, then the throughput is 0. If $y_1^\mathsf{T} N^0 > y_2^\mathsf{T} N^0$, the throughput is $y_2^\mathsf{T} N^0 / \pi_2\tau$. The case $y_1^\mathsf{T} N^0 = y_2^\mathsf{T} N^0$ depends on the value of the different coordinates of N^0.

4.6 Concluding remarks

We introduced a hybrid dynamical system model for continuous Petri nets having both free choice and priority places, and showed that there is a correspondence between the stationary solutions of the continuous dynamics and the discrete one. An advantage of the continuous setting is that some pathologies of the discrete model (failure of convergence to a stationary solution) may vanish. This is the case in our two examples, and we therefore may apply this dynamics to model the behavior of our different emergency call center architectures.

Regarding the model introduced in this paper, further investigations still need to be done. First, as already mentioned in Section 2.3.2, we remark that, since the introduction of hybrid dynamics of Petri nets, no formal proof has been proposed that the markings reached by these differential equations belong to the set of reachable markings of the untimed Petri net with fractioned firings, analyzed for example in [RTS99]. We leave it for further work to propose such proof, for our dynamics, as for the original hybrid dynamics of [DA87].

Now, concerning our differential system, it still remains to see under which generality convergence towards the stationary solution can be established. In particular, one would like to know if stationary regimes with oscillations could exist, or if asymptotic throughputs different from the throughputs of the affine solutions could be reached for some systems.

Finally, we remark that forward Carathéodory solutions of a piecewise linear system do not exclude Zeno behaviors. Can such phenomena be observed for our class of differential equations?

4.7 Proof of Proposition 4.6

We recall here that we are interested in the policies δ such that, for any other policy π,

$$\forall q \in \mathcal{Q}, \quad (S_\delta w_0, C_\pi^- f_\delta^{(1)}(0), C_\pi^- f_\delta^{(2)}(0), \dots)_q \overset{\text{lex.}}{\leqslant} (S_\pi w_0, S_\pi D^{-1} m_\delta^{(1)}(0), S_\pi D^{-1} m_\delta^{(2)}(0), \dots)_q \,. \quad (4.31)$$

We build our proof of Proposition 4.6 on three lemmas.

▶ **Lemma 4.16.** *Let $x \in \mathbb{R}^{\mathcal{Q}}_{\geqslant 0}$. There exists, for any transition q, a subset of upstream places $\mathrm{up}^x(q) \subseteq q^{in}$, such that any policy σ associating transitions to their upstream places in up^x satisfies the relation*

$$\forall \pi, \quad C^-_\pi (C^-_\sigma)^{-1} S_\sigma x \leqslant S_\pi x \,, \tag{4.32}$$

and such that, for any other policy, the relation would not hold. Moreover, the quantity $(C^-_\pi)^{-1} S_\pi x$ is independent of π selecting places in the subsets up^x, so that, for these policies, the relation is an equality.

The result still holds if we restrict ourselves to policies which associate to each transition a place belonging to fixed subsets of q^{in}.

Proof. We proceed by necessary conditions. Suppose that σ is such that (4.32) holds. Then, for any policy π, using the decomposition of C^-_π and of its inverse on $\mathcal{Q}_{0+} \cup \mathcal{Q}_-$ given in Section 4.3.1, necessarily,

$$A^\pi_+ (A^\sigma_+)^{-1} (S_\sigma x)_{0+} \leqslant (S_\pi x)_{0+} \,,$$

where, for a vector z, z_{0+} designates the coordinates of z in $\mathcal{Q}_{0+}$.

Matrix A^π_+ is a diagonal matrix with positive entries, so that any σ such that (4.32) holds is also such that $(A^\sigma_+)^{-1}(S_\sigma x)_{0+}$ is minimal in the set $S_1 = \{(A^\pi_+)^{-1}(S_\pi x)_{0+} \mid \pi \text{ policy}\}$. Moreover, for $q \in \mathcal{Q}_{0+}$, it happens that the value of $((A^\pi_+)^{-1}(S_\pi x)_{0+})_q$ depends only on the value of $\pi(q)$ (and not on any other $\pi(q')$). We denote by $a(x, q, \pi(q))$ this quantity. Consequently, S_1 has a minimum which can be chosen componentwise according to the following procedure: for any $q \in \mathcal{Q}_{0+}$, let $\mathrm{up}^{x,+}(q) = \mathrm{argmin}\{a(x, q, p) \mid p \in q^{\mathrm{in}}\}$. Therefore, the set of policies such that (4.32) holds is included in the set of policies such that the upstream places of transitions $q \in \mathcal{Q}_{0+}$ are in $\mathrm{up}^{x,+}(q)$, and we denote by y_{0+} the common value $(A^\sigma_+)^{-1}(S_\sigma x)_{0+}$ (and y_{0+} is nonnegative).

Now, using again the decomposition of C^-_π and of its inverse, a policy σ such that (4.32) holds should also satisfy

$$A^\pi_{-+}(A^\sigma_+)^{-1}(S_\sigma x)_{0+} + A^\pi_-(A^\sigma_-)^{-1}\left((S_\sigma x)_- - A^\sigma_{-+}(A^\sigma_+)^{-1}(S_\sigma x)_{0+}\right) \leqslant (S_\pi x)_-$$

which, by replacing $(A^\sigma_+)^{-1}(S_\sigma x)_{0+}$ by y_{0+}, gives

$$A^\pi_{-+} y_{0+} + A^\pi_-(A^\sigma_-)^{-1}\left((S_\sigma x)_- - A^\sigma_{-+} y_{0+}\right) \leqslant (S_\pi x)_- \,.$$

After a few manipulations, we get

$$(A^\sigma_-)^{-1}\left((S_\sigma x)_- - A^\sigma_{-+} y_{0+}\right) \leqslant (A^\pi_-)^{-1}\left((S_\pi x)_- - A^\pi_{-+} y_{0+}\right) \,,$$

using the fact that A^π_- is diagonal, with positive diagonal entries.

Therefore, a policy σ satisfying (4.32) also attains the minimum in the set $S_2 = \{(A^\pi_-)^{-1}\left((S_\pi x)_- - A^\pi_{-+} y_{0+}\right) \mid \pi \text{ policy}\}$. Now, for a transition $q \in \mathcal{Q}^-$, remember that q is the non priority transition of an upstream place p_0 subject to priority. Entry (q, q') of matrix A^π_{-+} is different from 0 only if $\pi(q) = p_0$, and if q' is the priority transition of p_0. In this situation, the value of the entry is $a^-_{q'\pi(q)}$ and does not depend on the other values of π (in particular, it does not depend on $\pi(q')$). Therefore, again, for any transition $q \in \mathcal{Q}_-$, one can choose independently a subset $\mathrm{up}^{x,-}(q) \subseteq q^{\mathrm{in}}$ of places realizing the minimum for coordinate q, and a policy satisfying (4.32) should choose upstream places of transitions of $\mathcal{Q}_-$ in the sets $\mathrm{up}^{x,-}(q)$. The minimum of S_2 is denoted by y_- and is reached by any such policy.

To summarize, necessarily, a policy such that (4.32) holds associates with transitions $q \in \mathcal{Q}_{0+}$ places in the respective $\mathrm{up}^{x,+}(q)$ and with transitions $q \in \mathcal{Q}_-$ places in the respective $\mathrm{up}^{x,-}(q)$ (depending on the value obtained by selecting places in $\mathrm{up}^{x,+}(q)$). We can hence build a selection of upstream places $\mathrm{up}^x(q)$ for any $q \in \mathcal{Q}$, composed of the maps $\mathrm{up}^{x,+}$ and $\mathrm{up}^{x,-}$. The value of $(C^-_\sigma)^{-1} S_\sigma x$ is independent of a policy associating to each transition q a place of $\mathrm{up}^x(q)$, and equals $y := (y_{0+}, y_-)$. In addition, such y is constructed such that it satisfies, for any policy π, $C^-_\pi y \leqslant S_\pi x$. This concludes the proof in the general case.

Concerning the last assertion, we observe that, when constructing the sets $\mathrm{up}^x(q)$, it suffices to consider only the places belonging to the required subsets. ◀

By the previous lemma, we are able to compare the elements of (4.31) pairwise. The next lemma now compares the sequences of elements, up to an arbitrary index i.

▶ **Lemma 4.17.** *Let $i \geqslant 0$. For each transition q, there exists a non empty subset $\mathrm{up}^i(q) \subseteq q^{in}$ of upstream places of q such that, for each policy δ selecting for each q a place in $\mathrm{up}^i(q)$, for each policy π,*

$$\forall q \in \mathcal{Q}, \quad (S_\delta w_0, C_\pi^- f_\delta^{(1)}(0), \dots, C_\pi^- f_\delta^{(i-1)}(0))_q \overset{\text{lex.}}{\leqslant} (S_\pi w_0, S_\pi D^{-1} m_\delta^{(1)}(0), \dots, S_\pi D^{-1} m_\delta^{(i-1)}(0))_q \,, \quad (4.33)$$

and the inequality is an equality for each transition if and only if, for any q, policy π selects a place of $\mathrm{up}^i(q)$.

We note Π^i the set of policies associating with each q an upstream place in $\mathrm{up}^i(q)$.

Proof. We prove the lemma by induction on i. If $i = 0$, this simply corresponds just to selecting, for each transition q the upstream places whose marking $(w_0)_p$ is zero:

$$\forall q \in \mathcal{Q}, \quad \mathrm{up}^0(q) = \{p \in q^{\text{in}} \mid (w_0)_p = 0\} \,,$$

and none of these sets are empty, by the initial conditions (4.10). Moreover, it follows from the definition of $\mathrm{up}^0(q)$ that, if, for some policy π and transition q, $\pi(q) \notin \mathrm{up}^0(q)$, then $(w_0)_{\pi(q)} > 0$, so that the inequality (4.33) is strict for this transition. Finally, for any policies δ, $\sigma \in \Pi^0$, $S_\delta w_0 = S_\sigma w_0 = 0$.

Now, let $j \geqslant 0$ and suppose that the result of the lemma holds for j. Suppose, that, moreover, it holds that the vectors $f_\sigma^{(1)}(0), \dots, f_\sigma^{(j-1)}(0)$ are independent of $\sigma \in \Pi^j$. We note them $(d_1, \dots, d_{j-1})$. We prove these properties for $j+1$. First observe that the vectors $m_\sigma^{(1)}(0), \dots, m_\sigma^{(j-1)}(0)$ are independent of $\sigma \in \Pi^j$. Indeed, this is true for $m_\sigma(0) = m(0) = m_0$, and this can be shown by a straightforward induction, using Relation (4.11) and the fact that the result holds for the $f_\sigma^{(i)}$. We note these quantities $(g_1, \dots, g_{j-1})$. Moreover, we observe that, again because of the relation $m_\sigma^{(j)}(0) = C^+ d_{j-1} - D^{-1} g_{j-1}$, this is also true for $m_\sigma^{(j)}(0)$, which we denote g_j. We finally note that, for $\sigma \in \Pi^j$, $f_\sigma^{(j)} = (C_\sigma^-)^{-1} S_\sigma D^{-1} g_j$, by construction of f_σ.

By Lemma 4.16, the set of policies $\sigma \in \Pi^j$ such that, for any other policy $\pi \in \Pi^j$,

$$C_\pi^- (C_\sigma^-)^{-1} S_\sigma D^{-1} g_j \leqslant S_\pi D^{-1} g_j \,,$$

is exactly the set of policies associating to each transition q a place of $\mathrm{up}^{(D^{-1}g_j)}(q) \subseteq \mathrm{up}^j(q)$, and there is equality for two policies of this kind. We define $\mathrm{up}^{j+1}(q) := \mathrm{up}^{(D^{-1}g_j)}(q)$, and we denote by Π^{j+1} the corresponding set of policies. Moreover, by Lemma 4.16 again, for $\sigma \in \Pi^{j+1}$, $f_\sigma^{(j+1)}(0) = (C_\sigma^-)^{-1} S_\delta D^{-1} g_j$, and it is independent of $\sigma \in \Pi^{j+1}$.

By the induction assumption, as $\Pi^{j+1} \subseteq \Pi^j$, and by Lemma 4.16 for index j, any $\delta \in \Pi^{j+1}$ satisfies, for any other policy $\sigma \in \Pi^j$, for any q,

$$(C_\sigma^- d_1, \dots, C_\sigma^- d_{j-1}, C_\sigma^- (C_\delta^-)^{-1} S_\delta D^{-1} g_j) \overset{\text{lex.}}{\leqslant} (S_\sigma D^{-1} g_1, \dots, S_\sigma D^{-1} g_j)$$

and there is equality if $\sigma \in \Pi^{j+1}$.

Now, suppose that, for another policy $\pi \notin \Pi^j$, we have for a transition q, $\left(C_\pi^- (C_\sigma^-)^{-1} S_\sigma D^{-1} g_j\right) > (S_\pi D^{-1} g_j)_q$. Then, necessarily, $\pi(q) \notin \mathrm{up}^j(q)$. The sets $(\mathrm{up}^i(q))_{i \geqslant 0}$ form a nondecreasing sequence for the inclusion relation $\subseteq$, so that, either $\pi(q) \notin \mathrm{up}^0(q)$, which means that $w_{\pi(q)} > 0$, or there exists an index $i < j$ such that $\pi(q) \in \mathrm{up}^i(q)$ and $\pi(q) \notin \mathrm{up}^{i+1}(q)$. This implies, by construction of $\mathrm{up}^{i+1}(q)$, $\left(C_\pi^- (C_\sigma^-)^{-1} S_\sigma D^{-1} g_i\right)_q < (S_\pi D^{-1} g_i)_q$, and, for any $k < i$, $\left(C_\pi^- (C_\sigma^-)^{-1} S_\sigma D^{-1} g_k\right)_q < (S_\pi D^{-1} g_k)_q$. Therefore, using the induction assumption, the lexicographic ordering also holds for policies π outside Π^j.

Finally, the set of policies Π^{j+1} is such that (4.33) holds for any policy π, and equality holds only for pairs of transitions of Π^{j+1}. This completes the proof. ◀

The next lemma shows that, in Relation (4.31), it suffices to compare the $m+2$ first elements.

▶ **Lemma 4.18.** *If, for two policies π and σ, $f_\pi^{(i)}(0) = f_\sigma^{(i)}(0)$ for $i \in \{1, \dots, m+1\}$, where $m := |\mathcal{Q}|$, then equality holds for any $i > m+1$.*

Proof. Recall that, by (4.18) and by definition of matrix B_π, for any policy, and $i > 1$, $f_\pi^{(i)}(0) = B_\pi f_\pi^{(i-1)}(0) = B_\pi^{i-1} f_\pi(0)$.

Let us also note that, by the Cayley-Hamilton Theorem, B_π is a zero of its characteristic polynomial, so that B_π^m can be expressed in terms the smaller powers of B_π:

$$B_\pi^m = \sum_{i=0}^{m-1} \lambda_i B_\pi^i \,.$$

Therefore, by multiplying the relation by $f_\pi(0)$ on the right, the following holds:

$$f_\pi^{(m+1)}(0) = \sum_{i=0}^{m-1} \lambda_i f_\pi^{(i+1)}(0) \,. \tag{4.34}$$

Now, let π, σ be such that $f_\pi^{(i)}(0) = f_\sigma^{(i)}(0)$ for $i \in \{1, \dots, m+1\}$. We prove by induction that the equality is true for any $i \in \mathbb{N}$. Note that it is true for $i \in \{1, \dots, m+1\}$ by hypothesis. Now, take $l > m$ and suppose that the equality holds for any $i \leqslant l$. We prove the equality for $l+1$. By multiplying Relation (4.34) by B_π^{l-m}, we get

$$\begin{aligned} f_\pi^{(l+1)}(0) &= \sum_{i=0}^{m-1} \lambda_i f_\pi^{(l-m+i+1)}(0) \\ &= \sum_{i=0}^{m-1} \lambda_i f_\sigma^{(l-m+i+1)}(0) \\ &= B_\sigma^{l-m} \sum_{i=0}^{m-1} \lambda_i f_\sigma^{(i+1)}(0) \\ &= B_\sigma^{l-m} f_\sigma^{(m+1)}(0) = f_\sigma^{(l+1)}(0) \,. \end{aligned}$$

The second and fourth equalities follow from the induction hypothesis. ◀

Proof of Proposition 4.6. We apply Lemma 4.17 with $i = m+2$. For the policies of Π^{m+2}, that realize the equality in (4.33), equality further holds for any $j > m+2$ by Lemma 4.18. The set of policies such that (4.31) holds is therefore exactly the set Π^{m+2}, and equality holds in the relation for two policies of Π^{m+2}.

Now if Π^{m+2} contains two different policies σ and δ, the equality of the derivatives at time 0 implies that the quantities $e^{B_\sigma t} f_\sigma(0)$ and $e^{B_\delta t} f_\delta(0)$ are identical for any t, so that the two policies lead to the same dynamics on the interval on which they are valid. ◀

Chapter 5

A stochastic analysis of a network with two levels of service

This chapter is a joint work with Philippe Robert (Inria de Paris). This work has led to a pre-print [BR17], submitted to the journal *Mathematics of Operations Research*. It is included as such.

5.1 Introduction

The motivation of the model analyzed in this paper originates from a collaboration with "Préfecture de police de Paris", the police department of Paris, and "Brigade de sapeurs-pompiers de Paris", the fire department of Paris, to design an emergency call center in charge of receiving emergency calls for police *and* for firemen in Paris area. The previous organization had two independent call centers with a single level of operators. The new call center has an architecture with two levels of operators. A first-level pool of operators handles (numerous) non-urgent calls and has to detect and transfer calls classified as urgent to a second-level pool of more specialized operators, policemen or firemen, depending on the nature of the call. Second level operators may dispatch emergency means, if needed. The first level pool operates therefore as a filter so that the second-level pool can process efficiently urgent calls. An additional, natural, constraint is that if a first level operator has detected an urgent call, this operator releases the call only when a second level operator has handled it. In particular, the operator will wait when all servers of the second level are busy. In this situation there are two issues: firstly, the handling of the urgent call is delayed and, secondly, the server of the first level is blocked and, consequently, the processing capacity of the first level is reduced. The main problem in the

design of this new organization is of determining a minimal number of (expensive) second level operators necessary so that this blocking phenomenon has a small probability.

We will investigate the behavior of this architecture in stressed situations, i.e., when a large number of incoming calls is arriving at the first level. A key characteristic to analyze in this situation is the evolution of the number of blocked operators at level 1. This number should remain small in a convenient design. For this reason, it will be assumed that an infinite number of calls are waiting for processing in a queue. Calls require random processing time whose distribution depends on the level and the class of the call (urgent or non-urgent). We now give a quick description of this system in terms of a queueing model.

A Queueing Description of the System

As input, there is an infinite queue of jobs waiting to enter the system, this is the saturation assumption mentioned above. With probability $p \in [0,1]$ a job is of class 0, otherwise it is of class 1. A job of class 0 represents an urgent call, otherwise it is a non-urgent call.

1. The first level has C_1 servers.
   Every time a server of this level is idle, it immediately receives a job from the infinite queue. It is of class $i \in \{0,1\}$ with probability $p \in [0,1]$ and $1-p$, respectively. A job of class i requires an exponentially distributed service with rate μ_{i1} at this level.
   Class 0 jobs are urgent calls and have to be processed by level 2.
   (a) When a job of class 0 completes its service at level 1, it goes to the second level if there is at least one idle server there.
   (b) If there is no place then it remains at the first level and, consequently, blocks a server at this level. As soon as a job leaves the second level, a blocked job at the first level is sent to the second level and the server can take a new job in the infinite queue.
   When a job of class 1 completes its service, it leaves the system.
2. The second level has C_2 servers and receives only class 0 jobs. A job at this level requires a exponentially distributed service with rate μ_{02}.

See Figure 5.1. A key feature of this network is that blocked jobs of class 0 at level 1 reduce the capacity of the system since the corresponding servers at level 1 cannot process the calls waiting in the saturated queue.

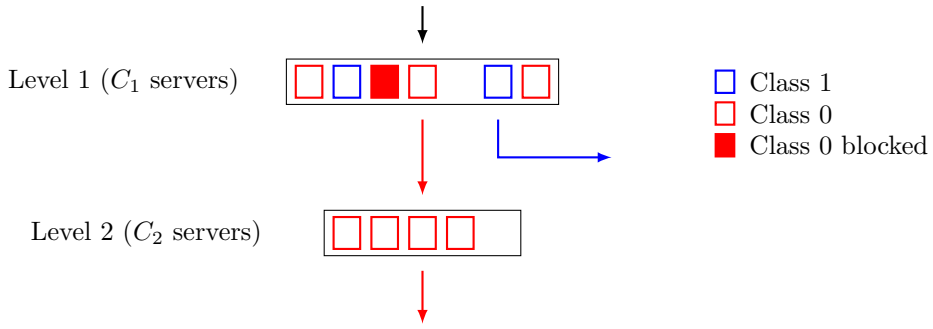


Figure 5.1 – Queueing System with Two Levels

Literature

Deterministic Modeling

In this paper, the classes of calls and their processing times are assumed to be random. In a non-random setting, some aspects of this system have been investigated in Allamigeon et al. [ABG15, ABG17] where a performance analysis was carried out using a deterministic Petri net modeling. A Petri net is a language describing systems in which resources circulate from place to place, incurring concurrency, synchronizations and bifurcations [Mur89, BW13]. The dynamics of a Petri net can be translated into a dynamical system, whose stability and stable points can be analyzed, see Cohen et al. [CGQ95]. In Allamigeon et al. [ABG15, ABG17], a

simplified model of the emergency call center is investigated. Computations on the stationary regimes of the dynamical system have shown a phase transition characterizing the different levels of congestion of the call center. The threshold is a critical ratio between the number of operators at level 1 and level 2.

The analysis of Petri net models may give general results for this class of systems in a deterministic framework. However, the dynamics investigated in the above articles do not take account of the random nature of the delays or the classes of calls in the call center for example. In contrast, the queueing network analysis adopted in the present article focuses on a simpler system describing the transfer or the blocking of calls from level 1 to level 2. As it will be seen, it provides a deeper understanding of the behavior of this system in a random context.

It should be noted that the differential equations resulting from the continuous Petri net modeling of Allamigeon et al. [ABG17] do correspond do the dynamical system that we obtain below as the scaling limit of our model. This highlights the consistency and strong relationship between both analyses.

