Conditional Dynamics of Non-Markovian, Infinite-Server Queues

by

Theophane Weber

S.M., Ecole Centrale Paris (2004)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

Signature of Author .....
Sloan School of Management
12 May 2005

Certified by .....
Jérémie Gallien
J. Spencer Standish Career Development Professor
Thesis Supervisor

Accepted by .....
John N. Tsitsiklis
Professor of Electrical Engineering and Computer Science
Co-Director, Operations Research Center

Conditional Dynamics of Non-Markovian, Infinite-Server Queues

by

Theophane Weber

Submitted to the Sloan School of Management
on 12 May 2005, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

We study the transient dynamics of a partially observed, infinite server queue fed with a Poisson arrival process whose controlled rate is changed at discrete points in time. More specifically, we define a state that incorporates partial information from the history of the process and write analytical formula for the dynamics of the system (state transition probabilities). Moreover, we develop an approximation method that makes the state finite-dimensional, and introduce techniques to further reduce the dimension of the state. This method could thus enable the formulation of tractable DPs in the future.

Thesis Supervisor: Jérémie Gallien
Title: J. Spencer Standish Career Development Professor

Acknowledgements

Thanks to my advisor, Jeremie Gallien, for giving me the opportunity of working on this project. His enthusiasm and his idea of what constitutes "good research" have guided me through these two years. He has always been here to motivate me when I felt lost or overwhelmed. Thanks for being a friend and for teaching me so many things.

Thanks to Lou, Brent and Kevin working at the anonymous e-retailer for their help and welcome in their company.

Thanks to the ORC staff, Paulette, Laura, Andrew and Veronica for getting me out of all the administrative problems I got myself into.

Thanks to the many friends that I have met at MIT. Thanks go to the "Tang gang", who are among the first people that I met at MIT: Lauren, Kelly, Srini, Jeremy, Jose, Stefano and others. Thanks to the Frenchies and Italians for bringing a bit of the old continent across the ocean. Special thanks to Flf. Thanks to my classmates who took the quals and generals with me. Thanks to my friends at the ORC: Katy, Margret, Mike, Juliane, Nelson, Hamed, Guillaume, Marion, Nico and all the others. Thanks to my fellow informs officers for making up for all my blunders and absences.

Thanks to Yann and Jean-Philippe for many conversations about research, life and other things, and for reading drafts of this thesis. Thanks to Mohamed for innumerable conversations about just everything, disagreeing with me on every single topic possible and for helping this thesis make the deadline.

A million thanks to my roomates, past or present: Meg, Mara, Quirin, Paolo, Khaldoun, for the conversations at ungodly hours, for letting me eat their food, and for all the things I learnt from them. Thanks to all the girls at the Wellesley French House. Thanks to my friends in

France for welcoming me like a king everytime I came back: special thanks to Alex, Amandine, Vaness, Sarah, Aurelien, Eglantine, Jessica.

Thanks to la petite Katy for so many things.

Last but not least, I would like to thank my parents, my brother Jean-Guillaume and my sister Margot for their love and support through all these years.

Contents

Chapter 1

Introduction

Everyday, several thousands of shipments are processed in the fulfillment centers (FC) of a large e-retailer we have interacted with. The outbound operation, which corresponds to the preparation of customers shipments, is a complex process whose performance critically affects the e-retailers' operational costs. Modeling outbound process using a queueing network, understanding its dynamics and developing a control algorithm for it is the initial motivation behind the research work described in this thesis.

There are four main steps in this process: first, the items of the various shipments are manually picked from shelves that cover most of the fulfillment center. They are then transported to a different part of the facility, where they are sorted according to which customer they are destined to. Finally, the order is manually packed and prepared for shipping.

The shipments are sorted using an expensive system of chutes. Each chute is either idle or assigned to a unique order and will be used to accept only items from that order. The main limitation of the system is that whenever an item corresponding to a new order arrives to the chute system, it should find an idle chute to be assigned to. If it is not the case, the corresponding order cannot be sorted. The unassigned items block other items from reaching the chute system, preventing them to reach and free chutes. While there are some simple methods to remove these blocking items from the conveyor belts and try to restore continuity of the flow in the FC, this event is usually the sign of an unrecoverable situation that will lead into the collapse of throughput; the system fails and has to be stopped for half a day, resulting

in considerable loss of performance for the day. This situation has been called "gridlock" by the engineers of the company.

The release of work in the system is the responsibility of engineers called "flowmeisters" who constantly choose how many orders should be released per unit of time while keeping in mind the trade-off between pushing more flow through the system and being more conservative to avoid gridlock.

It is a complex task for many reasons: the system is complex and highly stochastic, variability occurring at all stages of the process: picking time of the order, congestion of the conveyor belts, efficiency of the sorting team, etc.. Moreover, their is a long lag between the control (release of work in the system), and the parameter they are trying to affect (congestion at the chute facility).

Their decision is based on intuition and experience in the facility. According to the e-retailer's engineers, the balance between throughput and gridlock has not yet been reached: not only do they think gridlock is still reached too often, but also they believe the throughput could also be increased without risk. The company has therefore been very interested in a deeper understanding of the dynamics and phenomenon affecting the system (the process), the end goal being the development of quantitative guidelines for controlling flow in the FC.

While describing the full-fledged model is beyond the scope of this paper, its main characteristic is to be a queueing network whose stations are either Markovian (i.e. with memoryless service time) or non-Markovian (with general service time) with infinite number of servers. While the dynamics of the former are well known, those of the latter are more complex and not completely studied in the literature. In this paper, our objective is to understand in full generality the dynamics of an infinite-server queue fed with a time-varying, closed-loop controlled arrival rate. Such a queue will be called "the system" or "the queue" from now on.

Queues with general service time are complex stochastic systems faced with a phenomenon that we will call hysteresis. The word "hysteresis" has been used in physics to describe the dependence of the evolution of a system not only on its current observable "state", but also on the history of that state. Common examples are magnetic hysteresis or thermal hysteresis (the evolution of the system doesn't depend only on the temperature, but also on the fact that

the temperature is currently increasing or decreasing). In our case, queues with general service times also face hysteresis in the sense that the future evolution of the number of people in queue doesn't depend only on the current number of people in queue (as with a Markovian queue), but also on the history of the process (mainly, the arrival times).

Optimal control of queues with general service time is a hard problem for many reasons. First, the age of the customers in the queues provide information that in the general case should be used for optimal control, requiring in theory a state which dimension is as high as the number of customers in queue. Second, the state of the system is a set of mixed variables: the variables describing the number of customers in queue are discrete, while the variables describing the age of customers in queue are continuous. Because of these continuous variables, the state of the system evolves continuously over time, rendering the use of dynamic programming harder, and forcing us - in the general case - to use the more complex field of stochastic optimal control. Without making specific assumptions, it is not even possible to write closed form expression for the state transitions $P(X(t_2) = s_2 | X(t_1) = s_1,\ u(t) \text{ for } t_1 \leq t \leq t_2)$, making it impossible to write down the optimal control problem. The infinite number of servers in system will prove in our case to be the source of great tractability.

We believe that our contributions are the following:

- We define a framework where the system is partially observed and in which the state of the system is well defined. Our choice of the state is based on two criterion: first, it should not grow too fast with the complexity of the problem, and second, it should incorporate some knowledge of the past process (i.e. take into account the non-Markovian service time)
- For any two states (s_1, s_2), times t_1, t_2, and control policy $u(t)$ for $t_1 \leq t < t_2$, we show how to compute the state transition probabilities $P(X(t_2) = s_2 | X(t_1) = s_1,\ u(t) \text{ for } t_1 \leq t < t_2)$
- We develop a method that approximates the system arbitrarily closely and for which the states s are finite-dimensional for any level of approximation. The method uses Erlang approximation, and to our knowledge, it is the first time the Erlang decomposition is used with a focus on control, and the related results are new.

- We present methods to do the approximation and techniques to further reduce the size of the state-space.

The thesis is organized as follows.

In chapter 2, we present the current status of literature in the domain. In chapter 3, we explain further what our method and approach is. Chapter 4 deals with non-Markovian queues with infinite servers: in 4.1, we present general theorems that will be used in the rest of the paper. In 4.2 we recall the results from [13] concerning the analysis of the unobserved $M_t/G/\infty$ queue (and provide a different proof). In 4.3, we present our first results about the dynamics of the observed $M_t/G/\infty$; we define the state and the compute the state transition probabilities. In chapter 5, we look at a special case that we call the "hidden network". In chapter 6 we show how to use the hidden network to approximate any general system, and mention techniques to reduce the size of the state. Finally, in chapter 7, we conclude by mentioning possible extensions.

Chapter 2

Literature Review

The control of queueing systems is a difficult task for many reasons. For instance, for most queueing control problem, the state-space is infinite if the number of servers or the size of the buffers is infinite; however, most of the time, one can truncate the state-space and remove all states unlikely to be reached without any significant loss of performance.

Models with exponential service times (Markovian) and Poisson arrivals are often used because they offer many simplifications. First, the state of the system only contains the level of the queues. In that case, because of the memoryless property of the exponential distribution, the future of the process solely depends on the number of customers in each queue The arrival times of the customers do not affect the future. Second, the state of such a system is discrete, and changes -discretely- at random times. Using uniformization (see Bertsekas [6] or Gallager [17]), which amounts to scaling time, one can write a discrete-time dynamic program that solves the continuous-time optimal control of the queue under a wide variety of assumptions (controlled service rate, controlled arrival rate, etc..). One can find classical examples of the use of exponential service times in Bertsekas [6], Bertsekas and Tsitsiklis [7].

Other authors use the exponential service times to derive structural results (monotonicity of the policy with the queue length, convexity of the value function, etc..) for various queueing problems (controlled service rate, controlled arrival rate, etc..): see for instance Koole [27]. Stidham and Weber [43] also present general results of monotonicity of the optimal policy for average-cost exponential queues. They also approach general-service times in specific frame-

works and show insensitivity of the optimal policy to the service-time distribution for these special cases. Finally, Liu, Nain and Towsley [29] also present general properties of optimal policies using a different method, the sample path proofs, but once again, they require specific assumptions concerning the problem.

The control of a queue with general service times is a perilous exercise that is rarely tackled in the literature.

For extremely low volumes, one could decide to include the ages of customers in the state of the system, but this approach is very limited from a computational point of view.

By considering specific queues (one server) and restricting the control to certain families, one can find closed form solution for the optimal control. Heynman [23], and later Bell ([4],[5]) look at this type of problems, when the arrivals are Poisson and the policy is restricted to be stopping or re-starting the server at customers arrival or departure. In [42], Sobel looks at a very similar problem, except that inter-arrival times are now drawn from a general distribution; the resulting problem being a controlled random walk. All of these papers find solution because they make specific assumptions about the queue and about the problem.

Another approach is to make approximations of the dynamics of the system: very common approximations are the fluid and the diffusion approximation, which are respectively first order (deterministic) and second order (stochastic) approximations of the system. They correspond to situations where volumes (and utilization) are high. These approximations allow to apply the theory of optimal control much more easily. Optimal control of the fluid approximation of a queueing system is developed by Ricard and Bertsimas in [37]. Several papers in the domain by the same authors can be found. Fleischer and Sethuraman develops approximately optimal control of queues using optimization technique in [16]. Optimal control of the Brownian motion has its theory covered by Harrison in [19], [21] and by Harrison and Van Mieghem in [22]. The current status of the field is surveyed by Harrison and Nguyen in [18]. It is applied successfully in the case of tandem queues (most of the time, 2 queues) by Wein ([36], [44]), Veatch and Wein ([44]) and Reiman and Wein [36]; other control families, such as trajectory tracking, are developed by Maglaras ([31]). For applications of these techniques, see Plambeck ([33] and [34]), Carr and Duenyas [11], or Kushner [28].

In a sense, there are two cases that seem tractable: the low volumes systems, "atomic

level" situations that can be solved using classical dynamic programming, and the high-volume systems solved using optimal control of continuous systems. We believe that our results could be used in a medium to high volume framework, using neuro-dynamic programming techniques (see Bertsekas and Tsitsklis [7], and Pucci de Farias and Van Roy [35]) if necessary to further address the large-scale of the system.

Several papers derive transient dynamic results of infinite-server queueing systems; Eick, Massey and Whitt wrote a seminal paper [13] about the $M_t/G/\infty$ queue. Other examples are given by Nelson and Taafe [32], or Massey [30]. However, most of these paper consider that the system starts in an empty state. If it is not the case, they usually assume we know what the remaining time distribution for customers in system is. Two natural approaches are commonly used (see for instance Duffield [12])

· One assumes we know the age τ_i of the customer i in the system, in which case we can write the conditional residual time distribution using Bayes' rule. This distribution depends on τ_i, which for optimal control would force us to put all the ages τ_i in the state (as previously said), which is untractable.

· The other approach is to assume that the release rate is constant and that the system has attained steady state, in which case it is know that the residual time has attained for all customers some equilibrium distribution. Two reasons at least make this assumption invalid in our setting. First, assuming constant release rate is not possible for a system with controlled arrival rate. Second, making steady state assumptions correspond to a system in open-loop control; a closed loop control is never in steady state at the moment a control is chosen.

In our paper, we present results about what the remaining time distributions should be, given our partial knowledge of the history of the queueing process. This permits to describe what the conditional dynamics of the system are. We also develop an approximation based on Erlang decomposition that allows to have a finite state space for the dynamic programming problem, and try to limit the growth of the state-space as the order of the approximation increases. Keilson and Nunn ([25], [26]) present the mathematical tools for the Erlang basis decomposition, while Schassberger presents the needed approximation proof [41].Whitt and

Abate ([1] and [2]) give practical methods to decompose the service time distributions on the Erlang basis.

Another difficulty that arises in our case is that we are trying to solve a dynamic programming problem with constraints. Several papers deal with this field; an extensive summary and presentation of Lagrange-based methods can be found in Altman [3]; sample-path constrained dynamic programming is considered by Ross in [39] and [40].

Finally, Duffield and Whitt in [12] study a problem fairly similar problem to ours (control and recovery from rare congestion events in a large multi-server systems). While the objective (recovering from congestion) and method (strong use of the infinite-server assumption to derive transient dynamics of the queue) are fairly similar, we believe our paper differs from theirs on many very important points.

Their objective is to be able to recover from a high congestion, while our objective is to avoid it. In their case, the control only comes into play once the failure has happened. Control only occurs "once" (there is no optimal control), and they can assume the system is in steady state upon failure.

To summarize, the strong difference between the two papers is that we are trying to avoid a situation (while optimizing some performance measure) by using optimal control, while they recover from this situation using control policies restricted to certain classes.