Queueing Models of Blocking Phenomena

A natural class of stochastic models related to the system described above is that of call centers. There is a huge literature dealing with the problem of staffing these systems. To the best of our knowledge, few seem to have considered jobs going through a series of call centers as in our case. The closest models of this literature seem to be multi-skill call centers where jobs can have different levels of quality of service depending on the call center chosen. They are nevertheless addressing quite different problems than the ones considered in this paper. See Koole and Mandelbaum [KM02] for a survey.

The model that we are studying can be described in terms of finite capacity queues with blocking in tandem. The blocking has the effect that, when a server at level 1 completes the service of a class 0 job, it cannot be used again until a server at level 2 is available. At level 1 a fraction of the servers, and consequently the corresponding calls, may be blocked. Related models have been investigated in the literature, see the survey Balsamo [Bal11]. The papers study the corresponding finite Markovian models of these systems to express in particular the blocking probability at equilibrium. The corresponding equilibrium equations do not have, in general, a solution with a closed form expression. When the values of the capacities (the numbers of operators) are not small, the dimension of the state space can be quite large so that a numerical procedure can also be out of reach in practice. Some approximations have been proposed but, for the moment, without any convergence result which could give an idea of the accuracy of such estimations. Kelly [Kel86] has investigated the problem of blocking of a series of queues, the analysis is concentrated on the estimation, via bounds, of the achievable throughput of such a system. To conclude, the literature of rigorous mathematical results for finite capacity queues with blocking is therefore somewhat scarce.

When the blocking is replaced by the following mechanism defined as an exclusion process: a job blocked at some stage immediately repeats a service until the next stage can accommodate it, the situation is quite different. Some of the mathematical models related to the asymmetric simple exclusion process can give some insights on the performances of these systems(*e.g.*, throughputs). Due to its relative mathematical tractability, the literature investigating these processes is also huge. See, for example Liggett [Lig85] for a general presentation of these important processes and Liggett [Lig75] for a study of asymmetric simple exclusion process in finite dimension. These models are however quite different and do not seem to be usable since the blocking phenomenon of interest is not really taken into account.

Contributions

With the above notations for our system, one of the main results of the paper, Theorem 5.13, shows that, under appropriate scaling conditions, if r is the ratio of the capacities of the two levels, $r{=}C_2/C_1$, then the condition

$$r\left(\frac{p}{\mu_{01}}+\frac{1-p}{\mu_{11}}\right)>\frac{p}{\mu_{02}} \tag{5.1}$$

implies that there exists some fixed instant independent of the initial state such that after that time, with high probability, there are no blocked customers at level 1 on any finite time interval.

See Theorem 5.13 and Corollary 5.12.

Otherwise, if the opposite (strict) inequality

$$r\left(\frac{p}{\mu_{01}}+\frac{1-p}{\mu_{11}}\right)<\frac{p}{\mu_{02}} \tag{5.2}$$

holds then, Theorem 5.9 shows that, under appropriate scaling conditions, the fraction of blocked customers at level 1 is positive after some time almost surely and it converges to

$$1-\frac{\mu_{02}}{\mu_{01}}\frac{C_2}{C_1}\left(\frac{(1-p)\mu_{01}}{p\mu_{11}}+1\right).$$

See also Corollary 5.8.

Consequently, as the intuition suggests, if the ratio C_2/C_1 of the number of servers is larger than the ratio of the loads of the two levels, then the phenomenon of blocking will not occur with high probability. Relation (5.1) gives therefore a rule for a convenient design of such a system.

A Heuristic Picture

Assume that there is no blocking at level 1 of class 0 jobs. Level 1 can be seen as a simple birth and death process described by the number of jobs ($Q(t)$) of class 0. A birth (resp. death) occurs when a job of class 1 (resp. 0) completing its service is replaced by a job of class 0 (resp. 1). Therefore in state $x\in\{0,\dots,C_1\}$, the birth rate is $p(C_1-x)\mu_{11}$ and the death rate is $(1-p)\mu_{01}x$. At equilibrium these two rates should be of the same order and therefore that, for a large C_1, the number Q_0 of class 0 jobs is of the order of

$$Q_0\sim C_1\frac{p\mu_{11}}{(1-p)\mu_{01}+p\mu_{11}}.$$

To avoid congestion, the rate $\mu_{01}Q_0$ at which class 0 jobs enter level 2 must be smaller than the maximal output rate of the second level, that is $C_2\mu_{02}$. This gives exactly Condition (5.1).

Mathematical Aspects

Proving rigorously these intuitive results turns out to be, quite surprisingly, challenging. The Markov process associated with the queueing system has a finite state space, included in $\mathbb{N}^3$. It bears some similarity with classical loss networks of the literature but with a routing mechanism as in Jackson networks. See Kelly [Kel79]. As such, little can be said for this process, in particular its invariant probability distribution does not seem to have a simple closed form expression.

To get quantitative results on this system a scaling approach is used. It is assumed that the capacities C_1 and C_2 are both large so that C_2/C_1 is close to some fixed constant $r>0$. In this framework one investigates convergence of the distributions of the stochastic processes, when the scaling parameter C_1 goes to infinity. The main technical difficulties lie in the behavior of the processes at the boundaries of the state space, when there are no blocked customers at level 1 or when there are no idle servers at level 2. As always with processes behaving locally as random walks, getting convergence results of scaled process in this context with two boundaries may be difficult. This situation has some similarities with the reflected random walks associated with classical queueing networks where the convergence results can be, sometimes, obtained by using a Skorokhod problem formulation. See Harrison and Reiman [HR81], Chen and Mandelbaum [CM91] or Section 9.4 of Robert [Rob03] for example. There is no such global Skorokhod problem formulation for our model. An additional difficulty is the dependence on the scaling parameter of the location of time intervals where blocking (or no-blocking) occurs eventually.

To handle this complicated setting, we introduce two auxiliary processes which are first separately investigated in Section 5.3, for each of them, only one of the boundary conditions is involved. A generalized Skorokhod problem formulation is used in both cases. The final Section 5.4 establishes the main convergence results. Stochastic calculus with Poisson processes, coupling arguments and the results obtained on auxiliary processes are the main ingredients of the proofs. See the proof of Proposition 5.6 for example.

5.2 The Stochastic Model

To analyze the stability properties of this network, it will be assumed that the capacities of the two levels of service are large, proportional to a scaling parameter N. Qualitative and quantitative properties of the system when N gets large will be obtained. In particular we will determine the conditions on the parameters for which the blocking probability is negligible or not. We begin with a brief reminder on Poisson processes and some notations used in this domain.

Notations for Poisson processes

Throughout the paper, for $\xi>0$, one denotes by $\mathcal{N}_\xi=(t_n)$ a Poisson point process on $\mathbb{R}_+$ with rate ξ and $(\mathcal{N}_{\xi,i})$ denotes a sequence of i.i.d. such Poisson processes. In the following, we will use at some occasions the following coupling of Poisson processes, for $0<\alpha\leqslant\beta$, one can construct a version of $\mathcal{N}_\alpha$ and $\mathcal{N}_\beta$ such that, for all $0\leqslant s\leqslant t$,

$$\mathcal{N}_\alpha([s,t]) \stackrel{\text{def.}}{=} \int_s^t \mathcal{N}_\alpha(\mathrm{d}s) \leqslant \mathcal{N}_\beta([s,t]).$$

This can be done in the following way. If $\mathcal{P}$ is a Poisson process on $\mathbb{R}_2^+$ whose intensity measure is Lebesgue on this space, then for $\xi\in\{\alpha,\beta\}$, the order relation will hold if we take

$$\mathcal{N}_\xi(\mathrm{d}t) = \mathcal{P}([0,\xi]\times\mathrm{d}t).$$

The notation $\overline{\mathcal{N}}_\xi=(t_n,B_n)$ is for a marked point Poisson process on $\mathbb{R}_+\times\{0,1\}$, where (t_n) is a Poisson process with rate ξ on $\mathbb{R}_+$ and (B_n) is an i.i.d. sequence of Bernoulli random variables with parameter p. If f is some positive Borelian function on $\mathbb{R}_+\times\{0,1\}$, we will use the (usual) notation

$$\int f(t,b),\overline{\mathcal{N}}_\xi(\mathrm{d}s,\mathrm{d}b) = \sum_{n\geqslant 1} f(t_n,B_n),$$

$(\overline{\mathcal{N}}_{\xi,i})$ denotes a sequence of such i.i.d. marked point Poisson processes. Concerning marked point Poisson processes see Kingman [Kin93] for example. They can be interpreted as follows in our case, if $\xi\in\{\mu_{01},\mu_{11},\mu_{02}\}$, $u\in\{0,1\}$ and the quantity $\overline{\mathcal{N}}_{\xi,i}(\mathrm{d}t,\{u\})$ is not 0, then a completion of a service occurs at time t, and if a new job enters the first level at this occasion, u is the class of this job. Clearly the point process $\mathcal{N}_\xi(\mathrm{d}t)$ has the same distribution as $\overline{\mathcal{N}}_\xi(\mathrm{d}t,\{0,1\})$.

Scaling

The capacities C_1 and C_2 of levels 1 and 2 depend on a scaling parameter N, $C_1=C_1^N=N$ and $C_2=C_2^N$ such that the convergence

$$\lim_{N\to+\infty} C_2^N/C_1^N = r \tag{5.3}$$

holds for some $r>0$.

The evolution of the state of this system can described by the stochastic process $(X^N(t)) := (Y_*^N(t),Y^N(t),Z^N(t))$ with, for $t\geqslant 0$,

- $(Y_*^N(t))$ being the number of class 0 jobs blocked at level 1 at time t,
- $(Y^N(t))$, the number of class 0 jobs being served at level 1,
- $(Z^N(t))$, the number of idle servers at level 2.

For $t\geqslant 0$, remark that at least one of the variables $Y_*^N(t)$ or $Z^N(t)$ is null. It is not difficult to see that $(X^N(t))$ is an irreducible Markov process on the state space

$$\mathcal{S}_N := \left\{x=(y_*,y,z)\in\mathbb{N}^3 : y+y_*\leqslant C_1^N, z\leqslant C_2^N, y_*\cdot z=0\right\}$$

It will be assumed that the sequence of initial states satisfies the relation

$$\lim_{N\to+\infty}\frac{1}{N}(Y_*^N(0),Y^N(0),Z^N(0))=x_0=(y_{*0},y_0,z_0)\in[0,1]^2\times[0,r]. \tag{5.4}$$

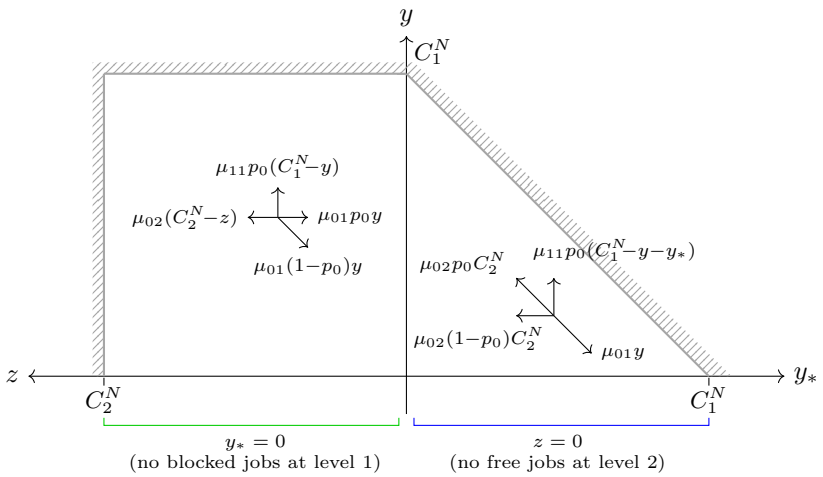


Figure 5.2 – A Representation of the Transitions Rates of $(X^N(t))$. The 3-dimensional process with $y^* \cdot z = 0$ is "unfolded" in two dimensions.

The vector x_0 will be referred to as the initial fluid state in the following. The transition rates are defined as follows, for $x{=}(y_*, y, z){\in}\mathcal{S}$,

$$x \mapsto \begin{cases} (y_*{+}1, y{-}1, 0) & \text{at rate} \quad \mu_{01} y \mathbb{1}_{\{z=0\}}, \\ (0, y{-}1, z{-}1) & " \quad \mu_{01} y(1{-}p) \mathbb{1}_{\{z>0\}}, \\ (0, y, z{-}1) & " \quad \mu_{01} yp \mathbb{1}_{\{z>0\}}, \\ (y_*, y{+}1, z) & " \quad \mu_{11} p(N{-}y_*{-}y), \\ (y_*{-}1, y, z) & " \quad (1{-}p)\mu_{02} C_2^N \mathbb{1}_{\{y_*>0\}}, \\ (y_*{-}1, y{+}1, z) & " \quad p\mu_{02} C_2^N \mathbb{1}_{\{y_*>0\}}, \\ (0, y, z{+}1) & " \quad \mu_{02}(C_2^N{-}z) \mathbb{1}_{\{y_*=0\}}. \end{cases} \tag{5.5}$$

Due to the constraints on the coordinates y_* and z of x (at least one of them is 0), the Markov process $(X^N(t))$ can be seen as a two-dimensional process as depicted in Figure 5.2.

Representation by Stochastic Differential Equations

From the transition rates (5.5), the process $(X^N(t))$ can also be seen as the unique solution of the following stochastic differential equations,

$$\mathrm{d}Y_*^N(t) = \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leqslant Y^N(t-), Z^N(t-)=0\}} \overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t, \{0,1\}) - \mathbb{1}_{\{Y_*^N(t-)>0\}} \sum_{i=1}^{C_2^N} \overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}t, \{0,1\}), \tag{5.6}$$

$$\begin{aligned} \mathrm{d}Y^N(t) = &- \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leqslant Y^N(t-)\}} \mathbb{1}_{\{Z^N(t-)=0\}} \overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t, \{0,1\}) \\ &- \mathbb{1}_{\{Z^N(t-)>0\}} \overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t, \{1\}) + \mathbb{1}_{\{Y_*^N(t-)>0\}} \sum_{i=1}^{C_2^N} \overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}t, \{0\}) \\ &+ \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leqslant N - Y_*^N(t-) - Y^N(t-)\}} \overline{\mathcal{N}}_{\mu_{11},i}(\mathrm{d}t, \{0\}), \end{aligned} \tag{5.7}$$

$$\mathrm{d}Z^N(t) = -\sum_{i=1}^{+\infty} \mathbb{1}_{\{i\leqslant Y^N(t-),Z^N(t-)>0\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{0,1\}) + \sum_{i=1}^{+\infty} \mathbb{1}_{\{i\leqslant C_2^N-Z^N(t-),Y_*^N(t-)=0\}}\overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}t,\{0,1\}), \quad (5.8)$$

starting from some fixed initial state. The notation $f(t-)$ stands for the left-limit of f at t.

Filtration

The σ-field $\mathcal{F}_t$ of the events up to time t is classically defined as the σ-field generated by the random variables

$$\overline{\mathcal{N}}_{\xi,i}([0,s]\times u), \text{ where } \xi\in\{\mu_{01},\mu_{11},\mu_{02}\}, s\in[0,t], u\in\{\{0\},\{1\}\} \text{ and } i\in\mathbb{N}.$$

With this definition the process $(Y_*^N(t), Y^N(t), Z^N(t))$ is clearly $(\mathcal{F}_t)$-adapted. The martingale properties mentioned in the following are understood to be with respect to this filtration.

Evolution equations

The rescaled process is denoted by

$$\left(\overline{X}^N(t)\right) \stackrel{\text{def.}}{=} \left(\overline{Y}_*^N(t),\overline{Y}^N(t),\overline{Z}^N(t)\right) := \frac{1}{N}\left(Y_*^N(t),Y^N(t),Z^N(t)\right), \quad (5.9)$$

the integration of the above SDEs and classical stochastic calculus give the relations

$$\overline{Y}_*^N(t) = \overline{Y}_*^N(0) + \mu_{01}\int_0^t \overline{Y}^N(s)\mathbb{1}_{\{\overline{Z}^N(s)=0\}}\,\mathrm{d}s - \mu_{02}\frac{C_2^N}{N}\int_0^t \mathbb{1}_{\{\overline{Y}_*^N(s)>0\}}\,\mathrm{d}s + \overline{M}_{Y_*}^N(t), \quad (5.10)$$

$$\overline{Y}^N(t) = \overline{Y}^N(0) - \mu_{01}\int_0^t \overline{Y}^N(s)\left(1-p\mathbb{1}_{\{\overline{Z}(s)>0\}}\right)\mathrm{d}s + p\mu_{02}\frac{C_2^N}{N}\int_0^t \mathbb{1}_{\{\overline{Y}_*^N(s)>0\}}\,\mathrm{d}s + p\mu_{11}\int_0^t (1-\overline{Y}_*^N(s)-\overline{Y}^N(s))\,\mathrm{d}s + \overline{M}_Y^N(t), \quad (5.11)$$

$$\overline{Z}^N(t) = \overline{Z}^N(0) - \mu_{01}\int_0^t \overline{Y}^N(s)\mathbb{1}_{\{\overline{Z}^N(s)>0\}}\,\mathrm{d}s + \mu_{02}\int_0^t \left(\frac{C_2^N}{N}-\overline{Z}^N(s)\right)\mathbb{1}_{\{\overline{Y}_*^N(s)=0\}}\,\mathrm{d}s + \overline{M}_Z^N(t), \quad (5.12)$$

where, for $V\in\{Y_*,Y,Z\}$, $(\overline{M}_V^N(t))$ is a martingale. We complete this section with a tightness result.

▶ **Proposition 5.1.** *The sequence of processes $(\overline{X}^N(t))$ defined by Relation (5.9) is tight and any of its limiting points is a continuous process.*

Proof. Since, for $t\geqslant 0$, one has $Y_*^N(t)+Y^N(t)\leqslant N$ and $Z^N(t)\leqslant C_2^N$, the variables $\overline{Y}_*^N(t)$, $\overline{Y}^N(t)$ and $\overline{Z}^N(t)$ are thus uniformly bounded. By using a similar procedure as in the proof of Theorem 6.13 page 159 of Robert [Rob03], one can show that the expected value of the previsible increasing process of the martingales $(\overline{M}_V^N(t))$, $V\in\{Y_*,Y,Z\}$, is of the order of $1/N$ and thus converges to 0. By Doob's Inequality, one gets that for any $\eta>0$ and $T>0$, the relation

$$\lim_{N\to 0}\mathbb{P}\left(\sup_{0\leqslant t\leqslant T}|\overline{M}_V^N(t)|\geqslant\eta\right)=0. \quad (5.13)$$

holds. Denote by $w_{f,T}$ the modulus of continuity of a function $(f(t))$ on $[0,T]$, i.e., for $\delta>0$

$$w_{f,T}(\delta) = \sup\left(|f(t)-f(s)| : 0\leqslant s\leqslant t\leqslant T, |t-s|\leqslant\delta\right).$$

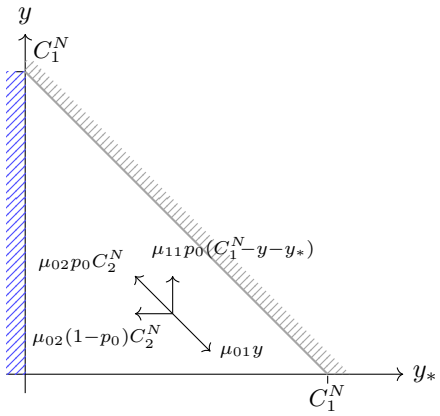


Figure 5.3 – The first auxiliary process, with saturation of level 2.

By using again that $(\overline{Y}_*^N(t))$, $(\overline{Y}^N(t))$ and $(\overline{Z}^N(t))$ are bounded and by Relation (5.13), Equations (5.10), (5.11) and (5.12) show that for any $\varepsilon>0$ and $\eta>0$, there exist $N_0\geqslant 1$ and $\delta_0>0$ such that if $N\geqslant N_0$ and $\delta<\delta_0$ then

$$\mathbb{P}\left(w_{V,T}(\delta)\geqslant\eta\right)\leqslant\varepsilon,\quad V\in\left\{\overline{Y}_*^N,\overline{Y}^N,\overline{Z}^N\right\}.$$

One concludes with Theorem 15.1 of Billingsley [Bil99]. ◀

5.3 Analysis of Auxiliary Processes

To study the asymptotic evolution of blocked customers, it is convenient to introduce two important stochastic processes. The first one describes the behavior of the system when the second level is permanently full, and the second one corresponds to the situation when there are no blocked class 0 customers at level 1.

5.3.1 A Process with Saturation of Level 2

The corresponding process is denoted by $(Y_{a*}^N(t),Y_a^N(t))$, it describes a system when level 2 is always saturated by class 0 jobs. The process $(Y_a^N(t))$ [resp. $(Y_{a*}^N(t))$] indicates the number of class 0 jobs [resp. blocked] at level 1. For this system blocked class 0 jobs are served at rate $\mu_{02}C_2^N$, otherwise the statistical assumptions are the same as before: see Figure 5.3.

This is a Markov process with transition rates defined by

$$(y_*,y)\mapsto\begin{cases}(y_*+1,y-1) & \text{at rate } \mu_{01}y,\\ (y_*-1,y) & " \quad (1-p)\mu_{02}C_2^N\mathbb{1}_{\{y_*>0\}},\\ (y_*-1,y+1) & " \quad p\mu_{02}C_2^N\mathbb{1}_{\{y_*>0\}},\\ (y_*,y+1) & " \quad p\mu_{11}(N-y_*-y).\end{cases}\tag{5.14}$$

The first transition is for a 0 job being blocked after its service at level 1. The second one corresponds to a 0 job leaving level 2 allowing a blocked 0 job to go to level 2 and a new 1 job is added at level 1. The third transition is similar except that a new 0 job enters level 1. The last transition corresponds to a 1 job leaving level 1 allowing a 0 job to enter level 1.

As long as $Y_*^N(t)>0$, this Markov process has the same transition rates as the process $(Y_*^N(t),Y^N(t))$, see Relation (5.5).

▶ **Proposition 5.2.** *If the initial condition of* $(Y_{a*}^N(t),Y_a^N(t))$ *is such that*

$$\lim_{N\to+\infty}\frac{1}{N}(Y_{a*}^N(0),Y_a^N(0))=(y_{a*}^0,y_a^0)\in[0,1]^2,\tag{5.15}$$

with $0\leqslant y_{a*}^0+y_a^0\leqslant 1$ *then, for the convergence in distribution, the relation*

$$\lim_{N\to+\infty}\frac{1}{N}(Y_{a*}^N(t),Y_a^N(t))=(y_{a*}(t),y_a(t))$$

holds, where $(y_{a*}(t), y_a(t))$ *is a couple of continuous functions such that*

$$y_{a*}(t) + y_a(t) \geqslant \overline{h}(t) := (y^0_{a*} + y^0_a)e^{-p\mu_{11}t} + \left(1 - \frac{(1-p)\mu_{02}r}{p\mu_{11}}\right)\left(1 - e^{-p\mu_{11}t}\right) \tag{5.16}$$

and $(y_{a*}(t), u(t))$ *is the unique solution of the following Skorokhod problem*

$$y_{a*}(t) = y^0_{a*} + \mu_{01}\int_0^t y_a(s)\,\mathrm{d}s - \mu_{02}rt + u(t) \tag{5.17}$$

where $(u(t))$ *is a non-decreasing continuous function such that* $u(0) = 0$ *and*

$$\int_0^{+\infty} y_{a*}(s)\ \mathrm{d}u(s) = 0.$$

Concerning the Skorokhod problem in dimension 1, see Skorokhod [Sko62], Chaleyat-Maurel and El Karoui [EKCM78]. The main trick is to express the couple $(y_{a*}(t), u(t))$ of Equation (5.17) as a regular functional of the free process

$$\left(y^0_{a*} + \mu_{01}\int_0^t y_a(s)\,\mathrm{d}s - \mu_{02}rt\right).$$

Note that, in our case, this free process depends on $(y_{a*}(t), y_a(t))$.

Proof. We will proceed as follows, first show that any limiting point $(y_{a*}(t), y_a(t))$ of $(Y^N_{a*}(t), Y^N_a(t))$ is such that $(y_{a*}(t))$ can be seen as the first coordinate of the solution of a Skorokhod problem associated with a free process. In a second step, we will show that the later process can be expressed as a regular functional of $(y_{a*}(t))$. One has then to use uniqueness results of Anderson and Orey [AO76] to conclude the proof.

From the transition rates (5.14), the process $(Y^N_{a*}(t), Y^N_a(t))$ can be seen as the solution of the stochastic differential equations (SDE)

$$\begin{aligned}
\mathrm{d}Y^N_{a*}(t) &= \sum_{i=1}^{+\infty} \mathbb{1}_{\{i\leqslant Y^N_a(t-)\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{0,1\}) - \mathbb{1}_{\{Y^N_{a*}(t-)>0, i\leqslant C^N_2\}}\overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}t,\{0,1\}),\\
\mathrm{d}Y^N_a(t) &= -\sum_{i=1}^{+\infty} \mathbb{1}_{\{i\leqslant Y^N_a(t-)\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{0,1\}) + \mathbb{1}_{\{Y^N_{a*}(t-)>0, i\leqslant C^N_2\}}\overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}t,\{0\})\\
&\quad + \sum_{i=1}^{+\infty} \mathbb{1}_{\{i\leqslant N - Y^N_{a*}(t-) - Y^N_{a*}(t-)\}}\overline{\mathcal{N}}_{\mu_{11},i}(\mathrm{d}t,\{0\}).
\end{aligned}$$

With the notation

$$\left(\overline{Y}^N_{a*}(t), \overline{Y}^N_a(t)\right) = \frac{1}{N}\left(Y^N_{a*}(t), Y^N_a(t)\right),$$

by integrating the above SDE, one gets the relations

$$\overline{Y}^N_{a*}(t) = \overline{Y}^N_{a*}(0) + \mu_{01}\int_0^t \overline{Y}^N_a(s)\,\mathrm{d}s - \mu_{02}\frac{C^N_2}{N}\int_0^t \mathbb{1}_{\{\overline{Y}^N_{a*}(s)>0\}}\,\mathrm{d}s + M^N_*(t) \tag{5.18}$$

$$\begin{aligned}
\overline{Y}^N_a(t) &= \overline{Y}^N_a(0) - \mu_{01}\int_0^t \overline{Y}^N_a(s)\,\mathrm{d}s + p\mu_{11}\int_0^t \left(1 - \overline{Y}^N_{a*}(s) - \overline{Y}^N_a(s)\right)\mathrm{d}s \\
&\quad + p\mu_{02}\frac{C^N_2}{N}\int_0^t \mathbb{1}_{\{\overline{Y}^N_{a*}(s)>0\}}\,\mathrm{d}s + M^N(t),
\end{aligned} \tag{5.19}$$

where $(M^N_*(t))$ and $(M^N(t))$ are local martingales. In the same way as in the proof of Proposition 5.1 of Section 5.2, one can prove that the sequence of processes $(\overline{Y}^N_{a*}(t), \overline{Y}^N_a(t))$ is tight and that any of its limiting points is a continuous process.

Let $(y_{a*}(t), y_a(t))$ be a limiting point, i.e., for some subsequence (N_k) the relation

$$\lim_{k\to+\infty}\left(\overline{Y}^{N_k}_{a*}(t), \overline{Y}^{N_k}_a(t)\right) = (y_{a*}(t), y_a(t))$$

holds for the convergence in distribution of processes. Denote

$$F_a^N(t) = \overline{Y}_{a*}^N(0) + \mu_{01}\int_0^t \overline{Y}_a^N(s)\,\mathrm{d}s - \mu_{02}\frac{C_2^N}{N}t + M_*^N(t). \tag{5.20}$$

Equation (5.18) can be written as

$$\overline{Y}_{a*}^N(t) = F_a^N(t) + \mu_{02}\frac{C_2^N}{N}\int_0^t \mathbb{1}_{\{\overline{Y}_{a*}^N(s)=0\}}\,\mathrm{d}s,$$

so that the couple

$$\left(\overline{Y}_{a*}^N(t), \mu_{02}\frac{C_2^N}{N}\int_0^t \mathbb{1}_{\{\overline{Y}_{a*}^N(s)=0\}}\,\mathrm{d}s\right)$$

is the solution of the Skorokhod problem associated with the free process $(F_a^N(t))$. See Skorokhod [Sko62] and Appendix D of Robert [Rob03] for a brief account.

For the convergence in distribution of processes, one has

$$\lim_{k\to+\infty}(F_a^{N_k}(t)) = (f_a(t)) := \left(y_{a*}^0 + \mu_{01}\int_0^t y_a(s)\,\mathrm{d}s - \mu_{02}rt\right), \tag{5.21}$$

since, as before, the martingales are vanishing as N gets large. From Proposition D.4 of the appendix of Robert [Rob03], one gets that $(y_{a*}(t))$ is the first coordinate of the solution of the Skorokhod problem associated with $(f_a(t))$ and $(y_{a*}(t))$ is differentiable almost everywhere for the Lebesgue measure on $\mathbb{R}_+$. In particular Relation (5.17) holds.

Since the free process $(f_a(t))$ depends on $(y_{a*}(t), y_a(t))$, there is no guarantee of the uniqueness of such a limit point $(y_{a*}(t), y_a(t))$. We now give a representation of $(f_a(t))$ in terms of $(y_{a*}(t))$. We proceed by getting rid of the process $(\overline{Y}_a^N(t))$ in the expression (5.20) of $(F_a^N(t))$. From Equations (5.18) and (5.19), we get the relation

$$\begin{aligned} p\overline{Y}_{a*}^N(t) + \overline{Y}_a^N(t) = {} & p\overline{Y}_{a*}^N(0) + \overline{Y}_a^N(0) + p\mu_{11}t - p\mu_{11}\int_0^t \overline{Y}_{a*}^N(s)\,\mathrm{d}s \\ & - ((1-p)\mu_{01} + p\mu_{11})\int_0^t \overline{Y}_a^N(s)\,\mathrm{d}s + pM_*^N(t) + M^N(t). \end{aligned}$$

By reordering the terms, one gets the relation

$$\begin{aligned} \overline{Y}_a^N(t) + \overline{\mu}\int_0^t \overline{Y}_a^N(s)\,\mathrm{d}s = {} & \left(p\overline{Y}_{a*}^N(0) + \overline{Y}_a^N(0)\right) + p\mu_{11}t - p\overline{Y}_{a*}^N(t) \\ & - p\mu_{11}\int_0^t \overline{Y}_{a*}^N(s)\,\mathrm{d}s + pM_*^N(t) + M^N(t), \end{aligned} \tag{5.22}$$

with $\overline{\mu} \overset{\text{def.}}{=} (1-p)\mu_{01} + p\mu_{11}$. Hence, by denoting

$$K^N(t) \overset{\text{def.}}{=} \frac{1}{\overline{\mu}}\left(p\overline{Y}_{a*}^N(0) + \overline{Y}_a^N(0)\right)\left(e^{\overline{\mu}t} - 1\right) + \frac{p\mu_{11}}{\overline{\mu}^2}\left(1 + (\overline{\mu}t - 1)e^{\overline{\mu}t}\right),$$

and $(k(t))$ its limit,

$$k(t) \overset{\text{def.}}{=} \frac{1}{\overline{\mu}}\left(py_{a*}^0 + y_a^0\right)\left(e^{\overline{\mu}t} - 1\right) + \frac{p\mu_{11}}{\overline{\mu}^2}\left(1 + (\overline{\mu}t - 1)e^{\overline{\mu}t}\right),$$

from Equation (5.22), trite calculations give the representation

$$\begin{aligned} \int_0^t \overline{Y}_a^N(s)\,\mathrm{d}s = {} & K^N(t)e^{-\overline{\mu}t} - p\int_0^t \left(\overline{Y}_{a*}^N(s) + \mu_{11}\int_0^s \overline{Y}_{a*}^N(u)\,\mathrm{d}u\right)e^{-\overline{\mu}(t-s)}\,\mathrm{d}s \\ & + \int_0^t \left(pM_*^N(s) + M^N(s)\right)e^{-\overline{\mu}(t-s)}\,\mathrm{d}s. \end{aligned}$$

Therefore, we can write the free process $(F_a^N(t))$ as

$$F_a^N(t) = G\left(\overline{Y}_{a*}^N\right)(t) + \overline{Y}_{a*}^N(0) + \mu_{01}K^N(t)e^{-\overline{\mu}t} - \mu_{02}\frac{C_2^N}{N}t + \mu_{01}\int_0^t \left(pM_*^N(s) + M^N(s)\right)e^{-\overline{\mu}(t-s)}\,\mathrm{d}s + M_*^N(t), \quad (5.23)$$

where $G(\cdot)$ is a functional on Borelian functions $(x(t))$ defined by

$$G(x)(t) = -p\mu_{01}\int_0^t \left(x(s) + \mu_{11}\int_0^s x(u)\,\mathrm{d}u\right)e^{-\overline{\mu}(t-s)}\,\mathrm{d}s.$$

This gives us an alternative representation of $(f_a(t))$ as

$$f_a(t) = \overline{G}(y_{a*})(t) \overset{\text{def.}}{=} G(y_{a*})(t) + y_{a*}^0 + \mu_{01}k(t)e^{-\overline{\mu}t} - \mu_{02}rt. \quad (5.24)$$

We have shown that $(y_{a*}(t))$ is the first coordinate of $(y_{a*}(t), u(t))$, the solution of a generalized Skorokhod problem associated to the functional $\overline{G}$,

$$y_{a*}(t) = \overline{G}(y_{a*})(t) + u(t) \text{ and } \int_0^{+\infty} y_{a*}(s)\,\mathrm{d}u(s) = 0,$$

with the usual assumptions on $(y_{a*}(t))$ and $(u(t))$. See Anderson and Orey [AO76]. For any Borelian functions $(a(t))$ and $(b(t))$ on $\mathbb{R}_+$, it is not difficult to check that

$$\|\overline{G}(a) - \overline{G}(b)\|_{\infty,t} \overset{\text{def.}}{=} \sup_{0\leqslant s\leqslant t}\|\overline{G}(a)(s) - \overline{G}(b)(s)\| \leqslant C_t\int_0^t \|a-b\|_{\infty,s}\,\mathrm{d}s,$$

with $C_t = p\mu_{01}(1+\mu_{11}t)$. Anderson and Orey [AO76] show that such $(y_{a*}(t))$ is unique. The convergence in distribution follows:

$$\lim_{N\to+\infty}\left(\overline{Y}_{a*}^N(t), \mu_{02}\frac{C_2^N}{N}\int_0^t \mathbb{1}_{\{\overline{Y}_{a*}^N(s)=0\}}\,\mathrm{d}s\right) = (y_{a*}(t), u(t)).$$

Consequently, Relations (5.18) and (5.19) give the relations

$$\begin{cases} y_{a*}(t) = y_{a*}^0 + \mu_{01}\displaystyle\int_0^t y_a(s)\,\mathrm{d}s - \mu_{02}rt + u(t), \\ y_a(t) = y_a^0 - \mu_{01}\displaystyle\int_0^t y_a(s)\,\mathrm{d}s + p\mu_{11}\int_0^t (1 - y_{a*}(s) - y_a(s))\,\mathrm{d}s \\ \qquad\qquad + p\mu_{02}rt - pu(t). \end{cases} \quad (5.25)$$

By using Relations (5.21) and (5.24), one deduce the uniqueness of $(y_a(t))$and, therefore, the convergence in distribution of the sequence of processes $(Y_{a*}^N(t), Y_a^N(t))$.

We now prove that the limit $(y_{a*}(t), y_a(t))$ satisfies necessarily $y_{a*}(t)+y_a(t)\geqslant\overline{h}(t)$ for all t, where $\overline{h}$ is the solution of

$$\overline{h}(t) = (y_{a*}^0 + y_a^0) - (1-p)\mu_{02}rt + p\mu_{11}\int_0^t \left(1 - \overline{h}(s)\right)\mathrm{d}s,$$

that is,

$$\overline{h}(t) = (y_{a*}^0 + y_a^0)e^{-p\mu_{11}t} + \left(1 - \frac{(1-p)\mu_{02}r}{p\mu_{11}}\right)(1 - e^{-p\mu_{11}t}).$$

First note that, for any N, the process $\overline{Y}_a^N$ is bounded above by 1, so that F_a^N is Lipschitz. Hence, again by Proposition D.4 of the appendix of Robert [Rob03], u is also Lipschitz, and thus continuous.

From Relations (5.25), one gets that the identity

$$(y_{a*}(t) + y_a(t)) = (y_{a*}^0 + y_a^0) - (1-p)\mu_{02}rt + p\mu_{11}\int_0^t (1 - (y_{a*}(s) + y_a(s)))\,\mathrm{d}s + (1-p)u(t)$$

holds, so that the difference $y_{a*}+y_a-\overline{h}$ satisfies the system

$$x(t)+p\mu_{11}\int_0^t x(s)\,\mathrm{d}s=(1-p)u(t),\ \text{with } x(0)=0$$

Any continuous solution $(x(t))$ of this system is non-negative. Suppose that there exists $t_1>0$ such that $x(t_1)<0$. Then, by continuity of $(x(t))$, there exists $t_0<t_1$ such that $x(t_0)=0$ and $x(t)<0$ for $t_0<t<t_1$. But

$$x(t_1)-x(t_0)=x(t_1)=(1-p)(u(t_1)-u(t_0))-p\mu_{11}\int_{t_0}^{t_1}x(s)\,\mathrm{d}s$$

and the right-hand-side of the equality is positive because $(u(t))$ is non-decreasing and $(x(t))$ is negative on this interval, which is a contradiction. Relation (5.16) is established. The proposition is proved. ◀

▶ **Proposition 5.3.** *Under Condition (5.1), there exists $t_0\geqslant 0$, independent of the initial state (5.15), such that for $t\geqslant t_0$, the functions $(y_{a*}(t))$ and $(y_a(t))$ introduced in Proposition 5.2 are differentiable at t and*

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}y_{a*}(t)=\mu_{01}y_a(t)-\mu_{02}r,\\[2ex] \dfrac{\mathrm{d}}{\mathrm{d}t}y_a(t)=-(\mu_{01}+p\mu_{11})y_a(t)-p\mu_{11}y_{a*}(t)+p(\mu_{02}r+\mu_{11})\end{cases}\tag{5.26}$$

Any solution $(y_{a*}(t),y_a(t))$ of the differential system (5.26) converges to

$$\left(1-\frac{\mu_{02}r}{\mu_{01}}\left(\frac{(1-p)\mu_{01}}{p\mu_{11}}+1\right),\frac{\mu_{02}r}{\mu_{01}}\right).\tag{5.27}$$

Proof. The above proposition shows that

$$\liminf_{t\to+\infty}y_{a*}(t)+y_a(t)\geqslant 1-\frac{(1-p)\mu_{02}r}{p\mu_{11}}.$$

Let

$$\varepsilon_0=\left(1-\frac{(1-p)\mu_{02}r}{p\mu_{11}}\right)-\frac{\mu_{02}r}{\mu_{01}},$$

then $\varepsilon_0>0$ by Condition (5.1). Let t_0 be such that if $t\geqslant t_0$ then

$$y_{a*}(t)+y_a(t)\geqslant 1-\frac{(1-p)\mu_{02}r}{p\mu_{11}}-\frac{\varepsilon_0}{2}.\tag{5.28}$$

The classical representation of the solution of one-dimensional Skorokhod problem, see Relation (D.1) p.376 of Robert [Rob03], gives the identity

$$y_{a*}(t)=\left(y_{a*}^0+\mu_{01}\int_0^t y_a(u)\,\mathrm{d}u-\mu_{02}rt\right)\vee\sup_{0\leqslant s\leqslant t}\left(\mu_{01}\int_s^t y_a(u)\,\mathrm{d}u-\mu_{02}r(t-s)\right).$$

If $t_1>t_0$ is such that $y_{a*}(t_1)=0$, then by continuity of $(y_a(t))$, one gets the relation

$$\mu_{01}y_a(t_1)\leqslant\mu_{02}r,$$

and, by using Relation (5.28),

$$1-\frac{(1-p)\mu_{02}r}{p\mu_{11}}-\frac{\varepsilon_0}{2}\leqslant y_{a*}(t_1)+y_a(t_1)=y_a(t_1)\leqslant\frac{\mu_{02}r}{\mu_{01}}$$

which leads to a contradiction. One concludes that $t\mapsto y_{a*}(t)$ is positive for $t>t_0$ and consequently that the measure $\mathrm{d}u(t)$ vanishes on the interval $(t_0,+\infty)$. The proposition is proved. ◀

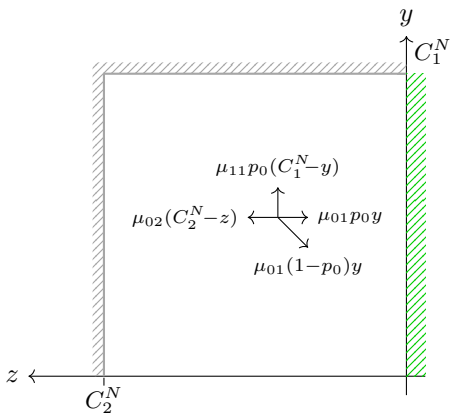


Figure 5.4 – The second auxiliary process, without blocked jobs.

5.3.2 A System without blocked jobs

A second auxiliary process is introduced, it is denoted by $(Y_b^N(t), Z_b^N(t))$. It describes the situation when there are no blocked jobs at level one: if a class 0 job finishes while level two is saturated, i.e., $Z_b^N(t)=0$, then it leaves the system, instead of being blocked. If there are free servers at level two, the process behaves in the same way as the main process under study, see Figure 5.4. The process $(Y_b^N(t), Z_b^N(t))$ is a Markov process, with the following transition rates:

$$(y,z) \mapsto \begin{cases} (y-1,0) & \text{at rate} \quad (1-p)\mu_{01} y \mathbb{1}_{\{z=0\}}, \\ (y-1,z-1) & " \qquad (1-p)\mu_{01} y \mathbb{1}_{\{z>0\}}, \\ (y,z-1) & " \qquad p\mu_{01} y \mathbb{1}_{\{z>0\}}, \\ (y+1,z) & " \qquad p\mu_{11}(N-y), \\ (y,z+1) & " \qquad \mu_{02}(C_2^N - z). \end{cases} \tag{5.29}$$

Note that, when $z>0$, this Markov process has the same transition rates as the process $(X^N(t))$ (see (5.5)).

▶ **Proposition 5.4.** *If the initial condition of $(Y_b^N(t), Z_b^N(t))$ is such that*

$$\lim_{N\to+\infty} \frac{1}{N}(Y_b^N(0), Z_b^N(0)) = (y_b^0, z_b^0) \in [0,1]\times[0,r], \tag{5.30}$$

then, for the convergence in distribution, the relation

$$\lim_{N\to+\infty} \frac{1}{N}(Y_b^N(t), Z_b^N(t)) = (y_b(t), z_b(t))$$

holds, where $(y_b(t))$ is given by

$$y_b(t) = y_b^0 e^{-(p\mu_{11}+(1-p)\mu_{01})t} + \frac{p\mu_{11}}{p\mu_{11}+(1-p)\mu_{01}} \left(1-e^{-(p\mu_{11}+(1-p)\mu_{01})t}\right) \tag{5.31}$$

and $(z_b(t))$ is the unique solution of the Skorokhod problem

$$z_b(t) = z_b^0 + \mu_{02} r t - \mu_{02}\int_0^t z_b(s)\,\mathrm{d}s - \mu_{01}\int_0^t y_b(s)\,\mathrm{d}s + u(t) \tag{5.32}$$

where $(u(t))$ is a non-decreasing continuous function such that $u(0)=0$ and

$$\int_0^{+\infty} z_b(s)\,\mathrm{d}u(s) = 0\,.$$

As before, the free process associated to the Skorokhod problem is

$$\left(z_b^0 + \mu_{02} r t - \mu_{02}\int_0^t z_b(s)\,\mathrm{d}s - \mu_{01}\int_0^t y_b(s)\,\mathrm{d}s\right),$$

it is also a functional of $(z_b(t))$, see the proof of Proposition 5.2.