Chapter 3

Proposed approach

3.1 Preliminary remarks

From a mathematical point of view, infinite server queues are very useful because they "decouple" their customers (customers don't affect each other as they all are served immediately). From a practical point of view, infinite server queues model well systems with large number of servers, as argued in [12]; obviously, they model even better systems like the one described by Whitt in [49] or the one we study in this paper, as the controller's objective is precisely to be able to immediately serve any arriving customer; making the queue equivalent to an infinite-server one. We call this constraint a "hot potato" constraint by analogy with the communications network literature.

Eick et Al ([13]) derive transient results for the $M_t/G/\infty$. They assume a given schedule of arrival rate $\lambda(t)$ for all $t \geq 0$, and they derive for all $T \geq 0$ a closed form solution for the distribution $N(T)$ of customers in system at time T. We illustrate this in figure 3.1 (only the first two moments are represented for graphical simplicity, but the whole distribution is known).

This result has two limitations:

First, it gives the answer for a given predetermined, endogenous schedule $\lambda(t)$..Even if we were trying to evaluate the performance of some closed-loop policy ($\lambda(t) = f(N(t))$ for instance), we would not be able to compute the steady state of the queue (not to mention the fact that solving an optimal control problem by writing a general form for all policies, evaluating the

performance over this general form and then optimizing over that performance measure is not the best way to solve the problem - it amounts to have a brute force approach in dynamic programming).

Second, it doesn't fully describe the queueing process, as we would require to know the distribution of $(N(T_1), N(T_2), ..N(T_n))$ for any $T_1 < T_2.. < T_n$ for the process to be well defined. The result doesn't describe the correlations between different times, or, said differently, it doesn't describe the conditional dynamics of the queueing process.

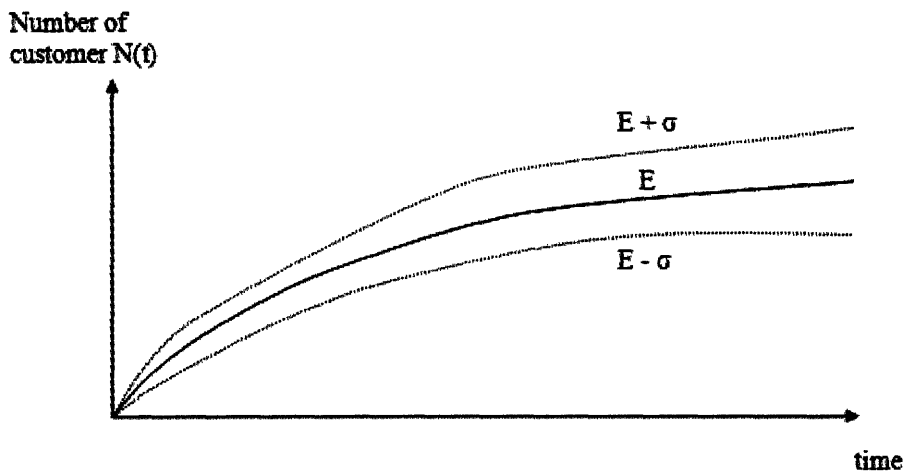


Figure 1

3.2 Proposed Model

We consider an infinite-server queue with general service time. The arrival rate is controlled by a closed-loop feedback of the observed information. The information is constituted by the history of the previous control, and the observation at discrete time steps of the level of the queue. The controlled rate is changed only at these discrete time steps.

Mathematically, if we denote the arrival rate and the level of the queue at time t by $\lambda(t)$ and $N(t)$, the state of the system at time T is $X(T) = (N(T), \lambda(u)$ for $u < T\}$. We will denote the general state-space by Σ, and its elements by s. The control $\lambda(s)$ for $T \leq s < T + \delta$ is constant, function of the state $X(T)$.

At time $T + \delta$, the state of the system $X(T + \delta)$ is observed again and an optimal decision is taken.

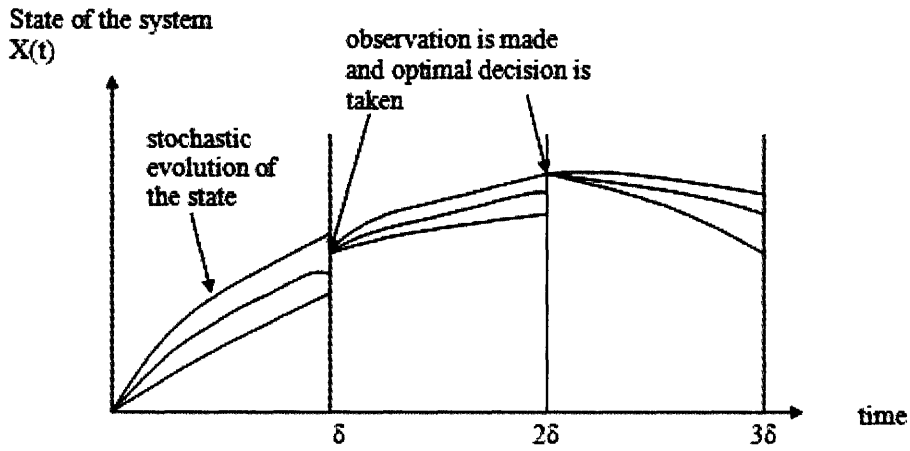


Figure 2

Our objective is the next section is to be able to derive the state transition probabilities for our model. As a remark, the results developed in section 4.3, are **exact**. It is because we are able to compute these conditional dynamics starting from **any** state that we can write and solve the closed-loop control problem (or optimal control with feedback). Because we can take δ as small as we want, we believe that the optimal dynamic programming policy will converge towards the optimal control policy when δ goes to zero.

Chapter 4

Non-Markovian, Infinite-Server Queues

In this chapter, we present the dynamics of queues with an infinite number of servers. "The queue" or "the system" will refer to a queue with general service time and infinitely many servers. The service time is denoted S, its distribution $G(t) = P(S \leq t)$ and complementary distribution $G^c(t) = P(S > t)$.

4.1 Preliminaries: General Poisson splitting

In this part, we present theorems that we will use in the proofs of all the other parts. These theorems are already known and can be found in Ross [38], or Foley [15] for the time-heterogeneous versions.

Unless stated otherwise, we assume that the queue started at time 0.

Theorem 1 (Classical Poisson Splitting) *Let $A(t)$ be a Poisson process with arrival rate λ. Let every arrival be considered of type 1 with probability p and of type 2 with probability $q = 1 - p$, and let $A_1(t)$ and $A_2(t)$ be the resulting processes. Then A_1 and A_2 are Poisson processes with respective rates λp and λq.*

This result is too weak for our purpose; however, this theorem extends to the non-homogeneous case.

Theorem 2 (Time-Heterogeneous Poisson Splitting) *Let $\lambda(t)$ be a continuous real function and $p(t)$ a continuous function such that $\forall t,\ 0 \leq p(t) \leq 1$. Let the $A(t)$ be a Poisson process of intensity $\lambda(t)$ and let every arrival (say, at time τ) be considered of type 1 with probability $p(\tau)$ and of type 2 otherwise. Let A_1 and A_2 be the resulting processes. Then $A_1(t)$ is heterogeneous Poisson with intensity $p(t).\lambda(t)$, and $A_2(t)$ with intensity $(1-p(t)).\lambda(t)$. Furthermore, A_1 and A_2 are independent.*

Theorem 3 (Intensity biasing) *Let $A(t)$ be a time-dependent Poisson process with arrival intensity $\lambda(t)$. Conditioning on the number of arrivals $A(T) = N$, the unordered arrival epochs are i.i.d with probability proportional to the arrival intensity. More precisely, if we denote $f(t)$ the probability density of the arrival time of a customer in system, we have:*

$$f(t) = \frac{\lambda(t)}{\int_{u=0}^{T} \lambda(u).du}$$

If we denote τ the "age" of a given customer in system, and f^a the probability distribution of the age of a customer in queue is, the equation above can be rewritten:

$$f^a(\tau) = \frac{\lambda(T-\tau)}{\int_{u=0}^{T} \lambda(u).du}$$

Moreover, if $A(t)$ is split into $A_1(t)$ and $A_2(t)$ with time-dependent probability $p_1(t)$ and $p_2(t)$, then, conditioning on $A(T) = N$,

$A_1(T)$ is binomial with with number of trials N and parameter

$$p = \frac{\int_{u=0}^{T} \lambda(u).p_1(u).du}{\int_{u=0}^{T} \lambda(u).du}$$

Under special condition, the theorems above can be extended to the case of a system that started infinitely long ago.

For instance theorem 3 has the following generalization:

Theorem 4 *Let $A(t)$ be a time-dependent Poisson process with arrival intensity $\lambda(t)$ that started infinitely long ago, i.e. $\lambda(t)$ is defined on $(-\infty;\infty)$.*

Assume that the total intensity up to time T is finite, i.e. $\int_{u=-\infty}^{T} \lambda(u).du < \infty$.

Then, conditioning on the number of arrivals $A(T) = N$, the unordered arrival epochs are i.i.d with probability proportional to the arrival intensity. More precisely, if we denote $f(t)$ the probability density of the arrival time of a customer in queue, we have:

$$f(t) = \frac{\lambda(t)}{\int_{u=-\infty}^{T} \lambda(u).du}$$

Moreover, if $A(t)$ is split into $A_1(t)$ and $A_2(t)$ with time-dependent probability $p_1(t)$ and $p_2(t)$, then, conditioning on $A(T) = N$,

$A_1(T)$ is binomial with with number of trials N and parameter

$$p = \frac{\int_{u=-\infty}^{T} \lambda(u).p_1(u).du}{\int_{u=-\infty}^{T} \lambda(u).du}$$

Note that theorem 3 is just the special case of theorem 4 in the case of an arrival rate with the following form (which we will call a causal form):

$$\begin{aligned} \lambda(t) &= 0 \text{ for } t < 0 \\ \lambda(t) &\geq 0 \text{ for } t \geq 0 \end{aligned}$$

4.2 The unobserved $M_t/G/\infty$ queue

We now expose the dynamic of the $M_t/G/\infty$ queue found in Eick et Al. [13]. To prove the result they implicitly use the classical Poisson splitting theorem by introducing a Poisson spatial measure on the space $\{(t,s) \in \mathbb{R}^2$,t is a Poisson arrival time, s is a service time sampled from $G\}$.

We present the proof in a slightly different way, using directly the most general forms of Poisson splitting theorems.

The arrival rate is denoted $\lambda(t)$. The service times are i.i.d random variables ; the reference random variable is denoted S. The density of S is denoted $g(t)$, its distribution $G(t)$, and the complimentary distribution $G^c(t) = 1 - G(t)$. $N(T)$ is the number of customers in system at time T.

$D(t_1, t_2)$ is the number of customers exiting system between t_1 and t_2.

If S has a finite first moment, then the equilibrium service time of S is defined as the random variable S_e which has probability density $g_e(t) = \frac{G^c(t)}{E[S]}$, distribution and complimentary distribution: $G_e(t) = \int_0^t \frac{G^c(u)}{E[S]}.du$, $(G_e)^c(t) = \int_t^\infty \frac{G^c(u)}{E[S]}.du$.

In the rest of this paper, we now assume that the process started infinitely long ago and that the following condition is verified at all times T.

Condition 1 (Queue Stability Condition) *For all* T, $\int_{u=0}^{\infty} \lambda(T-u).G^c(u).du < \infty$

The condition is verified, for instance, if the system actually started at time 0 (λ causal), or if λ is bounded and S has a finite first moment. It means that the queue length doesn't blow up to infinity.

We use the following definitions:

· Let $A(t)$ be a Poisson arrival process with intensity $\lambda(t)$. We split the process using the parameter $p(t) = P(S > T - t) = G^c(T - t)$.

Intuitively, $p(t)$ is the probability that an arrival at time t is still in system at time T. The resulting process is called the survival process up to T of an $M_t/G/\infty$ queue.

· Let $A(t)$ be an arrival process with intensity $\lambda(t)$. We split the process using the parameter $p(t) = P(t_1 - t < X \leq t_2 - t)$

Intuitively, $p(t)$ is the probability that an arrival at time t will leave the system between t_1 and t_2. The resulting process is called the leaving process between t_1 and t_2 of an $M_t/G/\infty$ queue.

The processes are, by theorem 2, Poisson processes. However, they are only well defined up to time T for the first, and up to t_2 for the second (they can be defined for subsequent times, but their intensity is always zero after these limits).

Theorem 5 (Eick et Al. [13]) *$N(T)$ is Poisson with mean*

$$\int_{u=0}^{\infty} \lambda(T-u).G^c(u).du = E[\int_{T-S}^{T} \lambda(u)\delta u] = E[\lambda(T-S_e)].E[S].$$

The departure process is a Poisson process with time dependent rate function δ, where $\delta(t) = E[\lambda(t-S)]$.

The proof is given in appendix A.

4.3 Conditional Dynamics: Evolution of the observed $M_t/G/\infty$ queue

Recall the definition of our state $X(T) = \{N(T), \lambda(s) \text{ for all } s < T\}$

We now want to adapt the results of the previous part to a queue which real state has just been observed. The difference between this part and the previous might be more clear if we talk about transition probabilities. If we denote s_0 the "empty state" - $s_0 = (N = 0$, no release history) - then the previous section gives the probability distribution of $X(\delta)$ given $X(0) = s_0$ and some schedule $\lambda(u)$ for $0 \leq u < \delta$. This allows to compute the transition probabilities $P(X(\delta) = s \mid X(0) = s_0, \lambda(u) \text{ for } 0 \leq u < \delta)$.

We will now detail what the distribution of $X(T+\delta)$ is, given $X(T)$ and the schedule $\lambda(u)$ for $T \leq u < T+\delta$.This allows to compute all transition probabilities $P(X(T+\delta) = t \mid X(T) = s, \lambda(u) \text{ for } T \leq u < T+\delta)$ for $(T, \delta) \in (\mathbb{R}^+)^2$, and $(s,t) \in \Sigma^2$.

The time-step δ is the time between observations are made and decision taken.

4.3.1 Definitions and Framework

In this subsection, we give a couple of definitions that will be used in the statement and proof of the main result (the computation of the state transition probabilities). We will state the result in terms of stochastic processes. In the appendix B, we give an alternate proof inspired from [13] that states the result in terms of random variables.

We first show how to "cut" a stochastic process in two parts, one being after the observation time T, and the other, after.

For any given function $t \rightarrow f(t)$, the past function of f is the function $t \rightarrow f^p(t)$ defined by:

$$\begin{aligned} f^p(t) &\triangleq f(t) \text{ for } t &< 0 \\ f^p(t) &\triangleq 0 \text{ for } t &\geq 0 \end{aligned}$$

The causal ("future") function $f^f(t)$ is defined by:

$$\begin{aligned} f^f(t) &\triangleq 0 \text{ for } t &< 0 \\ f^f(t) &\triangleq f(t) \text{ for } t &\geq 0 \end{aligned}$$

Often, the separation will be at some value T instead of zero ($f^p(t) = 0$ for $t < T$) and $f^p(t) = f(t)$ otherwise). To avoid heavy notations, we will specify the separation point beforehand. In the rest of this section, the separation point is at T.