Proof. From the transition rates (5.29) and as in the proof of Proposition 5.2, if

$$\left(\overline{Y}_b^N(t), \overline{Z}_b^N(t)\right) := \frac{1}{N}\left(Y_b^N(t), Z_b^N(t)\right),$$

then one gets the evolution equations

$$\overline{Y}_b^N(t) = \overline{Y}_b^N(0) - \mu_{01}(1-p)\int_0^t \overline{Y}_b^N(s)\,\mathrm{d}s + p\mu_{11}\int_0^t (1-\overline{Y}_b^N(s))\,\mathrm{d}s + M_Y^N(t),$$

$$\overline{Z}_b^N(t) = \overline{Z}_b^N(0) + \mu_{02}\frac{C_2^N}{N}t - \mu_{02}\int_0^t \overline{Z}_b^N(s)\,\mathrm{d}s - \mu_{01}\int_0^t \overline{Y}_b^N(s)\,\mathrm{d}s + M_Z^N(t) + R_Z^N(t),$$

with

$$R_Z^N(t) = \mu_{01}\int_0^t \overline{Y}_b^N(s)\mathbb{1}_{\{\overline{Z}_b^N(s)=0\}}\,\mathrm{d}s,$$

and $(M_Z^N(t))$ and $(M_Y^N(t))$ are local martingales. It is easily seen that these two martingales vanish when N gets large and hence that, with the criterion of the modulus of continuity, the sequence of processes $(\overline{Y}_b^N(t))$ is tight. Furthermore, any limiting point $(y_b(t))$ satisfies the integral equation

$$y_b(t) = y_b^0 - (p\mu_{11} + (1-p)\mu_{01})\int_0^t y_b(s)\,\mathrm{d}s + p\mu_{11}t,$$

so that Relation (5.31) holds.

Clearly, $(\overline{Z}_b^N(t), R_Z^N(t))$ is the solution of a generalized Skorokhod process associated with a free process which depends itself on $\overline{Z}_b^N(t))$. One concludes in the same way as in the proof of Proposition 5.2. ◀

We now prove that, under Condition (5.2), the reflecting part of the Skorokhod problem of the last proposition vanishes for t large enough.

▶ **Proposition 5.5.** *Under Condition (5.2), there exists $t_0 \geqslant 0$, independent of the initial state (5.30), such that for $t \geqslant t_0$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}z_b(t) = \mu_{02}r - \mu_{02}z_b(t) - \mu_{01}y_b(t). \tag{5.33}$$

Furthermore,

$$\lim_{t\to+\infty}(y_b(t), z_b(t)) = \left(\frac{p\mu_{11}}{p\mu_{11}+(1-p)\mu_{01}}, r - \frac{p\mu_{01}\mu_{11}}{\mu_{02}(p\mu_{11}+(1-p)\mu_{01})}\right).$$

Proof. By Proposition 5.4, $(y_a(t))$ converges to $\overline{y}_b$ as t gets large. By Condition (5.2), the relation $\overline{y}_b < \mu_{02}r/\mu_{01}$ holds. Consequently, there exists t_0 such that the inequality

$$\mu_{01}y_b(t) < \mu_{02}r$$

holds for $t \geqslant t_0$.

Theorem D.1 of Robert [Rob03] gives the relation

$$z_b(t) = \left(z_b^0 + \mu_{02}rt - \mu_{02}\int_0^t z_b(s)\,\mathrm{d}s - \mu_{01}\int_0^t y_b(s)\,\mathrm{d}s\right) \vee \sup_{0\leqslant s\leqslant t}\left(\mu_{02}r(t-s) - \mu_{02}\int_s^t z_b(u)\,\mathrm{d}u - \mu_{01}\int_s^t y_b(u)\,\mathrm{d}u\right).$$

Suppose that $z_b(t_1)=0$ for some $t_1 > t_0$, then in particular, for $0 \leqslant s \leqslant t_1$,

$$\mu_{02}r(t_1-s) - \mu_{02}\int_s^{t_1} z_b(u)\,\mathrm{d}u - \mu_{01}\int_s^{t_1} y_b(u)\,\mathrm{d}u \leqslant 0,$$

by continuity of $(z_b(t))$ and $(y_b(t))$, this gives that

$$\mu_{02}r - \mu_{01}y_b(t_1) = \mu_{02}r - \mu_{02}z_b(t_1) - \mu_{01}y_b(t_1) \leqslant 0,$$

contradiction. This implies that $z_b(t) > 0$ holds for $t \geqslant t_0$ and, consequently, the measure $\mathrm{d}u(t)$ vanishes on $[t_0, +\infty)$. The proposition is proved. ◀

5.4 Asymptotic Study of the Blocking Phenomenon

The goal of this section is of showing that if the ratio $r{\sim}C_2^N/C_1^N$ of the capacities of the two levels of our system is less than the quantity

$$r_c \stackrel{\text{def.}}{=} \frac{p}{\mu_{02}} \Big/ \left(\frac{p}{\mu_{01}}+\frac{1-p}{\mu_{11}}\right),$$

then there exists some fixed instant t_0 such that, with a probability converging to 1 as N gets large, the number of blocked servers at level 1 is of the order of N on any finite time interval after t_0. Otherwise, if $r{>}r_c$, then there exists $t_0{>}0$ such that the number of blocked servers is 0 with high probability on any finite time interval after t_0. These results are respectively proved in Sections 5.4.1 and 5.4.2. The proofs use the technical tools introduced in the last section and additional probabilistic arguments.

5.4.1 The Overloaded Regime

In this section, we will assume that Condition (5.1) holds, i.e., that $r{<}r_c$. Recall that $(Y_*^N(t), Y^N(t), Z^N(t))$ describes the state of our system. The following proposition is a technical result that shows that type 0 jobs occupy at least a fixed fraction of the first level after some time.

▶ **Proposition 5.6.** *Under Condition (5.1), for $\varepsilon{>}0$, there exists $t_0{>}0$ such that, for any initial fluid state (5.4), and $T{\geqslant}t_0$, then*

$$\lim_{N\to\infty} \mathbb{P}\left(\inf_{t\in[t_0,T]} \left(Y_*^N(t){+}Y^N(t)\right) \geqslant N(\bar{y}{-}\varepsilon)\right) = 1,$$

with

$$\bar{y} := \frac{p\mu_{11}}{p\mu_{11}{+}(1{-}p)\mu_{01}}.$$

Proof. Define, for $t \geqslant 0$,

$$H^N(t) = N\bar{y}{-}Y_*^N(t){-}Y^N(t),$$

we want to show that $H^N(t)$ is, with high probability, below $N\varepsilon$ on any finite time interval after some finite fixed instant.

The stochastic differential equations (5.6) and (5.7) give the relation

$$\mathrm{d}H^N(t) = -\sum_{i=1}^{+\infty} \mathbb{1}_{\{i\leqslant N(1-\bar{y})+H^N(t-)\}} \overline{\mathcal{N}}_{\mu_{11},i}(\mathrm{d}t,\{0\}) + \mathrm{d}D_0^N(t) \tag{5.34}$$

where $(D_0^N(t))$ is the process associated with the positive jumps of this SDE,

$$D_0^N(t) \stackrel{\text{def.}}{=} \sum_{i=1}^{+\infty} \int_0^t \mathbb{1}_{\{Y_*^N(s)>0, i\leqslant C_2^N\}} \overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}s,\{1\}) + \mathbb{1}_{\{i\leqslant Y^N(s), Z^N(s)>0\}} \overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}s,\{1\}).$$

In the following we will use repeatedly, without mentioning it explicitly, the coupling of Poisson process with ordered rates described at the beginning of Section 5.2.

If, for some $T{>}0$, one has that, for all $0{\leqslant}t{\leqslant}T$, $H^N(t){\geqslant}\lceil\varepsilon N\rceil$, then, on this time interval, $Y^N(t){\leqslant}N(\bar{y}{-}\varepsilon)$ and, $D_0^N(t){-}D_0^N(s){\leqslant}D_1^N(t){-}D_1^N(s)$ for all $0{\leqslant}s{\leqslant}t$, with

$$D_1^N(t) \stackrel{\text{def.}}{=} \int_0^t \sum_{i=1}^{+\infty} \mathbb{1}_{\{Y_*^N(s)>0, i\leqslant C_2^N\}} \overline{\mathcal{N}}_{\mu_{02},i}(\mathrm{d}s,\{1\}) \quad + \mathbb{1}_{\{i\leqslant N(\bar{y}-\varepsilon), Z^N(s)>0\}} \overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}s,\{1\}).$$

By using classical results on superposition and thinning of independent Poisson processes, see Kingman [Kin93] for example, we have $(D_1^N(t)) \stackrel{\text{dist.}}{=} (D_2^N(t))$, with

$$D_2^N(t) \stackrel{\text{def.}}{=} \int_0^t \mathbb{1}_{\{Y_*^N(s)>0\}} \mathcal{N}_{\mu_{02}(1-p)C_2^N}(\mathrm{d}s) + \mathbb{1}_{\{Z^N(s)>0\}} \mathcal{N}_{\mu_{01}(1-p)N(\bar{y}-\varepsilon)}(\mathrm{d}s).$$

It is easily seen that Condition (5.1) is equivalent to the relation $r\mu_{02}<\bar{y}\mu_{01}$, hence for $\eta>0$ there exists some N_0 such that

$$\mu_{02}C_2^N < \mu_{01}(\bar{y}+\eta)N, \quad \text{for } N\geqslant N_0.$$

By using this inequality and the relation $\{Y_*^N(t-)>0\}\subset\{Z^N(t-)=0\}$, one has

$$\int_0^t \mathbb{1}_{\{Y_*^N(s)>0\}}\mathcal{N}_{\mu_{02}(1-p)C_2^N}(\mathrm{d}s) \leqslant \int_0^t \mathbb{1}_{\{Z^N(s)=0\}}\mathcal{N}_{\mu_{01}(1-p)N(\bar{y}+\eta)}(\mathrm{d}s).$$

Moreover, since

$$\int_0^t \mathbb{1}_{\{Z^N(s)>0\}}\mathcal{N}_{\mu_{01}(1-p)N(\bar{y}-\varepsilon)}(\mathrm{d}s) \leqslant \int_0^t \mathbb{1}_{\{Z^N(s)>0\}}\mathcal{N}_{\mu_{01}(1-p)N(\bar{y}+\eta)}(\mathrm{d}s),$$

hence $D_2^N(t)-D_2^N(s)\leqslant D_3^N(t)-D_3^N(s)$, for $0\leqslant s\leqslant t$, with

$$D_3^N(t) \stackrel{\text{def.}}{=} \int_0^t \mathbb{1}_{\{Z^N(s)=0\}}\mathcal{N}_{\mu_{01}(1-p)N(\bar{y}+\eta)}(\mathrm{d}s) + \mathbb{1}_{\{Z^N(s)>0\}}\mathcal{N}_{\mu_{01}(1-p)N(\bar{y}+\eta),2}(\mathrm{d}s),$$

where $\mathcal{N}_{\mu_{01}(1-p)N(\bar{y}+\eta)}$ and $\mathcal{N}_{\mu_{01}(1-p)N(\varepsilon+\eta),2}$ are two independent Poisson point processes. The integer valued process $(D_3^N(t))$ has jumps of size 1 and it is easily checked that

$$\left(D_3^N(t)-\mu_{01}N(1-p)(\bar{y}+\eta)t\right)$$

is a martingale. From Watanabe's Theorem, see Watanabe [Wat64], one gets that $(D_3^N(t))$ is a Poisson process on $\mathbb{R}_+$ with rate $\lambda_H N$ with $\lambda_H \stackrel{\text{def.}}{=} \mu_{01}(1-p)(\bar{y}+\eta)$.

If, for all $0\leqslant t\leqslant T$, $H^N(t)\geqslant\lceil\varepsilon N\rceil$ then

$$D_0^N(t)-D_0^N(s) \leqslant D_3^N(t)-D_3^N(s) \text{ for } 0\leqslant s\leqslant t, \tag{5.35}$$

where $(D_3^N(t))$ is a Poisson process with rate $\lambda_H N$.

Assume that $H^N(0)\geqslant\lceil\varepsilon N\rceil$ and define the process $(\overline{H}^N(t))$, with the initial condition $\overline{H}^N(0)=H^N(0)-\lceil\varepsilon N\rceil$ and such that

$$\mathrm{d}\overline{H}^N(t) = \mathcal{N}_{\lambda_H N}(\mathrm{d}t) - \mathbb{1}_{\{\overline{H}^N(t-)>0\}}\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N(1-\bar{y})+\varepsilon N\}}\overline{\mathcal{N}}_{\mu_{11},i}(\mathrm{d}t,\{0\}) \tag{5.36}$$

holds for $t\geqslant 0$. Clearly, $(\overline{H}^N(t))$ has the same distribution as $(L_H(Nt))$, where $(L_H(t))$ is the Markov process associated with an $M/M/1$ queue whose arrival and services rates are respectively λ_H and $\mu_H \stackrel{\text{def.}}{=} p\mu_{11}(1-\bar{y}+\varepsilon)$. Note that, by definition of $\bar{y}$,

$$\lambda_H-\mu_H = \mu_{01}(1-p)\eta - p\mu_{11}\varepsilon.$$

One can choose $\eta>0$ so that $\lambda_H-\mu_H<-p\mu_{11}\varepsilon/2$, hence $(L_H(t))$ is an ergodic Markov process in this case.

Let $T_\varepsilon^N=\inf\{t\geqslant 0 : H^N(t)<\varepsilon N\}$, in particular $H^N(t)\geqslant\lceil\varepsilon N\rceil$ for $0\leqslant t\leqslant T_\varepsilon^N$. From Relations (5.34), (5.35) and (5.36), we get therefore the inequality

$$H^N(t)-\lceil\varepsilon N\rceil \leqslant \overline{H}^N(t), \tag{5.37}$$

for $t<T_\varepsilon^N$, hence $T_\varepsilon^N\leqslant\tau_L^N$, with

$$\tau_L^N = \inf\left\{t>0 : \overline{H}^N(t)=0\right\} = \inf\{t>0 : L_H(Nt)=0\}.$$

Since $\overline{H}^N(0)\leqslant H^N(0)\leqslant N$, Proposition 5.16 of Robert [Rob03] gives that, for any t_0 such that $t_0>1/(\mu_H-\lambda_H)$ then

$$\lim_{N\to+\infty}\mathbb{P}\left(T_\varepsilon^N\leqslant t_0\right) \geqslant \lim_{N\to+\infty}\mathbb{P}\left(\tau_L^N\leqslant t_0\right) = 1.$$

By using the strong Markov property, up to a change of time origin, one can assume that $H^N(0)\leqslant\lceil\varepsilon N\rceil$. If $(\overline{H}^N(t))$ is defined as before with the initial condition $\overline{H}^N(0)=0$, then it is not difficult to show that Relation (5.37) holds for all $t\geqslant 0$. For the excursions of $(H^N(t))$ below $\lceil\varepsilon N\rceil$ this is clear and for the excursions above this level it has just been proved. In particular, for any $T>0$,

$$\mathbb{P}\left(\inf_{0\leqslant t\leqslant T} Y_*^N(t)+Y^N(t)\leqslant N(\bar{y}-2\varepsilon)\right)=\mathbb{P}\left(\sup_{0\leqslant t\leqslant T}\overline{H}^N(t)\geqslant 2\varepsilon N\right)$$
$$\leqslant \mathbb{P}_0\left(\sup_{0\leqslant t\leqslant NT} L_H(t)\geqslant\lfloor\varepsilon N\rfloor\right),$$

and the last quantity is the probability that the hitting time of $\lfloor\varepsilon N\rfloor$ by an $M/M/1$ queue starting from 0 is less that NT. Proposition 5.11 of Robert [Rob03] shows that this hitting time is of the order of $(\mu_H/\lambda_H)^{\lfloor\varepsilon N\rfloor}$ and therefore exponentially large in N (recall that $\lambda_H<\mu_H$). In particular, the last term of the right-hand-side of the above relation is converging to 0 as N gets large. The proposition is proved. ◀

The above proof relies on the comparisons of point processes associated to the counting processes $(D_i(t))$, $i\in\{0,1,2,3\}$. We have, for example, that the point process associated to $(D_0(t))$ is "smaller" that the one associated to $(D_1(t))$: $D_0^N(t)-D_0^N(s)\leqslant D_1^N(t)-D_1^N(s)$ for all $0\leqslant s\leqslant t$. In the following, for convenience, we will use the notation $D_0(\mathrm{d}t)\leqslant D_1(\mathrm{d}t)$.

▶ **Proposition 5.7.** *Under Condition (5.1), for any $\varepsilon>0$ small enough, there exists $t_1>0$ such that, for any initial fluid state (5.4), and for any $T\geqslant t_1$, the relation*

$$\lim_{N\to\infty}\mathbb{P}\left(\inf_{t\in[t_1,T]} Y_*^N(t)\geqslant\varepsilon N\right)=1.$$

holds.

Proof. From Equations (5.6) and (5.8), we get that

$$\mathrm{d}\left(Y_*^N-Z^N\right)(t)=\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant Y^N(t-)\}}\mathcal{N}_{\mu_{01},i}(\mathrm{d}t)-\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant C_2^N-Z^N(t-)\}}\mathcal{N}_{\mu_{02},i}(\mathrm{d}t).$$

By Proposition 5.6, there exists t_0 be such that, for $T\geqslant t_0$, for the events

$$\mathcal{A}_N\overset{\text{def.}}{=}\left\{\inf_{t\in[t_0,T]}\left(Y_*^N(t)+Y^N(t)\right)\geqslant N(\bar{y}-\varepsilon)\right\},$$

the sequence $(\mathbb{P}(\mathcal{A}_N))$ converges to 1.

Suppose that, for some time $t\in(t_0,T)$, $Y_*^N(t)<2\varepsilon N$, then, on the event $\mathcal{A}_N$, $Y^N(t)\geqslant N(\bar{y}-3\varepsilon)$ and consequently

$$\mathrm{d}\left(Y_*^N-Z^N\right)(t)\geqslant\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N(\bar{y}-3\varepsilon)\}}\mathcal{N}_{\mu_{01},i}(\mathrm{d}t)-\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant C_2^N\}}\mathcal{N}_{\mu_{02},i}(\mathrm{d}t)\tag{5.38}$$
$$\overset{\text{dist.}}{=}\mathcal{N}_{\mu_N}(\mathrm{d}t)-\mathcal{N}_{\lambda_N}(\mathrm{d}t).$$

with $\mu_N=\lfloor\mu_{01}N(\bar{y}-3\varepsilon)\rfloor$ and $\lambda_N=\mu_{02}C_2^N$. As noted before, Condition (5.1) is equivalent to the relation $\bar{y}\mu_{01}-r\mu_{02}>0$. Since

$$\lim_{N\to+\infty}\frac{\mu_N-\lambda_N}{N}=(\mu_{01}\bar{y}-\mu_{02}r)-\mu_{01}3\varepsilon,$$

one can find $0<\lambda<\mu$ and ε sufficiently small, such that for N sufficiently large $\lambda_N\geqslant\lambda N$ and $\mu_N\leqslant\mu N$ hold. If

$$T_N=\inf\{t>0:Y_*^N(t)\geqslant 2\varepsilon N\},$$

then since $\left(Y_*^N-Z^N\right)(0)\geqslant -C_2^N$, Relation (5.38) gives the existence of $t_1>0$ such that

$$\lim_{N\to+\infty}\mathbb{P}(T_N\leqslant t_1)=1.$$

We now assume that $\left(Y_*^N - Z^N\right)(0) = \lceil 2\varepsilon N \rceil$. By taking $T > t_1$, using Relation (5.38) and the estimates for λ_N and μ_N, we get that, on the event $\mathcal{A}_N$,

$$\left(Y_*^N - Z^N\right)(t) \geqslant \lceil 2\varepsilon N \rceil - X(Nt)$$

holds for all $t>0$, where $(X(t))$ is an $M/M/1$ queue with arrival [resp. service] rate λ [resp. μ] starting at 0. With the same argument as the end of the previous proof, since $\lambda<\mu$, we have that, for any $T>0$,

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leqslant t\leqslant NT} X(s) \geqslant \varepsilon N \right) = 0,$$

consequently

$$\lim_{N\to+\infty} \mathbb{P}\left(\left\{ \sup_{0\leqslant t\leqslant T} \left(Y_*^N - Z^N\right)(t) \geqslant \lceil \varepsilon N \rceil \right\} \bigcap \mathcal{A}_N \right) = 1.$$

We conclude the proof by using the fact that $Y_*^N(t)>0$ implies that $Z^N(t)=0$. ◀

The following corollary gives a more precise statement concerning the asymptotic behavior of the Z-component of the state vector. It is a simple consequence of the fact that $Y_*^N(t)>0$ implies $Z^N(t)=0$.

▶ **Corollary 5.8.** *Under Condition (5.1), there exists $t_1>0$ such that, for all $T > t_1$,*

$$\lim_{N\to+\infty} \mathbb{P}\left(Z^N(t) = 0, \forall t \in [t_1, T]\right) = 1$$

holds for any initial fluid state (5.4).

▶ **Theorem 5.9** (Saturated regime)**.** *Under Condition (5.1), there exists $t_1>0$ such that, for any initial fluid state (5.4), any limiting point $(y_\infty^*(t), y_\infty(t), z_\infty(t))$ of the sequence $(\overline{X}^N(t))$ defined by Relation (5.9) satisfies the following relations, for all $t\geqslant t_1$, $z_\infty(t)=0$ and the differential equations*

$$\frac{\mathrm{d}y_\infty^*}{\mathrm{d}t}(t) = \mu_{01} y_\infty(t) - \mu_{02} r,$$
$$\frac{\mathrm{d}y_\infty}{\mathrm{d}t}(t) = -\mu_{01} y_\infty(t) + p(\mu_{02} r + \mu_{11}(1 - y_\infty^*(t) - y_\infty(t))).$$

hold. Furthermore

$$\lim_{t\to+\infty} (y_\infty^*(t), y_\infty(t)) = \left(1 - \frac{\mu_{02} r}{\mu_{01}} \left(\frac{(1-p)\mu_{01}}{p\mu_{11}} + 1 \right), \frac{\mu_{02} r}{\mu_{01}}\right).$$

Proof. By Proposition 5.7, for some ε_0 sufficiently small, there exists t_1 such that, for any $T>t_1$, the event

$$\mathcal{E}_N \stackrel{\text{def.}}{=} \left\{ Y_*^N(t) \geqslant \varepsilon_0 N : \forall t \in [t_1, T] \right\}$$

has a probability arbitrarily close to 1 as N gets large. Consider the process $(Y_{a*}^N(t), Y_a^N(t))$ defined in Section 5.3.1 with initial state $(Y_*^N(t_1), Y^N(t_1))$, then by checking Q-matrix of both processes, it is easily seen that, on the event $\mathcal{E}_N$, the relation

$$\left((Y_*^N(t), Y^N(t)), t_1 \leqslant t \leqslant T\right) \stackrel{\text{dist.}}{=} \left((Y_{a*}^N(t), Y_a^N(t)), t_1 \leqslant t \leqslant T\right)$$

holds. By using the fact that the sequence of random variables

$$\left(\frac{1}{N} (Y_*^N(t_1), Y^N(t_1)) \right) \in [0,1]$$

is tight, one has only to use Proposition 5.2 to conclude the proof of the theorem. ◀

5.4.2 The Underloaded Regime

In this section, it will be assumed that Condition (5.2) holds.

▶ **Proposition 5.10.** *Under Condition (5.2), there exists $\eta_0>0$ and $t_1>0$ such that, for any initial fluid state (5.4) and for $T>t_1$,*

$$\lim_{N\to\infty}\mathbb{P}\left(\sup_{t\in[t_1,T]}\left(Y_*^N(t)+Y^N(t)\right)\leqslant N\left(\underline{y}-\eta_0\right)\right)=1,$$

with $\underline{y}=r\mu_{02}/\mu_{01}$.

Note that $\underline{y}<1$ by Condition (5.2).

Proof. By using the SDEs (5.6) and (5.7), we get that

$$\begin{aligned}\mathrm{d}\left(Y_*^N+Y^N\right)(t)=&\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N-Y_*^N(t-)-Y^N(t-)\}}\overline{\mathcal{N}}_{\mu_{11},i}(\mathrm{d}t,\{0\})\\&-\mathbb{1}_{\{Y_*^N(t-)>0\}}\overline{\mathcal{N}}_{\mu_{02}C_2^N}(\mathrm{d}t,\{1\})-\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant Y^N(t-),Z^N(t-)>0\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{1\})\end{aligned}\tag{5.39}$$

holds. The strategy of the proof is of deriving an upper bound for the process $(Y_*^N(t)+Y^N(t))$, as before we will work on the differential terms of the above relation.

We choose $\eta>0$ sufficiently small so that for N large enough the relation

$$\frac{\mu_{02}}{\mu_{01}}C_2^N>\left\lfloor N\left(r\frac{\mu_{02}}{\mu_{01}}-\eta\right)\right\rfloor$$

holds. Under this condition one has, with a convenient coupling of Poisson processes,

$$\overline{\mathcal{N}}_{\mu_{02}C_2^N}(\mathrm{d}t,\mathrm{d}u)\geqslant\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N(\mu_{02}r/\mu_{01}-\eta)\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\mathrm{d}u).\tag{5.40}$$

The relation $Y_*^N(t-)>0$ implies $Z^N(t-)=0$, consequently, we get the inequality

$$\mathbb{1}_{\{Y_*^N(t-)>0\}}\overline{\mathcal{N}}_{\mu_{02}C_2^N}(\mathrm{d}t,\{1\})\geqslant\mathbb{1}_{\{Z^N(t-)=0\}}\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N(\mu_{02}r/\mu_{01}-\eta)\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{1\}).$$

If the relation $Y^N(t-)+Y_*^N(t-)\geqslant N(\mu_{02}r/\mu_{01}-\eta)$ holds, then

$$\begin{aligned}\mathbb{1}_{\{Z^N(t-)>0\}}\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant Y^N(t-)\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{1\})&\\\geqslant\mathbb{1}_{\{Z^N(t-)>0\}}\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N(\mu_{02}r/\mu_{01}-\eta)\}}\overline{\mathcal{N}}_{\mu_{01},i}(\mathrm{d}t,\{1\}),&\end{aligned}\tag{5.41}$$

since $Y_*^N(t-)=0$ if $Z^N(t-)>0$.

By plugging Relations (5.40) and (5.41) into the SDE (5.39), we get that

$$\mathrm{d}\left(Y_*^N+Y^N\right)(t)\leqslant\mathcal{N}_{\lambda_N}(\mathrm{d}t)-\mathcal{N}_{\mu_N}(\mathrm{d}t)\tag{5.42}$$

holds on the event $Y^N(t-)+Y_*^N(t-)\geqslant N(\mu_{02}r/\mu_{01}-\eta)$, with

$$(\lambda_N,\mu_N)\overset{\text{def.}}{=}\left(\left\lfloor p\mu_{11}N\left(1-r\frac{\mu_{02}}{\mu_{01}}+\eta\right)\right\rfloor,\left\lfloor(1-p)\mu_{01}N\left(r\frac{\mu_{02}}{\mu_{01}}-\eta\right)\right\rfloor\right).$$

By Condition (5.2) we can take $\eta=\eta_0>0$ to be such that

$$2\eta_0<r\frac{\mu_{02}}{\mu_{01}}-\frac{p\mu_{11}}{p\mu_{11}+(1-p)\mu_{01}},$$

In this case if (λ,μ) is the limit of the sequence $((\lambda_N,\mu_N)/N)$, then $\lambda<\mu$. There exist $0<\lambda_0<\mu_0$, such that, for N sufficiently large, the relations $\lambda_N\leqslant\lambda_0 N$ and $\mu_N\geqslant\mu_0 N$ hold. Let $(X(t))$ be the (ergodic) $M/M/1$ queue with input [resp. service] rate given by λ_0 [resp. μ_0] and $X(0)=N$, then Equation (5.42) gives the coupling relation $(Y_*^N+Y^N)(t)\leqslant X(Nt)$ for t less than the hitting time of $N(\mu_{02}r/\mu_{01}-2\eta_0)$. Consequently, by ergodicity, there exists some $t_1\geqslant 0$ such that this hitting time is, with high probability, less than t_1N. Now, by taking the initial conditions $(Y_*^N+Y^N)(0)=X(0)=\lceil N(\mu_{02}r/\mu_{01}-2\eta_0)\rceil$, with the same argument as in the proof of Proposition 5.7, one gets that, for $T>0$, the process $((Y_*^N+Y^N)(t))$ remains below $N(\mu_{02}r/\mu_{01}-\eta_0)$ on the time interval $[0,T]$ with high probability. The proposition is proved. ◀

The following result is the analogue of Proposition 5.7 for the underloaded regime.

▶ **Proposition 5.11.** *Under Condition (5.2), for any $\varepsilon>0$ small enough, there exists a time $t_1\geqslant 0$ such that, for any initial fluid state (5.4) and for $T>t_1$,*

$$\lim_{N\to\infty}\mathbb{P}\left(\inf_{t\in[t_1,T]}(Z^N(t)-Y_*^N(t))\geqslant\varepsilon N\right)=1.$$

Proof. Since the proof follows the same lines as in the proof of Proposition 5.7, we sketch the main technical arguments. From the last proposition, one can chose η_0 and $t_1\geqslant 0$ such that the event

$$\mathcal{B}_N\overset{\text{def.}}{=}\left\{\sup_{t\in[t_1,T]}(Y_*^N(t)+Y^N(t))\leqslant N\left(r\frac{\mu_{02}}{\mu_{01}}-\eta_0\right)\right\}$$

has a probability converging to 1 when N gets large.

The SDEs (5.7) and (5.8) give the relation

$$\mathrm{d}\left(Z^N-Y_*^N\right)(t)=\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant C_2^N-Z^N(t^-)\}}\mathcal{N}_{\mu_{02},i}(\mathrm{d}t)-\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant Y^N(t-)\}}\mathcal{N}_{\mu_{01},i}(\mathrm{d}t),$$

by using again that $Z^N(t-)$ is null if $Y_*^N(t-)$ is positive.

One takes $\eta_0<\mu_{02}\varepsilon/(4\mu_{01})$ then, on the event $\mathcal{B}_N$, if $Z^N(t)\leqslant\varepsilon N$,

$$\mathrm{d}\left(Z^N-Y_*^N\right)(t)\geqslant\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant C_2^N-\varepsilon N\}}\mathcal{N}_{\mu_{02},i}(\mathrm{d}t)-\sum_{i=1}^{+\infty}\mathbb{1}_{\{i\leqslant N\mu_{02}(r-\varepsilon/2)/\mu_{01}\}}\mathcal{N}_{\mu_{01},i}(\mathrm{d}t).$$

Hence, the process $(Z^N-Y_*^N(t))$ can be compared with a (scaled) ergodic $M/M/1$ queue with arrival rate $\lfloor C_2^N-\varepsilon N\rfloor\mu_{02}$ and service rate $\lfloor N(r-\varepsilon/2)\mu_{02}/\mu_{01}\rfloor\mu_{01}$. We conclude the proof in the same way as in the proof of Proposition 5.6. ◀

▶ **Corollary 5.12.** *Under Condition (5.2), there exists $t_1>0$ such that, for all $T>t_1$,*

$$\lim_{N\to+\infty}\mathbb{P}\left(Y_*^N(t)=0,\forall t\in[t_1,T]\right)=1$$

holds for any initial fluid state (5.4).

Proof. The proof follows from the mutual exclusivity of the events $\{Y_*^N(t)>0\}$ and $\{Z^N(t)>0\}$ and from Proposition 5.11. ◀

We can now state the main result for the underloaded regime.

▶ **Theorem 5.13** (Underloaded Regime)**.** *Under Condition (5.1), there exists $t_1>0$ such that, for any initial fluid state (5.4), any limiting point $(y_\infty^*(t),y_\infty(t),z_\infty(t))$ of the sequence $(\overline{X}^N(t))$ defined by Relation (5.9) satisfies the following relations, for all $t\geqslant t_1$, $y_\infty^*(t)=0$ and the differential equations*

$$\frac{\mathrm{d}y_\infty}{\mathrm{d}t}(t)=-(p\mu_{11}+(1-p)\mu_{01})y_\infty(t)+p\mu_{11}$$
$$\frac{\mathrm{d}z_\infty}{\mathrm{d}t}(t)=-\mu_{02}z_\infty(t)-\mu_{01}y_\infty(t)+\mu_{02}r$$

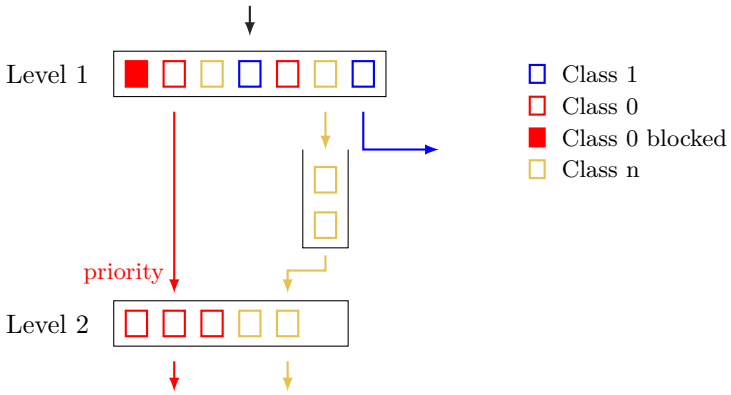


Figure 5.5 – A queueing network with a third class of calls

hold. Furthermore,

$$\lim_{t\to+\infty}(y_\infty(t), z_\infty(t)) = \left(\frac{p\mu_{11}}{p\mu_{11}+(1-p)\mu_{01}}, r-\frac{p\mu_{01}\mu_{11}}{\mu_{02}(p\mu_{11}+(1-p)\mu_{01})}\right).$$

Proof. In the same way as in the proof of Theorem 5.9, a coupling between the processes $(Y_b^N(t), Z_b^N(t))$ defined in Section 5.3.2 and the process $(Y_N(t), Z_N(t))$ can be constructed so that the convergence results of Proposition 5.5 can be used. ◀

5.5 Concluding remarks

In this chapter, we considered a bilevel call center with only one class of calls directed to level 2. In future work, we would like to extend these results to the case when there is a third class of calls, also needing a treatment by level 2. These calls, however, shall not be kept in line by level 1 operator, but are directed to an independent queue if level 2 operators are busy. This is the queueing model depicted in Figure 5.5.

This would allow a perfect correspondence between the results of this chapter, and the results of Chapter 3 and Chapter 4.

Part II

Case Study: simulations and analysis of an Emergency Call Center

Chapter 6

Case study: simulations and analysis of an emergency call center

In this chapter, we focus on our emergency call center application, which motivated the theoretical approach developed in the first part of this thesis. The key features of this emergency call center is a two-level architecture, with three-way conferences between operators of both levels and callers, and the priority treatment of some of the calls at the second level.

Our numerical simulations take into account many features that were not included in the simplified model described in Section 2.5.1, some of which could not be modeled with our formalism, and some others were dropped out for the sake of simplicity.

Among these additional features, the most significant one is certainly the heterogeneous nature of level 2. In the planned architecture of the Parisian emergency call center, level 2 is composed of firemen and policemen, who do not answer the same calls: level 1 operators identify which of the incoming calls should be dispatched to police and firemen. Moreover, in early versions of the system, policemen of different administrative areas do not use the same information systems and, hence, also form separated groups. This can still be modeled by the Petri nets with free choice and priorities described in Chapters 3 and 4, at the cost of complicating the models: see Section 6.2.1.

The other important characteristics of the emergency call center that we take into account in our simulation are the impatience of callers, who can give up their calls, the latencies between two successive call treatments by an operator, or the typical statistical distributions of the call service times. See our analysis in Section 6.1. Furthermore, we evaluate performance by measuring, not only call throughputs, but also classical quality of service criteria, such as waiting times or abandonment rates.

Note that call abandonment cannot be modeled by our Petri nets with free choice and priority. However, this does not invalidate these models, as we were interested in measuring capacities (maximal throughputs). In a first order (and rough) approximation, abandonment rates are low as long as the system does not work at full capacity. In our simulations, we ran more detailed analyses in order to take into account impatience of callers and its consequence in terms of sizing. See Section 6.3.1.

An important preliminary remark is the following: in this chapter, as in the rest of this dissertation, we do not intend to provide an optimal operator sizing, on a daily and hourly basis, in order to achieve some performance targets. This would require to resort to elaborate optimization tools, that we did not develop in this thesis. On this topic, see for example [ACG+10, ATLB16]. Our goal here is rather to further investigate and comment on the specific features and behaviors of this two-level architecture, especially in situations of congestion.

Most of the results presented in this chapter have been delivered to Préfecture de police de Paris and Brigade de sapeurs-pompiers de Paris in a more practitioner-focused presentation. In the present chapter, we give a general overview of this applied work, leaving apart some details of the real application, because of confidentiality imperatives.

6.1 Data analysis of an emergency call center

The new bilevel architecture of the emergency call center was set in year 2016. Our analysis focuses on the situation before this event, because our objective was then to foresee the behavior of the new call center. We had the opportunity to study several weeks of data from the '18-112' emergency call center of BSPP, and the '17' emergency call center of Préfecture de police de Paris, at the individual call level.

Some processing of the data was required. For example, before entering the queue or being assigned to an operator, all incoming calls are welcomed by an initial voice message, whose duration essentially depends on the emergency call number. We filtered out of our data set all calls that did abandon before the end of this message, and did not count this initial delay in our waiting times. It is a well-known fact by people in charge of the call center that the duration of this initial voice message is a key parameter to control the volume (and urgency) of incoming calls. We nevertheless did not analyze this phenomenon in this dissertation.

We refer to Brown *et al.* [BGM+05] for a detailed statistical analysis of a call center. We found our statistical analyses to yield similar results to the ones described by these authors. This was helped by the fact that the number of calls handled by our emergency call center represents the same order of magnitude than the number of calls treated by operators in the call center they analyzed. Statistical analyses of call centers is an active field of research. Other relevant references are avramidis2005modeling,matteson2011forecasting,ibrahim2016modeling,

In the following, we describe a few key characteristics of the two call centers we analyzed. Because of confidentiality issues, most numerical values are obfuscated.

Call arrival The call arrival rate varies continuously in a day, with amplitude between the hour with minimal rate and the hour with maximal rate going up to a factor 7. The volume of incoming calls can also vary up to 60% between two different days of the week. In contrast with the queueing models used in Chapter 5, the statistical distribution is significantly different from a Poisson distribution, even for small time periods. In Figure 6.1, we select 4-hour periods during weekdays, cumulate our observations on a few months, and compare the statistical distribution with the Poisson distribution with the same mean. It appears to have a larger dispersion than its mean, as in call centers observed by Jongbloed and Koole [JK01] or Avramidis *et al.* [ADL04] (with more accurate methods). We refer to Ibrahim *et al.* [IYLS16] for a recent survey on call arrivals analysis.

Service times Brown *et al.* [BGM+05] observed that the statistical distribution of service times in their case study fitted a lognormal distribution. This also contrasts with the standard queueing model, in which the service times are supposed to be exponential. In our case study, we observed that the service times were a mixture of two different lognormal distribution, one with very short service times (less than 10 seconds) and one with longer service time. See Figure 6.2. The weight of this second category of calls is much larger. According to our discussions with call center practitioners, the first category could correspond to error calls with nobody on line. This is a current phenomenon for emergency call centers, because an emergency call number is likely to be dialed by random moves of a phone in a pocket. We point out, however, that a

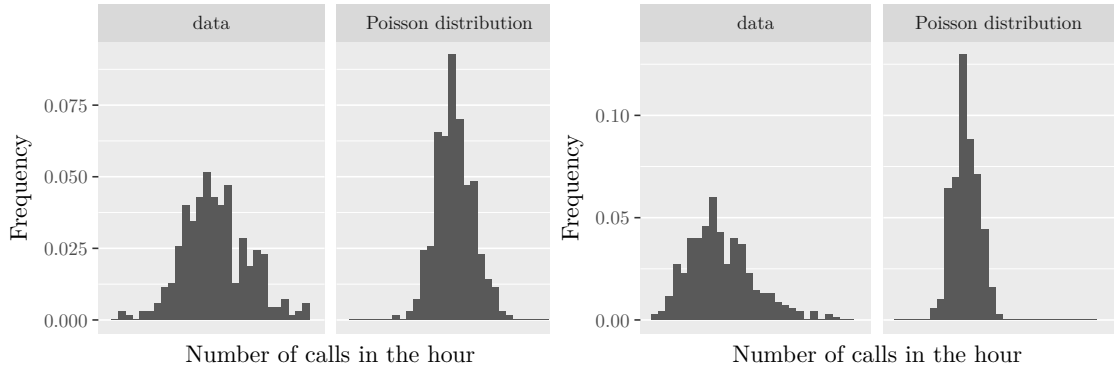


Figure 6.1 – Distribution of the number of calls received in one hour, counted on a few months during a given 4-hour period of the day, during regular weekdays, vs distribution of a data set drawn from a Poisson distribution with the same mean value. Right and left correspond to different moments of the day. Note that part of the variance of the data distribution could also be explained by other factors (*e.g.*, holidays, or days with special events).

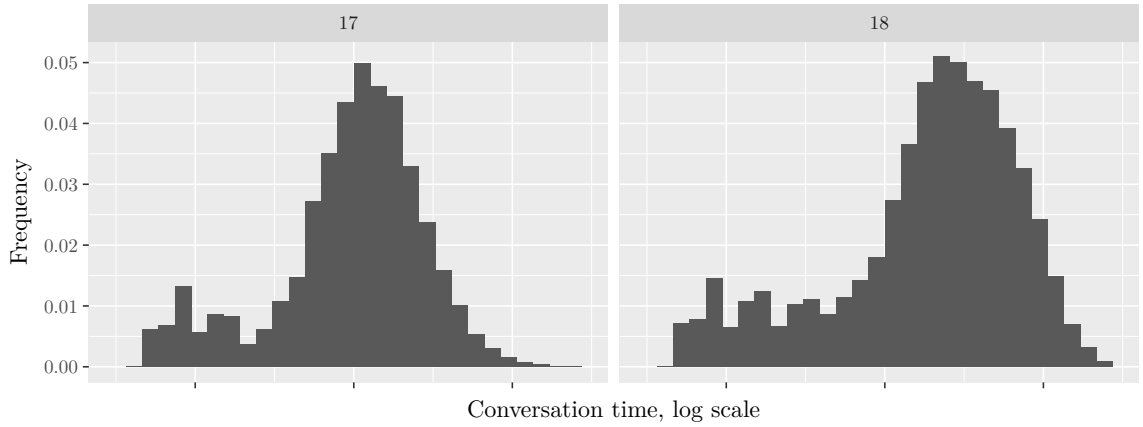


Figure 6.2 – Conversation times in log scale for calls to '17' and '18'. The graphs suggest that this statistical distribution corresponds the weighted sum of two lognormal distributions, one for short calls, the other one for long calls.

lognormal distribution modeling is a rough model that fails to account for more accurate results from call center data analysis, as the servers heterogeneity, or the interdependency of service times: see [ILST16] and references therein.

Service times vary depending on the hour of the day. Another interesting phenomenon we observed is that, for one of the two call centers we analyzed, there was a negative correlation between the number of calls received in an hour and the mean service time in this hour, while in the other call center, the two variables seemed to be independent. See Figure 6.3.

This suggests that, in the first call center, operators adapt their response to the affluence of calls. In the second call center, this is not the case. Two explanations were possible: either the type of service was different and the operators could not adapt their service time, or the operators of the first call center were better informed and alerted when several calls were queueing, or when the maximal waiting time in the queue went beyond some threshold.

The analysis of the new call center, in which all operators have the same information on calls queueing, and on their waiting times, shall allow to identify the correct explanation.

Call abandonment In trying to estimate the patience of callers, that is, their willingness to wait for service, one needs to take into account that, the longer a caller accepts to wait for service, the larger the probability is that he or she is served before abandoning. Consequently, the patience of callers cannot be measured by the distribution of waiting times, or by the

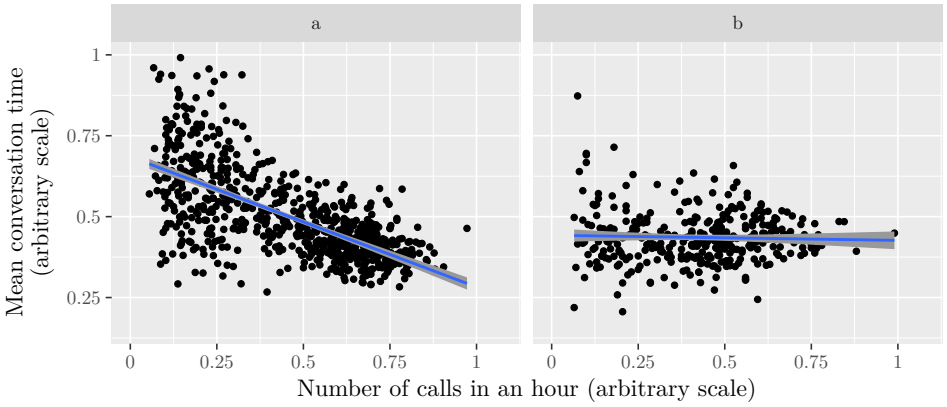


Figure 6.3 – Mean conversation time as a function of the number of calls in the hour, and linear regression: each point represent an hour, taken in a given month. Plots (a) and (b) correspond to our two different call centers.