A $M/G/\infty$ queueing process $Q(t)$ is defined by the number of customers $Q(t)$ in a $M/G/\infty$ queue.

The past queueing process $Q^{\rightarrow T}(t)$ ending at T is a $M/G/\infty$ queueing process with arrival rate $\lambda^p(t)$.

The causal queueing process $Q^{T\rightarrow}(t)$ starting at T is a $M/G/\infty$ queueing process with arrival rate $\lambda^f(t)$.

Proposition 1 (Superposition) *Let $Q_1(t)$ and $Q_2(t)$ be independent $M/G/\infty$ queueing process with arrival rate $\lambda_1(t), \lambda_2(t)$. Then $Q_1(t) + Q_2(t)$ is a $M/G/\infty$ queueing process with arrival rate $\lambda_1(t) + \lambda_2(t)$.*

Proof. This simply comes from the fact that for the sum of two Poisson random variables $Q_1(t)$ and $Q_2(t)$ is a Poisson random variable whose intensity is the sum of the two intensities of Q_1 and Q_2. ■

Proposition 2 (Decomposition) *Let's consider a $M/G/\infty$ queueing process $Q(t)$ with arrival rate function $\lambda(t)$. Then $Q(t)$ is the superposition (the sum) of its past and its causal queueing processes (defined at any time T)*

Proof. For any T, we have $\forall t$, $\lambda(t) = \lambda^f(t) + \lambda^p(t)$. We apply proposition 1 and the result follows. ■

The definition of a Bernoulli random variable and a binomial random variable are known: we give their discrete stochastic process equivalent.

Definition 1 (Bernoulli Process) *A Bernoulli process $B(t)$ with parameter G (where G is a distribution) is a discrete stochastic process defined by the following two properties:*

- *It is a point process with one arrival*
- *The arrival time has distribution G*

Said differently, one draws a random arrival time T from the distribution G and B is defined by

$$\begin{aligned} B(t) &= 0 \text{ if } t < T \\ &= 1 \text{ if } t \geq T \end{aligned}$$

Definition 2 (Binomial Process) *The binomial process $B(t)$ with parameters (N, G) (where N is a nonnegative number and G is a distribution function) can be defined either as the superposition of N independent Bernoulli processes with parameter G, or by the following:*

- *It has exactly N arrivals.*
- *The arrival times are i.i.d with distribution G*

Said differently, one draws N iid unordered arrival times $T_i, i = 1...N$ from the distribution G, and B is defined by:

$$B(t) = |\{i \ / \ T_i < t\}|$$

Definition 3 (Depleting binomial process) *A depleting binomial process $B^d(t)$ with parameters (N, G^c) is defined by $B^d(t) = N - B(t)$, where $B(t)$ is the binomial process with parameters $(N, 1 - G^c)$*

Said differently, one draws N iid unordered arrival times $T_i, i = 1...N$ from the distribution G, and B is defined by:

$$B(t) = N - |\{i \ / \ T_i < t\}|$$

Binomial (and depleting binomial) processes have a strong relation with Poisson processes: for instance, a stationary Poisson process $N(t)$ conditioned on $N(T) = N$ is a binomial process between 0 and T, with parameter N and uniform distribution (This is a known result of conditional Poisson processes, and a special case of theorem 3).

4.3.2 The main theorem

We now give the notations of the theorem.

Notation 1 *T is the observation time.*

$T + \delta$ is the prediction time.

The arrival rate to the queue is denoted $\lambda(t)$

N *is the number of customers in queue at time* T.

$N(t)$ *is the stochastic* $M/G/\infty$ *queueing process with rate* $\lambda(t)$.

Φ *is a vector with the complete history of arrival intensity* $\Phi = (\lambda(t), t \leq T)$. *It can be the result of a control policy.*

$N(T+\delta \mid T, \Phi)$ *is the stochastic process representing the number of customers in the queue at time* $T+\delta$, *conditioned on* $N(T) = N$ *and* Φ.

We finally state and prove the result:

Theorem 6 $N(T+\delta \mid T, \Phi)$ *is the sum of the causal* $M/G/\infty$ *process (the "future arrivals" process) and a depleting binomial process ("the current customers" process) with parameters* (N, G), *where* G *is computed by the following formula:*

$$G^c(x) = \frac{\int_{u'=0}^{\infty} G^c(u'+x).\lambda(T-u').\delta u'}{\int_{u'=0}^{\infty} G^c(u').\lambda(T-u').\delta u'}$$

We verify that for λ *=constant, we find* $G^c(x) = G_e^c(x)$

Proof. By proposition 2, $N(t)$ can be written as the sum of its past process up to time T and its causal process from time T.

$$N(t) = N^p(t) + N^c(t)$$

Conditioning on $N(T) = N$ and Φ

$$N(t|T, \Phi) = N^p(t|T, \Phi) + N^c(t|T, \Phi)$$

In particular:

$$N(T+\delta|T, \Phi) = N^p(T+\delta|T, \Phi) + N^c(T+\delta|T, \Phi)$$

The causal process doesn't depend on the state at time T (the causal process from time T, by definition, only depends on the arrival rate after time T). Therefore, $N^c(T+\delta|T,\Phi)$ is equal to $N^c(T+\delta)$

$N^p(T+\delta|T,\Phi)$ is a $M/G/\infty$ queueing process conditioned on the number of customers at time T.

We are going to prove it is a depleting binomial process with the parameters specified above. To do, we use the theorems from the previous section.

$N^p(t)$ is a $M/G/\infty$ queueing process with intensity $\lambda(t_\dagger)$. For this process, all arrivals occur before T.

We consider the survival process up to T of $N^p(t)$. By theorem 2, it is a $M/G/\infty$ queueing process with intensity $\lambda(t_\dagger)G^c(T-t)$

The survival process is then split into two subprocesses

- The first subprocess corresponds to customers gone at $T+\delta$. The corresponding probability, for a customer arriving at time t, is

$$\begin{aligned} p_1(t) &= P(X < T+\delta-t | T-t \leq X) \\ &= \frac{P(T-t \leq X < T+\delta-t)}{G^c(T-t)} = \frac{\{G^c(T-t) - G^c(T+\delta-t)\}}{G^c(T-t)} \end{aligned}$$

  By theorem 2, this is a Poisson process with intensity $\lambda(t_\dagger)\{G^c(T-t) - G^c(T+\delta-t)\}$.

- The second subprocess corresponds to customers still in system at time $T+\delta$. The corresponding probability is $1-p_1(t)$, and by theorem 2, the resulting process is a Poisson process with intensity $\lambda(t_\dagger)G^c(T+\delta-t)$.

We can now interpret $N^p(T+\delta|T) : N^p(T+\delta|T)$ is the number of arrivals of the second subprocess, given that the number of arrivals of the survival process is N. By theorem 4, it is a binomial random variable with parameter

$$p_\Phi(\delta) = \frac{\int_{t=-\infty}^{T} \lambda(t).G^c(T+\delta-t).dt}{\int_{t=-\infty}^{T} \lambda(t).G^c(T-t).dy} = \frac{\int_{t=0}^{\infty} \lambda(T-u).G^c(u+\delta).du}{\int_{t=0}^{\infty} \lambda(T-u).G^c(u).du}$$

■

The theorem above gives describes the conditional dynamics of the system from a stochastic process point of view. It is then easy to give the equations for the random variable point of view:

Corollary 1 *The random variable $N(T+\delta \mid T)$ is the sum of a Poisson random variable and a binomial random variable.*

More precisely, if $poisson(\lambda)$ denotes a Poisson r.v with parameter λ and $B^d(N,p)$ is equal to N minus a binomial random variable with N trials and probability p, we have:

$$N(T+\delta \mid T) = Poisson(\int_0^\delta \lambda(T+\delta-u).G^c(u).du) + B^d(N, \frac{\int_{t=0}^{\infty} \lambda(T-u).G^c(u+\delta).du}{\int_{t=0}^{\infty} \lambda(T-u).G^c(u).du})$$

We may also want to study the departure process of the observed queue. The method is the same as previously, and we state the result without proving it:

Corollary 2 *$D(T+\delta|T)$, the departure process of the queue conditioned on $N(T)=N$ and Φ, is the superposition of a Poisson process and a binomial process.*

At time $T+\delta$, the Poisson process has the following intensity: $\int_{t=0}^{\delta} \lambda(T+\delta-t).g(t).dt$

The binomial process has parameters N and $1-p_\Phi$ (where p_Φ was defined earlier).

4.3.3 Discussion

Concerning the definition of the state, we chose to discard the fine granularity information given by the exact arrival and departure times. However, it would be untrue to say that the controller doesn't have any information on arrivals and departures: the known history of control (the release rate) and the level of queues give information about arrival and departures. There are several possible reasons to make such an assumption: first, there are cases where individual arrivals and departures are actually not observed, and there is no easy way to do so. This is currently the case in the e-retailer's outbound process. Some other times, the observation is possible but the controller chooses to discard the information. Indeed, as volume increases, the relative variability of the system will decrease, and the value of the fine-granularity information of arrival times will decrease as well.

As for the main result, the evolution of the system depends on two stochastic processes, or two random variables.

The depleting binomial process represents the effects of both the past and the present on the future. As the horizon δ goes to infinity, these effects vanish: the depleting binomial process becomes zero with probability one. The parameter of the binomial process depends on the whole history of past release rate. This hysteresis could be a problem to formulate a control problem, as it would make the state of the control grow infinitely. Even in the case where we just predict the evolution of the system, it is an issue to have an infinite dimensional state. We would like to know if there are some special cases where there are "sufficient statistics" of the system, that is, a finite set of parameters that completely characterize the state of the system.

As for the Poisson process, it represents the effects of the future control on future states.

In the next section, we present two results: we are going to prove again theorem 6 in the special case of a certain type of distribution, the hyper-Erlang (mixture of Erlang) distributions. The result still applies, but as added bonus, we prove that there exists a set of sufficient statistics for that system. In a following part, we will argue that every distribution can be approximated by a mixture of Erlang distributions, and provide with a decomposition method to approach

the problem practically. This approximation will be called "the hyper-Erlang approximation" or "hidden network decomposition".

Chapter 5

Hyper-Erlang Service time

We now study a special case of the previous section: instead of assuming a general distribution for the station, we assume that the distribution is an hyper-Erlang distribution. In the rest of the paper, for a given type of distribution (for instance, exponential, Erlang), we call hyper-type (hyper-exponential, hyper-Erlang), as usually done in the literature, the set of any mixture of the distribution. The generalized hyper type (generalized hyper-Erlang) is a mixture which coefficients can be negative.

An Erlang random variable $X_{\mu,j}$ with scale $\mu \in \mathbb{R}^+$ and mode $n \in \mathbb{N}^+$ is equal to the sum of n exponential r.v with parameter μ.

It has the following density:

$$f_{\mu,n}(t) = e^{-\mu.t}.\frac{t^{n-1}.\mu^n}{n!}$$

The complementary distribution is:

$$P(X_{\mu,n} > t_0) = \int_{t=t_0}^{\infty} e^{-\mu.t}.\frac{t^{n-1}.\mu^n}{n!}.dt = \sum_{l=1}^{n} e^{-\mu.t_0}.\frac{(\mu.t_0)^{(l-1)}}{(l-1)!}$$

X has an hyper-Erlang distribution if and only if there exist $P_1..., P_k$ belonging to the probability simplex, and $(\mu, n) \in (\mathbb{R}^+ \times \mathbb{N}^+)^k$ such that:

$$f_X(t) = \sum_m P_m . f_{\mu_m, n_m}(t)$$

The class of hyper-Erlang distribution will be denoted hE; the one restricted to Erlang distributions of mode less than or equal to k will be denoted hE_k.The generalized classes (coefficients P_m can be negative) will be denoted $\widetilde{hE}$ and $\widetilde{hE_k}$.

We now assume that the service time G belongs to hE. There are k Erlang terms in the mixtures; the mixtures parameters are denoted $P_1, .., P_k$. The i^{th} $(i = 1..k)$ has scale μ_i and mode n_i.

The complimentary CDF is therefore

$$G^c(t) = \sum_{m=1..k} P_m \sum_{l=1}^{n_m} e^{-\mu_m . t} . \frac{(\mu_m . t)^{l-1}}{(l-1)!} \tag{5.1}$$

We can then imagine two systems: the first system is a single station queue, which service time distribution is G. The second system is a network of queues composed of k possible lines. The customer who enters the system is branched in line m with probability P_m; after entering line m, the customer goes through a sequence of n_m identical stations with exponentially distributed service time (parameter μ_m).

It suffices to write the service time distribution in the second system to realize that the two systems are equivalent for a given customer. However, it would be inexact to say that the two systems are identical without an additional remark: When a customer enters the equivalent network of queues, it is not possible for a controller to tell where exactly in the system the customer is..

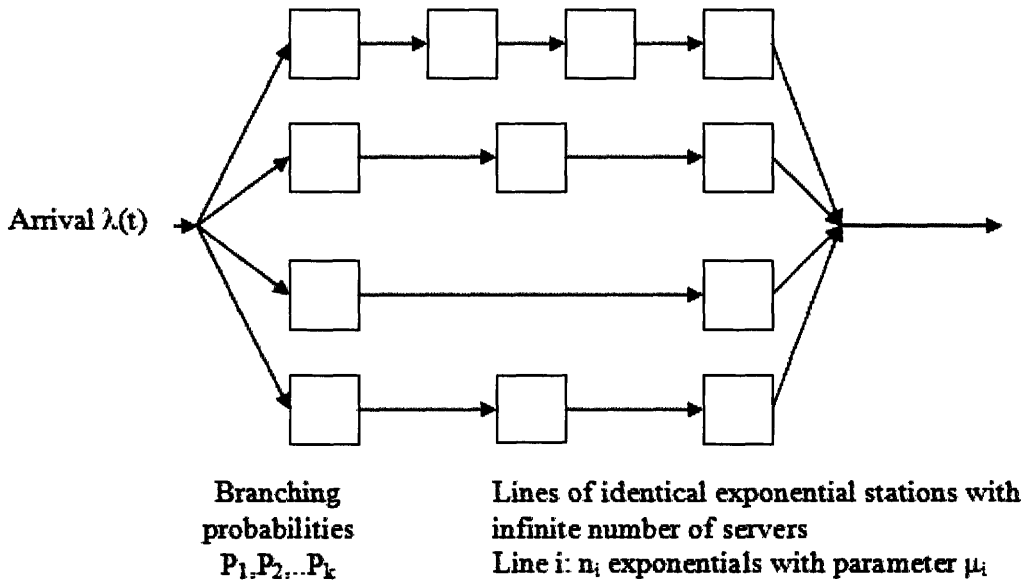


Figure 3

Likewise, if one is able to observe the level of the single station and knows how many customers are in the system, he will not be able to know the level of each individual station in the network. This is the reason why we call the network a "hidden" network.

However, using information about the service times (i.e. the structure of the network and its parameters) and the history of arrival intensity in the system, one is able to build a belief how were customers are in the system. For a customer just arrived in system, we know that he is located in one of the stations at the entrance of the network. Then, as time goes on, we believe that he is moving forward in the system till he finally leaves the system. Conditioning on the only information available, that is, the fact that the customer is still in the network, the belief of the location of the customer "flows" from the entrance to the exit of the graph. In the next section, we define more precisely what this belief is and write the equations of how does it flow in the system.

5.1 Flow equations

We only consider a single customer. He arrives in system at time 0, and we want to know at time τ what is the probability that he is in a given station.

Note that this is strictly equivalent to considering that we are at time 0 and looking at a customer who arrived at time $-\tau$, i.e. looking at a customer whose age in system is τ.

Definition 4 *The probability that at time τ, the customer still in system is in the branch m ($m = 1..k$) at station l ($l = 1..n_k$) will be denoted $P(m, l|\tau)$.*

One important remark is that this is a conditional probability since it is conditioned on the fact the customer is still in system after a time τ.

There are two approaches to solve that problem: one is faster, requires knowledge of Markov processes, and offers slightly less intuition. The second is a more tedious computation that provides more intuition. We quickly present the first method, present the results and prove them in the appendix using the second method.

5.2 The Markov Process approach

One will remark that, for a single customer, the system is also equivalent to a Markov process with $K+1$ nodes, where K transient nodes represent the stations, and a self looping, absorbant node representing the state of having left the system.

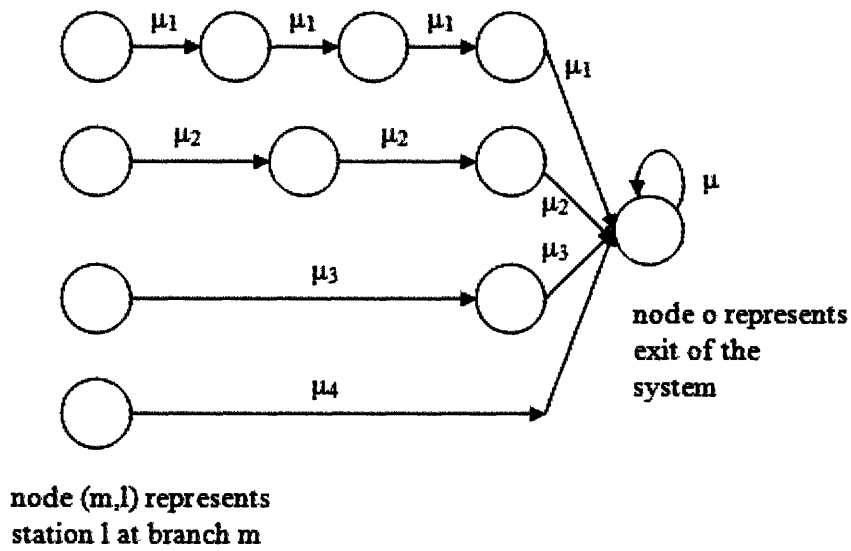


Figure 4

We denote $x(\tau)$ the state of the system at time τ.

The theory of Markov processes tells us the following:

- The structure of the network is described by the rate matrix Q. Q_{ij} is the rate of transition of i into j ; Q_{ii} is the opposite of the sum of all rates exiting from i. In our case, the network has a tree structure giving the rate transition matrix a special structure: every node exits into one node and node only. Node o is the only node to loop into itself. This system is, in some sense, quite deterministic as the path in the Markov chain is entirely determined by the position in the chain. The stochasticity entirely comes from the transition times, and is only present in the Markov process. This "flow" structure of the chain is at the origin of the properties that will be developed in the subsequent sections.

  For a branch i, there are n_i nodes; the n_i^{th} node exits into node o: by an abuse of notation, we also give to node o the label $(m, n_m + 1)$ for every m.

  Then, the transition matrix is described by:

$$\begin{aligned} Q([m,l],[m,l+1]) &= \mu_m \\ Q([m,l],[m,l]) &= -\mu_m \\ Q([m,l],.) &= 0 \text{ otherwise} \end{aligned}$$

- The system dynamics are entirely described by the state transition matrix $P(t)$ defined by $P_{ij}(t) = P(x(t) = j | x(0) = i)$

The state transition matrix is the solution to Chapman-Kolmogorov equation:

$$\begin{aligned} \frac{dP(t)}{dt} &= Q.P(t) = P(t).Q \\ P(0) &= I \end{aligned}$$

The Markov structure of the network gives the very important semi-group structure to the state transition matrices:

$$\forall (t_1, t_2), P(t_1 + t_2) = P(t_1).P(t_2)$$

Also, the probability vector $\pi(t)$ defined by $\pi_i(t) = P(x(t) = i)$ verifies:

$$\pi(t)' = \pi(0)'P(t)$$

- The branching probabilities $P_1, .., P_k$ can be interpreted as a state distribution for the initial state:

$$\begin{aligned} \pi_{(m,1)}(0) &= P_m \text{ (probability of being branched to the first station of branch m)} \\ \pi_{(m,l)}(0) &= 0 \end{aligned}$$

It is now clear that by solving the Chapman-Kolmogorov equation, one can compute the vector $\pi(t)$ for all t. This allows us to compute $P(m,l|\tau)$ by the following reasoning:

$$\begin{aligned} P(m,l|\tau) &= P(\text{customer is in (m,l) at time } \tau | \text{ customer hasn't left system at time } \tau) \\ &= P(x(\tau) = (m,l) \mid x(\tau) \neq o) \\ &= \frac{P(x(\tau) = (m,l)}{P(x(\tau) \neq o)} = \frac{\pi_{(m,1)}(\tau)}{\sum\limits_{m,l} \pi_{(m,1)}(\tau)} \end{aligned}$$

5.3 Direct computation

The location probability of our customer has a closed form solution; computing it does not require knowledge of Markov processes theory, and a proof of the direct computation is given in the appendix.

Proposition 3 *For branch m, station l, time τ, the location probability $P(m,l|\tau)$ is given by the following formula:*

$$P(m,l|\tau) = \frac{Pm.e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}}{\sum\limits_{j=1}^{k}\sum\limits_{i=0}^{n_j-1} P_j.e^{-\mu_j.\tau}.\frac{(\mu_j.\tau)^i}{i!}}$$

We define the location intensity at station l in branch m as :

$$\mu(m,l|\tau) \triangleq e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!} \text{ for } l \text{ greater than 1}$$
$$\mu(m,0|\tau) \triangleq 0$$

Location intensities are linked to the location probabilities by the following:

$$P(m,l|\tau) = \frac{P_m\mu(m,l|\tau)}{\sum\limits_{m,l} P_m\mu(m,l|\tau)} \tag{5.2}$$

We now give two very important result, that directly come from the Markov Process representation of the equation above

Proposition 4 *For all δ, $P(m,l|\tau+\delta)$ can be computed from the set of all $P(m,l|\tau)$ and δ.*

Proposition 5 *For all δ, $\mu(m,l|\tau+\delta)$ can be computed from the set of all $\mu(m,l|\tau)$ and δ. Furthermore, there exists a linear relation between $\mu(m,l|\tau+\delta)$ and the set of $\mu(m,l|\tau)$ and $\mu(m,l|\delta)$.More precisely:*

$$\mu(m,l|\tau+\delta) = \sum_{i=1}^{l} [\mu(m,i|\tau] \,.\, [\mu(m,l+1-i|\delta)] \tag{5.3}$$

The linear operation that updates the location intensity for station (m,l) is denoted $\phi_{m,l}(\delta)$. The overall update is denoted $\delta(\delta)$.

$$\begin{aligned} \mu(m,l|\tau+\delta) &= \phi_{m,l}(\delta)\,[\mu(.)|\tau] \\ \mu(.|\tau+\delta) &= \phi(\delta)\,[\mu(.)|\tau] \end{aligned}$$

Proof. See appendix C ■

This means that if ones knows the set of probabilities $P(m,l|\tau)$, the actual value of τ is not needed to compute $P(m,l|\tau+\delta)$.The fact the relation is linear for $\mu(m,l|\tau)$ will prove very useful in the case where the system is considered as a queue with infinite servers (several customers in system).

5.4 Queueing network equations

For this section, all proofs can be found in appendix C.

We now consider the system as a queueing network: There is a time-heterogeneous Poisson arrival process; upon arrival, every customer is branched to some line of the network, and goes

through all the stations before exiting. There are several customers in the system at the same time, but the only observation we can make is the total number of customers in system. In the spirit of the previous section, we wish to build a "belief" of the location of customers in the system. We will see that in our case, this belief have two important properties:

- We have the exact same location belief for every single customer in system.
- The locations of customers are independent.

These two properties rely on three main factors: the infinite number of servers, the Poisson arrival process, and the fact we only observe intensity history, not individual arrivals.

We first give new definitions; they correspond to location intensities that take into account the history of the past arrival rates.

Definition 5 $\mu^T(m,l) \triangleq \int_0^{+\infty} \lambda(T-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!} = \int_0^{+\infty} \lambda(T-\tau).\mu(m,l|\tau).d\tau$ *is the location intensity at time* T *for branch* m*, station* l*.*

Definition 6 $\mu^T \triangleq \int_0^{+\infty} \lambda(T-\tau).P(X>\tau) = \sum_{m,l} \mu^T(m,l)$ *is the total intensity at time* T*.*

Proposition 6 *The location of all customers in system at time* T *are i.i.d random variable. For any of such customer, we denote*

$$P^T(m,l) = P(\textit{customer in system at time } T \textit{ is in branch } m\textit{, station } l)$$

And the distribution is equal to

$$P^T(m,l) = \sum_{m,l} Pm.\frac{\mu^T(m,l)}{\mu^T} \tag{5.4}$$

One can note the similarity between eqn 5.2 and eqn 5.4.

We see that the location of a customer in system unsurprisingly depends on the history of the past arrival intensity (through the location intensities). This would in theory make the dimensionality of a control problem infinite.

We prove next that the special structure of our problem makes the dimension of the state is in fact finite.

Looking back at theorem 6, we see that to predict the future state, we need to know:

- The future release rates, which depend on future control and are not part of the state
- The current level of the queue
- The probability $p_\Phi(\delta) = \frac{\int_{t=0}^{\infty} \lambda(T-u).G^c(u+\delta).du}{\int_{t=0}^{\infty} \lambda(T-u).G^c(u).du}$ for a customer in system at time T is still in system at time $T+\delta$.

The above probability is the reason why someone should a priori include the history of arrivals in the state.

We can first note that because of the hidden network structure of our service time, we have a simplification.

For every station (m,l) in the queueing network we denote $X^*(m,l)$ the random variable corresponding to the remaining time in system for a customer in station (m,l). For instance, if a customer is in the first station of a branch with three stations with parameter μ, then the random variable is an Erlang$(\mu,3)$. In the general case, using the notations defined previously, we have $X^*(m,l) = X_{\mu_m, n_m - l + 1}$.

We have the following intuitive lemma:

Lemma 1 *The probability for a customer in queue at time T to be still in the queue at time $T+\delta$ is given by:*

$$p_\Phi(\delta) = \sum_{stations} P(station).P(X^*(station) > \delta)$$

This confirms our intuition of interpreting an hyper-Erlang service time as going through the hidden network.

If the controller was able to update the $P^{T+\delta}(.)$ from its current values $P^T(.)$ and the arrival intensity between T and $T+\delta$, there would be no need per se to store the history of arrivals intensity: the $P^T(.)$ parameters would capture everything we need to know about the system. Unfortunately, because of the arrivals between T and $T+\delta$, this can't be done. However, the location probabilities are equal to the normalized location intensities, so it would also be sufficient to update to station intensity $\mu^{T+\delta}(.)$ from $\mu^T(.), \delta$, and the controls between T and $T+\delta$. The next proposition proves this is possible:

Proposition 7 *The location intensities can be updated without knowledge of the past, that is, $\mu^{T+t_0}(m,l)$ can be computed from $\mu^T(m,l)$, t_0 and the arrival intensity between T and $T+t_0$.*

Proposition 8 *More precisely, if we denote:*

$$f_{m,l}(\lambda^\dagger, \delta) = \int_0^\delta \lambda(T+\delta-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}.d\tau$$

Then, we have

$$\begin{aligned} \mu^{T+\delta}(m,l) &= f_{m,l}(\lambda^\dagger, \delta) + \sum_{i=1}^{l} \left[\mu^T(m,i)\right].\mu(m, l+1-i|\delta) \\ &= f_{m,l}(\lambda^\dagger, \delta) + \phi(\delta)\left[\mu^T(.)\right] \end{aligned} \tag{5.5}$$

Remark 1 *The relation above can be differentiated:*

$$\frac{d}{dT}\left[\mu^T(m,l)\right] = \frac{d}{d\delta}\left[\mu^{T+\delta}(m,l)\right]_{|\delta=0} = \lambda(T).\delta(l) + \mu_m(\mu^T(m,l-1) - \mu^T(m,l)) \tag{5.6}$$

We give the following interpretation to the equations 5.5 and 5.6.

The $f_{m,l}(\lambda^\dagger, \delta)$ term represents the "replenishing" of the network with new customers. It is a term that depend on the future arrival rate and on the observation time.

The $\sum_{i=0}^{l-1} \left[\mu^T(m, i+1)\right] . \mu(m, l-i|\delta)$ term represents the movement of the customers in the network.

We mention some quick properties of the two terms, to give some intuition behind the dynamics of the system:

We first discuss properties of the replenishing term:

- $f_{m,l}(\lambda^\dagger, \delta)$ only depends on the values of $\lambda(t)$ for $T \leq t \leq T + \delta$
- $f_{m,l}(\lambda^\dagger, \delta)$ is increasing in the value of λ; that is, for $\lambda_1 \leq \lambda_2$,we have $f_{m,l}(\lambda_1^\dagger, \delta) \leq f_{m,l}(\lambda_1^\dagger, \delta)$
- In the case where $\lambda(t) = \lambda$ for $T \leq t \leq T + \delta$ and δ is small, we have $f_{m,l}(\lambda^\dagger, \delta) \approx \lambda.\delta$.if $l = 1$ and is equal to zero otherwise.

  This corresponds to the fact that a customer enters the system through the first station of each branch.