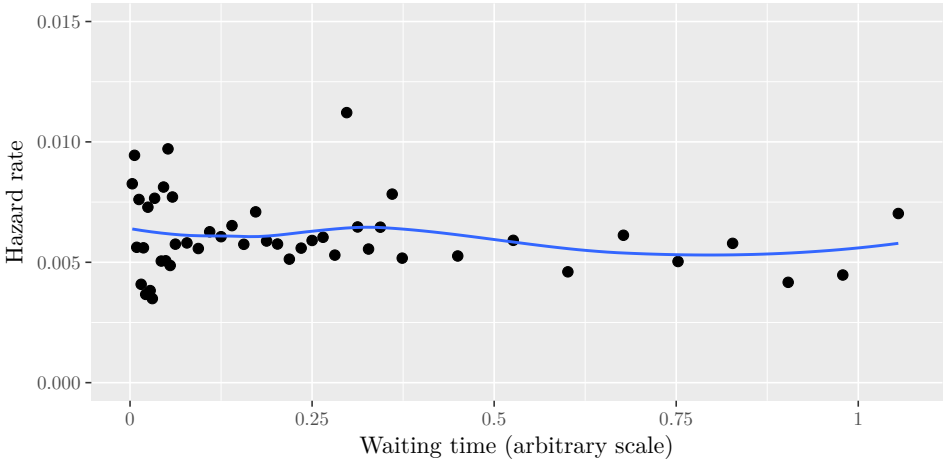


Figure 6.4 – Hazard rate as a function of waiting time. After the first 0.1 units, the hazard rate approaches a constant function. The outlier point at a waiting time of 0.3 units does not correspond to a large rate of abandonments at this time, but to a specific procedure of the call center which leads to loosing track of the call.

distribution of waiting times of those abandoning. In fact, for a given caller, one observes only the minimum of these two quantities, his or her patience and the time before he or she is served. However, by plotting the hazard rate function for this system (following the method in [BGM+05], first introduced by Palm [Pal53] for service systems), corresponding to the fraction of calls giving up in the next time interval among calls that are still in the queue, one can observe a rather constant function: see Figure 6.4. This suggests that an exponential distribution of callers' patience may be a reasonable approximation.

Inter-call delays A major challenge to improve simulations was to model inter-call delays, that is, the non-reducible periods of time during which an operator is not available, at the end of a call. The available data allows us to compute the periods during which an operator is not in communication with a caller, which comprises, but cannot be reduced to this inter-call delay. Indeed, these periods correspond to various situations, such as:

- After-call work (ACW): dispatching, communications with emergency services, with managers and co-workers.
- Wait: the operator is idle but no call is dispatched to him or her.
- Breaks: some are necessary, other can be delayed or adjusted depending on the frequency of call arrivals.
- End of service or beginning of service.

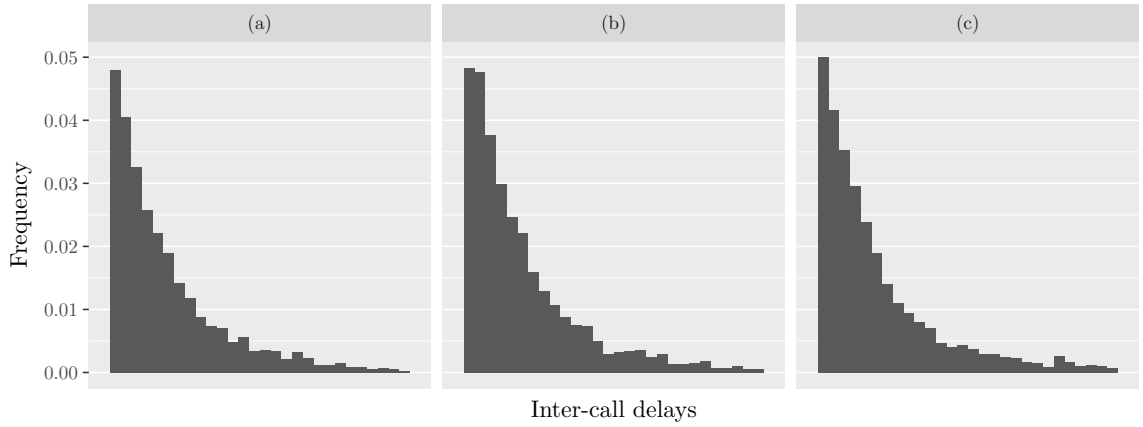


Figure 6.5 – Delays between two successive call handlings, for three different operator stations (arbitrary scale): the distribution is remarkably smooth, with no break between frequent small durations and unfrequent long durations.

In simulations, one would like to model only after-call work and minimal break periods, considering that the call center operators are in alert, so that the capacity of the call center is maximal (of course periods of waiting are unavoidable, and resort as outputs of the simulations). However, this information is not directly available. The distribution of inter-call delays observed in our data is given in Figure 6.5 for one of the call centers (for the other call center, we did not have information on which operator station answered the call). Its main characteristics is the continuity from short durations to long durations, so that the different categories of inter-call delays do not seem to be distinguishable. Still, we observed, as one can expect, that there is a negative correlation between the mean inter-call delay during a period and the number of incoming calls in this period.

Retrials We did not analyze *retrials* (situations in which a caller cannot reach an operator and calls again), in particular because call numbers were not available for privacy reasons. However, we point out the analysis of Aguir *et al.* [AAKD08], that these retrials can have a significant impact on the call center behavior in situations with large waiting times, which are of course rare in emergency call center, but may happen in situations of crisis.

6.2 The impact of having separated groups of operators at level 2

Recall that one of the key characteristics of the call center of our case study is that operators of level 2 are split between 2 groups (or even more), which do not handle the same kind of calls. We first describe the behavior of this organization by the analytical methods of Part I.

6.2.1 Analysis of a model with two separated groups at level 2

The Petri net in Figure 6.6 models a call center with two groups of operators of level 2. In comparison with the Petri net of Section 2.1 (see Figure 2.7), the itinerary of extremely urgent calls and urgent calls is duplicated: in addition to the level of urgency of a call, the operator of level 1 also determines if the call should be addressed to police or firemen. This yields four different tracks for calls directed to level 2, instead of two tracks before. In contrast, advice calls, which are handled at level 1, are not labeled as “police” or “firemen”.

This is a Petri net with free choice and priority, as the ones considered in Chapters 3 and 4. The asymptotic throughputs of the dynamics can be expressed as a piecewise linear function of the parameters of the system, following the methods of Section 3.3.

Without entering into the details of the computation, we exhibit here the asymptotic throughputs formulæ. The rates $\mu^f_{\text{ext}}, \mu^p_{\text{ext}}, \mu^f_{\text{ur}}, \mu^p_{\text{ur}}, \mu_{\text{adv}}$, with $\mu^f_{\text{ext}} + \mu^p_{\text{ext}} + \mu^f_{\text{ur}} + \mu^p_{\text{ur}} + \mu_{\text{adv}} = 1$,

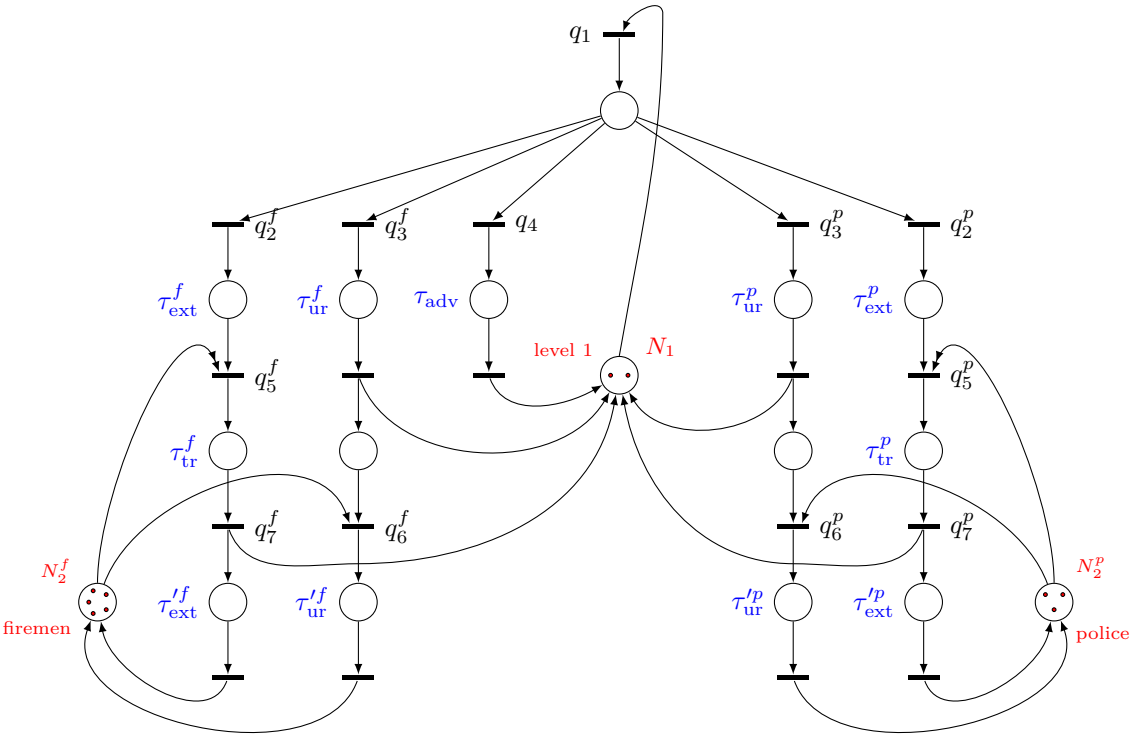


Figure 6.6 – Petri net modeling of a call center with two groups at level 2.

correspond to the fractions of calls classified in one of the different categories: f stands for 'firemen' and p stands for 'police'. The constant $\bar{\tau}$ characterizes here the average sojourn time at level 1, $\bar{\tau} := \mu_{\text{ext}}^f \tau_{\text{ext}}^f + \mu_{\text{ext}}^p \tau_{\text{ext}}^p + \mu_{\text{ur}}^f \tau_{\text{ur}}^f + \mu_{\text{ur}}^p \tau_{\text{ur}}^p + \mu_{\text{adv}} \tau_{\text{adv}}$.

As in the application of Part I (see Section 2.1), transitions q_5^f and q_5^p characterize the extremely urgent calls throughput at level 2 (firemen or police, resp.), and transitions q_6^f and q_6^p characterize the urgent calls throughput at level 2.

We define the constants r_1^f, r_1^p, r_2^f, r_2^p in a similar way to the constants r_1 and r_2 of Section 3.4, that is:

$$r_1^f := \frac{\mu_{\text{ext}}^f (\tau_{\text{tr}}^f + \tau_{\text{ext}}'^f)}{\bar{\tau}} \qquad r_1^p := \frac{\mu_{\text{ext}}^p (\tau_{\text{tr}}^p + \tau_{\text{ext}}'^p)}{\bar{\tau}}$$

$$r_2^f := \frac{\mu_{\text{ext}}^f (\tau_{\text{tr}}^f + \tau_{\text{ext}}'^f) + \mu_{\text{ur}}^f \tau_{\text{ur}}'^f}{\bar{\tau}} \qquad r_2^p := \frac{\mu_{\text{ext}}^p (\tau_{\text{tr}}^p + \tau_{\text{ext}}'^p) + \mu_{\text{ur}}^p \tau_{\text{ur}}'^p}{\bar{\tau}}$$

The stationary throughputs of the system are the following:

- In any situation, $\rho_5^f = \mu_{\text{ext}}^f \rho_1$ and $\rho_5^p = \mu_{\text{ext}}^p \rho_1$.
- If $N_2^f / N_1 \leqslant r_1^f$ and $N_2^f / N_2^p \leqslant r_1^f / r_1^p$, then

$$\rho_1 = N_2^f / \mu_{\text{ext}}^f \tau_{\text{tr}}^f$$

$$\rho_6^f = 0$$

$$\rho_6^p = \begin{cases} \dfrac{N_2^p}{\tau_{\text{ur}}'^p} - N_2^f \dfrac{\mu_{\text{ext}}^p \tau_{\text{ext}}^p}{\mu_{\text{ext}}^f \tau_{\text{ext}}^f \tau_{\text{ur}}'^p} & \text{if } N_2^p / N_2^f \leqslant r_2^p / r_1^f \\ \mu_{\text{ur}}^p \rho_1 & \text{if } N_2^p / N_2^f \geqslant r_2^p / r_1^f \end{cases}$$

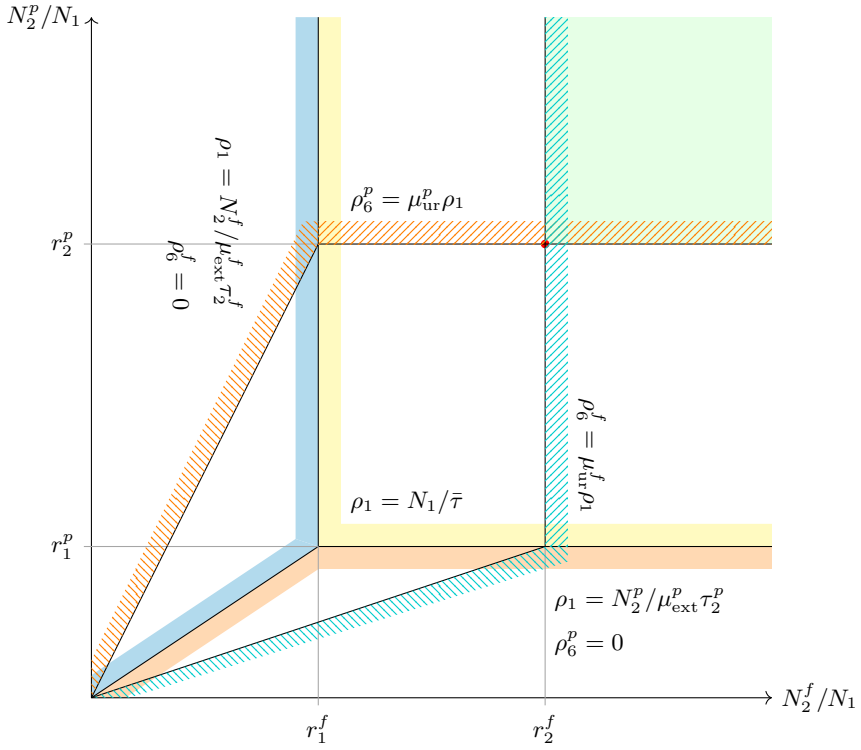


Figure 6.7 – Regimes of the call center with two groups at level 2. The yellow area is such that level 1 is fluid. The green area is such that every group of operators of level 2 is fluid. In the rest of the diagram, the throughput of level 1 is determined by the limiting group of operators at level 2.

- If $N_2^p/N_1 \leqslant r_1^p$ and $N_2^p/N_2^f \leqslant r_1^p/r_1^f$, then

$$\rho_1 = N_2^p/\mu_{\text{ext}}^p\tau_{\text{tr}}^p$$

$$\rho_6^f = \begin{cases} \dfrac{N_2^f}{\tau_{\text{ur}}'^f} - N_2^p\dfrac{\mu_{\text{ext}}^f\tau_{\text{ext}}^f}{\mu_{\text{ext}}^p\tau_{\text{ext}}^p\tau_{\text{ur}}'^f} & \text{if } N_2^f/N_2^p \leqslant r_2^f/r_1^p \\ \mu_{\text{ur}}^f\rho_1 & \text{if } N_2^f/N_2^p \geqslant r_2^f/r_1^p \end{cases}$$

$$\rho_6^p = 0$$

- If $N_2^f/N_1 \geqslant r_1^f$ and $N_2^p/N_1 \geqslant r_1^p$ then

$$\rho_1 = N_1/\bar{\tau}$$

$$\rho_6^f = \begin{cases} \dfrac{N_2^f}{\tau_{\text{ur}}'^f} - \mu_{\text{ext}}^f\dfrac{\tau_{\text{tr}}^f}{\tau_{\text{ur}}'^f}\rho_1 & \text{if } N_2^f/N_1 \leqslant r_2^f \\ \mu_{\text{ur}}^f\rho_1 & \text{if } N_2^f/N_1 \geqslant r_2^f \end{cases}$$

$$\rho_6^p = \begin{cases} \dfrac{N_2^p}{\tau_{\text{ur}}'^p} - \mu_{\text{ext}}^p\dfrac{\tau_{\text{tr}}^p}{\tau_{\text{ur}}'^p}\rho_1 & \text{if } N_2^p/N_1 \leqslant r_2^p \\ \mu_{\text{ur}}^p\rho_1 & \text{if } N_2^p/N_1 \geqslant r_2^p \end{cases}$$

The different regimes of this new organization are depicted in Figure 6.7. When $N_2^f/N_1 \geqslant r_1^f$ and $N_2^p/N_1 \geqslant r_1^p$, the throughput at level 1 is maximal, equal to $N_1/\bar{\tau}$, and the throughputs at both sides of level 2, police and firemen, are determined by the same kind of formulæ as in the case of Chapter 3, as if this level 2 was unique. The reason is that, in this situation, the two groups of level 2 are able to handle all extremely urgent calls, so that level 1 operators are not slowed down by level 2.

In contrast, when one of the two groups of level 2 is undersized in comparison to level 1, this affects level 1, and therefore, also the other group of level 2, even if this group was normally

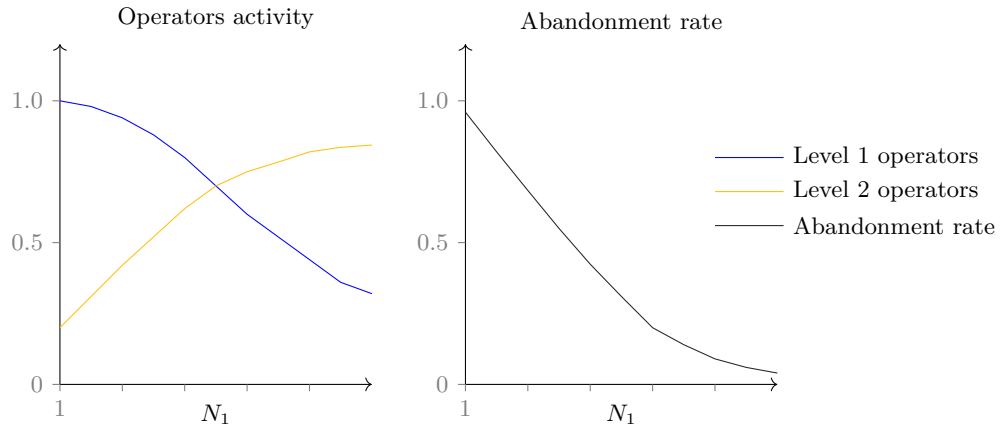


Figure 6.8 – Bottleneck level 1: when the number of level 1 operators diminishes, the fraction of abandoning calls increases, and the occupancy ratio of (remaining) level 1 operators increases, while level 2 operators become inactive. These graphs are "hand-drawn", but directly inspired by outputs of simulations. Therefore, the horizontal axis is intentionally not given values.

sized. We call this situation ***cross-congestion***, and it is identified as a critical issue for the call center practitioners. In this situation, the whole system is congested because of just one undersized group of operators. General bad performance for call treatment is then associated with some operators being underemployed in the other group at level 2. The next sections analyze this issue by the means of simulations.

6.2.2 Cross-congestion

Our simulations allow us to illustrate phenomena of cross-congestion exhibited by computations, and to point out their effects on operators activity and abandonment rates.

First, we recall that an undersized level 1 can become bottleneck for the system. As all calls have to go through level 1, the fluidity of this level is crucial for the call center. A direct illustration is given in Figure 6.8: if the number of operators of level 1 is too low, calls abandon before level 1, and cannot be treated by level 2 operators, who become inactive, which yields inefficient situations. Furthermore, calls giving up before level 1 are not qualified, so that urgent calls and extremely urgent calls can be lost at this step. Note that, in practice, when level 1 encounters congestion, level 1 operators can speed up conversation times of advice calls. This flexibility is crucial for the system.

Now, consider a case in which a large fraction of calls are accompanied to a group of level 2 (for example, firemen). Then, in a situation in which this group is saturated by incoming calls, level 1 operators are blocked, forced to wait with callers until an operator of this group becomes available. As a consequence, level 1 is slowed down, so that it also cannot feed the other group(s) of level 2, which become inactive. This is illustrated by Figure 6.9. From a practitioner point of view, this is a sensitive situation, in which police can hamper firemen, or firemen can hamper police.

We proposed a few methods to avoid such situations, like trunk reservation or cross-priorities. Inactive operators of level 2 could also be allowed to directly answer incoming calls. The first two methods are made possible by the fact that most calls from the emergency number 17 are eventually oriented to police, and most calls from the emergency number 18 are eventually oriented to firemen.

6.2.3 Arbitrating between the different levels of priority.

The role of three-way conferences between level 1 and level 2 is to improve quality of call treatment for the most urgent calls. An important question is the following: is it better to accompany as much calls as possible between the two levels?

While the answer shall be affirmative in fluid situations, our experiments tend to show that, if all calls are kept in line by level 1 operators while queuing, this can be counterproductive.

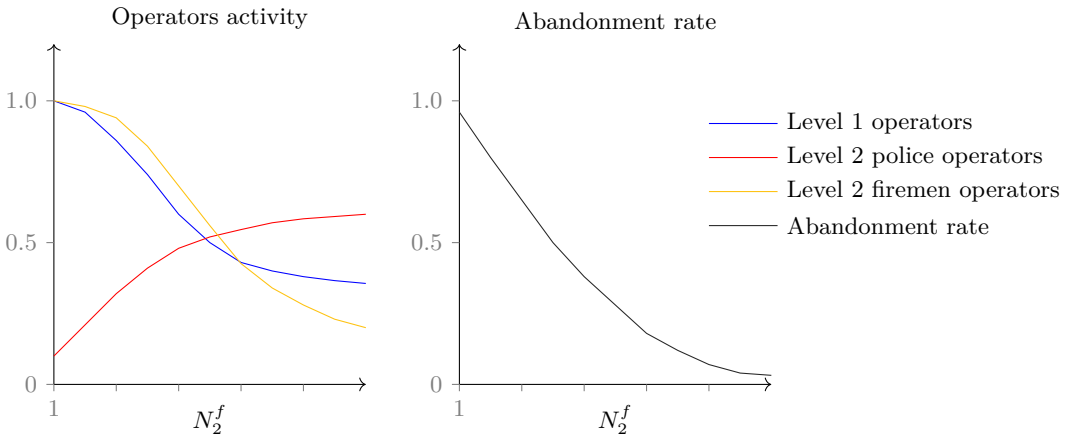


Figure 6.9 – Cross congestion: here, calls dispatched to the group of firemen of level 2 are all accompanied. When the number of firemen operators of level 2 decreases, waiting times between level 1 and level 2 increase, so that level 1 operators are more and more active. As a consequence, fewer calls are transferred to police operators of level 2, who become inactive. These graphs are "hand-drawn", but directly inspired by outputs of simulations. Therefore, the horizontal axis is intentionally not given values.

Indeed, this diminishes the range of the buffer regime in which level 1 is fluid, while a group of level 2 starts to be saturated. At the limit when all calls are accompanied, this buffer regime disappears, and as soon as no level 2 operator is available, level 1 begins to be slowed down. This can be negative for the whole system, and even for urgent calls accompanied between level 1 and level 2, because some of these calls are lost before level 1. See Figure 6.10.

Therefore, a subtle compromise should be found between these two extremes of accompanying no urgent call at level 2 (and taking the risk to loose some extremely urgent calls in the queue before level 2), and accompanying all urgent calls, which quickly slows down the whole system. Of course, in fluid situations, accompanying all urgent calls to level 2 is possible, as soon as procedures are set to handle congestion situations.

Our experiments tend to show that, as long as extremely urgent calls represent a small proportion of the urgent calls, the buffer regime is wide enough, and accompanying extremely urgent calls to level 2 does not hamper the system.

6.3 Other lessons from simulations

6.3.1 Efficiency of operators

The impatience of callers, associated with the random arrival of calls, result in the existence of a trade-off between idleness of operators and abandonment rate: one cannot achieve full-time activity of operators and no caller giving up with the same organization.

Our simulations allow us to draw the curve of the best trade-offs between abandonment rates and occupancy ratios (or efficiency) for our 17-18-112 emergency call center (see Figures 6.11, 6.12 and 6.13). For example, in order to achieve an abandonment rate of less than 10% of the calls entering the system, the best occupancy ratio that can be achieved in the system (with an appropriate distribution of operators in the different groups) is 70%. In other words, in order to achieve a certain performance in terms of call abandonments, one must expect that operators in the call center remain idle during significant periods of time.

Of course, this best trade-off between operators activity and abandonment is strongly related to the size of the call center. The larger the volume of incoming calls is, the more productive are operators of the call center, because the effect of the randomness of calls arriving is reduced in proportions, due to their large number.

We refer to the model of [GMR02] for a mathematical analysis of this trade-off between efficiency and abandonment.

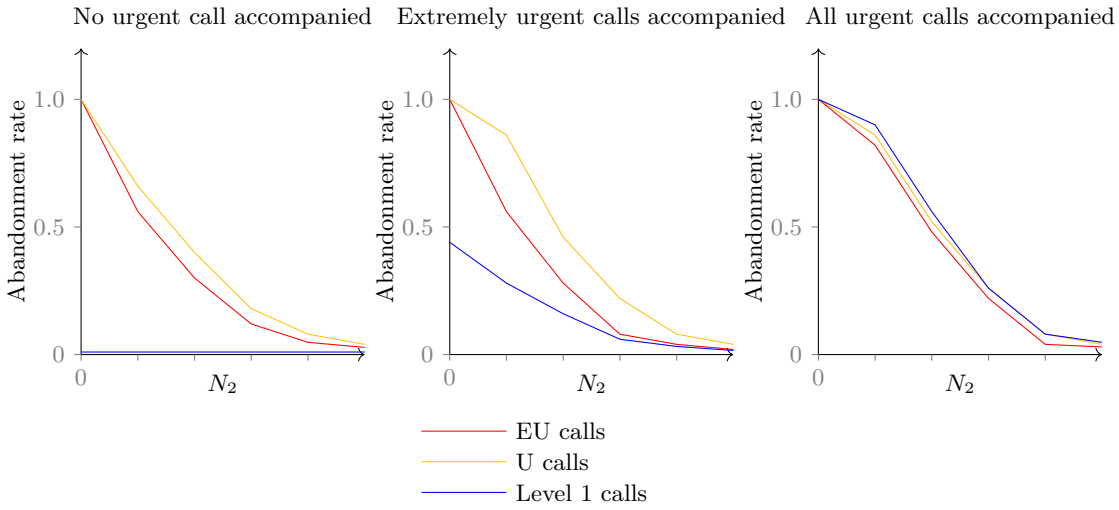


Figure 6.10 – Abandonment rates depending on the number of operators at the corresponding group of level 2, for extremely urgent (EU) calls, simply urgent (U) calls, and calls at level 1. The three graphs describe three strategies of call accompaniment. The second strategy (accompanying only extremely urgent calls) is the strategy that best protects EU calls in situations of low congestion. These graphs are "hand-drawn", but directly inspired by outputs of simulations.

6.3.2 Other qualitative observations

The sensitivity to conversation times An important part of our simulations and analyses were conducted before the new bilevel emergency call center became operational. At that moment, the conversation times of calls at level 1 and level 2 could only be estimated from the conversation times observed in the current single-level call centers. For this purpose, a conservative approach would be to consider that the conversation time at level 2 would be equal to the conversation time observed in the single-level call center. Then conversation time at level 1 must still be added to obtain the total conversation time experienced by the caller. The total of the conversation times of a call at level 1 and level 2 can also be bounded from below: it should not be inferior to the conversation time observed in the single-level call center.

In our simulations, we found our results to be very sensitive to the position of the cursor between these two extreme cases. The total conversation time of the new organization, if much larger than the conversation time of the ancient one, could lead to lower throughputs and lower performance, if the number of operators is equivalent, or to a larger number of operators, in order to reach the same level of (quantitative) performance, despite the advantages of pooling operators from police and firemen.

This can of course also be assessed from the classical analytical formulæ, which establish that the throughput of a bottleneck service place is proportional to the inverse of the mean service time.

We highlighted this in our reports delivered to the heads of the new emergency call center project: the success of the project was also conditioned to the ability of the new organization to staff more operators or to control conversation times, so that information is not repeated nor lost between level 1 and level 2, and that the call qualification at level 1 is as efficient as possible.

However, the advantages expected from the new organization of the emergency call center should not be measured only in terms of throughputs and quantitative performance (*e.g.*, volume of calls answered), but also in regard to other criteria of fundamental importance, such as quality of call handling, appropriateness of emergency means dispatching, or improved coordination between police and firemen (see Section 1.1).

The need for detailed metrics As in any operations center, there is a need for practitioners to evaluate the state of the emergency call center by a few eloquent metrics, easy to understand

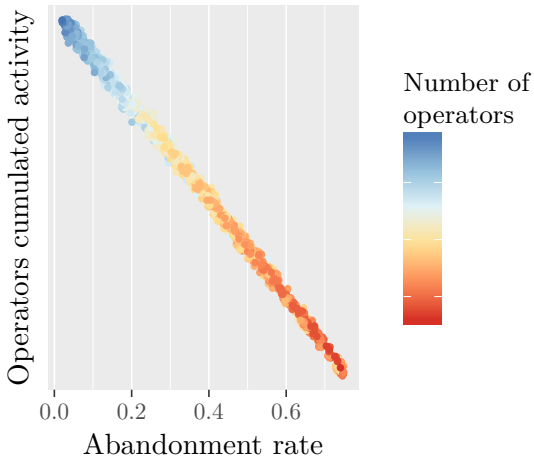


Figure 6.11 – Operators activity (total time spent in conversation with callers) and abandonment rates: observe that the relation is linear: the number of calls handled is proportional to the volume of time spent by operators in conversation. However, the number of operators required to answer these calls is much larger than the sum of the treatment times divided by the time period.

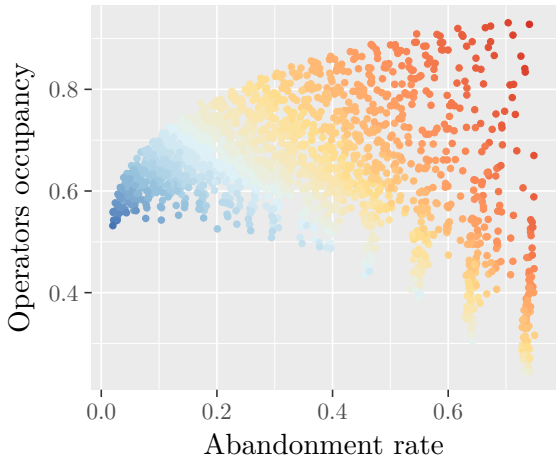


Figure 6.12 – Operators occupancy ratio and abandonment rate. In comparison with Figure 6.11, the vertical axis represents a ratio and not a volume. Each point corresponds to a given (naive) staffing of operators in the different groups of the call center. Note that the major part of these naive staffings are inefficient, *i.e.*, they do not reach a good activity - abandonment trade-off.

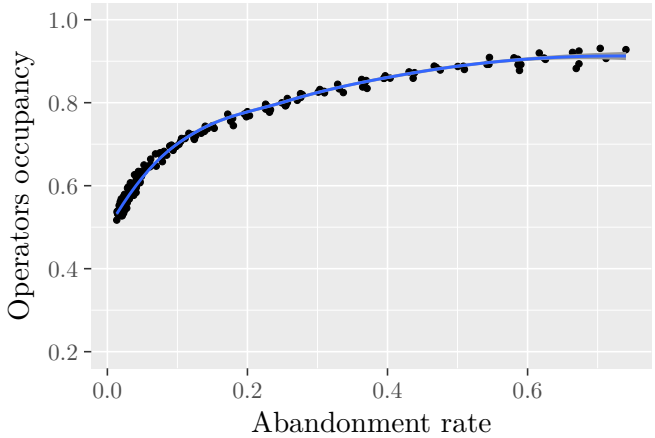


Figure 6.13 – Pareto optima of Figure 6.12: best trade-offs between occupancy ratio and abandonment rate.

and to support decision. Examples of these simple metrics are the volume of calls answered, the maximal waiting times, the maximal number of callers waiting in the queue, or the abandonment rates [1].

In the architecture of the new emergency call center, however, these simple metrics do not account for the different types of calls, and for their various itineraries. Firstly, the urgency of the calls being qualified by operators of level 1, it is possible and advised to consider metrics specific to the urgent calls and extremely urgent calls, which are supposed to be offered a better treatment.

Secondly, our simulations showed that, under some conditions, a good performance for calls in general could be associated with a much lower performance for a certain category of calls, raising fairness issues. As an example, suppose that, for one of the emergency numbers 17 or 18, whose urgent calls are essentially directed to, respectively, second level police operators and second level firemen operators, the ratio of extremely urgent calls is much lower than for the other one. Then, it can happen that the mean waiting time of extremely urgent calls is low, while it remains high for this specific emergency number. Such situation would lead to extremely urgent calls of one emergency number experiencing long waiting times.

While this could be the best configuration depending on the context, we strongly recommend to keep abreast of such unbalanced situations, and hence, to develop a certain number of indicators so that the performance measure is differentiated according to the different call types.

6.4 Concluding remarks

Complementary to our analytical developments of Part I, the large range of simulations that we ran was a crucial work in order to understand our complex, bilevel emergency call center, especially in critical configurations identified by the practitioners (meteorological events, for example), and to test the system at its limits.

A major limitation of our work, that we underlined regularly in our contacts with the call center project leaders, is that our results focus on congestion, circulation of calls, and, more generally, quantitative performance of the call treatment, aiming at optimizing the staffing of the call center and minimizing the number of lost calls. This is just one dimension of the whole criteria that practitioners have to keep in mind.

In particular, the quality of conversations, leading to better interactions with the caller, who may be in situations of distress, and to a better dispatching of emergency means, is crucial in an emergency call center. Moreover, some major benefits of the new organization could not be captured by our study. Thus, the existence of a first level group of operators allows to protect second level operators from calls with no one in line, or callers with inappropriate requests, and therefore to focus on real emergencies. Moreover, a key feature of this project was to gather together policemen and firemen, and hence, to improve coordination between services.

Another limitation of our work is that we did not model the human intelligence of the organization, which is illustrated, for example, in Figure 6.3: operators adapt their efficiency (and their breaks and inter-call delays) to the call center peak periods, but which of course comprises many other characteristics.

Despite these limitations, we were able to provide some quantitative and qualitative results to the emergency call center project leaders, on top of the classical call center theory, which already indicates that pooling operators is beneficial in general.

The major lesson that we exhibited is the cost of keeping calls in line with an operator between level 1 and level 2. Our analyses show that it may be harmless in fluid situations, but that, as soon as a group of operators of level 2 encounters some congestion, it can slow down level 1, and therefore, the whole system, generating "cross-congestion". However, we do not recommend to stop accompanying calls from level 1 to level 2 (remember that this operational choice was driven by the need to serve properly people in distress), but rather to keep this feature for the most important calls or, alternatively, to activate some downgraded modes as soon as this feature yields congestion at level 1. Simulations and numerical analyses show that keeping this feature only for a fraction of urgent calls permits to have a "buffer regime" in which a group of level 2 starts to be saturated, without impacting the rest of the call center. This

1. Examples of more "state-of-the-art" metrics can be found in [ACG+10, ATLB16]

coincides with our analytical results exposed in Part I, see for example the graph of Figure 1.5, in which Phase 2 corresponds to the buffer regime.

Other key issues were identified in this chapter, such as the unavoidable inactivity of operators during some periods of time, in order to be able to address more incoming calls, the importance of differentiating metrics indicating the quality of service for the various emergency number and types of urgencies, or the impact of conversation times, which controls at first order the number of operators needed to staff the call center.

Finally, this case study also pointed out, from a practitioner point of view, the importance for an emergency call center to identify quickly and properly the most urgent calls, so as to provide them a service as good as it can be. This was one of the reason of this two-level organization of our case study. More generally, it calls for other innovations and ideas, that should probably involve recent or future technologies, to serve this objective of identifying quickly and properly the most urgent calls. For these calls, some seconds spared can dramatically make the difference.

6.5 A few words on our simulations

For our simulations, we developed a program written in Python 3 and using only its standard libraries [2]. Our program allows us to design various call centers, in which calls circulate depending on their characteristics (emergency number, urgency). It allows to model situations in which several operators of different groups are in conversation with the same person. In case of concurrency between different type of calls in queues, calls are assigned to operators according to different levels of priority, and to a "first-in, first-out" rule, for calls of the same priority. For each queue of the network, it is possible to specify if calls can give up or not. Metrics (waiting times, abandonments) are collected for each queue of the network and for each type of calls. It is also possible to measure throughputs at the different stages of the calls itineraries.

Our program allows us to "replay" some days of activity of our call center, while varying the staffing of operators, the architecture of the call center, or other parameters of the system. In particular, we had access to data of a few crisis events (like meteorological events, celebrations and festivals), leading to stressed situations in which the volume of emergency calls forced the call center to work at full regime. It is also possible to generate a data set of "incoming calls" from scratch, following certain statistical distributions and parameters. Some insights on the statistics of our data are given in Section 6.1 (see also Avramidis and L'Ecuyer [AL05]).

Of these two cases, replaying real situations is the toughest part, as a lot of data reconstruction is required. For example, for a given call logged in the data, one cannot determine simultaneously its patience (its willingness to wait for service) and its conversation time: if the caller gives up, we have access to the call's patience, but not to the conversation time, and if the call is answered, we have access to the conversation times, but only to a lower bound of its patience. For the needs of our simulations, we re-constructed data according to the following lines: first, we considered that conversation times of calls giving up were similar in distribution to the conversation times of calls answered. Second, we assumed patience of callers to follow an exponential distribution, in accordance with the analysis developed in Section 6.1 (see Figure 6.4). This made it simple to compute patience of calls that were answered, just by adding an exponential draw to their waiting times.

The validation of our program (before the new call center became operative, and its data became available) was twofold. First, we modeled the previous call center organizations, so that comparing our simulations with data sets of these call centers allowed us to validate our model and to calibrate some parameters. Moreover, we benefited from the student project of Dejean de la Bâtie and Petroff [DdlBP16], who worked on the same case study and also simulated one of the architectures in project, in an other programming language: this double simulation allowed us to validate our program by pointing out and resolving differences in the outputs.

2. We plan to share some parts of this code online, so that the community can use it and verify it. However, some work is still required to clean the specificities from our call center case study.

Bibliography

[AAKD08] M Salah Aguir, O Zeynep Akşin, Fikri Karaesmen, and Yves Dallery. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*, 191(2):398–408, 2008.

[ABG15] Xavier Allamigeon, Vianney Bœuf, and Stéphane Gaubert. Performance evaluation of an emergency call center: tropical polynomial systems applied to timed Petri nets. In *FORMATS*, volume 9268 of *Lecture Notes in Computer Science*, pages 10–26. Springer, 2015.

[ABG16] Xavier Allamigeon, Vianney Bœuf, and Stéphane Gaubert. Stationary solutions of discrete and continuous Petri nets with priorities. In *VALUETOOLS'16*, Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools, 2016.

[ABG17] Xavier Allamigeon, Vianney Bœuf, and Stéphane Gaubert. Stationary solutions of discrete and continuous Petri nets with priorities. *Performance Evaluation*, 2017.

[ACG+10] Athanassios N Avramidis, Wyean Chan, Michel Gendreau, Pierre L'ecuyer, and Ornella Pisacane. Optimizing daily agent scheduling in a multiskill call center. *European Journal of Operational Research*, 200(3):822–832, 2010.

[ADL04] Athanassios N Avramidis, Alexandre Deslauriers, and Pierre L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

[AFG+14] Xavier Allamigeon, Uli Fahrenberg, Stéphane Gaubert, Axel Legay, and Ricardo Katz. Tropical Fourier-Motzkin elimination, with an application to real-time verification. *International Journal of Algebra and Computation*, 24(5):569–607, 2014.

[AG03] Marianne Akian and Stéphane Gaubert. Spectral theorem for convex monotone homogeneous maps, and ergodic control. *Nonlinear Anal.*, 52(2):637–679, 2003.

[AL05] Athanassios N Avramidis and Pierre L'Ecuyer. Modeling and simulation of call centers. In *Simulation Conference, 2005 Proceedings of the Winter*, pages 9–pp. IEEE, 2005.

[AN01] P. A. Abdulla and A. Nylén. Timed Petri nets and BQOs. In *Applications and Theory of Petri Nets '01*, volume 2075 of *LNCS*. Springer, 2001.

[AO76] Robert F. Anderson and Steven Orey. Small random perturbation of dynamical systems with reflecting boundary. *Nagoya Math. J.*, 60:189–216, 1976.

[ATLB16] Thuy Anh Ta, Pierre L'Ecuyer, and Fabian Bastin. Staffing optimization with chance constraints for emergency call centers. In *MOSIM 2016-11th International Conference on Modeling, Optimization and Simulation*, 2016.

[Avi00] David Avis. A revised implementation of the reverse search vertex enumeration algorithm. In *Polytopes—combinatorics and computation (Oberwolfach, 1997)*, volume 29 of *DMV Sem.*, pages 177–198. Birkhäuser, Basel, 2000.

[Bal11] Simonetta Balsamo. *Queueing Networks with Blocking: Analysis, Solution Algorithms and Properties*, pages 233–257. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[Bau93] Falko Bause. Queueing Petri Nets - a formalism for the combined qualitative and quantitative analysis of systems. In *Petri Nets and Performance Models, 1993. Proceedings., 5th International Workshop on*, pages 14–23. IEEE, 1993.

[BCH+05] B. Bérard, F. Cassez, S. Haddad, D. Lime, and O. H. Roux. Comparison of the expressiveness of timed automata and time Petri nets. In *FORMATS'05*, volume 3829 of *Lecture Notes in Computer Science*. Springer, 2005.

[BCOQ92] François Baccelli, Guy Cohen, Geert Jan Olsder, and Jean-Pierre Quadrat. *Synchronization and Linearity*. Wiley, 1992.

[BD90] Eike Best and Jörg Desel. Partial order behaviour and structure of Petri nets. *Formal aspects of computing*, 2(1):123–138, 1990.

[BD91] Bernard Berthomieu and Michel Diaz. Modeling and verification of time dependent systems using time Petri nets. *Software Engineering, IEEE Transactions on*, 17(3), 1991.

[Ben09] Pascal Benchimol. *Modélisation du service des urgences de l'Hôtel-Dieu*. Rapport de stage d'option, École Polytechnique, 2009.

[BF95] François Baccelli and Serguei Foss. On the saturation rule for the stability of queues. *J. Appl. Probab.*, 32(2):494–507, 1995.

[BFG96] François Baccelli, Serguei Foss, and Bruno Gaujal. Free-choice Petri nets-an algebraic approach. *IEEE Transactions on Automatic Control*, 41(12):1751–1778, 1996.

[BGM98] Fabio Balduzzi, Alessandro Giua, and Giuseppe Menga. Hybrid stochastic Petri nets: firing speed computation and FMS modelling. In *WODES'98, Proc. Fourth Workshop on Discrete Event Systems*, pages 432–438, 1998.

[BGM+05] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50, 2005.

[BGM06] Anne Bouillard, Bruno Gaujal, and Jean Mairesse. Extremal throughputs in free-choice nets. *Discrete Event Dyn. Syst.*, 16(3):327–352, 2006.

[Bil99] Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[BJS09] J. Byg, K. Y. Jørgensen, and J. Srba. Tapaal: Editor, simulator and verifier of timed-arc Petri nets. In *ATVA'09*, volume 5799 of *LNCS*. Springer, 2009.

[BK92] Eike Best and Maciej Koutny. Petri net semantics of priority systems. *Theoret. Comput. Sci.*, 96(1):175–215, 1992. Second Workshop on Concurrency and Compositionality (San Miniato, 1990).

[BM98] François Baccelli and Jean Mairesse. Ergodic theorems for stochastic operators and discrete event networks. *Idempotency (Bristol, 1994)*, 11:171–208, 1998.

[Bon99] Thomas Bonald. *Stabilite des systemes dynamiques a evenements discrets application au controle de flux dans les reseaux de telecommunication*. PhD thesis, 1999.