We now discuss properties of the movement term

- If $\lambda^\dagger = 0$, then: $\mu^{T+\delta}(m, l) \underset{\delta \to \infty}{\to} 0$
- Also, for $\lambda^\dagger = 0$, if $l < n_m$ or if $\mu_m \neq \inf_{j=1..k}(u_j)$ then $P^{T+\delta}(m, l) \underset{\delta \to \infty}{\to} 0$. For $m_s = \arg\min_{j=1..k}(\mu_j)$ $P^{T+\delta}(m_s, n_{m_s}) \underset{\delta \to \infty}{\to} 1$: all customers eventually leave the system, and after an infinitely long time, the customers are in the last station of the slowest branch with probability one.

These properties are easy to prove; for the ones concerning the movement term, we can remark that the update equations are the same as for the single customer case. Asymptotically, we will therefore have $\mu^T(m, l) = \mu(m, l|T)$. Once we have written that the properties follow.

5.5 Dimension and Control of the partially observed network

The previous section states that in the case of an hyper-Erlang service time, there exists a finite set of parameters that is sufficient to know completely what the dynamics of the system are.

Considering that the step δ is some constant time-step, we may write the dynamics equations of the system in the discrete case setting and we will then be in the framework of discrete-time dynamic programming.

We may want to know how complex the problem will be, and compare the complexity of a hidden network to the one of a fully observed network.

Let's suppose we are controlling a network of n Markovian queues and that we can observe the level of each queue. Then, the dimension of the state would be equal to n, and each dimension would represent the level of a queue.

Now, if we control a network of n Markovian queues where we can only observe the total number of customers in system, the previous section states that the state is $(n+1)$ dimensional: one dimension (discrete) represents the total number of customers in queues; the n other dimensions are for the location intensity of each queue. Indeed, we saw that for the queueing network it was not possible to update the location probabilities (which actually are n-1 parameters, not n), only the location intensities. It is therefore impossible to reduce the n location parameters to n-1.

In a sense, we have to pay a "price" for not being able to observe the level of each queue. Until we make further assumptions, this price is double:

- First, the dimension is increased by one
- Second, all dimensions but one become continuous

For example, for a station with service time which is an Erlang distribution with mode 2, the dimension of the problem is 3; one dimension is integer, the two other are continuous.

Chapter 6

Erlang Approximation and Extensions

6.1 Hyper-Erlang decompositions

We have seen that in the case of a $M/hE/\infty$ queue, a finite dynamic programming can be formulated for the control of the queueing system. Our objective is to control a $M/G/\infty$ queue. The natural question that arises is "can hyper-Erlang queueing system be used to approximated any queueing system?". The answer is yes and we will see several theorems to support that intuition.

But first we want to remark the following:

Contrary to lemma 1 developed in the hidden network case, theorem 6 holds in greater generality. For the hidden network, we require the distribution to be a well defined mixture of Erlang distributions, that is, coefficients should be non-negative and sum up to one. However, using the intuition from this theorem, if we wrote the distribution of the general theorem as a generalized mixture of Erlang distributions, the theorem would still hold and the same update equations would work. In others words, we do not really require the coefficients of the mixture

to be non-negative. While having non-negative coefficients helps building an intuition of the system, it is not required to write the dynamic equations in that case.

For that reason, we are interested in fitting our general distribution both using hyper-Erlang distributions and generalized hyper-Erlang distributions.

We now present the theorems that state how can a distribution be approximated using Hyper-Erlang distributions.

Theorem 7 (Schassberger) *Let G be a distribution. There exists a sequence of distribution $G_n \in hE$ such that $G_n \underset{n\to\infty}{\to} G$*

Moreover, the hyper-Erlang distributions can all be chosen with the same scale parameter μ.

The proof of the theorem can be found in [41]. If we denote $G_{\mu,k}$ the CDF of an Erlang(μ, k), the approximating sequence is the following:

$$G_n(x) = G(0) + \sum_{k=1}^{\infty} \{G(k/n) - G((k-1)/n)\}.G_{n,k}(t)$$

The intuition is simple: the Erlang variable with parameters (n, k) has expectation $\frac{k}{n}$ and variance $\frac{k}{n^2}$; $G_{n,k}$ is asymptotically equal to a step function in $\frac{k}{n}$ and it is used to approximate the increase $G(k/n) - G((k-1)/n)$.

This method is unfortunately not very stable as it "compresses" the good approximation terms on the left (for a fixed k, $\frac{k}{n} \underset{n\to\infty}{\to} 0$: if one truncates the sum and increases the order of the approximation, the approximation converges to zero). Very high orders both in the approximation and in the sum are required to have a good approximation.

The fact that the scale parameter can be chosen to be the same for all distributions is very important: we call distributions of this type to be "uniform hyper-Erlang distributions". The set of such distributions is denoted hE^{μ} .The general form of their distribution is

$$\sum_{n\geq 1} p_n.e^{-\mu.t}.\frac{t^{n-1}.\mu^n}{n!}$$

hE_n^{μ} is the set of homogeneous hyper-Erlang distributions which order is limited at n.

If coefficients are allowed to be negative (generalized distributions), the sets are respectively denoted $\widetilde{hE^{\mu}}$, $\widetilde{hE_n^{\mu}}$

Another method (and several others) is presented in [47], [1] and [2]; the methods to truncate the infinite sum are also discussed.

Proposition 9 (Abate, Whitt) *Let $g(t)$ be an analytic density on $[0, +\infty)$. We denote $h(t) = t.e^t.g(t)$ which is also analytic and admits the following development:*

$$h(t) = \sum_{n \geq 1} \frac{h^{(n)}(0)}{n!}.t^n$$

Then g admits the following decomposition:

$$\begin{aligned} g(t) &= \sum_{n \geq 1} h^{(n)}(0).\frac{e^{-t}t^{n-1}}{n!} \\ &= \sum_{n \geq 1} h^{(n)}(0).g_{1,n}(t) \end{aligned}$$

The generalization to another scale parameter is easy: one just has to scale the argument of the function h.

The above proposition is very similar to using a Taylor approximation to approximate a function by a polynomial; this method is known to be sometimes unstable, and we may want to use the following instead:

Lemma 2 *Let $g(t)$ be an analytic density on $[0, +\infty)$. Then there exists a sequence of $g_n(t) \in \widetilde{hE_n^{\mu}}$ such that $g_n(t) \to g(t)$*

Proof. Let's consider the function $h(t) = e^t.g(t)$. From the theory of approximation by polynomials, there exists a polynomial $P_n(t)$ such that $|P_n(t) - h(t)| \leq \frac{1}{n}$ for $t \in [0, n]$.

Then

$$\begin{aligned} |g(t) - P_n(t)e^{-t}| &= e^{-t}|P_n(t) - h(t)| \\ &\leq e^{-t}\frac{1}{n} \leq \frac{1}{n} \text{ for t} \in [0, n] \end{aligned}$$

We choose $g_n(t) = P_n(t)e^{-t}$ and the theorem follows ■

We may also want to fit moments instead; the advantage of the method is that it outputs a finite hyper–Erlang distribution without having to do a truncation operation.

Proposition 10 (Moment fitting for Generalized Hyper-Erlang distributions) *For a distribution G and a $k \geq 0$, denote $\mu^k(G)$ the k^{th} moment the distribution G, ie $\mu^k = \int_{t=0}^{\infty} t^k.dG(t)$.*

Then for some distribution G, there exists for all $\mu \geq 0$ and $n \geq 1$ some $G_n \in \widetilde{hE}_{n+1}^{\mu}$ such that $\mu^k(G) = \mu^k(G_n)$ for all $k \leq n$.

We say that G_n approximates G to the n^{th} order.

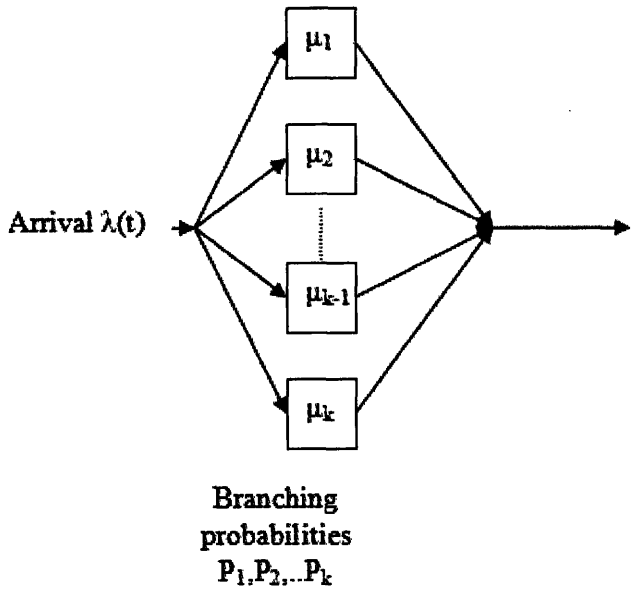
The speed parameter μ can be chosen arbitrarily; in some instances it can be chosen so that $G_n \in \widetilde{hE}_n^{\mu}$ instead.

Proof. See Appendix D ■

6.2 A special case: the hyper-exponential

The previous section focuses on using uniform hyper-Erlang distributions; another approach would be to use only exponential distributions but with different parameters. The corresponding distributions are called "hyper-exponential" and are a special case of hyper-Erlang. The advantage of the method is that for each exponential, the dimension only increases by one.

This corresponds to a hidden network that has the following structure:



Notation 2 *We denote M^d the set of all possible exponential distributions $M^d = \{t \to e^{-s.t} \mid s \in \mathbb{R}^{+*}\}$*

hM^d is the set of mixture of exponential distributions (hyper-exponential distributions).

$\widetilde{hM^d}$ is the set of generalized mixture of exponential distributions.

For a distribution in $\widetilde{hM^d}$, the update of location intensities have the following simple form: (the station number is omitted as it is always 1)

$$\mu^{T+\delta}(m) = f_{m,l}(\lambda^{\dagger}, \delta) + \mu^{T}(m).e^{-s.\delta}$$

The simplicity of the hyper-exponential distribution is attractive, however, how good is this class to approximate functions?

That is what we see next:

Definition 7 *A function f is said to be completely monotonous if is C^{∞} and verifies the following property:*

$$\forall n,\ (-1)^n \frac{d^n}{dx^n} f \geq 0$$

Such a function is therefore positive, decreasing, convex, etc...

A completely monotonous probability measure has the additional property of having finite measure of one.

The corresponding random variable is also said to be completely monotonous.

Lemma 3 *An exponential random variable is completely monotonous; so does any function in* hM^d.

The previous lemma (proving it just requires taking the derivatives of the Laplace transform) shows that a necessary condition for a probability measure to be in hM^d is that it should be a completely monotonous probability measure. It turns out that this condition is also sufficient, and a famous theorem by Bernstein proves it:

Theorem 8 *(Bernstein) Let f be positive function. Then it is the Laplace transform of a probability measure if and only if it completely monotonous and has value one in zero*

Corollary 3 *Let g be a probability measure. Then g belongs to* hM^d *if and only if it is completely monotonous.*

hM^d turns out to be a weak approximating class: only the function with the properties of decreasing exponential can be approximated arbitrarily closely . Fortunately, $\widetilde{hM^d}$ offers more flexibility, and can approximate almost all densities; a result is mentioned in [2]. In particular, moment fitting in $\widetilde{hM^d}$ is fairly easy and involves inversion of Vandermonde matrices.

6.3 A generalization: Phase type distributions

All the approximation class we have mentioned so far are all special cases of a very large approximation class, the phase type distribution.

Every phase-type distribution is described by two things:

- A Markov Process with one unique recurrence class constituted by an absorbing node o; for an example, cf fig 5.2.
- A probability distribution on the nodes of the Markov chain representing the initial state of the Markov process.

The corresponding phase type distribution is the distribution of the time to absorption starting from a random initial state.

The set of phase-type distributions has many properties:

- It is dense in the space of distributions
- It is convex
- Every phase-type distribution also admits a finite set of parameters to describe the state of the system (in the framework of section 6).

A commonly used phase-type distribution family is the set of Coxian distributions. The Markov graph of a Coxian distribution can be seen in fig5 below. Coxian distributions are also dense in the set of continuous distributions, and computing their location intensity can also be done in closed form.

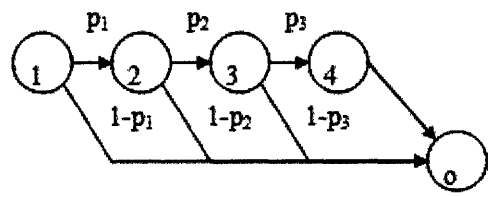


Figure 5

6.4 Uniform speed and dimensionality reduction of our problem

We now make a simple remark that will reduce the dimensionality of our problem in the state-space.

Indeed, we initially supposed in section 5 that the scale parameters of the Erlangs were different; theorem 7 tells us that the same scale parameter can be taken to be the same for all branches. This has an important implication:

Indeed, for two different branches m_1, m_2, the location intensities of the two stations (m_1, l) and (m_2, l) are both equal to

$$\mu^T(m_1, l) = \mu^T(m_2, l) = \int_0^{+\infty} \lambda(T-\tau).e^{-\mu.\tau}.\frac{(\mu.\tau)^{l-1}}{(l-1)!} \triangleq \mu^T(l)$$

The two stations have the same intensity; the intuition behind this result is that because we do not observe the individual level of the queues, we remove unnecessary fine granularity information and can estimate the scaled (by the branching probabilities) number of customers in each queue to be the same.

This reduces the dimensionality of our DP: If we use a distribution $G_n \in \widetilde{hE_n^{\mu}}$ (generalized mixture of n Erlangs of order less than n with the same scale parameter) to fit the distribution G, which means if we use a triangular network (see fig 6-1), then the dimensionality of the station is $n+1$, as opposed to $1 + \frac{n.(n+1)}{2}$.

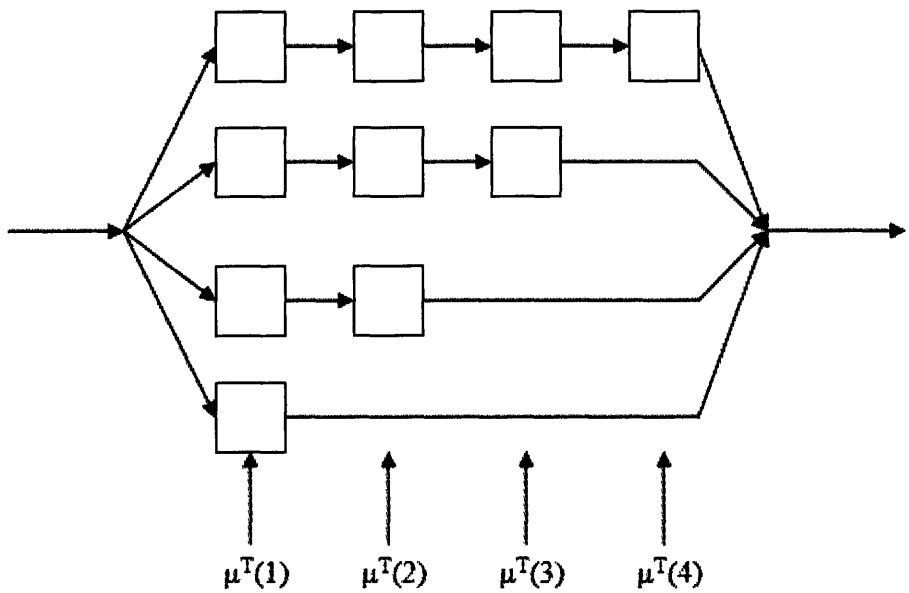


Figure 6-1: Figure 6

Chapter 7

Conclusion

In this paper, we have presented the general dynamics of the $M_t/G/\infty$ queue. We believe that from a theoretical point of view, our method has two benefits: first, it takes into account the fact that the service time has a general distribution. While this is also done in [12] by using the equilibrium distribution, they can only do so because they apply their control "once" on the system, as opposed to controlling the system optimally over the whole time period. Indeed, if we consider in our dynamic program that the probability of exiting a queue was always the same for all customers, it would have been equivalent (by choosing an appropriate parameter) to consider that we had an exponential queue. Second, our method also takes into account some past history of the process, which is important because a $M_t/G/\infty$ intrinsically has hysteresis.

Now that we have equation for the dynamics of the $M_t/G/\infty$ queue, we could write a dynamic program for one $M_t/G/\infty$ queue.

A first extension would be: can we extend it to networks of $M_t/G/\infty$? The answer would be yes if it wasn't for the departure process of an $M_t/G/\infty$, which is not Poisson. Indeed, we have seen it is the superposition of a Poisson process and a depleting binomial process. However, in the case of medium to high volumes, the depleting binomial process will be the superposition of several point processes; the theory of superposition of point processes, as explained by Cinlar in [9], is approximately Poisson. We can therefore make the approximation that the output process will be Poisson, and therefore the theory approximately applies to networks of $M_t/G/\infty$.

A second extension is that all theorems actually apply in the case of $M_t/G_t/\infty$. While it doesn't necessarily make sense to have the service time to depend on time, we can use it with a very special objective.

Suppose we can parametrize a class of distribution G with a parameter s, such that G is stochastically decreasing in s (the higher s is, the lower the service time is)

In section 3, we use the trick of observing the state of the system, changing the release rate and predicting the future state of the system given the state and the control to be able to have a controlled release rate. Because the theorems are still true with G_t distributions, one can use the same trick to change the service rate; it could be a function of the state (for instance if we have congestion dependent service rates, which is the case in the e-retailer's system) and also of a control (changing staff levels for instance). When we have these dynamics, nothing prevents us from writing a very general DP for the following problem: We have a network of $M/G/\infty$ queues (each of them corresponding to one hidden network). For each queue, the arrival process can be the superposition of a controlled arrival process, an exogenous arrival process, and the departure process of some other queue in the network (it can be just one of these processes, or a superposition of the three). For each queue, the service distribution can depend on congestion of the state of the network and also on some control. For every state control (s, u) we incur a cost. Finally, there can be congestion constraints on parts or totality of the network. While the previous problem has great generality (the only two main assumptions are: Poisson arrival processes, and infinity of servers for all queues), it is obvious that it complexity grows very fast if we write it in great generality.

Finally, one may wonder how can a fluid or diffusion approximation be applied to our problem. A first method would be to directly apply fluid approximation (resp. diffusion) approximation to each queue of the real network. One can also apply them to the extended network (the network of hidden networks), for greater modeling power, at the cost of greater complexity.

Appendices

Appendix A: Proof of theorem 5

$N(T)$ is the number of arrivals at time T of the survival process of the queue. Therefore, from theorem 2, it is a Poisson random variable with mean

$$\int_{t=-\infty}^{T} \lambda(t).G^c(T-t).dt = \int_{t=0}^{\infty} \lambda(T-t).G^c(t).dt$$

$D(t_1, t_2)$ is the number of arrivals at time t_2 of the leaving process between t_1 and t_2. It is therefore a poisson random variable with mean

$$\int_{t=-\infty}^{t_2} \lambda(t).P(t_1 - t < X \leq t_2 - t).dt \tag{7.1}$$

If we split the general process into the leaving process between t_1 and t_2, the leaving process between t_2 and t_3, and other arrivals, we can claim that the $D(t_1, t_2)$ and $D(t_2, t_3)$ are independent.

The process has independent increments and the number of arrivals over an interval is poisson distributed: the departure process is therefore a time-heterogeneous poisson process.

To find the intensity, we compute

$$\begin{aligned}
\lim_{h\to 0}\frac{E[D(T,T+h)]}{h} &= \lim_{h\to 0}\frac{\begin{array}{r}\int_{t=-\infty}^{T}\lambda(t).\{G(T+h-t)-G(t_1-t)\}.dt \\ +\int_{t=T}^{T+h}\lambda(t).G(T+h-t).dt\end{array}}{h} \\
&= \int_{t=-\infty}^{T}\lambda(t).\lim_{h\to 0}\frac{\{G(T+h-t)-G(T-t)\}}{h}.dt \\
&\quad +\lim_{h\to 0}\frac{\int_{t=T}^{T+h}\lambda(t).G(T+h-t)}{h}.dt \\
&= \int_{t=-\infty}^{T}\lambda(t).g(T-t).dt+\lambda(T).G(0) \\
&= \int_{t=-\infty}^{T}\lambda(t).g(T-t).dt=\int_{t=0}^{\infty}\lambda(T-t).g(t).dt \\
&= E[\lambda(t-S)]
\end{aligned}$$

Appendix B: Proof of theorem 6 using method from [13]

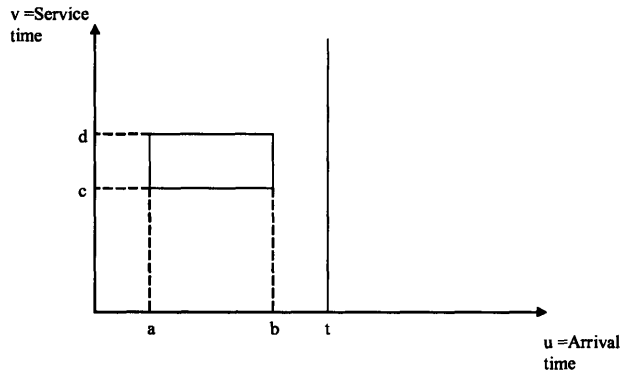
We give a proof a theorem 6 using the random poisson measure method introduced in [13]. This method proves results about the random variables, not the stochastic processes. Some work would be required to give the result about the processes.

In this whole appendix, the separation point is considered to be at T.

We plot in a 2D graph (u,v) a point for every customer arriving in the system. His arrival time is reported on the u-axis, while his service time is reported on the v axis.

Lemma 4 *The queueing process generates a Poisson random measure on the (u,v) graph.*

For all a, b, c and d real numbers, the number of points in the rectangle $(a,b]\times(c,d]$ is Poisson with mean $[G(d)-G(c)]\int_a^b\lambda(u)\delta u$



For a infinitely small rectangle, denoting $N(a,c)$ *number of points in* $(a, a+\delta a] \times (c, c+\delta c]$, *we have:*

$P(N(a,c) = 0) = 1 - \lambda(a).\delta a.\delta G(c)$

$P(N(a,c) = 1) = \lambda(a).\delta a.\delta G(c)$

$P(N(a,c) \geq 2) = o(\delta a.\delta c)$

The integral of the Poisson measure on a certain surface S *be denoted by* $PoissonArea(S)$.

The number of customers $N(T)$ in system at time t is the number of customers for which the relations $u \leq T$ and $u + v > T$ are verified

The number of departures $D(t_1, t_2)$ is the number of customers for which we have the following relation:

$t_1 \leq u + v \leq t_2$.

Using these results and the previous lemma, they prove:

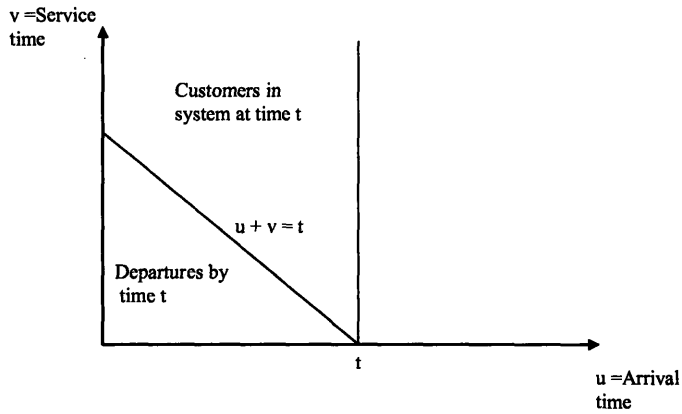
$N(T)$ is poisson with mean $m(T)$, with

$m(T) = E[\int_{T-S}^{T} \lambda(u)\delta u] = E[\lambda(T - S_e)].E[S]$

The departure process is a poisson process with time dependent rate function δ, where

$\delta(t) = E[\lambda(t - S)]$

We now use the following notations:



Notation 3 *T is the observation time.*

$T + \delta$ is the prediction time.

The arrival rate to the queue is denoted $\lambda(t)$

N is the number of customers in queue at time T.

Φ is a vector with the complete history of arrival intensity $\Phi = (\lambda(t), t \leq T)$. It can be the result of a control policy.

$N(T + \delta \mid T)$ is the random variable representing the number of customers in the queue at time $T + \delta$, conditioned on $N(T) = N$ and Φ.

Theorem 9 *$N(T + \delta|T)$ is the sum of a binomial random variable and a poisson random variable.*

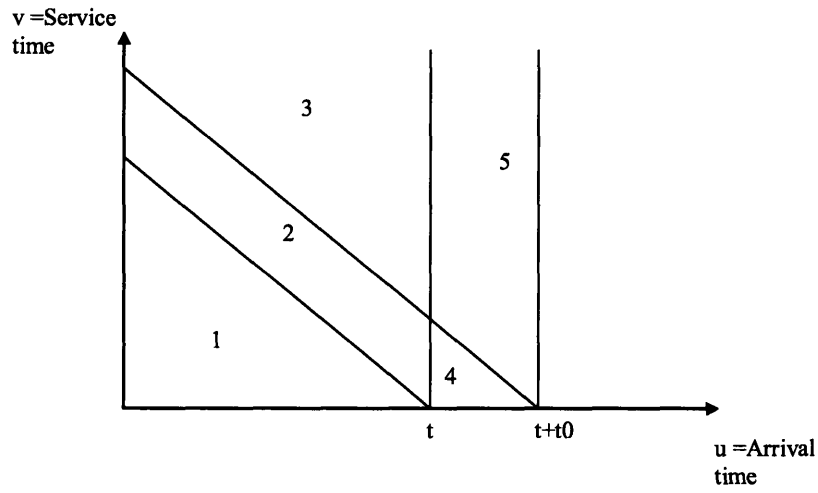
The binomial random variables has N trials and parameter $p_\Phi(\delta) = \frac{\int_{u'=0}^{\infty} G^c(u'+x).\lambda(T-u').\delta u'}{\int_{u'=0}^{\infty} G^c(u').\lambda(T-u').\delta u'}$

The poisson random variable has parameter $E[\lambda(t + t_0 - S^e)^\dagger].E[S]$)

To prove the result, we use the same graphical-probabilistic method as presented before:

In this picture, we have the following definitions for zones:

1: Customers arrived and departed by time T



2: Customers arrived by time T, departed between T and $T+\delta$

3: Customers arrived by time T, still in system at $T+\delta$

2+3: Customers in system at time T

4: Customers arrived and departed between T and $T+\delta$

5: Customers arrived between T and $T+\delta$, departing after $T+\delta$

Knowing $N(T)$, we know the exact number of points in $2+3$, and $N(T+\delta|T)$ is equal to the number of points in $3+5$.

Poisson splitting results say that, conditioned on the value of $2+3$, the number of points in 3 is a binomial random variable, with parameters N and $p = \frac{PoissonArea\ (3)}{PoissonArea(2+3)}$.

The equations of the zone 3 is:

$Zone3 = \{(u,v) \in \mathbb{R}^2 \mid u \leq T,\ u+v \geq T+\delta\}$.

$Zone(2+3) = \{(u,v) \in \mathbb{R}^2 \mid u \leq T,\ u+v \geq T\}$.

$$
\begin{aligned}
PoissonArea(3) &= \int_{u=-\infty}^{t} \left(\int_{v=T+\delta-u}^{\infty} g(v).dv \right) \lambda(u).du \qquad (7.2)\\
&= \int_{u=-\infty}^{t} G^c(T+\delta-u\lambda(u).du \\
&= \int_{u=0}^{\infty} G^c(u+\delta).\lambda(T-u).\delta u
\end{aligned}
$$

$$
\begin{aligned}
PoissonArea(2+3) &= \int_{u=-\infty}^{t} \left(\int_{v=T-u}^{\infty} g(v).dv \right) \lambda(u).du \qquad (7.3)\\
&= \int_{u=-\infty}^{t} G^c(T-u\lambda(u).du \\
&= \int_{u=0}^{\infty} G^c(u).\lambda(T-u).\delta u
\end{aligned}
$$

Combining 7.2 and 7.3, we get the desired result for $p_\Phi(\delta)$.

The number of points in 5 is stochastic and independent by definition of $N(t)$. It is a Poisson random variable; it is also equal, by definition, to the number of points in this zone for the "virtual" system with arrivals $\lambda(t^\dagger)$: and is therefore equal to $Poisson(E[\lambda(T+\delta-S^e)^\dagger]).E[S]$.

$N(T+\delta|T)$ is equal to the number of points in 3 and 5, is therefore the sum of a binomial and a poisson random variables, which achieves the proof.

Appendix C : Proof of the direct computation (Section 5)

Let's recall that we first consider a single customer that has been in system for a time τ (the age of the customer is τ).

The branching belief

For that customer, we denote I_τ the random variable which value is equal to the branching that the customer effectively took when it arrived in system. For instance, if, upon arrival, the customer went on the first (upmost) branch, the value of I_τ is 1. Because this movement is actually not observed, it is a random variable. We are given the probability of branching $P_1..P_k$, giving the initial distribution of $I : I_0$ has the following distribution: $P(I_0 = m) = P_m$

We do learn something from the fact the customer is still in system τ units of time after arriving: for instance, let's consider a system with two branches, one of them being very fast, the other very slow. If after a time τ the customer is still in system, the probability he was in the slow branch is strong.