[Bow00] Fred DJ Bowden. A brief survey and synthesis of the roles of time in Petri nets. *Math. and Comput. Model.*, 31(10):55–68, 2000.

[BR17] Vianney Boeuf and Philippe Robert. A stochastic analysis of a network with two levels of service. *arXiv preprint arXiv:1708.09590*, 2017.

[BV06] B. Berthomieu and F. Vernadat. Time Petri nets analysis with TINA. In *QEST'06*. IEEE, 2006.

[BW13] Eike Best and Harro Wimmel. Structure theory of Petri nets. In *Transactions on Petri Nets and Other Models of Concurrency VII*, pages 162–224. Springer, 2013.

[CCS91] Javier Campos, Giovanni Chiola, and Manuel Silva. Ergodicity and throughput bounds of Petri nets with unique consistent firing count vector. *IEEE Trans. Software Engrg.*, 17(2):117–125, 1991.

[CDF91] Giovanni Chiola, Susanna Donatelli, and Guiliana Franceschinis. Priorities, inhibitor arcs and concurrency in P/T nets. In *Proc. of ICATPN*, volume 91, pages 182–205, 1991.

[CE09] Pierre Collet and J-P Eckmann. *Iterated maps on the interval as dynamical systems*. Springer Science & Business Media, 2009.

[CFS12] Isaac P Cornfeld, Sergej V Fomin, and Yakov Grigorévič Sinai. *Ergodic theory*, volume 245. Springer Science & Business Media, 2012.

[CGQ95] Guy Cohen, Stéphane Gaubert, and Jean-Pierre Quadrat. Asymptotic throughput of continuous timed Petri nets. In *CDC*, 1995.

[CGQ98] Guy Cohen, Stéphane Gaubert, and Jean-Pierre Quadrat. Algebraic system analysis of timed Petri nets. In J. Gunawardena, editor, *Idempotency*, Publications of the Isaac Newton Institute, pages 145–170. Cambridge University Press, 1998.

[CHB14] B. Cottenceau, L. Hardouin, and J.L. Boimond. Modeling and control of weight-balanced timed event graphs in dioids. *IEEE Trans. Autom. Control*, 59(5), 2014.

[CM91] H. Chen and A. Mandelbaum. Discrete flow networks: bottleneck analysis and fluid approximations. *Mathematics of Operation Research*, 16(2):408–446, May 1991.

[Cor08] Jorge Cortes. Discontinuous dynamical systems. *IEEE control Systems*, 28(3), 2008.

[CS91] José Manuel Colom and M. Silva. Convex geometry and semiflows in P/T nets. A comparative study of algorithms for computation of minimal p-semiflows. In *Advances in Petri nets 1990 (Bonn, 1989)*, volume 483 of *Lecture Notes in Comput. Sci.*, pages 79–112. Springer, Berlin, 1991.

[CT80] Michael G. Crandall and Luc Tartar. Some relations between nonexpansive and order preserving mappings. *Proc. Amer. Math. Soc.*, 78(3):385–390, 1980.

[DA87] René David and Hassane Alla. Continuous Petri nets. In *8th European Workshop on Application and Theory of Petri nets*, volume 340, pages 275–294, 1987.

[DA10] René David and Hassane Alla. *Discrete, continuous, and hybrid Petri nets*. Springer Science & Business Media, 2010.

[DdlBP16] P. Dejean de la Bâtie and S. Petroff. *Projet de modélisation de la PFAU*. Rapport de projet de troisième année, École Polytechnique, 2016.

[DE95] Jörg Desel and Javier Esparza. Free choice Petri nets. 1995.

[DN08] R.W.R. Darling and J.R. Norris. Differential equation approximations for Markov chains. *Probab. Surv.*, 5:37–79, 2008.

[dpdP16] Préfecture de police de Paris. La plateforme des appels d'urgence 17-112-18. Brochure, 2016. www.prefecturedepolice.interieur.gouv.fr/content/download/26519/189631/file/dp-PFAU2016.pdf (pdf).

[EKCM78] N. El Karoui and M. Chaleyat-Maurel. *Temps locaux*, volume 52-53, chapter Un problème de réflexion et ses applications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$, pages 117–144. Société Mathématique de France, 1978. Exposés du Séminaire J. Azéma-M. Yor, Paris, 1976–1977.

[EN94] Javier Esparza and Mogens Nielsen. Decidability issues for Petri nets. *BRICS Report Series*, 1(8), 1994.

[FGQ11] Nadir Farhi, Maurice Goursat, and Jean-Pierre Quadrat. Piecewise linear concave dynamical systems appearing in the microscopic traffic modeling. *Linear Algebra and Appl.*, pages 1711–1735, 2011.

[FH15] Estíbaliz Fraca and Serge Haddad. Complexity analysis of continuous Petri nets. *Fundamenta informaticae*, 137(1):1–28, 2015.

[Fil88] Aleksei Fedorovich Filippov. Differential equations with discontinuous right-hand side. Kluwer Academic Publishers, 1988.

[FL15] Alain Finkel and Jérôme Leroux. Recent and simple algorithms for Petri nets. *Software & Systems Modeling*, 14(2):719–725, 2015.

[FLGD+11] Goran Frehse, Colas Le Guernic, Alexandre Donzé, Scott Cotton, Rajarshi Ray, Olivier Lebeltel, Rodolfo Ripado, Antoine Girard, Thao Dang, and Oded Maler. SpaceEx: Scalable verification of hybrid systems. In *CAV*, pages 379–395. Springer, 2011.

[FP96] Komei Fukuda and Alain Prodon. Double description method revisited. In *Combinatorics and computer science (Brest, 1995)*, volume 1120 of *Lecture Notes in Comput. Sci.*, pages 91–111. Springer, Berlin, 1996.

[Gau05] Stéphane Gaubert. Nonlinear perron-frobenius theory and discrete event systems. *Journal européen des systèmes automatisés*, 39(1/3):175, 2005.

[GG98] Stéphane Gaubert and Jeremy Gunawardena. The duality theorem for min-max functions. *C. R. Acad. Sci. Paris.*, 326, Série I:43–48, 1998.

[GG04a] Stéphane Gaubert and Jeremy Gunawardena. The perron-frobenius theorem for homogeneous, monotone functions. *Transactions of the American Mathematical Society*, 356(12):4931–4950, 2004.

[GG04b] Bruno Gaujal and Alessandro Giua. Optimal stationary behavior for a class of timed continuous Petri nets. *Automatica*, 40(9):1505–1516, 2004.

[GHM03] Bruno Gaujal, Stefan Haar, and Jean Mairesse. Blocking a transition in a free choice net and what it tells about its throughput. *Journal of Computer and System Sciences*, 66(3):515–548, 2003.

[GMR02] Ofer Garnett, Avishai Mandelbaum, and Martin Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[GRR04] G. Gardey, O. H. Roux, and O. F. Roux. Using zone graph method for computing the state space of a time Petri net. In *FORMATS'04*, volume 2791 of *Lecture Notes in Computer Science*. Springer, 2004.

[Gt16] Torbjörn Granlund and the GMP development team. *GNU MP: The GNU Multiple Precision Arithmetic Library*, 6.1.1 edition, 2016. http://gmplib.org/.

[Hac76] Michel Henri Théodore Hack. *Decidability questions for Petri Nets.* PhD thesis, Massachusetts Institute of Technology, 1976.

[HcVdSS02] WPMH Heemels, MK Çamlıbel, AJ Van der Schaft, and JM Schumacher. On the existence and uniqueness of solution trajectories to hybrid dynamical systems. *Nonlinear and hybrid control in automotive applications*, pages 391–422, 2002.

[Hen00] Thomas A Henzinger. The theory of hybrid automata. In *Verification of Digital and Hybrid Systems*, pages 265–292. Springer, 2000.

[HOvdW06] G. Heidergott, G. J. Olsder, and J. van der Woude. *Max Plus at work*. Princeton University Press, 2006.

[HR81] J.M. Harrison and M.I. Reiman. Reflected Brownian motion on an orthant. *Annals of Probability*, 9(2):302–308, 1981.

[ILST16] Rouba Ibrahim, Pierre L'Ecuyer, Haipeng Shen, and Mamadou Thiongane. Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research*, 250(2):480–492, 2016.

[IvdS00] Jun-ichi Imura and Arjan van der Schaft. Characterization of well-posedness of piecewise-linear systems. *IEEE Trans. Automat. Control*, 45(9):1600–1619, 2000.

[IYLS16] Rouba Ibrahim, Han Ye, Pierre L'Ecuyer, and Haipeng Shen. Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3):865–874, 2016.

[JJMS11] L. Jacobsen, M. Jacobsen, M. H. Møller, and J. Srba. Verification of timed-arc Petri nets. In *SOFSEM'11*, volume 6543 of *LNCS*. Springer, 2011.

[JK01] Geurt Jongbloed and Ger Koole. Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318, 2001.

[JRS03] Jorge Júlvez, Laura Recalde, and Manuel Silva. On reachability in autonomous continuous Petri net systems. In *ICATPN*, pages 221–240. Springer, 2003.

[JRS05] Jorge Júlvez, Laura Recalde, and Manuel Silva. Steady-state performance evaluation of continuous mono-t-semiflow Petri nets. *Automatica*, 41(4):605–616, 2005.

[Kel79] Frank P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons Ltd., Chichester, 1979. Wiley Series in Probability and Mathematical Statistics.

[Kel86] Frank P. Kelly. Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability*, 18:473–505, 1986.

[Kel91] Frank P. Kelly. Loss networks. *Annals of Applied Probability*, 1(3):319–378, 1991.

[Kin93] J. F. C. Kingman. *Poisson processes*. Oxford studies in probability, 1993.

[KM69] Richard M Karp and Raymond E Miller. Parallel program schemata. *Journal of Computer and system Sciences*, 3(2):147–195, 1969.

[KM02] Ger Koole and Avishai Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1):41–59, 2002.

[Koh80] Elon Kohlberg. Invariant half-lines of nonexpansive piecewise-linear transformations. *Mathematics of Operations Research*, 5(3):366–372, 1980.

[Kos82] S Rao Kosaraju. Decidability of reachability in vector addition systems. In *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pages 267–281. ACM, 1982.

[Lib96] L. Libeaut. *Sur l'utilisation des dioïdes pour la commande des systèmes à événements discrets*. Thèse, École Centrale de Nantes, 1996.

[Lig75] Thomas M. Liggett. Ergodic theorems for the asymmetric simple exclusion process. *Transactions of the American Mathematical Society*, 213:237–261, 1975.

[Lig85] Thomas M. Liggett. *Interacting Particle Systems*. Grundlehren der mathematischen Wissenschaften. Springer Verlag, New York, 1985.

[Lip76] Richard Lipton. The reachability problem requires exponential space. *Research Report 62. Department of Computer Science, Yale University*, 1976.

[LRST09] D. Lime, O. H. Roux, C. Seidner, and L.-M. Traonouez. Romeo: A parametric model-checker for Petri nets with stopwatches. In *TACAS'09*, volume 5505 of *Lecture Notes in Computer Science*. Springer, 2009.

[May84] Ernst W Mayr. An algorithm for the general Petri net reachability problem. *SIAM Journal on computing*, 13(3):441–460, 1984.

[MBC+94] Marco Ajmone Marsan, Gianfranco Balbo, Gianni Conte, Susanna Donatelli, and Giuliana Franceschinis. *Modelling with generalized stochastic Petri nets*. John Wiley & Sons, Inc., 1994.

[McM70] Peter McMullen. The maximum numbers of faces of a convex polytope. *Mathematika*, 17(2):179–184, 1970.

[Mey12] A Meyer. Discontinuity induced bifurcations in timed continuous Petri nets. *IFAC Proceedings Volumes*, 45(29):28–33, 2012.

[MF76] Philip Merlin and David Farber. Recoverability of communication protocols–implications of a theoretical study. *IEEE transactions on Communications*, 24(9):1036–1043, 1976.

[Mol82] Michael K. Molloy. Performance analysis using stochastic Petri nets. *IEEE Transactions on computers*, 31(9):913–917, 1982.

[MRS06] Cristian Mahulea, Laura Recalde, and Manuel Silva. On performance monotonicity and basic servers semantics of continuous Petri nets. In *Discrete Event Syst., 2006 8th International Workshop on*, pages 345–351. IEEE, 2006.

[MRTRS08] Cristian Mahulea, Antonio Ramírez-Treviño, Laura Recalde, and Manuel Silva. Steady-state control reference and token conservation laws in continuous Petri net systems. *IEEE Trans. on Autom. Sci. and Eng.*, 5(2):307–320, 2008.

[MRTT53] T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall. The double description method. In *Contributions to the theory of games, vol. 2*, Annals of Mathematics Studies, no. 28, pages 51–73. Princeton University Press, Princeton, N. J., 1953.

[Mur89] Tadao Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.

[NGRTS16] Manuel Navarro-Gutiérrez, Antonio Ramírez-Treviño, and Manuel Silva. Discontinuities and non-monotonicities in mono-t-semiflow timed continuous Petri nets. In *Discrete Event Systems (WODES), 2016 13th International Workshop on*, pages 493–500. IEEE, 2016.

[Pal53] Conny Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4(189208):4–5, 1953.

[Pet62] Carl Adam Petri. Kommunikation mit automaten. 1962.

[Pet81] James Lyle Peterson. *Petri net theory and the modeling of systems*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981.

[Plu99] M. Plus. Max-plus-times linear systems. In *Open Problems in Mathematical Systems and Control Theory*. Springer, 1999.

[Put94] Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1994. A Wiley-Interscience Publication.

[RHS10] Laura Recalde, Serge Haddad, and Manuel Silva. Continuous Petri nets: Expressive power and decidability issues. *International Journal of Foundations of Computer Science*, 21(02):235–256, 2010.

[RMS06] Laura Recalde, Cristian Mahulea, and Manuel Silva. Improving analysis and simulation of continuous Petri nets. In *2006 IEEE CASE*, pages 9–14. IEEE, 2006.

[Rob03] Philippe Robert. *Stochastic Networks and Queues*, volume 52 of *Stochastic Modelling and Applied Probability Series*. Springer, New-York, June 2003.

[RR15] S. Raclot and R. Reboul. Analysis of the new call center organization at PP and BSPP. Personal communication to the authors, 2015.

[RS00] Laura Recalde and Manuel Silva. PN fluidification revisited: Semantics and steady state. *J. Zaytoon S. Engell, Automation of Mixed Processes: Hybrid Dynamics Systems*, pages 279–286, 2000.

[RTS99] Laura Recalde, Enrique Teruel, and Manuel Silva. Autonomous continuous P/T systems. In *Petri Nets*, pages 107–126. Springer, 1999.

[SCC92] M. Silva, José Manuel Colom, and J. Campos. Linear algebraic techniques for the analysis of Petri nets. In *Recent advances in mathematical theory of systems, control, networks and signal processing, II (Kobe, 1991)*, pages 35–42. Mita, Tokyo, 1992.

[Sko62] A.V. Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory Probab. Appl.*, 7:3–23, 1962.

[SR02] Manuel Silva and Laura Recalde. Petri nets and integrality relaxations: A view of continuous Petri net models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 32(4):314–327, 2002.

[Srb08] J. Srba. Comparing the expressiveness of timed automata and timed extensions of Petri nets. In *FORMATS'08*, volume 5215 of *LNCS*. Springer, 2008.

[Tc11] Le Quang Thuan and M. Kanat Çamlıbel. Continuous piecewise affine dynamical systems do not exhibit Zeno behavior. *IEEE Trans. Automat. Control*, 56(8):1932–1936, 2011.

[TCS97] Enrique Teruel, José Manuel Colom, and Manuel Silva. Choice-free Petri nets: A model for deterministic concurrent systems with bulk services and arrivals. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(1):73–83, 1997.

[Tre88] Nicolas Treves. A comparative study of different techniques for semi-flows computation in place/transition nets. In *European Workshop on Applications and Theory in Petri Nets*, pages 433–452. Springer, 1988.

[TS96] Enrique Teruel and Manuel Silva. Structure theory of equal conflict systems. *Theoretical Computer Science*, 153(1-2):271–300, 1996.

[VMJS13] C Renato Vázquez, Cristian Mahulea, Jorge Júlvez, and Manuel Silva. Introduction to fluid Petri nets. In *Control of Discrete-Event Systems*, pages 365–386. Springer, 2013.

[Wat64] Shinzo Watanabe. On discontinuous additive functionals and Lévy measures of a Markov process. *Japanese Journal of Mathematics*, 34:53–70, 1964.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Dynamique d'un système bi-niveau avec priorités et application à un centre d'appels d'urgence

Mots clés : évaluation de performance, réseau de Petri, calcul stochastique, systèmes à événements discrets, systèmes hybrides, centres d'appels

Résumé : Nous analysons la dynamique de systèmes à événements discrets avec synchronisation et priorités, au moyen de réseaux de Petri et de réseaux de files d'attente. Nous appliquons cela à l'évaluation de performance d'un centre d'appels d'urgence.
Notre motivation est en premier lieu pratique. En 2016, un nouveau centre d'appels d'urgence a été mis en place pour l'agglomération parisienne, traitant les appels pour la police et les pompiers. La nouvelle organisation comporte deux niveaux de traitement. Un premier niveau d'opérateurs répond aux appels, identifie les appels urgents et traite les appels non urgents. Les opérateurs de second niveau sont spécialistes (policiers ou pompiers) et traitent les demandes d'intervention. De plus, certains appels sont identifiés comme très urgents et bénéficient d'un traitement prioritaire. Nous nous intéressons à l'évaluation de performance de divers systèmes correspondant à cette description générale.
Nous proposons trois modélisations différentes. Les deux premières sont des modèles de réseaux de Petri temporisés. Nous enrichissons le cadre classique des réseaux de Petri à choix libres en autorisant des situations de conflit où le routage est résolu par des priorités. La principale difficulté est alors que l'opérateur de la dynamique n'est plus monotone.
Dans un premier modèle, nous proposons une dynamique discrète pour cette classe de réseaux de Petri. Nous prouvons que les variables compteurs du réseau sont les solutions d'un système affine par morceaux avec retards. Nous étudions les régimes stationnaires de cette dynamique, et caractérisons les régimes affines comme solutions d'un système affine par morceaux, qui peut être vu comme un système d'équations rationnelles sur le semi-corps de germes tropical (min plus). Les applications numériques montrent cependant que la convergence ne se fait pas toujours vers ces régimes stationnaires affines.
Le second modèle est une version infinitésimale du précédent. Pour la même classe de réseaux de Petri, nous proposons une dynamique sous forme d'équations différentielles : les jetons et le temps deviennent continus. Pour ce système différentiel discontinu, nous établissons l'existence et l'unicité de la solution. L'avantage de cette modélisation continue est que les pathologies du temps discret disparaissent. Nous montrons cependant que les régimes stationaires sont les mêmes que ceux de la dynamique discrète. Les simulations numériques semblent montrer que la convergence s'obtient effectivement dans ce cas.
Nous modélisons aussi le centre d'appels d'urgence comme un réseau de files d'attente, prenant ainsi en compte le caractère aléatoire des différentes variables du centre d'appel. Pour ce système, nous prouvons que la dynamique, après une transformation d'échelle, converge vers une limite fluide, qui correspond au système d'équations différentielles de notre modèle de réseau de Petri. Cela conforte notre seconde modélisation. Les principaux outils de la preuve de convergence sont le calcul stochastique pour les processus de Poisson, des formulations en terme de problème de Skorokhod généralisé, ou encore des arguments de couplage.
Ainsi, nos trois modèles d'un même centre d'appels d'urgence définissent un même comportement asymptotique schématique, exprimé comme un système linéaire affine par morceaux, décrivant différentes phases de congestion du centre.

Dans une seconde partie de cette thèse, nous analysons des simulations poussées, prenant en compte les nombreux détails de notre étude de cas. Les simulations confirment le comportement schématique prédit par nos modèles mathématiques. Nous discutons aussi des interactions complexes provenant de la nature hétérogène du niveau 2.

Title : Dynamics of a two-level system with priorities and application to an emergency call center

Keywords : performance evaluation, Petri net, stochastic calculus, discrete event systems, hybrid systems, call centers

Abstract We analyze the dynamics of discrete event systems with synchronization and priorities, by means of Petri nets and queueing networks. We apply this to the performance evaluation of an emergency call center.
Our original motivation is practical. In 2016, a new emergency call center became operative in Paris area, handling emergency calls to police and firemen. The new organization includes a two-level call treatment. A first level of operators answers calls, identifies urgent calls and handles (numerous) non-urgent calls. Second level operators are specialists (policemen or firemen) and handle emergency demands. In this architecture, some calls are qualified as extremely urgent and receive a priority treatment. We are interested in the performance evaluation of bilevel systems corresponding to this general description.
We propose three different models addressing this kind of systems. The first two are timed Petri net models. We enrich the classical framework of free choice Petri nets by allowing conflict situations in which the routing is solved by priorities. The main difficulty in this situation is that the dynamics becomes non monotone.
In a first model, we consider discrete dynamics for this class of Petri nets. We prove that the counter variables of the Petri net are solutions of a piecewise linear system with delays. We investigate the stationary regimes of the dynamics, and characterize the affine ones as solutions of a piecewise linear system, which can be thought of as a system of rational equations over a tropical (min-plus) semifield of germs. Numerical experiments show that, however, convergence does not always holds towards these affine stationary regimes.
The second model is a infinitesimal version of the previous one. For the same class of Petri nets, we introduce a dynamics expressed by differential equations, so that the tokens and time events become continuous. For this differential system with discontinuous righthandside, we establish the existence and uniqueness of the solution. The benefit of this continuous model is that the discrete time pathologies disappear. We show however that the stationary regimes are the same as the stationary regimes of the discrete time dynamics. Numerical experiments tend to show that, in this setting, convergence effectively holds.
We also model the emergency call center described above as a queueing system, taking into account the randomness of the different call center variables. For this system, we prove that, under an appropriate scaling, the dynamics converges to a fluid limit which corresponds to the differential equations of our Petri net model. This provides support for the second model. Stochastic calculus for Poisson processes, generalized Skorokhod problems formulations and coupling arguments are the main tools used to establish this convergence.
Hence, our three models of an identical emergency call center yield the same schematic asymptotic behavior, expressed as a piecewise linear system of the parameters, and describing the different congestion phases of the system.

In a second part of this thesis, simulations are carried out and analyzed, taking into account the many details of our case study. The simulations confirm the schematic behavior described by our mathematical models. We also address the complex interactions coming from the heterogeneous nature of level 2.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France