Using Bayes, we can write:

$$P(I_\tau = m|X > \tau) = \frac{P(I_\tau = mX > \tau)}{P(X > \tau)} = \frac{P(X > \tau \mid I_\tau = m).P(I_\tau = m)}{P(X > \tau)} = \frac{P_m.P(X_{\mu_m,n_m} > \tau)}{\sum P_j.P(X_{\mu_j,n_j} > \tau)}$$

It goes to zero for every branch except the slowest one (highest mean time for the whole branch), for which it goes to 1.

Flow inside a branch

We assume here everything is conditioned on the fact that customer took some branch. We omit the branch number subscript in the parameters for notational simplicity.

The branch has parameter μ and a total of n stations. We want to compute the probability he is in station l, for $l \leq n$.

Lemma 5

$$P(X_1 + .. + X_l = t \mid X_1 + X_2.. + X_n > \tau) = g(\tau, t)$$

where

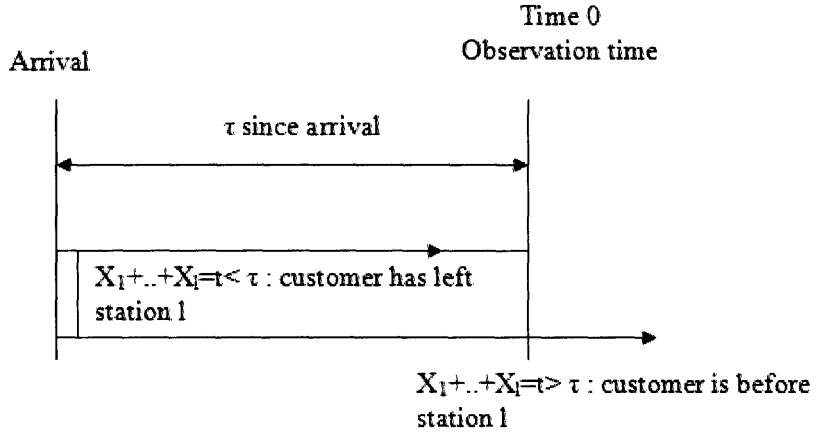
$$g(\tau,t) = \frac{e^{-\mu t}.\frac{t^{l-1}.\mu^L}{l!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^i}{i!}} \; if\ \tau \leq t$$

$$= \frac{e^{-\mu t}.\frac{t^{l-1}.\mu^L}{l!}.\sum_{i=1}^{n-l} e^{-\mu.(\tau-t)}.\frac{(\mu.(\tau-t))^{i-1}}{(i-1)!}}{\sum_{i=1}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i-1}}{(i-1)!}} \; otherwise$$

Proof.

$$P(X_1 + .. + X_l = t | X_1 + X_2 .. + X_n > \tau) = \frac{P(X_1 + .. + X_l = t \cap X_1 + X_2 .. + X_n > \tau)}{P(X_1 + X_2 .. + X_n > \tau)}$$

There are two possible cases for t :



$$P(X_1 + .. + X_l = t \cap X_1 + X_2 .. + X_n > \tau) = P(X_1 + X_2 .. + X_n > \tau | X_1 + .. + X_l = t).P(X_1 + .. + X_l = t)$$

$$P(X_1 + .. + X_l = t \cap X_1 + X_2 .. + X_n > \tau) = P(X_{l+1} + X_2 .. + X_n > \tau - t).P(X_1 + .. + X_l = t)$$

$$\begin{aligned} P(X_{l+1} + X_2 .. + X_n &> \tau - t) \\ &= 1 \text{ if } \tau \leq t \\ &= \sum_{i=0}^{n--l-1} e^{-\mu.(\tau-t)}.\frac{(\mu.(\tau-t))^i}{i!} \end{aligned}$$

Hence:

$$P(X_1 + .. + X_l = t \mid X_1 + X_2 .. + X_n > \tau) = g(\tau, t)$$

where

$$\begin{aligned} g(\tau,t) &= \frac{e^{-\mu t}.\frac{t^{l-1}.\mu^L}{l!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^i}{i!}} \text{ if } \tau \leq t \\ &= \frac{e^{-\mu t}.\frac{t^{l-1}.\mu^L}{l!}.\sum_{i=0}^{n--l-1} e^{-\mu.(\tau-t)}.\frac{(\mu.(\tau-t))^i}{i!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^i}{i!}} \text{ otherwise} \end{aligned}$$

■

Lemma 6 *Defining* $F(l) = P(X_1 + .. + X_l \geq t | X_1 + X_2 .. + X_n > \tau) = \int_{\tau}^{\infty} g(\tau, t).dt$, *we have:*

$$F(l) = P(X_1 + .. + X_l \geq t | X_1 + X_2 .. + X_n > \tau) = \int_{\tau}^{\infty} g(\tau, t).dt = \frac{\sum_{i=0}^{l-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^i}{i!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^i}{i!}}$$

The extreme values are $F(0) = 0$, $F(n) = 1$

Proof. The integral from τ to infinity is easy to compute:

$$\int_{\tau}^{\infty} g(\tau,t).dt = \int_{\tau}^{\infty} \frac{e^{-\mu t}.\frac{t^{l-1}.\mu^{L}}{l!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i}}{i!}}.dt = \frac{\int_{\tau}^{\infty} e^{-\mu t}.\frac{t^{l-1}.\mu^{L}}{l!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i}}{i!}} = \frac{\sum_{i=0}^{l-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i}}{i!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i}}{i!}}$$

■

Lemma 7 *Using the two previous lemmas, we can compute*

$$\begin{aligned} P(l|\tau) &= P(\textit{customer is installation } l \mid \textit{customer is still in system after a time } \tau) \\ &= \frac{e^{-\mu.\tau}.\frac{(\mu.\tau)^{l-1}}{(l-1)!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i}}{i!}} \end{aligned}$$

Proof.

$$\begin{aligned} For\ 1 \leq l \leq n, & \\ F(l) - F(l-1) &= P(X_1 + .. + X_l \geq t | X_1 + X_2 .. + X_n > \tau) \\ -P(X_1 + .. + X_{l-1} &\geq t | X_1 + X_2 .. + X_n > \tau) \\ &= P(X_1 + .. + X_l \geq t \cap X_1 + .. + X_{l-1} < t | X_1 + X_2 .. + X_n > \tau) \\ &= P(\textit{customer is in system } l | \textit{customer has been in system for } \tau) \\ &= \frac{e^{-\mu.\tau}.\frac{(\mu.\tau)^{l-1}}{(l-1)!}}{\sum_{i=0}^{n-1} e^{-\mu.\tau}.\frac{(\mu.\tau)^{i}}{i!}} \end{aligned}$$

■

Finally, using branching probabilities and the lemma above, we can conclude:

Proof of lemma 3.

$$\begin{aligned} P(m,l|\tau) &= P(I_\tau = m|X > \tau).P(l \mid m,\tau) \\ &= \frac{P_m.P(X_{\mu_m,n_m} > \tau)}{\sum P_j.P(X_{\mu_j,n_j} > \tau)} . \frac{e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}}{\sum_{i=0}^{n_m-1} e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^i}{i!}} \\ P(m,l|\tau) &= \frac{Pm.e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}}{\sum_{j=1}^{k}\sum_{i=0}^{n_j-1} P_j.e^{-\mu_j.\tau}.\frac{(\mu_j.\tau)^i}{i!}} \end{aligned}$$

■

We now prove the update equation:

Proof of prop. 4 and 5 .

$$\mu(m,l|\tau+\delta) = e^{-\mu_m.(\tau+\delta)}.\frac{(\mu_m.(\tau+\delta))^{l-1}}{(l-1)!}$$

Using the binomial development:

$$\begin{aligned} \mu(m,l|\tau+\delta) &= e^{-\mu_m.\tau+\delta}.\sum_{i=0}^{l-1}\frac{\mu_m^i\tau^i}{i!}.\frac{\mu_m^{l-1-i}.\delta^{l-1-i}}{(l-1-i)!} \\ &= \sum_{i=0}^{l-1}\left[e^{-\mu_m.\tau}\frac{\mu_m^i\tau^i}{i!}\right].\left[e^{-\mu_m.\delta}\frac{\mu_m^{l-1-i}.\delta^{l-1-i}}{(l-1-i)!}\right] \\ &= \sum_{i=1}^{l}[\mu(m,i|\tau].[\mu(m,l+1-i|\delta)] \end{aligned}$$

The linear operation that updates all the components of the matrix $\mu(.|\tau) = [\mu(m,l|\tau)]_{m,l}$ is denoted $\phi(\delta)$.It corresponds to a fourth order tensor, and we have

$$\mu(.|\tau+\delta) = \phi(\delta).[\mu(.|\tau)]$$

It is linear in the following sense:

$$\begin{aligned}\mu(.|\tau_1+\delta)+\mu(.|\tau_2+\delta) &= \phi(\delta).[\mu(.|\tau_1)+\mu_2(.|\tau_2)] \\ &= \phi(\delta).\mu(.|\tau_1)+\phi(\delta).\mu_2(.|\tau_2)\end{aligned}$$

For the probabilities, one updates the probabilities according to the following method: First, we compute the set of non-normalized probabilities $P'(m,l|\tau+\delta)$

$$\begin{aligned}P'(.|\tau+\delta) &= \phi(\delta).P'(m,l|\tau) \\ P'(m,l|\tau+\delta) &= \sum_{i=1}^{l}[P(m,i|\tau]\,.\,[\mu(m,l+1-i|\delta)]\end{aligned}$$

and then, the probabilities are obtained through normalization:

$$P(m,l|\tau+\delta)=\frac{P'(m,l|\tau+\delta)}{\sum_{m,l}P'(m,l|\tau+\delta)}$$

Because of the normalization step, this operation is non-linear. ■

Queueing network equations

Proof of proposition 6. All the customers in queue at time T correspond to the arrivals of the survival process of the $M/hE/\infty$ queue up to time T.

From theorem 4, their age in queue are i.i.d with distribution

$$f(\tau)=\frac{\lambda(\tau).P(X>T-\tau)}{\int\lambda(t).P(X>T-t)}$$

Because the location of a customer only depends on its age in the queue (no coupling between customers because of the infinite number of servers), the location of the customers in the queue also are i.i.d. The distribution of their location is given by:

$$
\begin{aligned}
P(\text{customer is in station (m,l) at T}) &= \int P(m,l|\tau).f(\tau).d\tau \\
&= \int_{\tau=o}^{\infty} \frac{Pm.e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}}{P(X>\tau)}.\frac{\lambda(T-\tau).P(X>\tau)}{\int \lambda(T-\tau).P(X>\tau)}.d\tau \\
&= Pm.\frac{\int \lambda(T-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}}{\int \lambda(T-\tau).P(X>\tau)} \\
&= Pm.\frac{\mu^T(m,l)}{\mu^T}
\end{aligned}
$$

■

Proof of lemma 1. We have:

$$
\begin{aligned}
p_\Phi(\delta) &= \frac{\int_{t=0}^{\infty} \lambda(T-\tau).G^c(\tau+\delta).d\tau}{\int_{t=0}^{\infty} \lambda(T-\tau).G^c(\tau).d\tau} \\
&= \frac{\int_{t=0}^{\infty} \lambda(T-\tau).\sum_{m=1..k} P_m \sum_{l=1}^{n_m} e^{-\mu_m.(\tau+\delta)}.\frac{(\mu_m.(\tau+\delta))^{l-1}}{(l-1)!}.d\tau}{\mu^T} \\
&= \sum_{m=1..k} P_m.\sum_{l=1}^{n_m} \frac{\int_{t=0}^{\infty} \lambda(T-\tau).e^{-\mu_m.(\tau+\delta)}.\frac{(\mu_m.(\tau+\delta))^{l-1}}{(l-1)!}.d\tau}{\mu^T}
\end{aligned}
$$

We rewrite the $\int_{t=0}^{\infty} \lambda(T-\tau).e^{-\mu_m.(\tau+\delta)}.\frac{(\mu_m.(\tau+\delta))^{l-1}}{(l-1)!}.d\tau$ term using eq 5.3:

$$\begin{aligned}
\int_0^{\infty} \lambda(T-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.(\tau+t_0))^{l-1}}{(l-1)!} &= .\int_0^{\infty} \lambda(T-\tau)\sum_{j=1}^{l}[\mu(m,j|\tau].[\mu(m,l+1-j|\delta)] \\
&= \sum_{j=1}^{l}[\mu(m,l+1-j|\delta)].\int_0^{\infty}\lambda(T-\tau).[\mu(m,j|\tau] \\
&= \sum_{j=1}^{l}\left[\mu^T(m,j)\right].\mu(m,l+1-j|\delta)
\end{aligned}$$

Using the result above:

$$\begin{aligned}
p_\Phi(\delta) &= \sum_{m=1..k}\sum_{l=1}^{n_m}\sum_{j=1}^{l} P_m.\frac{[\mu^T(m,j)]}{\mu^T}\mu(m,l+1-j|\delta) \\
&= \sum_{m=1..k}\left(\sum_{l=1}^{n_m}\sum_{j=1}^{l} P^T(m,j)\mu(m,l+1-j|\delta)\right) \\
&= \sum_{m=1..k}\left(\sum_{j=1}^{n_m} P^T(m,j).\sum_{l=j}^{n_m}\mu(m,l+1-j|\delta)\right)
\end{aligned}$$

$$\begin{aligned}
\sum_{l=j}^{n_m}\mu(m,l+1-j|\delta) &= \sum_{l=j}^{n_m} e^{-\mu_m.t}\frac{\mu_m^{l-j}.t^{l-j}}{(l-j)!} \\
&= \sum_{l=0}^{n_m-j} e^{-\mu_m.t}\frac{\mu_m^{l}.t^{l}}{l!} = P(X_{\mu_m,n_m-j+1}>\delta
\end{aligned}$$

■

Theorem 10 ***Proof.*** *And finally*

$$\begin{aligned}
p_\Phi(\delta) &= \sum_{m=1..k}\left(\sum_{j=1}^{n_m} P^T(m,j)P(X^*(m,j)>\delta\right) \\
&= \sum_{stations} P^T(station).P(X^*(station)>\delta)
\end{aligned}$$

■

Proof of proposition 7.

We have:

$$\begin{aligned}
\mu^{T+t_0}(m,l) &= .\int_0^{\infty}\lambda(T+t_0-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}.d\tau \\
&= \left[\int_0^{t_0}\lambda(T+t_0-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}.d\tau + .\int_{t_0}^{\infty}\lambda(T+t_0-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}.d\tau\right] \\
&= \left[\int_0^{t_0}\lambda(T+t_0-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}.d\tau + .\int_0^{\infty}\lambda(T-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.(\tau+t_0))^{l-1}}{(l-1)!}.d\tau\right]
\end{aligned}$$

The first term is $f_{m,l}(\lambda^{\dagger},\delta)$; the second term is:

$$\int_0^{\infty}\lambda(T-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.(\tau+t_0))^{l-1}}{(l-1)!}.d\tau = \int_0^{\infty}\lambda(T-\tau).\mu(m,l|\tau+\delta)$$

.Using eq 5.3, we can write:

$$\begin{aligned}
\int_0^{\infty}\lambda(T-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.(\tau+t_0))^{l-1}}{(l-1)!} &= .\int_0^{\infty}\lambda(T-\tau)\sum_{i=1}^{l}\left[\mu(m,i|\tau\right].\left[\mu(m,l+1-i|\delta)\right] \\
&= \sum_{i=1}^{l}\left[\mu(m,l+1-i|\delta)\right].\int_0^{\infty}\lambda(T-\tau).\left[\mu(m,i|\tau\right] \\
&= \sum_{i=1}^{l}\left[\mu^T(m,i)\right].\mu(m,l+1-i|\delta)
\end{aligned}$$

■

Proof of remark 1. We denote for the proof $g(\delta)=\mu^{T+\delta}(m,l)$. To obtain the derivative of the function $\mu^T(m,l)$ as a function of T, we have to take the derivative of $g(\delta)$ in zero

$$\frac{d}{d\delta}\left[\mu^{T+\delta}(m,l)\right] = \frac{d}{d\delta}\left[f_{m,l}(\lambda^{\dagger},\delta)\right] + \sum_{i=1}^{l}\left[\mu^T(m,i)\right].\frac{d}{d\delta}\left[\mu(m,l+1-i|\delta)\right]$$

The first term is equal to

$$\begin{aligned}\frac{d}{d\delta}\left[f_{m,l}(\lambda^{\dagger},\delta)\right]_{|\delta=0} &= \lim_{\delta\to 0}\left[\int_0^{\delta}\lambda(T+\delta-\tau).e^{-\mu_m.\tau}.\frac{(\mu_m.\tau)^{l-1}}{(l-1)!}.d\tau\right] \\ &= \lambda(T).\delta(l)\end{aligned}$$

where $\delta(x) = 1$ if $x = 0$ and is equal to 0 otherwise

$$\begin{aligned}\frac{d}{d\delta}\left[\mu(m,i|\delta)\right]_{|\delta=0} &= \frac{d}{d\delta}\left[e^{-\mu_m.\delta}\frac{\mu_m^{i-1}.\delta^{i-1}}{(i-1)!}\right]_{|\delta=0} \\ &= -\mu_m \text{ if } i=1 \\ &= \mu_m \text{ if } i=2 \\ &= 0 \text{ otherwise}\end{aligned}$$

$$\frac{d}{d\delta}\left[\mu^{T+\delta}(m,l)\right] = \lambda(T).\delta(l) + \mu_m(\mu^T(m,l-1) - \mu^T(m,l))$$

This equation is not surprising as the structure of Q appears in it. In a sense, we re-proved in a particular case the existence of the transition matrix Q. ■

Appendix D: Proof of theorem 10

For an Erlang $X_{\mu,n}$ with scale μ and mode n, we have the following moments:

$$\mu^k(X_{\mu,n}) = \frac{(n+k-1)!}{(n-1)!}.\frac{1}{\mu^k}$$

We write the a priori decomposition of a function G_n in $\widetilde{hE^{\mu}_{n+1}}$

$$g_n(t) = \sum_{m=1}^{n+1} P_m.g_{\mu,m}(t)$$

The equations of the moments can be written:

$$\mu^k(G_n) = \sum_{m=1}^{n+1} P_m.\frac{(m+k-1)!}{(m-1)!}.\frac{1}{\mu^k} = \mu^k(G) \text{ for } k = 0..n$$

This can be rewritten:

$$\sum_{m=1}^{n+1} P_m.\frac{(m+k-1)!}{(m-1)!}. = \mu^k.\mu^k(G) \tag{7.4}$$

For all k, one can write $\frac{(m+k-1)!}{(m-1)!} = m^k + P_k(m)$ where $\deg(P_k) < k$

We prove now by induction that there exists some coefficients $b_\mu(k)$ for $k = 0..n$ such that equation 7.4 is equivalent to

$$\sum_{m=1}^{n+1} P_m.m^k = b_\mu(k) \text{ for } k = 0..n$$

The proof works for $k = 0$ with $b_\mu(k) = 1$.

Let's assume it is true for k and prove it is also true for $k+1$. We have:

$$\sum_{m=1}^{n+1} P_m.\frac{(m+k)!}{(m-1)!}. = \sum_{m=1}^{n+1} P_m.(m^{k+1} + P_{k+1}(m))$$

Because $deg(P_{k+1}(m)) \leq k$, $P_{k+1}(m)$ can be written $\sum_{i \leq k} a_{k+1}(i).m^i$, yielding:

$$\begin{aligned}
\sum_{m=1}^{n+1} P_m.(m^{k+1} + P_{k+1,m}(m)) &= \sum_{m=1}^{n+1} P_m.m^{k+1} + \sum_{m=1}^{n+1} P_m.\sum_{i \leq k} a_{k+1}(i).m^i \\
&= \sum_{m=1}^{n+1} P_m.m^{k+1} + \sum_{i \leq k} a_{k+1}(i) \left[\sum_{m=1}^{n+1} P_m..m^i\right] \\
&= \sum_{m=1}^{n+1} P_m.m^{k+1} + \sum_{i \leq k} a_{k+1}(i).b_\mu(i)
\end{aligned}$$

writing $\sum_{m=1}^{n+1} P_m.\frac{(m+k)!}{(m-1)!} = \mu^{k+1}.\mu^{k+1}(G)$ finally yields the equivalent equation:

$$\sum_{m=1}^{n+1} P_m.m^{k+1} = \mu^{k+1}.\mu^{k+1}(G) - \sum_{i \leq k} a_{k+1}(i).b_\mu(i) = b_\mu(k+1) \tag{7.5}$$

which achieves the induction.

We now admit that we have the equivalent equation

$$\sum_{m=1}^{n+1} P_m.m^k = b_\mu(k) \text{ for } k = 0..n$$

This corresponds to finding the solution of a Vandermonde system.

More precisely, if we denote a matrix A with coefficients $A_{ij} = j^{i-1}$ for $i, j = 1..n+1$ and P the vector with coefficients P_m, the equation above can be rewritten:

$$A.P = b_\mu$$

A is a Vandermonde matrix and is therefore invertible; the equation has a solution $P(\mu)$ for all μ; because matrix multiplication is continuous and $b_\mu(k)$ is a continuous function of μ, $P(\mu) = A^{-1}.b_\mu$ is a continuous function of μ.

We now try to reduce the dimensionality by one:

We choose to disregard the first equation $\sum_{m=1}^{n+1} P_m = 1$, and we set $P_{n+1} = 0$. The system of equations for moments greater than one still has a solution by the same argument as before (n coefficients and n equations, for moments from 1 to n).

The corresponding mixture belongs to $\widetilde{hE_n^\mu}$ because $P_{n+1} = 0.P(\mu)$ exists for all μ.

One can also reformulate this result as an inversion for n rows of equation 7.4; and one can write $P(\mu) = C^{-1}d_\mu$ where $d_\mu(k) = \mu^k.\mu^k(G)$, $k \geq 1$.

Now we try to choose μ so that $e'P(\mu) = 1$

We have :

$$\begin{aligned} e'P(\mu) &= e'C^{-1}d_\mu \\ &= \sum_{1 \leq i \leq n} [(e'C^{-1})_i.\mu^i(G)]\mu^i \end{aligned}$$

The above function is a polynomial ; if there is some value for which $e'P(\mu) > 1$ (for instance if $(e'C^{-1})_n > 0$), then, because $e'P(0) = 0$, there exists some μ^* such that $e'P(\mu^*) = 1$ and the mixture corresponding to $P(\mu^*)$ belongs to $\widetilde{hE_n^\mu}$ and fits n moments

Bibliography

[1] Abate, J. and Whitt, W. Infinite series representation of Laplace transforms of probability density functions for numerical inversion (1999). Journal of Operations Research of Japan Vol 42, 268-285.

[2] Abate, J. and Whitt, W. Modeling Service-Time distributions with non-exponential tails: beta-mixtures of exponentials.

[3] Altman, E. Constrained Markov Decision Processes (1995). Rapport de recherche del'INRIA n2574.

[4] Bell, C.E. Optimal Operation of an M / G / 1 Priority Queue with Removable Server (1973), Operations Research, Vol 21, No 6, 281-1290.

[5] Bell, C.E. Characterization and Computation of Optimal Policies for Operating an M/G/1 Queuing System with Removable Server (1971). Operations Research, Vol 19, No 1, 208-218.

[6] Bertsekas, D.P. Dynamic Programming and Optimal Control, vol 1 (2000) and 2 (2001), Athena Scientific

[7] Bertsekas D.P. and Tsitsiklis, J.N. Neuro-Dynamic Programming (1996). Athena Scientific.

[8] Billingsley, P. Convergence of probability measures (1968). Wiley and sons, New York.

[9] Cinlar, E. Superposition of point processes (1972) in Stochastic Point Processes, 549-606.

[10] Dai, J.G. and Nguyen, V. On the convergence of multiclass queueing networks in heavy traffic (1994).

[11] Carr, S. and Duenyas, I. Optimal admission control and sequencing in a make-to-stock / make-to-order system (2000). Operations Research, Vol. 48, No. 5, 709-720.

[12] Duffield, N.G. and Whitt, W. Control and recovery from rare congestion events in a large-multiserver system (1997). Queueing Systems Vol 26, 69-104.

[13] Eick, S.G., Massey, W.A., and Whitt, W. The Physics of the $M_t/G/\infty$ queue (1993). Ops. Res. Vol 41, No 4, 731-742

[14] Eick, S.G., Massey, W.A., and Whitt, W. $M_t/G/\infty$ queues with sinusoidal arrival rates (1993). Management Sci. Vol 39, 241-252.

[15] Foley, R.D. Stationary Poisson departure process from non-stationary queues (1986). Journal of Applied probability Vol 23, 256-260.

[16] Fleischer L. and Sethuraman J. Approximately Optimal control of Fluid Networks. Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, January 12-14, 2003, Baltimore, Maryland.

[17] Gallager, R.G. Discrete Stochastic Processes (1996). Kluwer Academic.

[18] Harrison, J.M. and Nguyen V. Brownian models of multiclass queueing networks: Current status and open problems (1993). Queueing systems : Theory and Applications, vol 13, 5-40.

[19] Harrison, J.M. Brownian Motion and Stochastic Flow Systems (1985). Wiley.

[20] Harrison, J.M. and Wein L.M. Scheduling networks of queues: Heavy traffic analysis of a simple open network (1989). Queueing Systems Theory and Applications 5, 265-280.

[21] Harrison, J.M. Models of open processing networks: Canonical representation of workload (2000). The Annals of Applied probability, Vol 10, No 1, 75-103.

[22] Harrison, J.M and Van Mieghem, J.A. Dynamic Control of Brownian Networks: State-space collapse and equivalent workload formulations.

[23] Heyman, D. Optimal operating policies for M/G/1 queueing systems (1968). Operations Research, Vol 16, 362-382.

[24] Jennings, O.B, Mandelbaum, A., Massey, W.A., and Whitt, W. Server staffing to meet time-varying demand (1996). Management Science, vol 42, no 10, 1383-1394.

[25] Keilson J., Nunn W. H and Sumita U. The bilateral Laguerre transform (1981). Applied Math. and Computation Vol 8 137-174.

[26] Keilson J., Nunn W. H and Sumita U. The Laguerre transform (1979).

[27] Koole, G. Structural results for the control of queueing systems using event-based dynamic programming (1998). Queueing systems, Vol 30, 323-339.

[28] Kushner, H.J. Control and optimal control of assemble-to-order manufacturing systems (1999). Stochastics and stochastic reports, Vol 66, 233-272.

[29] Liu, Z., Nain, P. and Towsley, D. Sample paths methods in the control of queues (1994). Queueing systems: Theory and Application. Special issue of QUESTA on Optimization of Queueing systems.

[30] Massey, W.A, and Whitt, W. Networks of infinite-server queues with non-stationary Poisson input (1993). Queueing Systems: Theory and Appl.Vol 13. 183-250.

[31] Maglaras C. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality (2000). The Annals of Applied Probability vol 10, No 3 897-929

[32] Nelson, B.L. and Taafe, M.R. The Pht/Pht/∞ Queueing system: Part I - The single node(2004). INFORMS journal on computing, Vol 16, No3, 266-274.

[33] Plambeck, E. and Ward, A. Optimal control of high-volume assemble to order systems (2003), Stanford University.

[34] Plambeck, E. Optimal leadtime differentiation via diffusion approximations (2003). Operations Research, Vol. 52, No. 2, 213-228.

[35] Pucci de Farias, D. and Van Roy, B. The linear-programming approach to dynamic programming. Operations Research, Vol. 51, No. 6, p850-865.

[36] Reiman, M.I and Wein, L.M. Heavy traffic analysis of Polling Systems in Tandem.

[37] Ricard, M.J. Optimization of queueing networks; an optimal control approach (1995). Ph.D. Thesis, MIT.

[38] Ross, S.M. Stochastic Processes (1996). Wiley.

[39] Ross K. and Varadarajan R. Markov Decision Processes with Sample path constraints : the communicating case" (1989). Operations Research, 37, pp. 780-790.

[40] Ross K. and Varadarajan R. Multichain Markov Decision Processes with a Sample Path Constraint: A Decomposition Approach (1991). Math. of Operations Research, 16, pp.195-207.

[41] Schassberger, R.S. Warteschlangen (1973). Springer Verlag, Berling.

[42] Sobel, M.J. Optimal average-cost policy for a queue with start-up and shut-down costs (1969). Operations Research, vol 18, 145-162

[43] Stidham, JR. and Weber, R.R. Monotonic and insensitive policies for control of queues with undiscounted costs (1989). Operations Research, Vol 87, 611-625.

[44] Veatch M.H and Wein, L.M. Optimal control of a two-station tandem production/inventory system (1994). Operations Research, Vol 42 No 2, 337-350.

[45] Wein, L.M. Optimal Control of a two-station brownian network (1990). Mathematics of Operations Research, Vol 15 No 2 215-242.

[46] Wein, L.M. Scheduling networks of queues: Heavy traffic analysis of a two-station network with controllable inputs (1990). Operations Research, Vol 38. No. 6, 1065-1078.

[47] Whitt, W. Laplace Transform of probability density functions with series representations (1998).

[48] Whitt, W. Decomposition Approximation for time-dependent Markovian queueing networks (1999). Operations Research Letters Vol 24, 97-103.

[49] Whitt, W. Dynamic staffing in a telephone call center aiming to immediately answer all calls (1999). Operations Research Letters, Vol 24, 205-212.

[50] Wolff, R.D. Stochastic Modeling and the Theory of Queues (1989), Prentice-Hall.